PREDICTING HEPATITIS C VIRUS USING MACHINE LEARNING

BY

Shahida Akter
ID: 191-15-2500

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. S.M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2022

APPROVAL

This Project/internship titled **““Predicting Hepatitis C Virus Using Machine Learning ”**, submitted by Shahida Akter, ID No: 191-15-2500 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *05 February 20233.*

BOARD OF EXAMINERS

8|2|23

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Raja Tariqul Hasan Tusher
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Mr. Mushfiqur Rahman
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Ahmed Wasif Reza
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Dr. S.M. Aminul Haque, Associate Professor, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. S.M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

20.2.23

Mohammad Jahangir Alam
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:

Shahida

Shahida Akter
ID: 191-15-2500
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish our profound our indebtedness to Dr. S.M. Aminul Haque, Associate Professor Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Data science" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude to Professor Dr. Touhid Bhuiyan, Head Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

A virus that causes called hepatitis C damages and incites the liver. Enlargement triggered by inflammation happens whenever bodily parts are hurt and diseased. Body parts could be harmed by irritation. Most of the world's cases of hepatitis C are found in countries like Egypt. According to estimates, there are 3–4 million new cases each year, making it a public health concern that needs to be addressed with treatment and screening programs. In recent years, diagnostics-based technology has dramatically advanced. AI systems are able to create diagnostic models from a wide range of unusually complicated structures. I employed machine learning methods in this research to assess the presence or absence of the hepatitis C virus. The availability of very accurate risk prediction models would make it easier to identify those who need more intensive monitoring and treatment ahead of time. Individuals with chronic hepatitis C (CHC), the most common cause of cirrhosis worldwide, would benefit most from risk prediction models. Despite the availability of efficient antiviral treatment for CHC, the disease has yet to be eradicated. There are six ML algorithms are applied that are used Logistic Regression, K Neighbors Classifier, Random Forest Classifier, Cat Boost Classifier, Decision Tree Classifier, and Gradient Boosting Classifier. The gradient-boosting algorithm generated a number of the top 6 results. The answer is 94.31 percent. I also find out the confusion matrix of these algorithms and use a correlation matrix to calculate the following numbers: Individuals who are suspicious 5402. 75 healthy patients, 61.30% of whom are men and 38.70% of whom are women.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF Tables

CHAPTER 1

Introduction

1.1. Introduction

The hepatitis C virus is what produces the liver illness known as HCV. Interaction involving plasma out of an infected individual can transmit hepatitis C. These days, permissible- or even other injecting gear is how the majority of persons contract the virus hepatitis C.Antiviral medications may treat and over 95percent of people having hepatitis C virus, but accessibility to treatment and prevention is limited. According to the WHO WHO, 290 thousand individuals died from hepatitis C in 2019, largely from cirrhosis and hepatocellular carcinoma.[1]

Chronic and acute hepatitis C are both possible:

1. The disease known as acute hepatitis C is transient. As much as six months may pass between signs. Quite often the virus disappears because the system is capable of successfully combat this illness. However, the severe infection usually develops into a persistent infection.

1. Hepatitis C that is persistent is a persistent disease. If left untreated, it may follow a lifelong course & lead to serious health issues, such as liver failure, fibrosis (hepatic fibrosis), liver disease, or even mortality.[2]

Hepatitis C can take anywhere from two weeks to six months to fully develop. About 80 percent of the total of patients do not show any signs after the primary outbreak. Severe symptoms include:

1. exhaustion,
2. fever,
3. a lack of appetite,

4. light feces, vomiting,
5. nausea, muscle aches,
6. abdominal discomfort,
7. blood in the urine, 8. discoloration of the skin and
9. the whites of the eyes.

People are more likely to contract hepatitis C if what:

1. Are indeed a medical professional that has come into contact with contaminated plasma, that can occur when a contaminated syringe penetrates the body.
2. Have you previously shot and smoked unlawful substances?
3. Own HIV
4. Used to have a tattooing and pierce with unsanitary tools or within an unsanitary setting
5. Had a kidney transplant or had a transfusion prior to 1992
6. Until 1987, got coagulation factors concentrations
7. Benefited from dialysis for a considerable amount of time.
8. Were delivered to the a mother who had hepatitis C 9. Ever were incarcerated

were aged during 1945 to 1965, the age range in which hepatitis C infection occurs most frequently.

Patients with chronic liver disease should be concerned about the development of cirrhosis and the consequences that come with it. Cirrhosis progression rates might vary a lot from person to person. 1 and 2 Slow progress may be subjected to excessive monitoring, while rapid progress may be subjected to insufficient monitoring and treatment due to a lack of proper risk classification. The availability of very accurate risk prediction models would make it easier to identify those who need more intensive monitoring and treatment ahead of time. Individuals with chronic hepatitis C (CHC), the most common cause of cirrhosis worldwide, would benefit most from risk prediction models. Despite the availability of efficient antiviral treatment for CHC, the disease has yet to be eradicated.

I subsequently validated these models in a small cohort of individuals with minimal or no fibrosis and without prior treatment exposure. It remains unknown whether predictive modeling can accurately predict CHC progression in a large or heterogeneous sample outside of clinical trial settings.

Further, in our prior studies, there is a possibility of not obtaining the best prediction accuracy if we treat the cirrhosis outcome as a dichotomous outcome at a specified interval of time. Thus, in this analysis, I sought to address these questions by applying a time-to-event analysis to provide more accuracy.

1.2 Motivation

I use methods of machine learning to forecast the hepatitis C virus. If hepatitis C patients' symptoms are predicted in a timely way and particularly risky practices are eliminated, they can benefit from earlier diagnosis. So It encouraged that the condition was still in its very early stages and that there was little awareness of it. The disease is also exceedingly expensive to treat when it is severe, and early detection can lessen its global burden.

1.3 Objectives

Data mining and machine learning for hepatic disorders Ml has been used by many scientists to diagnose hepatic disorders, some of them to anticipate & stage fibrosis. In a study conducted, decision tree algorithm (dt) and nb were employed to determine the variables that would increase the risk of HCV vertical transmission amongst newborn Egyptian children. decision tree, genetic algorithms, multi-linear regression, and optimization using particle swarms were utilized to forecast severe fibrosis in adults by integrating blood indicators and medical evidence, again for the Egyptian population and in a prospective analysis for 39,569 Patients with hiv infection [3]. An AUC around 0.73 & 0.76 as well as an accuracy range between 66% to 84% were used to identify progressive fibrosis. The early diagnosis of severe illnesses & chronic illnesses may be aided by artificial Intelligence illness diagnostic or prediction systems. The goal of my research is to create a machine learning-based intelligent health monitoring system for such timely identification of hepatitis C.

1.4 Expected Outcome

1. I identify the most accurate and effective algorithm with accuracy.
2. I also find out the confusion matrix of these algorithms
3. Comparison with previous research work.
4. Calculate whether the following numbers represent suspicious patients or healthy patients using a correlation matrix.

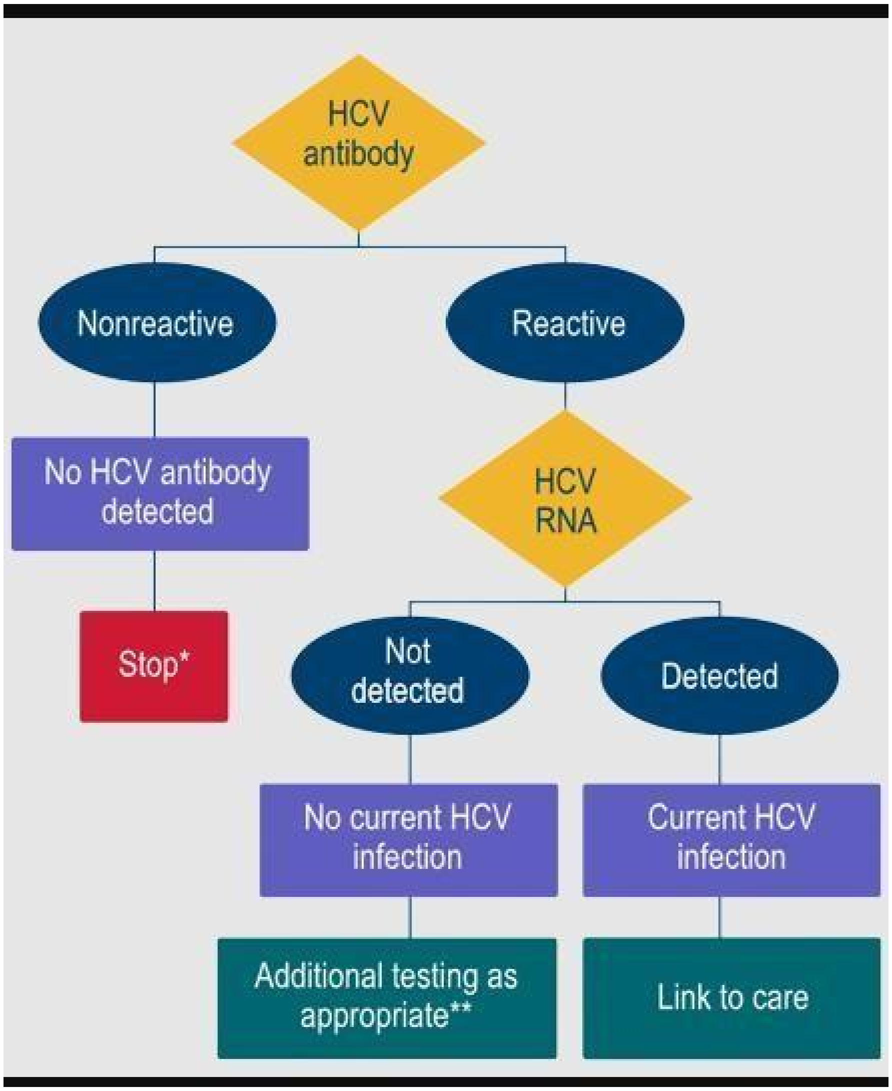


Figure 1.1: Expected outcome map

CHAPTER 2

Literature review

To detect hepatitis B, Chen et al. used SVM, k-nearest neighbor, & naive bayes (nb). With 513 individuals, 11 characteristics using genuine tissue elastography (rte) photos were used. In respect of the liver fibrosis score, the authors have provided enhanced results using rf, which surpassed previous approaches [4].

Safdari et al. suggested an ML-based system that can categorize the phases of the hepatic in hepatitis C disease individuals. From the UCI machine learning repository, researchers collected cases of hepatic fibrosis illness in Egyptian patients. To find significant features of the hepatitis C virus in this dataset, they used various selection techniques. Researchers stated that this outcome had already been helpful for examining & making decisions regarding the chronic virus that causes hepatitis C. Nearly 1385 hospital data with HCV infection were used in the investigation. Fake minority up sampling technology, which adds artificial cases of individuals, has been employed to equalize occurrences of several groups. Subsequently, they used various feature selection techniques towards this data to find relevant hepatitis C viral traits.Because GB, NB, and LR's accuracy in these papers was less than 50%, they were disregarded in this section. These models have significantly subpar outcomes when applied to main HCV patient data; KNN has the greatest accuracy (25.48%). The original HCV information was not utilized in this study since the classifiers did not produce better outcomes. SMOTE was therefore used eight times on the original HCV data.[5]

The study by Kasahif et al. aims to predict well how individuals with hepatitis C will respond to the medication "L-ornithine L-Aspartate (LOLA)". The study examined how important deep learning methods were in helping cancer patients identify how their medications were responding. In order to forecast how drugs will affect cancer cells, the researchers closely investigated the cancerous cells. In order to anticipate how drugs will affect cancer cells, the investigators closely investigated the cancerous cells. In order to anticipate esophagus illness, which is a side consequence of liver problems, scientists [18] employ Bayesian networks.additionally, These trials are carried out using the Weka tool. The effectiveness of such methods is evaluated using many metrics, including"Fmeasures, Precision, Accuracy and Recall." This was found that the train data, KNN, K*, and RF performed well on the accuracy measure, however the test data, DT, performed well. The author did not make apparent their plans for the future, and none of the sections were written with the same clarity or quality as those in other papers. [6]

Mendlowitz et al. calculate the expenditures of healthcare and the utilization of resources due to HCV infection amongst The First Nations populations in Ontario, stratified by sex and place of living (inside or outside of First Nations communities). In this study, Status First Nations people in Ontario who tested positive for HCV antibodies or RNA between 2004 and 2014 were identified using connected wellness organizational datasets, as were Status First Nations people who did not have documents of HCV testing as well as had only a bad test result (control group, matched 2:1 to case participants). There are more descriptions of researchers using administrative data sets from ICES. To estimate 95% confidence intervals, they employed clustered bootstrap sampling, where matched sets were resampled with 1000 replications (CIs). Results were divided into groups based on sex and whether a case participant lived inside or outside a First Nations community at the time of matching.[7]

Ahammed et al. suggested an ML technique that can categorize the phases of the liver in hepatitis C virus-infected patients. The influence of liver fibrosis is examined in this research utilizing a variety of feature selection and classification techniques. In order to establish the optimal machine learning model for identifying HCV, the experiment's results can therefore be examined in great detail. Waikato Environment for Knowledge Analysis (WEKA) was used to discover key features, and several classifications were integrated into the HCV dataset using the Python-based Sci-kit Learn module. Using data balance, feature selection, and classification algorithms, the dataset of an HCV patient was examined. The dataset itself is the main issue because it is unbalanced.[8]

Syafa'ah et al. assess the level of precision in detecting HCV to use the algorithm classification machine learning. The methodology used in this research is based on AI techniques and classification methods such as k-nearest neighbors, naive Bayes, neural networks, and random forests. The goal of this study is to test and evaluate the level of accuracy in detecting the disease HCV using the algorithm classification machine learning. The UCI dataset was used in this work, and it can be solved using the KNN, naive Bayes, NN, and RF approaches. Methods NN displays the accuracy results.In this article, the decision trees approach had an accuracy of 75.3%, the PSO method had 66.4%, the GA method had 69.4%, logistic regression had 79.4%, and SVM had 80%. The value of the accuracy of implementation techniques of KNN, naive Bayes, NN, and RF for the prediction of disease Hepatitis C can conduct repairs is indicated by the value of which is high. The rated accuracy of the methods KNN, naive Bayes, and RF in a row was 89.43%, 90.24%, and 94.31%, respectively. When compared to the methods KNN, naive Bayes, RF, and, the method NN has a high value of accuracy, namely 95.12%.[9]

The goal of Nandipati et al. is to see if equivalent classification performance can be observed for multi and binary different classifiers within the same dataset. In both multi and binary-class label datasets, the performance of the model on testing data is measured using measurement methods like accuracy, precision, and recall (macro average). To evaluate the effectiveness of Python and R tools, the average total ratings of evaluation measures are used. the effectiveness of Python and R tools in the contents of the multi and binary class labels using Hepatitis C Virus (HCV) from Egyptian patient's data from UCI. The multi and binary dataset performances of the HCV dataset were analyzed in this work using classification and feature selection methods built in the Python Scikit learn to package and the R-CARET package, respectively. The RF with an accuracy of 54.56% in binary classification outperforms the prior studies' classifier. In a multi-class label, the KNN, on the other hand, has an accuracy of 51.06%. [10]

The goal of Shousha et al. is to examine and compare the prediction accuracy of scoring systems. There are 16 features, and the REPTree chose the IL28B genotype as the greatest predictor of progressive fibrosis. They employed data mining analysis to build a decision tree using the reduced error (REP) technique, then used the Auto-WEKA tool to find the best classifier. This study comprises 427 HCV-related liver fibrosis patients. Based on the results of their FibroScan, the patients were separated into two groups. Group 1 included 204 patients (47.8%) with no, little, or moderate fibrosis, and Group 2 included 223 patients (52.2%) with severe fibrosis. There is relevant work there. However, the way they explain it is not very nice.[11]

KayvanJoo et al. find matching drivers of therapy outcome inside the complete HCV target dna using feature selection techniques like Chi-Squared, Gini Index, and ML algorithms and other bio informatics techniques. Characteristic weighted methods were used to assess the significance and significance of each beneficial feature in the construction of the variable. A mixture of methods have used chosen genetic properties to identify HCV sub types 1a and 1b treatment takers versus quasi at 75.00 percentage points and 85.00 percentage accuracy, correspondingly.[12]

In order to anticipate repurposed therapeutics targeting HCV quasi (NS) molecules, Kamboj et al. developed the "Anti-HCV" system combining machine learning and structure-property relationship (QSAR) techniques. To use the finest created algorithms, they were able to achieve Pearson's correlations ranging between 0.80 - 0.92 in 10-fold cross validation and comparable numbers on independent datasets. In this investigation, interesting repurposing drugs were found, including naftifine, butalbital (NS3), vinorelbine, epicriptine (NS3/4A), pipecuronium, trimethaphan (NS5A), olodaterol, and vemurafenib (NS5B), among others. In order to locate prospective candidates for reusing drugs, researchers also searched its "DrugBank" registry. The docking studies technology has demonstrated the efficacy of selectively remanufactured compounds.[13]

CHAPTER 3

Methodology

3.1 Explanation of the data

First of all I find the dataset from UCI Machine Learning Repository .The Dataset Contains an account of the Blood donation of 615 people. There are 14 types of attributes including Age, Sex and the rest are lab tasted Blood Donor. All attributes except category and sex are numerical .There are 5 Category (0= Blood Donor, 0s= suspected Blood Donor , 1 =Hepatitis c virus. 2=Fibrosis stage(first stage). The remaining 3= Cirrhosis). The target attribute for Classification is Category(blood donor vs Hepatitis C(including its progress ('just' Hepatitis C ,Fibrosis, Cirrhosis) Where suspected patients are 540 and Healthy patients are 75 .

3.2 Algorithm description

3.2.1. Logistic regression

This model is widely used in statistics to simulate the likelihood that a specific type of event will occur, such as the likelihood that a squad will be successful or that a patient would be in perfect condition. This can be extended to cover a variety of additional scenarios, such figuring out whether a picture depicts a dog, cat, lion, or something else entirely. One significant detected object was expected to be present in the image, with the significance of each object ranging from zero to one. The logistic model's log-odds for the value "1" are a linear synthesis of one or maybe more relationships between the predictions; each of the two parameters may be a binary classifier problem or any real value. The logistic function, hence the name, converts the data to likelihood; the labeling is necessary since the average diameter of the value labeled "1" may vary between 0 and 1. The logit, which derives from the logistic unit and is therefore relatively separate, is the accepted unit of measurement for the logarithmic scale. The distinguishing feature of the logistic regression model is that, with each predictor variable having its own parameter, increasing one of the individual predictor variables scales the likelihood of the specific result at a constant speed; for binary predictors, this extrapolates the hazard ratio. The likelihood of stagnation can be replaced by equivalent models, like the probit model. The dependent variable is represented using a binary logistic regression model with two levels.When there are more than two results in an output, multinomial logistic regression is used to model it. The logistic function, hence the name, translates file to likelihood; the average diameter of the value labeled "1" may swing between 0 and Ordinal logistic regression is used if the various categories are arranged in an ordered fashion.

3.2.2 Random Forest

Hyperparameters in call trees and maybe in fabric classification are similar to those in a random forest. The development of these plants in a Random Forest increases the unpredictability of the model. Instead than focusing on the most significant attribute while ripping a node, it searches for the most fundamental one among a variety of potential values. The Random Forest algorithm in action
The stages listed below can assist us in comprehending how the Random Forest algorithmic program works:

Picking random samples from a specific dataset is the first stage.

One set of stairs The wire tree for each sample can then be generated by this algorithmic software. Each call tree will then experience the influence of the forecast after that.

In step three, options are put into practice for each expected result.

Choose the anticipated outcome that garnered the most votes as the last stage.

The mean square error is used to show how your data branches out from each node when using the Random Forest approach for regress cases.MSE=1ni=1N(xi-yi)2

3.2.3 K Nearest Neighbors Classifier

One of the fundamental but crucial categorization methods in machine learning is KNearest Neighbors. This falls under the category of supervised learning and has numerous applications in data analysis, penetration testing, and pattern matching. Despite its ability to be applied to classification and regression problems issues, it really is commonly employed as both a classification model because it relies just on idea that comparable locations could be discovered close to each other.In summary, the k-nearest neighbor algorithm seeks to determine the closest neighbors of a particular probe image in order to categorize them. There seem to be numerous range measurements available, but this section would only talk about the following several:

The much more widely utilized maximum distance is called as Euclidean distance (p=2), which can only be employed with genuine matrices. The perfect line here between query location and the additional point being evaluated is determined using the formula following.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Formula for Manhattan distance:

$$\text{Manhattan Distance} = d(x,y) = \left(\sum_{i=1}^{m} \left| x_i - y_i \right| \right)$$

3.2.4 Decision tree

The non-leaf or decoration nodes in a decision tree represent decisions, while the leaf nodes in a decision tree reflect results or class labels. Each internal node examines one or more designations, resulting in one or more connections or branches. The government has given this relationship a decision value. The links listed above are inclusive and exclusive at the same time. This means that if you simply click on one of the links, you will be protected against any scenario.

Decision trees are the finest tools for comparing various options. from each choich. A binary decision tree's nodes each identify the comparison to be done or the preferred option. In and out of each node, there are two edges. One edge represents the outcome "yes" or "true," and the other edge represents the outcome "no" or "false."

On four coins, three of equal weight and one lighter, the letters A, B, C, and d appear to be printed. Find the coin that weighs more. Figure 3 displays the prediction model for this problem. At the root node, the body weights of A + B and C + D are compared and evaluated. If A+B is more than C+D, the left subsidiary is unquestionably true, hence the answer is "yes." A + D is, however, the more difficult choice because it necessitates using the correct branching strategy. The node on the left branch contrasts the weight-training regimens of a and b. If the answer to this question is "yes," then the greater load coin is chosen as an. If you say "no," it chooses the currency with the significantly higher value, which is b. If the main node returns "no," the same procedure is used for c and d.

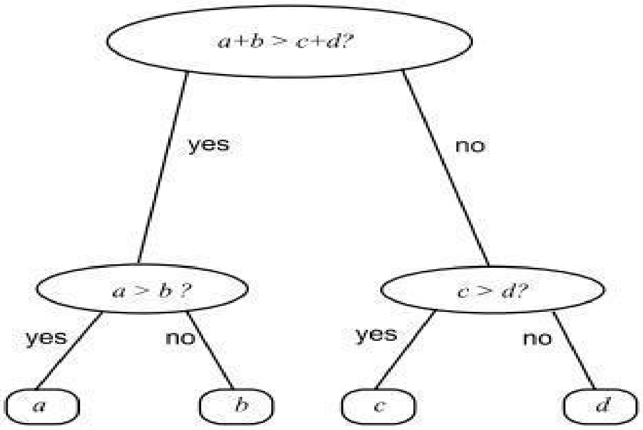


Figure 3.1: Decision tree

The illustration in Figure 3.1 illustrates a simple decision tree in the context of the organization. The following problems were brought up:

Four-leaf nodes represent each of the four potential outcomes. The weight of a coin is represented by each leaf node.

Two pairwise comparisons are necessary for a conclusive finding or a leaf node. Every weighing technique has a corresponding stage. There are numerous decision nodes, one at each leaf and root.

Every tree, from the root to the leaf, is subject to a set of rules. As an example, the condition for the followed by a judgment is "if A + B>C + D and A>B, then an is light."

3.2.5 CatBoost Classifier

Yandex created the open source library known as Cat Boost[14]. It provides a framework for gradient boosting that, among some other things, tries to address the problem of categorical data to use a possible combination approach as an alternative to the standard method. Among the most widely used ML frameworks worldwide, according to Kaggle, is Cat Boost. There in 2020 survey [15], it was ranked as the top-8 most used ML platform, while in the 2021 assessment, it must have been ranked even as top-7 highest rated ML template.

Broadly speaking, boost techniques have two major flaws.

1. Since boost techniques are forest, performance of the model is a frequent issue.
2. Due to the successive generation of new forests during the training stage of boost methods, parallelization this procedure remains difficult.

3.2.6. Gradient Boosting Classifier

Given huge and complicated data, gradient boosting is a strategy that stands out because of its forecast accuracy and speed. This program had provided as finest outcomes from across board, spanning Kaggle tournaments to machine learning solutions for companies.That method begins by creating a judgment stem, after which all of the pieces of information are given adequate weight. The values for each of the incorrectly classified items are therefore increased, while those are simple to identify or are categorized have their weights reduced.

Its first action can indeed be expressed by the equation by the equation as

$$F_0(x) = \underset{\gamma}{\arg\min} \sum_{i=1}^{n} L(y_i, \gamma).$$

3.3 Data Implementation

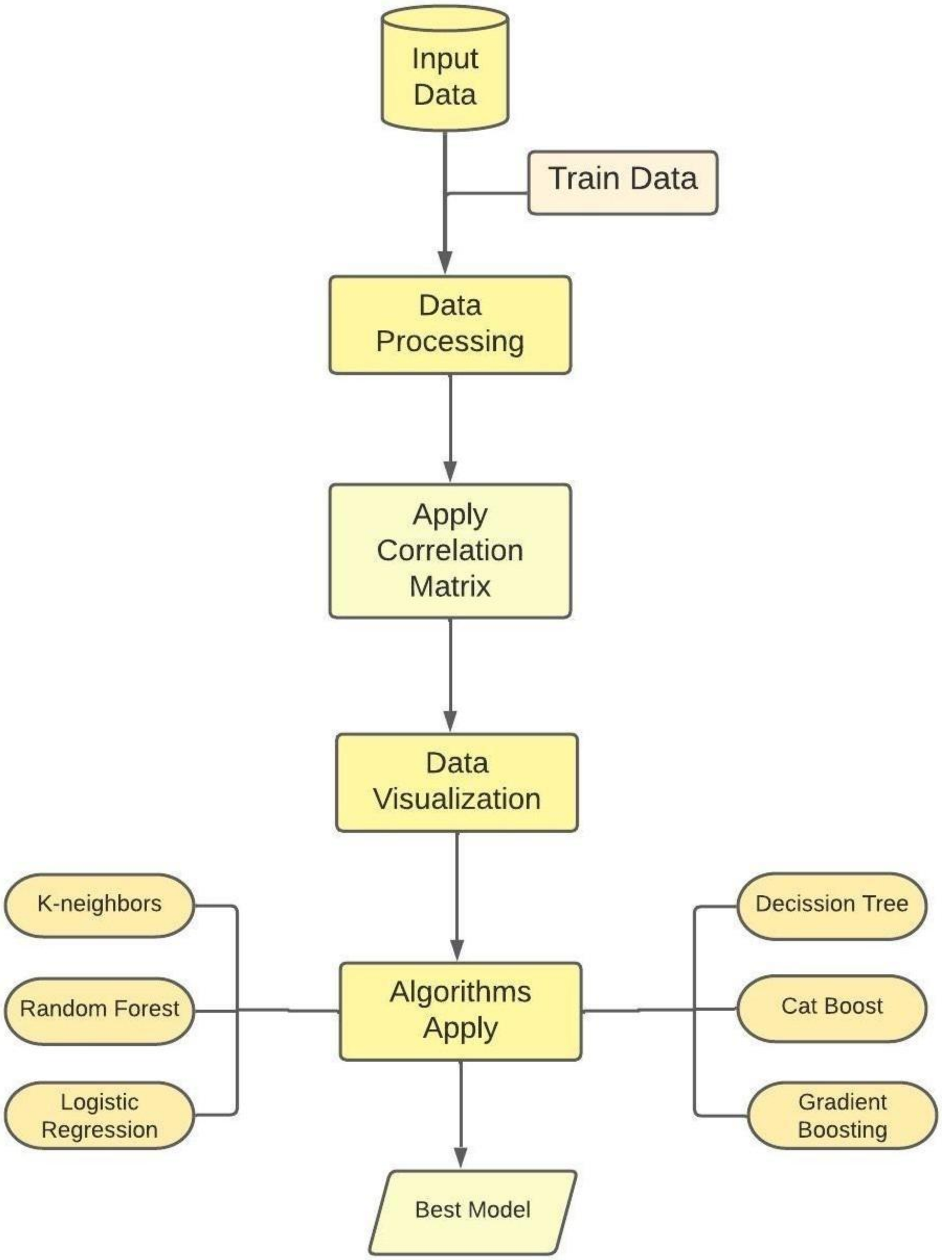


Figure 3.2: Data implementation

3.3.1 Input data

Input data is stored in a system file that will be used as an input by a system or piece of software. the original file The area of engineering technology known as computational engineering employs computers to investigate structures and processes that can be calculated. Separating data into different groups is a common strategy used in machine learning. Data is frequently divided into train and test sets in order to train the model. For this analysis 80% data uses for training and 20% data is using for testing. Total 615 data and 14 attribute are used. Attribute name are: patient id noted by X, category value are('0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'), age value are in year format, sex (Male as M and Female as F) and rest 10 attribute are laboratory attribute, alb means albumin its value are numeric, alp means alkaline phosphatase its value are numeric, alt means alanine aminotransferase its value are numeric, ast means aspartate aminotransferase its value are numeric, bil means bilirubin its value are numeric, che its value are numeric, chol means cholesterol its value are numeric, crea means Creatinine its value are numeric, ggt level its value are numeric, prot its value are numeric.

3.3.2 Data processing

Data processing is the method of transforming unprocessed, computer data into useful knowledge. In order to transform the basic information into useful information, processing of data entails gathering, collecting, categorizing, keeping, then modifying and amending it.

3.3.2. Apply correlation matrix

All the connections between the variables in a dataset can be quickly and easily summarized using a correlation matrix. Regression diagnostics employ a correlation matrix.

3.3.3 Algorithm

There are six algorithm are used and they are:

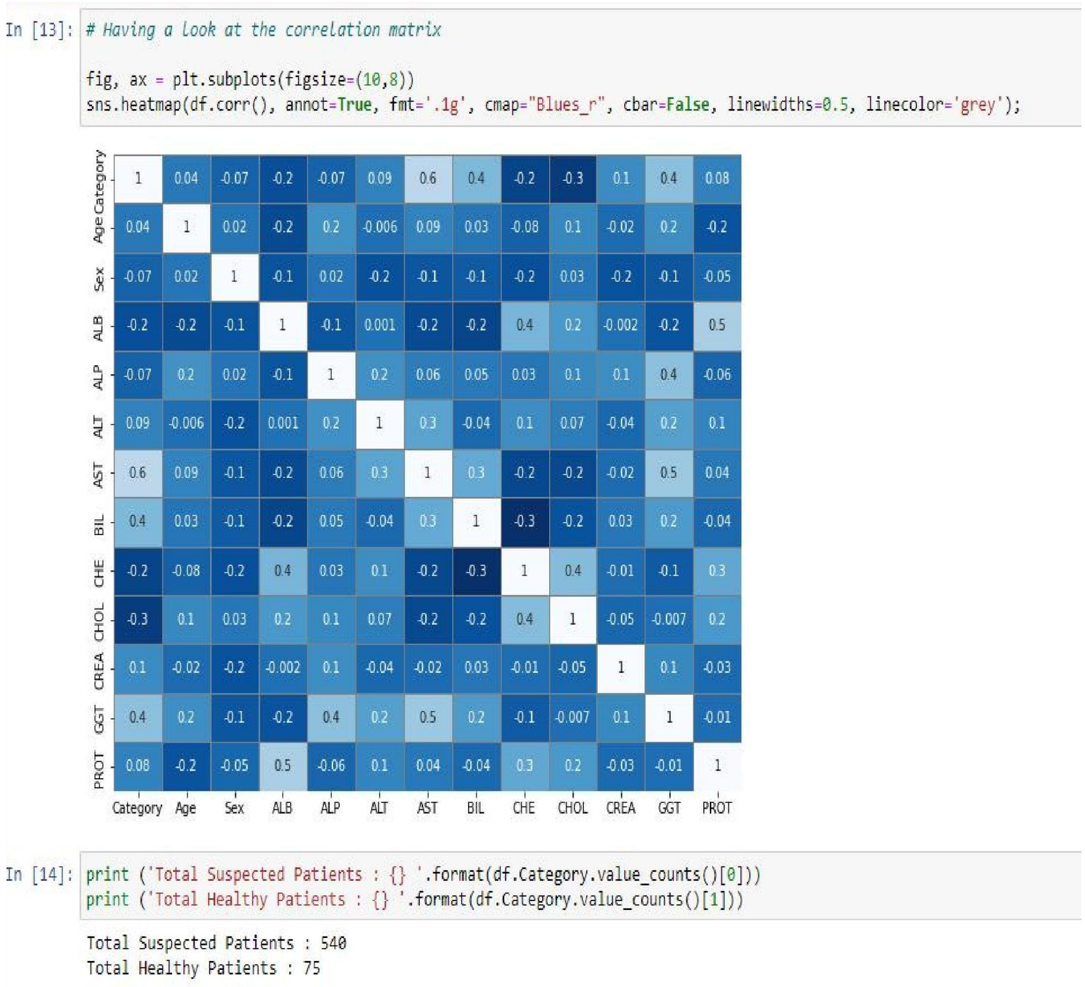
1. Logistic Regression.(LR)
2. Decision Tree
3. Random Forest
4. Cat Boost
5. Gradient Boosting
6. K Neighbors

3.3.4 Best Model

After utilizing these six algorithms, the best outcome is discovered.

The optimal algorithm is one that is applied after testing six different algorithms.

3.4. Corelation matrix



```
In [13]: # Having a look at the correlation matrix

         fig, ax = plt.subplots(figsize=(10,8))
         sns.heatmap(df.corr(), annot=True, fmt='.1g', cmap="Blues_r", cbar=False, linewidths=0.5, linecolor='grey');
```

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
Category	1	0.04	-0.07	-0.2	-0.07	0.09	0.6	0.4	-0.2	-0.3	0.1	0.4	0.08
Age	0.04	1	0.02	-0.2	0.2	-0.006	0.09	0.03	-0.08	0.1	-0.02	0.2	-0.2
Sex	-0.07	0.02	1	-0.1	0.02	-0.2	-0.1	-0.1	-0.2	0.03	-0.2	-0.1	-0.05
ALB	-0.2	-0.2	-0.1	1	-0.1	0.001	-0.2	-0.2	0.4	0.2	-0.002	-0.2	0.5
ALP	-0.07	0.2	0.02	-0.1	1	0.2	0.06	0.05	0.03	0.1	0.1	0.4	-0.06
ALT	0.09	-0.006	-0.2	0.001	0.2	1	0.3	-0.04	0.1	0.07	-0.04	0.2	0.1
AST	0.6	0.09	-0.1	-0.2	0.06	0.3	1	0.3	-0.2	-0.2	-0.02	0.5	0.04
BIL	0.4	0.03	-0.1	-0.2	0.05	-0.04	0.3	1	-0.3	-0.2	0.03	0.2	-0.04
CHE	-0.2	-0.08	-0.2	0.4	0.03	0.1	-0.2	-0.3	1	0.4	-0.01	-0.1	0.3
CHOL	-0.3	0.1	0.03	0.2	0.1	0.07	-0.2	-0.2	0.4	1	-0.05	-0.007	0.2
CREA	0.1	-0.02	-0.2	-0.002	0.1	-0.04	-0.02	0.03	-0.01	-0.05	1	0.1	-0.03
GGT	0.4	0.2	-0.1	-0.2	0.4	0.2	0.5	0.2	-0.1	-0.007	0.1	1	-0.01
PROT	0.08	-0.2	-0.05	0.5	-0.06	0.1	0.04	-0.04	0.3	0.2	-0.03	-0.01	1

```
In [14]: print ('Total Suspected Patients : {} '.format(df.Category.value_counts()[0]))
         print ('Total Healthy Patients : {} '.format(df.Category.value_counts()[1]))

         Total Suspected Patients : 540
         Total Healthy Patients : 75
```

Figure 3.3: Correlation matrix

To obtain the cross-correlation of the two matrices, compute and sum the component products for each point of the two columns relative to the first column. There are several limitations, but this can be used to work out the offset required to get two matrices of related values to overlap. An easy way to condense a datasets is to use a correlation matrix.

A correlation matrix can be used to quickly and readily summarize all the relationships between the variables in a datasets. A correlation matrix is used in regression diagnostics.

First off all ,I entered data and then mapping 0 and
1(0=suspected patients,1=Healthy patients).
Then use correlation matrix and get

1. Suspected patients:540

2.Healthy patients:75.

Of which 61.30% man and 38.70% female.

3.5. Data visualization

3.5.1. **Co**nfusion Matrix

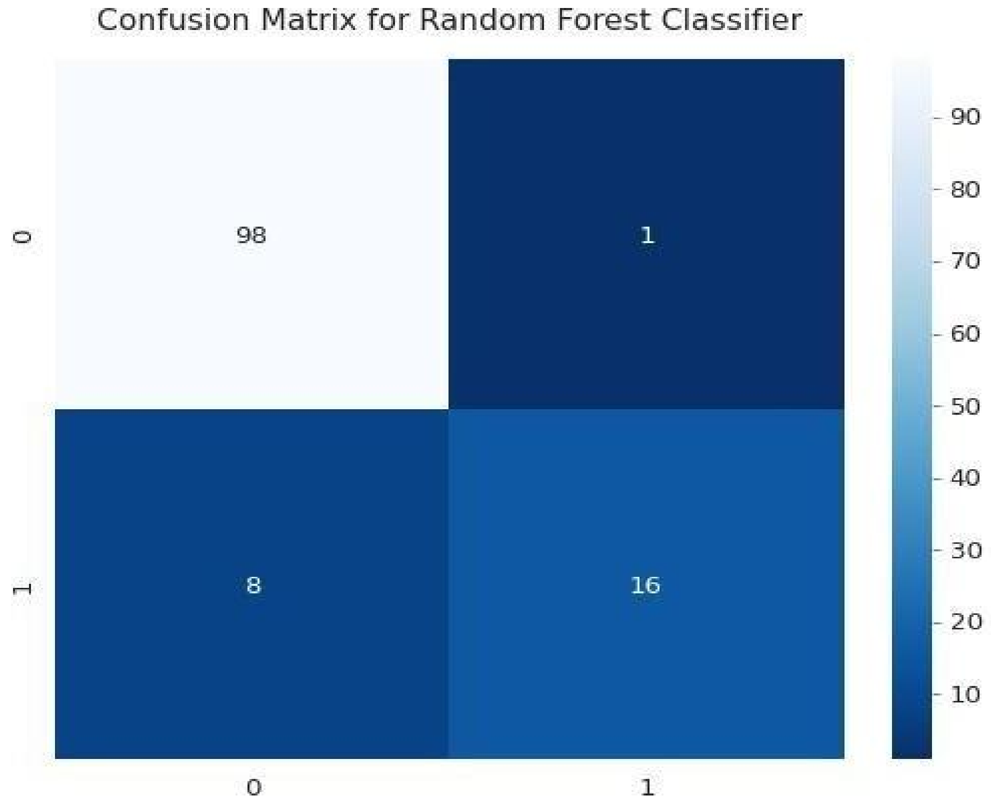


Figure 3.4: confusion matrix Random Forest Algorithm

Above this figure we can see the confusion matrix for Random Forest Algorith to perfectly classify the Performance of the problem

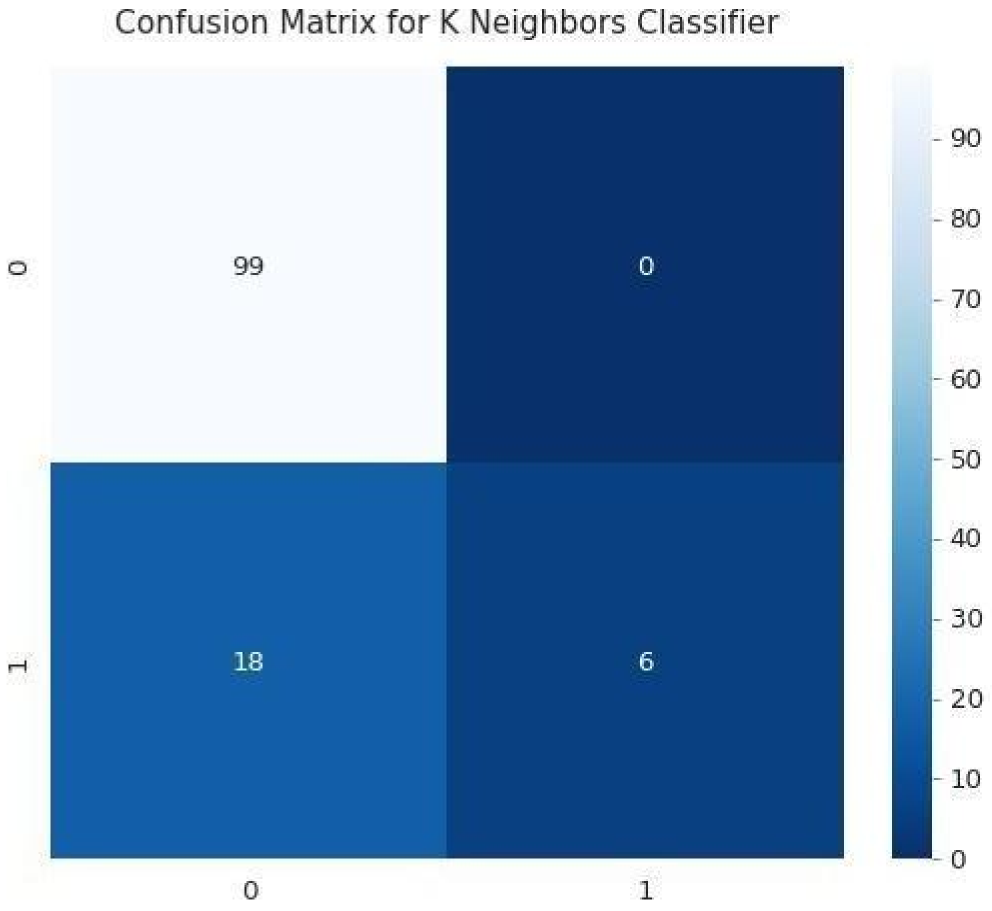


Figure 3.5: confusion matrix for K Neighbors algorithm

KNN determines the distances between a query and each example in the data, chooses the K instances closest to the query, and then votes for the label with the highest frequency (in the case of classification) or averages the labels (in the case of regression).

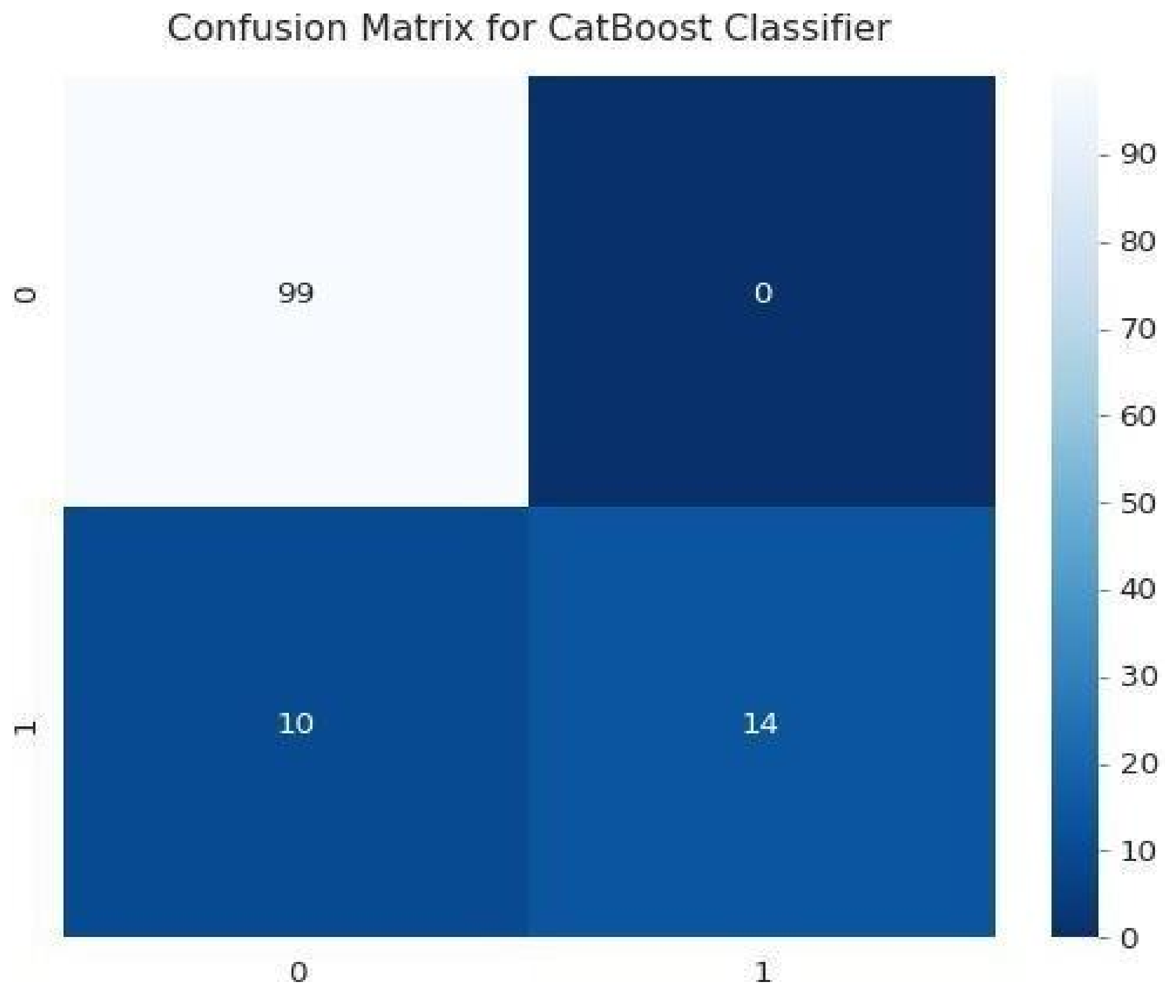


Figure 3.6: Confusion matrix for CatBoost algorithma

Above figure show the Accuracy of the CatBoost algorithm for confusion matrix as a graph. So anyone who have some basic knowledge understand the procce of work.

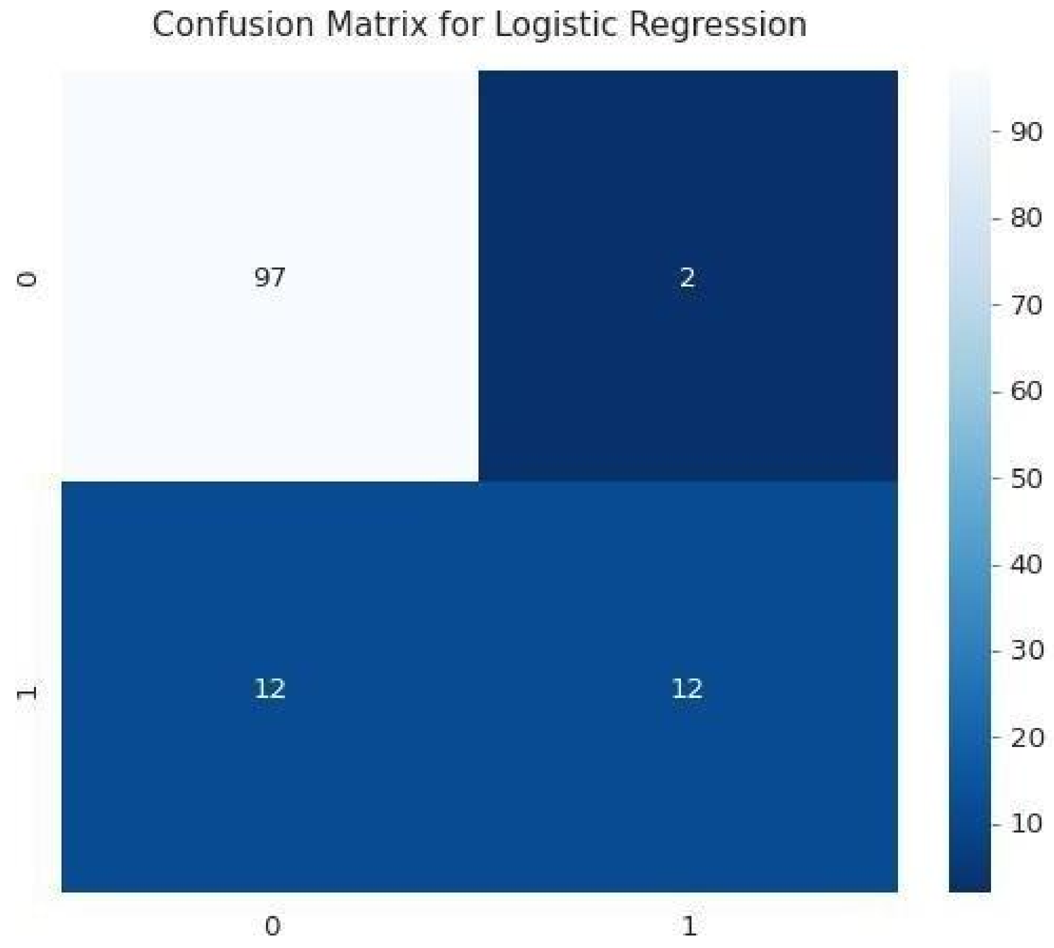


Figure 3.7: Confusion matrix for Logistic Regression algorithm

In order to verify the accuracy, we will classify our train data using a logistic regression model and forecast our test data. We may import accuracy score and classification report from the same library to determine the accuracy of a confusion matrix as well as all other metrics.

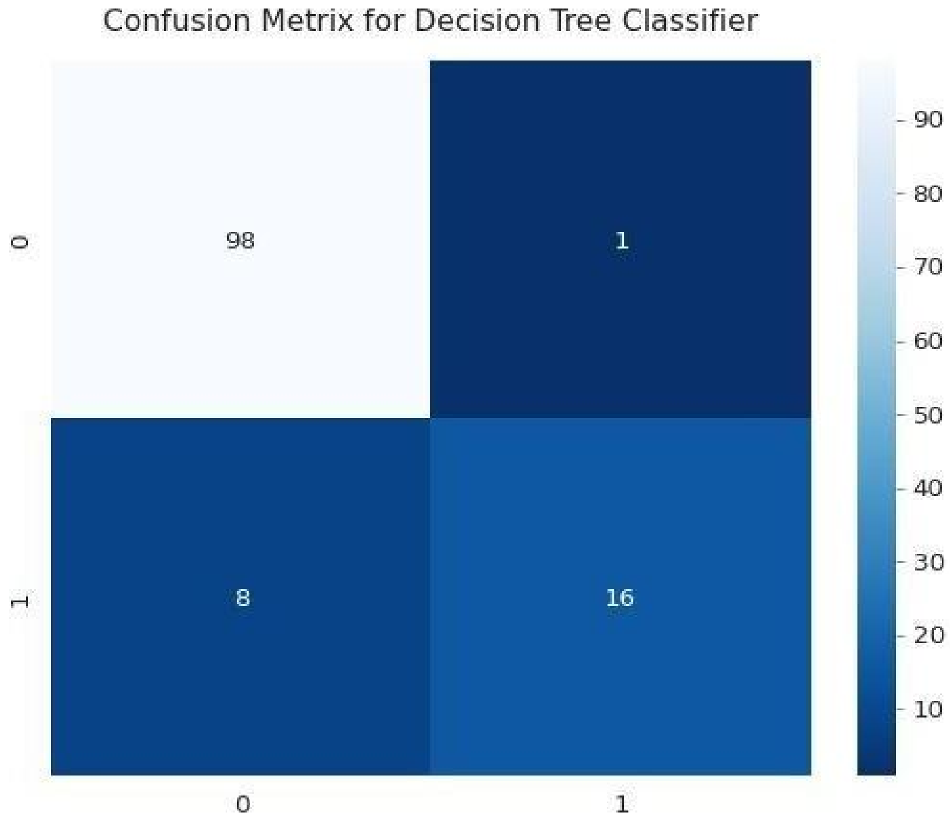


Figure 3.8: Confusion matrix for decision tree algorithm

In this figure ,we can see the confusion matrix for decision tree algorithm .Basically it's a simple procces of decision tree algorithm

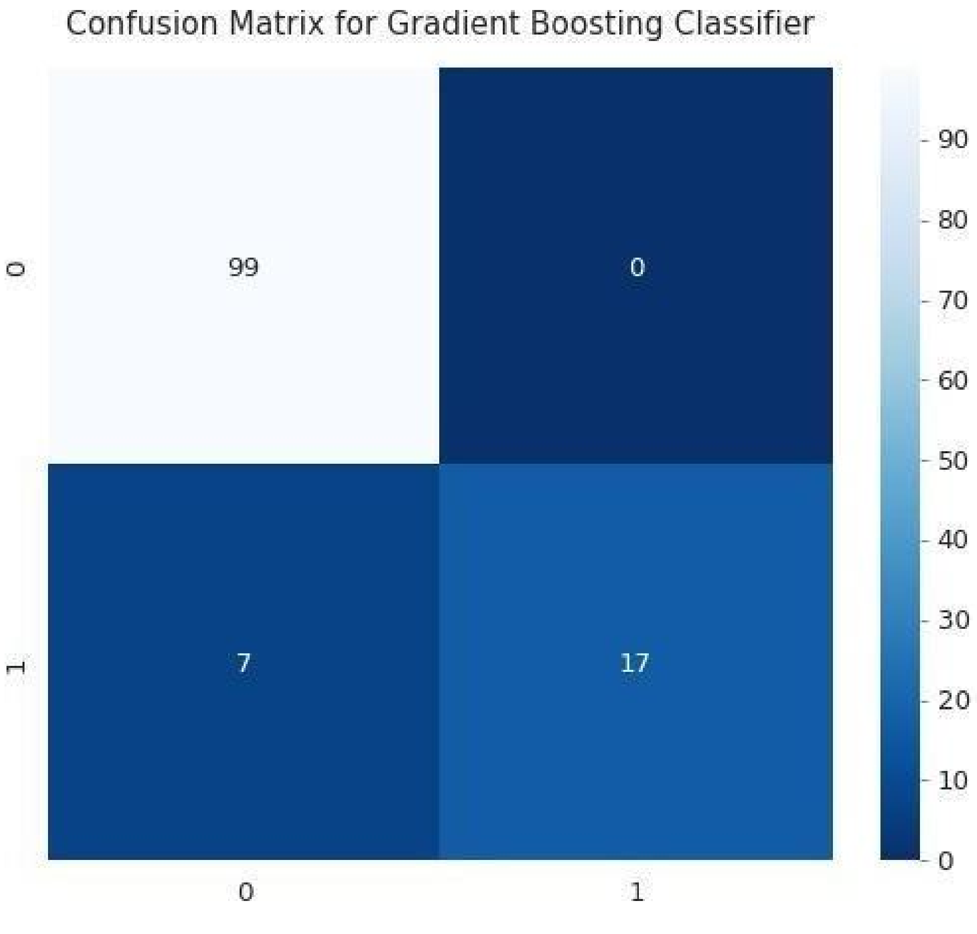


Figure 3.9: Confusion matrix for Gradient Boosting algorithm

In this figure ,we can see the specifically performance of Confusion matrix for Gradient Boosting Algorithm .It is categorizing into classes to the given set.

3.5.2. Flowchart of gender in datasets

A chart called a confusion matrix is employed to describe how well a categorization performs. The output of a prediction model is shown and summarized in a confusion matrix. Include the prediction and the real figures that evaluate inside the program while constructing your confusion matrix. Every anticipated category has a corresponding rows, so each number of classes has a corresponding columns. The matrix may determine difficulties with several classes or two classes, based upon the number of responses you get for every entry. In figure 3.4, figure 3.5, figure 3.6, figure 3.7, figure 3.8, figure 3.9 I find out the confusion matrix of these applying algorithm.

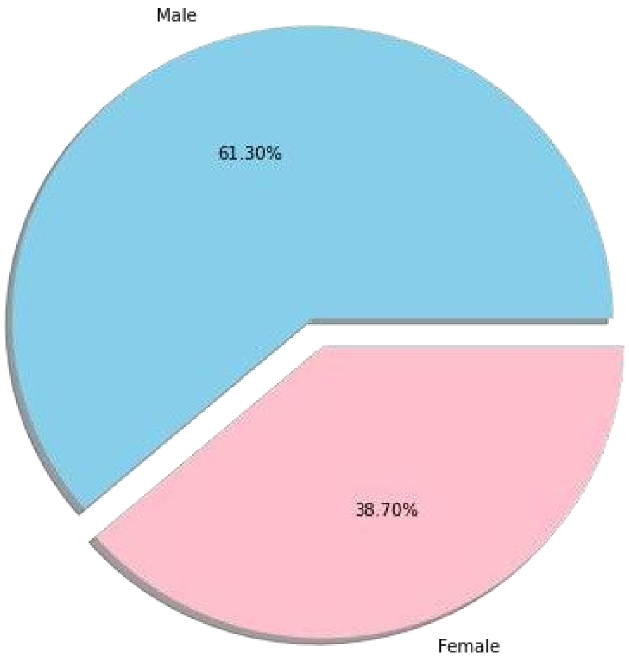


Figure 3.10: Percentage of gender

We can see a flowchart .This displays the proportion of genders excluded from data.

CHAPTER 4

Result analysis

In table 1 I calculated this same accuracy by evaluating the results of confusion matrices. and AUC value.The Accuracy of Random Forest Classifier 92% Logistic Regression Classifier 89%, Gradient Boosting Classifier 94%, Decision Tree Classifier 92% K Neighbors Classifier 85%, and Cat Boost classifier 91%. Comparing the result here Gradient Boosting algorithm deliver the best accuracy . In this table, I also find out the result of confusion matrix. I'm going to call it Actual Negative. and Actual Positive. Here logistic regression algorithm’s true positive is highest (99%) and false negative (0%).

In this table, I also find out the result of confusion matrix. A chart called a confusion matrix is utilized to describe how well a classification system performs. The output of a classification method is shown and summarized in a confusion matrix. The accuracy, sensitivities, & specificity of three crucial characteristics are found using the components of the confusion matrix. Here, I learn how accurate these six models are.The capacity of a diagnostic to properly distinguish between patients and healthy instances is a measure of its accuracy. Calculating overall percentage of true positive and true negative results in all analyzed cases is essential in determining a study's accuracy.
The formula of accuracy,

Table 4.1: Evaluated Result For all Algorithm and Confusion Matrices

Classification	Accuracy	Label	Predictive Negative(%)	Predictive Positive(%)
Logistic regression	89%	Actual Negative	12	2
		Actual Positive	12	97
Random Forest	92%	Actual Negative	16	1
		Actual Positive	8	98
K Neighbors Classifier	85%	Actual Negative	6	0
		Actual Positive	18	99
Decision tree	92%	Actual Negative	16	1
		Actual Positive	8	98
CatBoost Classifier	91%	Actual Negative	14	0
		Actual Positive	10	99
Gradient Boosting Classifier	94%	Actual Negative	17	0
		Actual Positive	7	99

In this table, I also find out the result of confusion matrix. A chart called a confusion matrix is utilized to describe how well a classification system performs. The output of a classification method is shown and summarized in a confusion matrix. The accuracy, sensitivities, & specificity of three crucial characteristics are found using the components of the confusion matrix. Here, I learn how accurate these six models are.The capacity of a diagnostic to properly distinguish between patients and healthy instances is a measure of its accuracy. Calculating overall percentage of true positive and true negative results in all analyzed cases is essential in determining a study's accuracy.

The formula of accuracy,

$$\text{Accuracy} = TP + TN\ TP + TN + FP + FN.$$

Here,

True positive (TP) is the proportion of instances where patients were accurately detected.

False positive (FP) = percentage of instances where patients were wrongly recognized

The number of instances accurately classified as normal is known as the true negative (TN)

False negatives (FN) are cases that were mistakenly classified as normal.

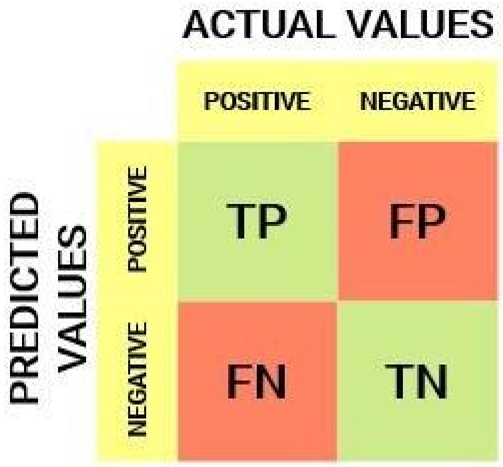


Figure 4.1:Confusion matrix chart

Positive or negative values can be assigned to the target variable.

The columns show the target variable's real values.

The rows display the target variable's anticipated values.

I'm going to call it Actual Negative. and Actual Positive. Here logistic regression algorithm’s true positive is highest (99%) and false negative (0%).

CHAPTER 6

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

In Bangladesh the highest HEV prevalence (53%), followed by HAV (39%), HBV (19%), and HCV (13%), suggesting that HEV and HAV are the most common in Bangladesh . Notably, HAV prevalence was very high (100%) among children (≤6 years of age). However, HEV infection was reported to be higher in adults (≥30 years of age) than in children. Another study using sera collected at various hospitals in and around Dhaka from 74 adult patients (aged 15–67 years) indicated concomitant infection in 28

patients with more than one type of hepatitis virus . Notably, 6.75% (5/74) of the cases were positive for HAV, 40.54% (30/74) for HBV, 17.56% for HCV, and 39.18% (29/74) for HEV [103]. Another recent study screened 998 suspected cases of acute hepatitis for anti-HAV IgM and anti-HEV IgM in 10 different hospitals across seven divisions (Dhaka, Chattogram, Rajshahi, Khulna, Sylhet, Barishal, and Rangpur) of Bangladesh and found that 19% (191/998) and 10% (103/998) were positive for HAV and HEV, respectively [4]. Notably, a recent study investigated 275 samples obtained from an outbreak of acute jaundice syndrome (AJS) among Rohingya refugees in Cox's Bazar, Bangladesh, and found that 154 (56%) samples were positive for hepatitis A, 1 (0.4%) for hepatitis E, 36 (13%) for hepatitis B, and 25 (9%) for hepatitis C . Coinfections with multiple etiologies were also reported in 24 samples (9%) [16].

5.2 Ethical Aspects

When conducting research, I must be truthful. Some academics exploited fictitious datasets or modified data to achieve better outcomes in their studies. It is completely unethical. Because in medical science, this type of study has the potential to endanger patients' lives. Patients can die as a result of poor care. I obtained data from the uci machine learning repository a reliable data collection for users of data science Program collects hepatitis c information in attempt to minimize the cancer burden in the United States. For predicting survival and cancer stage, I used 5402 patient datasets for analysis.

5.3 Sustainability Plan

When we wish to undertake something, we must consider its long-term viability. In this study, we attempted to develop models utilizing machine learning methodologies that can identify the disease stages and survival of Hepatitis C patients. This type of research has the potential to have a substantial influence on humans. That is why we considered a sustainability plan for our study. Every disease's symptoms have been altering day by day. This is also true for prostate cancer. That is why we have devised a solution to this problem. In the near future, we want to incorporate reinforcement learning into our model. This will acquire new forms of data, train our model, and continually deliver new models. According to who, they want people to have greater access to treatment and care, regardless of the form of hepatitis they have. The WHO has set a goal of eradicating hepatitis by 2030, and in order to do so, they are asking all nations to help reach particular objectives such as: 90% reduction in new hepatitis B and C infections

CHAPTER 5

Conclusion

5.1 Conclusion

Hepatitis C that is persistent is a persistent disease. When left untreated, it may last a lifespan, result in severe medical conditions, and perhaps even lead to mortality. Unless humans cure this same liver disease resultative as a ritualistic result at such a specific time interval, we may not achieve the highest predictive performance in previous research. Six machine learning algorithms are employed. Gradient Boosting Classifier,Decision Tree Classifier, K Neighbors Classifier, Classifier, Random Forest Classifier, Logistic Regression one of the gradient-boosting classifier's top six results. The percentage is 94.31. I'll use a correlation matrix to compute the values below: 1. 5402 accused treatments. healthy patients: 75., with men making up 61.30% and women 36.70% of the total. I also discover these algorithms' confusion matrix.

5.2 Future work

We used four algorithms to predicting Hepatitis C virus here, and we want to use more in the future to identify it. We can employ a variety of techniques to achieve better and ideal results. Deep learning or artificial neural networks should be used in our future study for the greatest results.

References

[1].https://www.who.int/news-room/fact-sheets/detail/hepatitisc#:~:text=Hepatitis%20C%20is%20an%20inflammation,including%20liver%20cirrhosis%20and%20cance r.

[2]. https://medlineplus.gov/hepatitisc.html

[3]. Barakat, N. H., Barakat, S. H., & Ahmed, N. (2019). Prediction and staging of hepatic fibrosis in children with hepatitis c virus: A machine learning approach. Healthcare Informatics Research, 25(3), 173181.

[4]. Chen, Y., Luo, Y., Huang, W., Hu, D., Zheng, R. Q., Cong, S. Z., ... & Yan, H. (2017). Machinelearning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B. Computers in biology and medicine, 89, 18-23.

[5]. Safdari, R., Deghatipour, A., Gholamzadeh, M., & Maghooli, K. (2022). Applying data mining techniques to classify patients with suspected hepatitis C virus infection. Intelligent Medicine.

[6].Kashif, A. A., Bakhtawar, B., Akhtar, A., Akhtar, S., Aziz, N., & Javeid, M. S. (2021). Treatment response prediction in hepatitis C patients using machine learning techniques. International Journal of Technology, Innovation and Management (IJTIM), 1(2), 79-89.

[7]. Mendlowitz, A., Bremner, K. E., Walker, J. D., Wong, W. W., Feld, J. J., Sander, B., ... & Krahn, M. (2021). Health care costs associated with hepatitis C virus infection in First Nations populations in Ontario: a retrospective matched cohort study. Canadian Medical Association Open Access Journal, 9(3), E897E906.

[8].Ahammed, K., Satu, M. S., Khan, M. I., & Whaiduzzaman, M. (2020, June). Predicting infectious state of hepatitis c virus affected patient's applying machine learning methods. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 1371-1374). IEEE.

[9]. Syafa'ah, L., Zulfatman, Z., Pakaya, I., & Lestandy, M. (2021). Comparison of machine learning classification methods in hepatitis C virus. Jurnal Online Informatika, 6(1), 73-78.

[10]. Nandipati, S. C., XinYing, C., & Wah, K. K. (2020). Hepatitis C virus (HCV) prediction by machine learning techniques. Applications of Modelling and Simulation, 4, 89-100.

[11]. Shousha, H. I., Awad, A. H., Omran, D. A., Elnegouly, M. M., & Mabrouk, M. (2018). Data mining and machine learning algorithms using IL28B genotype and biochemical markers best predicted advanced liver fibrosis in chronic hepatitis C. Japanese journal of infectious diseases, 71(1), 51-57.

[12].Syafa'ah, L., Zulfatman, Z., Pakaya, I., & Lestandy, M. (2021). Comparison of machine learning classification methods in hepatitis C virus. Jurnal Online Informatika, 6(1), 73-78.

[13]. Kamboj, S., Rajput, A., Rastogi, A., Thakur, A., & Kumar, M. (2022). Targeting non-structural proteins of Hepatitis C virus for predicting repurposed drugs using QSAR and machine learning approaches. Computational and Structural Biotechnology Journal, 20, 3422-3438.

[14]. "catboost/catboost". August 30, 2020 – via GitHub.

[15]. "State of Data Science and Machine Learning 2020".

APPENDIX

Abbreviation

ML = Machine Learning

AI = Artificial Intelligence

HCV = Hepatitis C Virus

CHC = chronic hepatitis C

LOLA = L-ornithine L-Aspartate

Appendix : Research reflection

We know a little bit about machine learning at the start of the project. We are unable to comprehend how algorithms function or how to anticipate the future. Our supervisor treated us with kindness and generosity. He provided us with helpful advice and was a great assistance. We learn a lot throughout this study period of work. We get new skills in technique, algorithm, and other methodologies.

Finally, through conducting the research, we have grown braver and motivated to conduct other research in the future.

Report

ORIGINALITY REPORT

22%	17%	10%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	6%
2	www.ncbi.nlm.nih.gov Internet Source	3%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
4	Submitted to Coventry University Student Paper	1%
5	www.researchgate.net Internet Source	1%
6	Submitted to University of Westminster Student Paper	1%
7	Submitted to Teaching and Learning with Technology Student Paper	1%
8	Submitted to Kingston University Student Paper	<1%
9	Submitted to Liverpool John Moores University	<1%