Ensembles

January 3, 2021

1 Mustererkennung/Machine Learning - Assignment 6

[68]:
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split
import numpy as np
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
import itertools
import random
import math
from pprint import pprint
import seaborn as sns
from sklearn import metrics
```

###Load the spam dataset:

[69]:
```
data = np.array(pd.read_csv('spambase.data', header=None))
feature_names=['word_freq_make',
'word_freq_address',
'word_freq_all',
'word_freq_3d',
'word_freq_our',
'word_freq_over',
'word_freq_remove',
'word_freq_internet',
'word_freq_order',
'word_freq_mail',
'word_freq_receive',
'word_freq_will',
'word_freq_people',
'word_freq_report',
'word_freq_addresses',
'word_freq_free',
'word_freq_business',
'word_freq_email',
```

```
'word_freq_you',
'word_freq_credit',
'word_freq_your',
'word_freq_font',
'word_freq_000',
'word_freq_money',
'word_freq_hp',
'word_freq_hpl',
'word_freq_george',
'word_freq_650',
'word_freq_lab',
'word_freq_labs',
'word_freq_telnet',
'word_freq_857',
'word_freq_data',
'word_freq_415',
'word_freq_85',
'word_freq_technology',
'word_freq_1999',
'word_freq_parts',
'word_freq_pm',
'word_freq_direct',
'word_freq_cs',
'word_freq_meeting',
'word_freq_original',
'word_freq_project',
'word_freq_re',
'word_freq_edu',
'word_freq_table',
'word_freq_conference',
'char_freq_;',
'char_freq_(',
'char_freq_[',
'char_freq_!',
'char_freq_$',
'char_freq_#',
'capital_run_length_average',
'capital_run_length_longest',
'capital_run_length_total']
```

[70]:
```
X = data[:,:-1] # features
#X_train = np.
 ↪array([[0,1,1],[1,0,0],[0,1,1],[1,1,1],[0,1,0],[0,0,0],[1,1,0],[0,1,0],[0,0,0],[1,1,0]])␣
 ↪# features
y = data[:,-1] # Last column is label
#y_train = np.array([1,0,1,1,0,1,0,0,1,1]) # Last column is label
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0,␣
 ↪shuffle=True, stratify=y)
total_train_samples=y_train.size #total samples
```

[71]:
```
#confusion matrix plotting
def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):
    """
    This function prints and plots the confusion matrix.
    Normalization can be applied by setting `normalize=True`.
    """
    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    print(cm)

    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    fmt = '.2f' if normalize else 'd'
    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, format(cm[i, j], fmt),
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

    plt.ylabel('True label')
    plt.xlabel('Predicted label')
    plt.tight_layout()
```

[72]:
```
# Create a leaf node value based on majority vote
def to_terminal(outcomes):
    return np.round(np.mean(outcomes))
```

[73]:
```
#This gives best split node for decision tree uisng gini impurity
def find_split(X_train, y_train):
    class_values = np.unique(y_train)
    X,Y=X_train,y_train
```

```
    #parameter initialization
    min_gini=999
    z=999
    col=999
    gini_total=0
    best_tree_childs=()
    #for each column
    for j in range(X_train.shape[-1]):
        #indices=np.argsort(X_train[:, j])
        #X,Y=X_train[indices],y_train[indices]
        #if we take unique value for j-th column into account then it reduces␣
↪the time computation instead of individual value
        unique_values_to_split=np.unique(X[:,j])
        #for each unique value of j-th column
        for i in unique_values_to_split:
            curr_z=i
            #splitted data based on current value of j-th column
            y_left,y_right=Y[X[:, j] < curr_z],Y[X[:, j] >= curr_z]
            c1=to_terminal(y_left)
            #print('c1',c1)
            c2=to_terminal(y_right)
            childs=(y_left,y_right)
            #weighted average gini impurity value for both groups left and right␣
↪child
            gini=gini_index(childs, class_values)
            #print(j,i,gini)
            #best gini value which is minimum among all values
            if min_gini>gini:
                min_gini=gini
                z=curr_z
                c1_temp=c1
                c2_temp=c2
                col=j
                best_tree_childs=childs
    #parent node gini impuirty calculation
    for class_val in class_values:
        p = [y for y in y_train].count(class_val) / y_train.size
        gini_total += p * p
    gini_parent=1-gini_total
    #gini gain if we split parent into left child and right child which is␣
↪basically will be used to calculate feature importance
    gini_gain=(gini_parent-min_gini)*y_train.size / total_train_samples
    #print('X'+str(j_temp)+' cutoff: '+str(z))
    return z,col,gini_gain,best_tree_childs
```

[74]:
```
# Calculate the Gini index for a split dataset
def gini_index(childs, classes):
    # count all samples at split point
    n_instances = float(sum([child_node.size for child_node in childs]))
    # sum weighted Gini index for each childs
    gini=0
    for child_node in childs:
        size = float(child_node.size)
        # avoid divide by zero
        if size == 0:
            continue
        score = 0.0
        # score the group based on the score for each class
        for class_val in classes:
            p = [row for row in child_node].count(class_val) / size
            score += p * p
        # weight the group score by its relative size
        gini += (1.0 - score) * (size / n_instances)
    return gini
```

[75]:
```
class DecisionTreeClassifier():
    #initialiazation with math depth(max_depth) and minimum no of asmples for␣
 ↪leaf(t)
    def __init__(self, max_depth=5,t=1):
        self.max_depth = max_depth
        self.t=t
    #fit method with parameters dataset, parent node information and depth of␣
 ↪the tree
    def fit(self, x, y, par_node={}, depth=0):
            t=self.t
            cutoff,col, gini_gain,best_tree_childs = find_split(x, y)     # find␣
 ↪best split given a gini impurity
            #best split information assignment for tree node
            par_node = {'col': 'X'+str(col), 'index_col':col,'cutoff':
 ↪cutoff,'gini_gain':gini_gain }
            y_left,y_right=best_tree_childs
            #print('y_left:',y_left.size,' y_right ',y_right.size)
            #if any of the child samples are zero then there is no need of␣
 ↪further split
            if y_left.size==0 or y_right.size==0:
                par_node['left'] = par_node['right'] = to_terminal(list(y_left)␣
 ↪+ list(y_right))
                return par_node
            # trif ee depth is greater than equals to max depth then we can stop␣
 ↪here
            if depth >= self.max_depth:
```

```
                par_node['left'], par_node['right'] = to_terminal(y_left),␣
↪to_terminal(y_right)
                return par_node
            #stop if leaf nodes have less samples than the specified value␣
↪otherwise split further
            if y_left.size<=t:
                par_node['left']=np.round(np.mean(y_left))
            else:
                par_node['left'] = self.fit(x[x[:, col] < cutoff], y_left, {},␣
↪depth+1)

            if y_right.size<=t:
                par_node['right']=np.round(np.mean(y_right))
            else:
                par_node['right'] = self.fit(x[x[:, col] >= cutoff], y_right,␣
↪{}, depth+1)
            self.trees = par_node
            return par_node
```

[76]:
```
#prediction using trained parameters of tree
def predict( m,x):
    tree = m
    results = np.array([0]*len(x))
    for i, c in enumerate(x):
        results[i] = get_prediction(m,c)
    return results

def get_prediction(m, row):
    cur_layer = m
    while cur_layer['cutoff'] is not None:
        if row[cur_layer['index_col']] <= cur_layer['cutoff']:
            if isinstance(cur_layer['left'], dict):
                cur_layer = cur_layer['left']
            else:
                return cur_layer['left']
        else:
            if isinstance(cur_layer['right'], dict):
                cur_layer = cur_layer['right']
            else:
                return cur_layer['right']
        #print('cutoff',cur_layer['cutoff'])
```

[77]:
```
def plot_model_report(y_test,y_pred):
    # Generate a classification report
    cm_plot_labels = ['Not Spam', 'Spam']
    # For this to work we need y_pred as binary labels not as probabilities
    #y_pred_binary = np.where(predictions > 0.5, 1, 0)
```

```
report = classification_report(y_test, y_pred, target_names=cm_plot_labels)

print(report)

# argmax returns the index of the max value in a row
cm = confusion_matrix(y_test, y_pred)

plot_confusion_matrix(cm, cm_plot_labels, title='Confusion Matrix')
```

2 (a) Assume that classifying a genuine E-Mail as spam is ten times worse than classifying spam as genuine. How would you change the design of your decision tree?

Answer: I guess differences in accuracies between class non spam and class spam come from the class_weight parameter you have used. Class spam will benefit from this overweighting towards class non spam. You could try to play on this parameter to re-balance your results in class non spam and class spam.

other solution could be, we can increase or decrese the minimum number of samples at leaf node by keeping max depth constant or we can increase or decrese max depth by keeping minimum number of samples as constant to overcome this problem.

3 (b) Use your tree to analyze feature importance. Plot the difference between the top 5 features (check spambase.names to check what features those belong to).

Answer: As part of feature importance calculation, We have calculated the gini gain at every best split as follows :

N_t / N * (impurity - N_t_R / N_t * right_impurity - N_t_L / N_t * left_impurity)

where N is the total number of samples, N_t is the number of samples at the current node, N_t_L is the number of samples in the left child, and N_t_R is the number of samples in the right child.

after training the model we have calculated the feature importance as sum of all gini gain where feature X was the cause of split.

[78]:
```
#feature importance calculation
class f_importance():
    def __init__(self, features=1):
        self.res = {k:0 for k in range(features)}
    #
    def f_importance_calculation(self, root):

            if root:
```

```
                self.res[root['index_col']]+=root['gini_gain']
                if isinstance(root['left'], dict):
                    self.f_importance_calculation(root['left'])
                if isinstance(root['right'], dict):
                    self.f_importance_calculation( root['right'])
            return self.res
```

[79]:
```
def call_decision_tree():
    #decision tree with max depth  as 10 amd minimum samples at leaf node as 1
    clf = DecisionTreeClassifier(max_depth=10, t=1)
    tree = clf.fit(X_train, y_train)
    pprint(tree) #to print tree

    #prediction on test set
    y_pred=predict(tree,X_test)

    #feature imprtance calculation
    feature_importance=f_importance(features=X_train.shape[-1]).
 ↪f_importance_calculation(tree)
    feat_imp_list=[]
    for k, v in zip(feature_names,feature_importance.items()):
        print(k,v)
        feat_imp_list.append((k,v[1]))
    Sorted_feature=[[key,value] for key, value in␣
 ↪sorted(feat_imp_list,reverse=True, key=lambda item: item[1])]
    feat=[]
    score=[]
    for row in Sorted_feature:
        feat.append(row[0])
        score.append(row[1])


    #sns.set_style('darkgrid')

    sns.barplot(score[0:5],feat[0:5])
    plt.xlabel('Feature importance value')
    plt.ylabel('Features')
    plt.show()

    plot_model_report(y_test,y_pred) #model report plotting
```

[80]:
```
call_decision_tree()
```

```
/home/suresh/.local/lib/python3.8/site-packages/numpy/core/fromnumeric.py:3372:
RuntimeWarning: Mean of empty slice.
  return _methods._mean(a, axis=axis, dtype=dtype,
/home/suresh/.local/lib/python3.8/site-packages/numpy/core/_methods.py:170:
```

```
RuntimeWarning: invalid value encountered in double_scalars
  ret = ret.dtype.type(ret / rcount)
{'col': 'X51',
 'cutoff': 0.079,
 'gini_gain': 0.1573079330035172,
 'index_col': 51,
 'left': {'col': 'X6',
          'cutoff': 0.05,
          'gini_gain': 0.036940589137910096,
          'index_col': 6,
          'left': {'col': 'X23',
                   'cutoff': 0.02,
                   'gini_gain': 0.013600533060163808,
                   'index_col': 23,
                   'left': {'col': 'X15',
                            'cutoff': 0.2,
                            'gini_gain': 0.0043473035328583055,
                            'index_col': 15,
                            'left': {'col': 'X52',
                                     'cutoff': 0.182,
                                     'gini_gain': 0.004177741889223485,
                                     'index_col': 52,
                                     'left': {'col': 'X3',
                                              'cutoff': 0.44,
                                              'gini_gain':
0.0010432403678429902,
                                              'index_col': 3,
                                              'left': {'col': 'X27',
                                                       'cutoff': 8.33,
                                                       'gini_gain':
0.0010458306605004967,
                                                       'index_col': 27,
                                                       'left': {'col': 'X24',
                                                                'cutoff': 0.13,
                                                                'gini_gain':
0.0009246440692747221,
                                                                'index_col': 24,
                                                                'left': {'col':
'X55',
'cutoff': 11.0,
'gini_gain': 0.002662657521746514,
'index_col': 55,
                                                                         'left':
{'col': 'X10',
  'cutoff': 0.29,
  'gini_gain': 0.0007089664779281465,
  'index_col': 10,
```

```
  'left': {'col': 'X7',
           'cutoff': 11.11,
           'gini_gain': 0.0005497897931703006,
           'index_col': 7,
           'left': 0.0,
           'right': 1.0},
  'right': {'col': 'X0',
            'cutoff': 0.26,
            'gini_gain': 0.0003864734299516908,
            'index_col': 0,
            'left': 1.0,
            'right': 0.0}},
'right': {'col': 'X4',
   'cutoff': 0.73,
   'gini_gain': 0.0028408620939035937,
   'index_col': 4,
   'left': {'col': 'X35',
            'cutoff': 0.24,
            'gini_gain': 0.00209379471663033,
            'index_col': 35,
            'left': 0.0,
            'right': 1.0},
   'right': {'col': 'X26',
             'cutoff': 0.39,
             'gini_gain': 0.0011690821256038644,
             'index_col': 26,
             'left': 1.0,
             'right': 0.0}}},
                                                                    'right': {'col':
'X10',
'cutoff': 1.36,
'gini_gain': 0.001143808643696067,
'index_col': 10,
'left': {'col': 'X56',
   'cutoff': 2510.0,
   'gini_gain': 0.00018835224813089504,
   'index_col': 56,
   'left': {'col': 'X48',
            'cutoff': 0.28300000000000003,
            'gini_gain': 2.4684055049824003e-05,
            'index_col': 48,
            'left': 0.0,
            'right': 0.0},
   'right': {'col': 'X0',
             'cutoff': 0.06,
             'gini_gain': 0.0003864734299516908,
             'index_col': 0,
             'left': 0.0,
```

```
            'right': 1.0}},
'right': {'col': 'X0',
    'cutoff': 0.0,
    'gini_gain': 0.0,
    'index_col': 0,
    'left': 1.0,
    'right': 1.0}}},
                                                    'right': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain':
0.0,
                                                              'index_col': 0,
                                                              'left': 1.0,
                                                              'right': 1.0}},
                                                 'right': {'col': 'X0',
                                                           'cutoff': 0.0,
                                                           'gini_gain': 0.0,
                                                           'index_col': 0,
                                                           'left': 1.0,
                                                           'right': 1.0}},
                                          'right': {'col': 'X54',
                                                    'cutoff': 3.272,
                                                    'gini_gain':
0.002591716396703609,
                                                    'index_col': 54,
                                                    'left': {'col': 'X12',
                                                             'cutoff': 5.55,
                                                             'gini_gain':
0.0004665286404416835,
                                                             'index_col': 12,
                                                             'left': {'col': 'X17',
                                                                      'cutoff': 4.83,
                                                                      'gini_gain':
0.0005383022774327124,
                                                                      'index_col':
17,
                                                                      'left': {'col':
'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 0.0,
'right': 0.0},
                                                                      'right': 1.0},
                                                             'right': 1.0},
                                                    'right': {'col': 'X33',
                                                              'cutoff': 0.13,
                                                              'gini_gain':
```

```
0.0006365444728616088,
                                                    'index_col': 33,
                                                    'left': {'col': 'X0',
                                                             'cutoff': 0.0,
                                                             'gini_gain':
0.0,
                                                             'index_col':
0,
                                                             'left': 1.0,
                                                             'right': 1.0},
                                                    'right': {'col': 'X0',
                                                              'cutoff':
0.19,
                                                              'gini_gain':
0.0003864734299516908,
                                                              'index_col':
0,
                                                              'left':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 0.0,
'right': 0.0},
                                                              'right':
1.0}}}},
                            'right': {'col': 'X4',
                                      'cutoff': 1.17,
                                      'gini_gain': 0.003098814229249011,
                                      'index_col': 4,
                                      'left': {'col': 'X44',
                                               'cutoff': 0.18,
                                               'gini_gain':
0.0012574738576443607,
                                               'index_col': 44,
                                               'left': {'col': 'X24',
                                                        'cutoff': 0.27,
                                                        'gini_gain':
0.0013564143853998932,
                                                        'index_col': 24,
                                                        'left': {'col': 'X15',
                                                                 'cutoff': 0.51,
                                                                 'gini_gain':
0.0012054467271858567,
                                                                 'index_col':
15,
                                                                 'left': {'col':
'X4',
```

```
'cutoff': 0.52,
'gini_gain': 0.000452976974716105,
'index_col': 4,
'left': {'col': 'X50',
   'cutoff': 0.059000000000000004,
   'gini_gain': 0.000527009222661397,
   'index_col': 50,
   'left': {'col': 'X0',
            'cutoff': 0.0,
            'gini_gain': 0.0,
            'index_col': 0,
            'left': 1.0,
            'right': 1.0},
   'right': 0.0},
'right': {'col': 'X2',
    'cutoff': 0.84,
    'gini_gain': 0.0003864734299516908,
    'index_col': 2,
    'left': {'col': 'X0',
             'cutoff': 0.0,
             'gini_gain': 0.0,
             'index_col': 0,
             'left': 0.0,
             'right': 0.0},
    'right': 1.0}},                                                    'right':
{'col': 'X9',
'cutoff': 0.3,
'gini_gain': 0.0006020066889632112,
'index_col': 9,
'left': {'col': 'X15',
    'cutoff': 4.0,
    'gini_gain': 0.00044851258581235645,
    'index_col': 15,
    'left': {'col': 'X13',
             'cutoff': 0.52,
             'gini_gain': 0.00043393507924400446,
             'index_col': 13,
             'left': 0.0,
             'right': 1.0},
    'right': {'col': 'X15',
              'cutoff': 7.69,
              'gini_gain': 0.00030917874396135266,
              'index_col': 15,
              'left': 1.0,
              'right': 0.0}},
'right': {'col': 'X0',
     'cutoff': 0.0,
```

```
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 1.0,
     'right': 1.0}}},
                                                    'right': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain':
0.0,
                                                              'index_col':
0,
                                                              'left': 0.0,
                                                              'right':
0.0}},
                                                  'right': {'col': 'X3',
                                                            'cutoff': 7.07,
                                                            'gini_gain':
0.0005626598465473155,
                                                            'index_col': 3,
                                                            'left': {'col': 'X0',
                                                                     'cutoff': 0.0,
                                                                     'gini_gain':
0.0,
                                                                     'index_col':
0,
                                                                     'left': 0.0,
                                                                     'right': 0.0},
                                                            'right': 1.0}},
                                                'right': {'col': 'X44',
                                                          'cutoff': 0.95,
                                                          'gini_gain':
0.0013094629156010225,
                                                          'index_col': 44,
                                                          'left': {'col': 'X12',
                                                                   'cutoff': 0.8,
                                                                   'gini_gain':
0.0005456095481670925,
                                                                   'index_col': 12,
                                                                   'left': {'col': 'X0',
                                                                            'cutoff': 0.0,
                                                                            'gini_gain':
0.0,
                                                                            'index_col':
0,
                                                                            'left': 1.0,
                                                                            'right': 1.0},
                                                                   'right': 0.0},
                                                          'right': {'col': 'X0',
                                                                    'cutoff': 0.0,
```

```
                                                    'gini_gain': 0.0,
                                                    'index_col': 0,
                                                    'left': 0.0,
                                                    'right': 0.0}}}},
                    'right': {'col': 'X24',
                              'cutoff': 0.17,
                              'gini_gain': 0.0034168294792912084,
                              'index_col': 24,
                              'left': {'col': 'X52',
                                       'cutoff': 0.013999999999999999,
                                       'gini_gain': 0.0027529427774375718,
                                       'index_col': 52,
                                       'left': {'col': 'X54',
                                                'cutoff': 4.333,
                                                'gini_gain':
0.001670117322291235,
                                                'index_col': 54,
                                                'left': {'col': 'X18',
                                                         'cutoff': 12.5,
                                                         'gini_gain':
0.0005383022774327124,
                                                         'index_col': 18,
                                                         'left': {'col': 'X0',
                                                                  'cutoff': 0.0,
                                                                  'gini_gain':
0.0,
                                                                  'index_col': 0,
                                                                  'left': 0.0,
                                                                  'right': 0.0},
                                                         'right': 1.0},
                                                'right': {'col': 'X11',
                                                          'cutoff': 1.31,
                                                          'gini_gain':
0.0004968944099378883,
                                                          'index_col': 11,
                                                          'left': {'col': 'X0',
                                                                   'cutoff': 0.0,
                                                                   'gini_gain':
0.0,
                                                                   'index_col':
0,
                                                                   'left': 1.0,
                                                                   'right': 1.0},
                                                          'right': 0.0}},
                                       'right': {'col': 'X26',
                                                 'cutoff': 0.16,
                                                 'gini_gain':
0.0009782608695652166,
```

```
                                                    'index_col': 26,
                                                    'left': {'col': 'X7',
                                                             'cutoff': 0.73,
                                                             'gini_gain':
0.0005203515263644802,
                                                             'index_col': 7,
                                                             'left': {'col': 'X56',
                                                                      'cutoff':
53.0,
                                                                      'gini_gain':
0.0005429755057716249,
                                                                      'index_col':
56,
                                                                      'left': 0.0,
                                                                      'right':
{'col': 'X0',
'cutoff': 0.36,
'gini_gain': 0.00027725267800882266,
'index_col': 0,
'left': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 1.0,
     'right': 1.0},
'right': {'col': 'X0',
      'cutoff': 0.36,
      'gini_gain': 0.0,
      'index_col': 0,
      'left': 0.0,
      'right': 0.0}}},
                                                             'right': 0.0},
                                                    'right': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain': 0.0,
                                                              'index_col': 0,
                                                              'left': 0.0,
                                                              'right': 0.0}}},
                                'right': {'col': 'X0',
                                          'cutoff': 0.0,
                                          'gini_gain': 0.0,
                                          'index_col': 0,
                                          'left': 0.0,
                                          'right': 0.0}}},
          'right': {'col': 'X26',
                    'cutoff': 0.16,
                    'gini_gain': 0.004669228309547926,
                    'index_col': 26,
```

```
                    'left': {'col': 'X24',
                             'cutoff': 0.21,
                             'gini_gain': 0.0015375386077453087,
                             'index_col': 24,
                             'left': {'col': 'X7',
                                      'cutoff': 2.22,
                                      'gini_gain': 0.00103754940711463,
                                      'index_col': 7,
                                      'left': {'col': 'X45',
                                               'cutoff': 0.29,
                                               'gini_gain':
0.001075977162933675,
                                               'index_col': 45,
                                               'left': {'col': 'X32',
                                                        'cutoff': 1.76,
                                                        'gini_gain':
0.0005530723734705912,
                                                        'index_col': 32,
                                                        'left': {'col': 'X9',
                                                                 'cutoff': 1.78,
                                                                 'gini_gain':
0.00017713944369775417,
                                                                 'index_col': 9,
                                                                 'left': {'col':
'X55',
'cutoff': 6.0,
'gini_gain': 0.00011036789297658651,
'index_col': 55,
'left': {'col': 'X2',
   'cutoff': 0.25,
   'gini_gain': 0.00046376811594202875,
   'index_col': 2,
   'left': {'col': 'X0',
            'cutoff': 0.0,
            'gini_gain': 0.0,
            'index_col': 0,
            'left': 1.0,
            'right': 1.0},
   'right': 0.0},
'right': {'col': 'X0',
    'cutoff': 0.0,
    'gini_gain': 0.0,
    'index_col': 0,
    'left': 1.0,
    'right': 1.0}},
                                                                 'right':
{'col': 'X9',
'cutoff': 1.96,
```

```
'gini_gain': 0.0003864734299516908,
'index_col': 9,
'left': 0.0,
'right': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 1.0,
     'right': 1.0}},
                                                        'right': 0.0},
                                                'right': {'col': 'X0',
                                                          'cutoff': 0.0,
                                                          'gini_gain': 0.0,
                                                          'index_col': 0,
                                                          'left': 0.0,
                                                          'right': 0.0}},
                                         'right': {'col': 'X0',
                                                   'cutoff': 0.0,
                                                   'gini_gain': 0.0,
                                                   'index_col': 0,
                                                   'left': 0.0,
                                                   'right': 0.0}},
                                 'right': {'col': 'X4',
                                           'cutoff': 0.68,
                                           'gini_gain': 0.0009275362318840581,
                                           'index_col': 4,
                                           'left': {'col': 'X0',
                                                    'cutoff': 0.0,
                                                    'gini_gain': 0.0,
                                                    'index_col': 0,
                                                    'left': 0.0,
                                                    'right': 0.0},
                                           'right': {'col': 'X4',
                                                     'cutoff': 0.8,
                                                     'gini_gain':
0.00017391304347826064,
                                                     'index_col': 4,
                                                     'left': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain': 0.0,
                                                              'index_col': 0,
                                                              'left': 0.0,
                                                              'right': 0.0},
                                                     'right': {'col': 'X0',
                                                               'cutoff': 0.0,
                                                               'gini_gain': 0.0,
                                                               'index_col': 0,
                                                               'left': 1.0,
```

```
                                                        'right': 1.0}}}},
                    'right': {'col': 'X0',
                              'cutoff': 0.0,
                              'gini_gain': 0.0,
                              'index_col': 0,
                              'left': 0.0,
                              'right': 0.0}}},
 'right': {'col': 'X52',
           'cutoff': 0.013000000000000001,
           'gini_gain': 0.046520469486797486,
           'index_col': 52,
           'left': {'col': 'X54',
                    'cutoff': 2.755,
                    'gini_gain': 0.026507781617062208,
                    'index_col': 54,
                    'left': {'col': 'X6',
                             'cutoff': 0.18,
                             'gini_gain': 0.012226442029898955,
                             'index_col': 6,
                             'left': {'col': 'X51',
                                      'cutoff': 0.8140000000000001,
                                      'gini_gain': 0.006907894254644194,
                                      'index_col': 51,
                                      'left': {'col': 'X4',
                                               'cutoff': 0.33,
                                               'gini_gain':
0.0031959751598086617,
                                               'index_col': 4,
                                               'left': {'col': 'X15',
                                                        'cutoff': 1.43,
                                                        'gini_gain':
0.002093023665501414,
                                                        'index_col': 15,
                                                        'left': {'col': 'X8',
                                                                 'cutoff': 0.93,
                                                                 'gini_gain':
0.0005239510897192364,
                                                                 'index_col': 8,
                                                                 'left': {'col':
'X9',
'cutoff': 7.55,
'gini_gain': 0.0005281094317011106,
'index_col': 9,
'left': {'col': 'X22',
   'cutoff': 2.24,
   'gini_gain': 0.0005323174749815887,
   'index_col': 22,
   'left': {'col': 'X12',
```

```
            'cutoff': 1.15,
            'gini_gain': 0.00038362679833277226,
            'index_col': 12,
            'left': 0.0,
            'right': 0.0},
   'right': 1.0},
'right': 1.0},
                                                                  'right': 1.0},
                                                        'right': {'col': 'X55',
                                                                  'cutoff': 7.0,
                                                                  'gini_gain':
0.0010118577075098815,
                                                                  'index_col':
55,
                                                                  'left':
{'col': 'X2',
'cutoff': 1.44,
'gini_gain': 0.00046376811594202875,
'index_col': 2,
'left': {'col': 'X0',
    'cutoff': 0.0,
    'gini_gain': 0.0,
    'index_col': 0,
    'left': 0.0,
    'right': 0.0},
'right': 1.0},
                                                                  'right':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 1.0,
'right': 1.0}}},
                                                  'right': {'col': 'X26',
                                                            'cutoff': 0.38,
                                                            'gini_gain':
0.0029644268774703534,
                                                            'index_col': 26,
                                                            'left': {'col': 'X20',
                                                                     'cutoff':
0.45,
                                                                     'gini_gain':
0.002740887132191482,
                                                                     'index_col':
20,
                                                                     'left':
{'col': 'X7',
'cutoff': 0.6,
```

```
'gini_gain': 0.0005410628019323668,
'index_col': 7,
'left': {'col': 'X0',
    'cutoff': 0.0,
    'gini_gain': 0.0,
    'index_col': 0,
    'left': 0.0,
    'right': 0.0},
'right': 1.0},
                                                                  'right':
{'col': 'X17',
'cutoff': 0.36,
'gini_gain': 0.000913629044063825,
'index_col': 17,
'left': {'col': 'X49',
     'cutoff': 0.147,
     'gini_gain': 0.0014229249011857715,
     'index_col': 49,
     'left': {'col': 'X56',
              'cutoff': 42.0,
              'gini_gain': 0.0007453416149068321,
              'index_col': 56,
              'left': 0.0,
              'right': 1.0},
     'right': {'col': 'X0',
               'cutoff': 0.0,
               'gini_gain': 0.0,
               'index_col': 0,
               'left': 0.0,
               'right': 0.0}},
'right': {'col': 'X4',
      'cutoff': 0.59,
      'gini_gain': 0.0005475040257648954,
      'index_col': 4,
      'left': 0.0,
      'right': {'col': 'X0',
                'cutoff': 0.0,
                'gini_gain': 0.0,
                'index_col': 0,
                'left': 1.0,
                'right': 1.0}}}},
                                                        'right': {'col': 'X0',
                                                                  'cutoff':
0.0,
                                                                  'gini_gain':
0.0,
                                                                  'index_col':
0,
```

```
                                                                    'left': 0.0,
                                                                    'right':
0.0}}},
                                        'right': {'col': 'X56',
                                                  'cutoff': 15.0,
                                                  'gini_gain':
0.003066289003327658,
                                                  'index_col': 56,
                                                  'left': {'col': 'X15',
                                                           'cutoff': 1.4,
                                                           'gini_gain':
0.001073537305421365,
                                                           'index_col': 15,
                                                           'left': {'col': 'X51',
                                                                    'cutoff':
4.3469999999999995,
                                                                    'gini_gain':
0.0008783487044356597,
                                                                    'index_col':
51,
                                                                    'left':
{'col': 'X51',
'cutoff': 0.925,
'gini_gain': 0.0005019135453918068,
'index_col': 51,
'left': 1.0,
'right': {'col': 'X51',
     'cutoff': 1.754,
     'gini_gain': 6.901311249137372e-05,
     'index_col': 51,
     'left': {'col': 'X0',
              'cutoff': 0.0,
              'gini_gain': 0.0,
              'index_col': 0,
              'left': 0.0,
              'right': 0.0},
     'right': {'col': 'X51',
               'cutoff': 2.272,
               'gini_gain': 0.00048309178743961335,
               'index_col': 51,
               'left': 1.0,
               'right': 0.0}}},
                                                                    'right':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 1.0,
```

```
'right': 1.0}},
                                                                  'right': {'col': 'X0',
                                                                            'cutoff':
0.0,
                                                                            'gini_gain':
0.0,
                                                                            'index_col':
0,
                                                                            'left': 1.0,
                                                                            'right':
1.0}},
                                                          'right': {'col': 'X11',
                                                                    'cutoff': 2.58,
                                                                    'gini_gain':
0.0012903945973098182,
                                                                    'index_col': 11,
                                                                    'left': {'col': 'X54',
                                                                             'cutoff':
1.2,
                                                                             'gini_gain':
0.0005951109658203482,
                                                                             'index_col':
54,
                                                                             'left':
{'col': 'X2',
'cutoff': 1.0,
'gini_gain': 0.0003864734299516908,
'index_col': 2,
'left': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 0.0,
     'right': 0.0},
'right': 1.0},
                                                                             'right':
{'col': 'X12',
'cutoff': 0.28,
'gini_gain': 0.0005305078553160381,
'index_col': 12,
'left': {'col': 'X24',
      'cutoff': 4.34,
      'gini_gain': 0.0005626598465473155,
      'index_col': 24,
      'left': {'col': 'X0',
               'cutoff': 0.0,
               'gini_gain': 0.0,
               'index_col': 0,
```

```
                'left': 1.0,
                'right': 1.0},
      'right': 0.0},
'right': 0.0}},
                                                          'right': {'col': 'X0',
                                                                    'cutoff':
0.0,
                                                                    'gini_gain':
0.0,
                                                                    'index_col':
0,
                                                                    'left': 0.0,
                                                                    'right':
0.0}}}},
                               'right': {'col': 'X45',
                                         'cutoff': 0.23,
                                         'gini_gain': 0.0010695132200274518,
                                         'index_col': 45,
                                         'left': {'col': 'X24',
                                                  'cutoff': 0.46,
                                                  'gini_gain':
0.0011100832562442157,
                                                  'index_col': 24,
                                                  'left': {'col': 'X0',
                                                           'cutoff': 0.0,
                                                           'gini_gain': 0.0,
                                                           'index_col': 0,
                                                           'left': 1.0,
                                                           'right': 1.0},
                                                  'right': {'col': 'X0',
                                                            'cutoff': 0.0,
                                                            'gini_gain': 0.0,
                                                            'index_col': 0,
                                                            'left': 0.0,
                                                            'right': 0.0}},
                                         'right': {'col': 'X4',
                                                   'cutoff': 1.67,
                                                   'gini_gain':
0.0004347826086956522,
                                                   'index_col': 4,
                                                   'left': {'col': 'X0',
                                                            'cutoff': 0.0,
                                                            'gini_gain': 0.0,
                                                            'index_col': 0,
                                                            'left': 0.0,
                                                            'right': 0.0},
                                                   'right': 1.0}}},
                     'right': {'col': 'X24',
```

```
                                'cutoff': 0.21,
                                'gini_gain': 0.006888702663029791,
                                'index_col': 24,
                                'left': {'col': 'X45',
                                         'cutoff': 0.56,
                                         'gini_gain': 0.003528676148213704,
                                         'index_col': 45,
                                         'left': {'col': 'X41',
                                                  'cutoff': 0.89,
                                                  'gini_gain':
0.003269214666727411,
                                                  'index_col': 41,
                                                  'left': {'col': 'X47',
                                                           'cutoff': 0.77,
                                                           'gini_gain':
0.0010255743154123976,
                                                           'index_col': 47,
                                                           'left': {'col': 'X48',
                                                                    'cutoff':
1.151,
                                                                    'gini_gain':
0.0009899010558082045,
                                                                    'index_col':
48,
                                                                    'left':
{'col': 'X12',
'cutoff': 0.55,
'gini_gain': 0.0006520122357384341,
'index_col': 12,
'left': {'col': 'X43',
    'cutoff': 2.04,
    'gini_gain': 0.0005124587996788213,
    'index_col': 43,
    'left': {'col': 'X51',
             'cutoff': 19.131,
             'gini_gain': 0.0005167115698006461,
             'index_col': 51,
             'left': 1.0,
             'right': 0.0},
    'right': 0.0},
'right': {'col': 'X12',
     'cutoff': 0.74,
     'gini_gain': 0.0008695652173913044,
     'index_col': 12,
     'left': {'col': 'X0',
              'cutoff': 0.0,
              'gini_gain': 0.0,
              'index_col': 0,
```

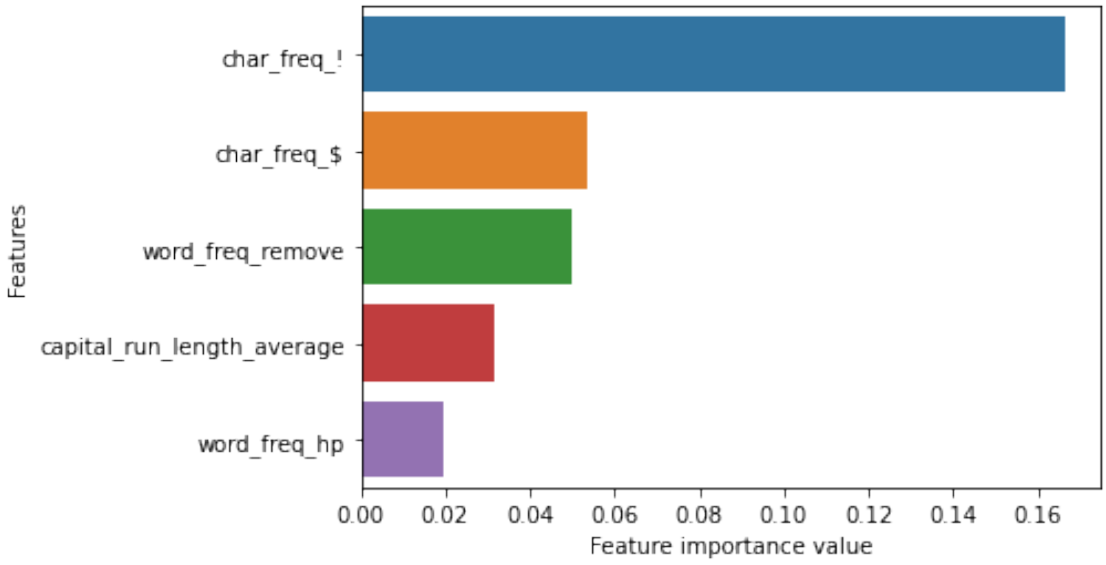
```
              'left': 0.0,
              'right': 0.0},
     'right': {'col': 'X0',
               'cutoff': 0.0,
               'gini_gain': 0.0,
               'index_col': 0,
               'left': 1.0,
               'right': 1.0}}},
                                                                  'right':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 0.0,
'right': 0.0}},
                                                        'right': {'col': 'X0',
                                                                  'cutoff':
2.32,
                                                                  'gini_gain':
0.0004347826086956522,
                                                                  'index_col':
0,
                                                                  'left':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 0.0,
'right': 0.0},
                                                                  'right':
1.0}},
                                                    'right': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain': 0.0,
                                                              'index_col': 0,
                                                              'left': 0.0,
                                                              'right': 0.0}},
                                          'right': {'col': 'X0',
                                                    'cutoff': 0.0,
                                                    'gini_gain': 0.0,
                                                    'index_col': 0,
                                                    'left': 0.0,
                                                    'right': 0.0}},
                                'right': {'col': 'X15',
                                          'cutoff': 0.98,
                                          'gini_gain': 0.0015217391304347826,
                                          'index_col': 15,
                                          'left': {'col': 'X0',
```

```
                                                'cutoff': 0.0,
                                                'gini_gain': 0.0,
                                                'index_col': 0,
                                                'left': 0.0,
                                                'right': 0.0},
                                     'right': {'col': 'X0',
                                               'cutoff': 0.0,
                                               'gini_gain': 0.0,
                                               'index_col': 0,
                                               'left': 1.0,
                                               'right': 1.0}}}},
            'right': {'col': 'X24',
                      'cutoff': 0.73,
                      'gini_gain': 0.00377994863327831,
                      'index_col': 24,
                      'left': {'col': 'X45',
                               'cutoff': 0.24,
                               'gini_gain': 0.003464897124220191,
                               'index_col': 45,
                               'left': {'col': 'X41',
                                        'cutoff': 0.96,
                                        'gini_gain': 0.0016726572070332592,
                                        'index_col': 41,
                                        'left': {'col': 'X8',
                                                 'cutoff': 3.23,
                                                 'gini_gain':
0.0005607521291871066,
                                                 'index_col': 8,
                                                 'left': {'col': 'X28',
                                                          'cutoff': 0.31,
                                                          'gini_gain':
0.000562363485880222,
                                                          'index_col': 28,
                                                          'left': {'col': 'X32',
                                                                   'cutoff':
0.46,
                                                                   'gini_gain':
0.00017770335611414195,
                                                                   'index_col':
32,
                                                                   'left':
{'col': 'X41',
'cutoff': 0.2,
'gini_gain': 0.00017924860268904284,
'index_col': 41,
'left': {'col': 'X11',
    'cutoff': 0.29,
    'gini_gain': 8.944185109456895e-05,
```

```
    'index_col': 11,
    'left': {'col': 'X11',
             'cutoff': 0.28,
             'gini_gain': 0.0003191493600278908,
             'index_col': 11,
             'left': 1.0,
             'right': 1.0},
    'right': {'col': 'X10',
              'cutoff': 1.31,
              'gini_gain': 0.00019205363304742755,
              'index_col': 10,
              'left': 1.0,
              'right': 1.0}},
'right': {'col': 'X0',
     'cutoff': 0.41,
     'gini_gain': 0.0003864734299516908,
     'index_col': 0,
     'left': {'col': 'X0',
              'cutoff': 0.0,
              'gini_gain': 0.0,
              'index_col': 0,
              'left': 1.0,
              'right': 1.0},
     'right': 0.0}},
                                                                  'right':
{'col': 'X0',
'cutoff': 0.23,
'gini_gain': 0.0003864734299516908,
'index_col': 0,
'left': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 1.0,
     'right': 1.0},
'right': 0.0}},
                                                        'right': 0.0},
                                                'right': 0.0},
                                        'right': {'col': 'X0',
                                                  'cutoff': 0.0,
                                                  'gini_gain': 0.0,
                                                  'index_col': 0,
                                                  'left': 0.0,
                                                  'right': 0.0}},
                                'right': {'col': 'X34',
                                          'cutoff': 0.33,
                                          'gini_gain': 0.0009275362318840575,
                                          'index_col': 34,
```

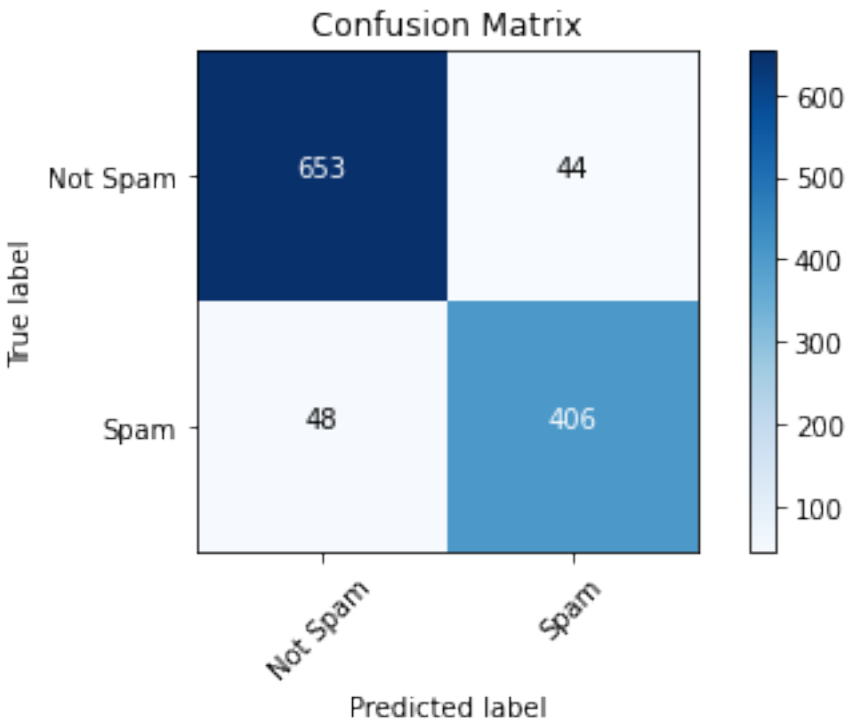
```
                                    'left': {'col': 'X0',
                                             'cutoff': 0.0,
                                             'gini_gain': 0.0,
                                             'index_col': 0,
                                             'left': 0.0,
                                             'right': 0.0},
                                    'right': {'col': 'X0',
                                              'cutoff': 0.0,
                                              'gini_gain': 0.0,
                                              'index_col': 0,
                                              'left': 1.0,
                                              'right': 1.0}}},
                          'right': {'col': 'X6',
                                    'cutoff': 0.15,
                                    'gini_gain': 0.0005152979066022544,
                                    'index_col': 6,
                                    'left': {'col': 'X0',
                                             'cutoff': 0.0,
                                             'gini_gain': 0.0,
                                             'index_col': 0,
                                             'left': 0.0,
                                             'right': 0.0},
                                    'right': 1.0}}}}
word_freq_make (0, 0.002644402436462929)
word_freq_address (1, 0)
word_freq_all (2, 0.0017004830917874392)
word_freq_3d (3, 0.0016059002143903058)
word_freq_our (4, 0.01167236436750024)
word_freq_over (5, 0)
word_freq_remove (6, 0.04968232907441131)
word_freq_internet (7, 0.0026487535285817775)
word_freq_order (8, 0.001084703218906343)
word_freq_mail (9, 0.001693728994313767)
word_freq_receive (10, 0.002044828754671641)
word_freq_will (11, 0.002195880218370166)
word_freq_people (12, 0.0034478502953873245)
word_freq_report (13, 0.00043393507924400446)
word_freq_addresses (14, 0)
word_freq_free (15, 0.010998741691175432)
word_freq_business (16, 0)
word_freq_email (17, 0.0014519313214965373)
word_freq_you (18, 0.0005383022774327124)
word_freq_credit (19, 0)
word_freq_your (20, 0.002740887132191482)
word_freq_font (21, 0)
word_freq_000 (22, 0.0005323174749815887)
word_freq_money (23, 0.013600533060163808)
word_freq_hp (24, 0.019576820940810764)
```

```
word_freq_hpl (25, 0)
word_freq_george (26, 0.00978099818218736)
word_freq_650 (27, 0.0010458306605004967)
word_freq_lab (28, 0.000562363485880222)
word_freq_labs (29, 0)
word_freq_telnet (30, 0)
word_freq_857 (31, 0)
word_freq_data (32, 0.0007307757295847331)
word_freq_415 (33, 0.0006365444728616088)
word_freq_85 (34, 0.0009275362318840575)
word_freq_technology (35, 0.00209379471663033)
word_freq_1999 (36, 0)
word_freq_parts (37, 0)
word_freq_pm (38, 0)
word_freq_direct (39, 0)
word_freq_cs (40, 0)
word_freq_meeting (41, 0.005121120476449713)
word_freq_original (42, 0)
word_freq_project (43, 0.0005124587996788213)
word_freq_re (44, 0.0025669367732453832)
word_freq_edu (45, 0.009139063655395022)
word_freq_table (46, 0)
word_freq_conference (47, 0.0010255743154123976)
char_freq_; (48, 0.0010145851108580284)
char_freq_( (49, 0.0014229249011857715)
char_freq_[ (50, 0.000527009222661397)
char_freq_! (51, 0.1666649059777205)
char_freq_$ (52, 0.053451154153458545)
char_freq_# (53, 0)
capital_run_length_average (54, 0.0313647263018774)
capital_run_length_longest (55, 0.003784883122232982)
capital_run_length_total (56, 0.00454295837213701)

/home/suresh/.local/lib/python3.8/site-packages/seaborn/_decorators.py:36:
FutureWarning: Pass the following variables as keyword args: x, y. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```



	precision	recall	f1-score	support
Not Spam	0.93	0.94	0.93	697
Spam	0.90	0.89	0.90	454
accuracy			0.92	1151
macro avg	0.92	0.92	0.92	1151
weighted avg	0.92	0.92	0.92	1151

```
Confusion matrix, without normalization
[[653  44]
 [ 48 406]]
```



[81]:
```
#bootstrap samples
def draw_bootstrap(X_train, y_train):
    bootstrap_indices = list(np.random.choice(range(len(X_train)), len(X_train),␣
 ↪replace = True))
    oob_indices = [i for i in range(len(X_train)) if i not in bootstrap_indices]
    X_bootstrap = X_train[bootstrap_indices,:]
    y_bootstrap = y_train[bootstrap_indices]
    X_oob = X_train[oob_indices,:]
    y_oob = y_train[oob_indices]
    return X_bootstrap, y_bootstrap, X_oob, y_oob
```

[82]:
```
# prediction with a list of  trees
def rf_predict(trees,feature_list, row):
        predictions = [get_prediction(tree, row[feature]) for tree,feature in␣
 ↪zip(trees,feature_list)]
        return max(set(predictions), key=predictions.count)

# Random Forest Algorithm
def random_forest(X_train, y_train, max_depth, min_size, n_trees,␣
 ↪n_features=None):
    trees = list()
    feature_list=list()
    random_n_features=None
```

```
    for i in range(n_trees):
        X_bootstrap, y_bootstrap, X_oob, y_oob=draw_bootstrap(X_train, y_train)
        if n_features is not None:
            random_n_features=random.sample(range(1, X_train.shape[-1]),␣
↪n_features)
        else:
            random_n_features=random.sample(range(1, X_train.shape[-1]), int(np.
↪round(math.sqrt(X_train.shape[-1]))))
        clf = DecisionTreeClassifier(max_depth, t=min_size)
        tree = clf.fit(X_bootstrap[:,random_n_features], y_bootstrap)
        trees.append(tree)
        feature_list.append(random_n_features)
    return trees,feature_list
```

[83]:
```
def main():

    call_decision_tree()# decision tree function called, for parameters you can␣
↪visit this function, defaut parameters are maxdeptth with 10 and minimum␣
↪samples as 1.

    #random forest
    AUC_list=[]
    F1_score_list=[]
    for n_t in range(1,200,10):
        print("Random forest model training in progress")
        trees,feature_list=random_forest(X_train, y_train, max_depth=10,␣
↪min_size=1, n_trees=n_t)
        y_pred = [rf_predict(trees,feature_list, row) for row in X_test]
        fpr, tpr, thresholds = metrics.roc_curve(y_test, y_pred)
        AUC_list.append(metrics.auc(fpr, tpr))
        F1_score_list.append(metrics.f1_score(y_test, y_pred))
    #save the results in a file
    f = open("RF-result.txt", "w")
    for auc,f1,nt in zip(AUC_list,F1_score_list,range(1,200,10)):
        f.write(str(auc)+','+str(f1)+','+str(nt)+'\n')
    f.close()
```

[84]:
```
#main()
```

```
/home/suresh/.local/lib/python3.8/site-packages/numpy/core/fromnumeric.py:3372:
RuntimeWarning: Mean of empty slice.
  return _methods._mean(a, axis=axis, dtype=dtype,
/home/suresh/.local/lib/python3.8/site-packages/numpy/core/_methods.py:170:
RuntimeWarning: invalid value encountered in double_scalars
  ret = ret.dtype.type(ret / rcount)
```

```
{'col': 'X51',
 'cutoff': 0.079,
 'gini_gain': 0.1573079330035172,
 'index_col': 51,
 'left': {'col': 'X6',
          'cutoff': 0.05,
          'gini_gain': 0.036940589137910096,
          'index_col': 6,
          'left': {'col': 'X23',
                   'cutoff': 0.02,
                   'gini_gain': 0.013600533060163808,
                   'index_col': 23,
                   'left': {'col': 'X15',
                            'cutoff': 0.2,
                            'gini_gain': 0.0043473035328583055,
                            'index_col': 15,
                            'left': {'col': 'X52',
                                     'cutoff': 0.182,
                                     'gini_gain': 0.004177741889223485,
                                     'index_col': 52,
                                     'left': {'col': 'X3',
                                              'cutoff': 0.44,
                                              'gini_gain':
0.0010432403678429902,
                                              'index_col': 3,
                                              'left': {'col': 'X27',
                                                       'cutoff': 8.33,
                                                       'gini_gain':
0.0010458306605004967,
                                                       'index_col': 27,
                                                       'left': {'col': 'X24',
                                                                'cutoff': 0.13,
                                                                'gini_gain':
0.0009246440692747221,
                                                                'index_col': 24,
                                                                'left': {'col':
'X55',
'cutoff': 11.0,
'gini_gain': 0.002662657521746514,
'index_col': 55,
                                                                         'left':
{'col': 'X10',
  'cutoff': 0.29,
  'gini_gain': 0.0007089664779281465,
  'index_col': 10,
  'left': {'col': 'X7',
           'cutoff': 11.11,
           'gini_gain': 0.0005497897931703006,
```

```
           'index_col': 7,
           'left': 0.0,
           'right': 1.0},
  'right': {'col': 'X0',
            'cutoff': 0.26,
            'gini_gain': 0.0003864734299516908,
            'index_col': 0,
            'left': 1.0,
            'right': 0.0}},
'right': {'col': 'X4',
   'cutoff': 0.73,
   'gini_gain': 0.0028408620939035937,
   'index_col': 4,
   'left': {'col': 'X35',
            'cutoff': 0.24,
            'gini_gain': 0.00209379471663033,
            'index_col': 35,
            'left': 0.0,
            'right': 1.0},
   'right': {'col': 'X26',
             'cutoff': 0.39,
             'gini_gain': 0.0011690821256038644,
             'index_col': 26,
             'left': 1.0,
             'right': 0.0}}},
                                                                  'right': {'col':
'X10',
'cutoff': 1.36,
'gini_gain': 0.001143808643696067,
'index_col': 10,
'left': {'col': 'X56',
   'cutoff': 2510.0,
   'gini_gain': 0.00018835224813089504,
   'index_col': 56,
   'left': {'col': 'X48',
            'cutoff': 0.28300000000000003,
            'gini_gain': 2.4684055049824003e-05,
            'index_col': 48,
            'left': 0.0,
            'right': 0.0},
   'right': {'col': 'X0',
             'cutoff': 0.06,
             'gini_gain': 0.0003864734299516908,
             'index_col': 0,
             'left': 0.0,
             'right': 1.0}},
'right': {'col': 'X0',
    'cutoff': 0.0,
```

```
    'gini_gain': 0.0,
    'index_col': 0,
    'left': 1.0,
    'right': 1.0}},
                                                    'right': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain':
0.0,
                                                              'index_col': 0,
                                                              'left': 1.0,
                                                              'right': 1.0}},
                                                 'right': {'col': 'X0',
                                                           'cutoff': 0.0,
                                                           'gini_gain': 0.0,
                                                           'index_col': 0,
                                                           'left': 1.0,
                                                           'right': 1.0}},
                                          'right': {'col': 'X54',
                                                    'cutoff': 3.272,
                                                    'gini_gain':
0.002591716396703609,
                                                    'index_col': 54,
                                                    'left': {'col': 'X12',
                                                             'cutoff': 5.55,
                                                             'gini_gain':
0.0004665286404416835,
                                                             'index_col': 12,
                                                             'left': {'col': 'X17',
                                                                      'cutoff': 4.83,
                                                                      'gini_gain':
0.0005383022774327124,
                                                                      'index_col':
17,
                                                                      'left': {'col':
'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 0.0,
'right': 0.0},
                                                                      'right': 1.0},
                                                             'right': 1.0},
                                                    'right': {'col': 'X33',
                                                              'cutoff': 0.13,
                                                              'gini_gain':
0.0006365444728616088,
                                                              'index_col': 33,
                                                              'left': {'col': 'X0',
```

```
                                                                'cutoff': 0.0,
                                                                'gini_gain':
0.0,
                                                                'index_col':
0,
                                                                'left': 1.0,
                                                                'right': 1.0},
                                                     'right': {'col': 'X0',
                                                               'cutoff':
0.19,
                                                               'gini_gain':
0.0003864734299516908,
                                                               'index_col':
0,
                                                               'left':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 0.0,
'right': 0.0},
                                                               'right':
1.0}}}},
                           'right': {'col': 'X4',
                                     'cutoff': 1.17,
                                     'gini_gain': 0.003098814229249011,
                                     'index_col': 4,
                                     'left': {'col': 'X44',
                                              'cutoff': 0.18,
                                              'gini_gain':
0.0012574738576443607,
                                              'index_col': 44,
                                              'left': {'col': 'X24',
                                                       'cutoff': 0.27,
                                                       'gini_gain':
0.0013564143853998932,
                                                       'index_col': 24,
                                                       'left': {'col': 'X15',
                                                                'cutoff': 0.51,
                                                                'gini_gain':
0.0012054467271858567,
                                                                'index_col':
15,
                                                                'left': {'col':
'X4',
'cutoff': 0.52,
'gini_gain': 0.000452976974716105,
'index_col': 4,
```

```
'left': {'col': 'X50',
   'cutoff': 0.059000000000000004,
   'gini_gain': 0.000527009222661397,
   'index_col': 50,
   'left': {'col': 'X0',
            'cutoff': 0.0,
            'gini_gain': 0.0,
            'index_col': 0,
            'left': 1.0,
            'right': 1.0},
   'right': 0.0},
'right': {'col': 'X2',
    'cutoff': 0.84,
    'gini_gain': 0.0003864734299516908,
    'index_col': 2,
    'left': {'col': 'X0',
             'cutoff': 0.0,
             'gini_gain': 0.0,
             'index_col': 0,
             'left': 0.0,
             'right': 0.0},
    'right': 1.0}},
                                                                  'right':
{'col': 'X9',
'cutoff': 0.3,
'gini_gain': 0.0006020066889632112,
'index_col': 9,
'left': {'col': 'X15',
    'cutoff': 4.0,
    'gini_gain': 0.00044851258581235645,
    'index_col': 15,
    'left': {'col': 'X13',
             'cutoff': 0.52,
             'gini_gain': 0.00043393507924400446,
             'index_col': 13,
             'left': 0.0,
             'right': 1.0},
    'right': {'col': 'X15',
              'cutoff': 7.69,
              'gini_gain': 0.00030917874396135266,
              'index_col': 15,
              'left': 1.0,
              'right': 0.0}},
'right': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 1.0,
```

```
     'right': 1.0}}},
                                                    'right': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain':
0.0,
                                                              'index_col':
0,
                                                              'left': 0.0,
                                                              'right':
0.0}},
                                                   'right': {'col': 'X3',
                                                             'cutoff': 7.07,
                                                             'gini_gain':
0.0005626598465473155,
                                                             'index_col': 3,
                                                             'left': {'col': 'X0',
                                                                      'cutoff': 0.0,
                                                                      'gini_gain':
0.0,
                                                                      'index_col':
0,
                                                                      'left': 0.0,
                                                                      'right': 0.0},
                                                             'right': 1.0}},
                                          'right': {'col': 'X44',
                                                    'cutoff': 0.95,
                                                    'gini_gain':
0.0013094629156010225,
                                                    'index_col': 44,
                                                    'left': {'col': 'X12',
                                                             'cutoff': 0.8,
                                                             'gini_gain':
0.0005456095481670925,
                                                             'index_col': 12,
                                                             'left': {'col': 'X0',
                                                                      'cutoff': 0.0,
                                                                      'gini_gain':
0.0,
                                                                      'index_col':
0,
                                                                      'left': 1.0,
                                                                      'right': 1.0},
                                                             'right': 0.0},
                                                    'right': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain': 0.0,
                                                              'index_col': 0,
                                                              'left': 0.0,
```

```
                                                    'right': 0.0}}}},
                    'right': {'col': 'X24',
                              'cutoff': 0.17,
                              'gini_gain': 0.0034168294792912084,
                              'index_col': 24,
                              'left': {'col': 'X52',
                                       'cutoff': 0.013999999999999999,
                                       'gini_gain': 0.0027529427774375718,
                                       'index_col': 52,
                                       'left': {'col': 'X54',
                                                'cutoff': 4.333,
                                                'gini_gain':
0.0016701173222912 35,
                                                'index_col': 54,
                                                'left': {'col': 'X18',
                                                         'cutoff': 12.5,
                                                         'gini_gain':
0.0005383022774327124,
                                                         'index_col': 18,
                                                         'left': {'col': 'X0',
                                                                  'cutoff': 0.0,
                                                                  'gini_gain':
0.0,
                                                                  'index_col': 0,
                                                                  'left': 0.0,
                                                                  'right': 0.0},
                                                         'right': 1.0},
                                                'right': {'col': 'X11',
                                                          'cutoff': 1.31,
                                                          'gini_gain':
0.0004968944099378883,
                                                          'index_col': 11,
                                                          'left': {'col': 'X0',
                                                                   'cutoff': 0.0,
                                                                   'gini_gain':
0.0,
                                                                   'index_col':
0,
                                                                   'left': 1.0,
                                                                   'right': 1.0},
                                                          'right': 0.0}},
                                       'right': {'col': 'X26',
                                                 'cutoff': 0.16,
                                                 'gini_gain':
0.0009782608695652166,
                                                 'index_col': 26,
                                                 'left': {'col': 'X7',
                                                          'cutoff': 0.73,
```

```
                                                        'gini_gain':
0.0005203515263644802,
                                                        'index_col': 7,
                                                        'left': {'col': 'X56',
                                                                 'cutoff':
53.0,
                                                                 'gini_gain':
0.0005429755057716249,
                                                                 'index_col':
56,
                                                                 'left': 0.0,
                                                                 'right':
{'col': 'X0',
'cutoff': 0.36,
'gini_gain': 0.00027725267800882266,
'index_col': 0,
'left': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 1.0,
     'right': 1.0},
'right': {'col': 'X0',
      'cutoff': 0.36,
      'gini_gain': 0.0,
      'index_col': 0,
      'left': 0.0,
      'right': 0.0}}},
                                                                 'right': 0.0},
                                                        'right': {'col': 'X0',
                                                                  'cutoff': 0.0,
                                                                  'gini_gain': 0.0,
                                                                  'index_col': 0,
                                                                  'left': 0.0,
                                                                  'right': 0.0}}},
                                 'right': {'col': 'X0',
                                           'cutoff': 0.0,
                                           'gini_gain': 0.0,
                                           'index_col': 0,
                                           'left': 0.0,
                                           'right': 0.0}}},
                    'right': {'col': 'X26',
                              'cutoff': 0.16,
                              'gini_gain': 0.004669228309547926,
                              'index_col': 26,
                              'left': {'col': 'X24',
                                       'cutoff': 0.21,
                                       'gini_gain': 0.0015375386077453087,
```

```
                            'index_col': 24,
                            'left': {'col': 'X7',
                                     'cutoff': 2.22,
                                     'gini_gain': 0.00103754940711463,
                                     'index_col': 7,
                                     'left': {'col': 'X45',
                                              'cutoff': 0.29,
                                              'gini_gain':
0.001075977162933675,
                                              'index_col': 45,
                                              'left': {'col': 'X32',
                                                       'cutoff': 1.76,
                                                       'gini_gain':
0.0005530723734705912,
                                                       'index_col': 32,
                                                       'left': {'col': 'X9',
                                                                'cutoff': 1.78,
                                                                'gini_gain':
0.00017713944369775417,
                                                                'index_col': 9,
                                                                'left': {'col':
'X55',
'cutoff': 6.0,
'gini_gain': 0.00011036789297658651,
'index_col': 55,
'left': {'col': 'X2',
   'cutoff': 0.25,
   'gini_gain': 0.00046376811594202875,
   'index_col': 2,
   'left': {'col': 'X0',
            'cutoff': 0.0,
            'gini_gain': 0.0,
            'index_col': 0,
            'left': 1.0,
            'right': 1.0},
   'right': 0.0},
'right': {'col': 'X0',
    'cutoff': 0.0,
    'gini_gain': 0.0,
    'index_col': 0,
    'left': 1.0,
    'right': 1.0}},
                                                                'right':
{'col': 'X9',
'cutoff': 1.96,
'gini_gain': 0.0003864734299516908,
'index_col': 9,
'left': 0.0,
```

```
'right': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 1.0,
     'right': 1.0}}},
                                                        'right': 0.0},
                                                'right': {'col': 'X0',
                                                          'cutoff': 0.0,
                                                          'gini_gain': 0.0,
                                                          'index_col': 0,
                                                          'left': 0.0,
                                                          'right': 0.0}},
                                        'right': {'col': 'X0',
                                                  'cutoff': 0.0,
                                                  'gini_gain': 0.0,
                                                  'index_col': 0,
                                                  'left': 0.0,
                                                  'right': 0.0}},
                                'right': {'col': 'X4',
                                          'cutoff': 0.68,
                                          'gini_gain': 0.0009275362318840581,
                                          'index_col': 4,
                                          'left': {'col': 'X0',
                                                   'cutoff': 0.0,
                                                   'gini_gain': 0.0,
                                                   'index_col': 0,
                                                   'left': 0.0,
                                                   'right': 0.0},
                                          'right': {'col': 'X4',
                                                    'cutoff': 0.8,
                                                    'gini_gain':
0.00017391304347826064,
                                                    'index_col': 4,
                                                    'left': {'col': 'X0',
                                                             'cutoff': 0.0,
                                                             'gini_gain': 0.0,
                                                             'index_col': 0,
                                                             'left': 0.0,
                                                             'right': 0.0},
                                                    'right': {'col': 'X0',
                                                              'cutoff': 0.0,
                                                              'gini_gain': 0.0,
                                                              'index_col': 0,
                                                              'left': 1.0,
                                                              'right': 1.0}}}},
                         'right': {'col': 'X0',
                                   'cutoff': 0.0,
```

```
                            'gini_gain': 0.0,
                            'index_col': 0,
                            'left': 0.0,
                            'right': 0.0}}},
 'right': {'col': 'X52',
           'cutoff': 0.013000000000000001,
           'gini_gain': 0.046520469486797486,
           'index_col': 52,
           'left': {'col': 'X54',
                    'cutoff': 2.755,
                    'gini_gain': 0.026507781617062208,
                    'index_col': 54,
                    'left': {'col': 'X6',
                             'cutoff': 0.18,
                             'gini_gain': 0.012226442029898955,
                             'index_col': 6,
                             'left': {'col': 'X51',
                                      'cutoff': 0.8140000000000001,
                                      'gini_gain': 0.006907894254644194,
                                      'index_col': 51,
                                      'left': {'col': 'X4',
                                               'cutoff': 0.33,
                                               'gini_gain':
0.0031959751598086617,
                                               'index_col': 4,
                                               'left': {'col': 'X15',
                                                        'cutoff': 1.43,
                                                        'gini_gain':
0.002093023665501414,
                                                        'index_col': 15,
                                                        'left': {'col': 'X8',
                                                                 'cutoff': 0.93,
                                                                 'gini_gain':
0.0005239510897192364,
                                                                 'index_col': 8,
                                                                 'left': {'col':
'X9',
'cutoff': 7.55,
'gini_gain': 0.0005281094317011106,
'index_col': 9,
'left': {'col': 'X22',
   'cutoff': 2.24,
   'gini_gain': 0.0005323174749815887,
   'index_col': 22,
   'left': {'col': 'X12',
            'cutoff': 1.15,
            'gini_gain': 0.00038362679833277226,
            'index_col': 12,
```

```
            'left': 0.0,
            'right': 0.0},
   'right': 1.0},
'right': 1.0},
                                                                  'right': 1.0},
                                                         'right': {'col': 'X55',
                                                                   'cutoff': 7.0,
                                                                   'gini_gain':
0.0010118577075098815,
                                                                   'index_col':
55,
                                                                   'left':
{'col': 'X2',
'cutoff': 1.44,
'gini_gain': 0.00046376811594202875,
'index_col': 2,
'left': {'col': 'X0',
    'cutoff': 0.0,
    'gini_gain': 0.0,
    'index_col': 0,
    'left': 0.0,
    'right': 0.0},
'right': 1.0},
                                                                   'right':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 1.0,
'right': 1.0}}},
                                                  'right': {'col': 'X26',
                                                            'cutoff': 0.38,
                                                            'gini_gain':
0.0029644268774703534,
                                                            'index_col': 26,
                                                            'left': {'col': 'X20',
                                                                     'cutoff':
0.45,
                                                                     'gini_gain':
0.002740887132191482,
                                                                     'index_col':
20,
                                                                     'left':
{'col': 'X7',
'cutoff': 0.6,
'gini_gain': 0.0005410628019323668,
'index_col': 7,
'left': {'col': 'X0',
```

```
    'cutoff': 0.0,
    'gini_gain': 0.0,
    'index_col': 0,
    'left': 0.0,
    'right': 0.0},
'right': 1.0},
                                                                  'right':
{'col': 'X17',
'cutoff': 0.36,
'gini_gain': 0.000913629044063825,
'index_col': 17,
'left': {'col': 'X49',
     'cutoff': 0.147,
     'gini_gain': 0.0014229249011857715,
     'index_col': 49,
     'left': {'col': 'X56',
              'cutoff': 42.0,
              'gini_gain': 0.0007453416149068321,
              'index_col': 56,
              'left': 0.0,
              'right': 1.0},
     'right': {'col': 'X0',
               'cutoff': 0.0,
               'gini_gain': 0.0,
               'index_col': 0,
               'left': 0.0,
               'right': 0.0}},
'right': {'col': 'X4',
      'cutoff': 0.59,
      'gini_gain': 0.0005475040257648954,
      'index_col': 4,
      'left': 0.0,
      'right': {'col': 'X0',
                'cutoff': 0.0,
                'gini_gain': 0.0,
                'index_col': 0,
                'left': 1.0,
                'right': 1.0}}}},
                                                 'right': {'col': 'X0',
                                                           'cutoff':
0.0,
                                                           'gini_gain':
0.0,
                                                           'index_col':
0,
                                                           'left': 0.0,
                                                           'right':
0.0}}},
```

```
                                    'right': {'col': 'X56',
                                              'cutoff': 15.0,
                                              'gini_gain':
0.003066289003327658,
                                              'index_col': 56,
                                              'left': {'col': 'X15',
                                                       'cutoff': 1.4,
                                                       'gini_gain':
0.001073537305421365,
                                                       'index_col': 15,
                                                       'left': {'col': 'X51',
                                                                'cutoff':
4.3469999999999995,
                                                                'gini_gain':
0.0008783487044356597,
                                                                'index_col':
51,
                                                                'left':
{'col': 'X51',
'cutoff': 0.925,
'gini_gain': 0.0005019135453918068,
'index_col': 51,
'left': 1.0,
'right': {'col': 'X51',
     'cutoff': 1.754,
     'gini_gain': 6.901311249137372e-05,
     'index_col': 51,
     'left': {'col': 'X0',
              'cutoff': 0.0,
              'gini_gain': 0.0,
              'index_col': 0,
              'left': 0.0,
              'right': 0.0},
     'right': {'col': 'X51',
               'cutoff': 2.272,
               'gini_gain': 0.00048309178743961335,
               'index_col': 51,
               'left': 1.0,
               'right': 0.0}}},
                                                                'right':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 1.0,
'right': 1.0}},
                                                       'right': {'col': 'X0',
                                                                 'cutoff':
```

```
0.0,
                                                                    'gini_gain':
0.0,
                                                                    'index_col':
0,
                                                                    'left': 1.0,
                                                                    'right':
1.0}},
                                                          'right': {'col': 'X11',
                                                                    'cutoff': 2.58,
                                                                    'gini_gain':
0.0012903945973098182,
                                                                    'index_col': 11,
                                                                    'left': {'col': 'X54',
                                                                             'cutoff':
1.2,
                                                                             'gini_gain':
0.0005951109658203482,
                                                                             'index_col':
54,
                                                                             'left':
{'col': 'X2',
'cutoff': 1.0,
'gini_gain': 0.0003864734299516908,
'index_col': 2,
'left': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 0.0,
     'right': 0.0},
'right': 1.0},
                                                                             'right':
{'col': 'X12',
'cutoff': 0.28,
'gini_gain': 0.0005305078553160381,
'index_col': 12,
'left': {'col': 'X24',
      'cutoff': 4.34,
      'gini_gain': 0.0005626598465473155,
      'index_col': 24,
      'left': {'col': 'X0',
               'cutoff': 0.0,
               'gini_gain': 0.0,
               'index_col': 0,
               'left': 1.0,
               'right': 1.0},
      'right': 0.0},
```

```
'right': 0.0}},
                                                    'right': {'col': 'X0',
                                                              'cutoff':
0.0,
                                                              'gini_gain':
0.0,
                                                              'index_col':
0,
                                                              'left': 0.0,
                                                              'right':
0.0}}}},
                              'right': {'col': 'X45',
                                        'cutoff': 0.23,
                                        'gini_gain': 0.0010695132200274518,
                                        'index_col': 45,
                                        'left': {'col': 'X24',
                                                 'cutoff': 0.46,
                                                 'gini_gain':
0.0011100832562442157,
                                                 'index_col': 24,
                                                 'left': {'col': 'X0',
                                                          'cutoff': 0.0,
                                                          'gini_gain': 0.0,
                                                          'index_col': 0,
                                                          'left': 1.0,
                                                          'right': 1.0},
                                                 'right': {'col': 'X0',
                                                           'cutoff': 0.0,
                                                           'gini_gain': 0.0,
                                                           'index_col': 0,
                                                           'left': 0.0,
                                                           'right': 0.0}},
                                        'right': {'col': 'X4',
                                                  'cutoff': 1.67,
                                                  'gini_gain':
0.0004347826086956522,
                                                  'index_col': 4,
                                                  'left': {'col': 'X0',
                                                           'cutoff': 0.0,
                                                           'gini_gain': 0.0,
                                                           'index_col': 0,
                                                           'left': 0.0,
                                                           'right': 0.0},
                                                  'right': 1.0}}},
                    'right': {'col': 'X24',
                              'cutoff': 0.21,
                              'gini_gain': 0.006888702663029791,
                              'index_col': 24,
```

```
                                'left': {'col': 'X45',
                                         'cutoff': 0.56,
                                         'gini_gain': 0.003528676148213704,
                                         'index_col': 45,
                                         'left': {'col': 'X41',
                                                  'cutoff': 0.89,
                                                  'gini_gain':
0.003269214666727411,
                                                  'index_col': 41,
                                                  'left': {'col': 'X47',
                                                           'cutoff': 0.77,
                                                           'gini_gain':
0.0010255743154123976,
                                                           'index_col': 47,
                                                           'left': {'col': 'X48',
                                                                    'cutoff':
1.151,
                                                                    'gini_gain':
0.0009899010558082045,
                                                                    'index_col':
48,
                                                                    'left':
{'col': 'X12',
'cutoff': 0.55,
'gini_gain': 0.0006520122357384341,
'index_col': 12,
'left': {'col': 'X43',
    'cutoff': 2.04,
    'gini_gain': 0.0005124587996788213,
    'index_col': 43,
    'left': {'col': 'X51',
             'cutoff': 19.131,
             'gini_gain': 0.0005167115698006461,
             'index_col': 51,
             'left': 1.0,
             'right': 0.0},
    'right': 0.0},
'right': {'col': 'X12',
     'cutoff': 0.74,
     'gini_gain': 0.0008695652173913044,
     'index_col': 12,
     'left': {'col': 'X0',
              'cutoff': 0.0,
              'gini_gain': 0.0,
              'index_col': 0,
              'left': 0.0,
              'right': 0.0},
     'right': {'col': 'X0',
```

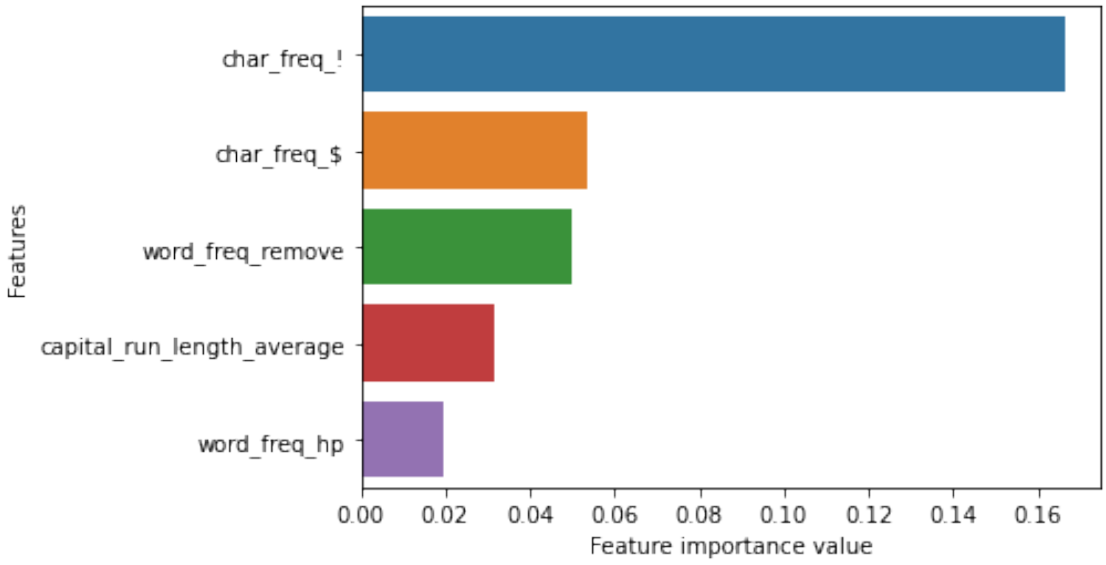
```
                'cutoff': 0.0,
                'gini_gain': 0.0,
                'index_col': 0,
                'left': 1.0,
                'right': 1.0}}},
                                                                    'right':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 0.0,
'right': 0.0}},
                                                          'right': {'col': 'X0',
                                                                    'cutoff':
2.32,
                                                                    'gini_gain':
0.0004347826086956522,
                                                                    'index_col':
0,
                                                                    'left':
{'col': 'X0',
'cutoff': 0.0,
'gini_gain': 0.0,
'index_col': 0,
'left': 0.0,
'right': 0.0},
                                                                    'right':
1.0}},
                                                  'right': {'col': 'X0',
                                                            'cutoff': 0.0,
                                                            'gini_gain': 0.0,
                                                            'index_col': 0,
                                                            'left': 0.0,
                                                            'right': 0.0}},
                                          'right': {'col': 'X0',
                                                    'cutoff': 0.0,
                                                    'gini_gain': 0.0,
                                                    'index_col': 0,
                                                    'left': 0.0,
                                                    'right': 0.0}},
                                 'right': {'col': 'X15',
                                           'cutoff': 0.98,
                                           'gini_gain': 0.0015217391304347826,
                                           'index_col': 15,
                                           'left': {'col': 'X0',
                                                    'cutoff': 0.0,
                                                    'gini_gain': 0.0,
                                                    'index_col': 0,
```

```
                                                'left': 0.0,
                                                'right': 0.0},
                                       'right': {'col': 'X0',
                                                 'cutoff': 0.0,
                                                 'gini_gain': 0.0,
                                                 'index_col': 0,
                                                 'left': 1.0,
                                                 'right': 1.0}}}},
           'right': {'col': 'X24',
                     'cutoff': 0.73,
                     'gini_gain': 0.00377994863327831,
                     'index_col': 24,
                     'left': {'col': 'X45',
                              'cutoff': 0.24,
                              'gini_gain': 0.003464897124220191,
                              'index_col': 45,
                              'left': {'col': 'X41',
                                       'cutoff': 0.96,
                                       'gini_gain': 0.0016726572070332592,
                                       'index_col': 41,
                                       'left': {'col': 'X8',
                                                'cutoff': 3.23,
                                                'gini_gain':
0.0005607521291871066,
                                                'index_col': 8,
                                                'left': {'col': 'X28',
                                                         'cutoff': 0.31,
                                                         'gini_gain':
0.000562363485880222,
                                                         'index_col': 28,
                                                         'left': {'col': 'X32',
                                                                  'cutoff':
0.46,
                                                                  'gini_gain':
0.0001777035611414195,
                                                                  'index_col':
32,
                                                                  'left':
{'col': 'X41',
'cutoff': 0.2,
'gini_gain': 0.00017924860268904284,
'index_col': 41,
'left': {'col': 'X11',
    'cutoff': 0.29,
    'gini_gain': 8.944185109456895e-05,
    'index_col': 11,
    'left': {'col': 'X11',
             'cutoff': 0.28,
```

```
            'gini_gain': 0.0003191493600278908,
            'index_col': 11,
            'left': 1.0,
            'right': 1.0},
    'right': {'col': 'X10',
              'cutoff': 1.31,
              'gini_gain': 0.00019205363304742755,
              'index_col': 10,
              'left': 1.0,
              'right': 1.0}},
'right': {'col': 'X0',
     'cutoff': 0.41,
     'gini_gain': 0.0003864734299516908,
     'index_col': 0,
     'left': {'col': 'X0',
              'cutoff': 0.0,
              'gini_gain': 0.0,
              'index_col': 0,
              'left': 1.0,
              'right': 1.0},
     'right': 0.0}},
                                                                  'right':
{'col': 'X0',
'cutoff': 0.23,
'gini_gain': 0.0003864734299516908,
'index_col': 0,
'left': {'col': 'X0',
     'cutoff': 0.0,
     'gini_gain': 0.0,
     'index_col': 0,
     'left': 1.0,
     'right': 1.0},
'right': 0.0}},
                                                          'right': 0.0},
                                                 'right': 0.0},
                                        'right': {'col': 'X0',
                                                  'cutoff': 0.0,
                                                  'gini_gain': 0.0,
                                                  'index_col': 0,
                                                  'left': 0.0,
                                                  'right': 0.0}},
                                'right': {'col': 'X34',
                                          'cutoff': 0.33,
                                          'gini_gain': 0.0009275362318840575,
                                          'index_col': 34,
                                          'left': {'col': 'X0',
                                                   'cutoff': 0.0,
                                                   'gini_gain': 0.0,
```

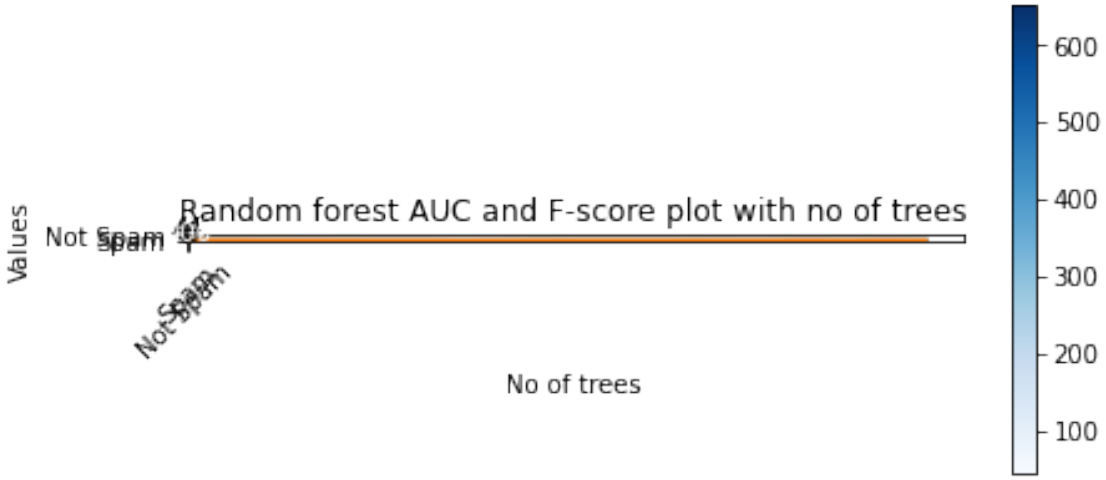
```
                                                'index_col': 0,
                                                'left': 0.0,
                                                'right': 0.0},
                                       'right': {'col': 'X0',
                                                 'cutoff': 0.0,
                                                 'gini_gain': 0.0,
                                                 'index_col': 0,
                                                 'left': 1.0,
                                                 'right': 1.0}}},
                            'right': {'col': 'X6',
                                      'cutoff': 0.15,
                                      'gini_gain': 0.0005152979066022544,
                                      'index_col': 6,
                                      'left': {'col': 'X0',
                                               'cutoff': 0.0,
                                               'gini_gain': 0.0,
                                               'index_col': 0,
                                               'left': 0.0,
                                               'right': 0.0},
                                      'right': 1.0}}}}
word_freq_make (0, 0.002644402436462929)
word_freq_address (1, 0)
word_freq_all (2, 0.0017004830917874392)
word_freq_3d (3, 0.0016059002143903058)
word_freq_our (4, 0.01167236436750024)
word_freq_over (5, 0)
word_freq_remove (6, 0.04968232907441131)
word_freq_internet (7, 0.0026487535285817775)
word_freq_order (8, 0.001084703218906343)
word_freq_mail (9, 0.001693728994313767)
word_freq_receive (10, 0.002044828754671641)
word_freq_will (11, 0.002195880218370166)
word_freq_people (12, 0.0034478502953873245)
word_freq_report (13, 0.00043393507924400446)
word_freq_addresses (14, 0)
word_freq_free (15, 0.010998741691175432)
word_freq_business (16, 0)
word_freq_email (17, 0.0014519313214965373)
word_freq_you (18, 0.0005583022774327124)
word_freq_credit (19, 0)
word_freq_your (20, 0.002740887132191482)
word_freq_font (21, 0)
word_freq_000 (22, 0.0005323174749815887)
word_freq_money (23, 0.013600533060163808)
word_freq_hp (24, 0.019576820940810764)
word_freq_hpl (25, 0)
word_freq_george (26, 0.00978099818218736)
word_freq_650 (27, 0.0010458306605004967)
```

```
word_freq_lab (28, 0.000562363485880222)
word_freq_labs (29, 0)
word_freq_telnet (30, 0)
word_freq_857 (31, 0)
word_freq_data (32, 0.0007307757295847331)
word_freq_415 (33, 0.0006365444728616088)
word_freq_85 (34, 0.0009275362318840575)
word_freq_technology (35, 0.00209379471663033)
word_freq_1999 (36, 0)
word_freq_parts (37, 0)
word_freq_pm (38, 0)
word_freq_direct (39, 0)
word_freq_cs (40, 0)
word_freq_meeting (41, 0.005121120476449713)
word_freq_original (42, 0)
word_freq_project (43, 0.0005124587996788213)
word_freq_re (44, 0.0025669367732453832)
word_freq_edu (45, 0.009139063655395022)
word_freq_table (46, 0)
word_freq_conference (47, 0.0010255743154123976)
char_freq_; (48, 0.0010145851108580284)
char_freq_( (49, 0.0014229249011857715)
char_freq_[ (50, 0.000527009222661397)
char_freq_! (51, 0.1666649059777205)
char_freq_$ (52, 0.053451154153458545)
char_freq_# (53, 0)
capital_run_length_average (54, 0.0313647263018774)
capital_run_length_longest (55, 0.003784883122232982)
capital_run_length_total (56, 0.00454295837213701)

/home/suresh/.local/lib/python3.8/site-packages/seaborn/_decorators.py:36:
FutureWarning: Pass the following variables as keyword args: x, y. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```

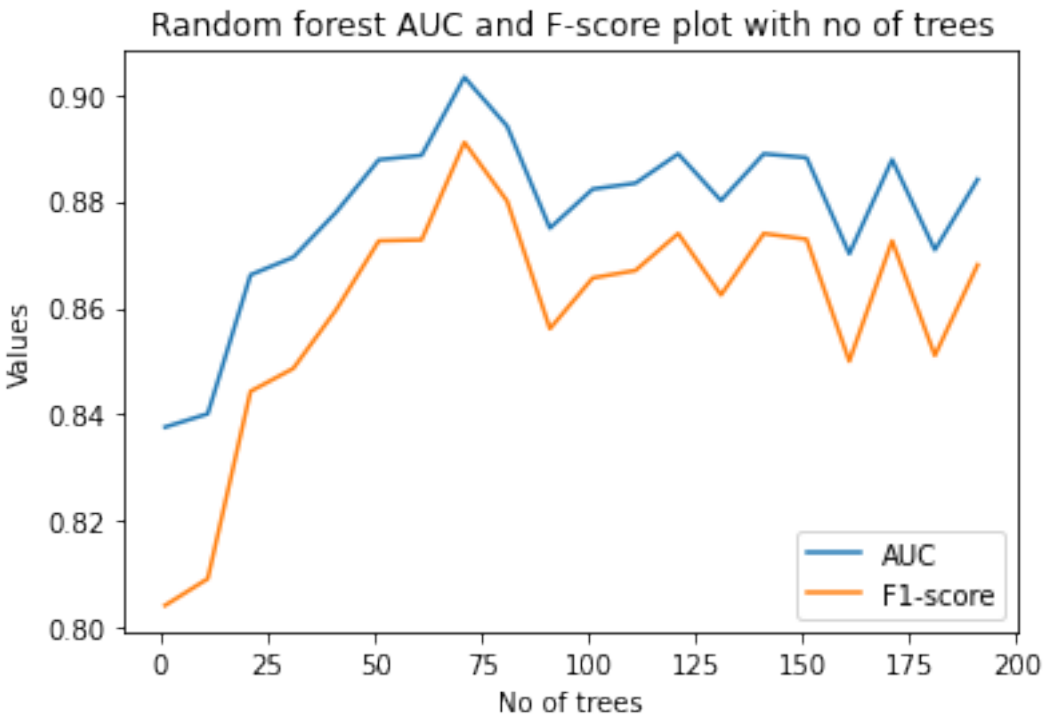



```
              precision    recall  f1-score   support

    Not Spam       0.93      0.94      0.93       697
        Spam       0.90      0.89      0.90       454

    accuracy                           0.92      1151
   macro avg       0.92      0.92      0.92      1151
weighted avg       0.92      0.92      0.92      1151

Confusion matrix, without normalization
[[653  44]
 [ 48 406]]
Random forest model training in progress

/home/suresh/.local/lib/python3.8/site-packages/numpy/core/fromnumeric.py:3372:
RuntimeWarning: Mean of empty slice.
  return _methods._mean(a, axis=axis, dtype=dtype,
/home/suresh/.local/lib/python3.8/site-packages/numpy/core/_methods.py:170:
RuntimeWarning: invalid value encountered in double_scalars
  ret = ret.dtype.type(ret / rcount)

Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
```

```
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
Random forest model training in progress
```



[90]:
```
data = pd.read_csv('RF-result.txt', header=None)
data.head()
#AUC plot
plt.plot(data[2], data[0], label = "AUC")

# plotting the line 2 points
plt.plot(data[2],data[1], label = "F1-score")
plt.title('Random forest AUC and F-score plot with no of trees')
plt.xlabel('No of trees')
plt.ylabel('Values')
plt.legend(loc=4)
plt.show()
```

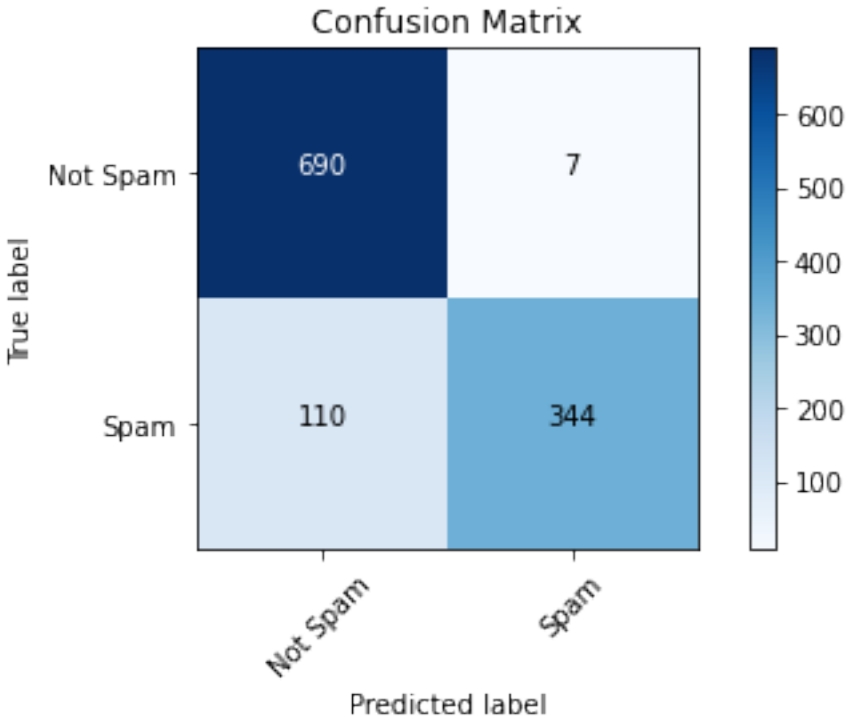


4 (a) Print a confusion matrix (you can use package implementations here).

[89]:
```
print("Best Random forest model with 71 trees ")
trees,feature_list=random_forest(X_train, y_train, max_depth=10, min_size=1,␣
 ↪n_trees=71)
y_pred = [rf_predict(trees,feature_list, row) for row in X_test]
plot_model_report(y_test,y_pred) #model report plotting
```

```
Best Random forest model with 71 trees

/home/suresh/.local/lib/python3.8/site-packages/numpy/core/fromnumeric.py:3372:
RuntimeWarning: Mean of empty slice.
  return _methods._mean(a, axis=axis, dtype=dtype,
/home/suresh/.local/lib/python3.8/site-packages/numpy/core/_methods.py:170:
RuntimeWarning: invalid value encountered in double_scalars
  ret = ret.dtype.type(ret / rcount)

              precision    recall  f1-score   support

    Not Spam       0.86      0.99      0.92       697
        Spam       0.98      0.76      0.85       454
```

```
    accuracy                           0.90      1151
   macro avg       0.92      0.87      0.89      1151
weighted avg       0.91      0.90      0.90      1151

Confusion matrix, without normalization
[[690   7]
 [110 344]]
```



5 (b) What is a good number of trees in the forest?

Answer: To get good number of trees we tried gridsearch with varying number of trees and plotted their respective accuracies(F1-score and AUC),which can be seen in above plotted figure we found 131 number of trees is best in the forest.