A Comparison of Naive Bayes and Random Forest on Predicting Liver Disease
Supplementary Material

Name: Sajeel Nadeem Alam
Email: sajeel.alam@city.ac.uk
Coursework: INM431 Machine Learning

Glossary

The goal of this study was to predict whether a patient has liver disease or not depending on their features. These features include their age, gender and certain blood tests which have been defined in this section. Since these are medical terms, for clarity and to avoid any confusion direct quotes have been mentioned from the sources that have been cited.

Term	Definition
Total Bilirubin (TB)	"Bilirubin is a substance produced during the breakdown of red blood cells. Bilirubin passes through the liver and is excreted in stool. Higher levels of bilirubin might mean liver damage or disease. At times, conditions such as a blockage of the liver ducts or certain types of anemia also can lead to elevated bilirubin" [1]
Direct Bilirubin (DB)	"Direct bilirubin (sometimes referred to as conjugated) is the form of bilirubin which has been conjugated with glucuronic acid and is excreted in the bile. Measurement of this metabolite is of assistance in diagnosis and monitoring of the many disease states associated with raised bilirubin" [2]
Alkaline Phosphatase (Alkphos)	"Alkaline Phosphatase is an enzyme found in the liver and bone and is important for breaking down proteins. Higher-than-usual levels of ALP may mean liver damage or disease, such as a blocked bile duct, or certain bone diseases, as this enzyme is also present in bones" [1]
Alanine Aminotransferase (Sgpt)	"Alanine Aminotransferase is an enzyme found in the liver that helps convert proteins into energy for the liver cells. When the liver is damaged, ALT is released into the bloodstream and levels increase. This test is sometimes referred to as SGPT" [1]
Aspartate Aminotransferase (Sgot)	"Aspartate Aminotransferase is an enzyme that helps the body break down amino acids. Like ALT, AST is usually present in blood at low levels. An increase in AST levels may mean liver damage, liver disease or muscle damage. This test is sometimes referred to as SGOT" [1]
Total Proteins (TP)	"A total protein test measures the amount of protein in your blood. Proteins are important for the health and growth of the body's cells and tissues. If your total protein level is low, you may have a liver or kidney problem, or it may be that protein isn't being digested or absorbed properly" [3]
Albumin (ALB)	"Albumin is one of several proteins made in the liver. Your body needs these proteins to fight infections and to perform other functions. Lower-than-usual levels of albumin and total protein may mean liver damage or disease. These low levels also can be seen in other gastrointestinal and kidney-related conditions" [1]

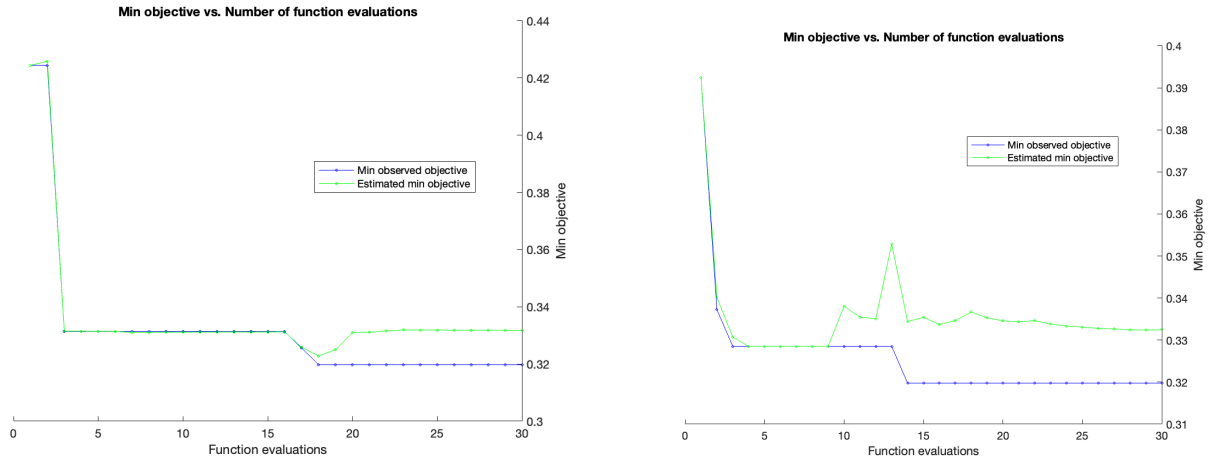
Albumin and Globulin Ratio (A/G Ratio)	"The A/G ratio is a measure of the amount of albumin proteins in blood compared to globulins. Typically, your body has slightly more albumin than globulins. A normal A/G ratio is slightly more than 1" [4]

Intermediate Results

Attempt at balancing the dataset

Initially this study was balancing the dataset to make the minority class equal to the majority class using random oversampling. I was making the mistake of doing this on the entire dataset, before splitting into training and testing. This resulted in a data leakage and gave false high accuracy values because some duplicates of the training data were present in the test data. In this case, for Random Forest a validation accuracy of 81.4% and testing accuracy of 81.5% was obtained with an AUC value of 0.9. After realising the mistake, I tried balancing only the training dataset. First I used random oversampling on the training set, and trained the model using 10 fold cross validation but this resulted in data leakage within the folds because the duplicates of some data points that were in the training set of the fold were also present in its validation set. In this case, for Random Forest a high validation accuracy of 86.2% and a poor testing accuracy of 63.16% was obtained because the model was unable to be evaluated effectively during training leading to poor hyperparameter fine-tuning. I tried balancing using SMOTE instead of random oversampling and got identical results. Eventually I realised that the appropriate way to balance the data without data leakage is to balance only the training set of each fold in cross validation and this would have to be done during cross validation. Due to time restrictions, it was decided to continue training with the unbalanced dataset.

Minimum Objective (Cross Validation Loss) vs Function Evaluations in Bayesian Optimization (without Feature Selection)



The graphs above (left: Naive Bayes, right: Random Forest) show the cross validation loss against function evaluations. In each function evaluation, different hyperparameter values are used to estimate which values give the lowest validation loss. The values that give the lowest observed and estimated losses are returned and used as starting points for the grid search for further fine-tuning. For Random Forest, the Bayesian Optimisation was giving NumLearningCycles: 447, MaxNumSplits: 1, NumVariablesToSample: 1 as the parameters for estimated minimum objective however this was giving

high validation error during training hence I selected the parameters used for the minimum observed objective instead as a starting point and did the same for Naive Bayes for a fair comparison.

Implementation Details

Initially the goal of the study was to compare Decision Trees and Random Forest, to see the effects of ensembling on evaluation metrics and whether a simple model can outperform a more complex one when dealing with a small dataset. However, after discussing with peers and professors it was decided to compare Naive Bayes and Random Forest since they are both probabilistic models thus allowing for a more fair comparison.

The Selector column in the dataset is the class column and it indicates whether the patient has liver disease or not. In the original dataset, a value of 2 meant no liver disease and 1 meant liver disease. However, for convenience and better interpretability this was replaced to 0 denoting no liver disease and 1 denoting liver disease. The Gender column contained 'Male' and 'Female' which were label encoded to 1 and 0 respectively so that machine learning algorithms could deal with them. There were 4 rows with missing values in the A/G Ratio column which were imputed with the mean of the column because it did not contain any outliers and a cited research paper was doing the same. Thirteen rows were duplicates which were dropped. After this there 164 rows in the 0 class and 406 in the 1 class. A pie chart of this data was made in Python as seen in the poster. After this the data was divided into training and test sets using a 70:30 as seen in a cited research paper. This ratio uses substantial data to train the models while also making sure that there is enough data in the test set for a robust evaluation. The test set is kept separate.

The training set is then checked for outliers where values that have a z score greater than 3 are considered as outliers and eliminated from the dataset. Although outliers can contain important insights about patterns in the data, this was done as seen in a cited research paper. The possibility of retaining outliers, given that they aren't errors using domain knowledge has been mentioned in the Future Work section. Boxplots representing the distribution of all the features for both the classes have been plotted using Python and displayed in the poster. The difference in distribution for each feature in both the classes helps to think about how each feature affects the classification process. Moreover, a correlation matrix between all the variables has also been plotted using Python and illustrated as a heatmap in the poster. This helps to see which features are highly correlated amongst each other and with the class variable.

This was followed by balancing the dataset trying both random oversampling and SMOTE (individually) however, it became apparent that balancing the dataset before training and then applying cross fold validation resulted in a data leakage as the cross fold validation loss was low i.e. the validation accuracy was high but the test accuracy was poor. This meant that some data points (or their augmentations) that were in the training subset of a model in a fold were also present in the validation subset resulting in an inflated validation accuracy and poor hyperparameter optimisation. Hence this step was skipped and it was decided to continue with the unbalanced dataset.

Feature selection was applied at this point on the training set to train the model with selected features. This step was skipped when training was done without feature selection. Features that had a correlation

coefficient greater than 0.15 with the class variable were selected and out of this set of features, any pair of features that had a correlation amongst each other greater than 0.8 were examined and the one with the higher correlation with the class variable was selected while the other one was removed. This resulted in the selection of six features: Age, DB, Alkphos, Sgpt, Sgot and A_GRatio. Their indices were saved in a .mat file so that the same features could be extracted from the test set to be used for testing.

Both models were trained separately but using the same steps in order to ensure a completely fair comparison. First hyperparameter fine-tuning was done automatically using Bayesian Optimisation with 30 function evaluations to see which values in the hyperparameter space would perform better. The hyperparameters chosen to be fine-tuned for each model have been mentioned in the poster. The values obtained from Bayesian Optimisation were then used as a starting point for further fine-tuning in a grid search. Since each model had 3 hyperparameters to be fine-tuned, the grid search was set up in such a way that for each unique set of hyperparameters, training was done using 10 fold cross validation. 10 fold uses 90% of the training set for training each fold which is essential for the model to learn patterns effectively when the dataset is small. Having more folds also allowed for a more robust comparison between hyperparameter values. In the case of random forests, bootstrapping is done internally to train different decision trees thus it serves a different purpose as compared to cross validation which is done to reduce bias in hyperparameter selection. The combined predictions made by each fold on its validation set resulted in the predictions of the entire training set. Using these predictions and the true labels, evaluation metrics such as validation accuracy, precision, recall, f1 score, AUC and cross validation loss were calculated for that specific set of hyperparameters and stored. In the next iteration, this was repeated for another set of hyperparameters where only one hyperparameter was updated while the other two remained constant. This was continued till all possible sets of hyperparameters were exhausted. At the end of the grid search, the set of hyperparameters with the lowest cross validation loss was selected for each model. Using these sets of hyperparameters, both models were retrained on the entire training set for optimisation and saved to be used for testing. During this retraining, the training times for each model were recorded.

The trained models were made to make predictions on the test set which were compared with the true labels and used to calculate evaluation metrics such as testing accuracy, precision, recall, f1 score and AUC. Confusion chart along with the true positive rate and true negative rate is plotted for each model. ROC curves of both models are also plotted on the same graph to see which model is better at differentiating between the classes.

References

1. “Liver function tests”, Mayo Clinic, https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595
2. “Bilirubin Direct (Conjugated)”, NHS choices, https://www.southtees.nhs.uk/services/pathology/tests/bilirubin-direct/
3. “Total protein test”, NHS choices, https://www.nhs.uk/conditions/total-protein-test/
4. C. C. medical professional, “Globulin blood test: What it is, procedure, results”, Cleveland Clinic, https://my.clevelandclinic.org/health/diagnostics/22365-globulin-blood-test