COMP 551 Assignment 1

Matthew Barg, Jack Kelly, Cleo Norris

January 31, 2024

Abstract

In this assignment, we analyzed the performance of two basic ML models: K-Nearest Neighbour (KNN) and Decision Trees (DT). The testing uses two UCI health datasets: Age Prediction Subset of National Health and Nutrition Health Survey, and Breast Cancer Wisconsin. Comparing KNN and Decision Trees, it was concluded that for the NHANES dataset, KNN achieved its best performance with k=8 and an accuracy of about 0.813596. DT achieved its best performance at a depth of 1, which gave an accuracy of about 0.813596. For the breast cancer dataset, KNN achieved its best performance with k=5 and an accuracy of about 0.948905. DT achieved its best performance at a depth of 3, which gave an accuracy of about 0.956204.

1 Introduction

The goal of this assignment was to familiarize ourselves with two classification techniques, KNN and DT, learn how to preprocess data, run experiments, and compare methods of implementation. To do so, we examined two datasets from the UC Irvine ML Repository: the NHANES age prediction dataset, and the Breast Cancer Wisconsin dataset. Both KNN and DT can be used for either classification or regression problems, and for both datasets we performed classification.

In existing literature, the NHANES dataset has been used to classify at-risk patients using models including logistic regression, Support Vector Machines (SVM), and ensemble models like Random Forest, Gradient Boosted Trees, and Weighted Ensemble Model. It has also been used to perform a demographic and dietary features analysis for the classification of subjects according to their oral health status and their age group. The methodology included logistic regression and a net reclassification improvement (NRI) for validation, along with AUC, ROC, and OR values to evaluate classification accuracy.

The Breast Cancer Wisconsin dataset has been used to study estimators of case prevalences under separate sampling using analytical and numerical methods, an important issue in the application of traditional statistical learning techniques to biomedical data. This paper also notes that under separate sampling, the unbiasedness property of k-fold cross-validation does not hold. Another paper uses the breast cancer dataset to propose a reweighted least squares algorithm for L1-Norm Principal Component Analysis.

2 Methods

The two ML algorithms implemented in this assignment were K-Nearest Neighbours (KNN) algorithm and Decision Tree (DT) algorithm.

KNN is a examplar-based learner/non-parametric method that predicts the label of a data point by referencing similar data points with known labels. To implement KNN, one must split the data into training, testing, and validation data. The amount of data points in the training set used for the prediction is dependent on the value of K. The algorithm determines the K nearest neighbours of the desired point and can make this prediction using a number of distance formulas (Manhattan, Hamming, Euclidean), depending on whether the input features are continuous or discrete/categorical. The value of K is a hyper-parameter that is determined empirically through testing on the validation dataset, and it controls the degree of smoothing in the sense that as K increases, there are fewer and larger regions of each class. For each new data point, class probabilities of each class are calculated.

A DT is a supervised learning algorithm that has a non-linear approach. Data is recursively split up through a sequence of binary selections until a full tree is built containing a root node, internal nodes and leaf nodes. Each leaf node contains one of the possible outcomes for the given dataset, and the number of leaf nodes is the maximum number of unique class probabilities that a DT can produce. Each split is evaluated by a cost function, and the feature and threshold for each split is chosen to minimize this cost function. Similarly to the KNN procedure, a

validation set is used to choose the tree depth.
Both algorithms were implemented using the Object Oriented Programming paradigm, with functions defined for the classes that accomplish the model training and prediction based on the input data. Additionally, a function was defined to evaluate the accuracy of both models. Finally, KNN is sensitive to feature scaling, so the scaling was performed on the data for KNN after the train/test split (the data is not normalized all at once to avoid giving the model any additional information on the testing data). We note that decision trees are not sensitive to the variance in the data and don't require feature scaling.

3 Datasets

Two datasets were analyzed in this assignment. Dataset 1 (DS1) is NHANES's Age Prediction Subset, which predicts age and categorized subjects into 'Adults' (below 65) or 'Senior' (65 and above). The dataset contained 7 features that were all taken into account (gender, exercise level, BMI, blood glucose after fasting, diabetic status, oral, blood insulin levels). After loading the dataset, it was discovered that this data set had no null entries. We then performed exploratory analysis, starting with computing the mean of each feature for each class. From this we hypothesized that some features were relatively similar for both classes, while others varied between the classes. We confirmed this by ranking the squared differences of the group means and found that there were 5 features with significant mean differences between classes, and 4 that had negligible differences (about 0.01 or less). The data was then standardized using sklearn's StandardScaler function, as KNN is sensitive to feature scaling. This was done after splitting the training and testing data so as to prevent data leaks.

Dataset 2 (DS2) is a Wisconsin Breast Cancer Database that identified subjects with breast cancer. This dataset had null entries. Since this is not time-series data, where dropping values would be a detriment to our analysis as opposed to imputing values based on averages, for example, we proceeded by dropping the null values from our data, reducing the amount of rows from 699 to 683. All 9 features were taken into account in our work (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses).

We noted that there was a substantial difference in the sizes of the two datasets (2278 instances in DS1 vs 683 instances in DS2), which should be considered when comparing accuracy between the datasets respective implementations of KNN and DT. Additionally, within the datasets there was an uneven distribution of data. For example, of the 2278 instances in dataset 1, 1914 were categorized as 'Adults' and only 364 were 'Senior'. This made the algorithms prone to a potential bias.

4 Results

In the analysis below, DS1 refers to the NHANES dataset, and DS2 refers to the Wisconsin Breast Cancer dataset.

4.1 KNN and DT Accuracy

The figures below show the testing accuracy of different K values in the KNN algorithm and different tree depths for the DT algorithm for both datasets. We ultimately selected the values of these hyperparameters that produced the highest test accuracy for each dataset and algorithm. For the Health and Nutrition data, $K = 8$ and a max tree depth of 1 were found to produce the best results, whereas $K = 5$ and a max tree depth of 3 were optimal for the Breast Cancer data. When comparing the two datasets, a significant difference in accuracy is visible. One possible reason for this is the difference in sample sizes.

4.2 ROC Curve Comparisons

As can be noted from the shape of ROC curve, the higher test accuracy and the higher AUROC values, both the KNN and Decision Tree algorithms performed much better on the Breast Cancer dataset than the Health and Nutrition dataset. KNN had a test accuracy of 81.3% on the Health and Nutrition dataset versus 94.9% on the Breast Cancer data, with AUROC values of 0.73 and 0.97 respectively. Similarly, the DT algorithm had a test accuracy of 80.4% on the Health and Nutrition dataset versus 95.6% on the Breast Cancer data, with AUROC values of 0.54 and 0.94 respectively. It is interesting to note the DT that had an AUROC of 0.54 (meaning this classifier was barely better than a random one) had a depth of 1. This depth was chosen via testing on a validation

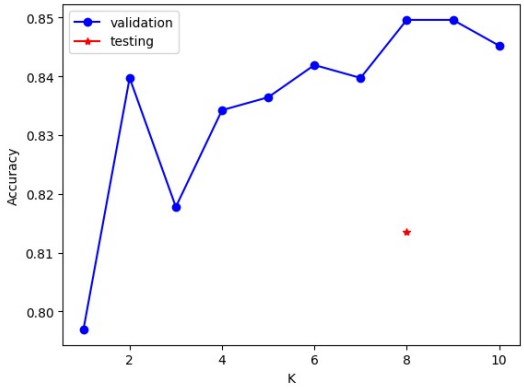

Figure 1: KNN, Euclidean distance, DS1

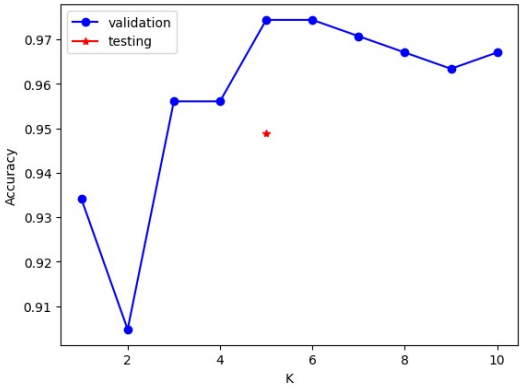

Figure 2: KNN, Euclidean distance, DS2

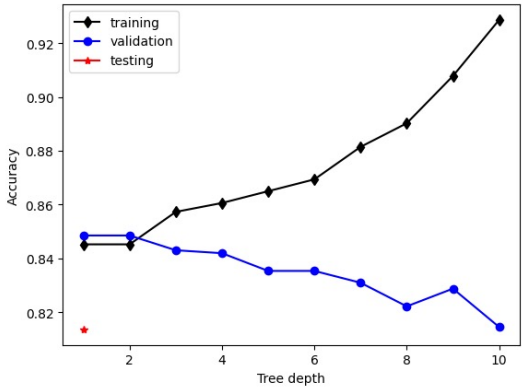

Figure 3: DT, Entropy cost, DS1

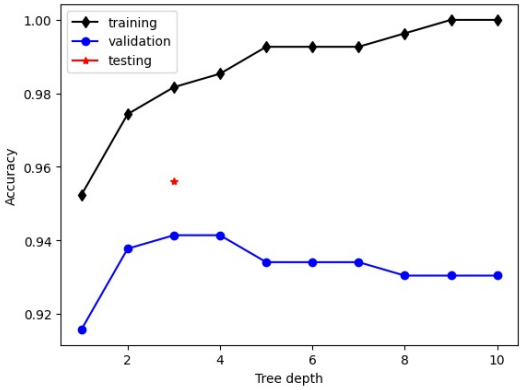

Figure 4: DT, Misclassification cost, DS2

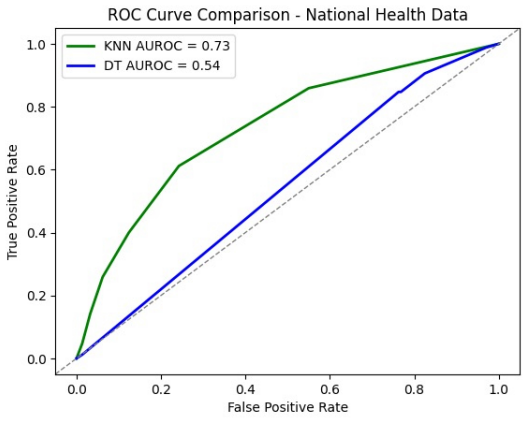

Figure 5: ROC Curve Comparison DS1

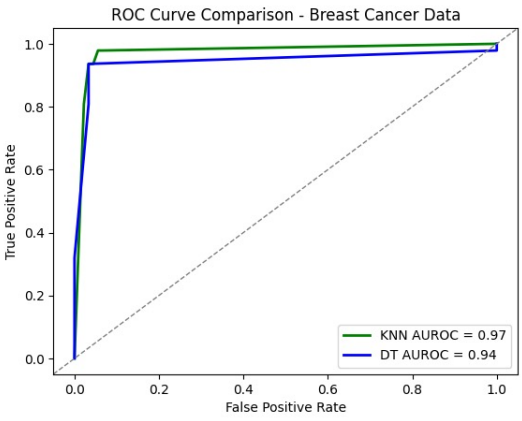

Figure 6: ROC DS2

set, but as shown by figure 3, deeper trees led to better training performance but worse validation results, likely due to overfitting.

4.3 Test different distance/cost functions

Below are the results of our experiments:

DS1 KNN Results		
Distance function	Accuracy	K
Euclidean	0.813596	8
Manhattan	0.809211	6

DS2 KNN Results		
Distance function	Accuracy	K
Euclidean	0.948905	5
Manhattan	0.948905	7

DS1 DT Results		
Cost function	Accuracy	Depth
Misclassification	0.804825	6
Entropy	0.813596	1
Gini Index	0.813596	1

DS2 DT Results		
Cost function	Accuracy	Depth
Misclassification	0.956204	3
Entropy	0.956204	3
Gini Index	0.956204	3

It is interesting to note that the Decision Tree for DS2 performed the same for all cost functions. It is also intersting to note that both KNN and DT performed better on average for DS2 than DS1, observed from the Accuracy columns.

4.4 Choice of Key Features for KNN

As the KNN algorithm does not come with any inherent system for ranking feature importance (as other methods like Decision Trees do) we calculated the Point-Biserial Correlation Coefficient between each of the numerical features and the binary category of interest for each data set. This was chosen because it is applicable to the correlation between numerical and a categorical variables, whereas other common measures of correlation like Pearson's correlation coefficient r are strictly intended for numerical variables. Similar to other correlation metrics, this coefficient ranges from -1 to +1, where -1 indicates a perfect negative association, +1 indicates a perfect positive association, and 0 indicates no association. For the Health and Nutrition data, calculating these coefficients shows us that the only variable with a strong association with either age category is "RIDAGEYR" (the respondent's age), which is not surprising. For the Breast Cancer data, all the variables were moderate to strongly associated with Malignant tumours. This could explain why both classification methods performed better on this dataset than the previous one. Although both methods are capable of creating non-linear decision boundaries, the presence of a definite difference in the predictor variables between the outcome groups would likely result in better model performance.

4.5 Feature Importance Scores

We defined a function `calc_importance` to compute a rough feature importance score for each feature. We see that for the DT with Misclassification cost for DS1, the top 5 features are BMXBMI (Importance Score = 15), RIAGENDR (Importance Score = 8), PAQ605 (Importance Score = 2), LBXGLU (Importance Score = 7), and LBXGLT (Importance Score = 1). This is interesting as there are some differences from the simple mean difference approach performed before - the top 5 were SEQN, RIDAGEYR, LBXGLT, LBXGLU, and LBXIN. Notably, BMXBMI and PAQ605 had relatively low differences between classes. These two functions measure different metrics, so it makes sense that there is some difference. Also, the difference in group means may take into account potential interactions between features, whereas the DT algorithm evaluates the features individually.

For the DT with Misclassification cost for DS2, the top 5 features are Uniformity_of_cell_size (Importance Score = 2), Bare_nuclei (Importance Score = 1), Normal_nucleoli (Importance Score = 1), Clump_thickness (Importance Score = 2), and Uniformity_of_cell_shape (Importance Score = 1). The mean differences approach gave Sample_code_number, Bare_nuclei, Uniformity_of_cell_size, Uniformity_of_cell_shape, and Normal_nucleoli as the top 5 that vary most between classes. These are almost the same, which perhaps demonstrates another reason why DT performed better for DS2.

5 Discussion and Conclusion

One of the most notable results from these experiments is that both ML models performed with higher accuracy on DS2 (Breast Cancer) than on DS1 (NHANES). As suggested above, this could be in part because all variables in DS2 were moderate to strongly associated with Malignant tumors, giving an unbalanced training set.

For DS1, the highest accuracy achieved by both KNN and DT was the same. Thus to choose between the methods, one must also consider factors such as model interpretability, robustness to noise, and computational efficiency. For DS2, DT achieved higher accuracy for all cost functions than KNN achieved.

One possible area of future investigation is feature engineering, in which only a few features are chosen from the given dataset for training and prediction, as a method of dimensionality reduction and to reduce the amount of noisy information. Another area of focus could be a more comprehensive hyperparameter tuning analysis by using search methods like grid search or Bayesian optimization. Finally, steps could be taken to ensure that the training data for both datasets is more balanced between classes. Taking these steps would help in determining exactly which ML method is better for each dataset.

6 Statement of Contributions

There was an equal contribution from team members across cleaning data, implementing the methods, experimenting, and the write-up.