# Tarea 5– Proyecto Análisis de Datos

Lindsay Andrea Quintero Hernández

202016908A_1701: Análisis de Datos

Grupo 20

Tutor. Ferley Medina Rojas

**Programa. Ingeniería de Sistemas**

**Universidad Nacional Abierta y a Distancia -UNAD**

**Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI**

**Zona Centro Oriente - ZCORI (CEAD Bucaramanga)**

**Mayo del 2024**

## Introducción

El análisis supervisado evalúan modelos etiquetados para llegar a conclusiones según modelos matemáticos en seguimiento del comportamiento de los datos, con el fin de llegar a conclusiones, estos procesos requieren de estudios previos de los mismos “EDA”, así como implementación de funciones y métodos que permitan el tratamiento de datos en marcos de modelos para el análisis de esta información; ahora bien, en este caso se realizó aplicación de regresión lineal para reconocer la relación entre variables, el comportamiento de variables dependientes con respecto a las variables independientes, esto aplicado a la base de datos Titanic que presenta datos sobre los tripulantes de esta embarcación, el objetivó fue analizar los datos para encontrar patrones en seguimiento a los sobrevivientes de la embarcación, en este caso se analizó según la edad, el sexo, la clase y la variable sobreviviente, se presentaron modelos y graficas que aportaron al proceso de reconocimiento de resultados, aplicando técnicas y conocimientos aprendidos en el desarrollo del curso análisis de datos, además se aportó con información al foro, lo qué brindó una visión clara y concisa de los modelos de aprendizaje supervisados, así como la aplicación en diversos contextos. Para finalizar, se proporciona enlaces a recursos a la evidencia en repositorio en GitHub.

# Objetivos

## Objetivo general

Aplicar algoritmos de Machine Learning supervisado según el caso de estudio en aplicación modelos de regresión y clasificación.

## Objetivo especifico

Identificar descargar y descargar base datos para realizar su análisis exploratorio o EDA.

Comprender conceptos relacionados con aprendizaje supervisado, así como modelos de regresión y clasificación.

Identificación de características de modelos de aprendizaje supervisado como regresión lineal, regresión logística, arboles de decisión.

Aplicar modelos de aprendizaje supervisado para llegar a conclusiones manipulación y análisis de la información.

Presentar evidencia y consolidar documentos en repositorio GitHub.

## Evidencia

**Enlace GitHub:** https://github.com/122309/Tarea5_LindsayQuintero_AD.git

## Interpretación de los resultados

### EDA

Según el análisis exploratorio de datos se reconoce que la base de datos está compuesta por identificación de pasajero, sobreviviente, nombre, edad, sexo, clase socio económica, parch, tiquete, tarifa, cabina, embarcación; de los cuales se tienen 891 registros, estos datos deben ser reestructurados para poder ser procesados y analizados, por ejemplo, en el caso de la edad, se deben completar los datos ya que la tabla está incompleta en algunos campos de esta columna, así como reemplazar los datos nulos del atributo Cabina.

Se observa que hay valores nulos o faltantes en la columna edad, cabina y embarcación, por otra parte se observa que los datos atípicos de la edad son 11, estos datos están fuera de los rangos inferiores y superiores del diagrama de cajas y bigotes, en este caso estos datos se reemplazaron con la media, ahora bien, una vez realizada la limpieza de datos, se presnta la información por medio de “info()”, donde se identifica qué id, sobreviviente, clase, sexo, SibSP y Parch son de tipo int64; el nombre y el ticket son de tipo objeto; la edad y la tarifa son de tipo float64 y la cabina y la embarcación son de tipo bool.

Se observan 891 registros, se observa que la final age está incompleta, así que hay que agregar datos al modelo, se observa que la variable sobreviviente es una variable categórica con valor 0 o 1 (no sobrevivió, sobrevivió), la PClase tiene velor 0, 1, 2 o 3, dependiendo de la clase socio económica de la persona, el sexo puede se femenino a masculino, el SibSp va de 0 a 8, el parch de 0 a 6; por otra parte,  las edades de la tripulación van de aproximadamente 5 meses o 0,47 años y 80 años, la mayoría de la tripulación de 29 a 30 años según la media, se podría decir que la mayoría está entre 28 y 30, debido a qué según el cuartil 2 o percentil 50 es el valor en posición central y no tiene afectaciones de valores extremos que la media aritmética, útil en valores atípicos, por otra parte, se observa que 577 son hombres y 314 son mujeres, ahora bien, la tarifa va de 0 a 512,3292 según el pasajero.

## Conclusiones

El aprendizaje supervisado evalúa modelos de datos etiquetados, para procesar esta información es necesario hacer un reconocimiento de los datos, esto se conoce como EDA o análisis exploratorio de datos para identificar la base de datos, qué tipo de datos la componen, así como información estadística relevante que aportará en el logro del objetivo, para esto también se requiere hacer una limpieza de los datos para eliminar o reducir valores nulos, valores ceros y valores atípicos que afecten el análisis; se debe proceder con la identificación de variables dependientes e independientes para aplicación de modelo de regresión lineal, se procede a dividir la base de datos en train y test para finalmente aplicar el modelo e identificar tendencia de datos en las graficas y resultados, esto aplica a gran cantidad de base de datos y permite reconocer patrones en datos etiquetados, así se estructura el aprendizaje máquina, además de apropiación de análisis de datos, así como aplicación de conceptos que aportan significativamente a mi formación como futura ingeniera de sistemas en profundización de sistemas de información.

## Referencias Bibliográficas

Dangeti, P. (2017). *Statistics for Machine Learning : Build Supervised, Unsupervised, and Reinforcement Learning Models Using Both Python and R. Packt Publishing.* Retrieved from Biblioteca virtual: https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1560931&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp_Cover Cap 8

Guiseppe Banaccorso. (2018). *Machine Learning Algorithms : Popular Algorithms for Data Science and Machine Learning, 2nd Edition: Vol. 2nd ed. Packt Publishing.* Retrieved from Biblioteca virtual: https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1881497&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp_Cover Cap 9 y 11

Maquina, C. (2023, 07 24). *¿Qué tan buenos son tus Clusters? Métricas para Clustering con Python: Silueta y Davies Bouldin*. Retrieved from YouTube: https://www.youtube.com/watch?v=b920s9nXGao&ab_channel=CodigoMaquina

Miguillon , J., Caihuelas Quiles, R., Casas Roma, J., & Gironés Roig, J. (2017). *Minería de datos: modelos y algoritmos. Editorial UOC.* Retrieved from E Libro: https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/58656. Cap 7 y 8

UNAD. (2024). *Analisis de Datos - (202016908A_1701)*. Retrieved from Guía de actividades y rúbrica de evaluación - Tarea 5 - Proyecto Análisis de datos: https://campus107.unad.edu.co/ecbti131/mod/folder/view.php?id=137

Veliz Capuñay , C. (2020). *Aprendizaje automático. Introducción al aprendizaje profundo. El Fondo Editorial de la Pontificia Universidad Católica del Perú*. Retrieved from Biblioteca virtual: https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2600876&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp_I Cap 3