# Executive Summary– Evaluating Regression and Classification Models on the IRIS data Set.

By Samuel Mbah Nde – DDS 8555 – Assignment 1

## Project Overview

This document provides a summary of the project to evaluate regression and classification models using the iris dataset (Fisher, 1936). It covers methodology, key findings and is intended to be used together with the python Jupiter Notebook (and its knitted pdf version) attached to this as part of the homework submission.

I loaded the data into a pandas dataframe and renamed the columns which contained spaces and special characters to camel cased words making them easy to access in code using pandas dot notation which reduces the amount of code I have to type and reduces the chance of types (A comprehensive review of tools for exploratory analysis of tabular industrial datasets, 2018).

Then following instructions provided for the project, I first computed a new column in the dataset named **New** using the provided formula. Next, I divided the dataset into training and test sets in the ratio of 80% to 20% (Ndung'u, 2022) before computing static values from the training set to which were used as predictors to evaluate their effectiveness of using such averages to predict the Sepal Length.

The first regression predictor which evaluated to a constant value of 4.64 (computed as the mean of (Sepal Length – Petal Width) proved to be more effective predictor of Sepal Length than the second predictor (a value of 3.77, computed as the average Petal Widths). I investigated why and determined that the former predictor was more efficient because values of Petal Widths are much smaller than values of Sepal Widths.

To evaluate classification models, I set up two classifiers following the distribution of values of by quartiles of Sepal Lengths as summarized in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **1st Quarter** | **2nd Quarter** | **3rd Quarter** | **4th Quarter** |
| **Model1** | 0 | 1 | 2 | 2 |
| **Model2** | 0 | 0 | 1 | 2 |

Table 1: Predictor Class for Model 1 vs Model 2.

Then I used classification report from sklearn to evaluate how these classifiers would perform on the test set and concluded that the second model was more efficiency based on the averages of classification metrics accuracy, precision, recall and F1 score (Nogales & Benalcázar, 2023).

More analysis is done in the attached notebook to explore reasons for each model being more efficient.

# Bibliography

A comprehensive review of tools for exploratory analysis of tabular industrial datasets. (2018). *Visual Informatics, 2*(4), 235-253. doi:10.1016/j.visinf.2018.12.004

Fisher, R. A. (1936). *IRIS*. doi:10.24432/C56C76

Ndung'u, R. (2022). Data Preparation For Machine Learning Modelling. *International Journal of Computer Applications Technology and Research, 11*, 231-235. doi:10.7753/IJCATR1106.1008}

Nogales, R. E., & Benalcázar, M. E. (2023). Analysis and Evaluation of Feature Selection and Feature Extraction Methods. *International Journal of Computational Intelligence Systems, 16*(1). doi:10.1007/s44196-023-00319-1