Mid Sem Lab Exam - Report

- Date : 27 March 2019
- CED15I031

Problem Statement

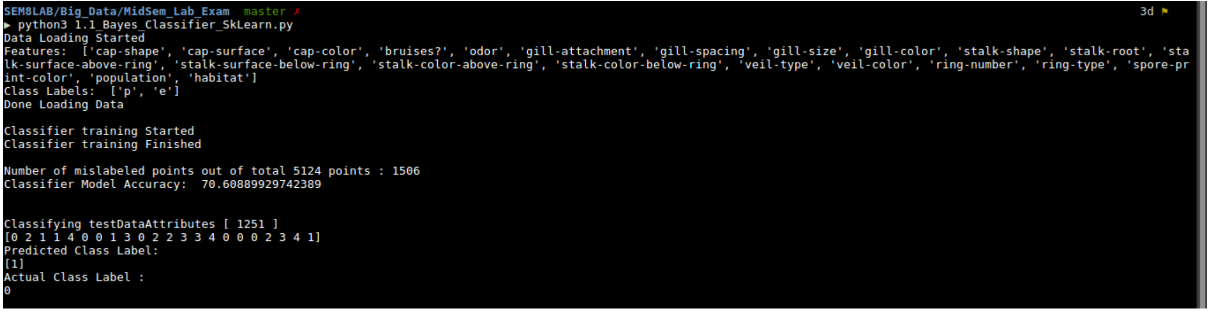
- Mushroom Data @ UCI

1. Test drive any four built in classifiers supported by your platform to classify the test data sets.
2. Compare the four classifiers for their performance measures (detailed measures).
3. Test drive the Association Rule based Classifier implemented as part of your lab exercise over this data set.
4. Compare the model in (3) to any one in (1) in terms of detailed performance measures.

Solution

1. Classifiers tested
   - Complement Naive Bayes Classifier from SciKit Learn

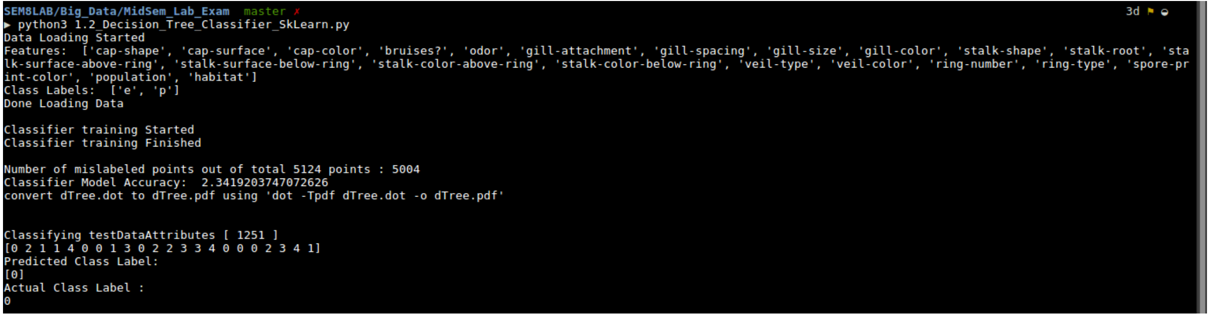
     > Output



```
SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                  3d ⚑
▶ python3 1.1_Bayes_Classifier_SkLearn.py
Data Loading Started
Features:  ['cap-shape', 'cap-surface', 'cap-color', 'bruises?', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'sta
lk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-pr
int-color', 'population', 'habitat']
Class Labels:  ['p', 'e']
Done Loading Data

Classifier training Started
Classifier training Finished

Number of mislabeled points out of total 5124 points : 1506
Classifier Model Accuracy:  70.60889929742389


Classifying testDataAttributes [ 1251 ]
[0 2 1 1 4 0 0 1 3 0 2 2 3 3 4 0 0 0 2 3 4 1]
Predicted Class Label:
[1]
Actual Class Label :
0
```

   - Decision Tree Classifier from SciKit Learn

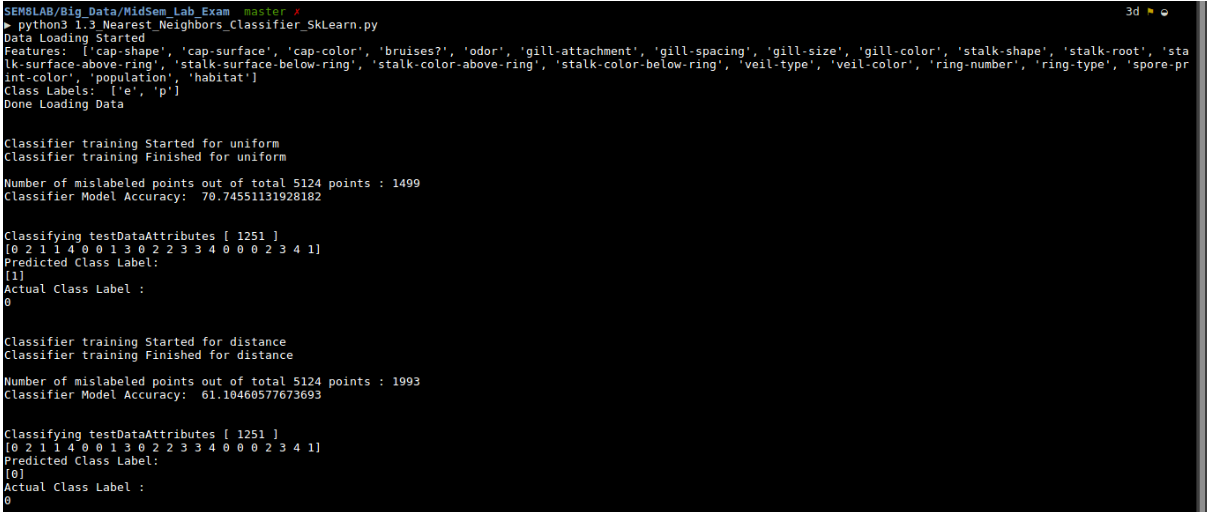
     > Output



```
SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                 3d ⚑ ◒
▶ python3 1.2_Decision_Tree_Classifier_SkLearn.py
Data Loading Started
Features:  ['cap-shape', 'cap-surface', 'cap-color', 'bruises?', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'sta
lk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-pr
int-color', 'population', 'habitat']
Class Labels:  ['e', 'p']
Done Loading Data

Classifier training Started
Classifier training Finished

Number of mislabeled points out of total 5124 points : 5004
Classifier Model Accuracy:  2.3419203747072626
convert dTree.dot to dTree.pdf using 'dot -Tpdf dTree.dot -o dTree.pdf'


Classifying testDataAttributes [ 1251 ]
[0 2 1 1 4 0 0 1 3 0 2 2 3 3 4 0 0 0 2 3 4 1]
Predicted Class Label:
[0]
Actual Class Label :
0
```

   - Nearest Neighbors Classifier from SciKit Learn

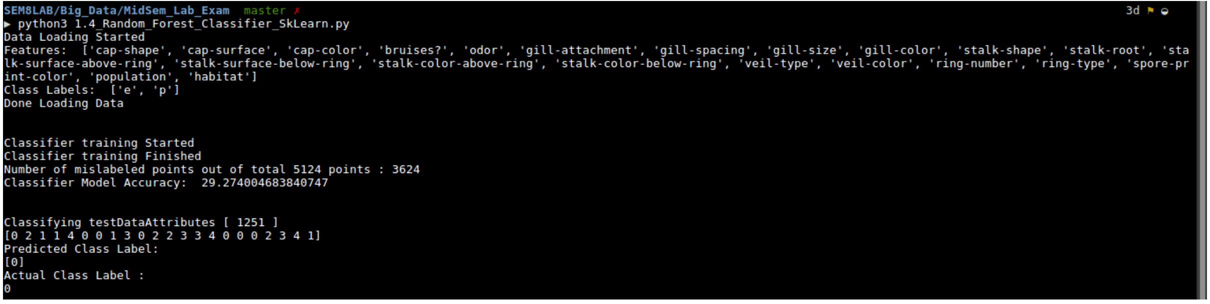
Output



```
SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                      3d
▶ python3 1.3_Nearest_Neighbors_Classifier_SkLearn.py
Data Loading Started
Features:  ['cap-shape', 'cap-surface', 'cap-color', 'bruises?', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'sta
lk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-pr
int-color', 'population', 'habitat']
Class Labels:  ['e', 'p']
Done Loading Data


Classifier training Started for uniform
Classifier training Finished for uniform

Number of mislabeled points out of total 5124 points : 1499
Classifier Model Accuracy:  70.74551131928182


Classifying testDataAttributes [ 1251 ]
[0 2 1 1 4 0 0 1 3 0 2 2 3 3 4 0 0 0 2 3 4 1]
Predicted Class Label:
[1]
Actual Class Label :
0


Classifier training Started for distance
Classifier training Finished for distance

Number of mislabeled points out of total 5124 points : 1993
Classifier Model Accuracy:  61.10460577673693


Classifying testDataAttributes [ 1251 ]
[0 2 1 1 4 0 0 1 3 0 2 2 3 3 4 0 0 0 2 3 4 1]
Predicted Class Label:
[0]
Actual Class Label :
0
```

- Random Forest Classifier from SciKit Learn

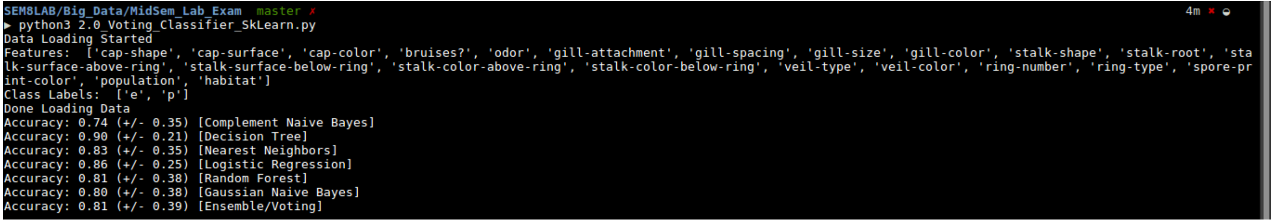
  Output



```
SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                      3d
▶ python3 1.4_Random_Forest_Classifier_SkLearn.py
Data Loading Started
Features:  ['cap-shape', 'cap-surface', 'cap-color', 'bruises?', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'sta
lk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-pr
int-color', 'population', 'habitat']
Class Labels:  ['e', 'p']
Done Loading Data


Classifier training Started
Classifier training Finished
Number of mislabeled points out of total 5124 points : 3624
Classifier Model Accuracy:  29.274004683840747


Classifying testDataAttributes [ 1251 ]
[0 2 1 1 4 0 0 1 3 0 2 2 3 3 4 0 0 0 2 3 4 1]
Predicted Class Label:
[0]
Actual Class Label :
0
```

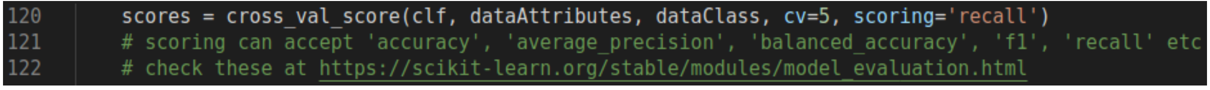
2. Used Voting Classifier from SciKitLearn

   Output



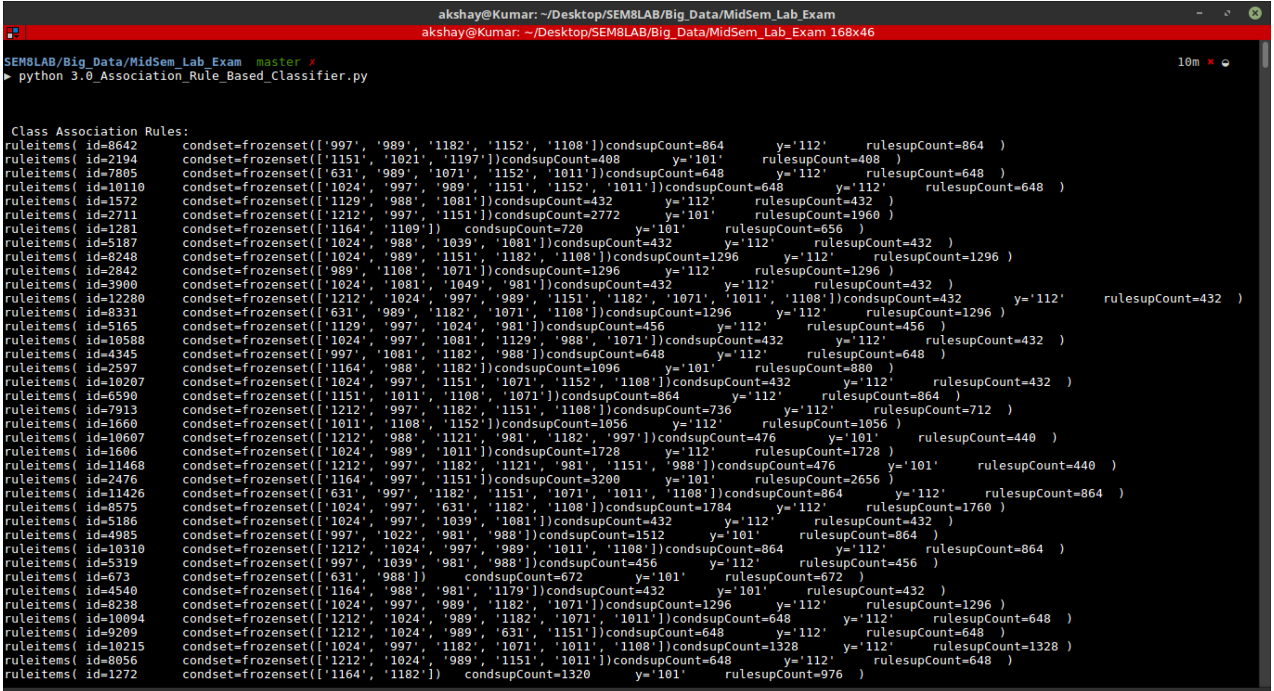
```
SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                      4m
▶ python3 2.0_Voting_Classifier_SkLearn.py
Data Loading Started
Features:  ['cap-shape', 'cap-surface', 'cap-color', 'bruises?', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'sta
lk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-pr
int-color', 'population', 'habitat']
Class Labels:  ['e', 'p']
Done Loading Data
Accuracy: 0.74 (+/- 0.35) [Complement Naive Bayes]
Accuracy: 0.90 (+/- 0.21) [Decision Tree]
Accuracy: 0.83 (+/- 0.35) [Nearest Neighbors]
Accuracy: 0.86 (+/- 0.25) [Logistic Regression]
Accuracy: 0.81 (+/- 0.38) [Random Forest]
Accuracy: 0.80 (+/- 0.38) [Gaussian Naive Bayes]
Accuracy: 0.81 (+/- 0.39) [Ensemble/Voting]
```

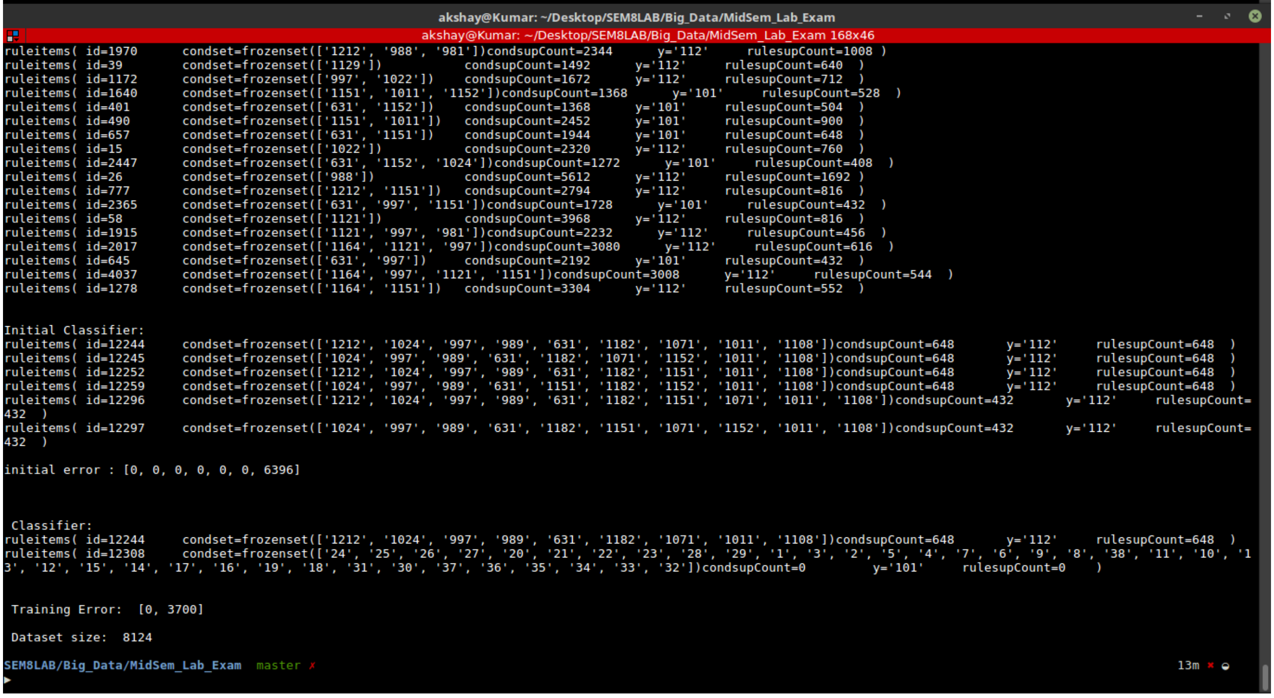
   - Made use of Cross Validation i.e. cross_val_score() and Model Evaluation Parameters
   - To compare any parameter of these classifiers, just change the scoring parameter in cross_val_score function @ Line 120



```
120        scores = cross_val_score(clf, dataAttributes, dataClass, cv=5, scoring='recall')
121        # scoring can accept 'accuracy', 'average_precision', 'balanced_accuracy', 'f1', 'recall' etc
122        # check these at https://scikit-learn.org/stable/modules/model_evaluation.html
```

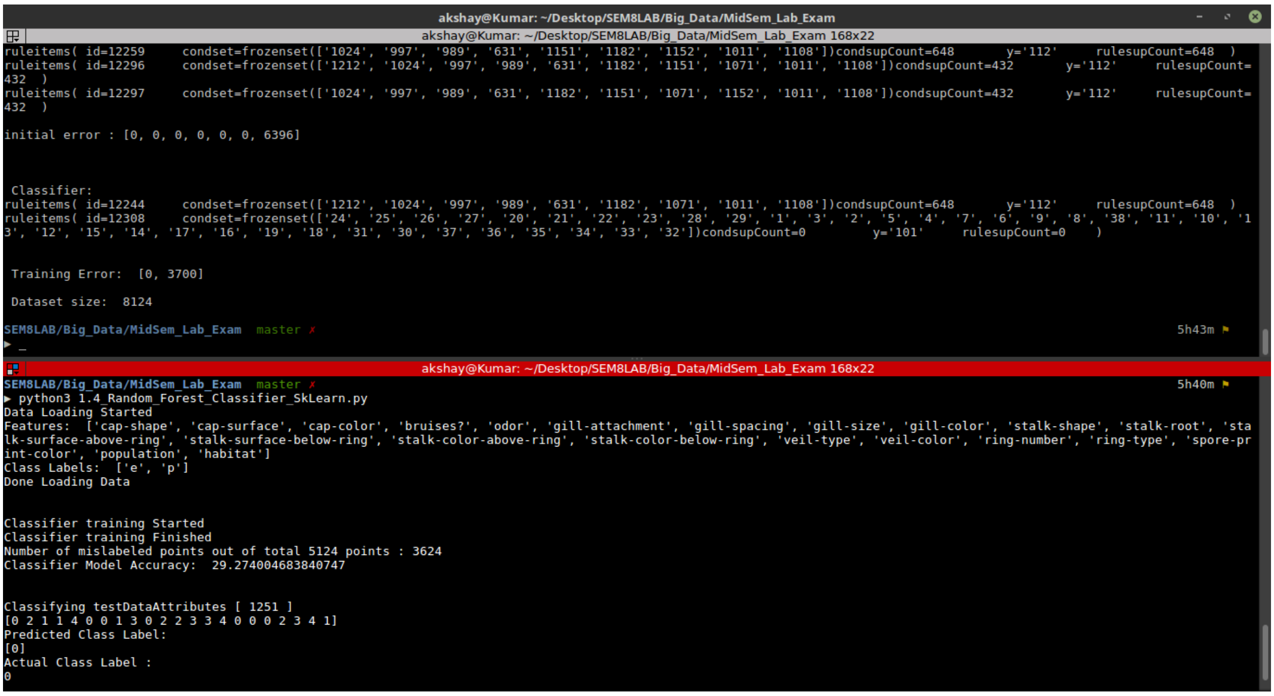
3. ARBC Implementation referred from Big Data - Ruchi09
   - Step 1 : Transform data so that similar value notaions in different attributes are considered as different, else the rule generation is affected
   - Step 2 : Selecting K Best attributes as 22 attributes will cause overfitting for the classifier
   - Step 3 : Run ARBC to find rules and their supports and confidence
     - Step 3.1 : Find the Rule with minimum error and then use it to find the class.

> Output
>
> 
>
>

akshay@Kumar: ~/Desktop/SEM8LAB/Big_Data/MidSem_Lab_Exam
> akshay@Kumar: ~/Desktop/SEM8LAB/Big_Data/MidSem_Lab_Exam 168x46
>
> ```
> SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                                                     10m ✖ ◒
> ▶ python 3.0_Association_Rule_Based_Classifier.py
>
>
>  Class Association Rules:
> ruleitems( id=8642      condset=frozenset(['997', '989', '1182', '1152', '1108'])condsupCount=864       y='112'     rulesupCount=864  )
> ruleitems( id=2194      condset=frozenset(['1151', '1021', '1197'])condsupCount=408       y='101'     rulesupCount=408  )
> ruleitems( id=7805      condset=frozenset(['631', '989', '1071', '1152', '1011'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=10110     condset=frozenset(['1024', '997', '989', '1151', '1152', '1011'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=1572      condset=frozenset(['1129', '988', '1081'])condsupCount=432       y='112'     rulesupCount=432  )
> ruleitems( id=2711      condset=frozenset(['1212', '997', '1151'])condsupCount=2772      y='101'     rulesupCount=1960 )
> ruleitems( id=1281      condset=frozenset(['1164', '1109'])   condsupCount=720       y='101'     rulesupCount=656  )
> ruleitems( id=5187      condset=frozenset(['1024', '988', '1039', '1081'])condsupCount=432       y='112'     rulesupCount=432  )
> ruleitems( id=8248      condset=frozenset(['1024', '989', '1151', '1182', '1108'])condsupCount=1296      y='112'     rulesupCount=1296 )
> ruleitems( id=2842      condset=frozenset(['989', '1108', '1071'])condsupCount=1296      y='112'     rulesupCount=1296 )
> ruleitems( id=3900      condset=frozenset(['1024', '1081', '1049', '981'])condsupCount=432       y='112'     rulesupCount=432  )
> ruleitems( id=12280     condset=frozenset(['1212', '1024', '997', '989', '1151', '1182', '1071', '1011', '1108'])condsupCount=432       y='112'     rulesupCount=432  )
> ruleitems( id=8331      condset=frozenset(['631', '989', '1182', '1071', '1108'])condsupCount=1296      y='112'     rulesupCount=1296 )
> ruleitems( id=5165      condset=frozenset(['1129', '997', '1024', '981'])condsupCount=456       y='112'     rulesupCount=456  )
> ruleitems( id=10588     condset=frozenset(['1024', '997', '1081', '1129', '988', '1071'])condsupCount=432       y='112'     rulesupCount=432  )
> ruleitems( id=4345      condset=frozenset(['997', '1081', '1182', '988'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=2597      condset=frozenset(['1164', '988', '1182'])condsupCount=1096      y='101'     rulesupCount=880  )
> ruleitems( id=10207     condset=frozenset(['1024', '997', '1151', '1071', '1152', '1108'])condsupCount=432       y='112'     rulesupCount=432  )
> ruleitems( id=6590      condset=frozenset(['1151', '1011', '1108', '1071'])condsupCount=864       y='112'     rulesupCount=864  )
> ruleitems( id=7913      condset=frozenset(['1212', '997', '1182', '1151', '1108'])condsupCount=736       y='112'     rulesupCount=712  )
> ruleitems( id=1660      condset=frozenset(['1011', '1108', '1152'])condsupCount=1056      y='112'     rulesupCount=1056 )
> ruleitems( id=10607     condset=frozenset(['1212', '988', '1121', '981', '1182', '997'])condsupCount=476       y='101'     rulesupCount=440  )
> ruleitems( id=1606      condset=frozenset(['1024', '989', '1011'])condsupCount=1728      y='112'     rulesupCount=1728 )
> ruleitems( id=11468     condset=frozenset(['1212', '997', '1182', '1121', '981', '1151', '988'])condsupCount=476       y='101'     rulesupCount=440  )
> ruleitems( id=2476      condset=frozenset(['1164', '997', '1151'])condsupCount=3200      y='101'     rulesupCount=2656 )
> ruleitems( id=11426     condset=frozenset(['631', '997', '1182', '1151', '1071', '1011', '1108'])condsupCount=864       y='112'     rulesupCount=864  )
> ruleitems( id=8575      condset=frozenset(['1024', '997', '631', '1182', '1108'])condsupCount=1784      y='112'     rulesupCount=1760 )
> ruleitems( id=5186      condset=frozenset(['1024', '997', '1039', '1081'])condsupCount=432       y='112'     rulesupCount=432  )
> ruleitems( id=4985      condset=frozenset(['997', '1022', '981', '988'])condsupCount=1512      y='101'     rulesupCount=864  )
> ruleitems( id=10310     condset=frozenset(['1212', '1024', '997', '989', '1011', '1108'])condsupCount=864       y='112'     rulesupCount=864  )
> ruleitems( id=5319      condset=frozenset(['997', '1039', '981', '988'])condsupCount=456       y='112'     rulesupCount=456  )
> ruleitems( id=673       condset=frozenset(['631', '988'])     condsupCount=672       y='101'     rulesupCount=672  )
> ruleitems( id=4540      condset=frozenset(['1164', '988', '981', '1179'])condsupCount=432       y='101'     rulesupCount=432  )
> ruleitems( id=8238      condset=frozenset(['1024', '997', '989', '1182', '1071'])condsupCount=1296      y='112'     rulesupCount=1296 )
> ruleitems( id=10094     condset=frozenset(['1212', '1024', '989', '1182', '1071', '1011'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=9209      condset=frozenset(['1212', '1024', '989', '631', '1151'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=10215     condset=frozenset(['1024', '997', '1182', '1071', '1011', '1108'])condsupCount=1328      y='112'     rulesupCount=1328 )
> ruleitems( id=8056      condset=frozenset(['1212', '1024', '989', '1151', '1011'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=1272      condset=frozenset(['1164', '1182'])   condsupCount=1320      y='101'     rulesupCount=976  )
> ```
>
> 
>
>

akshay@Kumar: ~/Desktop/SEM8LAB/Big_Data/MidSem_Lab_Exam
> akshay@Kumar: ~/Desktop/SEM8LAB/Big_Data/MidSem_Lab_Exam 168x46
>
> ```
> ruleitems( id=1970      condset=frozenset(['1212', '988', '981'])condsupCount=2344      y='112'     rulesupCount=1008 )
> ruleitems( id=39        condset=frozenset(['1129'])           condsupCount=1492      y='112'     rulesupCount=640  )
> ruleitems( id=1172      condset=frozenset(['997', '1022'])    condsupCount=1672      y='112'     rulesupCount=712  )
> ruleitems( id=1640      condset=frozenset(['1151', '1011', '1152'])condsupCount=1368      y='101'     rulesupCount=528  )
> ruleitems( id=401       condset=frozenset(['631', '1152'])    condsupCount=1368      y='101'     rulesupCount=504  )
> ruleitems( id=490       condset=frozenset(['1151', '1011'])   condsupCount=2452      y='101'     rulesupCount=900  )
> ruleitems( id=657       condset=frozenset(['631', '1151'])    condsupCount=1944      y='101'     rulesupCount=648  )
> ruleitems( id=15        condset=frozenset(['1022'])           condsupCount=2320      y='112'     rulesupCount=760  )
> ruleitems( id=2447      condset=frozenset(['631', '1152', '1024'])condsupCount=1272      y='101'     rulesupCount=408  )
> ruleitems( id=26        condset=frozenset(['988'])            condsupCount=5612      y='112'     rulesupCount=1692 )
> ruleitems( id=777       condset=frozenset(['1212', '1151'])   condsupCount=2794      y='112'     rulesupCount=816  )
> ruleitems( id=2365      condset=frozenset(['631', '997', '1151'])condsupCount=1728      y='101'     rulesupCount=432  )
> ruleitems( id=58        condset=frozenset(['1121'])           condsupCount=3968      y='112'     rulesupCount=816  )
> ruleitems( id=1915      condset=frozenset(['1121', '997', '981'])condsupCount=2232      y='112'     rulesupCount=456  )
> ruleitems( id=2017      condset=frozenset(['1164', '1121', '997'])condsupCount=3080      y='112'     rulesupCount=616  )
> ruleitems( id=645       condset=frozenset(['631', '997'])     condsupCount=2192      y='101'     rulesupCount=432  )
> ruleitems( id=4037      condset=frozenset(['1164', '997', '1121', '1151'])condsupCount=3008      y='112'     rulesupCount=544  )
> ruleitems( id=1278      condset=frozenset(['1164', '1151'])   condsupCount=3304      y='112'     rulesupCount=552  )
>
>
> Initial Classifier:
> ruleitems( id=12244     condset=frozenset(['1212', '1024', '997', '989', '631', '1182', '1071', '1011', '1108'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=12245     condset=frozenset(['1024', '997', '989', '631', '1182', '1071', '1152', '1011', '1108'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=12252     condset=frozenset(['1212', '1024', '997', '989', '631', '1182', '1151', '1011', '1108'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=12259     condset=frozenset(['1024', '997', '989', '631', '1151', '1182', '1152', '1011', '1108'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=12296     condset=frozenset(['1212', '1024', '997', '989', '631', '1182', '1151', '1071', '1011', '1108'])condsupCount=432       y='112'     rulesupCount=
> 432  )
> ruleitems( id=12297     condset=frozenset(['1024', '997', '989', '631', '1182', '1151', '1071', '1152', '1011', '1108'])condsupCount=432       y='112'     rulesupCount=
> 432  )
>
> initial error : [0, 0, 0, 0, 0, 0, 6396]
>
>
>
>  Classifier:
> ruleitems( id=12244     condset=frozenset(['1212', '1024', '997', '989', '631', '1182', '1071', '1011', '1108'])condsupCount=648       y='112'     rulesupCount=648  )
> ruleitems( id=12308     condset=frozenset(['24', '25', '26', '27', '20', '21', '22', '23', '28', '29', '1', '3', '2', '5', '4', '7', '6', '9', '8', '38', '11', '10', '1
> 3', '12', '15', '14', '17', '16', '19', '18', '31', '30', '37', '36', '35', '34', '33', '32'])condsupCount=0         y='101'     rulesupCount=0    )
>
>
>  Training Error:  [0, 3700]
>
>  Dataset size:  8124
>
> SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                                                     13m ✖ ◒
> ▶
> ```

4. Comparing ARBC with Random Forest Classifier
    - Step 1 : Run ARBC and find the misclassified points
    - Step 2 : Run Random Forest Classifier and see the number of misclassified points.



akshay@Kumar: ~/Desktop/SEM8LAB/Big_Data/MidSem_Lab_Exam
akshay@Kumar: ~/Desktop/SEM8LAB/Big_Data/MidSem_Lab_Exam 168x22

```
ruleitems( id=12259     condset=frozenset(['1024', '997', '989', '631', '1151', '1182', '1152', '1011', '1108'])condsupCount=648       y='112'     rulesupCount=648  )
ruleitems( id=12296     condset=frozenset(['1212', '1024', '997', '989', '631', '1182', '1151', '1071', '1011', '1108'])condsupCount=432       y='112'     rulesupCount=
432  )
ruleitems( id=12297     condset=frozenset(['1024', '997', '989', '631', '1182', '1151', '1071', '1152', '1011', '1108'])condsupCount=432       y='112'     rulesupCount=
432  )

initial error : [0, 0, 0, 0, 0, 0, 6396]



 Classifier:
ruleitems( id=12244     condset=frozenset(['1212', '1024', '997', '989', '631', '1182', '1071', '1011', '1108'])condsupCount=648       y='112'     rulesupCount=648  )
ruleitems( id=12308     condset=frozenset(['24', '25', '26', '27', '20', '21', '22', '23', '28', '29', '1', '3', '2', '5', '4', '7', '6', '9', '8', '38', '11', '10', '1
3', '12', '15', '14', '17', '16', '19', '18', '31', '30', '37', '36', '35', '34', '33', '32'])condsupCount=0         y='101'     rulesupCount=0    )


 Training Error:  [0, 3700]

 Dataset size:  8124

SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                                                  5h43m ⚑
▶ _
```

akshay@Kumar: ~/Desktop/SEM8LAB/Big_Data/MidSem_Lab_Exam 168x22

```
SEM8LAB/Big_Data/MidSem_Lab_Exam  master ✗                                                                                                                  5h40m ⚑
▶ python3 1.4_Random_Forest_Classifier_SkLearn.py
Data Loading Started
Features:  ['cap-shape', 'cap-surface', 'cap-color', 'bruises?', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'sta
lk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-pr
int-color', 'population', 'habitat']
Class Labels:  ['e', 'p']
Done Loading Data


Classifier training Started
Classifier training Finished
Number of mislabeled points out of total 5124 points : 3624
Classifier Model Accuracy:  29.274004683840747


Classifying testDataAttributes [ 1251 ]
[0 2 1 1 4 0 0 1 3 0 2 2 3 3 4 0 0 0 2 3 4 1]
Predicted Class Label:
[0]
Actual Class Label :
0
```

- Step 3 : Compare the error rate

```
Error rate in ARBC : (3700 / 8124) * 100 = 45.54
Error rate in RFC : (3624 / 5124) * 100 = 70.73
```

- As we can see above the error rate for ARBC is less implying more accuracy. But the time taken for the algorithm to run is high, and the scans happening over the transaction set is very higher that the RFC.