K-Nearest Neighbour and Decision Trees: Implementation and Analysis

Eamonn Lye

ABSTRACT

The purpose of this project was to implement two machine learning models on two datasets. A decision-tree model and a K-nearest neighbors model were implemented to analyze the datasets. I found that each model performed similarly on the datasets. Furthermore, both models provided more accurate results for the hepatitis dataset than the diabetes dataset. For the K-nearest neighbors model, the best K value was chosen according to different distance functions, euclidean and manhattan. For the decision tree model, three cost functions were analyzed: misclassification cost, gini index and entropy. I found that for the K-nearest neighbors model, a K value of 3 and the use of either euclidean distance or manhattan distance provided the best results, with an accuracy of 88.8% for the hepatitis dataset. For the diabetes dataset, a K value of 9 and euclidean distance provided the highest accuracy at 61%. For the decision tree model and the hepatitis dataset, the best tree depth was 2 and any one of the cost functions resulted in a prediction accuracy of 81.5%. For the diabetes set, a tree depth of 7 and the entropy cost function provided the best accuracy, 63.4%.

INTRODUCTION

The task of this assignment was to predict the presence or absence of two health conditions, as described in two datasets. One contains health information about hepatitis patients, where I was trying to predict the outcome of their life or death. The second dataset also contains health information and the aim is to predict whether or not a given individual has diabetic retinopathy, a complication of diabetes. The predictions were made using both a K-nearest-neighbors and a decision-tree model. Furthermore, multiple distance functions were compared for the K-nearest neighbors model, and cost functions were compared for the decision tree model. I found that the models provided similar results, with both having higher accuracy on the hepatitis dataset than the diabetes dataset. The K-nearest neighbors algorithm resulted in 88.8% correctly identified hepatitis patients and 61% correctly identified diabetes patients. The decision tree algorithm correctly identified 81.5% of hepatitis patients and 63.4% of diabetes patients.

METHODS

The K-nearest-neighbor model classifies testing data based on their distance from data in the training data. A given data point is classified according to the class membership of its neighbors. The number of neighbors to which a given data point is compared and the choice of distance function is determined heuristically by iterating over several K values and different distance functions, the hyperparameters are then chosen based on the highest accuracy. The decision tree model classifies testing data based on thresholds. Each training data is classified along a sequence of nodes, which follow one of two branches, depending on whether they are greater than or less than the threshold. These thresholds are determined recursively by trying to minimize the cost function in use.

DATASETS

The two datasets analyzed both contained information pertaining to diseases. One contains information about hepatitis patients, and the other information about diabetic retinopathy patients, a complication of diabetes. The hepatitis dataset contains 18 features, all but 6 of which are categorical. The diabetes dataset contains 19 attributes, 3 of which are categorical. The features of both datasets include other observations that have been made about the patient's health. For the hepatitis dataset, analysis began with cleaning, which involved removing all entries that had missing values in any feature. This did not need to be done for the diabetes dataset, as all of the entries were complete. Then, scatter plots were generated to understand and visualize relationships between variables. Finally, a correlation matrix was generated to see which variables correlated most closely with the classes. These figures were the main basis for choosing the features to be graphed in the decision boundary plots, along with the scatter plots. After the incomplete observations were removed, we standardized the continuous variables and left the categorical variables as-is. At this point, I also determined the number of observations belonging to each class for both datasets and found that the hepatitis dataset was very imbalanced while the diabetes dataset was not, on top of the diabetes dataset being much larger. Once the data is cleaned, but before the data is standardized, I split the data into training and testing datasets. It is important to create these two datasets so that the model's accuracy is calculated based on data it has never seen before.

RESULTS

The K-nearest neighbors algorithm resulted in a maximum of 88.8% correctly identified hepatitis patients and 61% correctly identified diabetes patients. The decision tree algorithm correctly identified a maximum of 81.5% of hepatitis patients and 63.4% of diabetes patients. For each model, results varied depending on the dataset being analyzed. The models were able to predict hepatitis patients with much more accuracy than the diabetes patients. There are multiple factors at play causing this difference. First of all, the diabetes dataset is much larger, and second, the hepatitis dataset is much less balanced, with over 80% of observations having the same class. For the K-nearest neighbors model, the effect of choice for hyperparameter differs for the two datasets. To find the suitable hyperparameters, K and distance function, the training set was further split into two sets, one for validation, which is used for testing the model, one for training, which is used to training the model to fine the K and distance combination with the highest accuracy. For the hepatitis dataset, the hyperparameter choice of K=2 provides the highest testing accuracy, regardless of distance function. The choice of K=9 combined with the use of manhattan distance provides the highest testing accuracy for the diabetes dataset. For the decision tree model, the training set was split into validation and training to find a suitable max depth and cost function much like what happened when choosing the hyperparameters for the KNN model. a maximum tree depth of 2 provides the highest accuracy, independent of the cost function for the hepatitis dataset. A maximum tree depth of 7 combined with the entropy cost function gives the highest accuracy for the diabetes dataset. As can be seen from *Fig 2.*, the results received from the testing data does follow the general trend provided by results from the validation data. For the hepatitis dataset, I used the correlation function to find the two most correlated numerical features to my target, which happened to be “PROTIME” and “ALBUMIN”, before using them to generate the decision boundaries. The same process was used to choose my two features for the diabetes dataset, which were “MA Detection (Alpha = 0.5)” and “Euclidean Distance of The Macula And The Center Of The Optic Disc”. Many aspects of the model can be changed, resulting in various increases or decreases to the models’ accuracy. For example, isolating either the categorical or numerical variables both result in modest improvements to the K-nearest neighbor model accuracy and decision tree model. These could be the result of luck, or it could be the product of the models' weakness at dealing with a mix of categorical and continuous variables. On the categorical-only hepatitis data, I used a third distance function, the hamming distance, but this measure performed significantly worse than both the euclidean and manhattan measures. This difference could also be the result of the way I cleaned the datasets, by only standardizing continuous features and leaving categorical features as-is.

To further understand my datasets, I used principal component analysis to reduce my multivariate data set to a smaller set of data consisting of 2 features to have a easier time visualizing the datasets through its decision boundaries. And to accommodate for the imbalance nature of the hepatitis dataset, a precision recall curve was made to show the quality of KNN model, which performed to a reasonable AUROC of 0.970.

DISCUSSION

The models had similar prediction accuracies for both datasets. The difference between results across the datasets emphasizes the effect that an imbalanced dataset has on both K-nearest neighbors and decision tree models. Furthermore, a large number of decisions need to be made on the part of the designer of a model in order to produce a final, working product. The choice of key features, cost/distance function, normalization, and principal component analysis all happen independent of code implementation. Another notable observation is the so-called "stability" of the different datasets according to the different models. When trained on the hepatitis dataset, the accuracy of the K-nearest neighbors model does not depend on the distance measure chosen. Similarly, the cost functions all agree on the best tree depth for the hepatitis dataset. On the other hand, the choice of K varies greatly depending on distance function for the diabetes dataset, and the tree depth changes based on the cost function. Due to the similarities of all other aspects of the datasets (number of features, type of features), it is fair to hypothesize that the differences are the product of the size and relative imbalance of the datasets. Creating these models also requires solid knowledge of the data being analyzed, as the decisions made when cleaning and formulating the data, such as selecting important features, permeate throughout the whole model. These decisions can also be made easier by an intimate knowledge of the data. Future study could analyze methods to balance datasets such as the hepatitis dataset, since the imbalance of the samples had a large negative effect on both the K-nearest neighbor and decision tree models. This would allow these easier-to-implement models to be used on a greater variety of datasets. More rigorous techniques to choose key features would also be of use.

Appendix

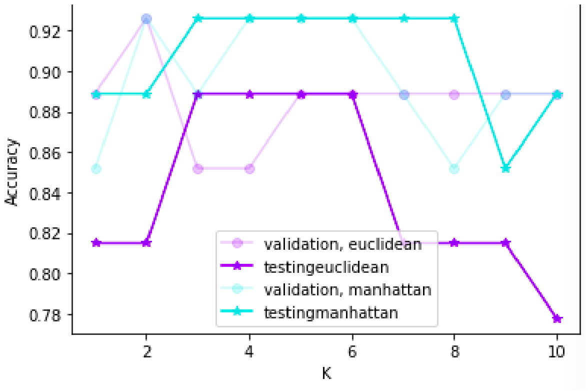
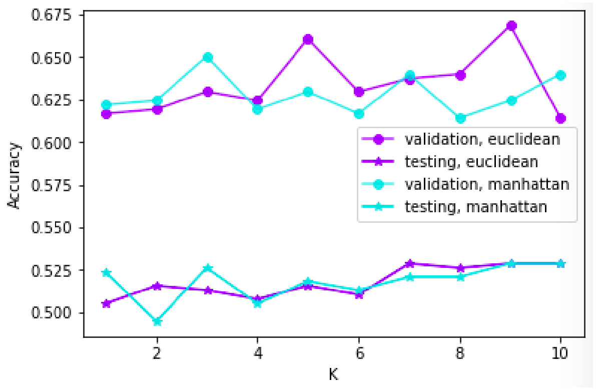
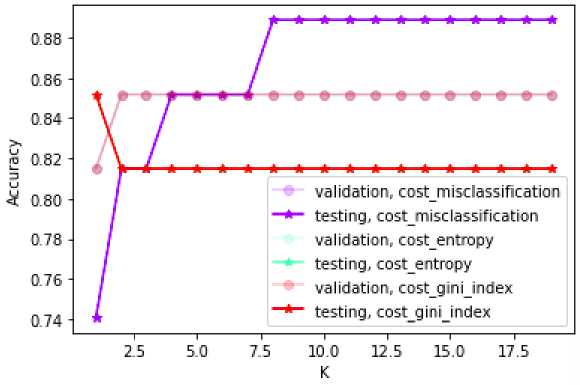
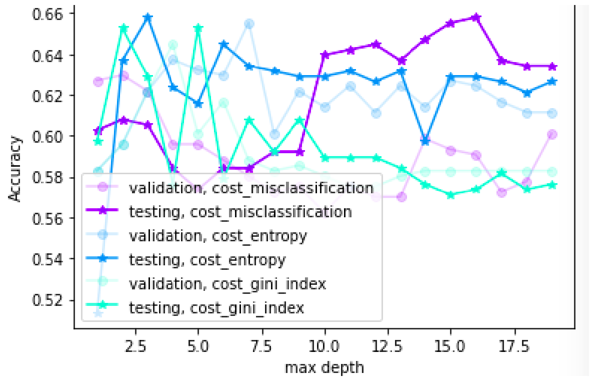
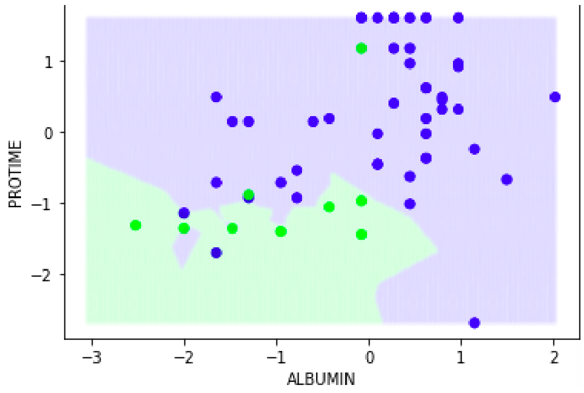
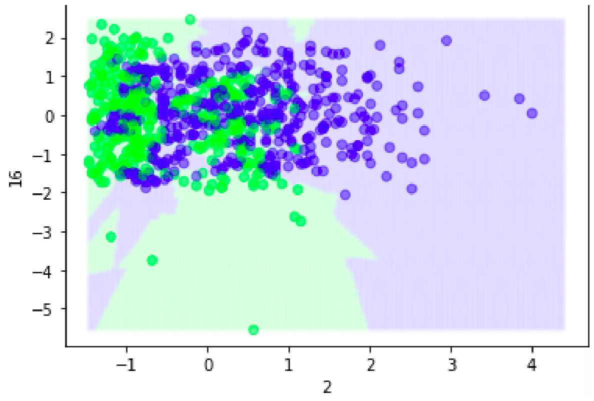
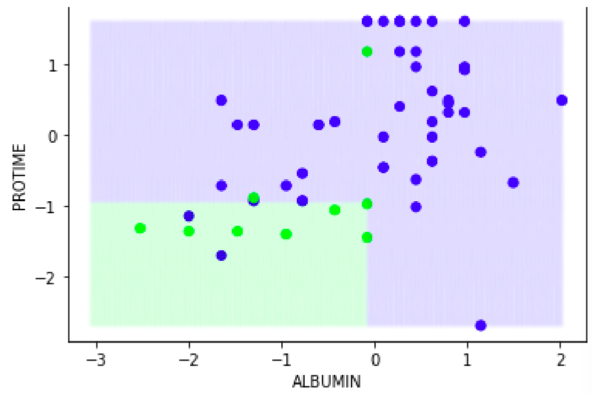
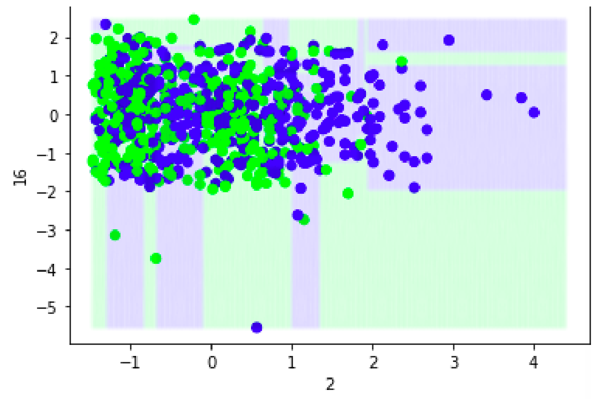
 	
Choice of hyperparameter K (hepatitis)	Choice of hyperparameter K (diabetes)
 	
Choice of tree depth (hepatitis)	Choice of tree depth (diabetes)

Fig. 1 Plot for choices of hyperparameters

 	
Hepatitis decision boundary (KNN)	Diabetes decision boundary (KNN)

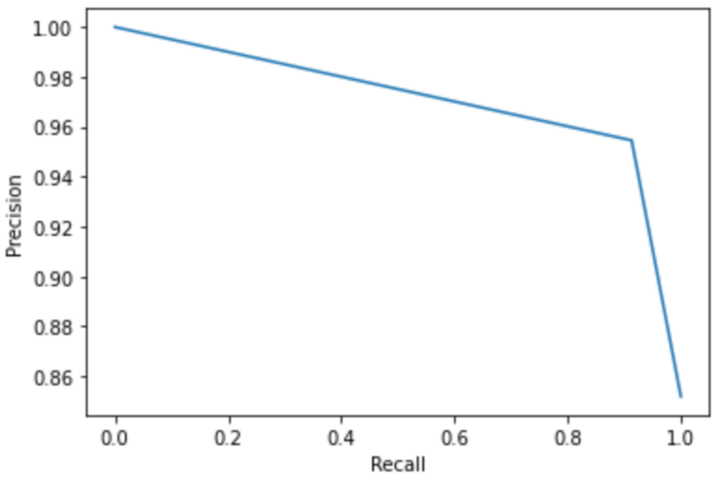


Hepatitis decision boundary (DT)



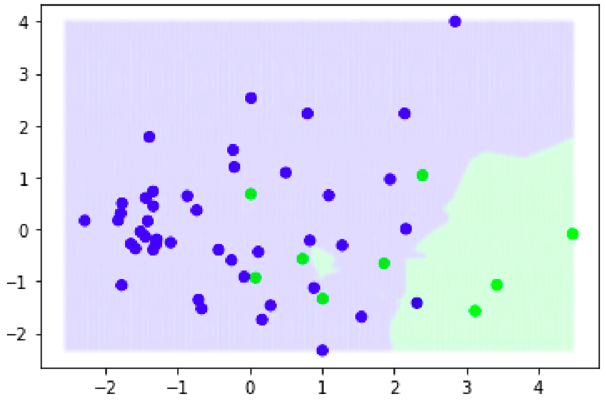
Diabetes decision boundary (DT)

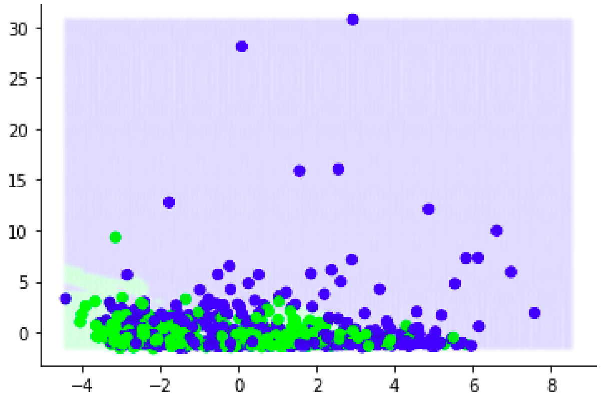
Fig 2. Decision Boundaries.



Precision recall curve for hepatitis dataset KNN model (AUROC: 0.970)

Fig 3. Precision recall curve





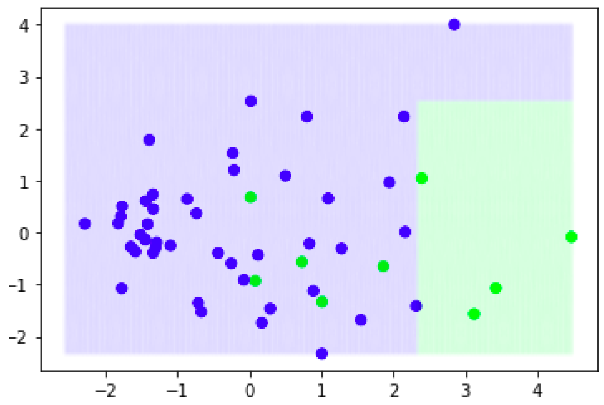
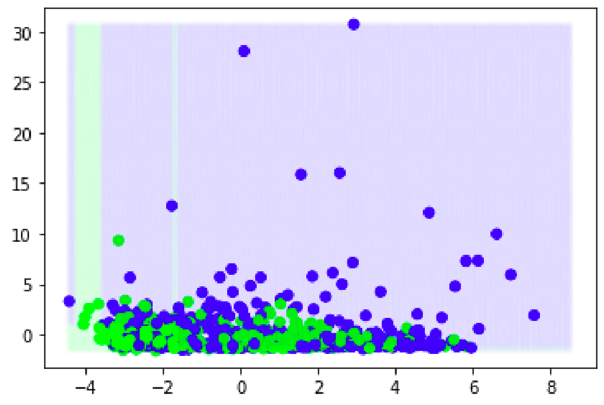
Hepatitis decision boundary (KNN) PCA	Diabetes decision boundary (KNN) PCA
	
Hepatitis decision boundary (DT) PCA	Diabetes decision boundary (DT) PCA

Fig 4. Decision Boundaries.