**Análisis De Datos - (202016908A\_1394)**

**Etapa 5 - Resultado de análisis de datos**

Pablo Emilio Escobar Ossa

Universidad Nacional Abierta y a Distancia – UNAD.

Escuela de Ciencias Básicas Tecnología e Ingeniería

Programa Ingeniería de sistemas

2023

**Tabla de Contenidos**

[Introducción 1](#_Toc152796986)

[Descripción del ejercicio 2](#_Toc152796987)

[Importación de bibliotecas 2](#_Toc152796988)

[Lectura de datos 2](#_Toc152796989)

[Cambio de tipo de valor de variables 2](#_Toc152796990)

[Selección de variables 2](#_Toc152796991)

[Dividir datos 3](#_Toc152796992)

[Crear el modelo Bosques Aleatorios 3](#_Toc152796993)

[Transformación de datos NaN 4](#_Toc152796994)

[Entrenamiento 4](#_Toc152796995)

[Evaluar modelo 4](#_Toc152796996)

[Matriz de confusión 4](#_Toc152796997)

[Conclusiones 6](#_Toc152796998)

[Lista de referencias 7](#_Toc152796999)

# Introducción

El objetivo de este es describir el proceso de implementación de un modelo de clasificación para predecir la probabilidad de sobrevivir a una situación como la que vivieron las personas en el Titanic.

# Descripción del ejercicio

## Importación de bibliotecas

Es la fase inicial del proceso, aquí se cargan las librerías necesarias para el desarrollo del ejercicio, tales como matplotlib, mlxtend, numpy, pandas y sklearn. Estas son las librerías necesarias para el funcionamiento de nuestro algoritmo.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente con confianza media

## Lectura de datos

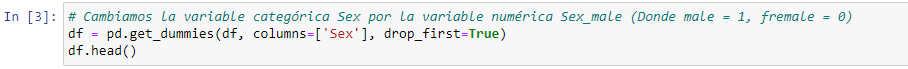
Se utiliza la función ***read\_csv()*** de pandas para leer el archivo de datos “train.csv” que se encuentra dentro de la carpeta “data”. Este archivo contiene la información necesaria para el funcionamiento de nuestro algoritmo.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

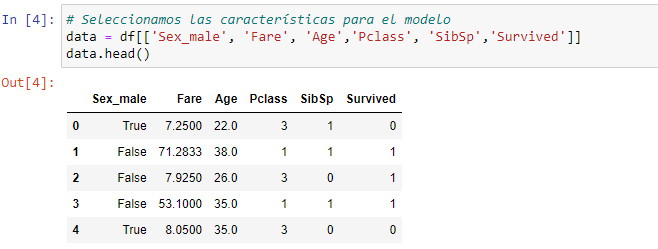
## Cambio de tipo de valor de variables

Cambiamos el tipo de valor de la variable “Sex” por la variable numérica “Sex\_male”, donde male será igual a 1 y female será igual 0. Esto con el fin de poder trabajar mejor esa variable haciendo su respectivo cast.



## Selección de variables

Ahora seleccionamos las variables que vamos a usar para nuestro modelo en el cual también llevamos la variable “Survived” que es la variable que debemos tener en cuenta para predecir en nuestro modelo.



## Dividir datos

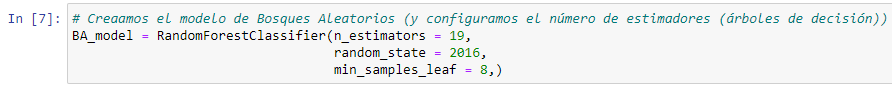
Se dividen los datos del modelo para sacar los datos de entrenamiento y los datos de prueba para nuestro modelo, donde “X” son nuestras variables independientes y “y” son nuestras variables dependientes. Hacemos una división donde el 75% de los datos del modelo son para el entrenamiento y el 25% son para los datos de prueba.

Texto

Descripción generada automáticamente

## Crear el modelo Bosques Aleatorios

Ahora se procede a crear el modelo de bosques aleatorios y configuramos el número de estimadores, es decir, los arboles de decisión. También se configura una semilla para la aleatoriedad y la mínima cantidad de muestras. Una semilla normalmente se usa para que un proceso aleatorio se pueda replicar. Es decir, existen procesos aleatorios (como por ejemplo cuando se divide un dataset) que se requieren replicar para efectos de depuraciones. Cuando pones una semilla te aseguras que, aunque la naturaleza del proceso es aleatorio, se pueda replicar cada vez que se corra el código.



## Transformación de datos NaN

Se transforman los datos para poder imputar aquello en los que tiene como valor NaN. Esto se hace para prevenir problemas con el dataset y tener los datos lo más limpios posibles antes del entrenamiento.

Texto, Aplicación

Descripción generada automáticamente

## Entrenamiento

Ahora se procede a realizar el entrenamiento en dónde enviamos las variables independientes y las variables dependientes.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

## Evaluar modelo

Usando los datos de prueba, procedemos a evaluar el modelo obteniendo el accuracy promedio.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

## Matriz de confusión

Procedemos a construir una matriz de confusión. Se hace una predicción usando los datos de prueba.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Enlace del repositorio:**

# Conclusiones

En conclusión, podemos determinar que el uso de árboles de decisiones a través de librerías ofrecidas por Python podemos mejorar la precisión y la exactitud del algoritmo a través de los arboles aleatorios. Así mismo, pudiendo saber cual es la probabilidad de que una persona se pueda salvar en una situación similar al titanic.

# Lista de referencias

Titanic - Machine Learning from Disaster. <https://www.kaggle.com/competitions/titanic>

KNIME. Open for Innovation.  
<https://www.knime.com>

Garriga Trillo, A. J. (2009). Introducción al análisis de datos. UNED - Universidad Nacional de Educación a Distancia.   
<https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/48460?page=130>

Carlos Veliz. (2020). Machine learning. Introduction to deep learning . The Editorial Fund of the Pontifical Catholic University of Peru. <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_113> . Capítulo 4.

Romero Villafranca, R. y Zúnica Ramajo, L. (2020). Métodos estadísticos para ingenieros. Editorial de la Universidad Politécnica de Valencia.   
<https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/129644?page=212>

Minguillón, J. Casas, J. y Minguillón, J. (2017). Minería de datos: modelos y algoritmos. Editorial UOC.   
<https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/58656?page=208>

Sepúlveda, J. F. D., & Morales, J. C. C. (2013). Comparación entre árboles de regresión CART y regresión lineal. Comunicaciones en Estadística, 6(2), 175-195. <https://dialnet.unirioja.es/servlet/articulo?codigo=7393722>