551 Mini Project 1

Yuhe Fan, Xinyu WAng, Yicheng Huang

February 9, 2022

1 Abstract

In this project we implemented two ML models perform them on two datasets, Hepatitis and Messidor. Cross-validation is performed to compare and determine the optimal parameters within each model, which shows different trends across different trails. Due to the relatively large amount of features, we run the experiments both with all feature and selected features to test if the model accuracy changes. Cross-model accuracy is also compared. The main finding is that two models have same level of accuracy. This accuracy is generally constant but under the subjection of random data-splitting.

2 Introduction

In this project we have test KNN and decision tree model on two datasets. Both model have same level of accuracy which is 5% to 10% apart. For the hepatitis dataset, there is mostly 80% accuracy, for the Messidor dataset, there is mostly 65% accuracy. It shows feature selection could increase the model accuracy. With our selected feature, these accuracy would be improved into 90% and 70% respectively. Further analysis is required to determine how does the selected features and data-splitting affect the optimal parameters, K and dist function for KNN, Depth and cost function on decision tree. We also test if the parameter leaf numbers could affect the accuracy, and how does pruning increases the decision tree accuracy.

3 Datasets

Hepatitis datasets: Starts with 155 observations in Hepatitis, after dropping records containing missing values there is 80 observations left. To understand the data, binary features are plotted as pie chart, numerical features are plotted as box plots. Thus we could observe if the range of numerical feature of the two class differs, a higher difference indicates greater correlation of the feature with the class. The plots could be found in Appendix as figure 1 and figure 2.

The correlated features observed from the plots are MALAISE, SPIDERS, ASCITES, HISTOLOGY, ALBUMIN, PROTIME.

Messidor datasets: Messidor datasets contains 1151 observations, after dropping None values 384 observations. The correlated features selected from the plots includes: feature 12, 13, 14, 15.

4 Results

4.1 Hepatitis

Both data are trained first with all features then with our selected features in the second run to see if our selection improves the model accuracy.

Hepatitis datasets:

All features: Hepatitis datasets is divided into 60 training cases and 20 test cases. To tune the parameters, the training cases is used to perform a ten-fold cross-validation.

Due to the random splitting of testing and training data, and considering the small sample size and imbalanced categorization, we tried several splitting, in the plotting we're showing trial 1, 2, 3. With the first row comes from KNN cross validation and the second row comes from decision tree cross validation. Thus, Models from the same trial (column) are derived from same set of data.(Figure 1)

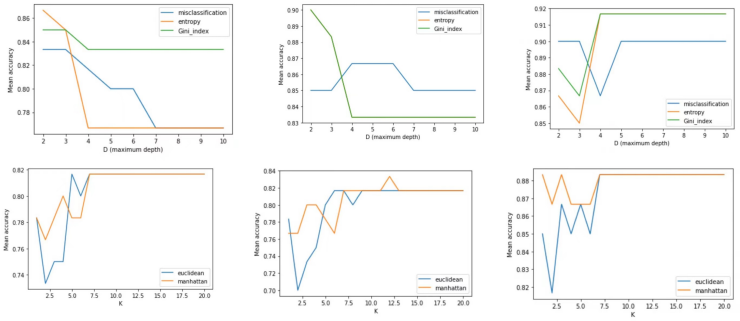

Figure 1: Cross validation -hepatitis -all features

Accordingly, each model perform on the test cases in three trials with parameters tuned according to its cross validation plots.(Figure 2)

	Trial 1	Trial 2	Trial 3
K	9	7	9
Dist function	Manhattan	Manhattan	Manhattan
Knn accuracy	70%	90%	90%
Depth	5	5	5
Cost function	Gini	Misclassification	Gini
Decision Tree accuracy	75%	85%	75%
After prune	75%	85%	80%

Figure 2: Test accuracy - hepatitis - all feature

The above experiments are repeated for selected features: MALAISE, SPIDERS, ASCITES, HISTOLOGY, ALBUMIN, PROTIME. (Figure 3, Figure 4)

When we select features, the cross validation plots shows the selection of parameters and trends follows the first run, except now it could be clearly observe that with increasing K, the accuracy would decrease due to under-fitting.

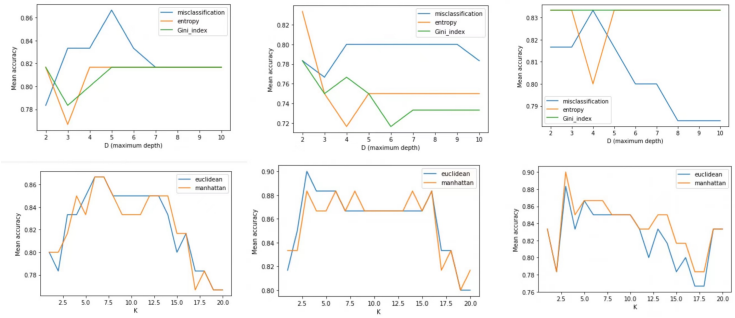

Figure 3: Cross validation -hepatitis -selected features

	Trial 1	Trial 2	Trial 3
K	5	3	5
Dist function	Manhattan	Euclidean	Manhattan
Knn accuracy	85%	85%	90%
Depth	4	5	5
Cost function	Misclassification	Misclassification	Gini
Decision Tree accuracy	95%	90%	90%
After prune	95%	90%	90%

Figure 4: Test cases - hepatitis -selected features

4.2 Messidor

First run with all features: (Figure 5,Figure 6)

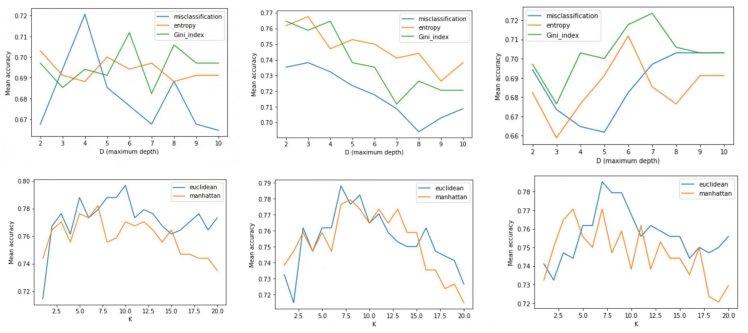

Figure 5: Cross validation - messidor - all features

	Trial 1	Trial 2	Trial 3
K	11	7	7
Dist function	Euclidean	Euclidean	Euclidean
Knn accuracy	74%	76%	91%
Depth	6	3	7
Cost function	Gini	Entropy	Gini
Decision Tree accuracy	69%	60%	81%
After prune	69%	60%	81%

Figure 6: Test accuracy - messidor - all features

Second run with selected features: 12, 13, 14, 15. (Figure 7, Figure 8)

4.3 Decision boundary on promising features

We use two promising features to plot decision boundary for KNN and decistion tree for each data sets. It is observed that with optimal parameters, KNN have a clearner decision boundary on decision tree with K = 5. Both achieve same performance as to distinguish class based on selected features in Messidor dataset. (Figure 9)

4.4 Messi -Leaf number

We further tested how the leaf instances affect the model accuracy, it is found when the depth remains constant, the accuracy increases first then decrease as the max leaf instance number increases. (Figure 10)

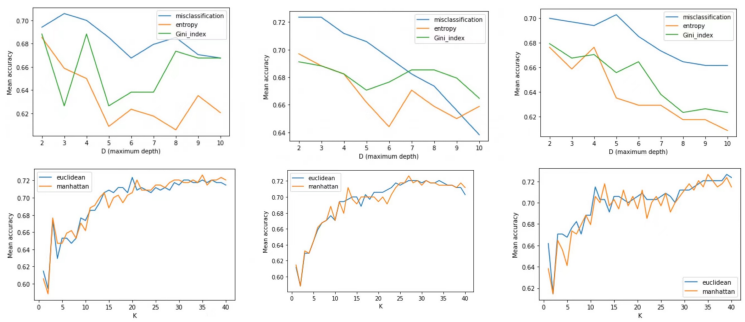

Figure 7: Cross validation - messidor - selected features

	Trial 1	Trial 2	Trial 3	
K	35	27	35	
Dist function	Manhattan	Manhattan	Manhattan	
Knn accuracy	67%	71%	67%	
Depth	3	3	5	
Cost function	Misclassification	Misclassification	Misclassification	
Decision Tree accuracy	71%	71%	69%	
After prune	71%	71%	71%	

Figure 8: Test accuracy - messidor - selected features

5 Discussion

5.1 Hepatitis

Across trials, We observe that Manhattan distance has greater accuracy when K is small. Both would goes down and up as K increase from 1 to around 7.5 and then reaches plateau and with same accuracy. The plateau may due to the imbalanced nature of the classification such that when a greater radius is considered, the proportion of classification as 1 would always exceeds 0.

For decision tree model, it is harder to determine the best cost function but it is consistent that the accuracy oscillates before reaches depth 5. It is observed when more trials is tested on cross validation, entropy and Gini index line may overlap, any one of them may have the greatest accuracy. This may again due to the small sample size and binary nature of most features.

The accuracy of decision tree and KNN model have level of accuracy with no more that 5% of deviation most of the time. It is also found the pruning process won't improve the accuracy significantly at this scale.

The main observation is that with selected features both models show increased accuracy which supports our selection of correlated features.

5.2 Messidor

It is observed with messidor dataset, with respect to KNN modal, Euclidean distance function is in favor overall. The shape of two cross-validation lines follows the up and down trend with increasing K showing that a moderate K is needed for avoiding under-fitting and over-fitting.

The accuracy when selecitng all features are around 0 mostly, with both modal have accuracy sudden jumps to 5% around, again suggests the accuracy depends on the data-splitting.

The observation with selected features are showing some patterns. With significant K required, both 35 K gives 67% accuracy. On the contrary, the depth of the tree decrease to 3 which gives 71% accuracy. For the last trials when using 5 maximum depth, the accuse is 69% but the pruning process bring the value back to 71%. The choice of parameters is also consistent, misclassification cost function is always chosen. Manhattan and Euclidean cost function have same trend.

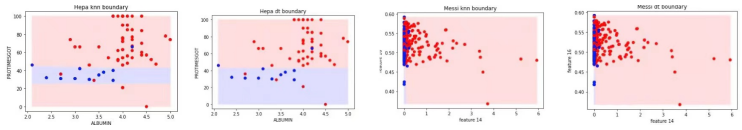

Figure 9: Decision boundary chosen from several parameters

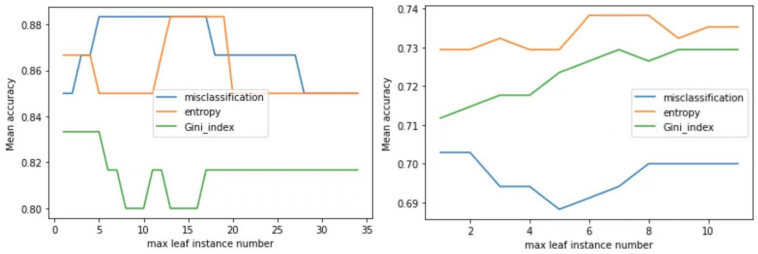

Figure 10: Max leaf instance on accuracy

It could be estimated with these feature selections, the distribution of underlying data sets lessen the effect of data-splitting and gives the consistent optimal prediction it could find.

6 Statement of Contributions

All works are distributed equally with group participation and discussion.

7 Appendix

Figure 11: Hepatitis binary features

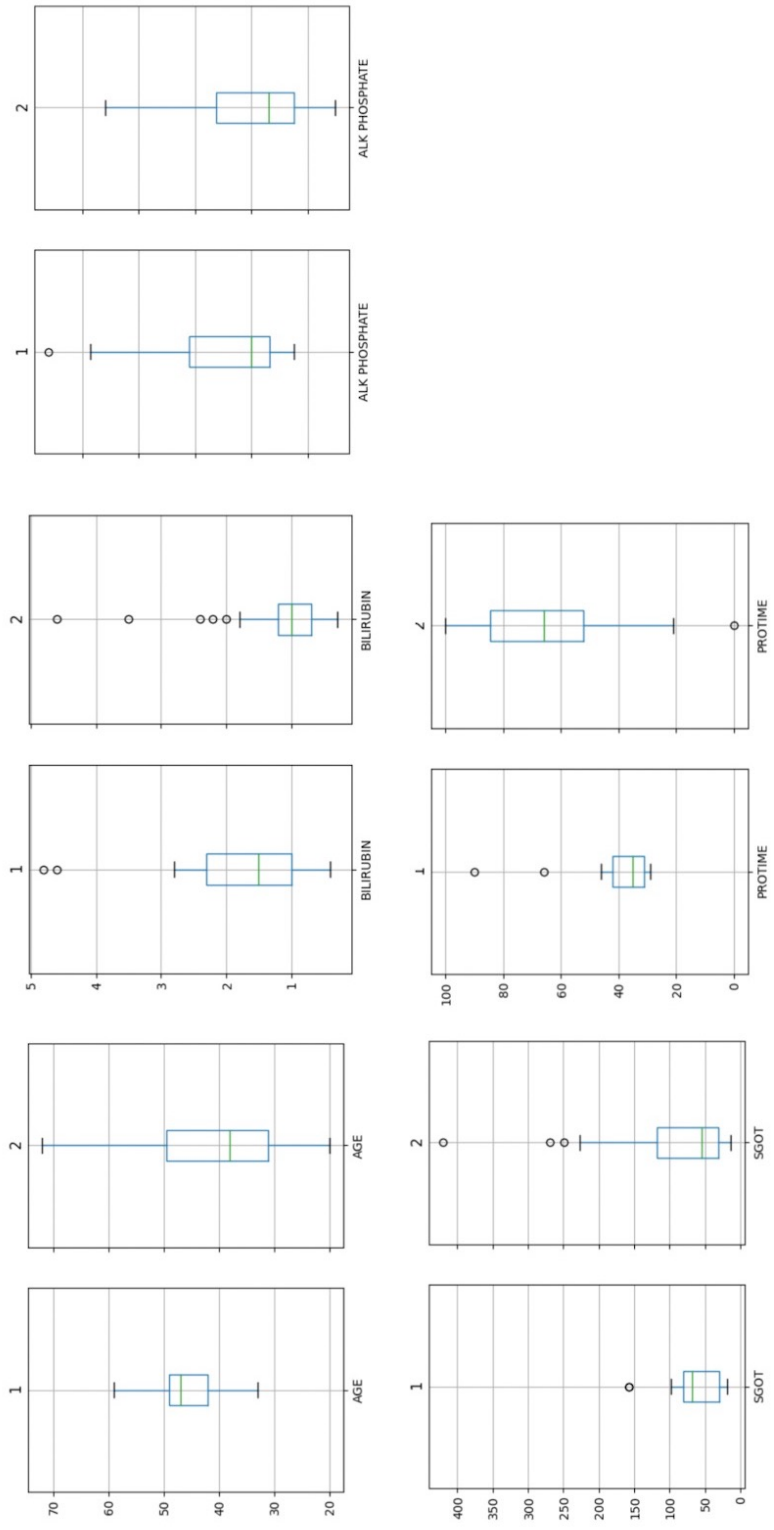

Figure 12: Hepatitis numerical features

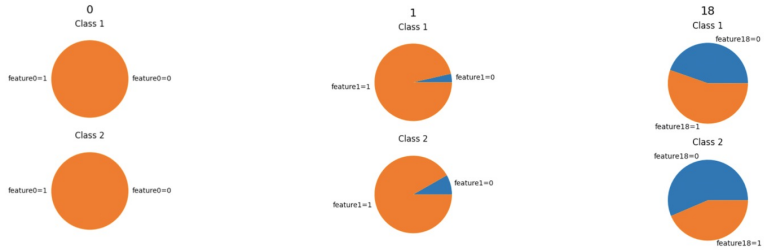

Figure 13: Messidor binary features

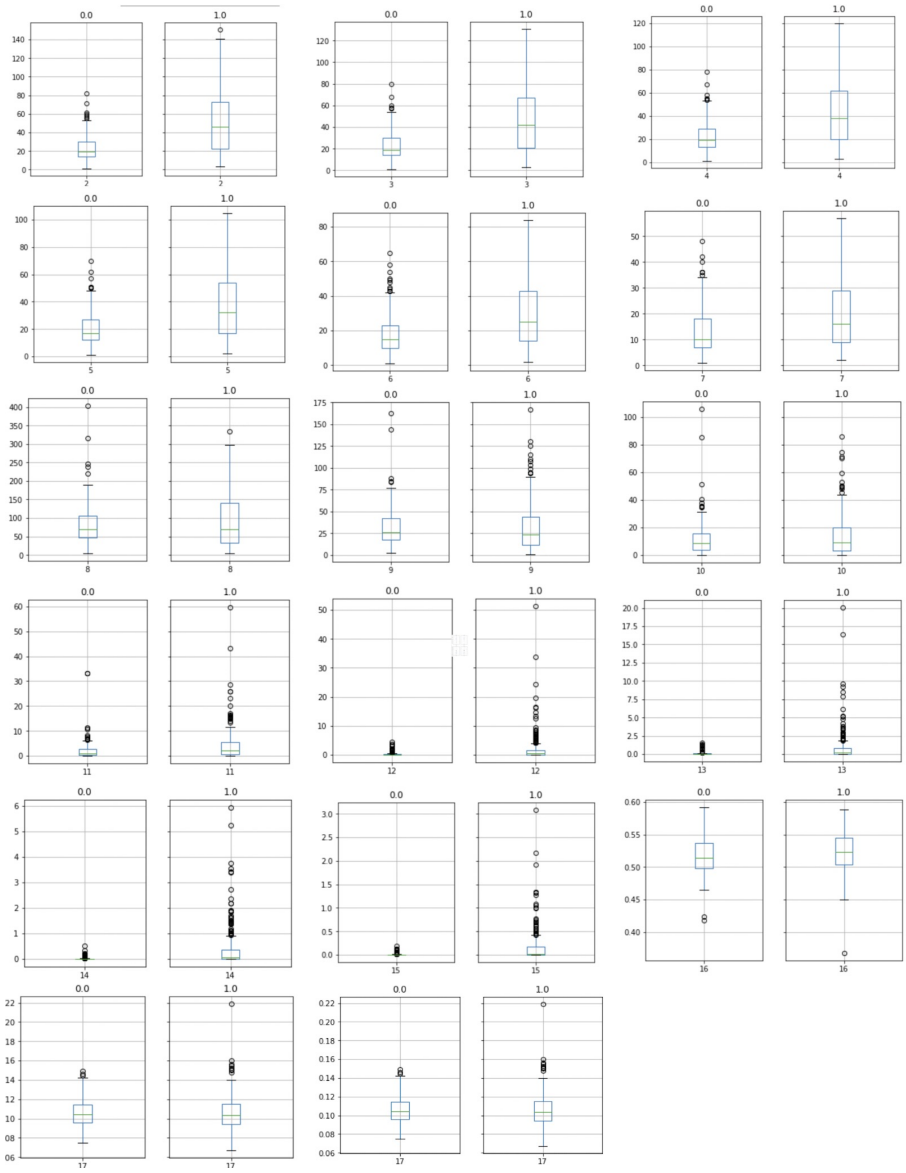

Figure 14: Messidor numerical features