Universidad el Bosque

Facultad de Ciencias
Programa de Estadística
Bogotá

Modelos para la Predicción de Deserción Universitaria de Estudiantes de Psicología de la Universidad el Bosque

Autor:
Nicolás Torres Acero

Junio de 2022

Universidad el Bosque

Facultad de Ciencias
Programa de Estadística
Bogotá

Modelos para la Predicción de Deserción Universitaria de Estudiantes de Psicología de la Universidad el Bosque

Participantes

Estudiante:
Nicolás Torres Acero

Director:
Jesús David Ramos Montaña

Junio de 2022

Aprobación

Nicolás Torres Acero
Autor

Jesús David Ramos Montaña
Director

Alejandro Duitama Leal
Jurado

Emiliano Rodríguez Arango
Jurado

Carlos Alberto Puentes Morales
Jurado

Junio de 2022

Índice

1. Lista de tablas

2. Lista de figuras

3. Resumen

Este proyecto busca implementar los modelos de clasificación supervisada Random Forest y XGBoost con el propósito de predecir deserción universitaria de los estudiantes de la carrera de Psicología de la Universidad El Bosque, utilizando información académica, demográfica, socio-económica y de personalidad de los mismos.

Dichos modelos serán comparados utilizando diferentes métricas para identificar el modelo con mayor potencia predictiva y así buscar factores de riesgo de deserción universitaria.

4. Abstract

This project seeks to implement the Random Forest and XGBoost supervised classification models in order to predict college dropout of students in the Psychology program at Universidad El Bosque, using their academic, demographic, socioeconomic and personality information.

These models will be compared using different metrics to identify the model with the highest predictive power and thus search for risk factors for college dropout.

5. Introducción

El fenómeno de deserción en el ambiente universitario según Guzmán et al.(2009) implica no sólo la inasistencia del individuo a la institución misma, sino a su completo retiro de la formación académica. De acuerdo con Guzmán et al.(2009) la deserción consiste en el abandono consciente del estudiante de sus estudios, independiente de su entorno y de las distintas fuerzas o presiones de toda índole que puedan estar afectando su decisión. Sin embargo, en algunos casos se confunde con el ausentismo o con el retiro forzoso, el cual está sujeto al fracaso en el rendimiento académico del estudiante.

La deserción es un fenómeno que debe ser revisado juiciosamente por las IES (Instituciones de Educación Superior), en Guzmán et al.(2009) se describe que en Colombia la Ley 1188 de 2008, estipula la acreditación de calidad y alta calidad en los programas universitarios, con el propósito de lograr que las instituciones educativas ajusten sus programas académicos de tal forma, que permitan garantizar la educación de calidad y el éxito en la culminación de los estudios por parte de sus estudiantes.

Según Guzmán et al.(2009) y MEN (2014) los principales determinantes a los que se le atribuye la deserción universitaria son: la deficiente preparación académica para el ingreso a la educación superior y el rendimiento académico dentro de ella; las condiciones socio-económicas particulares y del país; los aspectos institucionales del claustro universitario y los factores individuales del estudiante como su entorno familiar y su integración social.

Con base en lo anteriormente expuesto, este trabajo propone la implementación y validación de dos modelos predictivos desde la minería de datos, la técnica Random Forest y XGBoost (Gareth et al.(2013) y Hastie et al.(2009)) que permitan predecir, con base en un perfil psicológico, académico y económico, si un estudiante desertará o no del programa al que está adscrito.

La metodología a seguir para la implementación de estos modelos a grandes rasgos, consiste en construir cada uno de los modelos acorde a los perfiles mencionados en el párrafo anterior, validar que estos modelos cumplen con todos los requisitos para realizar una buena medición y compararlos con el propósito de determinar cuál de los dos modelos tiene mayor potencia de predicción de la deserción universitaria.

Como insumo para dichos modelos, se utilizará información de los estudiantes de psicología de la Universidad el Bosque, entre los años 2013 y 2017 de la jornada diurna, suministrados por el Laboratorio de psicometría de dicha Facultad, la cual consta de: información académica, resultados de dos pruebas de personalidad y parte de su información socio-económica.

6. Antecedentes

El presente estudio se basa en los desarrollos teóricos realizados por diferentes investigadores sobre las diferentes técnicas utilizadas para estudiar o evaluar la deserción en diferentes ámbitos, los cuales servirán de marco conceptual para conseguir el objetivo establecido.

En Aulck et al. (2017) se propone el uso de variables demográficas en una población heterogénea para la construcción de tres (3) modelos de aprendizaje automático para medir deserción universitaria: ELS (regresión logística regularizada), KNN (K vecinos más próximos) y Random Forest, en los cuales se utilizó el método de validación cruzada para calcular la fuerza de predicción en cada modelo y posteriormente construir unas curvas ROC para comparar los modelos. De los modelos construidos, la regresión logística obtuvo los mejores resultados en la aplicación de lo expuesto. Adicionalmente, se encontró que las curvas ROC que se construyeron pueden mejorarse para futuros proyectos. Finalmente, la conclusión del estudio es que se lograron unos buenos resultados preliminares para predecir la deserción de estudiantes en un conjunto de datos heterogéneo con datos demográficos y de expediente académico y además se propone un próximo paso que consiste en realizar discusiones con los administradores de la Universidad de Washington con el propósito de mejorar el modelo predictivo y posiblemente expandirlo a otras universidades.

Formia et al.(2013) plantea el uso de información personal, académica y laboral de estudiantes para evaluar esta información con métodos de selección wrapper y genético con el fin de seleccionar las mejores variables para alimentar un modelo. También hacen uso del método SOAP (Selección de atributos por proyección) para obtener una evaluación más detallada y finalmente con los resultados aplicar el método árbol de decisión para predecir deserción desarrollado con el algoritmo C4.5. Este estudio encontró que los métodos de selección redujeron en un 40 % los atributos a evaluar y demostró que los atributos más relevantes son aquellos que se relacionan con la situación laboral actual del estudiante y con su proyección laboral. Los resultados obtenidos por

el modelo concluyen que la Universidad Nacional de Río Negro - UNRN debe generar estrategias que contemplen los variables laborales de los estudiantes para influenciar su permanencia dentro de ella.

Chen et al.(2019) propone el uso de un algoritmo híbrido basado en árboles de decisión y máquinas de soporte vectorial conocido como DT-ELM (Decision Trees - Extreme Learning Machine) el cual es usado para predecir la deserción en los cursos MOOC (Massive Open Online Courses). Este algoritmo consiste en adaptar la estructura de un árbol de decisión a la de un ELM (Extreme Learning Machine), utilizando la teoría de la entropía, con el fin de permitir que la decisión que se obtiene en los nodos hoja del árbol se vea influenciada por los nodos internos. Los resultados indican que el algoritmo DT-ELM es más potente en términos de la entropía, el criterio AUC (area under curve) y el F1-score (media armónica de precisión y recuperación) observados por medio de curvas ROC. La conclusión a la que se llegó es que el algoritmo cumple con los requerimientos establecidos ya que obtuvo mejores resultados según las comparaciones realizadas.

7. Justificación

Todas las Instituciones de educación superior, buscan la acreditación de alta calidad en todos sus programas académicos, como lo estipula la Ley 1188 de 2008, referida anteriormente en Guzmán et al.(2009). Las instituciones que hacen parte de los programas de calidad y de alta calidad se comprometen a controlar la tasa de deserción de estudiantes, con el propósito de conseguir un registro calificado para sus programas, el cual garantiza que estos programas se ofrecen con calidad o alta calidad a los colombianos.

En Colombia, según Guzmán et al.(2009) se define que: «*El Sistema de Prevención de la Deserción en la Educación Superior SPADIES es una herramienta informática que permite a las instituciones y al sector hacer seguimiento a la deserción estudiantil, identificar y ponderar variables asociadas al fenómeno, calcular el riesgo de deserción de cada estudiante a partir de condiciones académicas y socio económicas, y facilitar la elección, seguimiento y evaluación de impacto de estrategias asociadas a disminuirlo.*»

El SPADIES y el Gobierno Nacional de Colombia, a través del Ministerio de Educación sugieren como factores que aumentan el riesgo de deserción: individuales, académicos, institucionales y socio-económicos para determinar

la deserción como se indica en Guzmán et al.(2009) y MEN (2014). En este proyecto, se propone además, utilizar variables extraídas de los perfiles de personalidad de los estudiantes con el fin de proponer nuevas alternativas. Los modelos basados de Random Forest permiten determinar las variables que más aportan o afectan a la variable respuesta, con lo cual, se obtienen los factores de riesgo.

Se busca aplicar métodos de aprendizaje supervisado dentro de un nuevo campo, como lo es la deserción de estudiantes universitarios. De otra parte, se tendrán en cuenta los resultados de las pruebas de personalidad practicadas a los estudiantes de psicología entre los años 2013 y 2017 jornada diurna, información innovadora debido a que resulta desconocido si la deserción de estudiantes se ve afectada por estas variables; por esta razón, si se logra algún resultado positivo con estas nuevas variables, pueden ser de apoyo para futuros proyectos de miembros pertenecientes a la comunidad científica dentro de este ámbito.

8. Objetivos

8.1. Objetivo General

Desarrollar dos modelos de clasificación supervisada para la predicción de deserción universitaria con base en perfiles psicológicos, académicos y socioeconómicos de estudiantes de psicología de la Universidad El Bosque, modalidad diurna, matriculados entre 2013 y 2017.

8.2. Objetivos Específicos

- Implementar un modelo de clasificación supervisada Random Forest para predicción de deserción universitaria.
- Implementar un modelo de clasificación supervisada XGBoost para predicción de deserción universitaria.
- Comparar los dos modelos implementados, en términos de su potencia para predecir deserción universitaria a través de diversas técnicas y métricas adecuadas.
- Identificar con base en el modelo con mayor potencia predictiva, factores de riesgo de deserción universitaria.

9. Notación

Suponemos una tabla de datos $X|Y$, de dimensiones $n \times (p+1)$ donde:

- n: Número de individuos, casos u observaciones del conjunto
- $X = \{X_1, X_2, ..., X_p\}$: Espacio de las p-variables predictoras que pueden ser tanto cuantitativas como cualitativas, generalmente continuas, discretas o categóricas
- X_j: j-ésima variable predictora (con $1 \leq j \leq p$) de tamaño $n \times 1$
- Y: Variable objetivo o de respuesta de tipo categórico que cuenta con dos niveles o clases $Y = 1$ o $Y = 0$ donde 1 es desertó y 0 no desertó el estudiante
- y_i: i-ésima respuesta de cada individuo la cual varia entre 1 y 0 si el individuo desertó o no desertó, respectivamente
- $x_i = \{x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{ip}, y_i\} \Rightarrow i$-ésimo individuo observado sobre las p-variables predictoras y la variable de respuesta Y (con $1 \leq i \leq n$)
- x_{ij}: Valor numérico o categórico correspondiente a la j-ésima variable predictora del i-ésimo individuo
- x_i^*: Es el perfil del individuo x_i compuesto por sus valores observados sobre j variables predictoras
- f: Función clasificadora o clasificador, tal que: $f = X \in \mathbb{R}^p \mapsto \{0, 1\}$ y $x \to f(x) \in \{0, 1\}$
- $\widehat{y_i} = f(x_i^*)$: Valor predicho en Y para el i-ésimo individuo; es 1 o 0 según corresponda
- $\widehat{f}$: Función clasificadora estimada
- P: Conjunto de prueba
- E: Conjunto de entrenamiento
- $x_0^* = \{x_{01}, x_{02}, ...x_{0p}\}$ es el perfil del nuevo individuo x_0 que no ha sido observado en la tabla $X|Y$

10. Marco Teórico

10.1. Deserción Universitaria

La deserción es un fenómeno difícil de interpretar pues sus definiciones varían respecto al contexto. Para este proyecto, la definición de deserción que se utiliza es la establecida por el Ministerio de Educación de Colombia en Guzmán et al.(2009), la cual se muestra como una situación a la que se enfrenta un estudiante que no pudo culminar su proyecto educativo y no presenta actividad académica durante dos semestres consecutivos, lo que equivale a un año de inactividad académica.

En Guzmán et al.(2009), se menciona algunos de los factores determinantes que afectan el fenómeno de deserción universitaria como se muestra en la figura 1:

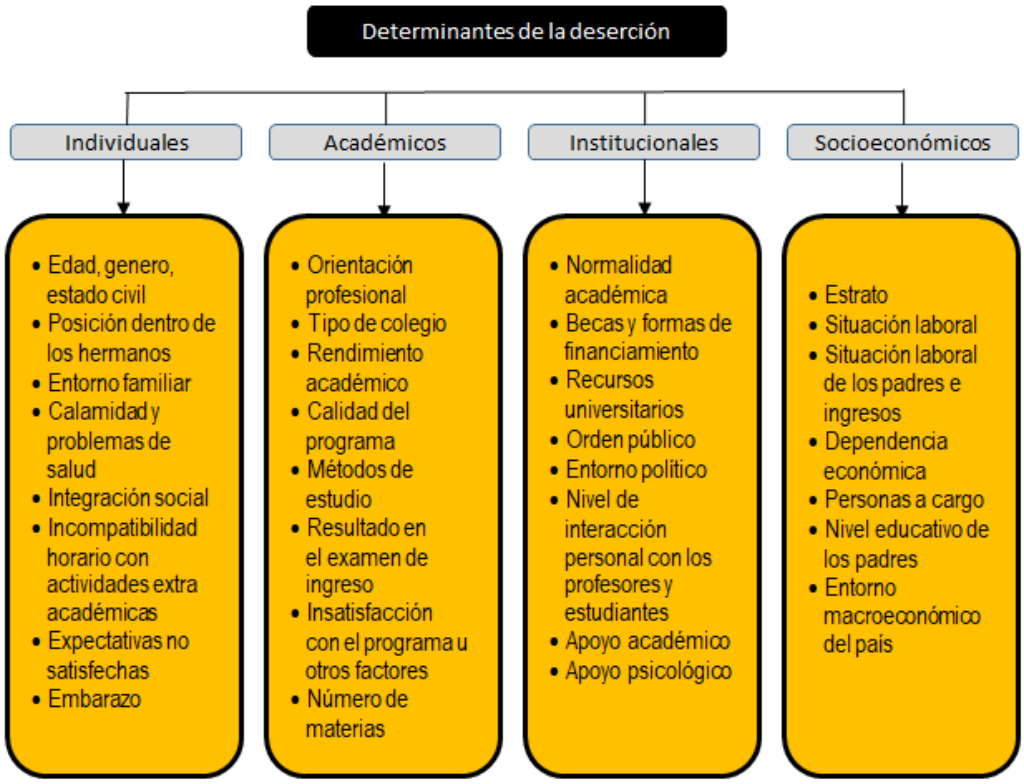

Figura 1: Determinantes de la deserción universitaria; Fuente: Guzmán et al.(2009)

Algunos de estos factores, son nombrados en Díaz (2009) y MEN (2014) como determinantes para algunos estudios citados dentro del mismo acerca de la deserción.

10.2. Aprendizaje Supervisado: Clasificación

El aprendizaje automático (Machine Learning), también conocido como aprendizaje de máquina o aprendizaje estadístico, proporciona a un sistema la capacidad de aprender y mejorar de manera automática (Gironés et al.(2017) y Kubat (2017)) a partir de tres posibles paradigmas: el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje semi-supervisado o de refuerzo.

En Gironés et al.(2017) y Kubat (2017) se indica que el aprendizaje supervisado se compone de un grupo de algoritmos, que en su proceso de aprendizaje, requieren de un conjunto de datos de entrenamiento donde se conocen los valores o las clases de una variable objetivo o de respuesta, y por lo que el algoritmo podrá realizar una predicción para dicha categoría, con base en un conjunto de variables X predictoras conocidas. En este contexto, se llama aprendizaje supervisado, pues las X's supervisan la respuesta Y. De otra parte, en el aprendizaje no supervisado, no hay una variable Y para supervisar dentro del conjunto de datos por lo que el propósito no es hacer predicción sino realizar tareas de tipo descriptivo como: el agrupamiento de información también conocido como clustering o segmentación, Detección de anomalías, reducción de dimensionalidad y detección de patrones. En paralelo a estos dos, esta el aprendizaje semi-supervisado, el cual según Zhu (2008) utiliza una gran cantidad de datos a los cuales se les puede o no conocer su valor con el fin de tener menos limitaciones y así construir mejores clasificadores. Por lo tanto, el aprendizaje semi-supervisado requiere menos esfuerzo humano y además esta diseñado para ser más versátil y brindar así una mayor precisión. En este escrito se dará énfasis al aprendizaje supervisado.

El aprendizaje supervisado, Gironés et al.(2017), se divide en dos tareas: regresión y clasificación. La regresión es un tipo de aprendizaje automático en donde la variable objetivo es de tipo numérico; en cambio, la clasificación supervisada se utiliza cuando la variable objetivo a predecir es de tipo categórico. Algunas técnicas del aprendizaje supervisado que permiten realizar clasificación son: la regresión logística, el Random Forest y las redes neuronales, entre otras. Las relaciones anteriormente explicadas se pueden ver dentro de la figura 2:

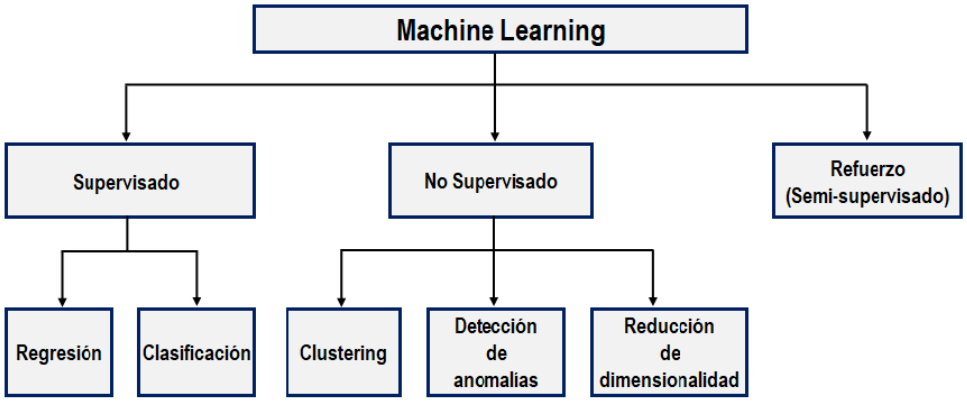

Figura 2: Tipos de aprendizajes y tareas del Machine Learning

En clasificación supervisada se pretende estimar una función f desconocida, la cual clasifica a los individuos con perfil x_i^* de manera correcta en una de las categorías de Y. La clasificación puede ser de dos tipos: clasificación en una vía, cuanto Y cuenta con dos categorías, generalmente 1 y 0, y clasificación multiclase cuando $Y = \{0, 1, 2, ..., m\}$ lo que indica que Y tiene m clases. La función f es comúnmente conocida como función clasificadora o clasificador (Gareth et al.(2013)).

Dentro de esta dinámica, se asume que existe algún tipo de dependencia entre Y y X, según Gareth et al.(2013) por esta razón, existe una relación expresada por el modelo:

$$Y = f(X) + \epsilon$$

La función f anteriormente indicada, es desconocida, por esta razón se hace uso de técnicas de machine learning para estimarla, teniendo en cuenta que Gareth et al.(2013) expresa que como X es un conjunto de datos conocido y $E[\epsilon] = 0$ entonces se puede predecir un valor para Y usando la función:

$$\widehat{Y} = \widehat{f}(X)$$

Donde $\widehat{f}$ se refiere a la estimación de f y $\widehat{Y}$ es el vector que se constituye por medio de las clases $\widehat{y_i}$ predichas del conjunto Y.

10.3. Árboles de Decisión para Clasificación

Los árboles de decisión para regresión o clasificación (Gareth et al.(2013), Hastie et al.(2009) y Tibshirani (2009)) también conocidos como modelos CART (Classificaction and Regression Trees) son modelos de aprendizaje supervisado utilizados para predecir una respuesta en la variable Y haciendo uso de las p variables predictoras de X. Se crea un árbol binario en donde cada uno de sus nodos internos crea una división en el espacio de las p variables X predictoras por medio de una pregunta o regla lógica, hasta llegar a un nodo hoja o nodo respuesta que arroja un valor predicho. Estas divisiones, se pueden representar gráficamente en un espacio con solo dos variables X_1 y X_2 como se observa en la Figura 3, un caso mas general se puede identificar en la figura 4 donde se observa la división de cada nodo:

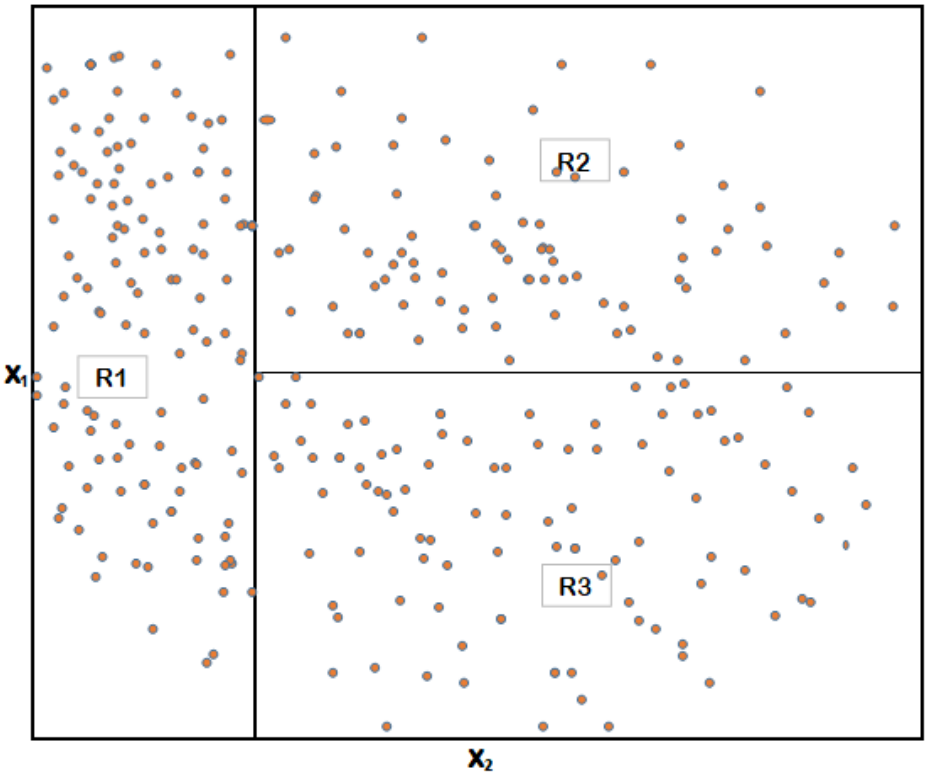

Figura 3: Representación gráfica de la división de una nube de puntos

Los árboles de decisión para clasificación, (Gareth et al.(2013), Hastie et al.(2009), Tibshirani (2009)) dan una predicción de tipo categórica, que corresponde a la clase más frecuente de las predicciones dadas por los nodos

hoja. Las predicciones dentro de los nodos hoja son el producto de todas las segmentaciones creadas por el árbol, las cuales, cuando la variable objetivo tiene dos categorías, serán un grupo de unos y ceros y la predicción final $\widehat{y_i}$, será la clase que más se repita de las dos dentro del nodo hoja.

El nodo principal de un árbol de decisión, es la primera segmentación del conjunto de datos, de ser posible en dos partes, con el fin de agrupar las observaciones más parecidas en la dos áreas divididas. Los nodos internos de los árboles de decisión se refieren a las intersecciones que existen entre las ramas, en donde se plantea una pregunta dentro de cada una con el fin de crear un división. Finalmente, los nodos hoja son las puntas de las ramas del árbol, en donde se encuentran las posibles predicciones resultantes de las reglas de los nodos internos. Esta estructura, se observa en la Figura 4, que muestra un árbol de decisión simple o pequeño, con su correspondiente organización.

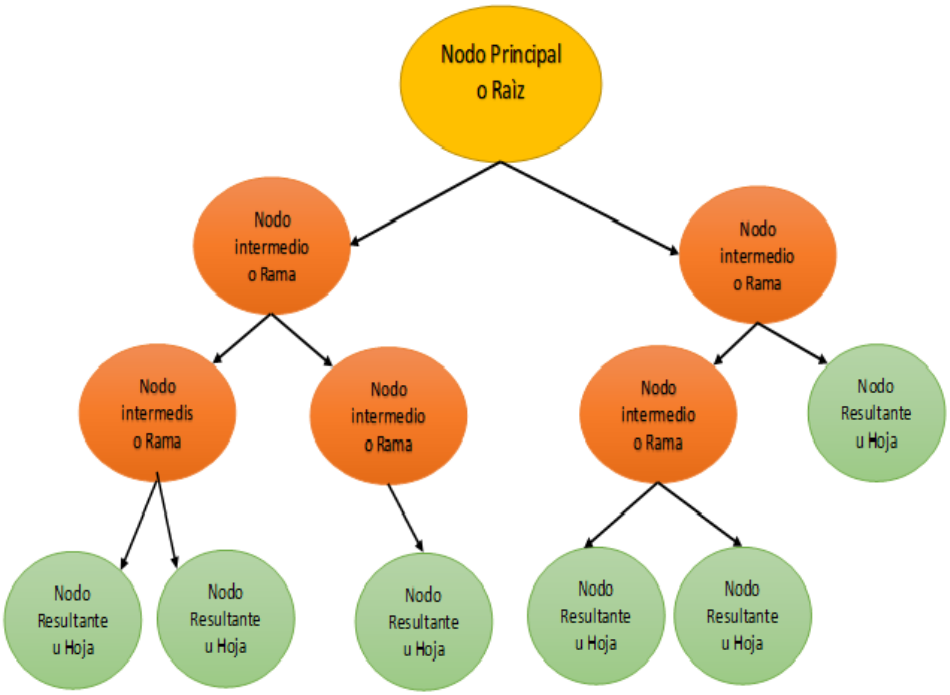

Figura 4: Representación gráfica de los nodos de un árbol de decisión

Las divisiones dentro de los nodos, se producen mediante un cálculo interno como se explica mas adelante; el cual, varia según el criterio del investigador, para así decidir la regla lógica que dividirá el grupo de datos. Algunos de estos criterios, (Gareth et al.(2013), Hastie et al.(2009), son: El error de mala

clasificación de entrenamiento, que se refiere a la proporción de observaciones que no pertenece a la clase más común dentro del conjunto de entrenamiento, buscando minimizar este valor; El índice de Gini, que mide la varianza total dentro de cada nodo, expresado como: $G = \sum_{i=1}^{2} P_i(1 - P_i)$ donde P_i es la proporción de observaciones que pertenecen a un nodo y a la clase i; y la entropía, que se refiere a la impureza dentro de los nodos, siendo la impureza una variedad de clases dentro del nodo, y su calculo es: $D = \sum_{i=1}^{2} P_i \log(P_i)$. Estos cálculos permiten encontrar las posibles divisiones del nodo, cuando su valor se minimiza siendo lo más cercano a cero; el valor de estas dos métricas fluctúa entre uno y cero y es más bajo entre más puro sea el nodo.

Luego de realizar las divisiones dentro de cada nodo, se debe controlar el número de nodos dentro del árbol, esta acción se efectúa mediante reglas de parada, (Gareth et al.(2013), Hastie et al.(2009), las cuales se introducen a manera de hiper-parámetros que son valores que ayudan a configurar el modelo durante el proceso de entrenamiento, los cuales no se obtienen directamente de los datos, sino a través de la experiencia y el criterio del investigador. Estos son: Las observaciones mínimas para división, que entre más grande sea el valor, menos flexible es el modelo lo que implica una árbol mas grande y riesgo de sobre ajuste. Las observaciones mínimas del nodo hoja, que tienen la misma consecuencia que el anterior; la profundidad máxima del árbol, que define el número máximo de divisiones de la rama más larga; el número máximo de nodos hoja, que indica un límite de nodos hoja teniendo consecuencias iguales que el anterior y finalmente, la reducción mínima del error, que define el mínimo error que se debe conseguir para realizar la división.

Es importante controlar el tamaño del árbol mediante técnicas de podado Gareth et al.(2013), porque puede causar sobre ajuste (overfitting), que se define como la reducción de la capacidad de predicción del árbol, porque el modelo se ajusta completamente a los datos de entrenamiento.

La tasa de mala clasificación de prueba, es uno de los criterios más usuales para validar este tipo de modelos, tal como se define más adelante en la sección de validación de modelos, Gareth et al.(2013) y Hastie et al.(2009) se expresa que, aunque resulta ser una medida bastante sencilla de calcular, puede resultar no ser lo suficientemente sensible al momento de validar el modelo, por lo tanto, podría no llegarse a los árboles más óptimos debido a, que si alguno de los dos valores de Y se repite en menor cantidad que el otro, la medida se sobrecarga hacia la clase mas frecuente; por esta razón, el uso de un método más potente como la validación cruzada, nos permite recalcular

esa tasa de mala clasificación, convirtiéndola en una tasa de mala clasificación de validación cruzada, que supera ese déficit de sensibilidad manteniendo la propiedad de la tasa de mala clasificación descrita por Gareth et al.(2013) y Hastie et al.(2009) que indican, que si el objetivo del algoritmo es conseguir la máxima precisión al momento de predecir, este indicador es considerado el mejor.

El aplicar este tipo de modelos, tiene ventajas como la fácil interpretación y visualización, cuando el árbol no es muy grande, y la exigencia es moderada en el arreglo de datos, además, permite trabajar con variables tanto cuantitativas como cualitativas. Sin embargo, también puede tener desventajas como cuando el árbol es muy grande ocasionando sobre ajustes, es inestable debido a que aumenta su varianza, no se puede garantizar que sea el mejor árbol para construir y se puede desbalancear si alguna de las clases de la variable respuesta es demasiado predominante, representando el 65 % o mas de la variable.

En general, el algoritmo para construir un árbol de decisión se puede observar en la Tabla 1:

Tabla 1 Algoritmo para construir un árbol para clasificación

1. Usar los criterios de división para construir el árbol de acuerdo a los datos de entrenamiento, para cuando el nodo hoja o terminal contenga el mínimo numero de observaciones posible.
2. Aplicar los métodos de podado para encontrar el sub-árbol mas óptimo, en función del parámetro α, que controla el ajuste a los datos de entrenamiento del árbol podado.
3. Utilizar la validación cruzada en K folios para buscar el mejor valor para α. Para $k = 1, ..., K$: (a) Repetir el paso 1 y 2 en las $\frac{K-1}{K}$ particiones del conjunto de entrenamiento, excluyendo el k-esimo folio o iteración. (b) Evaluar el error de clasificación dentro del modelo ajustado en los folios diferentes al k-esimo, en función de . Promedie los resultados, y tome el valor de α que minimice el valor del error de clasificación.
4. Retorne el sub-árbol del paso 2 utilizando el valor de α encontrado.

Fuente: Gareth et al.(2013).

10.4. Random Forest

El bagging dentro de la metodología de árboles de decisión para predicción se refiere a la aplicación del método de remuestreo bootstrap; que consiste en generar una cantidad k de sub-muestras a partir de un conjunto de datos $X|Y$, para realizar k veces un cálculo o un modelo para al final promediar todos los resultados y obtener un resultado final. Siguiendo esta metodología, el bagging, en Gareth et al.(2013) y Hastie et al.(2009) ajusta una gran cantidad de árboles de manera paralela con el fin de formar un bosque, como se muestra en la Figura 5. Todos los árboles que se encuentren dentro del bosque darán como resultado una predicción, en cuyo caso para los árboles de clasificación se toma la clase predicha más frecuente entre todas.

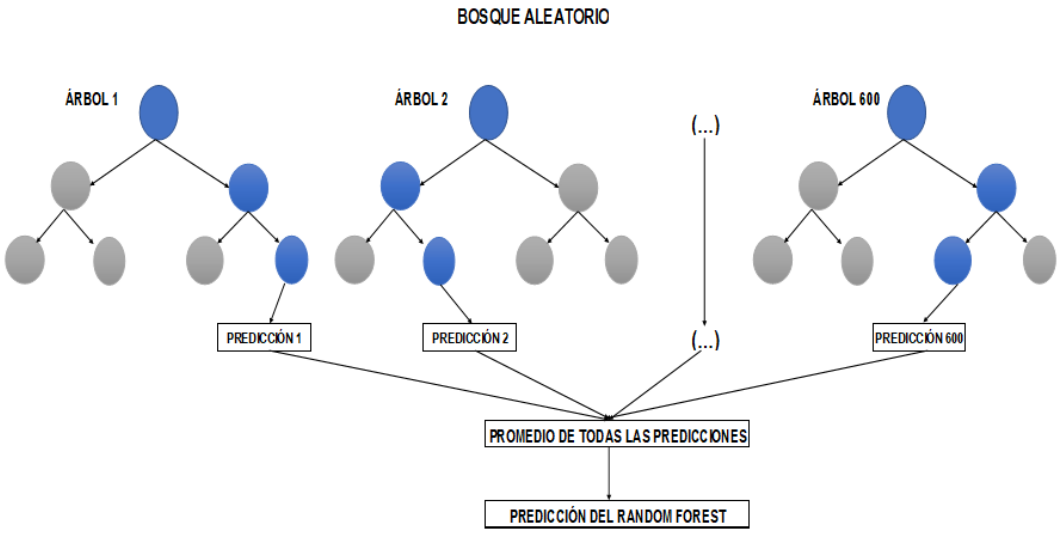

Figura 5: Bosque Aleatorio

El Bagging genera modelos distintos en cada ajuste debido a que en cada uno toma diferentes muestras aleatorias haciendo uso del método de remuestreo bootstrap, a partir de la muestra original; gracias a esto, por lo general genera muestras independientes.

El método Bagging por lo general, genera modelos de poco bias y alta varianza. En este contexto, el bias o sesgo, Gareth et al.(2013), se refiere a un error producido al aplicar un modelo demasiado sencillo para un problema que requiere métodos más complejos. Por otro lado, la varianza, Gareth et al.(2013), indica la cantidad en la que $\widehat{f}$ cambia si se utiliza un conjunto de

datos diferente para entrenamiento, de manera ideal, no debería cambiar la forma de $\widehat{f}$, por lo que a mayor varianza, mayor cambio y mayor flexibilidad; sin embargo, si se agrega una gran cantidad de modelos se tiende a reducir la varianza sin afectar mucho el bias. La estabilidad entre bias y varianza se mide a través del test MSE, el cual se define por notación como $E(y_0 - \widehat{f}(x_0))^2$.

El Random Forest propuesto por Breiman (2001) es un caso particular del Bagging, pero con la diferencia de que este obtiene mejores resultados, Gareth et al.(2013) y Hastie et al.(2009) pues en el Bagging solo se reduce la varianza si los modelos no son correlacionados; en el caso de existir correlación, la varianza disminuye de manera ínfima, en cambio el Random Forest, evita el problema seleccionando por medio del hiperparámetro m, los m predictores de los p que hay dentro de X tanto correlacionados como no correlacionados, antes de evaluar a cada una de las divisiones para que de esta forma un promedio de $\frac{(p-m)}{p}$ divisiones no contemplarán al predictor más influyente, para así permitir que otros predictores sean utilizados si se da el caso que alguno de ellos influya más que otro.

En Hastie et al.(2009) si $m = p$ los resultados del Random Forest y el Bagging son equivalentes debido a que el Random Forest adquiere las propiedades del Bagging y realiza los mismos procesos. Un valor recomendado para el hiperparámetro es $m \approx \sqrt{p}$, siendo p el número de predictores totales.

Aun así, para encontrar el valor óptimo de m se puede llegar a utilizar el método de validación cruzada para optimizar los hiper-parámetros; en comparación con métodos alternos al Bagging como el Boosting, el Random Forest tiene menos hiper parámetros, por lo tanto, es aun más sencillo de implementar, pero si es el caso de existir una proporción alta de predictores que puedan ser considerados irrelevantes, el Random Forest es menos eficiente pues Hastie et al.(2009) plantea que cabe la posibilidad que por el tamaño del m no se tome alguna información relevante; aun así, el Random Forest, facilita la paralelización del modelo por la independencia de los árboles y no sufre del overfitting (sobreajuste) sin importar cuantos árboles utilice.

El Bagging y el Boosting, son metodologías que conservan las propiedades de los árboles de decisión, pero tienen objetivos diferentes, por lo que su aplicación depende del problema; si lo que se busca es que el el modelo sea el más apropiado para representar un caso, el Bagging es mejor opción, pues al crear una gran variedad de modelos, disminuye la varianza del mismo. Por otro lado, si lo que se busca es mejorar el rendimiento de un modelo, es mejor

el Boosting, debido a que su metodología busca reducir el sesgo, optimizando las ventajas del modelo y reduciendo el riesgo, sacrificando algunos valores en la varianza.

10.5. XGBoost

El boosting según Gareth et al.(2013) y Hastie et al.(2009) genera multiples modelos predictivos de manera secuencial, para que el siguiente modelo, tome los parámetros del anterior y decida si debe alterarlos o permanecer con ellos, con el fin, de optimizar sus resultados y así ganar un mayor poder de predicción y una mayor estabilidad en sus resultados debido a, que al igual el Bagging, es un método que estabiliza la varianza, aunque en menor medida; este procedimiento se puede observar en la Figura 6.

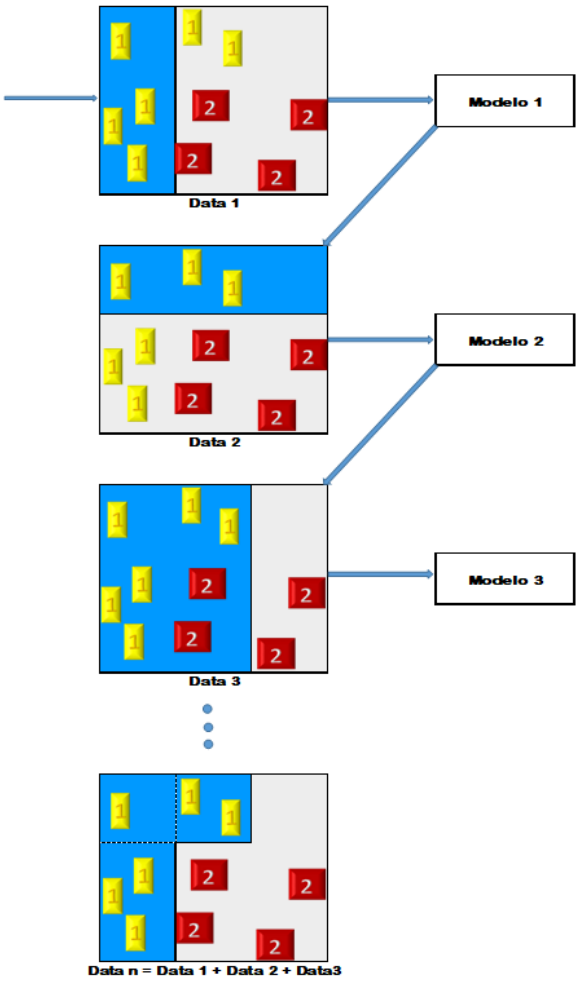

Figura 6: Boosting

XGBoost, Chen et al.(2016) es un sistema de boosting de árboles propuesto por Friedman (2001) comúnmente utilizado por los científicos de datos para lograr mejores resultados en problemas de aprendizaje automático. Se caracteriza por ser rápido, preciso, flexible y eficiente debido a que está diseñado según Chen et al.(2016) para consumir una menor cantidad de recursos por lo que obtiene buenos resultados con mínimo esfuerzo.

Continuando con el procedimiento, durante el entrenamiento de los modelos, los parámetros se ajustan de manera iterativa con el fin de minimizar métricas de validación, como lo es la tasa de mala clasificación.

Si un nuevo árbol resultante de la iteración tiene mejores resultados que el anterior, entonces se toman los parámetros de ese árbol y se utilizan para evaluarlos en nuevas iteraciones con diferentes conjuntos de entrenamiento. Pero si sucede lo contrario, donde el nuevo árbol obtiene peores resultados que el anterior, se conservan dichos parámetros y se prueban en un nuevo conjunto de entrenamientos. Este proceso se repite hasta que la diferencia entre dos árboles consecutivos sea ínfima o cuando se logre el número de iteraciones máximo definida dentro del algoritmo. Este procedimiento, se denomina disminución del gradiente.

Algunos de los parámetros importantes, para el desarrollo del XGBoost, son: El número de iteraciones que se va a permitir, el cual puede ser fijo o en función de los parámetros; el error mínimo, para realizar particiones dentro de los nodos y la profundidad máxima del árbol, que aporta al sobreajuste del modelo, entre otros, que limitan el algoritmo.

En resumen, el XGBoost cuenta con la ventaja de ser muy bueno mejorando el rendimiento de un modelo, y suele tener un muy buen desempeño en conjuntos de datos en donde se mezclan tanto variables numéricas como categóricas.

10.6. Validación de modelos para Clasificación

Los modelos para predicción, después de ser entrenados, independientemente del método que ajuste el modelo, deben ser validados para garantizar que se estén realizando unas buenas predicciones, en el caso de los modelos para clasificación, dichos modelos se validan en base a su potencia de clasificación, la cual se define como el contrario de el error de clasificación, por lo

tanto nos dice, que tan acertado es el modelo en el momento de predecir.

10.6.1. Métricas para evaluar potencia de clasificación

10.6.1.1. Matriz de confusión La matriz de confusión, Gironés et al.(2017), es una representación gráfica de los errores que comete un modelo ya entrenado, al hacer la clasificación de nuevos individuos, por lo tanto, es un método gráfico para estimar la potencia de predicción que tiene un modelo predictivo al realizar una predicción. En la Tabla 2, se puede observar un ejemplo de matriz de confusión.

		Clase Predicha		
		P	N	Total
Clase Observada	P	a	b	a+b
	N	c	d	c+d
Total		a+c	b+d	n

Tabla 2: Matriz de confusión binaria.

La Tabla 2 muestra las clases que han sido clasificadas correcta e incorrectamente por el modelo estimado, representadas por los valores: verdadero positivo (a) que corresponde al número de clasificaciones en la clase positiva u objetivo (P) realizadas de manera correcta; verdadero negativo (d) que es el número de clasificaciones en la clase negativa (N) realizadas de manera correcta; falso negativo (b) que indica el número de clasificaciones en la clase negativa (N) realizadas de manera incorrecta; y falso positivo (c) que corresponde al número de clasificaciones en la clase positiva u objetivo (P) realizadas incorrectamente.

10.6.1.2. Error de clasificación El Error de clasificación, Gironés et al.(2017), es la proporción de individuos nuevos mal clasificados por el modelo, se expresa de la siguiente manera:

$$Error = \frac{b+c}{a+b+c+d}$$

10.6.1.3. Accuracy o Precisión La Precisión, Gironés et al.(2017), es una métrica que proporciona información general acerca de un número de clases, las cuales, han sido correctamente clasificadas y se representa así:

$$Precision = \frac{a+d}{a+b+c+d}$$

10.6.1.4. Recall o Sensibilidad La sensibilidad, Gironés et al.(2017), se define como la probabilidad de clasificar a un individuo de manera correcta cuando su perfil lo cataloga realmente como un caso positivo. Esta probabilidad también es conocida usualmente como un verdadero positivo (a) y es contraria a lo que se conoce en estadística como error tipo 1 o falso positivo (c), además se calcula de la siguiente manera:

$$Sensibilidad = \frac{a}{a+c}$$

10.6.1.5. Especificidad La especificidad, Gironés et al.(2017), es la probabilidad de clasificar correctamente a un individuo cuando su verdadero estado es realmente negativo. Esta probabilidad se denomina también un verdadero negativo (d), el contrario del error tipo 2 o falso negativo (b); y se calcula de la siguiente forma:.

$$Especificidad = \frac{d}{b+d}$$

10.6.2. Curvas ROC y área bajo la curva

En López et al.(2001) para la construcción de las curvas ROC se tiene en cuenta los criterios de sensibilidad y especificidad, denotados anteriormente.

Cuando se realiza análisis de pruebas que cuentan con resultados continuos o de escala, es decir no son dicotómicas, se puede calcular una gran cantidad de valores de sensibilidad y especificidad que varía según la cantidad de opciones que tenga la respuesta, por lo tanto, se puede calcular un valor de sensibilidad y especificidad por cada posible respuesta.

Las curvas ROC (Receiver Operating Characteristic), Gironés et al.(2017), se encargan de medir el rendimiento frente a los falsos positivos y a los verdaderos positivos; la diagonal de la curva ROC, se debe interpretar como un modelo aleatorio y los valores inferiores, se consideran como resultados menos precisos que una estimación producida mediante el uso de datos aleatorios. Un ejemplo de curvas ROC, se puede observar en la Figura 7.

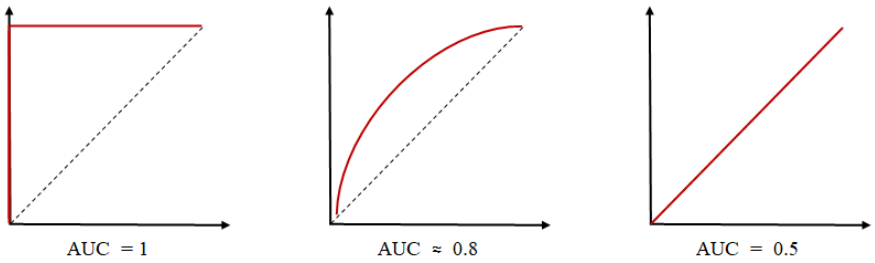

Figura 7: Curvas ROC. Fuente: Gironés et al.(2017)

Para el cálculo de curvas ROC, en la posición superior izquierda de la gráfica se debe encontrar una tasa de verdaderos positivos que sea igual a uno y una tasa de falsos positivos que sea igual a cero, las cuales conforman el clasificador perfecto. Con base en la curva, se calcula el área bajo la curva (AUC - Área Under de Curve) con la cual se caracteriza el rendimiento del modelo evaluado. La figura 6, presenta un ese orden, el rendimiento excelente, bueno y malo que puede tener una curva ROC.

En la implementación de curvas ROC, Gareth et al.(2013), se calcula un valor ajustado para cada observación, en donde el signo de dicho valor, determina a qué lado del límite de decisión se encuentra la observación; Si el valor ajustado excede de cero, entonces la observación se asigna a una clase determinado, pero si el valor es menor que cero se asigna a otra. Luego de tener estos valores ajustados, se puede producir un gráfico ROC.

10.6.3. Validación Cruzada

La validación cruzada según Gareth et al.(2013) es un método de remuestreo que consiste, al igual que el Bootstrap, en particionar una tabla de datos en K pliegues o submuestras no solapadas que sean seleccionadas de manera aleatoria y aproximadamente del mismo tamaño.

El modelo a utilizar es entrenado y validado con las K submuestras tomando una submuestra o pliegue por iteración con el fin de calcular K errores de estimación para promediarlos y dar un resultado final, como se muestra en la Figura 8:

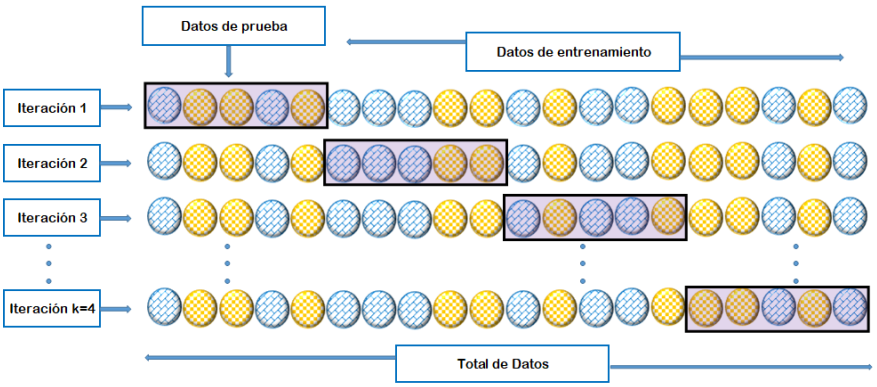

Figura 8: Validación Cruzada.

En la Figura 8, se puede observar como en cada iteración se toma unos datos de prueba dejando los demás para entrenamiento; los datos de prueba varían según la iteración con el fin de entrenar el modelo de diferentes maneras para al final, encontrar los mejores resultados en el promedio de los resultados de cada iteración.

Los modelos que usan validación cruzada, de acuerdo con Gareth et al.(2013) generalmente utilizan entre 5 y 10 pliegues y arrojan modelos con una varianza más reducida. Siguiendo este procedimiento, Gareth et al.(2013) plantea que se proporciona una mejor estimación en el caso de clasificación de la tasa de mala clasificación.

Al utilizar un solo pliegue, se puede calcular una sola tasa o error de mala clasificación y de igual manera, una sola especificidad, sensibilidad o cualquier otra métrica que se derive de ellas. Por otro lado, con K pliegues de prueba, se puede calcular K tasas o errores de clasificación, los cuales, al promediarse al final, conforman el Error de clasificación de validación cruzada, también conocido como $Error_{CV}$ de Clasificación.

La ventaja de utilizar este modelo es, que al tener K valores se puede estimar la varianza del error y generar, por ejemplo, un intervalo de confianza.

11. Descripción de los Datos

La información a analizar fue suministrada por Hernán Salazar, coordinador del Laboratorio de Psicometría de la Universidad el Bosque y consta de la información demográfica, académica y de personalidad de los estudiantes de

psicología para el periodo 2013 a 2017.

Originalmente, fueron obtenidas un total de 46 variables de las cuales se realizó un estudio detallado para identificar cuales cumplen con los requerimientos para realizar el análisis. Luego de realizar un análisis de datos faltantes durante la fase de limpieza, se decidió que solo se trabajaría con 39 de ellas, 38 que corresponden a variables X explicativas y 1 variable Y de respuesta $DESERTOR$; estas variables tienen un porcentaje de datos faltantes inferior al 31 %, y solo se encontró valores superiores a ese porcentaje, en la información correspondiente al desempeño académico por semestre.

La distribución de Y se muestra a continuación:

Descripción	**Cantidad**	**Porcentaje**
Desertores	112	20,40 %
No desertores	437	79,60 %

Tabla 3: Distribución de la variable Y.

Con el fin de lograr la completitud de la información, se realiza imputación multivariada, dejando los datos listos para continuar con el análisis. Para conocer mas detalles sobre la imputación remítase a las fases uno y dos de la metodología, donde se explica cual fue el método utilizado y la razón por la cual se utilizó.

Las variables X que fueron conservadas se pueden ver en la tabla anexa, al igual que sus correspondientes descripciones y categorías para 549 estudiantes resultantes. La distribución de las X_j numéricas que fueron conservadas, se puede observar en la tercera fase de la metodología, donde se realiza una exploración de los datos, para la categorización uniforme. Por otro lado la distribución de las X_j categóricas y su proporción frente cada categoría de Y se puede apreciar en la siguiente tabla:

				Desertor	
Variable	**Categoria**	**Frecuencia**	**Prop. Global**	**0**	**1**
Genero	**1**	195	22.2 %	9.0 %	13.2 %
	2	680	77.8 %	41.0 %	36.8 %
EST_CIVIL	**1**	870	99.5 %	49.5 %	50.0 %
	2	3	0.3 %	0.3 %	0.0 %
	3	1	0.1 %	0.1 %	0.0 %
CIU_NACIM	**0**	184	21.1 %	8.0 %	13.0 %
	1	680	77.8 %	41.3 %	36.5 %
	2	10	1.1 %	0.7 %	0.5 %
SIT_CARRERA	**1**	434	49.7 %	48.6 %	1.0 %
	2	3	0.3 %	0.3 %	0.0 %
	3	249	28.5 %	0.0 %	28.5 %
	4	162	18.5 %	0.0 %	18.5 %
	5	14	1.6 %	1.0 %	0.6 %
	6	4	0.5 %	0.0 %	0.5 %
	7	8	0.9 %	0.0 %	0.9 %
LOCALIDAD	**1**	243	27.8 %	12.4 %	15.4 %
	2	22	2.5 %	0.5 %	2.1 %
	3	5	0.6 %	0.6 %	0.0 %
	4	11	1.3 %	0.6 %	0.7 %
	5	6	0.7 %	0.3 %	0.3 %
	6	2	0.2 %	0.2 %	0.0 %
	7	1	0.1 %	0.0 %	0.1 %
	8	36	4.1 %	2.4 %	1.7 %
	9	26	3.0 %	2.3 %	0.7 %
	10	67	7.7 %	4.9 %	2.7 %
	11	230	26.3 %	14.3 %	12.0 %
	12	22	2.5 %	1.6 %	0.9 %
	13	14	1.6 %	0.8 %	0.8 %
	14	1	0.1 %	0.1 %	0.0 %
	15	2	0.2 %	0.2 %	0.0 %
	16	24	2.7 %	0.7 %	2.1 %
	17	2	0.2 %	0.0 %	0.2 %
	18	15	1.7 %	0.7 %	1.0 %
	19	3	0.3 %	0.3 %	0.0 %
	20	142	16.2 %	7.1 %	9.2 %

Tabla 4: Distribución de las X_j Categóricas.

				Desertor	
Variable	**Categoria**	**Frecuencia**	**Prop. Global**	**0**	**1**
RECURSO	**1**	583	66.7 %	34.0 %	32.7 %
	2	103	11.8 %	6.3 %	5.5 %
	3	50	5.7 %	3.0 %	2.7 %
	4	13	1.5 %	0.7 %	0.8 %
	5	11	1.3 %	1.3 %	0.0 %
	6	114	13.0 %	4.8 %	8.2 %
TIEMP_PERM	**0.5**	285	32.6 %	16.2 %	16.4 %
	1.0	151	17.3 %	5.7 %	11.6 %
	1.5	62	7.1 %	0.5 %	6.6 %
	2.0	105	12.0 %	4.5 %	7.6 %
	2.5	89	10.2 %	6.4 %	3.8 %
	3.0	58	6.6 %	5.1 %	1.5 %
	3.5	64	7.3 %	6.2 %	1.1 %
	4.0	60	6.9 %	5.4 %	1.5 %

Tabla 4: Distribución de las X_j Categóricas.

12. Metodología

En esta sección, se detallan los pasos a seguir para la consecución de los objetivos propuestos, siguiendo cada una de las actividades que se muestran en el procedimiento que se muestra a continuación:

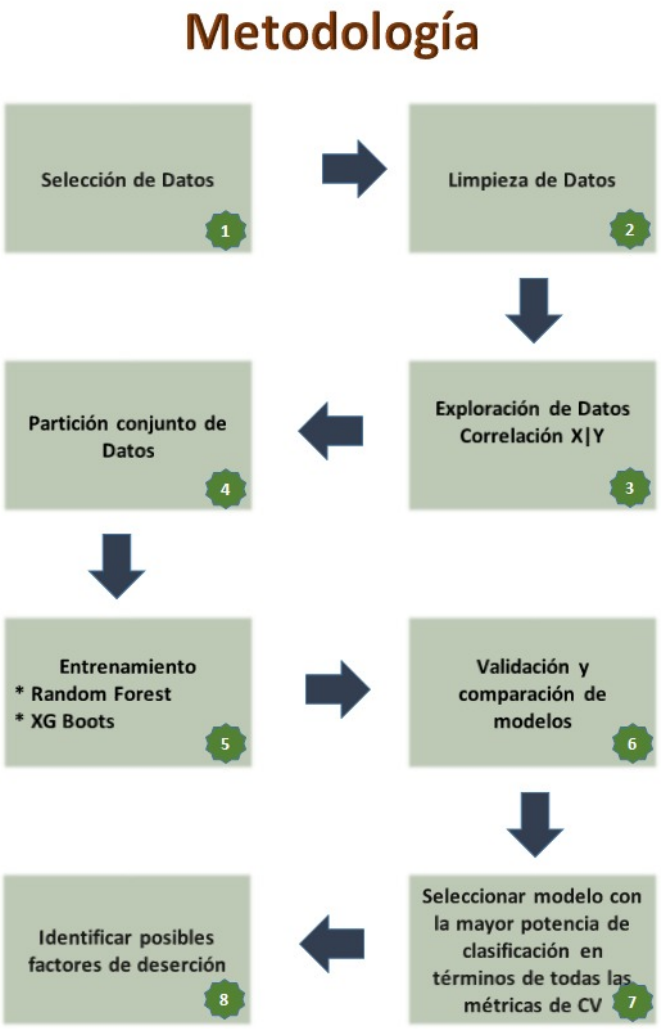

Figura 9: Metodología.

1. Selección de datos: de los datos entregados por el Laboratorio de Psicometría, se debe escoger qué variables son útiles para entrenar los modelos de deserción. A la información seleccionada, se debe buscar la existencia de datos faltantes. Si existen, se debe realizar un estudio de datos faltantes y se calculará el porcentaje de estos frente a los datos totales para cuantificar la cantidad y reconocer las variables que tienen mayor cantidad de los mismos. De ser necesario, se extraerán las variables con mayor cantidad de datos faltantes para de esta forma seleccionar las variables más completas.

Para realizar este procedimiento se hará uso de las librerías Amelia Honaker

et al.(2011) y DataExplorer Boxuan(2020) del software estadístico R(2022), las cuales permiten realizar un análisis de datos faltantes, de manera gráfica.

2. Limpieza de datos: se extraerán las variables que cuenten con un porcentaje de datos faltantes superior al 32 %. Las variables restante serán completadas utilizando imputación multivariada usando la librería "mice"de R Stef et al.(2011), debido a su flexibilidad al trabajar con variables numéricas y categóricas; y a que se puede considerar un método supervisado para imputación. La imputación se realizará para que el conjunto de datos $X|Y$ no contenga NA's.

Mice Stef et al.(2011), también conocido como método de imputaciones multivariadas por ecuaciones encadenadas, se basa en la especificación de un condicional, para que cada variable que contenga valores faltantes, sea imputada mediante un modelo separado, razón por la que permite trabajar con diferentes tipos de variables al mismo tiempo. Para este proyecto se tomó la decisión que el cálculo interno de mice Stef et al.(2011), fuera realizado con un Random Forest, con el fin de conectar la imputación con la metodología que se va a realizar en los siguientes fases.

Por otro lado, como se logró observar en la descripción de los datos, Y presenta un nivel alto de desbalance, debido a la concentración de estudiantes que se consideran no desertores; por ello, es necesario balancear Y, para evitar que los modelos descritos, se inclinen a clasificar una sola categoría durante la fase de prueba. Para esto se utilizará el método de sobre muestreo(oversample) de la librería *unbalanced* Andrea et al.(2015), con la función *ubBalance*, la cual se ajusta para crear nuevos x_i y y_i, utilizando la información de la tabla $X|Y$; que cumplan con pertenecer a la clase no dominante de Y, hasta conseguir un balance de 50 % en las dos categorías.

3. Exploración de datos - Correlación $X|Y$: consistirá en la exploración de la distribución de las X y posteriormente en la transformación de las variables X numéricas a categóricas, usando los cuartiles de cada variable, como una categorización uniforme. Teniendo todas las variables categóricas, se ajustará un modelo de árbol de clasificación a la tabla $X|Y$, con el fin de tener un primer acercamiento a los factores de deserción usando la función $varImp$ de *caret* Max(2021), la cual será explicada mas adelanté por ser parte importante del paso 8 de la metodología.

4. Partición del conjunto de datos: se particionará el conjunto de da-

tos $X|Y$ en un conjunto de prueba P y conjunto de entrenamiento E, los cuales comprenden el 40 % y el 60 % del total de observaciones de $X|Y$, respectivamente.

5.Entrenamiento de modelos: utilizando la técnica de validación cruzada con $K = 10$ folios, se entrenará un modelo de Random Forest utilizando las funciones *trainControl* y *train* de la librearía *caret* Max (2021); y uno de XGBoost usando las funciones *xgb.cv* y *xgboost* de la librearía *xgboost* Tianqi et al.(2022). Variando sus hiper-parámetros por medio de una grilla de búsqueda, para el caso del Random Forest con el hiper-parámetro m y para el XGBoost, con el número de iteraciones, y la tasa de aprendizaje utilizando como insumo el conjunto de entrenamiento E, con el fin de buscar el mejor modelo de cada uno.

6. Validación y comparación de modelos: se ejecutarán los dos modelos finales, resultantes del paso anterior en el conjunto de prueba P, con el fin de estimar su potencia de clasificación usando la tasa de mala clasificación de validación cruzada de prueba, el accuracy, la sensibilidad, y la especificidad, por medio de matrices de confusión, para los hiper-parámetros encontrados durante el entrenamiento, que permitirán la constricción de curvas ROC como sugiere Gareth et al.(2013). Estas métricas pueden verse a afectadas por el sobre muestreo de la fase 2 de la metodología, pero aun así de acuerdo con los antecedentes se consideran suficientes para tomar la decisión.

7. Seleccionar modelo con la mayor potencia de clasificación en términos de todas la métricas de validación cruzada descritas: escoger cuál de los dos modelos finales obtuvo la menor tasa de mala clasificación de validación cruzada de prueba. El modelo seleccionado, se utilizará para encontrar los posibles factores de deserción por medio de la función de R $varImp$ de la librería *caret* Max(2021) de R. y realizar la predicción de las tasas de deserción.

8. Identificar posibles factores de deserción: el modelo final debe ser ajustado a la base de datos $X|Y$, independientemente de las particiones, con el propósito de identificar el conjunto de variables predictoras X que tienen mayor influencia sobre la variable respuesta Y y así, identificar los posibles factores de deserción dados por la función $varImp$ del paquete *caret* Max(2021) de R.

La función $varImp$ de *caret* Max(2021), permuta de manera aleatoria variable

por variable y calcula la disminución del AUC (área bajo la curva) calculado con el conjunto de entrenamiento E. $varImp$ permite desarrollar varias permutaciones para mejorar la estimación, por defecto la función realiza 10 permutaciones y al final promedia las disminuciones de la AUC.

13. Resultados

En esta sección se representa cada uno de los resultados obtenidos durante las ocho fases de la metodología.

13.1. Selección de Datos

De acuerdo con la Metodología, el primer paso es realizar un estudio de datos faltantes, con el fin de reconocer qué variables van a ser conservadas del conjunto de datos original. Para ello, se utilizaron las funciones *missmap* de la librería *Amelia* Honaker et al.(2011), y *plot_missing* de *DataExplorer* Boxuan(2020) del software estadístico R(2022).

la funcion *missmap* de la librería *Amelia* Honaker et al.(2011), representa los siguientes resultados:

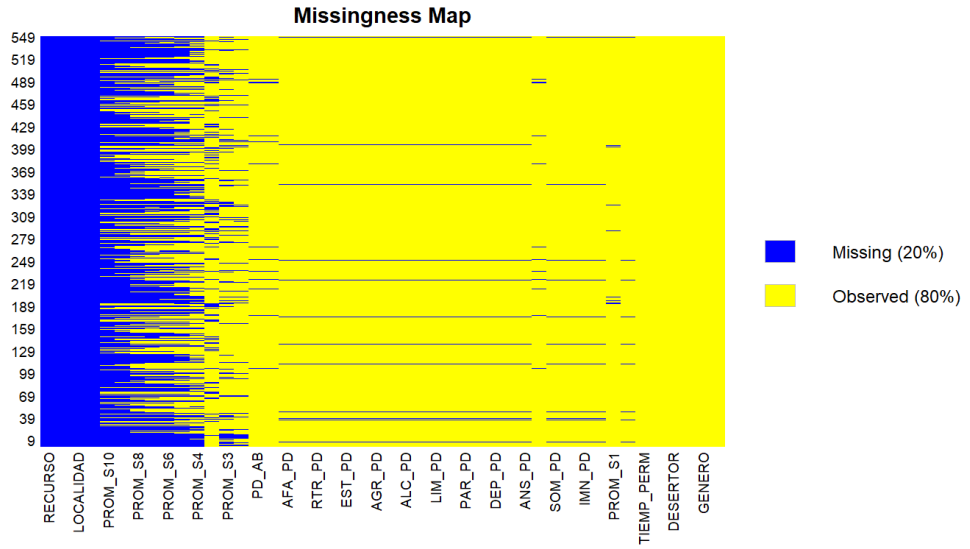

Figura 10: Missingness Map de *Amelia* Honaker et al.(2011).

El Missingness Map representa en azul la cantidad de datos faltantes que hay en cada una de las X, y en amarillo la información completa. De acuerdo con este criterio, los NA, representan el 20 % de la tabla $X|Y$.

Ahora bien, es necesario conocer el porcentaje de datos faltantes que existe, dentro de cada X_j. Para esto se realiza un *plot_missing* de *DataExplorer* Boxuan(2020), el cual se ve a continuación:

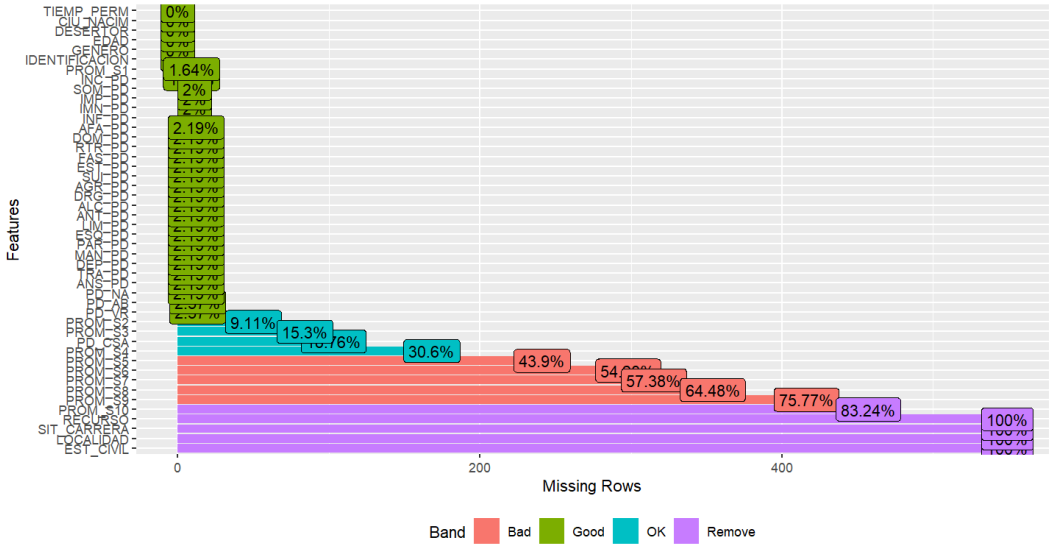

Figura 11: Missing Plot de *DataExplorer* Boxuan(2020).

El Missing Plot de *DataExplorer* Boxuan(2020), representa en morado las variables que deben ser removidas y en rojo las que tienen un porcentaje alto de datos faltantes. De acuerdo con este criterio, solo se van a conservar las variables que cuenten con un porcentaje de datos faltantes menor al 31 %.

13.2. Limpieza de Datos

Originalmente, la tabla $X|Y$ contenía 46 variables de las cuales se conservaron 39, debido a que se conservó algunas de las X que eran candidatas a ser removidas gracias a que fueron completadas con información adicional suministrada por el Laboratorio de Psicometría. 38 de estas corresponden a variables X explicativas y una variable Y de respuesta $DESERTOR$. Las X que fueron extraídas tiene un porcentaje de datos faltantes superior al 40 %, valor que supera el 31 %.

Las X que fueron conservadas fueron imputadas utilizando el método de imputaciones multivariadas por ecuaciones encadenadas *mice* función que pertenece al paquete del mismo nombre Stef et al.(2011). Se decidió emplear este método gracias a su flexibilidad al trabajar con variables numéricas y categóricas, y a los métodos que implementa que fueron mencionados en la fase dos de la metodología. Las X conservadas y su información se pueden ver en la tabla anexa.

El cálculo interno de *mice* Stef et al.(2011), fue realizado con un Random Forest, con el objetivo de conectar la imputación con la metodología, con la que se continuará en las próximas fases.

El siguiente paso de la limpieza de datos consiste en balancear la variable Y, cuya distribución se ve en la descripción de los datos. Para realizar este proceso se utiliza la función *ubBalance* de *unbalanced* Andrea et al.(2015), la cual aplica sobre-muestreo(oversample), método que crea nuevos X_j y y_i, que cumplan con pertenecer a la clase no dominante de Y, hasta conseguir un balance de 50 % en las dos categorías.

La nueva distribución de Y es la siguiente:

Descripción	**Cantidad**	**Porcentaje**
Desertores	437	50 %
No desertores	437	50 %

Tabla 5: Distribución de la Y balanceada.

13.3. Exploración de Datos y Análisis de Correlación

La fase número cuatro de la metodología indica que es necesario hacer una transformación de las X numéricas a categóricas con el fin de calcular con facilidad correlaciones y riesgos relativos entre cada X_j y Y. Para llegar a esto es necesario conocer la distribución de cada X_j para conocer su naturaleza.

Variable	Media	Var	SD	Min	Q1	Mediana	Q3	Max
EDAD	24.10	12.69	3.56	20.00	22.00	24.00	25.00	57.00
PD_VR	14.90	32.40	5.69	0.00	11.00	15.00	18.00	40.00
PD_AB	17.20	135.86	11.66	0.00	9.00	12.00	24.00	60.00
PD_NA	38.30	418.07	20.45	0.00	15.00	42.00	55.00	99.00
PD_CSA	49.90	157.76	12.56	0.00	41.00	50.00	57.00	90.00
INC_PD	10.50	18.68	4.32	0.00	7.00	10.00	13.00	30.00
INF_PD	4.70	5.75	2.40	0.00	3.00	4.00	6.00	24.00
IMN_PD	1.30	3.83	1.96	0.00	0.00	0.00	2.00	15.00
IMP_PD	17.80	18.18	4.26	2.00	16.00	18.00	20.00	27.00
SOM_PD	10.20	39.69	6.30	0.00	6.00	9.00	13.00	43.00
ANS_PD	18.60	72.53	8.52	0.00	13.00	18.00	24.00	51.00
TRA_PD	21.30	54.90	7.41	0.00	17.00	21.00	25.00	52.00
DEP_PD	11.90	38.42	6.20	0.00	7.00	12.00	15.00	35.00
MAN_PD	26.80	59.94	7.74	0.00	21.00	27.00	31.00	53.00
PAR_PD	23.10	54.44	7.38	2.00	17.25	23.00	28.00	52.00
ESQ_PD	15.20	55.22	7.43	0.00	10.00	15.00	20.00	39.00
LIM_PD	18.40	73.00	8.54	0.00	12.00	17.00	24.00	54.00
ANT_PD	15.10	40.46	6.36	1.00	10.00	14.00	19.00	40.00
ALC_PD	2.20	9.48	3.08	0.00	0.00	1.00	3.00	37.00
DRG_PD	3.00	10.40	3.22	0.00	0.00	3.00	5.00	22.00
AGR_PD	11.60	40.81	6.39	0.00	7.00	11.00	15.00	40.00
SUI_PD	1.70	9.17	3.03	0.00	0.00	0.00	2.00	35.00
EST_PD	5.60	8.36	2.89	0.00	4.00	5.00	7.00	16.00
FAS_PD	4.70	9.41	3.07	0.00	3.00	4.00	6.00	17.00
RTR_PD	15.50	16.70	4.09	0.00	13.00	15.00	18.00	61.00
DOM_PD	24.30	25.19	5.02	1.00	22.00	25.00	28.00	36.00
AFA_PD	22.50	22.38	4.73	4.00	19.00	23.00	26.00	34.00
PROM_S1	3.50	0.47	0.68	0.00	3.28	3.67	3.90	4.66
PROM_S2	3.50	0.46	0.68	0.05	3.20	3.67	3.98	4.75
PROM_S3	3.40	0.52	0.72	0.35	3.05	3.46	3.85	4.78
PROM_S4	3.20	0.64	0.80	0.60	2.97	3.42	3.82	4.55

Tabla 6: Distribución de las X_j numéricas.

Para categorizar, se decidió implementar una categorización por entropía usando arboles de decisión, utilizando la función de R "smbinning"del paquete "smbinning"Herman (2019). De tal manera que se buscaba que todas las X_j fueran categóricas, pero algunas resultaron no ser representativas para generar una partición, por dicha razón se decide conservar algunas X_j numéricas que no fueron adecuadas para el método de categorización.

Continuando con la metodología, para hacer un primer acercamiento a los factores de deserción se ajustó un modelo de árbol de clasificación sencillo el cual no tiene en cuenta la validación cruzada, ni la optimización de ningún parámetro. A este árbol se le aplica la función $varImp$ de *caret* Max(2021); de acuerdo con esto, las variables mas importantes para el modelo son: la situación de la carrera ($SIT\_CARRERA$), las notas promedio correspondientes a los semestre 1 al 3 ($PROM\_S1$, $PROM\_S2$, $PROM\_S3$), el tiempo de permanencia en la carrera ($TIEMP\_PERM$) y el puntaje directo de la prueba de personalidad PAI en el ámbito de depresión (DEP_PD), estos resultados se pueden ver en la tabla 7:

Variable	**Importancia**
SIT_CARRERA_4	100.0
PROM_S4_T04>3.3	51.4
SIT_CARRERA_3	45.8
PROM_S3_T02>3.32	32.9
PROM_S2_T03>3.67	24.5

Tabla 7: Resultados de $varImp$ del árbol de clasificación.

Teniendo en cuenta la tabla 7, solo se va realizar una análisis de Riesgo Relativo (RR) a las X_j situación de la carrera y al promedio de notas correspondiente al cuarto semestre las cuales obtuvieron los valores mas altos de importancia, por encima del 40 %.

	RR	lower	upper
1	1.00		
2	0.00	0.00	
3	43.50	23.60	80.30
4	43.50	23.60	80.30
5	15.50	6.10	39.50
6	43.50	23.60	80.30
7	43.50	23.60	80.30

Tabla 8: Riesgo Relativo de situación de la carrera.

De acuerdo con la información de la tabla 8, el riesgo de desertar aumenta en un 43.5 % cuando la persona tiene una situación de la carrera de pérdida de la calidad académica (3), pérdida de la calidad voluntaria (4), admitido (6) y reserva de cupo (7); por otro lado cuando esta en estado de prueba académica (5) el riesgo relativo es de 15.5 %.

	RR	lower	upper
01 <= 2.45	1.00		
02 <= 3.15	0.90	0.80	0.90
03 <= 3.3	0.60	0.40	0.70
04 > 3.3	0.30	0.30	0.30

Tabla 9: Riesgo Relativo de la nota promedio del cuarto semestre.

La tabla 9 indica que el riesgo de desertar como máximos se eleva en un 1 % cuando un estudiante tiene un promedio en el cuarto semestre menor o igual a 2.45 y como es de esperar el riesgo baja entre más alto es el promedio.

13.4. Partición del Conjunto de Datos

El cuarto paso de la metodología consiste en dividir el conjunto de datos $X|Y$, en un conjunto de entrenamiento E y un conjunto de prueba P con una proporción de 60 % y 40 % respectivamente, de tal manera que E se compone por 526 individuos x_i y P por 348 x_i. La distribución de Y en cada conjunto se puede ver en las tablas 10 y 11 que están a continuación:

Descripción	Cantidad	Porcentaje
Desertores	263	50 %
No desertores	263	50 %

Tabla 10: Distribución de la Y en el conjunto E.

Descripción	Cantidad	Porcentaje
Desertores	174	50 %
No desertores	174	50 %

Tabla 11: Distribución de la Y en el conjunto P.

13.5. Entrenamiento de modelos

13.5.1. Random Forest

En la etapa 5 de la metodología se describe el entrenamiento de los modelos, inicialmente se entrena un modelo de Random Forest, usando validación cruzada a 10 folios y una grilla de búsqueda para el hiper-parámetro m aplicando la función *train* del paquete *caret* Max(2021).

El resultado de esta metodología fue un modelo de Random Forest con $m = 10$; a continuación se puede apreciar la distribución de la presión para comprobar si el m encontrado es el mejor:

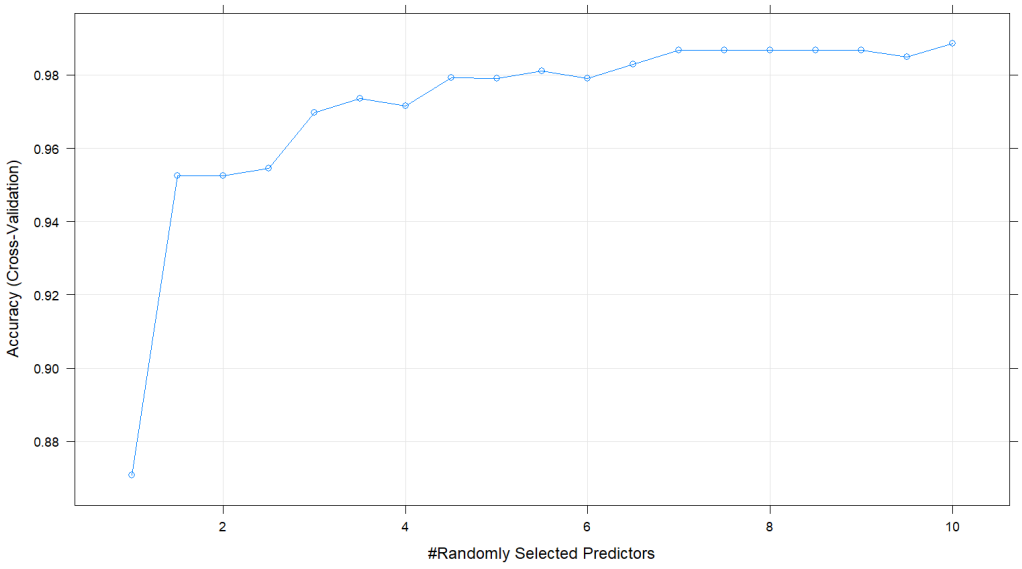

Figura 12: Distribución de la precisión de entrenamiento respecto a mtry.

La distribución del error indica que el cambio mas brusco dentro del modelo sucede cuando $m = 1$, si se tomara $m = 10$ el recomendado de *caret* Max(2021), el modelo podría empezar a sobre ajustar, por esta razón se considera que $m = 1$ es el más indicado.

La matriz de confusión del modelo propuesto por *caret* Max(2021) es la siguiente:

		Clase Predicha		
		1	0	Total
Clase Observada	1	4877	168	5045
	0	120	4829	4949
Total		5140	4997	9994

Tabla 12: Matriz de confusión de validación cruzada del modelo Random Forest.

13.5.2. XGBoost

Continuando con la segunda parte de la etapa 5 de la metodología, se entrena un modelo de XGBoost con el propósito de encontrar mejores resultados que el modelo de Random Forest. Para realizar el modelo de XGBoost, se utiliza la función *train* de *caret* Max(2021) conectada con la librería *xgboost* Tianqi(2022) de R, R (2022), el modelo sugerido por *caret* Max(2021) indica que se necesitan 14 iteraciones.

La distribución de la presión respecto al numero de iteraciones se ve en la figura 13:

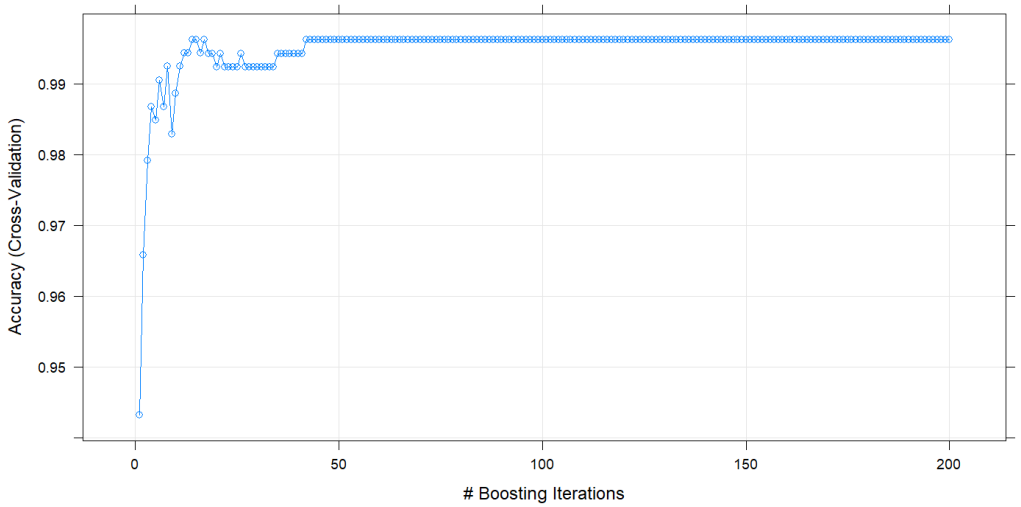

Figura 13: Distribución de la precisión de entrenamiento respecto al numero de iteraciones del XGBoost.

De acuerdo con la figura 13 el cambio mas considerable se encuentra cuando el numero de iteraciones es igual a 4. Por esta razón se decidió ajustar el modelo final de XGBoost usando las 4 iteraciones. La matriz de confusión de validación cruzada obtenida durante el entrenamiento del modelo XGBoost es representada en la tabla 13 que se encuentra a continuación:

		Clase Predicha		
		1	0	Total
Clase Observada	1	52140	67	52207
	0	460	52533	52993
Total		52600	52600	105200

Tabla 13: Matriz de confusión de validación cruzada del modelo XGBoost.

13.5.3. Métricas de Comparación

Los resultados obtenidos en el entrenamiento de los dos modelos se pueden observar en la tabla 14, a manera de medidas de validación calculadas con las matrices de confusión de validación cruzada descritas en las tablas 12 y 13.

Métrica	Random Forest	XGBoost
Precisión	0.97	0.99
Sensibilidad	0.97	0.99
Especificidad	0.96	0.99
Kappa	0.94	0.99

Tabla 14: Métricas de Comparación de la Fase de Entrenamiento.

Durante la fase de entrenamiento los dos modelos tuvieron un desempeño muy bueno, ya que las diferentes métricas de validación cruzada, reportaron valores óptimos. Se procede con el siguiente paso, utilizando los modelos finales Random Forest con $m = 1$ y XGBoost con 4 iteraciones.

13.6. Validación y Comparación de Modelos

13.6.1. Random Forest

Teniendo en cuenta los resultados obtenidos durante la fase de entrenamiento, se entrena nuevamente el modelo, con los datos de prueba con el fin de conocer el valor de m el cual distribuye la precisión como se observa en la figura 14:

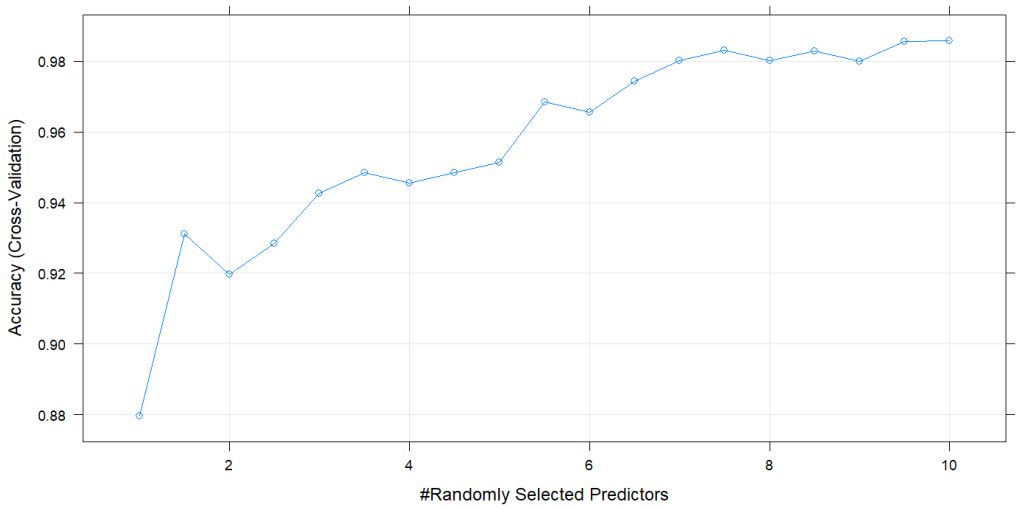

Figura 14: Distribución de la precisión de prueba respecto al hiper-parámetro m del Random Forest.

Teniendo en cuenta la figura 14, se optó por ajustar el modelo de Random Forest con $m = 1$ como modelo final, a los datos de prueba P con el fin de poder evaluarlo; proceso que dio como resultado la siguiente matriz de confusión:

		Clase Predicha		
		1	0	Total
Clase Observada	1	166	13	179
	0	8	161	169
Total		174	174	348

Tabla 15: Matriz de confusión de prueba del modelo Random Forest.

Los resultados del Modelo final de Random Forest, que hacen parte de la sexta etapa de la metodología, indican que el modelo clasifica muy bien alcanzando una Precisión (accuracy) del 93,97 %.

Con el fin de complementar los resultados se gráfica la curva ROC usando la función *roc* del paquete *pROC* Xavier et al.(2011) y se calcula su respectiva área bajo la curva.

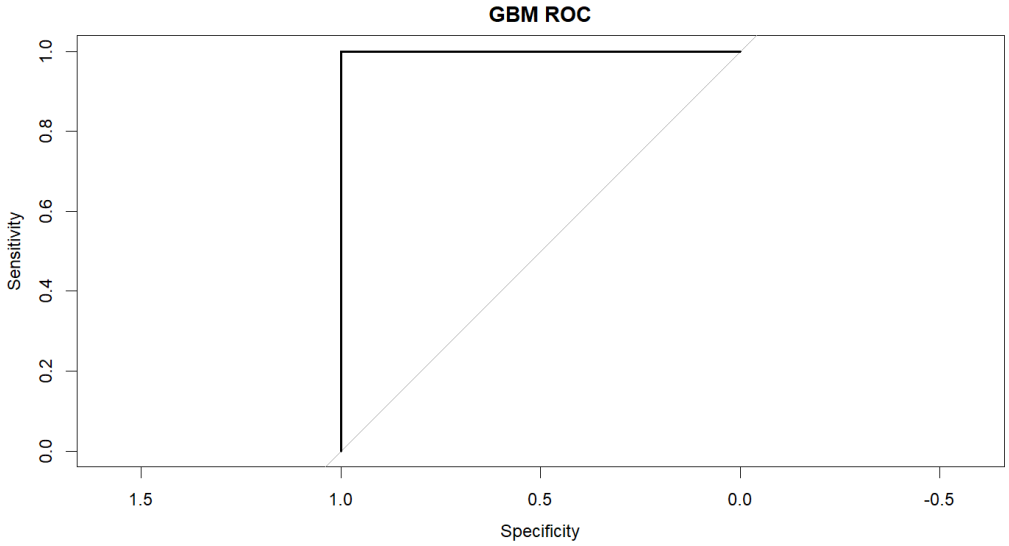

Figura 15: Curva ROC del Random Forest con el paquete *pROC*.

El área bajo de curva (AUB) de la curva ROC es de 1, valor que resulta cuando la sensibilidad y la especificidad del modelo son muy cercanos a 1.

13.6.2. XGBoost

Procediendo con la segunda parte de la fase 6, donde se realiza el ajuste del modelo XGBoost obtenido a los datos de prueba P; el proceso resultó la siguiente distribución de la precisión:

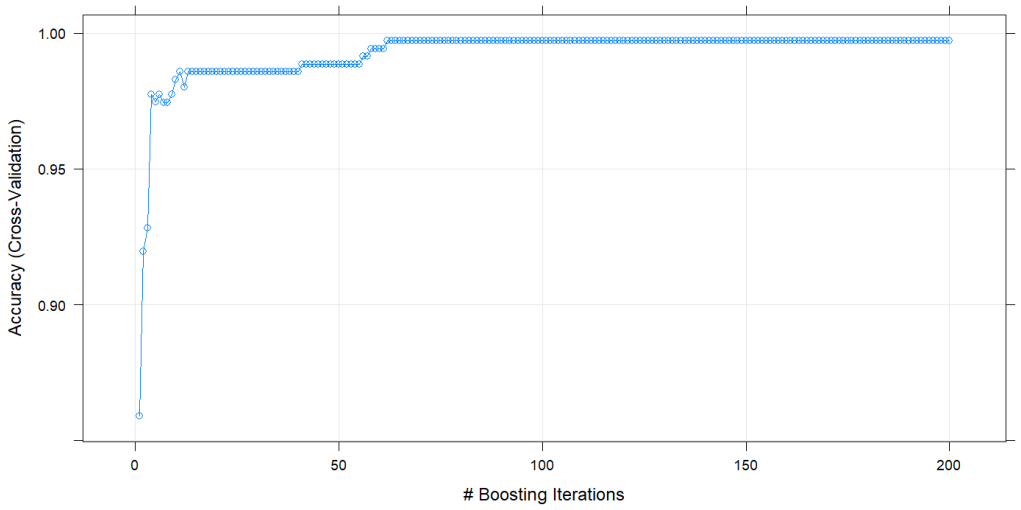

Figura 16: Distribución de la precisión de prueba respecto al numero de iteraciones del XGBoost.

Teniendo en cuenta la figura 16, se ajusta un modelo de XGBoost final a los datos P, con un número de iteraciones igual a 4, proceso que reportó la siguiente matriz de confusión:

		Clase Predicha		
		1	0	Total
Clase Observada	1	170	4	174
	0	4	170	174
Total		174	174	348

Tabla 16: Matriz de confusión de prueba del modelo XGBoost.

Los resultados del Modelo final de XGBoost, indican que el modelo clasifica de manera correcta, con una precisión (accuracy) del 97.7 %, valor que supera el del Random Forest. Para apoyar estos resultados a continuación se encuentra la curva ROC calculada con la función *roc* del paquete *pROC* Xavier et al.(2011) que tiene un área bajo la curva igual a 0.99:

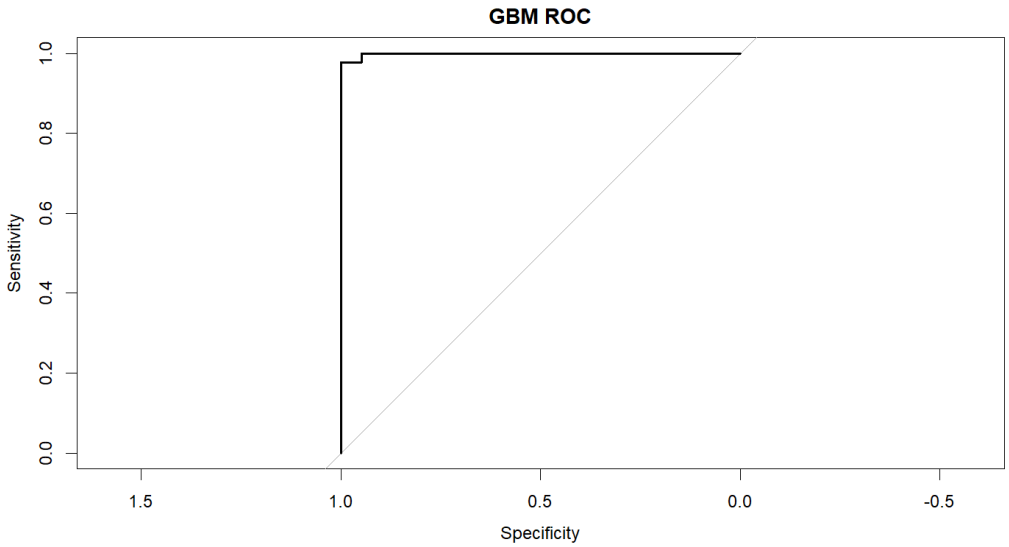

Figura 17: Curva ROC del XGBoost con el paquete *pROC*.

Al obtener los dos mejores modelos en la etapa 6 de la metodología, se continuó con la etapa 7 donde se busca compararlos, y seleccionar el que tenga mayor tasa de mala clasificación de validación cruzada. En la siguiente tabla se puede observar la exactitud, la métrica Kappa y otras métricas, correspondientes a cada modelo:

13.6.3. Métricas de Comparación

Los resultados obtenidos durante la prueba de los dos modelos se pueden observar en la tabla 17:

Métrica	Random Forest	XGBoost
Precisión	0.93	0.97
Sensibilidad	0.95	0.97
Especificidad	0.92	0.97
Kappa	0.87	0.95
AUC	1.0	0.99

Tabla 17: Métricas de Comparación de la Fase de Prueba.

La tabla 17, permite apreciar algunas de las métricas más importantes, que permiten la comparación y validación de cada uno de los modelos. Durante la fase de prueba los dos modelos tuvieron un desempeño muy bueno.

Teniendo en cuenta estos resultados, se observó que el Random Forest en general obtuvo un rendimiento inferior al del XGBoost para este ejercicio. Por esta razón, se tomó la decisión de ajustar dicho modelo a la tabla $X|Y$, para continuar con el octavo paso de la metodología.

13.7. Detección de Factores de Deserción Universitaria

Para realizar la detección de factores de deserción universitaria primero se ajustó el modelo de XGBoost obtenido en la etapa de validación y comparación al conjunto de datos $X|Y$, con el fin de obtener el modelo final.

Ya contando con el modelo final, se utilizó la función *varImp* del paquete *caret* Max(2021), para encontrar las X_j que tienen mayor impacto dentro de la $\widehat{f}$ seleccionada. Los resultados obtenidos se observan en la tabla 18 que se muestra a continuación:

Variable	**Importancia**
SIT_CARRERA_3	100.0
SIT_CARRERA_4	79.3
PROM_S4_T04>3.3	43.4
PD_CSA	11.0

Tabla 18: Resultados de *varImp* del XGBoost Final.

La tabla 18, indica que solo se implementará una análisis de Riesgo Relativo (RR) a la X_j situación de la carrera, y a la nota promedio del cuarto semestre, las cuales obtuvieron los primeros tres puestos en términos de importancia. El riesgo relativo de las dos X_j, ya fue expuesto en las tablas 9 y 10, en la sección de exploración de datos y análisis de correlación de los resultados.

De acuerdo con los resultados de las tablas 8, 9 y 18, se observó que el mayor factor de riesgo de deserción universitaria es la situación de la carrera. Por

otro lado, la tabla 18 indica que aunque la nota promedio del cuarto semestre pertenece a los tres aspectos que más afectan al modelo final $\widehat{f}$, su riesgo relativo nos es muy alto, por lo que no se puede considerar un factor crítico.

14. Discusión de los Resultados

14.1. Entrenamiento de modelos

Durante la fase de entrenamiento,se obtuvo dos resultados, el primero, son las matrices de confusión de validación cruzada de los dos modelos entrenados y el segundo resultado, son los hiper-parámetros encontrados para cada modelo. Estos resultados, fueron los esperados debido al buen desempeño de los dos modelos durante la etapa de entrenamiento.

De acuerdo con lo anterior, se esperaba que las métricas de validación representaran valores óptimos por encima del 75 %, pero no que fueran exactamente del 100 %, durante la validación de cada modelo.

14.2. Validación y Comparación de Modelos

Teniendo en cuenta el desempeño de los modelos durante la fase de entrenamiento, se esperaba que el desempeño de los modelos, fuera óptimo. Las matrices de confusión del Random Forest y del XGBoost demostraron que los dos métodos son efectivos para este ejercicio.

Para validar estos dos modelos, se utilizaron métricas como la precisión, la sensibilidad, la especificidad y las curvas ROC; métricas que alcanzaron su valor máximo o uno muy cercano a él.

Los resultados anteriores, hacen pensar que los modelos seleccionados, no sobre ajustaron en el entrenamiento ni en la prueba del modelo, aunque es una hipótesis que no fue comprobada.

De otra parte, las validaciones obtenidas de cada $\widehat{f}$ fueron los esperados, puesto que se esperaba que el XGBoost reportara mejores resultados que el Random Forest, aunque fue por una mínima diferencia.

Probablemente XGBoost, obtendría mejores resultados si se optimizarán todo los aspectos del algoritmo, por ejemplo, para este trabajo no se optimizó la

semilla que controla las iteraciones con el fin de utilizar la misma semilla para todo los procesos iterativos.

15. Conclusiones

- El modelo de Random Forest obtuvo buenos resultados, se considera una herramienta precisa para la predicción de deserción universitaria, por lo tanto cumple el primero de los objetivos específicos de tal manera que se concluye que el Random Forest superó las expectativas de la investigación.

- El XGBoost obtuvo los resultados esperados, se creía que seria un modelo mas potente que el Random Forest, afirmación que fue comprobada por los buenos resultados, por lo tanto, se llega a la conclusión de que cumple con el segundo objetivo especifico, gracias a su buen desempeño.

- Al comparar el Random Forest con el XGBoost, se esperaba que el XGBoost fuera mucho mejor en todos los aspectos, las métricas que se mencionaron en la metodología sugieren que así fue, por lo tanto se concluye que para este trabajo el XGBoost, se considera la mejor opción para predecir deserción universitaria.

- Se concluye que el aspecto que incrementa el riego de deserción en mayor medida, es la situación en la carrera, y lo aumenta aun más en dos situaciones particulares, la primera cuando el estudiante se encuentra en un estado crítico a nivel académico y la segunda cuando el estudiante no tiene decidido su ingreso y permanencia en la universidad. Esta conclusión da espacio a otra que consiste en que realmente puede que las variables X que fueron utilizadas, no fueron factores lo suficientemente riesgosos como para considerarse factores de deserción universitaria.

Referencias

Breiman, L. (2001). *Random Forests*. University of California, Berkeley.

Friedman, J. (2001). *Greedy Function Approximation: A Gradient Boosting MachineA*. Stanford University.

López de Ullibarri Galparsoro I, P. F. (2001). Curvas ROC. *Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña (España)*.

Xiaojin, Z. (2008). Semi-Supervised Learning Literature Survey. *University of Wisconsin, Madison*.

Díaz, C. J. (2009). Factores de Deserción Estudiantil en Ingeniería: Una Aplicación de Modelos de Duración. *Universidad Católica de la Santísima Concepción, Facultad de Ingeniería, Alonso de Ribera 2850, Concepción-Chile*.

Guzmán Ruiz Carolina, J. F. G., Diana Durán Muriel. (2009). *Deserción estudiantil en la educación superior colombiana*. Bogotá, Colombia: Ministerio de Educación Nacional.

Tibshirani, R. (2009). *Classification and Regression Trees*.

Trevor Hastie, J. F., Robert Tibshirani. (2009). *The elements of statistical learning Data Mining, Inference, and Prediction*. Stanford, California.

Honaker, J., King, G. & Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, *45*(7), 1-47. https://www.jstatsoft.org/v45/i07/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1-67. https://www.jstatsoft.org/v45/i03/

Formia Sonia, W. H., Laura Lanzarini. (2013). Caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos. Un caso de estudio. *RedUNCI-UNLP, Argentina*.

Gareth James, T. H., Daniela Witten. (2013). *An Introduction to Statistical Learning with Applications in R*.

Ministerio de educación nacional, M. (2014). *Acuerdo nacional para disminuir la deserción en educación superior*. Bogotá, Colombia: Ministerio de Educación Nacional.

Pozzolo, A. D., Caelen, O. & Bontempi, G. (2015). *unbalanced: Racing for Unbalanced Methods Selection* [R package version 2.0]. https://CRAN.R-project.org/package=unbalanced

Chen Tianqi, C. G. (2016). XGBoost: A Scalable Tree Boosting System.

Aulck Lovenoor, J. B., Nishant Velagapudi. (2017). Predicting Student Dropout in Higher Education. *DataLab, The Information School, University of Washington, Seattle, WA 98195, USA*.

Jordi Gironés, J. M., Jordi Casas. (2017). *Minería de Datos Modelos y Algoritmos*. Editorial UOC.

Kubat, M. (2017). *An Introduction to Machine Learning*. Springer.

Chen Jing, X. S., Jun Feng. (2019). MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine. *Hindawi*.

Jopia, H. (2019). *smbinning: Scoring Modeling and Optimal Binning* [R package version 0.9]. https://CRAN.R-project.org/package=smbinning

Cui, B. (2020). *DataExplorer: Automate Data Exploration and Treatment* [R package version 0.8.2]. https://CRAN.R-project.org/package=DataExplorer

Kuhn, M. (2021). *caret: Classification and Regression Training* [R package version 6.0-88]. https://CRAN.R-project.org/package=caret

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y. & Yuan, J. (2022). *xgboost: Extreme Gradient Boosting* [R package version 1.6.0.1]. https://CRAN.R-project.org/package=xgboost

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

16. Anexos

16.1. Anexo 1

Nombre	**Descripción**	**Tipo**	**Rango de Variación**
ID	Documento de identidad del estudiante	Cualitativa Nominal	No está definido
Desertor	Variable de respuesta, que representa si un estudiante deserto(1) o no(0)	Cualitativa Nominal Dicotómica	0 o 1
Genero	Genero del estudiante, que se representa como Femenino(0) y Masculino(1)	Cualitativa Nominal Dicotómica	0 o 1
Edad	Edad del estudiante	Cuantitativa Discreta	20 - 57
PD_VR	Puntaje directo de la prueba de personalidad conocida como DAT en la sección de razonamiento verbal (VR)	Cuantitativa Discreta	0 - 120
PD_AB	Puntaje directo de la prueba de personalidad conocida como DAT en la sección de razonamiento abstracto (AB)	Cuantitativa Discreta	0 - 120
PD_NA	Puntaje directo de la prueba de personalidad conocida como DAT en la sección de razonamiento numérico (NA)	Cuantitativa Discreta	0 - 120
PD_CSA	Puntaje directo de la prueba de personalidad conocida como DAT en la sección de razonamiento rápido y preciso (CSA)	Cuantitativa Discreta	0 - 120
INC_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de inconsistencia	Cuantitativa Discreta	0 - 120
INF_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de infrecuencia	Cuantitativa Discreta	0 - 120
IMN_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de impresión negativa	Cuantitativa Discreta	0 - 120

Nombre	**Descripción**	**Tipo**	**Rango de Variación**
IMP_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de impresión positiva	Cuantitativa Discreta	0 - 120
SOM_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de somatización	Cuantitativa Discreta	0 - 120
ANS_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de ansiedad	Cuantitativa Discreta	0 - 120
TRA_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de trastorno de ansiedad	Cuantitativa Discreta	0 - 120
DEP_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de depresión	Cuantitativa Discreta	0 - 120
MAN_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de manía	Cuantitativa Discreta	0 - 120
PAR_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de paranoia	Cuantitativa Discreta	0 - 120
ESQ_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de esquizofrenia	Cuantitativa Discreta	0 - 120

Nombre	Descripción	Tipo	Rango de Variación
LIM_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de trastorno límite	Cuantitativa Discreta	0 - 120
ANT_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de trastorno antisocial	Cuantitativa Discreta	0 - 120
ALC_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de problemas con el acohol	Cuantitativa Discreta	0 - 120
DRG_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de problemas con drogas	Cuantitativa Discreta	0 - 120
AGR_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de agresión	Cuantitativa Discreta	0 - 120
SUI_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de ideación suicida	Cuantitativa Discreta	0 - 120
EST_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de estrés	Cuantitativa Discreta	0 - 120
FAS_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de falta de apoyo social	Cuantitativa Discreta	0 - 120
RTR_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de relación con el tratamiento	Cuantitativa Discreta	0 - 120

Nombre	**Descripción**	**Tipo**	**Rango de Variación**
DOM_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de dominancia	Cuantitativa Discreta	0 - 120
AFA_PD	Puntaje directo de la prueba de personalidad conocida como PAI en el ámbito de afabilidad	Cuantitativa Discreta	0 - 120
PROM_S1	Nota promedio del estudiante en el primer semestre	Cuantitativa Discreta	0.0 - 5.0
PROM_S2	Nota promedio del estudiante en el segundo semestre	Cuantitativa Discreta	0.0 - 5.0
PROM_S3	Nota promedio del estudiante en el tercer semestre	Cuantitativa Discreta	0.0 - 5.0
PROM_S4	Nota promedio del estudiante en el cuarto semestre	Cuantitativa Discreta	0.0 - 5.0
EST_CIVIL	Estado civil del estudiante Soltero(1), Casado(2) o Union Libre(3)	Cualitativa Nominal	1 - 3
CIU_NACIM	Ciudad de Nacimiento del estudiante, que se representa como: Otros(0), Bogota(1) y Extranjero(2)	Cualitativa Nominal	0 - 2
LOCALIDAD	Localidad de la ciudad de Bogotá donde está el Estudiante; cada localidad es representada por su número correspondiente a acepción de Sumapaz la cual no se encuentra dentro de la base de datos, por lo tanto el número 20 corresponde a otro.	Cualitativa Nominal	1 - 20

Nombre	**Descripción**	**Tipo**	**Rango de Variación**
SIT_CARRERA	Situación académica del estudiante que puede ser: Normal(1), Graduado(2), Pérdida Cal. Académica(3), Pérdida Cal. Voluntaria(4), Prueba Académica(5), Admitido(6) y Reserva de Cupo(7)	Cualitativa Nominal	1 - 7
RECURSO	Recursos con los cuales el estudiante hace pago de su matricula los cuales pueden ser: Recursos Propios(1), Icetex(2), Préstamo Entidad(3), Auxilios Empresariales(4), Ser Pilo Paga(5) y Otros(6).	Cualitativa Nominal	1 - 6
TIEMP_PERM	Tiempo de permanecía en años que el estudiante aporta al estudio dentro del intervalo de tiempo de 2013-1 a 2017-2	Cuantitativa Discreta	0.0 - 5.0

Anexo 1: Diccionario de Datos.