**Major Project Guideline and Problem Statement**

**Objective**

The objective of this major project is to give students hands-on experience in solving a real-world problem, documenting their process in a professional project report, and presenting their work effectively.

**Project Requirements**

1. Project Implementation
2. Project Report
   1. **Format:** Your report should be neatly formatted as per the shared format
   2. **Length:** The report should be between 10-15 pages, excluding the cover page and references.
3. Presentation
   1. **Format:** Your presentation should be neatly formatted as per the shared format
4. Recording the Presentation
   1. Record a video of yourself delivering the presentation.
   2. Make sure both your slides and your voice are clearly visible and audible.
   3. You can use screen recording software such as Zoom, Teams, or PowerPoint’s built-in recording feature.
   4. Guidelines:
   * Practice before recording to ensure a smooth delivery.
   * Speak clearly and confidently.
   * Ensure that the background is quiet and free from distractions.
   * Keep the recording within 10 minutes.

**Submission Guidelines:**

You are required to submit the following:

1. Source code (jupyter Notebook)
2. Project report in PDF format
3. Presentation slides in PPT or PDF format
4. Presentation video recording

**Evaluation Criteria:**

|  |  |
| --- | --- |
| **Submissions** | **Marks Distribution** |
| Source code (jupyter Notebook) | **60** |
| Project report in PDF format | **20** |
| Presentation slides in PPT or PDF format | **10** |
| Presentation video recording | **10** |

**Major Project: Exploratory Data Analysis (EDA) and Machine Learning Integration**

**Problem statement:**

Perform a comprehensive analysis on the given dataset using Python, incorporating exploratory data analysis (EDA) and machine learning techniques. Your task is to preprocess the data, engineer features, select and train models, and evaluate their performance.

***Additionally, document your process with Python comments explaining your code, and for each section, provide detailed conclusions and observations.***

**Section 1: Understanding the Dataset [2 marks]**

* 1. Load Data: Import the dataset into your working environment using appropriate methods or libraries.
  2. Checking Data Shape: Determine the shape of your dataset, including the number of rows and columns.
  3. View Data: Display the first and last few rows of the dataset and summarize any initial insights.

**Section 2: Initial Data Examination [2 marks]**

* 1. Dataset Information: Provide a concise summary of the dataset, including the number of non-null entries, and explain what this reveals.

2.2 Inspect Data Types: Check data type of each column in the dataset. If columns need any data type conversion, update the data type accordingly and describe the rationale behind the conversions.

2.3 Summary Statistics: Generate summary statistics for the numerical columns and interpret what these statistics tell you about the data.

2.4 Provide detailed comments that explain your understanding of the data.

**Section 3: Data Cleaning [2 marks]**

3.1 Handling Missing Values: Identify missing values in the dataset and describe how you handled them, including your chosen method.

3.2 Handling Duplicates: Check for duplicate rows in the dataset and describe your approach to handling any duplicates found.

3.3 Outliers removal: Check if there are any outliers and remove them using graphical/non-graphical methods.

3.4 Add python comments to explain the observations.

**Section 4: Data Analysis [2 marks]**

4.1 Univariate Analysis of numeric features: Generate histograms for numerical data and infer insights from these visualizations.

4.2 Examine the skewness in the data and apply appropriate data transformation technique.

4.3 Apply appropriate standardization method wherever applicable.

4.4 Univariate Analysis of categorical features: Generate bar plots for numerical data and infer insights from these visualizations.

4.5 Encode categorical features

4.6 Bivariate and Multivariate Analysis: Calculate the correlation matrix for the numerical variable. Generate heatmap for the correlation matrix, and describe the evident relationships.

4.7 Provide detailed observations and conclusions.

**Section 5: Feature Selection [2 marks]**

5.1 Use correlation result for feature selection.

5.2 Select the features according to the K highest score.

5.3 Provide detailed insights about the selected features.

**Section 6: Model Selection and Training: [20 marks]**

6.1 Choose at least three different machine learning algorithms to train on the dataset.

6.2 Train the models and apply hyperparameter tunning.

6.3 Provide detailed observations and conclusions.

**Section 7: Model Evaluation: [20 marks]**

7.1 Evaluate the performance of each model using appropriate metrics (e.g., accuracy, precision, recall, F1-score for classification; RMSE, MAE, R square for regression).

7.2 Compare the performance of the models and select the best model based on the evaluation metrics.

7.3 Provide detailed comparison and analysis of the models’ performance.

**Section 8: Model Deployment with web app: [10 marks]**

8.1: Develop an interactive web application using Streamlit.

8.2 Integrate the best-performing machine learning model into the Streamlit app.

8.3 Provide an interface for users to input new data and obtain predictions from the model.