Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name: Charan C S **Email:** ccharancs543@gmail.com **Name:** Lokesh Tokas **Email:** lokesh.you@gmail.com **Name:**Saraswat Mukherjee **Email:**mae21saraswat@gmail.com **Name:** Shubham Sartape **Email:**shubhamns19.pumba@gmail.com **Contributor roles:** Charan C S - Project Summary Lokesh Tokas – Colab Notebook Saraswat Mukherjee - Technical Documentation Shubham Sartape - Presentation
Please paste the GitHub Repo link.

GithubLink:- https://github.com/Donein/Supervised-ML---Regression

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

We have presented our performance predictions on the basis of runtime by using different Machine Learning algorithms, exploratory data analysis, visualizations and lots of other interesting insights into the SSEMM GPU kernel.

The given dataset has total 241600 entries and 18 columns. The dataset measures the running time of matrix-matrix product (AB=C) which is in the form of 2048 x 2048 matrix. For each test combination 4 runs are performed which are present in the last 4 columns. There are 14 parameters in which first 10 are ordinal that take up to 4 different powers of two values and next 4 are binary.

The dataset does not contain null values. Initially in data cleaning we merged the last 4 columns by it's mean value into a column called 'runtime' and dropped those 4 columns.
In EDA, analysis is done on the dependent variable column 'runtime' if outliers are present by visualizing with boxplot and it is known that there are no extreme outliers. Then visualizing each column by plotting histogram plots to know it's frequency distributions. Transforming 'runtime' data by scaling it with log values for normalization, visualizing it and checking it with each variable column by plotting heat map to know the strength of correlation.

In machine Learning, the goal of supervised learning is to build a model that performs well and accurate predictions on new data. Hence the splitting the data into training set and testing test to avoid overfitting and to estimate the performance of the ML model on new data.

Calculating the R^2 and adjusted R^2 score to determine the strength of correlation between the predictions and the target values.
Interpreting the output coefficients by applying linear Regression algorithm between dependent and independent variable columns by calculating R^2 and adjusted R^2 score and visualizing using scatter plot. Then getting the same score for different ML algorithms like Lasso Regression, Ridge Regression and Decision Tree to get the predicted data and visualizing it by scatter plot for better understanding. Finally deciding which ML model to use based on R^2 and adjusted R^2 scores, then finding the feature importance and visualizing using bar graph.

Decision Tree Regression algorithm gives us the better estimation out of all the used algorithms with the largest adjusted R^2 score of 0.99. While all other algorithms has R^2 score of 0.55.