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](https://olympus.greatlearning.in/courses/13602/files/2234615/download?verifier=W5GKzi7L2HRhPARciZJ2UsMF6cIREKeFMA7xXfyV&wrap=1)

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 and Naïve Bayes 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:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

* Find the number of characters, words and sentences for the mentioned documents. – 3 Marks

(Hint: use .words(), .raw(), .sent() for extracting counts)

* 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 [ 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')  
"  
**Important Note: Please reflect on all that you have learned while working on this project. This step is critical in cementing all your concepts and closing the loop. Please write down your thoughts**[**here**](https://docs.google.com/forms/d/e/1FAIpQLSfqHHlmJyUkniiBiejtudIuRFk_TVCLe843wfX6Iu3QNRPmng/viewform?usp=sf_link)**.**