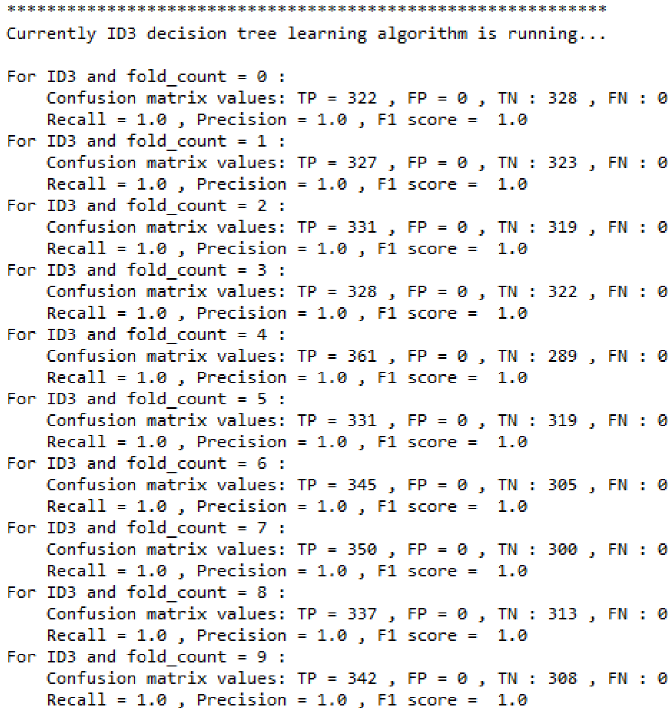
Project 1 report:

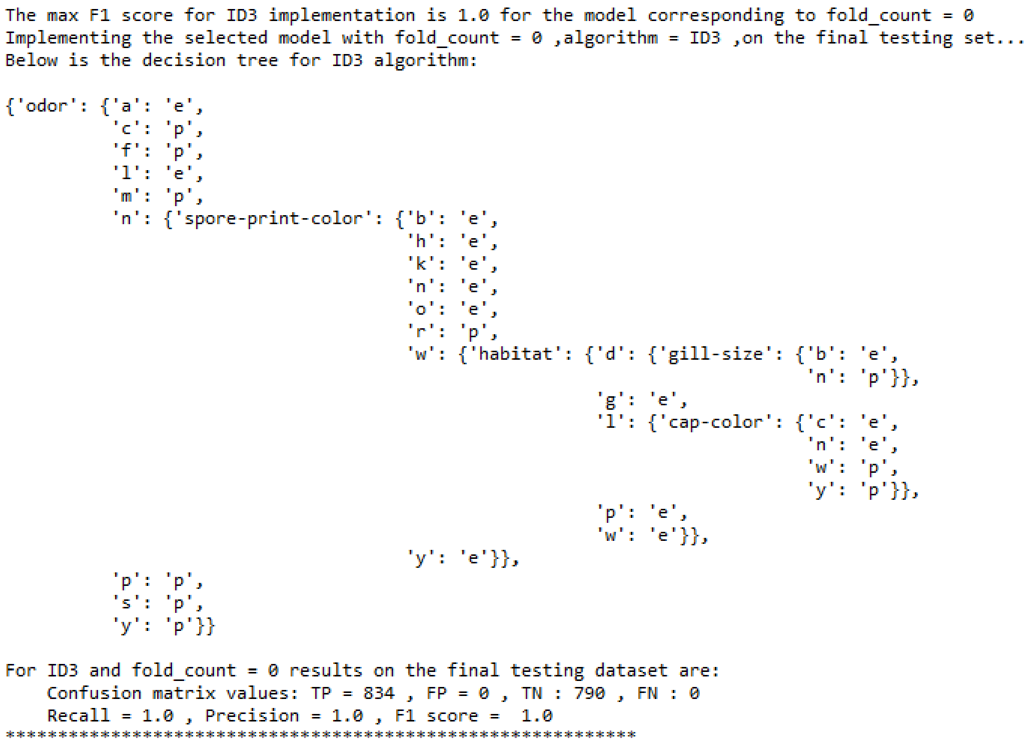
- As part of Project 1, for course CIS6930 – Applied Machine Learning using Python, Decision Trees using ID3 & C4.5 Algorithms is implemented.

- The implemented program follows the below steps:
    - Import the necessary python packages
    - Import the datasets and define the features and the target labels.
    - Functions implemented:
        - entropy: calculates the entropy of a dataset
        - information_gain: calculates the information gain of a dataset w.r.t. a feature
        - gain_ratio: calculates the gain ratio of a dataset w.r.t. a feature
        - create_decision_tree: this function creates a decision tree taking the input data("data" parameter) and following the algorithm specified("algorithm" parameter).
        - predict: it predicts target value for a new/unseen test data instance.
        - test: it runs the test data("data" parameter) on the decision tree & outputs the performance measures - precision, recall, F1_score,TP,FP,TN,FN
        - main() – Below are the steps performed,
            - 10-fold cross validation is performed using the 10 files that were provided.
            - A best model is selected on basis of the max F1 score. This model is then run on the final testing dataset.
            - Reports are printed in the following format.

Output for ID3 algorithm on IPython console

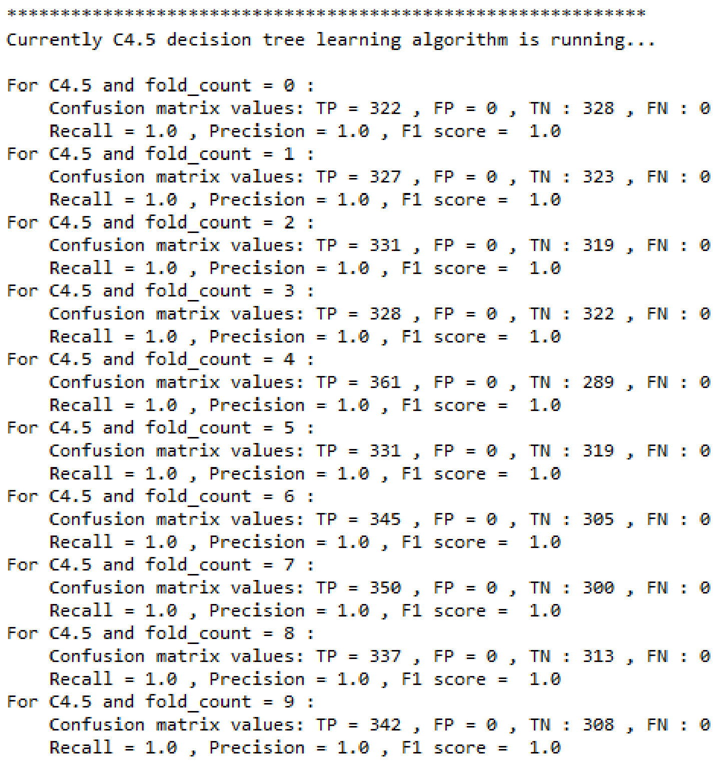


```
************************************************************
Currently ID3 decision tree learning algorithm is running...

For ID3 and fold_count = 0 :
    Confusion matrix values: TP = 322 , FP = 0 , TN : 328 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 1 :
    Confusion matrix values: TP = 327 , FP = 0 , TN : 323 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 2 :
    Confusion matrix values: TP = 331 , FP = 0 , TN : 319 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 3 :
    Confusion matrix values: TP = 328 , FP = 0 , TN : 322 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 4 :
    Confusion matrix values: TP = 361 , FP = 0 , TN : 289 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 5 :
    Confusion matrix values: TP = 331 , FP = 0 , TN : 319 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 6 :
    Confusion matrix values: TP = 345 , FP = 0 , TN : 305 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 7 :
    Confusion matrix values: TP = 350 , FP = 0 , TN : 300 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 8 :
    Confusion matrix values: TP = 337 , FP = 0 , TN : 313 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For ID3 and fold_count = 9 :
    Confusion matrix values: TP = 342 , FP = 0 , TN : 308 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
```

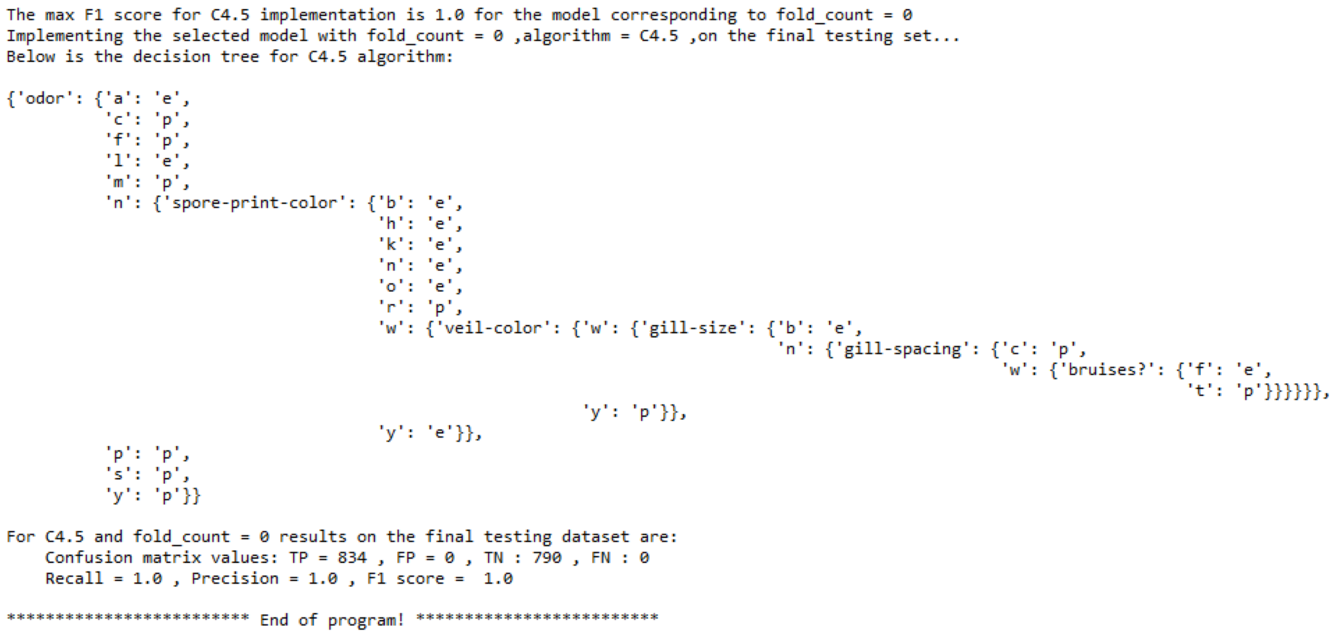


```
The max F1 score for ID3 implementation is 1.0 for the model corresponding to fold_count = 0
Implementing the selected model with fold_count = 0 ,algorithm = ID3 ,on the final testing set...
Below is the decision tree for ID3 algorithm:

{'odor': {'a': 'e',
          'c': 'p',
          'f': 'p',
          'l': 'e',
          'm': 'p',
          'n': {'spore-print-color': {'b': 'e',
                                      'h': 'e',
                                      'k': 'e',
                                      'n': 'e',
                                      'o': 'e',
                                      'r': 'p',
                                      'w': {'habitat': {'d': {'gill-size': {'b': 'e',
                                                                            'n': 'p'}},
                                                        'g': 'e',
                                                        'l': {'cap-color': {'c': 'e',
                                                                            'n': 'e',
                                                                            'w': 'p',
                                                                            'y': 'p'}},
                                                        'p': 'e',
                                                        'w': 'e'}},
                                      'y': 'e'}},
          'p': 'p',
          's': 'p',
          'y': 'p'}}

For ID3 and fold_count = 0 results on the final testing dataset are:
    Confusion matrix values: TP = 834 , FP = 0 , TN : 790 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
*************************************************************
```

Output for C4.3 algorithm on IPython console



```
*************************************************************
Currently C4.5 decision tree learning algorithm is running...

For C4.5 and fold_count = 0 :
    Confusion matrix values: TP = 322 , FP = 0 , TN : 328 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 1 :
    Confusion matrix values: TP = 327 , FP = 0 , TN : 323 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 2 :
    Confusion matrix values: TP = 331 , FP = 0 , TN : 319 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 3 :
    Confusion matrix values: TP = 328 , FP = 0 , TN : 322 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 4 :
    Confusion matrix values: TP = 361 , FP = 0 , TN : 289 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 5 :
    Confusion matrix values: TP = 331 , FP = 0 , TN : 319 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 6 :
    Confusion matrix values: TP = 345 , FP = 0 , TN : 305 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 7 :
    Confusion matrix values: TP = 350 , FP = 0 , TN : 300 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 8 :
    Confusion matrix values: TP = 337 , FP = 0 , TN : 313 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
For C4.5 and fold_count = 9 :
    Confusion matrix values: TP = 342 , FP = 0 , TN : 308 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0
```



```
The max F1 score for C4.5 implementation is 1.0 for the model corresponding to fold_count = 0
Implementing the selected model with fold_count = 0 ,algorithm = C4.5 ,on the final testing set...
Below is the decision tree for C4.5 algorithm:

{'odor': {'a': 'e',
          'c': 'p',
          'f': 'p',
          'l': 'e',
          'm': 'p',
          'n': {'spore-print-color': {'b': 'e',
                                      'h': 'e',
                                      'k': 'e',
                                      'n': 'e',
                                      'o': 'e',
                                      'r': 'p',
                                      'w': {'veil-color': {'w': {'gill-size': {'b': 'e',
                                                                               'n': {'gill-spacing': {'c': 'p',
                                                                                                      'w': {'bruises?': {'f': 'e',
                                                                                                                         't': 'p'}}}}}},
                                                           'y': 'p'}},
                                      'y': 'e'}},
          'p': 'p',
          's': 'p',
          'y': 'p'}}

For C4.5 and fold_count = 0 results on the final testing dataset are:
    Confusion matrix values: TP = 834 , FP = 0 , TN : 790 , FN : 0
    Recall = 1.0 , Precision = 1.0 , F1 score =  1.0

************************* End of program! *************************
```

- Test results on the Mushroom data for training (10-fold cross validation)) datasets.
  - The program was tested on each of the 10 cross validation files provided against the remaining 9 files.
  - For every fold count the recall and precision as seen in the above figure is equal to 1 for both ID3 and C4.3 algorithms. And hence the F1 score for every run equals 1.
  - F1 scores are stored in a list. Score and index of the max value in the F1_scores_list is computed. Using this the best model is selected.

- Test results on the Mushroom dataset for final testing dataset.
  - Best model selected in the above-mentioned step is utilized on the final testing set.
  - In this case since the first model itself gives the best score of 1 in case of both ID3 and C4.3 algorithms, the first model is selected to run on the final testing set.
  - As seen from the above figure, even for the final testing dataset the computed precision, accuracy and F1 score is equal to 1 for both ID3 and C4.3.