Assignment: Advanced Feature Engineering and Predictive Modeling

Objective

The aim of this assignment is to extend the work done in Assignment 1 by applying advanced feature engineering techniques and building predictive models using Linear Regression and Decision Tree algorithms. This assignment focuses on preparing a dataset for predictive tasks, engineering relevant features, and evaluating model performance.

Instructions

1. **Individual Work:**
   - This is an individual assignment.
   - You must complete the assignment independently. Collaboration is not permitted.
   - Clearly mention your name and roll number in your submission.
2. **Dataset Selection:**
   - Use the dataset cleaned and prepared in Assignment 1. If any additional changes are needed, document these changes clearly in your report.
   - Ensure the dataset includes at least one target variable (dependent variable) and relevant features (independent variables).
3. **Execution Platform:**
   - Complete the assignment in a Jupyter Notebook environment (e.g., Google Colab).
   - Ensure that all intermediate outputs and steps are clearly visible in the submitted notebook.
4. **Submission Requirements:**
   - Submit both the Jupyter Notebook file (.ipynb) and a well-formatted PDF with all code, outputs, and visualizations.
   - The PDF should include observations and justifications for each task, written in plain English.
   - Marks will be deducted for improperly formatted submissions or missing outputs.
5. **Plagiarism:**
   - Strictly avoid plagiarism. Any evidence of copied work will result in a penalty.

Tasks

1. Feature Engineering (3 Marks)

1.1 **Dimensionality Reduction (1 Mark)**

- Apply Principal Component Analysis (PCA) to reduce the dimensions of your dataset.
- Document and justify the number of components selected and explain their significance.

1.2 **Feature Selection (1 Mark)**

- Use one filter-based method (e.g., correlation-based) and one wrapper-based method (e.g., forward selection or backward elimination) to select features.
- Document the selected features and justify their inclusion in the predictive model.

1.3 **Handling Imbalanced Data (if applicable) (1 Mark)**

- If the dataset contains class imbalance, apply sampling methods (e.g., oversampling, undersampling) to address the issue.
- Document and justify the method used.

---

2. Predictive Modeling (3 Marks)

2.1 **Linear Regression (1.5 Marks)**

- Train a linear regression model using the processed dataset.
- Evaluate the model using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

2.2 **Decision Tree (1.5 Marks)**

- Train a decision tree model for regression or classification (depending on your dataset’s target variable).
- Visualize the tree structure and analyze the importance of features used in the model.

---

3. Model Evaluation (3 Marks)

3.1 **Comparison of Models (1.5 Marks)**

- Compare the performance of Linear Regression and Decision Tree models.

- Discuss the strengths and limitations of each model based on their results.

3.2 **Cross-Validation (1.5 Marks)**

- Perform k-fold cross-validation for both models.
- Report the mean and standard deviation of the evaluation metrics.

Expectations

- Provide clear observations and justifications for each step, supported by evidence from your analysis.
- Ensure that your notebook and PDF are well-structured, neat, and easy to follow.
- Highlight any challenges faced during the assignment and describe how you addressed them.

Grading Criterion

- **Feature Engineering**: 3 marks
  - Dimensionality Reduction: 1 mark
  - Feature Selection: 1 mark
  - Handling Imbalanced Data: 1 mark
- **Predictive Modeling**: 3 marks
  - Linear Regression: 1.5 marks
  - Decision Tree: 1.5 marks
- **Model Evaluation**: 3 marks
  - Comparison of Models: 1.5 marks
  - Cross-Validation: 1.5 marks
- **Submission Quality and Formatting**: 1 mark