COMP 551 A1: *Getting Started with Machine Learning*

Rebecca Mizrahi (260975001) rebecca.mizrahi@mail.mcgill.ca
Software Engineering

Amélie Barsoum (290988658) amelie.barsoum@mail.mcgill.ca
Software Engineering

Samantha Handal (260983914) samantha.handal@mail.mcgill.ca
Software Engineering

1. Abstract

In this study, we explored the performance of K-Nearest Neighbor (KNN) and Decision Trees (DT) algorithms on two health datasets: the NHANES age prediction dataset and the Breast Cancer Wisconsin dataset. Through implementation and analysis, we aimed to apply our learnings so far in machine learning; specifically, in data pre-processing using the pandas library, model implementation of two well known algorithms, and performance evaluation through conducting different experiments on the data. Our findings gave us valuable insight into the strengths and weaknesses of each model, and how they compare when applied across different datasets and scenarios. While KNN excels in scenarios with less complex and smaller datasets with fewer dimensions, DTs show superior performance in handling more complex data structures and larger datasets with more variables and more extensive data points by breaking down the data into simpler decision paths. This was exemplified in our analysis, in which the relatively less complex Breast Cancer Wisconsin dataset with fewer dimensions achieved more accurate classification when handled by KNN as opposed to DTs, and the greater-dimensional and more varied NHANES dataset achieved the best results with DTs.

2. Introduction

This assignment entailed implementing and comparing two fundamental machine learning models, K-Nearest Neighbor (KNN) and Decision Trees (DT), on two health datasets. The first dataset pertained to age prediction based on health and nutrition survey data (NHANES 2013-2014), while the second dealt with breast cancer classification. Our objective was to not only implement these algorithms from scratch but also to understand their behavior and performance on real-world data. Comparitive analyses of KNN and DTs have been published already. Further, these analyses have been applied to our exact datasets (NHANES and Breast Cancer Wisconsin), giving us valuable insight into what our accuracy results should look like after applying KNN and DTs, as well as what our ROC and AUROC curves should look like. The results presented in [2] confirm our findings showing that KNN has higher accuracy on the breast cancer dataset, while [1] confirms that DTs perform better than KNN for the NHANES dataset when the data has not been processed yet. Overall, the simplicity of KNN and its reduced risk of overfitting make it good for datasets where each feature heavily impacts the overall decision-making process, while DTs offer a clear decision path and the ability to handle non-linear relationships between features, making them suitable for complex datasets where feature interactions are important.

3. Methods

The K-Nearest-Neighbour and Decision Tree machine learning methods are relatively simple and intuitive algorithms used for classification.

K-Nearest Neighbors machine learning algorithm used for both classification and regression tasks. The main idea behind KNN is to make predictions based on most common or average value of the k-closest data points in the feature space. Calculating closeness is accomplished with a distance metric; for example, Euclidean or Manhattan distance. It is an instance-based, supervised learning classifier that has proved an effective machine learning method despite its simplicity.

Decision Trees are another versatile machine learning algorithm used for classification and regression. The algorithm recursively splits the dataset into subsets based on the most significant feature at each step. These splits form a tree-like structure, where each internal node represents a decision based on a feature, each branch represents an outcome of that decision, and each leaf node represents the final predicted class or value. The goal is to create a tree that effectively partitions the data, making predictions by traversing the tree from the root to a leaf. Decision Trees are interpretable, easy to understand, and can capture complex relationships in the data.

4. Datasets

There are two datasets analyzed with the aforementioned machine learning processes. In order to perform classification with the highest level of accuracy, the data must first be understood. To do this, the mean of each feature in the positive and negative groups were calculated, followed by the squared difference of group means to rank features, and finally a correlation matrix was created. The squared differences are positive values, and higher values indicate larger differences between group means. Correlation values range from -1 to 1 to identify the strength and direction of the linear relationship between each feature and the target. Correlation is more about the overall relationship, while squared differences highlight features where group means significantly differ, which may be important for classification. Both are taken into account to analyze the most and least important features.

The breast cancer dataset analysis indicated that the features in the positive group generally had higher mean values, potentially suggesting more aggressive characteristics than its lower counterparts in the negative group. Features like Bare_nuclei, Uniformity_of_cell_size, and Uniformity_of_cell_shape have higher squared differences of group means, indicating potential importance in distinguishing between classes. Both feature ranking and correlation analysis confirmed that all features show a significant difference between the positive and negative groups, and thus should all always be taken into account.

As for the NHANES dataset, adult features have higher mean values for features like LBXGLT and LBXGLU, whereas senior features have lower for the same. Both correlation and squared differences indicate that the features with low correlation to the target variable (age_group) are RIAGENDR, PAQ605, BMXBMI, DIQ010, and LBXIN. The correlation matrix still indicates an above 0 correlation, but relatively very low for these 5 feature groups.

5. Results

5.1 Accuracy and AUROC of KNN and DT

For the Breast Cancer Dataset, KNN with Euclidean distance shows consistently high performance with an average accuracy of 96%, but a low AUROC of 0.5 suggests issues with class discrimination or balance. In contrast, the NHANES dataset sees moderate success with KNN, achieving an average accuracy of 81.1% and a better-than-random AUROC of 0.645, indicating room for improvement in distinguishing age groups.

The DT model on the Breast Cancer dataset performs strongly, with average accuracies around 93.9% and a high AUROC of 93.8%, demonstrating effective class differentiation. However, on the NHANES dataset, the DT model shows moderate accuracy around 82.7% but a lower AUROC of 54.1%, suggesting limited effectiveness in differentiating outcomes.

5.2 Different K values impact on KNN

Analyzing the KNN performance graphs for both the Breast Cancer and NHANES datasets reveals their different behaviors affected by the choice of k. In the Breast Cancer dataset graph in Figure 2, there is a notable variance in testing accuracy with different k values, peaking at k=5, which suggests a specific k value can significantly enhance model generalization for this dataset. The training accuracy shows a decline as k increases, as most likely due to less overfitting.

For the NHANES dataset, the graph in Figure 1 shows a mostly flat testing accuracy curve across a range of k values, indicating that for this larger and more complex dataset, the choice of k does not heavily impact the model's performance on unseen data. The training accuracy, however, decreases consistently as k increases, suggesting an improvement in generalization but also a potential loss of detail from the training data.

In summary, the Breast Cancer dataset graph shows the sensitivity of small and less complex datasets to the choice of k, where optimal performance can be achieved with the right k. In contrast, the NHANES dataset graph suggests that larger, more complex datasets may require a broader approach to neighbor selection as they have stable performance across many k values, potentially due to inherent noise that overshadows the details of neighbor selection.

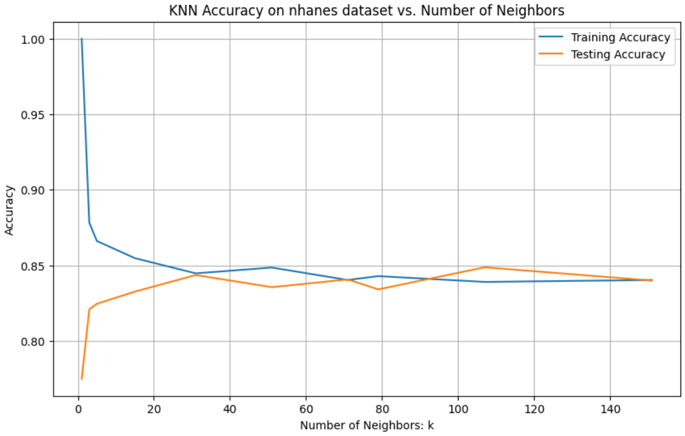

Figure 1: KNN Accuracy on NHANES dataset vs. Number of Neighbors

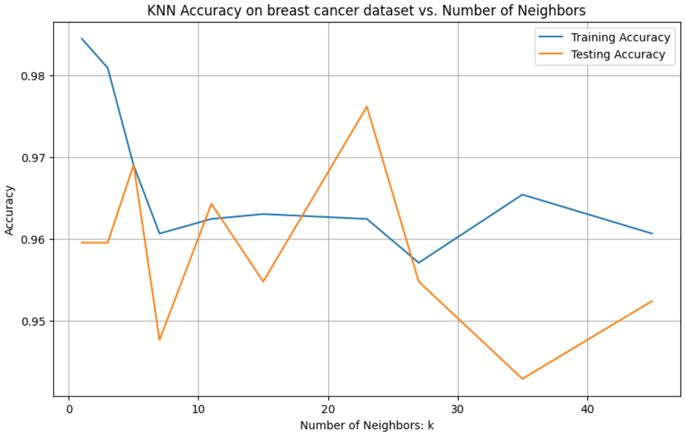

Figure 2: KNN Accuracy on breast cancer dataset vs. Number of Neighbors

5.3 Maximum tree depth impact on DT

Upon examining the graphs for Decision Tree accuracy versus maximum tree depth on both the NHANES and Breast Cancer datasets, we can understand the relationship between tree depth and model performance.

For the Breast Cancer dataset, the graph in Figure 3 shows that as tree depth increases, training accuracy improves, indicating a better fit to the training data. However, the testing accuracy peaks at a depth of 4 and then sharply declines, suggesting that depths beyond 4 lead to overfitting, where the model captures noise instead of underlying patterns.

In contrast, the NHANES dataset graph in Figure 4 displays a more consistent increase in training accuracy with tree depth, plateauing as the depth reaches around 12. The testing accuracy, after fluctuating a bit, remains relatively stable but begins to trend downwards after a depth of 8, hinting at the onset of overfitting, still at a slower rate compared to the Breast Cancer dataset.

The key takeaway from these graphs is that the smaller Breast Cancer dataset (with 9 features) is more sensitive to changes in tree depth, quickly exhibiting overfitting after a certain point, as seen by the drop in testing accuracy. On the other hand, the larger NHANES dataset (with 7 features) shows a more gradual impact to increasing tree depth, with a slower decline in testing accuracy, showing its higher tolerance for complexity before overfitting becomes detrimental.

These observations show that the optimal tree depth is very specific to the dataset. We must find a balance so that the model is complex enough to capture the data patterns, without becoming so complex that it no longer generalizes to unseen data.

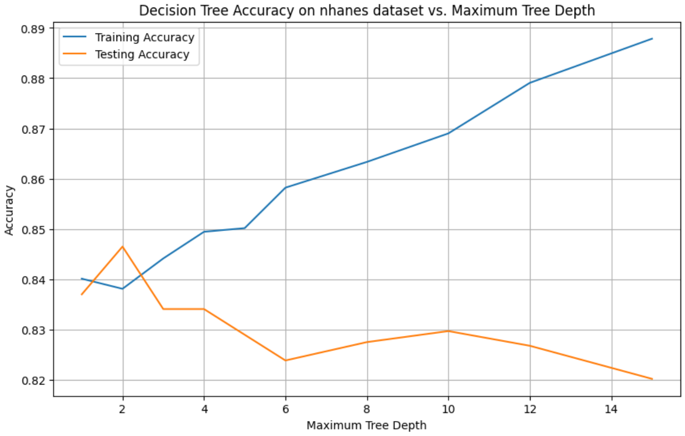

Figure 3: Decision Tree Accuracy on NHANES dataset vs. Maximum Tree Depth

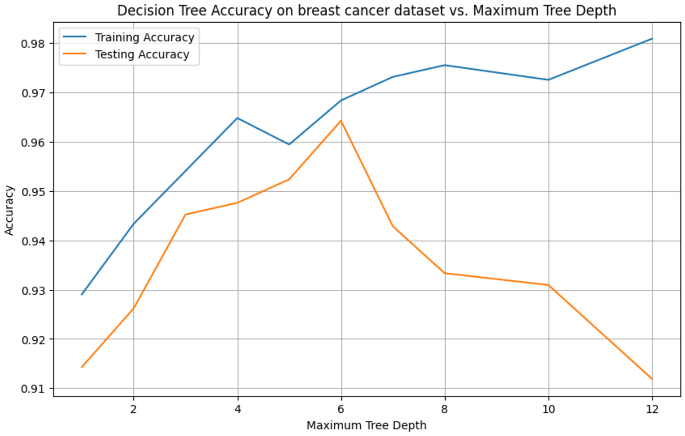

Figure 4: Decision Tree Accuracy on breast cancer dataset vs. Maximum Tree Depth

5.4 Different distance/cost functions

To evaluate the impact of different distance functions on KNN and various cost functions on DT, we conducted multiple runs on both the Breast Cancer and NHANES datasets.

For the Breast Cancer Dataset, KNN with the Euclidean distance function achieved a higher average accuracy (96.43%) compared to the Manhattan distance function (95.29%), showing us that the straight-line measure between data points is slightly more effective for this dataset. Even after removing the least important feature, the Manhattan distance maintained a stable accuracy (95.43%), indicating robustness when reducing features.

Conversely, for the NHANES Age Prediction Dataset, the Manhattan distance function outperformed the Euclidean with an average accuracy of 81.45% against 80.44%, indicating that the sum of absolute differences is more effective for this particular dataset. However, the removal of the least important feature led to a decrease in accuracy (79.43%), demonstrating a sensitivity to the feature set used.

Regarding cost functions for DT, the Misclassification model on the Breast Cancer dataset provided an average test accuracy of 92.57% and an AUROC of 92.01%, indicating good performance. For the NHANES dataset, the same cost function achieved a lower test accuracy of 82.81% and an AUROC of 53.64%, showing the Misclassification model's reduced effectiveness on this more complex dataset.

When using Entropy as the cost function, both datasets showed a slight decrease in test accuracy compared to the Misclassification model, with the Breast Cancer dataset at 92.43% and the NHANES dataset at 82.24%. This suggests that while Entropy is an effective measure, it might not provide significant benefits over Misclassification in these particular datasets.

Finally, the Gini index model showed similar performance to Entropy on the Breast Cancer dataset with a test accuracy of 92.43% but achieved a slightly better AUROC of 92.37%. For the NHANES dataset, the Gini index outperformed both Misclassification and Entropy models with a test accuracy of 83.51% and the highest AUROC of 55.65%, demonstrating its performance on datasets with more complex feature interactions.

In summary, the choice of distance function for KNN and cost function for DT can affect model performance, with Euclidean distance generally performing better on the Breast Cancer dataset and Manhattan distance on the NHANES dataset. For DT, while all three cost functions led to similar outcomes on the Breast Cancer dataset, the Gini index proved to be the most effective on the NHANES dataset. These results emphasize the need for picking the best cost/distance functions based on dataset characteristics to achieve optimal performance.

5.5 ROC for KNN and DT

To compare the performance of KNN and DT models, ROC curves for each model were plotted on the same graph for the Breast Cancer and NHANES datasets. The ROC curve is a graphical representation of the trade-off between the true positive rate (TPR) and false positive rate (FPR) at various thresholds.

For the Breast Cancer dataset, the ROC curve in Figure 6 illustrates that the DT model significantly outperforms the KNN model, proven by a much higher AUROC value of 0.92 for DT compared to an AUROC of 0.50 for KNN. An AUROC of 0.50 indicates that the KNN model's performance is no better than random chance, while the DT model's AUROC of 0.92 reflects power to accurately classify the positive class.

In contrast, the ROC curve for the NHANES dataset in Figure 5 shows that both models perform moderately, with KNN having a slightly higher AUROC value of 0.67 compared to DT's AUROC of 0.56. Neither model is particularly outstanding in this case, but KNN seems to have a slight edge over DT in distinguishing between age groups.

These ROC curves suggest that the DT model is better suited for the Breast Cancer dataset, likely due to its ability to capture the important features that distinguish between benign and malignant classes. However, for the NHANES dataset, the performance is relatively close between KNN and DT, and neither model demonstrates good predictive power. This could be due to the NHANES dataset's more complex feature space, which poses challenges for both models in effectively discriminating between the different age groups.

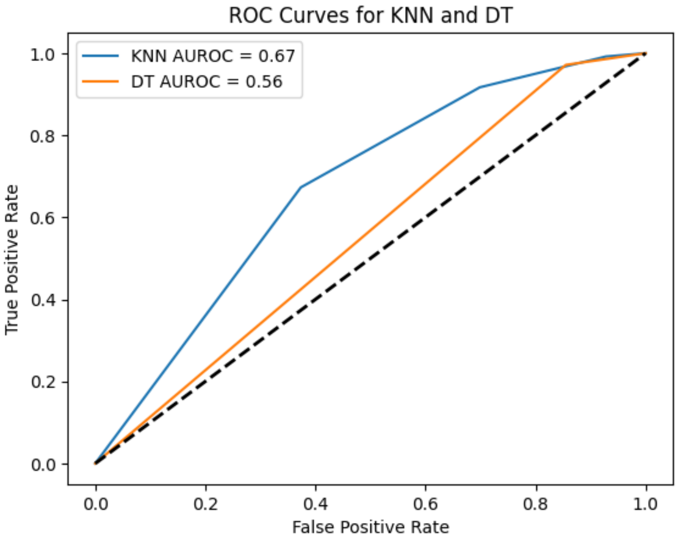

Figure 5: ROC Curves for KNN and DT on NHANES dataset

5.6 Key features used in KNN

As described in the *Dataset* section, the key features were calculated using a correlation matrix. This is done by first calculating the correlation coefficient between each feature and the target. 1 indicates a perfect positive linear relationship. -1 indicates a perfect negative linear relationship. 0 indicates no linear relationship. We may set a threshold to discard features whose correlation falls below the threshold. As mentioned in the *Dataset* section, the breast cancer dataset had a high correlation for all features, but the NHANES dataset had low correlation for the following 6 features:

1.	RIAGENDR	0.002767
2.	BMXBMI	0.004147
3.	DIQ010	0.026399
4.	LBXIN	0.064159
5.	PAQ605	0.094789

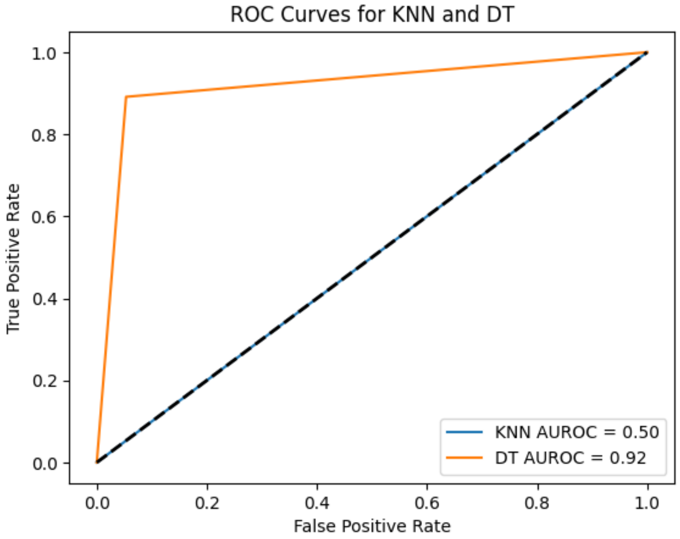

Figure 6: ROC Curves for KNN and DT on breast cancer dataset

In order to determine a threshold, one may test different thresholds, or the accuracy after dropping a certain subset of low-correlation features. Dropping features with low correlation allows for quicker classification, as less columns must be considered, resulting in less computations.

As an experiment, we performed KNN on 6 different versions of the NHANES dataset; the first, after dropping the least-correlated feature RIAGENDR; the second, after dropping both RIAGENDR and the second-least-correlated feature BMXBMI; and so on. The accuracy of each iteration of KNN was averaged over 10 iterations to determine which refined dataset performed the best. The results were:

1. Dropping 0 least-correlated features: 0.8142543859649123
2. Dropping 1 least-correlated feature: 0.813815789473684
3. Dropping 2 least-correlated features: 0.8107456140350877
4. Dropping 3 least-correlated features: 0.8153508771929824
5. Dropping 4 least-correlated features: 0.7929824561403509
6. Dropping 5 least-correlated features: 0.8054824561403506

We conclude that we may drop 3 columns, eliminating a significant portion of the data points and reducing the computations required, without negatively impacting the accuracy- in fact, in our test and reducing the number, the accuracy increased but about 0.11%.

5.7 Rough feature importance score for each feature d

The top 5 most important features for NHANES are: Feature 2 (used 58 times), Feature 3 (used 31 times), Feature 5 (used 24 times), Feature 6 (used 24 times), Feature 0 (used 15 times). For breast cancer dataset they are Clump_thickness (used 12 times), Marginal_adhesion (used 7 times), Uniformity_of_cell_size (used 5 times), Uniformity_of_cell_shape (used 5 times) and Bare_nuclei (used 4 times). However, for the breast cancer dataset, using the simple mean difference approach, we get (in this order): Bare_nuclei, Uniformity_of_cell_size, Uniformity_of_cell_shape, Normal_nucleoli, Clump_thickness. We can observe different results from the rough feature importance calculations for DTs compared to simple mean difference approach.

When comparing the top features identified by the decision tree model and those highlighted by the mean difference approach for the NHANES and breast cancer datasets, we can see that the methods offer different benefits. The decision tree considers the combined effect of multiple features on the prediction. It prioritizes features based on their contribution to the overall model accuracy and the purity of the data split at each node. This can elevate the importance of features that work well in conjunction with others, even if they don't show the largest mean differences between classes on their own.

In contrast, the mean difference approach focuses on the individual power of each feature to distinguish between classes based only on their mean values. This can highlight features that have a strong individual potential but may overlook their potential correlations or redundancies when considered in the context of other features.

For both the NHANES and breast cancer datasets, the overlap in identified features between the two methods suggests that some features are consistently influential, regardless of the analysis technique. However, the differences show the unique contributions of features that might be masked in a univariate analysis but are revealed to be significant through the decision tree's more complete evaluation of feature interactions. This highlights the value of using multiple methods to gain a better picture of feature importance in complex datasets.

5.8 Model selection using a Validation Set

We tried a strategic approach to optimize the hyperparameters of K-Nearest Neighbor (KNN) and Decision Tree (DT) models by dividing our dataset into training, validation, and testing segments. This division enabled us to fine-tune the 'k' value for KNN and the tree depth for DT, ensuring these parameters were chosen based on their performance on unseen data, which would help generalizability. We identified the optimal parameters using the validation set and retrained our models on a combined dataset of training and validation sets. Finally, we assessed the tuned models on the test set to evaluate their effectiveness on new data. This approach to hyperparameter tuning not only improved our models' accuracy but also showed the importance of a systematic process for achieving reliable predictive performance.

6. Discussion and Conclusion

In this assignment, we explored the performance of K-Nearest Neighbor (KNN) and Decision Trees (DT) algorithms on two health datasets: the NHANES age prediction dataset and the Breast Cancer Wisconsin dataset. Our objective was to apply our machine learning knowledge in data pre-processing, model implementation, and performance evaluation. The findings revealed insights into the strengths and weaknesses of each model, highlighting their applicability across different datasets and scenarios.

KNN excelled in scenarios with less complex and smaller datasets, demonstrating its effectiveness in breast cancer classification. On the other hand, DTs outperformed KNN in handling more complex data structures and larger datasets, showcasing their ability to capture non-linear relationships in the NHANES dataset.

The importance of parameter tuning was made clear in our analysis, particularly of different K values for KNN and maximum tree depth for DT. The Breast Cancer dataset showed sensitivity to the choice of k, while the NHANES dataset exhibited stability across a range of k values. For DT, the optimal tree depth varied between datasets, emphasizing the need to balance model complexity.

The impact of distance/cost functions on model performance was also shown to have an impact on accuracy, with Euclidean distance performing better for KNN on the Breast Cancer dataset and Manhattan distance for the NHANES dataset. DTs demonstrated varying performance with different cost functions, highlighting the need to select the most suitable function for each dataset.

ROC curves illustrated the trade-off between true positive and false positive rates for KNN and DT. DTs outperformed KNN in the Breast Cancer dataset, emphasizing their ability to discriminate between classes. However, the NHANES dataset posed challenges for both models, with KNN slightly outperforming DT in ROC analysis.

In experimenting with these two methods, we gained important knowledge on how we should approach machine learning models. Tweaking variables and trying different variations is a crucial step in finalizing a model. The importance of this step and the time commitment needed relative to designing the model was made clear by this assignment.

In conclusion, this assignment provided valuable insights into the application of KNN and DT algorithms on health datasets. The choice of algorithm, parameter tuning, and distance/cost functions significantly impacted model performance. The comparison of feature importance between KNN and DT highlighted the different perspectives offered by these models. Future investigations could explore more advanced machine learning techniques, feature engineering methods, and ensemble learning to further enhance predictive capabilities.

7. Statement of Contributions

For this project, all three of us brainstormed together before starting to decide which experiments we would be conducting. Once decided, Amélie and Rebecca worked on tasks 1 and 2 together, which were data acquisition, preprocessing and analysis, as well as DT and KNN implementation. Samantha then implemented the suite of experiments for task 3, as well as commented and adjusted the formatting of our Notebook so that it was ready for submission. All three team members then created an outline for the report together, then Amélie and Rebecca wrote it up.

References

[1] Ashley L. Merianos, E. Melinda Mahabee-Gittens, Timothy M. Stone, Roman A. Jandarov, Lanqing Wang, Deepak Bhandari, Benjamin C. Blount, and Georg E. Matt. Distinguishing exposure to secondhand and thirdhand tobacco smoke among u.s. children using machine learning: Nhanes 2013–2016. *Environmental Science & Technology*,

57(5):2042–2053, 2023. doi: 10.1021/acs.est.2c08121. URL `https://doi.org/10.1021/acs.est.2c08121`. PMID: 36705578.

[2] O. I. Obaid, M. A. Mohammed, M. K. A. Ghani, A. Mostafa, and F. Taha. Evaluating the performance of machine learning techniques in the classification of wisconsin breast cancer. *International Journal of Engineering & Technology*, 7(4.36):160–166, 2018.