**MT2002 Statistical Modeling**



|  |  |
| --- | --- |
| **Assignment No:** 02 | **CLO:** |
|  | **Semester:** Fall2023 |
| **Due date: 12**-Oct -2023 | **Marks:**  100 |

# Instructions

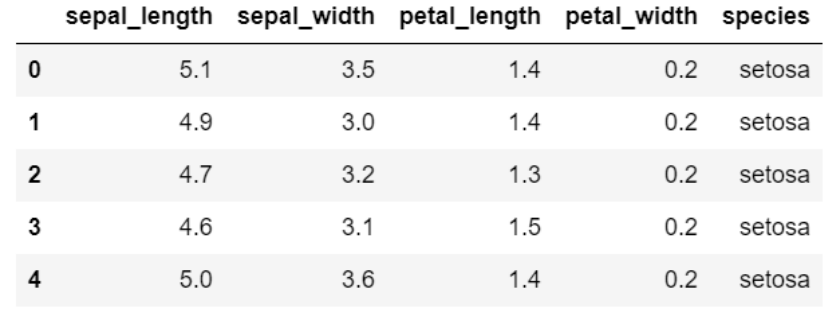
1. **Plagiarized work will result in zero marks.**
2. **No retake or late submission will be accepted.**
3. The submission should be a SINGLE UNZIPPED NOTEBOOK submitted on google classroom.
4. This notebook should properly document what you did? How you did it? And the source-code for each part as well as the generated outputs.
5. Your submission file should be according to the following **format: id\_section\_A2** e.g., i22123456\_A\_A2. (Note: A2 in the end denotes Assignment 2).

**Marks distribution for this assignment.**

|  |  |  |
| --- | --- | --- |
| Sepal\_length and **Petal\_width** | Pearson Correlation Coefficient? | 20 marks |
| Petal Length and **Petal\_width** | Pearson Correlation Coefficient? |
|  | |  |
| Model 1 | Model by selecting independent variable with high correlation from above | 40 marks |
| Model 2 | Model by selecting independent variable with low correlation from above |
|  | |  |
| Model 1 | R-squared value? | 20 marks |
| Model 2 | R-squared value? |
| Your conclusion based on above r-squared value and selection of which model is better and why? | ? | 20 |

# Problem Description

The Iris dataset is a well-known dataset in the field of machine learning and data analysis. It contains measurements of iris flowers, including sepal and petal lengths and widths, as well as the species of each iris. This dataset is often used for tasks such as classification, regression, and data exploration. The following is small portion of the dataset.



In this assignment, you will have the opportunity to explore the Iris dataset and apply linear regression analysis techniques using the PyMC library. Linear regression is a fundamental statistical method for modeling relationships between variables, and this assignment will help you gain hands-on experience in conducting such analyses. Before proceeding with the modeling, follow these steps:

Calculate the Pearson correlation coefficient for the following pairs of variables:

Sepal Length and **Petal Width**

Petal Length and **Petal Width**

# Linear Regression Model

After calculating the correlation coefficients, build two linear regression model using PyMC3. Here's what you need to do:

* Choose Petal Width as the dependent variable (Y) for your linear regression model.
* First, select one independent variable (X) from either Sepal Length or Petal Length based on the highest correlation coefficient observed in Task 1. Explain your choice clearly.
* Then, select another independent variable (X) from the remaining option (either Sepal Length or Petal Length) that results in a lower Pearson correlation coefficient.
* Construct two separate simple linear regression models using PyMC, one for each choice of independent variable.
* Calculate and report the R-squared values for both models to assess their quality of fit. Also explain which model results in better r-squared value.