MiniProject 1: Getting Started with Machine Learning

Ricardo Alvarez, 260926136, ricardo.alvarez@mail.mcgill.ca
Tanja Barath, 260852872, tanja.barath@mail.mcgill.ca
Amily Li, 260888273, yuxiao.li@mail.mcgill.ca

***Abstract*—In this project, we investigated the performance of k-nearest neighbours and decision tree models on two real-world data sets. The first investigated the survivability of hepatitis patients and the second was looking to predict whether an image contains signs of diabetic retinopathy or not. We found that the Decision Tree approach achieved better accuracy than K-Nearest Neighbour.**

I. Introduction

Even with recent advances in biotechnology, many diseases' origins and underlying mechanisms remain widely unclear. It has become clear that in order to truly manipulate morbidity, there is an essential need for a prediction system to brings awareness about illnesses. Machine learning provides prestigious support in predicting any kind of event, which takes training from gathered data and is the perfect tool that could be used to deal with this problem.

There has been many applications of machine learning models on medical use cases. A particularly interesting use case by Geamsakul W. et al. [1] involves classifying types of hepatitis using decision trees based on graph-based induction. This involves extracting patterns by pairwise induction and achieves 76.79% accuracy. A use case that is related to our diabetes dataset is Ali, Ameer, et al's [2] research on diagnosing diabetes by training a KNN model based on different implementations. This is insightful as it compares different KNN implementations such as weighted (99.8% accuracy) and cosine (85.6% accuracy).

In this project, we analyze the accuracy of machine learning algorithms, K-Nearest Neighbours and Decision Tree, for predicting important features on two distinct health data sets. In particular, the hepatitis data set contained data about the survival of hepatitis patients, personal attributes such as age and sex and medical attributes such as whether the patient was on steroids or were anorexic. It contained 155 instances. The second dataset contained information about apparent anatomical measurements such as exudates and the diameter of the optic disc, used to predict whether an image contains signs of diabetic retinopathy or not. The data set contained 1151 instances.

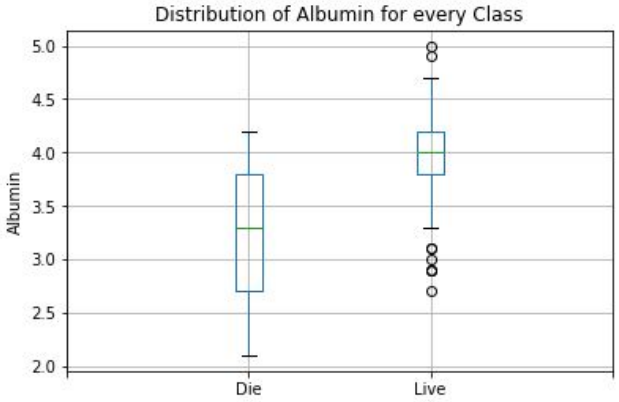
We found that the decision tree model performed better than the K-nearest Neighbour model for both given data sets. Experiments were conducted to find the optimal hyperparameters. The decision tree hyperparameters tuning results are discussed further below but the most important finding was that decision trees tend to overfit the data when presented with small datasets and low class correlation features, which was our case. The analysis on the KNN model found similar behaviour that had a bigger affect on its accuracy.

II. Datasets

A. Hepatitis Dataset

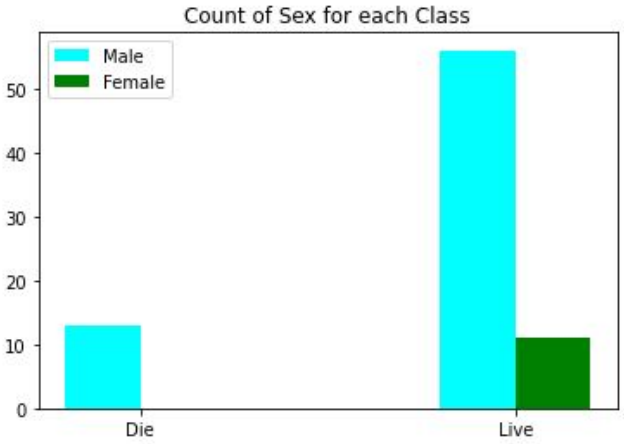
The first dataset that we worked with was a dataset about survival of hepatitis patients. The dataset contained personal attributes such as age and sex, medical attributes such as whether the patient was fatigued or had a firm liver and medical measurement attributes such as bilirubin and albumin. The dataset contained 155 instances but only 80 were used for training after dropping the instances with null values and converting everything to numeric values. Moreover, the binary values (1 and 2) were changed to 0 and 1 to facilitate the analysis and visualisation of the data.

The first observation we made was that there where many more patients who survived than did not and patients of all ages survived whereas non-surviving patients mostly ranged from 40 to 60 years old. A notable characteristic of the dataset is that the feature with the most class correlation, albumin, had a significant distribution difference between surviving and non-surviving patients. Non-surviving had an albumin distribution with a mean of only around 3.3 compared to the distribution of surviving patients that concentrated around a 4.0 mean. Their interquartile ranges also had no overlap.

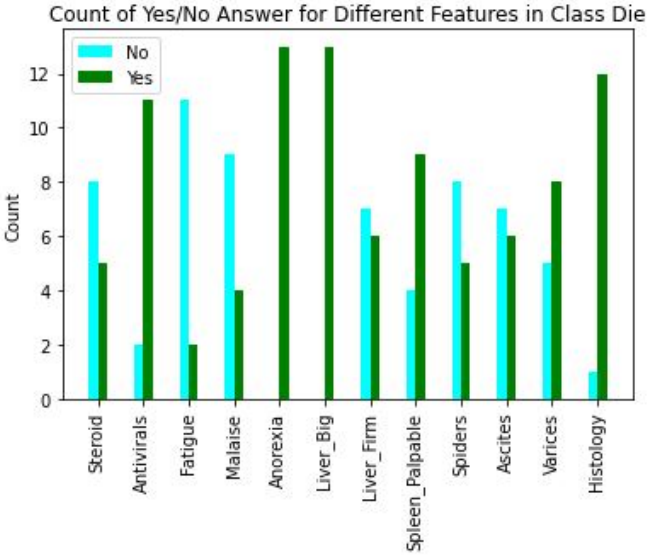


A potential limitation of this dataset is that all non-surviving patients were male. This might be because of

the small size of the dataset and may cause the model to not generalize well.



All medical binary attributes were present in both surviving and non-surviving patients except for having anorexia and a big liver where all patients who died presented those symptoms.



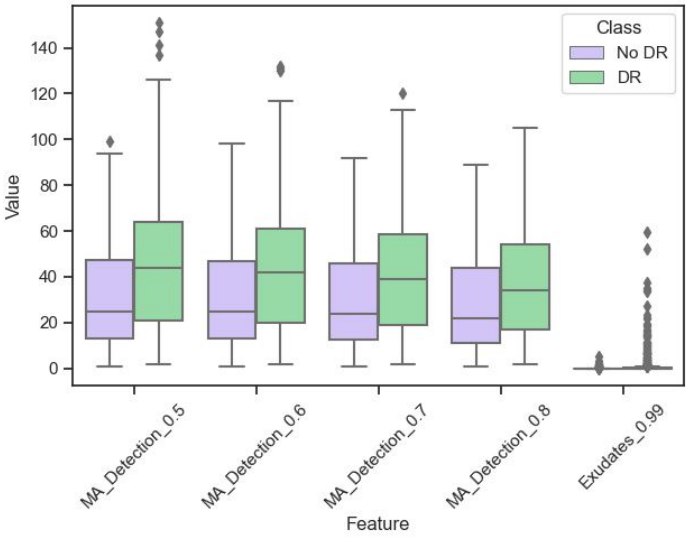
Most attributes had relatively high class correlation and the top 5 attributes (Ascites, Albumin, Protime, Histology and Bilirubin) had correlations that ranged in the absolute from 0.35 to 0.48.

There were no apparent ethical issues with this dataset.

B. Diabetic Retinopathy Dataset

The second dataset presented data about whether or not an image had signs of diabetic retinopathy. The features contained information about image characteristics such as quality and apparent anatomical measurements such as exudates and the diameter of the optic disc. The dataset contained 1151 instances and 1147 were used for training after dropping 4 rows with low quality of image (Quality = 0). No more cleaning was required.

This dataset had a significantly higher number of instances than the first but the attributes had much lower class correlations. The top 5 attributes (Exudates_0.99, MA_Detection_0.8, MA_Detection_0.7, MA_Detection_0.6, MA_Detection_0.5) had correlations that ranged in the absolute from 0.18 to 0.29. These distribution of each of these features grouped by class is shown below. Exudates 0.99 is scaled, each value multiplied by 10 for a better view of the spread.



People with signs of diabetic retinopathy have higher average values for Microaneurysm Detection. The interquartile ranges are still significantly overlapping meaning there is a relationship between Class and MA Detection but this correlation is low. There is a concentration of 0 values for Exudates at confidence level 0.99. However, for the data points that differ, the distribution for people with signs of DR is clearly more spread out.

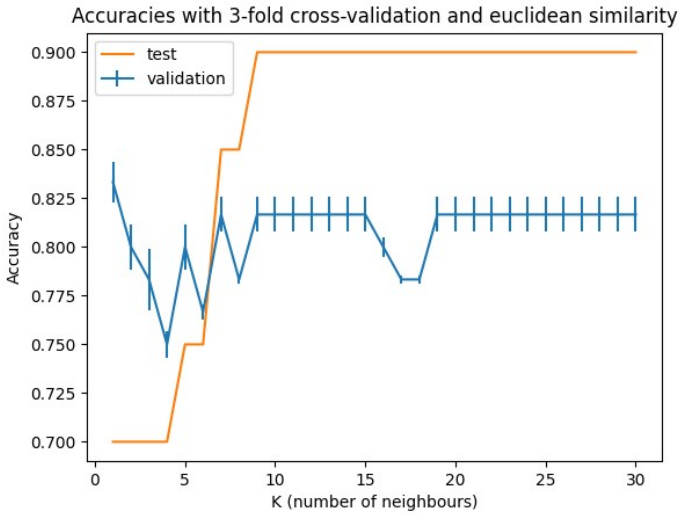
There were no apparent ethical issues with this dataset.

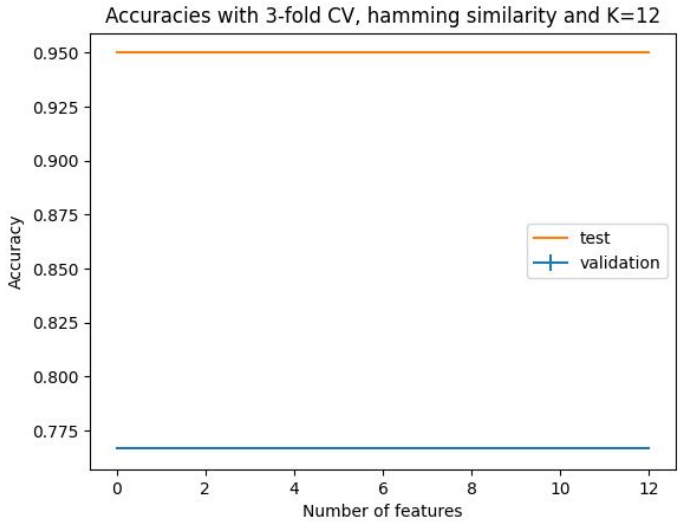
III. Results

A. KNN

The K-nearest neighbour classifier was implemented using an L-fold cross-validation. The average validation error is reported to give a better sense of the models performance.

1) Hepatitis Dataset: The hepatitis dataset is left with a total of 80 data points to use after deleting malformed data. 20 of these were randomly put aside to serve as the test set. The other 60 data points made up the training set. Hyper-parameter tuning was carried out on the two most correlated real-valued features (Albumin and Protime) with different distance metrics and different splits of the training set for cross-validation. The highest accuracy attained was the same for all combinations of splits and distance metrics. Neither attribute significantly changes the accuracy of the model. Still, the Cosine similarity distribution was slightly different from the rest. The training accuracy stabilized at its peak 81% from K = 5 and remained there. The other distance metrics (Euclidean, Manhattan and Chebyshev) output the same distribution with fluctuating accuracies throughout.



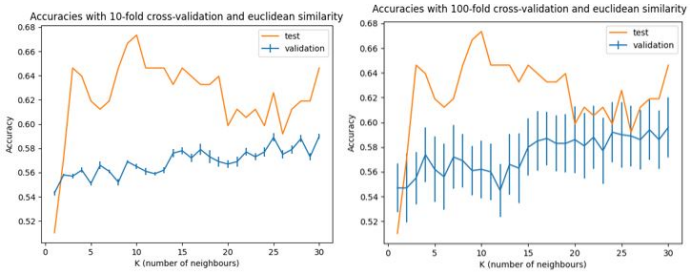


Without considering the test accuracy, the choice of K from the Cosine similarity could have been anything larger than 6. But the euclidean distance shows some undesired behaviour happening for K in between 14 and 20. Thus, K = 12 is a strong candidate. Testing performance with K = 12 on the test set results in a 90% accuracy for every distance metric used (euclidean, manhattan, chebyshev and cosine). At first glance, this seems very good but having such a limited sized dataset warrants more investigation. The test set used only had 2 data points that were labelled DIE but every single data point were predicted to be of class LIVE. This makes the number of correctly predicted data points 18 out of 20 equal to an accuracy of 90%. When the test set was changed to have 5 rows with class DIE and using K = 12, the accuracy was 75% because the 5 points belonging to class DIE were predicted to be of class LIVE.

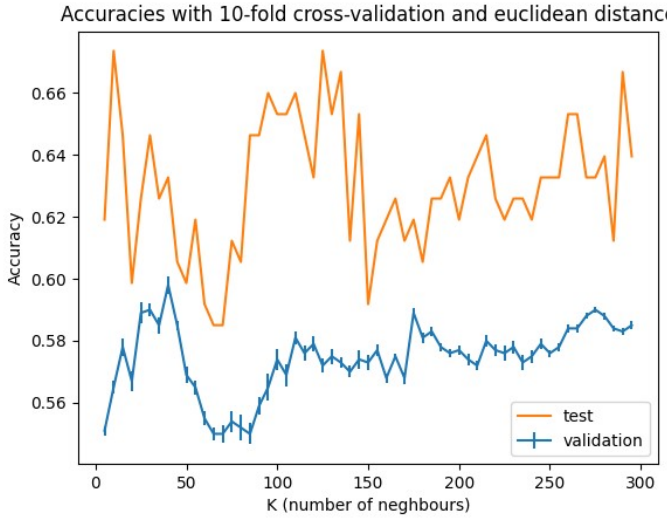
With increasing K, the people that lived ten to over-power the set of neighbours because they make up more than 80% of the data. In this case, reporting the models' recall, the amount of dead people that the model accurately predicted, is important. For both test sets, recall is 0%. This model would generalize poorly to truly unseen data as it is underfitting.

Another interesting observation is that using the discrete features of the dataset and the hamming distance decreases the training accuracy to 76.67% but increases that of the test set to 95%. The recall is 50% making the choice of discrete features more favorable. Also, adding less correlated features doesn't decrease the accuracy of neither the validation nor test sets. The plot below summarises the accuracy of the K = 12 model on an increasing number of binary features going from the most correlated to the least in the following order: (Ascites, Histology, Varices, Spiders, Malaise, Liver Big, Fatigue, Anorexia, Sex, Spleen Palpable, Antivirals, Steroid, Liver Firm).

2) Diabetic Retinopathy Dataset: This dataset was split into 1000 training data points and 147 test points. The 5 most correlated features (Exudates_0.99, MA_Detection_0.8, MA_Detection_0.7, MA_Detection_0.6, MA_Detection_0.5) are used for hyper-parameter tuning. To begin with, the data was tested on a few different splits and no change in the accuracies was observed. Rather, a larger number of splits gave rise to much higher variances in the validation accuracies. This difference is apparent in the 10-fold and 100-fold cross-validation graphs below.



Thus, average accuracy for the training set is reported using a 10-fold cross-validation where the variances are much smaller. Since the features chosen are real-valued, Euclidean, Chebyshev and Manhattan distances as well as Cosine similarity were tested. The first 3 output the same accuracy distribution that reached as high as 60% while the latter didn't go higher than 53%. Cosine similarity also makes it harder to choose an appropriate K value since it is mostly constant for long intervals. The distribution of accuracies given by Euclidean distance is preferred for hyper-parameter tuning.



From the graph above, the training accuracies shows that the best K is in between 20 and 50. Taking a closer at this interval, K = 40 is chosen. But notice that the highest test accuracy is actually reached at K = 125. For 10-fold cross-validation at K = 40 using the features given, the training accuracy is 59.8% and the test accuracy is 63.26%.

This model can generalize much better to truly unseen data because it had a larger dataset to learn from. However, the main drawback is its low accuracy, the result of all the features being poorly correlated with the Class.

B. Decision Tree

1) Hepatitis Dataset: The decision tree was trained with the top three most class correlated features (Albumin, Ascites and Histology). This design decision was taken after testing the impact of feature inclusion on validation set accuracy. The dataset was split in 50% training, 25% validation, 25% test sets. Three-fold cross-validation was performed to determine the best parameters to give to the decision tree. These parameters included the maximum depth of the tree, the cost function to use and the minimum instances that should be included in a tree leaf. Experiments were first undertaken by changing one parameter at a time. The default values of the tree were a maximum depth of three, a misclassification cost function and a minimum leaf instances of one.

The experiments demonstrated that the accuracy increased as we increased the depth of the tree and plateaued at 93%.

Maximum tree depth	Mean of accuracies
1	87%
3	88%
5	93%
10	93%
15	93%
20	93%
30	93%

This is an expected result as decision trees tend to overfit the data as their depth grows. The experiments on the cost function parameter showed that both the misclassification and Gini index cost function rendered a 88% validation accuracy and the entropy cost function a 90% validation accuracy. The experiments on the minimum leaf instances all resulted on the same validation accuracy of 88% across all values (1, 10, 20, 50).

Further, experiments were made by performing cross-validation of all combinations of these parameters. The best parameters (misclassification cost, maximum depth of 5 and minimum leaf instances of 1) resulted in a validation accuracy of 93%, but in a testing accuracy of 70%. As this indicated overfitting, a gini index cost function, maximum depth of 5 and minimum leaf instances of 10 were chosen. This resulted in a more appropriate validation accuracy of 92%, but in a testing accuracy of 90%. This design decision was taken because the gini cost function and a minimum leaf instance of 10 consistently provided good accuracy during validation trials and performance plateaued at a maximum depth of 5. It is important to note that the limited dataset size might cause our model to not generalize well.

2) Diabetic Retinopathy Dataset: The decision tree was trained with the top five most class correlated features (Exudates_0.99, MA_Detection_0.8, MA_Detection_0.7, MA_Detection_0.6, MA_Detection_0.5). This design decision was taken after testing the impact of feature inclusion on validation set accuracy. The dataset was split in 58% training, 29% validation, 13% test sets. The same three-fold cross-validation experiments as with the hepatitis dataset were performed to determine the best parameters.

In this case, the experiments demonstrated that the accuracy kept increasing as we increased the depth of the tree and without plateauing as early as with the hepatitis case. This might be related to the size of the dataset as the tree requires more tests and thus more depth to classify the instances individually and overfit.

Maximum tree depth	Mean of accuracies
1	60%
3	65%
5	70%
10	80%
15	90%
20	95%
30	97%

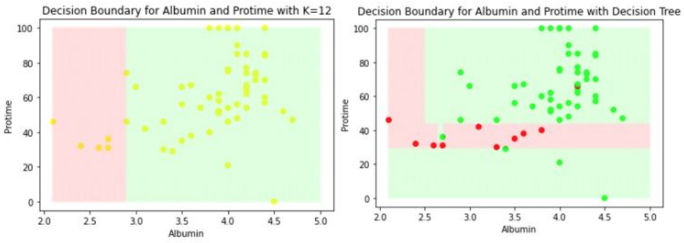
The experiments on the cost function parameter showed that all rendered a 65% validation accuracy. The experiments on the minimum leaf instances all resulted on the same validation accuracy of 65% across all values (1, 10, 20, 50).

Further, experiments were made by performing cross-validation of all combinations of these parameters. The best parameters (misclassification cost, maximum depth of 30 and minimum leaf instances of 1) resulted in a validation accuracy of 97%, but in a testing accuracy of 63.9%. As this indicated overfitting, a misclassification cost function, maximum depth of 10 and minimum leaf instances of 1 were chosen. This resulted in a more appropriate validation accuracy of 68%, but in a testing accuracy of 69.4%. This design decision was taken because these parameters consistently provided good accuracy during validation trials.

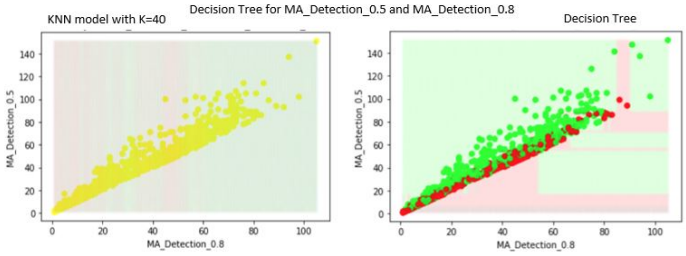
C. Final Results and Decision Boundaries

Summary of results

Accuracy	KNN	Decision Tree
Hepatitis validation	76.67%	92%
Hepatitis test	95%	90%
Diabetic Retinopathy validation	59.8%	68%
Diabetic Retinopathy testing	63.26%	69.4%



The decision boundary for K = 12 and features (Albumin and Protime) shows only one class predicted. This matches the analysis from the KNN results. A section of the graph is of class DIE but no points are chosen because there are so few neighbors in that part. The boundary given by the Decision Tree model is cleaner and does classify some points as people that have died whether this is correct or not.



The decision boundaries from the Decision Tree model show a linear relationship between micro aneurysm with confidence level 0.5 and 0.8. It also indicates that points located towards the bottom of this relationship are almost all classified into one Class while the top are classified in the other. The KNN model boundaries are much more confusing making it hard to find any patterns.

Taking the final accuracy results and decision boundary plots, it is clear that Decision Tree is better at predicting both datasets.

IV. Discussion and Conclusion

The prediction of the unbalanced parameter that reflects the survivability of hepatitis patients is a crucial issue when talking about the correction of the model. One important takeaway is that the results of the algorithms demonstrate that KNN is prone to overfitting with small data sizes but very computationally expensive with large ones. While KNN struggles to assign any data points to its correct value, Decision Tree manages to correctly predict a large amount of them. Therefore, it is suggested by current results to use Decision Tree in the prediction of imbalanced parameters and small datasets.

The experiments also show that K-Nearest Neighbor is worse than the Decision Tree in terms of accuracy when working with large data set. Since KNN uses the number of nearest neighbor “K” as one of the parameters in classifying an object, the larger the data set, the more memory it will require to store the training data and subsequently, larger distance calculations would be performed. This makes the classifier slow and inefficient. However, the accuracy results were not very far apart meaning KNN learned better but Decision Tree was still a better option.

Statement of Contributions

Ricardo Alvarez worked on implementing the decision tree and its hyper-parameter tuning experiments and cross-validation. He wrote the introduction, dataset and decision tree results sections of the report.

Tanja Barath read in the data and cleaned it. She implemented the KNN algorithm for these two datasets and wrote the cross-validation for it with the help of Amily. She wrote the KNN results and created the plots for it as well as the plots in the dataset section.

Amily Li implemented the cross-validation and the decision boundary. She assisted on writing the abstract, introduction, discussion and conclusion.

References

[1] Geamsakul W. et al. (2007) Analysis of Hepatitis Dataset by Decision Tree Based on Graph-Based Induction. In: Sakurai A., Hasida K., Nitta K. (eds) New Frontiers in Artificial Intelligence. JSAI 2003, JSAI 2004. Lecture Notes in Computer Science, vol 3609. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-71009-7_2

[2] Ali, Ameer, et al. “Diabetes Diagnosis Based on Knn.” Iium Engineering Journal, vol. 21, no. 1, 2020, pp. 175–181., doi:10.31436/iiumej.v21i1.1206.