COMP 551 - WINTER 2022
APPLIED MACHINE LEARNING

MiniProject 1 Report

Group 70

George Qiao	260779462
Atia Islam	260737946
Vraj Patel	261022581

School of Computer Science, McGill University

February 9, 2022

Abstract

In this project we implement k-nearest-neighbor (KNN) and decision tree to learn and classify on the Hepatitis dataset [1] and Diabetic Retinopathy Debrecen dataset [2]. Feature selection, cross-validation, hyperparameter tuning and model selection techniques were involved. We found that KNN overall provides higher accuracies than Decision Trees (Hepatitis: 85% to 75%; Diabetes: 68.6% to 62.4%). We also found that different distance/cost functions introduce little change to the accuracies. This project also reinforces that the quantity and quality of data has significant impact on results of machine learning models.

Introduction

The goal of this project is to implement and evaluate the performance of two different classifiers, namely k-nearest-neighbor (KNN) and decision tree,on two different datasets. Throughout the project we explored feature selection, hyperparameter tuning and model selection among other concepts. The two datasets are, respectively, Hepatitis dataset [1] and Diabetic Retinopathy Debrecen dataset [2]. The tasks for this project include importing and cleaning up the data, analyzing the data, selecting key features, implementing models, tuning hyperparameters through cross-validation, reporting test accuracy and plotting decision boundaries. The two datasets are also compared to each other in terms of model performance. The effects of hyperparameters (K for KNN and max tree depth for decision tree) have been analyzed and different distance functions (for KNN) and cost functions (for decision tree) have been experimented with the data sets.

Datasets

For both the data sets we use the same approach to process the data, we first load the files with the help of Panda data frame. We proceed to remove the instances which have missing values. The first important discovery is that the datasets contain both discrete (binary) as well as continuous data, so in order to perform same operations (distance function, etc.) on them, we convert discrete values into 0's and 1's, make sure they are all of the same type (float), and also normalize all continuous values to same scale so that large continuous values would not overwhelm the binary features. In order to understand the class distribution better, we need to find the important parameters which have the closest relationship with the label and help us get the better accuracy. So, we created T-tables (which contains basic statistics like mean, min, max and interquartile range), pairwise plots of features and the correlation matrix. T-tables give us a great initial idea of the key basic stats of dataset. The pairwise plots show direct pair-to-pair relationship between the features (with the labels color-coded) and also have histograms of each feature on the diagonal. We decided to choose the two best features of each dataset to perform experiments.

Results

The first step of running the experiments is the train-validation-test split. We first split off the test set to leave until the end for the estimation of generalization error. As for the rest of the data, we divide them into equal portions and perform L-fold cross-validation on them. Hyperparameter values are varied while training and running the models, and then we use the average validation error and its variance (uncertainty) to pick the best model. Finally, the results of the experiments are analyzed according to the Task 3 instructions.

Note: Because of our implementation of the cross_correlation function randomizes the shuffle of the data instances, each time the .ipynb file is run different outputs might be shown. The numbers in following paragraphs present one typical case that we have found.

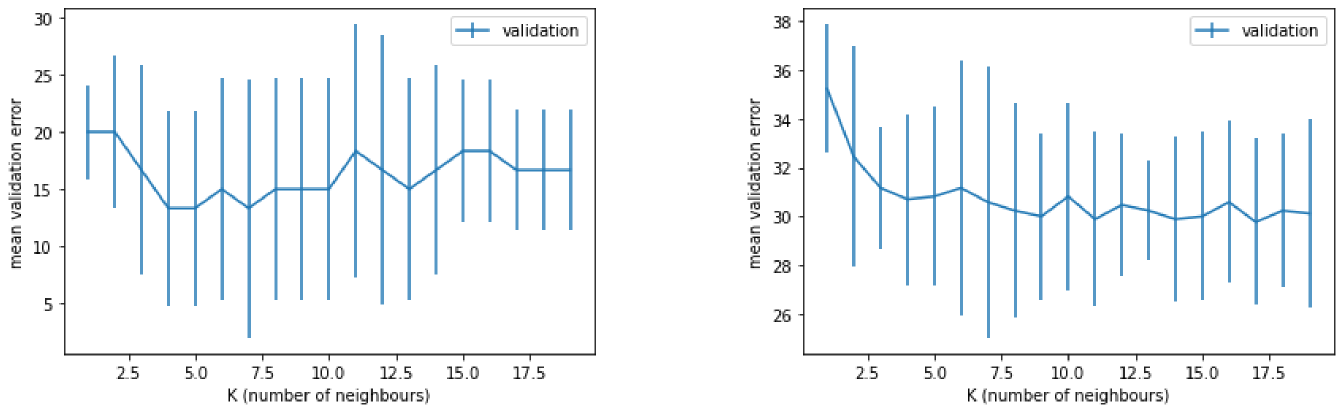
Accuracy of KNN and Decision Tree algorithm on the two datasets

For the Hepatitis dataset, the best K for KNN is determined to be 3, which gives an estimated generalization accuracy of 85%. The best max tree depth for decision tree is determined to be 10, which gives an accuracy of 75%. Note though, that the first dataset's small sample size (80 instances after empty entry deletion) make it difficult to tune hyperparameters very effectively.

For the Diabetes dataset, KNN with K=4 gives an estimated generalization accuracy of 68.6% while decision tree with max_depth=10 gives accuracy of 62.4%.

Effect of different K values

There were no significant effect of using k values on the first data set. But for the second data set as the value of k increases mean square error decreases.



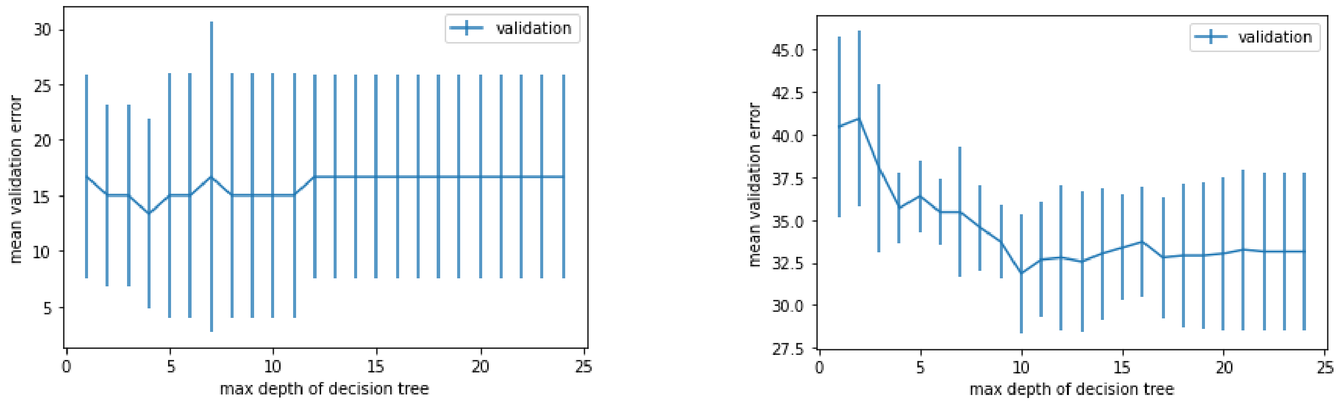
(a) Dataset 1

(b) Dataset 2

Figure 1: KNN Effect of K

Effect of maximum tree depth

There are no effects of tree depth on the results since this data set is very small. The second data set is huge compared to first one and also from the figure we can see the effect of depth on the results.



(a) Dataset 1

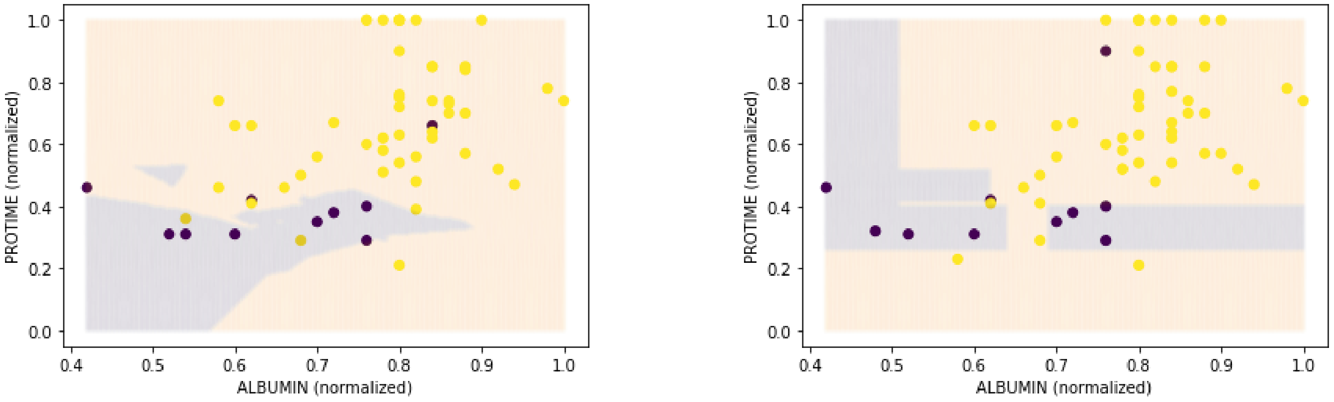
(b) Dataset 2

Figure 2: Decision Tree Effect of Max Depth

Effect of different distance/cost functions

There were no significant effect of using two different distances (Euclidean and Manhattan) on the results, nor did changing the cost function (between misclassification cost and gini index cost). See .ipynb sections ”Change Distance Function” and ”Change Cost Function”, where exact same results were obtained. Due to time constraints the second dataset is not investigated on this. This ”no effect” result once again might be attributed to the small sample size of dataset 1.

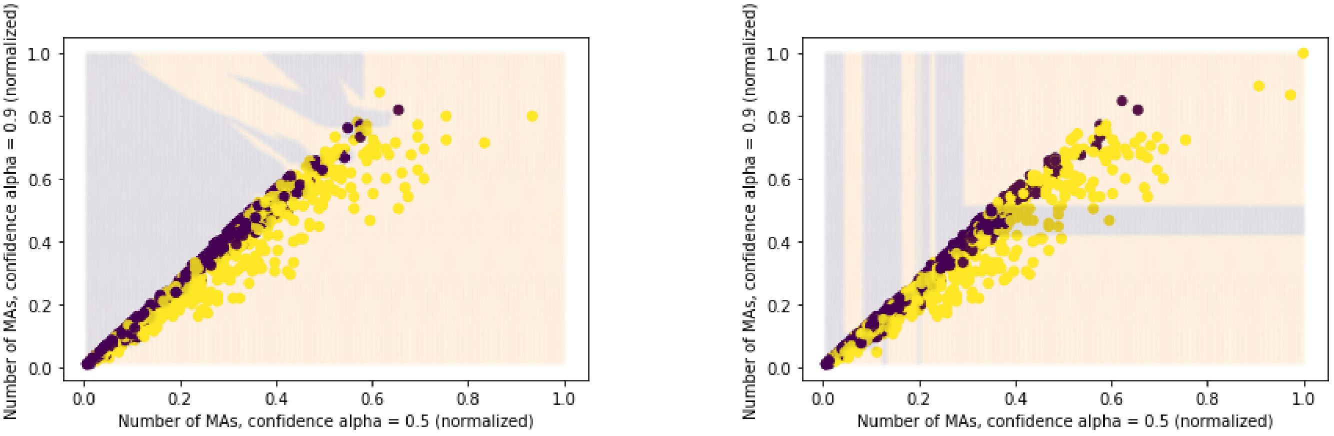
Decision boundary



(a) KNN

(b) Decision Tree

Figure 3: Decision Boundary, Dataset 1



(a) KNN

(b) Decision Tree

Figure 4: Decision Boundary, Dataset 2

Discussion and Conclusion

Key takeaways from the project:

- Data quantity and quality are very important for machine learning algorithm results.
- KNN, due to its lazy learning, are much faster to train than decision tree, but much slower in prediction as the computation of distances scale in complexity with numebr of instances.
- Different distance/cost function much less effect on accuracy than hyperparameters.

Possible directions for future investigation

- Perform the same experiments on other similar but larger datasets and compare results.
- Extend the models to include more or even all features.
- Change the models from classification to regression.
- Investigate other approaches to select the most important features.
- Investigate other approaches to integrate discrete and continuous data in the same model, e.g. in Naive Bayes likelihoods of different features use different distributions.
- Investigate other approaches for model evaluation/selection beyond the rule of thumb, e.g. AUC.

Statement of Contribution

Atiya Islam: Data processing, analyzing from the files .Implementing the KNN for both datasets, project report.

George Qiao: Implementing the decision tree for the both datasets, writing report.

Vraj Patel: Cross validation part, deciding of attributes and finalizing hyperparameters for the model, project report.

References

[1] M. Raymer, T. Doom, L. Kuhn, and W. Punch, “Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm,” *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 33, pp. 802–13, 02 2003.

[2] B. Antal and A. Hajdu, “An ensemble-based system for automatic screening of diabetic retinopathy,” *Knowledge-Based Systems*, vol. 60, 04 2014.