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

Pilot Study proposal:

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.

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.