Comparative Study of KNN and Decision Tree Classification Techniques on two health datasets

AUER Hans, HAWARRI Nadia, MENDILAHARZU Malena

Abstract: Classification algorithms in machine learning use input training data to predict the likelihood with which predetermined categories the subsequent data fall into. Classification algorithms can be applied to many fields such as the diagnosis of medical diseases, the analysis of social networks, classifying emails into "spam" or "non-spam" categories, and artificial intelligence. In this project we investigated the performance of two classification techniques K-Nearest Neighbor and Decision Trees on two health datasets. What we discovered was that both the K-NN algorithm and Decision Trees algorithm achieved high accuracies, however, we concluded that the Decision Tree algorithm would seem to be better at predicting classifications that K-NN for large databases.

Introduction: The rapid development of medical informatization has led to the use of classification algorithms for the detection of diseases. This paper focuses on the comparative study of KNN and Decision Tree for the detection of breast cancer and hepatitis. Research has shown "that KNN outperforms decision trees when it comes to rare occurrences ie. if you are classifying types of cancer in the general population, many cancers are quite rare, therefore a decision tree will almost certainly prune those important classes out of the model" [1]. Due to its simplicity, the lazy learning algorithm KNN, is commonly used for small databases. This algorithm is based on the principle that similar samples are generally close to each other. It requires less computation time during the training phase but more computation during the actual classification process. This is why, lazy learning is not always optimal for big data bases. Eager-learning algorithms' time complexity and efficiency is preferable for large datasets. The Decision Tree algorithm is an example of an Eager-learning algorithm. As expected, we found that the Decision Tree algorithm does better in large databases (in our case the Breast cancer database) compared to the KNN algorithm. This conclusion was drawn from the fact that there were more occurrences of misclassified data points in the decision boundaries for KNN than for the Decision Trees.

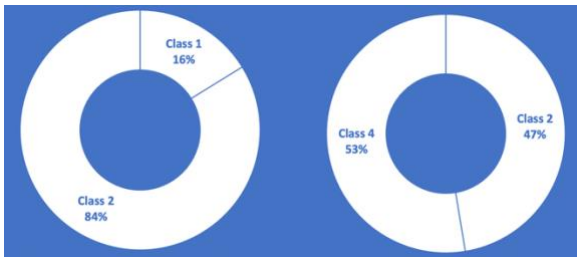


Fig: Percentage of each class in clean data (Breast Cancer (right), Hepatitis (left))

Datasets: The Breast Cancer dataset contained information about nine different features that allow to classify whether the cancer was benign or malignant. The hepatitis dataset described nineteen features for the classification between hepatitis 1 and 2. Both datasets were preprocessed by removing rows with missing values, duplicates and outliers (more than three std away from the mean). Outliers were identified through boxplots.

# of rows and dataset	Breast Cancer	Hepatitis
# rows original data	699	155
After removing '?'	683	80
Removing duplicates	449	80
Cleaning outliers	435	74

Since errors in the diagnosis of diseases should be avoided at all cost, we need to make sure that we are correctly classifying the data. In order to do so, we need a diverse dataset taking into account different genders and ethnicities. By doing so, we will be avoiding biases in the data that lead to misclassifications. Note that none of these databases provide this information.

	Clump_Thickness	Uniformity_of_Cell_Size	Uniformity_of_Cell_Shape	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses
mean	5.378619	4.222717	4.273942	3.746102	3.879733	4.806236	4.200445	3.828508	1.913140
std	2.869029	3.251280	3.141494	3.158413	2.456544	3.880509	2.651634	3.387146	2.068909
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
max	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
correlation	0.670230	0.758697	0.759500	0.630415	0.611432	0.760278	0.706738	0.645257	0.357184

Fig: Summary Statistics for Breast Cancer

	AGE	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA	LIVER_BIG	LIVER_FIRM	SPLEEN_PALPABLE	SPIDERS	ASCITES	VARICES	BILIRUBIN	ALK_PHOSPHATE	SGOT	ALBUMIN	PROTIME	HISTOLOGY
mean	40.662500	1.137500	1.525000	1.737500	1.350000	1.612500	1.850000	1.837500	1.525000	1.812500	1.687500	1.850000	1.875000	1.221250	102.912500	82.025000	3.843750	62.512500	1.412500
std	11.280030	0.346547	0.502525	0.442769	0.479979	0.490253	0.359324	0.371236	0.502525	0.392775	0.466437	0.359324	0.332805	0.875213	53.684779	71.599974	0.576292	23.427774	0.495390
min	20.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.300000	26.000000	14.000000	2.100000	0.000000	1.000000
max	72.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	4.800000	280.000000	420.000000	5.000000	100.000000	2.000000
correlation	-0.212769	0.175876	0.123830	-0.108776	0.181151	0.275595	-0.185042	-0.194030	0.055978	0.135643	0.287839	0.479211	0.345785	-0.351557	-0.189360	0.078731	0.477404	0.395386	-0.456856

Fig: Summary Statistics for Hepatitis

Splitting Data

One of the first issues we had to tackle was to decide what percentage of the data should be used for training/testing. According to Jeremy Jordan, "a typical train/test split would be to use 70% of the data for training and 30% of the data for testing". In order to verify that these percentages would be optimal for our data, we varied the train percentage and observed which percentage maximized the accuracy.

[2]

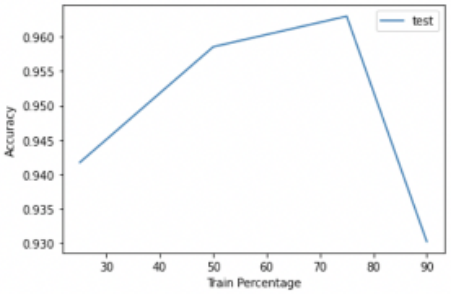


Fig 1.1: Accuracy vs Train Percentage for the Breast Cancer Data

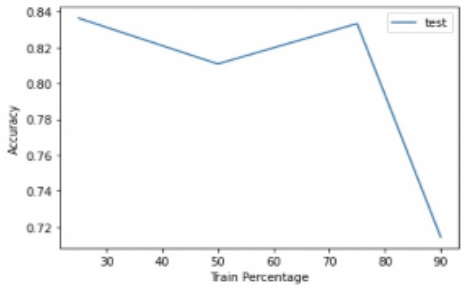


Fig 1.2: Accuracy vs Train Percentage for the Hepatitis Data

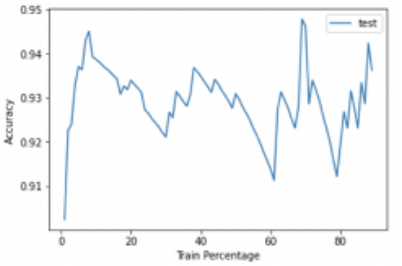


Fig 1.3: Accuracy vs Train Percentage for the Breast Cancer Data

In Fig1.1 and Fig1.2, we passed 25, 50, 75 and 90% of the data to the training model as the training set. We then observed the changes in accuracy which were aligned with our hypothesis: both figures present the maximum accuracy around the 70% mark. In Fig1.3, we decided to not only pass four percentages but to pass approximately 90 values. We can see that this is again consistent with our hypothesis.

KNN – k value

We wanted to test how different values for k affected the training and data accuracy: We expected that if K was too large there would be underfitting while if it was too small overfitting would occur. Underfitting means that the model is not efficient in modelling the training data nor in generalizing to the testing data. An underfit model will have poor performance on the training data. This is consistent with the graphs below: as K increases, the accuracy of the training and the testing data decreases. Overfitting refers to when the model models the training data too well. This leads to bad generalizations and therefore lower accuracy in our testing data. Again, this is consistent with our graphs. For very small values of K the training data accuracy is very high while the testing data accuracy is small.

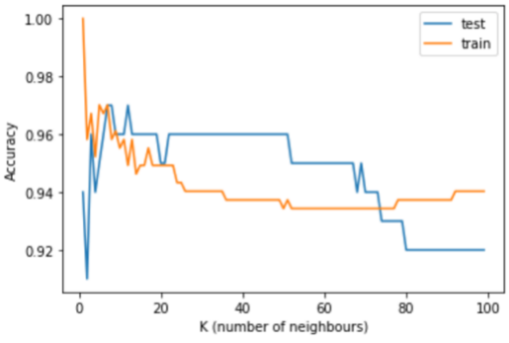


Fig 2.1: Accuracy vs Train Percentage for the Breast Cancer after removing outliers

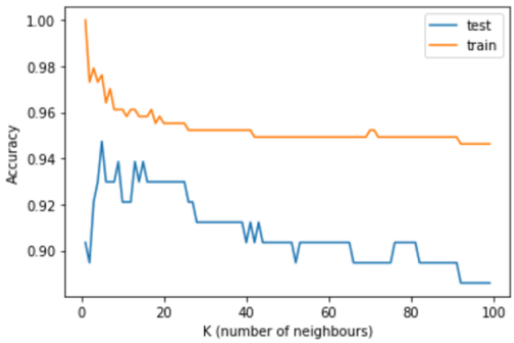


Fig 2.2: Accuracy vs Train Percentage for the Breast Cancer before removing outliers

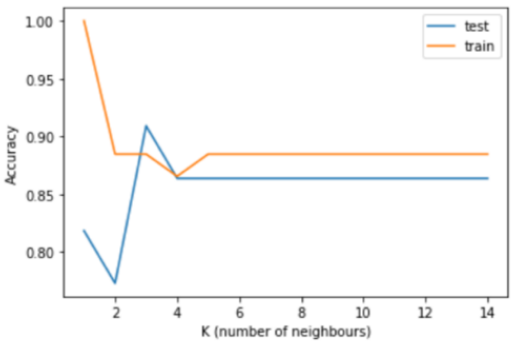


Fig 2.3: Accuracy vs Train Percentage for Hepatitis after removing outliers

Fig 3.1: Accuracy vs Distance Function for Breast

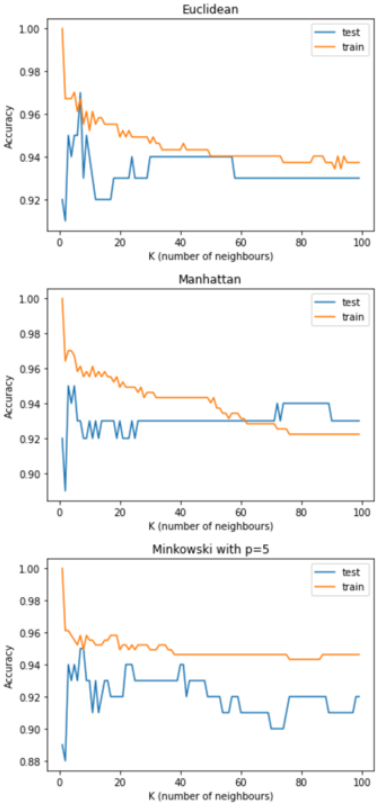
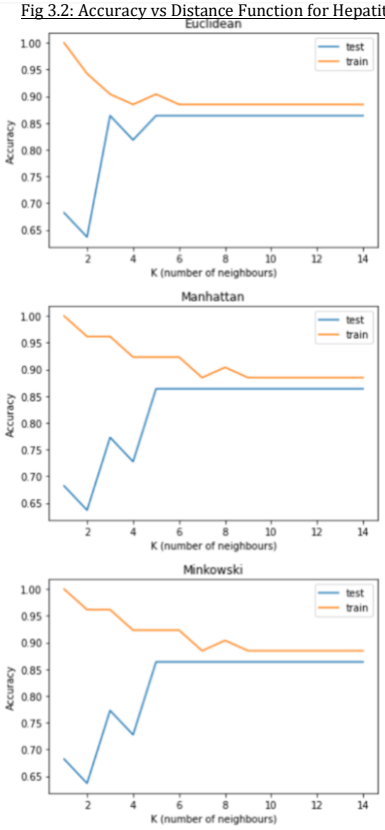


Fig 3.2: Accuracy vs Distance Function for Hepatitis



KNN - Distance Functions

Our goal was to test whether different distance functions affected the training and data accuracy. To do so, we ran the algorithm using three different distance functions: Euclidean, Manhattan and Minkowski (with p=5).

Both for the Breast Cancer and the Hepatitis dataset, no significant differences in the accuracy were observed. For this reason, we cannot conclude that there is one distance function that is better than the other two.

One thing we noticed was that there was less change for the smaller database (Hepatitis).

KNN - Features

We wanted to test whether the different features selected for the training process affected the test results.

For both datasets we monitored the change in accuracy when we considered all the features vs. the two features with the highest correlations vs. the two features with the lowest correlations.

For the Breast Cancer Dataset, we found that the accuracy did not significantly change when we considered the features with the two highest correlations vs all the features. This could be due to the fact that the majority of the features in the dataset have a high correlation. However, regarding the accuracy of the two features with the lowest correlations a decrease in the accuracy was observed which was consistent with our expectations.

For the Hepatitis dataset there was an increase in the accuracy when we considered the features with the highest correlations vs all the features. There was a decrease in the accuracy when using the features with the lowest correlations.

Fig 4.1: Accuracy vs Features selected Breast Cancer

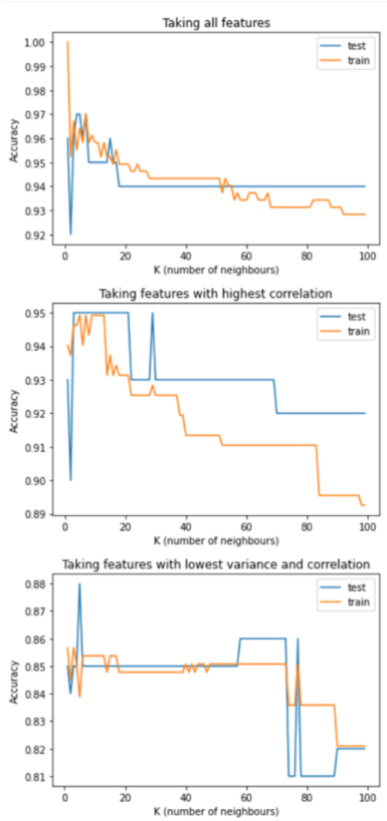
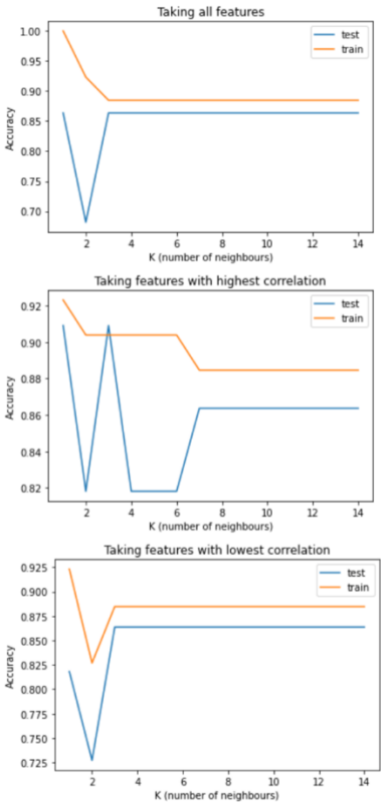


Fig 4.2: Accuracy vs Features selected Hepatitis



Decision Boundaries

For each algorithm and each data set, the decision boundaries were plotted using the two features with the highest correlations and lowest variance (uniformity of cell shape and size for breast cancer and albumin and protime for hepatitis). Points in the same-colored region belong to the same class. Points that are of a different color than their surrounding region correspond to misclassified data. Hence, decision boundaries allow us to drag conclusions about the accuracy of the model. Notice that only with the Decision Tree algorithm we get 0 misclassified points.

Fig 5.1: Decision boundary for Breast Cancer (K=2,4,11)

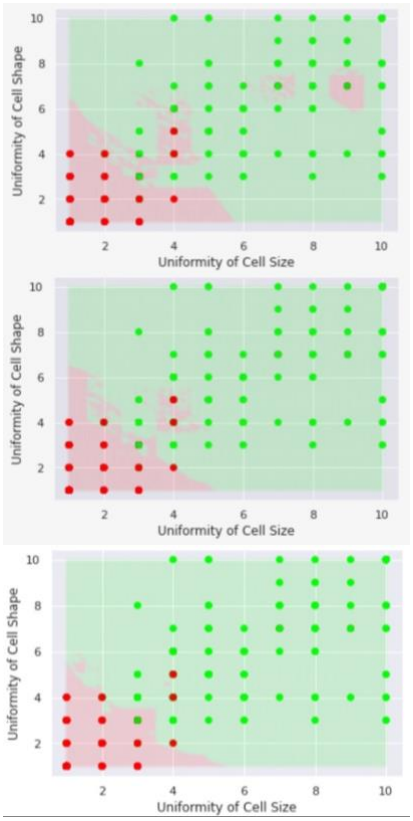
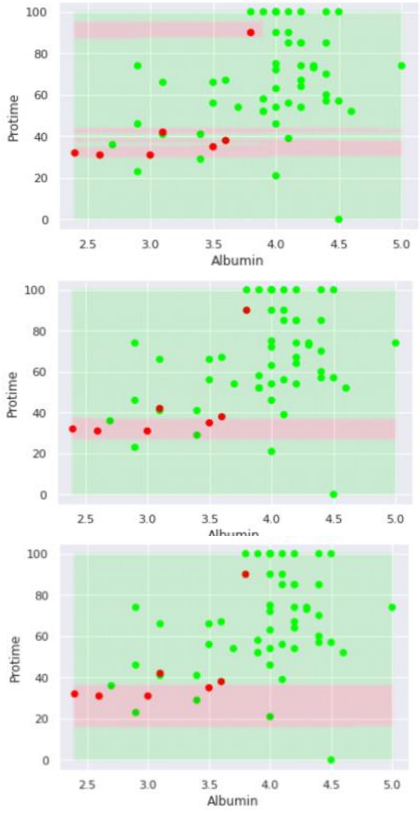


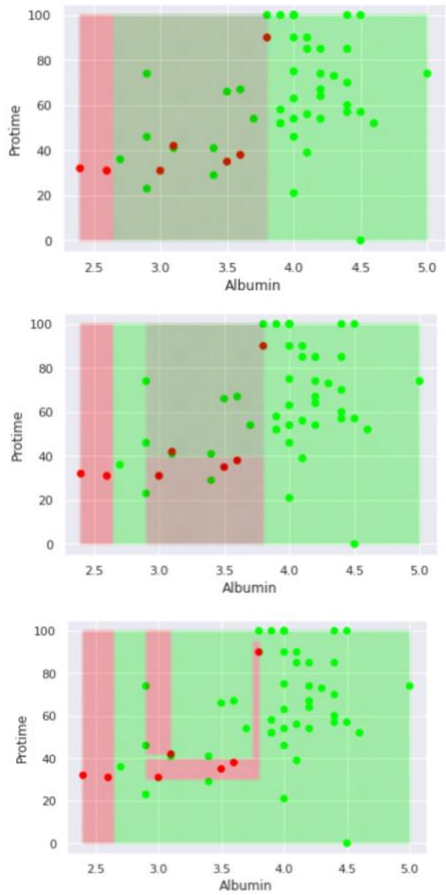
Fig 5.2: Decision boundary for Breast Cancer (K=1,3,7)



KNN - Decision Boundaries

We notice that the value of k affects the percentage of misclassified points and hence the accuracy of the model. For small/large values of k the accuracy is not as big as for moderate values of k.

Fig 5.3: Decision boundary for Hepatitis (Depth=2,4,10)



We analyzed the decision boundaries for the breast cancer and hepatitis datasets, at varying depths and at the Gini Index cost function. The boundary illustrations for the decision trees above have grid like boundaries for predicting the label of the data. This seems more intuitive for humans to rationalized/reason with, compared to a nearest neighbor analysis, further suggesting its higher interpretability. Moreover, the decision tree boundaries for both datasets have a consistent inverse correlation between the number of outliers (data points which are found within the wrong boundaries) and the depth of the tree. It is interesting to compare these tendencies with those of the KNN boundaries, where the number of outliers seems to drop and then increase again. This could be a result of the KNN boundaries having more difficulty with larger datasets as opposed to decision tree boundaries.

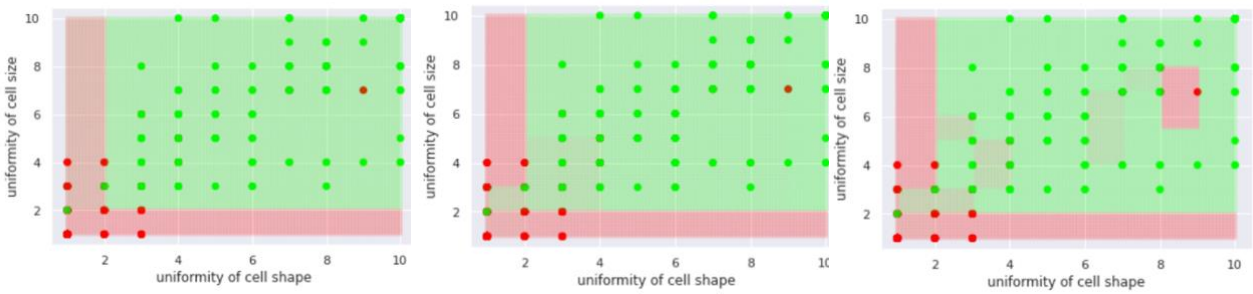


Fig 5.4: Decision boundary for Breast Cancer (Depth=2,4,10)

Decision Trees – Depth and Cost Functions

Fig 6.1: Accuracies vs Tree depth for Hepatitis

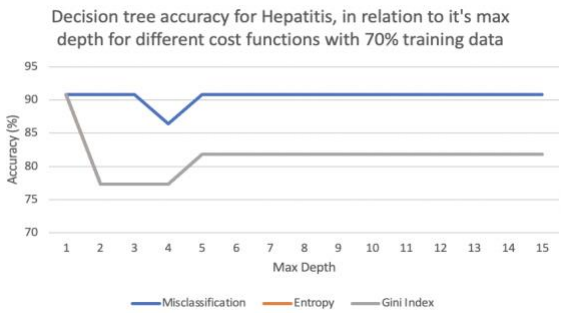
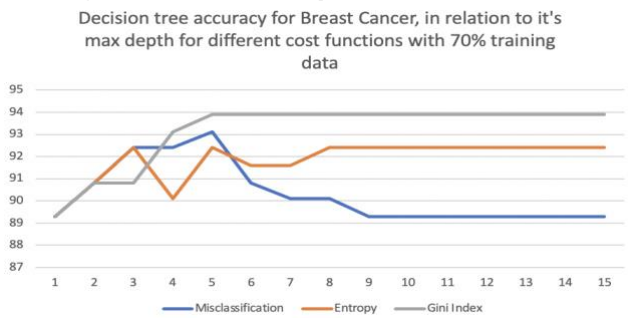


Fig 6.2: Accuracies vs Tree depth for Breast Cancer



***For Fig 6.1 the entropy and Gini Index are overlapping.**

By comparing the different tree depths within each respective cost function, the accuracy for breast cancer seems to always be highest around depth 4 and 5 and then decreases relative to the distance of those points. This is due to the tendency of the algorithm to overfit and become too specific to the training data when the tree depth is too high. Similarly, the smaller depths of the decision tree tend to underfit the data. On the other hand, we notice almost random trends for the hepatitis data. This data set, containing under 100 instances, is significantly smaller than the breast cancer dataset and therefore it is much harder for the algorithms to represent the data in a consistent manner, especially Decision Trees, whose design contains an affinity towards handling larger datasets. This also explains why we observed significant changes in the accuracies every time we generated new random subgroups of training and testing data. Furthermore, it was also a difficult decision for our team to decide whether we should get rid of outliers in the hepatitis data, since this reduces the already limited information.

We noticed for the breast cancer that the maximum accuracy throughout the tree depths were highest for Gini index cost function and lowest for Misclassification function. This is representative of the Gini Index's generally better heuristic.

Comparison of the two algorithms

Upon analysis of Fig 3 & 6 we noticed similarities in the breast cancer data between KNN and Decision trees. The test data for both algorithms reach global maximums within the first 10 k-distances and 10 tree depths. This suggests an equal susceptibility to under and overfitting for the algorithms. Furthermore, In a comparative study of KNN and Decision trees, Harikumar Rajaguru says: "The presence of irrelevant features (noise) in the input dataset will reduce the accuracy or precision of KNN algorithm even with its high efficiency of classification". Consequently, the graphs shown in Fig 3 & 6 were all representations of accuracy of our cleaned datasets. This explains why the KNN algorithms do not show a weaker accuracy compared to the more robust decision tree algorithm.

Conclusion & Possible Improvements

In this report we analyzed the effects of hyper-parameters on K-NN and decision tree classification. One thing we can add onto this analysis is finding the optimal value of the hyper-parameters by tuning them on a validation set. Furthermore, perhaps selecting a better database for Hepatitis would further show patterns and improve our classification. The reason for this is that the Hepatitis database consists of 84% class 2, and 16% class 1. There is clearly an imbalance between the amount of datapoint for class 1 and 2. This could lead to an increase in misclassifications. Furthermore, since one of the purposes of this report was to analyze the differences between KNN and decision trees, we could add to this by analyzing the time taken for each algorithm on each database. Finally, we could also consider using principal feature component analysis for feature selection.

In conclusion, we found that there is a slight improvement when using decision trees for large databases such as the Breast Cancer database compared to smaller ones like the Hepatitis one. However, it is difficult to know without further testing if these results are accurate or not. One should run the same experiments taking into consideration the aforementioned possible improvements as well as averaging multiple runs to avoid biases due to the randomness of the permutations of the data.

Works Cited:

[1] Stephanie Glen "Comparing Classifiers: Decision Trees, K-NN & Naive Bayes" *Data Science Central,* 19 June 2019.

[2] Jeremy Jordan "Evaluating a machine learning model" *Jeremy Jordan*, 21 July 2017.

[3] Rajaguru, Harikumar, and Sannasi Chakravarthy S R. "Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer." *Asian Pacific journal of cancer prevention: APJCP* vol. 20,12 3777-3781. 1 Dec. 2019, doi:10.31557/APJCP.2019.20.12.3777

Contributions

Nadia and Malena:

- Cleaning data and statistics
- KNN algorithm (code and analysis)
- Report

Hans:

- Decision trees algorithms (Code and analysis)
- Comparison of the two algorithms