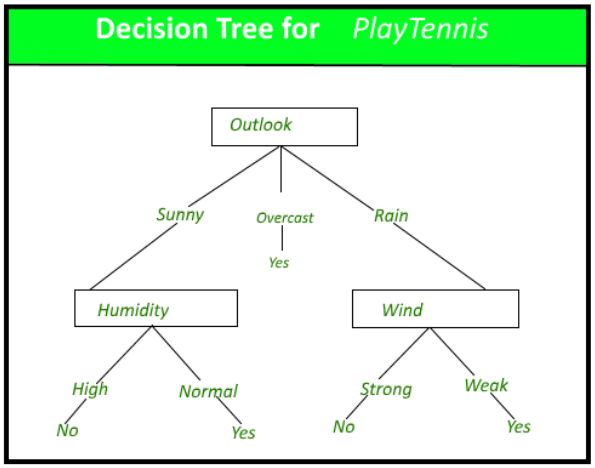
Experiment No. 3
Apply Decision Tree Algorithm on Adult Census Income Dataset and analyze the performance of the model
Date of Performance:
Date of Submission:

Aim: Apply Decision Tree Algorithm on Adult Census Income Dataset and analyze the performance of the model.

Objective: To perform various feature engineering tasks, apply Decision Tree Algorithm on the given dataset and maximize the accuracy, Precision, Recall, F1 score. Improve the performance by performing different data engineering and feature engineering tasks.

Theory:

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



Dataset:

Predict whether income exceeds $50K/yr based on census data. Also known as "Adult" dataset.

Attribute Information:

Listing of attributes:

>50K, <=50K.

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic,

Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Code:

Conclusion:

1. Discuss about the how categorical attributes have been dealt with during data pre-processing.

   While data preprocessing it is an important task to handle the categorical & numerical attributes for building accurate models. There are several preprocessing techniques-

   a. Label Encoding: The attributes like education, occupation, marital-status, relationship, sex etc were encoded into numerical values. Label encoding assigns a unique integer to each category, which allows Decision tree algorithm to work with these attributes.

   b. Missing values: Categorical values having missing values were imputed, either by median imputation or mode imputation for ordinal values.

   c. One Hot Encoder: The attributes with nominal categories like native-country, occupation, one hot encoder is applied to convert the categorical values into binary vectors & each category is represented as a binary column.

2. Discuss the hyper-parameter tunning done based on the decision tree obtained.

   The hyper-parameter tuning helps to optimize the Decision Tree model's performance which ensures that the model is not underfitting or overfitting the data.

   A systematic grid search method was applied to explore different combinations of parameters such as maximum tree depth, splitting criteria. This process aims to find the best hyper-parameter that maximize the model's performance.

3. Comment on the accuracy, confusion matrix, precision, recall and F1 score obtained.

   Accuracy: The accuracy of the model was calculated by taking the ratio of correctly predicted results to total number of instances. It describes the overall performance of

the model. The accuracy for this model is around 85% which shows that we got 84% correctly predicted results.

Confusion Matrix: The confusion matrix provides a detailed breakdown of the model's performance, showing true positives, true negatives, false positives, and false negatives. The confusion matrix for this model is TP=1031, TN=6564, FP=303, FN=1151.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance. The f1 score for this model is in between precision & recall.