1 Abstract

In this project, we investigated the performance of two classification models, namely k-nearest neighbours and decision trees, with two data sets: Adult Data Set and Breast Cancer Wisconsin (Diagnostic) Data Set. For the first data set, the task is to predict if the income of an adult exceeds $50K/yr based on various factors, such as workclass, education, etc. For the second dataset, the task is to determine if breast cancer is benign or malignant with information including clump thickness, marginal adhesion, normal nucleoli, etc. For both data sets, we found that the k-nearest neighbor classifier generated lower accuracy than decision trees and was significantly slower for predictions.

2 Introduction

Since each machine learning algorithm is biased to perform well on some class of problems, model selection can significantly affect the outcome of our tasks. In this project, we compared the performance of two classification models, k-nearest neighbors and decision trees with Adult Data Set and Breast Cancer Wisconsin (Diagnostic) Data Set, based on the hyperparameters of each model as well as the affections of the amount of data inputs, with the implementation of 5-fold cross-validation and gridsearch techniques.

3 Datasets

3.1 Adult dataset

This dataset contains 32,560 training samples and 16,280 testing samples. We dropped instances that had missing features during preprocessing, and the size of the training set was reduced to 30,161. ‘Education’ and ‘education-num’ are closely correlated, so we only kept ‘education’ as one of the features. Since the values of ‘capital-gain’ and ‘capital-loss’ are 0’s for 92.2% and 95.6% of instances, respectively, these two features were also dropped. Eight categorical features (‘Husband,’ ‘Not-in-family’, etc) were preprocessed with one-hot encoding. Two continuous features, namely ‘age’ and ‘hours-per-week’, were preprocessed using min-max scaling. We labeled the salary “<= 50K” as 0 and “> 50K” as 1.

3.2 Breast Cancer Wisconsin dataset (abbr. BCW)

Link: *https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29*
This dataset contains 699 samples. We dropped the instances with missing values and split the remaining 683 samples into 580 training and 103 testing samples. Since all features are numerical and in the range of 1 to 10, one-hot encoding is not needed to transform them. Therefore, we labeled ‘benign’ as 0 and ‘malignant’ as 1. Besides, KNN does not work well for data sets with large features due to the curse of dimensionality.

4 Results

4.1 Comparing performances between KNN and decision trees

KNN has average scores of 82.96% and 66.60% for the adult dataset and BCW dataset, respectively. Decision trees have average scores of 85.27% and 94.66% for the adult dataset and BCW dataset, respectively. We found that Decision trees overall achieved higher accuracy and were significantly faster to train. This is likely because KNN is cursed for its expensive real-time execution, whereas the Decision tree can automate feature interaction in the training section. Moreover, KNN is more sensitive to feature scaling since it detects and sorts its neighbors by the distances that calculate directly from scaled vectors. DicisonTree, on the other hand, handles collinearity by continuously generates the most important features; it also cares less about the level of irregularity of the distribution of the data.

4.2 Hyperparameters vs Performances

4.2.1 Decision Trees

We adjusted four hyperparameters of decision trees, namely max depth (maximum allowed depth for the tree, main feature that subjective to under and overfitting), min sample split, and min samples leaf (trying 10 different values for each), and criterion (with either ‘gini’ or ‘entropy’ as the cost function).

For adult dataset, the accuracy decreases drastically as max depth increases from 11 to 21, while the maximum and variance remain almost constant as max depth keeps increasing (Appendix A Figure 1). Min samples split does not have a huge impact on accuracy. The model with the highest accuracy (85.27%) has a max depth equal to 11, min sample split equal to 42, min samples leaf equal to 11, and entropy as the criterion. We applied the best model on the test set and got an accuracy of 0.857.

For BCW dataset, we first set the range of the max depth to 1-101, but f1-score remained almost constant as the max depth increases. We performed another grid search, choosing a simpler model instead, with the max depth between 2 and 12 (Appendix A Figure 2). The highest f1-score is 0.95, given by the model with depth equal to 4, min sample split equal to 2, min samples leaf equal to 1, and entropy as the criterion. We applied the best model on the test set and got an f1-score of 0.96.

Table 1: Decison trees with the highest validation accuracy for two datasets

Dataset	max depth	min sample split	min samples leaf	criterion	testing result
Adult	11	42	11	entropy	0.857 (accuracy)
BCW	4	2	1	entropy	0.960 (f1-score)

4.2.2 KNN

We changed two hyperparameters of KNN, namely the number of neighbors (from 2 to 50) and the weighted method (either uniform or by distance).

For adult dataset, in general, the accuracy increases as the number of neighbors increases (Appendix A Figure 3). The model with the highest accuracy (82.96%) has 29 neighbors and a uniform weight method. We applied the best models on the test set and got an accuracy of 0.835.

For BCW dataset, there is no general pattern for the correlation between the number of neighbors and the f1-score (Appendix A Figure 4). The model with the highest f1-score (66.60%) has 17 neighbors and a uniform weight method. We applied the best models on the testset and got f1-score of 0.59.

Table 2: KNNs with the highest validation accuracy for two datasets

Dataset	number of neighbors	weight method	testing result
Adult	29	uniform	0.835 (accuracy)
BCW	17	uniform	0.590 (f1-score)

4.3 Performances vs Amount of data

We used the difference amount (10%, 20%,...,100%) of data for both KNN and decision trees and explored how it affects the accuracy.
It turns out that for both frameworks, the accuracy goes up when the amount of data goes from 0% to 20% roughly. Between 20% and 40%, the increasing rate slows down. After 40%, the graph has reached a plateau (Figure 5 and 6).

5 Discussion and Conclusion

In this project, we explored the size of the dataset, preprocessing of data, the effect of hyperparameters, and performed model selection between k-nearest neighbors and decision trees.

We wonder how different choices of hyperparameters for the model tuning would impact the accuracy. In both of our examples, decision trees had a relatively better performance. For future investigation, we may apply these two models on other datasets to figure out if this is the case in general.

6 Contributions

Dailun Li: coding the GridSearch method and the 5-fold cross validation. Generates data output. Small fixes on the report.
Xinyi Zhu: datasets preprocessing, sample data testing, report writing, results visualization.
Yuyan chen: dataset preprocessing, code for BCW data set and result visualization, and report writing

Appendix A Figures

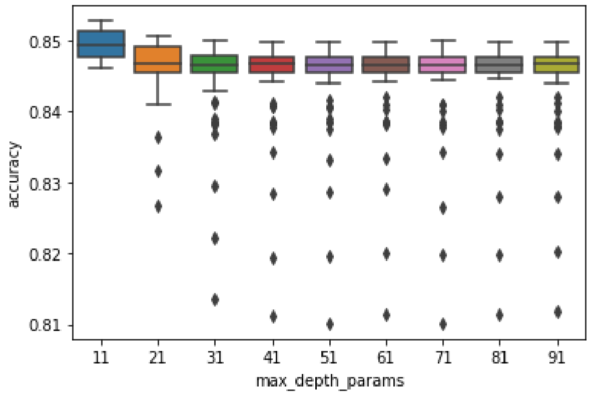

Figure 1: Accuracy of decision trees with different max depth (adult)

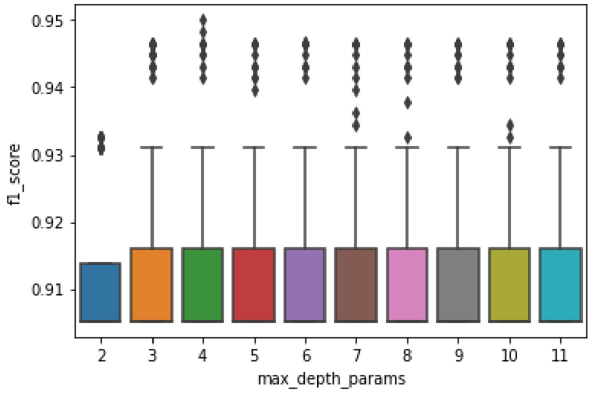

Figure 2: f1 score of decision trees with different max depth (BCW)

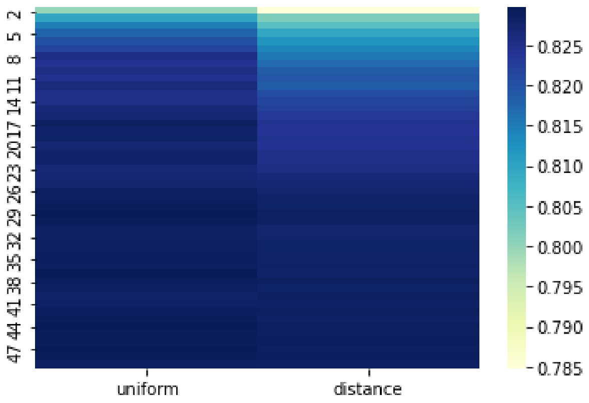

Figure 3: accuracy of KNN for adult

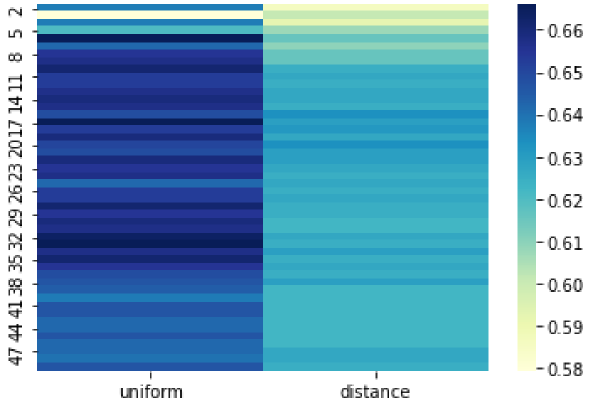

Figure 4: accuracy of KNN for BCW

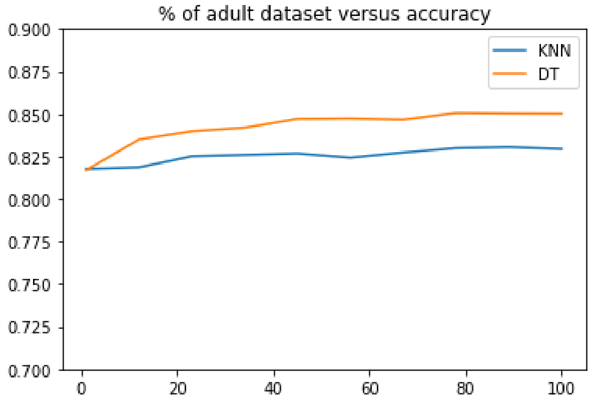

Figure 5: percentage of data vs accuracy for adult

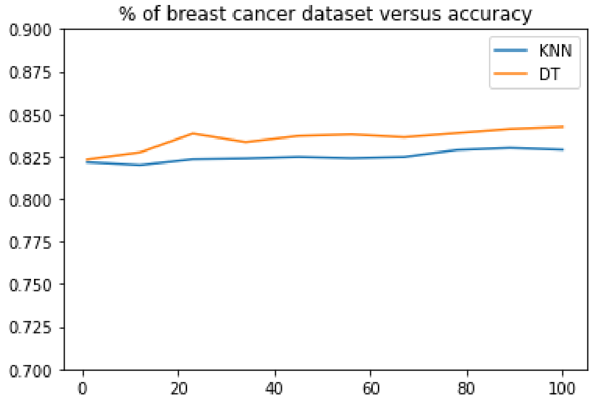

Figure 6: percentage of data vs accuracy for BCW