Abstract

In this assignment, we evaluated the accuracy of K-Nearest Neighbour (KNN) and Decision Tree (DT) machine learning models on the hepatitis dataset and Diabetic Retinopathy Debrecen dataset. We selected important features for the machine learning models by evaluating their correlation with the class. For the hepatitis dataset, we chose “albumin” and “protime” as the relevant features, and for the Diabetic Retinopathy Debrecen dataset, we chose MA with confidence level $\alpha = 0.5$ and MA confidence level $\alpha = 0.7$. By running KNN and DT on these two datasets, we established that KNN generally demonstrated around 5% greater accuracy in testing new data than DT. We also observed that the hepatitis dataset showed higher accuracy in both models than the Diabetic Retinopathy Debrecen dataset. The higher accuracy could be attributed to the large number of live patients in the hepatitis dataset, which makes it greatly unbalanced. Thus, it is more prone to classify new data points as having this label and getting it right.

Introduction

Machine learning algorithms have become the most efficient way to find patterns in data and to use these patterns to make predictions on new data. Two of these algorithms are K-Nearest Neighbor (KNN) and Decision Tree (DT). We applied these algorithms to 2 datasets: the hepatitis dataset and the Diabetic Retinopathy Debrecen dataset. The hepatitis dataset classifies individuals as dead or alive and provides 19 attributes, such as age, sex, steroids, antivirals, fatigue, malaise, and anorexia, and can help determine the survivability of a hepatitis patient based on these characteristics. In this analysis, we used the levels of albumin and protime as the two key features. Albumin is a peptide chain produced by the liver which contributes to maintaining the volume of circulating blood (Whicher and Spence, 1987). A decrease in the concentration of albumin can be indicative of liver illnesses (Whicher and Spence, 1987). Prothrombin time measures the time taken for the blood to clot. A high protime could be linked to liver dysfunction or disease (Yang and Moosavi, 2022). These characteristics of the two features can indicate the severity of hepatitis and thus help us predict whether the illness could be deadly. For KNN, the best number of neighbors found was 1, with an accuracy on the test set of 87.2%. As for the decision tree, we found that the best depth of the tree is one and that the algorithm's accuracy on the test dataset amounts to 82.1%, hence lower than the result obtained by the KNN algorithm.

The second dataset classifies retinal images as containing signs of diabetic retinopathy or not using features extracted from said images. Diabetic retinopathy is an effect of diabetic mellitus which affects the retina and, without detection and proper medication, can lead to blindness (Antal and Hajdu, 2014). To recognize early signs of DR, the analysis of color fundus images, also called retinal images, are used since microaneurysms (MAs) on the retina are a symptom of the illness and appear as small red dots on the photographs (Antal and Hajdu, 2014). We found that using the numbers of MAs of different confidence levels alpha as features in our analysis brought us the best results in terms of accuracy. For KNN, the best number of neighbors found was 9, with an accuracy on the test set of 69.1%. As for the decision tree, we found that the best depth of the tree is nine and that the algorithm's accuracy on the test dataset amounts to 66.7%, which is again lower than the result obtained by the KNN algorithm.

Methods

KNN works by storing all the training data so that when new data is presented, it can calculate the difference (or distance) between each new data point and the training data. The probabilities of classes for the new data are equal to that of the classes of all “K” closest training data points. K, the number of most similar data points, is a hyper-parameter that can be changed to yield better results, and the same applies to the distance function.

A decision tree works by analyzing all the features and determining which divides the data into the two different classes most efficiently. The features that best differentiate between the labels are selected for the whole experiment, along with the values that will become the thresholds when classifying the sets. The algorithm also displays which tree depth is the best for maximizing the accuracy, and this tree depth is then used on the testing set. Since the selection of attributes is NP-Hard, we used a greedy algorithm to get a “good enough” approximation.

Datasets

The hepatitis dataset contains 155 entries (i.e., patients) and 20 columns, of which one is for the class of the entries (whether the patient died or survived). The 19 others are the features of those patients, like their age, their sex, whether they experienced fatigue, malaise, or anorexia, and tests results on the levels of albumin and bilirubin, for instance, or even the prothrombin time (protime). As there were many entries for which some attributes were missing, only the complete ones were kept, totaling 80 entries. Since many of these features could have been irrelevant to the experiment, we chose to run a correlation test. We selected the two features with the highest correlation to the labels: albumin and protime (0.477 and 0.395, respectively). Another critical aspect of the dataset we noted is the imbalance in the number of members of each class. Indeed, there are 13 entries of class c = 1 (meaning that 13 of the 80 patients died) and 67 of class c = 2. We thus tried to balance the testing, training, and validation sets such that they would all have a similar ratio of members of both classes.

The Diabetic Retinopathy Debrecen dataset, which we also refer to as Messidor features dataset, contains 1151 entries and is more balanced regarding the ratio of sick to healthy patients. Indeed, there are 540 members of class c = 0 (healthy patients) and 611 of class c = 1. To match the other dataset and ensure our algorithms would run smoothly, we changed the values associated with the labels, such that c = 1 would indicate healthy patients and c = 2 would indicate sick patients. While the dataset was not missing any entries, a feature called 'quality assessment' gave a binary result on whether the retinal image of the patient was of good enough quality for further experimentation (Antal and Hajdu, 2014). We decided against removing these entries since there are only four of them. Their presence in the dataset is also more representative of reality, as retinal images may sometimes be ambiguous. However, if there were more of them, we would have had to remove them as they could skew the results. As for the features chosen, we ran a correlation test, and the features with the highest correlations were the number of MAs at confidence level α = 0.5 (MA, 0.5) with a correlation to the labels of 0.29 and the number of MAs at confidence level α = 0.7 (MA, 0.5) with a correlation to the labels of 0.27.

Results

We calculated the accuracy of our test sets for both datasets using the KNN and DT algorithms. We have consistently observed that the accuracy is higher when calculated using the KNN algorithm. Indeed, for the hepatitis dataset, the best number of neighbors found was 1, with an accuracy on the test set of 87.2%, compared to a depth of 1 with an accuracy of 82.1% using a decision tree. This pattern becomes apparent when we examine the Messidor features dataset, for which the accuracy amounts to 69.1% when calculated with KNN and drops to 65.5% when calculated with the DT.

We chose the best k neighbors value and the best depth by computing the accuracy on a validation set over a range of different values and taking the best k or depth that gave the highest accuracy. We can see the effects of changing these values on the accuracy in figure 1 and figure 2.

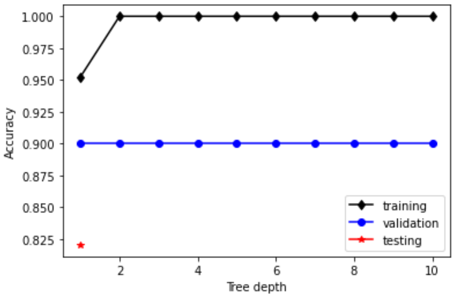

Figure 1. Accuracy calculated by the DT on the hepatitis dataset as the depth varies. *The figure shows how the accuracy on both the training set and the validation set varies when the tree depth changes. The star represents the depth at which we get the best accuracy on the validation set. The accuracy on the testing set is 82.1%*

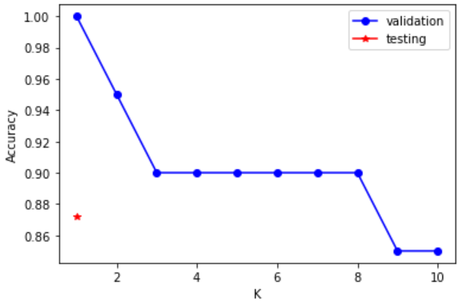

Figure 2. Accuracy calculated by KNN on hepatitis dataset as K varies. *The figure shows the accuracy of KNN for K values from 1 to 10 on the validation set, along with KNN on the test set for K = 1. At this K, we observe the best accuracy on the validation set. The accuracy on the test set is 87.2%*

For both algorithms, we took the value at which the accuracy of the validation set was optimized. We also experimented with the cost functions for the decision tree algorithm. While there was no difference in the accuracy of the test set for the hepatitis dataset, we observed interesting results on the Messidor dataset.

Table 1. *Comparison between the accuracy obtained on the test set using the three cost functions and the different distance functions*

Cost functions - DT	**Accuracy on the test set of the hepatitis dataset**	**Accuracy on the test set of the Messidor dataset**
Gini index	82.1%	65.5%
Misclassification cost	82.1%	66.7%
Entropy	82.1%	66.7%
Distance functions - KNN		
Euclidean distance	87.2%	69.1%
Manhattan distance	87.2%	69.1%

As we can see in table 1, the choice of the cost function in the DT does not change the accuracy value on the first dataset, and the same can be said for the choice of the distance functions for both datasets: the accuracy remains the same. However, for the Messidor features dataset, choosing entropy or the misclassification cost as cost functions result in an increase in the accuracy. For this reason, we chose the Gini index cost function when working with the hepatitis dataset, and for the other dataset, we chose entropy to measure the cost.

For the Messidor features dataset, we proceeded in the same way to choose the depth of the DT and the k value for KNN.

We found that the best depth is nine and the best k is nine as well. However, KNN has again a better accuracy value on the test set. Indeed, the accuracy calculated via KNN is 69.1%, compared to the 66.7% found with the DT.

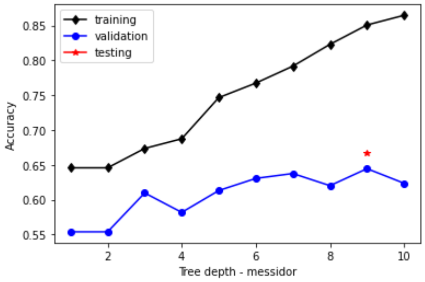

Figure 3. *Accuracy calculated by the DT on the Messidor dataset as the depth varies.*

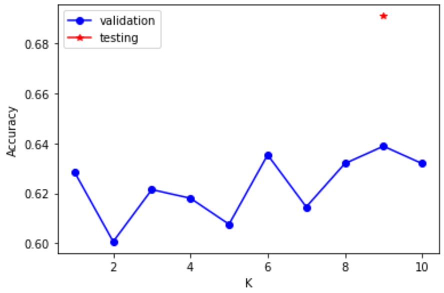

Figure 4. *Accuracy calculated by KNN on the Messidor dataset as K varies.*

We also plotted the decision boundaries for both datasets using the two algorithms.

As we can see in fig. 5 and fig. 6, the decision tree has a lower accuracy when using a depth of 1 as some data points end up in the wrong-colored regions. This difference in accuracy is also reflected in the number themselves as DT gave an accuracy of 82.1% compared to KNN with an accuracy of 87.2% on the same test set for the hepatitis dataset.

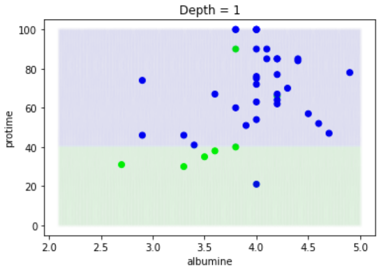

Figure 5. *Plot of the decision boundaries for the test set of the hepatitis dataset using DT at depth 1. The green region represents the class c = 1 (death of the patient).*

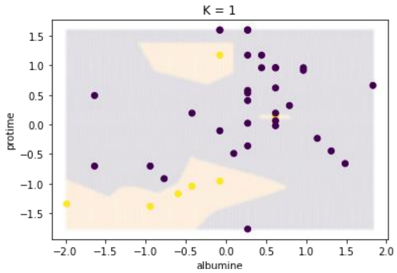

Figure 6. *Plot of the decision boundaries for the test set of the hepatitis dataset using KNN at k = 1. The pink region represents the class c = 1 (death of the patient).*

We have also plotted the classification results of both algorithms on the whole dataset. We can see that figure 8 has fewer misclassifications than figure 7, hence the higher accuracy given by the KNN algorithm.

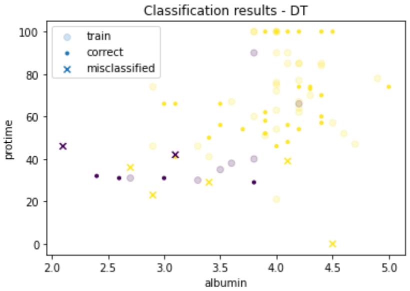

Figure 7. *Plot of the classification results on the hepatitis dataset using DT. Yellow data points represent entries of class c = 2, and purple points represent data points of class c = 1.*

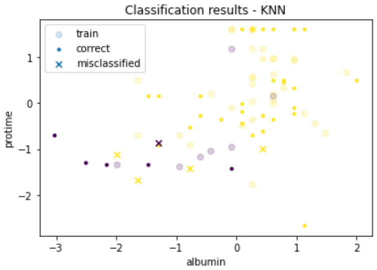

Figure 8. *Plot of the classification results on the hepatitis dataset using KNN. Yellow data points represent entries of class c = 2, and purple points represent data points of class c = 1.*

We can see from figures 9 and 10 that, once again, KNN is more accurate on the testing set. Indeed, many regions that should be green in fig. 9 are blue instead, which could be explained by the closeness of the data points despite having different labels. The decision tree gave an accuracy of 66.7%, while KNN gave one of 69.1%

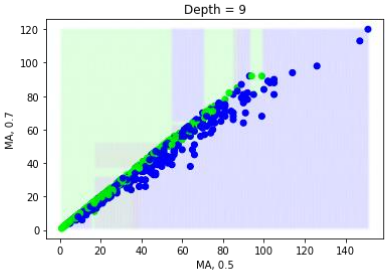

Figure 9. *Plot of the decision boundaries for the test set of the Messidor dataset using DT at depth 9. The blue region represents the class c = 1 (sickness).*

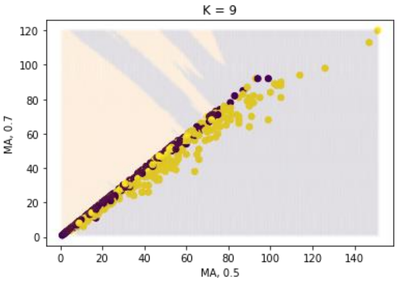

Figure 10. *Plot of the decision boundaries for the test set of the Messidor dataset using KNN at k = 9. The purple region represents the class c = 1 (sickness).*

Similarly, to the hepatitis dataset, we have plotted the classification results on the entire dataset, and, again, we can see that the results obtained with KNN have fewer misclassifications.

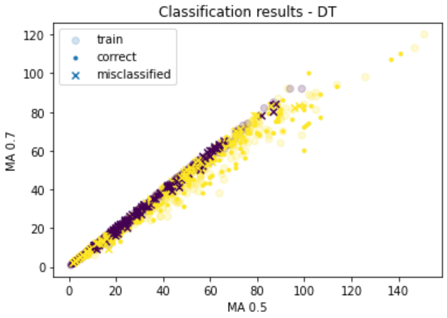

Figure 11. *Plot of the classification results on the Messidor dataset using DT. Yellow data points represent entries with diabetic retinopathy, and purple points represent data points with no signs of diabetic retinopathy.*

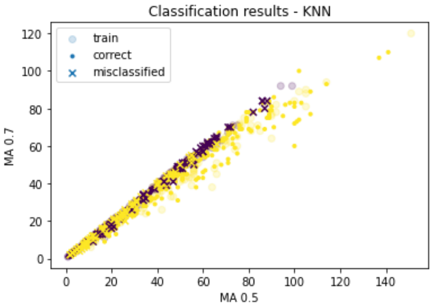

Figure 12. *Plot of the classification results on the Messidor dataset using KNN. Yellow data points represent entries with diabetic retinopathy, and purple points represent data points with no signs of diabetic retinopathy.*

We used the same features for both algorithms, as it can be seen from the fact that our decision boundary plots are very similar. We chose them by computing the correlation between every feature and the labels, and the ones with the highest correlations were used in the experimental part. In the first dataset, our choice was also supported by the fact that the algorithm for KNN allows the use of a single distance function at a time. Thus, for the algorithm to run smoothly, both features needed to be either discrete or continuous variables, as they each entail different distance functions. The features we thus worked with were albumin and protime, which have a correlation of 0.477 and 0.395, respectively.

We made an interesting observation about columns 2 to 7 in the Messidor features dataset, which represent the number of MAs found at different confidence levels alpha in an interval of [0.5, 1]. It is important to note that the higher the confidence level alpha is, the more confident we can be that the results we get are significant (Taylor, 2019). However, choosing an MA feature with a too-high confidence level alpha could result in a loss of information since a significant number of MAs could be discarded from the total count because of the high confidence level. We thus decided to work with the number of MAs at confidence level $\alpha = 0.5$ (MA, 0.5) with a correlation to the labels of 0.29, and the number of MAs at confidence level $\alpha = 0.7$ (MA, 0.5) with a correlation to the labels of 0.27. The rest of the features were ignored because of their lower correlation.

Discussion and Conclusion

Firstly, we observed that we obtained greater accuracy in both machine learning methods on the hepatitis dataset rather than the Diabetic Retinopathy Debrecen dataset. The reasoning for this most probably comes down to the datasets themselves and the more significant correlation between attributes and the class in the hepatitis dataset. However, this dataset suffers from an evident lack of data entries. The hepatitis dataset contained only 80 valid data entries that we could use, compared to the 1151 entries in the Diabetic Retinopathy Debrecen dataset. Furthermore, the hepatitis dataset contained far more entries with patients who survived. This could not only skew the results obtained with the models but could also indicate that the models are more likely to classify new patients as likely to survive. To remedy these problems, we think evaluating more data in this dataset would be relevant, primarily to ensure that we have a ratio of entries of each class more representative of reality. Furthermore, further research could evaluate the effectiveness of k-fold cross-evaluation since we are possibly working with insufficient data.

Statement of Contributions

Arien worked on the KNN code and analysis side of the assignment, whereas Soumaia worked on the Decision Tree code and analysis side of the assignment.

Bibliography:

Antal, Bálint, and András Hajdu. “An Ensemble-Based System for Automatic Screening of Diabetic Retinopathy.” Knowledge-Based Systems, Elsevier, 20 Jan. 2014, https://www.sciencedirect.com/science/article/abs/pii/S0950705114000021.

Taylor, Courtney. “It's Important to Know the Difference between Alpha and P-Values.” ThoughtCo, ThoughtCo, 30 Apr. 2019, https://www.thoughtco.com/the-difference-between-alpha-and-p-values-3126420.

Whicher, J, and C Spence. “When Is Serum Albumin Worth Measuring? - J Whicher, C Spence, 1987.” Sage Journal, Annals of Clinical Biochemistry: International Journal of Laboratory Medicine, Nov. 1987, https://journals.sagepub.com/doi/10.1177/000456328702400604.

Yang, Rocky, and Leila Moosavi. “Prothrombin Time.” National Library of Medicine, StatPearls, Jan. 2022, https://www.ncbi.nlm.nih.gov/books/NBK544269/.