UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI

FACULTAD DE INDUSTRIAS AGROPECUARIAS Y CIENCIAS AMBIENTALES

CARRERA DE COMPUTACIÓN

Tema: "Minería de Datos para mejorar los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi en el año 2022"

Trabajo de Integración Curricular previo a la obtención del título de Ingeniero en Ciencias de la Computación

AUTOR: Chugá Burbano Kevin Anderson

TUTOR: **MSc. Miranda Realpe Jorge Humberto**

Tulcán, 2023.

CERTIFICADO DEL TUTOR

Certifico que el estudiante Chugá Burbano Kevin Anderson con el número de cédula 0402046874 ha desarrollado el Trabajo de Integración Curricular: "Minería de Datos para mejorar los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi en el año 2022".

Este trabajo se sujeta a las normas y metodología dispuesta en el Reglamento de la Unidad de Integración Curricular, Titulación e Incorporación de la UPEC, por lo tanto, autorizo la presentación de la sustentación para la calificación respectiva.

_____

MSc. Miranda Realpe Jorge Humberto
TUTOR

Tulcán, febrero de 2023

AUTORÍA DE TRABAJO

El presente Trabajo de Integración Curricular constituye un requisito previo para la obtención del título de Ingeniero en la Carrera de computación de la Facultad de Industrias Agropecuarias y Ciencias Ambientales.

Yo, Chugá Burbano Kevin Anderson con cédula de identidad número 0402046874 declaro que la investigación es absolutamente original, auténtica, personal y los resultados y conclusiones a los que he llegado son de mi absoluta responsabilidad.

_____

Chugá Burbano Kevin Anderson

AUTOR

Tulcán, febrero de 2023

ACTA DE CESIÓN DE DERECHOS DEL TRABAJO DE INTEGRACIÓN CURRICULAR

Yo, Chugá Burbano Kevin Anderson declaro ser autor de los criterios emitidos en el Trabajo de Integración Curricular: "Minería de Datos para mejorar los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi en el año 2022" y eximo expresamente a la Universidad Politécnica Estatal del Carchi y a sus representantes de posibles reclamos o acciones legales.

_____

Chugá Burbano Kevin Anderson

AUTOR

Tulcán, febrero de 2023

AGRADECIMIENTO

Poder observar los resultados de este proyecto me llena de una inmensa satisfacción y gratitud hacia todas las personas que me han ayudado a alcanzar esta etapa en mi carrera académica. Principalmente quiero darle las gracias a toda mi familia, en especial a mi madre, quien me ha brindado todo el apoyo que he necesitado a lo largo de toda mi educación, me voy a sentir siempre en deuda porque sin su esfuerzo y consejos no habría logrado llegar hasta este punto en mi vida.

Me gustaría darle las gracias a todos los ingenieros que tuve el gusto de conocer durante toda la carrera, de cada uno aprendí cosas nuevas tanto académicas como valores y principios éticos que siempre llevare conmigo. Particularmente quiero darle mi sincero agradecimiento a mi tutor MSc. Jorge Miranda, por la paciencia que tuvo conmigo durante el desarrollo del proyecto, y por mostrar siempre la disponibilidad de su tiempo para brindarme su apoyo, indicaciones y orientaciones en la investigación en base a su experiencia y sabiduría.

Agradecer a mis amigos y compañeros de la universidad, por las alegrías, enfados, risas y muchos momentos que vivimos durante estos años que recordare como una de las mejores etapas de mi vida; de igual forma darles las gracias por los ánimos y consejos brindados durante las últimas etapas del proyecto.

Finalmente, agradezco a la Universidad Politécnica Estatal del Carchi y a la Carrera de Computación, que se convirtieron en un segundo hogar durante los últimos años y me dieron la oportunidad de estudiar y cumplir mis objetivos de formarme como profesional y como persona. Adicionalmente me gustaría agradecer al Ministerio de Turismo de la provincia del Carchi, y en concreto al Ing. Diego García y al Ing. Adrián Quesada por darme el apoyo y la oportunidad de desarrollar mi proyecto de investigación y por la cooperación brindada durante todo el proceso de desarrollo de la investigación.

De verdad, a todos ustedes gracias.

DEDICATORIA

Principalmente este proyecto se lo dedico a mi madre Fátima, porque es la persona que me ha acompañado durante todo mi crecimiento personal y académico, por siempre estar ahí apoyándome incondicionalmente, especialmente en los malos momentos que es cuando más la he necesitado y ha estado ahí; con sus consejos ha logrado convertirme en una persona de buenos sentimientos, hábitos y valores; sin mi madre nada habría sido posible, es mi ejemplo a seguir por el trabajo, lucha, dedicación y sacrificio que ha tenido que hacer para que mi hermano y yo cumplamos nuestras metas; es un orgullo y privilegio ser su hijo, siempre estaré agradecido y en deuda.

De igual forma esta tesis se la dedico a mi hermano Alejo por estar siempre presente, apoyándome moralmente durante esta etapa de mi vida. A mis abuelitos, tías, primas y a toda mi familia por las risas, consejos y ánimos que me han dado en momentos de mi vida que la he pasado mal.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

RESUMEN

La presente investigación denominada "Minería de Datos para mejorar los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi en el año 2022", tiene como objetivo principal crear un modelo de minería de datos para los procesos de control de la demanda turística en el ministerio de turismo de la provincia del Carchi, que sea capaz de mejorar los procesos de control que utiliza esta institución pública. Para dar cumplimiento a los objetivos planteados, se utilizó una metodología con enfoque investigativo, la investigación presento dos enfoques cualitativo y cuantitativo; en conjunto con estos enfoques se utilizó la investigación de exploración, documental, de campo, mediante la cuales se obtuvo características particulares relacionados a los procesos de alojamiento y gasto turístico como los tiempos de ejecución, el manejo y almacenamiento de la información y los usuarios que intervienen en los procesos. Para la recolección de toda esta información se aplicó una entrevista dirigida al Analista de Desarrollo y Promoción Turística de la Zona N°1 del Ministerio de Turismo y una encuesta dirigida a los sitios de alojamiento de la provincia del Carchi. Toman en cuenta los resultados conseguidos se desarrolló una propuesta utilizando la metodología CRISP-DM, donde se definió las actividades que involucran la creación de un modelado de minería de datos incluyendo aspectos técnicos e instrumentos de más utilidad. Por medio de un estudio de factibilidad se determinó los medios necesarios con los que cuenta el ministerio de turismo de la zona N°1 para adoptar la construcción de un proyecto de minería de datos. Finalmente, con relación a un aspecto más técnico se desarrolló una técnica de modelado de agrupamiento o clustering, con una base de datos en Microsoft Excel y el software Knime Analytics para construir el modelo, se utilizó la plataforma Power BI que permitió evaluar y analizar los datos resultantes del algoritmo K-means, técnica de modelado dedicada al agrupamiento o clustering. Con el uso de estas plataformas y algoritmos para crear un modelo de minería de datos y fusionando con las actividades de CRISP-DM, se obtuvo una documentación organizada que puede ser utilizada como referencia en proyectos futuros de investigación.

Palabras Claves: minería de datos, algoritmos, procesos de control, CRISP-DM.

ABSTRACT

The present research work named Data Mining to improve the control processes of tourist demand in the Ministry of Tourism of the Province of Carchi in the year 2022 aims to create a data mining model for the use of this public institution. To accomplish with the goals, it was applied the qualitative and quantitative approaches as well as the exploratory, documentary and field research to obtain particular characteristics related to the processes of accommodation, tourist expenses in the execution times, management and storage of information and the users involved in the processes. For data collection, an interview was performed to the Tourism Development and Promotion Analyst of Zone No. 1 of the Ministry of Tourism and a survey addressed to accommodation sites in the province of Carchi. Taking into account the results, a proposal was developed using the CRISP-DM methodology. Therefore, there were defined the activities that are involved in the creation of a data mining model including technical aspects and more useful instruments. Through a feasibility study, there were determined the necessary means available to the ministry of tourism of zone No. 1 to develop the construction of a data mining project. Finally, a clustering modeling technique was developed with a Microsoft Excel database and Knime Analytics software to build the model. Moreover, the Power BI platform was used, which allowed the evaluation and analysis of the data resulting from the K-means algorithm, a modeling technique dedicated to grouping or clustering. The use of these platforms and algorithms allow creating a data mining model. Also, merging with the CRISP-DM activities, an organized documentation was obtained that can be used as a reference in future research projects.

Keywords: data mining, algorithms, control processes, CRISP-DM.

INTRODUCCIÓN

Para el Estado Ecuatoriano el turismo es de gran relevancia debido a que es la fuente de ingresos económicos no petroleros más importante después de las exportaciones. Actualmente el visualizador del MINTUR recoge datos de cada provincia del país que corresponden directamente a demanda turística del Ecuador. Existe una gran problemática para obtener la información turística del país y es que en ciertas zonas recogen la información de forma tradicional y en algunos casos existe pérdida de datos que genera un error en él análisis de información con relación a la demanda turística.

El estudio tuvo como objetivo principal la propuesta de un modelo de minería de datos, mediante el cual buscaba mejorar los procesos de control que emplea el Ministerio de Turismo para la demanda turística de la provincia de la Carchi, la investigación inicio con bases en las necesidades que tenían las personas involucradas en estos procesos, y a partir de un primer encuentro, ir comprendiendo como se manejaban los procesos, su tiempo de ejecución, la manera de almacenar la información y si se realizaba un análisis de la información. Todo ello con el fin de conocer la realidad de cómo se gestionaban estos procesos para lograr tener una visualización inicial del problema.

El uso de un enfoque cualitativo y cuantitativo fue necesario dentro de la investigación, debido a la recolección de información que se requirió hacer, con el fin de formar un marco teórico y metodológico, que sirvió de referencia para la construcción de un modelado de minería de datos basándonos en los procesos que emplea el ministerio de turismo. El enfoque mixto de la investigación dio la posibilidad de aplicar técnicas e instrumentos mediante los cuales se generaron nuevos criterios sobre los procesos de la demanda turística, y se pudo observar detalladamente como se gestionaban estos procesos, además de ello se obtuvo un nuevo panorama de la investigación, entre otros aspectos.

Todo este proceso que se aplicó dio como resultado a una propuesta de modelado de minería de datos para un tener la posibilidad de mejorar los procesos que emplea el ministerio de turismo para determinar la demanda turística de la provincia del Carchi.

I. EL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

El Ministerio de Turismo del Ecuador cuenta con un visualizador donde proporciona la información turística que ha tenido el país, este tipo de sistema sirve como apoyo para el gobierno con el fin de brindar información acerca del turismo que se ha generados en el país.

Para el Estado Ecuatoriano el turismo es de gran relevancia debido a que es la fuente de ingresos económicos no petroleros más importante después de las exportaciones. Actualmente el visualizador del MINTUR recoge datos de cada provincia del país, donde muestra la información económica, entrada y salida de viajeros, ventas y recaudación del IVA de las actividades turísticas, entre otras acciones. Toda esta información está a disponibilidad de la ciudadanía, ya sean estudiantes u empresarios e incluso a turistas extranjeros y nacionales.

Existe una gran problemática para obtener la información turística del país y es que en ciertas zonas recogen la información de forma tradicional y en algunos casos existe pérdida de datos que genera un error en él análisis de información con relación a la demanda turística, otro punto problemático es que no le dan el valor que debe tener la información, debido a que les puede servir como guía e incluso ayuda para gestionar mejor estos procesos.

Por otra parte, el Ecuador y el mundo se vio atormentado por el coronavirus SARS-CoV2, causante de la enfermedad COVID-19. La propagación de este virus comenzó a finales de 2019 y, ha ocasionado crisis en diferentes países y aunque su incidencia ha bajado en los últimos dos años, dejo grandes impactos negativos en el Ecuador y uno de ellos es la crisis económica ocasionada generando una gran pérdida e incertidumbre hacia el futuro. Desde este punto de vista la información y datos que dejan los procesos para la demanda turística y distintos datos que obtiene el MINTUR a través de distintas plataformas se ha vuelto muy importante y de gran relevancia para el Ecuador, debido a que el turismo es la tercera fuente de ingresos económicos no petroleros en el país, por esta razón tener un conocimiento de los procesos de

gestión de la demanda turística de las distintas provincias, puede ayudar a entender cómo está la situación actual del turismo y cómo potenciar los procesos de la demanda turística, incluyendo predicciones sobre el turismo en un futuro. Con el fin tener bases sólidas para tomar decisiones y mejorar turismo en el Ecuador.

En la Zona de 8 del MINTUR del Ecuador, a principios del año 2021 surgieron problemas relacionados con los procesos de gestión de los servicios turísticos, siendo más específico se explica como una inadecuada organización y forma de almacenamiento de información de los procesos de gestión turística de esta zona, y lo que ocasionaba es que no se lograba estimar la operación e intermediación turística hasta que surgió la opción usar una plataforma que permita realizar los registros de; recategorización, reclasificación, actualización y reingreso de establecimientos en las actividades de alojamientos, alimentos y bebidas. La directora de la Zona 8 Troya Niza menciona que: "esta plataforma permite estar a la vanguardia de las Tecnologías de la Información y la Comunicación, creando la oportunidad de facilitar los trámites ciudadanos y tener un mejor control de la información y tener base de conocimiento sobre la cual tomar decisiones en base a la demanda turística " (Ministerio de Turismo, 2021, pág. 1).

En la Zona sur del Ecuador con el apoyo de diferentes socios entre ellos el Ministerio de Turismo y el Municipio de Loja, y con el fin de generar información turística necesaria para una toma de decisiones adecuadas de diferentes agentes económicos implicados en el sector turísticos se crea en el año 2016 un Observatorio Turístico. Este proyecto apoya el desarrollo de la industria turística que busca fomentar la práctica de turismo sostenible, a través de la construcción de una base de series estadísticas, se pretende establecer un marco de referencia para el análisis sistemático de la situación real y tendencias de la industria turística. Los datos que colecta este sistema son indicadores de alojamientos, perfil de visitantes, gasto turístico, motivación del viaje, nivel de satisfacción del turista.

Actualmente en la Zona 1 del Ministerio de Turismo que corresponde a Carchi Ecuador, cuentan con procesos para los feriados nacionales que tiene por objeto cuantificar el número de viajes realizados por turistas y excursionistas, además de estimar el gasto efectuado durante ese feriado. Estos procesos de alojamiento y gastos turísticos se los lleva de forma tradicional, se explica que; para los 30 lugares de alojamiento que posee la provincia del Carchi, los coordinadores o propietarios

de cada lugar tienen que llenar un documento con diferentes aspectos de información y luego ser enviada al Ministerio de Turismo, de este distrito.

Según el responsable de dirección de esta Zona Diego García explica que en ciertas ocasiones por parte del MINTUR han tenido que trasladarse a los lugares para hacer llenar la información, e incluso han ocurrido errores en los datos por motivos de que algunos coordinadores de los alojamientos han lleno datos incorrectos sobre la demanda turística que ha tenido el lugar.

Toda esta información y datos que son recopilados por parte de la Zona 1 del MINTUR deben ser analizada empíricamente para generar reportes generales e individuales sobre el alojamiento y gastos turístico en feriados nacionales que ha tenido el lugar. Estos procesos realizados muestran la demanda turística que ha tenido la provincia en las diferentes temporadas del año, el MINTUR de la provincia del Carchi no usa ni posee ningún tipo de nuevas herramientas tecnológicas que sirvan de apoyo para gestionar y administrar los procesos que determinan la demanda turística, siendo esto una grave desventaja para el sector turístico de esta provincia ya que la información y el tratamiento de los datos que se le puede dar son vitales para una toma de decisiones e incluso la optimización de recursos. Se debe tomar en cuenta que el sector turístico por su rápido crecimiento ha generado gran demanda en cuanto al uso de las Tecnología de Información y Comunicación debido a que necesitan información fiable que se convierta en una herramienta de gestión turística que aporte a la toma de decisiones y a la generación de nuevas políticas en turismo.

Uno de los problemas fundamentales dentro del MINTUR de la Zona N°1 es la forma en la que se almacena la información, en caso requerir la información de proceso de años anteriores no se va a tener lista, ya que se encuentra en documentos físicos, en grandes archivadores, y cabe mencionar que no existe ningún respaldo de los datos, no se puede contar con información rápida y eficaz de los procesos que determinan el nivel de turismo de la provincia. Por otra parte, para el análisis y estimación de la demanda turística de la provincia del Carchi no cuentan con un algún tipo de análisis de los datos o análisis estadístico para los datos de los procesos de la demanda turística de la provincia; así como la ausencia de indicadores que muestren el nivel de turismo que pueda tener la provincia.

1.2. FORMULACIÓN DEL PROBLEMA

El escaso uso de minería de datos provoca un ineficiente análisis de información lo que genera un inadecuado manejo sobre los procesos de control que emplea el Ministerio de Turismo para determinar la demanda turística en la provincia del Carchi en el período 2022.

1.3. JUSTIFICACIÓN

El sector turístico definido como una actividad económica tiene mucha dependencia de las tecnologías de información. En el año de 1960 se implementó el primer sistema de información de reservas de avión, la evolución que han tenido las tecnologías de la información y de la comunicación ha sido con el objetivo de mejorar diferentes ámbitos en distintos sectores como es el turismo, convirtiéndose en una herramienta de apoyo, para el desempeño de distintas funciones que son realizadas con efectividad y eficiencia (Miralbell, Lamsfus, Miquel, & Gonzáles, 2018, pág. 1).

Las ventajas que tienen las nuevas tecnologías dentro del sector turístico van desde el incremento de la competitividad, reducción de errores, optimización de recursos y generar nuevas funcionalidades que apoyen a distintos procesos, y no solo sirven de apoyo dentro del área turística, si no también son incuestionables en cualquier sector ya sea dentro de la medicina, agricultura, industrias de todo tipo, construcción, entre otros. Dentro del sector turístico para el desarrollo potencial de la informática y las comunicaciones se debe tomar en cuenta las tendencias actuales de las tecnologías de la información, las nuevas herramientas que se aplican en sectores como el área turística son con el fin de mejorar las calidad en dos vertientes, por un lado generar un ahorro costos y optimar procesos (mejora de gestión), y el otro aspecto es mejorar las condiciones de servicio y la incorporaciones nuevas funcionalidades (mayor satisfacción al cliente) (Valles, 2017, pág. 9).

El uso correcto de las nuevas tecnologías dentro del turismo en los últimos años, se han visto influenciados directamente en áreas sociales y económicas, se puede explicar que la influencia que ha tenido las TIC en el turismo está siendo positiva, han aportado beneficios esenciales, como es el incremento y mejora de los flujos de información y datos, recordando que la información es de gran relevancia para la actividad turística. Los proyectos de tecnologías de la información aplicadas al turismo se están desarrollando en varias líneas de actuación entre las más importantes

se destacan; Sistemas de Información, Bases de Datos Multimedia, Redes Especializadas, Minería de Datos, entre otros (Valles, 2017).

En la provincia del Carchi en el MINTUR para cumplir los procesos de control que determinan el nivel de turismo que se ha tenido, tienen que examinar la información recopilada y realizar un análisis estadístico para completar el proceso; actualmente todo se lo hace de forma manual, es un punto importante por mejorar; y es aquí donde la incorporación de técnicas como la Minería de Datos podrían ayudar ya que, se tendría la posibilidad de convertir esos datos en información útil, por medio de una adecuada gestión de la información, al igual que un análisis y evaluación de datos, mejorar el funcionamiento de los procesos de control de la demanda turística.

Las Tics y el sector turístico continuamente están creando nuevas oportunidades de desafíos técnicos y empresariales. Cada vez más las nuevas tecnologías de información y comunicación han incrementado su presencia en el turismo y en las organizaciones que manejan como el Ministerio de Turismo de Ecuador. Con el objetivo de contribuir con la construcción de proyectos que se consoliden y transparenten con la data que se genera, en el sector turístico, esta organización se ha visto en la necesidad de construir un Observatorio Turístico Nacional. Este observatorio ofrece diversas iniciativas dispersas, tanto públicas como privadas. Con el apoyo del Ministerio de Turismo en temas técnicos y metodológicos puedan construir un sistema de información cuantitativa y cualitativa, que sirva de insumo para toma de decisiones estratégicas por parte de todos los actores del sector turístico (MINTUR, 2020, pág. 19).

Si hablamos de nuevas herramientas tecnológicas dentro del sector turístico se debe explicar el beneficio que tiene la Minería de Datos, el uso potencial del Data Mining permite a las organizaciones o empresas identificar nuevas oportunidades les ayuda a potenciar su productividad, y gestionar de forma eficaz y eficiente las operaciones que llevan dentro de una organización. El objetivo de la minería de datos es producir un nuevo conocimiento para que los usuarios de un negocio o instituciones que utilicen esta herramienta puedan tomar decisiones, a partir de una base de construcción de una modelo del mundo real y basándose en datos de distintas fuentes. El modelo de Minería de Datos que se usa siempre va a estar condicionado por los objetivos que se plantea.

La incorporación de nuevas herramientas tecnológicas como es la Minería de Datos puede estar inmersas en distintos sectores, incluso en la gestión de los procesos que tiene el sector turístico, debido a que reducen costos, incrementan ventas, mejoran la productividad, aumentan las oportunidades de negocio, mejoran la producción, ayuda a tomar decisiones e incluso generan fuentes de empleo entre otros beneficios.

En el Ministerio de Turismo de la provincia del Carchi, existen problemas que pueden ser solventadas con el uso de estas nuevas tecnologías, el aporte que le puede dar la Minería de Datos al MINTUR de esta provincia, varía dependiendo de la necesidad, tomando en cuenta que el uso de estas técnicas es ilimitado. Una de las ventajas que tiene el Data Mining y principal utilidad que le podría dar es el tratamiento de datos para la toma decisiones; desde un punto de vista pragmático y asociando el proceso de control para la demanda turística con la Minería de Datos, permitirá reunir, depurar y transformar datos de información que no está estructurada, en información estructurada, con el objetivo de analizarla y convertirla en conocimiento y así tener un soporte para tomar decisiones. Otro de los beneficios que podrían mejorar los procesos de demanda turística de la provincia del Carchi es en el almacenamiento de información, al aplicar técnicas de minería de datos, todos los datos pueden ser gestionada de mejor manera, lo que le permitirá que la información este bien organizada y a la hora de necesitarla se presente de forma rápida con eficacia y eficiencia.

1.4. OBJETIVOS Y PREGUNTAS DE INVESTIGACIÓN

1.4.1. Objetivo General

Crear un modelo de minería de datos para los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi.

1.4.2. Objetivos Específicos

- Fundamentar bibliográficamente las variables de estudio para que sustente el desarrollo de la presente investigación.
- Diagnosticar el manejo de los procesos de control de la demanda turística de la provincia del Carchi.
- Proponer un modelo de minería de datos con la información de los procesos que maneja el Ministerio de Turismo para la demanda turística de la provincia de Carchi.

1.4.3. Preguntas de Investigación

¿Como se fortalece la investigación de minería de datos y procesos de control?

¿Cómo se manejan los procesos de control de la demanda turística en el Ministerio de Turismo?

¿Qué procesos intervienen en la generación de un modelo de minería de datos?

II. FUNDAMENTACIÓN TEÓRICA

2.1. ANTECEDENTES DE LA INVESTIGACIÓN

Esta sección del capítulo trata de recopilar y organizar información con el fin de fundamentar la investigación, se ha realizado una búsqueda de información acerca de antecedentes que tiene características similares con el tema de investigación que se está trabajando, esta información ha sido encontrada en repositorios digitales de libros, tesis realizadas por universidades, artículos de revistas, entre otros.

En una primera investigación realiza por la Universidad de Loja y con el apoyo de los departamentos de Administración de Empresas y Economía, de la misma universidad. En el año 2016 desde la Sección de Hotelería y Turismo de la UTPL, se crea el Observatorio Turístico, Región Sur de Ecuador. El objetivo de este proyecto es facilitar la información turística necesaria para la toma de decisiones de los distintos actores implicados en el sector turístico. Esta iniciativa apoya al desarrollo de la industria turística donde busca fomentar prácticas de turismo sostenible y elevar la competitividad del destino en base a información fiable y comprobada (Universidad Técnica de Loja, 2016, pág. 1).

El Observatorio Turístico de Loja utiliza bases estadísticas y con ello establecen un marco de referencia para el análisis sistemático de la situación real y tendencias de la industria turística, y esta información la usan como base para tomar decisiones y al mismo tiempo dar un seguimiento al impacto que puede tener el sector turístico en la Zona 7. Algunos datos que recopila el Observatorio son indicadores de alojamiento, perfil del visitante, gasto turístico, motivación del viaje, nivel de satisfacción turística, entre otros. Cabe señar que el alcance geográfico que tiene es solo la región sur del Ecuador (Universidad Técnica de Loja, 2016, pág. 1).

En otra investigación realiza por Uriarte del Águila, Ch. (2018). indica que: Con el objetivo de ayudar a tomar decisiones dentro del área de gestión al cliente, optaron

por aplicar nuevas herramientas de tecnología como es la minería de datos, este proyecto está orientado hacia una Telefónica de Perú y para cumplir su meta, inicialmente tuvieron que analizar los procesos con lo que se trabaja, y así tener claro cómo funcionan todos los procesos e incluso identificar posibles errores que existan, tomando en cuenta estos aspectos, el siguiente punto fue diseñar e implementar un sistema que aplique minería de datos, para poder medir los resultados e interpretar información para tomar decisiones en cuanto a como mejor los procesos y optimización de recursos e inclusive destacar falencias para poder corregirlas. Los resultados de esta investigación fueron positivos para el área de gestión de la telefónica debido a que el desempeño de las herramientas de minerías de datos mejoro en 100% los procesos de gestión al cliente, los aspectos que mejoraron fueron el tiempo, la eficacia y la eficacia en cuanto a la forma de como procesar la información (Uriarte del Águila, 2018, pág. 14).

Las herramientas Data Mining son muy útiles para procesar información en gran cantidad, debido a que puede detectar ciertos patrones que pueden ayudar a mejorar tareas o procesos donde exista una buena cantidad de información y sea vital del proceso. Estas nuevas herramientas trabajan con un análisis estadístico y tiene la posibilidad deducir tendencias que existen en los datos, todo esto que hace la minería de datos no se pude detectar mediante una exploración tradicional de la información porque, las relaciones son complejas por la gran cantidad de datos que regularmente se sabe manejar.

En otra investigación realizada por Martínez, C. (2018). mencionan que: La investigación se enfoca en los procesos de análisis de ingresos no percibidos en la empresa de Telecomunicaciones ENTEL, el primer paso que hicieron fue un análisis de los procesos que tienen que ver con los ingresos, entre lo que se puede mencionar, fueron provisión de servicios privados de telefonía, internet, comunicaciones, entre otros. Con el análisis realizado determinaron que estos procesos son controlados mediante indicadores de gestión, que son generados por la transformación de datos de clientes y servicios. Para obtener algunos de estos indicadores la empresa tiene que realizarlo de forma manual y esto es un trabajo que demanda tiempo y esfuerzo por parte de los analistas de la empresa. Desde este punto de vista aplicar minería de datos donde lograron diseñar nuevas métricas para los procesos que maneja ENTEL e introdujeron un nuevo proceso donde puede identificar a los cliente y servicio

que se encuentran en estado crítico, lo que les permitió tener datos con mayor exactitud de los ingresos no percibidos con el fin de aplicar nuevas estrategias para hacer el cobro a sus clientes (Martínez C. , 2018, pág. 2).

En otro proyecto desarrollado por el Ministerio de Turismo el cual es una herramienta tecnológica que ayuda actualizar información de establecimientos turísticos. Esta plataforma digital se la denomina SITURIN, esta herramienta se la desarrollo para realizar procesos de acreditación de los establecimientos de servicios turísticos a nivel nacional. Este sistema permite realizar le registro, recategorización, reclasificación, actualización, inactivación y reingreso de establecimientos en las actividades de alojamientos, alimentos y bebidas, operación e intermediación turística. Con este proyecto el Ministerio de Turismo busca disminuir trámites presenciales y regularizar establecimientos. La plataforma SITURIN fue lanzada el 04 de enero del año 2021, donde el Ministerio de Turismo explico como herramienta obligatoria para las acreditaciones de los prestadores de servicios turísticos a escala nacional (Ministerio de Turismo, 2021).

Un proyecto de investigación elaborado por Rodríguez & Coronel. (2016). indican que: Los problemas que suelen presentar empresas, organizaciones o instituciones para aplicar nuevas herramientas tecnológicas es el desconocimiento de las funciones y ventajas que aportan las nuevas tecnologías, un claro ejemplo es la aplicación de Minería de Datos, son pocas las organizaciones que emplean este tipo de técnicas y desconocen sus ventajas. El Data Mining tiene el objetivo de buscar y extraer conocimiento de grandes volúmenes de información histórica. Un aspecto importante que se menciona en esta investigación es que las instituciones públicas o privadas que están orientadas al sector turístico utilizan muchos servicios donde son aplicables nuevas herramientas como la minería de datos, debido a que el turismo enfoca sus estrategias en base a temporadas, es por ello que siempre es necesario anticiparse a la demanda con el fin de optimizar el uso de recursos y posibilita a tomar decisiones con bases teóricas (Rodríguez & Coronel, 2016, pág. 12).

"El sector turístico siempre va a requerir de gran cantidad de información para ser gestionada y comercializada, y son este tipo de sectores donde la tecnología los puede impulsar con aplicaciones orientadas a la automatización de tratamiento de datos" (Fernández, Díaz, & Martínez, 2017, pág. 11).

2.2. MARCO TEÓRICO

En sección de la documentación se define palabras técnicas que se usarán en todo el desarrollo del proyecto, mismas que serán de gran utilidad para el proceso de desarrollo de la investigación.

En la actualidad es común manejar gran cantidad de datos en entornos económicos, comerciales, científicos, entre otros, donde son almacenados de forma automatizada en grandes bases de datos. Toda esta información es muy difícil lograr analizarla en su totalidad debido a su excepcional cantidad, tradicionalmente se aplicaba distintas series de bases estadísticas, que consistía en aplicar técnicas estadísticas rutinarias como regresión, correlación, entre otros. El efecto que ha tenido la evolución de las nuevas tecnologías ha sido de beneficio para distintos sectores debido a conseguido la automatización de procesos, almacenamiento, redes, pruebas, entre otros; pero, así como también existen ventajas, se han encontrado con nuevos problemas como la generación de grandes cantidades de información. Las empresas guardan toda la información de forma empírica es decir almacenan la información como bitácoras de registros, datos de usuarios. Toda esta información que algunas empresas han determinado como inservible, y cabe mencionar que es una grave decisión, puesto que puede ser o convertirse en una gran fuente de información y conocimiento. El conocimiento es poder, el poder es la habilidad de controlar o influenciar eventos, el conocimiento es tener normas de como un hecho afectar a otro, tomando en cuenta la información disponible. La información se obtiene a partir de si los datos son útiles. Los datos son registros almacenados a partir de observaciones y hechos (Martínez B. , 2018, pág. 5).

2.2.1. Descubrimiento de Conocimiento de Bases de Datos

Knowledge Discovery in Data bases (KDD), es un proceso que tiene por objetivo la identificación de patrones dentro de un conjunto de datos que sean valiosos, interesantes y potencialmente útiles. El proceso general trata de transformar información de bajo nivel en conocimiento de alto nivel. KDD implica dos procesos:

- Búsqueda de regularidades interesantes entre los datos de inicio,
- Formulación de leyes que las describan.

2.2.1.1. Descubrimiento

Se pude definir de diferentes formas como recolección de datos, formular hipótesis con el fin de explicar observaciones, comparar hallazgos con los de otros investigadores y repetir el ciclo. Los ordenadores tienen la capacidad de realizar todas estas acciones es decir observar y recoger datos. Los programas estadísticos pueden generar agrupaciones de forma automática entre los datos recogidos; así como programas informáticos que tiene la capacidad de diseñar experimentos y otros sistemas robóticos realizan manipulaciones necesarias en los experimentos. Pero ningún computador tiene la capacidad de hacen todas estas capacidades juntas o simultáneamente, lo sígnica que ningún computador tiene la capacidad de descubrir. Pero debemos tomar en cuenta que el descubrimiento no requiere hacer estas actividades en conjunto; de igual forma que un investigador pueda lograr descubrir nuevos conocimientos a partir de un análisis de datos, una computadora puede inspeccionar datos que estén a su disponibilidad y encontrar relaciones y explicaciones que eran desconocidas, realizando así un descubrimiento en un sentido restringido. Una buena estrategia para obtener un descubrimiento de forma automática reside en la capacidad de los ordenadores, ya que tiene la capacidad de ejecutar búsquedas exhaustivas entre grandes volúmenes de datos, que tenga cierto grado de utilidad (Martínez B. , 2018, pág. 7).

2.2.1.2. Descubrimiento de conocimiento

Es la obtención de información que se encuentra implícita en un conjunto de datos, donde la información previamente es desconocida, pero tiene un potencial que puede resultar útil. Un sistema de descubrimiento será un programa que toma como entrada el conjunto de hechos y extrae las regularidades existentes. Si el conocimiento es extraído a partir de datos que se encuentran almacenados en una base de datos, se obtiene KDD. Los conceptos de lenguaje, certeza, simplicidad e interés con los que se define el descubrimiento de conocimiento, son suficientemente vagos como para que esta definición cubra la amplia variad de tendencias. Las ideas fundamentales que diferencia el KDD de otros sistemas de aprendizaje son:

- Lenguaje de alto nivel: Un leguaje de alto nivel se lo defino como un conocimiento descubierto, es un conocimiento inteligible desde el punto de vista humano. Dentro del KDD, representación de bajo nivel como las

generadas por redes neuronales, a pesar de que son métodos valido de minería de datos.

- Precisión: los descubrimientos representan el contenido de las bases de datos, lo que significa la realidad, el grado de incertidumbre medirá la confianza que el sistema o usuario puede asignar a cierto descubrimiento; si la confianza no es la necesaria, lo patrones no llegaran a ser conocimiento (Martínez B. , 2018, pág. 8).
- Interés: Es posible extraer numerosos patrones de cualquier base de datos, pero solo se puede considerar conocimiento datos que sean o resulten interesantes y esto se determina por medio de criterios generados por el usuario. Particularmente un patrón es interesante cuando es nuevo y potencialmente útil.
- Eficiencia: Son procesos de descubrimiento que puedan ser implementados en un computador con eficiencia. Se considera que un algoritmo es eficiente cuando su tiempo de ejecución y el espacio de memoria crece con el tamaño de los datos de entrada. Es imposible aprender con eficiencia cualquier concepto booleano, pero, si existen algoritmos para clases restringidas de conceptos. Otra opción que existe es aplicar heurísticas y algoritmos aproximados para la inducción de conocimiento.

El descubrimiento de conocimiento tiene sus inicios en el aprendizaje automático y la estadística, existen elementos que lo hacen diferente en gran mediad. La principal diferencia es que el objetivo principal es encontrar conocimiento útil, valido, relevante y nuevo sobre un fenómeno o actividad a través del uso de algoritmo eficientes. Generalmente el campo de Knowledge Discovery in Data bases (KDD), es la convergencia del aprendizaje automático, la estadística, el reconocimiento de patrones, IA, DB, la visualización de datos, los sistemas de apoyo a la toma de decisiones, recuperación de información, y otros campos (Martínez B. , 2018).

2.2.2. Extracción de Conocimiento

El proceso de extracción de conocimiento en ingles Knowledge Discovery from Databases, es como se lo llama al proceso o los procesos que se encargar de obtener conclusiones o información útil a partir de datos que regularmente están almenada en bases de datos. Los procesos son iterativos debido a sus fases ya que se pueda

necesitar que se regrese a una fase anterior y además es necesario iterar varias veces para lograr obtener el conocimiento de los datos.

Los procesos del KDD consta de 5 fases, para el presente proyecto nos centramos en una de ellas la cual es minería de datos, pero hablaremos de las demás por lo que no podemos hablar sobre minería de datos sin tomar en cuenta los conocimientos sobre cómo se seleccionan, limpian transforman y se almacenan estos datos, y por supuesto los más importante como evaluar, extraer e interpretar los resultados obtenidos de la investigación. (Cortina, 2018)

Procesos KDD

Los pasos para el proceso interactivo e iterativo del KDD son los siguientes:

- El primer paso requiere de cierta dependencia del usuario/analista, por razón de que intervienen factores como: saber que partes son susceptibles de un procesado automático y cuales no, criterios de rendimiento, para que se usaran los resultados que se obtengan, compromisos entre simplicidad y precisiones de conocimiento extraído. Este proceso se los denomina desarrollo y entendimiento del dominio de la aplicación, el conocimiento relevante.
- Creación del conjunto de datos objetivo, seleccionado el subconjunto de variables sobre los que se realizará el descubrimiento, todo esto implicará consideraciones sobre la homogeneidad de los datos, su variación, estrategia de muestreo, grados de libertad, entre otros.
- Transformación y reducción de los datos, en este proceso se realiza la búsqueda de característica útiles de los datos según sea el objetivo final, la reducción del número de variables y la proyección de los datos sobre espacios de búsqueda en los que sea más fácil encontrar una solución. Es un paso critico dentro del proceso global, que requiere un buen conocimiento del problema y una buena intuición, debido a que esto marcara la diferencia entre el éxito o fracaso de la minería de datos (Martínez B. , 2018, pág. 10).
- Elección del tipo de sistema para minería de datos. Depende de sí el objetivo del proceso de KDD es la clasificación, regresión, agrupamiento de conceptos (clustering), detección de desviación, entre otros.
- Elección del algoritmo de minería de datos.

- Minería de datos en este paso se realiza una búsqueda de conocimiento con una determinada representación de este. El éxito de la minería de datos depende de los pasos previos por parte del usuario.
- Interpretación del conocimiento extraído, tiene la posibilidad de iterar de nuevo des el primero paso. Para obtener resultados aceptables dependerá de diversos factores como: definición de medidas del interés del conocimiento que permitan el filtrado de forma automática, existencia de técnicas de visualización para facilitar la valoración de los resultados o búsqueda manual de conocimiento útil entre los resultados obtenidos (Martínez B. , 2018, pág. 11).
- Consolidación del conocimiento descubierto, incorporándolo al sistema o simplemente documentándolo y enviándolo a la parte interesada. Este paso incluye la revisión y resolución de posibles inconsistencias con otro conocimiento extraído previamente.

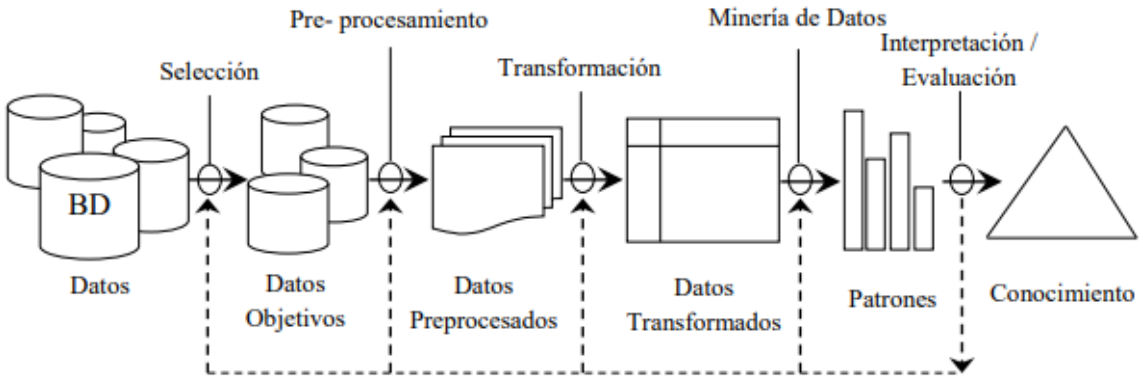


Figura 1. Proceso de KDD
Fuente: Martínez, B. (2018). Minería de Datos. Puebla: Benemérita Universidad Autónoma de Puebla.

Debemos tomar en cuenta que los procesos de KDD no son muy diferentes el uno del otro, lo que significa que si existe una alteración en cualquiera de los pasos puede afectar todo el proceso.

Fases del KDD

El descubrimiento de conocimiento tiene la función de identificación de relaciones y patrones existenciales en los datos. KDD consiste en la extracción no trivial de conocimiento previamente desconocido y potencialmente útil iniciando por un conjunto de datos.

En los procesos de KDD es posible definir 6 estados (Martínez B. , 2018, pág. 12):

Recolección de Datos

Estas primeras fases del descubrimiento de conocimiento determinan que las fases siguientes sean capaces de extraer conocimiento válido y útil a partir de la información original. Comúnmente, la información que se requiere investigar sobre una cierta parte de la organización se encuentra en bases de datos y otras fuentes muy diversas.

El análisis siguiente será si la fuente es unificada, accesible y desconectada del trabajo transaccional. De tal manera que el proceso subsiguiente de minería de datos:

- Depende mucho de la fuente:
    - OPLAP u OLTP
    - Datawarehouse o copia con el esquema original
    - ROLAP O MOLAP
- Depende también del tipo de usuario:
    - Granjeros: se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros controlar valores anómalos.
    - Exploradores: encargados de encontrar nuevos patrones significativos utilizando técnicas de minerías de datos.
- Recolección de la información externa: las informaciones externas pueden ser de gran importancia y se los puede encontrar en:
    - Demografía, gráficos web, información de otras organizaciones, entre otros.
    - Datos compartidos en una industria o área de negocios, organizaciones, entre otros.
    - Datos resumidos de áreas geográficas, distribución de la competencia.
    - Bases de datos externas compradas a otras compañías.

Selección, Limpieza y Transformación de Datos

Data cleansing y criba de datos: se debe eliminar el mayor número posible de datos erróneos o inconsistente e irrelevantes. Métodos estadísticos casi exclusivamente.

Minería de Datos

Aparte del gran volumen de los datos se deben tomar en cuenta las características especiales de los datos. A pesar de que algunos datos se pueden aplicar directamente, el interés dé la investigación en minería de datos está en su adaptación.

- Patrones por descubrir
    - Una vez recolectados los datos de interés, un explorador puede decidir que tipos de patrón quiere descubrir.
    - El tipo de conocimiento que se desea extraer va a marcar claramente la técnica de minería de datos a utilizar.
    - Según como sea la búsqueda del conocimiento se puede distinguir entre
        - Directed data Mining.- se sabe claramente lo que se busca, generalmente predecir unos ciertos datos o clases.
        - Undirected data Mining.- no se sabe lo que se busca, se trabaja con los datos.

Evaluación y Validación

La fase anterior produce una o más hipótesis de modelos. Para selecciona y validar estaos modelos es necesario el uso de criterios de evaluación de hipótesis.

- Primera fase: comprobación de la precisión del modelo en un banco de ejemplos independiente del que se ha utilizados para aprender el modelo. Se puede elegir el mejor modelo.
- Segunda fase: se puede realizar una experiencia piloto con ese modelo, por ejemplos si el modelo encontrado se quería usar para predecir la respuesta de los clientes a un nuevo producto, se puede envía un mailing a un subconjunto de clientes y evaluar la fiabilidad del modelo (Martínez B. , 2018, pág. 16).

Interpretación y difusión

El despliegue del modelo a veces es trivial pero otras veces requiere un proceso de implementación o interpretación:

- El modelo puede requerir implementación
- El modelo es descriptivo y requiere de interpretación

- El modelo puede tener muchos usuarios y necesita difusión; el modelo puede requerir ser expresado de una manera comprensible para ser distribuido en la organización.

Actualización y Monitorización

Los procesos derivan en un mantenimiento:

- Actualización: Un modelo válido puede cambiar el contexto ya sea económico, competencia, fuente de datos, entre otros.
- Monitorización: Consiste en ir revalidando el modelo con cierta frecuencia sobre nuevos datos, con el objetivo de detectar si el modelo requiere una actualización.

Producen realimentación en el proceso KDD

2.2.3. La Minería de Datos

Evolución Histórica

Debemos tomar en cuenta que algunos componentes que son parte del Data Mining, existen desde hace décadas en la investigación de sectores como: la inteligencia artificial, la estadística o el aprendizaje automático. Lo que ahora el mundo está presenciando es el reconocimiento de madurez de estas técnicas y junto con el desarrollo de los sistemas de gestión de bases de datos y herramientas para integrar información. El termino de minería de datos aparece en los años 50. Las áreas de informática en distintas empresas preparaban resúmenes de información generalmente de tipo comercial, que se encontraban almacenada en ficheros del ordenador central, con el fin de hacer más sencilla la tarea de la labor directiva. Este punto de vista fue importante puesto que así nacieron los sistemas de información para la dirección, estos primeros sistemas, eran voluminosos, pocos flexibles y difíciles de leer para los no informáticos. En los años 60 apareces los sistemas de gestión de bases de datos, que aún se mostraban rígidos y carecían de flexibilidad para realizar consultas. Por otra para solventar los problemas y a pesar de los informes resultaban muy complicados de preparar y depurar, aparecen los motores de relacionales. Un grave problema que se genero era la diversidad de bases de datos no integradas establecidas por las diferentes áreas de una organización (Martínez B. , 2018, pág. 17). Para darle solución a todos los problemas que se ocasionaron en los años 60, se crear el Data Warehouse a finales de los 80. El DW estimula el desarrollo de los

enfoques de Data Mining, en los que las tareas de análisis se automatizan y dan un paso más al posibilitar la extracción de conocimiento inductivo.

Tabla 1. Historia de las tecnologías enlazadas a la Minería de Datos

Etapa	Ejemplos de Uso	Tecnologías	Características
Recolección de datos. (1960)	Dime mis beneficios totales en los últimos 4 años	Ordenadores, cintas, discos	Retrospectivo, datos estáticos.
Acceso a los datos. (1980)	Ventas en Guayaquil durante las últimas navidades.	Bases de Datos Relacionales (SQL) ODBC	Retrospectivo, datos dinámicos a nivel de registro.
Data Warehouse y soporte a la toma de decisiones. (1990)	Ventas en Quito detalle por delegación y descender a nivel tienda.	OLAP; Bases de datos multidimensional, data warehouse	Retrospectivo, obtención dinámica de datos a múltiples niveles.
Data Mining	Justifica la tendencia de ventas en Tulcán para el próximo año.	Algoritmos avanzados, ordenadores, multiprocesadores, bases de datos masivas.	Prospectivos, obtención proactiva de información.

Fuente: Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

¿Qué es Minería de Datos?

Como se ha mencionado las empresas en años atrás estaban o tenían objetivo principal de alimentar los sistemas contables y financieros, de igual forma gestionar los procesos de inventarios, producción, recursos humano y ventas. Conforme a pasado el tiempo las empresas al igual que las tecnologías cambiaron y se han hecho más competitivas, y esto ocasionó que los datos cumplan un papel fundamental en las empresas ya que la información es vital y estratégica para la toma de decisiones. En este sentido las empresas han buscado la forma de darle un valor a la basta información que se tiene almacenado en los distintos sistemas de gestión de bases de datos. Por esta razón algunas empresas y organizaciones han buscado la forma de automatizar los procesos y poder así descubrir información útil o valiosa, ya que de otra manera esta información seguiría están en segundo plano o simplemente desperdiciarla.

Con la constate evolución que está teniendo la tecnología, las empresas ya disponen de herramientas de software y hardware más sofisticadas que tienen la posibilidad

de almacenar grandes de cantidades de información y el análisis de esta. El avance tecnológico junto con la aparición de la competitividad entre empresas le sugiere y no solo a las empresas sino también organización e instituciones públicas el mejorar continuamente sus esquemas de administración y toma de decisiones, de forma que puedan explotar las grandes fuentes de información con el fin de obtener un conocimiento que aporte al mejoramiento de la empresa (Martínez B. , 2018, pág. 18).

Existen distintas técnicas que tienen la posibilidad de sacar el máximo provecho de los datos, extrayendo información que o es detectada a simple vista. Unas de estas técnicas es la Minería de Datos, donde combina técnicas semiautomáticas de inteligencia artificial, análisis estadístico, bases de datos y visualización gráfica, para obtener información que se encuentra implícita es una gran cantidad de datos. La minería de datos descubre relaciones, tendencias, desviaciones, comportamientos, atípicos, patrones y trayectorias ocultas, con el objetivo principal de darle soporte a los procesos de toma de decisiones con mayor conocimiento. La minería de datos se puede ubicar en el nivel más alto de la evolución de los procesos tecnológicos de análisis de datos.

Según Martínez, B. (2018). explica que: "El Data Mining nace de una analogía entre una montaña y la gran cantidad de datos almacenado en cualquier empresa. Dentro de la montaña, ocultos entre piedras y tierra, se encuentra diamantes de gran valor que mediante actividades de minería son encontrados y aprovechados" (pág. 18).

Existen muchas definiciones para la técnica de Minería de Datos, hecha en distintos libros, por diferentes autores, es una bibliografía muy amplia y variada, con el finde orientar la investigación podemos resaltar algunas:

- Grupo de técnicas que buscan automatizar la detección de parones en un gran volumen de datos.
- El Data Mining es un proceso de descubrimiento muy útiles de nuevas correlaciones, patrones y tendencias de grandes cantidades de datos almacenados en repositorios, utilizando tecnologías de reconocimiento, así como también el uso de estadística y matemática.

- La minería de datos utiliza herramientas estadísticas avanzadas con el fin de analizar e investigar las bases de datos existentes en una empresa, organización e institución por medio de descubrimiento.
- Data Mining es el análisis, exploración y extracción por medio de técnicas automáticas y semiautomáticas, de grandes cantidades de datos con el fin de obtener conocimiento de una gran cantidad de datos almacenado en un SGBD.
- El término de minería de datos aparece en la integración de múltiples tecnologías como la estadística, el soporte a la toma de decisiones, el aprendizaje automático, entre otras. Para realizar este proceso se aplican técnicas procedentes de diversas áreas, como algoritmos genéticos, las redes neuronales, arboles de decisión, entre otros.
- DM es un proceso de análisis de archivos históricos como bitácoras, transacciones, procesos de gestión, entre otros, con el objetivo de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones que sean útiles para la toma de decisiones.
- La Minería de datos se la puede definir como actividades de extracción de información, tendencias y patrones de comportamiento que se encuentran ocultos en grandes volúmenes de información.
- Data Mining es un proceso que, mediante la cuantificación y descubrimiento de relaciones predictivas en los datos, nos permite transformar los datos en conocimiento útil.
- La técnica del DM ayuda a transformar información que se encuentra implícita y esta almacena en bases de datos, en un conocimiento donde puede ser utilizado para el mejoramiento de una negocio, empresa, organización e institución (Martínez B. , 2018, pág. 19).

Potencial de la Minería de Datos

La tecnología informática es infraestructura fundamental de las grandes organizaciones y permite, en la actualizada, registrar con lujo de detalle, los elementos de todas las actividades con asombrosa facilidad.

La tecnología de bases de datos permite almacenar grandes cantidades de información donde estos reflejan la interacción de la organización con todas las conexiones que tiene la empresa, ya sean otras organizaciones, sus clientes,

empleados, entre otros. Puesto que se tiene un registro bastante amplio de la organización, pero en cierto punto existe un problema, como interpretar esta gran cantidad de información en conocimiento y sabiduría corporativa que apoye efectivamente la toma de decisiones, especialmente a nivel gerencial que dirige el destino de las organizaciones (Martínez B. , 2018, pág. 22). ¿Cómo comprender el fenómeno, tomando en cuenta grandes volúmenes de datos?

La minería de datos ha surgido del potencial de análisis de grandes volúmenes de información, con el fin de obtener resúmenes y conocimiento que apoye la toma de decisiones y que pueda construir una experiencia a partir de grandes cantidades de información que se registra en una empresa u organización, ya sea en sus sistemas informáticos o documentos físicos.

Martínez, B. (2018). indica que: "Las DB parece ser más efectiva cuando los datos tienen elementos que pueden permitir una interpretación y explicación en concordancia con la experiencia humana" (Martínez B. , 2018, pág. 21).

La tecnología promete analizar con facilidad grandes volúmenes de datos y reconocer patrones en tiempo y espacio que soportarán la toma de decisiones y construirán un conocimiento corporativo de alto nivel. La tecnología de minería de datos parece robusta y lista para su aplicación, dado el gran crecimiento de empresa que comercializan software con diferentes técnicas. Más aún, gran parte de estas técnicas son una combinación directa de madurez en tecnología de bases de datos y data warehouse, con técnicas de aprendizaje automático y de estadística.
Sin embargo, la tecnología enfrenta aún varios retos.
La minería de datos es una herramienta exploratoria y no explicativa. Es decir, explora los datos para sugerir hipótesis

2.2.4.1. Proceso de minería de datos
El proceso se iniciar con identificar los datos, para esto debemos tener claro que datos se necesitan, donde se los puede encontrar puede ser en: bases de datos, papel, archivos, ficheros, entre otros; y saber cómo conseguirlos. Teniendo a disponibilidad los datos, se deben preparar, poniéndolo en bases de datos en un formato adecuado o construir una Waterhouse. Esta es una de las tareas más difíciles del Data Mining. Luego de esto los datos deben ser analizados en un formato adecuado para realizar una selección de datos esenciales y eliminación de los

innecesarios. Antes de esto paso es recomendable tener ideas de que lo que le interesa averiguar, que herramientas se necesitan y como proceder. Tras aplicar la herramienta elegida o construida por nosotros sí es el caso, haya que saber interpretar los resultados obtenidos para saber los que son significativos y como podarlos para extraer únicamente los resultados útiles. Luego de examinar solo los resultados útiles se debe identificar las acciones que deben ser tomada, discutirlas y pensar en los procedimientos para llevar a cabo e implementarlas (Martínez B. , 2018, pág. 27).

"Una vez han sido implementadas hay que evaluarlas para ello hay que observar los resultados, los beneficios y el coste para poder reevaluar el procedimiento completo. Para entonces los datos puede a ver cambiado, nuevas herramientas pueden estar disponibles y probablemente habrá que cambiar el ciclo de minería" (Martínez B. , 2018, pág. 27). La realidad de la Minería de Datos es un proceso que involucra ajustar modelos o determinar patrones a partir de datos. Los pasos que se debe tomar en cuenta para realizar un proyecto de minería de datos son siempre los mismos, independientemente de la técnica especifica de extracción de conocimiento usada.

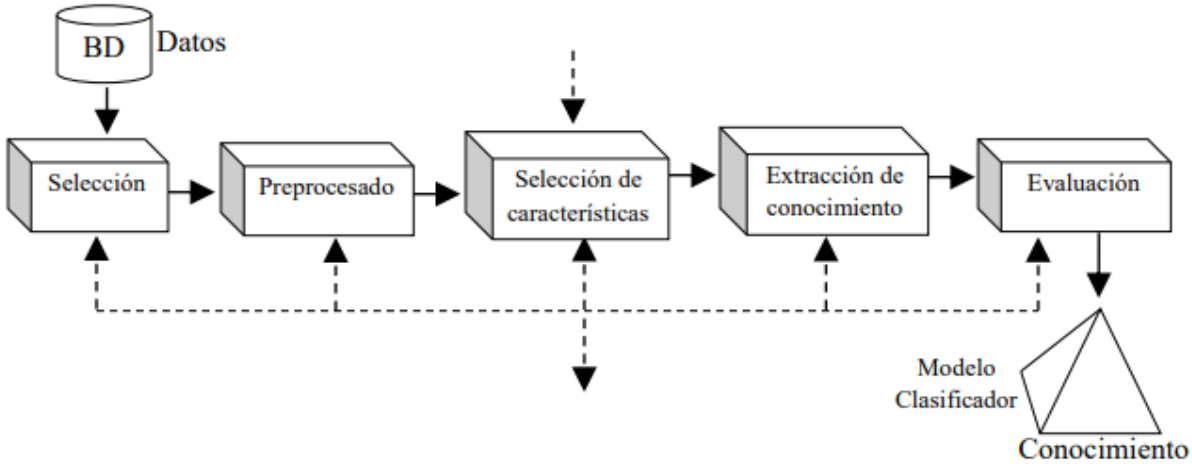


Figura 2. Visión general de los procesos del Data Mining
Fuente: Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

Los estados de los procesos de la minería de datos:

- Procesado de los Datos

  El formato de los datos contenido en la fuente de datos ya sea una base de datos, data Waterhouse, papel, archivos físicos, entre otros; generalmente nos es posible utilizar ningún algoritmo de minería sobre los datos y es aquí donde se involucra el preprocesado. Mediante el preprocesado, se puede filtrar los datos donde se

eliminan valores incorrectos, no válidos, desconocidos; esto depende de las necesidades y el algoritmo a usar, luego se obtiene muestra de los mismos en busca de una mayor velocidad de respuesta del proceso, o se reducen el número de valores posibles mediante técnicas de redondeo, clustering, ente otros (Martínez B. , 2018, pág. 28).

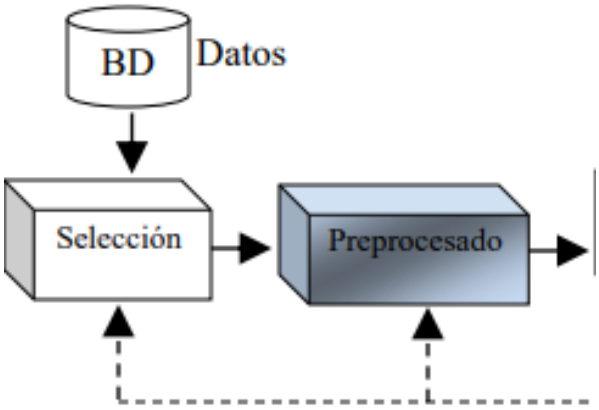


Figura 3. Preprocesado - Data Mining
Fuente: Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

- Selección de Características

Luego de haber preprocesados los datos, frecuentemente en la mayoría de los casos aún se tiene una cantidad considerable de los datos. La selección de características reduce el tamaño de los datos tomando en cuenta las variables que tiene más influencia en el problema, sin apena sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería. Los métodos para la selección de características son dos:

- Aquellos basados en la elección de los mejorar atributos del problema y,
- Aquellos que buscan variables independientes mediante test de sensibilidad, algoritmo de distan o heurísticos.

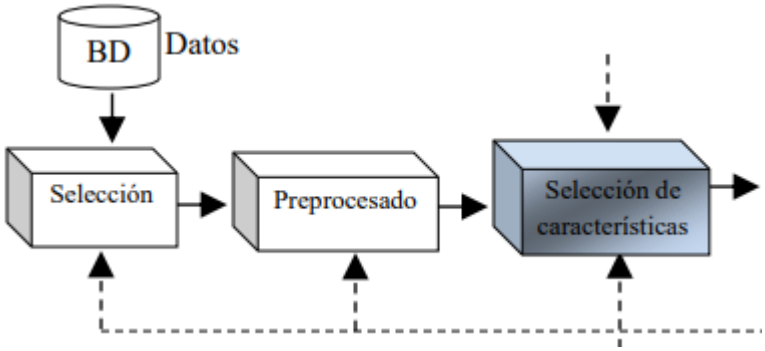


Figura 4. Selección de Características - Data Mining
Fuente: Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

- Algoritmos de Aprendizaje

Aplicando una técnica de minería de datos, se logra tener un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse distintas técnicas a la vez con el fin de generar diferentes modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

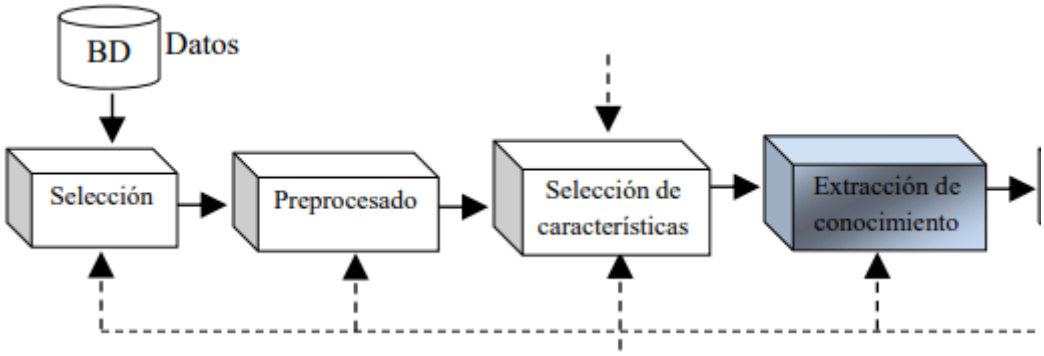


Figura 5 Algoritmos de Aprendizaje - Data Mining
Fuente: Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

- Evaluación y Validación

Con la obtención del modelo, el siguiente paso es su validación, comprobando que las conclusiones que arroja son validad y suficientemente satisfactorias. En un caso dado de haber tenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Sin ninguno de los modelos alcanza lo resultado esperados, debe

alterarse algunos de los pasos anteriores para generar nuevos modelos de conocimiento (Martínez B. , 2018, pág. 29).

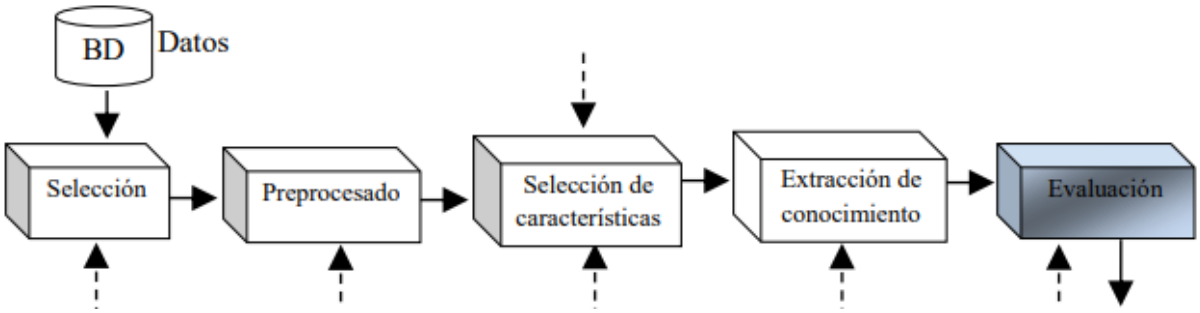


Figura 6. Evaluación y Validación - Data Mining
Fuente: Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

Los procesos de la minería de datos son analíticos, diseñados para explorar grandes cantidades de datos, con el fin de detectar patrones de comportamiento consistentes o relaciones entre diferentes variables para aplicarlos a nuevos conjuntos de datos.

2.2.4.3. Técnicas de Aprendizaje

Las técnicas de minería de datos crean modelos que son predictivos y/o descriptivos. Un modelo predictivo responde preguntas sobre datos futuros. Un modelo descriptivo proporciona información sobre las relaciones entre los datos y sus características.

Los algoritmos supervisados o predictivos predicen el valor de un atributo -etiqueta- de un conjunto de datos, conocidos como atributos descriptivos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos (Martínez B. , 2018, pág. 31).

Cualquier problema de aprendizaje inductivo se pude presentar de forma directa en distintas formas:

- Interpolación.- función continua sobre varias dimensiones.
- Predicción secuencial.- las observaciones están ordenadas secuencialmente. Se predice el siguiente valor de las secuencias. Caso particular de interpolación con 2 dimensiones, una discreta y regular.
- Aprendizaje Supervisado.- cada observación incluye un valor de la clase a la que corresponde.

- Aprendizaje no Supervisado.- el conjunto de observaciones no tiene clases asociadas. El objetivo es detectar regularidades en los datos de cualquier tipo: agrupaciones, contornos, asociaciones, valores anómalos (Martínez B. , 2018, pág. 34).

Algoritmos Predictivos y Descriptivos

Algoritmos Predictivos

Los algoritmos predictivos se los pueden encontrar en tres diferentes formas:

- Interpolación y Predicción Secuencial, donde se usan técnicas como:
    - Datos continuos reales:
        - Regresión Lineal
            - Regresión lineal clásica.
            - Regresión lineal ponderada localmente.
        - Regresión No Lineal: logarítmica, pick, entre otros.
    - Datos discretos:
        - No existen técnicas específicas, se suelen utilizar técnicas de algoritmos genéticos o algoritmos de enumeración refinados.
- Aprendizaje Supervisado

  Dependiendo la estimación que se obtenga si es:
    - Clasificación: se estima una función -clases disjuntas-
    - Categorización: se estima una correspondencias -las clases pueden solapar-.

  Dependiendo también del número y tipos de clases:
    - Clase discreta: se conoce como *clasificación* "Determinar el grupo sanguíneo a partir de los grupos sanguíneos de los padres".
    - Si tienen dos valores "VyF" se conoce como *concept learning* "Determinar si un compuesto químico es cancerígeno".
    - Clase continua se conoce como "estimación".
- Clasificación del Aprendizaje Supervisado
    - Técnicas
        - K-NN *vecino más cercano*
        - K-means *aprendizaje competitivo*
        - Aprendizaje de perceptrón *perceptrón learning*
        - Métodos ANN multicapa
        - Funciones de base radial

        - Árbol de decisiones
        - Clasificadores Bayesianos
        - Métodos de división Central.

Algoritmos Descriptivos

- Análisis Exploratorios
    - Técnicas
        - Estudios correlacionales
        - Dependencias
        - Detección datos anómalos
        - Análisis de dispersión.
- Aprendizaje no Supervisado *segmentación*
    - Técnicas de Agrupamiento *clustering*
        - K-means *aprendizaje competitivo*
        - DBSCAN
        - K-medoids
        - Redes Neuronales de Kohonen

2.2.4.4. Métodos de Minería de Datos

La función de los métodos de minería de datos tomando en cuenta el alto nivel es el realizar la predicción de datos desconocidos y la descripción de patrones. Se puede emplear diferentes criterios para la clasificación de sistemas de minería datos y en general, los sistemas de aprendizaje inductivo en computadoras:

- Dependiendo del objetivo para el que se realiza el aprendizaje.
  Pueden diferenciarse sistemas para *clasificación* clasificar datos en clases predefinidas, *regresión* función que convierte datos en valores de una función de predicción, *agrupamiento de conceptos* búsqueda de conjuntos en los que agrupar los datos, *compactación* búsqueda de descripciones más compactas de los datos, *modelado de dependencias* entre variables de los datos, *detección de desviaciones* búsqueda de desviaciones importantes de los datos respecto de valores anteriores o medios (Martínez B. , 2018, pág. 47).
- Dependiendo de la tendencia con que se aborde el problema.
  Se distinguen en tres líneas de investigación: *sistemas conexionistas* redes neuronales, *sistemas evolucionistas* algoritmos genéticos y *sistemas simbólicos.*
- Dependiendo del lenguaje utilizado para representar el conocimiento.

> Se las puede diferenciar en; representaciones basadas en la *lógica de proposiciones,* representaciones basadas en *lógicas de predicados* de primer orden, representaciones *estructuradas*, representaciones a través de ejemplos y representaciones *no simbólicas* como las redes neuronales (Martínez B. , 2018, pág. 47).

Ahora, explicaremos con más detalle los distinto métodos de representación del conocimiento que se emplean en la minería de datos, dado que el lenguaje de representación es uno de los aspectos importantes para el proceso de KDD.

- **Agrupamiento (Clustering)**

Conocido también como *segmentación,* es una herramienta que permite la identificación de tipologías o grupos donde los elementos guardan similitud entre sí y diferencias con aquellos de otros grupos. Para alcanzar las distintas tipologías o grupos existentes en una base de datos, estas herramientas requieren, como entrada, información sobre el colectivo a segmentar. La información corresponderá a los valores concretos, para cada elemento en un momento del tiempo, de una serie de variables *segmentación estática* o a través del comportamiento en el tiempo cada uno de los elementos del colectivo *segmentación dinámica.*

Usando esta técnica para el tratamiento de información, nos deriva en que estas herramientas presentan los distintos grupos detectados junto con los valores característicos de las variables. Este tipo de herramientas se basan en técnica de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros, para otro tipo de elementos a agrupar o segmentar, como el texto y documentos, se usan técnicas de reconocimiento de conceptos (Martínez B. , 2018, pág. 48).

Algunos de los algoritmos que son comúnmente aplicados con relación a tipos de aprendizajes no supervisados son:

- K-means

Es un algoritmo que intenta encontrar una partición de las muestras en un determinado valor de agrupaciones, de forma que cada agrupamiento que se genere pertenezca a una de ellas, concretamente a aquella cuyo centroide este más cerca. Un algoritmo de K-medias es relativamente eficiente y requieren pocos pasos para que el proceso se estabilice. Pero es necesario determinar el número de agrupaciones a priori, la generación de un modelo utilizando estos tipos de técnicas

de modelados son sensibles a la posición inicial de los K clústeres que se van a generar (Caparrini, 2020). A pesar de todo esto no existe un método teórico global que permita encontrar el valor óptimo de grupos iniciales ni las posiciones en las que debemos situar los centros, por loque se suele hacer una aproximación experimental.

- DBSCAN

Es un algoritmo de clustering, no paramétrico, basado en densidades y marca como outliers aquellos puntos que no superan un criterio de densidad establecido. Es uno de los algoritmos de agrupamiento más usados y citados, es relativamente eficiente cuando se usan estructuras de datos adecuados.

Generalmente, este algoritmo inicia en un punto con suficiente densidad y se construye en base a los puntos más cercanos relevantes al clúster, cuando termina la construcción de un agrupamiento con suficiente densidad, pasa a la creación de otro agrupamiento (Caparrini, 2020). En comparación con los algoritmos de K-medias y K-medoids, DBSCAN presenta ventajas como:

- No se necesita especificar número de clústeres deseados.
- Tiene la capacidad de detectar ruido en los agrupamientos generados.
- Solo necesita dos parámetros para la construcción de los clústeres.
- Una estructura adecuada de datos puede generar una implementación bastante rápida.

Este algoritmo también presenta desventajas importantes como:

- Las distancias euclidianas pueden dar resultados indeseados.
- Si dentro del conjunto de datos con densidades bajas o diferentes, no hay combinación de valores que puedan trabajar adecuadamente para todos ellos a la vez.

- K-medoids

El algoritmo K-medoid se basa en similitud, en lugar de usar medoids para representar los clusters. El medoid es un elemento del conjunto de datos y es el más centralizado del conjunto de datos, este algoritmo inicia con la selección aleatoria de k elementos de datos como centros iniciales para representar los k clusters, los elementos restantes se incluyen en el grupo que tiene el medoid más cercano a ellos y posteriormente se determina un nuevo centro que pueda representar mejor al grupo.

En cada iteración que se realice los elementos distintos a los centros se asigna nuevamente a los clusters que tiene el medoid más cercano, provocando que los centros alteren su ubicación. El algoritmo de K-medoid minimiza la suma de distancias

entre cada elemento de datos y su correspondiente medoid, esta acción se va a repetir hasta que ningún medoid cambie su colación, esto determinará el final del proceso y obtendrá los clusters finales (Salinas & Chavez, 2021, pág. 26).

- **Asociación (Descubrimiento de patrones de asociación)**

Estas herramientas establecen las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes, pudiendo reconocer como la ocurrencia de un suceso o acción puede incluir o generar la aparición de otros. Esta herramienta se fundamenta en técnicas estadísticas como los análisis de correlación y de variación.

- **Secuenciamiento (Descubrimiento de patrones secuenciales)**

Nos permite identificar como en el tiempo, la ocurrencia de una acción desencadena otras posteriormente. Similar a la anterior técnica, pero en este caso, el tiempo es una variable crítica e imprescindible introducir en la información que se tiene que analizar (Martínez B. , 2018, pág. 48).

- **Reconocimiento de Patrones**

Esta herramienta nos permite la asociación de una señal o información de entrada con aquella/as con la que guarda mayor similitud y están catalogadas en el sistema. Se las usa por elementos que son tan habituales como un procesador de texto o un despertador. Los patrones puede ser cualquier elemento de información que deseemos.

Dentro de la minería de datos este tipo de técnicas puede ayudar con la identificación de problemas e incidencias y de sus posibles soluciones toda vez que dispongamos de la base de información necesaria en la cual buscar. Estas herramientas se sustentan en las técnicas de Redes Neuronales y Algoritmo Matemáticos.

- **Previsión (Pronóstico)**

La previsión establece el comportamiento futuro más probable dependiendo de la evolución pasada y presente. Este tipo de técnica tiene un uso fundamental en el tratamiento de series temporales y las técnicas asociadas disponen de una importante madurez.

Las herramientas de forecasting utilizan bien la propia información histórica, o bien, la información histórica relativa a otras variables de las cuales depende la primera (Martínez B. , 2018, pág. 49).

- **Simulación**

Son herramientas que forman parte de un conjunto de herramienta experimentadas de la investigación científica. Estas técnicas se pueden definir como la generación de múltiples escenarios o posibilidades sujetas, normalmente a unas reglas o esquemas con el fin de analizar la idoneidad y comportamiento de una decisión o prototipo en un marco de posibles condiciones futuras o para analizar todas la posibles variaciones o alternativas a una decisión o situación, además que se la puede usar para el calculó numérico.

- **Optimización**

Al igual que las dos técnicas anteriores, la optimización tiene una amplia tradición de uso. Esta herramienta ha sido una de las más usadas en la resolución de los problemas asociados a la logística de distribución y a la gestión de *stocks* en los negocios y en la determinación de parámetros teóricos a partir de los experimentos en la investigación científica.

La optimización resuelve el problema de la minimización o maximización de una función que depende de una serie de variables, encontrando los valores de éstas que satisfacen esa condición de máximo, típicamente beneficios, o mínimo, normalmente costes. Habitualmente estos problemas conllevan, una serie de *ligaduras* o restricciones de forma que no todas las posibles soluciones son aceptables, se traduce en que debemos reducir nuestro universo de búsqueda a aquellas soluciones que satisfagan tales restricciones (Martínez B. , 2018, pág. 50).

- **Clasificación**

Esta técnica agrupa todas aquellas herramientas que permiten asignar a un elemento la pertenencia a un grupo o clase. Se realiza a través de la dependencia de la pertenencia a las clases en los valores de una serie de atributos o variables. Mediante un análisis colectivo de elementos, o casos de los cuales conocemos la clase a la que pertenecen, se establece un mecanismo que establece la pertenencia a tales clases en función de los valores de las distintas variables y nos permite establecer el grado de discriminación o influencia de éstas.

La clasificación se usa también para herramientas de **predicción** o **evaluación** en casos donde se apliquen técnicas, normalmente numéricas, que establecen para cada elemento un valor dependiente de los valores que tengan las variables en tal elemento. Las herramientas de clasificación hacen uso de técnicas como algoritmos matemáticos, análisis discriminante y de variaciones, sistemas expertos y sistemas de

conocimiento e inducción de reglas. Generalmente es necesario la conjunción e integración de varios tipos de herramientas a efectos de brindar una solución completa a nuestros problemas (Martínez B. , 2018, pág. 50).

Métodos Apropiados

- No estructurados
    - Métodos bayesianos
    - Otros métodos estadísticos
    - Métodos relacionales
- Semi estructurados
    - Gramaticales
    - Métodos relacionales con constructores

Métodos no apropiados

Sin una profunda transformación de los datos, muchas técnicas de aprendizaje automático son útiles para muchas aplicaciones.

- Métodos de clasificación *árboles de decisión* están basados, en una clase dependiente de un número de atributos predeterminados.
- Métodos numéricos *regresión, redes neuronales,* los datos son simbólicos, no numéricos.
- Métodos por casos *KNN* donde los tiempos de respuesta serían muy altos.

2.2.4.5. Técnicas de Minería de Datos

Las técnicas de minería de datos crean modelos que son predictivos y/o descriptivos. Un modelo predictivo responde preguntas sobre datos futuros. Un modelo descriptivo proporciona información sobre las relaciones entre los datos y sus características.

Los algoritmos supervisados o predictivos predicen el valor de un atributo -etiqueta- de un conjunto de datos, conocidos como atributos descriptivos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos (Martínez B. , 2018, pág. 31).

La minería de datos con la evolución que ha tenido ha ido opacando y de forma sosegada sustituyendo el análisis de datos dirigido a la verificación por un enfoque de análisis de datos dirigido al descubrimiento del conocimiento. La principal diferencia entre ambos se encuentra en que el último se descubre información sin

necesidad de formular previamente una hipótesis. La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, razón por la cual está técnica es mucho más eficiente que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad elevada. Estas técnicas emergentes se encuentran en continua evolución como resultado de la colaboración entre campos de investigación como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones (Martínez B. , 2018, pág. 51).

Los algoritmos de minería de datos se clasifican en dos categorías *supervisados* o *predictivos* y *no supervisados* o de *descubrimiento del conocimiento*.

Tabla 2. Técnicas de Minería de Datos

Supervisados	No Supervisados
Árboles de Decisión	Detección de Desviaciones
Inducción Neuronal	Segmentación
Regresión	Agrupamiento *Clustering*
Series Temporales	Reglas de Asociación
	Patrones Secuenciales

Fuente: Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

La aplicación de algoritmo de minería de datos requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada debido a que, en muchas ocasiones dichos datos proceden de fuentes heterogéneas, no tiene el formato adecuado o contienen ruido. Por otra parte, es necesario interpretar y evaluar los resultados obtenidos. Ahora se presentará lagunas de las técnicas de minería de datos más utilizadas:

Tabla 3. Técnicas más usadas en Data Mining

Métodos Estadísticos	Análisis de Varianza *ANOVA*
	Prueba Ji Cuadrado
	Análisis de Componentes Principales
	Análisis Clúster
	Análisis de Discriminante
	Regresión Lineal
	Regresión Logística
Árboles de Decisión	Detección Automática de Interacciones Mediante Chi-Cuadrado *CHAID*
	Clasificación y Regresión *CART*
Reglas de Asociación	
Redes Neuronales Artificiales	
Algoritmos Genéticos	
Otros Métodos	Lógica Difusa
	Series Temporales

Fuente: Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

- **Métodos Estadísticos**

La estadística es una herramienta que se ha usado para el tratamiento de grandes volúmenes de datos numéricos, además de tener una gran efectividad tiene un amplio conjunto de modelos de análisis para cubrir el tratamiento de todo tipo de poblaciones y serie de datos. Algunos de los métodos estadísticos más usados son:

- Análisis de Varianza *ANOVA*.- contrasta si existen diferencias significativas entre medidas de una o más variables continuas en grupos de población distintos.
- Prueba de Ji Cuadrado.- nos permite diferenciar las hipótesis que son independientes entre variables.
- Componentes Principales.- la función que tiene es que ayuda a reducir el número de variables observadas a un menor número de variables artificiales, conservando la mayor parte de la información sobre la varianza de las variables.
- Análisis Clúster.- permite clasificar una población en un número determinado de grupos, sobre la base de semejantes y diferencias de perfiles entre los diferentes componentes de dicha población.
- Análisis Discriminante.- se trata de un método de clasificación de individuos en grupos que previamente se han establecido, y que permite encontrar la regla

de clasificación de los elementos de estos grupos, y por tanto identificar cuáles son las variables que mejor definan la pertenencia al grupo.

- Regresión Lineal.- es la técnica más básica del Data Mining. Un modelo de regresión lineal se implementa identificando una variable dependiente (y) y todas las variables independientes x_1, x_2 ..., Se asume que la relación entre estas y aquella es lineal. Todas las variables han de ser continuas, el resultado es la ecuación de la recta que mejor se ajusta al juego de datos y esta ecuación se interpreta o se usa para predicción.
- Regresión Logística.- este método requiere que todas las variables sean lineales, además puede trabajar con variables discretas (Martínez B. , 2018, pág. 52).

Métodos Basados en Árboles de Decisión

Son herramientas analíticas empleadas para el descubrimiento de reglas y relaciones mediante la ruptura y subdivisión sistemática de la información contenida en el conjunto de datos. El árbol de decisión se construye a partir de la división del conjunto de datos en dos (Clasificación y Regresión *CART*) o más (Detección Automática de Interacciones Mediante Chi-Cuadrado *CHAID*) subconjunto de observaciones a partir de los valores que toman las variables predictoras. Cada uno de estos subconjuntos vuelven después a ser divididos utilizando el mismo algoritmo.

Este proceso continúa hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta. La raíz del árbol es el conjunto de datos íntegro, los subconjuntos conforman las ramas del árbol y un conjunto en el que se hace una partición se llama nodo.

El método *CHAID* es útil en ciertas situaciones en las que el objetivo es dividir una población en distintos segmentos basándose en algún criterio de decisión.

Reglas de Asociación

Son métodos donde se realizan análisis con el fin de extraer información por coincidencias. Este análisis permite descubrir correlaciones en los sucesos de la base de datos a analizar y se formaliza en la obtención de reglas de tipo *si…. entonces…*

Redes Neuronales

Las redes neuronales son una técnica inspirada en los trabajos de investigación, desde 1930, que intentaban modelar computacionalmente el aprendizaje humano llevado a cabo a través de las neuronas en el cerebro. Las redes neuronales son una nueva forma de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales, tienen la capacidad de detectar, aprender patrones y características dentro de los datos.

Se comportan de forma parecida a nuestro cerebro aprendiendo de la experiencia y el pasado, tal conocimiento es aplicado a la resolución de problemas. Teniendo el control de las redes neuronales se pueden hacer previsiones, clasificaciones y segmentación. Las *neural networks* se diseñan a partir de la estructura de una serie de capas o niveles que están compuestas por nodos (*neuronas*). Tienen dos formas de aprendizaje derivadas del tipo de paradigma que usan: supervisado y no supervisado.

Son técnicas que tienen un proceso numérico en paralelo que tienen el objetivo de modelizar el funcionamiento del cerebro. La red asigna tareas de forma aleatoria a cada variable independiente y determina si existe algún patrón predictivo en los datos. A la vez que encuentra un patrón la red lo optimiza reforzando las tareas de las variables y comparando con los datos del grupo de validación. A partir de esto sigue el proceso y aprende de los resultados una y otra vez. Por último, se puede aplicar el modelo que aprendió a cualquier nuevo conjunto de datos de entrada. Pueden manejar datos continuos y discretos, lineales y no-lineales simultáneamente. El único "inconveniente" que se puede presentar es que no se genera una ecuación o modelo que explica el comportamiento del sistema, siendo muy difícil determinar la influencia de cada variable en el comportamiento global del sistema (Martínez B. , 2018, pág. 54).

Algoritmo Genéticos

Este tipo técnica tiene sus bases o se puede decir que se inspira en la Biología. Son algoritmos que representan la modelización matemática de como los cromosomas en un marco evolucionista alcanzan la estructura y composición óptima. Entendiendo la evolución de un proceso de búsqueda y optimización de la adaptación de las especies que se plasma en mutaciones y cambios en los genes o cromosomas.

Los algoritmos genéticos hacen uso de las técnicas biológicas de reproducción *mutación y cruce* con el fin de ser usadas en todo tipo de problemas de búsqueda y optimización. Esta aproximación está enfocada a problemas de optimización. Se comienza con una población de partida y se va alterando y optimizando su composición para la solución de un problema particular mediante mecanismos tomados de la teoría de la evolución lo que significa que se va a introducir elementos aleatorios para la modificación de las variables. El material genético o información de los individuos puede ser transmitido a las siguientes generaciones, de diferentes formas que van optimizando el proceso. A través de la reproducción, los mejores segmentos perduran y su proporción crece de generación. Luego de ciertas iteraciones, la población estará constituida por buenas soluciones al problema de optimización (Martínez B. , 2018, pág. 54).

Esta herramienta se usa en las primeras fases del Data Mining, para seleccionar variables que luego se emplearán con otra técnica, como las redes de neuronas o la regresión logística.

Lógica Difusa

Este tipo de técnica nace de la necesidad de modelizar la realidad de una forma más exacta evitando la exactitud. La lógica permite el tratamiento probabilístico de la categorización de un colectivo.

Es aquella técnica que permite y trata la existencia de barreras difusas o suaves entre los distintos grupos en los que categorizamos un colectivo o entre los distintos elementos, factores o proporciones que concurren en una situación o solución.

Series Temporales

Es el estudio de una variable a través del tiempo, a partir de este conocimiento, y bajo el supuesto de que no van a producirse cambios estructurales, poder realizar predicciones. Suelen basarse en un estudio de la serie en ciclos, tendencias y estacionalidades, que se diferencian por el ámbito de tiempo abarcado. Se pueden aplicar enfoques híbridos con los métodos anteriores, en los que la serie se puede explicar no sólo en función del tiempo sino como combinación de otras variables de entorno más estables y, por lo tanto, más predecibles (Martínez B. , 2018, pág. 55).

Redes Bayesianas

Las redes bayesianas son una alternativa para minería de datos, la cual tiene varias ventajas.

- Aprender sobre relaciones de dependencia y casualidad.
- Combinar conocimiento de datos.
- Evitan el sobre ajuste de los datos.
- Pueden manejar bases de datos incompletos.

El obtener una red bayesiana a partir de datos es un proceso de aprendizaje, que se divide en dos aspectos.

- Aprendizaje paramétrico: dada una estructura, obtener las probabilidades a priori y condicionales requeridas.
- Aprendizaje estructural: obtener la estructura de la red bayesiana, es decir, las relaciones de dependencia e independencia entre las variables involucradas.

Las técnicas de aprendizaje estructural dependen del tipo de estructura de red: árboles poli árboles y redes multi-conectadas. Otra alternativa es combinar conocimiento subjetivo del experto con el aprendizaje. Para ellos se parte de la estructura dada por el experto, la cual se valida y mejora utilizando datos estadísticos.

Inducción de Reglas

Las técnicas de inducción de Reglas surgieron hace dos décadas y permiten la generación y contraste de árboles de decisión o reglas y patrones a partir de los datos de entrada. Como información de entrada, tendremos un conjunto de casos donde se ha asociado una clasificación o evaluación a un conjunto de variables o atributos. Con tal información estas técnicas obtienen el árbol de decisión o conjunto de reglas que soportan la evaluación o clasificación.

En los caos en que la información de entrada posee algún tipo de ruido o defecto estas técnicas pueden habilitar métodos estadísticos de tipo probabilístico para generar, para estos casos, árboles de decisión podados y recortados (Martínez B. , 2018, pág. 56).

Sistemas basados en el Conocimiento y Sistemas Expertos

Estos sistemas son un clásico de la IA. Son técnicas que permiten la formalización de árboles y reglas de decisión extraídas de la formalización del conocimiento de los expertos. Tienen motores o *motores de inferencia* los cuales cumplen la función de gestionar las distintas preguntas al ser realizadas de forma que el proceso de decisión sea lo más eficiente y rápido posible (Martínez B. , 2018, pág. 56).

Algoritmos Matemáticos

Existe una amplia gama de algoritmos matemáticos que son especialmente útiles y eficaces en la resolución y tratamiento de problemas muy específicos , puntuales; donde normalmente son incorporados en alguna de aquellas técnicas con el objeto de mejorarlas. No son técnicas que cumplen funciones específicas, a diferencia de las anteriores.

2.2.4. Almacenamiento de Datos

Desde el punto de vista organizacional, en cuanto a plataforma o sistemas orientados, siempre ha existido una problemática en relación la cantidad de información que se emplea dentro de una empresa u organización. Con el fin de darle solución a esta necesidad, aparecen herramientas orientadas al manejo de la información que se emplea dentro de una institución, una de las soluciones principales el almacén de datos (Gavídia, Visern, & Josep, 2017, pág. 7).

La capacidad principal de un almacén de datos es el de almacenar grandes cantidades de información homogénea y fiable; toda la información que se almacena tiene una estructura jerárquica, todo esto con el fin de que el usuario pueda solicitar o hacer consultas estratégicas de información donde pueda obtener un conocimiento, donde tenga la posibilidad de ayudar o solucionar un problema en una empresa u organización.

En 1994, Bill Inmon y R. D. Hackthorn, tienen la iniciativa de crear el termino almacén de datos, donde también lo conceptualizan como una colección de información o datos, toda la información almacenada se encuentra debidamente organizada; además de que, pueden ser datos no volátiles e historiados.

Un almacén de datos se puede explicar como un nuevo tipo de base de datos, y la diferenciación ventajosa que tiene es la importancia que le da a una organización, entorno al manejo de información, y desde un punto de vista estratégico un almacén de datos tiene la posibilidad de llevar un mejor manejo de la organización.

Para la construcción de un almacén de datos los usuarios comúnmente idealizan que es relativamente fácil, pero la creación de un Data Waterhouse tiene una dificultad, que se puede explicar como un problema común y es el de conocer o saber los tipos de datos o qué tipo de información tiene gran relevancia para la organización y

como sería la mejor forma de organizar la información de manera que resulte efectiva y eficiente al momento de hacer consulta, solicitudes, entre otros (Gavídia, Visern, & Josep, 2017, pág. 7).

2.2.5. Metodología CRISP-DM

Dentro del área de minería de datos existen diversas modelos de proceso que han sido propuestos para el desarrollo de proyectos de minería de datos tal como: Sample, Explore, Modify, Model, Asses, (SEMMA); Definir, Medir, Analizar, Mejorar, Controlar (DMAMC); o CRISP-DM (Cross Industry Standard Process for Data Mining); esta última metodología es uno de los modelos más usados en áreas tanto académicas como para grandes empresas.

La metodología CRISP-DM es una de las más usadas y es una referencia para los proyectos orientados a la data mining, y esto lo respaldo un estudio realizado por una página de investigación KDnuggets que indica que durante los últimos años esta metodología es una las principales guías de desarrollo en proyectos de minería de datos. CRISP-DM aparece en el año de 1999 cuando grandes empresas se unen para formar una guía de desarrollo de proyectos de minería de datos tomando como base de referencia las versiones que ya existían de la metodología KDD (Knowledge Discovery from Databases) (Gallardo, 2007, pág. 3).

Esta metodología organiza todo el proyecto de desarrollo de minería de datos en seis fases, cada fase tiene varias que acciones de segundo nivel, que se pueden explicar cómo acciones a realizar con el fin de completar la fase, no existe ningún tipo de formato o propuesta para realizar las tareas que completan la fase:

- Fase 1. Comprensión del Negocio o problema

Se puede explicar como la fase más importante del problema debido a que debe organizar conjuntamente la creación de objetivos y requerimientos desde una perspectiva diferente, específicamente con relación a una empresa o institución, con el fin de transformar esto en objetivos técnicos y en un plan de proyecto. Debemos tomar en cuenta el problema que debemos resolver y lo que queremos obtener mediante minería de datos (Gallardo, 2007, pág. 4).

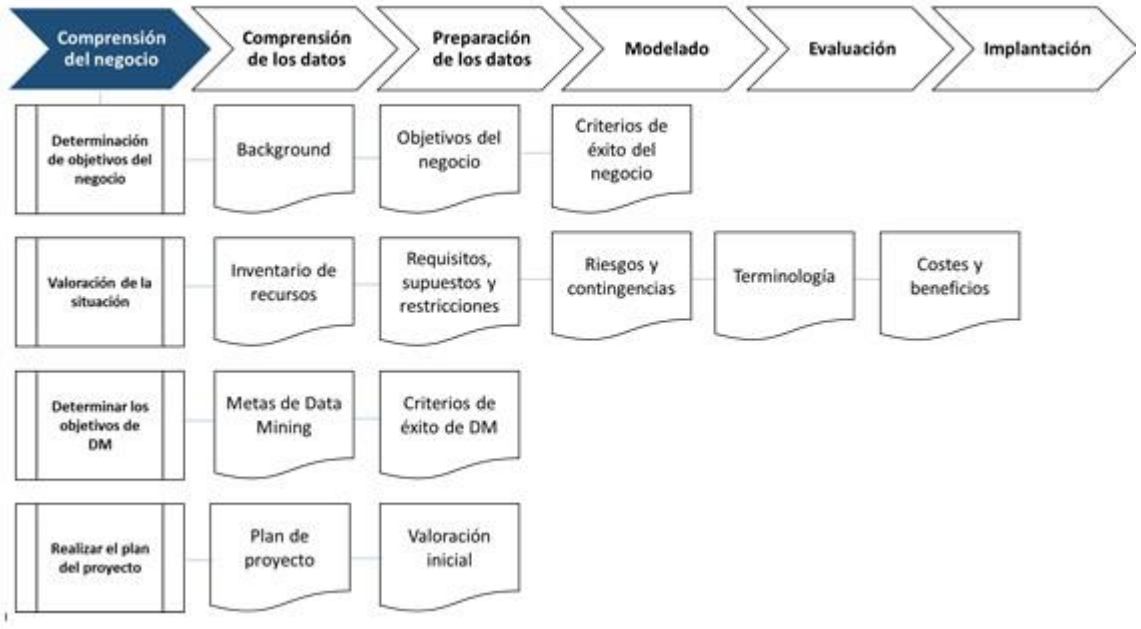


Figura 7. Fase 1 comprensión del negocio
Fuente: Gallardo, J. (2007). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM*. ER-DM.

Esta fase está compuesta por las siguientes tareas:

Definición de los objetivos del negocio

Esta tarea tiene como objetivo tener claro el problema que se tiene para solucionar, y determinar si la opción de usar minería de datos es la correcta, además de definir los criterios de éxito.

Evaluación de la situación

Esta fase ayuda a entender de cómo está la situación actual sin el uso de minería de datos, para hacer esta tarea se de formular algunas preguntas como: ¿Se cuenta con los datos requeridos para iniciar el proceso?, entre otros. Establecemos aquí los requerimientos que tiene el problema.

Determinar los objetivos de Minería de Datos

La función de esta tarea es presentar los objetivos de negocio como objetivos de metas del proyecto de minería de datos.

Plan de Proyecto

Esta es la última tarea de la primera fase, y es desarrollar un plan de proyecto que indica las actividades, técnicas y tiempos en el que se deben ejecutar para cumplir con las metas establecidas.

- Fase 2. Comprensión de los Datos

Dentro de esta fase se recogen los datos iniciales con los que se va a trabajar, se puede explicar como el primer contacto que se tiene con el problema, además, se verifica la calidad de los datos, y con un análisis de datos básico sin ningún tipo de aplicación, solamente observando la data que se tiene ir encontrado las posibles relaciones que podrían surgir. Es una de las fases donde requiere mayor cantidad de esfuerzo y demanda en cuanto a tiempo en un proyecto de minería de datos (Gallardo, 2007).

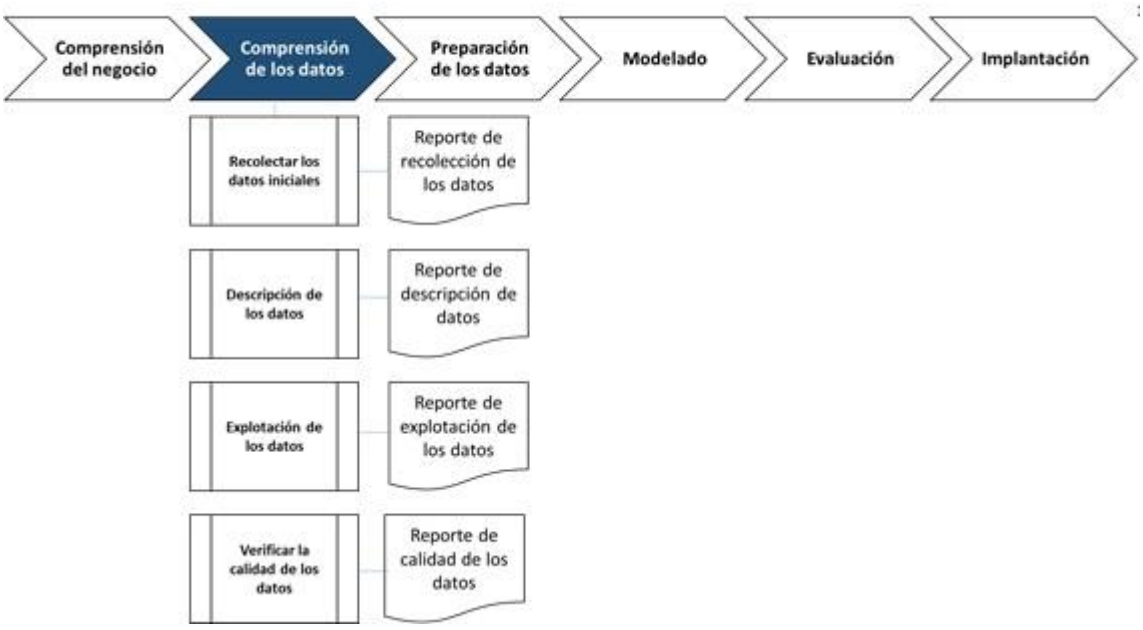


Figura 8. Fase 2 comprensión de los datos
Fuente: Gallardo, J. (2007). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* ER-DM.

Para completar esta fase se deben realizar las siguientes acciones:

Recolección de datos inicial

Es la primera tarea que se debe cumplir y como se menciona a la recolección inicial a los datos y no solo esto si no también transfórmalos a datos que estén listo para el futuro procesamiento que se le va a dar, el objetivo principal que tiene esta tarea es detallar los datos que se tiene, el lugar donde fueron encontrados, como están almacenados, añadiendo a esto describir las técnicas que uso para la recolección de datos.

Descripción de los datos

Luego de la recopilación de los datos iniciales, tienen que ser descritos detalladamente, haciendo referencia al tipo de dato, el formato en el que están, el significado de cada campo, y establecer la cantidad de los datos.

Exploración de los datos

Esta tarea tiene permite realizar un análisis básico a los datos que se tiene, mediante el uso de estadística básica, y a través de ellos permitirá conocer la estructura general de los datos. Al final de esta tarea podremos conocer las propiedades de los datos y un informe de exploración de los datos que contiene gráficos de histogramas, distribución, diagrama de barras, entre otros.

Verificación de la calidad de los datos

En esta tarea se deben analizar los datos, para saber si tienen consistencia en los valores individuales de los datos, si existen datos nulos, también se debe verificar los valores que no tengan consistencia con los datos o incluso valores fuera de rango. La idea de esta tarea en general es asegurar que los datos estén completos y realizar correcciones en caso de ser necesario (Gallardo, 2007).

- Fase 3. Preparación de los Datos

Una vez realizado la recolección inicial de los datos, se tienen que prepararlos y adaptarlos para aplicar las técnicas de minería de datos que van a utilizar conforme avance la metodología, algunas de ellas pueden ser buscar relaciones entre variables, visualización de los datos, u otras medidas para la exploración y análisis de los datos. Las tareas generales que se cumplen en esta fase de cierta manera es tener los datos lista para que posteriormente se le aplique las técnicas de modelado elegida, limpieza de datos, surgimiento de nuevas posibles variables, cambios de formato, entre otros.

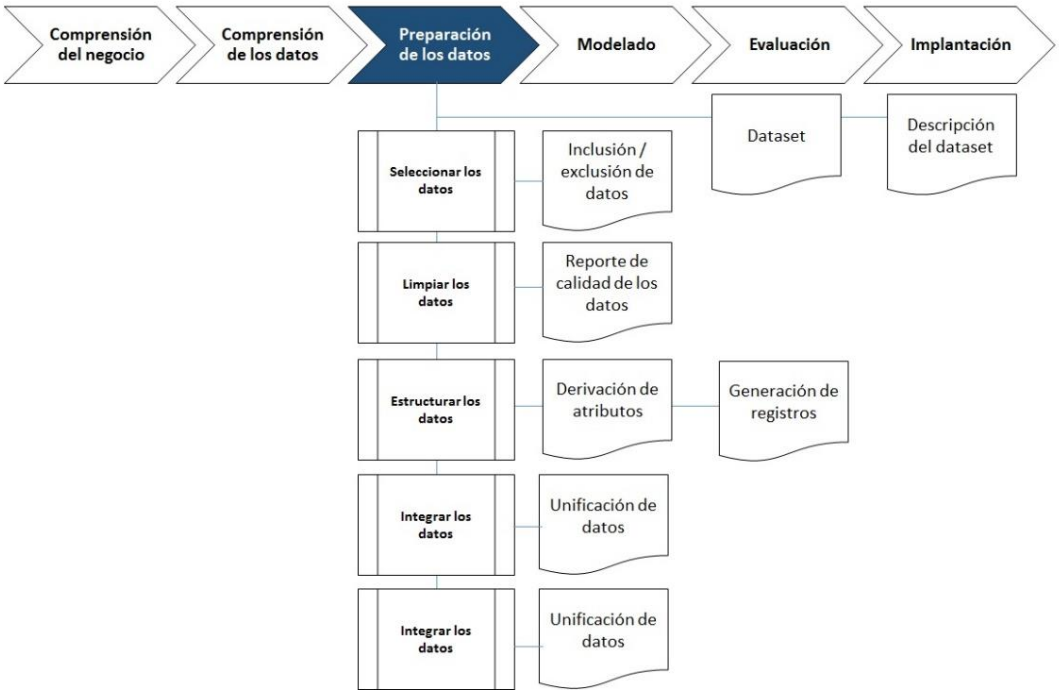


Figura 9. Fase 3 preparación de los datos
Fuente: Gallardo, J. (2007). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* ER-DM.

Esta fase tiene relación con la fase de modelado, en función a la técnica de modelado que se elige, debido a que esto va a ser el punto de inicio de cómo va a ser el procesado de los datos. Las tareas involucradas en fase son las siguientes:

Selección de los datos

Lo que se realiza en esta tarea es elegir un subconjunto de los datos que se adquirieron en la fase anterior y deben estar compuestos con criterios de: calidad de datos, corrección de datos y deben estar completos. Además, dentro del subconjunto que se selección debe existir una relación entre los datos y tipos de datos que va a ser procesados por las técnicas de minería de datos seleccionada.

Limpieza de los datos

Es una acción que complementa a la anterior, y el objetivo es mejorar la calidad de los datos y prepáralos para la fase de modelación, es una tarea que requiere de mucho esfuerzo y tiempo debido a la variedad de técnicas de minería de datos que existen para procesar los datos. Algunas técnicas que cumplen la función de limpieza de datos son: normalización de los datos, reducción de volumen de los datos, tratamiento de valores ausentes, ente otro; también depende del tipo de data que se tenga y en el lugar donde se encuentra almacenados (Gallardo, 2007).

Estructuración de los datos

Se puede explicar como una acción que prepara los datos a fin de tener la posibilidad de crear nuevas variables y nuevos atributos a partir de los datos que ya se seleccionó e incluso transformación de valores que ya existen.

Integración de los datos

Es un complemento de la anterior tarea puesto genera nuevas estructuras datos a partir de los datos ya seleccionados. Siendo más específico pueden crear un variable a partir de otra u otras variables, de igual forma con los atributos y tener la posibilidad de generar nuevas tablas, campos, registros a fin de realizar un resumen en los datos.

Formateo de los datos

El formateo de los datos consiste en realizar un resumen sintáctico de los datos que se elijo, es decir, realizar una transformación de los datos sin cambiar el sentido de los datos iniciales; esto puede ayudar a elegir la técnica de minería de datos adecuada para el procesamiento, se pueden hacer cambios entrono al ordenamiento de tablas y sus registros, ajustes de los atributos, los cambios que realice deben ajustarse a las limitaciones del modelado de minería de datos que se seleccionó.

- Fase 4. Modelado

En esta fase de la metodología CRISP-DM, es cuando se ya elijé de modelado de minería de datos que más se apegue al proyecto de minería de datos que se está trabajando. Algunos criterios de selección de técnicas de minería de datos pueden ser:

- Ser apropiada al problema
- Disponer de datos adecuados
- Cumplir los requisitos del problema
- Tiempo adecuado para obtener un modelo
- Conocimiento de la técnica.

Antes de ejecutar la fase del modelado en sí, existen tareas preliminares que hay que realizar y es de analizar y evaluar los modelados donde muestre cual podría ser la mejor opción para aplicar al proyecto y a los datos en sí. Luego hacer esta labor se iniciar con la generación y evaluación del modelo seleccionado, los parámetros que debe cumplir el modelo deben apegarse a las condiciones en el que están los datos, tomando en cuenta sus características principales, características de precisión y los que resultados que se quiere lograr con el modelo (Gallardo, 2007).

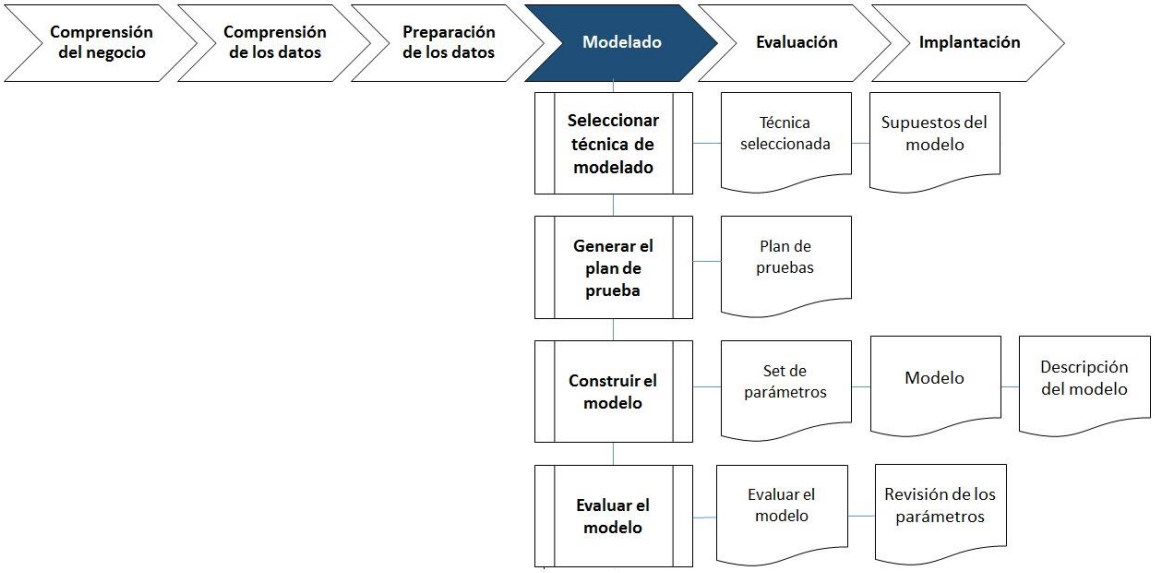


Figura 10. Fase 4 modelado
Fuente: Gallardo, J. (2007). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* ER-DM.

Antes de comenzar con las tareas principales, existen actividades que deben realizarse antes de las tareas que involucran el proceso del modelado en sí, y es que se necesita ya establecer previamente el método de evaluación de los modelos. Debemos tomar en cuenta que el modelado depende del tipo y las características que tienen los datos. Las tareas principales que interviene en la fase del modelado son la siguientes:

Selección de la técnica de modelado

Esta tarea consiste en la elección de la técnica de minería de datos, la técnica que se elija debe cumplir distintos criterios con la finalidad de tener eficacia en los resultados para poder darle solución al problema, los principales criterios para elegir la técnica de modelado apropiada es el objetivo principal del estudio y la relación con las herramientas de minería de datos, debido a que algunas herramientas de DM no contemplen librerías e incluso extensiones que el modelo requiera.

Generación del plan de prueba

Este proceso consiste en establecer un procedimiento donde tiene el objetivo de comprobar la calidad y si es válida la técnica de modelado de minería de datos. Es común en algunos modelos de minería de datos usar la razón de error como medida de calidad, algunos modelos separan los datos en dos conjuntos con el fin de crear dos pruebas, en la primera prueba se construye el modelo basado en el primer

conjunto de datos y mide la calidad del modelo con el segundo conjunto de los datos (Gallardo, 2007).

Construcción del modelo

Luego de tener clara la técnica de modelado, se debe aplicar sobre los datos que fueron preparados en las fases anteriores para crear un o más modelos, cada modelo que se va a generar tienen un conjunto de parámetros donde muestran sus características, la selección de un solo modelo se basa en los mejores parámetros y en los resultados generados.

Evaluación del modelo

Para evaluar el modelado generar, se involucran distintos factores, en primera instancia se encuentran los criterios de éxito preestablecidos y segundo lugar están criterios de técnicos de minería de datos, es decir, seguridad del conjunto de prueba, pérdida y ganancia de datos, entre otros.

- Fase 5. Evaluación

Dentro de esta fase de evalúa el modelo, teniendo en cuenta los criterios de éxito del problema, además de, la fiabilidad calculada del modelo, y solamente se aplica sobre los datos que se realizado el análisis. También es importante revisar el proceso, tomando como base los resultados obtenidos, con el fin de repetir alguna fase o proceso anterior, en el que se pudo haber cometido un error.

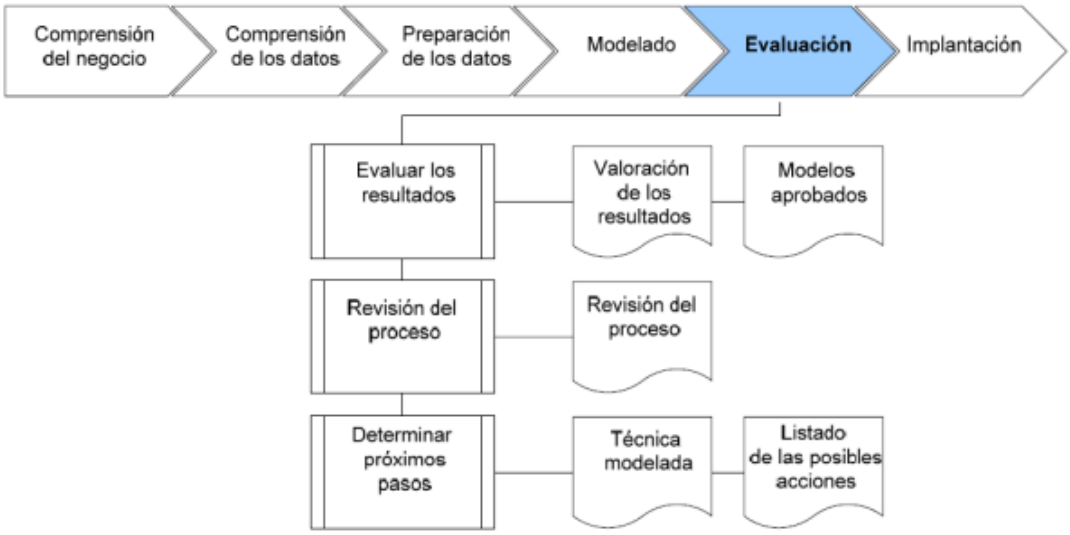


Figura 11. Fase 5 Evaluación
Fuente: Gallardo, J. (2007). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* ER-DM.

Con el fin de darle cumplimiento a esta fase, existen actividades que intervienen en el proceso de evaluación, las tareas son la siguientes:

Evaluación de los resultados

Anteriormente hubo proceso de evaluación, pero eran relación a factores donde se determinaba la exactitud y generalidades del modelo. Esta tarea también involucra evaluación al modelo, pero tiene una relación directa con los objetivos del estudio y determina si el modelo resultante es deficiente o si es aconsejable aplicar el modelo en un problema real.

Proceso de revisión

Esta actividad tiene el fin de darle una calificación al proceso entero de minería de datos, con el objetivo de conocer posibles elementos que pudieran ser mejorados.

Determinación de futuras fases

Es una tarea interesante e importante, y está relacionada a la situación de las fases que han sido aplicadas hasta el momento, si en las fases que se han generado ha dado resultados eficientes y satisfactorios, se podría pasar a la siguiente fase, en el caso contrario se podría hacer un iteración desde la fase de preparación de los datos o de modelación con otros parámetros; incluso en esta fase podría darse la situación de iniciar de cero con un nuevo proyecto de minería de datos (Gallardo, 2007).

- Fase 6. Implementación

Debemos tomar cuenta que en este punto el modelo ha sido construido y validado, luego de ello se transforma en conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el usuario recomiende las acciones basadas en la observación del modelo y sus resultados o cualquier otra situación el proyecto de minería de datos no termina con la implantación del modelo, pues se requiere presentar resultados de forma comprensible para el usuario, con el objetivo de generar un incremento de conocimiento en el usuario.

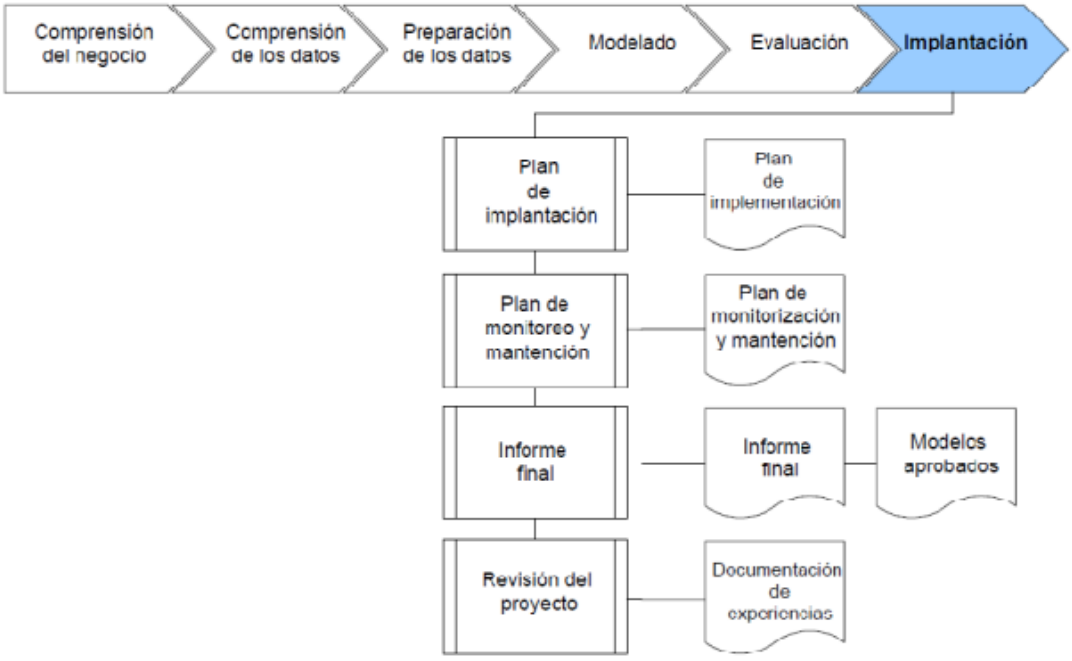


Figura 12 Fase 6 Implementación
Fuente: Gallardo, J. (2007). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* ER-DM.

Las tareas que interviene dentro de esta fase son la siguientes:

Plan de implementación

Lo que se realiza en esta tarea en analizar los resultados de la evaluación para concluir con una estrategia de implementación, donde se debe tomar en cuenta los procedimientos para crear el modelo.

Monitorización y Mantenimiento

Es recomendable generar propuestas de monitorización y mantenimiento debido a que si los modelos resultantes en el proceso de minería de datos son aplicados en un problema que forma parte de una rutina diaria puede ayudar a la retroalimentación del modelo y saber también si se está aplicando correctamente el modelo.

Informe Final

Es la conclusión del proyecto de minería de datos realizado, es un resumen de los puntos importantes de proyecto (modelos, conjuntos de datos, entre otros) y la experiencia conseguida, también se pueden presentar los resultados logrados con el proyecto.

Revisión del proyecto

Esta actividad se puede explicar como una evaluación con relación a lo correcto e incorrecto, es decir, que se hizo bien y que es lo que se requiere mejorar.

2.2.6. Sistema de Gestión de Base de Datos

Un sistema de gestión de base de datos se encarga de gestionar los datos, es el motor que permite a los usuarios tener acceso a los datos y a la información dentro de una base de datos estructurada. Un DBMS generalmente permite organizar archivos, registros, información que se tenga dentro de una empresa o institución, con la finalidad de proporcionar a los usuarios finales más acceso y control sobre sus datos. El sistema de gestión de base de datos les permite a los usuarios manipular los datos incluyendo crearlos y editarlos.

Algunes de los sistemas de gestión de bases de datos que se encuentran en tendencia y son los más utilizados en el mercado actual son:

- Microsoft SQL Server
- Oracle
- MySQL
- PostgreSQL

¿Por qué Excel?

Se ha seleccionado Excel con el fin de almacenar toda la información que corresponden a los procesos de alojamiento y gasto turístico mediante los cuales se obtiene la información acerca de la demanda turística que ha tenido la provincia. Una de las ventajas que provee Excel como base de datos es el grado de utilización que ha tenido a lo largo de los años, por muchas personas como un repositorio de datos. Si se diseña y organiza toda la información dentro del libro que provee Excel le brinda al usuario la posibilidad de hacer consultas, modificar y manipular los datos de manera fácil y eficiente como cualquier otro software dedicado al DBMS, además de ello permite tener la información estructurada y organizada.

Microsoft Excel tiene la capacidad de mejorar la productividad al facilitar el tratamiento de datos con las funciones que provee como: ordenar, filtrar, buscar, entre otros. Es una de las herramientas más utilizadas por el reporting de datos, con el fin de facilitar información de utilidad para tomar decisiones.

A continuación, se muestra una comparación de los sistemas de gestión de bases de datos, indicando que Microsoft Excel tiene la capacidad de estar a la altura de muchos motores de bases de datos e incluso mejor.

Tabla 4. Análisis comparativo de distintos sistemas de gestión de bases de datos

	Características	Ventajas	Desventajas
Excel	Capacidad para almacenar grandes cantidades de datos. Herramientas de análisis para segmentar y desglosar datos. Ejecutar cálculos sofisticados que devuelvan los datos que necesite.	Permite conectarse a una amplia variedad de orígenes de datos, como bases de datos de Access, SQL Server y Analysis Services. Herramienta de análisis de los datos. El análisis de hipótesis le permite ejecutar diferentes escenarios en los datos.	Excel es más adecuado para datos numéricos. Ineficiente en ahorro de recursos.
Microsoft SQL Server	Sistema de administración de bases de datos relacional comercial. Desarrollado en C y C++. Sistema de fácil integración del sistema de gestión de bases de datos con cualquier dispositivo móvil.	Reconocido por su sistema de protección, clasificación y supervisión de los datos. Tiene un enfoque disciplinado para la gestión de los datos.	Gestor de base de datos de pago. Pésima implementación de los tipos de datos y variables. Soporte solo para Windows.
Oracle	Sistema de administración de bases de datos multi-modelo. Capacidad para ejecutar transacciones online (OLTP) y almacenamiento de datos (Data warehousing). Desarrollado en Assembly language, C y C++	Puede ejecutarse en cualquier sistema operativo. Multiplataforma. Permite el uso de particiones para hacer consultas e informes.	Gestor de base de datos de pago. Incompatibilidad y complejidad.
MySQL	Sistema de administración de bases de datos relacional de código abierto y gratuito. Desarrollado en C y C++.	Puede ejecutarse en cualquier sistema operativo. Facilidad de configuración e instalación.	No existe mucha documentación con relación a las herramientas que posee. No es intuitivo.

PostgreSQL	Sistema de administración de bases de datos relacional. Desarrollado en C Estabilidad y confiabilidad, además de ello tiene gran variedad de extensiones .	Velocidad al realizar las operaciones, lo que le hace uno de los gestores con mejor rendimiento. Gran escalabilidad, capaz de ajustarse al número de CPU y a la cantidad de memoria disponible de forma óptima. Posee pgAdmin una herramienta gráfica con la que podemos administrar nuestras bases de datos de forma fácil e intuitiva.	Es relativamente lento en inserciones y actualizaciones en bases de datos pequeñas. Se necesita un nivel medio de conocimiento, para ser usado. No posee mucha documentación.

Una de las razones principales por las cuales se utilizado Microsoft Excel para almacenar la información de los procesos de control que maneja el Ministerio de Turismo de la Zona N°1, es porque tiene la capacidad de organizar la información y dar la posibilidad de hacer consultas conforme lo requiera, además por características técnicas de los ordenadores que maneja en esta institución y por el nivel de conocimiento y la experiencia que tiene con Microsoft Excel a diferencia de otros sistemas de gestión de bases de datos.

2.2.7. Herramientas de Minería de Datos

En esta sección observamos distintas herramientas que nos permiten aplicar técnicas relacionada a minería de datos, con el uso de librerías y funciones que proveen las herramientas, esto puede variar conforme a cada herramienta, debido a que existen herramientas con versiones gratuitas que en algunas ocasiones limitan el uso completo de sus funciones, y de igual forma existen software de minería de datos en versiones pagas, que provee más métodos de análisis y funciones extra a diferencia de la versión gratuita, el objetivo que tiene este apartado es describir las herramientas más populares para minería de datos con el fin de determinar la o las herramientas

que pueden brindar un mejor apoyo a la propuesta de la investigación y estrictamente a la parte de minería de datos.

En la siguiente figura podemos observar información acerca de un estudio realizado por KDnuggets, que se trata de una página web dedicado a la recopilación de información acerca de la Minería de Datos y Gestión del Conocimiento. En la investigación que realizó KDnuggets muestra las principales plataformas que se ha utilizado para el análisis y ciencia de datos y aprendizaje automático. Además de ello, muestra las tendencias de las mejores herramientas para minería de datos entre los años 2017 y 2019.

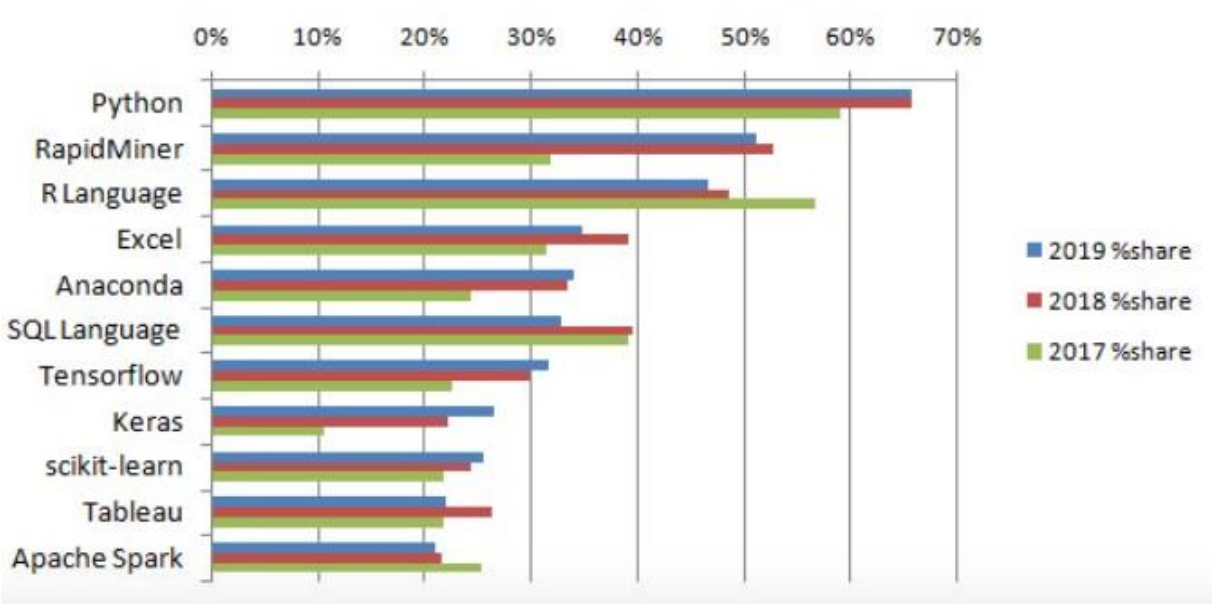


Figura 13. Herramientas de Minería de datos habitualmente usadas entre 2017-2019
Fuente: KDnuggets. (2019). *Herramientas de análisis/ciencia de datos en los años 2017-2019.*

Este tipo de investigaciones nos aclaran ideas con relación al uso de aplicaciones de data mining más populares que están usando los profesionales en el mercado actual, y nos brindan un panorama sobre que herramienta podría utilizarse, no obstante, la elección del software para minería de datos no necesariamente debe ser la más popular sino la que mejor se adapte a las necesidades y requerimientos que tengamos para que los resultados sean los óptimos y de calidad.

Las aplicaciones que se muestran en la Figura 13, son solamente una muestra de las diversas aplicaciones que existe. De ellos se destacan programas comerciales que

han tenido un impacto importante en el mercado para el análisis de datos y machine learning como:

RapidMiner

Es un software dedicado al análisis y minería de datos, se lo llamaba Yet Another Learning Environment *YALE*. Esta plataforma es utilizada en gran medida para realizar investigaciones y en aplicaciones para empresas y organizaciones, usa un entorno gráfico con el que le permite desarrollar procesos de análisis de datos a través del encadenamiento de operadores.

Consta de más de 500 operadores para realizar los procesos de minería de datos incluyendo las funciones que se requiere para la entrada y salida de resultados, y de igual forma la visualización de resultados. Este software tiene la posibilidad de utilizar algoritmos de otras herramientas dedicada al análisis de datos, como es el caso de Weka. Esto es una de las razones por las que RapidMiner es una las plataformas open source más populares en el mundo para la aplicación de minería de datos (Beltrán & Povéda, 2018).

Algunas otras características que tiene RapidMiner son:

- Software gratuito que posibilita el análisis de datos distribuido bajo una licencia GPL.
- Puede usarse de diversas formas como su línea de comandos o GUI, en incluso usar su librería para hacer llamadas en otras aplicaciones.
- Posee una interfaz gráfica mediante la cual permite visualizar los datos con el uso de sus operadores y herramientas.
- Consta de módulos donde le permiten al analista integrar leguajes de programación, con el fin mejorar el análisis de los datos, como es el caso de R y Python.
- Los procesos de análisis de datos que se realiza los puede representar mediante ficheros XML.

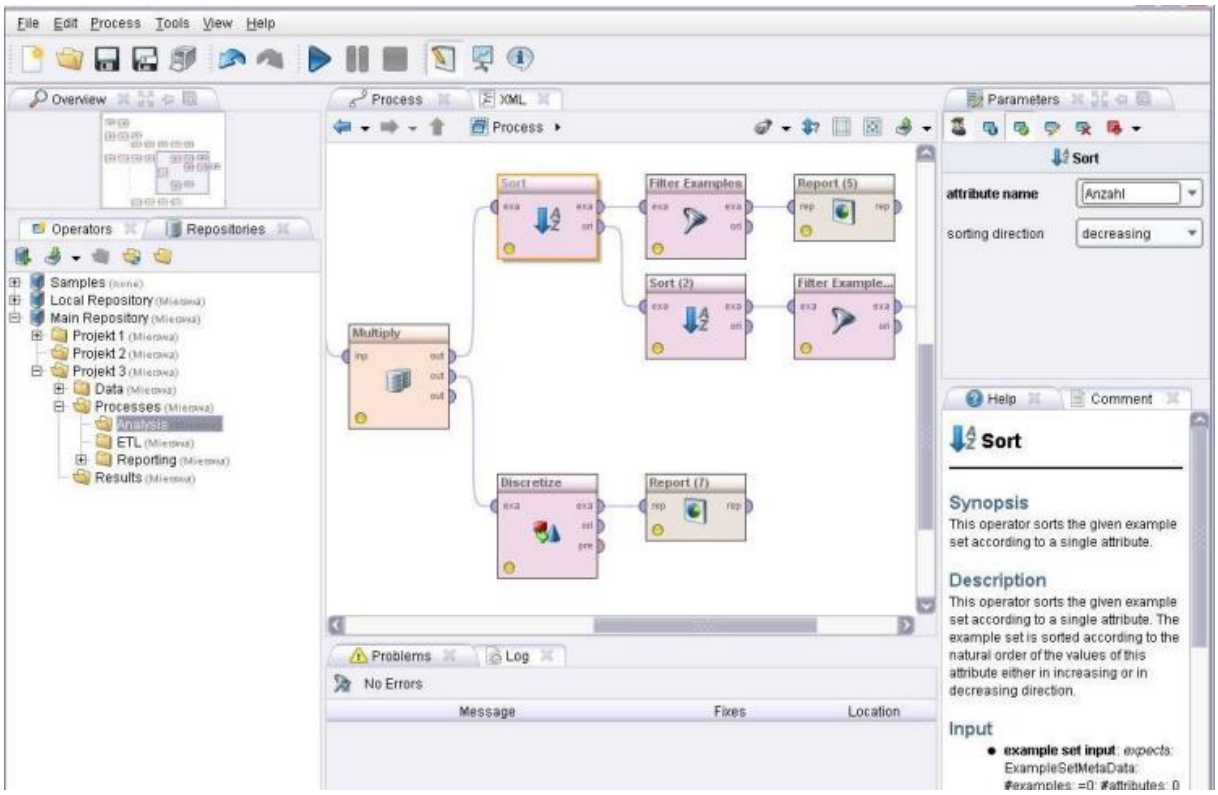


Figura 14. Interfaz gráfica que proporciona RapidMiner

Fuente: Bolaños, C., Cusba, J., Martínez, L., y Caicedo, C. (2018). *Herramientas de analítica para la exploración de datos.* Bogotá: Ministerios de Tecnologías de la Información y las Comunicaciones.

R Language

R es una herramienta de libre distribución, calificada como una plataforma eficiente para el análisis de datos, tiene sus bases en el software estadístico llamado S-plus, tiene una gran similitud a la aplicación de Matlab debido al manejo de las matrices y variables. Tiene gran utilidad para la aplicación de un análisis estadístico, transformación y manipulación de los datos. Debido a las bibliotecas que maneja, actualmente son 2337 librería desarrolladas con R que cubre multitud de campos para la aplicación de minería de datos (Rodríguez & Díaz, 2019).

El lenguaje R proporciona una amplio y profundo análisis de datos, que no cualquier software de orientados a la minería de datos posee, más allá si es versión gratuita o de pago. R es un lenguaje de programación Delaware, lo que significa que realiza una programación para cálculo de estadística de alto grado, lo que significa que está altamente calificado, y no está disponible para muchas aplicaciones. Cabe

mencionar que R tiene un bajo enfoque cuando se trata de investigaciones con motivos académicos (Recalde, Baldeón, Gaibor, & Toasa, 2020).

Características que posee R:

- Proporciona una amplia gama de herramientas para la generación de gráficos y el análisis de los datos.
- Tiene la posibilidad de cargar nuevas bibliotecas y paquetes que le permiten al usuario mejoran la funcionalidad de los análisis.
- R estudio se divide en cuatro cuadrantes, para facilidad del análisis; el primero es en referencial al historial del código, el segundo cuadrante el espacio de trabajo, luego la sección donde se introduce el código y el cuarto cuadrante es donde se puede visualizar los gráficos.
- R tiene la capacidad de integrar distintas bases de datos.
- R consta de un grupo de bibliotecas que le permiten al usuario usar lenguajes de programación como Python, Perl, Ruby.
- Permite el uso de proyectos generados con Java y .Net

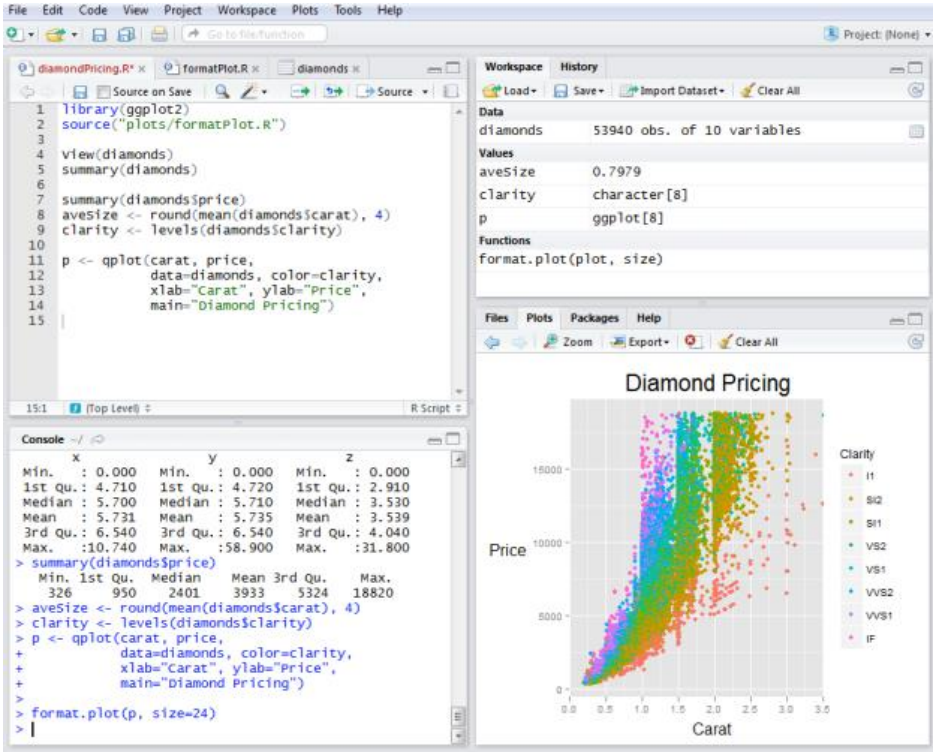


Figura 15. Interfaz gráfica de RStudio
Fuente: Bolaños, C., Cusba, J., Martínez, L., y Caicedo, C. (2018). *Herramientas de analítica para la exploración de datos.* Bogotá: Ministerios de Tecnologías de la Información y las Comunicaciones.

Anaconda

Es un software dedicado a la ciencia de datos, disponible para todos los usuarios que la necesiten, no que requiere de ninguno paga ya que es una plataforma comercial. Anaconda es open source, la distribución la realiza Python y R, se define como una aplicación de alto rendimiento. Constas de más de 100 librerías, dedicadas específicamente al análisis y ciencia de datos; además tiene la capacidad de acceder a más de 720 paquetes que permitan brindar nuevas funciones para el análisis de datos (Bolaños, Cusba, Martínez, & Caicedo, 2018, pág. 35).

La interfaz gráfica de Anaconda provee a los usuarios que realizan trabajos relacionados con la ciencia de datos, un mejor rendimiento con relación a recursos para desarrollar y compartir los resultados de los análisis de datos al público. Una de las características primordiales que tiene este software es poder tener la oportunidad de compartir los archivos que se generan en un proyecto de ciencia de datos con el fin de que el equipo de trabajo puede aportar mejoras o funciones para el análisis y sirva de apoyo para mejorar la productividad y pueda asegurar la consistencia en los flujos de trabajo.

A continuación, se destacan algunas de las funciones que tiene Anaconda:

- Posibilidad de generar un análisis de interacción.
- Generar y analizar flujos de trabajo.
- Software de alto rendimiento.
- Los análisis que realiza se los cataloga como alto rendimiento, ya que tiene bases de Python y R.
- Capacidad de generar una interfaz gráfica para los resultados de los análisis de datos.

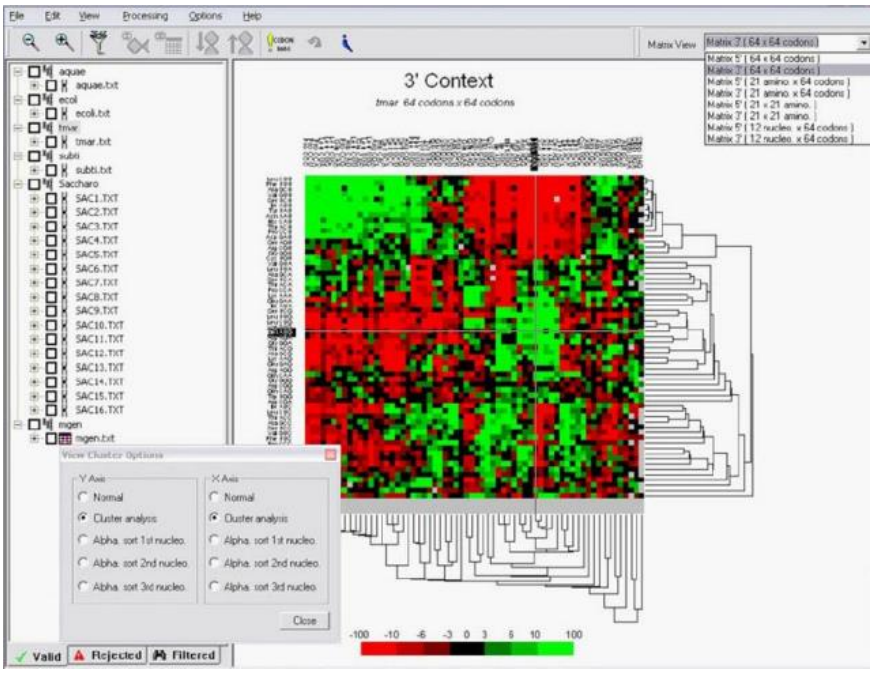


Figura 16. Interfaz gráfica de la herramienta Anaconda
Fuente: Bolaños, C., Cusba, J., Martínez, L., y Caicedo, C. (2018). *Herramientas de analítica para la exploración de datos.* Bogotá: Ministerios de Tecnologías de la Información y las Comunicaciones.

Se han descrito algunas de las herramientas que se han mantenido en tendencia en el mercado actual para el análisis de datos y machine learning, no obstante, existen otras herramientas que han sido tendencia en otros años, pero ha ido disminuyendo con relación a su usabilidad más no en su eficiencia para las ciencias de datos, ya que, en el mercado actual, siguen siendo aplicadas. En el año 2017 la plataforma KDnuggets dedicada a la investigación de Minería de Datos y Gestión del Conocimiento, público las herramientas que han sido tendencia entre los años 2015 y 2017.

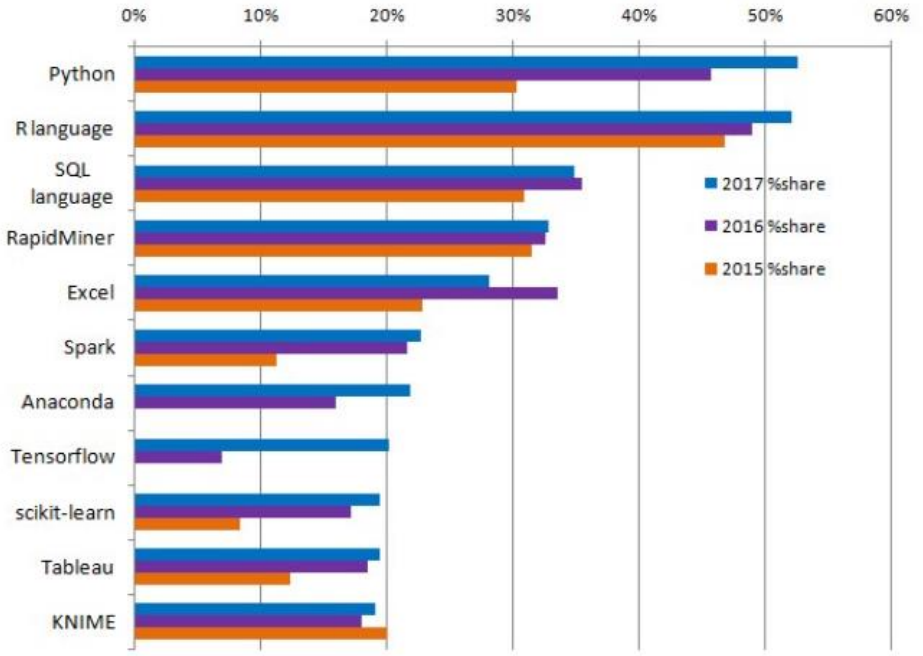


Figura 17. Herramientas principales para minería de datos entre los años 2015-2017
Fuente: KDnuggets. (2017). *Herramientas de análisis/ciencia de datos en los años 2015-2017*.

Se menciona esta investigación debido a una herramienta en particular la cual es Knime Analytics, es un software, que ha sido tendencia no solo entre estos años sino desde el 2002, ha sido una de principales herramientas para el data mining, y no solo por capacidad de ejecución que tiene, sino por la forma de generar los modelos y realizar el análisis de los datos, además como muchas otras herramientas es open source, es decir; es de una distribución comercial.

De igual forma en el año 2017 un informe de investigación proporcionado por la empresa Gartner que se dedica a analizar el mercado con relación a las tendencias en herramientas de ciencias datos, es decir; a herramientas de análisis predictivo, minería de datos, estadística y exploración; muestra las tendencias para la data science.

Figura 18 Tendencias de herramientas de ciencias de datos en el cuadrante mágico de Gartner, 2017
Fuente: Noelia, G. (2017): *Herramientas de ciencias de datos.* LinkedIn.

Como se puede evidenciar el software de análisis de datos Knime Analytics es uno de los lideres, para aplicación de estas técnicas, esta herramienta es una de las mejores en cuanto a su capacidad de ejecución y tiene una buena visión a largo plazo es decir van a tener un gran impacto en mercado en los próximos años, tomando en cuenta que se está hablando del año 2017.

Knime

Es una plataforma desarrollada para el análisis de datos, con la posibilidad de generar reportes e integrar la información. Knime Analytics es open source, y consta de varios componentes para machine learning y data mining, todo lo realiza mediate la generación de flujos de trabajo modulares. Posee una interfaz gráfica para la generación los flujos de trabajo y mediante la conexión de los distintos nodos que tienen, realiza el procesamiento, modelado, análisis y visualización de los datos (Bolaños, Cusba, Martínez, & Caicedo, 2018, pág. 26).

Este software consta de más de 1000 formas de realizar un análisis de datos, muchas de ellas son nativas y otras son mediante librerías de Weka y R. A continuación, se muestran algunos de los nodos que tiene Knime para realizar los procesos de ciencia de datos.

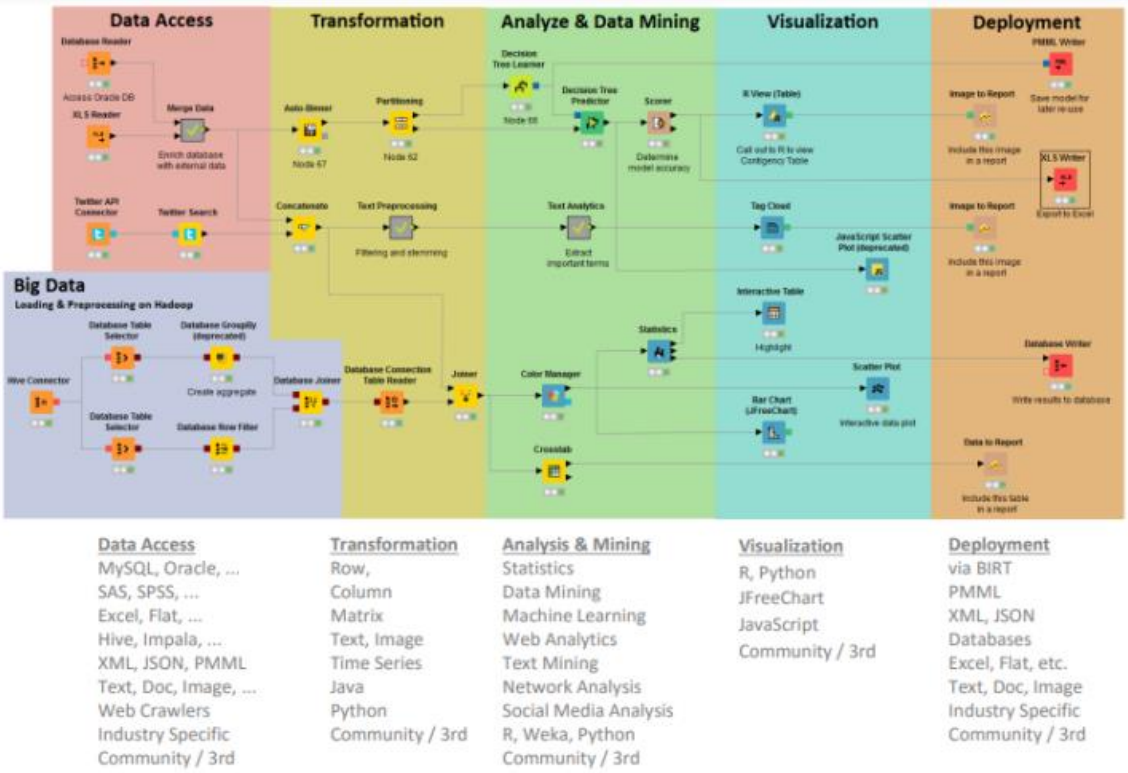


Figura 19. Tipos de nodos que tiene Knime Analytics
Fuente: Bolaños, C., Cusba, J., Martínez, L., y Caicedo, C. (2018). *Herramientas de analítica para la exploración de datos.* Bogotá: Ministerios de Tecnologías de la Información y las Comunicaciones.

Knime está desarrollado en Java, pero tiene bases de Eclipse, consta de mecanismos que le permite añadir extensiones y complementos para mejorar ay adoptar nuevas funcionalidades a los proyectos. La versión comercial consta de todos los módulos para integración, y los nodos que tiene para bases de datos, soportan todos los sistemas gestores de bases de datos más comunes. Alunas de las funcionalidades que tiene Knime son:

- Posibilidad de conexión a otras herramientas de análisis de datos (R, Python, SQL, Java, Weka, Excel, entre otros).
- Visualización de datos interactivos.
- Funciones básicas para matemática y estadística.
- Complementos para integración de métodos de minería de texto y de imágenes.
- Soporte para distintos tipos de datos.

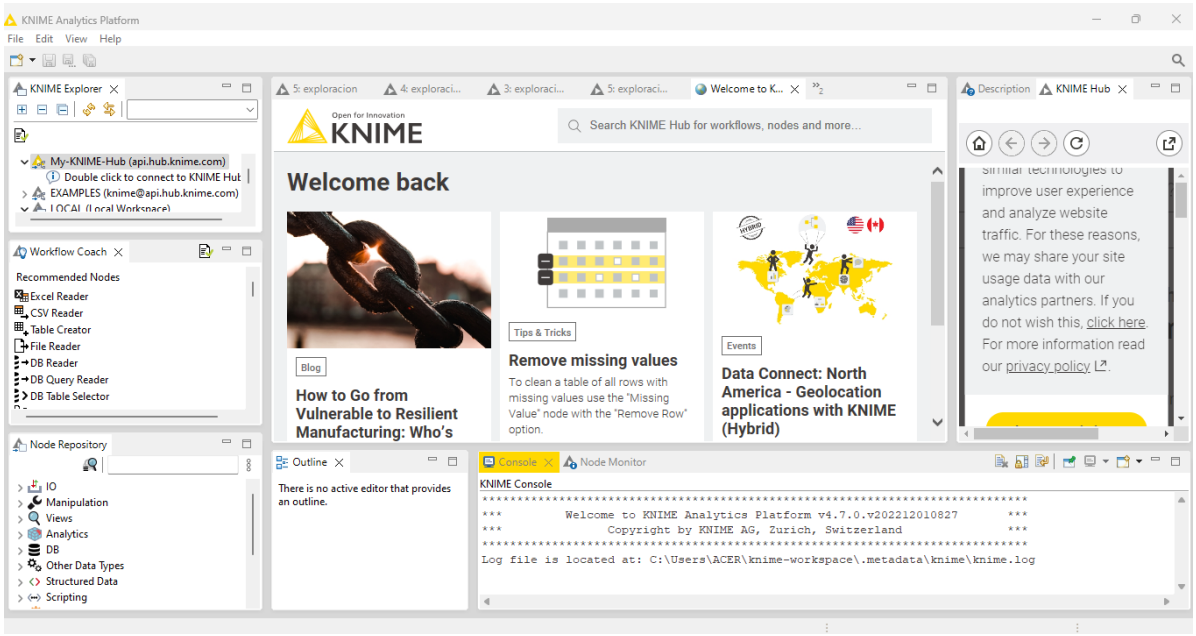


Figura 20. Interfaz gráfica de Knime Analytics

2.2.8. ¿Por qué Knime?

Knime Analytics Platform es una aplicación de escritorio, open source que permite crear aplicaciones y servicios de análisis de datos, es un software muy intuitivo y tiene la posibilidad de adaptarse a todos los desarrollos que se requiera debido a que está en constante evolución, por los creadores y la misma comunidad. Esta herramienta permite generar flujos de trabajo usando la ciencia end to end, de cierta manera se la pude definir como ciencia de datos sin la necesidad de utilizar programación.

Se ha seleccionado Knime en primera parte por algunas de las características mencionadas anteriormente y por la razón de que esta herramienta no solo proporciona una interfaz gráfica mediante la cual se generan los flujos sino como hace que la compresión de los datos, el diseño de los datos de los flujos de trabajo generados y los componentes reutilizables sean más accesibles, no solo esto la hace una herramienta eficaz para este estudio de minería de datos sino porque proporciona las siguientes características, que ayudan al modelado que se quiere proponer, tomando en cuenta el tipo de data que se tiene:

- Herramientas para el análisis de datos
- Manipulación de los datos
- Visualización y generación de informes, sobre los resultados obtenidos.

Knime es una herramienta muy completa y como se menciona está en constante evolución, por lo que está en la capacidad de mezclar herramientas de otros dominios con notas nativas de Knime, este software está en la capacidad de crear flujos de trabajo mediante programación con extensiones basadas en R, Python, Java, Weka, Keras, H20, minería de texto, entre otros.

La principal característica que tiene Knime a la hora de realizar el análisis de datos y la creación de los flujos son los nodos. Los nodos son la parte central del trabajo de Knime, las funciones de los nodos son en función directa a los flujos de trabajo que vayan creando y en ese momento cada nodo pasa por tres etapas diferentes:

- Rojo: el nodo aún no está configurado
- Amarillo: el nodo se ha configurado, pero no se ha realizado ninguna acción en los datos.
- Verde: esta fase se presenta cuando el usuario ejecute los nodos, y se realizan los procesos establecidos para los datos, además, al ejecutar los nodos se combinan los flujos de trabajo y se observa la salida de ejecución de los datos que intervienen.

Hay que mencionar que el software Knime también posee una versión de pago, que no tiene funcionalidades o herramientas adicionales, todos los análisis de datos que pueden hacerse en la versión "gratuita" es igual que en la versión pagada la diferencia esta que ayuda al desarrollo de la aplicación juntamente con la organización (Strate Bi Open Bussiness Intelligence, 2022).

Por otra parte, haciendo un poco más de énfasis en el software que se va a usar en los datos para esta investigación, se destacan varias partes importantes y que son de primera vista al usar Knime:

Knime Explorer

Dentro de esta sección se van a encontrar los archivos del proyecto en el que se está trabajando, e incluso algunos ejemplos practico que nos proporciona Knime.

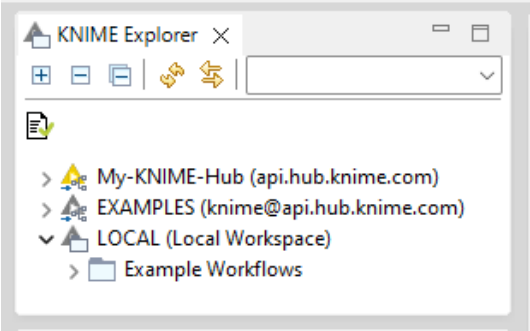


Figura 21. Knime Explorer

Workflow Coach

En esta sección nos presentan algunas sugerencias de nodos que se puede aplicar en los flujos de trabajo con referencia al proyecto con el que se está trabajando, las sugerencias son realizadas por toda la comunidad que ha usado el software, donde mediante un análisis nos muestren nodos que existen en relación con el proyecto que trabaja el usuario.

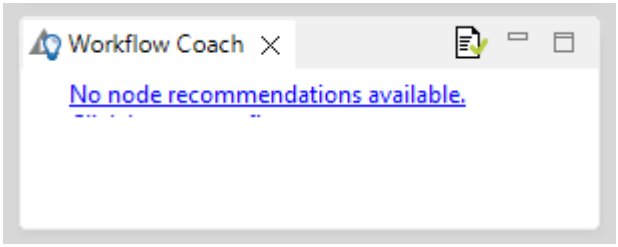


Figura 22. Workflow Coach

Node Repository

Esta ventana nos encontramos con todos los nodos, van a estar organizados mediante categorías.

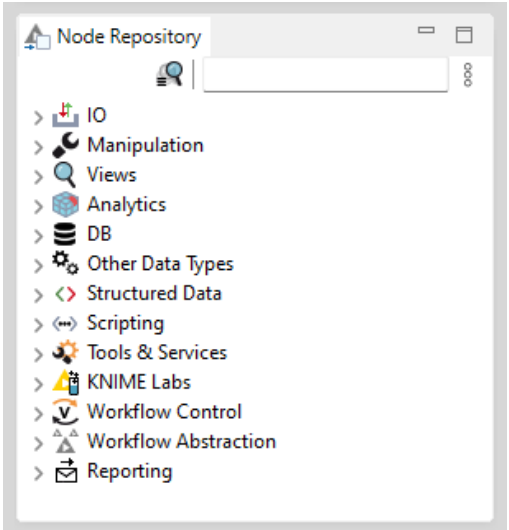


Figura 23. Node Repository

Description

Es una ventana de información, donde describe la utilidad de cómo es el uso de los nodos.

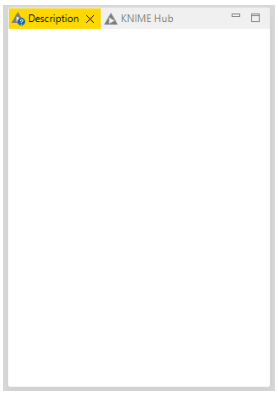


Figura 24. Ventana de Descripción

Outline

Es una sección de navegación, que está directamente enlazada al Workflow Coach para su funcionalidad.

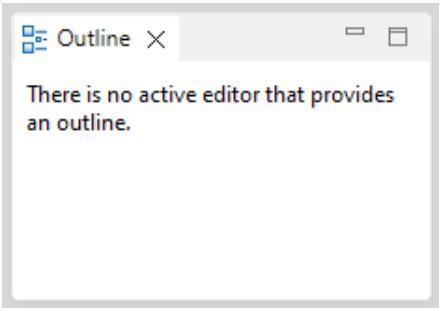


Figura 25. Sección de navegación

Console

Dentro de esta sección nos indicara los registros y errores que se tenga al ejecutar los flujos de trabajo y el proyecto en general.

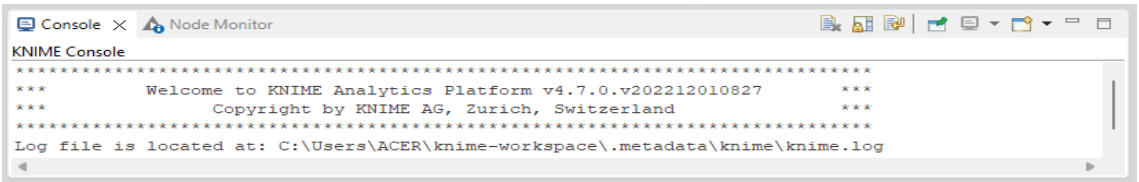


Figura 26. Consola

A continuación, se muestra un análisis comparativo de algunas de las herramientas más utilizadas en el mercado actual por profesionales en minería de datos y gestión del conocimiento:

Tabla 5. Análisis comparativo de las plataformas para el análisis y ciencia de datos

	Características	Funcionalidades	Ventajas	Desventajas
RapidMiner	Software gratuito que posibilita el análisis de datos distribuido bajo una licencia GPL. Puede usarse de diversas formas como su línea de comandos o GUI, en incluso usar su librería para hacer llamadas en otras aplicaciones.	Consta de más de 500 operadores para realizar los procesos de minería de datos incluyendo las funciones que se requiere para la entrada y salida de resultados	Este software tiene la posibilidad de utilizar algoritmos de otras herramientas dedicada al análisis de datos. Software gratuito que posibilita el análisis de datos distribuido bajo una licencia GPL.	Complejidad de la codificación semántica, es necesario unificar los estándares semánticos, otro laborioso proceso.
R Languaje	Proporciona una amplia gama de herramientas para la generación de gráficos y el análisis de los datos. R consta de un grupo de bibliotecas que le permiten al usuario usar lenguajes de programación como Python, Perl, Ruby.	Debido a las bibliotecas que maneja, actualmente son 2337 librería desarrolladas con R que cubre multitud de campos para la aplicación de minería de datos	R estudio se divide en cuatro cuadrantes, para facilidad del análisis; el primero es en referencial al historial del código, el segundo cuadrante el espacio de trabajo, luego la sección donde se introduce el código y el cuarto cuadrante es donde se puede visualizar los gráficos.	No soporta gráficos en tres dimensiones o dinámicos. Su lentitud le resta efectividad y competitividad. Los algoritmos no están unificados.

Anaconda	Los análisis que realiza se los cataloga como alto rendimiento, ya que tiene bases de Python y R. Capacidad de generar una interfaz gráfica para los resultados de los análisis de datos.	Constas de más de 100 librerías, dedicadas específicamente al análisis y ciencia de datos; además tiene la capacidad de acceder a más de 720 paquetes que permitan brindar nuevas funciones para el análisis de datos	Mejor rendimiento con relación a recursos para desarrollar y compartir los resultados de los análisis de datos al público. Los análisis que realiza se los cataloga como alto rendimiento, ya que tiene bases de Python y R.	Debe instalar los paquetes que no están instalados de forma estándar. No existe mucha documentación.
Knime Analytics	Posibilidad de conexión a otras herramientas de análisis de datos (R, Python, SQL, Java, Weka, Excel, entre otros). Visualización de datos interactivos.	Este software consta de más de 1000 formas de realizar un análisis de datos, muchas de ellas son nativas y otras son mediante librerías de Weka y R.	Es una plataforma desarrollada para el análisis de datos, con la posibilidad de generar reportes e integrar la información. Funciones básicas para matemática y estadística.	Mucha de su documentación esta es otro idioma (inglés).

2.2.9. Herramientas de Inteligencia de Negocios

En sección se habla acerca de las principales tendencias de herramientas de inteligencia de negocios, estas herramientas sirven de apoyo para visualizar dashboard de información o tableros de control, mediante los cuales faciliten tomar decisiones. Estas herramientas en la investigación servirán como apoyo para mejorar la descripción e interpretación de los análisis realizados por el software Knime.

La popularidad de las herramientas en inteligencia de negocios nace con el objetivo de convertir datos en información útil. Para tomar decisiones en una empresa o

institución, lo primero que deben hacer es estudiar su información mediante datos que son analizados, procesados, para que luego sean utilizados en beneficio de la organización. la inteligencia de negocios es una forma de aumentar el rendimiento de una institución y la efectivad mediante una inteligente organización de sus datos (Bolaños, Cusba, Martínez, & Caicedo, 2018, pág. 61).

- Tendencias de herramientas en Inteligencia de Negocios

Un estudio realizado por la empresa Gartner presenta un informe sobre las plataformas analíticas de Business Intelligence, a continuación, se muestra el cuadrante mágico de Gartner analizado en el mes de marzo del año 2022.

Figura 27. Cuadrante mágico de Gartner sobre las herramientas de análisis y business intelligence del 2022
Fuente: Noelia, G. (2017): *Herramientas de ciencias de datos.* LinkedIn.

En el informe propuesto de la empresa Gartner menciona que Power BI de Microsoft es la plataforma más completa y con mayor ejecución en el mercado en inteligencia de negocios y análisis de datos. Es el cuarto año seguido que Microsoft lidera las herramientas para BI. Uno de los criterios para determinar cuáles son las plataformas lideres es la capacidad que deben tener las herramientas de análisis y business intelligence para ofrecer insights automatizados para que el usuario final que

manipule la plataforma sea el encargado de decisiones en base a los datos proyectados.

Salesforce Tableau

Es una herramienta de BI que tiene la capacidad de poder tener una visualización interactiva en los datos, de forma que pueda comparar los datos, hacer una filtración de los datos, puede generar conexiones entre las variables de los datos, entre otros. Tableau puede tener accesos a distintas bases de datos como MySQL, Oracle, Microsoft, entre otras; acepta formatos Access, texto y Excel. Utiliza una API, que le ayuda en la extracción de datos, de forma sistemática (Bolaños, Cusba, Martínez, & Caicedo, 2018, pág. 65). Algunas de las características que tiene son:

- Actualización de los procesos de Extracción – Transformación y Carga (ETL).
- Posibilidad de ser gestionado mediante dispositivos móviles.
- Conexión de datos 2.0.
- Capacidad de compartir las visualizaciones de los tableros de contenido.

Figura 28. Interfaz de Tableau
Fuente: Bolaños, C., Cusba, J., Martínez, L., y Caicedo, C. (2018). *Herramientas de analítica para la exploración de datos.* Bogotá: Ministerios de Tecnologías de la Información y las Comunicaciones.

Qlik View

Es una herramienta que se enfatiza en el análisis visual de los datos, se pude conectar con aplicaciones para mejorar los procesos de accesos a los datos. Le proporciona

al usuario la facilidad de acceder a una visualización de los datos más limpia y de más fácil comprensión.

Por otra parte, Qlik posee una herramienta adicional denominada Qlik Indexing Engine *QIX*, esta herramienta le ayuda a asociar los datos, este motor de indexación está calificado como el más potente en el mundo. Es un motor que le permite combinar cualquier tipo y número de fuentes de datos para que sean analizadas y exploradas, además, de que puedan se transformadas en cualquier momento. Algunas de sus ventajas es que tiene una basta documentación, tutoriales, muestras de análisis colaborativos e interactivos, capacidad de generar informes automatizados (Bolaños, Cusba, Martínez, & Caicedo, 2018, pág. 64).

Figura 29. Interfaz Qlik View
Fuente: Bolaños, C., Cusba, J., Martínez, L., y Caicedo, C. (2018). *Herramientas de analítica para la exploración de datos.* Bogotá: Ministerios de Tecnologías de la Información y las Comunicaciones.

Power BI

Es una plataforma que permite realizar un análisis de un negocio, a diferencia de las otras herramientas le permite al usuario tener una vista de los datos más críticos del negocio. Esto que sígnica que Power BI permite, la gestión del estado de una institución o empresa, mediante paneles que se generan; tiene la capacidad de crear informes interactivos, donde el usuario puede manipular la información y tener

acceso a los datos y paneles con la aplicación nativa de Power BI en dispositivos móviles y de escritorio.

Con Power BI existe la posibilidad de usar una API de tipo REST, con el objetivo de integrar la aplicación o servicio con Power BI, lo que le permite tener soluciones con más rapidez. Esta herramienta da soporte al acceso con orígenes de datos locales, bases datos y servicios en la nube (Bolaños, Cusba, Martínez, & Caicedo, 2018, pág. 63). Las características que hacen diferente a Power BI de las demás plataformas pueden ser:

- Generación de gráficos interactivos mediante filtros.
- Capacidad para crear alertas cuando los datos sean más críticos para el negocio.
- Facilidad y agilidad para la creación de los informes.
- Basta documentación en español.
- Incremento de eficiencia.
- Capacidad para más de 65 fuentes de datos: SQL, CSV, XML, web, entre otros.
- Creación de paneles interactivos donde se puede manipular y evaluar la información.

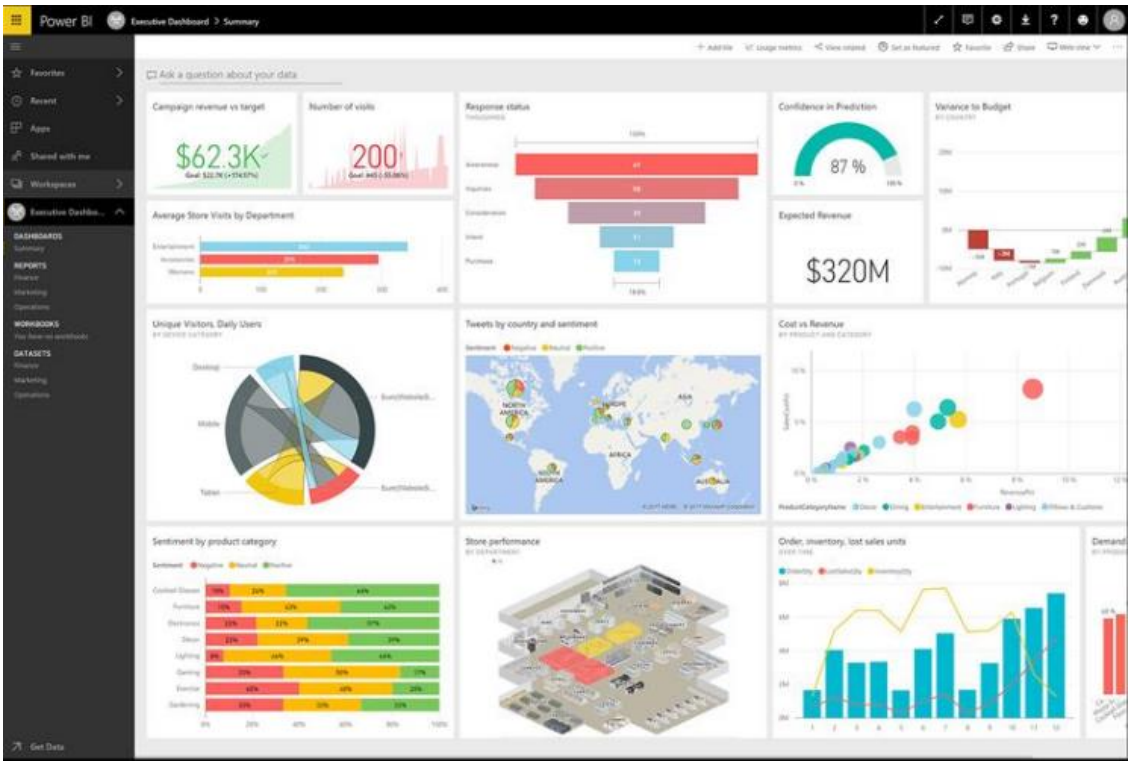


Figura 30. Interfaz de la web de Power BI
Fuente: Bolaños, C., Cusba, J., Martínez, L., y Caicedo, C. (2018). *Herramientas de analítica para la exploración de datos.* Bogotá: Ministerios de Tecnologías de la Información y las Comunicaciones.

Power BI es la herramienta que se ha elegido para evaluar los resultados del modelado que se generará en Knime con los motivos de que; esta plataforma genera paneles interactivos a modo que resulten más fácil de comprender y también por que tiene la capacidad de filtrar los datos y así facilita la toma decisiones sobre la situación de la demanda turística de la provincia en base a los datos de sus procesos. Otra razón por la cual se usa esta herramienta es su libre distribución significa que todas las funciones estarán a disponibilidad, y no habrá limitaciones.

Los paneles que se crearan y los posibles informes ayudaran a las personas que intervienen en los procesos de control de gasto y alojamiento turístico a observar el estado y situación actual de la demanda turística de la provincia del Carchi y también le servirá de apoyo para mejorar el manejo de los datos en sus procesos.

2.2.10. Procesos de Control

En una organización los procesos de control son ciclos repetitivos, están compuestos por etapas;

- **Primera Etapa.-** dentro del proceso de control la primera etapa está dirigida a la fijación de modelos a seguir para tomar de referencia.

Un modelo estándar es un resultado deseado, norma que se establece y que es de estricto cumplimiento, es una medida fiada con anticipación y que se tomará como referencia para medir el comportamiento futuro del proceso.

Para implantar un proceso de control existen diferentes tipos de estándares:

- Estándares de cantidad.- hace referencia a la cantidad de producción hora, aprovisionamiento de materias primas, número de horas de proceso, entre otras.
- Estándares de calidad.- son los patrones aprobados por la empresa con los cuales se verificarán, medirán y evaluará los productos fabricados o adquiridos por la empresa.
- Estándares de tiempo.- son las unidades de tiempos establecidas para el desarrollo de un proceso o una actividad en la organización.
- Estándares de costo.- son la definición de costos establecidos como referentes y óptimos para el desarrollo del proceso productivo (Asturias Corporación Universitaria, 2018, pág. 4).

El establecimiento de estándares constituye al primer paso del proceso de control, pero comúnmente esta tarea es realizada por el proceso de planeación, es decir, en la etapa de inicio del proceso administrativo, como mecanismos para establecer criterios de evaluación en el proceso de control.

- **Segunda Etapa.-** en esta etapa del proceso del control se trata de la evaluación del desempeño y está dirigida a medir el desempeño de los procesos que se está realizando.
- **Tercera Etapa.-** se trata de la comparación del desempeño con el estándar establecidos, este proceso consiste en comparar el proceso que se está realizando con el estándar establecidos, el cual permite identificar si hay variación, error o falla en el proceso.
- **Cuarta Etapa.-** implementa acciones correctivas, que cumplen una función y es el de desarrollar un proceso eficiente de correlación de las desviaciones, error o fallas encontradas en caso de que estos no estén bajo los parámetros establecidos (Asturias Corporación Universitaria, 2018, pág. 4).

Proceso

A un proceso se lo define como un conjunto de actividades de trabajo que están relacionadas entre, posee características particulares como: *inputs* productos o

servicios obtenidos de otros, cumplen con ciertas actividades específicas que implican valor, para obtener ciertos resultados *outputs.*

El proceso es una unidad donde tiene un objetivo, y es de completar un ciclo de actividades que se inicia con un usuario de una institución, empresa u organización y termina con un cliente o usuario común.

Las normas ISO 9000 corresponde a un conjunto de listas de referencia de las mejores prácticas de gestión con respecto a la calidad, que las define la Organización Internacional de Normalización. En una versión de años anteriores en la norma ISO 9001, que es parte de la familia ISO 9000, esta norma es la que más nos compete debido a que se concentra principalmente en los procesos usados para producir un servicio o producto, con el propósito de agregar valor para un tercero en esta transformación. Existen diferentes tipos de procesos como son los procesos industriales que tienen el objetivo de terminar un producto, usando máquinas, recursos humanos. También están los procesos de nuestro interés que son de tipo administrativo, son actividades donde se usa recursos, particularmente el tiempo de las personas, que se transforman, agregándoles valor y generando básicamente un servicio.

- Elementos del Proceso
    - Inputs.- son recursos por transformar, materiales por procesar, personas a formar, informaciones a procesar, conocimientos a elaborar y sistematizar, entre otros.
    - Recursos o factores que transforman.- actúan directamente sobre el anterior elemento, se dividen en dos secciones:
        - Factores dispositivos humanos.- planifican, organizan, dirigen y controlan operaciones.
        - Factores de apoyo.- infraestructura tecnológica como hardware, programas de software, computadoras, entre otras.
    - Flujo real de procesamiento o transformación.- la transformación puede ser física *mecanizado,* entre otros; de lugar, pero también puede modificarse una estructura jurídica de propiedad.

Si el input es información, puede tratarse de reconfigurarla, un ejemplo seria servicios financieros, o posibilitar difusión. Puede también tratarse una transferencia de

conocimientos como en la capacitación, o de almacenarlos en bases de datos, bibliotecas virtuales, entre otros.

En resumen, un proceso es un conjunto de actividades planificadas que implican la participación de recurso materiales y recursos humanos, todo debe estar coordinado para conseguir un objetivo que ya se ha identificado con anterioridad. Se estudia la forma en la que el servicio se diseña, gestiona y mejora sus procesos o acciones para apoyar la política y estrategia con el fin de satisfacer a sus cliente u grupos de interés.

Control

Es una función administrativa la cual se encarga medir, evaluar y corregir el desempeño de la gestión administrativa operativa, y también el desempeño de los empleados, todo con el fin de garantizar el cumplimiento de los objetivos que tiene la empresa, organización o institución. El control se encarga de verificar que los procesos de la empresa se hagan según lo planeado y organizado, que todo se haya realizado según las órdenes dadas, con el objetivo de identificar los errores y las respectivas causas que lo hayan ocasionado, y así corregir y reorientar el o los procesos (Asturias Corporación Universitaria, 2018, pág. 3).

La palabra control se la puede entender como:

- Comprobar o verificar.
- Regular: comparar con un patrón.
- Ejercer autoridad sobre alguien.
- Frenar o impedir el desarrollo de un proceso que este por fuera de los estándares establecidos.

2.2.11. Turismo

La actividad turística, de acuerdo con su planificación y desarrollo, puede ayudar a los pueblos a salir de la pobreza y a construir mejores vidas. La actividad turística tiene potencial para promover el crecimiento económico y la inversión a nivel local, lo cual a su vez se traduce en oportunidades de empleo. La Organización Mundial del Turismo, el turismo representa el 35% de las exportaciones mundiales y más del 70% en los países menos adelantados.

Es tanta la dinámica del turismo en la actividad económico, que la misma amerita ser temática de investigación desde diversos ámbitos *social, económico, administrativo, legal, ambiental, entre otros.* El turismo es una de las pocas actividades humanas que ha sido abordada desde diversas disciplinas *economía, ecología,*

psicología, geografía, sociología, historia, estadística, derecho y ciencias políticas y administrativas (Morillo, 2019, pág. 136). Por ello existen diferentes criterios para definir al turismo y entre las que más podemos destacar serian:

- Turismo es un conjunto de desplazamientos que generan fenómenos socioeconómicos, políticos, culturales y jurídicos.
- El turismo es una afición del hombre a viajar por el gusto de recorrer.
- Turismo es el medio por el que las personas buscan beneficios psicológicos, mediante la suma de tres factores: tiempo e ingresos libres y una consideración positiva o de tolerancia social hacia el hecho de viajar.
- El turismo es la oportunidad del individuo de colmar sus necesidades cuando se encuentra entregado a sus labores.
- Basado en la demanda turística con aspectos económicos, se define al turismo como un fenómeno social que consiste en un conjunto de relaciones por desplazamientos voluntarios y temporales de individuos o grupos que, fundamentalmente por motivos de recreación, descanso, cultural o salud, se trasladan de su lugar de residencia habitual a otro.
- El turismo existe sólo en condiciones de libertad donde los individuos tengan facultades para decidir cuándo y hacia donde desplazarse. No existe turismo cuando el individuo está obligado a desplazarse (Morillo, 2019, pág. 142).

2.2.11.1. Demanda Turística

Es un conjunto de posibles consumidores de bienes y servicios turísticos que buscan satisfacer sus necesidades de viajes. Sean éstos los turistas viajeros y visitantes, independientemente de las motivaciones que los animan a viajar y de lugar que visitan o planean visitar. La demanda turística comprende un grupo heterogéneo de personas con diferentes características sociodemográficas, motivaciones y experiencias, que, influenciadas por sus interés y necesidades particulares, desean, pueden y están dispuestos a disfrutar de las facilidades, atractivos, actividades, bienes o servicios turísticos. Esta demanda está ligada de forma directa con la toma de decisiones que los individuos realizan en la planificación de sus actividades de ocio (Socatelli, 2019, pág. 1).

Los turistas pagan por los servicios que necesitan para disfrutar de su tiempo libre y para sobrevivir en diferentes amientes, pero, ante todo; buscan experiencias y utilidades, explicado de diferente forma la demanda turística es el conjunto de

productos, facilidades, atractivos, servicios y actividades que satisfacen las necesidades, anhelos y deseos del turista (Rigol, 2018, pág. 4).

2.2.12. Ministerio de Turismo Ecuador

El ministerio de Turismo ejerce la rectoría, regulación, control, planificación, gestión, promoción y difusión, a fin de posicionar al Ecuador como un destino turístico preferente por su excepcional diversidad cultural, natural y vivencial en el marco del turismo consciente como actividad generadora de desarrollo socio económico y sostenible (Ministerio de Turismo Ecuador, 2022).

III. METODOLOGÍA

3.1. ENFOQUE METODOLÓGICO

3.1.1. Enfoque

- Enfoque cuantitativo

Según Sampieri, Fernández, Y Baptista (2014). indican que: "un enfoque cuantitativo tiene características particulares como los planteamientos acotados, utiliza estadística, mide fenómenos, hace una prueba de hipótesis" (pág. 3). Un enfoque cuantitativo tiene un proceso deductivo, secuencial, probatorio; cada etapa del desarrollo de la investigación es continua no se puede eludir pasos. El orden es riguroso en este tipo de enfoque, parte de una idea que va acotándose y una vez delimitada se derivan los objetivos y preguntas de investigación, se revisa la literatura y se construye una perspectiva teórica (Sampieri, Fernández, & Baptista, 2014).

- Enfoque cualitativo

En el mismo documento Sampieri, Fernández, Y Baptista (2014). mencionan al enfoque cualitativo explicando : "que este enfoque se conduce básicamente en ambientes naturales , los significados se extraen de los datos, y no se fundamenta en la estadística" (pág. 3). El enfoque cualitativo se guía por temas significativos de investigación. Pero en lugar de que la claridad sobre las preguntas de investigación e hipótesis procesa a la recolección y el análisis de los datos, los estudios cualitativos pueden desarrollar preguntas e hipótesis antes, durante o después de la recolección y el análisis de los datos (Sampieri, Fernández, & Baptista, 2014).

La presente investigación metodológicamente se basa en un enfoque mixto debido a que en primera instancia tiene bases de un enfoque cuantitativo ya que se debe realizar un análisis mediante técnicas como la encuesta y el uso de instrumentos como el cuestionario estructurado que deben ser aplicados a los usuarios que están involucrados en los procesos de control de la demanda turística con el fin de analizar y tabular esta información a través de la estadística descriptiva y por consiguiente interpretar estos datos que ayudaran a identificar algunos aspectos como: el manejo

de información, tiempo de ejecución de los procesos, si ha existido errores al realizar estos procesos, uso de técnicas para los controles de los procesos, entre otros.

Por otra parte, la investigación tiene un enfoque cualitativo, ya que se debe analizar y extraer cualidades con el fin de interpretar la información que posteriormente se transforme en conocimiento útil para los procesos de control de demanda turística que maneja el ministerio de turismo y esto con el objetivo de darle un mejor tratamiento a la información además de obtener detalles con respecto a la especificación de los requerimientos, si existe la necesidad de mejorar los procesos, comprender la forma en la que se almacena la información y comprensión de las necesidades del cliente.

3.1.2. Tipo de Investigación

3.1.2.1. Investigación Exploratoria

Según Arias, F. (2012). afirma que: "La investigación exploratoria es aquella que se efectúa sobre un tema u objeto desconocido o poco estudiado, por lo que sus resultados constituyen una visión aproximada de dicho objeto, es decir, un nivel superficial de conocimientos" (pág. 23). Este tipo de investigación están orientados a la formulación más precisa de un problema de investigación. Dado que se carece de información suficiente y de conocimiento previo del objeto de estudio, resulta lógico que la información inicial del problema sea precisa. Esta investigación permitirá obtener nuevos datos y elementos que pueden conducir a formular con mayor precisión las preguntas de investigación (Arias, 2012).

La exploración en la investigación nos aclara el problema de la investigación y hará posible que nos indique una mejor comprensión e identificación de las variables de estudio, además nos pueda explicar la conexión entre ellas, todo esto a través de las distintas herramientas de estudio como una entrevista directas a los usuarios encargadas del proceso, todo esto con el objetivo de analizar de forma más específica el control de los procesos de la demanda turística, donde se podrá determinar cómo se administra y se almacena la información, y hacer un análisis de cómo se manejan los procesos en el MINTUR con el propósito cubrir todos los requerimientos.

3.1.2.2. Investigación Descriptiva

Arias, F. (2012). indica que este tipo de investigación consiste: "en la caracterización de un hecho, fenómeno, individual o grupo, con el fin de establecer su estructura o comportamiento. Los resultados de este tipo de investigación se ubican en un nivel intermedio en cuanto a la profundidad de los conocimientos se refiere" (pág. 24).

La razón por la cual se va a trabajar con este tipo de investigación es que se deben describir los datos y características importantes recolectados mediante las técnicas e instrumentos como es la encuesta aplicados a todos los usuarios que intervienen en este proceso, donde nos permita caracterizar y recoger información útil con el fin de establecer las necesidades que tiene el MINTUR en cuanto al proceso de la demanda turística.

3.1.2.3. Investigación Documental

La investigación documental es un proceso basado en la búsqueda, recuperación, análisis, crítica e interpretación de datos secundarios, es decir, los obtenidos y registrados por otros investigadores en fuentes documentales: impresas, audiovisuales o electrónicas. Como en toda investigación, el propósito de este diseño es el aporte de nuevos conocimientos (Arias, 2012, pág. 27).

Este tipo de investigación permitirá recolectar toda información que sea útil, contribuyendo a sustentar la investigación; con toda la información que se recopile de fuentes primarias y secundarias, como: libros, trabajos de grado, tesis, informes de investigación, revistas científicas, páginas web, entre otros; nos dará una mejor visión del problema que se está trabajando, las variables de estudio y el tema de investigación en general. Por otra parte, también tendremos un panorama más amplio del tema y en general la investigación.

3.1.2.4. Investigación de Campo

Arias, F. (2012). indica que este tipo de investigación consiste: "La investigación de campo es aquellas que consiste en recolectar datos principalmente del sujeto investigado y del ambiente a su alrededor, y no tiene la necesidad manipular o controlar alguna variable" (Arias, 2012, pág. 31)

Se empleará la investigación de campo ya se requerira de la observación sobre los procesos que emplea en el MINTUR para determinar el nivel de turismo que tenido la

provincia, con el fin de realizar un diagnostico general para cubrir todas las necesidades resultante y así proporcionar soluciones al problema.

3.2. IDEA A DEFENDER

La aplicación de un modelo de minería de datos mejorará los procesos control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi.

3.3. DEFINICIÓN Y OPERACIONALIZACIÓN DE LAS VARIABLES

3.3.1. Definición de Variables

3.3.1.1. Minería de Datos

Las herramientas Data Mining son muy útiles para procesar información en gran cantidad, debido a que puede detectar ciertos patrones que pueden ayudar a mejorar tareas o procesos donde exista una buena cantidad de información y sea vital del proceso. Estas nuevas herramientas trabajan con un análisis estadístico y tiene la posibilidad deducir tendencias que existen en los datos, todo esto que hace la minería de datos no se pude detectar mediante una exploración tradicional de la información porque, las relaciones son complejas por la gran cantidad de datos que regularmente se sabe manejar.

3.3.1.2. Procesos de Control

Para entender la variable dependiente y definirlo correctamente, debemos definir "control" y se puede explicar como una función que verifica los procesos de una organización o empresa, de modo que las actividades se cumplan según como se haya organizado y todo esto con el objetivo de identificar posibles errores y las respectivas causas que lo hayan ocasionado, con el fin de corregir y reorientar el proceso. Tomando en cuenta esto, el proceso de control en una organización es repetitivo y constante está compuesto generalmente por cuatro fases; la primera está dirigida a la fijación de modelos a seguir para tomar referencia; la segunda fase está orientada a medir el desempeño de lo que se está realizando; en tercer lugar consiste en una comparación del proceso que se está realizado con un proceso que se ha estándar que se ha establecido, lo que permite identificar si existe alguna variación, erro o fallo en el nuevo proceso; la última fase tiene el objetivo de desarrollar un proceso eficiente de corrección es decir, darle solución de forma eficiente a las posibles fallas o errores encontrados (Asturias Corporación Universitaria, 2018).

3.3.2. Operacionalización de Variables

Tabla 6. Operacionalización de la variable independiente

	Variable	Dimensión	Indicadores	Técnica	Instrumento
Variable Independiente	Minería de Datos	Datos	• Cantidad • Calidad • Validez	• Análisis Documental Documentos, registros	• Computador y sus unidades de almacenaje
		Métodos de Minería de Datos	• Procesamiento • Análisis de resultados • Análisis de información		
		Técnicas de La Minería de Datos	• Algoritmos • Modelos • Evaluación de resultados. .	• CRISP-DM	• Herramientas de Minería de Datos

Observación: La tabla muestra la variable independiente y los indicadores e instrumentos que se aplicaran en la investigación del proyecto.

Tabla 7. Operacionalización de la variable dependiente

	Variable	Dimensión	Indicadores	Técnica	Instrumento
Variable Dependiente	Procesos de Control	Eficiencia	• Nivel de satisfacción del usuario • Tiempo de ejecución	• Encuesta	• Cuestionario
		Almacenamiento de los datos	• Capacidad de información • Fiabilidad • Rendimiento	• Entrevista Estructurada	• Guía de preguntas
		Organización	• Usuarios involucrados en los procesos. • Tiempo en el desarrollo del proceso. • Nivel de complejidad de los procesos.	• Entrevista Estructurada	• Guía de preguntas

Observación: La tabla muestra la variable dependiente y los indicadores e instrumentos que se aplicaran en la investigación del proyecto.

3.4. MÉTODOS UTILIZADOS

Como se ha mencionado dentro del enfoque de investigación, el presente proyecto cuenta con un enfoque cualitativo, por razón de que se describen aspectos de cómo es el manejo de los procesos que se emplea en el Ministerio de Turismo de la Zona N°1para determinar la demanda turística de la provincia del Carchi. Por otra parte, la investigación cuenta con un enfoque cuantitativo, debido a que se analizan e interpretan los datos recolectados mediante el uso de instrumentos previamente seleccionados con el fin tener un mejor enfoque acerca de la investigación e identificar las necesidades reales para dar una solución. Los métodos empleados dentro de la investigación fueron:

3.4.1. Encuesta

El uso de esta herramienta nos permitió conocer y recoger información sobre cómo se manejan los procesos que determinan el nivel de turismo de la provincia del Carchi, además de comprender de forma específica como almacena, administra e interpreta los datos que se recoge de este proceso, por otra parte, de igual forma este instrumento nos indica como es la interacción de los usuarios involucrados para ejecutar el proceso de demanda turística.

3.4.2. Entrevista

Este instrumento nos ayudó a describir las necesidades y comprender como se lleva a cabo los procesos donde se determina la demanda turística de la provincia, realizando la entrevista al responsable de la Dirección de la Zona N°1 del Ministerio de Turismo; dicha entrevista está construida con interrogantes, que amplíen el panorama del estudio, además de que son necesarios para el avance de la investigación. Todos los datos que se recolecten de esta herramienta nos muestran aspectos de cómo controla y maneja el MINTUR los procesos para la demanda turística, es decir, conocer el grado de desempeño, tiempos de procesamiento, nivel de cumplimiento, como se analiza e interpreta la información, entre otros.

3.4.3. Análisis Documental

Se empleó este instrumento con el objetivo de comprender como se analiza, almacena e interpreta la información de los procesos correspondientes al turismo de la provincia en el Ministerio de Turismo, permitió conocer y obtener indicadores acerca del proceso de alojamiento y gasto turístico. Además, se pudo comprender

el flujo de los procesos de interacción con los usuarios involucrados, específicamente los hoteles de la provincia del Carchi.

3.4.4. Análisis y síntesis

Permitió comprender y conocer de forma más específica la realidad de la investigación, encontrar nuevos aspectos e indicadores para realizar los procesos que emplea el MINTUR para determinar el nivel de turismo que ha tenido la provincia del Carchi. Se logró construir nuevos conocimientos y descubrir relaciones aparentemente ocultas, que son aplicables en los procesos; todo esto con la aplicación de las técnicas e instrumentos seleccionados *(entrevista, encuesta, análisis documental)*, además de la ayuda de la investigación bibliográfica sobre procesos, control, demanda turística, turismo, entre otros.

3.5. ANÁLISIS ESTADÍSTICO

El estudio requiere de un análisis estadístico, debido a que se obtuvieron datos de la encuesta aplicada a los usuarios con los que interactúa el MINTUR para realizar el proceso que determina el nivel de turismo de la provincia, la preguntas fueron planteadas mediante la herramienta que provee Google *(Google forms)* con el fin de facilitar la tabulación y obtención de gráficos estadísticos.

Además, se ha realizado la entrevista al responsable de la Dirección de la Zona N°1 del Ministerio de Turismo sobre los procesos que se emplea con la finalidad de profundizar el estudio de la variable de investigación.

3.4.5.1. Población y Muestra

La población está compuesta por 38 usuarios que corresponden a todos los hoteles de la provincia del Carchi, los registros fueron tomados de una base de información que tiene guardada el MINTUR con previos permisos de los usuarios y del mismo ministerio. Para la muestra se escogió un procedimiento de muestreo no probabilístico de tipo intencional debido a que los usuarios seleccionados están escogidos por razón de que ellos trabajan en los procesos junto con el Ministerio de Turismo para determinar la demanda turística que ha tenido la provincia del Carchi.

Para estimar el tamaño de la muestra, se utilizaron criterios estadísticos mediante el uso de la fórmula que nos presenta (Arias, 2012, pág. 86) para hacer el cálculo,

utilizando un nivel de confianza del 91% en zeta crítico, un margen de error del 7% y determinando que la probabilidad de éxito sea del 50%.

$$n = \frac{N \cdot Z_c^2 \cdot p \cdot q}{(N-1) \cdot e^2 + Z_C^2 \cdot p \cdot q}$$

Nomenclatura:

n = tamaño de la muestra

N = total de elemetos que integran la población

Z_c^2 = Zeta crítico → nivel de confianza

e = error muestral

p = probabilidad de éxito

q = probabilidad de fracaso

$$n = \frac{38 \cdot 1.69^2 \cdot 0.50 \cdot 0.50}{(38-1) \cdot 0.07^2 + 1.69^2 \cdot 0.50 \cdot 0.50}$$

$$n = 30.305 \approx \mathbf{30}$$

Usando la fórmula propuesta se obtuvo un tamaño muestral de 30 usuarios, donde se lo identifica como los sujetos de estudio, que específicamente representan a los gerentes de cada uno de los hoteles de la provincia del Carchi, con el fin de recopilar información mediante el uso de instrumentos y técnicas previamente establecidas *(encuesta y entrevista)*.

IV. RESULTADOS Y DISCUSIÓN

4.1. RESULTADOS

4.1.1. Resultados de la encuesta aplicada

Con el fin de avanzar con el estudio, se exponen los resultados obtenidos de la encuesta aplicada a los usuarios que intervienen en el proceso para determinar la demanda turística de la provincia del Carchi, es decir, a cada encargado de cada uno de los hoteles que se encuentran en la provincia. La encuesta estuvo estructurada con el propósito de obtener información necesaria para la construcción de una propuesta acerca de mejorar los procesos mediante minería de datos, a continuación, se presentan las respuestas y el respectivo análisis de cada una de las preguntas formadas:

1. El trato o actitud entre el Ministerio de Turismo hacia el usuario.

Tabla 8. Resultados de Encuesta - pregunta 1

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Muy Bueno	9	0,3	30%
Bueno	9	0,3	30%
Aceptable	9	0,3	30%
Malo	2	0,067	6,7%
Muy Malo	1	0,033	3,3%

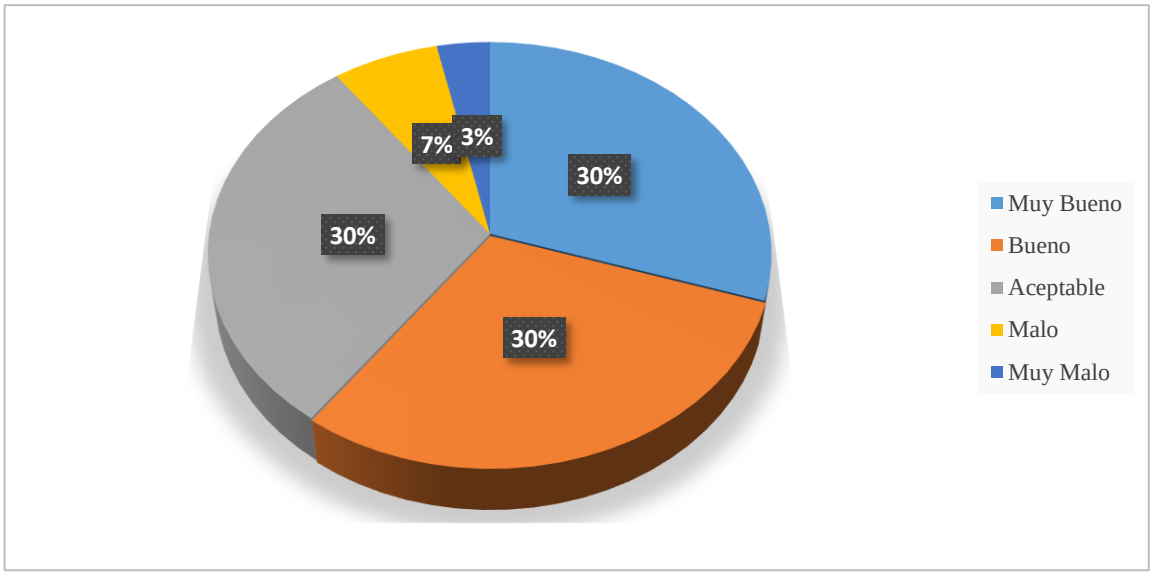


Figura 31. Gráfico de Resultados Encuesta - pregunta 1

Análisis

Tomando en cuenta lo que indica la pregunta acerca del trato y la actitud entre el MINTUR y las personas que se involucran en los procesos que trabajan conjuntamente se puede evidenciar e indicar que existe un nivel de relación estable y buena ya que entre los puntajes de muy buena, buena y aceptable son del 30%. Por otra parte, no significa que la relación alcancé la perfección en cuanto a trato y actitud ya que se obtuvieron datos que son negativos; malo 6,7% y muy malo 3,3% , lo cual significa que existen personas que están inconformes en cuanto a cómo es el trato que reciben, y se demuestra que hay que mejorar los criterios en cuanto a relación, comunicación, comportamiento, con las personas que se trabaja en los distintos procesos.

2. La forma en que se realiza el proceso de alojamiento y gasto turístico.

Tabla 9. Resultados de Encuesta - pregunta 2

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Muy Bueno	4	0,133	13,3%
Bueno	8	0,267	26,7%
Aceptable	9	0,3	30%
Malo	6	0,2	20%
Muy Malo	3	0,1	10%

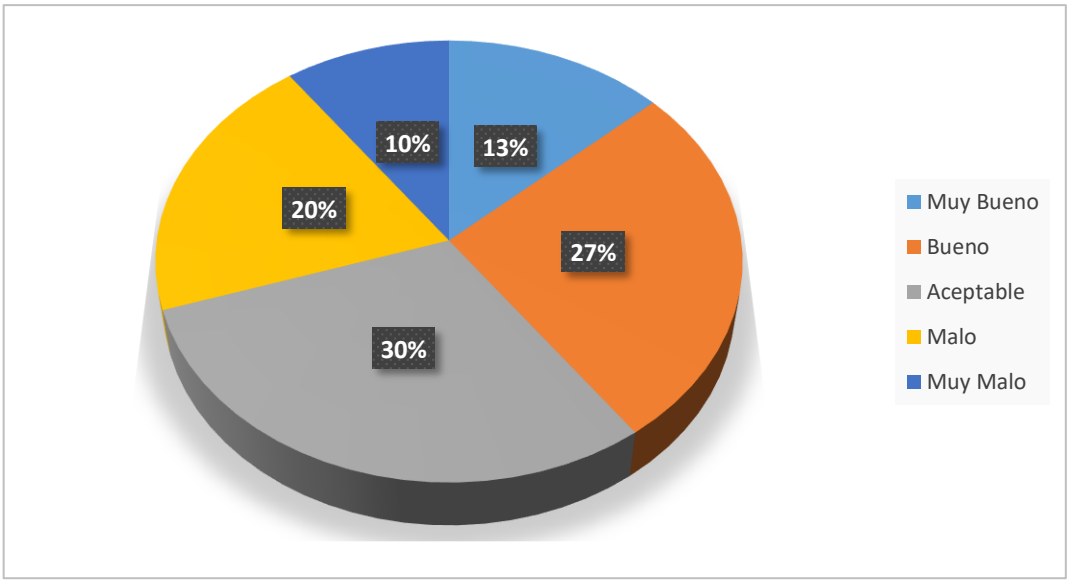


Figura 32. Gráfico de Resultados Encuesta - pregunta 2

Análisis

De acuerdo con los datos recolectados acerca de cómo se realizan los procesos de alojamiento y gasto turístico donde a través de esto se determina la demanda turística se puede explicar que es aceptable ya que es el puntaje con un mayor porcentaje siendo el 30%, se puede indicar que la forma en que se ejecuta este proceso está en un punto intermedio y no excelente debido al puntaje de muy bueno que no es el óptimo siendo de un 13%, evidenciando que hay ciertos aspectos y características acerca de proceso que hay que mejorar, y no solo por el porcentaje de muy bueno, que es bajo sino también que entre *malo* y *muy malo* suman un porcentaje del 30% que es un valor alarmante y de tomar en cuenta debido a que pueden ocasionar o generar errores e inconvenientes a la hora de realizar los procesos de alojamiento y gasto turístico.

3. Como califica la agilidad o rapidez con que el MINTUR resuelve problemas acerca del proceso de alojamiento y gasto turístico.

Tabla 10. Resultados de Encuesta - pregunta 3

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Muy Bueno	4	0,133	13,3%
Bueno	5	0,167	16,7%
Aceptable	9	0,3	30%
Malo	8	0,267	26,7%
Muy Malo	4	0,133	13,3%

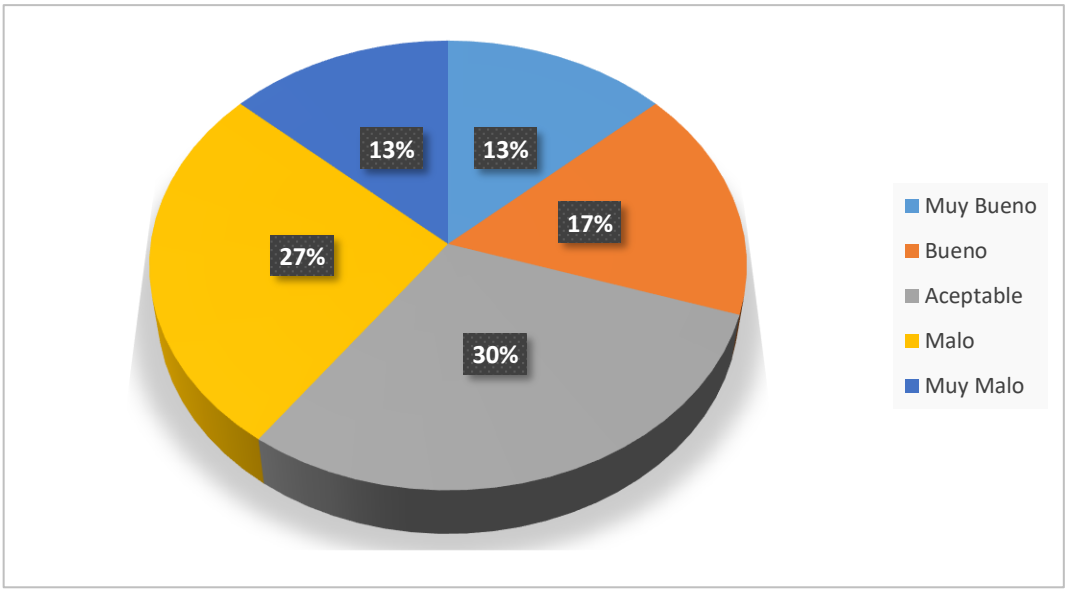


Figura 33. Gráfico de Resultados Encuesta - pregunta 3

Análisis

Tomando en cuenta que dentro de los procesos de alojamiento y gasto turístico se pueden generar problemas e incluso tener ciertas dudas de cómo realizar los procesos, el Mintur tiene la obligación de solventar los problemas con eficacia y agilidad, esta acción se la puede calificar regularmente mala, debido a la evidencia de los datos donde se muestra que en entre *malo* y *muy malo* suman un porcentaje del 40%, lo que ocasiona que los usuarios al tener un problema y no tener un respuesta ágil por parte del Mintur ejecuten los procesos con errores, son aspectos y características que hay que mejorar. Por otra parte, la situación no está todo en negativo, debido al porcentaje de *muy bueno y bueno* que se obtuvo siendo de 30%; esto significa que la forma y el tiempo en que solucionan problemas el Mintur acerca de los procesos de alojamiento y gasto tiene ciertos aspectos por corregir y mejorar con el objetivo de ayudar a todos los usuarios.

4. **La interacción con el MINTUR a través de medios de comunicación (email, redes sociales, página web) le ayuda a resolver problemas con el proceso de alojamiento y gasto turístico.**

Tabla 11. Resultados de Encuesta - pregunta 4

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Siempre	2	0,067	6,7%
Casi Siempre	9	0,3	30%
A veces	9	0,3	30%
Casi nunca	7	0,233	23,3%
Nunca	3	0,1	10%

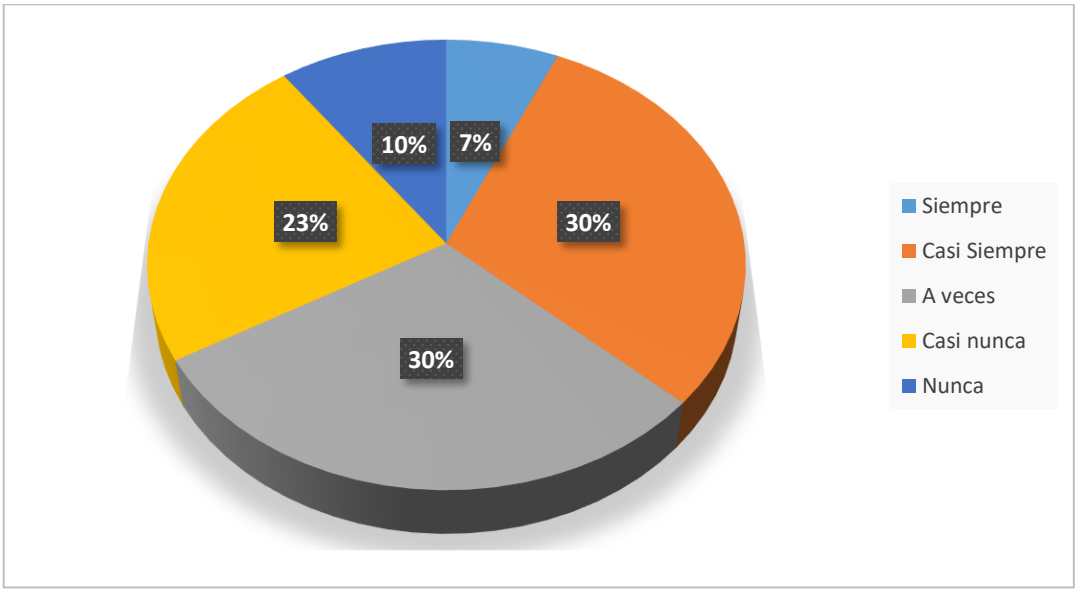


Figura 34. Gráfico de Resultados Encuesta - pregunta 4

Análisis

La comunicación que existen es importante, y aún más importante es la forma y el medio en que se realiza, se obtuvieron datos donde se evidencia que el uso de medios de comunicación, en algunos casos si le ayuda para poder solucionar dudas o problemas que puedan tener con los procesos de alojamiento y gasto turístico, debido a los resultados que se obtuvo el cual el porcentaje está en nivel aceptable siendo un 36,7% entre *siempre y casi siempre*. En general se puede indicar que la comunicación entre el Mintur y los usuarios a través de medios de comunicación en tendencia está en un nivel intermedio es decir que hay que aprender a manejar mejor el uso de estas nuevas fuentes de comunicación y sacar el máximo provecho para que la interacción sea la óptima con el fin que se requieres solucionar problemas o dudas que se puedan generar por parte de los usuarios que realizan el proceso de alojamiento y gasto turístico.

5. ¿Realiza frecuentemente preguntas o reclamos acerca del proceso de alojamiento y gasto turístico?

Tabla 12. Resultados de Encuesta - pregunta 5

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Diario	1	0,033	3,3%
Semanal	7	0,233	23,3%
Mensual	10	0,333	33,3%
Trimestral	6	0,2	20%
Anual	6	0,2	20%

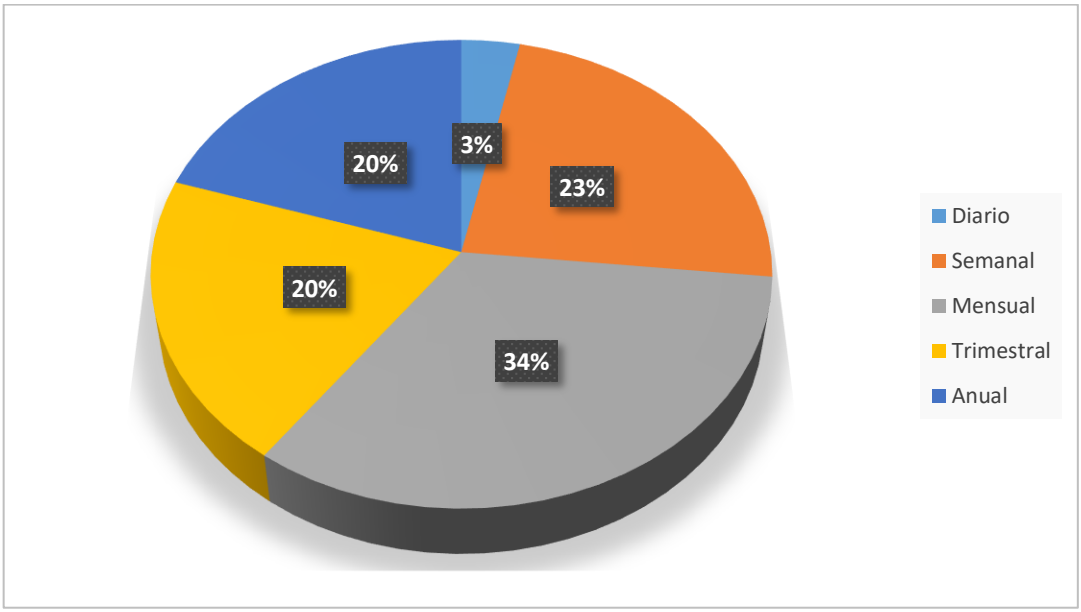


Figura 35. Gráfico de Resultados Encuesta - pregunta 5

Análisis

Los resultados evidencian el tiempo en el que se generan dudas, reclamos por parte de los usuarios los datos están muy diversos debido a que los procesos que se realizan tienen cierto grado de dificultad por tipo de información que se emplea y en diferentes tiempos ya preestablecidos. Donde más se generan preguntas y reclamos son cada mes de acuerdo con los datos que se obtuvieron siendo 34%, pero no se debe tomar solo este dato si no también se muestra que diario se genera un 3% en cuanto a reclamos o dudas y de forma trimestral un porcentaje del 20% donde los usuarios que realizan los procesos de alojamiento y gasto tienen reclamos o dudas, se debe tomar en cuenta cada uno de los usuarios para darle solución a todas estas acciones que se pueden generar.

6. **¿En el momento de realizar el proceso de alojamiento y gasto turístico ha tenido inconvenientes?**

Tabla 13. Resultados de Encuesta - pregunta 6

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Si	14	0,467	46,7%
No	16	0,533	53,3%

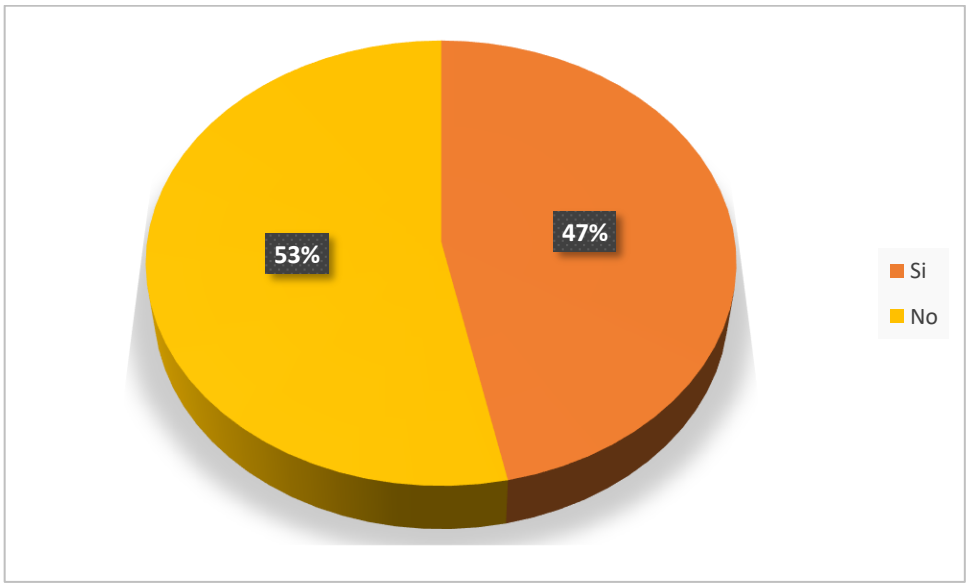


Figura 36. Gráfico de Resultados Encuesta - pregunta 6

Análisis

Tomando en cuenta de cómo se realizan los procesos de alojamiento y gasto turístico con el fin de determinar el nivel de turismo que ha tenido la provincia los resultados muestran que el 53% de los usuarios no tienen ningún tipo de problema para ejecutar este proceso, pero esto significa que todo está claro, y así lo demuestran los datos, ya que con un porcentaje de 47% dice *Si*, que se puede interpretar que todos ellos han tenido ciertos inconvenientes con este proceso, y el mayor grado de dificultad es la cantidad y tipo de información que se emplea en este proceso. En general se puede explicar que en este tipo de procesos siempre se ha de tener alguna dificultad o inconveniente para elaborarlo, y esto se puede evidenciar en el mínimo porcentaje de diferencia que existe entre tener o no tener inconvenientes siendo este del 3%.

7. **¿Aproximadamente cuánto tiempo emplea en realizar el proceso de alojamiento y gasto turístico?**

Tabla 14. Resultados de Encuesta - pregunta 7

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
De 5 a 10 minutos.	5	0,167	16,7%
De 10 a 20 minutos.	8	0,267	26,7%
De 15 a 20 minutos.	7	0,233	23,3%
Más de 20 minutos.	10	0,333	33,3%

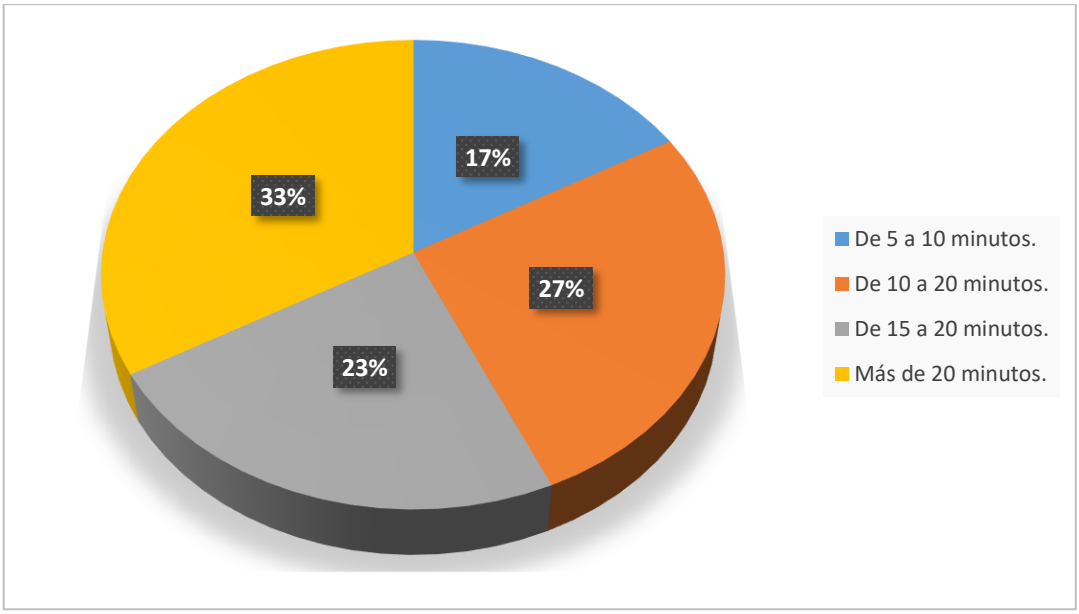


Figura 37. Gráfico de Resultados Encuesta - pregunta 7

Análisis

Los resultados que se obtuvieron nos muestran que para realizar los procesos de alojamiento y gasto turísticos a los usuarios les toma *más de 20 minutos*, corresponde al 33%, lo cual supone que durante los procesos se pueden generar errores, problemas e incluso dudas acerca del proceso, no obstante, 50% de los usuarios que realizan el mismo proceso en menor cantidad de tiempo, por lo que significa existen criterios para realizar un seguimiento acerca de lo que sucede, con el fin de mejorar el tiempos de procesos de alojamiento y gasto sin perder eficacia y eficiencia.

8. **¿Cómo califica el proceso de alojamiento y gasto turístico que actualmente emplea el MINTUR?**

Tabla 15. *Resultados de Encuesta - pregunta 8*

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Muy Bueno	2	0,067	6,7%
Bueno	5	0,167	16,7%
Aceptable	11	0,367	36,7%
Malo	4	0,133	13,3%
Muy Malo	8	0,267	26,7%

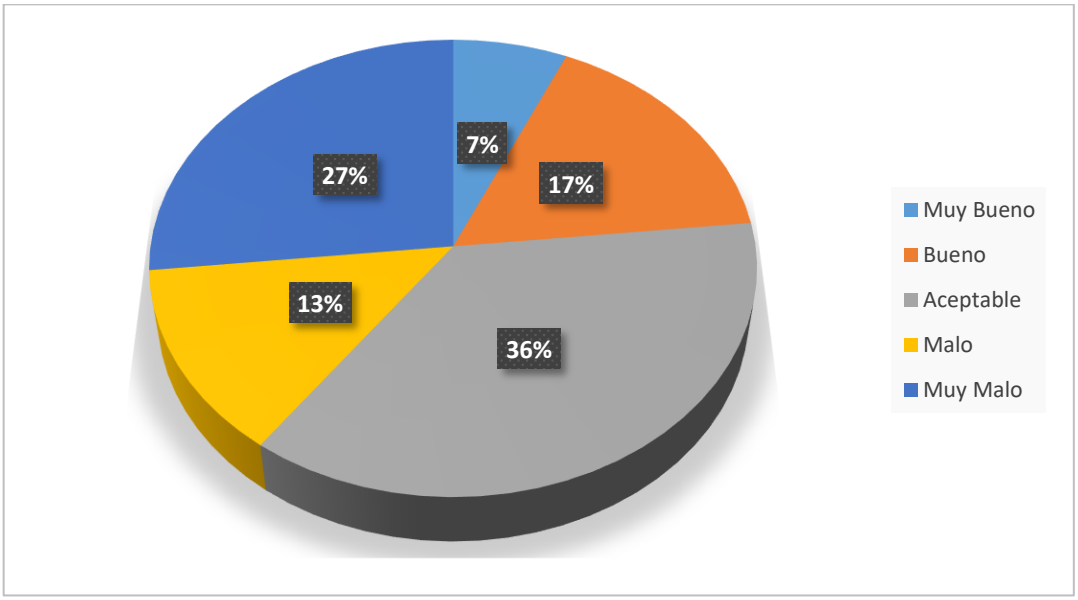


Figura 38. Gráfico de Resultados Encuesta - pregunta 8

Análisis

Actualmente el cómo se realiza el proceso de alojamiento y gasto turístico tiene una calificación aceptable e incluso se pude decir que es bueno de acuerdo con el porcentaje evidenciado por los usuarios siendo de 60,1%, no obstante los usuarios indican que pueden existir falencias dentro de esto proceso debido al porcentaje de 34% donde se puede evidenciar que hay aspectos por corregir con el objetivo de optimizar los procesos de gasto y alojamiento para realizar un análisis adecuado y determinar la demanda de la turística de la provincia con efectividad.

9. ¿Estaría de acuerdo con emplear herramientas tecnológicas con el fin de mejorar los procesos de alojamiento y gasto turístico?

Tabla 16. Resultados de Encuesta - pregunta 9

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Muy de acuerdo	4	0,133	13,3%
De acuerdo	10	0,333	33,3%
Ni acuerdo ni en desacuerdo	10	0,333	33,3%
Desacuerdo	6	0,2	20%

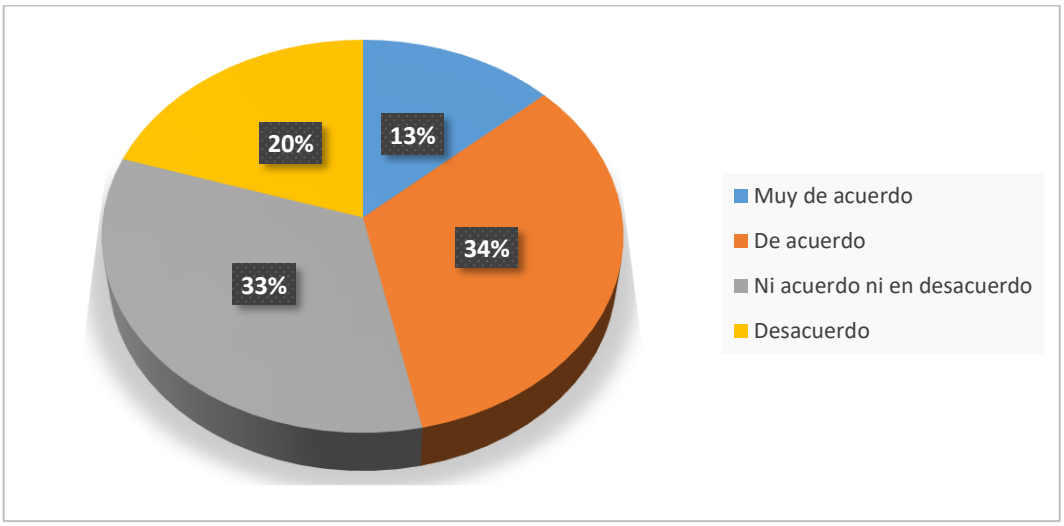


Figura 39. Gráfico de Resultados Encuesta - pregunta 9

Análisis

El uso de nuevas herramientas tecnológicas son cambios que pueden ayudar a mejorar procesos y gestionar mejor la información y según los resultados que se evidencian con relación a los procesos de alojamiento para los usuarios les es indiferente el aplicar nuevos métodos tecnológicos, el 33% de los usuarios consideran este criterio. No obstante, el 34% consideran que aplicar cambios a los procesos con el uso de nuevos métodos podría a ayudar a mejorar la condición actual de la ejecución actual de los procesos. En general tomando en cuenta los datos obtenidos el usuario tiene un criterio tradicional y es evidenciado por 53% de la muestra, es decir que para realizar los procesos de alojamiento y gasto donde a través se determina el nivel de turismo de la provincia no consideran el aplicar nuevos métodos tecnológicos.

10. ¿Cree usted que la información que se recoge del proceso de la alojamiento y gasto turístico debe ser almacenada adecuadamente?

Tabla 17 *Resultados de Encuesta - pregunta 10*

Opciones	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia %
Si	21	0,7	70%
No	9	0,3	30%

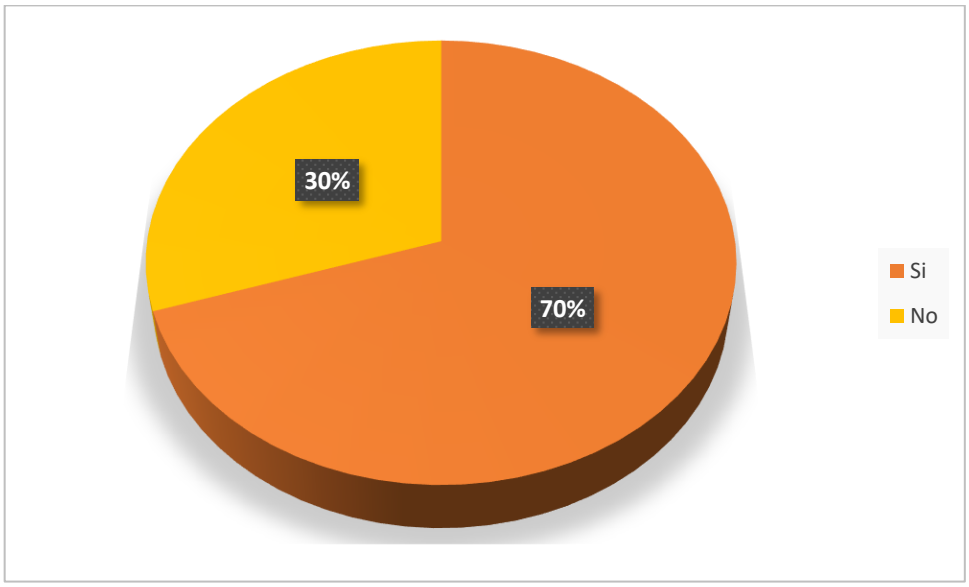


Figura 40. *Gráfico de Resultados Encuesta - pregunta 10*

Análisis

Toda la información que recoge el proceso de alojamiento y gasto turístico los usuarios consideran que es importante de acuerdo con los resultados obtenidos la mayor parte de los encuestados con el 70% indica que se debe almacenar correctamente, puesto esto información que corresponde directamente a la demanda turista que está teniendo y ha tenido la provincia del Carchi e incluso que se puede obtener conocimiento de cómo mejorar la situación turística e incluso perfeccionar los mismos procesos de alojamiento y gasto . Sin embargo, el 30 % de los usuarios afirman que la información que se recolecta y se obtiene no tiene mayor relevancia.

4.1.2. Resultados de la entrevista aplicada

A continuación, se presentan los resultados obtenidos de la entrevista aplicada al Analista de Desarrollo y Promoción Turística de la Zona N°1 del Ministerio de Turismo. La entrevista está estructurada con el propósito de analizar y entender

específicamente los procesos de alojamiento y gasto para la demanda turística de la provincia es decir tiempos, trámites, análisis de información, entre otros. Todo con el fin de servir de apoyo para una construcción adecuada de una propuesta acerca de mejorar los procesos mediante minería de datos, a continuación, se presentan las respuestas y el respectivo análisis de cada una de las preguntas formadas:

1. **¿Cree usted que los procesos que se maneja actualmente, donde se determina el nivel de turismo de la provincia del Carchi son organizados? ¿Por qué?**

" Los procesos para recopilar datos como Ministerio de Turismo se lo realiza a nivel nacional, nosotros si tenemos un organización debido a que se lo ejecuta mediante el segmento que se maneja el cual son los hoteles; para que una persona nosotros poder considerar como turista debemos entender que debe quedarse en sitio de destino, entonces desde este punto de vista las estadísticas se las realiza con los turistas que se quedaron en el lugar, y donde se quedan en los sitios de alojamiento que están registrados. Es por esto los datos para determinar la demanda turística se los levanta solamente en hoteles, porque si nosotros levantamos en algunos atractivos turísticos como por ejemplo el cementerio no tenemos la seguridad de si esa persona es visitante es decir solo viene por tiempo de determinado e incluso de paso o es turista, entonces podemos decir quien se quedó en un hotel, lo podemos considerar un turista y es la razón por la que recogemos información en los distintos tipo de alojamiento que existen en la provincia".

2. **¿Cómo se realiza el seguimiento (usando documentos, sistemas de gestión, ambas herramientas) de los procesos de alojamiento y gasto turístico?**

"No existe ningún tipo de seguimiento para los procesos de alojamiento y gasto de forma local, sin embargo si se realiza un seguimiento a nivel nacional donde se almacena toda la información acerca de este proceso, lo mínimo que se realiza aquí es mediante una encuesta directa a los empresarios de cada uno de los hoteles de la provincia y a través de un formulario o plantilla que envían desde Quito llenan la información y la remiten nuevamente a Quito, y es ahí donde se encargan de realizar el análisis estadístico en cuanto a demanda turística de la provincia.

3. **¿Existe un período específico donde se realizan los procesos de demanda turística de la provincia del Carchi?**

"Los procesos de alojamiento y gasto turístico solamente son aplicados en feriados nacionales, pero considero que deben ser realizados en otros periodos ya sea semanal, mensual, trimestral; sobre todo para tener datos no solo datos de los feriados sino tener información de otras épocas y poder hacer un análisis más específica sobre la demanda turística que ha tenido y puede llegar a tener la provincia".

4. **¿Qué tiempo toma la realización del proceso donde nos indica la demanda turística que ha tenido la provincia?**

"Existe una problemática debido a trámite de recopilación de los datos, por motivos que los empresarios de cada uno de los hoteles no remitan la información solicitada, pero generalmente todo el tema de proceso 1 día si todo se ha enviado correctamente, sino lo máximo ha sido 5 días".

5. **¿Existen estándares o técnicas previamente establecidas en los procesos para determinar la demanda turística de la Provincia?**

"En este momento no se cuenta con ningún estándar o parámetro que sean aplicados directamente a los procesos de alojamiento y gasto turístico".

6. **¿Han existido inconvenientes en la realización del proceso de alojamiento y gasto turístico?**

"Si han existido varios inconvenientes debido a que algunos de los empresarios no entregan la información, por ejemplo, el feriado pasado no hubo la entrega de dos sitios de alojamiento, y lo que sucede es que existen sanciones y multas para lo que no realizan estos procesos pero como Ministerio no queremos llegar a eso; otro problema que se ha generado es que la información que han enviado esta incorrecta y todo esto ocasiona que entorpezca el proceso de análisis acerca de la demanda turista de la provincia".

7. **¿Describa cómo se realiza actualmente los procesos para determinar el turismo que ha tenido la provincia?**

"Pues el primer paso, es construir los formularios para los procesos de alojamiento y gasto turístico por motivos de que algunos establecimientos de alojamiento no manejan correo electrónico, por lo que hay que entregar de manera física para que inicien en este proceso, y para el resto que ya manejan correo electrónico se les envía

por este medio, para poder realizar el proceso de correo electrónico y poder empezar con la recolección de información".

8. **¿Actualmente los procesos para determinar la demanda turística tienen algún costo?**

" Este proceso en el que se determina la demanda turística de la provincia para el Ministerio de Turismo no nos genera ningún costo".

9. **¿Cuántas personas intervienen en el proceso de alojamiento y gasto turístico?**

"Pues estarían los empresarios de cada uno de los hoteles se podría explicar como la primea fase de este proceso, y también depende de la magnitud del hotel, ya que podrían intervenir no solo los empresarios sino también los administradores e incluso la contadora de cada uno de los hoteles, y luego ya en el Ministerio de la provincia el Analista Zonal (yo), donde se aplican los análisis estadísticos correspondientes a los datos que se emitieron por parte de cada uno de los sitios de alojamiento".

10. **¿Cómo ser realiza la interacción con los usuarios que intervienen en este proceso?**

"Emplear el correcto electrónico para poder enviar datos sobre los datos que estamos solicitando para la determinar la demanda turística y también el uso de WhatsApp para cualquier solicitud o pregunta acerca del proceso de alojamiento y gasto turístico".

11. **¿Cómo es el cálculo para el proceso de la demanda turística y que herramienta utiliza para ello?**

"El cálculo son estadísticas básicas para obtener un consumo de alojamiento y gastos medios digamos así, y usamos técnicas tradicionales, es decir calculadora, hoja, papel, lápiz".

12. **¿Durante la ejecución de un proceso tuvo algún inconveniente que interrumpió el avance de actividades y aproximadamente cuánto tiempo fue de retraso?**

"Los problemas más comunes que se han presentado son de tiempos de entrega de los mis empresarios en cuanto a los datos solicitados hacer de como estuvo la demanda turística en el sitio de alojamiento".

13. ¿Actualmente cómo se almacena toda la información con respecto al proceso de demanda turística de la provincia?

" Toda la información acerca de los procesos de alojamiento y gasto turístico de cada uno de los hoteles se los tiene almacenados en físico en expedientes y algunos están en una hoja de cálculo de Excel, y otros datos se encuentran en el repositorio de la página del Ministerio de Turismo del Ecuador".

14. ¿Cuál es el mecanismo de seguridad que utiliza para proteger la información?

"Toda esta información es importante debido a que puede ayudar a mejorar tanto los procesos como la demanda turística de la provincia, pero no se tiene ninguno mecanismo para salvaguardar esta información".

15. ¿Cuál es el método que se emplea para respaldar los datos?

"Pues algunos datos, no todos, se encuentran enviados al Ministerio de Turismo, pero de Quito, no existe digamos un método local para tener respaldos de toda la información de todos los procesos que se maneja".

PROPUESTA

Tomando en cuenta las variables de estudio, los análisis de resultados de la investigación, la primera comunicación que hubo con el responsable de la Dirección de la Zona N°1 del Ministerio de Turismo, la propuesta consiste en crear un modelo de minería de datos para mejorar los procesos donde se determina el nivel de turismo de la provincia del Carchi, mediante el uso de datos históricos de los procesos de alojamiento y gasto turístico, aplicando la metodología CRISP-DM y la aplicación de herramientas y técnicas de minería de datos; obtener una nueva información o conocimiento limpio donde se pueda evaluar, extraer e interpretar para mejorar los procesos y en general la demanda turística de la provincia.

4.1.3. Estudio de Factibilidad

4.1.3.1. Institucionales

Tabla 18. Recursos humanos

Nombres	Actividad para realizar	Cargo
Chugá Burbano Kevin Anderson	Desarrollo de Tesis	Estudiante
MSc. Miranda Realpe Jorge Humberto	Tutor	Docente
Msc. Enríquez Herrera Jhony Vicente	Asesor	Dirección de TIC

4.1.3.2. Materiales

Tabla 19. Recursos Materiales

Cantidad	Descripción	Costo
1	Resma de papel A4	$ 3,50
	Útiles de Oficina	$ 50,00
Total		$ 53,50

4.1.3.3. Económicos

Tabla 20. *Recursos Económicos*

Cantidad	Descripción	Costo
1	Laptop	$ 850,00
175	Movilización (0,30 bus)	$ 52,50
	Imprevistos	$ 150, 00
Total		$ 1052,50

4.1.3.4. Tecnológicos

Tabla 21. Recursos tecnológicos

Cantidad	Descripción	Tipo
1	Laptop Aspire A315-55G Intel(R) Core (TM) i5-8265U CPU @ 1.60GHz 1.80 GHz	Hardware
1	Mouse Logitech G203	Hardware
1	Power BI	Software
1	Knime Analytics	Software
5	Herramientas ofimáticas	Software

4.1.4. Metodología CRISP-DM

En esta sección de la investigación tratamos la parte más práctica mediante la aplicación de las fases que conlleva la metodología, los procesos que va a ser aplicados ayudaran a la extracción, identificación y evaluación de los datos de los procesos que emplea el ministerio de turismo para el análisis de la demanda turística de la provincia. A continuación, se irán enumerando cada una de las fases que contiene esta metodología y aplicando todos los procesos que tiene intervienen en cada fase.

4.1.4.1. Comprensión del Negocio

En primera instancia tenemos la comprensión del negocio es quizá una de las partes más importantes de esta metodología debido a que se deben comprender y plantear nuevos objetivos de la investigación desde un punto vista empresarial e incluso institucional, y con el avance del estudio llegara a convertirlos en objetivos técnico y en un plan de proyecto. Si no se puede entender o comprender ninguno de los objetivos esta metodología nos indica que, por más que se aplique los algoritmos de datos correcto y herramientas sofisticadas nunca podrá obtener ningún resultado fiable, ni algún conocimiento útil que pueda usar

Para lograr conseguir un mejor beneficio y sacar el máximo provecho en cuanto a Minería de Datos, lo que se debe hacer es comprender el problema que se desea resolver, debido a que todo esto nos permitirá recopilar la información, datos válidos y a través de ellos nos permitirá interpretar correctamente los resultados (Gallardo, 2007).

- **Determinar los objetivos del Negocio**

El objetivo al que se quiere alcanzar es que por medio de la minería de datos que se va a aplicar en esta investigación es el mejorar los procesos que emplea en el Ministerio de Turismo tomando como referencia los datos que se generan al realizar estos procesos, que además se los encuentra en formatos tradicionales. El objetivo es encontrar un conocimiento útil donde nos indique las posibles problemáticas implícitas que tiene este proceso y así poder tomar decisiones en cuanto a mejora de la demanda turística de la provincia del Carchi.

<u>Contexto</u>

Tomando en cuenta la situación actual de del negocio de la organización (Zona n°1 Ministerio de Turismo), como inicio de esta investigación se puede explicar que el

ministerio cuenta con un proceso donde se determina la demanda turística de la provincia, donde los datos que se recolecta no están en ningún repositorio digital, sino se los encuentra en archivos físicos almacenado en ficheros, lo registros que tienen son históricos desde el año 2019. Hasta el presente año (2022) dentro de la institución pública no se ha hecho ningún tipo de investigación acerca de los procesos que se manejan, con el fin de realizar una analítica de datos para obtener un conocimiento de cómo está el estado de estos procesos.

Objetivos del Negocio

Como ya se ha descrito con anterioridad el objetivo es mejorar los procesos que realiza el Ministerio de Turismo, mediante la obtención de un conocimiento limpio y útil, y a través de ello tomar decisiones acerca de los mismos procesos y como encontrar un nuevo enfoque para mejorar la demanda turística de la provincia. Se podría encontrar todo tipo de nuevo conocimiento, dependiendo del enfoque y objetivos al que se quiere llegar, pero en esta investigación se han delimitado los siguientes objetivos:

- Analizar los datos de los procesos de la demanda turística que emplea el ministerio de turismo de la provincia.
- Extraer la información sobre los procesos de alojamiento y gasto turístico.
- Interpretar la información de modo que aporte un conocimiento útil donde se pueda obtener un nuevo enfoque acerca de cómo mejorar los proceso para la demanda turística de la provincia.

Esta información puede ser muy útil tanto para el ministerio de turismo como para la demanda turística de la provincia e incluso indirectamente para los sitios de alojamiento que existen en la provincia. Además, esto permitirá detectar problemas que se tenga en cuanto al manejo de los procesos de alojamiento y gasto turístico, incluyendo el estado de estos, y de cierta manera saber y entender la situación actual de los procesos que actualmente están manejando. Todo le permitirá al ministerio de turismo mejorar los procesos para la demanda turística de la provincia del Carchi.

Criterios de éxito de negocio

Desde la perspectiva de un negocio el criterio de éxito que se pretende es tener la posibilidad de obtener un conocimiento que se encuentre implícito en los datos de los procesos que maneja el ministerio de turismo para los procesos de la demanda turística, de forma que puedan dar consejos de cómo mejorarlos tanto en calidad

como en eficiencia. Otro criterio de éxito es ayudar a la demanda turística de la provincia, e incluso a los sitios de alojamiento que existen.

- **Evaluación de la situación actual**

En este momento se cuenta con los datos en físico de los procesos que se han registrado para determinar la demanda turística de la provincia con información detallada del proceso de alojamiento y gasto turístico, tomando en cuenta que los procesos son relativamente nuevos se cuenta con información detallada desde el año 2019 hasta el presente año 2022, de primera intención se puede decir que toda esta data es más que suficiente para poder resolver el problema. La información incluye datos directamente proporcionados por el ministerio de turismo, de los lugares de alojamiento que existen en la provincia y los gastos que se han generado, siendo específicos; la cantidad de personas nacionales y extranjeras registradas, pernoctaciones, tarifas, tipos de tarifas categorización de los sitios, entre otros datos que pueden resultar útiles para la aplicación de minería de datos.

Inventario de Recursos

En cuanto a recursos de software podemos disponer de todos los programas de minería de datos que se han mencionado en el capítulo II de la investigación, no obstante, si se encuentra habilitado el aplicativo de RapidMiner donde nos proporcionan hacer tareas de minería de datos sobre la base de datos que teníamos en físico, debido a que ya fue transcrita como una base de datos en Excel. Los recursos de hardware de los que disponemos son un computador portátil con las siguientes características técnicas:

- Marca: Acer
- Modelo: Aspire 3 A315-55G-51FB
- Procesador: Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz
- Memoria RAM: 8,00 GB (7,85 GB usable)
- Capacidad de almacenamiento: 128GB PCIe SSD + 1000GB HDD
- Tarjeta gráfica: NVIDIA GeForce MX230
- Sistema operativo: Windows 11 Home Single Language

Requisitos, supuestos, restricciones

Al tener un convenio entre la Universidad Politécnica Estatal del Carchi y el Ministerio de Turismo de la Zona N°1, además desde la primera comunicación directa con el ministerio se acordó que toda la información que se obtenga y se requiera será exclusivamente con propósitos investigativos.

Terminología

Ver Anexo 5: Glosario de términos acerca de minería de datos

Costes y Beneficios

Los costes de este proyecto no suponen ningún coste adicional a la universidad debido a que todo el estudio dedicado a este tema le pertenece a la propia institución de acuerdo con las normas establecidas.

Con respecto a los beneficios, el estudio no genera beneficios de tipo económico directamente a la universidad, pero si supone beneficios para el ministerio de turismo de la zona n°1 de la provincia y aunque no sean económicos, si son de mejora en cuanto a los procesos que se emplean en indirectamente ayudando a la demanda turística de la provincia del Carchi.

- **Determinar los objetivos de la Minería de Datos**

Los objetivos de minería de datos son:

- Analizar los datos de los procesos de alojamiento y gasto turístico mediante un algoritmo de agrupamiento.
- Interpretar la información resultante del modelado de agrupamiento con relación a la situación de la demanda turística de la provincia
- Obtener un conocimiento útil con el fin de generar recomendaciones o consejos de cómo mejorar los procesos para la demanda turística de la provincia.

Criterios de éxito de minería de datos

El criterio de éxito desde la perspectiva de minería de datos en este estudio es lograr encontrar un tipo de conocimiento que se encuentre implícito en la data que dejan los procesos de alojamiento y gasto turístico. Mediante el uso del algoritmo específico elegido y las técnicas adecuadas de minería de datos evaluar e interpretar la información resultante, para entender el estado de los procesos y a partir de ahí, formular consejos o recomendaciones que ayuden a los procesos para la demanda turística de la provincia, e incluso tener la posibilidad de ayudar a los establecimientos de alojamiento que existen en la provincia.

- **Plan de Proyecto**

Con el objetivo de cumplir un cronograma y que sirvan de apoyo a la organización de la investigación se ha estimado tiempos para la ejecución de tareas con relevancia practica:

- Primera Fase: Análisis del tipo de datos que se encuentran almacenados en ficheros dentro del Ministerio de Turismo: 1 semana.
- Segunda Fase: Transformar los datos en datos alojados en documentos físicos a datos digitales: 2 semanas.
- Tercera Fase: Preparación de los datos selección, limpieza, transformación y formateo si fuese necesario: 2 semanas.
- Cuarta Fase: Selección de técnicas de minería de datos de modelado y ejecución sobre los datos: 1 semana.
- Quinta Fase: Análisis e interpretación de los resultados obtenidos en la fase anterior, se puede repetir esta etapa una o dos veces en función de la anterior: 1 semana.
- Sexta Fase: Realizar informes, gráficos estadísticos, los resultados obtenidos haciendo referencia a los objetivos propuestos en la metodología CRISP-DM y la investigación en general: 1 semana.
- Séptima Fase: Presentación de los resultados finales: 1 semana.

Evaluación inicial de herramientas y técnicas

La herramienta que se va a usar para el modelado de minería de datos en este proyecto es Knime como se describió en el capítulo II de la investigación es un herramienta de código abierto para el análisis de Business Intelligence, Machine Learning y ETL, mediante el proceso de arrastrar y soltar para generar los flujos de trabajo con sus respectivos nodos, posee un interfaz gráfica para crear dichos flujos con facilidad, además de que no se requiere una gran experiencia para hacer un análisis de datos efectivo (Strate Bi Open Bussiness Intelligence, 2022). Algunas de las funciones que tiene Knime son:

- Procesos ETL
- Machine Learning
- Creación de modelos de Deep Learning
- Cálculo de analíticas potentes sobre los datos
- Se puede utilizar de distintos tipos de datos como series temporales, imágenes, textos, entre otros.

Centrándonos en la investigación y el tipo modelado de minería de datos que se pretende usar, Knime permite el uso de varias técnicas para el análisis de los datos y uno de ellos es el algoritmo de segmentación o comúnmente se le conoce algoritmo de clustering, que es un aprendizaje no supervisado, Knime consta de un repositorio

llamado Main-Clustering donde existen varios tipos de nodos a usar para clustering, pero el nodo que interesa es el de K-medias, debido a que es él se va a aplicar en los datos referentes al proceso de alojamiento y gasto turístico de la provincia.

Knime no solo consta de este tipo de aprendizaje o algoritmo si no también consta de un aprendizaje supervisado y al igual que en se le puede aplicar distintos análisis dependiendo a lo que se requiera obtener con la aplicación de minería de datos y los datos que se tiene, algunas de las características y funciones que tiene Knime y algunas de ellas resultan de gran utilidad para el estudio son:

- Preparación de los Datos
- Modificación de Valores
- Modelo de Datos: Arboles de Decisión
- Modelo de Datos: Redes Neuronales
- Modelo de Datos: Análisis de Regresión
  - Regresión Lineal
  - Agrupación (clustering)
- Flujo de Datos para reportes
- Creación de reportes

Algunas de las funciones de Knime se apegan a la metodología con la que se está trabajando en la presenta investigación, y no solo con la metodología, sino también las técnicas y análisis que posee se relacionan con los resultados que se pretende obtener al aplicar minería de datos en los procesos de demanda turística. Con respecto a la lectura de archivos es un software muy adaptable y tiene las funcionalidades precisas que esta investigación requieres debido a que puede leer bases de datos alojados desde archivos CSV hasta un archivo de Excel (xsl).

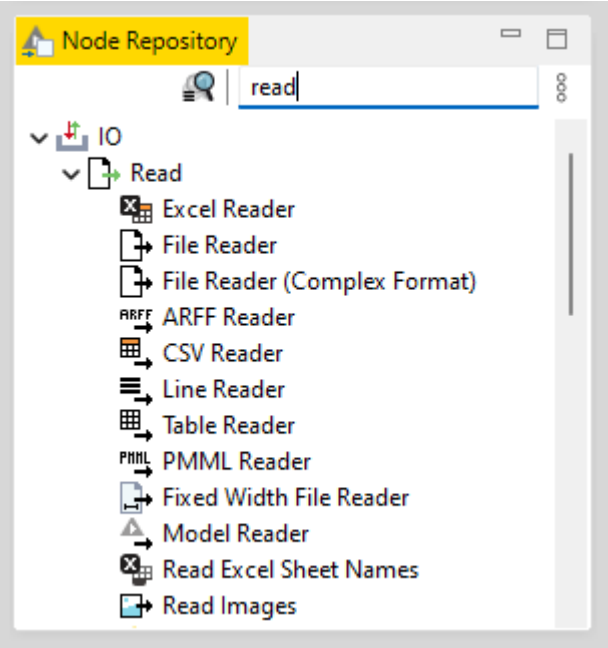


Figura 41. Tipos de lectura de archivos

4.1.4.2. Comprensión de los Datos

La comprensión de datos hace referencia a la segunda fase de la metodología CRISP-DM, donde se desarrolla la recolección inicial de los datos, para empezar a comprender y tener la primera interacción con el problema; además, analizar los tipos de datos que se está manejando en cuanto a calidad, para poder identificar las primeras relaciones más convenientes y ciertas con el fin de iniciar con la formulación de hipótesis entorno a los resultados que requiere el problema.

- **Recolectar los datos iniciales**

Para la recolección de los datos que necesita la investigación, se tomó entorno a la información de los procesos de alojamiento y gasto turístico que son aplicados desde el Ministerio de Turismo con el fin de determinar la demanda turística de la provincia, los datos recolectados constan de características específicas de los hoteles que existen dentro del Carchi, se tiene información de ubicación, categoría, nombre del sitio de alojamiento, entre otros aspectos. Al ser un proceso donde no se aplica ningún tipo de técnica o herramienta tecnológica de innovación para ejecutar estos procesos, se construye la base de información sobre los procesos de alojamiento y gasto turístico, tomando como referencia todos los documentos almacenados en ficheros de forma física, que resumen datos históricos desde el año 2019 hasta el año 2022; se puede decir que, se construyó desde cero una base de datos en una hoja de cálculo que contenga toda la información histórica de estos procesos. Para la construir la base de datos se tomó en cuenta variables específicas de los procesos,

que busquen una relación entre los atributos, esto debido al objetivo que busca este estudio el cual es obtener un conocimiento útil a través de una técnica de modelado de minería de datos, que busque de cierta forma ayudar los procesos de la demanda turística e indirectamente ayudar a un posible aumento de turismo en la provincia, y es importante pues para los ciudadanos de la provincia es un gran aporte económico en sus vidas.

A continuación, se muestran los datos detallados obtenidos que van a resultar de gran utilidad para el uso de las técnicas de minería de datos:

- Categoría

Es un tipo de dato numérico que es establece el Ministerio de Turismo a cada uno de los sitios de alojamiento que tiene la provincia, que directamente tiene relación a las tarifas que establecen los empresarios de cada sitio. La categoría se establece de acuerdo con parámetros entorno a la calidad, comportamiento y todos los permisos que establece el ministerio de turismo.

- Capacidad

Al igual que en el anterior el tipo dato que se emplea es numérico, y refiere al máximo de personas que pueden alojarse en el hotel.

- Tarifas

Es un tipo de dato de numérico con formato de moneda que hace referencia al costo de estancia en el sitio de alojamiento y dependiendo de cada uno de los sitios pueden ser por personas o por habitación.

- Fechas

Son un tipo de dato numérico con formato aaaa/mm/dd, que representa la fecha donde se registró el proceso de alojamiento y gasto turístico, que es realizado por el ministerio de turismo.

- Subtipo del sitio

Hace referencia en el que establece si el sitio de alojamiento es hotel, hostal y único, se puede como carácter en referencia al tipo de dato.

Por otra parte, tenemos también los atributos que van a hacer aplicados mediante el modelado y algoritmo de minería de datos seleccionado:

- Provincia
- Establecimiento
- Subtipo

- Categoría
- Número de Habitaciones
- Número de Plazas
- Alojamiento
    - Entrada de personas nacionales
    - Entrada de personas extranjeras
    - Pernoctaciones
- Gasto
    - Número de habitaciones ocupadas
    - Tarifa promedio
    - Tipo de tarifa
- **Descripción de los Datos**

Los datos con los que se cuenta corresponden a los procesos de gasto y alojamiento turístico, como se menciona anteriormente se encuentran digitalizados y almacenados en una hoja de cálculo de Excel. En la siguiente figura podemos observar la estructura del DataSet de estos procesos, dado que son datos históricos se presentan los datos desde el año 2019 hasta el año 2022.

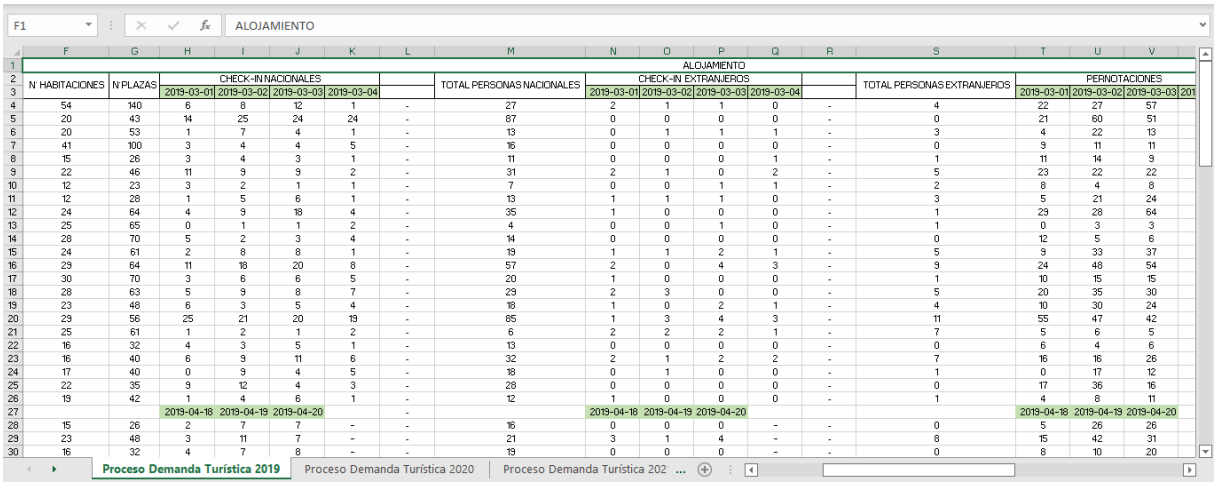
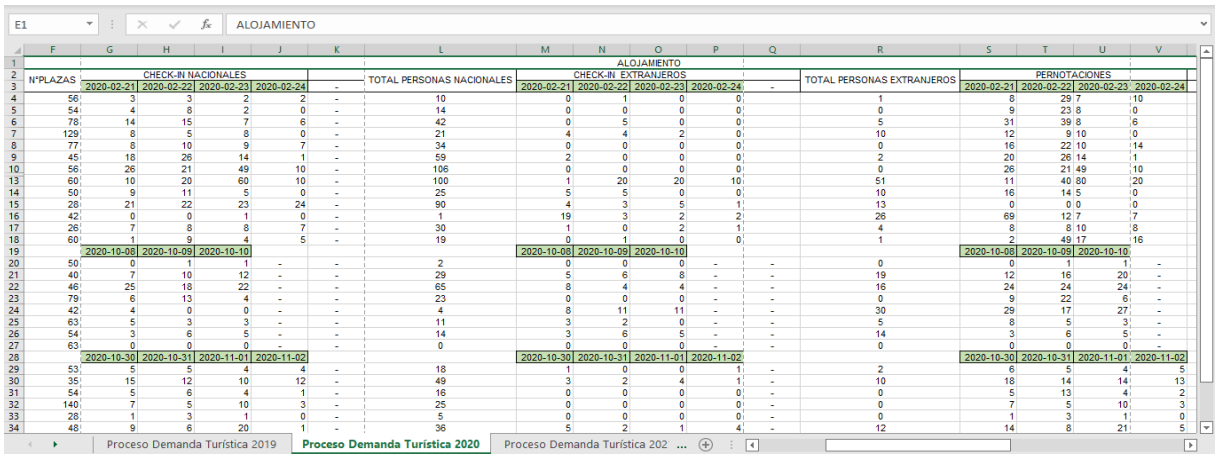


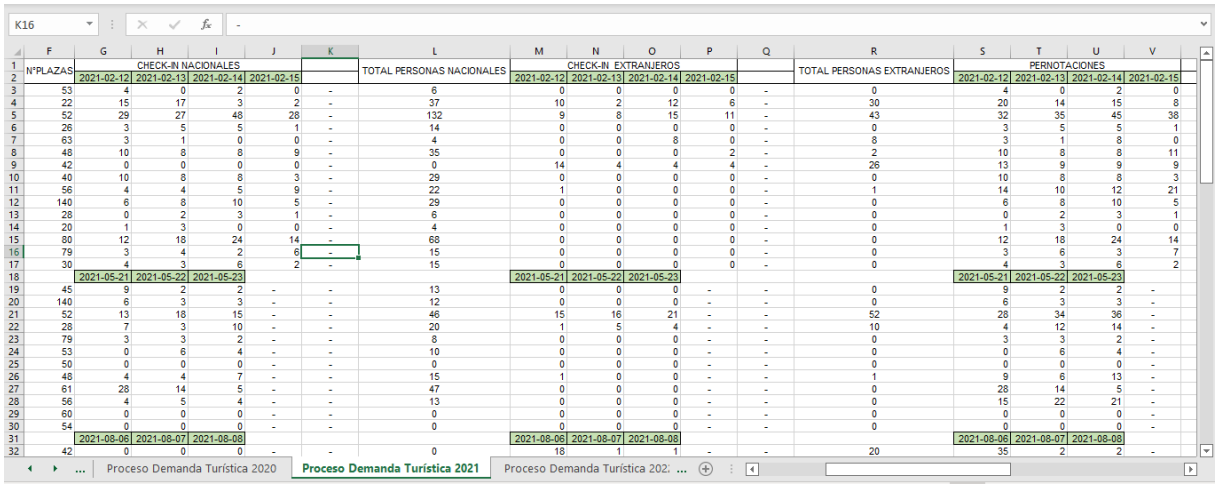
Figura 42. Data de los Procesos de Alojamiento y Gato Turístico en el año 2019



	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1								ALOJAMIENTO									
2	N°PLAZAS	CHECK-IN NACIONALES					TOTAL PERSONAS NACIONALES	CHECK-IN EXTRANJEROS					TOTAL PERSONAS EXTRANJEROS	PERNOTACIONES			
3		2020-02-21	2020-02-22	2020-02-23	2020-02-24	-		2020-02-21	2020-02-22	2020-02-23	2020-02-24	-		2020-02-21	2020-02-22	2020-02-23	2020-02-24
4	56	3	3	2	2	-	10	0	1	0	0		1	8	29	7	10
5	54	4	8	2	0	-	14	0	0	0	0		0	9	23	8	0
6	78	14	15	7	6	-	42	0	5	0	0		5	31	39	8	6
7	129	8	5	8	0	-	21	4	4	2	0		10	12	9	10	0
8	77	8	10	9	7	-	34	0	0	0	0		0	16	22	10	14
9	45	18	26	14	1	-	59	2	0	0	0		2	20	26	14	1
10	56	26	21	49	10	-	106	0	0	0	0		0	26	21	49	10
13	60	10	20	60	10	-	100	1	20	20	10		51	11	40	80	20
14	50	9	11	5	0	-	25	5	5	0	0		10	16	14	5	0
15	28	21	22	23	24	-	90	4	3	5	1		13	0	0	0	0
16	42	0	0	1	0	-	1	19	3	2	2		26	69	12	7	7
17	26	7	8	8	7	-	30	1	0	2	1		4	8	8	10	8
18	60	1	9	4	5	-	19	0	1	0	0		1	2	49	17	16
19		2020-10-08	2020-10-09	2020-10-10				2020-10-08	2020-10-09	2020-10-10				2020-10-08	2020-10-09	2020-10-10	
20	50	0	1	1	-	-	2	0	0	0	-	-	0	0	1	1	-
21	40	7	10	12	-	-	29	5	6	8	-	-	19	12	16	20	-
22	46	25	18	22	-	-	65	8	4	4	-	-	16	24	24	24	-
23	79	6	13	4	-	-	23	0	0	0	-	-	0	9	22	6	-
24	42	4	0	0	-	-	4	8	11	11	-	-	30	29	17	27	-
25	63	5	3	3	-	-	11	3	2	0	-	-	5	8	5	3	-
26	54	3	6	5	-	-	14	3	6	5	-	-	14	3	6	5	-
27	63	0	0	0	-	-	0	0	0	0	-	-	0	0	0	0	-
28		2020-10-30	2020-10-31	2020-11-01	2020-11-02			2020-10-30	2020-10-31	2020-11-01	2020-11-02			2020-10-30	2020-10-31	2020-11-01	2020-11-02
29	53	5	5	4	4	-	18	1	0	0	1	-	2	6	5	4	5
30	35	15	12	10	12	-	49	3	2	4	1	-	10	18	14	14	13
31	54	5	6	4	1	-	16	0	0	0	0	-	0	5	13	4	2
32	140	7	5	10	3	-	25	0	0	0	0	-	0	7	5	10	3
33	28	1	3	1	0	-	5	0	0	0	0	-	0	1	3	1	0
34	48	9	6	20	1	-	36	5	2	1	4	-	12	14	8	21	5

Proceso Demanda Turística 2019 | **Proceso Demanda Turística 2020** | Proceso Demanda Turística 202 ...

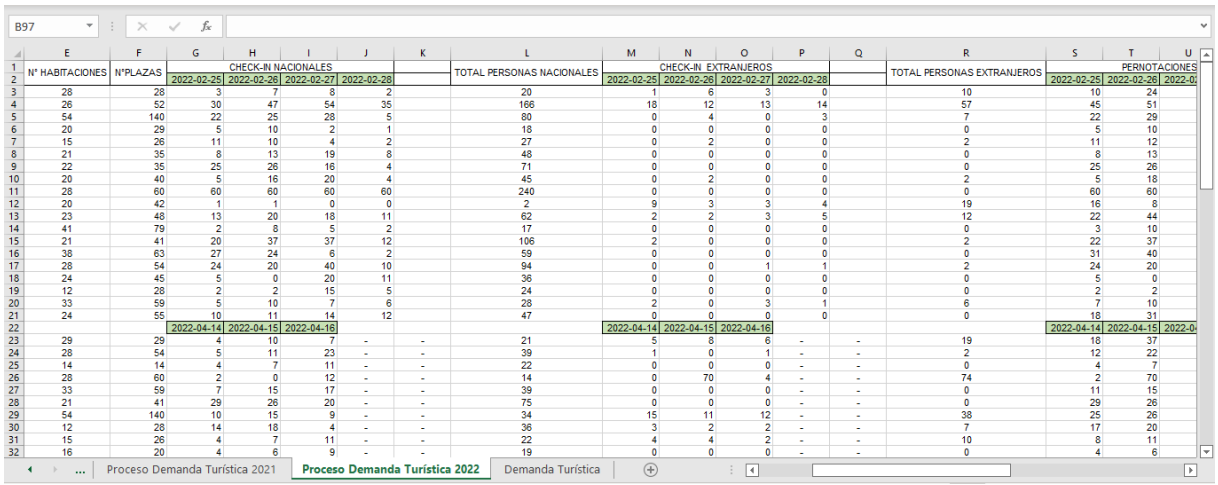
Figura 43. Data de los Procesos de Alojamiento y Gasto Turístico en el año 2020



	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	N°PLAZAS	CHECK-IN NACIONALES					TOTAL PERSONAS NACIONALES	CHECK-IN EXTRANJEROS					TOTAL PERSONAS EXTRANJEROS	PERNOTACIONES			
2		2021-02-12	2021-02-13	2021-02-14	2021-02-15			2021-02-12	2021-02-13	2021-02-14	2021-02-15			2021-02-12	2021-02-13	2021-02-14	2021-02-15
3	53	4	0	2	0	-	6	0	0	0	0	-	0	4	0	2	0
4	22	15	17	3	2	-	37	10	2	12	6	-	30	20	14	15	8
5	52	29	27	48	28	-	132	9	8	15	11	-	43	32	35	45	38
6	26	3	5	5	1	-	14	0	0	0	0	-	0	3	5	5	1
7	63	3	1	0	0	-	4	0	0	8	0	-	8	3	1	8	0
8	48	10	8	8	9	-	35	0	0	0	2	-	2	10	8	8	11
9	42	0	0	0	0	-	0	14	4	4	4	-	26	13	9	9	9
10	40	10	8	8	3	-	29	0	0	0	0	-	0	10	8	8	3
11	56	4	4	5	9	-	22	1	0	0	0	-	1	14	10	12	21
12	140	6	8	10	5	-	29	0	0	0	0	-	0	6	8	10	5
13	28	0	2	3	1	-	6	0	0	0	0	-	0	0	2	3	1
14	20	1	3	0	0	-	4	0	0	0	0	-	0	1	3	0	0
15	80	12	18	24	14	-	68	0	0	0	0	-	0	12	18	24	14
16	79	3	4	2	6	-	15	0	0	0	0	-	0	3	6	3	7
17	30	4	3	6	2	-	15	0	0	0	0	-	0	4	3	6	2
18		2021-05-21	2021-05-22	2021-05-23				2021-05-21	2021-05-22	2021-05-23				2021-05-21	2021-05-22	2021-05-23	
19	45	9	2	2	-	-	13	0	0	0	-	-	0	9	2	2	-
20	140	6	3	3	-	-	12	0	0	0	-	-	0	6	3	3	-
21	52	13	18	15	-	-	46	15	16	21	-	-	52	28	34	36	-
22	28	7	3	10	-	-	20	1	5	4	-	-	10	4	12	14	-
23	79	3	3	2	-	-	8	0	0	0	-	-	0	3	3	2	-
24	53	0	6	4	-	-	10	0	0	0	-	-	0	0	6	4	-
25	50	0	0	0	-	-	0	0	0	0	-	-	0	0	0	0	-
26	48	4	4	7	-	-	15	1	0	0	-	-	1	9	6	13	-
27	61	28	14	5	-	-	47	0	0	0	-	-	0	28	14	5	-
28	56	4	5	4	-	-	13	0	0	0	-	-	0	15	22	21	-
29	60	0	0	0	-	-	0	0	0	0	-	-	0	0	0	0	-
30	54	0	0	0	-	-	0	0	0	0	-	-	0	0	0	0	-
31		2021-08-06	2021-08-07	2021-08-08				2021-08-06	2021-08-07	2021-08-08				2021-08-06	2021-08-07	2021-08-08	
32	42	0	0	0	-	-	0	18	1	1	-	-	20	35	2	2	-

... | Proceso Demanda Turística 2020 | **Proceso Demanda Turística 2021** | Proceso Demanda Turística 202: ...

Figura 44. Data de los Procesos de Alojamiento y Gasto Turístico en el año 2021



	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	N° HABITACIONES	N°PLAZAS	CHECK-IN NACIONALES					TOTAL PERSONAS NACIONALES	CHECK-IN EXTRANJEROS					TOTAL PERSONAS EXTRANJEROS	PERNOTACIONES		
2			2022-02-25	2022-02-26	2022-02-27	2022-02-28			2022-02-25	2022-02-26	2022-02-27	2022-02-28			2022-02-25	2022-02-26	2022-02
3	28	28	3	7	8	2		20	1	6	3	0		10	10	24	
4	26	52	30	47	54	35		166	18	12	13	14		57	45	51	
5	54	140	22	25	28	5		80	0	4	0	3		7	22	29	
6	20	29	5	10	2	1		18	0	0	0	0		0	5	10	
7	15	26	11	10	4	2		27	0	2	0	0		2	11	12	
8	21	35	8	13	19	8		48	0	0	0	0		0	8	13	
9	22	35	25	26	16	4		71	0	0	0	0		0	25	26	
10	20	40	5	16	20	4		45	0	2	0	0		2	5	18	
11	28	60	60	60	60	60		240	0	0	0	0		0	60	60	
12	20	42	1	1	0	0		2	9	3	3	4		19	16	8	
13	23	48	13	20	18	11		62	2	2	3	5		12	22	44	
14	41	79	2	8	5	2		17	0	0	0	0		0	3	10	
15	21	41	20	37	37	12		106	2	0	0	0		2	22	37	
16	38	63	27	24	6	2		59	0	0	0	0		0	31	40	
17	28	54	24	20	40	10		94	0	0	1	1		2	24	20	
18	24	45	5	0	20	11		36	0	0	0	0		0	5	0	
19	12	28	2	2	15	5		24	0	0	0	0		0	2	2	
20	33	59	5	10	7	6		28	2	0	3	1		6	7	10	
21	24	55	10	11	14	12		47	0	0	0	0		0	18	31	
22			2022-04-14	2022-04-15	2022-04-16				2022-04-14	2022-04-15	2022-04-16				2022-04-14	2022-04-15	2022-0
23	29	29	4	10	7	-	-	21	5	8	6	-	-	19	18	37	
24	28	54	5	11	23	-	-	39	1	0	1	-	-	2	12	22	
25	14	14	4	7	11	-	-	22	0	0	0	-	-	0	4	7	
26	28	60	2	0	12	-	-	14	0	70	4	-	-	74	2	70	
27	33	59	7	15	17	-	-	39	0	0	0	-	-	0	11	15	
28	21	41	29	26	20	-	-	75	0	0	0	-	-	0	29	26	
29	54	140	10	15	9	-	-	34	15	11	12	-	-	38	25	26	
30	12	28	14	18	4	-	-	36	3	2	2	-	-	7	17	20	
31	15	26	4	7	11	-	-	22	4	4	2	-	-	10	8	11	
32	16	20	4	6	9	-	-	19	0	0	0	-	-	0	4	6	

... | Proceso Demanda Turística 2021 | **Proceso Demanda Turística 2022** | Demanda Turística

Figura 45. Data de los Procesos de Alojamiento y Gasto Turístico en el año 2022

Se debe tomar en cuenta que las figuras donde muestra la data de los procesos, es solo una muestra de cómo son los tipos de datos y algunas de las variables que tienen

estos procesos y en general como estructurada la base de datos en Excel. Por otra parte, también tenemos las variables con las que se va a trabajar el modelado de minería de datos:

Tabla 22. Variables e Indicadores de los Procesos de Alojamiento y Gasto Turístico

	Variable	Indicador	Tipo de Dato
Alojamiento Turístico	Provincia	Lugar donde se encuentra	Caracter
	Establecimiento	Nombre del sitio	Caracter
	Sub-Tipo	Hotel/Hostal	Caracter
	Categoría	Calificación del hotel en referencia a su calidad	Numérico
	Habitaciones	Habitaciones disponibles	Numérico
	Plazas	Cantidad de plazas disponibles	Numérico
	Check-in Nacionales	Personas nacionales registradas	Numérico
	Check-in Extranjeros	Personas extranjeras registradas	Numérico
	Pernoctaciones	Personas que pasaron la noche	Numérico
	Habitaciones Ocupadas	Número de habitaciones que se ocuparon	numérico
Gasto Turístico	Fecha	Registro del proceso de alojamiento y gasto.	Fecha (aaaa/mm/dd)
	Tarifa promedio	Costo del sitio de alojamiento	Moneda
	Tipo de tarifa	Por persona / por habitación	Caracter

- **Exploración de los Datos**

Una vez realizado las anteriores fases que corresponden a descripción y recolección de datos iniciales, y con el fin de continuar el estudio se debe realizar una exploración de los datos, para realizar este análisis usaremos la herramienta Knime ya que es un software que proporciona funciones que resultan útiles para esta exploración, en general lo que se va a realizar es un análisis estadístico "básico", que me va a permitir

encontrar relaciones entre las variables, corregir atributos, y verificar que los datos estén completo y sean consistentes, que tengan lógica. Este análisis inicial que se le va a hacer a los datos se los explicará mediante el uso de gráficos de distribución y la descripción de estos.

Ahora tomando los datos que se tiene, para realizar una exploración con más detalle se realizó una modificación en los datos. Tomando en cuenta que los datos son históricos, lo que se hizo es separar los procesos de gasto y alojamiento turístico por años y a su vez fueron estructurados de forma que se unan todos los procesos realizados de ese año en una sola data. Esta tarea se la realizo como ya se mencionó con el fin de encontrar las primeras relaciones entre datos y variables, además verificar la consistencia y completitud de los datos.

Procesos de Alojamiento y Gasto Turístico en el año 2019

La siguiente figura muestra la distribución de datos de turistas nacionales, extranjeros en sitios de alojamiento, nos indica un promedio de la demanda turística que ha tenido la provincia en base al subtipo de alojamiento que ha hecho uso el turista en el año 2019.

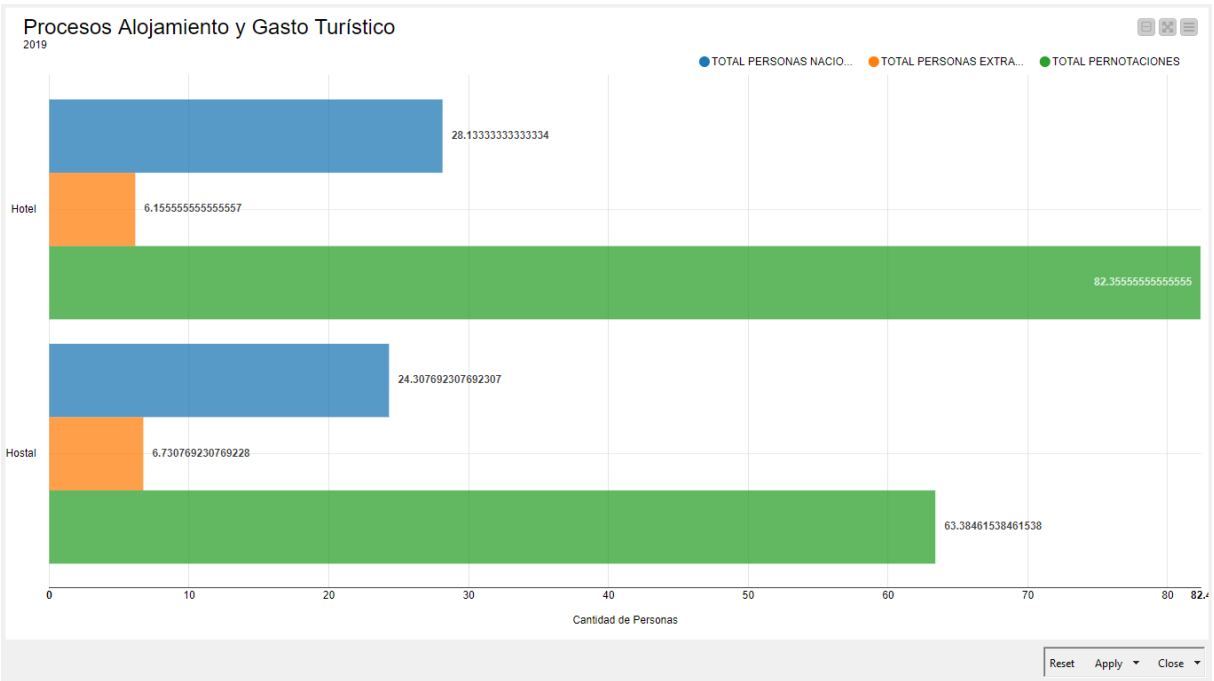


Figura 46. Promedio de alojamiento y gasto turístico en base al subtipo

En la siguiente distribución, se indica la evolución que ha tenido las tarifas de los sitios de alojamiento, tomando como referencia su categoría en el año 2019.

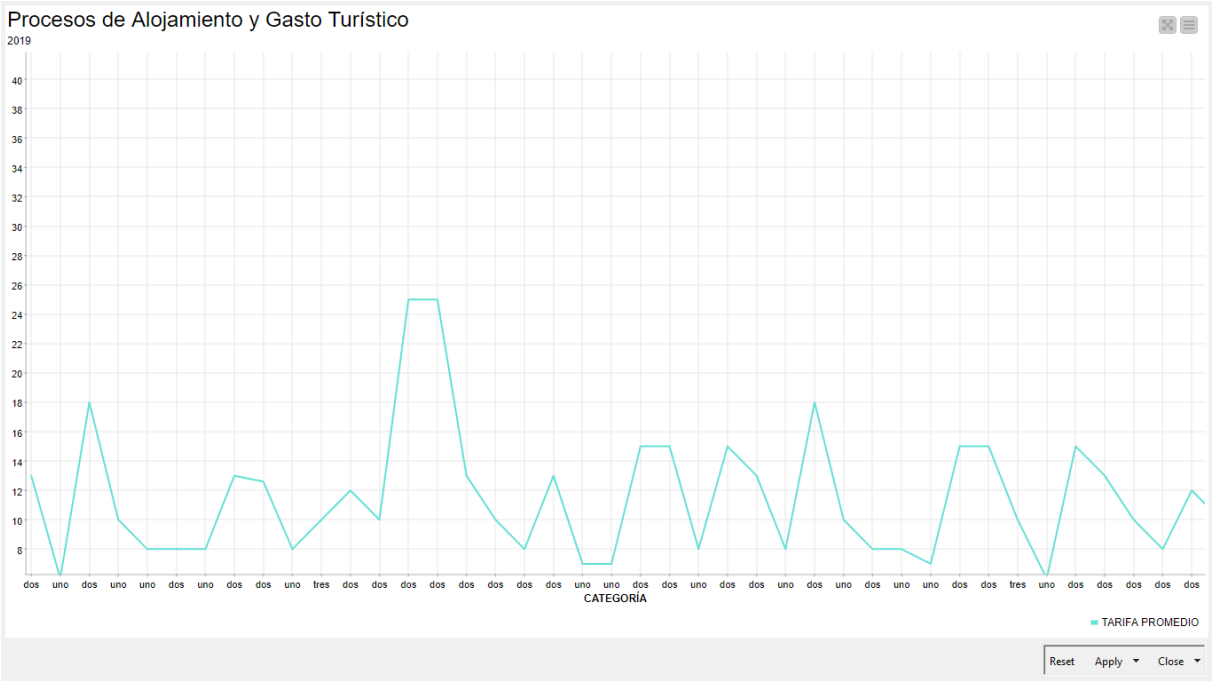


Figura 47. Evolución del costo de los sitios de alojamiento 1

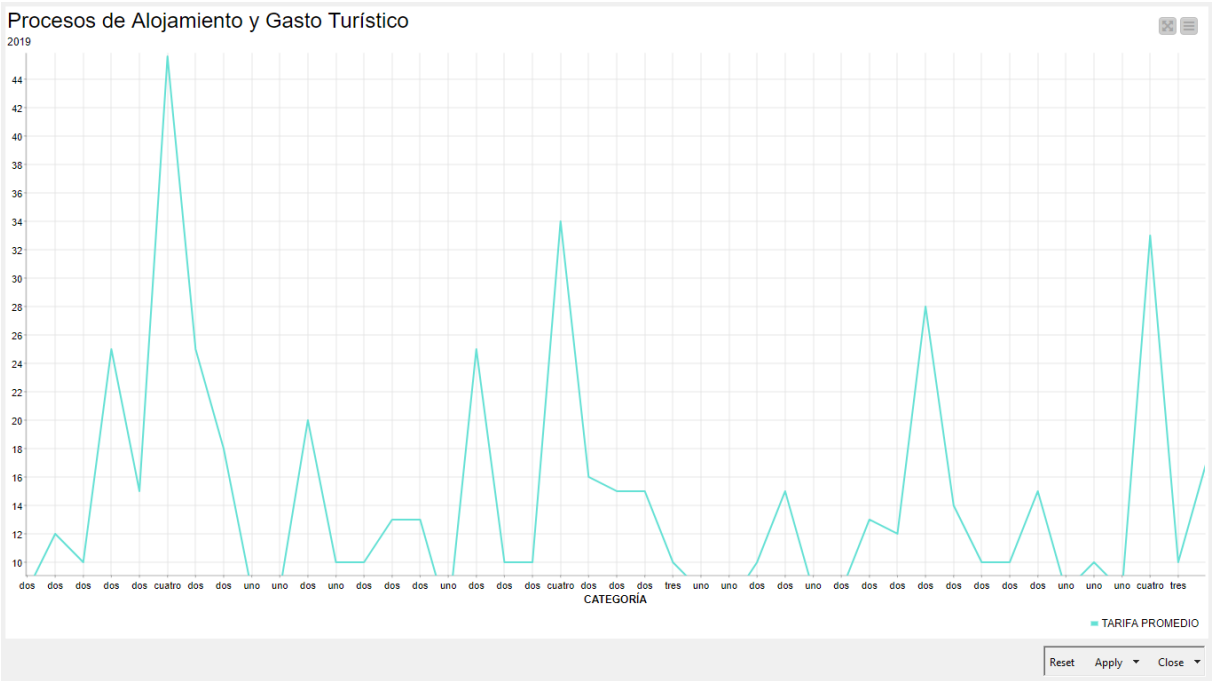


Figura 48. Evolución del costo de los sitios de alojamiento 2

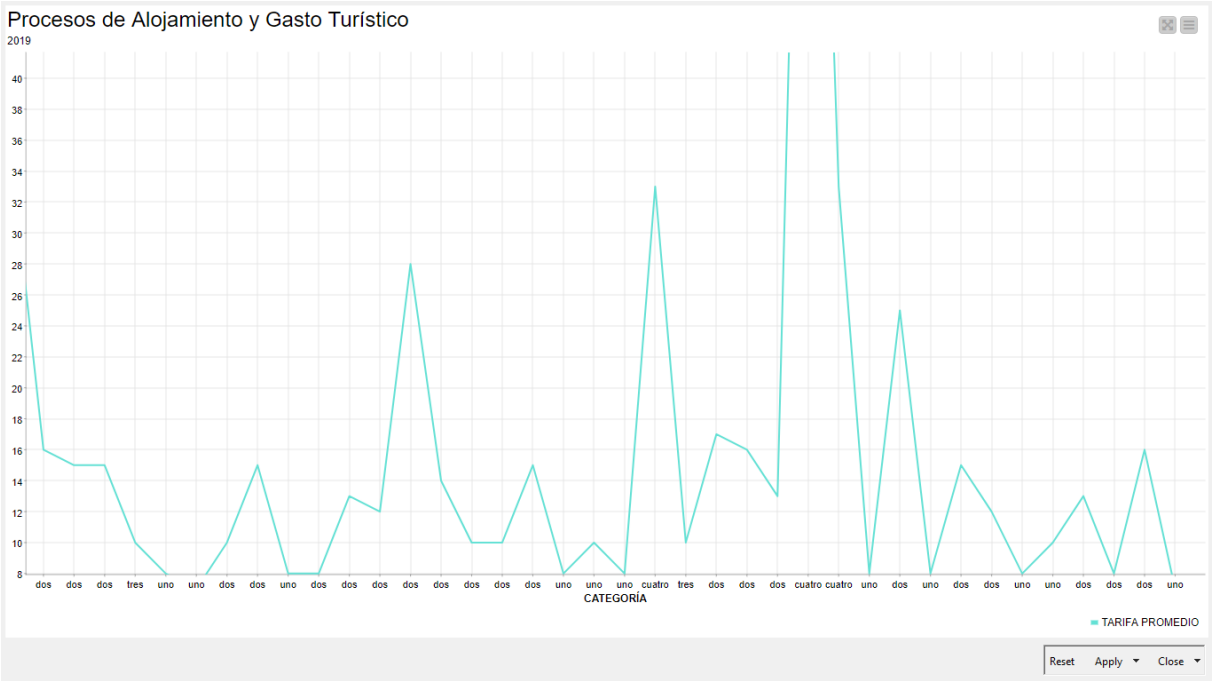


Figura 49. Evolución del costo de los sitios de alojamiento 3

El siguiente gráfico muestra el total de la demanda turística de personas nacionales y extranjeras, haciendo usos de sitios de alojamiento basados en la categoría.

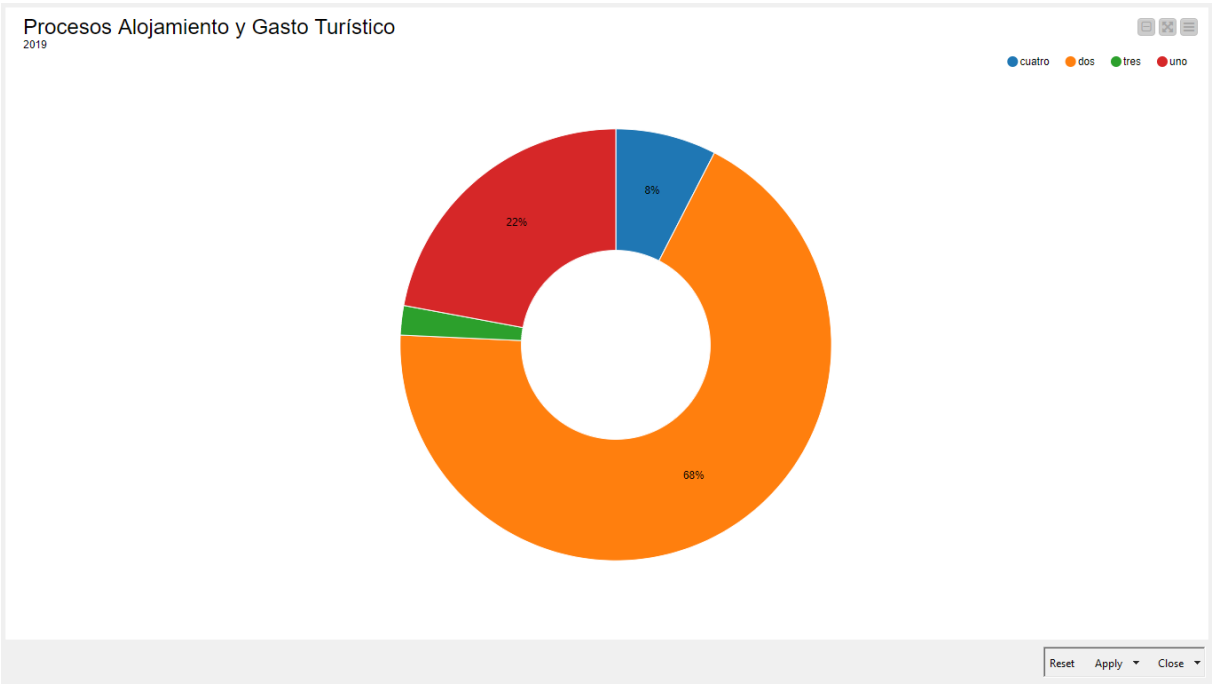


Figura 50. Datos de la demanda turística de sitios de alojamiento en el 2019

A continuación, los siguientes dos graficas de distribución, indican la demanda turística de la provincia con relación a personas nacionales y extranjeras.

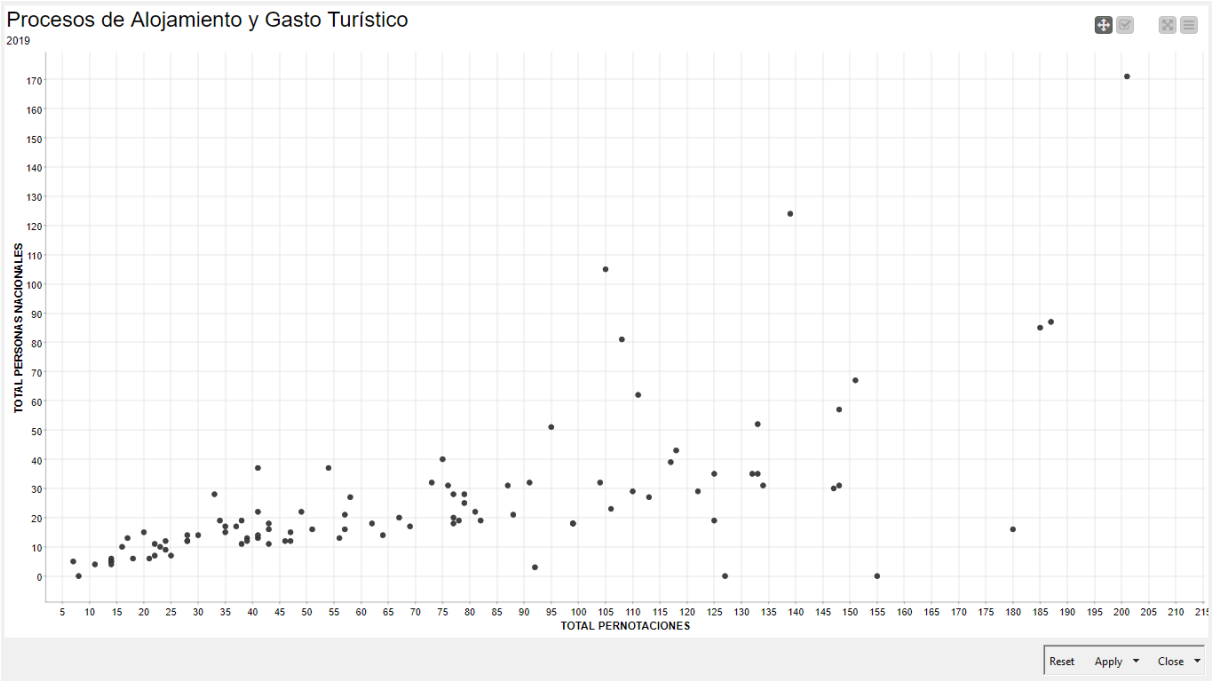


Figura 51. Distribución de la demanda turística de personas nacionales

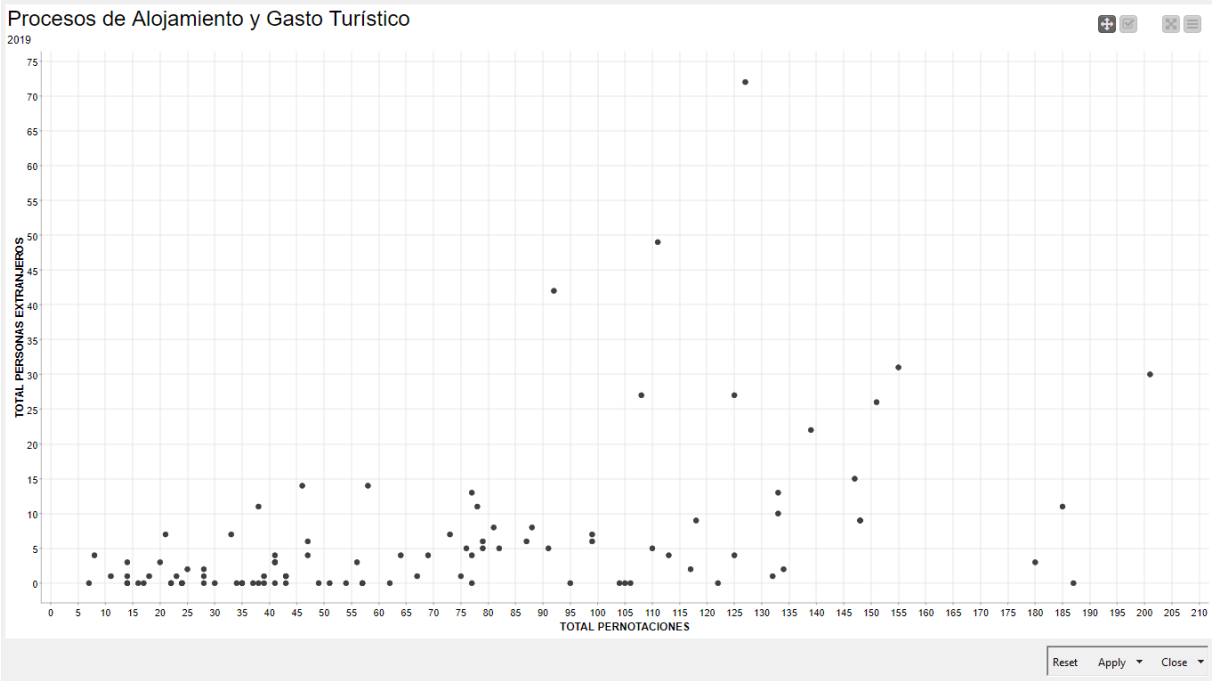


Figura 52. Distribución de la demanda turística de personas extranjeras

Procesos de Alojamiento y Gasto Turístico en el año 2020

La siguiente figura muestra la distribución de datos de turistas nacionales, extranjeros en sitios de alojamiento, nos indica un promedio de la demanda turística que ha tenido la provincia en base al subtipo de alojamiento que ha hecho uso el turista en el año 2020.

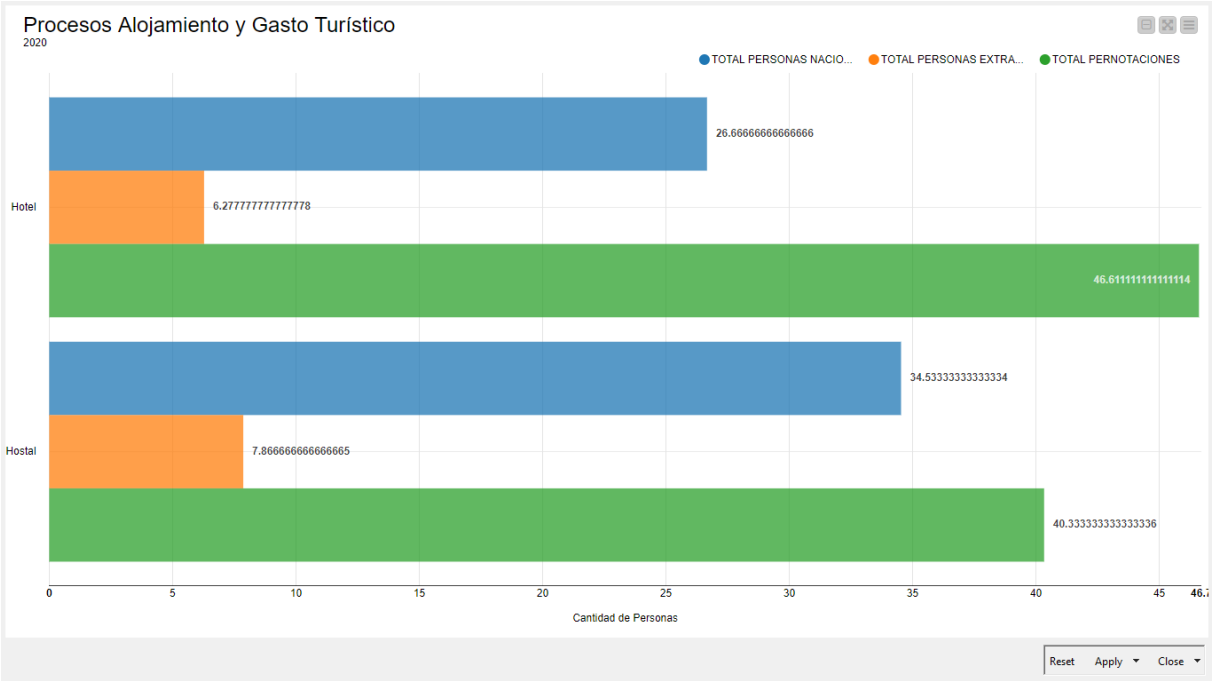


Figura 53. Promedio de la demanda turística en sitios de alojamiento

A continuación, se muestra de la evolución que ha tenido las tarifas de los sitios de alojamiento, tomando en base a la categoría en el año 2020.

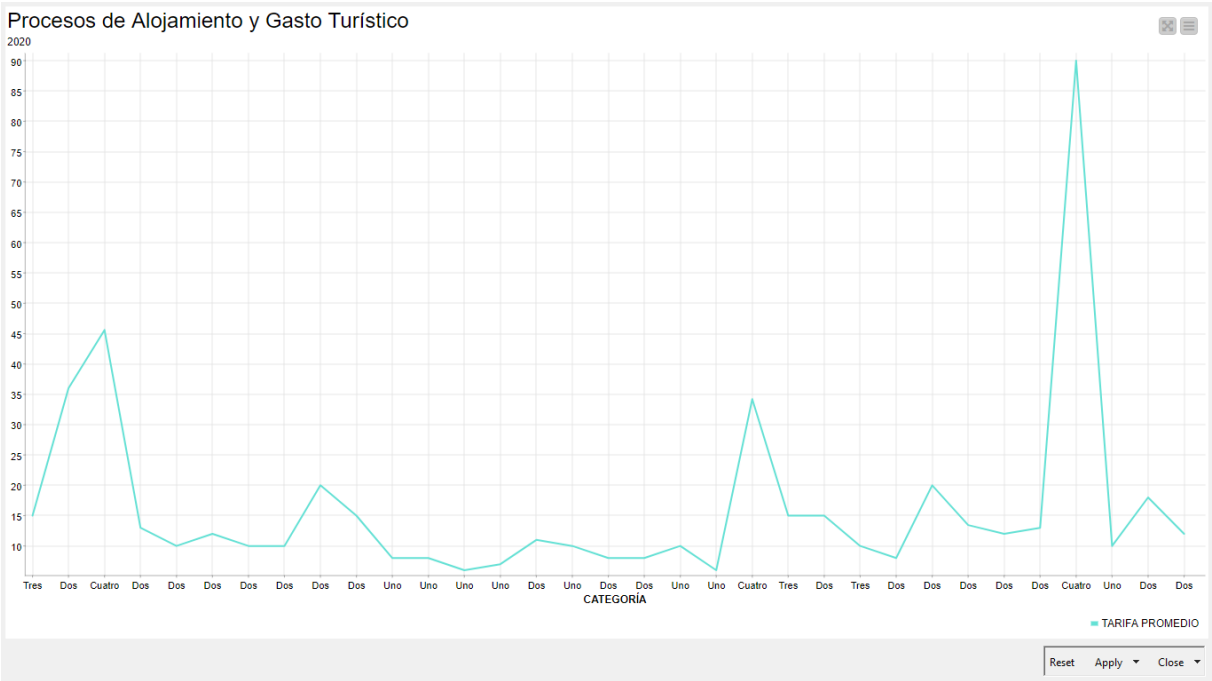


Figura 54. Evolución de las tarifas de los alojamientos turísticos

La siguiente figura nos enseña la demanda turística de personas nacionales y extranjeras, haciendo usos de sitios de alojamiento basados en la categoría en el año 2020.

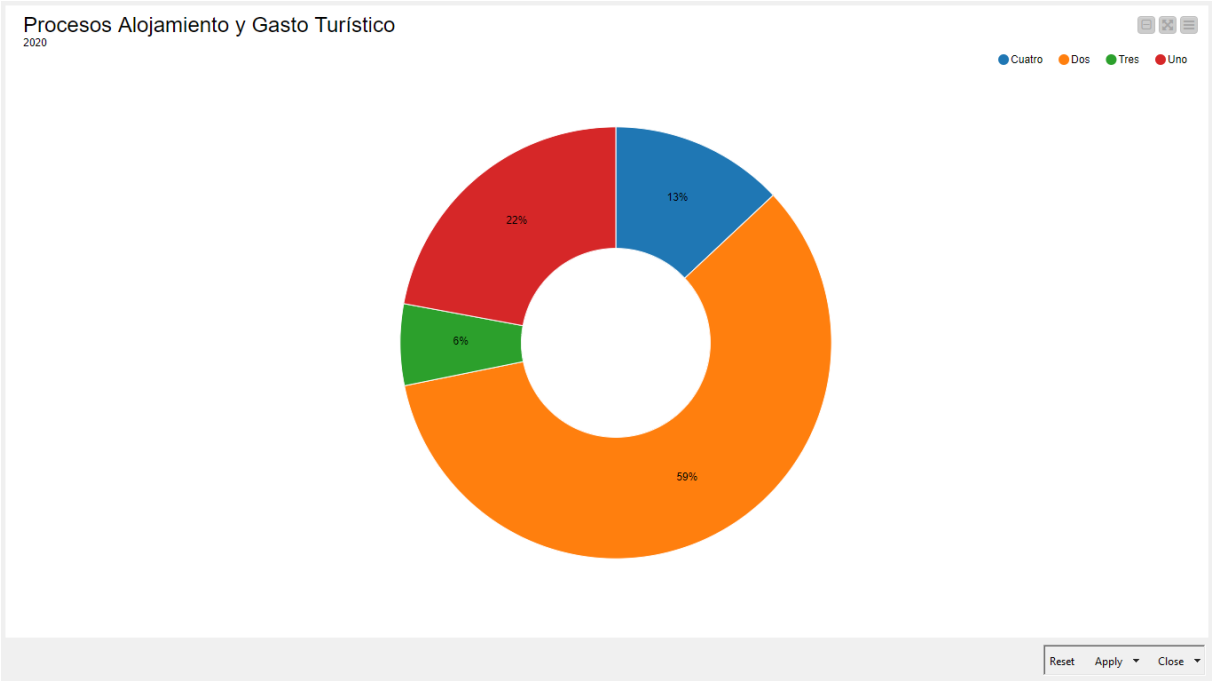


Figura 55. Porcentaje de demanda turística de personas nacionales y extranjeras

Las dos siguientes graficas de distribución, indican la demanda turística de la provincia con relación a personas nacionales y extranjeras en el año 2020.

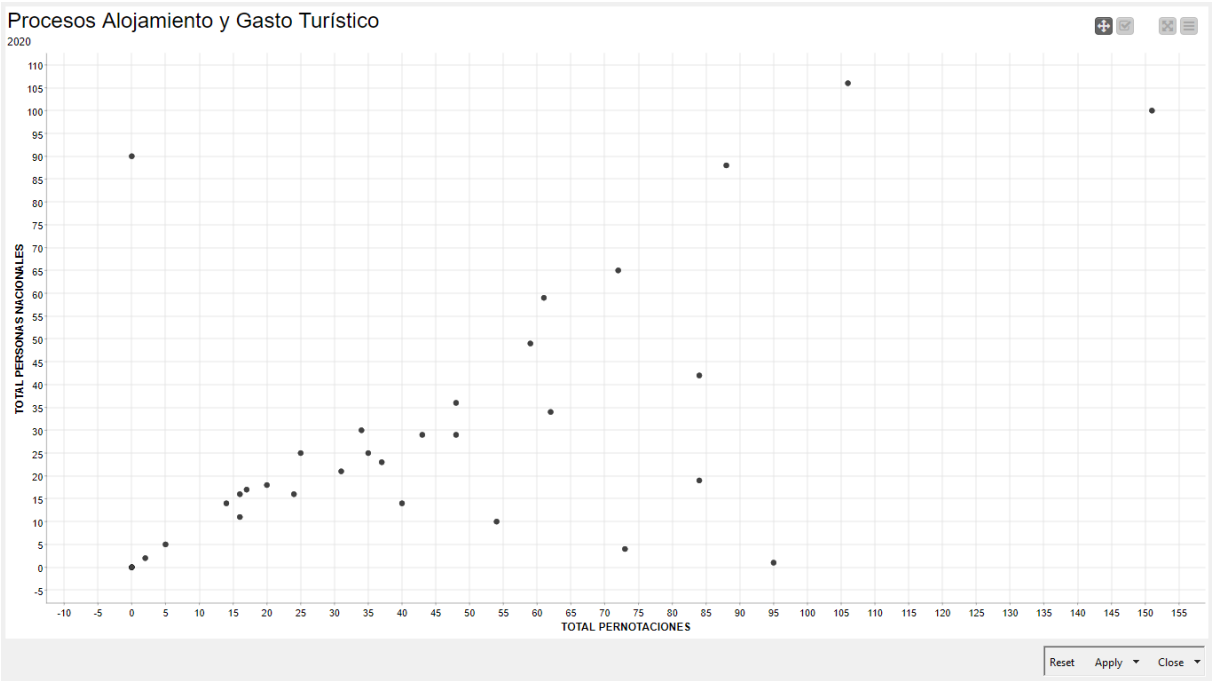


Figura 56. Demanda turística de personas nacionales en el 2020

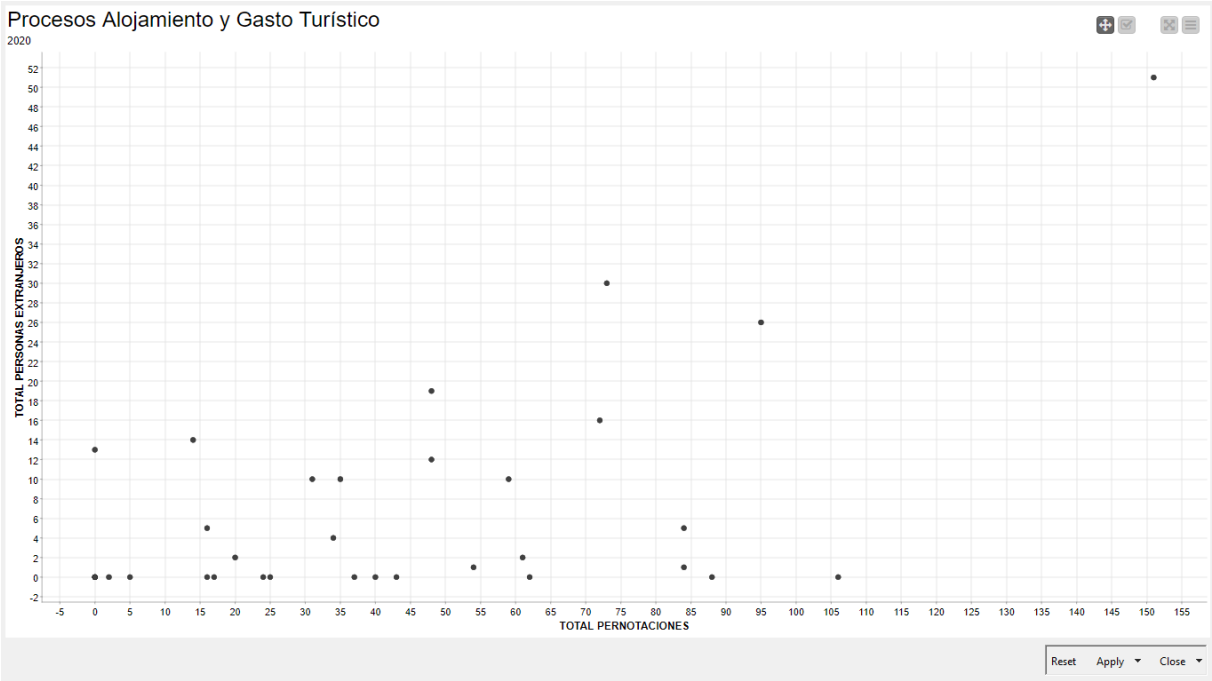


Figura 57. Demanda turística de personas extranjeras en el 2020

Procesos de Alojamiento y Gasto Turístico en el año 2021

El siguiente gráfico muestra una distribución de los datos de turistas nacionales, extranjeros en sitios de alojamiento, es un promedio de la demanda turística que ha tenido la provincia en base al subtipo de alojamiento que ha hecho uso el turista en el año 2021.

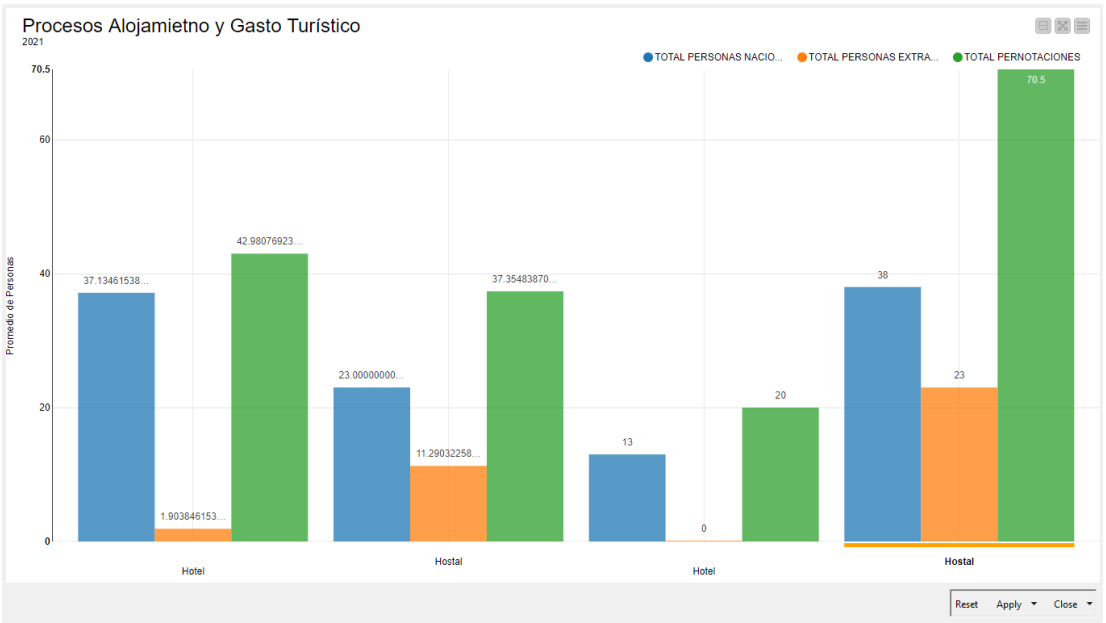


Figura 58. Promedio de demanda turística según el subtipo de alojamiento en el año 2021

El siguiente gráfico indica la evolución que ha tenido las tarifas de los sitios de alojamiento, tomando en cuenta sus categorías en el año 2021.

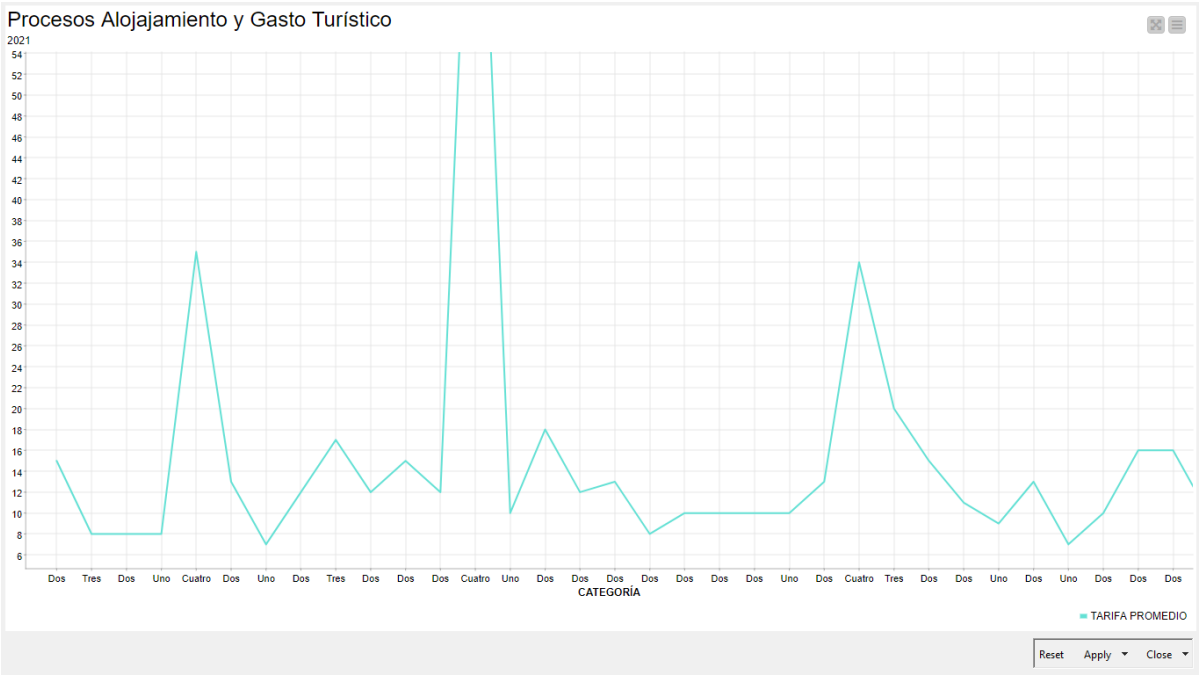


Figura 59. Tipos de tarifa de alojamiento turísticos – 1

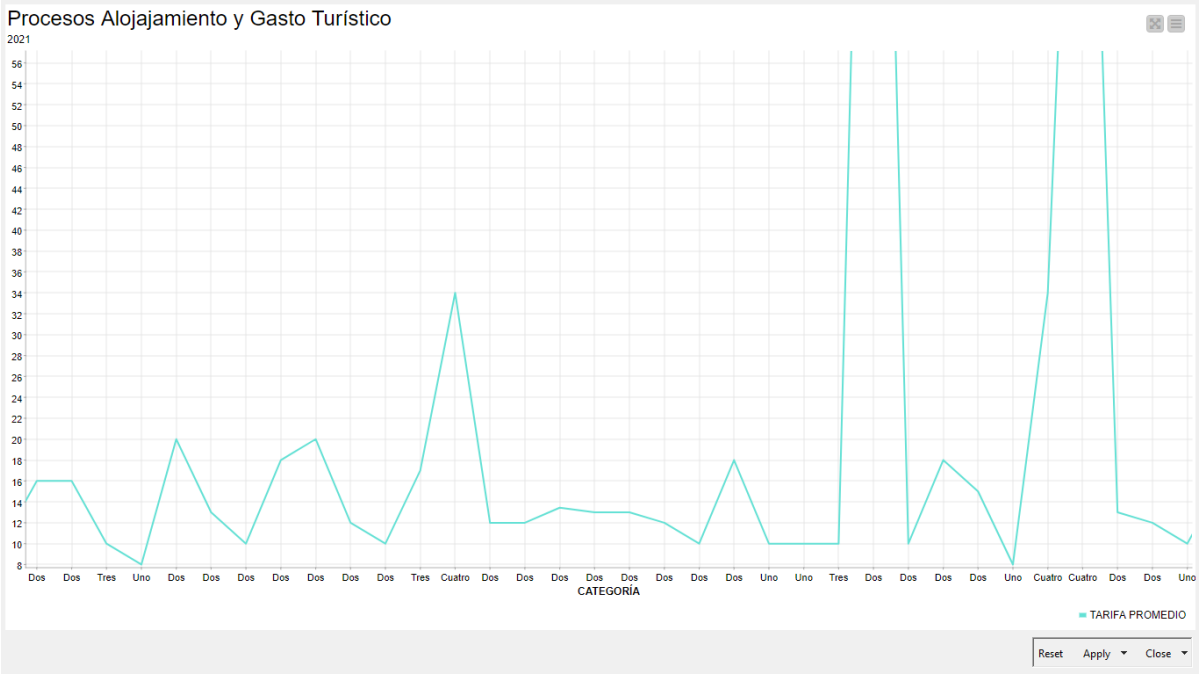


Figura 60. *Tipos de tarifa de alojamiento turísticos – 2*

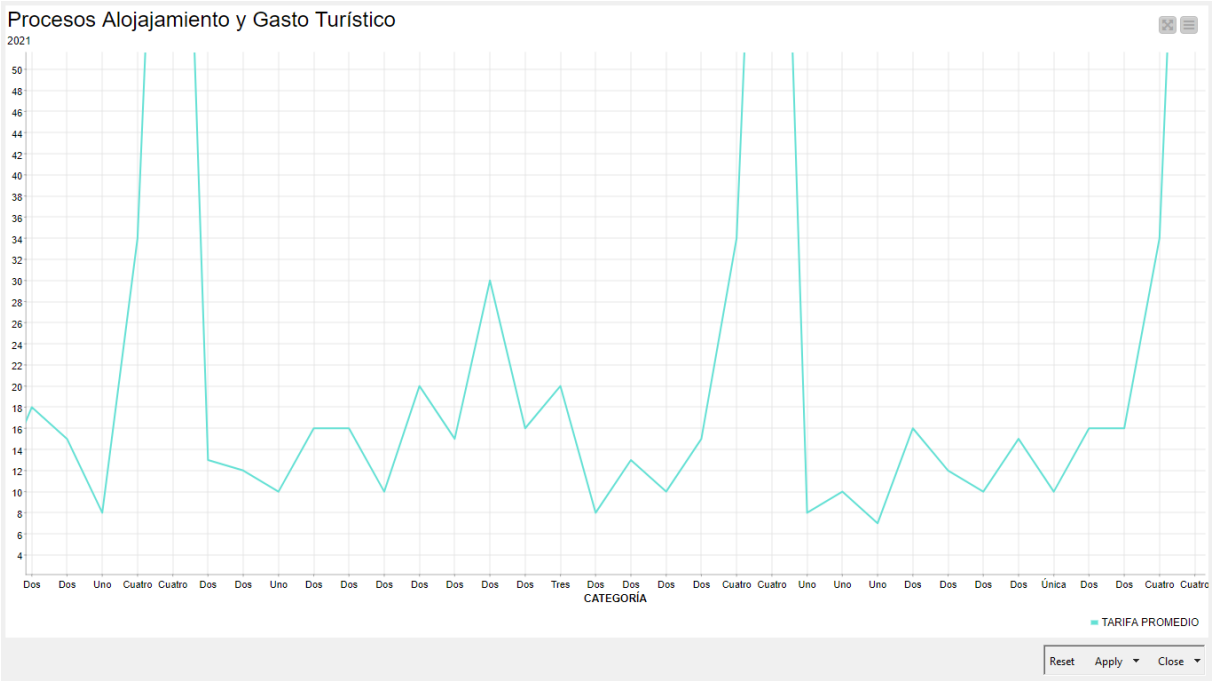


Figura 61. Tipos de tarifa de alojamiento turísticos – 3

El siguiente gráfico nos indica el porcentaje de demanda turística de personas nacionales y extranjeras, haciendo usos de sitios de alojamiento basados en la categoría en el año 2021.

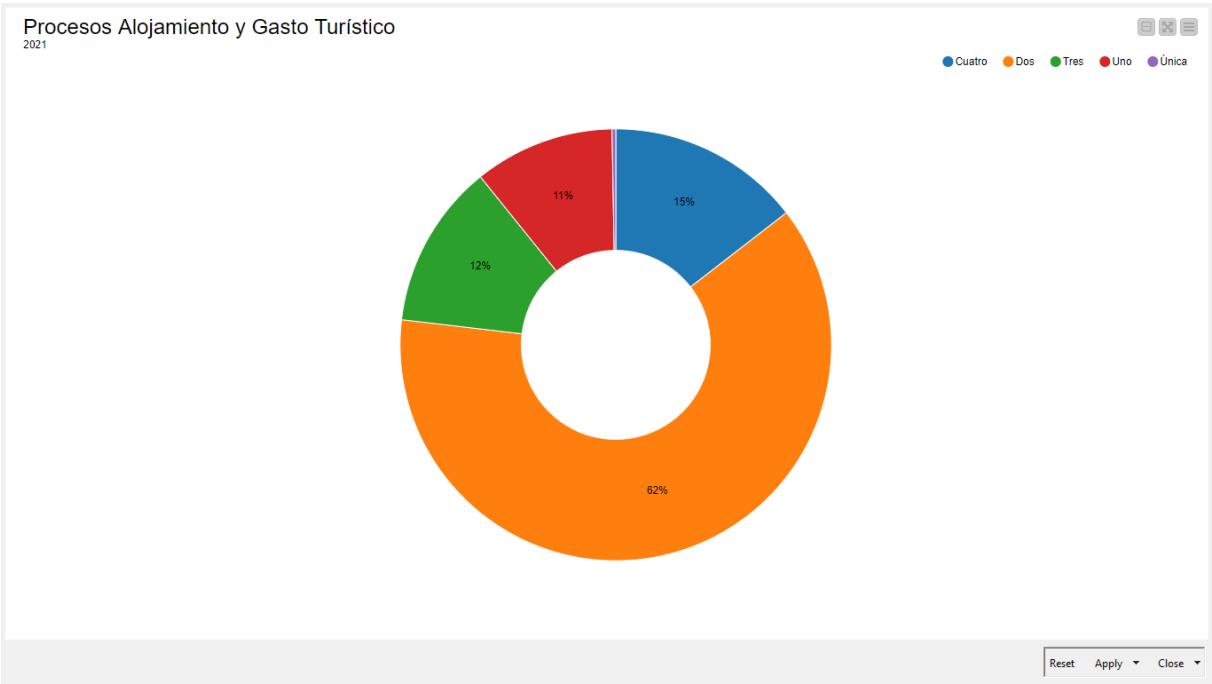


Figura 62. Porcentaje de demanda turística de personas nacionales y extranjeras en el año 2021

Las dos siguientes graficas de distribución, indican la demanda turística de la provincia con relación a personas nacionales y extranjeras en el año 2021.

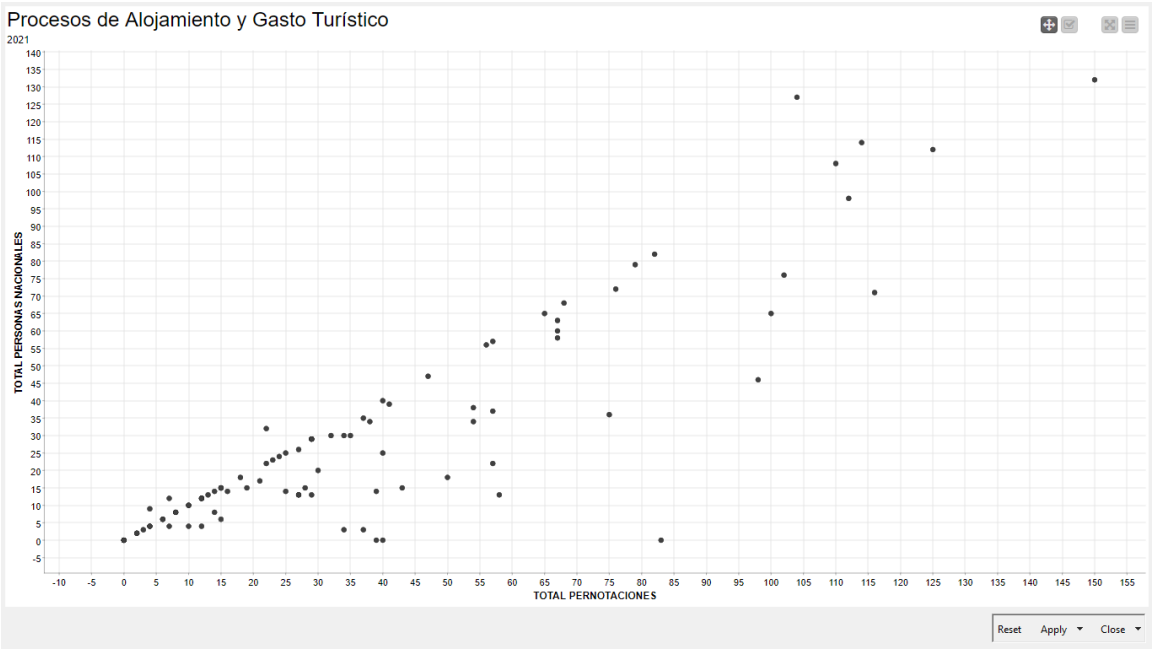


Figura 63. Demanda turística de personas nacionales en el 2021

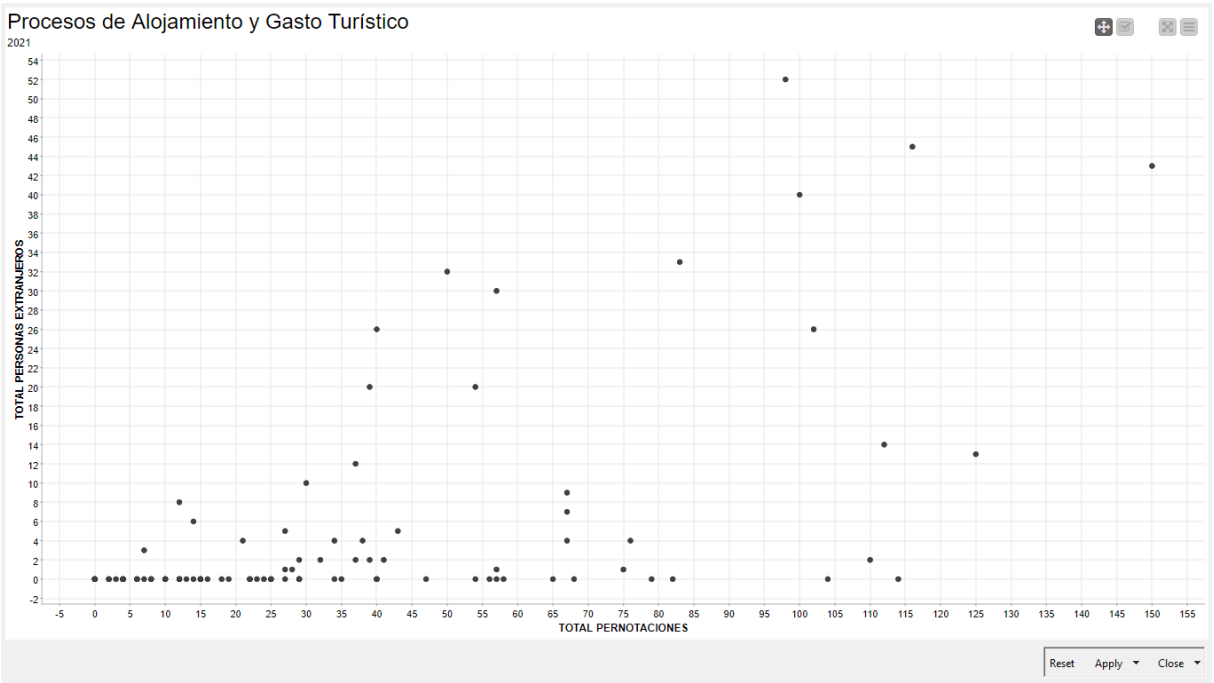


Figura 64. Demanda turística de personas extranjeras en el 2021

Procesos de Alojamiento y Gasto Turístico en el año 2022

En la siguiente distribución son datos de turistas nacionales, extranjeros en sitios de alojamiento, el promedio de la demanda turística que ha tenido la provincia es en base al subtipo de alojamiento que ha hecho uso el turista en el año 2022.

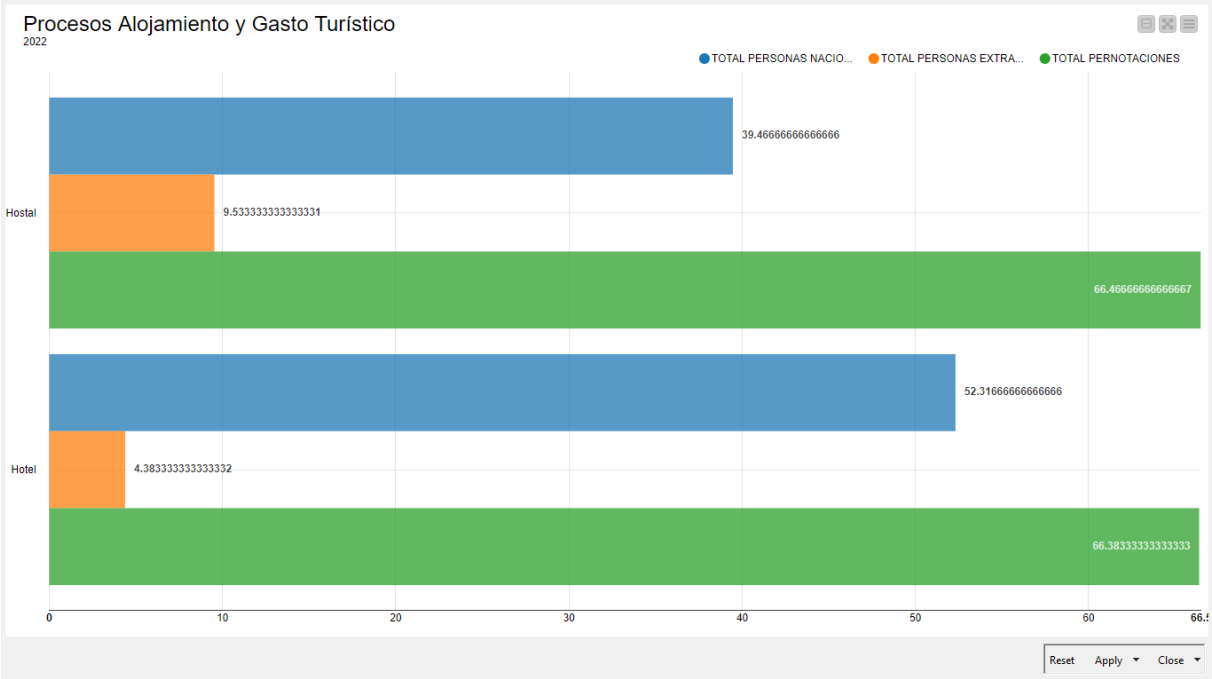


Figura 65. Promedio de la demanda turística en base al subtipo de alojamiento en el 2022

Los siguientes gráficos de distribución nos indican los cambios que ha tenido las tarifas de los sitios de alojamiento, tomando en cuenta sus categorías en el año 2022.

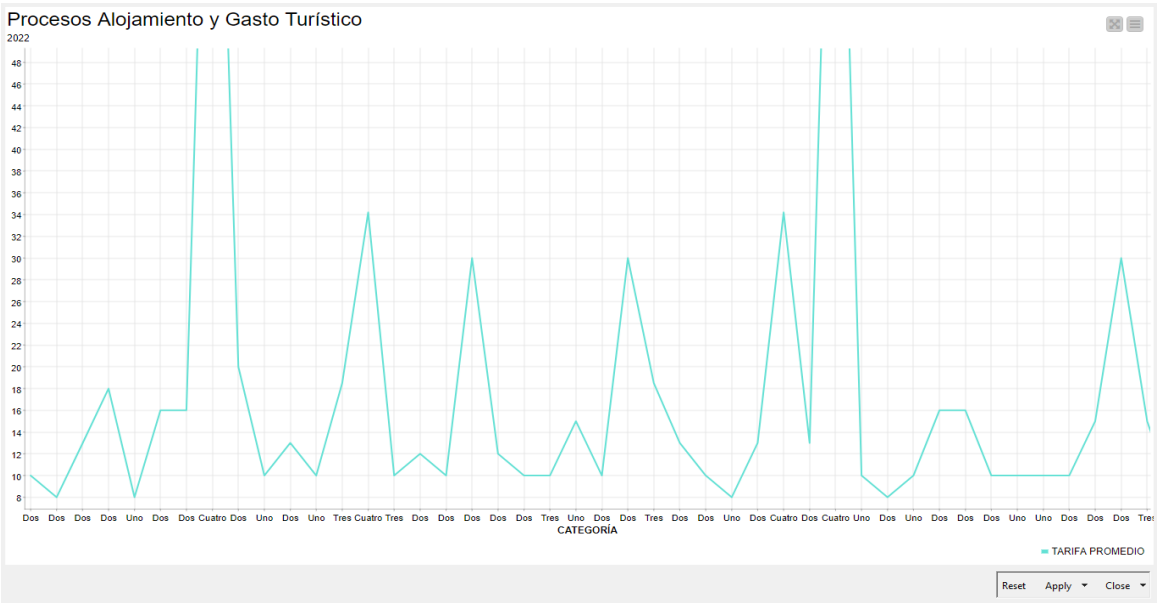


Figura 66. Tarifas promedio de los sitios de alojamiento basados en su categoría en el año 2022

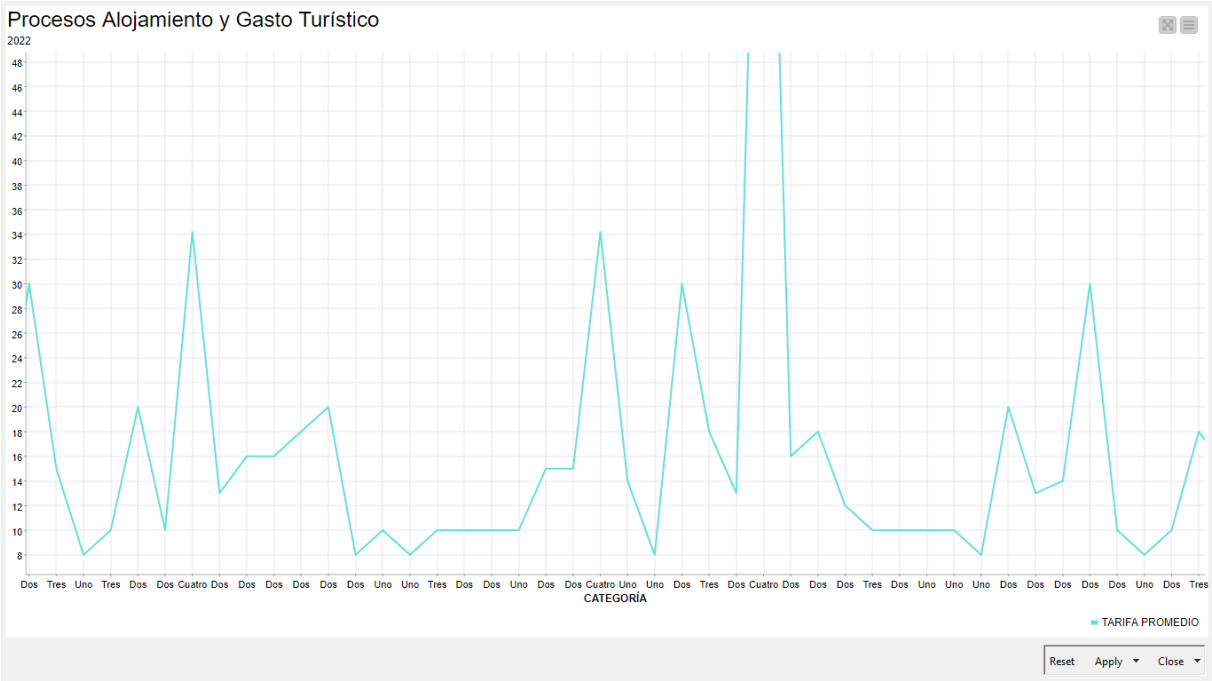


Figura 67. Tarifas promedio de los sitios de alojamiento basados en su categoría en el año 2022

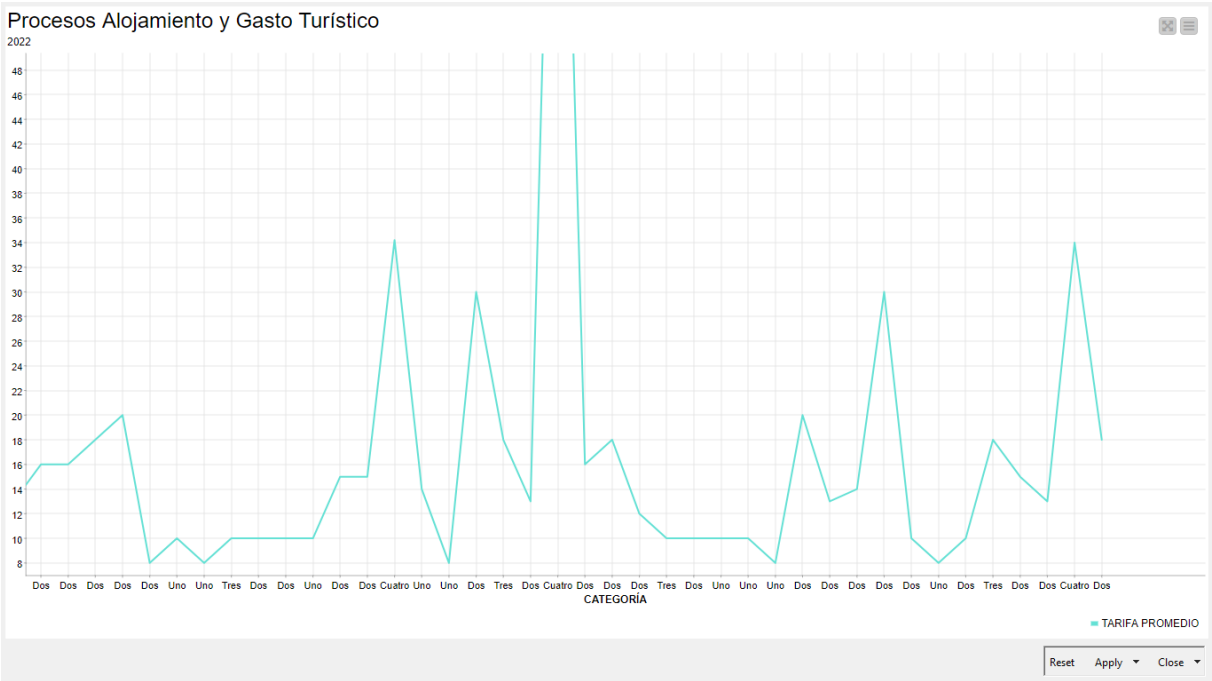


Figura 68. Tarifas promedio de los sitios de alojamiento basados en su categoría en el año 2022

En el siguiente gráfico se muestra el porcentaje de la demanda turística de personas nacionales y extranjeras, haciendo usos de sitios de alojamiento basados en la categoría en el año 2022.

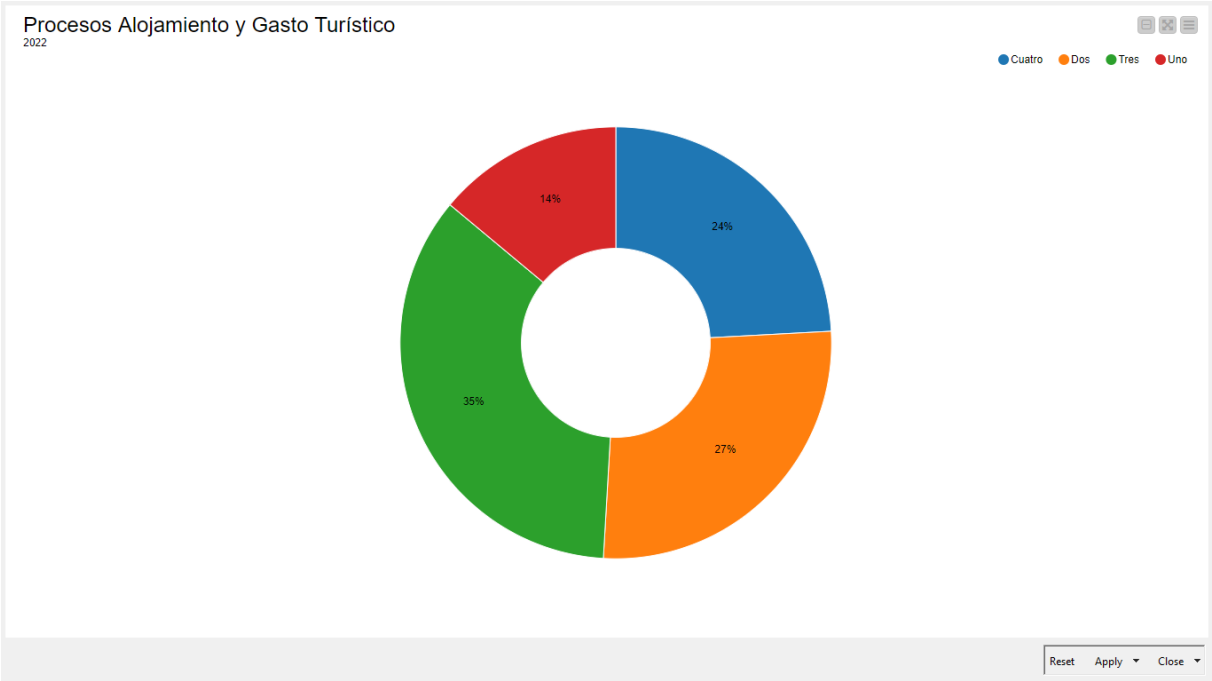


Figura 69. Porcentaje de demanda turística en los sitios de alojamiento según la categoría en el año 2022

En las siguientes gráficas no muestran la demanda turística de la provincia con relación a personas nacionales y extranjeras en el año 2022.

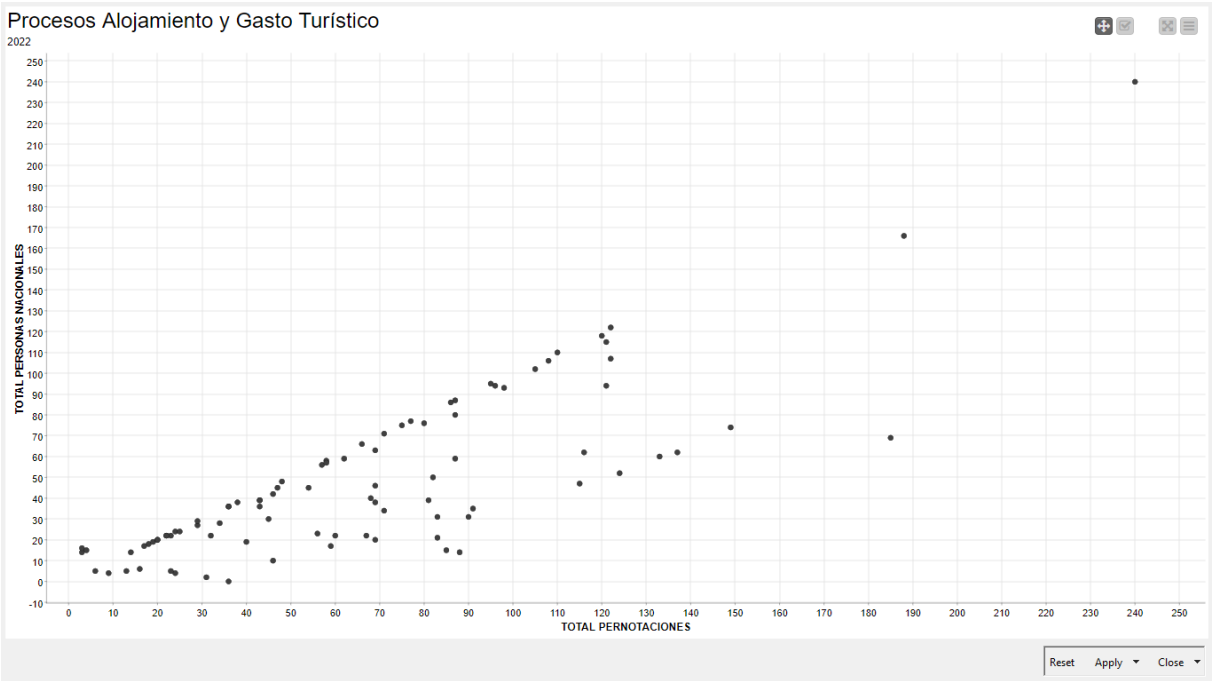


Figura 70. *Demanda turística de personas nacionales en el año 2022*

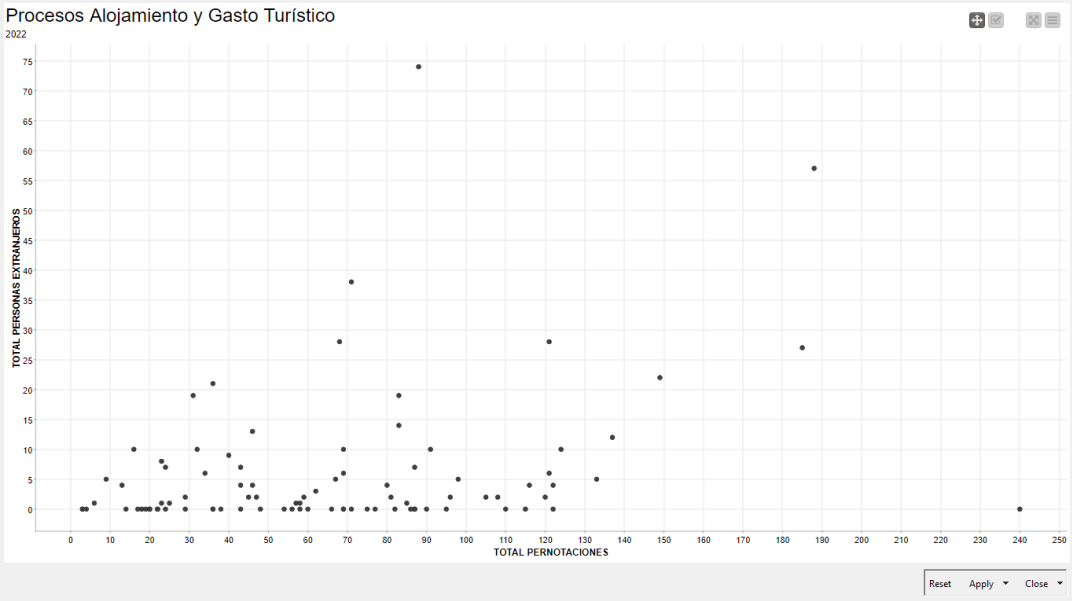


Figura 71. *Demanda turística de extranjeras nacionales en el año 2022*

- **Verificar Calidad de los Datos**

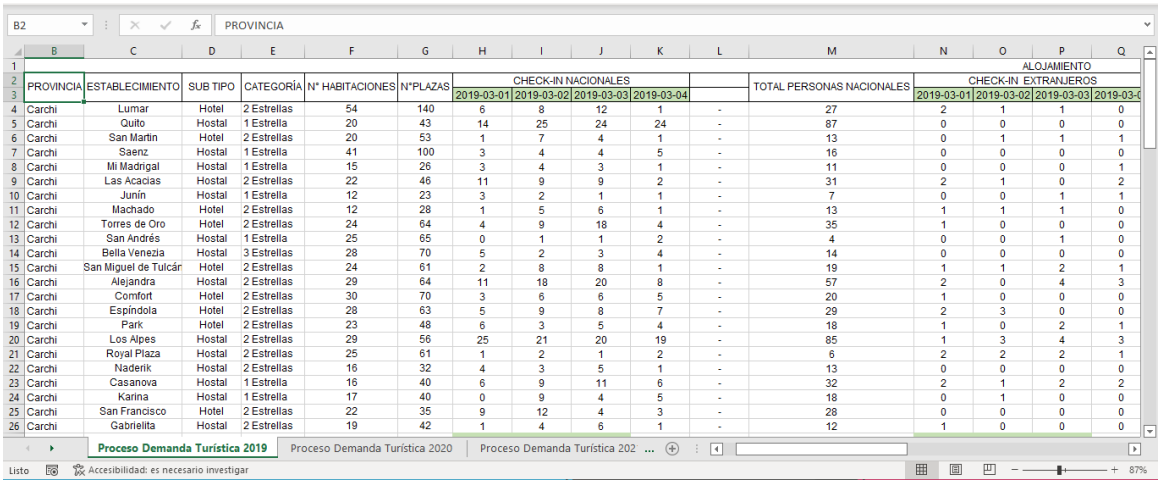
Tomando como referencia lo que se hizo con los datos iniciales para proceder a explorarlos con el fin de establecer criterios de calidad, podemos explicar en general que se encontraron algunos errores e inconsistencias en los datos de los procesos de alojamiento y gasto turístico los cuales determinan la demanda turística de la provincia, algunos de los que podemos mencionar es que se encontraron datos vacíos que generaban relaciones ilógicas entre los atributos y las variables, por lo cual algunos datos fueron eliminados y otros corregidos a fin de que todos los datos tengan lógica y se relacionen, cabe destacar que los datos donde existió el mayor grado de error fueron en los procesos que corresponden al año 2021. Durante la exploración se encontraron valores nulos en los distintos procesos a través de los años, lo que opto hacer con esos datos es no cambiar ni modificarlos, sino tomarlos en cuenta para la exploración, debido a que resultan útiles tomando en cuenta que son datos que intervienen directamente con la demanda turística que ha existido en la provincia. Una vez realizado la revisión y exploración de los datos iniciales con las correcciones realizadas en la data de los procesos de alojamiento y gasto turístico de todos los años, se puede decir que ahora los datos cubren la calidad de exigencia que deben tener un data en términos de completitud y consistencia para la aplicabilidad de un modelado de minería de datos, todo con el fin de cumplir los objetivos de la investigación y poder obtener los resultados esperados.

4.1.4.3. Preparación de los Datos

Esta fase de la metodología CRISP-DM, me permite preparar los datos de manera que estén listos para aplicar la técnica de modelado de minería de datos sobre la date que se tiene, en este caso se aplicara sobres los datos de los procesos de alojamiento y gasto turístico. Para lograr esta fase se deben cumplir tareas específicas que son propuestas por la metodología, en resumen, se debe iniciar analizando los datos aplicados en la exploración inicial, limpiar la data con el fin de mejorar su calidad, en caso de ser necesario agregar nuevas variables con nuevos atributos a partir de los datos originales, y otra parte importe adaptar los datos al modelado de minería de datos que se va a aplicar; ya sea cambios en el formato, tipos de datos, entre otros; con el fin de no generar inconsistencias o errores.

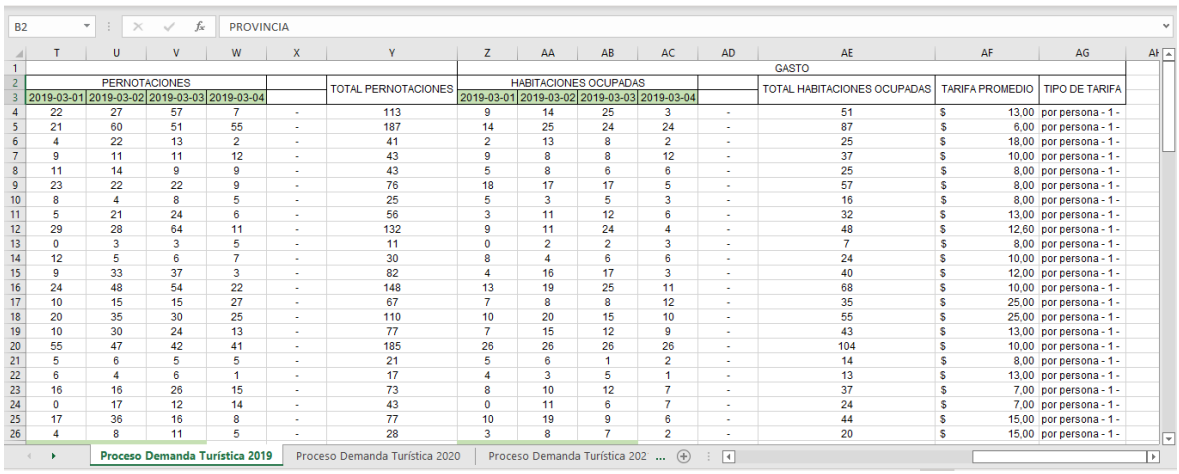
- **Seleccionar los Datos**

Para la elección de los datos que van a ser analizados en primera instancia se toma como referencia la base de datos original de los procesos de alojamiento y gasto turístico, donde son registros históricos desde el año 2019 hasta el año 2022. A continuación, se muestra los datos originales de los procesos de demanda turística desde los años ya descritos:



	PROVINCIA	ESTABLECIMIENTO	SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	CHECK-IN NACIONALES					TOTAL PERSONAS NACIONALES	ALOJAMIENTO CHECK-IN EXTRANJEROS			
							2019-03-01	2019-03-02	2019-03-03	2019-03-04			2019-03-01	2019-03-02	2019-03-03	2019-03-0
4	Carchi	Lumar	Hotel	2 Estrellas	54	140	6	8	12	1	-	27	2	1	1	0
5	Carchi	Quito	Hostal	1 Estrella	20	43	14	25	24	24	-	87	0	0	0	0
6	Carchi	San Martin	Hotel	2 Estrellas	20	53	1	7	4	1	-	13	0	1	1	1
7	Carchi	Saenz	Hostal	1 Estrella	41	100	3	4	4	5	-	16	0	0	0	0
8	Carchi	Mi Madrigal	Hostal	1 Estrella	15	26	3	4	3	1	-	11	0	0	0	1
9	Carchi	Las Acacias	Hostal	2 Estrellas	22	46	11	9	9	2	-	31	2	1	0	2
10	Carchi	Junín	Hostal	1 Estrella	12	23	3	2	1	1	-	7	0	0	1	1
11	Carchi	Machado	Hotel	2 Estrellas	12	28	1	5	6	1	-	13	1	1	1	0
12	Carchi	Torres de Oro	Hotel	2 Estrellas	24	64	4	9	18	4	-	35	1	0	0	0
13	Carchi	San Andrés	Hostal	1 Estrella	25	65	0	1	1	2	-	4	0	0	1	0
14	Carchi	Bella Venezia	Hostal	3 Estrellas	28	70	5	2	3	4	-	14	0	0	0	0
15	Carchi	San Miguel de Tulcár	Hotel	2 Estrellas	24	61	2	8	8	1	-	19	1	1	2	1
16	Carchi	Alejandra	Hostal	2 Estrellas	29	64	11	18	20	8	-	57	2	0	4	3
17	Carchi	Comfort	Hotel	2 Estrellas	30	70	3	6	6	5	-	20	1	0	0	0
18	Carchi	Espíndola	Hotel	2 Estrellas	28	63	5	9	8	7	-	29	2	3	0	0
19	Carchi	Park	Hotel	2 Estrellas	23	48	6	3	5	4	-	18	1	0	2	1
20	Carchi	Los Alpes	Hostal	2 Estrellas	29	56	25	21	20	19	-	85	1	3	4	3
21	Carchi	Royal Plaza	Hostal	2 Estrellas	25	61	1	2	1	2	-	6	2	2	2	1
22	Carchi	Naderik	Hostal	2 Estrellas	16	32	4	3	5	1	-	13	0	0	0	0
23	Carchi	Casanova	Hostal	1 Estrella	16	40	6	9	11	6	-	32	2	1	2	2
24	Carchi	Karina	Hostal	1 Estrella	17	40	0	9	4	5	-	18	0	1	0	0
25	Carchi	San Francisco	Hotel	2 Estrellas	22	35	9	12	4	3	-	28	0	0	0	0
26	Carchi	Gabrielita	Hostal	2 Estrellas	19	42	1	4	6	1	-	12	1	0	0	0

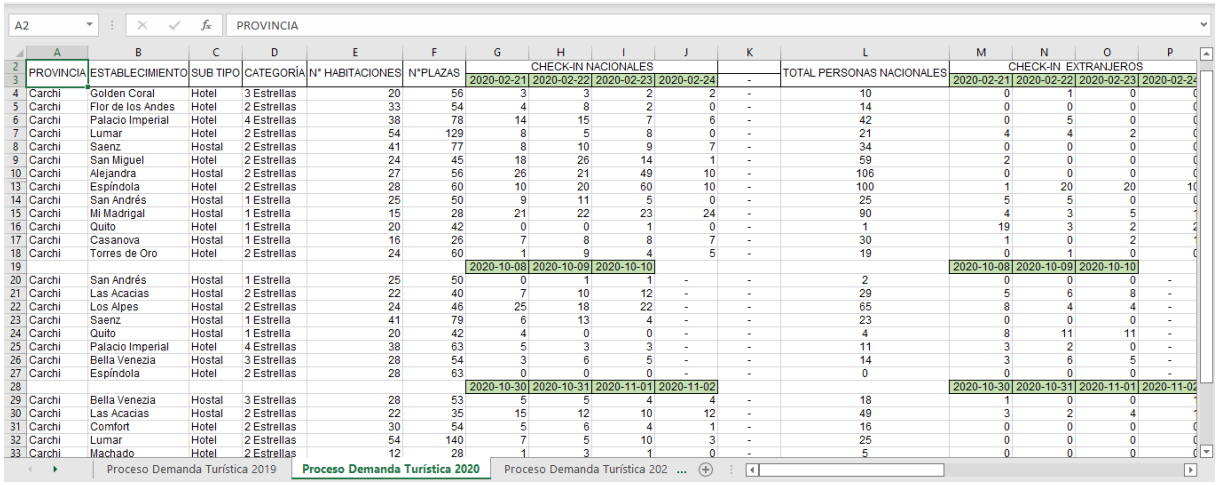
Figura 72. Datos originales de los procesos de demanda turística de la provincia del 2019



	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1												GASTO		
2	PERNOTACIONES					TOTAL PERNOTACIONES	HABITACIONES OCUPADAS					TOTAL HABITACIONES OCUPADAS	TARIFA PROMEDIO	TIPO DE TARIFA
3	2019-03-01	2019-03-02	2019-03-03	2019-03-04			2019-03-01	2019-03-02	2019-03-03	2019-03-04				
4	22	27	57	7	-	113	9	14	25	3	-	51	$ 13,00	por persona - 1 -
5	21	60	51	55	-	187	14	25	24	24	-	87	$ 6,00	por persona - 1 -
6	4	22	13	2	-	41	2	13	8	2	-	25	$ 18,00	por persona - 1 -
7	9	11	11	12	-	43	9	8	8	12	-	37	$ 10,00	por persona - 1 -
8	11	14	9	9	-	43	5	8	6	6	-	25	$ 8,00	por persona - 1 -
9	23	22	22	9	-	76	18	17	17	5	-	57	$ 8,00	por persona - 1 -
10	8	4	8	5	-	25	5	3	5	3	-	16	$ 8,00	por persona - 1 -
11	5	21	24	6	-	56	3	11	12	6	-	32	$ 13,00	por persona - 1 -
12	29	28	64	11	-	132	9	11	24	4	-	48	$ 12,60	por persona - 1 -
13	0	3	3	5	-	11	0	2	2	3	-	7	$ 8,00	por persona - 1 -
14	12	5	6	7	-	30	8	4	6	6	-	24	$ 10,00	por persona - 1 -
15	9	33	37	3	-	82	4	16	17	3	-	40	$ 12,00	por persona - 1 -
16	24	48	54	22	-	148	13	19	25	11	-	68	$ 10,00	por persona - 1 -
17	10	15	15	27	-	67	7	8	8	12	-	35	$ 25,00	por persona - 1 -
18	20	35	30	25	-	110	10	20	15	10	-	55	$ 25,00	por persona - 1 -
19	10	30	24	13	-	77	7	15	12	9	-	43	$ 13,00	por persona - 1 -
20	55	47	42	41	-	185	26	26	26	26	-	104	$ 10,00	por persona - 1 -
21	5	6	5	5	-	21	5	6	1	2	-	14	$ 8,00	por persona - 1 -
22	6	4	6	1	-	17	4	3	5	1	-	13	$ 13,00	por persona - 1 -
23	16	16	26	15	-	73	8	10	12	7	-	37	$ 7,00	por persona - 1 -
24	0	17	12	14	-	43	0	11	6	7	-	24	$ 7,00	por persona - 1 -
25	17	36	16	8	-	77	10	19	9	6	-	44	$ 15,00	por persona - 1 -
26	4	8	11	5	-	28	3	8	7	2	-	20	$ 15,00	por persona - 1 -

Proceso Demanda Turística 2019 | Proceso Demanda Turística 2020 | Proceso Demanda Turística 202 ...

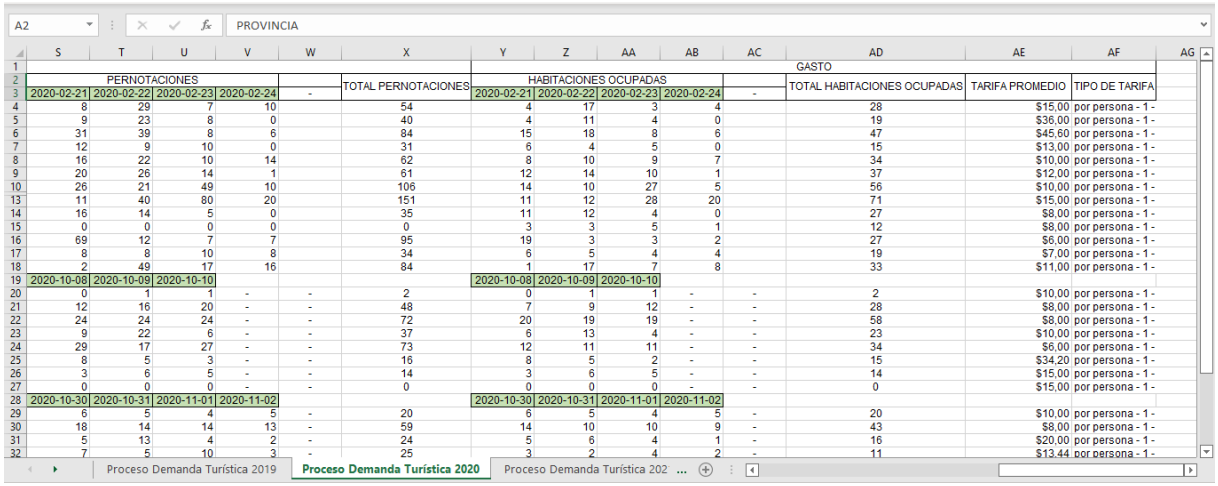
Figura 73. Datos originales de los procesos de demanda turística de la provincia del 2019



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
2	PROVINCIA	ESTABLECIMIENTO	SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	CHECK-IN NACIONALES					TOTAL PERSONAS NACIONALES	CHECK-IN EXTRANJEROS			
3							2020-02-21	2020-02-22	2020-02-23	2020-02-24	-		2020-02-21	2020-02-22	2020-02-23	2020-02-24
4	Carchi	Golden Coral	Hotel	3 Estrellas	20	56	3	3	2	2	-	10	0	1	0	0
5	Carchi	Flor de los Andes	Hotel	2 Estrellas	33	54	4	8	2	0	-	14	0	0	0	0
6	Carchi	Palacio Imperial	Hotel	4 Estrellas	38	78	14	15	7	6	-	42	0	5	0	0
7	Carchi	Lumar	Hotel	2 Estrellas	54	129	8	5	8	0	-	21	4	4	2	0
8	Carchi	Saenz	Hostal	2 Estrellas	41	77	8	10	9	7	-	34	0	0	0	0
9	Carchi	San Miguel	Hotel	2 Estrellas	24	45	18	26	14	1	-	59	2	0	0	0
10	Carchi	Alejandra	Hostal	2 Estrellas	27	56	26	21	49	10	-	106	0	0	0	0
13	Carchi	Espíndola	Hotel	2 Estrellas	28	60	10	20	60	10	-	100	1	20	20	10
14	Carchi	San Andrés	Hostal	1 Estrella	25	50	9	11	5	0	-	25	5	5	0	0
15	Carchi	Mi Madrigal	Hostal	1 Estrella	15	28	21	22	23	24	-	90	4	3	5	1
16	Carchi	Quito	Hotel	1 Estrella	20	42	0	0	1	0	-	1	19	3	2	2
17	Carchi	Casanova	Hostal	1 Estrella	16	26	7	8	8	7	-	30	1	0	2	1
18	Carchi	Torres de Oro	Hotel	2 Estrellas	24	60	1	9	4	5	-	19	0	1	0	0
19							2020-10-08	2020-10-09	2020-10-10				2020-10-08	2020-10-09	2020-10-10	
20	Carchi	San Andrés	Hostal	1 Estrella	25	50	0	1	1	-	-	2	0	0	0	-
21	Carchi	Las Acacias	Hostal	2 Estrellas	22	40	7	10	12	-	-	29	5	6	8	-
22	Carchi	Los Alpes	Hostal	2 Estrellas	24	46	25	18	22	-	-	65	8	4	4	-
23	Carchi	Saenz	Hostal	1 Estrella	41	79	6	13	4	-	-	23	0	0	0	-
24	Carchi	Quito	Hostal	1 Estrella	20	42	4	0	0	-	-	4	8	11	11	-
25	Carchi	Palacio Imperial	Hotel	4 Estrellas	38	63	5	3	3	-	-	11	3	2	0	-
26	Carchi	Bella Venezia	Hostal	3 Estrellas	28	54	3	6	5	-	-	14	3	6	5	-
27	Carchi	Espíndola	Hotel	2 Estrellas	28	63	0	0	0	-	-	0	0	0	0	-
28							2020-10-30	2020-10-31	2020-11-01	2020-11-02			2020-10-30	2020-10-31	2020-11-01	2020-11-02
29	Carchi	Bella Venezia	Hostal	3 Estrellas	28	53	5	5	4	4	-	18	1	0	0	1
30	Carchi	Las Acacias	Hostal	2 Estrellas	22	35	15	12	10	12	-	49	3	2	4	1
31	Carchi	Comfort	Hotel	2 Estrellas	30	54	5	6	4	1	-	16	0	0	0	0
32	Carchi	Lumar	Hotel	2 Estrellas	54	140	7	5	10	3	-	25	0	0	0	0
33	Carchi	Machado	Hotel	2 Estrellas	12	28	1	3	1	0	-	5	0	0	0	0

Proceso Demanda Turística 2019 | Proceso Demanda Turística 2020 | Proceso Demanda Turística 202 ...

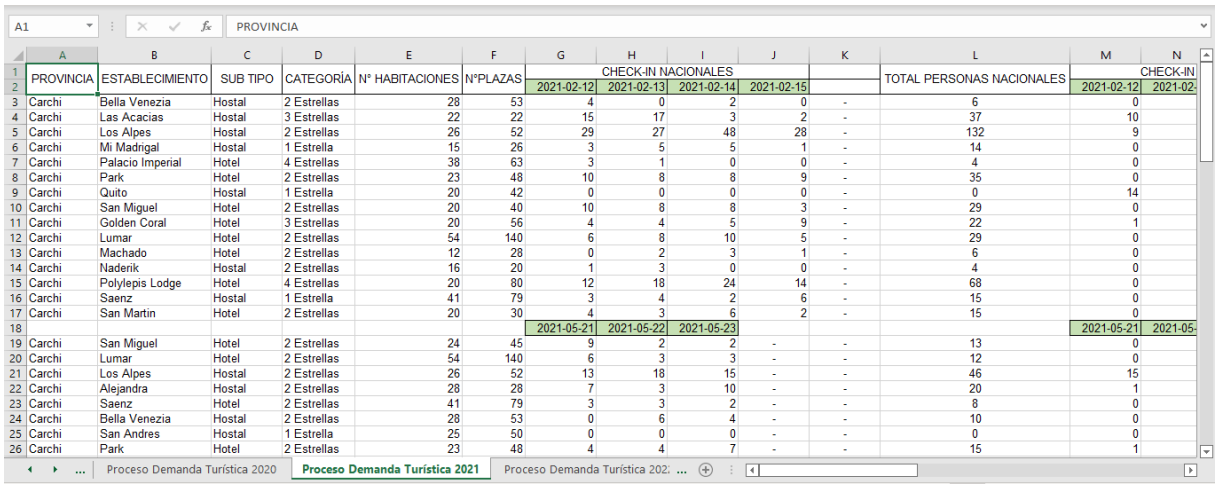
Figura 74. Datos originales de los procesos de demanda turística de la provincia del 2020



	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1												GASTO			
2	PERNOTACIONES					TOTAL PERNOTACIONES	HABITACIONES OCUPADAS					TOTAL HABITACIONES OCUPADAS	TARIFA PROMEDIO	TIPO DE TARIFA	
3	2020-02-21	2020-02-22	2020-02-23	2020-02-24	-		2020-02-21	2020-02-22	2020-02-23	2020-02-24	-				
4	8	29	7	10		54	4	17	3	4		28	$15,00	por persona - 1 -	
5	9	23	8	0		40	4	11	4	0		19	$36,00	por persona - 1 -	
6	31	39	8	6		84	15	18	8	6		47	$45,60	por persona - 1 -	
7	12	9	10	0		31	6	4	5	0		15	$13,00	por persona - 1 -	
8	16	22	10	14		62	8	10	9	7		34	$10,00	por persona - 1 -	
9	20	26	14	1		61	12	14	10	1		37	$12,00	por persona - 1 -	
10	26	21	49	10		106	14	10	27	5		56	$10,00	por persona - 1 -	
13	11	40	80	20		151	11	12	28	20		71	$15,00	por persona - 1 -	
14	16	14	5	0		35	11	12	4	0		27	$8,00	por persona - 1 -	
15	0	0	0	0		0	3	3	5	1		12	$8,00	por persona - 1 -	
16	69	12	7	7		95	19	3	3	2		27	$6,00	por persona - 1 -	
17	8	8	10	8		34	6	5	4	4		19	$7,00	por persona - 1 -	
18	2	49	17	16		84	1	17	7	8		33	$11,00	por persona - 1 -	
19	2020-10-08	2020-10-09	2020-10-10				2020-10-08	2020-10-09	2020-10-10						
20	0	1	1	-	-	2	0	1	1	-	-	2	$10,00	por persona - 1 -	
21	12	16	20	-	-	48	7	9	12	-	-	28	$8,00	por persona - 1 -	
22	24	24	24	-	-	72	20	19	19	-	-	58	$8,00	por persona - 1 -	
23	9	22	6	-	-	37	6	13	4	-	-	23	$10,00	por persona - 1 -	
24	29	17	27	-	-	73	12	11	11	-	-	34	$6,00	por persona - 1 -	
25	8	5	3	-	-	16	8	5	2	-	-	15	$34,20	por persona - 1 -	
26	3	6	5	-	-	14	3	6	5	-	-	14	$15,00	por persona - 1 -	
27	0	0	0	-	-	0	0	0	0	-	-	0	$15,00	por persona - 1 -	
28	2020-10-30	2020-10-31	2020-11-01	2020-11-02			2020-10-30	2020-10-31	2020-11-01	2020-11-02					
29	6	5	4	5	-	20	6	5	4	5	-	20	$10,00	por persona - 1 -	
30	18	14	14	13	-	59	14	10	10	9	-	43	$8,00	por persona - 1 -	
31	5	13	4	2	-	24	5	6	4	1	-	16	$20,00	por persona - 1 -	
32	7	5	10	3	-	25	3	2	4	2	-	11	$13,44	por persona - 1 -	

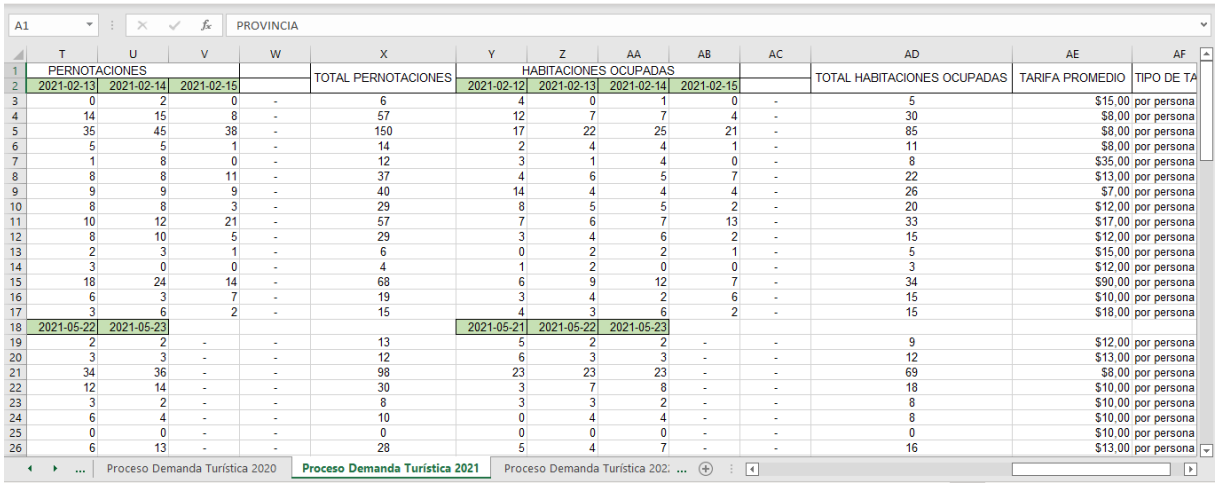
Proceso Demanda Turística 2019 | Proceso Demanda Turística 2020 | Proceso Demanda Turística 202 ...

Figura 75. Datos originales de los procesos de demanda turística de la provincia del 2020



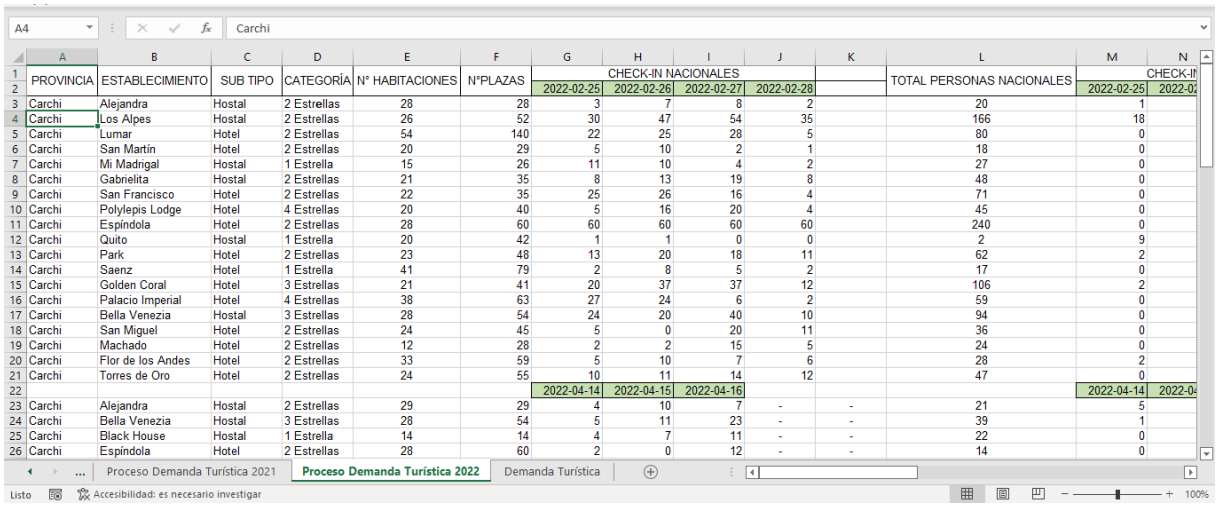
PROVINCIA	ESTABLECIMIENTO	SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	CHECK-IN NACIONALES					TOTAL PERSONAS NACIONALES	CHECK-IN
						2021-02-12	2021-02-13	2021-02-14	2021-02-15			2021-02-12
Carchi	Bella Venezia	Hostal	2 Estrellas	28	53	4	0	2	0	-	6	0
Carchi	Las Acacias	Hostal	3 Estrellas	22	22	15	17	3	2	-	37	10
Carchi	Los Alpes	Hostal	2 Estrellas	26	52	29	27	48	28	-	132	9
Carchi	Mi Madrigal	Hostal	1 Estrella	15	26	3	5	5	1	-	14	0
Carchi	Palacio Imperial	Hotel	4 Estrellas	38	63	3	1	0	0	-	4	0
Carchi	Park	Hotel	2 Estrellas	23	48	10	8	8	9	-	35	0
Carchi	Quito	Hostal	1 Estrella	20	42	0	0	0	0	-	0	14
Carchi	San Miguel	Hotel	2 Estrellas	20	40	10	8	8	3	-	29	0
Carchi	Golden Coral	Hotel	3 Estrellas	20	56	4	4	5	9	-	22	1
Carchi	Lumar	Hotel	2 Estrellas	54	140	6	8	10	5	-	29	0
Carchi	Machado	Hotel	2 Estrellas	12	28	0	2	3	1	-	6	0
Carchi	Naderik	Hostal	2 Estrellas	16	20	1	3	0	0	-	4	0
Carchi	Polylepis Lodge	Hotel	4 Estrellas	20	80	12	18	24	14	-	68	0
Carchi	Saenz	Hostal	1 Estrella	41	79	3	4	2	6	-	15	0
Carchi	San Martin	Hotel	2 Estrellas	20	30	4	3	6	2	-	15	0
						2021-05-21	2021-05-22	2021-05-23				2021-05-21
Carchi	San Miguel	Hotel	2 Estrellas	24	45	9	2	2	-	-	13	0
Carchi	Lumar	Hotel	2 Estrellas	54	140	6	3	3	-	-	12	0
Carchi	Los Alpes	Hostal	2 Estrellas	26	52	13	18	15	-	-	46	15
Carchi	Alejandra	Hostal	2 Estrellas	28	28	7	3	10	-	-	20	1
Carchi	Saenz	Hotel	2 Estrellas	41	79	3	3	2	-	-	8	0
Carchi	Bella Venezia	Hostal	2 Estrellas	28	53	0	6	4	-	-	10	0
Carchi	San Andres	Hostal	1 Estrella	25	50	0	0	0	-	-	0	0
Carchi	Park	Hotel	2 Estrellas	23	48	4	4	7	-	-	15	1

Figura 76. Datos originales de los procesos de demanda turística de la provincia del 2021



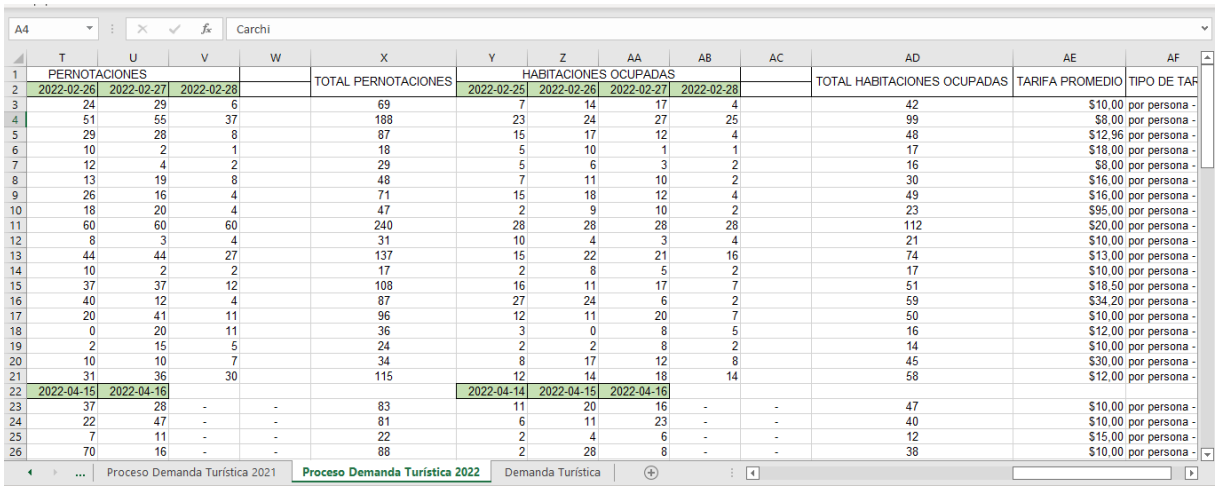
PERNOTACIONES				TOTAL PERNOTACIONES	HABITACIONES OCUPADAS					TOTAL HABITACIONES OCUPADAS	TARIFA PROMEDIO	TIPO DE TA
2021-02-13	2021-02-14	2021-02-15			2021-02-12	2021-02-13	2021-02-14	2021-02-15				
0	2	0	-	6	4	0	1	0	-	5	$15,00	por persona
14	15	8	-	57	12	7	7	4	-	30	$8,00	por persona
35	45	38	-	150	17	22	25	21	-	85	$8,00	por persona
5	5	1	-	14	2	4	4	1	-	11	$8,00	por persona
1	8	0	-	12	3	1	4	0	-	8	$35,00	por persona
8	8	11	-	37	4	6	5	7	-	22	$13,00	por persona
9	9	9	-	40	14	4	4	4	-	26	$7,00	por persona
8	8	3	-	29	8	5	5	2	-	20	$12,00	por persona
10	12	21	-	57	7	6	7	13	-	33	$17,00	por persona
8	10	5	-	29	3	4	6	2	-	15	$12,00	por persona
2	3	1	-	6	0	2	2	1	-	5	$15,00	por persona
3	0	0	-	4	1	2	0	0	-	3	$12,00	por persona
18	24	14	-	68	6	9	12	7	-	34	$90,00	por persona
6	3	7	-	19	3	4	2	6	-	15	$10,00	por persona
3	6	2	-	15	4	3	6	2	-	15	$18,00	por persona
2021-05-22	2021-05-23				2021-05-21	2021-05-22	2021-05-23					
2	2	-	-	13	5	2	2	-	-	9	$12,00	por persona
3	3	-	-	12	6	3	3	-	-	12	$13,00	por persona
34	36	-	-	98	23	23	23	-	-	69	$8,00	por persona
12	14	-	-	30	3	7	8	-	-	18	$10,00	por persona
3	2	-	-	8	3	3	2	-	-	8	$10,00	por persona
6	4	-	-	10	0	4	4	-	-	8	$10,00	por persona
0	0	-	-	0	0	0	0	-	-	0	$10,00	por persona
6	13	-	-	28	5	4	7	-	-	16	$13,00	por persona

Figura 77. Datos originales de los procesos de demanda turística de la provincia del 2021



PROVINCIA	ESTABLECIMIENTO	SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	CHECK-IN NACIONALES					TOTAL PERSONAS NACIONALES	CHECK-IN
						2022-02-25	2022-02-26	2022-02-27	2022-02-28			2022-02-25
Carchi	Alejandra	Hostal	2 Estrellas	28	28	3	7	8	2		20	1
Carchi	Los Alpes	Hostal	2 Estrellas	26	52	30	47	54	35		166	18
Carchi	Lumar	Hotel	2 Estrellas	54	140	22	25	28	5		80	0
Carchi	San Martín	Hotel	2 Estrellas	20	29	5	10	2	1		18	0
Carchi	Mi Madrigal	Hostal	1 Estrella	15	26	11	10	4	2		27	0
Carchi	Gabrielita	Hostal	2 Estrellas	21	35	8	13	19	8		48	0
Carchi	San Francisco	Hotel	2 Estrellas	22	35	25	26	16	4		71	0
Carchi	Polylepis Lodge	Hotel	4 Estrellas	20	40	5	16	20	4		45	0
Carchi	Espíndola	Hotel	2 Estrellas	28	60	60	60	60	60		240	0
Carchi	Quito	Hostal	1 Estrella	20	42	1	1	0	0		2	9
Carchi	Park	Hotel	2 Estrellas	23	48	13	20	18	11		62	2
Carchi	Saenz	Hotel	1 Estrella	41	79	2	8	5	2		17	0
Carchi	Golden Coral	Hotel	3 Estrellas	21	41	20	37	37	12		106	2
Carchi	Palacio Imperial	Hotel	4 Estrellas	38	63	27	24	6	2		59	0
Carchi	Bella Venezia	Hostal	3 Estrellas	28	54	24	20	40	10		94	0
Carchi	San Miguel	Hotel	2 Estrellas	24	45	5	0	20	11		36	0
Carchi	Machado	Hotel	2 Estrellas	12	28	2	2	15	5		24	0
Carchi	Flor de los Andes	Hotel	2 Estrellas	33	59	5	10	7	6		28	2
Carchi	Torres de Oro	Hotel	2 Estrellas	24	55	10	11	14	12		47	0
						2022-04-14	2022-04-15	2022-04-16				2022-04-14
Carchi	Alejandra	Hostal	2 Estrellas	29	29	4	10	7	-	-	21	5
Carchi	Bella Venezia	Hostal	3 Estrellas	28	54	5	11	23	-	-	39	1
Carchi	Black House	Hostal	1 Estrella	14	14	4	7	11	-	-	22	0
Carchi	Espíndola	Hotel	2 Estrellas	28	60	2	0	12	-	-	14	0

Figura 78. Datos originales de los procesos de demanda turística de la provincia del 2022



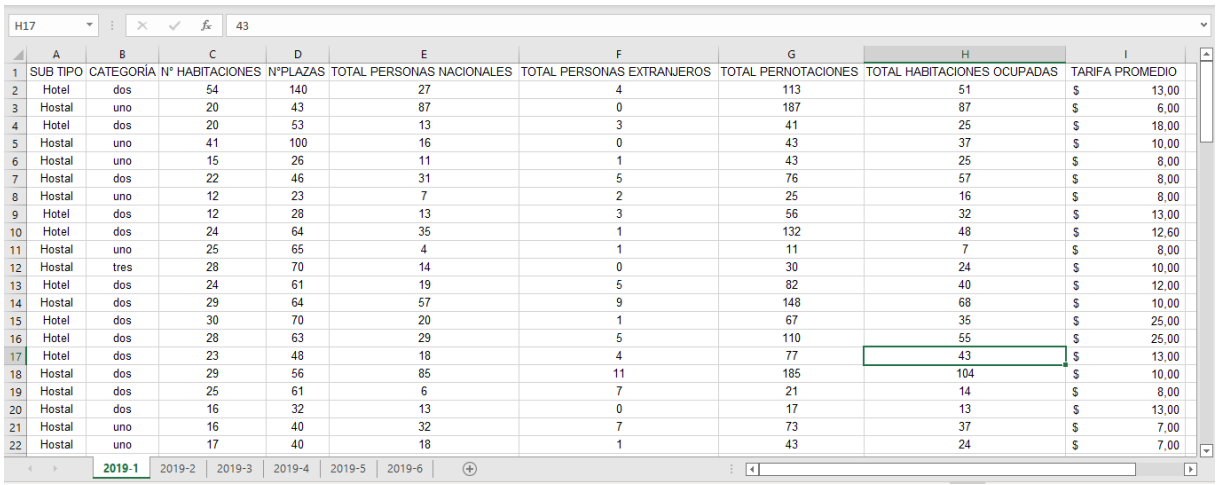
	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	PERNOTACIONES				TOTAL PERNOTACIONES	HABITACIONES OCUPADAS					TOTAL HABITACIONES OCUPADAS	TARIFA PROMEDIO	TIPO DE TAR
2	2022-02-26	2022-02-27	2022-02-28			2022-02-25	2022-02-26	2022-02-27	2022-02-28				
3	24	29	6		69	7	14	17	4		42	$10,00	por persona -
4	51	55	37		188	23	24	27	25		99	$8,00	por persona -
5	29	28	8		87	15	17	12	4		48	$12,96	por persona -
6	10	2	1		18	5	10	1	1		17	$18,00	por persona -
7	12	4	2		29	5	6	3	2		16	$8,00	por persona -
8	13	19	8		48	7	11	10	2		30	$16,00	por persona -
9	26	16	4		71	15	18	12	4		49	$16,00	por persona -
10	18	20	4		47	2	9	10	2		23	$95,00	por persona -
11	60	60	60		240	28	28	28	28		112	$20,00	por persona -
12	8	3	4		31	10	4	3	4		21	$10,00	por persona -
13	44	44	27		137	15	22	21	16		74	$13,00	por persona -
14	10	2	2		17	2	8	5	2		17	$10,00	por persona -
15	37	37	12		108	16	11	17	7		51	$18,50	por persona -
16	40	12	4		87	27	24	6	2		59	$34,20	por persona -
17	20	41	11		96	12	11	20	7		50	$10,00	por persona -
18	0	20	11		36	3	0	8	5		16	$12,00	por persona -
19	2	15	5		24	2	2	8	2		14	$10,00	por persona -
20	10	10	7		34	8	17	12	8		45	$30,00	por persona -
21	31	36	30		115	12	14	18	14		58	$12,00	por persona -
22	2022-04-15	2022-04-16				2022-04-14	2022-04-15	2022-04-16					
23	37	28	-	-	83	11	20	16	-	-	47	$10,00	por persona -
24	22	47	-	-	81	6	11	23	-	-	40	$10,00	por persona -
25	7	11	-	-	22	2	4	6	-	-	12	$15,00	por persona -
26	70	16	-	-	88	2	28	8	-	-	38	$10,00	por persona -

Figura 79. Datos originales de los procesos de demanda turística de la provincia del 2022

Cabe mencionar que en las imágenes donde se muestran los datos, son solamente una sección de todo el conjunto original de los datos de los procesos de alojamiento y gasto turístico. Por otra parte, para hacer la elección de un conjunto o subconjuntos de datos, como dice la metodología, se tomó como referencia los datos de la exploración que se realizado, donde el conjunto de datos original fue reestructura sin cambiar el sentido a fin de adaptarse a la herramienta mediante la cual se hizo el análisis exploratorio.

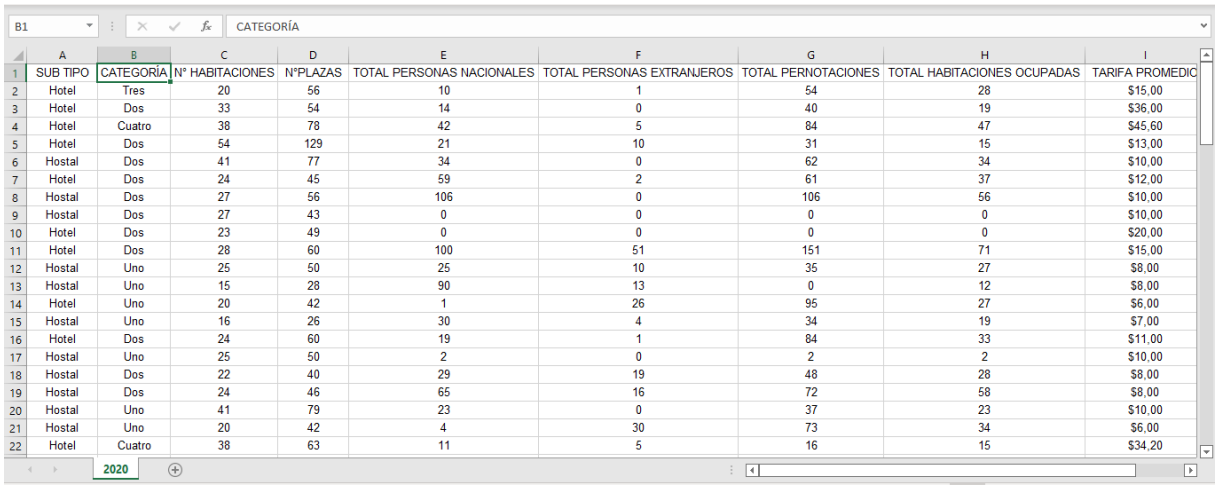
Ahora bien, en la nueva base de información que se tiene, se excluyeron algunas variables y atributos de la base original, debido a que se necesitaba que quedara solamente información que apoye a la investigación, a las técnicas de modelado de minería de datos y en general cumpla los objetivos propuestos en la investigación y en las primeras fases de la metodología. Otro de los casos por lo que se tuvieron que reemplazar y quitar atributos fue que existían datos en blanco y algunos generaban inconsistencias entre las variables que se tenía.

Los datos seleccionados para el análisis del modelado de minería de datos se muestran en las siguientes figuras:



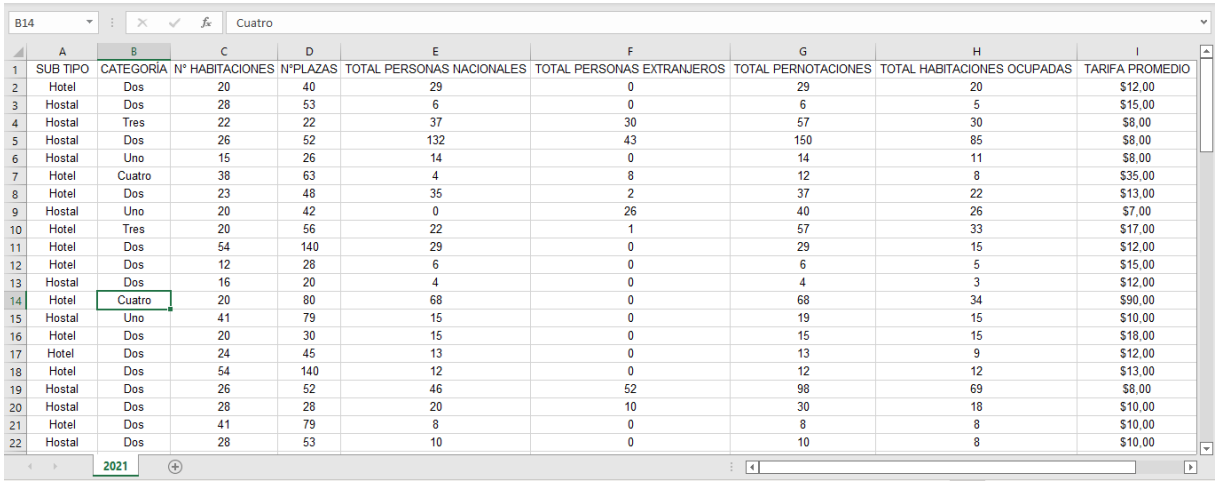
SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	TOTAL PERSONAS NACIONALES	TOTAL PERSONAS EXTRANJEROS	TOTAL PERNOTACIONES	TOTAL HABITACIONES OCUPADAS	TARIFA PROMEDIO
Hotel	dos	54	140	27	4	113	51	$ 13,00
Hostal	uno	20	43	87	0	187	87	$ 6,00
Hotel	dos	20	53	13	3	41	25	$ 18,00
Hostal	uno	41	100	16	0	43	37	$ 10,00
Hostal	uno	15	26	11	1	43	25	$ 8,00
Hostal	dos	22	46	31	5	76	57	$ 8,00
Hostal	uno	12	23	7	2	25	16	$ 8,00
Hotel	dos	12	28	13	3	56	32	$ 13,00
Hotel	dos	24	64	35	1	132	48	$ 12,60
Hostal	uno	25	65	4	1	11	7	$ 8,00
Hostal	tres	28	70	14	0	30	24	$ 10,00
Hotel	dos	24	61	19	5	82	40	$ 12,00
Hostal	dos	29	64	57	9	148	68	$ 10,00
Hotel	dos	30	70	20	1	67	35	$ 25,00
Hotel	dos	28	63	29	5	110	55	$ 25,00
Hotel	dos	23	48	18	4	77	43	$ 13,00
Hostal	dos	29	56	85	11	185	104	$ 10,00
Hostal	dos	25	61	6	7	21	14	$ 8,00
Hostal	dos	16	32	13	0	17	13	$ 13,00
Hostal	uno	16	40	32	7	73	37	$ 7,00
Hostal	uno	17	40	18	1	43	24	$ 7,00

Figura 80. Subconjunto de datos seleccionados de los procesos de demanda turística en el año 2019



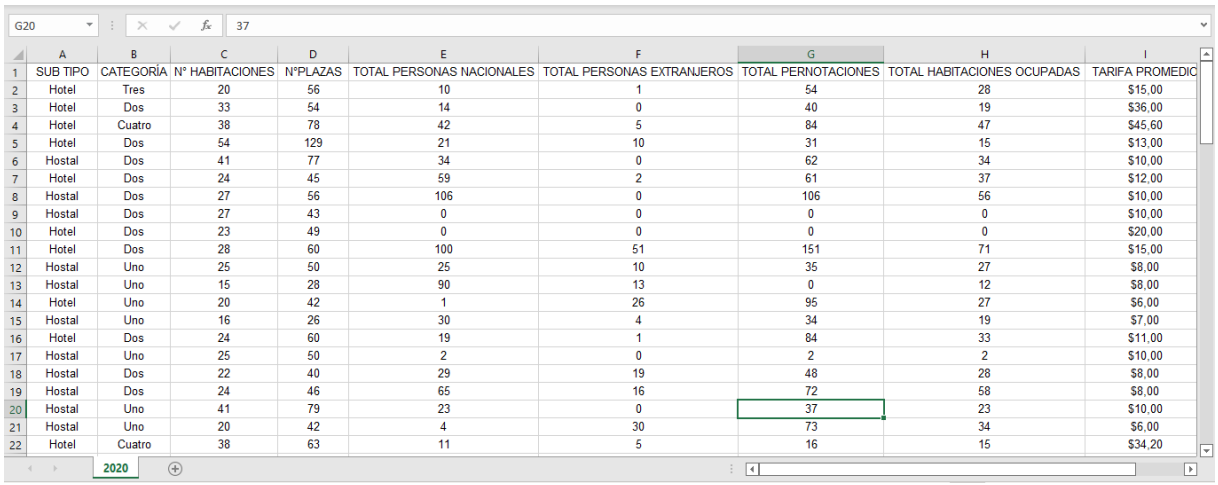
SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	TOTAL PERSONAS NACIONALES	TOTAL PERSONAS EXTRANJEROS	TOTAL PERNOTACIONES	TOTAL HABITACIONES OCUPADAS	TARIFA PROMEDIO
Hotel	Tres	20	56	10	1	54	28	$15,00
Hotel	Dos	33	54	14	0	40	19	$36,00
Hotel	Cuatro	38	78	42	5	84	47	$45,60
Hotel	Dos	54	129	21	10	31	15	$13,00
Hostal	Dos	41	77	34	0	62	34	$10,00
Hotel	Dos	24	45	59	2	61	37	$12,00
Hostal	Dos	27	56	106	0	106	56	$10,00
Hostal	Dos	27	43	0	0	0	0	$10,00
Hotel	Dos	23	49	0	0	0	0	$20,00
Hotel	Dos	28	60	100	51	151	71	$15,00
Hostal	Uno	25	50	25	10	35	27	$8,00
Hostal	Uno	15	28	90	13	0	12	$8,00
Hotel	Uno	20	42	1	26	95	27	$6,00
Hostal	Uno	16	26	30	4	34	19	$7,00
Hotel	Dos	24	60	19	1	84	33	$11,00
Hostal	Uno	25	50	2	0	2	2	$10,00
Hostal	Dos	22	40	29	19	48	28	$8,00
Hostal	Dos	24	46	65	16	72	58	$8,00
Hostal	Uno	41	79	23	0	37	23	$10,00
Hostal	Uno	20	42	4	30	73	34	$6,00
Hotel	Cuatro	38	63	11	5	16	15	$34,20

Figura 81. Subconjunto de datos seleccionados de los procesos de demanda turística en el año 2020



SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	TOTAL PERSONAS NACIONALES	TOTAL PERSONAS EXTRANJEROS	TOTAL PERNOTACIONES	TOTAL HABITACIONES OCUPADAS	TARIFA PROMEDIO
Hotel	Dos	20	40	29	0	29	20	$12,00
Hostal	Dos	28	53	6	0	6	5	$15,00
Hostal	Tres	22	22	37	30	57	30	$8,00
Hostal	Dos	26	52	132	43	150	85	$8,00
Hostal	Uno	15	26	14	0	14	11	$8,00
Hotel	Cuatro	38	63	4	8	12	8	$35,00
Hotel	Dos	23	48	35	2	37	22	$13,00
Hostal	Uno	20	42	0	26	40	26	$7,00
Hotel	Tres	20	56	22	1	57	33	$17,00
Hotel	Dos	54	140	29	0	29	15	$12,00
Hotel	Dos	12	28	6	0	6	5	$15,00
Hostal	Dos	16	20	4	0	4	3	$12,00
Hotel	Cuatro	20	80	68	0	68	34	$90,00
Hostal	Uno	41	79	15	0	19	15	$10,00
Hotel	Dos	20	30	15	0	15	15	$18,00
Hotel	Dos	24	45	13	0	13	9	$12,00
Hotel	Dos	54	140	12	0	12	12	$13,00
Hostal	Dos	26	52	46	52	98	69	$8,00
Hostal	Dos	28	28	20	10	30	18	$10,00
Hotel	Dos	41	79	8	0	8	8	$10,00
Hostal	Dos	28	53	10	0	10	8	$10,00

Figura 82. Subconjunto de datos seleccionados de los procesos de demanda turística en el año 2021



	A	B	C	D	E	F	G	H	I
1	SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	TOTAL PERSONAS NACIONALES	TOTAL PERSONAS EXTRANJEROS	TOTAL PERNOTACIONES	TOTAL HABITACIONES OCUPADAS	TARIFA PROMEDIO
2	Hotel	Tres	20	56	10	1	54	28	$15,00
3	Hotel	Dos	33	54	14	0	40	19	$36,00
4	Hotel	Cuatro	38	78	42	5	84	47	$45,60
5	Hotel	Dos	54	129	21	10	31	15	$13,00
6	Hostal	Dos	41	77	34	0	62	34	$10,00
7	Hotel	Dos	24	45	59	2	61	37	$12,00
8	Hostal	Dos	27	56	106	0	106	56	$10,00
9	Hostal	Dos	27	43	0	0	0	0	$10,00
10	Hotel	Dos	23	49	0	0	0	0	$20,00
11	Hotel	Dos	28	60	100	51	151	71	$15,00
12	Hostal	Uno	25	50	25	10	35	27	$8,00
13	Hostal	Uno	15	28	90	13	0	12	$8,00
14	Hotel	Uno	20	42	1	26	95	27	$6,00
15	Hostal	Uno	16	26	30	4	34	19	$7,00
16	Hotel	Dos	24	60	19	1	84	33	$11,00
17	Hostal	Uno	25	50	2	0	2	2	$10,00
18	Hostal	Dos	22	40	29	19	48	28	$8,00
19	Hostal	Dos	24	46	65	16	72	58	$8,00
20	Hostal	Uno	41	79	23	0	37	23	$10,00
21	Hostal	Uno	20	42	4	30	73	34	$6,00
22	Hotel	Cuatro	38	63	11	5	16	15	$34,20

Figura 83. Subconjunto de datos seleccionados de los procesos de demanda turística en el año 2022

Cabe mencionar que los datos que no fueron seleccionados no presentan un importante grado de relevancia con relación a los objetivos que se busca con el modelado de minería de datos.

- **Limpiar Datos**

La base de datos con las que cuenta este proyecto de investigación es en referencia a los procesos de alojamiento y gasto turístico de la provincia, los datos no se encuentran en repositorios digitales si no, se los encuentra en una serie de documentos físicos almacenados en ficheros, y para obtener la información de los procesos de los años anteriores y el actual, se tuvo que digitalizar toda esta información de modo que pueda ser usada para aplicar minerías sobre estos. Apegándonos a las características técnicas que posee el Ministerio de Turismo de la Provincia del Carchi, se optó por usar la herramienta ofimática Excel para almacenar todos los datos de los procesos, y sabiendo que la mayoría de las herramientas dedicadas a la aplicación de minería de datos soporta las distintas extensiones que maneja Excel (xlsx, CSV, xlsm, entre otros). Tomando en cuenta que los datos fueron transcritos directamente desde los ficheros originales hacia Excel no se requiere de una limpieza de datos en profundidad, más de allá de que en la base de datos exista datos nulos o datos ausentes, y a pesar de que existan datos faltantes con relación a la demanda turística de la provincia, no sería necesario estimar los valores sino excluir ese análisis, cabe mencionar que no existen datos faltantes en la base de datos que se tiene. Con relación a los datos nulos que, si se tiene en los datos, en el momento de aplicar la técnica de minería de datos simplemente se van ignoran debido a que no aportan ningún tipo de información a la investigación.

- **Estructuración de los datos**

Atributos modificados

La metodología de esta tarea nos dice que preparar los datos en base a la modificación de atributos a raíz de los atributos originales, en la base de datos que se tiene se modificó los atributos de una variable denominada *Categoría* su transformación conllevo a cambiar los atributos de tipo numérico y caracter a codificar los atributos de la variable en un formato de tipo caracter, no se cambió el sentido y la lógica que tenía la variable y sus atributos.

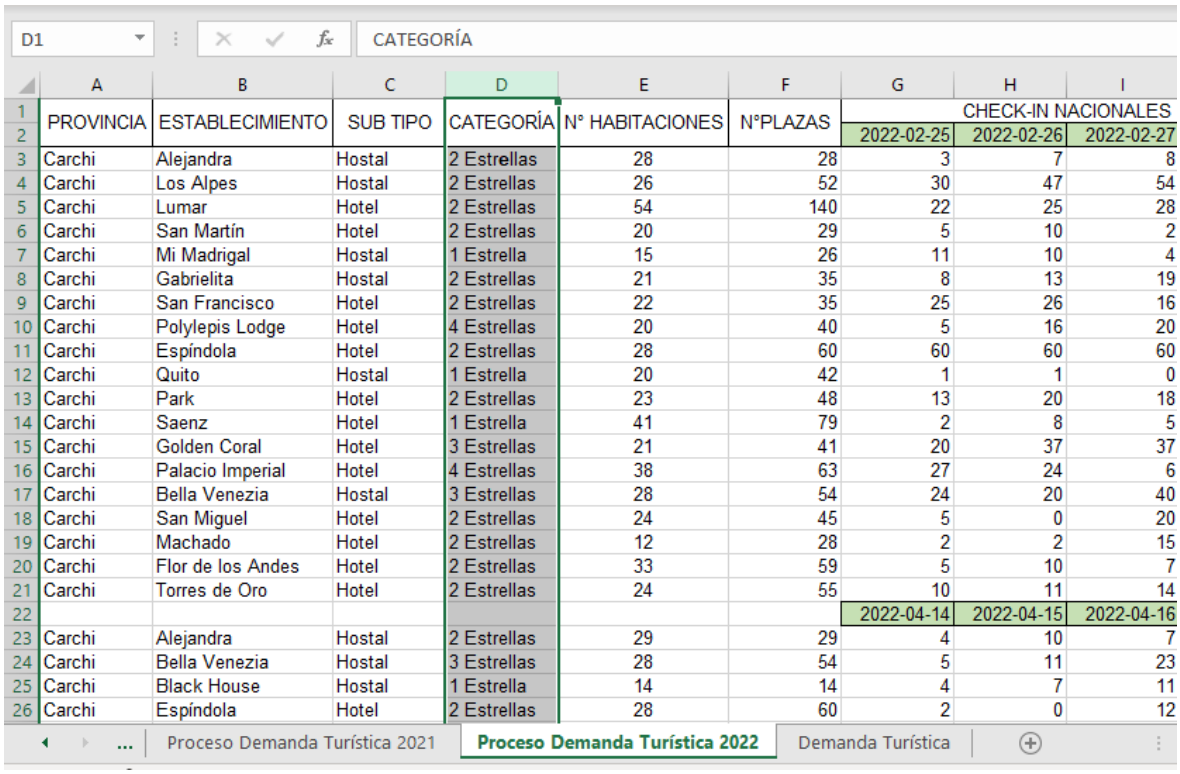
Mas allá al cambio de estructura de esta variable, no éxito la necesidad de crear ninguna nueva variable al igual que nuevos registros, debido a que la base de datos fue creada específicamente para el presente estudio y los propósitos que tiene.

- **Integrar Datos**

En los datos que se tiene no existió la necesidad de crear nuevas estructuras con relación a variables y atributos, ni la fusión de variables, para crear un nuevo campo, por la razón de que el subconjunto de datos seleccionado sobre la cual se va a aplicar minería de datos son los necesarios para obtener los resultados propuestos.

- **Formateo de Datos**

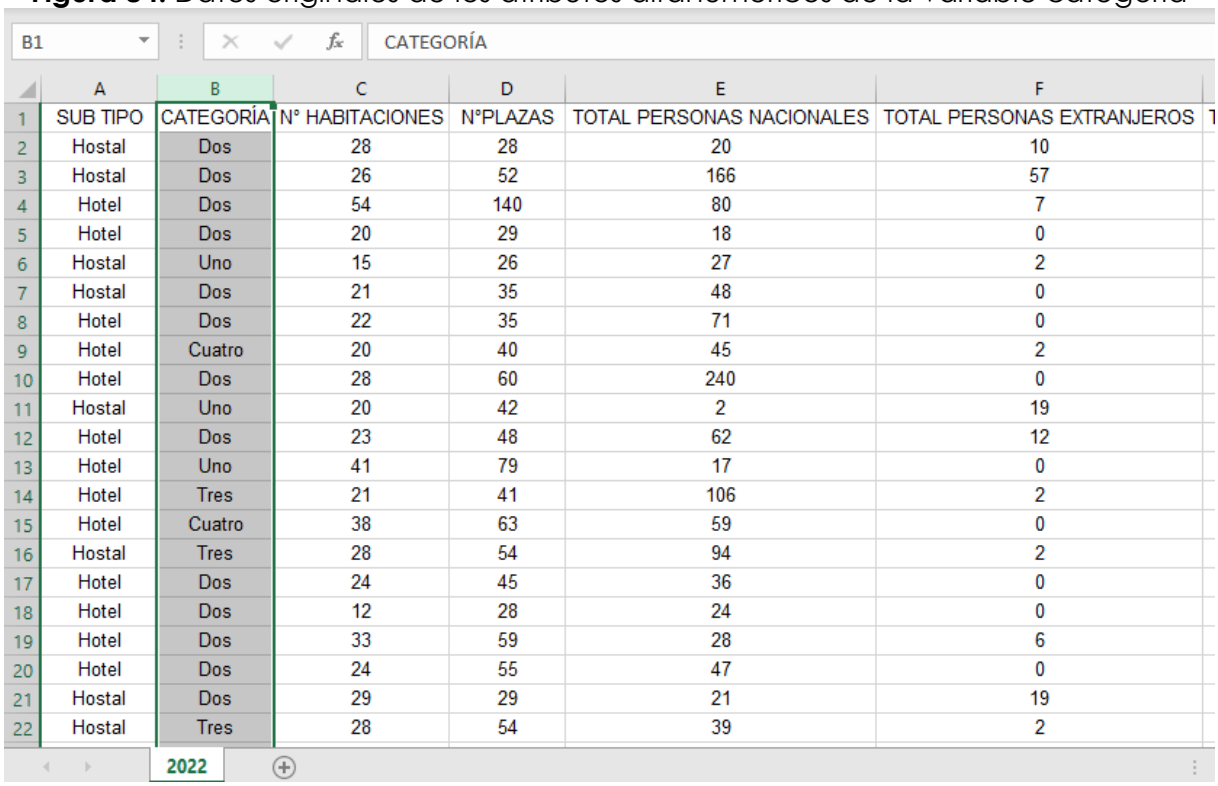
Esta tarea complementa a la tarea de estructuración de los datos, y como se mención hubo un cambio en los atributos de una variable llamada *categoría*, los datos de esta variable me muestran la clasificación de los hoteles y hostales que se encuentran en la provincia, lo datos se encuentran descritos con caracteres alfanuméricos y se los ha modificado a valores solamente numéricos por la exigencia de herramienta Knime y el análisis que se requiere con la técnica de minería de datos. Los datos de la variable quedaron de la siguiente manera:



	A	B	C	D	E	F	G	H	I
1	PROVINCIA	ESTABLECIMIENTO	SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	CHECK-IN NACIONALES		
2							2022-02-25	2022-02-26	2022-02-27
3	Carchi	Alejandra	Hostal	2 Estrellas	28	28	3	7	8
4	Carchi	Los Alpes	Hostal	2 Estrellas	26	52	30	47	54
5	Carchi	Lumar	Hotel	2 Estrellas	54	140	22	25	28
6	Carchi	San Martín	Hotel	2 Estrellas	20	29	5	10	2
7	Carchi	Mi Madrigal	Hostal	1 Estrella	15	26	11	10	4
8	Carchi	Gabrielita	Hostal	2 Estrellas	21	35	8	13	19
9	Carchi	San Francisco	Hotel	2 Estrellas	22	35	25	26	16
10	Carchi	Polylepis Lodge	Hotel	4 Estrellas	20	40	5	16	20
11	Carchi	Espíndola	Hotel	2 Estrellas	28	60	60	60	60
12	Carchi	Quito	Hostal	1 Estrella	20	42	1	1	0
13	Carchi	Park	Hotel	2 Estrellas	23	48	13	20	18
14	Carchi	Saenz	Hotel	1 Estrella	41	79	2	8	5
15	Carchi	Golden Coral	Hotel	3 Estrellas	21	41	20	37	37
16	Carchi	Palacio Imperial	Hotel	4 Estrellas	38	63	27	24	6
17	Carchi	Bella Venezia	Hostal	3 Estrellas	28	54	24	20	40
18	Carchi	San Miguel	Hotel	2 Estrellas	24	45	5	0	20
19	Carchi	Machado	Hotel	2 Estrellas	12	28	2	2	15
20	Carchi	Flor de los Andes	Hotel	2 Estrellas	33	59	5	10	7
21	Carchi	Torres de Oro	Hotel	2 Estrellas	24	55	10	11	14
22							2022-04-14	2022-04-15	2022-04-16
23	Carchi	Alejandra	Hostal	2 Estrellas	29	29	4	10	7
24	Carchi	Bella Venezia	Hostal	3 Estrellas	28	54	5	11	23
25	Carchi	Black House	Hostal	1 Estrella	14	14	4	7	11
26	Carchi	Espíndola	Hotel	2 Estrellas	28	60	2	0	12

Proceso Demanda Turística 2021 | Proceso Demanda Turística 2022 | Demanda Turística

Figura 84. Datos originales de los atributos alfanuméricos de la variable categoría



	A	B	C	D	E	F
1	SUB TIPO	CATEGORÍA	N° HABITACIONES	N°PLAZAS	TOTAL PERSONAS NACIONALES	TOTAL PERSONAS EXTRANJEROS
2	Hostal	Dos	28	28	20	10
3	Hostal	Dos	26	52	166	57
4	Hotel	Dos	54	140	80	7
5	Hotel	Dos	20	29	18	0
6	Hostal	Uno	15	26	27	2
7	Hostal	Dos	21	35	48	0
8	Hotel	Dos	22	35	71	0
9	Hotel	Cuatro	20	40	45	2
10	Hotel	Dos	28	60	240	0
11	Hostal	Uno	20	42	2	19
12	Hotel	Dos	23	48	62	12
13	Hotel	Uno	41	79	17	0
14	Hotel	Tres	21	41	106	2
15	Hotel	Cuatro	38	63	59	0
16	Hostal	Tres	28	54	94	2
17	Hotel	Dos	24	45	36	0
18	Hotel	Dos	12	28	24	0
19	Hotel	Dos	33	59	28	6
20	Hotel	Dos	24	55	47	0
21	Hostal	Dos	29	29	21	19
22	Hostal	Tres	28	54	39	2

2022

Figura 85. Formateo de los atributos de la variable categoría en caracteres

4.1.4.4. Modelado

Con la finalidad de darle continuidad al estudio y en particular a la metodología CRISP-DM, en esta fase se debe seleccionar la o las técnicas más apropiadas y

efectivas para cumplir con los objetivos planteados según los criterios enfocados a minería de datos. Luego de seleccionar las técnicas más optimas y realizado los planes de pruebas para los modelos elegidos, se iniciará con la aplicación de dichas técnicas sobre los datos que han sido previamente elegidos, donde nos permitirá obtener un modelo o varios modelos, que pasarán a hacer evaluados para verificar si ha cumplido o no los criterios de éxito que se establecieron.

- **Escoger la Técnica de Modelado**

Para la elegibilidad de la técnica de modelado como primera instancia debemos tomar en cuenta la herramienta de minería de datos que se va a utilizar. Knime Analytics es el software que se va a usar mediante el cual vamos a obtener el modelado, debido a que esta herramienta tiene la posibilidad de aplicar técnicas requeridas para darle cumplimiento a los objetivos con criterios de minería de datos descritos en la primera fase de la metodología.

De los modelos que tiene la posibilidad de generar Knime Analytics, el más adecuado para cumplir con los objetivos establecidos es la generación de una técnica de modelado de segmentación o clustering, debido a que el problema que se quiere resolver es la de mejorar procesos, las razones que existe para usar algoritmos de segmentación es que tiene la capacidad de detectar patrones inhabituales con el uso de técnicas como la agrupación en clústeres; y es lo que necesita la investigación debido a que toda la información que se tiene esta conjuntamente relacionada con la demanda turística de la provincia, y a través de la aplicación de este tipo de modelado obtener un conocimiento y comportamiento implícito en los datos de los procesos gasto y alojamiento turístico, y con esta nueva base de conocimiento tomar decisiones o acciones que le permita al Ministerio de Turismo mejorar la demanda turística de la provincia.

Como se ha mencionado se va a utilizar una técnica de modelado de aprendizaje no supervisado, denominada modelo de agrupamiento o clustering, este modelo consta de distintos algoritmos que permiten realizar el análisis de la base de información que se tiene acerca de los procesos de alojamiento y gasto turístico.

Ahora bien, se van a aplicar tres algoritmos que pertenecen a un modelado de agrupamiento o clustering con el fin de determinar, que algoritmo es el óptimo y aprueban los criterios de evaluación, que se establezcan, además de ello, que

modelo se enfoca más a los resultados que se quiera obtener. Los algoritmos que se va a aplicar en los datos de alojamiento y gasto turístico son:

- K-means
- K-medoids
- DBSCAN

- **Generar un Plan de Prueba**

Los procesos que se va a aplicar para probar la calidad y validez del modelo o modelos que se generen con el uso de algoritmos de clustering son pruebas estadísticas. Existen dos criterios propuestos para la evaluación de modelos de minería de datos orientados a la agrupación, mediante el cual se puede determinar un modelo óptimo. El primer criterio de evaluación es Cohesión (Sum of Squared Within - SSW), el segundo criterio de evaluación es Separación (Sum of Squared Between - SSB). Estos criterios de evaluación como se habían mencionado tienen bases estadísticas, los datos que necesitan cada criterio los proporcionan los parámetros de cada modelo de minería de datos que se genere. Con el fin de comprender más los criterios que se va a utilizar en los modelos a continuación se describe información que puede ser de utilidad:

Debemos tomar cuenta que el objetivo inicial de un algoritmo de clustering, es el de agrupar información u objetos similares en un mismo clúster, y distinta información colocarlos en diferentes clústeres, es por ellos que existen métricas de validación externas e internas, pero las que van a hacer aplicadas para la validación de los modelos de minería de datos del presente estudio son métricas internas, y se dividen en dos criterios (Guzmán, 2018).

- Cohesión (SSW): este criterio nos indica que el miembro de cada agrupación o clúster tienen que estar los más cerca posible a los otros miembros del mismo clúster, una medida común de la cohesión es la varianza, que debe minimizarse. Sum of Squared Within representa la medida interna que se usa para probar la Cohesión de los clústeres que el algoritmo de agrupamiento creo (Guzmán, 2018).

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

Nomenclatura:

$k = \text{número de clústeres}$

$x =$ un punto del clúster c_i

$m_i =$ el centroide del clúster C_i

- Separación (SSB): la principal característica que tiene es que sus clústeres deben estar ampliamente separados entre ellos. Este criterio lo que hace es medir la separación, existen distintos enfoques para medir dicha separación entre clústeres:
    - Enlace único: mide la distancia entre los miembros más cercanos de los grupos.
    - Enlace completo: mide la distancia entre los miembros más distantes.
    - Comparación de centroides: mide la distancia entre los centros de los grupos (Pastrán & Gongora, 2021).

  Sum of Squared Between es la medida interna de separación que utiliza para evaluar la distancia del clúster:

$$SSB = \sum_{j=1}^{k} n_j dist^2(c_j - \bar{x})$$

Nomenclatura:

$k =$ número de clústeres

$n_j =$ número de elementos en el clúster j

$c_j =$ el centroide del clúster j

$x =$ media del data set

Otro criterio de para probar los modelos es el Coeficiente de Silhouette al igual que los anteriores son métricas internas, y la razón es que combina las ideas de Cohesión y Separación, pero se las realiza para puntos individuales, y cuando se trata de sus clústeres se calcula el promedio y el promedio general del clustering. El coeficiente para un punto individual está definido como:

$$s(x) = \frac{b(x) - a(x)}{max\{a(x), b(x)\}}$$

Cabe mencionar que el valor de $s(x)$ puede variar entre $1\ y\ -1$.

- $-1 = mal\ agrupamiento$
- $0 = indiferente$
- $1 = bueno$

Nomenclatura:

- $a =$ distancia promedio de i a los puntos de su clúster
- $b =$ min (distancia promedio de i a puntos de otro clúster)

Para un plan de pruebas la metodología CRISP-DM sugiere que cuando se trata de modelos basados en técnicas de modelados de agrupamiento como es el de clustering es recomendable dividir los datos en dos conjuntos antes de crear el modelo; en el primer conjunto de datos se debe generar el modelo, comúnmente se lo suele llamar entrenamiento, en el segundo conjunto de datos se aplican la pruebas y se evalúa el modelo, a este conjunto de datos se los llama prueba o evaluación.

- **Construir el Modelo**

En esta sección de la fase de modelado se aplica el modelo elegido para este caso una técnica de modelado de segmentación va a ser aplicado directamente sobre el primer conjunto de datos. Se explicarán los parámetros establecidos en el software Knime para cada uno de los modelos que se genere, así como la salida de resultados que genere el modelo.

Para iniciar con la construcción del modelado de minería de datos debemos tomar en cuenta lo que queremos obtener con la aplicación de minería de datos, es decir tener claro los objetivos formados con criterios de minería de datos formados en las primeras fases de la metodología CRISP-DM. Centrándonos en el software de minería de datos que se va a aplicar directamente en los datos que han sido preparados previamente. Debemos generar cuatro nodos llamados Excel Reader, que va a permitir leer y cargar todos los datos que corresponden a los procesos de alojamiento y gasto turístico de la provincia. A cada nodo se le adaptara los parámetros que se han considerado pertinente para leer los datos.

Inicialización de lectura de datos del año 2019

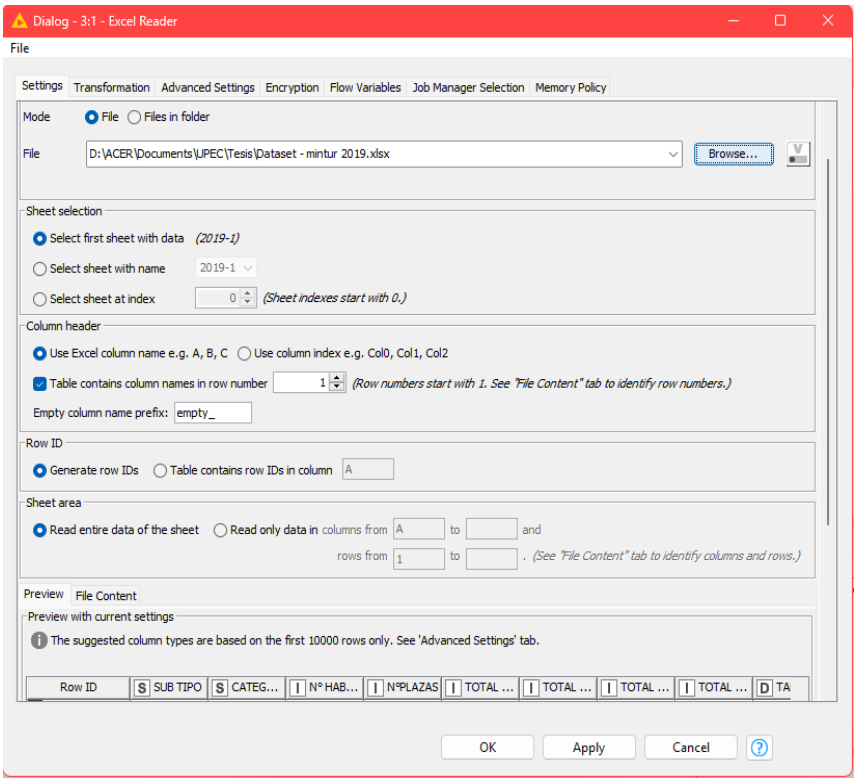
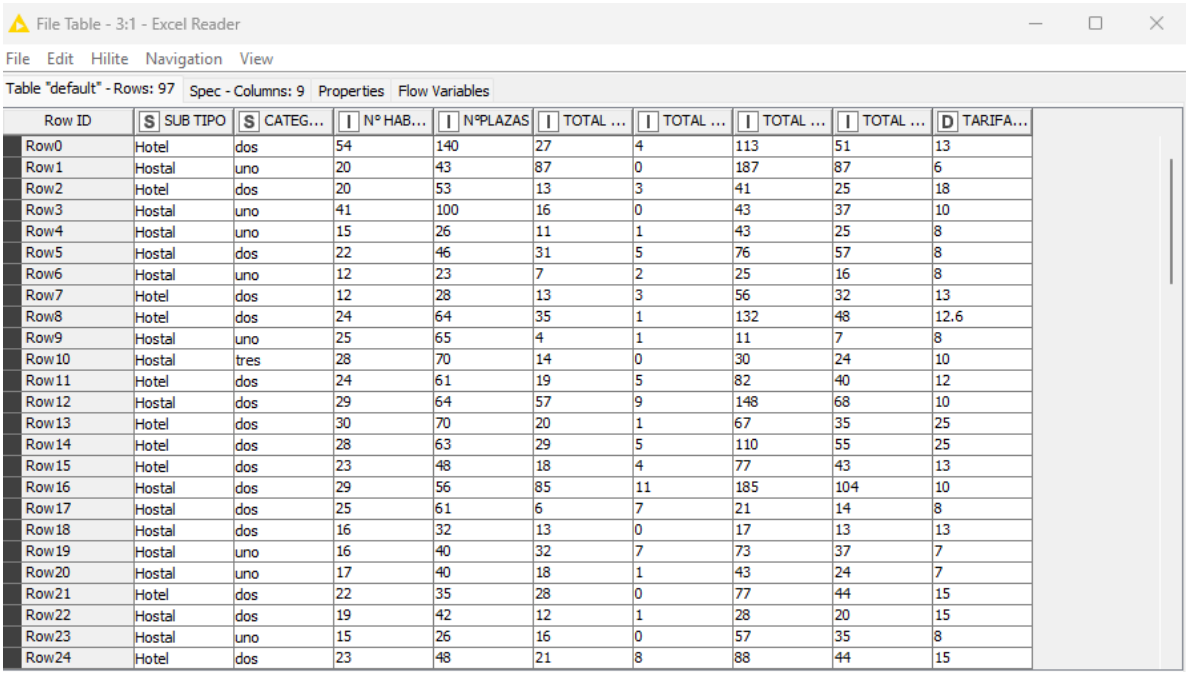


Figura 86. Configuración del nodo para leer datos del año 2019



Row ID	SUB TIPO	CATEG...	N° HAB...	NºPLAZAS	TOTAL ...	TOTAL ...	TOTAL ...	TOTAL ...	TARIFA...
Row0	Hotel	dos	54	140	27	4	113	51	13
Row1	Hostal	uno	20	43	87	0	187	87	6
Row2	Hotel	dos	20	53	13	3	41	25	18
Row3	Hostal	uno	41	100	16	0	43	37	10
Row4	Hostal	uno	15	26	11	1	43	25	8
Row5	Hostal	dos	22	46	31	5	76	57	8
Row6	Hostal	uno	12	23	7	2	25	16	8
Row7	Hotel	dos	12	28	13	3	56	32	13
Row8	Hotel	dos	24	64	35	1	132	48	12.6
Row9	Hostal	uno	25	65	4	1	11	7	8
Row10	Hostal	tres	28	70	14	0	30	24	10
Row11	Hotel	dos	24	61	19	5	82	40	12
Row12	Hostal	dos	29	64	57	9	148	68	10
Row13	Hotel	dos	30	70	20	1	67	35	25
Row14	Hotel	dos	28	63	29	5	110	55	25
Row15	Hotel	dos	23	48	18	4	77	43	13
Row16	Hostal	dos	29	56	85	11	185	104	10
Row17	Hostal	dos	25	61	6	7	21	14	8
Row18	Hostal	dos	16	32	13	0	17	13	13
Row19	Hostal	uno	16	40	32	7	73	37	7
Row20	Hostal	uno	17	40	18	1	43	24	7
Row21	Hotel	dos	22	35	28	0	77	44	15
Row22	Hostal	dos	19	42	12	1	28	20	15
Row23	Hostal	uno	15	26	16	0	57	35	8
Row24	Hotel	dos	23	48	21	8	88	44	15

Figura 87. Datos del año 2019 en Knime Analytics

Inicialización de lectura de datos del año 2020

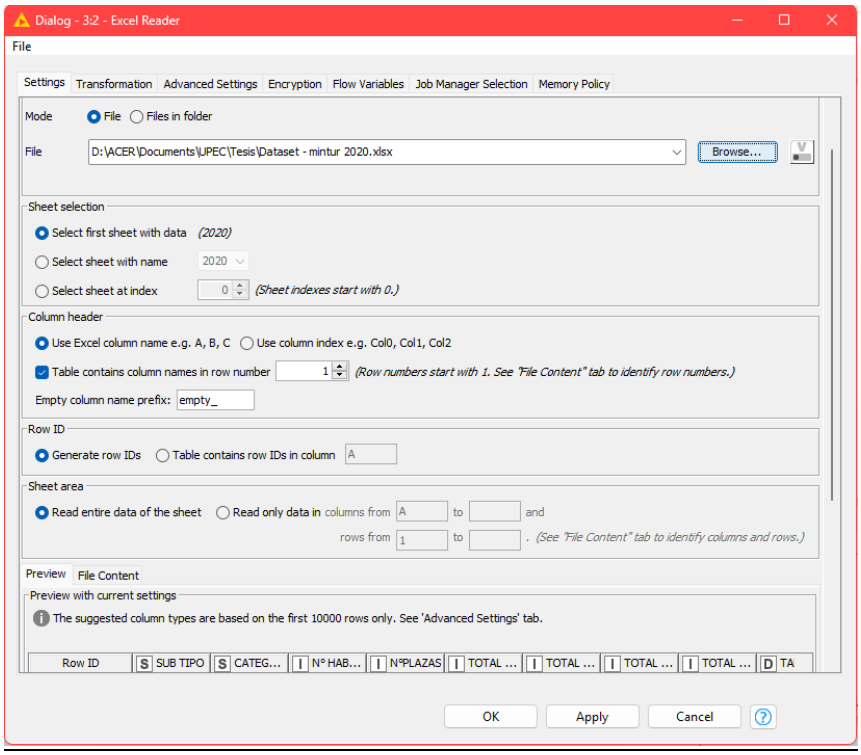
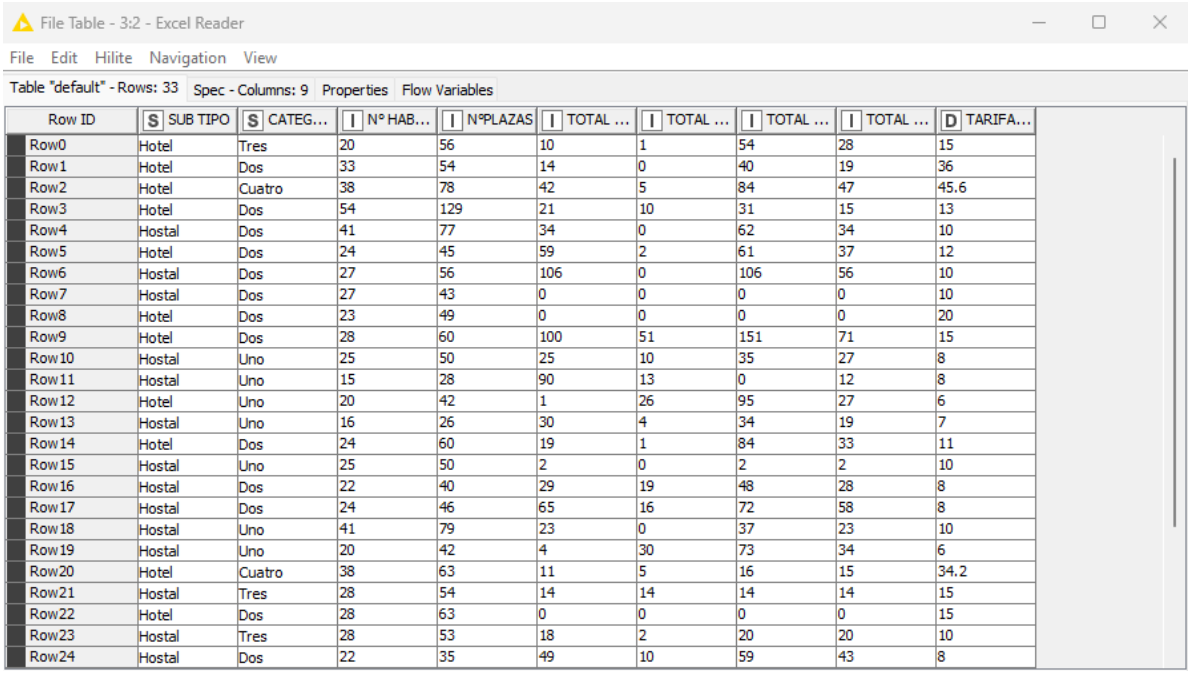


Figura 88. Configuración del nodo para leer datos del año 2020



Row ID	SUB TIPO	CATEG...	N° HAB...	NºPLAZAS	TOTAL ...	TOTAL ...	TOTAL ...	TOTAL ...	TARIFA...
Row0	Hotel	Tres	20	56	10	1	54	28	15
Row1	Hotel	Dos	33	54	14	0	40	19	36
Row2	Hotel	Cuatro	38	78	42	5	84	47	45.6
Row3	Hotel	Dos	54	129	21	10	31	15	13
Row4	Hostal	Dos	41	77	34	0	62	34	10
Row5	Hotel	Dos	24	45	59	2	61	37	12
Row6	Hostal	Dos	27	56	106	0	106	56	10
Row7	Hostal	Dos	27	43	0	0	0	0	10
Row8	Hotel	Dos	23	49	0	0	0	0	20
Row9	Hotel	Dos	28	60	100	51	151	71	15
Row10	Hostal	Uno	25	50	25	10	35	27	8
Row11	Hostal	Uno	15	28	90	13	0	12	8
Row12	Hotel	Uno	20	42	1	26	95	27	6
Row13	Hostal	Uno	16	26	30	4	34	19	7
Row14	Hotel	Dos	24	60	19	1	84	33	11
Row15	Hostal	Uno	25	50	2	0	2	2	10
Row16	Hostal	Dos	22	40	29	19	48	28	8
Row17	Hostal	Dos	24	46	65	16	72	58	8
Row18	Hostal	Uno	41	79	23	0	37	23	10
Row19	Hostal	Uno	20	42	4	30	73	34	6
Row20	Hotel	Cuatro	38	63	11	5	16	15	34.2
Row21	Hostal	Tres	28	54	14	14	14	14	15
Row22	Hotel	Dos	28	63	0	0	0	0	15
Row23	Hostal	Tres	28	53	18	2	20	20	10
Row24	Hostal	Dos	22	35	49	10	59	43	8

Figura 89. Datos del año 2020 en Knime Analytics

Inicialización de lectura de datos del año 2021

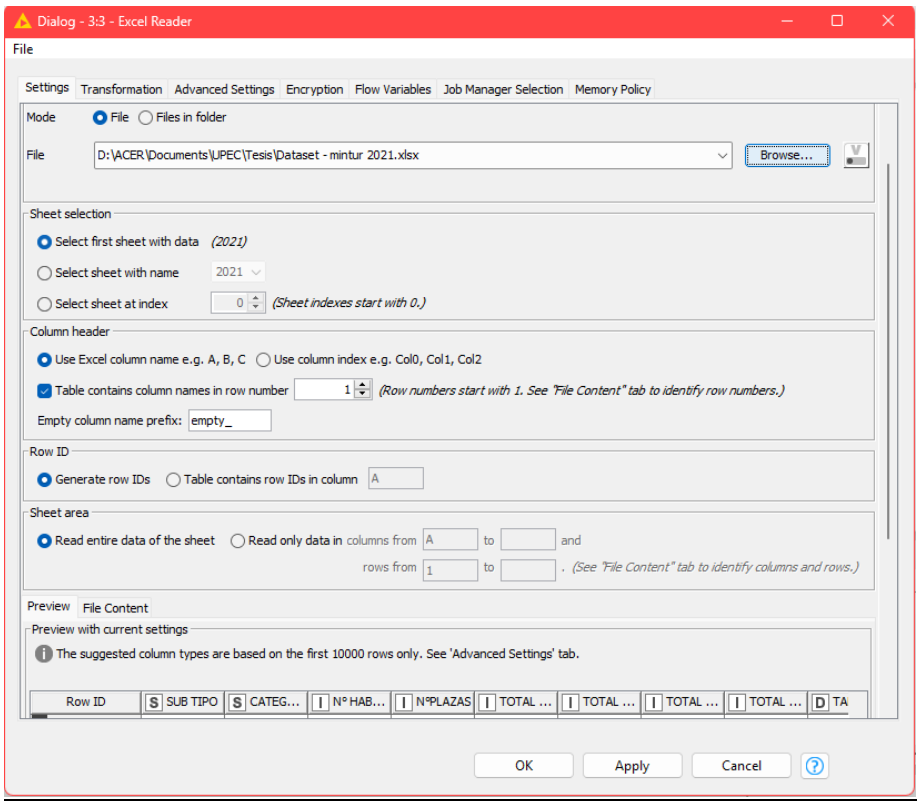
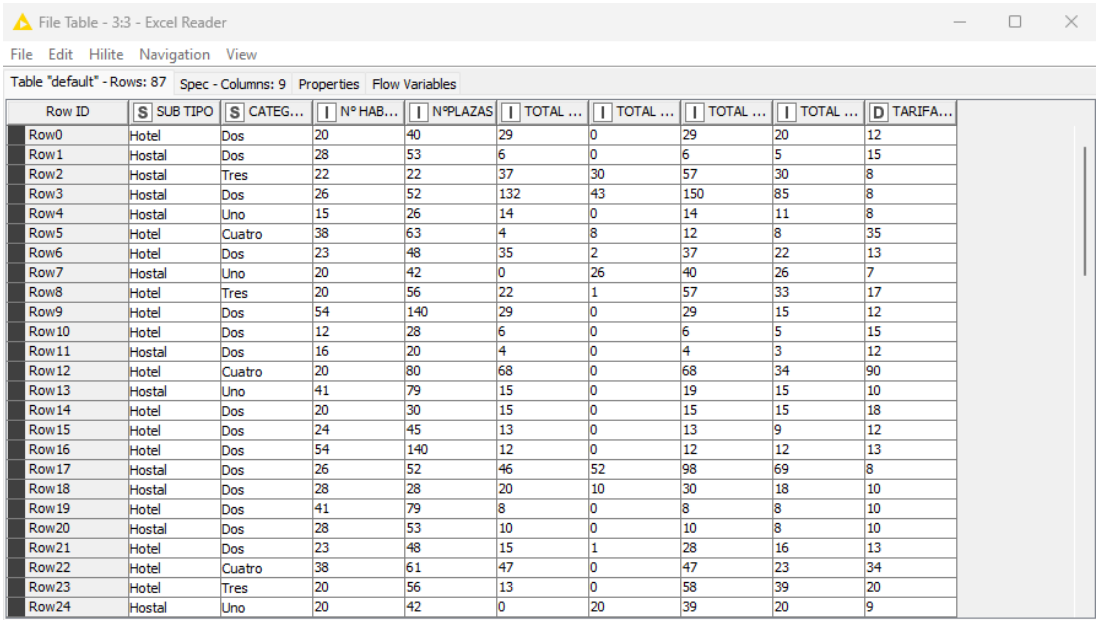


Figura 90. Configuración del nodo para leer datos del año 2021



Row ID	SUB TIPO	CATEG...	N° HAB...	NºPLAZAS	TOTAL ...	TOTAL ...	TOTAL ...	TOTAL ...	TARIFA...
Row0	Hotel	Dos	20	40	29	0	29	20	12
Row1	Hostal	Dos	28	53	6	0	6	5	15
Row2	Hostal	Tres	22	22	37	30	57	30	8
Row3	Hostal	Dos	26	52	132	43	150	85	8
Row4	Hostal	Uno	15	26	14	0	14	11	8
Row5	Hotel	Cuatro	38	63	4	8	12	8	35
Row6	Hotel	Dos	23	48	35	2	37	22	13
Row7	Hostal	Uno	20	42	0	26	40	26	7
Row8	Hotel	Tres	20	56	22	1	57	33	17
Row9	Hotel	Dos	54	140	29	0	29	15	12
Row10	Hotel	Dos	12	28	6	0	6	5	15
Row11	Hostal	Dos	16	20	4	0	4	3	12
Row12	Hotel	Cuatro	20	80	68	0	68	34	90
Row13	Hostal	Uno	41	79	15	0	19	15	10
Row14	Hotel	Dos	20	30	15	0	15	15	18
Row15	Hotel	Dos	24	45	13	0	13	9	12
Row16	Hotel	Dos	54	140	12	0	12	12	13
Row17	Hostal	Dos	26	52	46	52	98	69	8
Row18	Hostal	Dos	28	28	20	10	30	18	10
Row19	Hotel	Dos	41	79	8	0	8	8	10
Row20	Hostal	Dos	28	53	10	0	10	8	10
Row21	Hotel	Dos	23	48	15	1	28	16	13
Row22	Hotel	Cuatro	38	61	47	0	47	23	34
Row23	Hotel	Tres	20	56	13	0	58	39	20
Row24	Hostal	Uno	20	42	0	20	39	20	9

Figura 91. Datos del año 2021 en Knime Analytics

Inicialización de lectura de datos del año 2022

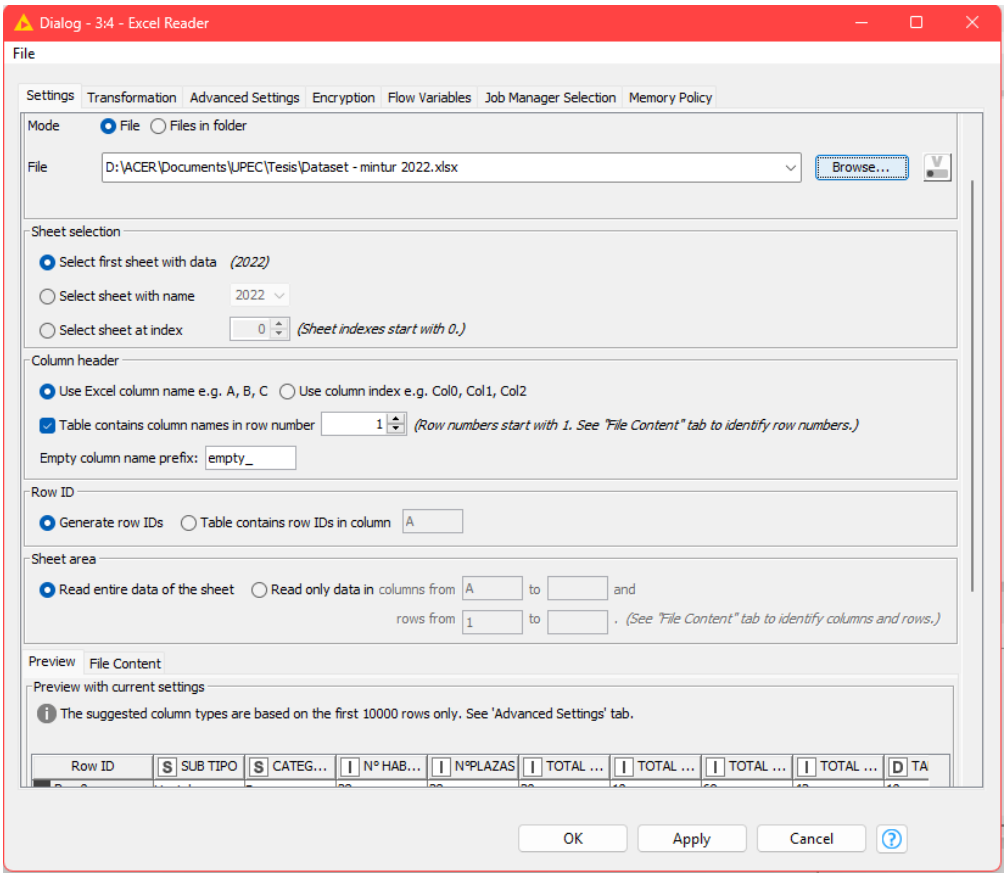
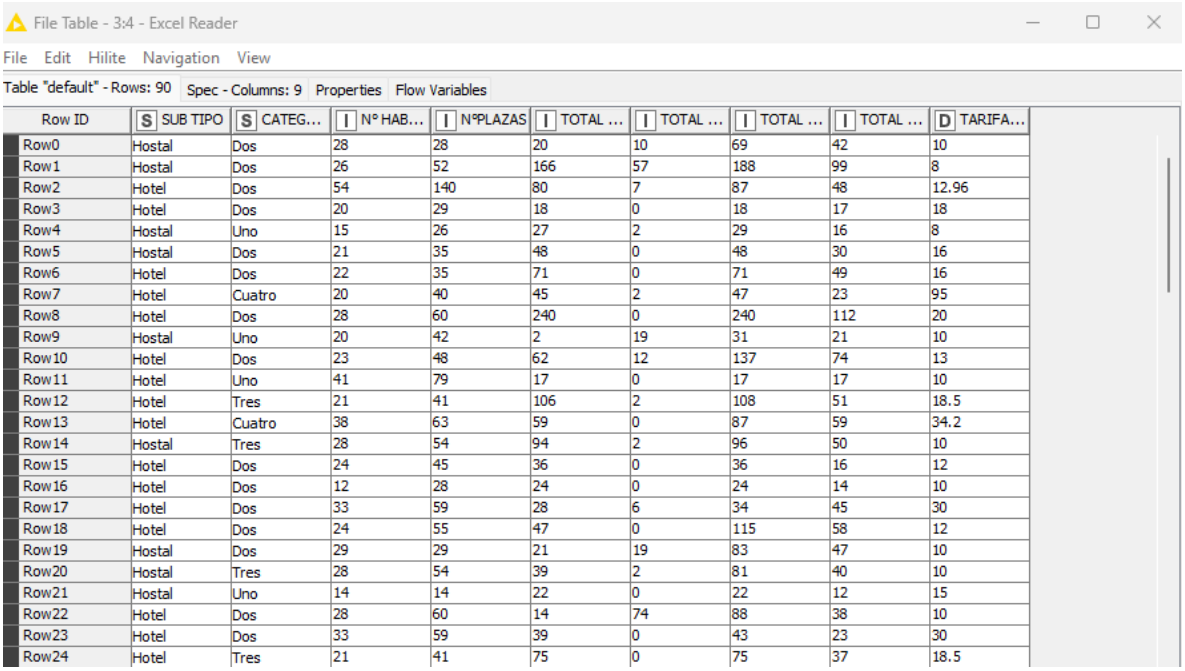


Figura 92. Configuración del nodo para leer datos del año 2022



Row ID	SUB TIPO	CATEG...	N° HAB...	NºPLAZAS	TOTAL ...	TOTAL ...	TOTAL ...	TOTAL ...	TARIFA...
Row0	Hostal	Dos	28	28	20	10	69	42	10
Row1	Hostal	Dos	26	52	166	57	188	99	8
Row2	Hotel	Dos	54	140	80	7	87	48	12.96
Row3	Hotel	Dos	20	29	18	0	18	17	18
Row4	Hostal	Uno	15	26	27	2	29	16	8
Row5	Hostal	Dos	21	35	48	0	48	30	16
Row6	Hotel	Dos	22	35	71	0	71	49	16
Row7	Hotel	Cuatro	20	40	45	2	47	23	95
Row8	Hotel	Dos	28	60	240	0	240	112	20
Row9	Hostal	Uno	20	42	2	19	31	21	10
Row10	Hotel	Dos	23	48	62	12	137	74	13
Row11	Hotel	Uno	41	79	17	0	17	17	10
Row12	Hotel	Tres	21	41	106	2	108	51	18.5
Row13	Hotel	Cuatro	38	63	59	0	87	59	34.2
Row14	Hostal	Tres	28	54	94	2	96	50	10
Row15	Hotel	Dos	24	45	36	0	36	16	12
Row16	Hotel	Dos	12	28	24	0	24	14	10
Row17	Hotel	Dos	33	59	28	6	34	45	30
Row18	Hotel	Dos	24	55	47	0	115	58	12
Row19	Hostal	Dos	29	29	21	19	83	47	10
Row20	Hostal	Tres	28	54	39	2	81	40	10
Row21	Hostal	Uno	14	14	22	0	22	12	15
Row22	Hotel	Dos	28	60	14	74	88	38	10
Row23	Hotel	Dos	33	59	39	0	43	23	30
Row24	Hotel	Tres	21	41	75	0	75	37	18.5

Figura 93. Datos del año 2022 en Knime Analytics

A continuación, se muestra el modelo general de los nodos que inician el modelo de los datos de los procesos de alojamiento y demanda turística de la provincia:

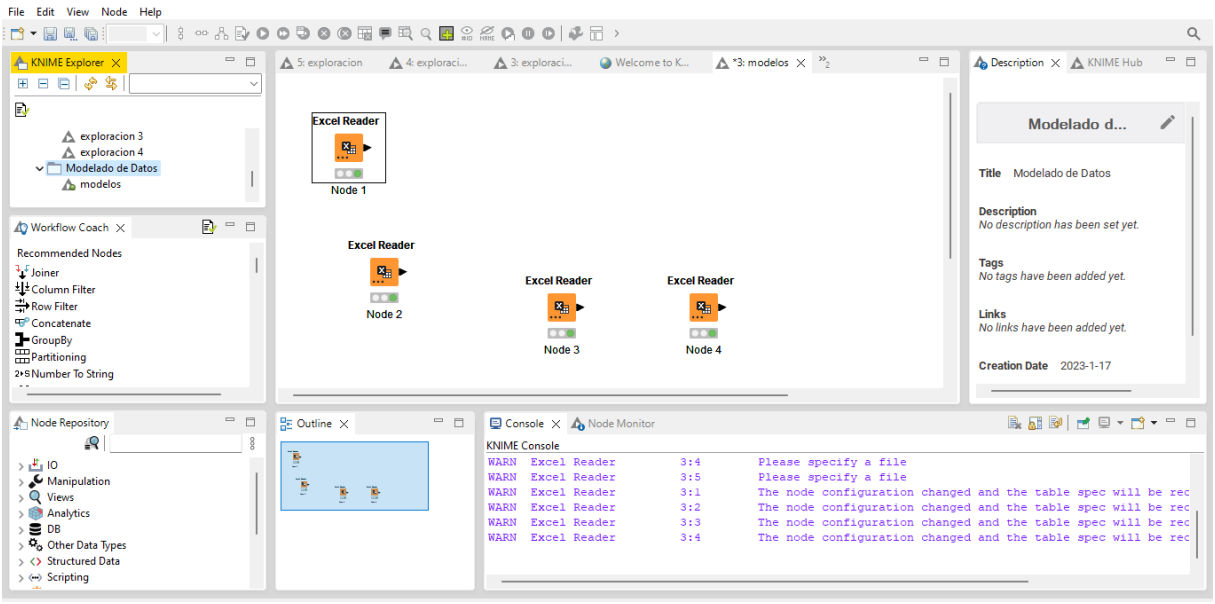


Figura 94 Nodos de lectura de los datos

Por otra parte, los modelos que se van a generar serán respectivamente a los registros históricos donde el ministerio de turismo ha realizado los procesos que determinan la demanda turística de la provincia, siendo específicos se genera un modelo en referencia a cada año donde se ha aplicado procesos de alojamiento y gasto turístico iniciando desde el 2019, ya que es el año donde se tiene los primeros registros de los procesos que determinan la demanda turística de la provincia del Carchi.

Para cada modelo que se va a generar se debe tomar en cuenta que es un tipo de técnica de Segmentación o Clustering, y se optado por usar un nodo como lo determina Knime o también llamado algoritmo K-means. K-means es un algoritmo de clasificación no supervisado, y el objetivo que tiene es agrupa x objetos o datos, basándose en las características que tengan los objetos, en este caso los datos. Ahora bien, cada modelo utilizara el mismo nodo, pero los parámetros para los datos que representan cada año de los procesos serán diferente y aleatorios, para que análisis que se realice sea objetivo, la modificación de los parámetros se la realiza en la herramienta de minería de datos Knime Analytics.

Para establecer la configuración del algoritmo de K-means lo primero debemos tomar en cuenta las variables y los objetos que tiene el conjunto de datos, para este caso datos del año 2019 hasta el año 2022, luego debemos establecer el número de clústeres. El algoritmo K-means se basa en la idea de establecer centroides, nodos ficticios, a partir de los cuales va a medir la distancia de todos los demás nodos, y los

va a agrupar conforme se encuentren más cerca, a estos nodos se los denomina centroides y tienen que inicializarse.

Para inicializar estos nodos existen diversas formas, y la que se opto es usar una inicialización aleatoria, y es por esta razón que cada modelo será distinto a otro, ya que tendrá su propia base de inicialización, las iteraciones que se va a aplicar son 99, representa el número de veces que va a repetir el algoritmo para cada uno de los puntos.

Ajuste de Parámetros Modelo 1

Como se había mencionado anteriormente los centroides se inicializarán de forma aleatoria, y se estableció para este modelo se generen 4 clúster.

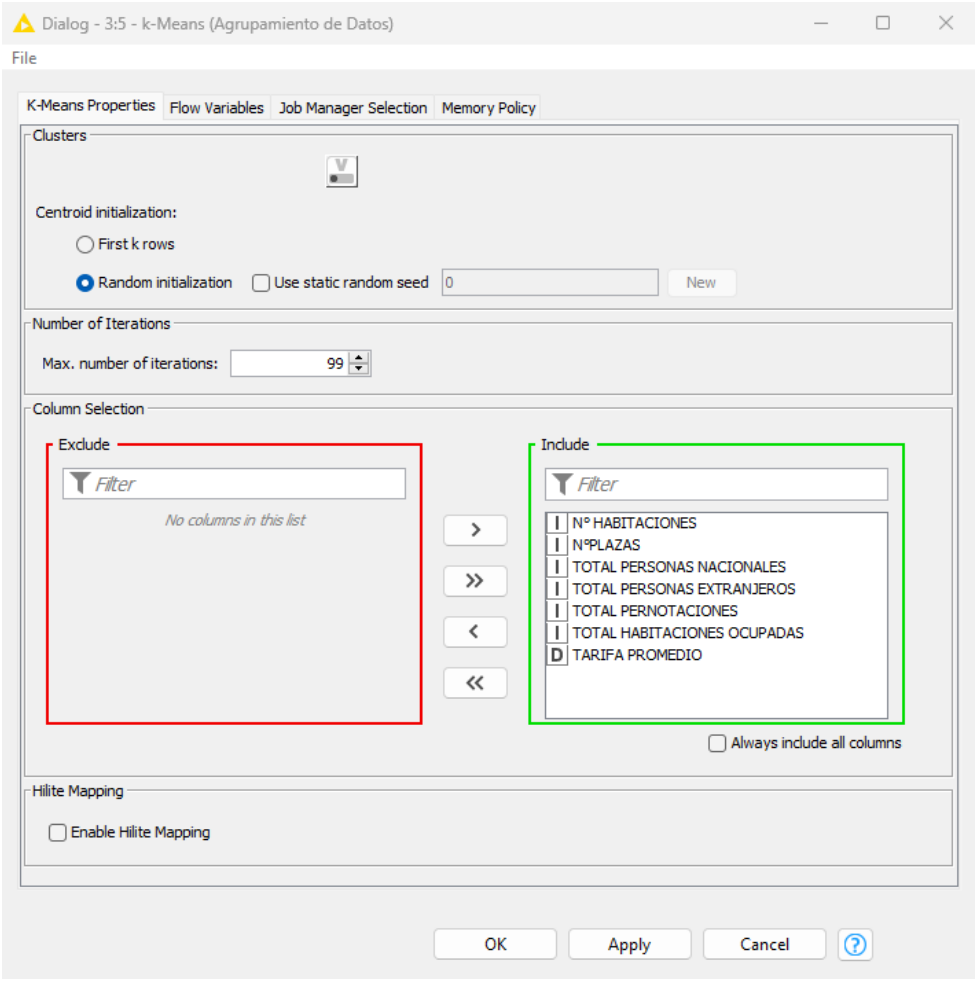


Figura 95. Parámetros para el modelo 1

Los resultados de esta configuración nos muestran se han generado 4 clústeres, en el clúster 0, a enracimado 31 de los puntos, en el clúster 1, 25 de los puntos, en el clúster 2, 24 de los puntos y en el clúster 3, 17 de los puntos. Cada una de ramas indica la

posición de los centroides, tomando en cuenta las escalas que se estableció de 0 a 100.

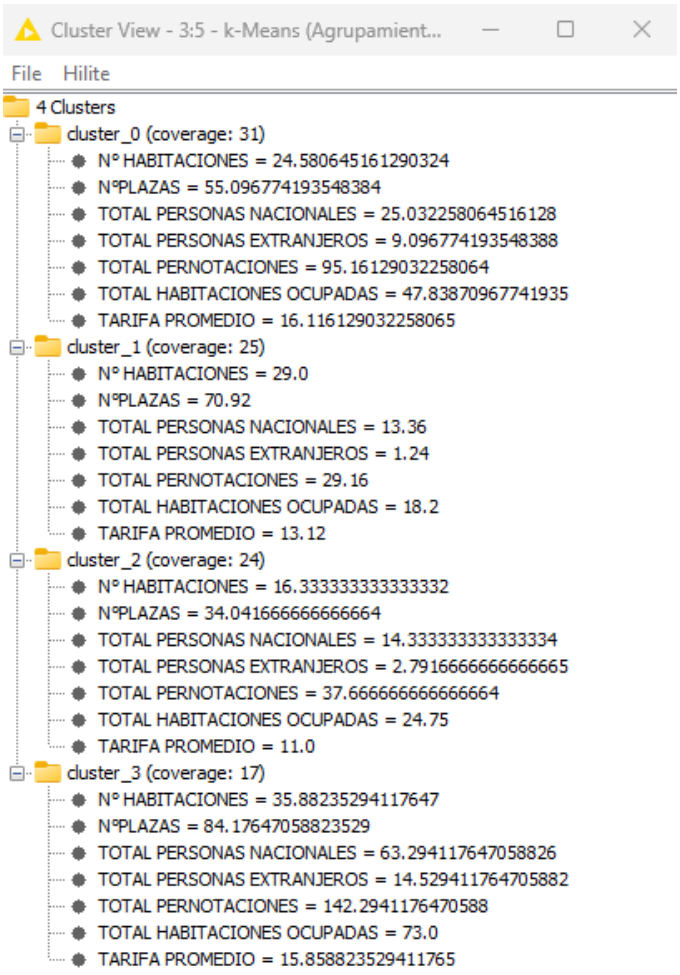


Figura 96. Clústeres generados para el modelo 1

Ajuste de Parámetros Modelo 2

Como en el anterior modelo los centroides se inicializarán de forma aleatoria, y se generarán 4 clústeres y sus escalas irán de 0 a 100.

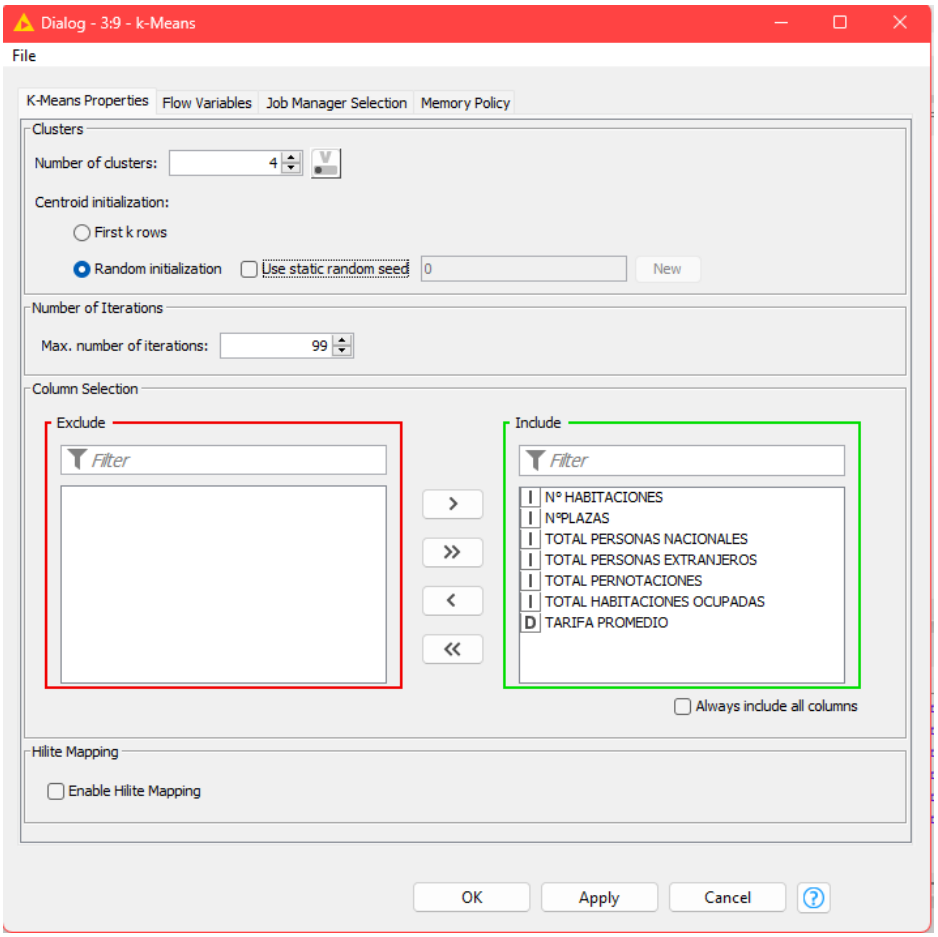


Figura 97. Parámetros para el modelo 2

Los resultados de esta configuración nos muestran se han generado 4 clústeres, en el clúster 0, a enracimado 10 de los puntos, en el clúster 1, 5 de los puntos, en el clúster 2, 11 de los puntos y en el clúster 3, 7 de los puntos. Cada una de ramas indica la posición de los centroides, tomando en cuenta las escalas que se estableció de 0 a 100.

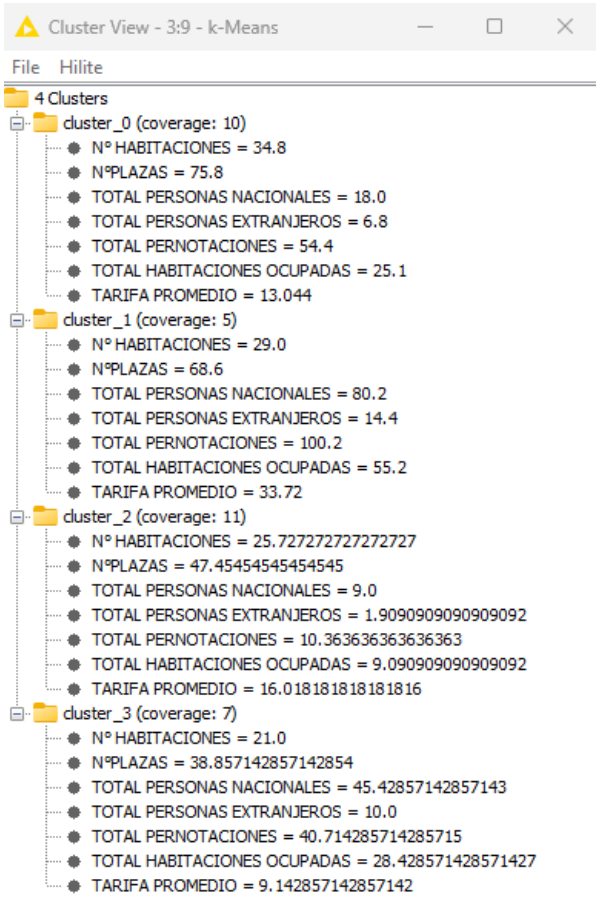


Figura 98. Clústeres generados para el modelo 2

Ajuste de Parámetros Modelo 3

Como en el anterior modelo los centroides se inicializarán de forma aleatoria, y se generarán 4 clústeres y sus escalas irán de 0 a 100.

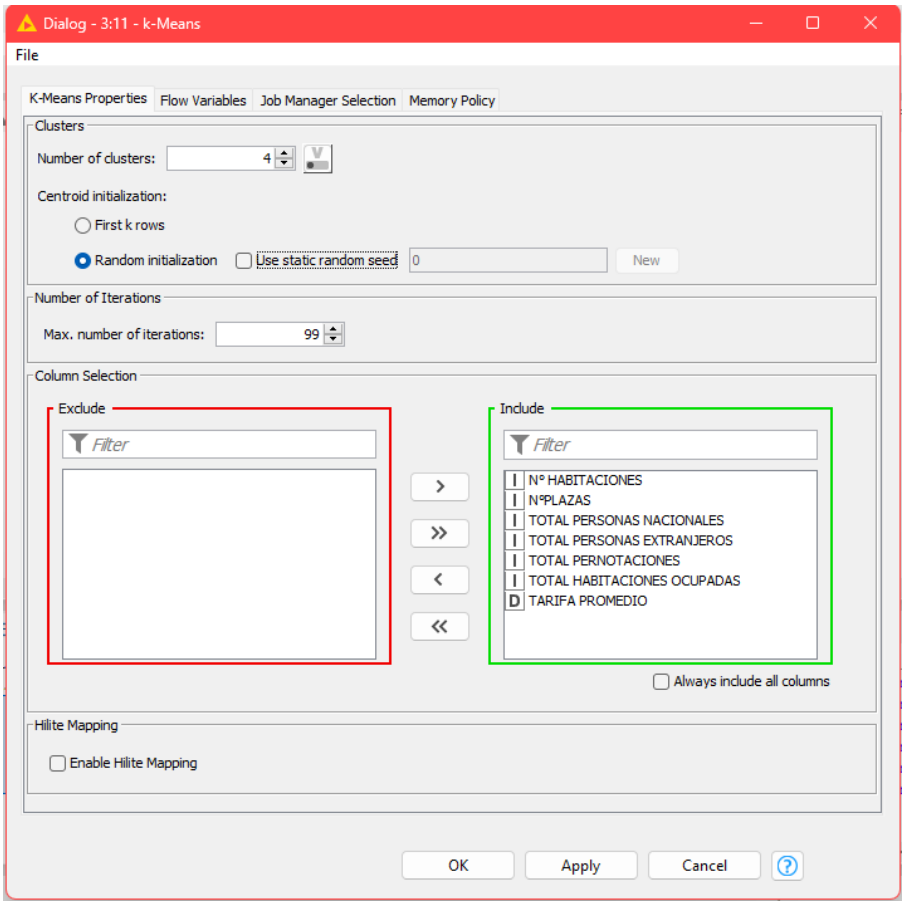


Figura 99. Parámetros para el modelo 3

Los resultados de esta configuración nos muestran se han generado 4 clústeres, en el clúster 0, a enracimado 39 de los puntos, en el clúster 1, 9 de los puntos, en el clúster 2, 17 de los puntos y en el clúster 3, 22 de los puntos. Cada una de ramas indica la posición de los centroides, tomando en cuenta las escalas que se estableció de 0 a 100.

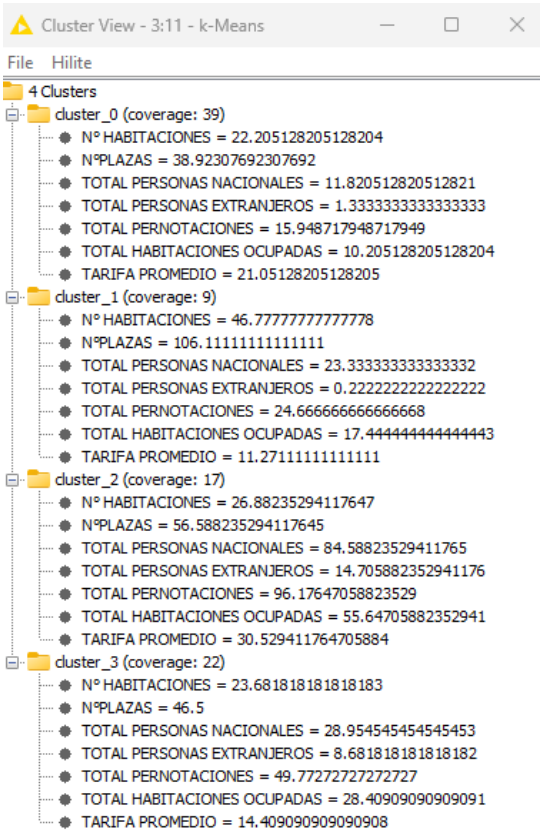


Figura 100. Clústeres generados para el modelo 3

Ajuste de Parámetros Modelo 4

Como en el anterior modelo los centroides se inicializarán de forma aleatoria, y se generarán 4 clústeres y sus escalas irán de 0 a 100.

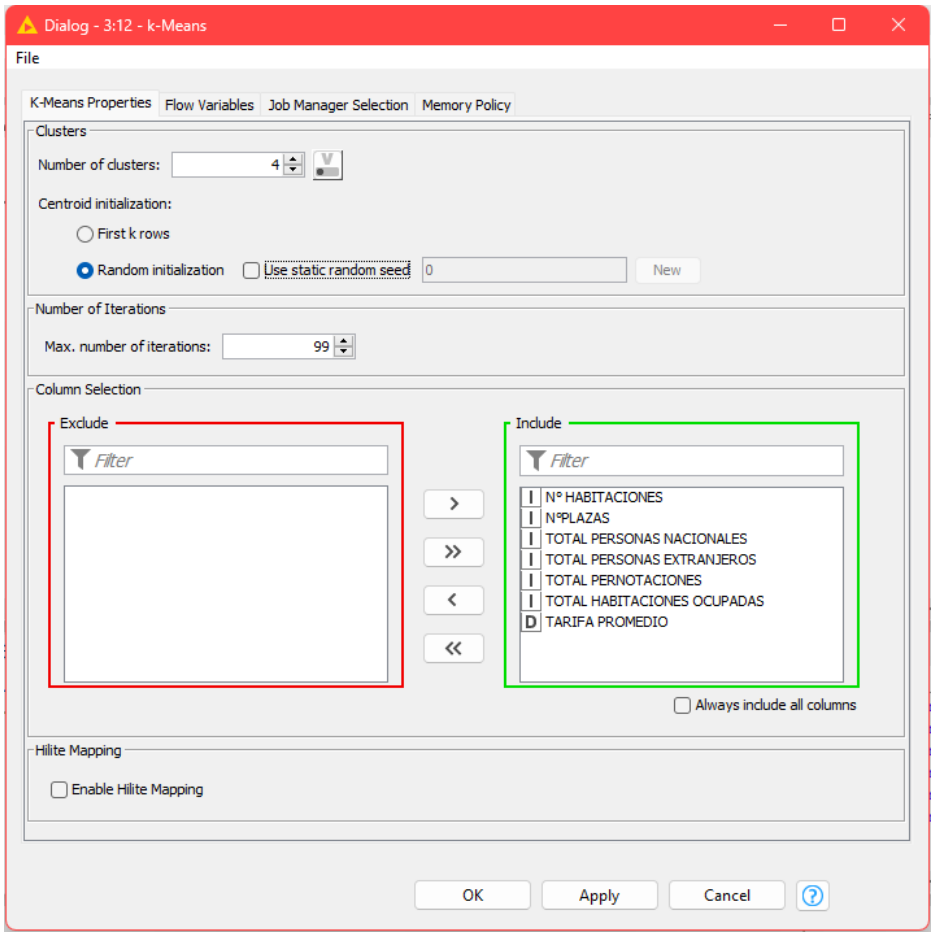


Figura 101. Parámetros para el modelo 4

Los resultados de esta configuración nos muestran se han generado 4 clústeres, en el clúster 0, a enracimado 19 de los puntos, en el clúster 1, 28 de los puntos, en el clúster 2, 13 de los puntos y en el clúster 3, 30 de los puntos. Cada una de ramas indica la posición de los centroides, tomando en cuenta las escalas que se estableció de 0 a 100.

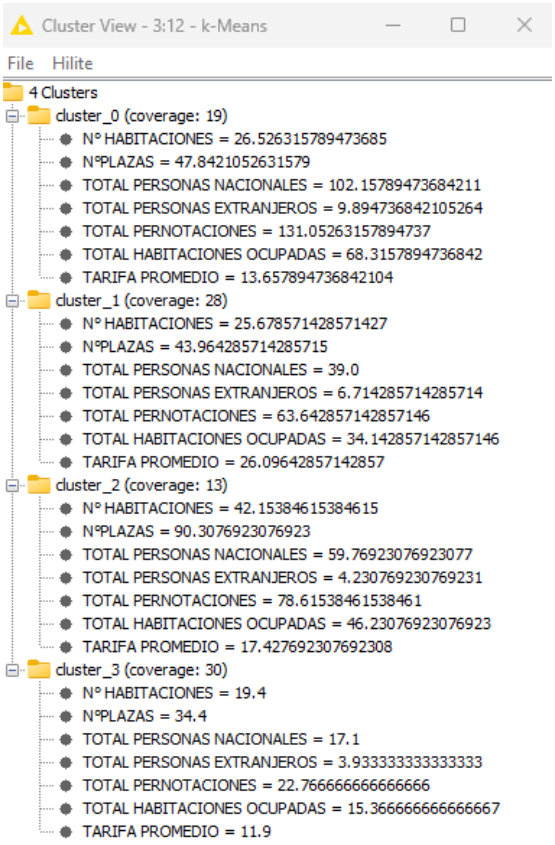


Figura 102. Clústeres generados para el modelo 4

Descripción del modelo Algoritmo K-means

Ahora, se va a describir brevemente los resultados de la ejecución, debido que más adelante en el apartado de evaluación se explicaran con más detalle, para cada modelo se aplicó el nodo K-medias, que representa una técnica de modelado de clustering. Basándonos en el coeficiente de Silhouette las medias de los resultados para cada clúster generado es la siguiente:

Modelo 1: datos de los procesos del año 2019.

Tabla 23. Media de Coeficiente de Silhouette - m1

Clúster Generados	Media de Coeficiente Silhouette
cluster_3	0,062
cluster_2	0,486
cluster_1	0,253
cluster_0	0,298
Media General del Coeficiente Silhouette	0,292

Modelo 2: datos de los procesos del año 2020

Tabla 24. Media de Coeficiente de Silhouette - m2

Clúster Generados	Media de Coeficiente Silhouette
cluster_0	0,035
cluster_1	0,055
cluster_3	0,252
cluster_2	0,52
Media General del Coeficiente Silhouette	0,246

Modelo 3: datos de los procesos del 2021

Tabla 25. Media de Coeficiente de Silhouette - m3

Clúster Generados	Media de Coeficiente Silhouette
cluster_3	0,23
cluster_0	0,356
cluster_2	0,136
cluster_1	0,323
Media General del Coeficiente Silhouette	0,278

Modelo 4: datos de los procesos del año 2022.

Tabla 26. Media de Coeficiente de Silhouette - m4

Clúster Generados	Media de Coeficiente Silhouette
cluster_1	0,104
cluster_0	0,289
cluster_2	0,188
cluster_3	0,562
Media General del Coeficiente Silhouette	0,308

Descripción del modelo Algoritmo DBSCAN

Los resultados del algoritmo DBSCAN muestran que no tiene la posibilidad de generar los clústeres adecuados, por lo cual no se puede generar conocimiento, la razón principal es la cantidad de datos que se está manejando, este tipo de algoritmo

necesita mayor cantidad de datos con mayor densidad. El algoritmo DBSCAN se lo aplico sobre los datos de los procesos de alojamiento y gasto turístico en el año 2019. A continuación, se muestra la configuración de parámetros y el único clúster generado.

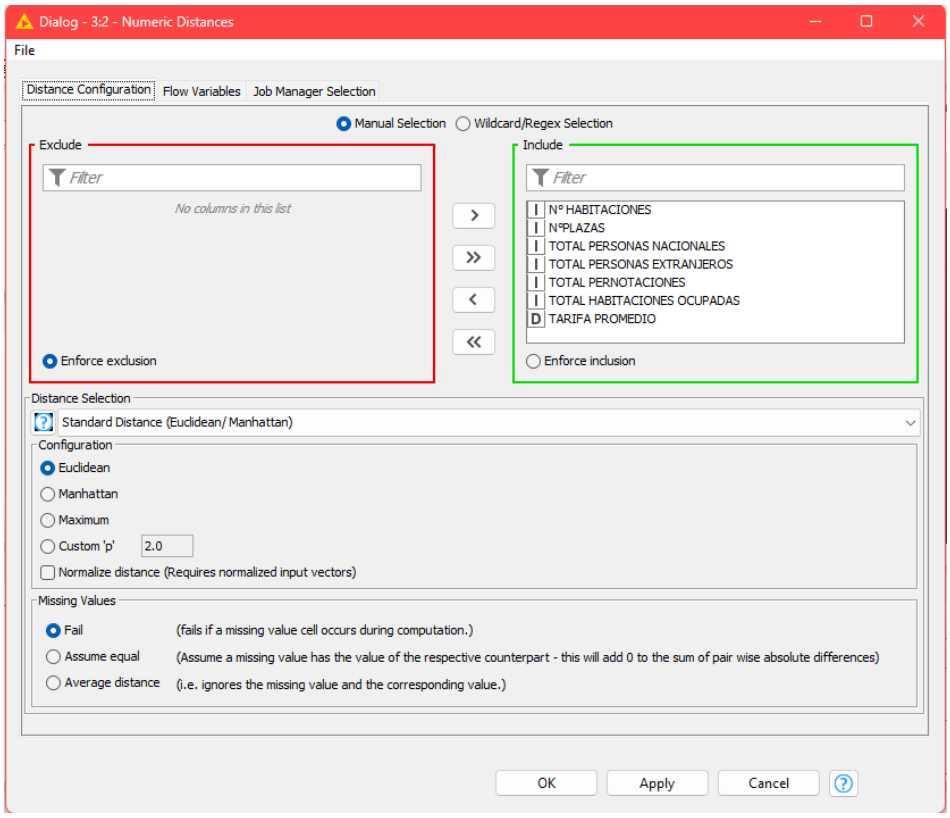


Figura 103. Parámetros de distancia para el algoritmo

A continuación, se muestra el valor de eps en el algoritmo, que indica lo cerca que deben estar los punto entre sí para ser considerados parte de un clúster. Lo que significa que, la distancia entre dos puntos es menor igual al valor de épsilon, estos puntos se los considera vecinos.

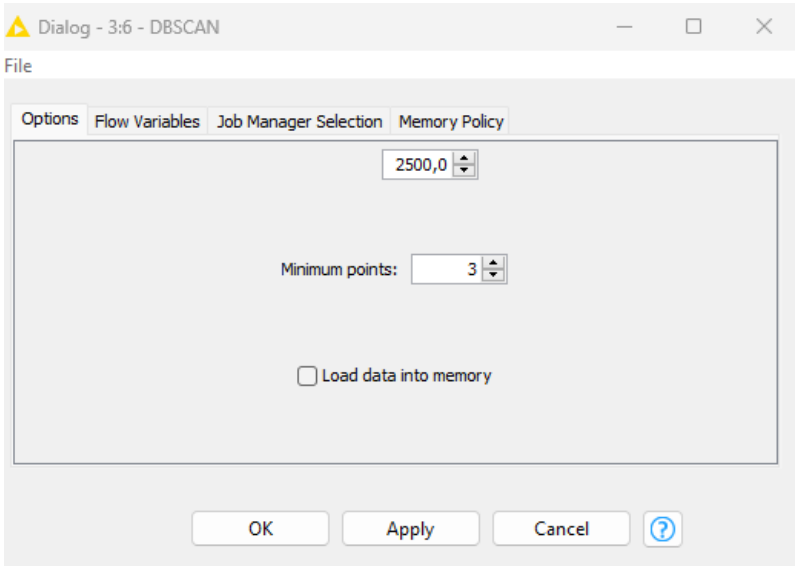


Figura 104. Valores de eps para el algoritmo DBSCAN

En la sección de anexos se puede evidenciar los resultados del algoritmo DBSCAN y el modelo realizado muestran solo un clúster generado, de los datos del proceso de alojamiento y gasto turístico, lo que indica que no se puede obtener ningún tipo de conocimiento útil que pueda ser utilizado para los procesos de control que emplea el ministerio de turismo con fin de determinar la demanda turística de la provincia del Carchi. La razón principal como se ha mencionado es principalmente la cantidad de datos que se está empleando, debido a que este algoritmo necesita gran cantidad de datos con mayor densidad, y es una de la roznes por las que este algoritmo pasa a la fase de evaluación del modelo.

Descripción del modelo Algoritmo K-medoids

El algoritmo K-medoids es muy similar al algoritmo de k-means y permitió el análisis de los datos generando 3 clústeres, y se realizó el análisis con el fin de comprobar la calidad y validez del modelo generado con los criterios previamente establecidos en el plan de prueba, a continuación, se muestra la tabla general de las medias de coeficiente de silhouette, donde se evidencia que este modelo no está apto y no pasa la prueba de validez para utilizarlo como modelo optimo.

Tabla 27. *Promedio del coeficiente de silhouette algoritmo K-medoids*

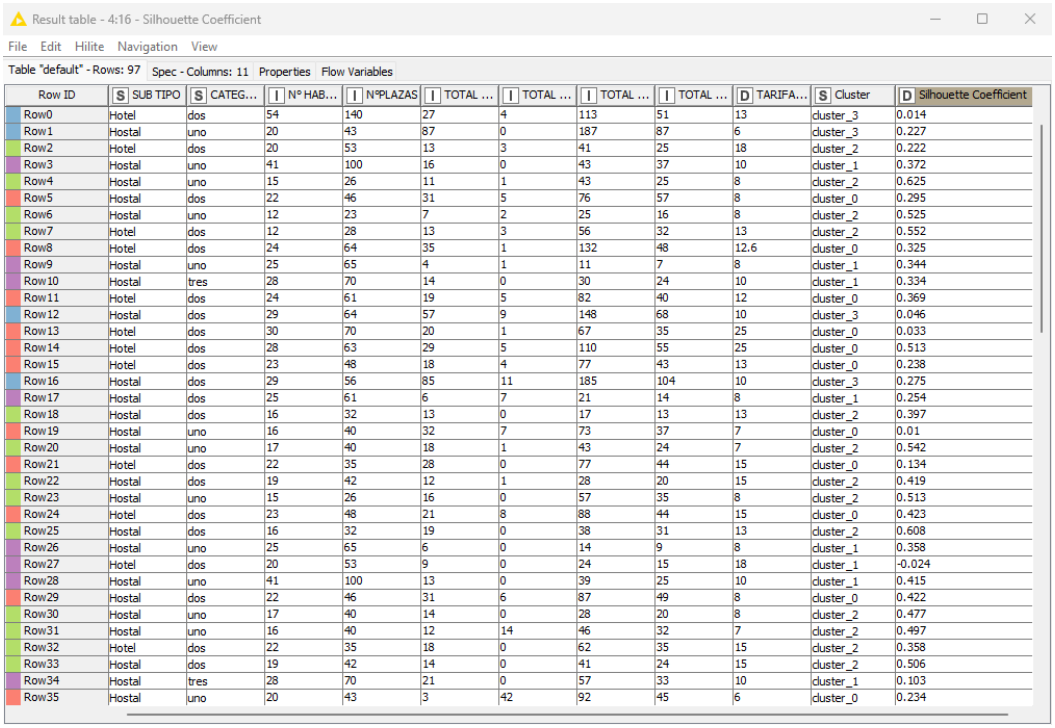
	Clúster 70	Clúster 12	Clúster 22
Modelo 1	-0,219	-0.136	-0.518
Modelo 2	-0.155	-0,088	-0424
Modelo 3	-0.042	-0.434	-0.315
Modelo 4	-0.106	-0.294	-0.533

- **Evaluar el Modelo**

Dentro de la fase 5 de la metodología CRISP-DM, se encarga de hacer una evaluación de los modelos que se generaros, esta sección también se evalúa, pero con criterios más orientados a los objetivos de minería de datos, mientras que en la fase de evaluación se orienta más por los objetivos con relación al negocio en este caso el Ministerio de Turismo, cabe mencionar que ambos criterios de evaluación se relacionan entre sí.

En términos de minería de datos, la mejor forma de evaluar la calidad, validez y efectividad de los modelos generados es utilizar y aplicar los indicadores que se establecieron en el plan pruebas de este documento, el parámetro que se va a utilizar para demostrar la validez del modelo es el Coeficiente de Silhouette, debemos tomar en cuenta que cuando se trata validación de modelos de clustering no existe un criterio totalmente definido, a diferencia con algoritmos que son supervisados. Desde este punto de vista el coeficiente que vamos a usar para determinar la validez nos indica que el proceso y agrupamiento de clustering en los datos esta ejecutado correctamente. Para realizar este proceso utilizamos un nodo de Knime Analytics el cual nos permitió obtener los coeficientes de cada dato individual del clúster para comprobar que el agrupamiento sea bueno.

En el primero modelo se evidencia que los coeficientes individuales de cada clúster se encuentran en un rango de 0 y 1; lo que se define como un buen agrupamiento, el coeficiente de Silhouette determina que mientras el valor se encuentre más cerca de 1 y más cerca de 0, se define como clustering válido. A continuación, se muestran los datos individuales y sus coeficientes para el modelo 1:



Result table - 4:16 - Silhouette Coefficient

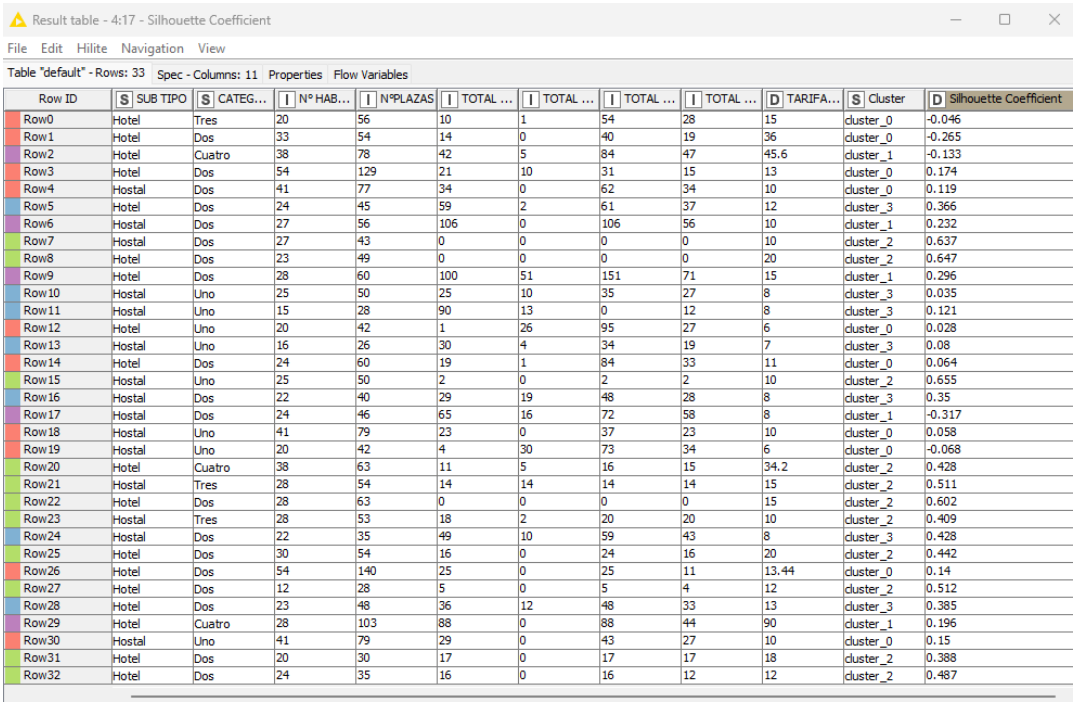
File Edit Hilite Navigation View

Table "default" - Rows: 97 Spec - Columns: 11 Properties Flow Variables

Row ID	SUB TIPO	CATEG...	N° HAB...	NºPLAZAS	TOTAL ...	TOTAL ...	TOTAL ...	TOTAL ...	TARIFA...	Cluster	Silhouette Coefficient
Row0	Hotel	dos	54	140	27	4	113	51	13	cluster_3	0.014
Row1	Hostal	uno	20	43	87	0	187	87	6	cluster_3	0.227
Row2	Hotel	dos	20	53	13	3	41	25	18	cluster_2	0.222
Row3	Hostal	uno	41	100	16	0	43	37	10	cluster_1	0.372
Row4	Hostal	uno	15	26	11	1	43	25	8	cluster_2	0.625
Row5	Hostal	dos	22	46	31	5	76	57	8	cluster_0	0.295
Row6	Hostal	uno	12	23	7	2	25	16	8	cluster_2	0.525
Row7	Hotel	dos	12	28	13	3	56	32	13	cluster_2	0.552
Row8	Hotel	dos	24	64	35	1	132	48	12.6	cluster_0	0.325
Row9	Hostal	uno	25	65	4	1	11	7	8	cluster_1	0.344
Row10	Hostal	tres	28	70	14	0	30	24	10	cluster_1	0.334
Row11	Hotel	dos	24	61	19	5	82	40	12	cluster_0	0.369
Row12	Hostal	dos	29	64	57	9	148	68	10	cluster_3	0.046
Row13	Hotel	dos	30	70	20	1	67	35	25	cluster_0	0.033
Row14	Hotel	dos	28	63	29	5	110	55	25	cluster_0	0.513
Row15	Hotel	dos	23	48	18	4	77	43	13	cluster_0	0.238
Row16	Hostal	dos	29	56	85	11	185	104	10	cluster_3	0.275
Row17	Hostal	dos	25	61	6	7	21	14	8	cluster_1	0.254
Row18	Hostal	dos	16	32	13	0	17	13	13	cluster_2	0.397
Row19	Hostal	uno	16	40	32	7	73	37	7	cluster_0	0.01
Row20	Hostal	uno	17	40	18	1	43	24	7	cluster_2	0.542
Row21	Hotel	dos	22	35	28	0	77	44	15	cluster_0	0.134
Row22	Hostal	dos	19	42	12	1	28	20	15	cluster_2	0.419
Row23	Hostal	uno	15	26	16	0	57	35	8	cluster_2	0.513
Row24	Hotel	dos	23	48	21	8	88	44	15	cluster_0	0.423
Row25	Hostal	dos	16	32	19	0	38	31	13	cluster_2	0.608
Row26	Hostal	uno	25	65	6	0	14	9	8	cluster_1	0.358
Row27	Hotel	dos	20	53	9	0	24	15	18	cluster_1	-0.024
Row28	Hostal	uno	41	100	13	0	39	25	10	cluster_1	0.415
Row29	Hostal	dos	22	46	31	6	87	49	8	cluster_0	0.422
Row30	Hostal	uno	17	40	14	0	28	20	8	cluster_2	0.477
Row31	Hostal	uno	16	40	12	14	46	32	7	cluster_2	0.497
Row32	Hotel	dos	22	35	18	0	62	35	15	cluster_2	0.358
Row33	Hostal	dos	19	42	14	0	41	24	15	cluster_2	0.506
Row34	Hostal	tres	28	70	21	0	57	33	10	cluster_1	0.103
Row35	Hostal	uno	20	43	3	42	92	45	6	cluster_0	0.234

Figura 105. Coeficientes de Silhouette clúster individual

Para el segundo modelo se evidencia que los coeficientes individuales de cada clúster se encuentran en un rango de 0 y 1; lo que se define como un buen agrupamiento. A pesar de que existen valores como el primero -0.046, sigue estimando como válido debido a que se hacer más al 0, de acuerdo con los criterios de Silhouette.



Result table - 4:17 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 33 Spec - Columns: 11 Properties Flow Variables

Row ID	SUB TIPO	CATEG...	N° HAB...	NºPLAZAS	TOTAL ...	TOTAL ...	TOTAL ...	TOTAL ...	TARIFA...	Cluster	Silhouette Coefficient
Row0	Hotel	Tres	20	56	10	1	54	28	15	cluster_0	-0.046
Row1	Hotel	Dos	33	54	14	0	40	19	36	cluster_0	-0.265
Row2	Hotel	Cuatro	38	78	42	5	84	47	45.6	cluster_1	-0.133
Row3	Hotel	Dos	54	129	21	10	31	15	13	cluster_0	0.174
Row4	Hostal	Dos	41	77	34	0	62	34	10	cluster_0	0.119
Row5	Hotel	Dos	24	45	59	2	61	37	12	cluster_3	0.366
Row6	Hostal	Dos	27	56	106	0	106	56	10	cluster_1	0.232
Row7	Hostal	Dos	27	43	0	0	0	0	10	cluster_2	0.637
Row8	Hotel	Dos	23	49	0	0	0	0	20	cluster_2	0.647
Row9	Hotel	Dos	28	60	100	51	151	71	15	cluster_1	0.296
Row10	Hostal	Uno	25	50	25	10	35	27	8	cluster_3	0.035
Row11	Hostal	Uno	15	28	90	13	0	12	8	cluster_3	0.121
Row12	Hotel	Uno	20	42	1	26	95	27	6	cluster_0	0.028
Row13	Hostal	Uno	16	26	30	4	34	19	7	cluster_3	0.08
Row14	Hotel	Dos	24	60	19	1	84	33	11	cluster_0	0.064
Row15	Hostal	Uno	25	50	2	0	2	2	10	cluster_2	0.655
Row16	Hostal	Dos	22	40	29	19	48	28	8	cluster_3	0.35
Row17	Hostal	Dos	24	46	65	16	72	58	8	cluster_1	-0.317
Row18	Hostal	Uno	41	79	23	0	37	23	10	cluster_0	0.058
Row19	Hostal	Uno	20	42	4	30	73	34	6	cluster_0	-0.068
Row20	Hotel	Cuatro	38	63	11	5	16	15	34.2	cluster_2	0.428
Row21	Hostal	Tres	28	54	14	14	14	14	15	cluster_2	0.511
Row22	Hotel	Dos	28	63	0	0	0	0	15	cluster_2	0.602
Row23	Hostal	Tres	28	53	18	2	20	20	10	cluster_2	0.409
Row24	Hostal	Dos	22	35	49	10	59	43	8	cluster_3	0.428
Row25	Hotel	Dos	30	54	16	0	24	16	20	cluster_2	0.442
Row26	Hotel	Dos	54	140	25	0	25	11	13.44	cluster_0	0.14
Row27	Hotel	Dos	12	28	5	0	5	4	12	cluster_2	0.512
Row28	Hotel	Dos	23	48	36	12	48	33	13	cluster_3	0.385
Row29	Hotel	Cuatro	28	103	88	0	88	44	90	cluster_1	0.196
Row30	Hostal	Uno	41	79	29	0	43	27	10	cluster_0	0.15
Row31	Hotel	Dos	20	30	17	0	17	17	18	cluster_2	0.388
Row32	Hotel	Dos	24	35	16	0	16	12	12	cluster_2	0.487

Figura 106. Coeficientes de Silhouette clúster individual modelo 2

En el tercer modelo se muestran que los coeficientes individuales de cada clúster se encuentran en un rango de 0 y 1; lo que se define como un buen agrupamiento.

Result table - 4:18 - Silhouette Coefficient

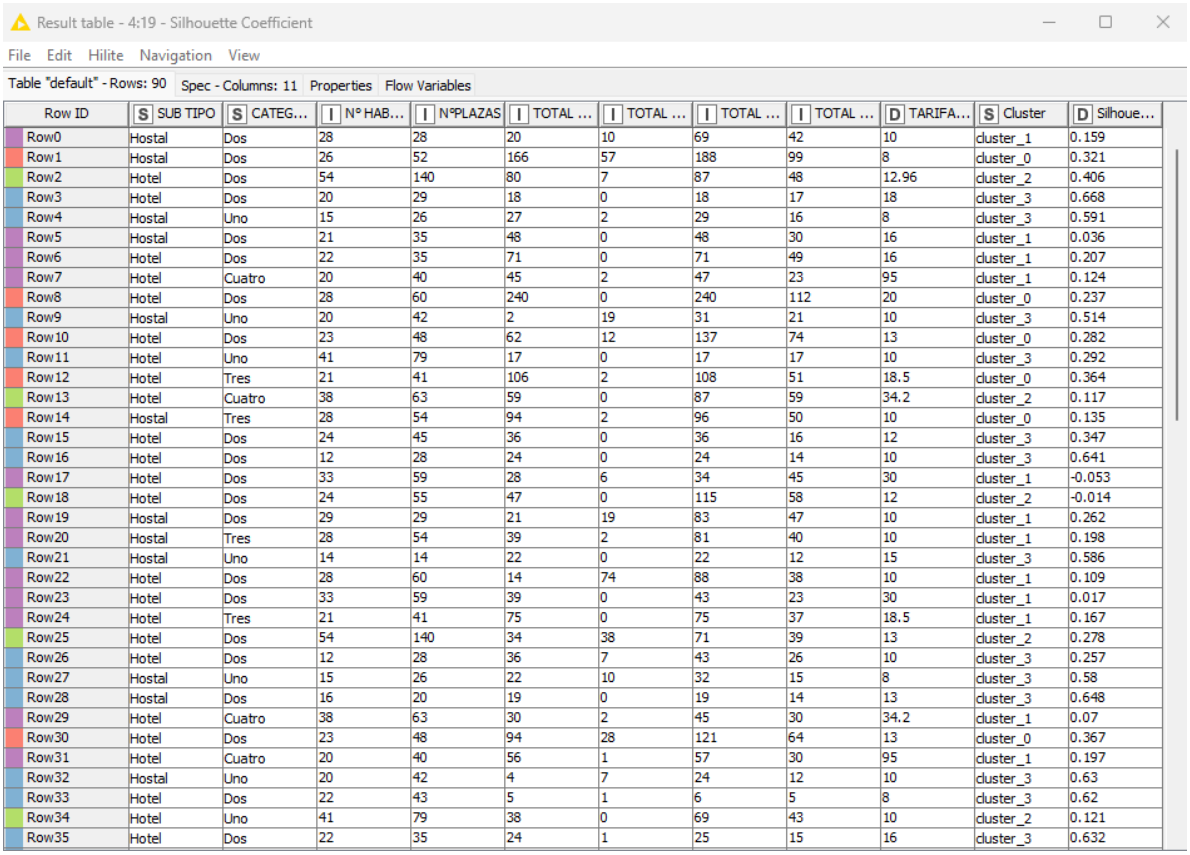
File Edit Hilite Navigation View

Table "default" - Rows: 87 Spec - Columns: 11 Properties Flow Variables

Row ID	SUB TIPO	CATEG...	N° HAB...	N°PLAZAS	TOTAL ...	TOTAL ...	TOTAL ...	TOTAL ...	TARIFA...	Cluster	Silhoue...
Row48	Hotel	Dos	20	29	9	0	4	6	18	cluster_0	0.542
Row49	Hotel	Uno	41	79	32	0	22	30	10	cluster_1	0.176
Row50	Hostal	Uno	20	42	0	33	83	33	10	cluster_3	0.288
Row51	Hostal	Tres	20	45	34	20	54	30	10	cluster_3	0.427
Row52	Hotel	Dos	33	54	17	4	21	12	140	cluster_0	0.066
Row53	Hostal	Dos	28	28	18	32	50	21	10	cluster_3	0.218
Row54	Hotel	Dos	20	29	12	0	7	8	18	cluster_0	0.537
Row55	Hotel	Dos	12	28	34	4	38	25	15	cluster_3	0.162
Row56	Hostal	Uno	15	26	24	0	24	17	8	cluster_0	0.239
Row57	Hotel	Cuatro	38	63	36	1	75	37	34	cluster_3	0.351
Row58	Hotel	Cuatro	20	40	108	2	110	56	114	cluster_2	0.245
Row59	Hotel	Dos	23	48	112	13	125	79	13	cluster_2	0.4
Row60	Hotel	Dos	54	140	98	14	112	62	12	cluster_2	0.153
Row61	Hostal	Uno	41	79	30	0	35	25	10	cluster_1	0.093
Row62	Hotel	Dos	22	35	82	0	82	60	16	cluster_2	0.151
Row63	Hostal	Dos	21	35	57	0	57	40	16	cluster_3	0.385
Row64	Hostal	Dos	28	54	13	5	27	15	10	cluster_0	0.182
Row65	Hotel	Dos	30	58	30	4	34	22	20	cluster_3	0.16
Row66	Hotel	Dos	28	60	2	0	2	2	15	cluster_0	0.381
Row67	Hotel	Dos	33	54	8	6	14	8	30	cluster_0	0.372
Row68	Hostal	Dos	21	35	2	0	2	1	16	cluster_0	0.529
Row69	Hotel	Tres	19	54	3	0	34	16	20	cluster_0	0.141
Row70	Hostal	Dos	27	43	71	45	116	67	8	cluster_2	0.265
Row71	Hotel	Dos	54	140	30	2	32	17	13	cluster_1	0.552
Row72	Hotel	Dos	12	28	4	3	7	5	10	cluster_0	0.508
Row73	Hostal	Dos	16	20	8	0	8	8	15	cluster_0	0.501
Row74	Hotel	Cuatro	38	63	6	0	15	6	34	cluster_0	0.271
Row75	Hotel	Cuatro	20	40	29	0	29	14	114	cluster_0	0.056
Row76	Hostal	Uno	20	42	3	12	37	14	8	cluster_0	0.079
Row77	Hotel	Uno	41	79	14	0	25	14	10	cluster_1	0.216
Row78	Hotel	Dos	22	35	23	0	23	20	16	cluster_0	0.255
Row79	Hotel	Dos	24	55	25	0	40	25	12	cluster_3	0.248
Row80	Hostal	Dos	24	63	14	2	39	21	10	cluster_3	0.111
Row81	Hotel	Dos	23	49	15	5	43	16	15	cluster_3	0.11
Row82	Hostal	Única	4	16	4	0	10	4	10	cluster_0	0.422
Row83	Hostal	Dos	21	35	10	0	10	7	16	cluster_0	0.544

Figura 107. Coeficientes de Silhouette clúster individual modelo 3

El último modelo de igual forma se indica que los coeficientes individuales de cada clúster se encuentran en un rango de 0 y 1; lo que se define como un buen agrupamiento. Cabe mencionar que, al momento de la descripción de los modelos mediante sus promedios, los resultados de sus agrupamientos fueron válidos, es por ellos que los coeficientes individuales de cada clúster tienen un buen agrupamiento y son válidos.



Result table - 4:19 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 90 Spec - Columns: 11 Properties Flow Variables

Row ID	SUB TIPO	CATEG...	N° HAB...	NºPLAZAS	TOTAL ...	TOTAL ...	TOTAL ...	TOTAL ...	TARIFA...	Cluster	Silhoue...
Row0	Hostal	Dos	28	28	20	10	69	42	10	cluster_1	0.159
Row1	Hostal	Dos	26	52	166	57	188	99	8	cluster_0	0.321
Row2	Hotel	Dos	54	140	80	7	87	48	12.96	cluster_2	0.406
Row3	Hotel	Dos	20	29	18	0	18	17	18	cluster_3	0.668
Row4	Hostal	Uno	15	26	27	2	29	16	8	cluster_3	0.591
Row5	Hostal	Dos	21	35	48	0	48	30	16	cluster_1	0.036
Row6	Hotel	Dos	22	35	71	0	71	49	16	cluster_1	0.207
Row7	Hotel	Cuatro	20	40	45	2	47	23	95	cluster_1	0.124
Row8	Hotel	Dos	28	60	240	0	240	112	20	cluster_0	0.237
Row9	Hostal	Uno	20	42	2	19	31	21	10	cluster_3	0.514
Row10	Hotel	Dos	23	48	62	12	137	74	13	cluster_0	0.282
Row11	Hotel	Uno	41	79	17	0	17	17	10	cluster_3	0.292
Row12	Hotel	Tres	21	41	106	2	108	51	18.5	cluster_0	0.364
Row13	Hotel	Cuatro	38	63	59	0	87	59	34.2	cluster_2	0.117
Row14	Hostal	Tres	28	54	94	2	96	50	10	cluster_0	0.135
Row15	Hotel	Dos	24	45	36	0	36	16	12	cluster_3	0.347
Row16	Hotel	Dos	12	28	24	0	24	14	10	cluster_3	0.641
Row17	Hotel	Dos	33	59	28	6	34	45	30	cluster_1	-0.053
Row18	Hotel	Dos	24	55	47	0	115	58	12	cluster_2	-0.014
Row19	Hostal	Dos	29	29	21	19	83	47	10	cluster_1	0.262
Row20	Hostal	Tres	28	54	39	2	81	40	10	cluster_1	0.198
Row21	Hostal	Uno	14	14	22	0	22	12	15	cluster_3	0.586
Row22	Hotel	Dos	28	60	14	74	88	38	10	cluster_1	0.109
Row23	Hotel	Dos	33	59	39	0	43	23	30	cluster_1	0.017
Row24	Hotel	Tres	21	41	75	0	75	37	18.5	cluster_1	0.167
Row25	Hotel	Dos	54	140	34	38	71	39	13	cluster_2	0.278
Row26	Hotel	Dos	12	28	36	7	43	26	10	cluster_3	0.257
Row27	Hostal	Uno	15	26	22	10	32	15	8	cluster_3	0.58
Row28	Hostal	Dos	16	20	19	0	19	14	13	cluster_3	0.648
Row29	Hotel	Cuatro	38	63	30	2	45	30	34.2	cluster_1	0.07
Row30	Hotel	Dos	23	48	94	28	121	64	13	cluster_0	0.367
Row31	Hotel	Cuatro	20	40	56	1	57	30	95	cluster_1	0.197
Row32	Hostal	Uno	20	42	4	7	24	12	10	cluster_3	0.63
Row33	Hotel	Dos	22	43	5	1	6	5	8	cluster_3	0.62
Row34	Hotel	Uno	41	79	38	0	69	43	10	cluster_2	0.121
Row35	Hotel	Dos	22	35	24	1	25	15	16	cluster_3	0.632

Figura 108. Coeficientes de Silhouette clúster individual modelo 4

En la siguiente tabla se muestran los cuatro modelos con los valores de los promedios del Coeficiente de Silhouette:

Tabla 28. Promedio de los coeficientes de Silhouette por cada clúster algoritmo K-means

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
Modelo 1	0,298	0,253	0,486	0,062
Modelo 2	0,035	0,055	0,52	0,252
Modelo 3	0,356	0,323	0,136	0,23
Modelo 4	0,289	0,104	0,188	0,562

4.1.4.5. Evaluación

Esta fase de la metodología se evalúan los modelos que se generaron, pero como se mencionó se analizan tomando en cuenta los objetivos con criterios de negocio, una vez realizada la evaluación se debe determinar si se han cumplidos los objetivos, y de ser así avanzar a la fase de implementación.

- **Evaluar los resultados**

Se estableció que, para un criterio de éxito en base a los objetivos de negocio, se analice y se evalué la información de los procesos de la demanda turística, por medio de la aplicación de técnicas de minería de datos, pero aún no se interpreta la información resultante. Y desde este punto de vista es inevitables basarse en los criterios de éxito de minería de datos para calificar como aceptable los resultados de los modelos generados.

Los modelos que se generaron fueron validados en base a los criterios establecidos en el plan de pruebas y se puede decir que estos métodos de evaluación son más precisos y específicos; los cuatro modelos generados fueron aprobados en bases a sus coeficientes Silhouette y cumplieron con los criterios de éxito. Es por ello, que, para cumplir con la etapa de análisis y exploración de datos con base en los modelos aprobados, se requiere pasar a la fase de implementación.

4.1.4.6. Implantación

Esta es la última fase de la metodología CRSIP-DM y el objetivo que tienen es mostrar a las personas que intervienen en los procesos de control mediante las cuales se determina la demanda turística de la provincia de Carchi, como poner en funcionamiento el proyecto que se ha construido. Desde este punto de vista se tienen que exponer los resultados obtenidos de los modelos de clustering de forma sencilla, de modo que le permita a la persona encarga de los procesos de alojamiento y gasto turístico, evaluar y analizar la información.

Con el fin de cumplir esta fase y para una mejor visualización de los datos, basándonos en los resultados de los modelos generados, se crearon Dashboard interactivos en la plataforma de Power BI, para que el analista Zonal pueda observar un tablero de datos dinámico, gráficos interactivos, y la capacidad de filtrar información según como lo requiera.

Se generaron cinco dashboard puesto que marcaban una tendencia de turismo común en cada uno de los años iniciando en el año 2019 y terminando en el 2022.

Cabe mencionar que cada tablero de contenido generado se basó en los modelos de clustering que se generaron en el software de Knime Analytics. El primer proceso de visualización de datos de la demanda turística se observa de la siguiente forma:

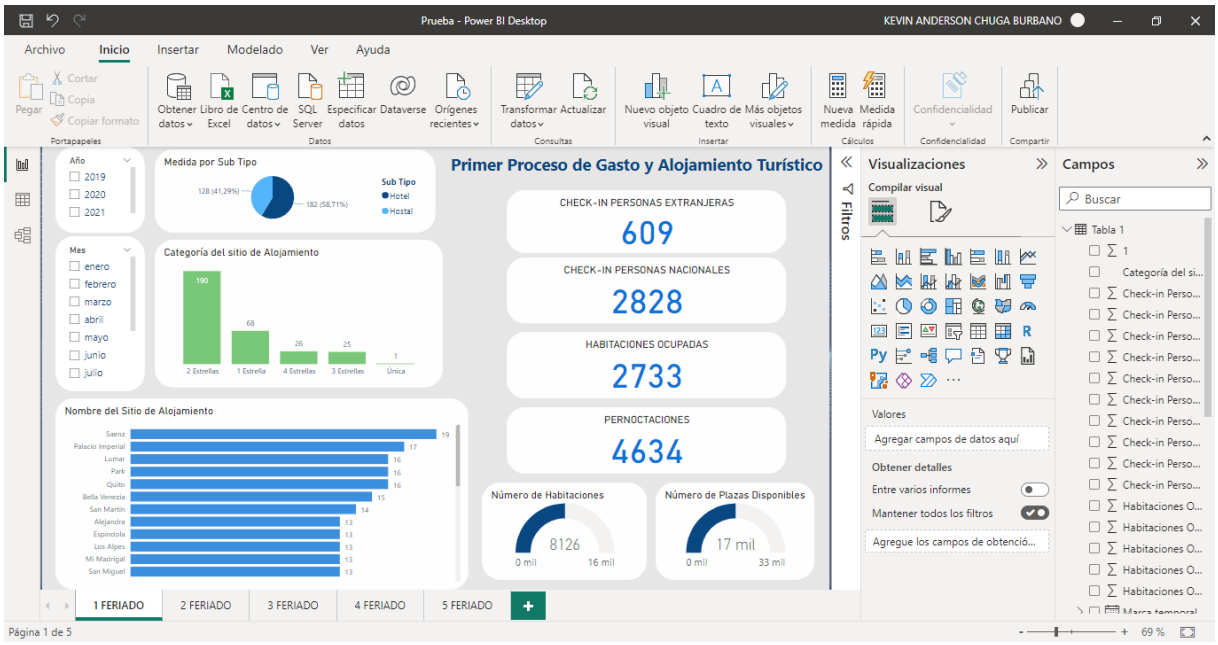


Figura 109. Dashboard del primer proceso de alojamiento y gasto turístico

Es un dashboard interactivo, cambia según las filtraciones que podría dar puede ser según el año, personas nacionales, extranjeras, entre otras filtraciones

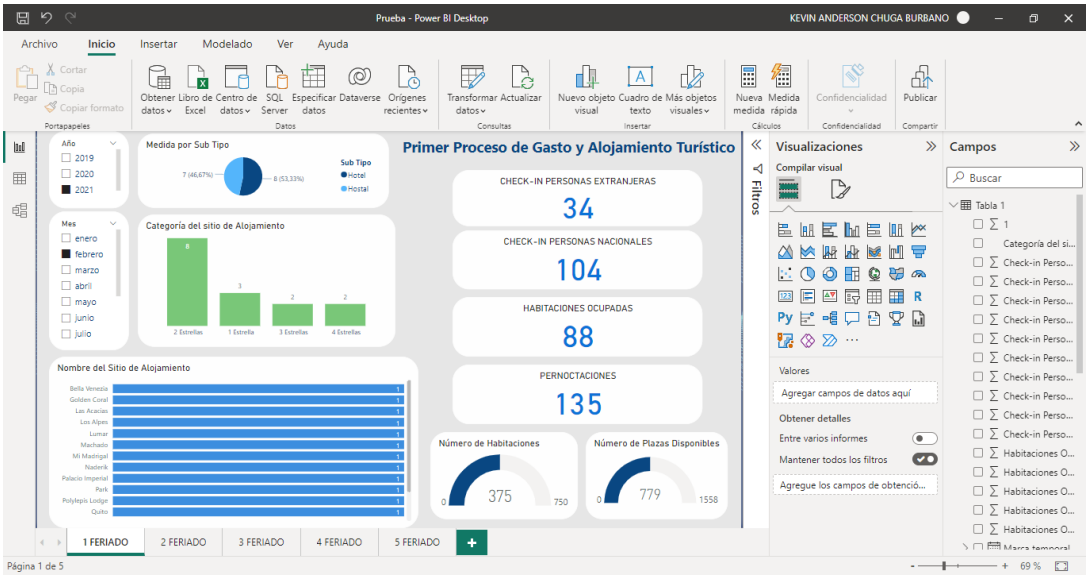


Figura 110. Filtrado según las fechas en proceso de alojamiento y gasto turístico

Algunos de los componentes que consta el Dashboard son:

- Información de alojamiento general que tiene relación con la demanda turística de la provincia.

Figura 111. Gráficos interactivos mostrando indicadores de alojamiento turístico

- Método de filtración de fechas según el año y el mes

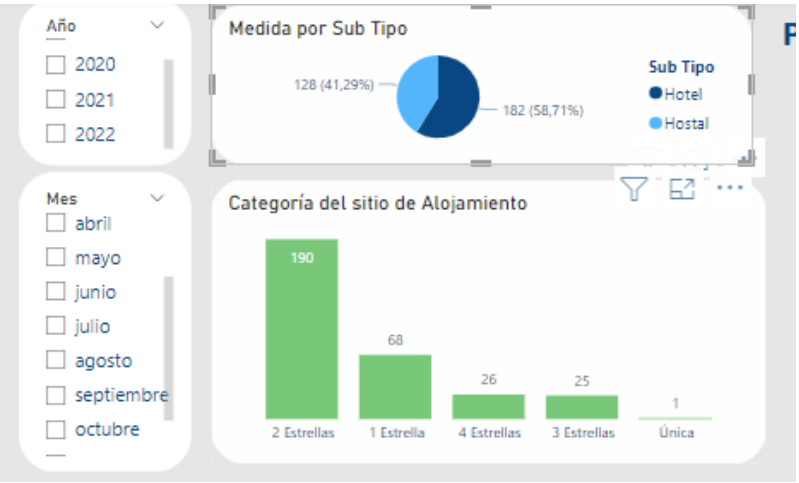


Figura 112. Gráfico interactivo según las fechas

Los demás procesos mantienen la misma estructura con relación a los tableros dinámicos, lo que varía son sus datos.

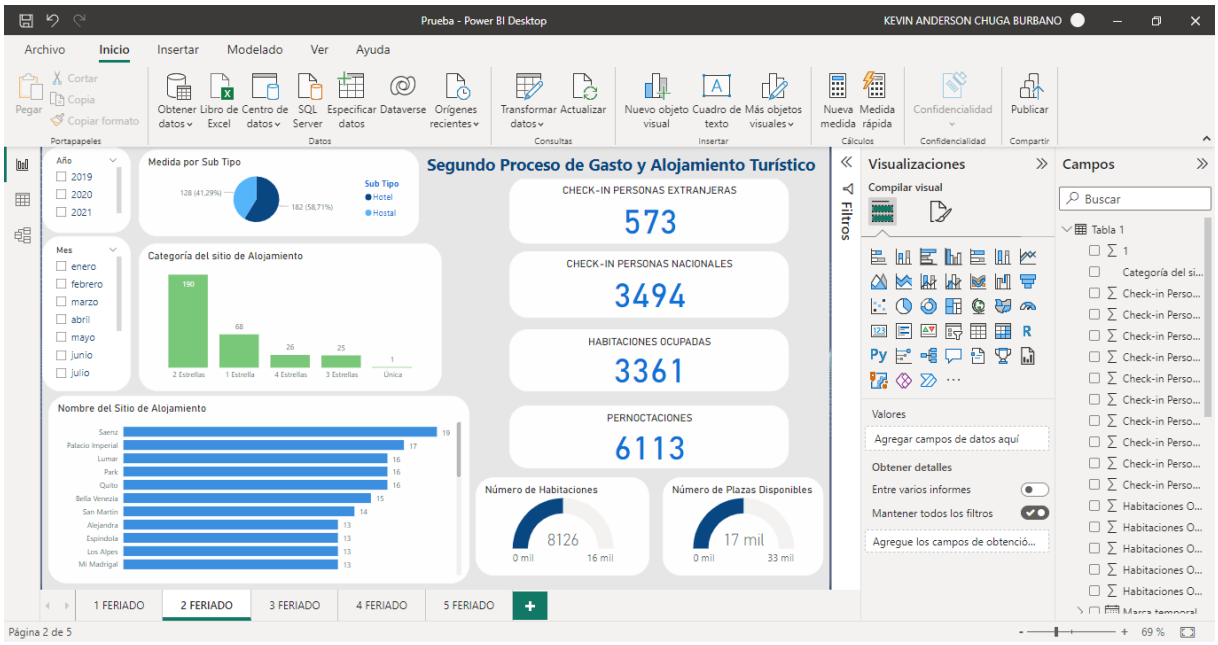


Figura 113. Vista general del segundo proceso de gasto y alojamiento turístico

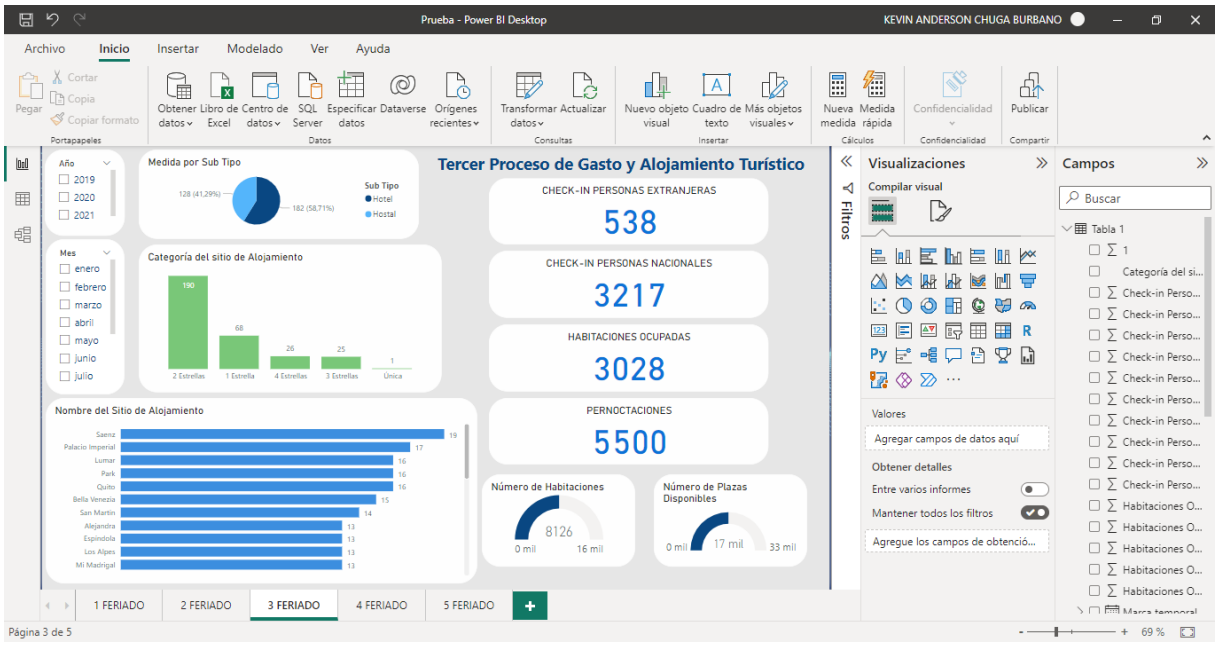


Figura 114. Vista general del tercer proceso de gasto y alojamiento turístico

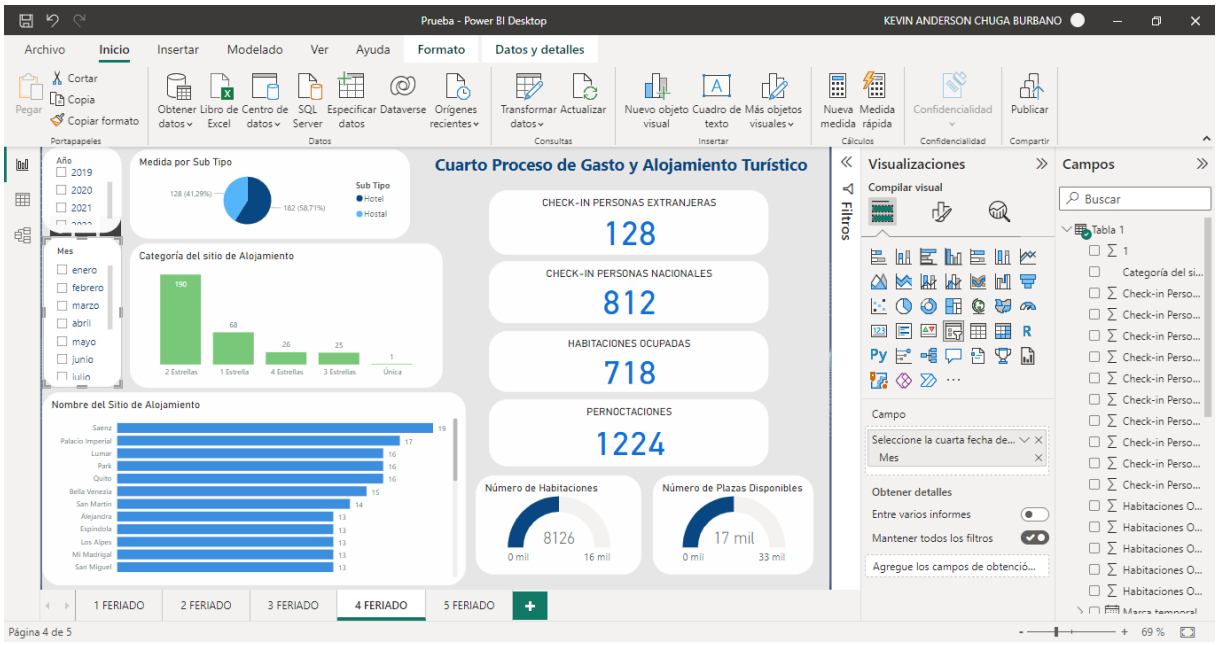


Figura 115. Vista general del cuarto proceso de gasto y alojamiento turístico

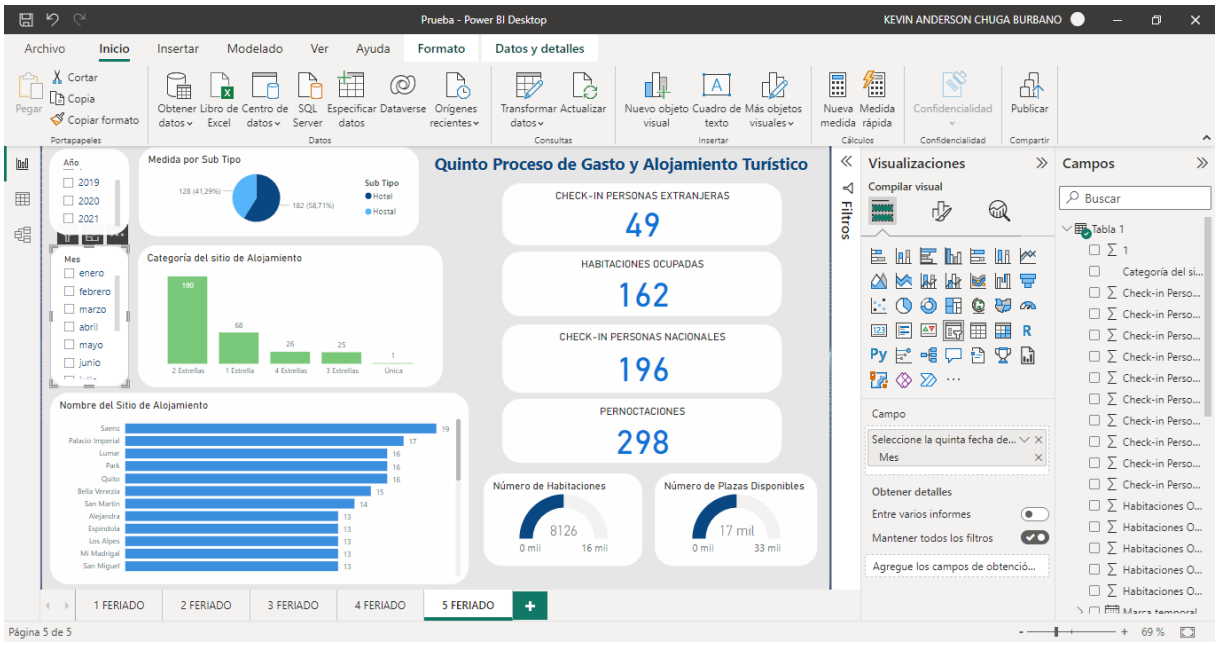


Figura 116. Vista general del quinto proceso de gasto y alojamiento turístico

- **Monitorización y mantenimiento**

La supervisión y el mantenimiento es una fase importante para un proyecto como el de minería de datos, debido a que los datos pueden cambiar con mucha frecuencia, y más cuanto se trata con procesos para demanda turística, debido a que los procesos se los sigue registrando continuamente, y esa la razón por que se generó un plan de supervisión y mantenimiento.

Debemos mencionar que se recolecto información de los procesos de alojamiento y gasto turístico desde el 2019 hasta fines del año 2022, este proceso va a continuar en el 2023, y esta es la razón por la que en base a los modelos generados con el algoritmo de clustering se creó una estructura en Google Forms con variables resultantes de los clústeres generados, para que por medio de esta encuesta se registren los nuevos procesos de alojamiento y gasto turístico.

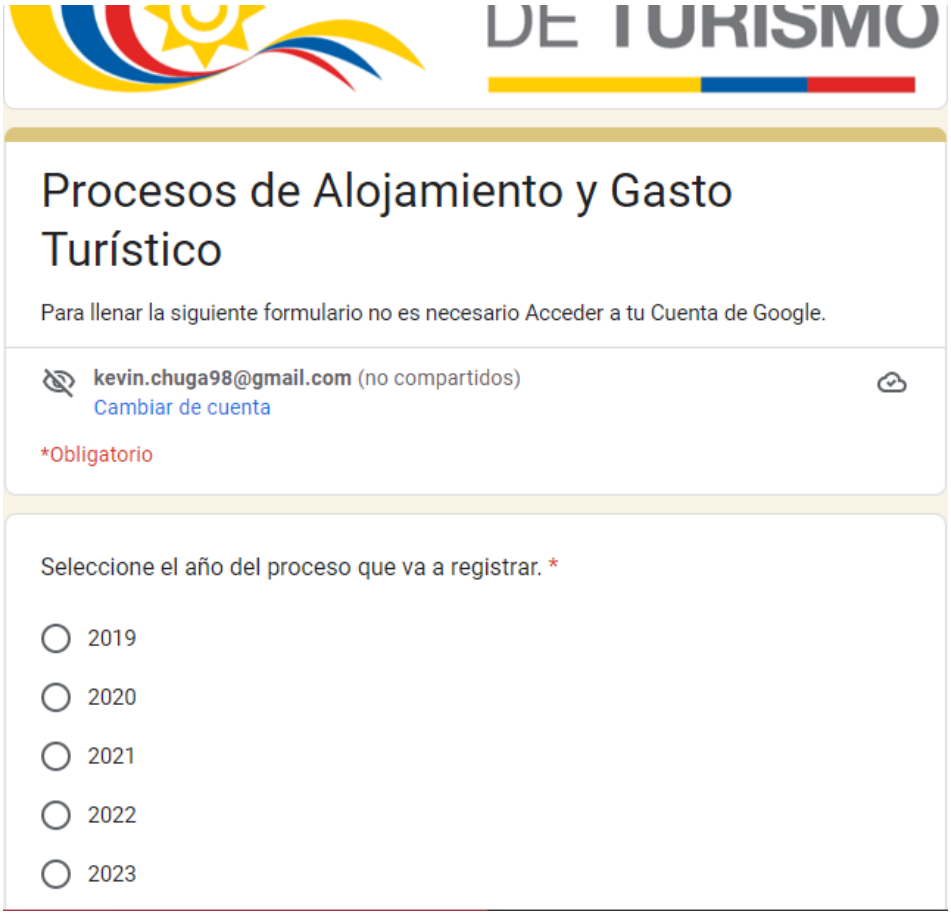


Figura 117. Formulario creado en base a los modelos generados en Knime Analytics

Este formulario de Google se conectó directamente con la plataforma de Power BI para que de alguna manera pueda seguir cargando información y monitorizando los procesos de la demanda turística de la provincia.

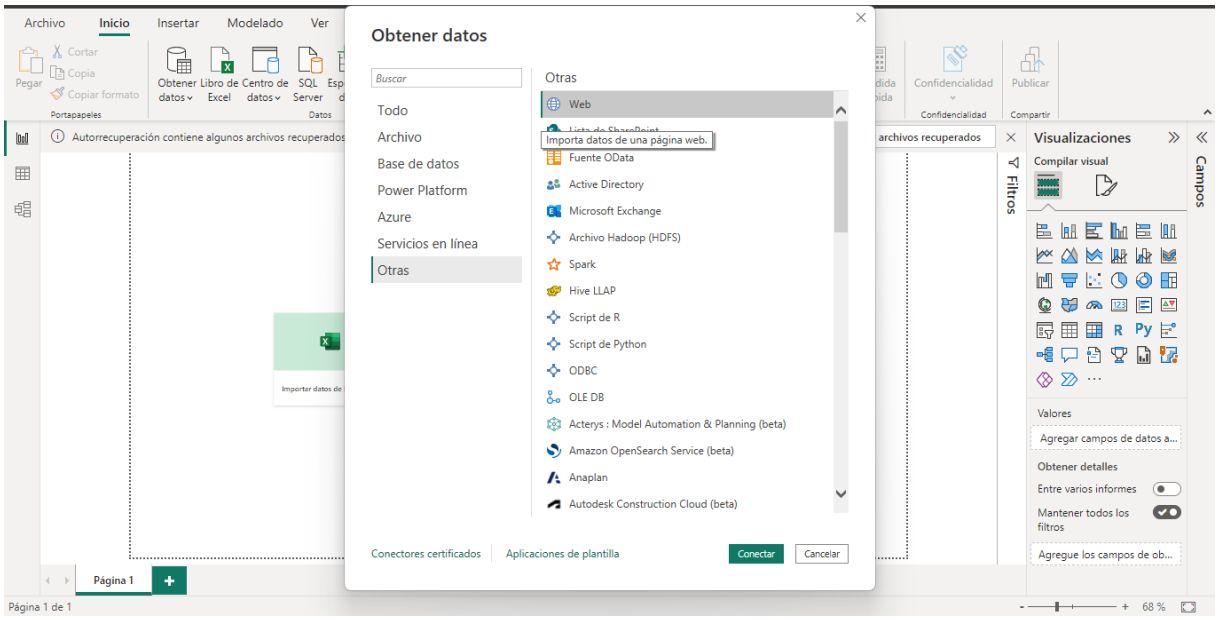


Figura 118. Conexión de Power BI con formulario de Google

4.2. DISCUSIÓN

El estudio tuvo como objetivo principal la propuesta de un modelo de minería de datos, mediante el cual buscaba mejorar los procesos de control que emplea el Ministerio de Turismo para la demanda turística de la provincia de la Carchi, la investigación inicio con bases en las necesidades que tenían las personas involucradas en estos procesos, y a partir de un primer encuentro, ir comprendiendo como se manejaban los procesos, su tiempo de ejecución, la manera de almacenar la información y si se realizaba un análisis de la información. Todo ello con el fin de conocer la realidad de cómo se gestionaban estos procesos para lograr tener una visualización inicial del problema.

El uso de un enfoque cualitativo y cuantitativo fue necesario dentro de la investigación, debido a la recolección de información que se requirió hacer, con el fin de formar un marco teórico y metodológico, que sirvió de referencia para la construcción de un modelado de minería de datos basándonos en los procesos que emplea el ministerio de turismo. El enfoque mixto de la investigación dio la posibilidad de aplicar técnicas e instrumentos mediante los cuales se generaron nuevos criterios sobre los procesos de la demanda turística, y se pudo observar detalladamente como se gestionaban estos procesos, además de ello se obtuvo un nuevo panorama de la investigación, entre otros aspectos.

Las técnicas y sus instrumentos que se usó para recoger información; en primera instancia una entrevista mediante una guía de preguntas realizada al Analista de Desarrollo y Promoción Turística de la Zona N°1, que era el encargado de ejecutar los procesos, luego se aplicó una encuesta mediante un cuestionario estructurado, que fue dirigido a todos los hoteles y hostales, registrados en la provincia del Carchi, debido a que en gran medida eran usuarios involucrados en los proceso de la demanda turística de la provincia. Por último, se hizo un análisis documental a través del cual, se pudo observar los registros y toda la información histórica de los procesos de la demanda turística de la provincia, además de ello, sirvió para comprender como se almacena y gestiona toda la información.

El presente estudio como se había mencionado tenía la tarea de crear un modelo de minería de datos que ayude de alguna manera con los procesos que se maneja dentro del ministerio de turismo para demanda de turismo de la provincia, los resultados fueron favorables con el uso de la metodología CRISP-DM, porque permitió gestionar todo el procesos para la creación de un modelo de minería de datos, iniciando con un análisis de lo que se quiere obtener aplicando técnicas de minería de datos, luego una recolección de requisitos basándonos en el hardware donde se va a crear el modelo, a partir de allí comprender los datos de los procesos de la demanda turística mediante tareas como realizar un evaluación inicial de los datos con el fin de establecer criterios de completitud y determinar posibles inconsistencias entre los datos recopilados, luego de ello pasa a una etapa de , preparación de datos que involucran tareas como selección, limpieza y formateo de datos, entre otras acciones.

Luego de ello intervienen fases de modelado y evaluación, donde en la primera tarea se construye el modelado con base en los datos que han sido previamente preparado, seguido se evalúa el modelo dependiendo de la técnica de modelo, en este caso una técnica de modelado de segmentación o clustering, y la teoría dice que para verificar la calidad y validez del modelo se toman en cuenta criterios de cohesión, separación o coeficiente de silhouette que es una mezcla de ambos criterios. Por último, entra a fase de implantación, donde se exponen los resultados del modelo generado de forma que puedan ser comprendidos y entendidos fácilmente por el Analista de Desarrollo y Promoción Turística de la Zona N°1, que es el encargado de los procesos de la demanda turística de la provincia.

Todo este proceso que se aplicó dio como resultado a una propuesta de modelado de minería de datos para un tener la posibilidad de mejorar los procesos que emplea el ministerio de turismo para determinar la demanda turística de la provincia del Carchi.

Previos trabajos investigativos como es el caso de Uriarte del Águila, que presenta un proyecto de investigación en el año 2018, con el objetivo de tomar decisiones, para mejorar la gestión de un cliente, mediante la aplicación de técnicas de minera de datos, las acciones iniciales que tuvo que hacer fue un análisis sobre cómo se gestiona los procesos de atención al cliente y los datos que se manejan dentro de los procesos, además de detectar las necesidades. Los resultados de la investigación fueron exitosos ya que el área de gestión al cliente mejoró en criterios de eficacia y eficiencia de como analizar y procesar la información. Una comparativa de este antecedente con el presente estudio es el grado de importancia que tiene la comprensión de los datos sobre los cuales se va a aplicar minería de datos, es una tarea que ayuda a conocer los tipos de datos que se va a emplear, así como su calidad.

En otra investigación realizada por Martínez Clemente que tenía el enfoque de aplicar minería de datos para el diseño de nuevas métricas en los procesos de la empresa ENTEL, el primer paso que hicieron fue análisis de lo proceso relacionados con los ingresos, luego se determinó que los procesos se los realizaba de forma manual, lo que generaba demanda, tiempo y esfuerzo de parte de los analistas de la empresa. Tomando en cuenta estos aspectos aplicaron data mining y lograron diseñar un nuevo proceso para identificar el servicio y el cliente que se encontraban en estado crítico, lo que les permitió tener datos con mayor exactitud con relación a los ingresos no percibidos, y generar nuevas estrategias para hacer el cobro. En comparación con la presente investigación es importante realizar una comprensión del negocio, esto permite recoger información de datos válidos relacionados al problema que se desea resolver, y otro aspecto importante es que nos ayudara a interpretar correctamente los resultados.

Un antecedente investigativo realizado por la Universidad de Loja del Ecuador, con la iniciativa de facilitar la información turística necesaria para tomar decisiones, en base a los indicadores que se presentan dentro del sector turístico crean un observatorio turístico, donde establece un marco de referencia para el análisis

sistemático de la situación real y tendencias del mercado turístico con el fin de tener bases fiables para tomar decisiones en base al fomento del turismo sostenible. En comparación con el estudio es importante determinar la forma en la que se evalúa y se analiza la información, cuando se tratan investigaciones de análisis de datos; existen herramientas en tendencias que se manejan en el mercado actual como Power BI que permiten visualizar la información de forma más comprensible mediante el uso gráficos y tableros interactivos, que le ayuden al usuario o cliente a tomar decisiones con mayor claridad.

V. CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

- La información recopilada durante todo el proceso de investigación permitió comprender como es el procedimiento para generar modelos de minería de datos con relación a los procesos de control en un ambiente vinculado al turismo, además se generó referencias sólidas acera del manejo de información de los procesos que emplea el ministerio de turismo para el análisis de la demanda turística en la provincia del Carchi.
- En el estudio se aplicaron técnicas e instrumentos para la recolección de datos, que sirvió de apoyo para clarificar la gestión de almacenamiento y administración de la información de los procesos para la demanda turística, y también se logró comprender en profundidad la forma de ejecución de estos procesos y la cantidad de personas que intervienen.
- Los resultados de la investigación enfatizado en la idea a defender formulada y el propósito de la investigación, demuestran que no se mejoraran los procesos de control de la demanda turística de la provincia del Carchi debido a la limitada cantidad de datos con la que se cuenta, tomando en cuenta que la base de información con la que se contaba eran datos actuales de los procesos de alojamiento y gasto turístico.
- El análisis realizado a los datos de los procesos de alojamiento y gasto turístico, describieron un comportamiento en los datos que indicaban características relacionadas a diferentes procesos que emplea el ministerio de turismo, por ello se digitalizo toda la información que era relevante para determinar la demanda turística de la provincia.
- El modelo de minería de datos que se presenta en la investigación está relacionado con los procesos de control de la demanda turística de la provincia del Carchi, para la creación se aplicaron distintos procesos estructurados propuestos por la metodología CRISP-DM, que permitió generar un modelo de minería de datos, adaptado a las necesidades y objetivos del estudio.

- El despliegue de los resultados obtenidos por el modelo de minería de datos creado, se lo realizó mediante Power BI una herramienta dedicada al análisis de negocios, esta plataforma sirvió de apoyo para la presentación de resultados; se diseñaron paneles, gráficos y tableros dinámicos e interactivos, que le permitieron al encargado de los procesos de la demanda turística tener una mejor comprensión de la información de los resultados de la investigación, además de ello le da la opción de manipular y filtrar los datos, conforme los requiera.

5.2. RECOMENDACIONES

- Dentro de la investigación se aplicó la técnica de modelado de segmentación o clustering para los procesos de la demanda turística, es por esta razón que se recomienda realizar un nuevo proceso de investigación donde busque analizar distintas técnicas de minería de datos que puedan ser aplicadas dentro del área del turismo.
- Es importante comprender que cuando se trata de proyectos de minería de datos tener claro lo que se quiere obtener al aplicar estos conjuntos de tecnologías, por esta razón es recomendable estudiar y comprender los datos que se va a manejar y los procesos que se involucran para obtener la información.
- Se recomienda al Ministerio de Turismo de la provincia del Carchi gestionar convenios y alianzas con las distintas instituciones públicas de la provincia con el objetivo de obtener nueva información e indicadores relacionados al sector turístico, de modo que puedan aumentar la información y lograr tener una base de datos estructurada relacionada directamente con el turismo de la provincia del Carchi y todos los indicadores que engloban el sector turístico.
- Para un proyecto de investigación relacionada con el análisis y ciencia de datos, la base de información que se tenga es importante debido a que es el medio principal para aplicar los procesos de análisis y exploración de datos, por ello se recomienda tener claro la información que se va a usar, los tipos de datos que maneja y la forma en la que se encuentra almacenada toda la información.
- La metodología CRISP-DM, proporciona a los proyectos de minería de datos resultados efectivos por su flexibilidad, organización y adaptabilidad a las necesidades de la investigación, es recomendable que cuando se tratan proyectos relacionados con minería de datos, tener una línea de guía base para gestionar los procesos que conlleva la creación de un modelado.

- Power BI es una herramienta que sirve para el análisis y exploración de datos y mediante la generación de dashboard dinámicos le ayudan al usuario a comprender fácilmente la información que se presenta. Es recomendable utilizar este tipo de plataformas para exponer los resultados que se obtiene cuando se desarrolla un proyecto de minería de datos.

VI. REFERENCIAS BIBLIOGRÁFICAS

Arias, F. (2012). *El Proyecto de InvestIgacIón Introducción a la metodología científica.* Caracas: Episteme.

Asturias Corporación Universitaria. (2018). *El Proceso de Control.* Asturias: Asturias.

Beltrán, D., & Povéda, D. (2018). *RapidMiner.* Bogotá: Universidad Nacional de Colombia.

Bolaños, C., Cusba, J., Martínez, L., & Caicedo, C. (2018). *Herramientas de analítica para la exploración de datos.* Bogotá: Ministerio de Tecnologías de la Información y las Comunicaciones.

Caparrini, F. (20 de Diciembre de 2020). *cs.us.es.* Obtenido de cs.us.es: http://www.cs.us.es/~fsancho/?e=230

Cortina, V. (2018). *APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO.* Madrid: Universidad Carlos III de Madrid.

Fernández, J., Díaz, Z., & Martínez, P. (2017). *UN MODELO TEÓRICO PARA LA COORDINACIÓN AUTOMATIZADA DE TRANSACCIONES. APLICACIÓN AL SECTOR TURÍSTICO.* Madrid: Universidad Complutense de Madrid.

Gallardo, J. (2007). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* ER-DM.

Gavídia, Á., Visern, M., & Josep, D. (2017). *Data Warehouse.* Catalunya: Universidad Oberta de Catalunya.

Guzmán, E. (2018). *Métricas para la validación de Clustering.* Bogotá: Universidad Nacional de Colombia .

Martínez, B. (2018). *Minería de Datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

Martínez, C. (2018). *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA MEJORAR EL PROCESO DE CONTROL DE GESTIÓN EN ENTEL.* Santiago de Chile: Universidad de Chile.

Ministerio de Turismo. (04 de Enero de 2021). *turismo.gob.ec*. Obtenido de turismo.gob.ec: https://www.turismo.gob.ec/siturin-herramienta-para-actualizar-informacion-de-establecimien

Ministerio de Turismo Ecuador. (20 de Septiembre de 2022). *Gob.ec*. Obtenido de Gob.ec: https://www.gob.ec/mintur#:~:text=El%20Ministerio%20de%20Turismo%20ejerce,actividad%20generadora%20de%20desarrollo%20socio

MINTUR. (2020). *turismo.gob.ec*. Obtenido de turismo.gob.ec: https://www.turismo.gob.ec/wp-content/uploads/2021/05/Informe-de-Rendicio%CC%81n-de-Cuentas-2020.pdf

Miralbell, O., Lamsfus, C., Miquel, J., & Gonzáles, F. (2018). *Estudio de las TIC y el Turismo en España. Análisis de las ponencias del congreso TURITEC entre 1999 y 2010.* Barcelona: Universitat Oberta de Catalunya.

Morillo, M. (2019). Turismo y producto turístico. Evolución, conceptos, componentes y clasificación. *Visión Gerencial*, 135-158.

Pastrán, L., & Gongora, S. (2021). *ALGORITMO DE SELECCIÓN Y VALIDACIÓN DEL MÉTODO DE CLUSTERIZACIÓN ÓPTIMO PARA DATOS NO SUPERVISADOS.* Bogotá: Universidad Tecnológica de Pereira.

Recalde, E., Baldeón, P., Gaibor, M., & Toasa, R. (2020). Minería de datos con R para información académica en instituciones de Educación Superior. *Revista Ibérica de Sistemas y Tecnologías de Informática*, 63-71.

Rigol, M. (2018). Conceptualización de la demanda turística. *Ciencias Holguín*, 1-8.

Rodríguez, T., & Coronel, M. (2016). *MINERÍA DE DATOS PARA LA IDENTIFICACIÓN DE PATRONES DE CONSUMO DE SERVICIOS TURÍSTICOS PARA ORIENTAR LA OFERTA EN EL SECTOR HOTELERO DE LA CIUDAD DE TRUJILLO.* Trujillo: UNIVERSIDAD PRIVADA ANTENOR ORREGO.

Rodríguez, Y., & Díaz, A. (2019). Herramientas de Minerías de Datos . *Revista Cubana de Ciencias Informáticas* , 73-80.

Salinas, W., & Chavez, L. (2021). Aplicación del algoritmo K-medoid para la segmentacion de los alumnos ingresantes de una universidad. *Perfiles*, 24-29.

Sampieri, R., Fernández, C., & Baptista, P. (2014). *Metodologías de la Investigación*. México D.F.: McGRAW-HILL.

Socatelli, M. (2019). Consumidores en Turismo. En M. Socatelli, *Mercadeo Aplicado al Turismo. La Comercialización de Servicios - Productos y Destinos Turísticos Sostenibles* (págs. 1-7). Costa Rica: InterMark.

Strate Bi Open Bussiness Intelligence. (2022). *Introducción a KNIME*. Stratebi.

Universidad Técnica de Loja. (2016). *vinculacion.utpl.edu.ec*. Obtenido de vinculacion.utpl.edu.ec: https://vinculacion.utpl.edu.ec/es/observatorios/obtur

Uriarte del Águila, C. (2018). *"Minería de datos para mejorar la toma decisiones en el área de gestión al cliente de telefónica del Perú zonal Tarapoto*. Tarapoto: UNIVERSIDAD NACIONAL DE SAN MARTÍN - TARAPOTO.

Valles, D. (2017). LAS TECNOLOGÍAS DE LA INFORMACIÓN Y EL TURISMO. *Estudios Turísticos*, 3-24.

VII. ANEXOS

Anexo 1. Acta de la sustentación de Predefensa del TIC

UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI

Educamos para transformar el mundo

FACULTAD DE INDUSTRIAS AGROPECUARIAS Y CIENCIAS AMBIENTALES

CARRERA DE COMPUTACIÓN

ACTA

DE LA SUSTENTACIÓN ORAL DE LA PREDEFENSA DEL TRABAJO DE INTEGRACIÓN CURRICULAR

ESTUDIANTE:	CHUGA BURBANO KEVIN ANDERSON	**CÉDULA DE IDENTIDAD:**	0402046874
PERIODO ACADÉMICO:	2022B		
PRESIDENTE TRIBUNAL	MSC. GEORGINA GUADALUPE ARCOS PONCE	**DOCENTE TUTOR:**	MSC. JORGE HUMBERTO MIRANDA REALPE
DOCENTE:	MSC. JEFFERY ALEX NARANJO CEDEÑO		
TEMA DEL TIC:	"Minería de Datos para mejorar los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi en el año 2022"		

No.	CATEGORÍA	Evaluación cuantitativa	OBSERVACIONES Y RECOMENDACIONES
1	PROBLEMA - OBJETIVOS	9,00	
2	FUNDAMENTACIÓN TEÓRICA	9,00	
3	METODOLOGÍA	9,00	
4	RESULTADOS	9,00	
5	DISCUSIÓN	9,00	Profundizar la discusión en función de los antecedentes
6	CONCLUSIONES Y RECOMENDACIONES	9,00	Indicar si se cumplió o nó la idea a defender
7	DEFENSA, ARGUMENTACIÓN Y VOCABULARIO PROFESIONAL	8,00	
8	FORMATO, ORGANIZACIÓN Y CALIDAD DE LA INFORMACIÓN	9,00	revisión normas APA, ortografía, márgenes, revisar títulos, tablas

Obtenido una nota de: 8,90 Por lo tanto, **APRUEBA** ; debiendo el o los investigadores acatar el siguiente artículo:

Art. 36.- De los estudiantes que aprueban el informe final del TIC con observaciones.- Los estudiantes tendrán el plazo de 10 días para proceder a corregir su informe final del TIC de conformidad a las observaciones y recomendaciones realizadas por los miembros del Tribunal de sustentación de la pre-defensa.

Para constancia del presente, firman en la ciudad de Tulcán el 15 de febrero de 2023

MSC. GEORGINA GUADALUPE ARCOS PONCE
PRESIDENTE TRIBUNAL

MSC. JORGE HUMBERTO MIRANDA REALPE
DOCENTE TUTOR

MSC. JEFFERY ALEX NARANJO CEDEÑO
DOCENTE

Anexo 2. Certificado del abstract por parte de idiomas

UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI
FOREIGN AND NATIVE LANGUAGE CENTER

ABSTRACT- EVALUATION SHEET

NAME: Chugá Burbano Kevin Anderson

DATE: 17 de febrero de 2023

TOPIC: "Minería de Datos para mejorar los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi en el año 2022"

MARKS AWARDED **QUANTITATIVE AND QUALITATIVE**

VOCABULARY AND WORD USE	Use new learnt vocabulary and precise words related to the topic	Use a little new vocabulary and some appropriate words related to the topic	Use basic vocabulary and simplistic words related to the topic	Limited vocabulary and inadequate words related to the topic
	EXCELLENT: 2 ■	GOOD: 1Vera Játiva Edwin Andrés,5 ☐	AVERAGE: 1 ☐	LIMITED: 0,5 ☐
WRITING COHESION	Clear and logical progression of ideas and supporting paragraphs.	Adequate progression of ideas and supporting paragraphs.	Some progression of ideas and supporting paragraphs.	Inadequate ideas and supporting paragraphs.
	EXCELLENT: 2 ■	GOOD: 1,5 ☐	AVERAGE: 1 ☐	LIMITED: 0,5 ☐
ARGUMENT	The message has been communicated very well and identify the type of text	The message has been communicated appropriately and identify the type of text	Some of the message has been communicated and the type of text is little confusing	The message hasn't been communicated and the type of text is inadequate
	EXCELLENT: 2 ■	GOOD: 1,5 ☐	AVERAGE: 1 ☐	LIMITED: 0,5 ☐
CREATIVITY	Outstanding flow of ideas and events	Good flow of ideas and events	Average flow of ideas and events	Poor flow of ideas and events
	EXCELLENT: 2 ☐	GOOD: 1,5 ■	AVERAGE: 1 ☐	LIMITED: 0,5 ☐
SCIENTIFIC SUSTAINABILITY	Reasonable, specific and supportable opinion or thesis statement	Minor errors when supporting the thesis statement	Some errors when supporting the thesis statement	Lots of errors when supporting the thesis statement
	EXCELLENT: 2 ☐	GOOD: 1,5 ■	AVERAGE: 1 ☐	LIMITED: 0,5 ☐

TOTAL/AVERAGE	9 - 10: EXCELLENT 7 - 8,9: GOOD 5 - 6,9: AVERAGE 0 - 4,9: LIMITED	TOTAL 9

1

UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI FOREIGN AND NATIVE LANGUAGE CENTER

Informe sobre el Abstract de Artículo Científico o Investigación.

Autor: Chugá Burbano Kevin Anderson
Fecha de recepción del abstract: 17 de febrero de 2023

Fecha de entrega del informe: 17 de febrero de 2023

El presente informe validará la traducción del idioma español al inglés si alcanza un porcentaje de: 9 – 10 Excelente.

Si la traducción no está dentro de los parámetros de 9 – 10, el autor deberá realizar las observaciones presentadas en el ABSTRACT, para su posterior presentación y aprobación.

Observaciones:

Después de realizar la revisión del presente abstract, éste presenta una apropiada traducción sobre el tema planteado en el idioma Inglés. Según los rubrics de evaluación de la traducción en Inglés, ésta alcanza un valor de 9, por lo cual se validad dicho trabajo.

Atentamente

EDISON ROANERGES PEÑAFIEL ARCOS

Ing. Edison Peñafiel Arcos MSc
Coordinador del CIDEN

2

Anexo 3. Informe de Originalidad de Turnitin

Proyecto de Investigación

INFORME DE ORIGINALIDAD

4%	%	4%	%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	A. J. O. Reyes, A. O. Garcia, Y. L. Mué. "System for Processing and Analysis of Information Using Clustering Technique", IEEE Latin America Transactions, 2014 Publicación	1 %
2	Yadira Robles Aranda, Anthony R. Sotolongo. "Integración de los algoritmos de minería de datos 1R, PRISM E ID3 A POSTGRESQL", Journal of Information Systems and Technology Management, 2013 Publicación	<1 %
3	Edwin Gerardo Acuña Acuña. "Aplicación de minería de datos e Internet de las Cosas (IoT) para Productos Biomédicos", TECHNO REVIEW. International Technology, Science and Society Review /Revista Internacional de Tecnología, Ciencia y Sociedad, 2023 Publicación	<1 %
4	Eloy Albaladejo Gutierrez, E. Pérez, J. C. Espín Jaime. "Elaboración de un modelo de herramienta de gestión (Ficha de Transición)	<1 %

Anexo 4. Certificado de Conformidad

República del Ecuador

Ministerio de Turismo

Tulcán, 24 enero 2023.

CERTIFICACIÓN

Ing. Diego X. García Mera, Responsable de la Oficina Técnica del Carchi Coordinación Zonal 1 - Ministerio de Turismo,

CERTIFICO:

QUE la Señor, Chugá Burbano Kevin Anderson, con cédula de ciudadanía 0402046874 estudiante de la carrera de Computación, de la Universidad Politécnica Estatal del Carchi, finalizo el proyecto de investigación denominado "Minería de datos para mejorar los procesos de control de la demanda turística en el ministerio de Turismo de la provincia del Carchi", mismo que se ha realizado con todo lo solicitado por el Ministerio de Turismo.

En tal virtud me permito agradecer al estudiante Chugá Burbano Kevin Anderson con CI: 0402046874 por el trabajo realizado en este proyecto alcanzando los objetivos propuestos.

El Señor, Chugá Burbano Kevin Anderson, puede hacer uso de la presente certificación en la forma que convenga a sus intereses personales, menos para trámites legales.

Particular que pongo en su conocimiento, para los fines académicos correspondientes

Atentamente,

MINISTERIO DE TURISMO Gestión y Promoción Turística ecuador CARCHI COORDINACIÓN ZONAL 1

Ing. Diego X. Garcia Mera
Responsable OT- Carchi
MINISTERIO DE TURISMO

Dirección: Av. Gran Colombia N11-165 y Gral. Pedro Briceño
Código postal: 170403 / Quito - Ecuador
Teléfono: 593 2 399 9333 - www.turismo.gob.ec

Gobierno del Encuentro | Juntos lo logramos

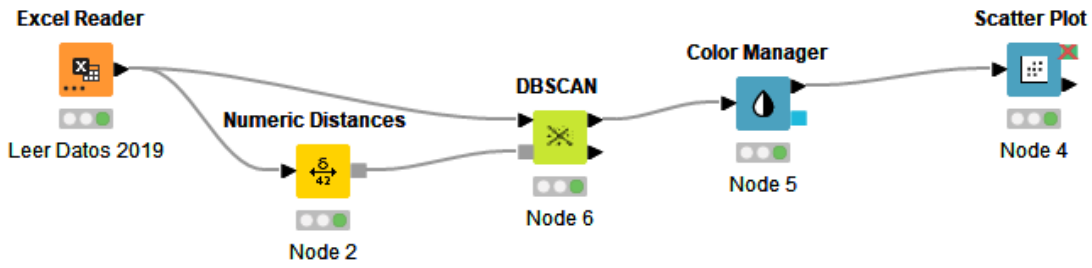
Anexo 5. Glosario de Terminología acerca de Minería de Datos

- Aprendizaje supervisado: un aprendizaje supervisado requiere datos de entra y salida con etiquetas, durante una fase de entrenamiento de datos, se le llama aprendizaje supervisado por qué parte del modelo que se cree requiere de supervisión humana, para etiquetar los datos.
- Aprendizaje no supervisado: es un modelo de datos que no se necesita procesar y realizar una etiqueta a los datos, como su nombre lo indica no requiere mucha intervención humana.
- Algoritmos: es un conjunto de heurísticas y cálculos, que permiten crear un modelo tomando como base la información que se otorga; para generar un modelo, el algoritmo utilizado inicialmente realiza un análisis de los datos proporcionados, con el fin de obtener patrones o tendencias, dependiendo del algoritmo utilizado.
- Data Mining: procesos de búsqueda y análisis en una gran cantidad de información con la finalidad de encontrar un patrón en una base de datos dispersa para conocimiento útil.
- KDD: descubrimiento de conocimiento en una gran cantidad de datos con potencial de utilidad con el fin de transformar la información de bajo nivel en conocimiento de alto nivel.
- Datawarehouse: es un gran almacén de datos, generalmente utilizado en empresas y organizaciones, con la finalidad de mantener los datos seguros, fiables, y tener la capacidad de administrar y gestionar de mejor manera toda la información.
- Data cleansing: también llamada limpieza de datos realiza la acción de depurar, identificar y realizar una corrección de erros en un conjunto de datos sin procesar.
- Modelo: un modelo en minería de datos en un conjunto de datos, estadísticas y patrones que pueden ser aplicados en un conjunto de datos para realizar predicciones y deducciones relaciones, dependiendo de los resultados que se quiera obtener al aplicar un modelo.
- SGBD: son aplicaciones o software que permiten administra un base de datos, algunas de las funciones que puede hacer es modificar, extraer, almacenar grandes cantidades de datos, entre otros.

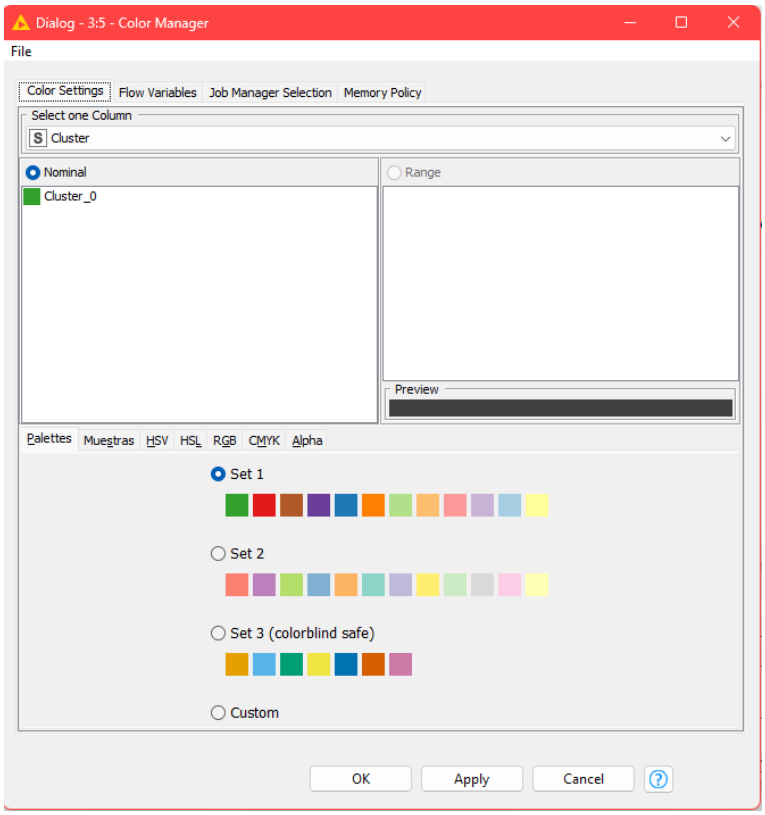
- K-NN: es un algoritmo clasificador, de aprendizaje supervisado, utiliza distancias de proximidad para clasificar los datos con el fin de generar predicciones o descubrir nuevos patrones de conocimiento.
- K-means: es un algoritmo de segmentación o clustering, y el objetivo que tiene es agrupar objetos en k grupos tomando en cuenta las características de la base de información proporcionada.
- Árbol de decisiones: algoritmos estadísticos, que permiten la creación de modelos predictivos para el análisis de datos tomando en cuenta características particulares con la relación entre variables dentro del a base de datos.
- Agrupamiento clustering: técnica de aprendizaje no supervisado, en la que, mediante un algoritmo de agrupamiento, realiza grupos que presentan características similares de un conjunto de datos.
- Varianza ANOVA: formula estadística, comúnmente utilizada para comparar varianzas entre los promedios de diferentes grupos.
- Clustering: principal técnica de modelado de minería de datos consiste en dividir información en distintos grupos e internamente los objetos de cada grupo presentan características similares unos de otros.
- CRISP-DM: metodo probado para orientar proyectos relacionas a la mineria de datos, proporciona una descripcion normalizada del cliclo de vida de un proyecto estandarizado de análsis de datos.
- Data Science: ligado a la inteligencia artificial, tiene la capacidad de analizar grandes cantidades de datos, con el fin de encontrar patrones, realizar pronósticos y principalmente tomar decisiones.
- OLAP: utilizado en ámbitos empresariales, es una tecnología de software que tiene la capacidad de hacer un análisis de datos tomando en cuenta diferentes criterios.
- Interpolación: tiene la capacidad de predecir valores a partir de una cantidad de dato limitada, poder ser utilizada para predecir valores que se desconozcan como: precipitaciones, concentraciones químicas, niveles de ruido, entre otros.
- Predicción secuencial: son procesos mediante los cuales se obtiene relaciones entre ocurrencias secuenciales con el fin de encontrar un orden específico en el que ocurren eventos, todo esto a partir de un conjunto de datos.

Anexo 6. Resultados del algoritmo DBSCAN en Knime Analytics

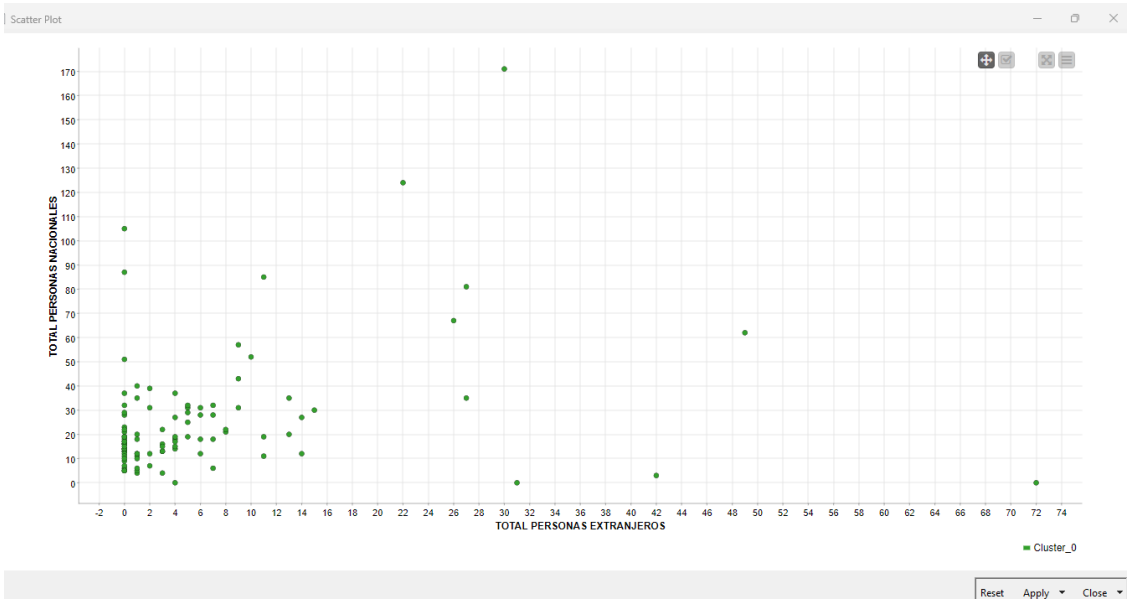
Se muestra inicialmente el modelo de agrupamiento o clustering generado tomando como base de información los datos de los procesos de alojamiento y gasto turístico.



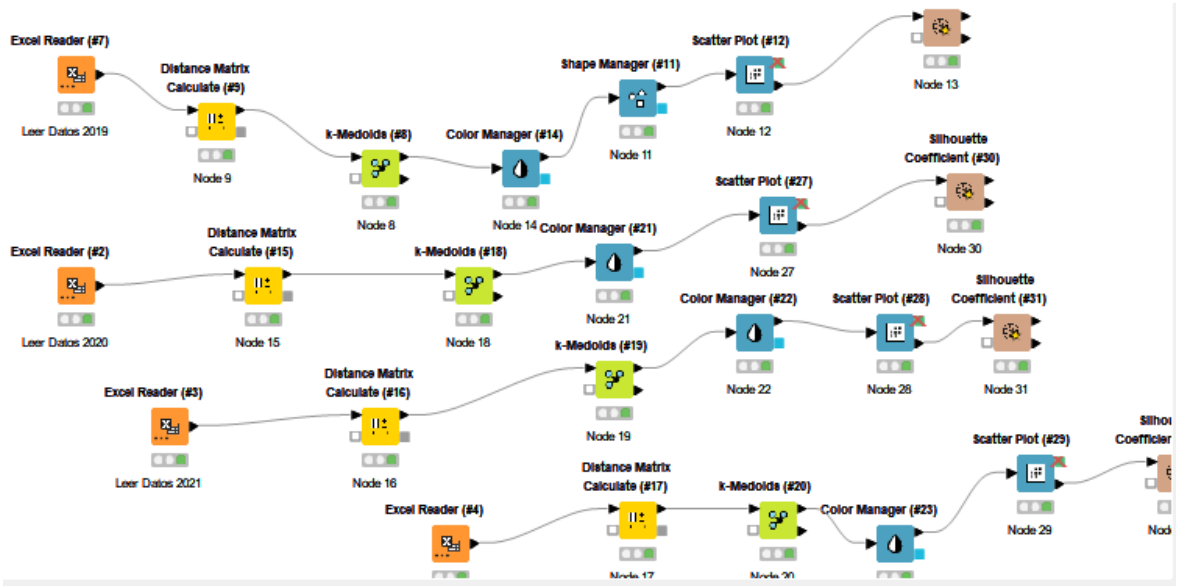
Generación de los clusters

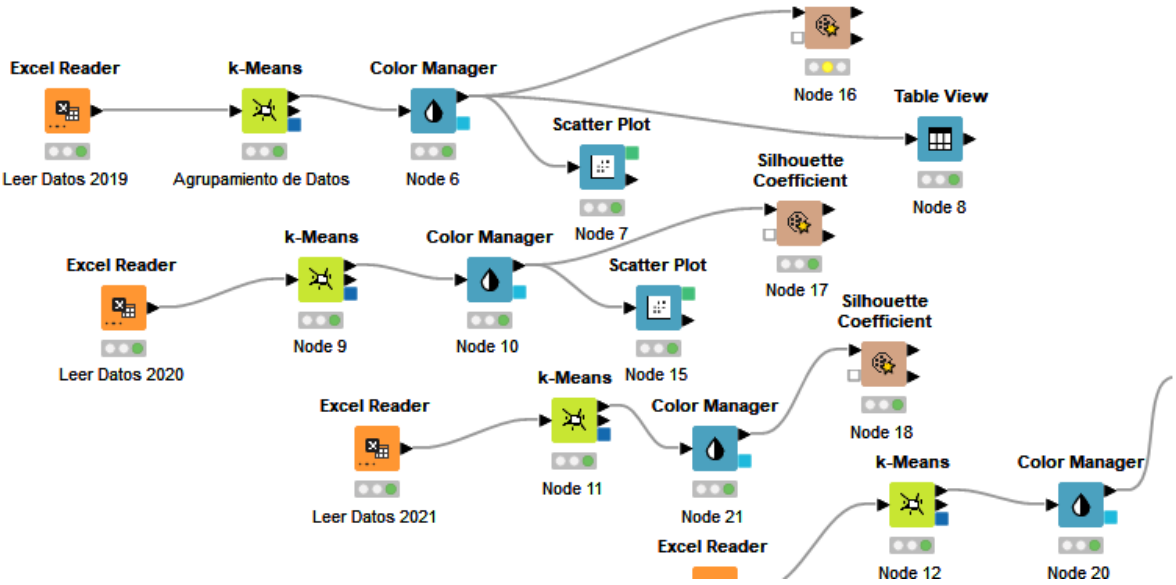


Por otra parte, se muestra la gráfica de resultados del algoritmo generado con un solo clúster, donde no muestra información que pueda ser analizada y explorada.



Anexo 7. Modelo del Algoritmo K-medoids y K-means en Knime Analytics




Excel Reader
k-Means
Color Manager
Node 16
Table View
Leer Datos 2019
Agrupamiento de Datos
Node 6
Scatter Plot
Node 7
Node 8
Silhouette Coefficient
Excel Reader
k-Means
Color Manager
Scatter Plot
Node 17
Leer Datos 2020
Node 9
Node 10
Node 15
Silhouette Coefficient
k-Means
Excel Reader
Color Manager
Node 18
Node 11
Leer Datos 2021
Node 21
k-Means
Color Manager
Excel Reader
Node 12
Node 20

Anexo 8. Formato Entrevista aplicada al Analista de Desarrollo y Promoción Turística de la Zona N°1 del Ministerio de Turismo

UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI

FACULTAD DE INDUSTRIAS AGROPECUARIAS Y CIENCIAS AMBIENTALES

CARRERA DE COMPUTACIÓN

Entrevista dirigida al Analista de Desarrollo y Promoción Turística de la Zona N°1 del Ministerio de Turismo Ing. Adrian Camilo Quesada Morelo.

Apreciado colaborador reciba un cordial saludo de parte del estudiante de la carrera de Ingeniería en Computación de la UPEC. La presente entrevista forma parte de la investigación denominada Minería de Datos para mejorar los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi, por lo cual le solicitamos comedidamente y agradecemos su disposición para completar cada uno de los ítems planteados.

Objetivo de la Entrevista: La presente entrevista tiene como finalidad la recolección de información correspondiente a los indicadores de la variables dependiente e independiente formuladas. Todos datos que se obtendrá harán referencia acerca de cómo se manejan los procesos de la demanda turística, y todo lo relacionado con esta actividad con la finalidad de facilitar la respuesta del entrevistado.

Toda la información que se obtenga será manejada con total confidencialidad, apoyando la realización de un ejercicio académico.

1. ¿Cree usted que los procesos que se maneja actualmente, donde se determina el nivel de turismo de la provincia del Carchi son organizados? ¿Por qué?

2. ¿Cómo se realiza el seguimiento (usando documentos, sistemas de gestión, ambas herramientas) de los procesos de alojamiento y gasto turístico?

3. ¿Existe un período específico donde se realizan los procesos de demanda turística de la provincia del Carchi?

4. ¿Qué tiempo toma la realización del proceso donde nos indica la demanda turística que ha tenido la provincia?

5. ¿Existen estándares o técnicas previamente establecidas en los procesos para determinar la demanda turística de la Provincia?

6. ¿Han existido inconvenientes en la realización del proceso de alojamiento y gasto turístico?

7. ¿Describa cómo se realiza actualmente los procesos para determinar el turismo que ha tenido la provincia?

8. ¿Actualmente los procesos para determinar la demanda turística tienen algún costo?

9. ¿Cuántas personas intervienen en el proceso de alojamiento y gasto turístico?

10. ¿Cómo ser realiza la interacción con los usuarios que intervienen en este proceso?

11. ¿Cómo es el cálculo para el proceso de la demanda turística y que herramienta utiliza para ello?

12. ¿Durante la ejecución de un proceso tuvo algún inconveniente que interrumpió el avance de actividades y aproximadamente cuánto tiempo fue de retraso?

13. ¿Actualmente cómo se almacena toda la información con respecto al proceso de demanda turística de la provincia?

14. ¿Cuál es el mecanismo de seguridad que utiliza para proteger la información?

15. ¿Cuál es el método que se emplea para respaldar los datos?

Anexo 9. Formato Encuesta aplicada a los sitios de alojamiento de la provincia del Carchi

UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI

FACULTAD DE INDUSTRIAS AGROPECUARIAS Y CIENCIAS AMBIENTALES

CARRERA DE COMPUTACIÓN

Encuesta dirigida a los Sitios de Alojamiento de la provincia del Carchi registrados en el Ministerio de Turismo de la Zona N°1.

1. El trato o actitud entre el Ministerio de Turismo hacia el usuario.
    a. Muy bueno
    b. Bueno
    c. Aceptable
    d. Malo
    e. Muy malo
2. La forma en que se realiza el proceso de alojamiento y gasto turístico.
    a. Muy bueno
    b. Bueno
    c. Aceptable
    d. Malo
    e. Muy malo
3. Como califica la agilidad o rapidez con que el MINTUR resuelve problemas acerca del proceso de alojamiento y gasto turístico.
    a. Muy bueno
    b. Bueno
    c. Aceptable
    d. Malo
    e. Muy malo
4. La interacción con el MINTUR a través de medios de comunicación (email, redes sociales, página web) le ayuda a resolver problemas con el proceso de alojamiento y gasto turístico.
    a. Siempre

b. Casi siempre
c. A veces
d. Casi nunca
e. Nunca

5. ¿Realiza frecuentemente preguntas o reclamos acerca del proceso de alojamiento y gasto turístico?
    a. Diario
    b. Semanal
    c. Mensual
    d. Trimestral
    e. Anual
6. ¿En el momento de realizar el proceso de alojamiento y gasto turístico ha tenido inconvenientes? Si tuvo algún problema indique cuál.
    a. Si
    b. No
7. ¿Aproximadamente cuánto tiempo emplea en realizar el proceso de alojamiento y gasto turístico?
    a. De 5 a 10 minutos.
    b. De 10 a 20 minutos.
    c. De 15 a 20 minutos.
    d. Más de 20 minutos.
8. ¿Cómo califica el proceso de alojamiento y gasto turístico que actualmente emplea el MINTUR?
    a. Muy bueno
    b. Bueno
    c. Aceptable
    d. Malo
    e. Muy malo
9. ¿Estaría de acuerdo con emplear herramientas tecnológicas con el fin de mejorar los procesos de alojamiento y gasto turístico?
    a. Muy de acuerdo
    b. De acuerdo
    c. Ni acuerdo ni en desacuerdo
    d. Desacuerdo

10. ¿Cree usted que la información que se recoge del proceso de la alojamiento y gasto turístico debe ser almacenada adecuadamente?

    a. Si
    b. No