2201010

CE807_Machine Learning (950 words)

Design and Application of a Machine Learning System for a Practical Problem

Assignment Report on the Investigation

Introduction:

This project aims to find the most appropriate machine learning model which can be used to solve an issue involving the forecast and management of a customer's electricity bill in AENERGY.

The study determines if a customer of the AENERGY company will struggle to pay the electricity bill when there is a rise in electricity cost based on features like habits, past payment, heating system efficiency, family, etc. The predictive task must be carried out in order to move forward with the forecast. Out of all the predictive tasks, we choose classification because it belongs to a class, which indicates whether certain classes of data points in the dataset belong to them. We choose the best-suited preprocessing techniques based on our dataset to perform regression classification.

Part 2: Comparative Study

In part 2 of the code, which is a comparative study, the given dataset is imported and we obtain a training dataset containing 21 features (F1 – F21) and a column named class which specifies bool values about which class each row belongs and represents in (True/False). We also obtain a test dataset to predict the class. Therefore, it is clearly a classification problem of supervised learning.

Several packages like NumPy, pandas, math, seaborn, and scikit-learn are imported in order to perform various operations on the given dataset and to apply different machine learning models. Exploratory data analysis is performed after preprocessing the given CE802 P2 Data.csv dataset.

Understanding the Data:

To understand the data better, a correlation matrix is driven to see how the variables are correlated with the target variable. The correlation matrix is shown below fig:

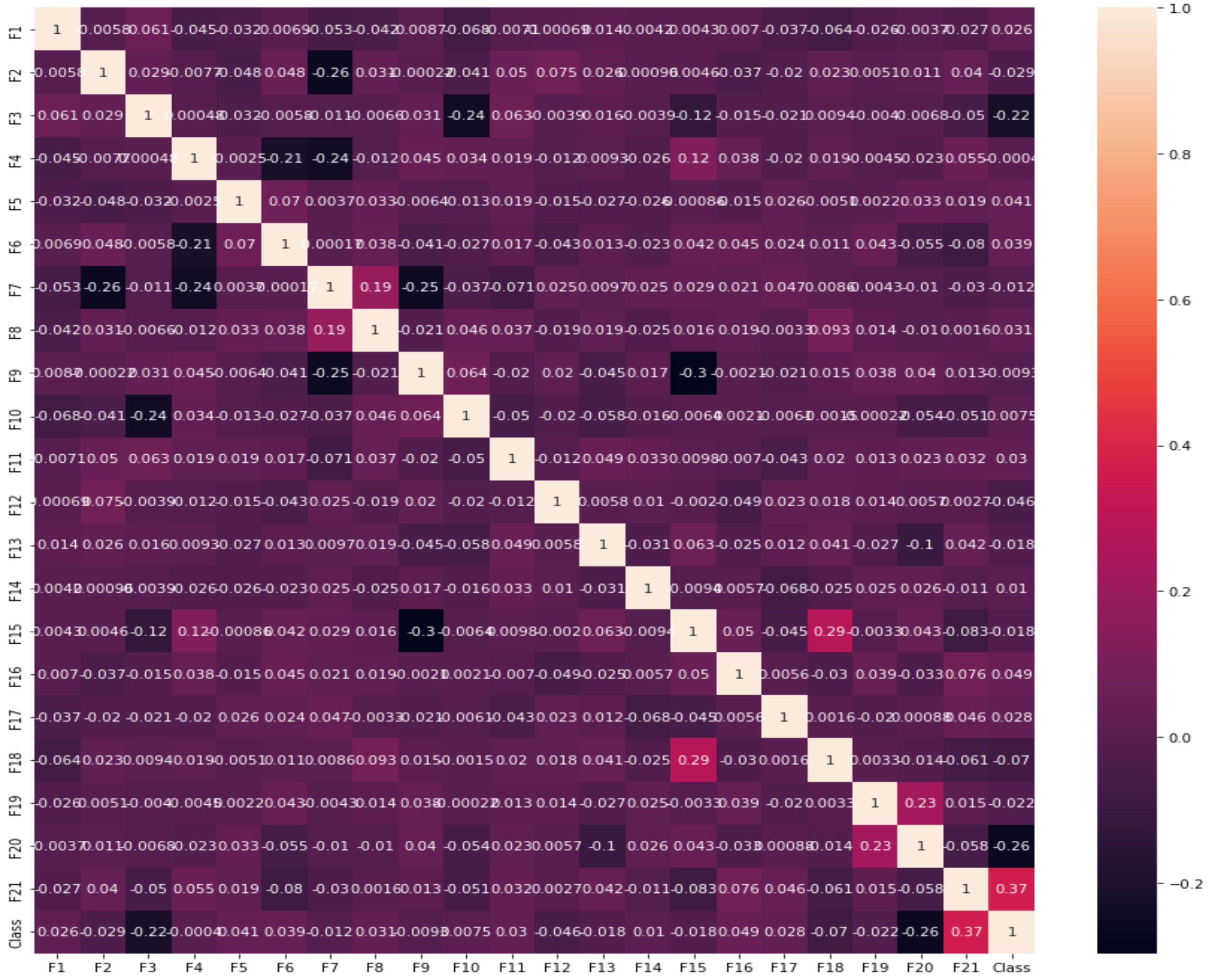


Fig.1: Correlation Matrix of variables with Target variable

Fig.1 shows that the features are not highly correlated with the target variable

Interpretation of features in boxplot:

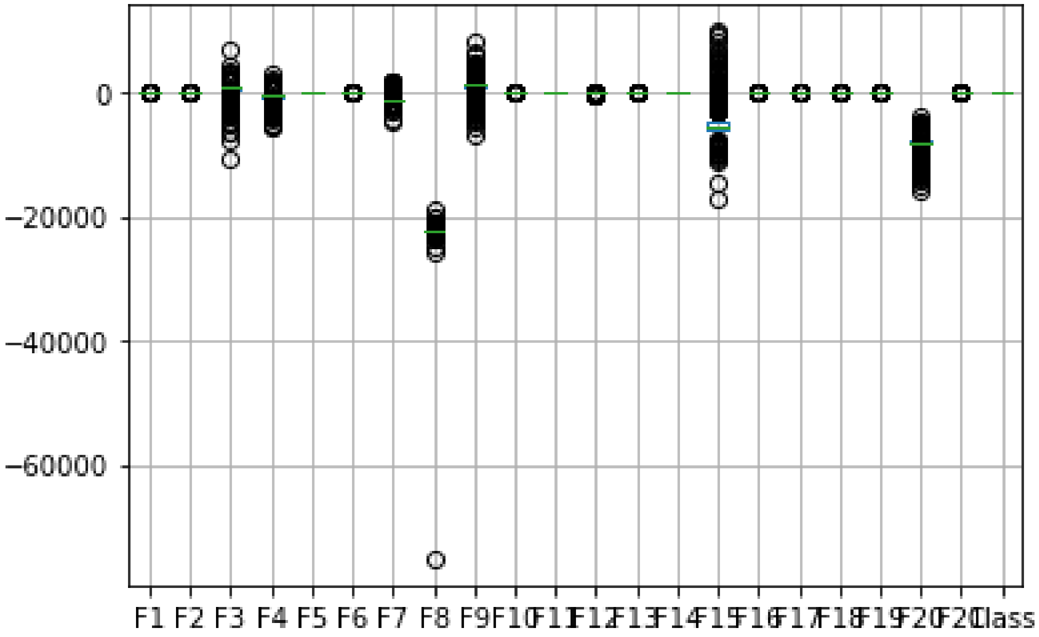


Fig.2: Boxplot showing features

As seen in the boxplot, the data is not normalized, and we see the True or False class in the below bar chart.

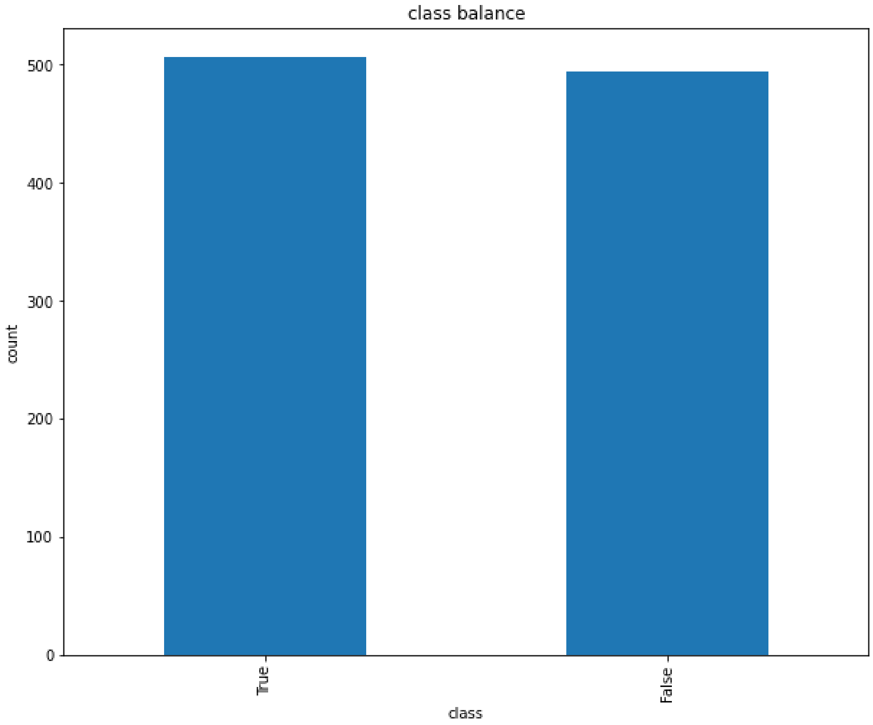


Fig.3: Bar graph showing no. of true/false values

The data is skewed toward true class.

Features of the training data and Handling the missing values using imputation:

There are missing/null values in feature 21(‘F21’) in the given dataset, which we need to preprocess in order to proceed further and get better results. The missing values can be handled by removing them or by using the imputation method to fill in all the missing values. The feature is dropped as it has more than 50% of missing values as a part of the baseline approach.

Splitting the data into train and test data:

In order to avoid overfitting, data splitting is done. For the same reason, once the null values column is dropped, the given data is separated into two parts X and Y. To obtain the necessary training and test sets for the model's training, these two X and Y are provided to the test train split function.

Classifiers used to evaluate the data:

1. Decision Tree
2. K Nearest Neighbor
3. Support Vector Machine

Decision Trees have been fitted on the given training and test data and generate decisions. The c lassification report is shown. The accuracy of the decision tree is 81%.

```
Classification report:
```

```
              precision    recall  f1-score   support

       False       0.78      0.84      0.81        96
        True       0.84      0.78      0.81       104

    accuracy                           0.81       200
   macro avg       0.81      0.81      0.81       200
weighted avg       0.81      0.81      0.81       200


Kappa Score: 0.6206070287539935

Confusion Matrix:

[[81 15]
 [23 81]]

The accuracy is:  0.81
```

K-Nearest Neighbour and Support Vector Machines:

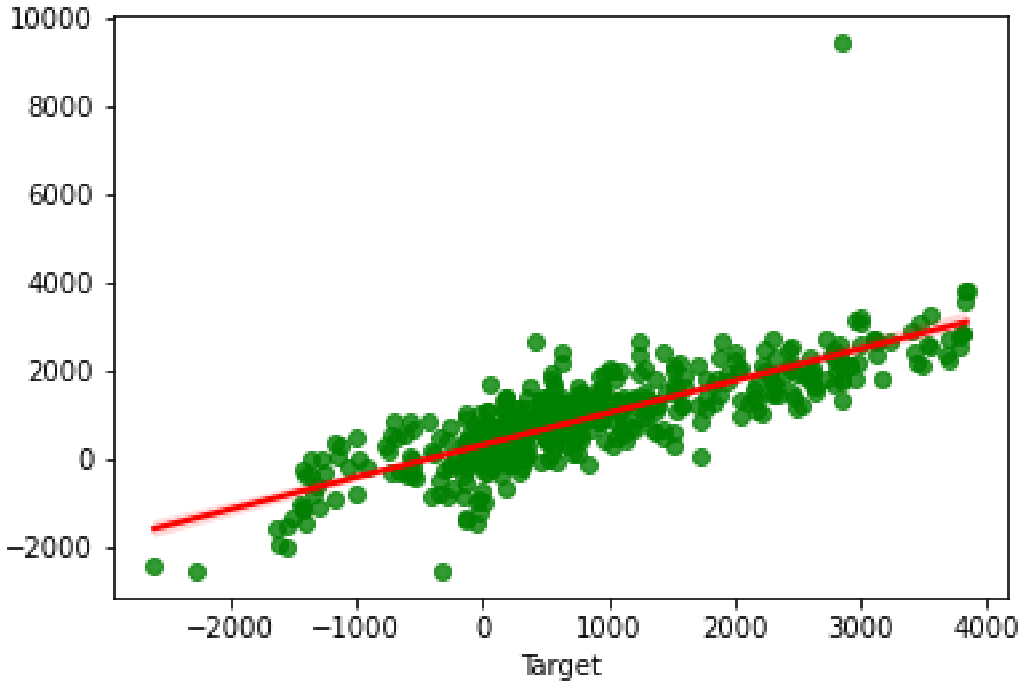
K-Nearest Neighbour and Support Vector Machine classification is performed on the data and the accuracy is 0.68 and 0.66 respectively. Therefore, Decision Tree classification has more accuracy than the other two classifiers.

Performing Mean Imputation:

Mean imputation is applied to the Class feature having missing values. Then the three machine learning models that were used before are performed again on the imputed data to check which imputation produces the best results. X and Y are set as a feature and target variables where x has all the features except class and the target variable class is set as y.
Finally, the three models are then compared by implementing them in a bar chart.

Part 3 Regression Models:

In the second task, the training set with numerical values and information about habits, past electricity prices, family composition, etc.,) are given. Now, Regression Model is built based on the training data and the target value is predicted for the given test data. The results will be forecasted using a **linear regression model**, which explains the relationship between dependent variable, target variable and input factors.



Preparing data for regression model:

Analysing the data, there are categorical data in two of the columns with feature names F5 and F21. The data present in these features are changed into ‘int’ or ‘float’ using mapping.

Splitting the data:

The data is split again as we did in part 2 by separating the target value from the dataset and creating x, and y. Since there are no missing values, imputation is not performed. The Random Forest Classifier is used to build the model, the r mean square is calculated and the accuracy is 0.68, whereas, for linear regression and lasso classifier, the accuracy is 0.63. Based on the results, it is seen that the random forest classifier is the best for predicting the model.

Conclusion:

After conducting the studies and performing various classifiers and regression models, it is observed that **random forest regression** is the best approach for predicting the model. Following the prediction, the P3 test predictions.csv which is the result will be published into a different Excel file. The findings of this study will be helpful to AENERGY proactively in understanding whether it is possible to foresee trouble in paying energy bills using machine learning models and they will also guide the creation of future predictive models.