# IDENTIFICACIÓN Y PREDICCIÓN DE ESTUDIANTES EN RIESGO DE DESERCIÓN ACADÉMICA POR MEDIO DE MODELOS BASADOS EN MACHINE LEARNING


# IDENTIFICATION AND PREDICTION OF STUDENTS AT RISK OF ACADEMIC DROPOUT THROUGH MODELS BASED ON MACHINE LEARNING

**Juan Carlos González Sánchez, jcgonzalezs03@libertadores.edu.co**

**Marco Javier Peñaloza Pérez, mjpenalozap@libertadores.edu.co**

## RESUMEN

En el ámbito de la educación universitaria virtual en Colombia existe una creciente preocupación por el tema de la deserción estudiantil, particularmente en las facultades de ingeniería dónde asignaturas relacionadas con las ciencias naturales y matemáticas tienen altos índices de mortalidad académica. El presente trabajo describe el proceso de identificación de las características más importantes que conllevan a que muchos estudiantes abandonen sus estudios en la asignatura Mecánica en la Universidad Nacional Abierta y a Distancia, para eso se tiene una base de datos entre los periodos académicos 2018 a 2020 y sobre la cual se realizó un análisis de predicción basado en técnicas de Machine Learning, cuyo fin ha sido obtener un pronóstico que permita identificar y prever posibles casos de deserción académica para tomar las medidas necesarias que eviten tal situación en futuros casos. El preprocesamiento de los datos y la aplicación de los modelos han ofrecido resultados satisfactorios que permiten efectuar recomendaciones para reducir el porcentaje de alumnos que abandonan sus estudios.



***Palabras clave:*** Árbol de decisión, Regresión Logística, Random Forest, Deserción Escolar.



## ABSTRACT

In the field of virtual university education in Colombia there is growing concern about the issue of student dropout, particularly in engineering schools where subjects related to natural sciences and mathematics have high academic mortality rates. This paper describes the process of identifying the most important characteristics that lead many students to abandon their studies in the Mechanics subject at the National Open and Distance University, for that there is a database between the academic periods 2018 to 2020 and On which a prediction analysis based on Machine Learning techniques was carried out, the purpose of which has been to obtain a forecast that allows identifying and anticipating possible cases of academic dropout to take the necessary measures to avoid such a situation in future cases. The pre-processing of the data and the application of the models have offered satisfactory results that allow recommendations to be made to reduce the percentage of students who drop out.

***Keywords:*** Decision tree, Logistic Regression, Random Forest, School Dropout.

# INTRODUCCIÓN

La deserción estudiantil es un fenómeno complejo que conlleva múltiples impactos negativos, tanto a nivel individual como colectivo: universidad, región y sociedad en general. Múltiples son los factores que conlleva la deserción a nivel universitario desde variables a nivel individual, académicas, institucionales y socioeconómicas.

En Colombia, los altos niveles de deserción estudiantil en el pregrado se presentan como una de las problemáticas más apremiantes del sistema de educación superior. Pese a que los últimos años se han caracterizado por aumentos de cobertura e ingreso de estudiantes nuevos, el número de alumnos que logra culminar sus estudios superiores está por debajo del número de estudiantes que entran por cohorte, dejando entrever que una gran parte de éstos abandona sus estudios, principalmente en los primeros semestres. Según estadísticas del Ministerio de Educación Nacional (Min educación, 2018), de cada cien estudiantes que ingresan a una institución de educación superior cerca de la mitad no logra culminar su ciclo académico y obtener la graduación.

La deserción estudiantil de igual manera es una realidad a nivel local y es un fenómeno complejo, ya que las pérdidas que representa son significativas en los ámbitos financiero y social, para el individuo, las familias, las universidades y la sociedad. Estudiar el problema de la deserción también proporciona políticas de control efectivas para proporcionar un aumento en la cobertura en educación con calidad y equidad, por ello, mediante el uso de técnicas de modelación se logra obtener predicciones a través del estudio de patrones ocultos en sistemas de información (base de datos) que permiten visualizar resultados para que puedan ser leídos y entendidos de manera inmediata, simple y efectiva y útiles para tomar decisiones estratégicas en la solución del problema.

Las causas de la deserción son tan diversas que requieren atención desde distintos frentes, tecnologías como las técnicas de modelos y la gestión de datos emergen como herramientas seguras y de alta precisión para ayudar a las instituciones de educación superior a tomar decisiones, en tiempo real, basadas en evidencia de datos, esto puede ayudar a mitigar y evitar deserciones futuras, ya que tomando datos históricos y con ellos arrojando resultados para ser analizados se brinda una base sólida en la toma de decisiones. Por tal motivo en la facultad

de ingeniería de la Universidad de educación virtual , específicamente en la asignatura física mecánica, se hace importante el desarrollo de un proyecto que ayude a pronosticar y prevenir posibles casos de deserción mediante el uso de algoritmos de Machine Learning, en la cual se toman eventos históricos con distintas variables de tipo social, académico, personal, laboral, etc. y posteriormente a estas variables se le ajustan modelos que permitan predecir las probabilidades de deserción de cada estudiante y posteriormente, con esta información se pueda alertar y aplicar medidas preventivas tempranas con la población estudiantil.

Este modelo con herramientas de minería de datos, permitirá: analizar los factores que afectan la deserción dentro de la facultad, y predecir la probabilidad de deserción que pueda tener un estudiante. Esto, podría generar conocimientos que la universidad pueda utilizar, para crear estrategias que promuevan la retención estudiantil, y, por lo tanto, lograr mayores tasas de graduación y menores de deserción.

# REFERENTES TEORICOS

## Deserción estudiantil

La deserción estudiantil en educación superior es un fenómeno de gran relevancia en américa latina, esta tiene gran incidencia en los indicadores de calidad y eficiencia de las instituciones de educación; por su complejidad y diversidad de causas, no ha sido posible llegar a una solución de este fenómeno. (Muñoz-Camacho et al. ,2018).

En la deserción se entrecruzan factores macroeconómicos y de política educativa estatal con factores institucionales (incluyen consideraciones sobre excelencia académica, educación como servicio o derecho, objetivos mismos de la formación académica, patrones de relacionamiento entre estudiantes, profesores y personal administrativos; entre muchos otros aspectos), factores sociales y factores de orden individual o personal ( motivación, resistencia a la frustración, historia personal y familiar, proyecto de vida) . La deserción dista de ser un fenómeno homogéneo. (Rodríguez-Urrego, 2019).

La deserción o desvinculación voluntaria se concreta cuando el estudiante decide abandonar sus estudios; en tanto que la desvinculación forzada se presenta cuando el estudiante suspende sus estudios por razones estipuladas en la institución, como requerimientos académicos mínimos para garantizar su permanencia en calidad de estudiante (Castaño et al., 2006).Desde el punto de vista de la temporalidad, la deserción estudiantil se considera de tres tipos: se considera precoz, cuando el estudiante a pesar de haber sido admitido en la institución educativa, no se matricula. Temprana, cuando abandona los estudios en los primeros semestres del programa al que se ha matriculado y tardía, cuando abandona sus estudios en los últimos semestres (Durán et al., 2007).

Respecto al espacio, se considera deserción institucional cuando el estudiante abandona la institución educativa; en tanto que, se habla de deserción interna o de programa académico, cuando realiza cambio de programa dentro de la misma institución. Algunos autores consideran que las transferencias de programa en la misma universidad, no se puede considerar deserción porque se trata simplemente de un traslado interno. La dificultad para medir de manera precisa el fenómeno de deserción, se fundamenta en el seguimiento deficiente a los traslados de estudiantes entre instituciones (Montes et al., 2010).

**Factores asociados a la deserción estudiantil**

Cuando se toman en cuenta los factores de riesgo o determinantes de la deserción, se encuentra que estos se abordan con diferentes enfoques: Donoso y Schiefelbein (2007) clasifican estos riesgos de deserción en cinco categorías y según la disciplina en que se desarrollan: psicológico, sociológico, económico, organizacional e interaccionista. Los estudios explicativos toman en cuenta cuatro aspectos como los más relevantes: individuales, académicos, socioeconómicos y los institucionales. El modelo más completo y referenciado es el interaccionista; en este se vincula variables propias de la institución y condiciones particulares del estudiante (Tinto, 1987). La aplicación de estos modelos para estudios de deserción, no depende solo del enfoque teórico, sino de la disponibilidad de información de los estudiantes inscritos en los diferentes programas de las instituciones educativas.

**Deserción en entornos virtuales de aprendizaje**

El uso de entornos virtuales de aprendizaje permite mayor acceso a la educación y mejora las posibilidades de obtener información en diversas fuentes; sin embargo, estos no sustituyen a los recursos pedagógicos tradicionales. Estos entornos amplían y diversifican las posibilidades de la enseñanza y aprendizaje; en este sentido, el reto para las universidades que ofertan este sistema virtual de educación es muy grande. Se busca que estos entornos de aprendizaje fomenten la construcción del conocimiento y este se logre de manera flexible, autónoma y con calidad; disminuir los índices de deserción en estos entornos, se convierte en un desafío que empieza por considerar los roles de los participantes en el proceso formativo, las propuestas y estrategias de enseñanza , los medios u objetos de enseñanza , la estructura y diseño de la plataforma educativa y su capacidad de adaptación a las necesidades presentes y futuras de sus usuarios (La Madriz, 2016).

La educación virtual ha afrontado diversos ataques y discusiones que buscan revisar su diseño, por la falta de contacto entre profesores y estudiantes; asociado esto a los niveles de calidad e indicadores de deserción presentados. En los últimos años, el avance hacia modelos educativos más complejos, la implementación de sistemas de aseguramiento de la calidad en educación virtual, las nuevas formas de interacción y mejores recursos didácticos han permitido una mayor articulación entre la virtualidad y la presencialidad; esto conlleva a

mejores indicadores de deserción y una apuesta hacia un modelo educativo virtual de calidad (La Madriz, 2016).

**Las matemáticas y la deserción estudiantil**

La incidencia de algunas áreas del conocimiento o cursos específicos en la deserción estudiantil universitaria, se evidencia en estudios realizados para matemáticas y áreas afines: Entre otros factores asociados a la deserción se destaca la carencia o ausencia de conocimientos previos, complejidad de abstracción entre la educación media y la universitaria (Castillo-Sánchez,2019).

La matemática a nivel universitario ejerce un impacto en todas las carreras que incluyen algún componente matemático en su currículo; por su grado de abstracción y su negativa percepción, se ha convertido en factor decisivo para la elección o cambio de programa académico. Se escoge los programas con bajo o ningún contenido de cursos en esta área (Cabanzo,2017).

## METODOLOGÍA

La base de datos con la cual se trabajó está compuesta por 904 observaciones sobre diferentes estudiantes de la asignatura Física-Mecánica de la Universidad de educación virtual, se cuenta con información referente a los periodos académicos 20181 al 20202 y algunas variables como ciudad, programa, género, estrato social, deserción , la idea principal es encontrar un modelo estadístico bajo el estudio de Machine Learning que permita clasificar y pronosticar posibles desertores en base a los datos disponibles, el tratamiento de la información y la construcción de los modelos se resumen en las siguientes fases:

1. Inicialmente, se realiza un análisis exploratorio de los datos, esto se hace para encontrar patrones y características de los datos, antes de realizar el modelo. Para nuestro caso utilizamos Sweetviz; esta es una biblioteca de visualización automática de Python de código abierto que se centra en visualizar la relación de los datos, mediante la generación de diferentes tipos de gráficos. El reporte generado incluye: descripción general del conjunto de datos, propiedades de las variables, asociaciones categóricas, asociaciones numéricas, valores más frecuentes, más grandes y más pequeños de las variables numéricas.

   La variable ciudad, presenta cuatro categorías; destacándose Bogotá con un porcentaje del 49 %, seguida por Neiva con 14 %, Zona caribe 13 %, otras ciudades 13% y Medellín con el 11%.

   En programas, ingeniería industrial representa el 39%, ingeniería de sistemas el 17%, ingeniería electrónica 11%, ingeniería de telecomunicaciones 10%, ingeniería ambiental 9%, ingeniería de alimentos 7% y otros programas 6%.

   El estrato socioeconómico se distribuye: estrato 3 con el 46%, estrato 2 con 43% y estrato 1 con 11%.

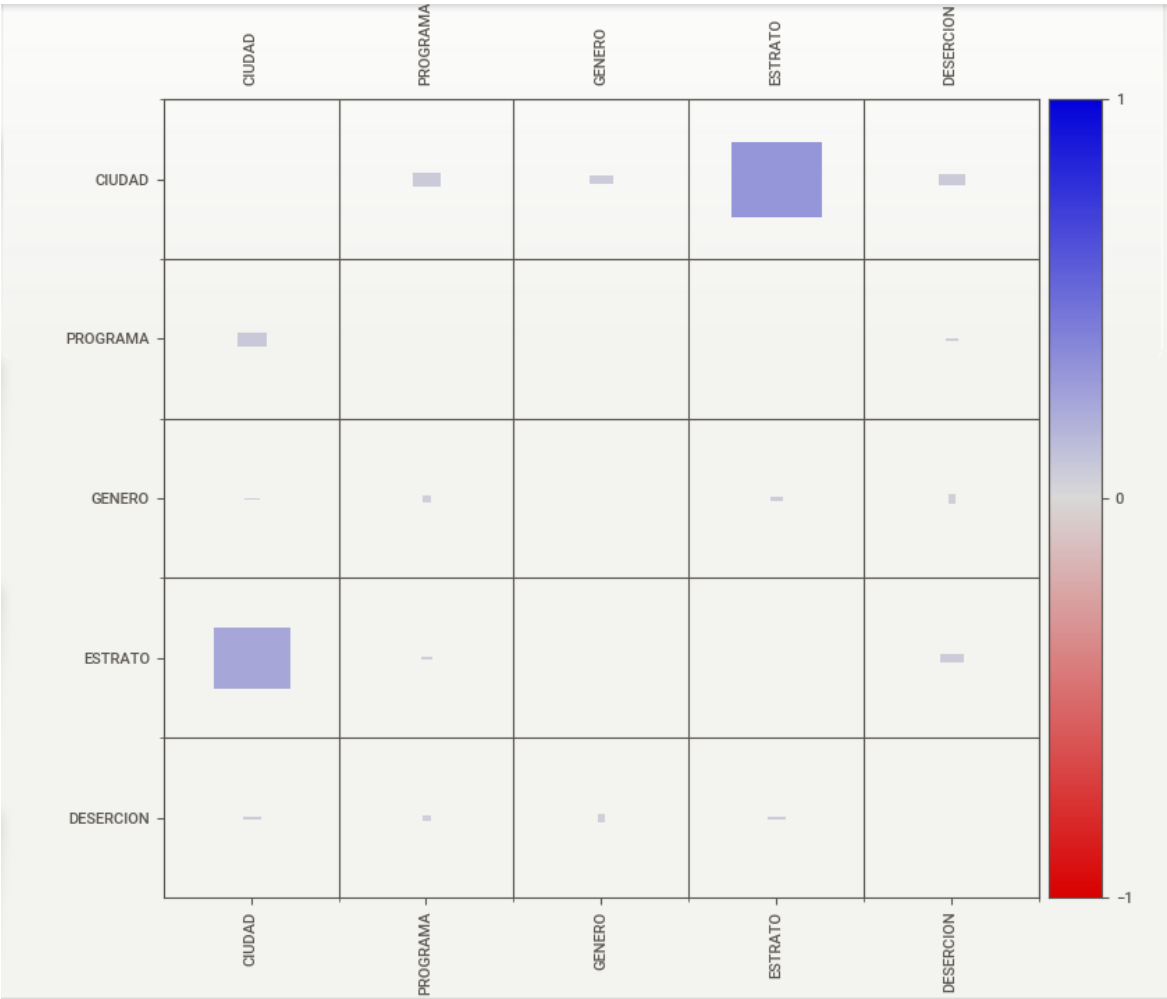
   Por género, los estudiantes de la muestra se distribuyen en femenino el 37% y masculino el 67%.

   En cuento a deserción, el 57% se considera posibles desertores y el 43% no se encuentran en riesgo de deserción.

   En la figura 1 se muestra la matriz de correlación de las variables analizadas.

Figura1.

Matriz de correlación de las variables incluidas en la base de datos.




Se evidencia una baja correlación entre las variables estrato socioeconómico y ciudad; las demás variables no muestran correlación en este estudio.

2. En la segunda fase se realiza una limpieza y extracción de características sobre la base de datos; la base inicial incluye las causas de deserción y estas las convertimos en la variable objetivo: Deserción.
3. La tercera fase nos lleva a definir las variables a utilizar en la elaboración del modelo, se tienen las variables explicativas: Género, Ciudad, Estrato Socioeconómico y Programa, mientras que la variable explicada es Deserción.

4. La cuarta fase requiere utilizar la base de datos ya organizada para proceder a implementar los diferentes modelos de clasificación y pronóstico, teniendo las variables a utilizar, definimos los datos que serán de entrenamiento (80%) y los de pruebas (20%).

5. Finalmente, se desarrollan los modelos de **Árbol de Decisión, Random Forest y Regresión Logística**. Para cada modelo se realiza la respectiva evaluación, utilizando la matriz de confusión relacionada para poder comparar y escoger el modelo que mejores resultados ofrece y presentar un pronóstico para los posibles casos de deserción en la población de estudio.

## RESULTADOS

Utilizando matriz de confusión para realizar la evaluación de los tres modelos implementados, encontramos que los valores de exactitud (Accuracy); Esto es la proporción de predicciones que cada modelo clasificó correctamente, es cercana a 0,6 para los tres modelos. Los mejores resultados corresponden al modelo Random Forest.

Tabla 1. Resultados comparativos de los modelos al ser evaluados por matriz de confusión

| | Árbol de decisión | Random Forest | Regresión logística |
|---|---|---|---|
| **Verdaderos positivos** | 0.43 | 0.26 | 0.33 |
| **Verdaderos negativos** | 0.59 | 0.79 | 0.68 |
| **Falsos negativos** | 0.57 | 0.74 | 0.67 |
| **Falsos positivos** | 0.41 | 0.21 | 0.32 |
| **Accuracy (Exactitud)** | 0.52 | 0.59 | 0.54 |

Acorde a lo especificado en la implementación del modelo, se ha definido con valor uno (1) a los desertores y con valor cero (0) a los no desertores. El modelo ha clasificado al 79% de los desertores de manera correcta (verdaderos negativos); en tanto que a los no desertores (verdaderos positivos) los ha clasificado de manera correcta en un 26%. Los no desertores, que fueron clasificados como desertores (falsos negativos) corresponde al 74% y los desertores, clasificados como no desertores (falsos positivos) representa el 21 %.

Análisis de resultados

La matriz de confusión aplicada al modelo Random Forest para evidenciar la forma como está realizando la clasificación de los datos, nos muestra que este modelo clasifica de manera correcta al 79% de los estudiantes desertores o en riesgo de deserción y al 21% restante de este grupo los clasifica como no desertores (falso positivo). Este resultado indica que el

modelo implementado puede ser utilizado como herramienta de predicción de la deserción; esto permite tomar las medidas necesarias para garantizar la retención y permanencia de los estudiantes que se encuentran en riesgo de deserción.

Para el caso de los estudiantes no desertores, el modelo está clasificando como desertores o en riesgo de deserción al 74% (falsos negativos); esto puede implicar mayores gastos en programas de prevención enfocado a estudiantes que realmente no están en riesgo de deserción.

# CONCLUSIONES

A través de la revisión del estado del arte sobre deserción, logramos identificar la importancia de este fenómeno en términos de la afectación que genera al sistema educativo y las múltiples causas que lo generan.

El modelo encontrado en este proyecto, contribuye a mitigar la problemática de la deserción estudiantil en la institución que facilitó la información para ser analizada. Se realiza una predicción del 79% sobre la deserción o abandono de estudios universitarios de los estudiantes que realmente están en riesgo.

Los modelos utilizados para este estudio fueron: árbol de decisión, Random Forest y regresión logística; a partir de las métricas de evaluación de los resultados obtenidos para cada modelo, encontramos que el algoritmo de clasificación Random Forest realizó una mejor predicción sobre la deserción estudiantil con una exactitud del 59%; la predicción de los posibles desertores se ubica en el 79%.

El modelo implementado puede ser utilizado como herramienta de predicción de la deserción, permitiendo acciones necesarias para garantizar la retención y permanencia de los estudiantes en la institución; esta labor es compartida por las instituciones y los padres de familia de los educandos.

## REFERENCIAS BIBLIOGRÁFICAS


Cabanzo, E. (2017). Las matemáticas y su influencia en la deserción universitaria. (Trabajo de grado). Universidad Militar Nueva Granada, Colombia.

Castaño, E., Gallón, S., Gómez, K. y Vásquez, J. (2006). Análisis de los factores asociados a la deserción y graduación estudiantil universitaria. Lecturas de Economía, 65, 9-36.

Castillo-Sánchez, Mario, Gamboa-Araya, Ronny, & Hidalgo-Mora, Randall. (2020). Factores que influyen en la deserción y reprobación de estudiantes de un curso universitario de matemáticas. *Uniciencia*, *34*(1), 219-245

Castro López, R. (2020). Aplicación de técnicas de Machine Learning para el estudio de deserción temprana y egreso oportuno en estudiantes de Ingeniería de la Facultad de Ciencias Físicas y Matemática

Donoso, S. y Schiefelbein, E. (2007). Análisis de los modelos explicativos de retención de estudiantes en la universidad: una visión desde la desigualdad social. *Estudios Pedagógicos, 33 (1), 7-27.*

Durán, D., Pérez-Almonacid, R., Rodríguez, A., Reverón, C. y Pinto, M. (2007). Cuestión de Supervivencia. Graduación, deserción y rezago en la Universidad Nacional de Colombia. Bogotá: Universidad Nacional de Colombia.

La Madriz, Jenniz (2016). Factores que promueven la deserción del aula virtual. Orbis. Revista Científica Ciencias Humanas, 12(35),18-40.

Montes, I., Almonacid, P., Gómez, S., Zuluaga, F. y Tamayo, E. (2010). Análisis de la deserción estudiantil en los programas de pregrado de la universidad EAFIT. Cuadernos de Investigación, 81.

Muñoz-Camacho, Samaria V., Gallardo, Teresita, Muñoz-Bravo, Meridalba, & Muñoz-Bravo, Carlos A. (2018). Probabilidad de Deserción Estudiantil en Cursos de Matemáticas Básicas en Programas Profesionales de la Universidad de Los Andes-Venezuela. *Formación universitaria*, *11*(4), 33-42.

Rodríguez-Urrego, Marcela. (2019). La investigación sobre deserción universitaria en Colombia 2006-2016. Tendencias y resultados. *Pedagogía y Saberes*, (51), 49-66

Sistema Nacional de Información de la Educación Superior. (2018). Ministerio de educación. Nacional. Recuperado de https://www.mineducacion.gov.co/sistemasinfo/spadies/Informacion-Institucional/357549: Estadísticas-de-Deserción

Tinto, V. (1987). Leaving college: Rethinking the causes and cures of student attrition. Chicago: University of Chicago Press