Analysis of NHANES and BREAST CANCER Datasets

Bonaventure Dossou, Mat Vallejo, Minjae Kim

January 31, 2024

Abstract

This report examines the performance of two machine learning models, K-Nearest Neighbor (KNN) and Decision Tree (DT), on two distinct datasets: NHANES and BREAST CANCER. The evaluation was conducted by exploring various aspects of model tuning such as hyperparameters, distance functions (KNN), and cost functions (DT), across both datasets and comparing the results. The results show strong capabilities of both models to accurately predict target values, with both models outperforming each other, under certain conditions. Notably, the DT model was able to achieve perfect accuracy on the NHANES dataset due to the nature of the input variables.

1 Introduction

The purpose of this assignment is to implement K-Nearest Neighbor (KNN) and Decision Tree (DT) machine learning models across two distinct datasets, NHANES and BREAST CANCER. The tasks include importing and cleaning the datasets, splitting the datasets into train/validation/test sets, establishing accuracy across all input variables as well as specific subsets of highly relevant input features. Additionally, we have to build the models, and experiment them with various hyperparameters, in order to access their efficiency and accuracy in predicting target values. Hyperparameters include different values of k and different distance functions in the KNN model and cost functions in the DT model as well as K-fold cross validation.

2 Methods

The KNN method is widely used in the machine learning field for classification problems, and involves measuring test data against training data that has been fit to the model by use of distance functions (see more in KNN Experiments). The test data is evaluated by taking the test input variable(s) 'X' and finding the closest k-number of training target values 'y'. By calculating the distance between these points, the model can infer the most likely label for the target value.

The DT method can be used for classification, regression, or a combination of both. It functions by passing input values 'X' through a combination of conditional nodes that determine whether or not a value should be passed either "left" or "right" by implementing certain cost functions using the training datam (see more in DT Experiments). When the input variable reaches either a pre-designated maximum "depth" or the last node in the set, known as a leaf node, the decision 'y' is made.

3 Dataset Description and Statistics

The sub-dataset of the NHANES, managed by the CDC, selectively extracts features such as physiological data, lifestyle choices, and biochemical markers from the broader dataset to predict the age of a diverse U.S. population, addressing its usual breadth for more specific analysis. The Breast Cancer dataset is the original Wisconsin Breast Cancer Database.

NHANES contains no missing values and has no duplicate values. Breast Cancer has one variable called `Bare_nuclei` which has 16 missing values, representing 0.023% of its total number of samples. Consequently, we decided just to dynamically drop those values. Additionally, this dataset also has 234 duplicate values (out of 683), that have been dropped to avoid contamination.

3.1 Breast Dataset Statistics

Feature	Malignant	Benign	Mean Square Diff
Bare_nuclei	7.627615	1.346847	**39.448049**
Uniformity_of_cell_size	6.577406	1.306306	**27.784490**
Uniformity_of_cell_shape	6.560669	1.414414	**26.483941**
Normal_nucleoli	5.857741	1.261261	**21.127622**
Marginal_adhesion	5.585774	1.346847	17.968504
Clump_thickness	7.188285	2.963964	17.844884
Bland_chromatin	5.974895	2.083333	15.144255
Single_epithelial_cell_size	5.326360	2.108108	10.357144
Mitoses	2.602510	1.065315	2.362969

Table 1: Statistical summary of Breast Cancer Dataset features.

- Top-4 Ranked by Squared Difference: `[Bare_nuclei, Uniformity_of_cell_size, Uniformity_of_cell_shape, Normal_nucleoli]`
- Top-4 With Spearman Correlation: `[Uniformity_of_cell_size, Uniformity_of_cell_shape, Bare_nuclei, Single_epithelial_cell_size]`
- Top features associated with the target variable: `[Bare_nuclei, Uniformity_of_cell_shape, Uniformity_of_cell_size]`

3.2 NHANES Dataset Statistics

	Senior	Adult	mean_square_diff
Age	73.425824	35.780564	**1417.165594**
Oral	141.208791	109.990596	**974.575736**
Glucose	104.329670	98.644723	**32.318625**
Blood Insulin	10.405247	12.106661	**2.894810**
Activness	1.909341	1.806165	0.010645
BMI	27.886264	27.968286	0.006728
Diabetic	2.027473	2.014107	0.000179
Gender	1.508242	1.512017	0.000014

Table 2: Statistical analysis of NHANES dataset.

- Top-4 Ranked by Squared Difference: `[Age, Oral, Glucose, Blood Insulin]`
- Top-4 With Spearman Correlation: `[Age, Oral, Glucose, Activeness]`
- Top features associated with the target variable: `[Oral, Age, Glucose]`

4 Results

4.1 Feature Importance and Impact on Performance

Based on the results in Table 3, it is overall obvious that using top-3 features gives the best results, on both models. Moving on next, we will stick to those 3 features. Those features were selected as follows:

- Compute the square difference between both classes for all features and select the top ones.
- Compute the Spearman Rank Correlation between all features and the target variable.
- The final set of features used is the intersection of features of the two steps described above.

Model	Dataset	With all features	With top-3 features
Decision Tree	NHANES	1.0	**1.0**
KNN	NHANES	0.9672	**0.9803**
Decision Tree	BREAST	0.9111	**0.9333**
KNN	BREAST	0.9556	**0.9666**

Table 3: KNN Accuracy with All Initial Features vs With top-3 Best Features.

4.2 KNN Experiments

4.2.1 Evaluation of different values of k hyperparameter for KNN

k	acc_nhanes	auc_nhanes	acc_breast	auc_breast
10	0.980263	0.997418	**0.966667**	**0.995769**
20	0.973684	0.996374	0.966667	0.996765
30	**0.969298**	**0.995437**	0.955556	0.997760
40	0.971491	0.994800	0.944444	0.997511
50	0.969298	0.993385	0.944444	0.997511
60	0.971491	0.992677	0.933333	0.997511
70	0.967105	0.992094	0.944444	0.997511
80	0.967105	0.992058	0.933333	0.997760
90	0.980263	0.991952	0.933333	0.997511
100	0.980263	0.991616	0.933333	0.997511

Table 4: KNN performance with different k values.

We decided to stick to AUC to select the optimal k. This is because AUC is better when the problem is imbalanced (which is the case here). Given this, AUC is also more robust to changes in the threshold and can capture the model's performance across the entire range of probabilities. Finally, AUC is also useful when the cost of false positives and false negatives are different and need to be balanced. Overall, AUC will allow us to select models that achieve false positive and true positive rates that are significantly above random chance, which is not guaranteed for accuracy. Consequently, based on the results of Table 4, for NHANES and BREAST CANCER datasets, the chosen best k values are respectively $k = 10$ and $k = 30$. To compute the AUROC, since we need the probabilities of the positive class, we have implemented and used the `predict_proba` in both models' classes. These metrics are computed using the Euclidean distance function. In the KNN class, we have also implemented the cosine similarity and the Manhattan distance.

4.2.2 KNN Comparison of Several Distance Functions with the Best Hyperparameters

Metric	Dataset	Euclidean Distance	Cosine Similarity	Manhattan Distance
Accuracy	NHANES	0.980263	0.721491	**0.982456**
AUCROC	NHANES	0.997418	0.076376	**0.998320**
Accuracy	BREAST	**0.955556**	0.366667	0.944444
AUCROC	BREAST	**0.997760**	0.298656	0.997511

Table 5: KNN Model Performance As Function of the Distance Function

Following the results in Table 5, we can see that in the case of NHANES, the Manhattan distance works better than the Euclidean distance; as opposed to the BREAST Cancer dataset. The following might be possible reasons:

- When the data is high-dimensional, Manhattan distance can better capture the similarity between points than Euclidean distance (Source).
- When the data has different scales or units, Manhattan distance can be more robust to outliers or irrelevant features than Euclidean distance (Source).
- When the data is discrete or categorical, Manhattan distance can be more appropriate than Euclidean distance, which assumes a continuous space (Source).
- Manhattan distance uses the L1 Norm which encourages sparsity while Euclidean Distance uses the L2 norm. Manhattan distance works better input variables are not similar in type (e.g. in NHANES they are Glucose, Oral, and Age which are highly likely unrelated vs in BREAST they are Bare_nuclei, Uniformity_of_cell_size, and Uniformity_of_cell_shape which are very likely to be related). Moreover, due to the curse of dimensionality, Euclidean distance becomes a poor choice as the number of dimensions increases. (MIT Lesson).

For the remaining KNN experiments, we have used the Manhattan and Euclidean distances, respectively for the NHANES and BREAST CANCER datasets.

4.2.3 Exploring Impact of K-Fold Cross Validation

We implemented the k-fold cross validation technique and leveraged it important on final model's performance on the test set. Below are the results:

Dataset	Metrics	k=5	k=10	k=15	k=20
BREAST	Evaluation Accuracy	0.9155	0.9143	0.9130	0.8824
BREAST	Test Accuracy	0.9333	0.9444	0.9444	0.9444
BREAST	Evaluation AUC	0.9601	1.0	1.0	1.0
BREAST	Test AUC	0.9973	0.9973	0.9980	0.9980
NHANES	Evaluation Accuracy	0.9718	0.9714	1.0	1.0
NHANES	Test Accuracy	0.9583	0.9627	0.9649	0.9649
NHANES	Evaluation AUC	0.9985	0.9967	1.0	1.0
NHANES	Test AUC	0.9891	0.9895	0.9891	0.9903

Table 6: KNN K-Fold Cross-Validation Results.

Comparing Tables 6 and 5 (experiments without cross validation), we can notice that K-Fold cross-validation helps. However, compared to the best hyperparameters results, these new results are statistically insignificant (in the case of the BREAST CANCER dataset) and not better in the case of the NHANES dataset.

4.3 Evaluation of DT model

The evaluation of the decision tree model was measured for a range of maximum depth k for $k = [1, 20]$ using three separate training costs and finding the best depth for each. The range of k depth values was chosen to give general context to the results after preliminary tests showed significant diminishing returns in accuracy variance after a max depth of $k \approx 7$. The training costs implemented were chosen as misclassification rate, entropy, and gini index. The implementation of the same decision tree iteration using various training cost calculations highlights the strengths, weaknesses, and similarities of each in the results. General attributes of each training cost are as follows:

- **Misclassification** measures the total proportion of misclassified samples and has the potential to be insensitive as a training cost as a result (Source).
- **Entropy** refers to the amount of uncertainty present at a given node and is calculated by evaluating the potential split from each input variable across the possible outcomes. The lower the entropy, the more information is gained increasing decision accuracy (Source).

- **The gini index** is commonly used as the default cost function for decision tree architectures (such as in sklearn) and calculates how likely a variable is to be misclassified. In this case, low values also equate to increased decision accuracy (Source).

Dataset	Model	Misclassification	Entropy	Gini Index
NHANES	Accuracy	1.0	1.0	1.0
NHANES	AUCROC	1.0	1.0	1.0
NHANES	Best Depth	1	1	1
BREAST	Accuracy	0.955556	0.933333	0.955556
BREAST	AUCROC	0.95	0.98	0.95
BREAST	Best Depth	4	4	4

Table 7: Decision Tree Results with Best Hyperparameters

Considering results in Table 7, in the case of the NHANES dataset, the model performed optimally under all conditions. We see subtle shifts in accuracy in the BREAST CANCER implementation, notably when using entropy as the training cost, which resulted in a slight decrease in accuracy measurements but a slight increase in AUCROC. Overall, the accuracy of the model is functionally consistent across all hyperparameters and cost functions.

5 Conclusion

Experiments conducted for the NHANES and BREAST CANCER datasets reflect the varying results that can be achieved when implementing KNN and Decision Tree machine learning models. As shown in Figure 1 (see Appendix), these experiments yielded high accuracy across both models and datasets, achieving perfect accuracy for all implementations except for the implementation of DT on the BREAST CANCER dataset. Notably, both models perform comparably well under most circumstances, and there is often no perfect model choice for every scenario. As a result, fine-tuning the respective models using various cost/distance functions and other tools such as value-weighting proves critical for determining the relative performance of a given model under certain conditions and preparing for optimal model selection.

6 Statement of Contributions

Data cleaning, KNN implementation, Report - Bonaventure Dossou
Data cleaning, DT implementation, Report - Mat Vallejo
Review and Report - Minjae Kim

7 Appendix

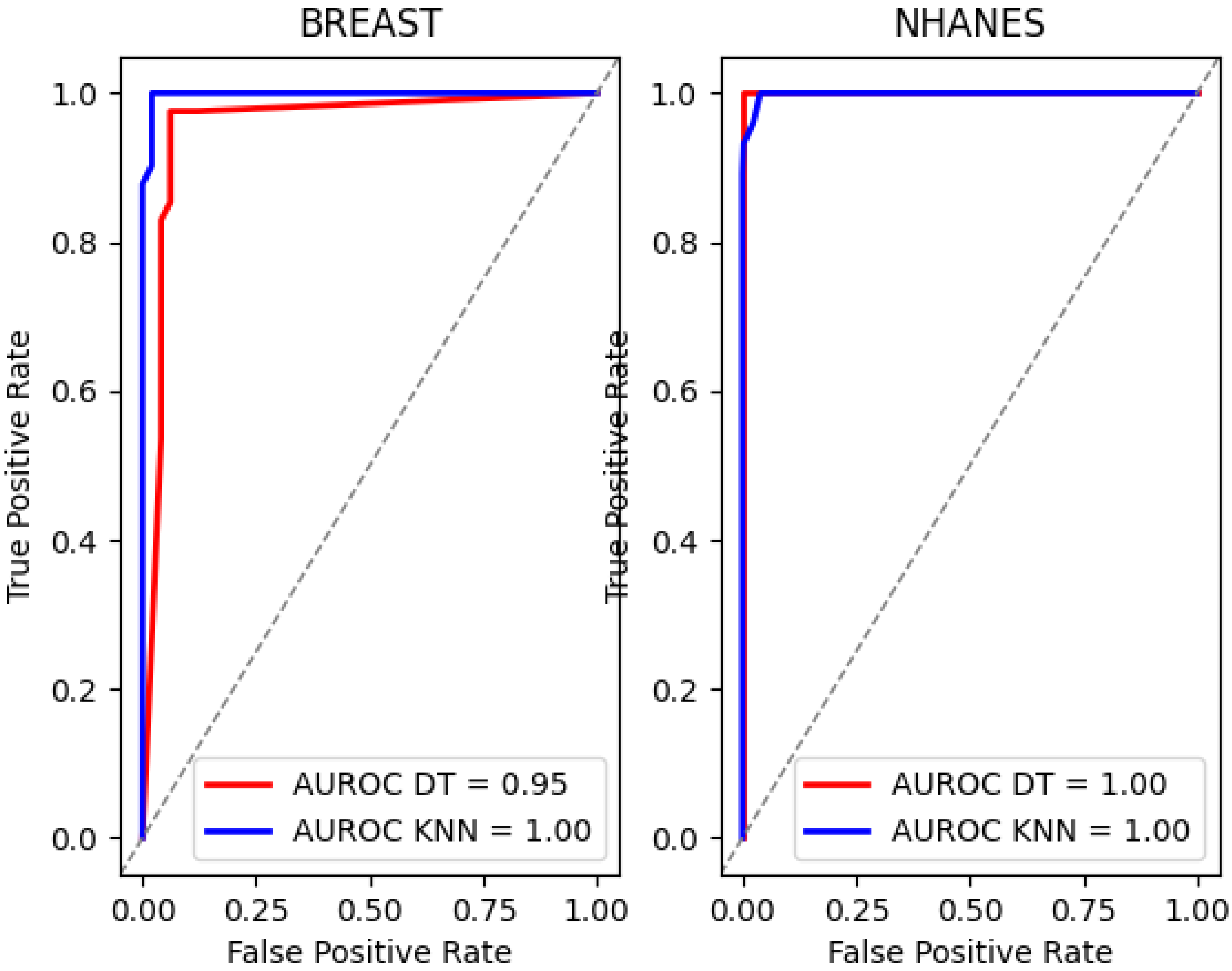

Figure 1: Comparison of AUCROC scores KNN vs DT cross both datasets