**Prediction of Baseline Histological Staging of Patients Infected by Hepatitis C Virus using Machine Learning Algorithms**

**A Project Presented to**

##### The Faculty of the Department of Industrial and Systems Engineering

**San Jose State University**

**In Partial Fulfillment of the Requirements**

**For the degree Master of Science in Engineering Management**

**By**

**Naman Ashar**

**April 2020**

**SAN JOSE STATE UNIVERSITY**

**The undersigned Project Committee Approves the Project Titled**

**"Prediction of Baseline Histological Staging of Patients Infected by Hepatitis C Virus using Machine Learning Algorithms."**

**By**

**Naman Ashar**

**APPROVED FOR THE DEPARTMENT OF INDUSTRIAL & SYSTEMS ENGINEERING**

**Professor Niranjani Patel, Department of Industrial and Systems Engineering**

TABLE OF CONTENTS

# Introduction…………………………………………..1

# Project Introduction……………………...............1

# Need for the Study……………………………….1

# The Objective of the Study………………………2

# Scope of the Study……………………………….3

# Limitations of the Study…………………………4

# Literature Review……………………………….…….5

# Terminologies and Packages…………….…...……….7

# Solution Methodology…………………..………..…..12

#### Data Analysis and Interpretation…………….……….15

# Data Acquisition………………………..………...15

# Exploratory Analysis…………………...………...18

# Data Cleaning and Shaping…..…...……………....26

# Model Construction and Evaluation….….……….33

# Interpretation and Results of the Models Created..41

# Forecasting…………………………………………...46

# Conclusion…………………………………………...47

# References……………………………………………50

# Appendix……………………………………………..53

# LIST OF FIGURES:

Figure 1:Reading the Dataset

Figure 2: Summary of Dataset

Figure 3: str function for the HCV Dataset

Figure 4: Histogram for Age

Figure 5: Histogram for WBC Count

Figure 6: Histogram for Aspartate Transaminase Ratio

Figure 7: Histogram for Stages of Fibrosis

Figure 8: Histogram for Body Mass Index

Figure 9: Collinearity Plot to check for Biological Significance

Figure 10: Correlation Plot

Figure 11: Detection of Outliers With Code Snippet

Figure 12: Checking for Normality of Data

Figure 13: SVM Model Construction and Evaluation With Snippets

Figure 14: Confusion Matrix of SVM Model

Figure 15: Confusion Matrix of ctree Model with Code Snippet

Figure 16: Confusion Matrix of Naïve Bayes Model

Figure 17: Confusion Matrix of kNN Model

Figure 18: New Feature Creation Code Snippet

Figure 19: Hold Out Method ( Dividing Data into Training and Testing)

Figure 20: Naïve Bayes Model With Code Snippet

Figure 21: Confusion Matrix for Naïve Bayes Model

Figure 22: Confusion Matrix for kNN Model

Figure 23: Confusion Matrix for SVM Model

###### Figure 24: Confusion Matrix for ctree Model

Figure 25: Neural Network and Accuracy

Figure 26: Confusion Matrix for Logistic Regression Model

Figure 27: Obtaining Statistically Significant Columns through Backtracking

Figure 28: Summary for Boosting Method

Figure 29: Boosting Method Plot

Figure 30: Bagging via treebag and rf

Figure 31: Bagging Method Plot

###### Figure 32: Forecasting ALT until Week 96

Figure 33: Conclusion

## ABSTRACT

Hepatitis C virus is known to affect more than 170 million people worldwide. It is prevalent in the Mediterranean region, and a very high incidence is seen in Egypt, with over 13-15% of the population being affected annually. High invasiveness and expensive costs of medical procedures are the two key factors responsible for patients opting out of HCV testing, which can lead to the progression of the disease from mild fibrosis to liver cirrhosis. Therefore, a non-invasive and effective method needs to be in place, which helps doctors determine the progress and histological staging based on factors which hold biological significance in terms of the disease. For this purpose, data mining and analysis are carried out using the cross-industry process for data mining (CRISP-DM) framework, which is the standard for data mining practices in the industry.

UCI ML Repository dataset on HCV for Egyptian Patients Dataset was used for analysis and various machine learning algorithms. Algorithms like kNN, Naive Bayes, SVM, Decision Tree, and Neural Network were applied using both the whole Dataset and hold-out method, which involves dividing the Dataset into training and validation dataset. Factors that were found to be statistically significant through Logistic Regression included Body Mass Index (BMI), Epigastric pain, RNA Elongation Factor (ENA EF), and were used for prediction of baseline histological staging. It was found out that kNN is well suited for the Dataset with an accuracy of over 97% in predicting the baseline histological stage of the patient as mild (which includes level or scores 1 and 2) or severe (which provides for levels or score 3 and 4). Applying kNN over the Dataset makes for a great model for prediction because of its high accuracy, p-value consideration, and Kappa statistics. Furthermore, there is scope for deployment of this particular model using an R Shiny dashboard. The above model can act as a breakthrough in countries like Egypt, where liver biopsies are expensive and highly invasive, keeping in mind the incidence of the disease.

1. Introduction

1.1 Project Introduction

HCV or the Hepatitis C virus is known to infect the liver.

It is estimated that globally, 71 million people have chronic hepatitis C infection. It is known to affect 2.3% of the population in the Eastern Mediterranean region in 2015. WHO also estimated that in 2016 399,000 people died from hepatitis C, which was mostly due to cirrhosis and hepatocellular carcinoma. HCV is responsible for causing acute and chronic hepatitis in patients who can range from having a mild illness which is prevalent for a few weeks to a serious lifelong illness. It is regarded as the primary cause of liver cancer. HCV, being a blood-borne virus, can infect through the exposure of blood, which can happen due to injection drug use, unsafe injection practices, unsafe health care, transfusion of unscreened blood and blood products, and through sexual practices that lead to exposure to the blood. Multiple strains of the virus are known to exist, and their distribution varies by region. There is a difference seen in strains found in the Mediterranean region and the European region even though the incidence in the European area is far lower than compared to the Mediterranean region[1].

1.2 Need for study

HCV infections are usually asymptomatic. There is no effective vaccine against hepatitis C. Antiviral medicines can cure more than 95% of infections, thereby reducing the risk of death from cirrhosis. Taking into consideration the current healthcare system in place, and day-to-day medical advancements, access to diagnosis and treatment is shallow [1].

Due to the asymptomatic nature of the virus, it is difficult to ascertain the infection for decades after the symptoms appear, leading to severe liver damage. The testing majorly involves two steps- a) Testing for anti-HCV antibodies with a serological test.

b) A nucleic acid test includes testing for HCV ribonucleic acid for confirmation.

Another technique includes testing liver tissue through biopsy. But this technique is considered highly invasive and is an expensive procedure. Further, it adds to the complications a patient might be already suffering from and is inadvisable for patients above the age of 45. Therefore, a cheaper and reliable technique based on symptoms or other relevant factors needs to be developed to reduce the invasiveness of procedures, medical costs, and for early detection and prevention of liver damage.

1.3 Objective of the study

Due to the expensive and invasive nature of the biopsies needed in countries with few to no technical and medical advancements, a computational technique that aims at predicting the class of fibrosis of patients based on parameters carrying statistical and biological significance need to be developed. For this purpose, machine learning algorithms like KNN, Naive Bayes classifier, SVM, and neural network are deployed on the HCV dataset of Egyptian Patients from the UCI Machine Learning Repository. A machine learning method does not need computationally intensive resources unless it is on a large dataset. Zeroing down the essential factors observed during disease progression should act as inputs for the algorithms being used to predict the baseline histological staging of the patients. Predicting the baseline histological staging helps determine the medications and or therapies needed to prevent the disease progression, and ultimately cirrhosis in worst-case scenarios. The objective is to develop a usable model that helps predict the stage of the disease progression in patients, thereby helping develop a method which is economical, fast, and easy to use.

1.4 Scope of the study

This study aims to develop a model that predicts the baseline histological staging of a patient taking into consideration all the 28 variables available in the Dataset. A total of 1385 patients were screened, and their observations were recorded. For data analysis, the hold-out method can be used which divides the data into training (90%) and testing (10%), and then applying the algorithms and checking their accuracy of prediction using a confusion matrix. It is to be done on the normalized data set for which min-max normalization is used, which will be explained in the report. Even though the outlier detection is done, the outliers are still included in the data set for analysis since the removal of outliers make the Dataset very concise. Hence, this poses a problem for most classification algorithms, which might give false-positive predictions. It may lead to a reduction in the accuracy of the prediction. Further, new features are engineered, which helped categorize the baseline histological staging into two parts, namely, one and two. 1 includes stages and 1 and 2 (mild fibrosis) and two, which includes stages 3 and 4 (severe fibrosis). Through collinearity analysis, we come to understand that the variables in the data set are not correlated. It is made apparent through the collinearity plots seen further in the report. Hence, we try and apply as many algorithms as possible, and check for the most reliable one with the best statistics, using both, the Dataset as a whole, and the training-testing split that is created through random sampling. Further, the algorithms are used on the Dataset, and the most reliable algorithm for the prediction of the baseline histological staging is chosen as a deployable and can be used for decision making while predicting the stages of fibrosis is a patient infected by HCV. Algorithms like neural network and SVM are attempted but cannot be deployed with ease since they are computationally intensive and slow. Even after tuning models by adding and subtracting model parameters, some models show better performance, while some algorithms tend to perform slower or worse. Choosing the fastest, most accurate, and the most economical (in terms of computational cost) algorithm defines the scope of this project.

1.5 Limitations of Study

Even though machine learning algorithms are known to be reliable, one cannot rely on the predictions given by algorithms. It certainly can be considered as a basis for decision making with proper statistical facts and figures, but it cannot be regarded as the complete source for decision making. False positives have to be ruled out. Deployment and adoption of machine learning algorithms in the medical community is a time-consuming task. Relevant results and trials have to be in place to establish proof since machine learning and Artificial Intelligence are an emerging concept and might not be adopted as readily as they have been in other settings. Training the medical personnel to carry out predictions by using the ML algorithms poses another challenge. Proper training has to be provided, which is a time-consuming task, and supervision is needed. Machine Learning algorithms require a significant amount of data. The problem of overfitting can arise if algorithms are used on a small data set. These are the possible limitations one can think of throughout the CRISP-DM workflow usage until the deployment stage.

2. Literature Review

The HCV or Hepatitis C Virus has serious health implications and symptoms. It can transfer through a re-useable syringe needle. More than 170 million have been affected by the virus.[2] One of the most affected places in the world due to Hepatitis C is Egypt, as stated in the research paper titled, "A Novel Model-Based On Non Invasive Methods for Prediction of Liver Fibrosis." The following research paper and its suggestions and used and implemented throughout the project. It states that in Egypt, HCV has become a considerable problem to the extent where it has economic and health issues nationwide[3]. The virus initially begins as fibrosis (a complication in the liver functioning). It can lead to cirrhosis(liver failure) or liver cancer if not identified and treated in a timeframe of up to 2 years[4]. No research suggests a pattern as to how quickly liver fibrosis turns into cirrhosis, and identifying and treating the virus is difficult and expensive.

The Egyptian government is trying to treat patients with chronic(long-term) Hepatitis C. Therefore, the amount of people affected, and their financial constraints, makes it necessary that an alternate solution to a liver biopsy is presented[5]. The most accurate way to study fibrosis (liver damage) is to have back-to-back liver biopsies(extraction of the liver tissue) conducted. It is a costly process and can cause a lot of inconvenience to patients, as it is an invasive process that can cause further complications[6][7]. A lot of alternative solutions like imaging techniques that include 3-D imaging of liver using software via a CT scan, and serum markers that signify the presence of biological indicators in the bloodstream, have been used to diagnosing liver fibrosis.[8],[9].

Through the project, one can develop, evaluate, and validate a prediction model that can help prevent the use of invasive techniques altogether, by using data as a means of diagnosis. If a patient has liver fibrosis, it is essential to understand the amount of damage done to the liver. Doctors would typically assign a stage by looking at the scariness of the liver. This process is subjective, as some doctors would categorize the same liver in different stages. Hence, it acts as a limitation. The process helps other doctors understand the damage inflicted on a particular liver.

One of the most adopted scoring systems is the METAVIR scoring system. A score is assigned to a liver based on the progression of fibrosis severity in it. The score is usually assigned one; a doctor takes a liver tissue sample or a biopsy. The score ranges from A0 to A3. A0 meaning no activity, A1 meaning mild activity, A2 meaning moderate activity, and A3 meaning severe activity. The above categories comprise half the score, which states the fibrosis activity, the other half being the stage of fibrosis (damage is already done to the liver). The score ranges from F0 to F4. F0 meaning no fibrosis, F1 meaning portal fibrosis without septa (elements of the liver that make it septic), F2 meaning portal fibrosis with few septa, F3 meaning numerous septa without cirrhosis(liver failure), F4 meaning cirrhosis. Thus, someone with mild disease would have a score of A2, F2 METAVIR score, and someone with severe disease would have A3, F4 METAVIR score [10].

Historical clinical data can be used along with Artificial Intelligence techniques to aid physicians as well as patients in the diagnostic process [11]

Decision Support Systems are developed to support the solution of unstructured management issues to improve the decision-making process[12]. With the progress of statistical learning and artificial intelligence, intelligent decision systems are being designed[13]. Algorithms like, kNN, neural networks, support vector machines (SVMs), random forests, etc., are a few suggestions that can be developed into an excellent model to process the data on[14][15].

3.Terminologies and Packages:

Packages

psych- A general purpose toolbox for personality, psychometric theory, and experimental psychology. Functions are primarily for multivariate analysis and scale construction using factor analysis, principal component analysis, although others provide basic descriptive statistics[18]. (Package definition from CRAN)

Corrplot - A graphical display of a correlation matrix or general matrix. It also contains some algorithms to do matrix reordering.[18] (Package definition from CRAN)

E1071 - Functions for latent class analysis, short-time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, etc.[18]. (Package Definition from CRAN)

Caret – Miscellaneous functions for training and plotting classification and regression models[18].(Package definition from CRAN )

Party - A computational toolbox for recursive partitioning. The core of the package is ctree(), an implementation of conditional inference trees that embed tree-structured regression models into a well-defined theory of conditional inference procedures[18]. (Package definition from CRAN)

Class – Various functions for classification, including a k-nearest neighbor, Learning Vector Quantization, and Self-Organizing Maps[18]. (Package definition from CRAN)

Neuralnet – Attempts usage of an artificial neural network, training of neural networks using backpropagation, resilient backpropagation with or without weight backtracking, or the modified globally convergent version. The package allows flexible settings through a custom choice of error and activation function[18]. (Package definition from CRAN)

Mlbench -A collection of artificial and real-world machine learning benchmark problems, including, e.g., several data sets from the UCI repository[18]. (Package definition from CRAN)

caretEnsemble- Functions for creating ensembles of caret models: caretList() and caretStack(). caretList() is a convenience function for fitting multiple caret::train() models to the same dataset. caretStack() will make linear or non-linear combinations of these models, using a caret::train() model as a meta-model, and caretEnsemble() will make a robust linear combination of models using a GLM[18].

Terminologies

HCV- Hepatitis C Virus- **Hepatitis C** is a liver infection caused by the hepatitis C virus (HCV). Hepatitis C is a blood-borne virus[16].

ML- Machine Learning. Machine learning is considered as an application of artificial intelligence (AI) that provides an agent with the ability to systematically learn and improve performance measures at a specific learning rate with no learning and improvement from experience.[17].

Str- Str is a compact way to display the structure of an R object. It allows the user to see the various data structures and data types of all the columns in the Dataset in a condensed manner[18].

Hist- A histogram is a graphical display of data using bars of different heights [18].

BMI- Body Mass Index. Body mass index is a value derived from the mass and height of a person. Its SI unit is kg/m², resulting from weight in kilograms and height in meters[19].

WBC- White blood cell. White blood cells (WBCs) are also called leukocytes. Their function is to protect the body against harmful substances of both foreign and domestic origin[20].

RBC- Red blood cells help transport nutrition, oxygen, and everything that the cells need to survive, hormones, etc. It is the body’s internal fluid transport system[21].

HGB- Hemoglobin. It is the oxygen-carrying pigment found in the red blood cells.[22].

Plat- Platelets. During an injury, the platelets rush to the spot and prevent excessive blood loss by combining with other platelets[23].

AST 1- Aspartate transaminase ratio. An enzyme found in the liver. A damaged liver releases AST in the bloodstream. AST levels in the blood can be measured through a blood test. This test is then used by doctors to diagnose liver diseases or damage[24].

ALT 1- Alanine transaminase ratio for one week. ALT (alanine aminotransferase): It is an enzyme that is ONLY released in the blood during heart or liver damage. ALT levels are measured using a blood test and are elevated in case there is liver or heart damage. In this case, if a doctor has to check for Hepatitis C, finding the ALT levels is significant as it is a primary indicator of liver or heart damage.[25].

ALT 4- Alanine transaminase ratio (4 weeks)

ALT 12- Alanine transaminase ratio (12 weeks)

ALT 24- Alanine transaminase ratio (24 weeks)

ALT 36- Alanine transaminase ratio (36 weeks)

ALT 48- Alanine transaminase ratio (48 weeks)

RNA EOT- Ribonucleic acid at end-of-treatment.

RNA EF- Ribonucleic acid Elongation Factor.[26].

PCA- PCA stands for Principal Component and is as an orthogonal linear transformation that transforms the data to a new coordinate system, thereby reducing the dimensionality of the data. It is used when there are a large number of variables present in the data. [27].

kNN- k nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure by assigning a value (k) and then used for classification problems. [28].

Naïve Bayes- Naïve Bayes classifier is a classification machine learning algorithm. This classifier is responsible for assigning a class to a finite set. The standard assumption that the classifier does is that there is no dependent value for a feature [29].

SVM- A support vector machine (SVM) is a machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method responsible for plotting each data item as a single entity in n-dimensional space where n stands for the number of features with the value of each feature being the value of that particular coordinate[30].

Neural Network- A neural network can be vaguely defined as a system that is loosely modeled on the working of neurons in the human brain. They are used to understand and learn about both classification and regression problems. A neural net uses a network function to understand and predict information. [31].

Decision Tree- A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences[32].

Logistic Regression- It is a predictive analysis. It studies the relationship between categorical variables and a set of independent variables[33].

Bagging and Boosting- Bagging helps decrease the variance in the prediction by using additional data for training from Dataset using combinations with repetitions to produce multi-sets of the original Dataset. Boosting is an iterative technique that helps adjust the weight of an observation based on the previous classification[34].

Tuning- Tuning is a process of maximizing a model's performance without overfitting or creating too high of a variance. In machine learning, this is attained by selecting other parameters[35].

4. Solution Methodologies:

Data collection- Egypt had launched a nationwide government-sponsored campaign to treat patients with chronic hepatitis. The large pool of patients and the financial constraints necessitate prioritizing therapy to those most likely to progress rapidly to liver fibrosis. The data of these patients were collected over some time and hence form the Dataset, which is available from the UCI Machine Learning Repository.

Tools used- The UCI ML Repository defines the Hepatitis C Virus for Egyptian patients' Dataset as a Classification dataset with 1385 instances across 29 variables. It is a multivariate dataset, meaning there is more than one variable, and they include integers. It contains categorical variables with factors 0 and 1 wherein zero means absent, and one means present. There is no missing data, so data imputation is not required. For data exploration and analysis, R programming language was used. RStudio is the IDE used for R. The classification and other machine learning algorithms include kNN, Naïve Bayes, SVM, Decision Tree, Neural Network, and Logistic Regression. Fine-tuning of these algorithms was carried out to check the change in performance and accuracy of the algorithms. Further, bagging and boosting of algorithms was also carried out to check performance measure and accuracy.

kNN stands for k- nearest neighbors' algorithm and is chosen for analysis since it is a classification algorithm. K nearest neighbors is an algorithm that stores all available cases and classifies new cases based on a similarity measure by assigning a value (k) and then used for classification problems. The value of k chosen was 4. kNN is chosen since it gives the class of a particular point of datum in the whole Dataset, which is what exactly we are trying to predict, i.e., baseline histological staging.

Naïve Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is most comfortable to understand when described using binary or categorical input values. It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis is simplified to make their calculation tractable. Since the Dataset shows no collinearity amongst the variables, we try and attempt all classification algorithms and check for the accuracy of the models.

A Support Vector Machine (SVM), when given labeled training data (supervised learning), gives outputs as an optimal hyperplane, which categorizes new examples. Put simply, SVM is responsible for the separation of 2 or more classes.

An algorithm based on decisions, and their possible consequences, is known as Decision Tree. It uses tree-like graphs to represent the test and the possible outcomes of each test. The nodes in a decision tree represent a “test” conducted on attributes. Attributes like the number that occurs when a dice is rolled, whether flipping a coin will result in heads or tails, etc. The branches of that node represent the outcome of the test. The leaf nodes represent a class label, which is the decision taken after computing all the attributes. All the paths from the leaf to root will represent classification rules.

Neural Network (or Artificial Neural Network) can learn by examples. A neural network can be vaguely defined as a system that is loosely modeled on the working of neurons in the human brain. They are used to understand and learn about both classification and regression problems. A neural net uses a network function to understand and predict information. Adaptive means it can change its internal structure by adjusting weights of inputs. Neural networks were initially designed for real-world problems related to image recognition. Its application can be an object detection or pattern recognition.

The most appropriate analysis to conduct when the dependant variable is binary ( 0 or 1) is Logistic Regression. It is a predictive analysis technique. It is used to compare one dependant variable with one or multiple independent variables of different types(interval, ratio, nominal), and explain the relationship between them. Adding more independent variables to a model that runs logistic regression can lead to better variance (R2), which is also how one can determine how good the model is. An R2 value of 0.7-0.8 on a scale of 0 to1 is a good model, while a value of 0.9 is an excellent fit. It is why adding more independent variables is essential. However, this may result in overfitting and may reduce the accuracy of the model.

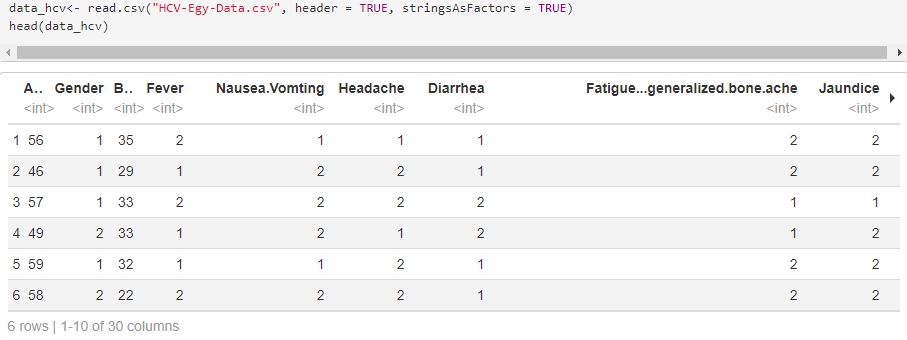
There are a lot of pseudo R2 values that have been developed for logistic regression. Due to computational issues, this value can sometimes be artificially low or high. A better way to present a model is to measure the goodness of fit. Based on the Chi-Square test, the Hosmer-Lemeshow is often used as a measure of goodness of the fit.

Sampling method- Random sampling was used to choose samples to create testing and training Dataset randomly. It is done through the set.seed function in R, which generates a series of random numbers. A 90-10 split is done wherein 90 percent of the patients are a part of the training dataset, and 10 percent of patients are a part of the validation dataset.

5. Data Analysis and Interpretation:

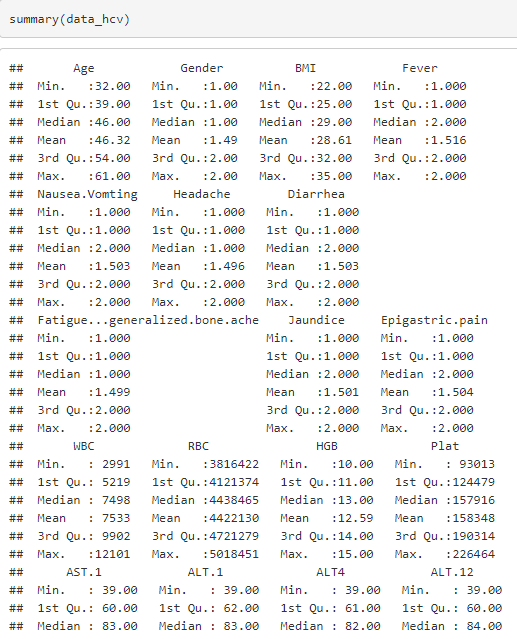
5.1 Data Acquisition:

Data Acquisition is the process in which data is acquired. In this project, read.csv function is used to read the CSV file.



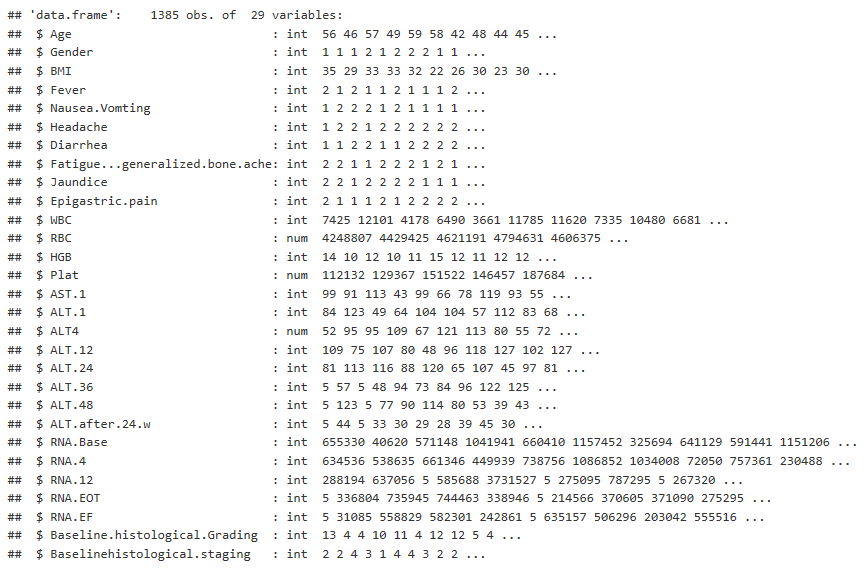
###### Figure 1 – Reading the Dataset (Refer to Appendix 1 for R code that applies “head” function to the Variable in which Dataset is stored)

Analysis and Interpretation: In Figure 1, we can see that the head part of the data\_hcv, which contains the HCV-Egy-Data.csv, is seen. data\_hcv contains multiple columns. Some columns seen in the head part are Age, Gender, BMI, Fever, Nausea and Vomiting, Headache, Diarrhea, Jaundice, etc.



###### Figure 2 – Summary of Dataset (Refer to Appendix 2 for usage of “summary” function in R code to give statistical summary)

Analysis and Interpretation: The summary function is shown in Figure 2, gives us a gist of important statistical figures of all the columns in the Dataset. As seen, each column has integer values. The columns which have the minimum value is 1, and the maximum value is 2 are categorical variables. 1 stands for absent, and 2 stands for present. Other columns, which include counts, are seen with differing minimum, maximum, and quartile figures, which can be considered as discrete variables. As one can observe in Figure 3, there are 29 columns which signify that there 29 factors or variables with respect to which observations are recorded. 1385 records signifies that there are 1385 patients who were screened. The CSV file has 1385 rows and 29 columns accounting for the whole dataset. The information about these variables can be found using the str command.



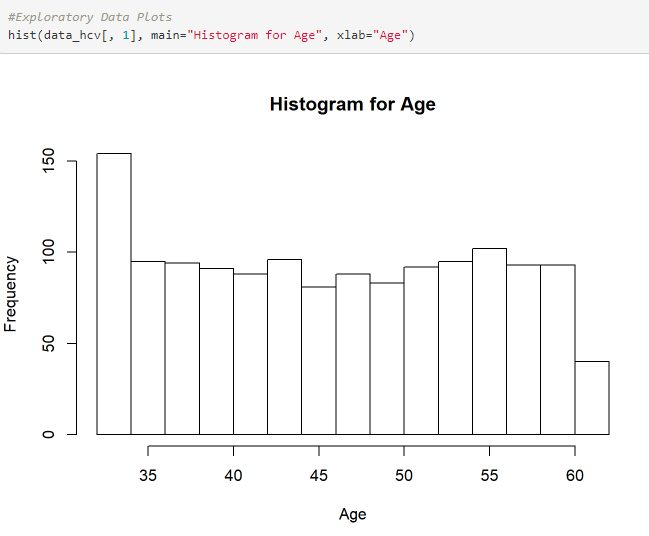
###### Figure 3 – str function for the HCV Dataset ( Refer to Appendix 2 for R code for usage of “str” function to show structure of the data\_hcv variable)

The columns and their datatypes depicted in Figure 3 are as follows-

Age- numeric, gender- categorical variable where 1 stands for Male and 2 stands for female and other columns like Fever, Nausea and Vomiting, Headache, Diarrhea, Fatigue, and Generalized bone ache, Jaundice, Epigastric pain is another set of categorical variables where 1 stands for absent and 2 stands for present. The other columns which include amounts of a biological entity include White blood cells, Red blood cells, Hemoglobin, Platelets, ALT 1, ALT 4, ALT 12, ALT 24, ALT 36, ALT 48, ALT after 24 weeks, RNA base, RNA 4, RNA, 12, RNA at the end of treatment, RNA Elongation factor, baseline histological grading. These are all real numbers, num datatype. Baseline histological staging includes factor datatype with factors 1, 2, 3, and 4, which are classified according to the METAVIR scoring system referenced above.

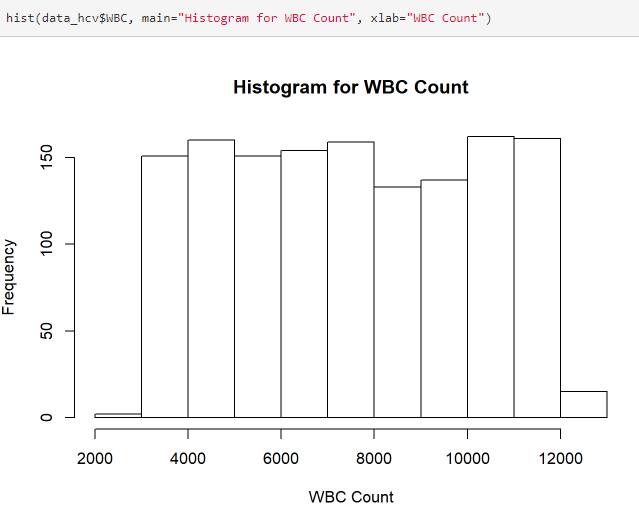
5.2 Exploratory Analysis:

Through exploratory analysis, we try and examine trends in the Dataset in the forms of plots.



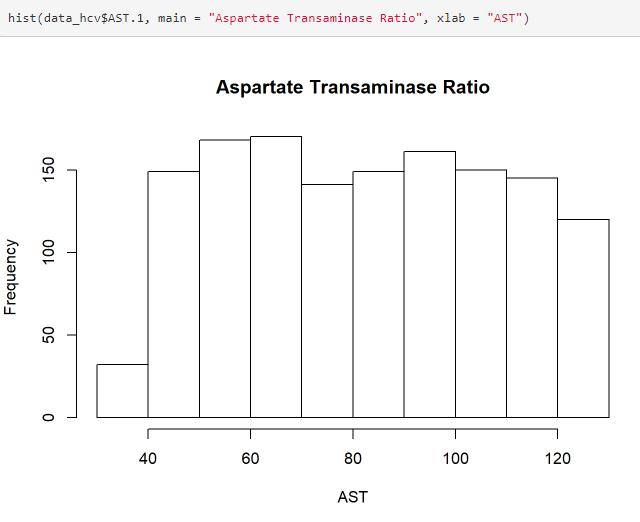
###### Figure 4 – Histogram for Age ( Refer to Appendix 3 for R code using “hist” function to plot a Histogram in R)

Analysis and Interpretation: Through this histogram in Figure 4, it is seen that most patients are in the age bracket of 35-60 and that the data is not normally distributed.



###### Figure 5 -Histogram for WBC Count ( Refer to Appendix 3 for R code using “hist” function to plot a Histogram in R)

Analysis and Interpretation: Through the histogram shown in Figure 5, it can be seen that the WBC count range is within the range of 2000-12000. There are relatively fewer patients in the range of 8000-10000 as compared to the other brackets seen in the data. WBC forms an integral part of the human immune system, and other components of the immune system are responsible for fighting fibrosis.



###### Figure 6 – Histogram for Aspartate Transaminase Ratio ( Refer to Appendix 3 for R code using “hist” function to plot a Histogram in R)

Analysis and Interpretation: Figure 6 explains AST, which is the Aspartate Transaminase enzyme, that is checked when the patients are screened for liver-related disorders. A rise in the AST levels is seen since the liver fails to proteolyze this enzyme. It can be considered as a biomarker for disease progression or diagnosis. As seen through this histogram, the data is not normally distributed.

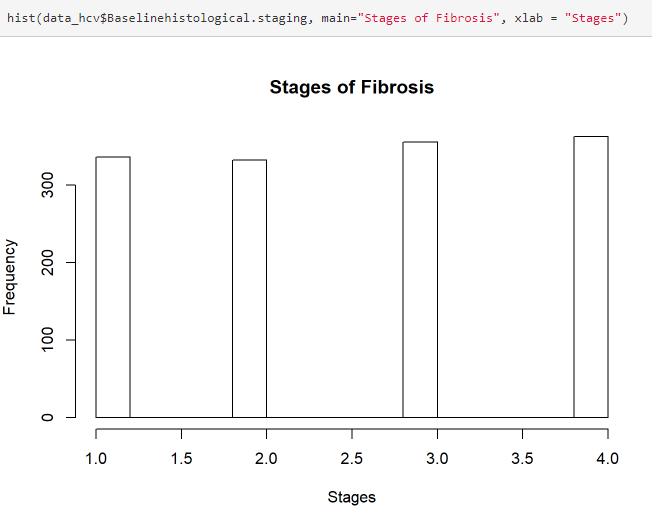
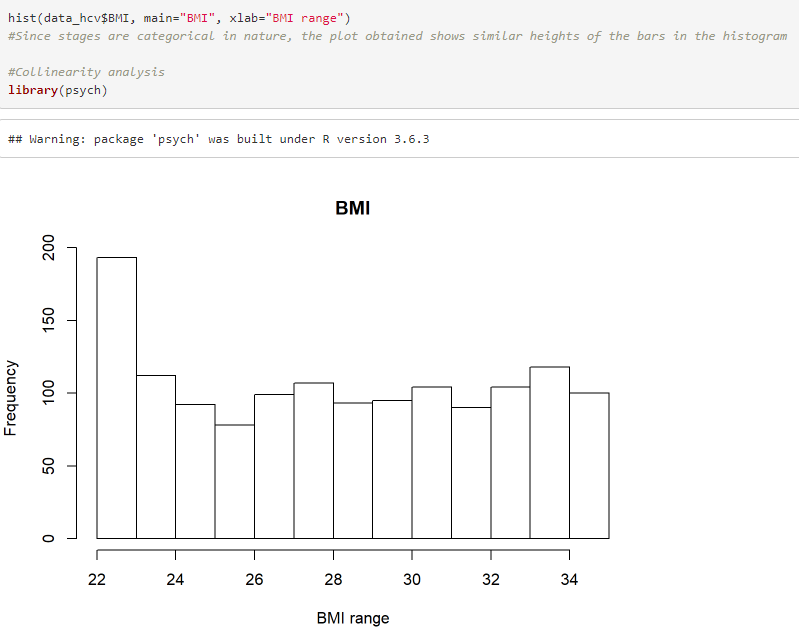


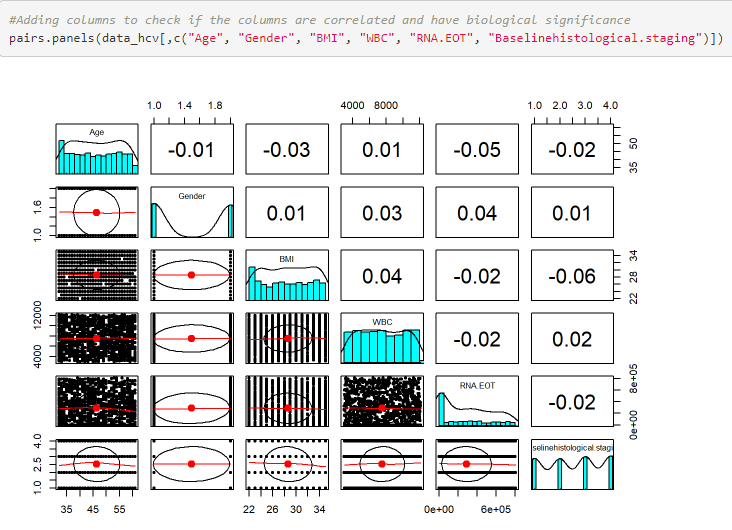
Figure 7 - Histogram for Stages of Fibrosis ( Refer to Appendix 3 for R code using “hist” function to plot a Histogram in R)

Analysis and Interpretation: 4 stages of fibrosis are seen in patients. As seen through the histogram in Figure 7, there is roughly the same number of patients in all 4 stages (of 1385 patients).



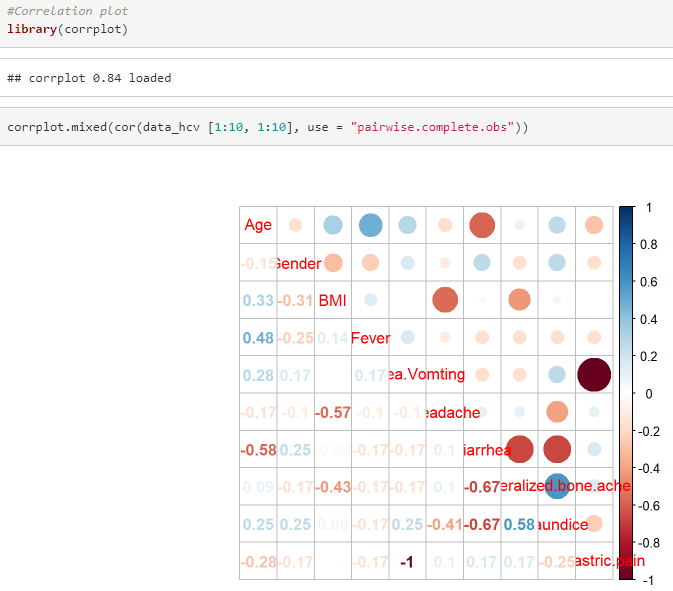
###### Figure 8- Histogram for Body Mass Index ( Refer to Appendix 3 for R code using “hist” function to plot a Histogram in R)

Analysis and Interpretation: As seen in the histogram in Figure 8, BMI stands for Body Mass Index. Most people are under the index of 22, while over 800 patients are above the BMI of 25, which is considered obese.



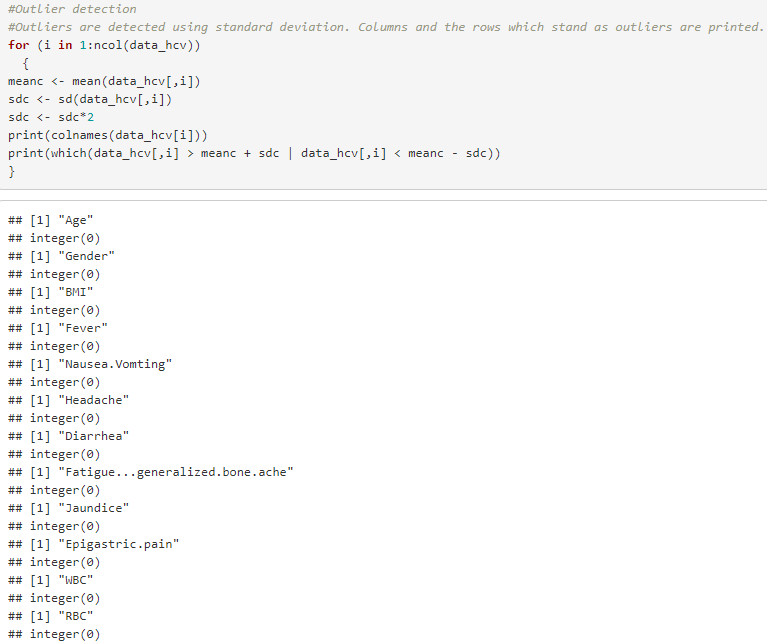
###### Figure 9 – Collinearity Plot to check for Biological Significance ( Refer to Appendix 3 for usage of “psych” package and “pairs.panels” function using R)

Analysis and Interpretation: In Figure 9, the Psych package is used to construct a collinearity plot. The columns are chosen for building the collinearity plot using the pairs.panels function are Age, Gender, BMI, WBC, RNA EOT, and Baseline histological staging. These columns were chosen to examine the correlation between age, gender and BMI, RNA EOT, and baseline histological staging. Here, these columns were chosen because we tried to determine if age. BMI, WBC count, and RNA EOT have any sort of correlation so that they can be used as a statistically viable predictor for predicting the baseline histological staging. But as we can see through the collinearity plot, there is very low or no correlation between the chosen factors. So, these factors cannot be directly used to predict the baseline staging of a patient.



###### Figure 10 – Correlation Plot ( Refer to Appendix 3 for usage of “corrplot.mixed” function from corrplot Package in R)

Analysis and Interpretation: As seen in Figure 10, the first ten columns are chosen for analysis. These columns, which act as factors, are plotted using the corrplot package available from CRAN. The examination of the first ten columns, which play a role in the disease progression, reveals that the columns or factors have no or a very low correlation. It makes us conclude that the factors in this Dataset are not collinear, neither positively or negatively. There are some in the range of 0 to -1, which state that the variables taken into consideration are not collinear.



###### Figure 11 – Detection of Outliers with Code Snippet ( Refer to Appendix 4 for R code which uses 2\*standard deviation for Outlier detection)

Outlier Detection

Analysis and Interpretation: As seen in Figure 11, outlier detection is taken into consideration in the Dataset. The presence of outliers in data can induce a bias in the data and hamper the prediction once algorithms are run. However, the consideration or removal of the outliers in the data is left to the data analyst. In this Dataset, we tried detecting outliers using standard deviation. We used 2\*sd, which is two times the standard deviation as the measure for outliers. Any point in data that is two levels of standard deviation away from the mean for each data point is considered as an outlier. In the case of this Dataset, outliers are not removed. They are considered during data analysis because the size of the Dataset is significantly reduced, and supervised machine learning algorithms do not perform well if the size of the data is small. It is why we keep the outliers in the data, and they form a part of the data analysis. It suggests that data points from each factor in the Dataset,  both smaller and more significant than two times the corresponding standard deviation are present. The columns or factors which are outliers obtained through this method are- Age, Gender, BMI, Fever, Nausea and Vomiting, Headache, Diarrhea, Fatigue, and generalized bone ache, Jaundice, Epigastric pain, WBC and RBC. Finer tuning of outliers can be done if we consider 3\*sd, which will not rule out many observations, but we have chosen to keep the outliers in the Dataset.

5.3 Data cleaning and Shaping (Feature Scaling):

Data Normalisation

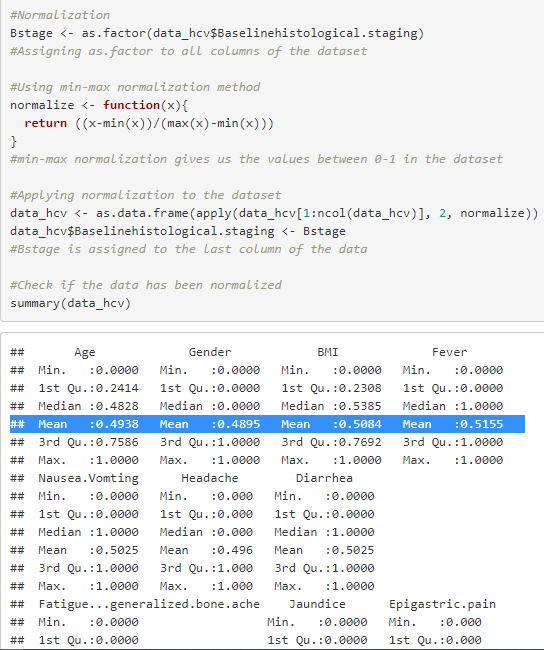


Figure 12 – Checking for Normality of Data ( Refer to Appendix 5 for R code which uses “min-max” method for normalization and “summary” function for checking normalization)

Analysis and Interpretation: Dataset in Figure 12 is normalized so that all the factors have their values that fall between the range of 0 to 1. To normalize the Dataset for this range, the min-max method of normalization is used.  Min-max normalization attempts a linear transformation on the data. It ensures that all the data points of the Dataset do not extensively differ from each other and are scaled to a particular range so that analysis can be quickly done. If normalization is carried out, the attributes of a data are scaled, and then the models can perform predictions properly. If the normalization is proper or not can be checked with the use of summary function in R. As seen in the columns containing categorical variables, the minimum and maximum values have changed. Hence, we infer that the data has been normalized. In the case of categorical variables, where 1 and 2 are present, the mean value is 0.5 for factors, and 1 is the max value. Through this, we can infer that the Dataset has been appropriately normalized.

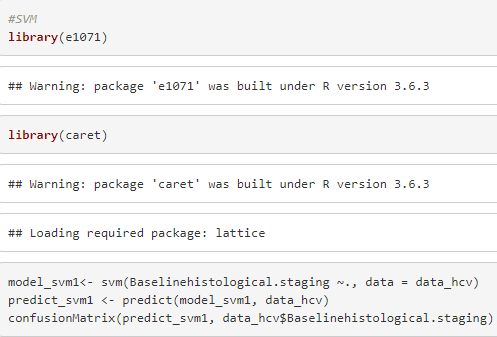
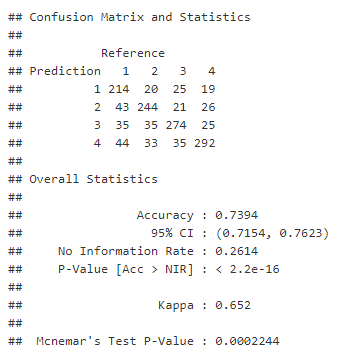


Figure 13 – SVM Model Construction and Evaluation with Snippets ( Refer to Appendix 6 for R code using “svm” function from “caret” package)

Prediction models are constructed using machine learning algorithms like KNN, Naive Bayes, SVM, decision tree, neural network, and logistic regression in R using RStudio IDE . The results of the prediction of these models are represented using a confusion matrix (which is created using ConfusionMatrix function). It gives us a performance measure of the classification algorithm along with its accuracy, p-value, Kappa statistics, and other figures. Accuracy defines how accurately the model has predicted a particular data point in the Dataset given the whole Dataset or the training part for the learning and testing part for testing the model. P-value stands for the probability value, which tells us how likely our value stands, assuming that the null hypothesis considered is true. Kappa statistics or the Kappa figure tells us the measure of how much accuracy the system possess compared to a random system. When a prediction is a true positive or of total accuracy, it attains a kappa value of 1. Using these terminologies from the confusion matrix, we grade how well the prediction models have performed. 

###### Figure 14 – Confusion Matrix of SVM Model ( Refer to Appendix 6 for R code to generate confusion matrix using “confusionMatrix” function in R.)

Analysis and Inference: As seen in Figure 14, the SVM algorithm is run on the whole Dataset using the caret package from CRAN. The function “svm” runs the algorithm. SVM is run, and predict function is used to predict the baseline histological staging of the HCV infected patients. The above confusion matrix is that of the Support Vector Machine. Here the whole Dataset is used for prediction, and the confusion matrix is created. According to the confusion matrix, the accuracy obtained is 73.94%, and the results can be considered statistically significant since the p-value is less than 0.05. The prediction and reference tell us that the SVM algorithm has predicted both true negatives and false negatives, along with false positives. 1, 2, 3, and 4 represent the baseline histological stages, and the exact predictions (correct) can be seen diagonally. If there are numbers scattered around the true positives (i.e., correct predictions), they represent a mistake in the prediction. As we see that there are some incorrect predictions, the accuracy of the model in predicting the baseline histological staging decreases.

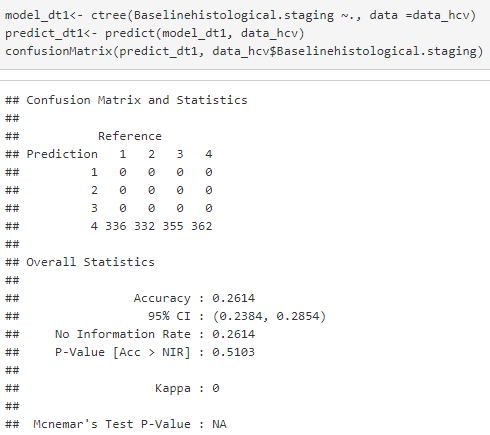
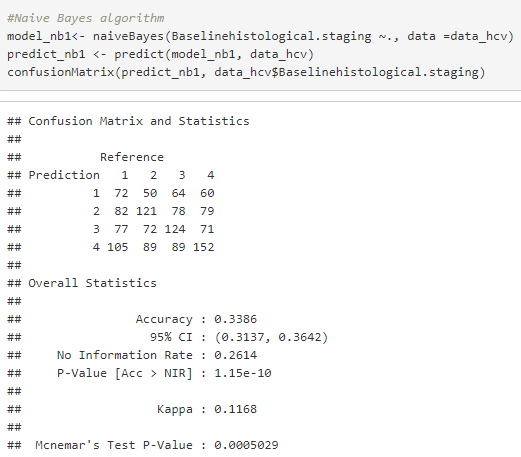


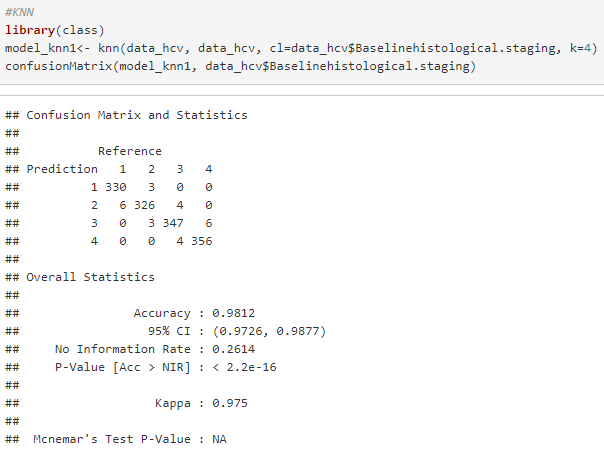
Figure 15 – Confusion Matrix of ctree Model with Code Snippet (Refer to Appendix 6 R code for creating confusion matrix using “confusionMatrix” function in R for ctree model using “party” package)

Analysis and Interpretation: ctree is the core function of the CRAN R package party. As seen in Figure 15, it is the package used for recursive partitioning. Recursive partitioning helps create a decision tree from the input dataset to classify points of a dataset by splitting it into smaller parts based on other variables. A decision tree is obtained using this recursive partitioning. As seen through the confusion matrix, all the predictions are a part of baseline histological stage 4, which is not the case in reality. The only accuracy obtained in this model is through when the model correctly predicts the number of patients who are suffering through level 4 of fibrosis. The accuracy obtained is 26.14%, and a very poor p-value is seen. Kappa statistics give us a figure of 0.26, which is not a great value, and hence the model can be considered a poor prediction and classification model.



###### Figure 16 – Confusion Matrix of Naïve Bayes Model (Refer to Appendix 6 R code for creating confusion matrix using “confusionMatrix” function in R using “e1071” package)

Analysis and Interpretation: Naive Bayes algorithm is called so because- 1) it is based on Bayes algorithm and 2) the naive is added because of the assumption that it makes- it assumes that the features of a certain measurement are independent of each other which might not always be the case. Its use case is mostly seen in text classification, e.g., spam message classification. In this model, the Naive Bayes algorithm is attempted with the whole Dataset for training and whole Dataset for prediction of the results, again using the whole Dataset for testing too. It yields an accuracy of 33.86%, which makes it a poor model for the prediction of baseline histological staging. It has an extremely low kappa statistics figure, further suggesting that it is not a great model for prediction.



###### Figure 17 – Confusion Matrix of kNN Model (Refer to Appendix 6 R code for creating confusion matrix using “confusionMatrix” function in R using “class” package)

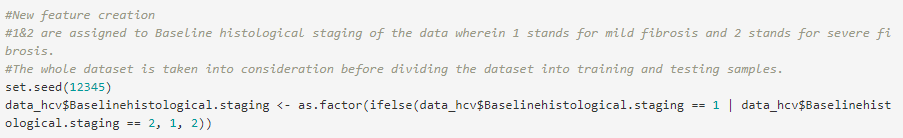
Analysis and Interpretation: The kNN algorithm is a supervised machine learning algorithm that is used to solve both classification and regression problems. kNN assumes that the data points within a certain distance are similar. This distance is defined as the Euclidean distance. The value of k used in the model constructed with this Dataset is 4. As seen in Figure 17, we observe that we get an accuracy of 98.12%, which makes it a great model. The kappa statistics give us a figure of 0.975, which is very close to 1, thereby making it the best model so far. A low p-value suggests that the model is statistically significant and highly sensitive. It can be regarded as a highly reliable model and can be used for making predictions of the baseline histological staging considering the whole Dataset

5.4 Model Construction and Evaluation, model tuning, test train dataset, feature addition (Data Engineering)

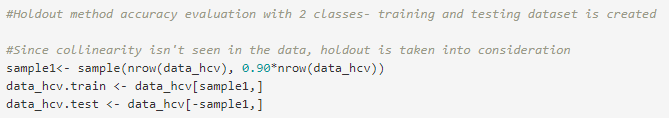
Feature Addition

A new feature is constructed by clustering the two baseline histological stages into one for predictions. Further, the hold-out method, which involves splitting the Dataset into training and testing set, is to check if there are any variations in the prediction models. A 90-10 split for training and the testing dataset is adopted. Generally, it is suggested when the Dataset is huge (number of observations). But since the Dataset has only 1385 observations, a larger training dataset is created.

The tuning of models is carried out to increase the accuracy, if possible, and results are compared. Another reason for model tuning is because there is very low or low collinearity seen in the factors of this Dataset.

Figure 18 – New Feature Creation Code Snippet (Refer to Appendix 7 for R code which uses “ifelse” condition to cluster 1 and 2 into one and 3 and 4 into two.)

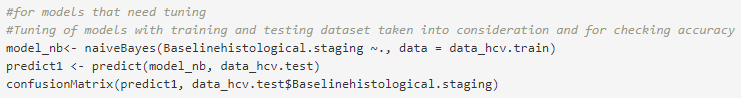
Explanation: A new feature is created, as seen in Figure 18, using the baseline histological staging column. Stages 1 and 2 are a part of 1, and stages 3 and 4 are a part of 2.



###### Figure 19 – Hold Out Method ( Dividing Data into Training and Testing) with Code Snippet

###### ( Refer to Appendix 7 for R code which divides data\_hcv into 90-10 split for training and testing randomly using “sample” function)

In Figure 19, a 90-10 split for the training and testing dataset is done. Since there is no collinearity observed in the data, the hold-out method is considered. The model tuning approach is used to try and improvise the accuracy of the models.



###### Figure 20 – Naïve Bayes Model with Code Snippets ( Refer to Appendix 7 for R code which uses “naiveBayes” function from “e1071” package)

The code snippet in Figure 20 represents that naive Bayes function using the e1071 package, and a confusion matrix is constructed for with the prediction.

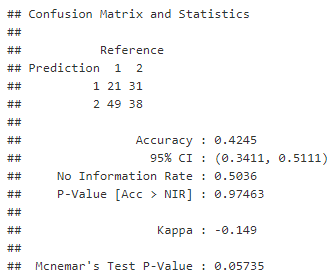


Figure 21 – Confusion Matrix for Naive Bayes Model (Refer to Appendix 7 for R code which uses “confusionMatrix” function for Confusion Matrix creation for Naïve Bayes model)

Analysis and Interpretation: As seen from the confusion matrix for the Naive Bayes algorithm, the model accuracy using the hold out method is 42.45%, which is an increase from the previous 33.86%. But, the p-value is close to 0.98, which makes this model a poor performer. The kappa statistics give us a negative figure, which tells us that the model has predicted and hence performed poorly. It makes it clear that the Naive Bayes algorithm isn’t the best choice for this particular Dataset after adopting the hold-out method.

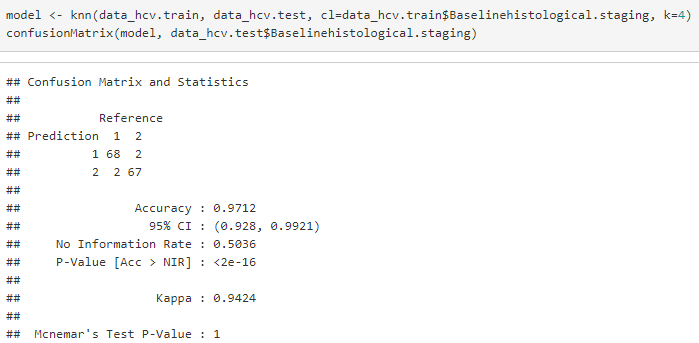
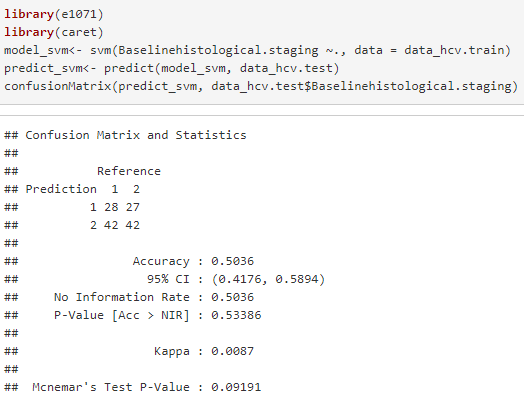


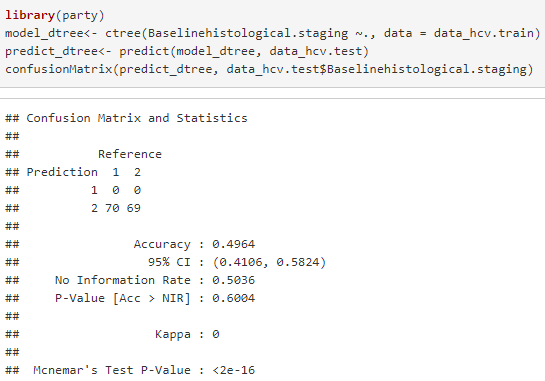
Figure 22 – Confusion Matrix for kNN Model (Refer to Appendix 7 for R code which uses “confusionMatrix” function for Confusion Matrix creation for kNN model)

Analysis and Interpretation: KNN still performs extremely well with the value of k as 4. It attains an accuracy of 97.12%, even in the hold-out method, as clearly seen in Figure 22. It is statistically significant since the obtained value is less than 0.05, which is considered as the cutoff. Even with the hold-out method taken into consideration, there is very little change in the accuracy, suggesting that this is the best model for this particular Dataset. It can be used while using the whole dataset and using training and testing datasets while modeling construction.



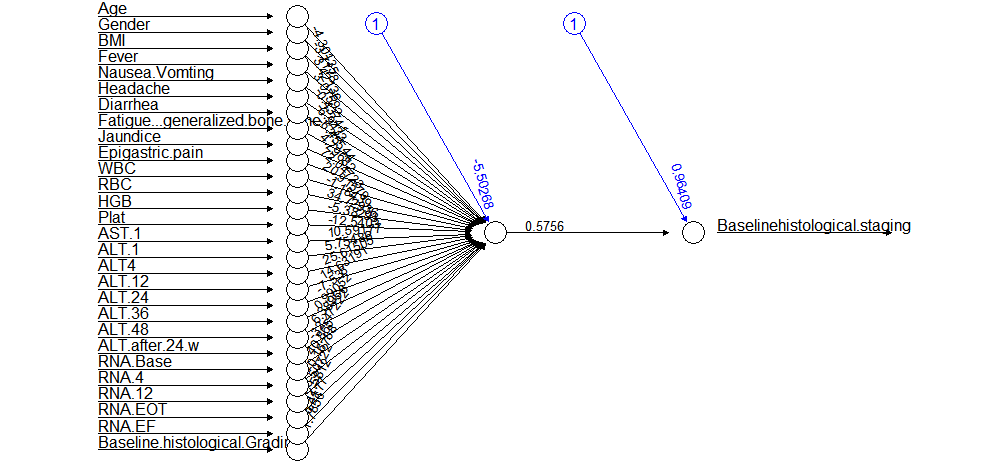
###### Figure 23 –Confusion Matrix for SVM Model (Refer to Appendix 7 for R code which uses “confusionMatrix” function for Confusion Matrix creation for SVM model)

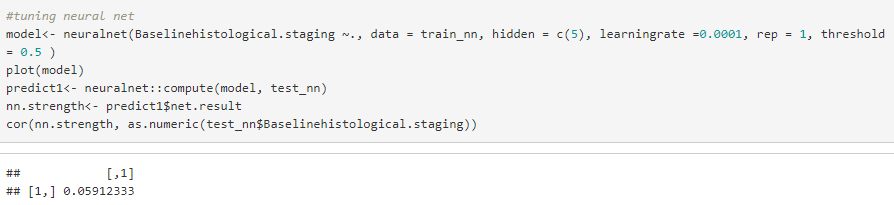
Analysis and Interpretation: A decrease in the accuracy of the SVM algorithm is seen when compared to the accuracy wherein the whole Dataset is considered. The results are not statistically significant, as seen through the p-value in Figure 23. So if one decides to work with SVM for this particular Dataset, one has to stick to considering the whole Dataset for prediction. There is a significant decrease in the accuracy, and a poor p-value score is obtained, suggesting that it is not the best model to be used for this Dataset.



###### Figure 24 – Confusion Matrix for ctree Model (Refer to Appendix 7 for R code which uses “confusionMatrix” function for Confusion Matrix creation for ctree Model)

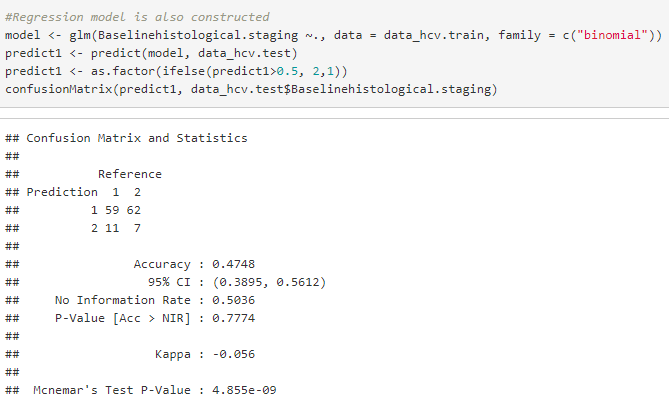
Analysis and Interpretation:  The accuracy of the ctree sees a significant increase when compared to the decision tree model, which considers the whole Dataset. However, the p-value is high, and this model cannot be considered for reliable predictions. The main observation here is that Kappa statistics give us a value of zero, which is unacceptable. This model cannot be used for prediction using both the hold-out method and when considering the whole dataset





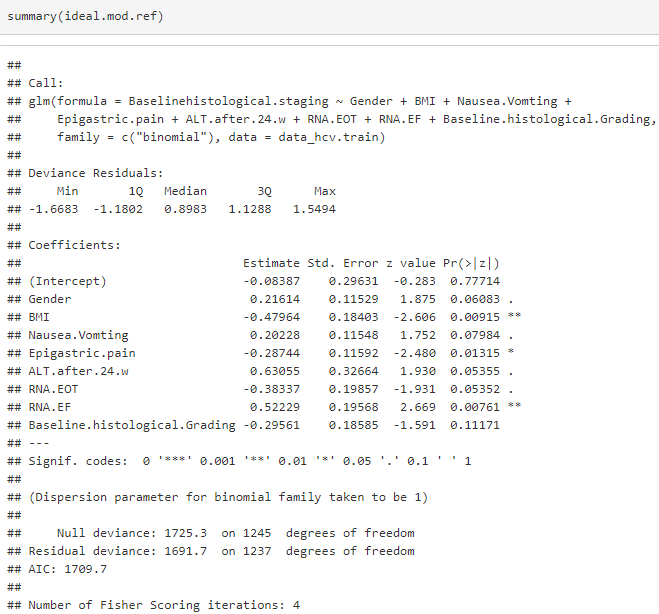
###### Figure 25 – Neural Network and Accuracy ( Refer to Appendix 7 for R code for contructing a neural network using “neuralnet” package)

Analysis and Interpretation: An accuracy of 59% is seen when a learning rate of 0.0001 is included, and the number of hidden layers is used as a vector of 5 in Figure 25. Attempt to model showed very negligible accuracy and hence cannot be plotted properly. However, after adding a learning rate that determines the rate at which a neural net absorbs information and increasing the number of nodes of the network increases the accuracy to 59% from a negative 25% (negative accuracy is not statistically significant). The neural network above shows us all the columns, and each node gives us a threshold value and weights. A threshold value of 57.56 is seen on the baseline histological grading node, which we are predicting in this project.



###### Figure 26 – Confusion Matrix for Logistic Regression Model (Refer to Appendix 7 for R code which uses “confusionMatrix” function for Confusion Matrix creation for Logistic Regression Model using “glm” function in R)

Analysis and Interpretation: Logistic Regression predicts the binary response (in the form of 1s and 0s or True or False), taking into consideration one or more independent variables. If represented graphically, it divides the space in a manner wherein one line divides the space wherein the data can be linearly classified. We perform logistic regression since we have divided the baseline histological staging into 2 clusters 1 and 2. The accuracy is then tested using the test dataset. The accuracy obtained is 47.48%, as clearly seen in Figure 26. Even though the accuracy is not great, this method is carried out so that we can backtrack and determine important factors in the Dataset through this method.



###### Figure 27 – Obtaining Statistically Significant Columns through Backtracking ( Refer to Appendix 7 for R code which uses “step” function for backtracking the Logistic Regression Model)

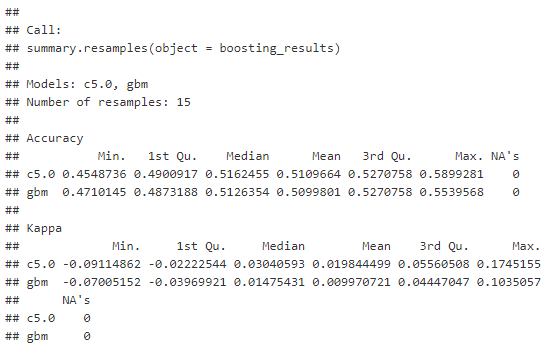
Analysis and Interpretation: Backtracking here is a line search, and it helps determine the important factors in this Dataset given the whole Dataset. It uses logistic regression as a reference and gives us statistically significant answers using the summary function of the reference model. According to the inputs from logistic regression, we get BMI and RNA Elongation factor

5.5 Interpretation and Results of the Models Created:

As seen, some models tend to perform better when the whole Dataset is considered, and some models perform better when the hold-out methods (that is when sects of the data are made), which include training and testing Dataset are used. Specifically, KNN is well suited for this Dataset as it gives us 97% accuracy, and SVM tends to perform well with the whole Dataset. With tuning, which increased the number of hidden layers and inserting a learning rate, a neural net is seen to perform better. The accuracy of the decision tree is almost doubled, as seen once the hold-out method is taken into consideration. As far as regression is concerned, it holds an accuracy of 47%. When backtracking is attempted, we understand that columns like BMI, Gender, RNA EF (RNA Elongation factor), and Epigastric pain are the ones that can be considered as significant. These can be considered as markers of disease during disease progression. However, Jaundice, Aspartate Transaminase are other important markers one can consider during disease progression.

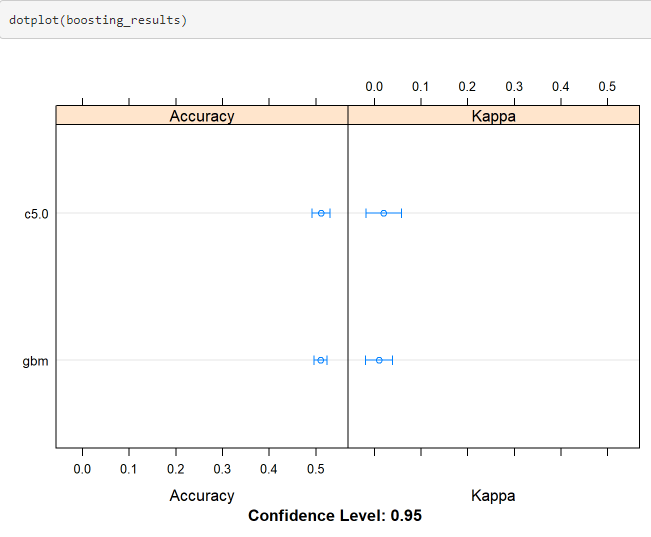
###### Bagging and Boosting

Boosting is another machine learning algorithm (ensemble method- consists of multiple algorithms) that aims to increase accuracy by reducing variance and, most importantly, bias. It helps turn poor predictions into better ones, in turn, increasing its accuracy. The miscalculations which might have occurred while predictions are collected, weights are assigned, and then preference is given to these calculations once the boosting algorithm is run while model creation. For these purposes, mlbench, caret, and caretEnsemble packages from CRAN are used.



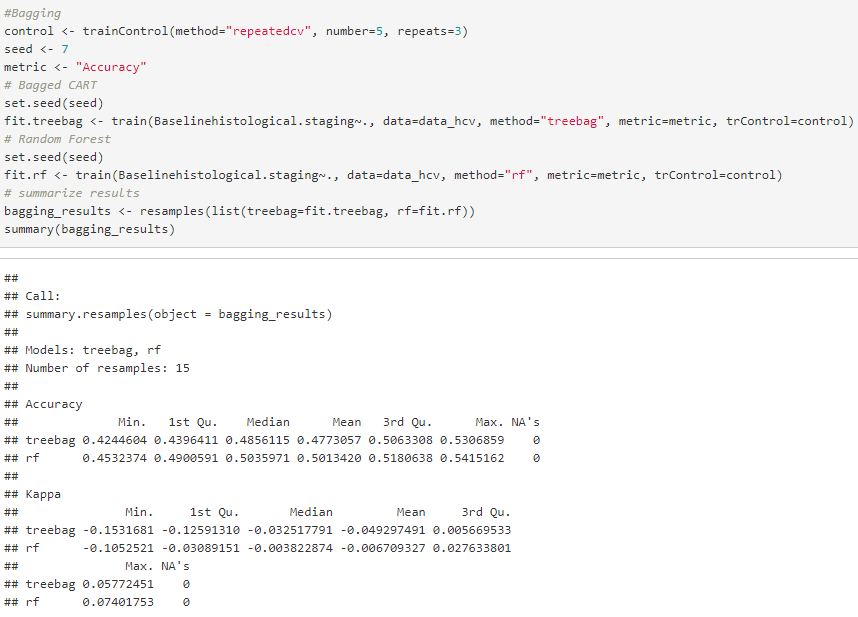
###### Figure 28 – Summary for Boosting Method ( Refer to Appendix 8 for R code which uses “caretEnsemble” package for boosting and “summary” function for statistical summary of the model)

Analysis and Interpretation: The c5.0 and gbm are used on the whole Dataset. These are two algorithms that consist of boosting methods. Gbm stands for the gradient boosting method and inserts steps for a regression model during recalculation. c5.0 obtains an accuracy of 58.99% and gbm an accuracy of 55%. Boosting is hence considered successful. The plot below in figure 29 explains the accuracy, while Figure 28 summarizes the results.



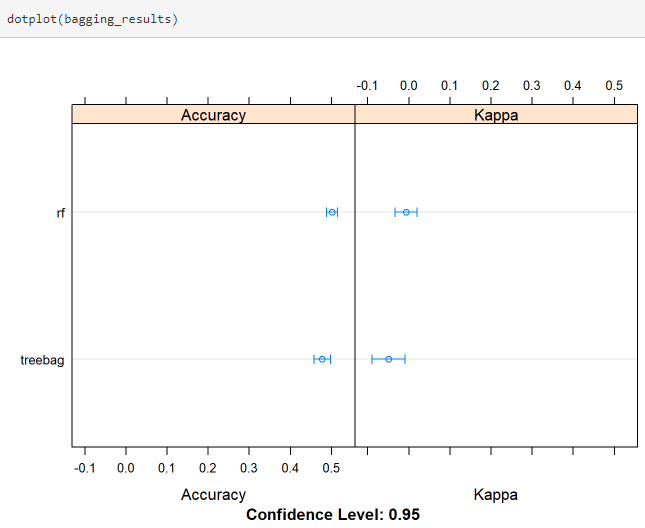
###### Figure 29 – Boosting Method Plot ( Refer to Appendix 8 for R code which uses “dotplot” function to create a Dot Plot for result of Boosting Model)

After boosting is executed, the accuracy of both c5.0 and gbm tends towards 0.5, which suggests that boosting is successful.



###### Figure 30 – Bagging via treebag and rf ( Refer to Appendix 8 for R code which uses “mlbench” and “caretEnsemble” packages for Bagging and “summary” function for statistical summary)

Analysis and Interpretation: Bagging is a method which is similar to boosting. The two algorithms used here are treebag and rf. Rf stands for random forest and tweaks, which results in a powerful classifier. After bagging, the accuracy for treebag and rf are 53% and 54% respectively



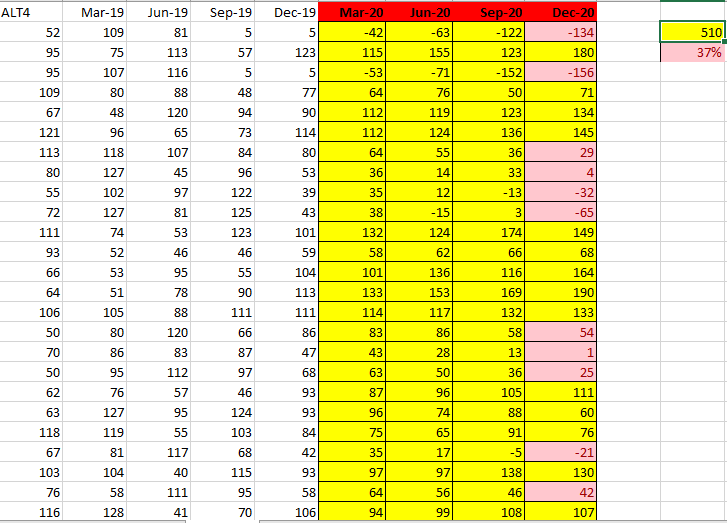


###### Figure 31 – Bagging Method Plot ( Refer to Appendix 8 for R code which uses “dotplot” function to create a Dot Plot for result of Bagging Model)

Analysis and Interpretation: The accuracy of both random forest and tree bag tend towards 0.5 as seen from the plot in Figure 31. The blue dot tends towards 0.5, thereby explaining successful bagging.

Forecasting :

The presence of ALT in the bloodstream can indicate liver complications, so we try to use forecasting methods to predict data until 72weeks and conclude if a liver transplant is necessary or not. Since the data for ALT were available, in an increment of 12 weeks, it was necessary to see the trends that could be observed. A patient with liver fibrosis can live for up to 2 years without a liver transplant before its treatment becomes difficult, and eventually, it turns into liver cirrhosis. In the project, prediction of ALT levels for weeks 60, 72, 84, and 96 were also carried out. ALT levels less than 56 units per liter are considered to be normal. Through the predicted data, one can observe the effectiveness of the treatment given to patients, based on the data available for ALT week 96[36][37]. The seasonality index has been used as a forecasting method where excel automatically calculates.

Figure 32 – Forecasting ALT till week 96

Explanation: Using the FORECAST function in excel, the forecast values were calculated. To perform the forecast, instead of ALT4, ALT 12, ALT24, real dates were required. Therefore, the first date was assumed, and then there was an increment of 12 weeks taken into consideration. The dates Mar-20 to Dec-20, which are highlighted in red in the above Figure 32, are the forecasted weeks. The rest of the data was already available. The idea was to predict the data over the next 48 weeks and determine whether the ALT levels were reducing or not. If they did and had any value below 56 units per liter, that means that the given treatment was successful or aiding the patient's cure. Five hundred ten people were found to have an ALT level of fewer than 56 units per liter, which is 37% of the total number of patients. It concludes that the provided treatments were not efficient, and an alternate treatment/therapy could be more beneficial.

Conclusion:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Name and method information | Packages/functions  used | Accuracy | Other statistics |
| Method 1 | Considering the whole Dataset:  Support Vector Machine  Ctree  Naive Bayes  kNN | Caret package in R  party package in R  e1071 package in R  class package in R | 73.94%  26.14%  33.86%  98.12% | p-value < 0.05  p-value > 0.05  p-value < 0.05  p-value < 0.05 |
| Method 2 | Hold-out Method:  Support Vector Machine  Ctree  Naive Bayes  kNN  Neural Network  Logistic Regression | Caret package in R  party package in R  e1071 package in R  class package in R  neuralnet package in R  glm function in R | 50.36%  49.64%  42.45%  97.12%  59%  47.48% | p-value > 0.05  p-value > 0.05  p-value > 0.05  p-value < 0.05  N/A  p-value > 0.05 |
| Method 3 | Boosting:  c5.0  gbm  Bagging:  Treebag  rf | Mlbench, caret, and caretEnsemble  Mlbench, caret, and caretEnsemble | 58.99%  55.39%  53%  54% |  |

Figure 33 - Conclusion

Based on these machine learning models, the report has been formed. Predicting the baseline histological staging of patients based on various factors, it can be concluded that kNN performs the best for this Dataset. The accuracy obtained was 97.98%, which is very high. This model can be deployed (as per CRISP-DM) and can be used by various medical professionals to grade the progression of diseases based on other factors (non-invasive) like BMI, RNA Elongation Factor, ALT ratio, Jaundice, and epigastric pain[38]. kNN models , while considering the whole Dataset or while considering hold-out method(90-10) for analysis, exhibit the same. These factors carry biological significance in disease progression and can be used as biomarkers during liver disease examination and prognosis.

References:

1) <https://www.cdc.gov/hepatitis/hcv/index.htm>

2) Mohamed, Amal Ahmed et al. Hepatitis C Virus: A Global View. World Journal of Hepatology 7.26 (2015): 26762680. PMC. Web. 4 Nov. 2017.

3) Elgharably, Ahmed et al. Hepatitis C in Egypt Past, Present, and Future. International Journal of General Medicine 10 (2017): 16. PMC. Web. 4 Nov. 2017

4) Bourliere, M., et al." Validation and comparison of indexes for fibrosis and cirrhosis prediction in chronic hepatitis C patients: proposal for a pragmatic approach classification without liver biopsies." Journal of viral hepatitis 13.10 (2006): 659-670.

5) Trooskin, Stacey B., Helen Reynolds, and Jay R. Kostman." Access to costly new hepatitis C drugs: medicine, money, and advocacy." Clinical Infectious Diseases 61.12 (2015): 1825-1830.

6) Gebo, Kelly A., et al." Role of liver biopsy in management of chronic hepatitis C: a systematic review." Hepatology 36.S1 (2002).

7) Chalasani, Naga, et al." The diagnosis and management of nonalcoholic fatty liver disease: Practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association." Hepatology 55.6 (2012): 2005-2023.

8) Cobbold, J. F. L., et al." Optimal combinations of ultrasoundbased and serum markers of disease severity in patients with chronic hepatitis C." Journal of viral hepatitis 17.8 (2010): 537-545.

9) Kamal, Sanaa M., et al." Progression of Fibrosis in hepatitis C with and without schistosomiasis: correlation with serum markers of fibrosis." Hepatology 43.4 (2006): 771-779.

10) <https://www.researchgate.net/publication/323130913>

11) Keltch, Brian, Yuan Lin, and Coskun Bayrak." Comparison of AI techniques for prediction of liver fibrosis in hepatitis patients." Journal of medical systems 38.8 (2014): 60.12)

12) Power, Daniel J." A brief history of decision support systems." DSSResources. COM, World Wide Web, http://DSSResources. COM/history/dsshistory. html, version 4 (2007).

13) S. Guerlain, D.E. Brown and C. Mastrangelo, Intelligent Decision Support Systems, Proc. IEEE International Conference on Systems, Man, and Cybernetics 3 (2000), 193438

14) E. Turban, J.E. Aronson and T-P. Liang, Decision Support Systems, and Intelligent Systems (7th ed.), Prentice-Hall of India, 2006.

15) T. Hastie, R. Tibshirani and R. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.), Springer, 2009.

16) <https://www.cdc.gov/hepatitis/hcv/index.htm>

17) <https://expertsystem.com/machine-learning-definition/>

18) https://cran.r-project.org/

19) <https://en.wikipedia.org/wiki/Body_mass_index>

20) <https://en.wikipedia.org/wiki/White_blood_cell>

21) <https://www.medicinenet.com/script/main/art.asp?articlekey=9988>

22) <https://www.medicinenet.com/script/main/art.asp?articlekey=15738>

23) <https://www.medicinenet.com/script/main/art.asp?articlekey=4941>

24) <https://medlineplus.gov/lab-tests/ast-test/>

25) <https://www.medicinenet.com/script/main/art.asp?articlekey=6585>

26) <https://en.wikipedia.org/wiki/EF-G>

27) <https://en.wikipedia.org/wiki/Principal_component_analysis>

28) <https://www.saedsayad.com/k_nearest_neighbors.htm>

29) <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>

30) <https://www.techopedia.com/definition/30364/support-vector-machine-svm>

31) <https://www.investopedia.com/terms/n/neuralnetwork.asp>

32)<https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb>

33) <https://www.statisticssolutions.com/what-is-logistic-regression/>

34) <https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/>

35) <https://riskspan.com/news-insight-blog/tuning-machine-learning-models/>

36) <https://www.emedicinehealth.com/liver_blood_tests/article_em.htm>

37) <http://www.natap.org/2002/Oct/103002_2.htm>

38) <https://ieeexplore.ieee.org/document/8289800>

Appendix:

Appendix 1: Data Acquisition

#Data Acquisition

data\_hcv<- read.csv("C:/Users/Naman/Documents/HCV-Egy-Data.csv", header = TRUE,stringsAsFactors = TRUE)

head(data\_hcv)

Appendix 2: Data Exploration

#Data Exploration

summary(data\_hcv)

str(data\_hcv)

Appendix 3: Data Visualization

#Histograms – For generating Histograms.

hist(data\_hcv[, 1], main="Histogram for Age", xlab="Age")

hist(data\_hcv$WBC, main="Histogram for WBC Count", xlab="WBC Count")

hist(data\_hcv$AST.1, main = "Aspartate Transaminase Ratio", xlab = "AST")

hist(data\_hcv$Baselinehistological.staging, main="Stages of Fibrosis", xlab = "Stages")

hist(data\_hcv$BMI, main="BMI", xlab="BMI range")

#Collinearity plots – For showing collinearity in columns of the Dataset.

library(psych) – Both psych and corrplot packages help in determining and plotting collinearity

pairs.panels(data\_hcv[,c("Age", "Gender", "BMI", "WBC", "RNA.EOT", "Baselinehistological.staging")])

library(corrplot)

corrplot.mixed(cor(data\_hcv [1:10, 1:10], use = "pairwise.complete.obs"))

Appendix 4: For determining Outliers

for (i in 1:ncol(data\_hcv))

{

meanc <- mean(data\_hcv[,i])

sdc <- sd(data\_hcv[,i])

sdc <- sdc\*2

print(colnames(data\_hcv[i]))

print(which(data\_hcv[,i] > meanc + sdc | data\_hcv[,i] < meanc - sdc))

}

Appendix 5: For normalizing the Dataset

Bstage <- as.factor(data\_hcv$Baselinehistological.staging)

normalize <- function(x){

return ((x-min(x))/(max(x)-min(x)))

}

data\_hcv <- as.data.frame(apply(data\_hcv[1:ncol(data\_hcv)], 2, normalize))

data\_hcv$Baselinehistological.staging <- Bstage

summary(data\_hcv)

Appendix 6: Various models (kNN, Naïve Bayes, etc.) Construction and Evaluation using the whole Dataset.

library(e1071)

library(caret)

model\_svm1<- svm(Baselinehistological.staging ~., data = data\_hcv)

predict\_svm1 <- predict(model\_svm1, data\_hcv)

confusionMatrix(predict\_svm1, data\_hcv$Baselinehistological.staging)

library(party)

model\_dt1<- ctree(Baselinehistological.staging ~., data =data\_hcv)

predict\_dt1<- predict(model\_dt1, data\_hcv)

confusionMatrix(predict\_dt1, data\_hcv$Baselinehistological.staging)

model\_nb1<- naiveBayes(Baselinehistological.staging ~., data =data\_hcv)

predict\_nb1 <- predict(model\_nb1, data\_hcv)

confusionMatrix(predict\_nb1, data\_hcv$Baselinehistological.staging

library(class)

model\_knn1<- knn(data\_hcv, data\_hcv, cl=data\_hcv$Baselinehistological.staging, k=4)

confusionMatrix(model\_knn1, data\_hcv$Baselinehistological.staging)

Appendix 7:

1) New Feature Construction- Clustering Baseline Histological Staging into 2 stages.

2) Hold-out method Implementation- Splitting Dataset into training and testing.

3) Different Models (kNN, Naïve Bayes, etc.) Reconstruction and reevaluation using hold-out method.

set.seed(12345)

data\_hcv$Baselinehistological.staging <- as.factor(ifelse(data\_hcv$Baselinehistological.staging == 1 | data\_hcv$Baselinehistological.staging == 2, 1, 2))

sample1<- sample(nrow(data\_hcv), 0.90\*nrow(data\_hcv))

data\_hcv.train <- data\_hcv[sample1,]

data\_hcv.test <- data\_hcv[-sample1,]

model\_nb<- naiveBayes(Baselinehistological.staging ~., data = data\_hcv.train)

predict1 <- predict(model\_nb, data\_hcv.test)

confusionMatrix(predict1, data\_hcv.test$Baselinehistological.staging)

model <- knn(data\_hcv.train, data\_hcv.test, cl=data\_hcv.train$Baselinehistological.staging, k=4)

confusionMatrix(model, data\_hcv.test$Baselinehistological.staging)

library(e1071)

library(caret)

model\_svm<- svm(Baselinehistological.staging ~., data = data\_hcv.train)

predict\_svm<- predict(model\_svm, data\_hcv.test)

confusionMatrix(predict\_svm, data\_hcv.test$Baselinehistological.staging)

library(party)

model\_dtree<- ctree(Baselinehistological.staging ~., data = data\_hcv.train)

predict\_dtree<- predict(model\_dtree, data\_hcv.test)

confusionMatrix(predict\_dtree, data\_hcv.test$Baselinehistological.staging)

test\_nn <- data\_hcv.test

train\_nn <- data\_hcv.train

train\_nn$Baselinehistological.staging <- as.numeric(train\_nn$Baselinehistological.staging)

test\_nn$Baselinehistological.staging <- as.numeric(test\_nn$Baselinehistological.staging)

model<- neuralnet(Baselinehistological.staging ~., data =train\_nn)

plot(model)

predict1<- neuralnet::compute(model,test\_nn)

nn.strength<- predict1$net.result

cor(nn.strength, as.numeric(test\_nn$Baselinehistological.staging))

model<- neuralnet(Baselinehistological.staging ~., data = train\_nn, hidden = c(5), learningrate =0.0001, rep = 1, threshold = 0.5 )

plot(model)

predict1<- neuralnet::compute(model, test\_nn)

nn.strength<- predict1$net.result

cor(nn.strength, as.numeric(test\_nn$Baselinehistological.staging)

model <- glm(Baselinehistological.staging ~., data = data\_hcv.train, family = c("binomial"))

predict1 <- predict(model, data\_hcv.test)

predict1 <- as.factor(ifelse(predict1>0.5, 2,1))

confusionMatrix(predict1, data\_hcv.test$Baselinehistological.staging)

summary(model)

ideal.mod.ref<- step(model, direction = "backward")

summary(ideal.mod.ref)

pred.glm<- predict(ideal.mod.ref, data\_hcv.test, type = "response")

pred.glm <- as.factor(ifelse(pred.glm>0.5, 2,1))

confusionMatrix(pred.glm, data\_hcv.test$Baselinehistological.staging)

Appendix 8: Bagging and Boosting for increasing accuracy.

library(mlbench)

library(caret)

library(caretEnsemble)

control <- trainControl(method="repeatedcv", number=5, repeats=3)

seed <- 7

metric <- "Accuracy"

set.seed(seed)

fit.c50 <- train(Baselinehistological.staging~., data=data\_hcv, method="C5.0", metric=metric, trControl=control)

set.seed(seed)

fit.gbm <- train(Baselinehistological.staging~., data=data\_hcv, method="gbm", metric=metric, trControl=control, verbose=FALSE)

boosting\_results <- resamples(list(c5.0=fit.c50, gbm=fit.gbm))

summary(boosting\_results)

dotplot(boosting\_results)

control <- trainControl(method="repeatedcv", number=5, repeats=3)

seed <- 7

metric <- "Accuracy"

set.seed(seed)

fit.treebag <- train(Baselinehistological.staging~., data=data\_hcv, method="treebag", metric=metric, trControl=control)

set.seed(seed)

fit.rf <- train(Baselinehistological.staging~., data=data\_hcv, method="rf", metric=metric, trControl=control)

bagging\_results <- resamples(list(treebag=fit.treebag, rf=fit.rf))

summary(bagging\_results)

dotplot(bagging\_results)