BRAC UNIVERSITY

Inspiring Excellence

CSE422

LAB PROJECT REPORT

Calories Burnt Prediction Using Machine Learning

Submitted By

Group: 05

Husne Mubarak, 20101336

Fahim Ahsan, 20101545

Ha-mim Ahmad, 20101286

Fahim Irfan Ahmed, 20101508

INTRODUCTION

People typically only consider eating or weight loss when they think of calories. But a calorie is frequently a measurement of heat energy. The amount of energy required to raise 1 gram(g) of water by 1°C is measured in calories. A range of energy-releasing systems unrelated to the human body may be evaluated using the technique. From the viewpoint of the human body, calories are defined as the quantity of energy required by the body to complete a task. Food has calories. Since different foods have different calorie counts, each and every item has a unique amount of energy packed in it. When we exercise or do a rigorous workout, our body temperature and heart rate will start to rise. The breakdown of the carbohydrates yields glucose, which is then processed by O2 into energy (oxygen). The training period, average heart rate, and temperature are the variables taken into consideration here. Then, calculate the person's height, weight, gender, and age to estimate how much energy they expend each day. Exercise duration, average heart rate, temperature, height, weight, and gender are all variables that can be taken into account. Based on factors including length, heart rate, body temperature, height, weight, and age, some machine learning regression algorithms are utilized to forecast how many calories will be expended. The purpose of the study is to provide a method for estimating the number of calories burnt using machine learning algorithms. Linear regression, ridge regression, lasso regression, decision tree regressor and random forest regressor are the learning techniques taken into account. This study's objective is to assess which algorithms are the most accurate in estimating how many calories a person would burn given their data. The study's model may be incorporated into or utilized in conjunction with already available technologies to provide a more accurate estimate of the number of calories burnt by people during various types of physical activity.

METHODOLOGY

I. Dataset Description

People typically only consider eating or weight loss when they think of calories. But a calorie is frequently a measurement of heat energy. The amount of energy required to raise 1 gram (g) of water by 1°C is measured in calories. A range of energy-releasing systems unrelated to the human body may be evaluated using the technique. From the viewpoint of the human body, calories are defined as the quantity of energy required by the body to complete a task. Food has calories. Since different foods have different calorie counts, each and every item has a unique amount of energy packed in it.

When we exercise or do a rigorous workout, our body temperature and heart rate will start to rise. The breakdown of the carbohydrates yields glucose, which is then processed by O2 into energy (oxygen). The training period, average heart rate, and temperature are the variables taken into consideration here. Then, calculate the person's height, weight, gender, and age to estimate how much energy they expend each day. Exercise duration, average heart rate, temperature, height, weight, and gender are all variables that can be taken into account. Based on factors including length, heart rate, body temperature, height, weight, and age, some machine learning regressor algorithms are utilized to forecast how many calories will be expended.

II. Pre-Processing Techniques

Handling Missing Values:

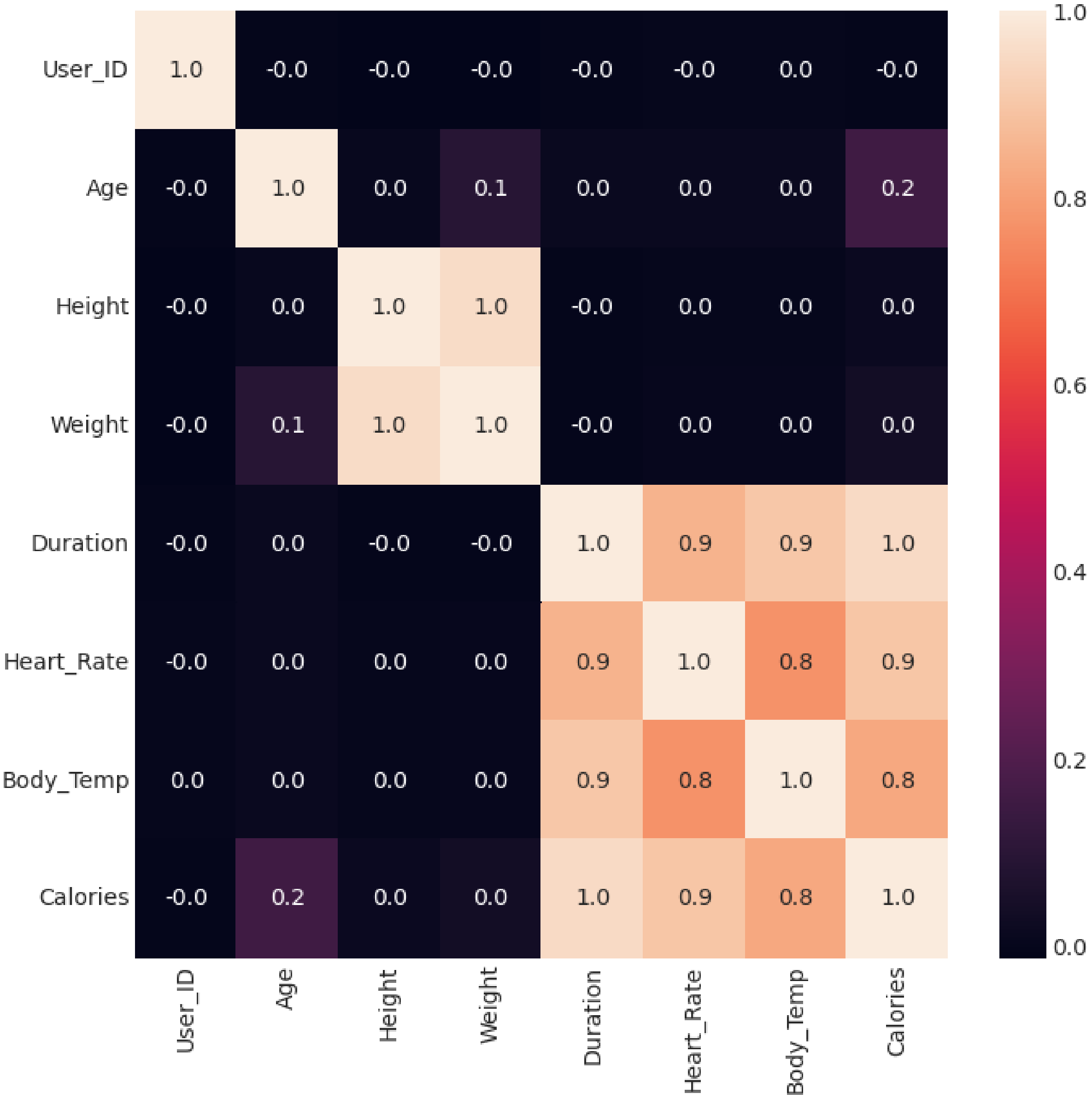
Data cleaning is the process of identifying and correcting inaccurate or incomplete data that are included in the dataset. One of these necessary steps is dealing with the data that are absent from the dataset. As many datasets in everyday life may contain a significant number of missing values, dealing with them is an important task.

If some of the values from the dataset is missing or null or NaN, then operating with the datasets will result in error. For this reason, we have to either remove the values or randomize the values or fill the values using mean, median etc strategies. Fortunately, in our dataset there were no missing or null values so we did not require to use any strategies stated above.

Feature Selection:

Machine learning models only require a narrow percentage of the dataset's variables; the others are either redundant or pointless. If the dataset contains all these unnecessary and redundant information, the model's overall performance and accuracy may deteriorate. Machine learning's feature selection technique makes it possible to find and choose the most relevant features from the data, which is vital to remove redundant or less important information. For pairs characteristics in this study, the correlation coefficient is determined in order to better comprehend the features and investigate the link between the features. The correlation coefficient's range is from **-1** to **1**, with **-1** denoting strong negative (“***y***” declines for “***x***” decrement), **1** denoting strong positive (“***y***” increases for “***x***” increment), and **0** denoting no correlation at all between the features.

If two features are correlated, we can predict one from the other. From our correlation matrix, we can see that "***User_ID***" has no correlation with any of the other features. So, we dropped the "***User_ID***" from our feature list. "***Height***" and "***Weight***" have mutual correlation and any of them could have been dropped. But we did not drop any of the remaining features.



	User_ID	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
User_ID	1.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0
Age	-0.0	1.0	0.0	0.1	0.0	0.0	0.0	0.2
Height	-0.0	0.0	1.0	1.0	-0.0	0.0	0.0	0.0
Weight	-0.0	0.1	1.0	1.0	-0.0	0.0	0.0	0.0
Duration	-0.0	0.0	-0.0	-0.0	1.0	0.9	0.9	1.0
Heart_Rate	-0.0	0.0	0.0	0.0	0.9	1.0	0.8	0.9
Body_Temp	0.0	0.0	0.0	0.0	0.9	0.8	1.0	0.8
Calories	-0.0	0.2	0.0	0.0	1.0	0.9	0.8	1.0

Encoding Categorical Features:

Datasets for machine learning occasionally include words to make the data comprehensible or in human-readable form. Choosing how to include this data into the study is difficult. Without further modification, many machine learning algorithms can accommodate categorical values, but there are many more algorithms that cannot. Therefore, this data must be transformed into numeric form in order for machine learning algorithms to decide how to operate those labels more effectively. The categorical "**Gender**" feature in the dataset we utilized indicates whether the population of the dataset is "***Male***" or "***Female***". We changed the genders of "***Male***" and "***Female***" using Label Encoder to "**1**" and "**0**" respectively.

III. Applied Models

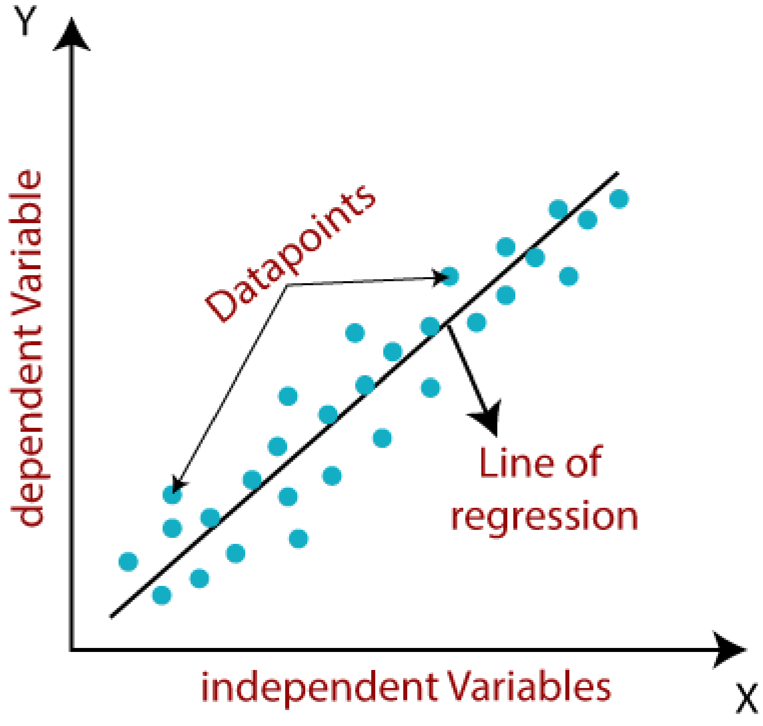
Linear Regression:

In order to model the relationship between two variables, linear regression fits a linear equation to the observed data. The first variable is regarded as an independent variable, whereas the second is regarded as a dependent variable. The strength of the association between two variables may be assessed using a scatter plot. A meaningful model will probably not be produced by fitting a linear regression model to the data if it appears that there is no correlation between the suggested explanatory and dependent variables.

The correlation coefficient, which has a value between **-1** and **1**, is a useful numerical indicator of the strength of the link between two variables in the observed data.

$\mathbf{Y = b_0 + b_1X}$, where **X** is the independent variable and **Y** is the dependent variable, is the equation of a linear regression line. $\mathbf{b_0}$ is the intercept (the value of **Y** when **X = 0**), and $\mathbf{b_1}$ is the line's slope.

$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ldots . + b_nx_n$



Ridge Regression:

In situations when the independent variables are strongly correlated, ridge regression is a technique for estimating the coefficients of multiple-regression models. It applies the L2 regularization method. Ridge regression updates the feature weights due to the added squared term in the loss function. Ridge regression minimizes overfitting and decreases the size of the weight values as a whole during optimization.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

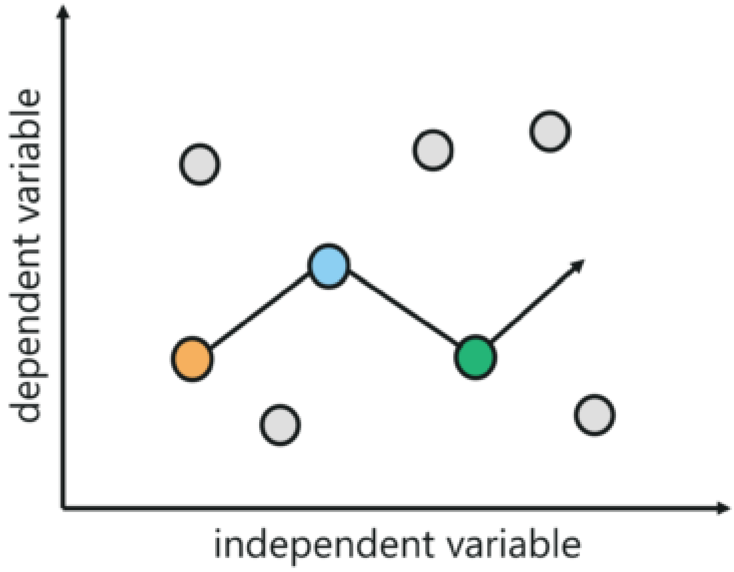
The penalty term is lambda. The ridge function's alpha parameter serves as a placeholder for the provided here. So, we may regulate the penalty term by varying the values of alpha. The penalty is greater with larger alpha values, which reduces the magnitude of coefficients.

Lasso Regression:

A lasso regression is a type of regularization method. For a more accurate forecast, it is preferred over regression techniques. Shrinkage is used in this model. When data values shrink toward the mean, this is referred to as shrinkage. Lasso approach is mainly implemented in simple models. When a model exhibits a high degree of multicollinearity or when you wish to automate some steps in the model selection process, such as variable selection and parameter removal, this specific sort of regression is ideally suited.

L1 regularization is employed by Lasso Regression. Because it does feature selection automatically, it is employed when there are more features.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

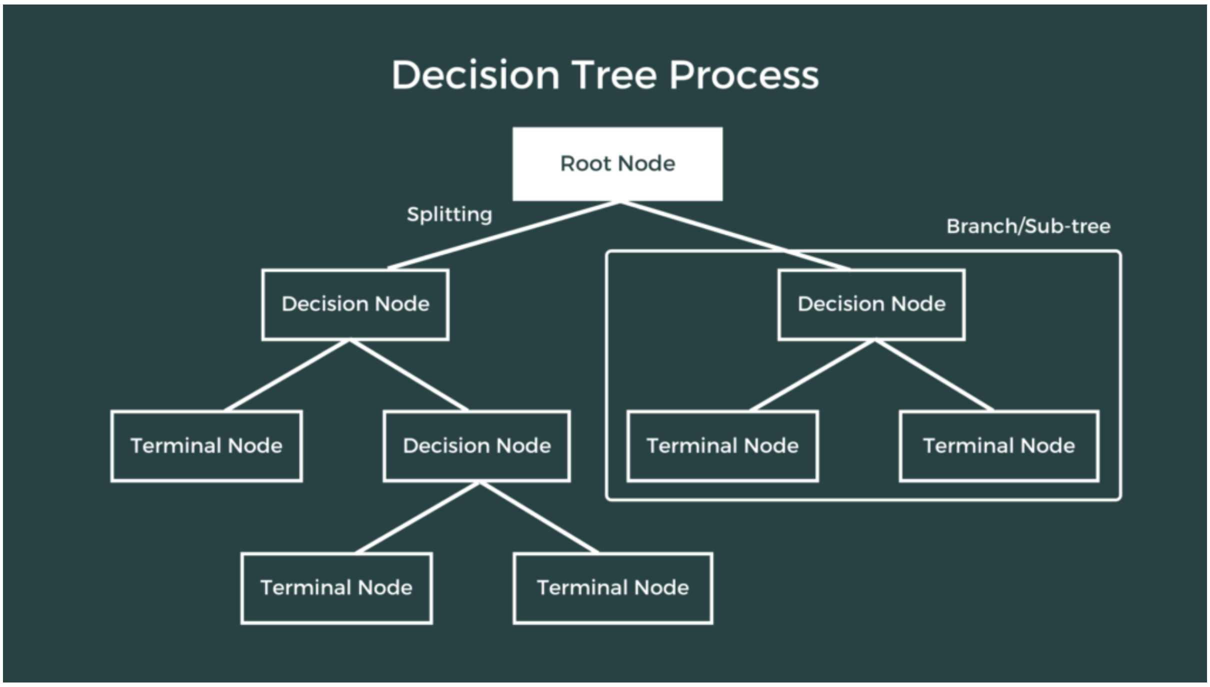


Decision Tree Regressor:

One of the most popular Machine Learning techniques for handling classification and regression issues is the decision tree. The technique, as its name implies, employs a tree-like model of choices to either forecast the target value or the target class.

The splitting process begins at the root node and proceeds through a branching tree, ending at a leaf node or terminal node, which carries the prediction or algorithmic result. The process of building a decision tree typically proceeds top-down, with each step selecting the variable that best divides the set of objects. A binary tree may be used to represent each subtree of the decision tree model, where a decision node divides into two nodes depending on the circumstances.

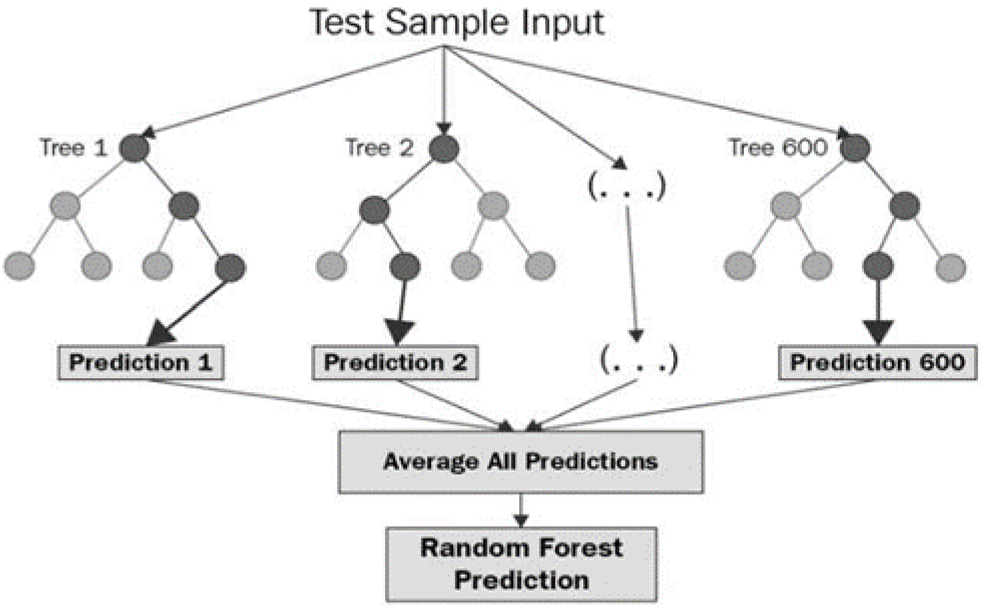
Regression trees are decision trees where the target variable or the terminal node can accept continuous values.



Random Forest Regressor:

One of the supervised machine learning approaches called random forest regression is frequently applied to classification and regression issues. On various samples gathered from datasets, decision trees are constructed, and the majority vote for categorization and average are selected. The random forest algorithm's capacity to analyze datasets with continuous variables, like those used in regression, is one of its most crucial features. It's possible that some decision trees anticipate the right outcome while others don't since a random forest mixes numerous trees to forecast classes in a dataset. But when all the trees are combined, they anticipate the right result.

Even for bigger datasets, which occasionally lose significant sections of data, the random forest method predicts with greater accuracy and requires less training time than other algorithms. As the name implies, a random forest is made up of several distinct decision trees. Every single tree in the forest makes a prediction or produces a result, and the outcome that receives the majority of votes is the model's ultimate finding. In essence, the random forest method provides the machine the average best-voted solution after gathering the knowledge of the population.

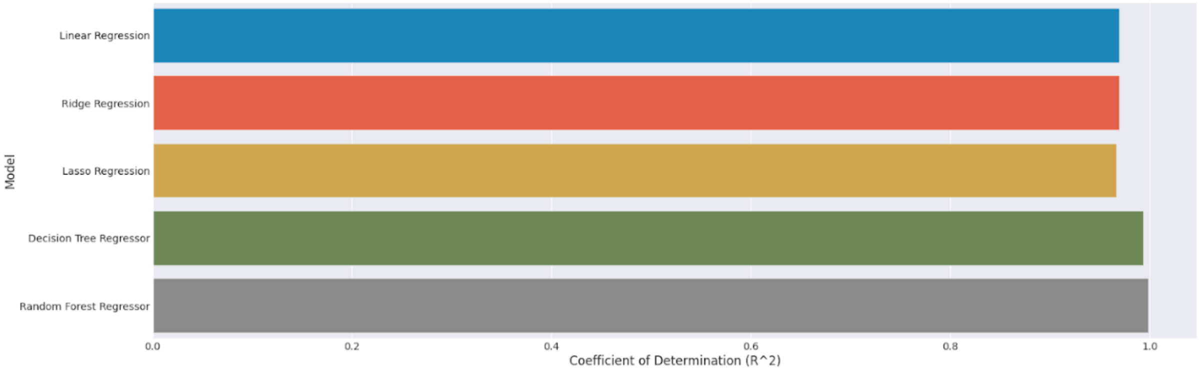


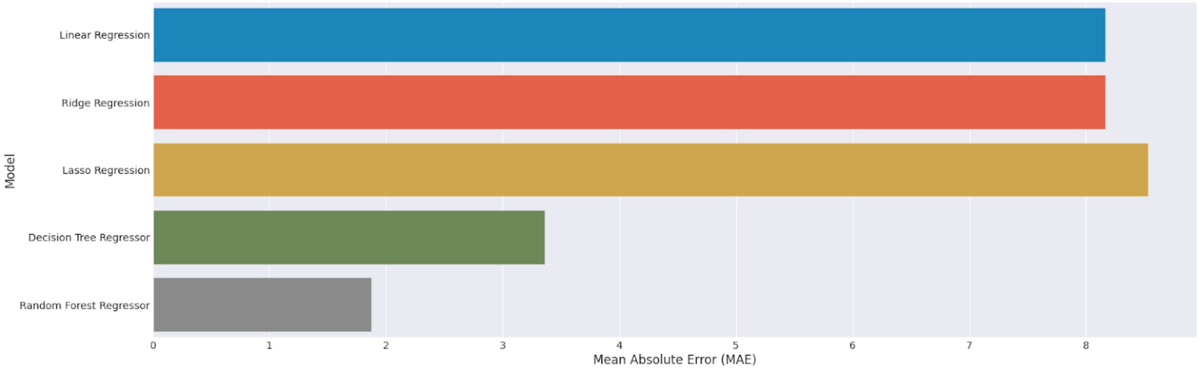
RESULTS

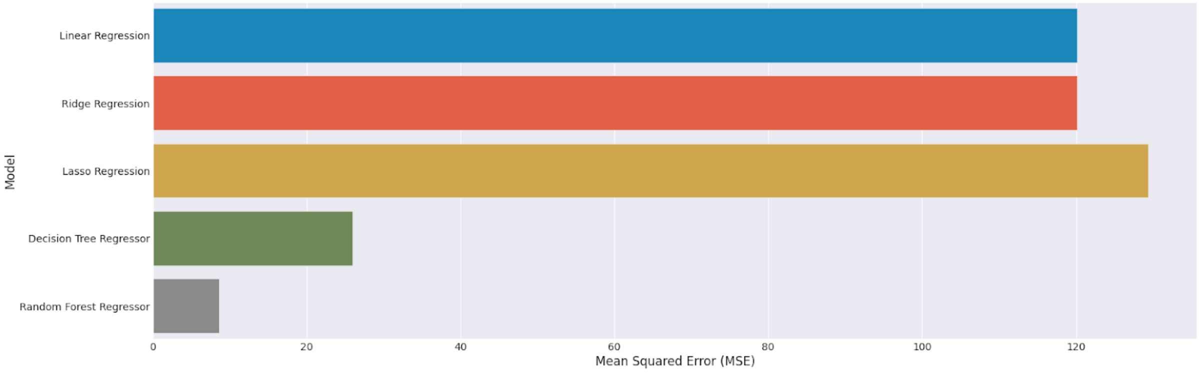
After Applying Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor and Random Forest Regressor, we have achieved predicted results of the training dataset.

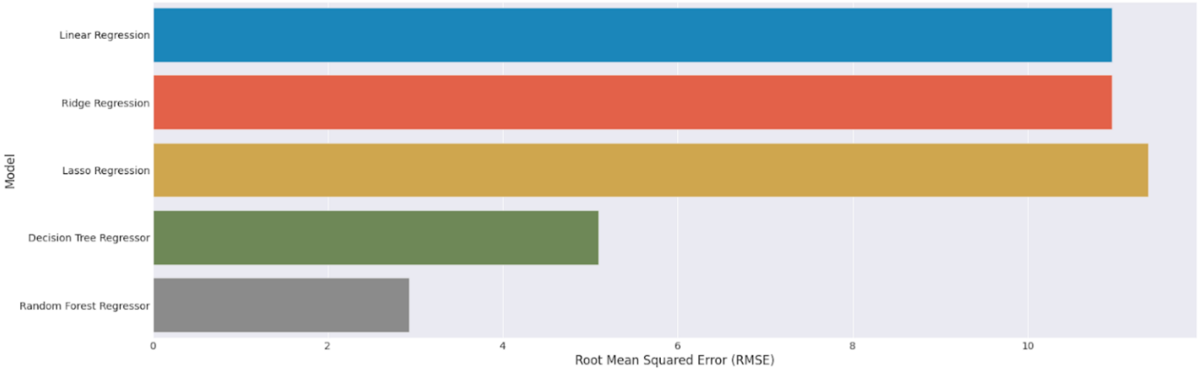
Model	R^2	MAE	MSE	RMSE
Linear Regression	0.969	8.165	120.163	10.962
Ridge Regression	0.969	8.165	120.162	10.962
Lasso Regression	0.966	8.529	129.319	11.372
Decision Tree Regressor	0.993	3.358	25.926	5.092
Random Forest Regressor	0.998	1.872	8.606	2.934

Bar Graphs:

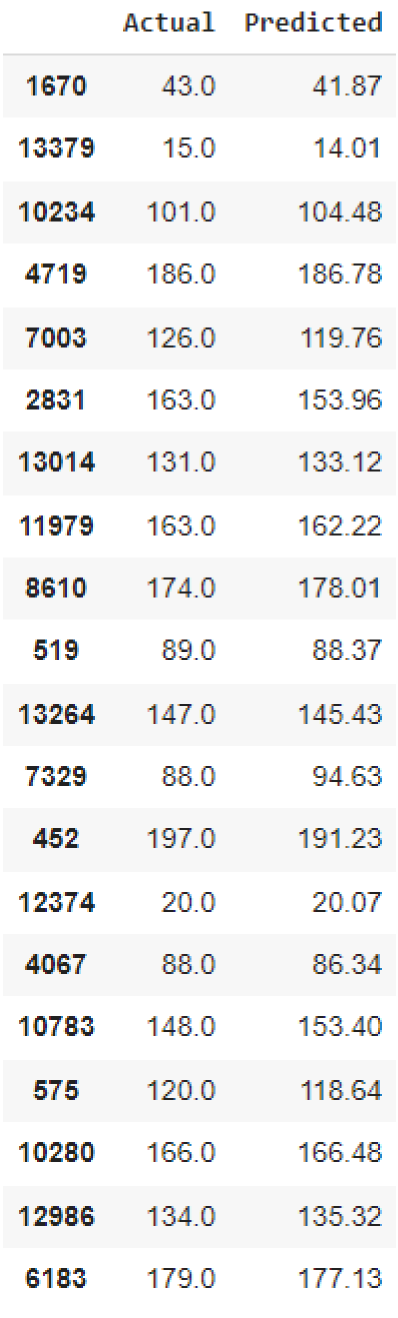

Model
Linear Regression
Ridge Regression
Lasso Regression
Decision Tree Regressor
Random Forest Regressor
0.0
0.2
0.4
0.6
0.8
1.0
Coefficient of Determination (R^2)


Model
Linear Regression
Ridge Regression
Lasso Regression
Decision Tree Regressor
Random Forest Regressor
0
1
2
3
4
5
6
7
8
Mean Absolute Error (MAE)


Model
Linear Regression
Ridge Regression
Lasso Regression
Decision Tree Regressor
Random Forest Regressor
0
20
40
60
80
100
120
Mean Squared Error (MSE)


Model
Linear Regression
Ridge Regression
Lasso Regression
Decision Tree Regressor
Random Forest Regressor
0
2
4
6
8
10
Root Mean Squared Error (RMSE)

From the above result table, we can observe that Random Forest Regressor has the most accurate predicted scores among the five ML models. Now, we will look at the predicted values generated by Random Forest Regressor vs the actual values in the dataset:



	Actual	Predicted
1670	43.0	41.87
13379	15.0	14.01
10234	101.0	104.48
4719	186.0	186.78
7003	126.0	119.76
2831	163.0	153.96
13014	131.0	133.12
11979	163.0	162.22
8610	174.0	178.01
519	89.0	88.37
13264	147.0	145.43
7329	88.0	94.63
452	197.0	191.23
12374	20.0	20.07
4067	88.0	86.34
10783	148.0	153.40
575	120.0	118.64
10280	166.0	166.48
12986	134.0	135.32
6183	179.0	177.13

REFERENCES

1. https://www.kaggle.com/datasets/fmendes/fmendesdat263xdemos
2. https://www.javatpoint.com/linear-regression-in-machine-learning
3. https://dataaspirant.com/ridge-regression/#t-1605117100775
4. https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/
5. https://www.theclickreader.com/decision-tree-regression/
6. https://www.geeksforgeeks.org/random-forest-regression-in-python/
7. https://www.javatpoint.com/data-preprocessing-machine-learning