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.

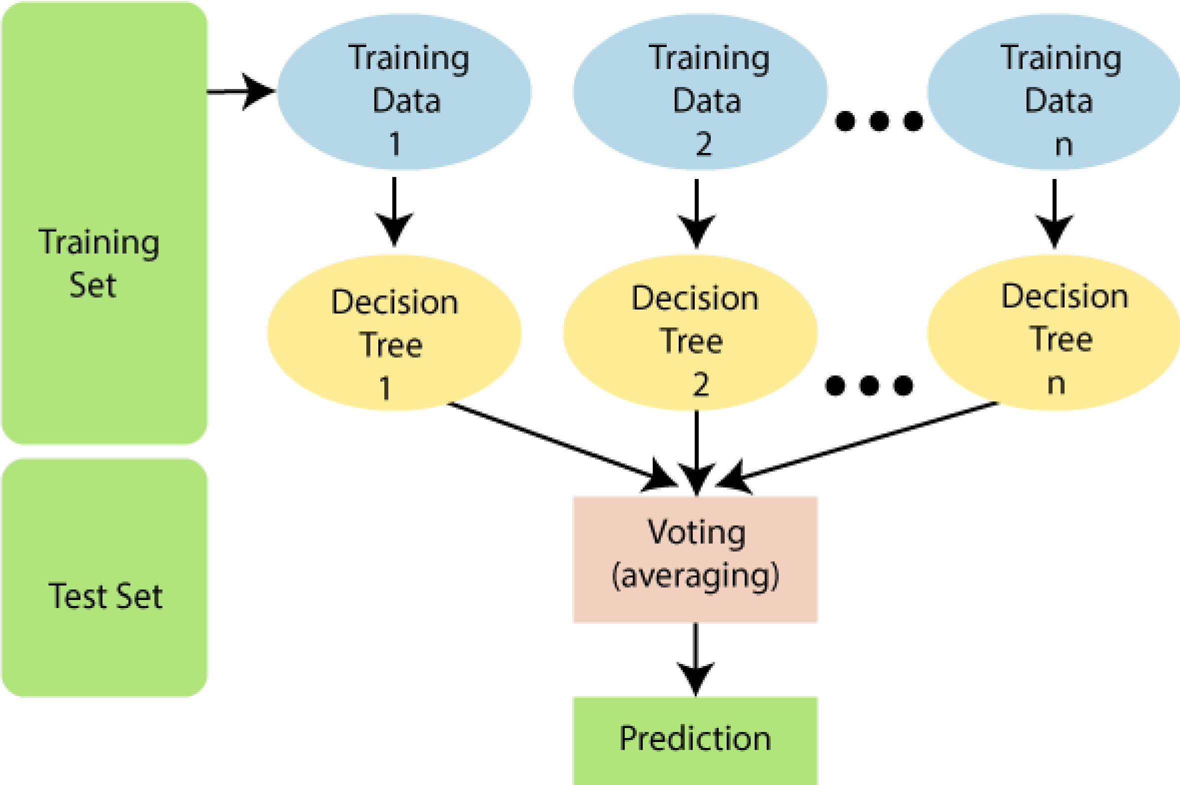
Theory:

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:

```
import pandas as pd

import seaborn as sns

import numpy as np

import matplotlib.pyplot as plt

from sklearn.preprocessing import LabelEncoder

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.linear_model import LogisticRegression

from sklearn.naive_bayes import GaussianNB

from sklearn.model_selection

import train_test_split,cross_val_score,KFold,GridSearchCV

from sklearn.metrics

import confusion_matrix,classification_report,accuracy_score

import scikitplot as skplt

dataset=pd.read_csv("adult.csv")

print(dataset.isnull().sum())
```

print(dataset.dtypes)

dataset.head()

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0

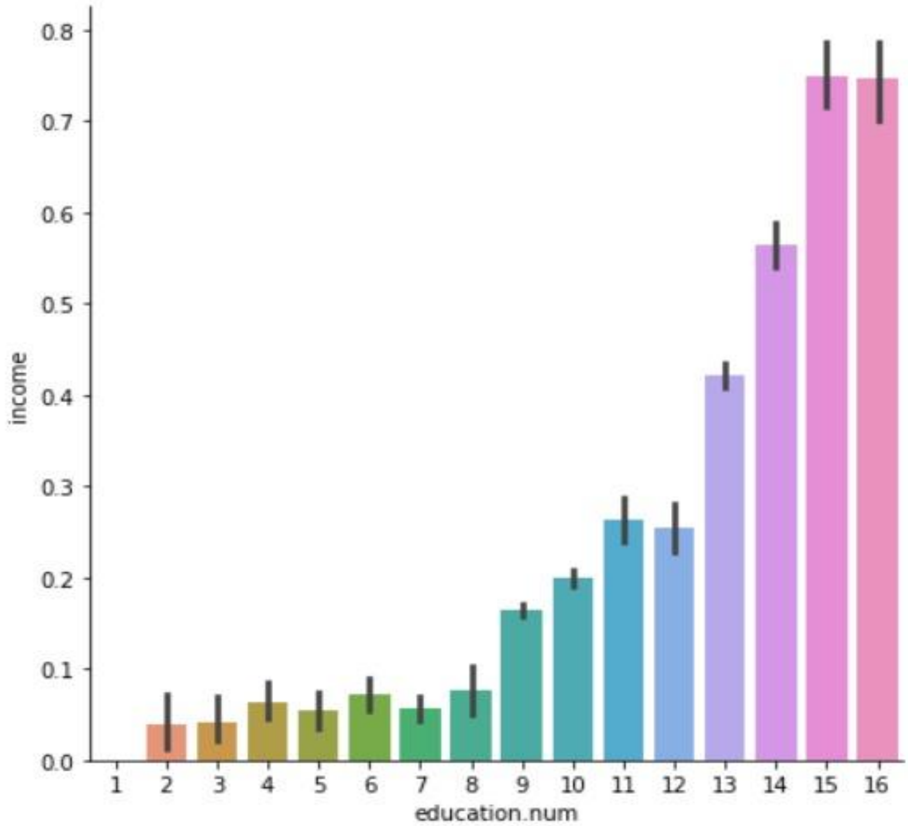
dataset = dataset[(dataset != '?').all(axis=1)]

#label the income objects as 0 and 1

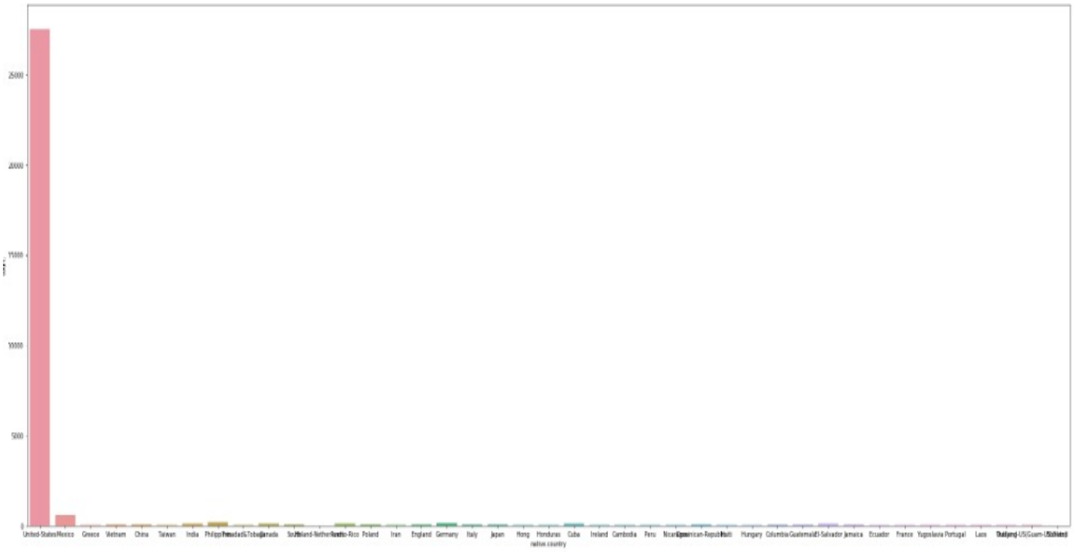
dataset['income']=dataset['income'].map({'<=50K': 0, '>50K': 1})

sns.catplot(x='education.num',y='income',data=dataset,kind='bar',height=6)

plt.show()

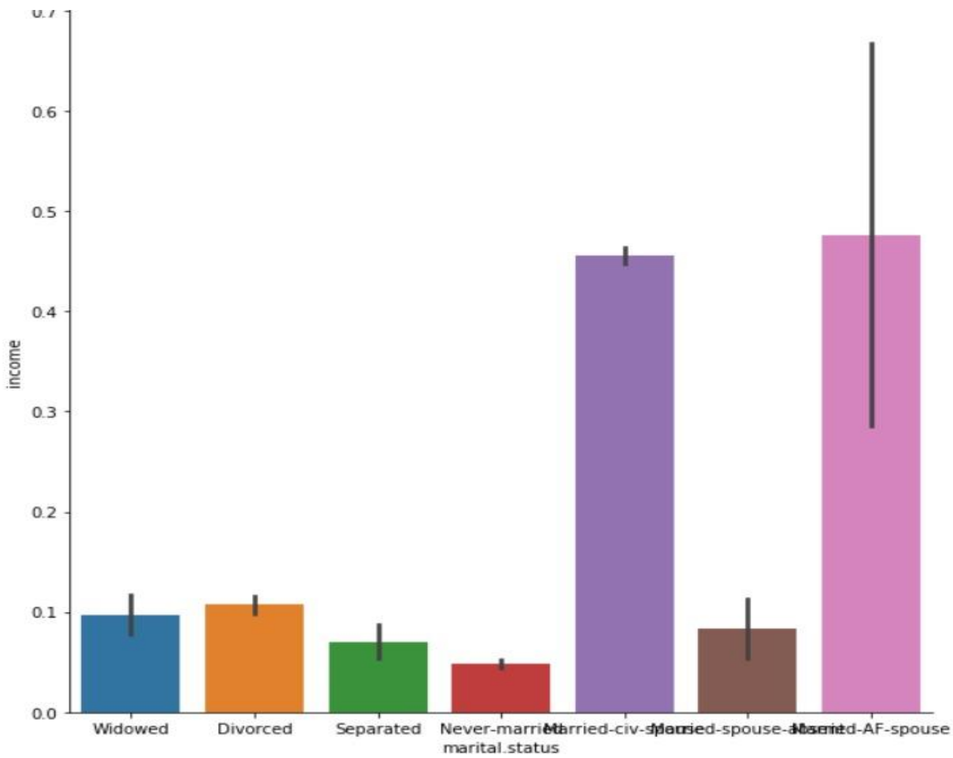



```
plt.figure(figsize=(38,14))

sns.countplot(x='native.country',data=dataset)

plt.show()
```

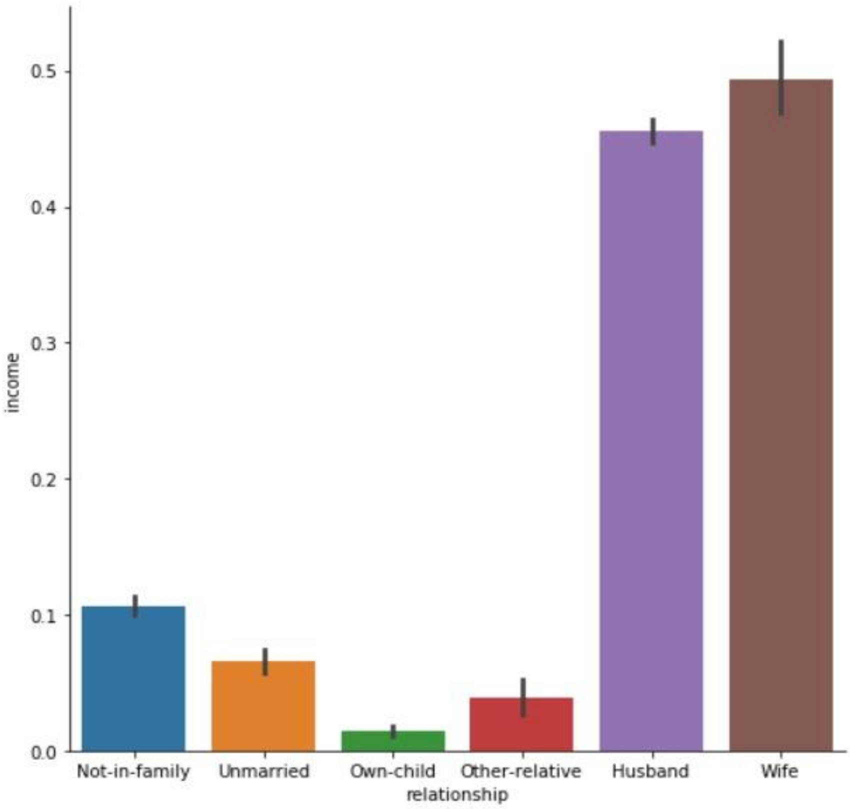


sns.factorplot(x='marital.status',y='income',data=dataset,kind='bar',height=8)

plt.show()



sns.factorplot(x='relationship',y='income',data=dataset,kind='bar',size=7) plt.show()



dataset['marital.status']=dataset['marital.status'].map({'Married-civ-spouse':'Married', 'Divorced':'Single', 'Never-married':'Single', 'Separated':'Single',

'Widowed':'Single', 'Married-spouse-absent':'Married', 'Married-AF-spouse':'Married'})

```
for column in dataset:

    enc=LabelEncoder()

    if dataset.dtypes[column]==np.object:

    dataset[column]=enc.fit_transform(dataset[column])
```

plt.figure(figsize=(14,10))

sns.heatmap(dataset.corr(),annot=True,fmt='.2f')

plt.show()

Conclusion:

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

   For all adult Census income dataset from the heatmap education and "education.num" are highly collected same can be said about the "marital-status" and "relationship" thus, we can develop "relationship " and " education" .

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

Accuracy: 85.011051

	precision	recall	f1-score	support
0	0.87	0.95	0.91	7942
1	0.70	0.45	0.55	2012
micro avg	0.85	0.85	0.85	9954
macro avg	0.79	0.70	0.73	9954
weighted avg	0.84	0.85	0.84	9954

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

   Random Forest or Decision Tree, is more accurate depending on various factors, including the specific dataset.
   We achieved 0.85 accuracy using random forest algorithm and using decision tree we got 84 .