Estimación del riesgo de deserción de los beneficiarios en Sapiencia

Seminario de Analítica y Ciencia de Datos

# Resumen Descriptivo del Proyecto

Desarrollar modelos predictivos con el fin de mejorar la gestión del recurso económico al estimar la proyección financiera para cada estudiante en su ciclo académico e identificar qué beneficiarios pueden ser susceptibles de abandonar sus estudios postsecundarios y por ende su crédito o beca en Sapiencia.

Marco Teórico:

1. Revisión de Literatura.

Dada la gravedad de la deserción en la educación superior en Colombia, es esencial explorar proyectos y artículos que ofrezcan soluciones efectivas. Por ejemplo, el trabajo de Kang y Wang (2018) propone un modelo basado en regresión logística que permite perfilar a estudiantes de cursos virtuales en riesgo de deserción, utilizando variables sociodemográficas y datos de interacción con la plataforma. Este enfoque ha demostrado una precisión del 81,8%. (Kang & Wang, 2018). Este porcentaje es significativo, ya que indica que el modelo es capaz de identificar correctamente a un alto porcentaje de estudiantes en riesgo. Sin embargo, el 18,2% de los estudiantes en riesgo podrían no ser identificados, lo que podría resultar en un número considerable de deserciones no anticipadas. En conclusión, este enfoque resalta la importancia de aplicar técnicas de minería de datos para detectar patrones de abandono, incluso antes de que los estudiantes ingresen a la institución, lo que se alinea con el objetivo de nuestro proyecto de identificar potenciales deserciones desde el inicio y permitir intervenciones proactivas que mejoren la retención estudiantil.

El segundo modelo, presentado por Albán y Mauricio (2018), se centra en datos cualitativos que reflejan los hábitos de comportamiento de los estudiantes. En este estudio, se utilizan respuestas de encuestas aplicadas a los alumnos y se aplica la técnica de árboles de decisión, logrando una precisión del 97,95% al predecir la deserción (Albán & Mauricio, 2018). Sin embargo, es importante señalar que este alto porcentaje de precisión se basa en variables derivadas de formularios en los que los estudiantes reportan comportamientos adictivos, como el uso de drogas, alcohol y videojuegos, esto implica que la fiabilidad de las predicciones depende de la sinceridad con la que los estudiantes respondan a estas encuestas. Los resultados del proceso experimental indican que factores como la adicción a las redes sociales, el apego emocional, el estado civil y la edad influyen en el proceso de deserción (Albán & Mauricio, 2018). Con una tasa de precisión del 97,95%, el método Chi-squared Automatic Interaction Detector (CHAID) se establece como óptimo en términos de calidad y efectividad para predecir la deserción en universidades (Albán & Mauricio, 2018). Sin embargo, dado que no contamos con encuestas que capten este tipo de información específica, la aplicación directa de este modelo en nuestro contexto podría ser limitada. La falta de datos sobre comportamientos adictivos y otros factores cualitativos podría impedir una evaluación precisa del riesgo de deserción.

Además, se presenta una investigación que utiliza un clasificador bayesiano simple (SBND) con un enfoque de Markov, enfocándose en la deserción escolar de estudiantes de la Facultad de Ciencias de la Ingeniería de la Universidad de Mumbai, India, durante el período 2017-2018. Este estudio se basa en diversos datos socioeconómicos, como carrera, curso, discapacidad, costo de la educación y situación de vivienda (Vilora et al., 2019). Sin embargo, los resultados de esta metodología no fueron los mejores, dado que apenas logró detectar el 33.35% de los estudiantes que probablemente desertarán mediante el uso de un clasificador bayesiano (BayesNet con K2) (Vilora et al., 2019), el porcentaje de detección es bastante bajo, lo que sugiere que esta metodología puede no ser la ideal para nuestro proyecto. Además, el costo, medido como el porcentaje de estudiantes considerados potenciales desertores entre los no desertores, es solo del 1.49%, lo que dificulta la identificación precisa de los desertores (Vilora et al., 2019). Por lo tanto, se deben considerar otros enfoques que puedan proporcionar una mayor tasa de detección y eficacia en la identificación de estudiantes en riesgo de abandono, asegurando así un apoyo más efectivo para fomentar su permanencia en el sistema educativo.

Un estudio relevante realizado por Guzmán-Castillo (2022) se enfoca en la creación de un sistema de información predictiva (SI) destinado a prevenir la deserción en instituciones de educación superior. Este sistema permite calcular el riesgo de que cada estudiante abandone sus estudios, lo que resulta esencial para identificar a aquellos que podrían necesitar apoyo adicional (Guzmán Castillo et al., 2022). Para el desarrollo de su metodología tuvieron en cuenta la disponibilidad de datos en las diferentes dependencias de la institución objeto, incluyendo información socioeconómica y demográfica, registros de desempeño académico y puntajes de pruebas estandarizadas nacionales (ICFES) (Guzmán Castillo et al., 2022). El sistema de información predictiva (SI) presentado es una herramienta eficaz para anticipar, monitorear y gestionar los factores de riesgo relacionados con la deserción estudiantil. Sin embargo, su aplicación en estudiantes que aún no han ingresado a la universidad presenta limitaciones significativas, dado que se centran en un seguimiento a los estudiantes que ya ingresaron a la Universidad.

En “Predicting Student Dropout in Higher Education” (Aulck et al., 2024), se presenta un modelo para predecir deserción utilizando una base de datos de 32.000 estudiantes, una de las más grandes empleadas para este fin en comparación con otros trabajos analizados. Las técnicas aplicadas incluyeron regresión logística, k-Nearest Neighbors y árboles aleatorios, abarcando tanto variables cualitativas como cuantitativas. Se consideraron características sociodemográficas, incluyendo información demográfica (raza, género, fecha de nacimiento, estado de residencia, hispano o no), escolaridad previa, y los puntajes obtenidos en las pruebas SAT (Scholastic Assessment Test) y ACT (American College Testing), que evalúan habilidades en lectura, escritura, matemáticas y ciencias, y se utilizan para la admisión a universidades (Aulck et al., 2024). Esta metodología puede ser muy importante para nuestro proyecto, ya que se asemeja bastante a los datos sociodemográficos que manejamos. Además, los test SAT y ACT tienen similitudes con el ICFES realizado en nuestro país, el cual también se tendrá en cuenta en nuestro modelo. De las técnicas utilizadas en dicho estudio la que mejor desempeño mostró a la hora de predecir la deserción fue la regresión logística; sin embargo, el valor de precisión logrado fue de tan sólo 66.59%, por lo cual los mismos autores proponen utilizar otras técnicas con una mejor capacidad predictiva como son las redes neuronales y las máquinas de soporte vectorial.

Otro proyecto de investigación es “*Diseño de un modelo predictivo para generar alertas tempranas de deserción universitaria en los programas de pregrado presenciales de la Facultad de Ingeniería de la Universidad de Antioquia*”, el cual busca desarrollar una herramienta para identificar tempranamente a los estudiantes en riesgo de deserción, utilizando datos históricos de la universidad y la plataforma del ICFES (Quintero, 2022). En dicho estudio aplicaron técnicas de machine learning, específicamente redes neuronales artificiales (RNA) y Xtreme Gradient Boosting (XGBoost), para crear modelos que abordan la deserción en los primeros tres semestres académicos. El modelo del primer semestre se basó en datos de ingreso, mientras que los de segundo y tercer semestre incluyeron variables de desempeño académico, mostrando que la precisión de los modelos mejora a medida que los estudiantes avanzan en su formación. El modelo de XGBoost logró clasificar a un estudiante como potencial desertor con una precisión del 74,91%, superando a otros estudios que no alcanzaron el 68% (Quintero, 2022). Este método, aunque potencialmente útil, presenta la limitante de ser más efectivo para aquellos estudiantes que ya se encuentran en la universidad, dado que los modelos de segundo y tercer semestre incorporan variables relacionadas con el desempeño académico, lo que mejora la precisión de las predicciones (Quintero, 2022). Por lo tanto, aunque el modelo puede ofrecer algunas indicaciones iniciales sobre el riesgo de deserción, su efectividad aumenta considerablemente una vez que los estudiantes han comenzado sus estudios. Esto se aleja del objetivo principal de nuestro proyecto, que busca identificar a los posibles desertores antes de que ingresen a la universidad.

1. Modelos y Métodos ya existentes.

**2.1. Regresión Logística:**

La regresión logística es un grupo de técnicas estadísticas que tienen como objetivo comprobar hipótesis o relaciones causales entre una variable dependiente categórica y otras variables independientes que pueden ser categóricas y cuantitativas (Martínez Pérez & Pérez Martín, 2024).

La regresión logística ha sido utilizada en diversos estudios para predecir la deserción estudiantil, ofreciendo una herramienta valiosa para entender este fenómeno. Por ejemplo, en el estudio realizado por Kang y Wang, se aplicó este enfoque en el contexto de cursos virtuales, donde lograron alcanzar una precisión del 81,8% (Kang & Wang, 2018).

En nuestro proyecto, se busca predecir si un estudiante desertará antes de ingresar a la educación postsecundaria. La regresión logística podría llegar a ser una buena herramienta, dado que se podrían analizar datos como el rendimiento académico previo (resultados ICFES) y factores socioeconómicos, e identificar a aquellos beneficiarios que tienen una mayor probabilidad de abandonar sus estudios incluso antes de comenzar. Esta identificación temprana es fundamental para que instituciones como Sapiencia puedan seleccionar a aquellos estudiantes que son más propensos a completar su formación, optimizando así la asignación de recursos para becas condonables.

Sin embargo, la regresión logística presenta dificultad para capturar relaciones no lineales y puede resultar en una identificación incompleta de estudiantes en riesgo, lo que podría afectar la precisión de nuestras proyecciones.

**2.2. Árboles de decisión**

Los árboles de decisión han demostrado ser efectivos al basarse en hábitos de comportamiento. En un estudio reciente, se analizaron variables como la adicción a redes sociales, adicción a drogas, adicción al alcohol, adicción a juegos, apego emocional y edad (Albán & Mauricio, 2018). Este enfoque logró una precisión del 97,95%, lo que resalta su capacidad para manejar datos complejos y ofrecer interpretaciones claras (Albán & Mauricio, 2018).

Los árboles de decisión podrían complementar la regresión logística al proporcionar una visión más detallada sobre los factores que influyen en la probabilidad de deserción de estudiantes antes de ingresar a la educación postsecundaria. Identificando patrones de comportamiento que podrían indicar un riesgo elevado de abandono, Sapiencia podría tomar decisiones más informadas sobre la selección de beneficiarios para las becas condonables.

Sin embargo, es importante considerar que el enfoque de árboles de decisión puede verse limitado por la calidad y la representatividad de los datos utilizados (Albán & Mauricio, 2018). Por lo tanto, es esencial garantizar que se cuente con información precisa y relevante para maximizar la efectividad del modelo en la identificación de estudiantes en riesgo.

**2.3. Clasificador Bayesiano con Enfoque de Markov**

El Clasificador Bayesiano con Enfoque de Markov combina la teoría bayesiana y la dependencia de Markov para clasificar datos secuenciales, mejorando la precisión en tareas como el procesamiento de lenguaje natural (Díaz Bonilla, 2019).

En el clasificador bayesiano simple con un enfoque de Markov, se consideraron variables como carrera, curso, discapacidad, costo de la educación, situación familiar (vivir separado de la familia, ser propietario de la vivienda), servicios básicos (televisión por cable, tarjeta de crédito, Internet) y transporte. Sin embargo, este modelo solo logró detectar el 33,35% de los estudiantes en riesgo de deserción (Vilora et al., 2019). Sin embargo, este modelo solo logró detectar el 33,35% de los estudiantes en riesgo de deserción (Vilora et al., 2019). Este bajo rendimiento indica que esta metodología podría no ser adecuada para nuestro proyecto, dado su bajo porcentaje en la identificación de desertores.

**2.4. Sistema de Información Predictiva (SI)**

Un sistema de información predictiva (SIP) es una herramienta que utiliza modelos estadísticos y algoritmos de machine learning para analizar datos históricos y predecir comportamientos futuros (Guzmán Castillo et al., 2022). En el estudio de Guzmán-Castillo, se desarrolló un SIP que integra una variedad de datos, incluyendo información socioeconómica y demográfica, registros de desempeño académico y puntajes de pruebas estandarizadas nacionales (ICFES) (Guzmán Castillo et al., 2022). Este enfoque permite una evaluación más completa del riesgo de deserción, identificando factores críticos que afectan la permanencia estudiantil (Guzmán Castillo et al., 2022). No obstante, la metodología utilizada en el SIP podría adaptarse para analizar datos de estudiantes potenciales, lo que permitiría predecir comportamientos futuros y optimizar la identificación de aquellos con mayor probabilidad de abandonar sus estudios.

**2.5. Regresión logística, k-Nearest Neighbors (KNN) y árboles aleatorios**

* **k-Nearest Neighbors (KNN):** Es un algoritmo de clasificación que asigna una etiqueta a un nuevo punto de datos en función de las etiquetas de los k puntos más cercanos en el espacio de características (Aulck et al., 2024).
* **Árboles aleatorios:** Este método combina múltiples árboles de decisión para mejorar la precisión y reducir el riesgo de sobreajuste. Al agrupar las decisiones de varios árboles, se logra una mayor robustez en las predicciones (Aulck et al., 2024).

Un estudio reciente (Aulck et al., 2024) empleó regresión logística, k-Nearest Neighbors (KNN) y árboles aleatorios para predecir la deserción estudiantil, utilizando variables como información demográfica (raza, género, fecha de nacimiento, estado de residencia, hispano o no), escolaridad previa y puntajes obtenidos en las pruebas SAT y ACT (Aulck et al., 2024). Las precisiones obtenidas fueron de 62,24% para árboles aleatorios, 66,59% para regresión logística y 64,60% para KNN (Aulck et al., 2024). Aunque las precisiones alcanzadas no son óptimas, su enfoque permite identificar una gama más amplia de factores que influyen en la deserción. Por lo tanto, este estudio es particularmente útil para nuestro proyecto, ya que proporciona una base sólida para mejorar la identificación de estudiantes en riesgo antes de su ingreso a la educación postsecundaria.

**2.6. Redes Neuronales y XGBoost**

Redes neuronales son modelos computacionales que simulan el funcionamiento de las neuronas del cerebro humano, compuestos por capas de nodos interconectados que permiten capturar patrones complejos en grandes volúmenes de datos (*¿Qué Es Una Red Neuronal? - Explicación De Las Redes Neuronales Artificiales*, n.d.). Por otro lado, XGBoost es un algoritmo de boosting que combina múltiples árboles de decisión para optimizar la precisión de las predicciones, destacándose por su rapidez y eficacia en diversas aplicaciones de *machine learning* (Quintero, 2022).

En el trabajo de Quintero (Quintero, 2022), se utilizaron tanto redes neuronales artificiales como XGBoost, considerando una amplia gama de variables, que incluyen rendimiento académico, calidad del programa, norma académica, resultados de pruebas saber, situación laboral del estudiante y de los padres, estrato socioeconómico, edad, género, estado civil y apoyo académico y psicosocial (Quintero, 2022). Con una precisión del 74,91%, XGBoost se presenta como un enfoque prometedor al integrar múltiples factores relevantes para la predicción de la deserción estudiantil (Quintero, 2022). Al igual que en este estudio, nuestro objetivo es identificar a los estudiantes en riesgo, pero lo hacemos desde una etapa anterior, antes de que ingresen a la educación superior. Adaptar este enfoque nos permitiría analizar datos relevantes de posibles estudiantes, como rendimiento académico previo y factores socioeconómicos, mejorando así la capacidad de anticipar quiénes podrían enfrentar dificultades y optimizando la asignación de recursos en becas condonables.

Una vez analizados los diferentes métodos utilizados para predecir la deserción estudiantil antes de ingresar a la educación postsecundaria, se puede concluir que las redes neuronales son la opción más viable. Aunque la regresión logística y los árboles de decisión mostraron resultados útiles, las redes neuronales destacan por su capacidad para manejar grandes volúmenes de datos y detectar patrones complejos, siendo esto fundamental para nuestro proyecto.

1. Diferenciación del Proyecto

[Elabore una discurso basado en los capítulos anteriores que muestre por qué su proyecto es innovador, ya sea por la vía de plantear algo totalmente nuevo o por la vía de mejorar significativamente algo existente. Sea recursivo, use su propio trabajo y el de los demás]

Cómo se ha constatado, la deserción postsecundaria es una problemática que puede ser abordada desde distintas perspectivas, ya sea desde la implementación de modelos de regresión, árboles de decisión, redes neuronales, sistemas de información predictiva; a través de la utilización de la estadística bayesiana o frecuentista. Donde principalmente se ha buscado predecir el riesgo asociado a la deserción en los estudiantes que aspiran o adelantan sus estudios de educación superior

En Sapiencia, se ha buscado desde hace mucho tiempo un modelo que busque predecir, con la información que se recoge del beneficiario tanto en el proceso de legalización del crédito como a lo largo de su ciclo en la agencia de educación, su probabilidad o nivel de riesgo de desertar de su crédito condonable. Por ello, el proyecto pretende integrar, a través de la implementación de redes neuronales, el modelo que logre abstraer toda esa información que proporciona el beneficiario tales como la información socioeconómica, demográfica, académica e información del deudor solidario. Con el objetivo de utilizar la mayor cantidad de información del beneficiario que ayude a predecir un comportamiento futuro, en este caso, el escenario en el que el estudiante se plantea la posibilidad de abandonar sus estudios.

Así como lo presentó Quintero en su trabajo desarrollado en 2022, donde se clasificaron variables de un mismo tipo en varios factores académicos, socioeconómicos, familiares e individuales. Nuestra perspectiva busca también identificar características académicas que son las que presenta el estudiante cada vez que renueva su crédito ya sea semestral o anualmente. Así como características demográficas que son las que se recolectan en el proceso de legalización del crédito, tales como el estrato, comuna o corregimiento de residencia, puntaje del SISBEN, etnia, entre otras. Sumado a ello, se considera incluir otras variables propias de sapiencia que son cruciales a lo largo de la ejecución del crédito, tales como el número de suspensiones temporales, total de semestres de financiación proyectados, si efectúo renuncias a modalidad o cambios de programa, entre otras.

El factor diferenciador que se desea emplear en el proyecto, es la utilización de información a priori de beneficiarios antiguos que ya han desertado de sus estudios para ser utilizada como factores que o condiciones a priori con las que cuenta un beneficiario con el fin de que los pesos o variables predictoras sean más conscientes con la realidad del beneficiario. Además de agregar variables de contexto que se presentan en la Agencia, como el número de suspensiones o el reglamento bajo el cuál se legalizó el crédito. Las cuales pueden influir considerablemente en la probabilidad de deserción del beneficiario, además de ser variables y situaciones particulares que solo se presentan en la entidad y en el tipo de crédito que legaliza el beneficiario. Por consiguiente, requieren un tratamiento diferenciado al manejo que se le puede dar a las variables que corresponden a características propias del beneficiario

En síntesis, nuestro modelo busca aprovechar la robustez que ofrecen las redes neuronales y la disponibilidad de información que se tiene sobre el beneficiario para buscar predecir una probabilidad de riesgo de deserción del crédito condonable con el añadido que supone predecir adicionalmente el costo que puede acarrear este para la entidad.

1. Referencias.

Albán, M., & Mauricio, D. (2018). Decision trees for the early identification of university students at risk of desertion. *International Journal of Engineering Technology*, *7*(4.44), 51-54. https://doi.org/10.14419/ijet.v7i4.44.26862

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2024). Predicting Student Dropout in Higher Education. *Technology in Society*, *76*. https://doi.org/10.1016/j.techsoc.2024.102474

Díaz Bonilla, R. E. (2019). *Metodos Bayesianos para Modelos ´ Ocultos de Markov en series de tiempo con conteo*. [Tesis de Maestría, Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/bitstream/handle/unal/75888/1032432796.pdf

Guzmán Castillo, S., Körner, F., Pantoja García, J. I., Nieto Ramos, L., Gómez Charris, Y., Castro Sarmiento, A., & Romero Conrado, A. R. (2022). Implementation of a Predictive Information System for University Dropout Prevention. *Procedia Computer Science*, *198*, 566-571. https://doi.org/10.1016/j.procs.2021.12.287

Kang, K., & Wang, S. (2018). Analyze and predict student dropout from online programs. *ACM International Conference Proceeding Series*, 6–12. https://doi.org/10.1145/3193077.3193090

Martínez Pérez, J. A., & Pérez Martín, P. S. (2024). Regresión logística. Medicina de Familia. *SEMERGEN*, *50*, 1. https://doi.org/10.1016/j.semerg.2023.102086

*¿Qué es una red neuronal? - Explicación de las redes neuronales artificiales*. (n.d.). AWS. Retrieved October 10, 2024, from https://aws.amazon.com/es/what-is/neural-network/

Quintero, Y. A. (2022). *Diseño de un modelo predictivo para generar alertas tempranas de deserción universitaria en los programas de pregrado presenciales de la Facultad de Ingeniería de la Universidad de Antioquia*. [Tesis de maestría ,Universidad de Antioquia]. https://bibliotecadigital.udea.edu.co/handle/10495/29368

Ramírez Oliveros, L. C., & Zuluaga Maya, J. J. (2018, 01 05). ANÁLISIS DE LAS CARACTERÍSTICAS DEMOGRÁFICAS QUE DIFERENCIAN A LOS ELEGIBLES DEL PROGRAMA “SER PILO PAGA” DE LOS DEMÁS APLICANTES. https://repositorio.uniandes.edu.co/server/api/core/bitstreams/1beefde6-df7c-4ca0-845d-30f3a7dd8160/content

Vilora, A., Pineda Lezama, O. B., & Varela, N. (2019). Bayesian classifier applied to higher education dropout. *Procedia Computer Science*, *160*, 573-577. https://doi.org/10.1016/j.procs.2019.11.045