

**UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA**

FACULTAD DE ECONOMIA Y PLANIFICACIÓN



**"MEJORA DEL INDICADOR DE RETENCIÓN EN UNA
UNIVERSIDAD PRIVADA A PARTIR DE LA CLASIFICACIÓN DE
ALUMNOS UTILIZANDO UN MODELO PREDICTIVO "**

**TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO
DE INGENIERO EN ESTADÍSTICO E INFORMÁTICO**

JEAN PIERRE URIBE MOSTACERO

LIMA - PERÚ

2022

Document Information

Analyzed document	Trabajo TSP_VF_09092022_Jean Pierre_Uribe_Original.pdf (D143964938)
Submitted	2022-09-12 21:41:00
Submitted by	ivan soto rodriguez
Submitter email	ivans@lamolina.edu.pe
Similarity	0%
Analysis address	ivans.unalm@analysis.orkund.com

Sources included in the report

	URL: https://dokumen.pub/probability-statistics-and-random-processes-with-queueing-theory-and-queue... Fetched: 2021-10-31 02:15:12	 1
---	--	---

Entire Document

1 RESUMEN La presente investigación tiene propósito predecir la deserción estudiantil de una universidad privada, aplicando dos técnicas de la minería de datos la regresión logística binaria y árbol de clasificación CART. Para el estudio se utilizó datos la base de datos de alumnos en los periodos 2019-2 con 32176 registros con datos relacionados a factores socio-demográfica, académicos y económicos. Se aplicó el balanceo de datos con la técnica de submuestreo a fin de mejorar la capacidad predictiva. El árbol de clasificación CART resultó con mayores valores para la exactitud, sensibilidad, especificidad y AUC de 73,8%, 97,3%, 50,3% y 73,8% respectivamente para predecir la deserción universitaria en comparación de la regresión logística binaria cuyos valores fueron 66,4%, 71,2%, 65,8% y 72,4% respectivamente. El árbol resultó identificó las variables más importantes: TAS_NOM_A, TAS_NOM_P, SEDE, TAS_ASI_A; con un tamaño de 13 nodos, con siete nodos terminales, de los cuales tres para predecir la clase SI y cuatro para la clase NO; así mismo, obtuvo cuatro reglas de decisión asociadas a la clase que no se matriculan. Palabras clave: deserción universitaria, regresión logística binaria, datos desbalanceados, árbol de clasificación CART.

ABSTRACT The purpose of this research is to predict student dropout from a private university, applying two data mining techniques: binary logistic regression and CART classification tree. For the study, data was used from the student database in the period 2019-2 with 32,176 records with data related to socio-demographic, academic, and economic factors. Data balancing was applied with the subsampling technique in order to improve the predictive capacity. The CART classification tree resulted in higher values for accuracy, sensitivity, specificity, and AUC of 73.8%, 97.3%, 50.3%, and 73.8%, respectively, to predict college dropout compared to logistic regression. binary whose values were 66.4%, 71.2%, 65.8% and 72.4% respectively. The resulting tree identified the most important variables: TAS_NOM_A, TAS_NOM_P, SEDE, TAS_ASI_A; with a size of 13 nodes, with seven terminal nodes, of which three to predict the SI class and four to predict the NO class; likewise, he obtained four decision rules associated with the class that are not enrolled. Keywords: college dropout, binary logistic regression, unbalanced data, CART classification tree.

**UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA**

FACULTAD DE ECONOMIA Y PLANIFICACIÓN

**"MEJORA DEL INDICADOR DE RETENCIÓN EN UNA
UNIVERSIDAD PRIVADA A PARTIR DE LA CLASIFICACIÓN DE
ALUMNOS UTILIZANDO UN MODELO PREDICTIVO "**

PRESENTADO POR

JEAN PIERRE URIBE MOSTACERO

**TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR POR EL
TÍTULO DE INGENIERO EN ESTADÍSTICO E INFORMÁTICO**

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

**Dr. Rino Nicanor Sotomayor Ruiz
PRESIDENTE**

**Mg. Sc. Ivan Dennys Soto Rodríguez
ASESOR**

**Mg. Grimaldo José Febres Huamán
MIEMBRO**

**Mg. Sc. Aldo Richard Meza Rodríguez
MIEMBRO**

LIMA - PERÚ

2022

DEDICATORIA

A mis hijas Gianella y Antonella por ser mi motivo por seguir avanzando.

A mi esposa Vanesa por todo el apoyo que día a día me ha demostrado.

A mi madre María de los Ángeles, por toda la dedicación y el apoyo que me dio para lograr culminar mis estudios.

AGRADECIMIENTO

Por todo el apoyo técnico a los profesores:

Ph.D. Cesar Higinio Menacho Chiok

Mg. Sc Ivan Dennys Soto Rodríguez

ÍNDICE GENERAL

I. INTRODUCCIÓN.....	1
1.1 Problema de investigación	3
1.2 Justificación de la investigación	3
1.3 Objetivos de la investigación	4
4.1.2 Objetivo General	4
1.3.2 Objetivos Específicos	4
II. REVISIÓN DE LITERATURA	6
2.1. Antecedentes	6
2.2 La minería de datos.....	9
2.2.1 El proceso de la minería de datos	9
2.2.2 Técnicas de minería de datos para la clasificación.....	10
2.2.3 Técnicas para el balanceo de datos.....	12
2.2.4 Técnicas para la evaluación de clasificadores	13
2.3 Modelo de Regresión logística binaria.....	17
2.4 Modelo del árbol de clasificación	21
III. DESARROLLO DEL TRABAJO.....	25
3.1. La recolección de datos	26
3.2. Técnicas estadísticas para el procesamiento de datos	31
IV. RESULTADOS Y DISCUSIÓN.....	33
4.1. Pre procesamiento de datos.	33
4.1.1 Análisis exploratorio de datos	33
4.1.2 Manejo de datos faltantes y atípicos.....	37
4.1.3 Balanceo de la base de datos	39
4.2 Técnicas de minería de datos.....	40
4.2.1 Regresión logística binaria	40
4.2.2 Árbol de clasificación CART	44
4.3 Comparación de la capacidad predictiva de la regresión logística binaria y del árbol de clasificación CART.....	49
V. CONCLUSIONES.....	50
VI. RECOMENDACIONES	52
VII. REFERENCIAS BIBLIOGRÁFICAS	53

ÍNDICE DE TABLAS

Tabla 1 : Evolución de la deserción por semestre	3
Tabla 2: Matriz de confusión	15
Tabla 3:Estructura de la base de datos de análisis.....	28
Tabla 4: Estructura de la codificación de las variables	29
Tabla 5: Distribución de alumnos según estado de matrícula.....	37
Tabla 6:Distribución de alumnos según estado de matrícula depurado	38
Tabla 7:Distribución de alumnos según estado de matrícula balanceada	39
Tabla 8: Distribución de alumnos para el conjunto de entrenamiento y prueba	40
Tabla 9: Estimación de los coeficientes de la regresión logística binaria	41
Tabla 10: Matriz de confusión de la regresión logística binaria	42
Tabla 11: Medidas de la eficiencia predictiva para la regresión logística.....	43
Tabla 12: Distribución de reglas por nodo del árbol.....	45
Tabla 13: Matriz de confusión del árbol de clasificación CART	48
Tabla 14 Medidas de la eficiencia predictiva del árbol de clasificación CART	48
Tabla 15 Comparación de la eficiencia predictiva de los modelos predictivos CART	49

ÍNDICE DE FIGURAS

Figura 1:Proceso del descubrimiento del conocimiento en base de datos (KDD).....	10
Figura 2: Clasificación de las técnicas de minería de datos.....	11
Figura 3: La curva ROC.....	17
Figura 4: Estructura de un árbol de decisión	22
Figura 5: Población objetivo.	25
Figura 6: Factores relacionados a la deserción universitaria.....	26
Figura 7: Fuentes para la integración de la base de datos.....	27
Figura 8: Distribución de alumnos matriculados (SI y NO).....	33
Figura 9: Análisis de las variables cualitativas según la variable de matrícula.....	34
Figura 10: Análisis de las variables cuantitativas según la variable de matrícula	35
Figura 11: Matriz de correlaciones.....	36
Figura 12: Distribución de valores perdidos.....	38
Figura 13: Curva ROC para la comparación de la sensibilidad y especificidad.....	43
Figura 14: Determinación del parámetro de complejidad óptimo.....	44
Figura 15: Árbol de clasificación CART	46

RESUMEN

La presente investigación tiene propósito predecir la deserción estudiantil de una universidad privada, aplicando dos técnicas de la minería de datos la regresión logística binaria y árbol de clasificación CART. Para el estudio se utilizó datos la base de datos de alumnos en los periodos 2019-2 con 32176 registros con datos relacionados a factores socio-demográfica, académicos y económicos. Se aplicó el balanceo de datos con la técnica de submuestreo a fin de mejorar la capacidad predictiva. El árbol de clasificación CART resultó con mayores valores para la exactitud, sensibilidad, especificidad y AUC de 73,8%, 97,3%, 50,3% y 73,8% respectivamente para predecir la deserción universitaria en comparación de la regresión logística binaria cuyos valores fueron 66,4%, 71,2%, 65,8% y 72,4% respectivamente. El árbol resultó identificó las variables más importantes: TAS_NOM_A, TAS_NOM_P, SEDE, TAS_ASI_A; con un tamaño de 13 nodos, con siete nodos terminales, de los cuales tres para predecir la clase SI y cuatro para la clase NO; así mismo, obtuvo cuatro reglas de decisión asociadas a la clase que no se matriculan.

Palabras clave: deserción universitaria, regresión logística binaria, datos desbalanceados, árbol de clasificación CART.

ABSTRACT

The purpose of this research is to predict student dropout from a private university, applying two data mining techniques: binary logistic regression and CART classification tree. For the study, data was used from the student database in the period 2019-2 with 32,176 records with data related to socio-demographic, academic, and economic factors. Data balancing was applied with the subsampling technique in order to improve the predictive capacity. The CART classification tree resulted in higher values for accuracy, sensitivity, specificity, and AUC of 73.8%, 97.3%, 50.3%, and 73.8%, respectively, to predict college dropout compared to logistic regression. binary whose values were 66.4%, 71.2%, 65.8% and 72.4% respectively. The resulting tree identified the most important variables: TAS_NOM_A, TAS_NOM_P, SEDE, TAS_ASI_A; with a size of 13 nodes, with seven terminal nodes, of which three to predict the SI class and four to predict the NO class; likewise, he obtained four decision rules associated with the class that are not enrolled.

Keywords: college dropout, binary logistic regression, unbalanced data, CART classification tree.

I. INTRODUCCIÓN

Las instituciones de educación superior en cada semestre admiten a nuevos estudiantes que tienen como objetivo lograr una carrera profesional que les permita mejorar su estatus de vida. Así mismo, los estudiantes deben matricularse en cada ciclo académico en los diferentes cursos; sin embargo, muchos de estos estudiantes no se matriculan en algunos de los semestres o en ninguno, lo cual impide que no puedan acabar su carrera universitaria. El caso más álgido, es cuando deja de estudiar. Este abandono de los estudiantes en continuar sus estudios superiores, es conocido como la deserción universitaria que está afectado por múltiples factores; tales como, académicos, económicos, familiares, personales, salud, etc. La deserción universitaria, se ha convertido en un problema para las universidades; por lo cual están haciendo los mayores esfuerzos para identificar los factores que los originan a fin de monitorearlos y realizar acciones preventivas durante el proceso académicos que redunden a disminuir y controlar esta deserción universitaria. Según (Peralta, 2008), la deserción universitaria es un fenómeno que está latente en el sistema educativo, que deben ser evaluados desde los procesos de selección, rendimiento académico y de la propia eficiencia del sistema educativo en general. En (Ferreya et al., 2017), en un estudio sobre la educación en América Latina y el Caribe, afirma que “la mitad de la población de 25-29 años de edad que comenzaron la educación superior en algún momento no finalizaron sus estudios porque desertaron”; así mismo, se ha estimado que cerca del 30% de estudiantes que empiezan un programa universitario abandonarán el sistema de educación superior y más del 30% lo harán al final del primer año.

Frente al aumento de la deserción universitaria, muchas universidades están implementando unidades de retención que recogen información académica, social y económica del estudiante, con la finalidad de realizar acciones estratégicas que apoyen a minimizar los factores que influyen y motivan a un estudiante a desertar en la institución educativa. Así mismo, las instituciones universitarias son conscientes que los datos académicos y no académicos recopilados de los estudiantes, permiten desarrollar modelos analíticos en base de indicadores cuantitativos que pueden apoyar la prevención de la deserción universitaria, que no sólo va en

prejuicios del estudiante con respecto lo económico y personal, sino también afecta a la universidad por el esfuerzo e inversión que se pierde con el abandono de los estudiantes.

El uso de la estadística aplicando las técnicas de minería de datos para desarrollar modelos analíticos, es una propuesta que se está implementando en muchas universidades públicas y privadas para enfrentar el problema de la deserción universitaria. Existen modelos estadísticos capaces de predecir si el estudiante tiene una alta probabilidad de desertar en el siguiente ciclo de su matrícula. Estudios como el de (Yukselturk et al., 2014), cuyo objetivo era predecir la deserción de estudiantes de un curso online bajo las técnicas de minería de datos, es uno de los casos de aplicación de modelos estadísticos.

En el presente trabajo de investigación se plantea la implementación modelos estadísticos predictivos para la deserción universitaria, aplicando las técnicas de minería de datos de la regresión logística binaria y el árbol de clasificación CART. Se usó los datos los social-demográficas, académicos y económicos que el Área de Retención que se almacena semestralmente de los estudiantes de una universidad privada del Perú. La metodología se desarrolló bajo el enfoque conocido como KDD (Knowledge Discovery in Data bases) que comprende un conjunto de etapas para construir los modelos predictivos. Se inició con el pre procesamiento y el análisis exploratorio de los datos con la finalidad de manejar los datos atípicos y faltantes, la selección de variables y el problema del desbalanceo de los datos. Se dividió los datos seleccionando en forma aleatoria el conjunto de entrenamiento para la estimación y el conjunto de prueba para la validación de los modelos predictivos propuestos. Se evaluaron y compararon los modelos predictivos a partir de los indicadores de eficiencia predictiva como la exactitud, la sensibilidad y la especificidad.

Para el estudio se utilizó datos la base de datos de alumnos en los periodos 2018-2 al 2019-2 con 32176 registros con datos relacionados a factores socio-demográfica, académicos y económicos. El árbol de clasificación CART resultó con mayores valores para la exactitud, sensibilidad, especificidad y AUC de 73,8%, 97,3%, 50,3% y 73,8% respectivamente para predecir la deserción universitaria en comparación de la regresión logística binaria cuyos valores fueron 66,4%, 71,2%, 65,8% y 72,4% respectivamente. El árbol resultó identificó las variables más importantes: TAS_NOM_A, TAS_NOM_P, SEDE, TAS_ASI_A; con un tamaño de 13 nodos, con siete nodos terminales, de los cuales tres para predecir la clase Si y cuatro

apara la clase NO; así mismo, obtuvo cuatro reglas de decisión asociadas a la clase que no se matriculan. Finalmente, árbol de clasificación CART fue implementado para el seguimiento semestral de matrícula de los estudiantes; con lo cual el Área de Retención cuenta con una herramienta cuantitativa para predecir en términos de probabilidad los niveles (alta, media o baja) de deserción posible de un estudiante y así mismo realizar acciones de monitoreo durante el semestre académico de deserción universitaria a los estudiantes.

1.1 Problema de investigación

La deserción universitaria es uno de los problemas más importantes que deben enfrentar las universidades en el Perú y en el mundo, cientos de alumnos ingresan cada periodo sin embargo no se puede saber a certeza su permanencia, ya que existen diferentes motivos asociados a la deserción de un alumno, entre los motivos más conocidos tenemos los factores relacionados al ámbito económicos, académicos y socio demográficos. El presente trabajo de investigación se centra en el problema de la deserción de estudiantes en una universidad privada que se matricularon en un periodo (X) y no llegaron a matricularse en el periodo siguiente (X+1), considerando X como un periodo regular. En la Tabla 1, se presenta el indicador de deserción en los tres periodos regulares de estudio (2018-2, 2018-3 y 2019-2) con un promedio 9.8% de deserción con respecto al total de alumnos matriculados.

Tabla 1
Evolución de la deserción por semestre

Estado	2018-2	2018-3	2019-2
Matrícula	18,400	25600	32,176
Deserción	1,343	3,200	4536
% Deserción	7,3	12,5	14,1

En la Tabla 1, se muestra muestra la evolución de la deserción a lo largo de tres periodos.
Fuente de elaboración propia.

1.2 Justificación de la investigación

El área de retenciones de la universidad en estudio, fue creada para realizar gestión y seguimiento pre y post a la deserción de un alumno, el área en los último cuatro periodos a pesar de su gestión no ha podido controlar el crecimiento de la deserción en la universidad. La diversidad de los factores motivadores y el dinamismo de la deserción estudiantil, se hace difícil identificar con anticipación dichos factores a fin de implementar medidas y acciones de monitoreo y seguimiento. Es necesario contar con modelos estadísticos predictivos que permitan identificar las posibles causas relacionadas a la deserción universitaria, convirtiéndose en una herramienta para el área de retenciones de la universidad; además, que permita dar seguimiento de manera semestral a los estudiantes, a partir de la clasificación de estudiantes con alta probabilidad de deserción; así, como revertir gradualmente el indicador de deserción el mismo que está acompañado de un impacto económico negativo a la universidad. Los modelos predictivos de la regresión logística binaria y el árbol de clasificación propuestos permitirán cuantificar niveles de probabilidad que un estudiante deserte y a su vez tomar acciones anticipadas a fin de disminuir la deserción universitaria de la universidad.

1.3 Objetivos de la investigación

El objetivo general y los específicos de la presente investigación son los siguientes:

4.1.2 Objetivo General

Desarrollar un modelo de estadístico que permita predecir la deserción universitaria aplicando regresión logística binaria y árboles de clasificación CART que permita al Área de Retención contar con una herramienta para apoyar las acciones con respecto al abandono de los estudiantes.

1.3.2 Objetivos Específicos

Los objetivos específicos son los siguientes:

- Identificar las principales variables que influyen en la deserción universitaria.
- Construir los modelos estadísticos predictivos aplicando la regresión logística binaria y el árbol de clasificación CART para la deserción de los estudiantes.

- Comparar la capacidad predictiva de los modelos predictivos de la regresión logística binaria y árbol de clasificación CART para implementar el mejor modelo en el Área de Retención.

II. REVISIÓN DE LITERATURA

2.1. Antecedentes

La deserción universitaria es el acto de abandonar las actividades académicas por parte de los estudiantes en cualquier momento del periodo de la carrera profesional. Para la presente investigación se define la deserción universitaria como el acto de no matricularse en el periodo semestral siguiente del que se encuentra cursando el alumno. Según (Díaz Peralta, 2008), indica que la deserción universitaria es un fenómeno que está latente en el sistema educativo, con los procesos de selección, rendimiento académico y de la propia eficiencia del sistema en general. En (Eckert y Suenaga, 2014), se define el concepto de deserción universitaria como una situación a la que se enfrenta un estudiante cuando sus proyectos educativos no logran concretarse, podemos considerar como desertor a aquel estudiante que no presenta actividad académica durante tres semestres académicos consecutivos. Según (Tudela, 2014), la deserción estudiantil tiene consecuencias sociales, emocionales y económicas no solo en el propio estudiante sino en su entorno más cercano. Adicionalmente, quienes no concluyen sus estudios son subempleados y no obtienen los ingresos económicos deseados.

A continuación, se realizará un recuento de algunas investigaciones que han abordado el tema de la deserción universitaria apoyándose en las técnicas de minería de datos.

En (Yukselturk et al., 2014), se realizó un estudio con el objetivo de predecir la deserción de estudiantes de un curso online bajo las técnicas de minería de datos. El estudio fue realizado a 189 estudiantes entre el 2007 y 2009, recolectándose los datos por cuestionarios en línea, donde se obtuvo 10 variables como género, edad, nivel educativo, experiencia previa en línea, ocupación, autoeficacia, preparación, conocimiento previo, lugar de control y Deserción (Si/No). Para la construcción del modelo se aplicó validación cruzada 10 veces y cuatro modelos distintos como en k-Nearest Neighbour (k-NN), Decision Tree (DT), Naive Bayes (NB) y Neural Network (NN), finalmente se analizan los indicadores de eficiencia predictiva

encontrando que en la sensibilidad de los cuatro clasificadores 3-NN, DT, NN y NB fueron 87%, 79.7%, 76.8% y 73.9% respectivamente. Siendo está considerado alta y con un valor predictivo. En (Pal, 2012), se realizó un estudio en una institución superior de la India usando los datos de los estudiantes de la carrera de ingeniería, con el objetivo de crear modelos predictivos utilizando minería de datos. La aplicación se basó en el uso del algoritmo Naive Bayes como herramienta de clasificación de estudiantes con riesgo de deserción, siendo las variables más significativas el Medio de enseñanza, Sexo, Calificación de ingreso del estudiante, Tipo de admisión, Lugar de residencia del estudiante, Profesión de la madre y Nivel educativo del Padre o Madre); el modelo predictivo final obtuvo un accuracy de 91.7%, siendo los resultados satisfactorios para la predicción de alumnos nuevos.

En (Tan y Shao, 2015) realizaron un estudio en con el fin de generar un modelo de deserción dirigido a alumnos de un curso online una universidad de China, donde existían alta tasa de deserción universitaria. El estudio fue aplicado a una muestra de 62,375 estudiantes, con los que se construyeron modelos aplicando tres diferentes algoritmos de minería de datos, algoritmos de redes neuronales (ANN), Árboles de decisión (DT) y las redes bayesianas (BN), finalmente los modelos fueron evaluados en base a tasas de precisión, precisión, recuperación y medida. Encontrando que la efectividad general de predicción fueron DT (71,91%), BN (69,19%) y ANN (65, 65%). Concluyendo que los tres modelos fueron efectivos para la predicción de la deserción universitaria, pero DT presentó un mejor desempeño.

En (Eckert y Suenaga, 2014), se realizó un estudio con el objetivo de identificar factores que influyen sobre la deserción de los estudiantes de la carrera de Ingeniería en Informática de la Universidad Gastón Dachary en Argentina, mediante la aplicación de la minería de datos. El estudio se realizó con una muestra de estudiantes de la carrera de Ingeniería en Informática, modalidad presencial del año 2000 a 2009 totalizando 855 alumnos analizados, los datos de análisis fueron datos personales y antecedentes del alumno. En la construcción del modelo se realiza la selección y depuración de datos, encontrándose como variables influyentes en la deserción, asignaturas aprobadas, cantidad y resultado de asignaturas cursadas, procedencia y edad de ingreso del estudiante. Los modelos aplicados en el estudio fueron los algoritmos de árboles de decisión (C4.5) y redes bayesianas. Los resultados respecto a la tasa de predicción

fueron 79.7% para el C4.5 y 81.1% para las redes bayesianas, ambos eficientes para el estudio de la deserción.

En (Sivakumar et al., 2016), se realizó un estudio con el objetivo de identificar los factores de mayor importancia sobre la deserción universitaria de los estudiantes de pregrado de una universidad de la India. Para este estudio las variables estuvieron relacionadas a factores sociodemográficos y académicos. La técnica utilizada para modelar la deserción fue el árbol de clasificación a partir del algoritmo ID3. Los resultados demostraron que el árbol de decisión proporciona una mejor precisión de predicción en los datos educativos que la de los algoritmos de clasificación tradicionales. Se propuso un algoritmo de decisión mejorado que mejora la capacidad de formar árboles de decisión y, por lo tanto, demostrar que la precisión de clasificación del algoritmo de decisión mejorado en el conjunto de datos educativos es mayor.

En (Martín et al., 2019), se realizó una investigación aplicada a estudiantes del instituto tecnológico de Costa Rica con el fin de modelar la deserción universitaria a partir de datos registrados por la institución. Para esta investigación se utilizaron 6 algoritmos de aprendizaje automático (Random Forest, boosted Trees, Redes Neuronales, SVM, sigma, R, Logística), con el fin de contrastar los niveles de sensibilidad. Los resultados obtenidos fueron proporcionados por el algoritmo random forest ya que clasificó correctamente al estudiante desertor con una probabilidad de 0.83 y por capturar al 34% de deserción real. Así mismo se obtuvo como resultado que la deserción universitaria está altamente relacionada a las variables sociodemográficas, programa de estudio, beneficios obtenidos al ingresar, historial académico y rendimiento en el primer semestre de estudio.

En (Manhães et al., 2014), se realiza una investigación aplicada a estudiantes de pregrado de una universidad de Brasil, con el objetivo de identificar a los alumnos en riesgo a partir del uso de técnicas de minería de datos. El estudio fue realizado a 14000, los cuales fueron clasificados bajo tres categorías diferentes, referente a la trayectoria académica. Las técnicas utilizadas fueron los algoritmos C4.5, Simple Cart, SVM, Naïve Bayes y Multilayer Perceptron. Los resultados fueron satisfactorios con el algoritmo Naïve Bayes ya que obtuvo una precisión global superior a 80%. La calidad de los resultados abre las posibilidades novedosas de nuevas investigaciones, el desarrollo de un sistema de información capaz de facilitar la gestión de las universidades académicas.

2.2 La minería de datos

La minería de datos (MD), comprende un conjunto de técnicas provenientes de los campos de la estadística y el aprendizaje automatizado que son aplicadas con la finalidad de extraer conocimiento en las bases de datos. Los algoritmos de MD se enmarcan en el proceso conocido como El descubrimiento del Conocimiento en Base de Datos (*KDD: Knowledge Discovery from Databases*). Sin embargo, muchas veces se considera la MD como el proceso en sí de KDD. La técnica de minería de datos a aplicar dependerá en general del tipo de aprendizaje supervisado o no supervisado y del tipo de conocimiento que se quiere descubrir (descriptivo o predictivo). En (Klenzi y López, 2017), la minería de datos se puede definir inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas tendencias especializadas englobadas bajo el nombre de minería de datos. Las técnicas de minería de datos (TMD) persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Las TMD, se están aplicando en diferentes áreas, como la educación generando toda una comunidad denominada Minería de Datos Educativos.

2.2.1 El proceso de la minería de datos

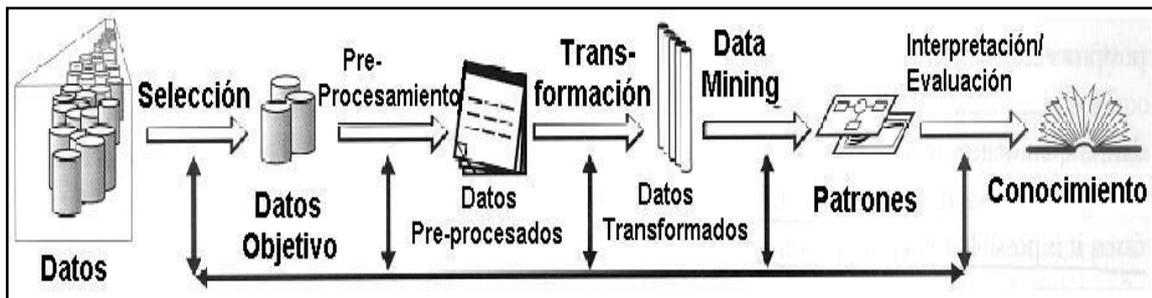
La aplicación de las técnicas de minería de datos bajo un enfoque del KDD, consiste en un proceso iterativo que comprende la ejecución de un conjunto de etapas con la finalidad de buscar conocimiento valioso, novedoso, útil y entendible de una base de datos.

En la Figura 1, se presenta las etapas del KDD que inicia con la etapa de la selección de los datos provenientes de las diversas fuentes internas o externas. La etapa del Pre procesamiento de datos, que permite mejorar la calidad de los datos para obtener los conocimientos válidos de las bases de datos. Comprende técnicas para la transformación y limpieza de los datos, que incluye el manejo de datos atípicos, datos faltantes, discretización de datos, selección de atributos y balanceo de datos. La etapa de la propia aplicación de las TM. Según El tipo de

conocimiento que se desea descubrir en los datos se determina la técnica de MD a aplicar. Las metas de las técnicas de MD pueden ser: extraer patrones, tendencias y regularidades para describir y comprender mejor los datos, extraer patrones y tendencias para predecir comportamientos futuros, extraer patrones y tendencias para clasificar nuevos datos. La etapa de la validación las TM, aplicando técnicas de bondad de ajuste o inferencia estadística para corroborar las hipótesis propuestas. La interpretación de los patrones descubiertos, puede beneficiarse grandemente usando visualización. Puede borrar patrones redundantes o irrelevantes. Los patrones pueden compararse con conocimiento previamente almacenado o repetir el proceso con otros datos u otros algoritmos, usando métodos de evaluación y validación.

Figura 1

Proceso del descubrimiento del conocimiento en base de datos (KDD)



En la Figura 1, se muestra las etapas que sigue de la metodología KDD para la obtención del conocimiento.

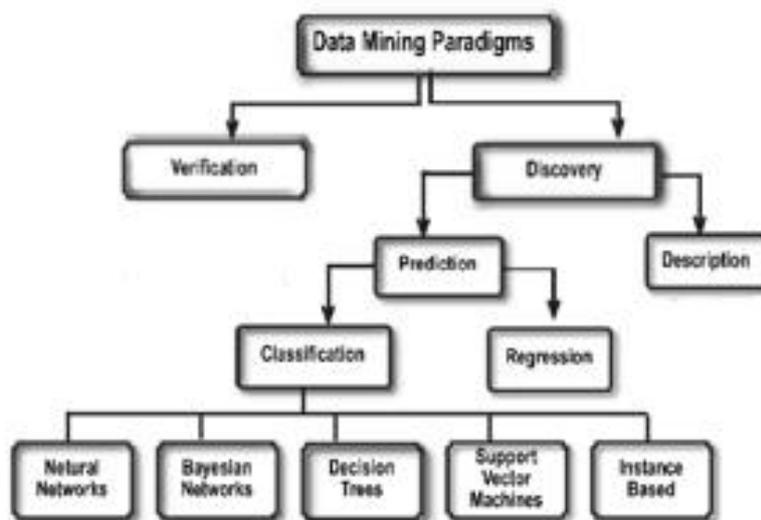
Fuente de elaboración propia.

2.2.2 Técnicas de minería de datos para la clasificación

Existen una gran variedad de técnicas de minería de datos (TMD) que son aplicadas según el propósito y objetivo del conocimiento que se desea extraer o descubrir. Las variedades de las TMD pueden ser clasificadas en dos grupos, según el tipo de aprendizaje que se desea realizar a partir de los datos:

- **Aprendizaje supervisado.** La fuente de información para el aprendizaje, es un conjunto de ejemplos clasificados o etiquetados en clases y suministrados por el experto. El sistema debe obtener los conceptos en base de la descripción para cada clase.
- **Aprendizaje no supervisado.** El sistema debe agrupar los conceptos sin ayuda alguna de un experto. El sistema recibe ejemplos, que debe observar y buscar características en común que permitan formar grupos. El resultado de este aprendizaje es un conjunto de descripciones de cada clase (clase de interés).

Figura 2
Clasificación de las técnicas de minería de datos



En la Figura 2, se muestra una taxonomía de las TMD clasificadas según el objetivo técnica. Fuente de elaboración propia.

Las TMD, se caracterizan por tener la capacidad de descubrir e identificar automáticamente los patrones de comportamiento en los datos. Las TMD pueden ser enfocadas al descubrimiento de tareas o problemas de la predicción versus a los de la descripción. Las TMD descriptivas están orientados a la interpretación de datos, que se centra en la comprensión por visualización, por ejemplo, la forma en que los datos subyacentes se relacionan con sus partes. Mientras que las TMD orientadas a la predicción tienen como objetivo construir automáticamente un modelo de comportamiento, que tienen la capacidad de predecir los valores de una variable dependiente

cuantitativa (tarea de regresión) o cualitativa (tarea de clasificación) en función de un conjunto de variables independientes o predictoras. También desarrolla patrones, que forman el descubrimiento conocimiento de una manera que sea comprensible y fácil de operar. El aprendizaje no supervisado está asociado a las TMD descriptivas, mientras las TMD de predicción se relacionan al aprendizaje supervisado (Regresión y clasificación).

2.2.3 Técnicas para el balanceo de datos

El desbalanceo de datos se presenta en el aprendizaje supervisado, cuando alguna clase se presenta en forma mayoritaria. El problema de datos desbalanceados, afecta la tabla de confusión en los clasificadores, tendiendo a sobre clasificar las observaciones de la clase mayoritaria (minimizando su tasa de error). Según (Cortes y Vapnik, 1995), los datos desbalanceados son aquellos que presentan una desproporción notable en el número de instancias pertenecientes a cada clase. Ello provoca un sesgo en el desempeño de los clasificadores estándares hacia el reconocimiento de las clases más numerosas. Entre las aplicaciones donde se puede observar prevalencia de datos desbalanceados se pueden citar: detección de fraude e intrusión, manejo de riesgo, clasificación de texto, detección de fallas en procesos industriales, diagnóstico y monitoreo médico, etc. En (Hoyos Osorio, 2019), se manifiesta la gran importancia de considerar el problema del desbalance de clases en diversos campos de aplicación, distintos algoritmos y métodos que han sido propuestos en la última década para abordar el problema de clasificación de datos desbalanceados.

Frente a la existencia de datos desbalanceado, en el cual se requiere que el clasificador proporcione una alta precisión para la clase minoritaria, pero sin poner en peligro la precisión de la clase mayoritaria. En tal sentido, se han propuesto tres estrategias básicas: i) métodos de re muestreo, los cuales son técnicas de pre proceso que intentan equilibrar las distribuciones al considerar las proporciones representativas de los ejemplos de clase en la distribución, ii) métodos de aprendizaje costo-sensitivos, los cuales consideran los costos asociados con la clasificación errónea de las muestras y iii) métodos de ensamble que consisten en la

combinación de dos o más clasificadores. En general se plantea dos enfoques para el desbalanceo que se basan en el re muestreo: sub muestreo y sobre muestreo.

Método de submuestreo

Se basa en reducir los datos de la clase mayoritaria; esto es, descartar muestras de la clase mayoritaria de acuerdo a algún criterio. El método más simple es el submuestreo aleatorio (RUS), que implica la eliminación aleatoria de los ejemplos de la clase mayoritaria hasta balancear la base de datos. Recientemente han sido implementados algoritmos más avanzados que hacen dicha eliminación basada en agrupamiento (k-means). Los métodos de submuestreo basados en agrupamiento buscan seleccionar las muestras más representativas de la clase mayoritaria, particionado la base de datos en un número k de grupos usando algoritmos de clustering. Una vez se ha hecho esto, se selecciona un número adecuado de muestras mayoritarias de cada conglomerado considerando la relación entre el número de muestras minoritarias y mayoritarias en cada grupo. Por su parte, los métodos basados en distancia establecen algunas reglas de selección de muestras de la clase mayoritaria, de tal forma que se preserven aquellas observaciones cuya distancia promedio a una cantidad de muestras más cercanas de la clase minoritaria son las más pequeñas o, por el contrario, seleccionan los ejemplos cuya distancia promedio a las l-muestras de la clase minoritaria más lejanas son las más pequeñas. El submuestreo en comparación con los métodos de sobremuestreo, usualmente ofrece un mayor rendimiento en tareas de clasificación, sin embargo, por la necesidad de balancear la base de datos se pierde mucha información y se modifica la distribución original de los datos, lo que podría ocasionar sobre-entrenamiento.

2.2.4 Técnicas para la evaluación de clasificadores

Las TMD referidos a la clasificación son evaluadas a través de medir su capacidad predictiva (tasa de buena clasificación). El objetivo de los métodos de evaluación es medir la precisión del clasificador como una medida de la bondad de ajuste de la TMD con el conjunto de datos de entrenamiento. Así mismo, permiten comparar entre varias técnicas de clasificación y

seleccionar la que tenga la mayor precisión. Para la evaluación de las TMD, se proponen el uso de la matriz de confusión, área bajo la curva ROC (AUC) y el coeficiente Kappa.

Métodos para la selección del conjunto de entrenamiento y prueba

Para aplicar las técnicas de validación, antes se debe usar algún método que permita dividir la base de datos, seleccionando en forma aleatoria un conjunto de entrenamiento (CE) y un conjunto de prueba (CP). El CE se usa para la estimación del modelo (**el entrenamiento**) y el CP para probar el ajuste del modelo (**la validación**). A partir del CP se construye la tabla de confusión y métricas de evaluación. Entre los métodos para dividir la base de datos se tiene:

- **Método de retención (Hold-Out).** El método se basa en dividir el conjunto de aprendizaje, en dos conjuntos independientes: Entrenamiento (CE) y Test (CT). El tamaño del CE normalmente es mayor que el CT (2/3,1/3),(4/5,1/5),etc. Los elementos del CE suelen obtenerse por muestreo sin reemplazo y el CT se forma por los datos no incluidos en el CE. Con el CE se realiza el aprendizaje del método de clasificación y con el CT se realiza su validación. Este método se usa con conjuntos de datos grandes.
- **Validación cruzada (cross-validation).** Este método se basa en dividir la base de datos en k subconjuntos (folds) de igual tamaño. Con cada una de las partes se realiza el aprendizaje usando un CE distinto de tamaño (k-1) y el resto como CT. La tasa de acierto (error), se calcula como el promedio obtenido de las k iteraciones. Valores típicos de k=5 y 10 pliegues. Suele utilizarse para conjuntos de datos moderado.
- **Dejar-uno-afuera (Leave-one-out).** Esta es una técnica de validación cruzada de n pliegues, donde n es el número de ejemplos del conjunto de datos. Por turnos, cada uno de los ejemplos se queda afuera y se entrena el clasificador con el resto (n-1). Los resultados de las n evaluaciones se promedian para determinar la proporción de error.

La Matriz de confusión

La matriz de confusión es una tabla de contingencia que muestra la distribución de la clasificación observada (real) y la predicha (clasificador) para las distintas categorías de la variable clase. En la Tabla 2 se muestra la matriz de confusión para el caso de dos clases.

Tabla 2
Matriz de confusión

Clasificación observada	Clasificación predicha		Total (Observado)
	Positiva (Clase 0)	Negativa (Clase 1)	
Positiva (Clase 0)	VP	FN	VP + FN
Negativa (Clase 1)	FP	VN	FP + VN
Total (Predicho)	VP + FP	FN + VN	N=VP+FN+FP+VN

En la Tabla 2, se muestra las fórmulas del cálculo de la matriz de confusión.

Fuente de elaboración propia.

- **Vverdadero positivo (VP).** Es el número de observaciones que predice correctamente el clasificador para la clase positiva.
- **Falsos positivos (FP).** Es el número de observaciones que se predice incorrectamente como la clase positiva siendo de la clase negativa.
- **Verdaderos negativos (VN).** Es el número de observaciones que predice correctamente el clasificador para la clase negativa.
- **Falsos negativos (FN).** Es el número de observaciones que se predice incorrectamente como la clase negativa siendo de la clase positiva.

A partir de la matriz de confusión, se pueden determinar las siguientes métricas:

$$TA = \frac{VP + VN}{N}$$

La tasa de buena clasificación. Mide la proporción de observaciones que el clasificador predice correctamente la clase positiva y negativa.

$$TE = \frac{FN + FP}{N} = 1 - TA$$

La tasa de mala clasificación. Es la proporción de observaciones que el clasificador predice incorrectamente.

$$TVP = \frac{VP}{VP + FN}$$

La tasa de verdaderos positivos. Es la proporción o porcentaje de observaciones que se predice correctamente la clase positiva con respecto a su total observado (totales de filas). Esta métrica es conocida como **la sensibilidad**.

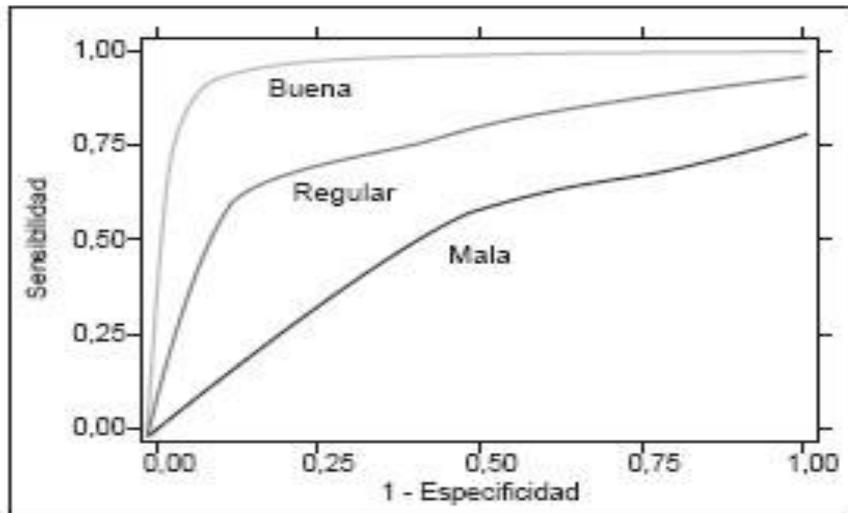
$$TVN = \frac{VN}{FP + VN}$$

La tasa de verdaderos negativos. Es la proporción o porcentaje de observaciones que se predice correctamente la clase negativa con respecto a su total observado (totales de filas). Es conocida como **la especificidad**.

Análisis de la curva ROC

El análisis ROC (Receiver Operating Characteristic), es usado para comparar y evaluar si un clasificador es mejor que otro, seleccionando aquel que tenga el mejor porcentaje de buena clasificación. La curva ROC representa para cada posible elección del valor de corte, la (1-*especificidad*) en el eje x y la *sensibilidad* en el eje y. En la Figura 3, se muestra la curva ROC, es creciente al modificar los valores de corte, para obtener mayor sensibilidad, se debe disminuir al mismo tiempo la especificidad. Si el clasificador no discrimina entre las clases, la curva ROC sería la diagonal. La exactitud de la prueba aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo. Si la clasificación fuera perfecta (100% de sensibilidad y 100% de especificidad) la curva pasaría por dicho punto.

Figura 3:
La curva ROC



En la Figura 3, se muestra la curva ROC y las diferentes clasificaciones según el valor del indicador.

Fuente de elaboración propia.

Área bajo la curva ROC (AUC)

Es toda el área bidimensional por debajo de la curva ROC. AUC se usa como un índice de la performance del clasificador (la exactitud máxima correspondería a un valor del área bajo la curva de 1 y la mínima a un valor de 0.5). El AUC proporciona una medición agregada del rendimiento en todos los umbrales de clasificación posibles. Una forma de interpretar el AUC es como la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio. La AUC se calcula con la siguiente expresión:

$$AUC = \frac{1 + TVP - TFP}{2}, \quad 0 \leq AUC \leq 1$$

2.3 Modelo de Regresión logística binaria

El modelo de regresión logística es una extensión del modelo de regresión lineal, donde la variable dependiente Y es binaria (0 ó 1) que define la posibilidad de ocurrencia de sólo dos eventos o valores (Éxito o Fracaso); de tal manera que la ecuación estimada busca predecir la probabilidad de ocurrencia de cada uno de sus valores. Según (Agresti, 2018), los datos binarios

son la forma más común de datos categóricos. El modelo más popular para datos binarios es la regresión logística, este modelo como un modelo lineal generalizado (GLM) para un componente aleatorio asociado a la distribución Binomial y con la función de enlace Logit. La regresión logística binaria es una TMD asociada a un problema de clasificación con aprendizaje supervisado que se aplica a muchos campos o áreas del mundo real. En el campo de la salud se usa para estudiar el riesgo que tiene una persona de tener o no una determinada enfermedad a partir de su historia médica; en finanzas se utiliza para estudiar el riesgo que tiene una entidad bancaria de que sus clientes cumplan con pagar las cuotas de su préstamo a partir de su perfil crediticio, etc. Según (Mount et al., 2014) regresión logística es el miembro más importante (y probablemente el más utilizado) de una clase de modelos llamados modelos lineales generalizados.

El modelo de la regresión logística binaria

El modelo de regresión logística binaria, trata de explicar la variable respuesta binaria Y (1 o 0) en términos de que tan probable suceda el evento de interés ($Y=1$), en función de un conjunto de variables predictoras. El modelo de regresión logístico binario, usa como función de enlace la función Logit que permite relacionar el logaritmo de la razón de la probabilidad $P(Y=1)=\pi$ y $P(Y=0)=1-\pi$ con el predictor lineal. Entonces se tiene el modelo Logit:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = X' \beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

El modelo de regresión logístico permite predecir la probabilidad $Y=1$ usando la función de enlace Logit para relacionarla con el predictor lineal. El modelo logístico binario se expresa por:

$$P(Y = 1) = \pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p))}$$

Interpretación de los coeficientes de regresión

En los modelos de regresión logística binaria la interpretación de los coeficientes estimados no es directa como el caso de la regresión lineal; debido a la naturaleza del modelo no lineal. La

interpretación de los coeficientes dependerá de los cambios en la escala de medida y en el tipo de la variable explicativa (factor, dicotómicas, categórica, covariable, continua, etc). Se usa el concepto de **Cociente de ventaja (OR: Odds ratios)**. El cociente o razón de ventaja, es una medida que indica que si es más probable (coeficiente de regresión positivo) o menos probable (coeficiente de regresión negativo) para la variable dependiente se presente el evento de interés (Y=1) con respecto a que no se presente (Y=0), y considerando un valor determinado de una variable explicativa. Entonces se tiene definido el coeficiente de ventaja:

$$\text{El Odds ratios para la variable } X_j : OR_j = e^{\hat{\beta}_j}$$

De acuerdo al signo del coeficiente de regresión, la interpretación del OR asociado a la variable predictora indicaría:

$$\left\{ \begin{array}{l} \text{Si } \hat{\beta}_j > 0 \text{ se tiene que } e^{\hat{\beta}_j} > 1 \text{ , entonces la variable } X_j \\ \text{incrementa la razón de ventaja para el valor } Y = 1. \\ \text{Si } \hat{\beta}_j < 0 \text{ se tiene que } e^{\hat{\beta}_j} < 1 \text{ , entonces la variable } X_j \\ \text{disminuye la razón de ventaja para el valor } Y = 1. \end{array} \right.$$

Para una variable predictora cualitativa, el valor de OR indica que tan probable es que aumente o disminuya el evento de interés de la variable dependiente (Y=1), cuando la variable explicativa toma el valor X=1 con respecto a que tome el valor X=0.

Prueba de hipótesis de los coeficientes de regresión

Se usa la prueba de Wald para probar la significación de un coeficiente de regresión y por lo tanto la influencia de la variable predictora asociada.

- **Formulación de hipótesis.** Se plantea una prueba bilateral para probar la significación de cada coeficiente de regresión. La hipótesis nula (H₀) y alterna (H₁) son:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

- **Prueba estadística.** Se usa la estadística Z (Normal estándar) como estadístico de prueba.

$$Z_c = \frac{b_j}{S_{b_j}} \quad |Z_c| > Z_{1-\frac{\alpha}{2}}$$

- **La decisión estadística.** Para tomar la decisión estadística, se rechaza H_0 , si:

Prueba de hipótesis del modelo

Es la medida más usada en los MLG para evaluar el ajuste de un modelo es la Deviance. Es la distancia entre el logaritmo de la función verosimilitud del modelo saturado (con N parámetros) y el modelo en investigación (con p parámetros).

- **Formulación de hipótesis.** Las hipótesis que se formulan para la estadística de deviance son:

H_0 : El modelo de regresión se ajusta a los datos

H_1 : El modelo de regresión no se ajusta a los datos

- **Prueba estadística.** Se usa Deviance como estadística de prueba.

- **La decisión estadística.** Para tomar la decisión estadística, se rechaza H_0 , si:

$$D_c \geq \chi^2_{(N-p)}$$

Un valor pequeño de la Deviance, indica que, para un número menor de parámetros, se obtiene un ajuste tan bueno como cuando se ajuste con un modelo saturado.

Coefficiente de determinación.

Se define como la reducción proporcional en la incertidumbre debido a la inclusión de los regresores. El Pseudo R^2 (McFadden), se ha propuesto como una medida relativa de la mejora de la log-verosimilitud. A mayor log-verosimilitud, mejor será el modelo ajustado. Compara el log verosimilitud del modelo de interés con respecto al modelo mínimo (sólo intercepto).

$$Pseudo R^2 = 100 \times \left(1 - \frac{D(y, \hat{y})}{D(y, \hat{y}_0)} \right) = 100 \times \left(1 - \frac{l(b; y)}{l(b_{\min}; y)} \right)$$

Son las desviaciones del modelo ajustado y el modelo nulo o mínimo (sólo con intercepto) respectivamente. El Pseudo R^2 , mide el porcentaje de cuanto se reduce la desviación del modelo nulo (con intercepto), cuando se adiciona las p variables predictoras.

Criterio de información Akaike (AIC=Akaike information criterios). Es una medida relativa para evaluar la bondad de ajuste de los modelos estadísticos. Se selecciona el modelo con menor AIC. En general, el AIC se calcula:

$$AIC = 2k - 2\ln(L)$$

Donde:

k= Números de parámetros del modelo

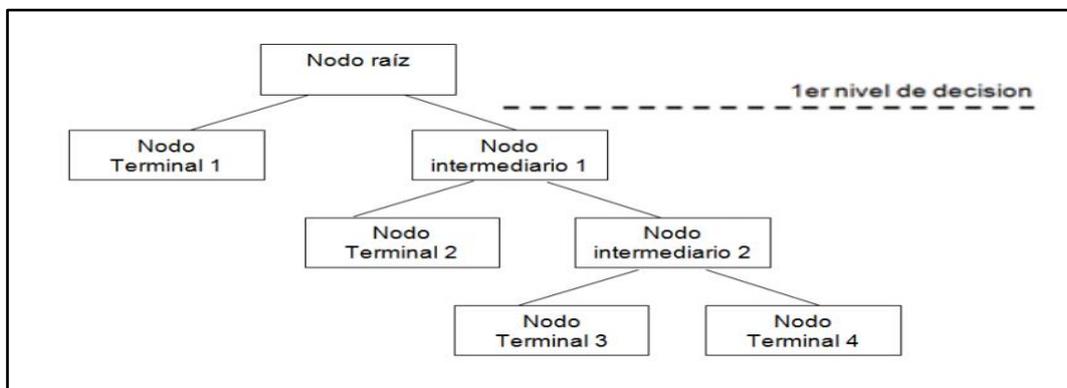
L =Valor máximo de la función verosimilitud para el modelo estimado.

2.4 Modelo del árbol de clasificación

Los árboles de clasificación son TMD para el aprendizaje supervisado que se basa en ir dividiendo el conjunto de la base de datos en dos o más grupos de acuerdo a las categorías que presenta un atributo. Un árbol de clasificación presenta una estructura jerárquica en forma de árbol, en donde las ramas representan conjuntos de decisiones; estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos de datos disjuntos y exhaustivos. Las ramificaciones se generan de forma recursiva hasta que se cumplan ciertos criterios de parada. Los árboles de decisión son empleados para clasificar y pronosticar, es decir identificar el resultado categórico atendiendo a una serie de criterios dados y pronosticar el resultado según una futura serie de criterios o variables independientes. El objetivo de este método es obtener individuos u objetos más homogéneos con respecto a la variable discriminadora dentro de cada subgrupo y heterogéneos entre los subgrupos. Para la construcción del árbol se requiere información de variables explicativas a partir de las cuales se va a realizar la discriminación de la población en subgrupos. Dentro de los métodos basados en árboles se pueden distinguir dos tipos dependiendo de tipo de variable a discriminar: Árboles de clasificación. Este tipo de árboles se emplea para variables categóricas, tanto nominales como ordinales. Árboles de regresión. Este tipo de discriminación se aplica a variables continuas. Diferentes algoritmos pueden ser usados para construir arboles de decisión tales como la Detección Automática de Interacciones (CHAID) y Árboles de Regresión (CART), QUEST y C5.0. El Árbol de Clasificación comienza con un nodo al que pertenecen todos los

casos de la muestra a clasificar (nodo raíz), el resto de nodos se dividen en nodos intermedios o no terminales y nodos hojas o nodos terminales. Un árbol de decisión consta de los siguientes elementos: **Nodo intermedio:** se generan dos o más segmentos descendientes inmediatos (dependiendo del método empleado). También llamados segmentos intermedios. **Nodo terminal:** Es un nodo que no se puede dividir más., como se muestra en la Figura 4.

Figura 4
Estructura de un árbol de decisión



En la Figura 4, se muestra la estructura de un árbol de decisiones.
Fuente de elaboración propia.

Árbol de clasificación CART

El procedimiento para este modelo no utiliza un modelo estadístico formal fue desarrollado por matemáticos de la universidad de Berkeley y Stanford (Breiman, Friedman, Olshen y Stone) a mediados de los 80, para clasificar, se utilizan particiones binarias sucesivas de los valores de una variable, trabaja con variables de todo tipo. El corte en cada nodo viene dado por reglas de tipo binario.

Medidas de Impureza

Una medida cuantitativa de la homogeneidad es la noción de impureza, medida de la siguiente manera:

$$Imp. de un nodo = \frac{\text{Número de sujetos que cumplen la característica en el nodo}}{\text{Número total de sujetos en el nodo}}$$

Para decidir qué variable va a utilizarse para hacer la partición en un nodo se calcula primero la proporción de observaciones que pasan por el nodo para cada uno de los grupos. Si se denomina a los nodos como $t = 1, 2, \dots, T$ y $p(g/t)$ a las probabilidades de que las observaciones que lleguen al nodo t pertenezcan a cada una de las clases, se define la impureza del nodo t como

$$i(t) = \phi \left(p \left(\frac{1}{t} \right), p \left(\frac{2}{t} \right), \dots, p \left(\frac{G}{t} \right) \right)$$

Dónde: ϕ es la función de impureza y, $p(g/t)$ puede calcularse empíricamente como la proporción de casos de clase g en el nodo t . Es decir:

$$p(g/t) = \frac{n_g(t)}{n(t)}$$

La variable que se introduce en un nodo es la que minimiza la heterogeneidad o impureza que resulta de la división en el nodo. La clasificación de las observaciones en los nodos terminales se hace asignando todas las observaciones del nodo al grupo más probable en ese nodo, es decir, el grupo con máxima $p(g/t)$. Si la impureza del nodo es cero, todas las observaciones pertenecerían al mismo nodo, en caso contrario puede haber cierto error de clasificación. Cuando el número de variables es grande, el árbol puede contener un número excesivo de nodos por lo que se hace necesario definir procedimientos de poda o simplificación del mismo.

Bondad de una Partición (Índice de GINI)

Se tiene el índice de Gini en el nodo t , $i(t)$, definida como:

$$i(t) = g(t) = 1 - \sum_{g=1}^G p(g/t)^2$$

Este índice es una medida de impureza en la clasificación de los datos, a medida que se van clasificando correctamente los datos, el índice de Gini va tomando valores cercanos a 0.

La función del criterio Gini $\phi(s,t)$ para la división s en el nodo t se define como:

$$\phi(s/t) = g(t) - p_L g(t_L) - p_R g(t_R)$$

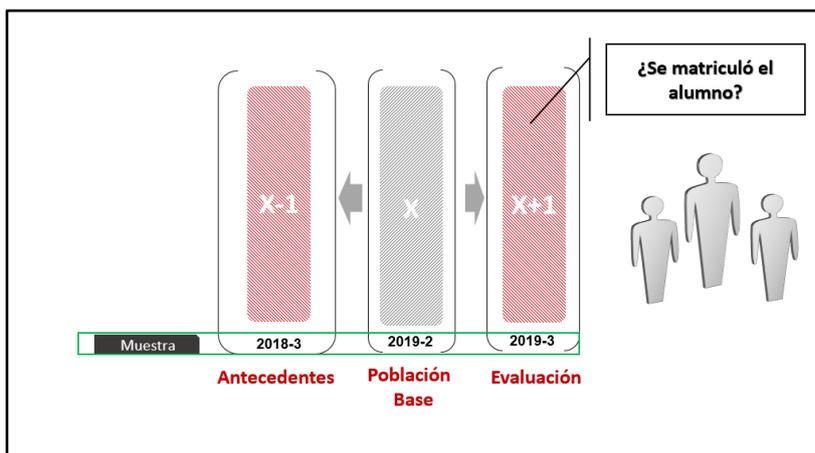
Así, como conocemos cómo calcular $g(t)$, podemos calcular $\emptyset(s,t)$ para cada partición s y seleccionar la mejor partición como la que proporciona la mayor bondad $\emptyset(s,t)$. Para establecer el efecto que produce la selección de la mejor partición en cada nodo sobre el árbol final necesitamos una medida de la impureza global del árbol. Este valor, ponderado por la proporción de todos los casos del nodo t , es el valor del que se informa en el árbol como mejora.

III. DESARROLLO DEL TRABAJO

El Área de Retenciones de la universidad, es la que recopila los datos con la finalidad de realizar acciones para monitorear la deserción de los estudiantes. La universidad en estudio cuenta con tres periodos académicos a lo largo del año, esto es: en el año 2019 se tiene tres semestres 2019-1 (Ciclo de Verano), 2019-2 (Ciclo regular 1) y 2019-3 (Ciclo regular 2).

La población objetivo son alumnos de la modalidad de pregrado de la universidad que se han reinscrito en un periodo previo X-1 y que se encuentren matriculados en el periodo X, donde X es el periodo presente y X-1 el antecedente. En la Figura 5, se muestra la población objetivo que comprende alumnos que se matricularon en el 2019-2 del que se obtendrán datos antecedentes (periodo 2018-3) y se realiza la verificación para determinar si los alumnos se matricularon en el periodo siguiente (periodo 2019-3)

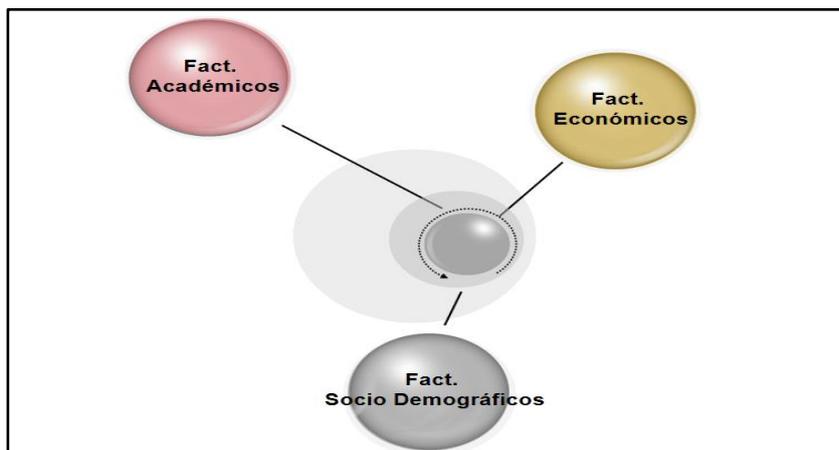
Figura 5
Población objetivo.



En la Figura 5, se muestra cómo está siendo considerada la población objetivo, el año x como un año base, X-1 como un año antecedente y el X+1 un año de evaluación de matrícula. Fuente de elaboración propia.

Los factores considerados que influyen en la deserción universitaria y a los cuales se aplicarán los modelos predictivos de la regresión logística binaria y el árbol de clasificación CART; estos están distribuidas en tres factores: factores académicos, factores económicos y factores socio demográficos. Gráfico de elaboración propia.

Figura 6
Factores relacionados a la deserción universitaria



En la Figura 6, se muestra los factores utilizados en la muestra de datos para realizar las estimaciones.

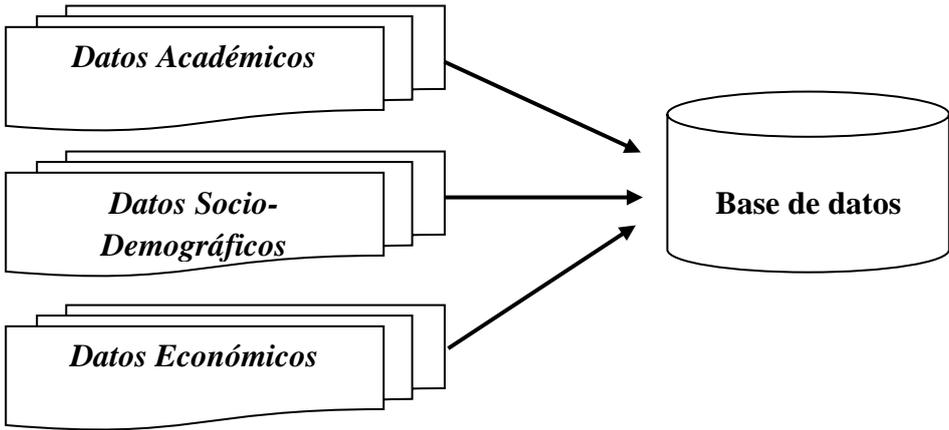
Fuente de elaboración propia.

3.1. La recolección de datos

Los datos de los alumnos de la universidad privada en estudio están almacenados en diferentes bases de datos y en diferentes servidores que utilizan los servicios de la nube (Cloud – Azure) disponibles que funcionan con el software SQL Server 2019. Para acceder a estos datos se generarán consultas en lenguaje SQL que permitan la extracción e integración de datos en una sola. Para el desarrollo de los modelos predictivos se utilizó la base de datos provenientes de los servidores de información de la universidad privada, donde se obtuvo acceso a tres fuentes de información principales para el estudio: socio-demográficos, económicos y académicos. La base de datos utilizada contiene del registro de alumnos matriculados en el periodo 2019-2, con información del periodo en 2019-2 presente y del 2018-3 como antecedente del alumno en el periodo X-1. Así mismo se pudo evaluar si los alumnos se matricularon en el periodo académico

siguiente (2019-3), variable que será descrita por ser objeto de estudio. Finalmente, las bases de datos fueron integradas, estructuradas y codificadas con el fin de contar con los datos en una sola matriz donde cada fila representa un registro, logrando obtener 32176.

Figura 7
Fuentes para la integración de la base de datos



En la Figura 7, se muestra la integración de las tres fuentes de datos para integrar la base de datos para la aplicación de los modelos predictivos propuestos.
Fuente de elaboración propia.

Se definió la estructura de la base de datos que permita manejar los diferentes tipos de variables que se almacenarán. Esta estructura de datos definida para cada uno de los tres tipos de información se presenta en la Tabla 3.

Tabla 3
Estructura de la base de datos de análisis

Fuente de datos	Variable	Posición	Descripción de la Variable	Tipo de Variable
Datos de Trazabilidad	ID_ALUM	1	Código identificador del alumno	Numerico
	PER	2	Periodo de matrícula del alumno	Cadena
Datos Socio demográficos	SEXO	3	Género del alumno	Cadena
	EDAD	4	Edad cronológica del alumno	Numerico
	PROC	5	Lugar de procedencia del alumno	Cadena
	CE	6	Tipo de centro educativo del alumno	Cadena
	MOD	7	Modalidad de ingreso a la institución	Cadena
Datos Académicos	CICLO	8	Ciclo académico del alumno	Numerico
	SEDE	9	Sede en la que esta matriculado	Cadena
	Y_MATR	10	Descripción de matrícula en el siguiente periodo	Cadena
	CUR_TRI	11	Cantidad de cursos en trica y cuatrica	Numerico
	TAS_APR_P	12	Tasa de aprobado en el periodo en curso	Numerico
	TAS_APR_A	13	Tasa de aprobado en el periodo anterior	Numerico
	TAS_ASI_P	14	Tasa de asistencia en el periodo en curso	Numerico
	TAS_ASI_A	15	Tasa de asistencia en el periodo anterior	Numerico
Datos Socio económicos	CUR_TT_P	16	Cantidad de cursos totales matriculados	Numerico
	CUR_TT_A	17	Cantidad de cursos totales matriculados en el periodo anterior	Numerico
Datos Socio económicos	TAS_NOM_P	16	Tasa de no morosidad en el periodo matriculado	Numerico
	TAS_NOM_A	19	Tasa de no morosidad en el periodo anterior	Numerico

En la Tabla 3, se muestra las variables que serán utilizadas para el análisis, divididas según la fuente de datos.

Fuente de elaboración propia.

Tabla 4**Estructura de la codificación de las variables**

Variable	Valores	Valores
ID_ALUM	Identificación alumno	
PER	2019-2	2019-2
Y_MATR	NO	1
	SI	0
SEXO	M:Masculino	1
	F:Femenino	0
EDAD	M:Masculino	1
PROC	Lima	1
	Provincia	0
CE	Nacional	1
	Privado	2
	Otros	3
MOD	Cepre	1
	Entrevista	2
	Ex. Extraordinario	3
	Otros	4
CICLO		[3-8]
SEDE	Arequipa	1
	Chiclayo	2
	Lima	3
CUR_TRI	NO	0
	SI	1
TAS_APR_P		[0-1]
TAS_APR_A		[0-1]
TAS_ASI_P		[0-1]
TAS_ASI_A		[0-1]
CUR_TT_P		[1-8]
CUR_TT_A		[1-8]
TAS_NOM_P		[0-1]
TAS_NOM_A		[0-1]

En la Tabla 4, se detalla la estructura de la base de datos codificada de la variable dependiente Y_MATR para explicar la deserción universitaria en función de 16 variables predictoras con sus respectivos valores. Fuente de elaboración propia.

A continuación, se describen las variables:

- **Y_MATR:** Variable dependiente. Se categoriza si el alumno no está matriculado (Y_MATR=0) o si se matriculó (Y_MATR=1).
- **CICLO:** Variable que contiene ciclo académico del estudiante, teniendo como rango del 3 al 8, ya que se exhiben los alumnos nuevos y por concluir, predominando la proporción del ciclo en el que tiene la mayor cantidad de cursos con promedio final.
- **SEDE:** Variable que contiene el local que eligió para llevar sus estudios, registrado en el proceso de matrícula.
- **CUR_TRI:** Variable que contiene el número de veces que el alumno ha tenido un curso con trica o cuatrica. (Trica es haber repetido tres veces un curso y cuatrica, cuatro veces un curso)
- **TAS_APR_P:** Variable calculada a partir del cociente entre la cantidad de cursos totales aprobados y el total de cursos matriculados en el periodo académico presente (Periodo X).
- **TAS_APR_A:** Variable calculada a partir del cociente entre la cantidad de cursos totales aprobados y el total de cursos matriculados en el periodo académico anterior (Periodo X-1).
- **TAS_ASI_P:** Variable calculada a partir del cociente entre la cantidad de veces totales que el alumno a asistido a clases y el total de veces que el alumno debe ir a clases en el periodo académico presente (Periodo X).
- **TAS_ASI_A:** Variable calculada a partir del cociente entre la cantidad de veces totales que el alumno a asistido a clases y el total de veces que el alumno debe ir a clases en el periodo académico anterior (Periodo X-1).
- **CUR_TT_P:** Variable que contiene la cantidad de cursos totales que el alumno se matriculo en el periodo presente (Periodo X).
- **CUR_TT_A:** Variable que contiene la cantidad de cursos totales que el alumno se matriculo en el periodo presente (Periodo X).

Respecto a las variables relacionadas al factor socio demográfico, se cuenta con cinco variables que se detallan a continuación:

- **SEXO:** Variable que contiene el género del alumno registrado en el proceso de matrícula.
- **EDAD:** Variable la edad cronológica del alumno a partir de la fecha de nacimiento declarada en el proceso de matrícula.
- **PROC:** Variable establece la procedencia geográfica del alumno, registrado en el proceso de matrícula.
- **CE:** Variable establece de qué tipo de centro educativo proviene el alumno, registrado en el proceso de matrícula.
- **MOD:** Variable establece la forma de ingreso a la institución, registrado en el proceso de matrícula.

Respecto a las variables relacionadas al factor económico, se cuenta con dos variables que se detallan a continuación:

- **TAS_NOM_P:** Variable calculada a partir del cociente entre la cantidad de veces totales que el alumno ha presentado puntualidad en sus pagos a y el total de cuotas asignadas al alumno en el periodo académico presente (Periodo X).
- **TAS_NOM_A:** Variable calculada a partir del cociente entre la cantidad de veces totales que el alumno ha presentado puntualidad en sus pagos a y el total de cuotas asignadas al alumno en el periodo académico anterior (Periodo X-1).

3.2. Técnicas estadísticas para el procesamiento de datos

La investigación propone el análisis de deserción académica bajo un enfoque de KDD, para lo cual se realizan las siguientes fases:

Fase 1. Pre procesamiento de datos. En busca de obtener una base de datos de calidad para aplicar los modelos predictivos se realiza un análisis exploratorio para identificar posibles datos faltantes y atípicos y manejar el problema del desbalanceo de datos.

- **Análisis exploratorio de los datos.** Se obtienen medidas estadísticas y gráficos para el análisis estadístico descriptivo de las variables cuantitativas y cualitativas, evaluando la presencia de datos atípicos y posibles datos perdidos.

- **Manejo de datos faltantes y atípicos.** El manejo de datos atípicos se realiza por su eliminación y el manejo de datos faltantes por imputación simple, si lo hubiera.
- **Balanceo de datos.** Se realiza el proceso de balanceo de datos aplicando el Sub muestreo (RUS), eliminando registros de la clase mayoritaria. Se extrae de clase mayoritaria en forma aleatoria de una muestra de aproximadamente igual tamaño de la clase minoritaria. El resultado será una base de datos balanceada.
- **Partición de la base de datos.** Se realiza la partición de los datos en dos grupos: Entrenamiento y Prueba. El conjunto de entrenamiento con el 70% aproximadamente de los datos que servirá para el ajuste del modelo predictivos y el 30% para definir el conjunto de prueba que se usa para evaluar la capacidad predictiva el modelo predictivo y obtener la tabla de confusión y sus respectivas métricas (exactitud, especificidad y especificidad).

Fase 2. Ajuste de los modelos predictivos. Se ajuste los datos de entrenamiento a una regresión logística binaria con función de enlace logit y a un árbol de clasificación CART usando el mejor parámetro de complejidad y con poda, considerando como variable dependiente (atributo objetivo) Y_MATR definiendo las categorías 0 (NO) y 1 (SI).

Fase 3. Evaluación de los modelos predictivos. Pata la regresión logística binaria, con la finalidad de identificar las variables más importantes, se aplica el método de selección de variables Backward con el criterio Akaike (AIC). Además, para ambos modelos usando los datos de prueba se obtiene la matriz de confusión y a partir de ella se calculan las medidas para evaluar la capacidad predictiva de la regresión logística y árbol de clasificación CART, tales como la exactitud, sensibilidad, especificidad, el AUC bajo la curva ROC y el coeficiente de concordancia de Kappa.

Fase 4. Comparación de los modelos predictivos. Con la finalidad de identificar el mejor modelo con la mayor capacidad predictiva para predecir la deserción universitaria se comparan las medidas de exactitud, sensibilidad, especificidad y AUC para evaluar su eficiencia predictiva.

IV. RESULTADOS Y DISCUSIÓN

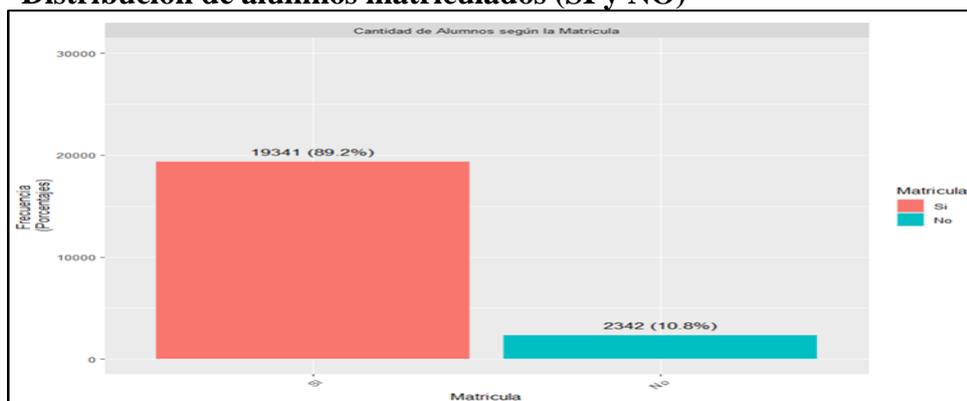
4.1. Pre procesamiento de datos.

El Pre procesamiento, es una etapa muy importante dentro del proceso de KDD. Esta etapa determina muchas veces que las sucesivas sean capaces de extraer conocimiento válido y útil a partir de los datos minados. Las bases de datos suelen contener datos con ruido, atípicos, faltantes y atributos irrelevantes y redundantes que afectan la precisión y distorsionan los resultados de las TMD. El pre procesamiento mejora la calidad de los datos minados; comprende un conjunto de procesos: la limpieza de datos, la transformación y la selección de atributos.

4.1.1 Análisis exploratorio de datos

A continuación, se realizar el análisis descriptivo de las variables cuantitativas y cualitativas según la variable dependiente (Y_MATR).

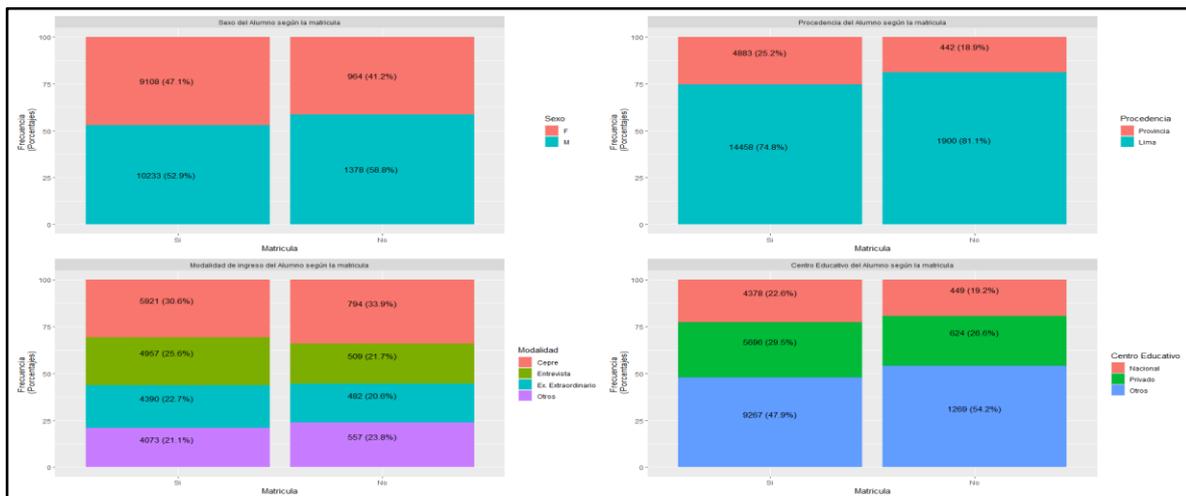
Figura 8
Distribución de alumnos matriculados (SI y NO)



En la Figura 8, se muestra la distribución de alumnos que se si y los no se matricularon. Fuente de elaboración propia.

En el análisis, se muestra la distribución de la matrícula de los alumnos. El 89.2% de los alumnos se matricularon el periodo siguiente (SI) y el 10.8% que representa los alumnos que no se matricularon el periodo siguiente (No).

Figura 9
Análisis de las variables cualitativas según la variable de matrícula

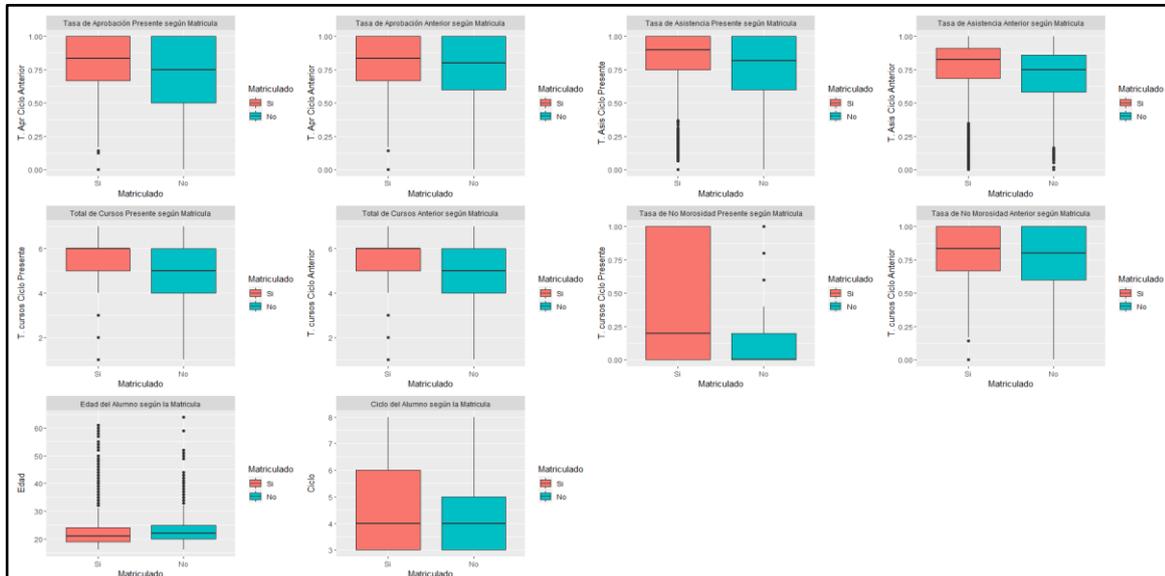


En la Figura 9, se muestra el análisis exploratorio cuantitativo que se le aplicó a las diferentes variables estudiadas.

Fuente de elaboración propia.

El análisis muestra la distribución de las variables cualitativas. En el SEXO, se evidencia ligeramente mayor porcentaje de alumnos que si (52,9%) y no (58,8%) se matricularon en comparación de las alumnas (47,9% y 41,2%). La PROCEDENCIA muestra que el mayor porcentaje de alumnos que si (74,8%) y no (81,1%) es Lima, mientras para los que provienen de provincias los que sí y no se matricularon son 25,2% y 18,9% respectivamente. En la modalidad de ingreso, el mayor porcentaje de alumnos que si (30,6%) y no (33,9%) se matricularon provienen del centro pre universitario de la universidad. Respecto al centro educativo CE que provienen los alumnos, los porcentajes para nacional, privado y otras instituciones que si se matricularon son 22,6%, 29,5% y 47,9% y para los que no son 19,2%, 26,6% y 54,2% respectivamente.

Figura 10
Análisis de las variables cuantitativas según la variable de matrícula



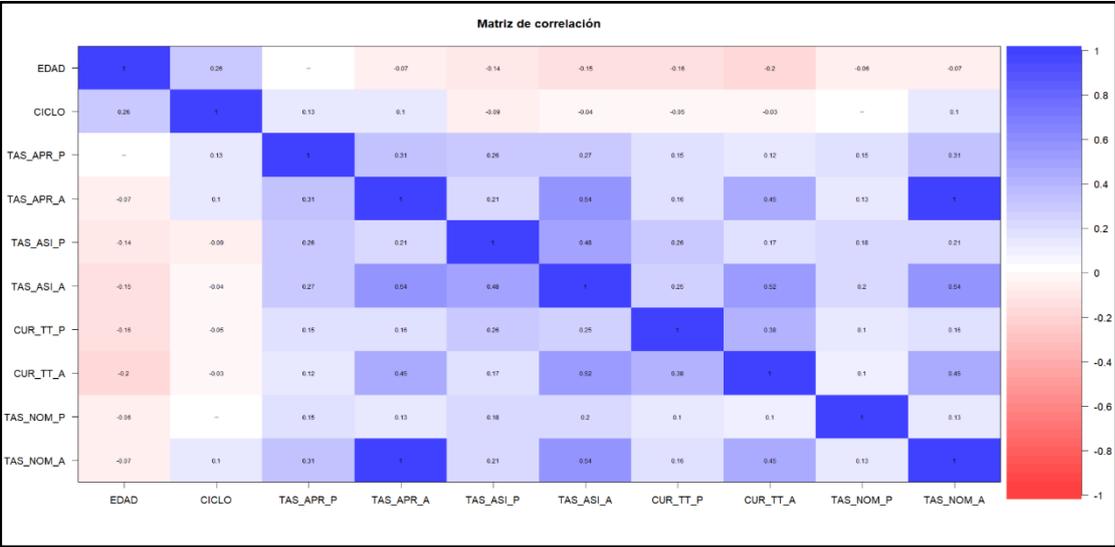
En la Figura 10, se muestra el análisis exploratorio cuantitativo que se le aplicó a las diferentes variables estudiadas.

Fuente de elaboración propia.

En el análisis se muestra la distribución de las variables cuantitativas. La variable TAS_APR_P y TAS_APR_A que representa la proporción de cursos aprobados del alumno con respecto al total matriculados en el periodo presente y anterior respectivamente, se evidencia una mayor variabilidad en el grupo no matriculados, mientras que el grupo de matriculados se puede observar un comportamiento más homogéneo, así mismo presenta algunos valores extremos inferiores en ambos casos sin embargo por la naturaleza de la variable son congruentes por tal motivo permanecerá dentro del análisis. La variable TAS_ASIS_P y TAS_ASIS_A que representa la proporción de veces que ha asistido un alumno al aula de clases respecto al total de veces que debería haber asistido en total en el periodo presente y anterior respectivamente, se evidencia una mayor variabilidad en el grupo no matriculados, mientras que el grupo de matriculados se puede observar un comportamiento más homogéneo, así mismo presenta algunos valores extremos inferiores en ambos casos sin embargo por la naturaleza de la variable son congruentes por tal motivo permanecerá dentro del análisis. La variable CUR_TT_P y CUR_TT_A que representa el total de cursos que el alumno se ha matriculado en el periodo

presente y anterior respectivamente, los que si se matriculan llevan entre 5 y 6 cursos y los que no se matriculan llevan entre 4 y 6 cursos, siendo un valor medio llevar 5 cursos. La variable TAS_NOM_P y TAS_NOM_A que representa la proporción de veces en las que el alumno no ha caído en morosidad respecto al total de las cuotas asignadas en el periodo presente y anterior respectivamente, evidenciándose que los alumnos si matriculados tenían un mejor comportamiento de pago explicado por la amplitud del rango intercuartílico versus los alumnos que no se matricularon. Respecto a la variable EDAD, que representa la edad cronológica del alumno, el promedio es de 25 años, en ambos grupos. Respecto a la variable CICLO, que representa el ciclo que cursan los alumnos, se tiene un promedio de 4 ciclos académicos y no se aprecian valores extremos, la mediana de ambos grupos es semejante sin embargo la variabilidad de ciclos es mayor en el grupo de Si matricula.

Figura 11
Matriz de correlaciones



En la Figura 11, se muestra la matriz de correlaciones aplicada a las variables cuantitativas para evaluar el grado de correlación y evitar la existencia de colinealidad.
Fuente de elaboración propia.

En el análisis de correlaciones realizado a las variables cuantitativas, al no existir valores superiores al 0.7 no se podría afirmar la presencia de co linealidad entre las variables en

estudio.

4.1.2 Manejo de datos faltantes y atípicos

Una vez integrados los datos e identificados los tipos de variables (numérica y de texto), finalmente se cuenta con una base de datos de 32176 registros con una variable dependiente que determina si el alumno se matriculo o no el siguiente ciclo (Y_MATR) el mismo que es de naturaleza dicotómica 18 variables independientes.

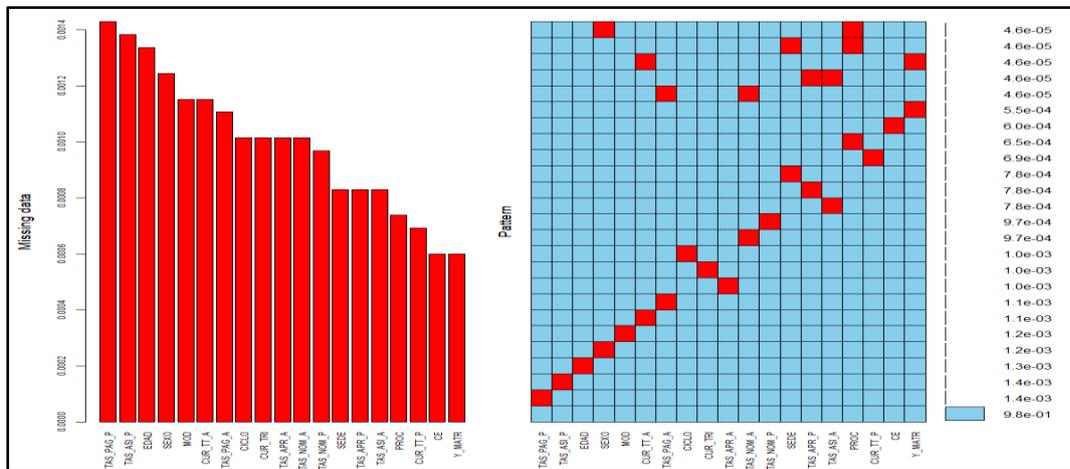
Tabla 5
Distribución de alumnos según estado de matrícula

Matricula	Número	Porcentaje (%)
SI	28849	89,7
NO	3327	10,3
Total	32176	100,0

En la Tabla 5, se muestra la distribución de los alumnos que si se matricularon y no. Se observa que el 89,7% se matricularon y un 10,3% que no lo hicieron.
Fuente de elaboración propia.

En la primera fase del análisis exploratorio se analizan los valores perdidos de la matriz construida identificándose que la proporción respecto al total de registros de la base de datos es de menos del 2%, siendo esta una pequeña proporción para el estudio son considerados errores propios del sistema, mostrándose en la Figura 10.

Figura 12
Distribución de valores perdidos



En la

Figura 12, se muestra la presencia de datos perdidos (Vacíos) en la base de datos analizada y los representa con color rojo.

Fuente de elaboración propia.

Tabla 6
Distribución de alumnos según estado de matrícula depurado

Matricula	Número	Porcentaje (%)
SI	19341	89,2
NO	2342	10,8
Total	21277	100,0

En la Tabla 6, se muestra la distribución de los alumnos que si se matricularon y no. Post realización del proceso de depuración.

Fuente de elaboración propia.

Entonces se puede apreciar que la base de datos esta desbalanceada, por lo tanto, se procederá aplicar la técnica de sub muestreo para obtener una base de datos desbalanceada.

4.1.3 Balanceo de la base de datos

Para el desarrollo de los modelos se tuvo en cuenta el problema del desbalanceo de datos con respecto a la variable dependiente (atributo clase), por lo cual si no se toma en cuenta afectaría la capacidad predictiva de los modelos ajustados sesgando a la clase mayoritaria. La base de datos presenta un desbalance en la variable respuesta (Y_MATR), existe una mayor proporción de alumnos que “Si” se matriculan (89,2%) comparado al grupo “No” se matriculan (10,8%). Una de las alternativas para mejorar la eficiencia del modelo es equilibrar la cantidad de datos mediante el método de Sub muestreo que se basa en una eliminación de datos de la clase mayoritaria. El método consiste en extraer una muestra aleatoria de la clase mayoritaria (SI) del mismo tamaño de la clase minoritaria (NO) para conseguir el balanceo de las dos clases y por ende de la base de datos. En la Tabla 7, se presenta la distribución de la base de datos balanceada con 4684 datos de los estudiantes y resultando con un 50,0% el porcentaje tanto para la clase “SI” y “NO”.

Tabla 7
Distribución de alumnos según estado de matrícula balanceada

Matricula	Número	Porcentaje (%)
SI	2342	50,0
NO	2342	50,0
Total	4684	100,0

En la Tabla 7, se muestra la distribución de los alumnos que si se matricularon y no. Post realización del proceso de balanceo de datos.
Fuente de elaboración propia.

Con la finalidad de ajustar y evaluar los modelos predictivos propuestos, se divide la base de datos en dos: Entrenamiento y Prueba. En forma aleatoria de la base de datos, se seleccionar el conjunto de entrenamiento para ajustar los modelos predictivos propuestos y el conjunto de prueba servirá para la validación de dichos modelos. En la Tabla 8, se presenta la distribución de la base de datos con una muestra aleatoria de un 70% aproximadamente para el conjunto de entrenamiento y un 30% para el conjunto de prueba considerando para los SI y NO se

matricularon. Se observa que para el conjunto de entrenamiento se seleccionó el 64,7% y para el conjunto de prueba el 35,3% de alumnos. Además, se puede indicar que ambos conjuntos están balanceados con 50% para la clase Si y 50% para la NO.

Tabla 8
Distribución de alumnos para el conjunto de entrenamiento y prueba

Partición	Matricula	Número	Porcentaje (%)
Entrenamiento 64,7	SI	1515	50,0
	NO	1516	50,0
	Total	3031	100,0
Prueba 35,3	SI	827	50,0
	NO	826	50,0
	Total	1653	100,0

En la Tabla 8, se muestra la distribución de los alumnos que si se matricularon y no. Post realización del proceso de partición de datos, donde un grupo es 64.7% y el otro 35.3%. Fuente de elaboración propia.

4.2 Técnicas de minería de datos

Se aplican las TMD de la regresión logística binaria y el árbol de clasificación CART para obtener el modelo para predecir la deserción de los estudiantes.

4.2.1 Regresión logística binaria

Se ajusta los datos del conjunto de entrenamiento para obtener el modelo predicho con la regresión logística binaria con función de enlace Logit.

Tabla 9
Estimación de los coeficientes de la regresión logística binaria

VARIABLES	Estimate	Std. Error	z	value	Pr(> z)
(Intercept)	2.501712	0.369157	6.777	1.23e-11	***
SEX01	0.157469	0.082142	1.917	0.05523	.
EDAD	0.025639	0.009617	2.666	0.00768	**
CICLO	-0.078210	0.027809	-2.812	0.00492	**
SEDE2	0.208921	0.201961	1.034	0.30092	
SEDE3	0.750578	0.122877	6.108	1.01e-09	***
CUR_TRI	0.783039	0.131747	5.944	2.79e-09	***
TAS_APR_P	-1.241916	0.193396	-6.422	1.35e-10	***
TAS_ASI_P	-1.174445	0.216580	-5.423	5.87e-08	***
TAS_ASI_A	0.670338	0.276624	2.423	0.01538	*
CUR_TT_P	-0.247089	0.034500	-7.162	7.95e-13	***
CUR_TT_A	-0.084619	0.032813	-2.579	0.00991	**
TAS_NOM_P	-0.510178	0.183328	-2.783	0.00539	**
TAS_NOM_A	-1.128038	0.135203	-8.343	< 2e-16	***

En la Tabla 9, se muestra el cuadro de coeficientes de regresión estimados y su significación con la prueba de Wald para las variables predictoras.

Fuente de elaboración propia.

En el análisis, se muestra los coeficientes de regresión estimados y su significación con el método de selección de variable Backward y usando la estadística AIC. Se identifican las variables significativas la Edad, Ciclo, CUR_TRI, TAS_APR_P, TAS_ASI_P, TAS_ASI_a, CUR_TT_P, CUR_TT_A, TAS_NOM_P y TAS_NOM_A.

A continuación, se calcula el Pseudo R^2 de McFalden, que resulta una aproximación basada en una comparación de la verosimilitud del modelo solo con la constante, con la verosimilitud del modelo con todos los parámetros. Se obtuvo un R^2 de 14,6%, indica que la regresión logística binaria aumenta en aproximadamente 14,6% el logaritmo de la función de verosimilitud en comparación al modelo estimado con el modelo nulo.

Matriz de confusión

Con la finalidad de evaluar la bondad de ajuste de la regresión logística se obtiene con el conjunto de prueba la matriz de confusión. En la Tabla 10, se presenta la distribución de la clasificación predicha (fila) con la regresión logística y la observada (columna) que se tiene en la base de datos para las dos clases, alumnos matriculados (SI) y no matriculados (NO). Se observa que el conjunto de prueba tiene un total de 1653 alumnos, asignando 827 a los matriculados y 826 a los no matriculados. Para el caso de los 827 alumnos que, si se matricularon, la regresión logística los clasificó correctamente a 537 y 290 los predijo incorrectamente como no matriculados. Así mismo, de los 826 alumnos que no se matricularon, la regresión logística los clasificó correctamente a 562 y 264 los predijo incorrectamente como matriculados.

Tabla 10
Matriz de confusión de la regresión logística binaria

Predicho	Observado		Total
	SI	NO	
SI	537	264	801
NO	290	562	852
Total	827	826	1653

En la Tabla 10, se muestra el resultado de la predicción del modelo Logit representado en la matriz de confusión, donde es comparado lo predicho con lo observado (Real).
Fuente de elaboración propia.

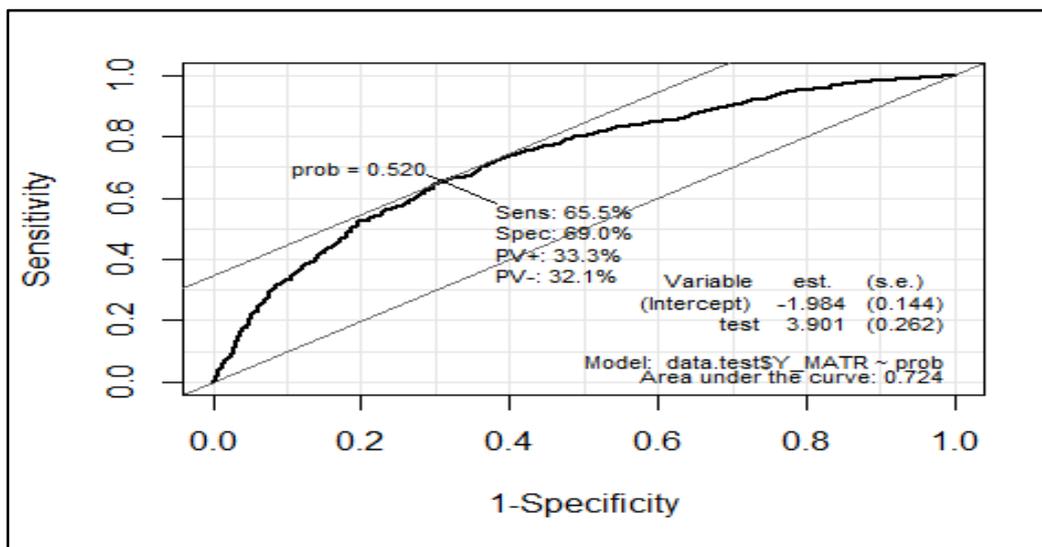
En la Tabla 11, se presenta los tres indicadores para evaluar la eficiencia predictiva para la regresión logística binaria con respecto a la buena clasificación de los alumnos matriculados y no matriculados, calculados a partir de la matriz de confusión. La exactitud indica que la regresión logística binaria predice el 66,4% de los alumnos correctamente tanto a los alumnos que se matricularon y no matricularon. Mientras que la Sensibilidad indica que la regresión logística binaria predice el 71,2% correctamente a los alumnos que si se matricularon y la Especificidad indica que la regresión logística binaria predice el 65,8% correctamente a los alumnos que se no matricularon.

Tabla 11
Medidas de la eficiencia predictiva para la regresión logística

Medida	Valor
Exactitud	66,4%
Sensibilidad	71,2%
Especificidad	65,8%

En la Tabla 11, se muestra el resultado de los indicadores de eficiencia predictiva que se obtuvo posterior a la realización de la predicción aplicando el modelo Logit.
Fuente de elaboración propia.

Figura 13
Curva ROC para la comparación de la sensibilidad y especificidad.



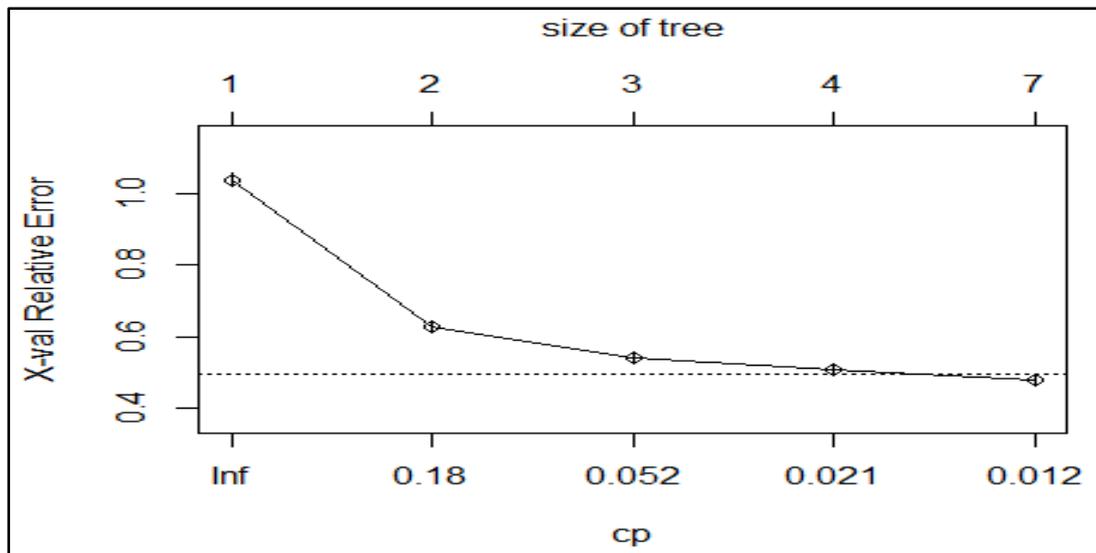
En la Figura 13, se muestra la curva ROC con los valores de la sensibilidad y la especificidad para un punto de corte de 0.52.
Fuente de elaboración propia.

El valor de la sensibilidad, indica que la regresión logística binaria está prediciendo correctamente al 65,5% de los alumnos que, si se matricularon y la especificidad, indica que el clasificador está prediciendo correctamente al 69,0% de los alumnos que no se matricularon. También un valor bajo la curva ROC de 0,724.

4.2.2 Árbol de clasificación CART

Se generó el árbol de clasificación con el algoritmo CART, utilizando las variables que fueron seleccionadas y aplicando la misma base de datos de entrenamiento balanceada, identificando el parámetro de complejidad (CP) óptimo y con la opción de una poda. Para la aplicación se ha utilizado la función `rpart` que proporciona el software R, obteniéndose los siguientes resultados:

Figura 14:
Determinación del parámetro de complejidad óptimo.



En la Figura 14, se muestra la distribución de errores relativos para varios valores de parámetros de complejidad. Se observa que el valor óptimo para el parámetro de complejidad que permita ajustar el árbol de clasificación con el menor error relativo es el valor de 0,021. Este valor se usa para ajustar el árbol de clasificación CART.

Fuente de elaboración propia.

Tabla 12
Distribución de reglas por nodo del árbol

Rama	Nodos de decisión	SI (1)	NO (0)	Clase estimada	P(Si)	P(No)
1)	root	3031	1515	1	0.4998350	0.5001650
2)	TAS_NOM_A>=0.1	1214	326	0	0.7314662	0.2685338
4)	TAS_NOM_P< 0.7	1079	191	0	0.8229842	0.1770158
8)	TAS_NOM_P>=0.5	619	22	0 (*)	0.9644588	0.0355412
9)	TAS_NOM_P< 0.5	460	169	0	0.6326087	0.3673913
18)	TAS_NOM_A< 0.5	166	0	0 (*)	1.0000000	0.0000000
19)	TAS_NOM_A>=0.5	294	125	1	0.4251701	0.5748299
38)	SEDE=1	42	9	0 (*)	0.7857143	0.2142857
39)	SEDE=2	252	92	1 (*)	0.3650794	0.6349206
5)	TAS_NOM_P>=0.7	135	0	1 (*)	0.0000000	1.0000000
3)	TAS_NOM_A< 0.1	1817	627	1	0.3450743	0.6549257
6)	TAS_ASI_A< 0.347	46	0	0 (*)	1.0000000	0.0000000
7)	TAS_ASI_A>=0.347	1771	581	1 (*)	0.3280632	0.6719368

En la Tabla 12, se presenta el resultado del árbol de clasificación CART. Se muestra las reglas obtenidas y las probabilidades calculadas para las clases SI y NO, con el resultado del nodo terminal para alguna de estas dos clases. Se observa que existen tres reglas para la clasificación de la clase SI y cuatro para la clase NO. Fuente de elaboración propia.

R1: TAS_NOM_A>=0.1, TAS_NOM_P< 0.7, TAS_NOM_P< 0.5 460, TAS_NOM_A>=0.5, SEDE=2

R2: TAS_NOM_A>=0.1, TAS_NOM_P>=0.7

R3: TAS_NOM_A<0.1, TAS_ASI_A>=0.3468725

Las reglas para la clase que NO se matricularon son:

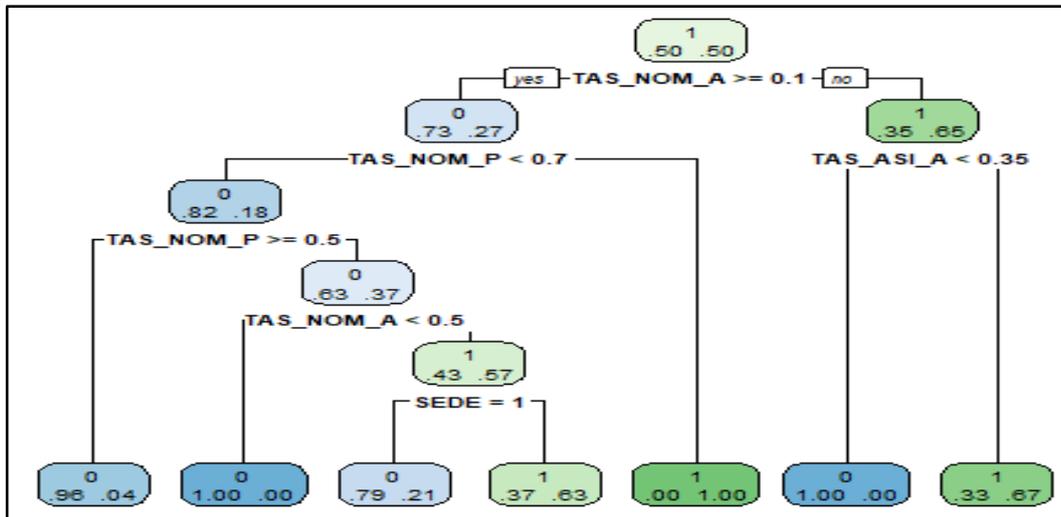
R1: TAS_NOM_A>=0.1, TAS_NOM_P< 0.7, TAS_NOM_P>= 0.5 460

R2: TAS_NOM_A>=0.1, TAS_NOM_P< 0.7, TAS_NOM_P< 0.5 460, TAS_NOM_A<0.5

R3: TAS_NOM_A>=0.1, TAS_NOM_P< 0.7, TAS_NOM_P< 0.5 460, TAS_NOM_A>=0.5, SEDE=1

R4: $TAS_NOM_A < 0.1$, $TAS_ASI_A < 0.3468725$

Figura 15
Árbol de clasificación CART



En la Figura 15, se muestra el árbol de clasificación CART. El nodo raíz es la variable TAS_NOM, con cinco nodos intermedios (TAS_NOM_P, TAS_ASI_A, TAS_NOM_A, SEDE) y con tres nodos finales asociados a la clase SI se matricularon y cuatro a la clase No se matricularon.

Fuente de datos, elaboración propia.

Interpretación de los nodos terminales

Primer Nodo Terminal

- La variable TAS_NOM_P es considerado el mejor predictor para la variable dependiente matricula del alumno.
- El 30% de los estudiantes que presentan un valor en la variable TAS_NOM_P mayor o igual a 0.5 se matricula en el periodo siguiente.

Segundo Nodo Terminal

- El 10% de los estudiantes que se matriculan el siguiente periodo pertenecen a cualquiera de las sedes 1,2 o 6, están matriculados en 6 o más cursos y tienen TAS_NOM_P mayor o igual a 0.5.

Tercer Nodo Terminal

- El 5% de los estudiantes que se matriculan el siguiente periodo tiene una TAS_ASI_A mayor o igual a 0.91, no pertenecen a las sedes 1,2 o 6, tienen 6 o más cursos matriculados y no tienen TAS_NOM_P mayor o igual a 0.5.

Cuarto Nodo Terminal

- El 15% de los estudiantes que no se matriculan el siguiente periodo no tienen una TAS_ASI_A mayor o igual a 0.91, no pertenecen a las sedes 1,2 o 6, tienen 6 o más cursos matriculados y no tienen TAS_NOM_P mayor o igual a 0.5.

Quinto Nodo Terminal

- El 40% de los estudiantes que no se matriculan el siguiente periodo tienen menos de 6 cursos matriculados y tienen una TAS_NOM_P menor a 0.5.

Matriz de confusión

Con la finalidad de evaluar la bondad de ajuste del árbol de clasificación CART se obtiene con el conjunto de prueba la matriz de confusión. En la Tabla 13, se presenta la distribución de la clasificación predicha (fila) con el árbol de clasificación y la observada (columna) que contiene la base de datos para las dos clases, alumnos matriculados (SI) y no matriculados (NO). Se observa que el conjunto de prueba tiene un total de 1653 alumnos, asignando 827 a los matriculados y 826 a los no matriculados. Para el caso de los 827 alumnos que, si se matricularon, árbol de clasificación los clasificó correctamente a 416 y 411 los predijo incorrectamente como no matriculados. Así mismo, de los 826 alumnos que no se matricularon, árbol de clasificación los clasificó correctamente a 804 y 22 los predijo incorrectamente como matriculados.

Tabla 13
Matriz de confusión del árbol de clasificación CART

Predicho	Observado		Total
	SI	NO	
SI	416	22	438
NO	411	804	1215
Total	827	826	1653

En la Tabla 13, se muestra el resultado de la predicción del modelo CART representado en la matriz de confusión, donde es comparado lo predicho con lo observado (Real).

Fuente de elaboración propia.

En la Tabla 14, se presenta los tres indicadores para evaluar la eficiencia predictiva para el árbol de clasificación CART con respecto a la buena clasificación de los alumnos matriculados y no matriculados, calculados a partir de la matriz de confusión. La exactitud indica que el árbol de clasificación predice el 73,8% de los alumnos correctamente tanto a los alumnos que se matricularon y no matricularon. Mientras que la Sensibilidad indica que el árbol de clasificación predice el 97,3% correctamente a los alumnos que si se matricularon y la Especificidad indica que el árbol de clasificación predice el 50,3% correctamente a los alumnos que se no matricularon.

Tabla 14
Medidas de la eficiencia predictiva del árbol de clasificación CART

Medida	Valor
Exactitud	73.8%
Sensibilidad	97,3%
Especificidad	50,3%

En la Tabla 14, se muestra el resultado de los indicadores de eficiencia predictiva que se obtuvo posterior a la realización de la predicción aplicando el modelo CART.

Fuente de elaboración propia.

4.3 Comparación de la capacidad predictiva de la regresión logística binaria y del árbol de clasificación CART

En la Tabla 16, se presenta los resultados de las medidas para evaluar la capacidad predictiva de los modelos propuestos de la regresión logística binaria y el árbol de clasificación CART. Se puede evidenciar que el árbol de clasificación CART tiene las mayores medidas de la exactitud y sensibilidad, esto indica que el árbol tiene la mayor capacidad predictiva para predecir a los estudiantes que se matricularon y que no se matricularon (73,8%) en comparación de la regresión logística (66,4%), y el árbol predice correctamente a los alumnos que si se matricularon (97,3%) en comparación de la regresión logística (71,2%). Con respecto a la especificidad, la regresión logística predice un mayor porcentaje de alumnos que no se matricularon (65,8%) que el árbol de clasificación CART (50,3%).

Tabla 15
Comparación de la eficiencia predictiva de los modelos predictivos

Medida	Regresión logística binaria	Árbol de clasificación CART
Exactitud	66,4%	73,8%
Sensibilidad	71,2%	97,3%
Especificidad	65,8%	50,3%
AUC	72,4%	73,8%

En la Tabla 15, se muestra la comparación de los indicadores de eficiencia predictiva de los modelos LOGT y CART.

Fuente de elaboración propia.

V. CONCLUSIONES

1. La regresión logística binaria aplicando el método de selección de Backward, identificó como variables significativas para predecir la deserción universitaria: Edad, Ciclo, CUR_TRI, TAS_APR_P, TAS_ASI_P, TAS_ASI_A, CUR_TT_P, CUR_TT_A, TAS_NOM_P y TAS_NOM_A. El árbol de clasificación CART con un parámetro de complejidad de 0.021 y con poda, identificó a las variables TAS_NOM_A, TAS_NOM_P, SEDE, TAS_ASI_A como las más importantes para predecir la deserción universitaria.
2. Las TMD puede afectar su eficiencia predictiva, cuando existe un desbalanceo de la base de datos de la variable dependiente. Con una base de datos desbalance, el 89,2% de registros para la clase Si y solo un 10,8% para la NO, se mejoró la capacidad predictiva al aplicar un submuestreo para el balanceo.
3. El árbol de clasificación CART resultó con mayores valores para la exactitud, sensibilidad, especificidad y AUC; siendo 73,8%, 97,3%, 50,3% y 73,8% respectivamente, mientras la regresión logística los valores fueron del 66,4%, 71,2%, 65,8% y 72,4% respectivamente; por lo tanto, el árbol de clasificación CART demostró el de mejor capacidad predictiva para predecir la deserción universitaria.
4. El árbol de clasificación CART, resultante tuvo un tamaño de 13 nodos (un nodo raíz, 5 nodos intermedios y 7 nodos terminales). De los 7 nodos terminales, 3 nodos están asociados a la clase SI (si se matricularon) y 4 a la clase NO (no se matricularon). Además, demostró una exactitud del 73,8%, indicando tener una capacidad de predecir correctamente a los alumnos que si y que no se matriculan.
5. El árbol de clasificación CART, permite obtener 3 reglas de decisión para predecir a los alumnos que si se matriculan y 4 reglas para predecir a los alumnos que no se matriculan, siendo las siguientes

6.

R1: $TAS_NOM_A \geq 0.1$, $TAS_NOM_P < 0.7$, $TAS_NOM_P \geq 0.5$ 460

R2: $TAS_NOM_A \geq 0.1$, $TAS_NOM_P < 0.7$, $TAS_NOM_P < 0.5$ 460,
 $TAS_NOM_A < 0.5$

R3: $TAS_NOM_A \geq 0.1$, $TAS_NOM_P < 0.7$, $TAS_NOM_P < 0.5$ 460,
 $TAS_NOM_A \geq 0.5$, SEDE=1

R4: $TAS_NOM_A < 0.1$, $TAS_ASI_A < 0.3468725$

7. El Área de retenciones de la universidad, decidió implementar el modelo del árbol de clasificación CART con las reglas de decisión obtenidas como una herramienta cuantitativa que permita mejorar y apoyar acciones preventivas y de monitoreo para enfrentar el problema de la deserción universitaria.

VI. RECOMENDACIONES

1. Aplicar métodos o técnicas de selección de atributos (variables) como los filtrados o Wrapper, que permiten seleccionar la gran variedad de dimensiones de factores o variables que influyen en la deserción universitaria, antes o simultáneamente de aplicar las TMD.
2. Aplicar otras TMD tales como las redes neuronales, redes bayesianas o máquinas de soporte vectorial; a fin de identificar el mejor modelo con la mayor capacidad predictiva.
3. El Área de Retenciones, debe de implementar acciones o estrategias automatizadas de prevención y monitoreo que estén relacionados con el modelo predictivo, con la finalidad que permita identificar a los alumnos con una probable deserción universitaria.
4. El área de retenciones debería cada cierto tiempo recalibrar el modelo, revisando nuevamente los datos con el fin de disminuir o agregando nuevas.
5. El área de retenciones deberá procesar el modelo cada semana, con el fin de calcular la lista de alumnos con alta probabilidad de deserción, con el fin de aplicar acciones de retención.
6. Se recomienda que la Universidad implemente un datawarehouse, con el fin de centralizar y garantizar la calidad de datos para mejorar la predicción.

VII. REFERENCIAS BIBLIOGRÁFICAS

- Díaz Peralta, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estudios pedagógicos (Valdivia)*, 34(2), 65-86. <http://dx.doi.org/10.4067/S0718-07052008000200004>
- Ferreira, M. M., Avitabile, C., Botero Álvarez, J., Haimovich Paz, F., & Urzúa, S. (2017). *Momento decisivo: la educación superior en América Latina y el Caribe*.
- Yukselturk, E., Ozekes, S., & Turel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-learning*, 17(1), 118-133.
- Eckert, KB; Suenaga, R. (2014). Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. *Formación universitaria* 8:03-12. DOI: <https://doi.org/10.4067/S0718-50062015000500002>.
- Tudela, H. E. V. (2014). Una aproximación teórica a la deserción estudiantil universitaria. *Revista digital de Investigación en Docencia universitaria*, 59-76.
- Tan, M., & Shao, P. (2015). Prediction of student dropout in e-Learning program through the use of machine learning method. *International journal of emerging technologies in learning*, 10(1).
- Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, 9(4), 1-5.
- Fernández-Martín, T., Solís-Salazar, M., Hernández-Jiménez, M. T., & Moreira-Mora, T. E. (2019). Un análisis multinomial y predictivo de los factores asociados a la deserción universitaria. *Revista Electrónica Educare*, 23(1), 73-97.
- Manhães, L. M. B., da Cruz, S. M. S., & Zimbrão, G. (2014, March). WAVE: an architecture for predicting dropout in undergraduate courses using EDM. *In Proceedings of the 29th annual acm symposium on applied computing* (pp. 243-247).
- Klenzi, R. O., & López, M. (2017, August). Detección de ataques DoS con herramientas de minería de datos. In XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires).

Cortes, C; Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20(3):273-297. DOI: <https://doi.org/10.1007/BF00994018>.

Osorio, H., & Keider, J. (2019). *Metodología de clasificación de datos desbalanceados basado en métodos de submuestreo*.

Agresti, A., & Tarantola, C. (2018). Simple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica*, 72(3), 210-223.

Mount, J; Karsy, M; Huang, T; Kleinman, G; Karpel-Massler, G. 2014. Practical Data Science with R. *s.l., s.e., vol.19*. 1065-1087 p. DOI: <https://doi.org/10.5326/50.5.toc>.

ANEXOS

Códigos en R para la Regresión Logística

- **Librerías de R**

```
library(psych)
library(MASS)
library(klaR)
library(gains)
library(caret)
library(Boruta)
library(gmodels)
library(mice)
library(VIM)
library(caTools)
library(MASS)
```

- **Lectura de datos**

```
library(readxl)
Data_Desercion<- read_excel("Data.xlsx")
Data<-Data_Desercion[,c(1:17)]
Data_cuanti<-Data[,c(3,7,9,10,11,12,13,14,15,16,17)]
attach(Data)
Data$Y_MATR<-as.factor(Data$Y_MATR)
Data$SEXO<-as.factor(Data$SEXO)
Data$PROC<-as.factor(Data$PROC)
Data$MOD<-as.factor(Data$MOD)
Data$SEDE<-as.factor(Data$SEDE)
Data$CE<-as.factor(Data$CE)
str(Data)
```

```

contrasts(Data$Y_MATR)

##    1
## 0 0
## 1 1

Data$Y_MATR=relevel(Data$Y_MATR,ref="0")
contrasts(Data$Y_MATR)

##    1
## 0 0
## 1 1

```

- **Gráficos Descriptivos**

```

library(dplyr)

Tabla <- Data %>% group_by(Y_MATR) %>% summarise(Total=n()) %>%
  dplyr::mutate(Porcentaje = round(Total/sum(Total)*100, 1))

M_1<-ggplot(Tabla, aes(x = Y_MATR, y=Total,fill=Y_MATR) ) +
  geom_bar(width = 0.9, stat="identity", position = position_dodge())+

  ylim(c(0,30000))+
  labs(x="Matricula", y= "Frecuencia \n (Porcentajes)") + #17
  labs(fill = "")+
  geom_text(aes(label=paste0(Total," ", "", "(", Porcentaje, "%",")")), #18
            vjust=-0.9,
            color="black",
            hjust=0.5,
            position = position_dodge(0.9),
            angle=0,
            size=4.0

  ) +
  scale_fill_discrete(name = "Matricula", labels = c("Si", "No")) + #19

```

```

theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1)) + #20
#theme_bw(base_size = 14) +
scale_x_discrete(labels=c("Si", "No")) +
facet_wrap("Cantidad de Alumnos según la Matricula")

```

M_1

```

par(mfrow=c(3,3))
library(dplyr)
library(ggplot2)
library(patchwork)
Tabla_SEXO <- Data %>% group_by(Y_MATR,SEXO) %>% summarise(Total=n()) %>%
  dplyr::mutate(Porcentaje = round(Total/sum(Total)*100, 1))

A_1<-ggplot(data=Tabla_SEXO, aes(x=Y_MATR, y=Porcentaje, fill=SEXO)) +
  geom_bar(width = 0.9, stat="identity")+
  ylim(c(0,100))+
  labs(x="Matricula", y= "Frecuencia \n (Porcentajes)") + #17
  labs(fill = "")+
  scale_x_discrete(labels=c("Si", "No")) +
  geom_text(aes(label=paste0(Total, " ", "", "(", Porcentaje, "%", ")")), #18
            vjust=0.5,
            color="black",
            hjust=0.7,
            position = position_stack(vjust =0.5),
            angle=0,
            size=4.0
  ) +

```

```

scale_fill_discrete(name = "Sexo", labels = c("F", "M")) +
facet_wrap(~"Sexo del Alumno según la matricula")
Tabla_PROC <- Data %>% group_by(Y_MATR,PROC) %>% summarise(Total=n()) %>%
  dplyr::mutate(Porcentaje = round(Total/sum(Total)*100, 1))
A_2<-ggplot(data=Tabla_PROC, aes(x=Y_MATR, y=Porcentaje, fill=PROC)) +
  geom_bar(width = 0.9, stat="identity")+
  ylim(c(0,100))+
  labs(x="Matricula", y= "Frecuencia \n (Porcentajes)") + #17
  labs(fill = "")+
  scale_x_discrete(labels=c("Si","No")) +
  geom_text(aes(label=paste0(Total," ", "", "(", Porcentaje, "%",")"), #18
    vjust=-0.9,
    color="black",
    hjust=0.7,
    position = position_stack(vjust =0.5),
    angle=0,
    size=4.0
  ) +
  scale_fill_discrete(name = "Procedencia", labels = c( "Provincia","Lima")) +
  #19
  facet_wrap(~"Procedencia del Alumno según la matricula")
Tabla_MOD <- Data %>% group_by(Y_MATR,MOD) %>% summarise(Total=n()) %>%
  dplyr::mutate(Porcentaje = round(Total/sum(Total)*100, 1))
A_3<-ggplot(data=Tabla_MOD, aes(x=Y_MATR, y=Porcentaje, fill=MOD)) +
  geom_bar(width = 0.9, stat="identity")+
  ylim(c(0,100))+
  labs(x="Matricula", y= "Frecuencia \n (Porcentajes)") + #17
  labs(fill = "")+
  scale_x_discrete(labels=c("Si","No")) +

```

```

geom_text(aes(label=paste0(Total," ", "", "(" , Porcentaje, "%",")")), #18
  vjust=-0.9,
  color="black",
  hjust=0.7,
  position = position_stack(vjust =0.5),
  angle=0,
  size=4.0
) +
  scale_fill_discrete(name = "Modalidad", labels = c("Cepre","Entrevista","Ex.
Extraordinario","Otros")) + #19
  facet_wrap(~"Modalidad de ingreso del Alumno según la matricula")

Tabla_CE <- Data %>% group_by(Y_MATR,CE) %>% summarise(Total=n()) %>%
  dplyr::mutate(Porcentaje = round(Total/sum(Total)*100, 1))
A_4<-ggplot(data=Tabla_CE, aes(x=Y_MATR, y=Porcentaje, fill=CE)) +
  geom_bar(width = 0.9, stat="identity")+
  ylim(c(0,100))+
  labs(x="Matricula", y= "Frecuencia \n (Porcentajes)") + #17
  labs(fill = "")+
  scale_x_discrete(labels=c("Si","No")) +
  geom_text(aes(label=paste0(Total," ", "", "(" , Porcentaje, "%",")")), #18
par(mfrow=c(3,3))
library(dplyr)
library(ggplot2)
library(patchwork)
G_1<-ggplot(Data, aes(x=Y_MATR, y=TAS_APR_P, fill=Y_MATR)) +
  geom_boxplot() +
  labs(x="Matriculado", y="T. Apr Ciclo Anterior", fill="Matriculado") + # ti
tulo ejes y leyenda

```

```

scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
scale_fill_discrete(labels=c("Si","No"))+ # etiquetas claves leyenda
facet_wrap(~"Tasa de Aprobación Presente según Matricula")
G_2<-ggplot(Data, aes(x=Y_MATR, y=TAS_APR_A, fill=Y_MATR)) +
geom_boxplot() +
labs(x="Matriculado", y="T. Apr Ciclo Anterior", fill="Matriculado") + # ti
tulo ejes y leyenda
scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
scale_fill_discrete(labels=c("Si","No"))+ # etiquetas claves leyenda
#ggtitle("Tasa de Aprobación Anterior según Matricula")
facet_wrap(~"Tasa de Aprobación Anterior según Matricula")
G_3<-ggplot(Data, aes(x=Y_MATR, y=TAS_ASI_P, fill=Y_MATR)) +
geom_boxplot() +
labs(x="Matriculado", y="T. Asis Ciclo Presente", fill="Matriculado") + # t
itulo ejes y leyenda
scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
scale_fill_discrete(labels=c("Si","No"))+# etiquetas claves leyenda
#ggtitle("Tasa de Asistencia Presente según Matricula")
facet_wrap(~"Tasa de Asistencia Presente según Matricula")
G_4<-ggplot(Data, aes(x=Y_MATR, y=TAS_ASI_A, fill=Y_MATR)) +
geom_boxplot() +
labs(x="Matriculado", y="T. Asis Ciclo Anterior", fill="Matriculado") + # t
itulo ejes y leyenda
scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
scale_fill_discrete(labels=c("Si","No"))+ # etiquetas claves leyenda
#ggtitle("Tasa de Asistencia Anterior según Matricula")
facet_wrap(~"Tasa de Asistencia Anterior según Matricula")
G_5<-ggplot(Data, aes(x=Y_MATR, y=CUR_TT_P, fill=Y_MATR)) +
geom_boxplot() +

```

```

  labs(x="Matriculado", y="T. cursos Ciclo Presente", fill="Matriculado") + #
titulo ejes y leyenda

  scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
  scale_fill_discrete(labels=c("Si","No"))+
  facet_wrap(~"Total de Cursos Presente según Matricula")
G_6<-ggplot(Data, aes(x=Y_MATR, y=CUR_TT_A, fill=Y_MATR)) +
  geom_boxplot() +
  labs(x="Matriculado", y="T. cursos Ciclo Anterior", fill="Matriculado") + #
titulo ejes y leyenda

  scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
  scale_fill_discrete(labels=c("Si","No"))+ # etiquetas claves leyenda
  #ggtitle("Total de Cursos Anterior según Matricula")
  facet_wrap(~"Total de Cursos Anterior según Matricula")
G_7<-ggplot(Data, aes(x=Y_MATR, y=TAS_NOM_P, fill=Y_MATR)) +
  geom_boxplot() +
  labs(x="Matriculado", y="T. cursos Ciclo Presente", fill="Matriculado") + #
titulo ejes y leyenda

  scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
  scale_fill_discrete(labels=c("Si","No"))+# etiquetas claves leyenda
  #ggtitle("Tasa de No Morosidad Presente según Matricula")
  facet_wrap(~"Tasa de No Morosidad Presente según Matricula")

G_8<-ggplot(Data, aes(x=Y_MATR, y=TAS_NOM_A, fill=Y_MATR)) +
  geom_boxplot() +
  labs(x="Matriculado", y="T. cursos Ciclo Anterior", fill="Matriculado") + #
titulo ejes y leyenda

  scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
  scale_fill_discrete(labels=c("Si","No"))+ # etiquetas claves leyenda
  #ggtitle("Tasa de No Morosidad Anterior según Matricula")
  facet_wrap(~"Tasa de No Morosidad Anterior según Matricula")

```

```

G_9<-ggplot(Data, aes(x=Y_MATR, y=EDAD, fill=Y_MATR)) +
  geom_boxplot() +
  labs(x="Matriculado", y="Edad", fill="Matriculado") + # titulo ejes y leyenda
da
  scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
  scale_fill_discrete(labels=c("Si","No"))+ # etiquetas claves leyenda
  #ggtitle("Edad del Alumno según la Matricula")
facet_wrap(~"Edad del Alumno según la Matricula")

G_10<-ggplot(Data, aes(x=Y_MATR, y=CICLO, fill=Y_MATR)) +
  geom_boxplot() + labs(x="Matriculado", y="Ciclo", fill="Matriculado") + #
titulo ejes y leyenda
  scale_x_discrete(labels=c("Si","No")) + # etiquetas del eje x
  scale_fill_discrete(labels=c("Si","No"))+ # etiquetas claves leyenda
  #ggtitle("Ciclo del Alumno según la Matricula")
facet_wrap(~"Ciclo del Alumno según la Matricula")

G_1+G_2+G_3+G_4+G_5+G_6+G_7+G_8+G_9+G_10

```

- **Balanceo de datos**

```

Data.NO<-subset(Data,Y_MATR=="1")
Data.SI<-subset(Data,Y_MATR=="0")
RNGkind(sample.kind="Rounding")
set.seed(10000)
n<-dim(Data.NO)[1]
Indices <- sample( 1:nrow(Data.SI), n )
Data.muestra.SI <- Data.SI[Indices,]
Data.model<-rbind(Data.NO,Data.muestra.SI)
dim(Data.model)

```

- **Balanceo de datos**

```

Data.model <- data.frame(Data.model)
attach(Data.model)
library(caTools)
RNGkind(sample.kind="Rounding")
set.seed(10000)
muestra<- sample.split(Data.model, SplitRatio = 0.70)

```

```
data.train <- subset(Data.model, muestra == TRUE)
data.test <- subset(Data.model, muestra == FALSE)
```

```
table(data.train$Y_MATR)
```

```
##
##    0    1
## 1515 1516
```

```
table(data.test$Y_MATR)
```

```
##
##    0    1
##  827  826
```

- **Selección de variables**

```
library(MASS)
m1 <- glm(Y_MATR~., data = data.train, family=binomial)
m2 <- stepAIC(m1, trace=F, direction="backward")
summary(m1)
```

- **Modelamiento con Logit**

```
model_Log_1 <- glm(Y_MATR ~ ., family="binomial", data=data.train)
model_Log_2 <- stepAIC(model_Log_1, trace=F, direction="backward")
summary(model_Log_1)
```

- **Predicción**

```
prob <- predict(model_Log_2, data.test, type = "response",)
pred <- ifelse(prob >= 0.5, "1", "0")
```

- **Evaluación de la capacidad predictiva**

```
library(caret)
confusionMatrix(data = as.factor(pred), reference = as.factor(data.test$Y_MATR), positive="1")
```

- **Matriz de Confusión**

```
library(gmodels)
CrossTable(x = data.test$Y_MATR, y = pred, prop.t=FALSE, prop.c=FALSE, prop.chisq = FALSE)
```

- **Curva ROC**

```
library(Epi)
ROC(prob, data.test$Y_MATR)
```

- **Calculo de la Devianza**

```
Alfa=0.05
Chi_Tab=qchisq(1-Alfa,model_Log_2$df.residual); Chi_Tab
p_valor=1-pchisq(model_Log_2$deviance,model_Log_2$df.residual); p_valor
```

- **Calculo del pseudo R2**

```
R2= (1-model_Log_2$deviance/model_Log_2$null.deviance)*100; R2
```

- **Modelando con árbol**

```
library(foreign)
library(rpart)
library(rpart.plot)
library(ggplot2)
par(mfrow=c(1,1))
RNGkind(sample.kind="Rounding")
set.seed(10000)
fit <- rpart(Y_MATR ~ . , data = data.train, method = 'class')
rpart.plot(fit,
           digits=-1,
           type=2,
           extra=101,
           cex=0.7,nn=TRUE)

print(fit)
```

- **Seleccionando Nodo de Corte**

```
plotcp(fit)
rpart.plot(fit, extra = 4)
```

- **Predicción**

```
prob1 <- predict(fit,data.test, type = "class")
pred1 <- ifelse(prob1 >= 0.5, "1", "0")
```

- **Evaluando Capacidad predictiva con el árbol.**

```
library(caret)
confusionMatrix(data = as.factor(prob1), reference = as.factor(data.test$Y_MATR),positive="1")
```

- **Matriz de confusión con el árbol**

```
library(gmodels)
CrossTable(x =data.test$Y_MATR, y = pred1,prop.t=FALSE, prop.c=FALSE, prop.chis
q = FALSE)
```

- **Calculando la curva ROC con el árbol**

```
par(mfrow=c(1,1))
library(Epi)
ROC(as.numeric(pred1),data.test$Y_MATR)
```