AI ASSIGNMENT 4 — Machine Learning
Deadline : 11:59 PM, 2/12/2024

Total Marks : 50 Marks **Weightage: 5%**

Instructions:

1. Assignments are to be attempted individually.
2. Submit the assignment as a single zipped folder (**A4_⟨RollNumber⟩.zip**) containing a pdf file (**A4_⟨RollNumber_*Report*⟩.pdf**) for all the theory questions, graphs, and code analysis, and single ipynb file (**A4_⟨RollNumber_*Code*⟩.ipynb**)for programming questions as per the format provided in the Coding section of the assignment. Add visualizations should be included where relevant. Discussion and interpretations of results should be included after each section if asked.
3. Please read the instructions given in the questions carefully. In case of any ambiguity, post your queries on Google Classroom before the deadline. **No TA will be responsible for responding to the queries after this.**
4. All the TAs will strictly follow the rubric provided. **No requests will be entertained related to scoring strategy.**
5. **The use of generative tools (such as ChatGPT, Gemini, etc.) is strictly prohibited.** Failure to comply may result in severe consequences related to plagiarism.
6. **Extension and Penalty clause:**
   - Even a 1 minute late submission on google classroom will be considered as late. Please turn-in your submissions atleast 5 minutes before the deadline.
   - Not explaining the answers properly will lead to zero marks.

Theory (10 marks)

1. Consider the training set given below for determining whether a loan application should be approved or rejected. Draw the full decision tree obtained using entropy as the impurity measure. Show all steps and calculations clearly. Compute the training error of the decision tree.

Long-Term Debt	Unemployed	Credit Rating	Down Payment < 20%	Class
No	No	Good	Yes	Approve
No	No	Bad	No	Approve
No	No	Bad	Yes	Approve
No	No	Bad	No	Approve
Yes	No	Good	No	Approve
No	Yes	Good	Yes	Reject
Yes	No	Bad	No	Reject
Yes	No	Bad	Yes	Reject
Yes	No	Bad	Yes	Reject
Yes	Yes	Bad	No	Reject

Coding (40 marks)

The objective is to build a decision tree model using the provided real estate dataset, which involves price prediction. You'll need to preprocess the data, deal with data imbalance, train the model, optimize it, and evaluate its performance.

The dataset to be used for this question is provided. The dataset is already split into **train.csv** and **test.csv**. Use **train.csv** for training and validation and **test.csv** for testing your model.

NOTE: You can use **Python libraries** like NumPy, Pandas, Scikit-learn, Imbalanced-learn, Matplotlib, and Seaborn to perform the required tasks. Additionally, other libraries can be utilized if needed, and you are free to use inbuilt functions whenever applicable. Below are some links to help you gain a better understanding of Exploratory Data Analysis (EDA)

`https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/`

`https://www.geeksforgeeks.org/smote-for-imbalanced-classification-with-python/`

2. **Data Preprocessing and Exploratory Data Analysis** (15 Marks)

    1. **Task 1: Understanding the Dataset:** (2 Marks)

        (a) Provide a dataset overview, summarizing the unique values in each column. Perform a detailed statistical analysis on the numerical columns, including calculations for mean, standard deviation, minimum, maximum, and percentiles (25th, 50th, and 75th).

    2. **Task 2: Drop Irrelevant Columns:** (1 Mark)

        (a) Remove the columns identified through correlation analysis with a correlation coefficient within the range of -0.1 to 0.1, as well as those that lack predictive power and do not contribute meaningfully to the target variable. Provide reasons for dropping each of these columns.

3. **Task 3: Encoding Categorical Features :** (2 Marks)
   (a) Use the label encoding technique to transform categorical columns. Discuss the impact of high cardinality on certain categorical variables and how to mitigate them.
4. **Task 4: Feature Scaling:** (3 Marks)
   (a) Scale numerical data using Standard Scaler and analyze its impact on model performance after training, particularly focusing on whether scaling affects the performance of Decision Tree models.(Analysis can be done after training).
5. **Task 5: Target Variable Imbalance Detection:** (4 Marks)
   (a) Since this is a regression problem, first analyze the distribution of the target variable, 'Price', by plotting it in bins of size 10. After understanding the distribution, convert the target variable into categories by creating price brackets using fixed binning. Define four fixed price categories: 'Low', 'Medium', 'High', and 'Very High', based on specified price ranges. Analyze the distribution of properties across these categories and visualize it using histograms or bar charts. Finally, discuss the level of imbalance across the different price brackets.
6. **Task 6: Handling Imbalanced Data:** (3 Marks)
   (a) Use random undersampling and random oversampling techniques to address data imbalance. Explain the benefits and limitations of each method.

3. **Building Decision Tree Model** (15 Marks)

1. **Task 1: Model Training:** (3 Marks)
   (a) Train a Decision Tree Regressor using the training data.
   (b) Visualize the decision tree and explain the model structure, including the depth and splitting decisions.(You can use plot_tree function from sklearn.tree module)
2. **Task 2: Feature Importance and Hyperparameter Tuning:** (4 Marks)
   (a) Extract and plot the feature importances from the trained decision tree model.
   (b) Discuss why certain features are more important than others and whether it matches your expectations.
   (c) Perform any method for hyperparameter optimization (e.g., Grid Search or Randomized Search) to find the best hyperparameters for the decision tree. The focus should be on:
      - max_depth
      - min_samples_split
      - min_samples_leaf
      - max_features

      Compare the performance of the tuned model with the default one.

3. **Task 3: Pruning Decison Tree:** (4 Marks)
   (a) Prune the decision tree using pre-pruning/post-pruning techniques like minimal cost-complexity pruning.
   (b) Visualize and discuss the difference between the pruned and unpruned trees.
4. **Task 4: Handling Overfitting:** (4 Marks)
   (a) Use cross-validation to assess model generalization and detect overfitting.
   (b) Implement learning curves and evaluate overfitting by comparing training and validation errors.
   (c) Discuss the role of cross-validation in controlling overfitting for Decision Trees.

4. **Model Evaluation and Error Analysis** (10 Marks)

1. **Task 1: Model Evaluation:** (4 Marks)
   (a) Evaluate the model(use the tuned model with best parameters from previous question) on test data using appropriate regression metrics:
   - Mean Squared Error (MSE)
   - Mean Absolute Error (MAE)
   - R-squared (R^2)

   (b) Report and interpret the model's performance on both the training and test datasets.
2. **Task 2 : Residual and Error Analysis:** (4 Marks)
   (a) Analyze the residuals (difference between predicted and actual prices).
   (b) Visualize the residuals to check for patterns. Are there groups of data where the model consistently underperforms? If yes, then propose possible improvements for the model based on the error analysis.
3. **Task 3 : Feature Importance based analysis:** (2 Marks)
   (a) Analyze how top 3 important features affect the target variable Prices individually. Calculate the RMSE.

5. **Bonus Challenge** (6 Marks)

1. **Task 1:Advanced Imbalance Handling** (3 Marks)
   (a) Experiment with advanced data balancing techniques like ADASYN (Adaptive Synthetic Sampling). Compare it with SMOTE and discuss its effectiveness in handling imbalanced data.
2. **Task 2:Ensemble Learning: Random Forest** (3 Marks)
   (a) Train a Random Forest Regressor on the same dataset.Compare the performance of the Random Forest model with your Decision Tree model. Discuss the tradeoffs between using a single decision tree versus an ensemble of trees in Random Forest.