Assignment No:- 7

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: https://www.kaggle.com/mohansacharya/graduate-admissions
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.

Software Library Package:

For this task, we'll use the following Python libraries:

- `pandas` for data manipulation.
- `scikit-learn` for machine learning algorithms (specifically, `DecisionTreeClassifier`).
- `numpy` for mathematical operations.

Theory:

i) Methodology:

- Decision Tree Classifier: Decision trees are a popular supervised learning method used for classification tasks. They partition the feature space into regions and assign a class label to each region based on majority voting. Decision trees are constructed recursively by selecting the best feature to split the data at each node based on a criterion (e.g., Gini impurity or entropy).
- Advantages: Easy to interpret and understand, can handle both numerical and categorical data, requires little data preprocessing.
- Applications: Used in various domains such as finance, healthcare, marketing, etc., for classification tasks such as customer segmentation, fraud detection, etc.
- Limitations/Example: Prone to overfitting, sensitive to noisy data and outliers, may not capture complex relationships in the data. For example, a decision tree

may create overly complex decision boundaries if the data is highly dimensional or contains many features.

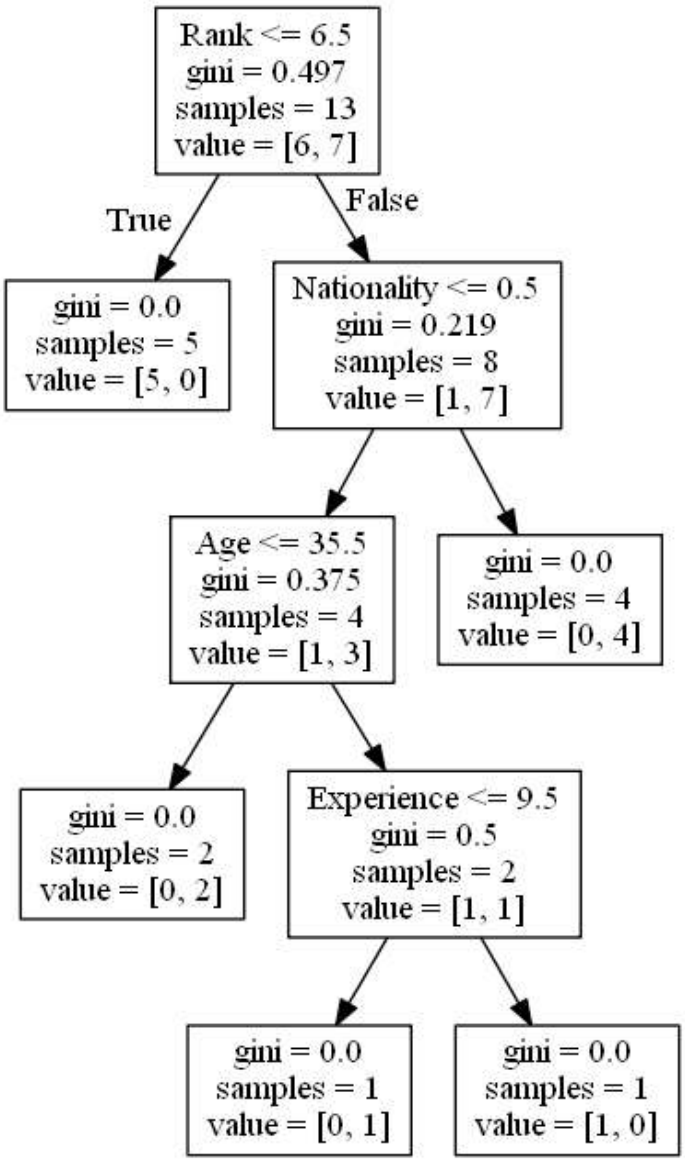

Fig 1: Decision Tree Diagram

ii) Advantages and Applications and Limitations/Example:

- Advantages: Decision trees are easy to interpret and understand, handle both numerical and categorical data, and require little data preprocessing. They can capture non-linear relationships between features and the target variable.
- Applications: Used in various domains such as finance, healthcare, marketing, etc., for classification tasks such as customer segmentation, fraud detection, etc.
- Limitations/Example: Prone to overfitting, sensitive to noisy data and outliers, may not generalize well to unseen data. For example, a decision tree with deep

branching may create overly complex decision boundaries, leading to poor generalization performance.

Working/Algorithm:

- Load the dataset into Python using `pandas`.
- Perform data preprocessing if necessary, such as label encoding for categorical variables.
- Split the data into training and testing sets using train-test split.
- Train a decision tree classifier model using the training data.
- Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, and F1-score.
- Visualize the decision tree if desired for better interpretation.

Diagram:

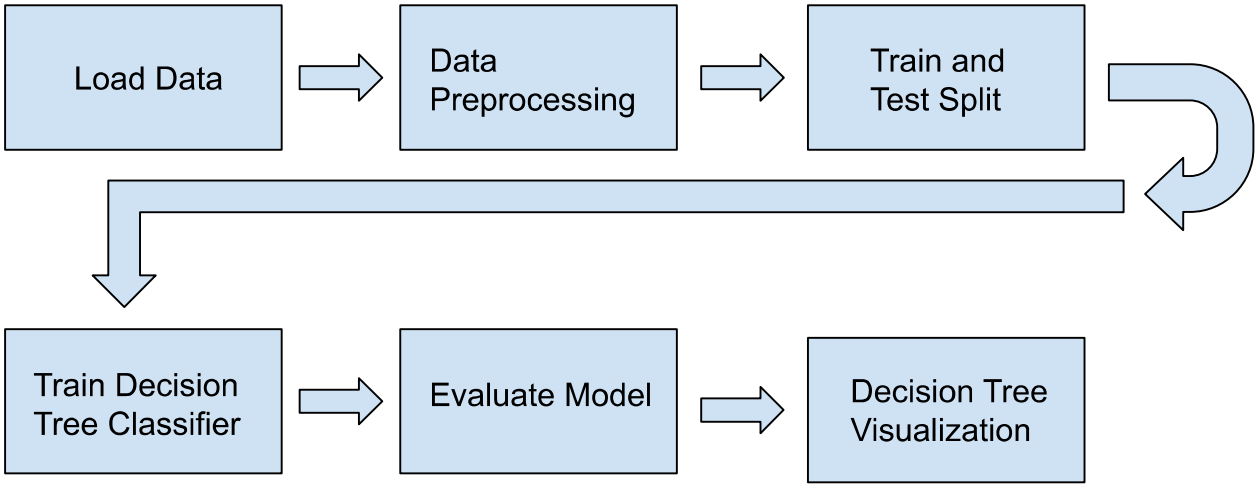

Fig 2: Workflow Diagram

Conclusion:

Decision trees are a powerful and interpretable machine learning algorithm for classification tasks. By constructing a tree-like model of decisions based on features, they provide insights into the decision-making process. However, it's important to handle overfitting and interpretability issues carefully, especially with deep decision trees. Overall, decision trees can be valuable tools for predicting admission outcomes in foreign universities based on student scores.