Exploration of KNN's and Decision Trees on Two Benchmark Datasets

Lambert Francis, Patrick Janulewicz, Kevin Sohn

Abstract

In this paper, we examine the effect of multiple design choices for the k-Nearest Neighbour (KNN) and decision tree (DT) machine learning algorithms. The algorithms were trained and tested on UC Irvine's hepatitis and diabetic retinopathy datasets. Hyperparameter tuning, filtering, normalizing, and scaling were all taken into consideration. The hepatitis dataset saw a greater accuracy overall for both models. Overall, KNN tended to be the more accurate algorithm, and hepatitis tended to be the more accurate dataset. For hepatitis, the highest accuracy for the KNN model was 0.90 with $K = 5$ and a normalized dataset, while that of the decision tree algorithm was 0.80 with a maximum depth of 11. While many distance functions performed well for the KNN model, the Euclidean, Minkowski, and Hamming distances performed best on hepatitis data, while Manhattan and Minkowski worked best for DR data. The optimal cost function for decision trees was found to be misclassification, though entropy also performed well. A decision boundary plot was also demonstrated on the bilirubin and albumen levels of hepatitis patients, which showed that people with unusual levels of said quantities were more likely to succumb to the disease. In all, the efforts were a success, as the optimization of the data and parameters resulted in an accuracy of up to 90%.

Introduction

In this paper, we investigate the performance of the k-Nearest Neighbour (KNN) and decision tree (DT) algorithms on two benchmark datasets. The first is the Hepatitis dataset, which is composed primarily of boolean features. The second is the diabetic retinopathy dataset, also known as Messidor-2, which is composed of primarily continuous features. The purpose of this endeavour was to develop KNN and decision tree models that gave optimal performances, as well as understanding their behaviour.

To do this, precedent was taken into consideration. For instance, a 2015 paper by Yldirim [1] considered filtering the set to remove redundant, noisy features in the hepatitis dataset. For the Messidor-2 dataset, a notable 2015 study by Antal and Hajdu [2] features a more advanced approach, an ensemble of different machine learning classifiers, with an accuracy of 90%. We drew inspiration from these works to optimize the algorithms' performances on the data.

The data was first processed by removing incomplete entries. After processing, three strategies were implemented to increase the performance of the model. These were normalizing all values between 0 and 1, filtering out parameters with weak correlation, and correlation scaling. It was found that filtering substantially improved the accuracy of both models. Scaling appeared to have a small positive effect, but was generally inconclusive. Normalizing greatly improved the accuracy of the KNN model but not the decision tree; this is to be expected, since normalizing will not affect the decision tree's branches. Finally, before beginning the primary analysis, hyperparameter tuning was performed by L-fold cross validation.

The data analysis then began by finding the accuracy of the models subject to different constraints. Overall, the KNN model had a higher accuracy for both datasets, with the highest accuracy of 0.90 coming from the normed hepatitis data. The highest accuracy of the decision tree model was calculated to be 0.80, also for the hepatitis data.

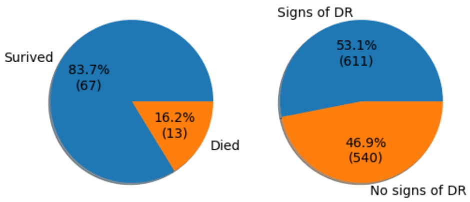
Now, consider the hyperparameter's effect on accuracy. For the KNN model, the k value with the highest accuracy was found to be 5 for the hepatitis data and about 7-9 for the diabetic retinopathy data. For the decision tree, the ideal maximum depth was found to be 11 for hepatitis and 7 for diabetic retinopathy.

We also studied different distance and cost functions. The differences were moderate between different distance functions. The Euclidean, cosine similarity, and Hamming distances each had the highest accuracy for the hepatitis set with 0.80. However, the hamming distance often performed best during testing. For the diabetic retinopathy set, most performances were similar, though the Euclidean and Minowski distances had the edge. Looking at the decision tree, the hepatitis data had the highest accuracy when the misclassification cost was used, while the diabetic retinopathy was optimized by choosing an entropy cost function.

Finally, decision boundaries were plotted for the bilirubin and albumen levels of the hepatitis dataset. The decision boundary plots showed that the survivors tended to cluster around the same bilirubin and albumen levels. Those who died generally had abnormally high bilirubin levels or abnormally low albumen levels. In the end, the decision boundary plots provided an intuitive look at the algorithms' implementations and effectiveness.

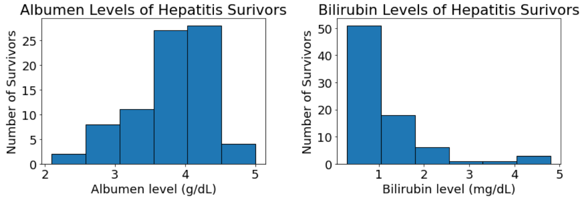
Datasets

The first datatset for hepatitis classifies patients as either living or deceased. Futhermore, the diabetic retinopathy dataset classifies patients as either showing signs of DR or showing no signs. For both sets, any rows with missing information were discarded. After preparation, the hepatitis set and the DR dataset contained 80 and 1151 entries, respectively. Below are charts displaying the classification for hepatitis (left) and diabetic retinopathy (right).



As we can see, the class distribution for hepatitis shows that the overwhelming majority of hepatitis victims survived. For the DR dataset, the distribution is far more even, with only about 53% of individuals showing signs.

Prior to the main analysis, we examined the distribution of each feature and their respective cross correlation's with the labels. Due to KNN's sensitivity to noise, only features with correlations of magnitude > 0.15 were selected for this model. Taking the hepatitis dataset as an example, one of the most promising features was albumen levels, which had a correlation coefficient of about 0.48. Furthermore, bilirubin levels had a negative correlation coefficient of -0.35. The opposing correlation of these values with survival rate made them prime candidates for decision boundary figures as will be seen in the results section. Below are histograms of these two features.



On the other hand of the spectrum for the hepatitis dataset, liver firmness showed the lowest correlation of all, measuring only 0.06.

For the Messidor-2 dataset, the correlations were generally lower. Features with the strongest correlations were results of microaneurysm (MA) detection, which had coefficients as high as 0.29. The lowest correlation coefficient found in this dataset was a meager 0.00048.

Before measuring the accuracy of the models, hyperparameter tuning was performed on both the KNN and decision tree models for each dataset. This was done via L-fold cross-validation. For the KNN model, the hyperparameter considered was the number of neighbours; for the decision tree model, it was the depth of the tree.

For the hepatitis data, the training and validation set was chosen to have size 60, leaving 20 points to serve as the test set. L-fold cross validation was then performed for $L = 10$. The mean squared error was then plotted as a function of the hyperparameter to form a validation curve. These curves can be found in the appendix. For the KNN tuning, the first 30 values of k were taken into consideration. For the decision tree tuning, it becomes far more computationally complex. As a result, the maximum depth was set to 10.

The Messidor-2 data was simililarly tuned for optimal hyperparameters. The training and validation set had size 800, while the test set had size 351. We selected $L = 10$ for the cross validation. The hyperparameter ranges were the same as for the hepatitis data, with 30 for KNN and 10 for decision trees.

With this methodology, the optimal hyperparameter was found for each dataset and each model. For the KNN model, the effects of filtering, scaling, and normalizing were also taken into account. See table 1 for the complete set of results.

Finally, despite the good that can come from these predictions, it is important to discuss some ethical issues that come along with it. One notable fact is that the hepatitis data has a disproportionate number of men compared to women. There are about 5 times as many men as women in this study. Since the number of women was small, the number of women that died was zero. As a result, the algorithm may decide that women are not at risk of dying from hepatitis even if they are. This could prevent women from seeking medical advice, thus allowing their conditions to worsen.

Results

The final KNN models, scaled with both distance and using selected features, had accuracies of 0.90 and 0.72, for the hepatitis and diabetic retinopathy datasets, respectively. These are improvements over the initial accuracies of 0.8 and 0.67 with no filtering or scaling. The hyperparameters used for each setup can be found in table 1 along with other performance metrics. The diabetic retinopathy performance is significantly worse than the more advanced methods mentioned in the introduction. The hepatitis dataset, however, has similar accuracies to those mentioned in the dataset description. A sample ROC plot for the best-performing KNN hepatitis model can be found in Figure 1 below.

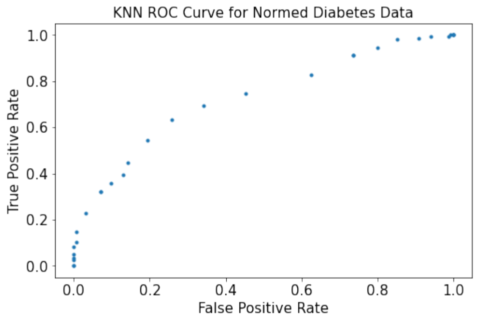


Figure 1: ROC plot for normalized diabetic retinopathy data

Algorithm	Dataset	Filtered	Scaled	Normed	K	Accuracy	Precision	Recall	F1 Score	AUC
KNN	Hep	0	0	0	5	0.80	0.83	0.94	0.88	0.80
KNN	Dia	0	0	0	7	0.67	0.78	0.57	0.66	0.72
KNN	Hep	1	0	0	7	0.80	0.88	0.88	0.88	0.88
KNN	Dia	1	0	0	5	0.72	0.76	0.64	0.69	0.68
KNN	Hep	1	1	0	9	0.85	0.84	1.0	0.91	0.92
KNN	Dia	1	1	0	5	0.67	0.76	0.64	0.69	0.68
KNN	Hep	0	0	1	5	0.90	0.94	0.94	0.94	0.94
KNN	Dia	0	0	1	26	0.66	0.78	0.55	0.64	0.76
DT	Hep	0	0	0	11	0.80	0.88	0.88	0.88	0.44
DT	Dia	0	0	0	7	0.63	0.73	0.53	0.61	0.69
DT	Hep	0	0	1	11	0.80	0.88	0.88	0.88	0.44
DT	Dia	0	0	1	7	0.63	0.73	0.53	0.61	0.69

Table 1: Model information and preparation steps for the different models shown in this paper. Select Features and Scaled Distance have 1 if those steps were taken for the model, otherwise 0. K refers to the max depth of the decision tree or the number of nearest neighbors. All hyperparameters were chosen using cross-validation with validation sets of size 10.

The decision trees performed worse than their best performing KNN counterparts in both datasets. With accuracy values of 0.80 versus 0.90 and 0.63 versus 0.72 for the hepatitis and diabetic retinopathy datasets. This is unsurprising in the diabetic retinopathy case as we expect KNN to perform well with large amounts of data, however, in the hepatitis case it is surprising for the KNN to perform better given the limited data. We also note that the norming of the data had no effect on the decision trees as expected.

In order to determine the effect of the number of nearest neighbors on the model's effectiveness, different base (unfiltered, unnormed) KNN model's were created with hyperparameter K ranging from 1 to 30. A plot of the accuracies is shown in Figure 2. We note that the accuracy of the hepatitis model quickly becomes stuck at 80%. This is likely because the dataset is so skewed. Say there are 9 deaths in the training set. Then as soon as we have 19 neighbors, every single prediction will result in predicting life. This might help explain the flat line present in the hepatitis dataset in contrast to the much larger diabetic retinopathy's fluctuations. We also note that typically in the range of the first few, accuracy slightly increases or stays the same as training points are still in the vicinity, and then quickly deteriorates or stagnates as we start considering points far away which are unlikely to resemble the data point. The initial significant swings of the hepatitis dataset are likely because the test set is small ($n = 20$) and so even small variations cause significant swings in the accuracy, making it difficult to draw conclusions.

We also compare the effect of the maximum tree depth on the accuracy. A plot of this accuracy can be found in Figure 3. Here both datasets are more consistent with what we would expect. Initial max depth increases see a big increase in accuracy in in the diabetic retinopathy model. However, in both cases the accuracies quickly stagnate as adding more regions does little to separate outcomes. This might suggest that not many features are needed to effectively model the outcomes as both perform fairly well with $maxdepth = 2$, which corresponds to 4 regions. This is something that could be explored in a future analysis.

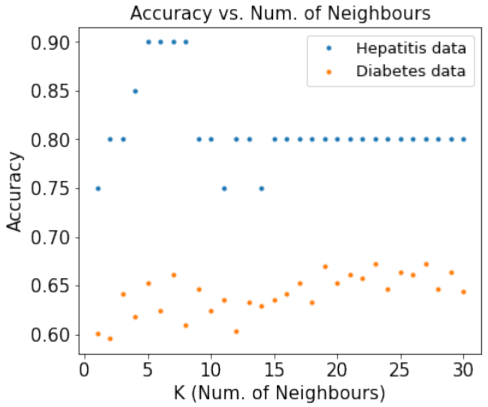


Figure 2: Accuracy of the unfiltered, unnormed KNN model for different number of neighbors values.

d

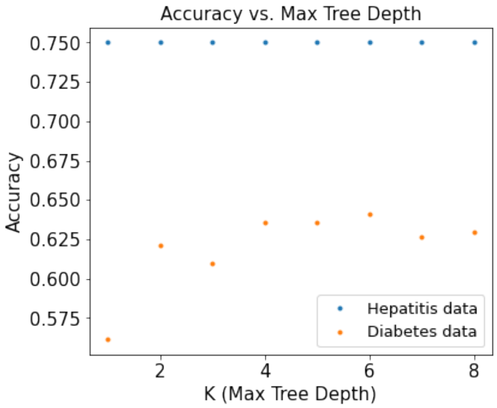


Figure 3: Accuracy of the unfiltered, unnormed decision tree model for different values of max depth.

The two distance functions that performed best for KNN in the hepatitis dataset were cosine similarity and Hamming distance as can be seen in Table 2. Given that the majority of features in the hepatitis dataset are discrete, it's unsurprising that the Hamming distance performs well here. For KNN on the diabetic retinopathy dataset, the Euclidean distance function performs best. Given that this dataset is comprised mostly of continuous features, this also fits our expectations. We note that Cosine Similarity performs poorly at an accuracy of 56%, fairly close to random. Surprisingly the Hamming distance performed on par with the other continuous distance functions.

The best performing tree cost functions were misclassification for hepatitis and a tie between all three for diabetic retinopathy.

Algorithm	Dataset	Distance/Cost Fnc.	Accuracy
KNN	Hep	Euclidean	0.8
KNN	Dia	Euclidean	0.67
KNN	Hep	Manhattan	0.7
KNN	Dia	Manhattan	0.65
KNN	Hep	Minkowski	0.8
KNN	Dia	Minkowski	0.67
KNN	Hep	Cosine Similarity	0.8
KNN	Dia	Cosine Similarity	0.56
KNN	Hep	Hamming	0.8
KNN	Dia	Hamming	0.65
DT	Hep	Misclassification	0.8
DT	Dia	Misclassification	0.65
DT	Hep	Entropy	0.75
DT	Dia	Entropy	0.65
DT	Hep	Gini	0.75
DT	Dia	Gini	0.65

Table 2: Table of accuracy per distance/cost function per algorithm. We chose to use the unfiltered and unscaled version for both datasets. Both algorithms were initialized with their respective optimal hyper-parameters listed in 1. We set the Minkowski distance free parameter, $p = 3$.

Shown in Figure 4 and Figure 5 are the decision boundaries of the KNN and decision tree models for the hepatitis data. The KNN hepatitis regions are fairly clean with only one deep intrusion into the life region. The hepatitis decision tree boundary also appears simple and without too many sharp or sudden deep intrusions. This suggests the models are relatively simple as desired. In fact, these two parameters were carefully chosen to produce informative boundary decision plots. The choice of parameters has two motivations. The first is that they are continuous values rather than binary ones. Continuous values make more informative decision boundary plots, as they have a wide range of values for each axis. On the other hand, binary features force the data into two categories, which makes for a rather uninformative figure. The second motivation for our choice is that the two features had a strong opposing correlation. Bilirubin's high positive correlation paired nicely with albumen's high negative correlation, creating plot with well defined sections.

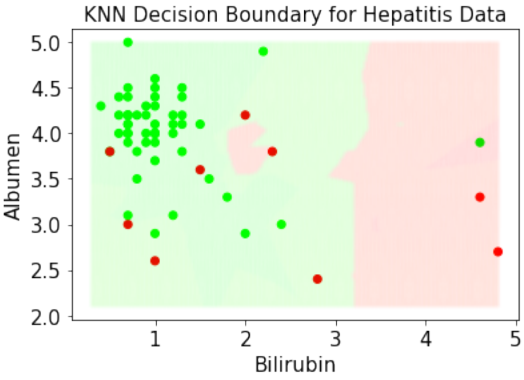


Figure 4: Decision boundaries of the unnormed, unfiltered KNN hepatitis model. These parameters were selected as they were continuous with a high cross-correlation with the outcome class.

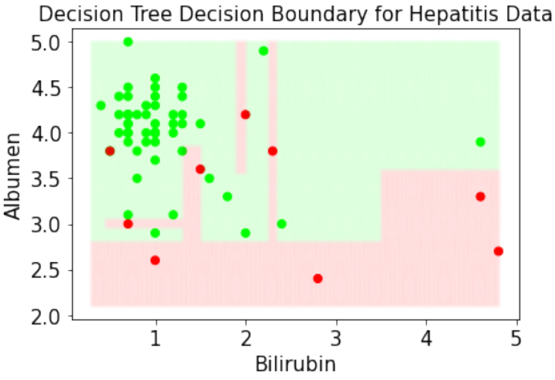


Figure 5: Decision boundaries of the unnormed, unfiltered decision tree hepatitis model. These parameters were selected as they were continuous with a high cross-correlation with the outcome class.

As a final experiment, we noticed upon inspection of the hepatitis dataset that the majority of missing values were from the Protime feature. Because the size of the hepatitis set is small, it was believed that having more data may increase accuracy. Removing the Protime feature increased the size of the intact data from 80 to 112. This resulted in a 2% increase in accuracy upon applying it to one of the models. However, given the limited size of the test data set, more significant testing is likely needed to determine if dropping this feature is due to the increased data, or just random variation due to the new test set.

Scaling the features proportionally to their correlation coefficients also proved inconclusive, its only effect being a slight lowering of the diabetic retinopathy accuracy for KNN. Further experiments are needed to see if we can improve the performance further using scaling.

Discussion and Conclusion

We found that the best performing model for the hepatitis dataset was a KNN with an accuracy of 90%. Its preparation included hyper-parameter tuning with L=10 cross validation, filtered features, and a normalized scale for the continuous features. The decision tree reported an accuracy of 80% with a simple decision boundary. For the diabetic retinopathy dataset the KNN with selected features and untouched scale and normalization performed best with an accuracy of 0.72 and a similar cross validation.

Accuracies dependence on the hyperparameters of KNN's and decision trees behaved largely as expected, with inconsistencies in the hepatitis dataset attributed to the small sample size of the test set. In particular the decision trees stagnate as expected where increasing the maximum depth does not cause significant change in the chosen tests.

Distance and cost functions for the data were also largely in line with expectations, with continuous distance functions performing best for the largely continuous dataset and vice versa. We also found norming the data was largely helpful for KNN so as to not weigh any features more than others simply due to the units they happened to appear in.

Attempting to scale the distances according to correlation was not conclusive and should be researched further. Further study should go into studying how the performance of the KNN could be enhanced by tweaking the scaling of features.

We also found the decision boundaries of our models to be reasonable, with well defined regions.

In summary, the areas of interest for further exploration include research into selecting scale factors for KNN features, possibly increasing the effectiveness of the hepatitis dataset by increasing the training size at the cost of eliminating a feature, and more in depth research as to the slight differences between the various distance functions.

Statement of Contributions

Lambert Francis, Patrick Janulewicz, and Kevin Sohn each contributed to this project fairly, evenly, and to the best of their abilities. Lambert focused on algorithm implementation and optimization, Patrick focused on data preparation and decision boundaries, and Kevin focused on hyperparameter tuning and accuracy. All members contributed to the writing of the report.

References

[1] Andras Hajdua Balint Antal. “An ensemble-based system for automatic screening of diabetic retinopathy”. In: *Knowledge-Based Systems* (2014). DOI: http://dx.doi.org/10.1016/j.knosys.2013.12.023.

[2] Al Khaldy et el. “Improve Class Prediction By Balancing Class Distribution For Diabetes Dataset”. In: *INTERNATIONAL JOURNAL OF SCIENTIFIC TECHNOLOGY RESEARCH VOLUME 9, ISSUE 04, APRIL 2* 9.4 (2020).

[3] Pinar Yldirim. “Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease”. In: *International Journal of Machine Learning and Computing* 5.4 (2015). DOI: http://dx.doi.org/10.7763/IJMLC.2015.V5.517.

Appendix: Validation Curves

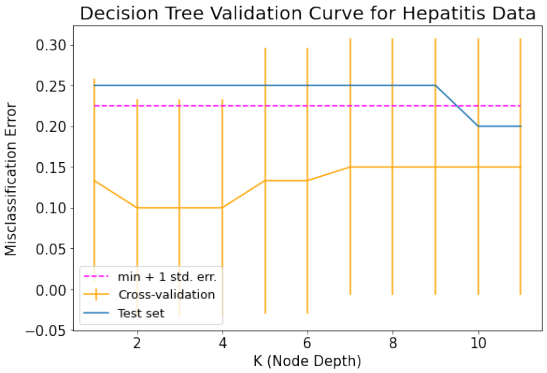


Figure 6: Cross-validation plot used to tune the max depth of the hepatitis decision tree. Specifics of the cross validation are discussed in the Dataset section.

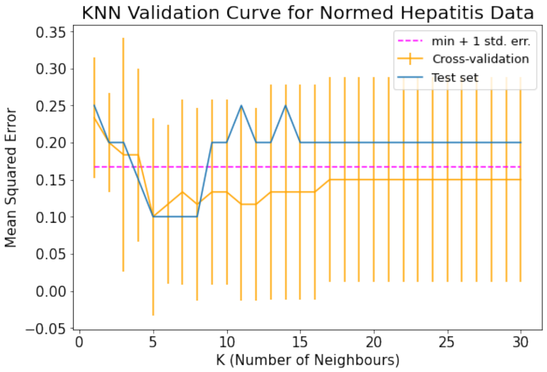


Figure 7: Cross-validation plot used to tune the number of neighbors for the hepatitis normed KNN, the best performing KNN of those studied. The mechanics of the cross-validation are discussed further in the Dataset section.