COMP 551 Assignment 1

Group 86: Akos Borbath, Aryan Chaturvedi, Filicia Hang

February 1, 2024

1. Abstract

For our first assignment, we were tasked with implementing two common supervised learning models and test their accuracies on 2 different datasets. Through the first dataset, we understood the tasks as building a model which could successfully classify whether a datapoint, i.e. a person was a “Senior” or an “Adult” based on several continuous factors. Our second dataset led us to another task of classifying whether a data point, i.e. a Tumour, was in “Malignant” or “Benign” for predicting Breast Cancer. We implemented a KNN model and a Decision Tree model and ran experiments on different associated distance/cost functions. We found that these distance/cost functions had an impact on the accuracy of the model.

2. Introduction

In this assignment we worked with two data sets: the “2013-2014 National Health and Nutrition Health Survey Age Prediction Subset” and “Breast Cancer Wisconsin”. The first one contains various health metrics as features, which are then used to guess if a person is an “Adult” or a “Senior”. The second one contains various DNA related features which are used to predict if the tumor is “Benign” or “Malignant”. We implemented both Decision Tree and KNN models for both datasets. Specifically, we experimented with Euclidean Distance and Manhattan Distance as the better “Similarity Measure” for the KNN Model and found that both functions lead to a similar level of accuracy, however Euclidean Distance led to much lower optimal K Hyperparameter, which reduced prediction times of the model. Similarly, we also experimented with Misclassification Rate, and concluded that our Decision Tree Model was slightly less accurate than if we used an Entropy or Gini Index to fit the data. We also found that our max depth of 5 was an ideal depth, beyond which our models were prone to overfitting into the training data.

3. Methods

We implemented 2 supervised learning models consisting of an instance-based classifier called KNN, as well as a tree-based Classifier, called a Decision Tree Model.

The KNN Model, as the name suggests, stores each training input data point in a multidimensional array, and maps to its true class label, essentially memorizing the values of each input data point, to use as a reference later when predicting a new data point. This is based on the idea of finding neighbouring data points within the trained dataset that are most similar to the new data point, which is to be predicted. We defined similarity based on the Euclidean distance, and the Manhattan distance between two data points. Through our validation process, we would then select the K hyper-parameter, to set a limit based on the K nearest neighbouring data points to our new datapoint, and then mapping our new data point to the model label, i.e. among K of the closest data points, the most mapped “output” class would then be the predicted “class"”of the new datapoint.

In contrast, the decision tree model tries to understand the distribution of the training data points by distinguishing them by different thresholds in features of each class and dividing the data points recursively in the form of a binary tree, with the sole aim of maximizing the probability of obtaining only a singular “label” in each leaf node or distinct region. The features and thresholds selected to partition data into subsets is determined through a greedy heuristic search at each split node and thereby explores the entire Combinatorial Tree space to select the optimal distinguishing features and thresholds.

4. Data Sets

In this assignment we worked with two data sets: the “2013-2014 National Health and Nutrition Health Survey Age Prediction Subset” and “Breast Cancer Wisconsin”. The first one contains various health metrics as features, which are then used to guess if a person is an “Adult” or a “Senior”. The second one contains various DNA related features which are used to predict if the tumor is “Benign” or “Malignant”.

For both datasets we looked at the squared difference in feature means between classes. In the first data set, the bottom four have an almost trivial difference. We then expect these to not play a significant role in the prediction algorithm, so we disregarded them for the rest of the assignment.

The second one’s features have a smaller disparity between themselves, so we kept them all, except “Mitosis”, which was discarded because it had the lowest squared mean difference. However, 16 data points had a null “Bare Nuclei”, so those were discarded as well.

5. Results

Question 1:

As seen from Figure 7 in Question 5, the AUROC for KNN was slightly higher (0.58) than for DT (0.5) on the first dataset. This is also seen in the Accuracy of KNN being slightly higher with about 0.8816, whereas DT only had an accuracy of 0.7216.

On the other hand, Figure 8 shows the 2 models performed identically with an AUROC of (0.95) for both models on the second dataset. This is inline with the high prediction accuracy of both models where the KNN model is only marginally better with 0.95, in contrast to 0.94 for the DT. This marginal difference is also bridged if DT was using Entropy or Gini Index.

Question 2:

We carried out two experiments per data set using two different distance functions. For more optimal results, we split the data sets for KNN like so: 10% of the data points for the testing subset, 18% for the validation subset and the rest (72%) was used for training. We used a range from 1 to 10 to look for the best K value per experiment. The best K picked was the one with the highest validation accuracy.

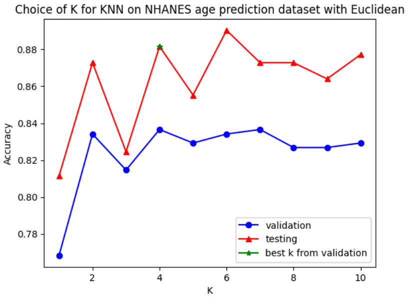

Figure 1.

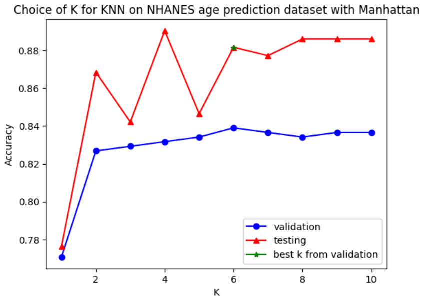

Figure 2.

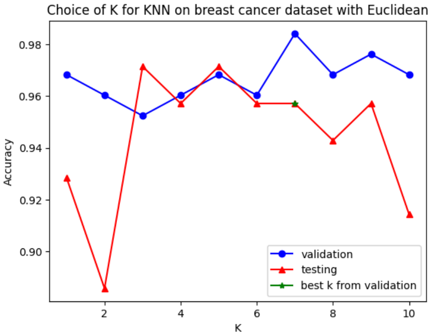

Figure 3.

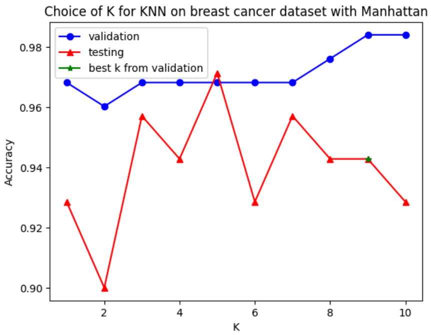

Figure 4.

Figures 1 and 2 show the results for the experiments done on the age prediction data set, while Figures 3 and 4 show the results for the breast cancer data set. Evidently, the best K varied depending on the distance function and on the data sets. Interestingly, the validation accuracy plots in Figures 1 and 2 are lower than the testing accuracy plots. Usually, the contrary is more observed like in Figures 3 and 4.

Question 3:

For testing the effect of maximum tree depth on performance, which was measured by the accuracy of the prediction, we used the same cost function, misclassification cost, and we tried maximum tree depths of 1, 5, 10, 15, and 20. The results can be seen in Table 1.

Table 1: Maximum Tree Depth and Corresponding Accuracies

Depth	Accuracy on Age Prediction Data Set	Accuracy on Breast Cancer Data Set
1	0.8358	0.9257
5	0.8358	0.9514
10	0.8288	0.94
15	0.8103	0.94
20	0.7893	0.94

We can see from Table 1 that for both data sets, increasing the maximum level after a certain point decreases the accuracy of the prediction. However, for the age prediction data set, the accuracy progressively gets worse, while it stays constant for the other, after 5.

Question 4:

For KNN, we tested the Manhattan and Euclidean distance functions in our experiments. For each of the datasets, two experiments were done to find the best K value using the validation subset while testing a different distance function. The test accuracy based on distance function and K value used results are compiled in the table below.

Table 2: Distance Functions with Best K and Corresponding Accuracies

Age Prediction Data Set

	K value	Accuracy
Euclidean	4	0.8816
Manhattan	6	0.8816

Breast Cancer Data Set

	K value	Accuracy
Euclidean	7	0.9571
Manhattan	9	0.9429

There was surprisingly no difference in test accuracy between the use of the different distance functions for the age prediction data set. However, when testing on the breast cancer data set, the Euclidean distance function came on top with a difference of approximately 0.0142 in accuracy. With both data sets having numerical features, we can understand why Euclidean distance is more widely used for those cases, although the difference was not very significant.

For the Decision Table, 3 cost functions were tested: misclassification, entropy, and gini index. All three were tested with a maximum tree depth of 20. The results can be seen in Table 3.

Table 3: Different Cost Functions and Corresponding Accuracies

	Accuracy on Age Prediction Data Set	Accuracy on Breast Cancer Data Set
Misclassification	0.7226	0.94
Entropy	0.7594	0.9514
Gini Index	0.7550	0.9543

We can see from Table 3 that the cost function does indeed influence performance. However, for different data sets, different cost functions are optimal. As a matter of fact, for the age prediction data set, entropy has the highest accuracy, while for the breast cancer data set, it's the Gini Index that's optimal.

Question 5:

Classification Accuracy of the KNN and DT Models can be seen through the Confusion Matrix on both Datasets:

Confusion Matrix for KNN Model on the NHANES Dataset predicted very low True Positive and False Positive Rates:
TPR = 0.20 and FPR = 0.0345

	Predicted Negative	Predicted Positive
Actual Negative	196	7
Actual Positive	20	5

The same can also be observed from the Confusion Matrix of the Decision Tree Model:
TPR = 0.00536 and FPR = 0.00945

	Predicted Negative	Predicted Positive
Actual Negative	943	9
Actual Positive	186	1

Due to this, we can see that both models had a low True Positive Rate, and high False Positive Rate, and thus the ROC Curve was mostly a diagonal line, or close to a diagonal line on the NHANES dataset, with the AUROC being 0.58 and 0.5 for KNN and DT respectively.

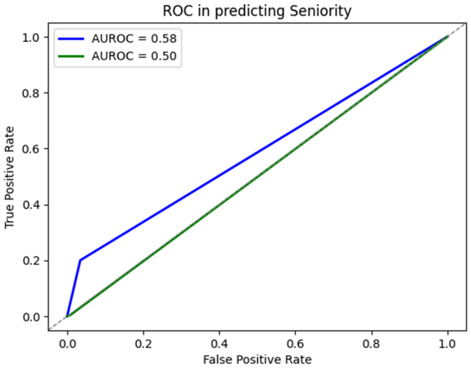

Figure 7: ROC Curves for Age Prediction

However, this was not the case for the second dataset, as both models yielded high True Positive rates and low False Positive Rates. This can be seen by looking at the confusion matrices for both models:

Confusion Matrix for KNN Model on the Breast Cancer dataset:
TPR = 0.917 and FPR = 0.0217

	Predicted Negative	Predicted Positive
Actual Negative	1	45
Actual Positive	22	2

Confusion Matrix for Decision Tree Model on the Breast Cancer dataset:
TPR = 0.859 and FPR = 0.0155

	Predicted Negative	Predicted Positive
Actual Negative	254	4
Actual Positive	13	79

Having far greater True Positive Rate and much lower False Positive Rates for all thresholds, we could see the ROC Curve approaching closer to the Top Left corner, indicating a much more reliable prediction model in this case:

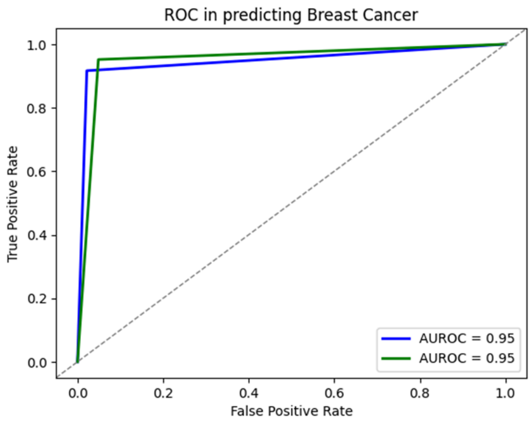

Figure 8: ROC Curves for Breast Cancer

Question 6:

As explained in the Data Sets section above, we computed the squared difference on each feature mean between classes to measure their significance. The key features became the features that were kept after discarding the least significant ones.

Question 7:

We also looked at the top 5 most used features on non-leaf nodes in each data set. We did this by printing out the used features at each node, and counting them. We used a tree depth of 10 for both data sets.

The five most used features for the breast cancer data base, from most to least used, were “Clump Thickness”, “Uniformity of cell shape”, “Uniformity of cell size”, “Single epithelial cell size”, and “Bare Nuclei”. However, the top five features by biggest mean difference, in order of biggest to smallest difference, were “Bare nuclei”, “Uniformity of cell size”, “Uniformity of cell shape”, and “Normal Nuclei”.

The five most used features for the age prediction database, from most to least used, were “Body mass index”, “Glucose level after fasting”, “Gender”, “Oral”, and “Insulin Level”. However, the top five features by biggest mean difference, in order of biggest to smallest difference, were “Oral”, “Glucose level after fasting”, “Insulin level”, “Activity level”, and “Body mass index”.

It can be seen that for either data set the two rankings do not match. There is some overlap as some features are in both top fives but not at the same rank. This can be explained by the fact that different cost functions and tree depths will generate different trees, and therefore will not use the same features in the same proportions.

6. Discussion and Conclusion

From the ROC curves, we understood how the performances of both models varied drastically on both datasets, where in the case of predicting Malignant Tumours based on the breast cancer data, both KNN and Decision Models were able to predict the test datasets to a high degree of accuracy and obtain an AUROC of 0.95, which gave us confidence in in our implementation of both models working correctly.

A possible reason as to why our AUROC was lower on the NHANES dataset is the fact that a portion of the test data consisted of a larger concentration of outliers than observed in the training data, thus resulting in many more False negatives than True Positives.

We also saw how different cost and distance functions can drastically change the accuracy of either model.

Further improvements can also be done at a later stage by implementing a weighted KNN algorithm, and performing K-fold Cross Validation on our datasets before splitting them to get a higher accuracy.

7. Statement of Contributions

Filicia: Task 1, KNN, and related experiments except ROC and AUROC, Documentation

Akos: DT and related experiments except ROC and AUROC, Documentation and Formatting

Aryan: ROC, AUROC, More Documentation