Mini Project 1 - COMP 551 W2022

Sung Jun Lee, Joseph Boehm, Namdar Kabolinejad
February 8, 2022

1 Abstract

The usage of Machine Learning is ever-increasing in every industry, especially in healthcare. Machine Learning algorithms can be used in medical diagnoses to help with the early detection of diseases. In this project, our group studies how two different Machine Learning algorithms, Decision Trees and K-Nearest Neighbours, perform on two medical data sets relating to Diabetic Retinopathy Debrecen (DRD) and Hepatitis. To conduct our experiments, we first preprocess the datasets based on conventional methods for statistical analysis and then use data visualisation techniques to better understand the distribution and characteristics of the datasets. From our experiments we found that the K-Nearest Neighbours approach outperformed the Decision Trees approach nearly every time for the DRD dataset, and that the Decision Trees approach outperformed the KNN model nearly every time for the Hepatitis dataset. In our experiments, we attempt to creatively explore new boundaries by highlighting key attributes and using additional cost functions, on top of the required baseline.

2 Introduction

The goal of this project was to compare the classification performance of Decision Trees and K-Nearest Neighbours models using two medical datasets: Diabetic Retinopathy Debrecen and Hepatitis. Another important factor we set out to measure was how different hyperparameters, such as the "K" value for KNNs and "max depth" for Decision Trees, and feature engineering techniques could affect the classification accuracy across both models. The Diabetic Retinopathy Debrecen [1] dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. The most recent research done by Bangladeshian computer scientists and software engineers in 2021 using the DRD dataset tested classification using 8 different models. They were able to get a lowest score of 57% using Naive Bayes and a highest score of 69% using Random Forests. An interesting note here is that they achieved 62% accuracy using Decision Trees [4]. The Hepatitis [2] dataset was created to predict patients' survivability from hepatitis. Previous research using the Hepatitis dataset in 2013 was able to test 7 different models and reach a lowest classification accuracy of 70.41% using neural networks and a highest accuracy 96.52% using Naive Bayes [3]. Therefore, our goal is to train our models such that it is able to achieve accuracies similar to the previous works' highest scores while using arguably much simpler models such as KNNs and Decision Trees. To do this, we initially preprocess, clean, and divide the provided datasets using statistical analyses of the data such as its correlation scores, means, and class distribution. As aid in the preprocessing step, we visualise the data to get a more intuitive understanding of how the data is distributed. After preprocessing, we create three similar sets of each dataset each with slightly less features in order to test many possibilities. We hypothesised that moderately large values of "K", around 7 to 10 will perform the best for KNNs, and moderately low values of "max_depth" around 5 will perform best for Decision Trees. Additionally, we believed that the Tree model will outperform the KNN in both datasets. In the end, we were able to achieve an accuracy of 95.7% on the Hepatitis dataset using Decision Trees, and an accuracy of 70.4% on the DRD dataset using KNNs.

3 Datasets

The Hepatitis dataset is a rather small dataset, consisting of 155 instances and 19 attributes. The Diabetic Retinopathy Debrecen dataset, on the other hand, is a larger data set, consisting of 1151 instances and 20 attributes. The data of both data sets are a combination of binary and continuous values. Observing the Hepatitis dataset, we noticed that by removing all the instances from the dataset which contained a null value for a certain attribute, only 80 instances were left of the original 155 instances. We noticed that the attribute Protime was the only missing value in a lot of the instances, so we elected to remove that attribute from the dataset before removing all the incomplete instances. This

left us with 112 instances and 19 attributes which is significantly more data than we would have had had we removed all incomplete instances. The Diabetic Retinopathy Debrecen dataset didn't need much for cleaning, although one of the attributes was a binary result expressing whether the instance was of good quality. This meant that if the attribute had a value of 0, then the rest of the data in the instance didn't mean anything. Obviously we removed all the bad quality instances as well as the quality attribute as it became ineffective. Thus leaving us with 1147 instances and 19 attributes for the Diabetic Retinopathy Debrecen dataset.

To better understand each dataset, we computed some basic statistics including correlation scores, means, data distributions per class, histograms per feature, and pairwise scatterplots for every possible pair of features. With the correlation scores, we wanted to observe which features were the most correlated with the class of the instance. By observing this metric, we believed that we could determine which features were the most important and the least important in determining the class of a given instance. We discovered that the "Ascites" and "Albumin" features were the most important for the Hepatitis dataset with correlation scores of 0.48 and 0.43 respectively, and that "MA1" and "MA2" features were the most important for the DRD dataset with correlation scores of 0.29 and 0.26. This also seemed to indicate that the data for the DRD dataset would be harder to learn as the correlation scores were significantly lower than the Hepatitis set. We also determined that there was a significant imbalance in the number of instances per class for both datasets. The Hepatitis dataset we used had 19 entries (17%) labelled Class 1 ("DIE") and 93 (83%) entries labelled Class 2 ("LIVE"). The DRD set had 536 entries (47%) labelled Class 0 ("No signs of DR") and 611 entries (53%) labelled Class 1 ("Signs of DR"). Having a highly imbalanced dataset is not ideal for training, and the histogram data showed that both binary and continuous features were heavily imbalanced or skewed.

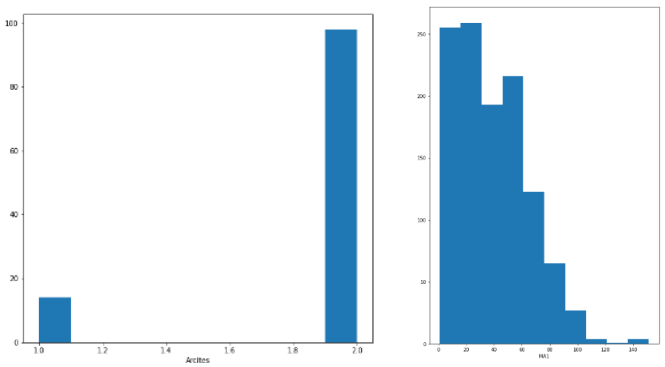

Figure 1: Example of the data imbalance in the "Ascites" feature of the Hepatitis dataset (left) and "MA1" feature of the DRD dataset (right)

Additionally, when we modelled the pairwise scatterplots between features, we noticed that a lot of the data would be hard to differentiate as they were mostly cluttered.

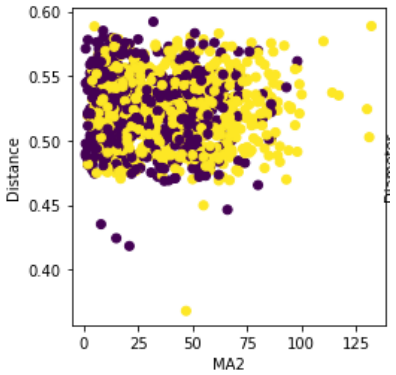

Figure 2: Example of cluttered scatterplot for the "Distance" and "MA2" features. Different colours represent instances of different classes.

From our statistical analysis, we therefore created 3 different datasets for each dataset for a total of 6: the original, one without attributes which have absolute correlation scores less than 0.05, and one without attributes which have absolute correlation scores less than 0.15. Then each of the 6 datasets were split 60:20:20 for training, validation, and

testing. We did not notice any ethical concerns as the data was anonymized, however, if each instance contained personal information of the patient such as name, address, and more, this could be a violation in ethical principles.

4 Results

We use the Decision Trees and K-Nearest Neighbours models that we wrote on both data sets. For the KNN model, we use the validation set with 4 different distance functions (Euclidean, Manhattan, Minkowski3, Minkowski4) and 15 different values of K to find the model with the best K and distance function, and eventually test the model using the best performing parameters. For the Decision Tree model, we take a similar approach, but with different combinations of max depth and cost functions. As explained in the datasets section, we create a total of 6 total data sets and we run 6 experiments for each of them.

4.1 Diabetic Retinopathy Debrecen

Table 1 demonstrates the best results achieved for each of the data sets with each of the models. We concluded that we achieve optimal accuracies on the Many Traits Removed dataset with K=1 and using Manhattan distance, resulting in a 70.4% testing accuracy. This is higher than what was done in previous works. In this experiment we discovered many notable results. Firstly, the KNN model performs better when more features are removed, whereas the Decision Tree model performs worse. Secondly, the KNN outperforms the Decision Tree in every test. Thirdly, as a general trend, we see that as the number of features go down, so does the best performing K value and max depth of the models. Lastly, another interesting thing to note is that Manhattan distance is preferred by the model over Euclidean distance, which is more widely used.

	K Nearest Neighbours		Decision Trees	
	Accuracy	Parameters	Accuracy	Parameters
Original Dataset	64.3%	K=15, Manhattan	62.6%	Depth=9, Entropy
Few Traits Removed	66.1%	K=11, Minkowski4	61.7%	Depth=3, Gini
Many Traits Removed	**70.4%**	K=1, Manhattan	60.0%	Depth=5, Misclassification

Table 1: Results of experiment for DRD Dataset, highest accuracy in **bold**

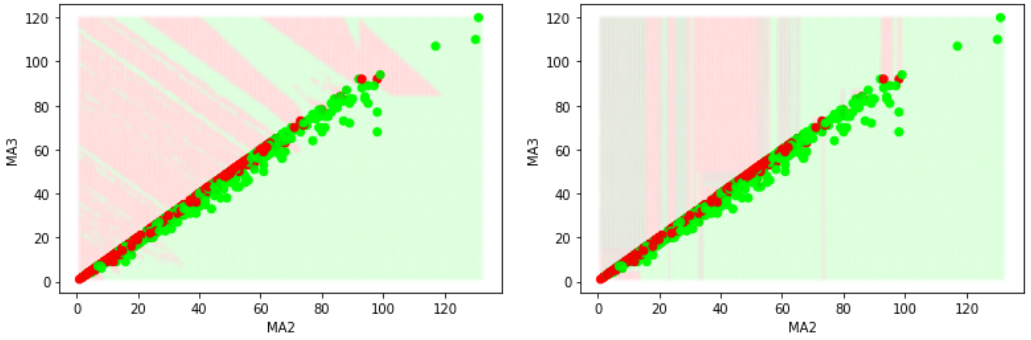

Figure 3: Decision boundaries between the features with the two highest correlation scores to the Class attribute. Left is KNN. Right is the Decision Tree.

Plotting the decision boundaries for both models, we see that they are rather messy, which is to be expected due to the messy nature of the data. It appears that the decision boundaries for the KNN are slightly neater, which seems to agree with the fact that the KNN performed better.

4.2 Hepatitis

In Table 2 we can see the results for the Hepatitis datasets. The optimal accuracy for the Hepatitis dataset was 95.7% on the Original dataset with Decision Trees using a max depth of 1 and the Misclassification cost function. Again, there are some interesting results we can draw from Table 2. Firstly, it is evident that the accuracies drop significantly as features are removed for both models. Secondly, low values for K and max depth tend to outperform higher values. Lastly, the KNN seems to favour Euclidean distance for this dataset and the Decision Tree appears to favour Misclassification cost. While these scores are not higher than previous works', a 95.7% accuracy is very close.

	K Nearest Neighbours		Decision Trees	
	Accuracy	Parameters	Accuracy	Parameters
Original Dataset	91.3%	K=3, Euclidean	**95.7%**	Depth=1, Misclassification
Few Traits Removed	87.0%	K=7, Euclidean	87.0%	Depth=1, Entropy
Many Traits Removed	87.0%	K=5, Euclidean	82.6%	Depth=3, Misclassification

Table 2: Results of experiment for Hepatitis Dataset, highest accuracy in **bold**

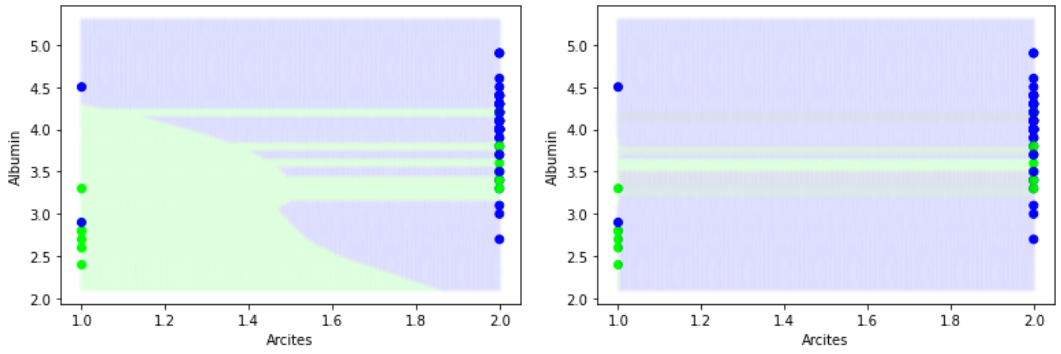

Figure 4: Decision boundaries between the features with the two highest correlation scores to the Class attribute. Left is KNN. Right is the Decision Tree.

Plotting the decision boundaries for both models, we see that they are rather strange, which is because "Ascites" is a binary feature. It appears that the decision boundaries for the KNN and Decision Trees are similarly accurate. While it is hard to tell, the Decision Tree boundaries correctly classify green points with "Ascites" = 1.0.

5 Discussion & Conclusion

The results of our experiments were mostly as expected and slightly surprising. Immediately noticeable was the effect of data imbalance and correlation scores for each feature. The models tested much better on the Hepatitis dataset which was heavily skewed and had higher correlation scores for each feature to the class. Whereas the Hepatitis dataset had features which were more distinct to a class, the DRD dataset had lower correlation scores and thus their accuracies were lower. The second conclusion we can draw is that the KNN models performed better when there were a lot more instances in the dataset, whereas the Decision Trees performed better for smaller datasets. This is to be expected as KNNs perform better when datasets are large. One of our creative approaches that worked well is the removal of uncorrelated features from the dataset. It appears that removing a lot of "insignificant" features from the DRD dataset improved the classification accuracy of the KNN by more than 6% compared to the original. However, the exact opposite was the case for the Hepatitis dataset; removing a lot of features ended up decreasing the accuracy. We therefore came to the conclusion that there must be a large number of instances in the dataset if we wanted to effectively remove features. If the dataset is too small and we remove too many features, this may "oversimplify" our data and make our models worse. Additionally, the effects of varying K values were observed. It appears that a K

value of 3 to 7 is ideal for most instances. When it is too low or too high, accuracies tend to drop, however, the exact value of K is best determined through validation. More on the value of K, it appeared that the more features, or dimensionality a dataset had, the model preferred a higher K value. This is most likely because the added dimensionality requires more neighbours for an accurate classification. As for the distance functions of the KNN, while we tried four functions (Manhattan, Euclidean, Minkowski3, Minkowski4), Manhattan and Euclidean were strongly favoured. From the results, when the dataset was large, Manhattan distance was preferred, and when the dataset was small, the Euclidean distance was favoured. For the Decision Trees, it was observed that a high value for max depth was preferred for large datasets and that a low value was preferred for smaller datasets. This may be explained by the fact that larger datasets tend to require more decisions to accurately model the data. Interestingly, the max depth did not seem to depend heavily on the number of features in the dataset as the highest performing models in each of the six datasets had sporadic values for max depth. Lastly, the preferred cost function was Misclassification cost, which performed best in 3 of the 6 datasets. In summary, we thus conclude that KNNs are superior for large datasets with a low number of features with a K value between 3 to 7 and a distance function of either Manhattan or Euclidean. Additionally, Decision Trees were superior for small datasets with a high number of features with a max depth value between 1 to 9 and the Misclassification cost function. In the future, it would be interesting to test additional models such as Naive Bayes or even neural networks to compare their performance. Additionally, feature scaling may be of interest for increasing KNN accuracy in the future. Finally, to comment on the decision boundary plots, the decision boundary for KNNs were irregular and were composed of straight lines and curves, whereas the boundaries for the Decision Trees were straight and rectangular in nature. This was the case because the boundary for KNNs were determined geometrically by calculating the distance functions, whereas the boundary for Decision Trees were determined via answers to binary questions. In both cases, the decision boundaries served as an aid in understanding the results but were not significantly impactful as they were quite messy, due to the cluttered-ness of the original data, which is a problem inherent to the original datasets.

6 Statement of Contributions

- Sung Jun Lee
    - Contributed to “Data Preprocessing” code
    - Wrote entirety of “Data Visualisation” code
    - Trained and tuned KNN and Decision Tree Models
    - Wrote validation algorithm for picking best models
    - Contributed to Abstract, Introduction, Dataset
    - Wrote majority of Results, Conclusion, and Contributions

- Joseph Boehm
    - Wrote entirety of “Imports” code
    - Wrote majority of “Data Preprocessing” code
    - Trained KNN and Decision Tree Models
    - Generated decision boundaries plot
    - Contributed to Datasets

- Namdar Kabolinejad
    - Contributed to”KNN” code
    - Wrote majority of Abstract
    - Wrote majority of Introduction
    - Contributed to Datasets
    - Wrote majority of References

References

[1] Balint Antal, Andras Hajdu: An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60 (April 2014), 20-27.
The dataset is based on features extracted from the Messidor image dataset: [Web Link].
[2] 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.
[3] Ba-Alwi, F.M., & Hintaya, H.M. (2013). Comparative Study for Analysis the Prognostic in Hepatitis Data : Data Mining Approach.
[4] Emon, M.U., Zannat, R., Khatun, T., Rahman, M., Keya, M.S., & Ohidujjaman (2021). Performance Analysis of Diabetic Retinopathy Prediction using Machine Learning Models. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 1048-1052.