1. Introduction

This is a report on a machine learning application which predicts whether a customer will suffer by the increase in energy price or not. To address this problem, the research and finding are divided into two sections.

1. Comparative Study of various classification techniques to classify whether a customer will suffer or not by the increase in energy price.
2. Comparative Study of various regression techniques to predict the value of the variation in annual expenditure for each customer.

The author looks for the best classification techniques to move forward because this problem fits into a supervised learning strategy. There are numerous algorithms that can be used, with decision tree classifier, support vector machine classifier, K closest neighbour classifier etc

2. Exploratory Data Analysis

Initial data exploration is carried out by detailing the data. There are some graphs and other statistical methods used to understand the data in a better way.

F3	F4	F5	F6	F7	F8	F9	F10	...	F13	F14	F15	F16	F17	F18	F19	F20	F21	Class
.051752	-6.99630	1473.45	0	4.40510	16004.16	1	-11645.820	...	-13.7350	5.9854	14.24730	1.4892	9959.04	-3199.350	6.8670	-4850.820	NaN	False
.153000	-4.27404	783.15	0	4.54500	16041.48	1	-9759.420	...	-5.1710	4.6222	14.99820	3.1206	10107.44	-3064.950	9.4710	378.780	NaN	True
.440000	-8.25900	2573.95	0	3.59962	16422.78	0	-10775.220	...	-38.6500	4.4096	17.79000	1.9818	10971.04	-3638.850	2.0373	-5215.480	NaN	False
.386600	-5.57460	1499.68	0	3.56675	16270.04	1	-9416.714	...	-8.6390	8.5080	12.98424	11.5620	9681.10	724.950	2.0220	-4378.420	NaN	False
.650200	-4.33848	1409.15	1	5.07000	16548.78	1	-9797.820	...	-14.7300	5.5188	13.57260	2.5440	10965.64	-2607.150	4.3140	-1919.220	NaN	True

Figure1: Data Description

From the details, it is clearly visible that the feature ‘Class’ have a Boolean data type which is our predictor variable which holds only Boolean values and the remaining all other variables and independent variables.

2.1 Handling Missing Values

The next step is to understand the data in detail by checking the missing values. Figure 3 gives a perfect picture of missing values of the classification problem.

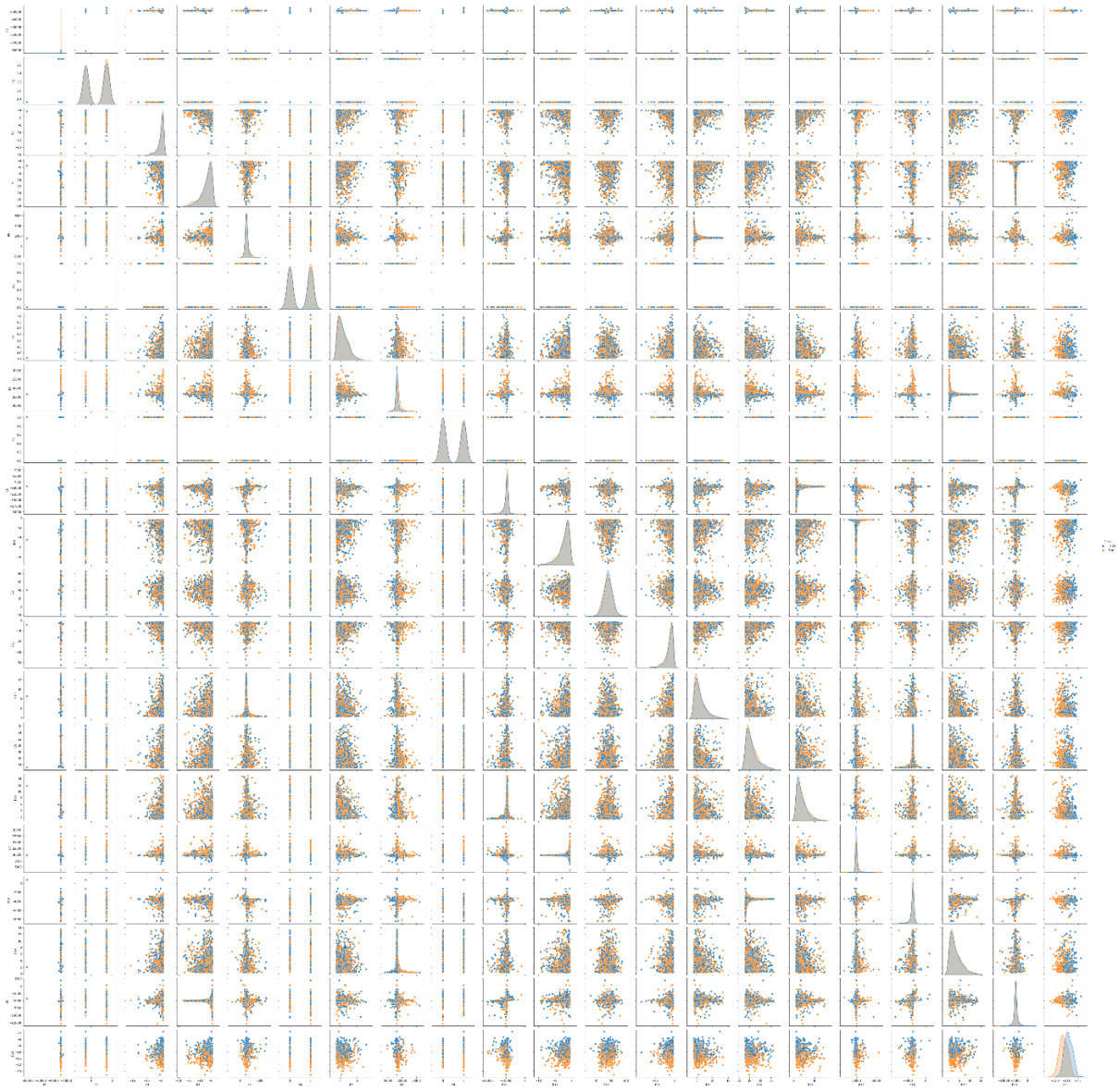

Figure 2: Pair plot

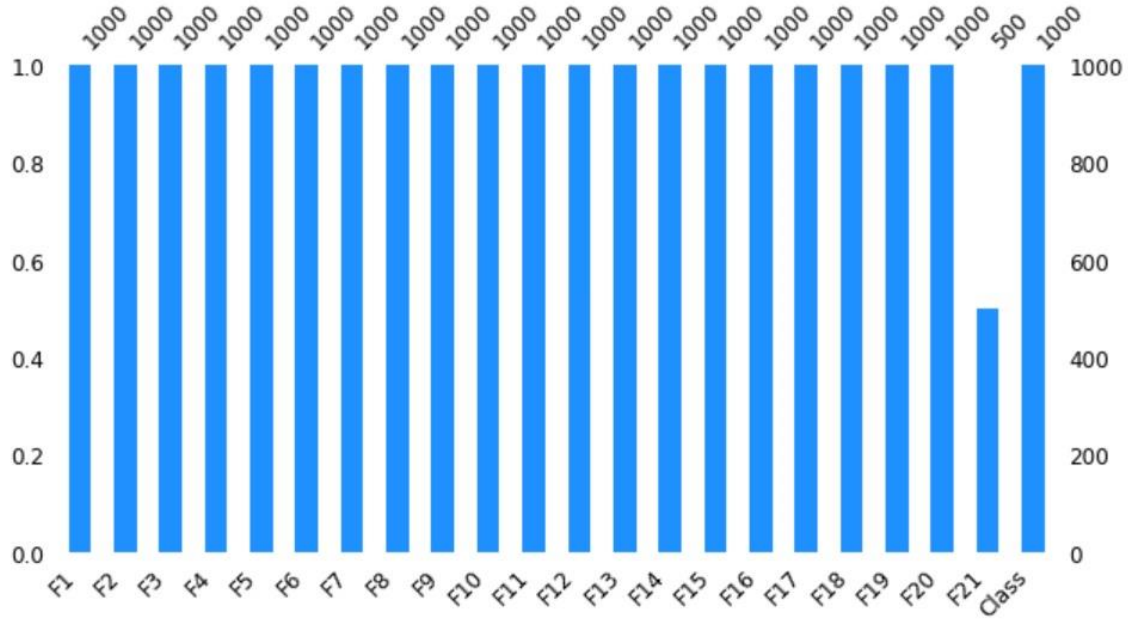

Figure 3: Missing Value (Data_P2)

The dataset only has one column, "F21," which has 500 missing values which can be dropped or filled using any one imputation method.

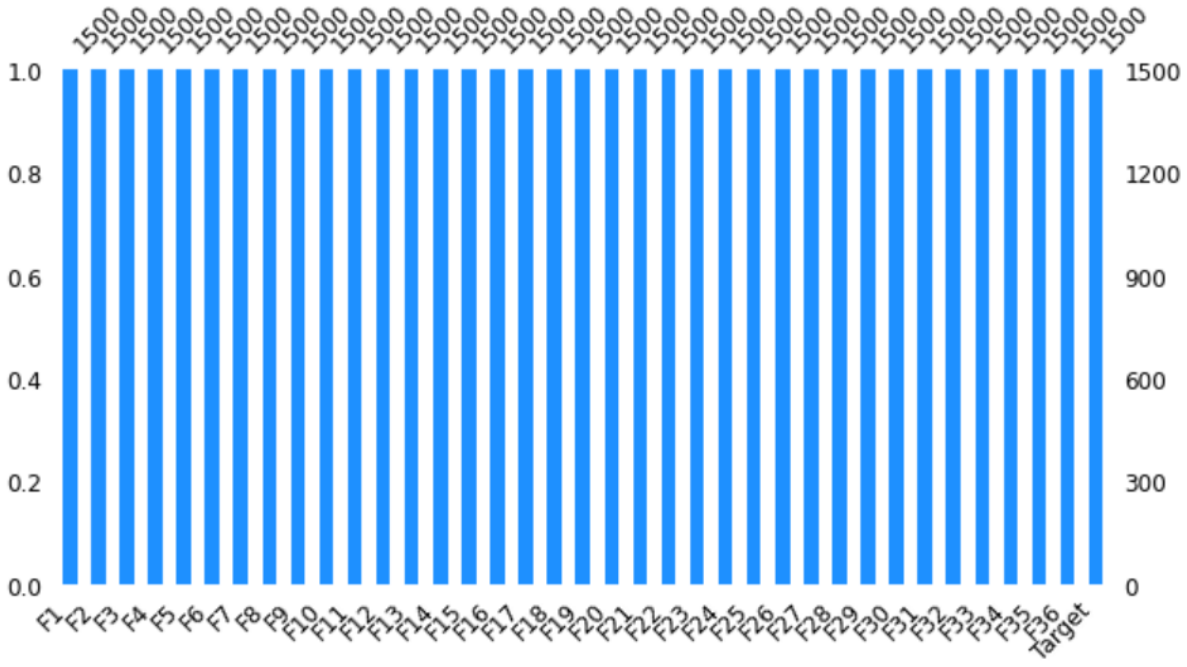

Figure 4: Missing Value (Data_P3)

It is always better to apply any imputation method than dropping the column as it has 50 % of its data. Figure 4 shows the missing values of data in the prediction problem. Since only one column

has a missing value and mean value can provide a superior level of accuracy, this method is the first one that is advised.

2.2 Handling of Categorical Value

When looking at figure 2, the feature "Class" has True or False as categorical [3] values. It must be modified to 0 and 1, as the target variable must either be 0 or 1 to indicate whether the customer will suffer when the price of energy rises.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	...	F28	F29	F30	F31	F32
0	-69.21	16536.84	6	Europe	-16720.42	735.74	22.14	15.48	Medium	124.05	...	-506.31	-3077.82	15.35	-53.04	-112.43
1	-168.67	28434.21	12	UK	-8070.93	91.35	-1.86	18.60	Very high	57.78	...	-731.28	-4559.70	20.74	-36.44	-44.06
2	-76.19	22895.97	18	Europe	-12126.02	145.64	-68.28	14.22	High	-65.88	...	-355.08	-3965.13	16.65	-76.06	-81.66
3	-103.19	22926.51	12	Europe	-10050.95	218.39	-40.58	14.99	Very high	132.12	...	-518.52	-1509.87	21.99	-128.34	-142.53

Figure 5: Describing Categorical Value

Interpreting Figure 5 also gives us a clear picture that 2 feature have categorical value which need to be replaced to numeric value. There are different methods to handle categorical data. One of the simplest yet efficient method called label encoding is suggested by the author.

2.3 Feature Scaling

Feature scaling [6] is a method to normalise independent variables to a uniform pattern. This involves rescaling of features to have their properties as standard distribution with mean zero and variance one.

3. Investigating the Performance of Algorithms

3.1 Classification Methods

Classifying each customer whether he will suffer by the increase of energy bills is being implemented by the following three classification algorithms.

1. Decision Tree (DT) Classification
2. K Nearest Neighbour (KNN) Classification
3. Support Vector Machine (SVM) Classification

The likelihood of obtaining the greatest result for any classification challenge is determined by the method of choice. Out layers are discovered but not frequent. As a result, the author selects SVM [1] with RBF kernel to evaluate the highly overlapping data The author's next choice was to use a decision tree. DT [2] is one of the finest models to predict the classes without many complexities. Additionally, this permits the classification of subsets and the creation of rules. In addition, it requires very little accurate data preparation and pre-processing. KNN is the

third one considered. KNN employs distance-based categorization to minimise the impact of these out layers even when the data is overlapped with them.

Test Set Accuracy						
Approach	Decision Tree		KNN		SVM	
Imputation Method	Mean Value Based	KNN Based	Mean Value Based	KNN Based	Mean Value Based	KNN Based
Accuracy	80.30%	65.15%	62.73%	65.15%	73.03%	65.15%

Figure 6: Test Set Accuracy

Any model can be evaluated using several methods. The author compares the performance by calculating the ratio of correct predictions to total predictions represents, a model's overall accuracy. An accuracy score ranges from 0 to 1, with 1 representing the ideal model. Through comparison, decision trees with mean value imputation are the most accurate way to predict categorization, with accuracy scores of 80.30% for test sets.

Training Set Accuracy						
Approach	Decision Tree		KNN		SVM	
Imputation Method	Mean Value Based	KNN Based	Mean Value Based	KNN Based	Mean Value Based	KNN Based
Accuracy	85.07%	79.10%	78.06%	79.10%	72.99%	79.10%

Figure 7: Training Set Accuracy

Let's now investigate other approaches to thoroughly assess the model to understand the statistics. Another way is to print the confusion matrix of the same to know about the truth values and this can identify where it gone wrong.

147 (TP)	**18 (FN)**
47 (FP)	**118 (TN)**

Figure 8: Confusion Matrix

The model predicts true out of positive class 147 times and 118 times false where the actual was false. Through classification report, a detailed report con precision, recall can be analysed. The decision tree model hit 76 % of precision, which means it retrieved that many percentages of relevant document out of 100. At the same time, 89% of relevant instance is being created.

	precision	recall	f1-score	support
0	0.76	0.89	0.82	165
1	0.87	0.72	0.78	165
accuracy			0.80	330
macro avg	0.81	0.80	0.80	330
weighted avg	0.81	0.80	0.80	330

Figure 9: Classification Report

3.2 Classifying the Test Set

To predict the class in the test set, the author repeated all the steps loading the data, cleaning the data including categorical value conversion and missing value imputation. Then applied the same process followed as in the training data. As the last step applied the test data in the trained model and written [9] back all the new prediction class into the CE802_P2_Test_Predictions.csv. The author rechecks and compares with the data given to ensure that none of the columns except the target class changed.

3.3 Regression Algorithms

Regression algorithms forecast the values of the dependent variable by using input characteristics from the data that is provided to the system. This method is used to build a model based on the traits of training data, and that model is then used to predict the value of new data.

There are many regressive algorithms to carry out the experiment. By analysing the size and nature of the data it is decided to move forward with the following three algorithms.

1. Linear Regression
2. Random Forest Regression
3. Ridge Regression

Multiple linear regression model[2] is used when we must find a strong relationship between dependent and independent variables. It is also one of the strongest techniques to predict the variable.

Random Forest regression model is opted when it is necessary to predict a continuous value. The algorithms work in 2 steps begins with constructing ‘n’ decision trees and secondly finds the average prediction across the estimators.

The third method, Ridge Regression the chance of overfitting is much less than that of linear regression and it reduces the complexity of the model. After pre-processing the data, it is ready to build the model using each of the prediction algorithms. The results obtained is shown below.

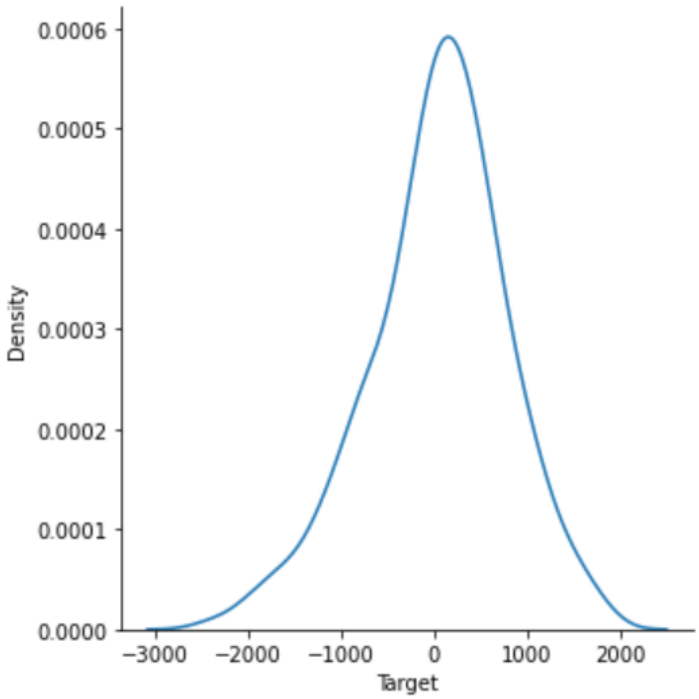

Figure 10: Actual vs Predicted Value

Figure 10 depicts the difference in actual and predicted value, which lie in between -3000 to 2000 and more. This gives a correct meaning that the predictor is worst with the given dataset.

Performance Matrix						
Approach	Linear Regression		Random Forest		Ridge Regession	
Criteria	R2 Value	Mean Square Error	R2 Value	Mean Square Error	R2 Value	Mean Square Error
Value	61.18%	559715.62	28.71%	1027826.73	61.18%	559631.05

Figure 11: Performance Matrix

From the above table Ridge Regression with the lowest mean square error 559631.05 and 61.18 % with R2 value choose to be the best model.

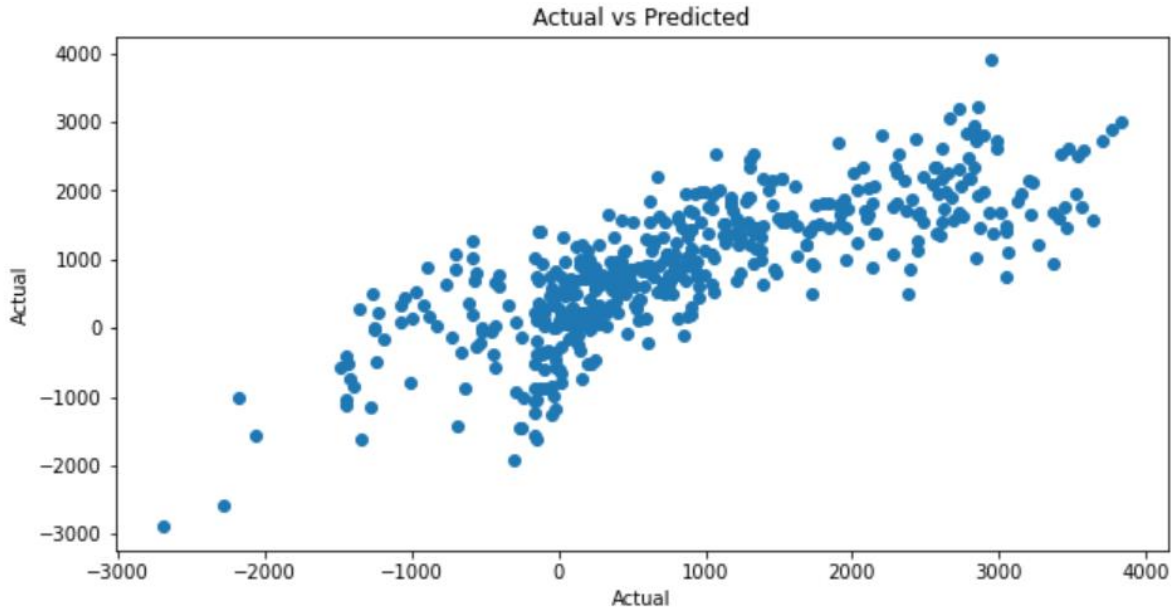

Figure 12: Test Data Vs Predicted Data

Figure 12[4] gives an overall idea about the test data and the predicted data. The graph says its little crowdy in the mid portion which is a good sign that the predicted values come in a line. Figure 12 [7]and 13 picturises data distribution among various algorithms.

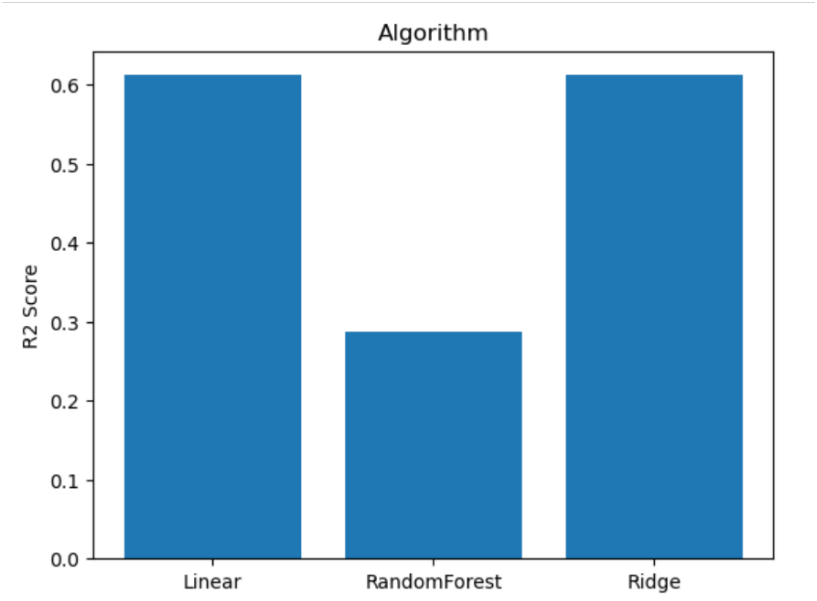

Figure 13: R^2 Value Vs Regression Algorithm

After comparing the results obtained through all the three models, Ridge [5] obtained the highest r2 value with lowest error, the algorithms work better that other algorithm.

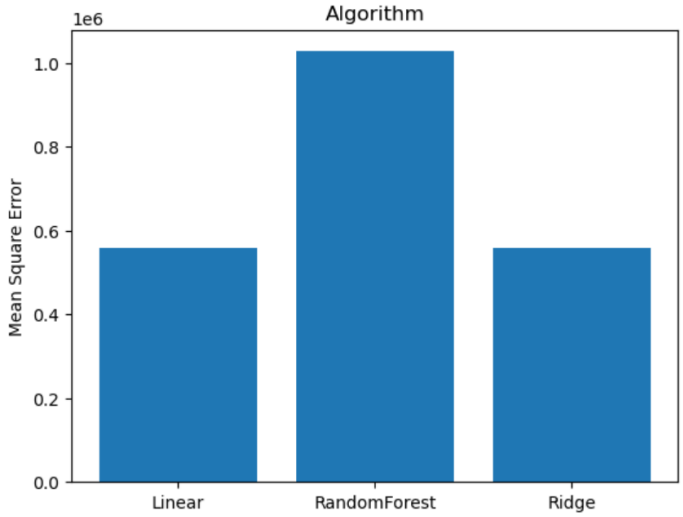

Figure 14: Mean Square Error Vs Regression Algorithm

3.4 Prediction on the Test Set: Regression

To predict the value for the test set, the author from loading date to scaling. As the next step, test data is applied the in the trained model and written [6] back the predicted values into the CE802_P3_Test_Predictions.csv.

4. Summary

Classification and regression algorithms are implemented and tested in the given dataset. For the classification problem, Decision tree classifier tops among others and Ridge regression predicted the best accurate values. There are many further improvisations can be implemented, which is not covered in this experiment.

References

[1] https://www.linkedin.com/learning/applied-machine-learning-algorithms/what-are-the-key-hyperparameters-to-consider-boosting?autoSkip=true&autoplay=true&resume=false&u=51088249

[2] https://www.linkedin.com/learning/machine-learning-with-python-k-means-clustering/choosing-the-right-number-of-clusters?autoSkip=true&autoplay=true&resume=false&u=51088249

[3] https://www.linkedin.com/learning/applied-machine-learning-foundations/why-do-we-need-to-explore-and-clean-our-data?autoplay=true&resume=false&u=51088249

[4] https://scikit-learn.org

[5] https://machinelearningmastery.com/feature-selection-machine-learning-python/

[6] Lecture notes on machine learning and Lab notes on scikit-learn, pandas.

[7] https://medium.com/codex/how-to-identify-and-visualize-missing-values-with-python-3b304b06fe21#0af8