KNN vs. Decision Tree

Yu Yun Liu; Ruoyu Deng; Pengyu Xue

February 9, 2022

Abstract

In this project, we investigated the performance of two machine learning models, K-Nearest Neighbour and Decision Trees. We implemented these two machine learning models from scratch and used two different datasets to compare/measure the performance of each model. Python works as the principal tool for executing these two models for this project; with the support of the assigned libraries in such as Matplotlib, NumPy and Scipy, it clearly shows the difference of accuracy between the two models, which helps us to determine the appropriate one to analyze to data sets. By testing different K values and maximum depth on the implanted models, we record the performances of the K-Nearest Neighbour and Decision Tree model in the table and observe the difference. We found that the K-Nearest Neighbour model results with better accuracy than the Decision Tree in both data sets with appropriate k-values by observing the data we got.

1 Introduction

The comparison is made by implementing two machine learning models, K-Nearest Neighbour and Decision Trees. The basic theory behind K-Nearest Neighbour is to explore the data points neighbour with different assigned k-value, whereas a Decision Tree is a tree-based algorithm derived with nodes with a sequence of conditions. We used two data sets, the Hepatitis dataset and Diabetic Retinopathy Debrecen dataset, to train the two models. Furthermore, we trained the model using different distance functions: Manhattan and Euclidean distance and different k-values, the number of nearest neighbours used in the training process. We thus observed that different accuracy occurs under different k-values and distance functions. For Decision Trees, we trained the model using different maximum tree depths and three cost functions: misclassification cost, entropy cost, and gini index cost. We picked the best values of hyper-parameters for both models in the training process. We then tested the performance of our models on both datasets and noted the results of the predictions. The predictions obtained from testing the models show that the K-Nearest Neighbour model with the appropriate k-value results in better accuracy than the Decision Tree for both data sets.

2 Datasets

We are given two datasets, the Hepatitis dataset and Diabetic Retinopathy Debrecen dataset (DR dataset). Both datasets have a total of 19 features describing symptoms of the associated illness. The first one describes 80 cases of whether a person with different combinations of the 19 features has or does not have hepatitis. The latter describes 1151 cases of different combinations of the 19 other features associating whether or not an image contains signs of diabetic retinopathy or not. The Hepatitis dataset was a .csv file whereas the DR dataset was originally a .arff file. We then loaded the datasets into pandas dataframes which allowed better manipulation and comprehension of the datasets. We then removed all examples (rows) with missing data, represented with a "?" in the dataframes. Furthermore, we noticed that some of the data in the Hepatitis dataset were strings instead of numbers. We then proceeded to transform these data into floats for all columns. To better understand the datasets, we used the describe method from the pandas library to get basic statistics on the features of both datasets. Moreover, we plotted the distribution of the positive vs. negative classes distribution of both datasets. For the Hepatitis dataset, we observed a highly larger amount of negative cases (die) than positive ones (live).

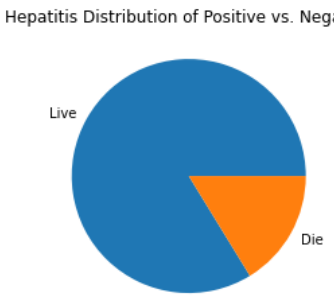

Figure 1: Sector Diagram of Hepatitis Distribution of Positive vs. Negative

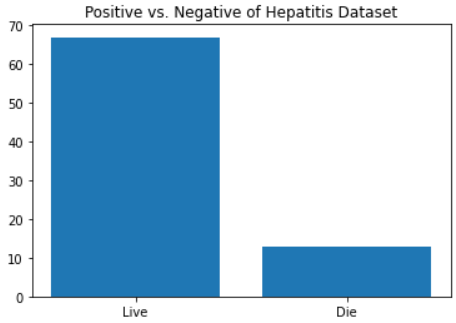

Figure 2: Bar Chart of Hepatitis Distribution of Positive vs. Negative

For the DR dataset, the data was more evenly distributed.

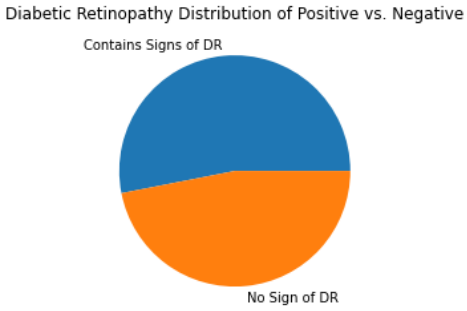

Figure 3: Sector Diagram of Diabetic Retinopathy Distribution of Positive vs. Negative

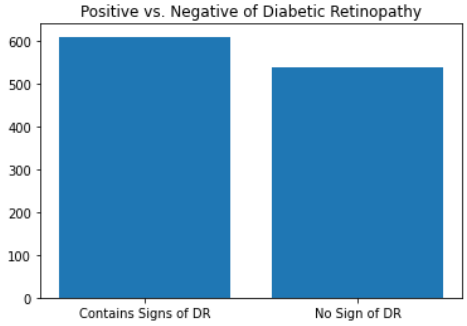

Figure 4: Bar Chart of Diabetic Retinopathy Distribution of Positive vs. Negative

Moreover, we computed the distributions of the numerical features. For the Hepatitis dataset, we noticed that data with label=live did not cover all the possible ages available. It is the same case with the feature Protime.We also did not have data for which the label=live for the features Anorexia and Liver Big. For the other features, there is less data with class label=live than die. This is mostly due to the unevenness of the distribution of the classes.

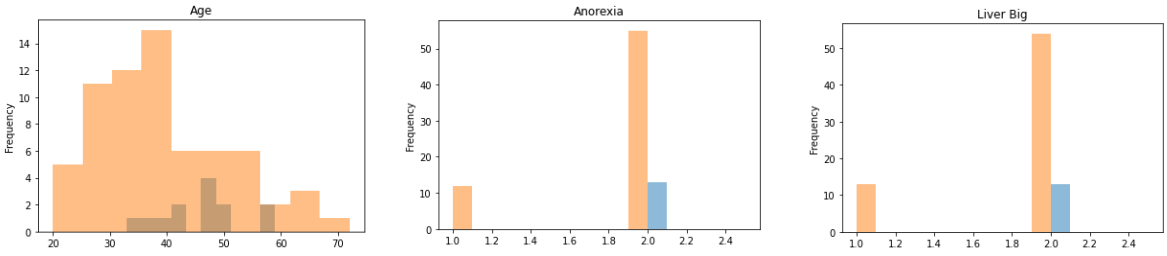

Figure 5: Bar Chart of . (a) Age; (b)Anorexia; (c) Liver Big.

For the DR dataset, most of the features were evenly distributed between the classes. However, the class with label=no signs of DR has no datapoint for features Exudates 0.7, 0.8, 0.9, and 1.0.

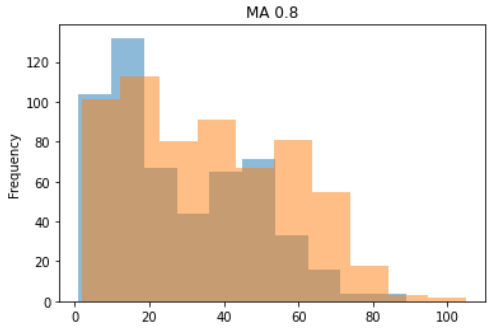

Figure 6: Bar Chart of MA 0.8

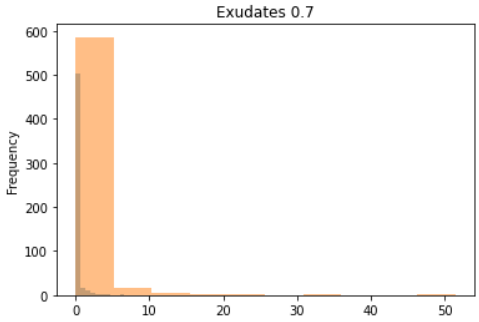

Figure 7: Bar Chart of Exudates 0.7

3 Results

We start by modifying hyper-parameters of KNN model and DT model to observe their behaviours on each dataset. Using the euclidean function from K-Nearest Neighbour and entropy cost in Decision Tree, we observed the highest test accuracy of the model performed on the two datasets.

	Test Accuracy	Train Accuracy	Accuracy Diff
K = 7	0.958333	0.839286	0.119048

	Test Accuracy	Train Accuracy	Accuracy Diff
K = 7	0.669565	0.743176	0.073611

	Test Accuracy	Train Accuracy	Accuracy Diff
Max_depth = 2	0.958333	0.928571	0.029762

	Test Accuracy	Train Accuracy	Accuracy Diff
Max_depth = 25	0.660870	0.995037	0.334168

Figure 8: The best Accuracy of KNN on Euclidean Function and DT on Entropy Function on Two Data Set

In the plots above, we have trained our K-Nearest Neighbour model with euclidean function and Decision Tree model with entropy cost function. We can observe that for the Hepatitis dataset, the K-Nearest Neighbour model can get higher test accuracy predictions with an appropriate choice of k-value than the Decision Tree model. On the other hand, for the DR dataset, K-Nearest Neighbour model results with higher test accuracy than the Decision Tree model.
In addition, we also found that the K value dramatically impacts the accuracy. As K increases to a certain threshold, the K-Nearest Neighbour can obtain the best prediction on its test accuracy. The k-value does not necessarily give the smallest gap between test accuracy and train accuracy. In addition, as k-value increments, train accuracy gradually decreases. We list out the diagrams for K-Nearest Neighbour models. With the most appropriate k-value in each function, they provide the best test accuracy in each dataset.

	Test Accuracy	Train Accuracy	Accuracy Diff
K = 7	0.958333	0.839286	0.119048
K = 9	0.958333	0.767857	0.190476
K = 10	0.958333	0.767857	0.190476
K = 11	0.958333	0.767857	0.190476
K = 12	0.958333	0.767857	0.190476
K = 13	0.958333	0.767857	0.190476
K = 14	0.958333	0.767857	0.190476
K = 8	0.916667	0.857143	0.059524
K = 3	0.833333	0.910714	0.077381
K = 5	0.833333	0.839286	0.005952
K = 6	0.833333	0.821429	0.011905
K = 1	0.750000	1.000000	0.250000
K = 4	0.750000	0.857143	0.107143
K = 2	0.708333	0.928571	0.220238

Figure 9: Performance of KNN Euclidean Function on Hepatitis Data Set

	Test Accuracy	Train Accuracy	Accuracy Diff
K = 7	0.669565	0.743176	0.073611
K = 6	0.657971	0.754342	0.096371
K = 8	0.657971	0.730769	0.072798
K = 14	0.655072	0.688586	0.033513
K = 9	0.649275	0.719603	0.070328
K = 12	0.649275	0.700993	0.051717
K = 13	0.646377	0.704715	0.058338
K = 10	0.643478	0.705955	0.062477
K = 11	0.637681	0.702233	0.064552
K = 5	0.631884	0.764268	0.132384
K = 1	0.628986	1.000000	0.371014
K = 4	0.626087	0.777916	0.151829
K = 3	0.620290	0.827543	0.207254
K = 2	0.594203	0.805211	0.211008

Figure 10: Performance of KNN Euclidean Function on Diabetic Retinopathy Debrecen Data Set

Different from the k-value, maximum tree depth barely affects the performance of the Decision Tree model when it is large enough. When we start the max depth with one, there is a change of test accuracy rapidly. However, it stabilizes after reaching a certain depth. In addition, the train accuracy will fix at 100% after max depth meets a threshold, which implies that the model is overfitting the training dataset. The diagrams of Decision Tree models show down below. We selected the appropriate depth from each function. The depth can not be one, since it's unpersuasive. It is chosen by the best test accuracy with the appropriate accuracy difference.

	Test Accuracy	Train Accuracy	Accuracy Diff
Max_depth = 2	0.958333	0.928571	0.029762
Max_depth = 3	0.958333	0.964286	0.005952
Max_depth = 4	0.916667	1.000000	0.083333
Max_depth = 5	0.916667	1.000000	0.083333
Max_depth = 6	0.916667	1.000000	0.083333
Max_depth = 7	0.916667	1.000000	0.083333
Max_depth = 8	0.916667	1.000000	0.083333
Max_depth = 9	0.916667	1.000000	0.083333
Max_depth = 1	0.875000	0.821429	0.053571

Figure 11: Performance of DT Entropy Cost on Hepatitis Data Set

	Test Accuracy	Train Accuracy	Accuracy Diff
Max_depth = 2	0.958333	0.875000	0.083333
Max_depth = 5	0.958333	0.982143	0.023810
Max_depth = 6	0.958333	0.982143	0.023810
Max_depth = 1	0.916667	0.857143	0.059524
Max_depth = 3	0.916667	0.892857	0.023810
Max_depth = 4	0.916667	0.946429	0.029762
Max_depth = 7	0.916667	1.000000	0.083333
Max_depth = 8	0.916667	1.000000	0.083333
Max_depth = 9	0.916667	1.000000	0.083333

Figure 12: Performance of DT Entropy Cost on Diabetic Retinopathy Debrecen Data Set

Furthermore, for Hepatitis, the K-Nearest Neighbour model has two distance functions that make roughly the exact predictions. However, for Messidor features dataset, the Manhattan function predicts better than the Euclidean function in K-Nearest Neighbour. The cost function Entropy generates the most precise prediction among all cost functions in either Hepatitis or Messidor features datasets.

	Test Accuracy	Train Accuracy	Accuracy Diff
Max_depth = 25	0.660870	0.995037	0.334168
Max_depth = 30	0.657971	1.000000	0.342029
Max_depth = 20	0.652174	0.967742	0.315568

	Test Accuracy	Train Accuracy	Accuracy Diff
Max_depth = 20	0.594203	0.933002	0.338800
Max_depth = 25	0.585507	0.970223	0.384716
Max_depth = 30	0.585507	0.998759	0.413252

	Test Accuracy	Train Accuracy	Accuracy Diff
Max_depth = 20	0.646377	0.996278	0.349901
Max_depth = 25	0.646377	1.000000	0.353623
Max_depth = 30	0.646377	1.000000	0.353623

	Test Accuracy	Train Accuracy	Accuracy Diff
K = 7	0.669565	0.743176	0.073611
K = 6	0.657971	0.754342	0.096371
K = 8	0.657971	0.730769	0.072798
K = 14	0.655072	0.688586	0.033513
K = 9	0.649275	0.719603	0.070328
K = 12	0.649275	0.700993	0.051717
K = 13	0.646377	0.704715	0.058338
K = 10	0.643478	0.705955	0.062477
K = 11	0.637681	0.702233	0.064552
K = 5	0.631884	0.764268	0.132384
K = 1	0.628986	1.000000	0.371014
K = 4	0.626087	0.777916	0.151829
K = 3	0.620290	0.827543	0.207254
K = 2	0.594203	0.805211	0.211008

	Test Accuracy	Train Accuracy	Accuracy Diff
K = 9	0.692754	0.727047	0.034294
K = 8	0.689855	0.733251	0.043396
K = 10	0.678261	0.719603	0.041342
K = 7	0.675362	0.758065	0.082702
K = 14	0.672464	0.699752	0.027288
K = 13	0.669565	0.715881	0.046316
K = 6	0.666667	0.761787	0.095120
K = 12	0.666667	0.715881	0.049214
K = 11	0.663768	0.719603	0.055835
K = 4	0.646377	0.758065	0.111688
K = 3	0.643478	0.827543	0.184065
K = 5	0.631884	0.785360	0.153476
K = 2	0.628986	0.794045	0.165059
K = 1	0.623188	1.000000	0.376812

Figure 13: From the Top Left to Top Right: DT Cost Entropy, Cost Misclassification, Cost and Gini index on Diabetic Retinopathy Debrecen Data Set. From Bottom Left to Right: KNN Euclidean Function, Manhattan function on Diabetic Retinopathy Debrecen Data Set.

In order to plot the decision boundary graph, we need to first adjust the key features we would like to plot. We choose the two features with highest correlations with Class of each dataset. In detail, the most 2 important key features for Hepatitis dataset are Ascites and Albumin, and the most 2 important key features for DR dataset are MA 0.5 and MA 0.6.
We can then plot the decision boundary with these key features:

	Class
Ascites	0.479211
Albumin	0.477404

	Class
MA 0.5	0.292603
MA 0.6	0.266338

Figure 14: Top Table relate to Hepatitis Data Set and Bottom Table relate to Diabetic Retinopathy Debrecen Data Set

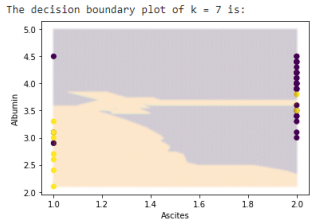

Figure 15: Modeling Hepatitis Data Set Using Euclidean Function with K = 7

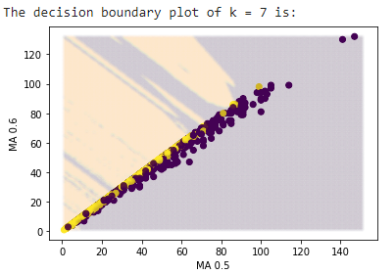

Figure 16: Modeling Diabetic Retinopathy Debrecen Data Set Using Euclidean Function with K = 7

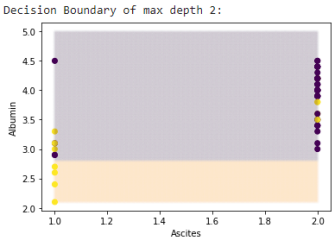

Figure 17: Modeling Hepatitis Data Set Using Cost Entropy with Depth of 2

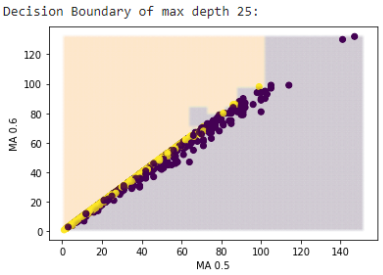

Figure 18: Modeling Diabetic Retinopathy Debrecen Data Set Using Cost Entropy with Depth of 2

4 Discussion and Conclusion

K-Nearest Neighbour generally performs better than Decision Tree but it takes extra work to find which hyper-parameter gives the best prediction performance, whereas Decision Tree performs not as good as K-Nearest Neighbour, but we can choose the hyper-parameter relatively easier. On the other hand, when handling large size of dataset, Decision Tree does not perform as fast as K-Nearest Neighbour does. Such problem can be improved through changing the initial test for nodes in Decision Tree. In this project, we calculated the possible threshold for a test using the mean between two values of the same feature on a sorted list of the given dataset However, it is possible to optimize these thresholds such that we can run the DT algorithm faster.

5 Statement of Contributions

Everyone in the group participated in working on this mini-project. We collaborated and helped out on each other. However, everyone has their primary focus. Yu Yun analyzed the dataset and implemented the Decision Tree model; Ruoyu implemented the K-Nearest Neighbour model and handled the experiment running; Pengyu focused on the report with Latex and analyzed the project's procession and conclusion.

6 Reference

Varghese, D. (2019, May 10). Comparative study on classic machine learning algorithms. Medium. Retrieved February 9, 2022, from https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222