Problem 1:

You are hired by one of the leading news channel CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Dataset for Problem: Election_Data.xlsx

Data Ingestion: 12 marks
1. Read the dataset. Do the descriptive statistics and do null value condition check. Write an inference on it. (5 Marks)
2. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

Data Preparation: 5 marks
1. Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). (5 Marks)

Modelling: 26 marks
1. Apply Logistic Regression and LDA (linear discriminant analysis). (5 marks)
2. Apply KNN Model, Naïve Bayes Model and support vector machine (SVM) model. Interpret the results. (7 marks)
3. Model Tuning, Bagging (Random Forest should be applied for Bagging) and Boosting. (7 marks)
4. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized. (7 marks)

Inference: 5 marks
1. Based on these predictions, what are the insights? (5 marks)

Problem 2:

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

President Franklin D. Roosevelt in 1941
President John F. Kennedy in 1961
President Richard Nixon in 1963
Find the number of characters, words and sentences for the mentioned documents. – 3 Marks
Remove all the stopwords from all the three speeches. – 3 Marks
Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) – 3 Marks
Plot the word cloud of each of the speeches of the variable. (after removing the

stopwords) – 3 Marks [in case of doubts for this question, please refer to the End-to-End Case Study done in the Mentored Learning Session]
Code Snippet to extract the three speeches:

"
```
import nltk
nltk.download('inaugural')
from nltk.corpus import inaugural
inaugural.fileids()
inaugural.raw('1941-Roosevelt.txt')
inaugural.raw('1961-Kennedy.txt')
inaugural.raw('1973-Nixon.txt')
```
"

Scoring guide (Rubric) - Machine Learning Project

Criteria	Points
1.1) Read the dataset. Do the descriptive statistics and do null value condition check.	5
1.2) Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.	7
1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).	5
1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis).	5
1.5) Apply KNN Model, Naïve Bayes Model and Support Vector Machine (SVM) model.	7
1.6) Model Tuning, Bagging and Boosting.	7
1.7) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.	7
1.8) Based on these predictions, what are the insights?	5
2.1) Find the number of characters, words and sentences for the mentioned documents	3
2.2) Remove all the stopwords from the three speeches.	3
2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)	3
2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)	3
Points	60