**INTRODUCTION:**

In this assignment, we are going to perform exploratory data analysis and inferential data analysis on the bank dataset, and we are going to build a machine learning model in order to predict the terms of a deposit subscription by the clients in the bank. It will be a supervised machine learning and this model will try to solve the classification problem like whether the client will subscribe the term deposit in a bank or not.

Let us now understand the dataset we have been given. What is there is in the data file and what are the columns present? Basically, this dataset includes the data retrieved and uploaded into the UCI Machine learning repository from a classic marketing bank. It has all the information belonging to a marketing campaign of a bank and we will have to analyze this data to come up with strategical planning to improve the marketing campaigns in the future for this financial institute. In the dataset given, the target column is ‘DEPOSIT’.

**Dataset Description:**

This dataset consists of the data such as client’s educational information, loans, balance, contact information of the client and so on. Based on these details, we can predict if the customer is interested in the term deposit and these values are shown in the form of YES or NO in our dataset.

Now let us understand what exactly term deposit means, it is a deposit system which is a little different from the normal deposit account in a bank. This is specially designed for a specific period where one will receive back the amount after the certain period ends with a fixed rate of interest.

**Analysis Part:**

After closely analyzing the dataset for its data structure that is analyzing its rows and columns, we found out that it has 11,000 plus rows and 17 columns.

**Numerical features** include age, balance, day, duration, campaign, pdays and previous.

We will further understand the **categorical features** and its unique values, and they include job, education, default, marital status, housing, loan, contact, month, poutcome and deposit.

**Brief description of the numerical and categorical features:**

1. Age: It is represented in years
2. Job: It is categorized into various jobs such as ‘admins’, ‘services’, ‘management’, ‘housemaid’ etc.
3. Marital: Status of marriage is categorized by ‘single’, ‘divorced’ or ‘divorced’.
4. Education: It is categorized as ‘primary’, ‘secondary’ or ‘teritary’
5. Default: Has credit in default is described as yes or no.
6. Balance: account balance
7. Housing: Is it loan based or not and is categorized by yes or no.
8. Loan: Did the client take a personal loan or not in yes or no.
9. Contact information: Client’s contact recorded as ‘cellular’, ‘telephone’ or ‘unknown’.
10. Day: It represents last contact day of the week and is represented as ‘mon’, ‘tue’, ‘wed’ etc.
11. Month: It represents last month of the year and is represented as ‘jan’, ‘feb’, ‘mar’ etc.
12. Duration: It represents the last call duration in seconds.
13. Campaign: It gives out the information on no. of contacts that were performed during the campaign and for that particular client.
14. Pdays: It is the information about the number of days that has been passed by after the client has been contacted from the last campaign.
15. Previous: For this client it represents the number of times we have contacted them during the previous campaign.
16. Poutcome: this is the outcome from the previous campaign

**Label:**

Deposit: It described if the client has subscribed to the term deposit or not and it is mentioned in the form of ‘yes’ or ‘no’.

**Methods used:**

For this assignment we have used methods such as decision tree algorithm, random forest and linear regression analysis. We have plotted various graphs such as histograms, box plots etc.

* **Decision tree:** It is one of the oldest yet very simplest and useful machine learning algorithm that have been there is use for a long time now. It can be broadly used for two purposes. One is for classification and the other is for regression. Regression is really about determining the set of continuous numbers, so if you wanted to predict a number or something numerical that’s when we use regression. Classification is when you want to predict a class as into which class of the given data in the dataset does it actually belong to. The combination of both these terms is known to us and is referred to as CART. Having a good representative set for the training data, the data that you feed into your training clearly influences the results that you get off your predictive models.
* **Random Forest:** It is the most powerful supervised machine learning algorithm that is capable of performing both regression and classification tasks. As the name implies, this algorithm creates forest with a number of decision trees. In general, the more trees in the forest, the more robust is the prediction and thus gives us higher accuracy to model multiple decision trees to create the forest. We are going to construct the decision with the information gained amongst other algorithms.
* **Linear Regression:** It attempts to model the relationship between two variables by setting a linear equation to the observed data. One variable is explanatory, and the other variable is dependent. Dependent is a variable whose values we want to explain or forecast. The explanatory variable is something that explains the other variables and to find if the values are independent.

Now, let us talk about regularization. One very powerful play that you use regularization is regression. Regularization is a method for automatically penalizing the extra features that you use in your model. There is one type called as LASSO regression and here is the formula:

Method for automatically penalizing extra features:

Minimize SSE +

Here, lambda is the parameter and beta is the coefficient of regression. A regular linear regression would say that I want to just to minimize the sum of the squared errors in my fit. I want to minimize the distance between the fit and any given data point, or the square of that distance. What LASSO says is that we want a small sum of squared error. In addition to minimizing the SSE, we also need to minimize the number of features used. So we add the second part which has penalty parameter and also coefficient regression. When we are performing the fit we consider both the errors that come from that fit and also the number of features that are being used. If we are comparing two different fits, which has different number of features in them, the one that has more features included in them will almost certainly have a smaller SSE because it can fit more precisely to the points. But we pay penalty for that extra feature used. So the gain that we get in terms of the precision, the goodness of fit of my regression, has to be bigger gain than the loss that we take as a result of having that additional feature n my regression. This is thus a mathematical way to say having small errors and having a simpler fit using fewer regressions. And what LASSO does is that it automatically takes into account this penalty parameter. In so doing it actually helps you figure out which features that are the ones that have the most important effect on your regression. Once found, it can eliminate them or set it to zero, the coefficients of the features that actually don’t help.

**Nalini’s individual Part:**

For this banking dataset my part was to perform exploratory data analysis. I have tried to find the unwanted columns, let’s say, if there is an ID sequence or serial numbers out there. In our dataset there is no unwanted columns. Later we will try to find out the missing values. If there are any missing values, we will try to find the relation between them and the targeted variable. But we find that there are no missing values from our code. We will also find the features with one value because if there is only one value then there is no use of them in the dataset. We run the code and check with values with one, we find no feature with one value. We then find the categorical features such as job, marital, education etc. in this dataset. To find the categorical features we try to write a comprehensive list, where we are actually trying to find the features only with the datatype that starts with ‘O’and the column not of deposit which is our target column. We find that there are total of nine categorical features and the ‘job’ and ‘month’ features has the highest categorical values there is. We next find the distribution of these categorical features. We have plotted the graphs to find the distribution. After we find the categorical features next we need to find the relation between these features and the labels. After we are done with categorical features, we shift to finding numerical features such as age, balance, day etc. in the dataset. First, we are going to find the discrete numerical features in the labels and then we will find the continuous numerical features and its distribution and also its relationship with the labels. Finally, we will try to find the outliers of these features and further we try to explore the correlation factor between them. I have also replaced the unnecessary values in the dataset with mean values. We find a pair plot and finally we check for the targeted values in the classification and see if they are balanced or not.