Simple Diagnoses using KNN and Decision Trees

Nicholas Dahdah (260915130), Jonayed Islam (260930902), Jasper Yun (260651891)

Abstract

Machine learning models are increasingly applied to classification tasks in which they can achieve greater accuracy than expert human equivalents. However, complex models are computationally expensive; simpler models may offer similar performance at a reduced cost. We implemented K-nearest neighbors (KNN) and decision tree (DT) models which were trained on benchmark hepatitis and diabetes datasets. In the training process, we examined the cross-correlation between features and outcomes to determine the best features to use. We used 10-fold cross-validation to compare the training accuracy against the validation accuracy. On the hepatitis dataset, we achieved testing accuracies of 76.7% and 83.8%, for the KNN and DT, respectively. On the diabetes dataset, we achieved testing accuracies of 64.4% and 68.6%, for the KNN, and DT, respectively. While this performance does not surpass human expert performance, the results of our classifiers can be used to aid doctors in the automated rapid classification of patients.

Introduction

Simple machine learning models can be used for classification with surprisingly good accuracy. We implemented K-nearest neighbors (KNN) and decision tree models on two benchmark datasets, hepatitis [1] and diabetes [2].

Various previous papers have used the hepatitis dataset to compare different ML models for classification accuracy. For instance, [3] implemented a classifier using a nonlinear-weighted Bayes discriminant function; their best results on the Hepatitis dataset achieved a 95.4% training accuracy with 79.4% test accuracy when using 6.5 features, on average, of the 19 total features in the dataset. In work by the same authors using a genetic algorithm combined with the KNN, their classifier achieved 86.0% training and 69.6% test accuracies using 8.1 features.

The diabetes dataset has been used in research by [2] in which an ensemble system of ML models was applied to predict the presence of DR. The ensemble combines the outputs of multiple models to produce a more accurate prediction. Using this method, the ensemble system achieved 94% sensitivity, 90% specificity, and 90% accuracy.

Datasets

The two datasets used in this project are the Hepatitis Data Set and the Diabetic Retinopathy (DR) Debrecen Data Set, both pulled from the UC Irvine (UCI) Machine Learning Repository (MLR). The Hepatitis Data Set contains information from 155 subjects with hepatitis regarding the fatality of the disease, along with information about each subject and their symptoms. Unfortunately, this dataset contains missing values, labeled with a "?," which were removed before training any model, so as to not skew classification. The DR Debrecen Dataset contains information extracted from 1151 samples of the Messidor image set pertaining to the characteristics of retinal abnormalities and whether the sample contains signs of DR. This dataset does not have any missing values.

To better understand the data, our team computed the correlation between the feature we aimed to predict (fatality of hepatitis and signs of diabetic retinopathy) and each of the other features in the dataset and compiled it into Table 1. This identified the features that were most related to our feature of interest. For the hepatitis data, the four most correlated features were ascites, albumin, histology, and protime. For the diabetes data, the four most correlated features were the first four MA detection features. This correlation information also allowed us to drop certain features when training our KNN model and DT, as relatively little information about the feature we aim to predict is gained from features that do not correlate strongly with it. This will be addressed more at the end of the Results section.

It is worth noting that when training the KNN model, the hepatitis and diabetes datasets were normalized since KNN is sensitive to scaling. We did this to equally weigh each feature before training. Normalizing the dataset for KNN improves accuracy. We did not normalize the datasets for the DT.

Table 1. Correlation of features to diagnosis of hepatitis fatality (left) and DR (right)

AGE	-0.212769		
SEX	0.175876	SPIDERS	0.287839
STEROID	0.123830	ASCITES	0.479211
ANTIVIRALS	-0.108776	VARICES	0.345785
FATIGUE	0.181151	BILIRUBIN	-0.351557
MALAISE	0.275595	ALK PHOSPHATE	-0.189360
ANOREXIA	-0.185042	SGOT	0.078731
LIVER BIG	-0.194030	ALBUMIN	0.477404
LIVER FIRM	0.055978	PROTIME	0.395386
SPLEEN PALPABLE	0.135643	HISTOLOGY	-0.456856

pre-quality	0.062816	Exudate3	0.038281
pre-screening	-0.076925	Exudate4	0.104254
MA1	0.292603	Exudate5	0.142273
MA2	0.266338	Exudate6	0.151424
MA3	0.234691	Exudate7	0.184772
MA4	0.197511	Exudate8	0.177313
MA5	0.161631	MaculaDist	0.008466
MA6	0.127861	OpticDics	-0.030868
Exudate1	0.058015	AMFM	-0.042144
Exudate2	0.000479		

According to the donation policy of the UCI MLR, all datasets require explicit permission to be made public and all personally identifiable information must be removed. While this policy protects subjects from violations of privacy, which is an important ethical consideration, it does not ensure that the datasets themselves are representative of present-day reality. Verification by independent third parties, as well as the revisitation of older datasets to ensure the data is still in-line with current medical standards, would substantiate the credibility of these datasets.

Results

After tuning the hyperparameters of the KNN and DT models, K and maximum depth, respectively, via cross-validation, we evaluated our models using a test set of data unique from the training and validation data. The accuracy of each model on each dataset is presented in Table 2. The DT model outperforms the KNN model by a margin of 7.1% and 4.2% on the hepatitis dataset and diabetes datasets, respectively. Tables 3 and 4 show the confusion matrix of both the KNN and DT models run on test sets of the hepatitis and diabetes datasets.

Table 2. KNN and DT test accuracy

	KNN	DT
Hepatitis	76.7% (K = 7, Euclidean)	83.8% (Max depth = 5, Entropy)
Diabetes	64.4% (K = 22, Euclidean)	68.6% (Max depth = 8, Gini)

Table 3. KNN confusion matrix from test set

	PREDICTED		
ACTUAL	Total: 30	Live	Die
	Live	22	6
	Die	1	1

	PREDICTED		
ACTUAL	Total: 275	DR+	DR-
	DR+	82	41
	DR-	57	95

Table 4. DT confusion matrix from test set

	PREDICTED		
ACTUAL	Total: 30	Live	Die
	Live	23	1
	Die	4	2

	PREDICTED		
ACTUAL	Total: 275	DR+	DR-
	DR+	118	0
	DR-	56	101

KNN

To determine the effect of K on the KNN model's testing accuracy, we ran a 10-fold cross-validation on the hepatitis and diabetes datasets while varying the value of K. For the hepatitis and diabetes datasets, K was varied from 1 to 25 and 1 to 100, respectively; the maximum value of K was limited by the size of the available dataset. Figure 1 shows the validation accuracy obtained from this process as the value of K was varied. There is no specific trend when varying K, but we note that the optimal values for K are 7 and 22 for the hepatitis and diabetes datasets, respectively.

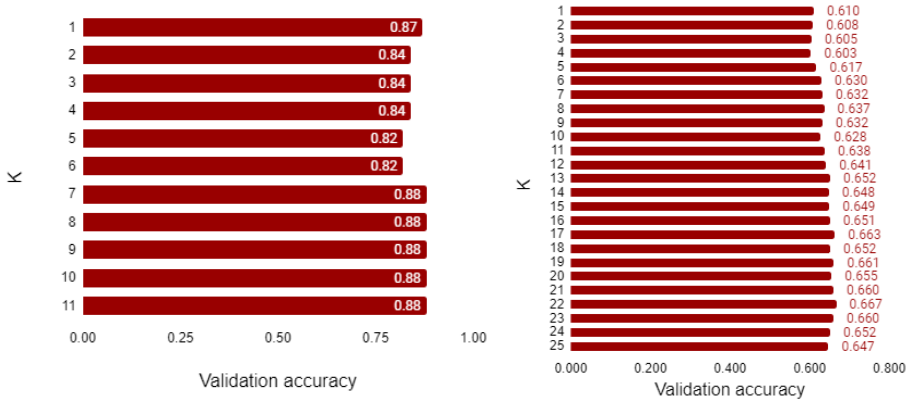

Fig. 1. Validation accuracy with varying K for the hepatitis (left) and diabetes (right) datasets

We also examined the effect of varying the model's distance function. A 10-fold cross-validation with varying distances yielded the validation accuracies shown in Table 5.

Table 5. KNN validation accuracy versus distance function

Distance Function	Hepatitis	Diabetes
Euclidean	88.0% (K = 7)	66.7% (K = 22)
Manhattan	88.0% (K = 7)	65.1% (K = 22)

Figures 2 and 3 show the decision boundaries (of the most correlated features) obtained from training the KNN model on the hepatitis and diabetes datasets. These decision boundaries form complex regions, which may indicate overfitting of the model. However, reducing the numerous features to a 2-dimensional decision boundary introduces a large loss of information, but including more than 2 features is difficult to illustrate in a meaningful way.

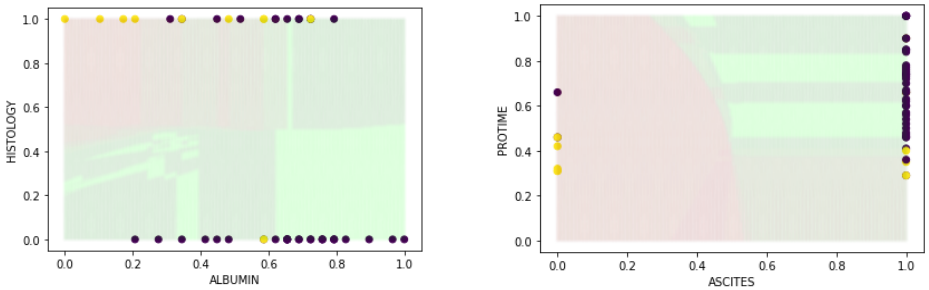

Fig. 2. KNN decision boundary of the hepatitis dataset

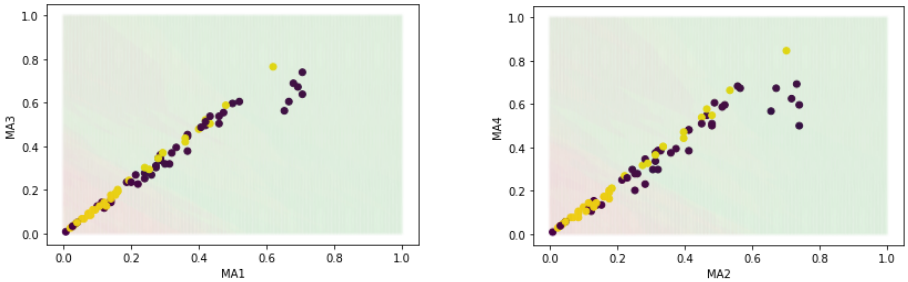

Fig. 3. KNN decision boundary of the diabetes dataset

Using the results of the experimentation, we determined the features that when removed, led to an increase in accuracy, and we removed these features before training our models. As a result, the validation accuracy on the hepatitis and diabetes datasets became 86.67% and 67.33%, respectively, using 12 of 19 and 12 of 20 features. This reduces the complexity and run-time of the machine learning models while improving the overall validation accuracy. However, we note that specific combinations of removed features may yield even better validation accuracies, as evidenced in Table 7.

Table 7. Validation accuracy versus removed features for the hepatitis (left) and diabetes (right) datasets.

```
Testing hepatitis dataset by dropping one column:
        dropping column #0      max accuracy = 0.90000   k = 5
        dropping column #1      max accuracy = 0.90000   k = 5
        dropping column #2      max accuracy = 0.93333   k = 7
        dropping column #3      max accuracy = 0.90000   k = 8
        dropping column #4      max accuracy = 0.90000   k = 5
        dropping column #5      max accuracy = 0.90000   k = 3
        dropping column #6      max accuracy = 0.90000   k = 6
        dropping column #7      max accuracy = 0.96667   k = 5
        dropping column #8      max accuracy = 0.93333   k = 5
        dropping column #9      max accuracy = 0.86667   k = 5
        dropping column #10     max accuracy = 0.90000   k = 5
        dropping column #11     max accuracy = 0.86667   k = 5
        dropping column #12     max accuracy = 0.96667   k = 7
        dropping column #13     max accuracy = 0.90000   k = 7
        dropping column #14     max accuracy = 0.93333   k = 8
        dropping column #15     max accuracy = 0.90000   k = 5
        dropping column #16     max accuracy = 0.90000   k = 5
        dropping column #17     max accuracy = 0.93333   k = 8
        dropping column #18     max accuracy = 0.93333   k = 5
average accuracy: 0.912281
```

```
Testing diabetes dataset by dropping one column:
        dropping column #0      max accuracy = 0.65347   k = 29
        dropping column #1      max accuracy = 0.65347   k = 29
        dropping column #2      max accuracy = 0.65347   k = 29
        dropping column #3      max accuracy = 0.65347   k = 29
        dropping column #4      max accuracy = 0.65347   k = 33
        dropping column #5      max accuracy = 0.65347   k = 29
        dropping column #6      max accuracy = 0.66337   k = 31
        dropping column #7      max accuracy = 0.68317   k = 31
        dropping column #8      max accuracy = 0.67327   k = 25
        dropping column #9      max accuracy = 0.66337   k = 27
        dropping column #10     max accuracy = 0.65347   k = 1
        dropping column #11     max accuracy = 0.67327   k = 31
        dropping column #12     max accuracy = 0.67327   k = 31
        dropping column #13     max accuracy = 0.67327   k = 31
        dropping column #14     max accuracy = 0.67327   k = 31
        dropping column #15     max accuracy = 0.67327   k = 31
        dropping column #16     max accuracy = 0.65347   k = 1
        dropping column #17     max accuracy = 0.64356   k = 1
        dropping column #18     max accuracy = 0.70297   k = 1
average accuracy: 0.664409
```

Decision Tree

To determine the effect of tree depth on the accuracy of the DT model, we ran a 10-fold cross-validation on the hepatitis and diabetes datasets, varying max depth from 1 to 20 to find its optimal value. The validation accuracies achieved are shown in Fig. 4. The optimal values for maximum tree depth are 5 and 8 for the hepatitis and diabetes datasets, respectively. It is worth noting that we saw no change in accuracy for depth greater than 12.

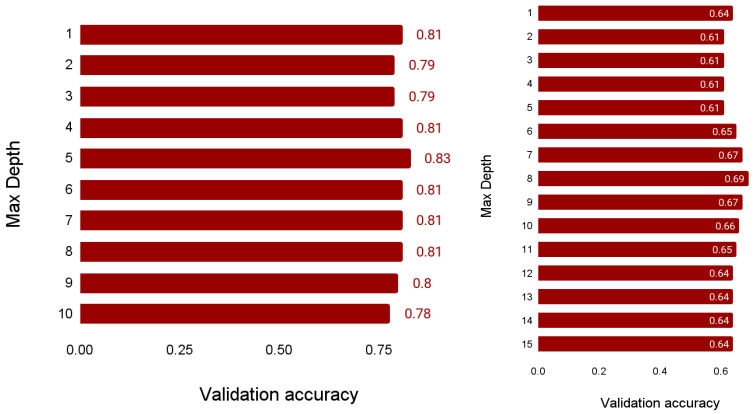

Fig. 4. Validation accuracy with varying max depth for the hepatitis (left) and diabetes (right) datasets

For both datasets, we found that varying the cost function did not produce significantly different results. A 10-fold cross-validation with varying cost functions yields the validation accuracies shown in Table 6.

Table 6. DT validation accuracy versus cost function

Cost Function	Hepatitis	Diabetes
Misclassification	82.5% (Max depth = 5)	64.2% (Max depth = 8)
Entropy	80.0% (Max depth = 5)	64.5% (Max depth = 8)
Gini Index	83.8% (Max depth = 5)	63.6% (Max depth = 8)

Figures 5 and 6 show the decision boundaries (of the most correlated features) obtained from training the DT model on the hepatitis and diabetes datasets. Unlike the KNN decision boundaries, the DT decision boundaries are much clearer, indicating distinct boundaries for specific diagnoses. Low albumin levels and medium ascites levels were indicative of death from hepatitis.

The decision boundary plots of the diabetes data show two groups. Those on the top left with higher MA3 or MA2 values and lower MA1 and MA4 values were more likely to have DR.

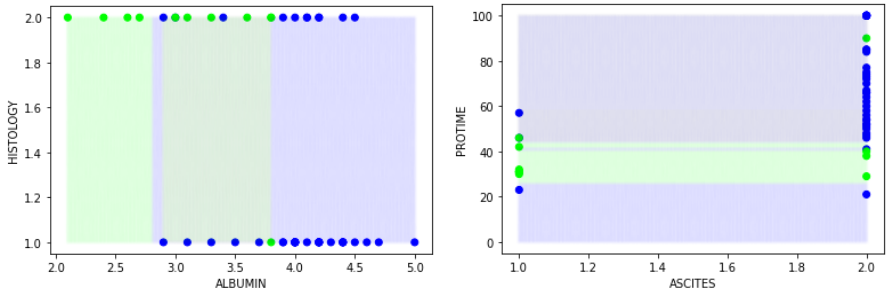

Fig. 5. DT decision boundary of the hepatitis dataset

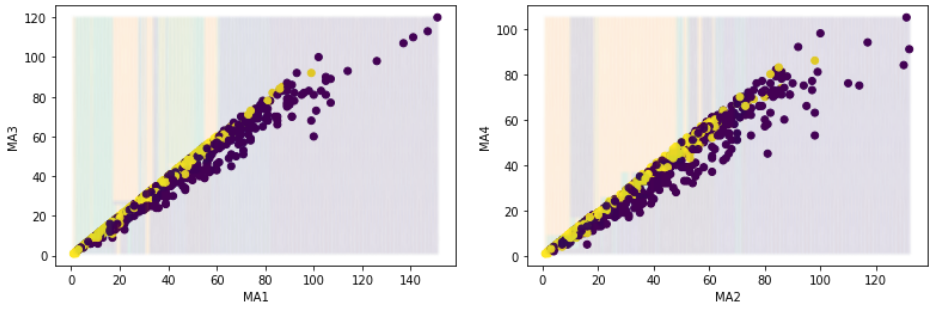

Fig. 6. DT decision boundary of the diabetes dataset

Discussion and Conclusion

We can extract certain trends from the results presented above. For additional insight, we also noted the effect of dropping certain features that were uncorrelated to the feature we aimed to predict.

Firstly, we note there is a strong link between the accuracy of both the KNN and DT models and the correlation of the features. Since the hepatitis dataset has features with higher correlations to the feature we aim to predict, we obtain test accuracies that are relatively higher than those of the diabetes dataset, which has features that are not as highly correlated with the output.

For the KNN model, we examined the effect of different distance functions and values of K on validation accuracy. For the DT model, we examined the effect of different cost functions and tree depths on validation accuracy. There are only minor differences in accuracy when using different distance and cost functions. We conclude that in the scope of these datasets, it is not the most important hyperparameter to tune for our models. However, the value of K for KNN and the tree depth for DT are more important. For KNN models, smaller values of K cause the model to overfit to the training data, while very large values of K cause the model to underfit. As we varied K in our experiment, we were able to find an optimal value that maximized validation accuracy. For DT models, when the depth is too shallow the model tends to underfit the data, whereas overly deep trees tend to overfit. As we varied the tree depth in our experiment, we were able to find an optimal value that maximized validation accuracy.

To improve the accuracy of these models, it would be interesting to work with larger datasets to provide more information to our models. Examining the dropping features to a greater extent would also prove useful, as the results we achieved were already promising. We may want to investigate variations of our models, such as a weighted KNN model where the classification task weighs each of the K-nearest neighbors by the distance from the test point or a DT model with pruning. These more advanced techniques are able to capture more information from the data, and as such, would likely yield higher test accuracies.

Overall, the KNN and DT models were both successful in diagnosing both the fatality of hepatitis and signs of diabetic retinopathy. Given that both models have better prediction accuracy than random selection, these models can help medical professionals make better decisions. However, since their accuracy is not sufficiently higher for diagnosing patients with certainty, these models cannot be solely relied on. Rather, an approach of using collaborative intelligence is advisable; doctors can use the models to get an initial idea before prescribing further tests. These models can thus help rapidly flag more serious cases of hepatitis or diabetes.

Statement of Contributions

Nicholas handled the cross-validation and hyperparameter optimization of the models. Jonayed handled the training and experimentation of the decision trees. Jasper handled the training and experimentation of the KNN model.

References

[1] D. Dua and C. Graff. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] B. Antal and A. Hajdu. “An ensemble-based system for automatic screening of diabetic retinopathy,” arXiv:1410.8576 [cs], Oct. 2014.

[3] M. L. Raymer, T. E. Doom, L. A. Kuhn and W. F. Punch. “Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm,” in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 33, no. 5, pp. 802-813, Oct. 2003, doi: 10.1109/TSMCB.2003.816922.