ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

TRABAJO FINAL DE LA MATERIA INTEGRADORA

“Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior”

Previo la obtención del Título de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

Presentado por:
MARÍA JOSÉ JURADO MANTILLA

GUAYAQUIL - ECUADOR

Año: 2019

DEDICATORIA

A Dios, por ser mi inspiración y guía en cada uno de los pasos que doy.

A mi familia, por ser mi principal soporte y quiénes me impulsan a ser mejor.

A mi madre, por los consejos y principios inculcados.

AGRADECIMIENTOS

A Dios, por permitirme culminar esta etapa en mi desarrollo profesional.

A mi familia, por alentarme para alcanzar la meta

A mi querida tutora Mgtr. Wendy Plata, por su dedicación, tiempo y paciencia para culminar este trabajo.

DECLARACIÓN EXPRESA

"Los derechos de titularidad y explotación, corresponde conforme al reglamento de propiedad intelectual de la institución. Yo, María José Jurado Mantilla doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

María José Jurado Mantilla

EVALUADORES

Ph.D. Sandra García Bustos

PROFESOR DE LA MATERIA

Mgtr. Wendy Plata Alarcón

PROFESOR TUTOR

RESUMEN

La deserción estudiantil es un problema no menor para una unidad de postgrado objeto de estudio del presente proyecto, problema que afecta al área financiera, sus estándares de calidad y al ámbito académico. El objetivo principal de este proyecto es diseñar un modelo predictivo de la deserción estudiantil. Durante el desarrollo de este proyecto integrador se tomaron en cuenta cuatro modelos de clasificación que son: Potenciación del Gradiente Estocástico (GBM), Bosques Aleatorios (RF), Redes Neuronales (NNET) y Regresión Logística (GLM); además, se utilizó la metodología KDD (Descubrimiento de Conocimiento en Bases de Datos) cuyas etapas especificadas de forma clara permiten la implementación correcta de los modelos estadísticos en conjunto con técnicas computacionales con el fin de obtener el óptimo desempeño de los mismo. Para definir el mejor modelo se evaluó cuantitativamente cada uno de ellos por medio de la matriz de confusión, sus indicadores de desempeño, la curva ROC y el AUC. Los principales resultados mostraron que los modelos Bosques Aleatorios y Redes Neuronales fueron los más destacados, pero se escogió el modelo Bosques Aleatorios por su facilidad de uso.

Palabras Clave: modelos estadísticos, eficiencia terminal, precisión, sensibilidad, especificidad.

ABSTRACT

The student dropout is a problem not less for a postgraduate unit under study of this project, a problem that affects the financial area, its quality standards and the academic field. The main objective of this project is to design a predictive model of student dropout. During the development of this project, four classification models were taken in consideration: Stochastic Gradient Boosting (GBM), Random Forests (RF), Neuronal Networks (NNET) and Logistic Regression (GLM); in addition, the KDD (Knowledge Discovery in Databases) methodology was used whose clearly specified stages allow the correct implementation of the statistical models together with computational techniques in order to obtain the optimum performance. To define the best model, each one was quantitatively evaluated through the confusion matrix, its performance indicators, the ROC curve and the AUC. The main results showed that the Random Forests and Neural Networks models were the most prominent, but the Random Forests model was chosen because of its ease of use.

Keywords: statistics models, terminal efficiency, accuracy, sensitivity, specificity.

ÍNDICE GENERAL

RESUMEN I

ABSTRACT II

ÍNDICE GENERAL III

ABREVIATURAS V

ÍNDICE DE FIGURAS VI

ÍNDICE DE TABLAS VIII

CAPÍTULO 1 9

1. INTRODUCCIÓN 9

1.1 Descripción del problema 10

1.2 Justificación del problema 11

1.3 Objetivos 12

1.3.1 Objetivo General 12

1.3.2 Objetivos Específicos 12

1.4 Marco teórico 12

1.4.1 Deserción estudiantil 12

1.4.2 Modelos explicativos de la deserción 14

1.4.3 Tasa de Eficiencia Terminal 18

1.4.4 Descubrimiento de conocimiento en bases de datos (KDD) 18

1.4.5 Coeficiente de correlación 20

1.4.6 Modelos de clasificación 21

1.4.6.1 Potenciación del Gradiente Estocástico 21

1.4.6.2 Bosques aleatorios 25

1.4.6.3 Redes neuronales 28

1.4.6.4 Regresión logística 32

1.4.7 Validación cruzada 34
1.4.8 Matriz de confusión 34
1.4.9 Curva Característica Operativa del Receptor 36
CAPÍTULO 2 38
2. METODOLOGÍA 38
2.1 Selección 38
2.1.1 Análisis de situación actual 38
2.1.2 Selección de variables 39
2.2 Preprocesamiento 40
2.3 Transformación 41
2.4 Modelización 42
2.5 Interpretación y evaluación 44
CAPÍTULO 3 45
3. RESULTADOS Y ANÁLISIS 45
3.1 Análisis de situación actual 45
3.1.1 Análisis estadístico univariado 45
3.1.2 Análisis estadístico bivariado 54
3.2 Interpretación y evaluación de los modelos 60
3.2.1 Modelos estadísticos 60
3.2.2 Comparación de los modelos 68
CAPÍTULO 4 71
4. CONCLUSIONES Y RECOMENDACIONES. 71
4.1 Conclusiones 71
4.2 Recomendaciones 72
BIBLIOGRAFÍA 74
ANEXOS 77

ABREVIATURAS

IES	Institución de Educación Superior
GBM	Potenciación del Gradiente Estocástico
RF	Bosques Aleatorios
NNET	Redes Neuronales
GLM	Regresión Logística
KDD	Descubrimiento de Conocimiento en Bases de Datos
ET	Eficiencia Terminal

ÍNDICE DE FIGURAS

Figura 1.1 Niveles y tipos de deserción (Himmel, 2002) 14
Figura 1.2 Modelo basado en la teoría del suicidio (Spady, 1970) 15
Figura 1.3 Modelo basado en la teoría de intercambio (Tinto,1975) 16
Figura 1.4 Modelo basado en la productividad del ambiente laboral (Bean,1985) 17
Figura 1.5 Proceso de descubrimiento del conocimiento en bases de datos (KDD).. 19
Figura 1.6 Algoritmo GBM (Friedman, 1999) 25
Figura 1.6 Algoritmo RF (Hastie, Tibshirani & Friedman, 2009) 28
Figura 1.6 Ejemplo de la estructura de una red neuronal. 29
Figura 1.7 Ejemplo de una red neuronal artificial. 31
Figura 3.1 Gráfico de Barras - Desertor 45
Figura 3.2 Gráfico de Barras - Programa 46
Figura 3.3 Gráfico de Barras - Ciudad de nacimiento 47
Figura 3.4 Gráfico de Barras - Sexo 47
Figura 3.5 Gráfico de Barras – Etnia 48
Figura 3.6 Gráfico de Barras – Estado civil 49
Figura 3.7 Gráfico de Barras - Cargo 49
Figura 3.8 Gráfico de Barras – Ingreso mensual promedio 50
Figura 3.9 Diagrama de cajas e histograma - Edad 51
Figura 3.10 Diagrama de cajas e histograma – Promedio general 52
Figura 3.11 Diagrama de cajas e histograma – Número de cursos tomados 53
Figura 3.12 Gráfico de Barras – Desertores por Programa 54
Figura 3.13 Gráfico de Barras – Desertores por Ciudad de nacimiento 55
Figura 3.14 Gráfico de Barras – Desertores por Sexo 55
Figura 3.15 Gráfico de Barras – Desertores por Etnia 56
Figura 3.16 Gráfico de Barras – Desertores por Estado civil 57
Figura 3.17 Gráfico de Barras – Desertores por Cargo 58
Figura 3.18 Gráfico de Barras – Desertores por Ingreso mensual promedio 59
Figura 3.19 Gráfico de barras – ET por Carrera 60
Figura 3.19 Curva ROC – GBM 61

Figura 3.20 Curva ROC – RF..... 64
Figura 3.21 Curva ROC – NNET..... 66
Figura 3.22 Curva ROC – GLM..... 67
Figura 3.22 Curva ROC 70

ÍNDICE DE TABLAS

Tabla 1.1 Matriz de confusión 35
Tabla 1.2 Interpretación de valores AUC por Swets 37
Tabla 3.1 Parámetros de los modelos seleccionados 43
Tabla 3.1 Estadística descriptiva - Edad 50
Tabla 3.2 Estadística descriptiva – Promedio General 52
Tabla 3.3 Estadística descriptiva – Número de cursos tomados 53
Tabla 3.4 Matriz de correlación 59
Tabla 3.5 Matriz de confusión – GBM 61
Tabla 3.6 Indicadores de desempeño – GBM 61
Tabla 3.7 Matriz de confusión – RF 63
Tabla 3.8 Indicadores de desempeño – RF 63
Tabla 3.9 Matriz de confusión – NNET 65
Tabla 3.10 Indicadores de desempeño – NNET 65
Tabla 3.11 Matriz de confusión – GLM 67
Tabla 3.12 Indicadores de desempeño – GLM 67
Tabla 3.13 Indicadores de desempeño 69

CAPÍTULO 1

1. INTRODUCCIÓN

El proyecto puesto en consideración del lector, se desarrolló con los datos de un unidad de postgrado en una institución pública de educación superior, ubicada en Guayaquil, en la que se ofrecen carreras de grado y programas de postgrado, las primeras son financiadas por el estado ecuatoriano, mediante lo que se denomina "gratuidad de la educación", mientras que, los programas, en especial de trayectoria profesional, son autofinanciados por los estudiantes admitidos, generando recursos que son reinvertidos en el campo académico, ya sea en infraestructura, infoestructura, tecnología e investigación científica.

La detección del fracaso de los estudiantes en su programa de estudios es un problema social serio y se ha vuelto muy importante para los profesionales de la educación entender mejor por qué existen personas que no completan sus estudios. Sin embargo, este es un problema complejo de resolver debido a la gran cantidad de factores de riesgo o características de los estudiantes que pueden influir en el fracaso de sus estudios, como la demografía, los antecedentes culturales, familiares, sociales o educativos, el nivel socioeconómico, el progreso académico y el perfil psicológico. En las últimas décadas, se han realizado una gran variedad de estudios e investigaciones sobre la identificación de las principales variables que influyen en el rendimiento de los estudiantes en los diferentes niveles de educación, desde el nivel básico hasta el nivel superior, considerando que la detección y prevención del fracaso estudiantil con su correspondiente intervención tiene un mejor resultado que la remediación.

En consecuencia, se realizó este proyecto con el objetivo de diseñar un modelo estadístico para predecir la deserción estudiantil en una unidad de postgrado y así entregar información que facilite y oriente la toma de decisiones para la aplicación de políticas educacionales.

En el primer Capítulo, se presentará la definición del problema y se mencionará la justificación para llevar a cabo este proyecto. Se presentarán tanto el objetivo general

como los objetivos específicos que se desean alcanzar y luego, se finalizará el capítulo con el marco teórico requerido para la elaboración de este proyecto.

Dentro del Capítulo 2, se aplicará la metodología KDD (Descubrimiento de Conocimiento en Bases de Datos) para el desarrollo de esta investigación, el cual consta de cinco pasos que son: Selección, Preprocesamiento, Transformación, Modelización y Evaluación.

El Capítulo 3 mostrará los métodos descriptivos para identificar las variables que describen el fenómeno de la deserción y luego mostrará los modelos predictivos para la detección de los posibles desertores de las carreras de postgrado, para posteriormente comparar el desempeño de cada uno y recomendar el mejor modelo.

Finalmente, se presentará en el Capítulo 4 las conclusiones a las que se llegaron durante el desarrollo de este proyecto, así como también se presentarán recomendaciones que serán de interés para la unidad de postgrado.

1.1 Descripción del problema

La deserción estudiantil en una unidad de postgrado es un problema no menor para una institución de educación superior (IES). Investigadores de diferentes áreas, como economía, psicología, sociología y recientemente el área de la ciencia de datos, han mostrado interés en el desarrollo de estudios para contribuir a la comprensión de este problema, y en conjunto a los costos asociados son principalmente la motivación de la investigación.

Para la IES, objeto de estudio, es importante la correcta administración de todos sus recursos con el fin de ofrecer un servicio de calidad acorde a lo presupuestado en cado programa de postgrado; sin embargo, en caso de presentarse la deserción del estudiante, esto afectaría no solamente en el ámbito académico con respecto a la disminución de la eficiencia terminal, sino también financieramente, dado que se corre el riesgo de no percibir los rubros necesarios para que el programa de

postgrado sea sustentable en el tiempo; además, no se tendría un profesional que contribuya al país, afectando el objeto para el que fue diseñado el postgrado.

La deserción de los estudiantes se convierte en un riesgo para la unidad académica, puesto que no conoce cuales son los factores y variables que inciden en la misma. Consecuentemente, al no conocer esta información la unidad académica no puede tomar decisiones y aplicar políticas educacionales para garantizar la retención de sus estudiantes y disminuir los efectos negativos ocasionados por el abandono del programa de estudios.

1.2 Justificación del problema

El conocimiento del comportamiento de la deserción y el análisis de factores que la originan es una oportunidad para aplicar nuevas estrategias que responden al desafío de mejorar la gestión académico-administrativa para reducir la probabilidad de abandono del programa de estudio, lo cual contribuirá a que la inversión en educación se refleje en beneficio para la sociedad; y, al mismo tiempo, genere un desarrollo nacional.

Actualmente, existe una gran variedad de técnicas estadísticas que son empleadas en el ámbito educativo, es por esto que en el presente proyecto se usaran modelos de clasificación y se escogerá al que tenga mejor desempeño con el objetivo de predecir a los posibles desertores con base en los registros administrativos existentes entregados por la unidad de postgrado, los cuales serán transformados y analizados para así obtener información relevante relacionada a la detección de la deserción.

Además, no solo la unidad de postgrado será el único beneficiario, sino también sus estudiantes, ya que con la información recibida la unidad académica podrá poner en marcha planes de contingencia para enfrentar el problema desde su raíz y así lograr convertirlos en profesionales con un título de cuarto nivel.

1.3 Objetivos

1.3.1 Objetivo General

Diseñar un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior.

1.3.2 Objetivos Específicos

- Analizar la situación actual de la unidad de postgrado para la identificación de variables que podrían influir en la deserción estudiantil.
- Construir modelos predictivos para la detección de la deserción de los estudiantes de las carreras de postgrado de la IES.
- Comparar el desempeño de cada uno de los modelos predictivos y recomendar el mejor de ellos.
- Determinar el perfil del estudiante desertor con base en la caracterización de las variables significativas del modelo predictivo.

1.4 Marco teórico

1.4.1 Deserción estudiantil

La deserción estudiantil está definida como el abandono anticipado de un servicio educativo previo a alcanzar el título o grado (Himmel, 2002). Tinto (1989) menciona que la deserción es un fenómeno muy complejo, por lo que su definición no abarca la totalidad del problema. Por otro lado, se tiene la teoría Spady (1970) donde define la deserción como el resultado de la falta de integración del estudiante en su entorno educativo. De forma general, se entiende que los autores definen la deserción como acto definitivo en que el estudiante deja el programa de estudios el cual se encontraba cursando. El término deserción debe diferenciarse de la palabra reprobación, puesto que, éste hace referencia cuando el estudiante no alcanza el puntaje mínimo de aprobación determinado por la institución educativa, aun agotando todas las instancias. Las entidades académicas que estudian la deserción de sus estudiantes pueden escoger la

definición más conveniente de acuerdo a sus intereses y objetivos, tomando en cuenta que el objetivo principal es la educación del individuo.

De acuerdo al artículo publicado por Himmel en el año 2002, la autora aborda la deserción desde el ámbito conceptual, considerando investigaciones internacionales previamente realizadas y así abordar las diferentes aristas del fenómeno en estudio. En el artículo, Himmel menciona que dentro de la deserción existen dos categorías: 1) Involuntaria y 2) Voluntaria. La deserción involuntaria se presenta cuando la entidad educativa toma la decisión de desligar al alumno del programa de estudios por causas académicas o disciplinarias, mientras que, la deserción voluntaria se produce cuando el estudiante toma la decisión de renunciar a la carrera, donde dicho estudiante puede informar al centro de estudios como no hacerlo (Himmel, 2002).

Según las dos categorías y los niveles de la deserción, es factible realizar un mapa conceptual para facilitar la comprensión del fenómeno como se observa en la Figura 1.1.

Figura 1.1 Niveles y tipos de deserción (Himmel, 2002)

El aporte de Himmel ha hecho posible conocer uno de los fundamentos teóricos en relación a los niveles, tipos y visiones del estudio de la deserción, así como el reconocimiento de los modelos explicativos más destacados de la deserción, lo que posibilitará la identificación de las potenciales variables que explicarían la deserción.

1.4.2 Modelos explicativos de la deserción

El fundamento teórico de la deserción ha ido evolucionando con respecto al tiempo, por lo que se presentará de forma cronológica tres de los principales modelos explicativos usados por la mayoría de los expertos en el área:

- Modelo basado en la teoría del suicidio.
- Modelo basado en la teoría de intercambio.
- Modelo basado en la productividad del ambiente laboral.

Modelo basado en la teoría del suicidio

Spady (1970) realizó uno de los primeros estudios afín a la deserción, el mismo que se encuentra sustentado en base a los principios del suicidio dictaminados por Durkheim en el año 1951. Durkheim estableció que la decisión de suicidarse es un hecho social ocasionado por la ruptura de la persona con su entorno, puesto que está imposibilitado a integrarse al mismo, por ende, la decisión de suicidarse no puede ser explicado únicamente por factores particulares del individuo. (Durkheim, 1951). Con fundamento en este argumento, Spady instaura que la deserción es la consecuencia de la no integración del sujeto con el medio educativo; además, insinúa que el entorno familiar con sus respectivas características produce fuertes efectos en el alumno, causados por las demandas, influencias y expectativas que pueden alterar tanto el rendimiento académico como la integración de la persona con el ambiente educativo.

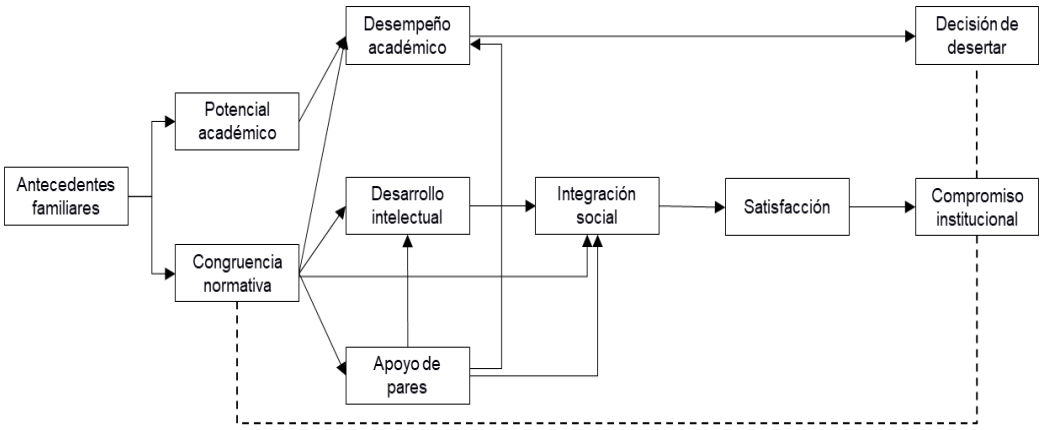

Figura 1.2 Modelo basado en la teoría del suicidio (Spady, 1970)

Se puede observar en la Figura 1.2. que Spady propone un modelo donde los antecedentes familiares causan un impacto directo tanto en su potencia académico como en su congruencia normativa del individuo, que hace referencia a la relación coherente entre los intereses, la disposición personal y las actitudes del estudiante con respecto a su entorno. A su vez, la gráfica muestra que estas dos variables afectan el apoyo a pares, el desarrollo intelectual y el desempeño académico. A partir del desarrollo intelectual y el desempeño académico, la integración social de la persona es afectada, donde luego directamente influye en la satisfacción del estudiante y de forma consecuente impacta con el compromiso hacia la institución educativa. Todas estas variables mostradas en la gráfica incurrirán en la decisión final del alumno en desertar o no, ya sea esto de manera involuntaria o voluntaria.

Modelo basado en la teoría de intercambio

Tinto (1975) publicó una revisión de las teorías planteadas hasta ese momento relacionadas a la deserción estudiantil, en esta publicación se hizo referencia al trabajo desarrollado por Spady y a su vez lo complementa, integrando la teoría del intercambio diseñada por Nye. La teoría del intercambio expone que las personas evaden los comportamientos que les generen costos de alguna forma y buscan beneficios en las interacciones, relaciones y estados emocionales con la institución educativa y sus iguales (Nye, 1976). Por medio de este criterio, para Tinto los alumnos perdurarían hasta el final en el programa de estudios siempre

que estos les genere un beneficio que supere a la dedicación, el esfuerzo y los diferentes costos personales; y, en caso de que se presente una actividad de otro tipo que le proporcione un beneficio mayor, el estudiante podría tomar la decisión final de desertar del programa de estudios causada por dicho beneficio. Véase la Figura 1.2.

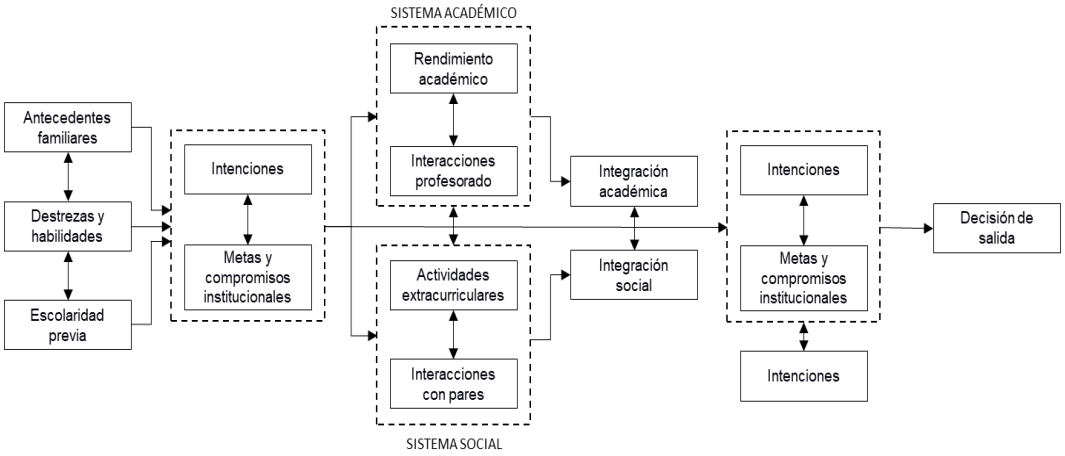

Figura 1.3 Modelo basado en la teoría de intercambio (Tinto,1975)

De acuerdo a la Figura 1.2., el modelo de Tinto considera tres variables bases que son: antecedentes familiares, destrezas y habilidades, y la escolaridad previa, las mismas que contribuyen en la adaptación de la persona en la institución ya que afectaran en su futura experiencia dentro de su carrera estudiantil. A partir de estas variables, se desprende las intenciones del estudiante y, las metas y compromisos institucionales que luego desencadenan en su relación y posterior integración con el sistema académico y con el sistema social, que abarcan su desempeño académico y su relación con profesores y compañeros. Luego, el modelo muestra una reevaluación de sus compromisos incluyendo los externos, que de forma aditiva pueden desencadenar una posible deserción.

Modelo basado en la productividad del ambiente laboral

Bajo esta misma línea de estudio, Bean realiza dos publicaciones: 1980 y 1985. En la primera publicación menciona los trabajos realizados por Spady y Tinto para probar si estos modelos poseen alguna evidencia empírica. Luego de cinco años, durante su segundo proyecto, amplía el trabajo tomando en cuenta a estudiantes

no convencionales como resultado del cambio en el ingreso a una institución educativa de la época que hasta ese momento se encontraba restringida para las personas que no pertenecían a un grupo élite. Gracias a este cambio aumentó la diversidad del grupo estudiantil, creando nuevas interrogantes en los modelos de las teorías de Spady y Tinto.

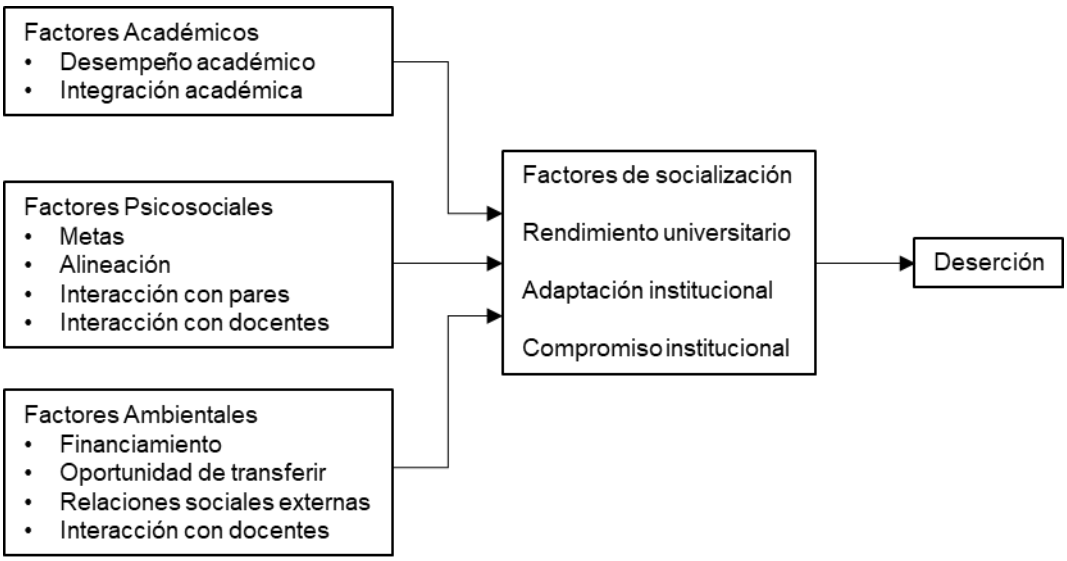

Figura 1.4 Modelo basado en la productividad del ambiente laboral (Bean,1985)

Por medio de la Figura 1.3. se observa que Bean plantea un modelo dónde la satisfacción con los estudios se explica de forma similar a la satisfacción con el trabajo, dicho modelo tiende a ser variable y con una influencia directa en las intenciones de desertar del programa de estudios. Bean otorga un gran peso a los factores no cognitivos como los psicosociales y ambientales que incluyen variables como: metas, alineación, interacción con pares, interacción con docentes, financiamiento, oportunidad de transferir y relaciones sociales externas. Luego se observa que todas estas variables afectan los factores de socialización, rendimiento universitario, adaptación y compromiso institucional que finalmente desembocan en el síndrome de deserción.

1.4.3 Tasa de Eficiencia Terminal

Según la SEP (2012) define la eficiencia terminal como la relación entre el número de estudiantes que ingresan con el número de los que se gradúan de la misma cohorte, tomando en cuenta el año de ingreso y el año de graduación acorde a la duración del programa de estudios. Por otra parte, Martínez (2001) definió la eficiencia terminal como la proporción de estudiantes que concluyen un programa de estudios en relación con aquellos que la iniciaron y considera que es un área de la calidad que debe tomarse en cuenta debido a que el costo depende de los productos de la educación superior.

Para obtener la eficiencia terminal se realiza la siguiente ecuación:

$$ET = \frac{AEg_t}{ANI_{t-2}}\ x\ 100$$

Dónde:

AEg_t	Alumnos egresados del ciclo escolar t
ANI_t	Alumnos ingresados a 1° hace t-2
t	Ciclo escolar

La tasa de eficiencia terminal es usada comúnmente para obtener la proporción de los estudiantes que concluyeron de manera regular su programa de estudios, es decir, en el tiempo ideal establecido. La tasa de eficiencia terminal toma valores entre cero y 100, lo que quiere decir que, los valores cercanos a cero indican que pocos estudiantes que ingresaron al programa de estudios dos ciclos atrás finalizaron dentro del tiempo regular esperado.

1.4.4 Descubrimiento de conocimiento en bases de datos (KDD)

El Descubrimiento de Conocimiento en Bases de Datos o KDD (por sus siglas en inglés *Knowledge Discovery in Databases*) es un conjunto de procesos metodológicos que se encarga de la extracción de los datos, de la preparación de la información y de la interpretación de los resultados obtenidos (Molina & García,

2006). Estos autores señalan que se trata de encontrar e interpretar patrones o relaciones en los datos recolectados usando técnicas de aprendizaje automático, estadísticas y de visualización.

El KDD realiza un proceso interactivo e iterativo en la búsqueda de patrones, parámetros y modelos, los mismos que deben ser potencialmente válidos y útiles. Se deben establecer medidas cuantitativas que ayuden a considerar la validez y utilidad de los patrones escogidos, con el fin de integrar el conocimiento adquirido y aplicarlo en la toma de decisiones de algún sistema real con la ayuda de los resultados alcanzados.

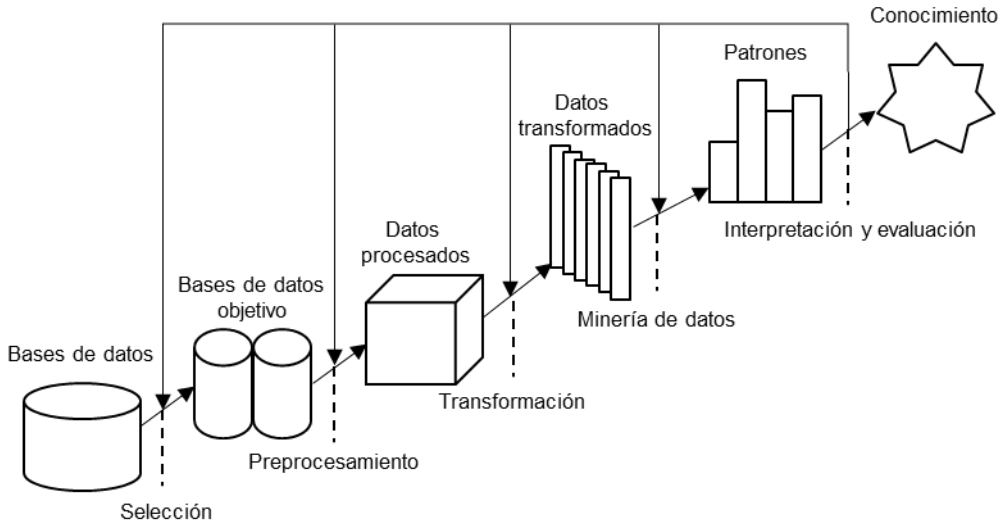

Figura 1.5 Proceso de descubrimiento del conocimiento en bases de datos (KDD).

En la Figura 1.5. se observa los pasos del método de descubrimiento de conocimiento en bases de datos, los cuales se describen a continuación:

- **Selección:** Es la primera etapa del proceso dónde se establece el objetivo del estudio y se realiza una elección coherente de los datos con su respectiva fuente, con el fin de poder definir una meta antes de iniciar el estudio como tal. Además, se escogen las variables que serán usadas para el posterior análisis.
- **Preprocesamiento:** En esta etapa, se eliminan las observaciones inconsistentes, a fin de separar los datos del ruido y así evitar sesgo durante el procesamiento en el paso de Modelización. Así mismo, se busca que los datos

sean completos, en caso de tener observaciones faltantes se considerará un método de imputación de datos para asegurar que los datos escogidos cumplen un nivel mínimo de calidad.

- **Transformación:** En la etapa de transformación se busca cambiar los datos procesados a un apropiado formato para un óptimo desempeño en el paso de Modelización. Cabe resaltar que, las primeras dos etapas del KDD son las que más tiempo consumen, puesto que se debe tener cuidado en el tratamiento que se le da a los datos, ya que podría afectar en la calidad de estos.
- **Modelización:** Este paso es caracterizado como la parte más relevante del proceso KDD, el cual tiene como objetivo poder identificar los patrones de comportamiento tanto descriptivo como predictivo a partir de los datos obtenidos.
- **Interpretación y evaluación:** En este paso se establecen las medidas cuantitativas, para verificar si el modelo es válido y a su vez identificar los patrones más relevantes. También se aplican técnicas de visualización con el propósito de facilitar la comprensión de la nueva información con los interesados.

1.4.5 Coeficiente de correlación

El coeficiente de correlación mide el grado de asociación o relación entre dos variables aleatorias cuantitativas que provienen de una distribución normal bivariada conjunta (Restrepo & González, 2007). Se define como:

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y}, \qquad -1 \leq \rho \leq 1$$

Cuando $\rho > 0$ existe una relación directa entre las variables. Si $\rho < 0$, significa que la relación entre las variables es inversa. Finalmente, si $\rho = 0$ las variables son independientes.

1.4.6 Modelos de clasificación

En el campo de la estadística y el aprendizaje automático, los modelos de clasificación consisten en identificar a cuál conjunto de categorías o subpoblaciones pertenece una nueva observación, por medio de una base de entrenamiento donde es conocida la categoría de pertenencia de cada uno de los datos (Alpaydin, 2014). Los modelos de clasificación son considerados un caso del aprendizaje supervisado, es decir, las observaciones de un conjunto de entrenamiento se encuentran identificadas correctamente.

1.4.6.1 Potenciación del Gradiente Estocástico

Es una herramienta estadística de aprendizaje supervisado que se utiliza para problemas de regresión como de clasificación, el cual crea un modelo predictivo ajustando secuencialmente una función parametrizada simple a los pseudo residuales actuales por mínimos cuadrados en cada iteración (Friedman, 1999). El modelo de Potenciación del Gradiente Estocástico (GBM por sus siglas en inglés) se lo construye de forma escalonada similar a los métodos de *boosting* y luego se generaliza optimizando una función de pérdida diferenciable. De manera formal se tiene que:

Potenciación del gradiente

En el problema de estimación de funciones, se tiene un sistema que consiste en una "salida" aleatoria o variable de "respuesta" y y un conjunto de una "entrada" aleatoria o variables "explicativas" $x = \{x_1, x_2, \dots, x_n\}$. Dada una muestra de "entrenamiento" $\{y_i, x_i\}_1^N$ de valores conocidos (y, x), el objetivo es encontrar una función $F^*(x)$ que mapee x a y, de modo que sobre la distribución conjunta de todos los valores (y, x), el valor esperado de alguna función de pérdida especificada $\Psi(y, F(x))$ es minimizada:

$$F^*(x) = \arg\min_{F(x)} E_{y,x}\ \Psi(y, F(x))$$

Aplicando boosting aproxima $F^*(x)$ a la forma:

$$F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m)$$

donde la función $h(x; a)$ (“base aprendiz”) generalmente se elige como funciones simples de x con parámetros $a = \{a_1, a_2, \ldots\}$. Los coeficientes de expansión $\{\beta_m\}_0^M$ y los parámetros $\{a_m\}_0^M$ se ajustan conjuntamente a los datos de entrenamiento de manera progresiva "por etapas". Se comienza con una aproximación inicial $F_0(x)$, y para $m = 1,2,\ldots,M$

$$(\beta_m, a_m) = \arg\min_{\beta,a} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(x_i) + \beta h(x_i; a))$$

y

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m)$$

Potenciación del gradiente resuelve aproximadamente las funciones de pérdida arbitrarias (diferenciables) $\Psi(y, F(x))$ con un procedimiento de dos pasos. Primero, la función $h(x; a)$ se ajusta por mínimos cuadrados

$$a_m = \arg\min_{a,p} \sum_{i=1}^{N} [\tilde{y}_{im} - p h(x_i; a)]^2$$

a los actuales pseudo - residuales

$$\tilde{y}_{im} = -\left[\frac{\partial\, \Psi\big(y_i, F(x_i)\big)}{\partial\, F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

entonces, dado que $h(x_i; a_m)$, se determina el valor óptimo del coeficiente β_m

$$\beta_m \arg\min_{\beta} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m))$$

Esta estrategia reemplaza un problema de optimización de funciones potencialmente difícil por uno basado en mínimos cuadrados, seguido de una optimización de un solo parámetro basada en un criterio general.

La potenciación del gradiente de árbol especializa este enfoque en el caso en que la base aprendiz $h(x; a)$ es un nodo L-terminal de un árbol de regresión. En cada iteración m, un árbol de regresión divide el espacio X en L-disjuntas regiones $\{R_{lm}\}_{l=1}^{L}$ y predice un valor constante separado en cada uno

$$h(x; \{R_{tm}\}_{l=1}^{L}) = \sum_{l=1}^{L} \bar{y}_{lm} 1(x \in R_{lm})$$

Aquí $\bar{y}_{lm} = media_{x_i \in R_{lm}}(\tilde{y}_{im})$ es la media de cada región R_{lm}. Los parámetros de esta base aprendiz son las variables de división y los puntos de división correspondientes que definen el árbol, que a su vez definen las regiones correspondientes $\{R_{lm}\}_1^L$ de la partición en la iteración m-enésima. Estos se inducen de una mejor manera de arriba hacia abajo utilizando un criterio de división de mínimos cuadrados. Con los árboles de regresión, se puede resolver por separado dentro de cada región R_{lm} definida por el nodo terminal l correspondiente al árbol m-enésimo. Debido a que el árbol predice un valor constante $\bar{y}_{lm}$ dentro de cada región R_{lm}, la solución se reduce a una estimación de ubicación simple basada en el criterio Ψ

$$\gamma_{lm} = \arg\min_{\gamma} \sum_{x_i \in R_{lm}} \Psi(y_i, F_{m-1}(x_i) + \gamma)$$

La aproximación actual de $F_{m-1}(x)$ se actualiza por separado en cada región correspondiente

$$F_m(x) = F_{m-1}(x) + v \cdot \gamma_{lm} 1(x \in R_{lm})$$

El parámetro de contracción $0 < v < 1$ controla la velocidad de aprendizaje del procedimiento. Empíricamente, se encontró que los valores pequeños $(v < 0.1)$ conducen a un error de generalización mucho mejor.

Friedman presentó algoritmos específicos basados en esta plantilla para varios criterios de pérdida, incluidos los mínimos cuadrados, la desviación mínima absoluta y, para la clasificación, la probabilidad de registro negativa multinomial de clase K.

Potenciación del gradiente estocástico

Con el procedimiento de *bagging*, Breiman introdujo la noción de que inyectar aleatoriedad en los procedimientos de estimación de funciones podría mejorar su rendimiento. Propuso un procedimiento híbrido de refuerzo de *bagging* destinado al ajuste de mínimos cuadrados de expansiones aditivas. Reemplaza la base aprendiz en el refuerzo regular para incorporar la aleatoriedad como parte integral del procedimiento. Específicamente, en cada iteración se extrae una submuestra de los datos de entrenamiento al azar (sin reemplazo) del conjunto completo de datos de entrenamiento. Esta submuestra seleccionada al azar se usa, en lugar de la muestra completa, para ajustarse a la base aprendiz y calcular la actualización del modelo para la iteración actual.

Sea $\{y_i, x_i\}_1^N$ la muestra completa de datos de entrenamiento y $\{\pi(i)\}_1^N$ sea una permutación aleatoria de los enteros $\{1, \dots, N\}$. Entonces, $\left\{y_{\pi(i)}, x_{\pi(i)}\right\}_1^{\widetilde{N}}$ da una submuestra aleatoria de tamaño $\widetilde{N} < N$. Por lo tanto, el algoritmo de potenciación del gradiente estocástico es:

1	$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} \Psi(y_i, \gamma)$
2	Desde $m = 1$ a M hacer:
3	$\{\pi(i)\}_1^N = permutación\ aleatoria\ \{i\}_1^N$
4	$\tilde{y}_{\pi(i)m} = -\left[\frac{\partial\, \Psi\left(y_{\pi(i)}, F\left(x_{\pi(i)}\right)\right)}{\partial\, F\left(x_{\pi(i)}\right)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \ldots, \tilde{\text{N}}$
5	$\{R_{lm}\}_1^L = nodo\ L - terminal\ de\ un\ árbol\left(\left\{\tilde{y}_{\pi(i)m}, x_{\pi(i)}\right\}_1^{\bar{N}}\right)$
6	$\gamma_{lm} = \arg\min_{\gamma} \sum_{x_{\pi(i)} \in R_{lm}} \Psi(y_{\pi(i)}, F_{m-1}\left(x_{\pi(i)}\right) + \gamma)$
7	$F_m(x) = F_{m-1}(x) + v \cdot \gamma_{lm} 1(x \in R_{lm})$
8	Fin

Figura 1.6 Algoritmo GBM (Friedman, 1999)

1.4.6.2 Bosques aleatorios

Los bosques aleatorios es un modelo de aprendizaje supervisado no paramétrico, donde usa una técnica de clasificación basado en un conjunto de árboles de decisiones (Medina & Ñique, 2017). El modelo escoge de manera aleatoria un cierto número de variables con las cuales se construye individualmente cada uno de los árboles y realizando las predicciones con las variables escogidas, para luego ser ponderadas por medio del cálculo de la categoría más votada de los árboles que fueron generados, para finalmente realizar la predicción por medio los bosques aleatorios. Para definir de manera formal este modelo se debe conocer lo siguiente:

Árboles de clasificación:

Los árboles de clasificación y regresión (CART), creados por Breiman, Friedman, Olshen y Stone en el año 1984, consisten en un modelo de aprendizaje supervisado, donde se sigue la idea de la estructura de un árbol compuesto por hojas, ramas, nodos y una raíz. Un árbol de clasificación y regresión inicia desde un nodo inicial llamado raíz, luego se extiende hacia abajo y usualmente se desarrolla de izquierda a derecha; los nodos son la

posición donde se dividen las ramas y son representados en círculos, mientras que las ramas conectan cada nodo del árbol y son representadas por segmentos de recta. Las hojas del árbol son los nodos externos de la cadena. (Ali, Khan, Ahmad y Maqsood, 2012).

Para los árboles de clasificación, donde el resultado toma valores discretos, los nodos deben seguir criterios por el cual el conjunto de datos será dividido, estos criterios son cuantificados por varias medidas. Para un nodo m, que representa una región R_m con N_m observaciones, se tiene que:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

es la estimación de la proporción de observaciones de clase k en el nodo m. Se clasifica las observaciones en el nodo m a la clase $k(m) = argmax_k\, \hat{p}_{mk}$, que es la clase mayoritaria en el nodo m. Las diferentes medidas de $Q_m(T)$ para medir la impureza de un árbol incluyen las siguientes:

- Índice de clasificación errónea: Mide la proporción de observaciones mal clasificadas cuando todos los datos del nodo son asignados a la mayoría de la que ellos pertenecen, de acuerdo al criterio de voto por mayoría. Este índice es calculado por la siguiente ecuación:

$$Error\ de\ clasificación = 1 - \ \hat{p}_{mk(m)}$$

- Índice de entropía: Mide las diferencias de las distribuciones de probabilidad de cada uno de los grupos de clasificación. Se puede calcular de la siguiente manera:

$$Entropía = \ -\sum_{k=1}^{K} \hat{p}_{mk} log\ \hat{p}_{mk}$$

- Índice de Gini: Mide la impureza existente luego de ramificar por una categoría a la base de casos de entrenamiento. Su ecuación viene dada por:

$$Gini = -\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Bagging

Bagging es el acrónimo de *bootstrap aggregation*, es una técnica para reducir la varianza de una función de predicción estimada (Cárdenas-Montes, 2015). Básicamente, la idea es remuestrear las observaciones y obtener las predicciones en base al conjunto remuestreado. Al obtener varios modelos y promediar estos, se logra un mejor ajuste causado por la mitigación de los modelos con alta varianza y los modelos con sesgo. Si el caso es de clasificación, entonces el "voto" de la mayoría proporcionará la categoría dominante o con mejor predicción.

Suponiendo que existe un árbol que produce un clasificador $\hat{G}(x)$ para una respuesta de K categorías. Es aquí donde es útil tomar en cuenta una función de tipo indicador-vector subyacente $\hat{f}(x)$, con valores de $K-1$ ceros y un solo valor de uno, de tal manera que $\hat{G}(x) = argmax_k \hat{f}(x)$. Entonces, el bagging estimado $\hat{f}_{bag}(x)$ es un K vector $[p_1(x), p_1(x), \dots, p_k(x)]$, con $p_k(x)$ igual a la proporción de árboles que predicen la lase K en x. Por ende, el clasificador bagging escoge la clase con más "votos" de los árboles B, dando la siguiente ecuación:

$$\hat{G}_{bag} = argmax_k \hat{f}_{bag}(x)$$

Bosques aleatorios

Los bosques aleatorios es una modificación sustancial del *bagging*, que construye un gran conjunto de árboles descorrelacionados y luego los promedia (Breiman, 2001). La idea esencial, consiste combinar los árboles predictivos usándolos como clasificadores débiles para luego obtener un clasificador robusto o fuerte. El algoritmo para bosques aleatorios es:

1	Desde $b = 1$ a B hacer:
2	Realice muestras bootstrap de tamaño N de los datos de entrenamiento
3	Para cada muestra, crezca un árbol realizando recursivamente los siguientes pasos para cada hoja del árbol hasta que el mínimo nodo de tamaño n_{min} es alcanzado
4	1. Seleccione aleatoriamente m variables de la p variables
5	2. Elija la mejor variable/punto de división entre las m
6	3. Divida el nodo en dos nodos hijas
7	Muestre el conjunto de árboles $\{T_b\}_1^B$
8	Para realizar la clasificación: Sea $\hat{C}_b(x)$ la predicción de clase del b-enésimo bosque aleatorio. Entonces $\hat{C}_{rf}^B(x) = voto\ mayoritario\ \{\hat{C}_b(x)\}_1^B$

Figura 1.7 Algoritmo RF (Hastie, Tibshirani & Friedman, 2009)

Cuando se usa para la clasificación, un bosque aleatorio obtiene un voto de clase de cada árbol y luego se clasifica usando el voto mayoritario. Para obtener el valor de m, según los autores, recomiendan calcular $\sqrt{p}$ y el tamaño mínimo de nodo es uno.

1.4.6.3 Redes neuronales

Las redes neuronales artificiales se crearon a partir de la necesidad de los investigadores de recrear un sistema que pueda representar las funciones del sistema nervioso de una persona (Hastie, Tibshirani & Friedman, 2009). El desarrollo de una red neuronal artificial se basa en un modelo de

optimización no lineal. Están formados por nodos llamados neuronas. Estos nodos reciben un conjunto de información de otros nodos y proporcionan información. Esta información saliente puede provenir de tres funciones:

- **Función de propagación:** esta función se compone de la suma de la información de entrada multiplicada por un peso de interconexión.
- **Función de activación**: esta función se encarga de modificar la función anterior, es decir, se utiliza como método de aprendizaje.
- **Función de transferencia:** es la función que se aplica al valor que ofrece la función de activación.

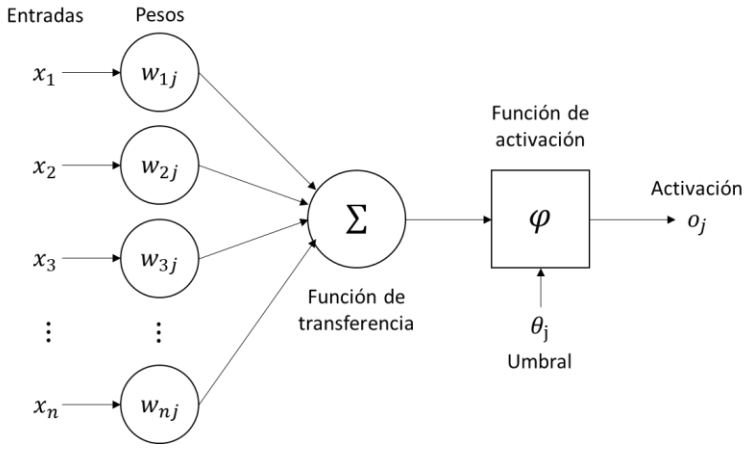

Figura 1.8 Ejemplo de la estructura de una red neuronal.

Perceptrón

Un perceptrón es la forma más simple de una red neuronal artificial que tiene una neurona de salida. Los valores de entrada $x_1, x_2, \dots, x_n$ se multiplican con diferentes pesos $w_1, w_2, \dots, w_n$ y entregan una función de salida $f(x)$ (Hastie, Tibshirani & Friedman, 2009).

Suponiendo que los pesos se establecieron durante el proceso de capacitación, la predicción de una observación se puede establecer en los siguientes pasos. Primero, calcule la combinación lineal de los pesos con las observaciones de entrada para crear una nueva observación, esto se calcula de la siguiente manera:

$$w_1x_1 + w_2x_2 + \cdots + w_nx_n - \varepsilon = \boldsymbol{w'x} - \varepsilon$$

Por lo tanto, la predicción $f(x)$ se obtiene como:

$$f(x) = g(w_1x_1 + w_2x_2 + \cdots + w_nx_n - \varepsilon) = g(\boldsymbol{w'x} - \varepsilon)$$

Donde $g(\cdot)$ es la función de activación y su objetivo es mapear la combinación lineal previamente calculada con el conjunto de valores posibles de la variable dependiente. Después de la implementación de la función de activación, se ejecuta un algoritmo iterativo que establece los valores de los pesos w_i, examinando uno por uno secuencialmente las observaciones del vector $\boldsymbol{x}$

Redes de prealimentación multinivel

Las redes neuronales artificiales pueden adoptar una estructura más compleja, que es el tipo de prealimentación de varios niveles, que tiene tres componentes principales:

- **Capa de entrada:** conjunto de neuronas que representan las variables predictoras.
- **Capa oculta:** conjunto de neuronas que reciben información de la capa de entrada, analiza sus patrones y establece el peso de cada uno correspondiente a cada uno de forma iterativa.
- **Capa de salida:** conjunto de neuronas que entregan la respuesta de la red neuronal artificial.

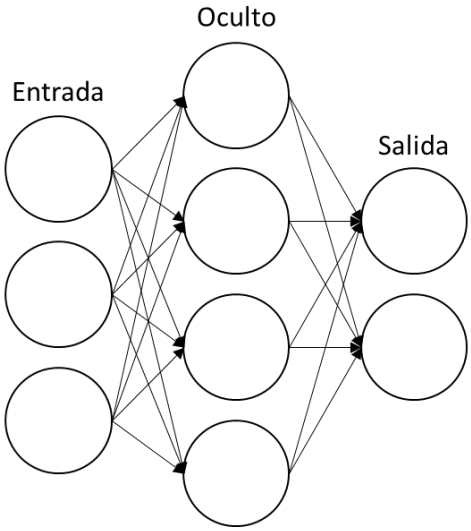

Figura 1.9 Ejemplo de una red neuronal artificial.

Cada neurona funciona de forma independiente, los pesos se asociarán con cada borde que conecta las neuronas, y cada neurona se asocia con un coeficiente de distorsión ε y una función de activación $g(\cdot)$. El método utilizado para establecer los coeficientes de distorsión se denomina algoritmo de retropropagación.

El algoritmo de retropropagación comienza con los pesos aleatorios w_i. Luego, cada momento del entrenamiento se analiza secuencialmente, creando una predicción para cada observación y, por lo tanto, comienza a obtener predicciones correctas e incorrectas. Estas respuestas se utilizan como retroalimentación para ajustar los valores de los pesos y nuevamente realizar el análisis con pesos y coeficientes de distorsión ajustados.

Las redes neuronales artificiales tienen un rendimiento que se puede convertir en "cajas negras", dado que a diferencia de otros modelos es complejo llegar a comprender e interpretar con claridad cada uno de los parámetros y sus funciones, por lo que solo es posible evaluarla a través de sus resultados.

1.4.6.4 Regresión logística

Es un modelo lineal de clasificación dentro del aprendizaje supervisado que es utilizado para predecir un dato de tipo categórico, es decir, que toma valores discretos finitos, en función de variables predictoras o independientes. La regresión logística se encuentra dentro del conjunto de Modelos Lineales Generalizados (por sus siglas en inglés GLM) dónde se hace uso de la función Logit, el cual modela las probabilidades de un único ensayo describiendo el posible resultado por medio de una función logística. (Hastie, Tibshirani & Friedman, 2009).

La ecuación de partida para un modelo de regresión logística es:

$$P(y = 1|x) = \frac{e^{b_0 + \sum_{i=1}^{n} b_i x_i}}{1 + e^{b_0 + \sum_{i=1}^{n} b_i x_i}}$$

Donde:

$P(y = 1\|x)$	Es la probabilidad de que y sea igual a 1, en presencia de las variables predictoras $x_1, x_2, \dots, x_j$ perteneciente al conjunto X.
b_0	Es el término independiente o promedio del modelo.
j	Es el número de variables predictoras.
b_i	Es el coeficiente de las variables predictoras.
x_j	Son las variables predictoras que pertenecen al modelo.

Transformando logarítmicamente la función previa se tiene que:

$$\log\left(\frac{P(y = 1|X)}{1 - P(y = 1|X)}\right) = b_0 + \sum_{i=1}^{n} b_i x_i$$

En la expresión previamente mencionada se puede identificar a la izquierda de la ecuación la función Logit, es decir, el logaritmo natural del *odd ratio*

de la variable de respuesta. Por el otro lado de la ecuación se tiene la expresión de una regresión lineal.

Ajuste del modelo

La técnica utilizada para obtener la estimación de los coeficientes regresores es el de máxima verosimilitud, el mismo que radica en maximizar la función de verosimilitud que se puede observar a continuación:

$$\ell(\beta) = \prod_{i=1}^{n} P(y = 1|X)^{y_i}(1 - P(y = 1|X))^{1-y_i}$$

Generando $k + 1$ ecuaciones de verosimilitud que son obtenidas mediante la diferenciación de la función log-verosímil con relación a los $k + 1$ coeficientes. Para este caso, las ecuaciones de verosimilitud son:

$$\sum_{i=1}^{n} \big(y_i - P(y = 1|X)\big) = 0$$

$$\sum_{i=1}^{n} x_{ij}\big(y_i - P(y = 1|X)\big) = 0, \;\; j = 1,2, \dots, q$$

Resolviendo las ecuaciones por medio del método de Newton – Raphson se llega a la siguiente ecuación de estimación:

$$b^{(t)} = b^{(t-1)} + \left[X' Diag\left[n_j p_j^{(t-1)}(1 - p_j^{(t-1)})\right]X\right]^{-1} X'(y - m^{(t-1)})$$

siendo $m^{(t-1)} = n_j p_j^{(t-1)}$ y $p_j^{(t-1)}$ la anteriormente mencionada.

Finalmente, la estimación por máxima verosimilitud de p_j viene dada por:

$$\widehat{p_J} = \frac{e^{\hat{b}_0 + \sum_{i=1}^{n} \hat{b}_i x_i}}{1 + e^{\hat{b}_0 + \sum_{i=1}^{n} \hat{b}_i x_i}}$$

Donde $\hat{b}_0$ y $\hat{b}_i$ son los estimadores de máxima verosimilitud de los parámetros.

1.4.7 Validación cruzada

La validación cruzada es una técnica estadística que sirve para evaluar el desempeño de un modelo dividiendo las observaciones en dos segmentos: el primero es usado para el entrenamiento del modelo y el segundo es para comprobar la validez de este (Refaeilzadeh *et al*, 2008). En una típica validación cruzada, los segmentos de entrenamiento y comprobación deben cruzarse en rondas sucesivas de tal manera que cada observación tenga la opción de ser validado. La forma básica de la validación cruzada es la de k-iteraciones, donde a partir de esta nacen los demás casos.

La validación cruzada de k-iteraciones consiste en dividir en k subconjuntos, donde se escoge uno para usarlo de prueba y los otros se los usa de entrenamiento. Este proceso se repite k veces con cada uno de los subconjuntos y estima los k errores de predicción. Finalmente, se obtiene la media de los todos los k errores de predicción. A pesar de que este tipo de validación cruzada puede alcanzar un buen estimador por las k combinaciones del proceso, hay que tomar en cuenta que el tiempo computacional puede ser extenso.

1.4.8 Matriz de confusión

La matriz de confusión es aquella que almacena información sobre las clasificaciones reales y predichas mediante un modelo de clasificación. El rendimiento del modelo es comúnmente evaluado por medio de los datos que se encuentran dentro de la matriz. En la Tabla 1.1 se puede observar una matriz de confusión para dos categorías

Tabla 1.1 Matriz de confusión

		Predicción	
		Negativo	Positivo
Real	Negativo	Verdaderos Negativos (VN)	Falsos Positivos (FP)
	Positivo	Falsos Negativos (FN)	Verdaderos Positivos (VP)

Las entradas en la matriz de confusión expresan lo siguiente en el contexto de este estudio:

- VN es el número de predicciones que fueron clasificadas como negativos de forma correcta.
- FN es el número de predicciones que fueron clasificadas como negativos de forma incorrecta.
- FP es el número de predicciones que fueron clasificadas como positivos de forma incorrecta.
- VP es el número de predicciones que fueron clasificadas como positivos de forma correcta.

A partir de la matriz de confusión de dos clases, se puede obtener los siguientes indicadores:

- Exactitud: Es la proporción del número total de predicciones que fueron correctas, es obtenida mediante la siguiente ecuación:

$$Exactitud = \frac{VP + VN}{VN + FN + FP + VP}$$

- Sensibilidad: Es una fracción que indica si el modelo puede identificar correctamente cuando la clasificación es positiva, se la obtiene por medio de la siguiente ecuación:

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

- Especificidad: Es una proporción que muestra si el modelo es capaz de identificar de forma correcta cuando la clasificación es negativa, es obtenida a través de la siguiente ecuación:

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

1.4.9 Curva Característica Operativa del Receptor

La Curva Característica Operativa del Receptor o Curva ROC (por sus siglas en inglés *Receiver Operating Characteristic*) es una representación gráfica que ayuda a evaluar la capacidad discriminatoria de un modelo de clasificación para asignar correctamente tomando 1-especificidad frente a la sensibilidad para cada posible valor del punto de corte (Del Valle, 2017). Es decir:

$$ROC(c)\begin{cases} y = S(c) \\ x = 1 - E(c) \end{cases}$$

Dado que, en ambos ejes se tienen probabilidades, la curva ROC se encontrará en un cuadrado [0,1] x [0,1]. Además, se considera que la curva estará contenida en el triángulo {(x, y) | 0 ≤ x ≤ y ≤ 1} por convención.

Una medida que ayuda a medir el desempeño discriminante de la Curva ROC es el área bajo la curva o AUC (por sus siglas en inglés, *area under curve*). Además, sirve como medida de comparación entre modelos y así determinar cuál de ellos es el más eficaz. De manera formal el AUC está definido como:

$$AUC = \int_0^1 ROC(t)\delta t$$

Dónde el rango de valores parte del 0.5, correspondiente a un modelo sin capacidad de discriminar, hasta 1, que representa un modelo capaz de clasificar

perfectamente los dos grupos en sus categorías. Por ende, se puede decir que cuanto mayor sea el AUC mejor desempeño tendrá el modelo.

Tabla 2.2 Interpretación de valores AUC por Swets

Baja exactitud	[0.5,0.7)
Útiles para algunos propósitos	[0.7,0.9)
Alta exactitud	[0.9,1]

El criterio propuesto por Swets dictamina que un AUC menor a 0.7 demuestra que el modelo discrimina con una exactitud baja, cuando el AUC se encuentra por debajo del 0.9 significa que el modelo puede ser útil para algunos propósitos y los mayores a 0.9, el modelo tiene una alta exactitud (Swets, 1988).

CAPÍTULO 2

2. METODOLOGÍA

Los modelos estadísticos son aplicables en diferentes contextos, son útiles para una extensa diversidad de ciencias, desde la física hasta las ciencias de la salud, desde las ciencias sociales hasta el control de calidad, incluso se utilizan para clasificar individuos según características predominantes y de ser posible, predecir su comportamiento con la finalidad de tomar decisiones. Aplicando los modelos estadísticos junto a técnicas computacionales es posible resolver problemas complejos en cualquier tipo de proyecto o investigación. En este sentido, se usará la metodología KDD planteada por Fayyad, cuyas etapas especificadas de forma clara permiten la implementación correcta de los modelos con el fin de obtener su óptimo desempeño, en este caso predecir con el mínimo error los posibles estudiantes que vayan a desertar de un programa de estudios de postgrado de la IES.

2.1 Selección

Dentro de esta etapa, se realiza el proceso de descubrimiento a través de los datos proporcionados por la Unidad de Postgrado de la IES, objeto de estudio del presente proyecto. Es aquí donde se identifica la información relevante y prioritaria, para lograr este objetivo se realizó la selección de las variables potenciales para posteriormente realizar un análisis de situación actual y un análisis de las variables, para finalmente construir la base de datos objetivo.

2.1.1 Análisis de situación actual

La situación actual de la Unidad de Postgrado se realizó mediante un análisis estadístico univariado de cada una las potenciales variables. Luego, para el desarrollo de este proyecto se definieron indicadores con el fin de medir el comportamiento de los desertores, aplicando un análisis estadístico bivariado y para las variables numéricas se obtuvo el coeficiente de correlación entre cada una de ellas. Los indicadores a considerar son:

- Porcentaje de desertores por programa
- Porcentaje de desertores por ciudad de nacimiento
- Porcentaje de desertores por etnia
- Porcentaje de desertores por estado civil
- Porcentaje de desertores por ocupación
- Porcentaje de desertores por ingreso mensual promedio

La forma de cálculo junto a su definición de cada indicador puede ser consultadas en el Manual de Indicadores dentro de la sección de Anexos.

2.1.2 Selección de variables

A partir de los datos proporcionados por la unidad de postgrado, se realizó la selección de las variables potenciales para el análisis estadístico, posteriormente se construyó una base de datos de trabajo, de la cual se escogieron las siguientes variables:

- **Variables categóricas:** Desertor, Programa, Ciudad de nacimiento, Sexo, Etnia, Estado civil, Cargo, Ingreso mensual promedio y Estado académico del estudiante.
- **Variables numéricas:** Edad, Promedio general y Número de cursos tomados.

Luego de escoger las variables, se tomaron en cuenta sólo los registros que cumplían el requisito de que tenía que haber ingresado en el primer semestre del año 2017; además, se seleccionaron los programas de estudios que iniciaron dentro de esta cohorte y al menos hayan tenido un estudiante desertor. Una vez seleccionada la fuente de datos y analizado los atributos de cada uno de estos, se conduce a la integración en un solo repositorio para así hacer uso de los datos y que los analistas en el área puedan verlos como un colectivo que proviene de una sola fuente bien definida y granulada (Inmon et al, 2008).

Es importante mencionar que la variable categórica *Estado académico del estudiante* es usada para obtener durante la etapa de Modelización la tasa de eficiencia terminal, mas no para ser usada en la creación de los modelos por alta relación con la variable de respuesta Desertor.

2.2 Preprocesamiento

La calidad del conocimiento por descubrir no solo depende del modelo escogido, sino también de la calidad de los datos por analizar. Por esta razón, luego de haber realizado la selección del subconjunto de datos, la siguiente etapa del KDD es el preprocesamiento de estos para luego ser evaluados.

Una vez recopilados los datos, es usual que aparezcan datos faltantes o datos inconsistentes, los cuales se deben sustituir de alguna forma, por varias razones: el modelo a usar puede no desempeñarse de la forma esperada, el análisis exploratorio de datos no entregaría una estimación válida, el investigador podría llegar a conclusiones erradas, etc. Los datos faltantes pueden ser reemplazados obteniendo valores por diferentes métodos, sin embargo, para el desarrollo de este trabajo se escogió el criterio de imputación de registros que posean al menos una observación faltante, es decir, estos registros no son considerados dentro de la base de datos a usar.

Dentro de esta etapa, luego de realizar un análisis exhaustivo de los datos, se pudo identificar lo siguiente con respecto a las variables categóricas:

- En la variable “Programa” se renombraron las categorías puesto que sus etiquetas constaban de nombres muy largo, por lo que se procedió a tomar la palabra clave que identifique cada una estas categorías. Por ejemplo, se tiene que la categoría Maestría en Economía y Dirección de Empresa fue renombrada como “Economía”.
- La variable “Ciudad de nacimiento” contaba con 45 niveles, sin embargo, sus registros se encontraban acumulados en el nivel “Guayaquil”, por ende, se agrupó las categorías restantes un uno solo, el cual fue llamado “Otro”.

- La variable “Etnia” poseía ocho niveles, pero la mayoría de sus registros se encontraban en el nivel “Mestizo”, por lo que se procedió a agrupar el resto de los niveles en un solo grupo renombrado como “Otro”.
- En la variable “Cargo” se puede encontrar inicialmente 12 niveles, de los cuales cuatro de ellos contaban con 5 o menos observaciones, por lo que se decidió reasignar estos registros al nivel “Otros”. Es decir, ahora la variable cuenta con ocho niveles.
- Dentro de la variable “Ingreso mensual promedio”, el nivel “No Aplica” contaba con una sola observación, por lo que se decidió imputar todo el registro y descartar el nivel. Es decir, que la variable pasó de tener ocho niveles a siete.
- La variable “Estado civil”, inicialmente contaba con cuatro niveles, pero se identificó que el nivel “Unido” poseía una sola observación, así que se decidió eliminar todo el registro en conjunto con el nivel. Por lo tanto, la variable ahora cuenta con tres categorías.

2.3 Transformación

Generalmente en las funciones que son aplicadas dentro de la etapa de Modelización, se da por sentado que los datos cumplen con ciertos requisitos o supuestos y técnicas de desempeño, es por esto que es necesario transformar los datos para que éstos al ser usados en los modelos estadísticos proporcionen resultados consistentes.

Uno de los requisitos más comunes es que todas las variables deben contener variables numéricas, pero, en la realidad las variables de una base de datos pueden ser de tipo categórica, por ende, no cumplen esta condición. Este proyecto, dentro del subconjunto de datos previamente seleccionado, cuenta con ocho variables categóricas, las cuales fueron transformadas a variables dicotómicas de la siguiente manera:

- **Variables Binomiales:** Se transformaron las variables que toman valores discretos binarios, que son “Sexo”, “Ciudad de nacimiento” y “Etnia”, reasignándoles sus observaciones a valores de cero y uno.
- **Variables Polinomiales:** Se transformaron las variables que poseen más de dos categorías, que son “Programa”, “Estado civil”, “Cargo” e “Ingreso mensual promedio”. Para este caso se crearon tantas variables como categorías menos uno y se renombran sus observaciones tomando valores de cero y uno.

En el caso de las variables numéricas, es necesario también tomar una medida correctiva con el fin de que los datos no violen los supuestos establecidos de los modelos, es por esto que es recomendable estandarizar los variables ya que permite llevar a una misma escala los datos independientemente de su unidad de medida. Dentro de este proyecto, existen tres variables numéricas las cuales fueron estandarizadas de la siguiente manera:

- **Estandarización de variables:** Se transformaron las tres variables numéricas de la base de datos objetivo que son: “Edad”, “Promedio general” y “Número de cursos tomados”. Esta transformación consiste en el proceso de centrar la variable restando la media de esta a cada una de sus observaciones y a su vez reduciendo la variable dividiendo el resultado de la resta por su desviación típica respectiva. Por la tanto, la variable centrada tiene una media igual a cero y la desviación típica de la variable reducida es igual a uno.

2.4 Modelización

Uno de los objetivos de este proyecto de materia integradora es la construcción de modelos predictivos para la detección de la deserción de los estudiantes de las carreras de postgrado de la IES. Bajo este enfoque, se escogieron cuatro modelos de clasificación: Potenciación del Gradiente (GBM), Bosques Aleatorios (RF), Redes Neuronales (NNET) y Regresión Logística (GLM), los cuales fueron definidos de manera formal en el Capítulo 1.

Para este análisis de los modelos de clasificación se dividió la base de datos de manera aleatoria en dos conjuntos de datos. La primera está conformada por el 75% de los datos que se usarán para el entrenamiento de los modelos y el 25% de los datos restantes son usados para la validación de estos y así, verificar su precisión. Este método es conocido como la técnica Training-Test.

Adicionalmente, usando la técnica de validación cruzada se realizaron tres diferentes particiones aleatorias, generando así tres bases de entrenamiento y 3 bases para validación a partir de la base original. Esta técnica es usada para demostrar que los resultados de la predicción son independientes de la división aleatoria del conjunto de datos de entrenamiento y validación. Una vez realizado las evaluaciones con cada conjunto de datos se toma la media aritmética de las mismas.

Luego de haber realizado las particiones de forma conjunta con la validación cruzada se obtuvieron los siguientes parámetros:

Tabla 3.1 Parámetros de los modelos seleccionados

Modelo	**Parámetros**
GBM	Número de árboles: 50 Máximo de nodos por árbol: 2 Tasa de aprendizaje: 0.1 Número mínimo de observaciones en los nodos terminales: 10
RF	Número de árboles: 500 Número de variables seleccionadas aleatoriamente en cada rama: 15
NNET	Número de unidades en la capa oculta: 1 Parámetro de regularización: 0.1
GLM	Número de parámetros: 30

2.5 Interpretación y evaluación

En esta última etapa de la metodología KDD, se evalúa el desenvolvimiento de cada uno de los modelos de clasificación escogidos en la etapa anterior. El desempeño de los modelos puede depender de muchos factores, tales como la manipulación realizado a los datos o la elección de las variables. Por consecuente, es importante implementar técnicas para evaluar el desempeño de cada modelo usado y así poder obtener un sustento objetivo del patrón identificado.

Bajo este enfoque, la literatura menciona diferentes métricas para cuantificar el desenvolvimiento del modelo predictivo, para este proyecto se usó la Matriz de Confusión en conjunto con sus indicadores más comunes que son: Precisión, Especificidad y Sensibilidad; considerando un umbral de discriminación o punto de corte de 0.5 para la clasificación de cada clase.

Adicionalmente, se implementó el gráfico de la Curva ROC con el fin de facilitar la interpretación de los patrones de los modelos seleccionado y a su vez, se calculó el área bajo la curva o AUC para obtener de forma métrica y precisa quien tuvo el mejor desempeño. Todos estos resultados serán mostrados y analizados con mayor detalle en el Capítulo 3

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

3.1 Análisis de situación actual

3.1.1 Análisis estadístico univariado

Se realizó un análisis estadístico univariado de las variables escogidas dentro del Capítulo 2, con el propósito de representar, describir y analizar la base de datos objetivo haciendo uso de los métodos estadísticos y gráficos que presenten y resuman la información obtenida. A continuación, se mostrarán los resultados de cada una de las variables sujetas a investigación:

- **Desertor:** La primera variable de estudio pertenece a “Desertor”, que representa la variable dependiente dentro del proyecto. Se puede observar que, dentro de la base de datos objetivo, el 9% corresponde a los estudiantes que desertaron su programa de estudios, mientras el 91% aproximadamente no lo hizo.

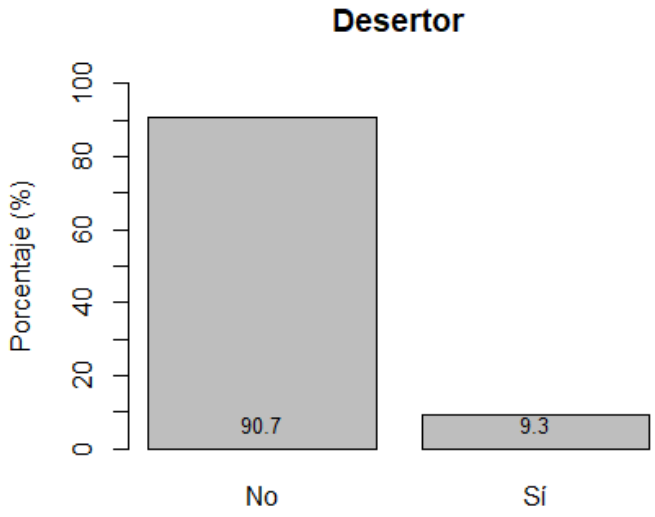

Figura 3.1 Gráfico de Barras - Desertor

- **Programa:** La variable “Programa” indica que el 17% de los estudiantes pertenecen al programa de estudios de “Diseño”, el 14% de los estudiantes pertenecen a Administración y otro 14% pertenecen a Automatización. De la misma manera se tiene que el 12% corresponden a Finanzas y un 11% del total les corresponde a los programas de Agronegocios y Proyectos respectivamente. Finalmente, los programas de Economía, Estadística y Logística tienen una proporción menor al 10%.

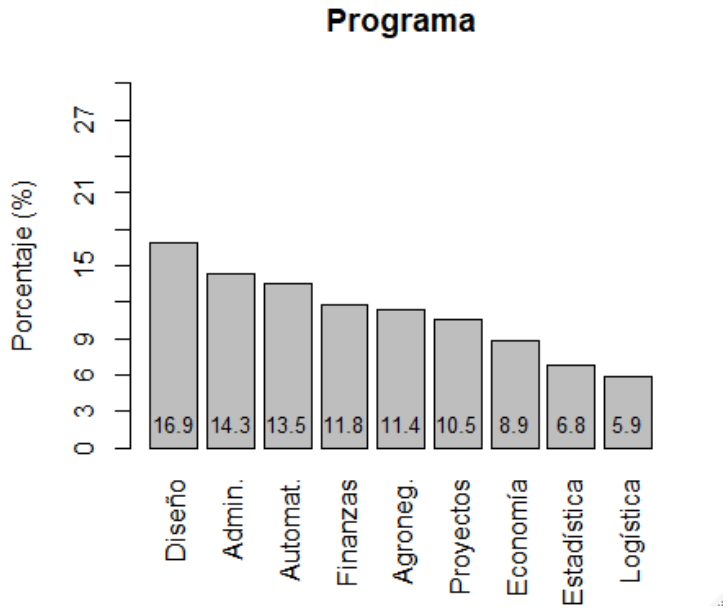

Figura 3.2 Gráfico de Barras - Programa

- **Ciudad de nacimiento:** Dentro de la base de datos se encontró que el 54% de los estudiantes pertenecen a la ciudad de Guayaquil, el 46% restante corresponde a 44 ciudades diferentes del país.

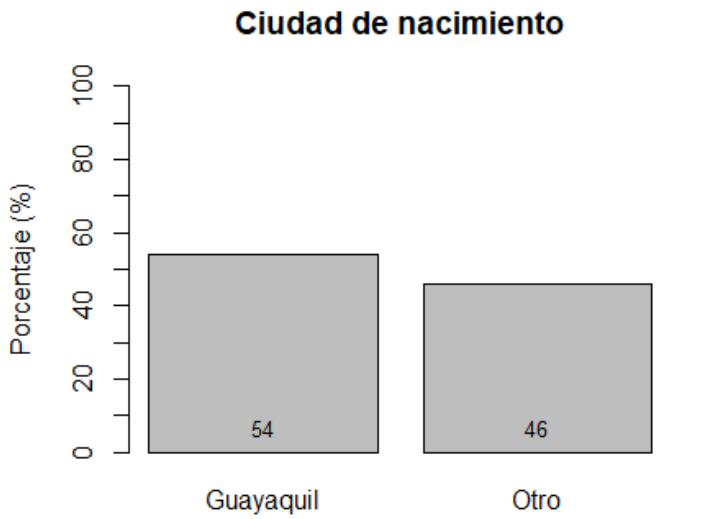

Figura 3.3 Gráfico de Barras - Ciudad de nacimiento

- **Sexo:** Se tiene dentro de esta variable que el 59.1% de los estudiantes desertores son de sexo masculino, en tanto el 40.9% pertenecen al sexo femenino.

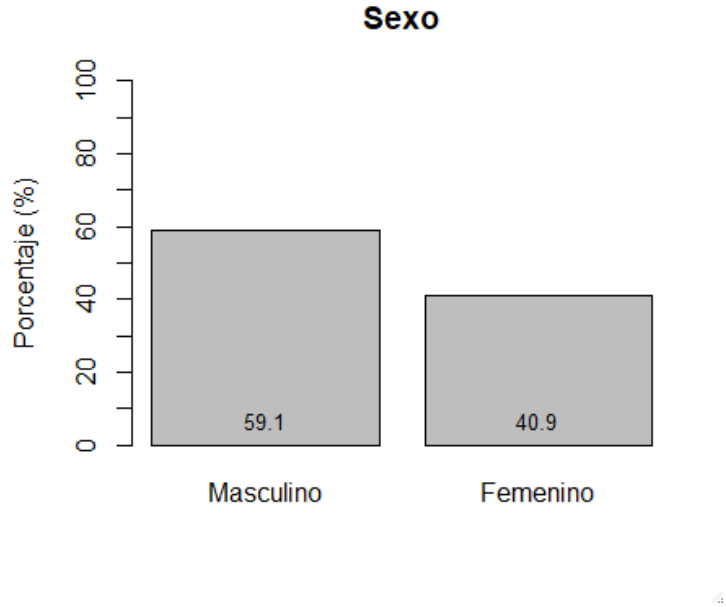

Figura 3.4 Gráfico de Barras - Sexo

- **Etnia:** Se observa que el 84.8% de los estudiantes que se encuentra en la base de datos objetivo se identifican como mestizos, el 7.2% se identifican como blancos, el 3.4% se identifican como montubios y el 1.7% se identifican como afroecuatorianos. Las demás categorías como indígena, mulato, negro y otros les corresponde menos del 1% del total de los estudiantes.

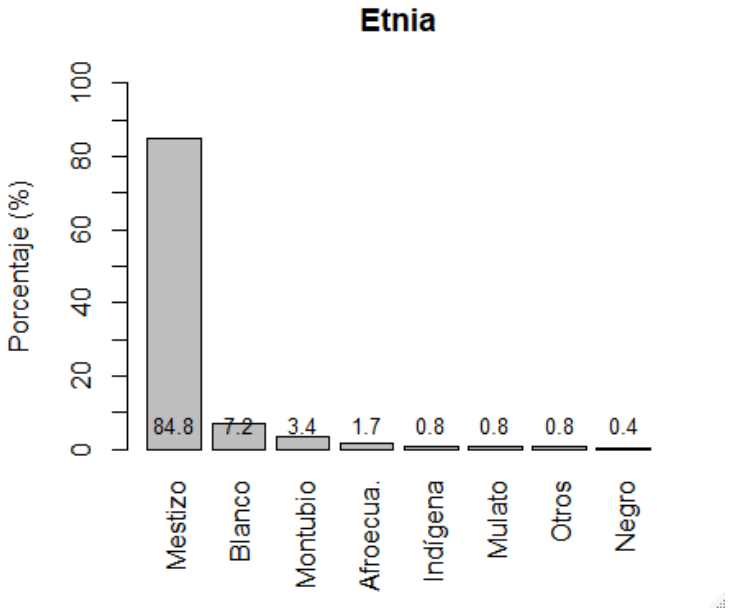

Figura 3.5 Gráfico de Barras – Etnia

- **Estado Civil:** En la base de datos objetivo se identificó que el 57% son estudiantes solteros, el 38.8% corresponde a estudiantes casados y el 4.2% de los estudiantes se encuentran divorciados.

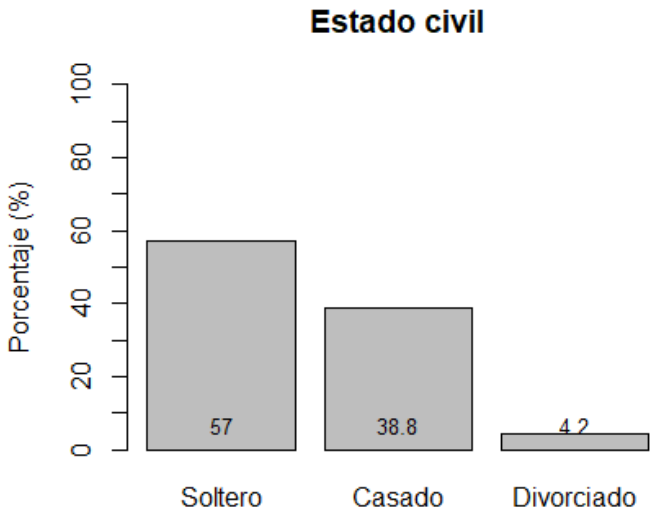

Figura 3.6 Gráfico de Barras – Estado civil

- **Cargo:** En el análisis descriptivo se evidencia que el 23.6% los estudiantes tienen un cargo de jefe, el 21.5% se encuentran en la categoría otros, el 16% ejercen un puesto como ingeniero, el 14.3% se desempeñan en un cargo de analista. Las categorías diseño gráfico, asistente, supervisor y gerente poseen una proporción menor al 10% del total cada uno.

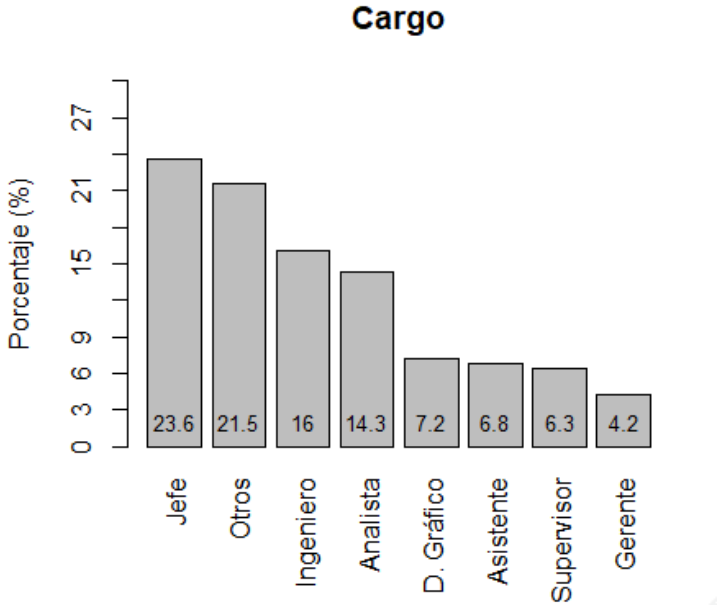

Figura 3.7 Gráfico de Barras - Cargo

- **Ingreso mensual promedio:** Se puede evidenciar por medio de esta variable que el 35% de los estudiantes de la Unidad de Postgrado tienen un ingreso mensual promedio de $1000 - $1500. Luego, se puede observar que el 27.4% de los estudiantes tienen un ingreso mensual promedio de $500 - $1000, le sigue el 11.4% de los estudiantes tienen un ingreso mensual promedio de $1500 - $2000. Los demás niveles tienen una representación menor al 10% cada uno.

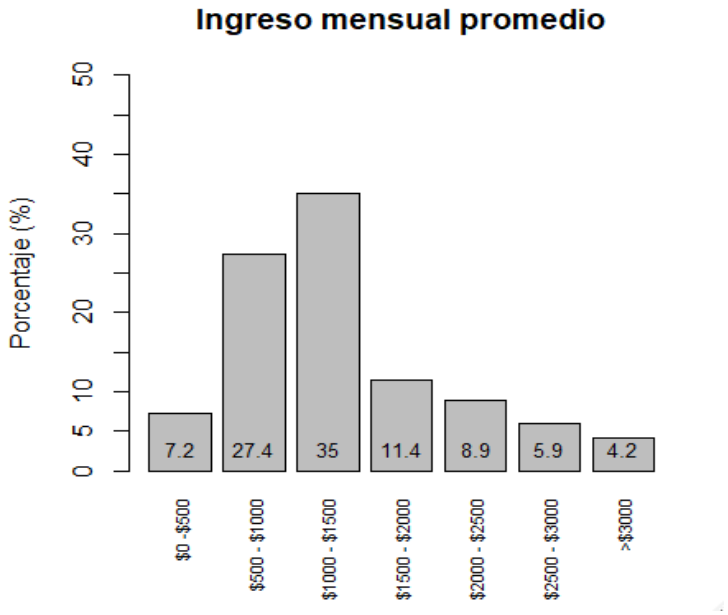

Figura 3.8 Gráfico de Barras – Ingreso mensual promedio

- **Edad:** En esta variable se tiene que en promedio los estudiantes tienen 33.19 años. Dentro del primer cuartil se observa que la edad mínima es 25 años hasta 29 años. Luego, en el cuartil dos, que es la mediana, muestra un valor de 32 años. Para el último 25% de los datos, se tiene que el tercer cuartil es 36 hasta el máximo que es 58. En las medidas de dispersión se tiene que la varianza muestra un valor de 30.92 y la desviación estándar tiene un valor de 5.56.

Tabla 4.1 Estadística descriptiva - Edad

Mínimo	25

Primer cuartil	29
Mediana	32
Media	33.19
Tercer cuartil	36
Máximo	58
Varianza	30.92
Desviación estándar	5.56
Desviación estándar de la media	0.36

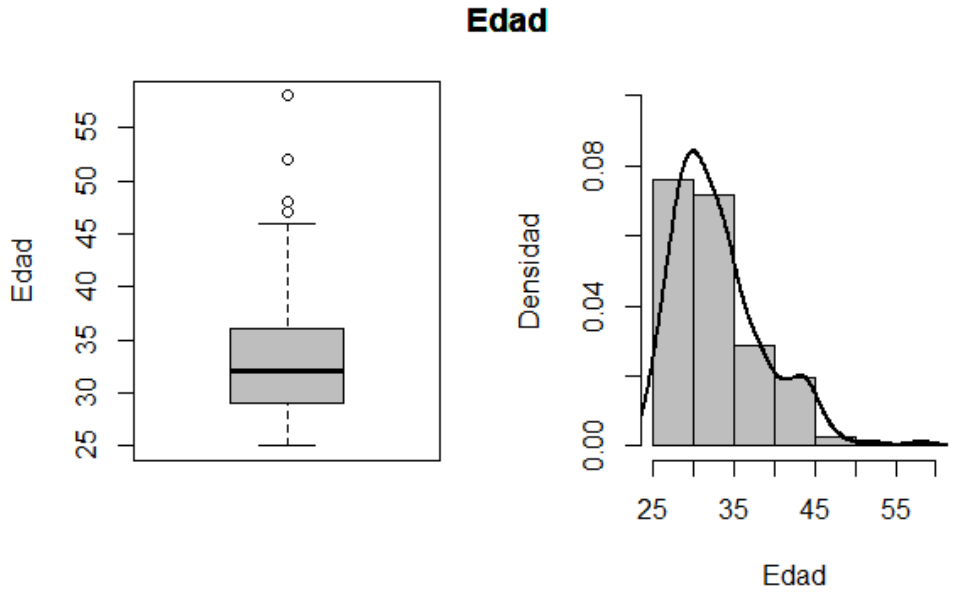

Figura 3.9 Diagrama de cajas e histograma - Edad

- **Promedio general:** Se puede evidenciar que los estudiantes de la unidad de postgrado tienen 8.49 como media de promedio general. Dentro del primer 25% de los datos se tiene 0 de promedio general mínimo hasta 8.34 de promedio general correspondiente al valor del primer cuartil. En el segundo cuartil se tiene de promedio general 8.65 y en el tercer cuartil se tiene 8.89. Como valor máximo del promedio general se tiene 9.79. En las medidas de dispersión, se observa 1.39 en la varianza y 1.17 en la desviación estándar.

Tabla 5.2 Estadística descriptiva – Promedio General

Mínimo	0
Primer cuartil	8.34
Mediana	8.65
Media	8.49
Tercer cuartil	8.89
Máximo	9.79
Varianza	1.39
Desviación estándar	1.17
Desviación estándar de la media	0.08

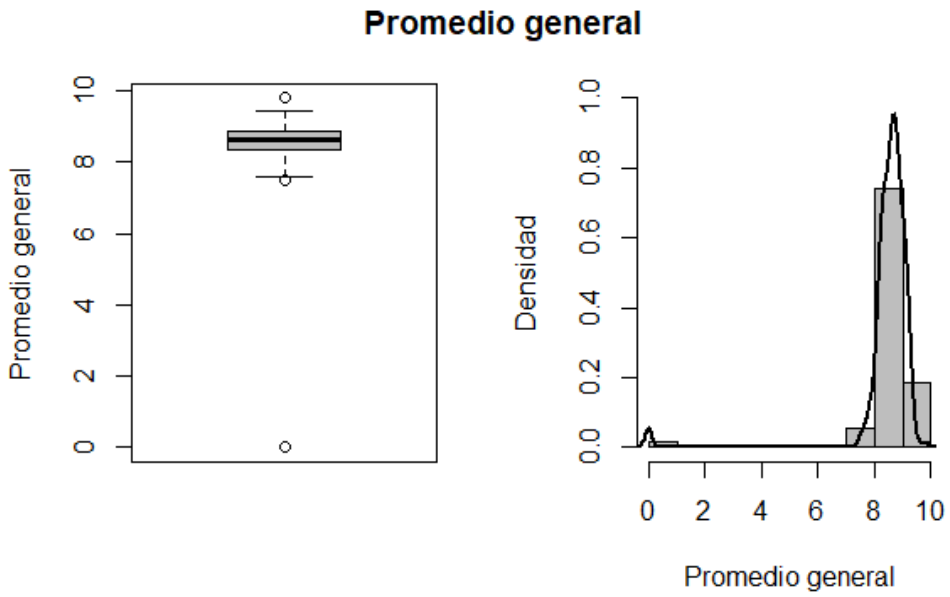

Figura 3.10 Diagrama de cajas e histograma – Promedio general

- **Número de cursos tomados:** En esta variable se puede observar que en promedio los estudiantes han tomado 15.19 materias. Como valor mínimo se tiene 0 materias tomadas y como valor máximo se tiene 26 materias tomadas. En los cuartiles se puede observar que el primero tiene el valor de 14, en el segundo 16 y en el tercero 17. En cuanto a las medidas de dispersión, la varianza tiene un valor de 24.47 y la desviación estándar 4.95.

Tabla 6.3 Estadística descriptiva – Número de cursos tomados

Mínimo	0
Primer cuartil	14
Mediana	16
Media	15.19
Tercer cuartil	17
Máximo	26
Varianza	24.47
Desviación estándar	4.95
Desviación estándar de la media	0.32

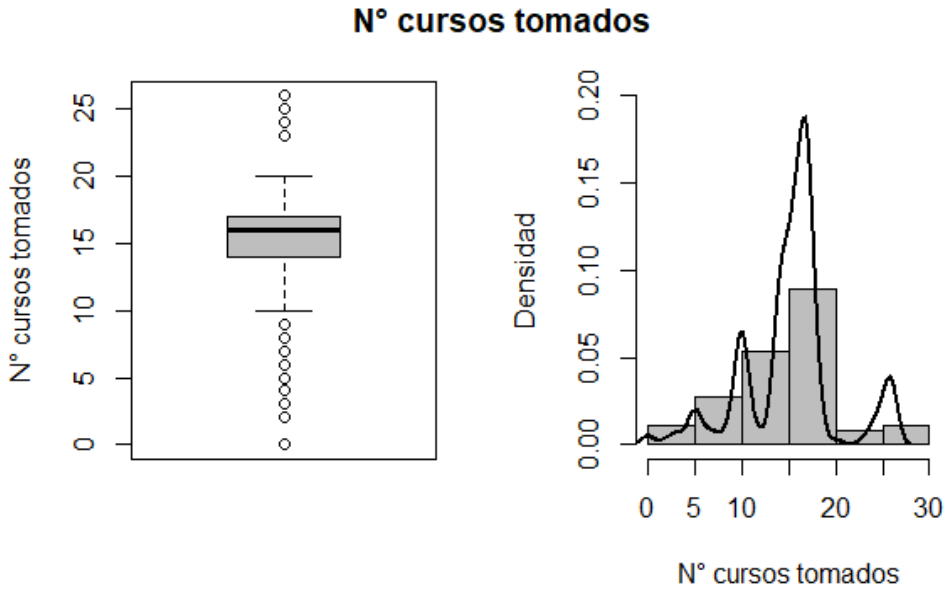

Figura 3.11 Diagrama de cajas e histograma – Número de cursos tomados

- **Tasa de eficiencia terminal:** Se calculó la tasa de eficiencia terminal, la cual dio como resultado 52.32%, esto quiere decir que los estudiantes que ingresaron durante el primer término del 2017 sólo el 52.32% pudo egresar con éxito dentro del tiempo mínimo establecido.

$$ET = 52.32\%$$

3.1.2 Análisis estadístico bivariado

- **Desertores por Programa**: Dentro de la variable “Programa” se puede observar que la categoría que más desertores tuvo fue el programa de Automatización con un 32%. Luego, se tiene que el programa de Estadística tuvo un 18% del total de desertores. Después, se puede identificar que sigue el programa de Logística con un 14% de los desertores. Los programas de Finanzas y Diseño tienen poseen el mismo porcentaje de 9% de desertores. Finalmente, se tiene que las carreras de Proyectos, Economía, Agronegocios y Administración cuentan con un porcentaje del 5%.

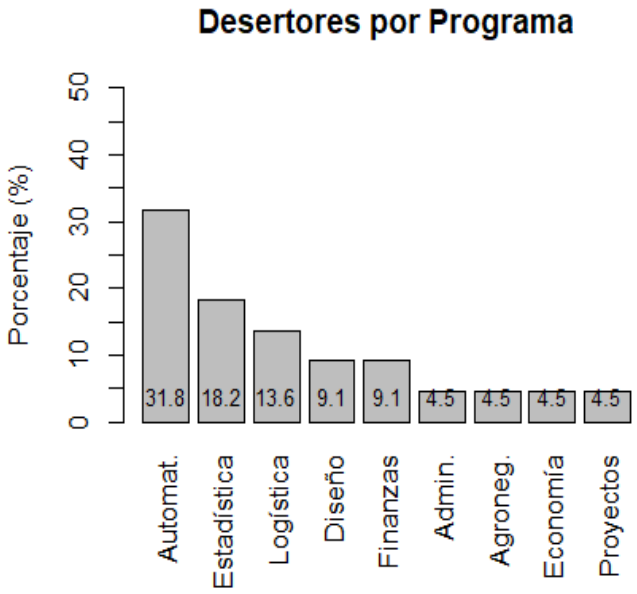

Figura 3.12 Gráfico de Barras – Desertores por Programa

- **Desertores por Ciudad de nacimiento:** En esta variable se puede identificar que el 59.1% de los desertores provienen de la ciudad de Guayaquil, el 40.9% restante provienen de las siguientes ciudades: Loja, Quevedo, Portoviejo, Chone, Quito, Bolívar y Esmeraldas. Es relevante mencionar que de las 45 ciudades que muestra la base de datos, sólo ocho tuvieron desertores, las cuales son mostradas en el gráfico.

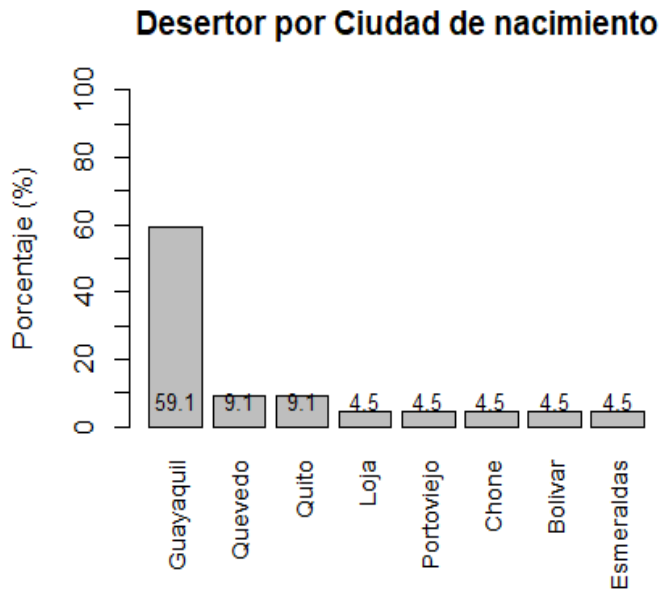

Figura 3.13 Gráfico de Barras – Desertores por Ciudad de nacimiento

- **Desertores por Sexo:** La variable “Sexo” muestra que el 72,7% de los desertores son de sexo masculino y por complemento el 27,3% pertenecen al sexo femenino.

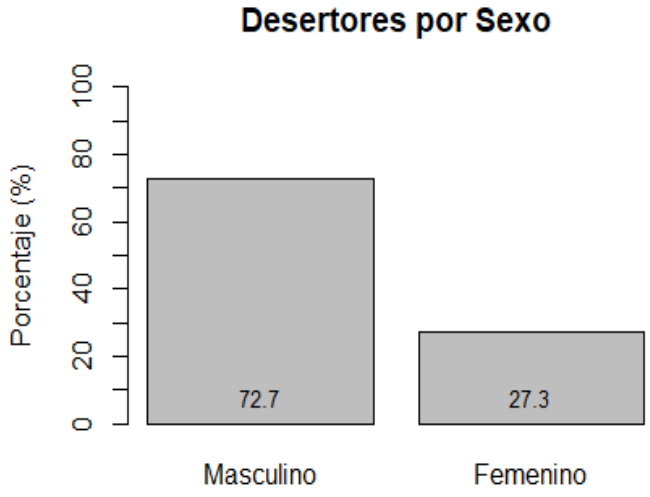

Figura 3.14 Gráfico de Barras – Desertores por Sexo

- **Desertores por Etnia:** Se puede identificar en la variable “Etnia” que el 81.8% de los desertores se identifican como Mestizos. Luego, se puede observar el 9.1% se identifica como Montubio. Finalmente, se tiene que el 4.5% se identifica como Blanco y otro 4.55% se identifica en la categoría Otros. De manera intuitiva, se puede observar que las categorías Blanco, Indígena, Mulato y Negro no constan con desertores.

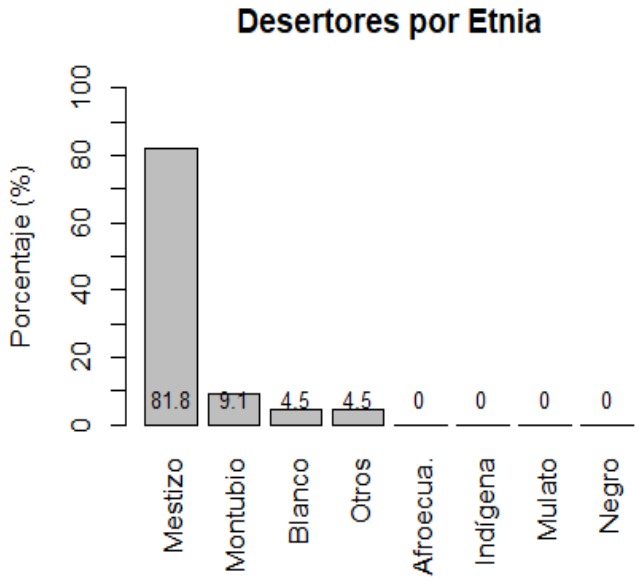

Figura 3.15 Gráfico de Barras – Desertores por Etnia

- **Desertores por Estado civil:** Se puede observar que los desertores que se encuentran en la categoría de soltero tienen un porcentaje de 54.5% y los que se constan en la categoría de casado poseen el 45.5% restante, por lo que se puede identificar que no existen desertores dentro de la clase divorciado.

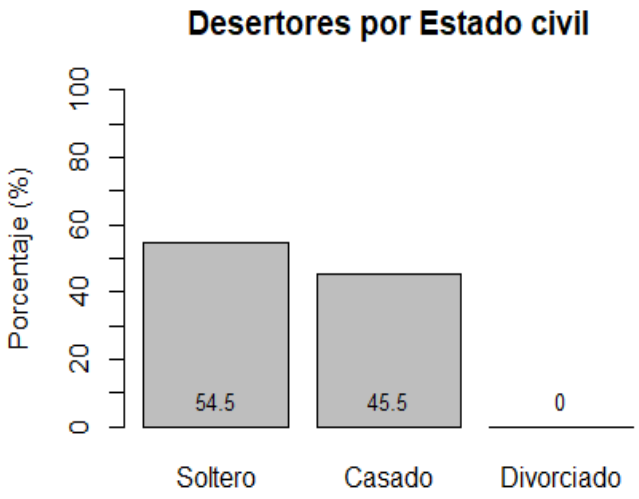

Figura 3.16 Gráfico de Barras – Desertores por Estado civil

- **Desertores por Cargo:** La variable ocupación demuestra que los desertores que se encuentran en la clase Otros representan el 31.8%. Después, se puede observar que el 27.3% de los desertores constan en la categoría Ingeniero. Luego, los desertores que tienen un cargo como Jefe forman parte del 18.2%. Las categorías Asistente, Analista y Supervisor tienen una frecuencia menor al 10% cada uno. Además, se puede distinguir que no existen desertores con cargos de tipo Gerente y Diseñador Gráfico.

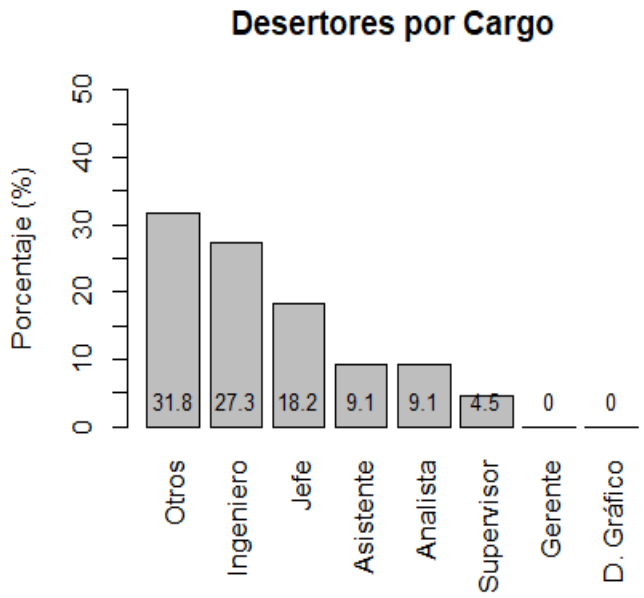

Figura 3.17 Gráfico de Barras – Desertores por Cargo

- **Desertores por Ingreso promedio mensual:** Los desertores que tienen un ingreso mensual promedio de $1000 a $1500 representan el 45%. Luego, los desertores que tienen un ingreso promedio mensual promedio de $500 a $1000 figuran el 36% del total. Después, se tiene que los desertores que tienen un ingreso promedio mensual de $1500 a $2000 representan el 14%. Finalmente, se tiene que los desertores con un ingreso promedio de $2000-$2500 forman parte del 5% del total. También se puede destacar que no existen desertores en los estudiantes que tiene como ingreso promedio mensual de 0 a $500 e ingresos promedios mensuales superiores a $2500.

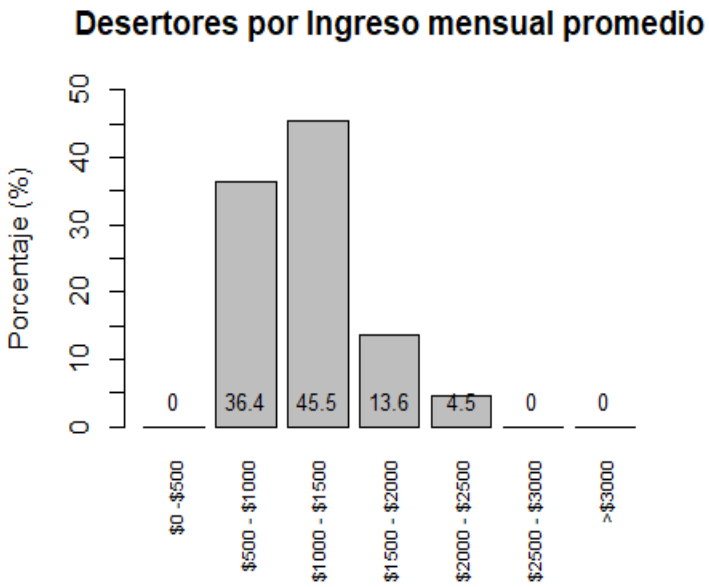

Figura 3.18 Gráfico de Barras – Desertores por Ingreso mensual promedio

- **Correlación de las variables numéricas:** Se puede observar que ninguna de las tres variables numéricas están correlacionadas entre sí. La variable Edad con la variable Promedio general tienen un coeficiente de correlación de -0.005, mientras que la variable Edad y la variable Número de cursos tomados tienen una correlación de 0.012. También se evidencia que las variables Promedio general y Número de cursos tomados tienen una correlación de 0.297. Por lo tanto, estas variables miden diferentes características.

Tabla 7.4 Matriz de correlación

	Edad	**Promedio general**	**N° cursos tomados**
Edad	1	-0.005	0.012
Promedio general	-0.005	1	0.297
N° cursos tomados	0.012	0.297	1

- **Tasa de eficiencia terminal por Carrera:** Se calculó la tasa de eficiencia terminal por cada una de las carreras que se tomaron en cuenta para este estudio. La carrera de Finanzas obtuvo una tasa del 92.9%, siendo la carrera con más alumnos egresados. Después, se tiene a la carrera de Diseño con el 82.5%, le sigue la carrera de Logística con el 78.6% y luego, la carrera de

Proyectos con el 74.1%, convirtiéndose en las carreras que tienen la mayor proporción de egresados. Se observa también que las carreras de Automatización y Estadística cuentan con un poco más del 50% de egresados con respecto a los alumnos que ingresaron durante el primer término del 2017. La última carrera que cuenta con una proporción de estudiantes egresados es la carrera de Agronegocios con un 24%. Finalmente, como se observa en la gráfica, las carreras de Administración y Economía no cuentan con alumnos que hayan finalizado el período de estudio.

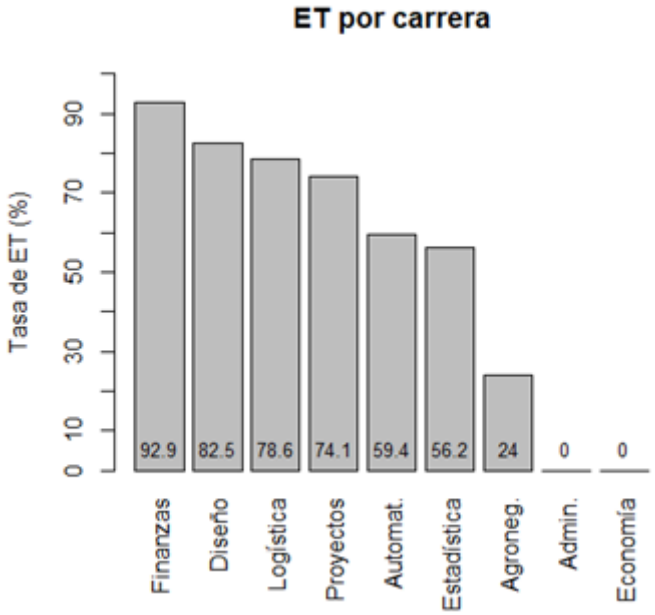

Figura 3.19 Gráfico de barras – ET por Carrera

3.2 Interpretación y evaluación de los modelos

3.2.1 Modelos estadísticos

Se realizó un análisis para los cuatro modelos de clasificación, dónde sus parámetros fueron definidos dentro del Capítulo 2 con el propósito de predecir a partir de las observaciones obtenidas en las variables independientes. A continuación, se mostrará la validación de cada uno de los modelos:

- **Potenciación del gradiente estocástico:** Este modelo presenta, con una validación cruzada de 3 capas, muestra una exactitud del 94.83% con una sensibilidad del 96.23% y una especificidad del 80%, por lo tanto, el modelo sí muestra una concordancia entre los valores de la predicción con los reales. Por otro lado, se observa que la curva ROC del modelo muestra un área bajo la curva de 0.94 y de acuerdo a la literatura previamente mencionada, se considera al modelo con una alta exactitud para realizar la predicción.

Tabla 8.5 Matriz de confusión – GBM

		Predicción	
		No	Sí
Real	No	51	2
	Sí	1	4

Tabla 9.6 Indicadores de desempeño – GBM

Indicadores	Porcentaje
Exactitud	94.83%
Sensibilidad	96.23%
Especificidad	80.00%

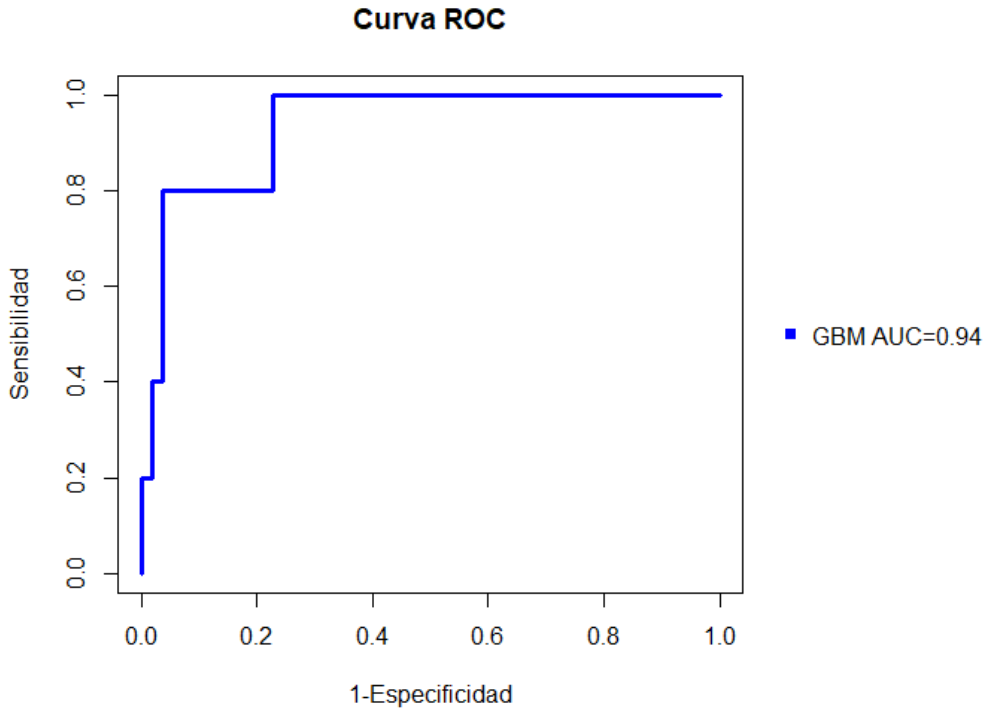

Figura 3.20 Curva ROC – GBM

Dentro de las variables importantes del modelo GBM se puede evidenciar que el modelo tomo en cuenta solo ocho variables de las 29 en total. La variable que mayor contribuye significativamente para el modelo es “Cursos tomados” y luego le sigue la variable “Promedio general” en menor medida. Las otras seis variables entregan una importancia pequeña para el modelo que son: “Programa Automatización”, “Edad”, “Ingreso mensual prom. $500-$1000”, “Cargo Jefe”, “Cargo Ingeniero” y “Cargo Analista”.

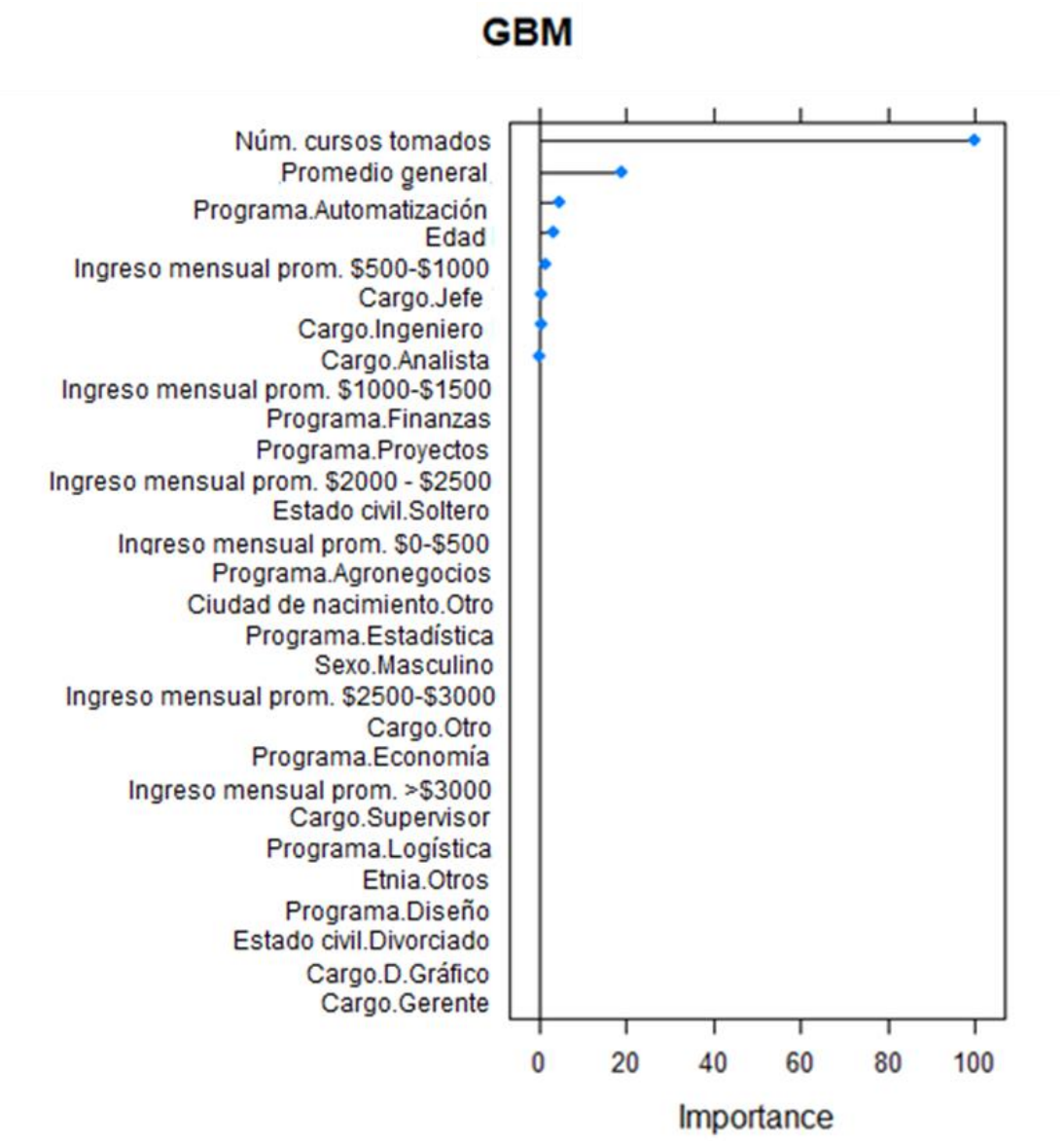

Figura 3.21 Variables importantes – GBM

- **Bosques aleatorios:** El modelo de bosques aleatorios presentó una exactitud del 96.55%, es decir, que el modelo pudo predecir casi con exactitud las observaciones que se encontraban en el conjunto de datos para la validación. Los indicadores de sensibilidad y especificidad se mostraron porcentajes del 98.11% y 80% respectivamente. Se puede decir que el modelo realiza las predicciones con un excelente desempeño y concordancia con respecto a las observaciones reales. En el gráfico de la curva ROC, se observa que el área bajo la curva tiene un valor de 0.98, ratificando de esta manera el excelente desempeño del modelo para clasificar.

Tabla 10.7 Matriz de confusión – RF

		Predicción	
		No	Sí
Real	No	52	1
	Sí	1	4

Tabla 11.8 Indicadores de desempeño – RF

Indicadores	Porcentaje
Exactitud	96.55%
Sensibilidad	98.11%
Especificidad	80.00%

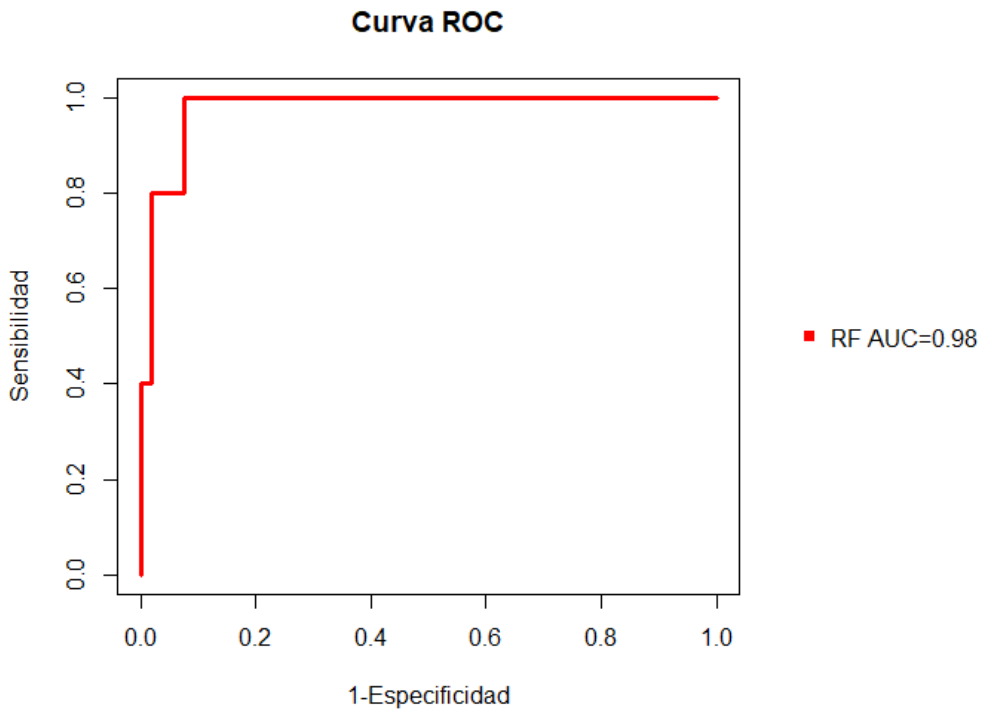

Figura 3.22 Curva ROC – RF

El modelo RF muestra que la variable que mayor contribuye a su desempeño es la variable “Número de cursos tomados”, después le sigue la variable “Promedio general” y luego la variable “Edad”. En base a lo mencionado, se puede decir que el modelo toma en cuenta las tres variables numéricas de la base de datos como las más importantes. Posteriormente, se observa que el modelo toma 24 variables con una importancia pequeña y finalmente, el modelo no considera dos variables que son “Ingreso mensual prom. $2500-$3000” e “Ingreso mensual prom. $2000-$2500”

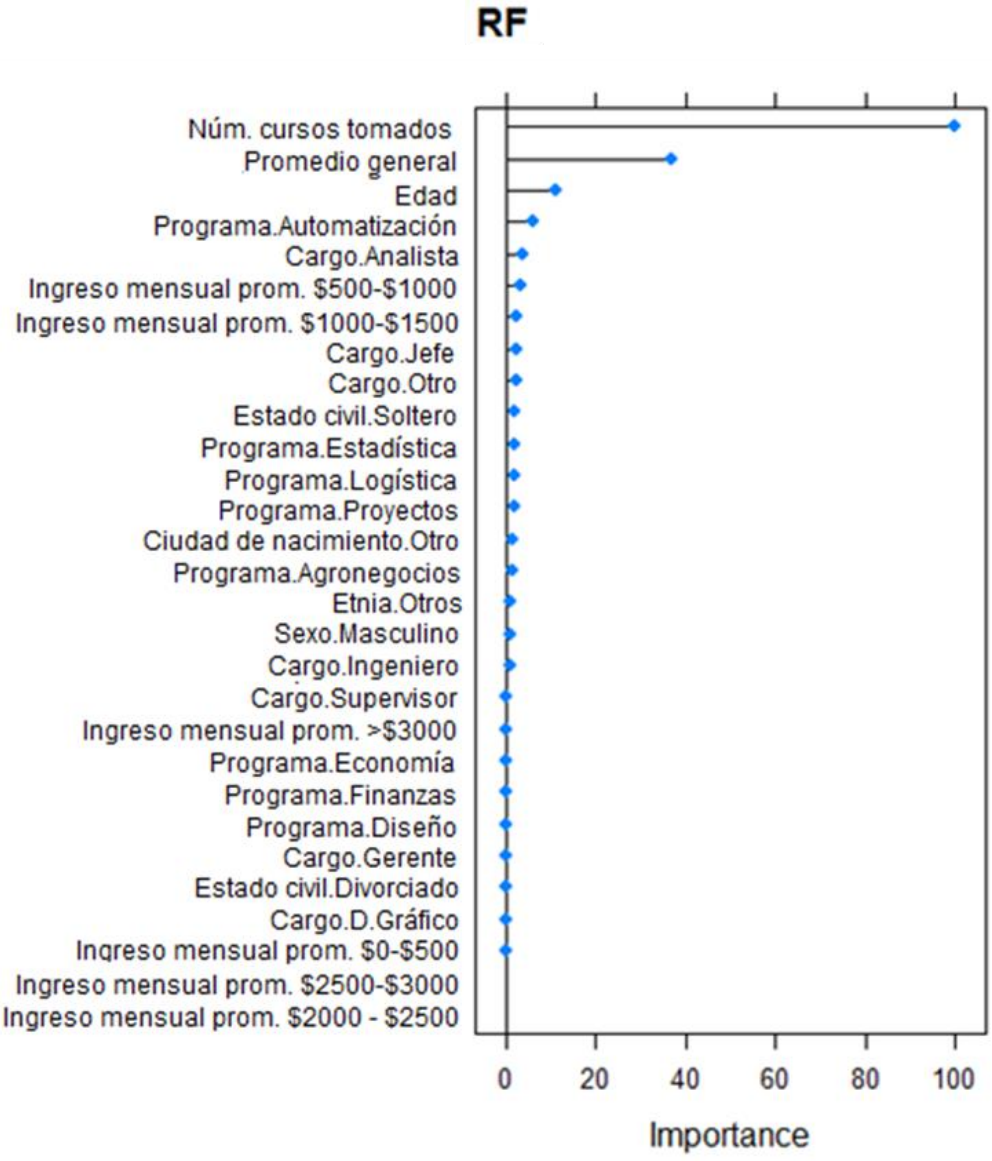

Figura 3.23 Variables importantes – RF

- **Redes neuronales:** Dentro de este modelo se evidencia una exactitud del 96.55%, la sensibilidad muestra un porcentaje de 98.11% y la especificidad muestra un 80%. Estos indicadores demuestran que las predicciones realizadas por el modelo tienen una alta precisión con las observaciones reales. Esto también se puede evidenciar en el gráfico de la curva ROC, dónde se muestra un área bajo la curva de 0.98, mostrando que el modelo NNET tiene un alto poder de clasificación.

Tabla 12.9 Matriz de confusión – NNET

Real \ Predicción	No	Sí
No	52	1
Sí	1	4

Tabla 13.10 Indicadores de desempeño – NNET

Indicadores	Porcentaje
Exactitud	96.55%
Sensibilidad	98.11%
Especificidad	80.00%

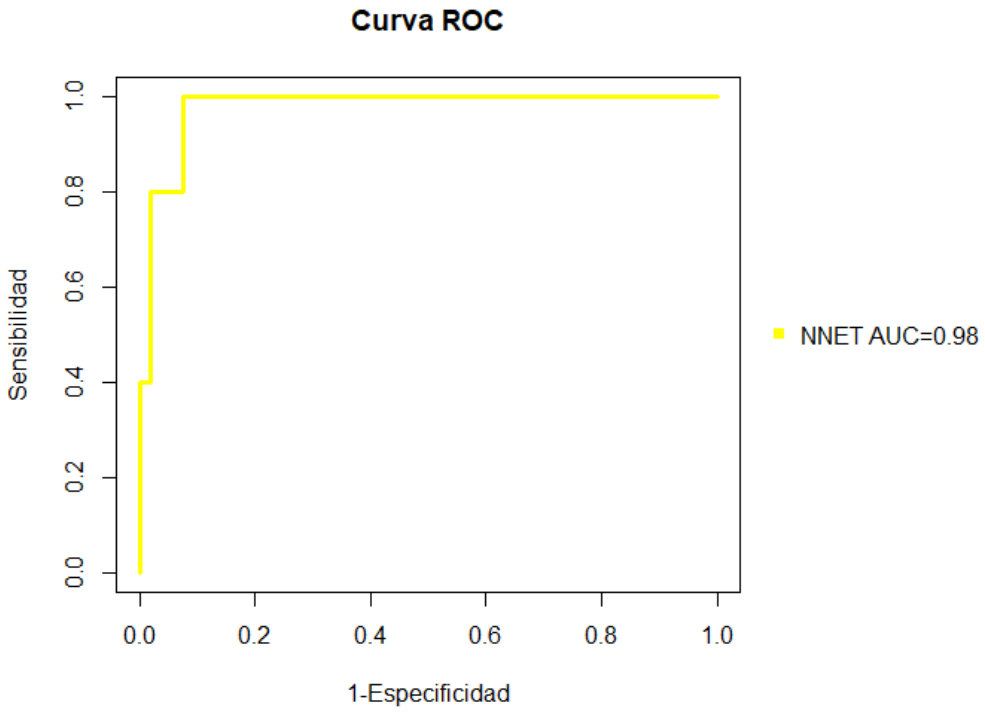

Figura 3.24 Curva ROC – NNET

En la Figura 3.24 se puede observar que el modelo NNET usa casi todas las variables. Las primeras cinco variables “Núm. Cursos tomados”, “Promedio general”, “Ingreso mensual prom. $500-$1000”, “Programa.Agronegocios” e “Ingreso mensual prom. >$3000” son las que más contribuyen al modelo de manera significativa. Después, se observa un segundo grupo de siete variables que aportan al modelo, pero en menor medida; estas variable son: “Programa.Logística”, “Etnia.Otros”, “Estado civil.Divorciado”, “Ingreso mensual prom. $2000-$2500”, “Estado civil.Soltero”, “Programa.Proyectos” y “Programa.Economía”. Finalmente, las variables restantes aportan poco o nada al modelo NNET.

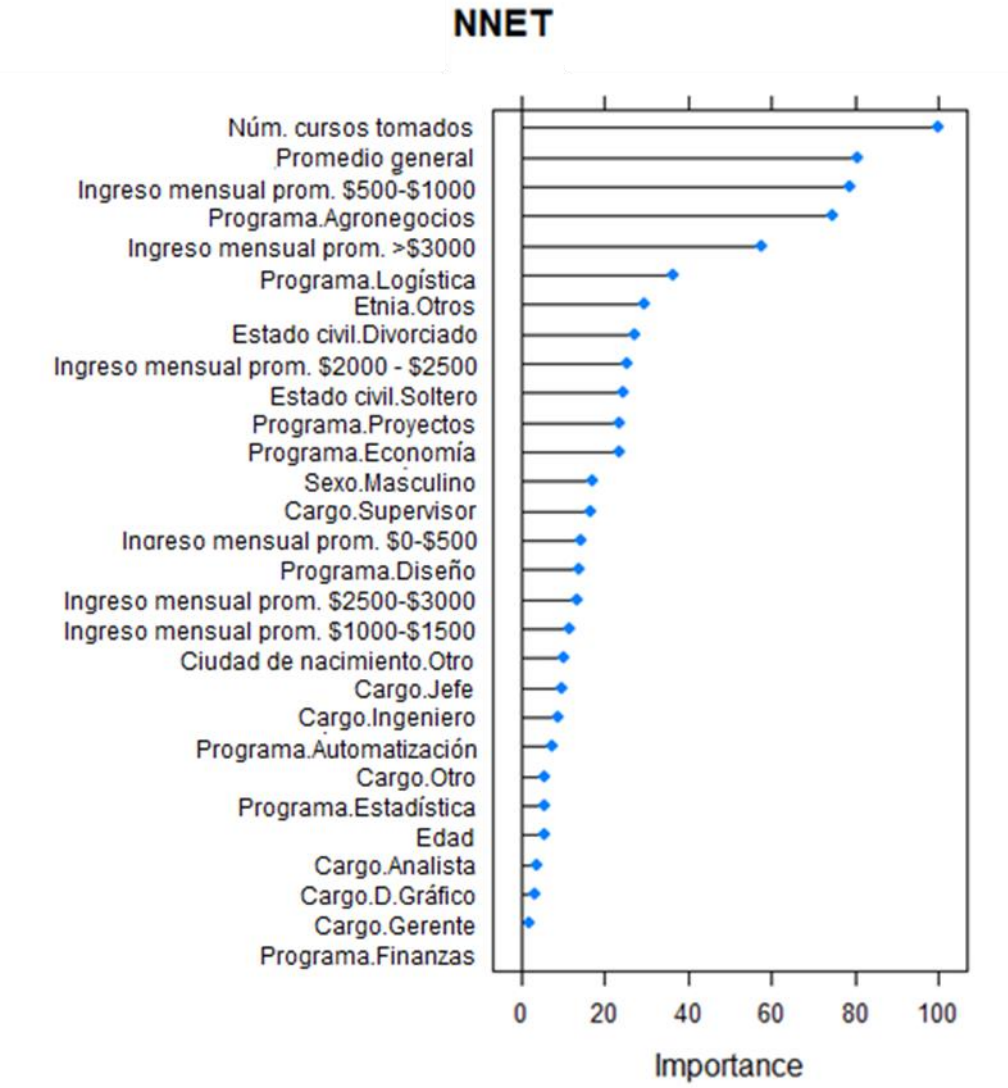

Figura 3.25 Variables importantes – NNET

- **Regresión logística:** El modelo de regresión logística muestra una exactitud de 86.21%. La sensibilidad muestra un porcentaje del 86.79% y la especificidad un 80%. Tanto los indicadores como la matriz de confusión demuestran un desempeño regular de este modelo, mostrando cierta preferencia de clasificar las observaciones como positivos. En cuanto a la curva ROC, se observa un área bajo la curva de 0.85, confirmando un poder de clasificación regular.

Tabla 14.11 Matriz de confusión – GLM

		Predicción	
		No	Sí
Real	No	46	7
	Sí	1	4

Tabla 15.12 Indicadores de desempeño – GLM

Indicadores	Porcentaje
Exactitud	86.21%
Sensibilidad	86.79%
Especificidad	80.00%

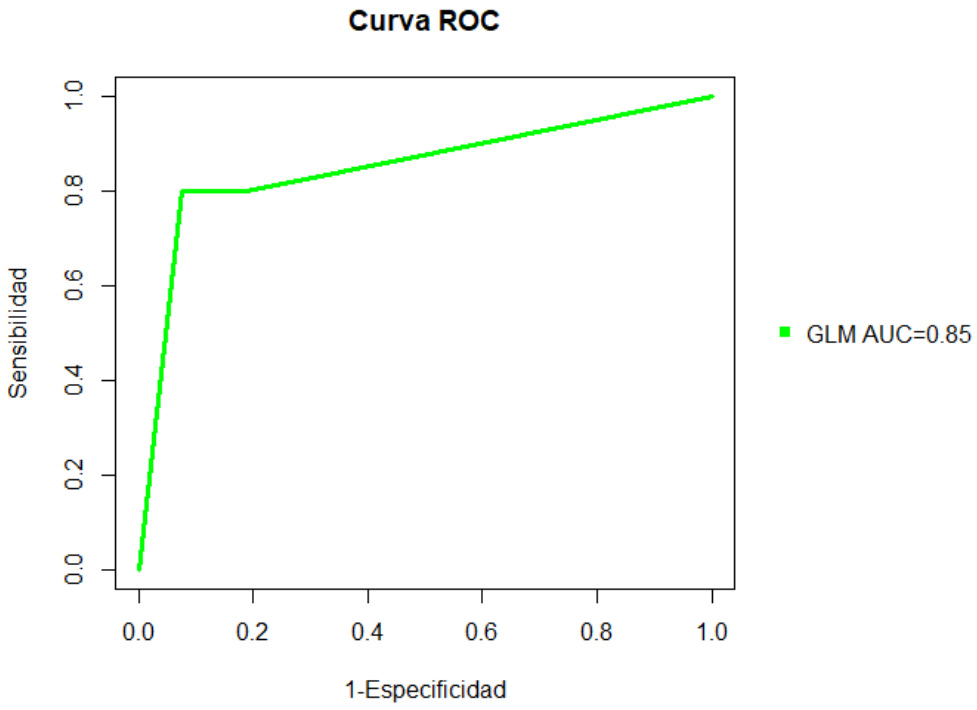

Figura 3.26 Curva ROC – GLM

En la Gráfica 3.26 se puede evidenciar que las cuatro variables “Promedio general”, “Sexo.Masculino”, “Núm. cursos tomados” y “Programa.Proyectos” aportan significativamente al modelo. Después, se observa que otras cuatro variables aportan al modelo en menor medida en comparación a las cuatro primeras, que son: “Programa.Agronegocios”, “Estado civil.Soltero”, “Programa.Automatización” e “Ingreso mensual prom. $500-$1000”. Finalmente, el resto de las 21 variables aportan poco o nada al modelo.

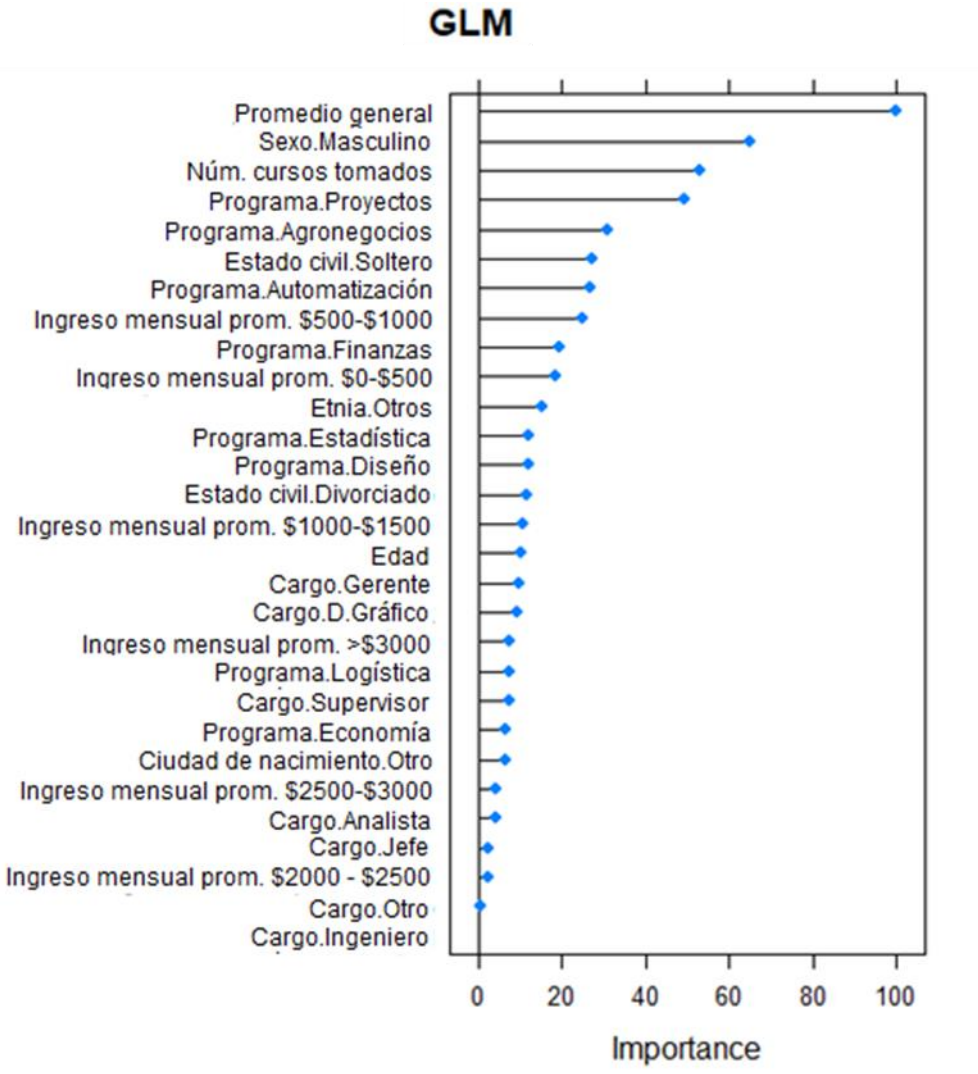

Figura 3.27 Curva ROC – GLM

3.2.2 Comparación de los modelos

Luego de haber realizado la reflexión sobre cada uno de los modelos, se realizó la comparación del desempeño de cada uno de ellos, donde se puede observar

que todos tuvieron buenos indicadores, los cuales presentaron una exactitud que va desde el 85% al 95%. Los modelos RF y NNET presentaron el mismo poder de clasificación, siendo estos los dos mejores presentando un 96.55% de exactitud cada uno, en segundo lugar, se encuentra el modelo GBM presentando una exactitud de 94.83% y en último lugar, el modelo GLM presentando un 86.21% de exactitud. También, se evidencia que con la sensibilidad se mantiene el mismo orden de los modelos. En cuanto a la especificidad, los cuatro modelos obtuvieron el 80%.

Tabla 16.13 Indicadores de desempeño

	Exactitud	**Sensibilidad**	**Especificidad**
GMB	94.83%	96.23%	80.00%
RF	96.55%	98.11%	80.00%
NNET	96.55%	98.11%	80.00%
GLM	86.21%	86.79%	80.00%

Se realizó una gráfica de curva ROC para los cuatro modelos y así facilitar la interpretación sobre el desempeño de cada uno, el mismo que ratifica el excelente poder de clasificación que tienen RF y NNET contando con un AUC de 0.98 cada uno. Luego, se puede observar que el modelo GBM cuenta con el AUC de 0.94 y finalmente, le sigue el modelo GLM con un AUC de 0.85.

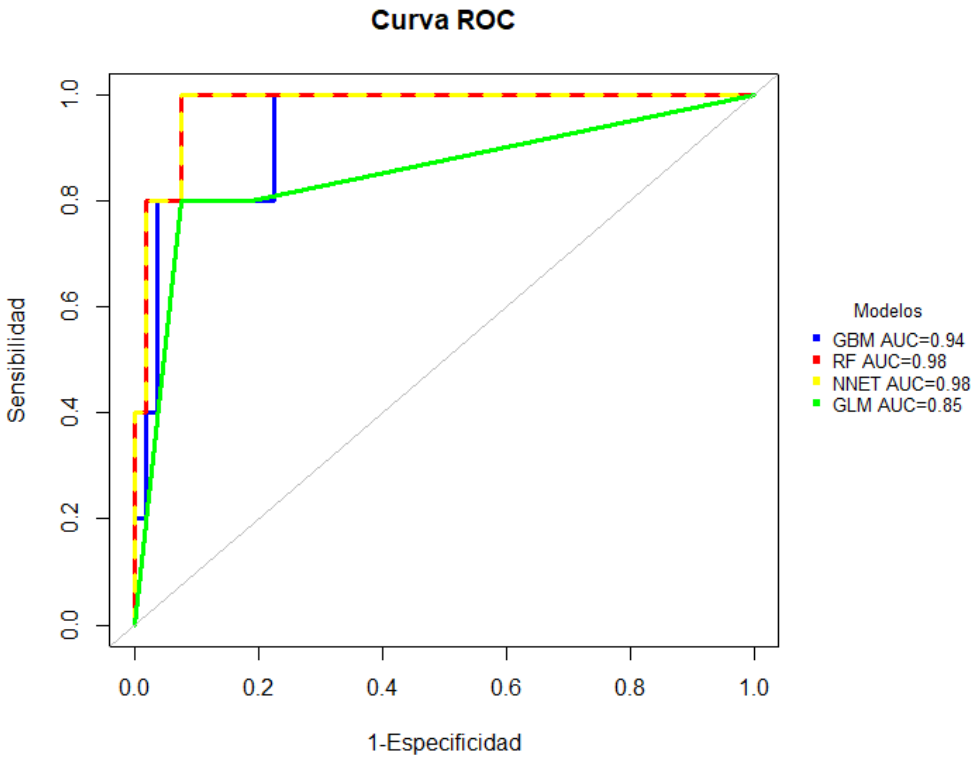

Figura 3.28 Curva ROC

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES.

4.1 Conclusiones

El principal objetivo de este proyecto integrador es el diseñar un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior; para lo cual, se realizó un análisis de situación actual para la selección de las variables, luego se construyeron los potenciales modelos predictivos para la detección de los posibles desertores y después, se mostró el mejor modelo de acuerdo a su desempeño para finalmente determinar el perfil del estudiante desertor de acuerdo a las variables escogidas, llegando así a las siguientes conclusiones:

- En el análisis de situación actual se tomaron las variables Programa, Ciudad de nacimiento, Sexo, Etnia, Estado Civil, Cargo, Ingreso mensual promedio, Edad, Promedio general, Número de cursos tomado y Estado académico del estudiante. La variable Estado académico del estudiante se la tomó en cuenta para determinar la tasa de eficiencia terminal, la cual dio como resultado un 52.32%, mostrando así que un poco de más de la mitad de los estudiantes que ingresaron dentro del primer semestre del 2017 pudieron egresar de sus estudios dentro del tiempo mínimo establecido. Las variables restantes fueron tomadas en cuenta como variables predictoras para el proceso de Modelización.
- Se usó la metodología KDD para desarrollar los cuatro modelos de clasificación escogidos que son: Potenciación del Gradiente Estocástico (GBM), Bosques Aleatorios (RF), Redes Neuronales (NNET) y Regresión Logística (GLM). Para proceder en el diseño de cada uno de ellos se dividió la base de datos en dos partes: una para entrenar representando el 75% de las observaciones y la segunda para contrastar el poder de clasificación de cada uno de los modelos usando el 25% restante. Las técnicas de validación que se tomaron en cuenta

son: matriz de confusión, indicadores de desempeño, curva ROC y su AUC. Se observó que los modelos RF y NNET presentaron la misma exactitud cada uno con un 96.55%; en segundo lugar, el modelo GBM muestra una exactitud de 94.83% y en último lugar, el modelo GLM muestra un 86.21% de exactitud.

- Se comparó el desenvolvimiento de cada uno de los modelos de manera cuantitativa por medio de las técnicas de validación, dando como resultado que los cuatro modelos tuvieron una precisión aproximada al 90%, de igual manera en su sensibilidad y su especificidad mostró en todos el 80%. Los modelos RF y NNET fueron los mejores presentando el mismo poder clasificación, el cual finalmente se escogió a RF como el mejor modelo por su sencillez en comparación a NNET.
- En general, se puede identificar que el perfil del desertor es una persona que proviene de la ciudad de Guayaquil, de sexo masculino, se identifica como mestizo, su estado civil es soltero, su cargo que desempeña en su área laboral es de tipo otros y su ingreso mensual promedio es de $1000 - $1500.

4.2 Recomendaciones

Acorde a los resultados presentado y a las conclusiones previamente mostradas, se presentan las siguientes recomendaciones:

- De acuerdo al análisis de situación actual, se recomienda a la unidad de postgrado incluir más variables que identifiquen al estudiante en general, así como información previa del mismo y no permitir que ellos dejen espacios sin completar en sus registros. También, actualizar de forma periódica la información contenida en los registros administrativos, ya que esta va cambiando según las necesidades de los estudiantes, por lo que es indispensable tener los datos ajustados a la realidad y así no afectar la toma de decisiones.
- Para trabajos futuros, tomar en cuenta otros modelos de clasificación como SVM (Máquinas de soporte vectorial) con el objetivo de ampliar la gama de modelos; así también investigar nuevas herramientas tecnológicas que faciliten el proceso de KDD.

- Investigar si existen técnicas estadísticas o indicadores que puedan desempatar el desempeño de los modelos RF y NNET y así tomar una decisión con base en información cuantitativa.
- Utilizar la información obtenida para aplicar políticas educacionales con el objetivo de disminuir la proporción de los desertores y lograr minimizar tanto el impacto económico como social.

BIBLIOGRAFÍA

Himmel, E. (2002). Modelo de análisis de la deserción estudiantil en la educación superior. Calidad en la Educación, (17), 91-108.

Tinto, V (1989). Definir la deserción: una cuestión de perspectiva. Revista de Educación Superior N° 71, ANUIES, México.

Spady, W. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. Interchange. Vol. 19, N° 1: 109-121

Durkheim, E. (1951). Suicide: A study in sociology (G. Simpson, Ed. J.A. Spaulding $ G Simpson, Trans.). New York: Free Press.

Tinto, V (1975). Dropout From Higher Education: A Theoretical Synthesis of Recent Research, Journal of Higher Education. N° 45: 89-125.

Nye, J. (1976). Independence and Interdependence. Foreign Policy. Spring, N° 22: 130-161.

Bean, J. P and B. S. Metzner (1985). A conceptual model non-traditional undergraduate student attrition. Review of Educational Research. Vol. 55, N° 4: 485-540.

Secretaría de Educación Pública (SEP) (2012) Subsecretaría de Educación Superior. Glosario de términos. Recuperado de www.ses.sep.gob.mx

Molina J. M. y García J. "Técnicas de Análisis de Datos - Aplicaciones prácticas utilizando Microsoft Excel y Weka" (en línea), 31 de agosto de 2019, http://www.giaa.inf.uc3m.es/docencia/II/ADatos/apuntesAD.pdf, (2006)

Restrepo B., L. F., & González L., J. (2007). De Pearson a Spearman. Revista Colombiana de Ciencias Pecuarias, 20(2), 183 - 192.

Alpaydin, E. (2014). Introduction to Machine Learning. Adaptive Computation and Machine Learning (Cambridge, Massachusetts: The MIT Press), 3era edición.

Friedman, J. H. (1999) Stochastic Gradient Boosting. Technical report, Dept. of Statistics Stanford University.

Medina-Merino, R., & Ñique-Chacón, C. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases, 0(010), 165-189.

Ali, J., Khan, R., Ahmad, N., y Maqsood, I. (2012). Random forests and decision trees. IJCSI International Journal of Computer Science Issues, 9(5), 272-278.

Cárdenas-Montes, M. (2015) Bagging. Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas.

Breiman, L. Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324

Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference and prediction. Springer.

Refaeilzadeh, P. & Tang, L. & Liu, H. (2009). Cross-Validation. Encyclopedia of Database Systems. 532–538. 532-538. Doi: 10.1007/978-0-387-39940-9_565.

Del Valle, A. R. (2017). Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones. Universidad de Sevilla. Recuperado de https://idus.us.es/xmlui/bitstream/handle/11441/63201/Valle%20Benavides%20Ana%20Roc%C3%A Do%20del%20TFG.pdf?sequence=1

Swets J. A. (1988): 'Measuring the accuracy of diagnostic systems'. Science 240: 1.285-1.293.

Inmon W., Strauss D., Neushloss G. (2008), DW 2.0. The Architecture for the Next Generation of Data Warehousing, Morgan Kaufmann, Burlington, p. 9-14.

ANEXOS

UNIDAD DE POSTGRADO

MANUAL DE INDICADORES

UNIDAD DE POSTGRADO

MANUAL DE INDICADORES

Nombre del Indicador:

Porcentaje de desertores por Programa

Objetivo:

Conocer el porcentaje de estudiantes que desertaron por cada programa de estudios ofertado en relación al total de desertores de la institución de educación superior.

Expresión matemática:

$$Porcentaje\ de\ desertores\ por\ programa = \frac{Número\ de\ desertores\ del\ programa_i}{Total\ de\ desertores} * 100$$

Unidad de medición:

Porcentaje

Fuente de Información:

Unidad de Postgrado

Periodicidad:

Semestral

UNIDAD DE POSTGRADO

MANUAL DE INDICADORES

Nombre del Indicador:

Porcentaje de desertores por Ciudad de Nacimiento

Objetivo:

Conocer el porcentaje de estudiantes que desertaron con respecto a la ciudad dónde provienen

Expresión matemática:

$$Porcentaje\ de\ desertores\ por\ ciudad\ de\ nacimiento = \frac{Número\ de\ desertores\ de\ la\ ciudad_i}{Total\ de\ desertores} * 100$$

Unidad de medición:

Porcentaje

Fuente de Información:

Unidad de Postgrado

Periodicidad:

Semestral

UNIDAD DE POSTGRADO

MANUAL DE INDICADORES

Nombre del Indicador:
Porcentaje de desertores por Sexo

Objetivo:

Conocer el porcentaje de estudiantes que desertaron con respecto al sexo

Expresión matemática:

$$Porcentaje\ de\ desertores\ por\ sexo = \frac{Número\ de\ desertores\ del\ sexo_i}{Total\ de\ desertores} * 100$$

Unidad de medición:

Porcentaje

Fuente de Información:

Unidad de Postgrado

Periodicidad:

Semestral

UNIDAD DE POSTGRADO

MANUAL DE INDICADORES

Nombre del Indicador:

Porcentaje de desertores por Etnia

Objetivo:

Conocer el porcentaje de estudiantes que desertaron con respecto a la etnia con la que se identifican

Expresión matemática:

$$Porcentaje\ de\ desertores\ por\ etnia = \frac{Número\ de\ desertores\ de\ la\ etnia_i}{Total\ de\ desertores} * 100$$

Unidad de medición:

Porcentaje

Fuente de Información:

Unidad de Postgrado

Periodicidad:

Semestral

UNIDAD DE POSTGRADO

MANUAL DE INDICADORES

Nombre del Indicador:
Porcentaje de desertores por Estado Civil

Objetivo:

Conocer el porcentaje de estudiantes que desertaron con respecto a su estado civil

Expresión matemática:

$$Porcentaje\ de\ desertores\ por\ estado\ civl = \frac{Número\ de\ desertores\ por\ estado\ civil_i}{Total\ de\ desertores} * 100$$

Unidad de medición:

Porcentaje

Fuente de Información:

Unidad de Postgrado

Periodicidad:

Semestral

UNIDAD DE POSTGRADO

MANUAL DE INDICADORES

Nombre del Indicador:

Porcentaje de desertores por Ocupación

Objetivo:

Conocer el porcentaje de estudiantes que desertaron con respecto al cargo profesional que desempeñan

Expresión matemática:

$$Porcentaje\ de\ desertores\ por\ ocupación = \frac{Número\ de\ desertores\ por\ ocupación_i}{Total\ de\ desertores} * 100$$

Unidad de medición:

Porcentaje

Fuente de Información:

Unidad de Postgrado

Periodicidad:

Semestral

UNIDAD DE POSTGRADO

MANUAL DE INDICADORES

Nombre del Indicador:

Porcentaje de desertores por Ingreso Mensual Promedio

Objetivo:

Conocer el porcentaje de estudiantes que desertaron con respecto al ingreso mensual promedio

Expresión matemática:

$$Porcentaje\ de\ desertores\ por\ ocupación = \frac{Núm.\ de\ desertores\ por\ ingreso\ mensual\ prom}{Total\ de\ desertores} * 100$$

Unidad de medición:

Porcentaje

Fuente de Información:

Unidad de Postgrado

Periodicidad:

Semestral