Comp 551 Assignment 1 Group 24

Angele Park Collin, Jessica Dekker, Adam Garay

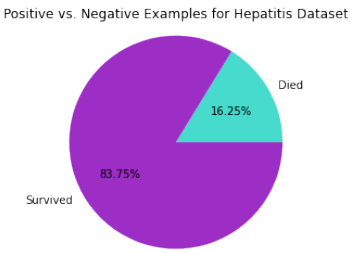
October 6th, 2022

1 Abstract

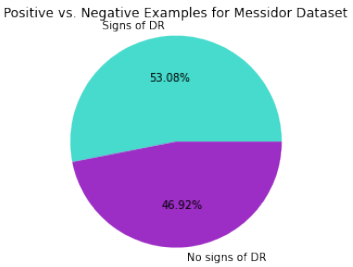
In this assignment, we investigated the performance of two elementary machine learning models on two benchmark data sets. This assignment was an introduction to Machine Learning concepts with the objective of gaining familiarity with running experiments and comparing the performance of different models. The first data set provided is a multivariate data set on the disease hepatitis, and the second is based on Diabetic Retinopathy Debrecen Data features extracted from the Messidor image set [1]. Initially, we performed fundamental statistical analysis on the data sets to better understand the correlation between potential vital features and the positive/negative diagnosis of the diseases. We implemented both the K-Nearest Neighbours and Decision Trees classification techniques to evaluate both data sets and compare which method achieves higher accuracy results in the potential diagnosis of hepatitis and diabetes. At the end of our analysis, we concluded that the KNN approach achieved better accuracy for the Hepatitis data set, and the KNN was also more effective for the Messidor data set.

2 Introduction

The task provided was to implement two classification techniques - K-Nearest-Neighbour (KNN) and Decision Trees (DT) - and compare the two algorithms on two distinct health data sets. The first data set was created by G.Gong from Carnegie-Mellon University. It has survival classes (Die or Live) and describes various attributes associated with hepatitis, such as age, sex, anorexia, steroids, etc. From running basic statistics on this data, we found the percentage of patients who survived hepatitis was 83.75%. In addition, we found that the age range of the patients who died from hepatitis was concentrated in their late-thirties the to early fifties.



The second dataset was created by Dr. Balint Antal and Dr. Andras Hajdu, both from the University of Debrecen in Hungary. This data set contains features and measurements extracted from the Messidor image set to predict whether an image has signs of diabetic retinopathy. Once again, we explored this data set by implementing some elementary statistics to understand the variables better. We found that over half of the imaging in the data set (53.08%) contained signs of diabetic retinopathy. We also explored the range of exudate levels, a fluid that leaks out of blood vessels into nearby tissues, concerning positive signs of diabetic retinopathy. From the scatter plot we generated, there seems to be a trend between higher exudate levels and symptoms of diabetic retinopathy developing.



3 Methods

The two machine learning methods we implemented to analyse the datasets are K Nearest Neighbours (KNN) and Decision Trees(DT).

The K Nearest Neighbours method is a non-parametric model that uses a supervised learning algorithm for classification and regression. It takes as input point K, and predicts a label for that point by using the K number of neighbouring data points. As KNN is a distance based classifier, we used both Euclidean and Manhattan distance formulas to calculate the distance metrics. Euclidean distance is simply the length of the line segment between any pair of data points. Whereas the Manhattan distance uses the absolute value of the difference between two points. Manhattan distance is thus preferred over Euclidean distance when using high dimensional datasets [3].

Like KNN, the Decision Tree method is also a non-parametric model that uses a supervised learning algorithm for classification and regression. However, it differs from the previous method as it models the data by breaking it down into smaller sets. The DT method takes in as input a dataset then it partitions the various attributes into smaller sets and calculates the entropy of each attribute. The algorithm then uses the entropy for each branch in comparison to the total entropy for the final step. Then, the DT has for output the information gain, which is the branch of the tree (ie. the attribute) with the best accuracy. [2].

4 Datasets

As mentioned above, we sorted through the data sets and found some basic trends that helped inform our decisions for implementing the models and splitting the data correctly. For the machine learning models to perform without

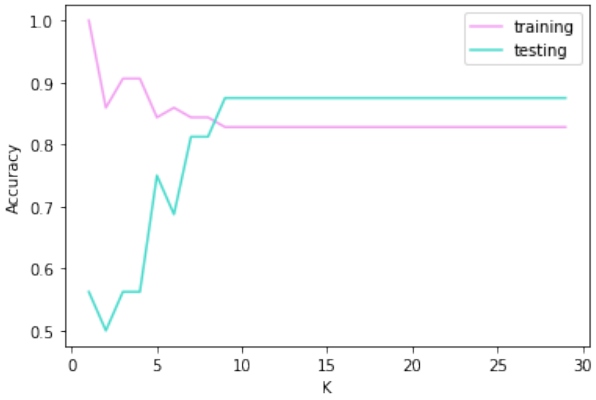
bias, for both data sets, we used the Train-Test split process to split the data into 2 sets, a training and testing set. We took the process from the sckitlearn package. Normally, the training set accounts for 80% of the original data and is used to create a model that predicts the data accuracy with additional unseen data. The remaining 20% is called testing set and is used to evaluate the model to see how accurately the model performed. However, for the Hepatitis data set, due to the lack data points, we had to split the data differently to have enough validation points to properly perform the testing accurately. For the Messidor data set we were able to use the normal splitting method of 80% training to 20% validation. This is when we found the hyper-parameters. Following this methodology, we also tested without the validation data to find how each parameter performed in reality.

5 Results

From performing multiple exploratory experiments using K-Nearest Neighbours and Decision Trees on hepatitis and diabetic retinopathy patient data, we observed some interesting patterns and results.

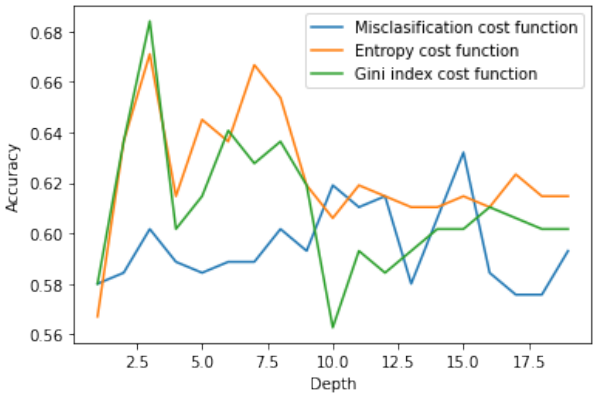
5.1 Hepatitis Data Set

The results from performing the 50/50 split on the Hepatitis data yielded a KNN model where K=5 and the test accuracy was 0.75 when we included the validation data. Then we performed a true test of the best K value, and found that although the accuracy spiked at K=5, K=9 had a slightly higher test accuracy. For DT, we tested both with and without validation data, but in this case the best depth remained the same with a value d=3. The next step in our experimental process was evaluating different distance and cost functions. For K=5, Manhattan distance performed better, and for K=9, Euclidean distance was more accurate. Misclassification function was the most accurate, and the entropy and gini index yielded the exact same accuracy. We found the most correlated features with positive hepatitis status were Ascites, Albumin, and Histology. After finding these key features, we selected two features that were not binary values as they gave us a better visualization to graph the classification models and decision boundaries.



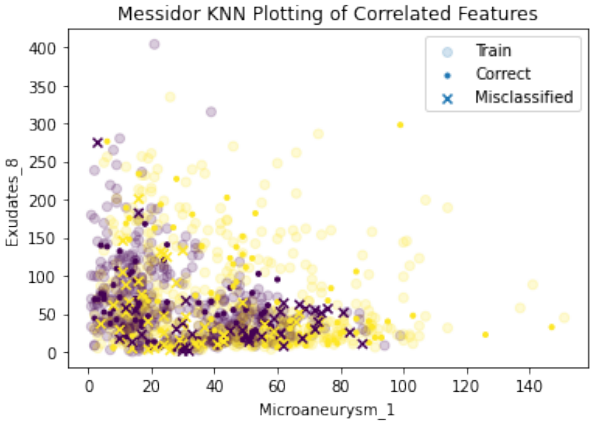
5.2 Diabetes Retinopathy Data Set

For the Messidor data, we were able to implement the normal 80/20 split, and using the validation data we got a K-value of 14, and the true test gave us at K-value of 12. We observed that the difference between these two analysis were very similar to each other, unlike the hepatitis data. For DT, we tested both with and without validation data, but in this case the best depth remained the same with a value d=3. The next step in our experimental process was evaluating different distance and cost functions. For K=12, Euclidean distance was more accurate, but for K=14 there was virtually not difference between the two. Gini index yielded the best accuracy for d=3, but for d=12 the gini index got out performed by the entropy and misclassification functions. We found the most correlated features with signs of diabetic retinopathy to be micro-aneurysm 1-3. After finding these key features, we selected two features that were not binary values as they gave us a better visualization to graph the classification models and decision boundaries.



6 Discussion and Conclusion

One of the key takeaways from the assignment was that larger data sets such as the Messidor data (see graph below), can be visualized better, but do not necessarily give more accurate modelling accuracy. Another takeaway from learning throughout this experiment is the importance of truly exploring your data sets and understanding their classes prior to implementing machine learning models as it makes it easier to understand where things go wrong or how you can improve the model. Initially, we were confused about some of the variables and feature, but by gaining a bigger picture of the data it allowed for us to complete Task 3 with greater ease. To summarize, we were able to conclude that for our simulation, K-Nearest-Neighbours was the most accurate model for both patient data sets. One component we could explore in the future would be hyper-parameters for minimum leaf instances for Decision Trees. In addition, we could perform cross validation tests to compare between data sets.



7 Statement of Contributions

- Angele : Task 1 - Loading, cleaning and distribution of the data; Task 2 - Implemented the KNN class; Task 3 - Testing K values, Testing Distance Functions, KNN Key Features; Deliverables - Abstract, Introductions, Methods, Datasets
- Adam : Task 1 - Cleaning and Basic Statistics; Task 2 - Implemented the DT class; Task 3 - Experiments 1-7; Deliverables - Results, Discussion & Conclusion
- Jess: Task 1 - Cleaning and Basic Statistics; Task 2 - Implemented the DT class; Task 3 - Experiments 1-7; Deliverables - Abstract, Introduction, Data sets, Results, Discussion & Conclusion

References

[1] Balint Antal and Andras Hajdu. http://messidor.crihan.fr/index-en.php.

[2] Chirag Sehra. Decision trees explained easily. https://chirag-sehra.medium.com/decision-trees-explained-easily-28f23241248, Jan 2018.

[3] Jason Wong. K-nearest neighbors algorithm. https://towardsdatascience.com/k-nearest-neighbors-algorithm-d4a8bb1926a3, Dec 2020.