Introduction

This report focuses on the comparative analysis of two issues: the design of a system to forecast the variation in a customer's annual expenditure due to an increase in energy costs, and a system to forecast whether a customer will have trouble paying the increasing cost of electricity based on a few characteristics. A machine learning technique based on regression is used to predict annual expenditures. To determine if a consumer will have trouble paying the bill, classification models were utilized.

1.Target Class Prediction Using Classification Methods

1.1Data Collection

A data file is available to examine the effectiveness of various machine learning techniques. It is referred to as historical data, and the AENERGY manager provides it.

1.2.Data Exploration

Data should be described and visualized to be understood. At this point, we can investigate the data set's data types, missing values, and outliers. Techniques for data visualization are employed in this procedure to find the problems[2].

- **Data types of the features** in the given data set.

```
energyprice_sample.dtypes
```

```
F1        float64
F2        float64
F3        float64
F4        float64
F5        float64
F6        float64
F7        float64
F8        float64
F9        float64
F10         int64
F11       float64
F12       float64
F13       float64
F14       float64
F15       float64
F16         int64
F17       float64
F18       float64
F19       float64
F20       float64
F21       float64
Class        bool
dtype: object
```

Figure 1: Data Description- Data Types

- The distribution of the target classes in the data set prior to the application of modelling techniques is shown in the accompanying bar chart[4]. The data shows that 46% of consumers were unaffected by the increase in energy prices, while 53% of customers were.

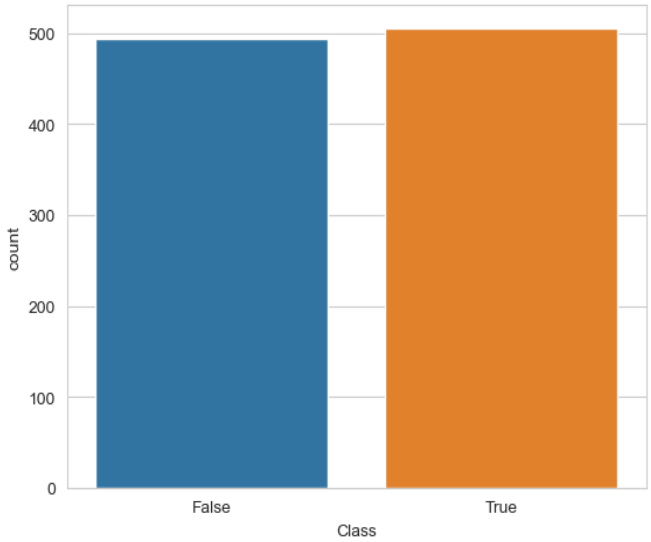

Figure 2: Distribution of Target Class

Before using any method based on the nature of the dispersed data, the pair plot provides information about the sort of algorithm that can solve this classification challenge. KNN and decision tree algorithms work best for the graph shown above because it has a lot of overlapped data. [1]We can use SVM and random forest as well.

Figure 3: Pair Plot of the given data set

1.3.Data Preparation

Data Normalisation

Normalize the data to bring all the variables into the same range because many algorithms are sensitive to the scale of the data. This enhances the model's accuracy as well. The following scaling method was incorporated into the code:

Standard Scaler: The data are scaled to have a distribution centred at a mean of 0 and a standard deviation of 1[2].

Dealing with Missing Values

With the aid of msno.bar, the missing values in the dataset are explored using the Missingno package.

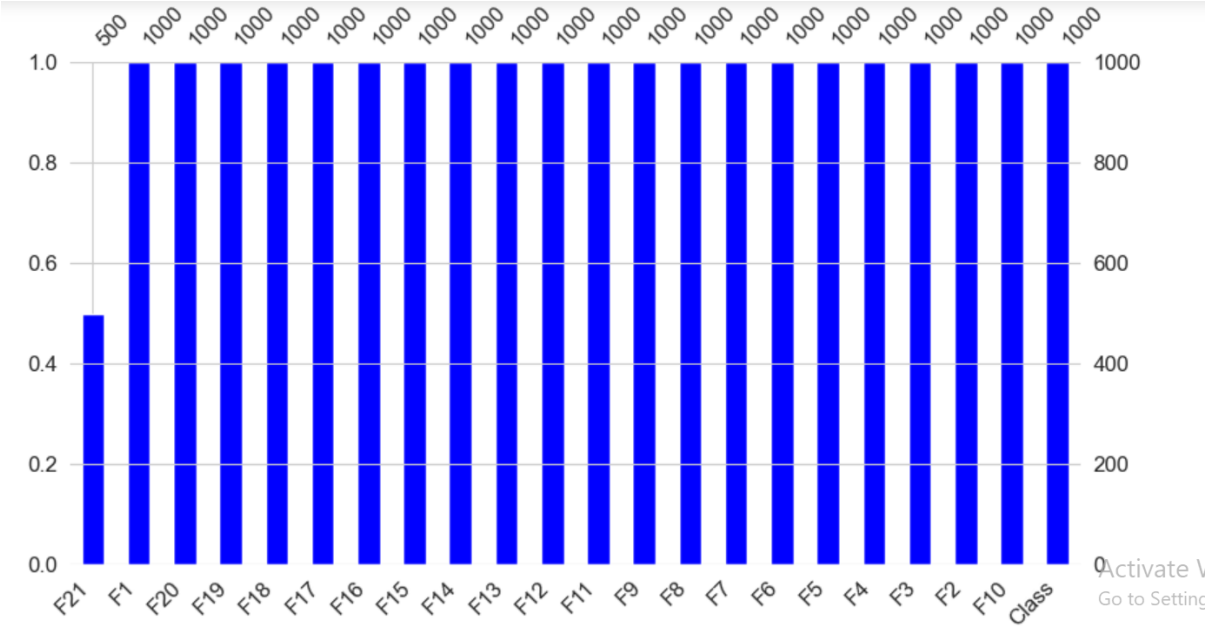

Figure 4: Missing Value Data Plot

In the graph above, it is evident that the F21 column has missing values in the upper half of the values, which can be fixed in one of two ways: either by deleting the column or by imputing the values using various methods. F21 column should not be deleted because it contains 50% of the data is filled. KNN based and Mean value imputations are used to deal with missing values.

1.4.Modeling with Classification Methods

Supervised Machine learning approaches used to train a model for predicting the target class in this problem are:

i. Decision Tree Classification

Decision tree is a supervised learning algorithm. This indicates that they train an algorithm that can make predictions using prelabelled data. Each leaf node of the tree representation represents a class label, whereas each internal node stands for an attribute.

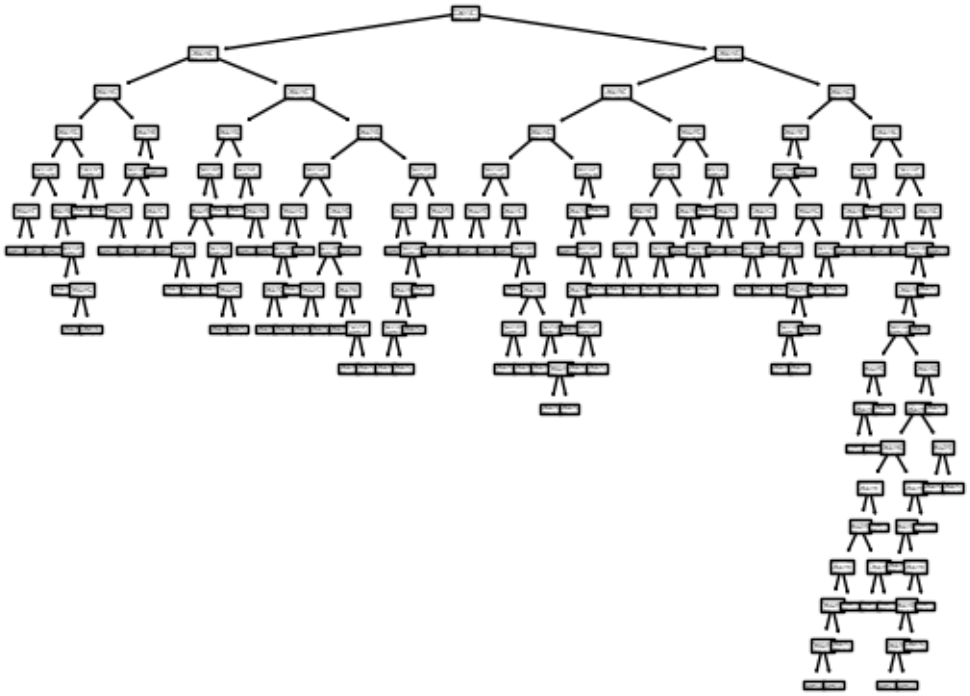

Figure 5: Decision Tree obtained for the given data set after modeling

ii. K Nearest Neighbours (KNN) Algorithm

After examining all available data points, the KNN (K Nearest Neighbors) algorithm assigns new instances to already defined categories. The Euclidean Distance is used to express the distance between two points, and K indicates how many neighbors there are.

iii. Support Vector Mechanism (SVM)

It is an algorithm for supervised classification. It is necessary to switch the hyperplane dimension from one dimension to the Nth dimension. This is known as a kernel. Radial Basis Function(rbf) Kernel is employed in this task to categorize the data.

1.5.Model Evaluation

In this data set, missing values are handled using mean imputation and KNN imputation. After these imputations, the given data is normalized using a conventional scalar approach, and one of three classification algorithms is used to train the model. The accuracy of the various algorithms' results in the provided data set is displayed in the tables below.

Classification Methods	Confusion Matrix	Precision Score	Accuracy Score
Decision Tree	[[304, 76], [165, 255]]	0.77	0.745
SVM	[[342, 38], [242, 178]]	0.82	0.65
KNN	([[244, 136], [198, 222]]	0.82	0.58

Table 1: Performance Evaluation Models with KNN Imputation

Classification Methods	Confusion Matrix	Precision Score	Accuracy Score
Decision Tree	[[277, 118], [86, 319]]	0.54	0.77
SVM	[[257 138] [108 297]]	0.69	0.69
KNN	[[220 175] [163 242]]	0.68	0.57

Table 2: Performance Evaluation Models with Mean Imputation

Analysis of Findings

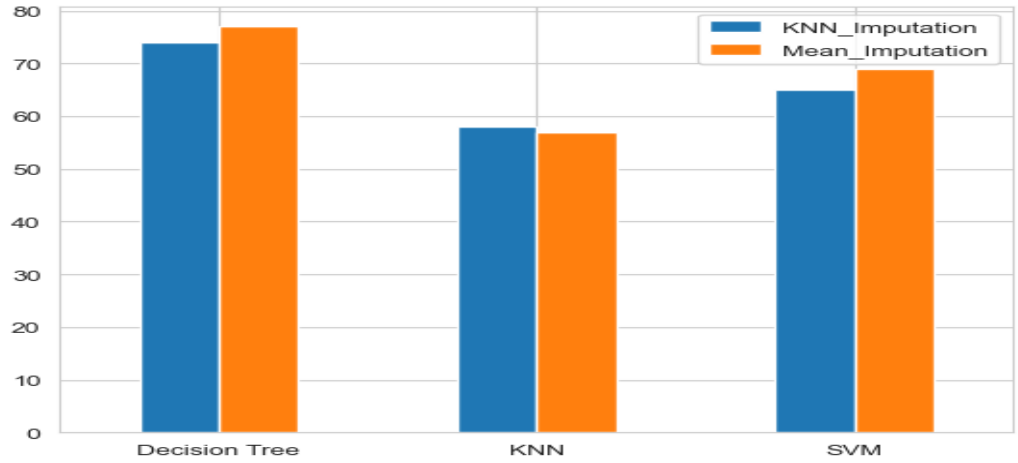

Figure 6: Model Comparison Chart with respect to accuracy

The Decision Tree Classifier has the best accuracy out of all the previous evaluations, with a rate of roughly 77%. For both imputation techniques, the decision tree classifier's accuracy rating was the highest. All classifiers performed better with the mean imputation approach than they did with the KNN imputation method, except for the K Nearest Neighbour (KNN) classifier. The KNN classifier has the lowest accuracy, at roughly 57%. SVM classifier achieved accuracy of 65% and 69% with KNN and mean imputation, respectively. Out of 1000 samples, the decision tree classification model's confusion matrix accurately predicts 277 samples as true positive and 319 samples as true negative.

1.6. Implementing Selected Model on the Test Data

The decision tree classifier has the best accuracy performance, according to the model comparison. The test set data were used to train the decision tree classifier. The mean imputation approach is used to fill in the missing values in the data set before applying the model, and the standard scalar method is used to normalise the data. The model is then used, and a forecast is made and saved in the desired column.

2.Fluctuation In Annual Expenditure Prediction Using Regression

2.1.Data Preparation

Historical information is provided, together with customer characteristics and a number indicating the fluctuation in annual spending for each client.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	...	F28	F29	F30	F31	F32	F33	F34	F35	F36
0	271.78	-21.92	UK	-10.80	1899.57	17.98	-21749.82	91.94	855.61	10.01	...	2098.80	500.16	-6325.53	1.36	-33.14	177.34	-141.97	101.52	5
1	202.54	43.82	USA	-16.94	1941.57	-9.16	-27668.04	93.50	975.44	7.29	...	1668.70	434.97	-6172.05	2.59	-58.87	87.48	-154.11	623.22	6
2	220.26	88.90	Europe	-18.76	2298.12	-18.38	-11548.56	65.16	1114.28	12.05	...	2604.56	252.93	-10132.68	2.94	-40.89	271.00	-279.84	284.96	2
3	141.00	140.72	Europe	-19.86	-133.32	-57.00	-16200.96	-14.00	910.12	4.54	...	2595.56	154.83	-7862.04	0.86	-117.03	201.66	-153.93	532.19	4
4	165.04	2.74	Europe	-21.34	3077.07	-20.50	-25683.06	29.08	216.24	10.10	...	1066.80	316.68	-6093.81	3.59	-63.84	211.82	-182.34	373.14	5

5 rows × 37 columns

Figure 7: Description of Data

Categorical value transformation

Columns F3 and F11 in the provided data set feature categorical values that need to be converted to numerical values. The label encoding approach involves transforming the labels into a numeric form, in order to make them machine-readable and is used to manage various category data.

Feature Selection

The accuracy of the model is increased by using the correlation technique to choose the best features from the data set. Simply said, a correlation matrix is a table that shows the correlation coefficients for various variables. The coefficient threshold is initially set to 0.7. However, there isn't anything over this level here. As a result, the data set is not modified based on the correlation coefficient of any attribute[3].

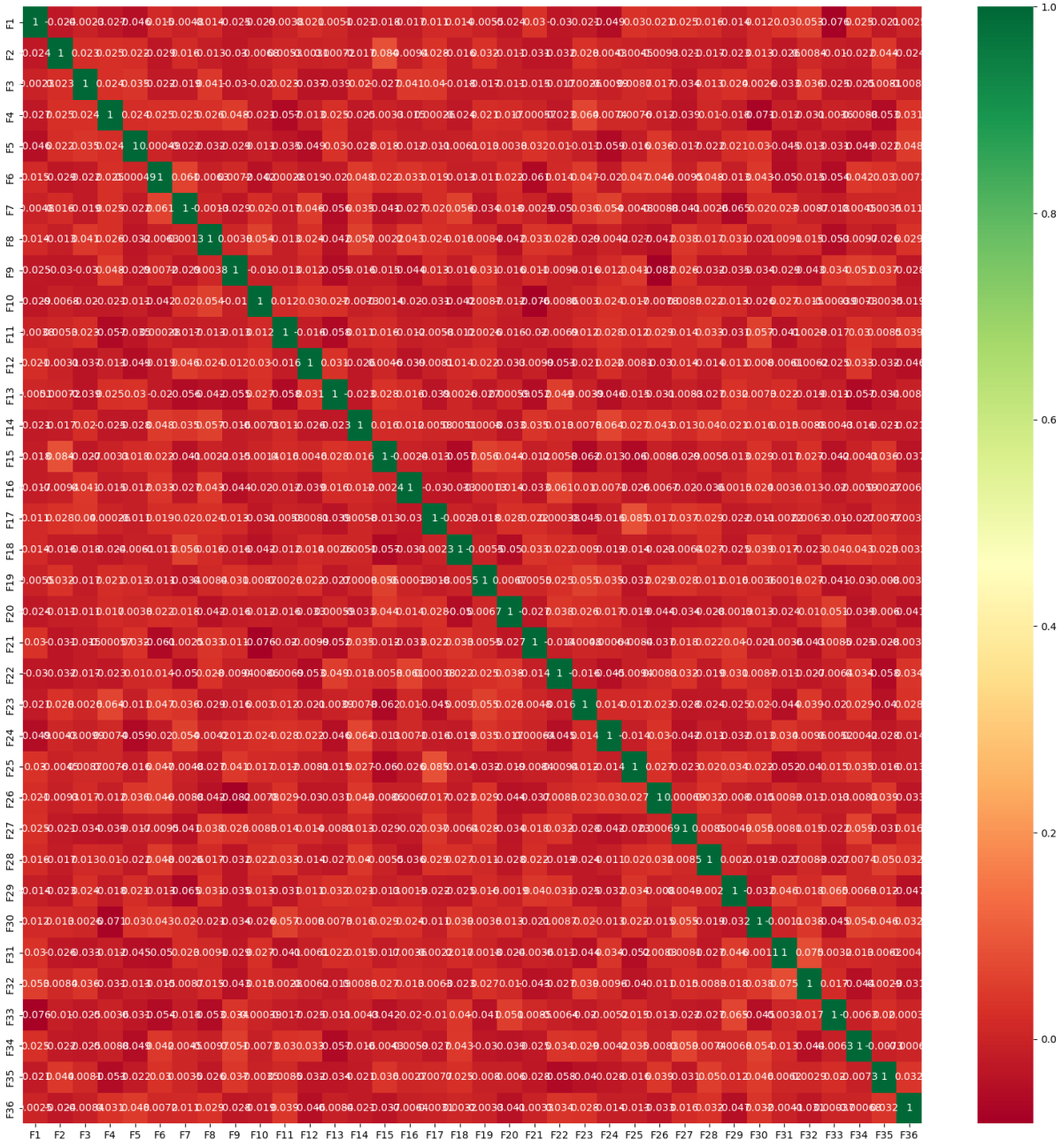

Figure 8: Correlation Matrix

2.2.Modeling Using Regression Methods

i. Liner Regression

Linear approach for modelling the relationship between dependent and independent variables.

ii. Random Forest implementation

This approach uses multiple decision tree results for learning the model.

iii. Decision Tree Regression

One of the most used algorithms for solving regression-based problems.

iv. Ridge Regression

This is one of the regression methods used to obtain low mean squared error[5].

2.3.Model Evaluation

Regression Models	Mean Squared Error	R Squired Value
Linear Regression	61856	0.66
Random Forest Implementation	88839	0.52
Decision Tree Regression	161394	0.13
Ridge Regression	61752	0.66

Table 3: Model Comparison

It is evident from the research that the mean squared error and R squared values for Ridge Regression and Linear Regression are roughly equal. The mean squares error for the Random Forest Implementation is 88839. In this data set, Decision Tree Regression performs the least well.

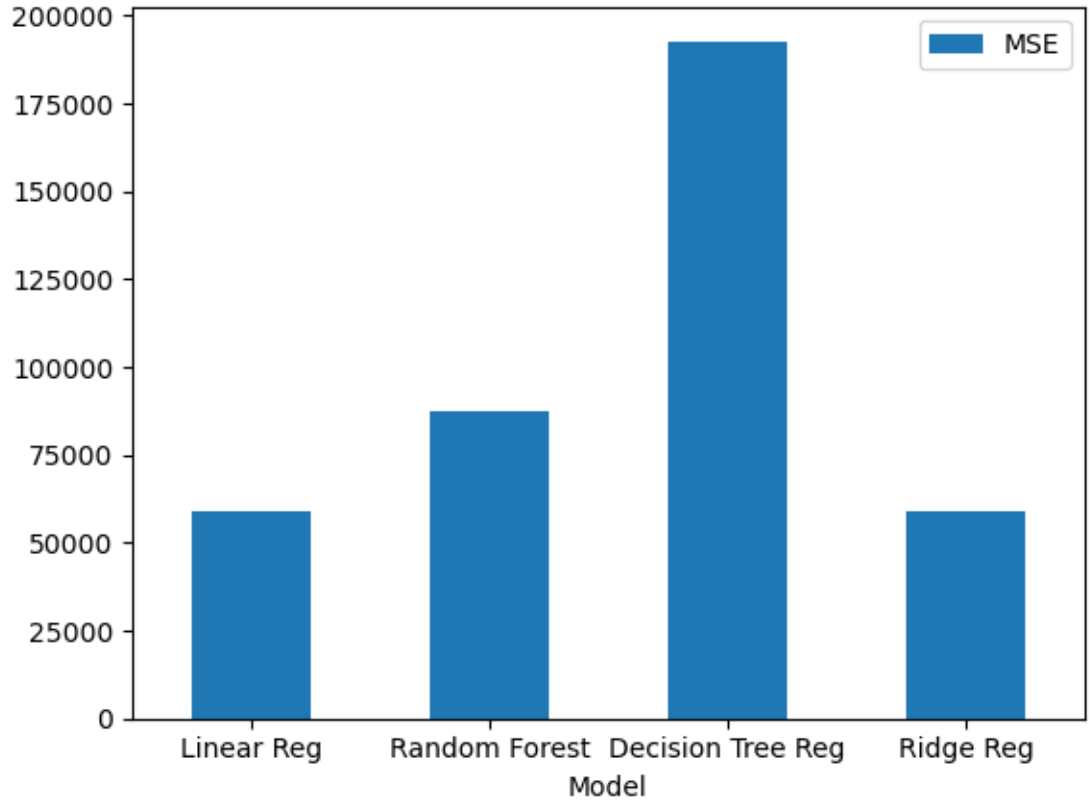

Figure 9: Model Comparison With Respect to Mean Squared Error

2.4. Implementing the Selected Model on the Test set.

Ridge regression is chosen to train the model in the test data set based on the examination of the performance matrix mentioned earlier. The categorical value features are transformed using label encoding during the data preparation phase, and the data is normalised using a traditional scalar approach. The Ridge Regression model is then used to make the forecast, which is then saved in the destination file.

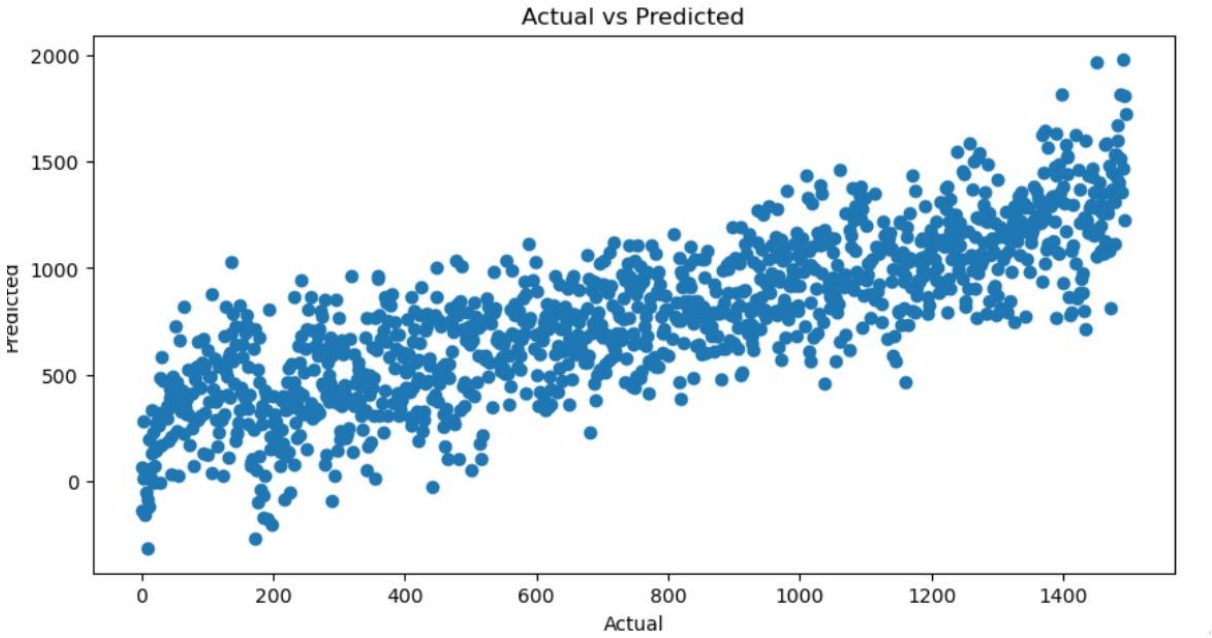

Figure 10: actual vs predicted values to see the difference

Summary

In conclusion, decision tree classifier is used to forecast if a customer will have trouble paying a growing energy bill, and ridge regression is used to predict changes in the annual expenditure of customers.

References

1. https://scikit-learn.org.
2. Lecture notes on machine learning and Lab notes on scikit-learn, pandas
3. Feature Selection in Machine Learning: Correlation Matrix | Univariate Testing | RFECV | by Zipporah Luna | Geek Culture | Medium
4. https://www.linkedin.com/learning/machine-learning-with-python-foundations/what-is-supervised-learning?autoplay=true&resume=false&u=51088249
5. Introduction to Ridge Regression - Statology