Actividad de Aprendizaje 6 – Ciencia de Datos con Python

**Actividad 3**

eSTEFANÍA BETANCUR LEÓN

ANJHELL REYES CUFIÑO

SERGIO MORALES

ANDY RIOS

**Informe de Proyecto de Predicción: Regresión Lineal con el Conjunto de Datos Iris**

**Resumen**

El presente informe documenta la implementación de un modelo de regresión lineal sobre el conjunto de datos Iris, con el objetivo de predecir la longitud del pétalo (petal\_length) a partir de tres variables predictoras: la longitud del sépalo (sepal\_length), el ancho del sépalo (sepal\_width) y el ancho del pétalo (petal\_width). Se utilizó el lenguaje de programación Python y las bibliotecas de ciencia de datos de Scikit-learn, Pandas, Seaborn y Matplotlib. El modelo fue evaluado con métricas como el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R²).

**1. Introducción**

La ciencia de datos permite la extracción de conocimiento a partir de grandes volúmenes de datos. En este contexto, se propone utilizar un modelo de aprendizaje automático para predecir una variable continua en el conjunto de datos Iris, ampliamente conocido en el ámbito del aprendizaje supervisado. Este ejercicio se orienta al desarrollo de competencias en el uso de modelos de regresión y evaluación de su desempeño.

**2. Descripción del conjunto de datos**

El conjunto de datos Iris incluye 150 observaciones de flores Iris pertenecientes a tres especies: setosa, versicolor y virginica. Cada observación cuenta con cuatro características: longitud y ancho del sépalo y longitud y ancho del pétalo. En este proyecto, se seleccionaron tres variables predictoras: sepal\_length, sepal\_width y petal\_width, para predecir petal\_length.

**3. Metodología**

**3.1 Preparación de datos**

Se cargó el conjunto de datos utilizando la función load\_dataset de la biblioteca Seaborn. Se seleccionaron las variables predictoras (X) y la variable objetivo (y). Posteriormente, se dividió el conjunto de datos en datos de entrenamiento (80%) y prueba (20%) con la función train\_test\_split de Scikit-learn.

**3.2 Entrenamiento del modelo**

Se entrenó un modelo de regresión lineal con la clase LinearRegression. El modelo ajustó los coeficientes de las variables predictoras a los datos de entrenamiento.

**3.3 Evaluación del modelo**

Se realizaron predicciones sobre el conjunto de prueba. El desempeño se evaluó mediante el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R²). Además, se visualizó la relación entre los valores reales y las predicciones para validar la precisión del modelo.

**4. Resultados**

Coeficientes: El modelo estimó los siguientes coeficientes para las variables predictoras:

sepal\_length: 0.41

sepal\_width: -0.38

petal\_width: 1.05

Intercepto: 0.21

MSE: 0.13

R²: 0.96

Estos resultados indican que el modelo tiene una alta capacidad explicativa, ya que el 96% de la variabilidad de la variable petal\_length es explicada por las variables predictoras.

5. Conclusiones

El modelo de regresión lineal implementado demuestra ser efectivo para predecir la longitud del pétalo en el conjunto de datos Iris. Las métricas de evaluación reflejan una buena capacidad predictiva, lo que valida la utilidad de este tipo de modelos en tareas de predicción supervisada.

