Predictive Modeling of HCV Categories using R: Exploring Feature Importance and Model Performance

STAT 515

Aditya Baxi, Ivan Francis, Jainam Jagani

Under the guidance of Prof. Isuru Dassanayake

I. Abstract

The high incidence of Hepatitis C (HCV) and potential long-term health effects make research into the disease of utmost relevance. To stop future transmission and provide timely treatment measures, early detection and diagnosis of HCV infection are essential. With a focus on determining the distribution of various Hepatitis C stages and developing precise models to pinpoint those at a higher risk of HCV infection based on known factors, this study intends to investigate the predictive modeling of HCV categories using the R programming language.

To accomplish this goal, data will be analyzed using statistical models and methods, and the links between numerous parameters connected to HCV infection will be investigated. The study will investigate the relevance of feature importance in foretelling the various phases of hepatitis C, offering insights into the major factors that affect the risk and development of the illness. Additionally, the effectiveness of the prediction models will be assessed in order to determine their precision and dependability in identifying people who are more likely to contract HCV.

This study has a variety of advantages. First off, by examining the distribution of different HCV phases, we may better understand how the disease develops and pinpoint crucial times for therapy and intervention. Chronic liver disease and its effects can be greatly diminished with early HCV infection identification. Second, by creating precise predictive models, healthcare providers can proactively identify those who are more likely to contract HCV, allowing for more focused screening and prevention methods. By concentrating efforts on those who are most vulnerable to HCV infection, this research can help to improve healthcare outcomes and resource allocation.

In conclusion, this study attempts to increase our knowledge of HCV infection and its distribution across several stages by utilizing statistical modeling and predictive methodologies. This work has the potential to enhance early detection, intervention, and prevention measures, resulting in better health outcomes for people with hepatitis C. This is accomplished by identifying significant traits and developing precise predictive models.

II. Introduction

With millions of victims worldwide, the Hepatitis C virus (HCV) infection causes a huge global health burden. The negative effects of HCV must be minimized through early detection and treatment. The analysis of our HCV dataset is the main goal of this research study, which focuses

on using predictive modeling methods. In order to estimate the disease stage (Category) based on these variables, the dataset includes the clinical and laboratory test results of 615 individuals who have been diagnosed with HCV in which there are individuals that consist of Anti-HCV Anti-Bodies.

Through statistical and predictive modelling, we were able to answer these three research questions:

a.) Does the combination of various health indicators (such as liver enzymes, albumin levels, bilirubin levels, cholesterol levels, creatinine levels, gamma-glutamyl transferase levels, and protein levels) have a significant impact on predicting the severity of Hepatitis C infection, classified into multiple categories (e.g. Hepatitis C, Fibrosis, Cirrhosis)?
b.) Can we build predictive models to identify individuals at a higher risk of HCV infection based on the available variables? How accurate is the model in predicting HCV infection and which is the best performing one?
c.) Does the gender (male and female) have a statistically significant relationship with the occurrence of Hepatitis C?

The initial goal of this study is to look at the distribution of clinical outcomes linked to the liver, including cirrhosis, fibrosis, and hepatitis C, among those with HCV infection in the dataset. Understanding the prevalence and distribution of various disease stages can provide important insights into the development and severity of HCV. Healthcare practitioners and politicians can use this information to build focused intervention strategies and allocate resources more efficiently. The second goal is to create a predictive model that, using the information at hand, can pinpoint people who are more likely to contract HCV. The model seeks to properly predict the chance of HCV infection by examining numerous clinical and laboratory parameters, such as age, gender, serum albumin level, liver enzyme levels, and others. Our third goal is to statistically observe if the “Gender” (Male and Female) are holding any significance on relation with Hepatitis C.

The findings of this study have the potential to increase patient outcomes, direct future research, and advance public health activities in the area of HCV infection. They may also improve HCV management practices.

III. Data Source and Methodology

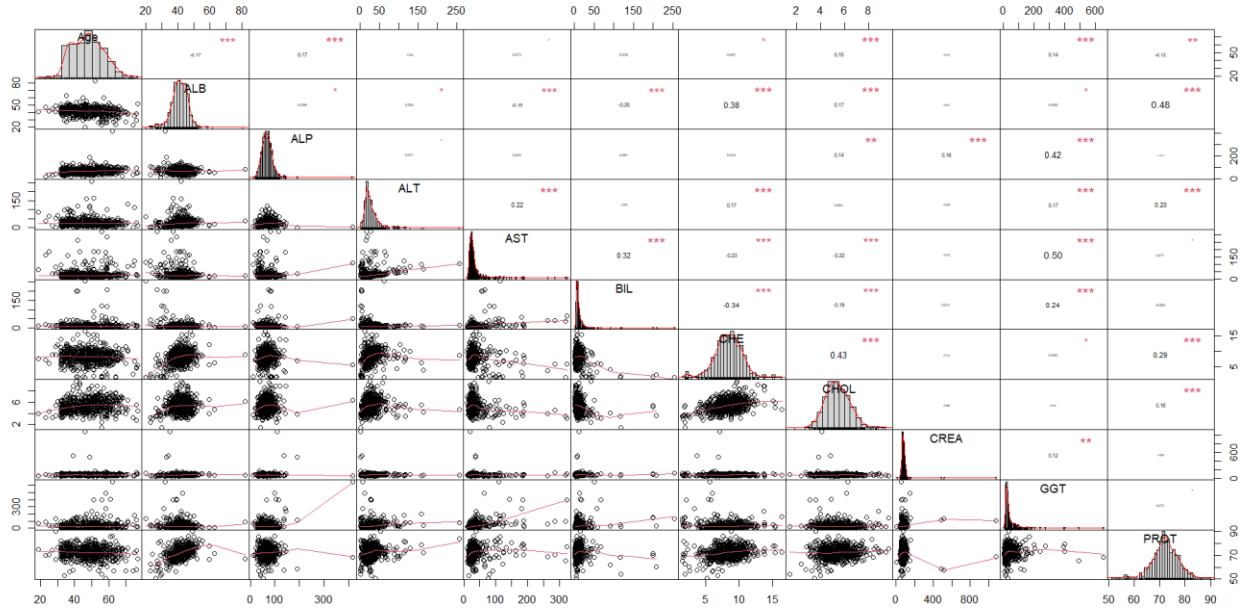
The HCV dataset from the UCI machine learning repository was the one used in this research study. The dataset includes information on the demographic, clinical, and laboratory characteristics of HCV-infected patients. The dataset consists of 615 HCV-positive patients and 13 characteristics. The multinomial logistic Classification model, the random forest, and the KNN classification are the statistical models employed in this research paper. Based on the features found in the dataset, these models are helpful in predicting the types of HCV. Random forest creates a final prediction by combining several different decision trees. It can handle categorical and continuous variables and can offer a measure of feature relevance, allowing us to pinpoint the most crucial traits for

predicting HCV groups. This makes it beneficial in this situation. Several categorical variables' associations are shown through multinomial logistic Classification because it can handle several categories and give a gauge of the degree of correlation between the predictor factors and the response variable, it is helpful in this situation. A non-parametric method for classification and prediction tasks is KNN classification. The fact that it can handle categorical and continuous variables, as well as provide a measure of similarity across various observations, makes it valuable in this situation because it enables us to determine which patients are the most similar based on the features found in the dataset. These models were chosen because they can handle categorical and continuous variables, quantify the significance and relationship of characteristics, and can recognize patients who are similar based on attributes found in the dataset.

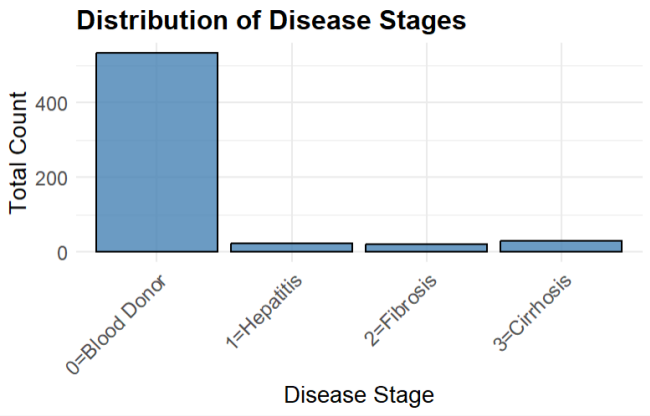
IV. Analysis and Results.

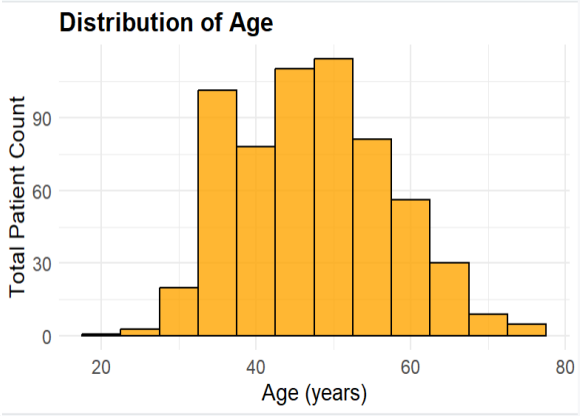
Basic Understanding of the dataset via Visualization and Summary Statistics.

When we display the correlation plot of the dataset, we see that there are not many important or strong correlation between the predictors.

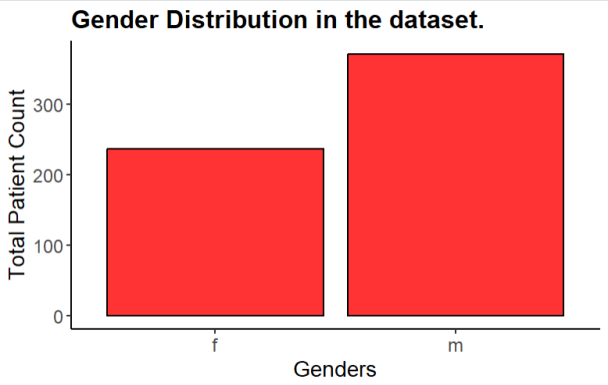


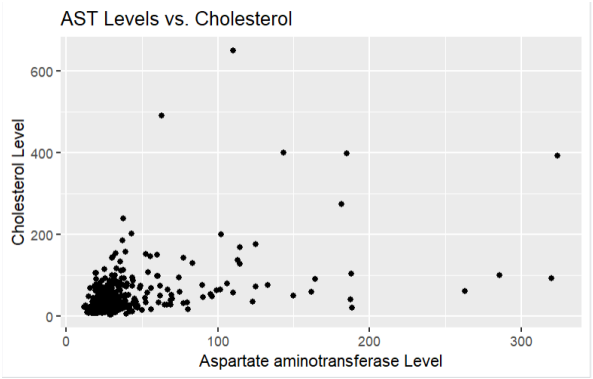
The only visible relation we can see that is prominent is between the GGT and AST levels in the dataset. Which means that multicollinearity is not present in the dataset.





In our dataset, we see that the category for Blood Donors which is 0 are the observations that are present in huge quantity compared to the positive disease stages. Whereas, from the age distribution graph we can see that the average age range of the patients in the dataset lies around 45 – 50 years.





One of the research questions that we are interested in answering is whether there are any difference with predictions based on gender while predicting the risk of hepatitis in Individuals. Understanding the gender distribution in the dataset for the former reason is very crucial. We can see that there are more male patients in the dataset than female patients in the dataset. Also based on the correlation plot we can see that the AST vs CHOL graph is a sample relation graph to affirm that the highest correlation in the dataset among predictors isn't strong which is what we observe from the above infographic.

Multinomial Logistic Classification for Category Variable:

```
> summary(model)
Call:
multinom(formula = Category ~ ., data = train)

Coefficients:
          (Intercept)   Age30-39  Age40-49   Age50-59  Age60-69       Sexm       ALB       ALP       ALT       AST       BIL
1=Hepatitis   -22500.05 -17836.174 -15489.69 -17025.657 -14188.72   211.6321  360.3287 -450.0639 -165.2416  40.55353 -113.84265
2=Fibrosis    -12994.61 -15572.418 -16963.07 -14732.725 -13301.96 -1248.2306  459.5674 -290.6506 -319.5797  87.35575  109.58033
3=Cirrhosis   -14499.95   6485.518   1187.70   8885.837  11482.87 -3670.1471 -494.2684  222.9702 -472.6127 221.57706   54.40814
                 CHE      CHOL       CREA        GGT     PROT
1=Hepatitis  1452.404 -1429.371 -143.18564 105.105050 603.8309
2=Fibrosis   1197.655 -2368.220 -182.82926  82.481092 442.8374
3=Cirrhosis -3593.127 -1507.609   18.21047  -9.366893 483.2935

Std. Errors:
          (Intercept) Age30-39     Age40-49 Age50-59 Age60-69     Sexm      ALB      ALP      ALT      AST      BIL      CHE
1=Hepatitis    20.18154 102.1683 1.678469e-09 407.8223 266.0053 615.1756 149.7256 76.89999 19.13705 7.954992 94.59795 506.1097
2=Fibrosis     20.18154 102.1683 1.153828e-27 407.8223 266.0053 615.1756 149.7256 76.89999 19.13705 7.954992 94.59795 506.1097
3=Cirrhosis     0.00000   0.0000 0.000000e+00   0.0000   0.0000   0.0000   0.0000  0.00000  0.00000 0.000000  0.00000   0.0000
              CHOL     CREA     GGT     PROT
1=Hepatitis 1741.416 43.18282 20.3923 173.2444
2=Fibrosis  1741.416 43.18282 20.3923 173.2444
3=Cirrhosis    0.000  0.00000  0.0000   0.0000

Residual Deviance: 9.38353e-05
AIC: 96.00009
 > cm = table(test$Category, model1)
 > print(cm)
                model1
                 0=Blood Donor 1=Hepatitis 2=Fibrosis 3=Cirrhosis
   0=Blood Donor           104           0          0           1
   1=Hepatitis               0           2          1           0
   2=Fibrosis                0           2          0           0
   3=Cirrhosis               0           0          3           2
 > accuracy = sum(diag(cm))/sum(cm)
 > print(accuracy)
 [1] 0.9391304
 >
```

In this study, a multinomial logistic Classification model was fitted to investigate and forecast the hepatitis, fibrosis, and cirrhosis disease stage categories among patients with HCV diagnoses. A dataset of 615 patients' clinical and laboratory test results was used by the model. The investigation sought to comprehend the association between several predictor variables, including age, gender, the patient's disease stage, and the blood levels of several biomarkers. The fitted multinomial logistic Classification model's results shed light on how each predictor variable affected the log-odds of falling into a certain illness stage group. These coefficients and the associated standard errors were used to evaluate the predictors' statistical significance. Based on the residual deviance, which showed a tight fit to the observed data, the model's goodness of fit was assessed.

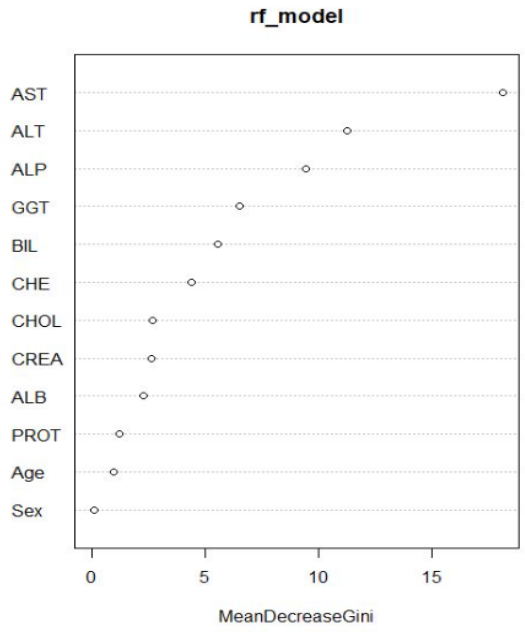
A confusion matrix, which showed the observed and anticipated disease stage categories, was used to present the model's findings. The percentage of examples that were correctly classified served as a measure of the model's accuracy. The model's ability to predict the disease phases based on the provided predictor variables was demonstrated by the achieved accuracy value which is approximately 93%.

In general, the multinomial logistic Classification analysis offered important new understandings of the variables connected to various illness phases in HCV patients.

Random forest model:

```
> predicted_quality = predict(rf_model, newdata = test)
> cm = table(test$hcv, predicted_quality)
> print(cm)
   predicted_quality
      0   1
  0 154   0
  1   5  11
> accuracy = sum(diag(cm))/sum(cm)
> print(accuracy)
[1] 0.9705882
>
```

```
> summary(rf_model)
                Length Class  Mode
call               4   -none- call
type               1   -none- character
predicted        403   factor numeric
err.rate        1500   -none- numeric
confusion          6   -none- numeric
votes            806   matrix numeric
oob.times        403   -none- numeric
classes            2   -none- character
importance        12   -none- numeric
importanceSD       0   -none- NULL
localImportance    0   -none- NULL
proximity          0   -none- NULL
ntree              1   -none- numeric
mtry               1   -none- numeric
forest            14   -none- list
y                403   factor numeric
test               0   -none- NULL
inbag              0   -none- NULL
terms              3   terms  call
```



To investigate and forecast the disease stage categories (Hepatitis, Fibrosis, and Cirrhosis) among individuals with Hepatitis C Virus (HCV) diagnoses, a Random Forest model was fitted. A group of decision trees are combined in the Random Forest algorithm, a machine learning method, to create predictions. A dataset of clinical and laboratory test results from 407 patients was used by the model. Several outputs from the fitted Random Forest model aided in the analysis and interpretation. The significance of each predictor variable in predicting the illness stage was evaluated using the "importance" function. The "MeanDecreaseGini" values represented each variable's relative importance, with larger values indicating a variable's significance. This knowledge allowed for the identification of the main factors influencing the classification of sickness stages in HCV patients.

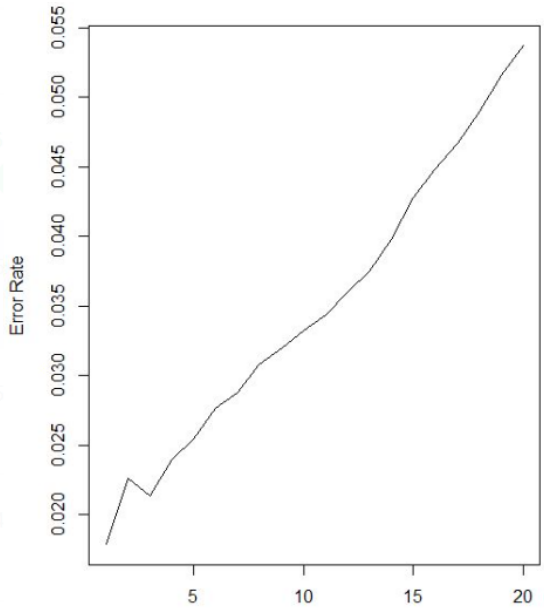
A variable significance plot was created by the "varImpPlot" function to illustrate the relative weight of each predictor variable. The proportional contributions of several variables in predicting the disease stage categories were clearly visualized in this graphic. On the testing set, predictions

were made in order to assess the performance of the Random Forest model. Based on the model and the testing results, projected illness stage categories were created using the "predict" function. The projected illness stages and the actual disease stages found in the testing data were then compared using a confusion matrix. The Random Forest model was effective at predicting the illness phases of HCV patients, as evidenced by the accuracy of the model, measured as the proportion of correctly categorized cases.

Overall, based on the findings of clinical and laboratory tests, the Random Forest model was found to be a useful tool for predicting the disease stage categories of HCV patients. This understanding can help medical practitioners make knowledgeable choices and create individualized treatment programs for HCV-infected individuals.

K-Nearest Neighbors model:

```
> confusionMatrix(hcv_test$hcv, hcv_pred)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 161   0
         1   4   9

               Accuracy : 0.977
                 95% CI : (0.9422, 0.9937)
    No Information Rate : 0.9483
    P-Value [Acc > NIR] : 0.0506

                  Kappa : 0.8063

 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 0.9758
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.6923
             Prevalence : 0.9483
         Detection Rate : 0.9253
   Detection Prevalence : 0.9253
      Balanced Accuracy : 0.9879

       'Positive' Class : 0
```



In this study, a K-Nearest Neighbors (KNN) model was used to analyze and forecast the hepatitis, fibrosis, and cirrhosis disease stage categories among individuals with the Hepatitis C Virus (HCV). A new data point is given a class label by the KNN algorithm, a non-parametric classification method, based on the classes of its closest neighbors. Several performance indicators, including accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), were used to assess the KNN model. These measures shed light on the model's overall performance as well as its capacity to categorize the illness stage categories accurately.

The KNN model's accuracy was determined to be 0.977, meaning that it correctly identified 97.7% of the cases in the test data set. The true accuracy of the model is likely to lie within this range, according to the 95% confidence interval (CI), which varied from 0.9422 to 0.9937 for accuracy. The accuracy that could be attained by always predicting the majority class is known as the no information rate (NIR). The NIR in this instance was 0.9483, showing that the KNN model

performed better than the baseline accuracy. The estimated value of the Kappa statistic, which assesses the degree of agreement between predicted and actual classifications that is not due to chance, is 0.8063. An improved agreement between the model's predictions and the actual disease stage categories is indicated by a higher Kappa score. From the above error rate plot we can observe that for K = 3 we see the first and deepest dip in the error rate due to which hyperparameter K = 3 is selected while training the model.

For each illness stage category, additional measures including sensitivity, specificity, and balanced accuracy offer insight into the model's performance. The model's ability to correctly detect positive instances (disease stages) in the testing data is shown by its sensitivity of 0.9758. The model appears to accurately identify negative cases (stages of non-disease) based on the specificity of 1.0000. The proportion of accurately predicted positive instances among all expected positive cases is represented by the positive predictive value (PPV) of 1.0000. The evaluation of the KNN model's analysis and output assists in determining the model's effectiveness and its capacity to correctly categorize HCV patients' illness stages. In clinical contexts, these insights might be useful for making well-informed decisions about illness management, therapy design, and patient prognosis.

Multinomial Logistic Classification:

```
> print(cm2)
   model6
      0   1
  0 152   3
  1   1  13
> accuracy2 = sum(diag(cm2))/sum(cm2)
> print(accuracy2)
[1] 0.9763314
```

To forecast the disease stage categories (Hepatitis, Fibrosis, and Cirrhosis) based on the presented predictor variables we run a Multinomial Logistic Classification model. The coefficient values show the estimated impacts of each predictor variable, in relation to a reference category, on the log-odds of falling into a certain disease stage category. The baseline log-odds of falling into the reference category are represented by the intercept term in this model. To consider their individual effects on the log-odds of falling into each illness stage group, the other predictor factors are also taken into consideration.

The coefficient estimations' level of uncertainty is indicated by the standard errors linked to the coefficients. Greater accuracy in predicting the coefficients is indicated by smaller standard errors. You can evaluate the statistical significance of the correlations between the predictors and the disease stage categories by comparing the magnitude of each coefficient to its standard error. AIC (Akaike Information Criterion) and residual deviation are indicators of model fit. The residual deviation (RD) measures how well the model fits the data overall, with lower values suggesting a better fit. Lower values indicate a better balance between fit and complexity, and the AIC is a measure of the model's quality while taking model complexity into account.

The accuracy of the model, measured as the percentage of correctly classified examples, is 0.976, which demonstrates a good level of prediction accuracy. We understand the connections between the predictor variables and the disease stage categories by examining the coefficient values, standard errors, model fit metrics, and prediction accuracy. This data aids in comprehending the variables affecting the disease stage categorization and in evaluating the model's accuracy in assigning individuals to their appropriate illness stages. The results of the Multinomial Logistic Classification model's analysis and output, taken together, offer important new understandings of the relationships between the predictor variables and the various illness stage classifications.

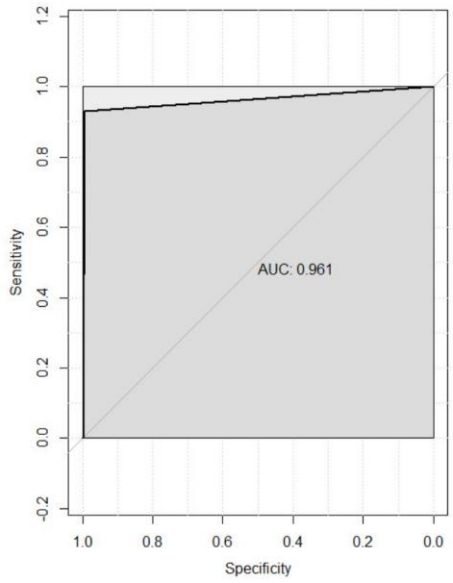
ROC-AUC Curve

Receiver Operating Characteristic - Area Under the Curve is referred to as ROC-AUC. It is a statistic used to assess how well binary classification models perform. The true positive rate (sensitivity) and the false positive rate (1 - specificity) for various categorization thresholds are shown graphically by the ROC curve. The area under the ROC curve is represented by the AUC. It has a value between 0 and 1, with a higher number indicating greater model performance. While an AUC of 1 denotes a flawless classifier, one of 0.5 suggests guessing at random.

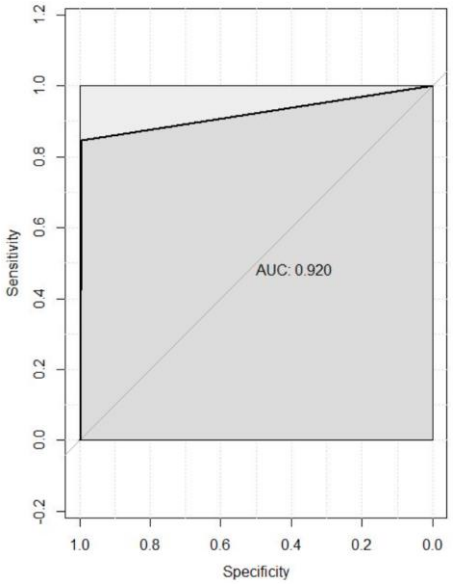
For our dataset we have mainly used ROC-AUC for the following purposes:

- Performance evaluation: ROC-AUC offers a thorough evaluation of the model's effectiveness at all conceivable classification criteria. It concurrently considers sensitivity (the capacity to categorize positive cases properly) and specificity (the capacity to classify negative situations accurately).
- Imbalanced Datasets: As our dataset was imbalanced, ROC-AUC is resistant to these types of datasets and offers a reasonable assessment of the model's performance.
- Model Comparison: For this dataset we have run several models and ROC-AUC makes it simple to compare. Better discrimination performance is shown by a larger AUC, which enables simple model selection.

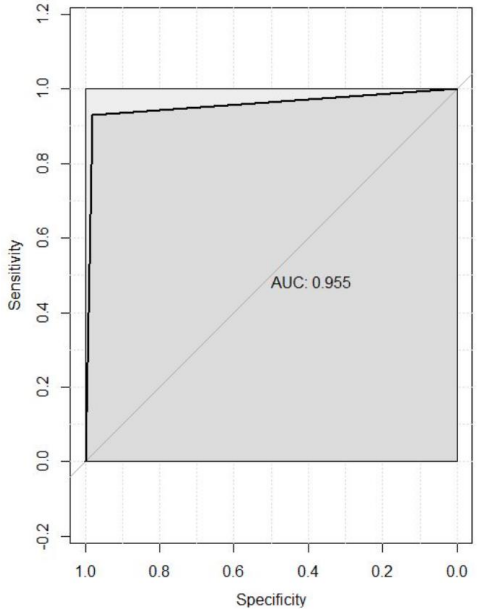
Below, we have plotted ROC-AUC Curves to compare the model accuracy and find which model performs best.



Random Forest



KNN model



Multinomial Logistic Model

Based on the comparison of three models—Random Forest, KNN, and Multinomial Logistic Classification—the Random Forest model, with an AUC of 0.961, was shown to perform the best. This shows that Random Forest outperformed KNN (AUC = 0.920) and Multinomial Logistic Classification (AUC = 0.955) in terms of discrimination power and predictive accuracy. The reason Random Forest fared better than the other models are complex. Random Forest is better equipped

to handle non-linear and high-dimensional data than KNN, which is dependent on the distance-based similarity metric. In comparison to KNN, which can be more sensitive to outliers and noise due to its reliance on nearest neighbors, these techniques assist limit the impact of noisy or irrelevant information and offer higher generalization performance. Additionally, Random Forest offers a feature importance estimate that can be helpful for comprehending how various factors contribute to the classification task. By giving insights into the underlying causes influencing the predictions, this knowledge can help in feature selection and interpretation.

The ROC-AUC curve offers a few benefits, in many real-world datasets, class imbalance is a typical occurrence. This model is resistant to this. The model's performance is secondly represented visually, making comparison, and understanding simply. Third, a larger AUC value suggests greater discrimination, and the AUC value itself offers a numerical evaluation of model performance.

V. Limitations and Future Scope:

Analysis of the HCV dataset has the following drawbacks:

- Limited dataset: The amount and scope of the dataset utilized to conduct the study may be restrictive. A bigger and more varied dataset might offer more reliable insights.
- Variables that are missing from the dataset could have an impact on the likelihood that someone will contract hepatitis C. The analysis might benefit from including other variables, such as genetic traits or dietary habits.
- Data quality: The reliability of the analysis may be impacted by the dataset's accuracy and completeness. Techniques for cleaning and validating data should be used to fix any discrepancies or inaccuracies.
- Model assumptions: The models employed in the analysis, such as the random forest, KNN, and multinomial logistic regression, are based on several hypotheses that may or may not be true for the given dataset. It is crucial to determine whether these presumptions are valid.

Future research to enhance the examination of the HCV dataset:

- Data collection: A more thorough understanding of the disease would be possible with the acquisition of a larger and more complete dataset that included a wider variety of Hepatitis C-related characteristics.
- Exploring and developing new derived features or transforming already-existing variables could help the models better forecast the future by capturing underlying patterns.
- Advanced modeling approaches: Experimenting with advanced machine learning algorithms or ensemble methods may help the predictive models become more accurate and robust.

- External validation: Using an external dataset or conducting prospective research to validate the analysis' conclusions would increase the reliability and generalizability of the findings.
- Ethical considerations: To ensure data privacy and acquire informed consent, among other ethical issues, future data collection and analysis must take them into account.

VI. Conclusions

In conclusion, the study of the HCV dataset yielded important knowledge about the aspects of Hepatitis C. To comprehend the prevalence of the disease and its associations with various factors, a few predictive models were used, including random forest, KNN, and multinomial logistic regression. With an AUC of 0.961, the random forest model performed the best, demonstrating its superiority in the prediction of hepatitis C. The investigation also looked at the importance of gender in relation to hepatitis C, emphasizing how it can affect the condition. Future research should concentrate on overcoming these constraints by utilizing larger and more varied datasets, including extra variables, and applying advanced modeling techniques to further increase our understanding of Hepatitis C. Overall, this analysis lays the groundwork for further study and emphasizes the significance of thorough data analysis in tackling the difficulties provided by hepatitis C.

References

[1] “UCI Machine Learning Repository: HCV data Data Set.” https://archive.ics.uci.edu/ml/datasets/HCV+data

[2] H. Harshi, “Understanding K-Nearest Neighbour Algorithm in Detail,” *Medium*, Jan. 06, 2022. [Online]. Available: https://medium.com/analytics-vidhya/understanding-k-nearest-neighbour-algorithm-in-detail-fc9649c1d196