Lambton College | In Mississauga

AML-2203 Advanced Python AI and ML Tools

Assignment#3:

Predicting Disease Progression Using Linear Regression

Use : Google colab

Objective

To understand and apply linear regression using Scikit-learn by analyzing the diabetes dataset. Students will explore data preprocessing, model training, evaluation, and interpretation of results.

Background

The diabetes dataset is a classic dataset provided by Scikit-learn that includes medical information for diabetes patients. The goal is to predict disease progression one year after baseline measurements, using ten baseline variables.

Dataset Features

The dataset includes the following input variables (features):

age

sex

body mass index (BMI)

average blood pressure

six blood serum measurements

The target variable (y) is a quantitative measure of disease progression after one year.

Instructions:

Exploratory Data Analysis (EDA)

Load the diabetes dataset using Scikit-learn’s load_diabetes() function, or import from diabetes.csv if using a CSV version.

Display the first few rows and understand the data distribution.

Visualize relationships between features and the target variable using scatter plots or correlation heatmaps.

Data Preprocessing

Normalize or standardize the features if necessary.

Check for missing values and handle them appropriately.

Split the dataset into training and testing sets (e.g., 80/20 split).

Model Building

Use LinearRegression from Scikit-learn to train your model.

Fit the model on the training data.

Model Evaluation

Predict values on the test set.

Evaluate performance using:

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

R-squared (R^2) score

Plot the predicted vs actual values

Model Interpretation

Print out model coefficients and intercept.

Discuss which features have the most significant impact on diabetes progression.

Reflect on any assumptions of linear regression and whether they seem to be satisfied.

Report/Presentation

Provide a summary of findings, including:

Which variables were most predictive

Overall model performance

Insights from the visualization

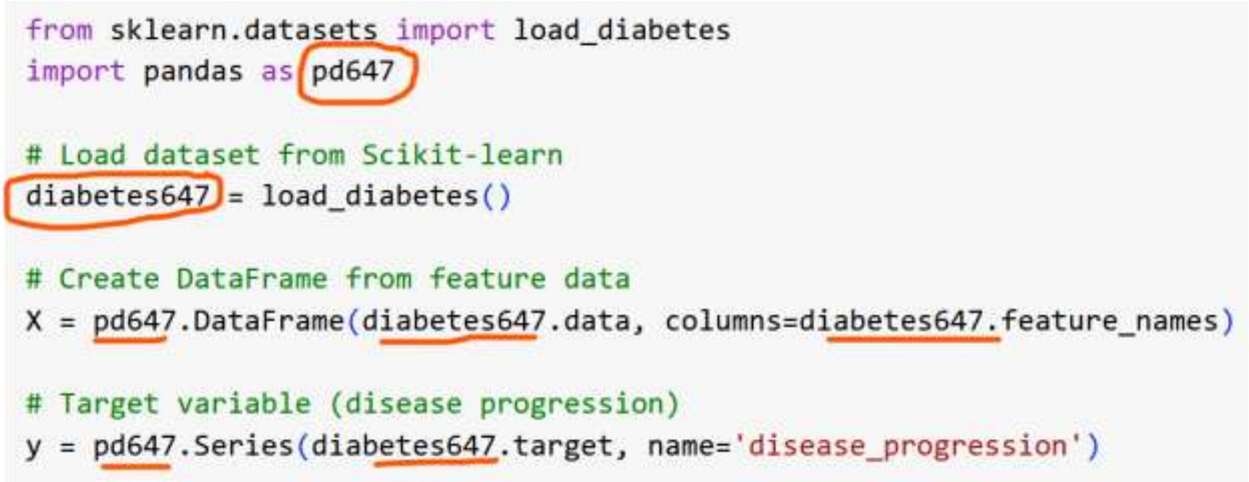
Note:

Create comments with your code blocks to make it unique

Explain the function used using your own words (not AI generated)

Suffix last 3 digis of your student number to variable to make it unique

Sample screen to show how to use customized variables:



```
from sklearn.datasets import load_diabetes
import pandas as pd647

# Load dataset from Scikit-learn
diabetes647 = load_diabetes()

# Create DataFrame from feature data
X = pd647.DataFrame(diabetes647.data, columns=diabetes647.feature_names)

# Target variable (disease progression)
y = pd647.Series(diabetes647.target, name='disease_progression')
```

- Do this Assignment step by step with relevant screen shots and brief explanation of important steps / code block(s) in your own words
- Paste all screen shots to a word document with explanations then convert the word document to PDF format
- Upload the complete PDF to Moodle before the deadline