CLASSIFICATION OF LIVER DISEASE USING THE INDIAN LIVER PATIENT DATASET

STAT441/841 (Statistical Learning - Classification)
Final Project Submission

Authors (TEAM AAM):

Angela Wang
Asad Amiruddin
Marcus Di Renzo

University of Waterloo
Department of Statistics & Actuarial Science

Fall 2019
Dr. Peijun Sang

Contents

1 Introduction

1.1 Background Information

The purpose of this report is to apply the classification framework on the Indian Liver Patient Dataset, taken from the UCI Machine Learning Repository. This dataset is a collection of 583 records taken from a cross-sectional study conducted in the late 2000s within Andhra Pradesh in India. The response variable of interest is determining whether or not a patient has liver disease, coded as a binary variable. Liver disease is a broad medical diagnosis which encompasses various ailments, and is characterized by a minimum of 75% of liver tissue being affected. The primary function of the liver is detoxification of substances within the blood, metabolization of drugs, the creation of bile to aid in digestion, and the secretion of proteins to promote blood clotting (Hoffman, 2019). Loss of liver function can lead to various health complications, including cancer, liver failure, and death. Dependent on the particular cause of liver disease, different diagnostic tests can be used. These tests include CT scans or ultrasounds in order to assess the tissue quality of the tumour and surrounding cells, a liver biopsy (in which a portion of the liver is removed and examined to assess damage), a complete blood count test to assess the quality and composition of blood, or more specialized medical tests in the case of rare or genetic diseases (Raypole, 2019).

Common diseases of the liver include cirrhosis, hemochromatosis, nonalcoholic fatty liver disease, and hepatitis (Wedro, 2019). Liver disease is of particular interest in India, as it is the 10th leading cause of death per year, resulting in nearly 260,000 deaths annually (WHO, 2017). Though liver disease can be the result of genetics, many of the risk factors are environmental in nature as well. Individuals who heavily drink alcohol, use contaminated needles, are diabetic or obese, or those who engage in risky sexual behaviours are at increased risk of developing various hepatic disease (Mayo Clinic, 2018). These risk factors are particularly important within the context of India, as cultural norms and lifestyle choices put many individuals at higher risk. Over 135 million individuals in India were classified as obese (Ahirwar & Mondal, 2018), the prevalence of binge and underage drinking have been increasing in prevalence (Alcoholism in India, 2019), and up to 46% of hepatitis B and 38% of hepatitis C cases in India are the related to improper disposal and reuse of used needles and syringes (Vora, 2017).

Compounding the issues related to the risk factors of liver disease in India, there is also less societal focus on health promotion, autonomy and education compared to some western societies (Pati, Sharma, Zodpey, Chauhan, & Dobe, 2012). The result is that many individuals do not recognize they are at risk of developing liver disease, and consequently many cases go undiagnosed. This is particularly problematic as the best treatment for most liver diseases includes lifestyle changes, such as limiting alcohol use, increasing exercise and improving diet, reducing the use of prescription and injection-based drugs, and consistent follow-ups with healthcare professionals. If an individual is not aware of their diagnosis, they may not make the necessary changes required to maintain their health. The majority of hepatic diseases are irreversible, and if left untreated, can lead to other health complications like cirrhosis and a life expectancy of 2 to 3 years (Cunha, 2019). In extreme cases, health interventions such as liver transplants can be used in order to maintain the necessary liver function required for life. However, liver transplants are especially difficult within the Indian healthcare system as a result of privatization of healthcare. Referrals for liver transplantation are carried out through the private healthcare sector rather than the public, which can lead to inconsistent referral times, quality of care, excessive financial cost, and the lack of a centralized organ transplant database (Narasimhan, 2016). Further complicating this issue is the type of organ donors available within India; in the context of organ donation, typically cadavers (deceased individuals) are preferred as the donor in order to reduce the number of operations and organ quality. However in India, cadaver donors only make up 3% of liver transplants, meaning the remaining 97% of liver donations come from living individuals, which introduces additional strain on the healthcare system in the form of increased cost, recovery time, and differences in post-operative health (Times of India, 2017).

Clearly liver disease is a highly prevalent and serious healthcare issue within India. By using data-driven approaches and introducing machine learning framework within the healthcare system, improvement in the quality of care can be made. Classification can be used to supplement the decisions made by physicians, resulting in fewer misdiagnosed patients. Additionally, through adequate data exploration, different risk

factors can be assessed to best identify the strong risk factors of various liver diseases. This can then be used to guide medical research, help implement public health education and communication strategies, and provide more insight in the best methods of treatment and prevention. The consequences could include improved wait times related to medical care, reduced strain on healthcare professionals, and an overall improvement in the knowledge translation framework between physicians and data scientists.

1.2 Justification and Motivation of Approaches

Though the Indian Liver Patient dataset has been extensively studied within the field of machine learning, the primary focus of many previous studies using this dataset have been on the prediction accuracy of various algorithms. Though many of these algorithms have been shown to be successful classifiers in the context of prediction, less attention has been made to the interpretability and communication of these results within healthcare. Consideration must be made to the audience in which this dataset relies on within the context of any data science problem. Physicians and other healthcare professionals have specialized knowledge related to the diagnosis and treatment of liver disease, but often have limited knowledge or training in statistics. This can result in gaps in communication if the suggested models being used are non-intuitive or lack a suitable interpretation. When considering different choices of models, we must ensure that the underlying framework of assumptions, results, and fitting are all capable of being explained to someone who does not have extensive training in statistics.

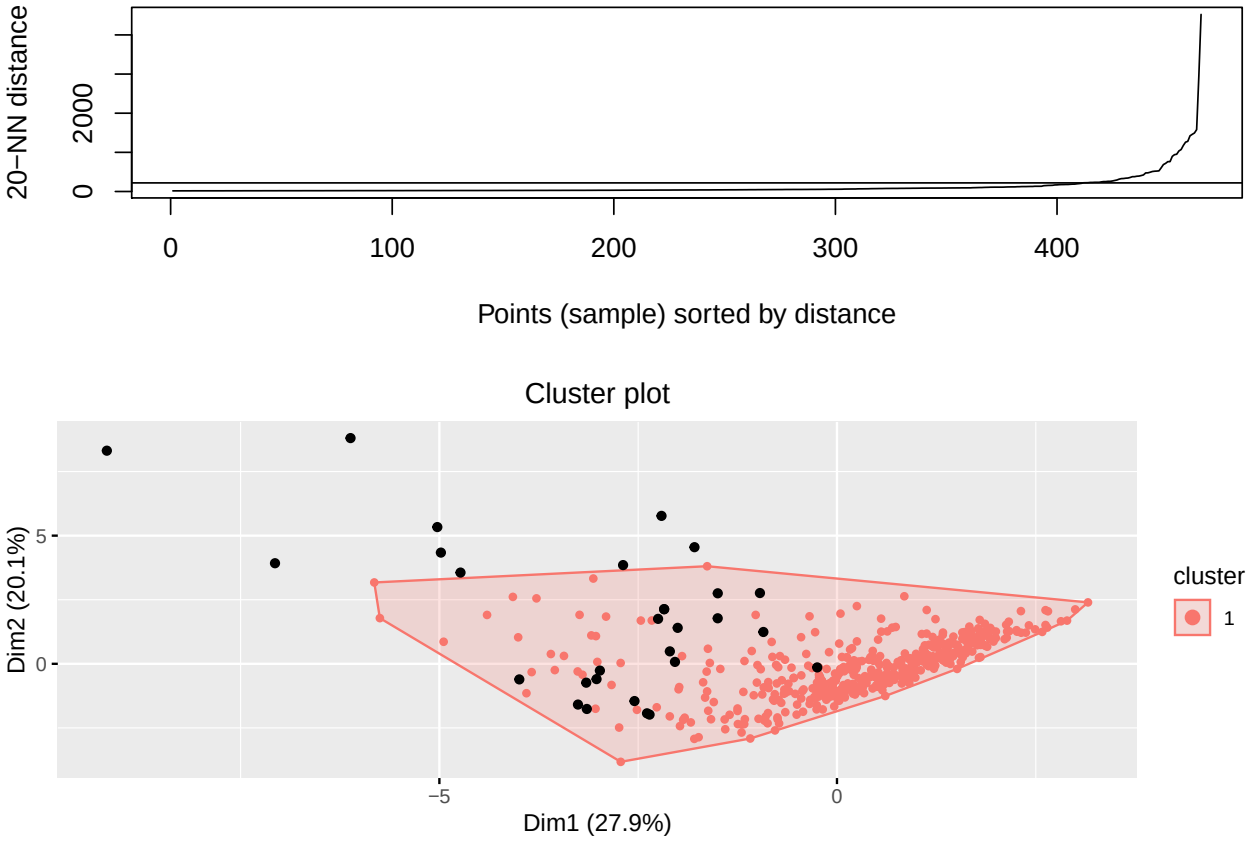
In the context of this report, our focus is on models algorithms that are intepretable and can therefore be easily communicated to professionals within the healthcare setting. Consequently, we have chosen to examine 6 different approaches for determining whether or not a patient will have liver disease - Logistic Regression, K-Nearest Neighbours, Discriminant Analysis, Random Forest, Naive Bayes, and Extreme Gradient Boosted Decision Tree ensembles. Though we are still concerned with the predictive accuracy of various approaches, it is equally important to ensure that the underlying method could be justified and explained to our intended audience. Though not all of these models are directly interpretable, the idea of our report is to compare simple interpretable models to more complex statistical approaches, and determine if an optimal trade-off can be determined between interpretability and predictive accuracy.

2 Data Pre-Processing

2.1 Outlier Assessment

To detect outliers, we use Density-Based Spatial Clustering of Applications with Noise (DBSCAN). This density-based clustering algorithm clusters data based on regions with high density of data points and categorizes data points based on their location with respect to other points (Lutins, 2017). Clusters are formed as an n-dimensional shape about a point and only qualify as a cluster when the minimum number of observations lie within said cluster region. Core points lie within ϵ distance of another point, border points lie within a cluster region but are not within ϵ distance of another point, and noise points neither lie within a cluster region nor within ϵ distance of another point. Noise points are identified as outliers.

There are two parameters to set up DBSCAN: ϵ and the minimum number of points in a cluster. Sander et al. suggest that a heuristic approach to selecting the minimum number of points in a cluster is setting it as twice the dimension of the data (Sander, Ester, Kriegel & Xu, 1998). From there, use a plot of the k-nearest-neighbor distances, computed for each point plotted against the sorted distances to identify where the curve begins to drastically rise. The k-nearest-neighbor distance at which the plot begins to ascend rapidly is a suggested value for ϵ.



Black points indicate outliers. Based on the DBSCAN plot, there are 28 outliers in the training set. Since excluding outliers could lead to biased results and inflate predictive accuracy, we did not remove any of the outliers. More subject matter expertise is required to decide how to treat and identify the outliers. In the Data Exploration section, certain predictors are log transformed so that extreme values have a lesser impact on our assessment.

2.2 Variable Standardization

Variable standardization was performed for the K-nearest neighbours and linear discriminant analysis methods. Training observations were centered and scaled before fitting the K-nearest neighbours classifier so that the relative distance of observations along different directions are accounted for when considering the nearest neighbouring points. For ease of model interpretation, training observations were scaled before fitting the linear discriminant analysis classifier.

To aid data exploration, the variables DB, TB, ALP, ALT, AST and TP were log transformed so that any potential symmetry in the distribution of these variables were more readily recognizable.

2.3 Missing Data

There were several approaches that could be undertaken in order to deal with the missing data such as case deletion, mean substitution and regression imputation (Kang, 2013). For data that are missing at random and data that are missing not at random, mean substitution may potentially lead to inconsistent bias and underestimates standard error as it alters the distribution of the data by concentrating more points at the mean of the sample containing no missing data (Kang, 2013). A better approach for remedying missing data is multiple imputation, which involves "creating several different plausible imputed data sets and

appropriately combining results obtained from each of them" such that "the uncertainty about the missing data" is accounted for (Sterne et al., 2009).

The data contained four observations out of the 583 observations which had missing values solely in the RAG variable. Since a low proportion of the data were missing, the bias introduced by mean substitution would be minimal, so the missing data were ultimately treated by mean substitution.

2.4 Training and Test Sets

We used a 80/20 split for the training and test set split with a validation set within the training set. The training set contains 465 observations while the test set contains 118 observations.

3 Data Exploration

3.1 Variable Summary

The dataset contains a total of 10 predictor variables, each explained in the subsequent sections below. The 11th variable in the dataset refers to the diagnosis status of the patient (=1 if the patient has liver disease diagnosis, and 0 otherwise). **Note that the entirety of this data visualization and exploration occurs on the training set only**. This is done intentionally to ensure the data within the test set is totally hidden, and does not 'contaminate' our knowledge of the problem at hand.

In the context of this report, 'cases' refer to individuals with a liver disease and 'controls' refer to those with no disease. In the training set, a total of 142 individuals did not have liver disease and 323 have a reported disease. The total sample size is 465 observations. For data with a highly skewed distribution, the natural logarithmic transform was applied to spread out the measured values. These variables are denoted with the "log" prefix. Most of the measurements are reported in terms of g/dL (grams per deciliter) or IU/L (international units per liter), unless otherwise stated.

Developing data visualizations provides a more thorough understanding of the underlying data, particularly in terms of the relationship between each predictor and the response variable of interest. By comparing the densities of each predictor between the two diagnosis groups, different trends can be assessed which can provide intuition for variable importance and effects during model fitting. Of primary interest is identifying covariates with considerably different densities between the two groups, as this implies that covariate has a strong discrimination effect and may therefore be useful in classifying patients.

3.1.1 Age

Age refers to the age of the patient at the time of data collection. In the methodology of the study, all individuals aged 90 years and older were assigned an age label of 90.

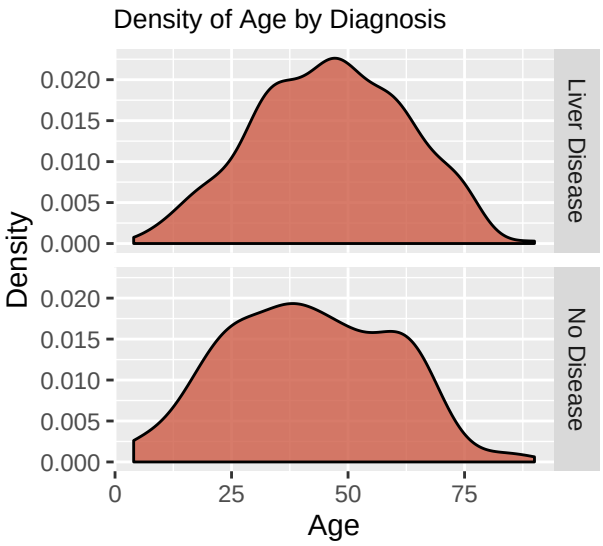


Table 1: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value
Minimum	4.00
1st Quartile	28.00
Median	42.00
Mean	41.62
3rd Quartile	55.75
Maximum	85.00

Summary Statistic	Value
Minimum	7.00
1st Quartile	34.00
Median	46.00
Mean	46.19
3rd Quartile	60.00
Maximum	90.00

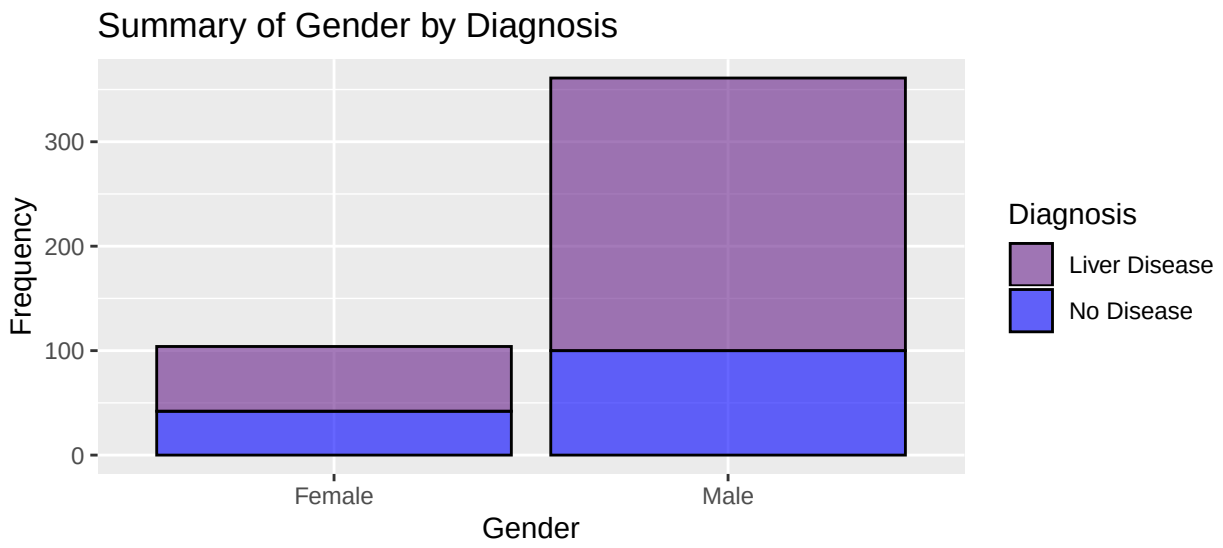
The majority of the study participants were between 25 to 60 years of age, with the maximum age being 90 years and the minimum being 4 years. The overall distribution is relatively symmetric and appears normally distributed. In general, the cases appear to have a slighter higher age than the controls within the training set, with the average age being 41.62 years in the controls VS 46.19 years in the cases. Research suggests that in general, the risk of liver disease increases with age (Kim, Kisseleva, & Brenner, 2015), so these results are relatively expected. Given the ages densities are not drastically different between groups, age alone may not be an important discriminator for the disease presence. However, given age is potentially related to many other predictors of interest, which may result in interaction, age is further explored in the subsequent section of the report.

3.1.2 Gender

Gender refers to the gender of the patient, classified as either male or female.

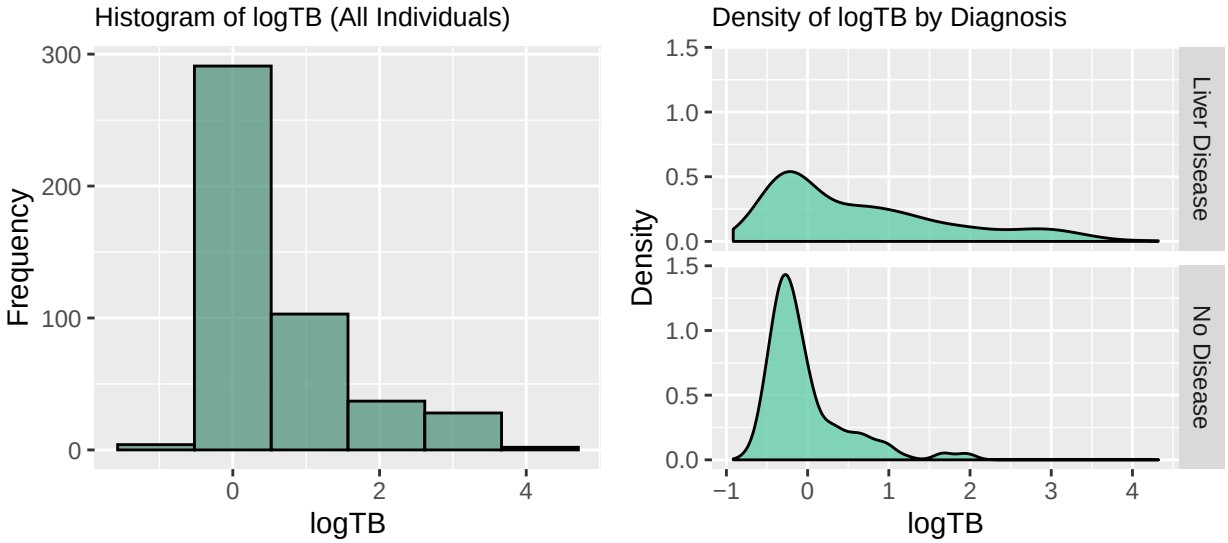
Table 2: Contingency Table of Gender and Diagnosis

	Liver Disease	No Disease
Female	62	42
Male	261	100



The majority of study participants were male (361/465) in the training set, resulting in a majority in both the case and control group. Of the 104 females, we see that 62 individuals had liver disease (60%) whereas 72% of the males in the training data had liver disease. This may suggest an association between gender and liver disease, particularly that males are more susceptible than females. Many of the associated risk factors for liver diseases, such as obesity, alcohol intake, drug use and biological differences in protein levels are dependent on sex, which may partially explain this relationship (Guy & Peters, 2013). Interactions involving sex and other covariates are explored later in the report, prior to model fitting.

3.1.3 Total Bilirubin (TB)



Bilirubin is a chemical which occurs as a result of red blood cells breaking down, measured through clinical

Table 3: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value	Summary Statistic	Value
Minimum	-0.69	Minimum	-0.92
1st Quartile	-0.36	1st Quartile	-0.22
Median	-0.22	Median	0.34
Mean	-0.04	Mean	0.66
3rd Quartile	0.10	3rd Quartile	1.28
Maximum	1.99	Maximum	4.32

tests in milligrams per deciliter (mg/dL). Total bilirubin is assessed by summing together the indirect and direct bilirubin measurements on a patient. Often bilirubin levels are used to assess liver function, since the liver takes in bilirubin and breaks it down for excretion (Mayo Clinic, 2018). Normal measurements for adults are approximately 1.2 mg/dL (0.18 on the log scale), which is where the mode of the data is contained roughly for cases and controls.

In general, total bilirubin levels which deviate greatly from typical measurements have been found to be associated with liver issues (Mayo Clinic, 2019), particularly those which are above average. Based on training data, we can see that total bilirubin levels appear to higher in individuals with reported liver disease. Evidently the cases have a longer right tail in the distribution, which is seen by the relatively large difference in the group means. In general, nearly all of the patients with very high logTB levels seem to be diseased, which means within this range of the variable, classification should be more deterministic.

3.1.4 Direct Bilirubin (DB)

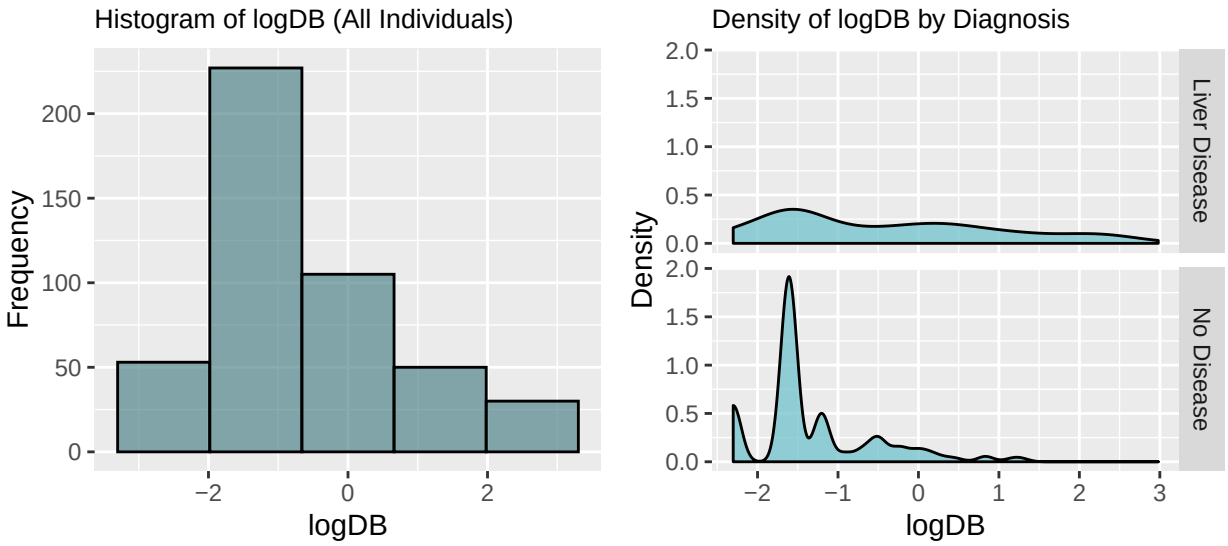


Table 4: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value	Summary Statistic	Value
Minimum	-2.30	Minimum	-2.30
1st Quartile	-1.61	1st Quartile	-1.61
Median	-1.61	Median	-0.69
Mean	-1.32	Mean	-0.40
3rd Quartile	-1.20	3rd Quartile	0.56
Maximum	1.28	Maximum	2.98

Direct bilirubin (measured in mg/dL) is similar to total bilirubin, but excludes the amount of indirect bilirubin in the blood. The results are similar to what was observed for total bilirubin levels, in that cases appear to have higher values than the controls. Again, the difference in distribution between the two groups may suggest that bilirubin is a useful predictor in determining the presence of liver disease.

3.1.5 Alkaline Phosphotase (ALP)

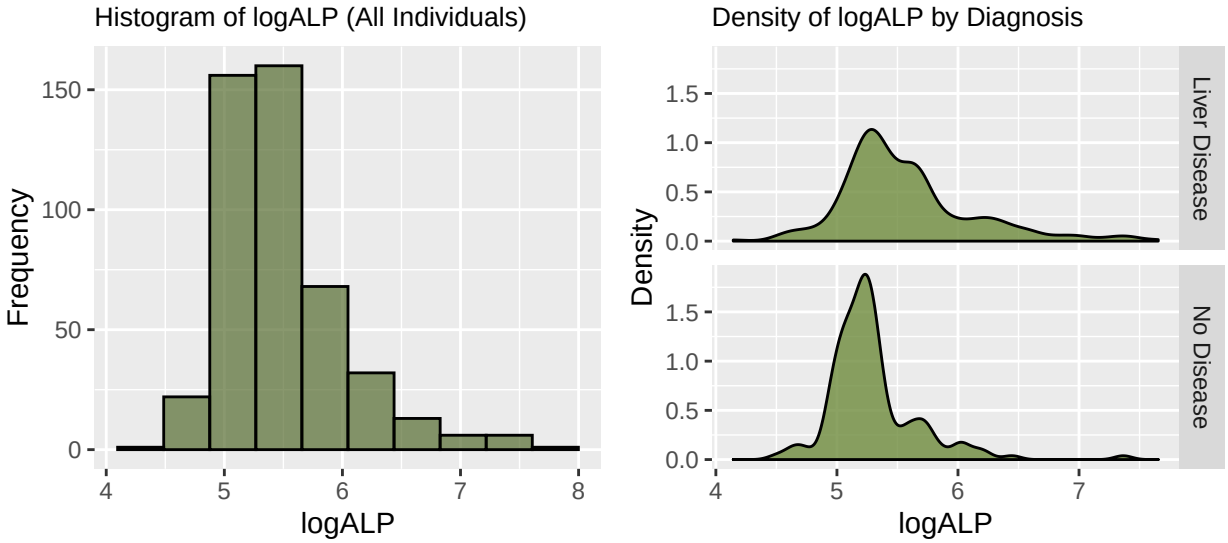


Table 5: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value
Minimum	4.50
1st Quartile	5.08
Median	5.23
Mean	5.29
3rd Quartile	5.37
Maximum	7.37

Summary Statistic	Value
Minimum	4.14
1st Quartile	5.24
Median	5.44
Mean	5.57
3rd Quartile	5.74
Maximum	7.65

Alkaline phosphotase (ALP measured in IU/L) is an enzyme found within bone cells and the liver, and is also a common biomarker used to assess liver function, measured via a blood test. Typically higher levels of ALP are associated with liver dysfunction, as damaged or diseased liver cells often secrete increased amounts of ALP into the bloodstream (Lab Tests Online, 2019). Similar to other measurements, ALP is measured in mg/dL.

In the context of the training data, we see that ALP levels tend to appear somewhat similar between the cases and controls, though there is a higher degree of spread in the cases and a longer right tail. In general, the spread is relatively large in the combined data set which is unlike some of the previous measurements. Despite the differences in distribution between ALP levels in the cases and the controls, caution should be made from these results alone, as ALP can also be a marker of other diseases. Since many potential confounders were not collected in this study, the results we see related to ALP should be interpreted with caution as we do not know the diagnosis of other diseases within this cohort of patients.

3.1.6 Alanine Aminotransferase (ALT)

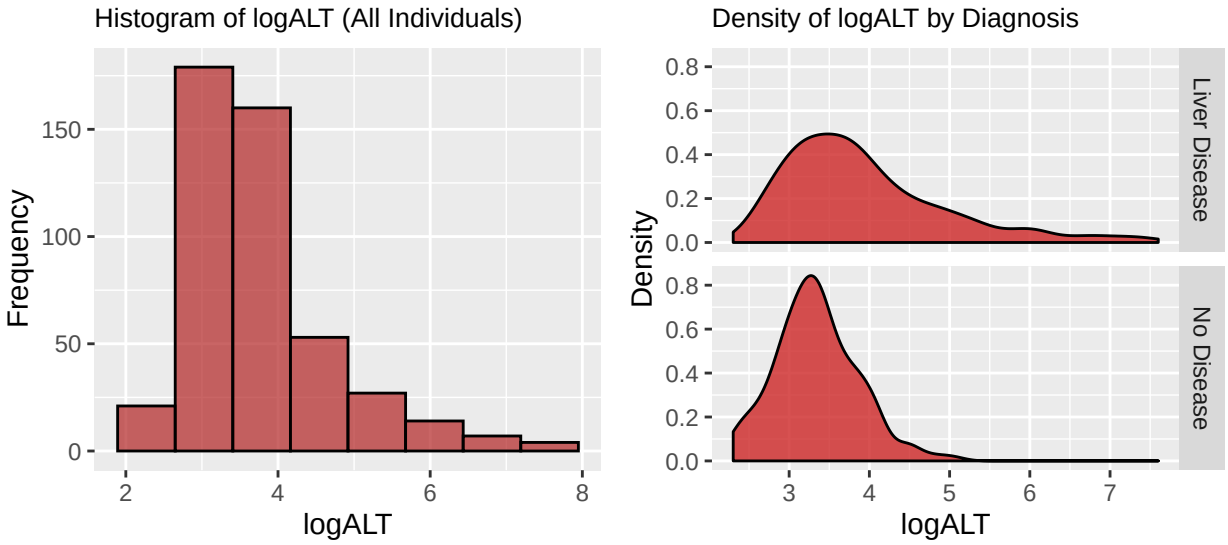


Table 6: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value
Minimum	2.30
1st Quartile	3.01
Median	3.31
Mean	3.34
3rd Quartile	3.64
Maximum	5.02

Summary Statistic	Value
Minimum	2.48
1st Quartile	3.22
Median	3.71
Mean	3.95
3rd Quartile	4.38
Maximum	7.60

Similar to ALP, alanine aminotransferase is an enzyme found in liver and protein cells. ALT is an enzyme responsible for adequate liver function, and damaged cells will leak this enzyme into the bloodstream. The result is that individuals with liver disease tend to have increased ALT levels relative to individuals with healthy livers (Medline Plus, 2019).

In the context of the training data, this trend is also apparent. Individuals with liver disease tend to have higher levels of ALT relative to controls, and the spread in observed ALT levels tends to be much wider in the cases. Normal ALT levels range from 19 to 33 IU/L, which is roughly where we see the mode in each of the datasets (roughly 3 to 3.5 on the log-scale), though ALT levels often vary by sex and age (Blocka, 2018).

3.1.7 Aspartate Aminotransferase (AST)

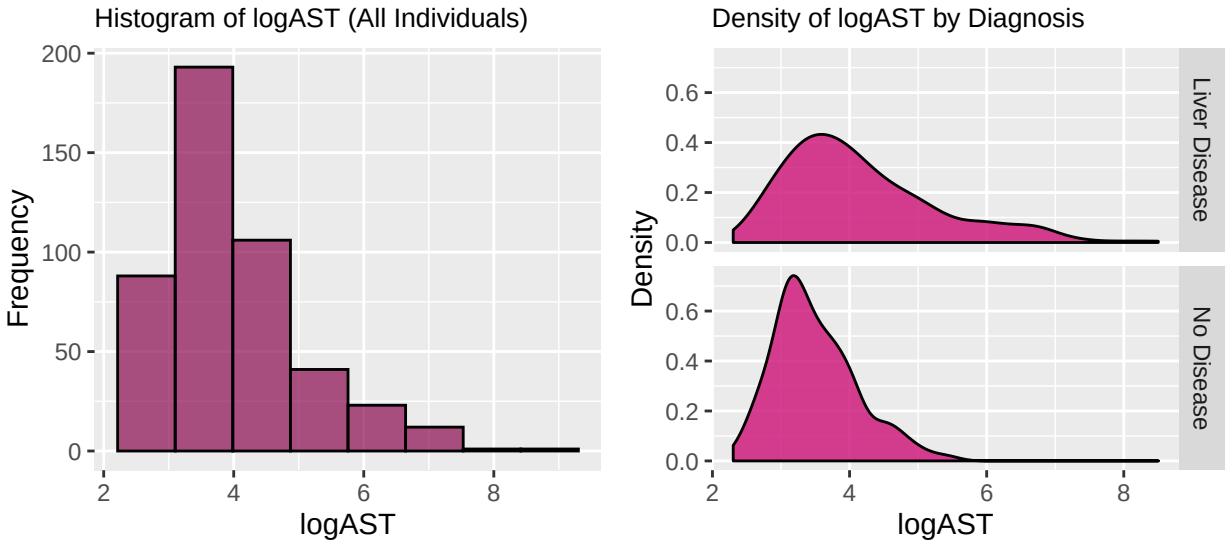


Table 7: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value
Minimum	2.30
1st Quartile	3.09
Median	3.37
Mean	3.49
3rd Quartile	3.82
Maximum	5.44

Summary Statistic	Value
Minimum	2.40
1st Quartile	3.37
Median	3.91
Mean	4.17
3rd Quartile	4.73
Maximum	8.50

Aspartate Aminotransferase is measured in IU/L and is an enzyme found in the blood and liver cells, similar to the previous proteins. Like AST and ALP, increased levels of AST result from damaged liver cells, and so individuals with liver disease often have higher blood levels of AST relative to those with a healthy liver (WebMD, 2019).

This trend is evident in the training set, as we see considerably higher average levels of AST in cases rather than controls. The evident difference in distribution among the different groups also gives insight that AST is an important predictor of liver health, and therefore may be an important predictor in the classification problem of interest.

3.1.8 Total Protiens (TP)

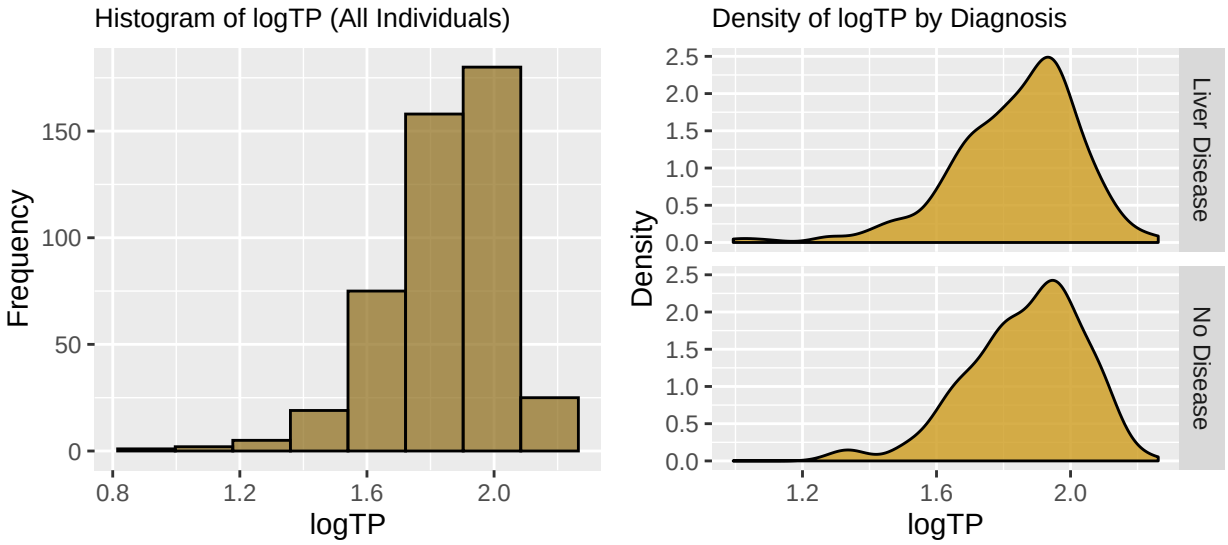


Table 8: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value
Minimum	1.31
1st Quartile	1.77
Median	1.90
Mean	1.87
3rd Quartile	1.99
Maximum	2.22

Summary Statistic	Value
Minimum	0.99
1st Quartile	1.72
Median	1.87
Mean	1.84
3rd Quartile	1.96
Maximum	2.26

Total protein refers to the combined totals of albumin and globulin in the bloodstream, measured in g/dL. Normal levels fluctuate between 6 to 8 grams per deciliter (roughly 1.8 to 2.1 on the log-scale), which is where the mode of the training set is observed for each group. Unlike most other blood proteins, liver diseases may cause elevated or decreased total protein levels dependent on the underlying diagnosis (Health Line, 2016).

Within the data, we see that the cases and control have extremely similar distributions of total protein levels. This similarity in distribution may suggest that total protein may not be a strong predictor of liver disease, as the values seem similar even after conditioning on diagnosis. Furthermore, the observed range of values is very small which may make class separation difficult to accomplish.

3.1.9 Albumin (ALB)

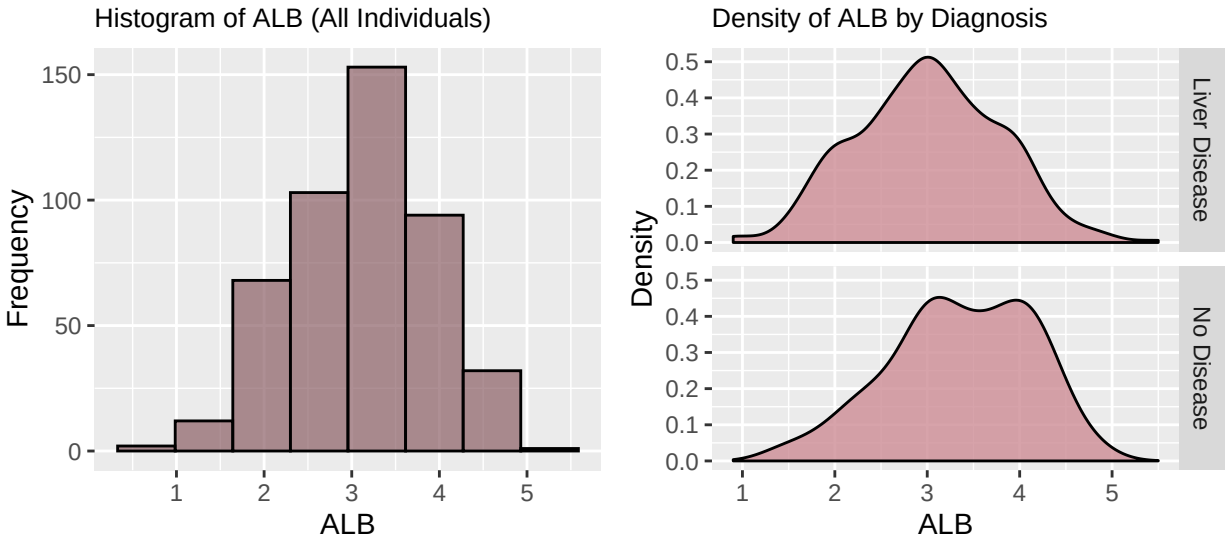


Table 9: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value
Minimum	1.40
1st Quartile	2.90
Median	3.40
Mean	3.36
3rd Quartile	4.00
Maximum	4.90

Summary Statistic	Value
Minimum	0.90
1st Quartile	2.50
Median	3.00
Mean	3.01
3rd Quartile	3.55
Maximum	5.50

Albumin is a protein produced by the liver and secreted into the bloodstream in order to retain fluid and transport substances like vitamins or enzymes. Albumin levels tend to be lower in individuals with liver diseases as a result of decreased production in liver cells, however damage to the liver must be fairly severe before a considerable drop in albumin levels are observed (Lab Tests Online, 2019).

The levels of ALB tend to be somewhat lower for the cases relative to the controls, though this difference is not relatively large. This may suggest that of the individuals in the study who had observed liver disease, the severity of their prognosis may have been relatively mild since ALB was still being adequately produced by the liver. As a result, ALB alone may not be a strong indicator of liver disease aside from the cases in which their liver disease has progressed significantly over time.

3.1.10 Ratio of Albumin to Globulin (RAG)

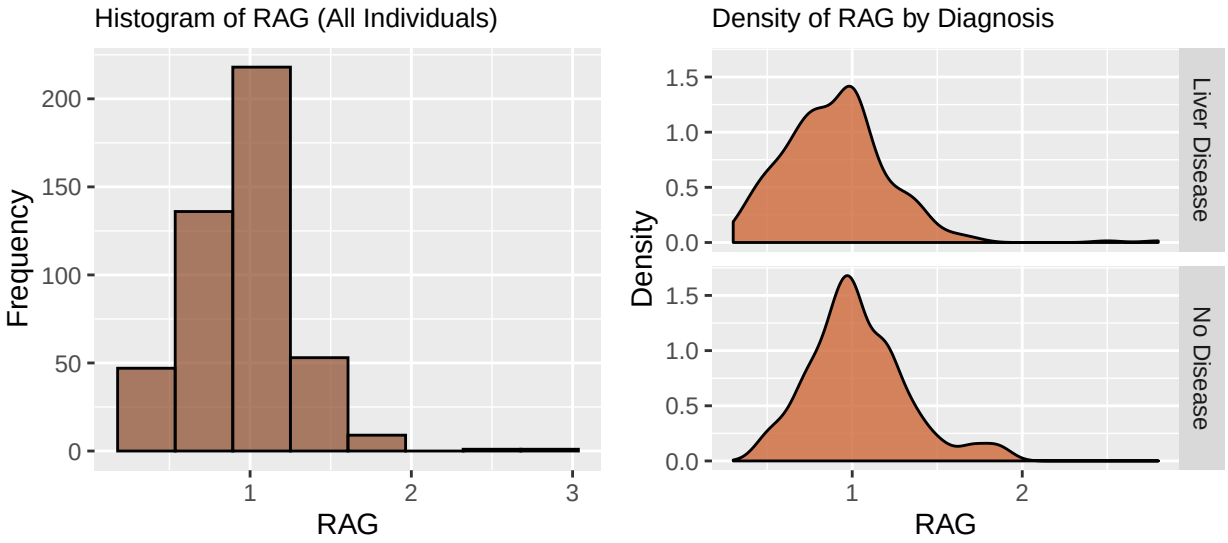


Table 10: Summary for Non-Diseased Patients (left) VS Diseased (right)

Summary Statistic	Value
Minimum	0.45
1st Quartile	0.90
Median	1.00
Mean	1.03
3rd Quartile	1.20
Maximum	1.90

Summary Statistic	Value
Minimum	0.3
1st Quartile	0.7
Median	0.9
Mean	0.9
3rd Quartile	1.1
Maximum	2.8

The ratio of albumin to globulin is similar to the measurement of total protein, but instead reflects the ratio of albumin to globulin levels observed in the bloodstream. The expected ratio of albumin:globulin is expected to be near 1, and values considerably below 1 indicate poor liver function (Health Line, 2019).

Looking to the training data, the results follow what is expected. We see that the cases tend to have a lower ratio than the controls (though the magnitude of difference is small), and the modes fall within the typical expected range. The dissimilarity in RAG values implies it will be useful in the context of prediction, albeit less so than some of the previously examined covariates. In the cases, we note some outlying points on the right hand side in which individuals have considerably higher RAG values, which may be the result of a combination of medical conditions or measurement error.

4 Analysis

4.1 Methodology

Multiple models of each class will be fit in order to consider different customizable options, such as hyperparameter tuning, the effect of PCA, influence of variable selection, etc. Of these fitted models within each class, the optimal choice will be chosen and justified. Each of the chosen models will then be compared in the final section of the report to select the final classifier of interest. The models considered in this report are discriminant analysis, Naive Bayes, K-Nearest Neighbours, XGBoost and Random Forest. Logistic regression was also considered but performed very poorly (test accuracy of only 0.2) and was therefore put into the appendix instead.

When fitting each of the models, we use repeated 5-fold cross-validation on the training set. This gives 25 out-of-sample prediction accuracy estimates, which can then be examined to determine the volatility and generalizability potential of the model. The goal is to find the optimal trade-off between interpretability, parsimony and interpretability for a clinician. The metrics used to assess each of the models are the out-of-sample prediction accuracy (which we want to maximize), and the the associated Cohen's κ values. κ is important in this context since our data is highly skewed; 70% of the training set observations are diseased. Cohen's Kappa value accounts for this by modeling both the observed and expected accuracy (originally refered to as *agreement* between responses), and therefore accounting for accuracy attributed to random chance. This means in the simplest terms, κ measures how much better a classifier is relative to one that randomly guesses the labels. The values themselves are not directly interpretable, but Cohen suggested values ≤ 0 as indicating no agreement and 0.01 - 0.20 as none to slight, 0.21 - 0.40 as fair, 0.41 - 0.60 as moderate, 0.61 - 0.80 as substantial, and 0.81 - 1.00 as almost perfect agreement (McHugh, 2012).

Variable selection will be dependent on the chosen method, but will usually be implemented via recursive feature elimination (RFE) with Random Forest via the caret() package. In this approach, a model is fit to the full training set and its performance is assessed, along with variable importance. Then for specified subset size S of the predictors, the S most important variables are used to fit a nested model, in which the predictive performance is also assessed. By doing this for all values of S in 1 to p (the number of predictors), the optimal value of S can be chosen. The final selected variables are the S most important variables based on the initial fit, and then the model gets re-fit with only these S important predictors (Kuhn, 2009). Other models, such as logistic regression, use a forward stepwise approach to conduct variable selection.

4.2 K-Nearest Neighbours (KNN)

4.2.1 About K-Nearest Neighbor (KNN)

KNN is a non-parametric approach that aims to improve on other parametric classifiers by eliminating the dependence on parametric assumptions. For a given observation, this method examines the class labels for the k-nearest observations and assigns a label based on the majority vote among these k neighbors. In this context, it implies that the diagnosis label for an individual will depend upon the diagnosis label for other individuals who have similar characteristics and medical measurements as them. Therefore, the decision boundaries created by this method are irregular and jagged.

4.2.2 Model Fitting

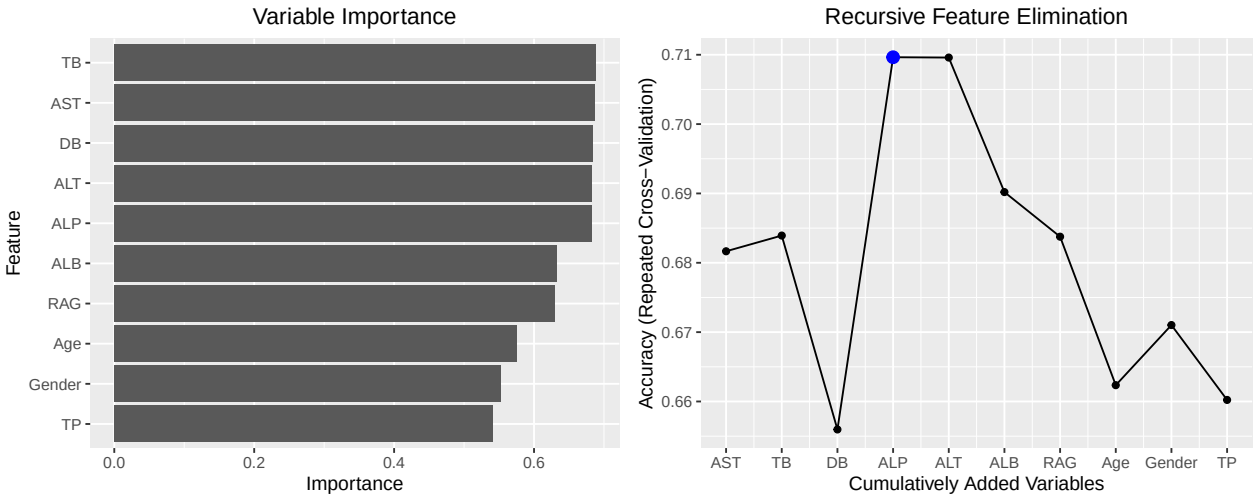
Three different variations of the KNN model will be considered in this report, and their performance based on repeated 5-fold cross validation on the training set will be assessed. These models include KNN with all the original predictor variables, KNN with selected variables from variable selection, and KNN using principal component analysis on the predictor variables. All three models will be tuned for the parameter of interest k which specifies the number of nearest neighbors to consider when classifying each observation. When k is high, the model fit will have a low variance and high bias with smoother boundaries, and when k is low, the model fit will have a high variance and low bias with very rough boundaries. Hence, the parameter k will be tuned with respect to accuracy for each of the three individual models. The values for out-of-sample prediction errors and kappa for each of these three models will be considered then to select the final KNN model. Log transformations on predictors resulted in worse performance, and thus were not used.

4.2.3 Model Assumptions

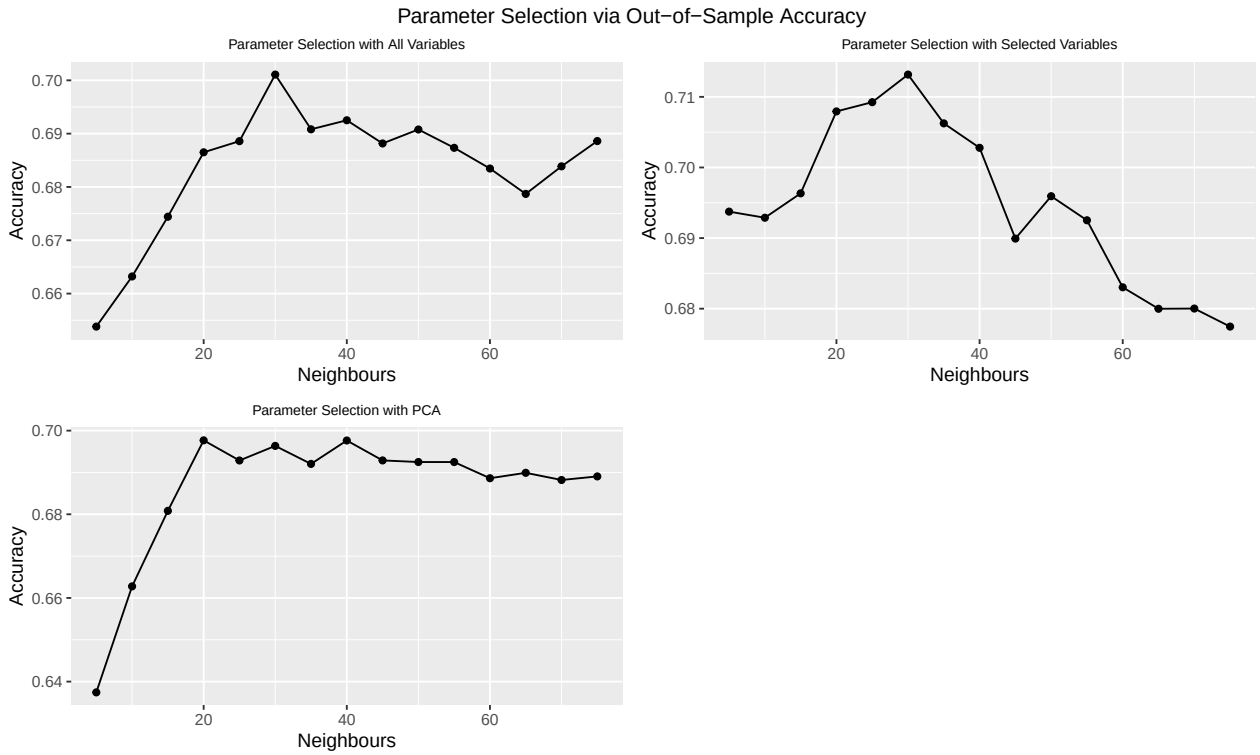
KNN assumes that the entire neighbourhood is uniform with respect to the probability distribution for the classes. Therefore, the model is subject to the curse of dimensionality. This becomes a problem when the number of covariates (p) is high, as the points are more and more distant from each other. Removing irrelevant features could potentially solve the problem therefore variable selection will be performed for one of

the variations of KNN, to only retain the most important and relevant predictor variables. Another potential problem that can arise from this assumption would be due to the multicollinearity between the predictor variables in our data. This means that some variables might contain extra weight on the prediction than we desire. As a potential solution, KNN will be performed after using principal component analysis on the data for one of the variations. Additionally, all the predictor variables will be first standardized to ensure that distance has same meaning in all directions.

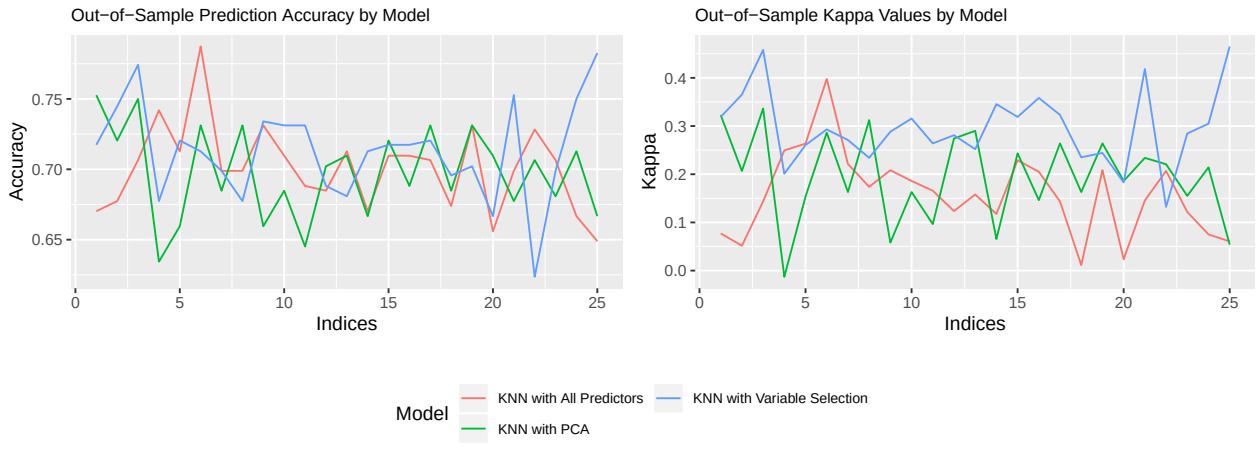
4.2.4 Results



It can be seen in the left plot that the variables contain different levels of importance for the KNN model. This indicates that the model can be potentially improved by making it more parsimonious with only the most important and relevant predictor variables getting considered for the model fit. The plot on the right is an output from conducting a recursive feature elimination on a KNN model using 5-fold cross validation. This plot shows that the accuracy of the model increases initially as more variables are cumulatively added to the model, however, it starts to reduce eventually once more and more predictors are added to perform the model fit. The potential reason for this pattern could be the curse of dimensionality. We can achieve the best model in terms of accuracy by including the variables "AST", "TB", "DB", and "ALP".



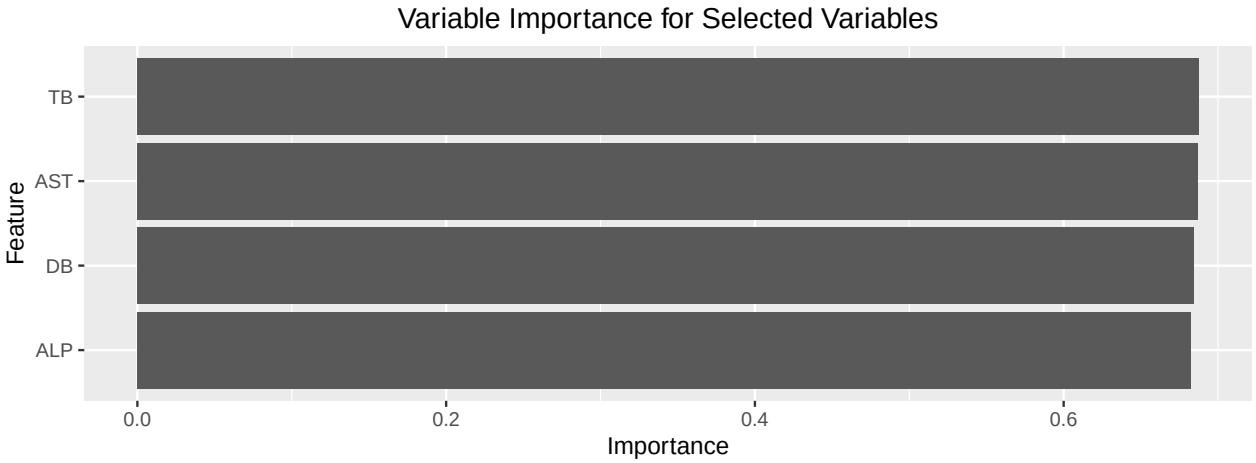
In tuning each of the models for the parameter k, we see that the accuracy initially increases as we increase the value of k and starts to decrease eventually when the value of k becomes too high. This is expected since the model becomes highly biased with a large k. KNN with all variables is more sensitive to low values of k compared to KNN with selected variables because for KNN with all variables the points are more distant from each other, so a smaller k might cause the proportion of points considered from the opposite class to be higher than a larger k. KNN with PCA is relatively robust to higher values of k. The predictive accuracy for KNN with selected variables seems to be relatively higher compared to the other two models under the selection of optimal tuning parameter k.



Model	Mean Out of Sample Accuracy	Mean Kappa Value
KNN with All Predictors	0.701	0.159
KNN with Variable Selection	0.713	0.297
KNN with PCA	0.698	0.194

The graph shows the out-of-sample predictive accuracies for all 3 models to be very similar to each other. KNN with PCA has the lowest predictive accuracy, and KNN with variable selection has the highest predictive accuracy out of all three models. In terms of the Kappa values, KNN with variables selection does a significantly better job than the other two models. Overall, KNN with variable selection has the highest predictive accuracy and highest kappa values. This shows that variable selection proved to be useful in terms of dealing with curse of dimensionality which potentially caused the results of the other two models to be worse.

Therefore, we choose KNN with variable selection as our final model for KNN and it contains the variables "AST", "TB", "DB", and "ALP", of which we plot the importance below. The optimal tuning parameter selected for the model was $k = 25$.



All 4 variables used for this model have high and equal importance. The other variables dropped had lower importance compared to these variables, hence the subset of variables selected seems to be very relevant and important.

4.3 Discriminant Analysis (DA)

4.3.1 About Discriminant Analysis

Discriminant analysis is a parametric statistical learning method that uses a Bayesian approach to calculate the posterior probability of each class $P(Y = k|\mathbf{X} = \mathbf{x})$. It assumes that we have a priori knowledge of the distribution of classes and the functional form for the conditional distribution of the predictors within each class, then applies Bayes' rule to calculate the posterior probability of each class. As such, discriminant analysis is a generative classifier.

Discriminant analysis is a method frequently used for classification involving three or more classes. It is of particular interest here as it is a relatively simple model, with low computational complexity and the ability to be highly customized. LDA can be combined with methods such as PCA, regularization, non-parametric smoothing or feature selection in order to improve performance.

4.3.2 Model Fitting

Three variants of discriminant analysis were considered and were assessed based on their performance on repeated 5-fold cross-validation on the training set. We considered linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and regularized discriminant analysis. These methods were also considered with variable selection, but the performance of these models were significantly worse and thus

have been omitted from the report. Similarly, log transformations resulted in worse performance and thus the original predictors were used.

4.3.3 Model Assumptions

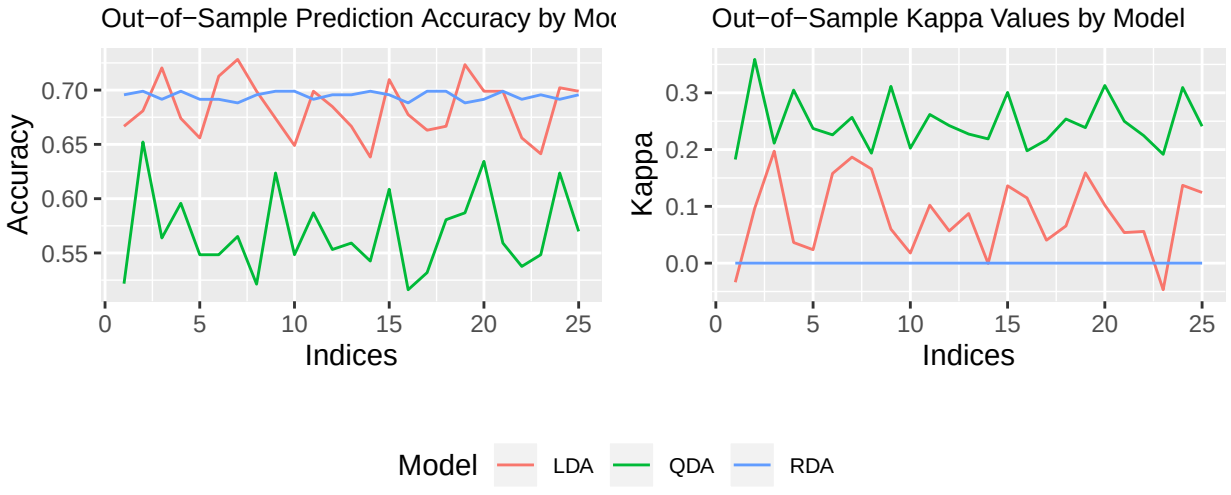
In general, discriminant analysis methods assume that each class probability densities of $X|Y = k$ are multivariate Gaussian distributed (Hastie, Tibshirani, Friedman, 2009). LDA further assumes homoscedasticity of all the class probabilities. Since all class probability densities share a common variance-covariance matrix, this implies that a given variable X_j will have the same variance in all the classes and the correlations between two given variables X_{j1} and X_{j2} are the same in all classes. Unlike LDA, QDA does not assume that the variance-covariance matrix of all the class densities are the same. As a result, QDA is a more flexible method than LDA.

Regularized discriminant analysis can be viewed as a compromise between LDA and QDA in that it assumes the variance-covariance matrix of each class probability density is a linear combination of the common pooled variance-covariance matrix used in LDA and of the class variance-covariance matrix used in QDA.

We use the Box's M-test, which is a multivariate statistical test to check the equality of multiple variance-covariance matrices, to assess whether the variance-covariance matrices of the group with the positive liver disease diagnosis and the group with the negative liver disease diagnosis are equal (Warner, 2013). The null hypothesis H_0 : covariance matrices of the predictors are equal across all groups based on the diagnosis label.

Executing the Box's M-test gives a Chi-square statistic value of 2288.68 with 55 degrees of freedom, which corresponds to an extremely small p-value. Hence, based on Box's M-test for homogeneity of covariance matrices, there is strong evidence against the null hypothesis that the class probabilities share a common variance-covariance matrix. This is expected to influence the results.

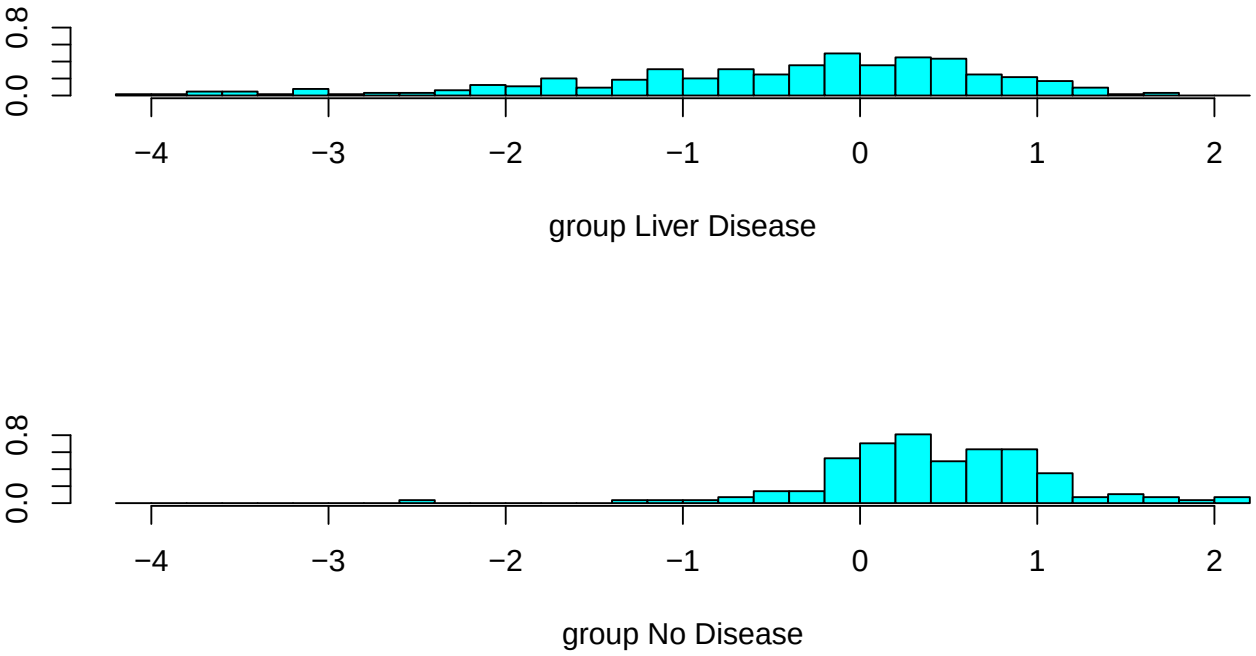
4.3.4 Results



Model	Mean Out of Sample Accuracy	Mean Kappa Value
LDA	0.683	0.084
QDA	0.569	0.247
RDA	0.695	0.000

Based on the mean out-of-sample accuracy, LDA and RDA perform much better than QDA in spite of Box's M-test suggesting that variance-covariance matrix for the liver disease diagnosis class and the no liver disease diagnosis class were different. Since the accuracy of the LDA classifier and the RDA classifier are similar, we compare the two classifiers based on the mean Kappa value. The LDA classifier has a higher Kappa value than the RDA classifier, so we ultimately choose the LDA classifier.

We can graph the histogram of the first canonical variable, which is the only discriminant function, to see how well defined the separation of the two classes are.



Based on the histograms, we can see that the values of the first canonical variable overlap for both class labels at values from approximately -1.5 to 2. This suggests that the LDA classifier would have difficulty distinguishing what the diagnosis of a given observation would be if the value of the first canonical variable lies in the range (-1.5,2). However, if the value of the first canonical variable is lower than -1.5, the LDA classifier will be able to correctly identify observations as receiving a positive liver disease diagnosis.

4.4 Random Forest

4.4.1 About Random Forest

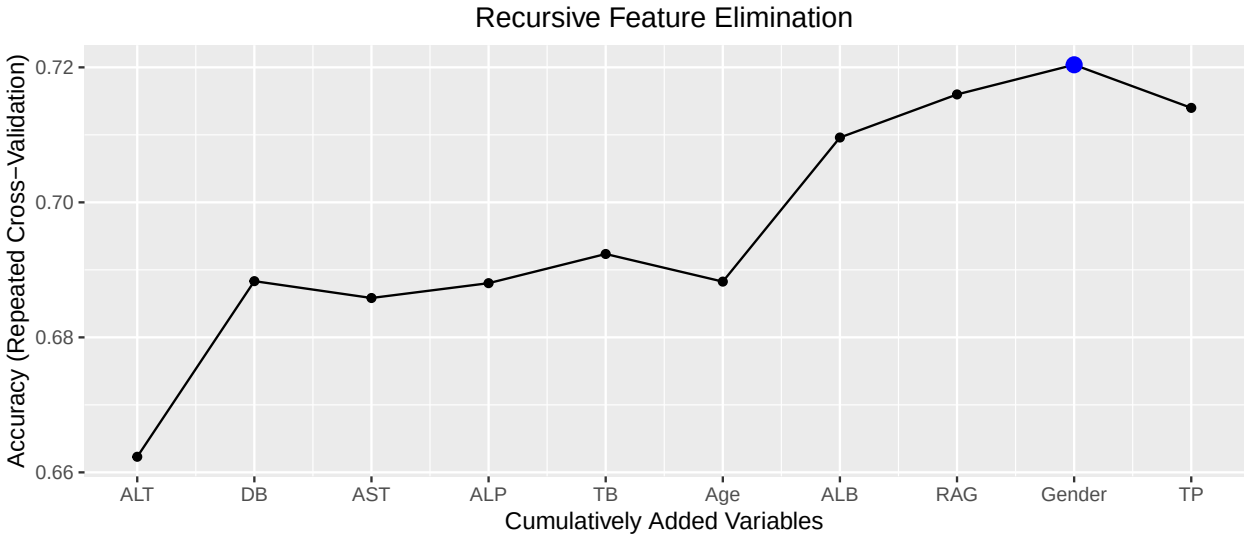
Random Forest is a tree-based classifier. The idea is based on the fact that each individual tree would have a low bias, but they are highly variable therefore we use an ensemble approach to reduce the high variability. Moreover, it aims to improve the approach of Bagging (Boostrap Aggregation) of trees by reducing the correlation among weak leaners. It randomly selects a subsample of predictor variables to reduce the overlap between resamples. This allows all the different variables to contribute towards the creation of splits within the tree. This method can become very complex, specially as the number of trees used increases. However, trees are relatively interpretable for individuals with non-technical backgrounds and they are therefore a useful approach to consider. As with KNN and DA, a log transform on predictors was considered but made the performance worse, and thus was not used in our final model consideration.

4.4.2 Model Fitting

Prior to fitting the random forest model, the effect of cumulatively adding variables to the model will be examined, using 5-fold cross validation to perform variable selection. The Random Forest model is not as sensitive to irrelevant features as some other models, however retaining only the important variables will help with reducing the complexity of the model. There are two parameters that will be tuned to select the best model. These two parameters are the number of variables randomly selected for each iteration (mtry), and the number of trees used to build the model (ntree). As we increase the number of trees, the out-of-bag error will reduce up to a point until it becomes stable because a higher number of trees enables the detection of all patterns in the data and reduces the bias.

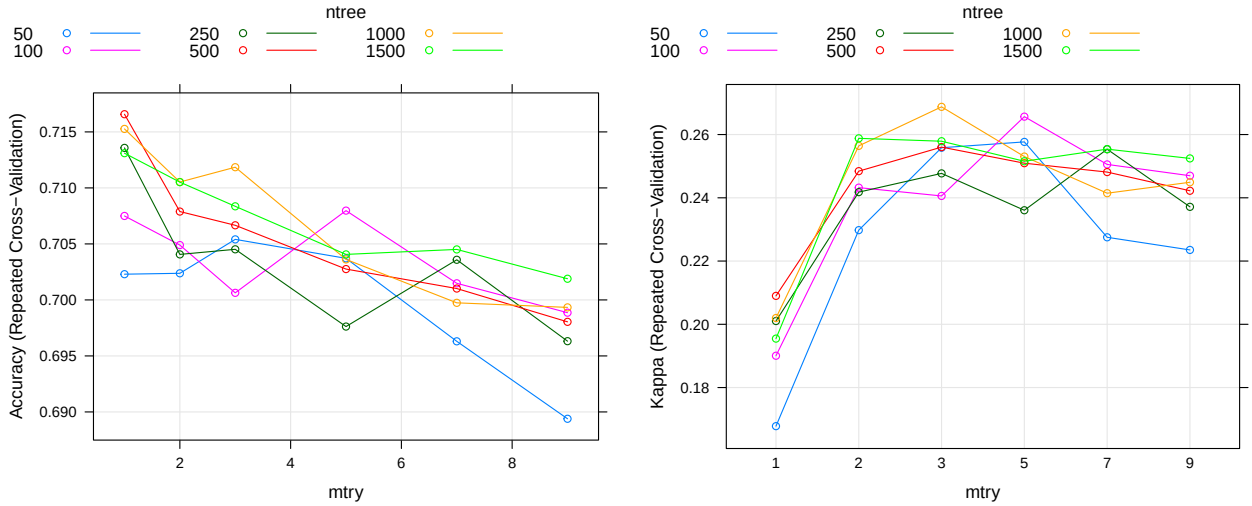
4.4.3 Model Assumptions

The model does not make any specific assumptions because it does not use any parameters and it directly creates splits in the data. The only assumptions that automatically comes with a tree structure, and that might be a drawback in some cases, is that trees automatically consider interaction among the predictor variables. If the underlying model was additive, trees might not perform that well with classification. Additionally, every single classification tree in the random forest model will be a weaker learner than a tree in the bagging method because fewer variables are used to create one, however, the overall performance might be better due to the reduction in correlation between the trees.



4.4.4 Results

Recursive feature elimination was conducted on the random forest model that was fit to the data using 5-fold cross validation. The plot shows that the accuracy increases as we increase the number of variables used to fit the model until we add the last variable "TP" to the data. This shows that the random forest model is not as sensitive, as some other models, towards features that are not highly relevant. Indeed, the performance becomes better as the variables used are increased. As indicated in the plot, we will used 9 variables, by excluding "TP", to fit the random forest model.

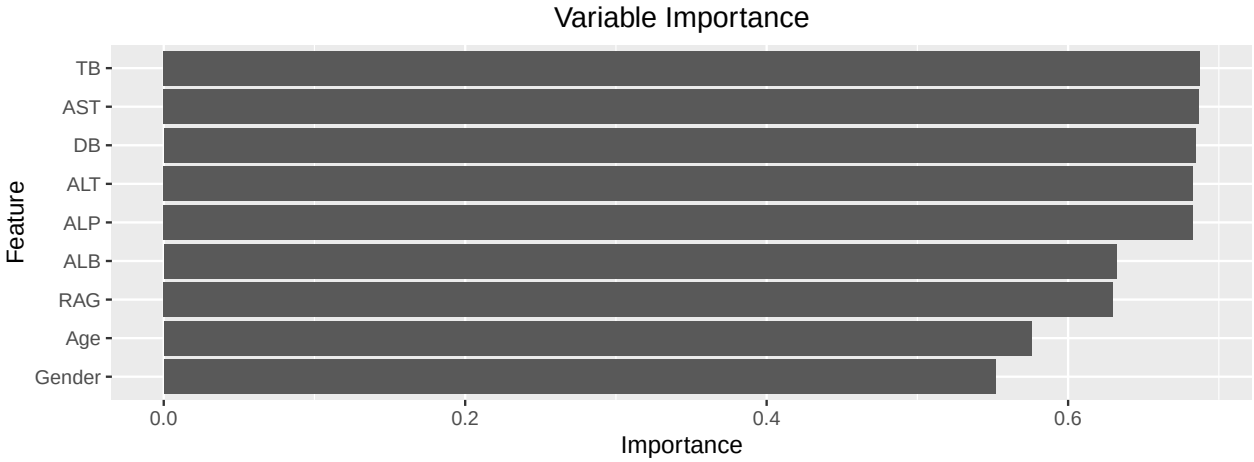


From the left plot, for all different number of trees used to fit the model, the accuracy decreases gradually as more variables are used for each iteration. However, from the right plot, the values for Kappa significantly increase as more variables are added initially and starts to become stable and gradually decrease as more and more variables are added. This shows that if fewer variables are used, specifically just 1 variable, the accuracy is high but just because the classes are imbalanced. If only 1 variable is used, most observations are classified as diseased, which is the actual label for majority of observations in our data, hence the high accuracy. This fact can be deduced from the plot of Kappa which shows the accuracy normalized by the imbalance of the classes in data.

The value of Kappa is highest for the case where $mtry = 3$ and $ntree = 1000$, and this is also the highest point in the accuracy plot apart from the points that come from the variable $mtry = 1$. Therefore, this selection can result in our model performing the best, given the imbalance between classes.

Model	Mean Out of Sample Accuracy	Mean Kappa Value
Random Forest	0.712	0.269

The table shows the results from fitting the model with $mtry = 3$ and $ntree = 1000$. These results are comparable with other models that we have seen so far.



The plot for variable importance for the selected model shows that all the variables contain almost similar importance to one another. “TB”, “AST”, “DB”, “ALT”, and “ALP” seem to be the most important, followed by “ALB”, “RAG”, “Age”, and “Gender” which also seem to be relevant to fit the random forest model.

4.5 Naive Bayes (NB)

4.5.1 About Naive Bayes

Naive Bayes is a classification approach that relies on non-parametric density estimation for continuous predictors. In approaches such as discriminant analysis, the interest is on modelling $f_k(\mathbf{x})$ which refers to how the covariates are distributed within the k^{th} response class. As the dimension of covariates increases, estimating this high dimensional density is difficult and usually enforces assumptions which may not be valid (such as a p dimensional Gaussian we see in discriminant analysis). Naive Bayes improves on this by allowing $f_k(\mathbf{x}) = \prod_{j=1}^{p} f_{kj}(x_j)$; this means we can partition the multivariate density into the product of independent, univariate densities. Each of these densities can then be modeled using a kernel-based approach with given bandwidth parameter h, or by enforcing assumptions of their form (such as assuming they are Gaussians). Both of these approaches are explored below.

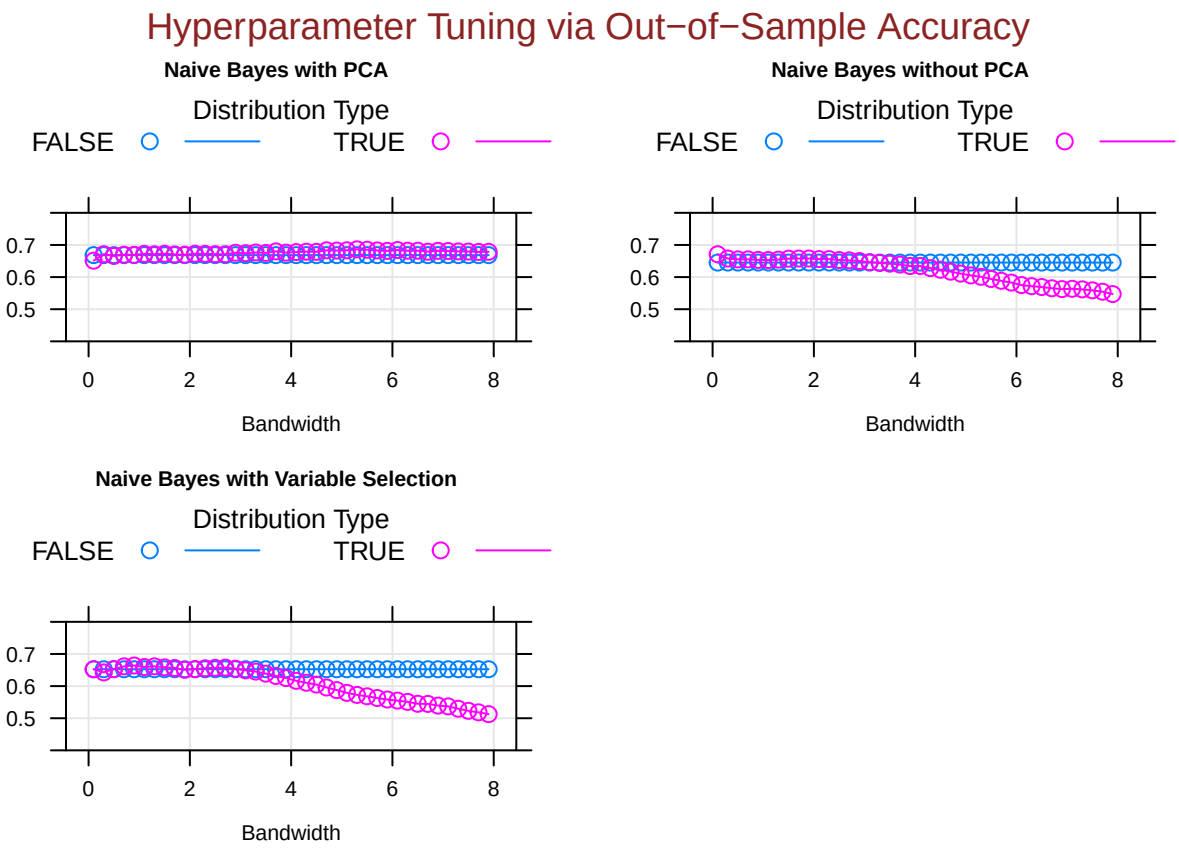
4.5.2 Model Fitting

Three major classes of Naive Bayes models were considered - one utilizing all of the original predictor variables, another using recursive feature elimination to reduce the number of covariates, and lastly using principal component analysis on the predictor set. The tuning parameters of interest in Naive Bayes depend on the kernel density estimation method being used. Each model was fit using a Gaussian density for the continuous predictors, and also fit using a more flexible non-parametric kernel density estimation approach. For the kernel density estimation iterations, the bandwidth parameter h was also tuned to achieve the optimal fit. The prediction accuracy was assessed on the out-of-sample data points using 5-fold cross-validation repeated 5 times in total, resulting in a total of 25 predictive error estimations. Both the mean and individual values of these out-of-sample prediction errors were considered to select the final Naive Bayes model.

4.5.3 Model Assumptions

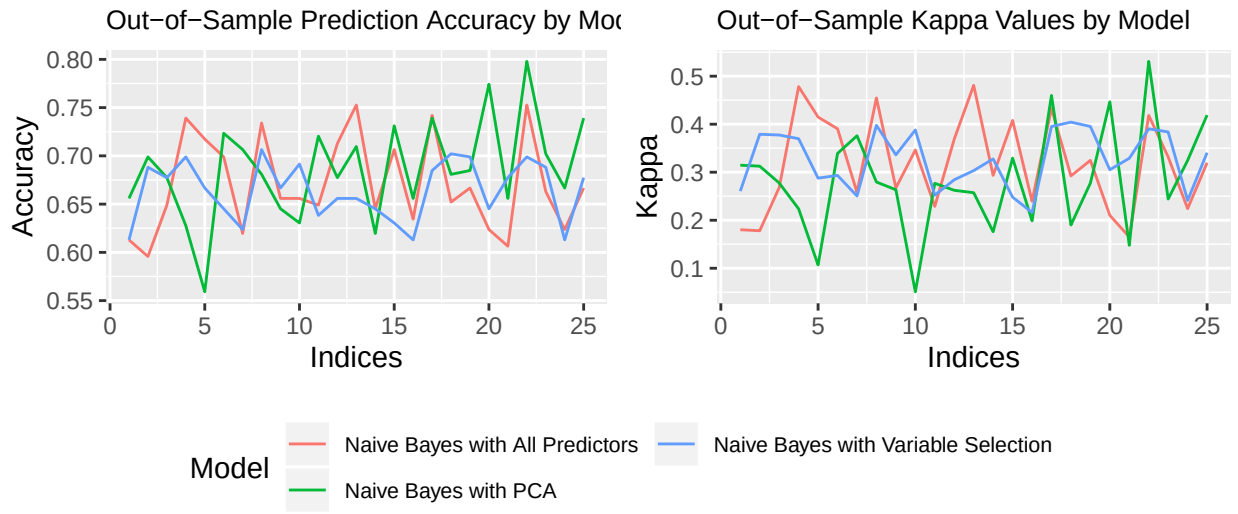
Naive Bayes assumes that within a given class, the predictor variables are independent of one another. This allows the density estimation to be partitioned into multiple univariate marginal densities, rather than estimating a high dimensional kernel density function. Within the data visualization section of the report, we see some covariates are not independent, for example total and direct bilirubin are highly correlated. This suggests the assumptions of Naive Bayes are not fully satisfied in this context. However, this can be potentially circumvented by introducing variable selection, as one variable from a correlated pair could be dropped. Additionally, many recent research papers suggest that Naive Bayes can still provide strong prediction performance despite violation of independence, as seen in "The Optimality of Naive Bayes" (Zhang, 2004).

4.5.4 Results



In tuning each model, we see the methods relying on kernel density estimation in pink and those relying on Gaussian densities in blue. Gaussian-based models do not rely on the bandwidth hyperparameter and thus we see a straight line in each case. In general, we see that Naive Bayes with PCA is relatively robust to the specification of the tuning parameters while the other two models are more sensitive. As we increase the bandwidth parameter we see a clear decline in predictive accuracy, which is expected as this can lead to oversmoothed densities. All 3 models achieve comparable predictive accuracies under optimal tuning parameter selection.

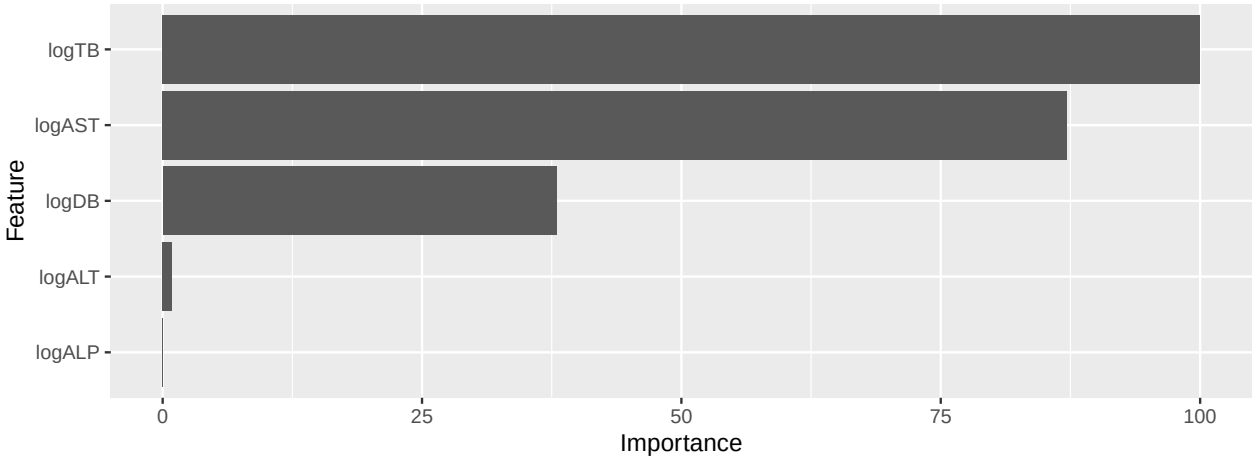
Naive Bayes with variable selection uses a kernel density estimation approach with $h = 0.9$, while the Naive Bayes without PCA uses $h = 0.1$ as the optimal tuning parameters. For Naive Bayes with PCA we increase the bandwidth to $h = 5.3$ as the optimal value, though the difference in prediction accuracy seems relatively robust to specification of h here.



Model	Mean Out of Sample Accuracy	Mean Kappa Value
Naive Bayes with All Predictors	0.671	0.319
Naive Bayes with PCA	0.686	0.283
Naive Bayes with Variable Selection	0.664	0.326

Each of the line graphs gives insight into the volatility in both predictive accuracy and the values of Kappa. The predictive accuracy for the model with all predictors seems relatively volatile in terms of prediction accuracy, and using PCA seems to give the most stable predictions across folds. The κ values relatively stable for all 3 models with some outlying points. The mean κ values show larger differences than out-of-sample accuracy, which are similar across all models.

Overall, for the implementation of Naive Bayes it seems using variable selection is the optimal choice here. This ensures our model is more parsimonious and interpretable relative to the other two choices. The predictive accuracy is comparable to both of the other models. This model contains the covariates logALT, logDB, logALP, logTB, logAST, of which we plot the importance below. The optimal tuning parameter selection involving using a kernel density estimation approach with a bandwidth value $h = 0.9$.



We can see that logTB and logAST are very important predictors, while logDB is moderately important.

logALT and logALP have low relative importance values, while age, gender, ALB and RAG were dropped using recursive feature elimination.

4.6 Boosting with XGBoost

4.6.1 About XGBoost

Extreme gradient boosting (XGBoost) is an extension of gradient boosting that implements regularization on model complexity to prevent potential overfitting. Gradient boosting is used in the scope of decision trees to minimize the loss function of the tree model. XGBoost has been a popular and successful choice in data competitions, which has resulted in it gaining a reputation as a strong learning method. XGBoost is a tree-based method similar to Random Forest (and thus is subject to similar assumptions), however it relies on boosting instead of bagging to develop the model. Boosting differs from bagging because trees are grown *sequentially*, with heavier emphasis on observations that were previously misclassified in previous iterations. This can be a particularly effective at improving the overall performance.

4.6.2 Model Fitting

We used 5-fold cross-validation on the training set to tune the hyperparameters: shrinkage/learning rate, penalty coefficient, the maximum depth of the trees, minimum number of observations per node, number of trees (iterations), sampling proportion of the training data, and the proportion of predictors to use to train the trees at every iteration. The original predictors were used instead of log-transformed predictors as these gave better prediction accuracy for each of the models considered. Variable selection using RFE was also considered.

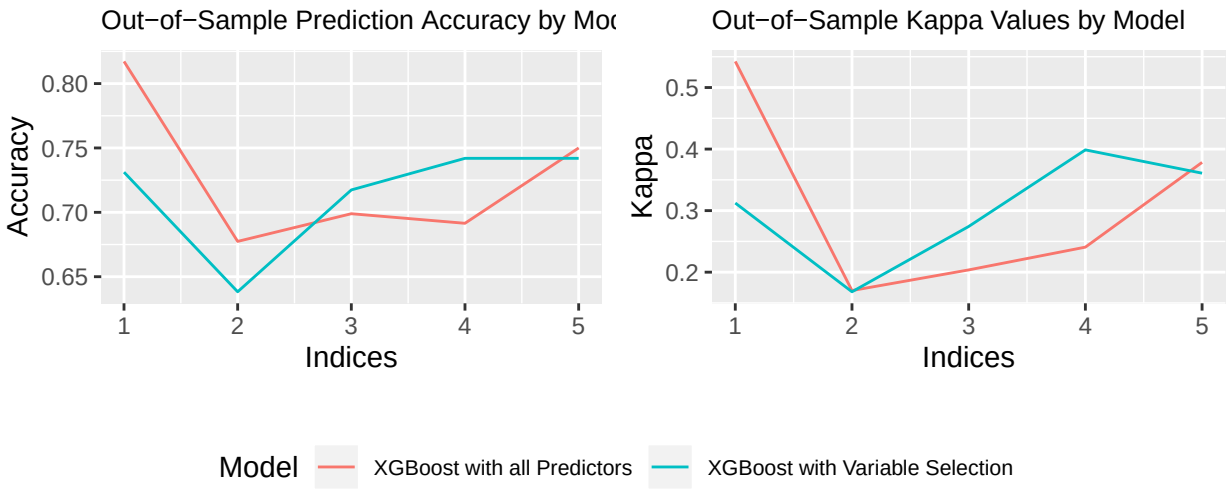
4.6.3 Results

Based on the output of the training, the optimal parameters which were selected based on the highest accuracy for the classifier on all predictors are: a shrinkage of 0.5, a penalty coefficient of 0.1 maximum depth of 5 terminal nodes, a minimum of 15 observations per node, 100 trees in the ensemble, randomly sampling 75% of the training data for each tree, and randomly selecting 50% of the training data for each tree. The training accuracy and the Kappa value of the XGBoost classifier with these set of classifiers was 0.727 and 0.307 respectively.

We may also consider using RFE to select features before fitting the XGBoost classifier on the training data.

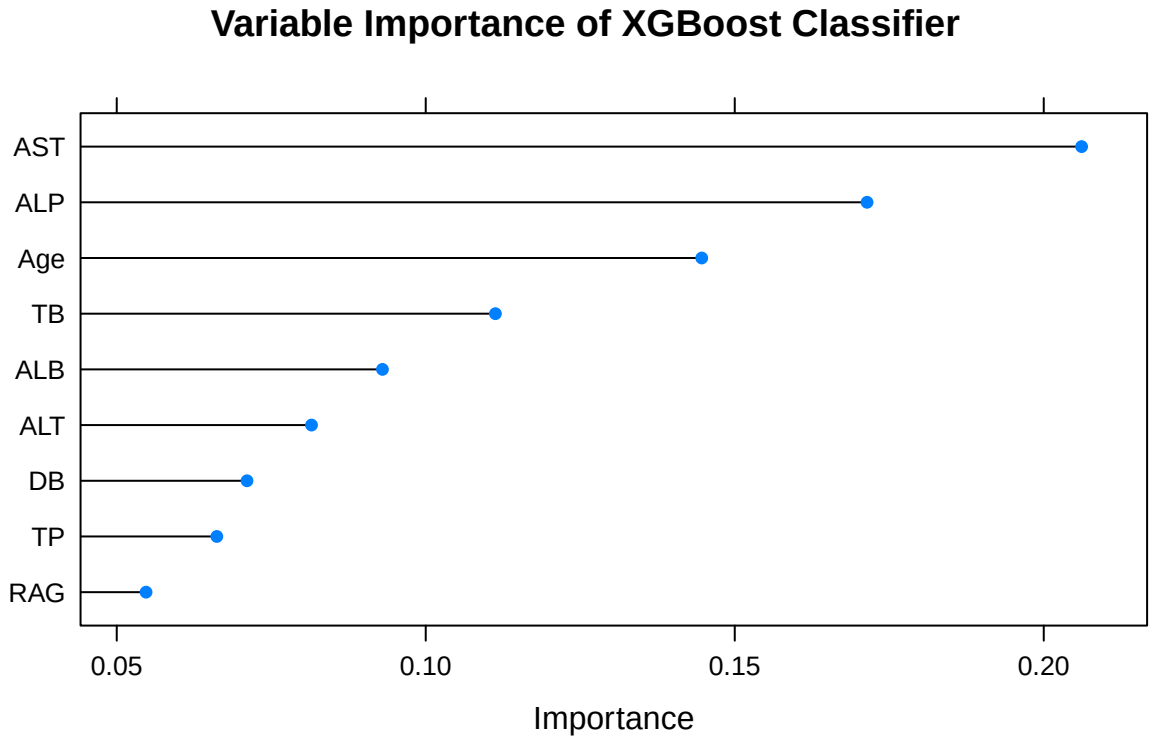
Applying RFE gives that the optimal variables are ALP, ALT, AST, Age, ALB, TB and TP.

Based on the output of the training, the optimal parameters which were selected based on the highest accuracy for the XGBoost classifier with variable selection are: a shrinkage of 0.1, a penalty coefficient of 0.1 maximum depth of 5 terminal nodes, a minimum of 5 observations per node, 300 trees in the ensemble, randomly sampling 75% of the training data for each tree, and randomly selecting 75% of the training data for each tree. The training accuracy and the Kappa value of the XGBoost classifier with these set of classifiers was 0.714 and 0.303 respectively.



Model	Mean Out of Sample Accuracy	Mean Kappa Value
XGBoost with all Predictors	0.727	0.307
XGBoost with Variable Selection	0.714	0.303

Since both the mean out-of-sample accuracy and the mean Kappa value for the XGBoost classifier containing all the predictors perform better than the XGBoost classifier with variables selected based on RFE, we choose the former classifier as the better performing one.



We note that AST is the most important predictor; after AST, the decrease in importance is almost similar proportionally, with RAG being the least important predictor.

5 Statistical Conclusions

To determine the best model out of the 5 models that were discussed, we will look at the mean out-of-sample accuracy from cross validation on the training set, mean Kappa values from cross validation on the training set, and test accuracy for each model.

We consider mean out-of-sample accuracy values for the purpose of checking the consistency of the model by comparing the value with the test accuracy value. If the two values are close enough it indicates that the model was highly consistent.

We use mean Kappa values because they are useful at normalizing the accuracy values with respect to the imbalance in our classes. A low Kappa value indicates that a high proportion of values for one or both of the classes were classified incorrectly. Since proportions are considered, instead of the actual numbers, class imbalance does not effect this metric. We will not use Kappa values directly because they are from the cross-validation on training set, but they give us an idea of what the confusion matrix would look like.

Lastly, we consider test accuracy values because these values have not been used at all for fitting the models, and therefore, they give us an unbiased view of the accuracy for the model.

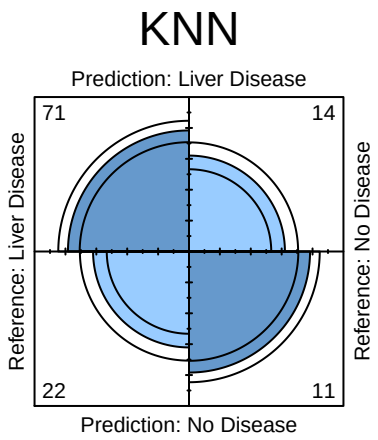
Table 11: Summary of final results

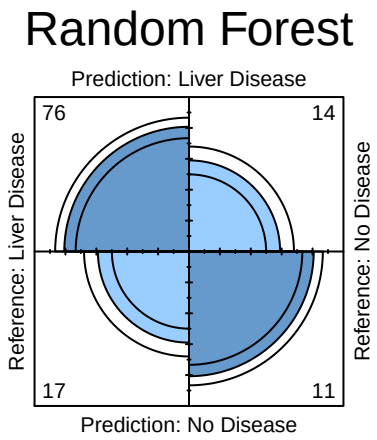
Model	Out-of-Sample Accuracy	Mean Kappa Value	Test Accuracy
Naive Bayes	0.664	0.326	0.661
KNN	0.713	0.297	0.695
LDA	0.683	0.084	0.729
Random Forest	0.712	0.269	0.737
XGBoost	0.727	0.307	0.746

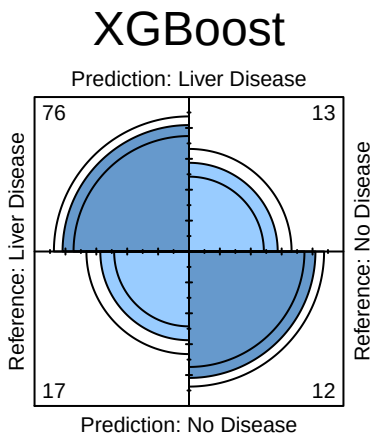
In terms of the mean out of sample accuracy and test accuracy, Naive Bayes has the lowest numbers compared to the other 4 models. It does have a high value for the mean of Kappa values. Since we have a combination of comparatively lower accuracy, with a higher Kappa, it indicates that the model performs well with classifiying patients with no disease correctly, but it is not as accurate with classifying patients with liver disease. This might cause serious problems, hence we try to look at other models that might perform better than Naive Bayes.

We look at LDA next since it has a higher out-of-sample accuracy and test accuracy than Naive Bayes. Even though the values for accuracy are comparatively high, they seem to be a little inconsistent. On top of that, the Kappa value for LDA is very low compared to the other models. A combination of low Kappa value with high accuracy indicates that the model is assigning most of the observations to the class that we have in majority. This would result in high accuracy because most of the majority class will be assigned correctly, but it comes at the cost of a very high proportion of values from the minority class being assigned to the majority class as well. In our case, this will lead to a prediction of Liver Disease for a very high number of individuals that actually do not have the disease, which is again a serious problem.

Therefore, after eliminating Naive Bayes and LDA due to comparatively low accuracy and low kappa value respectively, we look at the other three models that seem to have comparatively higher and very similar accuracy and kappa. KNN, Random Forest, and XGBoost seem to have very similar out-of-sample accuracy levels. In terms of test accuracy, the results for Random Forest seem to be the most consistent to out-of-sample accuracy. KNN has a lower test accuracy, and XGBoost has a higher test accuracy compared to the out-of-sample accuracy. This could just be due to the specfic test set that we have in this case, and performance might vary a bit with a different test set, however all three models perform at a similar level overall.







The plot shows the confusion matrices for the three models. XGBoost performs the best at classifying individuals with Liver Disease, followed by Random Forest, and then KNN. For individuals with no disease, XGBoost again does the best job, followed by KNN, and then Random Forest. Even though the numbers differ a little, they are very close to each other.

Another thing to consider here would be the interpretability of the model we choose. Our goal here is to find the optimal trade-off between interpretability, parsimony and interpretability for a clinician. Out of the three models, KNN is the easiest to interpret and to explain to a clinician. The model simply uses the blood test results of an individual to identify the diagnosis label based on the diagnosis label of other individuals with similar blood test results.

Moreover, the model selected for KNN was after variable selection, so it uses only 4 variables. On the other hand, the model used for Random Forest makes use of 9 variables and the model for XGBoost uses all 10 variables. Therefore, KNN also allows for a parsimonious model with all variables containing high and equal levels of importance.

Since, KNN allows for higher interpretability and parsimony for a slight decrease in accuracy, we will select the KNN model as our final model with variables "AST", "TB", "DB", and "ALP" and tuning parameter $k = 25$.

6 Conclusions and Future Work

Our chosen model, KNN, uses the variables "AST", "TB", "DB", and "ALP". In the data exploration section, we saw all these variables to be very important because they had an evident trend in the training data. AST (Aspartate Aminotransferase) is an enzyme found in the blood and liver cells and is generally higher among

the individuals with liver disease. TB (Total Bilirubin) and DB (Direct Bilirubin) are both measurements related to the Bilirubin levels in blood, which is a chemical that occurs as a result of red blood cells breaking down. Total Bilirubin includes Direct Bilirubin and Indirect Bilirubin, therefore the correlation among these two variables is very high. Even with the high correlation, both these variables were considered important variables for the model. This means that Bilirubin levels in blood are very strong predictors for Liver Disease because it is essentially considered with double the amount of weight as any other variables in our model. ALP (Alkaline Phosphotase) is an enzyme found within bone cells and the liver and is also generally higher among the individuals with Liver Disease.

Most of these function tests are useful indicators of liver disease, however, some of them might also indicate diseases other than liver disease. For example, increased AST levels might indicate liver damage but it could also indicate muscle damage. Similarly, increased bilirubin levels might indicate liver damage or certain types of anemia. Increased ALP levels might indicate liver damage or certain bone diseases (Mayo Clinic, 2019). Therefore, we need to be careful when determining the diagnosis of an individual using the model, because the blood test levels might be an indicator of diseases other than liver damage too. It might help to have the data on any other diseases that the patients had other than liver damage. It would be useful to explain the variations in blood test levels even more, and make better classifications, as to if an individual had liver disease, some other disease, multiple diseases, or no disease. Another variation for this idea is to have the data on the severity of the liver damage. Some of these measurements in the blood test are an indicator for liver disease, but also liver damage. Knowing the measurements associated with the specific case, and the severity of the damage might help with classifcation too, by having different classes for different levels of severity of damage or disease. This way the cases predicted as high severity might be worth looking into more.

Some other additions worth considering are to include other measurements from blood test levels that might also be an indicator of liver disease or damage. These measurements can include Gamma-glutamyltransferase (GGT), L-lactate dehydrogenase (LD), and Prothrombin time (PT) (Mayo Clinic, 2019). Additionally, covariates ascertained from patient's medical records, such as height, weight, family history of disease, smoking status, etc could be included too. These covariates could potentially act as confounders and may be useful in improving prediction accuracy. Currently, the data size is also not that large, specially the observations for no disease, which also causes an imbalance in the training data. A larger dataset, with balanced classes, might be helpful in improving the performance of the trained model.

Moreover, the purpose of this report was to find a method that creates an optimal trade-off between accuracy, parsimony, and interpretability therefore complex models were not considered in this report. Models such as support vector machines or neural networks are becoming increasingly popular in recent work due to their strong predictive performance. These could be compared relative to interpretable models to see how drastic the increase in performance is.

7 Contributions

Each group member contributed equal amounts of time and effort to the final report; two models were fit per team member, and then a body section of the report was completed by each group member.

- **Discriminant Analysis, XGBoost, Data Pre-processing**: Angela
- **KNN, Random Forest, Conclusion**: Asad
- **Introduction, Naive Bayes, Logistic Regression, Data Visualization + Interaction**: Marcus

8 Appendices

8.1 Appendix A: Variable Interaction

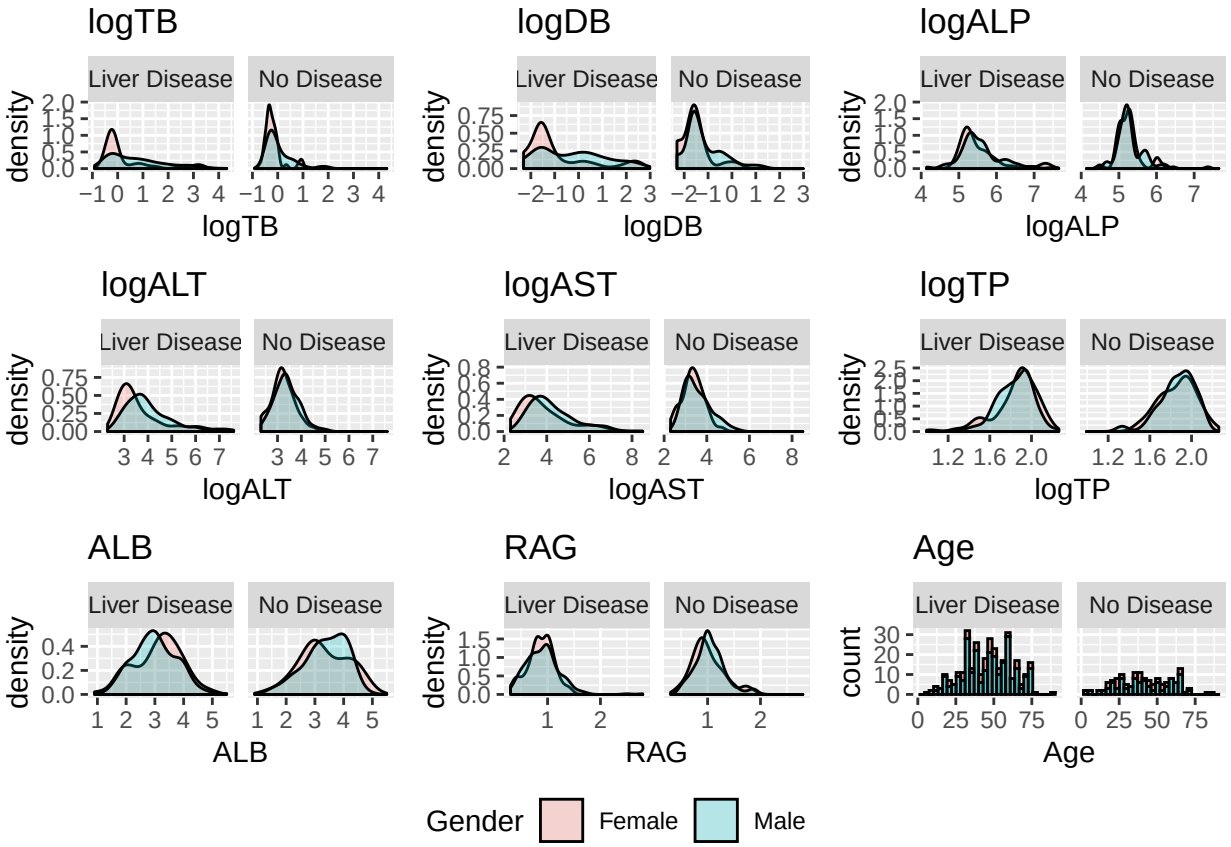
8.2 Interactions, Correlations, and Relations Among Predictors

In addition to looking at the univariate distributions of predictor variables, we may also be considering in two dimensional (or higher) conditional distributions. This can help identify variables which are correlated with one another, potential interactions between variables, or trends otherwise not apparent from a marginal distribution alone.

Within this report, the primary interaction of interest lies between the predictor variables and sex, as well as age. These features are commonly assessed in biomedical studies, since biological differences resulting from sex or age can lead to differences in blood protein levels. To avoid *data dredging* (fitting without first considering some underlying hypothesis or relationship in the data, like inserting all two-way interactions into a model without considering their interpretation or sensibility), interactions are considered before model fitting and supported through external research.

To get an idea of how interaction with sex and age may occur, the training data is partitioned into 4 groups, based on disease status and either sex (male/female) or age (splitting the distribution of ages into 3 distinct groups). The densities between the controls and cases are then compared; if we see a similar trend in how the disease status changes the density between each sex or age group, we do not have strong evidence of interaction between that covariate and sex or age with respect to disease status. However, if we notice that the change in distribution of a covariate between disease groups is dependent on sex or age, this suggests an interaction.

8.2.1 Interaction with Sex



Looking at each of the plots:

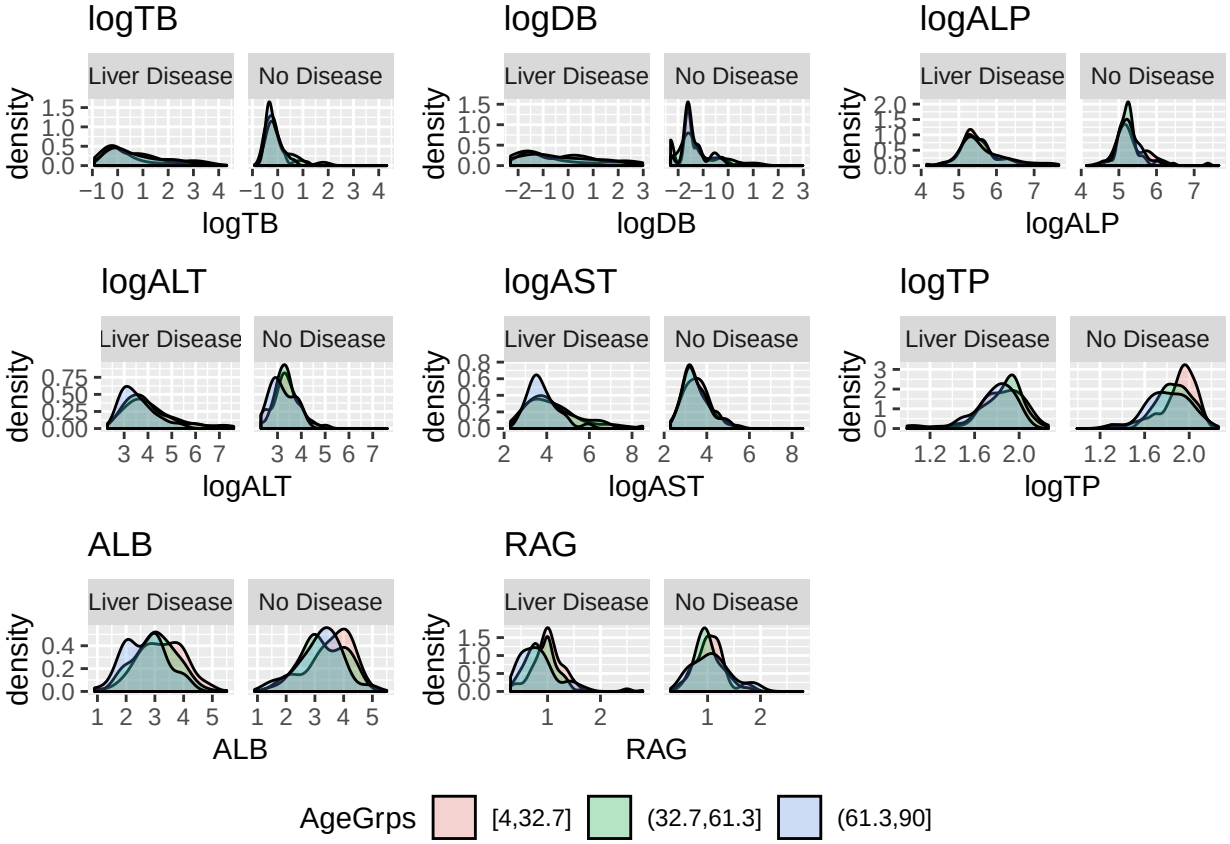
- As seen before, logTB and logDB tend to increase in cases relative to controls. However, the magnitude of increase appears to be greater for males than it does for females, as evident by the greater degree of elongation. This suggests that there may be an interaction between logTB and sex with respect to predicting disease.
- The change in logALP appears to be similar for both males and females, suggesting no major interaction. Similarly, we see that logALT, logTP, RAG and age appear to have consist trends between genders.
- For logAST, there is a slight trend in that the increase for men among controls to cases seems to be larger than that of females. This is evident because among the controls, the densities almost perfectly overlap (though males are slightly shifted to the right), but in the cases there is a more considerable horizontal shift in males relative to females.
- ALB also appears to have sex-based interaction; looking at the individuals with no disease, males tend to have higher ALB levels than females. However for diseased patients, females tend to have higher ALB levels than males. This may suggest the magnitude of increase is larger for females who develop liver disease.

In conclusion, it appears that interaction between sex and logTB, logDB, logAST and ALB should be considered in the initial model fitting.

8.2.2 Relationship with Age

To examine age, we use a simplified approach to partition age into 3 equally sized groups, though the sample size may differ between groups. The age groups of interest are [4,32.7] (red), (32.7,61.3] (green), and (61.3,90] (blue). Similar to sex, we examine whether or not the change between controls and cases of a given covariate

changes in a similar manner for each age group.

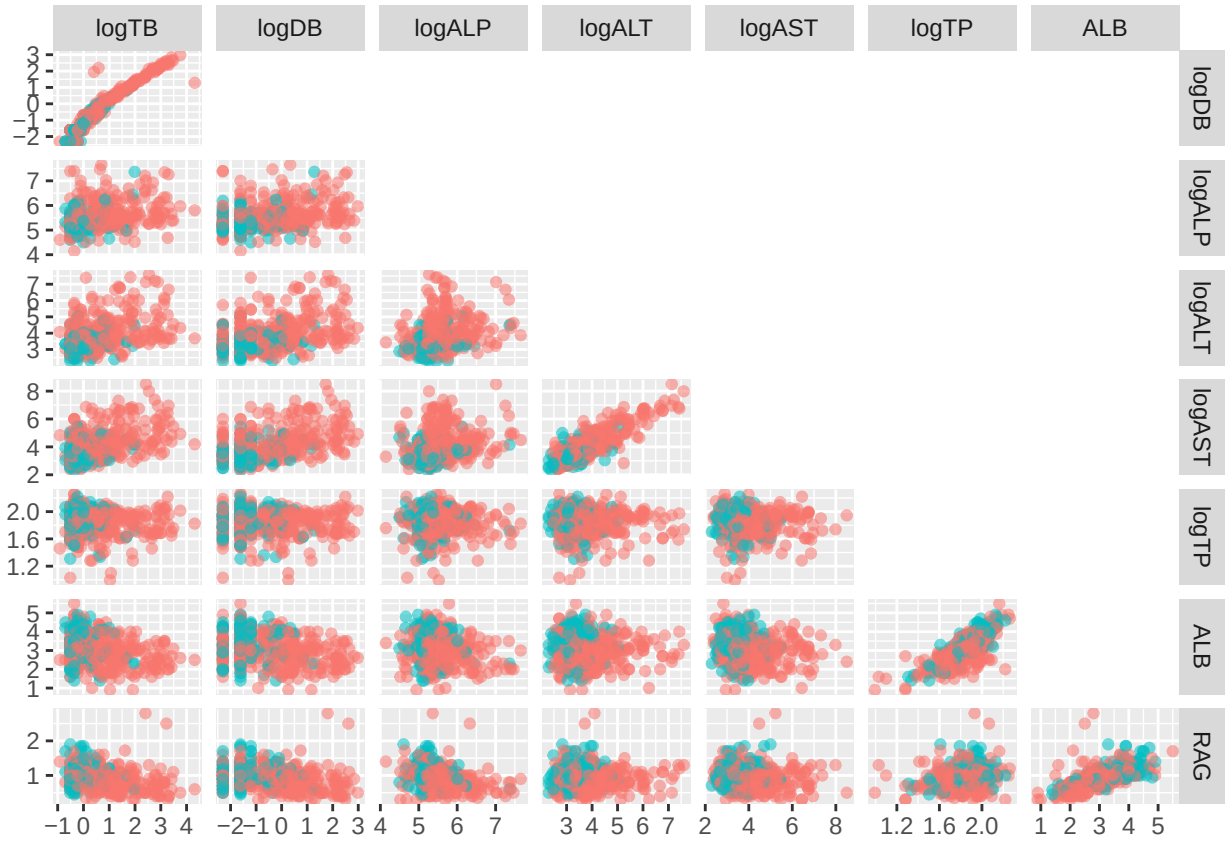


- Looking at logTB, we notice that the change between groups seems dependent on age. For the group in green (intermediate ages), we see that the bimodality in the controls becomes masked in the diseased group, whereas in the red group (lowest age group) still maintains bimodality. The difference in change in distribution between groups suggests age and logTB interact.
- Similarly, logDB appears to interact with age as the change in densities varies greatly between the two groups. The magnitude in which logDB increases from cases to controls appears to be much larger for the lowest and highest age groups compared to the intermediate group.
- The change in logALP, logALT, logAST,appears to be relatively consistent among ages.
- For logTP, we see that in the controls there seems to be a clear increasing gradient of logTP levels as age decreases. However, when we look at the controls, we see that this order changes, and there is considerably more overlap among the densities. In particular, we see that for the youngest group, the decrease in logTP is the most dramatic. This suggests an age-based interaction.
- Similar to logTP, for ALB levels we see differences in the change among age groups. For older individuals, liver disease seems to lower ALB levels, but for the intermediate and oldest age group, the disease appears to slightly increase ALB levels. The magnitude of change also seems different for the 3 groups, meaning an interaction should be considered.
- Lastly, for RAG we see that younger and intermediate ages seem to have relatively consistent RAG levels between groups. However for the oldest age group, the RAG levels tend to decrease considerably for diseased patients relative to controls.

In conclusion, we should consider interaction between age and logTB, logDB, logTP, ALB and RAG within the initial model fitting.

8.2.3 Correlation between Covariates

Based on what variables were collected, we expect to see some degree of correlation among some of the predictor variables in the dataset. For example, total bilirubin levels are expected to be correlated with direct bilirubin levels since total bilirubin is given by direct + indirect bilirubin levels. Correlated predictors can be an issue as this can lead to collinearity, and ultimately very large variances in our estimates. Exploring the pairs plot below can give insight into what variables are correlated. Similar to before, we have cases in red and controls in blue.



- Clearly total bilirubin and direct bilirubin are higher correlated, as there is almost a perfect linear relationship between them.
- ALT and AST levels, ALB and logTP also appear to be positively correlated with a strong degree of correlation.
- RAG and ALB levels have a weak positive correlation.

Based on these results, we have some strategies for dealing with these correlations. Each of these approaches will be applied to the models (where applicable) and then compared to a standard model in which none of the approaches are used:

1) Dropping one of the covariates in a pair of highly correlated measurements, especially total and direct bilirubin. These variables measure essentially the same thing, though arguably total bilirubin is more informative since it also includes indirect bilirubin levels as well. By only including one of these predictors, we reduce the correlation among them while still retaining the underlying information. This can be done in the variable selection step when fitting each of the models.

2) Implementing regularization to reduce the standard error of the estimated coefficients (dependent on the model of interest being considered). The advantage of this approach is the reduction of variance, but the cost is the introduction of bias in the estimation of parameters. In order to assess the improvement that regularization will provide, a regularized model will be compared against a non-regularized model. The

advantage of regularization over other approaches is that we can still interpret the resultant coefficients.

3) Introducing Principal Component Analysis. This approach is straightforward to implement, and will ensure our resultant predictors (principal components) are uncorrelated. However, in the context of this report, using PCA may not be suitable. By implementing PCA, the result is a linear combination of the original predictor variables, which can make interpretation difficult to impossible. This is especially an issue since the intended purpose of the fitted models is to not only predict liver disease, but also communicate these results to health professionals. By losing interpretation, a key goal of this report is violated. Therefore the results of PCA will be examined, but it will only be used if the final model is a significant improvement over the other previous approaches. This would suggest that maintaining both high predictive accuracy and interpretation in the final model may not be achievable.

Table 12: Correlation Matrix

	logTB	logDB	logALP	logALT	logAST	logTP	ALB	RAG
logTB	1.00	0.96	0.36	0.45	0.53	-0.06	-0.31	-0.27
logDB	0.96	1.00	0.36	0.44	0.52	-0.05	-0.29	-0.26
logALP	0.36	0.36	1.00	0.32	0.30	-0.02	-0.20	-0.30
logALT	0.45	0.44	0.32	1.00	0.85	-0.04	-0.07	-0.04
logAST	0.53	0.52	0.30	0.85	1.00	-0.07	-0.19	-0.12
logTP	-0.06	-0.05	-0.02	-0.04	-0.07	1.00	0.78	0.22
ALB	-0.31	-0.29	-0.20	-0.07	-0.19	0.78	1.00	0.66
RAG	-0.27	-0.26	-0.30	-0.04	-0.12	0.22	0.66	1.00

As expected, we see identical correlation patterns to what was observed before. The logDB and logTB values are almost perfectly correlated ($\rho = 0.96$), while logTP and ALB and ALB and RAG also have relative large correlations ($\rho = 0.78$ and $\rho = 0.66$, respectively).

8.3 Appendix B: Logistic Regression

8.4 Logistic Regression

Note: The final test set accuracy using the selected logistic model was very poor, and was therefore not considered in the final comparison of all models.

8.4.1 About Logistic Regression

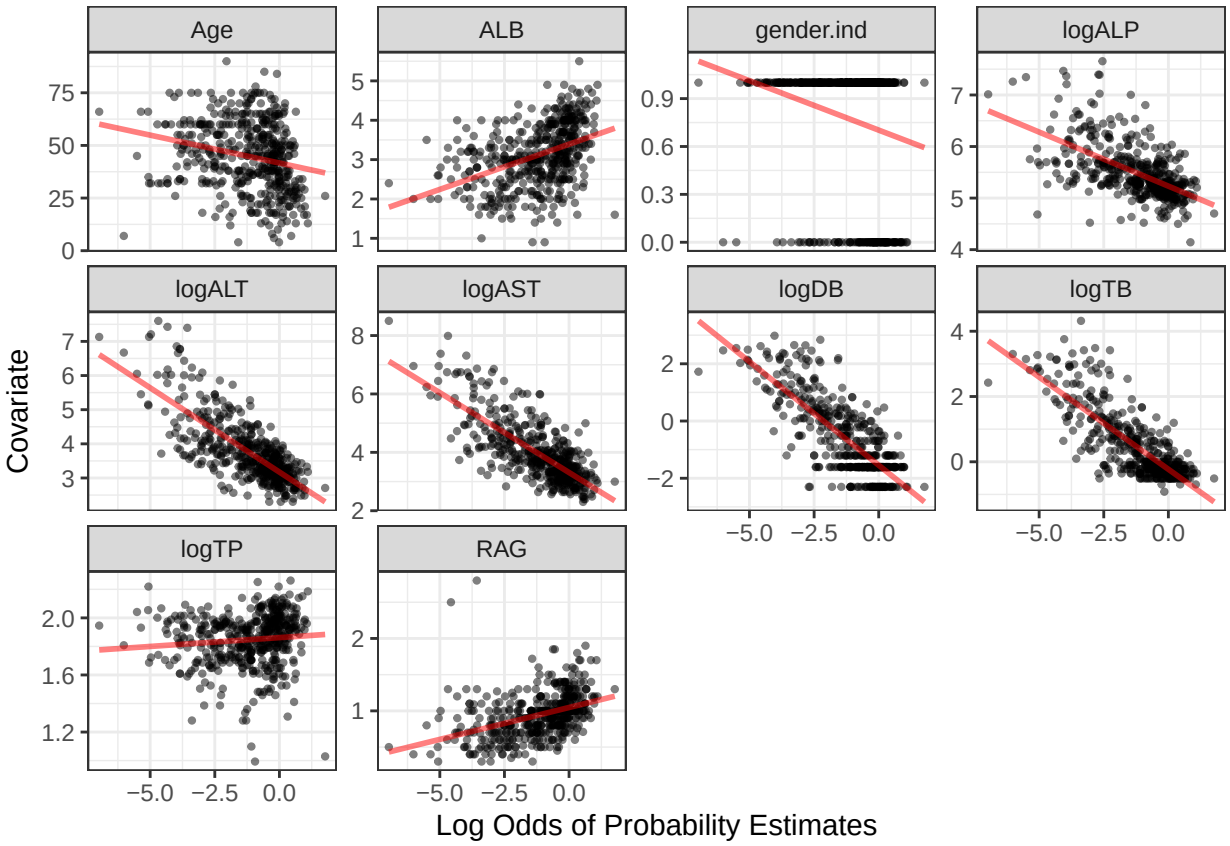
Logistic regression is a fully parametric approach which models the conditional distribution of $Y|X \sim Bernoulli(p)$. Though relatively simple in comparison to many modern machine learning techniques, logistic regression remains popular in biomedical and clinical research as a result of its easy implementation and interpretation. By using logistic regression, features such as coefficient interpretation, variable importance, hypothesis testing and visualization are relatively straightforward when compared to more complex models seen later in the report. Logistic models are also highly customizable, and can be incorporated with tools such as principal component analysis, regularization, feature selection, and non-parametric smoothing.

8.4.2 Model Fitting

5 types of logistic models were considered in this report, and then their performance based on repeated 5-fold cross-validation on the training set was assessed. These models included logistic regression with all of the covariates in the dataset, logistic regression with two-way interactions (in which the two-way interactions were chosen based off of the corresponding appendix section), logistic regression using forward stepwise regression, using PCA, and lastly non-parametric logistic regression using splines. The goal was to select the model with the optimal 'trade-off' between prediction performance, parsimony, and interpretability.

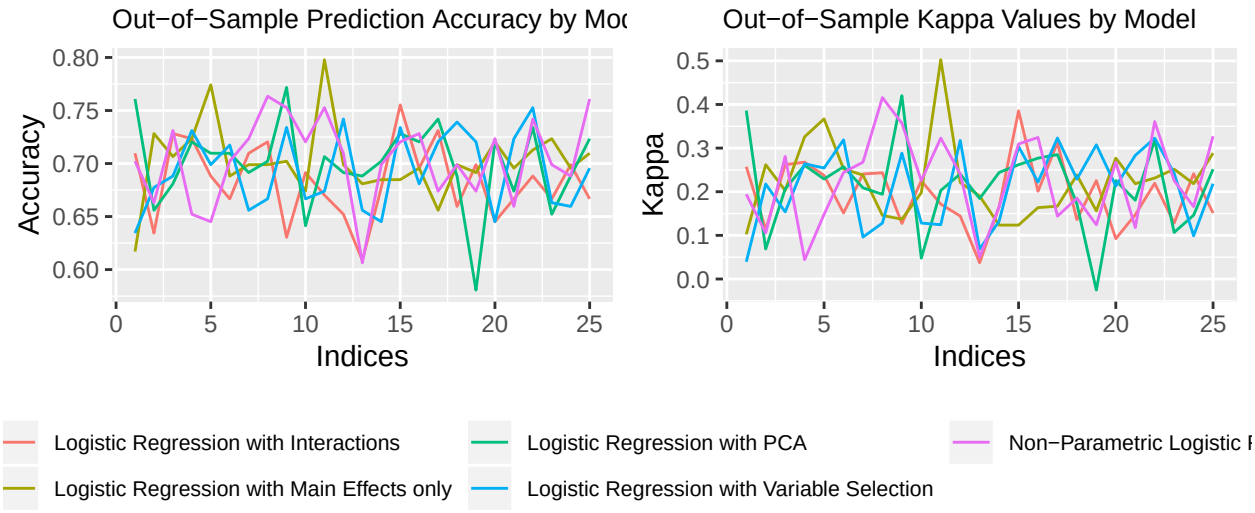
8.4.3 Model Assumptions

Logistic regression is a fully parametric approach which assumes a linear relationship between the log-odds of response and the predictor variables. In practice, this assumption may be relatively restrictive; occasionally we will need to incorporate higher order polynomial terms, introduce flexible basis functions like splines, or carry out other data transformations. In order to assess the validity of the linearity assumption, a logistic model was fit on the entire training set and each continuous covariate was plotted against the log-odds of response, denoted by the x axis. Excerpts of code from this article were used to develop these diagnostic plots:



In red we can see a simple linear regression line overlaid to each plot. With the exception of age and logTP, the linearity assumption seems relatively reasonable in all cases. This means the assumptions behind logistic regression are satisfied, and therefore it seems like a reasonable model to try. However, due to the potential non-linearity in a few covariates, we also try a non-parametric generalized additive model approach in order to potential accommodate more flexible shapes.

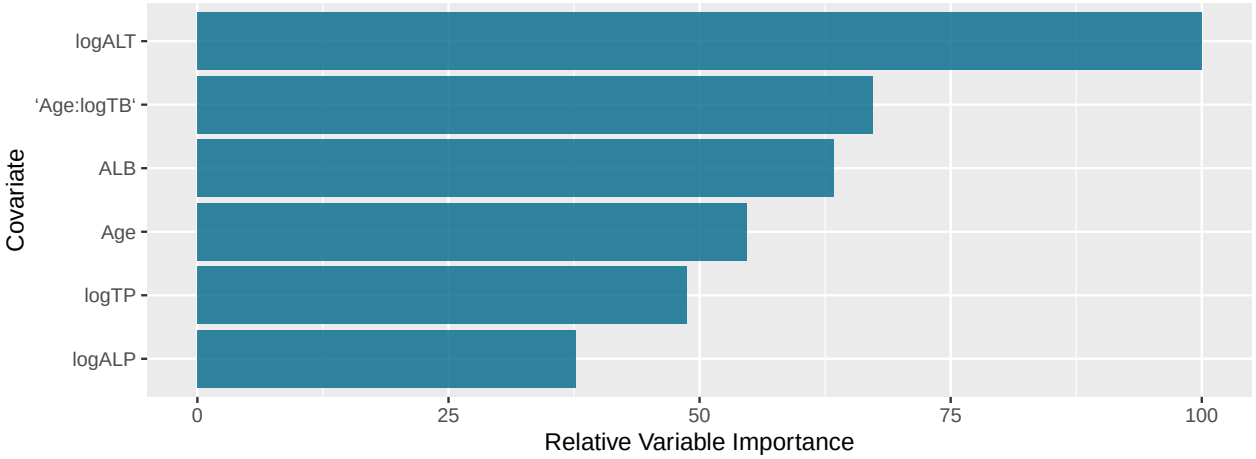
8.4.4 Results



Model	Mean Out of Sample Accuracy	Mean Kappa Value
Logistic Regression with Main Effects only	0.702	0.224
Logistic Regression with Interactions	0.683	0.195
Logistic Regression with Variable Selection	0.693	0.212
Logistic Regression with PCA	0.699	0.214
Non−Parametric Logistic Regression	0.704	0.225

We see very similar results between all 5 models in terms of mean out-of-sample accuracy among the 25 iterations. In terms of both the accuracy and κ trend lines, we note that these results are quite volatile relative to some other models examined, particularly in terms of the κ values. Though the prediction accuracy is relatively high, the κ values suggest none of these models are particularly effective predictors here. The test errors are relatively similar but it seems that introducing the two-way interactions seems to lead to a sizable increase in accuracy.

Since the results are similar across all 5 models, the most sensible choice is to select the logistic model involving variable selection. This model is more parsimonious than the main effect or interaction models, and more interpretable than a model relying on PCA or non-parametric methods. The trade-off is relatively minimal in terms of κ and prediction accuracy compared the non-parametric logistic regression, but we retain a more interpretable form that could be better explained to clinicians and patients. These results are relatively expected given the linearity assumption was relatively satisfied. Examining the chosen model more thoroughly, we see that Age, logALP, logALT, logTP, ALB and Age:LogTB were maintained in the final model. Looking at variable importance:



Clearly logALT is the most important variable, however the remaining variables are all moderately important which is unlike some of the other models examined. This means the forward variable selection approach worked rather well here, as we maintain good predictive accuracy and only model predictors which are important to the response.

8.5 Appendix C: Code

Please see the attached file with all of the code used to generate this report.

9 References

Ahirwar, R., & Mondal, P. R. (2018, September 21). Prevalence of obesity in India: A systematic review. Retrieved November 16, 2019, from https://www.sciencedirect.com/science/article/pii/S1871402118303655.

Albumin. (2019, July 29). Retrieved November 19, 2019, from https://labtestsonline.org/tests/albumin.

Alcoholism in India. (2019, October 30). Retrieved from https://alcoholrehab.com/alcoholism/alcoholism-in-india/.Cunha, J. P. (2019, July 3).

Alkaline Phosphatase (ALP). (2019, November 13). Retrieved November 19, 2019, from https://labtestsonline.org/tests/alkaline-phosphatase-alp.

ALT Blood Test: MedlinePlus Lab Test Information. (2019, November 1). Retrieved November 19, 2019, from https://medlineplus.gov/lab-tests/alt-blood-test/.

Aspartate Aminotransferase (AST) Test (aka SGOT): High vs. Low Levels. (2019, May 15). Retrieved November 19, 2019, from https://www.webmd.com/a-to-z-guides/aspartate_aminotransferse-test#1.

Bilirubin test. (2018, November 6). Retrieved November 19, 2019, from https://www.mayoclinic.org/tests-procedures/bilirubin/about/pac-20393041.

Blocka, K. (2018, July 26). ALT (Alanine Aminotransferase) Test. Retrieved November 19, 2019, from https://www.healthline.com/health/alt.

Cunha, J. P. (2019, July 3). What Is Cirrhosis of the Liver? Symptoms,Treatment, Causes & Stages. Retrieved November 16, 2019, from https://www.medicinenet.com/cirrhosis/article.htm.

Guy, J., & Peters, M. G. (2013, October). Liver disease in women: the influence of gender on epidemiology, natural history, and patient outcomes. Retrieved November 19, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992057/.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. New York, NY: Springer.

Hoffman, M. (2019, May 18). Liver (Anatomy): Picture, Function, Conditions, Tests, Treatments. Retrieved November 16, 2019, from https://www.webmd.com/digestive-disorders/picture-of-the-liver#1.

Is liver disease the next major lifestyle disease of India after diabetes and BP? - Times of India. (2017, April 11). Retrieved November 16, 2019, from https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/is-liver-disease-the-next-major-lifestyle-disease-of-india-after-diabetes-and-bp/articleshow/58122706.cms.

Kang, H. (2013, May 24). The Prevention and Handling of the Missing Data. Retrieved November 28, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/.

Kim, I. H., Kisseleva, T., & Brenner, D. A. (2015, May). Aging and liver disease. Retrieved November 19, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4736713/.

Kuhn, M. (2009, June 30). Variable Selection Using The caret Package. Retrieved November 26, 2019, from https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/caret/inst/doc/caretSelection.pdf?revision=77&root=caret&pathrev=90.

Liver disease. (2018, March 13). Retrieved November 16, 2019, from https://www.mayoclinic.org/diseases-conditions/liver-problems/symptoms-causes/syc-20374502.

Liver Disease in India. (n.d.). Retrieved from https://www.worldlifeexpectancy.com/india-liver-disease.

Liver function tests. (2019, June 13). Retrieved November 19, 2019, from https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595.

Lutins, E. (2017, September 5). DBSCAN: What is it? When to Use it? How to use it. Retrieved December 1, 2019, from https://towardsdatascience.com/best-clustering-algorithms-for-anomaly-detection-d5b7412537c8

McHugh M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276-282.

Narasimhan, G. (2016, April). Living donor liver transplantation in India. Retrieved November 16, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4824736/.

Pati, S., Sharma, K., Zodpey, S., Chauhan, K., & Dobe, M. (2012, June 24). Health promotion education in India: present landscape and future vistas. Retrieved November 16, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4776916/.

Raypole, C. (2019, March 21). Liver Diseases: List of Problems, General Symptoms, Diagnosis, More. Retrieved November 16, 2019, from https://www.healthline.com/health/liver-diseases#treatment.

Sander, J., Ester, M., Kriegel, H., & Xu, X (1998, June). Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data Mining and Knowledge Discovery 2, 2 (1998), 169-194. Retrieved December 1, 2019, from https://link.springer.com/article/10.1023%2FA%3A1009745219419.

Slightam, C. (2016, December 1). Total Protein Test. Retrieved November 19, 2019, from https://www.healthline.com/health/total-protein#proteins.

Sterne et al. (2009, January 30). Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. Retrieved November 28, 2019, from https://www.bmj.com/content/338/bmj.b2393.

Vora, P. (2017, January 3). India is discarding needles but reusing syringes and this is spreading disease. Retrieved November 16, 2019, from https://scroll.in/pulse/813346/docs-reusing-syringes-improper-biomedical-waste-disposal-raise

Wedro, B. (2019, November 7). Liver Disease Symptoms, Signs, Diet & Treatment. Retrieved November 16, 2019, from https://www.medicinenet.com/liver_disease/article.htm.

What Is Cirrhosis of the Liver? Symptoms,Treatment, Causes & Stages. Retrieved November 16, 2019, from https://www.medicinenet.com/cirrhosis/article.htm.