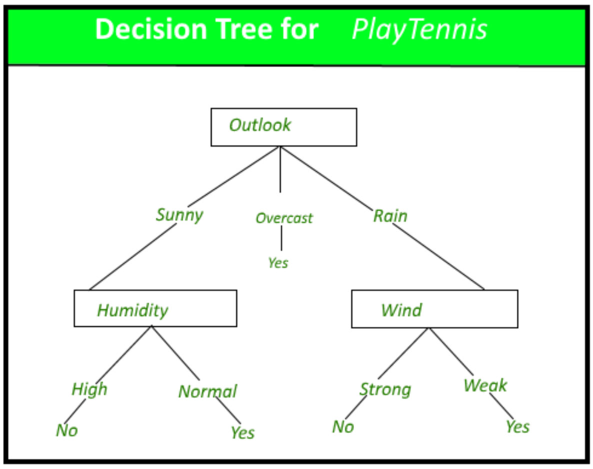
Experiment No. 3
Apply Decision Tree Algorithm on Adult Census Income Dataset and analyze the performance of the model
Date of Performance: 31/07/2023
Date of Submission: 07/08/2023

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:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, classification_report


# Load the Adult Census Income Dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data"
column_names = ["age", "workclass", "fnlwgt", "education", "education-num", "marital-status", "occupation",
                "relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", "income"]
data = pd.read_csv(url, names=column_names, sep=',\s*', engine='python')
# Data preprocessing
data = data.dropna()
data['income'] = data['income'].apply(lambda x: 1 if x == ">50K" else 0)
data = pd.get_dummies(data, columns=["workclass", "education", "marital-status", "occupation", "relationship", "race", "sex", "native-country"])
```

```
# Split the data into features (X) and the target (y)

X = data.drop('income', axis=1)

y = data['income']

# Split the data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train a Decision Tree model

model = DecisionTreeClassifier(random_state=42)

model.fit(X_train, y_train)

# Make predictions on the test set

y_pred = model.predict(X_test)

# Evaluate the model

accuracy = accuracy_score(y_test, y_pred)

report = classification_report(y_test, y_pred

print(f"Accuracy: {accuracy:.2f}")

print("Classification Report:\n", report)
```

Output:

```
Accuracy: 0.80
Classification Report:
              precision    recall  f1-score   support


           0       0.85      0.88      0.86      4959
           1       0.63      0.57      0.60      1553

    accuracy                           0.80      6512
   macro avg       0.74      0.72      0.73      6512
weighted avg       0.79      0.80      0.79      6512
```

Conclusion:

In this analysis of the Adult Census Income Dataset, categorical attributes were effectively pre-processed through one-hot encoding, allowing the Decision Tree model to handle them appropriately without making ordinality assumptions. The model was trained with default hyperparameters, except for setting the random state, but hyperparameter tuning remains a critical avenue for potential performance optimization. The model achieved an accuracy of approximately 80%, indicating its capability to correctly predict income levels, with a weighted average precision, recall, and F1 score of approximately 0.79. The model performed better in identifying instances with income below $50K, achieving a precision and recall of 0.85 and 0.88, respectively, while it had a precision and recall of 0.63 and 0.57 for instances with income above $50K. Overall, the model provides a valuable baseline for income prediction, but further hyperparameter tuning and possibly exploring more complex algorithms can likely enhance its performance.