Experiment No. 4
Apply Random Forest Algorithm on Adult Census Income Dataset and analyze the performance of the model
Date of Performance:
Date of Submission:

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

Objective: Able to perform various feature engineering tasks, apply Random Forest Algorithm on the given dataset and maximize the accuracy, Precision, Recall, F1 score.

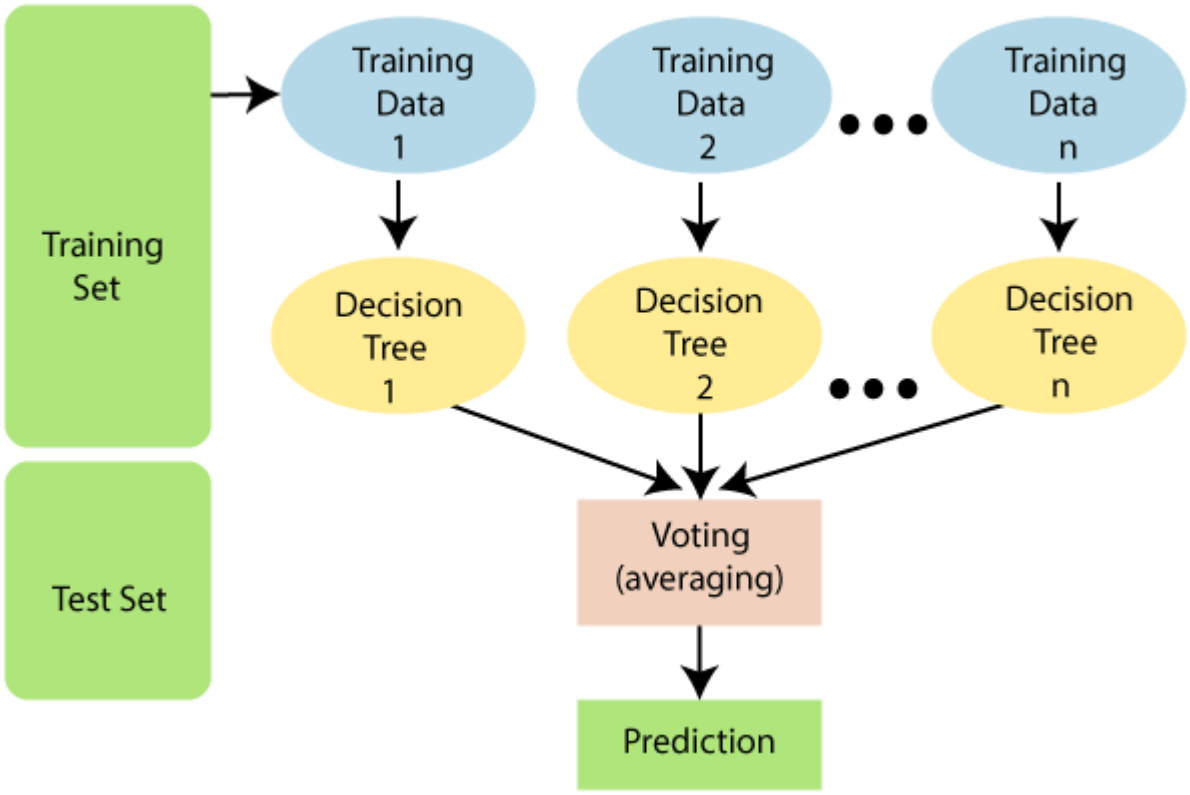
Theory:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:



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. State the observations about the data set from the correlation heat map.

   Positive correlations indicate that as one feature increases, the other tends to increase as well. The observation shows positive correlation between "education_num" and "income" concluding that higher education levels tend to be associated with higher income.

   Negative correlations indicate that as one feature increases, the other tends to decrease. The observation shows negative correlation between "age" and "hours_per_week" suggesting that older employees tend to work fewer hours per week.

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

   The accuracy of the model is approximately 85%, this means that the model correctly predicts the income level (>50K or <=50K) for 85% of the samples in the test dataset.

   The confusion matrix provides a detailed analysis of the model's predictions. For this dataset, the model correctly predicted 802 instances where income is >50K (TP), and it correctly predicted 4319 instances where income is <=50K (TN). However, it made 181 false positive predictions (FP) and 731 false negative predictions (FN).

   Precision for class 1(income > 50K) is approximately 0.82, indicates that when the model predicts a positive outcome (income > 50K), it is correct about 82% of the time.

   The recall for the income class 1 is approximately 0.52, which means that the model correctly identifies about 52% of all instances where income is actually greater than 50K.

   The F1-score for class 1 is approximately 0.64, indicating a reasonable balance between precision and recall for predicting incomes greater than 50K.

3. Compare the results obtained by applying random forest and decision tree algorithm on the Adult Census Income Dataset

   The observation between Decision Tree and Random Forest model is that they both have the similar accuracy of around 85%. This indicates that they had correctly predicted the income levels of the individuals. But they have lower recall for higher income claas 1, which shows the tendency to miss some individuals with income greater than 50K.