Mini Project 1 Report

Getting Started with Machine Learning

COMP 551: Applied Machine Learning

Maxime Buteau	260868661
Charles Couture	260923463
Eric Pelletier	260895863

McGill University
February 8th, 2022

Abstract

A key element in machine learning is the ability to predict outcomes with the highest accuracy possible. In this project, we investigated the performance of two machine learning models —K-Nearest Neighbour (KNN) and Decision Tree— on two benchmark datasets —hepatitis and diabetic retinopathy—. During the experiments, we tested different numbers of neighbours to consider for the KNN model to evaluate differences in training data accuracy and test data accuracy, as well as tested maximum tree depth to appraise the performance of the Decision Tree model. In addition, we evaluated the affects that different cost functions have on the models. Overall, we found that both KNN and Decision Tree perform very similarly on the given datasets, with the latter having a slight edge with the larger data set —hepatitis data set—.

1 Introduction

During this project, we implemented two classification techniques —K-Nearest Neighbour (KNN) and Decision Tree— and compared these two algorithms on two distinct health datasets. The datasets analyzed during this process were the *Hepatitis Dataset* released by Gail Gong from Carnegie-Mellon University and the *Diabetic Retinopathy Debrecen* dataset issued by Dr. Balint Antal and Dr. Andras Hajdu from the University of Debrecen. In the context of the hepatitis dataset, these algorithms try to predict whether an individual with hepatitis will live or die based on 19 given attributes of the individual. Similarly, in the context of the diabetic retinopathy debrecen dataset, these techniques try to predict whether "an image contains signs of diabetic retinopathy or not"[Antal and Hajdu, 2014].

The KNN algorithm simply stores all the training examples it receives during its training, and then tries to find the training point that is closest to the input we want a prediction for. On the other hand, the Decision Tree algorithm tries to find conditions that efficiently split the training data. This creates a tree with two branches for every condition. Once a certain depth is reached, the prominent data class in the leaf branch will tell us which class this leaf associates with. These same conditions are then applied to the new input, and the branch it ends on will give us our prediction.

By testing different values of neighbours to consider in the K-Nearest Neighbour model, we evaluated the effects that this value has on the training data accuracy and test data accuracy. As well, by testing different maximum tree depths, we evaluated the effects that this value has on the performance of the Decision Tree model. Furthermore, we tested the effects that different distance/cost functions have on both models.

Overall, we found that the decision tree algorithm seems to perform slightly better with large datasets with many features that do not necessarily have a high correlation with the class, such as the Diabetic dataset. For smaller datasets with better feature correlation such as the hepatitis dataset, both algorithms had a very high precision.

2 Datasets

As stated above, the two datasets processed during this project are the hepatitis dataset and the diabetic retinopathy debrecen dataset. Following the information provided on the UC Irvine Machine Learning Repository, the hepatitis dataset contains 155 entries with each entry containing 19 attributes as well as a class attribute corresponding to whether the individual lives or dies. This dataset contains some missing values in which the entries associated with these missing values were eliminated before splitting the dataset into training and testing sets. On the other hand, the UC Irvine Machine Learning Repository states that the diabetic retinopathy debrecen dataset contains 1151 entries with each entry containing 19 attributes as well as a class label attribute which corresponds to whether there are signs of diabetic retinopathy or not. However, this dataset does not contain any missing values, therefore no adjustments were needed before splitting it into training and testing sets.

Ethical concerns can arise while working with health datasets. One ethical concerns that emerges is patient consent. In most cases, patients won't understand how these algorithms work which leads to them not being completely informed. Therefore, even though the patients consent to the use of their personal data, it's not ethical due to the fact they are not completely informed. Another ethical concern that arises is the correctness of the algorithms. If an algorithm is implemented incorrectly, it can be dangerous for the patients.

3 Results

3.1 Hepatitis Dataset

We started off by creating a heatmap of cross correlations in our hepatitis dataset to select appropriate features. The ideal two features would have high correlation with the target column, but a low correlation with each other. Figure 1 illustrates the heatmap associated with the cross correlations of the attributes in the hepatitis dataset.

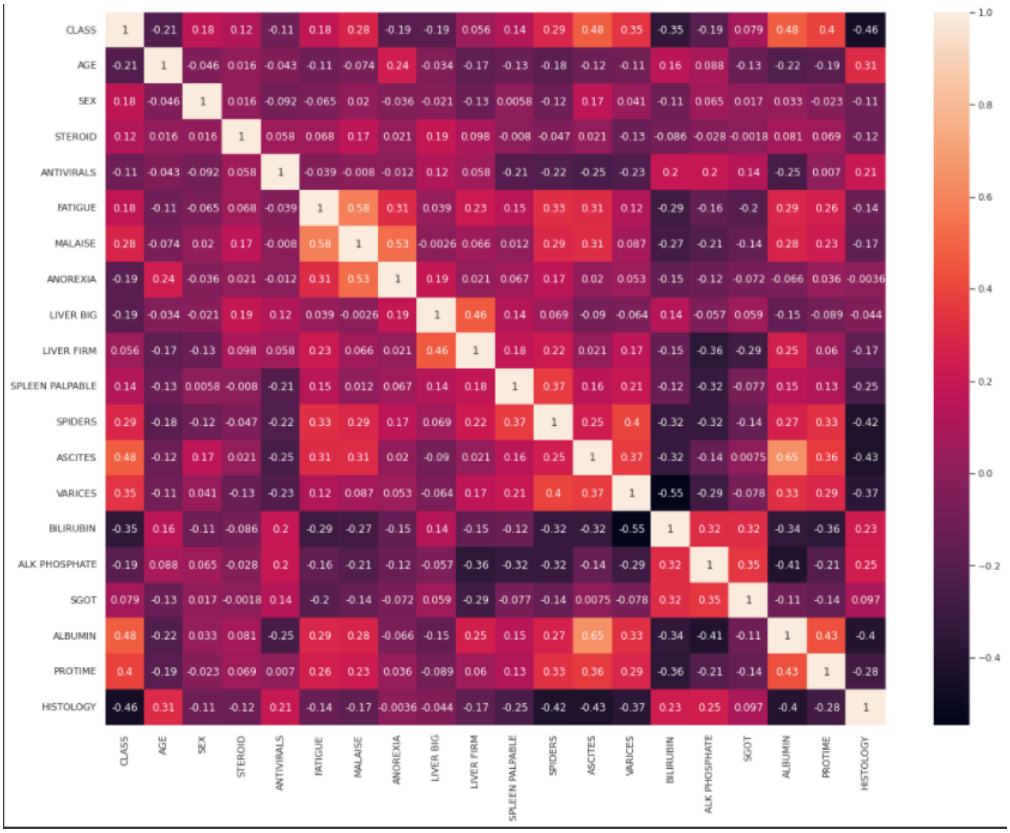
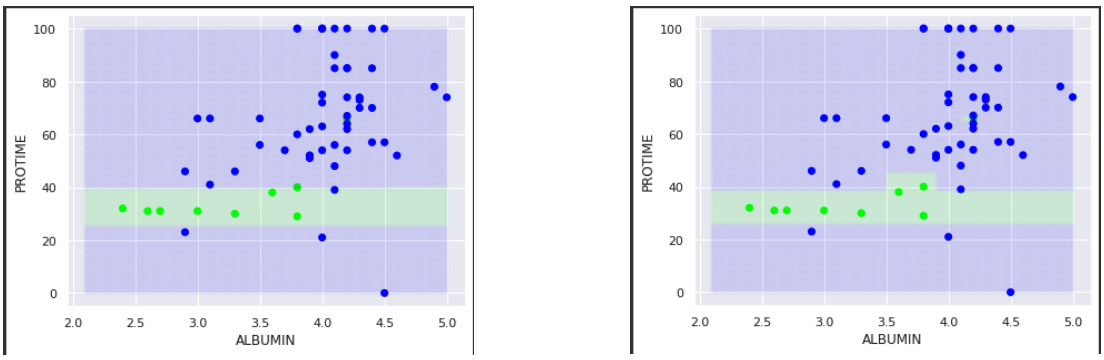


Figure 1: Hepatitis Dataset Heatmap of Cross Correlations

For our first test, we decided to use the attributes *Protime* and *Albumin*. As seen in figure 1, *Protime* has a correlation of 0.40 with the target attribute *Class* and *Albumin* has a correlation of 0.48 with the same target attribute. Between themselves, they have a correlation of 0.43.

After running our K-Nearest Neighbour algorithm with these two attributes and a K value of 3, the accuracy was found to be 75.0. In addition, after running our Decision Tree algorithm on the previously mentioned attributes with a maximum depth value of 20, the accuracy was found to be 75.0 as well. The following figure illustrates the decision boundaries of both models.



(a) K-Nearest Neighbour Model

(b) Decision Tree Model

Figure 2: Decision Boundaries of KNN model and Decision Tree model with *Protime* and *Albumin*

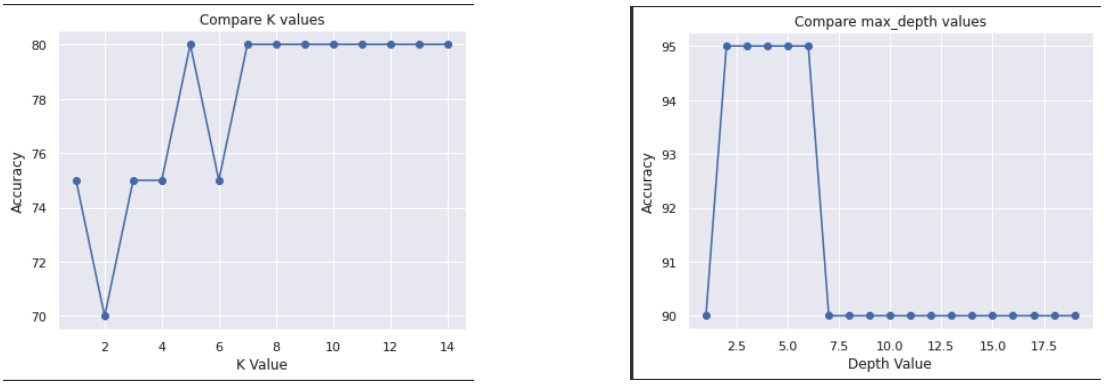
While comparing the decision boundary plots of our K-Nearest Neighbour model and our Decision Tree model from figure 2, we noticed that they both cover a similar amount of area in the graph. With this observation, we could conclude that the two algorithms give a similar accuracy when the decision boundaries are almost identical.

Following this test, we decided to test out the attributes *Albumin* and *Alk Phosphate* due to their different correlations with the target attribute and between themselves. As seen in figure 1, *Albumin* and *Alk Phosphate* have correlations of 0.48 and -0.19 with the target attribute *Class* respectively. As well, they have a correlation of -0.41 between themselves.

After running our K-Nearest Neighbour algorithm with a K value of 3 and Decision Tree algorithm with a maximum depth value of 20 on the *Albumin* and *Alk Phosphate* attributes, the accuracies were found to be 85.0 for both models.

From these two tests, we are able to conclude that both models perform very similarly on the hepatitis dataset as they both resulted in the same accuracies for both test.

We then decided to test out how various K values affected the accuracy of our K-Nearest Neighbour model as well as how various maximum depth values affected the accuracy of our Decision Tree model. The K values were tested in a range from 1 to 14 and the maximum depth values were tested in a range from 1 to 19. In addition, we picked the Albumin and Alk Phosphate pair for this test, as they seem to result in high accuracy. Figure 3 demonstrates the plots of the different accuracies.



(a) Accuracy of K-Nearest Neighbour model with K Values from 1 to 14

(b) Accuracy of Decision Tree Model with Maximum Depth Values from 1 to 19

Figure 3: Comparison of various K values and maximum depth values for hepatitis dataset

In the figure 3, we can notice the accuracies of the K-Nearest Neighbour model only vary between 70, 75 and 80. Also, smaller values of K lead to smaller accuracies compared to the other values and accuracies associated

with values of K greater than 7 are constant at 80. On the other hand, the accuracies of the Decision Tree model only vary between 90 and 95. Maximum depth values from 2 to 6 all lead to the greater accuracy and the others lead to the smaller accuracy. While the two graphs in figure 3 might not seem very intuitive (one would think that accuracies would drop back down when values of K are too high), it is important to note that the dataset has relatively few entries to begin with, and the majority of entries survive. Thus, adding more neighbours to consider should make it increasingly likely to get a survival prediction, but that is already the case for most entries in our dataset.

Next, we tested different cost and distance functions for both of our models. We used the attributes *Albumin* and *Alk Phosphate* as they are high in accuracy, a K and and depth values of 6 (these seemed to give us our best results in our previous experiment), and averaged our results over 10 runs.

For distance functions, we obtained the exact same results of 83.0. This was to be expected, with the large discrepancy between our axis scales. Therefore, the points on the smaller axis are much closer to each other, and both distance functions give similar results since they are mostly calculating straight line distances without diagonals.

For cost functions, we found that using Misclassification cost gave us a better accuracy than Entropy cost and Gini cost. Since the classification cost is based on the prediction of the most frequent label for each region and in this test there are only two regions that are reasonably split, we could conclude that it is the reason why Missclassification cost gave us the better accuracy.

3.2 Diabetic Retinopathy Dataset

For testing on the diabetic retinopathy dataset, we created a cross correlation heatmap as seen in section 3.1 for our feature selection. For our first test on the diabetic retinopathy dataset, we chose the attribute *Ma Detection 1* which has a correlation of 0.29 with the target class *Class Label* as well as the attribute *Ma Detection 6* which has a correlation 0.13 with the same target class. Between themselves, they have a correlation of 0.86.

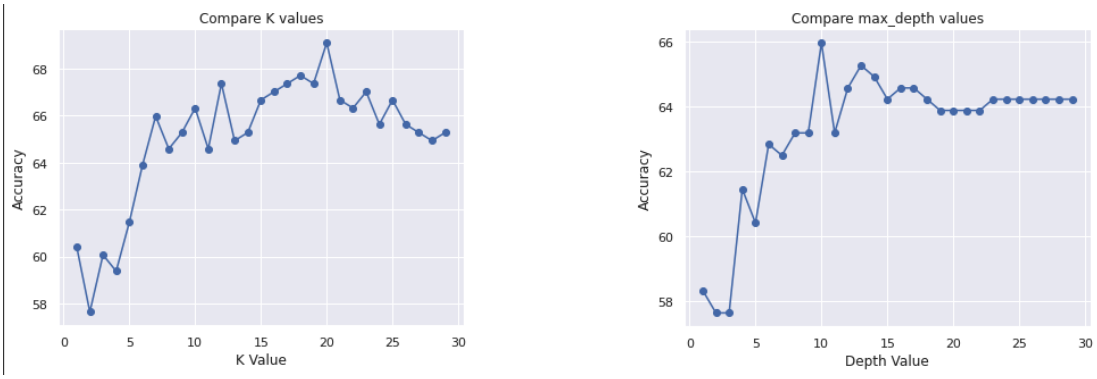
With the *Ma Detection 1* and *Ma Detection 6* attributes, our K-Nearest Neighbour model with a K value of 3 reported an accuracy of 58.68 and our Decision Tree model with a maximum depth value of 20 recorded an accuracy of 59.37.

For our second test on this dataset, we chose the attributes *Diameter* and *Ma Detection 4* which have correlations of -0.031 and 0.20 respectively with the target class *Class Label* respectively. In addition, between themselves, they have a correlation of 0.017. We chose they attributes since they have a relatively similar correlation with the target class but have a very different correlation between themselves.

With the *Ma Detection 2* and *Exudate Detection 3* attributes, our K-Nearest Neighbour model with a K value of 3 recorded an accuracy of 56.60 and our Decision Tree model with a maximum depth value of 20 reported an accuracy of 57.64.

From these tests, we can claim that our Decision Tree model results in a slightly higher accuracy than the K-Nearest Neighbour model while working on the diabetic retinopathy data set.

In addition, we tested how various K values and maximum depth values affected our models using the Ma Detection 1 and *Ma Detection 6* attributes since they seem to result in high accuracy. The K values utilized in this test ranged from 1 to 29 and the maximum depth value ranged from 1 to 29 as well. Figure 4 demonstrates the plot of the different accuracies.



(a) Accuracy of K-Nearest Neighbour model with K Values from 1 to 29

(b) Accuracy of Decision Tree Model with Maximum Depth Values from 1 to 29

Figure 4: Comparison of various K values and maximum depth values for diabetic retinopathy dataset

In figure 4, we can notice that values of K above 5 provide similar accuracies with the peak value being at K equal to 20. In addition, values of K below 5 provide fairly poor accuracy compared to the other values. On the other hand, we can observe that maximum depth values below 6 provide fairly poor accuracy compared to the other values. Also, maximum depth values above 15 result in accuracies that vary from each other minimally. The peak accuracy is at a maximum depth value equal to 10.

Figure 5 demonstrates the average results after 10 runs for the distance functions *Euclidean* and *Manhattan* as well as the cost functions *Misclassification Cost*, *Entropy Cost* and *Gini Cost*.

```
After 10 runs, we have the following average results for distance functions:
Euclidean: Accuracy = 63.854166666666664
Manhattan: Accuracy = 63.75

After 10 runs, we have the following average results for cost functions:
Misclassification cost: Accuracy = 62.01388888888889
Entropy cost: Accuracy = 60.972222222222214
Gini cost: Accuracy = 61.666666666666664
```

Figure 5: Average results for distance and cost functions after 10 runs

4 Discussion and Conclusion

Overall, the K-Nearest Neighbour and Decision Tree models performed very similarly on the given datasets. The Decision Tree model provided a slightly better accuracy while working with the smaller data set as we noticed when comparing various accuracies of maximum depth values. On the other hand, the K-Nearest Neighbour model provided a slightly better accuracy while working with the larger data set as we found while comparing the accuracies of various K values. As we were not able to test out all the attributes in the given data set due to the large number of attributes, we could further expand our research on the models by testing other combinations of attributes and possibly investigate a model that is capable of incorporating all of the features at once.

5 Statement of Contributions

We collectively contributed to the entire project.

References

Bálint Antal and András Hajdu. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-Based Systems*, 60:20–27, Apr 2014. ISSN 0950-7051. doi: 10.1016/j.knosys.2013.12.023. URL `http://dx.doi.org/10.1016/j.knosys.2013.12.023`.