IE 7374 – Machine Learning in Engineering

Project Report

Classification of patients with Hepatitis C Virus

Submitted by :

Project Group 4

Sai Bhavana Atluri

Meghana Morey

Chaitanya Swan

Shah Zeb

TABLE OF CONTENTS

1. Abstract

A virus is a small collection of genetic code, either DNA or RNA, surrounded by a protein coat. A virus cannot replicate , they must infect cells and use components of the host cell to make copies of themselves. Often, they kill the host cell in the process, and cause damage to the host organism. Hepatitis C virus is one such virus that causes liver infection on the host. It can be spread through contact with the blood of the infected person. This virus can lead to either a short-term illness or a long-term illness (chronic infection). More than half of the people infected with Hep C have a chronic infection and most of the times they are asymptomatic. When symptoms appear, they often are a sign of advanced liver disease. A chronic illness can lead to serious, even life-threatening health problems like cirrhosis and liver cancer and the sad part is that there is no vaccine. Best way to prevent it is by avoiding behaviours that can spread the disease, especially injecting drugs. Good part is that Hep C is treatable, and treatments can cure most people with hepatitis C in 8 to 12 weeks. So, to help cure Hepatitis C, getting tested for it is important. In this project, we aim to build a model to classify the patients with Hep C into different stages based on the severity of the symptoms and various other aspects related to their health so that they can be administered with appropriate treatment.

2. Problem Description

Hepatitis C is a liver infection caused by the hepatitis C virus (HCV). In some cases, it is curable while it is fatal in others. These days most people are prone to be infected due to the usage of common medical equipment such as needles that are in direct contact with blood. For the patients undergoing the treatment for HCV, they are classified into stages based on the severity and the various other physical aspects of the body.

The aim of the project is to study all the features leading towards severity and provide a clear classification of the stages so that the patient receives proper care based on the need. Here the patients are intended to be classified into baseline histological stage of the liver from stages Portal Fibrosis, Few Septa, Many Septa and Cirrhosis. Where, serious damage is expected in case of cirrhosis.

3. Data source and description

Dataset used: Hepatitis C Virus (HCV) patient's dataset

https://archive.ics.uci.edu/ml/datasets/Hepatitis+C+Virus+%28HCV%29+for+Egyptian+patients

Data Description:

The data that is being discussed here is collected from patients who underwent treatment dosages for HCV for about 18 months from Egypt. Based on the expert recommendations and domain knowledge, we are looking at the following attributes and their impact on classification of stages of the disease.

The following are the prime characteristics of the dataset considered for the study,

- No. of records = 1385 nos.
- Response variable , Baseline histological staging , No. of classes = 4 classes, i.e., 1 – Portal Fibrosis , 2 – Few Septa , 3 – Many septa , 4 - Cirrhosis
- Predictor variables, No. of variables = 28 Variables

Feature Names	Feature Values	Discretization (Items)	Feature description
Age	32:61	[0; 32],]32; 37],]37; 42],]42; 47],]47; 52],]52; 57],]57; 62]	Provides the age of the patient
Gender	Male, Female	[Male], [Female]	Specifies the gender of the patient
BMI(Body Mass Index)	22:35	[0; 18:5[[18:5; 25[, [25; 30[, [30; 35[, [35; 40[	Gives the measure of Body Mass Index of patient
Fever	Absent, Present	[Absent], [Present] -	Specifies if the patient has fever or not
Nausea/Vomiting	Absent, Present	[Absent], [Present] -	specifies if the patient has had history of vomiting or not
Headache	Absent, Present	[Absent], [Present] -	specifies if the patient has had history of headache or not
Diarrhea	Absent, Present	[Absent], [Present] -	specifies if the patient has had history of diarrhea or not
Fatigue	Absent, Present	[Absent], [Present] -	Specifies if the patient has had history of fatigue or not
Bone ache	Absent, Present	[Absent], [Present] -	Specifies if the patient has history of bone ache or not
Jaundice	Absent, Present	[Absent], [Present] -	Specifies if the patient has history of fatigue or not
Epigastria pain	Absent, Present	[Absent], [Present] -	Specifies if the patient has history of epigastria pain or not
WBC(White Blood Cells)	2991:12101	[0; 4000[, [4000; 11000[, [11000; 12101]	WBC count from blood test
RBC(Red Blood Cells)	3816422:5018451	[0; 3000000[, [3000000; 5000000[,[5000000; 5018451]	RBC count from blood test
HGB (Hemoglobin)	2:20	If (Gender==[Male]):[2; 14[, [14; 17:5],]17:5; 20] If(Gender==[Female]):[2; 12:3[, [12:3; 15:3],]15:3; 20]	HGB count from blood test
Plat(Platelet)	93013:226464	[93013; 100000[, [100000; 255000[,[255000; 226465[	Platelet count from blood test
AST1(1 week)	0.088888889	[0; 20[, [20; 40],]40; 128]	Aspartate transaminase ratio
ALT1(1 week)	0.088888889	[0; 20[, [20; 40],]40; 128]	Alanine Transaminase ratio 1 weeks
ALT4(4 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]	alanine transaminase ratio 4 weeks
ALT12(12 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]	alanine transaminase ratio 12 weeks

ALT24(24 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]	alanine transaminase ratio 24 weeks
ALT36(36 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]	alanine transaminase ratio 36 weeks
ALT48(48 weeks)	0.088888889	[0; 20[, [20; 40],]40; 128]	alanine transaminase ratio 48 weeks
RNA Base	0:1201086	[0; 5],]5; 1201086]	RNA Base
RNA 4	0:1201715	[0; 5],]5; 1201715]	RNA at 4 weeks
RNA 12	0:3731527	[0; 5],]5; 3731527]	RNA at 12 weeks
RNA EOT	0:808450	[0; 5],]5; 808450]	RNA at End Of Treatment
RNA EF(Elongation Factor)	0:808450	[0; 5],]5; 808450]	RNA Elongation factor
Baseline Histological Grading	1:16	[1]; [2]; [3]; :::[16]	Grading of the stage of disease
Baseline Histological	F0:F4	[No Fibrosis], [Portal Fibrosis],Staging (Class Label) [Few Septa], [Many Septa], [Cirrhosis]	Stage of disease, i.e., disease classification

Figure : Snapshot of the dataset

	Age	Gender	BMI	Fever	Nausea/Vomting	Headache	Diarrhea	Fatigue & generalized bone ache	Jaundice	Epigastric pain	...	ALT 36	ALT 48	ALT after 24 w	RNA Base	RNA 4	RNA 12
0	56	1	35	2	1	1	1	2	2	2	...	5	5	5	655330	634536	288194
1	46	1	29	1	2	2	1	2	2	1	...	57	123	44	40620	538635	637056
2	57	1	33	2	2	2	2	1	1	1	...	5	5	5	571148	661346	5
3	49	2	33	1	2	1	2	1	2	1	...	48	77	33	1041941	449939	585688
4	59	1	32	1	1	2	1	2	2	2	...	94	90	30	660410	738756	3731527
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1380	44	1	29	1	2	2	2	1	1	1	...	63	44	45	387795	55938	5
1381	55	1	34	1	2	2	1	1	1	1	...	97	64	41	481378	152961	393339
1382	42	1	26	2	2	1	1	1	2	1	...	87	39	24	612664	572756	806109
1383	52	1	29	2	1	1	2	2	2	1	...	48	81	43	139872	76161	515730
1384	55	2	26	1	2	2	2	1	2	1	...	64	71	34	1190577	628730	5

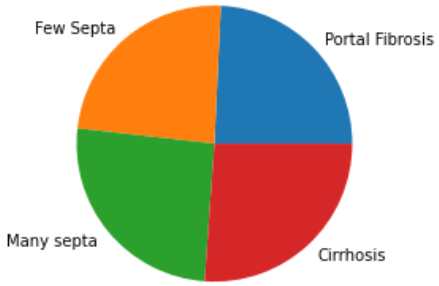
4. Exploratory Data Analysis (EDA)

One of the initial stages of a data analytics or Machine learning project is that of an exploratory data analysis. Likewise, for this project on classification of stages of patient suffering from Hepatitis C virus, an extensive exploratory data analysis is conducted. So for this project, we began this by analysing the nature of the predictor variables. As can be seen from, Figure 1, the variables in the dataset are the physical details of patient such as age, gender along with details from the blood test reports that includes, RBC, WBC, Platelet etc.,. which gives an integer or float value as best suited for the variable of interest. Also, present is the data on history of any symptoms tested for such as fever, headache, pain etc., this data is binary in nature, i.e., yes or no. So, we can conclude that the dataset considered here for the study has a mixture, of binary and continuous data, with response variable of categorical nature. Then, we performed a check for missing values, and identified that there are no missing values in the dataset.

Post identifying the nature of dataset, it is now the time to check for equal representation of classes of response variable. Therefore, for establishing this, an attempt was made to check on the variables for balanced class representation, it is as follows,

Class	Count	% of class representation
1	336	24.26
2	332	23.97
3	355	25.63
4	362	26.13

Figure : Visualization depicting the class representation



Further, an attempt is made towards identifying correlation between predictor variables. Some of the visualizations like pair plot, Heat map and Correlogram have been explored to identify these relationships.

Figure : Pair plot for the dataset

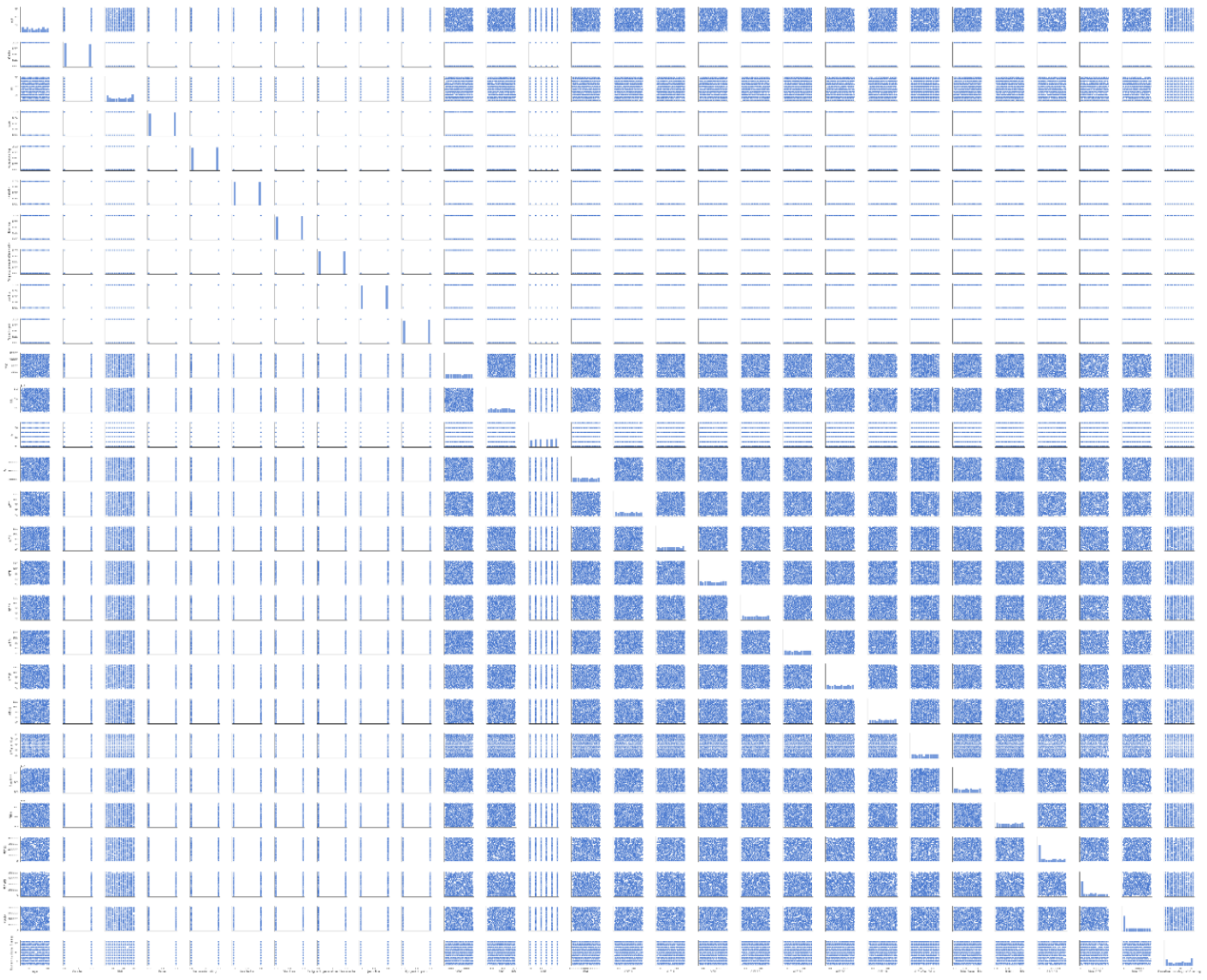
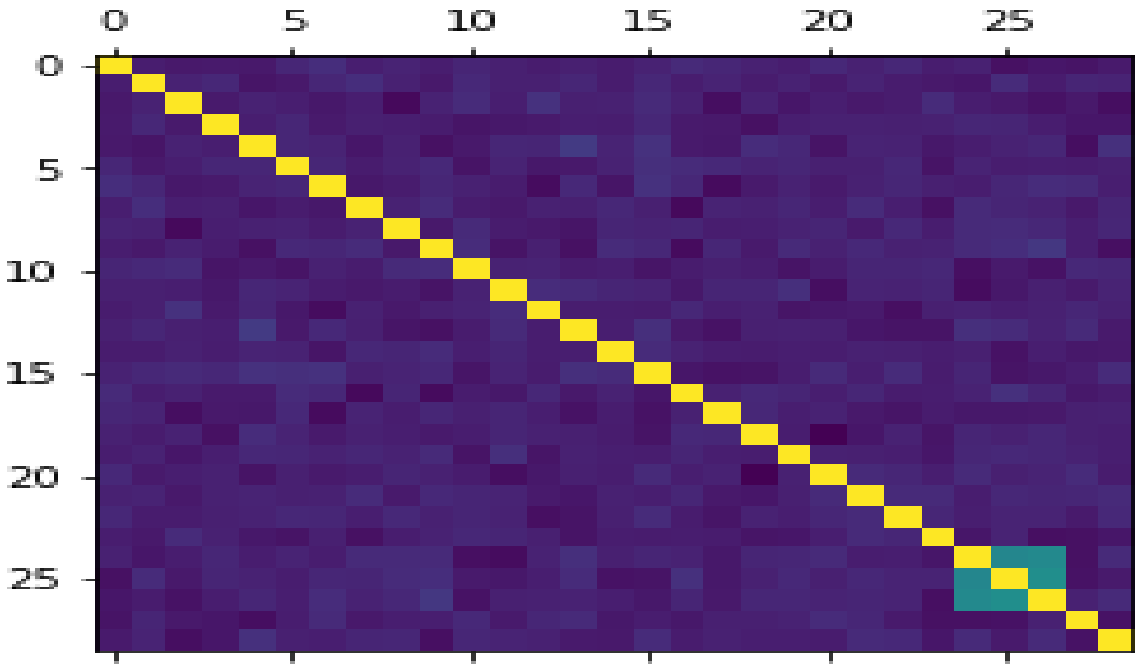
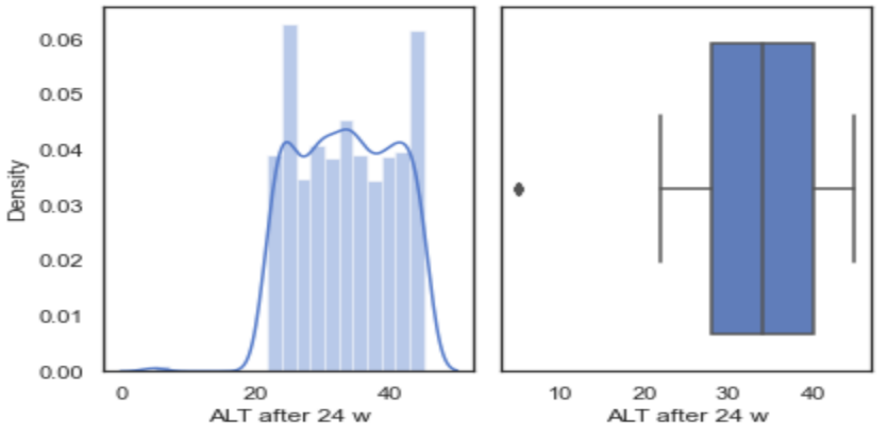
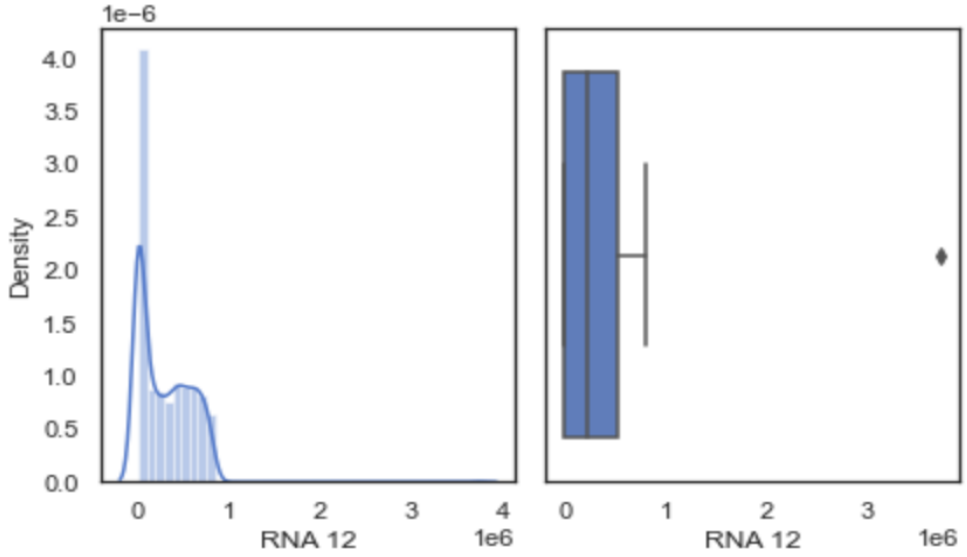


Figure : Correlogram for the dataset



As a part of further exploration of data, we check for outliers in the data that seem to influence determining the correlation coefficient of the variables. For, determining this, we resorted to perform a visual analysis using box plots. So, we were able to identify outliers in the dataset, and decided upon eliminating the rows of data that had highly skewed outliers. We have even tried visualizing the distribution of the dataset, and identified that most of the variables followed normal distribution, whereas the distribution looks skewed for the variables with outliers. The variables like, RNA 12 and Alt after 24 W had outliers.





For the ease of implementation of the project, for ALT after 24 w, we have deleted all the rows greater than 22 and for RNA 12 we have deleted all the rows containing values less than 1000000. Now, the dataset contains 1324 rows, instead of the prior 1385 rows.

On performing the correlation analysis on the variables, we decided to remove all the variables that are highly correlated with each other. For this operation, we considered a threshold of 80%, and checked for the same, it was identified that there are no columns with such high correlation and hence all the variables will be considered for the analysis.

In the context, of easing the implementation of classification models, we decided upon binning the response variables into two bins, instead of having 4 classes based on the severity

measures. So, we have assigned 1- Portal Fibrosis and 2 – Few Septa to one class and 3 – Many septa and 4 – Cirrhosis to other class. On implementing and reviewing the same, it was identified that the data belonging to each of the bins is equally distributed, i.e., the data is sampled equally and there will not be any further need to oversample the data.

On checking for correlation, from the variables set available, the following variables, 'BMI', 'Nausea/Vomiting', 'Epigastric pain', 'ALT 1', 'ALT after 24 w', 'RNA 4', 'RNA 12' and 'RNA EF' seem to have considerably a high correlation with the response variables over the rest.

The following are some of the plots visualised to check for representation of data within each of the variables,

Figure : BMI

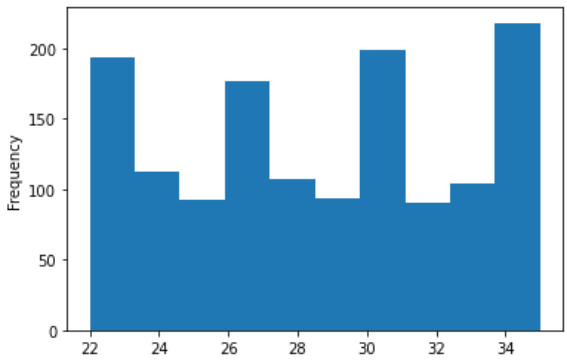


Figure : Age

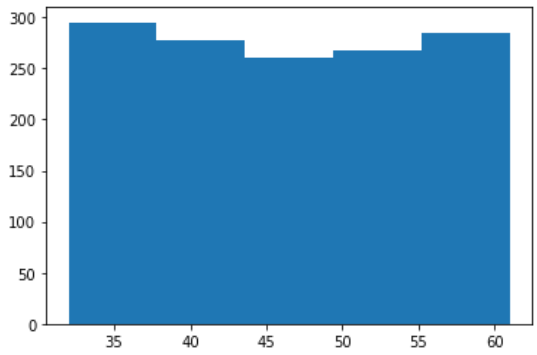


Figure : RNA 4

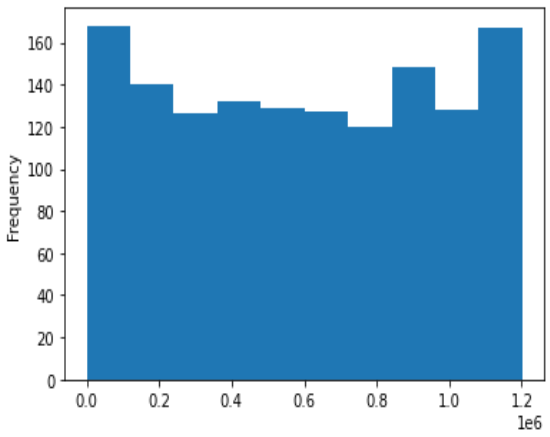
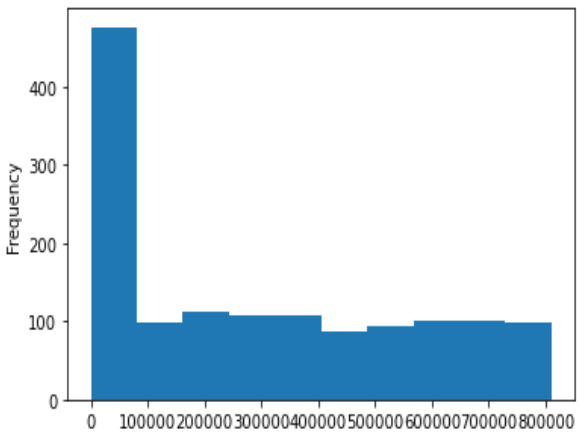


Figure : RNA EF



5. Model Selection and Implementation

One of the first tasks that is thought of before deploying any models is partitioning the data available at hand. Here, for the study considering to partition the data into training dataset and test dataset. The partitioning followed the usual 70:30 rule, where 70% of the data was used for training the model and 30% of the data was used for testing.

Considering the dataset available for the study, this is a classification Data Mining models such as logistic regression, Bayesian classifiers, decision tree, Neural networks, K- Nearest Neighbour, support vector machines etc., could be used for classification purposes, in this case the classification of stages of Hepatitis C virus, into two bins.

On a brief understanding of the Machine Learning algorithms and the data available at hand, the algorithms like Logistic regression, K- Nearest Neighbour model , Naïve Bayes model and Hard margin SVM are considered for this study. Taking into account the complexity of the models and runtime, we concluded that the above-mentioned models could give out the best possible results. The models deployed and implemented for the study are discussed below.

Logistic Regression

Logistic regression, is a predictive modelling algorithm that is used when the Y variable is binary or categorical. This algorithm determines a mathematical equation that predicts the probability of Event 1 and once the equation is determined, it can predict the Y value for the given X values.

Logistic regression is applied on our data as we believe that it is a good starting point for the kind of the dataset that we have. The model implemented, determines the severity of the

Hepatitis C as Cirrhosis or Portal Fibrosis**.** The outcome variable is coded as 1 for Cirrhosis and 0 for Portal Fibrosis. There are 28 predictors ranging from fever, nausea, headache, and several other health factors.

The model has been deployed and checked for a few performance metrics such as accuracy, precision and recall. Further, as a part to better understand the model and the prediction patterns, it is evaluated on various sizes of the test data such as 10%, 20%, 30%, 40% of the dataset and the model yielded the results are as follows:

Test size: 10.0 %

Solving using gradient descent
100%|■■■■■■■■■■| 10000/10000 [00:02<00:00, 4461.23it/s]
Accuracy: 0.57
Precision: 0.59
Recall: 0.58

Evaluate for Testing data:

Accuracy: 0.46
Precision: 0.48
Recall: 0.37

Test size: 20.0 %

Solving using gradient descent
100%|■■■■■■■■■■| 10000/10000 [00:01<00:00, 5126.38it/s]
Accuracy: 0.58
Precision: 0.6
Recall: 0.58

Evaluate for Testing data:

Accuracy: 0.48
Precision: 0.49
Recall: 0.44

Test size: 30.0 %

Solving using gradient descent
100%|■■■■■■■■■■| 10000/10000 [00:01<00:00, 5571.03it/s]
Accuracy: 0.58
Precision: 0.6
Recall: 0.58

Evaluate for Testing data:

Accuracy: 0.48
Precision: 0.51
Recall: 0.46

Test size: 40.0 %

Solving using gradient descent
100%|■■■■■■■■■■| 10000/10000 [00:01<00:00, 6031.78it/s]
Accuracy: 0.58
Precision: 0.59
Recall: 0.57

Evaluate for Testing data:

Accuracy: 0.5
Precision: 0.53
Recall: 0.48

Test size: 50.0 %

Solving using gradient descent
100%|■■■■■■■■■■| 10000/10000 [00:01<00:00, 7107.22it/s]
Accuracy: 0.58
Precision: 0.59
Recall: 0.57

Evaluate for Testing data:

Accuracy: 0.51
Precision: 0.55
Recall: 0.47

For the various % of test sizes, it can be seen that the run time of the model kept increasing for the increase in the data considered for test set. On seeing the metrics, a 30% of dataset yields better performance metrics.

K- Nearest Neighbour

The KNN algorithm assumes that there exists similarity in the data and better segregate the unlabelled data points into well-defined groups. Choosing the number of nearest neighbours i.e., determining the value of k plays a significant role in determining the efficacy of the model. Thus, selection of k will determine how well the data can be utilized to generalize the results of the KNN algorithm. A large k value has benefits which include reducing the variance due to

the noisy data; the side effect being developing a bias due to which the learner tends to ignore the smaller patterns which may have useful insights.

Since we only have 1385 instances in our dataset, another good model to implement is the KNN Algorithm which is great for smaller data. This model is applied on our data to determine the severity of the Hepatitis C as Cirrhosis or Portal Fibrosis**.** The outcome variable is coded as 1 for Cirrhosis and 0 for Portal Fibrosis. There are 28 predictors ranging from fever, nausea, headache, and several other health factors.

The results of the model are as follows:

```
Neighbours 1
Evaluate for Test data:

Accuracy: 0.53
Precision: 0.54
Recall: 0.56

Neighbours 2
Evaluate for Test data:

Accuracy: 0.49
Precision: 0.53
Recall: 0.28

Neighbours 3
Evaluate for Test data:

Accuracy: 0.5
Precision: 0.52
Recall: 0.53

Neighbours 4
Evaluate for Test data:

Accuracy: 0.51
Precision: 0.54
Recall: 0.36

Neighbours 5
Evaluate for Test data:

Accuracy: 0.52
Precision: 0.54
Recall: 0.52

Neighbours 6
```

```
Evaluate for Test data:

Accuracy: 0.52
Precision: 0.55
Recall: 0.41

Neighbours 7
Evaluate for Test data:

Accuracy: 0.51
Precision: 0.53
Recall: 0.57

Neighbours 8
Evaluate for Test data:

Accuracy: 0.47
Precision: 0.49
Recall: 0.38

Neighbours 9
Evaluate for Test data:

Accuracy: 0.5
Precision: 0.52
Recall: 0.52

Neighbours 10
Evaluate for Test data:

Accuracy: 0.49
Precision: 0.52

Recall: 0.41
```

Naïve Bayes

Bayes' Theorem provides a way that we can calculate the probability of a piece of data belonging to a given class, given our prior knowledge. Bayes' Theorem is stated as:

P(class|data) = (P(data|class) * P(class)) / P(data)

where P(class|data) is the probability of class given the provided data.

Naïve Bayes is based on Bayes' Theorem with an assumption of independence among predictors.

Naïve bayes model is also applied on our data to determine the severity of the Hepatitis C as Cirrhosis or Portal Fibrosis as most of our predictors are categorical**.** The outcome variable is coded as 1 for Cirrhosis and 0 for Portal Fibrosis. There are 28 predictors ranging from fever, nausea, headache, and several other health factors. Detailed list of predictors is described in data source description above.

The model is evaluated on various sizes of the test data as 10%, 20%, 30%, 40%, 50% of the data and the results are as follows:

```
Test size: 10.0 %

Evaluate for trainning data:

Accuracy: 0.56
Precision: 0.57
Recall: 0.63


Evaluate for Testing data:

Accuracy: 0.46
Precision: 0.48
Recall: 0.44


Test size: 20.0 %

Evaluate for trainning data:

Accuracy: 0.58
Precision: 0.58
Recall: 0.65


Evaluate for Testing data:

Accuracy: 0.46
Precision: 0.48
Recall: 0.5


Test size: 30.0 %

Evaluate for trainning data:

Accuracy: 0.59
Precision: 0.6
Recall: 0.65
```

```
Evaluate for Testing data:

Accuracy: 0.48
Precision: 0.5
Recall: 0.51


Test size: 40.0 %

Evaluate for trainning data:

Accuracy: 0.59
Precision: 0.59
Recall: 0.6


Evaluate for Testing data:

Accuracy: 0.49
Precision: 0.52
Recall: 0.48


Test size: 50.0 %

Evaluate for trainning data:

Accuracy: 0.61
Precision: 0.61
Recall: 0.61


Evaluate for Testing data:

Accuracy: 0.5
Precision: 0.54

Recall: 0.48
```

Hard Margin SVM

SVM is a supervised Machine Learning Model in which we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. We perform classification by finding the hyper-plane that differentiates the two classes efficiently. In addition, distance between the hyperplane to the closest x, or the margin, must be maximized.

As we are dealing with critical data pertaining to human life, one misclassification could lead to giving wrong treatment which could be life threatening, we do not want any misclassifications. One good model to avoid misclassifications is the Hard Margin SVM and so it is also applied on our data to determine the severity of the Hepatitis C as Cirrhosis or Portal Fibrosis**.** The outcome variable is coded as 1 for Cirrhosis and 0 for Portal Fibrosis. There are 28 predictors ranging from fever, nausea, headache, and several other health factors. Detailed list of predictors is described in data source description above.

the results are as follows:

```
Test size: 10.0 %

Training Data
Accuracy: 0.56
Precision: 0.57
Recall: 0.63

Testing Data
Accuracy: 0.48
Precision: 0.51
Recall: 0.46

Test size: 20.0 %

Training Data
Accuracy: 0.57
Precision: 0.57
Recall: 0.64

Testing Data
Accuracy: 0.47
Precision: 0.49
Recall: 0.51

Test size: 30.0 %

Training Data
Accuracy: 0.58
Precision: 0.58
Recall: 0.64

Testing Data
Accuracy: 0.48
Precision: 0.5
Recall: 0.5
```

```
Test size: 40.0 %

Training Data
Accuracy: 0.58
Precision: 0.58
Recall: 0.6

Testing Data
Accuracy: 0.52
Precision: 0.55
Recall: 0.5

Test size: 50.0 %

Training Data
Accuracy: 0.59
Precision: 0.59
Recall: 0.59

Testing Data
Accuracy: 0.5
Precision: 0.53
Recall: 0.47
```

6. Model outcomes and Inferences

The performance metrics we used to compare the models are accuracy, precision, and recall. Accuracy specifies us the measure of a percentage of correct predictions for the train and test data. for the dataset under consideration, Logistic regression, Naïve bayes gives and Hard margin SVM gives the test accuracy of 48% while KNN gives 52%.

Precision is an indicator of quality of a positive prediction made by the model. For the given dataset Logistic regression and Naïve Bayes gives a precision of 0.6, Hard margin 0.58 and KNN gives 0.54.

Recall is an indicator of the sensitivity; it indicates how many of the true positives were recalled. For our dataset, Logistic regression has a recall of 0.46, Naïve Bayes has 0.51, Hard Margin SVM has a recall of 0.5 and KNN has a recall of 0.56.

Below is a table comparing the Accuracy, precision and recall for Logistic regression, Naïve Bayes, KNN and Hard Margin SVM with the test and train data split at 30%.

LOGISTIC REGRESION	NAÏVE BAYES	KNN	HARD MARGIN SVM
Training data: Accuracy: 0.58 Precision: 0.6 Recall: 0.58 **Testing data:** Accuracy: 0.48 Precision: 0.51 Recall: 0.46	**training data:** Accuracy: 0.59 Precision: 0.6 Recall: 0.65 **Testing data:** Accuracy: 0.48 Precision: 0.5 Recall: 0.51	**Neighbours** 6 **Test data:** Accuracy: 0.52 Precision: 0.55 Recall: 0.41	**Training Data** Accuracy: 0.58 Precision: 0.58 Recall: 0.64 **Testing Data** Accuracy: 0.48 Precision: 0.5 Recall: 0.5

7. Discussion of Results

The best model would be the KNN model as it gives the highest accuracy of 0.53. For the models selected, as both the train and test accuracy is close, the model is neither overfitting nor underfitting as its error is low and bias variance are at their lowest.

We have calculated our evaluation metrics for different test and train splits ranging from 10%, 20%, 30%, 40%, and 50%. This this process, we have observed that the metrics have always been stable and there are no fluctuations. So, we can say that our model is generalized.

8. Conclusion

In this project we have implemented four different models i.e., Logistic Regression, K Nearest Neighbours (KNN), Naïve Bayes, and Hard Margin SVM to predict if the patients' symptoms are either Portal Fibrosis or Cirrhosis based on different attributes pertaining to their health and finalised that the best performing model would be KNN. This model is useful in the healthcare sector to determine which stage of Hepatitis C the patient is in so that they can get appropriate treatment.

9. Future scope

In our current project we are dealing with a binary classification problem as we are predicting if the patients' symptoms are Cirrhosis or Portal Fibrosis. For a future extension of this project, we can build prediction models that can tackle multiclass problems i.e., we can build a model to determine which stage (1 – Portal Fibrosis , 2 – Few Septa , 3 – Many septa , 4 - Cirrhosis) their symptoms are at. This will help in making the treatment process even quicker and there can be a considerable decrease in the fatality rate.

Links:

Github:

https://github.com/chaitanyaswan/Machine-Learning/tree/master/IE_7374_Final_Project_Group4

Dataset:

https://drive.google.com/file/d/1GALcBNy801BE3BdzHCaEbfhcA6BPUtTA/view?usp=sharing

google colab file:

https://colab.research.google.com/drive/1aHfVZTd1OgYj9RXPDCWHMtnsC7gt0qLw?usp=sharing