Assignment No: - 7

Decision Tree

Problem Statement: -

Assignment on Classification technique

Every year many students give the GRE exam to get admission in foreign Universities. The data set contains GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose strength (out of 5), Letter of Recommendation strength (out of 5), Undergraduate GPA (out of 10), Research Experience (0=no, 1=yes), Admitted (0=no, 1=yes). Admitted is the target variable.

Data Set: Graduate Admission 2 (kaggle.com)

The counselor of the firm is supposed to check whether the student will get an admission or not based on his/her GRE score and Academic Score. So to help the counselor to take appropriate decisions, build a machine learning model classifier using a Decision tree to predict whether a student will get admission or not.

a) Apply Data pre-processing (Label Encoding, Data Transformation....) techniques if necessary.

b) Perform data-preparation (Train-Test Split)

c) Apply Machine Learning Algorithm

d) Evaluate Model.

Objective: The objective of this project is to develop a Decision Tree classifier to assist counselors in predicting student admissions to foreign universities based on their GRE scores and academic performance. By leveraging machine learning techniques, we aim to streamline the admissions process, providing counselors with a reliable tool to make informed decisions efficiently.

S/W Packages and H/W apparatus used: OS: Windows, Kernel: Python 3, Tools: Google Colab

Libraries and packages used: NumPy, Pandas, Matplotlib, Scikit-Learn

Theory:

- **Classification**: Classification is the process of organizing a dataset into classes or categories. This can be applied to both structured and unstructured data, with the goal of predicting the class of given data points based on their features.

- **Decision Tree:** It uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

  Root Nodes – It is the node present at the beginning of a decision tree. From this node the population starts dividing according to various features.

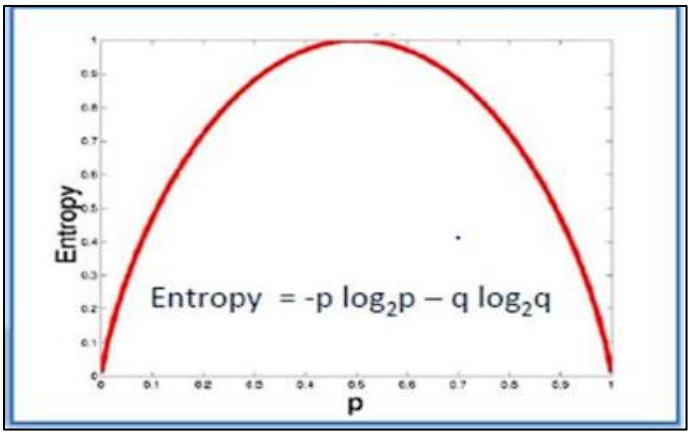
  Decision Nodes – the nodes we get after splitting the root nodes are called Decision Node.

  Leaf Nodes – the nodes where further splitting is not possible are called leaf nodes or terminal nodes

  Sub-tree – just like a small portion of a graph is called sub-graph similarly a subsection of this the decision tree is called a sub-tree.

  Pruning – It is cutting down some nodes to stop overfitting

- **Entropy**: Entropy is a measure of the randomness or impurity in a dataset. In the context of Decision Trees, entropy is used to quantify the homogeneity of a sample. A sample with low entropy is more homogeneous, while a sample with high entropy is more diverse.



- **Information Gain:** The information gain is based on the decrease in entropy after a dataset is split on an attribute.

- **Constructing a Decision Tree:**

1. Calculate the entropy of the target variable.

2. Split the dataset based on different attributes and calculate the entropy for each branch.

3. Calculate the information gain for each attribute and select the attribute with the highest information gain as the decision node.

4. Continue this process recursively until all data is classified, with branches either becoming leaf nodes or further split.

- **Pruning:** Pruning is a technique used to prevent overfitting in Decision Trees by removing nodes or branches that are not significant. It helps improve the performance of the tree by eliminating unnecessary complexity. Pruning can be done during tree construction (pre-pruning) or after the tree is built (post-pruning).

  We will utilize these principles of Decision Trees, including entropy calculation, information gain, tree construction, pruning, and transformation to decision rules, to develop a predictive model for determining student admissions to foreign universities based on GRE scores and academic performance.

Advantages and Applications & Limitation/Example:

Advantages:

1. **Interpretability:** Decision trees are highly interpretable models, making them suitable for explaining the reasoning behind predictions to stakeholders and domain experts. The decision rules learned by the model can be easily understood and visualized, providing insights into the factors influencing the outcome.

2. **Handling Both Numerical and Categorical Data:** Decision trees can handle both numerical and categorical features without requiring extensive data preprocessing. This versatility simplifies the data preparation process and makes decision trees applicable to a wide range of datasets without the need for feature engineering.

3. **Robustness to Outliers and Missing Values:** Decision trees are robust to outliers and missing values in the data. They partition the feature space based on thresholds, minimizing the impact of outliers on the overall model performance. Additionally, missing values can be handled naturally during the tree construction process.

4. **Feature Importance:** Decision trees provide a measure of feature importance, indicating the extent to which each feature contributes to the model's predictive performance. This information can be valuable for feature selection, identifying the most relevant predictors, and gaining insights into the underlying data generating process.

Applications:

1. **Classification and Regression Tasks:** Decision trees find applications in both classification and regression tasks across various domains, including finance, healthcare, marketing, and education. They can predict categorical outcomes (e.g., admission or rejection) as well as continuous variables (e.g., sales revenue or house prices).

2. **Customer Segmentation:** Decision trees are used for customer segmentation, where they partition customers into distinct groups based on demographic, behavioral, or transactional attributes. This segmentation enables targeted marketing campaigns, personalized recommendations, and customer retention strategies.

3. **Credit Risk Assessment:** In finance, decision trees are employed for credit risk assessment, where they predict the likelihood of default or delinquency based on borrower characteristics such as credit score, income, and debt-to-income ratio. This helps financial institutions make informed lending decisions and manage risk effectively.

4. **Medical Diagnosis:** Decision trees are utilized in healthcare for medical diagnosis and prognosis, where they analyze patient data (e.g., symptoms, lab results) to predict diseases, recommend treatments, or assess patient outcomes. Decision trees provide interpretable models that can aid physicians in clinical decision-making.

5. **Product Recommendation Systems:** Decision trees play a role in product recommendation systems, where they analyze user preferences and behavior to suggest relevant products or services. By understanding the features driving user preferences, decision trees enable personalized recommendations that enhance user experience and drive sales.

Limitations:

1. **Overfitting:** Decision trees are prone to overfitting, especially when the tree depth is not appropriately controlled or when the dataset is imbalanced. For example, if the Decision Tree algorithm creates a deep tree with many branches and leaf nodes to fit the training

data perfectly, it may not generalize well to unseen data, leading to poor performance on the test set.

2. **High Variance:** Decision trees are sensitive to small variations in the training data, leading to high variance in the model predictions. As a result, decision trees may produce unstable results when trained on different subsets of the data or with slight changes in the input features.

3. **Bias Towards Features with Many Levels:** Decision trees tend to favor features with many levels or categories during the tree construction process. Features with a large number of levels may be overrepresented in the decision rules, potentially leading to biased predictions or suboptimal model performance.

4. **Inability to Capture Complex Relationships:** Decision trees may struggle to capture complex relationships or interactions between features in the data. For example, if the relationship between the target variable and predictors is nonlinear or involves higher-order interactions, a single decision tree may not be able to model it accurately.

Working / Algorithm:

Step 1: Initialization

- Select a decision tree algorithm.
- Instantiate the decision tree classifier with specified parameters.

Step 2: Model Training

- Train the decision tree classifier using the training dataset (x_train, y_train).
- The algorithm recursively partitions the feature space based on the target variable to create a tree structure.

Step 3: Prediction

- For each instance in the testing dataset (x_test):
- Traverse the decision tree by following the learned rules.
- Determine the predicted class based on the final leaf node reached.

Step 4: Evaluation

- Compare the predicted labels with the true labels from the testing dataset to assess model performance.

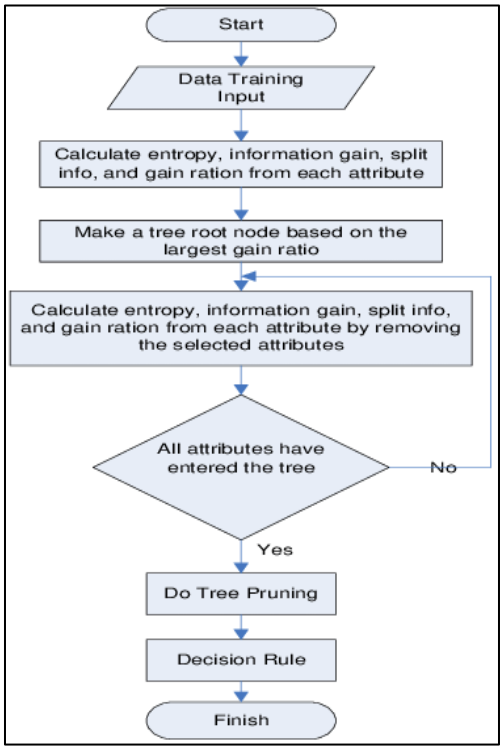
- Calculate evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

Step 5: Interpretation

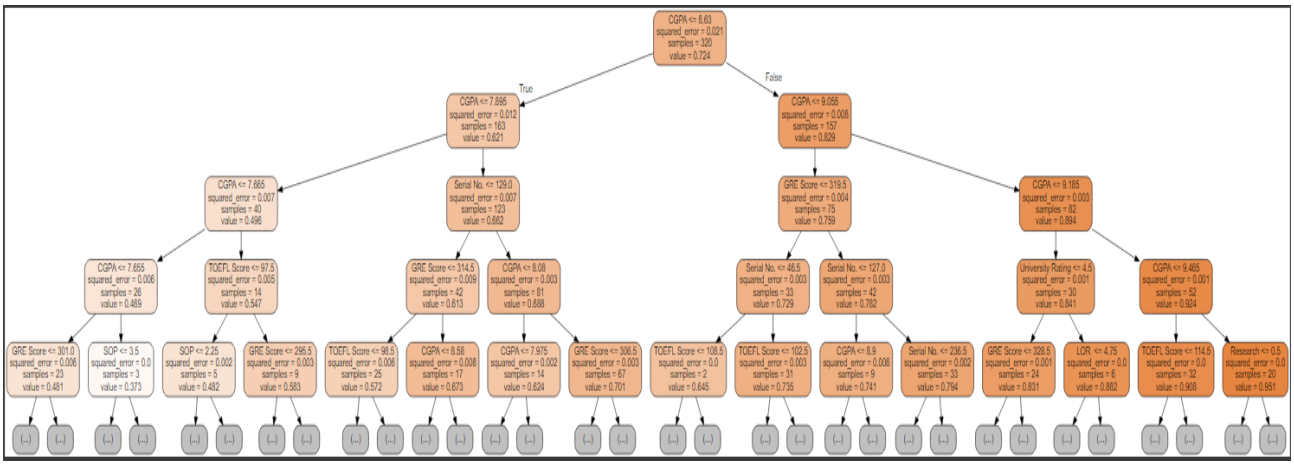
- Visualize the decision tree graphically to understand the rules learned by the model.
- Analyze feature importance to identify the most influential features in decision-making.

Step 8: The model is ready.

Diagram:



Max depth: 4



Conclusion:

In conclusion, the Decision Tree classifier for predicting student admissions to foreign universities based on GRE scores and academic performance offers a valuable solution for efficient decision-making. Its interpretability, simplicity, and versatility make it an effective tool for counselors. Through model evaluation, we gained insights into admission factors, guiding future improvements. The model contributes to transparent and equitable admission practices. Moving forward, ongoing monitoring, maintenance, and updates will ensure its continued relevance and effectiveness. The successful development and deployment of the Decision Tree classifier represents a significant advancement in improving admission procedures, leading to better outcomes for both students and educational institutions.