



UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE POSGRADO

MAESTRÍA EN CIENCIAS DE DATOS

Proyecto de investigación previo a la
obtención del Grado Académico de Magíster
en Ciencia de Datos

TEMA

Detección de patrones académicos en curso de nivelación con deserción en
la Universidad Técnica Estatal de Quevedo basado en los factores
socioeconómicos

AUTOR

ING. ALMEIDA MURILLO JEAN CARLOS

DIRECTOR

ING. EDUARDO SAMANIEGO MENA, M.Sc

QUEVEDO - ECUADOR

2023 - 2024

CERTIFICACIÓN

Ing. Eduardo Samaniego Mena, M.Sc. en calidad de director del proyecto de investigación, previo a la obtención del Grado Académico de Magíster en Ciencia de Datos.

CERTIFICA:

Que el Ing. JEAN CARLOS ALMEIDA MURILLO, autor del proyecto de investigación titulado “DETECCIÓN DE PATRONES ACADÉMICOS EN CURSO DE NIVELACIÓN CON DESERCIÓN EN LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO BASADO EN LOS FACTORES SOCIOECONÓMICOS”; ha sido revisada en todos sus componentes, la misma que está apta para la presentación ante el tribunal respectivo.

Quevedo, 28 de mayo del 2024



Ing. Eduardo Samaniego Mena, Msc.

DIRECTOR

AUTORÍA

Yo, Ing. Almeida Murillo Jean Carlos declaro que el presente proyecto de investigación y desarrollo titulado **"DETECCIÓN DE PATRONES ACADÉMICOS EN CURSO DE NIVELACIÓN CON DESERCIÓN EN LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO BASADO EN LOS FACTORES SOCIOECONÓMICOS"**, es de exclusiva autoría y extendiendo los derechos a la Universidad Técnica Estatal de Quevedo, hacer uso del documento como material académico y de consulta.

A handwritten signature in blue ink, appearing to read 'Almeida', is written over a horizontal line.

ING. JEAN CARLOS ALMEIDA MURILLO

AUTOR

DEDICATORIA

A Dios, mi guía y sustento, le dedico este humilde logro. En cada paso incierto, encontré fortaleza en su amor incondicional y en la fe que ha iluminado mi camino. Agradezco su constante gracia que me ha sostenido durante los desafíos y sus bendiciones que han hecho posible este momento. En mi búsqueda de conocimiento, reconozco que su divina sabiduría ha sido la brújula que ha guiado mis decisiones y ha inspirado mi perseverancia. Este logro es un testimonio de su gracia abundante y a él le entrego el fruto de mis esfuerzos, reconociendo que todo es posible con su infinito amor.

Quiero expresar mi más sincero agradecimiento a mis padres, cuyo amor incondicional y apoyo constante han sido la fuerza impulsora detrás de este logro. Su sacrificio y dedicación han sido mi inspiración a lo largo de este viaje académico. A ustedes les debo no solo mi educación, sino también los valores fundamentales que me han guiado en cada paso de esta travesía.

Finalmente, dedico este trabajo a mis amigos y seres queridos, quienes han sido mi red de apoyo emocional. Su ánimo, comprensión y alegría compartida han aliviado los momentos difíciles y han hecho que cada logro sea más significativo. Este logro no solo es mío, sino también de todos aquellos que han formado parte de mi vida y han contribuido a mi crecimiento personal y académico. A todos ustedes, mi más profundo agradecimiento.

AGRADECIMIENTO

Siempre he creído que Dios revela su presencia de diversas maneras. A lo largo de la realización de esta tesis, he sentido la mano divina prodigando abundantes bendiciones en mi vida. Entre ellas, destaco la maravilla de compartir un mismo cielo que nos cobija a todos, y reconocer que somos instrumentos de la voluntad divina en la tierra cuando trabajamos con dedicación, entrega y pasión desde nuestra misión específica.

Agradezco por el tiempo y el espacio que me ha sido otorgado en esta existencia, ya que en el presente encuentro la realización de mis sueños. Este es un tiempo de gracia, el momento en que puedo levantar mis manos y hacer escuchar mi voz en favor de una educación que será una oportunidad para las niñas y jóvenes de mi Institución, inspirándolas a ser constructoras de un mundo mejor, un mundo que anhela protagonistas y sueña con líderes.

A mi director de tesis, les agradezco por su orientación experta y paciencia inagotable. Sus enseñanzas han ampliado mi comprensión del mundo académico y han contribuido significativamente al desarrollo de mis habilidades. Cada desafío superado y cada conocimiento adquirido ha sido posible gracias a sus enseñanzas y aliento constante.

Adicionalmente, quiero expresar mi profundo agradecimiento a mi familia, mis principales pilares, mis padres y hermanos. Su amor, apoyo incondicional y comprensión han sido la base sólida que ha sustentado cada paso de este viaje académico. A ellos les dedico este logro, reconociendo que su influencia ha sido esencial para alcanzar este momento significativo en mi vida.

PRÓLOGO

La educación es un pilar fundamental en el desarrollo individual y social, y su acceso equitativo es esencial para la construcción de sociedades inclusivas y prósperas. Las instituciones de educación superior desempeñan un papel crucial en la preparación y formación de los futuros profesionales. Sin embargo, la deserción estudiantil constituye un desafío persistente que afecta no solo la calidad de la educación, sino también la realización personal y profesional de los individuos. En esta investigación se adentra en el análisis de los patrones académicos asociados a la deserción estudiantil en el curso de nivelación de la Universidad Técnica Estatal de Quevedo (UTEQ), centrándose en la influencia de los factores socioeconómicos.

El estudio de la deserción estudiantil en el ámbito universitario es de suma importancia, ya que no solo impacta en la trayectoria educativa individual, sino que también tiene repercusiones económicas y sociales significativas. Por ende, esta investigación se enmarca en el compromiso de contribuir al diseño de estrategias efectivas para mitigar la deserción estudiantil y promover la equidad educativa.

La UTEQ constituye el escenario ideal para esta investigación, dado su compromiso con la excelencia académica y su enfoque en la inclusión de estudiantes de diversos contextos socioeconómicos. Al profundizar en la comprensión de cómo estos factores influyen en la deserción estudiantil en el curso de nivelación.



ING. ALMEIDA MURILLO WILMER FABRICIO

RESUMEN

Actualmente las universidades de Ecuador enfrentan grandes desafíos debido a la deserción estudiantil, por lo cual es apremiante tomar medidas para identificar los factores que inciden en este problema. Para lograr los objetivos, se realizó un análisis descriptivo y exploratorio de un conjunto de datos obtenido del departamento de Sistema de Gestión Académico de la Universidad Técnica Estatal de Quevedo. A través este proyecto, se propone abordar el problema mediante la implementación de algoritmos de Minería de Datos con el objetivo de descubrir patrones que ocasionan la deserción estudiantil. Se utilizó la metodología de descubrimiento de conocimiento en base de datos (KDD), la cual consiste en fases de selección, preprocesamiento, transformación, Minería de Datos y evaluación de los datos.

Para la preparación y depuración del conjunto de datos se usó RStudio y Weka para aplicar los algoritmos de J48, DecisiónStump, RandomTree, RandomForest, HoeffdingTree, LMT y RepTree. Para elegir el algoritmo óptimo para el estudio se evaluaron cuantitativamente cada uno de ellos mediante la precisión. En los resultados obtenidos, se destacó que el algoritmo de Random Tree tuvo mejor resultado en los análisis, seguido del del Random Forest, este algoritmo permitió identificar los patrones asociados directamente con la deserción estudiantil.

Palabras claves: Deserción estudiantil, KDD, Weka, RStudio

ABSTRACT

Currently, universities in Ecuador face great challenges due to student dropout, so it is urgent to take measures to identify the factors that affect this problem. To achieve the objectives, a descriptive and exploratory analysis was carried out on a set of data obtained from the Academic Management System department of the State Technical University of Quevedo. Through this project it is proposed to address the problem by implementing Data Mining algorithms with the aim of discovering patterns that cause student dropout. The knowledge discovery in database (KDD) methodology was used, which consists of selection, preprocessing, transformation, data mining and data evaluation phases.

For the preparation and purification of the data set, RStudio and Weka are used to apply the J48, DecisionStump, RandomTree, RandomForest, HoeffdingTree, LMT and RepTree algorithms. To choose the optimal algorithm for the study, each of them will be quantitatively evaluated by precision. In the results obtained, it was highlighted that the Random Tree algorithm had better results in the analyses, followed by the Random Forest, this algorithm allowed us to identify patterns directly associated with student dropout.

Keywords: Student Dropout, KDD Methodology, Weka, RStudio

ÍNDICE

CERTIFICACIÓN	¡Error! Marcador no definido.
DEDICATORIA	iv
AGRADECIMIENTO	v
RESUMEN	vii
ABSTRACT.....	viii
INTRODUCCIÓN	xvi
CAPÍTULO I	1
MARCO CONTEXTUAL DE LA INVESTIGACIÓN	1
1.1. UBICACIÓN Y CONTEXTUALIZACIÓN DE LA PROBLEMÁTICA	2
1.2. SITUACIÓN ACTUAL DE LA PROBLEMÁTICA	3
1.3. PROBLEMA DE INVESTIGACIÓN	4
1.3.1.Problema General	4
1.3.2.Problemas Derivados.....	4
1.4. DELIMITACIÓN DEL PROBLEMA.....	5
1.5. OBJETIVOS	6
1.5.1.Objetivo General	6
1.5.2.Objetivos Específicos	6
1.6. JUSTIFICACIÓN	7
CAPÍTULO II	9
MARCO TEÓRICO DE LA INVESTIGACIÓN	9
2.1. FUNDAMENTACIÓN CONCEPTUAL	10
2.1.1.Deserción.....	10

2.1.2. Factores que inciden en la deserción estudiantil	10
2.1.3. Big Data.....	12
2.1.4. Minería de Datos	12
2.1.5. Aprendizaje supervisado	12
2.1.6. Árboles de decisión	13
2.1.7. Agrupación	14
2.1.8. Bosque aleatorio	14
2.1.9. J48	15
2.1.10. Decisión Stump.....	15
2.1.11. Random Tree	15
2.1.12. Rep Tree	16
2.1.13. LMT	16
2.1.14. Reglas de asociación.....	17
2.1.15. Weka.....	17
2.1.16. RStudio	17
2.1.17. Metodología KDD	18
2.2. FUNDAMENTACIÓN TEÓRICA	19
2.3. FUNDAMENTACIÓN LEGAL.....	22
CAPÍTULO III.....	24
METODOLOGÍA DE LA INVESTIGACIÓN	24
3.1. TIPO DE INVESTIGACIÓN	25
3.1.1. Investigación correlacional	25
3.1.2. Investigación descriptiva.....	25
3.2. MÉTODOS Y TÉCNICAS A USAR EN LA INVESTIGACIÓN	25

3.2.1.Método deductivo.....	25
3.3. CONSTRUCCIÓN METODOLÓGICA DEL OBJETO DE INVESTIGACIÓN	25
3.3.1.Técnicas de investigación	25
3.3.2.Instrumentos de Investigación	26
3.3.2.1.Google colab.....	26
3.3.2.2.WEKA	27
3.3.2.3.RStudio.....	27
3.3. ELABORACIÓN DEL MARCO TEÓRICO	27
3.3.1.Revisión de la literatura.....	27
3.3.2.Obtención de la literatura	28
3.3.3.Extracción y recopilación de la información de interés	28
3.4. RECOLECCIÓN DE LA INFORMACIÓN.....	28
3.4.1.Selección de fuente primaria	29
3.4.2.Adquisición del conjunto de datos	29
3.4.3.Evaluación preliminar de la informació	29
3.5. PROCESAMIENTO Y ANÁLISIS.....	30
3.5.1.Selección	30
3.5.2.Preprocesamiento	39
3.5.3.Transformación	40
3.5.4.Minería de Datos	45
3.5.5.Evaluación.....	45
CAPÍTULO IV	47
RESULTADOS Y DISCUSION	47

4.1. VARIABLES SOCIOECONÓMICAS Y ACADÉMICAS CON MAYORES INCIDENCIAS EN LA DESERCIÓN ESTUDIANTEL EN LOS CURSOS DE NIVELACIÓN	48
4.1.1. Selección de variables usando el método de correlación al conjunto de datos completos considerando variables Académicas y Socioeconómicas.	48
4.1.2. Selección de variables al conjunto de datos académicos	51
4.1.3. Selección de variables al conjunto de datos socioeconómicas.....	52
4.2. TÉCNICAS PARA EL ANÁLISIS DE PATRONES QUE CAUSAN LA DESERCIÓN DE LOS ESTUDIANTES DE NIVELACIÓN CONSIDERANDO FACTORES ACADÉMICOS Y SOCIOECONÓMICOS.....	54
4.2.1. Resultado del primer análisis con selección del conjunto de datos socioeconómicos y académicos.....	55
4.2.1.1. <i>Discusión</i>	58
4.2.2. Resultado del segundo análisis con selección del conjunto de datos.....	58
socioeconómicos	58
4.2.2.1. <i>Discusión</i>	61
4.2.3. Resultado del tercer análisis con selección del conjunto de datos académicos.....	62
4.2.3.1. <i>Discusión</i>	64
4.2.4. Factores que influyen en la deserción estudiantil	65
4.3. REGLAS DE ASOCIACIÓN A LOS PERFILES SOCIOECONÓMICOS Y ACADÉMICOS SIMILARES PARA ENCONTRAR PATRONES QUE AYUDEN O CONTRIBUYAN EN LA PREVENCIÓN DE LA DESERCIÓN ESTUDIANTEL.....	65
4.3.1. <i>Discusión</i>	68

CAPÍTULO V	70
CONCLUSIONES Y RECOMENDACIONES.....	70
5.1. CONCLUSIONES	71
5.2. RECOMENDACIONES.....	72
REFERENCIAS BIBLIOGRAFICAS	74
ANEXOS	80

INDICE DE TABLAS

Tabla 1: Conjunto de datos completo	31
Tabla 2: Conjunto de datos actualizado	35
Tabla 3: Conjunto de datos completo	40
Tabla 4: Rango de notas.....	45
Tabla 5: Variables obtenidas de la correlación.....	49
Tabla 6: Variables obtenidas de la correlación.....	52
Tabla 7; Variables obtenidas de la correlación.....	53
Tabla 8: Tabla comparativa primer análisis.....	55
Tabla 9: Tabla comparativa del segundo análisis	59
Tabla 10: Tabla comparativa del tercer análisis	62
Tabla 11: Factores que causan la deserción estudiantil	65

INDICE DE ILUSTRACIONES

Ilustración 1: Metodología KDD	30
Ilustración 2: Matriz de correlación de variables Académicas y Socioeconómicas	49
Ilustración 3: Matriz de correlación de datos académicos	51
Ilustración 4: Matriz de correlación de datos socioeconómicas	53
Ilustración 5: Valores de precisión	56
Ilustración 6: Resultados de valores de kappa	57
Ilustración 7: Resultados de errores.....	57
Ilustración 8: Resultados de valores de precisión.....	59
Ilustración 9: Resultados de valores kappa.....	60

Ilustración 10: Resultados de errores.....	61
Ilustración 11: Resultados de precisión	63
Ilustración 12: Resultados de kappa	63
Ilustración 13: Resultados de errores.....	64
Ilustración 14: Nodos del conjunto de datos académicos y socioeconómico	66

INTRODUCCIÓN

La deserción estudiantil es un problema alarmante en el ámbito de la educación superior, que afecta negativamente el desarrollo académico y socioeconómico de los estudiantes. En Ecuador, donde se promueve el acceso equitativo a la educación superior como un derecho humano, la deserción representa un obstáculo significativo para garantizar la igualdad de oportunidades. En los últimos años, a pesar de los esfuerzos por fomentar el ingreso masivo a la educación superior en todos los estratos sociales, este problema persiste (Gobierno de la República del Ecuador, 2018).

La deserción estudiantil representa una fuente significativa de desilusión para los jóvenes que se embarcan en el sistema de educación superior, ya que, por diversas razones, muchos no logran culminar el curso de nivelación. Esta situación no solo impacta negativamente en su autoestima y perspectivas de futuro, sino que también contribuye al aumento de la brecha educativa y social. Abordar la deserción estudiantil no solo implica garantizar el acceso inicial a la educación superior, sino también implementar estrategias efectivas de retención y apoyo que permitan a los estudiantes superar los desafíos que enfrentan a lo largo de su trayectoria académica.

La deserción estudiantil tiene relevancia al observar que la mayoría de los estudiantes provienen principalmente de estratos socioeconómicos medios y bajos, además de pertenecer a zonas rurales, lo cual les coloca en una situación de desventaja en comparación con el resto de la población estudiantil. Este contexto tiene repercusiones significativas en la deserción universitaria, que se percibe como un problema social con efectos a nivel personal, familiar e institucional (Cárdenas Matute et al., 2023).

La problemática de la deserción estudiantil se presenta como un desafío significativo en el ámbito educativo superior, ya sea en instituciones técnicas, tecnológicas o universitarias. Este fenómeno no solo impacta de manera adversa en los aspectos políticos, económicos, sociales y culturales del desarrollo nacional, sino que también genera consecuencias perjudiciales para los procesos educativos en dichos niveles (Armas, 2020).

Ante este escenario, este proyecto busca desarrollar un modelo que permita detectar los patrones que causan la deserción estudiantil en los cursos de nivelación, centrándose en el caso de la Universidad Técnica Estatal de Quevedo (UTEQ), considerando los factores socioeconómicos y académicos considerando los algoritmos de clasificación que ofrece la herramienta weka, teniendo el algoritmo de clasificación J48 (C4.5) siendo versátil, maneja datos numéricos y categóricos eficientemente, es fácil de interpretar y adecuado para conjuntos de datos de cualquier tamaño, incluyendo datos ruidosos y faltantes. DecisionStump es rápido y útil para conjuntos de datos con muchas características o desbalanceados. RandomTree y RandomForest ofrecen predicciones precisas y evitan el sobreajuste en conjuntos de datos grandes. HoeffdingTree se adapta bien a datos en flujo continuo, ideal para entornos de cambio rápido. RepTree es interpretativo y visualmente accesible, adecuado cuando se necesita un modelo explicativo.

Este documento está estructurado con los siguientes apartados:

Capítulo I- Marco contextual de la investigación, integrado por: ubicación y contextualización de la problemática, problema general y derivados, delimitación del problema, objetivos y justificación.

Capítulo II- Marco teórico de la investigación, describe: fundamentos conceptuales y teóricos relativos al estudio, y los sustentos legales.

Capítulo III- Metodología de la investigación, dentro de este apartado se especifica: tipo y métodos de investigación, población y muestra, técnicas de investigación, instrumento de investigación, elaboración del marco teórico, recolección de información y procesamiento y análisis.

Capítulo IV- Resultados, en este se detalla el análisis e interpretación de la información recolectada a través de los diferentes algoritmos utilizados.

Capítulo V- Conclusiones y recomendaciones, se detalla en que se culminó la investigación, además de dar recomendaciones a investigaciones futuras.

CAPÍTULO I

MARCO CONTEXTUAL DE LA INVESTIGACIÓN

1.1. UBICACIÓN Y CONTEXTUALIZACIÓN DE LA PROBLEMÁTICA

Un desafío al que deben enfrentarse la mayoría de las instituciones de educación superior en toda América Latina es la deserción estudiantil. A través de muchos estudios se ha identificado un número importante de estudiantes que no terminan sus estudios de pregrado, lo que perjudica a la sociedad (Mauricio et al., 2023)(Urbina-Nájera et al., 2021). Para aumentar la retención de estudiantes durante los primeros años de educación universitaria, muchos países han comenzado a implementar estrategias integrales de mejora (Navarro, 2018).

El abandono estudiantil ha adquirido una relevancia global en la investigación académica, dado su impacto negativo tanto en los estudiantes, generando sentimientos de frustración y fracaso, como en las instituciones educativas, obstaculizando la realización plena de su visión. Este problema no solo afecta a nivel individual, dejando a los estudiantes desalentados, sino que también tiene implicaciones significativas a nivel institucional, dificultando el cumplimiento de metas educativas y socavando el avance del progreso social, económico y tecnológico. La retención de los estudiantes en el sistema educativo no solo es un indicador del bienestar individual de los jóvenes, sino que también es una condición indispensable para el desarrollo de las comunidades y las naciones. Al asegurar que los estudiantes completen sus estudios, las instituciones educativas contribuyen a la formación de ciudadanos críticos, creativos y competentes, capaces de impulsar el cambio social y económico (Ponce-Sanchez et al., 2019).

Las universidades enfrentan desafíos únicos relacionados con el abandono del proceso de nivelación. Estos podrían incluir problemas con la estructura del programa de nivelación, la falta de apoyo y orientación estudiantil adecuados, así como posibles barreras económicas que impidan que ciertos estudiantes continúen sus estudios.

Las barreras económicas son una preocupación crucial, donde muchos estudiantes pueden tener dificultades para cubrir los costos asociados con la educación superior, como libros, transporte y gastos personales (Mena et al., 2021).

Una creciente demanda de acceso a la educación superior está impulsando un cambio en el sistema de educación superior del Ecuador con base en las políticas de asignación de cupos universitarios. Esto ha llevado a un aumento en la selección de cursos de nivelación disponibles como medio para preparar a los estudiantes para el ingreso a las universidades. Sin embargo, no todos los estudiantes que inician este proceso lo terminan, lo que genera preocupación sobre la equidad y el estándar educativo (Peña, 2021).

A nivel micro, es esencial comprender que cada estudiante que decide abandonar el curso de nivelación en la UTEQ tiene una historia única y motivos personales detrás de su decisión. Entre estas razones, se pueden destacar dificultades a la vida universitaria donde algunos estudiantes pueden sentirse abrumados por el cambio en la dinámica educativa, la carga de trabajo o la independencia que conlleva la vida universitaria, lo que puede llevarlos a optar por abandonar el curso de nivelación (Ferrin, 2016). Además, problemas familiares o personales también pueden influir significativamente en la decisión de un estudiante de abandonar el curso. Factores como problemas económicos en el hogar, responsabilidades familiares o desafíos personales pueden hacer que la continuidad en la educación sea difícil de mantener.

1.2. SITUACIÓN ACTUAL DE LA PROBLEMÁTICA

El presente estudio se realiza en la UTEQ de la ciudad de Quevedo provincia de Los Ríos en la con la finalidad de visualizar los patrones académicos en curso de nivelación que abandonan sus estudios.

Uno de los principales factores que inciden en la deserción es la falta de preparación académica previa por parte de los estudiantes. La presencia de deficiencias en los conocimientos previos de los estudiantes puede tener un impacto adverso en su capacidad para asimilar el contenido del curso de nivelación, lo que a su vez puede aumentar la probabilidad de abandono.

Los costos asociados con la educación superior que se genera siendo gastos relacionados con los materiales de estudio, pueden generar una barrera financiera significativa. Los estudiantes que enfrentan dificultades económicas pueden sentirse obligados a trabajar a tiempo parcial o incluso a tiempo completo para mantenerse a sí mismos o a sus familias, lo que puede interferir con su capacidad para dedicar tiempo y esfuerzo al curso de nivelación.

La motivación y el sentido de propósito académico representan un componente importante. La falta de una clara visión de sus metas académicas y profesionales puede hacer que los estudiantes pierdan el interés en el curso de nivelación y, en última instancia, abandonen. La percepción del valor de la educación superior y la comprensión de cómo se relaciona con sus objetivos personales son factores clave en la retención estudiantil.

1.3. PROBLEMA DE INVESTIGACIÓN

1.3.1. Problema General

¿Cuáles son los patrones académicos que contribuyen a las altas tasas de deserción en los cursos de nivelación de la Universidad Técnica Estatal de Quevedo?

1.3.2. Problemas Derivados

- ✓ ¿Cuáles son las variables socioeconómicas y académicas más significativas que están relacionadas con la deserción en los cursos de nivelación?

- ✓ ¿Cómo se podrá detectar los patrones que más interfieren en la deserción de los estudiantes del curso de nivelación?
- ✓ ¿Como asociar el conjunto de datos para prevenir la deserción considerando los factores socioeconómicos y académicos?

1.4. DELIMITACIÓN DEL PROBLEMA

- **Campo:** Tecnología de la información y la comunicación
- **Área:** Tecnología de la información y la comunicación
- **Línea:** Ciencia de datos
- **Lugar:** Quevedo
- **Tiempo:** Septiembre del 2023

1.5. OBJETIVOS

1.5.1. Objetivo General

Analizar patrones académicos para predecir la deserción estudiantil en los cursos de nivelación de la Universidad Técnica Estatal de Quevedo basado en los factores socioeconómicos.

1.5.2. Objetivos Específicos

- ✓ Identificar las variables socioeconómicas y académicas con mayores incidencias en la deserción estudiantil en los cursos de nivelación
- ✓ Seleccionar la técnica para el análisis de patrones que causan la deserción de los estudiantes de nivelación considerando factores académicos y socioeconómicos
- ✓ Aplicar técnicas de asociación a los perfiles socioeconómicos y académicos similares para encontrar patrones que ayuden o contribuyan en la prevención de la deserción estudiantil.

1.6. JUSTIFICACIÓN

La deserción en instituciones universitarias constituye un desafío global que causa preocupación en el ámbito educativo, tanto en universidades públicas como privadas. A pesar de su relevancia, este tema aún no ha sido completamente explorado a nivel local, lo que resalta la necesidad urgente de identificar y comprender los factores subyacentes que contribuyen a este problema. La deserción universitaria no solo afecta a los individuos directamente involucrados, sino que también tiene un impacto negativo en las familias, la economía y la sociedad en general, convirtiéndolo en un problema social de gran importancia, que requiere una atención prioritaria y estratégica.

El propósito central de este estudio es identificar los factores que ocasionan el abandono en los cursos de nivelación de la UTEQ. Al realizar este proyecto, se espera beneficiar a diversas partes interesadas. En primer lugar, los administradores y profesores de la institución podrán utilizar los resultados obtenidos para diseñar e implementar políticas y programas de apoyo específicos destinados a reducir la tasa de deserción en los cursos de nivelación. En la cual los estudiantes se verán favorecidos al recibir un respaldo adicional para superar las barreras académicas, financieras y personales que puedan estar enfrentando, lo que aumentará sus posibilidades de éxito académico y personal.

Además, esta investigación contribuirá significativamente al avance del conocimiento académico y la práctica educativa en el ámbito universitario. Los hallazgos y las estrategias desarrolladas podrían ser de utilidad no solo para la UTEQ, sino también para otras instituciones educativas que enfrentan desafíos similares en cuanto a la deserción estudiantil en cursos de nivelación.

Es esencial destacar que previamente se han llevado a cabo investigaciones en sistemas educativos de otras instituciones universitarias que guardan similitudes con el contexto presente. Estos estudios, mayormente realizados en entidades educativas similares, han subrayado la relevancia de la minería de datos y el análisis inteligente en el ámbito educativo. Dichas investigaciones han proporcionado evidencia sobre la eficacia de utilizar métodos de análisis de datos avanzados para comprender los patrones de comportamiento de los estudiantes, identificar tendencias y factores que influyen en el éxito académico

CAPÍTULO II

MARCO TEÓRICO DE LA INVESTIGACIÓN

2.1. FUNDAMENTACIÓN CONCEPTUAL

2.1.1. Deserción

La UNESCO define la deserción estudiantil “como el porcentaje de estudiantes que habiendo estado matriculado en un año, dejan de estudiar y no vuelven a matricularse en el siguiente semestre (UNESCO, 2021).

El nivel de motivación que posee un estudiante juega un papel crucial en su decisión de permanecer o abandonarla (Rascón-Larios, 2022), considerando que el mismo autor menciona que si la motivación es alta, es más probable que los estudiantes continúen su educación, mientras que, si es baja, aumenta la probabilidad de que abandonen sus estudios.

2.1.2. Factores que inciden en la deserción estudiantil

Diversas circunstancias y situaciones que pueden llevar a un estudiante a abandonar sus estudios universitarios donde enfrentan elementos que pueden impulsarlos a completar su carrera o aumentar las posibilidades de fracaso y, como resultado, la interrupción de sus estudios en la institución (Albarrán Peña, 2019). Además, el mismo autor indica que los factores que influyen en la permanencia de un estudiante en la universidad pueden dividirse en dos categorías: los internos, que están relacionados principalmente con la situación económica del estudiante, y los externos, que tienen que ver con la motivación y la elección vocacional de las carreras.

Dentro de este contexto, se representan los siguientes factores especificado por el autor (Albarrán Peña, 2019):

Factores individuales y motivacionales: Es cuando un joven o adulto decide emprender una carrera universitaria, lo hace con la esperanza de asegurar oportunidades laborales al obtener su título profesional de manera inmediata. No obstante, se suele observar que la elección de carrera no suele coincidir con las competencias y habilidades personales del estudiante, lo que resulta en dificultades al enfrentar asignaturas o seminarios desafiantes, llevándolos a reprobaciones. Algunos estudiantes se alargan en la universidad por las barreras que encuentran al tratar de comprender los contenidos teóricos y prácticos de cursos obligatorios para su formación.

Factores socioeconómicos: La limitación de recursos económicos suele obligar a los estudiantes universitarios a buscar empleos a medio tiempo para cubrir sus gastos, lo que puede interferir con su rendimiento académico. Una opción es obtener becas para aliviar la carga financiera. Sin embargo, esta situación financiera genera desánimo y frustración, lo que disminuye el compromiso con los estudios y puede llevar a buscar empleos más lucrativos o emigrar en busca de mejores oportunidades, impactando negativamente en la finalización de sus estudios universitarios y su desarrollo profesional.

Factores institucionales: El abandono de los estudios es una conducta frecuente en el entorno universitario latinoamericano, y su comprensión requiere un enfoque que considere tanto aspectos individuales, sociales como institucionales que afectan a los estudiantes. Tratar de simplificar la causa de la deserción a un solo factor sería reduccionista, ya que este fenómeno está influido por una amplia gama de factores internos y externos. Además, no es apropiado comparar realidades socioeconómicas de diferentes países con la situación específica de una nación.

2.1.3. Big Data

Big Data se refiere a conjuntos de datos extremadamente grandes y complejos que superan la capacidad de las herramientas de procesamiento de datos tradicionales para capturar, almacenar, gestionar y analizar de manera eficiente (Arturo Amaya, Flaklin Huerta, 2020). Además, menciona el autor que estos conjuntos de datos pueden ser de diversa naturaleza, incluyendo datos estructurados, no estructurados y semiestructurados, y pueden provenir de una variedad de fuentes, como transacciones comerciales, redes sociales, dispositivos móviles, sensores, etc.

2.1.4. Minería de Datos

La gestión de bases de datos, el análisis estadístico y el modelado de inteligencia artificial sirven para la minería de datos, que está muy relacionada con la informática. Su objetivo es generar conocimiento utilizando bases de datos (Martínez Moncaleano & Palencia Fajardo, 2021). Además, (Panizzi, 2019) comenta que Minería de Datos es una disciplina que se ocupa del desarrollo de métodos para extraer información útil de datos generados en entornos educativos y utilizarla para mejorar esos entornos. La información así se transforma en el insumo necesario para tomar decisiones y utilizar la Minería de Datos para identificar a los estudiantes que están en riesgo de abandonar la escuela.

2.1.5. Aprendizaje supervisado

El aprendizaje supervisado es un enfoque en el aprendizaje automático donde el algoritmo es entrenado en un conjunto de datos que contiene ejemplos de entrada junto con sus correspondientes salidas deseadas. Durante el entrenamiento, el algoritmo aprende a mapear las entradas a las salidas mediante la observación de ejemplos etiquetados.

El objetivo principal del aprendizaje supervisado es desarrollar un modelo que pueda generalizar correctamente a nuevas entradas no vistas, es decir, ser capaz de predecir la salida correcta para datos no vistos basándose en su entrenamiento previo con datos etiquetados (Sabharwal & Selman, 2020).

2.1.6. Árboles de decisión

Los árboles de decisión son un método de clasificación común, que es un algoritmo de aprendizaje automático con una lógica sencilla. El modelo de árbol de decisión se basa en un conjunto de reglas que se obtienen utilizando datos de entrenamiento, si es posible. Es sencillo de desarrollar, altamente interpretable, completamente consistente con la intuición humana y capaz de manejar cantidades masivas de datos. Para abordar los problemas de clasificación, se han introducido muchas estructuras de toma de decisiones, como la estructura iterativa de clasificación y retorno (Li et al., 2023). Un árbol de decisión es un grupo de reglas estructuradas de forma jerárquica y siguiendo estas reglas desde la raíz del árbol hasta alguna de las hojas se llega a una decisión final, es decir, se puede analizar un problema y siguiendo la estructura del árbol de manera adecuada se llega a una sola decisión. Los elementos de un árbol se conocen como nodos y cada uno representa una regla, de tal manera que siguiendo un nodo a otro nodo descendiente se llega, eventualmente, a un nodo terminal donde se realiza la predicción (Rico Páez & Gaytán Ramírez, 2022).

Una técnica de aprendizaje guiado conocida como árboles de decisión ha ganado popularidad en los últimos años debido a su simplicidad de comprensión y efectividad computacional. Estos modelos permiten tareas de clasificación y regresión, representando las decisiones utilizando una estructura de bellota en la que cada nodo interno denota prueba de una característica de entrada y cada hoja una etiqueta de clase o un valor.

A la hora de elegir un algoritmo de aprendizaje supervisado, los árboles de decisión tienen una serie de ventajas y desventajas que conviene tener en cuenta como la facilidad de interpretación y visualización, manejo de datos heterogéneos incluidos atributos numéricos y categóricos (Alaminos F A F, 2023).

2.1.7. Agrupación

El análisis de clústeres es un método que permite descubrir asociaciones y estructuras en datos que no son evidentes de inmediato pero que pueden ser útiles una vez descubiertas. Es una técnica no supervisada que se utiliza para clasificar grandes colecciones de datos en grupos relacionados y tiene aplicaciones en muchos campos (Jaramillo Feijo et al., 2020), además otras técnicas para el análisis de big data incluyen el aprendizaje automático y el reconocimiento de patrones considerando el propósito de agrupar y categorizar cosas relacionadas en grupos distintos, o dividir una colección de datos en cosas relacionadas con características similares.

2.1.8. Bosque aleatorio

Un bosque aleatorio (Random Forest) es un algoritmo de aprendizaje supervisado utilizado principalmente para problemas de clasificación y regresión. Se basa en la construcción de múltiples árboles de decisión durante el entrenamiento y la combinación de sus resultados para mejorar la precisión y la robustez del modelo (Balyan et al., 2022). Además, indica el autor Balyan que cada árbol de decisión en el bosque se entrena con una muestra aleatoria del conjunto de datos original y realiza predicciones independientes.

2.1.9. J48

Este algoritmo produce un árbol de decisión que se puede utilizar para clasificar datos en clases y también predecir la clase de datos. El algoritmo J48, también conocido como C4.5, es un algoritmo de aprendizaje automático utilizado en minería de datos para la construcción de árboles de decisión. Se basa en el concepto de selección recursiva de atributos y división de datos en subconjuntos homogéneos, con el objetivo de clasificar instancias desconocidas. J48 utiliza medidas como la ganancia de información para determinar qué atributos son más relevantes para la clasificación y cómo dividir los datos de entrenamiento en nodos de árbol. Este algoritmo es ampliamente utilizado debido a su capacidad para manejar datos categóricos y numéricos, su eficacia en la generación de modelos interpretables para la toma de decisiones (Maulana & Defriani, 2020).

2.1.10. Decisión Stump

Un Decisión Stump es un modelo de aprendizaje automático que consiste en un árbol de decisión de un solo nivel. En otras palabras, es un árbol de decisión con un único nodo interno (la raíz) que se conecta directamente a sus nodos terminales (las hojas). Este tipo de modelo realiza predicciones basadas únicamente en el valor de una característica de entrada. A menudo se les llama también reglas 1. Dependiendo del tipo de característica de entrada, existen varias variaciones posibles. Por ejemplo, para características nominales (Kawelah & Abdala, 2019).

2.1.11. Random Tree

Random Tree es un tipo de clasificador supervisado que se deriva del algoritmo de árbol de decisión, y se caracteriza por su rapidez y eficacia. Este algoritmo emplea la técnica de embolsado, que consiste en generar conjuntos aleatorios de datos para construir múltiples árboles de decisión. Posteriormente, el algoritmo lleva a cabo una poda del

árbol utilizando un método conocido como poda de error reducido con ajuste posterior (Alpan & Ilgi, 2020).

2.1.12. Rep Tree

El algoritmo de árbol de poda de error reducido (REP) es un enfoque de aprendizaje automático que se basa en los principios de calcular la ganancia de información mediante entropía y minimizar el error causado por la variación. Es considerado como un clasificador rápido que genera múltiples árboles. Utilizando información de varianza y resultados, construye un árbol de decisión para la regresión. Además, este algoritmo poda los árboles utilizando métodos de ajuste posterior y poda de errorreducido. Similar al algoritmo C4.5, también puede manejar valores faltantes o incompletos y dividir instancias idénticas en grupos (Hamoud et al., 2018).

2.1.13. LMT

Este enfoque implica la generación de árboles utilizando el modelo logístico, siguiendo una estructura de árbol de decisión convencional, pero incorporando funciones de regresión logística en los nodos terminales. Este modelo es ampliamente utilizado en la clasificación y es capaz de manejar atributos con valores numéricos, categóricos y faltantes. LMT, por sus siglas en inglés, también conocido como Tree Augmented Naive Bayes (TAN), es un algoritmo de minería de datos que combina características de los algoritmos de árbol de decisión (DT) y regresión logística (LR) (Duran et al., 2023).

2.1.14. Reglas de asociación

Las reglas de asociación son modelos condicionales que tienen la forma de “si tal cosa...entonces”. Buscan relaciones existentes dentro de un conjunto de datos, y estas relaciones, en muchas ocasiones, permiten descubrir patrones ocultos (Martínez Moncaleano & Palencia Fajardo, 2021).

2.1.15. Weka

Weka es una herramienta de software de código abierto ampliamente empleada para llevar a cabo el análisis de datos y la extracción de información valiosa mediante procesos de minería de datos. Weka proporciona una amplia gama de herramientas y algoritmos para tareas de aprendizaje automático, incluyendo clasificación, regresión, agrupamiento, asociación y selección de atributos, entre otros. La interfaz de usuario de Weka es intuitiva y permite a los usuarios cargar conjuntos de datos, aplicar algoritmos de aprendizaje automático, evaluar modelos y visualizar resultados de manera eficiente. Además, Weka es altamente extensible, lo que significa que los usuarios pueden desarrollar y agregar sus propios algoritmos o funciones personalizadas (Dou & Meng, 2023).

2.1.16. RStudio

RStudio es el principal entorno de desarrollo integrado o IDE¹ para R, que es un software libre disponible para los sistemas operativos Windows, MacOS, y Linux (RStudio Team, 2020). De manera sencilla, un IDE es un programa compuesto por cuatro herramientas: Editor de texto, Compilador, Depurador, Interfaz gráfica (Vargas & Mesa, 2021).

2.1.17. Metodología KDD

Es un proceso automático basado en los requerimientos del usuario final, en el que se combinan descubrimiento y análisis, con el objeto de extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice (Menéndez Domínguez et al., 2022).

Además, el autor menciona las siguientes fases (Menéndez Domínguez et al., 2022):

Selección: Desde la perspectiva del usuario final se identifican conocimientos importantes y prioritarios. Al elegir todos los datos o diseñar un experimento para usar una muestra representativa de estos, se crea un conjunto objetivo de datos sobre el que se realiza el proceso de descubrimiento.

Preprocesamiento o limpieza: La calidad de los datos se examina mediante operaciones fundamentales como la eliminación de datos inútiles, la elección de métodos para gestionar datos no reconocidos, nulos y duplicados y la definición de técnicas estadísticas para su sustitución.

Transformación o reducción: A partir del objetivo del proceso se buscan características para representar los datos. Los métodos de reducción o transformación de dimensiones se utilizan para reducir el número de variables relevantes bajo consideración o para identificar representaciones no lineales de los datos. Para ello se utilizan técnicas de reducción como agregaciones, compresión de datos, histogramas, segmentación y discretización.

Minería de datos: Se busca y descubren patrones, aplicando tareas de descubrimiento como clasificación, agrupamiento, patrones secuenciales y asociaciones, entre otras.

Interpretación o evaluación de datos: Se interpretan los patrones descubiertos y en ocasiones, se retorna a las anteriores etapas para posteriores iteraciones. En esta etapa se incluye la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario.

2.2. FUNDAMENTACIÓN TEÓRICA

En el proyecto titulado “Aplicación de técnicas de Minería de Datos para predecir la deserción estudiantil: un estudio de caso”, se concentra en la predicción de la deserción de estudiantes universitarios inscritos en el programa de ingeniería de sistemas de una universidad colombiana, específicamente después de 7 años de matrícula (Perez et al., 2018). Además, el autor se realizó ingeniería de características y empleo de dos enfoques para identificar los factores cruciales relacionados con la deserción. Teniendo en cuenta que el autor siguió la metodología CRISP-DM y aplicó modelos predictivos como Árboles de Decisión, Regresión Logística y Naive Bayes.

Además, el proyecto denominado “Integración de tecnología de datos para analizar la deserción universitaria” se enfocó en llevar a cabo un análisis de la deserción estudiantil en el ámbito universitario, con el objetivo central de reducir la tasa de abandono que ocurre durante el primer año de estudios teniendo el objetivo de obtención de un título académico (Viloria et al., 2019). El autor planteo la evaluación de las condiciones académicas de ingreso, que abarcan las calificaciones obtenidas en la escuela secundaria, el nivel de conocimientos previos y el puntaje obtenido en la prueba de aptitud universitaria (UAT), influyen en la deserción de estudiantes que cursan el grado de ingeniería en la Universidad de Mumbai.

En el contexto del proyecto “Predecir el abandono de los estudiantes en las clases universitarias utilizando un enfoque de aprendizaje automático conjunto de dos capas: una novedosa generalización apilada” se aborda de manera global la problemática de la deserción estudiantil en entornos universitarios (Niyogisubizo et al., 2022). El autor propuso un enfoque de aprendizaje conjunto para anticipar la deserción de estudiantes en cursos universitarios, aprovechando una variedad de algoritmos de aprendizaje automático. Dentro de esto Niyogisubizo uso los algoritmos de Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB) y Redes Neuronales Feed-forward (FNN). Y por último el autor tuvo un modelo de apilamiento con el propósito de mejorar la precisión de las predicciones.

En otro proyecto titulado “Perspectivas para predecir la deserción escolar en estudiantes universitarios con aprendizaje automático” se investiga el rendimiento de cuatro algoritmos de aprendizaje automático distintos en relación con la predicción de la deserción estudiantil en el entorno universitario (Solis et al., 2018). Además, los autores usaron algoritmos como Random Forest, Neural Networks, Support Vector Machines y regresión logística considerando que su enfoque central del estudio se centra en la capacidad predictiva de estos algoritmos en lo que concierne a la deserción estudiantil.

“Ciencia de datos orientado a la Predicción del Rendimiento Académico”, se lleva a cabo una investigación con el propósito de aplicar técnicas de ciencia de datos para predecir el rendimiento académico de estudiantes de nivel básico (Zoila Jessenia Fabara-Sarmiento,JanethPilar Diaz-Vera, 2022). Estos autores tuvieron la iniciativa de evaluar la efectividad de los métodos de enseñanza virtual en este contexto educativo. Donde ellos implementaron un diseño metodológico desarrollado con la herramienta PMI, que conste de fases secuenciales que conducen a obtener resultados concretos.

Donde se enfocaron en analizar los diversos factores que influyen en el rendimiento académico, haciendo uso de técnicas de ciencia de datos y herramientas de aprendizaje automático.

En otro proyecto (Rivera Vergaray, 2021) “Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica” se analiza la deserción universitaria donde compararon cuatro modelos predictivos (regresión logística, árboles de decisión, KNN y una red neuronal) para prever la deserción académica en la Universidad Nacional Intercultural de la Amazonía. Además, el autor empleó un conjunto de datos extraído del sistema de gestión académica de la universidad, que incluía información socioeconómica y de rendimiento académico. Para el análisis, los autores procesaron los datos utilizando técnicas de codificación mediante consultas Transac Sql, y los modelos se aplicaron utilizando el software Knime y Python en Google Colab.

Para (Castrillón-Gómez et al., 2020) en su artículo científico con el tema: “Predicción de las principales variables que conllevan al abandono estudiantil por medio de técnicas de minería de datos” realiza un estudio sobre las variables que afectan el abandono estudiantil, utilizando 25 variables independientes agrupadas en categorías como personales, económicas, sociales, familiares y académicas. Donde llevaron a cabo en una población de 410 estudiantes, utilizando análisis estadístico multivariado y el algoritmo J48 de Weka en la cual los autores obtuvieron variables importantes como la pedagogía de los profesores, sentimiento de frustración, importancia del programa, expectativas insatisfechas, motivación del programa y procrastinación como las más influyentes en el abandono estudiantil, con una efectividad superior al 80%.

De acuerdo a (Loza & Vinueza, 2021) en su artículo de “Diseño de un modelo matemático para estimar la deserción estudiantil mediante técnicas de análisis multivariado en una institución de educación superior tecnológica” Se implementó un modelo de predicción utilizando regresión logística, que logró clasificar correctamente el 83% de los datos de entrenamiento y el 79% de los datos de prueba, además se destaca el uso de la metodología KDD (Knowledge Discovery in Databases) para diseñar un modelo matemático con el objetivo de entender la deserción estudiantil. Los resultados revelaron que el 60,18% de los estudiantes eran mujeres, el 89,75% solteros, el 32,50% trabajaba, y solo el 1,8% estaba divorciado. En la cual los autores obtuvieron un promedio de renta económica de \$560.21, y la repetición de asignaturas fue la única variable relevante en cuanto a la profesión de los alumnos, destacaron que la variable de repetición mostró que los estudiantes asociaban el fracaso escolar con el deseo de abandonar la universidad. Además, la probabilidad de repetición aumentó en un 7%, indicando que los estudiantes que repetían tenían mayor probabilidad de desertar. La variable "ocupación" demostró que los estudiantes de gastronomía desertaban con más frecuencia, mientras que, para la profesión ganadera, los estudiantes presentaban un 58% y 68% menos de deserción.

2.3. FUNDAMENTACIÓN LEGAL

La legislación en Ecuador ha avanzado con el fin de resguardar y proteger la información personal de sus habitantes en un contexto cada vez más digital y conectado. En el artículo 66, punto 19, de la Constitución de la República de Ecuador (Ecuador, 2020) se reconoce y asegura a los individuos: *“El derecho a la protección de datos de carácter personal, que incluye el acceso y la decisión sobre información y datos de este carácter, así como su correspondiente protección.*

La recolección, archivo, procesamiento, distribución o difusión de estos datos o información requerirán la autorización del titular o el mandato de la ley”. Este derecho engloba la capacidad de acceder y tomar decisiones sobre la información y los datos personales y la garantía de su protección adecuada. Es relevante resaltar que cualquier acción relacionada con la recopilación, almacenamiento, tratamiento, distribución o divulgación de estos datos o información debe contar con la autorización explícita del titular de los datos o debe estar respaldada por un fundamento legal.

En el desarrollo del proyecto de detección de patrones académicos en curso de nivelación con altas deserción en la Universidad Técnica Estatal de Quevedo basado en la variable socioeconómica, Se encuentra dentro del ámbito del respeto a la intimidad y la salvaguarda de los datos personales, asegurando que cualquier tratamiento de información delicada se efectúe conforme a las leyes y los estándares de privacidad más adecuados. Adherirse a estas regulaciones es de vital importancia para garantizar la confidencialidad y la protección de los datos de los ciudadanos en un entorno tecnológico en constante cambio.

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

3.1. TIPO DE INVESTIGACIÓN

3.1.1. Investigación correlacional

Se estudió la relación entre las variables del conjunto de datos para poder identificar las variables que van a permitir analizar los factores que inciden en la deserción en el curso de nivelación en la UTEQ. Se determinó determinar si existe alguna asociación entre los factores socioeconómicos y académicos.

3.1.2. Investigación descriptiva

La investigación es descriptiva porque tiene como objetivo principal describir las características que se están estudiando. En este caso, se describieron tanto los factores socioeconómicos y académicos de los estudiantes del curso de nivelación.

3.2. MÉTODOS Y TÉCNICAS A USAR EN LA INVESTIGACIÓN

3.2.1. Método deductivo

El método deductivo se aplica en esta investigación sobre la detección de patrones académicos en un curso de nivelación con deserción en la UTEQ basado en factores socioeconómicos y académicos. A través de la recopilación y análisis de datos específicos sobre los factores socioeconómicos y académicos del estudiante.

3.3. CONSTRUCCIÓN METODOLÓGICA DEL OBJETO DE INVESTIGACIÓN

3.3.1. Técnicas de investigación

3.3.1.1. Estadística Descriptiva

La estadística descriptiva se utilizó para examinar en profundidad los datos relacionados con la deserción estudiantil en cursos de nivelación, para comprender la distribución de variables como la edad de los estudiantes y sus calificaciones previas.

Además, se elaboraron gráficos de barras para poder visualizar la distribución de estas variables y detectar posibles anomalías.

3.3.1.2. Análisis Exploratorio de Datos (EDA)

Fue esencial para profundizar en la exploración de los datos. Se realizó un análisis de correlación para identificar relaciones entre variables. Esto permitió una comprensión más completa de las interacciones entre los factores que contribuyen a la deserción estudiantil.

3.3.1.3. Minería de Datos

La Minería de Datos se erige como un componente esencial en la identificación de patrones latentes en los registros de deserción estudiantil. En este estudio, se utilizaron diversos algoritmos de minería de datos, incluyendo J48, DecisionStump, RandomTree, RandomForest, HoeffdingTree, LMT y RepTree, así como técnicas de asociación usando algoritmo a priori.

3.3.1.4. Visualización de Datos

Se crearon gráficos interactivos que muestran las tendencias de deserción a lo largo del tiempo, desglosadas por variables. Estos gráficos facilitaron la comunicación de información.

3.3.2. Instrumentos de Investigación

3.3.2.1. Google colab

Es una plataforma gratuita proporcionada por Google que permite a los usuarios escribir y ejecutar código Python directamente en el navegador. Está especialmente diseñada para la investigación y desarrollo en machine learning y ciencia de datos, proporcionando acceso a recursos de computación en la nube, como GPUs y TPUs.

Sin necesidad de configuraciones complejas

3.3.2.2. WEKA

La elección de Weka para este estudio se justifica por su robustez y versatilidad en Minería de Datos y aprendizaje automático. Weka ofrece una amplia gama de algoritmos y herramientas para procesamiento y análisis de datos, facilitando tareas como identificación de patrones, clasificación y predicción de manera eficiente. Su interfaz amigable facilita la implementación de técnicas de aprendizaje automático, lo que resulta útil para el análisis de datos socioeconómicos y la segmentación de estudiantes con perfiles académicos similares.

3.3.2.3. RStudio

Para llevar a cabo la limpieza del conjunto de datos, se utilizó RStudio como herramienta principal. La versatilidad de RStudio permitió realizar tareas clave en el proceso de limpieza, como la identificación y tratamiento de valores atípicos, la imputación de datos faltantes y la verificación de la coherencia de los datos.

3.3. ELABORACIÓN DEL MARCO TEÓRICO

3.3.1. Revisión de la literatura

Se llevó a cabo una actividad de consulta y análisis exhaustivo mediante una búsqueda sistemática de artículos relevantes, los cuales servirán como sustento para el desarrollo del proyecto, se seleccionaron únicamente aquellos artículos publicados en un intervalo de tiempo específico, comprendido entre los años 2018 y 2023, lo que asegura la relevancia y actualidad de la información recopilada. Este enfoque temporal permite capturar los avances y tendencias más recientes en el campo de estudio.

3.3.2. Obtención de la literatura

El proceso de recopilación de literatura fue exhaustivo y abarcó diversas fuentes de información, incluyendo páginas web, libros, revistas académicas y artículos científicos especializados. Se realizaron búsquedas en plataformas como Scopus, IEEE Xplore, ScienceDirect, así como consultas en Google Académico. Esta amplia gama de recursos garantizó la obtención de una variedad de perspectivas y datos relevantes para el tema de investigación. La exploración de estas fuentes se llevó a cabo con el objetivo de recabar información actualizada.

3.3.3. Extracción y recopilación de la información de interés

Este proceso de investigación permitió recopilar la literatura esencial para elaborar un marco teórico completo que sustentará el estudio. Al explorar una variedad de fuentes, desde plataformas académicas especializadas hasta libros y artículos científicos relevantes.

3.4. RECOLECCIÓN DE LA INFORMACIÓN

El Sistema de Gestión Académica (SGA) de la UTEQ es donde se alojan la totalidad de los recursos académicos digitales y que estarán disponibles para los estudiantes de manera permanente. El SGA también permitió recolectar los datos para la ejecución del proyecto, considerando los factores socioeconómicos.

El SGA de la UTEQ constituye un repositorio integral de información, donde se albergan todos los recursos académicos digitales. Estos recursos están disponibles de manera permanente para los estudiantes, facilitando así su acceso y consulta. Este sistema desempeñó un papel fundamental al permitir recolectar los datos necesarios para la ejecución del proyecto, particularmente al considerar los factores socioeconómicos pertinentes.

La amplia gama de información disponible en esta plataforma contribuyó significativamente a enriquecer nuestro análisis y comprensión del contexto socioeconómico relacionado con nuestra investigación

3.4.1. Selección de fuente primaria

Los datos utilizados en este estudio fueron extraídos de la plataforma SGA de la UTEQ, la cual sirve como el repositorio principal de información relacionada con el desempeño estudiantil, registros académicos y demás datos relacionados al estudiante durante su pertenencia académica en la UTEQ. Teniendo en cuenta que los datos seleccionados fueron de dos años siendo el 2023 y 2022, considerando que en el año 2022 tuvieron 25 carreras y 11 materias, a diferencia del año 2023 donde en el primer periodo tuvieron un total de 24 carreras y 18 materias, y en el segundo periodo del mismo año, se tuvo un total de 30 carreras y 18 materias.

3.4.2. Adquisición del conjunto de datos

Para obtener los datos, se elaboró un oficio dirigido a la Vicerrectora Académica de la UTEQ (ver Anexo 2). En este oficio, se solicitó formalmente la autorización para acceder y descargar los registros académicos y demás información relevante de la plataforma de SGA. Esta autorización fue esencial para garantizar la legalidad y la confidencialidad en la obtención de los datos, además de asegurar el cumplimiento de las regulaciones de privacidad de datos y políticas institucionales. La colaboración de la Vicerrectora Académica en este proceso refuerza la validez y la ética en la utilización de los datos para la investigación.

3.4.3. Evaluación preliminar de la información

Tras obtener los datos, estos fueron evaluados para verificar las variables disponibles y los registros disponibles.

Este paso es crucial para comprender la calidad y la idoneidad de los datos antes de someterlos a procesamiento y análisis. La revisión incluyó la identificación de posibles valores atípicos, la detección de datos faltantes o incompletos, así como la evaluación de la consistencia y la coherencia de las variables. Esta evaluación preliminar garantiza que los datos estén en condiciones adecuadas para llevar a cabo un análisis confiable y significativo

3.5. PROCESAMIENTO Y ANÁLISIS

Para la elaboración del proyecto, se usó la metodología KDD (Menéndez Domínguez et al., 2022) como se muestra en la ilustración 1:

Ilustración 1: Metodología KDD



Fuente: (Menéndez Domínguez et al., 2022)

3.5.1. Selección

Se realizó un proceso de identificación y adquisición de las fuentes de datos relevantes para abordar la detección de patrones académicos y su relación con la deserción en cursos de nivelación en la UTEQ, enfocándonos en la variable socioeconómica. Esto implicó la recopilación de datos académicos que incluyan el rendimiento de los estudiantes en cursos de nivelación, así como datos socioeconómicos que proporcionen información detallada sobre la situación financiera y social de los estudiantes. También es esencial considerar factores contextuales, como fechas de inicio y otros datos relevantes que puedan influir en la deserción.

La selección cuidadosa de las variables garantiza que el proceso de descubrimiento de conocimiento se centre en los aspectos más significativos de los estudiantes y su rendimiento académico. Por último, esta fase inicia con un conjunto de datos completo, que cuenta con 99 variables y un total de 23008 datos, como se presenta en la tabla 1.

Tabla 1: Conjunto de datos completo

VARIABLES	TIPO	DESCRIPCION
Periodo	Categórica	Período académico
Facultad	Categórica	Facultad a la que pertenece el estudiante
Carrera	Categórica	Carrera del estudiante
Estudiante	Categórica	Identificación del estudiante
Asignatura	Categórica	Asignatura del estudiante
Paralelo	Categórica	Grupo o paralelo de la asignatura
Profesor	Categórica	Identificación del profesor
Nivel	Categórica	Nivel académico del estudiante
Matriculas	Numérica	Número de matrículas
Corte 1	Numérica	Calificación del corte 1
Corte 2	Numérica	Calificación del corte 2
Examen final	Numérica	Calificación del examen final
Examen supletorio	Numérica	Calificación del examen supletorio
Nota 1	Numérica	Nota 1 del estudiante
Nota supletorio	Numérica	Nota del examen supletorio
Nota final	Numérica	Nota final del estudiante
Asistencia %	Numérica	Porcentaje de asistencia del estudiante
Estado	Categórica	(Aprobado,Reprobado)
Edad	Numérica	Edad del estudiante
Etnia	Categórica	(Mestizo, Afrodescendiente,etc.)

Provincia	Categórica	Provincia de origen del estudiante
Familiar parentesco F1, F2, ..., F8	Categórica	Estudiante con hasta 8 familiares
Fallecido F1, F2, ..., ¿F8?	Categórica	Indica si el familiar está fallecido (Sí/No)
¿Convive F1, F2, ..., F8?	Categórica	convive con el familiar (Sí/No)
Nivel de Titulación F1, F2, ..., F8	Categórica	Nivel de titulación del familiar
Tipo de trabajo F1, F2, ..., F8	Categórica	Tipo de trabajo del familiar
¿El estudiante es cabeza de familia?	Categórica	Indica si el estudiante es cabeza de familia (Sí/No)
¿El estudiante depende económicamente de sus padres u otras personas?	Categórica	Dependencia económica del estudiante (Sí/No)
¿Quién cubre los gastos del estudiante?	Categórica	Financiamiento de los gastos del estudiante
¿Cuál es el nivel de instrucción del jefe del Hogar?	Categórica	Nivel de instrucción del jefe del hogar
¿Cuál es la ocupación del jefe del Hogar?	Categórica	Ocupación del jefe del hogar
¿Cuál es el tipo de vivienda?	Categórica	Tipo de vivienda del estudiante
¿Su vivienda es?	Categórica	Propiedad de la vivienda (Propia, Alquilada, etc.)
¿Alguien en el hogar compra vestimenta en centros comerciales?	Categórica	Alguien en el hogar compra vestimenta en centros comerciales (Sí/No)
¿En el hogar alguien ha usado internet en los últimos 6 meses?	Categórica	Alguien en el hogar ha usado internet en los últimos 6 meses (Sí/No)

¿En el hogar alguien utiliza correo electrónico que no es del trabajo?	Categórica	Indica si alguien en el hogar utiliza correo electrónico que no es del trabajo (Sí/No)
¿En el hogar alguien está registrado en una red social?	Categórica	Indica si alguien en el hogar está registrado en una red social (Sí/No)
Exceptuando los libros de texto o manuales de estudio y lecturas de trabajo	Categórica	Indica si alguien en el hogar ha leído algún libro completo exceptuando los libros de texto o manuales de estudio y lecturas de trabajo en los últimos 3 meses
¿Alguien del hogar ha leído algún libro completo en los últimos 3 meses?	Categórica	(Sí/No)
¿Tiene el hogar servicio de teléfono convencional?	Categórica	Indica si el hogar tiene servicio de teléfono convencional (Sí/No)
¿Tiene el hogar cocina con horno?	Categórica	Indica si el hogar tiene cocina con horno (Sí/No)
Tiene el hogar refrigerador?	Categórica	Indica si el hogar tiene refrigeradora (Sí/No)
Tiene el hogar lavadora?	Categórica	Indica si el hogar tiene lavadora (Sí/No)
¿Tiene el hogar equipo de sonido?	Categórica	Indica si el hogar tiene equipo de sonido (Sí/No)
¿Cuántos TV a color tienen en el hogar?	Numérica	Número de televisores a color en el hogar
¿Cuántos vehículos de uso exclusivo tiene el hogar?	Numérica	Número de vehículos de uso exclusivo en el hogar
¿Tiene el hogar servicio de internet?	Categórica	Indica si el hogar tiene servicio de internet (Sí/No)
¿Tiene computadora de escritorio?	Categórica	Indica si el hogar tiene computadora de escritorio (Sí/No)

escritorio?		
¿Tiene computadora portátil?	Categórica	¿Tiene computadora portátil(Sí/No)
¿Cuántos celulares activados tienen en el hogar?	Numérica	Número de celulares activados en el hogar
Tareas En El Hogar	Categórica	Descripción de las tareas realizadas en el hogar
Trabajo Doméstico	Categórica	¿Trabajo doméstico en el hogar(Sí/No)
Trabajo Fuera De Casa	Categórica	Indica si se realiza trabajo fuera de casa (Sí/No)
Actividades de recreación	Categórica	¿Tiene actividades de recreación en el hogar
Cuanto tiempo promedio se dedica para hacer sus tareas	Numérica	¿Cuánto tiempo se dedica a realizar tareas en el hogar
Donde realiza las tareas	Categórica	¿Donde se realizan las tareas en el hogar
Folletos	Categórica	Indica si se utilizan folletos en el hogar (Sí/No)
Computador	Categórica	¿Tiene un computador en el hogar(Sí/No)
Enciclopedias	Categórica	¿Tiene enciclopedias en el hogar(Sí/No)
Otros	Categórica	¿Tiene otras actividades o elementos en el hogar (Sí/No)
Cyber	Categórica	Indica si se utiliza un ciber en el hogar (Sí/No)
Biblioteca	Categórica	¿Tiene biblioteca en el hogar(Sí/No)
Museo	Categórica	Indica si se utiliza un museo en el hogar (Sí/No)
Área de Recreación	Categórica	¿Tiene área de recreación en el hogar(Sí/No)

Otros	Categórica	Otras actividades o elementos en el hogar (Sí/No)
-------	------------	---

Fuente: Autor

Después de llevar a cabo un análisis exhaustivo del conjunto de datos completo, se ha implementado un proceso de selección riguroso para identificar únicamente las variables pertinentes que serán objeto de estudio. Es importante destacar que este proceso de filtrado ha sido meticuloso, focalizándose en los datos correspondientes exclusivamente a aquellos estudiantes que han sido reprobados. Esta selección estratégica no solo garantiza la relevancia y la coherencia de las variables elegidas, sino que también enfoca nuestra investigación en el segmento específico de la población estudiantil que requiere una atención particular. Por consiguiente, el conjunto de datos resultante, como se detalla en la Tabla 2.

Tabla 2: Conjunto de datos actualizado

VARIABLES	TIPO	DESCRIPCION
Periodo	Categórica	Período académico
Carrera	Categórica	Carrera del estudiante
Asignatura	Categórica	Asignatura del estudiante
Matriculas	Numérica	Número de matrículas
Corte 1	Numérica	Calificación del corte 1
Corte 2	Numérica	Calificación del corte 2
Examen final	Numérica	Calificación del examen final
Examen supletorio	Numérica	Calificación del examen supletorio
Nota 1	Numérica	Nota 1 del estudiante
Nota supletorio	Numérica	Nota del examen supletorio
Nota final	Numérica	Nota final del estudiante

Asistencia %	Numérica	Porcentaje de asistencia del estudiante
Estado	Categórica	(Aprobado,Reprobado, etc.)
Edad	Numérica	Edad del estudiante
Etnia	Categórica	(Mestizo, Afrodescendiente,etc.)
Provincia	Categórica	Provincia de origen del estudiante
Familiar parentezcoF1, F2 Y F3	Categórica	Estudiante con hasta 8familiares
Fallecido F1, F2 YF3	Categórica	Indica si el familiar está fallecido (Sí/No)
Convive F1, F2 YF3	Categórica	Convive con el familiar(Sí/No)
TitulaciónF1, F2 Y F3	Categórica	Nivel de titulación del familiar
de trabajo F1,F2 Y F3	Categórica	Tipo de trabajo del familiar
¿El estudiante es cabeza de familia?	Categórica	Indica si el estudiante es cabeza de familia (Sí/No)
¿El estudiante depende económicamente desus padres u otras personas?	Categórica	Dependencia económica del estudiante (Sí/No)
¿Quién cubre losgastos del estudiante?	Categórica	Financiamiento de los gastos delestudiante
¿Cuál es el nivel de instrucción del jefe del Hogar?	Categórica	Nivel de instrucción del jefe del hogar
¿Cuál es la ocupación del jefe del Hogar?	Categórica	Ocupación del jefe del hogar
Tipo devivienda?	Categórica	Tipo de vivienda del estudiante
¿Su vivienda es?	Categórica	Propiedad de la vivienda (Propia, Alquilada, etc.)

¿En el hogar alguien ha usado internet en los últimos 6 meses?	Categórica	Alguien en el hogar ha usado internet en los últimos 6 meses (Sí/No)
¿Tiene el hogar servicio de teléfono convencional?	Categórica	hogar tiene servicio de teléfono convencional (Sí/No)
¿Tiene el hogar cocina con horno?	Categórica	Indica si el hogar tiene cocina con horno (Sí/No)
Tiene el hogar refrigerador?	Categórica	Indica si el hogar tiene refrigeradora (Sí/No)
Tiene el hogar lavador?	Categórica	Indica si el hogar tiene lavadora (Sí/No)
¿Tiene el hogar equipo de sonido?	Categórica	Indica si el hogar tiene equipo de sonido (Sí/No)
¿Cuántos TV a color tienen en el hogar?	Numérica	Número de televisores a color en el hogar
¿Cuántos vehículos de uso exclusivo tiene el hogar?	Numérica	Número de vehículos de uso exclusivo en el hogar
¿Tiene el hogar servicio de internet?	Categórica	Indica si el hogar tiene servicio de internet (Sí/No)
¿Tiene computadora de escritorio?	Categórica	Tiene la computadora de escritorio (Sí/No)
¿Tiene computadora portátil?	Categórica	¿Tiene computadora portátil? (Sí/No)
¿Cuántos celulares activados tienen en el hogar?	Numérica	Número de celulares activados en el hogar
¿Cuanto tiempo promedio se dedica a realizar tareas en el hogar?	Numérica	Tiempo dedicado a realizar tareas en el hogar

para hacer sustareas		
Donde realiza lastareas	Categórica	Lugar donde se realizan las tareasen el hogar
Computador	Categórica	Tiene computador en el hogar(Sí/No)
Enciclopedias	Categórica	Enciclopedias en el hogar(Sí/No)
Cyber	Categórica	Indica si se utiliza un ciber en el hogar (Sí/No)
Biblioteca	Categórica	una biblioteca en el hogar(Sí/No)
Área de Recreación	Categórica	área de recreación en el hogar(Sí/No)

Fuente: Autor

3.5.2. Preprocesamiento

Se llevó a cabo la limpieza y preparación de los datos. Se identificaron y abordaron los valores faltantes, duplicados y posibles datos erróneos en los conjuntos de datos académicos y socioeconómicos. Después de obtener el conjunto de datos con las variables seleccionadas, se llevó a cabo un proceso de normalización y manejo de valores faltantes para asegurar la consistencia y calidad de los datos utilizando la herramienta RStudio. En primer lugar, se aplicó una técnica de normalización de texto mediante el uso de la función `iconv` para transformar caracteres acentuados a su forma no acentuada. Posteriormente, se realizó una imputación de valores faltantes, asignando valores de 0 a las columnas 6, 7, 8, 9 y 10 del conjunto de datos. Además, en las columnas 17 hasta la 31, se sustituyeron las celdas vacías por la etiqueta "NO RELLENO". Se procedió a asignar valores aleatorios a aquellas horas que excedían un umbral predefinido, tomando como referencia el promedio de las horas ya existentes en la variable.

En adición, se llevaron a cabo dos procesos de refinamiento en el conjunto de datos utilizando técnicas específicas para mejorar su estructura y legibilidad. En primer lugar, se aplicó una técnica de limpieza de texto para eliminar cualquier contenido encerrado entre corchetes dentro de la variable "Asignatura". Esta acción busca mejorar la claridad de los datos al eliminar información adicional que podría ser redundante o irrelevante para el análisis. En segundo lugar, se utilizó una técnica de filtrado de texto para eliminar la presencia de la palabra "nivelación" en la variable "Carrera". Esta decisión se tomó con el propósito de consolidar la representación de las distintas carreras, promoviendo la uniformidad y el control en esta variable.

Se realizaron acciones específicas para mejorar la coherencia y relevancia de la variable. "Periodo" en el conjunto de datos. Inicialmente, se eliminaron los primeros 23 caracteres, y luego se procedió a eliminar desde el cuarto hasta el decimoquinto carácter. Esta serie de operaciones busca estructurar y estandarizar la información temporal contenida en la variable "Periodo", facilitando así su interpretación y permitiendo análisis más precisos en futuras etapas del proyecto. La clarificación de la representación temporal contribuye a una mejor comprensión y contextualización de los datos, fortaleciendo la calidad del conjunto de datos en su conjunto.

Además, se aplicó un filtro de datos según el estado de "Reprobados", con el objetivo de emplear técnicas específicas de análisis para identificar patrones de deserción en la UTEQ.

3.5.3. Transformación

Tras la selección inicial de las variables, se llevó a cabo un proceso de refinamiento enfocado en simplificar las preguntas subyacentes, transformándolas en variables más compactas y representativas. Este proceso se encuentra detalladamente descrito en la tabla 3. El objetivo fundamental de esta transformación radica en optimizar la eficacia del análisis de datos.

Tabla 3: Conjunto de datos completo

Variables	Variable transformada
Periodo	Acad1
Carrera	Acad2
Asignatura	Acad3

Matriculas	Acad4
Corte 1	Acad5
Corte 2	Acad6
Examen final	Acad7
Examen supletorio	Acad8
Nota 1	Acad9
Nota supletorio	Acad10
Nota final	Acad11
Asistencia %	Acad12
Estado	Acad13
Edad	Demog1
Etnia	Demog2
Provincia	Demog3
iliar parentezcoF1	Demog4
Fallecido F1	Demog5
Convive F1	Demog6
Nivel de titulaciónF1	Demog7
Tipo de trabajo F1	Demog8
iliar parentezcoF2	Demog9
Fallecido F2	Demog10
Convive F2	Demog11
Nivel de titulaciónF2	Demog12
Tipo de trabajo F2	Demog13

Familiar parentesco F3	Demog14
Fallecido F3	Demog15
Convive F3	Demog16
Nivel de titulaciónF3	Demog17
Tipo de trabajo F3	Demog18
¿El estudiante es cabeza de familia?	SocioEcon1
¿El estudiante depende económicamente de sus padres u otras personas?	SocioEcon2
¿Quién cubre los gastos del estudiante?	SocioEcon3
¿Cuál es el nivel de instrucción del jefe del Hogar?	SocioEcon4
¿Cuál es la ocupación del jefe del Hogar?	SocioEcon5
¿Cuál es el tipo de vivienda?	SocioEcon6
¿Su vivienda es?	SocioEcon7

¿En el hogar alguien ha usado internet en los últimos 6 meses?	SocioEcon8
¿Tiene el hogar servicio de teléfono convencional?	SocioEcon9
¿Tiene el hogar cocina con horno?	SocioEcon10
¿Tiene el hogar refrigerador?	SocioEcon11
¿Tiene el hogar lavadora?	SocioEcon12
¿Tiene el hogar equipo de sonido?	SocioEcon13
¿Cuántos TV a color tienen en el <u>hogar</u> ?	SocioEcon14
¿Cuántos vehículos de uso exclusivo tiene el hogar?	SocioEcon15
¿Tiene el hogar servicio de internet?	SocioEcon16
¿Tiene computadora de escritorio?	SocioEcon17

¿Tiene computadora portátil?	SocioEcon18
¿Cuántos celulares activados tienen en el hogar?	SocioEcon19
Cuanto tiempo promedio se dedica para hacer sus tareas de realiza lastareas	SocioEcon20
Computador	SocioEcon21
Enciclopedias	SocioEcon22
Cyber	SocioEcon23
Biblioteca	SocioEcon24
Area de Recreación	SocioEcon25
	SocioEcon26

Fuente: Autor

En el proceso de elaboración de las reglas de asociación para analizar los datos académicos, se optó por transformar las variables numéricas en categorías más descriptivas y significativas, tales como "Insuficiente", "Suficiente", "Bien", "Notable" y "Sobresaliente" como se muestra en la tabla 4. Esta transformación se llevó a cabo con el objetivo de mejorar la calidad y la relevancia de los resultados obtenidos en el análisis. Al categorizar los datos de esta manera, se facilita la identificación de patrones y relaciones entre las diferentes variables, lo que permite una comprensión más profunda y una interpretación más precisa de los factores que influyen en el rendimiento académico.

Esta estrategia de preprocesamiento de datos contribuye a obtener insights más significativos y relevantes para la toma de decisiones en el ámbito educativo, al destacar las asociaciones y tendencias subyacentes entre las diferentes categorías de rendimiento estudiantil.

Tabla 4: Rango de notas

Rango	Variable
0 – 3.99	Insuficiente
4 – 4.99	Suficiente
5 – 6.99	Bien
7 – 8.99	Notable
9 – 10	Sobresaliente

Fuente: Autor

3.5.4. Minería de Datos

Para abordar la detección de patrones académicos en cursos de nivelación con altas tasas de deserción en la UTEQ, centrando el análisis en la variable socioeconómica y académicas, se emplearon algoritmos de Minería de Datos para el análisis de los resultados, se utilizaron varios algoritmos de aprendizaje automático como el J48, DecisionStump, RandomTree, RandomForest, HoeffdingTree, LMT y RepTree.

3.5.5. Evaluación

Para evaluar el mejor algoritmo, se emplearon diversas métricas fundamentales en el análisis de rendimiento de modelos de aprendizaje automático. Estas métricas incluyeron la precisión, que representa la proporción de predicciones correctas realizadas por el modelo; el coeficiente Kappa, que mide la concordancia entre las predicciones del modelo y los resultados reales.

Teniendo en cuenta la posibilidad de que las predicciones sean correctas por azar; el error absoluto medio, que indica la magnitud promedio de las desviaciones entre las predicciones del modelo y los valores reales; y el error cuadrático medio, que calcula el promedio de los errores al cuadrado entre las predicciones y los valores reales, proporcionando una medida de la precisión de las predicciones en términos de la varianza de los datos. Estas métricas se seleccionaron para obtener una comprensión completa del rendimiento de cada algoritmo y facilitar la comparación objetiva entre ellos.

CAPÍTULO IV

RESULTADOS Y DISCUSION

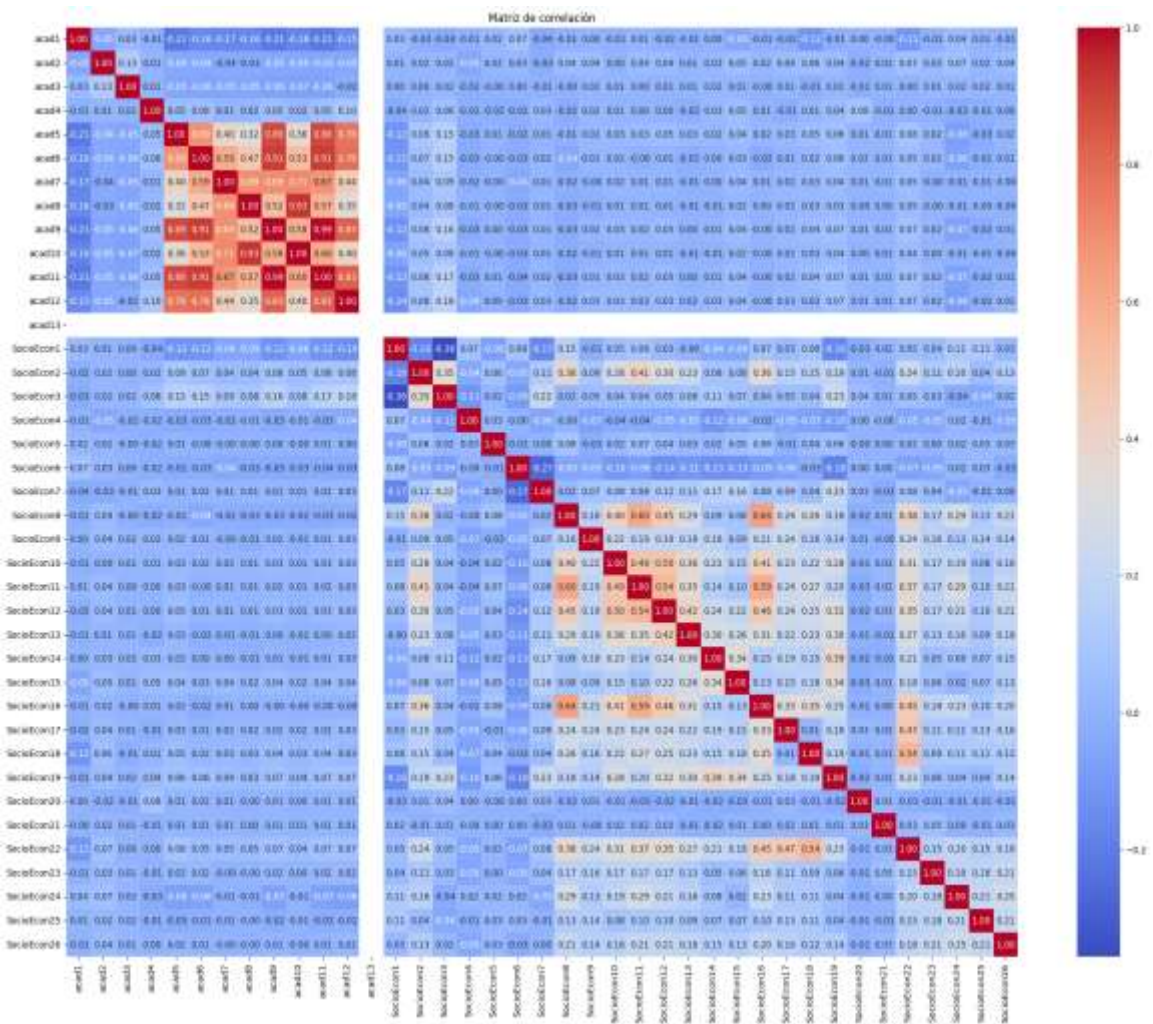
4.1. VARIABLES SOCIOECONÓMICAS Y ACADÉMICAS CON MAYORES INCIDENCIAS EN LA DESERCIÓN ESTUDIANTIL EN LOS CURSOS DE NIVELACIÓN

Se llevó a cabo un proceso de selección de variables utilizando Python como herramienta principal siendo el enfoque el análisis de correlación, una técnica estadística fundamental que permite identificar la relación entre pares de variables en un conjunto de datos. Este enfoque proporciona asociaciones entre las variables, lo que facilita la identificación de aquellas que tienen un impacto significativo en el fenómeno estudiado. Este proceso de selección guiado por correlación no solo contribuye a la eficiencia computacional, al reducir la dimensionalidad del conjunto de datos, sino que también mejora la interpretabilidad y la calidad general del modelo resultante.

4.1.1. Selección de variables usando el método de correlación al conjunto de datos completos considerando variables Académicas y Socioeconómicas.

Se realizó el primer análisis tomando en cuenta todo el conjunto de datos teniendo en cuenta las variables Académicas y Socioeconómicas, dando como resultado los valores como se muestran en la ilustración 2.

Ilustración 2: Matriz de correlación de variables Académicas y Socioeconómicas



Fuente: Autor

Una vez que se obtuvo el gráfico de correlación, la tabla 4 se muestra el resultado final del análisis de correlación con el objetivo de reducir variables teniendo en cuenta las Académicas y Socioeconómicas.

Tabla 5: Variables obtenidas de la correlación

Variables	Descripción
Acad5	Corte 1
Acad6	Corte 2

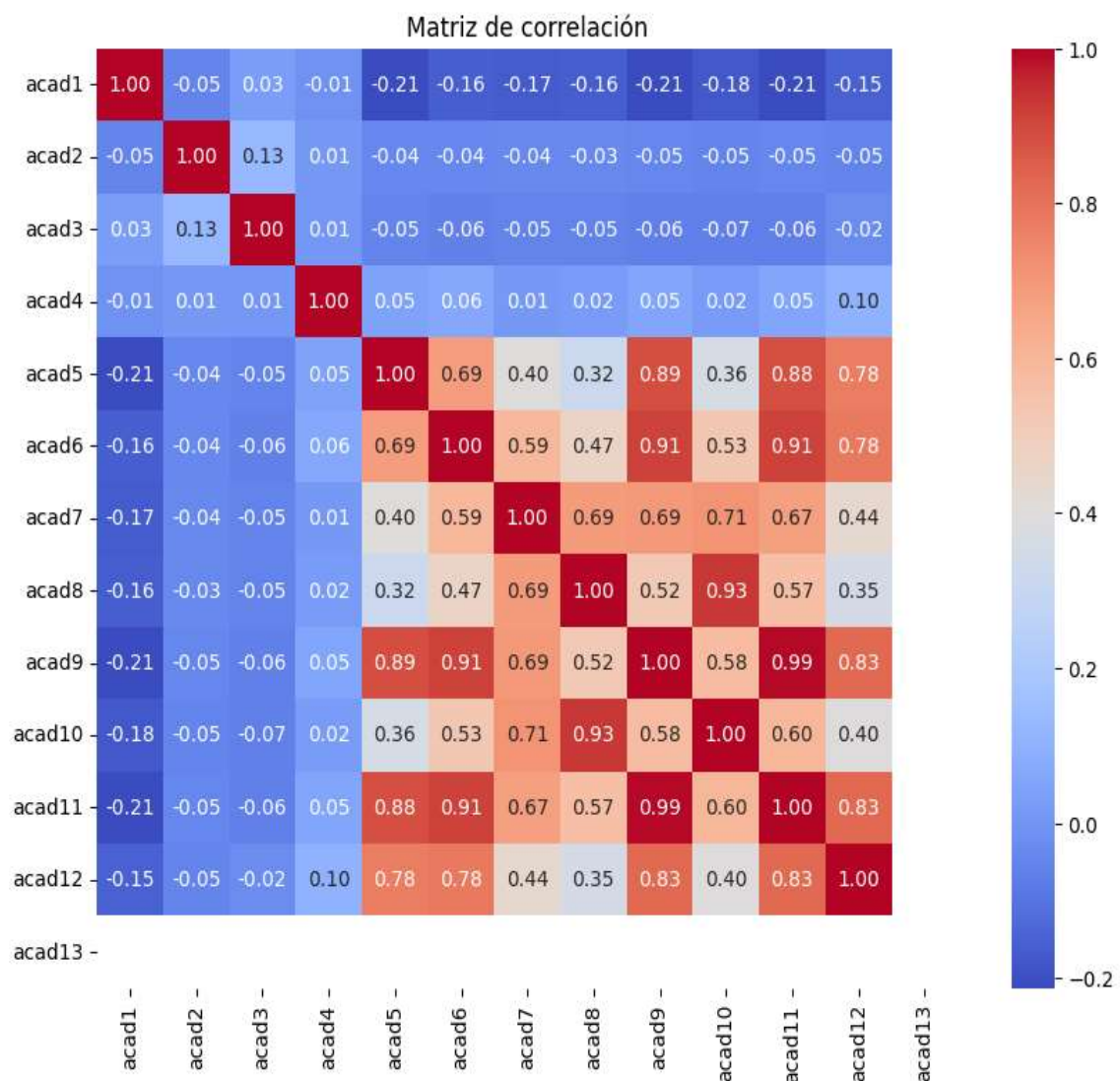
Acad7	Examen final
Acad8	Examen supletorio
Acad9	Nota 1
Acad10	Nota supletorio
Acad11	Nota final
Acad12	Asistencia %
SocioEcon2	¿El estudiante depende económicamente de sus padres u otras personas?
SocioEcon3	¿Quién cubre los gastos del estudiante?
SocioEcon7	¿Su vivienda es?
SocioEcon8	¿En el hogar alguien ha usado internet en los últimos 6 meses?
SocioEcon9	¿Tiene el hogar servicio de teléfono convencional?
SocioEcon16	¿Tiene el hogar servicio de internet?
SocioEcon17	¿Tiene computadora de escritorio?
SocioEcon18	¿Tiene computadora portátil?
SocioEcon19	¿Cuántos celulares activados tienen en el hogar?
SocioEcon20	Cuanto tiempo promedio se dedica para hacer sus tareas
SocioEcon21	Donde realiza las tareas
SocioEcon22	Computador
SocioEcon23	Enciclopedias
SocioEcon24	Cyber
SocioEcon25	Biblioteca
SocioEcon26	Area de Recreación

Fuente: Autor

4.1.2. Selección de variables al conjunto de datos académicos

Como segundo análisis de correlación, se realizó tomando en cuenta todo el conjunto de datos de las variables Académicas, dando como resultado los valores que se muestran en la ilustración 3

Ilustración 3: Matriz de correlación de datos académicos



Fuente: Autor

Una vez que se obtuvo el gráfico de correlación, la tabla 5 se muestra el resultado final del análisis de correlación con el objetivo de reducir variables teniendo en cuenta solamente las académicas.

Tabla 6: Variables obtenidas de la correlación

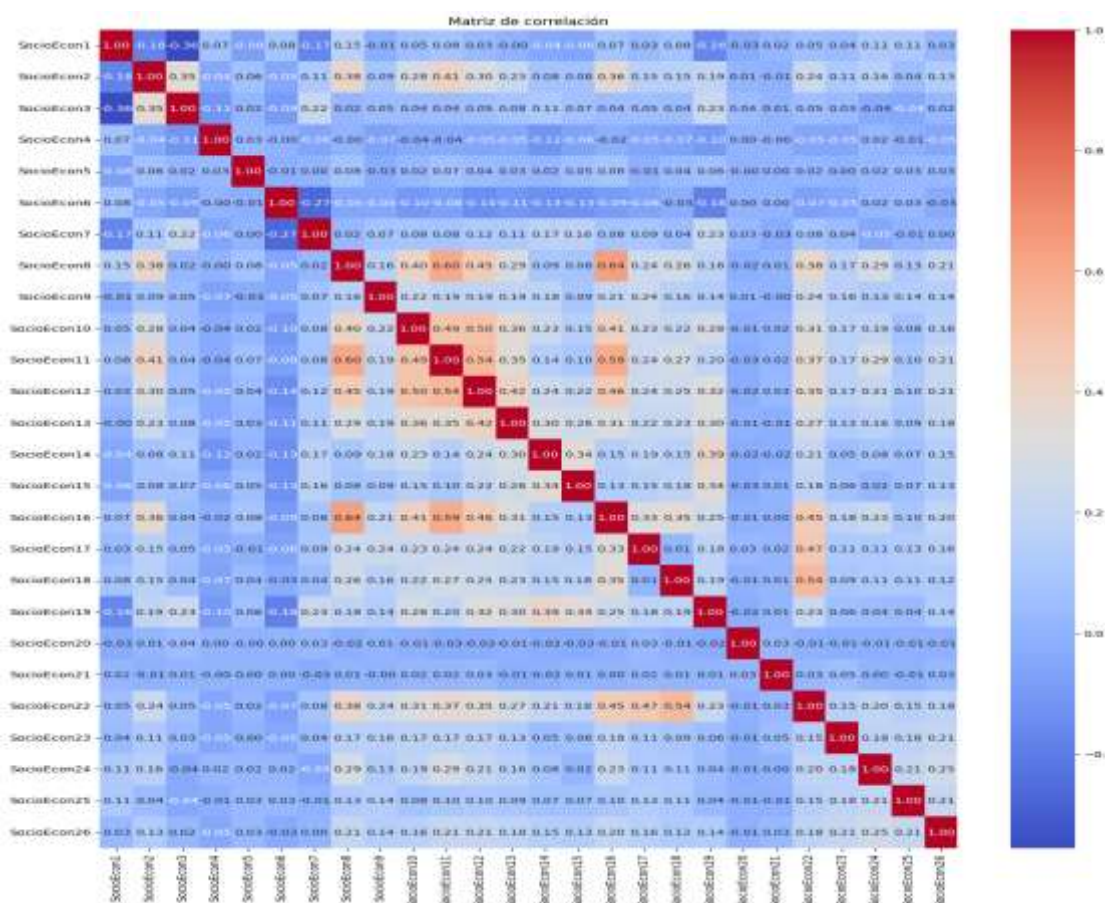
Variables	Descripción
acad5	Corte 1
Acad6	Corte 2
Acad7	Examen final
Acad8	Examen supletorio
Acad9	Nota 1
Acad10	Nota supletorio
Acad11	Nota final

Fuente: Autor

4.1.3. Selección de variables al conjunto de datos socioeconómicas

Como tercer análisis de correlación, se realizó tomando en cuenta todo el conjunto de datos de las variables socioeconómicas, dando como resultado los valores como se muestran en la ilustración 4

Ilustración 4: Matriz de correlación de datos socioeconómicas



Fuente: Autor

Una vez que se obtuvo el gráfico de correlación, la tabla 6 se muestra el resultado final del análisis de correlación con el objetivo de reducir variables teniendo en cuenta las variables Socioeconómicas.

Tabla 7; Variables obtenidas de la correlación

Variables	Descripción
SocioEcon1	¿El estudiante es cabeza de familia?
SocioEcon2	¿El estudiante depende económicamente de sus padres u otras personas?
SocioEcon8	¿En el hogar alguien ha usado internet en los últimos 6 meses?
SocioEcon9	¿Tiene el hogar servicio de teléfono convencional?

SocioEcon16	¿Tiene el hogar servicio de internet?
SocioEcon22	Computador
SocioEcon25	Biblioteca

Fuente: Autor

4.2. TÉCNICAS PARA EL ANÁLISIS DE PATRONES QUE CAUSAN LA DESERCIÓN DE LOS ESTUDIANTES DE NIVELACIÓN CONSIDERANDO FACTORES ACADÉMICOS Y SOCIOECONÓMICOS.

Para la implementación se utilizaron diferentes algoritmos de clasificación como J48, DecisiónStump, RandomTree, RandomForest, HoeffdingTree, LMT, RepTree. Estos algoritmos son utilizados en el campo del aprendizaje automático y la minería de datos. J48, una implementación del algoritmo C4.5, se emplea para la clasificación mediante la construcción de árboles de decisión basados en las características del conjunto de datos. DecisionStump es un tipo simple de árbol de decisión con una sola división, comúnmente utilizado en algoritmos de conjunto como AdaBoost. RandomTree, por otro lado, introduce aleatoriedad durante la construcción del árbol, lo que reduce el riesgo de sobreajuste. RandomForest es una técnica de conjunto que utiliza múltiples árboles de decisión, cada uno entrenado en una muestra aleatoria del conjunto de datos, y combina sus resultados para mejorar la precisión predictiva. HoeffdingTree es una alternativa diseñada para manejar flujos de datos en tiempo real, adaptándose dinámicamente a los cambios en los datos. LMT, o Logistic Model Trees, combina árboles de decisión con modelos lineales, permitiendo la interpretación y generalización de los resultados obtenidos a partir de datos complejos. Por último, RepTree es una variante de árbol de decisión que construye modelos mediante la repetición de un algoritmo de partición en subconjuntos aleatorios del conjunto de datos.

Proporcionando una opción más robusta para conjuntos de datos ruidosos o incompletos.

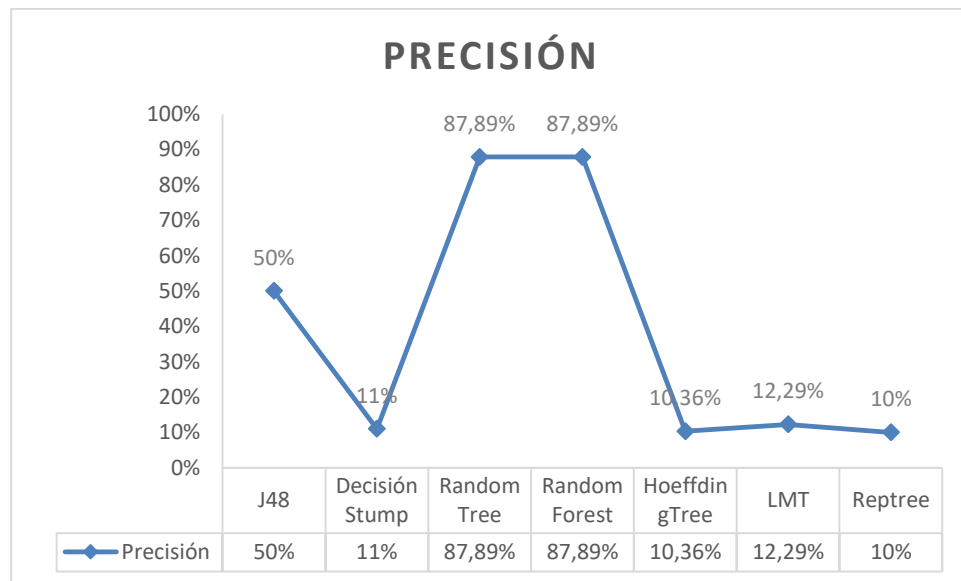
4.2.1. Resultado del primer análisis con selección del conjunto de datos socioeconómicos y académicos.

Dentro de estos algoritmos, el primer análisis teniendo en cuenta la elección de variables del conjunto de datos como se indica en la ilustración 2. Donde se obtuvieron los diferentes resultados usando los algoritmos ya mencionados, como se indica en la tabla 8.

Tabla 8: Tabla comparativa primer análisis

Algoritmo	Precisión	Kappa	Error medio	absoluto	Error cuadrático medio
J48	50%	0.49%	0.01		0.074
DecisiónStump	11%	0.05%	0.01		0.097
RandomTree	87.89%	0.87%	0.002		0.035
RandomForest	87.89%	0.87%	0.008		0.048
HoeffdingTree	10.36%	0	0.019		0.098
LMT	12.29%	0.064	0.018		0.096
Reptree	10%	0	0.019		0.098

Ilustración 5: Valores de precisión

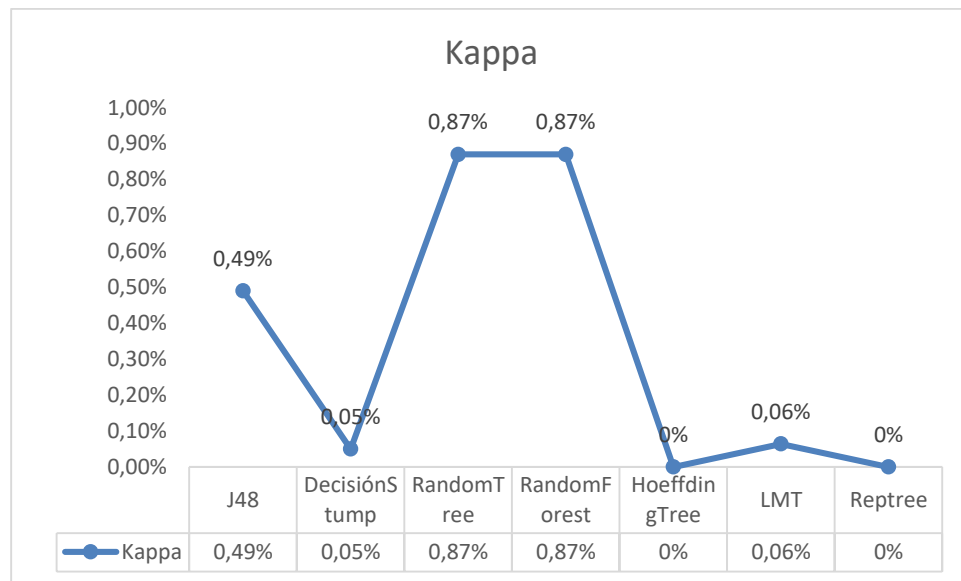


Fuente: Autor

Cada algoritmo se evaluó utilizando métricas, como la precisión, Kappa, el error absoluto medio y el error cuadrático medio. Los resultados revelaron que el algoritmo RandomForest y RandomTree destacaron como el más eficiente en términos de precisión, obteniendo un impresionante 87.89%. Esta alta precisión indica una concordancia significativa entre las predicciones del modelo y las observaciones reales. Además, RandomForest y RandomTree exhibió la estadística Kappa más alta, alcanzando un valor de 0.87, lo que sugiere una excelente capacidad de clasificación. RandomTree también presentó el error absoluto medio más bajo, con solo 0.002, lo que indica una precisión notable en sus predicciones. Además, su error cuadrático medio fue el más bajo entre todos los algoritmos evaluados, registrando un valor de 0.035.

Además, en la Ilustración 6 se muestra los valores de Kappa donde el de RandomForest y RandomTree tienen los valores altos.

Ilustración 6: Resultados de valores de kappa



Fuente: Autor

Además de obtener los valores de precisión y kappa, también se obtuvo los resultados de errores como se muestra en la Ilustración 7.

Ilustración 7: Resultados de errores



Fuente: Autor

4.2.1.1. *Discusión*

Dentro del árbol proporcionado en el anexo 3 y los grupos generados en el anexo 4. El análisis de los datos del segundo corte proporciona información valiosa para comprender los factores que podrían estar contribuyendo a la deserción estudiantil. La nota final es un indicador directo del rendimiento académico, y es que los estudiantes que no logran buenos resultados pueden enfrentar mayores riesgos de abandonar sus estudios. Sin embargo, más allá de las calificaciones, las preguntas relacionadas con la vivienda y el acceso a Internet arrojan desafíos adicionales que podrían estar afectando la retención estudiantil.

La falta de acceso a Internet o la disponibilidad limitada de dispositivos tecnológicos en el hogar pueden dificultar significativamente la participación en actividades educativas virtuales, lo que a su vez podría desmotivar a los estudiantes y aumentar la probabilidad de abandono escolar. Del mismo modo, el hecho de que los estudiantes dediquen un tiempo considerable a sus tareas escolares, pero no cuenten con un entorno adecuado para realizarlas, como una computadora, podría impactar negativamente su capacidad para completar las asignaciones de manera efectiva y, en última instancia, influir en su decisión de abandonar la escuela.

4.2.2. **Resultado del segundo análisis con selección del conjunto de datos**

socioeconómicos

Para el segundo análisis considerando los mismos algoritmos, se tomaron en cuenta solamente las variables socioeconómicas como se muestra en la ilustración 4. Además, como se muestra en la tabla 9 indica que el algoritmo RandomTree destacó como el más eficiente en términos de precisión, teniendo un valor de 86.9%. Esta alta precisión indica una concordancia significativa entre las predicciones del modelo y las observaciones

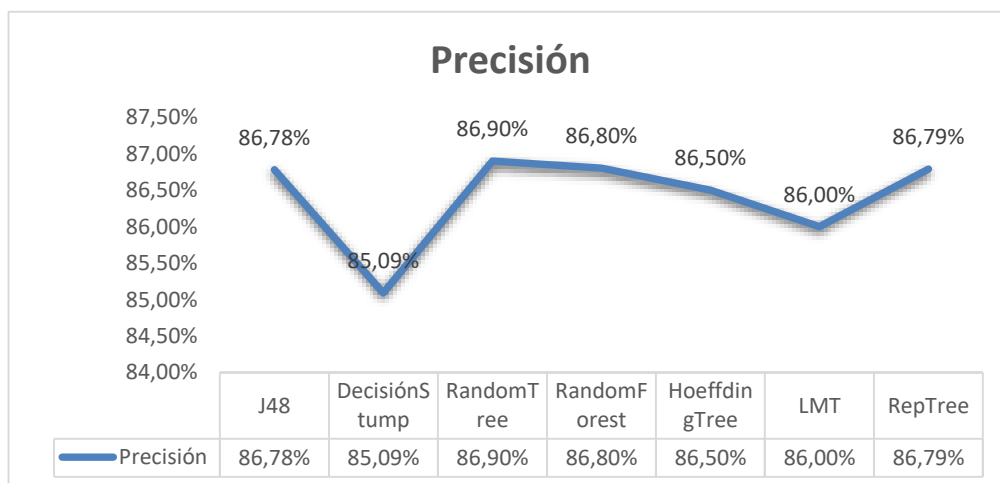
reales. Además, RandomTree exhibió un valor de Kappa de 0.67 al igual que el random forest.

Tabla 9: Tabla comparativa del segundo análisis

Algoritmo	Precisión	Kappa	Error absoluto medio	Error cuadrático medio
J48	86.78%	0.66%	0.22	0.33
DecisiónStump	85.09%	0.63%	0.24	0.32
RandomTree	86.90%	0.67%	0.20	0.31
RandomForest	86.80%	0.67%	0.20	0.31
HoeffdingTree	86.50%	0.66%	0.20	0.32
LMT	86.00%	0.67%	0.20	0.32
RepTree	86.79%	0.66	0.20	0.32

Fuente: Autor

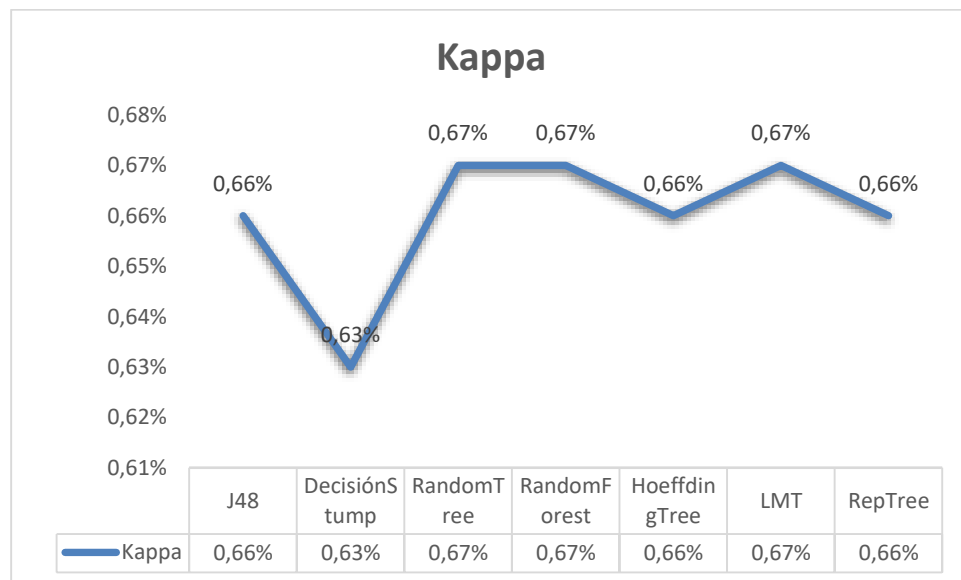
Ilustración 8: Resultados de valores de precisión



Fuente: Autor

Además, en la Ilustración 9 se muestra los valores de Kappa donde el de RandomTree y Random Forest son los que tienen un valor más alto.

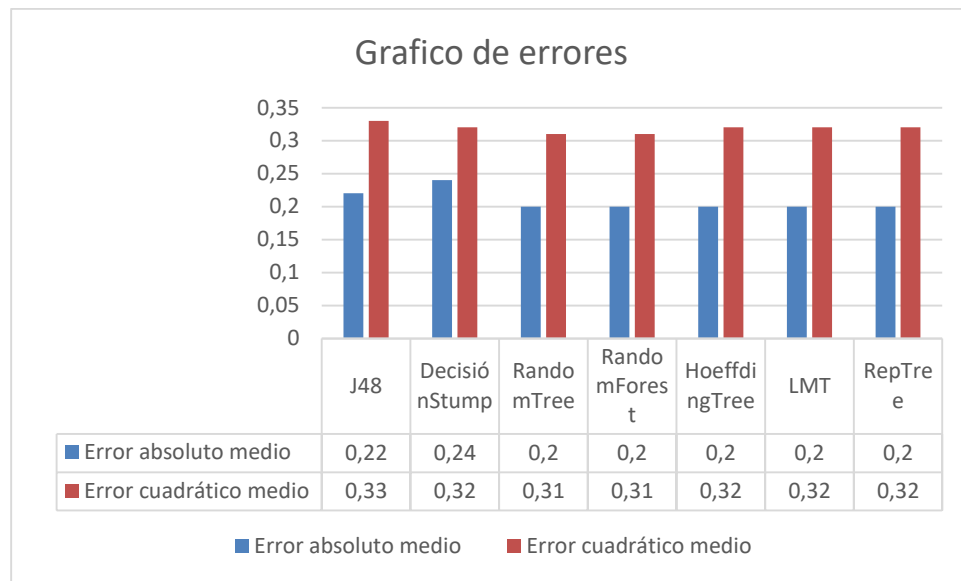
Ilustración 9: Resultados de valores kappa



Fuente: Autor

Además de obtener los valores de precisión y kappa, también se obtuvo los resultados de errores como se muestra en la Ilustración 10

Ilustración 10: Resultados de errores



Fuente: Autor

4.2.2.1. *Discusión*

Dentro del árbol proporcionado en el anexo 7 y los grupos creados en el anexo 8. Muchos estudiantes se enfrentan a desafíos derivados de su situación personal, los cuales influyen significativamente en su capacidad para completar con éxito el curso de nivelación. Uno de los obstáculos más comunes es la falta de acceso a recursos tecnológicos adecuados, como un computador o una conexión estable a internet, especialmente en el contexto actual de educación virtual. Esta limitación impide que los estudiantes participen plenamente en las clases en línea, accedan al material de estudio y completen las tareas asignadas, lo que repercute negativamente en su desempeño académico y aumenta la probabilidad de desertar.

Además, para algunos estudiantes, la responsabilidad de ser cabeza de familia agrega una carga adicional a sus compromisos diarios. La necesidad de encontrar empleo para satisfacer las necesidades financieras del hogar compite directamente con el tiempo y la energía que podrían dedicar a sus estudios. Esta situación se ve agravada cuando el

estudiante no cuenta con el apoyo económico de la familia, lo que lo obliga a buscar oportunidades laborales para cubrir sus gastos personales y educativos. En este sentido, la falta de estabilidad financiera y la necesidad de equilibrar múltiples responsabilidades pueden desviar la atención y el esfuerzo que los estudiantes podrían dedicar a su formación académica.

4.2.3. Resultado del tercer análisis con selección del conjunto de datos académicos

Para el tercer análisis considerando los mismos algoritmos, se tomaron en cuenta las variables Académicas como se muestra en la ilustración 3. Los resultados de la tabla 10 revelaron que el algoritmo RandomTree y RandomForest destacaron como lo más eficiente en términos de precisión, obteniendo un impresionante 57.59%.

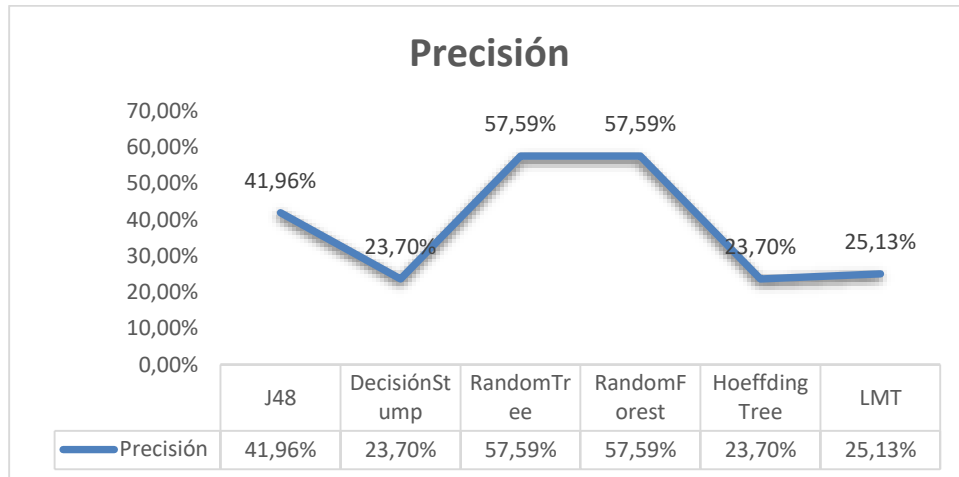
Esta alta precisión indica una concordancia significativa entre las predicciones del modelo y las observaciones reales. Además, RandomTree y RandomForest tuvieron un valor de kappa de 0.49.

Tabla 10: Tabla comparativa del tercer análisis

Algoritmo	Precisión	Kappa	Error absoluto medio	Error cuadrático medio
J48	41.96%	0.29%	0.06	0.17
DecisiónStump	23.70%	0	0.07	0.19
RandomTree	57.59%	0.49	0.04	0.14
RandomForest	57.59%	0.49	0.05	0.15
HoeffdingTree	23.70%	0	0.07	0.19
LMT	25.13%	0.036	0.07	0.19

Fuente: Autor

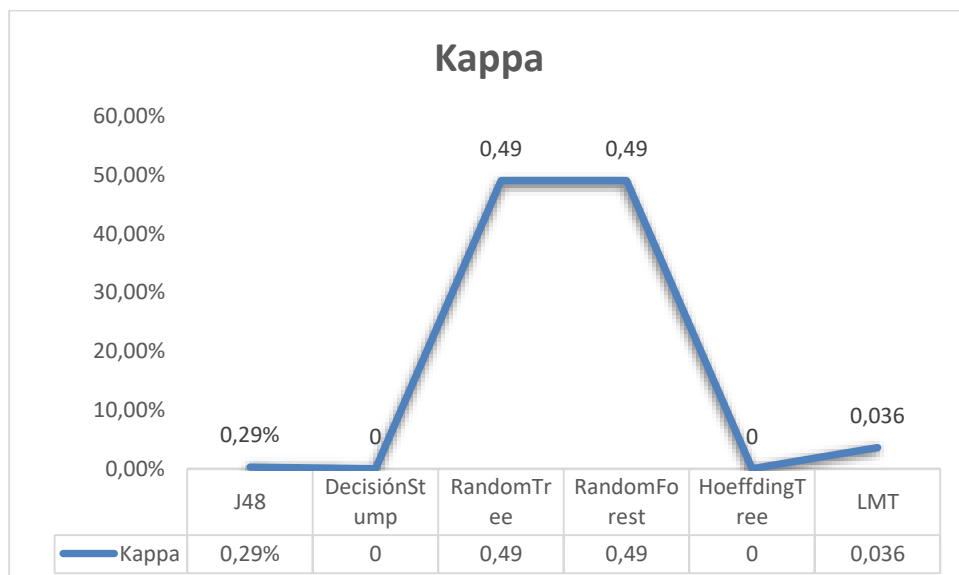
Ilustración 11: Resultados de precisión



Fuente: Autor

Además, en la Ilustración 12 se muestra los valores de Kappa donde el de RandomTree y Random Forest son los que tienen un valor alto a diferencia de los otros algoritmos.

Ilustración 12: Resultados de kappa

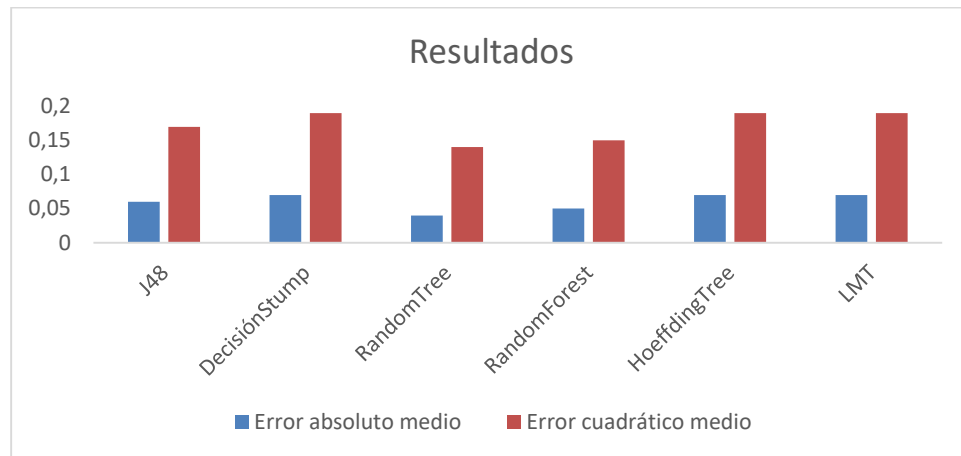


Fuente: Autor

Además de obtener los valores de precisión y kappa, también se obtuvo los resultados de

errores como si indica en la Ilustración 13

Ilustración 13: Resultados de errores



Fuente: Autor

4.2.3.1. Discusión

Dentro del árbol proporcionado en el anexo 5 y los grupos creados en el anexo 6. El análisis de los datos académicos revela patrones significativos que influyen en la deserción estudiantil durante el curso de nivelación. Uno de los factores más destacados es el bajo rendimiento académico, evidenciado por promedios finales inferiores a 7 al concluir el curso de nivelación, así como calificaciones por debajo de 6.25 durante el corte 1, estas cifras sugieren que los estudiantes que no logran alcanzar un nivel satisfactorio en estas etapas iniciales del proceso educativo enfrentando una mayor probabilidad de abandonar sus estudios.

Además, el rendimiento académico en asignaturas específicas juega un papel crucial en la toma de decisiones de los estudiantes respecto a su continuidad en el programa educativo. Entre las materias de mayor desafío se encuentran las relacionadas con las ciencias exactas, especialmente matemáticas, seguida de introducción a la psicología y química. La complejidad inherente a estas asignaturas puede generar desmotivación y dificultades adicionales para los estudiantes, lo que contribuye significativamente a la tasa de

deserción.

4.2.4. Factores que influyen en la deserción estudiantil

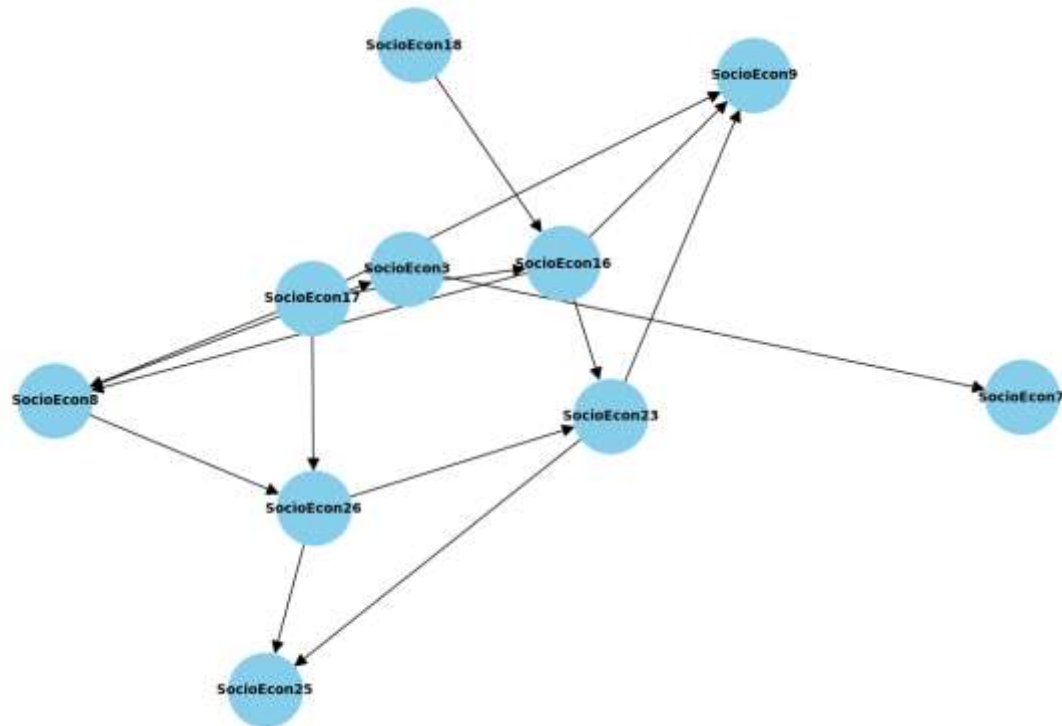
Tabla 11: Factores que causan la deserción estudiantil

VARIABLES	DESCRIPCION
Acad3	Asignatura
Acad5	Corte 1
acad11	Nota final
SocioEcon1	¿El estudiante es cabeza de familia?
SocioEcon2	¿El estudiante depende económicamente de sus padres u otras personas?
SocioEcon6	¿Cuál es el tipo de vivienda?
SocioEcon16	¿Tiene el hogar servicio de internet?
SocioEcon22	Computador

4.3. REGLAS DE ASOCIACIÓN A LOS PERFILES SOCIOECONÓMICOS Y ACADÉMICOS SIMILARES PARA ENCONTRAR PATRONES QUE AYUDEN O CONTRIBUYAN EN LA PREVENCIÓN DE LA DESERCIÓN ESTUDIANTIL

Dentro del análisis para identificar los nodos y poder interpretar como están asociados las variables, se muestra las siguientes variables obtenidas en el análisis como se muestra en la ilustración 14

Ilustración 14: Nodos del conjunto de datos académicos y socioeconómico



Fuente: Autor

Al observar la ilustración 14, se revela una representación visual de los nodos seleccionados en la red bayesiana. Esta representación proporciona una visión clara de las variables socioeconómicas clave que están interconectadas y que influyen en el entorno educativo y doméstico de los estudiantes. Cada nodo representa un aspecto crucial, como la cobertura de gastos estudiantiles, la conectividad digital, la disponibilidad de recursos educativos y el ambiente de vivienda. Al identificar y visualizar estos nodos, se puede apreciar la compleja interacción entre las variables socioeconómicas y su impacto en el desarrollo integral de los estudiantes.

Nodos Seleccionados

SocioEcon3: ¿Quién cubre los gastos del estudiante?

SocioEcon7: ¿Su vivienda es?

SocioEcon8: ¿En el hogar alguien ha usado internet en los últimos 6 meses?

SocioEcon9: ¿Tiene el hogar servicio de teléfono convencional?

SocioEcon16: ¿Tiene el hogar servicio de internet?

SocioEcon17: ¿Tiene computadora de escritorio?

SocioEcon18: ¿Tiene computadora portátil?

SocioEcon23: Enciclopedias

SocioEcon25: Biblioteca

SocioEcon26: Área de recreación

Además de los nodos seleccionados, también se muestran las aristas

(SocioEcon3, SocioEcon7): Quién cubre los gastos del estudiante y el tipo de vivienda.

(SocioEcon23, SocioEcon25): El uso de enciclopedias y la biblioteca.

(SocioEcon23, SocioEcon9): El uso de enciclopedias y tener servicio de teléfono convencional.

(SocioEcon26, SocioEcon23): Área de recreación y el uso de enciclopedias.

(SocioEcon26, SocioEcon25): Área de recreación y la biblioteca.

(SocioEcon8, SocioEcon26): Uso de internet en los últimos 6 meses y tener un área de recreación.

(SocioEcon8, SocioEcon3): Uso de internet en los últimos 6 meses y quién cubre los gastos del estudiante.

(SocioEcon16, SocioEcon8): Servicio de internet y el uso de internet en los últimos 6 meses.

(SocioEcon16, SocioEcon23): Servicio de internet y el uso de enciclopedias.

(SocioEcon16, SocioEcon9): Servicio de internet y tener servicio de teléfono convencional.

(SocioEcon17, SocioEcon16): Tiene computadora de escritorio y tener servicio de

internet.

(SocioEcon17, SocioEcon26): Tiene computadora de escritorio y tener un área de recreación.

(SocioEcon17, SocioEcon9): Tiene computadora de escritorio y tener servicio de teléfono convencional.

(SocioEcon17, SocioEcon8): Tiene computadora de escritorio y el uso de internet en los últimos 6 meses.

(SocioEcon18, SocioEcon16): Tiene computadora portátil y tener servicio de internet.

4.3.1. Discusión

El conjunto de nodos y aristas seleccionadas revela interconexiones significativas entre diversas variables socioeconómicas. Se observa una correlación entre el acceso a servicios digitales, como internet y computadoras, y otros aspectos del hogar, como el tipo de vivienda y la presencia de áreas de recreación. Esto sugiere que el nivel de desarrollo tecnológico en un hogar puede influir en su estilo de vida y entorno físico. Además, la relación entre el uso de recursos educativos, como enciclopedias y bibliotecas, y la disponibilidad de servicios digitales indica cómo se accede y utiliza la información en el hogar, lo que puede tener implicaciones en el acceso a la educación y el conocimiento.

Otro aspecto destacado es la conexión entre la responsabilidad financiera, representada por quién cubre los gastos del estudiante, y el tipo de vivienda. Esta relación sugiere que los recursos financieros disponibles en un hogar pueden influir en su capacidad para acceder a ciertos tipos de vivienda, lo que a su vez puede tener implicaciones en la calidad de vida y el bienestar general de los miembros del hogar. Además, la relación entre el uso de internet en los últimos 6 meses y quién cubre los gastos del estudiante

señala cómo los recursos financieros pueden estar relacionados con el acceso a servicios digitales, lo que puede tener implicaciones en la participación en la educación en línea y otras oportunidades educativas.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

La presente investigación fue realizada con la finalidad de poder detectar patrones académicos en curso de nivelación con deserción en la Universidad Técnica Estatal de Quevedo basado en los factores socioeconómicos, estos factores fueron hallados entre 57 variables que fueron analizadas para obtener información relevante de los estudiantes, concluyendo lo siguiente:

- El estudio realizado con la herramienta Google Colab ayudo con la correlación donde se obtuvieron variables académicas (corte 1, corte 2, examen final, examen supletorio, nota 1, nota supletorio y nota final) y socioeconómicas (jefe de familia, Dependencia económica, Acceso a internet, computador y biblioteca) para comprender el rendimiento académico, donde aparecen diferentes factores como el acceso a internet, recursos educativos en el hogar y apoyo familiar pueden influir significativamente.
- Se emplearon diversos modelos de árboles de decisión usando la herramienta weka, tales como Stump, Hoeffding Tree, J48, LMT, Random Forest, Random Tree y REP Tree, considerando 3 conjuntos de datos: el primero al conjunto completo, como segundo a los datos académicos y por ultimo a los datos socioeconómicos considerando que en la parte académica los estudiantes que tienen una baja calificación en el corte 1 tiene posibilidad que deserten en cualquier momento del curso de nivelación, además de considerar que si no tienen acceso a internet y no poseen una computadora en casa y dependen económicamente de alguien tiene una posibilidad alta que abandone el curso de nivelación siendo un gran factor negativo.

- El análisis a través de reglas de asociación permitió identificar factores internos como externos que influyen en el rendimiento académico considerando que si la primera nota es insuficiente es lo más probable que deserte el curso donde su estatus económico influye en el acceso a recursos y que el rendimiento académico en un curso de nivelación está relacionado con las calificaciones previas y la probabilidad de abandono del curso

5.2. RECOMENDACIONES

El proyecto de investigación es muy importante para el ámbito educativo considerando su fuerte en las universidades, por lo que este proyecto puede ser usado en un futuro procesos de investigación considerando las siguientes recomendaciones:

- Es recomendable continuar utilizando Google Colab u otras herramientas similares para realizar análisis detallados de correlación entre variables académicas y socioeconómicas. Sin embargo, para obtener una comprensión más profunda, se sugiere ampliar el análisis e incluir un mayor número de variables socioeconómicas que puedan influir en el rendimiento académico, como el nivel educativo de los padres, el acceso a programas de tutoría, entre otros.
- Se recomienda continuar utilizando diversos modelos de árboles de decisión con herramientas como Weka para analizar conjuntos de datos académicos y socioeconómicos. Sin embargo, sería útil realizar una evaluación comparativa más exhaustiva de los diferentes modelos considerados, teniendo en cuenta métricas de rendimiento como precisión, sensibilidad y especificidad. Además, se sugiere explorar la posibilidad de utilizar técnicas de ensemble, como Random Forest, para mejorar la robustez y precisión de las predicciones.

- Profundizar en la identificación de factores internos y externos que influyen en el rendimiento académico. Esto podría implicar la inclusión de variables adicionales relacionadas con el entorno familiar, el contexto escolar y las características individuales de los estudiantes. Además, sería beneficioso utilizar técnicas de minería de datos más avanzadas, como redes neuronales o modelos de aprendizaje profundo, para identificar relaciones complejas y no lineales entre las variables estudiadas.
- Que los resultados obtenidos en el proyecto de investigación, las autoridades responsables del proceso de nivelación, consideren estos resultados para efectuar estrategias o soluciones hacia los estudiantes y brindarle el apoyo necesario para que no deserten de los estudios en su inicio o a mediado de las clases y todos los estudiantes puedan aprobar o a su vez plantear estrategias académicas para que los estudiantes puedan aprobar el curso de nivelación.

REFERENCIAS BIBLIOGRAFICAS

- Alaminos F A F. (2023). *Árboles de decisión en R con Random Forest*.
- Albarrán Peña, J. (2019). La deserción estudiantil en la Universidad de Los Andes (Venezuela). *Educación y Humanismo*, 21(36), 60–92.
<https://doi.org/10.17081/eduhum.21.36.2806>
- Alpan, K., & Ilgi, G. S. (2020). Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach. *4th International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2020 - Proceedings*.
<https://doi.org/10.1109/ISMSIT50672.2020.9254720>
- Armas, S. (2020). Análisis de los factores que inciden en la deserción estudiantil de los niños, niñas y adolescentes ecuatorianos en el periodo 2009-2019. *Escuela Politécnica Nacional*, 1–64. <http://bibdigital.epn.edu.ec/handle/15000/21111>
- Arturo Amaya, Flaklin Huerta, C. hidalgo. (2020). *Vista de Big Data, una estrategia para evitar la deserción escolar en las IES.pdf*.
- Balyan, A. K., Ahuja, S., Lilhore, U. K., Sharma, S. K., Manoharan, P., Algarni, A. D., Elmannai, H., & Raahemifar, K. (2022). A Hybrid Intrusion Detection Model Using EGA-PSO and Improved Random Forest Method. *Sensors*, 22(16), 1–20.
<https://doi.org/10.3390/s22165986>
- Castrillón-Gómez, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Predicción de las principales variables que conllevan al abandono estudiantil por medio de técnicas de minería de datos. *Formación Universitaria*, 13(6), 217–228.
<https://doi.org/10.4067/s0718-50062020000600217>
- Dou, Y., & Meng, W. (2023). Comparative analysis of weka-based classification algorithms on medical diagnosis datasets. *Technology and Health Care: Official*

<https://doi.org/10.3233/THC-236034>

Orclever Proceedings of Research and Development, 3(1), 401–416.

Ecuador, corte constitucional del. (2020). *EL PLENO DE LA CORTE*

CONSTITUCIONALES Y LEGALES, EMITE LA SIGUIENTE. 2507(February), 1–9.

influyen en la deserción de los estudiantes de la unidad de admisión y nivelación de la

<http://dspace.unitru.edu.pe/bitstream/handle/UNITRU/10947/Miñano> Guevara%2C

Anali.pdf?sequence=1&isAllowed=y%0Ahttps://repository.upb.edu.co/bitstream/handle

SU.pdf?sequence=1&isAllowed=

LOES. *Registro Oficial Suplemento* 298 de 12-Oct.-2010, 106, 39–40.

Educacion-Superior-LOES.pdf

Performance in Higher Education Institutions Using Decision Tree Analysis.

<https://doi.org/10.9781/ijimai.2018.02.004>

Jaramillo Feijo, L. E., Galindo Villardon, M. P., & Real Cotto, J. J. (2020). Análisis clúster para big data: una aplicación con variables demográficas en provincias del Ecuador Cluster analysis for big data: an application with demographic variables in provinces of Ecuador. *J. Health Med. Sci*, 6(1), 45–50.

Kawelah, W., & Abdala, A. (2019). A Comparative Study on Machine Learning Tools Using WEKA and Rapid Miner with Classifier Algorithms C4. 5 and Decision Stump for Network Intrusion Detection. *European Academic Research*, VII(2/May), 852–861. https://www.academia.edu/download/59385696/EAR_396820190524-80802-1ia119q.pdf

Li, Y., Herrera-Viedma, E., Kou, G., & Morente-Molinera, J. A. (2023). Z-number-valued rule-based decision trees. *Information Sciences*, 643(October 2022), 119252. <https://doi.org/10.1016/j.ins.2023.119252>

Loza, E., & Vinueza, C. (2021). *Diseño de un modelo matemático para estimar la deserción estudiantil mediante técnicas de análisis multivariado en una institución de educación superior*. 1–63. <https://repositorio.uta.edu.ec/jspui/handle/123456789/32219>

Martínez Moncaleano, C. J., & Palencia Fajardo, O. (2021). Modelo de minería de datos para el análisis de la productividad y crecimiento personal en las mujeres emprendedoras: el caso de la Asociación las Rosas. *Suma de Negocios*, 12(26), 23–30. <https://doi.org/10.14349/sumneg/2021.v12.n26.a3>

Maulana, M. F., & Defriani, M. (2020). Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period. *PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic*, 8(1), 39–48. <https://doi.org/10.33558/piksel.v8i1.2018>

Mauricio, D., Orrego, M., Royero, L. F., & Caballero, H. B. (2023). *Minería de datos*

para la identificación de la Deserción Estudiantil en Universidades Públicas Data mining for the identification of Student Dropout in Public Resumen. 1464–1476.

Mena, M., Godoy, W., & Tisalema, S. (2021). Analysis of causes of early dropout of students higher education. *Minerva*, 2(6), 79–89.
<https://doi.org/10.47460/minerva.v2i6.45>

Menéndez Domínguez, V. H., Guerrero Sosa, J. D. T., Castellanos Bolaños, M. E., & Cervera Pérez, J. W. (2022). Uso de la minería de datos para la caracterización de investigadores y cuerpos académicos. In *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo* (Vol. 12, Issue 24).
<https://doi.org/10.23913/ride.v12i24.1144>

Navarro, L. A. T. (2018). Factores que impactan en la deserción estudiantil en las universidades de Latinoamérica del año 2009 al 2019”: una revisión de la literatura científica. *Universidad Privada Del Norte*, 2016–2017.
<https://repositorio.upn.edu.pe/handle/11537/11291>

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3(March). <https://doi.org/10.1016/j.caeai.2022.100066>

Panizzi, M. D. (2019). Establecimiento del estado del arte sobre la Minería de Datos Educacional en el Nivel Superior: Un Estudio de Mapeo Sistemático. *Ciencias Exactas, Naturales, Experimentales y de La Salud*, 51–60.

Peña, G. A. (2021). *Caracterización de la educación superior ecuatoriana en el contexto de la producción científica nacional y latinoamericana*. 6.

Perez, B., Castellanos, C., & Correal, D. (2018). Applying Data Mining Techniques to

- Predict Student Dropout: A Case Study. *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence, ColCACI 2018 - Proceedings*, 1–6.
<https://doi.org/10.1109/ColCACI.2018.8484847>
- Ponce-Sanchez, C., Andrade-Arenas, L., & Sotomayor-Beltran, C. (2019). Admission options and their implications for student desertion at the National University of Engineering. *Proceedings of the 2019 International Symposium on Engineering Accreditation and Education, ICACIT 2019*, 1–4.
<https://doi.org/10.1109/ICACIT46824.2019.9130250>
- Rascón-Larios, A. M. (2022). Deserción Universitaria; un estudio preliminar en los Programas Educativos de Administración y Contaduría Pública. In *Revista de Investigaciones Universidad del Quindío* (Vol. 34, Issue 2, pp. 188–199).
<https://doi.org/10.33975/riuq.vol34n2.969>
- Rico Páez, A., & Gaytán Ramírez, N. D. (2022). Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería. *IE Revista de Investigación Educativa de La REDIECH*, 13, e1426.
https://doi.org/10.33010/ie_rie_rediech.v13i0.1426
- Rivera Vergaray, K. (2021). Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica. *Revista Innovación y Software*, 2(2), 6–13.
- Sabharwal, A., & Selman, B. (2020). Book review. *Artificial Intelligence*, 175(5–6), 935–937. <https://doi.org/10.1016/j.artint.2011.01.005>
- Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. *2018 IEEE International Work Conference on Bioinspired Intelligence, IWOBI 2018 -*

Proceedings. <https://doi.org/10.1109/IWOBI.2018.8464191>

UNESCO. (2021). De los efectos de la pandemia COVID -19 sobre la deserción universitaria: desgaste docente y bienestar psicológico estudiantil. *Revista Educación Superior y Sociedad (ESS)*, 33(2), 402–421. <https://doi.org/10.54674/ess.v33i2.360>

Urbina-Nájera, A. B., Téllez-Velázquez, A., & Cruz Barbosa, R. (2021). Patrones que identifican a estudiantes universitarios desertores aplicando minería de datos educativa. *Revista Electrónica de Investigación Educativa*, 23, 1–15. <https://doi.org/10.24320/redie.2021.23.e29.3918>

Vargas, L., & Mesa, E. (2021). Introducción Al Análisis De Datos Con RStudio. In *Cenipalma*. www.cenipalma.org

Viloria, A., Padilla, J. G., Vargas-Mercado, C., Hernández-Palma, H., Llinas, N. O., & David, M. A. (2019). Integration of data technology for analyzing university dropout. *Procedia Computer Science*, 155(2018), 569–574. <https://doi.org/10.1016/j.procs.2019.08.079>

Zoila Jessenia Fabara-Sarmiento, Janeth Pilar Diaz-Vera, A. K. R.-R. (2022). *Data Science para la Predicción del Rendimiento Académico | Fabara-Sarmiento | Dominio de las Ciencias*. <https://dominiodelasciencias.com/ojs/index.php/es/article/view/2481/5522>

ANEXOS

ANEXO 1: RESULTADO DE ANTIPLAGIO DE LA HERRAMIENTA ANTI-SIMILITUD COMPULATIO

CERTIFICADO DEL SISTEMA ANTIPLAGIO COMPILATIO

Quevedo, 5 de junio del 2024

Ing. Byron Oviedo Bayas, PhD

DECANO DE POSGRADO UTEQ

Presente. -

En su despacho,

Por medio de la presente se da a conocer que el proyecto de investigación titulado **“Detección de patrones académicos en curso de nivelación con deserción en la Universidad Técnica Estatal de Quevedo basado en los factores socioeconómicos”**, del posgradista **Ing. Jean Carlos Almeida Murillo** en la **Maestría en Ciencia de Datos**, se obtuvo un 1% en la herramienta anti-similitud COMPILATIO; en consecuencia, el proyecto de investigación se encuentra listo para seguir con el proceso de defensa.

The screenshot shows a certificate from the 'CERTIFICADO DE ANÁLISIS magister' system. The title of the document is 'TESIS_MAESTRIA_CIENCIA_DATOS_ALMEIDA'. The analysis results show '1% Textos sospechosos' (Suspicious texts) and '2% Similitudes' (Similarities). The similarities are broken down as: '< 1% similitudes entre capítulos (ignorado)' (ignored), '< 1% entre las fuentes mencionadas (ignorado)' (ignored), and 'Δγ < 1% Idiomas no reconocidos' (unrecognized languages). The document details include: 'Nombre del documento: TESIS_MAESTRIA_CIENCIA_DATOS_ALMEIDA.pdf', 'ID del documento: 9556fb861468e4e2ff324eb3399671c906a009fc', 'Tamaño del documento original: 2.2 MB', 'Depositante: EDUARDO AMABLE SAMANIEGO MENA', 'Fecha de depósito: 5/6/2024', 'Tipo de carga: Interface', 'Fecha de fin de análisis: 5/6/2024', 'Número de palabras: 25.130', and 'Número de caracteres: 164.788'.

Particular que comunico para los fines pertinentes.



Firmado electrónicamente por:
**EDUARDO
AMABLE
SAMANIEGO
MENA**

Ing. Eduardo Samaniego Mena, MSc.

Director Proyecto de Investigación

ANEXO 2: FIRMA DE PRIVACIDAD PARA OBTENCIÓN DE LOS DATOS

	UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO	
	ACUERDO DE CONFIDENCIALIDAD Y NO DIVULGACIÓN DE LA INFORMACIÓN DE LA UTEQ	
	VERSIÓN: 1.0	PÁGINA: 1 de 3

ACUERDO DE CONFIDENCIALIDAD Y NO DIVULGACIÓN DE LA INFORMACIÓN DE LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

I. - ANTECEDENTES:

La Universidad Técnica Estatal de Quevedo al ser la entidad que administra la plataforma informática interna del Sistema de Gestión Académica (SGA), está facultada para registrar y dar buen uso de los datos personales de los aspirantes y estudiantes, en virtud de aquello y al ejercer el rol como responsable define las finalidades concretas sobre las cuales versará el tratamiento de datos que realizarán las personas a quienes se les facilite información, con el fin de llevar a cabo el correcto y adecuado tratamiento que únicamente serán para fines académicos e institucionales.

II. – BASE LEGAL:

- 2.1. La Constitución de la República del Ecuador establece en su artículo 11.9, que el deber más alto del Estado consiste en respetar y hacer respetar los derechos garantizados en la Constitución.
- 2.2. El artículo 226 de la Constitución de la República del Ecuador prevé que: *"Las instituciones del Estado sus organismos y dependencias, y las servidoras o servidores públicos, tendrán el deber de coordinar acciones para el cumplimiento de sus fines y hacer efectivo el goce y ejercicio de los derechos reconocidos en la Constitución"*.
- 2.3. El artículo 66 numeral 19 de la Constitución de la República del Ecuador dispone: *"Se reconoce y garantizará a las personas: (...) El derecho a la protección de datos de carácter personal, que incluye el acceso y la decisión sobre información y datos de este carácter, así como su correspondiente protección. La recolección, archivo, procesamiento, distribución o difusión de estos datos o información requerirán la autorización del titular o el mandato de la Ley"*.
- 2.4. La Ley Orgánica de Protección de Datos Personales en su artículo 10, determina los principios que se rige esta ley, entre los que se determinan:
 - g) Confidencialidad. - El tratamiento de datos personales debe concebirse sobre la base del debido sigilo y secreto, es decir, no debe tratarse o comunicarse para un fin distinto para el cual fueron recogidos, a menos que concurra una de las causales que habiliten un nuevo tratamiento conforme los supuestos de tratamiento legítimo señalados en esta Ley.



- h) Seguridad de datos personales. - Los responsables y encargados de tratamiento de

Para la aplicación de sanciones se tomará en cuenta lo establecido en la Ley Orgánica de Protección de Datos Personales, Código de Ética de la UTEQ y demás normativa aplicable; sin perjuicio de las acciones civiles y penales que procedan en cada caso.

IX. – DOMICILIO Y CONTACTO DEL RECEPTOR DE LA INFORMACIÓN:

El domicilio legal del RECEPTOR DE LA INFORMACIÓN se encuentra establecido en la dirección: SECTOR 20 DE FEBRERO, QUEVEDO - LOS RÍOS - ECUADOR. Contacto: SAMANIEGO MENA EDUARDO AMABLE, con correo electrónico esamaniego@uteq.edu.ec

El domicilio legal del RECEPTOR DE LA INFORMACIÓN se encuentra establecido en la dirección: CALLE CUBA Y B, sector JOSEFINA 1, QUEVEDO - LOS RÍOS - ECUADOR. Contacto: ALMEIDA MURILLO JEAN CARLOS, con correo electrónico jean.almeida2015@uteq.edu.ec

XI. - NOTIFICACIONES:

En el evento de que se produzca el incumplimiento de alguna de las cláusulas estipuladas en el presente acuerdo, la parte afectada, notificará del incumplimiento a la máxima autoridad de la institución involucrada, sin perjuicio de las acciones y sanciones previstas en la normativa vigente.

Una vez comprendido por los comparecientes el contenido y efectos del presente instrumento expresamente se ratifican en él, para fe y constancia se firma el presente documento por quienes en él intervinieron.

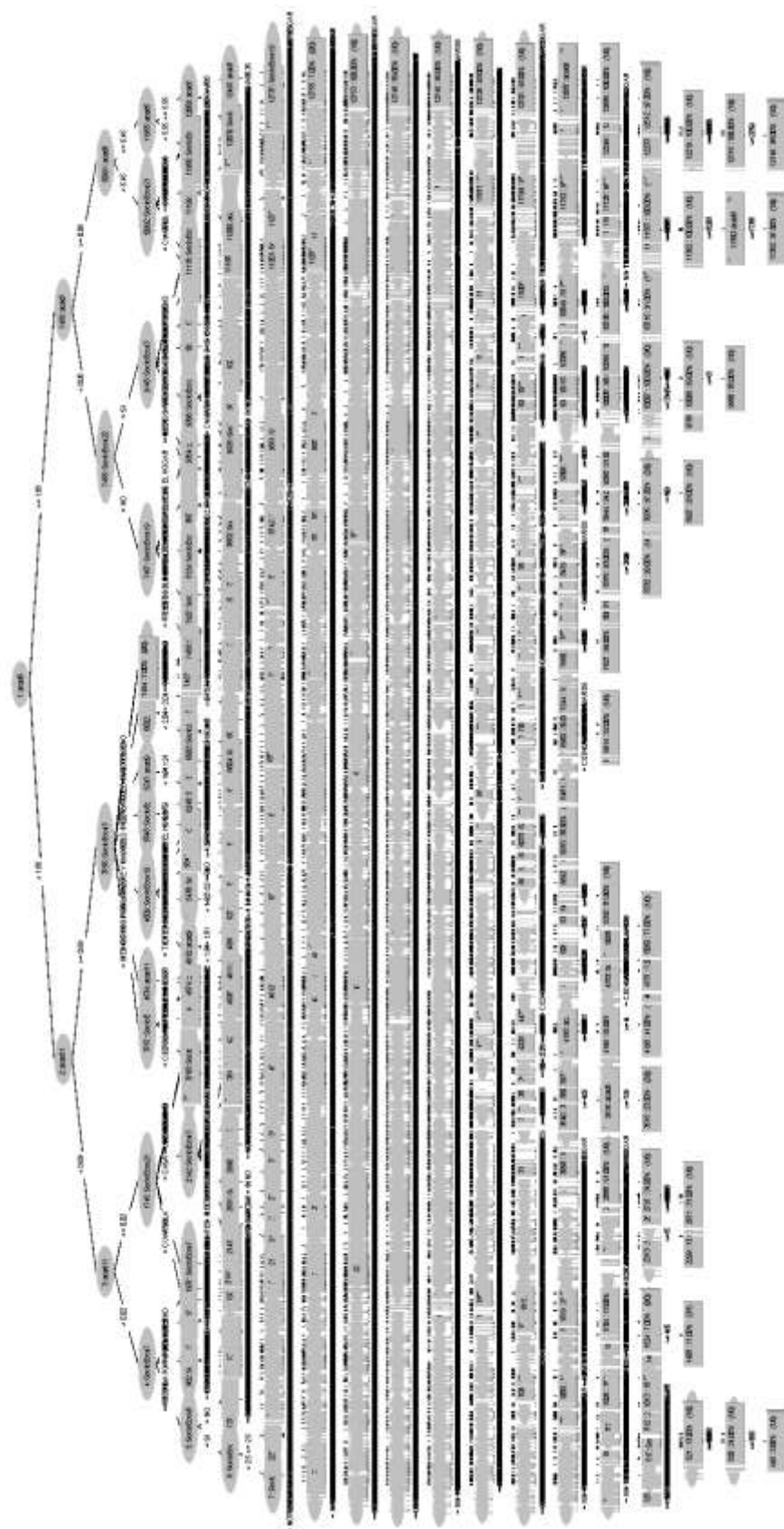

SAMANIEGO MENA EDUARDO AMABLE
CC.: 1718071929
Receptor de la Información: Profesor



Quevedo, 16 de enero de 2024

ALMEIDA MURILLO JEAN CARLOS
CC.: 1206774604
Receptor de la Información: Estudiante

84



ANEXO 4: GRUPOS DEL CONJUNTO DE DATOS DEL PRIMER ANALISIS

acad6 < 1.85

| acad11 < 0.69

| | acad11 < 0.02

| | | SocioEcon7 = PROPIA

| | | SocioEcon8 = SI

| | | SocioEcon20 < 2.5

| | | SocioEcon19 = TIENE 3 CELULARES

| | | SocioEcon21 = CUARTO

| | | SocioEcon3 = MEDIOS PROPIOS

| | | SocioEcon22 = NO : 16.00% (3/2)

| | | SocioEcon22 = SI

| | | SocioEcon17 = NO : 11.00% (1/0)

| | | SocioEcon17 = SI : 30.00% (4/3)

| | | SocioEcon3 = OTROS FAMILIARES : 7.00% (0/0)

| | | SocioEcon3 = PADRE Y MADRE

| | | SocioEcon20 < 1.5 : 52.00% (1/0)

| | | SocioEcon20 >= 1.5 : 10.00% (4/3)

| | | SocioEcon3 = SOLO PADRE : 7.00% (0/0)

| | | SocioEcon3 = CONYUGE : 7.00% (0/0)

| | | SocioEcon3 = SOLO MADRE : 29.00% (2/1)

| | | SocioEcon3 = OTROS : 7.00% (0/0)

| | | SocioEcon3 = NO RELLENO : 7.00% (0/0)

| | | SocioEcon21 = SALA

| | | SocioEcon24 = SI : 30.00% (2/1)

| | | SocioEcon24 = NO : 22.00% (2/1)

| | | SocioEcon21 = CASA

| | | SocioEcon26 = NO

| | | SocioEcon3 = MEDIOS PROPIOS

| | | SocioEcon16 = SI : 30.00% (4/3)

| | | SocioEcon16 = NO : 10.00% (4/3)

| | | SocioEcon3 = OTROS FAMILIARES : 7.00% (0/0)

| | | SocioEcon3 = PADRE Y MADRE

| | | SocioEcon18 = SI : 19.00% (2/1)

| | | SocioEcon18 = NO

| | | SocioEcon2 = NO

| | | SocioEcon22 = NO : 29.00% (3/2)

| | | SocioEcon22 = SI : 4.00% (1/0)

| | | SocioEcon2 = SI

| | | SocioEcon17 = NO : 1.00% (1/0)

| | | SocioEcon17 = SI

| | | SocioEcon22 = NO : 22.00% (2/1)

| | | SocioEcon22 = SI : 24.00% (1/0)

| | | SocioEcon3 = SOLO PADRE

| | | SocioEcon9 = NO

| | | SocioEcon18 = SI

| | | SocioEcon16 = SI : 7.00% (3/2)

| | | SocioEcon16 = NO : 17.00% (1/0)

| | | SocioEcon18 = NO

| | | SocioEcon22 = NO : 16.00% (2/1)

| | | SocioEcon22 = SI : 21.00% (1/0)

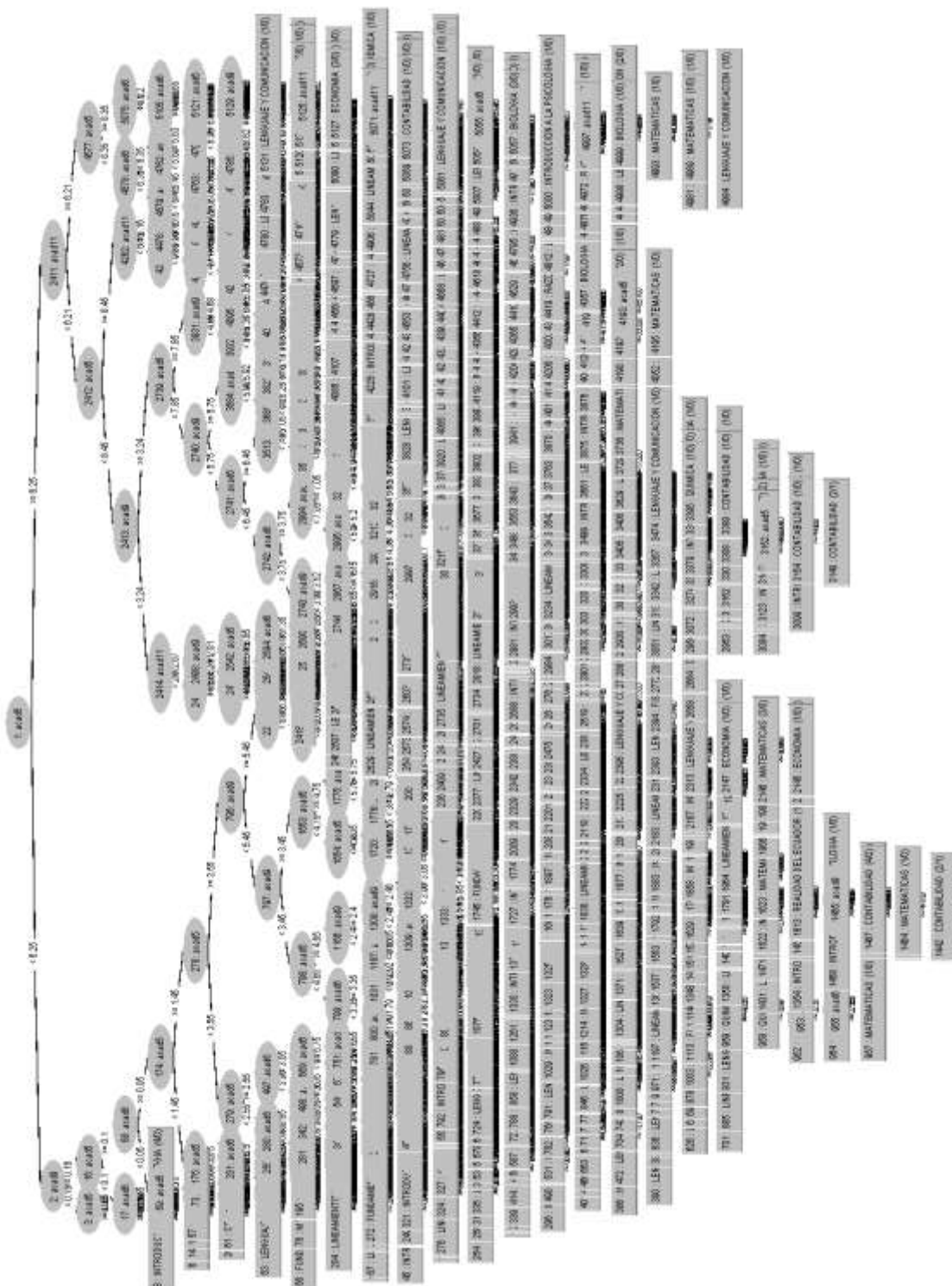
| | | SocioEcon9 = SI : 20.00% (2/1)

| | | SocioEcon3 = CONYUGE : 21.00% (1/0)

		SocioEcon25 = SI : 20.00% (1/0)
		SocioEcon17 = SI
		SocioEcon23 = NO
		SocioEcon20 < 1.5 : 7.00% (4/3)
		SocioEcon20 >= 1.5 : 22.00% (2/1)
		SocioEcon23 = SI : 7.00% (4/3)
		SocioEcon2 = SI
		SocioEcon9 = NO
		SocioEcon22 = NO
		SocioEcon26 = NO : 30.00% (1/0)
		SocioEcon26 = SI : 7.00% (3/2)
		SocioEcon22 = SI
		SocioEcon3 = MEDIOS PROPIOS : 7.00% (0/0)
		SocioEcon3 = OTROS FAMILIARES : 7.00% (0/0)
		SocioEcon3 = PADRE Y MADRE
		SocioEcon18 = SI : 31.00% (1/0)
		SocioEcon18 = NO
		SocioEcon20 < 1.5 : 23.00% (2/0)
		SocioEcon20 >= 1.5 : 10.00% (2/0)
		SocioEcon3 = SOLO PADRE : 7.00% (0/0)
		SocioEcon3 = CONYUGE : 7.00% (0/0)
		SocioEcon3 = SOLO MADRE
		SocioEcon18 = SI : 16.00% (2/1)
		SocioEcon18 = NO : 13.00% (3/2)
		SocioEcon3 = OTROS : 7.00% (0/0)
		SocioEcon3 = NO RELLENO : 7.00% (0/0)
		SocioEcon9 = SI : 12.00% (1/0)
		SocioEcon24 = NO
		SocioEcon22 = NO
		SocioEcon2 = NO : 7.00% (3/2)
		SocioEcon2 = SI : 21.00% (1/0)
		SocioEcon22 = SI
		SocioEcon20 < 1.5 : 54.00% (2/1)
		SocioEcon20 >= 1.5
		SocioEcon18 = SI : 18.00% (1/0)
		SocioEcon18 = NO : 26.00% (4/3)
		SocioEcon21 = OTROS : 25.00% (2/1)
		SocioEcon21 = UNIVERSIDAD : 7.00% (0/0)
		SocioEcon21 = BIBLIOTECA : 7.00% (0/0)
		SocioEcon21 = NO RELLENO : 7.00% (0/0)
		SocioEcon21 = PATIO : 7.00% (0/0)
		SocioEcon19 = TIENE 2 CELULARES
		SocioEcon3 = MEDIOS PROPIOS
		SocioEcon26 = NO
		SocioEcon18 = SI : 7.00% (4/3)
		SocioEcon18 = NO : 37.00% (1/0)
		SocioEcon26 = SI
		SocioEcon22 = NO : 20.00% (3/2)
		SocioEcon22 = SI : 48.00% (2/1)
		SocioEcon3 = OTROS FAMILIARES : 7.00% (0/0)
		SocioEcon3 = PADRE Y MADRE
		SocioEcon18 = SI
		SocioEcon22 = NO

	SocioEcon21 = CASA : 13.00% (3/2)
	SocioEcon21 = OTROS : 7.00% (0/0)
	SocioEcon21 = UNIVERSIDAD : 7.00% (0/0)
	SocioEcon21 = BIBLIOTECA : 7.00% (0/0)
	SocioEcon21 = NO RELLENO : 7.00% (0/0)
	SocioEcon21 = PATIO : 7.00% (0/0)
	SocioEcon25 = SI : 23.00% (2/0)
	SocioEcon9 = SI : 26.00% (2/1)
	SocioEcon18 = NO
	SocioEcon24 = SI
	SocioEcon17 = NO
	SocioEcon16 = SI
	SocioEcon2 = NO
	SocioEcon25 = NO
	SocioEcon9 = NO : 13.00% (3/2)
	SocioEcon9 = SI : 78.00% (1/0)
	SocioEcon25 = SI : 33.00% (3/2)
	SocioEcon2 = SI : 33.00% (1/0)
	SocioEcon16 = NO
	SocioEcon20 < 5.5 : 10.00% (2/1)
	SocioEcon20 >= 5.5
	SocioEcon20 < 7.5 : 19.00% (1/0)
	SocioEcon20 >= 7.5
	SocioEcon20 < 8.5 : 27.00% (1/0)
	SocioEcon20 >= 8.5 : 13.00% (1/0)
	SocioEcon17 = SI : 80.00% (1/0)
	SocioEcon24 = NO
	SocioEcon16 = SI
	SocioEcon25 = NO
	SocioEcon20 < 3.5 : 13.00% (3/2)
	SocioEcon20 >= 3.5 : 16.00% (2/1)
	SocioEcon25 = SI : 16.00% (1/0)
	SocioEcon16 = NO : 100.00% (3/2)
	SocioEcon19 = TIENE 1 CELULAR
	SocioEcon22 = NO
	SocioEcon20 < 4.5
	SocioEcon23 = NO
	SocioEcon2 = NO : 9.00% (4/2)
	SocioEcon2 = SI : 5.00% (3/2)
	SocioEcon23 = SI
	SocioEcon16 = SI : 7.00% (1/0)
	SocioEcon16 = NO : 44.00% (1/0)
	SocioEcon20 >= 4.5
	SocioEcon2 = NO : 57.00% (1/0)
	SocioEcon2 = SI
	SocioEcon20 < 6 : 18.00% (1/0)
	SocioEcon20 >= 6 : 16.00% (1/0)
	SocioEcon22 = SI
	SocioEcon17 = NO
	SocioEcon21 = CUARTO : 18.00% (3/2)
	SocioEcon21 = SALA : 7.00% (0/0)
	SocioEcon21 = CASA : 0.00% (2/0)
	SocioEcon21 = OTROS : 7.00% (0/0)

					SocioEcon21 = UNIVERSIDAD : 7.00% (0/0)
					SocioEcon21 = BIBLIOTECA : 7.00% (0/0)
					SocioEcon21 = NO RELLENO : 7.00% (0/0)
					SocioEcon21 = PATIO : 7.00% (0/0)
					SocioEcon17 = SI
					SocioEcon9 = NO : 42.00% (3/2)
					SocioEcon9 = SI
					SocioEcon26 = NO : 14.00% (4/3)
					SocioEcon26 = SI : 100.00% (4/3)
					SocioEcon19 = NO TIENEN CELULAR EN EL HOGAR : 7.00% (0/0)
					SocioEcon19 = NO RELLENO



ANEXO 6: GRUPOS DEL CONJUNTO DE DATOS DE VARIABLES ACADÉMICAS

```

acad5 < 4.25
| acad7 < 1.5
| | acad5 < 2.05
| | | acad5 < 0.65
| | | | acad9 < 0.57
| | | | | acad9 < 0.12 : 0 (1723/0)
| | | | | acad9 >= 0.12
| | | | | | acad9 < 0.3 : 0.18 (137/0)
| | | | | | acad9 >= 0.3
| | | | | | | acad5 < 0.2 : 0.39 (32/0)
| | | | | | | acad5 >= 0.2 : 0.48 (19/0)
| | | | | acad9 >= 0.57
| | | | | | acad9 < 1.52
| | | | | | | acad6 < 2.4
| | | | | | | | acad9 < 0.67 : 0.62 (6/0)
| | | | | | | | acad9 >= 0.67 : 0.77 (8/0)
| | | | | | | acad6 >= 2.4
| | | | | | | | acad6 < 2.85 : 0.94 (6/0)
| | | | | | | | acad6 >= 2.85
| | | | | | | | | acad9 < 1.19 : 1.14 (4/0)
| | | | | | | | | acad9 >= 1.19 : 1.27 (3/0)
| | | | | acad9 >= 1.52
| | | | | | acad6 < 6.15 : 1.79 (2/0)
| | | | | | acad6 >= 6.15 : 2.45 (4/0)
| | | | acad5 >= 0.65
| | | | | acad5 < 1.45
| | | | | | acad9 < 0.82
| | | | | | | acad9 < 0.44 : 0.35 (374/0)
| | | | | | | acad9 >= 0.44
| | | | | | | | acad9 < 0.61 : 0.49 (88/0)
| | | | | | | | acad9 >= 0.61 : 0.7 (28/0)
| | | | | acad9 >= 0.82
| | | | | | acad9 < 1.25
| | | | | | | acad9 < 1 : 0.94 (13/0)
| | | | | | | acad9 >= 1
| | | | | | | | acad9 < 1.18 : 1.06 (5/0)
| | | | | | | | acad9 >= 1.18 : 1.23 (1/0)
| | | | | acad9 >= 1.25
| | | | | | acad9 < 1.58
| | | | | | | acad9 < 1.36 : 1.29 (4/0)
| | | | | | | acad9 >= 1.36 : 1.43 (7/0)
| | | | | acad9 >= 1.58
| | | | | | acad6 < 4.2 : 1.67 (2/0)
| | | | | | acad6 >= 4.2
| | | | | | | acad6 < 4.45 : 1.98 (2/0)
| | | | | | | acad6 >= 4.45 : 1.85 (1/0)
| | | | acad5 >= 1.45
| | | | | acad9 < 1.31
| | | | | | acad9 < 0.82 : 0.62 (308/0)
| | | | | | acad9 >= 0.82
| | | | | | acad5 < 1.95

```



```

| | | | | | | | | acad6 < 3.75 : 2.36 (12/0)
| | | | | | | | | acad6 >= 3.75 : 2.51 (3/0)
| | | | | | | | | acad9 >= 2.54
| | | | | | | | | | acad5 < 3.45
| | | | | | | | | | | acad6 < 4.6 : 2.59 (5/0)
| | | | | | | | | | | acad6 >= 4.6
| | | | | | | | | | | | acad6 < 5.05 : 2.83 (4/0)
| | | | | | | | | | | | acad6 >= 5.05 : 2.96 (4/0)
| | | | | | | | | | | | acad5 >= 3.45
| | | | | | | | | | | | | acad6 < 4.4 : 2.7 (1/0)
| | | | | | | | | | | | | acad6 >= 4.4 : 2.9 (2/0)
| | | | | | | | | | | | | acad5 >= 3.65
| | | | | | | | | | | | | | acad5 < 4.05
| | | | | | | | | | | | | | | acad6 < 3.65
| | | | | | | | | | | | | | | | acad6 < 2.95 : 2.3 (6/0)
| | | | | | | | | | | | | | | | acad6 >= 2.95 : 2.46 (8/0)
| | | | | | | | | | | | | | | | acad6 >= 3.65
| | | | | | | | | | | | | | | | | acad9 < 2.87 : 2.77 (12/0)
| | | | | | | | | | | | | | | | | acad9 >= 2.87 : 2.93 (4/0)
| | | | | | | | | | | | | | | | | acad5 >= 4.05
| | | | | | | | | | | | | | | | | | acad5 < 4.15
| | | | | | | | | | | | | | | | | | | acad9 < 2.71
| | | | | | | | | | | | | | | | | | | | acad9 < 2.48 : 2.38 (2/0)
| | | | | | | | | | | | | | | | | | | | acad9 >= 2.48 : 2.63 (2/0)
| | | | | | | | | | | | | | | | | | | | acad9 >= 2.71
| | | | | | | | | | | | | | | | | | | | | acad6 < 4.1 : 2.8 (2/0)
| | | | | | | | | | | | | | | | | | | | | acad6 >= 4.1 : 2.96 (3/0)
| | | | | | | | | | | | | | | | | | | | | acad5 >= 4.15
| | | | | | | | | | | | | | | | | | | | | | acad6 < 3.4
| | | | | | | | | | | | | | | | | | | | | | | acad9 < 2.52 : 2.41 (4/0)
| | | | | | | | | | | | | | | | | | | | | | | acad9 >= 2.52 : 2.6 (2/0)
| | | | | | | | | | | | | | | | | | | | | | | acad6 >= 3.4
| | | | | | | | | | | | | | | | | | | | | | | | acad6 < 3.9 : 2.76 (5/0)
| | | | | | | | | | | | | | | | | | | | | | | | acad6 >= 3.9 : 2.91 (2/0)
| | | | | | | | | | | | | | | | | | | | | | | | acad9 >= 3
| | | | | | | | | | | | | | | | | | | | | | | | | acad8 < 1.85
| | | | | | | | | | | | | | | | | | | | | | | | | | acad9 < 3.52
| | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 < 3.22 : 3.11 (32/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 >= 3.22
| | | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 < 3.38 : 3.31 (13/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 >= 3.38 : 3.43 (10/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 >= 3.52
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 < 3.9
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 < 3.7 : 3.59 (11/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 >= 3.7 : 3.79 (25/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad9 >= 3.9
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad6 < 8.25 : 4.01 (9/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad6 >= 8.25 : 4.31 (2/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad8 >= 1.85
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad6 < 7.7 : 6.88 (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad6 >= 7.7 : 5.21 (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad7 >= 1.5
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | acad5 < 3.45 : 4.99 (2/0)

```

```

| | acad5 >= 3.45
| | | acad6 < 7.45
| | | | acad7 < 4.5 : 5.26 (1/0)
| | | | acad7 >= 4.5 : 5.53 (1/0)
| | | acad6 >= 7.45
| | | | acad7 < 4.3 : 4.99 (1/0)
| | | | acad7 >= 4.3
| | | | | acad5 < 4.1
| | | | | | acad7 < 6.5 : 5.94 (2/0)
| | | | | | acad7 >= 6.5 : 6.46 (1/0)
| | | | | acad5 >= 4.1 : 6.79 (1/0)
acad5 >= 4.25
| acad9 < 4.24
| | acad9 < 2.88
| | | acad6 < 0.95
| | | | acad9 < 2.12
| | | | | acad9 < 1.81
| | | | | | acad5 < 4.65 : 1.57 (70/0)
| | | | | | acad5 >= 4.65 : 1.72 (94/0)
| | | | | acad9 >= 1.81
| | | | | | acad9 < 1.94 : 1.87 (75/0)
| | | | | | acad9 >= 1.94 : 2.02 (65/0)
| | | | acad9 >= 2.12
| | | | | acad9 < 2.49
| | | | | | acad9 < 2.29 : 2.21 (74/0)
| | | | | | acad9 >= 2.29 : 2.39 (88/0)
| | | | | acad9 >= 2.49
| | | | | | acad6 < 0.65
| | | | | | | acad9 < 2.67 : 2.58 (42/0)
| | | | | | | acad9 >= 2.67 : 2.77 (40/0)
| | | | | | acad6 >= 0.65
| | | | | | | acad6 < 0.85
| | | | | | | | acad9 < 2.7 : 2.55 (10/0)
| | | | | | | | acad9 >= 2.7 : 2.78 (3/0)
| | | | | | | acad6 >= 0.85
| | | | | | | | acad9 < 2.65 : 2.56 (1/0)
| | | | | | | | acad9 >= 2.65 : 2.78 (2/0)
| | | | acad6 >= 0.95
| | | | | acad5 < 5.65
| | | | | | acad9 < 2.46
| | | | | | | acad5 < 4.75
| | | | | | | | acad6 < 1.75
| | | | | | | | | acad6 < 1.2 : 1.92 (7/0)
| | | | | | | | | acad6 >= 1.2 : 2.12 (21/0)
| | | | | | | | acad6 >= 1.75
| | | | | | | | | acad9 < 2.33 : 2.26 (10/0)
| | | | | | | | | acad9 >= 2.33 : 2.39 (15/0)
| | | | | | | acad5 >= 4.75
| | | | | | | | acad9 < 2.29 : 2.21 (14/0)
| | | | | | | | acad9 >= 2.29 : 2.39 (26/0)
| | | | | | acad9 >= 2.46
| | | | | | | acad9 < 2.67 : 2.56 (35/0)
| | | | | | | acad9 >= 2.67 : 2.77 (43/0)

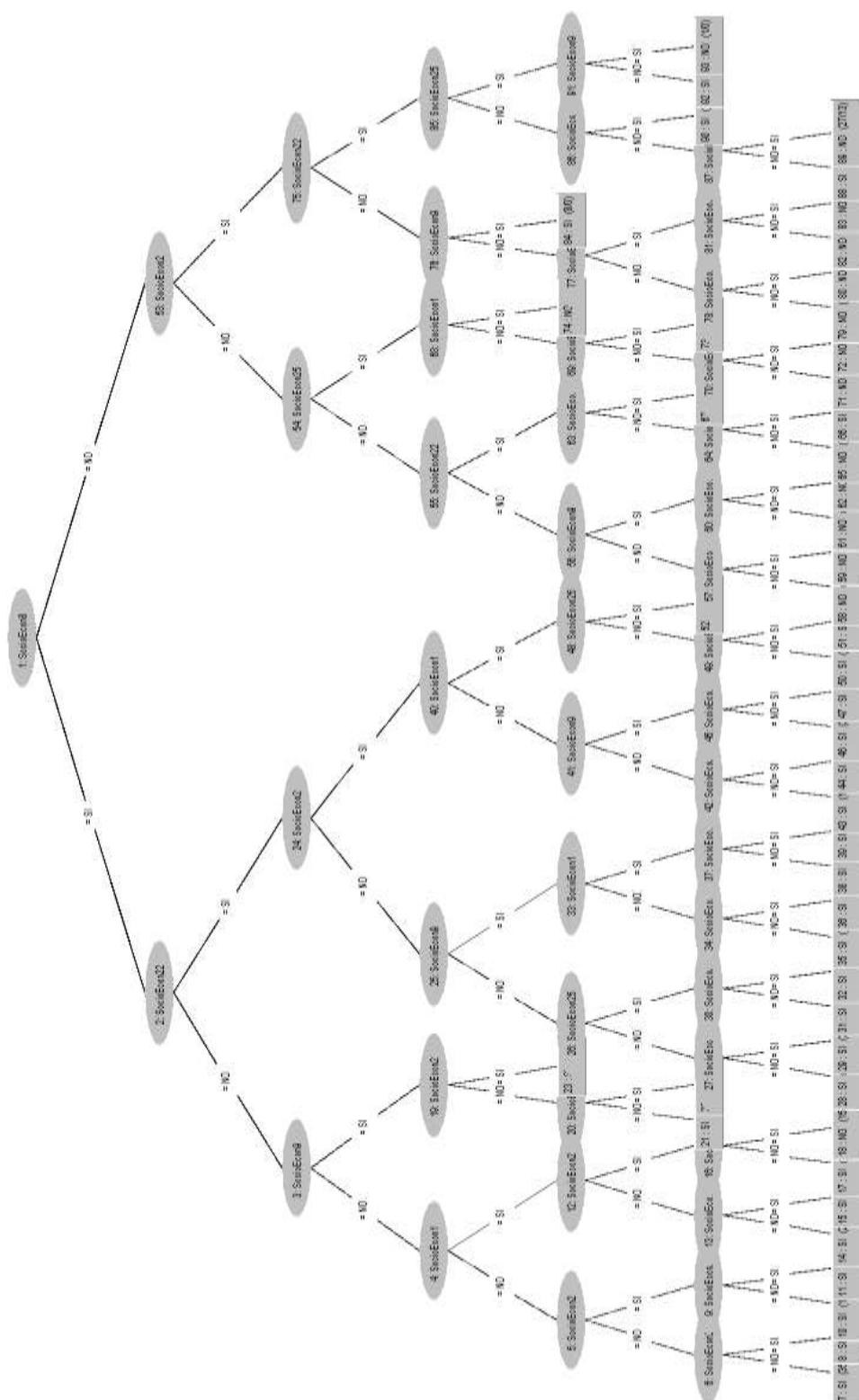
```

```

| | | | acad5 >= 5.65
| | | | | acad5 < 6.35
| | | | | | acad5 < 6.15
| | | | | | | acad6 < 1.75
| | | | | | | | acad6 < 1.05 : 2.35 (2/0)
| | | | | | | | acad6 >= 1.05 : 2.53 (22/0)
| | | | | | | | acad6 >= 1.75
| | | | | | | | acad9 < 2.79 : 2.72 (10/0)
| | | | | | | | acad9 >= 2.79 : 2.84 (9/0)
| | | | | | | | acad5 >= 6.15
| | | | | | | | acad6 < 1.55 : 2.65 (9/0)
| | | | | | | | acad6 >= 1.55 : 2.8 (5/0)
| | | | | | | | acad5 >= 6.35 : 2.78 (28/0)

```

ANEXO 7: ARBOL CON SELECCIÓN DE VARIABLES SOCIOECNOMICAS



ANEXO 8: GRUPOS DEL CONJUNTO DE DATOS DE VARIABLES

SOCIOECONOMICAS

SocioEcon8 = SI
| SocioEcon22 = NO
| | SocioEcon9 = NO
| | | SocioEcon1 = NO
| | | | SocioEcon2 = NO
| | | | SocioEcon25 = NO : SI (350/91)
| | | | SocioEcon25 = SI : SI (24/7)
| | | | SocioEcon2 = SI
| | | | SocioEcon25 = NO : SI (1021/199)
| | | | SocioEcon25 = SI : SI (61/16)
| | | SocioEcon1 = SI
| | | | SocioEcon2 = NO
| | | | SocioEcon25 = NO : SI (283/85)
| | | | SocioEcon25 = SI : SI (11/0)
| | | | SocioEcon2 = SI
| | | | SocioEcon25 = NO : SI (129/25)
| | | | SocioEcon25 = SI : NO (15/5)
| | SocioEcon9 = SI
| | | SocioEcon2 = NO
| | | | SocioEcon1 = NO : SI (37/4)
| | | | SocioEcon1 = SI : SI (12/0)
| | | SocioEcon2 = SI : SI (94/0)
| SocioEcon22 = SI
| | SocioEcon2 = NO
| | | SocioEcon9 = NO
| | | | SocioEcon25 = NO
| | | | SocioEcon1 = NO : SI (389/31)
| | | | SocioEcon1 = SI : SI (304/19)
| | | | SocioEcon25 = SI
| | | | SocioEcon1 = NO : SI (41/5)
| | | | SocioEcon1 = SI : SI (53/11)
| | | SocioEcon9 = SI
| | | | SocioEcon1 = NO
| | | | SocioEcon25 = NO : SI (117/1)
| | | | SocioEcon25 = SI : SI (31/0)
| | | | SocioEcon1 = SI
| | | | SocioEcon25 = NO : SI (51/3)
| | | | SocioEcon25 = SI : SI (16/0)
| | SocioEcon2 = SI
| | | SocioEcon1 = NO
| | | | SocioEcon9 = NO
| | | | SocioEcon25 = NO : SI (1442/86)
| | | | SocioEcon25 = SI : SI (118/5)
| | | | SocioEcon9 = SI
| | | | SocioEcon25 = NO : SI (388/20)
| | | | SocioEcon25 = SI : SI (70/0)
| | | SocioEcon1 = SI
| | | | SocioEcon25 = NO
| | | | SocioEcon9 = NO : SI (149/9)

```

| | | | SocioEcon9 = SI : SI (33/0)
| | | | SocioEcon25 = SI : SI (80/0)
SocioEcon8 = NO
| SocioEcon2 = NO
| | SocioEcon25 = NO
| | | SocioEcon22 = NO
| | | | SocioEcon9 = NO
| | | | SocioEcon1 = NO : NO (1214/93)
| | | | SocioEcon1 = SI : NO (80/10)
| | | | SocioEcon9 = SI
| | | | SocioEcon1 = NO : NO (29/12)
| | | | SocioEcon1 = SI : NO (4/0)
| | | SocioEcon22 = SI
| | | | SocioEcon9 = NO
| | | | SocioEcon1 = NO : NO (133/60)
| | | | SocioEcon1 = SI : SI (18/8)
| | | SocioEcon9 = SI : SI (23/5)
| | SocioEcon25 = SI
| | | SocioEcon1 = NO
| | | | SocioEcon9 = NO
| | | | SocioEcon22 = NO : NO (13/2)
| | | | SocioEcon22 = SI : NO (11/4)
| | | SocioEcon9 = SI : SI (7/0)
| | SocioEcon1 = SI : NO (4/0)
SocioEcon2 = SI
| SocioEcon22 = NO
| | SocioEcon9 = NO
| | | SocioEcon1 = NO
| | | | SocioEcon25 = NO : NO (251/90)
| | | | SocioEcon25 = SI : NO (5/0)
| | | | SocioEcon1 = SI
| | | | SocioEcon25 = NO : NO (38/4)
| | | | SocioEcon25 = SI : NO (3/0)
| | SocioEcon9 = SI : SI (8/0)
| SocioEcon22 = SI
| | SocioEcon25 = NO
| | | SocioEcon9 = NO
| | | | SocioEcon1 = NO : SI (156/42)
| | | | SocioEcon1 = SI : NO (27/13)
| | | SocioEcon9 = SI : SI (23/0)
| | SocioEcon25 = SI
| | | SocioEcon9 = NO : SI (4/0)
| | SocioEcon9 = SI : NO (1/0)

```