Assignment CE802 Machine Learning**:**
Design and Application of a Machine Learning System for a Practical Problem.

Report on the Investigation:

1.) This study aims to identify and implement a suitable machine learning technique appropriate for a problem regarding the prediction of a new hotel of a large hotel chain if opened in a given location to be profitable or not.

Location is one of the most important factors affecting the business of a hotel. The location of the hotel determines whether the business of the hotel would seek profit of not. The location along with the geographical data and socio-economic data such as health, crime, population, availability to public transport etc. would be some of the most important factors required for us to proceed with the machine learning process. Luckily, we are provided with the historical data of successful and unsuccessful hotels opened under the chain's brand on similar locational conditions and also provided with the geographical and socio-economic data about the locations and neighborhoods.

To proceed with the prediction, we must perform predictive task on the given set of data. There are various predictive tasks available for designing a machine learning system. Some of them are,

- Classification
- Regression
- Clustering
- Rules mining etc.

We will be proceeding with the Classification type of prediction technique because classification predicts the belonging to a class. Classification technique is preferred when the results of the model need to return the belongingness of data points in a dataset to specific classes. In our case, we want to find out whether the hotel would turn out to be as either class, profit which is true or class loss which is false based on the given set of data.

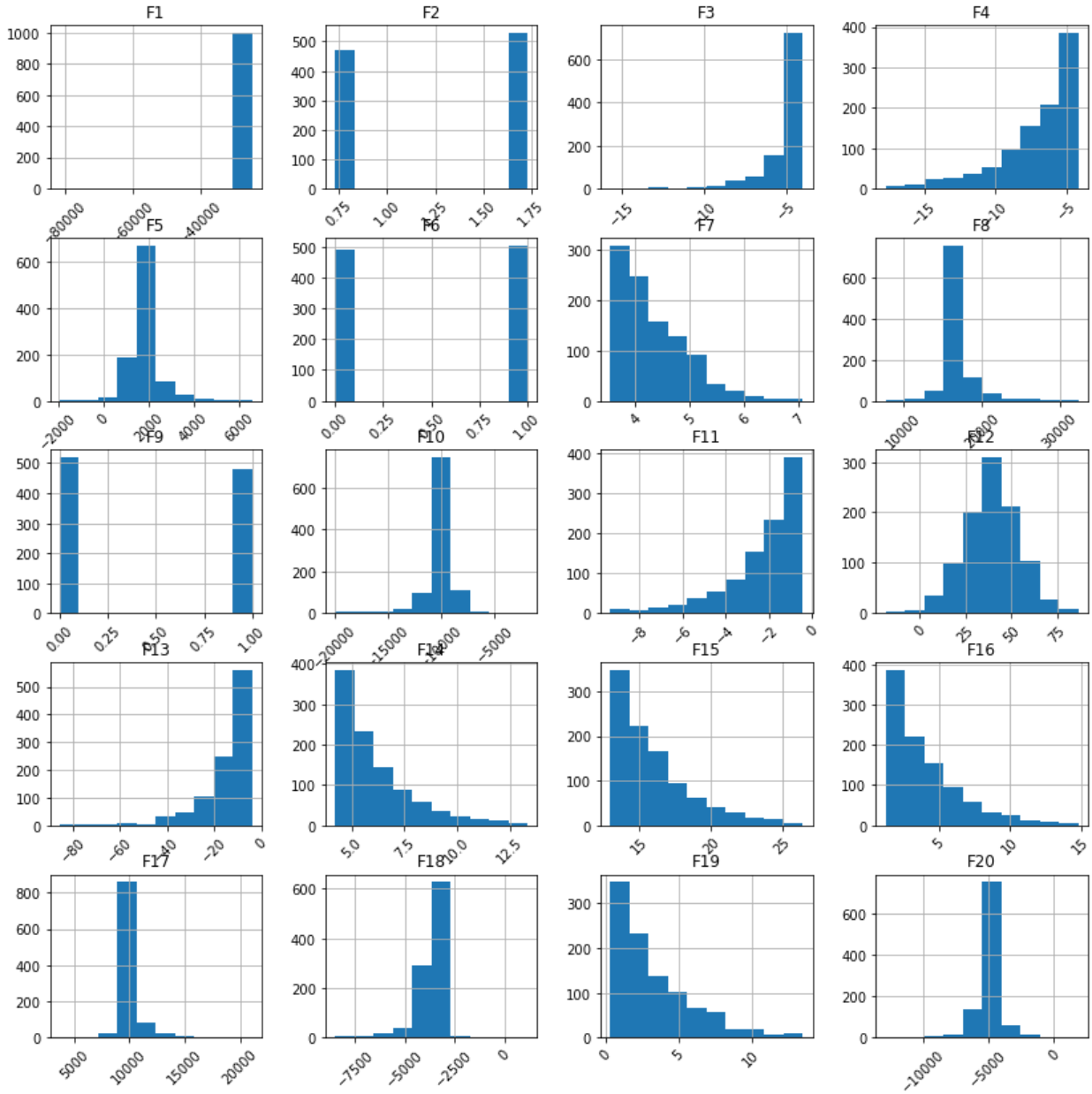

Fig 1.1: The values of the coloumns from F1 to F20.
Here we can only plot the values of the coloumns from F1 to F20 because before preprocessinh, F21 and class has missing values.

After selecting the type of technique, we are going to proceed with the preprocessing process. The preprocessing process helps to improve the accuracy of the model. There are various preprocessing methods, and we must select the ones that we need based on our dataset. In, our case we have over five hundred missing values in the column F21. To fill the missing values, we can find either one of these mean, median and mode and fill up the spaces of the missing values. We are using mean of the coloumn to fill up the missing values.

The next step is we have to pick a suitable classifier for our data. There are various types of classifiers available to perform the classification process. Some of them are,

- Decision Tree classifiers
- K-Nearest neighbor
- K-means
- Random Forest
- Support vector machine
- Naïve Bayes etc.

We will be implementing some of these classifiers to see which one of those classifiers predicts the most accurate result for our dataset. We'll be using Decision tree mode, Support vector machine model and Naïve Bias model and find which one of those classifiers predicts the more accurate data.

First we will use the **Decision tree model** to predict the outcomes. A Decision tree gets its name as it uses a tree like model to make decisions and the decision's possible consequences. First we split the P2_Data csv into train and test data. We'll train the model using the train data and use the test data to predict the output. For the Decision tree algorithm, we get a prediction accuracy score of 81.2%.

Then we will use the **Support Vector Machine model** to predict the outcomes. It uses classification algorithms for the process of group classification. In our case, the group to be true or false. First we split the P2_Data csv into train and test data. We'll train the model using the train data and use the test data to predict the output. For the SVM algorithm, we get a prediction accuracy score of 64.4%.

And then we will use the **Naive Bayes model** to predict the outcomes. It uses Naïve Bayes's theorem to get the predictions. First we split the P2_Data csv into train and test data. We'll train the model using the train data and use the test data to predict the output. For the Naïve Bias algorithm, we get a prediction accuracy score of 65.6%.

Out of these algorithms, Decision tree algorithms seem to predict a more accurate result compared to the other two models. So, we'll use decision tree algorithm for our dataset. Now, we'll implement the decision tree on P2_test_data csv file to predict the class colomn. After the prediction, we'll print out the predictions into another excel under the name, P2_test_predictions.csv which would be our desired output.

Fig 1.2: The values of input coloumns plotted with respect to class after prediction.

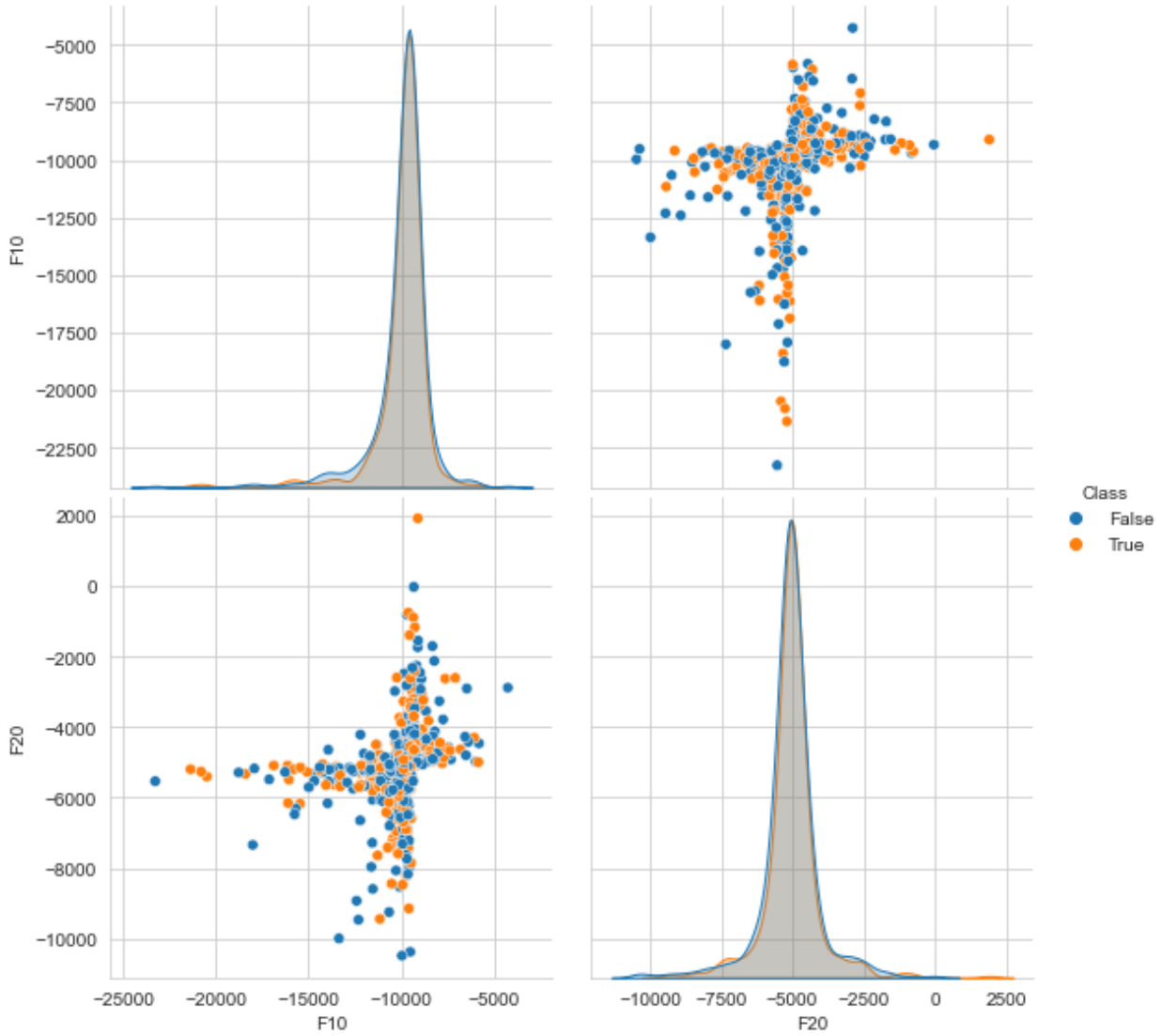

Fig 1.3: The values of input coloumns plotted with respect to class after prediction.

After prediction the values of the prediction data are plotted as a graph for the purpose of understanding.

2.) Now we'll have to build a similar Machine learning system for a different company. But in this case the company provided me with the data of each of the company's business along with the numerical data of the company's profit or loss margin. Here the company wants us to build an algorithm to check if the new business would be profitable along with it's annual profit.

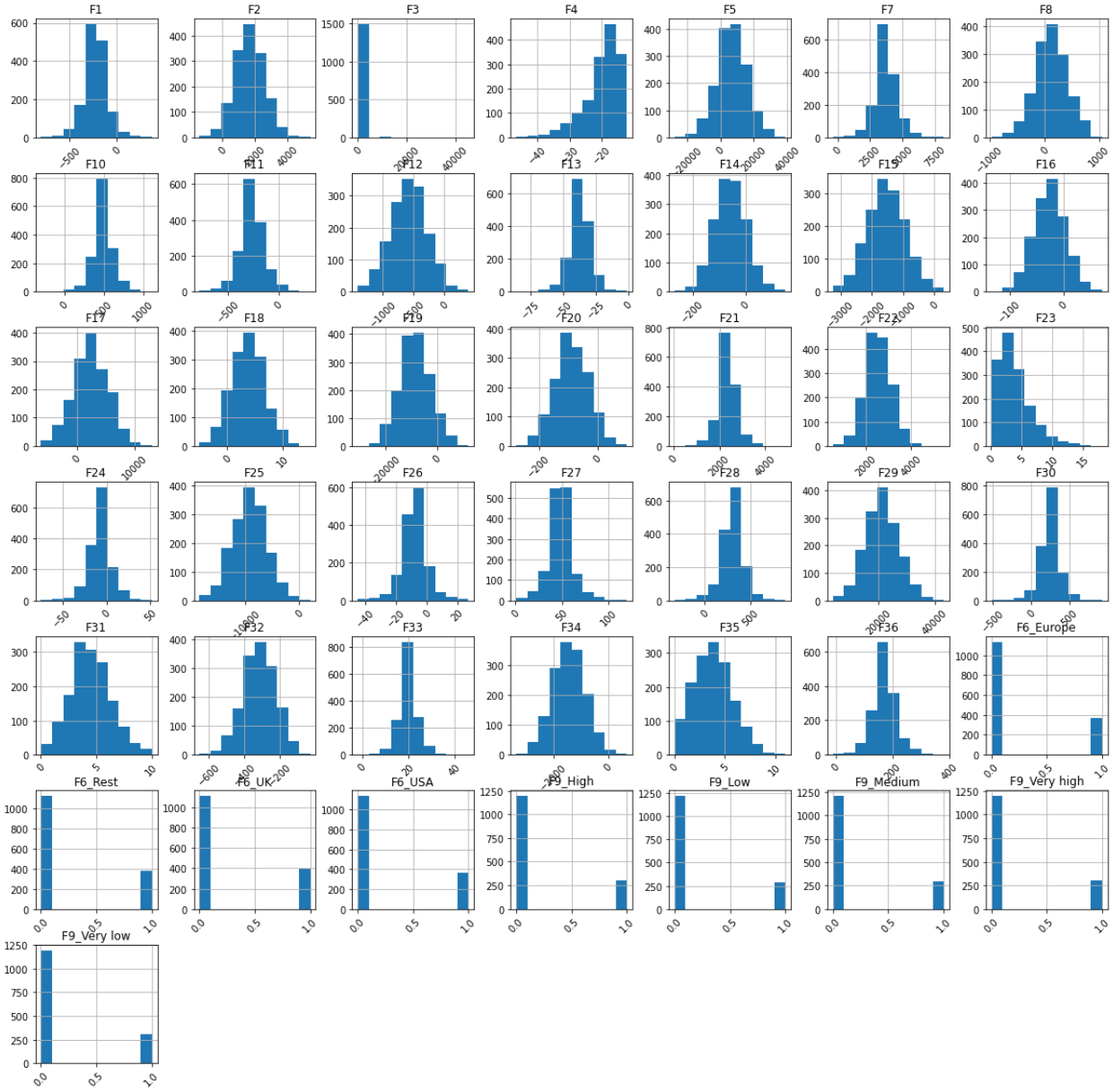

Fig 2.1: Plotting all the input coloumns in the form of Histogram

Since we have to do predict the values of annual profit and the term annual profit is a continuous outcome, we'll use **Regression** technique to perform the prediction function. There are various types of regression algorithms available, some of them are:

- Linear regression algorithm
- Elastic net algorithm
- Lasso regression algorithm
- Ridge regression algorithm.
- Logistic regression algorithm
- Decision tree regression algorithm etc.

Before selecting the type of technique, we are going to proceed with the preprocessing process. The preprocessing process helps to improve the accuracy of the model. There are various preprocessing methods, and we must select the ones that we need based on our dataset. In, our case we have Categorical values in coloumns **F6** and **F9** we have to change the datatype from **String** to **Float**.

The next step in pre processing we'll proceed with correlating the input coloumns with the **Target** coloumn and drop the coloumns that are least affecting the target in order to improve the accuracy.

Now, that the preprocessing is completed, we'll proceed with the regression process. First we will use the **Linear regression model** to predict the outcomes. It describes the relation between dependent variable which is Target and one or more input variables. First we split the P3_Data csv into train and test data. We'll train the model using the train data and use the test data to predict the output.

Then we will use the **Decision tree model** to predict the outcomes. A Decision tree gets its name as it uses a tree like model to make decisions and the decision's possible consequences. First we split the P3_Data csv into train and test data. We'll train the model using the train data and use the test data to predict the output.

And then we will use the **Elastic Net model** to predict the outcomes.It is used in fitting the linear and logistic models of regression. First we split the P3_Data csv into train and test data. We'll train the model using the train data and use the test data to predict the output.

We'll get the accuracy and root mean square values of each model to determine which model is more accurate to proceed with.

Model Type	Accuracy Score	Root mean square error
Linear regression	68.8%	689
Elastic Net	56.3%	810
Decision tree	23.5%	1071

Fig 2.2: Table containing the values of Accuracy and Root mean square error for their respective models.

From Fig 2.2, we see that **linear regression model** yields more accuracy compared to the rest so we proceed to perform linear regression to predict the Target coloumn of P3_test data. After the prediction, we'll print out the predictions into another excel under the name, P3_test_predictions.csv which would be our desired output. The company can use this model for their future prediction of Annual incomes.