ANALITICA DE DATOS APLICADA AL ESTUDIO DE DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD PEDAGÓGICA Y TECNOLÓGICA DE COLOMBIA – UPTC

DATA ANALITICS APPLIED TO THE STUDY OF STUDENT DROPOUT AT THE PEDAGOGICAL AND TECHNOLOGICAL UNIVERSITY OF COLOMBIA – UPTC

Marco Suarez-Barón[1]
Carolina Tinjaca Cristancho[2]
Juan González-Sanabria[3]

Resumen

Este trabajo presenta la aplicación de técnicas de ciencia de datos orientada a la predicción de patrones de deserción estudiantil cuyo caso de estudio corresponde a información estructurada en la UPTC seccional-Duitama. En la aplicación de la ciencia de datos se aplicaró algoritmos especializados para el desarrollo de modelos de predicción y se hace uso del análisis de datos. Adicionalmente, se estructuró un conjunto de datos cuyo contenido ha sido preparado para ser entrenado. El resultado final de la investigación presenta un modelo predictivo obtenido por medio de técnicas de ciencia de datos y que fue validado por varias métricas de calidad que evidencian la calidad del modelo final obtenido.

Palabras clave: Análisis social, anotación semántica, API, Indexación de información

Abstract

This work presents the application of data science techniques aimed at the prediction of student dropout patterns whose case study corresponds to structured information in the sectional UPTC-Duitama. In the application of data science, specialized algorithms were applied for the development of prediction models and data analysis is used. Additionally, a data set was structured whose content has been prepared to be trained. The final result of the research presents a predictive model obtained by means of data science techniques and that was validated by several quality metrics that show the quality of the final model obtained.

Keywords: Social analysis, semantic annotation, API, Information indexing

Fecha de recepción: Junio de 2020 / Fecha de aceptación en forma revisada: Septiembre de 2020

[1] PhD en Planeación Estratégica y Dirección de Tecnología. Estancias Posdoctorales en Analítica de datos y análisis social. Investigador Asociado I de Colciencias. Profesor de tiempo completo en la Escuela de Ingeniería de Sistemas y Computación de la Universidad Pedagógica y Tecnológica de Colombia – UPTC. Boyacá-Colombia. Email: marco.suarez@uptc.edu.co. ORCID: https://orcid.org/0000-0003-1656-4452

[2] Ingeniero de Sistemas. Universidad Pedagógica y Tecnológica de Colombia – UPTC. Boyacá – Colombia. Email: carolina2.crist@uptc.edu.co. ORCID: https://orcid.org/0000-0002-2881-2890

[3] Magister en Ingeniería de Software. Docente en la Universidad Pedagógica y Tecnológica de Colombia – UPTC. Boyacá – Colombia. Email: juansebastian.gonzalez@uptc.edu.co. ORCID: https://orcid.org/0000-0002-1024-6077

285

Introduction

La deserción académica en las instituciones de educación superior en Latinoamérica ha venido en un constante crecimiento, y Colombia no es ajeno a la problemática, lo cual lo evidencia el Banco Mundial en su informe "Momento decisivo: La educación superior en América Latina y el Caribe" posiciona al país como el segundo de América Latina con mayor tasa de deserción en educación superior (Figura 1), estimando que el 42% de los estudiantes que ingresan deserten en los primeros años de la formación (Ferreyra, Álvarez, Paz, & Urzúa, 2017). Lo anterior, ha llevado al Ministerio de Educación Nacional-MEN, a buscar estrategias para mitigar el impacto de deserción en el país mediante la implementación de un sistema de monitoreo semestral para evaluar y controlar la deserción estudiantil (Mubarak, Cao, & Zhang, 2020).

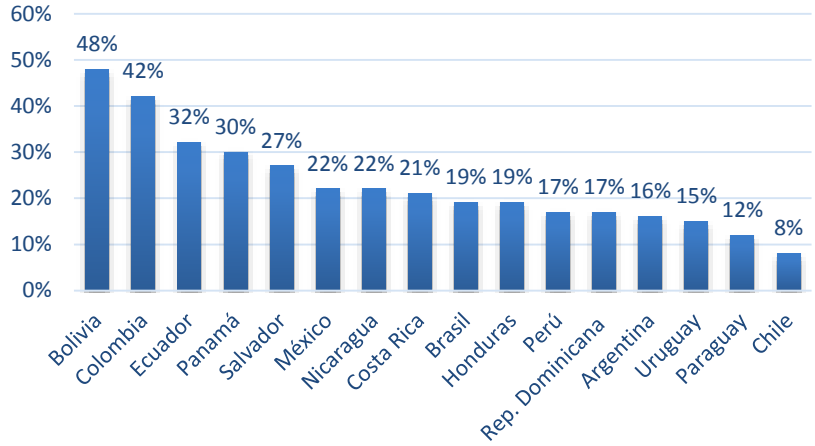

Figura 1. *Deserción universitaria en Latinoamérica. Adaptada (Ferreyra, Álvarez, Paz, & Urzúa, 2017)*

Es de aclarar, que en la deserción influyen diversos factores, entre ellos resalta el aspecto económico, el cual ha venido influyendo desde mediados de la década de los años 90, pues el estado y los sistemas financieros públicos y privados han enfrentado este problema con diferentes estrategias (Castrillón-Gómez, Sarache, & Ruiz-Herrera, 2020), relacionadas con el crédito y financiación de estudios superiores sin que ello haya impactado positivamente el fenómeno de baja cobertura y alta deserción en la educación superior como un problema de equidad social y un desafío ético para la sociedad en general (Bedregal-Alpaca, Cornejo-Aparicio, & Zárate-Valderrama, 2020).

El MEN a través del CEDE (Centro de Estudios sobre el Desarrollo Económico) de la Universidad de los Andes, realizó un proyecto el cual actualmente se encuentra vigente en donde se elabora un programa de seguimiento individualizado a cada estudiante que ingresa a la universidad desde el enfoque de riesgo y de prevención de la deserción, este proyecto no tuvo avance en la formulación comprensiva ni las argumentaciones discursivas como fenómeno social (Carvajal, Gonzalez, & Sarzosa, 2018).

Adicionalmente, no existe en el país suficientes evidencias ni bases de datos disponibles para realizar una comparación respecto a factores y tamaño de la deserción estudiantil (López,

Tulcan, 2018), la ausencia de reglas generales para mitigar el abandono por parte del estudiantado ha obligado a las IES en construir su propia fuente de análisis al fenómeno, generando una apertura de enfoques metodológicos que indican la importancia de las bases de datos en las instituciones, como necesidad a una contextualización adecuada a través de investigaciones cualitativas que permitan una ampliación como campo investigativo, mientras el Estado intenta normalizar los indicadores de deserción a nivel nacional (Isaza, Lubert, & Montoya, 2016).

En otros países, como Chile se han realizado estudios acerca de la deserción académica, pues este tipo de investigación es escaso y la magnitud del fenómeno es alarmante, para lo que se decidió desarrollar un modelo de análisis haciendo énfasis a las variables económicas, psicológicas, sociológicos, organizacionales o aspectos de las interacciones entre el estudiante y la institución, de tal forma que estos resultan predictivos al abandono estudiantil. Para ello se efectuaron diversos conjuntos de modelos que permiten analizar los factores que resultaban más predictivos dentro de las universidades (Himmel, 2018). Tambien si tienen casos mucho mas cerca como es el del instituto Tecnologico Superior de Misantla-Mexico, quienes realizaron un estudio comparativo de algoritmos para crear un sistema de predicción de desersion academica dentro de sus estudiantes (Hernandez et al., 2016), donde demuestran que una regresión logística puede ser una de las técnicas más adecuadas para realizar este tipo de análisis de datos.

Bajo este panorama, se debe recurrir a técnicas ya conocidas a nivel de muchas áreas del conocimiento, como la minería de datos (Lu, Huang & Yang, 2018), la cual se ha aplicado para el desarrollo de modelos predictivos en entornos educativos, financieros, sociales, entre otros. Para crear estrategias que ayuden a la implementación de modelos y métodos que controlen la deserción académica dentro de las instituciones de educación superior en Colombia, se toma como modelo de prueba el caso de la Universidad Pedagógica y Tecnológica de Colombia (Higuera-Martinez, 2017), seccional Duitama, donde no se cuenta con un sistema de información o registro físico que permita controlar y hacer seguimiento a los estudiantes que desertan académicamente dentro de cada programa (Guerrero, 2016).

Por lo anterior, se planteó el desarrollo de un modelo predictivo basado en datos recopilados de estudiantes en deserción académica bajo la experiencia de (Núñez-Naranjo, Ayala-Chauvin, Riba-Sanmartí, 2021), cuyo caso de estudio se da en la Universidad Pedagógica y Tecnológica de Colombia UPTC - Sede Duitama, para luego aplicar conceptos y técnicas de minería de datos, en el que se describen las metodologías, técnicas y herramientas utilizadas. Por último, se muestran los resultados obtenidos a lo largo de la investigación para inferir las conclusiones y posibles recomendaciones acerca de los factores más relevantes encontrados en cada programa académico.

Metodología

Para el desarrollo del algoritmo predictivo se plantea una metodología CRISP dividida las siguientes fases (Martínez & Mateus, 2020): Comprensión del problema, comprensión de los datos, preparación de los datos, modelado y evaluación; como se muestra en la Tabla 1.

Tabla 1.
Fases aplicadas para el desarrollo del modelo predictivo

Fase	Descripción
Comprensión de los datos	Comprender y verificar la consistencia de la información obtenida.

Preparación de los datos	Aplicar técnicas de refinamiento de información para crear un plan de contingencia para la incoherencia en las tablas de información.
Modelado	Desarrollo del modelo de predicción en base a las necesidades del problema.
Evaluación	Realizar un plan de revisión para validar el perfecto funcionamiento del modelo de predicción desarrollado.

A. Comprensión de los datos

Los datos usados para la estructuración del dataset fue requerida a la oficina de registro académico en la Sede Duitama, solicitando aquellos casos que presentaran abandono estudiantil entre el primer semestre de 2010 y primer semestre de 2018 sin importar la modalidad y jornada de todos los estudiantes de la Sede; para ello, se realiza una preparación, refinamiento y filtrado a los datos, para solucionar información errónea como: la existencia de fechas de nacimiento de 1900, detectando la incompatibilidad en el formato de fechas, lo que puede generar problemas a la hora del análisis de la información y la ejecución de las predicciones (Gómez, 2018).

Para ofrecer una clara disposición de la información adquirida, se plantea la metodología que muestra la Figura 2, donde se especifica paso a paso como fue la recolección de la información hasta la estructuración del dataset. Como resultado se obtiene un archivo delimitado por comas (csv) con el fin de lograr una mayor flexibilidad al momento de exportar datos, con un total de 1650 instancias con 18 atributos, cabe resaltar que estos datos no poseen ningún tipo de información categorizada como vulnerable.

Figura 2. *Recolección, preparación y limpieza de datos.*

B. Preparación de los datos

Para obtener un análisis exploratorio de datos conciso se usó la herramienta RStudio, para lograr un resultado más refinado y robusto. A la hora del refinamiento de información y preparación de los datos, se realizaron funciones tales como: cambio de tipo de variable de cada uno de los atributos obtenidos, ya que el formato original no era el adecuado para la manipulación estadística ni representación gráfica. Por otro lado, se realizaron los respectivos cálculos de edad para cada estudiante, la agrupación de algunos datos fue modificadas debido a que los 1650 datos se encontraban en un dataframe (matriz) lo que impedía aplicar funciones propias de R, lo que llevo a convertirlos en una lista de valores (vector). Por último, se ejecuta un cambio de variables cualitativas a cuantitativas, en el que se cambian y cortan las fechas de nacimiento, ingreso y retiro extrayendo únicamente el año, con el fin de obtener información concreta acerca del periodo de permanencia de cada estudiante para promediar los datos y manifestar un resultado final.

Al ejecutar el proceso de refinamiento y clasificación, se aplican técnicas y funciones básicas de estadística descriptiva, agrupamiento y ordenamiento de datos con el propósito de obtener una idea general de cada variable, así como una información general del dataset.

C. Modelado

La minería de datos posee diferentes técnicas para caracterizar métodos de clasificación y agrupamiento (Ahuja & Kankane, 2018), los cuáles son de vital importancia para la elaboración del modelo predictivo, cada una se ajusta según su necesidad o tipo de información obtenida, ya que no se aplican los mismos algoritmos a una variable discreta, continua o categórica para el desarrollo de un modelo predictivo (Timaran & Caicedo, 2017). Para la ejecución, se seleccionan tres técnicas de minería de datos las: árboles de decisión, clusters, teorema de Bayes y regresión logística. Cada una de estas metodologías, cuentan con el 70% de los datos como insumos de entrenamiento y el 30% restante serán los datos de prueba pues al realizar una predicción no es aconsejable tomar el 100% de los datos ya que la información de prueba busca mejorar la deducción de base, debido a que poseen comportamientos similares.

D. Evaluación

Con base a las técnicas mencionadas anteriormente se inicia el periodo de prueba con cada uno de los algoritmos de clasificación seleccionados, aplicando la librería ROC la cual ilustra la especificidad y la sensibilidad en cada uno de los puntos obtenidos, estos diagramas se ejecutan sobre la comparación de variables dicotómicas a partir de casos positivos o casos negativos, en donde su principal objetivo es demostrar los niveles de predicción dadas por el área bajo la curva, en donde el área que sea superior a 70 % será el modelo predictivo que mejor se acople al problema planteado, por otro lado, se ejecutarán matrices de confusión para observar las predicciones vs el modelo actual, además se evaluaran los resultados de las métricas previamente mencionadas, realizando una comparación entre ellas para así elegir la más robusta.

Para seleccionar el algoritmo que más se aproxime a un resultado eficiente sobre el modelo predictivo, se plantea el proceso de evaluación mostrado en la Figura 3, donde se ejecutan funciones métricas suministradas por R a cada una de las técnicas nombradas, midiendo el porcentaje de eficiencia que cada una de ellas presente respecto a los otros algoritmos, las funciones ofrecen resultados como ajuste, porcentaje de asertividad, exactitud, puntajes de predicción y comportamiento al momento de agrupar y clasificar los datos recolectados de la Sede Duitama con base a estos efectos, se extraen aquellos de mejor puntaje para realizar el modelo predictivo.

Figura 3. *Proceso de evaluación. Adaptado (Boehmke, & Greenwell, 2019).*

Desarrollo

A. Conjunto de datos

Con la información recolectada desde archivos de texto y presentación hasta gran cantidad de datos en archivos .csv o archivos planos no estructurados. Se comienza con la definición o estructuración de cada una de las columnas que se manejan dentro del dataset a construir, obteniendo cada una de las tupas como un caso de deserción académica, la estructuración final se muestra en la Tabla 2, donde se observa que con el estudio de algunas columnas podemos identificar problemas con la movilidad, jornada y modalidad de estudio y problemas relacionados con el proceso académico.

Tabla 2.
Estructuración del dataset.

Nombre	Descripción	Tipo
Id_caso	Identificación del caso de deserción.	Numero
Nombre_facultad	Nombre de la facultad a la que pertenece el caso de deserción	Texto
Fecha_nacimiento	Fecha de nacimiento del estudiante que presento el caso.	Date
Ciudad_origen	Ciudad de origen del estudiante que presento el caso.	Texto
Ciudad_residencia	Ciudad de residencia del estudiante que presento el caso.	Texto
Sexo	Sexo del estudiante que presento el caso.	Texto
Id_Programa	Identificador programa al que pertenecía el estudiante que presento el caso	
Programa	Programa al que pertenecía el estudiante que presento el caso.	Texto
Modalidad	Modalidad del programa al que pertenecía el estudiante que presento el caso.	Texto
Jornada	Jornada del programa al que pertenecía el estudiante que presento el caso.	Texto
Id_estado	Identificador del estado de deserción,	Numero
Nombre_estado	Descripción o nombre del estado de deserción.	Texto
Id_periodo_ingreso	Identificador del periodo en el que ingreso el estudiante.	Numero
Descripcion_periodo	Descripción o nombre del periodo en que ingreso el estudiante.	Texto
Id_periodo_final	Identificador del periodo en el que se retiró el estudiante.	Numero
Descripcion_periodo	Descripción o nombre del periodo en el que se retiró el estudiante.	Texto
Promedio_semestre_final	Promedio del semestre en el que se reportó el retiro	Decimal
Promedio_acumulado	Promedio final acumulado presentado al realizar el retiro	Decimal

Al estructurar y almacenar cada una de las tuplas encontradas con la recolección de información, se inicia la etapa de refinación para obtener un dataset de datos de calidad y completos, debido que uno de los problemas más comunes de aplicación de minería de datos y machine learning, es la calidad de los datos analizados, teniendo en cuenta que pueden tener problemas sintácticos y lingüísticos, que generan errores de coherencia y emparejamiento con los algoritmos de predicción y analítica. Las tareas realizadas en esta fase partieron desde la limpieza de datos vacíos y el manejo de herramientas de corrección de los datos para evitar tener diferentes grupos de información que se relacionen con el mismo objetivo (Pérez, Ramos, & Mejía, 2018).

B. Análisis exploratorio de datos

Para la ejecución de técnicas de predicción a partir de minería de datos, es recomendable hacer primero una exploración adecuada de la muestra, y así detectar que parámetros son los que más influyen para que un estudiante llegue a declararse en deserción; por lo anterior, se plantearon los algunos casos donde se demuestran puntualmente algunas circunstancias de deserción.

Como medida inicial se detento que nivel de deserción tiene cada una de las carreras que se prestan en la UPTC – Seccional Duitama, como lo muestra la Figura 4, donde en los últimos 8 años de academia la Ingeniería Electromecánica con 54% de deserción y Administración de Empresas Agropecuarias con 68%, obtienen el mayor índice de deserción académica de la sede, teniendo cada una un total de 16% y 15% de abandono respecto a los demás programas académicos.

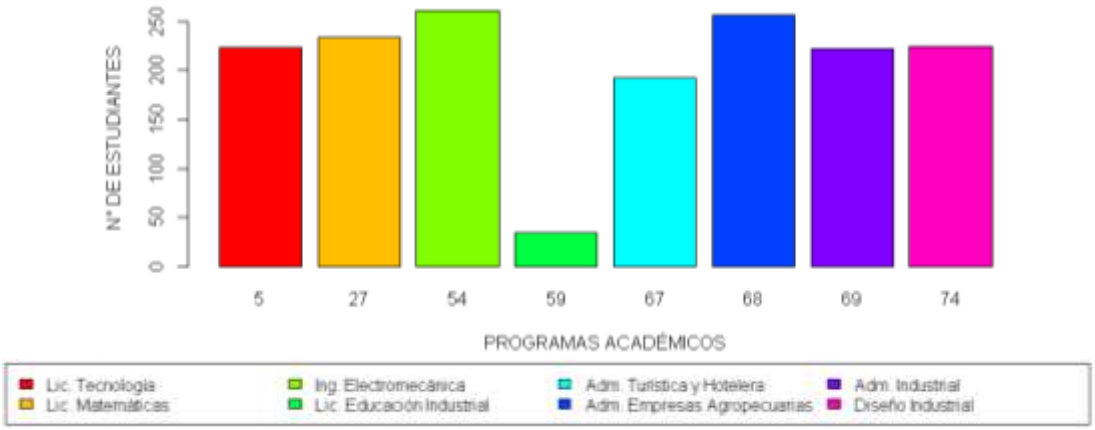

Figura 4. *Porcentaje de deserción por programa académico.*

Uno de los factores que se detectaron en la exploración fue el rango de edades que más presentan deserción académica como se muestra en la Figura 5, obteniendo que la mayoría de los estudiantes que se retiraron tienen entre los 18 y 21 años, teniendo como edad mínima 15 años y máxima de 39 años, la población presenta una media y mediana de 20 años, en donde el 25% de los estudiantes desertores tenían 18 años y el 75% tenían 22.

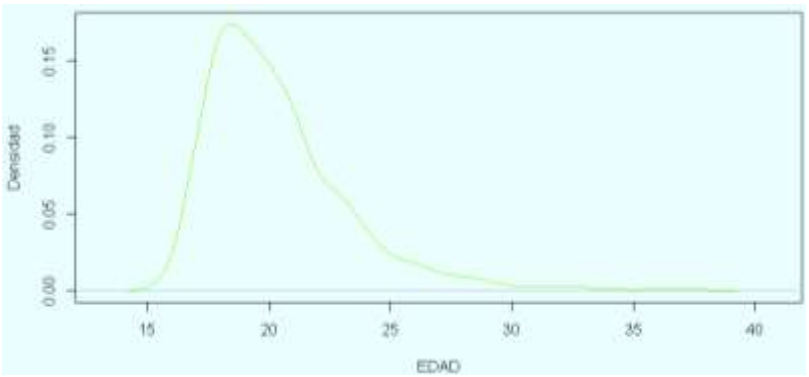

Figura 5. *Deserción académica por rango de edad.*

En algunos casos académicos como (Barberá, 2017) han demostrado que el género juega un papel importante a la hora de demostrar diferentes habilidades necesarias para el progreso positivo dentro de un programa académico, por esta razón se detectaron se plantea el análisis de desertores con relación al género como se puede ver en la Figura 6, donde es visible que los hombres han mostrado un mayor grado de deserción con un total de 1036 estudiantes retirados y las mujeres un total de 614 estudiantes, Administración de Empresas Agropecuarias (68) y Administración Turística y Hotelera (67) son los únicos programas que presentan mayor abandono por parte del género femenino cada uno de ellos presentando un total de mujeres de 122 y 126 equivalentes a un 20% y 21%.

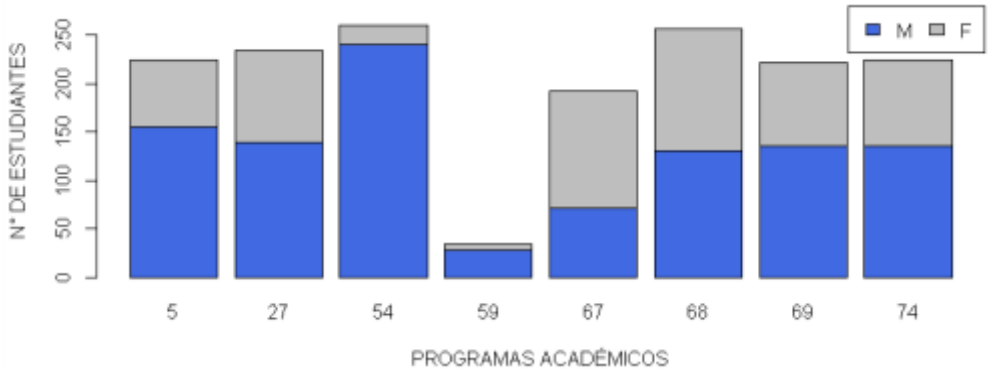

Figura 6. *Deserción académica con relación al género de cada uno de los programas académicos.*

Una de las principales razones de abandono por parte de los estudiantes sucede por ámbitos académicos en el cual se presentan los siguientes casos de deserción según el reglamento estudiantil de la UPTC, donde se establece que el estudiante perderá el cupo por:

a) Retirado con cupo reservado (Id 2)
b) Retirado definitivamente (Id 3)
c) No matriculado (Id 8)
d) Promedio acumulado inferior a 3.0 (Id 17)
e) Promedio semestral y acumulado inferior a 3.0 (Id 18)
f) Materia perdida por 2da vez con promedio inferior a 3.0 (Id 19)
g) Materias acumuladas perdidas (Id 20)
h) Materia perdida 3 veces (Id 21)

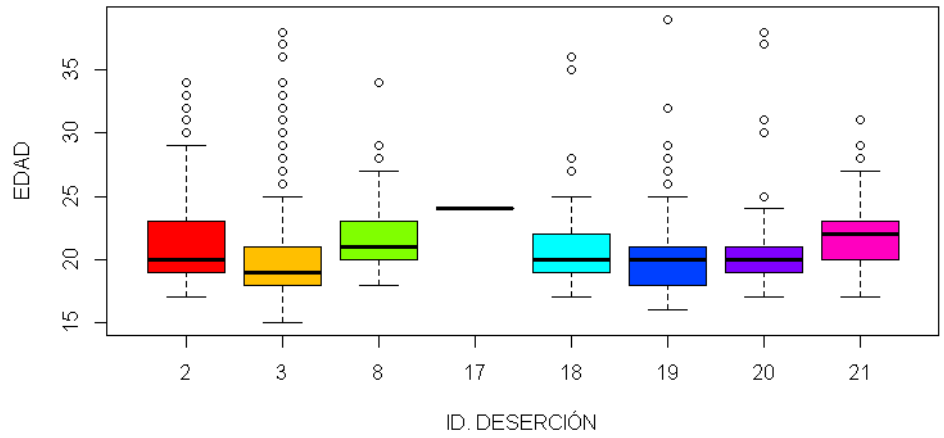

Figura 7. *Deserción derivada de problemas académicos.*

Los gráficos de cajas y bigotes permiten observar la simetría de los datos, en este caso, el literal D es el único que presenta datos simétricos debido a que la mediana se encuentra en el centro de la caja. Con base al Figura 7, el literal que mayor número de estudiantes presenta es retirado definitivamente con un total de 729 estudiantes, obteniendo una mediana de 19 años, en el cual, el 25% se retiró a una edad de 18 años mientras que el 75% se retira a los 21; con base a la edad en la que ésta muestra deserta se puede indicar que el estudiantado ingresa sin estar seguro de haber elegido el programa académico o universidad correcta, ya que es una edad en la que no se sabe realmente que se quiere.

El motivo (a) presenta un total de 299 estudiantes, con una mediana de 23 años, este factor fue a causa de una(s) materia(s) con una nota mínima lo que ocasiona un promedio bajo por 3 semestres consecutivos lo que lleva a la expulsión inmediata del estudiante.

C. Caracterización de métodos de clasificación

Para generar un modelo predictivo robusto y conciso para la deserción académica en la Sede Duitama, se debe realizar un procedimiento de caracterización de métodos para la clasificación de la información obtenida (Perez et al., 2018), para así ayudar a mitigar el fenómeno que en los últimos años se ha venido presentado de manera abrupta y que hasta el día de hoy no se cuenta ni dimensiona la magnitud de este. A continuación, se evidencian los resultados de algoritmos utilizados para el proceso de caracterización, para estas operaciones sólo se emplearon aquellas variables que poseen mayor impacto en temáticas de deserción, es decir, el género, programa académico, promedio semestral y acumulado, jornada, región de origen, tiempo cursado y edad.

1. Arboles de decisión

Para la aplicación y ejecución del árbol de decisión se tuvo en cuenta el uso del 30% de la dataframe procesada o datos de prueba, para así lograr el árbol mostrado en la Figura 8.

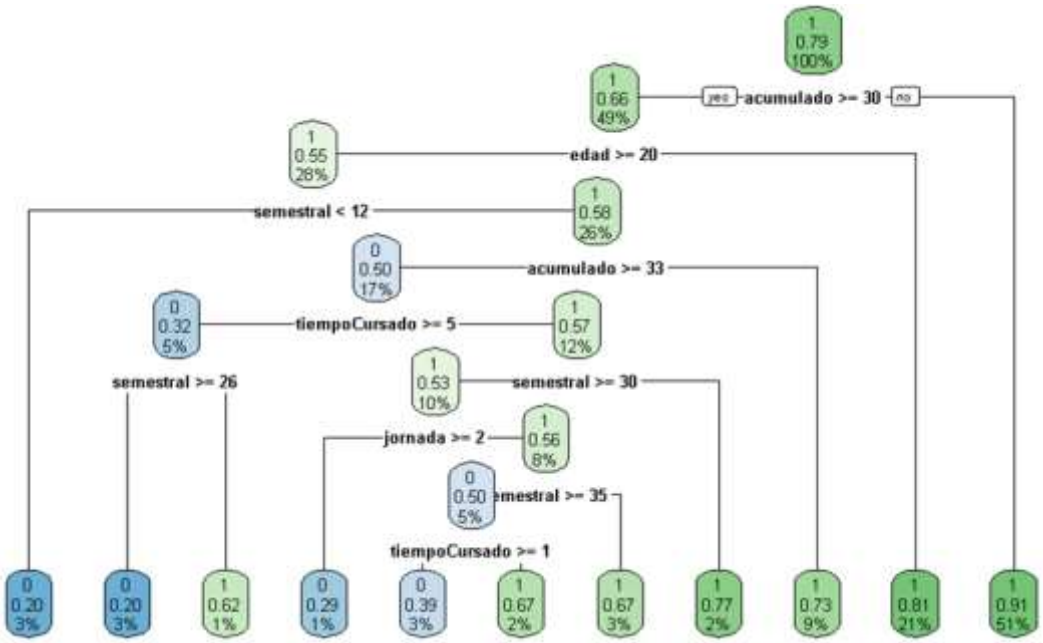

Figura 8. *Árbol de decisión ejecutado con datos de prueba.*

Para el caso de aplicación se toma como variable dependiente el estado de deserción para resaltar aquellas características que son fundamentales dentro del análisis, en donde el promedio acumulado de cada estudiante es primordial para conocer si un estudiante se encuentra en riesgo de perder el cupo, de tal forma que si el estudiante presenta un promedio inferior a 3.0 tiene una probabilidad de desertar en un 49% como se muestra en la primera ramificación (Lee et al., 2020).

Por último, el tiempo de permanencia llega a ser una variable significativa ya que la mayoría de los estudiantes que presentan un promedio semestral mayor a 3.3 cursaron más de 4 semestres, es decir, su deserción fue por motivos académicos mientras que el porcentaje restante que es mayor, permanecieron en la universidad menos de 3 semestres lo que concluye que pudieron ser ámbitos económicos o sociales (Quiñones et al., 2020).

2. Aplicación del algoritmo de Regresión Logistica.

La regresión logística fue aplicada con el método binomial mostrando que la variable dependiente es dicotómica y su comportamiento vario respecto a otro tipo de variables categóricas o continuas, en este caso se pudo observar que de llegar a implementarse el modelo predictivo se lograra proporcionar el porcentaje de deserción de cada estudiante que se inscribe dentro de la universidad. La Figura 9 demuestra que los valores son asimétricos o incluyen valores atípicos dentro del modelo predictivo, se distribuyen de igual forma a lo largo de los rangos predictores (Pérez et al., 2018).

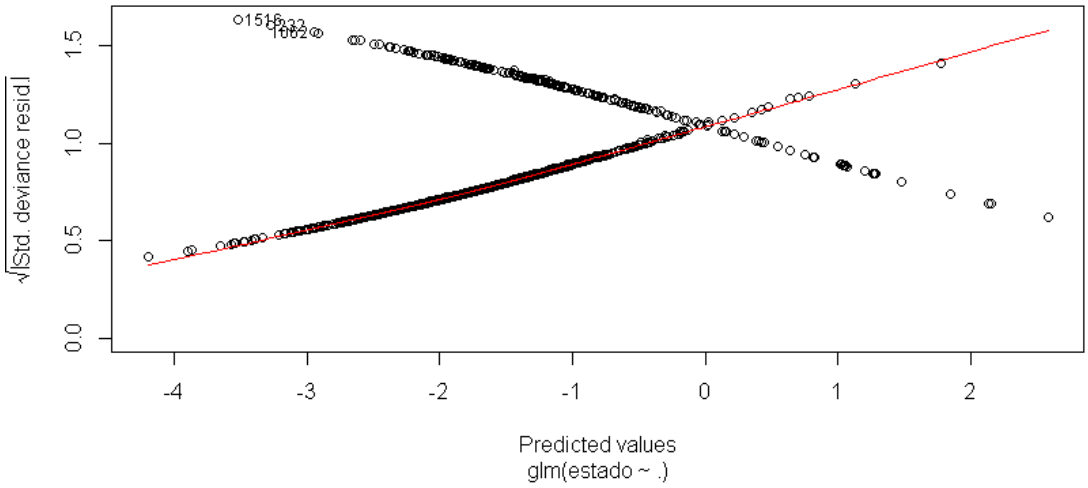

Figura 9. *Ubicación de escalada.*

Con lo anterior se comprueba la homogeneidad de los datos respecto a la varianza de cada uno de ellos la cual es constante; por otro lado, la línea exponencial infiere que las magnitudes de los promedios estandarizados varían en función con los valores ajustados.

3. Aplicación del algoritmo de Teorema de Bayes.

Para lograr predicciones basadas en la aplicación del teorema de Bayes (Lacave, Molina & Cruz, 2018), se implementaron predicciones o sucesos según eventos anteriores, con lo anterior, lograr determinar la probabilidad de éxito en futuros comportamientos evaluando efectos particulares, para obtener las probabilidades de eventos anteriores empleando la ecuación (1).

$$P(A_i \mid B) = \frac{P(B \mid A_i) * P(A_i)}{P(B)} \quad (1)$$

Resultados y discusión

A continuación, se presentarán los resultados obtenido de la ejecución y aplicación de los diferentes modelos de predicción como: Arbol de decisión, regresión logistica y teorema de Bayes, los cuales fueron analizaron teniendo en cuenta varios factores y al mismo tiempo se validó la calidad de los datos que se generaron como resultado, para poder lograr describir de una mejor manera la efectividad de cada uno de los modelos y al mismo tiempo también para tener varios factores de evaluación que ayudaran a determinar el modelo más adecuado.

El análisis exploratorio muestra que el rendimiento academico de los estudiantes es una variable altamente incidente y correlacional en el estado de predicción de deserción para un estudiante. El genero, el domicilio y los conocimientos traídos desde la educación secundaria son atributos son variables categoricas que evidencian su alta correlación. Por otro lado, las habilidades blandas yapropiación de tematicas relacionadas como las matemáticas, lenguaje y segundo idioma contribuyen en la explicación de la variación del rendimiento general. Adicionalmente, el apoyo que reciben los estudiantes por parte de sus familias, como la forma de

selección la carrera no son significativas y por consiguiente no contribuyen en el desempeño del rendimiento académico en general.

De otro lado, todos los modelos fueron evaluados implementando técnicas como: curva ROC, matriz de confusión y métricas de rendimiento exportadas de R, el resultado de cada técnica de minería de datos aplicada fue analizada detenidamente y así resaltar aquella que más se ajuste a la predicción de tendencias por las cuales los estudiantes abandonan los programas de la seccional Duitama de la UPTC,

El primer modelo de calificación a analizar es el árbol de decisión, al aplicar la matriz de confusión, en la Tabla 3, se puede observar el rendimiento del algoritmo dentro del conjunto de datos teniendo como resultado que un 75% hace parte de los verdaderos positivos mientras que los falsos negativos obtienen un 19%, lo cual determina que no es viable ya que la cantidad de verdaderos negativos no supera el número de falsos negativos (Olaya, Vásquez & Maldonado ,2020).

Tabla 3.
Matriz de confusión árbol de decisión.

	0	**1**	**SUM**
0	0.75	0.19	0.94
1	0.02	0.04	0.60
SUM	0.77	0.23	1.00

Al extraer las métricas que provee R como lo muestra la Tabla 4, el árbol de decisión obtiene una exactitud de clasificación del 80%, una identificación de elementos positivos del 80%, la métrica recall obtiene un 98% el cual indica la proporción más relevante de los resultados, es decir, que porcentaje de casos presentados son positivos y por último obtenemos la tasa de error con un 21%.

Tabla 4.
Métricas R árbol de decisión.

Exactitud	**Precisión**	**Recall**	**F1_Score**	**Tasa de Error**
0.80	0.80	0.98	0.88	0.21

Para terminar el modelo de árbol de decisión se grafica la curva ROC como lo muestra la Figura 10, obteniendo un área bajo la curva de 0.65 lo cual indica un buen nivel de predicción por parte de esta metodología ya que ofrece un mayor nivel de sensibilidad; es decir, captura una mayor cantidad de casos positivos que en nuestro caso sería un mayor nivel de predicción sobre los estudiantes que desertan.

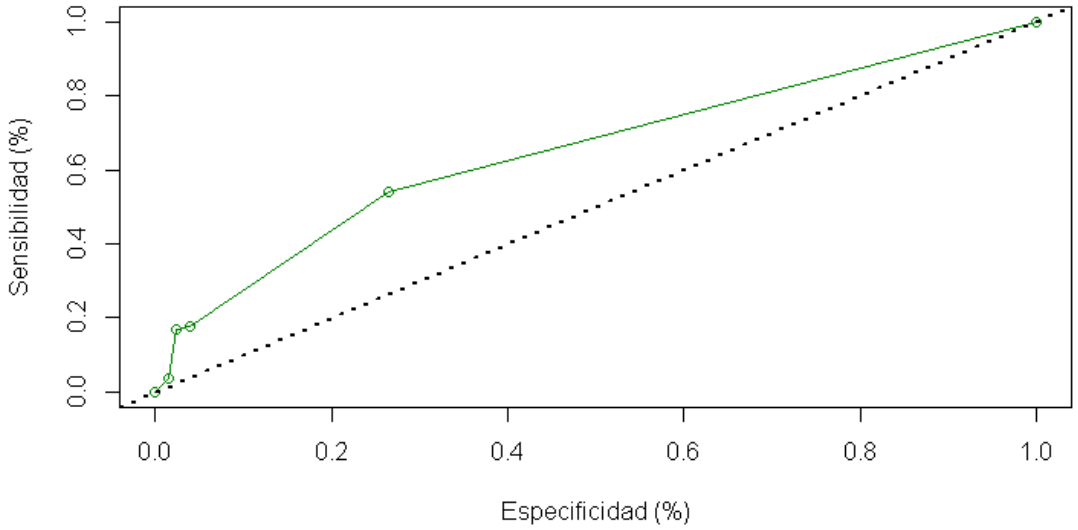

Figura 10. *Curva ROC árbol de decisión.*

El segundo modelo de calificación a analizar es la regresión LOGIT, donde al aplicar la matriz de confusión mostrada en la Tabla 5, se logra un 76% de verdaderos positivos mientras que los falsos negativos obtienen un 21%, lo cual deja en mejor posición los árboles de decisión ya que obtienen un porcentaje más bajo en aquellos estudiantes que no desertaban, pero que durante el proceso termina desertando.

Tabla 5.
Matriz de confusión regresión LOGIC.

	0	**1**	**SUM**
0	0.76	0.21	0.97
1	0.01	0.02	0.03
SUM	0.77	0.23	1.00

Las métricas exportadas por R mostradas en la Tabla 6 evidencian que la regresión LOGIC obtiene una exactitud de clasificación del 79%, una identificación de elementos positivos del 99%, una métrica recall de 78%, el cual indica la proporción más relevante de los resultados, es decir, el porcentaje de casos presentados que son positivos y por último obtenemos la tasa de error con un 22%.

Tabla 6.
Métricas R regresión LOGIC.

Exactitud	**Precisión**	**Recall**	**F1_Score**	**Tasa de Error**
0.79	0.99	0.78	0.87	0.22

Al analizar la curva ROC mostrada en la Figura 11, se observa que se obtiene una clasificación de las clases muy cercana a las mostradas en el árbol de decisión, pero no obtiene la medida de separabilidad de variables óptima.

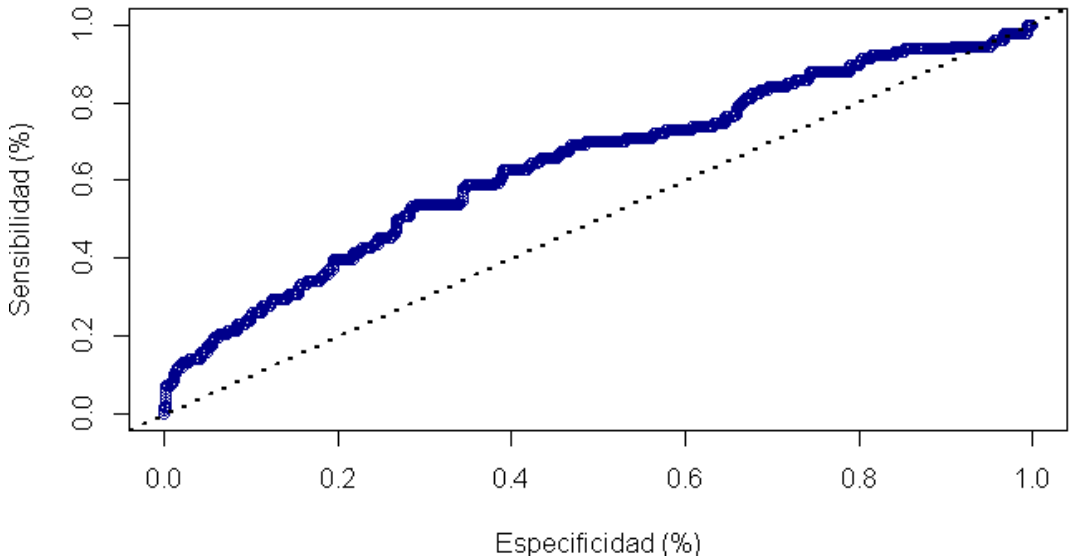

Figura 11. *Curva ROC regresión LOGIC.*

Por último, al analizar la aplicación de teorema de bayes se obtiene la matriz de confusión mostrada en la Tabla 7, desde el comienzo se detecta que no es el mejor **método** para la predicción necesaria, debido a sus probabilidades mínimas de predicción y mayor número de positivos negativos con un 18% a pesar de que los casos verdaderos positivos obtuvieron un 72%, porcentaje mucho menor a las técnicas aplicadas con anterioridad; por último, los sucesos exitosos basados en probabilidades anteriores como lo indica el teorema de Bayes obtuvo un 89%.

Tabla 7.
Matriz de confusión teorema de bayes.

	0	**1**	**SUM**
0	0.72	0.18	0.89
1	0.06	0.05	0.11
SUM	0.78	0.22	1.00

Al realizar la extracción de las métricas en la Tabla 8, en los resultados del modelo predictivo basado en teorema de bayes es evidente que la tasa de error es superior con un 0.23, una exactitud de clasificación de 0.76, una tasa de precisión de 0.80.

Tabla 8.
Métricas R teorema de bayes.

Exactitud	**Precisión**	**Recall**	**F1_Score**	**Tasa de Error**
0.76	0.80	0.92	0.86	0.24

Para finalizar el análisis de los modelos de clasificación, se analiza la curva de ROC mostrada en la Figura 12 para el teorema de bayes, donde los niveles de predicción son muy bajos ya que se obtiene un área bajo la curva inferior a 60, cuya área se superaba en modelos anteriores;

por lo tanto, su nivel de predicción no aporta los resultados que se requieren, dejando esta metodología descartada por sus resultados muy bajos en comparación a los demás.

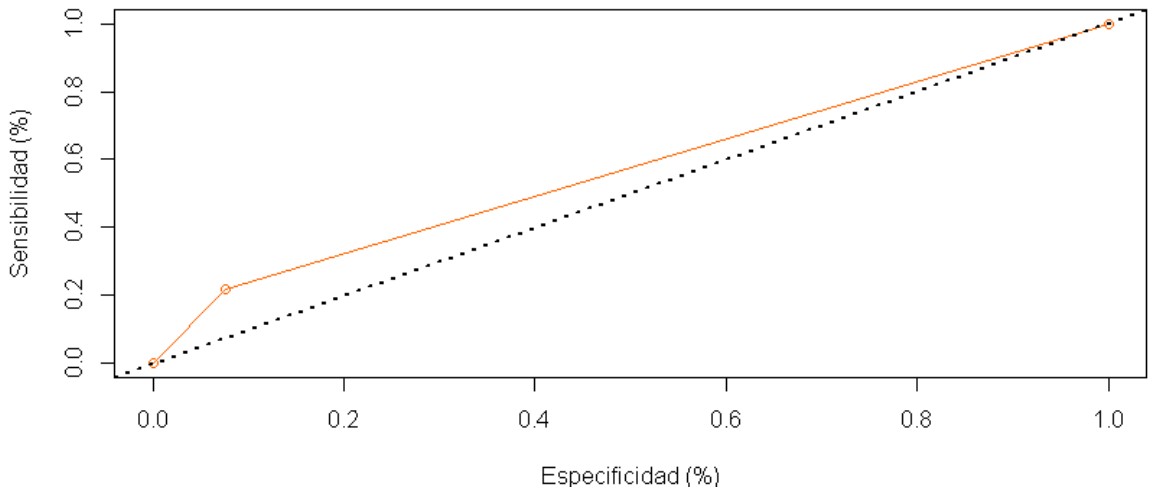

Figura 12. *Curva ROC teorema de bayes.*

Como se observa en el desarrollo y el planteamiento de los resultados de la aplicación de cada técnica de predicción, se determina que los árboles de decisión son los modelos mas eficientes para este caso de estudio, en comparación del teorema de Bayes que no cumplió con las expectativas debido a que los datos manejados no tienen una relación clara entre ellos y la ampliación del algoritmo no encontraba las mejores opciones. Cabe aclarar que la elección y la aplicación de un algoritmo de predicción como los elegidos siempre dependen de varios factores para su correcta aplicación, es por estas razones que la comparación de diferentes modelos de predicción siempre es necesaria antes de elegir el más adecuado para una implementación final.

Por lo anterior, al finalizar el análisis, categorización y verificación del modelo predictivo se registraron hallazgos que detectan que una gran parte de los estudiantes admitidos, provienen de la región andina, presentando un total de 1567 estudiantes lo cual indica que no influye mucho el lugar de procedencia. La edad con mayor índice oscila entre los 17 y 20 años, validando que una de la circunstancia es por la ausencia de vocación profesional.

Otro factor de gran impacto a tener en cuenta son aquellos estudiantes que lleven un tiempo cursado superior a cinco semestres y su promedio semestral sea inferior a 3.0, es un perfil de posible deserción como lo evidencia el análisis de resultados. Por último, la jornada extendida presenta un promedio semestral alarmante, puesto que la media de las calificaciones es de 2.0/5.0 aproximadamente, lo cual lleva a inferir que la no obligación o constante permanencia dentro de la Universidad influye en la deserción académica en programas a distancia.

Resultados como los anteriores representan la efectividad de la aplicación de modelos de predicción en conjuntos de datos trabajados en instituciones de educación superior y más en factores clave como lo es el estudio de la deserción académica. También demuestra que la correcta recolección y manipulación de los datos e información recolectada es importante para que los modelos predictivos puedan ejecutar de una la mejor manera los procesos y algoritmos correspondientes y logren generar resultados notables y de gran aplicación para futuras tomas de decisiones.

Finalmente, se demuestra que los modelos de predicción aplicados en casos relacionados con la deserción académica, son un mecanismo computación riguros y a su vez herramientas clave y útil para la toma de decisiones de los diferentes estamentos académicos; por ende, así se podrán plantear mejor estrategias y metodologías que ayuden a prevenir y mitigar estos casos que abundan en la mayoría de instituciones de educación superior.

Conclusiones

Al finalizar la investigación y algunas pruebas en los datos suministrados se logra determinar que el modelo que más se ajusta a las necesidades es el árbol de decisión ya que el resultado de sus puntos de precisión es más elevado en comparación a los demás, lo cual indica mayor exactitud. Pero al mismo tiempo, se detectó que el proceso de extracción de datos es un proceso clave para la gestión de información en la ejecución de los modelos desarrollados, debido a que su calidad en los daros puede consolidar resultados mas precisos y exactos; dado que se puede generar patrones inadecuados y poco importantes para la toma decisiones.

También se comprueba que la minería de datos es una técnica apropiada para obtener un resultado conciso sobre las principales causas de deserción que presentaba la universidad, así como los diferentes patrones y tendencias que ayudaban a clasificar y agrupar la información para obtener un mejor resultado. Todo esto lleva a refutar la necesidad de implementar mas modelos como los trabajados en los diferentes campos de acción, sean académicos, administrativos o financieros; ya que, se comprobó el alto nivel de eficiencia de las respuestas del correcto manejo e implementación de un modelo de predicción.

Cabe resaltar que la ejecución de los modelos de predicción basados en los datos recogidos no basta para convertirse en un insumo real para la toma de decisiones y la creación de estrategias que ayuden a mitigar los problemas relacionados con la deserción académica dentro de la institución. Lo anterior conlleva al siguiente paso que es promover el desarrollo de aplicaciones que implementen estos modelos y se instalen como herramienta primordial en los diferentes estamentos académicos que necesiten revisar y analizar los datos en mención de una manera mucho más legible y entendible, esto ayudará a que se cree una herramienta útil y al mismo tiempo se garantizara que las fuentes de datos se seguirán alimentando para lograr a futuro resultados mucho mas precisos.

Con el desarrollo de investigaciones como esta se espera que, a futuro, entidades tanto educativas o empresariales consideren la iniciativa de implementar y aplicar modelos predictivos que ayuden a mitigar situaciones como el de deserción académica detectado en la UPTC, creando estrategias de control para problemas detectados y como ayuda para la generación de alternativas que sirvan de insumo para la toma de decisiones.

Referencias bibliográficas

Ahuja, R., & Kankane, Y. (2018). Predicting the probability of student's degree completion by using different data mining techniques. Paper presented at the 2017 4th International Conference on Image Information Processing, ICIIP 2017, 2018-January 474-477. doi:10.1109/ICIIP.2017.8313763.

Barberá, M. (2017). Análisis de los factores asociados a la elección de estudios universitarios utilizando técnicas de agrupamiento. Universitat Politècnica de València. Escola Tècnica Superior d'Enginyeria Informàtica Universitat Politècnica de València. Departamento de Informática de Sistemas y Computadores - Departament d'Informàtica de Sistemes i Computadors.

Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zárate-Valderrama, J., & Yanque-Churo, P. (2020). Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students. Journal of Advanced Computer Science and Applications (IJACSA), 11(1).

Boehmke, B., & Greenwell, B. M. (2019). Hands-on machine learning with R. CRC Press.

Castrillón-Gómez, O., Sarache, W., & Ruiz-Herrera, S. (2020). Predicción de las principales variables que conllevan al abandono estudiantil por medio de técnicas de minería de datos. Formación universitaria, 13(6), 217-228. https://dx.doi.org/10.4067/S0718-50062020000600217.

Carvajal, C. M., González, J. A., & Sarzoza, S. J. (2018). Variables sociodemográficas y académicas explicativas de la deserción de estudiantes en la Facultad de Ciencias Naturales de la Universidad de Playa Ancha (Chile). Formación universitaria, 11(2), 3-12.

Ferreyra, M. M., Álvarez, J., Paz Haimovich, F., & Urzúa, S. (2017). Momento decisivo: la educación superior en América Latina y el Caribe.

Gómez, J. G. L. (2018). Comparación en el porcentaje de deserción entre alumnos de preparatorias públicas y privadas que ingresaron en el ciclo escolar 2015-2016 a la UASLP. Revista educativa, 4(11), 1-14.

Guerrero, S. C. (2016). Estimación y estrategias sobre el abandono en la educación superior en la Universidad Pedagógica y Tecnológica de Colombia. Congresos CLABES.

Hernandez Gonzalez, A. G., Melendez Armenta, R. A., Morales Rosales, L. A., Garcia Barrientos, A., Tecpanecatl Xihuitl, J. L., & Algredo, I. (2016). Comparative study of algorithms to predict the desertion in the students at the ITSM-Mexico. IEEE Latin America Transactions, 14(11), 4573-4578. doi:10.1109/TLA.2016.7795831.

Higuera Martínez, O. I. (2017). Deserción estudiantil en Colombia y los programas de Ingeniería de la UPTC Seccional Sogamoso. Revista Ingeniería Investigación y Desarrollo; Vol. 17, núm 1 (2017).

Himmel, E. (2018). Modelo de análisis de la deserción estudiantil en la educación superior. Calidad en la Educación, (17), 91-108. doi: https://doi.org/10.31619/caledu.n17.409

Isaza, L. G., Lubert, C. D., & Montoya, D. M. (2016). Caracterización de la deserción estudiantil en la universidad de caldas el período 2009-2013. análisis a partir del sistema para la prevención de la deserción de la educación superior –spadies. Latinoamérica de Estudios, pp. 132-158.

Lacave, C., Molina, A. I., & Cruz-Lemus, J. A. (2018). Learning analytics to identify dropout factors of computer science studies through bayesian networks. Behaviour and Information Technology, 37(10-11), 993-1007. doi:10.1080/0144929X.2018.1485053.

Lee, L. E., Martínez, S. I., Rocha, J. A. C., Villanueva, J. D. T., Menchaca, J. L., Berrones, M. G. T., & Rocha, E. C. (2020). Evaluation of Prediction Algorithms in the Student Dropout Problem. Journal of Computer and Communications, 8(03), 20.

López Cerón, A. N., & Tulcán Cuasapud, J. V. (2018). Factores que inciden en la tasa de deserción y repitencia de la carrera de nutrición y salud comunitaria de la Universidad Técnica del Norte en el periodo 2009-2017.

Lu, O. H. T., Huang, A. Y. Q., & Yang, S. J. H. (2018). Benchmarking and tuning regression algorithms on predicting students'academic performance. Paper presented at the ICCE 2018 - 26th International Conference on Computers in Education, Workshop Proceedings, 477-486.

Martínez, J. C., & Mateus, S. P. (2020). Propuesta de un Modelo Predictivo utilizando Aprendizaje Profundo para el análisis de deserción estudiantil en Universidades Colombianas Virtuales. Revista Innovación Digital y Desarrollo Sostenible-IDS, 1(1), 51-57.

Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. Interactive Learning Environments, 1-20.

Núñez-Naranjo, A. F., Ayala-Chauvin, M., & Riba-Sanmartí, G. (2021). Prediction of University Dropout Using Machine Learning. In International Conference on Information Technology & Systems (pp. 396-406). Springer, Cham

Olaya, D., Vásquez, J., Maldonado, S. (2020). Uplift Modeling for preventing student dropout in higher education. Decision Support Systems, 134, 113320.

Pérez, A., Grandón, E. E., Caniupán, M., & Vargas, G. (2018). Comparative analysis of prediction techniques to determine student dropout: Logistic regression vs decision trees. Paper presented at the Proceedings - International Conference of the Chilean Computer Science Society, SCCC, 2018-November doi:10.1109/SCCC.2018.8705262

Pérez, M. A., Ramos, M. B., & Mejía, C. S. (2018). Estudio sobre la deserción estudiantil universitaria y sus implicaciones académicas, económicas y sociales. Bolentín de Coyuntura, (19), 9-13.

Quiñones, L., Jara, D. M., Carrasco, N. A., Pino, M. M., & Gamarra, O. (2020). Modelo para la estimación de la deserción estudiantil Awajún y Wampis empleando minería de datos. Revista de Ciencia y Tecnología: RECyT, 34(1), 45-50.

Timaran Pereira, R., & Caicedo Zambrano, J. (2017). Application of decision trees for detection of student dropout profiles. Paper presented at the Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, 2017-December 528-531. doi:10.1109/ICMLA.2017.0-107.