*Type of the Paper (Article)*

# Desarrollo de un modelo de predicción de casos de deserción universitaria en el área de ciencias de la salud en Colombia

**Hector Julian Robayo Barreto, Juan Sebastian Arias Ayala**

[1] Affiliation 1; hectorj-robayob@unilibre.edu.co
[2] Affiliation 2; juans-ariasa@unilibre.edu.co
*Received: 14/01/2021; Accepted: 18/04/2021.; Published: 30/06/2021

**Abstract:**
El poder predecir qué cantidad de estudiantes de educación superior pertenecientes al área de salud desertaran bajo un periodo estudiantil es una problemática que abarca varios factores de los cuales se tomará en cuenta el periodo estudiantil y el año en que curso, para estipular y predecir bajo que variables el estudiante abandonaría sus estudios, con el fin de mostrar los resultados y que entidades hagan un plan de acción, de una data que contenía estudiantes de varios programas, se hizo el filtrado de estudiantes que solo pertenecieran al área de la salud, numerando los periodos estudiantiles ya sean I o II, una vez tratada la data, se aplica el algoritmo de predicción con regresión, discriminando en una gráfica y así poder ingresar datos para su valoración.

**Keywords:** Estudiantes de Salud; Regresión; Periodos Estudiantiles

## 1. Introducción

El estudio de la deserción de estudiantes de la salud viene fundamentado en hechos que parten de la escasez de médicos profesionales en el ámbito, además de la importancia que abarca el interés de personas por el estudio y la correcta elección de su vida profesional [1][2], significa poder entender bajo que periodos el estudiante toma la decisión de disciplinarse en otro ámbito o enfocarse en cualquier otro tipo de cosas, así poder dar motivaciones o una dinámica de inclusión, una vez aplicado y con resultados positivos, la diversidad del programa hará que sea escalable a otros ámbitos estudiantiles, en base a lo que los ministerios de educación otorguen los datos, siendo código libre [3].

El siguiente contenido viene enfocado también en el aprendizaje de algoritmos de regresión para su práctica y posibilidad de aplicación y mejora en otros ámbitos de predicción donde se verá el procedimiento de ajuste de la data y agrupamiento, donde la organización del documento viene por los materiales y métodos proporcionados y efectuados en el segundo capítulo, después el resultado adquirido en el tercer capítulo, en el cuarto capítulo donde se discutirá la veracidad y eficacia recibida por la data, después de una conclusión en base a lo anterior en el quinto capítulo y la exposición de la patente como sexto capítulo.

## 2. Materiales y Métodos

Se busca en el proyecto realizar un modelo predictivo que permita predecir la cantidad de deserciones estudiantiles en las IES en el área de ciencias de la salud. En la Figura 1 se puede visualizar el paso a paso que se utilizó como referencia para poder desarrollar y probar un modelo predictivo [4]. A continuación, se explicará cada una de estas etapas.

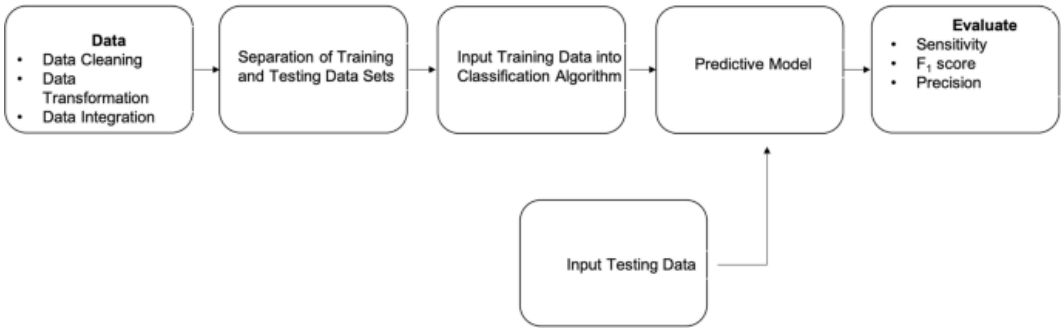



**Figura 1.** Representación del desarrollo de un modelo predictivo

### *2.1. Recolección de datos*

Los datos utilizados se obtuvieron del sistema de información SPADIES en su versión 3.0, versión donde se corrigen problemas de optimización y calidad de los datos, este sistema es una herramienta que permite realizar un seguimiento sobre las cifras de deserción estudiantil en la educación superior [3]. Las consultas en SPADIES se pueden realizar de forma pública, facilitando el acceso a la información sin necesidad de tener permisos especiales para acceder al sistema. Al momento de acceder al sistema, para realizar una consulta pública, se le pide al usuario que seleccione una de las variables que el sistema ofrece, estas variables se encuentran repartidas en las siguientes categorías:

- Características de las instituciones
- Características de los individuos (Saber 11)
- Características de los individuos (Saber Pro)
- Características de los programas académicos
- Programas de apoyo a los estudiantes
- Eventos cronológicos del estudiante

Este proyecto utilizó la variable "Área de conocimiento" que se encuentra en la categoría "características de los programas académicos", ya que se requieren estadísticas generales de los casos de deserción en las diferentes universidades de Colombia, se debe especificar que el tipo de cálculo que se requiere es en base a las IES (Instituciones de Educación Superior). El resultado de la consulta que muestra el sistema, como se muestra en la Figura 2, es un gráfico de líneas desde el año 1998-I hasta el año 2020-I, en base a las especificaciones que eligió el usuario. El sistema permite elegir qué tipo de datos consultar, porcentajes o cantidades, para el proyecto se eligieron cantidades, como lo requiere el problema.

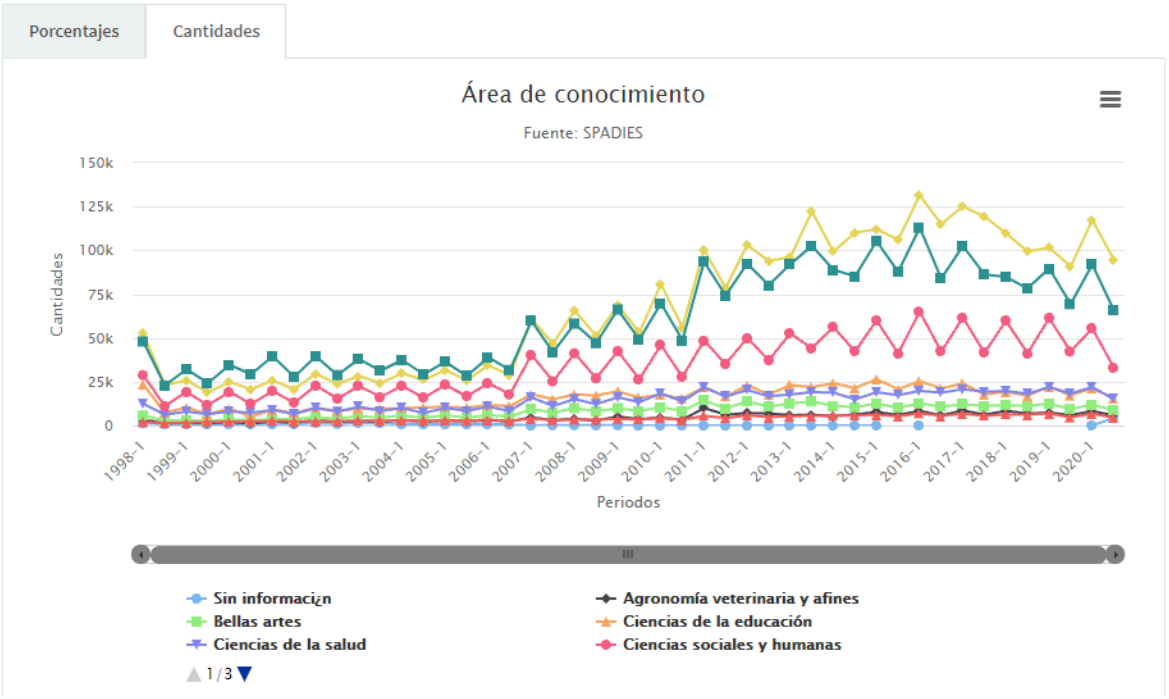



**Figura 2.** Grafica de líneas sobre los casos de deserción en las IES según el Área de conocimiento

En la parte posterior de la página se puede visualizar el DataFrame con todos los datos resultantes de la consulta (Figura 3), además de la opción de exportar los datos a un archivo de valores separados por comas (csv) y que es de interés para que el programa pueda manipular los datos que ahí aparecen.

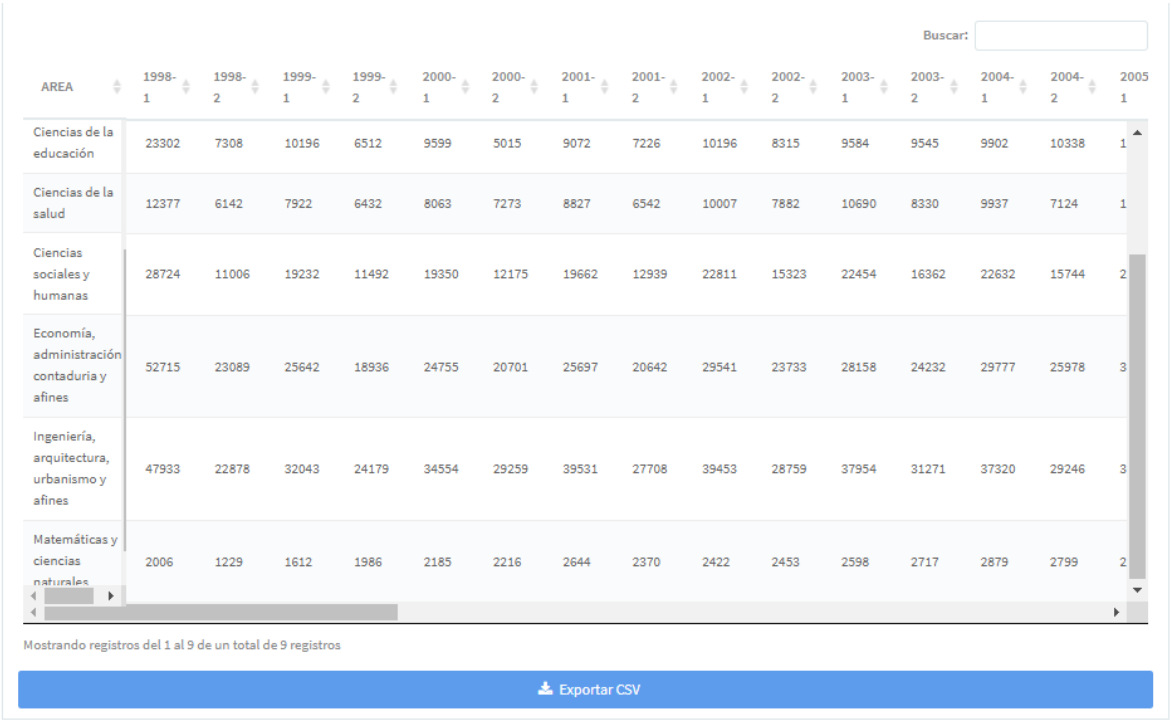



**Figura 3.** Marco de datos (DataFrame) de los casos de deserción estudiantil en las IES

*2.2. Preprocesamiento de datos*

Para las etapas posteriores se utilizó el lenguaje de programación Python, es de código abierto y es ampliamente utilizado en la ciencia de datos debido a su simple sintaxis y a la gran variedad de librerías que pueden utilizarse para la aplicación de la ciencia de datos [5]. Las librerías de Python que se utilizaron para el preprocesamiento de datos fueron NumPy y Pandas, junto a la librería Matplotlib que permite graficar los datos.

Después de importar las librerías se cargó el dataset, es decir, el archivo csv que se obtuvo del sistema SPADIES 3.0, luego se realizó el proceso de limpieza de datos, que consiste en detectar y eliminar errores e inconsistencias de los datos, así como datos innecesarios, para mejorar la calidad de los datos. Aplicando esto en el proyecto, se seleccionó únicamente del DataFrame el área de Ciencias de la salud con sus respectivos periodos y cantidad de casos de deserción, luego se verificó que no hubiese datos erróneos o nulos. Finalmente se agregaron los encabezados y se reiniciaron los índices como se ve en la Figura 4.

| | Año | Numero de deserciones |
|---|---|---|
| 1 | 1998-1 | 12382 |
| 2 | 1998-2 | 6262 |
| 3 | 1999-1 | 8016 |
| 4 | 1999-2 | 6509 |
| 5 | 2000-1 | 8149 |

**Figura 4**. DataFrame luego de aplicar la limpieza de datos

*2.3. Separación de los conjuntos de datos de entrenamiento y de prueba*

Los datos de entrenamiento son un conjunto de ejemplos, del mismo tipo de la data original, que permite ajustar parámetros del modelo [6]. Los datos con los que el algoritmo entrenó fueron tomados de el dataset que se cargó previamente. Los datos de prueba, debido a los limitados datos que se pudieron encontrar, fueron tomados del mismo DataFrame. La distribución, en porcentaje, de los datos que se tomaron del DataFrame para el entrenamiento y las pruebas fue: 80% para los datos de entrenamiento y 20% para los datos de prueba.

*2.4. Implementación del algoritmo de regresión*

Los anteriores pasos permitieron ordenar los datos que necesita el algoritmo, ahora lo que se busca es implementar un algoritmo de regresión que se adecue al comportamiento que tienen los datos, el comportamiento de los datos se puede visualizar por medio de las gráficas de dispersión (Figura, para ello se utilizó la librería Matplotlib previamente mencionada.

Modelo de regresión lineal simple:

$$y_i = \beta_0 + \beta_1 \mathrm{X}_i + \varepsilon_i \qquad (1)$$

En la ecuación (1) se muestra el modelo matemático de la regresión lineal simple donde $y_i$ corresponde a la variable dependiente, $\beta_0$ es el coeficiente o intercepto en el eje $y$, $\beta_\mathrm{i}$ es la pendiente

de la regresión, $X_i$ es la variable independiente, y $\varepsilon_i$ es el margen de error que utiliza valores aleatorios.

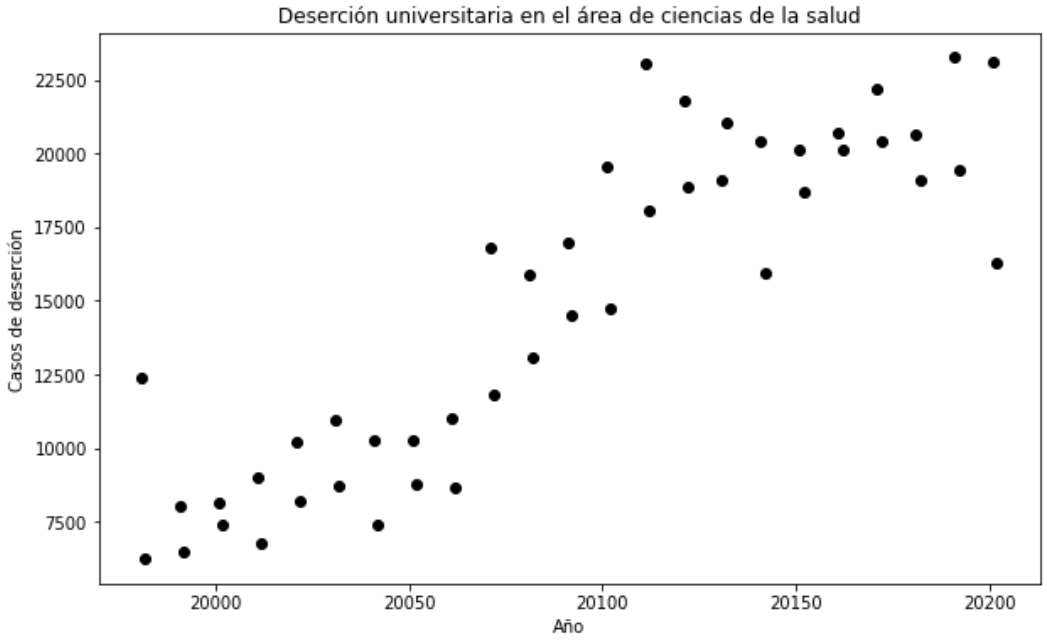



**Figura 5**. Gráfica de dispersión de los casos de deserción universitaria en el área de ciencias de la salud

En la anterior grafica (Figura 5) se puede identificar que según el comportamiento de los datos una de las opciones optimas a utilizar es la regresión lineal. La regresión lineal utiliza una ecuación lineal que mejor se adapte a los datos de entrenamiento y así predecir con una mayor probabilidad el valor de la variable dependiente [7], que en este caso son el número de deserciones universitarias en el área de ciencias de la salud.

*2.5. Evaluación del modelo*

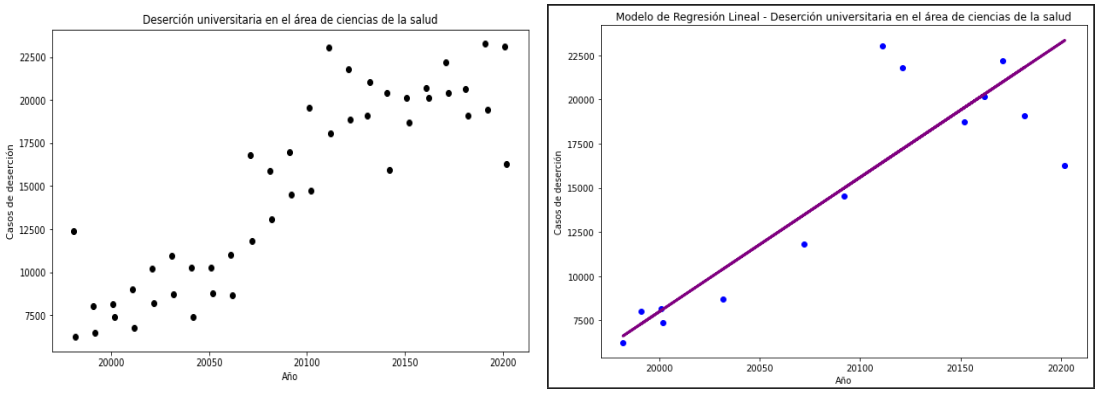
Anteriormente en la etapa de separación de los conjuntos de datos de entrenamiento y de pruebas se indicó que se tomaron un porcentaje de los datos de entrenamiento para que posteriormente se probará el algoritmo de regresión, a partir de este conjunto de datos se puede analizar la precisión que tiene el modelo de regresión que se desarrolló, esta precisión se da en porcentaje y lo que se busca con este dato es analizar la fiabilidad del algoritmo utilizado, puede que el algoritmo seleccionado no sea la mejor opción para el data set que se trabaja por lo que esta medición permitirá escoger el algoritmo óptimo para esos datos.

## 3. Resultados

Los resultados se consiguieron utilizando un valor predefinido de la variable dependiente, en este caso el año junto al periodo, a partir de esto haciendo uso del algoritmo se realizó una predicción de los casos de deserción para ese periodo de tiempo en concreto. La precisión del algoritmo, que también se muestra en el programa, se calculó a partir de la efectividad que tuvo el algoritmo para predecir los valores de prueba que se le asignaron, así se tuvo un punto de referencia para comparar los resultados de las predicciones y los valores reales a los que estaban asociados.

Debido a la aleatoriedad a la hora de escoger cuales eran los datos tanto de entrenamiento como de prueba, el porcentaje de precisión que obtiene el algoritmo varía con cada ejecución del programa, sin embargo, aunque esto ocurra si se utiliza el correcto modelo de predicción su precisión se mantendrá en un rango similar la mayor parte de las veces que se ejecute el programa.

Se obtuvieron datos de predicción con una margen de precisión del modelo en un 81.31%, teniendo un índice de porcentaje obtenido mediante la regresión lineal, prediciendo que para el año 2030, periodo 1; habrán 30864.06 casos de deserción universitaria en el área de ciencias de la salud. El programa arrojó las correspondientes graficas de dispersión de los datos y del modelo de regresión lineal aplicado a los datos de pruebas. Véase la Figura 6.




(a) (b)

**Figura 6.** (a) En esta grafica se denota un número grande de datos ubicados según la estadística de las tablas, tomando como margen los periodos de los años en eje X y la cantidad de personas que desertan en el eje Y. (b) En esta grafica se denota el resultado de la regresión línea en los datos de prueba que el algoritmo seleccionó.

Finalmente, para realizar una medición de la precisión del algoritmo se optó por realizar en total diez pruebas, véase la Tabla 1, donde variaba el periodo de la predicción para así realizar un promedio de los datos de precisión arrojados en las diez pruebas. El resultado en promedio fue del 80.67%, este resultado era de esperarse ya que como anteriormente se mencionó los datos con los que se trabajaron contaban con ruido, lo cual dificultó obtener una predicción más precisa al algoritmo.

**Tabla 1.** Resultado de las pruebas donde se muestran los valores obtenidos en cada una de las diez pruebas junto con el promedio del porcentaje de precisión de cada prueba.

| N° de prueba | Periodo (Año-periodo) | Predicción (Casos de deserción) | Precisión (%) |
|---|---|---|---|
| 1 | 2030-1 | 30864.06 | 81.31 |
| 2 | 2035-2 | 34719.57 | 81.27 |
| 3 | 2043-1 | 40448.53 | 84.34 |
| 4 | 2046-2 | 44813.69 | 85.18 |
| 5 | 2055-2 | 46265.12 | 71.87 |
| 6 | 2059-1 | 50606.1 | 77.04 |
| 7 | 2062-2 | 59722.26 | 84.68 |
| 8 | 2067-1 | 58501.93 | 80.21 |

| | | | |
|---|---|---|---|
| 9 | 2070-1 | 58874.4 | 79.56 |
| 10 | 2080-2 | 70877.3 | 81.28 |
| | | **Promedio** | 80.67 |

## 4. Discusión

Siendo una cifra considerable para entidades educativas. Afirmando que mientras se alcance un mayor rango de eficiencia del algoritmo, se puede dar una mejor definición y declaratoria a instituciones educativas con el fin de promover el cambio desde el proceso de admisión hasta el mejor rendimiento prestando los recursos educativos. Por otro lado, el uso del algoritmo presentado sujeta a datos equivalentes en tamaño al trabajado no llega a cumplir con el requerimiento de acertar lo más cercano al 95% o mayor, siendo ineficiente, pero estipulando que mediante el uso de algoritmos de predicción y datos que apunten a funciones del tipo no lineal, se puede considerar viable el uso de ordenamiento, lectura y hallazgo del presente trabajo.

## Referencias

1. C. Miguel and M. Amparo, "Determinantes de la deserción estudiantil en estudiantes universitarios." Universidad de Cartagena, Panorama Económico, 2019
2. A. Mauricio and B. María," Estudio sobre Deserción Estudiantil Universitaria y sus Implicaciones Académicas, Económicas y Sociales" Boletín de Coyuntura, 2018
3. Sistema Para la Prevención de la Deserción de la Educación Superior, Ministerio de Educación (2021). https://www.mineducacion.gov.co/sistemasinfo/spadies/Informacion-Institucional/363411:SPADIES-3-0
4. Bradley, J., Rajendran, S. Increasing adoption rates at animal shelters: a two-phase approach to predict length of stay and optimal shelter allocation. BMC Vet Res 17, 70 (2021). https://doi.org/10.1186/s12917-020-02728-2
5. Chandan07, Python for Data Science. GeeksforGeeks (2021). https://www.geeksforgeeks.org/python-for-data-science/
6. Xu, Yun; Goodacre, Royston (2018). "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning". Journal of Analysis and Testing. https://doi.org/10.1007%2Fs41664-018-0068-2
7. IBM, Regresión lineal (2020). https://www.ibm.com/co-es/analytics/learn/linear-regression