COMP 551: Assignment 1 Report

Emile Riberdy (260985547), Shubham Vashisth (261155310), Mohaddeseh Yaghoubpour (261094819)

31/Jan/2024

Abstract

This study evaluates K-Nearest Neighbors (KNN) and Decision Trees (DT) against the Breast Cancer Wisconsin (BC) and NHANES datasets. While both models excel with BC data, NHANES analysis is complicated by feature overlaps among age groups. Preprocessing included data visualization through pair plots and statistics for positive and negative class groups, alongside the removal of missing data. Emphasizing hyperparameter tuning's significance and employing metrics like the F1-Score, this research underlines strategies for enhancing model accuracy in datasets with class imbalances.

1 Introduction

This study delves into the comparative analysis of K-Nearest Neighbors (KNN) and Decision Trees (DT) algorithms, applied to two prominent datasets: the National Health and Nutrition Examination Survey (NHANES) and the Breast Cancer (BC) Wisconsin (Original) dataset. The NHANES dataset, a rich repository of public health information, is critical for evaluating age-based health indicators [3]. Conversely, the BC dataset is pivotal in oncology for distinguishing between benign and malignant tumors through image-derived features [2].

We commenced our exploration with data preprocessing, followed by an in-depth examination of feature interrelations. Following this, we implement the KNN and DT algorithms, tuning them to optimize performance on the given datasets. Feature significance was quantified using Mutual Information scores, guiding our feature selection and model refinement strategies [1]. Our report concludes with a presentation of our findings and a discussion on their implications for predictive modeling in healthcare.

2 Methods

Our study employs two machine learning techniques for predictive analysis:

K-Nearest Neighbors (KNN): A method that classifies each new instance based on the majority vote of its 'k' nearest neighbors from the training set. It operates under the premise that similar instances are likely to be in close proximity. The algorithm's effectiveness is contingent on the choice of 'k' and the distance metric used.

Decision Trees (DT): An approach that models decisions as a tree structure, segmenting the feature space into regions with homogenous labels. Nodes represent decision points, and leaves represent outcomes. DTs are favored for their interpretability but can be prone to overfitting without proper pruning.

3 Datasets

Our study investigates two datasets: the BC Dataset, with measurements from breast mass images for cancer diagnosis, and the NHANES dataset, featuring a suite of health indicators aimed at age classification.

Our exploratory analysis, shown in Figure 1, indicates clear separability between benign and malignant cases within the Breast Cancer dataset. Conversely, the NHANES dataset displays complex feature overlaps between age groups, posing a more intricate classification challenge. The class distribution in the Breast Cancer dataset includes 458 benign instances against 241 malignant cases. For the NHANES dataset, there are 1914 adult instances in comparison to 364 seniors.

We partitioned the datasets into training and testing sets to validate the predictive performance of our models.

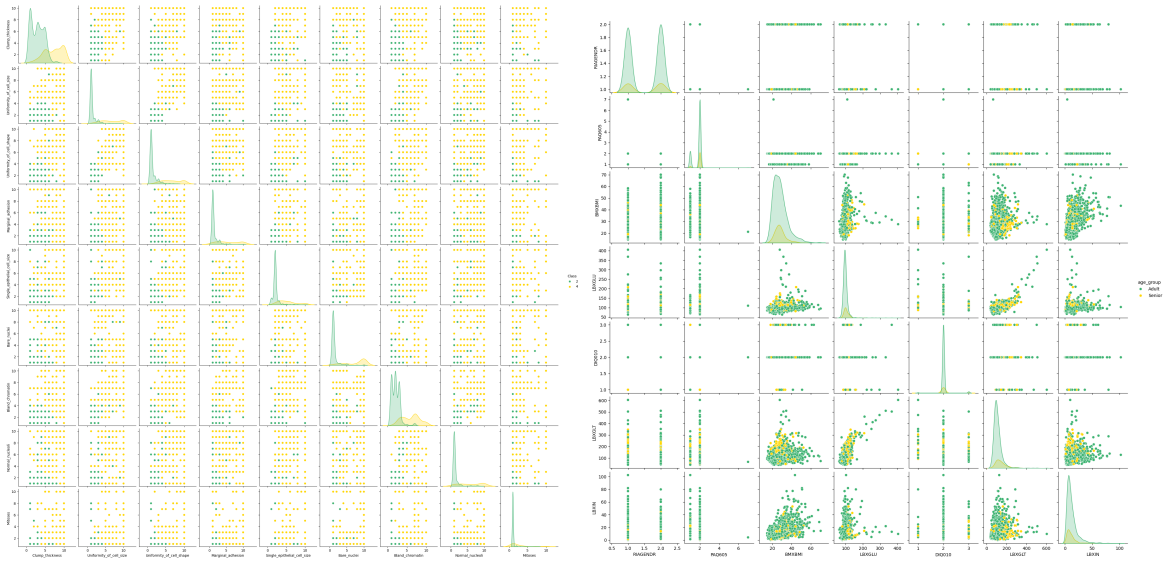


Figure 1: Pairplot visualizations for the BC dataset (left) and NHANES dataset (right).

4 Results

In this section, we will present the results obtained from both algorithms and address the questions outlined in Task 3 of the assignment.

4.1 Q1: Accuracy and AUROC Comparison of KNN and DT

Comparing the performance metrics of K-Nearest Neighbors (KNN) and Decision Trees (DT) on the NHANES and BC datasets provides insights into the efficacy of these algorithms. Figures 2, and 3 present a comprehensive overview of the test accuracy and Area Under the Receiver Operating Characteristic (AUROC) for each algorithm and dataset.

For the BC dataset, both KNN and DT algorithms exhibit a high level of test accuracy, each achieving a score of 0.96, as shown in Figure 3. This suggests that both algorithms are highly effective for this dataset. Similarly, the AUROC for both algorithms is 0.98, indicating excellent discriminative ability between the classes.

In contrast, the NHANES dataset shows a slight variation in performance between the two algorithms. The KNN achieves a slightly higher test accuracy of 0.84 compared to the DT's 0.83, as observed in Figure 2. The AUROC values are close, with KNN at 0.68 and DT at 0.69, indicating comparable classification performance.

These results underscore the importance of algorithm selection based on dataset characteristics and the performance metrics of interest.

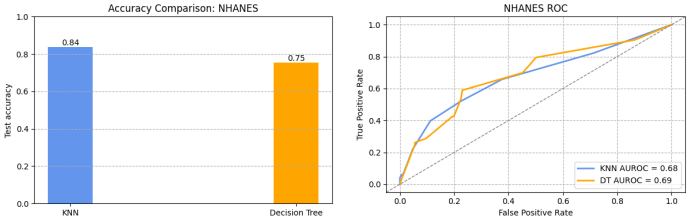


Figure 2: NHANES Accuracy and ROC

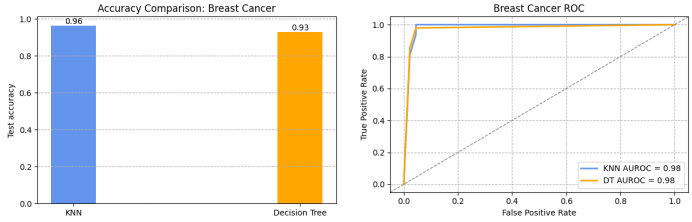


Figure 3: BC Accuracy and ROC

4.2 Q2: Impact of 'K' on KNN Accuracy

The efficacy of the K-Nearest Neighbors (KNN) algorithm is closely linked to the choice of K, the hyperparameter defining the count of neighbors influencing predictions. Through empirical evaluation demonstrated in Fig. 4, we observed that low K values yield high training accuracy due to tight fitting to training data, potentially leading to overfitting as indicated by diminished test accuracy in the NHANES dataset. Conversely, increased K values

enhance generalization, a trend that stabilizes test accuracy even with larger K, particularly noted in the BC dataset due to its clearer class distinctions. This analysis highlights the criticality of selecting an appropriate K to ensure a balance between overfitting and underfitting, crucial for model's generalization capability across varied datasets.for KNN accuracy over K when k=1 training accuracy is 100% for both dataset because for training its actually referring to the same instance which its trying to predict hence 100%.

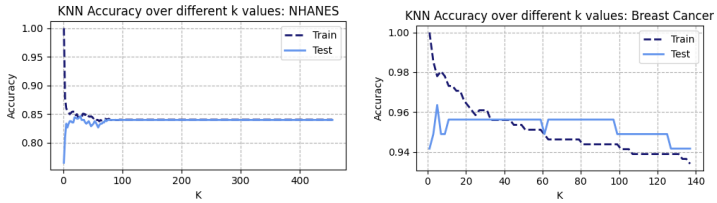


Figure 4: KNN accuracy over K for NHANES and BC.

4.3 Q3: The effect of maximum tree depth on the performance of DT.

The maximum depth of a Decision Tree (DT) is an essential hyperparameter that significantly influences the model's complexity and its generalization capabilities. To understand the effects of this parameter, we analyzed the DT's performance on the NHANES and BC datasets by varying the maximum depth.

As depicted in Fig. 5, for the NHANES dataset, a consistent increase in runtime with increasing tree depth is observed, which indicates a higher computational cost for deeper trees. However, accuracy shows a different trend; it increases initially but then experiences fluctuations and finally decreases sharply, implying overfitting at greater depths.

Similarly, for the BC dataset, as shown in Fig. 6, the runtime increases with the depth of the tree, while accuracy initially increases and then decreases, indicating a peak point of optimal depth before the model starts to overfit.

These findings reinforce the need for careful tuning of the tree depth in DT models to achieve a balance between learning the training data patterns and maintaining the ability to generalize well to unseen data.

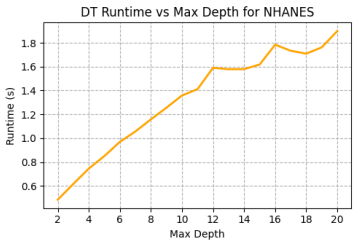
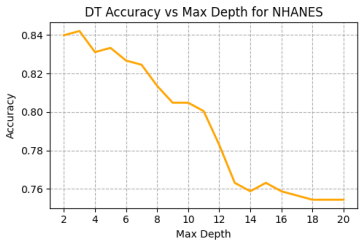
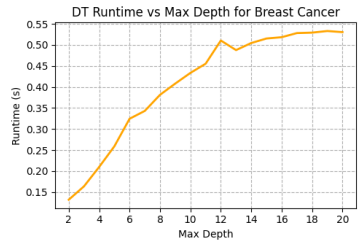


Figure 5: DT runtime and accuracy vs max depth for NHANES





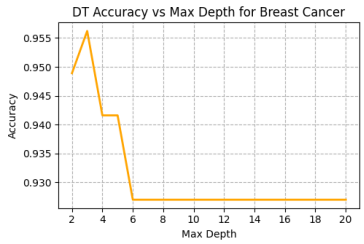


Figure 6: DT runtime and accuracy vs max depth for BC

4.4 Q4: Different distance/cost functions for both models.

Exploring different distance functions for K-Nearest Neighbors (KNN) and cost functions for Decision Trees (DT) can reveal how these choices impact the models' performance. We evaluated the performance of KNN using Euclidean, Manhattan, Minkowski, and Cosine similarity distance functions, and the performance of DT using cost functions such as misclassification, entropy, and the Gini index.

For KNN applied to the NHANES dataset, the various distance functions yield similar Area Under the ROC Curve (AUROC) values, with slight variations. In contrast, for the BC dataset, the Manhattan distance function shows marginally better AUROC, indicating better performance. Similarly, for DT, the choice of cost function affects performance on the NHANES dataset, with misclassification rate resulting in a lower AUROC, suggesting less effective split quality. However, all cost functions perform well on the BC dataset, with entropy and the Gini index showing the highest AUROC.

These findings underscore the importance of choosing suitable distance and cost functions to optimize model performance based on the dataset's characteristics.

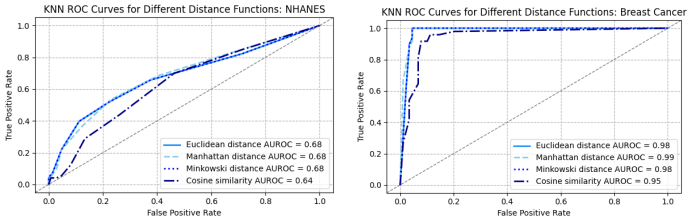


Figure 7: KNN ROC for NHANES (left) and Breast Cancer (right)

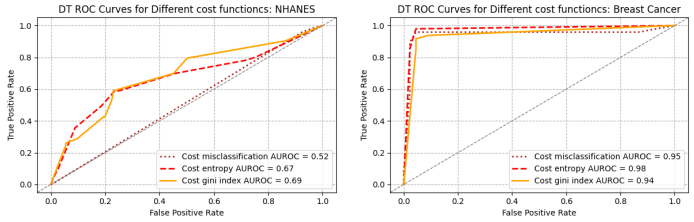


Figure 8: DT ROC for NHANES (left) and Breast Cancer (right)

4.5 Q5: ROC Analysis for KNN and DT

The ROC curve serves as a pivotal tool in evaluating binary classification models by plotting the true positive rate against the false positive rate under varied threshold settings. The AUROC quantitatively captures overall model performance. As detailed in Fig. 9, ROC curves for both KNN and DT across NHANES and BC datasets reveal nuanced insights into model effectiveness. Specifically, AUROC scores—0.68 for KNN and 0.69 for DT on NHANES—indicate nearly equivalent model performance. In contrast, the BC dataset sees both models achieving high AUROC scores of 0.98, signifying exceptional discriminative capacity. Such outcomes not only affirm the models' capabilities in distinguishing positive from negative classes in both contexts but also highlight a pronounced disparity in performance favoring the BC dataset. This disparity showcases the critical importance of model and dataset compatibility, underscoring how KNN and DT excel in scenarios with distinct class separability, as exemplified by the BC dataset.

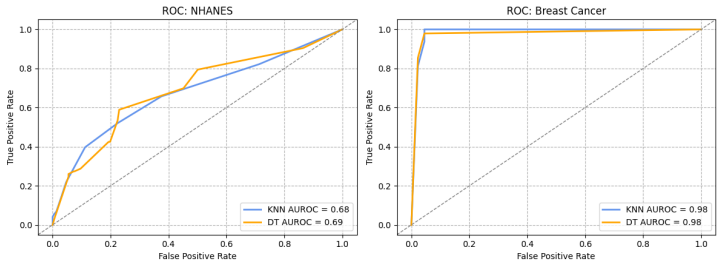


Figure 9: Comparison of ROC curves for KNN and DT on NHANES and BC datasets, illustrating the models' discriminative performances with corresponding AUROC values.

4.6 Q6: Key Feature Selection in KNN

Feature selection plays a pivotal role in the performance of K-Nearest Neighbors (KNN), by identifying the most predictive features and eliminating redundancy. We utilized mutual information to select key features that share the highest amount of information with the target variable.

Fig. 10, indicates the significance of each feature for the NHANES dataset the mutual information scores. It also shows the impact of the number of selected features on the KNN model's accuracy. A similar analysis for the BC dataset is depicted here, where certain features drastically influence the accuracy of the model.

These figures demonstrate the importance of feature selection in machine learning, directly affecting the efficacy and interpretability of the resulting models.

4.7 Q7: Decision Tree Feature Importance

The importance of features within a Decision Tree (DT) can be assessed by counting the occurrences of each feature in non-leaf nodes, which gives us a rough measure of feature importance. This approach can highlight the most significant features for making splits in the DT and hence, for the classification task.

The top five features according to their occurrence in non-leaf nodes of NHANES and BC datasets are shown in Fig. 11. These features are key to the decision-making process within the tree and may differ from those identified by the simple mean difference approach discussed in subtask 3 of Section 2. The difference in feature importance

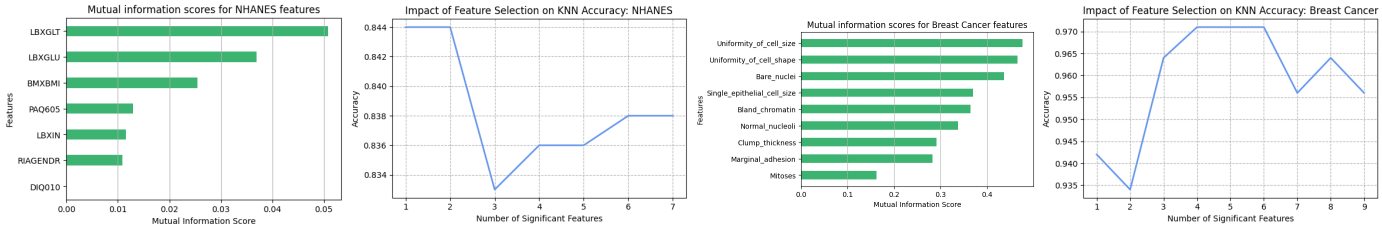


Figure 10: Left to right: NHANES Mutual Information, NHANES Accuracy, Breast Cancer Mutual Information, Breast Cancer Accuracy.

rankings can arise due to the DT's inherent method of evaluating feature effectiveness at reducing classification entropy, as opposed to the univariate statistical measure of mean difference.

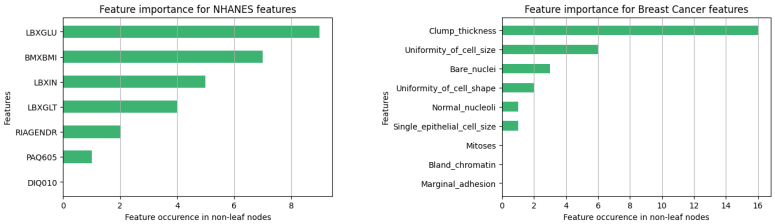


Figure 11: Feature importance for NHANES features (left) and BC features (right).

4.8 Additional experiment

As shown in Figures 12 and 13, for additional tests, we computed a Confusion Table and F1-Score for NHANES due to its imbalanced nature. Accuracy alone may not adequately reflect the model's performance when dealing with imbalanced datasets. The Confusion Table provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions, offering insights into how well the model distinguishes between classes. F1-Score, a metric that considers both precision and recall, is particularly useful in imbalanced scenarios, offering a balanced assessment of the model's ability to correctly identify positive instances while minimizing false positives and false negatives. (KNN F1-Score for NHANES = 0.21 and for BC = 0.95, and DT F1-Score for NHANES = 0.32 and for BC = 0.95.) These supplementary metrics provide a more comprehensive understanding of the model's effectiveness, especially in scenarios where accurate classification of minority classes is critical like NHANES.

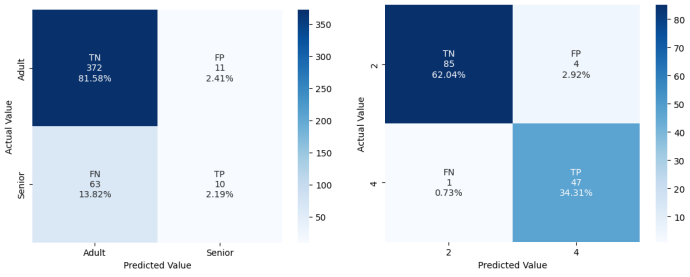


Figure 12: KNN F1-Score confusion tables for NHANES (left) 0.21 and BC (right). NHANES F1= 0.21 and BC F1 = 0.95.

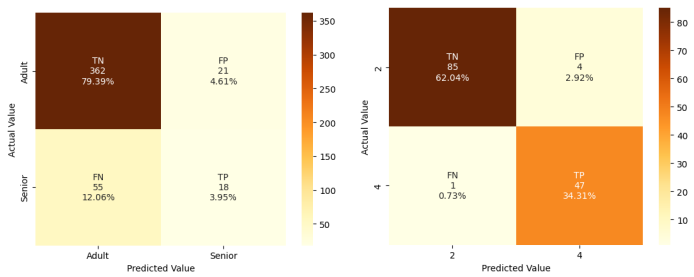


Figure 13: DT F1-Score confusion tables for NHANES (left) and BC (right). NHANES F1= 0.32 and BC F1 = 0.95.

5 Statement of Contributions

Each team member actively participated in all aspects of the assignment, utilizing in-person meetings, online discussions, Slack, and shared platforms like Colab, GitHub and Overleaf. However, there was a slight emphasis on specific areas: Shubham took the lead in KNN, Emile in DT, and Mohaddeseh in the report.

References

[1] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[2] William H Wolberg, W Nick Street, and Olvi L Mangasarian. Statistical features of cell nuclei in breast tissue. *Cancer letters*, 77(2-3):163–171, 1994.

[3] George Zipf, Michele Chiappa, Kathryn S Porter, Yechiam Ostchega, Brenda G Lewis, and Jennifer Dostal. National health and nutrition examination survey: Plan and operations, 1999-2010. *Vital and health statistics. Series 1, Programs and collection procedures*, (56):1–37, 2013.