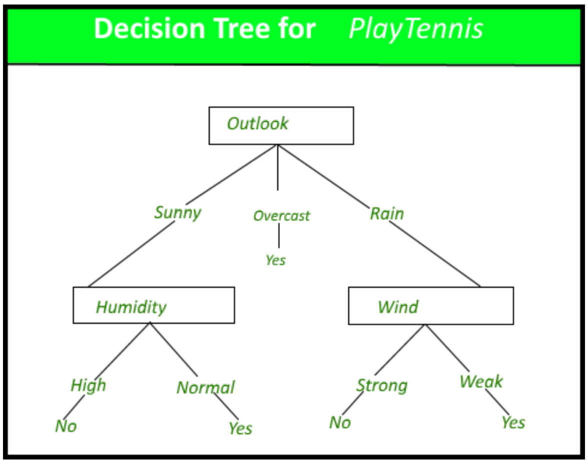
Experiment No. 3
Apply Decision Tree Algorithm on Adult Census Income Dataset and analyze the performance of the model
Date of Performance:
Date of Submission:

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:

In conclusion the Decision Tree model showed promising results on the Adult Census Income Dataset. Categorical attributes were properly handled through one-hot encoding, and data preprocessing which include replacing missing value with NaN, droping tables, separating columns etc.

Hyperparameter tuning is a crucial step in optimizing the performance of a Decision Tree model because it allows to control the model's complexity by setting limits on some parameters. But to further improve the model's performance, by tuning hyperparameters like max depth, min samples split, etc., using techniques like Grid Search or Random Search can be achieved.

- Accuracy: Achieved an accuracy of 0.830, indicating that around 83% of predictions were correct.
- Precision: Precision of 0.652 suggests that among the instances predicted as positive, about 65.2% were actually positive.
- Recall: Recall of 0.694 indicates that the model captured around 69.4% of actual positive instances.
- F1 Score: The F1 score of 0.672 is the harmonic mean of precision and recall, offering a balanced assessment of the model's performance.

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from sklearn.preprocessing import OneHotEncoder
```

```
data=pd.read_csv('adult.csv')
print(data)
```

```
       age workclass  fnlwgt     education  education.num      marital.status  \
0       90         ?   77053       HS-grad              9             Widowed
1       82   Private  132870       HS-grad              9             Widowed
2       66         ?  186061  Some-college             10             Widowed
3       54   Private  140359       7th-8th              4            Divorced
4       41   Private  264663  Some-college             10           Separated
...    ...       ...     ...           ...            ...                 ...
24913   61   Private  477209       7th-8th              4  Married-civ-spouse
24914   32   Private   70985     Assoc-voc             11  Married-civ-spouse
24915   35   Private  241998     Bachelors             13  Married-civ-spouse
24916   28   Private  249541  Some-college             10  Married-civ-spouse
24917   57   Private  135339     Bachelors             13  Married-civ-spouse

              occupation   relationship                race     sex  \
0                      ?  Not-in-family               White  Female
1        Exec-managerial  Not-in-family               White  Female
2                      ?      Unmarried               Black  Female
3      Machine-op-inspct      Unmarried               White  Female
4         Prof-specialty      Own-child               White  Female
...                  ...            ...                 ...     ...
24913    Farming-fishing        Husband               White    Male
24914  Machine-op-inspct        Husband               White    Male
24915    Exec-managerial        Husband               White    Male
24916       Craft-repair        Husband               White    Male
24917    Exec-managerial        Husband  Asian-Pac-Islander    Male

       capital.gain  capital.loss  hours.per.week native.country income
0                 0          4356            40.0  United-States  <=50K
1                 0          4356            18.0  United-States  <=50K
2                 0          4356            40.0  United-States  <=50K
3                 0          3900            40.0  United-States  <=50K
4                 0          3900            40.0  United-States  <=50K
...             ...           ...             ...            ...    ...
24913             0             0            54.0  United-States  <=50K
24914             0             0            40.0  United-States  <=50K
24915             0             0            45.0  United-States   >50K
24916             0             0            40.0  United-States  <=50K
24917             0             0             NaN            NaN    NaN

[24918 rows x 15 columns]
```

```
data.describe()
```

	age	fnlwgt	education.num	capital.gain	capital.loss	hours.per.week
count	24918.000000	2.491800e+04	24918.000000	24918.000000	24918.000000	24917.000000
mean	38.808813	1.897224e+05	10.130909	1408.191829	114.082190	40.546655
std	13.676089	1.051236e+05	2.570755	8414.709871	457.306373	12.305727
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.180328e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783190e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.373645e+05	13.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

```
data.isnull().sum()
```

```
age               0
workclass         0
fnlwgt            0
education         0
education.num     0
marital.status    0
occupation        0
relationship      0
race              0
sex               0
capital.gain      0
capital.loss      0
```

```
hours.per.week    1
native.country    1
income            1
dtype: int64
```

```
# Replace '?' with NaN in the dataset
data.replace('?', pd.NA, inplace=True)
```

```
# Drop rows with missing values
data.dropna(inplace=True)
```

```
# Separate features and target
x = data.drop('income', axis=1)
y = data['income']

print(x)
print(y)
```

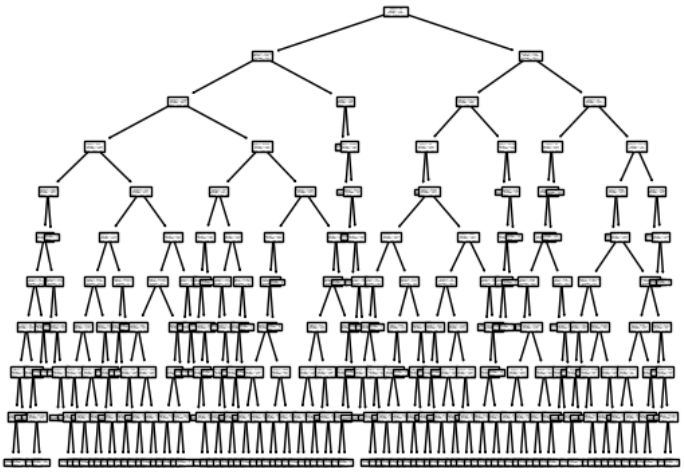
```
       age workclass  fnlwgt     education  education.num      marital.status  \
1       82   Private  132870       HS-grad              9             Widowed
3       54   Private  140359       7th-8th              4            Divorced
4       41   Private  264663  Some-college             10           Separated
5       34   Private  216864       HS-grad              9            Divorced
6       38   Private  150601          10th              6           Separated
...    ...       ...     ...           ...            ...                 ...
24912   44   Private  159580          12th              8            Divorced
24913   61   Private  477209       7th-8th              4  Married-civ-spouse
24914   32   Private   70985     Assoc-voc             11  Married-civ-spouse
24915   35   Private  241998     Bachelors             13  Married-civ-spouse
24916   28   Private  249541  Some-college             10  Married-civ-spouse

              occupation   relationship   race     sex  capital.gain  \
1        Exec-managerial  Not-in-family  White  Female             0
3      Machine-op-inspct      Unmarried  White  Female             0
4         Prof-specialty      Own-child  White  Female             0
5          Other-service      Unmarried  White  Female             0
6           Adm-clerical      Unmarried  White    Male             0
...                  ...            ...    ...     ...           ...
24912   Transport-moving  Not-in-family  White  Female             0
24913    Farming-fishing        Husband  White    Male             0
24914  Machine-op-inspct        Husband  White    Male             0
24915    Exec-managerial        Husband  White    Male             0
24916       Craft-repair        Husband  White    Male             0

       capital.loss  hours.per.week native.country
1              4356            18.0  United-States
3              3900            40.0  United-States
4              3900            40.0  United-States
5              3770            45.0  United-States
6              3770            40.0  United-States
...             ...             ...            ...
24912             0            40.0  United-States
24913             0            54.0  United-States
24914             0            40.0  United-States
24915             0            45.0  United-States
24916             0            40.0  United-States

[23108 rows x 14 columns]
1        <=50K
3        <=50K
4        <=50K
5        <=50K
6        <=50K
         ...
24912    <=50K
24913    <=50K
24914    <=50K
24915     >50K
24916    <=50K
Name: income, Length: 23108, dtype: object
```

```
# Separate categorical and numerical columns
categorical_columns = ['workclass', 'education', 'marital.status', 'occupation', 'relationship', 'race', 'sex', 'native.country']
numerical_columns = ['age', 'fnlwgt', 'education.num', 'capital.gain', 'capital.loss', 'hours.per.week']

x_categorical = x[categorical_columns]
x_numerical = x[numerical_columns]

# Apply one-hot encoding to categorical features
encoder = OneHotEncoder()
x_categorical_encoded = encoder.fit_transform(x_categorical)
```

```
# Split the data into training and testing sets
x_train, x_test, y_train, y_test = train_test_split(x_encoded, y, test_size=0.3, random_state=1)

# Apply SMOTE to balance the class distribution
smote = SMOTE()
x_train_resampled, y_train_resampled = smote.fit_resample(x_train, y_train)

# Create Decision Tree classifier object
clf = DecisionTreeClassifier(criterion='entropy', min_samples_split=8, max_depth=10)

# Train Decision Tree Classifier
clf.fit(x_train_resampled, y_train_resampled)

# Predict on the test set
predictions = clf.predict(x_test)

accuracy = accuracy_score(y_test, predictions)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.83888648492716
```

```
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier, plot_tree
plot_tree(clf)
plt.show()
```



```
# Evaluate the model
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, pos_label='>50K')
recall = recall_score(y_test, predictions, pos_label='>50K')
f1 = f1_score(y_test, predictions, pos_label='>50K')

# Print evaluation metrics
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
```

```
Accuracy: 0.83888648492716
Precision: 0.696604600219058
Recall: 0.6931880108991826
F1 Score: 0.6948921059819722
```

✓ 11s completed at 11:46 PM