Data Science Project Assignment

Objective:

The goal of this project is to practice working with real-world data using R, to explore and preprocess datasets, and to apply machine learning algorithms to solve problems in classification, regression, or clustering. You will gain experience with data preparation, exploratory analysis, model evaluation, and data visualization.

Assignment Outline:

- Select a **public dataset** from **Kaggle** that is suitable for either classification, regression, or clustering.
- **Data Exploration & Preprocessing**: Use R to conduct a thorough descriptive analysis of the data and prepare it for machine learning.
- **Modeling**: Apply two different algorithms to solve a problem (choose two from classification, regression, or clustering, based on your chosen problem type).
- **Evaluation**: Evaluate the performance of your models and compare results.
- **Report**: Summarize your findings, methodology, and results in a well-structured report.

Deliverables:

1. **Project Code & Files**:
   - Submit the **R script** (.R file) or **R Markdown** (.Rmd file) containing all the necessary code to perform data loading, exploration, preprocessing, modeling, and evaluation.
   - Include a **data dictionary** or a brief description of the dataset, including the source, variables, and their meaning.
2. **Report (PDF/Word)**:
   - A **well-structured report** detailing the entire process, from dataset selection to final model evaluation. This should include:
     A. **Introduction**:
       - Overview of the dataset (source, problem type: classification, regression, or clustering).
       - Justification for selecting this dataset.

     B. **Data Exploration**:
       - Descriptive statistics (summary statistics) of the dataset (e.g., mean, median, standard deviation, and range).
       - Visualizations (e.g., histograms, box plots, charts) to better understand the data.

     C. **Data Preprocessing**:
       - Data cleaning steps: handling missing values, outliers, duplicates, etc.
       - Feature engineering (e.g., encoding categorical variables, scaling/normalizing data).

D. **Modeling**:

- Description of the two machine learning algorithms you applied.
- Code snippets and explanation of how you trained the models.

E. **Results and Evaluation**:

- If you implemented any additional analyses or exploratory techniques (e.g., feature selection, cross-validation), describe them in the report.
- Performance metrics.
- Comparison of the results from the two algorithms.
- Interpretation of results.

F. **Conclusion**:

- Summary of your findings and recommendations.
- Suggestions for improving the model or potential next steps (e.g., tuning hyperparameters, trying other algorithms, etc.).

G. **References**

Grading Rubrics:

The project will be graded based on the following criteria:

Criteria	Description	Weight
Dataset Selection & Problem Understanding	Clear and appropriate selection of the dataset. Problem is clearly categorized as classification, regression, or clustering.	10%
Data Exploration & Preprocessing	Thorough exploratory analysis (descriptive statistics, visualizations) and well-documented data preprocessing steps (e.g., missing data handling, encoding).	20%
Modeling (Algorithm Selection & Application)	Selection of two relevant algorithms. Clear explanation of how the models were implemented, trained, and applied.	20%
Model Evaluation & Comparison	Use of appropriate evaluation metrics and clear comparison between the performance of the two models. Justification for the chosen metrics.	20%
Report Quality	Report is well-organized, clear, and concise. All sections are present with sufficient detail and clarity.	15%
Code Quality	Code is well-documented, clean, and structured. Reproducibility of analysis.	10%
Creativity & Additional Efforts	If any additional analysis or creative approaches were taken (e.g., feature engineering, hyperparameter tuning, model improvement).	5%

Additional Guidelines:

- **Dataset Selection**: Choose a dataset that fits into one of the following categories: classification, regression, or clustering. Ensure the dataset is of reasonable size and quality, and contains enough data for effective analysis.
  - Datasets must be sourced from **Kaggle** or other publicly available repositories.
  - Ensure that the dataset includes enough information for your chosen problem and allows you to apply two distinct machine learning algorithms.
- **Machine Learning Algorithms**: You are required to apply **two different machine learning algorithms**. You can choose from the following (but are not limited to):
  - **Classification**: Logistic Regression, Decision Trees, Random Forests, Support Vector Machine (SVM), k-NN, Naive Bayes.
  - **Regression**: Linear Regression, Decision Trees, Random Forest, Support Vector Regression, k-NN.
  - **Clustering**: k-Means, Hierarchical Clustering, DBSCAN.
- **Tools & Libraries**:
  - R must be used for data analysis and modeling.
- **Code Documentation**: Your code should be well-commented, explaining the purpose of each step. This will help reviewers understand your thought process and make the code easier to follow.
- **Submission**:
  - **Deadline**: Nov. 30, 2024
  - Submit the following on the course page in LMS:
    - R script or R Markdown file
    - Report (PDF or Word)
    - Any additional files (e.g., dataset if applicable, additional analysis files)

Plagiarism Policy:

- All work should be original. You are allowed to use publicly available code or libraries but should cite any external sources or references you use. Any form of plagiarism will result in penalties according to the academic integrity policy of the course/institution.