COMP551Mini Project1:
Classification on health dataset via KNN and decision Tree

Group 66
Yao Chen (260910483)
Chuyang Zhang (260840707)
Hanzhi Zhang (260908416)

Abstract

This mini-project aims to implement two classification algorithms: K-Nearest Neighbor (KNN) and Decision Trees, and increase the accuracy by modifying the models. We use datasets from Hepatitis and Diabetic Retinopathy Debrecen patients and clean them. The initial accuracies of the two models are observed and reported. To improve the prediction model, our group used 8-fold and 10-fold cross-validation to pick the best hyperparameter (K) for the KNN algorithm. Moreover, we also used cross-validation for the Decision Tree algorithm and compared three different cost functions (misclassification cost, entropy cost, and Gini index) to maximize the training set of the prediction accuracy. Besides, we plot and visualize the impact on accuracy by changing the hyperparameters (number of neighbours and maximum tree depth). After optimizing the hyperparameter and the distance/cost function, we find that the Decision Tree algorithm is better for predicting the Hepatitis dataset with an accuracy of 86.0%; the KNN model achieves a higher accuracy (66.3%) for predicting the Diabetic Retinopathy Debrecen datasets. Finally, we explore how dropping dataset features can improve prediction and make the model more accurate.

Introduction

Hepatitis and Diabetic Retinopathy Debrecen datasets are two well-known datasets published by UCL Machine Learning Repository, with binary label feature commonly used for training and testing machine learning models [1][2]. In this mini-project, our group investigated the performance of two classification techniques, K-Nearest Neighbor and Decision Trees. After data cleaning, the initial model was built on hepatitis and diabetic retinopathy datasets. Then our goal is set to optimize the accuracy by performing a series of experiments. Our experiment includes: comparing the accuracy of KNN and Decision Tree algorithm; testing the effect of change in different hyperparameters; observing the impact of maximum decision tree depth; observing the difference of using different distance/cost functions for both models; generating the decision boundary plots for each model. After running the above experiments, we find that, in general, with the increase of hyperparameter k value in the K-NN model, the training set's accuracy decreases due to the under-fitting. In contrast, with more depth in the decision tree, the accuracy converges to 1 due to the over-fitting. After adjusting parameters, we find that, on average, the Decision Tree algorithm better predicts the Hepatitis dataset, whereas KNN has a greater accuracy on Diabetic Retinopathy Debrecen datasets. In addition, we explore other possibilities to increase the accuracy, which we try to find the key features by comparing the correlation and dropping the features manually.

Datasets

Our classification model uses two data sets: 1) the Hepatitis dataset, with 155 instances and 19 attributes, 2) the Diabetic Retinopathy Debrecen dataset, with 1151 instances and 20 attributes. Then the *Pandas dataframe* package is used to manipulate the dataset. In the Hepatitis dataset, we preprocess the datasets by removing the missing values indicated by question marks from the dataset ("?"). After processing, only 80 instances are remaining (there is no missing field in the Diabetic Retinopathy Debrecen dataset).

When reviewing our data set, the ‘Class’ attribute is treated as our binary label in both datasets, and we modify the ‘Class’ value in the Hepatitis dataset from DIE/LIVE to 1/0 for modelling. Also, the dataset has binary features (with values 1 or 2) and continuous features (like ages). We convert the binary features to values 0 or 1 and rescale and normalize the continuous features into the range between 0 and 1 for more straightforward distance/cost calculations. At the same time, to better understand the (statistical tests) the data, we calculated the correlation between attribute ‘Class’ and other attributes in Table I in Appendix Section.

Finally, we split each dataset into a training set (80%) and a testing set (20%) randomly so that we can apply our classification model to them in the future.

Indeed, using these databases and machine learning can help the real-world medical system[3]. However, there are a lot of possible ethical concerns that can arise when working with this kind of data, such as evaluating whether a drug or other intervention does work[4]. Some researchers have indicated that there is still the issue with privacy and surveillance, bias, and discrimination in dataset and modelling. Furthermore, it should never be neglected that the role of human judgement is also potentially brought additional bias[3].

Results:

We set up an initial run of both the KNN and the Decision Tree algorithm on two datasets to better understand the models and find and compare the initial accuracies. As for the initial setting, in the KNN algorithm, the hyperparameter K is set to be 3. In the Decision Tree classification, the initial max depth of the decision tree is 20. The initial accuracies are shown below in Table I. From the table, better initial accuracy is achieved by the Decision Tree model, which is about 3% better in both datasets.

Table I: The accuracy of running two unmodified classification algorithms on two datasets

	Hepatitis dataset	Diabetic Retinopathy Debrecen dataset
KNN	78.6%	58.7%
Decision Tree	81.2%	61.5%

Then, we want to study how different k values affect the accuracy. The algorithm of L-fold cross-validation is used to determine a better estimate of hyperparameter for the KNN model for both datasets.

Since two datasets have different sizes, we divide them into L=8 and L=10 equal parts, respectively. In a total of L runs, one subset among these equal parts is chosen as the validation set for each run, and the remaining training set is used to train our model. After the L runs, the average validation error and its variance were calculated and used to define the best model, as shown in Figure 1.

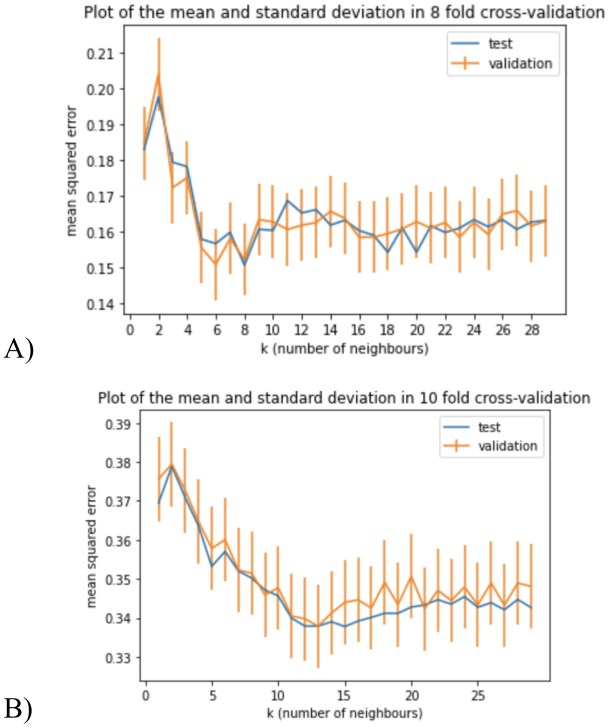


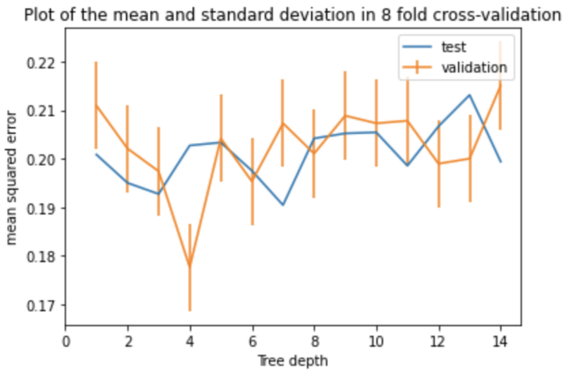
Figure1: The mean squared error for test and validation sets for a different number of neighbours. A: 8-fold cross-validation and B: 10-fold cross-validation was used to find the mean squared error (MSR).

Observing from Figure1(A), the overall trend found is that the mean square error for both the validation and the test sets decreased as K increased at interval k = [1,5], and a minimum error is reached around the interval of k = [6,8], after that the mean square error increase again. A similar trend can also be found in Figure1(B), where the minimum locate around interval k = [11,14]. Finally, by the rule of thumb, K = 8 and K = 13 are chosen for the Hepatitis and Diabetic Retinopathy KNN modelling, respectively. With the cross-validated hyperparameter, the KNN accuracy of modelling the Hepatitis dataset is 84.6% and 66.1% for the Diabetic Retinopathy dataset.

Next, considering the Decision Tree algorithm, 8-fold and 10-fold cross-validation are used to find the optimal tree depth to maximize accuracy, respectively. It has a similar logic with the K value for the KNN algorithm.

From cross-validation on the Hepatitis dataset, we did not find the correlation between the mean square error and the Tree depth this time, but when the tree depth was 4, the validation set had the minimum mean squared error. Thus, we used 4 tree depths to find the decision boundary shown below in Figure 2A).

A)



B)

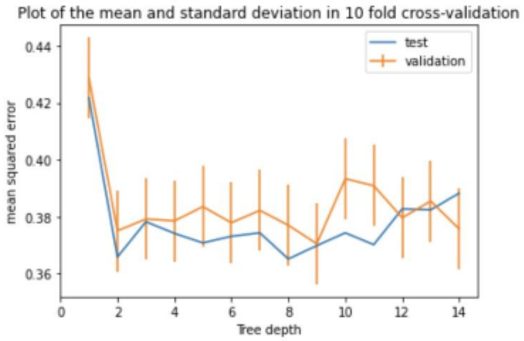


Figure 2: The mean squared error for test and validation sets for different Tree depths. A) 8-fold cross-validation and B) 10-fold cross-validation was used to find the mean squared error (MSR) for tree depth.

In figure 2B), the trend found was that the mean square error for both the validation and the test sets decreased for the first two K. After that, the mean squared error alternated around a value. From figure 2B, we used tree depth at 2 to find the decision boundary.

Aside from optimization on hyperparameter-like variables, the third experiment is about the choice of distance/cost functions. For KNN, The accuracy of using Euclidean’s distance function and Manhattan’s distance function is compared by increasing the K values. The comparison is shown in Figure 3, two plots on the left. It was observed that the two distancing functions gave similar accuracies; given different k values, we can observe an increasing trend as the number of neighbours increased, which may be due to the potential overfitting. Meanwhile, Manhattan's accuracies were slightly higher than the Euclidean distancing function for almost all K values from the plot at the upper left. Then consider the cost function used in the Decision Tree algorithm. There are three cost functions (misclassification, entropy and Gini index) being tested. Observing from the upper right of Figure 3, there is not much difference between the trend of the three cost functions, where they all give a maximum at k=3, but it is the only plot with a decreasing trend with the increasing value of k. Looking at the bottom right of Figure 3, it is noticeable that the misclassification function's accuracy is significantly lower than the other two cost functions.

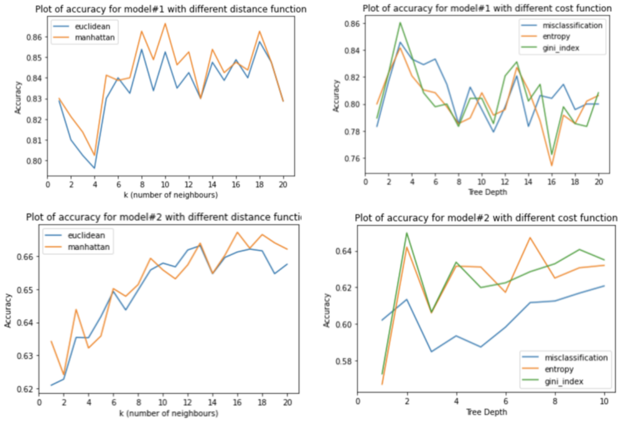


Figure 3: The accuracy of applying different distance/cost functions on two classification models. Upper left: apply distance function to Hepatitis dataset; bottom left: apply distance function to diabetic Retinopathy dataset; Upper right: apply cost function to Hepatitis dataset; bottom right: apply cost function to diabetic Retinopathy dataset

When plotting the decision boundary, we notice that there are 19 features for the Hepatitis dataset, which the dimension is too high to make an appropriate demonstration. If we want to graph the decision boundary for all the nineteen features, there would be 19 dimensions for the figure, which was hard to visualize on a 2-D plot. To better visualize the decision boundary, we loop through all combinations of two features and choose the one with the highest accuracy. If the selected two features are binary, the data points would only be plotted to four corners and not give much information, so we wanted two continuous features to visualize the decision boundary better. The two features that gave the highest accuracy are Age and Protime for KNN. We found the decision boundary and shown below in Figure 4A). Figure 4B) is the decision boundary using the Decision Tree model. We choose the features (Age and Protime) to make them consistent with what we get from the KNN model.

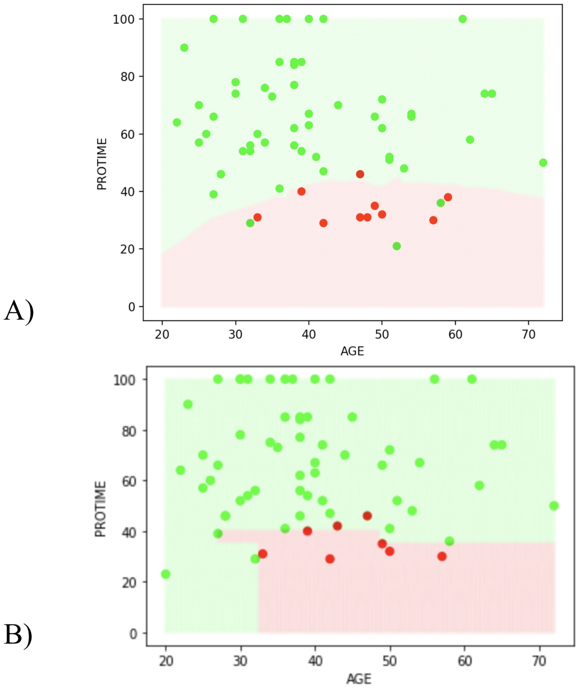


Figure 4: The KNN decision boundary for features Age and Protime. The number of neighbours used was six. The red dots indicated the actual case for death, and the green dots indicated the actual case for survival. The red area predicted death, and the green area predicted survival.

After we optimized the hyperparameters K and tree depth and selected the best distancing/cost function to increase the accuracy, we applied a final round of the KNN and the Decision Tree algorithms to the two datasets and tested the final accuracy. It is shown in table II.

Table II: The final accuracy of running two classification algorithms on two datasets

	Hepatitis dataset	Diabetic Retinopathy Debrecen dataset
KNN	85.3%	66.3%
Decision Tree	86.0%	63.8%

In general, after optimization, all the accuracies increased compared to the unadjusted model. We found that the Decision Tree model was better for predicting the Hepatitis dataset, and the KNN model was better for predicting the Diabetic Retinopathy Debrecen dataset.

Further Experiment

Lastly, we wanted to test if removing some features could increase the accuracy for both classification algorithms with the optimized hyperparameter and distancing function. We used the Hepatitis dataset and looped around nineteen times, each time removing one feature. It was found that the accuracy effectively increased by removing the feature SGOT and the ALK PHOSPHATE. The final accuracy after removing these two features for prediction was 87.6%. The appendix showed that the SGOT feature had a 0.079 correlation value with the class, and the ALK PHOSPHATE feature had a -0.189 correlation value with the class. We also applied the same method to the Diabetic Retinopathy Debrecen dataset. After removing the features “MA-1.0” and “Exudates-2”, the accuracy was 67.3%, which increased slightly.

Discussion and Conclusion

In general, we learn how a hyperparameter, a distance/cost function affects a classification model's accuracy by this project.

Overall, the accuracy is improved by the optimized hyperparameters (K in KNN and depth in Decision Tree). Whereas the change in serval distance/cost functions does not show a significant difference, at least in our dataset. The result found that the KNN classification shows a higher accuracy on the Diabetic Retinopathy Debrecen dataset. Whereas training on the Decision Tree algorithm achieves a better prediction for the Hepatitis dataset. Since our initial data is a result of random selection, we can not tell which classification algorithm improved performance on average.

Reviewing the entire project, we only have two datasets, with a limited amount of data points. A larger dataset may be required for generalizing and comparing the behaviour of the two classification models, especially for the dataset related to health and diseases. Then in the data preprocessing part, we tried to figure out where the key feature is by calculating the correlations (Appendix Table I, Table II). However, no feature shows a strong association with the label (>50%) [5]. Therefore, we test and delete some features only at the end. During the experiment, the hyperparameters that improve the accuracy of models were found, but we implemented the algorithms in a hard-code way, which brings some inefficiency. A better program design pattern may be considered in future experiments.

After that, some possible direction for future investigation could be how to utilize the correlation between features and the label to reduce the dimensionality of the dataset, which may improve the accuracy of the classification model. This may be realized by Principal Component Analysis (PCA)[6]. From another approach, finding some other model that fits the high-dimension dataset may also be considered[7].

In conclusion, applying the validation algorithm of hyperparameter on both KNN and Decision Tress can increase the accuracy of their classification, yet there are still other approaches that can optimize the accuracy of classification.

Contributions

The workload was distributed equally among the group members.
Yao: data preprocessing, implement “cross-validation” for value k in KNN and Decision Tree depth, graph generation, creativity.
Chuyang: graph generation, data generation, write-up, creativity.
Hanzhi: data preprocessing, write-up, creativity.

Citation

[1] G. Gong, “Hepatitis Data Set,” *UCI Machine Learning Repository: Hepatitis Data Set*, 01-Nov-1988. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Hepatitis. [Accessed: 07-Feb-2022].

[2] B. Antal and A. Hajdu, “Diabetic Retinopathy Debrecen Data Set Data Set,” *UCI Machine Learning Repository: Diabetic retinopathy debrecen data set data set*, 03-Nov-2014. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set. [Accessed: 07-Feb-2022].

[3] S. Humphreys, “Healthcare datasets: Ethical concerns,” *The British journal of general practice : the journal of the Royal College of General Practitioners*, Jun-2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3662435/. [Accessed: 07-Feb-2022].

[4] C. Pazzanese, “Ethical concerns mount as AI takes bigger decision-making role,” *Harvard Gazette*, 04-Dec-2020. [Online]. Available: https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/. [Accessed: 07-Feb-2022].

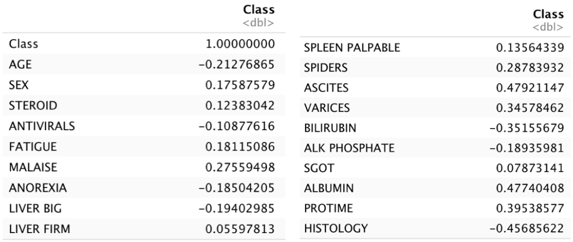
[5] M. Kozak, “What is Strong Correlation?” *Teaching Statistics*, 12-Aug-2009. [Online]. Available:https://doi.org/10.1111/j.1467-9639.2009.00387.x. [Accessed: 08-Feb-2022].

[6] J. Zheng and C. Rakovski, *Data Science Journal*, 18-Aug-2021. [Online]. Available: https://datascience.codata.org/articles/10.5334/dsj-2021-026/. [Accessed: 08-Feb-2022].

[7] V. Pappu and P. M. Pardalos, “High-dimensional data classification - springer,” *SpringerLink*, 2014. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4939-0742-7_8. [Accessed: 08-Feb-2022].

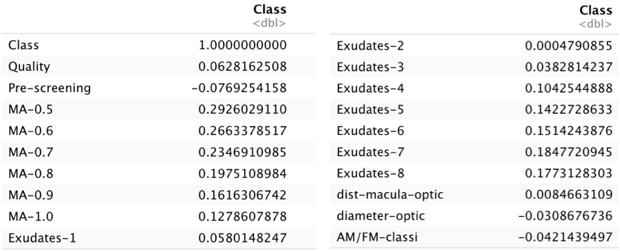
Appendix:

Table I: Correlation between label and features in the Hepatitis Dataset



	Class
	<dbl>
Class	1.00000000
AGE	−0.21276865
SEX	0.17587579
STEROID	0.12383042
ANTIVIRALS	−0.10877616
FATIGUE	0.18115086
MALAISE	0.27559498
ANOREXIA	−0.18504205
LIVER BIG	−0.19402985
LIVER FIRM	0.05597813
SPLEEN PALPABLE	0.13564339
SPIDERS	0.28783932
ASCITES	0.47921147
VARICES	0.34578462
BILIRUBIN	−0.35155679
ALK PHOSPHATE	−0.18935981
SGOT	0.07873141
ALBUMIN	0.47740408
PROTIME	0.39538577
HISTOLOGY	−0.45685622

Table II: Correlation between label and features in the Diabetic Retinopathy Debrecen Dataset



	Class
	<dbl>
Class	1.0000000000
Quality	0.0628162508
Pre−screening	−0.0769254158
MA−0.5	0.2926029110
MA−0.6	0.2663378517
MA−0.7	0.2346910985
MA−0.8	0.1975108984
MA−0.9	0.1616306742
MA−1.0	0.1278607878
Exudates−1	0.0580148247
Exudates−2	0.0004790855
Exudates−3	0.0382814237
Exudates−4	0.1042544888
Exudates−5	0.1422728633
Exudates−6	0.1514243876
Exudates−7	0.1847720945
Exudates−8	0.1773128303
dist−macula−optic	0.0084663109
diameter−optic	−0.0308676736
AM/FM−classi	−0.0421439497