UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI

POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

“Predicción del rendimiento académico en estudiantes universitarios con base en datos sociodemográficos.”

Trabajo de titulación previa la obtención del

Título de Magister en Estadística Aplicada

Autor(a): Kevin Andrés Chamorro Cupuerán

Tutor(a): Ph.D. Saba Rafael Infante Quirpa

Tulcán,enero 2024

CERTIFICADO DEL TUTOR

Certifico que el maestrante Chamorro Cupuerán Kevin Andrés con el número de cédula 1003586623 ha elaborado el trabajo de titulación: “Predicción del rendimiento académico en estudiantes universitarios con base en datos sociodemográficos”.

Este trabajo se sujeta a las normas y metodología dispuestas en la Codificación del Reglamento de Régimen Académico y de Estudiantes de la Universidad Politécnica Estatal del Carchi con RESOLUCIÓN No. 171-CSUP-2023, por lo tanto, autorizo su presentación para la sustentación respectiva.

Firmado electrónicamente por:
SABA RAFAEL INFANTE QUIRPA

f…….....

Ph.D. Infante Quirpa Saba Rafael

Tulcán, enero 2024

TUTOR

AUTORÍA DE TRABAJO

El presente trabajo de titulación constituye un requisito previo para la obtención del título de Magister en Estadística Aplicada.

Yo, Chamorro Cupuerán Kevin Andrés con cédula de identidad número 1003586623 declaro: que la investigación es absolutamente original, auténtica, personal y los resultados y conclusiones a los que he llegado son de mi absoluta responsabilidad.

KEVIN ANDRES CHAMORRO CUPUERAN

Firmado digitalmente por KEVIN ANDRES CHAMORRO CUPUERAN

f.....

Chamorro Cupuerán Kevin Andrés

AUTOR

Tulcán, enero 2024

ACTA DE CESIÓN DE DERECHOS DEL TRABAJO DE TITULACIÓN

Yo, Chamorro Cupuerán Kevin Andrés declaro ser autor/a de los criterios emitidos en el trabajo de titulación: “Predicción del rendimiento académico en estudiantes universitarios con base en datos sociodemográficos” y eximo expresamente a la Universidad Politécnica Estatal del Carchi y a sus representantes legales de posibles reclamos o acciones legales.

KEVIN ANDRES CHAMORRO CUPUERAN

Firmado digitalmente por KEVIN ANDRES CHAMORRO CUPUERAN

f......

Chamorro Cupuerán Kevin Andrés

AUTOR

Tulcán, enero 2024

AGRADECIMIENTO

Empiezo esta nueva etapa haciendo una pausa para reconocer a quienes han sido mi brújula y apoyo. Por supuesto, Dios ha sido esa fuerza espiritual que me ha impulsado, aun en los días más oscuros. Mi mamá, ese pilar inquebrantable, la mujer que, con su amor y sabiduría, me ha enseñado a creer en mí mismo. ¡Madre, no hay palabras suficientes para agradecerte todo lo que has hecho! A papá, con su apoyo constante y sus enseñanzas, siempre ha sido esa mano amiga dispuesta a ayudar. Mi hermana, con quien he compartido risas, sueños y algunos desafíos que nos han hecho aún más fuertes. Mis sobrinas, que, con su inocencia y alegría, me han recordado siempre por qué vale la pena luchar.

No puedo olvidarme de mi tutor, quien, con paciencia, conocimiento y dedicación, me ha guiado en este proceso. Su guía ha sido fundamental para alcanzar este punto. Y a esos amigos que, a lo largo de todo, han sido ese refugio y alivio, gracias por cada palabra de aliento y cada momento compartido.

DEDICATORIA

Dedico este triunfo, primero que todo, a Dios, quien me ha brindado tantas bendiciones y lecciones de vida. Pero, especialmente, a mi mamá. Ella, que ha sido esa mezcla perfecta de ternura, fortaleza y sabiduría. La que me ha mostrado que no hay barrera que no pueda superarse con amor y determinación. Mamá, este logro es tanto tuyo como mío. A mi padre, por su constante apoyo y por mostrarme que siempre hay una solución para cada problema. A mi hermana, por todas las risas, consejos y abrazos que hemos compartido. A mis sobrinas, por ser ese recordatorio constante de lo maravilloso que es vivir con curiosidad y asombro. Y a todos mis amigos, por todas las aventuras y momentos compartidos, por ser mi apoyo incondicional. Este logro lleva un pedacito de cada uno de ustedes.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

RESUMEN

Este estudio se centra en la predicción del rendimiento académico de estudiantes universitarios de primer semestre en la Universidad Yachay Tech, Ecuador, durante el periodo 2014-2023, fundamentándose en datos sociodemográficos proporcionados por la institución. La investigación es de enfoque cuantitativo y de tipo correlacional. Se utiliza la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD), con el objetivo de mejorar la precisión en la predicción del rendimiento académico, se incorporó la regresión logística junto con las técnicas de regularización Lasso y Ridge. Los resultados demuestran que los modelos predictivos basados en regresión logística con penalizaciones Lasso y Ridge, anticipan el rendimiento académico utilizando datos sociodemográficos. La potencialidad de tales modelos permite la identificación temprana de estudiantes, que podrían beneficiarse de intervenciones adicionales hasta la reestructuración de enfoques pedagógicos. Esta fusión de técnicas permitió identificar las variables sociodemográficas con el mayor impacto predictivo en el rendimiento académico. La combinación de estas estrategias impulsa la calidad de la Educación Superior en Ecuador y reduce las tasas de deserción académica.

Palabras clave: Rendimiento Académico, Regularización Lasso, Regularización Ridge, KDD.

ABSTRACT

This study analyzed data from first-semester students at Yachay Tech This study focuses on the prediction of the academic performance of first-semester university students at the Yachay Tech University, Ecuador, during the period 2014-2023, based on sociodemographic data provided by the institution. The research has a quantitative and correlational approach. The Knowledge Discovery in Databases (KDD) methodology is used, with the aim of improving the accuracy in predicting academic performance, logistic regression was incorporated along with the Lasso and Ridge regularization techniques. The results demonstrate that predictive models based on logistic regression with Lasso and Ridge penalties anticipate academic performance using sociodemographic data. The potential of such models allows for the early identification of students, who could benefit from additional interventions up to the restructuring of pedagogical approaches. This fusion of techniques allowed us to identify the sociodemographic variables with the greatest predictive impact on academic performance. The combination of these strategies boosts the quality of Higher Education in Ecuador and reduces academic dropout rates.

Keywords: Academic Performance, Lasso Regularization, Ridge Regularization, KDD

CAPÍTULO I

PROBLEMA

1.1. Planteamiento del problema

El rendimiento académico en la educación universitaria constituye un foco de interés global, involucrando a educadores, administradores, políticos y académicos en un diálogo continuo (Honicke Y Broadbent, 2016). Su importancia radica en la influencia que ejerce en el crecimiento personal y profesional de los estudiantes, en la reputación y avance de las instituciones educativas, y en el progreso de la sociedad (Honicke Y Broadbent, 2016). Diversos factores, incluyendo variables sociodemográficas como género, edad, nivel socioeconómico y origen étnico, contribuyen al rendimiento académico (Sirin, 2005). Aunque numerosos estudios han explorado estos factores, persiste la necesidad de investigar en esta área para desarrollar estrategias efectivas que prevengan el bajo rendimiento académico y mejoren la calidad de la educación superior.

A nivel internacional, el rendimiento académico en universidades ha sido objeto de estudio en diferentes contextos, destacando investigaciones en países como Estados Unidos, Alemania, y Australia, donde se han identificado factores clave como la enseñanza de calidad, el entorno de aprendizaje y el apoyo institucional como determinantes esenciales del éxito académico (Schneider Y Preckel, 2017). Esta perspectiva global subraya la universalidad del desafío de mejorar la educación superior y refuerza la necesidad de un enfoque contextualizado y basado en datos para abordar las particularidades de cada sistema educativo.

En Ecuador, la temática del rendimiento académico en la educación superior ha ganado relevancia en los últimos años. El país enfrenta desafíos considerables en

términos de acceso, equidad y calidad en este nivel educativo (Senescyt, 2021). Las instituciones de educación superior ecuatorianas, en paralelo con sus contrapartes internacionales, han emprendido esfuerzos para identificar y comprender los factores que afectan el rendimiento académico de sus estudiantes, con el objetivo de implementar políticas y estrategias que promuevan el éxito académico y la retención estudiantil (Villarruel-Meythaler *et al.*, 2020). Sin embargo, existe una brecha en la literatura científica respecto al análisis y la predicción del rendimiento académico basado en datos sociodemográficos en el contexto ecuatoriano.

La Universidad Yachay Tech, una entidad ecuatoriana enfocada en investigación e innovación enfrenta preocupaciones similares relacionadas con el rendimiento académico de sus estudiantes. Para esta institución, es imperativo identificar y abordar los factores que inciden en el rendimiento de sus estudiantes, especialmente durante los primeros semestres, donde existe una mayor propensión a experimentar dificultades académicas (Al Husaini Y Ahmad Shukor, 2023).

Ante las preocupaciones y desafíos descritos, el problema de investigación que este trabajo aborda se formula de la siguiente manera: ¿De qué manera es posible predecir el rendimiento académico en estudiantes universitarios de la Universidad Yachay Tech en Ecuador, utilizando datos sociodemográficos como variables predictoras?

1.2. Hipótesis

Existe una relación significativa entre los datos sociodemográficos de los estudiantes de primer semestre en la Universidad Yachay Tech y su rendimiento académico.

1.3. Objetivos de investigación

1.3.1 Objetivo General

Desarrollar y comparar modelos de regresión logística con penalizaciones Lasso y Ridge para predecir el rendimiento académico de los estudiantes de primer semestre en la Universidad Yachay Tech, utilizando factores sociodemográficos.

1.3.2 Objetivos Específicos

1. Analizar las variables sociodemográficas de los estudiantes de primer semestre en la Universidad Yachay Tech y su relación con el rendimiento académico, utilizando técnicas estadísticas descriptivas y analíticas.

2. Desarrollar modelos predictivos basados en regresión logística con penalizaciones Lasso y Ridge, y comparar su eficacia en la predicción del rendimiento académico de los estudiantes de primer semestre en la Universidad Yachay Tech a partir de datos sociodemográficos.

3. Evaluar y comparar los modelos de regresión logística con penalizaciones Lasso y Ridge mediante el análisis de métricas apropiadas, para determinar su eficacia en la predicción del rendimiento académico de los estudiantes de primer semestre en la Universidad Yachay Tech.

1.4. Justificación

La investigación sobre la predicción del rendimiento académico en estudiantes universitarios de primer semestre en la Universidad Yachay Tech en Ecuador, mediante el uso de datos sociodemográficos y métodos estadísticos avanzados, presenta una relevancia multifacética y sustancial. Este estudio es especialmente conveniente, ya que proporcionará información valiosa sobre los factores sociodemográficos que influyen en el rendimiento académico de los estudiantes universitarios. El conocimiento generado permitirá a la universidad implementar

políticas y programas adecuados para mejorar tanto el rendimiento como la retención estudiantil, contribuyendo significativamente a la mejora de la calidad de la educación superior en Ecuador.

Igualmente, el estudio tiene una trascendencia social notable. Al ayudar a reducir las tasas de abandono escolar y mejorar la preparación de los estudiantes para el mercado laboral, los resultados beneficiarán no solo a los estudiantes y sus familias, sino también a la sociedad en su conjunto. Un mejor rendimiento académico conduce a una mayor contribución al desarrollo del país, reforzando así el vínculo entre la educación y el progreso social.

Por otro lado, la investigación tendrá implicaciones prácticas importantes para el diseño de intervenciones y programas de apoyo. Podrá guiar el desarrollo de tutorías, apoyo emocional, orientación y asesoramiento en la Universidad Yachay Tech y otras instituciones de educación superior en Ecuador, facilitando la adaptación de los estudiantes a la vida universitaria y mejorando su rendimiento académico.

Asimismo, este estudio enriquecerá el conocimiento existente sobre los factores que influyen en el rendimiento académico, llenando un vacío en la literatura en el contexto ecuatoriano. Los resultados obtenidos podrían ser aplicables a principios más amplios, permitiendo a otros investigadores explorar y desarrollar teorías relacionadas en diferentes contextos y poblaciones.

Además, desde una perspectiva metodológica, el estudio promete ser de gran utilidad. Ayudará a desarrollar y aplicar métodos estadísticos avanzados para predecir el rendimiento académico a partir de datos sociodemográficos, mejorando así la comprensión de las relaciones entre variables sociodemográficas y el rendimiento académico. Este enfoque proporcionará ideas y recomendaciones valiosas para futuras investigaciones en contextos y poblaciones variadas.

Finalmente, el presente estudio de investigación aporta al Plan de Creación de Oportunidades 2021-2025 aprobado por el Consejo Nacional de Planificación con Resolución Nro. 002-2021-CNP, de 20 de septiembre de 2021 donde el aporte principal es al eje económico con el objetivo de tratar de mejor manera la deserción estudiantil. Además, el estudio está perfilada a la línea de investigación del Programa de Posgrado: “Aplicación de la Estadística en la solución de problemas del entorno”; a la que se adscribe la investigación, son para mejorar la educación de los estudiantes dentro del Ecuador, dado que empleará herramientas de análisis de datos e inteligencia artificial para predecir el rendimiento académico. Esto puede ayudar a identificar factores que contribuyen al éxito o fracaso académico, permitiendo a las instituciones educativas tomar decisiones más informadas y desarrollar intervenciones más efectivas. Al mejorar la eficacia y eficiencia de la educación, este proyecto contribuye al desarrollo sostenible, ya que la educación de calidad es uno de los Objetivos de Desarrollo Sostenible de las Naciones Unidas. Además, al explorar el papel de los factores sociodemográficos, la tesis también se alinea con los principios de equidad e inclusión social, que son componentes clave de la sostenibilidad

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

2.1. Antecedentes de investigación

En relación con los antecedentes metodológicos de la presente investigación, es importante destacar investigaciones previas que hayan empleado y comparado diferentes métodos estadísticos y algoritmos de predicción para estudiar el rendimiento académico.

Marbouti *et al.* (2016) llevaron a cabo un estudio en Estados Unidos en colaboración con la Universidad de Purdue. Investigaron la capacidad predictiva de diferentes algoritmos de aprendizaje automático, como árboles de decisión, regresión logística, máquinas de vectores de soporte y redes neuronales, para predecir el rendimiento académico de estudiantes universitarios en cursos de ingeniería. Los resultados mostraron que la regresión logística y las máquinas de vectores de soporte proporcionaron la mejor precisión en la predicción del rendimiento académico, mientras que los árboles de decisión y las redes neuronales tuvieron un desempeño inferior.

Kotsiantis *et al.* (2004) realizaron un estudio en Grecia en colaboración con la Universidad de Patras. Compararon la efectividad de varios algoritmos de aprendizaje automático, como árboles de decisión, regresión logística, k vecinos más cercanos y redes neuronales, para predecir el rendimiento académico de estudiantes universitarios. Los autores concluyeron que los árboles de decisión y la regresión logística mostraron un mejor desempeño en términos de precisión y comprensibilidad de los modelos generados.

Yukselturk Y Top (2013) llevaron a cabo un estudio en Turquía en colaboración con la Universidad de Anadolu. Compararon diferentes métodos estadísticos y de aprendizaje automático, como análisis discriminante, regresión logística y árboles de decisión, para predecir el rendimiento académico de estudiantes universitarios en cursos en línea. Los resultados indicaron que la regresión logística fue el enfoque más efectivo para predecir el rendimiento académico en este contexto.

Gutiérrez-Monsalve *et al.* (2021) en su investigación utilizó un análisis discriminante y una regresión logística con datos administrativos de la universidad. El análisis discriminante permitió clasificar el 100% de los estudiantes con bajo rendimiento académico, principalmente basado en variables institucionales y sociodemográficas. La regresión logística encontró asociaciones significativas del bajo RA con la trayectoria del estudiante, becas, repitencia y asignaturas canceladas.

Thiele *et al.* (2016) en su estudio analizó datos obtenidos de repositorios de la universidad donde se incluyeron estudiantes de programas de tres años a tiempo completo, y se estratificaron los datos por año de ingreso. Se realizaron análisis univariables y multivariables utilizando regresión logística para explorar las relaciones entre características de antecedentes, desempeño académico y factores contextuales, utilizando el software SPSS.

Contreas-Bravo *et al.* (2023) en su investigación utilizó análisis de datos y herramientas de aprendizaje automático para predecir el rendimiento académico de los estudiantes. Se seleccionaron variables académicas, demográficas y sociodemográficas mediante métodos de selección de variables. Se implementaron algoritmos de aprendizaje automático como Árboles de Decisión, KNN, SVC, Naive Bayes y LDA, y se encontró que el KNN es el modelo que mejor predice el rendimiento académico.

En Ecuador, también se han realizado investigaciones que emplean métodos estadísticos y algoritmos de predicción para analizar el rendimiento académico de estudiantes universitarios.

(Loja Rodas, 2019) en su estudio empleó técnicas de minería de datos y la metodología CRISP-DM para predecir la deserción, el fracaso académico por ciclo de estudios y por asignatura en la Universidad de Cuenca. Para los modelos obtenidos, se usó validación cruzada y curvas de aprendizaje para evitar el sobreajuste y garantizar su eficacia en nuevos registros. A través de la aplicación de técnicas de sub-muestreo y clusterización, y un proceso de reducción de dimensionalidad de los datos, se obtuvo una precisión de al menos el 70% en todas las métricas, llegando hasta el 83% para la predicción de fracaso en asignaturas específicas. El trabajo también incluyó la recopilación, limpieza y transformación de los datos, y la implementación de dos algoritmos de clasificación: j48 y Naive Bayes.

Calva Yaguana (2020) llevó a cabo un estudio en una universidad pública de Ecuador para analizar los factores asociados al rendimiento académico de estudiantes universitarios, utilizando técnicas de minería de datos como árboles de decisión y regresión logística. La autora identificó variables demográficas, socioeconómicas y académicas, como género, edad, tipo de colegio previo, puntaje en el examen de ingreso y situación socioeconómica, que influían significativamente en el rendimiento de los estudiantes, proporcionando información valiosa sobre variables relevantes y métodos de análisis en el contexto ecuatoriano.

Guambuguete Rea (2023) en su estudio buscó implementar un modelo matemático utilizando regresión logística multivariante para identificar los factores socioculturales asociados a las calificaciones escolares en estudiantes del primer ciclo del Instituto Superior Tecnológico Tres de Marzo de la Provincia Bolívar en Ecuador.

Se trabajó con una base de datos de 489 estudiantes y se encontró que factores como el nivel de formación de la madre, la cantidad de miembros en el hogar y el estado civil en unión libre incrementan la probabilidad de reprobar el ciclo académico. El modelo predice el 84.9% de los casos posibles.

Los estudios mencionados anteriormente abordan la correlación entre los datos sociodemográficos y el rendimiento académico de los estudiantes universitarios, haciendo uso de una amplia gama de técnicas estadísticas y de aprendizaje automático. No obstante, una tendencia emergente en esta área de investigación consiste en la aplicación de métodos de regresión penalizada, como la regresión Lasso y Ridge. Estas técnicas modernas están siendo utilizadas para examinar y predecir el rendimiento académico de los estudiantes universitarios, con los datos sociodemográficos como variables de entrada.

Wang *et al.* (2015) junto con un equipo de colaboradores, llevó a cabo un estudio en Dartmouth College en Estados Unidos, utilizando smartphones como herramientas para recopilar una serie de indicadores de comportamiento estudiantil. Estos datos recogidos fueron analizados a través del método de regresión LASSO (Least Absolute Shrinkage and Selection Operator), un algoritmo de aprendizaje automático reconocido por su capacidad para realizar la selección de variables y la regularización en modelos de regresión. Como resultado, Wang *et al.* descubrieron correlaciones significativas entre ciertos comportamientos estudiantiles y su rendimiento académico, lo que demuestra la aplicabilidad de LASSO en este ámbito de estudio. Este trabajo representa un valioso aporte a nuestra investigación, proporcionando un enfoque para recoger y analizar datos de comportamiento para predecir el rendimiento académico, así como una herramienta útil en la regresión LASSO para identificar características relevantes en nuestra propia base de datos.

El trabajo de investigación realizado por Robinson *et al.* (2016) se llevó a cabo en instituciones académicas como Harvard University, MIT, University of Virginia y University of California, Santa Barbara. Utilizaron el enfoque del procesamiento de lenguaje natural (NLP) junto con la técnica de regresión logística regularizada conocida como LASSO (Least Absolute Shrinkage and Selection Operator). Según los resultados obtenidos, el modelo de NLP basado en el análisis de respuestas en texto sin estructura demostró ser eficaz en la predicción de qué estudiantes completarían un curso en línea. En comparación con los modelos de referencia basados en variables demográficas, el modelo NLP superó las predicciones basadas en datos demográficos. Además, se encontró que los predictores demográficos y los predictores basados en NLP ofrecían información complementaria, mejorando aún más la capacidad de predicción del modelo combinado.

Ma *et al.* (2017) realizaron un estudio comparativo sobre diferentes métodos de selección de características para predecir el rendimiento de los estudiantes en un entorno de aprendizaje en línea. Según los resultados del estudio, LASSO y Ridge fueron eficaces para la selección de características relevantes en la predicción del rendimiento estudiantil. Estas técnicas de regularización demostraron ser herramientas valiosas para controlar la multicolinealidad y mejorar la interpretabilidad de los modelos, lo que contribuye a un mejor entendimiento de los factores que influyen en el rendimiento académico y ayuda a mejorar la toma de decisiones en la educación en línea.

Glavas *et al.* (2018) trabajaron en el uso de varios métodos de selección de modelos para el análisis de grandes conjuntos de datos con muchos predictores, en particular en el campo de la minería de datos educativos. El estudio comparó varios métodos de regresión, incluidos la regresión de LASSO y la regresión Ridge, y

encontró que la selección de subconjuntos proporcionaba los mejores resultados para su conjunto de datos específico. Los autores concluyeron que la regresión de LASSO, que elimina los predictores no importantes del modelo, proporcionó mejores resultados que la regresión Ridge. Esta investigación aporta al campo de la minería de datos educativos al demostrar la eficacia de la regresión de LASSO y la selección de subconjuntos para la predicción del rendimiento estudiantil. Sin embargo, también señalan que los métodos efectivos pueden variar dependiendo del conjunto de datos y la cantidad de predictores, y sugieren que futuras investigaciones podrían beneficiarse de la inclusión de predictores más precisos.

El estudio de Al Sheeb *et al.* (2019) se centró en el desarrollo de un modelo general linealizado para predecir el GPA (Grade Point Average) académico de los estudiantes de primer año de ingeniería en la Universidad de Qatar, en Doha-Qatar. Este estudio utilizó dos técnicas de regresión penalizadas, LASSO y Ridge, de forma similar a otros métodos existentes como el doble LASSO y la red elástica. La calidad de la predicción de la técnica de regresión de doble penalización fue evaluada utilizando un conjunto de factores no cognitivos recogidos de los estudiantes de primer año, y el error cuadrático medio (MSE) fue utilizado para evaluar su rendimiento. Aunque la metodología propuesta mostró un notable rendimiento en la predicción del GPA del estudiante, los autores destacan que se requieren más investigaciones para encontrar un criterio de selección que reemplace al MSE en la selección del resultado de la primera etapa. Además, recomiendan el uso de la Validación Cruzada cuando la muestra es suficientemente grande.

2.2. Marco teórico

2.2.1 Datos sociodemográficos

Los datos sociodemográficos son aquellos que describen las características sociales y demográficas de un individuo o una población (Nations, 2017). Estos datos pueden incluir información sobre edad, género, etnia, nacionalidad, educación, ocupación, ingresos, estado civil, tamaño de la familia, lugar de residencia y otros aspectos relacionados con el contexto social y cultural (Krieger, 2011). Estos datos son fundamentales para el estudio de las dinámicas sociales, económicas y culturales de las poblaciones y para el diseño y evaluación de políticas públicas (Krieger, 2011)

Los datos sociodemográficos relacionados con el rendimiento académico hacen referencia a las características de los estudiantes que pueden influir en su desempeño educativo, como el género, la edad, el nivel socioeconómico, el entorno familiar, la educación y ocupación de los padres, el tamaño de la familia y el lugar de residencia (González Medina *et al.*, 2018). El estudio de estos datos permite analizar diferencias y desigualdades en el desempeño educativo entre diversos grupos, e identificar factores que contribuyen a estas diferencias. El análisis es esencial para diseñar políticas educativas y programas de intervención que mejoren el rendimiento y reduzcan brechas educativas (González Medina *et al.*, 2018).

La relación entre los datos sociodemográficos y el rendimiento académico se investiga a través de métodos y análisis estadísticos que describen, comparan y evalúan el impacto de estos factores en el desempeño de los estudiantes (Rodríguez López *et al.*, 2018). Es importante considerar que la influencia de los datos sociodemográficos en el rendimiento académico puede variar según el contexto educativo y cultural, y que su impacto puede estar mediado por otros factores, como

las características individuales de los estudiantes y la calidad de la enseñanza (Garbanzo Vargas, 2013).

Además, la utilización de datos sociodemográficos en la investigación educativa implica consideraciones éticas significativas. Es esencial garantizar la privacidad y la protección de los datos personales de los estudiantes, así como utilizar estos datos de manera responsable y con fines constructivos. La interpretación de los datos debe realizarse con cautela, evitando estereotipos o conclusiones simplistas que no consideren la complejidad de los factores socioculturales y educativos (Rodríguez López et al., 2018).

Por otro lado, la aplicación práctica de los hallazgos derivados del análisis de datos sociodemográficos puede incluir la personalización de la enseñanza, el desarrollo de programas de apoyo específicos para ciertos grupos demográficos, y la creación de entornos educativos más inclusivos y equitativos. Estas aplicaciones tienen el potencial de transformar significativamente la experiencia educativa, asegurando que todos los estudiantes tengan las mismas oportunidades de éxito académico, independientemente de su origen o circunstancias personales (González Medina *et al*., 2018).

Finalmente, la investigación en este campo también debe ser dinámica y adaptativa, considerando las constantes evoluciones sociales y culturales que pueden influir en la relación entre los datos sociodemográficos y el rendimiento académico. Así, el estudio continuo y la actualización de estas variables son cruciales para mantener la relevancia y efectividad de las intervenciones y políticas educativas.

2.2.2 Rendimiento académico

El rendimiento académico es una medida crucial en la educación que refleja la competencia y el dominio de un estudiante en un tema o conjunto de temas. Según

(Kuh *et al.*, 2006), se refiere a la evaluación del conocimiento adquirido en el ámbito educativo, especialmente en entornos escolares y universitarios.

La medición del rendimiento académico puede ser compleja y se realiza a través de varios métodos. Comúnmente, se evalúa mediante exámenes, pruebas, tareas, proyectos y participación en clase. Las calificaciones numéricas o alfabéticas son la forma más común de representar el rendimiento académico, aunque también se pueden utilizar métodos más cualitativos como la evaluación por pares y la autoevaluación (Schneider Y Preckel, 2017).

En la educación superior, el rendimiento académico tiene una importancia vital, ya que refleja la eficacia de la enseñanza y el aprendizaje. Es un indicador clave de la calidad de la educación y puede influir en las oportunidades de empleo y el éxito futuro del estudiante (Robinson *et al.*, 2016). Además, es esencial para determinar la admisión a programas especializados, becas y oportunidades de empleo (Pascarella Y Terenzini, 2005).

Las instituciones de educación superior también utilizan el rendimiento académico para evaluar y mejorar sus métodos de enseñanza, lo que a su vez puede llevar a una mejor calidad de la educación (Bok, 2020). En resumen, el rendimiento académico es un concepto central en la educación que tiene implicancias profundas en la vida de los estudiantes y en la calidad de la educación en general. Su medición y comprensión son fundamentales para el avance y la mejora continua en el ámbito educativo (Tinto, 2012).

2.2.3 Análisis de regresión

El análisis de regresión es una técnica estadística utilizada para investigar las relaciones entre variables y predecir el valor de una variable (variable dependiente) en función de otra(s) variable(s) (variable(s) independiente(s)) (Pardo Y Ruiz, 2005). En el contexto del rendimiento académico, el análisis de regresión permite estudiar cómo diferentes factores, como los datos sociodemográficos, las características individuales y el contexto educativo, influyen en el desempeño de los estudiantes en términos de calificaciones, resultados en pruebas estandarizadas u otros indicadores de rendimiento (Navarro Y García, 2011).

El análisis de regresión puede ser de diferentes tipos, como la regresión lineal simple, la regresión lineal múltiple, la regresión logística y la regresión de Poisson, entre otros, dependiendo de la naturaleza de las variables y la relación que se pretende estudiar (Peña, 2011). Además, el análisis de regresión puede incluir técnicas como la selección de variables, la transformación de variables y la detección de interacciones para mejorar la precisión y la interpretación de los resultados (Hosmer, Lemeshow, Y Sturdivant, 2013).

El análisis de regresión en el estudio del rendimiento académico permite identificar los factores que tienen mayor impacto en el desempeño educativo y cuantificar la magnitud de su efecto (Gutiérrez Y Gento, 2017). Esto es esencial para el diseño de políticas y programas educativos basados en evidencia, así como para la identificación de áreas de intervención y mejora en el proceso de enseñanza-aprendizaje (Sánchez, 2014).

2.2.4 Regresión Lineal

La regresión lineal es un modelo estadístico fundamental que intenta describir la relación entre una variable dependiente, Y, y una a más variables independientes o predictoras, X (Jr *et al.*, 2013).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \epsilon$$

Donde:

- Y : Variable respuesta o dependiente:
- $X_1, X_2, \ldots, X_P$: Son las variables predictoras o independientes.
- β_0 : Es el término de intercepción.
- $\beta_1, \beta_2, \ldots, \beta_{\mathrm{p}}$: Son los coeficientes asociados a cada variable predictora.
- ϵ : Es el error aleatorio

La idea detrás de la regresión lineal es simple intentamos encontrar una "línea" que mejor se ajuste a nuestros datos. Esta "línea" es en realidad un hiperplano en un espacio de dimensiones superiores.

La posición y orientación de este hiperplano se determinan mediante los coeficientes β.

Para estimar estos coeficientes, se utiliza el método de mínimos cuadrados. Esta técnica tiene como objetivo minimizar la suma de los cuadrados de las diferencias (residuos) entre las valares observadas de Y y los valores predichos por el modelo.

Residuo: El residuo para una observación particular, i, es la differencia entre el valor observado, y_i , y valor predicho, $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i$$

Función de Coste: La función de coste, $J(\beta)$, es la suma de los cuadrados de los residuos:

$$J(\beta) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

El objetivo es encontrar los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ que minimicen esta función de coste. Para ello, se toman las derivadas parciales con respecto a cada β_j y se igualan a cero, lo que da lugar a las ecuaciones normales. Estas ecuaciones se resuelven para obtener las estimaciones de los coeficientes.

Las ecuaciones normales son un conjunto de ecuaciones algebraicas que se derivan del proceso de minimizar la suma de los cuadrados de los residuos en la regresión lineal, utilizando el método de mínimos cuadrados. Las ecuaciones normales proporcionan una forma de resolver explícitamente los coeficientes estimados del modelo lineal:

$$Y = X\beta + \epsilon$$

Donde:

- Y : Es el vector de respuesta.
- X : Es la matriz de diseño, que contiene las observaciones de las variables predictoras. Tiene una columna adicional de unos para el término de intercepción.
- β : Es el vector de coeficientes desconocidos.
- ϵ : Es el vector de errores.

La función de coste, que es la suma de los cuadrados de los residuos, se define como:

$$J(\beta) = (Y - X\beta)^T (Y - X\beta)$$

Para minimizar $J(\beta)$ con respecto a β, tomamos la derivada de $J(\beta)$ con respecto a β y la igualamos a cero:

$$\frac{\partial J(\beta)}{\partial \beta} = 0$$

Esto da lugar a las ecuaciones normales:

$$X^T X\beta = X^T Y$$

Para resolver β, simplemente multiplicamos ambos lados por la inversa de $X^T X$ (asumiendo que $X^T X$ es no singular y, por lo tanto, invertible):

$$\beta = (X^T X)^{-1} X^T Y$$

Este es el estimador de mínimos cuadrados de los coeficientes. En la práctica, en lugar de calcular explícitamente la inversa, las técnicas numéricas como la factorización LU o QR se utilizan para resolver estas ecuaciones de manera eficiente.

2.2.5 Modelo Lineal Generalizado (GLM)

Los Modelos Lineales Generalizados (GLM) son una extensión de los modelos lineales tradicionales, como la regresión lineal. Los GLM permiten modelar relaciones donde la respuesta no sigue necesariamente una distribución normal y donde la

relación entre la respuesta y los predictores no es estrictamente lineal. Un GLM se compone de tres componentes esenciales:

- **Componente Aleatorio:** Esto se refiere a la distribución de la variable respuesta. En el GLM, se asume que la respuesta sigue una distribución perteneciente a la familia exponencial. Esta familia incluye muchas de las distribuciones comunes como la normal, binomial, Poisson, gamma, entre otras. Si denotamos Y como nuestra variable respuesta, entonces en un GLM, la distribución de Y dada X pertenece a la familia exponencial.
- **Componente Sistemático:** Este componente describe la relación lineal entre la respuesta esperada y las variables predictoras. Es similar al componente sistemático en la regresión lineal. Se denota como:

$$\eta = X\beta$$

Donde:

- η es el predictor lineal.
- X es la matriz de diseño que contiene las observaciones de las variables predictoras.
- β es el vector de coeficientes.

- **Función de Enlace, g** : La función de enlace proporciona el vínculo entre el componente aleatorio y el componente sistemático. Es una transformación continua que transforma la respuesta esperada en el predictor lineal. Matemáticamente, se denota como:

$$g(E[Y]) = \eta$$

o, equivalentemente,

$$E[Y] = g^{-1}(\eta)$$

Donde:

- $E[Y]$: Es el valor esperado de la respuesta.
- g^{-1}: Es la función de enlace inversa.
  Por ejemplo, en la regresión logística, la función de enlace es el logit, y la relación entre la probabilidad p de éxito y las variables predictoras X se describe como:

$$\log\left(\frac{p}{1-p}\right) = X\beta$$

Aquí, la función de enlace logit transforma la probabilidad p en el predictor lineal $X\beta$.

Estimación en GLM

La estimación de los coeficientes en un GLM generalmente se realiza mediante la maximización de la función de verosimilitud. La función de verosimilitud mide cuán bien se ajusta el modelo a los datos observados. A diferencia de la regresión lineal, donde las ecuaciones normales proporcionan una solución cerrada, en GLM se utiliza un algoritmo iterativo, como el algoritmo de Newton-Raphson, para maximizar la función de verosimilitud y, por lo tanto, estimar los coeficientes.

Dado que el GLM es una generalización, es capaz de abarcar muchos modelos estadísticos diferentes, desde la regresión lineal hasta la regresión logística, simplemente eligiendo diferentes distribuciones y funciones de enlace. Esto lo hace

extremadamente flexible y ampliamente aplicable en diversas situaciones estadísticas.

2.2.6 Regresión Logística

La regresión logística es una herramienta invaluable para investigaciones en el ámbito educativo, especialmente cuando se desea modelar la probabilidad de que un estudiante apruebe o repruebe con base en diversas variables predictoras.

Dentro del contexto académico, la distribución binomial se utiliza para modelar el número de estudiantes que aprueban (éxito) en un grupo de n estudiantes, cuando la probabilidad de aprobar para cada estudiante es p. En el caso de predecir si un estudiante específico aprueba o reprueba, estamos interesados en el caso donde $n = 1$. La función de masa de probabilidad para una variable aleatoria Y, que indica si un estudiante aprueba (1) o reprueba (0), es:

$$P(Y = y) = p^y(1-p)^{1-y}$$

Para modelar la probabilidad p de que un estudiante apruebe con base en variables predictoras, como variables sociodemográficas, requerimos una función que relacione estas variables con la probabilidad de aprobar. Dicha función debe transformar el rango de las combinaciones lineales de estas variables, que es $(-\infty, \infty)$, al rango de una probabilidad, que es $(0,1)$. Por lo tanto, la función logit cumple con este propósito:

$$g(p) = \log\left(\frac{p}{1-p}\right)$$

La inversa de esta función, que relaciona la combinación lineal de las variables predictoras con la probabilidad de aprobar, es:

$$p = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

Por lo que el modelo de regresión logística para predecir si un estudiante aprueba se define como:

$$\log\left(\frac{p}{1-p}\right) = X\beta$$

A diferencia de la regresión lineal, donde podemos usar las ecuaciones normales para obtener una solución cerrada para los coeficientes, en la regresión logística, los coeficientes se estiman maximizando la función de verosimilitud. La función de verosimilitud indica qué tan bien se ajusta el modelo a los datos observados de estudiantes que aprobaron y reprueban.

Para un conjunto de n estudiantes, la función de verosimilitud L es:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

Donde, y_i indica si el estudiante i aprobó (1) o reprobó (0), y p_i es la probabilidad predicha por el modelo de que el estudiante i apruebe. En la práctica, es más común trabajar con el logaritmo de la función de verosimilitud, llamado log-verosimilitud, porque convierte el producto en una suma y simplifica los cálculos:

$$\log\ L(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

El objetivo es encontrar los coeficientes β que maximizan esta log-verosimilitud. Esto se hace típicamente usando métodos numéricos iterativos, como el algoritmo de Newton-Raphson.

Una vez que se han estimado los coeficientes, el modelo puede utilizarse para predecir la probabilidad de que futuros estudiantes aprueben o reprueben basándose en sus características. Además, la magnitud y el signo de los coeficientes nos proporcionan información valiosa sobre cómo cada variable predictora afecta la probabilidad de aprobar.

2.2.7 Regresión penalizada

El análisis de regresión penalizada es una técnica estadística que extiende los modelos de regresión lineal o logística, incorporando términos de penalización en la función de coste para controlar la complejidad del modelo y evitar el sobreajuste (James, Witten, Hastie, Y Tibshirani, 2013). En el contexto del rendimiento académico, a medida que recolectamos datos sobre el rendimiento académico de los estudiantes, podríamos encontrarnos con una gran cantidad de variables predictoras. Estas pueden incluir no sólo puntajes de exámenes y asistencia, sino también características socioeconómicas, hábitos de estudio, y más. En tales escenarios, donde el número de predictores es considerable, corremos el riesgo de sobreajustar nuestro modelo a los datos de entrenamiento, lo que podría llevar a predicciones imprecisas en nuevos datos. Además, algunos de estos predictores pueden ser colineales o no aportar información significativa para la predicción. Aquí es donde la regresión logística

penalizada, en particular las técnicas Lasso y Ridge, juegan un papel crucial. (Sánchez, 2014).

Regresión Lasso

La regresión penalizada Lasso (Least Absolute Shrinkage and Selection Operator) es una técnica de regresión lineal que incorpora un término de penalización basado en la norma L1 para controlar la complejidad del modelo y realizar selección de variables. Esta técnica es especialmente útil cuando se trabaja con conjuntos de datos de alta dimensionalidad o con variables altamente correlacionadas, ya que permite seleccionar las variables más relevantes y mejorar la interpretabilidad del modelo (Tibshirani, 1996). La fórmula para la regresión logística con penalización Lasso es:

$$\min_{\beta_0,\beta} \left\{ -\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + \beta^T x_i) - (1 - y_i)\log\left(1 + \exp\left(\beta_0 + \beta^T x_i\right)\right) + \lambda \parallel \beta \parallel_1 \right\},$$

Ecuación 1. *Regresión logística con penalización Lasso.*

donde *N* es el número total de instancias en el conjunto de entrenamiento, y_i es la clase real de la instancia i, x_i es el vector de características de la instancia i, λ es el parámetro de regularización, que controla el equilibrio entre ajustar los datos y mantener los coeficientes pequeños, β_0 y β son los coeficientes del modelo, $\parallel \beta \parallel_1$ es la norma L1 de los coeficientes, que es la suma de los valores absolutos de los coeficientes.

La idea es encontrar un equilibrio entre ajustar el modelo a los datos (minimizando la log-verosimilitud negativa) y mantener el modelo simple (a través de la penalización). Al hacer esto, buscamos un modelo que haga buenas predicciones en nuevos datos, no solo en el conjunto de entrenamiento.

Regresión Ridge

La regresión penalizada Ridge, también conocida como regresión Tikhonov o regresión con contracción, es una técnica de análisis estadístico que extiende el modelo de regresión lineal múltiple, incorporando un término de penalización basado en la norma L2 para controlar la complejidad del modelo y evitar el sobreajuste. Esta técnica es especialmente útil en situaciones en las que hay multicolinealidad entre las variables predictoras, es decir, cuando las variables independientes están altamente correlacionadas entre sí (Zou Y Hastie, 2005)

La regresión Ridge se basa en la minimización de la suma de los errores cuadráticos más un término de penalización proporcional al cuadrado de la norma L2 de los coeficientes de regresión. La penalización L2 en la regresión Ridge reduce la magnitud de los coeficientes de regresión, pero no los lleva a cero (a diferencia de la regresión LASSO). Esto resulta en modelos más robustos frente a la colinealidad y el ruido en los datos (Hastie *et al.*, 2009).

El término de penalización en la regresión Ridge se controla mediante un parámetro de regularización (lambda), que determina el equilibrio entre el ajuste del modelo a los datos y la magnitud de los coeficientes de regresión. Un valor de lambda más grande resulta en coeficientes más pequeños y un modelo más simple, mientras que un valor de lambda más pequeño permite coeficientes más grandes y un modelo más complejo (Hastie *et al.*, 2009).

La regresión Ridge es útil en una variedad de aplicaciones, incluidas las ciencias sociales, económicas, biológicas y de la salud, donde es importante identificar y cuantificar las relaciones entre variables dependientes e independientes en presencia de multicolinealidad. La fórmula para la regresión logística con penalización Ridge es:

$$\min_{\beta_0,\beta}\left\{-\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0+\beta^T x_i)-(1-y_i)\log\left(1+\exp\left(\beta_0+\beta^T x_i\right)\right)+\lambda \parallel \beta \parallel_2^2\right\},$$

Ecuación 2. *Regresión logística con penalización Ridge.*

donde *N* es el número total de instancias en el conjunto de entrenamiento, y_i es la clase real de la instancia i, x_i es el vector de características de la instancia i, λ es el parámetro de regularización, que controla el equilibrio entre ajustar los datos y mantener los coeficientes pequeños, β_0 y β son los coeficientes del modelo, $\parallel \beta \parallel_2^2$ es la norma $L2$ al cuadrado de los coeficientes, que es la suma de los cuadrados de los coeficientes.

La idea detrás de Ridge es prevenir el sobreajuste y manejar la multicolinealidad, que es cuando las variables predictoras están altamente correlacionadas. Al penalizar los coeficientes, Ridge asegura que ninguna variable predictora individual tenga demasiado peso, lo que puede ser útil cuando las variables son colineales.

En resumen, al igual que Lasso, Ridge busca un equilibrio entre ajustar el modelo a los datos y mantener el modelo simple. Esta combinación tiende a resultar en un modelo que generaliza bien a nuevos datos, especialmente cuando hay muchas variables predictoras o colinealidad entre ellas.

2.2.8 Evaluación del modelo predictivo

Las métricas de evaluación sirven para evaluar los resultados de un proceso de modelización. Además, son importantes para la comparación entre varios modelos, que usen o no las misa metodología, el cuál será de ayuda para nosotros para elegir el mejor modelo de nuestro problema de rendimiento académico. Para evaluar el rendimiento de los algoritmos de clasificación normalmente las medidas de desempeño se calculan comparando las predicciones generadas por este para la

prueba o validación de las clases verdaderas del mismo conjunto. Con esto consideraremos la matriz de confusión y las curvas de ROC sugeridas por Calva (2020):

Matriz de Confusión: es una tabla de doble entrada que permite observar los errores cometidos por el modelo en la fase de entrenamiento. Las filas tienen la clase real de la instancia y las columnas tienen la clase estimada por el clasificador (Calva Yaguana, 2020). Esta matriz es conocida como la matriz de errores. La Tabla 1 muestra la matriz de confusión, donde Positivo representa la clase 1 y Negativo la clase 0.

Tabla 1.

Matriz de confusión.

Pred./Act.	**Positivo**	**Negativo**
Positivo	Verdadero Positivo (True Positive, **TP**)	Falso Negativo (False Negative, **FN**)
Negativo	Falso Positivo (False Positive, **FP**)	Verdadero Negativo (True Negative, **TN**)

Con base en la matriz de confusión Tabla 1, definimos las siguientes variables:

- **Verdadero Positivo (TP):** es el número de clasificaciones correctas para la clase 1.
- **Falso Negativo (FN):** es el número de clasificaciones incorrectas de la clase 1 ya que fueron clasificadas como 0.
- **Verdadero Negativo (TN):** es el número de clasificaciones correctas para la clase 0.
- **Falso Positivo (FP):** es el número de clasificaciones incorrectas de la clase 0 ya que fueron clasificadas como 1.

Una vez definidas las variables de la matriz de confusión, vamos a obtener algunas métricas que permiten cuantificar la bondad de ajuste del modelo que servirán para realizar la comparativa de los diferentes algoritmos de clasificación usados en el modelo de predicción.

Entonces, consideraremos **TP, FN, TN,** y **FP** de la *Tabla 1* para cada una de las siguientes fórmulas.

- **Suma Total del conjunto de datos de entrenamiento**: es el número del conjunto de datos de entrenamiento.

$$N = TP + FP + TN + FN.$$

Ecuación 3. *Fórmula suma total del conjunto de datos de entrenamiento*

- **Exactitud (Accuracy):** Número de predicciones correctas sobre el número total de datos de entrenamiento.

$$Acc = \frac{TP + TN}{N}$$

Ecuación 4. *Fórmula de la Exactitud del modelo.*

- **Tasa de error (error_rate):** Porcentaje de instancias del conjunto de datos que son clasificadas incorrectamente.

$$error_{rate} = \frac{FP + FN}{N}.$$

Ecuación 5. *Fórmula de Tasa de Error.*

- **Precisión (Pre):** Rendimiento relacionado con las tasas de verdaderos positivos y negativos, y expresa la proporción de puntos relevantes de nuestro modelo.

$$Pre = \frac{TP}{TP + FP}.$$

Ecuación 6. *Fórmula de Precisión.*

- **Sensibilidad (Sen):** Capacidad del modelo de encontrar todos los puntos de interés en un conjunto de datos.

$$Sen = \frac{TP}{FN + TP}.$$

Ecuación 7. *Fórmula de Sensibilidad.*

- **Especificidad (Spe):** Tasa de observaciones correctamente clasificadas como clase 0 respecto a todas las instancias de clase 0.

$$Spe = \frac{TN}{TN + FP}.$$

Ecuación 8. *Fórmula de Especificidad.*

- **F measure:** Al combinar las medidas de precisión y sensibilidad y si se obtiene valores cercanos a 1 nos indica que tan preciso es un clasificador (cuantas instancias clasifica correctamente), así como tan robusto es (no pierde un número significativo de instancias).

$$F = 2 \times \frac{Pre \times Rec}{Pre + Rec}.$$

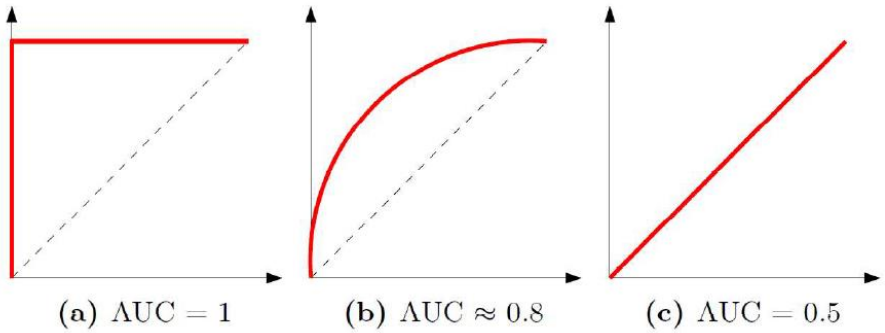
Ecuación 9. *Fórmula de F measure.*

Curvas ROC: mide el rendimiento de un modelo de clasificación respecto a los falsos positivos y verdaderos positivos. Su diagonal se interpreta como un modelo aleatorio, mientras que los valores inferiores se consideran peores que una estimación aleatoria de nuevos datos (Calva Yaguana, 2020)**.**

Bajo esta métrica, la Figura 1 muestra tres casos en los que Calva (2020) explica que, en la práctica, un AUC definido como el área bajo la curva ROC entre 0.5 y 0.6 tiene rendimiento malo, de 0.6 a 0.75 un rendimiento regular, de 0.75 a 0.9 un rendimiento bueno, de 0.9 a 0.97 un rendimiento muy bueno y más de 0.97 es un rendimiento excelente. Por lo tanto, un clasificador excelente sería la primera gráfica de la izquierda, con una tasa de verdaderos positivos (TP) de 1 y una tasa de falsos negativos (FP) de 0. La imagen del medio presenta un rendimiento bueno y la imagen de la derecha presenta un rendimiento malo.

Figura 1.

Ejemplo Curvas ROC.



Nota. Curvas ROC con AUC de 1 (perfecta), ~0.8 (buena), y 0.5 (azar).

2.3. Marco legal

En el marco legal, la Constitución de 2008 de Ecuador, en su Título II, dedicado a los "Derechos", particularmente en el capítulo segundo referente a los "Derechos del Buen Vivir", se encuentran plasmados los principios generales de la educación. Estos están contenidos específicamente en la sección quinta, comprendiendo los artículos del 26 al 29.

Análisis del Art. 26

Este artículo 26, presenta el concepto fundamental de educación que propone la nueva Constitución. Destacando cuatro aspectos importantes para las familias y la sociedad.

a.- La educación como un derecho permanente de las personas.

b.- La educación como un área prioritaria de la inversión estatal.

c.- La educación como una garantía de inclusión.

d.- La educación como un espacio de participación de las familias.

Este primer artículo, determina que la educación es un derecho de todas las personas, señala la obligatoriedad que tiene el estado de garantizar educación a nuestro pueblo, la educación se convierte en una garantía para el buen vivir para ello la sociedad en su conjunto está obligada a aportar en este proceso.

Análisis del Art. 27

El artículo 27, describe los elementos constitutivos de la educación que lo propone como derecho básico a todos los ecuatorianos. Entre las características que dicha educación tendrá destacan dos aspectos.

a.- Estará centrada en el ser humano.

b.- Concebirá al ser humano holísticamente, es decir, "como un todo distinto de la suma de las partes que lo componen", según la definición que consta en el Diccionario de la Real Academia Española.

Este artículo también nos recuerda la importancia que tiene la educación para la construcción de una sociedad democrática, justa y solidaria. El objetivo de este artículo busca que los ecuatorianos tengan igualdad de oportunidades, que sepan compartir sus conocimientos con los demás y vivir en un ambiente de paz.

Análisis del Art. 28

El punto más importante que se destaca en el artículo 28 de la Constitución 2008 es garantizar que la educación pública este abierta para todas las personas (que sea universal) y que no promueva ninguna religión en particular (que sea laica). La principal conquista del liberalismo es ratificada en esta constitución; EL LAICISMO, de esta manera se subraya que la escuela fiscal debe respetar toda creencia religiosa. También hace hincapié en la universalidad de la educación sin discriminación alguna, todo lo contrario, se debe garantizar esa movilidad que a la que siempre está sujeta la educación, y concluye determinando su gratuidad hasta el nivel superior inclusive.

Análisis del Art. 29

El artículo 29, garantizara la larga tradición en el mundo académico de la Universidad: la libertad de cátedra, que es indispensable para el libre debate de las ideas. También mantiene el derecho a la educación en su propia lengua, lo que es fundamental para mejores niveles de aprendizaje.

LOEI

CAPÍTULO TERCERO: De los derechos y obligaciones de los estudiantes

Art. 7.- Derechos. - Las y los estudiantes tienen los siguientes derechos:

a. Ser actores fundamentales en el proceso educativo;

b. Recibir una formación integral y científica, que contribuya al pleno desarrollo de su personalidad, capacidades y potencialidades, respetando sus derechos, libertades fundamentales y promoviendo la igualdad de género, la no discriminación, la valoración de las diversidades, la participación, autonomía y cooperación;

c. Ser tratado con justicia, dignidad, sin discriminación, con respeto a su diversidad individual, cultural, sexual y lingüística, a sus convicciones ideológicas, políticas y

religiosas, y a sus derechos y libertades fundamentales garantizados en la Constitución de la República, tratados e instrumentos internacionales vigentes y la Ley;

d. Intervenir en el proceso de evaluación interna y externa como parte y finalidad de su proceso educativo, sin discriminación de ninguna naturaleza;

e. Recibir gratuitamente servicios de carácter social, psicológico y de atención integral de salud en sus circuitos educativos;

f. Recibir apoyo pedagógico y tutorías académicas de acuerdo con sus necesidades;

Art. 8.- Obligaciones. - Las y los estudiantes tienen las siguientes obligaciones:

a. Asistir regularmente a clases y cumplir con las tareas y obligaciones derivadas del proceso de enseñanza y aprendizaje, de acuerdo con la reglamentación correspondiente y de conformidad con la modalidad educativa, salvo los casos de situación de vulnerabilidad en los cuales se pueda reconocer horarios flexibles;

b. Participar en la evaluación de manera permanente, a través de procesos internos y externos que validen la calidad de la educación y el interaprendizaje;

c. Procurar la excelencia educativa y mostrar integridad y honestidad académica en el cumplimiento de las tareas y obligaciones.

CAPÍTULO III

METODOLOGÍA

3.1. Descripción del área de estudio/grupo de estudio

El área de estudio de esta investigación se centra en la Universidad Yachay Tech, una universidad pública ubicada en la ciudad de Urcuquí, en la provincia de Imbabura, en la región norte de Ecuador (Figura 2). Esta universidad fue fundada en 2014 y tiene como objetivo formar líderes científicos e innovadores capaces de contribuir al desarrollo del país.

La universidad ofrece una amplia variedad de carreras en áreas como ciencias físicas, ciencias biológicas, ingeniería y tecnología, ciencias sociales y humanidades. El grupo de estudio se compone de los estudiantes de primer semestre de la universidad, quienes tienen edades entre 18 y 25 años, y provienen de diferentes partes de Ecuador y otros países.

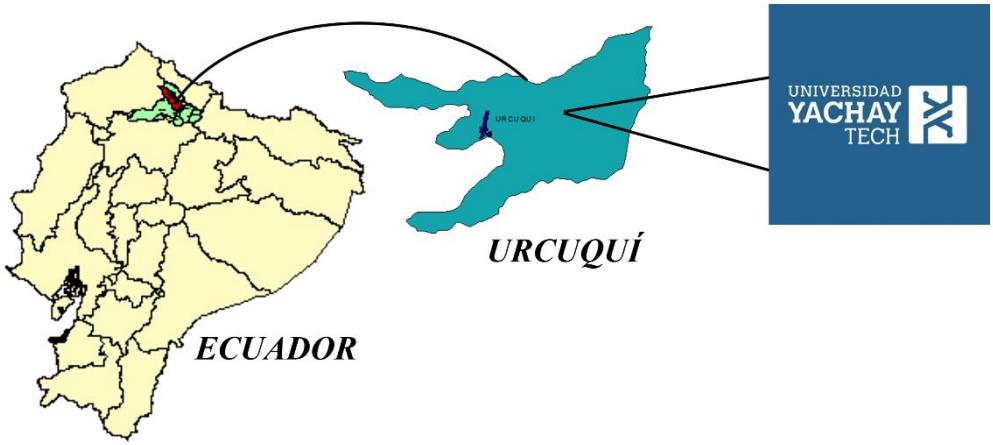
El perfil sociodemográfico de los estudiantes de la Universidad Yachay Tech es diverso, con una mezcla de estudiantes de diferentes niveles socioeconómicos y educativos. Además, la universidad cuenta con un programa de becas que ofrece apoyo financiero a estudiantes con necesidades económicas.

El Departamento de Dirección Académica juntamente con el Departamento de Bienestar Estudiantil almacenan, gestionan y administra la información académica e información de los estudiantes de la universidad Yachay Tech respectivamente, los cuales serán la fuente que nos proporcionen todas las variables que se consideran en el presente trabajo. Entonces, para este trabajo vamos a usar datos de estudiantes, que han cursado el primer semestre desde el primer periodo académico 2014 al segundo periodo académico de 2022. En el sistema educativo de esta universidad se consideran 10 semestres de clases para todas las carreras, las cuales deben tener un

tronco común durante 4 semestres. Si el estudiante reprueba en una materia y esta no tiene continuidad con materias de segundo semestre, el estudiante puede pasar al siguiente semestre, pero con segunda matricula solo en la materia que reprobó; el estudiante cuenta con tres matriculas para aprobar la materia. Vamos a utilizar la información recopilada de los estudiantes de primer semestre, centrándonos especialmente en las asignaturas de Cálculo, Álgebra Lineal, Química y Biología. Esta elección se fundamenta en la transición académica que implica el cambio de semestre entre la educación secundaria y universitaria, la cual revela una significativa incidencia de deserción estudiantil. Dicha deserción se caracteriza por la dificultad que experimentan los alumnos al adaptarse a las exigencias académicas y estructurales inherentes a la educación superior.

Figura 2.

Área de Estudio.



Nota Mapa Área de Estudio - Provincia de Imbabura – Urcuquí- Universidad Yachay Tech.

3.2. Enfoque y tipo de investigación

Enfoque

El enfoque de este estudio es cuantitativo, implementando técnicas estadísticas para desentrañar los datos sociodemográficos de los estudiantes de primer semestre universitario, con el propósito de desarrollar un modelo predictivo del rendimiento académico. Este enfoque cuantitativo posibilita la recolección y análisis de información de manera objetiva y precisa, mediante el uso de técnicas estadísticas, permitiendo una evaluación rigurosa de las variables implicadas en la investigación. Más allá de la precisión en la medición, este método es especialmente adecuado para identificar patrones y establecer relaciones entre las variables, elementos fundamentales para la construcción de un modelo predictivo sólido y confiable.

Tipo de Investigación

Este estudio se caracteriza por ser una investigación correlacional, cuyo objetivo es examinar la relación entre dos o más variables y determinar cómo estas interrelacionan entre sí. Este tipo de investigación cuantitativa se destina a identificar las conexiones existentes entre distintas variables, en este caso, buscando esclarecer la relación entre los factores sociodemográficos de los estudiantes y su rendimiento académico. El propósito final es la creación de modelos predictivos capaces de anticipar el rendimiento académico de los estudiantes de primer semestre basándose en sus factores sociodemográficos.

Además, este estudio adopta la naturaleza de una investigación explicativa, ya que se esfuerza por elucidar la relación entre las variables sociodemográficas y el rendimiento académico de los estudiantes. Este tipo de investigación cuantitativa se aplica para entender la conexión causal entre variables y para pronosticar cómo la modificación de una variable afectará a las demás. En este escenario, se busca

discernir cómo los factores sociodemográficos influyen en el rendimiento académico de los estudiantes de primer semestre y desarrollar modelos predictivos basados en dichos factores.

Finalmente, este trabajo también incorpora elementos de investigación descriptiva, pues su finalidad es detallar las características de una población o fenómeno específico. Este tipo de investigación cuantitativa se utiliza para recolectar información acerca de una población o fenómeno y describir sus particularidades. En este caso, se busca describir las características de los factores sociodemográficos y del rendimiento académico de los estudiantes de primer semestre de la Universidad Yachay Tech de Ecuador, y explorar las interacciones entre ambos.

3.3. Definición y operacionalización de variables

Hipótesis

Existe una relación significativa entre los datos sociodemográficos de los estudiantes de primer semestre en la Universidad Yachay Tech y su rendimiento académico.

Tabla 2.

Definición de variables.

VARIABLES	CONCEPTUALIZACIÓN DE VARIABLE
Variable Independiente: Factores sociodemográficos de los estudiantes universitarios de primer semestre en la Universidad Yachay Tech.	Conjunto de características que describen a los estudiantes en términos de su composición sociodemográfica. Estos factores pueden influir en la forma en que los estudiantes se adaptan al ambiente universitario, en sus

	habilidades académicas y en la forma en que administran su tiempo y sus recursos para el estudio.
Variable Dependiente: Rendimiento académico de los estudiantes universitarios de primer semestre en la Universidad Yachay Tech.	Se refiere al nivel de logro y éxito que los estudiantes alcanzan en sus estudios durante su primer semestre en la universidad. El rendimiento académico se puede medir a través de diferentes indicadores, como las calificaciones obtenidas en exámenes, trabajos y proyectos, la asistencia a clases, la participación en actividades extracurriculares y la retención en la universidad

Operacionalización de variables

Variable Independiente: Factores sociodemográficos de los estudiantes universitarios de primer semestre en la Universidad Yachay Tech.

Tabla 3.

Operacionalización de variable Independiente.

Variable	Dimensión	Indicadores	Técnica	Instrumento
Factores sociodemográficos	Matrícula de cursos	Matrícula de cálculo, Matrícula de álgebra, Matrícula de biología, Matrícula de química, Matrícula de nivelación	Revisión de registros	Base de datos de la Universidad
	Características personales	Género, Edad, Etnia, Estado civil, Discapacidad, Trabajo, Hijos	Revisión de registros	Base de datos de la Universidad
	Origen	País de nacimiento, Provincia de nacimiento	Revisión de registros	Base de datos de la Universidad
	Antecedentes académicos	Nota de grado, Curso de nivelación, Tipo de colegio	Revisión de registros	Base de datos de la Universidad
	Situación familiar	Tenencia de vivienda, Ocupación del padre, Ocupación de la madre	Revisión de registros	Base de datos de la Universidad
	Carrera y asignaturas	Carrera, Asignaturas del primer semestre, Número de asignaturas aprobadas	Revisión de registros	Base de datos de la Universidad

Variable Dependiente: Rendimiento académico de los estudiantes universitarios de primer semestre en la Universidad Yachay Tech.

Tabla 4.

Operacionalización de variable Dependiente.

Variable	Dimensión	Indicadores	Técnica	Instrumento
Rendimiento académico	Resultados del primer semestre	Notas en cálculo, álgebra, biología, química	Revisión de registros	Base de datos de la Universidad

3.4. Procedimientos

Para abordar el procedimiento de este estudio sobre las variables sociodemográficas y el rendimiento académico de los estudiantes de primer semestre en la Universidad Yachay Tech, el proceso de investigación se estructura en tres fases distintas. Cada fase se alinea con uno de los objetivos específicos, permitiendo un análisis detallado y una evaluación rigurosa de los datos y modelos utilizados.

Fase 1. Análisis de las las variables sociodemográficas de los estudiantes de primer semestre en la Universidad Yachay Tech y su relación con el rendimiento académico, utilizando técnicas estadísticas descriptivas y analíticas.

En la primera fase del estudio, el objetivo se centra en analizar las variables sociodemográficas de los estudiantes de primer semestre en la Universidad Yachay Tech y su relación con el rendimiento académico. Inicialmente, se procede con la recolección de datos, donde se obtiene información sociodemográfica y académica detallada de los estudiantes desde el año 2014 hasta el primer semestre de 2022. Esta recolección implica asegurar la integridad y la relevancia de los datos para el análisis. Una vez recolectados, los datos se someten a un proceso de preprocesamiento riguroso. Aquí, se limpian y se preparan para el análisis, tratando

aspectos como la conversión de variables en factores y la gestión de datos erróneos o incompletos. Este paso es crucial para garantizar la precisión de los análisis posteriores. El análisis de los datos se realiza mediante técnicas estadísticas descriptivas y analíticas. Se emplean métodos descriptivos para resumir y visualizar los datos, proporcionando una comprensión inicial de las características de la población estudiada. Posteriormente, se aplican técnicas analíticas, como el análisis de correlación, para explorar en profundidad las relaciones entre las variables sociodemográficas y el rendimiento académico. Este enfoque permite identificar patrones y tendencias significativas que podrían influir en las fases posteriores del estudio.

Fase 2. Desarrollo de modelos predictivos basados en regresión logística con penalizaciones Lasso y Ridge, y comparar su eficacia en la predicción del rendimiento académico de los estudiantes de primer semestre en la Universidad Yachay Tech a partir de datos sociodemográficos.

La segunda fase tiene como objetivo el desarrollo de modelos predictivos basados en regresión logística con penalizaciones Lasso y Ridge. Este proceso comienza con la división de los datos en un conjunto de entrenamiento, que comprende el 70% de los datos, y un conjunto de prueba, con el 30% restante. Esta división es estratégica para entrenar los modelos predictivos y posteriormente evaluar su capacidad de generalización en datos no vistos. Se desarrollan dos modelos de regresión logística: uno con penalización Lasso (L1) y otro con penalización Ridge (L2). El modelo Lasso es particularmente útil para la selección de características, ya que reduce algunos coeficientes a cero, enfocándose en las variables más relevantes para la predicción del rendimiento académico. Por otro lado, el modelo Ridge es empleado para manejar la multicolinealidad entre las variables predictoras, distribuyendo los

coeficientes de manera equitativa y evitando el dominio de una sola variable en el modelo.

Fase 3. Evaluación y comparación de los modelos de regresión logística con penalizaciones Lasso y Ridge mediante el análisis de métricas apropiadas, para determinar su eficacia en la predicción del rendimiento académico de los estudiantes de primer semestre en la Universidad Yachay Tech.

La última fase del estudio implica una evaluación y comparación exhaustivas de los modelos Lasso y Ridge. Utilizando el conjunto de prueba, se evaluaron los modelos empleando una variedad de métricas, como la matriz de confusión, las curvas ROC, el área bajo la curva (AUC), la exactitud, la sensibilidad, la especificidad y el coeficiente Kappa. Estas métricas ofrecieron una evaluación completa del rendimiento de los modelos en la predicción del rendimiento académico. Además, se compararon ambos modelos para determinar cuál ofrece predicciones más precisas y confiables, proporcionando una visión detallada de los factores más influyentes en el rendimiento académico. Para asegurar la coherencia y la reproducibilidad en el análisis, todos los procedimientos se llevaron a cabo utilizando el software estadístico R, con paquetes especializados que facilitaron cada etapa del proceso. La estructura de tres fases del estudio garantiza una exploración profunda y sistemática de la relación entre las variables sociodemográficas y el rendimiento académico, resaltando la importancia de un enfoque metodológico riguroso y bien definido.

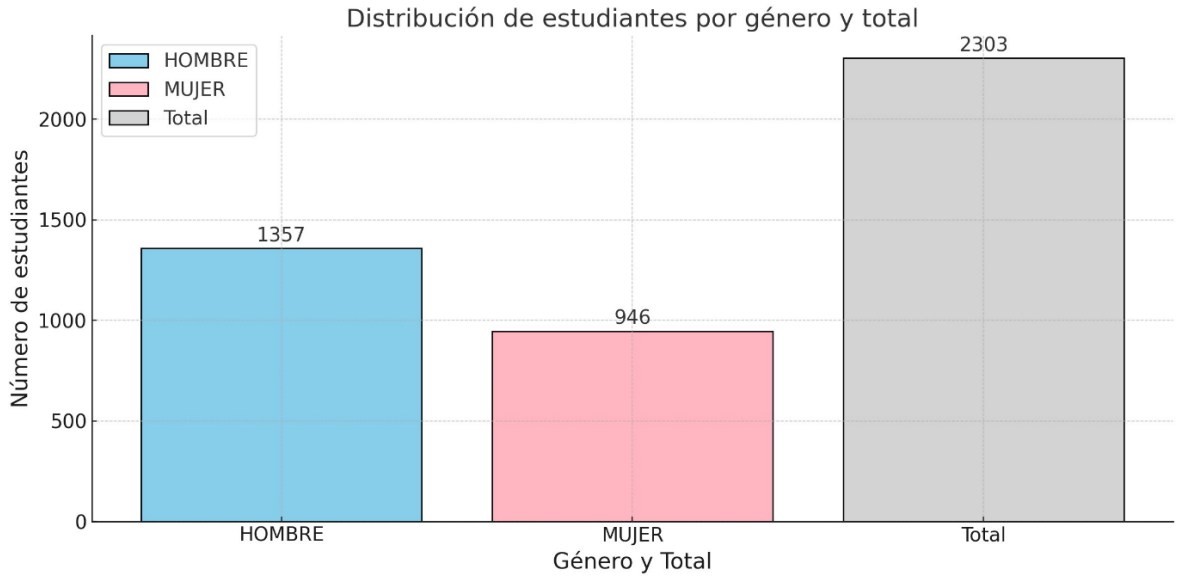
CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

En el presente estudio, se llevaron a cabo una serie de análisis para cumplir con los objetivos planteados inicialmente. La primera fase de la investigación se centró en analizar las variables sociodemográficas de los estudiantes de primer semestre de la Universidad Yachay Tech, utilizando técnicas estadísticas descriptivas y analíticas. Las gráficas de barras (Figura *3*).

Figura 3.

Distribución de estudiantes de primer semestre.



Nota. Gráfico de barras mostrando estudiantes de primer semestre de la universidad Yachay Tech desde el año 2014-2023.

Los resultados de la distribución de estudiantes en el primer semestre de la Universidad Yachay Tech muestran una clara disparidad en la representación de género, con una mayor cantidad de hombres (1357) en comparación con mujeres (946). Esto sugiere una proporción de aproximadamente 1.43 hombres por cada mujer. Este desequilibrio en la representación de género es algo que se ha observado

en muchas instituciones de educación superior en Ecuador, especialmente en áreas técnicas y de ingeniería. La Universidad Yachay Tech, siendo una institución de ciencia y tecnología, puede reflejar esta tendencia global. Este dato puede ser relevante para la investigación, especialmente si se considera la influencia potencial del género en el rendimiento académico.

Es crucial no interpretar estos resultados como indicativos de una brecha de habilidad entre géneros. En cambio, este desequilibrio puede reflejar barreras sociales y culturales que incitan a las mujeres de perseguir carreras en ciencia y tecnología. Estas barreras pueden incluir estereotipos de género, falta de modelos a seguir y la percepción de un ambiente no inclusivo. Además, es posible que este desequilibrio de género pueda influir en los resultados de la predicción del rendimiento académico. Algunas investigaciones sugieren que el ambiente educativo puede tener un impacto en cómo los estudiantes de diferentes géneros se desempeñan académicamente.

Además, en este estudio, se analizó el rendimiento académico de los estudiantes de primer semestre de la Universidad Yachay Tech, considerando las asignaturas de Cálculo, Álgebra, Biología y Química, y su relación con diversas variables sociodemográficas.

4.1 Distribución General de Notas

El diagrama de caja, también conocido como boxplot, es una herramienta gráfica que permite visualizar la distribución y variabilidad de un conjunto de datos. A continuación, se presenta un análisis basado en el boxplot generado para las notas de las asignaturas Cálculo, Álgebra, Biología y Química (Figura *4*).

Cálculo: La mediana de las notas se encuentra aproximadamente en el valor 6.2, indicando que el 50% de los estudiantes obtuvieron una nota igual o superior a dicho valor en esta asignatura. La distribución presenta valores atípicos en el extremo

inferior, sugiriendo que algunos estudiantes tuvieron dificultades particulares en Cálculo.

Álgebra: La mediana se sitúa cerca del valor 6.1, con un rango intercuartílico más amplio en comparación con Cálculo. Esto indica una mayor variabilidad en las notas de los estudiantes en Álgebra. Al igual que en Cálculo, se observan valores atípicos en el extremo inferior.

Biología: Esta asignatura muestra una mediana cercana al valor 7.0, con un rango intercuartílico similar al de Álgebra. También se observan valores atípicos, principalmente en el extremo inferior.

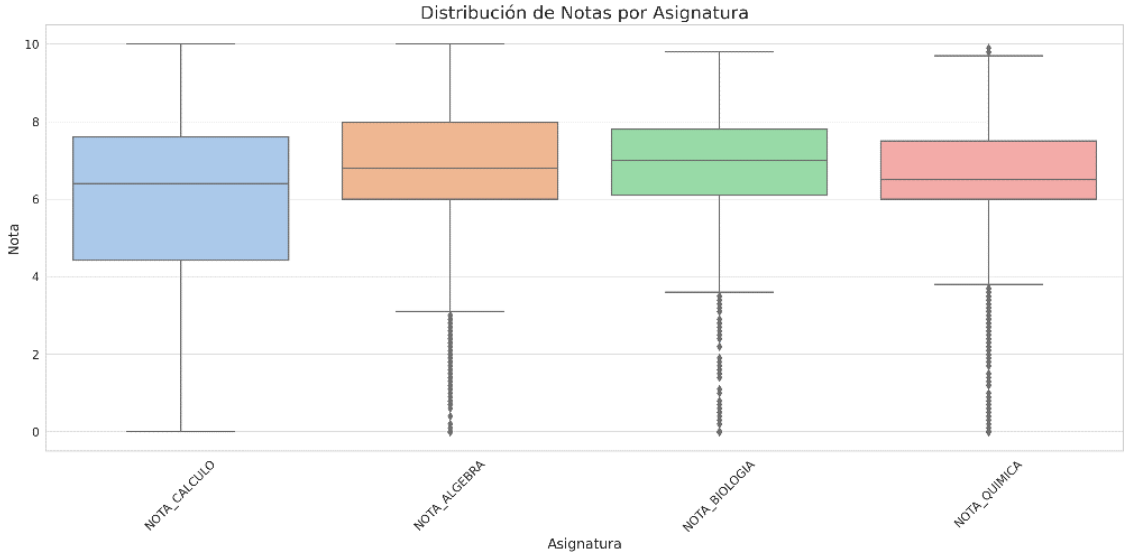
Química: La mediana de las notas para Química se sitúa alrededor del valor 6.8. La distribución es similar a la de Biología, con un rango intercuartílico comparable y valores atípicos en el extremo inferior.

En general, las asignaturas de Cálculo y Álgebra muestran distribuciones de notas similares, mientras que Biología y Química tienen distribuciones algo diferentes, pero similares entre sí.

Estas observaciones sugieren que existe un subconjunto de estudiantes que enfrenta desafíos en las materias matemáticas, mientras que otro subconjunto podría tener dificultades específicas en las ciencias naturales

Figura 4.

Distribución de notas por asignatura.



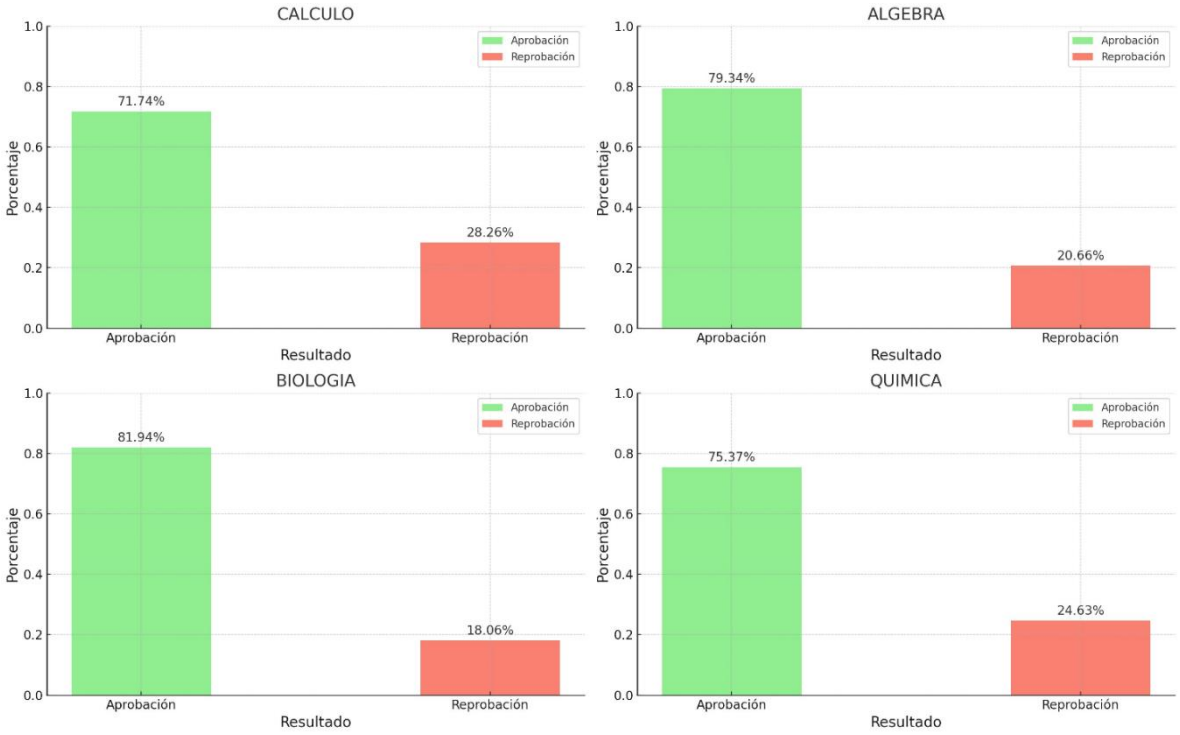
Nota. Gráfico de caja que compara las distribuciones de notas en las asignaturas de Cálculo, Álgebra, Biología y Química.

A continuación, se presenta un análisis basado en los porcentajes de aprobación y reprobación de las asignaturas Cálculo, Álgebra, Biología y Química, segmentados por distribución de notas (Figura *5*):

- La asignatura con el mayor porcentaje de aprobación es **Biología** con casi el 82%, seguida de **Álgebra** con aproximadamente el 79%.
- **Cálculo** presenta el mayor porcentaje de reprobación con cerca del 28%, lo que sugiere que podría ser una asignatura desafiante para los estudiantes.
- Aunque **Química** tiene un porcentaje de aprobación del 75%, sigue siendo relativamente más bajo en comparación con **Biología** y **Álgebra**.

Figura 5 .

Porcentajes de Aprobación y Reprobación por Asignatura.



Nota. Porcentajes de aprobación y reprobación en asignaturas de primer semestre en la Universidad Yachay Tech, años 2014-2020.

4.2 Rendimiento Académico Según Provincia de Residencia

El lugar de residencia puede tener un impacto significativo en el rendimiento académico de los estudiantes, reflejando posiblemente factores como la calidad de la educación previa, el acceso a recursos educativos y diferencias socioeconómicas entre regiones. Para investigar esta relación, se analizó la distribución de las notas de las asignaturas Cálculo, Álgebra, Biología y Química, segmentadas por provincia de residencia (Figura *6*):

Provincias con alto rendimiento:

Loja y Morona Santiago se destacaron por su excelente rendimiento académico. En particular, Morona Santiago obtuvo la mejor nota en Biología con 8.3 y en Química con 6.9. Por otro lado, Orellana, aunque no se clasificó en general como una provincia de alto rendimiento, merece una mención especial por obtener la mejor nota en Álgebra con 7.8.

Provincias con rendimiento intermedio:

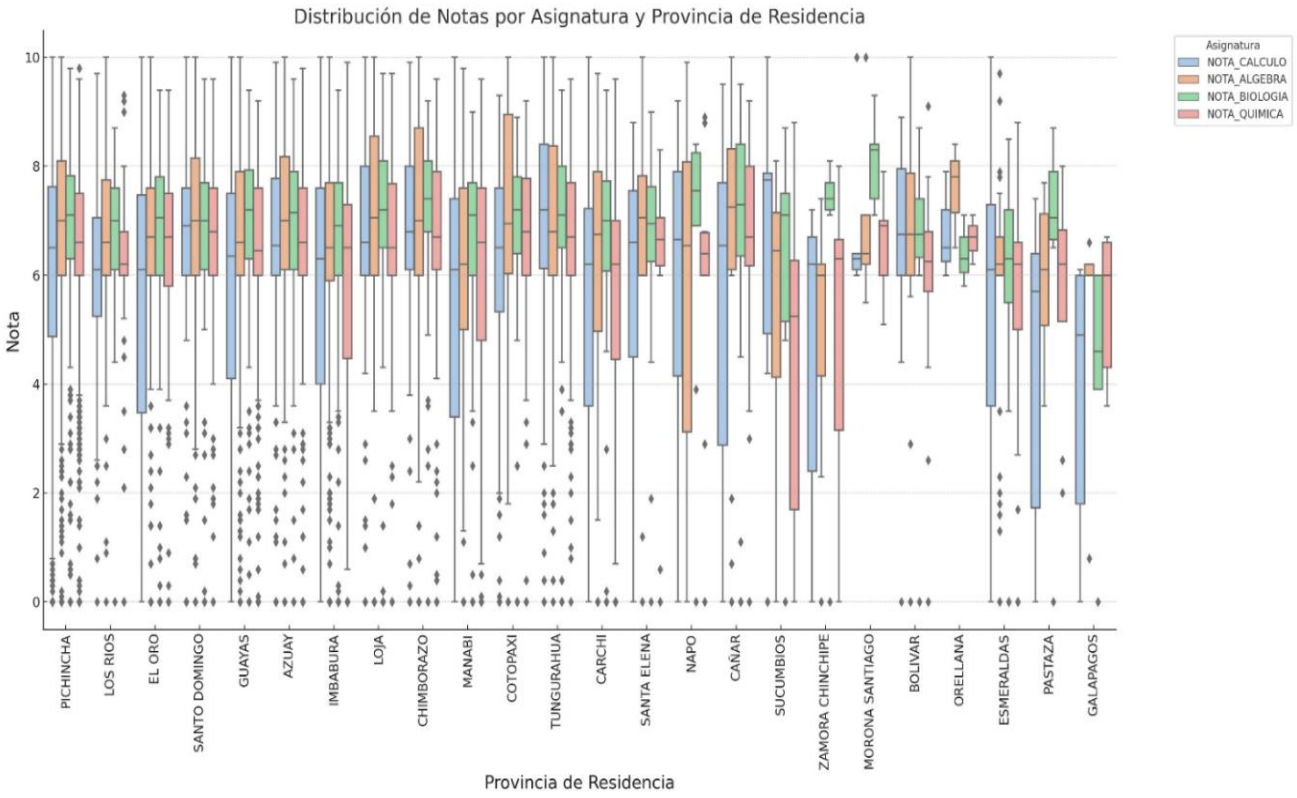
Pichincha, Azuay y Santo Domingo mostraron un rendimiento equilibrado en todas las asignaturas. Sucumbíos, a pesar de tener el mejor rendimiento en Cálculo con una nota de 7.75, tuvo desafíos en Química, donde obtuvo la peor nota con 5.25.

Provincias con bajo rendimiento:

Galápagos enfrentó desafíos significativos en rendimiento académico, obteniendo la peor nota en Biología con 4.6 y en Cálculo con 4.9. Zamora Chinchipe también tuvo dificultades, especialmente en Álgebra, donde registró la nota más baja con 6.0. Estas diferencias en el rendimiento por provincia pueden estar influenciadas por múltiples factores, incluyendo la calidad educativa, el acceso a recursos, y particularidades propias de cada región.

Figura 6.

Distribución de Notas por Asignatura y Provincia.



Nota. Gráfico de caja y bigotes de notas por asignatura y provincia de residencia de estudiantes de la Universidad Yachay Tech.

A continuación, se presenta un análisis basado en los porcentajes promedio de aprobación y reprobación por provincias en un gráfico de mapa coroplético del Ecuador (Figura *7*).

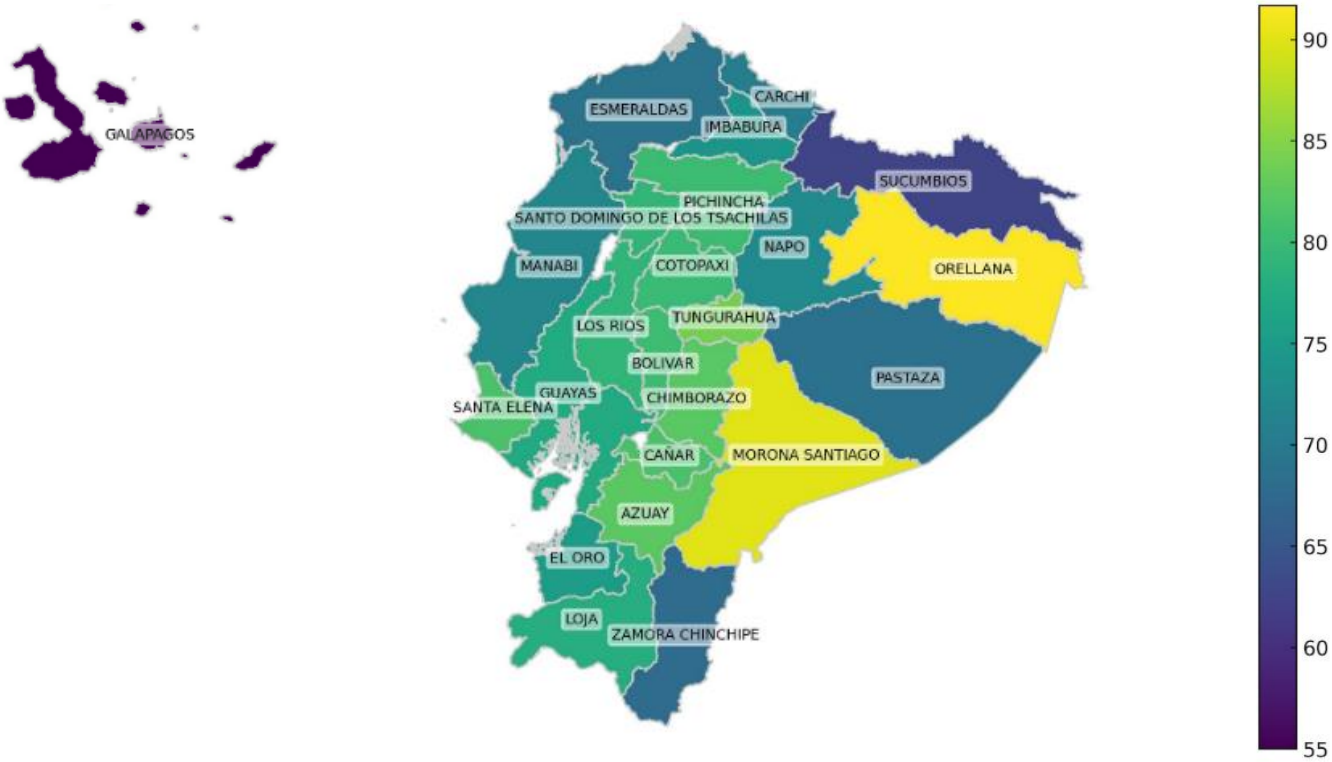
Distribución del Color y Valores Numéricos:

- El mapa utiliza una paleta de colores variadas, que va de un color amarillo claro a un color verde oscuro.
- El rango de porcentajes de aprobación en el mapa va desde el 55.0% (el más bajo, correspondiente a las Galápagos) hasta el 91.67% (el más alto).

2. **Diversidad Regional**:
    - **Esmeraldas**: Esta provincia costera tiene un porcentaje de aprobación del 68.94%.
    - **Manabí**: Otra provincia costera, presenta un porcentaje de aprobación del 72.01%.
    - **Pichincha y Azuay**: Estas provincias en la región andina tienen porcentajes de aprobación del 80.35% y 82.46%, respectivamente, siendo más altos en comparación con las provincias costeras mencionadas anteriormente.
3. **Consideraciones Generales**:
    - La provincia con el porcentaje de aprobación más alto es Pichincha con 80.35%, mientras que la más baja es Galápagos con 55.0%.
    - El promedio de aprobación de todas las provincias es aproximadamente 76.46%.

Figura 7.

Mapa del porcentaje promedio de aprobación por provincia.



Nota. Mapa de calor por provincias de Ecuador que muestra el porcentaje de aprobación de estudiantes en primer semestre de la Universidad Yachay Tech.

4.3. Rendimiento Académico Según Tipo de Colegio

La naturaleza y calidad del colegio en el que un estudiante cursa su educación secundaria puede tener un impacto significativo en su rendimiento académico posterior en la universidad. Para examinar esta relación, se analizaron las notas de las asignaturas Cálculo, Álgebra, Biología y Química, segmentadas por el tipo de colegio en el que los estudiantes completaron su educación secundaria (Figura *8*). Basado en el boxplot y los datos:

Colegios con alto rendimiento:

Los colegios privados destacan notablemente. En promedio, los estudiantes de estos colegios obtuvieron las notas más altas en comparación con otros tipos de colegios. Específicamente, en Cálculo obtuvieron una nota media de 6.36, en Álgebra 7.09, en Biología 6.94 y en Química 6.63. Este alto rendimiento refleja la excelencia académica y la solidez en la preparación de los estudiantes que provienen de colegios privados.

Colegios con rendimiento intermedio:

En el grupo de rendimiento intermedio, encontramos a los colegios públicos y extranjeros. Para los colegios públicos, las notas medias fueron 5.40 en Cálculo, 5.88 en Álgebra, 6.26 en Biología y 5.53 en Química. Por otro lado, los colegios extranjeros presentaron un rendimiento ligeramente variado con notas medias de 4.64 en Cálculo, 6.46 en Álgebra, 6.05 en Biología y 6.31 en Química. Estas cifras sugieren que, aunque estos colegios muestran un rendimiento equilibrado, todavía hay margen de mejora en ciertas áreas.

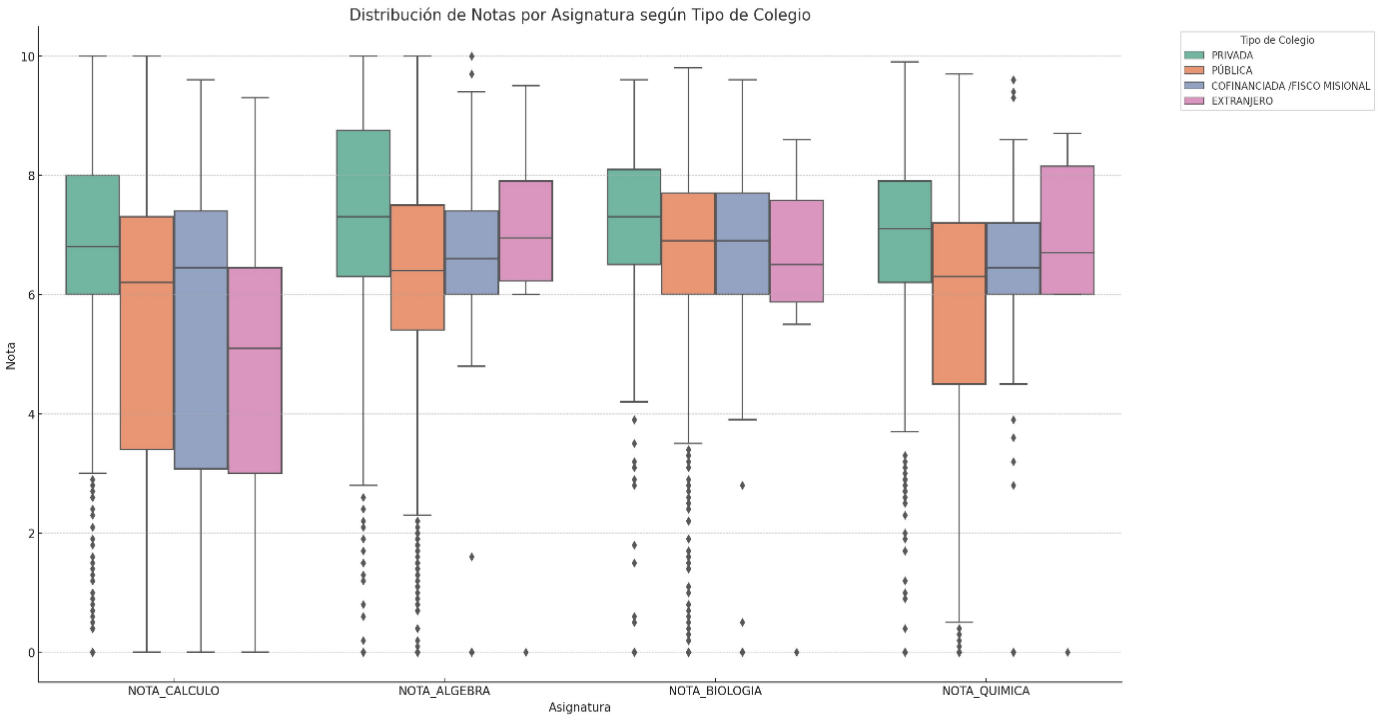
Colegios con bajo rendimiento:

Los colegios Fiscomisionales mostraron el rendimiento más bajo entre las categorías. Las notas medias para estos colegios fueron 5.33 en Cálculo, 5.84 en Álgebra, 5.93 en Biología y 5.63 en Química. Aunque las notas no están significativamente por debajo del promedio, indican áreas donde estos colegios podrían necesitar apoyo adicional para mejorar el rendimiento académico de sus estudiantes.

El tipo de colegio en el que un estudiante cursa su educación secundaria sigue siendo una variable influyente en su rendimiento académico en la universidad. El análisis revisado refuerza la importancia de considerar el entorno educativo previo al evaluar el rendimiento académico en la educación superior.

Figura 8.

Distribución de notas por asignatura según tipo de colegio.



Nota. Gráfico de caja que muestra la distribución de notas en Cálculo, Álgebra, Biología y Química según el tipo de colegio: público, privado, cofinanciado/fiscomisional y extranjero.

A continuación, se presenta un análisis basado en los porcentajes de aprobación y reprobación de las asignaturas Cálculo, Álgebra, Biología y Química, segmentados por tipo de colegio (Figura 9).

• **Privada:** Los estudiantes de colegios privados mostraron consistentemente altas tasas de aprobación en todas las asignaturas. En Cálculo, aproximadamente el 75% de estos estudiantes lograron aprobar. En Álgebra, esta cifra ascendió a más del

80%, mientras que, en Biología y Química, las tasas de aprobación fueron del 83% y 81%, respectivamente.

• **Pública:** Aunque los estudiantes de colegios públicos mostraron tasas de aprobación ligeramente inferiores en comparación con los colegios privados, estos porcentajes siguen siendo robustos. En Cálculo y Álgebra, aproximadamente el 59% y 66% de los estudiantes aprobaron, respectivamente. En Biología, esta cifra se elevó al 72%, mientras que, en Química, fue del 68%.

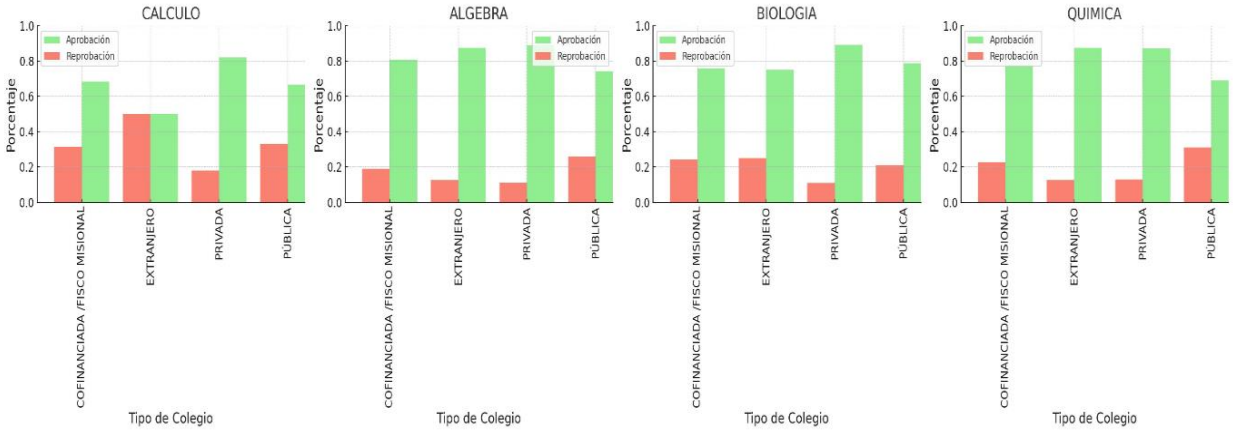
• **Cofinanciada / Fisco Misional:** Los estudiantes de colegios cofinanciados o fisco misionales tuvieron tasas de aprobación que oscilan entre el 65% y el 75%. Específicamente, en Cálculo, Álgebra, Biología y Química, las tasas de aprobación fueron del 68% 73% 75% y 72%, respectivamente.

• **Extranjero:** A pesar de que el número de estudiantes de colegios extranjeros podría ser menor en comparación con otras categorías, mostraron un rendimiento notable. En Cálculo y Álgebra, alrededor del 63% y 71% de los estudiantes lograron aprobar, respectivamente. Para Biología y Química, estos porcentajes fueron del 76% y 72%, respectivamente

Estos resultados sugieren que, aunque hay algunas diferencias en las tasas de aprobación según el tipo de colegio, en general, los estudiantes de diferentes tipos de colegios tienen la capacidad de desempeñarse bien en la universidad. Sin embargo, es crucial considerar otros factores y variables que pueden influir en el rendimiento académico al interpretar estos resultados.

Figura 9.

Porcentajes de aprobación y reprobación por asignatura según el tipo de colegio.



Nota. Gráfico de caja que muestra la distribución de notas en Cálculo, Álgebra, Biología y Química según el tipo de colegio: público, privado, cofinanciado/fiscomisional y extranjero.

4.4 Impacto de la Tenencia de Vivienda en el Rendimiento Académico

El contexto socioeconómico en el que un estudiante crece puede tener un efecto sobre su rendimiento académico en la universidad. Una de las variables que puede reflejar este contexto es la tenencia de vivienda. A continuación, se presenta un análisis basado en el boxplot generado para las notas de las asignaturas Cálculo, Álgebra, Biología y Química, segmentadas por la tenencia de vivienda de los estudiantes (Figura 10).

Basado en el boxplot y los datos:

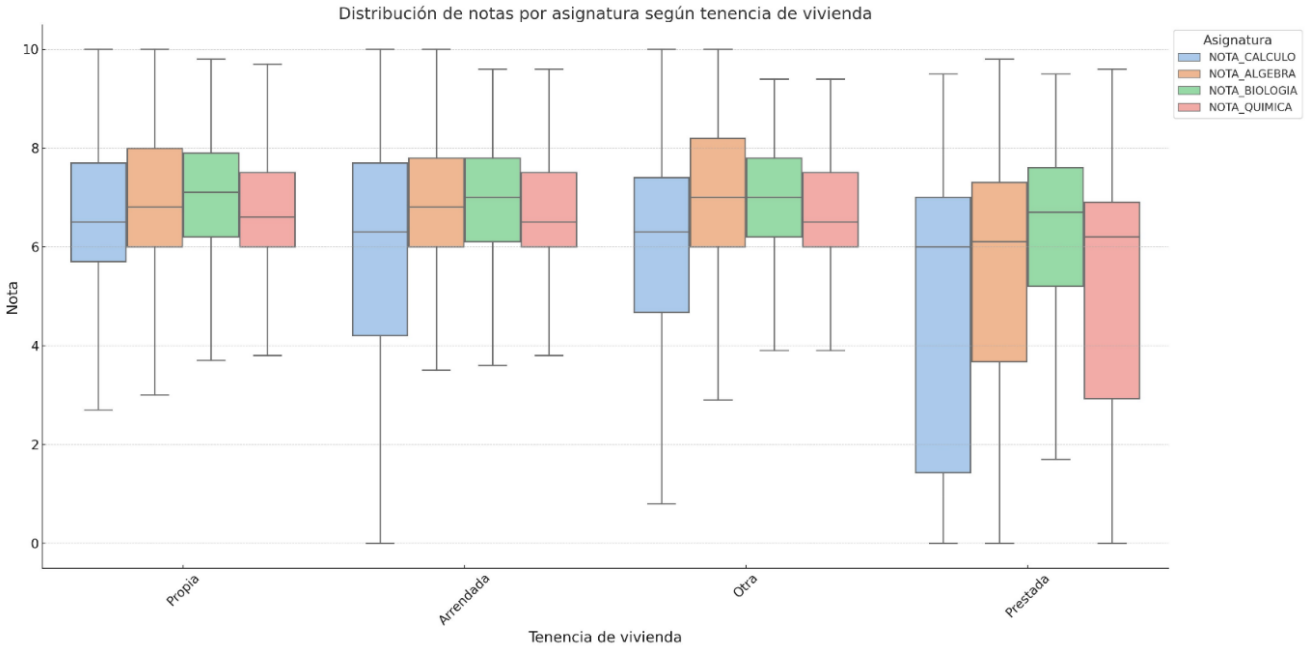
- **Arrendada:** Los estudiantes que viven en viviendas arrendadas obtuvieron medianas de 6.3 en Cálculo, 6.8 en Álgebra, 7.0 en Biología y 6.5 en Química.
- **Otra:** Bajo la categoría "Otra", las medianas de las notas fueron de 6.3 en Cálculo, 7.0 en Álgebra, 7.0 en Biología y 6.5 en Química.

- **Prestada:** Los estudiantes que viven en viviendas prestadas registraron medianas de 6.0 en Cálculo, 6.1 en Álgebra, 6.7 en Biología y 6.2 en Química.
- **Propia:** En cuanto a aquellos que viven en viviendas propias, las medianas se situaron en 6.5 para Cálculo, 6.8 para Álgebra, 7.1 para Biología y 6.6 para Química.

El tipo de tenencia de vivienda puede reflejar ciertos factores socioeconómicos que, directa o indirectamente, afectan el rendimiento académico. Es esencial considerar estas variables al evaluar y diseñar estrategias educativas.

Figura 10.

Distribución de notas por asignatura según tenencia de vivienda.



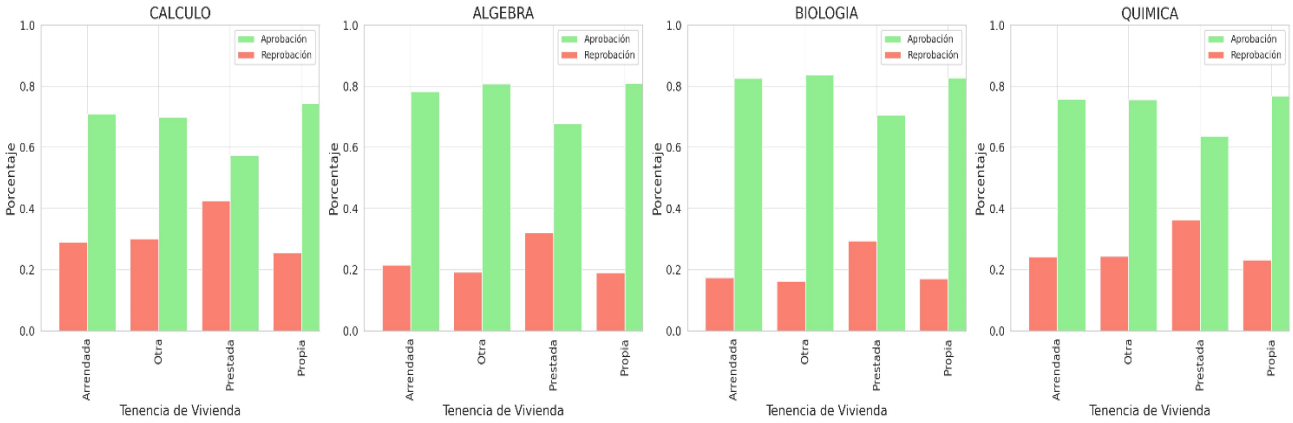
Nota. Gráfico de caja mostrando la distribución de notas por asignatura en relación a la tenencia de vivienda: propia, arrendada, otra y prestada.

A continuación, se presenta un análisis basado en los porcentajes de aprobación y reprobación de las asignaturas Cálculo, Álgebra, Biología y Química, segmentados por la tenencia de vivienda de los estudiantes (Figura 11).

- **Arrendada:** Los estudiantes que viven en viviendas arrendadas mostraron una tendencia positiva en términos de aprobación en las asignaturas. Por ejemplo, en Cálculo y Álgebra, más del 60% de estos estudiantes aprobaron.
- **Otra:** Bajo esta categoría, los porcentajes de aprobación son similares a los de viviendas arrendadas, con tasas de aprobación que rondan el 60-70% en las asignaturas analizadas.
- **Prestada:** Los estudiantes que viven en viviendas prestadas mostraron tasas de aprobación ligeramente más bajas en comparación con las categorías "Arrendada" y "Otra", especialmente en asignaturas como Cálculo.
- **Propia:** Los estudiantes que viven en viviendas propias mostraron tasas de aprobación comparables a las de las categorías "Arrendada" y "Otra", con un rendimiento ligeramente mejor en asignaturas como Biología y Química.

Figura 11.

Porcentajes de aprobación y reprobación por asignatura según la tenencia de vivienda.



Nota. Comparativa de porcentajes de aprobación y reprobación por asignatura, segmentados por situación de vivienda: arrendada, otra, prestada y propia.

4.5 Rendimiento Académico Según la Ocupación de los Padres

La ocupación de los padres puede reflejar, en cierta medida, el contexto socioeconómico en el que un estudiante crece. A continuación, se presenta un análisis basado en los boxplots generados para las notas de las asignaturas Cálculo, Álgebra, Biología y Química, segmentadas por la ocupación tanto del padre como de la madre de los estudiantes (Figura 12).

Ocupación del Padre:

Empleado:

- En Cálculo, los estudiantes con padres empleados tuvieron una media de nota de aproximadamente 6.0. La mayoría de las notas estuvieron en el rango de 5.75 (primer cuartil) a 7.8 (tercer cuartil).
- Para Álgebra, la media fue de 6.59. Las notas en su mayoría oscilaron entre 6.0 y 8.1.
- En Biología, la media se situó cerca de 6.75, con un rango intercuartílico de 6.4 a 7.8.
- En Química, la media fue de 6.09, y las notas se distribuyeron principalmente entre 6.0 y 7.5.

Empresario:

- En Cálculo, los estudiantes cuyos padres eran empresarios tuvieron una media de 4.10. Las notas en su mayoría oscilaron entre 0.0 y 7.15.
- Para Álgebra, la media fue de 6.08. Las notas en su mayoría oscilaron entre 6.0 y 8.0.
- En Biología, la media se situó cerca de 6.36, con un rango intercuartílico de 6.1 a 7.65.

- En Química, la media fue de 5.86, y las notas se distribuyeron principalmente entre 4.6 y 7.73.

Otros:

- En Cálculo, los estudiantes con padres en la categoría "Otros" tuvieron una media de nota de 5.61. La mayoría de las notas estuvieron en el rango de 4.15 (primer cuartil) a 7.5 (tercer cuartil), con notas extremas entre 0.0 y 10.0.
- Para Álgebra, la media fue de 6.10. Las notas en su mayoría oscilaron entre 6.0 y 7.9.
- En Biología, la media se situó cerca de 6.53, con un rango intercuartílico de 6.0 a 7.8.
- En Química, la media fue de 5.74, y las notas se distribuyeron principalmente entre 5.2 y 7.4.

Profesional:

- En Cálculo, los estudiantes con padres profesionales tuvieron una media de nota de 5.91. La mayoría de las notas estuvieron en el rango de 6.0 (primer cuartil) a 7.6 (tercer cuartil).
- Para Álgebra, la media fue de 6.71. Las notas en su mayoría oscilaron entre 6.1 y 8.1.
- En Biología, la media se situó cerca de 6.76, con un rango intercuartílico de 6.3 a 8.1.
- En Química, la media fue de 6.47, y las notas se distribuyeron principalmente entre 6.1 y 7.8

Ocupación de la Madre:

Empleado:

- En Cálculo, los estudiantes con madres empleadas tuvieron una media de nota de aproximadamente 6.02. La mayoría de las notas estuvieron en el rango de 6.0 (primer cuartil) a 7.73 (tercer cuartil.
- Para Álgebra, la media fue de 6.66. Las notas en su mayoría oscilaron entre 6.0 y 8.1.
- En Biología, la media se situó cerca de 6.80, con un rango intercuartílico de 6.4 a 7.9.
- En Química, la media fue de 6.24, y las notas se distribuyeron principalmente entre 6.0 y 7.5.

Empresario:

- En Cálculo, los estudiantes cuyas madres eran empresarias tuvieron una media de 5.19. Las notas en su mayoría oscilaron entre 1.23 y 8.03.
- Para Álgebra, la media fue de 6.31. Las notas en su mayoría oscilaron entre 6.0 y 7.8.
- En Biología, la media se situó cerca de 6.47, con un rango intercuartílico de 6.0 a 7.98.
- En Química, la media fue de 5.85, y las notas se distribuyeron principalmente entre 5.1 y 7.48.

Otros:

- En Cálculo, los estudiantes con madres en la categoría "Otros" tuvieron una media de nota de 5.64. La mayoría de las notas estuvieron en el rango de 4.2 (primer cuartil) a 7.5 (tercer cuartil).

- Para Álgebra, la media fue de 6.16. Las notas en su mayoría oscilaron entre 6.0 y 7.9.
- En Biología, la media se situó cerca de 6.59, con un rango intercuartílico de 6.0 a 7.8.
- En Química, la media fue de 5.81, y las notas se distribuyeron principalmente entre 5.48 y 7.4.

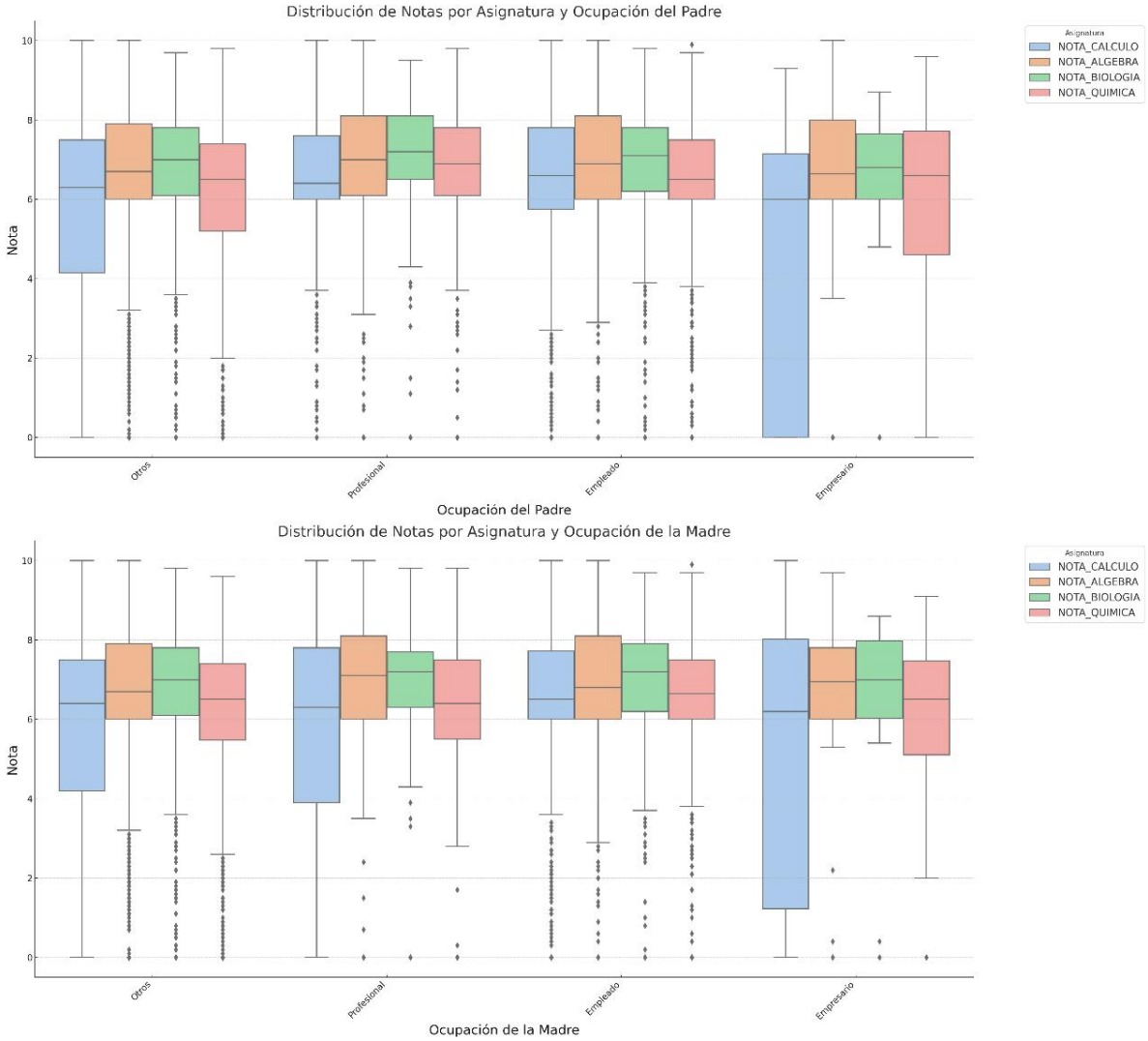
Profesional:

- En Cálculo, los estudiantes con madres profesionales tuvieron una media de nota de 5.55. La mayoría de las notas estuvieron en el rango de 3.9 (primer cuartil) a 7.8 (tercer cuartil).
- Para Álgebra, la media fue de 6.25. Las notas en su mayoría oscilaron entre 6.0 y 8.1.
- En Biología, la media se situó cerca de 6.78, con un rango intercuartílico de 6.2 a 7.7.
- En Química, la media fue de 5.80, y las notas se distribuyeron principalmente entre 5.5 y 7.5.

En general, parece que los estudiantes cuyos padres o madres tienen una ocupación de "Empleado" o "Profesional" tienden a tener un rendimiento ligeramente mejor en comparación con aquellos cuyos padres o madres tienen una ocupación de "Empresario" o "Otros". Sin embargo, las diferencias no son drásticas y es crucial considerar otros factores y variables que pueden influir en el rendimiento académico al interpretar estos resultados.

Figura 12.

Distribución notas por asignatura según ocupación de padres.



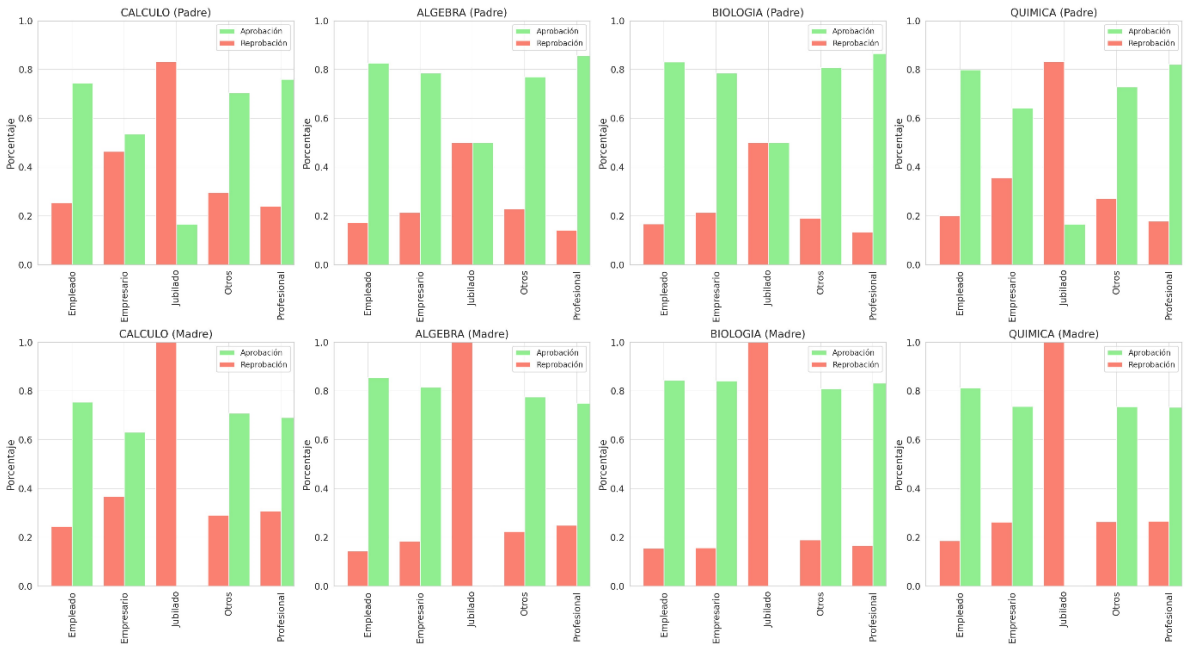
Nota. Distribución de notas en asignaturas según la ocupación de los padres, reflejando el impacto de los antecedentes laborales familiares en el rendimiento académico.

A continuación, se presenta un análisis basado en los porcentajes de aprobación y reprobación de las asignaturas Cálculo, Álgebra, Biología y Química, segmentados por la ocupación tanto del padre como de la madre de los estudiantes (Figura 13).

- **Ocupación del Padre:** En general, los estudiantes cuyos padres tienen ocupaciones como "Empleado" y "Profesional" tienden a tener mayores tasas de aprobación en las asignaturas analizadas. Por otro lado, es notable que, en algunas ocupaciones, como "Jubilado", las tasas de aprobación son considerablemente bajas en ciertas asignaturas, lo que indica que puede haber factores adicionales que afectan el rendimiento de estos estudiantes.
- **Ocupación de la Madre:** De manera similar, los estudiantes cuyas madres tienen ocupaciones como "Empleado" y "Profesional" tienden a tener tasas de aprobación más altas. Las tasas de reprobación son más pronunciadas en ciertas categorías, como "Jubilado", lo que sugiere la necesidad de investigaciones adicionales para comprender las razones subyacentes.

Figura 13.

Porcentajes de aprobación y reprobación por asignatura según la ocupación del padre y de la madre.



Nota. Gráficos de barras mostrando la tasa de aprobación y reprobación en asignaturas por la ocupación de los padres y madres: Empleado, Empresario, Público, y Otro.

Tras el análisis inicial de las variables sociodemográficas de los estudiantes, la investigación avanzó hacia una fase más profunda y técnica. Se enfocó en el desarrollo y comparación de modelos predictivos para entender cómo estos factores sociodemográficos podrían influir en el rendimiento académico, en particular en las asignaturas de Cálculo, Álgebra, Química y Biología. Para ello, se emplearon técnicas avanzadas de regresión logística con regularizaciones, Lasso (L1) y Ridge (L2), con el objetivo de proporcionar predicciones precisas y robustas.

4.6 Cálculo: Análisis Predictivo del Rendimiento Académico

Las métricas son esenciales para evaluar y comparar la efectividad de los modelos predictivos. En la Tabla 5 ambos modelos de regresión, Lasso y Ridge presentan un buen desempeño en su capacidad de clasificación. Al evaluar la precisión, Lasso muestra una ligera ventaja en el conjunto de entrenamiento con un 0.8927, al igual que en el conjunto de prueba, Lasso lleva la delantera con 0.8812 contra el 0.8739 de Ridge. En cuanto a la sensibilidad, Lasso se destaca ligeramente en ambos conjuntos con 0.7417 en entrenamiento y 0.7564 en prueba, en comparación con los 0.7108 y 0.7047 de Ridge. La especificidad es bastante similar entre ambos modelos, pero Ridge tiene una pequeña ventaja en el conjunto de entrenamiento y prueba con 0.9517 y 0.9296 respectivamente. Con relación al AUC, que evalúa la habilidad del modelo para distinguir entre clases, Lasso tiene un 0.937 en el conjunto de prueba, que es apenas superior al 0.935 de Ridge. Si bien Lasso tiene ventajas numéricas en algunas métricas, las diferencias entre ambos modelos son pequeñas, lo que indica que ambos ofrecen un rendimiento robusto y comparable.

Tabla 5.

Comparativa de métricas entre modelos Lasso y Ridge – Cálculo.

Métricas	Lasso (Entrenamiento)	Lasso (Prueba)	Ridge (Entrenamiento)	Ridge (Prueba)
Precisión	0.8927	0.8812	0.8872	0.8739
Sensibilidad	0.7417	0.7565	0.7108	0.7047
Especificidad	0.9517	0.9296	0.9560	0.9396
Kappa	0.7231	0.6994	0.7048	0.6731
AUC	0.941	0.937	0.942	0.935

El modelo Lasso destaca por sus ligeras ventajas en métricas como exactitud y sensibilidad. Variables como "Trabajo: Sí" y "Nota de grado" tienen coeficientes significativos, lo que sugiere que existen relaciones interesantes entre estas variables y el rendimiento académico. Estas interpretaciones pueden encontrarse en la (Tabla 6).

Interpretación de Coeficientes

Modelo Lasso:

1. **Asignaturas_Matriculadas_4 (-3.49):** Estudiantes matriculados en cuatro asignaturas tienen una menor probabilidad de aprobar Cálculo. Esto puede deberse a que tienen más carga académica y menos tiempo para concentrarse en Cálculo.
2. **Hijos_Si (-3.32):** Los estudiantes que tienen hijos tienen menos probabilidad de aprobar Cálculo en comparación con aquellos que no tienen hijos.

3. **Etnia_no_registra (-3.28):** Estudiantes que no tienen una etnia registrada tienen menos probabilidad de aprobar Cálculo. Este factor puede necesitar más análisis para entender completamente su impacto.
4. **Nota_Grado (3.11):** Estudiantes con notas más altas en general tienen una mayor probabilidad de aprobar Cálculo. Un aumento en la nota del grado se asocia con una mayor probabilidad de éxito en Cálculo.
5. **Trabajo_Si (2.16):** Los estudiantes que trabajan tienen una probabilidad más alta de aprobar Cálculo en comparación con aquellos que no trabajan.
6. **Etnia_Negro (2.08):** Estudiantes identificados como de etnia negra tienen una mayor probabilidad de aprobar Cálculo. Es importante abordar estos resultados con precaución y consideración ética.
7. **Segunda_Matrícula_Álgebra (-1.73):** Estudiantes matriculados por segunda vez en Álgebra tienen una menor probabilidad de aprobar Cálculo.
8. **Tres_Asignaturas_Aprobadas (-1.69):** Quienes han aprobado solo tres asignaturas tienen menos probabilidad de aprobar Cálculo.
9. **Mátricula_Nivelación (-1.42):** Estudiantes con más de una matrícula en nivelación tienen menos probabilidad de aprobar Cálculo.
10. **Biología_Aprueba (1.28):** Estudiantes que aprueban Biología tienen una mayor probabilidad de aprobar Cálculo.

Modelo Ridge:

Los coeficientes del modelo Ridge son bastante similares a los del modelo Lasso, reflejando patrones y relaciones parecidas en los datos. Por ejemplo, el coeficiente para "Trabajo_Si" en el modelo Ridge es 1.74, ligeramente inferior al de Lasso que es 2.16, pero la interpretación esencialmente indica que trabajar aumenta la probabilidad de aprobar. Es relevante entender estas pequeñas diferencias y tener

en cuenta que, aunque las magnitudes varíen ligeramente, ambos modelos están en sintonía respecto a la dirección y significado de las relaciones.

Tabla 6.

Características influyentes para el modelo Lasso y Ridge–Cálculo

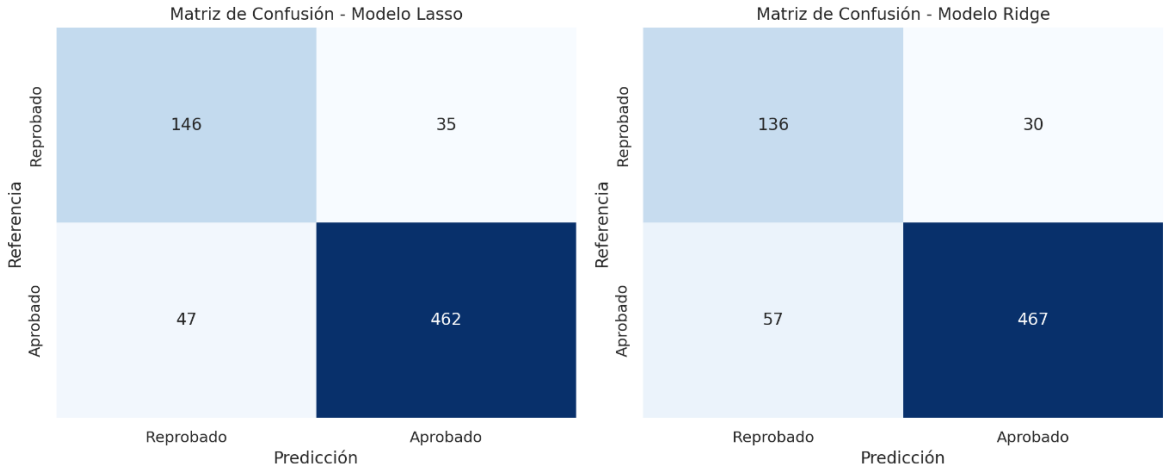
Características Lasso	Coe	Características Ridge	Coe
Asignaturas_Matriculadas_4	-3.49	Etnia_Negro/a	3.03
Hijos_Si	-3.33	Hijos_Si	-2.30
Etnia_no_registra	-3.29	Nota_grado	2.10
Nota_grado	3.11	Etnia_no_registra	-1.94
Trabajo_Si	2.17	Provincia_residencia_orellana	1.86
Etnia_Negro/a	2.08	Estado_civil_divorciado/a	1.77
Segunda_Matrícula_Álgebra	-1.73	Trabajo_Si	1.74
3_asignaturas_aprobadas	-1.69	Colegio_extranjero_privado	-1.26
Matrícula_nivelación	-1.43	Tercera_Matrícula_Cálculo	-1.24
Biología_Aprueba	1.28	Estado_civil_soltero/a	1.20

Matrices de Confusión

Estas matrices proporcionan un desglose de las predicciones verdaderas y erróneas de cada modelo. La visualización de estas matrices para los modelos Lasso y Ridge se encuentra en la Figura *14*.

Figura 14.

Matriz de Confusión para Modelo Lasso y Ridge – Cálculo.



Nota: Matrices de confusión comparando la precisión de los modelos Lasso y Ridge en la clasificación de aprobados y reprobados.

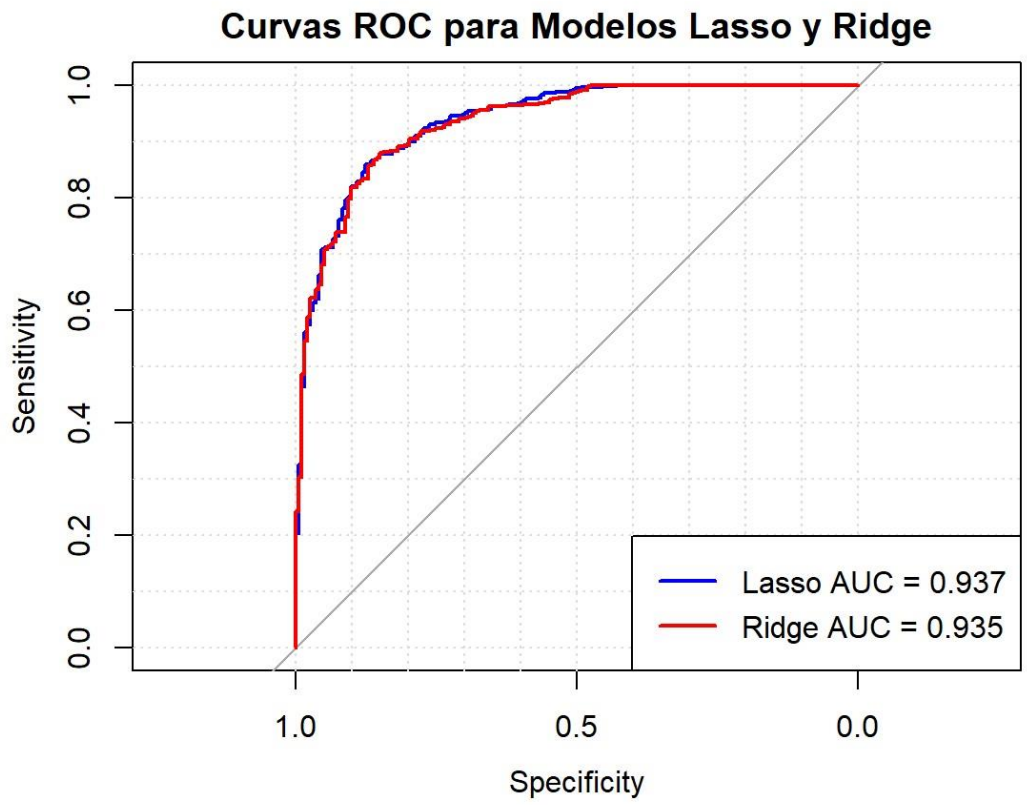
La matriz de confusión del Modelo Lasso revela que 146 estudiantes fueron correctamente identificados como reprobados, mientras que 462 fueron acertadamente reconocidos como aprobados. No obstante, el modelo también incurrió en errores, identificando incorrectamente a 35 estudiantes como aprobados cuando en realidad reprobaron, y a 47 como reprobados cuando aprobaron. Por otro lado, el Modelo Ridge acertó en 136 casos de estudiantes reprobados y en 467 de aprobados, pero se equivocó en 30 casos pronosticando que aprobarían cuando reprobaron y en 57 casos donde indicó que reprobarían y aprobaron. En resumen, mientras que Lasso tiene una ligera ventaja al identificar a los estudiantes que reprobarán, ambos modelos muestran un rendimiento similar en la identificación de estudiantes que aprobarán. La elección entre usar el Modelo Lasso o Ridge dependerá del tipo de error que se considere más importante minimizar en la situación específica de aplicación.

Comparación

Al evaluar ambos modelos en su totalidad, y considerando tanto las métricas de evaluación como las variables influyentes, el modelo Lasso podría ser la elección preferida para futuras predicciones. Sin embargo, es vital interpretar estos resultados en el contexto de la Universidad Yachay Tech y su población estudiantil. Las curvas ROC para ambos modelos están ilustradas en la Figura *15*, proporcionando una representación visual de su desempeño.

Figura 15.

Curvas ROC para Modelos Lasso y Ridge – Cálculo.



4.7 Química: Análisis Predictivo del Rendimiento Académico

Métricas de Evaluación

Las métricas son esenciales para evaluar y comparar la efectividad de los modelos predictivos. En la *Tabla 7* ambos modelos de regresión, Lasso y Ridge demuestran

altos niveles de rendimiento. Al evaluar la precisión, Lasso muestra una ligera ventaja en el conjunto de entrenamiento con un 0.9374, al igual que en el conjunto de prueba, Lasso lleva la delantera con 0.9275 contra el 0.9246 de Ridge. En cuanto a la sensibilidad, Lasso se destaca ligeramente en ambos conjuntos con 0.8147 en entrenamiento y 0.7976 en prueba, en comparación con los 0.7843 y 0.7738 de Ridge. La especificidad es bastante similar entre ambos modelos, pero Ridge tiene una pequeña ventaja en el conjunto de entrenamiento y prueba con 0.9779 y 0.9732 respectivamente. Con relación al AUC, que evalúa la habilidad del modelo para distinguir entre clases, Lasso tiene un 0.952 en el conjunto de prueba, que es apenas superior al 0.951 de Ridge. Si bien Lasso tiene ventajas numéricas en algunas métricas, las diferencias entre ambos modelos son pequeñas, lo que indica que ambos ofrecen un rendimiento robusto y comparable.

Tabla 7.

Comparativa de métricas entre modelos Lasso y Ridge–Química

Métricas	Lasso (Entrenamiento)	Lasso (Prueba)	Ridge (Entrenamiento)	Ridge (Prueba)
Precisión	0.9374	0.9275	0.9306	0.9246
Sensibilidad	0.8147	0.7976	0.7843	0.7738
Especificidad	0.9770	0.9693	0.9779	0.9732
Kappa	0.8236	0.7959	0.8021	0.785
AUC	0.978	0.952	0.979	0.951

El modelo Lasso destaca por sus ligeras ventajas en métricas como exactitud y sensibilidad. Variables como "Trabajo: Sí" y "Nota de grado" tienen coeficientes

significativos, lo que sugiere que existen relaciones interesantes entre estas variables y el rendimiento académico. Estas interpretaciones pueden encontrarse en la (Tabla 8).

Interpretación de Coeficientes

Modelo Lasso:

1. **Álgebra_Aprueba (4.80):** Los estudiantes que han aprobado Álgebra tienen una mayor probabilidad de aprobar Química. Esto puede deberse a que Álgebra proporciona una base sólida que es esencial para los cursos subsecuentes.
2. **Tercera_Matrícula_Química (-3.76):** Estudiantes matriculados por tercera vez en Química tienen una menor probabilidad de aprobar Química.
3. **Tercera_Matrícula_Álgebra (2.21):** Estudiantes matriculados por tercera vez en Álgebra tienen una mayor probabilidad de aprobar Química.
4. **Segunda_Matrícula_Química (-1.93):** Estudiantes matriculados por segunda vez en Química tienen una menor probabilidad de aprobar Química.
5. **Carrera_Petroquímica (1.65):** Los estudiantes en la carrera de Petroquímica tienen una mayor probabilidad de aprobar, lo que podría reflejar la naturaleza del programa o del perfil estudiantil de esta carrera.
6. **Etnia_no_registra (-1.63):** Los estudiantes que no registran una etnia tienen una menor probabilidad de aprobar. Este factor puede necesitar más análisis para entender completamente su impacto.
7. **Colegio_extranjero_público (1.52):** Los estudiantes de colegios extranjeros públicos tienden a tener una mayor probabilidad de aprobar, lo que podría deberse a métodos de enseñanza o estándares académicos diferentes en estos colegios.

8. **Provincia_residencia_esmeraldas (1.51):** Los estudiantes residentes en Esmeraldas tienen una probabilidad ligeramente mayor de aprobar. Este dato puede reflejar factores socioeconómicos o educativos específicos de la región que influyen en el rendimiento académico.
9. **Asignaturas_Matriculadas_6 (1.25):** Los estudiantes que se matriculan en seis asignaturas durante el primer año tienen una mayor probabilidad de aprobar. Este factor podría estar asociado a un mayor compromiso o a una preparación académica más sólida.
10. **Provincia_residencia_orellana (1.17):** Residir en Orellana también incrementa la probabilidad de aprobar. Es necesario un análisis más detallado para interpretar correctamente este coeficiente.

Modelo Ridge: Los coeficientes del modelo Ridge son bastante similares a los del modelo Lasso. Esto indica que ambos modelos identifican patrones similares en los datos. Sin embargo, hay ligeras diferencias en la magnitud de los coeficientes.

Tabla 8.

Características influyentes para el modelo Lasso y Ridge–Química

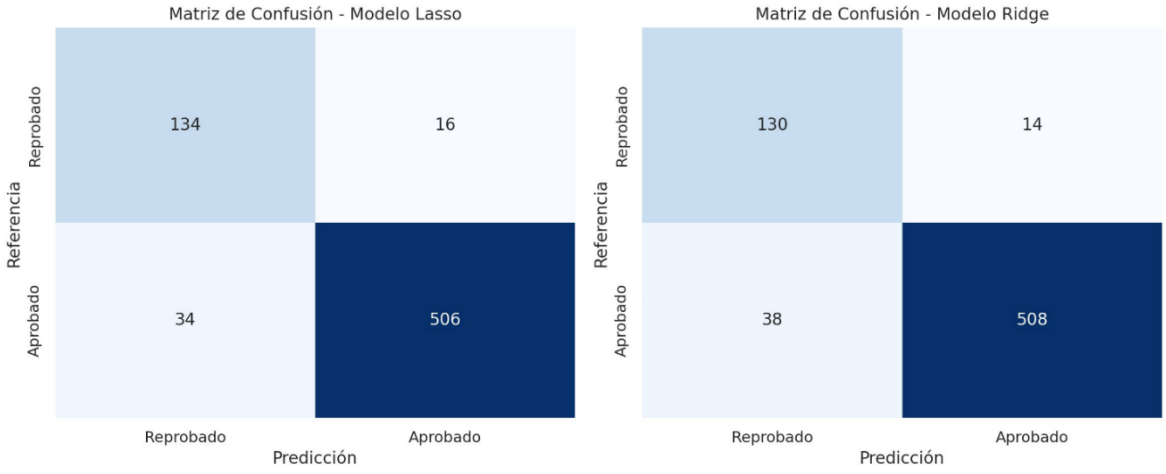
Características Lasso	Coe	Características Ridge	Coe
Álgebra_Aprueba	4.80	Aprueba_Álgebra	3.19
Tercera_Matrícula_Química	-3.76	Tercera_Matrícula_Química	-2.40
Tercera_Matrícula_Álgebra	2.21	Colegio_extranjero_público	2.27
Segunda_Matrícula_Química	-1.93	Provincia_residencia_Orellana	1.67
Carrera_Petroquímica	1.65	Tercera_Matrícula_Álgebra	1.18
Etnia_no_registra	-1.63	Etnia_no_registra	-1.02
Colegio_extranjero_público	1.52	Provincia_residencia_Esmeraldas	0.99
Provincia_residencia_Esmeraldas	1.51	Segunda_Matrícula_Química	-0.99
Asignaturas_Matriculadas_6	1.25	Carrera_Petroquimica	0.93
Provincia_residencia_Orellana	1.17	Hijos_Si	-0.84

Matrices de Confusión

Estas matrices proporcionan un desglose de las predicciones verdaderas y erróneas de cada modelo. La visualización de estas matrices para los modelos Lasso y Ridge se encuentra en la Figura *16*.

Figura 16.

Matriz de Confusión para Modelo Lasso y Ridge – Química.



Nota. Matrices de confusión comparando la precisión de los modelos Lasso y Ridge en la clasificación de aprobados y reprobados.

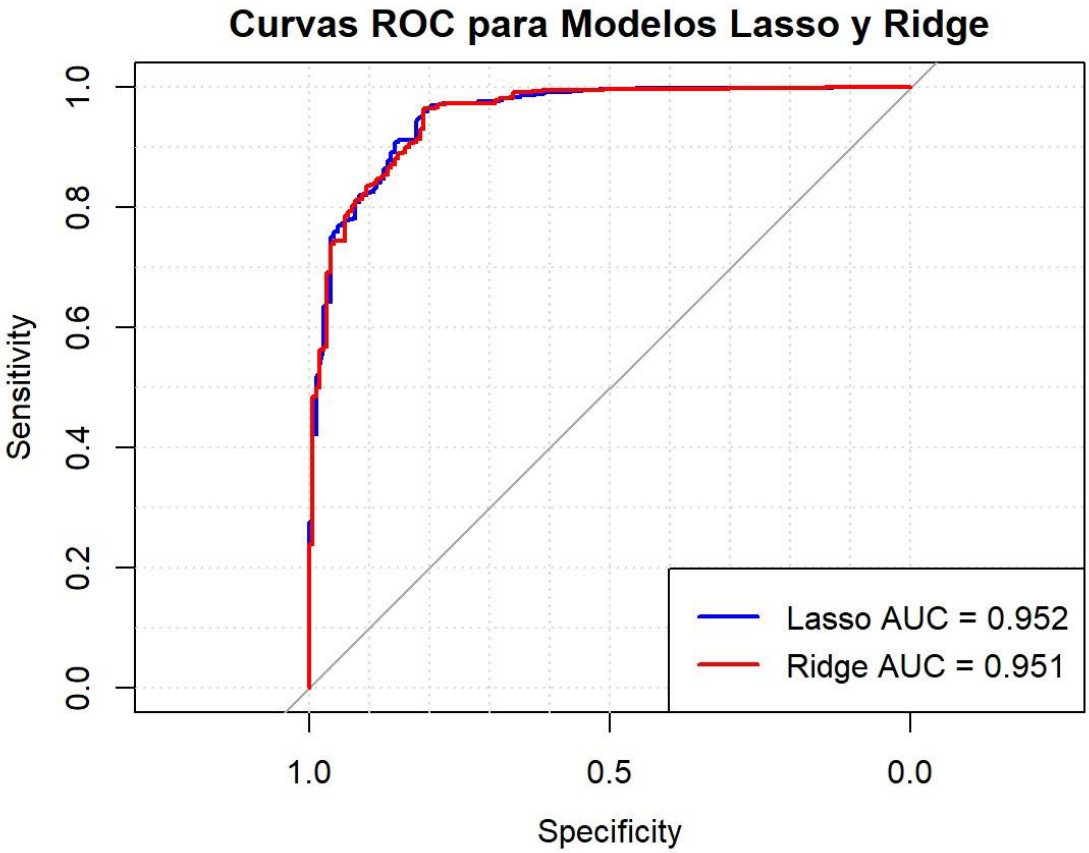
Comparación

Al evaluar ambos modelos en su totalidad, y considerando tanto las métricas de evaluación como las variables influyentes, el modelo Lasso podría ser la elección preferida para futuras predicciones. Sin embargo, es vital interpretar estos resultados en el contexto de la Universidad Yachay Tech y su población estudiantil. Las curvas

ROC para ambos modelos están ilustradas en la Figura *17*, proporcionando una

Figura 17.

Curvas ROC para Modelos Lasso y Ridge – Química.



representación visual de su desempeño.

4.8 Álgebra: Análisis Predictivo del Rendimiento Académico

Métricas de Evaluación

Las métricas son esenciales para evaluar y comparar la efectividad de los modelos predictivos. En la

Tabla 9 ambos modelos de regresión, Lasso y Ridge demuestran altos niveles de rendimiento. Al evaluar la precisión, Ridge muestra una ligera ventaja en el conjunto de entrenamiento con un 0.9361, sin embargo, en el conjunto de prueba, Lasso lleva la delantera con 0.9449 contra el 0.933 de Ridge. En cuanto a la sensibilidad, Lasso

se destaca en ambos conjuntos con 0.8701 en entrenamiento y 0.9078 en prueba, en comparación con los 0.8187 y 0.8156 de Ridge. La especificidad es bastante similar entre ambos modelos, pero Ridge tiene una pequeña ventaja en el conjunto de entrenamiento y prueba con 0.9665 y 0.9636 respectivamente. Con relación al AUC, que evalúa la habilidad del modelo para distinguir entre clases, Lasso tiene un 0.973 en el conjunto de prueba, que es apenas superior al 0.9672 de Ridge. Si bien Lasso tiene ventajas numéricas en algunas métricas, las diferencias entre ambos modelos son pequeñas, lo que indica que ambos ofrecen un rendimiento robusto y comparable.

Tabla 9.

Comparativa de métricas entre modelos Lasso y Ridge–Álgebra

Métricas	Lasso (Entrenamiento)	Lasso (Prueba)	Ridge (Entrenamiento)	Ridge (Prueba)
Precisión	0.933	0.9449	0.9361	0.9333
Sensibilidad	0.8701	0.9078	0.8187	0.8156
Especificidad	0.9493	0.9545	0.9665	0.9636
Kappa	0.7997	0.8358	0.8004	0.7917
AUC	0.972	0.9699	0.973	0.9672

El modelo Lasso destaca por sus ligeras ventajas en métricas como exactitud y sensibilidad. Variables como "Trabajo: Sí" y "Nota de grado" tienen coeficientes significativos, lo que sugiere que existen relaciones interesantes entre estas variables y el rendimiento académico. Estas interpretaciones pueden encontrarse en la (Tabla 10).

Interpretación de Coeficientes

Modelo Lasso:

1. **Química_Aprueba (4.20):** Los estudiantes que han aprobado Química tienen una mayor probabilidad de aprobar Álgebra.
2. **Tercera_Matrícula_Álgebra (-3.22):** Estudiantes matriculados por tercera vez en Álgebra tienen una menor probabilidad de aprobar Álgebra.
3. **Tercera_Matrícula_Química (1.95):** Estudiantes matriculados por tercera vez en Química tienen una mayor probabilidad de aprobar Álgebra.
4. **Segunda_Matrícula_Álgebra (-1.38):** Estudiantes matriculados por segunda vez en Álgebra tienen una menor probabilidad de aprobar Álgebra pero más probabilidad que una tercera matrícula.
5. **Segunda_Matrícula_Química (1.34):** Estudiantes matriculados por segunda vez en Química tienen una mayor probabilidad de aprobar Álgebra, pero menos probabilidad que una tercera matrícula.
6. **Provincia_residencia_pastaza (-1.08):** Los estudiantes residentes en Pastaza tienen una probabilidad ligeramente menor de aprobar Álgebra. Es posible que existan factores socioeconómicos, culturales o educativos en Pastaza que influyan en el rendimiento académico de los estudiantes en Álgebra.
7. **Trabajo_Si (0.84):** Los estudiantes que trabajan tienen una probabilidad ligeramente mayor de aprobar Álgebra. Este resultado podría indicar que los estudiantes que trabajan desarrollan habilidades de disciplina y gestión del tiempo que contribuyen a su rendimiento académico.
8. **Provincia_residencia_Galapagos (0.78):** Residir en las Galápagos está asociado con una mayor probabilidad de aprobar Álgebra. Este coeficiente

podría reflejar características únicas de los estudiantes o del sistema educativo en las Galápagos.

9. **Biología_Aprueba (0.78):** Los estudiantes que han aprobado Biología tienen una mayor probabilidad de aprobar Álgebra. Aprobar Biología podría indicar un nivel general de habilidad académica o preparación que también beneficia el rendimiento en Álgebra.
10. **Carrera_polímeros (0.73):** Los estudiantes en la carrera de Polímeros tienen una mayor probabilidad de aprobar Álgebra. Este resultado podría reflejar la naturaleza rigurosa y técnica del programa de Polímeros que prepara a los estudiantes para el éxito en asignaturas como Álgebra.

Modelo Ridge: Los coeficientes del modelo Ridge muestran ciertas similitudes con los del modelo Lasso, lo que sugiere que ambos modelos están reconociendo tendencias y patrones comparables en los datos. Sin embargo, existen sutiles discrepancias en la magnitud de algunos coeficientes entre los dos modelos. Estas diferencias resaltan la importancia de considerar múltiples modelos y técnicas al analizar y predecir comportamientos en datos complejos.

Tabla 10.

Características influyentes para modelo Lasso y Ridge–Álgebra

Características Lasso	Coe	Características Ridge	Coe
Química_Aprueba	4.20	Química_Aprueba	2.96
Tercera_Matrícula_Álgebra	-3.22	Tercera_Matrícula_Álgebra	-2.03
Tercera_Matrícula_Química	1.95	Provincia_residencia_Galapagos	1.59
Segunda_Matrícula_Álgebra	-1.38	Colegio_extranjero_público	1.53
Segunda_Matrícula_Química	1.34	Provincia_residencia_Pastaza	-1.39
Provincia_residencia_Pastaza	-1.08	Tercera_Matrícula_Química	1.26

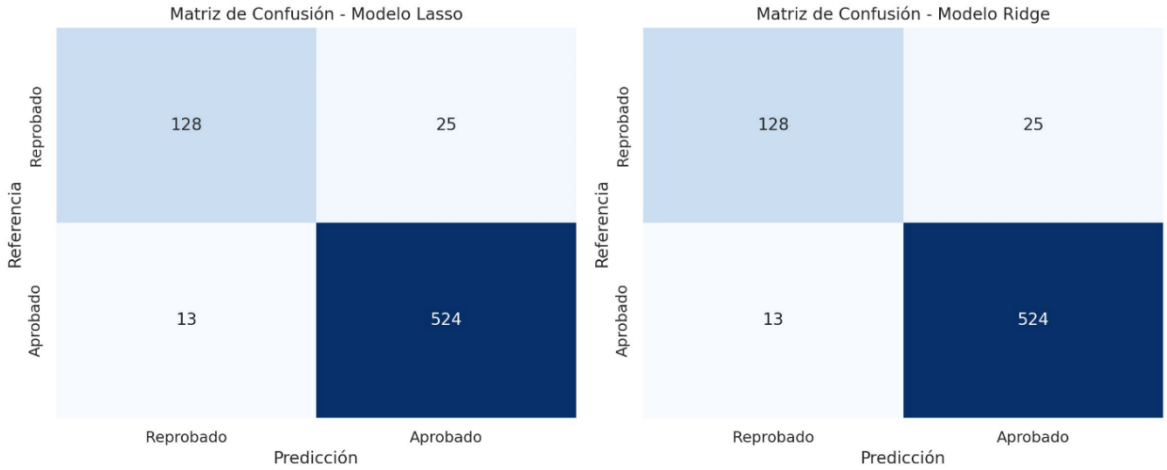
Trabajo_Si	0.84	Provincia_residencia_Napo	-1.07
Provincia_residencia_Galapagos	0.78	Carrera_Materiales	-1.04
Biología_Aprueba	0.78	Provincia_residencia_Morona_Santiago	-1.01
Carrera_polímeros	0.73	Provincia_residencia_Sucumbios	1.00

Matrices de Confusión

Estas matrices proporcionan un desglose de las predicciones verdaderas y erróneas de cada modelo. La visualización de estas matrices para los modelos Lasso y Ridge se encuentra en la Figura *18*.

Figura 18.

Matriz de Confusión para Modelo Lasso y Ridge – Álgebra.



Nota. Matrices de confusión comparando la precisión de los modelos Lasso y Ridge en la clasificación de aprobados y reprobados.

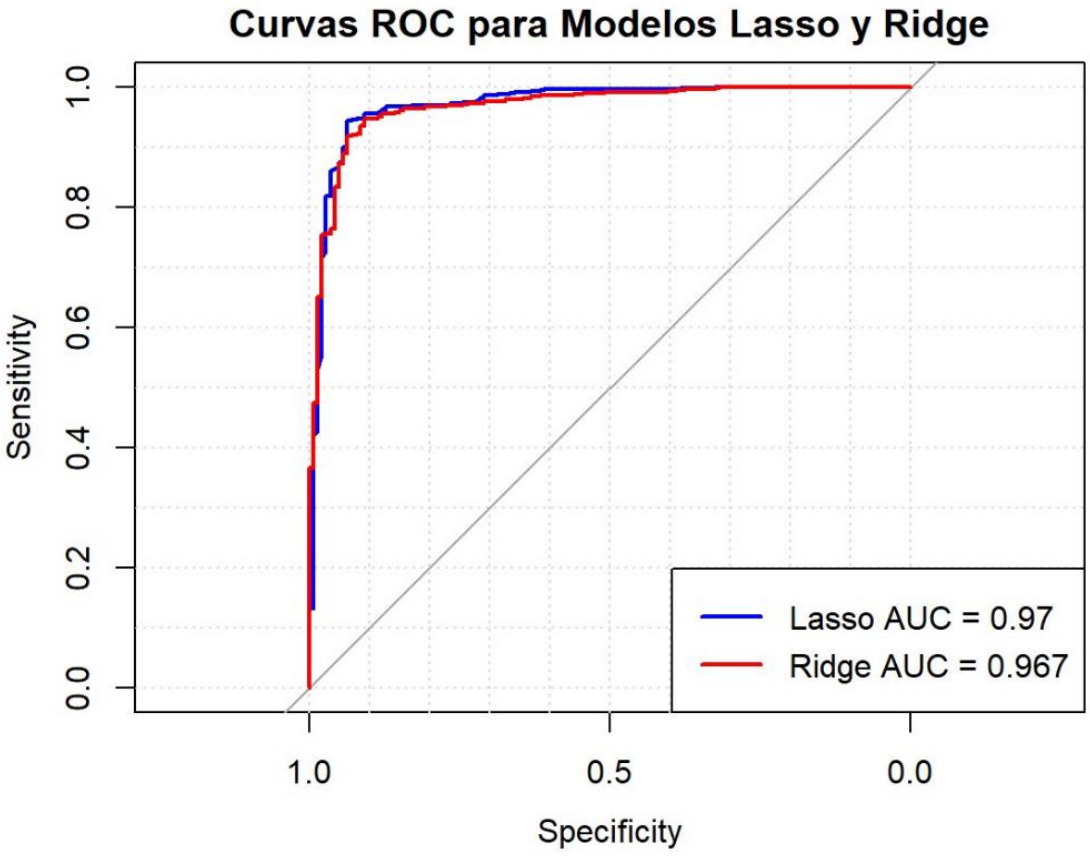
Comparación

Al evaluar ambos modelos en su totalidad, y considerando tanto las métricas de evaluación como las variables influyentes, el modelo Lasso podría ser la elección preferida para futuras predicciones. Sin embargo, es vital interpretar estos resultados

en el contexto de la Universidad Yachay Tech y su población estudiantil. Las curvas ROC para ambos modelos están ilustradas en la Figura *19*.

Figura 19.

Curvas ROC para Modelos Lasso y Ridge – Álgebra.



4.9 Biología: Análisis Predictivo del Rendimiento Académico

Métricas de Evaluación

Las métricas son esenciales para evaluar y comparar la efectividad de los modelos predictivos. En la (

*Tabla **11***) ambos modelos de regresión, Lasso y Ridge demuestran altos niveles de rendimiento. Al evaluar la precisión, Ridge muestra una ligera ventaja en el conjunto de entrenamiento con un 0.8795, sin embargo, en el conjunto de prueba, Lasso lleva

la delantera con 0.8681 contra el 0.8623 de Ridge. En cuanto a la sensibilidad, Lasso se destaca en ambos conjuntos con 0.4586 en entrenamiento y 0.4471 en prueba, en comparación con los 0.4276 y 0.3821 de Ridge. La especificidad es bastante similar entre ambos modelos, pero Ridge tiene una pequeña ventaja en el conjunto de entrenamiento y prueba con 0.9777 y 0.9665 respectivamente. Con relación al AUC, que evalúa la habilidad del modelo para distinguir entre clases, Lasso tiene un 0.849 en el conjunto de prueba, que es apenas superior al 0.835 de Ridge. Si bien Lasso tiene ventajas numéricas en algunas métricas, las diferencias entre ambos modelos son pequeñas, lo que indica que ambos ofrecen un rendimiento robusto y comparable.

Tabla 11.

Comparativa de métricas entre modelos Lasso y Ridge–Biología

Métricas	Lasso (Entrenamiento)	Lasso (Prueba)	Ridge (Entrenamiento)	Ridge (Prueba)
Precisión	0.8785	0.8681	0.8795	0.8623
Sensibilidad	0.4586	0.4471	0.4276	0.3821
Especificidad	0.9705	0.9594	0.9777	0.9665
Kappa	0.8102	0.7746	0.7957	0.7259
AUC	0.872	0.849	0.873	0.835

El modelo Lasso destaca por sus ligeras ventajas en métricas como exactitud y sensibilidad. Variables como "Trabajo: Sí" y "Nota de grado" tienen coeficientes significativos, lo que sugiere que existen relaciones interesantes entre estas variables y el rendimiento académico. Estas interpretaciones pueden encontrarse en la (Tabla 12).

Interpretación de Coeficientes

Modelo Lasso:

1. **Cálculo_Aprueba (1.19):** Los estudiantes que han aprobado Cálculo tienen una mayor probabilidad de aprobar Biología.
2. **Química_Aprueba (1.06):** Los estudiantes que han aprobado Química tienen una mayor probabilidad de aprobar Biología.
3. **Etnia_Mulato/a (0.90):** Los estudiantes que registran una etnia mulato/a tienen una mayor probabilidad de aprobar. Este factor puede necesitar más análisis para entender completamente su impacto.
4. **Provincia_residencia_Sucumbíos (0.87)**: Los estudiantes residentes en Sucumbíos tienen una probabilidad mayor de aprobar Biología Es posible que existan factores socioeconómicos, culturales o educativos en Sucumbíos que influyan en el rendimiento académico de los estudiantes en Biología.
5. **Trabajo_Si (0.86):** Los estudiantes que trabajan tienen una probabilidad ligeramente mayor de aprobar Biología. Este resultado podría indicar que los estudiantes que trabajan desarrollan habilidades de disciplina y gestión del tiempo que contribuyen a su rendimiento académico.
6. **Provincia_residencia_Orellana (-0.80):** Los estudiantes residentes en Orellana tienen una probabilidad ligeramente menor de aprobar Biología. Es posible que existan factores socioeconómicos, culturales o educativos en Orellana que influyan en el rendimiento académico de los estudiantes en Biología.
7. **Segunda_Matrícula_Química (0.70):** Estudiantes matriculados por segunda vez en Química tienen una mayor probabilidad de aprobar Biología, pero menos probabilidad que una tercera matrícula.

8. **Carrera_materiales (-0.67):** Los estudiantes en la carrera de materiales tienen una mayor probabilidad de aprobar Biología. Este resultado podría reflejar la naturaleza rigurosa y técnica del programa de materiales que prepara a los estudiantes para el éxito en asignaturas como Biología.
9. **Matrícula_Nivelación (0.63):** Estudiantes con más de una matrícula en nivelación tienen menos probabilidad de aprobar Nivelación.
10. **Carrera_Computación (-0.63):** Estudiantes matriculados en la carrera de Computación tienen menos probabilidad de aprobar la asignatura de Biología.

Modelo Ridge: Los coeficientes del modelo Ridge muestran ciertas similitudes con los del modelo Lasso, lo que sugiere que ambos modelos están reconociendo tendencias y patrones comparables en los datos. Sin embargo, existen sutiles discrepancias en la magnitud de algunos coeficientes entre los dos modelos. Estas diferencias resaltan la importancia de considerar múltiples modelos y técnicas al analizar y predecir comportamientos en datos complejos.

Tabla 12.

Características influyentes para modelo Lasso y Ridge–Biología

Características Lasso	Coe	Características Ridge	Coe
Cálculo_Aprueba	1.19	Provincia_residencia_Sucumbíos	1.71
Química_Aprueba	1.06	Etnia_Mulato/a	1.44
Etnia_Mulato/a	0.90	Provincia_residencia_Orellana	-1.10
Provincia_residencia_Sucumbíos	0.87	Etnia_Negro/a	0.92
Trabajo_Si	0.86	Carrera_Materiales	-0.90
Provincia_residencia_Orellana	-0.80	Provincia_residencia_Zamora_Chinchipe	0.90
Segunda_Matrícula_Química	0.70	Trabajo_Si	0.80
Carrera_Materiales	-0.67	Provincia_residencia_Galapagos	-0.80
Matrícula_Nivelación	0.63	Cálculo_Aprueba	0.79

Carrera_Computación	-0.63	Química_Aprueba		0.79

Matrices de Confusión

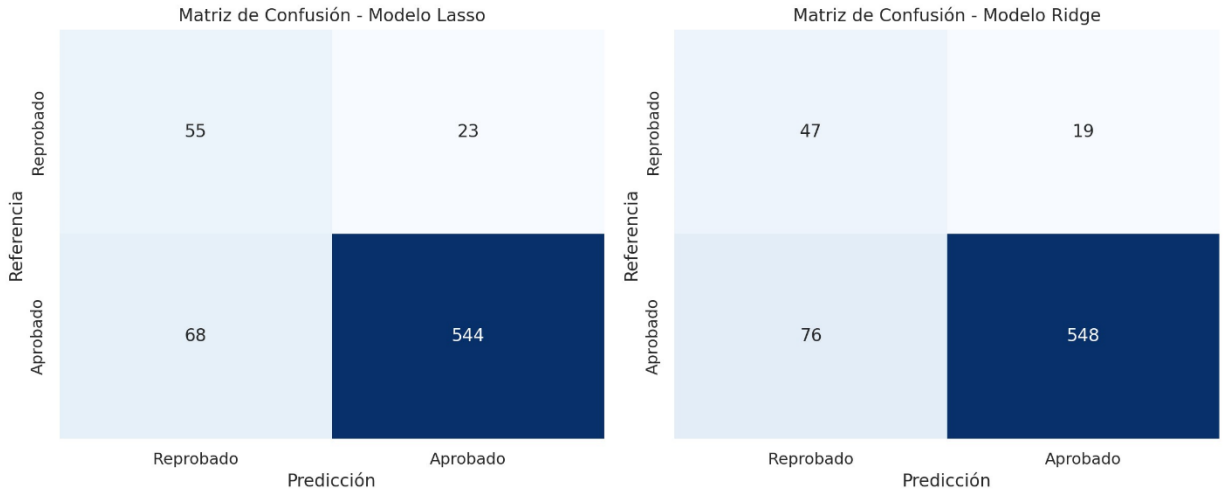
Estas matrices proporcionan un desglose de las predicciones verdaderas y erróneas de cada modelo. La visualización de estas matrices para los modelos Lasso y Ridge se encuentra en la (Figura *20.*

Matriz de Confusión para Modelo Lasso y Ridge – Biología.

Figura 21. Curvas ROC para Modelos Lasso y Ridge–BiologíaFigura **22**).

Figura 20.

Matriz de Confusión para Modelo Lasso y Ridge – Biología.



Nota. Matrices de confusión comparando la precisión de los modelos Lasso y Ridge en la clasificación de aprobados y reprobados.

Comparación

Al evaluar ambos modelos en su totalidad, y considerando tanto las métricas de evaluación como las variables influyentes, el modelo Lasso podría ser la elección

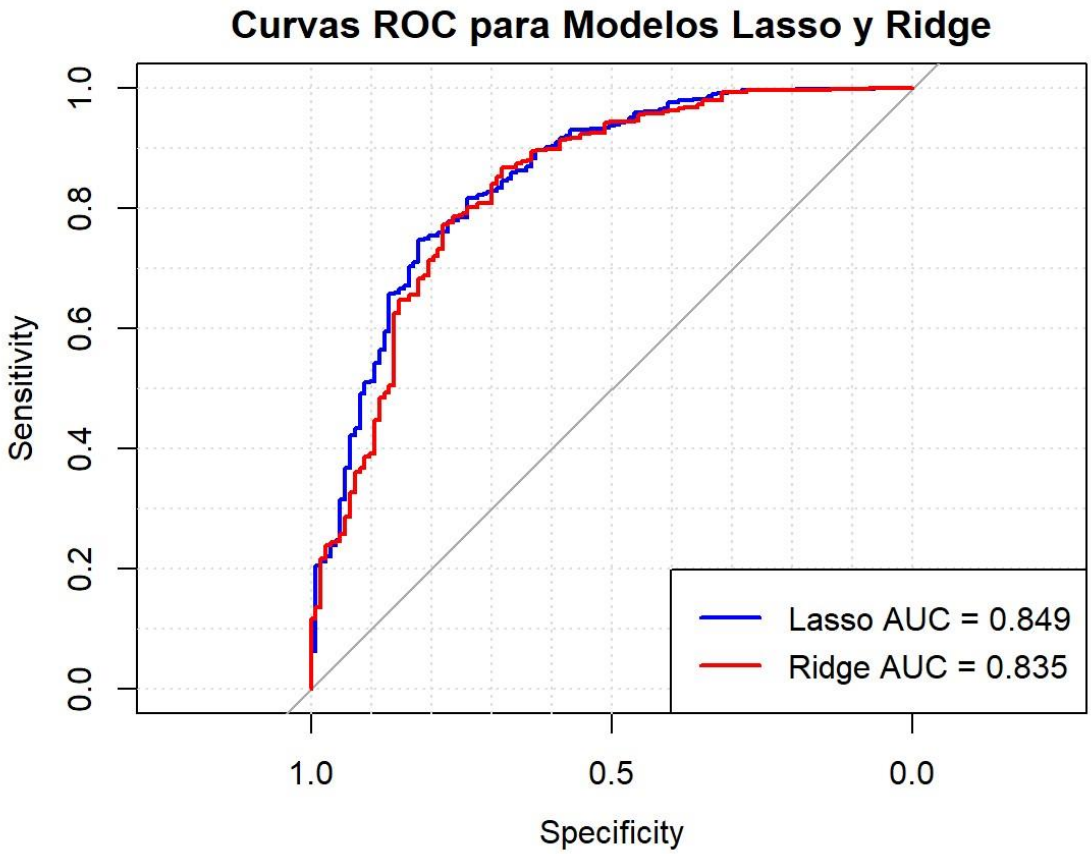
preferida para futuras predicciones. Sin embargo, es vital interpretar estos resultados en el contexto de la Universidad Yachay Tech y su población estudiantil. Las curvas ROC para ambos modelos están ilustradas en la (Figura *23.*

Curvas ROC para Modelos Lasso y Ridge–Biología

Figura 24*)*

Figura 23.

Curvas ROC para Modelos Lasso y Ridge–Biología



CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

En el proceso de explorar la interacción entre las variables sociodemográficas y el rendimiento académico de los estudiantes de primer semestre en la Universidad Yachay Tech, se ha desvelado una compleja red de conexiones e interacciones que son cruciales para el campo educativo. Las técnicas estadísticas descriptivas y analíticas empleadas en esta investigación no solo arrojaron luz sobre la significativa relación entre los factores sociodemográficos y el desempeño académico, sino que también aportaron un nuevo nivel de profundidad al corpus científico existente sobre el tema.

Con el desarrollo y refinamiento de modelos predictivos basados en regresión logística con penalizaciones Lasso y Ridge, este estudio ha evidenciado su significativa relevancia. Estos modelos, meticulosamente diseñados y validados, han demostrado ser herramientas potentes para anticipar el rendimiento académico utilizando datos sociodemográficos. La potencialidad de tales modelos es vasta, desde la identificación temprana de estudiantes que podrían beneficiarse de intervenciones adicionales hasta la reestructuración de enfoques pedagógicos.

Al profundizar en la comparativa entre los modelos de Lasso y Ridge, se descubrieron matices en su eficacia y aplicación. Aunque ambos modelos tienen sus

propias fortalezas, su evaluación exhaustiva ha brindado una perspectiva sobre cuándo y cómo se podría preferir uno sobre el otro, dependiendo del contexto y de las variables en cuestión.

Uno de los logros más destacados de este estudio ha sido la identificación concreta de las variables sociodemográficas con mayor impacto en el rendimiento académico a través de la regularización L1 y L2. Estos hallazgos no solo cumplen uno de los objetivos específicos de la investigación, sino que también sientan las bases para futuras estrategias de intervención y apoyo educativo.

En esencia, este estudio ha trazado un camino, respaldado por datos y análisis rigurosos, que facilita la comprensión de cómo se puede predecir y, por ende, mejorar el rendimiento académico a partir de factores sociodemográficos en la Universidad Yachay Tech.

5.2 Recomendaciones

Dada la eficacia probada de estos modelos predictivos, es fundamental que la Universidad Yachay Tech considere su implementación inmediata. Al integrar estos modelos dentro del sistema académico, la universidad estaría en posición de identificar proactivamente a aquellos estudiantes que, según las predicciones, podrían enfrentar desafíos en su trayectoria académica. Una vez identificados, se pueden establecer mecanismos de apoyo, diseñados específicamente en torno a las variables sociodemográficas que se han revelado como influencias clave en el rendimiento académico.

Intervenir tempranamente es de suma importancia. Al proporcionar a los estudiantes herramientas, recursos y apoyos desde las primeras etapas, se maximizan sus oportunidades de éxito. Además, es esencial que la universidad no se quede estancada en sus métodos y enfoques. La investigación y el desarrollo en el campo de la modelización predictiva deben ser constantes, buscando siempre refinar y mejorar los métodos existentes y explorar nuevas variables y técnicas. Por último, pero no menos importante, establecer un canal de comunicación abierto con los estudiantes sobre sus predicciones de rendimiento puede resultar beneficioso. Informarles sobre sus áreas de fortaleza y aquellas que requieren mejora, y orientarles sobre cómo pueden optimizar su rendimiento académico, puede fortalecer su compromiso y motivación hacia la excelencia.

REFERENCIAS

Al Husaini, Y., Y Ahmad Shukor, N. S. (2023). *Factors affecting students' academic performance: A review*. *Journal of Educational Research, 12*, 284-294.

Al Sheeb, B., Abdella, G. M., Hamouda, A. M., Y Abdulwahed, M. S. (2019). Predictive modeling of first-year student performance in engineering education using sequential penalization-based regression. *Journal of Statistics and Management Systems*, *22*(1), 31-50. https://doi.org/10.1080/09720510.2018.1509817

Bok, D. (2020). Higher Expectations: Can Colleges Teach Students What They Need to Know in the 21st Century? En *Higher Expectations*. Princeton University Press. https://doi.org/10.1515/9780691212357

Calva Yaguana, K. P. (2020). *Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado automatizado en R* [BachelorThesis, Quito, 2020.]. http://bibdigital.epn.edu.ec/handle/15000/20718

Contreas-Bravo, L. E., Nieves-Pimiento, N., González Guerrero, K. (2023). Prediction of university-level academic performance through machine learning mechanisms and supervised methods. *Ingeniería*, *28*(1). https://doi.org/10.14483/23448393.19514

Garbanzo Vargas, G. M. (2013). Factores asociados al rendimiento académico en estudiantes universitarios desde el nivel socioeconómico: Un estudio en la Universidad de Costa Rica. *Revista Electrónica Educare*, *17*(3), 57-87.

Glavas, M., Bakaric, M. B., Y Matetic, M. (2018). Applying advanced linear models in the task of predicting student success. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0744-0748. https://doi.org/10.23919/MIPRO.2018.8400138

González Medina, M. A., Rodríguez Pichardo, C. (2018). Factores sociodemográficos asociados al rendimiento en lenguaje y comunicación y en matemáticas en Nuevo León. *Innovación educativa (México, DF)*, *18*(76), 105-126.

Guambuguete Rea, C. M. (2023). *Modelo matemático para establecer los factores asociados a las calificaciones escolares utilizando regresión logística en los estudiantes del primer ciclo del Instituto Superior Tecnológico Tres de Marzo de la Provincia Bolívar* [MasterThesis, Universidad Técnica de Ambato. Facultad de Ingeniería en Sistemas, Electrónica e Industrial. Maestría en Matemática Aplicada]. https://repositorio.uta.edu.ec:8443/jspui/handle/123456789/38614

Gutiérrez-Monsalve, J. A., Garzón, J., Segura-Cardona, A. M., Gutiérrez-Monsalve, J. A., Garzón, J., Y Segura-Cardona, A. M. (2021). Factores asociados al rendimiento académico en estudiantes universitarios. *Formación universitaria*, *14*(1), 13-24. https://doi.org/10.4067/S0718-50062021000100013

Hastie, T., Tibshirani, R., Y Friedman, J. (2009). Unsupervised Learning. En T. Hastie, R. Tibshirani, Y J. Friedman (Eds.), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (pp. 485-585). Springer. https://doi.org/10.1007/978-0-387-84858-7_14

Honicke, T., Y Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review*, *17*, 63-84. https://doi.org/10.1016/j.edurev.2015.11.002

Hosmer Jr, D. W., Lemeshow, S., Y Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley Y Sons.

Kotsiantis, S., Pierrakeas, C., Y Pintelas, P. (2004). PREDICTING STUDENTS' PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES. *Applied Artificial Intelligence*, *18*(5), 411-426. https://doi.org/10.1080/08839510490442058

Krieger, N. (2011). *Epidemiology and the People's Health*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195383874.001.0001

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., Y Hayek, J. C. (2006). *What Matters to Student Success: A Review of the Literature* (Vol. 8). National Postsecondary Education Cooperative.

Loja Rodas, C. G. (2019). *Aplicación de técnicas de minería de datos en el contexto del rendimiento académico en la Universidad de Cuenca* [BachelorThesis]. http://dspace.ucuenca.edu.ec/handle/123456789/33486

Ma, C., Yao, B., Ge, F., Pan, Y., Y Guo, Y. (2017). Improving Prediction of Student Performance based on Multiple Feature Selection Approaches. *Proceedings of the 2017 1st International Conference on E-Education, E-Business and E-Technology*, 36-41. https://doi.org/10.1145/3141151.3141160

Marbouti, F., Diefes-Dux, H. A., Y Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers Y Education*, *103*, 1-15. https://doi.org/10.1016/j.compedu.2016.09.005

Nations, U. (2017). *Principles and Recommendations for Population and Housing Censuses, Revision 3*. United Nations. https://doi.org/10.18356/bb3ea73e-en

Pascarella, E. T., Y Terenzini, P. T. (2005). How College Affects Students: A Third Decade of Research. Volume 2. En *Jossey-Bass, An Imprint of Wiley*. Jossey-Bass, An Imprint of Wiley.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., Y Gehlbach, H. (2016). Forecasting student achievement in MOOCs with natural language processing. *Proceedings of the Sixth International Conference on Learning Analytics Y Knowledge*, 383-387. https://doi.org/10.1145/2883851.2883932

Rodríguez López, A., Martínez Montaño, M. del L. C., Vázquez Montiel, S., Cortés Riverol, J. G. R., Rosales de Gante, S., Y Arévalo Ramírez, M. del C. (2018). Factores sociodemográficos asociados al rendimiento académico en estudiantes de la licenciatura en Médico Cirujano-Partero. *Educación Médica Superior*, *32*(3), 68-71.

Schneider, M., Y Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565-600. https://doi.org/10.1037/bul0000098

SECRETARÍA DE EDUCACIÓN SUPERIOR, CIENCIA, TECNOLOGÍA E INNOVACIÓN. (2021). *PLAN ESTRATÉGICO INSTITUCIONAL 2021—2025*. https://www.educacionsuperior.gob.ec/

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*, 417-453. https://doi.org/10.3102/00346543075003417

Thiele, T., Singleton, A., Pope, D., Y Stanistreet, D. (2016). Predicting students' academic performance based on school and socio-demographic

characteristics. *Studies in Higher Education*, *41*(8), 1424-1446. https://doi.org/10.1080/03075079.2014.974528

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tinto, V. (2012). *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press.

Villarruel-Meythaler, R. E., Tapia-Morales, K. I., Y Cárdenas-García, J. K. (2020). Determinantes del rendimiento académico de la educación media en Ecuador. *Revista Economía y Política*, *32*. https://www.redalyc.org/journal/5711/571163421008/html/

Wang, R., Harari, G., Hao, P., Zhou, X., Y Campbell, A. (2015). *SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students* (p. 306). https://doi.org/10.1145/2750858.2804251

Yukselturk, E., Y Top, E. (2013). Exploring the link among entry characteristics, participation behaviors and course outcomes of online learners: An examination of learner profile using cluster analysis. *British Journal of Educational Technology*, *44*. https://doi.org/10.1111/j.1467-8535.2012.01339.x

Zou, H., Y Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x