COMP 551 Mini Project 1

Lansu Dai, Zichuan Guan, and Lan Xi Zhu

Abstract. In this project, we investigated the performance of K-nearest neighbour (K-NN) and decision tree (DT) for classification on two benchmark datasets on Hepatitis (HEP) and Diabetic Retinopathy Debrecen (DRD). The two models performed similarly on the same datasets. However, the accuracy of both models is highly influenced by the datasets. Both models performed better on the Hepatitis dataset, which only has 80 instances and is highly imbalanced, than on the DRD dataset, which has 1147 instances and is fairly balanced. We preprocessed the dataset with normalization and performed 10-fold cross-validation to select the best hyperparameters for both models. We found that, on HEP dataset, K-NN has the highest validation accuracy when $K = 9$ using Euclidean distance, with normalization applied. This gives a test accuracy of 0.75. Similarly, for the DRD dataset, $K = 13$ using Manhattan distance and no preprocessing resulted in a test accuracy of 0.61. For DT classifier, the best validation depth was found to be 4 on both datasets, with misclassification and entropy as loss function, and a test accuracy of 0.75 and 0.6, for HEP and DRD datasets respectively.

1 Introduction

We implement and investigate the difference between K-nearest neighbour (K-NN) and decision tree classifier. The Hepatitis dataset (HEP) [1] and the Diabetic Retinopathy Debrecen (DRD) dataset[2] are used. The HEP dataset enumerates personal information, different symptoms and medical indices of tests in order to predict the risk of hepatitis. The DRD dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not.

Both datasets contain greater than ten features. The HEP has particularly unbalanced class distribution. Previous works on the Hepatitis dataset revealed multiple problems when processing this dataset. Goh[3] highlighted the importance of a balanced dataset and the usefulness of domain expertise when it comes to feature selection. The accuracy of several classification models improved after applying an oversampling technique to the dataset.

Our project task is to compare both algorithms on the two datasets for their performance. We perform 10-fold cross validation to explore the validation accuracy under different hyperparameters (K in K-NN and maximum depth in DT) and different distance/cost functions. We found that with HEP dataset, K-NN performs better with normalized features using Euclidean distance function; whereas with DRD dataset, K-NN performs better with non-normalized features using Manhattan distance function. As for the decision tree algorithm, the maximum validation accuracy was reached at a maximum depth of 4, using misclassification cost for HEP dataset, and entropy for DRD dataset.

2 Datasets

2.1 Hepatitis Dataset

The Hepatitis dataset (HEP) consists of 155 binary labeled data with 19 features. This dataset contains several missing values, notably for the feature PROTIME which has 67 missing entries, consisting of 43% of the entire dataset. Instances with missing values are removed, which leaves us with 80 data points.

The dataset is highly imbalanced, with 13 deaths and 67 alive. To cover the most data points, we chose a train-test split of 90% training data. Due to this split ratio, the test set is limited and thus the test accuracy varies significantly depending on the randomly selected test data.

2.2 Diabetic Retinopathy Debrecen (DRD) Dataset

The DRD dataset is considerably larger. It consists of 1151 records with 19 features and binary labels. The dataset does not have any missing entry. All instances with bad quality (the value of the binary 'Quality' feature is 0) are removed, which leave us with 1147 instances. The 'Quality' feature is thus removed from subsequent analysis.

This dataset is relatively more balanced than the HEP dataset. 611 out of the 1147 instances contain signs of diabetic retinopathy. 1053 of the cases have a pre-screening result of 1, indicating severe retinal abnormality. Surprisingly, only 549/1053 are observed to have signs of DR.

Although medical-related datasets are generally not attributable to individual patients, some ethical concerns might arise from the distribution of the features. For example, 69/80 of the instances in the HEP dataset are from male patients. Without rigorous studies, this dataset might lead to the presumption that males are at higher risk of hepatitis. Such biases might lead to discrimination, especially when the diseases in question are transmissible.

3 Results

3.1 K-NN

The K-NN model is known to be sensitive to feature scaling. Since the values of the features in our datasets differ by orders of magnitude (with a mixture of categorical and numerical features), we normalize all features by mean-removal and scaling to standard deviation prior to applying the K-NN classifier. (We have also attempted scaling all numerical feature values to within 0-1, which resulted in lower accuracy.)

We use 10-fold cross-validation to determine the value of hyperparameter K. The results are plotted in Fig. 1 and Fig. 2. For the HEP dataset with normalized features, the model reaches a validation accuracy of 0.92 when $K = 9$ while using Euclidean distance function, and an accuracy of 0.90 when $K = 7$ while using Manhattan distance function. Using original non-normalized features result in lower accuracy (0.83) with both distance functions. For the DRD dataset, however, the model has better performance with original features. We obtain a maximum accuracy of 0.68 when $K = 13/15$ while using Manhattan/Euclidean distance function, respectively. Using normalized features result in a validation accuracy of 0.66 in both cases.

The test accuracy using different K values are plotted in Fig. 3 and Fig. 4. The resulting test accuracy for the DRD dataset (0.61) is comparable to its validation accuracy shown earlier, whereas the HEP dataset

has lower test accuracy (0.75) as compared to its validation accuracy.

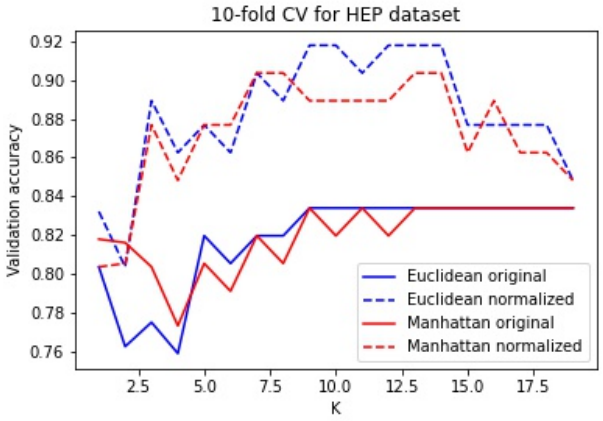

Figure 1: 10-fold cross-validation accuracy with respect to hyperparameter K for the HEP dataset.

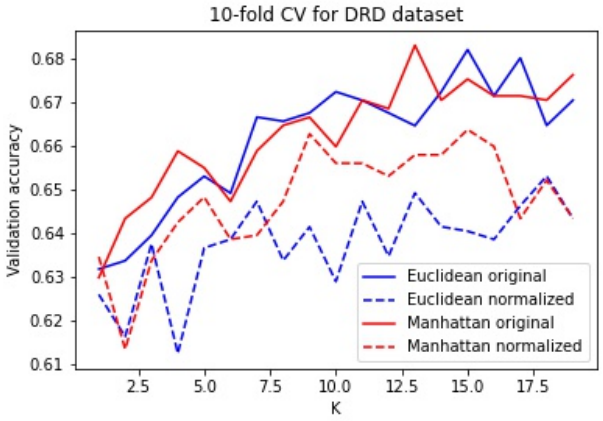

Figure 2: 10-fold cross-validation accuracy with respect to hyperparameter K for the DRD dataset.

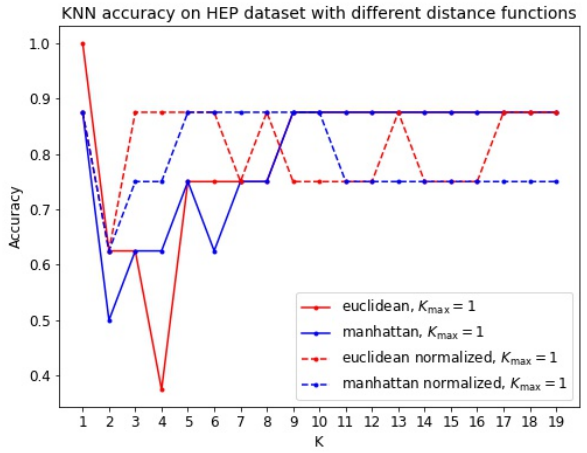

Figure 3: Test accuracy with respect to K for the HEP dataset.

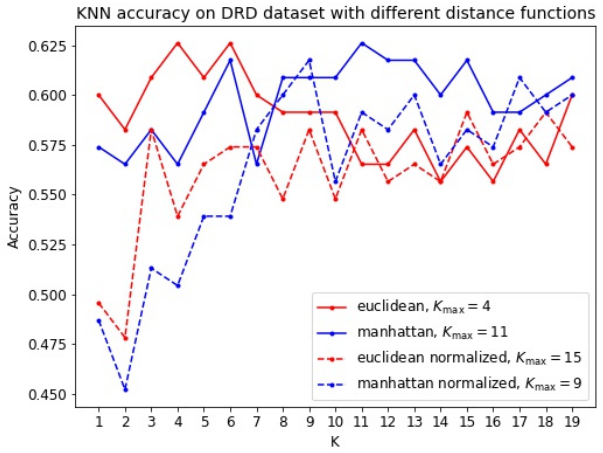

Figure 4: Test accuracy with respect to K for the DRD dataset.

3.2 Decision Tree

Unlike the K-NN model, decision tree does not require data normalization. In Fig. 5 and Fig. 6, we compare three different cost functions: misclassification cost, entropy, and Gini index. For the HEP dataset, the misclassification cost method results in a maximum validation accuracy of 0.86 (depth = 4). For the DRD dataset, the entropy method results in a maximum validation accuracy of 0.64 (depth = 4).

As shown in Fig. 7, the test accuracy on the HEP dataset results in a constant value (0.75) for different maximum decision tree depth. This is likely due to the lack of number of test instances (as mentioned before, the HEP dataset has limited size). For the DRD dataset, the test accuracy is again comparable (0.6) to validation accuracy in Fig. 6.

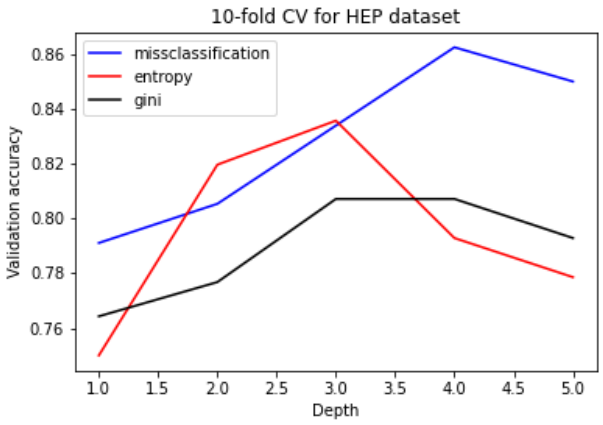

Figure 5: 10-fold cross-validation accuracy with respect to maximum depth for the HEP dataset.

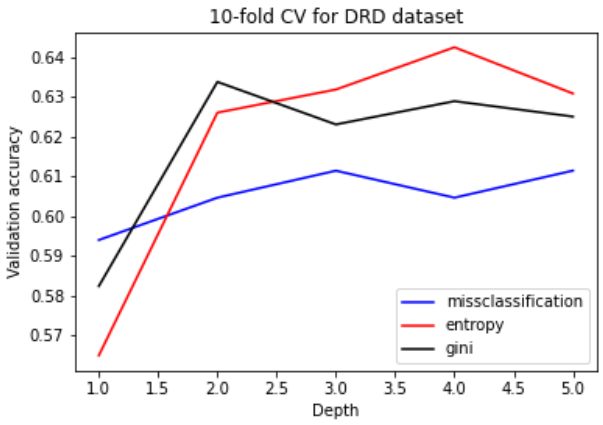

Figure 6: 10-fold cross-validation accuracy with respect to maximum depth for the DRD dataset.

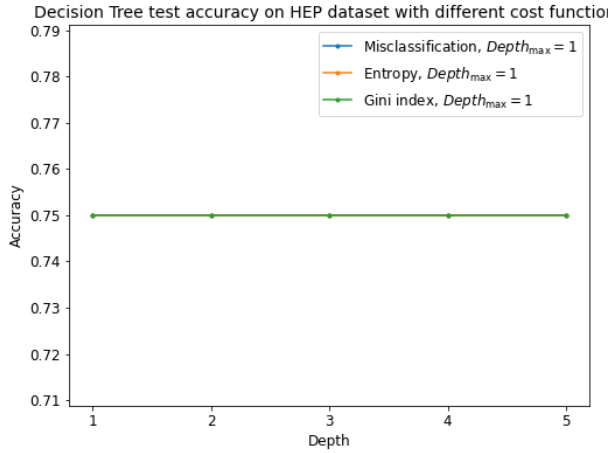

Figure 7: Test accuracy with respect to maximum depth for the HEP dataset.

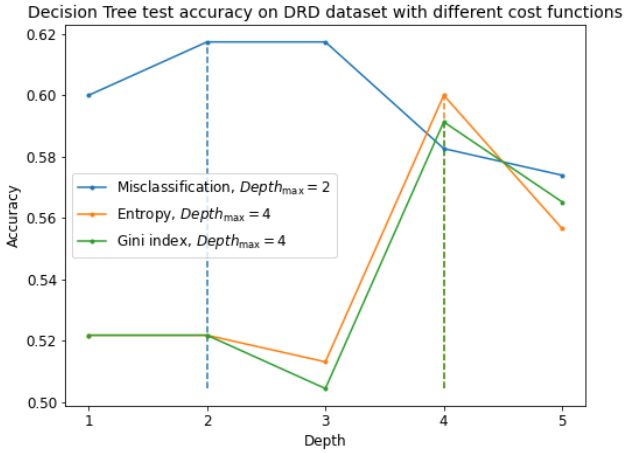

Figure 8: Test accuracy with respect to maximum depth for the DRD dataset.

3.3 Decision Boundaries

To plot the decision boundaries of both models, both datasets are reduced to two-dimensional. As shown in Fig. 9, for all pairs of numerical features in the HEP dataset, the instances belonging to either classes are closely mixed together. Instances in the DRD dataset are also indistinguishable in 2D plots. Therefore, we do not expect clear decision boundaries under a two-dimensional context. Our attempts to plot 2D decision boundaries are shown in Fig. 10.

4 Discussion and Conclusion

We investigate the performance of two models, K-NN and Decision tree, on two datasets with different sizes. With a total of 80 instances and therefore limited test set size, we observe overfitting in Fig. 3 where the test accuracy peaks at $K = 1$. The effect of this limitation is more clearly shown in Fig. 7, where the test accuracy at different maximum depth form a straight line (i.e. no further splitting is needed even

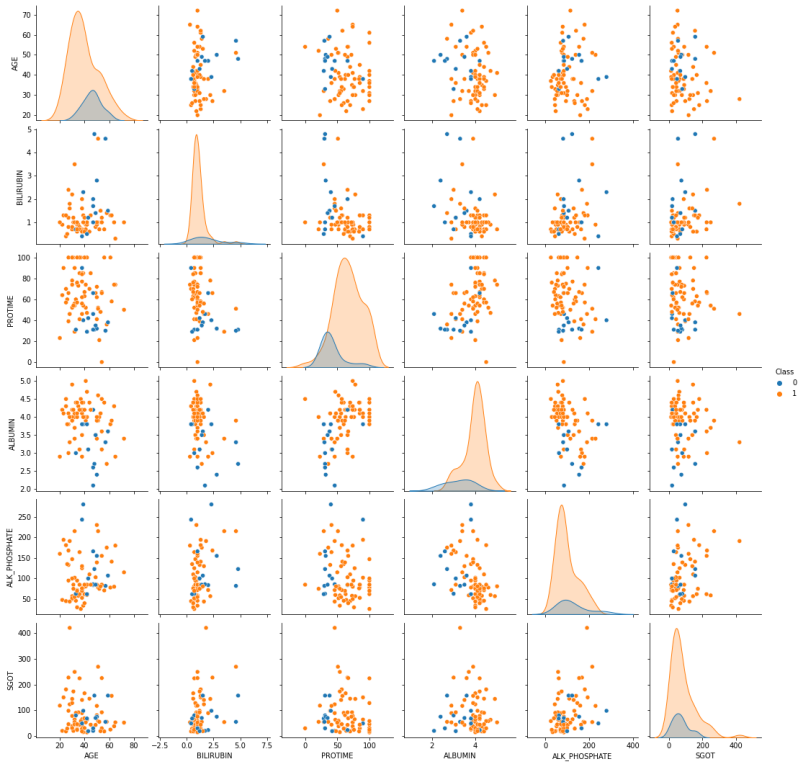

Figure 9: Pair plot showing the distribution of instances with respect to the six numerical features in the HEP dataset. Data points with different colors belong to different classes.

when a larger depth is allowed). The hyperparmeters we chose from cross-validation also did not result in the best test accuracy for HEP. However, with the limited data size of HEP, this is to be expected. We believe that the chosen hyperparameters will result in better generalization with larger datasets.

In Fig. 2 for DRD, the larger dataset, we observe that the validation accuracy increases with K and plateaus after reaching maximum accuracy, which agrees with our expectation.

The effect of normalizing the data is significant for K-NN, while not observable for DT. It agrees with the fact that K-NN is sensitive and scaling while DT is robust against noises. The performance of both models could be improved by applying more rigorous methods while pre-processing the input features. For example, we could assign different weights to the features based on deeper understanding of their properties.

Statement of Contributions

Everyone contributed equally to the coding component and the project write-up. We have not assigned specific part of the project to individual group member.

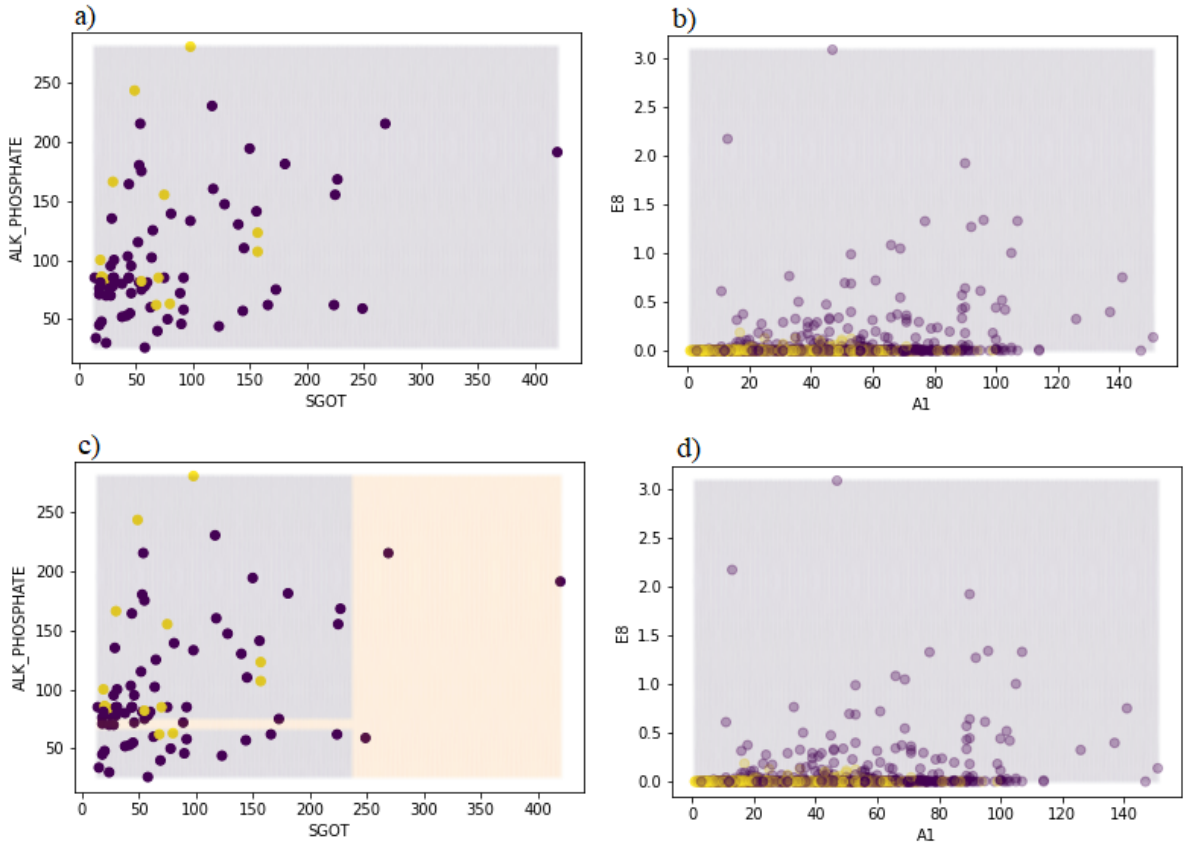

Figure 10: 2D Decision boundaries. a) HEP dataset, K-NN (K=9, Euclidean); b) DRD dataset, K-NN (K=13, Manhattan); c) HEP dataset, Decision tree (depth=4, misclassification cost); d) DRD dataset, Decision tree (depth=4, entropy).

References

[1] G.Gong, "UCI machine learning repository," 1988.

[2] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," Oct 2014.

[3] M. Goh, "Predicting hepatitis patient survivability (uci dataset)," Nov 2020.

[4] R. Safdari, A. Deghatipour, M. Gholamzadeh, and K. Maghooli, "Applying data mining techniques to classify patients with suspected hepatitis c virus infection," *Intelligent Medicine*, 2022.

[5] J. Novaković, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav Journal of Operations Research*, vol. 21, no. 1, 2016.