Project Description

Week-5 Graded Project (70 Marks)

Dataset Info

This data is taken from UCI.

https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Objective: to build a classifier that can predict whether a person is going to default or not based on their credit history details

1. Import required libraries
2. Read the provided csv file and check shape, info, and statistical summary of the data (5 marks)
3. Select columns having data type as object and save it in a new data frame (5 marks)
4. Check the shape and info of the above data frame having only object columns (5 marks)
5. Check for correlation among the predictors in the original dataset. (5 marks)
6. Drop irrelevant columns (subjective task) (5 Marks)
7. Encode above categorical data using get dummies and drop first. (hint - get dummies (data, drop first=True) (5 marks)
8. Separate target column from the features (5marks)
9. Split the data (Use 70:30 split) (5 marks)
10. Apply Random forest model (5 marks)
11. Evaluate above model using accuracy and confusion matrix (5 marks)
12. Check feature importance (Hint - RF.feature_importances_) (5 marks)
13. Apply Grid Search to tune important hyperparameters like,n_estimators,criterion, max_depth, and min_samples_leaf (5 marks)
14. Apply K-fold cross validation and check score for random forest model with the best parameters from grid search (5 marks)
15. Comment your findings (5marks)