551 MiniProject1

Fynn Schmitt-Ulms, Joon Hwan Hong, and Daniel Korsunsky

February 5, 2021

Abstract

In this project, we investigated the performance of two machine learning models on two benchmark healthcare datasets from the UCI Machine Learning Repository [1]. A numpy RNG seed of '551' was set for this exploration. For the breast cancer dataset, we found that the K-Nearest Neighbour (KNN) approach achieved better 5-fold average validation accuracy than the Decision Tree (DT) approach (0.9725 vs. 0.9571). However, DT achieved slightly better test set accuracy (0.99 vs. 0.98) compared to KNN. On the hepatitis dataset, the KNN model outperformed DT in best 5-fold average validation accuracy (0.8846 vs. 0.8461) while DT outperformed in the test set accuracy (0.8 vs. 0.775).

1 Introduction

The Breast Cancer and Hepatitis Datasets used in this analysis have been used in the past to develop DT methodologies. Bredensteiner and Bennett (1996) used the breast cancer dataset for feature minimization with decision trees [2], while Esposito et al. (1997) used the hepatitis dataset for decision tree pruning methods [3] . Similarly, our work attempts to compare the performance of DT and KNN models on these tabular datasets. In addition, we investigated trends in the performance of the models, under varying hyperparameter values.

The datasets were loaded into python as Pandas dataframe objects. 16 datapoints with missing features in the breast cancer dataset were removed, while the missing features found in 75 out of 155 hepatitis datapoints, were replaced with the mean or mode of the feature depending on whether it was continuous or categorical.

The KNN and DT models were implemented as classes with a `fit` and `predict` function. In addition, an `evaluate_acc` function was defined to evaluate model accuracies. For our experiments, different K values and distance functions were explored for the KNN model, while different cost functions and max tree depth were explored for the DT model. The experiments are explained in detail in Section 4. KNN achieved higher accuracy score in the 5-fold average validation accuracy (0.8846 vs. 0.8461) and DT obtained higher test set accuracy (0.8 vs. 0.775) against the KNN model for the hepatitis dataset. KNN achieved better accuracy than the DT approach in term of best 5-fold average validation accuracy (0.9725 vs. 0.9571) for the breast cancer dataset, while achieving a lower test set accuracy value of 0.98 compared to 0.99 of the DT approach.

2 Datasets

2.1 Breast Cancer

The "Bare Nuclei" feature contained 16 missing values; the other features did not appear to have missing values. The ids containing missing values were removed. The feature distribution is available in *Figure 1 (a)* as a matrix of histograms for each feature in the breast cancer dataset. We also explored the linear correlation between different features and the output class, seen in *Figure 1 (b)*.

2.2 Hepatitis

75 out of 155 data-points were missing at least one feature. Due to the high number of missing values, instead of simply removing the rows with missing values which would greatly reduce the size of our dataset, we decided to substitute the missing values with either the mean or mode of the feature (depending on whether the data is continuous or categorical). The distribution of the categorical features is available in *Figure 2 (a)*. Distributions of continuous features are available in *Figure 2 (b)*. We also explored the feature-class correlation in *Figure 2 (c)*.

2.3 Ethical Considerations

There are certainly ethical considerations when working with sensitive medical information, so it is imperative that patient confidentiality is protected. While healthcare data is typically well-protected, this security is certainly not foolproof. We felt comfortable using the data because all patient identities are obfuscated with IDs we do not have access to, and because our objective was to compare model performance, rather than using the information to inform decisions about individuals like qualification for medical treatment or hiring.

Additionally, 139/155 rows in the Hepatitis dataset have sex=1 (male). Due to the strong bias of the dataset, it is not guaranteed that our models, optimized on this data, will generalize well when making predictions for women. This must be kept in mind when interpreting the model's output.

3 Results

In the breast cancer dataset, 100 out of the 683 rows were set aside for testing, compared to 40 out of the 155 rows in the hepatitis dataset. Such a split allowed us to have enough data to train on while still being able to test properly. It is important to note that it was more difficult to create a powerful model for the hepatitis dataset because of it's small sample size (and would be even smaller if the 75 rows with missing features were removed instead of replaced with the feature mean or mode).

3.1 Experiment 1: Algorithm Accuracy

We found that for the breast cancer dataset, the 5-fold average validation accuracy was higher in the KNN approach than that in the DT approach (0.9725 vs. 0.9571), while the test set accuracy was higher in the DT approach (0.98 to 0.99).

For the hepatitis dataset, we likewise found that the 5-fold average validation accuracy was higher in the KNN approach than that in the DT approach (0.8846 vs. 0.8461), while the test set accuracy was higher in the DT approach (0.775 vs. 0.8).

3.2 Experiment 2: K-values

Both datasets were evaluated across a range of K-values from 1 to 10.

For the breast cancer dataset, the maximum 5-fold average validation accuracy occurred at K = 3 using the Euclidean distance function (with the maximum accuracy for the other two distance function also occurring at K = 3). The minimum accuracy for all three distance functions occurred at K = 2. See *Figure 3 (a)* for details.

For the hepatitis dataset, the maximum 5-fold average validation accuracy occurred at K = 8, 9, and 10 using the Manhattan distance function. For the Manhattan and Euclidean distances, the minimum accuracy occurred at K = 2. For K $\geq$ 4, the Manhattan distance outperformed the other two. See *Figure 4 (a)* for details.

3.3 Experiment 3: Maximum Tree Depth

Both datasets were evaluated across a range of $MaxDepths$ from 1 to 10.

For the breast cancer dataset, the maximum 5-fold average validation accuracy occurred at $MaxDepth = 5$ using the Misclassification cost function, which outperformed the other two cost functions across all Max-Depths. For all three cost functions, the accuracy started lowest at $MaxDepth = 1$, then rose consistently and peaked at $MaxDepth = 4$ or $MaxDepth = 5$ before slightly dropping off and remaining fairly constant from $MaxDepth = 5$ to $MaxDepth = 10$. See *Figure 3 (b)* for details.

For the hepatitis dataset, the maximum 5-fold average validation accuracy occurred at $MaxDepth = 1$ using the Misclassification cost function, which outperformed the other two cost functions across all Max-Depths besides $MaxDepth = 2$. The maximum accuracy for all three cost functions occurred at either $MaxDepth = 1$ or $MaxDepth = 2$, after which the Entropy and Gini cost functions dropped off significantly. See *Figure 4 (b)* for details.

3.4 Experiment 4: Distance and Cost Functions

Three different distance functions were tested (Euclidean, Manhattan, Minkowski($p = 4$)). Three different cost functions were tested (Misclassification, Entropy, Gini).

For the KNN model: On the breast cancer dataset, the varying performances of each distance functions are presented in *Figure 3 (a)*. The Euclidean distance function achieved the highest accuracy performance. Minkowski generally underperformed in comparison to the Euclidean and Manhattan distance functions with the exception at $K = 4$ and 5. The Manhattan distance function achieved the highest accuracy score for the hepatitis dataset, seen in *Figure 4 (a)*. The Minkowski function had the least fluctuations between different K values, however it generally had the lowest accuracy score out of the three functions.

For the DT model: On the breast cancer dataset, the varying performances of each cost function are presented in *Figure 3 (b)*. Misclassification consistently outperformed the other two functions. Entropy had the lowest accuracy score. Function performances on the hepatitis dataset is available in *Figure 4 (b)*. A similar result follows: Misclassification outperforms the other functions with the exception at $MaxDepth = 2$, where Gini achieved the highest accuracy score. Entropy again consistently achieved the lowest accuracy score.

3.5 Experiment 5: Decision Boundaries

We plotted the decision boundary for each model using the AGE and BILIRUBIN features from the hepatitis dataset. Both decision boundaries can be seen in *Figure 5*. For the KNN decision boundary, which you can see in *Figure 5 (a)*, we chose K = 1 and the Euclidean distance function as this hyper-parameter choice accentuates the behavior of the decision boundary around points. As expected, the decision boundary for the DT is very simple, with two dividing lines that split the sample space into four sections. The KNN boundary is much more complex (especially with K=1) which causes the predicted class of each region to be mapped to the class of the closest training point.

4 Discussion and Conclusion

While the KNN and Decision Tree models are fairly simple, they yielded extremely precise predictions in classifying test data (the two models achieved 0.98 and 0.99 test set accuracy, respectively, on the breast cancer dataset). Their performance was also comparable; neither model significantly or consistently outperformed the other. It was also imperative that we had a good understanding of the data and features being tested in order to appreciate the significance of the models' results, especially since we were working with sensitive and consequential patient healthcare data.

One of the biggest shortcomings of our analysis was its reliance on very limited biased data. In the future, we hope to extend our investigation to use larger datasets with more diverse subjects to improve generalization. In particular, the small Hepatitis dataset required us to use a particularly small test set in order to leave enough data for training. The result was very low precision in the reported test set accuracies.

5 Statement of Contributions

Contributions by Schmitt-Ulms: Cleaning data; K-fold; Decision boundaries; Report writing.
Contributions by Hong: KNN model; Distance functions; Report writing.
Contributions by Korsunsky: DT model; Cost functions; Report writing.

References

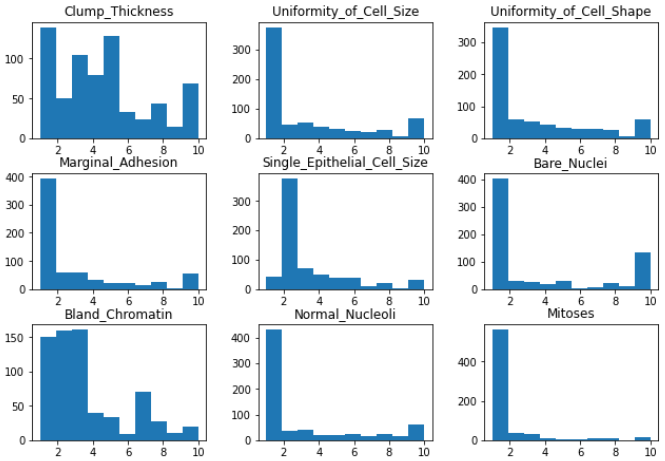
[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Erin J. Bredensteiner and Kristin P. Bennett. Feature Minimization within Decision Trees. National Science Foundation. 1996.

[3] Floriana Esposito and Donato Malerba and Giovanni Semeraro. A Comparative Analysis of Methods for Pruning Decision Trees. IEEE Trans. Pattern Anal. Mach. Intell, 19. 1997.

Figure 1: Breast Cancer Data Exploration

(a) Breast Cancer Feature Distributions



(b) Breast Cancer Feature-Class Correlation

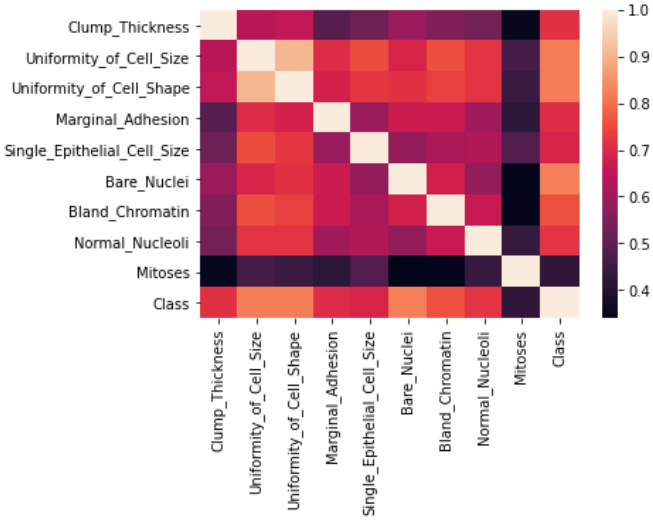


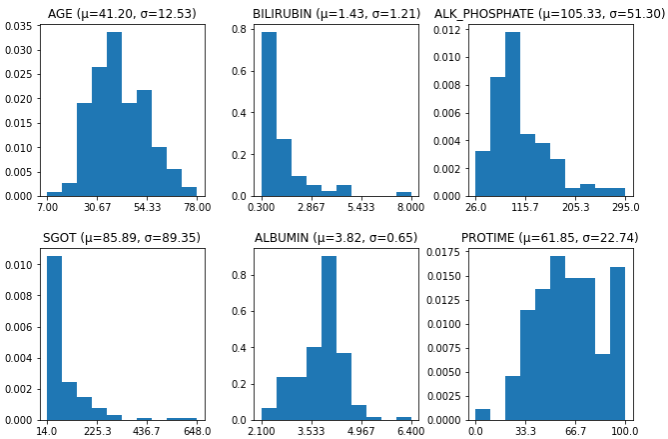
Figure 2: Hepatitis Data Exploration

(a) Hepatitis Categorical Feature Distributions

Distribution of Categorical Values

Feature	1	2	?
SEX	139	16	0
STEROID	76	78	1
ANTIVIRALS	24	131	0
FATIGUE	100	54	1
MALAISE	61	93	1
ANOREXIA	32	122	1
LIVER_BIG	25	120	10
LIVER_FIRM	60	84	11
SPLEEN_PALPABLE	30	120	5
SPIDERS	51	99	5
ASCITES	20	130	5
VARICES	18	132	5
HISTOLOGY	85	70	0

(b) Hepatitis Continuous Feature Distributions



(c) Hepatitis Feature-Class Correlation

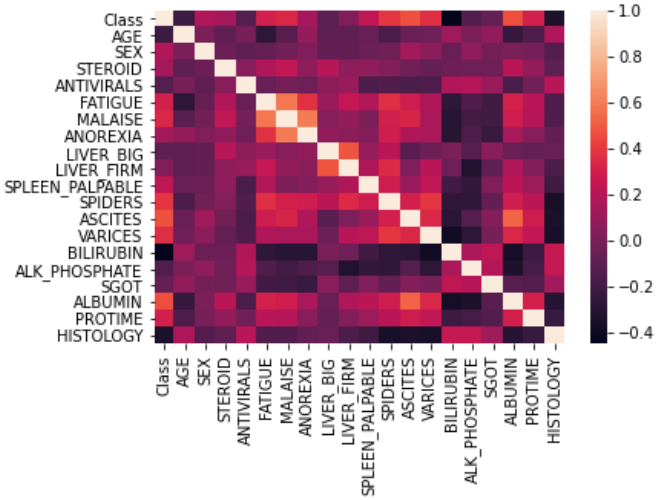
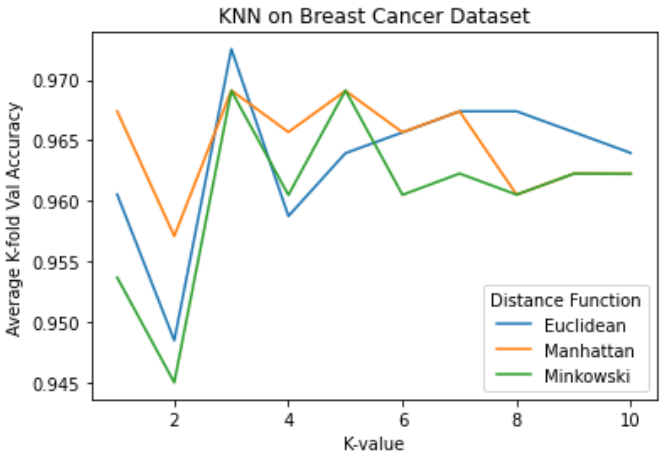


Figure 3: Model Performances on Breast Cancer Dataset

(a) KNN on Breast Cancer Dataset



(b) Decision Tree on Breast Cancer Dataset

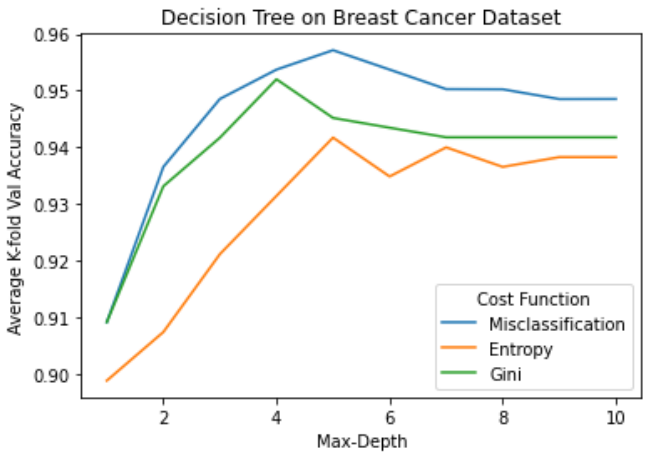
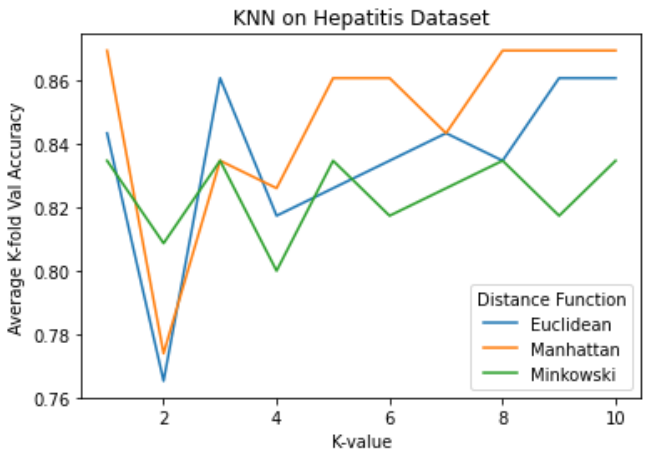


Figure 4: Model Performances on Hepatitis Dataset

(a) KNN on Hepatitis Dataset



(b) Decision Tree on Hepatitis Dataset

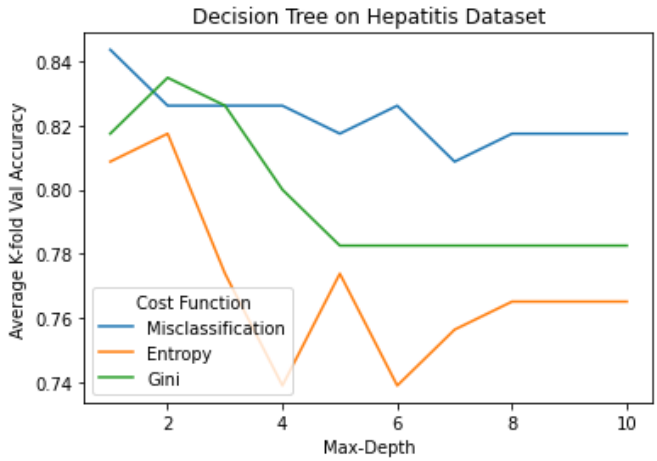
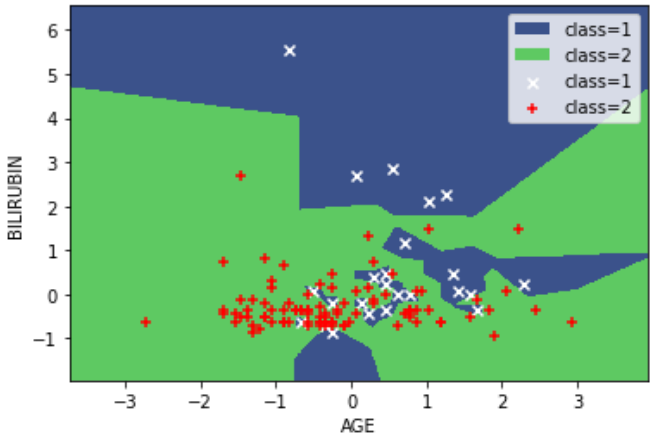


Figure 5: Decision Boundary of the Hepatitis Dataset

(a) KNN H Decision Boundary



(b) DT H Decision Boundary