Assignment 1: Customer Subscription Prediction with Machine Learning

- Introduction:

    o The goal of this assignment is to apply the basic concepts of data science to a real-world problem.

    o The dataset you will be using is the Bank Marketing dataset: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing, which contains information about customers of a Portuguese bank who were contacted by telemarketers about a term deposit.

    o Your task is to answer the following questions:

        ▪ Can you build a model to predict whether a customer will subscribe to a term deposit?

        ▪ What are the most important factors that influence a customer's decision to subscribe?

        ▪ What are the limitations of your model?

- Data Preparation:

    o The first step is to load the dataset into a data frame.

    o You will need to clean the data by removing missing values and outliers.

    o You may also need to transform some of the data, such as converting categorical variables to numerical values.

- Exploratory Data Analysis (EDA):

    o Once the data is clean, you can start exploring it.

    o This involves creating visualizations and performing statistical analyses to understand the data.

    o Some of the questions you might want to answer during EDA include:

        ▪ What is the distribution of the customer ages?

        ▪ What is the relationship between customer age and subscription?

        ▪ Are there any other factors that are correlated with subscription?

- Model Building:

    o Once you have a good understanding of the data, you can start building a model to predict subscription.

    o There are many different machine learning algorithms that you can use.

- Some of the most common algorithms for classification problems (like predicting subscription) include logistic regression, decision trees, and random forests.

- Model Evaluation:

  - Once you have built a model, you need to evaluate its performance.

  - This involves using a holdout dataset to test the model on unseen data.

  - You can use metrics such as accuracy, precision, and recall to evaluate the model's performance.

- Conclusion:

  - In the conclusion, you should summarize your findings and discuss the limitations of your work.

  - You should also discuss the implications of your findings for the real-world problem.

Dataset Download Link:

The Bank Marketing dataset can be downloaded from the UCI Machine Learning Repository website: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

The 10 questions that students need to solve are:

1. What is the distribution of the customer ages?

2. What is the relationship between customer age and subscription?

3. Are there any other factors that are correlated with subscription?

4. What is the accuracy of the logistic regression model?

5. What are the most important features for the logistic regression model?

6. What is the precision of the logistic regression model?

7. What is the recall of the logistic regression model?

8. What is the f1-score of the logistic regression model?

9. How can you improve the performance of the logistic regression model?

10. What are the limitations of the logistic regression model?