Pharos University in Alexandria Faculty of Computer Science & Artificial Intelligence Level-1 – Spring 2024/2025

Intro to data science
Course Code: DS 102
Lecturer : Dr. Shereen Shawky

TA: Eng. Gasser

Data science final project description

Prepared by:

1- Mahmoud soso Abdallah	:	202400659
2- Omar Moustafa Mahmoud	:	2024002701
3- Mahmoud Hossam	:	202400738

Dataset Details:

Dataset Name: [iris.csv]

Source: [Kaggle]

Description:

This dataset was retrieved from Kaggle. It contains 150 records about flowers and

SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species

Data Exploration

- Number of rows: 150

- Number of columns: 6

- No missing values or duplicates.

The dataset was loaded using the **Pandas** library to explore its structure and contents. The following steps were performed during the exploration phase:

- The data was read from a CSV file named **"iris.csv"**.
- We used the head() function to preview the first few rows of the dataset and understand the format of each column.
- We used shape and info() to inspect the number of rows and columns, as well as the data types and presence of any null values.
- We discovered that some columns like **"species"**, **"SepalLengthCm"**
-
- We dropped the **"ID"** column, as it doesn't provide useful information for analysis.
- After processing, the dataset was ready for modeling with all features in numeric format.

Data Processing:

During data processing, we focused on handling outliers and encoding

categorical data:

● To identify species outliers, we calculated the Interquartile Range

(IQR) as follows:

○ Calculated Q1 (25th percentile) and Q3 (75th percentile).

○ Computed IQR = Q3 - Q1.

○ Defined lower and upper bounds to detect potential outliers.

● The detected outliers were inspected to decide whether to retain or

remove them.

Data Modeling – Linear Regression or SVC

To predict **species**, we chose **Linear Regression** as our primary model due to its effectiveness in handling numerical data prediction tasks.

- The model was trained using the training portion of the dataset with the following steps:
  ○ The data was split into training and testing sets (80% for training, 20% for testing).
  ○ A **Linear Regression** model or super vector classiver was fitted using the training data.
  ○ We used the trained model to predict car prices from the test set and stored the results in a variable called y_pred.
  ○ We calculated the error by comparing the **actual vs. predicted** price values using metrics like **Mean Squared Error (MSE)** and **R² score** to evaluate model performance.

Data Visualization:

We used various visualization techniques to explore the relationships

```
import matplotlib.pyplot as plt

import plotly.express as px

import seaborn as sns
```

between variables:

Scatter Plot

Pie plot

Box plot

Histogram

Line

Bar