Group 1: Mini Project 1

COMP 551 Winter 2021

Arneet Kalra, Nicholas Kiriazis and Sierra Schena

1 Abstract

The discussion surrounding the value machine learning algorithms can add in improving the accuracy of medical diagnoses is becoming evermore popular. In this project, we investigated how well two machine learning models, K-Nearest Neighbours and Decision Trees, performed on two medical data sets: Breast Cancer and Hepatitis. Both models performed best on the Breast Cancer set, a larger and more complete dataset, achieving over 95% in both training and testing accuracies. The models were not as successful in classifying instances from the Hepatitis dataset, as both algorithms resulted in an overfitting model, which we suspect is due to the lack of usable instances and numerous missing entries. Common to both datasets, we found that the Decision Tree algorithm achieved worse accuracy than the K-Nearest Neighbour.

2 Introduction

The goal of this project is to implement and compare the two classification techniques of K-Nearest Neighbours and Decision Trees using two distinct health datasets, Breast Cancer and Hepatitis. The Breast Cancer dataset represents the work of Dr. Wolberg, whose goal was to accurately diagnose breast masses as benign or malignant based solely on a Fine Needle Aspiration (FNA) (Mangasarian and Wolberg, 2005). The Hepatitis dataset was donated by Gail Gong of Carnegie-Mellon University in order to try to predict patient survivability from hepatitis (UCI Hepatitis Data Set). The very high testing and training accuracies obtained by both approaches on the Breast Cancer dataset demonstrates the importance of a large, complete dataset for optimal classification. The algorithms could not obtain nearly the same accuracy on the Hepatitis dataset, outputting a testing accuracy less than 80%, due to the restriction of instances and missing values. Investigation of the relevance of the 'ALK PHOSPHATE' feature was inconclusive since the testing accuracy between these subsets varied by over 15% using the Decision Tree, but were similar using KNN. The disparity between the training and testing accuracies provided by the Decision Trees approach shows that the KNN models are better for classifying these datasets.

3 Datasets

The Breast Cancer data set originally consists of 699 instances and 10 attributes: the first one being an ID number followed by nine integer valued attributes that describe characteristics from an FNA sample (UCI Breast Cancer Data Set). The Hepatitis data set, in constrast, is a smaller data set originally consisting of 155 instances and 19 attributes: 6 of which are continuous variables, while 13 are binary categorical variables relating to an individual and their liver function (UCI Hepatitis Data Set).

To begin processing the data sets, we viewed how many blank entries were in each column. The Breast Cancer dataset only had 16 instances containing blank entries, and only from the 'Bare Nuclei' column, thus we eliminated the rows corresponding to these blank entries. Conversely, the cleaning process for the Hepatitis dataset was much more involved. If all of the rows that contained blank entries were to be removed from the Hepatitis dataset, 75 instances or 48.4% of the dataset would have been removed. We observed that 67 entries from the 'PROTIME' column were blank. Therefore, we decided to remove this attribute from the data. As well, we noticed that the attribute 'ALK PHOSPHATE' also contained 29 blank entries, and since our dataset contains only 155 instances, keeping this attribute in the data would cause us to remove 18.7% of the instances in our data. Therefore, we chose to save two versions of the cleaned Hepatitis dataset to use in our analysis - one keeping the 'ALK PHOSPHATE' column, and one removing it to see whether or not this attribute would be significant in terms of improving accuracy. The last step we took in cleaning both versions of the cleaned Hepatitis data was to remove the rest of the rows that contained blank entries.

After the data was processed, we were ready to examine the distribution of our attributes and labels. Beginning with the Breast Cancer dataset, there were almost twice as many Benign samples compared to Malignant samples (444 Benign labels and 239 Malignant labels) as can be seen in Figure 1. As can be seen in Figure 2, the attributes the most highly correlated with Class are Uniformity of Cell Size, Uniformity of Cell Shape, and Bare Nuclei.

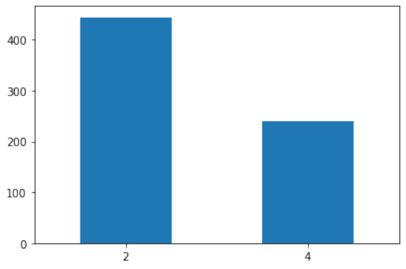


Figure 1: Distribution of Benign and Malignant labels

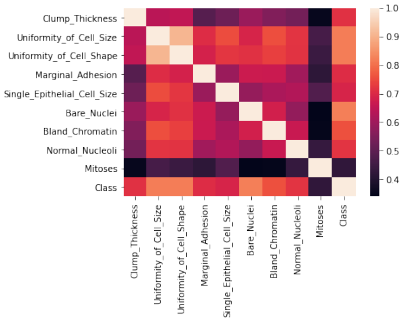


Figure 2: Breast Cancer Heat Map

The distributions of Uniformity of Cell Size and Shape appear to have the same association with Class: if cell size and shape are less uniform, then these samples are more strongly associated with being benign, but if cell size and shape are more uniform, then the samples are more associated with being malignant. The distribution of Bare Nuclei also follows the same pattern; low counts of bare nuclei are associated with benign samples, whereas high counts of bare nuclei are associated with malignant samples.

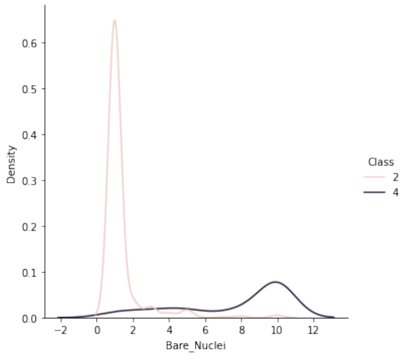


Figure 3: Distribution of Bare Nuclei

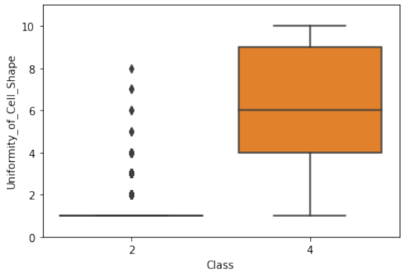


Figure 4: Distribution of Uniformity of Cell Shape

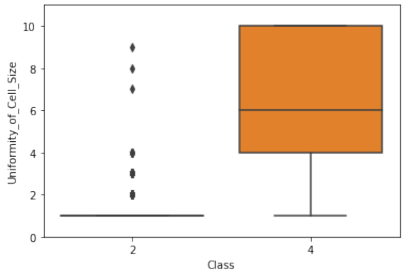


Figure 5: Distribution of Uniformity of Cell Size

We then began to analyze the relationships between the attributes and labels of the Hepatitis dataset. The proportion of instances that correspond to the 'Die' Class label is about 18% in either of the cleaned versions of the dataset, which corresponds to about 82% of the instances having the 'Live' Class label. As can be seen from Figure 6, the feature 'ASCITES' has the strongest correlation with Class. If we observe directly the distribution of 'ASCITES' in relation to Class, we notice that 'ASCITES' and patient death does not have a strong relationship, but 'ASCITES' being marked as 'yes' is strongly associated with patient survival.

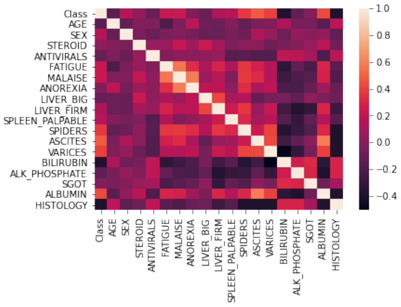


Figure 6: Hepatitis Heat Map

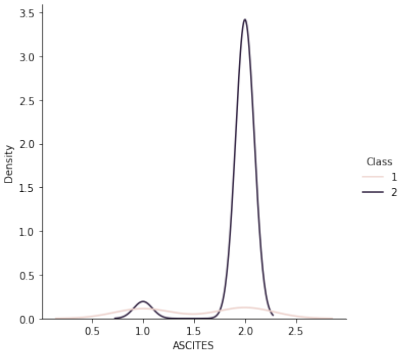
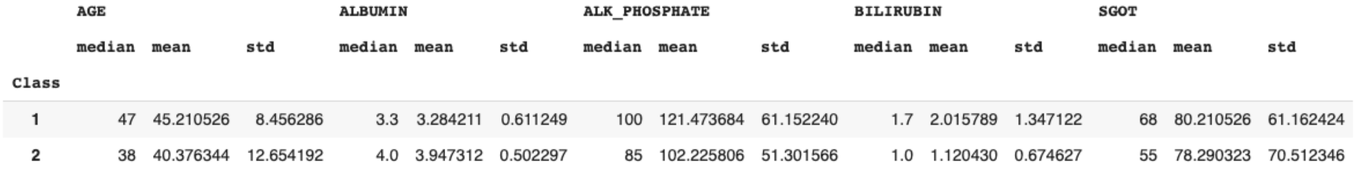
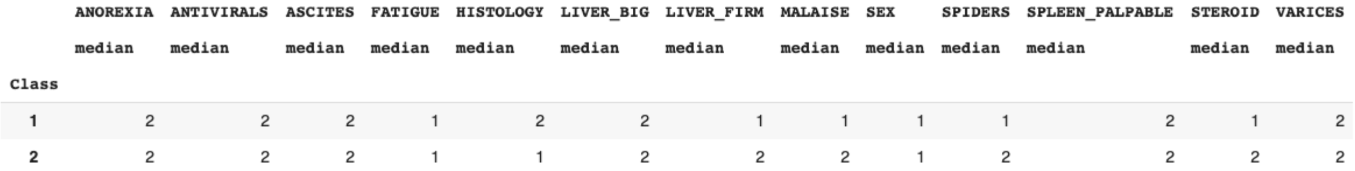


Figure 7: Distribution of Ascites

Since the Hepatitis Dataset contained both continuous and categorical variables, we also calculated various numerical summaries to analyze the relationships between each feature with Class that can be found in the tables below.



	AGE			ALBUMIN			ALK_PHOSPHATE			BILIRUBIN			SGOT		
	median	mean	std	median	mean	std	median	mean	std	median	mean	std	median	mean	std
Class															
1	47	45.210526	8.456286	3.3	3.284211	0.611249	100	121.473684	61.152240	1.7	2.015789	1.347122	68	80.210526	61.162424
2	38	40.376344	12.654192	4.0	3.947312	0.502297	85	102.225806	51.301566	1.0	1.120430	0.674627	55	78.290323	70.512346



	ANOREXIA	ANTIVIRALS	ASCITES	FATIGUE	HISTOLOGY	LIVER_BIG	LIVER_FIRM	MALAISE	SEX	SPIDERS	SPLEEN_PALPABLE	STEROID	VARICES
	median	median	median	median	median	median	median	median	median	median	median	median	median
Class													
1	2	2	2	1	2	2	1	1	1	1	2	1	2
2	2	2	2	1	1	2	2	2	1	2	2	2	2

It is important to note the potential ethical concerns that may arise when working with data surrounding patient's medical history. We must acknowledge the distribution of the data we have collected, and how it may lead to biased predictions. If we do not collect a representative sample of our population to train our models with, we will not make accurate predictions for those not largely represented in our data. As well, we must ensure that the data collection process is unbiased as to not perpetuate human bias into the data (Vayena,Blasimme and Cohen, 2018). Therefore, it is vital that these ethical concerns are thought about when analyzing data in order to avoid potential inaccurate diagnoses.

4 Results

4.1 Breast Cancer

We tested a variety of K-values for our KNN model. Figure 8 demonstrates the effect of changing K on the accuracy of classification on the validation set using different cost functions, namely Euclidean, Manhattan and Minkowski (p=3) distances. We concluded that K=10 provides our optimal solution, along with the Euclidean distance, although each of the distances achieved similar accuracy on our validation set. The training accuracy we obtained was 97.48% and our testing accuracy 97.81%. We can conclude that our model is fitting well as the accuracy on the training and test sets are high, and almost equivalent.

Figure 10 depicts the effects of changing the value of maximum depth and cost functions on the accuracy of the model. We found that a Decision Tree using the Gini Index as its cost function with a maximum depth of 4 provided the best model for our data, where we obtained 95.62% testing accuracy and 98.62% training accuracy.

The decision boundaries for both KNN and Decision trees can be found as Figures 9 and 11 respectively.

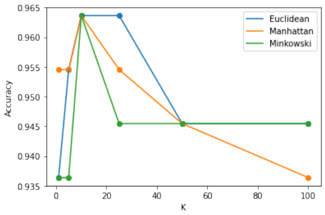


Figure 8: Accuracy of varying K-values

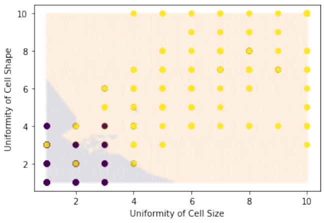


Figure 9: Decision Boundary of KNN

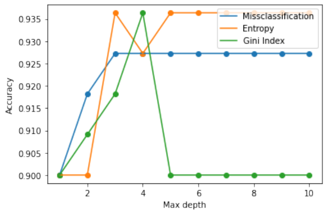


Figure 10: Accuracy vs Max Depth

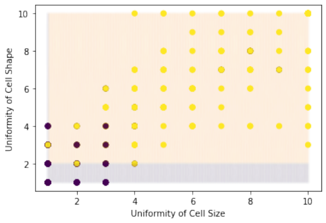


Figure 11: Decision Boundary of Decision Tree

4.2 Hepatitis

First, we will analyze the KNN algorithm applied to the Hepatitis dataset including the feature 'ALKPHOSPHATE'. As can be seen in Figure 12, K=5 and the Manhattan distance provide optimal parameters for our model. The testing accuracy found was 78.26% , while the training accuracy found was 83.10%, which seems to suggest that our model overfits the data.

If we now analyze the KNN algorithm applied to the Hepatitis dataset that does not include feature 'ALKPHOSPHATE', we chose K=10 as our optimal parameter. As shown in Figure 14, we do not find much of a difference with the different distance function, and simply chose the Euclidean distance. The test accuracy was found to be 76.92%, while the training accuracy was 84.14%. Again, this suggests some overfitting of our training data.

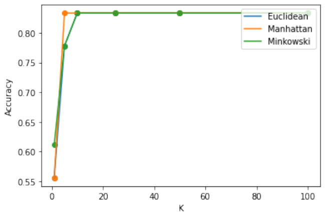


Figure 12: Accuracy vs K: Including AlkPhosphate

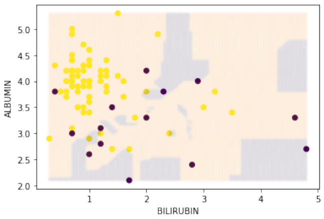


Figure 13: Decision Boundary: KNN Including AlkPhosphate

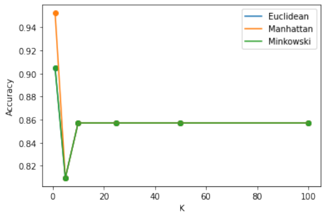


Figure 14: Accuracy vs K: Excluding AlkPhosphate

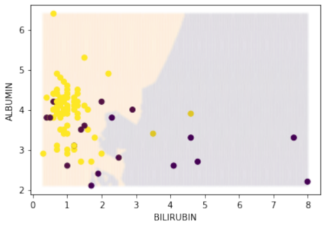


Figure 15: Decision Boundary: KNN Exluding AlkPhosphate

The chosen parameters for the dataset including the 'ALKPHOSPHATE' column were maximum depth equal to 3 and missclassification as the cost function. This model provided 69.57% for the test accuracy and 90.14% for the training accuracy. The same optimal parameters were found for the dataset excluding the 'ALKPHOSPHATE' column. The test accuracy found was 84.62% and 90.24% for the training accuracy.

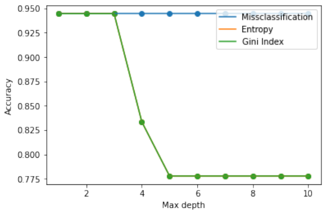


Figure 16: Accuracy vs Max Depth: Including AlkPhosphate

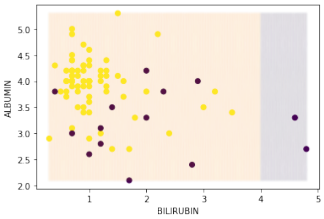


Figure 17: Decision Boundary: DT Including AlkPhosphate

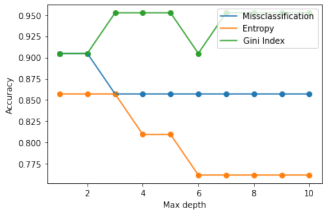


Figure 18: Accuracy vs Max Depth: Excluding AlkPhosphate

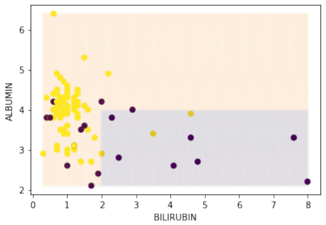


Figure 19: Decision Boundary: DT Exluding AlkPhosphate

5 Discussion and Conclusion

As can be seen from the results presented, both the KNN and Decision Tree algorithms performed best on the Breast Cancer dataset, each achieving above 95% accuracy on the unseen test data. As well, we saw that there was a smaller difference between the training and testing accuracies when using the Breast Cancer dataset compared to the Hepatitis dataset. The accuracies for the Decision Trees were generally lower than the KNNs for both datasets.

The results found from the comparison of the two cleaned versions of the Hepatitis dataset were inconclusive as to whether or not we should remove the feature 'ALK PHOSPHATE' in our dataset. The testing accuracy from using KNN on the data including 'ALK PHOSPHATE' was 1.34% higher, but more importantly it had closer testing and training accuracies (4.83% compared to 7.22%). However, the test accuracy found with the Decision Boundary algorithm was 15.05% higher for the data excluding 'ALK PHOSPHATE', with comparable training accuracies for both models. Such a drastic difference must be subject to further investigation before forming a conclusion. Due to the size of the Hepatitis dataset, it was difficult to tune the hyperparameters as the resulting accuracies were always very similar for different values of K or maximum depth. The resulting accuracies were then not as reliable due to the lack of data. As well, the Hepatitis dataset comprised of both categorical and continuous attributes, and only one measure of distance was used for classification. Potentially coding separate distance functions for the categorical and continuous variables would lead to more accurate predictions. As well, it would be of interest to investigate if pruning the Decision Tree and/or controlling for a number of internal nodes would increase the testing accuracy of the algorithm, or lessen the difference between testing and training accuracies, especially since the maximum depths chosen within our models were quite small.

As seen through the Decision Boundaries for the Breast Cancer dataset, Figures 9 and 11, the data is structured in a consistent manner. This allows for the model to predict the test set classes rather accurately. Furthermore, the boundary itself is not extremely different between the two algorithms. On the contrary, the decision boundaries for the Hepatitis datasets showcase the lower accuracies computed by the two algorithms. The plotted points are mixed together compared to the points of the Breast Cancer set, making it more difficult to make accurate predictions. Furthermore, the KNN boundary is much better than the Decision Tree boundary for this dataset as it is separated into more distinct sections.

6 Statement of Contributions

Arneet Kalra ran test experiments and assisted in the writing of the report. Nicholas Kiriazis implemented the KNN and Decision Tree algorithms. Sierra Schena cleaned both datasets and provided exploratory data analysis.

7 References

Dua, D. and Graff, C. (2017a). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Hepatitis]. Irvine, CA: University of California, School of Information and Computer Science.

Dua, D. and Graff, C. (2017b). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29]. Irvine, CA: University of California, School of Information and Computer Science.

Mangasarian, Prof. Olvi L. and Wolberg, Dr. William H. (2005). Machine Learning for Cancer Diagnosis and Prognosis. *University of Wisconsin-Madison.* http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html

Effy Vayena, Alessandro Blasimme and I. Glenn Cohen. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine.* https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002689