STATS 503 Project Report

Predicting Liver Disease Presence and Progression

Yukiko Yamazaki,Shreya Mittal,Mingcan Zhao

July 4, 2022

Abstract

Liver biopsy represents the "gold standard" for the assessment and quantification of liver fibrosis in chronic hepatitis C virus (HCV) patients. However, they are invasive, expensive and pose serious consequences leading to a low patient acceptance. This paper attempts to provide an alternative to liver biopsies by creating a classification model based on non-invasive markers to predict the 4 stages of liver disease: Healthy, Hepatitis, Fibrosis and Cirrhosis. Using the HCV data set sourced from UCI machine learning repository, employing SMOTE to overcome unbalanced class samples, and variety of model implementations, this study finds that the rate of false negatives can be significantly minimized in prediction by using Multinomial Logistic Regression for the fibrosis class and Radial Kernel SVM for the rest.

1 Background and Motivation

About 3.5 million in the United States have chronic hepatitis C virus (HCV). Yet most people infected with HCV don't know they have it. Over years, HCV infection can cause major damage to the liver. For every 75 to 85 people who have chronic HCV infection, between 5 and 20 of them will develop cirrhosis. HCV infection is the leading cause of cirrhosis and liver cancer (Case-Lo, n.d.).

Even today, liver biopsy represents the "gold standard" for the assessment and quantification of liver fibrosis. However, its invasiveness engenders pain, considerable cost and significant potential complications, leading to poor patient acceptance (Chin et al., 2016). These factors lead to the need of developing a classification algorithm that can identity the level of liver disease progression, especially the latter stages such as cirrhosis, which require immediate care, using non-invasive markers of liver fibrosis. This study attempts to achieve this goal in order to aid medical professionals discern when liver biopsies are an absolute necessity, allowing patients avoid unnecessary invasive and expensive procedures. Results of this study and models can also serve as a "second opinion" to medical professionals as they make their diagnoses. Given these goals, the report aims to maximize prediction performance over interpretation, and especially emphasizes on minimizing the rate of false negatives in all stages of liver disease.

2 Data

2.1 Data Description

The data set was sourced from UCI Machine Repository (Kamal et al., n.d.), and was donated by Medical University Hannover, Helmholtz Centre for Infection Research and Georg Hoffmann in 2019. It contains laboratory tests values and patient demographics for a total of 615 patients; both blood donors and those infected with Hepatitis C virus. The data set has 14 features (as shown in table 1) and 26 patients with at least one laboratory test value missing.

Table 1: Input Features and Definitions

Input Feature	Definition
Patient ID	Patient ID/No.
Age	Age of the patient in years
Sex	Sex of the patient
ALB	Albumin Blood Test; measures the amount of albumin (protein made by liver) in blood
ALP	Alkaline Phosphatase Test; measures the amount of ALP in blood, higher-than-normal (above 140) indicates liver disease (Balingit, n.d.)
ALT	Alanine Transaminase Test; measures the amount of ALT in the blood, high (above 55) ALT indicates liver disease (Daniel, n.d.)
AST	Aspartate Aminotransferas; measures the amount of AST in your blood, high (above 40) AST indicates liver disease (Daniel, n.d.)
BIL	Bilirubin Test; measures the amount of bilirubin in your blood, high (above 1.2 for adults, 1 for under 18) BIL indicates liver disease ("Bilirubin test", n.d.)
CHE	Acetylcholinesterase Test; shows pesticide exposures
CHOL	Cholestrol Test; measures cholesterol and triglycerides in blood, high (above 120 for adults, 170 for under 19) indicates liver disease ("Cholesterol Levels: What You Need to Know", n.d.)
CREA	Creatinine Test; measures creatinine levels in blood and/or urine to assess baseline serum creatinine concentration lower than 35-75 mol/l
GGT	Gamma-Glutamyl Transferase; measures GGT leaks in the bloodstream which occurs when liver is damaged
PROT	Total Protein Test; measures the total amount of two classes of proteins in blood

The motivation of this study is contained within the 'Category' variable and will be set as the response variable for the models. The liver disease stage or response variable is ordinal

and ranges as follows:
0: Blood donor, 1: Hepatitis, 2: Fibrosis, 3: Cirrhosis
Cirrhosis is the worst outcome out of the 4 and implies severe scarring of the liver (Nall, n.d.). Blood Donor implies a healthy patient.

2.2 Data Processing

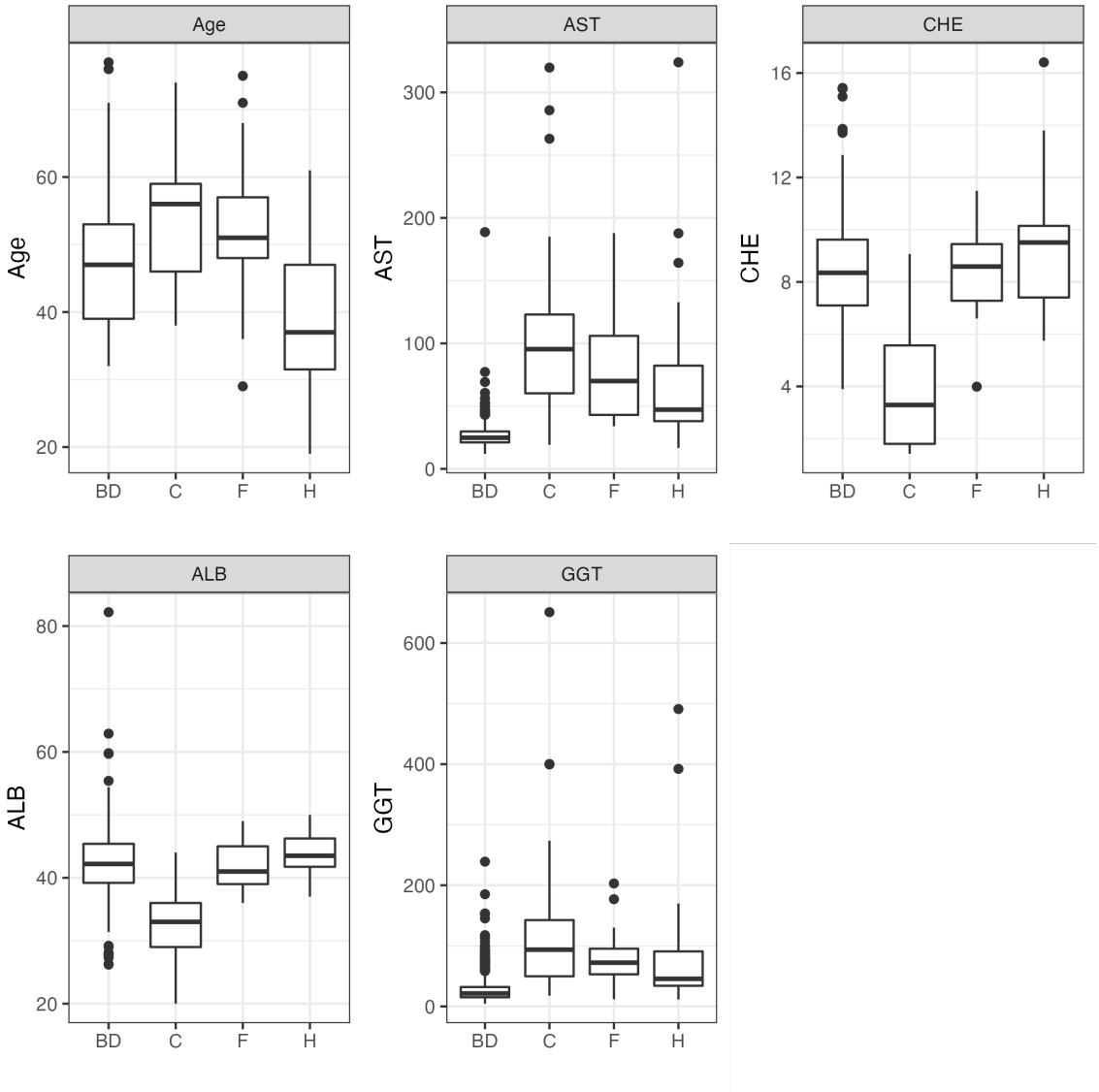
Data was subsetted into training (70%) and test (30%) data sets, which maintained similar proportions per class between the two. There were 19 patients (rows) who had at least one missing input feature value in the training data set and 7 patients (rows) who has at least one missing input feature value in the test data set. Given the small number of observations to begin with, missing input feature values were imputed by first grouping by the HCV category (defined response variable of this problem) and then taking the median. Both training and test data set missing values were imputed separately, using their own respective data sets to avoid any data leakage. Random seed is set to 0 for all models to ensure reproducibility and fair comparisons. Figure 1 also shows the existence of potential outliers, however they are few in number and not extreme in values. Majority of the models employed within this study such as random forest, MLP are also outlier resistant. Given the small size of the data set, outliers are therefore not excluded.

The problem at hand is also heavily unbalanced, where the majority class of Blood Donor accounts for 88% of the data, while the other three minority classes make up the rest 12% all together . In order to balance the data set, SMOTE (Synthetic Minority Oversampling TEchnique) was employed with different sampling strategies applied to each class. SMOTE oversamples from the minority classes by creating synthetic samples from the training data set only. Since cirrhosis represents the worst stage of liver disease progression and there is a high cost associated with false negatives, cirrhosis class was oversampled by SMOTE to equal 300 samples. The second worst stage is fibrosis, and was therefore oversampled to equal 200 and lastly hepatitis was oversampled to equal 100. Blood Donor class wasn't inflated or deflated, and remains at 378 patients in the training data set.

2.3 Exploratory Data Analysis

Since the data set contains 14 attributes, five variables are chosen for EDA based on the importance of variables from random forest model (see appendix figure 3). After selecting five variables, there was only one missing value; thus, that data point was eliminated and the further EDA was proceeded. The selected variables are Age, AST, CHE, ALB, and GGT, which are all numerical variables. Box plots and a pair plot of these variables are shown in figure 1.

Figure 1: Box Plots of the variables against Liver Disease Progression



It is clear to see the significant differences in medians for most variables between cirrhosis and other groups. For some of those variables, such as ALB and CHE, that difference is highlighted further as blood donors, hepatitis, and fibrosis patients share similar medians, while patients with cirrhosis stand out compared to those three groups. For AST and GGT, the median for Blood Donors sits significantly below the others, indicating these variables may be good indicators for distinguishing healthy patients from patients infected with HCV. The 'Age' boxplot indicates that worst cases of liver disease such as cirrhosis and fibrosis are found most in older individuals. Descriptions from table 1 can be visualized here within the data as well, with AST over 40 and high GGT values representing all three stages of liver disease.

Figure 2: Pair plot of the variables against Category

From the scatter plots, we see that blood donors dominates the space and are the most tightly packed cluster for all variables. Cirrhosis ranges widely among most variables, followed by fibrosis and hepatitis. Looking at the correlations, the correlation between Age and GGT for hepatitis, and ALB and CHE for cirrhosis are positively high, which indicate that those variables are highly correlated for those classes. Since each class has different combinations of variables that give high correlations, it provides an intuition for why each class needs to be analyzed separately.

3 Results

3.1 Models Considered

Five models are considered and employed. Ranging from the level of complexity they are:

- Multinomial Logistic Regression
- Random Forest

- Radial Kernel SVM
- Multilayer Perceptron
- Mixed Model Ordinal Classifier

3.1.1 Multinomial Logistic Regression

Multinomial logistic regression is the least complex model employed in this study and by definition is a simple extension of binary logistic regression that allows for more than two categories of the dependent variable. It does not assume normality, linearity, or homoscedasticity, is easy to implement, efficient to train and does not require tuning parameters (Starkweather & Moske, n.d.).

There are 5 types of solvers for the multinomial logistic regression classifier - 'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'. Only algorithms with 'newton-cg' and 'liblinear' solvers converged with not extremely large numbers of iterations. Compared to the model with 'newton-cg' solver, the model with 'liblinear' solver showed higher accuracy. Given that there are only 615 observations in the data set, which can be seen as relatively small, the model with 'liblinear' solver was selected.

3.1.2 Random Forest

Random Forest is one of the most efficient tools for performing classification tasks. It can handle both categorical and numerical variables, performs well in classification for high dimensional data set, and compared to decision trees, improves performance by using only a subset of attributes at each split, therefore decreasing variance while maintaining a low bias (Kho, n.d.). It is also robust to outliers and it is shown by EDA that there are outliers in this data set.

The Random Forest model was tuned using 5-fold cross validation to restrict overfitting. Entropy criterion was used to evaluate models. The hyperparameters of the best model are 'max_depth' = 12 (i.e. the maximum depth of the tree) and 'n_estimators' = 50 (i.e. the number of trees in the forest), and were picked from possible combinations of different values of 'n_estimators' and 'max_depth'.

3.1.3 Radial SVM

Kernel SVM is also one of the preferable options for multiple-class classifications given the moderately number of features. SVMs can handle categorical variables by making them into binary variables. All three SVM kernels (linear, polynomial, radial) were implemented on this 4-class classification. All hyper parameters corresponding to each method were tuned with fairly wide ranges using 5-fold cross validation. Given the high cost associated with a false negative in this study, radial SVM was selected out of the three due to its high recall. The hyper parameters for the best radial SVM are $c = 3.981072$ and $\gamma = 0.0006$. Since the parameter c is quite small, it encourages a larger margin. Also, γ parameter is minuscule, which implies that the shape of the hyperplane boundary is more linear and this model does not fully capture the complex shape of the high dimensional data.

3.1.4 Neural Network: Multilayer Perceptron

Table 5 - 8 indicate that Random Forest and Radial Kernel SVM outperform Multinomial Logistic Regression, especially for the class we prioritize the most; cirrhosis. Given the model performance of these non-parametric methods, a Multilayer Perceptron (MLP) model was also employed to discern if a further increase in flexibility would aid overall performance per class, especially because the response is not a binary case. MLP was chosen over other kinds of neural networks because the data set in hand is tabular and doesn't comprise of a time series or image component.

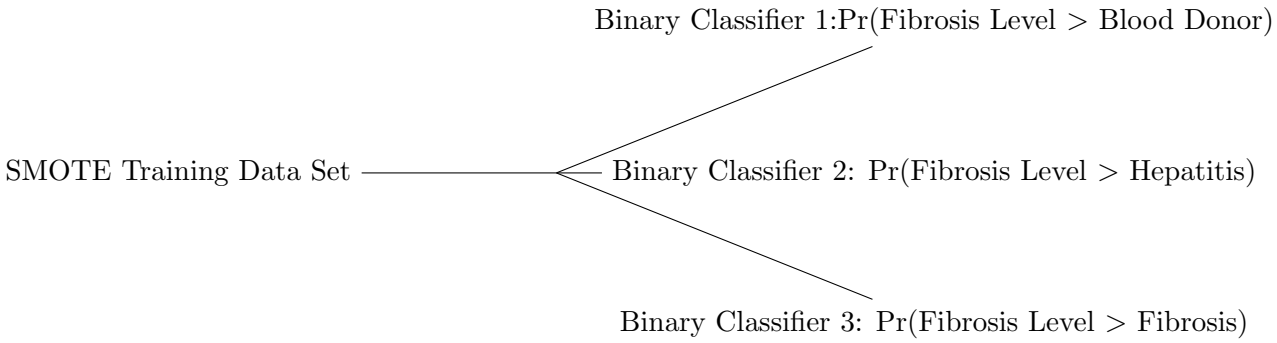
MLP was fine tuned using a 10 fold cross validation and a variety of parameter specifications, which resulted in the best model employing the stochastic gradient descent (SGD) optimizer with a learning rate of 0.01, sparse categorical cross entropy loss function, 32 epochs, batch size of 10 and a 15% dropout layer after the first two fully connected layers. Table 2 lists the details of all the layers.

Table 2: MLP Parameters

Parameter	Layer 1	Layer 2	Output Layer
Units	200	500	4
Activation	ReLU	ReLU	Softmax

3.1.5 Mixed Model Ordinal Classifier

Multiclass classification methods typically ignore the ordinal structure of the response variable and instead treat the outcome variable as nominal, meaning the ordinal information is left unused in prediction. This question of study is an example of such as case, where liver disease progression is in fact an ordinal variable. To introduce ordering information of class attributes in prediction, the training data is transformed from a $4-$class ordinal problem to 3 binary class problems, predicting the following probabilities (Frank & Hall, n.d.):



To predict the class value of a test instance, we need to estimate the probabilities of the k original ordinal classes using our $k - 1$ binary models. Every test instance is processed by each of the $k - 1$ classifiers and the probability of each of the k ordinal class values is calculated (Frank & Hall, n.d.) as follows:

Pr(Y=Blood Donor) = 1−Pr(Fibrosis Level>Blood donor)
Pr(Y=Hepatitis) = Pr(Fibrosis Level > Blood Donor) − Pr(Fibrosis Level > Hepatitis)
Pr(Y=Fibrosis) = Pr(Fibrosis Level>Hepatitis) − Pr(Fibrosis Level>Fibrosis)
Pr(Y=cirrhosis) = Pr(target>Fibrosis)

The class with maximum probability is assigned to the test instance. Since the ordinal classifier method requires only probabilities to call a class label, each binary classifier can be a different algorithm, as long as it yields class probabilities. Given this the following algorithms were used per classifier:

Table 3: Algorithms used within the Ordinal Classifier

Binary Classifier	**Algorithm**
Binary Classifier 1	Random Forest
Binary Classifier 2	MLP
Binary Classifier 3	Radial Kernel SVM

3.2 Model Results

Table 4: Model Accuracies on Test Data

Model	**Accuracy**
Multinomial Logistic Regression	93.5%
Random Forest	92.9%
Radial Kernel SVM	92.4%
MultiLayer Perceptron (MLP)	94.0%
Mixed Model Ordinal Classifier	94.0%

Table 4 shows the models' accuracies on the test data, all of which are upwards of 92%. While model accuracy is important, it may not be the best measure for this unbalanced data set since 88% of the data is blood donors, while only 12% is the true, target liver disease categories. Precision and recall of each class also take priority in this study because there is a high cost associated here with a False Negative. Tables 5 - 8 compare the five models employed with those two performance metrics, alongside their harmonic mean (F1 score) in mind.

Table 5: Model Results-Blood Donor Class

Blood Donor			
	Precision	**Recall**	**F1 Score**
Multinomial Logistic Regression	0.99	0.98	0.98
Random Forest	0.99	0.98	0.98
Radial Kernel SVM	0.99	0.97	0.98
MultiLayer Perceptron (MLP)	0.99	0.99	0.99
Mixed Model Ordinal Classifier	0.99	0.99	0.99

Table 6: Model Results-Hepatitis Class

Hepatitis			
	Precision	**Recall**	**F1 Score**
Multinomial Logistic Regression	0.57	0.57	0.57
Random Forest	0.50	0.57	0.53
Radial Kernel SVM	0.62	0.71	0.66
MultiLayer Perceptron (MLP)	0.63	0.71	0.66
Mixed Model Ordinal Classifier	0.42	0.60	0.50

Table 7: Model Results-Fibrosis Class

Fibrosis			
	Precision	**Recall**	**F1 Score**
Multinomial Logistic Regression	0.40	0.67	0.50
Random Forest	0.43	0.50	0.46
Radial Kernel SVM	0.25	0.17	0.20
MultiLayer Perceptron (MLP)	0.38	0.50	0.43
Mixed Model Ordinal Classifier	0.67	0.44	0.53

Table 8: Model Results-Cirrhosis Class

Cirrhosis			
	Precision	**Recall**	**F1 Score**
Multinomial Logistic Regression	0.83	0.56	0.67
Random Forest	0.67	0.67	0.67
Radial Kernel SVM	0.54	0.78	0.64
MultiLayer Perceptron (MLP)	0.83	0.55	0.66
Mixed Model Ordinal Classifier	0.67	0.67	0.67

The blood donor class shows extremely high precision and recall all throughout. All models seem to do equally well, with marginal differences. The fibrosis class stands out, as all models show a lower performance compared to any other minority class, indicating that the fibrosis is the hardest class to separate. The radial kernel SVM especially performs terribly for fibrosis. The mixed model ordinal classifier performs slightly better than the multinomial logistic

regression for the fibrosis class in the f1 score, but in lieu of minimizing false negatives first and therefore maximizing recall, the multinomial logistic regression ultimately has an edge. Given the supreme performance of the multinomial logistic regression, it is also fair to assume that the boundary separating the fibrosis class is linear in nature.

The radial kernel SVM and MLP perform similarly for the hepatitis class, indicating they correctly predict 71% of the patients who truly have hepatitis and are correct about 63% of the time when they label a patient with hepatitis. Hepatitis is the earliest stage of liver disease and this particular precision metric ought to be maximized in particular because it is important to not start treating a patient who actually doesn't have liver disease simply because the model predicted it.

Cirrhosis is the worst outcome of all the classes, and we see the most balanced performance in terms of both recall and precision from the mixed model ordinal classifier and random forest. It is also important to note, that although the radial kernel svm doesn't achieve a balanced performance, it outputs a 11% better recall. This is again an important metric because cirrhosis implies the worst outcome and therefore its vital the model correctly predicts and informs patients who truly have cirrhosis.

4 Conclusions

In an ideal scenario, a single model would maximize both precision and recall so patients aren't falsely predicted to have liver disease when they aren't infected and vice-versa, they are not predicted to be healthy, when they are actually infected with HCV. In reality different models outperform in different classes. Some complex models such as MLP perform equally well as the radial SVM, but a simpler model is prioritized overall.

This study was devoted to predicting various stages of liver disease progression with an emphasis on minimizing the rate of false negatives. In employing SMOTE to overcome the unbalanced class problem and various model implementations to maximize recall performance while maintaining model simplicity, this report recommends using a combination of two models for prediction of liver disease: Radial Kernel SVM for hepatitis, cirrhosis and blood donor class, and Multinomial Logistic Regression for the fibrosis class.

5 Limitations

It can be seen from the results that almost all models cannot predict Fibrosis as well as other classes even with the application of SMOTE. This is due to the heavy class imbalance of the original data set with only 15 observations available in this class for training. This limited the classifiers significantly, proving their incapacity to learn the patterns and relationships between features and the response variable.

A limited number of hyperparameters were selected in the tuning process for the models such as random forest, radial kernel SVM and MLP and by extension the mixed model ordinal classifier due to computational time and cost. Thus, it is possible that the hyperparameters of the "best model" that produces the smallest test errors are absent from the hyperparameter

tuning arguments. In addition to extending the range of hyperparameters tested, boosting can also be employed for random forest to check for further improvement in predictions.

The other limitation would be the presence of outliers. SVMs are not necessarily robust to outliers. Although the outliers are limited and not extreme in this data set,the performance may be improved by eliminating outliers.

The MLP model was also run without setting the seed at every step in R's Keras API, given the complex implementation. Although multiple runs of the MLP produce similar results, making model results trustworthy, methods should be employed to get exact reproducibility.

6 Team Contributions

6.1 Shreya Mittal

Shreya contributed the set up of the problem such as the background, motivation and data preprocessing along with the MLP and the Mixed Model Ordinal Classifier. She was responsible for tuning those models in order to find the best performing one, along with listing their limitations and next steps.

6.2 Yukiko Yamazaki

Yukiko contributed the EDA and implementation of the Radial Kernel SVM. She was also responsible for exploring all the SVM models available, hyperparameter tuning for those models and the limitations which come with SVM.

6.3 Mingcan Zhao

Mingcan contributed Random Forest and Multinomial Logistic Regression of the report. She was also responsible for exploring all modifications of those two models, tuning their hyperparameters and listing their limitations.

6.4 Collaboration

All team members collaborated to find the best interpretation, conclusion and limitations of this study.

References

Balingit, A. (n.d.). *Alkaline phosphatase level (alp) test*. https://www.healthline.com/health/alp (accessed: 04.20.2022)

Bilirubin test. (n.d.). https://www.mayoclinic.org/tests-procedures/bilirubin/about/pac-20393041 (accessed: 04.20.2022)

Case-Lo, C. (n.d.). *Cirrhosis and hepatitis c: Their connection, prognosis, and more*. https://www.healthline.com/health/cirrhosis-and-hepatitis-c (accessed: 04.19.2022)

Chin, J. L., Pavlides, M., Moolla, A., & Ryan, J. D. (2016). Non-invasive markers of liver fibrosis: Adjuncts or alternatives to liver biopsy? *Frontiers in Pharmacology*, *7*. https://doi.org/10.3389/fphar.2016.00159

Cholesterol levels: What you need to know. (n.d.). https://medlineplus.gov/cholesterollevelswhatyouneedtoknow.html (accessed: 04.20.2022)

Daniel, C. (n.d.). *Overview of alt and ast liver enzymes*. https://www.verywellhealth.com/liver-enzymes-1759916 (accessed: 04.20.2022)

Frank, E., & Hall, M. (n.d.). *A simple approach to ordinal classification* (Report). University of Waikato.

Kamal, S., ElBahnasy, K. A., ElEleimy, M. H., Hegazy, D., & Nasr, M. (n.d.). UCI machine learning repository. http://archive.ics.uci.edu/ml (accessed: 04.19.2022)

Kho, J. (n.d.). *Why random forest is my favorite machine learning model*. https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706 (accessed: 04.20.2022)

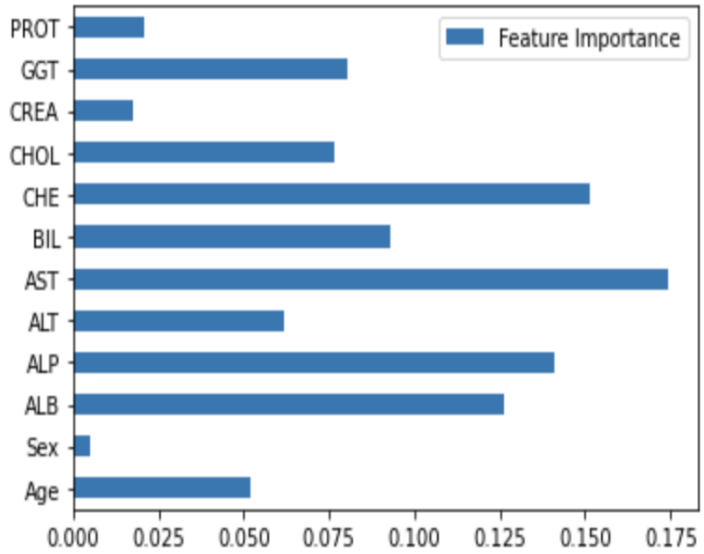
Nall, R. (n.d.). *Liver fibrosis*. https://www.healthline.com/health/liver-fibrosis (accessed: 04.19.2022)

Starkweather, D. J., & Moske, D. A. K. (n.d.). *Multinomial logistic regression*. https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf (accessed: 04.20.2022)

7 Appendix

7.1 Additional Infographics

Figure 3: Random Forest-Feature Importance



7.2 Code Repository

All the code regarding this report can be found on the following Github Repository:

Repository Name: HCV-Classification
Repository Location: https://github.com/umichshreyaM/HCV-Classification.git