Random Forest Assignment

support@intellipaat.com

+91-7022374614

US: 1-800-216-8930 (Toll-Free)

Problem Statement:

1. Census-income data plays the most important role in the democratic system of government, highly affecting the economic sectors. Census-related figures are used to allocate federal funding by the government to different states and localities.
2. Census data is also used for post census residents estimates and predictions, economic and social science research, and many other such applications. Therefore, the importance of this data and its accurate predictions is very clear to us. The main aim is to increase awareness about how the income factor actually has an impact not only on the individual lives of citizens but also an effect on the nation and its betterment. You will have a look at the data pulled out from the 1994 Census bureau database, and try to find insights into how various features have an effect on the income of an individual.
3. The data contains approximately 32,000 observations with over 15 variables.
4. The strategy is to analyze the data and perform a predictive task of classification to predict whether an individual makes over 50K a year or less by using a logistic regression algorithm. .

Data Description:

Column Names	Description
Age	Age of the individual
Workclass	department of the working individual
fnlwgt	Final weight of the individual
education	The education degree of the individual
education-num	Number of years of education
marital-status	Marital status of the individual
occupation	Occupation of the individual
relationship	Relation value
race	Ethnicity of the individual
sex	Female, Male
capital-gain	capital gain of the individual

capital-loss	capital loss of the individual
hours-per-week	number of working hours
native-country	The native country of the individual
Annual-Income	Annual income either >50K or <=50K

1. What is the biggest advantage that helps random forest classifiers to triumph over the decision trees?
   A. It has shown great predictive results over decision tree models.
   B. It Combines all positive predictions from all decision trees
   C. It works on the bagging method(bootstrap method)
   D. All of the above

2. In a given problem where you have a very large dataset with both continuous and categorical features, why would you choose the random forest classifier?
   A. Random forest can work on both regression and classification problem
   B. High accuracy with less need for interpretation
   C. Works well with the high dimensional data
   D. All of the above

3. Which of the following techniques is used in the Random Forest model?
   A. Bagging
   B. Boosting
   C. Ensemble
   D. None of the these

4. Choose the total population with income greater than 50% income?
   A. 75%
   B. 25%
   C. 24.08%
   D. 35%

5. Compute how many samples of the population are unmarried and working hours less than 20 hours?
   A. 134
   B. 145
   C. 127
   D. 123

6. Choose the correct list of age with minimum age , maximum and 50th Percentile of the age group?

A. [17,90,36]
B. [15,95,37]
C. [17,90,37]
D. All

7. From above census data which country has the highest population and the lowest population?

A. United-States and scotland
B. United-States and Holland-Netherlands
C. Scotland and Holland-Netherlands
D. Mexico and Holland-Netherlands

8. How does n_estimators work in the random forest classifier?

A. Number of random forests for the classifier.
B. Number of iterations
C. Training epochs
D. Number of decision trees

9. Can the target data for the random forest model be categorical or continuous value?

A. Yes
B. No

10. How can you use hyperparameter tuning to your advantage while working with the random forest classifier?

A. Improve the model's performance
B. Normalizes the features
C. Standardization of the data
D. All of the above

11. Select the best hyperparameters by RandomSearchCV and fit the model with the best hyperparameters and compute the accuracy score of the model.

A. 90% and above
B. 50% to 70%
C. 30% to 50&
D. None of the above

12. Which of the following Two features are most important in Random forest model?

A. Predict_proba
B. Correlation between 2 trees and how strong an individual tree is
C. sensitivity and specificity
D. None of the above

13. Based on what values, the feature importance will be calculated?
    A. mean increase gini and mean decrease accuracy
    B. Mean decrease gini and mean decrease accuracy
    C. mean increase gini and mean increase accuracy
    D. All of the above

14. From the above model, state the disadvantage of the random forest?
    A. It is a time consuming model building process
    B. It is same as all other model
    C. It's training time is huge due to the complexity of the model
    D. None of the above

15. Which are the two methods used for hyperparameter tuning and cross-validation?
    A. RandomForestCLassifier
    B. RandomizedSearchCV
    C. GridSearchCV
    D. RandomizedSearchCV and GridSearchCV