Abstract

In this report, the performance of supervised machine learning models, K-Nearest Neighbors, weighted K-Nearest Neighbors, and Decision Tree was investigated on two benchmark data sets: Hepatitis Data Set and Diabetic Retinopathy Debrecen Dataset. It was found that the K - Nearest Neighbor (KNN) algorithm achieved a 3.1% better accuracy than the Decision Tree model on the Hepatitis dataset and 0.8% better on the Diabetic Retinopathy Debrecen dataset. KNN was also 3.1% more accurate than Weighted KNN on the Hepatitis dataset and 2.6% the DRD dataset.

1. Introduction

The goal of this experiment was to use the K-Nearest Neighbors (K-NN), weighted K-NN, and Decision Tree (DT) machine learning (ML) models to make simple diagnoses of diseases based on patients' health data. The experiment was performed on two health datasets, namely Hepatitis and the Diabetic Retinopathy Debrecen (DRD). Before starting the report, it was useful to review papers that investigated similar datasets in order to get a general idea of the accuracy numbers the algorithms could obtain. [1] concluded that on the DRD dataset, the K-NN classifier performs best at an accuracy of 76%, while DT has a performance accuracy of 69%. For the Hepatitis dataset, [2] reported a K-NN accuracy of 85.29%, while [3] reported a DT accuracy of 82.05%. In both cases, the K-NN outperformed DT. The results of this experiment concluded that on a feature scaled and filtered Hepatitis dataset, K-NN outperforms both weighted K-NN and DT at a reported accuracy of 90.6%. For the feature filtered DRD dataset, K-NN also performs best with a prediction accuracy of 65.1%. The detailed results of each model performance can be found in the Reports, Part 1 section.

2. Methods

The K-Nearest neighbors (K-NN) algorithm is an exemplar-based classification technique that works by "remembering" the training data set it is provided and performing a categorical output prediction on a new data point based on the labels from the K-most similar examples in the training set. The weighted K-NN is similar to K-NN but in addition to calculating the K distances from the test point, the nearest neighbors are assigned the highest weight. The models are sensitive to feature scaling and noise, do not do well with multivariate datasets, and work poorly on high-dimensional data.

Unlike K-Nearest Neighbors, the decision tree algorithm (DT) learns the model from the data. The main positive aspects of the DT model is that it is capable of handling multivariate inputs well, is insensitive to feature scaling and outliers, performs automated variable selection, and can easily handle missing input features. The best DT model is chosen based on how it splits the data with a specified cost function. It is important to note that the features and thresholds at each split in the tree node are chosen to minimize the cost function.

3. Datasets

Hepatitis Data Set

Hepatitis is an inflammation of the liver. The Hepatitis dataset classifies whether an individual will live or die from hepatitis based on various qualitative/quantitative characteristics. After filtering out the rows with unknown values ('?'), dataset rows were cut down to 80, so it was useful to obtain the class distributions and compare them to the original dataset. The class distribution of this dataset was imbalanced: 16.3% of the data had a DIE outcome and 83.6% had a LIVE outcome.

When calculating useful statistics, measures of central tendency, variability, and correlation between pairs of data were considered. In particular, it was useful to see the maximum and minimum values for future standardization and feature scaling steps. The standard deviation and skewness were useful to observe how data deviated from the mean. The full report of data analytics can be observed in SECTION 1 Data Analysis (Hepatitis Dataset) part of the Collab code.

Diabetic Retinopathy Debrecen Dataset
Diabetic Retinopathy (DR) is a disease that causes a loss of vision from a prolonged case of diabetes mellitus. After filtering the bad-quality instances, the dataset was cut down to 1147 instances. The Pre-Screening result and Quality Assessment features were eventually excluded from the dataset as they provide poor quality information for the classification algorithms. Moreover, the Quality Assessment class distribution was highly uneven while the Class label distribution was even: 53.3% of the patients show signs of the DR and 46.7% show no signs of the disease. The full report of data analytics can be observed in SECTION 1 Data Analysis (Diabetic Retinopathy Debrecen Dataset) part of the Collab code.

4. Results

Part 1 - Comparing the accuracy of K-Nearest Neighbors and Decision Tree algorithms.
To perform the comparison analysis, an initial and final test was performed. The initial test contained the default hyperparameter values for K-NN and DT constructors, while the final test used the best-performing hyperparameters. Overall, it was concluded that the K-NN model outperforms Weighted K-NN and the Decision Tree algorithm for both datasets. The details and accuracies of each experiment are detailed below.

Hepatitis Dataset

Test	K	Distance Function	Depth	Cost Function	Accuracy
K-NN Initial	1	euclidean	N/A	N/A	71.9 %
K-NN Final	3	manhattan	N/A	N/A	90.6 %
Weighted K-NN Initial	1	euclidean	N/A	N/A	71.9 %
Weighted K-NN Final	3	manhattan	N/A	N/A	87.5 %
DT Initial	N/A	N/A	3	Cost Misclassification	87.5%
DT Final	N/A	N/A	2	Cost Misclassification	87.5%

Diabetic Retinopathy Debrecen Dataset

Test	K	Distance Function	Depth	Cost Function	Accuracy
K-NN Initial	1	euclidean	N/A	N/A	53.6 %
K-NN Final	12	manhattan	N/A	N/A	65.1 %
Weighted K-NN Initial	1	euclidean	N/A	N/A	53.6 %
Weighted K-NN Final	3	manhattan	N/A	N/A	62.5 %
DT Initial	N/A	N/A	3	Cost Misclassification	59.5%
DT Final	N/A	N/A	13	Cost Misclassification	64.3%

Part 2 - K-Nearest Neighbors: testing different K values and observing the effect on training data accuracy and test data accuracy.
To find the best hyperparameter K, it was necessary to split the data into training, validation, and testing. The best prediction was made by running the K-NN/Weighted K-NN classifiers on a range of K values [1, 15], while keeping the distance function as the default euclidean. From printing the test output, as the value of K increased, the prediction accuracy improved. At some point, too many nearest neighbors caused the accuracy to stagnate and even decrease as the models began to overfit. Additionally, when the models used training data to make predictions, the accuracy reported was higher than when the models predicted on unseen testing data.

- **K-NN: Hepatitis dataset:** best K = 3 with an accuracy of 70.8% on the testing set and 87.2% on the training set.

- **K-NN: DRD dataset:** best K = 12 with an accuracy of 58.3% on the testing set and 75.6% on the training set.
- **Weighted K-NN: Hepatitis dataset**: best K = 3 with an accuracy of 66.7% on the testing set and 71.8% on the training set.
- **Weighted K-NN: DRD dataset**: best K = 3 with an accuracy of 53.0% on the testing set and 53.8% on the training set.

Part 3 - Decision Tree: testing different tree depths and observing the effect on training data accuracy and test data accuracy.

To extract the most optimal depth, the data was split into training, validation, and testing. The best prediction accuracy was obtained at tree depths 2 and 13 for Hepatitis and DRD respectively. With regards to the Hepatitis dataset, as depth increased, accuracy increased and reached a plateau of 81.2% accuracy. For the DRD dataset, the same effect was observed; the final reported accuracy was 68.8%. Just as with KNN, when the DT model used training data to make predictions, the accuracy reported was higher than when the model predicted on unseen testing data.

- **DT: Hepatitis dataset:** best Depth = 2 with an accuracy of 81.2% on the testing set and 95.8% on the training set.
- **DT: DRD dataset:** best Depth = 13 with an accuracy of 68.8% on the testing set and 83.6% on the training set.

Part 4 - Testing the effect of different distance and cost functions.

1. **Distance Functions**

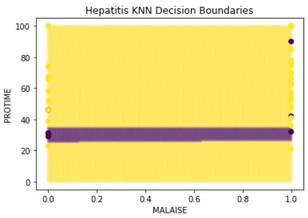
Much like the tests in Part 2, to find the best distance function (euclidean/manhattan), it was necessary to split the data into training, validation, and testing. The K values selected were the best-performing ones in Part 2. The best distance function for both datasets and models turned out to be the manhattan distance. Using manhattan distance, for the Hepatitis dataset, the K-NN model reported 70.8% while the weighted K-NN model reported an accuracy of 83.3%. For the DRD dataset, the K-NN reported an accuracy of 63.5% and the weighted K-NN reported 60.9%.

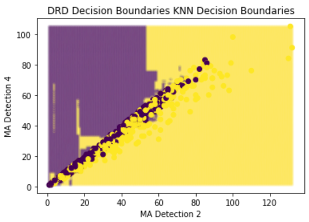
2. **Cost Functions**

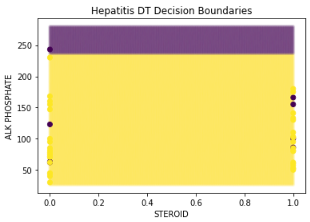
To find the best cost function (Misclassification/Entropy/Gini Index), the data was split into training, validation, and testing. The selected depth values were chosen from Part 3 to maximize performance. The best cost function for both datasets was the cost misclassification function with an accuracy of 81.2% for the Hepatitis dataset and 68.8% for the DRD dataset. It is also important to note that the reported high accuracy of the misclassification cost function for the Hepatitis dataset is due to features like STEROID and ANOREXIA being very closely correlated with the Class label of the data point.

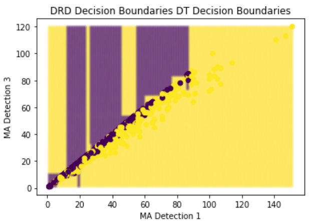
Part 5 - Decision Boundaries:

To obtain the best two features for decision boundary plots it was useful to loop over all combinations of features and obtain those producing the highest accuracy predictions. For the Hepatitis dataset, the best performing features were MALAISE and PROTIME for KNN and ALK PHOSPHATE and STEROID for DT. For the DRD dataset, the best performing features were MA DETECTION 2 and MA DETECTION 4 for KNN and MA DETECTION 1 and MA DETECTION 3 for DT.









Part 6 - Feature selection and standardization for both algorithms:

The K-Nearest Neighbors, Weighted K-Nearest Neighbors, and Decision Tree all used the same variation of the features because the models were made to fit multivariate data. Additionally, to accurately calculate a performance analysis of the models, the datasets fitted by each model should remain the same. This way, it is possible to observe which model is able to better adapt to the various types of input features. In order to decrease the dimensionality of the Hepatitis dataset, the AGE and SEX columns were removed because their observed correlation with the Class label was small. In the case of the DRD dataset, two features were excluded: Quality Assessment and Pre-Screening Result. In both datasets, the features were all within varying ranges so the outlier-robust standardization technique was performed over the train, validation, and test data. It is important to note that standardization was only useful for the K-NN algorithm. Decision trees are not sensitive to feature scaling, so it was unnecessary to standardize the data.

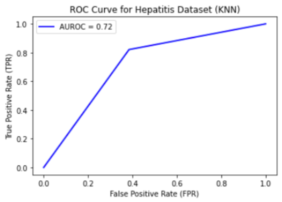
Part 7. K-Fold Cross Validation

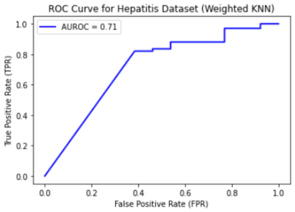
K-Fold cross validation was implemented in order to explore a new validation method. In the previous experiments, the dataset was split into training, validation, and testing partitions. However, this makes the testing portion go to 'waste' since it is not used for training the model. Using K-fold cross validation is especially important for small datasets like the Hepatitis dataset. The following k-fold cross validation results were obtained using the best performing model (K-NN) with a split number of 5. However, it did not produce better results than previous tests with regular validation.

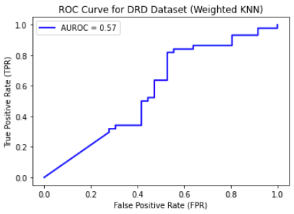
- K-NN: Hepatitis dataset: K = 3 with an accuracy of 86.25 %
- K-NN: DRD dataset: K = 3 with an accuracy of 87.5 %
- Weighted K-NN: Hepatitis dataset: K = 12 with an accuracy of 65.21 %
- Weighted K-NN: DRD dataset: K = 3 with an accuracy of 62.25 %

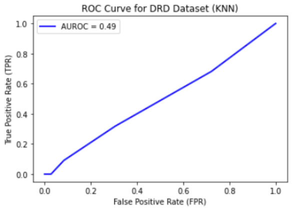
Part 8. Receiver Operating Characteristic Curve

Part 5 of this section used decision boundaries as a performance visualization metric. Decision boundaries only provide a comprehensive graph for a limited two-feature input. As observed from the graph below, the calculated ROC curves for both K-NN and weighted K-NN for the hepatitis dataset are higher than the random classifier that predicts with 50% accuracy; the random classifier would resemble a diagonal line from (0.0, 0.0) to (1.0, 1.0). However, the same cannot be said for the DRD datasets. The ROC curves for this dataset are very close to the diagonal, making its prediction close to that of a random classifier.









5. Discussion and Conclusion

Although the models perform well with both datasets, it would be useful to improve them to work with any type of data input of any feature arrangement. The Decision Tree model in particular, does differentiation of categorical and numerical data from the column names, while it could instead work by distinguishing the type of data in each column (e.g. float, integer, string). The K-NN models could also be improved by writing a function that is able to distinguish between categorical and numerical features and associate the correct distance function accordingly.

Using euclidean and manhattan distance for multivariate features was fine for these two datasets because the number of categories of each categorical feature was 2 and it was possible to one-hot encode them so that the calculated distance is similar to the hamming output. However, this would not work for categorical features with multiple labels.

Despite these potential improvements, the overall task of testing the performance of two supervised machine learning models was accomplished. The data was acquired, thoroughly analyzed, and preprocessed to include only what were considered to be the most useful features and instances. The K-NN, Weighted K-NN, and Decision Tree models were implemented to best iterate over the feature-scaled and standardized datasets. Finally, the algorithms were tested on the Hepatitis and DRD datasets and compared/contrasted with one another to find the best performing model.

6. Statement of Contributions

Everyone contributed an equal amount of effort. All team members initially implemented each ML model individually, however towards the end, each member took a specific model, improved it, and performed the testing experiments. The report was also written and previewed by everyone in the group.

7. References

[1] W. Jaisingh, R. K. Kavitha and S. Kowsalya, "Certain Investigation on Various Machine Learning Techniques and Feature Selection Methods using Diabetic Retinopathy Features," *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2021, pp. 1-5, doi: 10.1109/ICAECA52838.2021.9675771.

[2] V. K. Yarasuri, G. K. Indukuri and A. K. Nair, "Prediction of Hepatitis Disease Using Machine Learning Technique," *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2019, pp. 265-269, doi: 10.1109/I-SMAC47947.2019.9032585.

[3] Bhargav, K. S., Thota, D. S. S. B., Kumari, T. D., & Vikas, B. (2018). Application of machine learning classification algorithms on hepatitis dataset. *International Journal of Applied Engineering Research*, *13*(16), 12732-12737.