**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

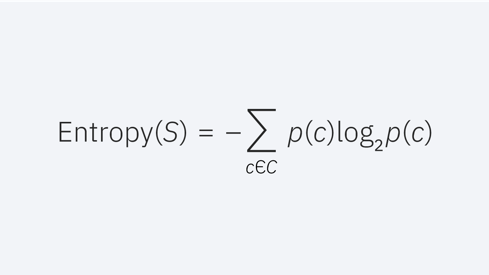
d) Evaluate Model.

**Theory**:-

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

**Entropy and Information Gain:**

Entropy is a concept that stems from information theory, which measures the impurity of the sample values. It is defined with by the following formula, where:

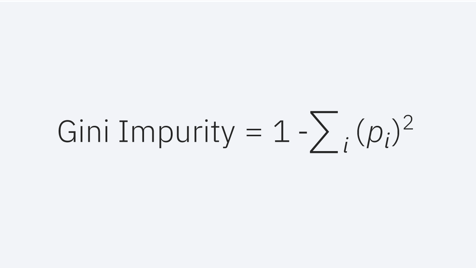


* S represents the data set that entropy is calculated
* c represents the classes in set, S
* p(c) represents the proportion of data points that belong to class c to the number of total data points in set, S

Entropy values can fall between 0 and 1. If all samples in data set, S, belong to one class, then entropy will equal zero. If half of the samples are classified as one class and the other half are in another class, entropy will be at its highest at 1.

**Gini Impurity:**

Gini impurity is the probability of incorrectly classifying random data point in the dataset if it were labeled based on the class distribution of the dataset. Similar to entropy, if set, S, is pure—i.e. belonging to one class) then, its impurity is zero. This is denoted by the following formula:



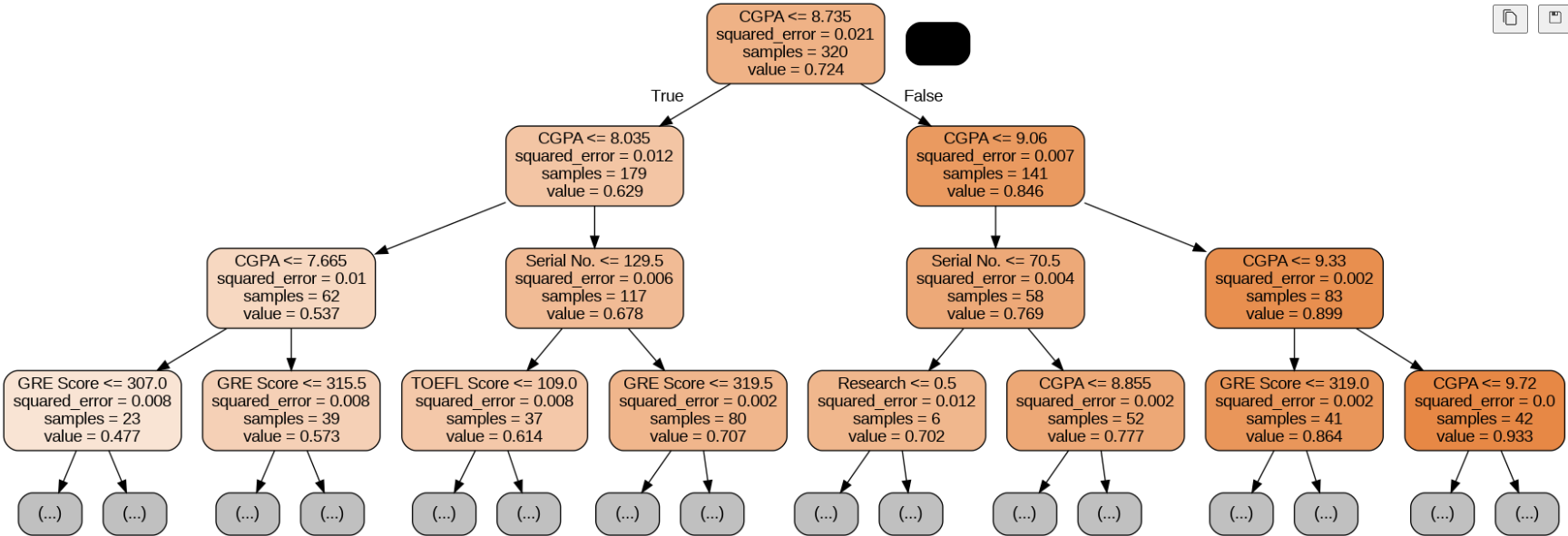
Advantages and Disadvantages & Limitation/Example:

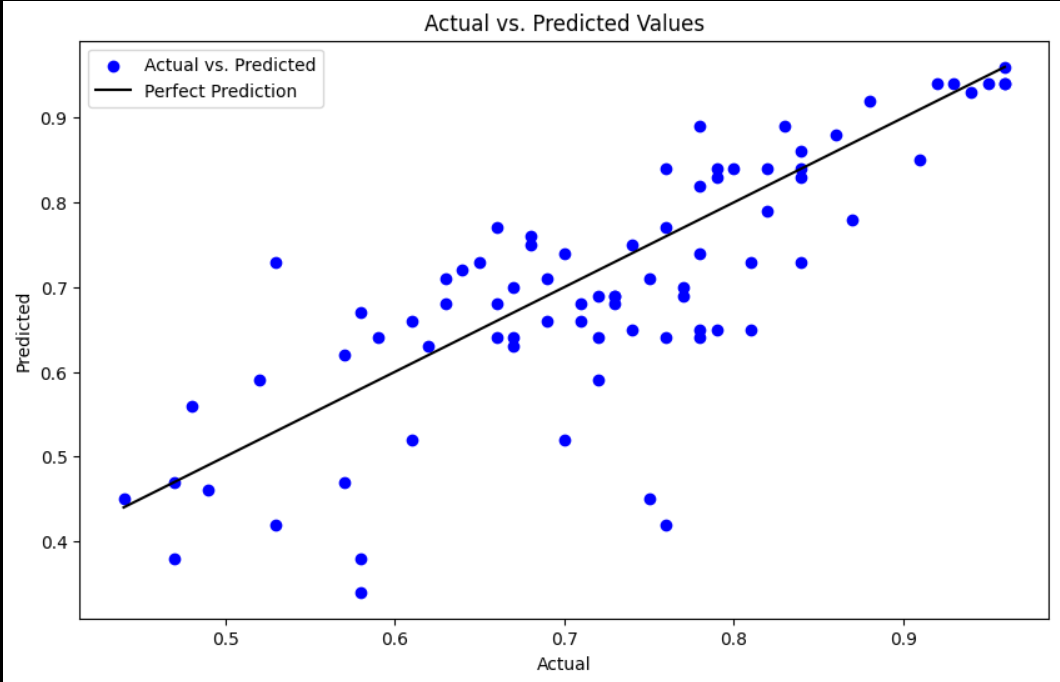
1. Advantages:
   * Interpretability: Decision trees are easy to interpret and understand, making them suitable for explaining the reasoning behind classification decisions.
   * Non-linear Relationships: Decision trees can capture non-linear relationships between features and the target variable.
   * Feature Importance: Decision trees provide information about feature importance, helping in feature selection and identifying influential variables.
2. Disadvantages & Limitations/Example:
   * Overfitting: Decision trees are prone to overfitting, especially when the tree depth is not properly controlled.
   * Instability: Decision trees are sensitive to small changes in the data, leading to different trees being generated for similar datasets.
   * Limited Expressiveness: Decision trees may not perform well on complex datasets with intricate relationships between features**.**

**Methodology:**

1. Data Pre-processing:
   * Perform label encoding for categorical variables (if any) to convert them into numerical format.
   * Check for missing values and handle them using techniques like imputation or deletion.
   * Scale or normalize the numerical features to ensure that all features contribute equally to the model.
2. Data Preparation (Train-Test Split):
   * Split the dataset into training and testing sets to train the model on a subset of the data and evaluate its performance on unseen data.
   * Typically, around 70-80% of the data is used for training, and the remaining 20-30% is used for testing.
3. Applying Machine Learning Algorithm (Decision Tree):
   * Use a suitable machine learning algorithm such as Decision Tree classifier to train the model on the training data.
   * Decision trees are a popular choice for classification tasks due to their simplicity and interpretability.
4. Model Evaluation:
   * Evaluate the performance of the trained model using evaluation metrics such as accuracy, precision, recall, and F1-score.
   * Plot the confusion matrix to visualize the performance of the model in terms of true positive, false positive, true negative, and false negative predictions.

**Diagram**

****

****

**Conclusion**  
In conclusion, employing decision tree classification on the admission dataset offers a straightforward approach to predicting students' admission likelihood based on GRE and academic scores. Despite its simplicity and interpretability, decision trees may suffer from overfitting and sensitivity to data variations. However, they provide valuable insights into feature importance and decision-making processes, aiding counselors in making informed admission decisions.