

# **Modelo de Aprendizaje Automático para Análisis de Datos**

## **Guillermo Romeo Cepeda Medina**

### **A01284015**

#### Descripción

Este repositorio incluye un modelo de Aprendizaje Automático (Machine Learning) diseñado para el análisis de datos. El modelo está implementado en Python y utiliza diferentes funciones y librerías para crear un algoritmo de machine learning. En particular, se ha desarrollado un algoritmo de clasificación utilizando regresión logística para múltiples variables.

#### Dataset

El modelo utiliza el conjunto de datos "DatasetML" para entrenar y evaluar su rendimiento. Este dataset se compone de datos etiquetados recopilados de <https://www.kaggle.com/datasets/alexandrepetit881234/fake-bills>. El dataset tiene 1500 filas y 7 columnas con las siguientes variables:

is\_genuine: booleano  
diagonal: número decimal  
height\_left: número decimal  
height\_right: número decimal  
margin\_low: número decimal  
margin\_upper: número decimal  
length: número decimal

Se utilizan 6 variables independientes y una variable dependiente con el objetivo de predecir si un billete es falso o verdadero.

#### Cambios Implementados

En esta sección se describen los cambios implementados en el modelo en comparación con la versión anterior para abordar los comentarios del docente:

Cambio Indicado: El docente sugirió cambiar los valores de las  $\theta$ s de las variables independientes.

Acciones Tomadas:

Se modificó la función de hipótesis para que fuera usada en cada  $x_i$  como la suma de un producto punto de ( $\theta$ s,  $x$ ) y  $\theta_0$ .

Se realizó un análisis de datos previo para identificar las variables más importantes en el modelo.

Se empleó el modelo de Gradient Boosting para identificar las variables significativas: "Length", "Margin Low" y "Margin Upper".

Se aplicó la función MinMaxScaler de sklearn para normalizar los datos y tenerlos en un rango de 0 a 1.

Se realizó un análisis de diferentes  $\theta$ s para mejorar la predicción.

Estructura del Código

En el README, se puede leer de mejor manera la implementación del código, mientras que aquí se realizará una discusión de lo implementado en este repositorio.

Se logró implementar con éxito un algoritmo de regresión logística con sus debidas modificaciones al dataset, así como la implementación de otros algoritmos para su correcto funcionamiento, además se compararon resultados con la librería de SKlearn y se concluyó que ambos tienen un comportamiento bastante similar.