Comparing K-Nearest Neighbour and Decision Trees On Two Distinct Health Datasets

Jacob Lerner 260958030, Jack Denton 260948222, Melody Bucchino 260971904

1. Abstract

In this assignment, we investigated the performance of two machine learning models, K-Nearest Neighbour (KNN) and Decision Trees, on two datasets, provided by the National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset and Breast Cancer Wisconsin. The assignment is structured into tasks including data acquisition, preprocessing, algorithm implementation, and running experiments. It emphasizes learning through practical implementation and includes both coding and reporting components. This assignment is designed to develop skills in machine learning programming, algorithm understanding, and comparative analysis of model performance. We found that KNN performed better than DT for both datasets when data was split into 80% train and 20% test.

2. Introduction

The assignment explores the efficiency of two machine learning models, K-Nearest Neighbour (KNN) and Decision Trees (DT), in analyzing two distinct health datasets: the National Health and Nutrition Examination Survey (NHANES) 2013-2014 Age Prediction Subset and the Breast Cancer Wisconsin dataset. These datasets are important in understanding the application of machine learning in diverse healthcare scenarios. The KNN model demonstrated an increasing trend in test accuracy as the number of neighbours (K) increased for the age prediction task. In contrast, for the breast cancer prediction task, the accuracy remained relatively high and stable across different K values. The highest test accuracy for the KNN model in age prediction was observed at K=15 with 82.7%, while for breast cancer prediction, the highest test accuracy was at K=5 and K=7 with 97.1%. The Decision Tree models exhibited different behaviour for age prediction, with increasing depth correlating to worse test accuracy, and similar behaviour for breast cancer prediction, with relatively high and stable accuracy across different depths. The highest test accuracy for the Decision Tree model in age prediction was observed at depth 1 with 82.5%, while for breast cancer prediction, the highest test accuracy was at depth 3 with 94.9%. Our testing and validation process emphasized the importance of selecting the right K value and adopting different distance measures for enhanced prediction accuracy. This aligns with the findings of Shahadat Uddin's comprehensive study in Nature, 'Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction.' Our data also helped explain how a shallow DT might not capture sufficient complexity, leading to underfitting, whereas a deep DT can overfit the data, reducing its generalizability to new data, as discussed in the study 'Evaluating algorithms of decision tree, support vector machine and regression for anode side catalyst data in proton exchange membrane water electrolysis'.

3. Methods

The two machine learning algorithms we implemented were the KNN and DT algorithms. The K-Nearest Neighbours algorithm is a non-parametric supervised learning classifier that uses proximity to make predictions about the group of an individual data point. The algorithm aims to identify the K nearest neighbours of a given query point so that we can assign a class label to that point. It can calculate the distance between the query point and all the data points using various distance metrics. Once the nearest neighbours are identified, the algorithm uses them to predict the output for the query point. The KNN algorithm is also referred to as a memory-based learning method as it stores its training data. A Decision Tree is a non-parametric supervised learning algorithm used for both classification and regression tasks. It has a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes. A decision tree starts with a root node; the outgoing branches from the root node feed into the internal nodes, also known as decision nodes and using the given features, each node assesses the data and puts it into subgroups represented by leaf nodes. The leaf nodes represent all the possible outcomes within the dataset. Decision tree learning employs a divide-and-conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated recursively until all or the majority of records have been sorted into specific class labels.

4. Datasets

This project involved the analysis of two distinct data sets. The first, derived from the National Health and Nutrition Examination Survey (NHANES), predicts the age group of respondents as 'Senior' (65 years and above) and 'Adult' (below 65 years). Upon loading the data into Google Collab, we assessed the correlation of each feature with the age group and removed all features with a correlation below the threshold of 0.1. This data set contained no null values, and after processing, was refined to include six features as seen in 5.0.4. By calculating and ranking the squared difference of the group means, the top five features associated with the target variable were determined to be the respondent's oral (LBXGLT), blood glucose after fasting (LBXGLU), blood insulin levels (LBXIN), weekly physical activity level (PAQ605), and BMI (BMXBMI), as seen in 5. The second data set originates from the Original Wisconsin Breast Cancer Database, identifying breast cancer patients. This data set was loaded into Google Collab, initially in two parts – one for features and another for targets. After merging these into a single DataFrame and removing null values, we split the data back up again and performed our correlation analysis, which determined all features to be relevant. The final data set comprised of nine features as seen in 5.0.4. The top five features associated with the target variable, as determined using the same method used for the first data set, are the bare nuclei, uniformity of cell size, uniformity of cell shape, normal nucleoli, and marginal adhesion, as seen in 8. Exploratory analysis of our data revealed a significant class imbalance in both data sets — the NHANES dataset had 1914 seniors to 634 adults, and the breast cancer dataset contained 444 patients with cancer versus 239 without. This imbalance posed a risk of bias in model training, potentially impairing performance. The mean values of each feature by target group for each data set are visualized in 3, 4, 6, and 7. The discrepancy between the top five features identified through squared differences of group means for each data set (as stated above) and the top five primary predictors for each data set in 5.0.4 can be attributed to the method's emphasis on amplifying larger disparities, which highlight outliers that may not necessarily align with predictive accuracy. Additionally, the magnitude of difference may not fully capture the more subtle relationships between the features and the target variable we aim to predict.

5. Results

5.0.1 Comparing accuracy of different values of K & Comparing DT vs KNN

Table 1: KNN and DT Accuracy for Datasets

KNN Accuracy

K	**Age Pred.**		**Breast Cancer**	
	Train	Test	Train	Test
1	100%	76.8%	100%	94.9%
3	89.1%	78.7%	98.2%	96.4%
5	86.9%	80.7%	97.8%	97.1%
7	85.6%	81.6%	97.6%	97.1%
9	85.7%	82.0%	96.9%	95.6%
11	85.8%	81.8%	96.9%	95.6%
15	85.3%	82.7%	97.1%	94.9%

DT Accuracy

Depth	**Age Pred.**		**Breast Cancer**	
	Train	Test	Train	Test
1	84.4%	82.5%	91.8%	90.5%
3	85.2%	81.6%	97.4%	94.9%
5	86.5%	81.1%	97.6%	92.7%
7	88.9%	80.3%	98.9%	92.7%
9	91.4%	78.5%	99.5%	94.2%
11	94%	75.9%	100%	92.7%
15	96.9%	75.4%	100%	92.7%

This table shows us both the test and train accuracy across different K values in the KNN algorithm. We can see across the Age prediction Dataset as K increases, train accuracy decreases while test increases.This could imply overfitting for a small value of K. This means when K is small the model is too closely fit to the training data and thus performs poorly on unseen test data. Although, as it grows the model becomes far more generalized and performs better.

For the breast cancer dataset, we see a notably higher overall accuracy. Also important to note the accuracy's seem much more stable as K increases. The perfect 100% training accuracy at K=1 occurs because the nearest neighbor of any data point in the training set is the point itself when K is set to 1. Therefore, the prediction for any training data point is always correct, as it's being compared to its own label.

5.0.2 Tree Depth on Decision Tree Performance

When comparing the accuracy of KNN and DT we can see that for the first dataset, as depth increases train data gets better but test data gets significantly worse. When looking at the second data set, there is a similar change. For the test set, we don't see a large enough change in accuracy to make a definitive answer on which performed better over time. Although, it is clear that the overall accuracy was better for DT in the second data set

5.0.3 Comparing ROC and AUROC of KNN and DT

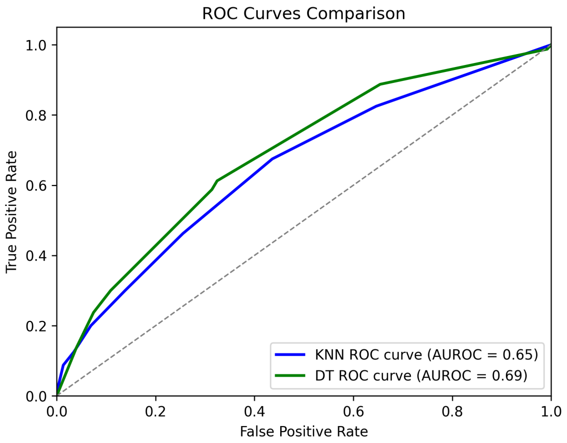

Figure 1: AUROC for Age Prediction

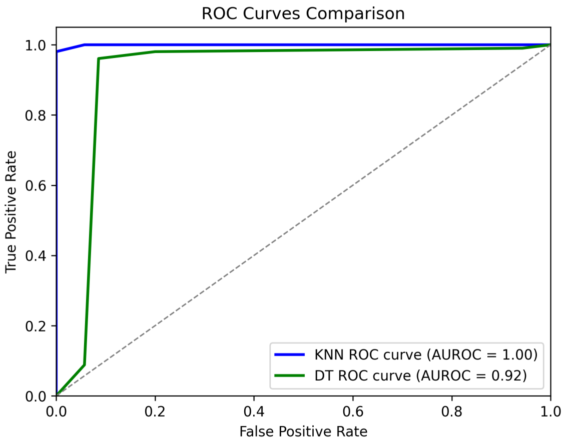

Figure 2: AUROC for Breast Cancer

In the comparative analysis of ROC (Receiver Operating Characteristic) curves for age prediction and breast cancer diagnosis models, the performance of two classifiers, KNN (K-Nearest Neighbors) and DT (Decision Tree), is evaluated. For age prediction, the AUROC (Area Under the ROC Curve) values are 0.65 for KNN and 0.69 for DT, indicating a moderate predictive ability with the Decision Tree model slightly outperforming the KNN.
In contrast, the breast cancer diagnostic models demonstrate superior performance, with the KNN achieving a near-perfect AUROC of 1.00 and the DT model reaching 0.92. The KNN's ROC curve hugs the top-left corner, implying an excellent true positive rate with a negligible false positive rate, which is ideal for medical diagnostic tests where the cost of false negatives is high.

5.0.4 Correlation between features and target

Table 2: Data Set 1 and 2 — Correlation of Features importance with Age and Breast Cancer

Feature	**Correlation with Age**
LBXGLT	0.318044
LBXGLU	0.229624
BMXBMI	0.147163
LBXIN	0.091879
DIQ010	0.049970
PAQ605	0.025973
SEQN	0.008806
RIAGENDR	0.006398

Feature	**Correlation**
Bare_nuclei	0.822696
Uniformity_of_cell_shape	0.821891
Uniformity_of_cell_size	0.820801
Bland_chromatin	0.758228
Normal_nucleoli	0.718677
Clump_thickness	0.714790
Marginal_adhesion	0.706294
Single_epithelial_cell_size	0.690958
Mitoses	0.423448

Correlation analysis between features and a target variable is done to understand the strength and direction of the linear relationship between two variables. In the context of health data, like age and breast cancer features, knowing which features are strongly correlated with the target can be crucial for predictive modeling. Features with higher correlation may have a more significant impact on the model's predictions and can help in feature selection, reducing dimensionality, and improving model performance.

5.0.5 Different distance and cost functions for KNN and DT

Analyzing the data from the KNN and Decision Tree (DT) models reveals insights into the best-performing functions based on test accuracy and AUROC. For Dataset 1, the KNN with Euclidean distance function and k=15 showed the highest test accuracy at 82.7% and an AUROC of 0.65, while the DT with the Misclassification cost function at a depth of 5 yielded comparable test accuracy at 82.9% and a lower AUROC of 0.56. The slightly higher test accuracy for DT might be due to its ability to capture more complex patterns without overfitting. In contrast, for Dataset 2, the KNN with Manhattan distance function and k=3 achieved the highest test accuracy at 97.1% and a perfect AUROC score of 1.00, indicating an exceptional fit to the test data and outstanding classification capabilities. Similarly, the DT with Misclassification cost function at depth 7 showed a test accuracy at 99.3% and a high AUROC of 0.99, suggesting that the DT model used the most telling features well at this level, fine-tuning its decision-making for highly accurate results. The function testing results can be found in tables 9 and 10.

5.0.6 Training, Validation, and Testing To Select Best K / Depth

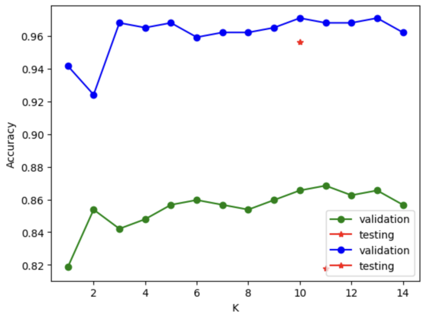

Figure 3: KNN for Dataset 1 & 2

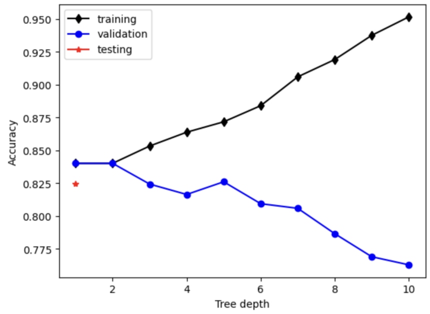

Figure 4: DT for Dataset 1

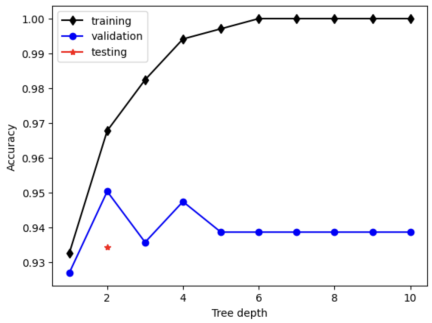

Figure 5: DT for Dataset 2

After splitting the data into 50% training, 25% validation, and 25% testing we found the most accurate K value for age prediction to be 11 with 81.80% test accuracy and the most accurate K value for breast cancer prediction to be 10 with 95.625% test accuracy. We found the most accurate depth for age prediction to be 1 with 82.46% test accuracy and the most accurate depth for breast cancer prediction to be 2 with test accuracy 93.43%.

The training set (50%), is used to teach the model to predict or classify, learning the relationship between features and target variables. The validation set (25%) helps fine-tune model parameters, such as K values for KNN or tree depth, acting as a checkpoint to avoid overfitting and ensure generalizability. The test set (25%) evaluates the final model performance, providing an unbiased assessment of its effectiveness on unseen data. This is done in addition to the original 80% training and 20% testing done earlier.

5.0.7 K-fold analysis

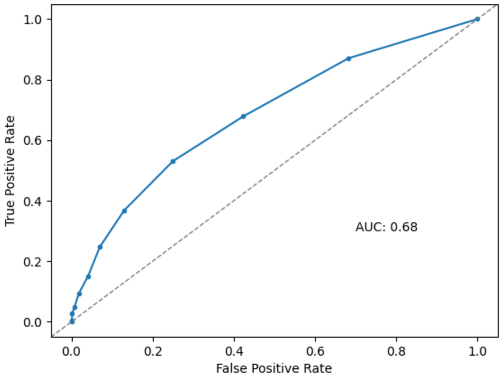

Figure 6: K-fold Nutrition Dataset

We employed k-fold cross-validation in our experimental design to enhance the accuracy and generalizability of our model. This method partitions the data into k subsets, allowing the model to be trained and validated K times, with each subset serving as a validation set once. This approach mitigates the risk of overfitting, as the model must prove robust across multiple independent data sets. It also maximizes the use of our data, ensuring that every data point contributes to both training and validation, leading to a more reliable performance estimate than a single train-test split. By adopting K-fold cross-validation, we aimed to produce a model with improved predictive accuracy when applied to new, unseen data.

6. Discussion and Conclusion

Several important conclusions are drawn from the performance examination of the KNN and DT models on two datasets. As the value of K grew, the KNN model for age prediction showed a pattern of increasing test accuracy and falling training accuracy, indicating a decline in over-fitting. In contrast, the breast cancer data set showed excellent accuracy for all K values, with perfect training accuracy at K=1 because the point itself is the nearest neighbour. For the Decision Tree models, trees with larger depths showed better training accuracy but worse test accuracy for age prediction, suggesting that over-fitting may occur at higher depths. However, the breast cancer dataset showed good accuracy at all depths; the best test accuracy, 99.3%, was found at a depth of 7. When we split data into training, validation and testing, the DT model significantly outperforms the KNN model in terms of predictive ability for age prediction, according to the AUROC values. The KNN model performed exceptionally well for the breast cancer data, with an almost perfect AUROC score and a high true positive rate with few false positives. The results of our correlation study highlighted the significance of key features in predictive modeling, such as the bare nuclei for breast cancer and LBXGLT (respondent's mouth) for age, which showed strong correlations with the target variables. For future investigations, focusing on model parameter tuning, such as standardizing all features and exploring and implementing more sophisticated machine learning algorithms, could improve the models' predictive capabilities.

7. Statement of Contribution

Jacob Lerner - Responsible for coding revolving around KNN and results with KNN.
Melody Bucchino - Responsible for calculations with basic data, methods, datasets and part of results.
Jack Denton - Responsible for coding revolving DT and results with DT.

8. Appendix

Table 3: Data Set 1 - Mean of Features for Age of Adults

Feature	Mean Value for Adults
PAQ605	1.81
BMXBMI	27.97
LBXGLU	98.64
DIQ010	2.01
LBXGLT	110.99
LBXIN	12.11

Table 4: Data Set 1 - Mean of Features for Age of Seniors

Feature	Mean Value for Seniors
PAQ605	1.91
BMXBMI	27.89
LBXGLU	104.33
DIQ010	2.03
LBXGLT	141.21
LBXIN	10.41

Table 5: Ranked Features for Dataset 1

Feature	Squared Difference
LBXGLT	974.58
LBXGLU	32.32
LBXIN	2.89
PAQ605	0.011
BMXBMI	0.007
DIQ010	0.00018

Table 6: Data Set 2 - Mean of Features for Patients with Breast Cancer

Feature	Mean Value for Breast Cancer
Clump Thickness	7.19
Uniformity of Cell Size	6.58
Uniformity of Cell Shape	6.56
Marginal Adhesion	5.59
Single Epithelial Cell Size	5.33
Bare Nuclei	7.63
Bland Chromatin	5.97
Normal Nucleoli	5.86
Mitoses	2.60
Class	4.00

Table 7: Data Set 2 - Mean of Features for Patients Without Breast Cancer

Feature	Mean Value for No Breast Cancer
Clump Thickness	2.96
Uniformity of Cell Size	1.31
Uniformity of Cell Shape	1.41
Marginal Adhesion	1.35
Single Epithelial Cell Size	2.11
Bare Nuclei	1.35
Bland Chromatin	2.08
Normal Nucleoli	1.26
Mitoses	1.07
Class	2.00

Table 8: Ranked Features for Dataset 2

Feature	Squared Difference
Bare Nuclei	39.45
Uniformity of Cell Size	27.78
Uniformity of Cell Shape	26.48
Normal Nucleoli	21.13
Marginal Adhesion	17.97
Clump Thickness	17.84
Bland Chromatin	15.14
Single Epithelial Cell Size	10.36
Class	4.00
Mitoses	2.36

Table 9: KNN Distance Function Testing Results

Distance Function	k Value	Train Accuracy	Test Accuracy	AUROC
		Dataset 1		
Euclidean	15	85.3%	82.7%	0.65
Manhattan	9	85.6%	82.0%	0.66
		Dataset 2		
Euclidean	7	97.6%	97.1%	0.99
Manhattan	3	98.2%	97.1%	1.00

Table 10: Decision Tree Cost Function Testing Results

Cost Function	Depth	Train Accuracy	Test Accuracy	AUROC
		Dataset 1		
Gini Index	1	84.4%	82.5%	0.69
Misclassification	5	85.7%	82.9%	0.56
Entropy	3	84.4%	82.5%	0.69
		Dataset 2		
Gini Index	3	97.4%	94.9%	0.92
Misclassification	7	97.3%	99.3%	0.99
Entropy	3	97.6%	95.6%	0.98

References

[1] Arjmandi, M., Fattahi, M., Motevassel, M., & Rezaveisi, H. (2023). Evaluating algorithms of decision tree, support vector machine and regression for anode side catalyst data in proton exchange membrane water electrolysis. *Scientific Reports*, Article number: 20309. Available at: `https://www.nature.com/articles/s41598-023-47174-w`

[2] Shahadat Uddin, Ibtisham Haque, Haohui Lu, Mohammad Ali Moni Ergun Gide (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*. Available at: `https://www.nature.com/articles/s41598-022-10358-x`