COMP 551 Mini Porject 1 Write-up

Group 33: Yuxuan Liu, Junxiang Mao, Kaicheng Yan

February 9, 2022

Abstraction

In this project, we investigate the performance of two machine learning models covered in class (K-Nearest Neighbour and Decision Tree) on two benchmark datasets (Hepatitis and Diabetic Retinopathy Debrecen Dataset). After necessary data cleaning, we analyze two datasets distribution to find two most important attributes as our training data features. By feeding two models with training data splitted from the original two datasets, we acquire the predict test labels and compare it with true labels to get each model's accuracy. After experimenting models using different hyperparameters, we found that with the best hyperparameters, the Decision Tree model achieves worse accuracy than the K-Nearest Neighbour model.

Introduction

Basic Introduction

K-Nearest Neighbour and Decision Tree are two widely-used machine learning techniques. They behave differently under various data environments. In this project, we compare the performance of these two models trained by two different datasets and further explore how hyperparameters would influence the models' behaviors respectively. The two datasets used in this process are patients' data related to two diseases, including individual information (their medical test results), and our predict mission: class labels. Among these information, we select two features that are most correlated with target labels as our training data. After feeding these data to our models, we observe the accuracy of K-Nearest Neighbour model with different K values and that of Decision Tree model with ascending values of maximum depth, and eventually compare the accuracy of these two models. We found that with the best hyperparameters, the accuracy of Decision Tree is worse than K-Nearest Neighbour model. Also, we found if removing the original data's outliers, the performance of two models both improve.

Related Work

In the article "Application of Data Mining Algorithms for Feature Selection and Prediction of Diabetic Retinopathy" (ICCSA 2019)(1), the authors discovered using feature selection on different machine learning algorithm could improve prediction accuracy, and the dataset that article used is same as we used in our project (Diabetes Retinopathy). Inspired by this article and considering the relatively large number of features two assigned dataset contain, we decide to choose only two features which are most correlated with target class to feed two models we are going to test. By this way, the models could make the work complexity decline but keep the necessary accuracy.

Datasets

Both Hepatitis dataset and Diabetic Retinopathy Debrecen dataset are downloaded from the urls provided in the instruction pdf, and transfered into csv files handly. We first clean the data by eliminating rows with missing values and dropping all rows containing invalid characters. Then, we transformed all the columns to specific types which lead to easier computation later on (Figure 1 and 2). We then process the data by identifying correlations among all the features and drop features which were barely correlated with dependent variables. In order to prevent multicollinearity, we also drop highly correlated attributes($\text{corr}(x_1, x_2) > 0.5$). From figure 3, the feature "euclidean distance of macula to optic disc" in Diabetic dataset shows a normal

distribution. In Hepatitis dataset, "die" label in the class attribute shows bernoulli distribution with probability 0.74. Similarly, in Diabetic Retinopathy dataset, the label which contains signs of DR in the class attribute shows bernoulli distribution with probability 0.54. For ethical challenges regarding datasets, it's unlawful and unethical to collect someone's personal data without their consent, especially in hepatitis dataset, which may involve personal health information.

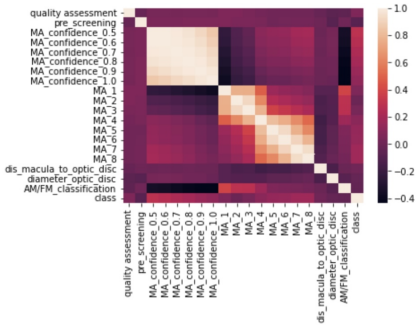

Figure 1

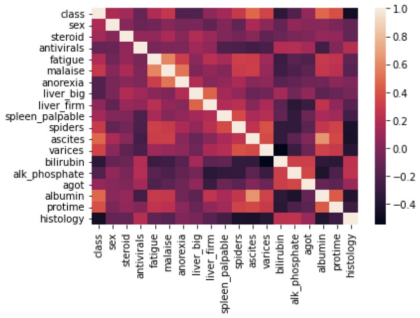

Figure 2

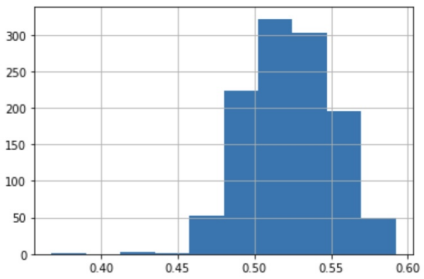

Figure 3

Results

As for the Hepatitis dataset (small data amount), we choose two features (ascites, albumin) which are most correlated with target labels. The result we got from running the K-Nearest Neighbour model under Hepatitis dataset is as follows: When the hyperparameter K is equal to 3 and when the cost function is euclidean, we get a 90% accuracy; the Decision Tree gives an accuracy of 85% under the same dataset with maximum depth equal to 4 and the misclassification cost function.

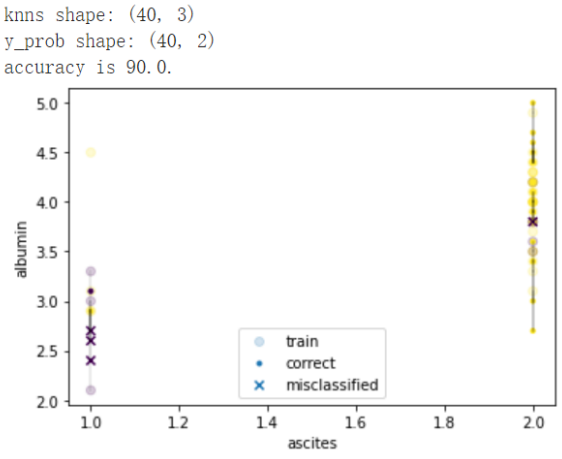

Figure 4: KNN's prediction

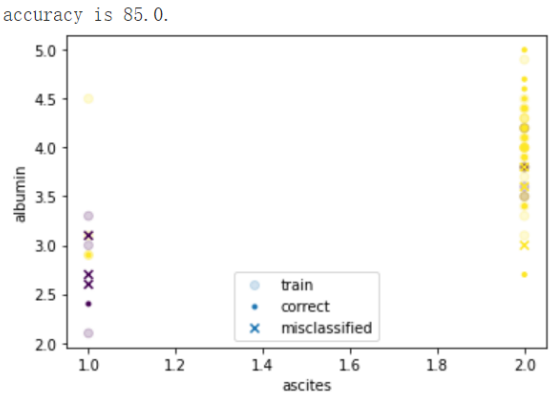

Figure 5: DecisionTree's prediction

From the above result, it is easy to see that the K-Nearest Neighbour approach achieves better accuracy than the Decision Tree approach .
Then we discovered if multiple choices of hyperparameters (K values and maximum depth) could alter the performance of two models. The following is what we got:

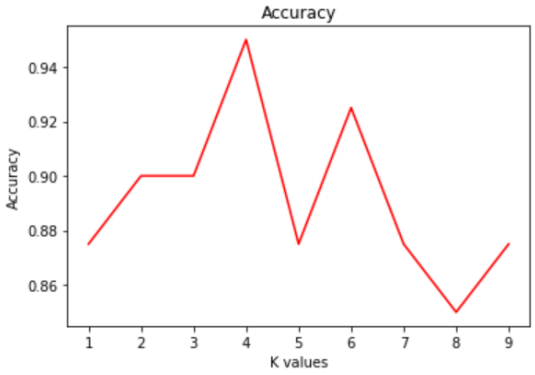

Figure 6: KNN accuracy

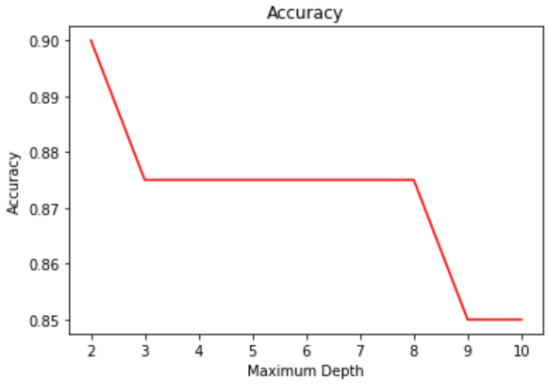

Figure 7: DecisionTree accuracy

From Figure 6 and Figure 7, we can see that in the K-Nearest Neighbour model with Euclidean cost function, the largest accuracy it can arrive (when K = 4) within the range we set is higher than the highest accuracy the Decision Tree model (misclassification cost function) can make (90%) when maximum depth is 2.
After trying Manhattan distance function in KNN model (K=3) on the second dataset, the accuracy does not improve (from 90% to 90%). If using Gini cost function in Decision Tree model in the second dataset, we find the accuracy raise from 85% to 87.5%.
The decision boundaries of both models are as follows. We can see that the live class mainly lies on the lower left part of the graphs, which means that a patient with lower albumin and negative result of ascites may have smaller probability of getting hepatitis.

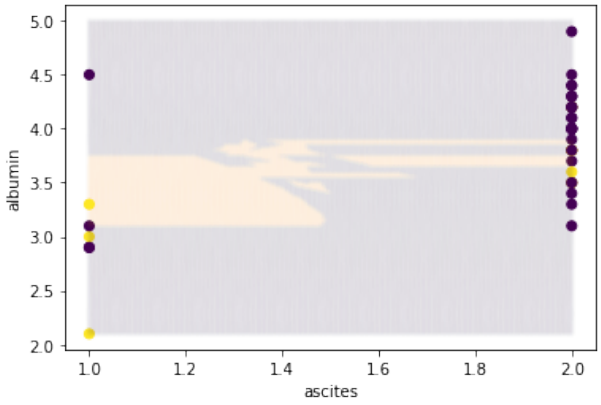

Figure 8: KNN decision boundary

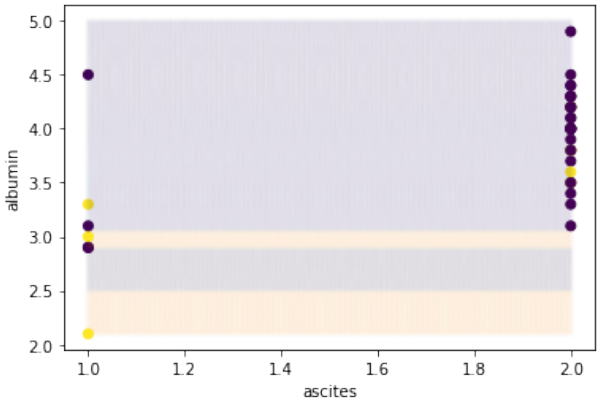

Figure 9: DecisionTree decision boundary

As for the Hepatitis Diabetic Retinopathy Debrecen Dataset (large data amount), we choose two features (MA_confidence_0.5, MA_confidence_0.6) which are most correlated with target labels. The result we got from running the K-Nearest Neighbour model under Diabetic Retinopathy Debrecen Dataset is as follows: When the hyperparameter K is equal to 3 and when the cost function is euclidean, we get a 61.8% accuracy; the Decision Tree gives an accuracy of 57.6% under the same dataset with maximum depth equal to 4 and the misclassification cost function.

```
knns shape: (576, 3)
y_prob shape: (576, 2)
accuracy is 61.8.
```

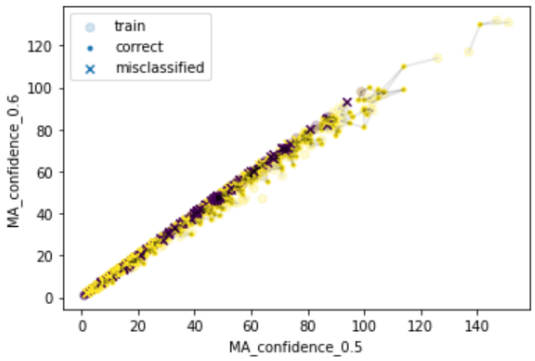

Figure 10: KNN's prediction

```
accuracy is 57.6.
```

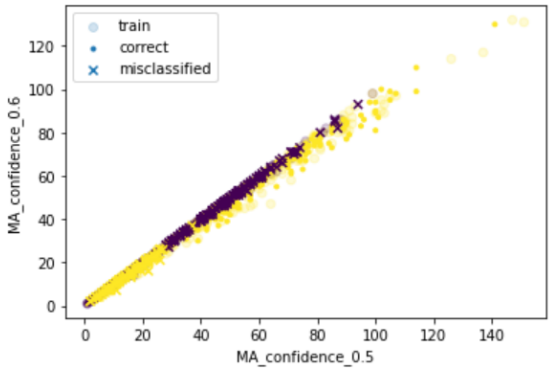

Figure 11: DecisionTree's prediction

We can see in this dataset, the K-Nearest Neighbour still achieves better accuracy than the K-Nearest Neighbour approach, even if their accuracy decrease in the same place when encountering such big data volume.

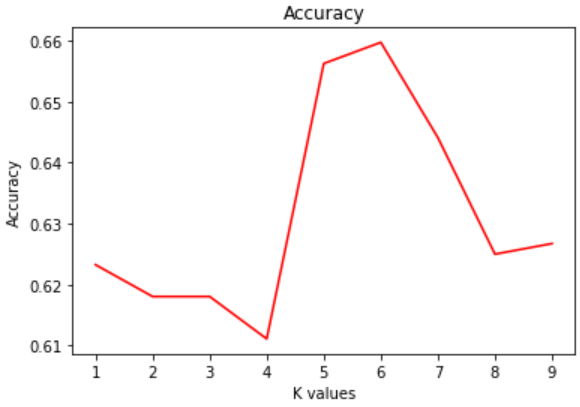

Figure 12: KNN accuracy

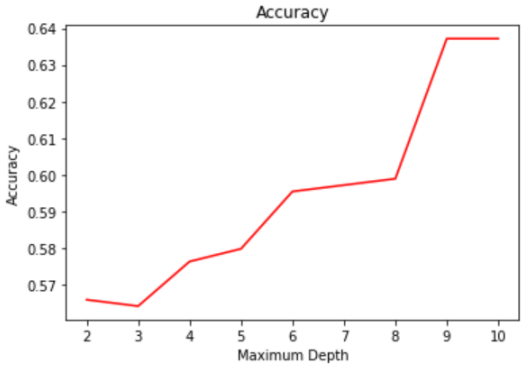

Figure 13: DecisionTree accuracy

From Figure 12, we can conclude that in the K-Nearest Neighbour model, if we use the Euclidean cost function, then when we set the value of K to be 6, we will get the largest accuracy of 66% within the K range we select; from Figure 13, we can conclude that in the Decision Tree model, if we use misclassification cost function, then a maximum depth of 9 or 10 can lead us to the largest accuracy which is 64%. Seeing the upward trending of figure 13, we suppose as maximum depth increase, the accuracy would raise as well. In addition, after trying manhatton distance function in KNN model (K=3) on the second dataset, the accuracy does not improve (from 61.8% to 61.8%). If using gini cost function in Decision Tree model in the second dataset, we find the accuracy decrease from 57.6% to 56.4%.
The decision boundaries of both models are as follows.

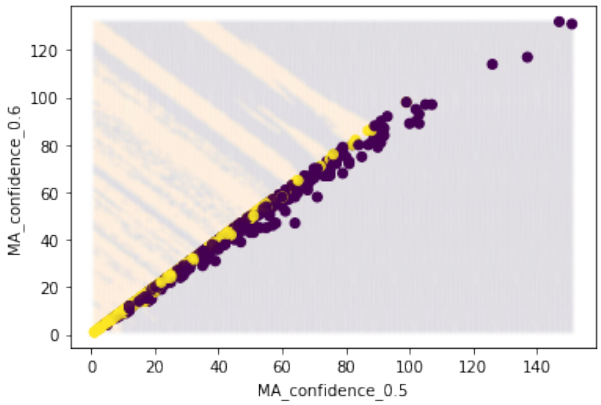

Figure 14: KNN decision boundary

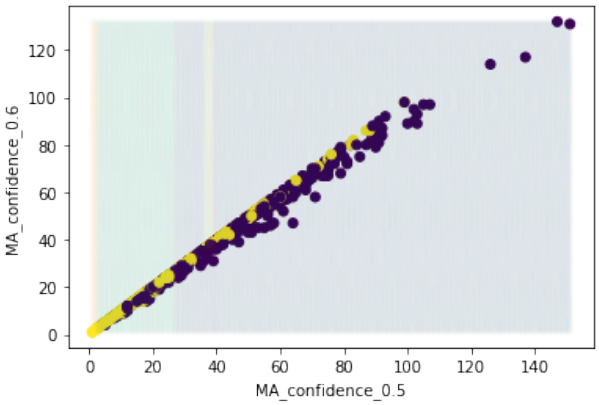

Figure 15: DecisionTree decision boundary

Discussion and Conclusion

Therefore, we can see from this project that different machine learning models have distinct predictions using different datasets, and that the optimization of hyperpatameters can greatly improve the final results. In this project, the K-Nearest Neighbour method always has better performance than the Decision Tree method. In future investigation, we may further explore the performances of these two models by taking more features into account. In addition, we see for smaller data set, the greater maximum depth shows declining accuracy, whereas for the larger data set, the deeper depth we choose for the decision tree, the greater accuracy we could have. As for the datasets themselves, we can see from our result of the Hepatitis Dataset that if a patient has ascites, it is more likely that this patient falls into the "DIE" class than patient who does not have ascites. However, after trained by the Diabetic Retinopathy Debrecen Dataset, it seems that the decision boundary plot does not give us direct conclusion. Thus in our future investigation, we could try different combinations of features, or including more features.

Statement of Contributions

In this project, we break the tasks down to the following parts: Junxiang is in charge of handling task 1, i.e. she acquires, pre-processes, analyzes, and cleans the original datasets; Kaicheng implements the K-Nearest Neighbour model and Yuxuan implements the Decision Tree models and both of them are in charge of running the experiments; and the whole group writes this report together.

References

[1] T. O. Oladele, R. O. Ogundokun, A. A. Kayode, A. A. Adegun, and M. O. Adebiyi, "Application of data mining algorithms for feature selection and prediction of diabetic retinopathy," in *International Conference on Computational Science and Its Applications*. Springer, 2019, pp. 716–730.