Project1

Group 20

February 8, 2022

1 Abstract

In this project we studied the performance of two machine learning models, K-Nearest Neighbors (KNN) and Decision Tree (DT), on two health datasets. Overall, the Decision Tree approach achieved better training and test data accuracy, but both models did not generalize well on the large size dataset. The hyper-parameter (K value and maximum tree depth) values greatly affected the model performance, but the distance/cost function choice was not significant. Furthermore, we experimented on various methods to improve model performance, like reducing redundant features for KNN and using cross-validation for DT.

2 Introduction

In this project, we first pre-processed the two datasets and statistically analyzed them in order to gain a better understanding of the distribution of different classes of data. Then we implemented the algorithms from scratch and ran several experiments to compare the model performance and explore the impact of some parameters.
For the Messidor dataset, the best models were KNN with K=21 and DT with depth 11. For the Hepatitis dataset, the best models were KNN with K=7 and DT with depth 4. Each model had its own choice of hyper-parameter and distance/cost function. Generally speaking, both models did not generalize well on Messidor dataset, but we can choose a suitable way to improve their accuracy. For the KNN model, we have found that choosing a proper distance function and K value contribute to archiving better performance. For the Decision tree model, we have seen a steady improvement by cross-validation.
For many years, classification has been an extensively studied machine learning task. To study classification, these two dataset has been cited in many papers. As an example, the *Extracting symbolic rules from trained neural network ensembles* used in Hepatitis dataset [1] ran experiments on this dataset to test accuracy of extracting rules. The *An ensemble-based system for automatic screening of diabetic retinopathy* [2] used Messidor dataset to run experiments on proposed ensemble-based method for the screening of diabetic retinopathy.

3 Datasets

3.1 Dataset Description

Messidor dataset contains features to predict whether an image contains signs of diabetic retinopathy or not. There are 1151 instances, 19 features and 1 class label. Two features are binary, the others are continuous. The class label 1 = contains signs of DR, 0 = no signs of DR.
Hepatitis dataset discovers the relationship between different feature of the patients such as sex, age and fatigue and their death. There are 19 features, including 13 binary features, and 2 classes. The class label "1" represents "die" while label "2" represents "live".

3.2 Data cleaning and pre-processing

First, we removed all the missing values in the two datasets. There are 1151 instances in Messidor and 80 instances in Hepatitis after cleaning. Then, we converted class labels to binary values. There

is no need to convert regarding Messidor dataset. For Hepatitis, we converted class label "2" into "0" to represent "live", and "1" was unchanged, representing "die". Finally, we divided the datasets into training set and test set by a ratio of approximately 2:1.
For KNN train/test data, we normalized the continuous features to between [0, 1] so that the continuous features will not weigh more than categorical features.

3.3 Data Analysis

In Messidor dataset, 540 out of 1151 are labelled "0" and 611 are labelled "1". Hepatitis dataset 67 are labelled "0" and 13 are labelled "1".
We also analyzed the data distribution of each attribute for the negative and positive classes respectively. We found that for Messidor dataset, some continuous features are highly skewed, and the distributions looked similar between the two classes. This may cause low accuracy using KNN for classification because it is difficult to distinguish the data points. In comparison, for Hepatitis dataset, the distributions are apparently different between two classes. Figures are in the Appendix.

3.4 Ethical Concerns

There are ethical concerns with collecting biological data. For Messidor dataset, there may be issues regarding consent before gathering patients' privacy information. For Hepatitis dataset, there may also be concerns on the bias of gender and age information. These incorrect information may lead to erroneous predictions for different distribution of training dataset and real dataset.

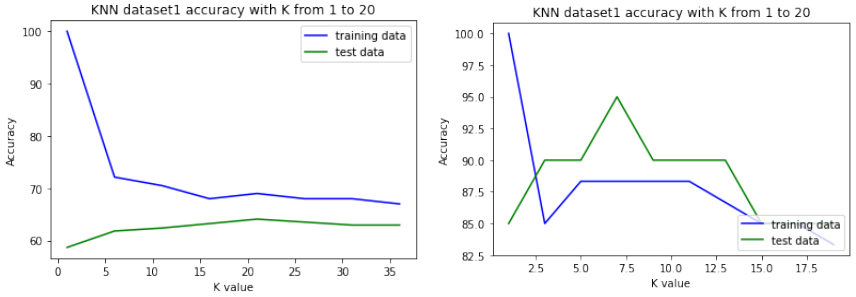
4 Results

4.1 Comparison of model performance

The decision trees gave the best training data accuracy (100%) on both datasets. For Messidor dataset, the decision tress gave slightly higher test data accuracy (69% vs. 64%). For Hepatitis dataset, KNN and DT performed equally well (95%).

4.2 Impact of hyper-parameters

For KNN model, both the training and testing data accuracy are lower when the K value is either too small or too large, since small k overfits and large k underfits. One exception is when k=1, in this case the training accuracy is 100 percent because KNN just choose the training point itself as the nearest neighbour (distance is zero). We got the highest test accuracy when K=20 for Messidor and K=7 for Hepatitis.

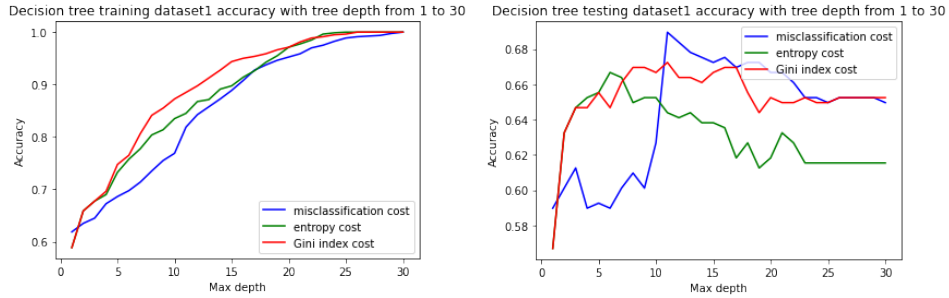


(a) Messidor dataset

(b) Hepatitis dataset

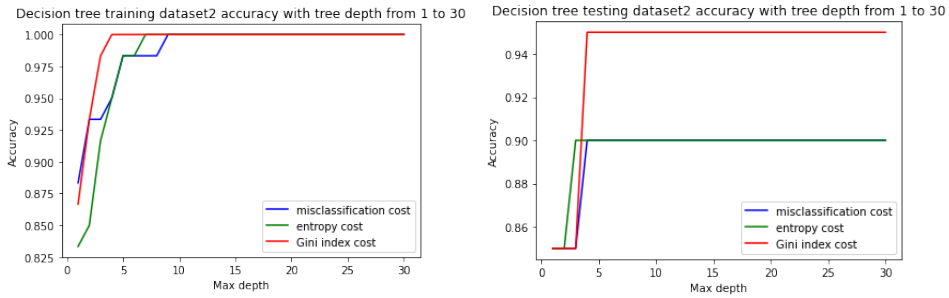
Figure 1: KNN training and test accuracy

For the Decision Tree, with increasing maximum tree depth, training data accuracy always increases to 100 percent, but test data accuracy may decrease because of overfitting. We got the highest test accuracy when depth=11 for Messidor and depth=4 for Hepatitis.



(a) Training accuracy (b) Test accuracy

Figure 2: DT training and test accuracy for Messidor dataset



(a) Training accuracy (b) Test accuracy

Figure 3: DT training and test accuracy for Hepatitis dataset

4.3 Impact of different distance/cost functions

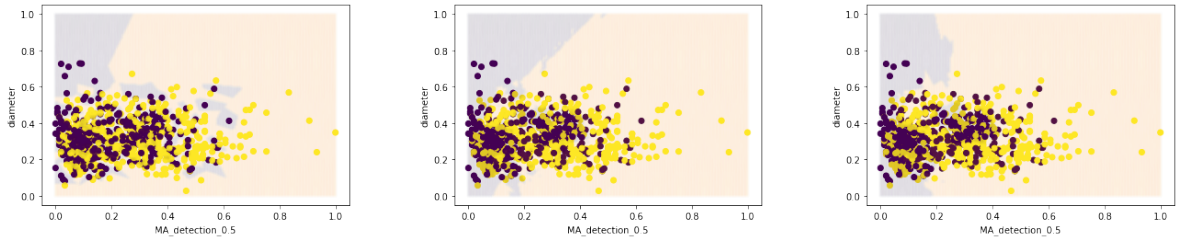
For distance functions, we experimented on euclidean, manhattan and hamming distances with the same set of K values. Euclidean and manhattan gave similar results, except they achieved the highest accuracy at different K values. Hamming gave much lower accuracy because it should be used for discrete features, but many of our features were continuous.

For cost functions, we experimented on misclassification cost, entropy cost and Gini index function. While misclassification cost function got the highest test accuracy for Messidor, Gini index function performed better for Hepatitis.

In conclusion, there is no distance/cost function that works well in all cases, we need to choose an appropriate one based on the datasets.

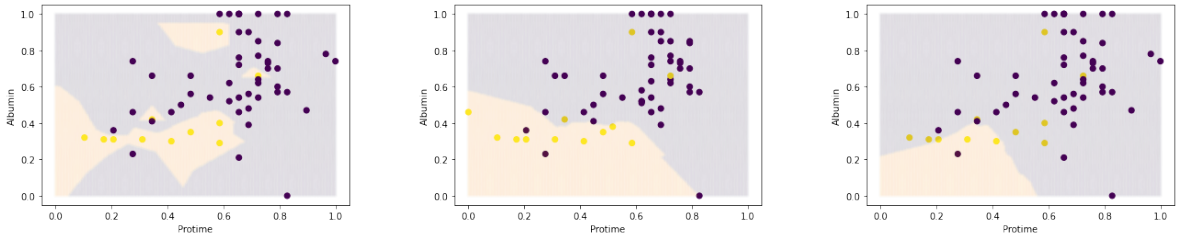
4.4 Plot the decision boundary

To visualize the data in 2D plots, we selected only two continuous features as x and y axis. Figure 4 and 5 demonstrate the plots of decision boundary for KNN. We find that when K is small, KNN tries to separate negative training points from positive ones completely. On the other hand, when K is large, an unseen data is more likely to be predicted as the category that dominates the training set, so area of this category becomes larger. Figure 6 demonstrates plots of decision boundary for DT with best tree depth, compared to KNN plots, the key feature of this plot is that the decision boundaries are linear.



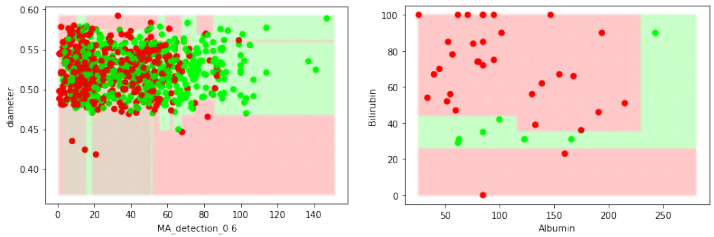
(a) $K = 1$ (b) $K = 21$ (c) $K = 40$

Figure 4: KNN: Plots of decision boundary for Messidor



(a) $K = 1$ (b) $K = 7$ (c) $K = 15$

Figure 5: KNN: Plots of decision boundary for Hepatitis



(a) for Messidor (b) for Hepatitis

Figure 6: DT: Plots of decision boundary

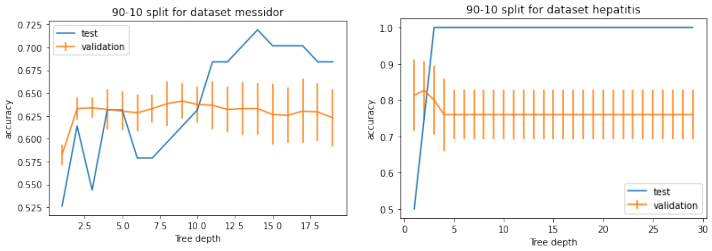
4.5 Other trials to improve accuracy

Select features for KNN training process

In Messidor dataset, feature 2-7 are highly correlated, having correlated features will increase the distance between points. To improve the performance of KNN, we kept the feature that has the highest correlation with "Class". Similar techniques were used for feature 8-15. Unfortunately, the accuracy of our KNN model did not improve. We think the reason is that the feature distributions are highly similar between the positive and negative classes, and the data points are cluttered together in the feature space. Thus it is difficult for KNN to distinguish them.

Use cross-validation for Decision Tree

To improve the performance of the decision tree, we tried to use 10-fold cross-validation. The best test accuracy was improved to 0.72 for Messidor and 1.0 for Hepatitis.



(a) for Messidor (b) for Hepatitis

Figure 7: DT: 10-fold cross validation

5 Discussion and Conclusion

For Messidor dataset, the decision tree performed slightly better, while for Hepatitis dataset KNN and DT performed equally well. Each model has its best K value or tree depth, too small or too large hyper-parameters can cause problems. Performance of different distance/cost functions is based on dataset, no function works well for all the cases.
For future investigation, it would be meaningful to explore how other hyper-parameters can affect the performance of the two models. For KNN, we would like to explore how data with mixed feature types can be better pre-processed to improve accuracy. For DT, we can explore the impact of the number of leaf nodes and different stopping criteria.

6 Statement of Contributions

Wen Cui did the datasets processing and analysis, Evelyn Cao did the KNN implementation and Zhengxin Chen did the decision tree implementation. For the report, Wen Cui wrote the Introduction and Dataset section. Abstract, Results, Discussion and Conclusion were split between Evelyn Cao and Zhengxin Chen.

7 Appendix

	quality_assessment	pre_screening	MA_detection_0.5	MA_detection_0.6	MA_detection_0.7	MA_detection_0.8	MA_detection_0.9	MA_detection_1.0	MA_normalized_1
count	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000
mean	0.996525	0.918332	38.428323	36.909644	35.140747	32.297133	28.747176	21.151173	64.096674
std	0.058874	0.273977	25.620913	24.105612	22.805400	21.114767	19.509227	15.101560	58.485289
min	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.349274
25%	1.000000	1.000000	16.000000	16.000000	15.000000	14.000000	11.000000	8.000000	22.271597
50%	1.000000	1.000000	35.000000	35.000000	32.000000	29.000000	25.000000	18.000000	44.249119
75%	1.000000	1.000000	55.000000	53.000000	51.000000	48.000000	43.000000	32.000000	87.804112
max	1.000000	1.000000	151.000000	132.000000	120.000000	105.000000	97.000000	89.000000	403.939108

	Class	Age	Sex	Antivirals	Histology
count	80.000000	80.00000	80.000000	80.000000	80.00000
mean	1.837500	40.66250	1.137500	1.737500	1.41250
std	0.371236	11.28003	0.346547	0.442769	0.49539
min	1.000000	20.00000	1.000000	1.000000	1.00000
25%	2.000000	32.00000	1.000000	1.000000	1.00000
50%	2.000000	38.50000	1.000000	2.000000	1.00000
75%	2.000000	49.25000	1.000000	2.000000	2.00000
max	2.000000	72.00000	2.000000	2.000000	2.00000

Figure 8: total statistics

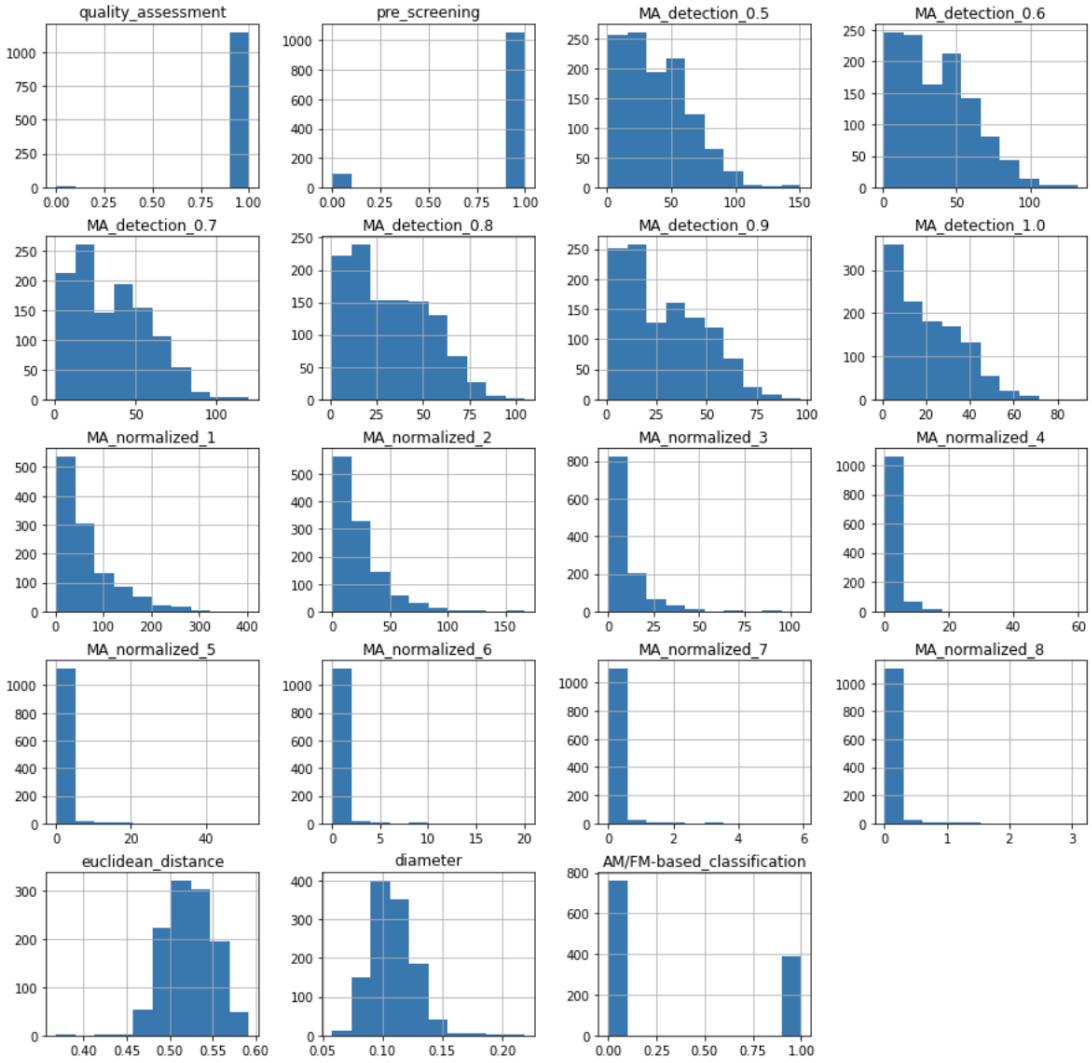


Figure 9: Diabetic Retinopathy Debrecen dataset class "0" distribution

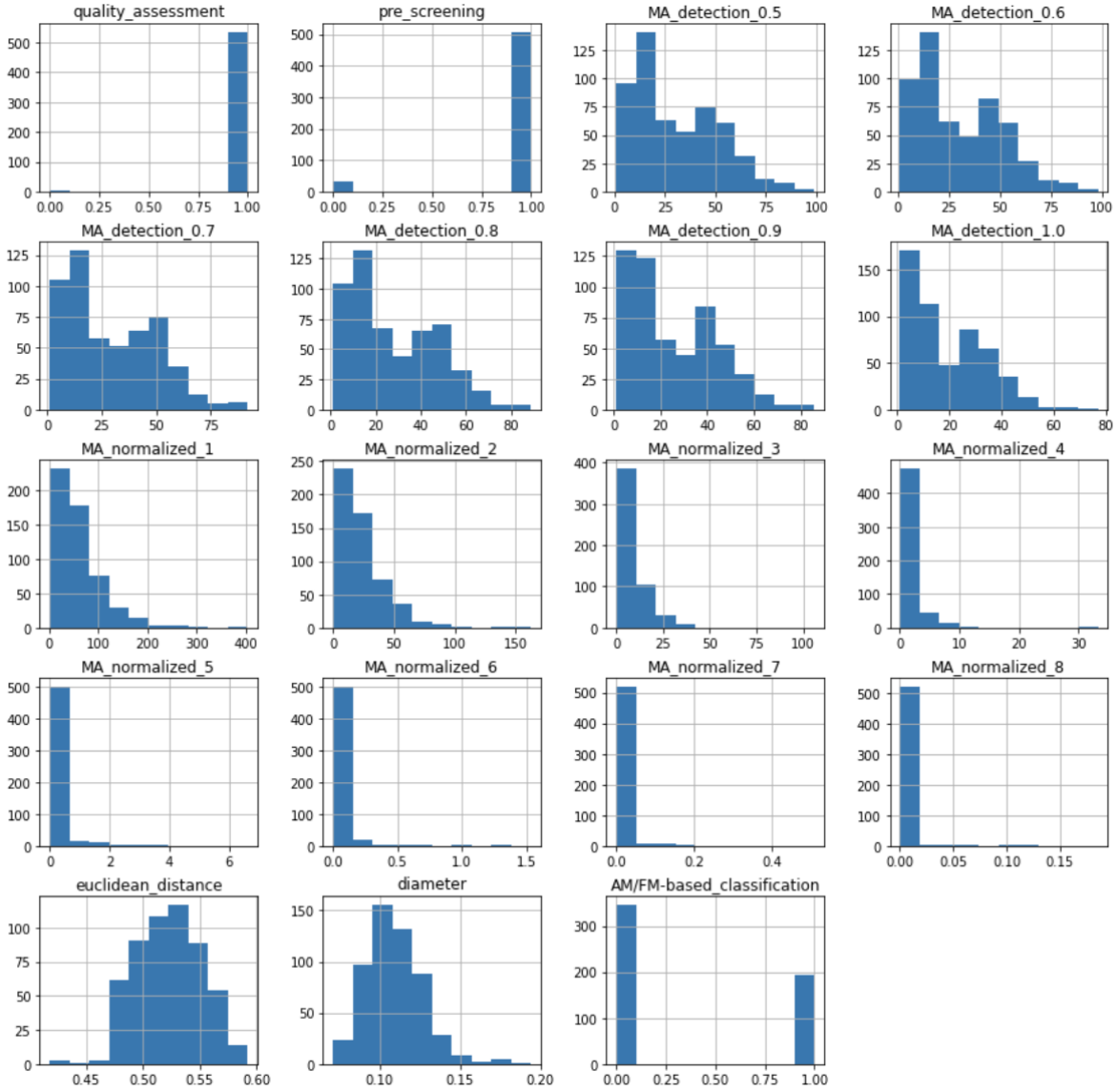


Figure 10: Diabetic Retinopathy Debrecen dataset class "1" distribution

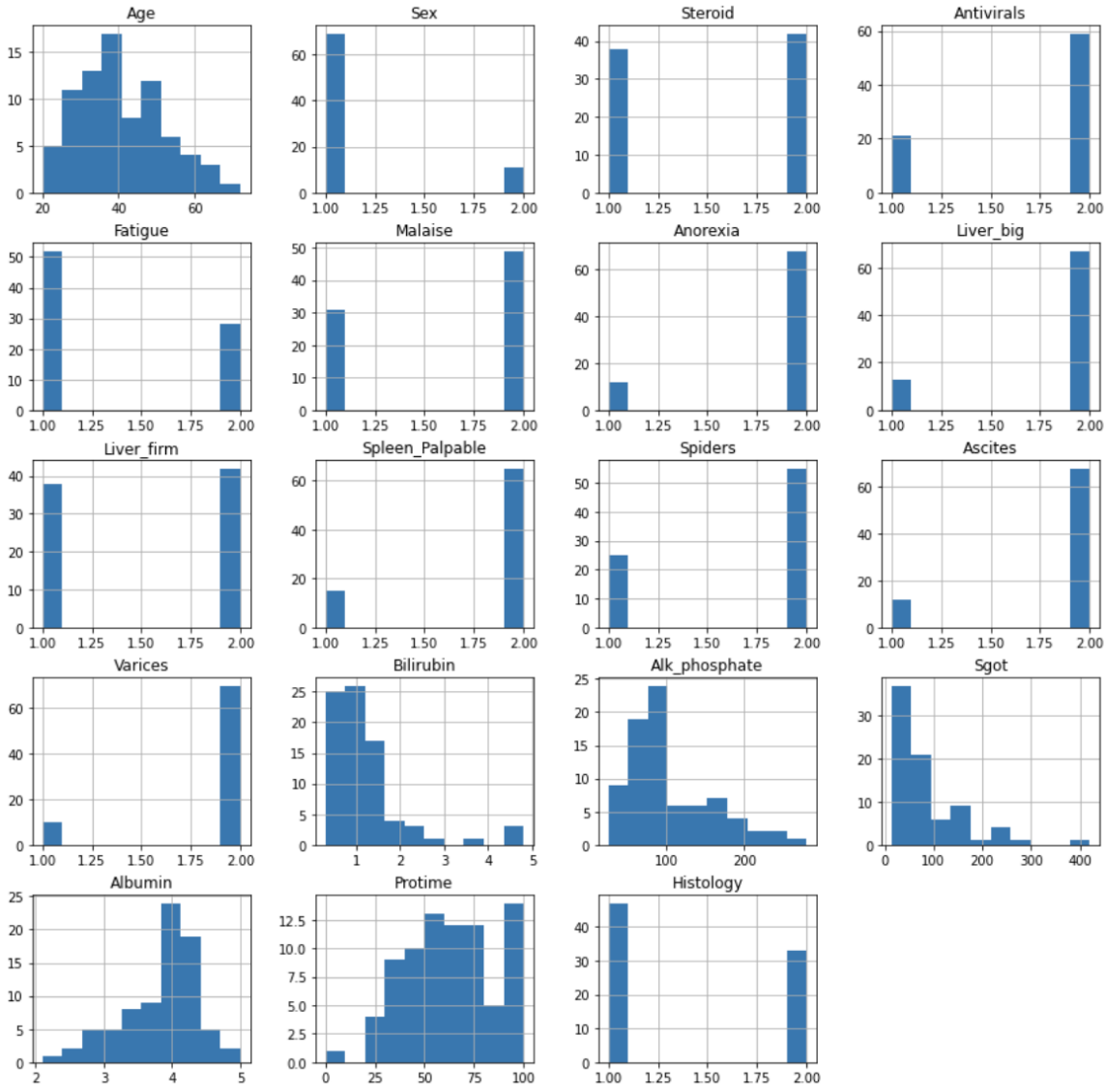


Figure 11: Hepatitis dataset class "0" distribution

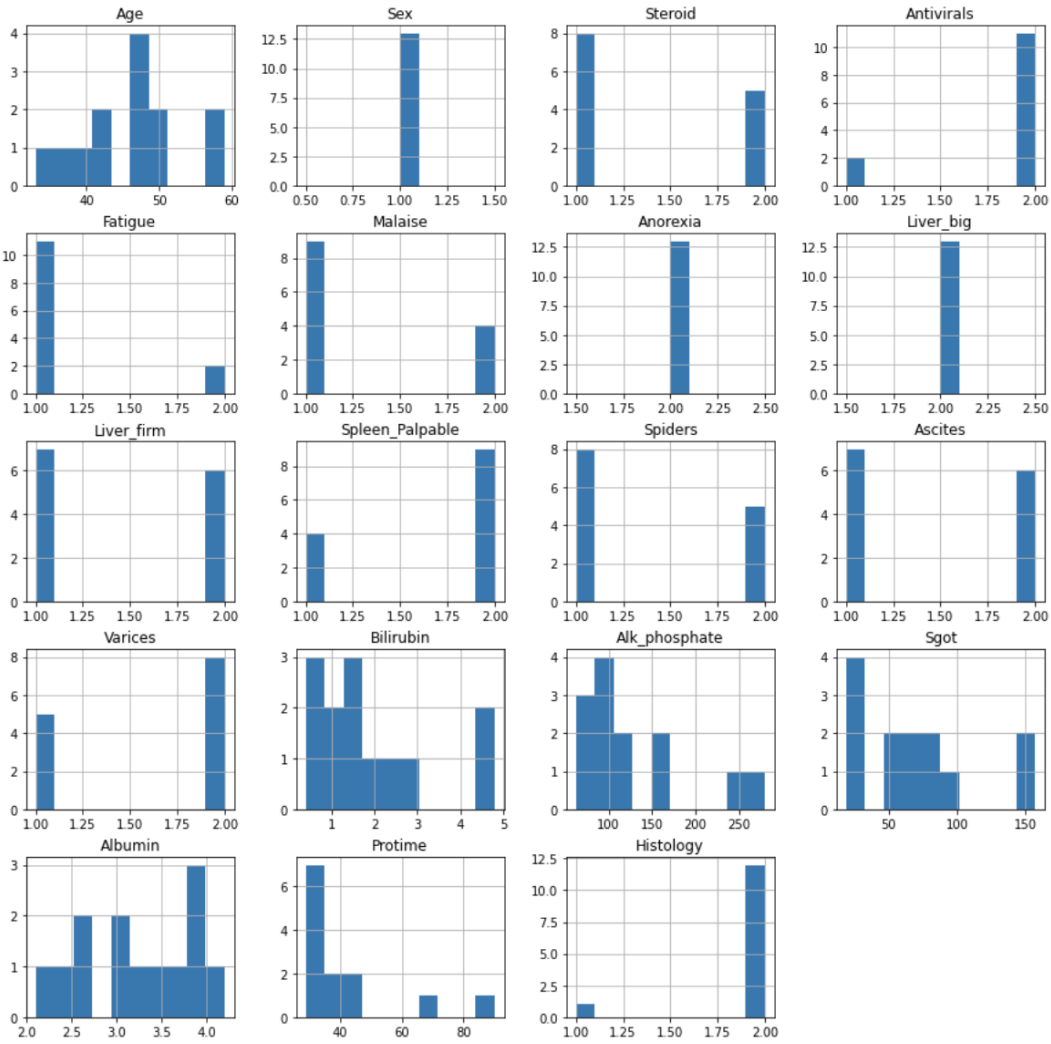


Figure 12: Hepatitis dataset class "0" distribution

References

[1] Y. C. S. Zhou, Zhi-Hua; Jiang, "Extracting symbolic rules from trained neural network ensembles," *IOS Press*, 2003.

[2] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-Based Systems*, vol. 60, pp. 20–27, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705114000021