SISTEMA DE ALERTAS TEMPRANAS PARA LA PREVENCIÓN DE LA DESERCIÓN UNIVERSITARIA CON EL USO DE TÉCNICAS DE MACHINE LEARNING

Oscar Andrés Ramírez Avendaño

Marco Javier Peñaloza Pérez

Miguel Ernesto Velandia Feria

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

David Arango Londoño

Director David Arango

Jurado1 Gloria Álvarez

Jurado 2 Jaime Aguilar

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de datos.

Camilo Rocha

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 08 de julio de 2023

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 08 de julio de 2023

Autor: Oscar Andrés Ramírez Avendaño
Marco Javier Peñaloza Pérez
Miguel Ernesto Velandia Feria

Título del Trabajo de Grado: SISTEMA DE ALERTAS TEMPRANAS PARA LA PREVENCIÓN DE LA DESERCIÓN UNIVERSITARIA CON EL USO DE TÉCNICAS DE MACHINE LEARNING

Director: David Arango Londoño

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

Firma del director del Trabajo de Grado

Santiago de Cali, 29 de mayo de 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado SISTEMA DE ALERTAS TEMPRANAS PARA LA PREVENCIÓN DE LA DESERCIÓN UNIVERSITARIA CON EL USO DE TÉCNICAS DE MACHINE LEARNING, el cual será realizado por los estudiantes Oscar Andrés Ramírez Avendaño con código 8971502, Marco Javier Peñaloza Pérez con código 8972227 y Miguel Ernesto Velandia Feria con código 8972922.

Perteneciente al énfasis en N/A, bajo la dirección del profesor David Arango Londoño.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Miguel Velandia F
Miguel Ernesto Velandia Feria
C.C. 78.381.829 de San Andras de sotavento
Estudiante

David Arango Londoño
David Arango Londoño
C.C. 1.130.586.950 de Cali
Director

Oscar Andrés Ramírez Avendaño
C.C. 1.118.863.919 de Riohacha
Estudiante

Marco J. Peñaloza
Marco Javier Peñaloza Pérez
C.C. 1.065.616.547 de Valledupar
Estudiante

Sincelejo, 16 de noviembre de 2022

ACTA DE ENTREGA DE INFORMACIÓN

En la ciudad de Sincelejo, en las oficinas de la Corporación Universitaria Del Caribe – CECAR, para fines pertinentes, se hace entrega formal de los documentos relacionados a continuación:

Nombre Documento	Ubicación	Numero de Relacionados
Caracterización Sociodemográfica	Hoja 1	1.957
Caracterización Académica	Hoja 2	97.907
Incentivos	Hoja 3	1.544
Créditos Directos	Hoja 5	1.023

Con esta acta se cierra la presente entrega para todos los efectos a que haya lugar y que los procesos continúen con normalidad en coherencia con el resto de documentos aportados a la Institución.

Para Constancia de lo anterior se firma a los 16 días del mes de noviembre del año 2022.

Quien entrega

Andres Alberto Viloria Sequeda
C.C. 1102802886
Decano
FCBIA

Quien recibe

Miguel Ernesto Velandia Feria
C.C. 78381829
Investigador

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: "SISTEMA DE ALERTAS TEMPRANAS PARA LA PREVENCIÓN DE LA DESERCIÓN UNIVERSITARIA CON EL USO DE TÉCNICAS DE MACHINE LEARNING"

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Sector educativo
4. ESTUDIANTE (S): Oscar Andrés Ramírez Avendaño, Marco Javier Peñaloza Pérez y Miguel Ernesto Velandia Feria
5. CORREO ELECTRÓNICO: oaramireza@javerianacali.edu.co , mjpenalozap@javerianacali.edu.co y miguelvelandia29@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Carretera Troncal de Occidente Km. 1, Vía Corozal - Sincelejo, Colombia Cel: 3126516547
7. DIRECTOR: David Arango Londoño
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co
10. CO-DIRECTOR(ES) (Si aplica): N/A
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): CECAR
12. OTROS GRUPOS O EMPRESAS: N/A
13. PALABRAS CLAVE (al menos 5): deserción universitaria, machine learning, ciencia de datos, predicción y sistemas de alertas tempranas
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): N/A
15. FECHA DE INICIO (Desarrollo del proyecto): 13/06/2022
16. RESUMEN (máximo 400 palabras).

La deserción escolar universitaria es un problema global que tiene un impacto negativo en el progreso social y científico de un país o región. Las Instituciones de Educación Superior (IES) tienen la responsabilidad de prevenir e intervenir en esta problemática. En este sentido, este estudio presenta un marco conceptual de la deserción universitaria, basado en investigaciones que abordan tanto enfoques cualitativos como cuantitativos en el uso de la ciencia de datos. A continuación, se realiza un análisis exploratorio descriptivo de los datos de deserción correspondientes a los periodos de 2019A-2022B. Este análisis se enfoca en comprender y examinar el fenómeno de la deserción en la Facultad de Ciencias Básicas e Ingenierías de la Corporación Universitaria del Caribe (Cecar).

Finalmente, se entrenaron varios modelos de machine learning, como la regresión logística, las máquinas de soporte vectorial, los bosques aleatorios de decisión y las redes neuronales simples. Estos modelos permiten predecir y emitir alertas sobre los riesgos de deserción en los programas de ingeniería de sistemas e industrial. Este logro se lleva a cabo mediante el desarrollo y despliegue de un modelo a través de una API y una interfaz gráfica que integra el análisis exploratorio y el modelo predictivo. De esta manera, utilizando los datos de entrada, el sistema puede predecir la probabilidad de deserción para nuevos estudiantes, configurando un sistema de alertas tempranas. Este sistema de alertas se convierte en un apoyo crucial para la toma de decisiones, ya que contribuye a la comprensión y mitigación de la deserción universitaria, así como a la promoción de políticas institucionales que buscan la permanencia de los estudiantes.

SISTEMA DE ALERTAS TEMPRANAS PARA LA PREVENCIÓN DE LA DESERCIÓN UNIVERSITARIA CON EL USO DE TÉCNICAS DE MACHINE LEARNING

Oscar Andrés Ramírez Avendaño

Marco Javier Peñaloza Pérez

Miguel Ernesto Velandia Feria

Proyecto Aplicado para optar al título deMagister en

Ciencia de Datos

Director

David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, MAYO DE 2023

RESUMEN

La deserción escolar universitaria es un problema global que tiene un impacto negativo en el progreso social y científico de un país o región. Las Instituciones de Educación Superior (IES) tienen la responsabilidad de prevenir e intervenir en esta problemática. En este sentido, este estudio presenta un marco conceptual de la deserción universitaria, basado en investigaciones que abordan tanto enfoques cualitativos como cuantitativos en el uso de la ciencia de datos. A continuación, se realiza un análisis exploratorio descriptivo de los datos de deserción correspondientes a los periodos de 2019A-2022B. Este análisis se enfoca en comprender y examinar el fenómeno de la deserción en la Facultad de Ciencias Básicas e Ingenierías de la Corporación Universitaria del Caribe Cecar.

Finalmente, se entrenaron varios modelos de aprendizaje automático, como la regresión logística, las máquinas de soporte vectorial, los bosques aleatorios de decisión y las redes neuronales simples. Estos modelos permiten predecir y emitir alertas sobre los riesgos de deserción en los programas de ingeniería de sistemas e industrial. Este logro se lleva a cabo mediante el desarrollo y despliegue de un modelo a través de una API y una interfaz gráfica que integra el análisis exploratorio y el modelo predictivo. De esta manera, utilizando los datos de entrada, el sistema puede predecir la probabilidad de deserción para nuevos estudiantes, configurando un sistema de alertas tempranas. Este sistema de alertas se convierte en un apoyo crucial para la toma de decisiones, ya que contribuye a la comprensión y mitigación de la deserción universitaria, así como a la promoción de políticas institucionales que buscan la permanencia de los estudiantes.

TABLA DE CONTENIDO

INTRODUCCIÓN 10
1. DEFINICIÓN DEL PROBLEMA 11
1.1 PLANTEAMIENTO DEL PROBLEMA 11
1.1 FORMULACIÓN DEL PROBLEMA 12
2. OBJETIVOS DEL PROYECTO 14
2.1 OBJETIVO GENERAL 14
2.2 OBJETIVOS ESPECÍFICOS 14
2.3 RESULTADOS ESPERADOS 14
3. ALCANCE 15
4. JUSTIFICACIÓN 16
5. MARCO DE REFERENCIA 18
5.1 MARCO TEÓRICO 18
5.1.1 Deserción escolar 18
5.1.1.1 Deserción escolar en la educación superior 19
5.1.1.2 Factores claves para la deserción universitaria 20
5.1.2 Machine Learning (ML) 23
51.2.1 Categorías y técnicas de Machine Learning 24
5.1.2.1.1 Regresión Logística 26
5.1.2.1.2 Regresión Logística aplicada a la predicción de la deserción escolar 28
5.1.2.1.3 Redes Neuronales 29
5.1.2.1.4 Redes neuronales aplicadas a la predicción de la deserción escolar 30
5.1.2.1.5 Arboles De Decisión 31
5.1.2.1.6 Arboles de decisión aplicados a la predicción de la deserción escolar 32
5.1.2.1.7 Máquina de Soporte Vectorial (SVM) 33
5.1.2.1.8 Bosques aleatorios – Random Forest (RF) 34
5.1.2.1.9 Bosques aleatorios aplicados a la predicción de la deserción escolar 35
5.1.3 Minería de datos (Data Mining) 37
5.1.3.1 Data Ware Housing DWH 37
5.1.4 Metodología de proyecto de ciencia de datos. (CRISP-DM) 41
5.1.4.1 Entendimiento del negocio 43
5.1.4.2 Entendimiento de los datos 43

5.1.4.3 Preparación de los datos....43
5.1.4.4 Modelamiento....44
5.1.4.5 Evaluación....44
5.1.4.6 Despliegue....44
5.1.5 Sistemas de alertas tempranas SAT....45
5.2 ANTECEDENTES....46
6. METODOLOGÍA....52
7. DESARROLLO....55
7.1 Planeación....55
7.1.1 Caracterización del conjunto de datos....55
7.1.2 Modelo bidimensional de datos....57
7.2 Desarrollo del modelo....58
7.2.1 Preparación y limpieza de datos....58
7.2.2 Análisis exploratorio de datos....60
7.2.3 Selección de variables para el modelo....76
7.3 Implementación del modelo....81
7.3.1 División de los datos en entrenamiento....81
7.3.2 Preprocesado de los datos....82
7.3.3 Modelo de regresión logística....83
7.3.4 Modelo de máquina de soporte vectorial (SVM)....83
7.3.5 Modelo de bosques aleatorios (Random Forest)....87
7.3.6 Modelo de redes neuronales simple (NNET)....91
7.3 Comparación y validación de los modelos....93
7.4 Arquitectura del sistema de alertas tempranas....95
CONCLUSIONES....99
Referencias....102

LISTA DE FIGURAS

Ilustración 1 Requisitos básicos de un sistema de DWH 38
Ilustración 2. Modelo de arquitectura de DWH 39
Ilustración 3 El ciclo de vida del proyecto de minería de datos 42
Ilustración 4. Fases del proyecto de ciencia de datos SATDU 52
Ilustración 5. Modelo bidimensional del análisis de la deserción 58
Ilustración 6. Porcentaje de datos faltantes por variable 59
Ilustración 7 .Resumen del modelo regresión logística 35 variables 78
Ilustración 8. Resumen del modelo regresión logística 7 variables por importancia 79
Ilustración 9. Distribución de valores de deserción datos de entrenamiento 82
Ilustración 10. Resumen del modelo de regresión logística 8 predictoras de la clase deserción 83
Ilustración 11. SVMLienal - Valores de validación (Accuracy y Kappa) obtenidos en cada partición y repetición. 84
Ilustración 12 Resumen del modelo con la validación cruzada SVMLienal 85
Ilustración 13. SVM Lineal - Matriz de confusión y error de clasificación 86
Ilustración 14 Resultado Máquinas de vectores soporte con núcleo de función de base radial. 87
Ilustración 15. Random Forest - Valores de validación (Accuracy y Kappa) obtenidos en cada partición y repetición. 88
Ilustración 16 Random Forest - Resumen del mejor modelo 89
Ilustración 17 Resumen del modelo con la validación cruzada Bosques aleatorios 89
Ilustración 18. RandomForest - Matriz de confusión y error de clasificación 90
Ilustración 19. NNET - Valores de validación (Accuracy y Kappa) obtenidos en cada partición y repetición. 91
Ilustración 20. NNET- Matriz de confusión y error de clasificación 92
Ilustración 21. Resultado de la prueba de Friedman 94
Ilustración 22 Resultado de la prueba de Wilcoxon 94
Ilustración 23 Arquitectura candidata del sistema de alertas tempranas Azure con Synapses 96

LISTA DE TABLAS

Tabla 1 Técnicas de Machine Learning 25
Tabla 2 Diferencias entre las bases de datos OLAP y OLTP 40
Tabla 3 Principales estudios de predicción de la deserción usando ciencia de datos 46
Tabla 4 Detalle fuente de datos caracterización socio demográfica 55
Tabla 5 Detalle fuente de datos caracterización académica 56
Tabla 6 Detalle fuente de datos caracterización Incentivos 56
Tabla 7 Detalle fuente de datos caracterización créditos directos 56
Tabla 8 Detalle fuente de datos caracterización deserción 57
Tabla 9 Número de asignaturas aprobadas por desertores 74
Tabla 10 Métricas de accuracy y ecuaciones por cada modelo de regresión 78
Tabla 11 Resumen de los datos estandarizados y binarizados 82
Tabla 12 Comprobación de variables con varianza próxima a cero 82
Tabla 13 Valor promedio de accuracy y kappa por modelo 93
Tabla 14 Métrica accuracy de los modelos para los datos de entrenamiento y prueba 95

LISTA DE GRÁFICAS

Gráfica 1 Porcentaje de valores de deserción 60
Gráfica 2 Porcentaje de estudiantes por estrato socioeconómico 61
Gráfica 3 Porcentaje de como pagan los estudios los estudiantes 61
Gráfica 4 Porcentaje de estado civil de los estudiantes 62
Gráfica 5 Porcentaje de ocupaciones de los estudiantes 62
Gráfica 6 Porcentaje de tipo de vivienda de los estudiantes 63
Gráfica 7 Porcentaje de motivación del retiro por factores económicos 64
Gráfica 8 Porcentaje de nivel educativo de los acudientes que pagan los estudios 65

Gráfica 9 Promedio de edad de quien paga sus estudios 65
Gráfica 10 Promedio de egresos mensuales del núcleo familiar 66
Gráfica 11 Total de incentivos económicos por porcentaje 66
Gráfica 12 Porcentaje de estudiantes por población a la que pertenece 67
Gráfica 13 Distribución de promedio general por porcentaje de estudiantes 67
Gráfica 14 Porcentaje de deserción por estrato socioeconómico 68
Gráfica 15 Porcentaje total de deserción por estrato socioeconómico 69
Gráfica 16 Porcentaje total de deserción por forma de pago de los estudios 70
Gráfica 17 Porcentaje total de deserción por estado civil 70
Gráfica 18 Porcentaje total de deserción por nivel educativo del que paga los estudios 71
Gráfica 19 Porcentaje total de deserción por edades del que paga los estudios 71
Gráfica 20 Porcentaje total de deserción por población de pertenencia de los estudiantes 72
Gráfica 21 Porcentaje total de deserción por situación laboral 73
Gráfica 22 Diagrama de caja número de asignaturas aprobadas vs deserción 74
Gráfica 23 Diagrama de caja promedio general vs deserción 75
Gráfica 24 Diagrama de caja número de asignaturas reprobadas vs deserción 75
Gráfica 25 Porcentaje de desertores por cantidad de incentivos económicos 76
Gráfica 26 Correlación entre variables numéricas 77
Gráfica 27 Lista de variables ordenadas de mejor a peor desempeño predictor de la deserción 80
Gráfica 28 SVMLineal-Evolución de los modelos según el valor de los hiperparámetros 85
Gráfica 29 Random Forest-Evolución de los modelos según el valor de los hiperparámetros 88
Gráfica 30 NNET-Evolución de los modelos según el valor de los hiperparámetros 92
Gráfica 31 Validación: Accuracy medio repeated-CV modelos ordenados por media 93
Gráfica 32 Métrica accuracy de entrenamiento y prueba por modelo 95

INTRODUCCIÓN

Las instituciones de educación superior tienen una responsabilidad apremiante en el desarrollo de la sociedad, constituida en la tarea de la graduación de los estudiantes en términos de calidad, oportunidad e inclusión. Esta tarea enfrenta un problema importante: la deserción universitaria, lo cual impacta negativamente en el progreso de los países en los ámbitos sociales y científicos [1]. La problemática constituye uno de los mayores retos que tienen las universidades en la actualidad, para prevenir e intervenir en el desarrollo de mecanismos óptimos de retención estudiantil. La naturaleza de este problema es diversa y compleja, así como también las repercusiones que trae a nivel personal, institucional y nacional.

Para afrontar la necesidad mencionada, se diseñó y aplicó una investigación sobre el fenómeno de la deserción escolar, apoyada mediante las técnicas de ciencia de datos. El proyecto se desarrolló en la Corporación Universitaria del Caribe - CECAR y se creó un sistema de alertas tempranas, utilizando técnicas de ciencia de datos, específicamente con aprendizaje automático y estadística. El proyecto constó de 6 fases, las tres primeras permitieron el entendimiento del negocio, los datos y la preparación de los mismos, mientras que las tres restantes abordaron el modelado, la evaluación y el despliegue del sistema de alertas tempranas.

Como resultado, se obtuvo la definición descriptiva de los factores históricos de la deserción universitaria en los programas de ingeniería de sistemas e industrial. Además, se desarrolló un modelo predictivo de la probabilidad de deserción de los estudiantes y se implementó un sistema de alertas tempranas que integra la visualización de la información a través de análisis gráficos y el uso del modelo predictivo. Este sistema permite notificar a las directivas para que desarrollen acciones institucionales oportunas con el objetivo de fomentar la permanencia de los estudiantes.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

La deserción escolar es un problema que afecta a todas las regiones del mundo y estos patrones globales muestran que entre más temprano se produce esa deserción escolar y se tengan menos años de estudio el impacto negativo en los estudiantes es mayor, según el programa piloto para la prevención de la deserción escolar (SDPP) regional en Asia y Oriente Medio de USAID [2].

Esta condición de abandono de la escolaridad genera impactos negativos en los estudiantes, los cuales van desde la reducción significativa de los ingresos de por vida, la dificultad de conseguir empleo, una mayor probabilidad de involucrarse en conductas antisociales, incluyendo la criminalidad y el encarcelamiento. Así mismo, se incrementa la tasa de abuso de sustancias ilícitas, se deteriora la salud y se observa una mayor incidencia en conductas agresivas y violentas.

A nivel internacional, diferentes investigaciones, indican que "la deserción universitaria en los países pertenecientes a la OCDE alcanza un 31% [3]. Para el caso de las instituciones universitaria europeas, la tasa de deserción varía de un 20% a un 55%. De acuerdo con lo consultado, en el ámbitoIberoamericano el 66 % de los estudiantes permanecieron activos y el 33 % abandonaron la carrera [4].

En Colombia, particularmente, la deserción es un fenómeno de gran relevancia. El Ministerio de Educación Nacional (MEN), como entidad encargada de suministrar datos precisos para hacer seguimiento a la deserción escolar en todos los niveles educativos, presentó en el año 2018 una nueva versión de su sistema, conocido como SPADIES 3.0. Esta nueva versión íntegra el SNIES con el objetivo de mejorar la precisión y calidad de la información, así como también estudiar la movilidad de los estudiantes entre programas y carreras universitarias. Al comparar los dos sistemas, se evidencia que las tasas de deserción escolar han variado del 9,89% al 8,25%, siendo esta última correspondiente al año 2019. Estas cifras reflejan una preocupante situación que

afecta a los sistemas educativos y tiene un impacto negativo en la permanencia y graduación de los estudiantes. Según estadísticas del Ministerio de Educación Nacional, aproximadamente la mitad de los estudiantes que ingresan a una institución de educación superior no logra culminar su ciclo académico y obtener la graduación [3].

De acuerdo al diagnóstico realizado, tomando como referente un estudio sobre tasa y tiempo promedio de la graduación de los estudiantes del programa ingeniería de sistemas de la Corporación Universitaria del Caribe CECAR, los estudiantes de este programa presentan un índice de deserción a 2019 primer semestre de 9,41% según reporte, la tasa de graduación para el programa correspondiente es del 37,8%, siendo un punto de interés indagar las motivaciones de los estudiantes por abandonar el programa de formación, pues no alcanza por lo menos un promedio por encima del 50% [5].

Finalmente, es de suma importancia reflexionar acerca de la responsabilidad que tienen las instituciones de educación superior sobre las estrategias que deben adoptar para plantear soluciones que permitan conocer, analizar, abordar e intervenir el fenómeno, con el fin de plantear soluciones contundentes que remedien esta problemática de carácter educativo y social y coadyuvar en el proceso de graduación de sus estudiantes. Dada la relevancia que tiene la problemática de la deserción escolar se hace necesario que las instituciones educativas identifiquen con precisión, sensibilidad y detección preventiva de los patrones de deserción en todos los niveles educativos.

1.1 FORMULACIÓN DEL PROBLEMA

El presente trabajo tiene como objetivo desarrollar una herramienta digital basada en algoritmos de aprendizaje automático para pronosticar y prevenir casos de deserción estudiantil en estudios universitarios. Se emplearon eventos históricos y diversas variables de tipo social, académico, personal y laboral. Estas variables fueron utilizadas para ajustar diferentes modelos capaces de predecir las probabilidades de deserción de cada estudiante. Posteriormente, con esta

información, se busca alertar y recomendar medidas preventivas tempranas para la población estudiantil.

Pregunta de investigación:

¿La implementación de un sistema de alertas tempranas de deserción universitaria permitirá la identificación de los factores determinantes para prevenir la deserción estudiantil de manera efectiva?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

- Implementar un sistema de alertas tempranas de deserción universitaria, con el uso de técnicas de minería de datos para la identificación de los factores determinantes de deserción universitaria.

2.2 OBJETIVOS ESPECÍFICOS

- Identificar las variables relevantes que inciden en la deserción estudiantil a través del análisis exploratorio de datos.
- Desarrollar un modelo de aprendizaje automático para la predicción de la deserción y realizarvalidación cruzada para el desempeño del modelo.
- Implementar un sistema de información para la administración y gestión de la información que arroja el modelo predictivo.

2.3 RESULTADOS ESPERADOS

Los resultados esperados son los siguientes:

- Desarrollo de un análisis estadístico descriptivo del histórico de deserción en la universidad, proporcionando una visión completa de los datos.
- Creación de un modelo predictivo de la deserción universitaria mediante la utilización de técnicas estadísticas y de aprendizaje automático, con el objetivo de prever con precisión la probabilidad de deserción de los estudiantes.
- Implementación de un sistema de alertas tempranas de deserción universitaria basado en el modelo predictivo desarrollado, que permita tomar decisiones informadas para mitigar el fenómeno de la deserción.

3. ALCANCE

El presente proyecto se desarrollaró en la Corporación Universitaria del Caribe, CECAR en la facultad de Ciencias Básicas, Ingenierías y Arquitectura para los programas de Ingeniería de Sistemas e Industrial, tomando como fuentes de datos para el análisis de información la caracterización del estudiante (prueba de ingreso), pruebas saber 11, información académica y evaluación docente.

Los resultados esperados son los siguientes:

- Desarrollo de un análisis estadístico descriptivo del histórico de deserción en la universidad, proporcionando una visión completa de los datos.
- Creación de un modelo predictivo de la deserción universitaria mediante la utilización de técnicas estadísticas y de aprendizaje automático, con el objetivo de prever con precisión la probabilidad de deserción de los estudiantes.
- Implementación de un sistema de alertas tempranas de deserción universitaria basado en el modelo predictivo desarrollado, que permita tomar decisiones informadas para mitigar el fenómeno de la deserción.

El modelo se enfocó mayoritariamente en lo cuantitativo, el análisis cualitativo sólo se enfocará en la descripción de las variables a partir del análisis exploratorio inicial, quedando para futuros proyectos ampliar el estudio cualitativo hacia modelos comportamentales, rasgos de personalidad y dinámicas del entorno estudiantil.

Restricciones del proyecto:

- Acceso de los datos bajo la política y el uso de proyección de información.
- Desarrollo de la plataforma SATDU basado en el modelo de interoperabilidad de sistemas de información de la universidad.

4. JUSTIFICACIÓN

La implementación de la educación es un desafío mundial la cual debe responder al pluralismo y las condiciones socioeconómicas donde se desarrolla y se hace urgente atender la exclusión de oportunidades educativas a los estudiantes más vulnerables, donde cerca del 70% de estos se retiran en los estratos 1 y 2 sumado a los demás factores socioeconómicos que hacen parte de los determinantes de la deserción universitaria [6].

Las universidades como instituciones activas en el conocimiento y con una responsabilidad social que propenda garantizar la educación para todos como derecho fundamental y que permita que los estudiantes permanezcan, finalicen sus estudios y logren una participación significativa en los ámbitos económicos, políticos, académicos y culturales.

El Ministerio de Educación colombiano, indica que es un "problema dual, ya que no solo es importante conocer sus causas, sino la manera de disminuirla, lo cual significa aumentar la retención estudiantil" [7] , característica que eleva la complejidad y se hace difícil de combatir a pesar de que es un fenómeno altamente estudiado.

Con la implementación del sistema de alertas tempranas para la deserción universitaria – SATDU, se aborda el problema desde los dos enfoques desde la perspectiva de la analítica de datos y la de negocios inteligentes, esto conlleva a la toma de decisiones oportunas basadas en las inferencias, patrones y evidencia de los datos, logrando así hacer frente a la dualidad del problema, porque permite la identificación de las causales del abandono universitario y el planteamiento de recomendaciones para la toma de decisiones efectivas que mitiguen las tasas de abandono universitario en CECAR.

El presente trabajo de investigación se enfoca en la comprensión del fenómeno de la deserción universitaria a partir de una visión de analítica de datos, la cual conlleva a la comprensión de fenómeno desde lo cuantitativo, cualitativo y estratégico, que conduzcan a la permanecía estudiantil desde las realidades de exclusión y participación en la vida educativa universitaria.

Enfoque cuantitativo:

- Relaciones de causalidad, predicción y explicación del fenómeno
- Construcción de cifras e indicadores e históricos de la deserción
- Cálculo de la probabilidad de que un estudiante abandone dado el conjunto de variables explicativas del fenómeno.

Enfoque cualitativo:

- Análisis del fenómeno socio demográfico

Enfoque estratégico para el fortalecimiento de los programas de retención [1]**:**

- Programas de asistencia educativa
- Programas de asistencia financiera y comunitaria
- Programa de estímulo y fomento
- Programas de guía y consejería psicológica
- Programas de respaldo a profesores

Finalmente se desarrolló un componente tecnológico donde se podrán visualizar las variables de estudio, las alertas generadas y las recomendaciones, que tiene como finalidad el desarrollo de un sistema inteligente para la toma de decisiones basado en técnicas de aprendizaje automático (machine learning), lo que permitirá la toma de decisiones y el desarrollo de nuevas estrategias, la detecciónde riesgos y amenazas futuras, la identificación de problemas en tiempo real y la reducción de costos y tiempo en la gestión del abandono estudiantil.

5. MARCO DE REFERENCIA

5.1 MARCO TEÓRICO

5.1.1 Deserción escolar

La deserción escolar se refiere al abandono de las actividades académicas por parte de un individuo debido a diversas circunstancias, como factores económicos, políticos, sociales, familiares, ambientales o de salud. Este fenómeno suele estar influenciado tanto por decisiones personales como por una serie de factores que obstaculizan la continuidad en la formación de los estudiantes. Como resultado, los estudiantes pueden mostrar un bajo interés o falta de motivación para continuar con su proceso de aprendizaje.

Uno de los factores que tiene mayor incidencia en la deserción escolar son los problemas socioeconómicos de las familias menos favorecidas, que ven necesario que los niños y jóvenes dejen sus estudios escolares por perseguir ofertas laborales que les den un sustento a sus necesidades de alimentación y de mejorar la calidad de vida, es frecuente ver casos de deserción cuando se labora mientras se estudia, pues las dificultades aumentan y termina ganando el abandono de los estudios. [8]

En el ámbito social, la deserción escolar suele manifestarse en situaciones en las que los jóvenes se ven forzados a abandonar sus estudios debido a dificultades legales, actividades delictivas o comportamientos destructivos que perjudican a la sociedad y provocan la interrupción de su educación. Estos problemas a menudo están asociados con circunstancias familiares y personales que llevan a la persona a tomar decisiones equivocadas, muchas veces derivadas de la falta de atención por parte de sus padres, familiares o responsables. Esto provoca que busquen compañía en individuos desconocidos, quienes les enseñan valores contrarios y formas inapropiadas de comportarse.

A nivel universitario, la deserción se manifiesta cuando los estudiantes presentan un bajo rendimiento académico, lo cual se traduce en una falta de motivación y atención hacia las directrices impartidas por los docentes. Este fenómeno también puede estar asociado con problemas de asistencia a clases y puede resultar en la interrupción de la trayectoria educativa. Es fundamental que los docentes estén atentos a las señales indicativas de un posible abandono de los estudios, ofreciendo alternativas que, en colaboración con las instituciones educativas, promuevan la retención de los estudiantes universitarios.

Finalmente, la pandemia de Covid-19 ha ejercido una notable influencia en numerosos casos de deserción escolar. La escasez de recursos económicos para cubrir necesidades básicas, como alimentación y servicios, ha llevado a muchos jóvenes a abandonar sus estudios y buscar empleos como una alternativa para generar ingresos. [10]

5.1.1.1 Deserción escolar en la educación superior

La deserción, definida por el Ministerio de Educación Nacional [7], como el abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno [3].

Es una problemática de gran relevancia a nivel público que afecta tanto a Colombia como al resto del mundo. Dicha situación socava los esfuerzos colectivos para elevar el nivel educativo de los estudiantes, buscando mejorar su competitividad y lograr una integración exitosa en la sociedad del conocimiento. Este desafío debe ser abordado de manera integral y colaborativa, involucrando no solo a los estudiantes, sino también a las diversas instituciones de educación superior. Es crucial que estas instituciones se esfuercen por comprender y atender las dificultades que conllevan a la deserción escolar, brindando apoyo y recursos necesarios para superarlas. Es fundamental recopilar información precisa y actualizada para comprender a fondo las causas y efectos de este fenómeno.

5.1.1.2 Factores claves para la deserción universitaria

Según diversos análisis realizados, se ha observado que la deserción en instituciones de educación superior es el resultado de la combinación de varios factores, entre los cuales se destacan aspectos sociales, políticos, institucionales, personales y relacionados con el proceso de aprendizaje. Los estudios y teorías que han examinado esta problemática se han centrado en evaluar las características académicas, socioeconómicas, psicológicas y familiares de los estudiantes al ingresar a la educación superior. Algunos investigadores señalan que las principales causas de la deserción están relacionadas con las deficiencias en la preparación académica previa de los estudiantes, lo cual dificulta su integración social en el nuevo nivel educativo. Además, factores como la situación financiera, los problemas familiares y el bajo rendimiento académico también influyen en la decisión de abandonar el sistema educativo.

Entre las variables sociales y económicas estudiadas, se han identificado la desigualdad social y económica como elementos relevantes. Asimismo, se destaca la responsabilidad que recae en los programas académicos y su falta de actualización como factores que contribuyen a la deserción. Sobre los factores económicos, cabe destacar que es uno de los factores más decisivos y que representa un peso significativo, en muchos casos afecta la percepción del estudiante acerca de su capacidad o incapacidad para solventar los costos asociados a los estudios universitarios, lo cual puede ser atendido mediante créditos o becas, dicho factor está relacionado con la pertenencia a niveles socioeconómicos bajos, falta de dinero para manutención y sostenimiento, pocas posibilidades para laborar paralelamente al estudio, o cualquier problema financiero que no permita hacer frente a los gastos debidos al desempeño académico requerido para concluir una carrera.

En este sentido, los problemas relacionados con la deserción escolar son causados por factores Inter sistémicos e intrasistémicos: los primeros se refieren a la oferta educativa, la desigualdad en la calidad de los servicios educativos y los mecanismos de acceso, asociados a la asignación de

plantel, modalidad y turno; los factores intrasistémicos se vinculan con las prácticas pedagógicas inadecuadas, formación docente limitada y condiciones laborales precarias, infraestructura y equipamiento insuficiente, incompatibilidad entre la cultura juvenil y escolar, currículo poco pertinente, gestión escolar deficiente, y participación limitada de padres y estudiantes en la escuela [3].

En América Latina, otro factor comúnmente asociado a la deserción escolar es el embarazo durante la adolescencia. Esta situación no solo afecta al estudiante que abandona, sino también a su entorno, ya que los futuros padres se ven obligados a asumir responsabilidades familiares que pueden interferir con su desarrollo académico normal.

Es importante resaltar que las mujeres tienen una menor probabilidad de completar sus estudios en estas circunstancias. Cuando un estudiante abandona sus estudios, se crea un vacío que podría haber sido ocupado por otro estudiante que, de persistir y completar su educación, contribuiría al sistema. Esto genera una pérdida económica tanto para el individuo como para la institución educativa, al crear inestabilidad en sus recursos.

Otra causa muy común en América Latina tiene que ver con el embarazo en la adolescencia, lo que afecta tanto al desertor mismo como a su entorno, obligando a los futuros padres a tener que asumir responsabilidades familiares que pueden afectar el desarrollo normal de sus actividades académicas. En este contexto, vale destacar que son las mujeres quienes tienen menos probabilidad de concluir sus estudios. Un estudiante que abandona sus estudios crea un espacio que pudo ser ocupado por otro estudiante que, si persistiría y concluiría en su educación, causando una pérdida económica también para la institución.

Asimismo, la deserción escolar puede estar relacionada con la predisposición o falta de motivación del estudiante hacia las asignaturas y programas académicos. En este sentido, es crucial el papel desempeñado tanto por las instituciones educativas con sus programas de estudio, como por los docentes con sus enfoques pedagógicos.

Es cierto que los problemas de actitud de los estudiantes pueden originarse en su desarrollo personal, pero también es verdad que la trayectoria escolar influye en la forma en que enfrentan el currículo académico. En otras palabras, todos los miembros de la comunidad educativa deben mostrar interés por los desafíos que enfrentan los estudiantes y abordarlos a través de diversas actividades que fomenten la integración.

Si logramos generar un cambio positivo en la mentalidad de los alumnos y tomamos en cuenta su desarrollo individual, ellos mismos contarán con las herramientas necesarias para afrontar no sólo los retos académicos, sino también los personales. A nivel institucional, el desafío consiste en prevenir la deserción escolar, lo cual requiere un análisis detallado de los diferentes factores que pueden contribuir a ella. [11]

En esta lógica y de acuerdo con los enfoques teóricos a nivel mundial, las causas de la deserción enla educación superior en Colombia que retoma el [3] son:

SOCIOECONÓMICAS	ACADÉMICAS	INDIVIDUALES	INSTITUCIONALES
✓ Miedo al endeudamiento por parte de los estudiantes o de sus padres.	✓ Falta de preparación desde la educación media en competencias generales	✓ La edad de inicio de los alumnos incide positivamente cuando son menores de edad.	✓ Falta de preparación desde la educación media en competencias generales
✓ Subestimar los costos de estudiar un programa de pregrado.	✓ Poca orientación profesional y vocacional antes del ingreso a la universidad	✓ Las personas casadas debido a la menor disponibilidad de tiempo.	✓ Poca orientación profesional y vocacional antes del ingreso a la universidad
✓ Pertenecer a estrato bajo.	✓ Bajo rendimiento académico	✓ Presiones familiares y sociales Costos monetarios y de tiempo que se deben afrontar al estudiar en otra ciudad.	✓ Bajo rendimiento académico
✓ Bajos ingresos familiares y desempleo de los padres.	✓ Baja calidad del programa al que se accede Métodos de estudio y metodologías de aprendizaje obsoletas	✓ Calamidad y problemas de salud Discriminación social por razones de orientación sexual o raza.	✓ Baja calidad del programa al que se accede Métodos de estudio y metodologías de aprendizaje obsoletas
✓ Dependencia económica de sí mismo.	✓ Insatisfacción con el programa	✓ Incompatibilidad horaria con actividades extracurriculares	✓ Insatisfacción con el programa
✓ Nivel educativo bajo de los padres (ninguno o primaria).	✓ Estrés por la carga académica	✓ Expectativas no satisfechas o no le encuentran un futuro a loque están estudiando.	✓ Estrés por la carga académica

Fuente: *SPADIES [3]*

5.1.2 Machine Learning (ML)

El aprendizaje automático (ML) es una rama de la inteligencia artificial (IA) y la informática que se centra en el uso de datos y algoritmos. Su aplicación práctica consiste en imitar la forma en que los humanos aprenden y mejorar gradualmente su precisión [12]. Otros autores, como [13], lo definen como un conjunto de métodos que pueden detectar automáticamente patrones en los datos y, utilizando esos patrones, descubrir patrones futuros en los datos para respaldar la toma de decisiones en las organizaciones.

La UC Berkeley [12] divide el sistema de aprendizaje de un algoritmo de aprendizaje automático en tres partes principales:

- **Un proceso de decisión:** En general, los algoritmos de aprendizaje automático se utilizan para hacer una predicción o clasificación. Basándose en unos datos de entrada, que pueden estar etiquetados o sin etiquetar, su algoritmo producirá una estimación sobre un patrón en los datos.
- **Una función de error:** Una función de error sirve para evaluar la predicción del modelo. Si hay ejemplos conocidos, una función de error puede hacer una comparación para evaluar la precisióndel modelo.
- **Un proceso de optimización del modelo:** Si el modelo puede ajustarse mejor a los puntos de datos del conjunto de entrenamiento, se ajustan los pesos para reducir la discrepancia entre el ejemplo conocido y la estimación del modelo. El algoritmo repetirá este proceso de evaluación y optimización, actualizando de manera autónoma los pesos hasta llegar a un umbral de precisión establecido. Este enfoque iterativo permite que el modelo mejore su rendimiento a medida que se le proporciona más información y datos de entrenamiento. Con cada iteración, se busca una mayor precisión y capacidad de generalización del modelo, lo cual resulta fundamental para su utilidad y aplicabilidad en diferentes escenarios y problemas.

51.2.1 Categorías y técnicas de Machine Learning

El aprendizaje automático suele dividirse en dos tipos principales, aprendizaje predictivo o supervisado y el aprendizaje no supervisado.

Aprendizaje supervisado: En el contexto del aprendizaje predictivo o supervisado, el propósito radica en adquirir el conocimiento sobre la relación entre las entradas, representadas como "X", y las salidas, representadas como "Y", a partir de un conjunto de pares de entrada-salida. El objetivo es encontrar patrones y regularidades que permitan predecir con precisión las salidas correspondientes a nuevas entradas.

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

D se denomina conjunto de entrenamiento y **N** es el número de ejemplos de entrenamiento.

Con el conjunto de datos etiquetados se entrena el algoritmo para clasificar los datos o predecir resultados con precisión, dado que a medida que los datos se introducen en el modelo este ajusta sus ponderaciones hasta lograr un nivel de ajuste adecuado. [13]

Aprendizaje no supervisado: En el aprendizaje descriptivo o no supervisado en este caso, solo se dan entradas, el objetivo es encontrar patrones en los datos lo cual se denomina descubrimiento de conocimiento, dado que se trata de un problema menos definido y no se cuenta con la definición de los patrones a buscar y no existe una métrica de error a utilizarlo que lo diferencia del aprendizaje supervisado. [13]

$$D = \{x_i\}_{i=1}^{N}$$

Este tipo de aprendizaje es útil para analizar y agrupar conjuntos de datos no etiquetados, permitiendo descubrir agrupaciones, similitudes y diferencias en la información sin la intervención humana, su aplicabilidad ideal para la realización de análisis exploratorio de datos, segmentación de clientes y reconocimiento de imágenes y patrones.

A continuación, se describen algunas de las técnicas más usadas en el aprendizaje automático, **Tabla** .

Tabla 1. Técnicas de Aprendizaje Automático

Técnica de ML	**Descripción**
Redes neuronales (con diferentes parámetros)	"Una red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa lainformación: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas". [14]
Support Vector Machine	"SVM funciona asignando datos a un espacio de características de alta dimensión para que los puntos de datos se puedan clasificar, incluso cuando los datos no se pueden separar linealmente." [14]
Gradient Boosting	"Es un método de aprendizaje conjunto que combina un conjunto de aprendices débiles en un aprendiz fuerte para minimizar los errores de entrenamiento. En el boosting, se selecciona una muestra aleatoria de datos, se le aplica un modelo y se entrena secuencialmente, es decir, cada modelo intenta compensar las debilidades de su predecesor [15]

Random Forest	"El bosque aleatorio es un algoritmo de aprendizaje automático comúnmente utilizado y registrado por Leo Breiman y Adele Cutler, que combina la salida de múltiples árboles de decisión para llegar a un único resultado." [15]
Regresión logística	"Es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo" [14]

El aprendizaje automático (ML) se cataloga como un componente importante de la ciencia de los datos, el cual por medio del uso de métodos estadísticos permite el entrenamiento de los algoritmos y estos a su vez generan resultados como clasificaciones o predicciones, descubrimiento de ideas clave dentro de los proyectos de minería de datos. Estos nuevos conocimientos impulsan posteriormente la toma de decisiones dentro de las organizaciones, lo que impacta positivamente en la generación de valor, la mejora de los indicadores claves de rendimiento y en general el crecimiento de las compañías.

5.1.2.1.1 Regresión Logística

Es una técnica de modelado estadístico ampliamente utilizada en diferentes campos de investigación, como la medicina, la psicología, la economía, la ingeniería y la ciencia de datos. Esta técnica permite analizar la relación entre una variable dependiente binaria y una o más variables independientes continuas o categóricas. La regresión logística se utiliza para predecir la probabilidad de que la variable dependiente tome un valor determinado en función de las variables independientes.

Según Hosmer y Lemeshow [16], la regresión logística es "un modelo estadístico utilizado para analizar la relación entre una variable dependiente binaria y una o más variables independientes continuas o categóricas. El modelo estima la probabilidad de que la variable dependiente tome un valor determinado en función de las variables independientes".

Según James et al. [17], la regresión logística es "una técnica estadística utilizada para modelar la probabilidad de un evento binario en función de una o más variables explicativas. Se utiliza en diversas aplicaciones, como la predicción de la deserción universitaria, el análisis de riesgos médicos y la evaluación del crédito".

La regresión logística puede ser utilizada para analizar la relación entre una variable dependiente binaria y una o más variables independientes continuas o categóricas. El modelo estima la probabilidad de que la variable dependiente tome un valor determinado en función de las variables independientes. En otras palabras, la regresión logística permite predecir la probabilidad de que un evento ocurra o no, dadas las características de los datos.

La regresión logística se basa en la función logística, que es una función sigmoide que transforma una variable lineal en una probabilidad que varía entre 0 y 1. La ecuación de la función logística es:

$$p = \frac{1}{1 + e^{-z}}$$

donde p es la probabilidad de que ocurra un evento, z es la variable lineal y e es la función exponencial.

Uno de los principales beneficios de la regresión logística es que proporciona una forma de cuantificar el efecto de cada variable independiente en la probabilidad del evento binario. Esto se logra mediante el cálculo de los coeficientes de regresión, que indican el cambio en la probabilidad de la variable dependiente por unidad de cambio en cada variable independiente.

Uno de los primeros autores en proponer la regresión logística fue Joseph Berkson [18] en su artículo de 1944 titulado "Application of the Logistic Function to Bioassay". Berkson utilizó la función logística para modelar la relación entre la dosis de un compuesto y la respuesta biológica en una prueba de toxicidad.

Otro autor importante en el desarrollo de la regresión logística es David Cox. En su artículo de 1958 titulado "The Regression Analysis of Binary Sequences (with Discussion)", Cox presentó una formulación matemática de la regresión logística y proporcionó una forma de estimar los coeficientes de regresión utilizando el método de máxima verosimilitud.

En el ámbito de la estadística aplicada, uno de los libros más influyentes sobre la regresión logística es "Applied Logistic Regression" de Hosmer y Lemeshow [16]. En su libro, los autores proporcionan una introducción detallada a la regresión logística y discuten diversos temas relacionados, como la selección de variables y la validación del modelo.

5.1.2.1.2 Regresión Logística aplicada a la predicción de la deserción escolar

La regresión logística es una técnica estadística ampliamente utilizada en la predicción de la deserción escolar universitaria. La deserción universitaria es un problema importante en todo el mundo y puede tener graves consecuencias para los estudiantes, las instituciones educativas y la sociedad en general. La regresión logística se ha utilizado con éxito para identificar las variables que influyen en la deserción universitaria y para desarrollar modelos de predicción precisos.

En un estudio llevado a cabo por Chen y Su [19], se aplicó la regresión logística para predecir la deserción universitaria en Taiwán. Los autores encontraron que la edad, el género, el nivel de ingresos familiares, el tipo de institución educativa, el campo de estudio y el rendimiento académico eran variables significativas para predecir la deserción universitaria. Además, el modelo de regresión logística desarrollado en el estudio tuvo una alta precisión en la predicción

de la deserción universitaria, con una tasa de precisión del 83%.

Otro estudio llevado a cabo por Liu [19] utilizó la regresión logística para predecir la deserción universitaria en una universidad de Malasia. Los autores encontraron que la edad, el género, la etnia, el estado civil, el nivel socioeconómico, el tipo de alojamiento, el rendimiento académico y la satisfacción con el entorno educativo eran variables significativas para predecir la deserción universitaria.

En resumen, La regresión logística es útil en la predicción de la deserción universitaria al identificar variables que influyen en ella y desarrollar modelos precisos. Su aplicación permite que las instituciones educativas intervengan tempranamente y mejoren las tasas de retención de estudiantes.

5.1.2.1.3 Redes Neuronales

Son una técnica de modelado de datos inspirada en el funcionamiento del cerebro humano. Utilizan una red de nodos interconectados para procesar información y generar salidas en base a entradas previas. A menudo se utilizan en problemas de clasificación y predicción, como el reconocimiento de patrones, la detección de fraudes, la identificación de imágenes, entre otros. Las redes neuronales se han convertido en una herramienta muy poderosa para analizar grandes cantidades de datos y obtener patrones complejos que de otra manera serían difíciles de detectar. En este sentido, su uso se ha popularizado en diversos campos como la medicina, la economía, la industria, la ciencia de datos y la inteligencia artificial. En esta era de la tecnología y el big data, las redes neuronales ofrecen un enfoque prometedor para la solución de problemas complejos en una amplia gama de aplicaciones.

Según Haykin [20], "Las redes neuronales artificiales son un modelo matemático inspirado en la estructura y función de los sistemas neuronales biológicos, con el objetivo de emular su

capacidad para el aprendizaje y la generalización en la resolución de problemas". Las redes neuronales consisten en una serie de capas de nodos que procesan y transforman la información para producir una salida. Cada nodo está conectado a otros nodos en la red a través de conexiones ponderadas que se ajustan durante el proceso de aprendizaje.

Según Rumelhart [21] "Una red neuronal es una máquina de procesamiento paralelo compuesta por unidades de procesamiento simples, que tienen la propiedad de almacenar conocimiento adquirido a través del aprendizaje y utilizar ese conocimiento para resolver problemas", estas unidades de procesamiento, también conocidas como neuronas, están conectadas en una estructura de red y trabajan juntas para producir una salida. El aprendizaje en una red neuronal se produce mediante la modificación de los pesos de las conexiones entre las neuronas, lo que permite a la red ajustar su comportamiento para adaptarse a los datos de entrada.

5.1.2.1.4 Redes neuronales aplicadas a la predicción de la deserción escolar

Las redes neuronales son una técnica estadística poderosa y cada vez más utilizada en la predicción de la deserción escolar universitaria. Estas redes son un modelo computacional que se inspira en el cerebro humano y su capacidad de aprender a partir de ejemplos.

Un estudio, realizado por Chen [19], utilizó una red neuronal de aprendizaje profundo para predecir la deserción universitaria en una universidad de Taiwán. Los autores utilizaron variables como la calificación del examen de ingreso, la asistencia a clase y la satisfacción del estudiante con la universidad. El modelo alcanzó una precisión del 93.4%, lo que indica que la red neuronal es una técnica estadística útil en la predicción de la deserción universitaria.

En conclusión, las redes neuronales son una técnica estadística poderosa y cada vez más utilizada en la predicción de la deserción escolar universitaria. Los estudios revisados demuestran que las redes neuronales son capaces de identificar las variables relevantes y desarrollar modelos precisos de predicción. La aplicación de redes neuronales en la predicción de la deserción

universitaria puede ser útil para que las instituciones educativas puedan intervenir tempranamente y mejorar las tasas de retención de los estudiantes.

5.1.2.1.5 Arboles De Decisión

Los árboles de decisión son una técnica estadística popular y ampliamente utilizada en la minería de datos y el aprendizaje automático. Son una herramienta visual que ayuda a tomar decisiones basadas en una serie de reglas y variables, donde se busca identificar la mejor opción posible para un problema dado. El árbol se construye a partir de un conjunto de datos de entrenamiento, donde se busca crear un modelo que pueda predecir la respuesta correcta para nuevas observaciones. Los árboles de decisión son útiles en una variedad de campos, desde la toma de decisiones empresariales hasta la detección de fraudes en la banca, entre otros.

Otra definición de árboles de decisión es la de Breiman [22], quienes los definen como "un método para aproximar funciones que mapean entradas a salidas". Los árboles de decisión funcionan dividiendo los datos en subconjuntos más pequeños a través de una serie de preguntas y reglas, para finalmente llegar a una predicción. A medida que se avanza en la estructura de ramificación del árbol, se reduce la complejidad del problema.

Los árboles de decisión son una técnica popular y poderosa para modelar relaciones complejas entre variables en el análisis de datos. Según Loh [23], los árboles de decisión son "un método de clasificación que usa un conjunto de reglas de decisión para asignar etiquetas de clase a las instancias". En otras palabras, los árboles de decisión son una estructura de datos de tipo árbol que se utiliza para modelar relaciones no lineales entre variables de entrada y salida.

La estructura de un árbol de decisión consiste en un nodo raíz, nodos internos y hojas. Cada nodo interno representa una regla de decisión y cada hoja representa una etiqueta de clase o una salida de predicción. Para construir un árbol de decisión, se utiliza un algoritmo que busca la mejor regla de decisión para dividir los datos en subconjuntos más pequeños, de modo que se

maximice la pureza de las clases en cada subconjunto.

Uno de los principales beneficios de los árboles de decisión es que permiten visualizar la estructura de decisión de un problema de clasificación, lo que facilita la interpretación y la explicación de los resultados. Además, los árboles de decisión son relativamente fáciles de implementar y de usar, y no requieren una gran cantidad de recursos computacionales.

Aunque los árboles de decisión son una técnica poderosa, pueden presentar algunos desafíos en la práctica. Uno de los principales desafíos es la tendencia a sobre ajustar el modelo a los datos de entrenamiento, lo que puede llevar a una baja capacidad de generalización del modelo a nuevos datos. Para abordar este problema, se pueden utilizar técnicas de poda de árboles, que implican la eliminación de ramas innecesarias del árbol para mejorar su capacidad de generalización.

En conclusión, los árboles de decisión son una técnica valiosa para el modelado de relaciones no lineales en los datos. Su capacidad para representar las reglas de decisión de manera visual y para manejar múltiples variables de entrada los convierte en una herramienta popular en el análisis de datos. Sin embargo, es importante tener en cuenta los desafíos asociados con el sobreajuste del modelo y la necesidad de técnicas de poda para mejorar su capacidad de generalización.

5.1.2.1.6 Arboles de decisión aplicados a la predicción de la deserción escolar

La predicción de la deserción escolar universitaria es un problema importante en la educación superior. Los árboles de decisión son una técnica de modelado de datos popular y poderosa que se ha utilizado con éxito para la predicción de la deserción escolar universitaria en varios estudios. A continuación, se presentan tres estudios que han utilizado los árboles de decisión para la predicción de la deserción escolar universitaria y han reportado diferentes porcentajes

de precisión.

El primer estudio es de Cheng y Li [19], quienes utilizaron árboles de decisión para predecir la deserción universitaria en una universidad en China. Utilizando datos de 2.871 estudiantes, los autores identificaron un conjunto de factores clave que influyen en la deserción universitaria, incluyendo el rendimiento académico, la participación en actividades extracurriculares y la situación financiera del estudiante. El modelo de árbol de decisión resultante tuvo una tasa de precisión del 78,5%.

En resumen, los árboles de decisión son una técnica de modelado de datos útil para la predicción de la deserción escolar universitaria. Los estudios revisados han demostrado que los árboles de decisión pueden identificar las variables clave que influyen en la deserción universitaria y desarrollar modelos precisos de predicción. Las tasas de precisión de los modelos de árbol de decisión en la predicción de la deserción universitaria varían de estudio en estudio, pero en general, los modelos tienen una tasa de precisión superior al 75%. La aplicación de los árboles de decisión en la predicción de la deserción universitaria puede ser útil para que las instituciones educativas puedan intervenir tempranamente y mejorar las tasas de retención de los estudiantes.

5.1.2.1.7 Máquina de Soporte Vectorial (SVM)

Los algoritmos de Máquinas de Vectores de Soporte (SVM) son un tipo de algoritmo de aprendizaje automático supervisado utilizado para resolver problemas de clasificación y regresión. Fueron desarrollados por Vapnik y sus colegas en los años 90 en Bell Labs [24]. En el aprendizaje predictivo o supervisado, el objetivo es aprender una correspondencia entre las entradas "X" y las salidas "Y", dado un conjunto de pares de entrada-salida.

La idea detrás de SVM es encontrar un hiperplano que separe los datos en diferentes clases de

manera óptima. El hiperplano es una superficie de separación que divide el espacio de características en dos regiones, cada una correspondiente a una clase. La optimización se logra mediante la maximización de la margen, que es la distancia perpendicular desde el hiperplano a los puntos de datos más cercanos de cada clase. Los puntos de datos más cercanos se llaman vectores de soporte.

SVM se puede utilizar para clasificar datos linealmente separables o no linealmente separables. En el caso de datos no linealmente separables, se utilizan trucos de kernel para mapear los datos a un espacio de características de mayor dimensión en el que los datos son linealmente separables. Además de la clasificación binaria, SVM también se puede utilizar para la clasificación multiclase y la regresión. En la clasificación multiclase, se utilizan diferentes estrategias de clasificación, como uno contra todos y uno contra uno. En la regresión, se ajusta una función lineal o no lineal a los datos para predecir valores continuos. SVM tiene varias ventajas, como la capacidad de manejar datos de alta dimensionalidad y la capacidad de generalizar bien a datos nuevos. Sin embargo, SVM también tiene algunas limitaciones, como la sensibilidad a la selección de parámetros y la dificultad para manejar grandes conjuntos de datos.

5.1.2.1.8 Bosques aleatorios – Random Forest (RF)

Random Forest es un algoritmo de aprendizaje automático basado en un conjunto de árboles de decisión, que se utiliza para abordar problemas de clasificación y regresión en diversos campos, desde la ciencia de datos hasta la inteligencia artificial. Se destaca por su capacidad para mejorar la precisión y la capacidad de generalización del modelo, al combinar las predicciones de múltiples árboles y aprovechar la aleatoriedad tanto en la selección de características como en la construcción de los árboles.

En Random Forest, se construye un conjunto de árboles de decisión mediante un proceso de muestreo aleatorio y selección de características. Para cada árbol, se selecciona aleatoriamente

un subconjunto de datos de entrenamiento y un subconjunto de características del conjunto completo. A través de la construcción de los árboles individuales, se busca maximizar la pureza de las clases en cada subconjunto y obtener un conjunto diverso de modelos.

La predicción final en Random Forest se obtiene combinando las predicciones individuales de cada árbol. En problemas de clasificación, se utiliza la votación mayoritaria para determinar la clase final asignada a una instancia, mientras que, en problemas de regresión, se realiza un promedio de las predicciones de los árboles para obtener un valor continuo. Esta combinación de predicciones permite reducir el sesgo y mejorar la precisión del modelo, al tiempo que proporciona una mayor resistencia a los datos ruidosos y atípicos.

Además, Random Forest ofrece medidas de importancia de características, que permiten identificar qué características son más influyentes en el modelo y brindan información valiosa para el análisis de datos y la toma de decisiones. El algoritmo también se beneficia de la paralelización, lo que acelera el tiempo de entrenamiento y predicción en conjuntos de datos grandes.

En resumen, Random Forest es una técnica versátil y poderosa en el campo del aprendizaje automático, que combina múltiples árboles de decisión mediante estrategias aleatorias. Proporciona modelos más precisos y robustos, con capacidad de generalización y resistencia a los datos ruidosos. La capacidad de interpretación de los resultados y la capacidad de manejar características diversas hacen de Random Forest una herramienta invaluable en la investigación y aplicación de la ciencia de datos.

5.1.2.1.9 Bosques aleatorios aplicados a la predicción de la deserción escolar

La técnica de Random Forest aplicada a modelos de predicción de la deserción escolar es una estrategia efectiva para abordar este problema y ayudar a identificar a los estudiantes en riesgo

de abandonar sus estudios.

Al utilizar Random Forest, se construye un conjunto de árboles de decisión que analizan diversas características de los estudiantes, como el rendimiento académico, el comportamiento, la asistencia, el contexto socioeconómico, entre otros factores relevantes. Cada árbol dentro del bosque toma en cuenta un subconjunto aleatorio de características, lo que ayuda a mejorar la diversidad y la robustez del modelo.

La capacidad de Random Forest para manejar una amplia gama de características y capturar relaciones no lineales entre ellas lo convierte en una técnica adecuada para la predicción de la deserción escolar. Además, Random Forest proporciona medidas de importancia de características, lo que permite identificar qué factores son más influyentes en la toma de decisiones del modelo.

Una ventaja clave de Random Forest en el contexto de la predicción de la deserción escolar es su capacidad para lidiar con datos desequilibrados, donde la proporción de estudiantes que abandonan es generalmente menor en comparación con los que no abandonan. El enfoque de votación mayoritaria utilizado en Random Forest ayuda a manejar este desequilibrio y proporcionar predicciones más precisas.

Además, Random Forest ofrece una interpretación más clara de los resultados en comparación con otros algoritmos de aprendizaje automático más complejos. Los árboles de decisión individuales pueden ser visualizados y analizados para comprender las reglas de decisión utilizadas por el modelo en relación con la deserción escolar.

Para aplicar la técnica de Random Forest a la predicción de la deserción escolar, es necesario recopilar y preparar adecuadamente los datos relevantes de los estudiantes, identificar las

características más influyentes y entrenar el modelo utilizando conjuntos de datos de entrenamiento y validación. Posteriormente, el modelo se puede utilizar para predecir la probabilidad de deserción de nuevos estudiantes y así tomar medidas preventivas y de intervención temprana.

Es importante destacar que, si bien Random Forest es una técnica poderosa, el éxito de la predicción de la deserción escolar no se limita únicamente al modelo utilizado. Es fundamental complementar el análisis con una comprensión profunda del contexto educativo, la implementación de políticas de apoyo y la colaboración entre educadores, estudiantes y familias para abordar eficazmente las causas de la deserción escolar y fomentar el éxito académico de todos los estudiantes.

5.1.3 Minería de datos (Data Mining)

La minería de datos, también llamada como descubrimiento de conocimiento de datos es un conjunto de técnicas, tecnologías que permiten el proceso de descubrir patrones, tendencias o reglas que expliquen comportamiento en determinados contextos, a través de procesos de exploración de un grande conjunto de datos [25]. Autores como [26] describen la minería de datos como una técnica emergente, la cual permite agregar valor a los negocios desde varios ámbitos, inicialmente permite determinar puntos de partida para la gestión de la innovación permitiendo la alineación de los objetivos del negocio, de igual forma permite ahorro en términos monetarios y de tiempo, finalmente ayuda a identificar nuevas oportunidades de negocio.

A continuación, se describen las tecnologías que dan soporte al análisis masivo de datos o mineríade datos.

5.1.3.1 Data Ware Housing DWH

El primero en hablar de data wherehousing fue Ralph Kimball en los años de 1996 con la

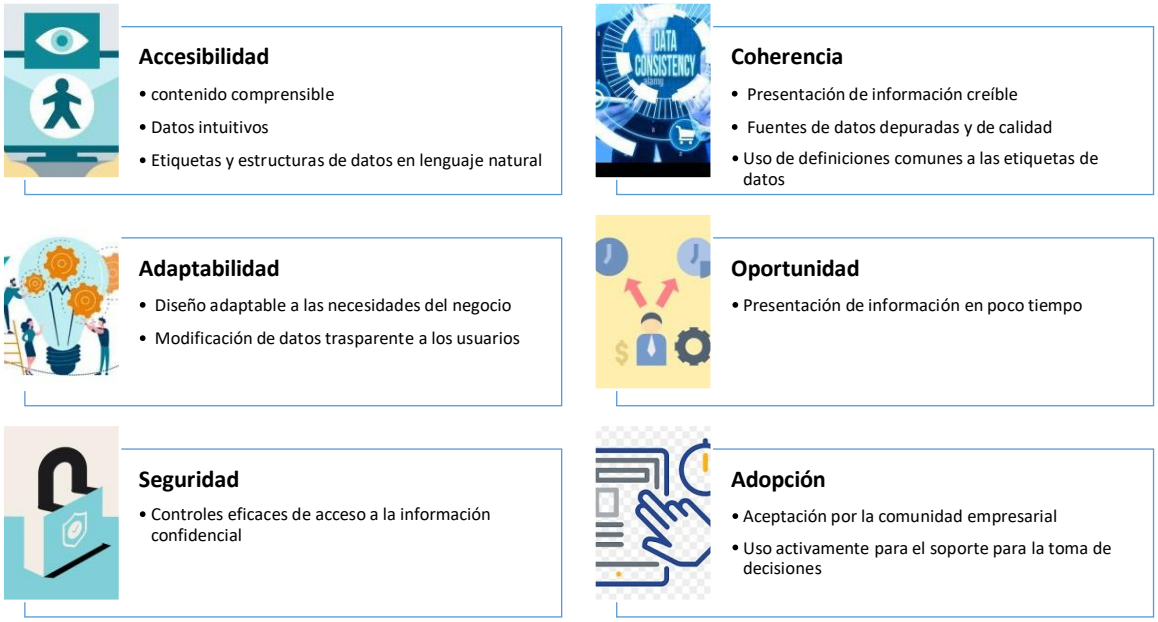
primera publicación The Data Warehouse Toolkit (Wiley), desde entonces las compañías ha ido adoptando este modelo de datos para el almacenamiento, distribución y análisis de estos.

Conceptualmente el DWH empresas como IBM [27] lo definen como un almacén de datos empresariales, que está compuesto por la agregación de diferentes fuentes de información bajo el esquema de almacén de datos centralizado, guardando una coherencia de los datos y se usa para apoyar el análisis de datos, la minería de datos, la inteligencia artificial y el aprendizaje automático, permitiendo así ejecutar potentes análisis sobre enormes volúmenes de informaciónen el orden de petabytes, característica que una base de datos estándar no soporta.

El DWH (Data Warehouse) se ha convertido en una pieza fundamental en la infraestructura tecnológica de las empresas modernas. Su capacidad de almacenar, integrar y gestionar grandes volúmenes de datos provenientes de diversas fuentes facilita la toma de decisiones basada en análisis profundos y precisos. Además, su arquitectura optimizada para consultas complejas y su capacidad de escalar horizontalmente hacen posible el procesamiento eficiente de petabytes de información, brindando a las organizaciones una ventaja competitiva en un mundo cada vez más impulsado por los datos.

El DWH para su implementación debe estar soportado por una arquitectura y una lógica de negocio para lograr dar respuesta a los requerimientos de análisis de grandes volúmenes de datospara esto se debe garantizar que se cumplan los siguientes requisitos como se muestra en la **Ilustración *1*** e **Ilustración 2**.

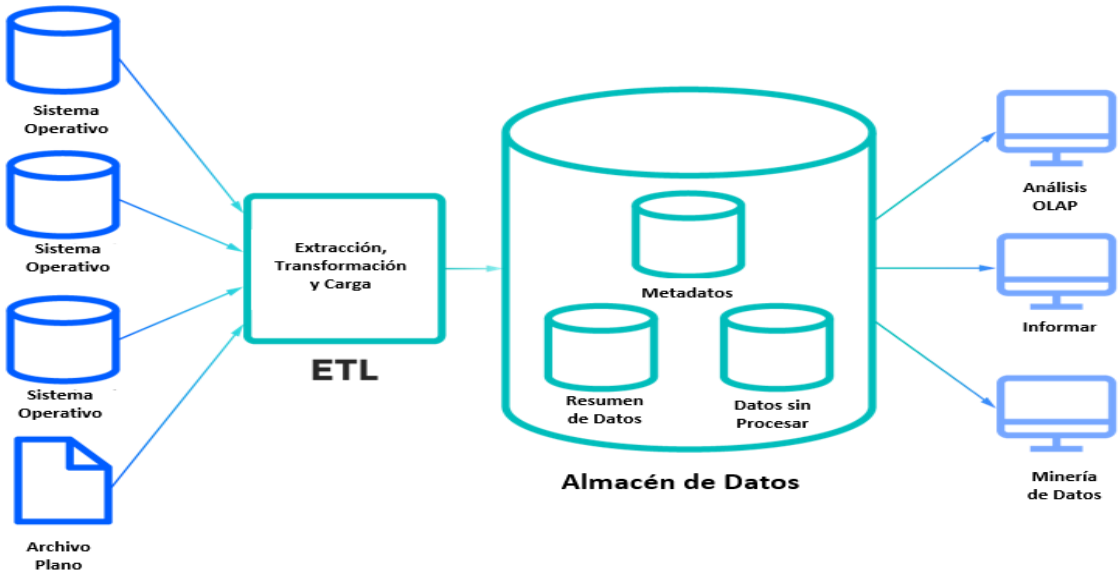
Ilustración 1 Requisitos básicos de un sistema de DWH



Fuente: elaboración propia basado en [28]

A nivel de arquitectura los DWH generalmente tienen tres niveles [27]:

Ilustración 2. Modelo de arquitectura de DWH



Fuente: IBM [27]

Nivel inferior: El nivel inferior consiste en un servidor de almacén de datos, normalmente un sistema de base de datos relacional, que recoge, limpia y transforma los datos de múltiples fuentes de datos a través de un proceso conocido como Extract, Transform, and Load (ETL) el cual se consiste en extraer los datos de las fuentes primarias, de la transformación y limpieza de estos y finalmente el proceso de carga en el almacén central DWH

Nivel intermedio: El nivel intermedio consiste en un servidor OLAP (procesamiento analítico en línea) que permite velocidades de consulta rápidas, el cual usan base de datos orientadas a procesos transaccionales OLTP (online transaction procesing), y base de datos OLAP cuya finalidad el proceso analítico de un gran volumen de datos.

Tabla 2. Diferencias entre las bases de datos OLAP y OLTP

Característica	Sistema OLTP	Sistema OLAP
Fuente de los datos	Datos operacionales, las bases de datos OLTP son las fuentes de los Datos.	Los datos de una base de datos OLAP provienen de diversas fuentes de datos OLTP.
Propósito de los datos	Control de tareas básicas del Negocio.	Ayuda a la planeación, solución de problemas y soporte de decisiones
Que revelan los datos	Visualización de los procesos actuales del Negocio.	Vistas multidimensionales de varios tipos de actividades del negocio.
Inserciones y actualizaciones	Las inserciones y actualizaciones realizan de forma corta y rápida siendo estas realizadas por usuarios finales	Se utilizan Jobs periódicos para actualizar los datos.

Queries	Relativamente estandarizados los cuales además retornan relativamente pocos datos.	Usualmente complejos debido a que envuelven diversas Agregaciones.
Velocidad de procesamiento	Muy rápido	Esta depende de la cantidad de datos envueltos, siendo las actualizaciones de los datos y las consultas complejas las que suelen tener velocidades de procesamiento de horas.
Requerimientos de espacio	Puede ser relativamente pequeño	Grande, debido a la existencia de estructuras de historiales de datosagregación.
Diseño de base de datos	Altamente normalizado con muchas tablas	Desnormalizada con pocas tablasutilizándose un esquema de estrella o copo de nieve.
Backup y recuperación de datos	Realización de backups continúa, debido a que los datos son críticos para el negocio	Se recargan los datos de la fuente de datos OLTP como método derecuperación de datos

***Fuente**:* adaptada de [29]

Nivel superior: El nivel superior está representado por algún tipo de interfaz de usuario front-end o herramienta de informes, que permite a los usuarios finales realizar análisis de datos ad-hoc sobre sus datos empresariales.

5.1.4 Metodología de proyecto de ciencia de datos. (CRISP-DM)

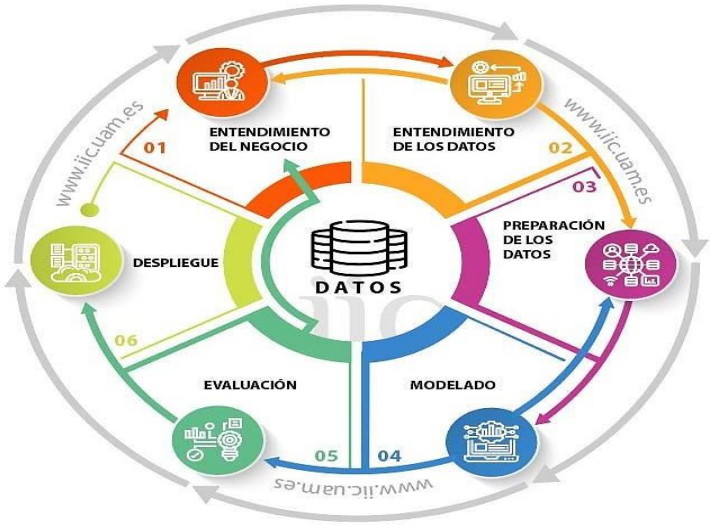
[30] “Las técnicas de Data Science o Data Analytics, que tanto interés despiertan hoy en día, en realidad surgieron en la década de los 90, cuando se usaba el término KDD (Knowledge Discoveryin Databases) para referirse al concepto de hallar conocimiento en los datos. En un intento de normalización de este proceso de descubrimiento de conocimiento, de forma similar a como se hace en ingeniería software para normalizar el proceso de desarrollo software, surgieron a finales de las 90 dos metodologías principales: CRISP-DM (Cross Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, and Assess)”.

CRISP-DM (Proceso Estándar Transversal para la Minería de Datos) presenta una descripción estandarizada del ciclo de vida de un proyecto de análisis de datos, de manera similar a cómo se hace en la ingeniería de software con los modelos de ciclo de vida del desarrollo de software. El modelo CRISP-DM abarca las etapas del proyecto, sus respectivas tareas y las interrelaciones entre estas tareas. Es importante destacar que no es posible identificar todas las relaciones a este nivel de descripción, ya que las relaciones pueden existir entre cualquier tarea, dependiendo de los objetivos, el contexto y el interés del usuario en los datos.

La metodología CRISP-DM considera el proceso de análisis de datos como un proyecto profesional, lo que implica un contexto mucho más completo que influye en la construcción de los modelos. Este contexto toma en cuenta la presencia de un cliente que no forma parte del equipo de desarrollo y reconoce que el proyecto no finaliza una vez que se encuentra el modelo óptimo. Además, se establece una relación con otros proyectos, lo que implica la necesidad de documentar de manera exhaustiva para que otros equipos de desarrollo puedan utilizar el conocimiento adquirido y trabajar sobre él.

Las etapas de la metodología se describen en la Ilustración 3 que sigue a continuación.

Ilustración 3 El ciclo de vida del proyecto de minería de datos



Fuente: [31]

5.1.4.1 Entendimiento del negocio

La etapa inicial de la guía de referencia CRISP-DM, conocida como fase de comprensión empresarial o del problema (ver Ilustración 3), representa un paso crucial y abarca las tareas relacionadas con la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional. El objetivo es transformar estos objetivos en metas técnicas y en un plan de proyecto. Sin comprender plenamente estos objetivos, ningún algoritmo, sin importar cuán sofisticado sea, podrá generar resultados confiables. Para aprovechar al máximo la Minería de Datos, es esencial comprender a fondo el problema que se desea resolver. Esto implica recopilar los datos correctos e interpretar adecuadamente los resultados. Durante esta fase, resulta crucial convertir el conocimiento empresarial adquirido en un problema de Minería de Datos y en un plan preliminar que tenga como objetivo alcanzar los objetivos empresariales. A continuación, se presenta una descripción de las principales tareas que componen esta fase [32].

5.1.4.2 Entendimiento de los datos

Se encarga de la recolección de datos, se iniciará a partir de la Big Data empresarial la comprensión de su naturaleza y significado del análisis de calidad para reconocer su validez de análisis la exploración inicial de los datos o de los conjuntos de datos y el planteamiento de las primeras hipótesis sobre los mismos en esta fase se tiene en cuenta también la fuente de datos que hasta el momento no se estaba utilizando como por ejemplo las fuentes externas a la empresa. [32]

5.1.4.3 Preparación de los datos

Esta etapa abarca todas las acciones necesarias para construir el conjunto de datos final que será utilizado en el entrenamiento de las herramientas de aprendizaje automático. El proceso de preparación de datos constituye la fase más extensa en el ciclo de vida del proyecto, llegando a representar hasta el 80% del tiempo de desarrollo. Esta fase reviste gran importancia, ya que se

encarga de la preparación de los datos, incluyendo actividades como la selección, limpieza, construcción de variables o campos, integración y formateo de los datos. Una de las primeras tareas en esta fase es la creación del diccionario de datos, que resulta crucial para establecer una base sólida. Además, se lleva a cabo la generación de nuevos datos a partir de los campos existentes, lo cual constituye una actividad crítica. [32]

5.1.4.4 Modelamiento

En esta fase se aplican las técnicas de modelamiento que se consideren pertinentes para el problema detectado y cuyas propiedades aporten el mayor valor al negocio. También se van a desarrollar actividades específicas como el diseño del modelo, la selección de la plataforma y el entrenamiento un Machine Learning de los datos, las pruebas y la evaluación de la performance del modelo con relación a un modelo ideal este proceso puede ser repetitivo hasta encontrar un modelo de alta performance. [32]

5.1.4.5 Evaluación

Se evalúan los modelos de la fase anterior para determinar si son muy útiles a las necesidades del negocio y si cumplen con los objetivos del proyecto esta fase no debe confundirse con la actividad de la evaluación de la performance del modelo realizado en la fase anterior, por esto mismo en esta etapa los modelos ya están construidos y deben tener una alta calidad desde una perspectiva de análisis de datos. En esta se realizan tres actividades principales la Evaluación de los Resultados y el Logro de los Objetivos del Proyecto. [32]

5.1.4.6 Despliegue

En esta fase una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes

conjuntos de datos o como parte del proceso, como, por ejemplo, en análisis de riesgo crediticio, detección de fraudes, etc. Generalmente un proyecto de Data Mining no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento. [33]

5.1.5 Sistemas de alertas tempranas SAT

Es un conjunto de herramientas [34] que determinan la posibilidad de ocurrencia de un resultado considerado negativo hacia un fenómeno en un determinado contexto, la anticipación se da mediante el análisis de señales físicas que pueden ser causantes de un resultado desfavorable, su importancia se debe a la prevención de pérdidas materiales o humanas. Los sistemas de alertas tempranas pueden ser utilizados en diversos campos donde sea susceptible el pronóstico de una amenaza a una población vulnerable, ya sea por fenómenos naturales o sociales, por lo que se hace necesario realizar un monitoreo sobre las variables que intervienen en el proceso.

Particularmente, en el sector educativo, existe un fenómeno social que genera preocupaciones por las pérdidas que genera a nivel individual y colectivo, denominado deserción escolar, en dicho caso es importante identificar las variables cualitativas y cuantitativas que pueden llevar a que un estudiante abandone sus estudios, y por medio del monitoreo dar una respuesta de alerta que permita a las instituciones educativas, tomar medidas para fortalecer la permanencia de los estudiantes involucrados.

La finalidad de un SAT en educación es identificar a los estudiantes en riesgo de abandonar sus estudios, incluyendo aquellos que inicialmente no se consideran en ese estado, generando un menor impacto en el tema de deserción, pero también en la percepción de los mismos estudiantes, por lo que también es informado al mismo tiempo que lo es la institución de estudio, dando recomendaciones a ambas partes para atender la problemática y tomar medidas que eviten el abandono de los estudios.

5.2 ANTECEDENTES

A continuación, en la **Tabla 3**. Principales estudios de predicción de la deserción usando ciencia de datos , se describen los principales trabajos de investigación que guardanrelación con el objeto de estudio y presentan un marco de antecedentes en el uso de la ciencia de datos para la predicción de fenómenos como la deserción escolar.

Tabla 3. Principales estudios de predicción de la deserción usando ciencia de datos

Estudio	Aporte
HACIA LA CONSTRUCCIÓN DE UN MODELO PREDICTIVO DE DESERCIÓN ACADÉMICA BASADO EN TÉCNICAS DE MINERÍADE DATOS **Autores:** Jonny Sotomonte Castro, Cristian Camilo Rodríguez, Carlos Enrique Montenegro Marín, Paulo Alonso Gaona García, John Gabriel Castellanos	En el ámbito de la investigación, se ha desarrollado un modelo predictivo utilizando la metodología de Árboles de Decisión y el algoritmo J48 implementado en la herramienta WEKA. El objetivo de este modelo es determinar la probabilidad de deserción de un estudiante en el contexto de la facultad de ingeniería de la Universidad Distrital Francisco José de Caldas, abarcando el período comprendido entre 2009 y 2015. Para la construcción de este modelo se consideraron variables demográficas, socioeconómicas y académicas relevantes [35].
MODELO PARA LA PREDICCIÓN DE LA DESERCIÓN DE ESTUDIANTES DE PREGRADO, BASADO EN TÉCNICAS DE MINERÍA DE DATOS **Autores:** Aníbal José Camargo García	Este fue un proyecto de investigación realizado con el objetivo de crear un modelo para la predicción de la deserción de estudiantes de pregrado en la La Universidad de la Costa - CUC ha llevado a cabo un estudio en el que se analizan diversos factores socioeconómicos y académicos con el objetivo de comprender la deserción estudiantil. Este estudio se ha desarrollado a través de distintas etapas: caracterización, experimentación, desarrollo y evaluación. Durante la fase de caracterización, se ha construido un conjunto de datos (dataset) mediante la recopilación de información demográfica, cultural, social, familiar, educativa, estatus socioeconómico y perfil psicológico de cada estudiante. Los datos utilizados abarcan el período comprendido entre 2013-1 y 2018-2 [36].

MODELO PREDICTIVO PARA ESTIMAR LA DESERCIÓN DE ESTUDIANTES EN UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR **Autores:** Jonathan Vásquez	En el presente trabajo de investigación se llevó a cabo un análisis exhaustivo de la problemática de la deserción estudiantil en el entorno universitario de Chile. Se enfocó específicamente en el comportamiento de los estudiantes en un programa académico perteneciente a la Universidad de Chile, gestionado por la Escuela de Sistemas de Información y Auditoría de la Facultad de Economía y Negocios [37].
MODELO PREDICTIVO DE DESERCIÓN ESTUDIANTIL BASADO EN ARBOLES DE DECISIÓN **Autores:** Blanca Cuji, Wilma Gavilanes, Rina sanchez	En este artículo se presenta la elaboración de un modelo predictivo de deserción estudiantil, diseñado para predecir la probabilidad de que un estudiante abandone su programa académico, utilizando técnicas de clasificación basadas en árboles de decisión. La metodología empleada se fundamenta en el Descubrimiento de Conocimiento en Bases de Datos (KDD), y consta de cinco etapas: selección, procesamiento, transformación, minería de datos y evaluación. Mediante la aplicación del algoritmo Classification and Regression Tree (CART) en la herramienta R, se construyó un árbol de cuatro niveles de profundidad y cuatro reglas correspondientes, que evalúan a los posibles estudiantes desertores. Como resultado del análisis, se concluye que las variables nivel y notas son las que ejercen una mayor influencia en la deserción estudiantil. [38].
SISTEMA DE PREVENCIÓN Y ANÁLISIS DE LA DESERCIÓN EN LAS INSTITUCIONES DE EDUCACIÓN SUPERIOR Documento del ministerio de educación nacional en colaboración de la universidad de los andes	Se llevó a cabo una investigación con el propósito de examinar los factores clave que influyen en el fenómeno de la deserción en las Instituciones de Educación Superior en Colombia, abarcando el período comprendido entre el primer semestre de 1998 y el segundo semestre de 2013. El objetivo fue identificar los factores relacionados con la deserción y determinar cuáles han tenido un mayor impacto en su aumento. [6]
FACTORES QUE MOTIVAN EL ABANDONO ESTUDIANTIL EN LA UNIVERSIDAD: UN ESTUDIO DE CASO. **Autores:** Tahimi Achilie Valencia	El propósito de este estudio fue identificar los elementos que conducen al abandono de los estudios universitarios. Se empleó una metodología basada en un análisis de caso, sin diseño experimental, con un enfoque combinado. Se utilizaron técnicas de encuesta y entrevista mediante un cuestionario estructurado que constaba de 24 preguntas abiertas y cerradas. Los resultados mostraron que el factor determinante de

abandono de los estudios es la motivación personal, que abarca la actitud hacia el crecimiento profesional, el escaso interés en los estudios y las expectativas en relación con la carrera. A continuación, se encuentra el factor académico, y en última instancia, el factor institucional. Como conclusión, se recomienda llevar a cabo una intervención enfocada en el desarrollo de habilidades y confianza del estudiante, además de implementar estrategias de respaldo universitario como laboratorios, cursos de nivelación y acompañamiento para aquellos estudiantes en riesgo académico, con el objetivo de reducir la deserción y fortalecer la retención de los estudiantes. [39]

LA INVESTIGACIÓN SOBRE DESERCIÓN UNIVERSITARIA EN COLOMBIA 2006-2016. TENDENCIAS Y RESULTADOS

Autor: Marcela Rodríguez Urrego

Este artículo recoge los resultados más relevantes de una revisión documental de 28 investigaciones realizadas en Colombia entre el 2006 y el 2016, sobre deserción en educación superior. Esta revisión fue el punto de partida de una investigación sobre la deserción en la Licenciatura en Educación Comunitaria de la Universidad Pedagógica Nacional. Los resultados se presentan en seis apartados: 1) caracterización de la muestra en la que se tipifican los estudios revisados, según su metodología y objeto de investigación; 2) consideraciones generales sobre la deserción, donde se evidencia la concepción de deserción que sirve de punto de partida de los estudios; 3) conceptualización de la deserción, en donde se explicitan los conceptos centrales de los estudios revisados y los problemas asociados a una deficiente conceptualización; 4) factores asociados, en el que se reportan los modelos de deserción utilizados en los análisis estadísticos; 5) presenta de manera sucinta los resultados de las investigaciones empíricas con relación a la deserción en las unidades académicas o conjuntos de los estudiados, y 6) un aparte final con resultados y reflexiones referidos a la retención, permanencia y prevención de la deserción. [40]

PROPUESTA DE UN MODELO PREDICTIVO UTILIZANDO APRENDIZAJE PROFUNDO PARA EL ANÁLISIS DE DESERCIÓN ESTUDIANTIL EN UNIVERSIDADES COLOMBIANAS VIRTUALES

Autor: Martínez, Julio César Mateus, Sandra Patricia

En este artículo se presentó un modelo de predicción que brinda apoyo a las universidades colombianas para el análisis de la deserción de estudiantes, especialmente en programas de pregrado en modalidad virtual. Un modelo de predicción puede ser de utilidad para las organizaciones al generar ganancias y evitar pérdidas futuras, al utilizar datos históricos para prever resultados y respaldar la toma de decisiones. Este modelo se construye mediante la utilización de eventos pasados con diversas variables sociales, académicas, personales, laborales, ingreso a plataformas de aprendizaje en línea, entre otras. Posteriormente, se aplican algoritmos de aprendizaje profundo a estas variables. La variable objetivo es la probabilidad de deserción de cada estudiante, y con esta información se generan alertas y se implementan medidas preventivas tempranas con la población estudiantil. [41]

INVESTIGACIÓN EN DESERCIÓN ESTUDIANTIL UNIVERSITARIA: EDUCACIÓN, CULTURA Y SIGNIFICADOS

Autor: Floralba Barrero Rivera

En este artículo se realiza una reflexión sobre la responsabilidad de las instituciones de educación superior en relación con la graduación de los estudiantes. Se destaca el desafío que representa la deserción estudiantil universitaria para cumplir con esta tarea. El abandono de los estudiantes en la universidad tiene un impacto negativo en el progreso del país en diversos ámbitos sociales y científicos. La educación juega un papel fundamental en la prevención e intervención de este problema. Desde esta perspectiva, se presentan las tendencias actuales de investigación sobre la deserción estudiantil universitaria. Además, se proporciona una conceptualización del tema y se resumen brevemente algunos estudios relevantes, clasificándolos desde las perspectivas cuantitativa y cualitativa. Se reflexiona sobre la complementariedad de ambos enfoques, con el objetivo de comprender, analizar, abordar e intervenir en este fenómeno, a fin de proponer soluciones efectivas que aborden esta problemática educativa y social. Por último, se reflexiona sobre el papel de la educación superior en la comprensión y mitigación de este fenómeno. [42]

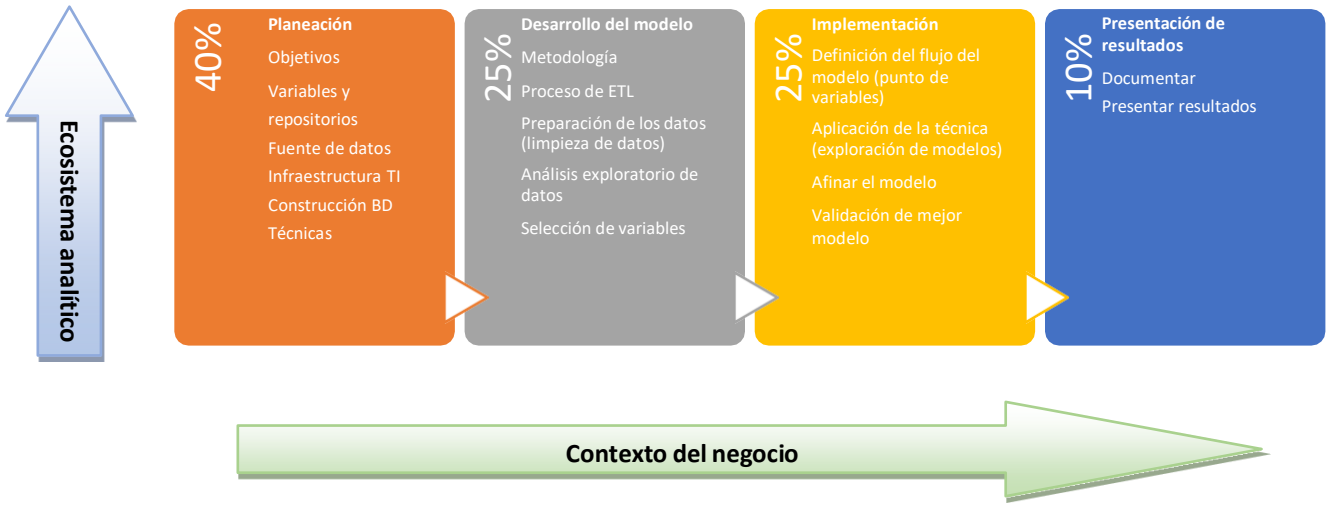
PREDICTING STUDENT'S DROPOUT IN UNIVERSITY CLASSES USING TWO-LAYER ENSEMBLE MACHINE LEARNING APPROACH: A NOVEL STACKED GENERALIZATION **Autores:** Jovial Niyogisubizoa, Lyuchao Liaoa, Eric Nziyumvaa, Evariste Murwanashyakab, Pierre Claver Nshimyumukizac	En este artículo se presenta un nuevo modelo conjunto que combina el Bosque Aleatorio (RF), el Aumento de Gradiente Extremo (XGBoost), el Aumento de Gradiente (GB) y la Red Neuronal de Avance (FNN) para predecir la deserción de los estudiantes en cursos universitarios. Utilizando un conjunto de datos recopilados de 2016 a 2020 por la Universidad de Filosofía de Nitra Constantine, bajo las mismas condiciones, se observó un rendimiento superior del método propuesto en comparación con el modelo base, utilizando la precisión de la prueba y el Área Bajo la Curva (AUC) como métricas de evaluación del rendimiento. Estos resultados permiten identificar a los estudiantes en riesgo de deserción en función de los factores que influyen, lo que brinda a las instituciones educativas la posibilidad de intervenir tempranamente en comportamientos desfavorables que podrían llevar al abandono. [43]
PREDICTING STUDENT DROP-OUT RATES USING DATA **MINING TECHNIQUES: A CASE STUDY** **Autores:** Boris Pérez, Camilo Castellanos, Darío Correal	En este artículo se exponen los descubrimientos de un estudio de caso centrado en datos educativos, con el objetivo de detectar la deserción de estudiantes de ingeniería de sistemas en una universidad colombiana después de 6 años. Los datos se enriquecen mediante un proceso de ingeniería de datos. Los resultados experimentales indican que un algoritmo sencillo logra una precisa identificación de los factores de abandono. Se comparan los resultados de diferentes algoritmos, como árboles de decisión, regresión logística, Naive Bayes y Random Forest, para determinar la opción más adecuada. También se evalúa la usabilidad de Watson Analytics para usuarios sin experiencia. Se presentan resultados preliminares para disminuir las tasas de deserción mediante la identificación de posibles causas, así como descubrimientos relacionados con la calidad de los datos para mejorar la recopilación de información de los estudiantes. [44]
ANALYSIS OF FIRST-YEAR UNIVERSITY STUDENT **DROPOUT THROUGH MACHINE LEARNING MODELS: A COMPARISON BETWEEN UNIVERSITIES** **Autores:** Diego Opazo, Sebastián Moreno, Eduardo	En este artículo se examina el abandono estudiantil en instituciones de educación superior a nivel global. Se emplean modelos de aprendizaje automático en dos universidades chilenas para predecir la deserción de estudiantes de ingeniería durante el primer año. Los resultados señalan que resulta más

Álvarez Miranda,3, Jordi Pereira	efectivo utilizar modelos específicos por universidad en lugar de combinar conjuntos de datos. El modelo óptimo identificado fue el árbol de decisión potenciado por gradiente. Un análisis adicional revela que obtener calificaciones altas en los exámenes de admisión, especialmente en matemáticas, se relaciona con una menor probabilidad de abandono, aunque en el caso de las pruebas de idioma, las calificaciones más altas aumentan la probabilidad de abandono. [45]

6. METODOLOGÍA

Para la realización del proyecto se implementará la metodología de proyecto de ciencia de datos (CRISP-DM) y se desarrollará en 4 fases con los aspectos a considerar más relevantes y el porcentaje de tiempo involucrado para el desarrollo de éste, como se nuestra en la **Ilustración 4**

Ilustración 4. Fases del proyecto de ciencia de datos SATDU



Fuente: elaboración propia

Contexto del negocio

El proyecto en sus cuatro etapas debe estar alienado al contexto del negocio, que garantice atender a las condiciones de la universidad a nivel de direccionamiento estratégico y flujo de trabajo (workflow), permitiendo así la integración a nivel tecnológico, de infraestructura y de personal, que garantice los objetivos del proyecto y generen impactos significativos posterior a su implementación.

Ecosistema analítico

Todo proyecto de ciencia de datos debe ser analizado a través de un comportamiento sistémico donde se determine el flujo de los datos a nivel de entradas, procesamiento y salidas, la interoperabilidad con otros sistemas, las variables y la capacidad de repuesta del mismo que garantice generación de valor en la organización.

I. Planeación

Es la etapa más importante del proyecto donde se terminan los objetivos, la selección de variables y sus fuentes de datos, la selección de la técnica a usar, la alienación estratégica con el contexto del negocio, el análisis y selección de la infraestructura tecnológica y la construcción de base de datos.

El proceso de selección de variables es fundamental porque determina la calidad de los datos, e inicia a partir de la categorización de las variables la identificación de si estas atienden a todos los clientes, accesibilidad fuentes de datos (temporal, aplicación de formularios, BD, terceros, conexiones con otros sistemas) La infraestructura TI, determina cómo se selecciona la tecnología y la comunicación entre sistemas dado que un proyecto de ciencia de datos agrupa un conjunto fuentes de datos y de uso de tecnologías y se debe garantizar la interoperabilidad entre estos sistemas.

Por último, la elección de la metodología posibilita la integración de los elementos de tecnología de la información, recurso humano, fuentes de información y entorno empresarial. En el caso específico, se consideraron dos opciones: la creación de reglas de programación y la implementación de un modelo encapsulado, siendo este último el seleccionado para su utilización. Además, se reconoce que dicha elección brinda la oportunidad de aprovechar al máximo las capacidades existentes y adaptarlas eficientemente a las necesidades del caso. Esto permite lograr una mayor sinergia entre los distintos componentes involucrados.

II. Desarrollo del modelo

Se basa fundamentalmente al desarrollo metodológico para ejecutar lo planificado basado en las condiciones de la compañía y sus flujos de trabajo a nivel de tecnología e infraestructura y personal, donde se parte de una buena limpieza de datos, la selección de atributos, estandarización y normalización. La construcción del DWH con los datos teniendo en cuenta la accesibilidad y disponibilidad de las fuentes de datos a través de un proceso de ETL.

Para el proceso de selección de variables este se realizó bajo dos criterios: el estadístico y de necesidad de negocio. Se aplicarán funciones para agregar nuevos atributos y técnicas de reducción dimensional para la selección de las variables óptimas para el modelo predictivo. El desarrollo del modelo se realizará con enfoque diferencial para el periodo de la pandemia de COVID 19 a partir del mes de marzo de 2020 y se evaluará estadísticamente relaciones de causalidad o dependencia asociado a este fenómeno.

III. Implementación

Se desarrollará a partir de enfoque estadístico para el desarrollo del modelo y la validación de este dividiendo los datos en 80% para entrenamiento y 20% para validación del modelo, las fuentes de datos a integrar son la caracterización del estudiante al ingreso, pruebas saber 11, sistema académico y evaluación docente.

IV. Presentación de resultados

Durante esta etapa de preparación de resultados, se llevará a cabo la elaboración de una presentación cuidadosamente diseñada que garantice una comprensión clara y precisa del proyecto y los logros alcanzados. Esta presentación se convierte en un elemento fundamental, ya que proporciona información relevante y valiosa para respaldar la toma de decisiones informadas.

7. DESARROLLO

7.1 Planeación

7.1.1 Caracterización del conjunto de datos

El conjunto de datos se entrega en un libro de Excel el cual tiene siguiente estructura hoja de caracterización socio demográfica la cual posee información de la tabulación de la encuesta de ingreso a la facultad de ingeniería de la universidad objeto de estudio, la cual tiene información desde el periodo académico 2017-1 hasta 2022-2, para los programas de ingeniería industrial y de sistemas con un total de 1957 registros y 97 variables como se detalla en la **Tabla 4**. , dentro de las variables a analizar se destacan información datos personales como edad, estado civil, número de hijos, residencia, estrato social, condiciones de trabajo del grupo familiar, ingresos y egresos y niveles educativos e información sobre habilidades en matemática, lenguaje, otros.

Tabla 4. Detalle fuente de datos caracterización sociodemográfica

Programa	Cantidad de datos	Numero de variables
INGENIERÍA DE SISTEMAS	754	97
INGENIERÍA INDUSTRIAL	1.203	
Total, general	1.957	97

Fuente: Elaboración propia

Hoja de caracterización académica la cual tiene información de los semestres cursados, asignaturas y notas, las medidas de cantidades de asignaturas aprobadas y reprobadas por cada semestre cursado, como se resumen en la **Tabla**

Tabla 5. Detalle fuente de datos caracterización académica

Programa	Cantidad de datos	Numero de variables
INGENIERÍA DE SISTEMAS	36.178	19
INGENIERÍA INDUSTRIAL	61.729	
Total, general	97.907	19

Fuente: Elaboración propia

Hoja de incentivos contiene la información sobre los programas, semestres y los tipos de incentivos y el valor por actividades deportivas, becas, bonos de descuentos y notas crédito como se observa en la *Tabla 6.*

Tabla 6. Detalle fuente de datos caracterización Incentivos

Programa	Cantidad de datos	Numero de variables
INGENIERÍA DE SISTEMAS	452	9
INGENIERÍA INDUSTRIAL	1.092	
Total, general	1.544	9

Fuente: Elaboración propia

Hoja de créditos directos esta fuente de datos contiene información de los estudiantes por programas que financiaron el semestre académico, donde se especifica el valor del crédito y la cuota inicial del mismo, esta fuente de datos sólo contiene los créditos directos con la universidad y no incluye la información de ICETEX u otros fondos y/o modalidades de crédito, como se detalla en la *Tabla 7.*

Tabla 7. Detalle fuente de datos caracterización créditos directos

Programa	Cantidad de datos	Numero de variables
INGENIERÍA DE SISTEMAS	317	6
INGENIERÍA INDUSTRIAL	706	
Total, general	1.023	6

Fuente: Elaboración propia

Finalmente, la última hoja del libro de los datos objeto de estudio es la información de los desertores por programa, periodo académico de la matrícula inicial, cantidad de asignaturas aprobadas y reprobadas, promedio del semestre, promedio general y el periodo de deserción como se detalla en la **Tabla 8.**

Tabla 8. Detalle fuente de datos caracterización deserción

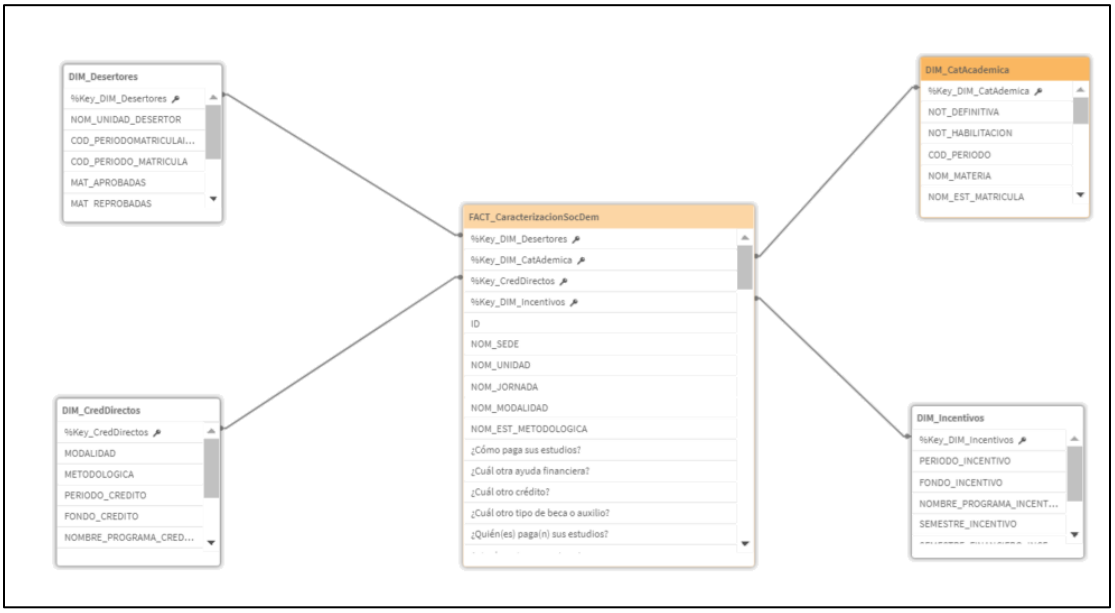
Programa	**Cantidad de datos**	**Numero de variables**
INGENIERÍA DE SISTEMAS	210	8
INGENIERÍA INDUSTRIAL	272	
Total, general	482	8

Fuente: Elaboración propia

7.1.2 Modelo bidimensional de datos

A partir de la información de las caracterizaciones sociodemográficas, académica, incentivos, créditos directos y deserción se construye un modelo de bodega de datos el cual agrupa los datos por hechos y dimensiones donde la tabla central del modelo es la caracterización demográfica y esta se detalla a partir de las dimensiones académica, créditos directos, incentivos y deserción como se muestra en la **Ilustración *1***, para la tabla de hechos se agregan las siguientes métricas calculadas, las cuales se transforman en variables de estudio por cada estudiante caracterizado: promedio general, último semestre cursado, máximo periodo cursado, total asignaturas aprobadas, total asignaturas reprobadas total incentivos, máximo periodo de incentivos, máximo semestre de crédito, promedio de crédito, promedio de cuota inicial, máximo periodo de deserción y la variable desertor como variable de estudio con valores binarios de 0 y 1.

Ilustración 5. Modelo bidimensional del análisis de la deserción



Fuente: elaboración propia

7.2 Desarrollo del modelo

7.2.1 Preparación y limpieza de datos

La base de datos resultante una vez aplicado la agrupación de datos por cada una de las dimensiones descritas en el modelo bidimensional (Ilustración 5), se resume en un set de datos compuesta de un conjunto de datos de 99 variables incluida la variable objetivo la deserción y 1.957 registros como se puede observar en el anexo 1 descripción de variables y datos. Inicialmente se determina el porcentaje de datos faltantes o nulos en el conjunto de datos como se muestra en la *Ilustración 6*, seguido a esto se procede a eliminar del conjunto de datos a aquellas variables cuyo porcentaje de nulos es mayor que 30%, quedando un set de datos de 36 variables.

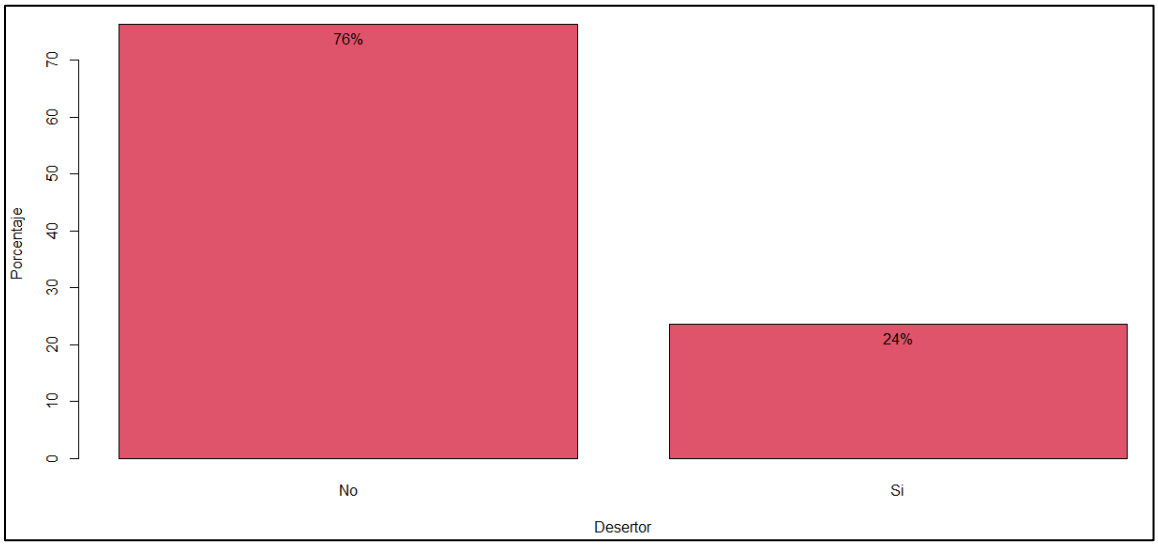
Ilustración 6. Porcentaje de datos faltantes por variable

v1	v2	v3	v4
27.8487481	25.8048033	96.9340828	96.9851814
v5	v6	v7	v8
40.2146142	58.3035258	39.4481349	85.4879918
v9	v10	v11	v12
61.3183444	37.4552887	52.0183955	96.6274911
v13	v14	v15	v16
99.5401124	83.1885539	99.3357179	94.0725600
v17	v18	v19	v20
96.2187021	22.3811957	22.5344916	22.3811957
v21	v22	v23	v24
22.3811957	93.0505876	84.8748084	65.4062340
v25	v26	v27	v28
98.9269290	94.1236587	96.6274911	22.6366888
v29	v30	v31	v32
22.3811957	22.3811957	52.0183955	52.0183955
v33	v34	v35	v36
96.8829842	22.6366888	22.3811957	52.0183955
v37	v38	v39	v40
22.5855902	52.0183955	97.0873786	22.3811957
v41	v42	v43	v44
39.4481349	85.4879918	22.3811957	97.8027593
v45	v46	v47	v48
63.0045989	22.6366888	96.7296883	22.3811957
v49	v50	v51	v52
22.3811957	22.3811957	85.4879918	22.3300971
v53	v54	v55	v56
94.0725600	22.2789985	65.7639244	22.3811957
v57	v58	v59	v60
22.3300971	96.8829842	22.3811957	96.9340828
v61	v62	v63	v64
22.6366888	52.0183955	98.4670414	39.3970363
v65	v66	v67	v68
85.4879918	97.7516607	61.2161472	98.4670414
v69	v70	v71	v72
22.6366888	39.3970363	85.4879918	97.8027593
v73	v74	v75	v76
63.0045989	52.0183955	65.8150230	96.9340828
v77	v78	v79	v80
94.1236587	22.3300971	22.6366888	22.3300971
v81	v82	v83	v84
22.3811957	22.3811957	52.0183955	52.0183955
v85	NOMBRE PROGRAMA	PROMEDIO GENERAL	ULTIMO SEMESTRE CURSADO
52.0183955	0.0000000	0.0000000	0.1532959
MAXIMO PERIODO CURSADO	TOTAL APROBADAS	TOTAL REPROBADAS	TOTAL INCENTIVOS
0.1532959	0.0000000	0.0000000	0.0000000
MAX SEMESTRE INCENTIVO	PROMEDIO INCENTIVO	MAX SEMESTRE CREDITO	PROMEDIO CREDITO
75.0638733	75.0638733	75.8814512	75.8814512
PROMEDIO CUOTA INICIAL	MAX PERIODO DE DESERCION	DESERTOR	
75.8814512	75.7281553	0.0000000	

Fuente: elaboración propia

Del nuevo set de datos con 36 variables se aplican las siguientes transformaciones a la variable v19 la cual corresponde a la edad se transforma a variable numérica, se crea un nuevo data set omitiendo los valores nulos y se grafica la proporción de datos de la variable objetivo como se muestra en la *Grafica 1*, se observa que el 76% equivalen a no desertores y el 24% restante a desertores, quedando el conjunto de datos para el desarrollo del análisis exploratorio de datos.

Gráfica 1. Porcentaje de valores de deserción



Fuente: elaboración propia

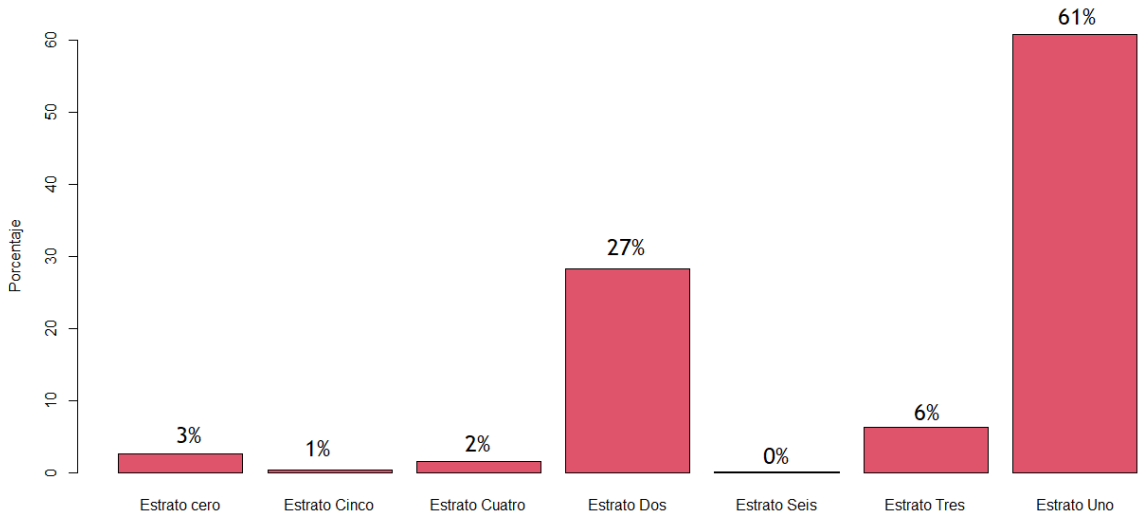
7.2.2 Análisis exploratorio de datos

Análisis Univariado

Una vez completada la etapa de limpieza de los datos se procedió a analizar de manera individual cada una de las variables que pueden influir sobre la variable explicada de la Deserción. A continuación, se describe las características generales de cada una de ellas.

La primera variable para analizar es el estrato socioeconómico, en el cual se evidencia que la mayoría de los estudiantes pertenecen a los estratos uno (61%) y estrato dos (27%) que corresponde al 88% del total, como se muestra en la *Grafica 2.*

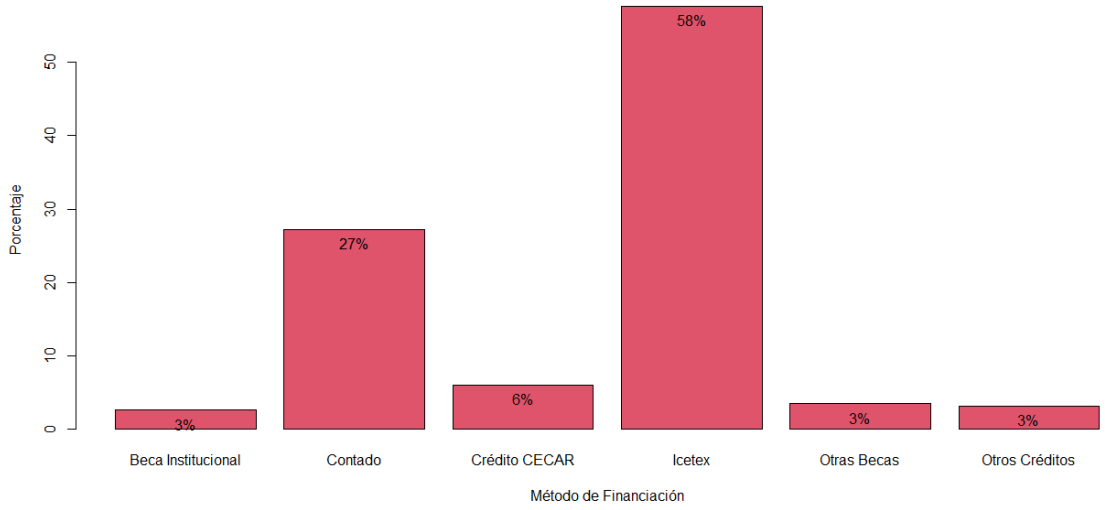
Gráfica 2. Porcentaje de estudiantes por estrato socioeconómico



Fuente: elaboración propia

En lo referente a cómo los estudiantes pagan sus estudios el ICETEX tiene el mayor porcentaje de financiación de los estudios con un 58% seguido de un 27% de pago de contado, los demás porcentajes corresponden a becas de otras naturalezas, como se observa en la *Grafica 3.*

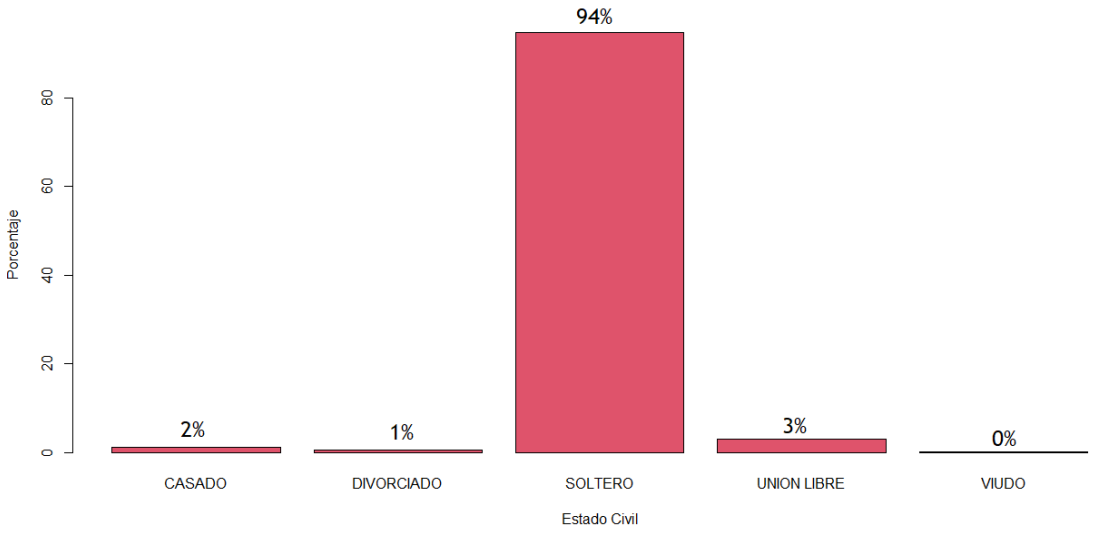
Gráfica 3. Porcentaje de cómo pagan los estudios los estudiantes



Fuente: elaboración propia

Al analizar el estado civil de los estudiantes, se observa que existe un porcentaje significativo de estudiantes con estado civil soltero (94%), el 6% restante se encuentra distribuidos entre las demás categorías, como se puede ver en la Porcentaje de estado civil de los estudiantes

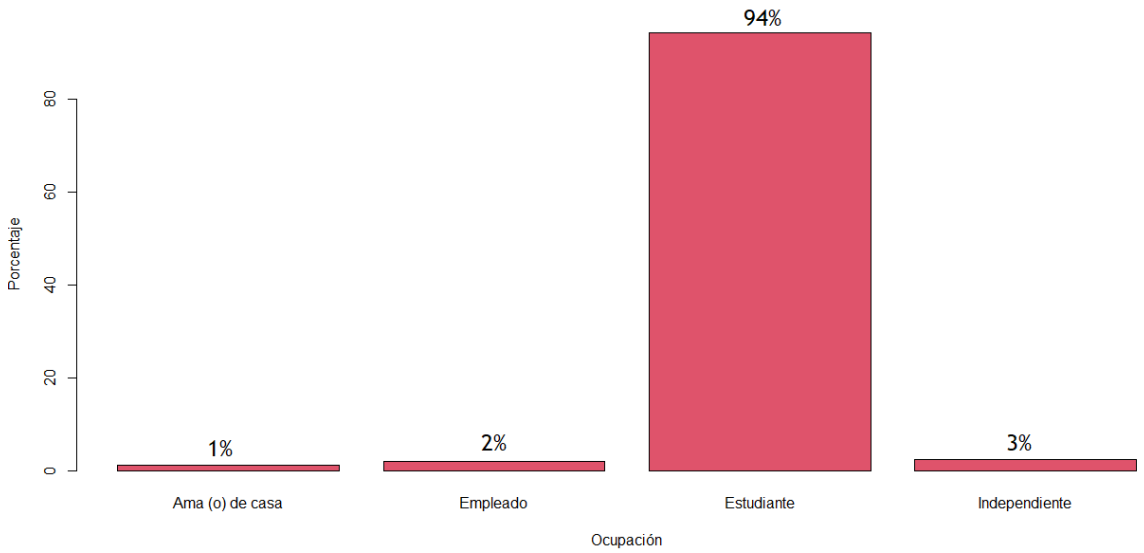
Gráfica 4. Porcentaje de estado civil de los estudiantes



Fuente: elaboración propia

Por otro lado, observamos la distribución de la ocupación de los estudiantes, donde el 94% de los estudiantes afirman dedicarse sólo a los estudios, frente a un 5% que asegura tener empleo o trabajo independiente, como se muestra en la *Gráfica 5*.

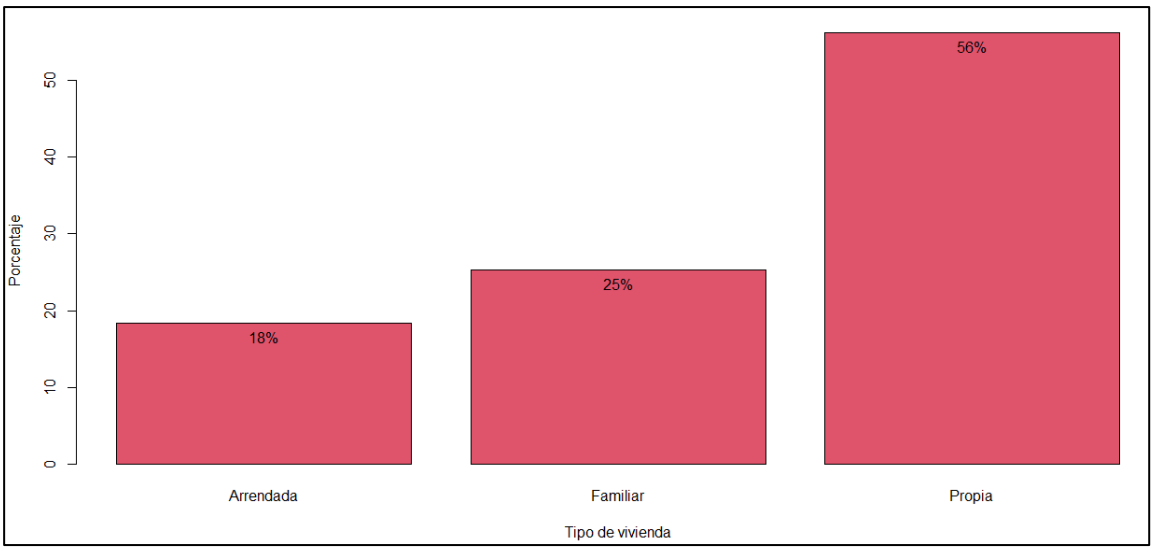
Gráfica 5. Porcentaje de ocupaciones de los estudiantes



Fuente: elaboración propia

Otra variable que se consideró dentro del estudio es el tipo de vivienda donde habitan los educandos, en el cual la mitad de los estudiantes (56%) tienen casa propia y el 25% viven con familiares, por lo que sólo el 18% debe pagar arriendo, como se muestra en la *Grafica 6.*

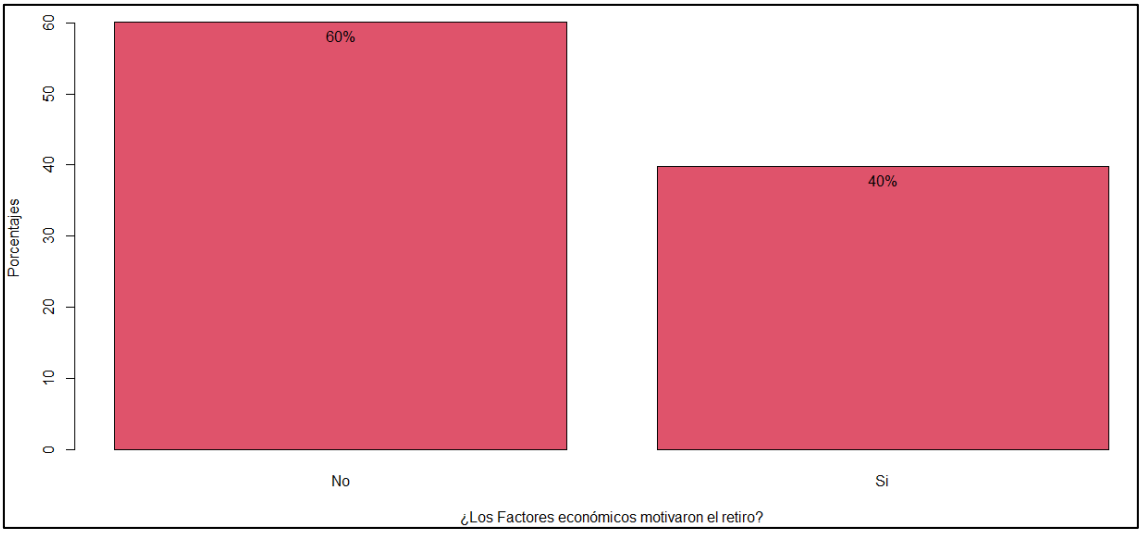
Gráfica 6. Porcentaje de tipo de vivienda de los estudiantes



Fuente: elaboración propia

En cuanto, a la proporción de si los factores económicos influyeron en el retiro de sus estudios, los datos revelan que el 60% que no fue el factor principal, mientras el 40% afirma que la principal causa de abandono fue el económico como se observa en la *Gráfica 7*.

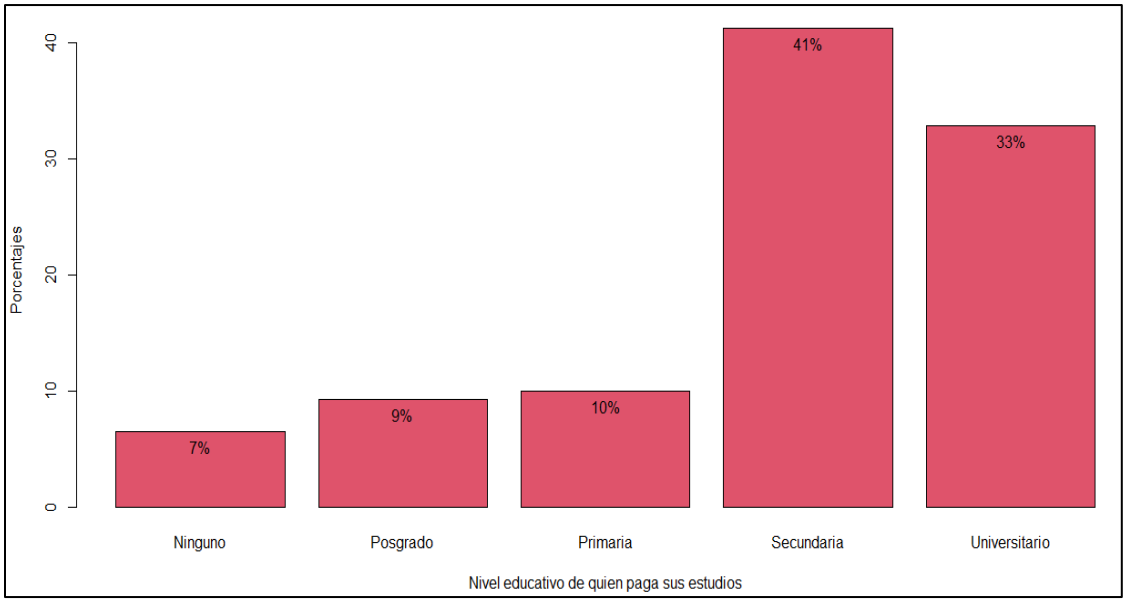
Gráfica 7. Porcentaje de motivación del retiro por factores económicos



Fuente: elaboración propia

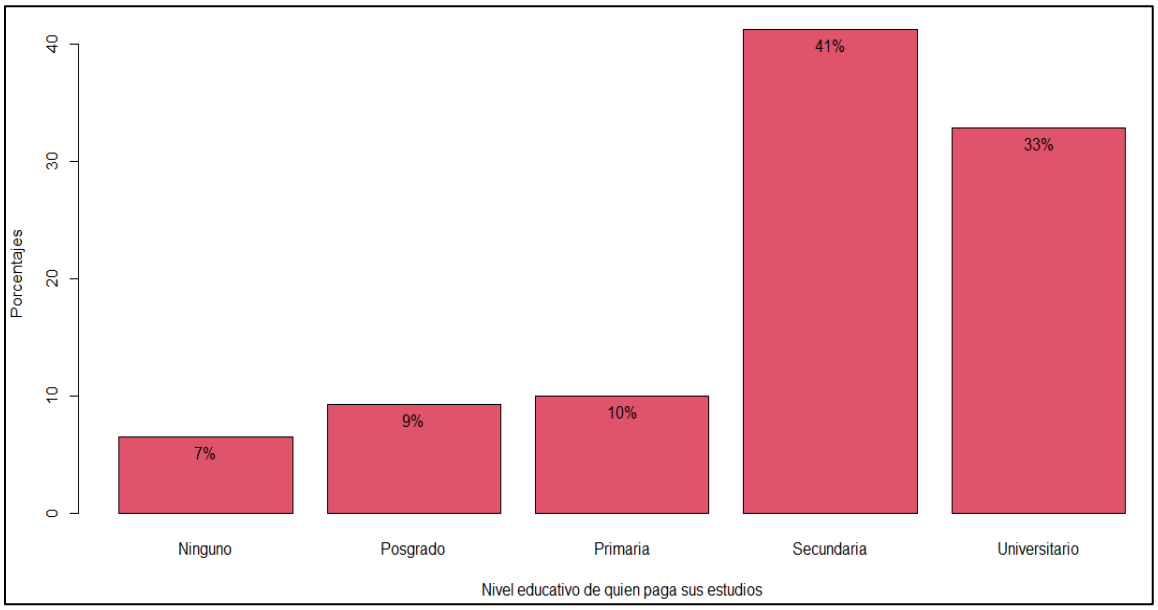
Sin embargo, al analizar el nivel educativo de los acudientes quienes pagan los estudios, nos encontramos que el 41% que son la mayoría tienen un nivel educativo de básica secundaria, seguido de un 33% de tienen universitaria y 9% posgrados, por otro lado, 7% y 10% no tienen estudios o sólo llegaron a primaria, respectivamente como se muestra en la *Grafica 8*, sin embargo, las edades de los acudientes que apoyan los estudios se distribuyen de la siguiente forma, el 53% de los estudiantes aseguran recibir apoyo económico de acudientes que tienen entre 31 y 49 años de edad, mientras que 35% de personas entre 35 y 50 años, finalmente, sólo 2% reciben apoyo de personas mayores a los 70 años ver en la *Grafica 9*. Dentro del núcleo familiar de los estudiantes, el promedio de gastos del 62% supera los 1,6 millones de pesos, lo que concuerda que la distribución socio económica de los estudiantes entre los estratos uno y dos, como se evidencia en la *Gráfica 10* y finalmente, en la *Grafica 11*, considerando el total de estudiantes el 74% de ellos no reciben incentivos económicos en sus estudios, 8.56% reciben uno y 16.44% reciben entre 2 y 10 incentivos

Gráfica 8. Porcentaje de nivel educativo de los acudientes que pagan los estudios



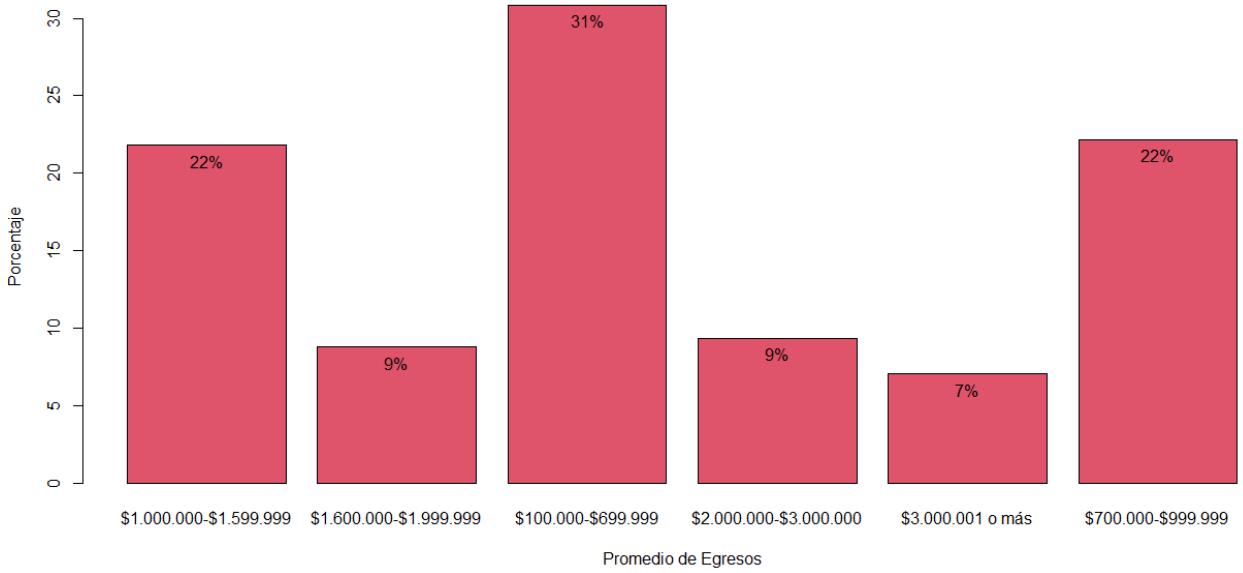
Fuente: elaboración propia

Gráfica 9. Promedio de edad de quien paga sus estudios



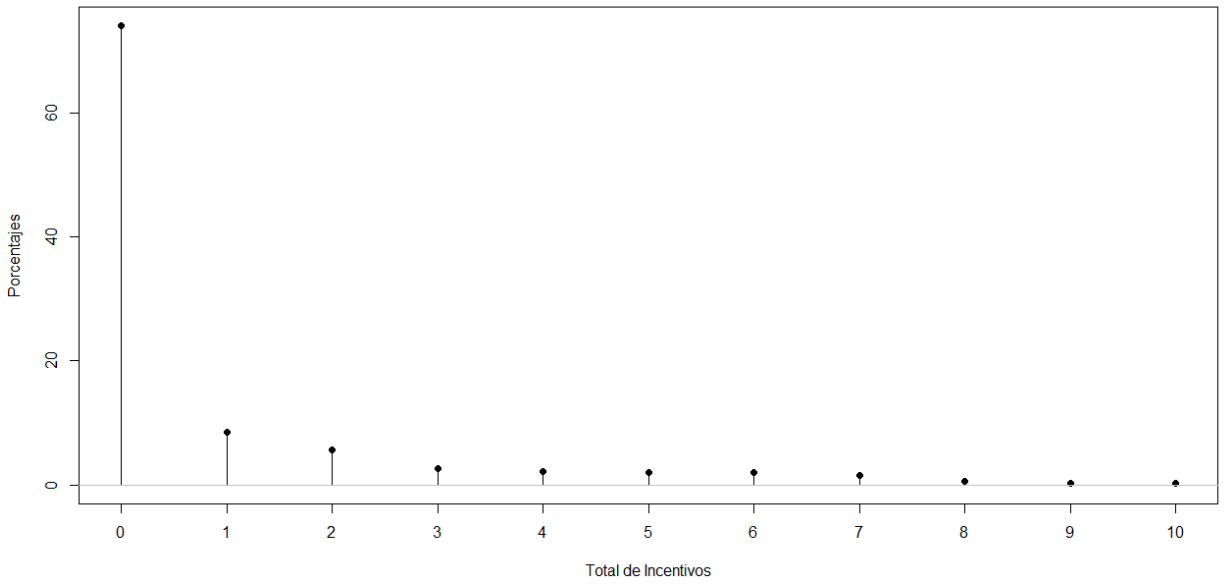
Fuente: elaboración propia

Gráfica 10. Promedio de egresos mensuales del núcleo familiar



Fuente: elaboración propia

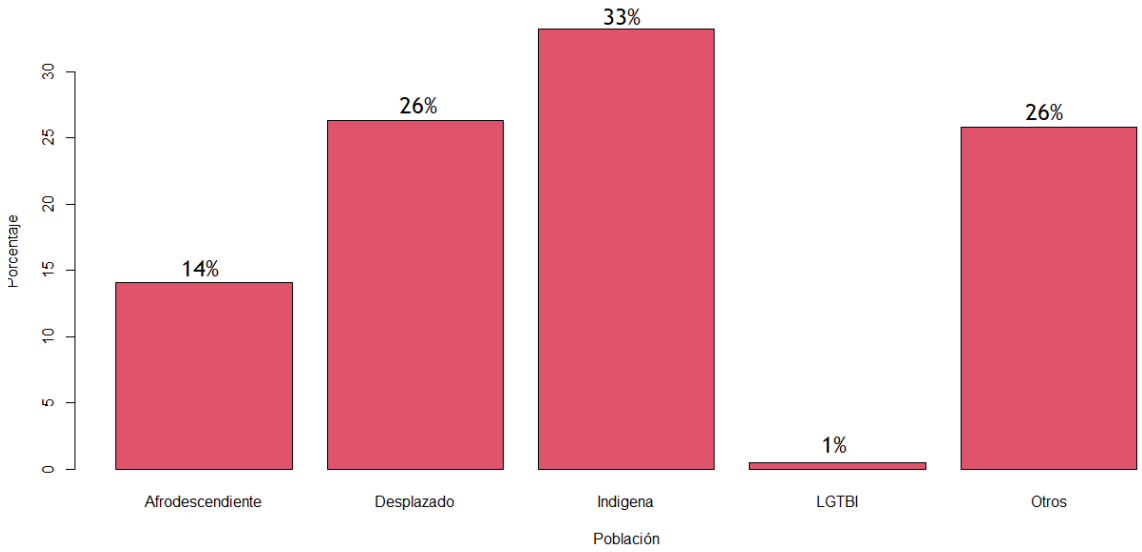
Gráfica 11. Total de incentivos económicos por porcentaje



Fuente: elaboración propia

En cuanto a la distribución por tipo de población de los estudiantes, la gran mayoría son de descendencia indígena con un 33%, afrodescendientes con 14%, desplazados con 26% y 26% son

de otras categorías, como se observa en la *Grafica 12*.

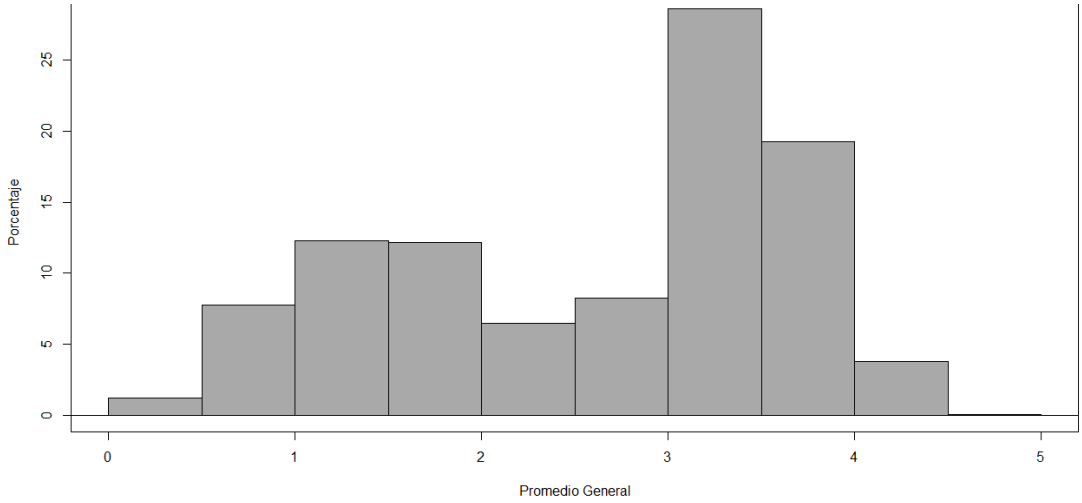


Gráfica 12. Porcentaje de estudiantes por población a la que pertenece

Fuente: elaboración propia

Finalmente se observa que la distribución del promedio general de los estudiantes está comprendida entre los rangos 3.0 y 4.0 lo cual corresponde a notas aprobatorias en la universidad, ver *Gráfica 13.*

Gráfica 13. Distribución de promedio general por porcentaje de estudiantes



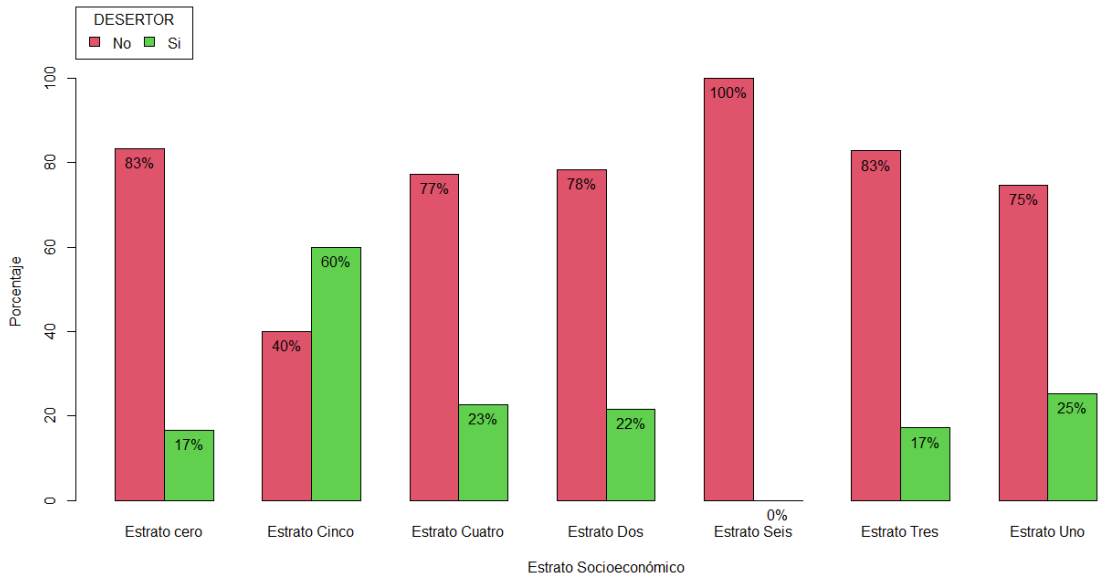
Fuente: elaboración propia

Análisis Bivariado

En esta sección se realiza una comparación entre la variable deserción y cada una de las variables predictoras, sean categóricas o cuantitativas, con el fin de identificar las posibles relaciones existentes, lo cual permite identificar patrones o incidencia de los factores sociodemográficos, educativos y financieros en la deserción universitaria de CECAR.

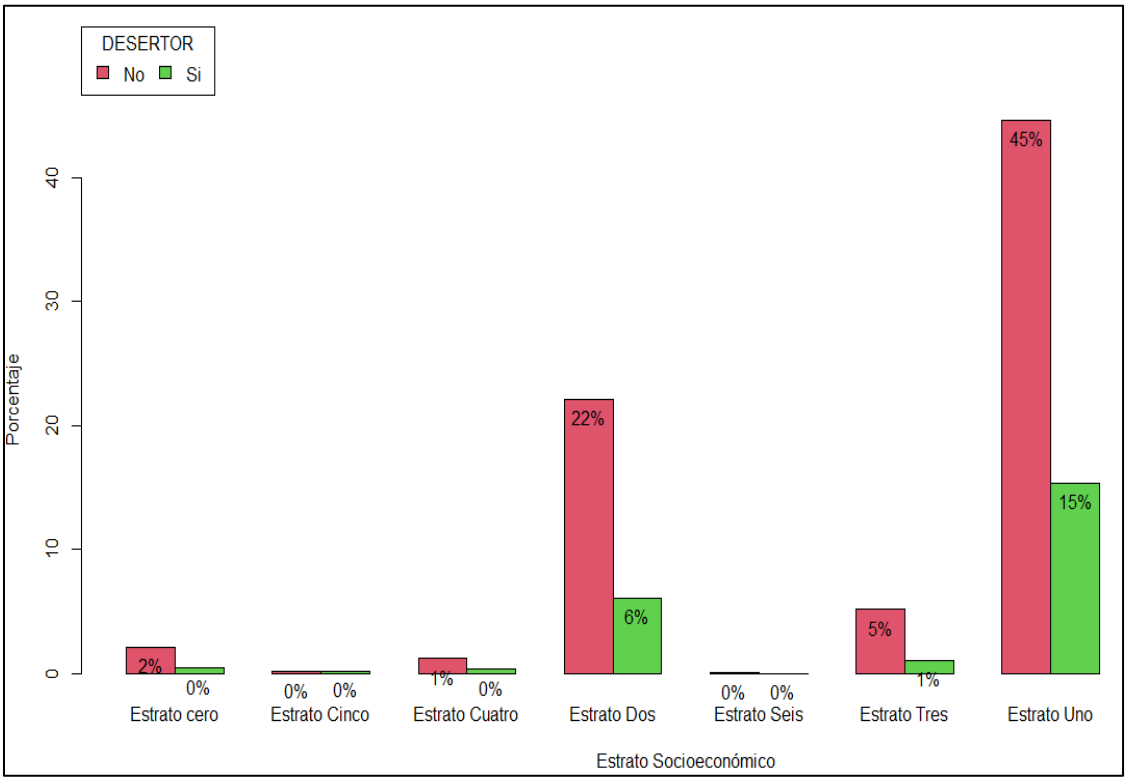
Iniciamos analizando el estrato socioeconómico frente a la deserción universitaria, donde se observa que en cada uno de los estratos socioeconómicos existe mayor porcentaje de estudiantes que siguieron con sus estudios frente aquellos que entraron en deserción, excepto el estrato 5 que 60% decidieron abandonarlos. Es importante resaltar que estudiantes de estratos 0, 1 y 2 tienen porcentajes de permanencias significativos (83%, 74% y 78%), finalmente, los estudiantes de estrato 6 ninguno desertó, como se muestra en *Gráfica 14*. De igual forma se analiza la deserción por porcentaje del total y esto muestra que el estrato uno presenta mayor número de estudiantes en deserción escolar con un 15% del total, seguido del estrato dos con 6%, los estratos 4, 5 y 6 sólo presentan menos del 1% de deserción, lo que es razonable dado que la mayor proporción de estudiantes pertenecen a esos estratos y perciben menos ingresos frente a estratos superiores a 4 como se muestra en la *Grafica 15.*

Gráfica 14. Porcentaje de deserción por estrato socioeconómico



Fuente: elaboración propia

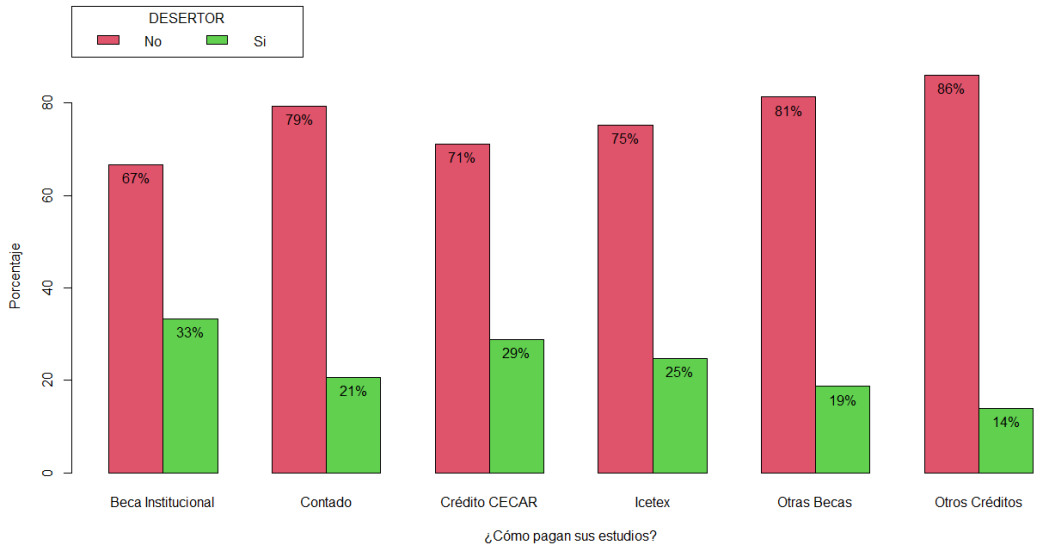
Gráfica 15 *Porcentaje total de deserción por estrato socioeconómico*



Fuente: elaboración propia

A continuación, se analiza la proporción de desertores frente a como pagan sus estudios, y se observa que los mayores porcentajes de deserción se encuentran en aquellos estudiantes que tienen créditos con ICETEX con un 14% sobre el total, seguido de 6% de aquellos que pagan de contado y 2% de quienes tienen crédito CECAR, lo cual va en concordancia con la capacidad de pago de los estudios como se evidencia en la *Gráfica 16*

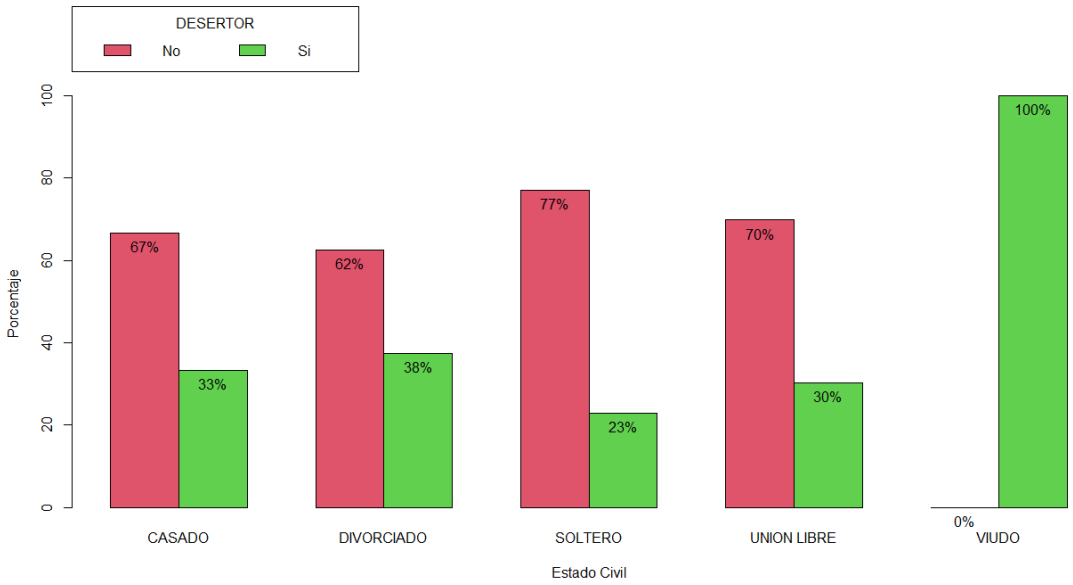
Gráfica 16. Porcentaje total de deserción por forma de pago de los estudios



Fuente: elaboración propia

Los datos también arrojan un hallazgo significativo el cual nos muestra que el estado civil en el cual más estudiantes desertan es el soltero a razón del 21 %, dado que va en relación con la preponderancia de la edad promedio de los estudiantes y la cantidad de estudiantes solteros como se observa en la *Gráfica 17*.

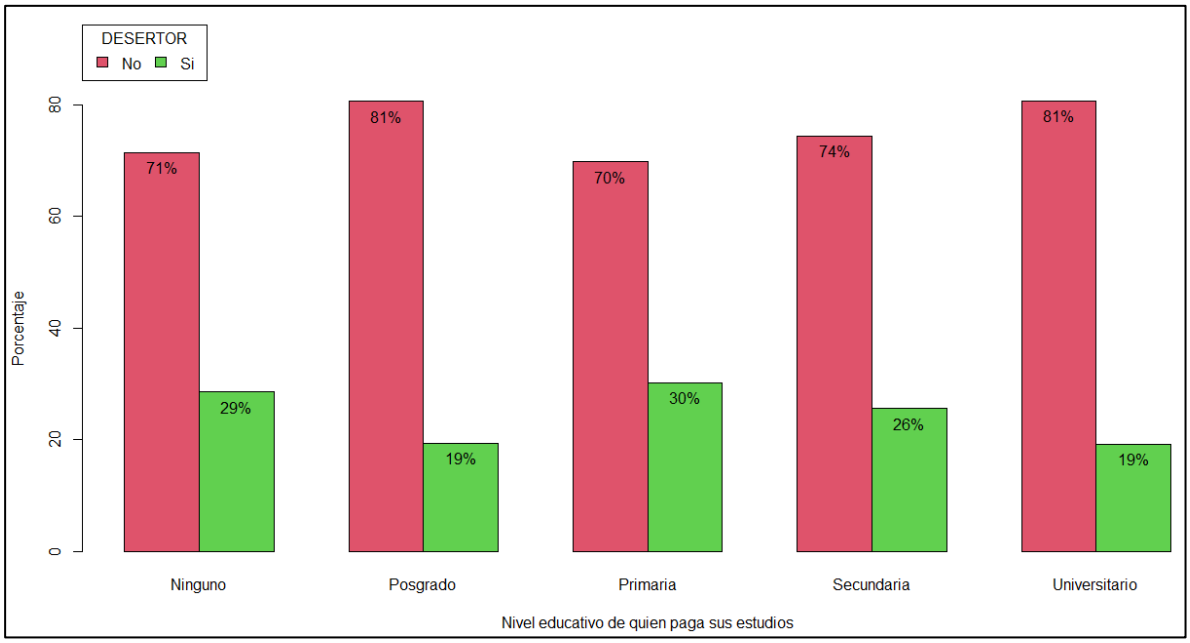
Gráfica 17. Porcentaje total de deserción por estado civil



Fuente: elaboración propia

Por otro lado, no se observan diferencias significativas del nivel educativo de las personas que pagan los estudios frente a la deserción dado que estas oscilan entre 20% y 30% según las categorías, se destaca el porcentaje más alto de 30% los del nivel primaria, como se observa en la *Gráfica 18*, de igual forma si analizando la edad promedio de los acudientes que pagan los estudios tampoco se observa diferencias significativas dado que se agrupan entre 22% y 25% de las categorías como se muestra en la *Gráfica 19*

Gráfica 18. Porcentaje total de deserción por nivel educativo del que paga los estudios

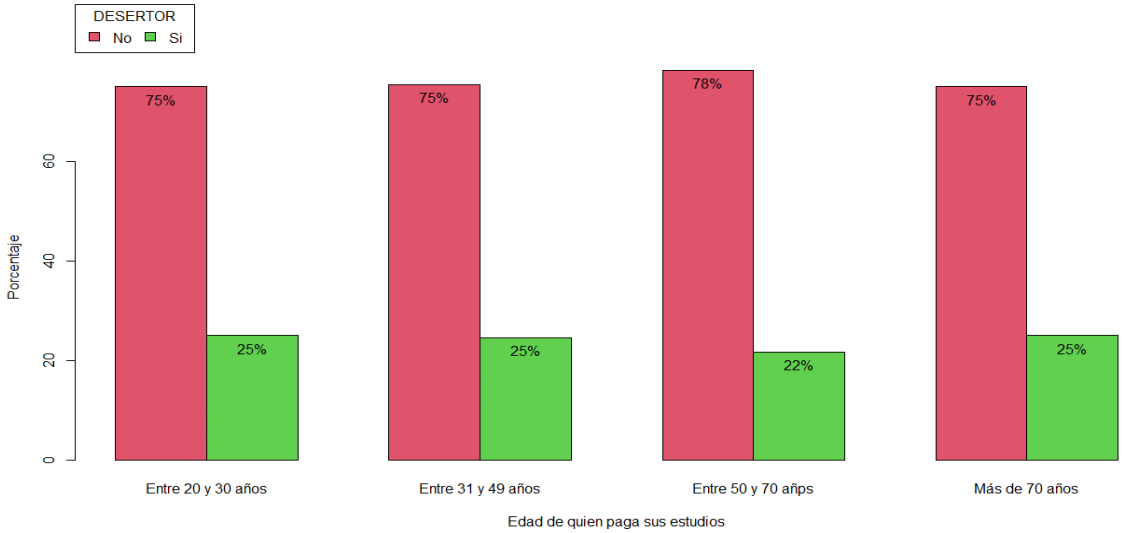


Fuente: elaboración propia

Gráfica 19. Porcentaje total de deserción por edades del que paga los estudios

Fuente: elaboración propia

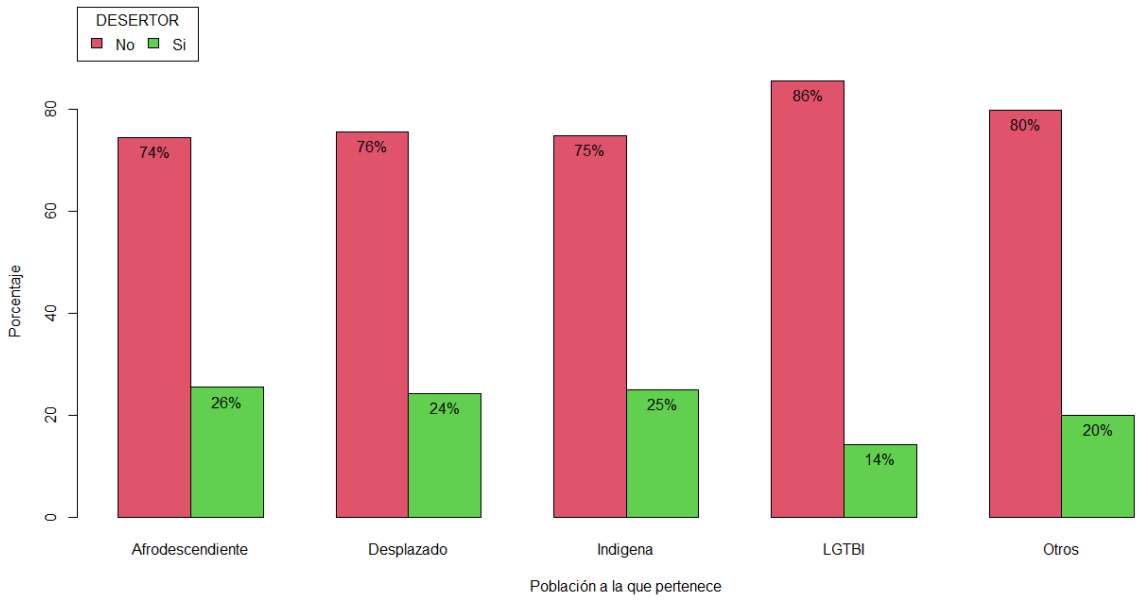
Con



respecto a la población a la que pertenecen los estudiantes, no se evidencia tendencias predominantes y estas obedecen a la relación de mayor proporción de estudiantes en esa población y por ende la pertenencia mayor de desertores, donde el grupo poblacional de los indígenas el 8% deserta, mientras el 5% de los desplazados, 4% para los afrodescendientes y 2% para estudiantes víctimas del conflicto armado, como se muestra en la

Gráfica **20**.

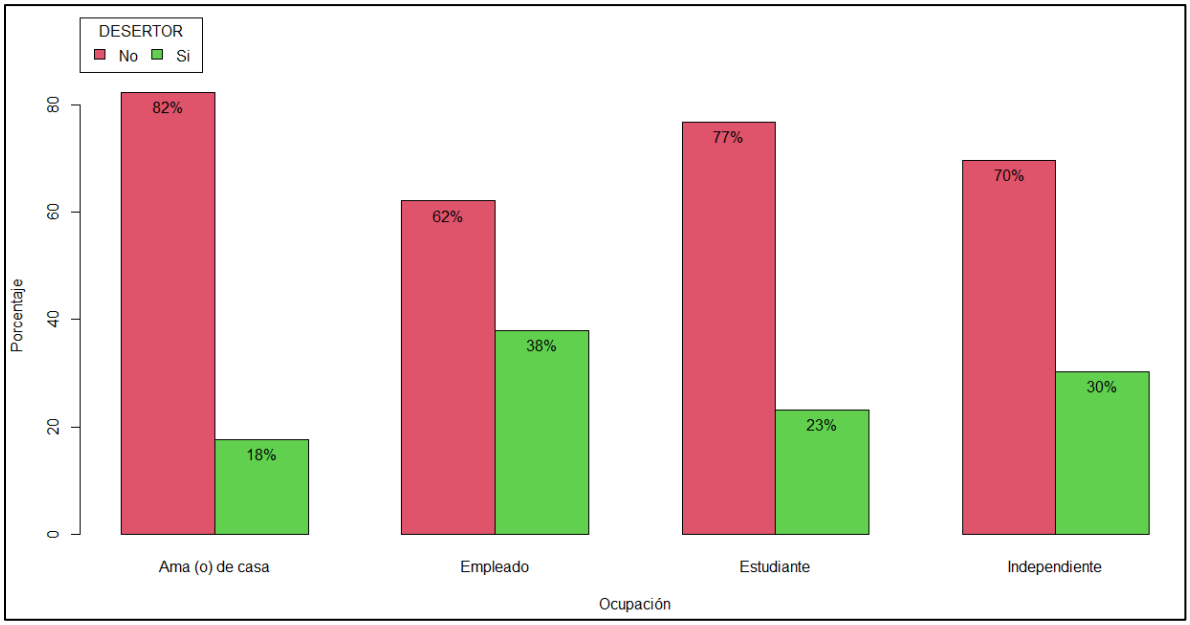
Gráfica 20. Porcentaje total de deserción por población de pertenencia de los estudiantes



Fuente: elaboración propia

Otro aspecto importante analizado fue la distribución de la deserción universitaria frente a la ocupación del estudiante, aquí se observa un patrón considerable, donde las personas que tienen empleo tanto informal como formal presentan mayor porcentaje de desertar, los que se categorizan como empleado desertan en un 38%, los trabajadores independientes en un 30% y frente a los que se dedican a solo estudiar con un 23% y trabajo como ama(o) de casa con un 18% ver *Gráfica 21*.

Gráfica 21. Porcentaje total de deserción por situación laboral



Fuente: elaboración propia

Otra variable que presenta relación significativa frente al fenómeno de deserción es la cantidad de asignaturas aprobadas como se observa en la *Tabla 9* y la *Grafica 17* Los estudiantes con promedio de 27 materias aprobada tienen mayor porcentaje de deserción frente a aquellos con promedio de 47 asignaturas aprobadas, eso hace pensar que la deserción tiene mayor frecuencia en los primeros semestres de la carrera, esto se evidencia en los gráficos de caja y medias donde se observa la diferencia entre las medias aritméticas de cada grupo, siendo inferior en aquellos que entraron en deserción. De igual forma el promedio general del estudiante se ve afectado por

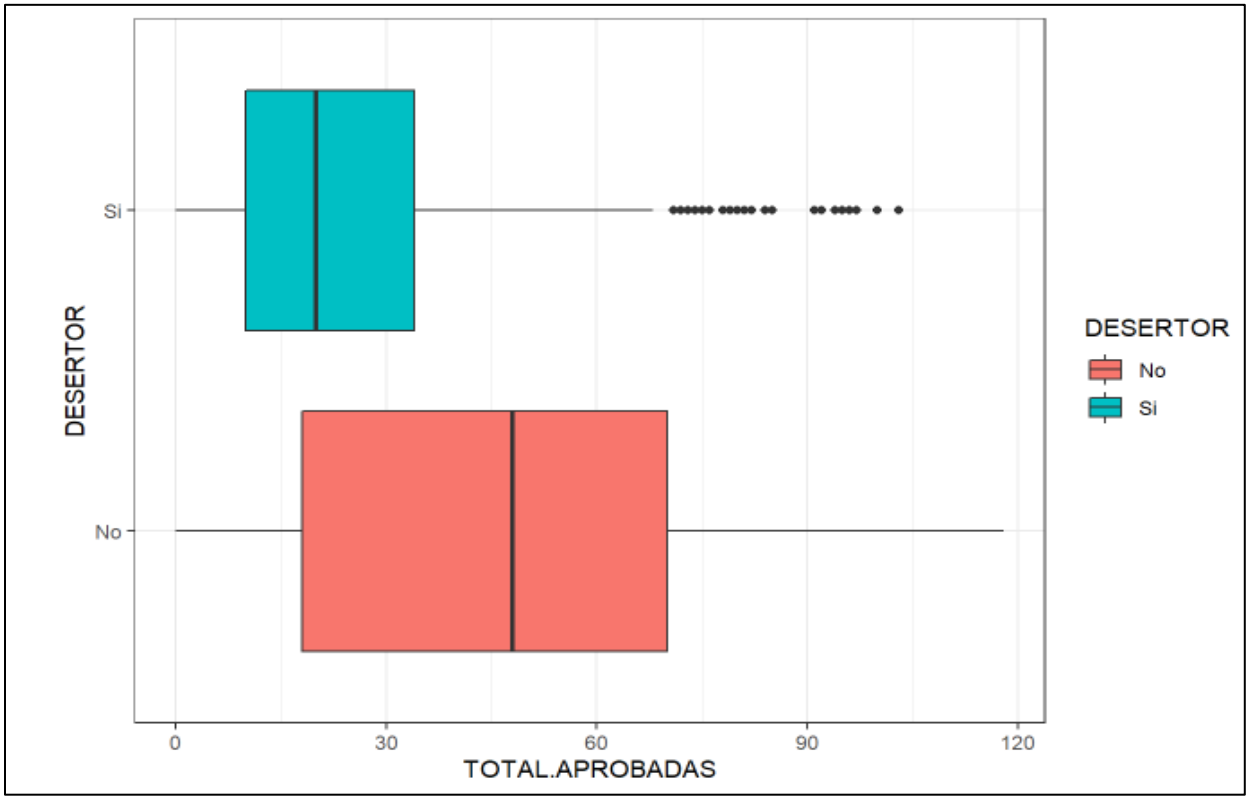
el total de asignaturas aprobadas y desaprobadas, como se muestra en *Gráfica 23*, en el cual los estudiantes con promedios de 3.0 o menores tienden a entrar en deserción, es importante mencionar que es el puntaje mínimo aprobatorio semestralmente.

Tabla 9. Número de asignaturas aprobadas por desertores

Deserción	Media	CV	Total
No	47	63%	1062
Si	27	88%	328

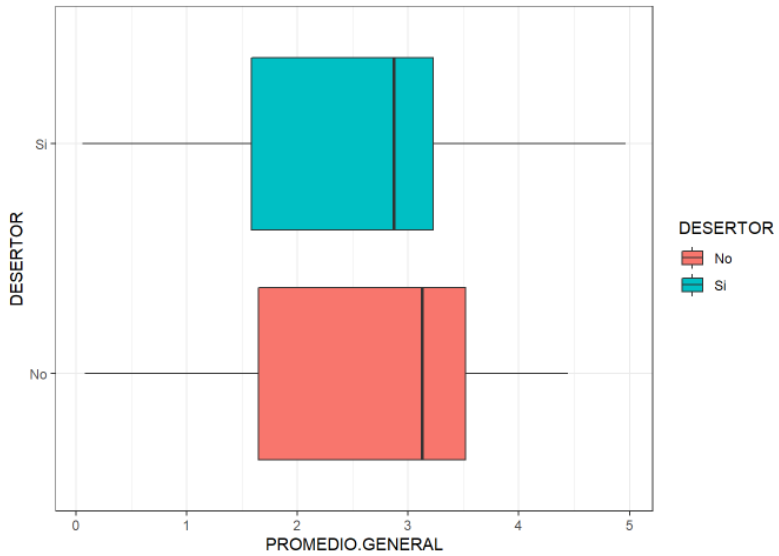
Fuente: elaboración propia

Gráfica 22. Diagrama de caja número de asignaturas aprobadas vs deserción



Fuente: elaboración propia

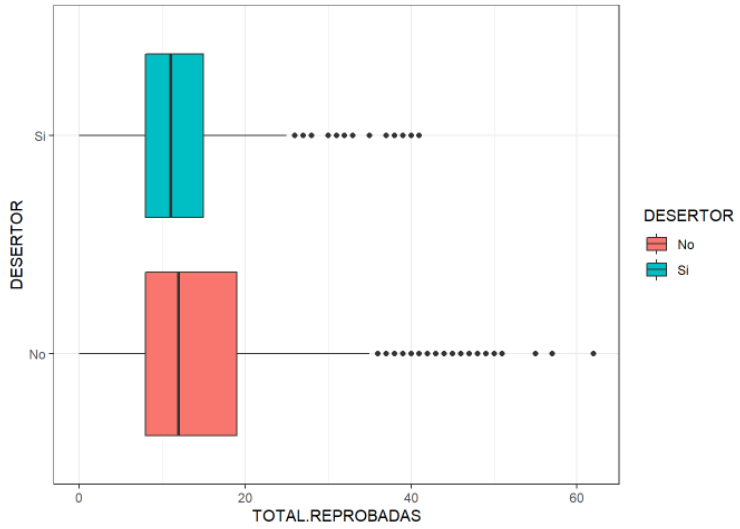
Gráfica 23. Diagrama de caja promedio general vs deserción



Fuente: elaboración propia

Continuando con el análisis se observa que el total de asignaturas reprobadas tiene directa sobre la deserción escolar, sin embargo, los datos y los gráficos revelan que no hay diferencia entre los promedios del número de asignaturas reprobadas en desertores y no desertores, como se observa en la Gráfica 24 “Diagrama de caja número de asignaturas reprobadas vs deserción”.

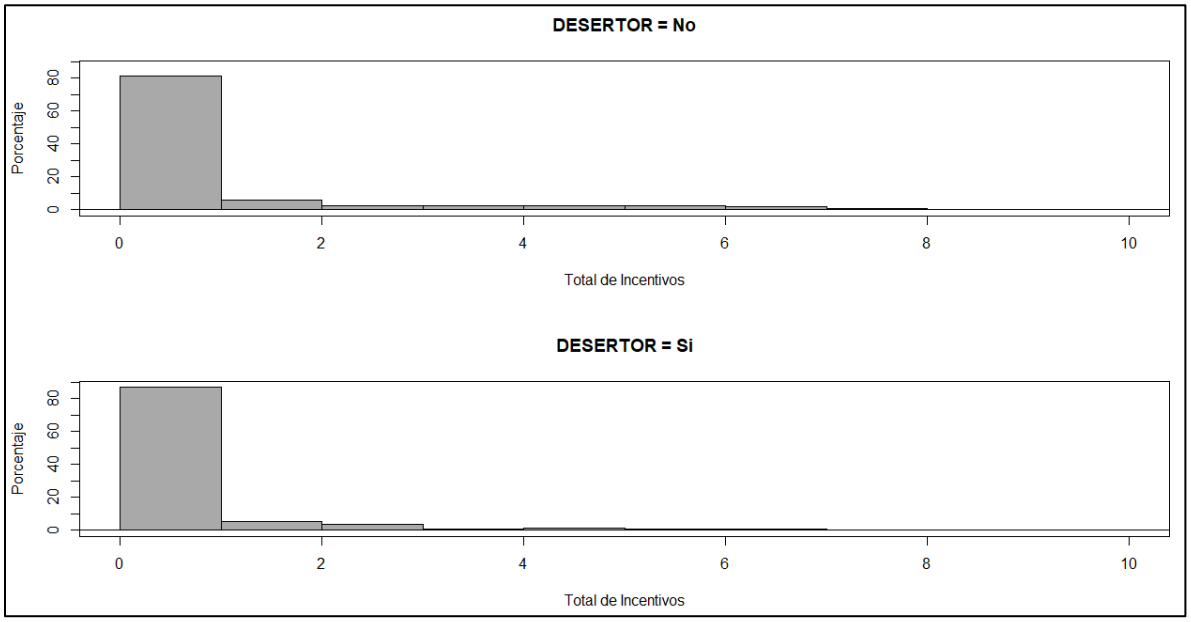
Gráfica 24. Diagrama de caja número de asignaturas reprobadas vs deserción



Fuente: elaboración propia

Finalmente, se analiza la incidencia de los incentivos académicos de los programas de Ingeniería de Sistemas e Industrial en la deserción, en el cual para el caso de estudio la mayoría de los estudiantes no tienen incentivos académicos, como se muestra en la *Gráfica 25* y en relación con la deserción se observan diferencias significativas que a mayor número de incentivos mejora la posibilidad de no desertar.

Gráfica 25. Porcentaje de desertores por cantidad de incentivos económicos



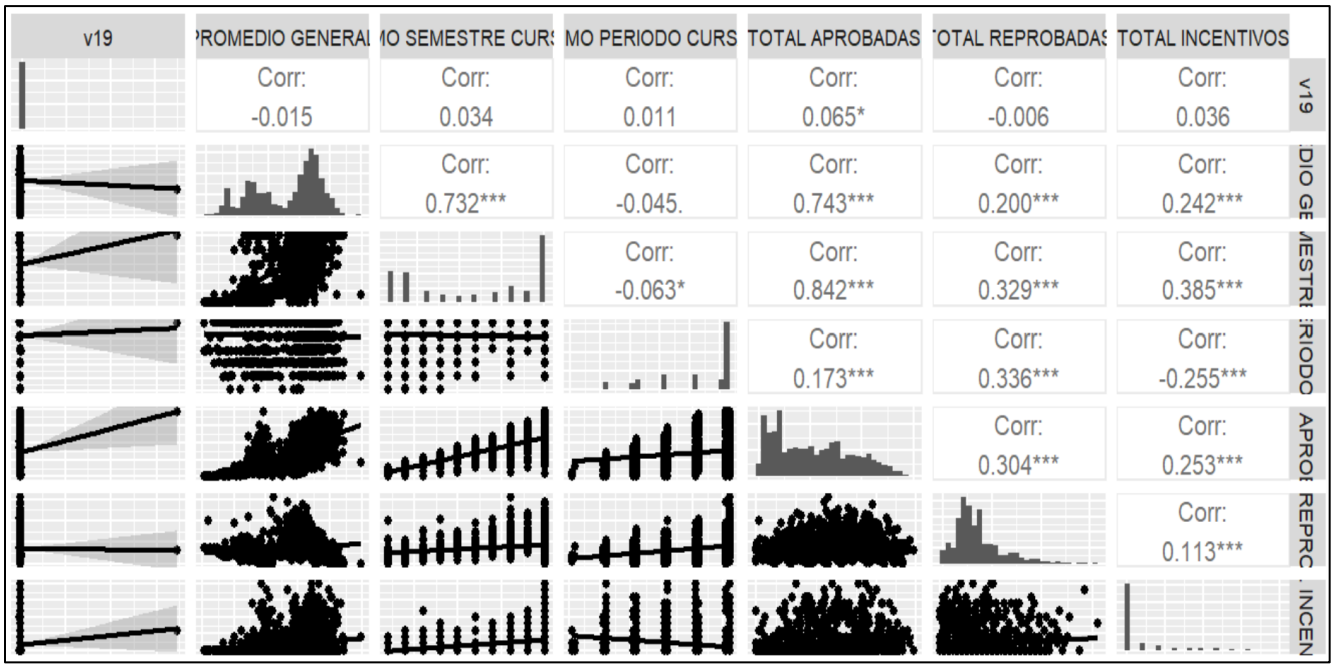
Fuente: elaboración propia

7.2.3 Selección de variables para el modelo

Para el proceso de selección de variables se empieza validando si dentro de las variables explicativas de la deserción están altamente relacionadas entre sí, es decir, presentan multicolinealidad entre ellas, para esto del conjunto de datos resultante se extraen las variables numéricas y se aplica una prueba de correlación de Pearson como se muestra en la

Gráfica 26 donde se observa que las variables total aprobadas y último semestre cursado tiene un alto grado de correlación pero es explicativo dado que a mayor número de semestres que los estudiantes cursen mayor será el total de asignaturas aprobadas y desaprobadas, sin embargo, no se elimina estas dos variables dado que para la predicción de la deserción el semestre cursado y la cantidades de asignaturas aprobadas y desaprobadas son un factor determínate para entender la deserción universitaria, dado que se esperaría que a más semestres cursados exista menos probabilidad de deserción, datos que son respaldados con en análisis de deserción universitario anexo 1 Informe autoevaluación facultad de ciencias básicas v 2.5

Gráfica 26. Correlación entre variables numéricas



Fuente: elaboración propia

Para la selección de las variables que expliquen mejor la deserción escolar se usaron dos técnicas la regresión y árboles aleatorios, donde la primera se corre una regresión logística con las 35 variables predictoras y la variable respuesta deserción, como se muestra en la *Ilustración 7* donde se observa que este modelo cataloga a 10 variables como las que mejor explican la variable

objetivo deserción, a partir de esas variables, se ejecutan 10 regresiones y se compara la métrica del accuracy para determinar las mejores predictoras como se muestra en la *Tabla* , donde se toma el valor más alto de las métricas y se seleccionan como mejores predictoras las siguientes variables: *MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+`TOTAL INCENTIVOS`+v19+v46.*

Ilustración 7 .Resumen del modelo regresión logística 35 variables

```
1390 samples
  35 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1112, 1112, 1112, 1112, 1112
Resampling results:

  Accuracy   Kappa
  0.8827338  0.6480845

glm variable importance

  only 20 most important variables shown (out of 96)

glm variable importance

  only 20 most important variables shown (out of 96)
```

	Overall <dbl>
`\\`MAXIMO PERIODO CURSADO\\``	100.000000
`\\`ULTIMO SEMESTRE CURSADO\\``	78.277581
`\\`TOTAL REPROBADAS\\``	67.889981
`\\`PROMEDIO GENERAL\\``	50.346313
`\\`TOTAL APROBADAS\\``	24.901176
v78Si	15.427294
v79Si	14.507912
`\\`TOTAL INCENTIVOS\\``	14.449183
`v1Crédito con CECAR`	14.128483
`v69No aplica`	13.554939
`v21Entre $100.000 y 699.999`	12.470597
`v1Beca o Auxilio de Bienestar Institucional`	12.424225
`v52Habitante de frontera`	11.551122
`v21Entre $700.000 y $999.999`	11.254565
v82Familiar	10.820398
`v29Estrato Cinco`	10.533550
v61Si	10.476172
`v21Entre $2.000.000 y $3.000.000`	10.458563
v34Masculino	10.122983
`v29Estrato Cuatro`	9.976538

Fuente: elaboración propia

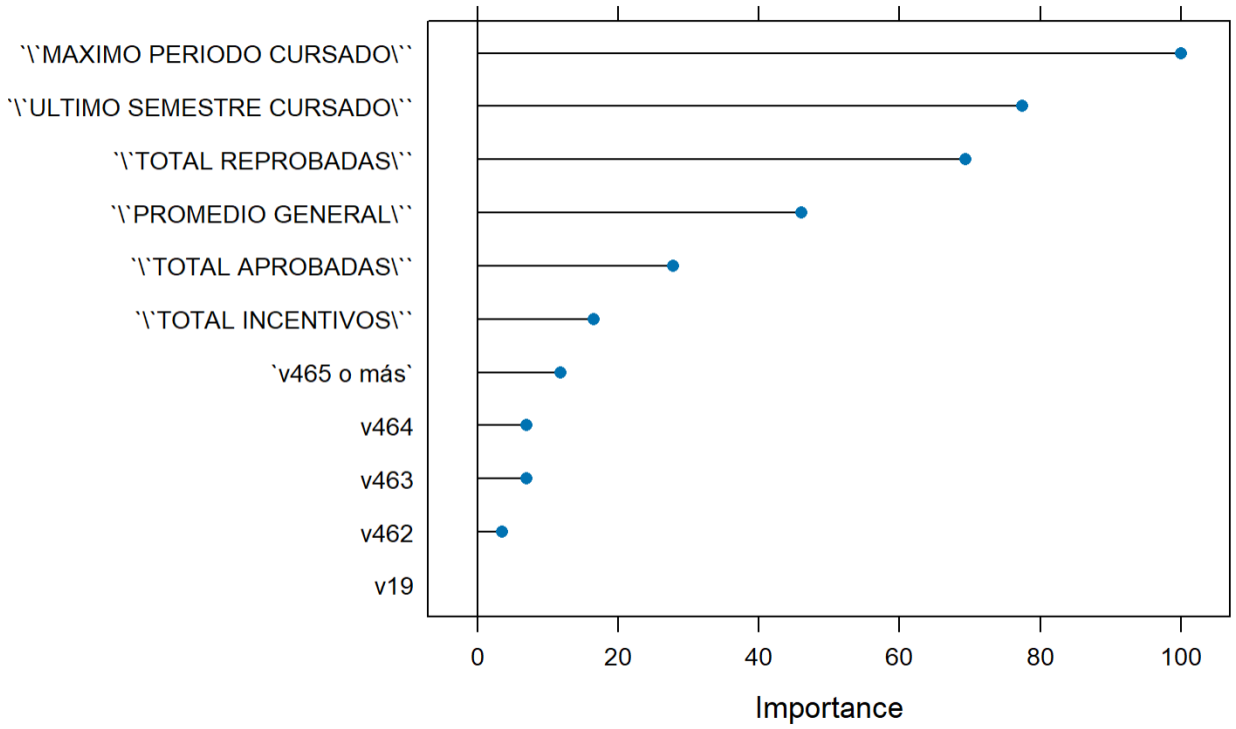
Tabla 10. Métricas de accuracy y ecuaciones por cada modelo de regresión

Accuracy	Kappa	Ecuación
0,8827338	0,6480845	DESERTOR~.
0,9129487	0,7322479	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+v78
0,9129496	0,7314488	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+v78+v79
0,9114922	0,7269984	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+v78+v79+`TOTAL INCENTIVOS`
0,9086482	0,7192479	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+v78+v79+`TOTAL INCENTIVOS`+v1+v21
0,9072016	0,7151886	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+v78+v79+`TOTAL INCENTIVOS`+v1+v21+v52

0,9086331	0,7193985	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+v78+v79+`TOTAL INCENTIVOS`+v1+v21+v52+v82
0,9064691	0,7148853	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+v78+v79+`TOTAL INCENTIVOS`+v1+v21+v52+v29+v34
0,9107569	0,7264214	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+v78+v79+`TOTAL INCENTIVOS`+v1+v21+v29
0,9129698	**0,7329681**	DESERTOR~`MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+`TOTAL INCENTIVOS`+v19+v46

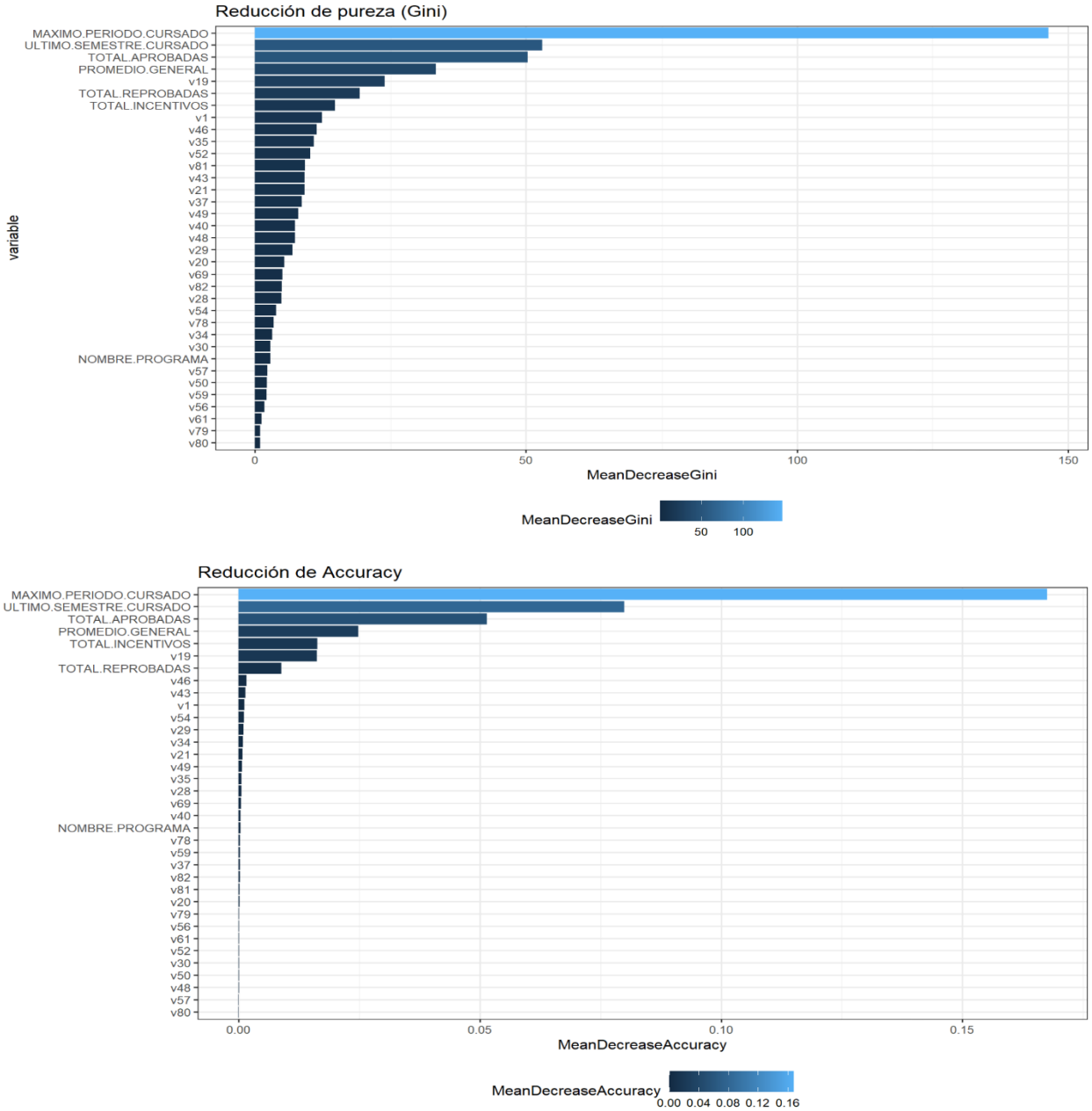
El modelo resultante con las 7 variables predictoras arroja un accuracy de 90% y Kappa de 0,72 mediante el método de validación cruzada como se muestra en la *Ilustración 8*, se puede observar la significancia de las variables y el criterio de información de Akaike (AIC) de calidad del modelo en 853.66 el cual determina que entre menos complejo el modelo es mejor.

Ilustración 8. Resumen del modelo regresión logística 7 variables por importancia



Fuente: elaboración propia

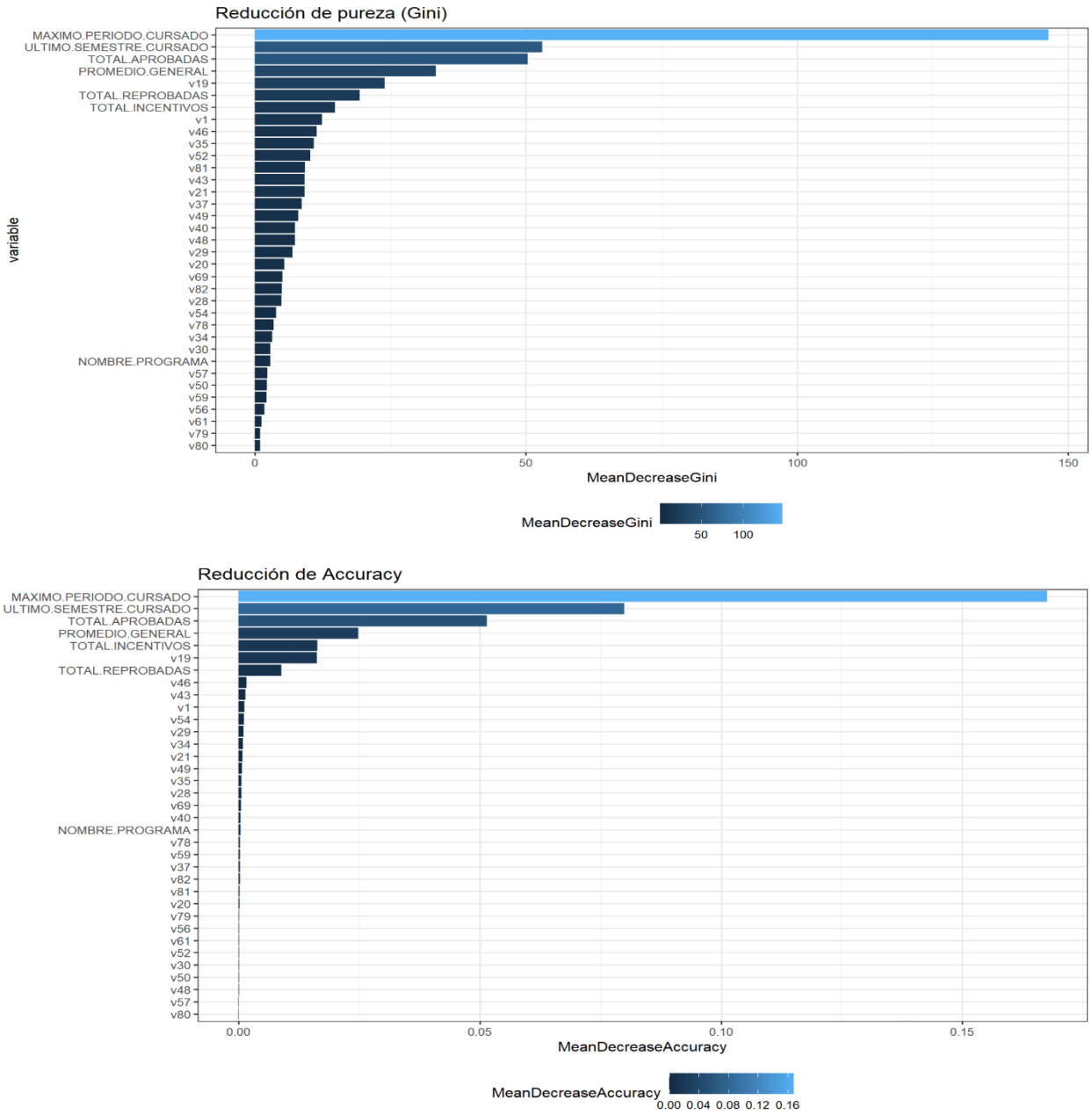
Seguido a esto, se usa la técnica de árboles aleatorios para contrastar las variables con mejor poder de predicción del paso anterior y nos arroja un valor muy similar como se muestra en la **Gráfica 27**. Lista de variables ordenadas de mejor a peor desempeño predictor de la deserción



, el cual confirma que las variables `MAXIMO PERIODO CURSADO`+`ULTIMO SEMESTRE CURSADO`+`TOTAL REPROBADAS`+`PROMEDIO GENERAL`+`TOTAL APROBADAS`+`TOTAL

INCENTIVOS`+v19+v46, donde v19 es la edad del estudiante y v46 corresponde al número de hermanos, son las mejores predictoras para la deserción universitaria del conjunto de datos analizado.

Gráfica 27. Lista de variables ordenadas de mejor a peor desempeño predictor de la deserción



Fuente: elaboración propia

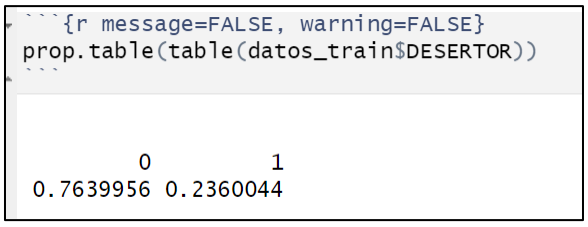
7.3 Implementación del modelo

Con base a la selección de las mejores variables predictoras, descritas en el punto anterior, se crea un nuevo conjunto de datos con estas variables y, a partir de este, se entrenan cuatro modelos de los cuáles el primero es de enfoque estadístico correspondiente a una regresión logística y tres de aprendizaje automático como las máquinas de soporte vectorial, arboles aleatorios de decisión y una red neuronal simple.

7.3.1 División de los datos en entrenamiento

Del conjunto de datos seleccionado con las mejores predictoras se dividen en dos grupos el 65,5% para entrenamiento de los modelos y el 34,5% para prueba y se valida que la distribución de los valores de deserción es homogénea del conjunto de datos completo como se muestra en la *Ilustración 9*.

Ilustración 9. Distribución de valores de deserción datos de entrenamiento



```
```{r message=FALSE, warning=FALSE}
prop.table(table(datos_train$DESERTOR))
```

        0   1
0.7639956 0.2360044
```

Fuente: elaboración propia

7.3.2 Preprocesado de los datos

Para garantizar el buen ajuste del modelo se realiza un pre-procesado de los datos a nivel de estandarización y escalado de las variables cuantitativas y binarización de las variables cualitativas, como se observa en la

Tabla 10

Tabla 10. Resumen de los datos estandarizados y binarizados

```
Rows: 911
Columns: 9
$ MAXIMO.PERIODO.CURSADO   <dbl> -2.186883799, 0.006339724, -1.455809291, -0.066767727, 0.737414232, -…
$ ULTIMO.SEMESTRE.CURSADO  <dbl> -1.10665532, 1.07669714, 1.07669714, 1.07669714, 0.53085902, 1.076697…
$ TOTAL.REPROBADAS         <dbl> -1.4614135, -1.4614135, -1.3575156, -1.3575156, -0.1107408, -1.253617…
$ PROMEDIO.GENERAL         <dbl> 2.233791, 1.739891, 1.729354, 1.641890, 1.579423, 1.557299, 1.548169,…
$ TOTAL.APROBADAS          <dbl> -0.5445700, 2.2878010, 0.7030220, 1.7820204, 1.5122708, 1.4111147, -0…
$ TOTAL.INCENTIVOS         <dbl> 0.61121746, 0.07735446, -0.45650854, 3.28053246, -0.45650854, 2.21280…
$ EDAD                     <dbl> 3.40663684, -0.32936034, 0.37113913, 0.37113913, -0.56286017, 0.13763…
$ NUM.HERMANOS             <dbl> 2.3761540, -1.0188841, -0.3398765, -1.0188841, -0.3398765, -1.0188841…
$ DESERTOR                 <fct> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, …
```

Fuente: elaboración propia

Finalmente, se comprueba si el conjunto de datos no tiene variables con varianza próximas a cero dado que estas no aportarían un buen ajuste al modelo, como se observa en la *Tabla 11*

Tabla 11. Comprobación de variables con varianza próxima a cero

	freqRatio <dbl>	percentUnique <dbl>	zeroVar <lgl>	nzv <lgl>
MAXIMO.PERIODO.CURSADO	6.704762	0.8633094	FALSE	FALSE
ULTIMO.SEMESTRE.CURSADO	2.165138	0.7194245	FALSE	FALSE
TOTAL.REPROBADAS	1.028571	3.9568345	FALSE	FALSE
PROMEDIO.GENERAL	1.250000	95.6834532	FALSE	FALSE
TOTAL.APROBADAS	1.085714	8.2733813	FALSE	FALSE
TOTAL.INCENTIVOS	8.647059	0.7913669	FALSE	FALSE
EDAD	1.540284	2.8776978	FALSE	FALSE
NUM.HERMANOS	1.035623	0.3597122	FALSE	FALSE

8 rows

Fuente: elaboración propia

7.3.3 Modelo de regresión logística

Una vez preprocesados los datos se entrenan los modelos con los hiperparametros, número de repeticiones y semillas por cada repetición y, finalmente, se selecciona el mejor modelo a partir de la validación cruzada, como resultado se obtiene de 911 muestras y las 8 variables predictoras un accuracy de 0.92 y un kappa de 0.74, donde se concluye que con el modelo tiene un 92% de exactitud para predecir la deserción, como se muestra en la **Ilustración 10**.

Ilustración 10. Resumen del modelo de regresión logística 8 predictoras de la clase deserción

```
Generalized Linear Model

911 samples
  8 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 819, 821, 820, 820, 820, 820, ...
Resampling results:

  Accuracy   Kappa
  0.9178947  0.7482142
```

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5255  -0.4636  -0.3211  -0.0889   4.0115

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.70016    0.12353 -13.763  < 2e-16 ***
MAXIMO.PERIODO.CURSADO   -2.25987    0.19483 -11.599  < 2e-16 ***
ULTIMO.SEMESTRE.CURSADO  -2.95194    0.33706  -8.758  < 2e-16 ***
TOTAL.REPROBADAS          1.16871    0.13602   8.592  < 2e-16 ***
PROMEDIO.GENERAL          1.13984    0.18937   6.019 1.75e-09 ***
TOTAL.APROBADAS           0.83028    0.30896   2.687   0.0072 **
TOTAL.INCENTIVOS         -0.35126    0.14586  -2.408   0.0160 *
EDAD                     -0.01655    0.12670  -0.131   0.8961
NUM.HERMANOS              0.17296    0.10987   1.574   0.1154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 995.60  on 910  degrees of freedom
Residual deviance: 525.71  on 902  degrees of freedom
AIC: 543.71

Number of Fisher Scoring iterations: 6
```

Fuente: elaboración propia

7.3.4 Modelo de máquina de soporte vectorial (SVM)

Para la selección del mejor modelo se usa la técnica de resampling a través de validación cruzada y el tuneo de los hiperparámetros donde se entrenan dos modelos svm Lineal y svm Radial, para el primero se establece el hiperparámetro de regularización(C) con valores de (0.001, 0.01, 0.1, 0.5, 1, 10) y se define la grilla de hiperparámetros de la siguiente forma: hiperparametros = data.frame(C = c(0.001, 0.01, 0.1, 0.5, 1, 10)) y se establece el número de repeticiones y semillas para cada repetición, donde se tienen 10 particiones y 5 repeticiones, el cual se utiliza para realizar la validación cruzada y este valor obedece a que el conjunto de datos es pequeño donde usualmente se toma valores de 5 a 10 particiones y para obtener una validación más robusta se establece el numero de 5 repeticiones, lo que significa que el modelo realizara la validación cruzada 5 veces con diferentes particiones aleatorias del conjunto de datos.

Seguido a este paso se selecciona los mejores hiperparámetros después de evaluar todas las combinaciones de la regularización (C), el cual nos arroja que el mejor modelo es el que se determina con el valor de C=10, como se muestra en la **Ilustración 11** donde se observan los valores obtenidos por cada partición y repetición, de igual forma en la se puede observar la

evolución de los valores de accuracy a partir de C donde el valor de 10 representan mejores valores de exactitud y disminución del error, finalmente se tomó el mejor modelo con métricas de accuracy de 94%, como se muestra en la **gráfica 28**.

Ilustración 11. SVM Lineal - Valores de validación (Accuracy y Kappa) obtenidos en cada partición y repetición.

```
Support Vector Machines with Linear Kernel

911 samples
  8 predictor
  2 classes: '0', '1'

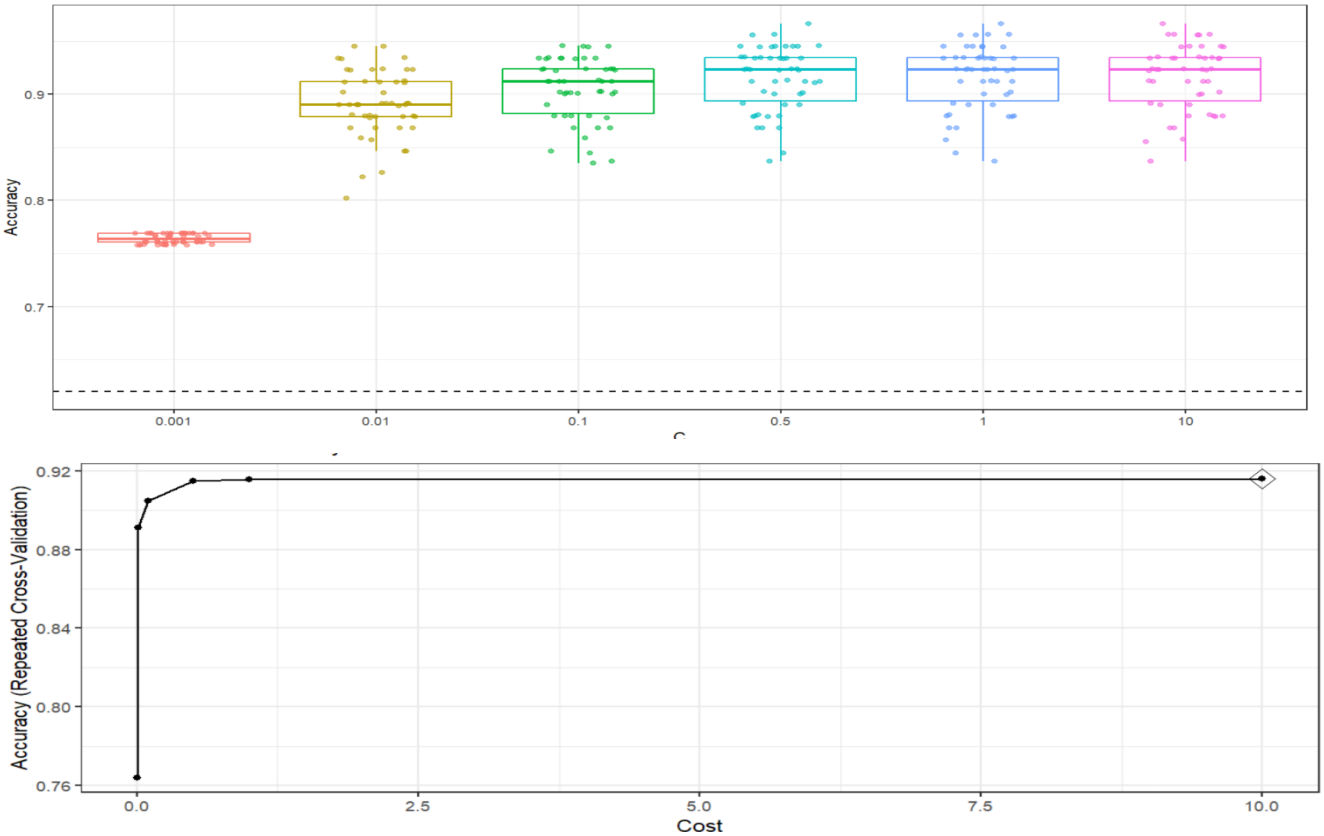
No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 819, 821, 820, 820, 820, 820, ...
Resampling results across tuning parameters:

  C      Accuracy   Kappa
  1e-03  0.7640092  0.0000000
  1e-02  0.8911188  0.6453307
  1e-01  0.9049367  0.6965385
  5e-01  0.9148225  0.7323472
  1e+00  0.9157017  0.7352744
  1e+01  0.9161437  0.7368998

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was C = 10.
```

Fuente: elaboración propia

Gráfica 28. SVMLineal-Evolución de los modelos según el valor de los hiperparámetros



Fuente: elaboración propia

Finalmente, se selecciona el modelo final con el parámetro de C=10 como se muestra en la ***Ilustración 12*** y se observan los valores promedio de accuracy, kappa y función de costo C, luego de la validación cruzada.

Ilustración 12 Resumen del modelo con la validación cruzada SVMLineal

```
```{r message=TRUE, warning=FALSE}
summary(modelo_svmlineal$resample)
```
```

```
    Accuracy          Kappa              C              Resample
 Min.   :0.7582   Min.   :0.0000   Min.   : 0.001   Length:300
 1st Qu.:0.8571   1st Qu.:0.5122   1st Qu.: 0.010   Class :character
 Median :0.9111   Median :0.7139   Median : 0.300   Mode  :character
 Mean   :0.8837   Mean   :0.5877   Mean   : 1.935
 3rd Qu.:0.9239   3rd Qu.:0.7800   3rd Qu.: 1.000
 Max.   :0.9565   Max.   :0.8726   Max.   :10.000
```

Fuente: elaboración propia

La validación del modelo se acompaña con la generación de la matriz de confusión y la estimación del error de clasificación de 9,39%, lo que lo constituye como un buen modelo de clasificación de la deserción universitaria como se observa en la **Ilustración 13**

Ilustración 13. SVM Lineal - Matriz de confusión y error de clasificación

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 364  43
         1   2  70

               Accuracy : 0.9061
                 95% CI : (0.8763, 0.9306)
    No Information Rate : 0.7641
    P-Value [Acc > NIR] : 6.978e-16

                  Kappa : 0.702

 Mcnemar's Test P-Value : 2.479e-09

            Sensitivity : 0.6195
            Specificity : 0.9945
         Pos Pred Value : 0.9722
         Neg Pred Value : 0.8943
             Prevalence : 0.2359
         Detection Rate : 0.1461
   Detection Prevalence : 0.1503
      Balanced Accuracy : 0.8070

       'Positive' Class : 1
```

```
### Error de clasificación

```{r message=FALSE, warning=FALSE}
error_test = mean(predicciones_raw != datos_test_prep$DESERTOR)
paste("El error de test del modelo:", round(error_test*100, 2), "%")
```

 [1] "El error de test del modelo: 9.39 %"
```

También se realiza una máquina de soporte vectorial radial, en la cual se usa los mismos valores para la selección de hiperparámetros del modelo lineal para la función de costo c y el modelo arroja tras la validación cruzada donde también se conservan los valores de 10 particiones y 5 repeticiones dando como resultado valores muy similares de accuracy y kappa del modelo SVMLineal como se muestra en **Ilustración *14***.

Ilustración 14 Resultado Máquinas de vectores soporte con núcleo de función de base radial.

```
Support Vector Machines with Radial Basis Function Kernel

910 samples
  8 predictor
  2 classes: 'NO', 'SI'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 818, 820, 819, 819, 819, 819, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.9171003  0.7385013
  0.50  0.9252374  0.7667834
  1.00  0.9333792  0.7955872

Tuning parameter 'sigma' was held constant at a value of 0.1446552
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.1446552 and C = 1.
```

7.3.5 Modelo de bosques aleatorios (Random Forest)

Para la selección del mejor modelo se usa la técnica de resampling a través de validación cruzada y el tuneo de los hiperparámetros donde se establece el número de árboles en el bosque de 500, la profundidad máxima de los árboles, el atributo el mtry toma valores de (3, 4, 5, 7) el cual establece el número de variables muestreadas aleatoriamente como candidatas en cada división y el número de nodos (2, 3, 4, 5, 10, 15, 20, 30) que establece el número mínimo de muestras requeridas para dividir un nodo interno y el parámetro de división con el método Gini.

El cual nos arroja que el mejor modelo es el que se determina mtry = 3, min.node.size = 30 y splitrule = "gini", como se muestra en **la Ilustración 15 e *Ilustración 16*** donde se observan los valores obtenidos por cada partición y repetición, de igual forma en la se puede observar la evolución de los valores de accuracy de los hiperparámetros representan mejores valores de exactitud y disminución del error, finalmente se tomó el mejor modelo con métricas de accuracy medio 94%, como se muestra en la **Ilustración *17***.

Ilustración 15. Random Forest - Valores de validación (Accuracy y Kappa) obtenidos en cada partición y repetición.

```
Random Forest

911 samples
  8 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 819, 821, 820, 820, 820, 820, ...
Resampling results across tuning parameters:

  mtry  min.node.size  Accuracy   Kappa
  3      2            0.9394172  0.8207064
  3      3            0.9405113  0.8241126
  3      4            0.9402914  0.8230089
  3      5            0.9400717  0.8219173
  3     10            0.9394148  0.8197904
  3     15            0.9389728  0.8180549
  3     20            0.9398495  0.8205237
  3     30            0.9409630  0.8233041
  4      2            0.9389776  0.8208935
  4      3            0.9391974  0.8212448
  4      4            0.9396369  0.8222692
  4      5            0.9389800  0.8209794
  4     10            0.9392023  0.8208676
  4     15            0.9405113  0.8240086
  4     20            0.9391902  0.8199672
  4     30            0.9409533  0.8248497
  5      2            0.9381033  0.8193531
  5      3            0.9398543  0.8238921
  5      4            0.9385429  0.8202247
  5      5            0.9385380  0.8199506
  5     10            0.9383159  0.8192511
  5     15            0.9380936  0.8186064
  5     20            0.9387578  0.8201015
  5     30            0.9398447  0.8221328
  7      2            0.9376589  0.8189104
  7      3            0.9374415  0.8182114
  7      4            0.9374391  0.8181238
  7      5            0.9378834  0.8197257
  7     10            0.9374463  0.8179182
  7     15            0.9374391  0.8172449
  7     20            0.9381009  0.8194641
  7     30            0.9394100  0.8226539

Tuning parameter 'splitrule' was held constant at a value of gini
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were mtry = 3, splitrule = gini and min.node.size = 30.
```

Fuente: elaboración propia

Ilustración 16 Random Forest - Resumen del mejor modelo

```
```{r message=FALSE, warning=FALSE}
modelo_rf$finalModel
```
```

```
Ranger result

Call:
 ranger::ranger(dependent.variable.name = ".outcome", data = x,      mtry = min(param$mtry,
ncol(x)), min.node.size = param$min.node.size,      splitrule = as.character(param$splitrule),
write.forest = TRUE,      probability = classProbs, ...)

Type:                             Classification
Number of trees:                  500
Sample size:                      910
Number of independent variables:  8
Mtry:                             3
Target node size:                 30
Variable importance mode:         none
Splitrule:                        gini
OOB prediction error:             6.04 %
```

Fuente: elaboración propia

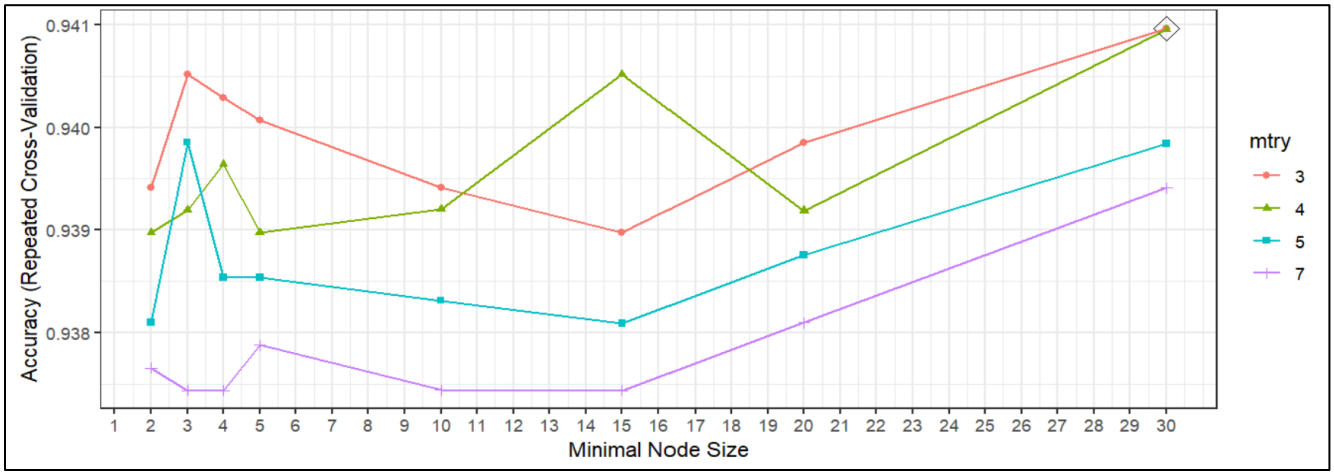
Ilustración 17 Resumen del modelo con la validación cruzada Bosques aleatorios

```
```{r message=TRUE, warning=FALSE}
summary(modelo_rf$resample)
```
```

```
    Accuracy          Kappa          Resample
 Min.   :0.8889   Min.   :0.6278   Length:50
 1st Qu.:0.9341   1st Qu.:0.8012   Class :character
 Median :0.9444   Median :0.8307   Mode  :character
 Mean   :0.9402   Mean   :0.8211
 3rd Qu.:0.9565   3rd Qu.:0.8726
 Max.   :0.9780   Max.   :0.9381
```

Fuente: elaboración propia

Gráfica 29. Random Forest-Evolución de los modelos según el valor de los hiperparámetros



Fuente: elaboración propia

De igual forma que el modelo SVM anterior, la validación del modelo se acompaña con la generación de la matriz de confusión y la estimación del error de clasificación de 6,05%, lo que lo constituye como un buen modelo de clasificación de la deserción universitaria y mejora significativamente el error de clasificación con respecto al modelo SVM como se observa en la *Ilustración 18*.

Ilustración 18. Random Forest - Matriz de confusión y error de clasificación

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 362  25
         1   4  88

               Accuracy : 0.9395
                 95% CI : (0.9142, 0.9591)
    No Information Rate : 0.7641
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8205

 Mcnemar's Test P-Value : 0.0002041

            Sensitivity : 0.7788
            Specificity : 0.9891
         Pos Pred Value : 0.9565
         Neg Pred Value : 0.9354
             Prevalence : 0.2359
         Detection Rate : 0.1837
   Detection Prevalence : 0.1921
      Balanced Accuracy : 0.8839

       'Positive' Class : 1
```

```
### Error de clasificación RF

```{r message=FALSE, warning=FALSE}
error_test = mean(predicciones_raw != datos_test_prep$DESERTOR)
paste("El error de test del modelo:", round(error_test*100, 2), "%")
```

[1] "El error de test del modelo: 6.05 %"
```

Fuente: elaboración propia

7.3.6 Modelo de redes neuronales simple (NNET)

Al igual que el entrenamiento de los modelos de SVM y Random Forest se unas la técnica de resampling a través de validación cruzada y el tuneo de los hiperparámetros como este entrenamiento se realiza con la librería Caret, solo es necesario establecer dos hiperparámetros donde el tamaño de la red comprende entre (1:5) el cual se establece como el número de capas ocultas y el decay (0, 0.1, 0.01, 0.001,0.5) que es el parámetro de regularización para evitar el sobre ajuste.

El cual nos arroja que el mejor modelo tiene un size=3 y un decay=0.1 , como se muestra en la Ilustración 19, donde se observan los valores obtenidos por cada partición y repetición, de igual forma en la se puede observar la evolución de los valores de accuracy de los hiperparámetros representan mejores valores de exactitud y disminución del error, finalmente se tomó el mejor modelo con métricas de accuracy de 94%, como se muestra en la Gráfica 30.

Ilustración 19. NNET - Valores de validación (Accuracy y Kappa) obtenidos en cada partición y repetición.

```
Neural Network

911 samples
  8 predictor
  2 classes: '0', '1'

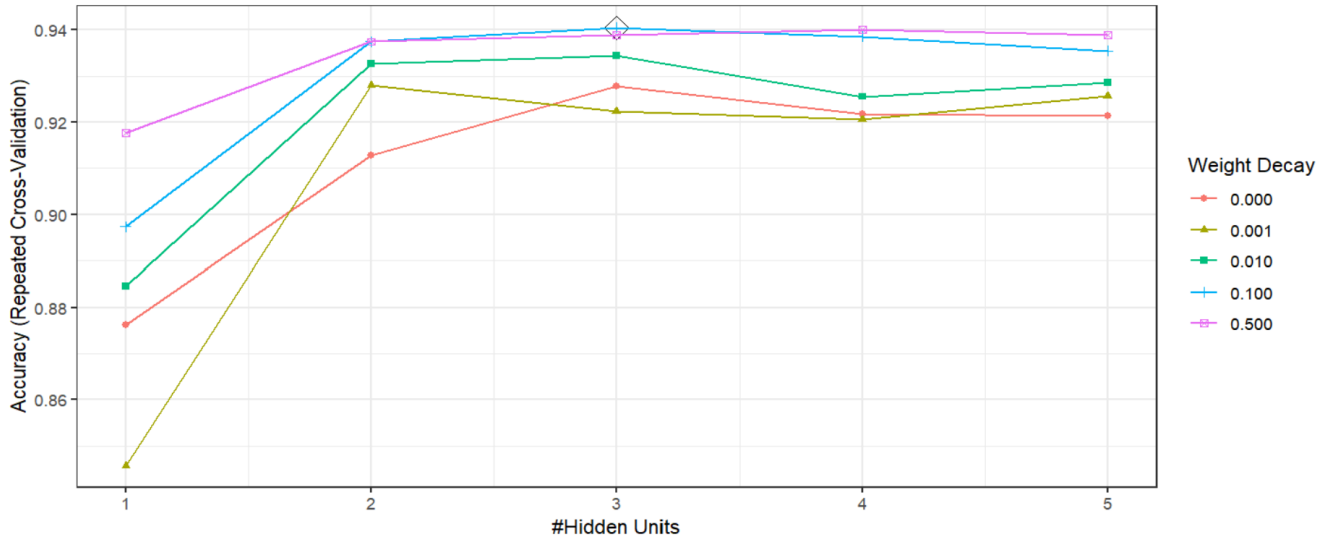
No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 819, 819, 821, 821, 820, 820, ...
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
  1     0.000  0.8762021  0.6599420
  1     0.001  0.8456646  0.6033636
  1     0.010  0.8845553  0.6832091
  1     0.100  0.8974143  0.7075639
  1     0.500  0.9176884  0.7499978
  2     0.000  0.9128073  0.7504087
  2     0.001  0.9280019  0.7865848
  2     0.010  0.9326152  0.8002889
  2     0.100  0.9374482  0.8147829
  2     0.500  0.9374507  0.8129411
  3     0.000  0.9277895  0.7891712
  3     0.001  0.9222873  0.7757649
  3     0.010  0.9343733  0.8107059
  3     0.100  0.9405108  0.8254097
  3     0.500  0.9389748  0.8164473
  4     0.000  0.9218355  0.7736531
  4     0.001  0.9205169  0.7709666
  4     0.010  0.9255675  0.7843012
  4     0.100  0.9385328  0.8209959
  4     0.500  0.9400761  0.8201223
  5     0.000  0.9214250  0.7735045
  5     0.001  0.9255987  0.7842261
  5     0.010  0.9286397  0.7943417
  5     0.100  0.9354509  0.8112183
  5     0.500  0.9389748  0.8170536

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were size = 3 and decay = 0.1.
```

Fuente: elaboración propia

Gráfica 30. NNET-Evolución de los modelos según el valor de los hiperparámetros



Fuente: elaboración propia

Finalmente, se realiza la validación del modelo con la generación de la matriz de confusión y la estimación del error de clasificación de 6,47%, lo que lo constituye como un buen modelo de clasificación de la deserción universitaria con un porcentaje de error similar al modelo Random Forest, como se observa en la **Ilustración 20**

Ilustración 20. NNET- Matriz de confusión y error de clasificación

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 361  26
         1   5  87

               Accuracy : 0.9353
                 95% CI : (0.9094, 0.9556)
    No Information Rate : 0.7641
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8082

 Mcnemar's Test P-Value : 0.000328

            Sensitivity : 0.7699
            Specificity : 0.9863
         Pos Pred Value : 0.9457
         Neg Pred Value : 0.9328
             Prevalence : 0.2359
         Detection Rate : 0.1816
   Detection Prevalence : 0.1921
      Balanced Accuracy : 0.8781

       'Positive' Class : 1
```

```
### Error de clasificación NNET

```{r message=FALSE, warning=FALSE}
error_test = mean(predicciones_raw != datos_test_prep$DESERTOR)
paste("El error de test del modelo:", round(error_test*100, 2), "%")
```

 [1] "El error de test del modelo: 6.47 %"
```

Fuente: elaboración propia

7.3 Comparación y validación de los modelos

7.3.1 Comparación de métricas de los modelos

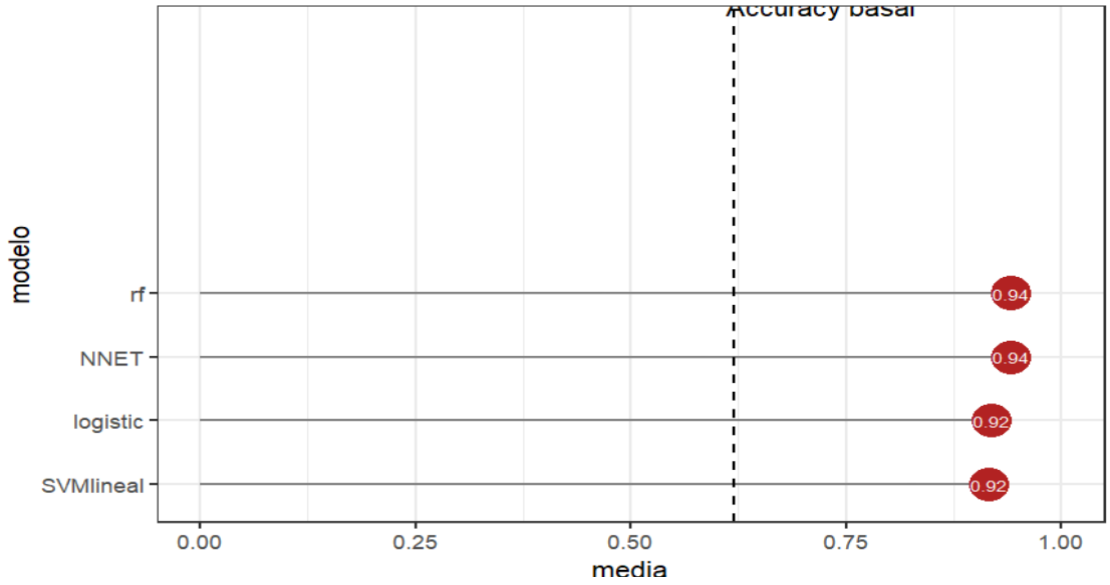
Los modelos para su comparación se realizan a través de las métricas accuracy y kappa promedio como se ve en la Tabla 13, se observa que el modelo que obtiene mayor accuracy promedio es el random forest, seguido de la red neuronal simple, en el caso de la regresión logística y la máquina de soporte vectorial lineal, los valores son muy similares y se evidencia que todos los modelos poseen métricas de exactitud por encima de la línea base de 0,62 como se observa en la **Gráfica 31.**

Tabla 13. Valor promedio de accuracy y kappa por modelo

modelo <chr>	Accuracy <dbl>	Kappa <dbl>
rf	0.9409630	0.8233041
NNET	0.9405108	0.8254097
logistic	0.9178947	0.7482142
SVMlineal	0.9161437	0.7368998

Fuente: elaboración propia

Gráfica 31. Validación: Accuracy medio repeated-CV modelos ordenados por media



Fuente: elaboración propia

Ahora bien, para comprobar si las diferencias entre las métricas de los modelos son significativas, se recurre al uso de dos pruebas estadísticas el Friedman y el de suma de rangos de Wilcoxon, en el primer test con un nivel de significancia de ($\alpha = 0.05$), no se encuentra evidencia estadística para determinar que existen diferencias entre la precisión de los modelos como se observa en la *Ilustración* **21**, para esto se recurre a la segunda prueba de contrastes post HOC (Wilcoxon) por pares de modelos, donde se evidencia que no existe suficiente evidencia estadística para considerar que existe diferencia en la capacidad predictiva de los cuatro modelos, como se observa en la **Ilustración 22.**

Ilustración 21. Resultado de la prueba de Friedman

```
        Friedman rank sum test

data:  matriz_metricas
Friedman chi-squared = 58.602, df = 3, p-value = 1.169e-12
```

Fuente: elaboración propia

Ilustración 22 Resultado de la prueba de Wilcoxon

modeloA <chr>	modeloB <chr>	p_value <dbl>
NNET	logistic	1.420737e-03
rf	logistic	5.616163e-08
rf	NNET	9.168052e-01
SVMlineal	logistic	7.763447e-01
SVMlineal	NNET	1.420737e-03
SVMlineal	rf	4.574392e-08

Fuente: elaboración propia

Finalmente, se realiza una prueba de estimaciones de error de los modelos a partir de los datos de prueba para aseguras que, durante la optimización y entrenamiento, no se haya generado sobre ajustes en los modelos, de la prueba se concluye que todos los modelos tienen una exactitud sobresaliente para los datos de entrenamiento y pruebas como se puede evidenciar en la Tabla 13 y la Gráfica *32*, lo cual reafirma que los modelos Random forest y red neuronal tienen

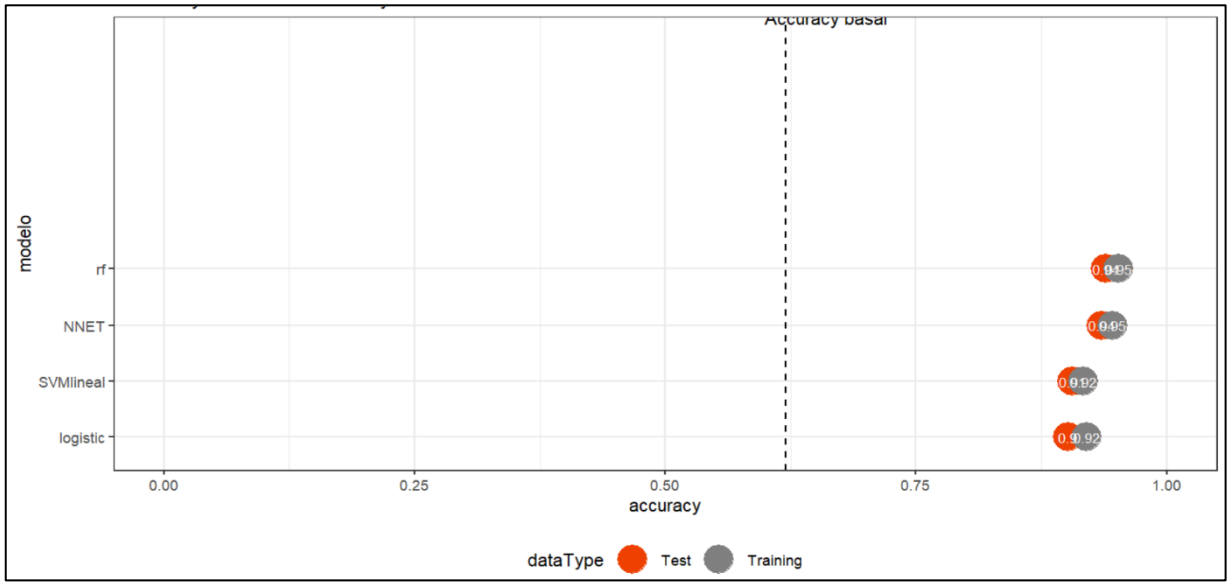
las mejores métricas de acierto de predicción de la deserción universitaria.

Tabla 14. Métrica accuracy de los modelos para los datos de entrenamiento y prueba

object <chr>	Test <dbl>	Training <dbl>
rf	0.9394572	0.9517014
NNET	0.9352818	0.9462130
SVMlineal	0.9060543	0.9165752
logistic	0.9018789	0.9198683

Fuente: elaboración propia

Gráfica 32. Métrica accuracy de entrenamiento y prueba por modelo



Fuente: elaboración propia

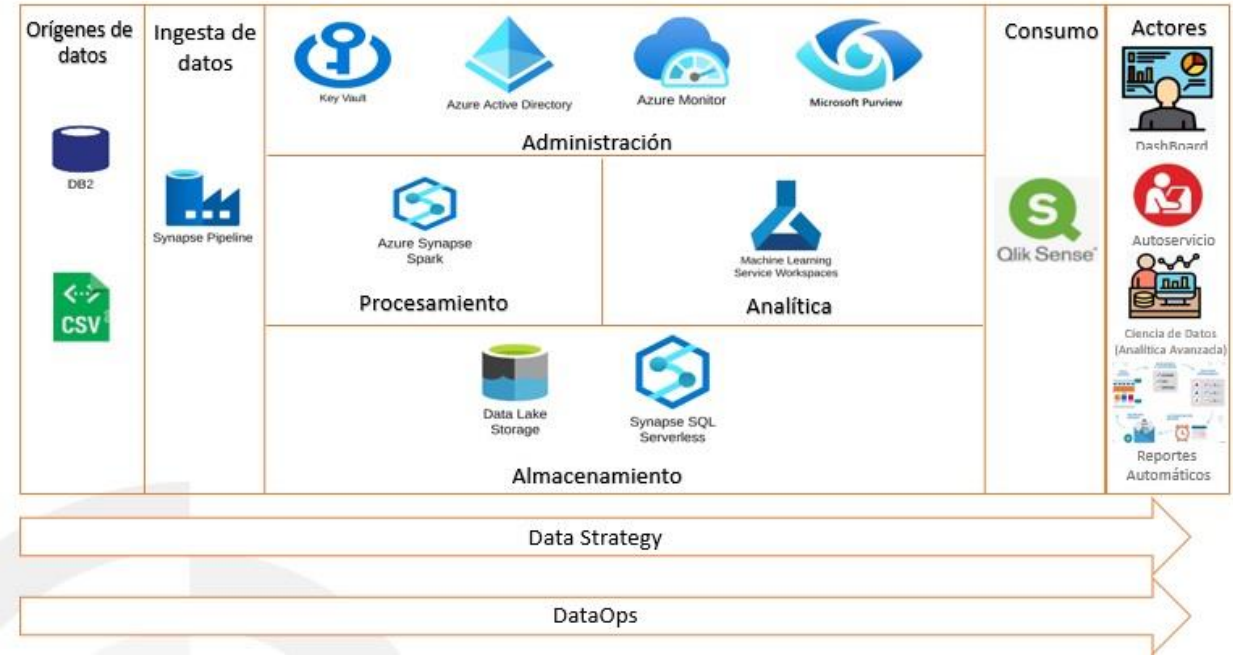
7.4 Arquitectura del sistema de alertas tempranas

Para la implementación del sistema se plantean el escenario arquitectura donde se definen las tecnologías, el flujo de los datos y la aplicación visual con los resultados del modelo y el sistema de reportaría automatizado, las tecnologías descritas a continuación se basan en el actual ecosistema de analítica de la Universidad de Cecar y las herramientas licenciadas en la misma,

como se muestra en la

Ilustración *23*, la cual se detalla a continuación.

Ilustración 23 Arquitectura candidata del sistema de alertas tempranas Azure con Synapses



Fuente: elaboración propia

Orígenes de datos: la información se tomará de la base de datos transaccional del sistema de gestión académico y financiero implementado el esquema de bodega propuesto en la **Ilustración 2**, a través de una conexión a Sql Server desde el servidor en tierra de la universidad, más el archivo de resumen y parametrización de la ingesta de datos en formato csv.

Almacenamiento de datos: se sugiere hacer uso de: Data Lake Storage Gen2 donde se establezcan tres capas de datos, Bronce: contiene datos sin procesar, Plata: contiene datos limpios y filtrados y Oro: almacena datos agregados que son útiles para el análisis empresarial, con el uso del formato Delta Lake el cual constituye la capa mantenida(oro) del lago de datos en formato parquet.

Procesamiento de datos: Se sugiere usar Azure Synapse el cual proporciona una solución de análisis integral al combinar **Big Data Analytics, Data Lake, Data Warehousing y Data Integration** en una sola plataforma unificada. Tiene la capacidad de consultar datos relacionales y no relacionales a escala de petabytes mediante la ejecución de consultas distribuidas inteligentes entre nodos en el backend de manera tolerante a fallas.

La arquitectura de Synapse consta de cuatro componentes: Synapse SQL, Spark, Synapse Pipeline y Studio.

1. **Synapse SQL** es un sistema de consultas distribuidas para T-SQL que permite escenarios de virtualización de datos y almacenamiento de datos y extiende T-SQL para abordar escenarios de transmisión y aprendizaje automático.

2. **Synapse SQL** ofrece modelos de recursos **dedicados** y **sin servidor (serverless)**. Para obtener un rendimiento y un costo predecibles, cree grupos de SQL dedicados para reservar la capacidad de procesamiento de los datos almacenados en las tablas de SQL. Para cargas de trabajo imprevistas o en ráfagas, use el punto final de SQL sin servidor y siempre disponible.

3. **Apache Spark** para **Azure Synapse** integra a la perfección **Apache Spark**, el motor de big data de código abierto más popular que se utiliza para la preparación de datos, la ingeniería de datos, ETL y el aprendizaje automático.

4. **Synapse Pipeline** contiene el mismo motor de integración de datos y las mismas experiencias que Azure Data Factory, lo que le permite crear **Pipelines ETL** enriquecidas a escala sin salir de **Azure Synapse Analytics**.

5. **Synapse Studio** elimina las barreras tecnológicas tradicionales entre el uso conjunto de SQL y Spark, el cual puede mezclar y combinar sin problemas según sus necesidades y experiencia.

Analítica: El modelo se implementa en Azure ML a través de un cuaderno denominado plumber.R, el cual define el API para el uso del modelo y tendrá como parámetros los siguientes

@param MAXIMO.PERIODO.CURSADO

@param ULTIMO.SEMESTRE.CURSADO

@param TOTAL.REPROBADAS

@param PROMEDIO.GENERAL

@param TOTAL.APROBADAS

@param TOTAL.INCENTIVOS @param EDAD

@param NUM.HERMANOS

@post /clasificador

Consumo: Se propone el uso de Qlik Sense donde se podrá visualizar el análisis exploratorio de datos donde se definen las métricas y los filtros asociados al análisis y el resultado del modelo estableciendo visualmente semaforización por colores del riesgo de deserción según los rangos de probabilidad de desertar por cada estudiante donde se muestran alertas en color verde = [0:0,3], naranja = (0,3:0,50] y rojo =(0,50:1], finalmente para la implementación de los reportes automáticos se sugiere el uso de Qlik NPrinting el cual es la herramienta de reportaría de la herramienta BI Qlik.

CONCLUSIONES

- Los modelos predictivos de aprendizaje estadístico y máquina permiten predecir la deserción escolar a partir de factores sociodemográficos, económicos y académicos, la aplicación de estos para la deserción universitaria ofrece beneficios significativos, tanto para las instituciones educativas como para los propios estudiantes. Las instituciones pueden utilizar estos modelos para identificar a los estudiantes en riesgo y diseñar intervenciones personalizadas, como tutorías académicas, asesoramiento emocional o programas de apoyo financiero. Al intervenir de manera proactiva, las instituciones pueden aumentar las tasas de retención y mejorar la experiencia estudiantil en general.

- La selección adecuada de características desempeña un papel fundamental en el desarrollo de modelos aprendizaje automático de predicción de la deserción escolar. Durante nuestra investigación, hemos observado que al identificar las variables más relevantes y descartar aquellas que carecen de impacto significativo, se logra mejorar tanto la precisión del modelo como su interpretación. Al centrarse en las características más influyentes, el modelo adquiere una mayor capacidad para identificar patrones y señales tempranas de deserción, lo que brinda a las instituciones educativas una base sólida para implementar intervenciones preventivas y personalizadas. Esta selección cuidadosa de características optimiza la utilidad práctica y la eficacia del modelo de predicción de deserción escolar en beneficio de los estudiantes y las comunidades educativas en general.

- Para la creación del modelo predictivo se tuvieron en cuenta variables 7 variables que explican mejor la deserción escolar para el caso de estudio, las dos primeras están asociadas a la temporalidad del suceso donde a mayor semestre y periodo cursado disminuye la probabilidad de desertar, las dos siguientes pertenecen a factores

académicos total de asignaturas aprobadas y reprobadas, la quinta variable determinante es la cantidad de incentivos monetarios y finalmente las dos últimas pertenecen a factores sociales como lo son el estrato y el número de hermanos, estas se seleccionaron a través de selección de importancia aplicando un modelo de regresión logística y el uso de las métricas reducción de pureza GINI y accuracy del modelo bosques aleatorios.

- Los modelos de regresión logística, máquina de soporte vectorial lineal y polinómica, bosques aleatorios y redes neuronales simples resultan ser buenos clasificadores para predecir la deserción escolar con una exactitud superior al 90%. Al comparar el rendimiento de estos modelos, se ha observado que todos demostraron ser válidos para obtener resultados confiables y ser eficaces en la predicción de la deserción. Esta amplia exploración de modelos ha permitido tener una visión más completa y respaldada por evidencia para seleccionar el enfoque más adecuado en función de los objetivos y las características de los datos de cada proyecto de predicción de deserción escolar.

- El uso de técnicas de estimación de hiperparámetros y la validación cruzada es fundamental para evaluar la capacidad de generalización de los modelos de predicción de la deserción escolar. Como resultado, se obtienen estimaciones más precisas del rendimiento del modelo y se disminuye el riesgo de sobreajuste.

- Se sugiere para la implementación del sistema de alertas tempranas bajo el ecosistema de Azure el cual cuenta con características de facilidad de uso y productividad, buenos niveles de procesamiento y escalabilidad, integración de lenguajes de programación para la ciencia de datos como R y Python para el despliegue y gestión de los modelos de machine learning en producción y finalmente la integración con plataformas de análisis y visualización de datos como PowerBI y Qlik Sense.

- Como futuros trabajos académicos sobre el estudio y la predicción de la deserción universitaria además de analizar factores como variables demográficas, socioeconómicas,

académicas relevantes y psicológicas, se deben ampliar el análisis a rasgos de personalidad y comportamental para poder determinar patrones de incidencia de la deserción con la personalidad y el compartimiento de los estudiantes. Sin embargo, es importante tener en cuenta que los modelos predictivos no son una solución completa por sí mismos, estos deben utilizarse en combinación con otros enfoques y considerarse como una herramienta adicional en un enfoque integral para abordar la deserción universitaria.

Referencias

[1] F. Barrero Rivera, «Investigación en deserción estudiantil universitaria: educación cultura y significados.,» *Revista de Educación y Desarrollo Social,* vol. 9(2), nº 8, pp. 6 -101, 2015.

[2] Creative Associates International, Inc., «PROGRAMA PILOTO DE PREVENCIÓN DE LA DESERCIÓN ESCOLAR,» Washington, DC, 2015.

[3] Ministerio de educación nacional, «SPADIES sistema para la prevención de la deserción de la educación superior,» 2020. [En línea]. Available: https://www.mineducacion.gov.co/sistemasinfo/spadies/. [Último acceso: 04 07 2022].

[4] V. Tinto, «Deserción universitaria en Iberoamérica alcanza,» El espectador, [En línea]. Available: https://www.elespectador.com/educacion/desercion-universitaria-en-iberoamerica-alcanza-el-33/. [Último acceso: 04 07 2022].

[5] Corporación Universitaria del Caribe, CECAR, «Estudio sobre tasa y tiempo promedio de la graduación de estudiantes del programa Ingeniería de Sistemas,» Sincelejo - Sucre, 2018.

[6] Universidad de los Andes, «Informe Determinantes de la deserción,» Centro de Estudios sobre Desarrollo Económico CEDE, Bogota, 2014.

[7] Ministerio de Eucación Nacional, Deserción estudiantil en la educación superior colombiana, Bogotá – Colombia: Colombia aprende, 2009.

[8] M. J. Pachay López y M. Rodríguez Gámez, «La deserción escolar: Una perspectiva compleja en tiempos de pandemia,» *Polo del conocimiento,* vol. vol. 6, pp. 130-155, 2021.

[9] M. A. Vélez y M. Rodríguez Gamez, «Escala de participación de los padres y su relación en la escolarización de los alumnos,» Tesis de maestria, España, 2020.

[10] L. Calderón Ruiz, «Factores y riesgo de deserción escolar durante la pandemia (Covid-19),» Tesis, Universidad Técnica de Ambato, Ambato-Ecuador, 2021.

[11] Observatorio de Educación Superior de Medellín, «Deserción en la Educación Superior,» Boletín N° 5, Medellín, Colombia, 2017.

[12] IBM, «learn/machine-learning,» IBM, [En línea]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Último acceso: 03 07 2022].

[13] K. P. Murphy, Machine Learning: A Probabilistic Perspective, Cambridge, Massachusetts: Massachusetts Institute of Technology - MIT, 2012.

[14] IBM, «SPSS Modeler,» 2021. [En línea]. Available: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model.. [Último acceso: 18 03 2023].

[15] IBM, «IBM Cloud Learn Hub,» 2021. [En línea]. Available: https://www.ibm.com/cloud/learn/boosting.. [Último acceso: 18 03 2023].

[16] D. HOSMER y S. LEMESHOW, Applied Logistic Regression, New York: Wiley, 2013.

[17] D. W. T. H. a. R. T. G. James, «An Introduction to Statistical Learning,» *Springer,* vol. 112, 2013.

[18] J. Berkson, «Journal of the American Statistical Association,» Journal of the American Statistical Association, 1944. [En línea]. Available: http://www.jstor.org/stable/2280041?origin=crossref.

[19] J. C. M. S. a. Q. L. H. Zhang, «A neural network model for short-term wind speed prediction using wavelet decomposition,» pp. 316-321, 2020.

[20] S. Haykin, Neural networks: A comprehensive foundation, Prentice Hall, 1994.

[21] G. E. H. a. R. J. W. D. E. Rumelhart, «Learning representations by back-propagating errors,» *Nature,* vol. 323, nº 6088, pp. 533-536, 1986.

[22] J. H. F. R. A. O. a. C. J. S. L. Breiman, «Classification and Regression Trees,» CRC PRESS, 1984.

[23] W. Y. Loh, «Classification and regression trees,» *Data Mining and Knowledge Discovery,* vol. 1, nº 1, pp. 14-23, 2011.

[24] E. Galindo, J. Perdomo y J. Figueroa García, «Estudio comparativo entre máquinas de soporte vectorial,» *Scielo,* vol. 31, nº 1, pp. 12-19, 2020.

[25] IBM, «learn/data-mining,» IBM, [En línea]. Available: https://www.ibm.com/cloud/learn/data-mining. [Último acceso: 03 07 2022].

[26] J. A. Barroso Salgado, «Modelo predictivo basado en machine learning de ordenes de trabajo riesgosas para mantenimiento de equipos mineros,» Universidad de Chile Facultad de Ciencias Físicas y Matemáticas , Santiago De Chile, 2018.

[27] IBM, «learn data-warehouse,» IBM, [En línea]. Available: https://www.ibm.com/cloud/learn/data-warehouse. [Último acceso: 02 07 2022].

[28] R. Kimball y M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition, Indianapolis, Indiana: John Wiley & Sons, Inc, 2013.

[29] D. A. Calle Sanchez, «Manual para el diseño e implementación de bases de datos OLAP y su aplicación en inteligencia de negocios,» Trabajo de grado Universidad EAFIT, Medellín, Colombia, 2009.

[30] SNGULAR, «SNGULAR,» 2021. [En línea]. Available: https://www.sngular.com/es/data-science-crisp-dm-metodologia/.. [Último acceso: 18 03 2023].

[31] P. Haya, «Instituto de Ingeniería del conocimiento,» 2021. [En línea]. Available: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/#_ftn3.. [Último acceso: 18 03 2023].

[32] R. Velasco, 02 2021. [En línea]. Available: https://www.linkedin.com/pulse/crisp-dm-metodolog%C3%ADa-para-proyectos-de- ciencia-datos-y-Velasco/?originalSubdomain=es.. [Último acceso: 12 02 2022].

[33] O. Rodríguez y R. Oldemar, «Oldemarrrodriguez,» [En línea]. Available: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037..

[34] H. F. Vargas Losada, M. F. Tovar Rubiano y J. C. Villanueva Muñoz, «Los SAT (Sistemas De Alertas Tempranas),» *Revista Científica CIDC,* vol. 26, pp. 21-28, 2016.

[35] J. E. Sotomonte Castro, C. C. Rodríguez Rodríguez, C. E. Montenegro Marín, P. A. Gaona García y J. Gabriel Castellanos, «Hacia la construcción de un modelo predictivo de deserción académica

basado en técnicas de minería de datos,» *Revista científica,* pp. 37-52, 2016.

[36] A. J. Camargo García, «Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos,» Corporación Universidad de la Costa, Barranquilla, 2020.

[37] «Modelo predictivo para estimar la deserción de estudiantes en una institución de educación superior,» Universidad de Chile, Santiago de Chile, 2016.

[38] B. Cuji , W. Gavilanes y R. Sanchez, «Modelo predictivo de deserción estudiantil basado en arboles de decisión,» *Espacios,* vol. 38, nº 55, p. 17, 2017.

[39] T. Achilie Valencia, «Factores que motivan el abandono estudiantil en la Universidad: Un estudio de caso,» *BOOK,* pp. 1-54, 2018.

[40] M. Rodríguez Urrego, «La investigación sobre deserción universitaria en Colombia 2006-2016. Tendencias y resultados,» *Pedagogía y Saberes,* nº 51, 2019.

[41] J. Zarate y S. Mateus, «Propuesta de un Modelo Predictivo utilizando Aprendizaje Profundo para el análisis de deserción estudiantil en Universidades Colombianas Virtuales,» *Innovación y Desarrollo Sostenible,* vol. 1, pp. 51-57, 2020.

[42] F. Barrero Rivera, «Investigación en deserción estudiantil universitaria: educación cultura y significados,» *Revista de Educación y Desarrollo Social,* vol. 9, nº 2, pp. 86-101, 2015.

[43] J. Niyogisubizo, E. Nziyumva, M. Evariste y P. C. Nshimyumukiza, «Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization,» *Computers and Education: Artificial Intelligence,* vol. 3, 2022.

[44] B. Pérez, C. Castellanos y D. Correal , «Applications of Computational Intelligence,» de *Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study*, IEEE Colombian Conference on Applications in Computational Intelligence, Communications in Computer and Information Science , 2018, p. 111–125.

[45] D. Opazo, S. Moreno, E. Álvarez Miranda y J. Pereira, «Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities,» *Mathematics,* vol. 9, nº 20, 2021.

[46] J. Ramón Martínez y Y. Ferrás Fernández, «Regresión logística y predicción del bajo rendimiento académico de estudiantes,» *Revista Electrónica Dr. Zoilo E. Marinello Vidaurreta,* vol. 45, nº 4, pp. 100-106, 2020.

[47] J. L. Sarmiento Ramos, «Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica,» *Revista UIS Ingenierías,* vol. 19, nº 4, pp. 1-18, 2020.

[48] H. Vite Cevallos, H. Carvajal Romero y S. Barrezueta Unda, «Aplicación de algoritmos de aprendizaje automático para clasificar la fertilidad de un suelo bananero,» *Scielo,* vol. 16, nº 72, pp. 15-19, 2020.