Assignment 1: Getting Started with Machine Learning

Maxx Railton[*], Dmitrii Vlasov[†]
COMP 551 Winter 2024, Mcgill University

January 31, 2024

Abstract

Summary. In this assignment, our team examined the performance of two primary machine learning models, K-Nearest Neighbour (KNN) and Decision Tree (DT), on two benchmark health datasets. Our examination involved a comprehensive process starting from preprocessing and analyzing the data, implementing the machine learning algorithms, and conducting experiments to compare their effectiveness. Our findings revealed that for the NHANES age prediction dataset, the KNN model achieved worse accuracy than the DT model. The opposite was true for the Breast Cancer Wisconsin dataset. Additionally, tuning the model's hyperparameters and implementing different cost and distance functions impacted the models' performance, with varying values of K in KNN and tree depth in DTs affecting accuracy and AUROC scores.
Keywords. Machine Learning, K - Nearest Neighbour, Decision Tree, Data Analysis

1 Introduction

This assignment involved the implementation and comparative analysis of two well-established machine learning models, K-Nearest Neighbour (KNN) and Decision Trees (DTs), applied to two distinct healthcare datasets. The first, the NHANES Age Prediction dataset, originates from the National Health and Nutrition Health Survey for 2013-2014, encompassing age-specific health data. The dataset is characterized by various features such as Body Mass Index (BMI), blood insulin levels, and gender, which vary with the target age variable. The second dataset, the Breast Cancer Wisconsin dataset, provides numerical data on benign and malignant tumour characteristics, with features including but not limited to clump thickness, bare nuclei, and marginal adhesion, each varying with the class of tumour. The initial phase of our study consisted of data acquisition and preprocessing, where we handled the issue of missing data in the Breast Cancer Wisconsin dataset, where rows with absent bare nuclei information were removed to maintain the integrity of the analysis. Our findings showed that in the NHANES dataset, the DT model exhibited better accuracy and a lower AUROC than the KNN model. In the Breast Cancer Wisconsin dataset, the KNN model outperformed the DT model in accuracy, and both models had equivalent AUROCs.

2 Methods

KNN works on the basis of similarity. Data points are plotted in a multi-dimensional space where each dimension represents a feature. Classification of new datapoint is determined on the class of the majority among its K nearest neighbors. The algorithm calculates distance (Euclidean, Manhattan etc) between the new data point and existing points, identifies the K nearest neighbors, and picks the class of the majority of nearest neighbors as a label for the new datapoint.

DT is structure where internal nodes represent a feature, branches represent a decision rule, and leaf nodes represent the outcome. The algorithm splits the each node based on the feature that results in largest information gain. This split continues until a maximum depth is reached.

[*]maximillian.railton@mail.mcgill.ca (260966381)
[†]dmitrii.vlasov@mail.mcgill.ca (261038431)

3 Datasets

The datasets used in this assignment were obtained using the ucimlrepo Python package. These Datasets contain various health-related features used to predict age groups classified as Senior or Adult and the class of tumours classified as Benign or 'Malignant. As shown in Tables 1 and 3, we examined the metadata to understand the structure and content of the datasets, including variable information. The datasets were split into feature data (X) and target labels (y). We then identified missing values in the datasets and removed the affected rows. To understand the distribution and differences between the age groups and tumour classes, we calculated the mean values of features for the Senior, Adult, Benign, and Malignant classifications. We further analyzed the features by computing the squared differences of the means between the two age groups and malignant and benign classes of Breast Cancer dataset, ranking them to identify the features with the most significant variation between classes. This can be seen in Table 2 and Table 4 For the experimental setup, the Dataset was shuffled and split into training, validation, and test sets following a 60/20/20 ratio, respectively.

3.1 Tables

Table 1: NHANES Age Prediction Dataset Metadata and Variables.

Name	Role	Type	Description	Missing Values
SEQN	ID	Continuous	Respondent Sequence Number	no
age group	Target	Categorical	Respondent's Age Group (senior/non-senior)	no
RIDAGEYR	Other	Categorical	Respondent's Age	no
RIAGENDR	Feature	Categorical	Respondent's Gender	no
PAQ605	Feature	Categorical	If respondent engages in moderate or intense activity	no
BMXBMI	Feature	Categorical	Respondent's Body Mass Index	no
LBXGLU	Feature	Categorical	Respondent's Blood Glucose after fasting	no
DIQ010	Feature	Categorical	If the Respondent is diabetic	no
LBXGLT	Feature	Categorical	Respondent's Oral	no
LBXIN	Feature	Categorical	Respondent's Blood Insulin Levels	no

Table 2: NHANES Age Prediction Dataset Features Ranking based on Squared Difference.

Rank	Name	Squared Difference of Group Means
1	LBXGLT	974.575736
2	LBXGLU	32.318625
3	LBXIN	2.894810
4	PAQ605	0.010645
5	BMXBMI	0.006728
6	DIQ010	0.000179
7	RIAGENDR	0.000014

Table 3: Breast Cancer Wisconsin Dataset Metadata and Variables.

Name	Role	Type	Description	Missing Values
Sample code number	ID	Categorical	None	no
Clump thickness	Feature	Integer	None	no
Uniformity of cell size	Feature	Integer	None	no
Uniformity of cell shape	Feature	Integer	None	no
Marginal adhesion	Feature	Integer	None	no
Single epithelial cell size	Feature	Integer	None	no
Bare nuclei	Feature	Integer	None	yes
Bland chromatin	Feature	Integer	None	no
Normal nucleoli	Feature	Integer	None	no
Mitoses	Feature	Integer	None	no
Class	Target	Binary	2 = Benign, 4 = Malignant	no

Table 4: Breast Cancer Wisconsin Dataset Features Ranking based on Squared Difference.

Rank	Name	Squared Difference of Group Means
1	Bare nuclei	39.448049
2	Uniformity of cell size	27.534017
3	Uniformity of cell shape	26.183019
4	Normal nucleoli	20.909380
5	Clump thickness	17.966483
6	Marginal adhesion	17.498234
7	Bland chromatin	15.045217
8	Single epithelial cell size	10.103929
9	Mitoses	2.328349

4 Results

After the data preprocessing, the datasets were divided into Training, Validation, and Test sets for choosing hyperparameters. The outcomes from our experiments revealed dataset-dependent performance variations between the models. In the NHANES dataset, the DT model demonstrated superior accuracy but lower AUROC scores than the KNN model, as illustrated in Figure 1. Conversely, the KNN model performed better than the DT model in accuracy and AUROC within the Breast Cancer Wisconsin dataset, as detailed in Figure 2. Hyperparameter optimization further varied the accuracy of the models, where the optimal number of neighbours (K) for KNN was determined to be 6 for the NHANES dataset and 9 for the Breast Cancer Wisconsin dataset. Similarly, the most effective maximum tree depth for DTs was 1 for NHANES and 3 for Breast Cancer Wisconsin. Moreover, we found that the Entropy cost function resulted in the best K value and accuracy for both datasets. The distance function varies between the two datasets, with the Manhattan function resulting in the most accurate tree depth for the NHANES dataset and the Euclidean function providing the highest accuracy for the Breast Cancer Wisconsin Dataset. These findings highlight the critical role of model configuration in maximizing performance. Finally, Key features for KNN were obtained using external feature selection by correlation with the labels (see Discussion section).

4.1 Figures

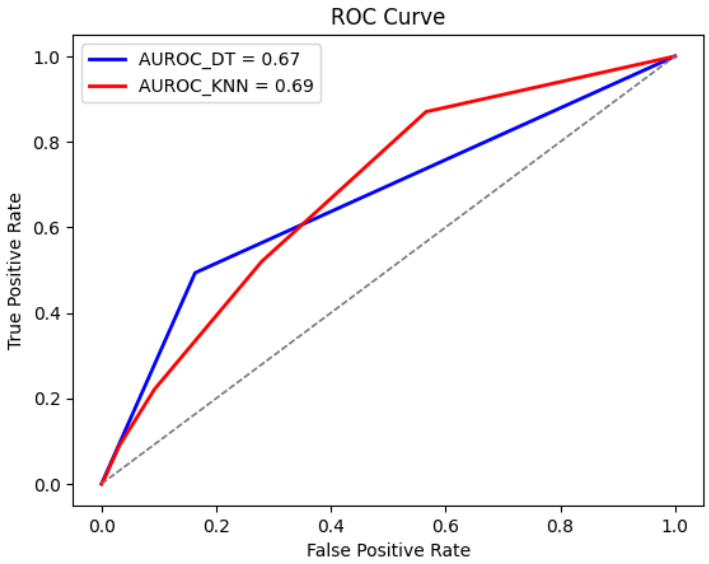

Figure 1: Plot of KNN and DT ROC Curves for NHANES Age Prediction Dataset.

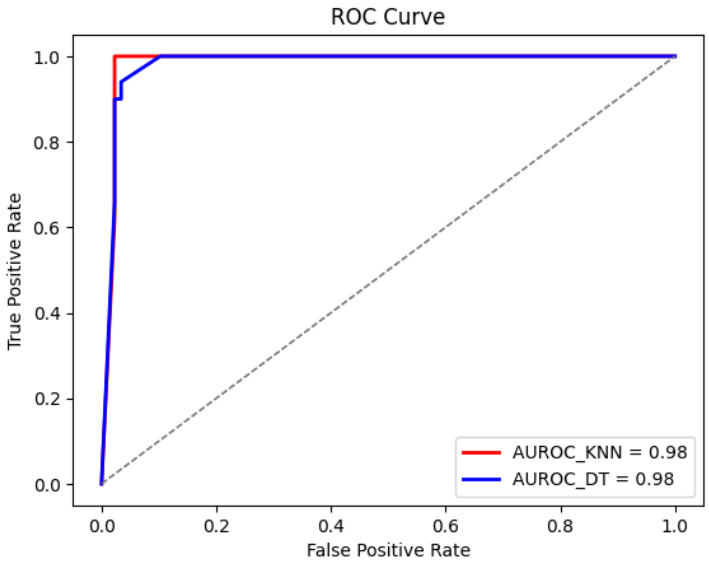

Figure 2: Plot of KNN and DT ROC Curves for Breast Cancer Wisconsin Dataset.

5 Discussion and Conclusion

Due to the nature of the data importation, the data was aggregated into a single dataset by aggregating the column vector of labels and the matrix of features. Both datasets were then shuffled to avoid potential issues with data splits. NHANES Age Prediction had no missing values and thus was left without change, while Breast Cancer Wisconsin had null values in the Bare nuclei column. These 16 rows containing null values were removed since the data was large enough to allow it (699 instances, 683 after removal). Cross validation was not considered as the data was abundant for general train-test-validation split (60%-20%-20%).

After performing preprocessing and implementing DT and KNN classes, slightly different approaches were utilized for both datasets. After calculating the correlation matrix for HANES Age Prediction Dataset, it was evident that only the following features had positive correlation with the age group, thus it only those features were used in KNN 'LBXGLT', 'LBXGLU', 'PAQ605'. The intuition to use only certain columns arose from the

squared difference of group means. At the same time, Breast Cancer Wisconsin demonstrated great performance for both KNN and DT with all the features present (around 98%), therefore all features were kept. Additionally, neither feature produced low squared difference of group means.

While training models for both datasets, multiple hyperparameters were explored on the validation set to avoid overfitting. For KNN, we chose the best performing pair of K and distance function, while for DT, we chose the best combination of depth and cost function. ROC curves were then plotted on the same graph. ROC curve proved to be a better representation of model performance for these datasets since there was an immense label imbalance: Adult - 1914, Senior - 364 (NHANES Age Prediction); benign - 458, malignant - 241 (Breast Cancer Wisconsin). ROC plotting and area under ROC were the only parts where sklearn library was used.

The performance of both models on Breast Cancer Wisconsin dataset was great (accuracy of 95.65 for KNN and 94.93 for DT), while model performance on NHANES Age Prediction demonstrated adequate results with room for improvement (accuracy of 82.06 for KNN and 83.15 for DT). Potential areas that could be modified include implementing weighted KNN, weighted sum based on the reduction cost for DT, or potentially using a different machine learning model for this dataset.

6 Statement Contributions

Dmitrii Vlasov: Data Preprocessing (NHANES), Implementation of KNN and DT, Execution of Experiments, Write-up
Maxx Railton: Data Preprocessing (Breast Cancer Wisconsin), Implementation of DT, Write-up .