Early Detection of Pancreatic Cancer using Machine Learning

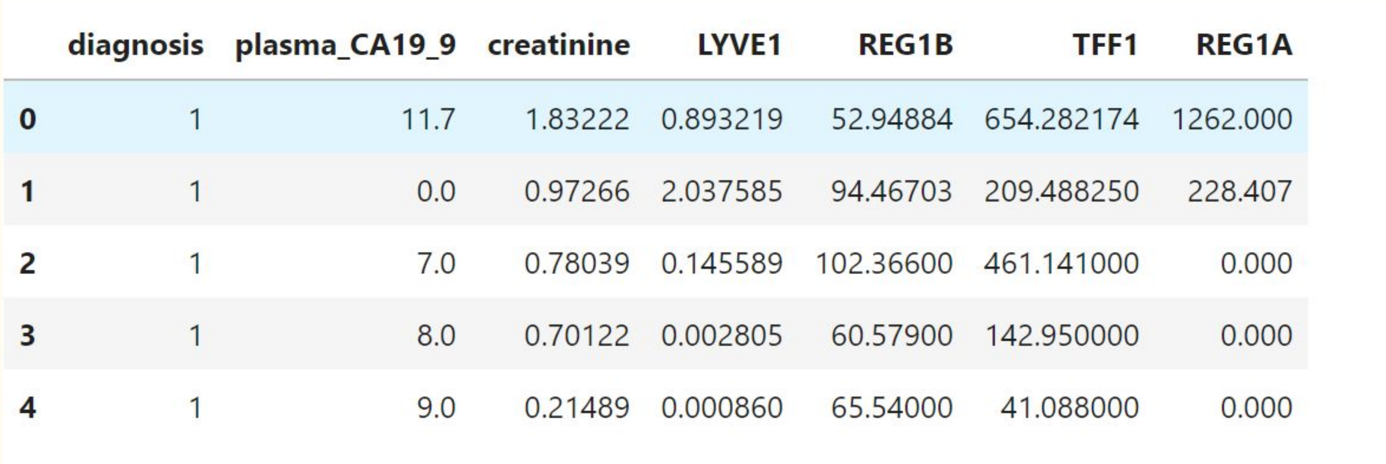
A supervised learning approach

Introduction

- Pancreatic Cancer is a deadly disease with low survival rates
- Early Detection is crucial for improving survival rates
- Examine the possibility of using ML to develop a model with currently available data to detect the cancer

Data Set

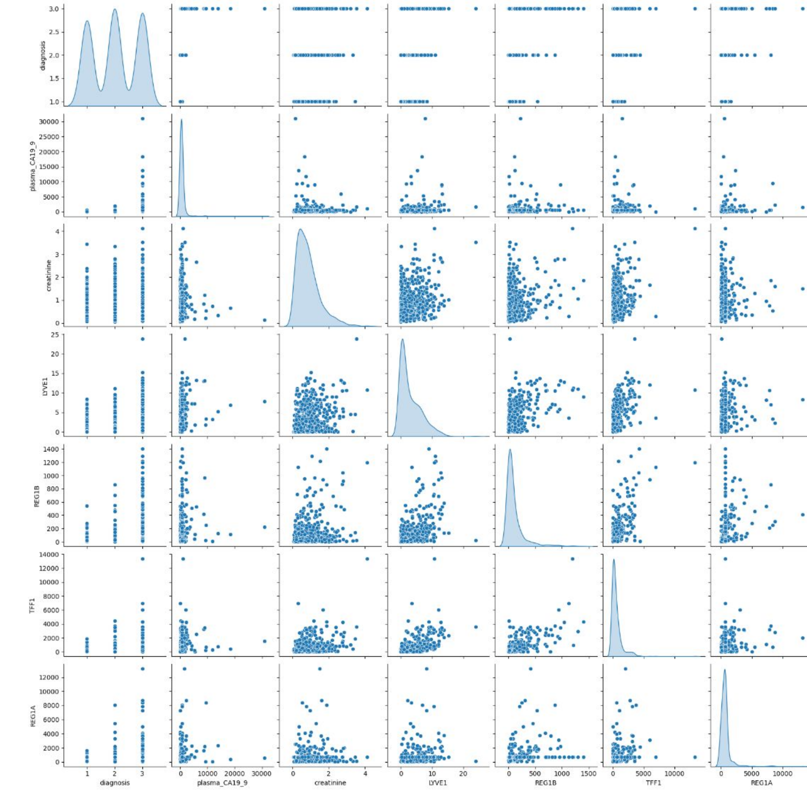
- 590 samples including healthy subjects, patients with non-cancerous pancreatic disease, and individuals with late stage pancreatic cancer
- Urinary biomarkers examined LYVE1, REG1B, TFF1, Plasma_CA19_9, and creatine



	diagnosis	plasma_CA19_9	creatinine	LYVE1	REG1B	TFF1	REG1A
0	1	11.7	1.83222	0.893219	52.94884	654.282174	1262.000
1	1	0.0	0.97266	2.037585	94.46703	209.488250	228.407
2	1	7.0	0.78039	0.145589	102.36600	461.141000	0.000
3	1	8.0	0.70122	0.002805	60.57900	142.950000	0.000
4	1	9.0	0.21489	0.000860	65.54000	41.088000	0.000

Exploratory Data Analysis

- Generally the pair plots show a random dispersion of data points with no correlation
- Density plots show the same results
- There are many outliers

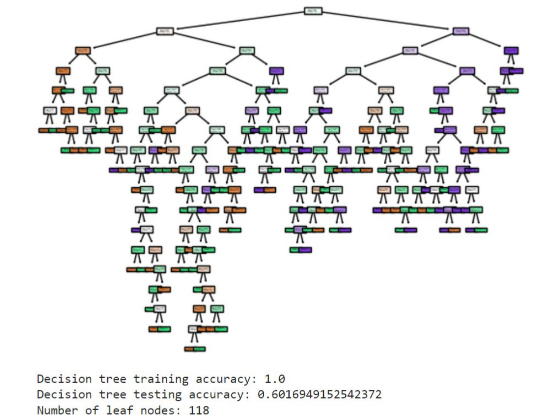


Model Building

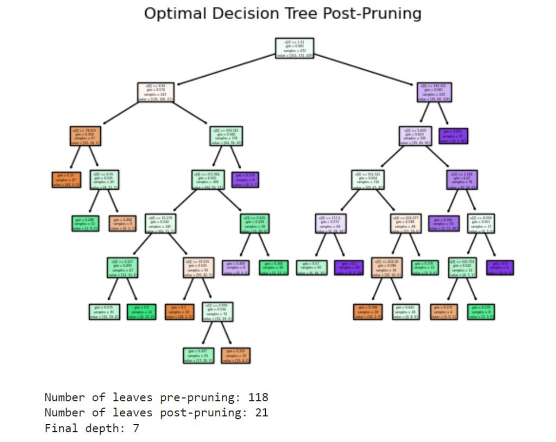
- Types of models used
    - Decision Tree
        - The unbalanced data set that we have may be covered by its ability to handle missing values which were inserted as 0 for the course of this analysis
    - Random Forest
        - This method boasts high accuracy and robustness of learning
    - AdaBoost
        - Since each biomarker has relatively weak correlation this method has the potential of boosting each element to have a great fit

Results- Decision Tree

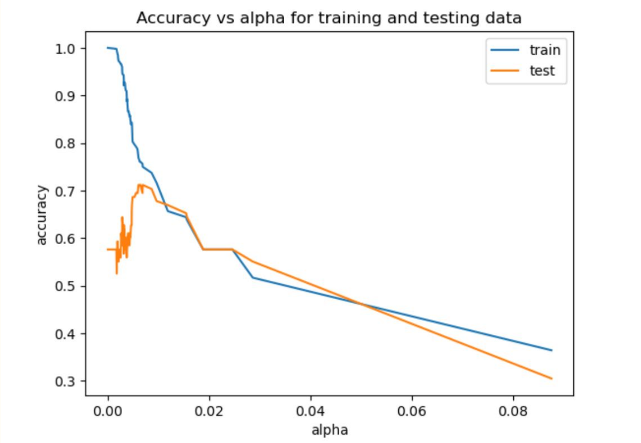
- Tree large and not accurate
- Optimized but highly unbalanced tree
- Accuracy is very low



```
Decision tree training accuracy: 1.0
Decision tree testing accuracy: 0.6016949152542372
Number of leaf nodes: 118
```

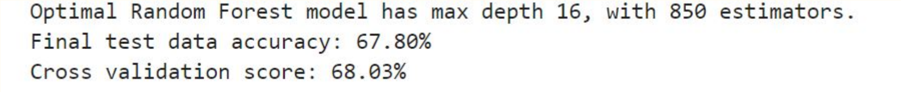


```
Number of leaves pre-pruning: 118
Number of leaves post-pruning: 21
Final depth: 7
```

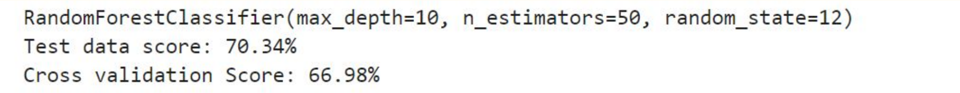


Results- Random Forest

- Unmodified
  - Decision Score 67.79%
- Optimal
  - Decision Score 67.80%
  - Cross Validation 68.03%
- Modified with specific criterion
  - Decision Score 69.49%
  - Cross Validation 65.70%



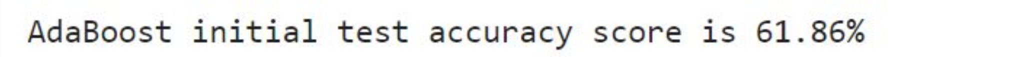
```
Optimal Random Forest model has max depth 16, with 850 estimators.
Final test data accuracy: 67.80%
Cross validation score: 68.03%
```



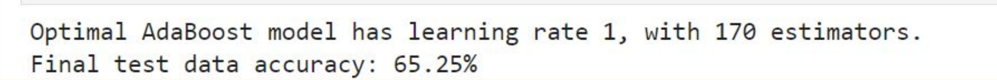
```
RandomForestClassifier(max_depth=10, n_estimators=50, random_state=12)
Test data score: 70.34%
Cross validation Score: 66.98%
```

Results- AdaBoost

- Initial
  - Accuracy 61.86%
- Final
  - Accuracy 65.25%



```
AdaBoost initial test accuracy score is 61.86%
```



```
Optimal AdaBoost model has learning rate 1, with 170 estimators.
Final test data accuracy: 65.25%
```

Conclusion

- What was found
  - Current data not enough to diagnose Pancreatic Cancer roughly 65% confidence
- Reason
  - The bio-markers are insufficient in diagnosis
  - Better diagnosis method