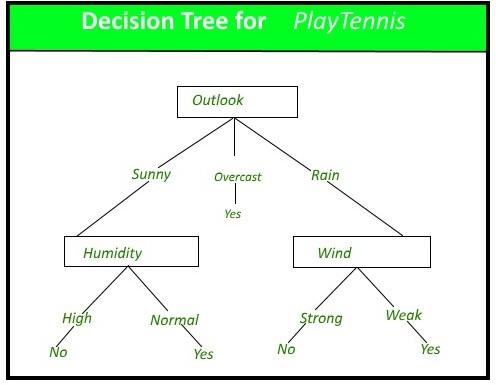
| Experiment No. 3 |
| --- |
| Apply Decision Tree Algorithm on Adult Census Income  Dataset and analyze the performance of the model |
| Date of Performance: 6/8/2024 |
| Date of Submission: 13/8/2024 |

**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.

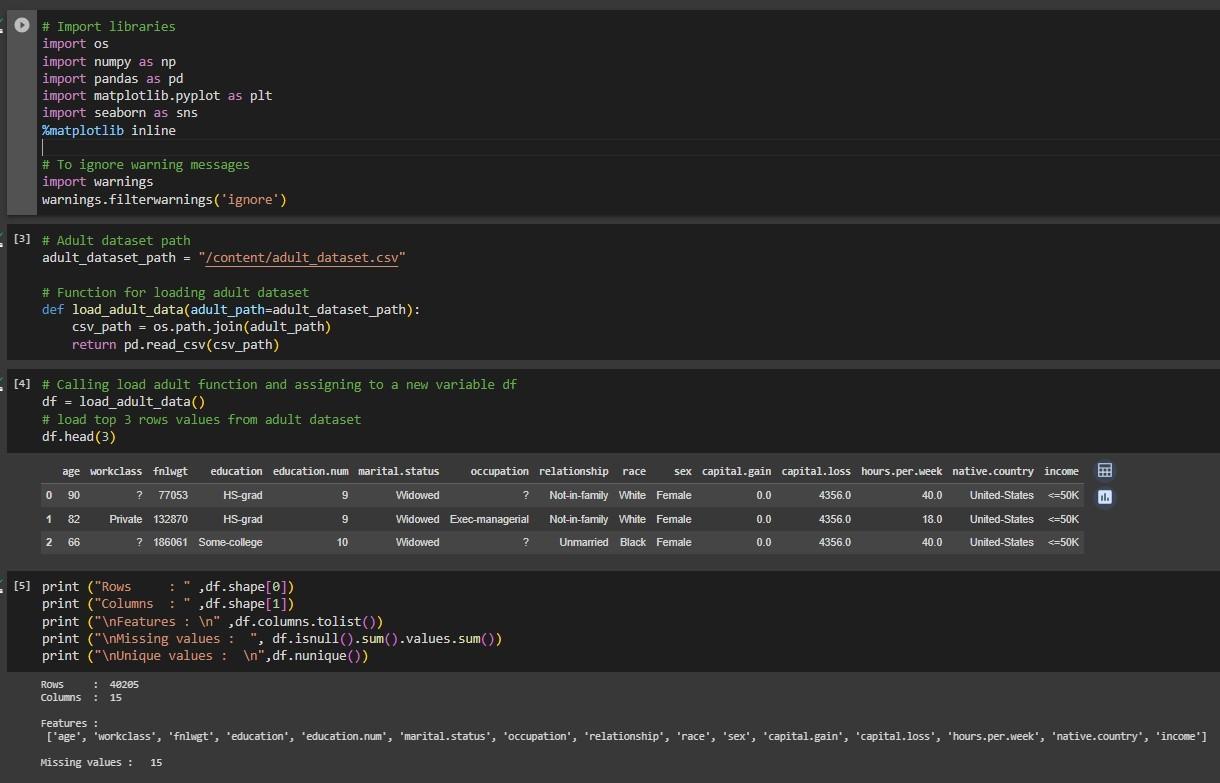
sex: Female, Male. capital-gain: continuous. capital-loss: continuous.

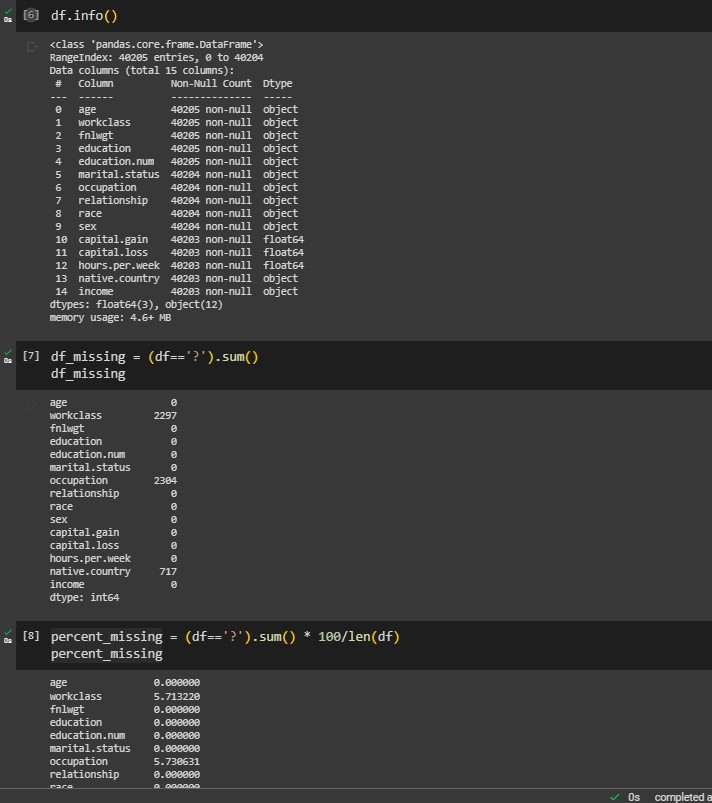
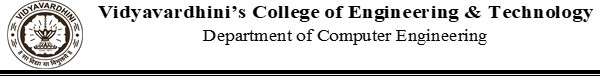
hours-per-week: continuous.

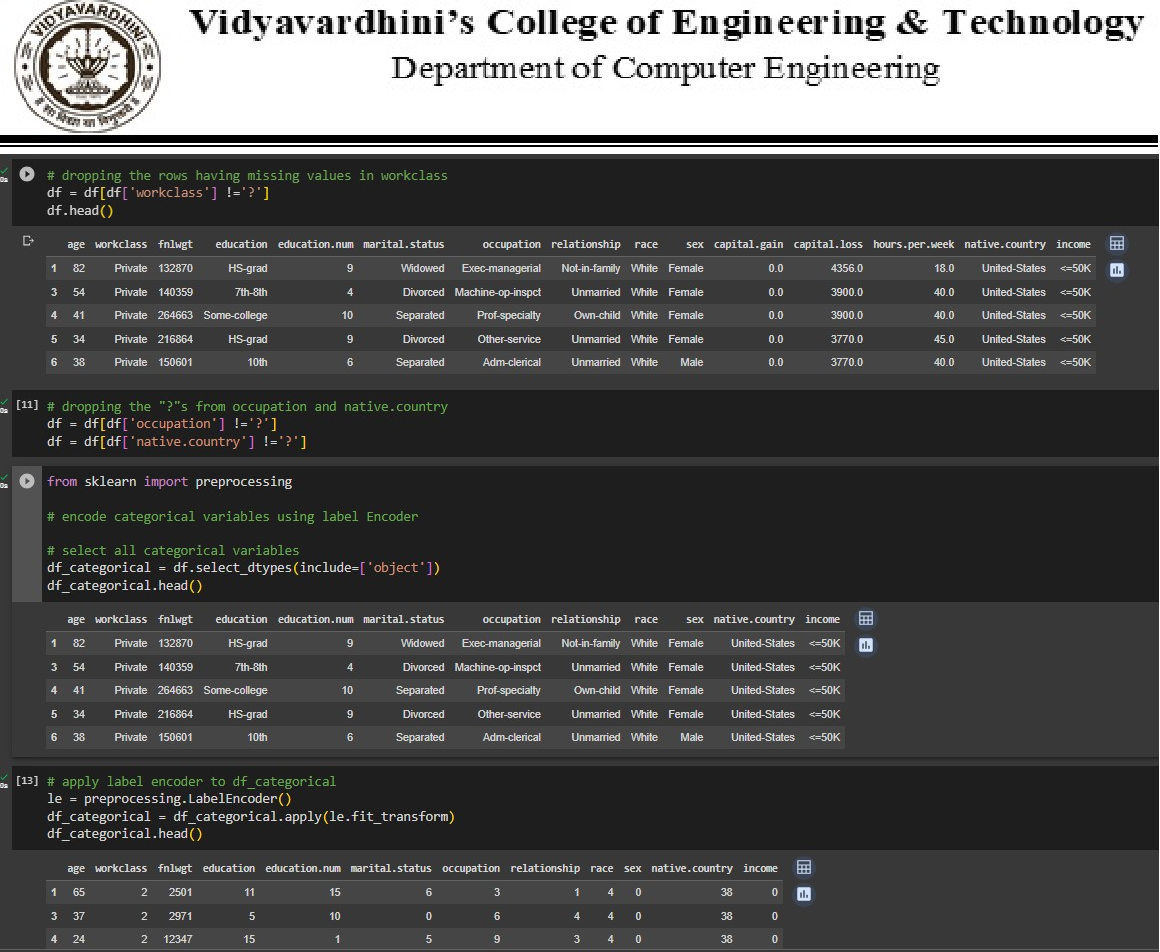
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican- Republic, Laos, Ecuador, Taiwan, Haiti, Columbia,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

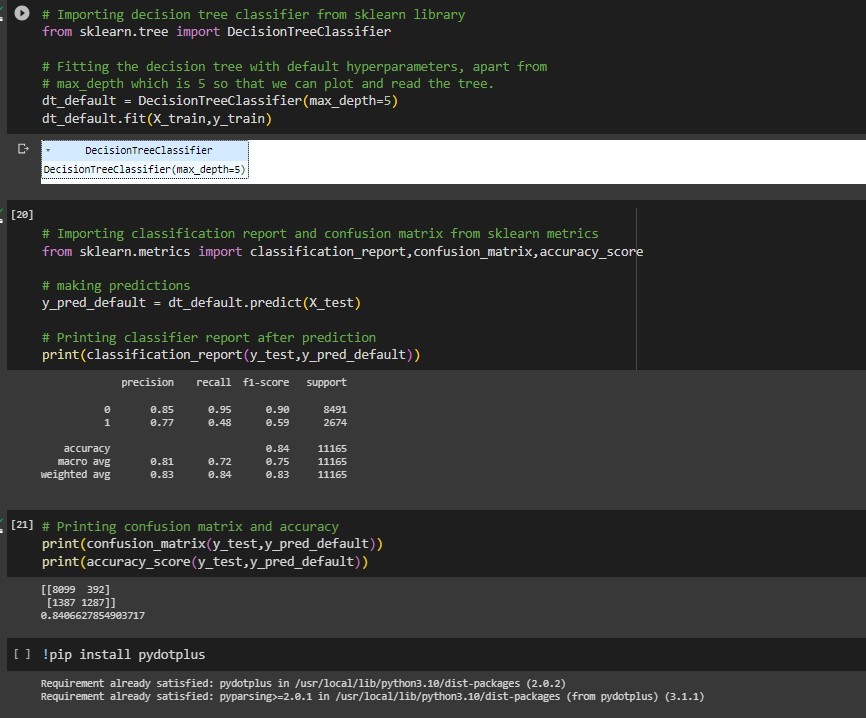
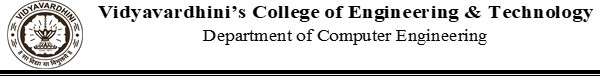
Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand- Netherlands.

# Code:









**Conclusion:**

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

=>Missing values with '?' are removed from specific columns (e.g., 'workclass,' 'occupation,' 'native.country').

Categorical attributes are label-encoded, converting them into numerical values. The target variable ('income') is converted to a categorical format.

The data is split into training and testing sets. A decision tree classifier is trained on the data

Model performance is evaluated using classification metrics like precision, recall, F1- score, and accuracy

**Conclusion:**

Discuss the hyper-parameter tunning done based on the decision tree obtained. Hyperparameter tuning is the process of finding the optimal values for the hyperparameters of a machine learning model to improve its performance.

In the provided code, only one hyperparameter, max\_depth, is manually set to 5. While this can be useful for creating a more interpretable tree, it may not necessarily result in the best predictive performance.

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

Accuracy measures the proportion of correctly classified instances out of all instances. In this case, the model has an accuracy of approximately 84.07%, indicating that it correctly predicts the income category for about 84.07% of the test data.

CONFUSION MATRIX:

The confusion matrix provides a more detailed view of the model's performance:

True Positives (TP): 1287 - The number of instances correctly classified as positive (income > 50K)

True Negatives (TN): 8099 - The number of instances correctly classified as negative (income <= 50K).

False Positives (FP): 392 - The number of instances incorrectly classified as positive (income > 50K).

False Negatives (FN): 1387 - The number of instances incorrectly classified as negative (income <= 50K).

PRECISION

In this case, precision for the positive class (income > 50K) can be calculated as 1287 / (1287 + 392), which is the ratio of correctly predicted high-income individuals to all predicted high-income individuals.

RECALL:

In this case, recall for the positive class is 1287 / (1287 + 1387), which is the ratio.