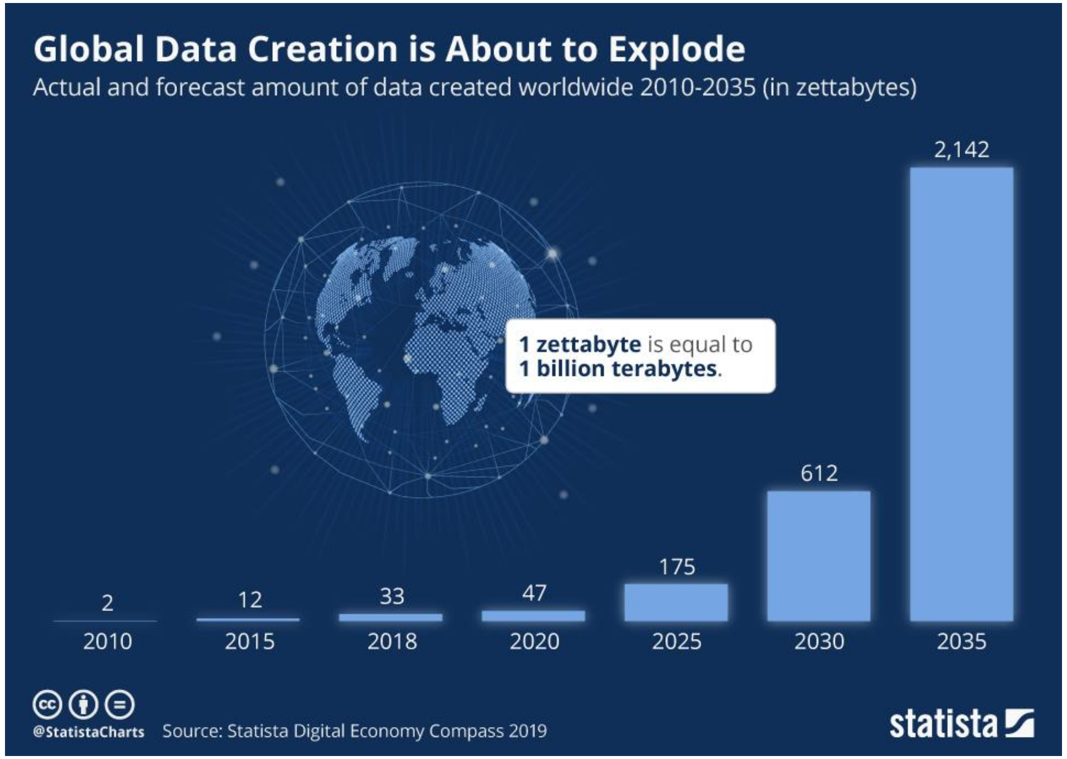
Machine Learning Assignment 1

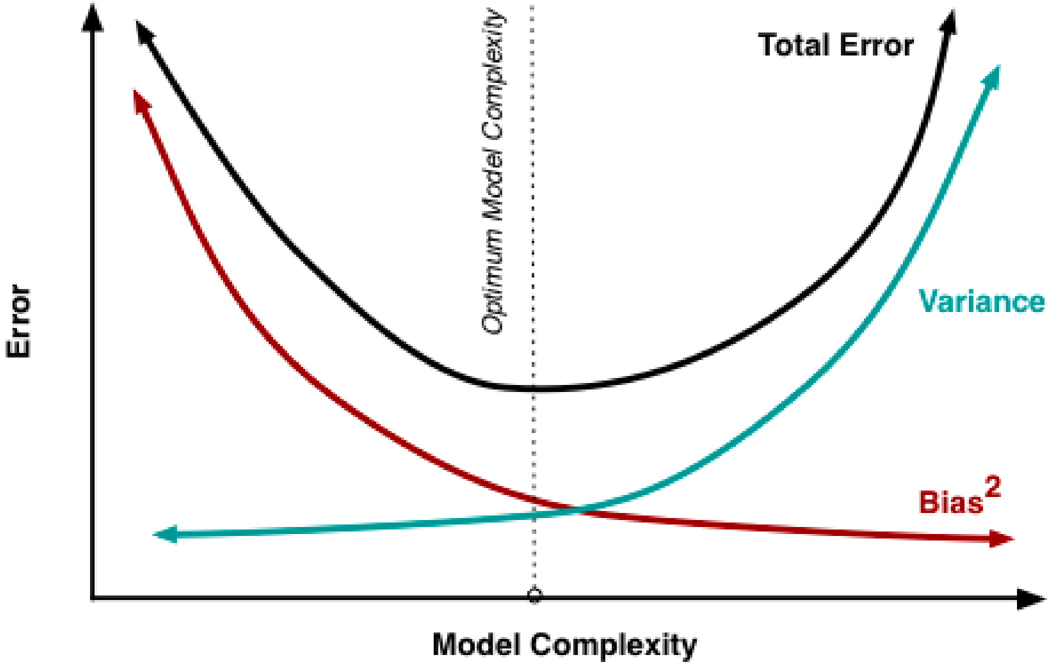
ANA RAQUEL MACEIRAS - UP200604342

GROUP B

Introduction

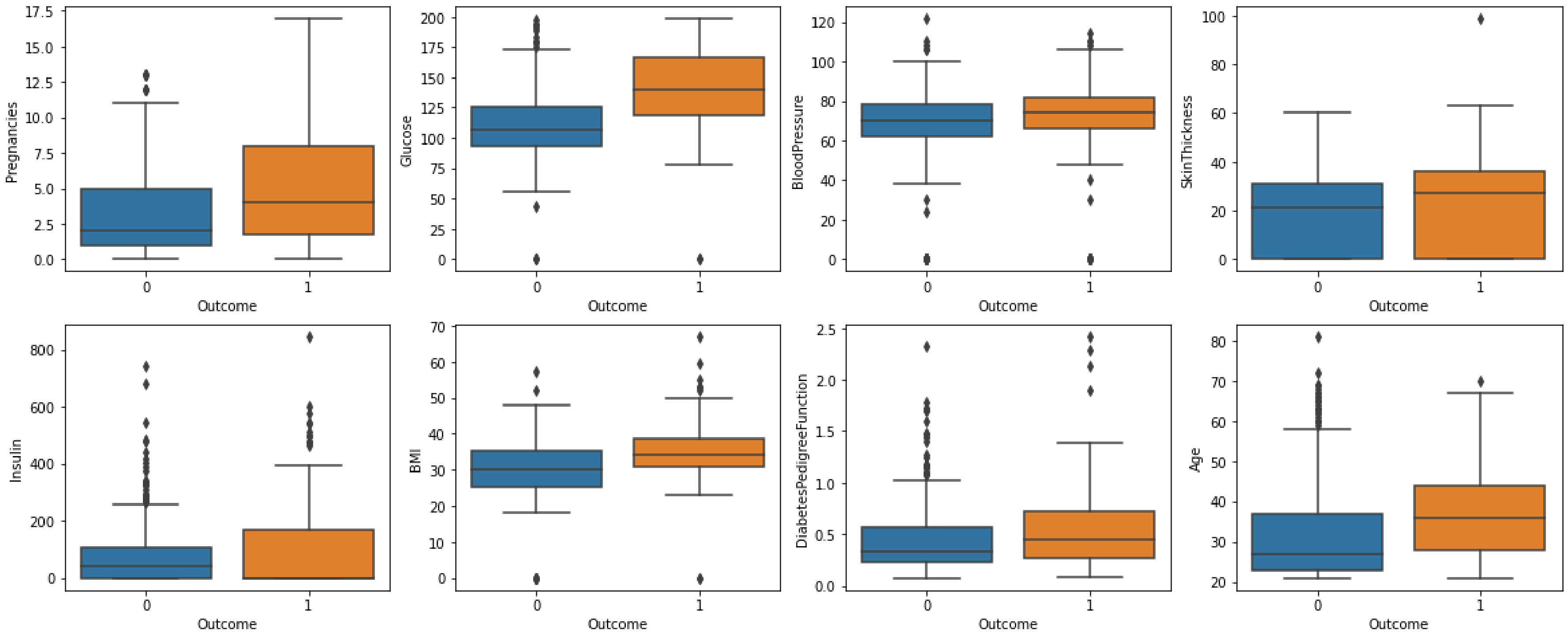


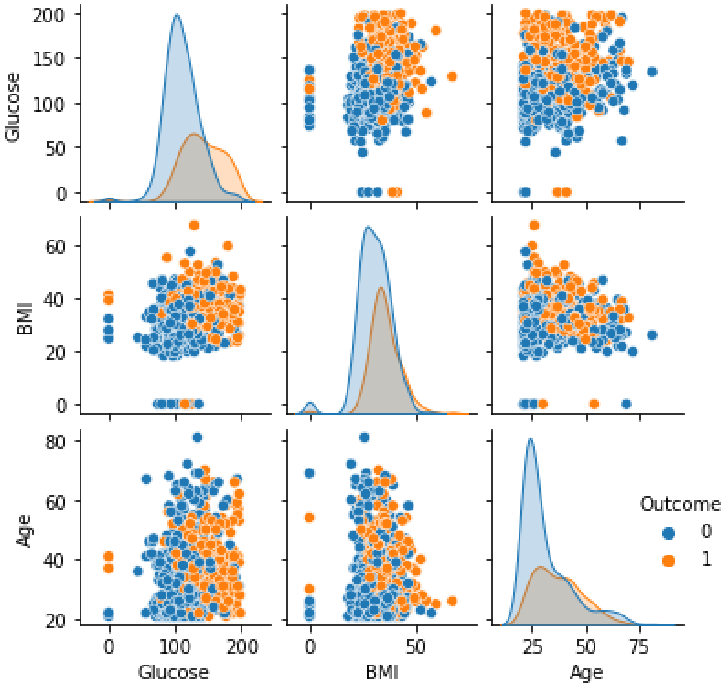
https://www.statista.com/chart/17727/global-data-creation-forecasts/



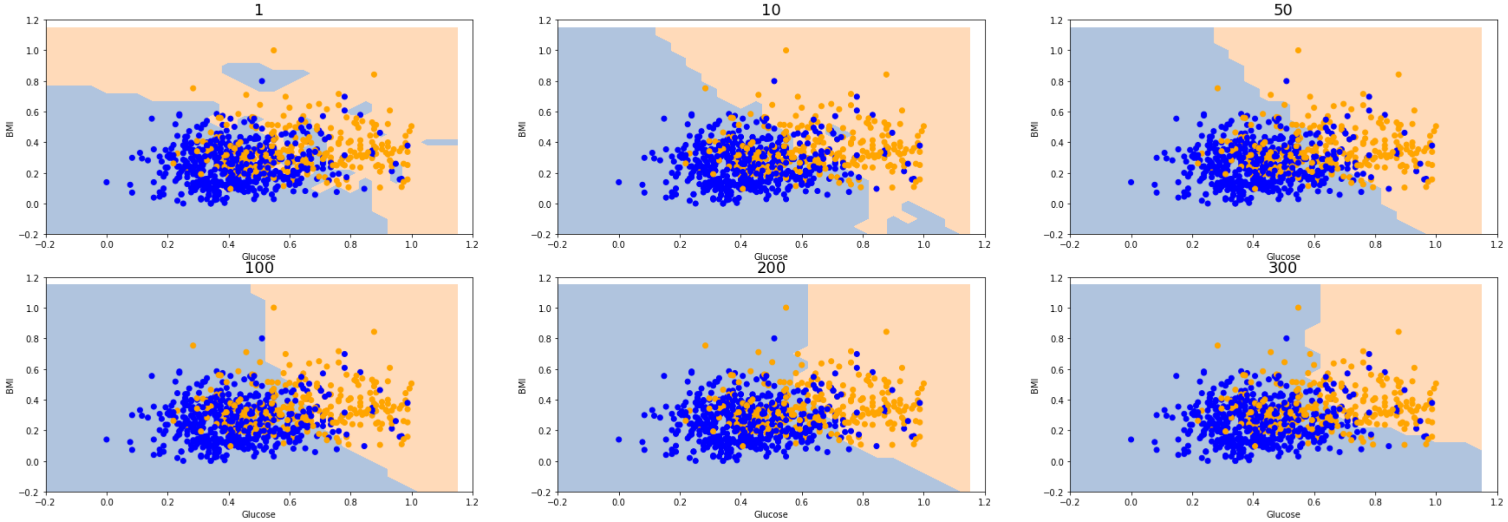
https://medium.com/30-days-of-machine-learning/day-3-k-nearest-neighbors-and-bias-variance-tradeoff-75f84d515bdb

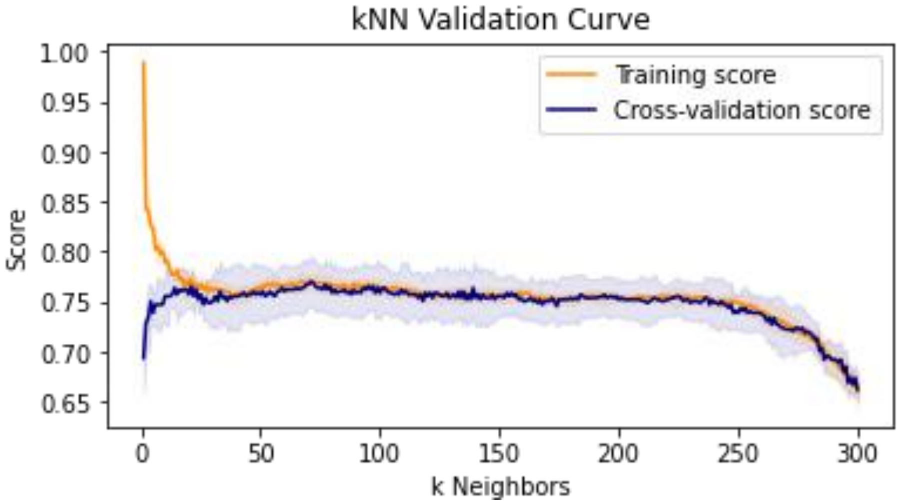
Attributes selection





kNN – k parameter hypertunning





Best k = 71

Logistic Regression, QDA and kNN models:

ROC curves, ROC AUC and decision boundaries

Logistic Regression

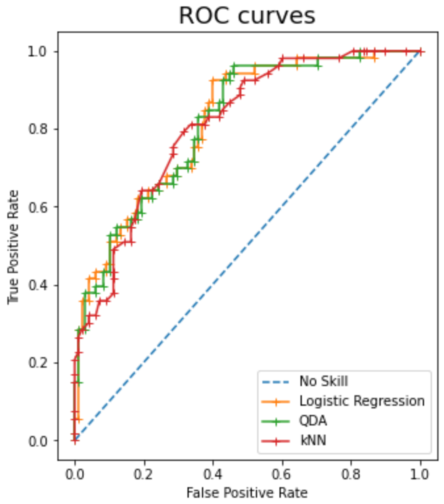
```
Accuracy score on training set: 0.7587
Accuracy score on test set: 0.7483
Accuracy score on 5-fold test data: 0.8015 +/- 0.0282

F1 score on training set: 0.5892
F1 score on test set: 0.5581
F1 score on 5-fold test data: 0.6339 +/- 0.1365
```

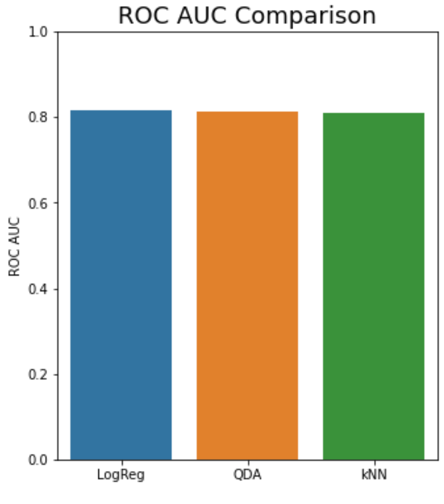
QDA

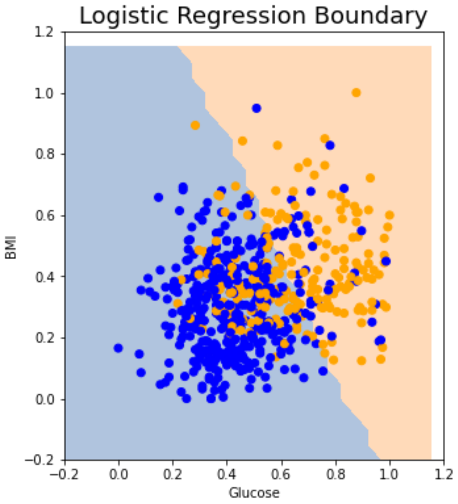
```
Accuracy score on training set: 0.7654
Accuracy score on test set: 0.755
Accuracy score on 5-fold test data: 0.7951 +/- 0.0363

F1 score on training set: 0.6137
F1 score on test set: 0.5843
F1 score on 5-fold test data: 0.6278 +/- 0.1356
```

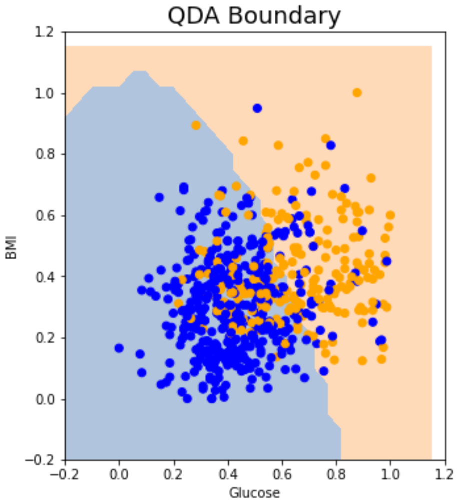
kNN

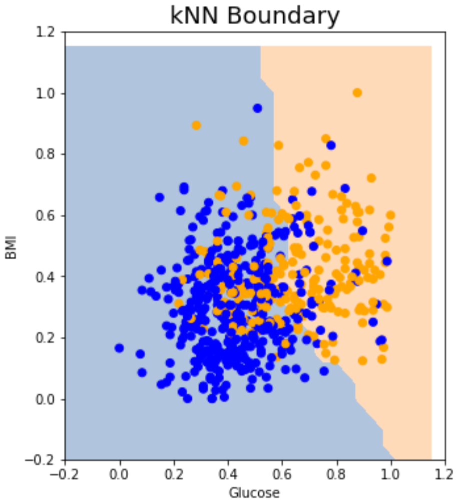
```
Accuracy score on training set:  0.7671
Accuracy score on test set:  0.7285
Accuracy score on 5-fold test data:  0.788 +/- 0.0274

F1 score on training set:  0.6154
F1 score on test set:  0.5287
F1 score on 5-fold test data:  0.6111 +/- 0.1113
```



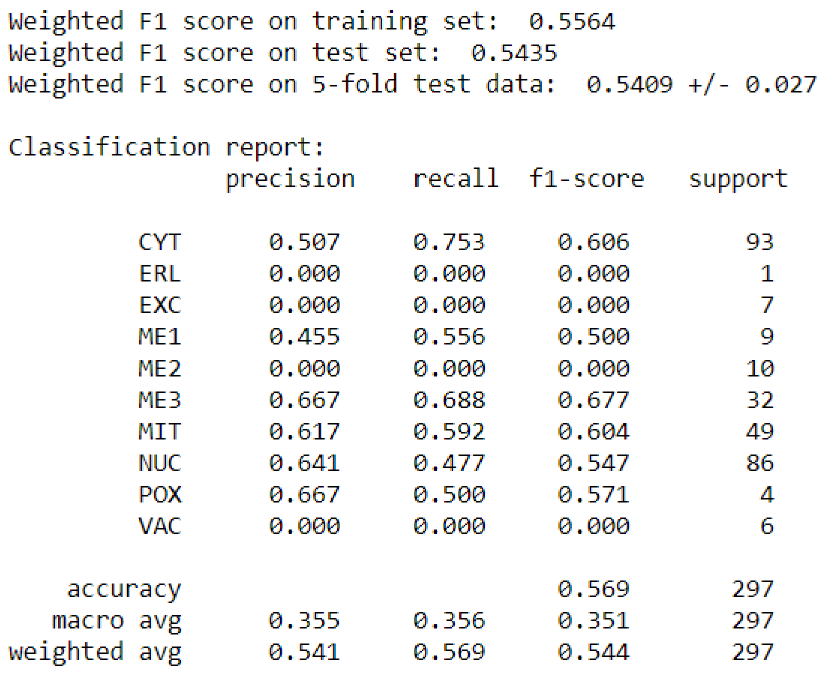






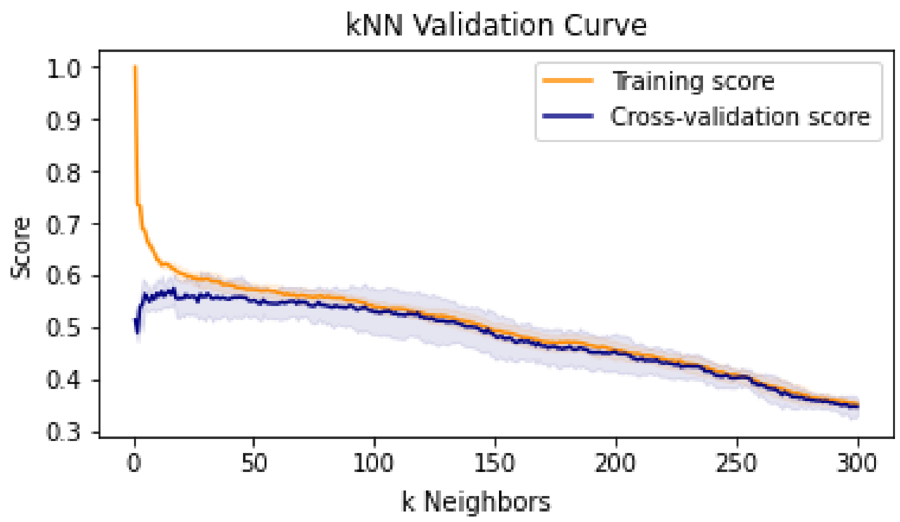


Logistic Regression

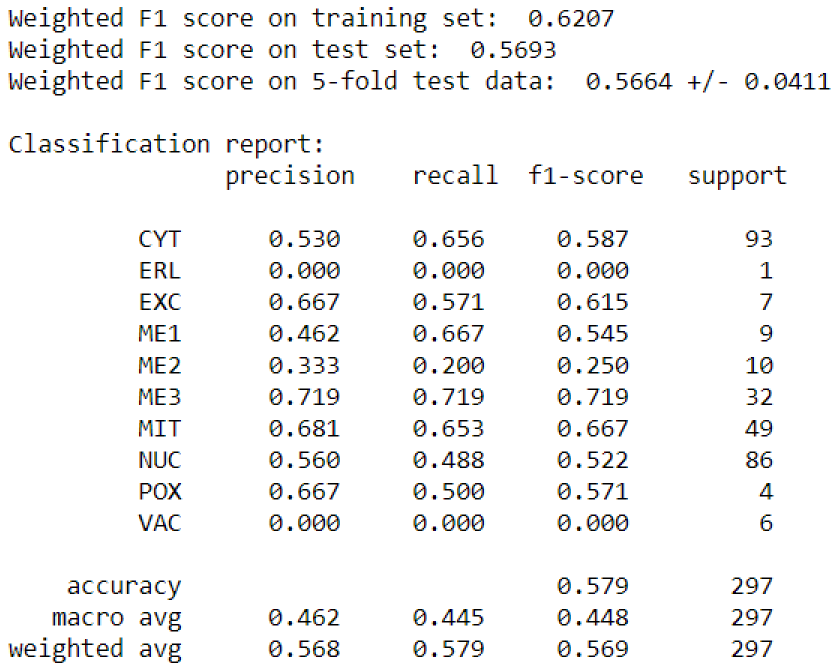


```
Weighted F1 score on training set:  0.5564
Weighted F1 score on test set:  0.5435
Weighted F1 score on 5-fold test data:  0.5409 +/- 0.027

Classification report:
              precision    recall  f1-score   support

         CYT      0.507     0.753     0.606        93
         ERL      0.000     0.000     0.000         1
         EXC      0.000     0.000     0.000         7
         ME1      0.455     0.556     0.500         9
         ME2      0.000     0.000     0.000        10
         ME3      0.667     0.688     0.677        32
         MIT      0.617     0.592     0.604        49
         NUC      0.641     0.477     0.547        86
         POX      0.667     0.500     0.571         4
         VAC      0.000     0.000     0.000         6

    accuracy                          0.569       297
   macro avg      0.355     0.356     0.351       297
weighted avg      0.541     0.569     0.544       297
```

kNN

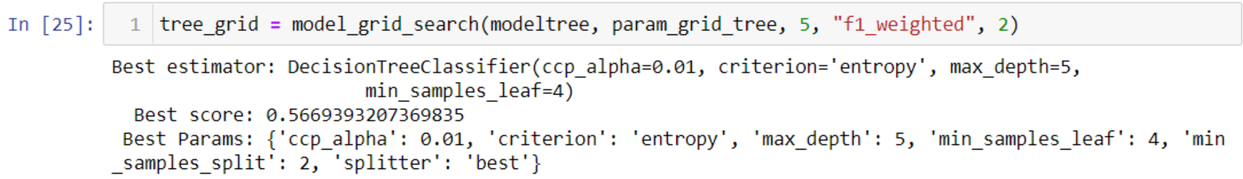


Best k = 17

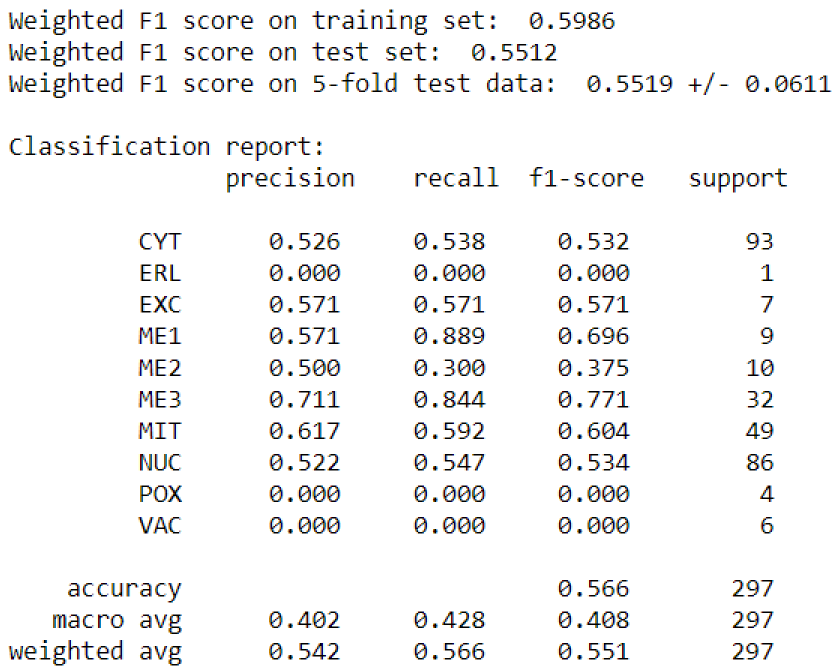


```
Weighted F1 score on training set:  0.6207
Weighted F1 score on test set:  0.5693
Weighted F1 score on 5-fold test data:  0.5664 +/- 0.0411

Classification report:
              precision    recall  f1-score   support

         CYT      0.530     0.656     0.587        93
         ERL      0.000     0.000     0.000         1
         EXC      0.667     0.571     0.615         7
         ME1      0.462     0.667     0.545         9
         ME2      0.333     0.200     0.250        10
         ME3      0.719     0.719     0.719        32
         MIT      0.681     0.653     0.667        49
         NUC      0.560     0.488     0.522        86
         POX      0.667     0.500     0.571         4
         VAC      0.000     0.000     0.000         6

    accuracy                          0.579       297
   macro avg      0.462     0.445     0.448       297
weighted avg      0.568     0.579     0.569       297
```

Decision Tree

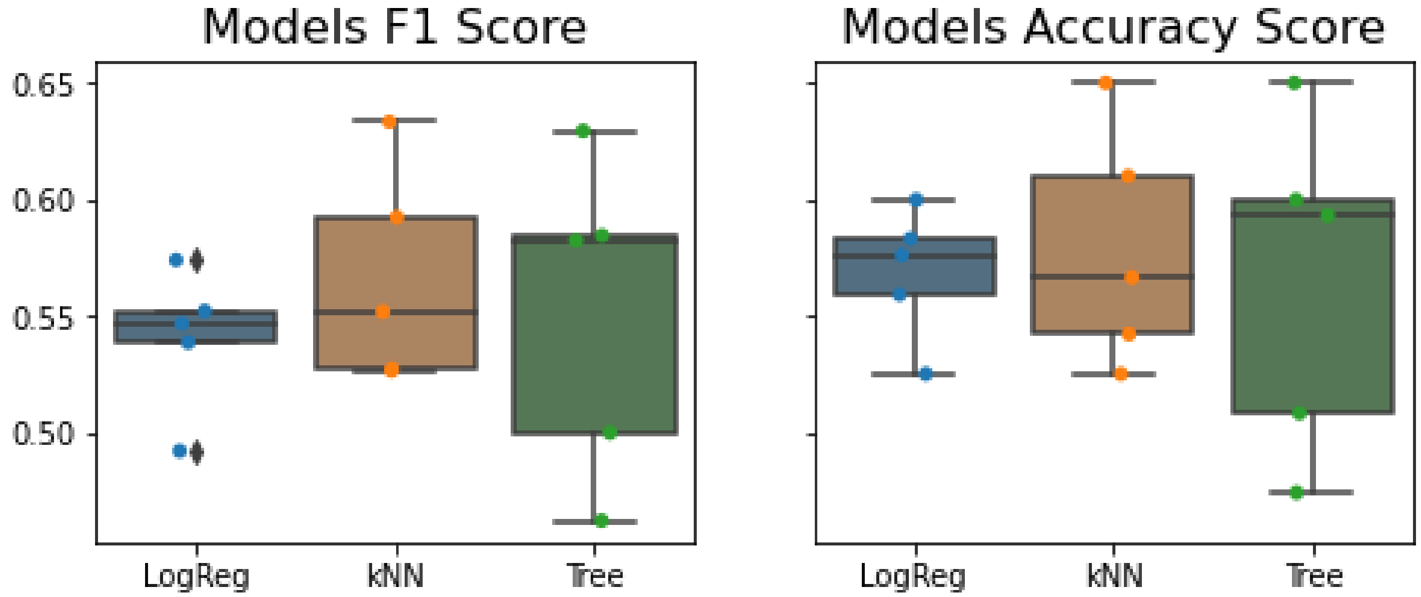


```
In [25]:  tree_grid = model_grid_search(modeltree, param_grid_tree, 5, "f1_weighted", 2)

Best estimator: DecisionTreeClassifier(ccp_alpha=0.01, criterion='entropy', max_depth=5,
                       min_samples_leaf=4)
  Best score: 0.5669393207369835
 Best Params: {'ccp_alpha': 0.01, 'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 4, 'min
_samples_split': 2, 'splitter': 'best'}
```



```
Weighted F1 score on training set:  0.5986
Weighted F1 score on test set:  0.5512
Weighted F1 score on 5-fold test data:  0.5519 +/- 0.0611

Classification report:
               precision    recall  f1-score   support

         CYT      0.526     0.538     0.532        93
         ERL      0.000     0.000     0.000         1
         EXC      0.571     0.571     0.571         7
         ME1      0.571     0.889     0.696         9
         ME2      0.500     0.300     0.375        10
         ME3      0.711     0.844     0.771        32
         MIT      0.617     0.592     0.604        49
         NUC      0.522     0.547     0.534        86
         POX      0.000     0.000     0.000         4
         VAC      0.000     0.000     0.000         6

    accuracy                          0.566       297
   macro avg      0.402     0.428     0.408       297
weighted avg      0.542     0.566     0.551       297
```

Achieved results comparison



Conclusions

- In classification problems, as in the case of the two problem here presented, machine learning models try to approximate the Bayes (true) Decision Boundary.

- Models that have hyperparameter that tune their performance, as was the case of kNN and decision trees, can be optimize in order to balance the variance and bias errors in order to not incur in overfitting or underfitting, respectively.

- In the case of the two problems presented in this work, we could not see significant differences between the performance of the proposed models, indicating that simpler models can perform as good as more complex model depending on the data.

- By using even more complex models or sampling models, especially in the yeast dataset which had unbalanced classes, one could expect to obtain better prediction scores.