# MODELO DE EVALUACIÓN TEMPRANA DEL RIESGO DE DESERCIÓN EN ESTUDIANTES UNIVERSITARIOS UTILIZANDO APRENDIZAJE DE MÁQUINA

Diana Gómez Botero
Vicerrectora

Genaro Daza Santacoloma
Coordinador Observatorio Social

Jeimy Higuera Molina
Santiago Marín
Angélica López Gómez
Observatorio social

Pereira
2023

# Introducción

La deserción estudiantil en la educación superior es un desafío global con importantes implicaciones sociales y económicas. Según Himmel (2013), la deserción universitaria se define como el abandono prematuro de un programa de estudios debido a factores generados, tanto dentro del sistema educativo, como por la sociedad, la familia y el entorno. Adicionalmente, como expresan Núñez-Naranjo et al. (2021) la deserción universitaria representa un desafío complejo que impacta a todas las instituciones de educación superior a nivel global. Este fenómeno se refleja en la elevada cantidad de estudiantes que no logran finalizar sus estudios universitarios, generando costos económicos y sociales asociados. En Colombia, la tasa de deserción interanual del año 2020 se ubicó en un 8,02% (SPADIES, 2020). Reducir la deserción tiene numerosos beneficios, como promover la movilidad social ascendente y evitar el desperdicio de recursos económicos por parte de los estudiantes, las familias, las universidades y el gobierno.

En relación a las estrategias para mitigar la deserción en la educación superior, Ramírez et al. (2016) resaltan la importancia de la detección temprana de estudiantes en riesgo, con el propósito de disminuir las posibilidades de abandono, además de la implementación de sistemas de alerta temprana para monitorear a aquellos estudiantes en peligro de desertar. Así mismo, Nuñez-Naranjo (2020) propone como estrategia de intervención la creación de un departamento de retención que cuente con profesionales en trabajo social y psicología, quienes se encarguen de realizar un seguimiento a los estudiantes desde el inicio de sus carreras. En consonancia con estas ideas, en la Universidad Tecnológica de Pereira, una institución universitaria pública del Estado Colombiano, con una población estudiantil de más de 17.000 personas en pregrado y posgrado, ha creado el Programa de Acompañamiento Integral (PAI). Este programa tiene como objetivo principal generar acciones enfocadas en promover modos de vida saludables, mejorar las condiciones y estilos de vida de la comunidad universitaria, y reducir la deserción estudiantil.

Una de las estrategias que se realiza como parte del PAI, es el monitoreo de los estudiantes ante diferentes características de riesgo de deserción, lo cual se puede realizar con modelos predictivos utilizando aprendizaje de máquina. En los últimos años, se ha observado un creciente interés y enfoque especializado por parte de las comunidades dedicadas a la minería de datos y el aprendizaje automático en el ámbito de la predicción de la deserción estudiantil (Pérez et al., 2018; Sandoval-Palis et al., 2020). Así mismo, Hasan & Aly (2019), Xu et al. (2019) y Aly & Hasan (2019) presentan trabajos en los cuales utilizan redes neuronales para la predicción, sin embargo los resultados no superan el 81% de exactitud. En Aulck et al. (2016), se utilizó un conjunto de datos de la Universidad de Washington que contiene información demográfica y académica de los estudiantes, se tomaron muestras al azar de 32.538 estudiantes y utilizaron tres modelos

de aprendizaje automático (regresión logística regularizada, *k*-vecinos más cercanos y bosques aleatorios) para predecir la variable binaria de abandono. La regresión logística regularizada proporcionó las mejores predicciones en términos de abandono estudiantil. Además, intentaron predecir el número de trimestres que los estudiantes tardan en terminar los cursos antes de retirarse, sin embargo las predicciones tenían un error medio de unos 5 trimestres.

Otro estudio realizado por Chen y Zhang (2017), busca predecir la tasa de abandono de cursos en línea (MOOCs), utilizando un algoritmo no supervisado basado en datos históricos. Los investigadores utilizaron los datos de un curso impartido por la Universidad de Pekín en Coursera y aplicaron el algoritmo de bosques aleatorios para clasificar los datos. La precisión del sistema de predicción se evaluó utilizando el índice F1. A partir de los resultados obtenidos, se realizaron sugerencias para mejorar la gestión del curso y prevenir el abandono. Estas sugerencias incluyen: proporcionar más oportunidades para realizar las pruebas y tareas, extender el plazo para completar las tareas calificadas, fomentar la participación en los foros de discusión y dividir los vídeos en fragmentos más cortos para las pruebas en vídeo.

En dos estudios distintos, se emplearon árboles de decisión y bosques aleatorios para abordar el desafío de predecir la deserción estudiantil. En Kemper et al. (2020), utilizaron regresión logística y árboles de decisión para analizar datos académicos de estudiantes de Ingeniería Industrial en el Instituto de Tecnología de Karlsruhe, logrando una exactitud de hasta el 95%. Por otro lado, Bello et al. (2020), se enfocaron en estudiantes de Ingeniería Informática en la Universidad de Santiago de Chile, empleando bosques aleatorios y árboles de decisión para seleccionar características relevantes y obtener una exactitud del 97,2% en la predicción de la deserción. Estos enfoques demuestran la efectividad de los árboles de decisión y de los bosques aleatorios para abordar el problema de la deserción estudiantil.

La bibliografía existente sobre predicción automática de deserción se ha centrado en el uso de medidas de exactitud para evaluar los sistemas de clasificación automática. Sin embargo, estas medidas no resultan adecuadas para el tipo de problema que estamos analizando debido a la naturaleza de los datos. No se pueden evaluar los errores de la misma forma, en el contexto de los modelos de predicción de deserción se debe dar prioridad a minimizar el error tipo II, también conocido como falso negativo, ocurre cuando un modelo o prueba no detecta una condición o efecto presente en los datos. Es decir, se falla al identificar una verdadera relación o resultado positivo que sí existe en realidad. En nuestro caso, el error tipo II es no detectar un estudiante que deserta, lo cual es mucho más importante para nosotros que el error tipo I, el cual sería detectar un supuesto desertor que finalmente no deserta, es por esto que Hoyos-Osorio y Daza-Santacoloma (2023) han desarrollado un modelo que

optimiza el parámetro de sensibilidad (equivalente al *recall*), pero que se enfoca en el análisis de los estudiantes con alto riesgo de deserción, sin embargo los mejores valores reportados no superan el 73% de sensibilidad y cuando se analiza sólo el grupo de estudiantes en bajo riesgo de deserción se tiene una medida promedio de desempeño inferior al 85%.

Es por lo anterior que, en nuestro caso, buscamos optimizar la medida de desempeño del sistema automático de clasificación conocida como *recall* o sensibilidad. La sensibilidad es la proporción de estudiantes que realmente desertaron y que el modelo identificó correctamente como estudiantes en riesgo de deserción, en relación con el total de estudiantes que efectivamente desertaron. Es decir, nos interesa maximizar la capacidad del modelo para detectar a los estudiantes que están en riesgo de abandonar sus estudios, aunque esto pueda implicar un mayor número de falsos positivos (estudiantes que el modelo identifica erróneamente como en riesgo de deserción, pero que finalmente no desertan).

En nuestro caso, buscamos crear un sistema de predicción sensible a la detección de estudiantes en riesgo, lo que nos permitirá intervenir y proporcionar el apoyo necesario para mejorar sus posibilidades de éxito académico y reducir la deserción universitaria.

## Metodología

Con el fin de predecir los estudiantes de primer semestre con alto riesgo de abandono, se ha conformado una base de datos que recopila información de estudiantes de doce semestres diferentes (2017-1 a 2022-2). La base de datos está compuesta por un conjunto de características de los estudiantes antes de su ingreso a la universidad, como edad, sexo, estrato socioeconómico, puntaje académico en la prueba estatal para el acceso a la educación superior, naturaleza del colegio de procedencia (pública o privada) y valor de la matrícula liquidado; además, de dos pruebas que miden el riesgo psicosocial del estudiante, la prueba de Ansiedad de Beck (Beck et al., 1993) y la prueba de depresión de Zung (Zung, 1986); y una prueba que mide el consumo de sustancias psicoactivas (ASSIST (WHO, 2002)). Esta base de datos se complementa con la información intersemestral de deserción de los estudiantes universitarios. Finalmente, se tiene una base de datos con 17.868 registros de estudiantes caracterizados por 28 variables, además de la etiqueta binaria: desertores o no desertores, los cuales están distribuidos así: 12.831 no desertores y 3.238 desertores.

Es evidente que existe un desequilibrio notable en la base de datos en términos de la proporción entre estudiantes matriculados y aquellos que abandonan sus estudios. Este desbalance puede tener implicaciones significativas en los análisis y modelos predictivos utilizados para abordar la deserción estudiantil.

En primer lugar, con el fin de mejorar la calidad de los datos empleados, se utilizaron diversas técnicas de preprocesamiento de datos. Estas técnicas incluyen la eliminación de valores atípicos, el relleno de valores nulos utilizando técnicas de regresión, la normalización de datos utilizando el escalador MinMax (Jain et al., 2005), el balanceo de clases utilizando SMOTE (Fernandez et al., 2018), y el análisis de relevancia mediante el análisis de componentes principales (Abdi, H., & Williams, L. J., 2010).

Se realizó la eliminación de valores atípicos con el fin de mejorar la robustez y precisión del modelo, excluyendo así observaciones que se alejan significativamente de la mayoría de los demás valores en el conjunto de datos. Por otro lado, se utilizaron técnicas de regresión para el relleno de valores nulos, permitiendo completar los valores faltantes mediante la relación entre las variables existentes.

Con el propósito de estandarizar las variables y garantizar que todas tengan la misma escala, se lleva a cabo la normalización de los datos numéricos mediante el uso del escalador MinMax. Esta etapa de normalización resulta fundamental antes de realizar un tipo posterior de análisis multivariado, tal como el análisis de componentes principales (PCA), dado que este tipo de métodos son generalmente sensibles a las diferencias de escala entre las variables. Al aplicar el escalador MinMax, se ajustan los valores de las variables para que estén en un rango específico, típicamente entre 0 y 1. Esto permite que todas las variables contribuyan de manera equitativa al cálculo de las componentes principales, evitando que una variable con una escala más grande domine el análisis. La normalización también ayuda a evitar sesgos y distorsiones en los resultados de PCA, garantizando que las variables estén en la misma escala y sean comparables entre sí.

Adicionalmente, se implementó la codificación *one-hot* para abordar las variables categóricas, con el propósito de obtener una representación numérica de dichas variables. Como resultado, se tienen 104 variables en el conjunto de datos final, dado que las categorías de algunas variables se convierten en variables independientes de estudio.

Para abordar el desequilibrio de clases en el conjunto de datos, es decir, el hecho de que haya una gran mayoría de estudiantes con etiqueta de no deserción en comparación con la cantidad de estudiantes desertores, se utilizó la técnica de SMOTE. Esta técnica genera nuevos ejemplos sintéticos para la clase minoritaria, equilibrando así la distribución de clases en el conjunto de datos. Los ejemplos sintéticos se generan a partir de los ejemplos existentes de la clase minoritaria, preservando la estructura subyacente de los datos.

Además, se realizó el análisis de componentes principales (PCA) para reducir la dimensión de los datos, transformando las variables originales en un conjunto reducido de componentes principales, lo que condujo a la

eliminación de la redundancia en los datos, otorgando mayor peso a aquellas variables que presentan una mayor variabilidad y, por ende, mayor carga informativa.

En cuanto al entrenamiento del modelo de clasificación, se utilizó la técnica conocida como bosques aleatorios (Pal, 2005). Los bosques aleatorios son un tipo de algoritmo de aprendizaje automático que construye múltiples árboles de decisión y los combina para realizar predicciones. Estos modelos son conocidos por su capacidad para manejar datos complejos y no lineales, así como por su capacidad para manejar características categóricas y numéricas.

En última instancia, los resultados presentados en el trabajo se centraron en la base de datos final de pruebas, la cual fue dividida previamente en tres conjuntos: 60% para entrenamiento, 20% para validación y 20% para las pruebas finales del modelo. Esta división permitió evaluar la capacidad del modelo para generalizar y realizar predicciones precisas en datos no vistos previamente.

En última instancia, se realizó una división predefinida de la base de datos en tres conjuntos distintos: entrenamiento, validación y pruebas finales del modelo. Para asegurar una distribución equitativa y representativa de los datos, se asignó un 60% de las observaciones al conjunto de entrenamiento, un 20% se destinó al conjunto de validación y el restante 20% se reservó para las pruebas finales. Además, para abordar la variabilidad y asegurar una evaluación robusta del modelo, se implementó una técnica conocida como validación cruzada mediante la generación de 10 particiones. Cada partición involucra la separación aleatoria de los datos en los conjuntos mencionados previamente, para luego realizar múltiples iteraciones del proceso de entrenamiento, validación y evaluación en diferentes configuraciones de datos. Esta estrategia refuerza el enfoque metodológico para evaluar la capacidad del modelo para generalizar y realizar predicciones precisas en datos no vistos previamente. El conjunto de entrenamiento se utilizó para ajustar los parámetros del modelo en cada iteración, mientras que el conjunto de validación se empleó para afinar los hiperparámetros y evitar el sobreajuste. Por último, el conjunto de pruebas finales, se usó exclusivamente para evaluar el rendimiento del modelo final en situaciones realistas.

## Resultados

En esta sección se presentan los resultados obtenidos después de realizar la validación cruzada sobre la base de datos de prueba con 10 particiones utilizando la técnica de PCA como estrategia de reducción de dimensión.

En el preprocesado de los datos, se inició con la eliminación de datos atípicos proceso en el cual se pasó de tener 17.987 sujetos a 16.069, posteriormente, se rellenaron los valores faltantes de: naturaleza del

colegio de procedencia (pública o privada), utilizando una regresión con el estrato socioeconómico y el valor matrícula liquidado de los estudiantes. En el caso particular de las pruebas psicológicas y de consumo de psicoactivos, de estudiantes que tenían valores faltantes, se adicionó la categoría “Inasistente”, lo cual equivale al hecho de no haber participado en estas actividades, pudiendo reflejar un comportamiento de desinterés, que tal vez se configure como una tendencia al abandono. En otros casos, en los cuales el registro del estudiante carecía de la mayoría de sus variables, se optó por la eliminación de tal registro, lo que llevó a tener una base de datos final con 16.069 registros.

El desbalance de clases señalado anteriormente persiste en la base de datos preprocesada, en la cual de los 16.069, 12.831 estudiantes están etiquetados como no desertores, mientras que 3.238 de ellos sí desertaron de la universidad. En el proceso de balanceo utilizando SMOTE, la base de datos final pasa de tener de 16.069 estudiantes a 25.662 observaciones en total, con igual cantidad de observaciones en cada clase.

La técnica de reducción de dimensión (PCA) disminuye el número total de variables originales a un conjunto reducido de componentes, particularmente para este caso, conservamos el número de componentes principales que explican el 90% de la varianza acumulada, obteniendo como resultado un total de 34 componentes principales.

Posteriormente, al emplear el algoritmo de aprendizaje de máquina denominado Bosques Aleatorios, se obtuvo una exactitud promedio 99,12%. Los resultados detallados de exactitud, sensibilidad, precisión e índice F1, para cada una de las particiones y su promedio general, así como su desviación estándar son presentados en la tabla 1.

Estos valores son evaluados con 5202 observaciones de los datos separados para pruebas, de los cuales 2601 son desertores y 2601 son no desertores.

Tabla 1: Resultados de modelo en 10 particiones

| **Medida** | **Particiones** | | | | | | | | | | **Media** | **Desviación estándar** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | | |
| **Exactitud** | 98,97% | 99,12% | 98,77% | 99,14% | 98,95% | 99,10% | 98,93% | 99,08% | 99,30% | 99,20% | 99,06% | 0,15% |
| **F-1** | 98,98% | 99,13% | 98,77% | 99,16% | 98,97% | 99,12% | 98,93% | 99,11% | 99,29% | 99,22% | 99,07% | 0,15% |
| **Precisión** | 97,98% | 98,27% | 97,61% | 98,33% | 97,95% | 98,25% | 97,92% | 98,24% | 98,59% | 98,45% | 98,16% | 0,28% |
| **Sensibilidad** | 100,00% | 100,00% | 99,96% | 100,00% | 100,00% | 100,00% | 99,96% | 100,00% | 100,00% | 100,00% | 99,99% | 0,02% |

Fuente: Elaboración propia

En nuestro caso específico, es importante optimizar el modelo para minimizar la tasa de falsos negativos, es decir, aquellos estudiantes que el modelo reconoce como no desertores, pero que en realidad sí desertan. Como lo expresa Hoyos-Osorio & Daza-Santacoloma (2023), es importante minimizar los falsos negativos, ya que estos casos representan estudiantes que, a pesar de ser considerados de bajo riesgo de deserción por el modelo, finalmente abandonaron sus estudios.

De la Tabla 1 se observa que la exactitud del modelo alcanzó un valor promedio de 99,06 %, lo que significa que 5.147 de las clasificaciones realizadas por el modelo fueron correctas. La precisión del modelo, con un valor de 98,16 %, representa la proporción de estudiantes clasificados como desertores por el modelo que realmente lo son. Es decir, de los 2.601 estudiantes que el modelo predijo como desertores, 2.553 efectivamente lo son. Esto destaca la capacidad del modelo para identificar de manera acertada a la mayoría de los estudiantes que están en riesgo de desertar. Una alta precisión es crucial, ya que indica que el modelo minimiza los falsos positivos, es decir, minimiza la clasificación errónea de estudiantes no desertores en la categoría desertores. Con una precisión del 98.2 %, podemos confiar en que la gran mayoría de las predicciones positivas realizadas por el modelo son correctas, lo que fortalece su utilidad para abordar la problemática de la deserción estudiantil de manera efectiva. La sensibilidad del modelo es igual a 99,99 %, lo que implicó que el modelo detectara a todos los 2.601 estudiantes que realmente desertan. El puntaje F-1, que combina precisión y sensibilidad, obtuvo un valor de 99,07 %, lo que indica un alto rendimiento conjunto en ambas métricas.

En resumen, el modelo demuestra ser altamente exacto, sensible, preciso y eficaz en la detección de deserciones y presenta una tasa muy baja de falsos negativos en la clasificación de los estudiantes desertores (0,04 %). Estos resultados muestran que el modelo es muy confiable para predecir qué estudiantes están en riesgo de desertar, lo que podría ser de gran utilidad para intervenir y brindarles el apoyo necesario para mejorar sus posibilidades de alcanzar el logro académico y reducir la deserción estudiantil.

## Conclusiones

En el marco de esta investigación, se ha logrado desarrollar un modelo de predicción de riesgo de deserción temprana basado en técnicas de aprendizaje automático. Los resultados obtenidos muestran un alto nivel de exactitud (99,1 % ± 0,1 %), sensibilidad (99,99% ± 0,02%), precisión (98,2% ± 0,28%), e índice F1 (99,07% ± 0,15%). El modelo desarrollado, tal como se señaló previamente, optimiza el indicador de desempeño denominado sensibilidad, el cual evalúa la capacidad del modelo para detectar de manera exhaustiva a todos los verdaderos desertores en el conjunto de datos. En otras palabras, la sensibilidad mide la proporción de estudiantes que realmente desertaron y que el modelo identificó correctamente como

en riesgo de deserción, en relación con el total de estudiantes que efectivamente abandonaron sus estudios. Una sensibilidad del 100 % indica que el modelo logra capturar a todos los verdaderos desertores, lo que es de gran importancia en el contexto de la prevención y retención estudiantil, ya que nos permite identificar de manera eficaz a los estudiantes en situación de riesgo y brindarles el apoyo necesario para mejorar sus posibilidades de éxito académico. Además, es importante destacar que el modelo también ha demostrado tener una tasa de falsos positivos mínima, con solo un 0,04% de casos clasificados incorrectamente como desertores.

Estos hallazgos son prometedores, ya que ofrecen una herramienta efectiva para identificar a aquellos estudiantes que enfrentan un mayor riesgo de abandonar sus estudios. El modelo de predicción de deserción se puede utilizar como herramienta de tamizaje para detectar a los estudiantes con el riesgo de deserción más alto, de manera que programas de intervención escolar, como el Programa de Acompañamiento Integral (PAI), puedan proporcionar un seguimiento y acompañamiento personalizado a los estudiantes identificados, ayudándoles a sortear los obstáculos y dificultades que propician la deserción.

Es importante destacar que el modelo desarrollado no solo proporciona beneficios a nivel individual para los estudiantes, sino que también tiene un impacto positivo en la institución educativa en su conjunto, dado que se cuentan con medidas objetivas para la evaluación del riesgo de deserción, lo que repercute en la mejora de las tasas de retención estudiantil, y promueve un ambiente académico más propicio para la consecución del logro académico y el crecimiento personal de los estudiantes.

Se plantea como trabajo futuro la ampliación del alcance del modelo de predicción para abarcar no solo a estudiantes del primer semestre, sino también a aquellos de todos los semestres. Sin embargo, para lograr esto, sería necesario un esfuerzo adicional por parte de la universidad para recolectar información en diferentes etapas de la carrera de los estudiantes. La recopilación de datos en varias etapas permitiría mejorar la precisión del modelo y brindar pronósticos más sólidos y personalizados para cada estudiante a lo largo de su trayectoria académica. Este proceso de recopilación y actualización continua de datos sería fundamental para asegurar la efectividad y relevancia del sistema de predicción en el futuro.

## Referencias

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

Aly, M., & Hasan, M. R. (2019). Improving stem performance by leveraging machine learning models. In *the Proceedings of the International Conference International Conference of Frontiers in Education (FECS'19)* (pp. 205-2011).

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. (1993). Beck anxiety inventory. Journal of consulting and clinical psychology.

Bello, F. A., Kóhler, J., Hinrechsen, K., Araya, V., Hidalgo, L., & Jara, J. L. (2020, November). Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout. In 2020 39th International Conference of the Chilean Computer Science Society (SCCC) (pp. 1-5). IEEE.

Chen, Y., & Zhang, M. (2017, May). MOOC student dropout: Pattern and prevention. In *Proceedings of the ACM Turing 50th Celebration Conference-China* (pp. 1-6).

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863-905.

Hasan, M., & Aly, M. (2019, December). Get more from less: a hybrid machine learning framework for improving early predictions in stem education. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 826-831). IEEE.

Himmel, E. (2013). University dropout: What are the reasons and what can be done? Journal of Economic Surveys, 27(4), 670-695.

Hoyos Osorio, Jhoan Keider, & Daza Santacoloma, Genaro. (2023). Predictive Model to Identify College Students with High Dropout Rates. *Revista electrónica de investigación educativa*, *25*, e13. Epub 26 de junio de 2023.https://doi.org/10.24320/redie.2023.25.e13.5398

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern recognition*, *38*(12), 2270-2285.

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28-47

Núñez-Naranjo, A. F., Ayala-Chauvin, M., & Riba-Sanmartí, G. (2021, January). Prediction of university dropout using machine learning. In *International Conference on Information Technology & Systems* (pp. 396-406). Cham: Springer International Publishing.

Ramírez, E., Espinosa, D. & Millán, E. (2016). Estrategia para afrontar la deserción universitaria desde las tecnologías de la información y las comunicaciones. Revista Científica,24, 52-62. Doi: 10.14483/udistrital.jour.RC.2016.24.a5

Thomas, Liz. (2002). Student Retention in Higher Education: The Role of Institutional Habitus. http://lst-iiep.iiep-unesco.org/cgi-bin/wwwi32.exe/[in=epidoc1.in]/?t2000=023091/(100). 17. 10.1080/02680930210140257.

Nuñez-Naranjo, A. (2020). Deserción y estrategias de retención: un análisis desde la universidad particular. *593 Digital Publisher CEIT*, *5*(5-2), 79-87. https://doi.org/10.33386/593dp.2020.5-2.306

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, *26*(1), 217-222.

Pérez, A. M., Escobar, C. R., Toledo, M. R., Gutierrez, L. B., & Reyes, G. M. (2018). Prediction model of first-year student desertion at Universidad Bernardo O´ Higgins (UBO). *Educação e Pesquisa*, *44*.

Sandoval-Palis, I., Naranjo, D., Vidal, J., & Gilar-Corbi, R. (2020). Early dropout prediction model: A case study of university leveling course students. *Sustainability*, *12*(22), 9314.

SPADIES (System for the Prevention and Analysis of Dropout in HEI). (2020). Recuperado de https://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-357549.html?_noredirect=1

WHO ASSIST Working Group. (2002). The Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): development, reliability and feasibility. Addiction, 97, 1183–1194. doi:10.1046/j.1360-0443.2002.00185

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166-173.

Zung, W. W. (1986). Zung self-rating depression scale and depression status inventory. In *Assessment of depression* (pp. 221-231). Berlin, Heidelberg: Springer Berlin Heidelberg.