Machine Learning

Olivetti Face Data – Random Forest Classifier

Download the Olivetti faces dataset.
Visit https://scikit-learn.org/0.19/datasets/olivetti_faces.html
There are 40 classes (corresponding to 40 people), each class having 10 faces of the individual; so there are a total of 400 images.
Here each face is viewed as an image of size 64 × 64 (= 4096) pixels; each pixel having values 0 to 255 which are ultimately converted into floating numbers in the range [0,1].

Split the dataset into train
and test parts. Do this splitting randomly 10 times and report the average accuracy.
You may vary the test and train dataset sizes.

Build a Random Forest Classifier using the training data. Tune the parameters corresponding to pruning the decision tree. Use the best decision tree to classify the test dataset and obtain the accuracy. Use misclassification impurity also.

CODE:

Please find the code committed for Random Forest Classifier as
RandomForestClassifier_OlivettiFaceData_Impl.py

- **Random Forest Classifier** from sklearn is used to solve this task.
- The training and test data is split randomly, and the test size is varied randomly in the range 0.2 to 0.35 to find the accuracies for 10 iterations.
- **The number of estimators or decision trees is evaluated for 100, 200,300 and 400.**
- **Feature importance is also extracted for 3 example scenarios.**

RESULT:

Pls find below a tabulation of results from each of the 10 iterations for different Tree sizes.

Number of Trees	Test Size	Accuracy
100	0.35	0.942857
	0.2	0.9625
	0.2	0.9625
	0.35	0.914286
	0.25	0.92
	0.35	0.921429
	0.35	0.914286
	0.3	0.925
	0.25	0.96
	0.3	0.916667
	Avg	0.933952

Number of Trees	Test Size	Accuracy
200	0.35	0.964286
	0.35	0.935714
	0.35	0.942857
	0.2	0.95
	0.35	0.928571
	0.2	0.9625
	0.35	0.914286
	0.25	0.92
	0.35	0.914286
	0.2	0.95
	Avg	0.93825

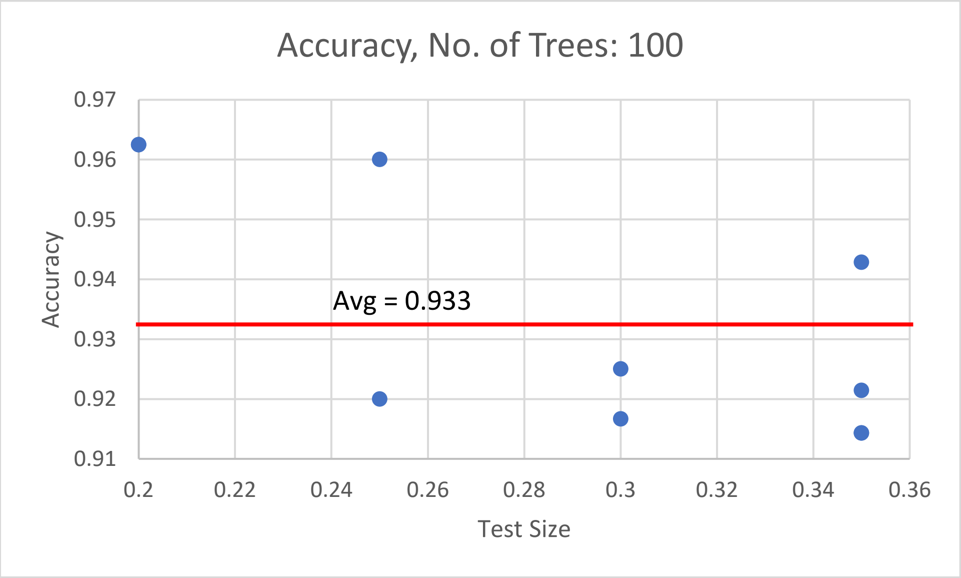
Number of Trees	Test Size	Accuracy
300	0.2	0.95
	0.2	0.9625
	0.35	0.935714
	0.35	0.9
	0.3	0.925
	0.3	0.941667
	0.3	0.933333
	0.25	0.98
	0.25	0.95
	0.3	0.966667
	Avg	0.944488

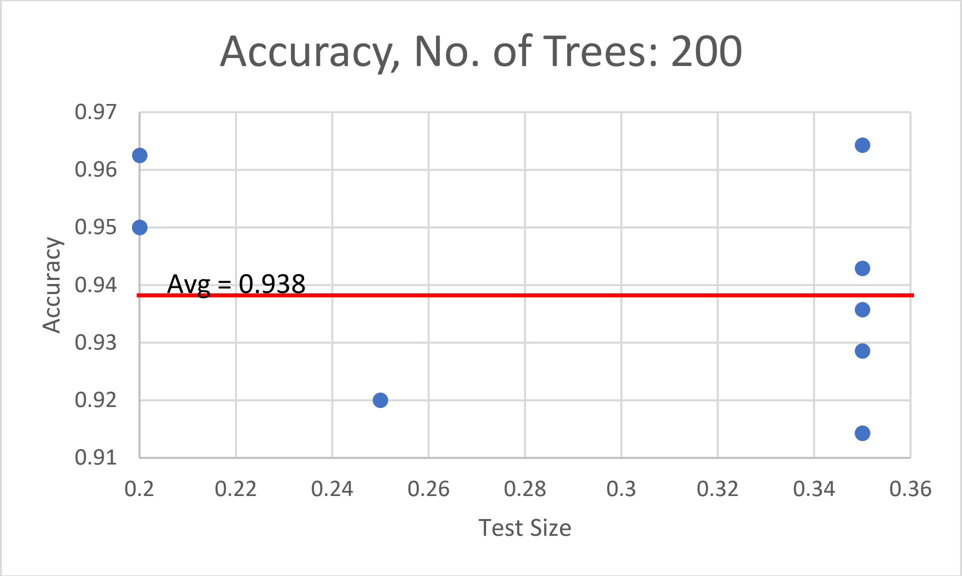
Number of Trees	Test Size	Accuracy
400	0.25	0.96
	0.35	0.921429
	0.25	0.96
	0.3	0.933333
	0.25	0.95
	0.2	0.975
	0.2	0.925
	0.25	0.99
	0.3	0.933333
	0.25	0.93
	Avg	0.94781

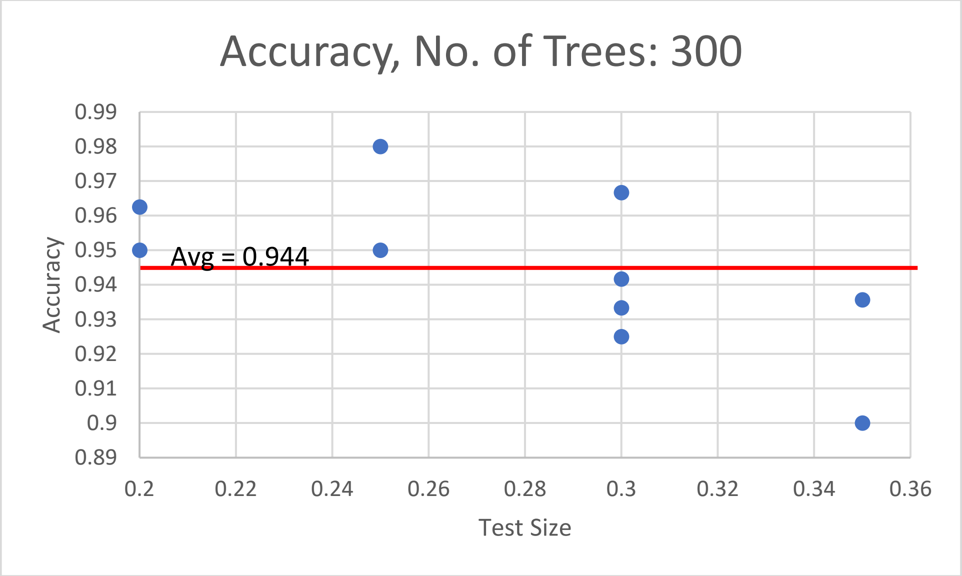
Number of Trees	Avg Accuracy
100	0.933952381
200	0.93825
300	0.944488095
400	0.947809524

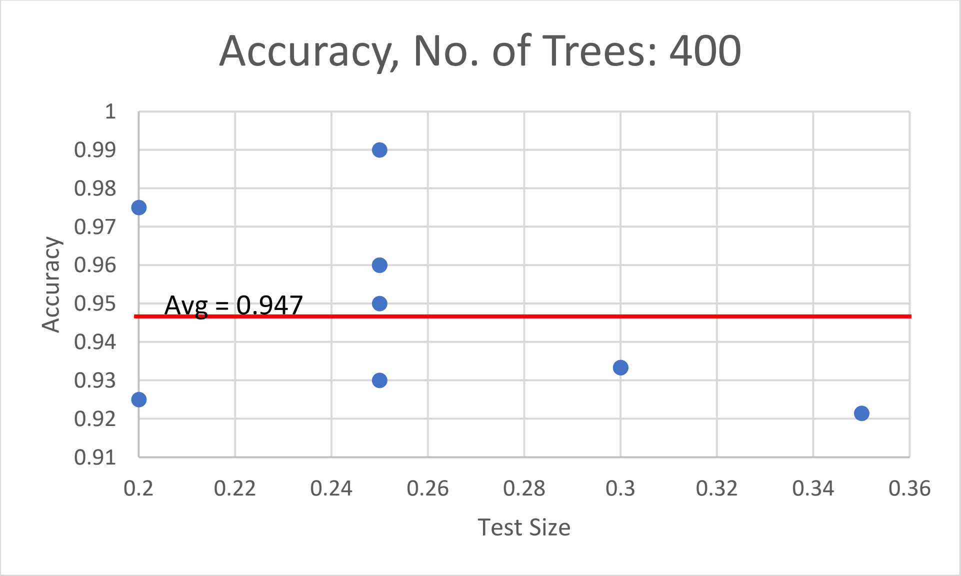
Plots are shown for each of these tables.
Plots are shown for 2 sets of Number of Trees and Test Size iterations.

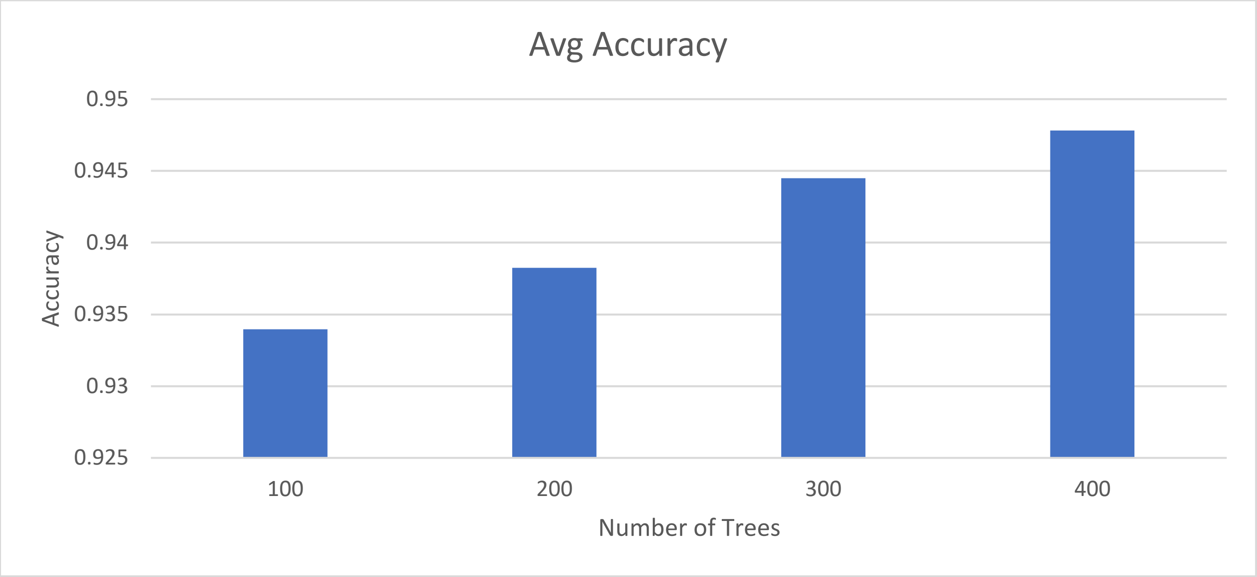
PLOT:

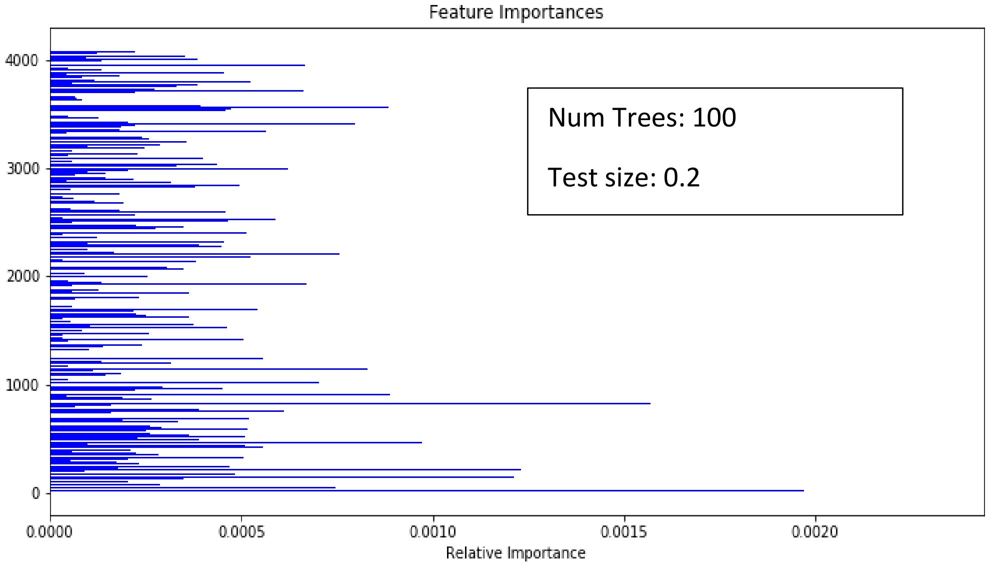


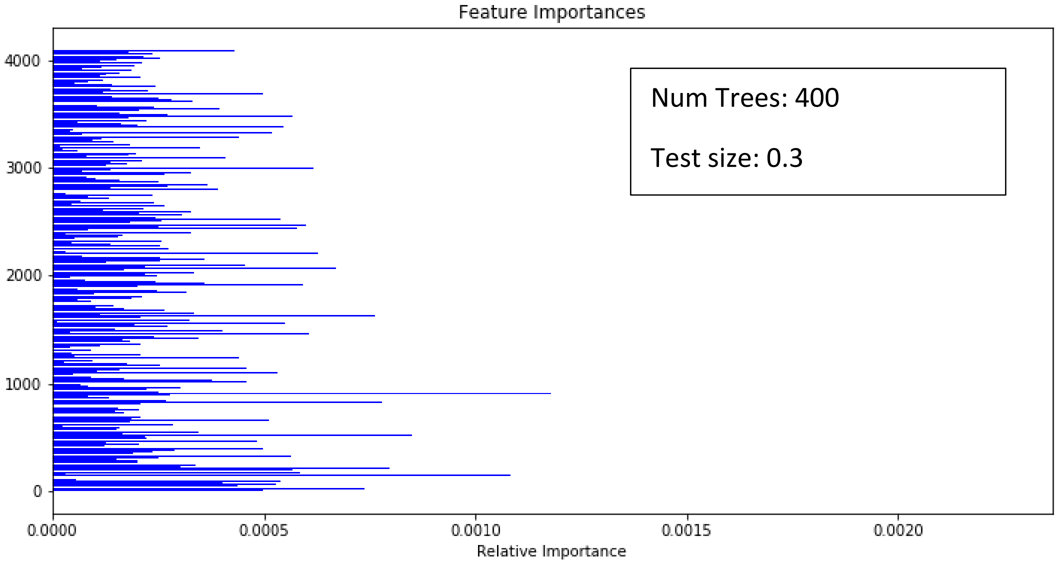

Accuracy, No. of Trees: 200
Accuracy
Test Size
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.2
0.22
0.24
0.26
0.28
0.3
0.32
0.34
0.36
Avg = 0.938


Accuracy, No. of Trees: 300
Accuracy
Test Size
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
0.89
0.2
0.22
0.24
0.26
0.28
0.3
0.32
0.34
0.36
Avg = 0.944


Accuracy, No. of Trees: 400
Accuracy
Test Size
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.2
0.22
0.24
0.26
0.28
0.3
0.32
0.34
0.36
Avg = 0.947


Avg Accuracy
0.95
0.945
0.94
0.935
0.93
0.925
Accuracy
100
200
300
400
Number of Trees


Feature Importances
4000
3000
2000
1000
0
0.0000
0.0005
0.0010
0.0015
0.0020
Relative Importance
Num Trees: 100
Test size: 0.2


Feature Importances
4000
3000
2000
1000
0
0.0000
0.0005
0.0010
0.0015
0.0020
Relative Importance
Num Trees: 400
Test size: 0.3

INFERENCE/ANALYSIS:

- Random Forest Classifier uses multiple decision trees and combines the output of the multiple Decision Trees to generate the final output.
- **Each node in the decision trees works on a random subset of features to calculate the output.**
- Random forest is able to prevent overfitting which is a general problem of a single decision trees.
- **Random forest is able to provide better accuracy than Decision Tree (Avg accuracy > 0.9)**
- **With the increase in number of trees from 100 to 400 , the average accuracy across 10 iterations of randomly generated test and train size also increases.**
- This shows that when number of trees used to determine within a random forest is tuned appropriately for the given dataset, accuracy improves**.**
- **Feature importance is depicted for each of the 4096 features and for this data set they are nearly similar.**

Decision tree does not give good accuracy for high dimensional featured dataset like an olivetti dataset with 4096 features due to risk of overfitting.
In order to solve this limitation of decision tree, classifiers like Random forest are preferred that use multiple decision tree and XGboost classifier which also uses multiple decision trees and a host of other parameters outperforms the simple decision tree despite best parameterization.

RESOURCES USED FOR THE ASSIGNMENT:

- **Environment:**
  Anaconda, Jupyter notebook
- **Software :**
  Python
  Python libraries/modules: Pandas, Numpy, SkLearn ,XGboost etc