COMP 551: Mini Project 1

1 Abstract

The goal of this project was to apply machine learning models to two health-related benchmark datasets and compare their respective accuracies. We tested the K-Nearest Neighbors and decision trees models on both of these datasets and performed hyperparameter tuning to obtain the optimal models with maximal accuracy. After applying both models to each dataset, we found that the decision tree model performed better than KNN on both datasets. In particular, it performed significantly better on a dataset for which KNN offered no prediction accuracy. Many different performance enchancing methods were tried.

2 Introduction

The main objective of this project is to predict various health-related outcomes based on the symptoms of several patients from two different labelled datasets; the Hepatitis dataset and the Diabetic Retinopathy Debrecen dataset. The former contains data relating various hepatitis symptoms to survival of patients, while the latter contains data relating various diabetic retinopathy symptoms to the actual presence of DR disease.

Our most important finding was that Decision trees performed slightly better on the hepatitis dataset (+5% accuracy) but performed much better on the diabetes dataset (+15-20% accuracy). Our KNN model had good results on the hepatitis datasat, but performed very poorly on the DR dataset, averaging 50% accuracy for all parameters.

Several classification models have already been applied by other researchers on both datasets with similar accuracy as ours. For the Hepatitis dataset, (1) achieved 86% accuracy using Naïve Bayes and KNN while on the other hand (2) achieved a best accuracy of 84% out of multiple different decision trees variations. Turning to the diabetic retinopathy dataset, (3) classified with multiple methods including KNN and decision trees achieving 62% and 70% respectively.

3 Datasets

In this section, we describe the respective attributes and construction of each dataset, as well as the results of preliminary exploratory analysis performed on each of them.

3.1 Hepatitis dataset

The Hepatitis dataset includes various patient data (Age, sex, presence of antivirals, anorexia, etc.) as well as a histology label, indicating the presence of hepatitis. and a class label, indicating patient survival. For this dataset as for the next one, the features contained various ranges of numerical values as well as categorical ones. We converted non-numerical feature values to numerical values. Two of the features ('ALK PHOSPHATE' and 'PROTIME') contained an abnormally high amount of missing values. Rather than drop all the corresponding patient data (representing over a third of our entire dataset) we chose to drop the features entirely as the dataset is already small and we expect the features to be less valuable than such a high number of our instances. For other missing values, we dropped the rows.

After cleaning out the hepatitis dataset, we obtained 24 deceased and 105 surviving patients out of a total of 129 patients. Note that these are only the people who lived or died regardless of their histology (i.e. whether or not they had hepatitis). To better understand this dataset, we looked at how the various features correlated with histology and the survival label. Out of the patients who were negative, 4 died anyway, while 68 survived. Out of the patients who had hepatitis, 20 died while only 37 survived.

We also found that age does not significantly impact the likelihood of hepatitis or survival. The age range of our patients is 7 to 78 years old, with a mean age of 40. For people whose histology was negative, the mean age is 39, while for those for who it was positive, it was 43. This indicates that age may not be a great predictor of hepatitis. Similarly, it appears not to be a great predictor of death, because the average age of surviving and deceased patients is respectively 40 and 45.

An ethical concern underlying the use of this dataset was found in the under-representation of women among all patients. The unbalanced nature of this dataset might lead our models to produce false connections between sex and health outcomes, because there are simply not enough female patients to derive a consequent conclusion on the correlation between sex and hepatitis-linked death. Another potential concern is privacy: healthcare data is inherently sensitive, and personal patient data can be notoriously difficult to anonymize for research purposes.

3.2 Diabetic retinopathy dataset

The features contained in the Diabetic Retinopathy Debrecen dataset are the numerical results of various health screenings, aside from f19 which is the class label. We found that the result of pre-screening (f1) does not strongly correlate with the class label (f19). Generally, this dataset has low correlation numbers between features and the class label, and proved challenging to predict to get much value out of overall.

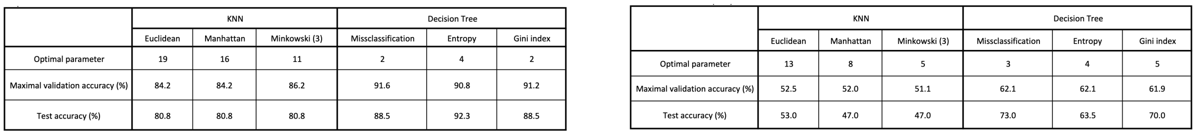
3.3 Data Processing

It is important to note that the size of the dataset being very small, dramatic fluctuations in accuracy will occur based on the train/test split. These papers deal with this in various ways, such as averaging out the accuracy over many runs of the model, or setting a random seed for the data splitting. We chose not to set a random seed for this assignment, since selecting the right seed could dramatically improve our model performance, and we want to display our results as honestly as possible. We will therefore present a randomly selected representative run of our models.

In order to give equal weight to all features for KNN (but not for decision trees), we normalized the features of both datasets so that every feature contains values with zero mean and unit variance. This was done to prevent out of scale features from disproportionately affecting the models.

4 Results

We now present the results of our various experiments performed on each dataset.



	KNN			Decision Tree		
	Euclidean	Manhattan	Minkowski (3)	Missclassification	Entropy	Gini index
Optimal parameter	19	16	11	2	4	2
Maximal validation accuracy (%)	84.2	84.2	86.2	91.6	90.8	91.2
Test accuracy (%)	80.8	80.8	80.8	88.5	92.3	88.5

(a) Hepatitis classification results

	KNN			Decision Tree		
	Euclidean	Manhattan	Minkowski (3)	Missclassification	Entropy	Gini index
Optimal parameter	13	8	5	3	4	5
Maximal validation accuracy (%)	52.5	52.0	51.1	62.1	62.1	61.9
Test accuracy (%)	53.0	47.0	47.0	73.0	63.5	70.0

(b) Diabetic retinopathy classification results

Figure 1: Model performances on both datasets

4.1 Classification results

In 1, we can observe that both our models performed much better on the hepatitis dataset than on the diabetes dataset. This is likely because the diabetic retinopathy dataset is bigger with less correlated features and more noise, and so is of much lower quality than the hepatitis dataset.

For both datasets, decision trees outperformed KNN, but for the diabetic retinopathy dataset the performance improvement is far more pronounced. This is expected, as Decision trees is less sensitive to noise and poor data quality than KNN.

To select hyperparameters for our models, we decided to perform 10-fold random split validations. In other words, a 20% portion of each dataset was reserved for testing the final models. For each fixed hyperparameter (number of neighbors$K$ or maximum tree depth D), the remaining data was split 10 times into random training-validation $(75\% - 25\%)$ partitions. For each such split, both KNN and decision tree models were applied and their validation accuracies were computed.

This process was repeated for three different distance functions for KNN: the Euclidean distance, the Manhattan (L_1) distance and the Minkowski distance (with $p = 3$). Similarly for decision trees, three different cost functions were tested; the misclassification cost, the entropy cost and the Gini index cost.

4.2 K-Nearest Neighbors

The resulting validation accuracies of KNN on the Hepatitis dataset and the Diabetic Retinopathy dataset are respectively shown in 3. For the hepatitis dataset, we observe that KNN performance tends to improve as we increase the value of K. Eventually, we reach a point of diminishing returns and model performance starts to decrease. No clear insight can be drawn from the diabetic retinopathy dataset, because its validation accuracy is stuck at around 50% for all distance functions. This indicates that the KNN model cannot predict any label from this dataset. This might be caused by the low quality of the data. Further along in the article we have attempted multiple methods to try and improve this accuracy.

4.3 Decision Trees

Similarly for decision trees, the validation accuracies on the Hepatitis dataset and the Diabetic Retinopathy dataset are shown in 4. For the hepatitis dataset, model performance reached a local maxima for a specific depth value, but does not follow any general trend as the depth increases. The decision tree model performs significantly better than the KNN model for the DR dataset, as it reaches around 62% validation accuracy for all cost functions.

4.4 Performance Summary

After having computed the 10-fold random split average validation accuracies of both models for each hyperparameter value and different distance functions, we then chose the optimal hyperparameters (number of neighbors K_{opt} for KNN and maximum tree depth D_{opt} for decision trees) as the parameters maximizing the average validation accuracy over all 10 splits. Having chosen these hyperparameters, we then applied the optimal models on the reserved test data and evaluated the test accuracy, as shown in Figure 1.

The exact optimal point in both algorithms depends heavily on our train/test split seed, as mentioned in the introduction, because the data set is very small and this leads to intense accuracy fluctuations. It is also dependent on the cost function used, although the change of cost function lead to only marginal differences in overall model performance for both models. The most pronounced of these differences is with entropy cost for decision trees, which performed significantly worse than the other two cost functions on the diabetes dataset (though it performed marginally better on the hepatitis dataset). However, while the accuracy differences may be marginal, we did observe more pronounced differences in hyperparameter value selected, particularly for KNN ($K = 19$ for Euclidean distance, $K = 11$ for Minkowski, see 1.

4.5 Decision boundaries

Note that for the Diabetic Retinopathy dataset, columns f2 to f15 represent the same measurement of MA detection under different confidence levels, and so were highly correlated to each other. The other features were categorical except f17, which has a low correlation of 0.03 with the class label. For this reason, we chose not to display the decision boundary for the DR dataset.

For the hepatitis dataset, we found the two continuous features with highest correlation with class label to be 'BILIRUBIN' (correlation -0.5069) and 'ALBUMIN' (correlation 0.4823). The hepatitis data classification results of KNN for the Euclidean distance function and optimal parameter $K_{opt} = 19$ over the two continuous features with highest correlation.

Finally, we can visualize the KNN decision boundary on the hepatitis data for both parameters $K = 1$ and $K = 5$ (see Figure 5) with the two continuous features with highest correlation with the class labels. As expected, the decision boundary is simplified as K increases. Note that we sliced off higher values of the Bilirubin feature for better visualization. Similarly, we can visualize the Decision tree decision boundary on the hepatitis data for the parameter $D = 2$, as shown below in Figure 2.

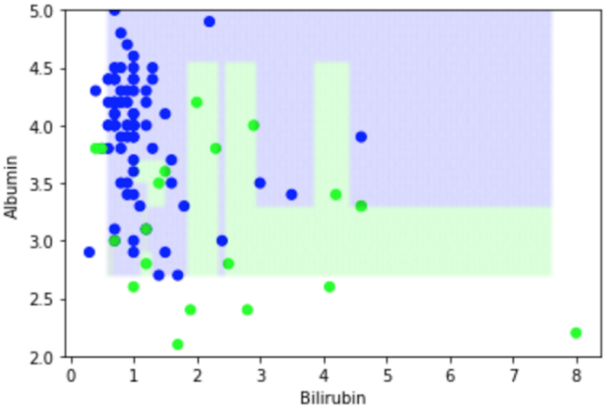


Figure 2: Decision tree decision boundary for $D = 2$

4.6 Feature selection

As a first experiment to see if it is possible to improve model performance by removing noise, we selected two features with highest correlation with the class label in the training set in both datasets. Accordingly, we chose the features 'ASCITES' (correlation 0.520441) and 'BILIRUBIN' (correlation 0.506906) for the hepatitis dataset, and the features 'f2' (correlation 0.2926) and 'f3' (correlation 0.266338) for the DR dataset. We then applied the KNN and decision tree models on a truncated dataset containing only these two features. However, the resulting accuracies of both models for the hepatitis dataset and the DR dataset offered no clear improvement over our previous experiments. Table omitted for space reasons.

4.7 Correlation-weighted KNN

As a second experiment, after normalizing feature values, we weighted all features by their correlation with the class label in the training set in order to give more importance to features with high correlation in the KNN model. Indeed, features with high correlation will have more impact on the distance function than those with low correlation, and so by weighting features by their correlation, high-correlation features are prioritized by the model. However, the resulting accuracies offered no clear improvement over our previous experiments. Table omitted for space reasons.

5 Discussion and Conclusion

We compared the performance of KNN and Decision trees with various cost functions and hyperparameter values on the hepatitis and diabetic retinopathy data provided. Overall, we found that the decision tree model performed slightly better on the hepatitis dataset (+5% accuracy) but performed much better on the diabetes dataset (+15-20% accuracy). This may be caused by the fact that the DR dataset is of lower quality and contains more noise, to which KNN models are more sensitive than decision trees. We found no large difference in performance for different cost functions, but we did find significant differences based on hyperparameter values.

We added 3 ideas of our own to the required experiments of this assignment. First, we use 10-fold random split validations to select hyperparameters. Then, we ran our models using only the 2 most correlated features, to see if this would allow KNN to bridge the gap with decision trees. Lastly, we ran the model weighting the features based on their level of correlation with the labels.

6 Statement of Contributions

The workload was evenly distributed among the 3 team members. Here are their respective contributions:

- **Dragos Secrieru**: Loaded datasets, cleaned datasets, implemented the decision trees model, ran simulations, created accuracy plots, wrote the Introduction sections.
- **Antoine Bonnet**: Implemented the dataset splitting method and KNN model, the L-fold cross validation method, ran simulations, created accuracy plots, decision boundaries, wrote the Dataset and Results sections.
- **Cyril Saidane**: Set up the notebook collaboration tools, loaded datasets, cleaned data, computed statistics and graphs for part 1, ran simulations, designed experiments and wrote the Results and Conclusion sections.

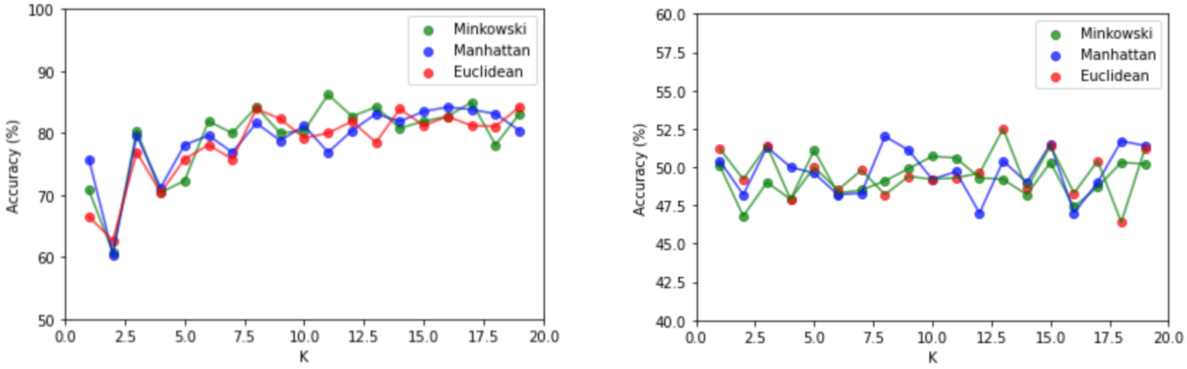
7 Appendix

References

[1] Ferdousy, E. Z., Islam, M. M., amp; Matin, M. A. (2013). Combination of naïve Bayes classifier and K-Nearest Neighbor (CNK) in the classification based predictive models. Computer and Information Science, 6(3). https://doi.org/10.5539/cis.v6n3p48

[2] Karthikeyan, T., amp; Thangaraju, P. (2013). Analysis of classification algorithms applied to hepatitis patients. International Journal of Computer Applications, 62(15), 25–30. https://doi.org/10.5120/10157-5032

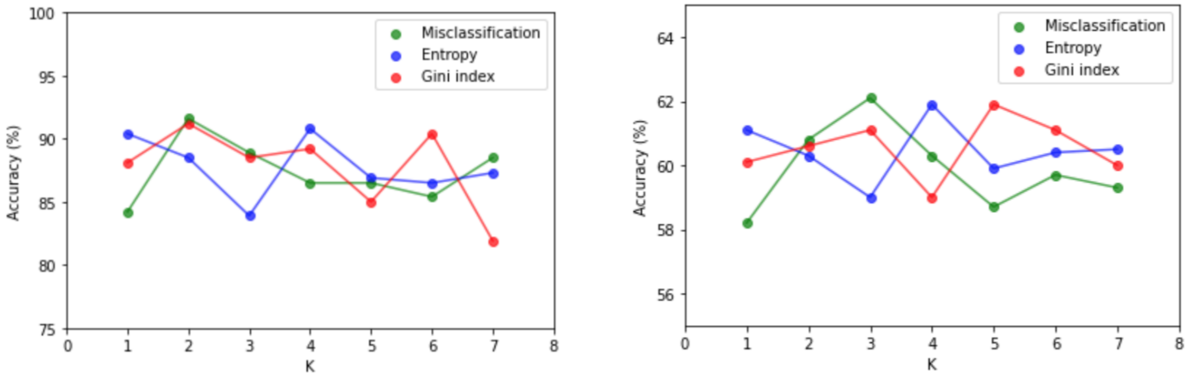
[3] Taveira-Gomes, T. (2016, June 25). Machine Learning on the Diabetic Retinopathy Debrecen Data Set Data Set. Retrieved February 8, 2022, from https://rstudio-pubs-static.s3.amazonaws.com/188757_8b0b10ee15a94850b9d3461496451618.html



(a). Hepatitis mortality prediction 10-fold random split validation accuracy using KNN with different distance functions

(b). Diabetic Retinopathy prediction 10-fold random split validation accuracy using KNN with different distance functions

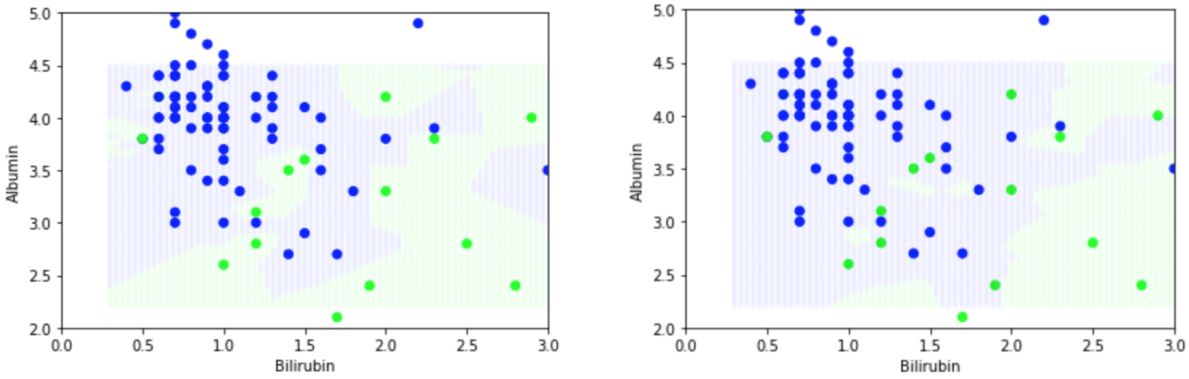
Figure 3: KNN validation accuracies



(a). Hepatitis mortality prediction 10-fold random split validation accuracy using Decision Tree with different cost functions

(b). Diabetic Retinopathy prediction 10-fold random split validation accuracy using Decision Tree with different cost functions

Figure 4: Decision tree validation accuracies



(a). Hepatitis dataset KNN decision boundary with Euclidean distance and $K = 1$

(b). Hepatitis dataset KNN decision boundary with Euclidean distance and $K = 5$

Figure 5: KNN Decision boundaries