READ ME

=====================================================================================

**Dataset:**

This dataset represents a series of student observations taken in a real classroom. The students are being observed using the BROMP protocol. The purpose of this protocol is to allow investigators to observe students working in real class situations and to code their affective states and on-task behaviors without direct interference in their activities

**Summary**:

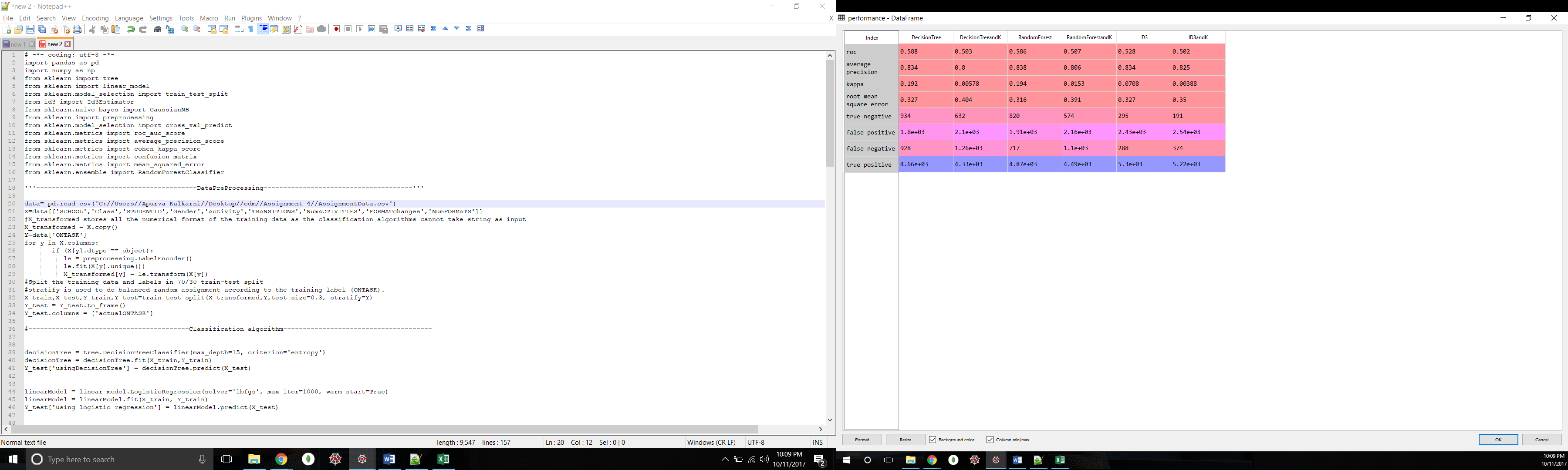
* In this Code, classification algorithm is implemented on the Dataset. Before classification, data is randomly balanced.
* Classification algorithms predict whether a student is ONTASK or OFFTASK. Which is done in two ways:
  + By splitting the data in 70/30 train-test split
  + By using 10-fold cross-validation
* Performance Evaluation is also performed to compare the results of all algorithms

**Detailed Steps Performed:**

1. The datafile is loaded in python Dataframe.
2. Then the data is transformed to numerical value using the LabelEncoder, as the classification algorithms cannot take string as input
3. Data is balanced randomly according to the column ‘ONTASK’ using stratify method
4. ID3 does feature selection by default. And, in classifiers, a parameter is passed to do feature selection by default criterion='entropy', hence feature selection is not required in data preprocessing.
5. The data is split into 70/30 train-test data
6. Classifiers are trained to predict ONTASK on a row-by-row basis
7. The same two classifiers are trained using 10-fold cross-validation (cross\_val\_predict function)
8. Training labels are converted into numeric value for performance evaluation
9. The performance of the classifiers is compared using ROC, Precision Average, Kappa, root mean square error and confusion matrix
10. **Which classifier performed better?**

According to the performance evaluation as shown below,

Random Forest Classifier performed better than others (High precision, roc, kappa and low root mean square error)



1. **How did the performance change from static sampling to cross-validation?**

The performance change was very minute but static sampling worked better than cross validation

1. **How did you control for overfitting in the dataset?**

Random Forest by default controls overfitting.

In Decision Tree and ID3, overfitting is controlled with the help of balanced random assignment(stratification) and by using K(10) fold cross validation