Assignment 1

An Ni Xu, Jia Song , YiWen Zhang

1 Abstract

In this assignment, we will analyse and compare two classification techniques seen in class: K-Nearest Neighbour (KNN) and Decision Trees (DTs). We first analyze the datasets and select important features to work with, then, to validate our implemented models, we try different hyper parameters and cost or distance functions for each of the models. The goal is to investigate the test accuracy and the performance of the two models on two visibly different datasets. We have concluded that KNN approach achieved better accuracy than DT approach for the Age_prediction NHANES dataset, and DT approach achieved better accuracy than KNN approach for the Breast Cancer dataset.

2 Introduction

This assignment consists of comparing two classification techniques (KNN and DTs) on the two datasets:

- National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset has two class labels : Adult (under 65 years old) and Senior (65 years old and older). The sample population is widely chosen from the U.S. population. A study, conducted by An Dinh, Stacey Miertschin, Amber Young, and S. MohantyUsing, used the NHANES dataset's features which can be used to predict and " [...] develop models for cardiovascular, prediabetes, and diabetes detection" (as cited on semantic scholar).
- Breast Cancer Wisconsin (Original) dataset has two class labels : 2 (benign) and 4 (malignant).

To evaluate the performance and quality of the models, we will compare the ROC curves and AUROC for both datasets of each model. To validate the models, we experiment with models using different hyper-parameters (K for KNN, and depth for DTs) as well as different distance/cost functions.

3 Methods

For each of the implemented methods, we need 3 functions:

- *fit*: this function stores the data points of the training set
- *predict*: this function predicts target label on the test datasets
- *evaluate_acc*: this functions evaluates the accuracy of the predictions
- KNN: This model is a "lazy learner", where no training occurs: only memorization of the data points. We first calculate distances between each pair of points of test sets and training sets , then select and store the K smallest distances for each test points. The points will be given the label of the class with higher neighbour percentage among its k-neighbours.
- DT: This model recursively splits the data by features while minimizing the cost function: we will be using a greedy function which compares all possible feature-value combinations and continuously split at the features with lowest cost function until a leaf node is reached. In addition, we will include *features_scores* which is a dictionary that includes the amount of time that each features were used for splits. We will use this dictionary to select top splitting features.

4 Datasets

1. We clean the dataset by removing NA values from both datasets since they are not values which our machine learning models can process.
2. All features are normalized to values between 0 and 1 for a more accurate and unbiased evaluation.
3. To have a better understanding of the input features, we computed features' means for positive and negative groups and the squared difference of the groups' means. For each feature, the largest its squared group means difference, the higher its association with the target variable. Also, correlation matrix and rough feature importance scores are used to determine the top features.
4. We further split both datasets into 33% training, 33% validation, and 34% test data for models evaluation.

5 Results

5.1 NHANES_age_prediction dataset

The KNN and DT models were trained on the NHANES age prediction dataset with two different distance functions(Euclidean and Manhattan) for KNN and with three different cost functions(Misclassification rate, Entropy and Gini index) for DT. Here are two summary tables of our results:

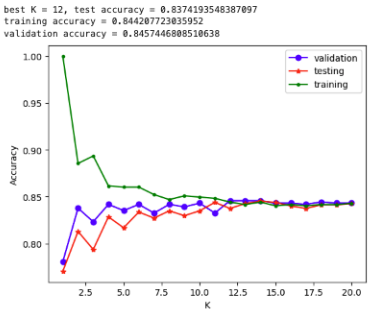
Distance Function	Test Accuracy	K Value
Euclidean	0.8374194	12
Manhattan	0.8438710	8

Cost Function	Test Accuracy	Best Depth
Misclassification	0.8387097	1
Entropy	0.8387097	1
Gini index	0.8387097	1

We conclude that KNN using Manhattan distance function returns the best accuracy rate of ≈ 84.39% and that DT using all cost functions returns the same test accuracy of ≈ 83.871%.

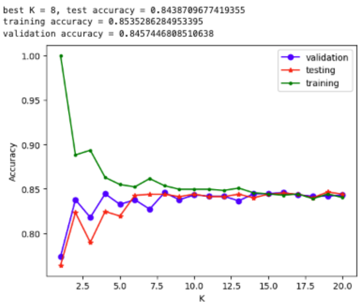
5.1.1 KNN on NHANES_age_prediction dataset

The graphs below show how different values of K affect the training data accuracy and test data accuracy. K = 12 for Euclidean distance function, and K = 8 for Manhattan distance function since it provides the best validation accuracy.



```
best K = 12, test accuracy = 0.8374193548387097
training accuracy = 0.844207723035952
validation accuracy = 0.8457446808510638
```

Figure 1: Euclidean



```
best K = 8, test accuracy = 0.8438709677419355
training accuracy = 0.8535286284953395
validation accuracy = 0.8457446808510638
```

Figure 2: Manhattan

5.1.2 Decision Tree on NHANES_age_prediction dataset

The graphs below show how different values of tree depths affect the training data accuracy and test data accuracy. For all cost functions, best depth = 1 and since it provides the best validation accuracy.

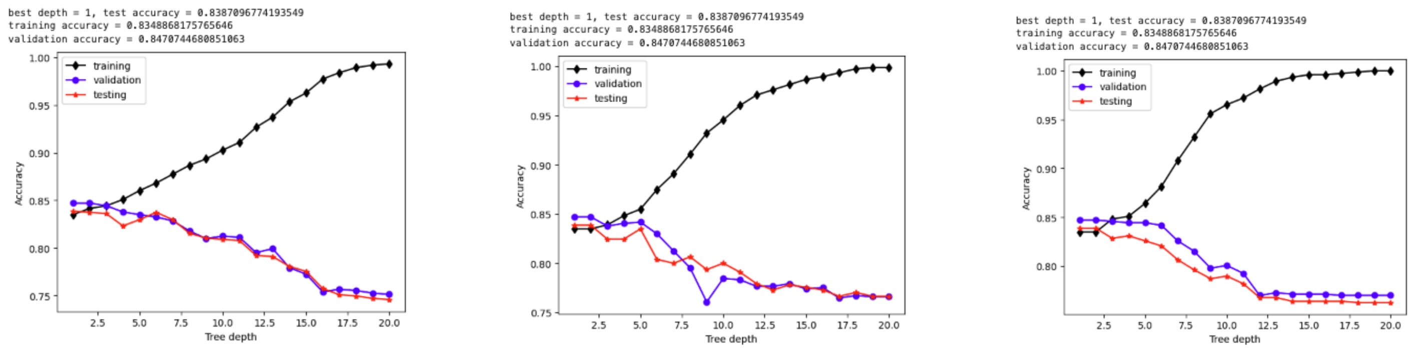


Figure 3: Misclassification,Entropy,Gini

5.2 Breast Cancer Wisconsin dataset

The KNN and DT models were trained on the Breast Cancer Wisconsin dataset with two different distance functions(Euclidean and Manhattan)for KNN and with three different cost functions(Misclassification rate, Entropy, and Gini index). Here are summary tables of our results:

Distance Function	Test Accuracy	K Value
Euclidean	0.9656652	1
Manhattan	0.9656652	1

Cost Function	Test Accuracy	Best Depth
Misclassification	0.9527897	2
Entropy	0.9656652	4
Gini index	0.9570815	3

We conclude that KNN using both distance functions return the same test accuracy rate of $\approx$ 96.567% and that DT using Entropy cost function returns the best accuracy rate of $\approx$ 96.567%.

5.2.1 KNN on Breast Cancer Wisconsin dataset

The graphs below show how different values of K affect the training data accuracy and test data accuracy. For both distance functions, K = 1 since it provides the best validation accuracy .

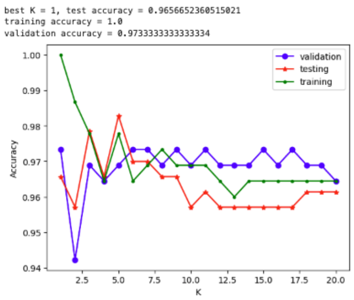


Figure 4: Euclidean

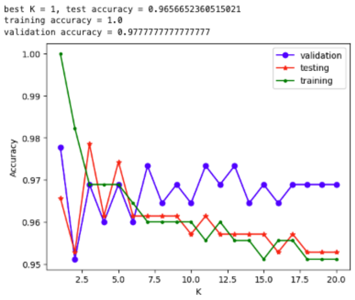
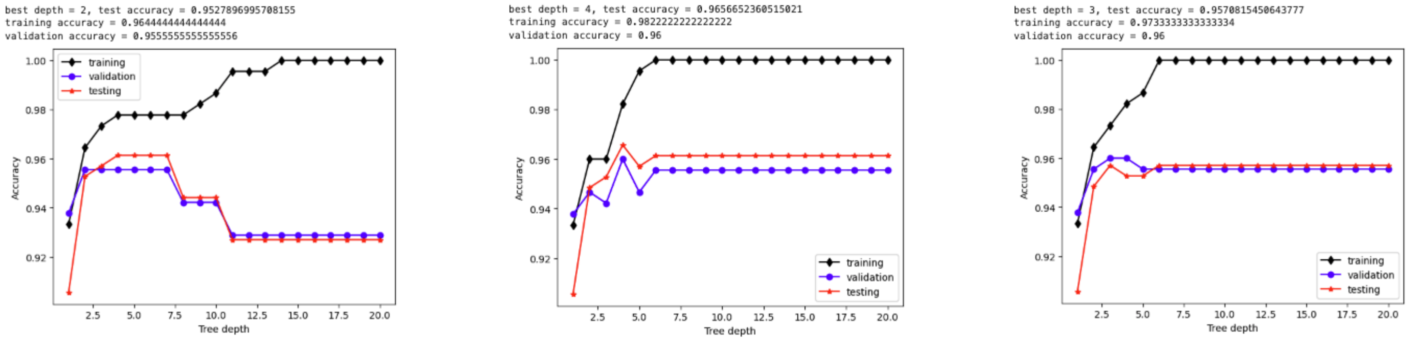


Figure 5: Manhattan

5.2.2 Decision tree on Breast Cancer Wisconsin dataset

The graphs below show how different tree depths affect the training data accuray and test data accuracy. For Misclassification rate cost functions, best depth = 2. For Entropy cost function, best depth = 4. For Gini index cost function, best depth = 3 since they provide the best validation accuracy.



```
best depth = 2, test accuracy = 0.9527896995708155
training accuracy = 0.9644444444444444
validation accuracy = 0.9555555555555556
```

```
best depth = 4, test accuracy = 0.9656652360515021
training accuracy = 0.9822222222222222
validation accuracy = 0.96
```

```
best depth = 3, test accuracy = 0.9570815450643777
training accuracy = 0.9733333333333334
validation accuracy = 0.96
```

Figure 6: Misclassification,Entropy,Gini

5.3 Model Comparison with AUROC/ROC

AUROC of KNN with Manhattan distance function is the best among the rest with a value of 0.55 on the Age Prediction dataset. AUROC of DT with Entropy cost function and AUROC of KNN with Euclidean distance function both have a value of 0.96 on the Breast Cancer dataset.

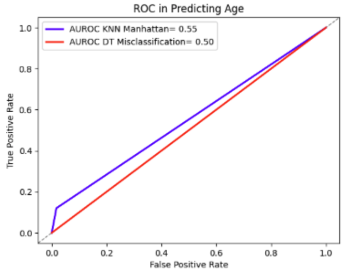


Figure 7: AUROC NHANES Dataset

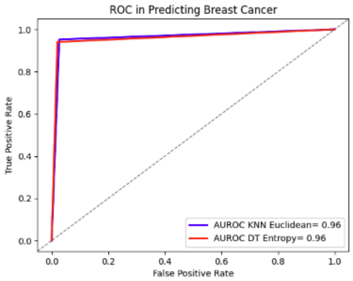
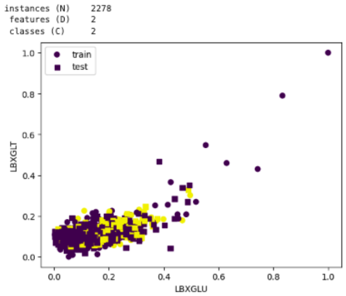


Figure 8: AUROC Breast Cancer Dataset

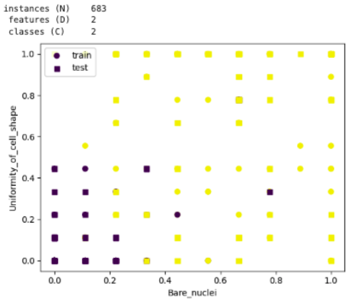
5.4 2D Scatterplot Data Visualization

Using the correlation and rough features importance scores(coded in Colab), we select LBXGLT and LBXGLU of NHANES Age Prediction dataset to be the two important features for plotting our 2D scatterplot. Similarly, we select Uniformity_of_cell_shape and Bare_nuclei of the Breast Cancer dataset to be the two important features for plotting the 2D scatterplot, even though many features are almost equally important. All features selected have positive correlation with the target variable, which means that data closer to the origin belong to negative(=0) group, whereas data farther away from the origin belong to positive(=1) group.



```
instances (N)    2278
 features (D)    2
 classes (C)     2
```

Figure 9: NHANES Age Prediction Dataset



```
instances (N)    683
 features (D)    2
 classes (C)     2
```

Figure 10: Breast Cancer Dataset

6 Discussion and Conclusion

6.1 Comparing KNN & DT on the NHANES_age_prediction dataset

The KNN model using the Manhattan distance function with K = 8 returns an AUROC score of 0.55, whereas the DT model using the Misclassification cost function returns an AUROC score of 0.50. Thus, KNN model with Manhattan distance function is better in this case. But we should keep in mind that this dataset is unbalanced, so it is very likely that KNN and DT might wrongly pick the majority class since they both have relatively low AUROC.

6.2 Comparing KNN & DT on the Breast Cancer dataset

The KNN model using the Euclidean distance function with K = 1 returns an AUROC score of 0.96. Whereas, the DT model using the Entropy cost function returns an AUROC score of 0.96. Both models return the same AUROC score. However, KNN model has best k = 1, which is too flexible since we classify data based on only one single feature. So, DT would be a better choice for this dataset.

6.3 Conclusion

From 6.1 and 6.2, we see that KNN performs better on the Age Prediction dataset while DT performs better on the Breast Cancer dataset. However, the performance of both models on the Age Prediction data is relatively low even though the dataset is relatively large. The reason may be that the Age Prediction dataset is unbalanced so KNN might wrongly pick the majority class. To solve this issue, one could balance the dataset before running experiment for a more accurate evaluation. For the Breast Cancer dataset, both KNN model and DT model have same value of AUROC and test accuracy. However, the best K value of KNN model for the Breast Cancer data is 1, which could be too flexible since data will be classified based on only one single feature. In other words, slight change of a feature value could result in quick change of its target variable. Therefore, we conclude that DT model is better for the Breast Cancer dataset.

7 Statement of Contributions

- Jia: Note book sections (code organiser and editor), report sections 5 and 6 main writer, report reviser (content and vocabulary check)
- An Ni: Main code writer, report editor (content check)
- Yi Wen: Main report writer and reviser, (vocabulary, grammar and syntax check), code writer procedures ideas provider

8 References

Dinh, A., Miertschin, S., Young, A., & Mohanty, S.D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19, 20 January 2024, https://www.semanticscholar.org/paper/A-data-driven-approach-to-predicting-diabetes-and-Dinh-Mi citing-papers

NA,NA. (2023). National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset. UCI Machine Learning Repository, 20 January 2024, https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+(nhanes)+age+prediction+subset

Wolberg,WIlliam. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository, 20 January 2024, `https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original`