Vol. 19, n.° 1, pp. 193-204, 2020
Revista UIS Ingenierías
Página de la revista: revistas.uis.edu.co/index.php/revistauisingenierias

Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico

Comparison of data mining techniques to identify signs of student desertion, based on academic performance

Boris Rainiero Pérez-Gutiérrez

Departamento de Sistemas e Informática, Universidad Francisco de Paula Santander, Cúcuta, Colombia.
Orcid: 0000-0001-9249-1756. Correo electrónico: borisperezg@ufps.edu.co

Recibido: 15 febrero, 2019. Aceptado: 12 agosto, 2019. Versión final: 30 diciembre, 2019.

Resumen

Uno de los grandes retos en las instituciones educativas consiste en poder establecer la posibilidad de retiro o deserción de sus estudiantes. En este artículo se presentan los resultados de un estudio de comparación de técnicas para apoyar la identificación de deserción estudiantil a partir del registro académico de los estudiantes de una Universidad en Colombia para el programa de Ingeniería de Sistemas. El registro académico se estableció para un periodo de 7 años. Árboles de decisión, regresión logística y Naive Bayes, fueron comparados para lograr establecer la mejor técnica de detección de desertores. Adicionalmente, la herramienta Watson Analytics de IBM fue utilizada para comparar su usabilidad y precisión para un usuario no experto. Nuestra experimentación demostró que el uso de algoritmos simples es suficiente para alcanzar niveles ideales de precisión. Estos resultados son presentados a la comunidad académica para ayudar en la disminución de la deserción estudiantil.

Palabras clave: datos estudiantiles; educación superior; minería de datos; modelos de predicción; deserción

Abstract

One of the great challenges in educational institutions is to be able to establish the possibility of retirement or desertion of their students. This article presents the results of a comparative study of techniques to support the identification of student dropouts using the academic record of students at a University in Colombia for the Systems Engineering program. The academic record was established for a period of 7 years. Decision trees, logistic regression, and Naive Bayes were compared to establish the best dropout detection technique. Additionally, IBM's Watson Analytics tool was used to compare its usability and accuracy to a non-expert user. Our experience has shown that the use of simple algorithms is sufficient to achieve ideal levels of accuracy. These results are presented to the academic community to help decrease student dropout.

Keywords: student data; higher education; data mining; prediction models; dropout.

1. Introducción

De forma globalizada, la educación está asociada al rápido crecimiento económico de un país [1]. Las Universidades juegan un papel muy importante como formadores de capital humano cualificado y como sistemas de innovación en un país [2]. Mejorar el capital humano requiere de estudiantes que completen sus

ISSN impreso: 1657 - 4583. ISSN en línea: 2145 - 8456,
B. Pérez, "Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico," *Rev. UIS Ing.*, vol. 19, no. 1, pp. 193-204, 2020. doi: 10.18273/revuin.v19n1-2020018

estudios, y de un alto nivel de educación. La formación de profesionales está correlacionada con el aumento de la esperanza de vida, el aumento de la condición social, la reducción del riesgo de desempleo, entre otros [3].

Por parte de las universidades, es claro el interés en comprender el comportamiento de los estudiantes exitosos. De acuerdo con Mishra et al. en [4], la reputación de las universidades se mide por el porcentaje de estudiantes que se gradúan. Es por esto, que la identificación temprana de los estudiantes que están en riesgo de abandonar sus estudios es crítica para el éxito de cualquier universidad. Una vez identificados, se pueden realizar distintas actividades de intervención con ellos encaminadas a reducir los niveles de abandono [5, 6].

En Colombia existe una iniciativa del Ministerio de Educación denominada SPADIES (Sistema de Prevención y Análisis de la Deserción en Instituciones de Educación Superior) [7]. Esta iniciativa fue diseñada por el Centro de Estudios Económicos (SEDE) de la Universidad de los Andes para dar seguimiento al problema de la deserción en la educación superior, calcular el riesgo de deserción de cada estudiante y clasificarlos por grupos. Esta iniciativa puede apoyar la evaluación de estrategias para cada una de las situaciones que influyen en la deserción escolar, tales como la situación de los estudiantes, el programa académico y la institución; y también promueve la consulta, consolidación, interpretación y uso de esta información.

Diversos trabajos se han presentado para enfrentar la deserción, encontrando que las características del estudiante y su contexto son fuertes influenciadores en su decisión de abandonar la universidad. Uno de ellos es el modelo de Tinto [8], el cual establece que la deserción escolar está fuertemente relacionada con el grado de integración académica (rendimiento académico y desarrollo intelectual) y social (interacciones entre pares e interacciones entre profesores) en la universidad. Bharadwaj y Pal en [9, 10] identificaron que otros factores, como la calidad en la puntuación en los exámenes, la residencia, el ingreso familiar anual y la situación familiar; son parámetros importantes para la deserción escolar. En otro trabajo, Kovacic [11] identificó que la asistencia de los estudiantes a clase, las horas de estudio después de clase, el ingreso familiar, la edad de la madre y la educación de la madre están significativamente relacionadas con la deserción escolar. Esto último es consistente con lo presentado por Devasia et al. en [12] donde establece que la educación de la madre y el ingreso familiar están altamente correlacionados con el desempeño del estudiante. Si bien existen múltiples estudios sobre el abandono en la universidad, hay un poco de investigación relacionada con el área de la informática. Además, la gran mayoría de los estudios se centran en variables estáticas, dejando de lado el componente dinámico de las calificaciones que el estudiante obtiene durante sus estudios. Es fundamental investigar las causas que llevan a los estudiantes de un programa de Ingeniería de Sistemas a retirar el curso antes de su finalización. Es por esto, que el objetivo del presente estudio es responder a las siguientes preguntas:

1. P1: ¿Cuáles son los principales factores determinantes de la deserción estudiantil de los estudiantes de pregrado en un programa de ingeniería de sistemas de una universidad privada colombiana?

2. P2: ¿Qué técnica de minería de datos es más adecuada para encontrar estos determinantes clave?

Para ello, se modelará la deserción estudiantil utilizando datos recopilados de bases de datos académicas de 2004 a 2010. Además, este artículo presenta las diferencias entre el uso de un enfoque programático para identificar los determinantes clave y el enfoque automático ofrecido por Watson Analytics de IBM. Los datos demográficos no son tenidos en cuenta.

La relevancia de los resultados para la industria está en la evaluación e identificación de modelos de clasificación para usar en un contexto universitario en Colombia. Los datos usados sirven de evidencia para que cualquier otra Universidad pueda entrenar dichos modelos con sus propios datos y de esta forma apoyar sus procesos de alertas tempranas de deserción. La relevancia de estos resultados para la academia está en el apoyo que se espera proveer para la Universidad donde se realizó la investigación, al entregar evidencia de los pasos realizados junto con las entradas y salidas. De esta forma se logra que la comunidad académica se pueda aprovechar de este estudio.

Este documento está organizado de la siguiente manera: En la Sección 2 se presenta la metodología que se seguirá en este trabajo y que es ampliamente utilizada en proyectos de analítica de datos. En la Sección 3 se presentan los datos empleados y el tratamiento que se le dará a los datos, de acuerdo con la metodología de trabajo. En la Sección 4 se presentan los modelos empleados, junto con la configuración necesaria para ponerlos a punta. Los resultados y discusión de este trabajo son presentados en la Sección 5. La Sección 6 se hace una revisión de trabajos que se han enfocado en este tema de investigación. Finalmente, las conclusiones son presentadas en la Sección 7.

2. Metodología

Para este trabajo se abordó una tarea de analítica basada en una clasificación binaria, donde el abandono (0, 1) es la variable objetivo o dependiente. Para esto, primero se debió seleccionar una metodología que permitiera la adecuada estructuración del proyecto, y segundo, se identificaron las actividades que correspondieran a nuestro proyecto. En este proyecto, por tanto, se adoptó CRISP-DM (Cross Industry Standard Process for Data Mining methodology/Proceso Estándar Multidisciplinario para la Minería de Datos) [13]. Esta metodología propone un modelo de proceso integral que proporciona un marco para la implementación de proyectos de minería de datos. Este modelo de procesos depende tanto del sector industrial como de la tecnología utilizada.

CRISP-DM es útil para planificar, comunicar al equipo del proyecto y documentar. Proporciona una lista de comprobación genérica que aconseja los pasos a seguir y proporciona consejos prácticos para todos los pasos. Esta metodología tiene como objetivo permitir que los proyectos de minería de datos se vuelvan menos costosos, más confiables, más repetibles, más manejables y más rápidos [14]. El modelo de referencia CRISP-DM proporciona una visión general de las fases del ciclo de vida de un proyecto de minería de datos [14]. Este se divide en seis fases como se presenta en la Figura 1.

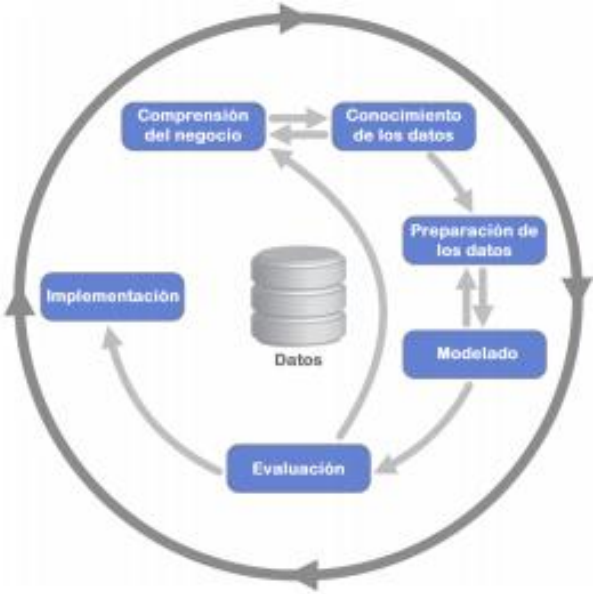

Figura 1. CRISP-DM: Proceso Estándar Multidisciplinario para la Minería de Datos

A continuación, se describen las fases de CRISP-DM, las cuales serán abordadas en las secciones siguientes:

- Comprensión del negocio: permite definir el objetivo del negocio, en nuestro caso el fenómeno de la deserción escolar, tratado en las secciones anteriores.
- Conocimiento de los datos: implica la recopilación de datos, la identificación de problemas de calidad de los datos y el descubrimiento de perspectivas.
- Preparación de los datos: cubre la extracción de características, el manejo de datos y puede requerir múltiples iteraciones.
- Modelado: consiste en la selección de técnicas, aplicación y calibración de parámetros: existe una estrecha relación entre la preparación de datos y el modelado. A menudo, se logran identificar problemas de datos mientras se modela y, de ahí, se obtienen ideas para construir nuevos modelos o datos.
- Evaluación: centrado en la evaluación del rendimiento de los modelos construidos en la fase anterior.
- Implementación: se ocupa de la puesta en ejecución del modelo dentro del contexto real.

Para la comprensión de datos, su preparación y modelado, se utilizó Jupyter Notebook. Para el análisis de datos se utilizó Pandas. Para la estructuración de los datos se usó Scikit-learn. Finalmente, la visualización de los análisis se realizó con Seaborn y Graphviz. Jupyter Notebook es una aplicación que se ejecuta dentro de un contenedor web, es de código abierto y permite crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto. Es una aplicación muy utilizada en el campo de la Ciencia de Datos para tareas tales como: limpieza y transformación de datos, simulación numérica, modelado estadístico, y visualización de datos. Las librerías seleccionadas se ejecutan sobre Jupyter Notebook.

3. Procedimiento

En esta sección se presenta la puesta en práctica de CRISP– DM. La Sección 3.1 hace una descripción de los datos usados en esta investigación. La Sección 3.2 presenta la estrategia usada para transformar los datos y prepararlos para el posterior análisis. Con esta transformación se busca entender cómo están relacionados los datos y qué características son más relevantes.

3.1. Datos

Los datos utilizados para este estudio fueron tomados de una universidad privada en Bogotá, Colombia. Debido por tanto a la naturaleza de la Institución, los datos no pueden ser compartidos, independiente del estado de anónimo en el que se encuentren.

Se recopilaron datos de 762 estudiantes matriculados en el Programa de Ingeniería de Sistemas. Los datos se recibieron organizados en cuatro tablas, y en conjunto, incluyen 43 columnas:

- Información de admisión, incluyendo información demográfica mínima (sexo, fecha de nacimiento, estado civil).
- Fechas de graduación, incluyendo la fecha de graduación y el programa académico.
- Registros de notas, incluyendo los cursos tomados y las calificaciones de cada uno de ellos, el programa académico y el promedio académico acumulativo.
- Ayudas económicas, incluyendo todas las ayudas económicas en los términos requeridos.

En esta investigación, el enfoque se hizo en los estudiantes que ingresan a la universidad desde el primer semestre de 2004 hasta el segundo semestre de 2010. Se eligió el año 2010 como el último año para el análisis, ya que la graduación de los estudiantes está definida a seis años a partir de la matrícula. La confidencialidad de los datos fue preservada al no utilizar datos personales como número de identificación, fecha de nacimiento, número de identificación del campus, ni nombre del estudiante. La tasa total de graduación en este conjunto de datos fue de 52,87 %. La tasa de abandono fue del 47,13 %.

Por parte de los datos, se asumió que son completamente independientes, lo que quiere decir, que el efecto de una variable sobre la variable destino no se ve afectado por alguna otra variable. Adicionalmente, para este trabajo se consideran como estudiantes desertores a aquellos que no completaron su carrera dentro de los seis años luego de matricularse. La deserción se manejó, en este dataset, como una variable binaria.

3.2. Preparación de los datos

En esta fase, se aplicaron principalmente dos operaciones, la depuración y la consolidación de datos, con el propósito de transformar el conjunto de datos originales en otro formato. Esto tuvo como objetivo volver los datos más apropiados y valiosos para el proceso de modelado. La consolidación de los datos consistió en la unión de los cuatro archivos fuente: información de admisión, registro de notas, ayudas económicas y fechas de graduación. Posteriormente, y con el propósito de entender mejor los datos, se realizó un perfilamiento de los mismos.

operación de consolidación en la cual se agruparon los cursos de acuerdo al campo académico al que pertenecen (por ejemplo, ingeniería de sistemas (SE), matemáticas (MATH), física (PH), lenguaje (LAN), administración (MGMT), biología (BIO), etc.). Adicionalmente, se agruparon las calificaciones de los cursos y las veces en que un curso ha sido visto, tanto por alumno como por facultad. Finalmente, se añadió la edad del estudiante al momento de su inscripción al Programa y la desviación estándar del promedio acumulado del semestre académico para reflejar la variación (irregularidad) del rendimiento académico. Como resultado se produjo un dataset con 31 columnas con los siguientes campos: género, estado civil, edad al inscribirse, semestres académicos (por campo académico), promedio académico acumulado, desviación estándar de los promedios de los semestres académicos, promedio de las calificaciones de los cursos, repetición de cursos por facultad e indicador de deserción (0,1).

Un primer resultado de esta consolidación permite visualizar (Figura 2) los promedios de calificaciones por campos académicos, segmentados por indicador de deserción escolar. Como era de esperar, los estudiantes que abandonan son aquellos con menor promedio académico, en general, en todas las materias. Los cursos de administración se presentan como los más desafiantes para ambos tipos de estudiantes, seguidos por las matemáticas y la física.

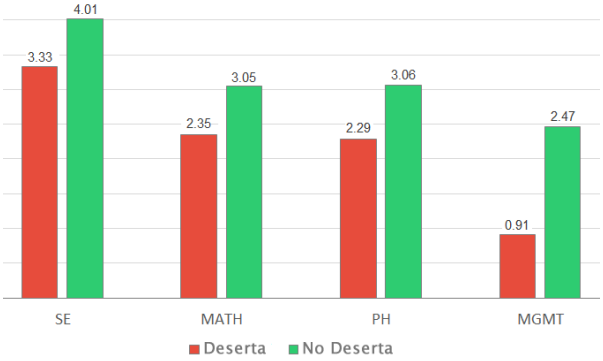

Figura 2. Promedio académico por campo académico de los estudiantes al final del programa

La Figura 3, Figura 4 y Figura 5 presentan un perfil demográfico claramente dominante de los estudiantes de SE: hombres solteros de entre 17 y 19 años de edad. Esta muestra representa el 71,5 % del conjunto de datos. La mayoría de los modelos de analítica requieren entradas numéricas para sus operaciones. Debido a esto, algunos campos categóricos como género y estado civil fueron transformados a valores numéricos. Esta transformación se llevó a cabo por medio de variables dummy que representaron cada valor categórico como una columna binaria. Además, se llevó a cabo un proceso de normalización de las magnitudes con el propósito de evitar que los campos de alta magnitud distorsionaran los pesos de los atributos (features) en los modelos de analítica empleados. Finalmente, como la frecuencia de

abandonos en el conjunto de datos es del 52,87 % para no (0) y del 47,13 % para sí (1), no fue necesario realizar un muestreo por encima o por debajo del conjunto de datos.

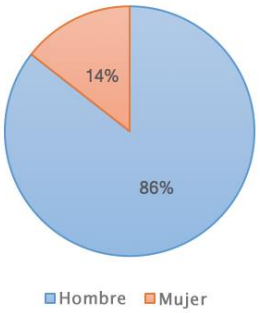

Figura 3. Género del estudiante

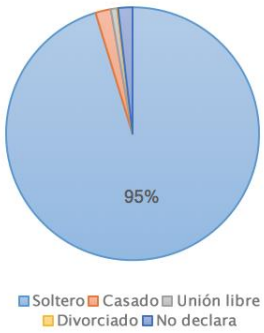

Figura 4. Estado civil del estudiante

Uno de los primeros análisis realizados consistió en la generación de un mapa de calor basado en la visualización de correlaciones entre las variables. Este mapa se presenta en la Figura 6. Como era de esperarse, se encontró una fuerte correlación entre los promedios académicos de los diferentes campos académicos, siendo o la más alta, la correlación entre matemáticas y física.

A partir de la Figura 6, también fue posible descubrir que el Promedio de Notas (GPA) está fuertemente influenciado por los cursos de ingeniería de sistemas (esto es evidente por cuanto componen la mayoría de las clases en el programa) y que además tiene una correlación negativa con la variable dependiente. También se descubrió la fuerte correlación entre la cantidad de veces que un estudiante pierde un curso de ingeniería de sistemas (Falló SE) y la cantidad en biología.

Otra operación aplicada en esta fase fue el Análisis de Componentes Principales (Principal Component Analysis - PCA) para evaluar la reducción de la dimensionalidad de este conjunto de datos. La Figura 7 muestra la proporción de varianza acumulada explicada para 20 Componentes Principales (PCs). Catorce PCs fueron necesarios para explicar más del 90 % (91,54 %) de la varianza, quince PCs explicaron el 93,21 % y veinte PCs acumularon el 97,66 % de la varianza.

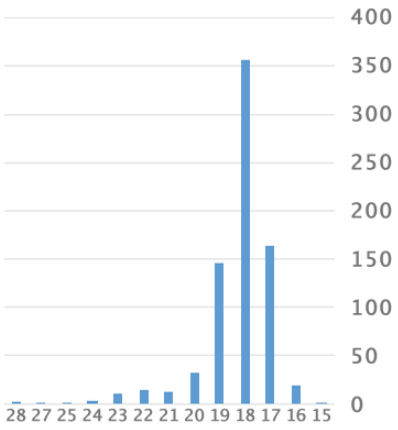

Figura 5. Edad del estudiante

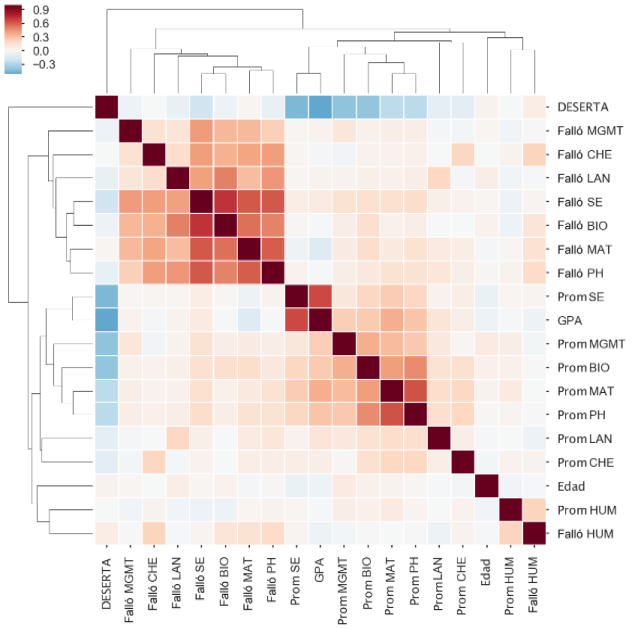

Figura 6. Mapa de calor de correlaciones

Finalmente, la Figura 8, Figura 9 y Figura 10 presentan la progresión del promedio de calificaciones durante los primeros tres semestres para ambos tipos de estudiantes (desertores y no desertores) para los grupos de Matemáticas, Física e Ingeniería de Sistemas. Como primera apreciación se puede apreciar algo evidente, y es que los estudiantes que desertan suelen tener un promedio de calificaciones menor que los que no desertan. Esto sin embargo no es una regla, pero sí un buen indicador. También es interesante evidenciar que las dificultades en los cursos afectan a ambos grupos por igual. Cuando las calificaciones de un grupo suben, las del otro grupo suben, y cuando bajan, las del otro grupo bajan por igual.

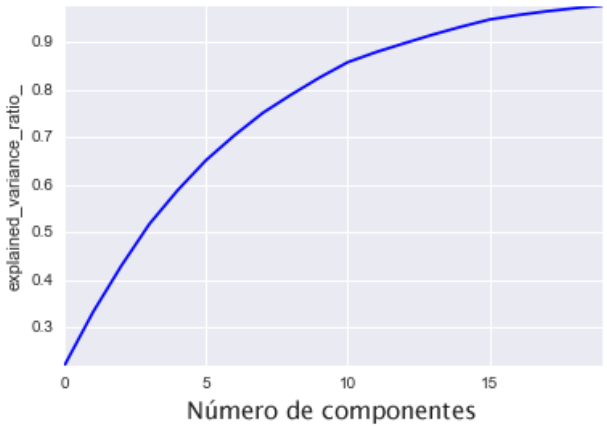

Figura 7. Relación de varianza acumulada para 20 componentes principales

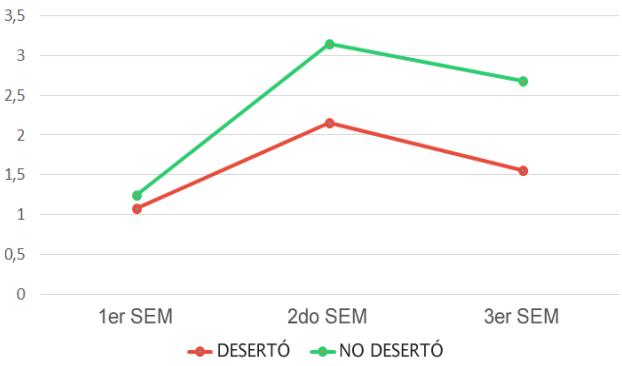

Figura 8. Comparación de calificaciones en cursos de Física

4. Modelamiento

El modelamiento, de acuerdo con CRISP-DM (Sección 2), consiste en la selección de técnicas, aplicación y calibración de parámetros. Este trabajo, basándose en la información presentada en la Sección 6, utilizó los siguientes modelos: Árbol de Decisión, Regresión Logística y Naive Bayes. Estos modelos son los que ofrecen mejor precisión [15, 16, 11]. Adicionalmente, se decidió utilizar un modelo *Bosques Aleatorios* (Random Forest) como técnica complementaria porque es ideal para casos de clasificación y regresión. Todos los modelos se aplicaron a los datos del primer, segundo, tercer y último semestre después de quedar inscritos.

El siguiente paso consistió en establecer cómo se haría la selección de los datos que servirán para entrenamiento y para prueba. Para esto se decidió usar Validación Cruzada, la cual es una técnica empleada para garantizar la independencia entre los datos de entrenamiento y prueba. Específicamente, se empleó el tipo *k-fold CV*, en el cual el conjunto de entrenamiento se divide en k conjuntos pequeños (folds). Para este trabajo, k=5 conjuntos. Donde, en cada conjunto, se dividen los datos de la siguiente manera

- Se entrena un modelo usando $k–1$ (4) conjuntos como datos de entrenamiento (4/5 de los datos).
- El modelo entrenado se valida con los datos restantes (1/5).

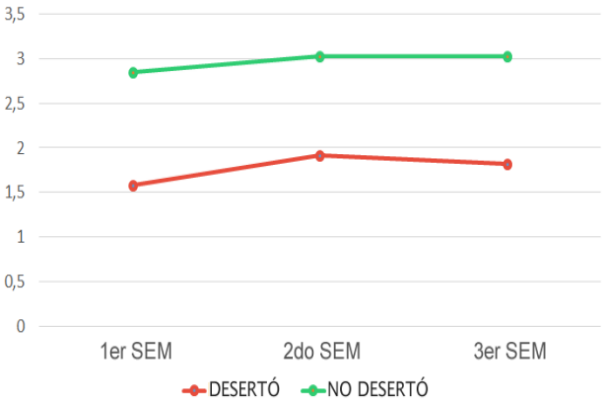

Figura 9. Comparación de calificaciones en cursos de Matemáticas

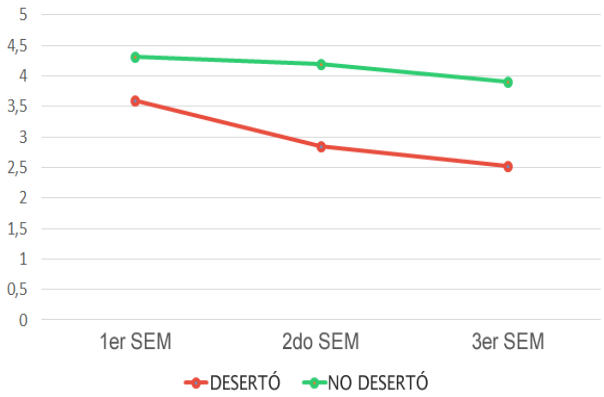

Figura 10. Comparación de calificaciones en cursos de Ingeniería de Sistemas

4.1. Configuración de los modelos

A continuación, se describe la manera en que cada uno de los cuatro modelos empleados fue configurado para su utilización:

- Árbol de Decisión. Se entrenó este modelo usando el criterio gini y Validación Cruzada. El árbol "sin podar" resultó en 12 niveles y 51 hojas. El proceso de "poda" se realizó configurando una profundidad máxima de 4 niveles y se obtuvo un árbol con 7 nodos hoja, tal y como se muestra en la Figura 11.
- Regresión Logística. Este modelo se entrenó con los siguientes parámetros: tolerancia del criterio de parada (tolerance for stopping criteria) = 0,0001, inverso de la fuerza de regularización (inverse of regularization strength) = 1,0, y solver = *liblinear.*
- Naive Bayes. Se utilizó el algoritmo Gaussian Naive Bayes.
- Bosques Aleatorios. Se entrenó el modelo con los siguientes parámetros: número de árboles en el bosque = 10, máxima profundidad del árbol = 4 y estado de aleatoriedad = 0, y un generador de validación cruzada = 6.

4.2. Watson analytics

Watson Analytics es un servicio inteligente que permite analizar y visualizar datos con el propósito de descubrir, rápidamente, patrones y significados en los datos, sin que se requieran conocimientos previos para su configuración. Watson Analytics permite realizar descubrimiento guiado de los datos, análisis predictivo automatizado y cuenta con capacidades cognitivas para interactuar con los datos y obtener resultados fáciles de entender.

4.3. Implementación

Esta fase de la metodología CRISP–DM implica operacionalizar el modelo en el entorno real para detectar riesgos y tomar decisiones para prevenir las tasas de deserción. Dependiendo de los requisitos, esta fase puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible. Lo ideal es que sea el usuario, y no el analista de datos, quien lleve a cabo los pasos de implementación.

Esta fase está por fuera del alcance de este trabajo, pero los hallazgos serán compartidos y discutidos con el programa de Ingeniería de Sistemas con el fin de validar y refinar el modelo, e implementarlo en un entorno productivo. Esta implementación puede desplegarse como un servicio web predictivo para centrarse en los posibles abandonos (basados en reglas de predicción), generar alertas tempranas y tratarlas adecuadamente. Posteriormente, la retroalimentación de las predicciones y los tratamientos de los datos, servirá como insumo para mejorar el modelo.

5. Resultados y discusión

Los modelos seleccionados fueron evaluados en términos del análisis de la curva ROC (Receiver Operating Characteristic). En una curva ROC, la tasa de verdaderos positivos, también llamada Sensibilidad, se representa en función de la tasa o razón de falsos positivos (es decir, 100-Especificidad) para diferentes puntos de corte de un parámetro. Cada uno de los puntos de la curva ROC representa un único par sensibilidad/especificidad correspondiente a un umbral de decisión. La curva ROC implica que el punto de la esquina superior izquierdo es el punto ideal (es decir, una tasa de falso positivo de cero y una tasa de verdadero positivo de uno), por lo que una mayor Área Bajo la Curva (Area Under the Curve - AUC) será mejor. El AUC–ROC es una medida de lo bien que un parámetro puede distinguir entre varios grupos. El resultado del ROC–AUC de los diferentes modelos evaluados por semestre se presenta en la Figura 12.

En general, todos los modelos mostraron buenos resultados, con valores superiores a 0,8 AUC a partir del segundo semestre. También fue posible evidenciar que cuanto más tiempo esté matriculado el estudiante, mejor será la predicción. De entre los modelos, Bosques Aleatorios se destacó por presentar los mejores resultados, con 0,91 AUC en el tercer semestre y 0,97 AUC en el último semestre, lo que lo convirtió en el modelo ideal para este estudio de caso.

A partir del resultado del modelo Árbol de Decisión (ver Figura 11), se destacan dos casos:

- Caso 1: Un estudiante cuyo promedio en los cursos de SE (PROM SE) sea menor o igual a 3,505, y que el número de veces en que ha fallado cursos de SE (Falló SE) sea menor o igual a 0,1389 (escalado entre 0 y 5), y un GPA menor o igual a 3,5164, tendrá el 100 % de probabilidad de desertar. Esto equivale a 143 estudiantes del dataset utilizado.
- Caso 2: Un estudiante cuyo promedio en los cursos de SE (PROM SE) sea menor o igual a 3,505 y que el número de veces en que ha fallado cursos de SE (Falló SE) sea menor o igual a 0,1389 (escalado entre 0 y 5), y un GPA mayor a 3,5164, tendrá 88,8 % de probabilidad de desertar. Esto equivale a 48 estudiantes del dataset utilizado.

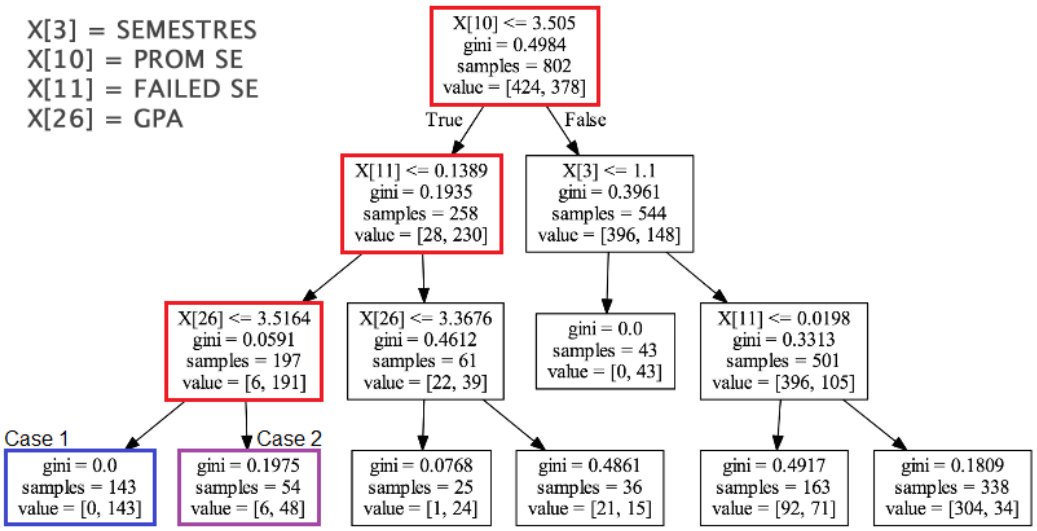

Figura 11. Árbol de decisión podado a tres niveles

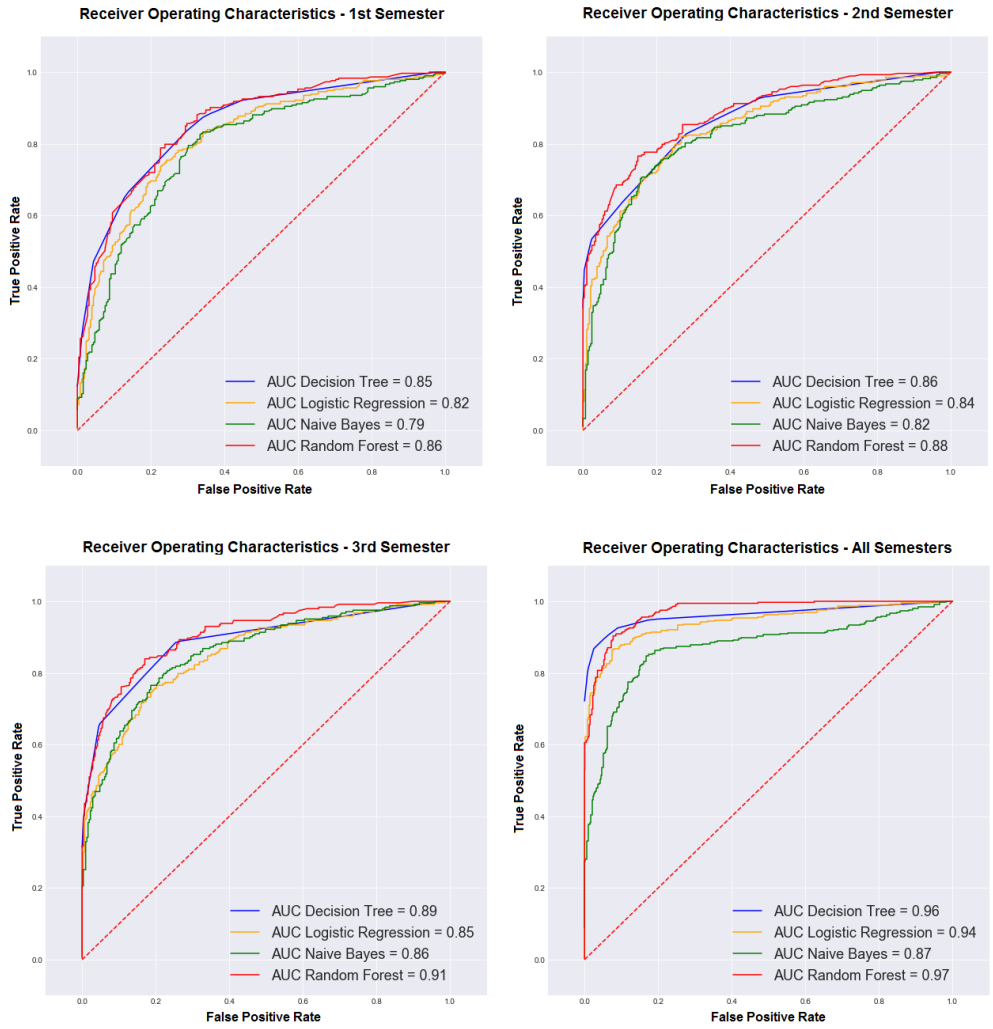

Figura 12. Modelos evaluados usando ROC - AUC

El Tabla 1 presenta los coeficientes de significancia para el modelo Regresión Logística. A partir de la información arrojada por el modelo se evidenció que las características *Falló MAT* y *Falló MGMT* aumentan la probabilidad de desertar. El caso de PROM SE, es opuesto, en el sentido de que reduce el riesgo de no obtener el grado. Algo inesperado fue descubrir la influencia negativa que tiene la característica *Falló SE* en la deserción. Al mismo tiempo, se evidenció la influencia negativa que tiene la característica *Semestres* (la cantidad de semestres que el estudiante tiene inscritos). Esto podría suponer que entre más tiempo esté el estudiando inscrito en la Universidad, menor será la probabilidad de que deserte. Con esto en mente, se podría establecer que como los cursos propios de SE están concentrados en los semestres superiores, entonces fallar estos cursos no implicaría necesariamente que el estudiante desertará. Finalmente, la desviación estándar de las notas de los estudiantes tiene el mayor impacto positivo en el factor de deserción de los estudiantes.

Tabla 1. Coeficientes de significancia para el modelo Regresión Logística con todos los semestres

Característica	Peso
Std GPA	2.84
Falló MAT	1.30
Falló MGMT	0.71
Prom SE	-1.28
Semestres	-1.62
Falló SE	-1.99

Por otra parte, Watson Analytics produjo una serie de gráficos orientados a explicar los hallazgos encontrados a partir de un análisis autónomo. En este trabajo el enfoque estuvo sobre el árbol de decisión presentado en la Figura 13, en el cual Watson usa la característica GPA como la variable raíz. Con esta organización, Watson detectó que el 99,9 % de los estudiantes con GPA por debajo de 3,57 desertan de la Universidad. Además, al tener la variable GPA entre 3,27 y 3,55 y la variable Failed SE igual y menor a 1, los estudiantes tendrán un 81 % de probabilidad de desertar. Watson utilizó las variables de una manera diferente a las encontradas por los diferentes modelos de aprendizaje supervisado, sin embargo, es evidente que los hallazgos encontrados por Watson sirven de igual manera para llevar a cabo el estudio.

5.1. Relevancia de la investigación

La relevancia de este trabajo se puede entender desde su contribución tanto para la industria como para la academia. Las universidades no solo son centros de investigación y formación, también son empresas, y como tal, requieren encontrar la mejor manera de administrarse. Para las universidades es muy importante mantener un flujo regular de estudiantes que permitan soportar sus gastos operativos. Del mismo modo, la Universidad invierte en cada estudiante que ingresa, y cuando un estudiante se retira, pues esa inversión se pierde. Es por esto que para la Universidad es crítico identificar a aquellos estudiantes que pueden estar desarrollando cuadros de deserción.

Independiente de si se hace con variables estáticas (datos demográficos) o dinámicas (calificaciones), la identificación de estos estudiantes es importante para poder realizarles un acompañamiento que permita que puedan continuar y finalizar sus estudios. Este trabajo ofrece esta oportunidad a las universidades, para apoyar y complementar los actuales procesos de alertas tempranas con los que cuente la universidad.

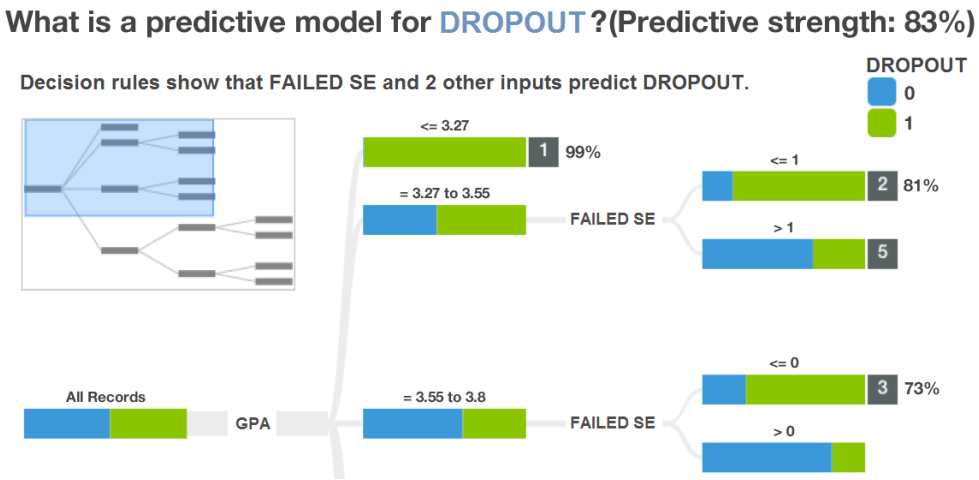

Figura 13. Árbol de Decisión generado por Watson Analytics

La relevancia para la academia se presenta en forma de conocimiento replicable que sienta las bases para que se puedan desarrollar experimentos incrementales que permitan encontrar datos de entrenamiento que provean modelos más precisos. Esta investigación aborda la problemática de la deserción desde un enfoque dinámico al emplear información incremental y cambiante cada semestre como lo son las calificaciones. Estos datos, complementados con información estática permitirá establecer mecanismos cada vez más robustos de identificación de la deserción.

6. Trabajos previos

La Minería de Datos Educativos (EDM) como una disciplina comprometida con el desarrollo de métodos y técnicas para explorar y analizar grandes repositorios de datos del contexto educativo para extraer patrones, asociaciones y relaciones entre ellos [12]. Estos datos pueden ser tomados de plataformas de aprendizaje colaborativo, o de calificaciones, ambiente académico y familiar, entre otros. Esta información puede usarse para anticipar el número de matriculados a un curso, para estimar la tasa de deserción escolar, detectar valores atípicos en las calificaciones de los estudiantes y mejorar los modelos estudiantiles [17, 18].

Bharadwaj y Pal [9] usaron EDM para evaluar el desempeño de los estudiantes entre 300 estudiantes de cinco universidades diferentes que estaban matriculados en un curso de pregrado. Emplearon un esquema de clasificación bayesiana de 17 atributos, de los cuales la puntuación en un examen de secundaria superior, la residencia, y los ingresos anuales de la familia demostraron ser parámetros importantes para el rendimiento académico. Un estudio similar fue propuesto por Kovacic en [11], quien utilizó los algoritmos CHAID y CART para identificar qué datos del proceso de inscripción y familiares se podrían usar para predecir el rendimiento académico de los estudiantes.

Al-Radaideh et al. [19] analizaron la información demográfica de los estudiantes para construir un clasificador basado en árboles de decisión, que permitió identificar que el grado obtenido en bachillerato correspondía al atributo con la mayor relevancia y así mismo, considerarlo el nodo raíz del árbol de decisión. En este trabajo, los métodos Retención (Holdout) y Validación Cruzada (KCross-Validation) fueron utilizados para evaluar el modelo.

Gerben et al. en [16] utilizaron técnicas de machine learning para predecir el éxito de los estudiantes utilizando información extraída de las calificaciones de los preuniversitarios. Sus resultados experimentales mostraron que los árboles de decisión, dan resultados adecuados con precisiones de entre el 75 % y el 80 %. Uno de los predictores más fuertes fue la calificación del curso de Álgebra Lineal. Luan en [20] construyó predictores utilizando Agrupación (clustering) como medio de exploración y clasificación de datos. En [21], Romero y Ventura evidenciaron que el uso de análisis por asociación es un enfoque con creciente popularidad. Finalmente, Herzog en [22] presentó evidencia, empleando conjuntos de datos menores, que las redes bayesianas y las redes neuronales son superadas por los algoritmos de Árbol de Decisión.

Sin embargo, y a pesar de estos estudios, aún no está claro qué algoritmos de minería de datos son ideales en este contexto educativo.

7. Conclusiones

La investigación en educación ha logrado aprovechar el campo de la minería de datos. El ritmo actual de aplicación de los métodos de minería de datos en este campo ha aumentado para una variedad de propósitos, por ejemplo, evaluar las necesidades de los estudiantes, predecir las tasas de deserción escolar, analizar y mejorar el rendimiento académico de los estudiantes. La predicción de la deserción escolar es una tarea importante y desafiante. En este trabajo, se mostraron los resultados preliminares para predecir el abandono de los estudiantes a partir de un conjunto de datos compuesto por datos demográficos estudiantiles y, principalmente, por los registros de transcripciones en diferentes puntos de sus carreras. Entre los hallazgos estuvo que el rendimiento de los cursos de ingeniería de sistemas está correlacionado con el rendimiento de los cursos de física y matemáticas. La irregularidad (desviación estándar de los promedios del semestre) se correlaciona positivamente con la deserción escolar.

Los resultados experimentales obtenidos en este trabajo mostraron que las mejores AUC se lograron por medio de Bosques Aleatorios, desde el tercer semestre de inscrito se obtuvo 0,91 AUC hasta el último semestre, donde el modelo dio 0,97 AUC. Cuatro características fueron necesarias (*Semestres, Prom SE, Falló SE, GPA*) para lograr esta precisión. Esto implica que los cursos relacionados con SE tienen el mayor impacto en la predicción de deserción escolar.

Como trabajo futuro se espera recopilar un conjunto de datos más amplio de toda la base de datos de estudiantes universitarios y aplicar el modelo utilizando dichos datos para ver cómo se generaliza a otros programas específicos. Además, se pueden aplicar otros métodos de

clasificación para encontrar el método más adecuado y ofrecer una mayor precisión de clasificación.

Los hallazgos deben ser compartidos y discutidos con el programa de Ingeniería de Sistemas a fin de validar y refinar el modelo e implementarlo en un entorno productivo.

Referencias

[1] D. Kim and S. Kim, "Sustainable Education: Analyzing the Determinants of University Student Dropout by Nonlinear Panel Data Models," *Sustainability*, vol. 10, no. 4, pp. 1–18, March 2018 [En línea]. Disponible en: https://ideas.repec.org/a/gam/jsusta/ v10y2018i4p954-d137969.html

[2] J. J. Brunner, J. Gacel-Avilà, M. Laverde, J. Puukka, J. Rubio, S. Schwartzman, Ó. Valiente et al., *Higher Education in Regional and City Development: Antioquia, Colombia 2012*. OECD, 2012. [En línea]. Disponible en: https://www.oecd-ilibrary.org/ content/publication/9789264179028-en

[3] S. d. O. Durso, J. V. A. d. Cunha, "Determinant Factors for Undergraduate Student's Dropout in an Accounting Studies Department of a Brazilian Public University," *Educação em Revista*, vol. 34, 00 2018. [En línea]. Disponible en: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982018000100142&nrm=iso

[4] T. Mishra, D. Kumar, and S. Gupta, "Mining students' data for prediction performance," in *Fourth International Conference on Advanced Computing & Communication Technologies*, ser. ACCT '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 255–262. doi: 10.1109/ ACCT.2014.105

[5] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: a case study with high school students," *Expert Systems*, vol. 33, no. 1, pp. 107–124, feb 2016. doi: 10.1111/exsy.12135

[6] A. Seidman, "Retention revisited: R= e, id+ e & in, iv." *College and University*, vol. 71, no. 4, pp. 18–20, 1996.

[7] "Spadies - sistema de prevención y análisis a la deserción en las instituciones de educación superior," Ministerios de Educación, [En línea]. Disponible en: www.mineducacion.gov.co/ 1621/article-156292.html

[8] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of educational research*, vol. 45, no. 1, pp. 89–125, 1975.

[9] B. K. Bhardwaj and S. Pal, "Data mining: A prediction for performance improvement using classification," *(IJCSIS) International Journal of Computer Science and Information Security*, vol. 9, no. 4, 2011.

[10] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, 2011.

[11] Z. Kovacic, "Early prediction of student success: Mining students' enrolment data," in *Informing Science & IT Education Conference (InSITE)*, vol. 10, 2010, pp. 647–665. doi: 10.28945/1281

[12] T. Devasia, Vinushree T P, and V. Hegde, "Prediction of students performance using Educational Data Mining," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) IEEE*, mar 2016, pp. 91–95. [En línea]. Disponible en: http://ieeexplore.ieee.org/document/7684167/

[13] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0," *CRISP-DM Consortium*, vol. 76, 2000. [En línea]. Disponible en: ftp://ftp.software.ibm.com/software/analytics/spss/support/ Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[14] R. Wirth, "Crisp-dm: Towards a standard process model for data mining," in *Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.

[15] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting Student Dropout in Higher Education," in *2016 ICML Workshop on Data4Good: Machine Learning in Social Good Applications*, 2016, pp. 16–20. [En línea]. Disponible en: http://arxiv.org/abs/1606.06364

[16] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting Students Drop Out: A Case Study," in *International Conference on Educational Data Mining (EDM)*, 2009, pp. 41–50. [En línea]. Disponible en: http://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf

[17] E. Yukselturk, S. Ozekes, and Y. K. Türel, "Predicting dropout student: an application of data mining methods in an online education program," European Journal of Open, Distance and Elearning, vol. 17, no. 1, pp. 118–133, 2014.

[18] A. Tekin, "Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach," *Eurasian Journal of Educational Research*, vol. 54, pp. 207–226, 2014. [En línea]. Disponible en: https://eric.ed.gov/?id=EJ1057301

[19] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees," in *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University*, Jordan, 2006, pp. 1–5.

[20] L. Jing, "Data mining and its applications in higher education," *New Directions for Institutional Research*, vol. 2002, no. 113, pp. 17–36, 2002. doi: 10.1002/ir.35

[21] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135 – 146, 2007. [En línea]. Disponible en: http://www.sciencedirect.com/science/article/pii/S0957417406001266

[22] H. Serge, "Estimating student retention and degreeâcompletion time: Decision trees and neural networks vis-á-vis regression," *New Directions for Institutional Research*, vol. 2006, no. 131, pp. 17–33, 2006. doi: 10.1002/ir.185