HEPATITIS C PREDICTION USING MACHINE LEARNING MODEL: LOGISTIC REGRESSION

Abstract:- **The hepatitis C virus (HCV), which affects millions of people globally, is a serious global health issue. Effective management and diagnosis of HCV infection depend on early detection and precise prediction of its effects.The historical patient data that comprises demographic data (e.g., age, gender), clinical data (e.g., symptoms, medical history), and laboratory results (e.g., viral load, liver function tests) is used to train logistic regression models. Logistic regression methods calculate the chance or probability of HCV infection, disease progression, or treatment response by analysing these factors.**

Risk assessment is one of the major uses of logistic regression models in HCV prediction. These algorithms can examine the data gathered and pinpoint groups or persons who are more prone to HCV infection. Logistic regression models provide targeted treatments and public health planning by prioritising prevention efforts and resource allocation by considering factors like age, gender, lifestyle, and location.

Additionally useful in the early identification of HCV are logistic regression models. These algorithms can find patterns and signs that indicate HCV infection at an early stage by using patient data. Early detection enables immediate intervention, such as starting a treatment plan or altering one's lifestyle, which can greatly enhance patient results.

Machine Learning is now widely used in various fields and it has made a significant impact in the field of disease diagnosis. Prediction and classification of diseases are essential in medical science, as it attempts to control the spread of the disease and discover the infected regions from the early stages. In order to accurately anticipate and categorise diseases, machine learning (ML) techniques are frequently applied, serving as a useful tool for doctors and specialists. In this study, we suggest that to improve accuracy, a different model should be trained for each scenario. We made use of the accessible real-world data. The 615 patients in the collected dataset have 13 different traits. For model evaluation, the logistic regression approach has been used. The accuracy of the logistic regression model was initially 88.66%; however, it was increased to 90.24% after the hyper parameter values were changed.

Keywords – Hepatitis C virus, Logistic Regression, Disease diagnosis, Hyper parameters, Machine learning

I. INTRODUCTION

Hepatitis refers to liver inflammation. The liver is a vital organ that filters blood, processes nutrients, and fights infections. The liver's function can be impacted by inflammation or injury. Hepatitis can be brought on by chemicals, drugs, some medical disorders, and heavy alcohol consumption. Hepatitis C is a dangerous condition that slowly advances, has the potential to cause cancer, and can lay dormant in the body for ten to twenty years.

Hepatitis most frequently results from a virus. Before the blood supply was routinely checked for HCV, it was typically spread by contact with contaminated blood,this may happen through sharing needles or other injecting equipment, obtaining blood transfusions or organ transplants, or, less frequently, through sexual contact with an infected person.

Hepatitis comes in a variety of forms, but in this case, the Hepatitis C virus is what we are focussing on. When HCV is initially contracted, a person may experience a very mild infection with few or no symptoms or a serious condition necessitating hospitalization. Less than 50% of those who have hepatitis C are able to get rid of the virus on their own during the first six months of illness, for unknown reasons. Most of the infected individuals have chronic, or lifelong, infections. Chronic hepatitis C, if left untreated, can lead to fatal liver cancer, liver disease, liver failure, and other major health issues.Even though Hepatitis C is harmful, a test for the virus only requires cc of blood. The majority of individuals do not realize they have hepatitis C, resulting in difficulty to prevent and cure the condition.

1.Hepatitis C antibody tests, which clarify whether the body is infected with the hepatitis C virus. If the test result is positive, it shows that the patient is currently infected with hepatitis C or has past history of hepatitis C.

2. A hepatitis C virus can accurately determine the duration of the infection and the quantity of virus in the patient's body through RNA test.

3.Liver puncture or ultrasound is the primary tool for estimating the severity of liver disease. For patients diagnosed, these two tests should typically be carried out to assess the course of liver disease, which is also critical to the current treatment process when the condition is serious or has a lengthy duration.

Only the LFTs are simple to administer at routine checkups; the others are similarly expensive and not suited for widespread use.Puncture tests or ultrasounds are typically used to track the disease's progression in people who already have the condition; they are not suitable for diagnosing the condition at an early stage.Therefore, making appropriate use of the information from liver function tests where the characteristics are the elements present in a human liver blood sample has become successful. Because of this, we require a reliable and accurate Hepatitis C diagnosis. Medical event interpretation is a specialty of machine learning algorithms.

The main aims of this research are as follows:

1.To make a machine learning model that can accurately classify patients as having the hepatitis C virus or not based on relevant medical information. The objective is to create a model that can correctly recognise Hepatitis C in patients.

2.Increased diagnostic accuracy: Compared to current diagnostic methods, the objective is to improve the accuracy of Hepatitis C detection. This may need testing with different machine learning algorithms and approaches to improve precision and recall rates.

3.Identify significant risk factors: The goal is to study the data and identify the key hepatitis C risk factors. This could involve feature selection or feature importance analysis to determine the components in the prediction model that have the most impact.

4.Giving advice to medical experts is goal of this project. The advice will be based on the findings of the machine learning model. For people who are atmost risk for Hepatitis C, this may suggest strategies for early detection, preventative measures, or specific medications.

II. LITERATURE

In a study by Neukam et al., a descriptive analysis of the demographic parameter data was conducted. Patients with HIV/HCV coinfection from four Spanish cohorts and one German cohort who were prospectively monitored in the Infectious Diseases Unit made up the study population. To confirm that HCV viral genotype, rs12979860 genotype, and baseline HCV-RNA load were independently linked with sustained virologic response (SVR) in the study population, binary logistic univariate and logistic regression analysis was carried out. The statistical software was classified into groups by randomly dividing the patients into two groups of 60 and 40. This study had 521 patients in total. In order to identify potential and predicted responders to therapy with Peg-IFN plus RBV in HIV/HCV coinfected patients, it provides a straightforward and reliable pre-treatment technique, including three blood parameters [1].

It is vital that clinicians could find these untreated patients with hepatitis C infection, a nowadays-treatable disease. With 120023 HCV patients and 9601900 non-HCV patients used as modeling data, Doyle et al[2]developed five kinds of AI models attempting to find undiagnosed hepatitis C infection out of which one is Logistic regression. Model mined data from demographics, symptoms, treatments,risks and procedures relevant to HCV from the patient's medical history. At different recall levels, the models had different diagnostic performances. For example, the stacked ensemble had a specificity of 0.99 and precision of 0.97 at a recall level of 0.50.[2]

The logistic regression (LR) model is the most preferred classification method in supervised machine learning algorithms [3]. It is used to predict the categorical dependent variable using a given set of independent variables. The model is developed based on the probability concept. By mapping probabilities, the output is produced through the logistic sigmoid function [3]. It uses the following formula for creating a logistic function:

$$p = 1/1 + e^{\wedge}y \text{ ......(1)}$$

This Sigmoid function maps the predicted values of probabilities between 0 and 1. With this ability, the sigmoid function is widely used in healthcare to analyze multivariate regression problems [3]. The LR model was used in our survey as one of the classifiers because of its aptitude for categorising categorical data.

In (Bhargav et al., 2018) the authors compared the performances of four classification algorithms namely;logistic regression, decision tree,linear support vector, and naive Bayes applied to classify the hepatitis dataset in terms of accuracy, precision,recall, and F1-score. They concluded that the ordinary logistic regression algorithm achieved the best accuracy of 87.17%.

III. METHODOLOGY

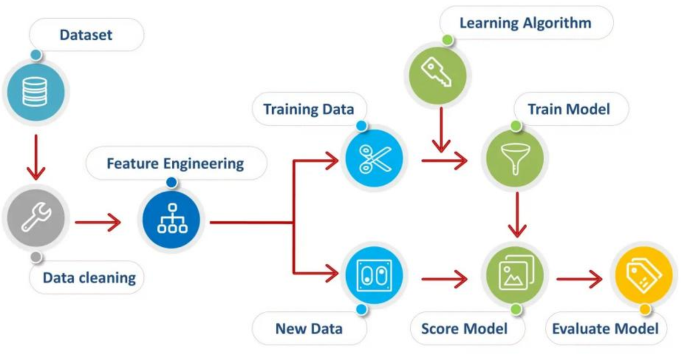


Fig 1: Workflow of machine learning

1.Data collection and preprocessing: Compile a dataset with pertinent elements, such as demographic data, medical history, and test results from labs. By handling missing values, encoding categorical variables, and normalizing numerical features, the data are preprocessed.

2.Feature selection: Identify the characteristics that are most useful and highly connected with hepatitis C. Stepwise regression, feature importance, and correlation analysis are a few examples of approaches that can be used for this.

3.Model training : Split training and testing sets from the dataset. Using the chosen features as input variables and the Hepatitis C status as the target variable, fit a logistic regression model to the training data. To improve the performance of the model, adjust hyperparameters like regularization strength.

4.Model assessment: Using appropriate evaluation metrics like accuracy, precision, assess the logistic regression model's performance. Examine the model's accuracy

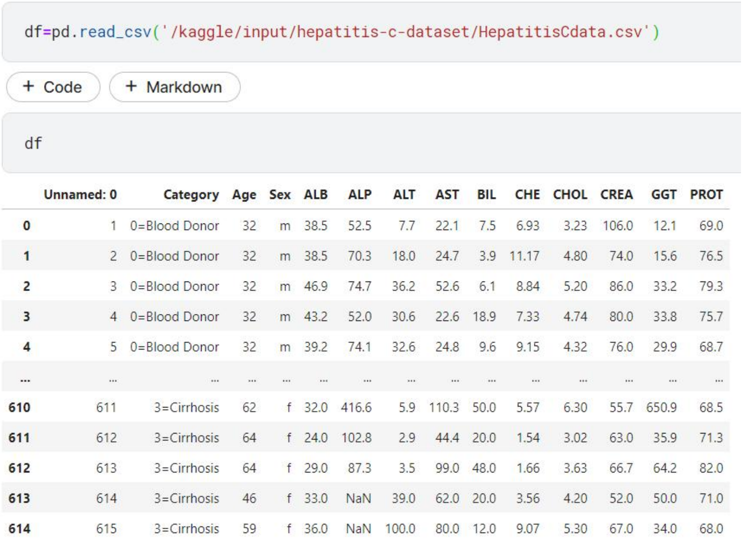
in identifying people as positive or negative for hepatitis C using the testing set.

5.Results interpretation: The logistic regression model's coefficients should be examined to determine each characteristic's impact on the risk of having hepatitis C. Using the size and direction of the coefficients, determine the main risk factors for the disease.

6.Model implementation and suggestions: After it has been developed and tested, think about applying the logistic regression model in the healthcare industry. Offer healthcare providers guidance in the form of early intervention methods or targeted screening programmes based on the model's predictions.

We have created a Machine Learning model which helps in predicting hepatitis C virus using Logistic Regression Algorithm. To develop this ML model, we used Kaggle, where we gathered the necessary dataset.

Here is how we created our ML model



```
df=pd.read_csv('/kaggle/input/hepatitis-c-dataset/HepatitisCdata.csv')
```

```
df
```

	Unnamed: 0	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
1	2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
2	3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
3	4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
4	5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
610	611	3=Cirrhosis	62	f	32.0	416.6	5.9	110.3	50.0	5.57	6.30	55.7	650.9	68.5
611	612	3=Cirrhosis	64	f	24.0	102.8	2.9	44.4	20.0	1.54	3.02	63.0	35.9	71.3
612	613	3=Cirrhosis	64	f	29.0	87.3	3.5	99.0	48.0	1.66	3.63	66.7	64.2	82.0
613	614	3=Cirrhosis	46	f	33.0	NaN	39.0	62.0	20.0	3.56	4.20	52.0	50.0	71.0
614	615	3=Cirrhosis	59	f	36.0	NaN	100.0	80.0	12.0	9.07	5.30	67.0	34.0	68.0

Fig.2: Data collection

In Fig. 2, we created a variable "df" to store the Hepatitis C dataset which we obtained from "Kaggle.com". It has 615 rows and 14 columns. As we can observe, there are certain missing(NaN) values in the above gathered dataset. We can the missing values with the mean values of their respective columns by using fillna() method and mean() method to fill and replace respectively.

Replacing categorical data values with numerical values to ensure that the data is in a format that can be used by the algorithm and also to simplify data representation and hence making the model suitable for analysis and modelling. We have replaced the values in the columns; ‘Category’ and ‘Sex’ .The values '0=Blood Donor' and '0s=suspect Blood Donor' are replaced with 0, while the values '1=Hepatitis', '2=Fibrosis', and '3=Cirrhosis' are substituted with “1” and the values 'm' and 'f' are replaced with ‘0’ and ‘1’, respectively.

We have plotted a graph including all the features on the x-axis and their respective counts on the y-axis as shown in Fig. 3.

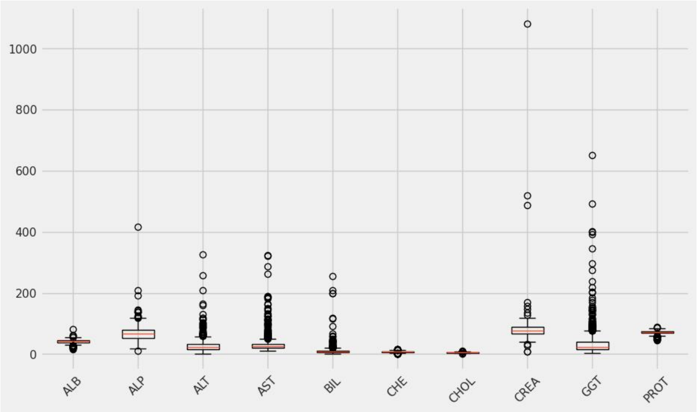


Fig. 3: Scatter plot

Removing Outliers:

As we can notice the above graph has a few outliers. Outliers can arise due to errors in data collection, measurement, or data entry. These errors can introduce noise and inaccuracies into the dataset.

Hence they have to be removed to improve data quality and reliability.Removing outliers can also help improving the model accuracy. We have accomplished this using ‘RobustScalar’ class from the sklearn.preprocessing module. Robust scaling method scales the features by taking the median out and dividing by the interquartile range (IQR), making it resistant to outliers.

IV.RESULTS AND DISCUSSION

Age Vs Count Graph:

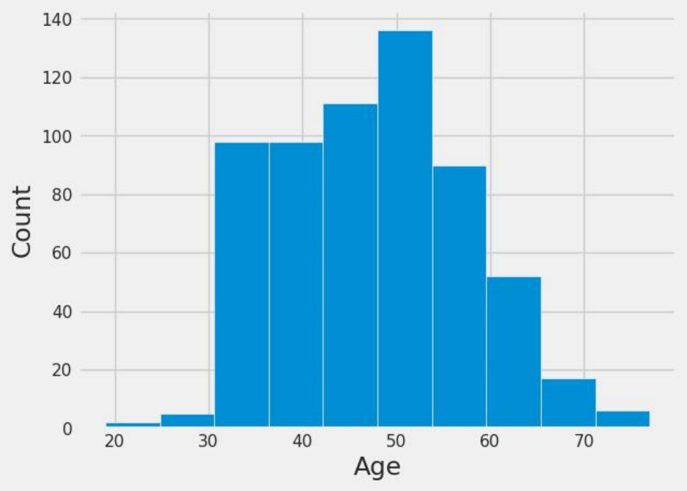


Fig. 4: Age vs count graph

In Fig. 4, Age of the patients has been plotted on x-axis and the corresponding count has been plotted on y-axis.

Ast Vs Categories Scatter Plot :

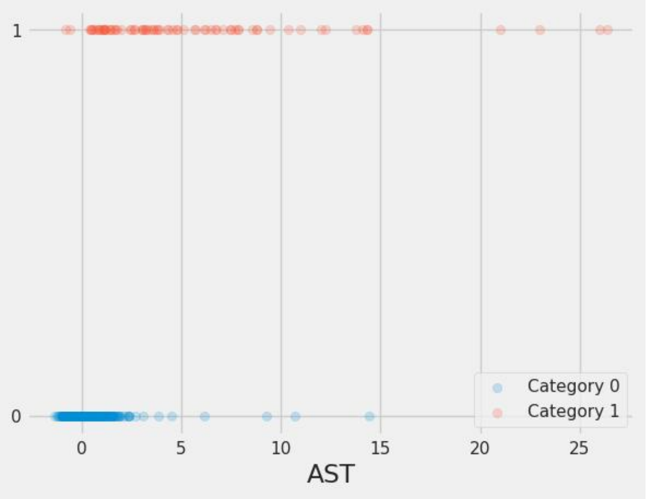


Fig 5: Ast vs Categories

In Fig, 5, a particular feature has been selected i.e, AST. The amount of AST (Aspartate transaminase) present in the blood sample of the patient has been plotted on the x-axis.

Category “0” represents healthy patients and Category “1” represents the patients infected by Hepatitis C.

The alpha=0.2 parameter has been used; It sets the transparency of the points, making them more visually distinguishable.

Healthy Vs Suspected Patients Pie Chart :

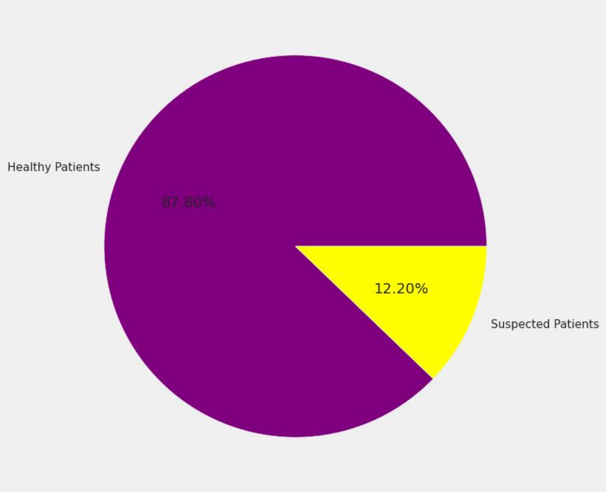


Fig 6: Healthy Vs Suspected Patients Pie Chart

Fig. 6 represents the suspected and healthy patients percentage. After calculating, the total number of suspected patients are 540 and total number of healthy patients are 75.

“plt.pie” has been used to plot this graph.

Male Vs Female Patients Pie Chart :

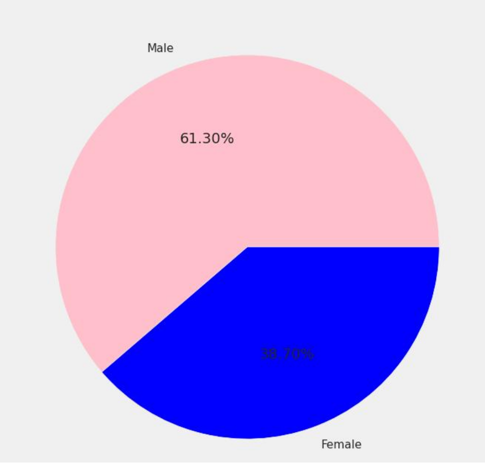


Fig 7: Male Vs Female Patients Pie Chart

The above chart represents the male and female patients percentage. The total percentage of male patients is 61.3% and total percentage of female patients is 38.7%.
“plt.pie” has been used to plot this graph.

Correlation Matrix Heatmap :
A Correlation Matrix Heatmap is a visual representation of the correlation coefficients between features in the dataset. It is commonly used to identify relationships or patterns between features and can provide insights into the strengths and direction of those relationships.

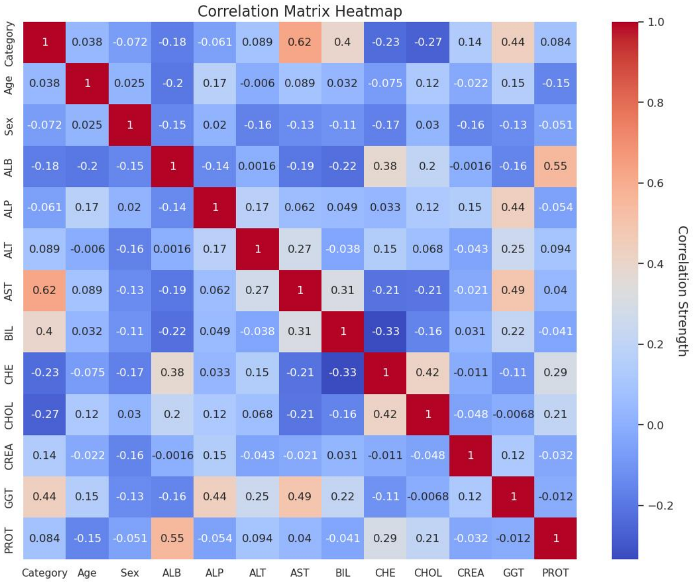


Fig 8: Correlation Matrix Heatmap

Category Vs Age Boxplot :

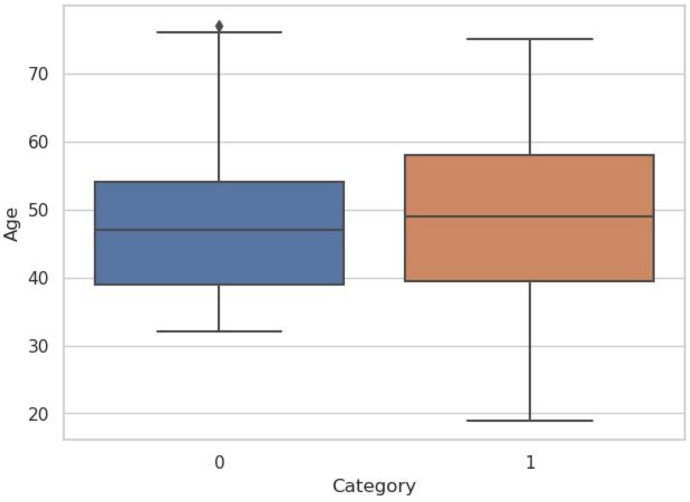


Fig 9: Category vs Age boxplot

In Fig. 9 a boxplot has been plotted between the categories and the age of the patients. It represents the age range of the patients in the 2 categories. The categories are “0” which represents the healthy patients and “1” which represents the suspected patients.

Histograms Of Features Vs Count :

Fig. 10 shows a set of histograms to visualize the distribution of various blood test results by category in a dataset. The categories are Age, ALP, ALB, ALT, AST,BIL, CHOL, CHE, CREA, GGT

Distribution of various blood test results by category have been shown in the below histograms in Fig 10.The categories and their respective counts have been plotted.

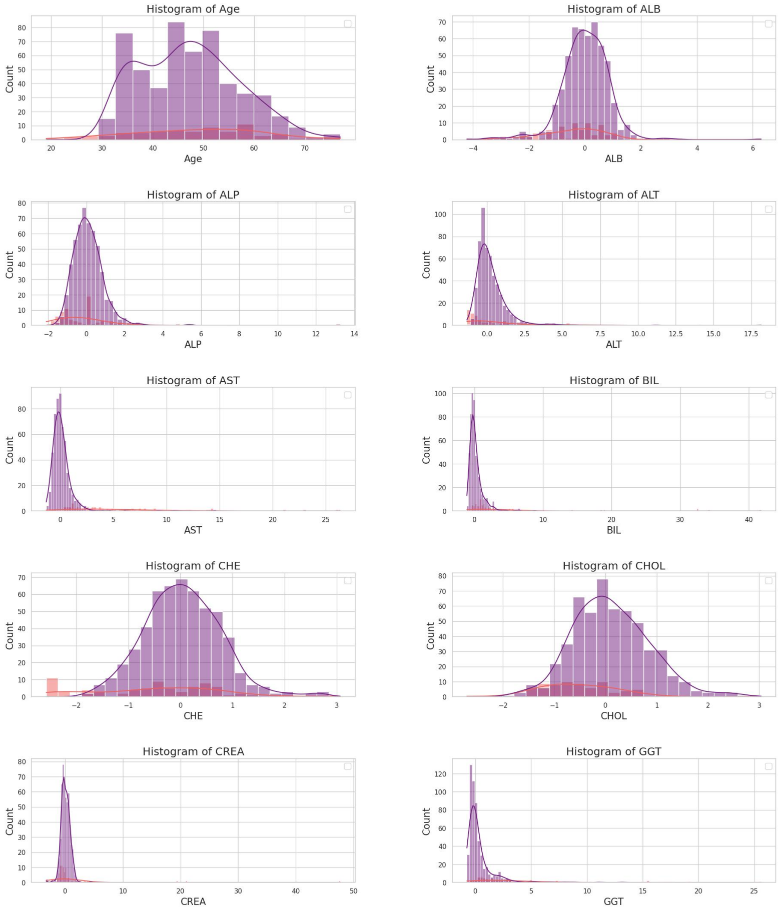


Fig 10: Histogram of Features vs Count

Step 3: Training And Testing Data :

Before Cross Validation:
Accuracy: 0.8861788617886179
Training set score: 0.98
Test set score: 0.8
Training Vs Tesing Accuracy

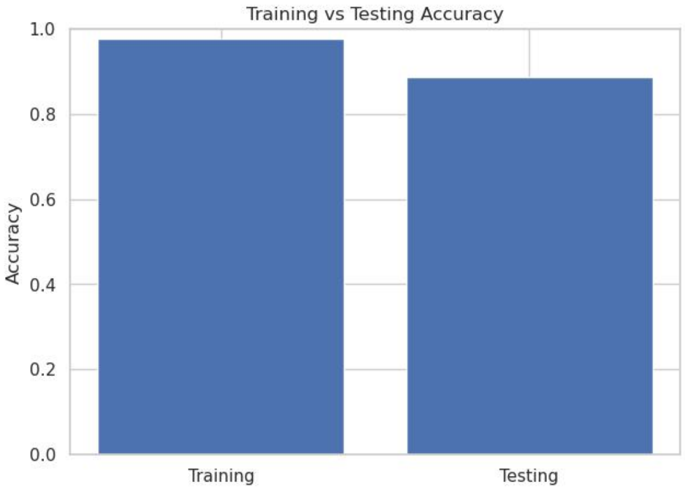


Fig 11: Training and testing bar graph

Discussion:
As we can see in Fig.11, the accuracy of the logistic regression model is **0.8861788617886179** or approximately **88.62%.**
Comparing the training set score (0.98 or 98%) to the test set score (0.89 or 89%), we can observe that the model's

performance is slightly better on the training data than on the test data. However, the difference between the training and test set scores is not substantial. Considering the relatively high training set score, there is a possibility of overfitting, where the model might have learned specific patterns or noise in the training data that do not generalize well to unseen data.

Overfitting occurs when the model learns the specific patterns or noise in the training data too well, which may cause it to struggle in generalizing to new, unseen data.

After Cross Validation:
Training Logistic Regression…
Best parameters for Logistic Regression: {'C': 10, 'max_iter': 500, 'penalty': 'l2'}
Training set score: 0.98
Test set score: 0.90
Accuracy for Logistic Regression: 0.9024390243902439
Training Vs Tesing Accuracy

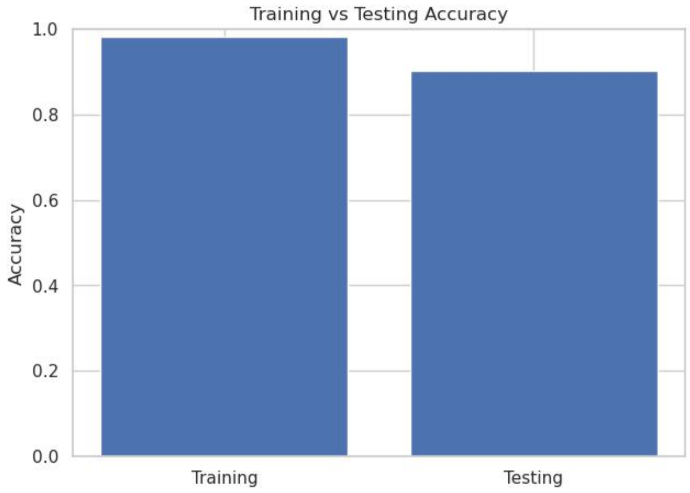


Fig 12: Training and testing bar graph

Discussion:

As we can see in Fig. 12, the accuracy of the logistic regression model is **0.9024390243902439** or approximately **90.24%.** Comparing the training set score (0.98 or 98%) to the test set score (0.90 or 90%), we can observe that the model's performance is slightly better on the training data than on the test data. However, the difference between the training and test set scores is not substantial. In the first case, the accuracy of the logistic regression model was 0.8861788617886179 or approximately 88.62%. However, in the second case, the accuracy has increased to 0.9024390243902439 or approximately 90.24%. This improvement in accuracy suggests that the model is performing better in terms of its ability to predict the correct outcome. The increase in accuracy could be attributed to various factors, such as the tuning of hyperparameters or the inclusion of more relevant features in the model. Additionally, it's worth noting that the training set score remains the same in both cases, at 0.98 or 98%. This indicates that the model is consistently performing well on the training data. By using GridSearchCV, the second code performs an exhaustive search over the specified hyperparameter values. It evaluates different combinations of hyperparameters using cross-validation and selects the best combination that maximizes the chosen scoring metric (accuracy in this case).

V.FUTURE SCOPE:

A logistic regression machine learning (ML) model has a good chance of accurately predicting the Hepatitis C virus (HCV). Here are some potential areas for study and application:

Early detection and diagnosis: Logistic regression models can be trained on historical patient data, including demographic, clinical, and laboratory characteristics, to predict the likelihood of HCV infection. These models can aid in the early detection and diagnosis of viruses, allowing for quick action and treatment.

Risk assessment: Based on multiple variables like age, gender, lifestyle, medical history, and location, logistic regression models can be used to evaluate the risk of HCV infection in diverse populations or people. This can assist healthcare providers in better allocating resources and prioritising preventative measures.

Logistic regression models can be trained using data from patients who have had HCV treatment to forecast the likelihood that a treatment would be successful or unsuccessful. This can help doctors choose the best treatment plans and keep track of patients' development over the course of therapy.

ML models, such as logistic regression, can be used to create personalised HCV treatment regimens for specific patients. These models can suggest customised treatment options that optimise outcomes and reduce negative effects by taking into account patient-specific variables, such as genetic markers and comorbidities.

Public health planning: To identify high-risk populations, transmission patterns, and risk factors for HCV infection, large-scale epidemiological data can be analysed using logistic regression models. This information can direct policy-making, targeted screenings, prevention programmes, and other public health actions.

Logistic regression algorithms can be used to forecast outbreaks and identify areas at high risk of rising HCV prevalence by analysing real-time data on HCV cases and related risk variables. With the help of these findings, preventative measures like resource allocation, preventive education, and quick response tactics can be implemented.

Integration with other ML methods: To improve prediction accuracy and give a more thorough insight of HCV dynamics, logistic regression can be integrated with other ML algorithms such ensemble methods or deep learning models. Performance can be optimised by utilising the advantages of several algorithms through this combination.

The logistic regression model, which implies a linear relationship between the predictors and the result variable, is vital to keep in mind. More complex ML algorithms can be necessary in cases with non-linear interactions. Additionally, for the successful application of logistic regression or any ML approach in predicting HCV, high-quality data accessibility, proper model validation, and evaluation are essential.

VI.CONCLUSION:

Although limited in their ability to predict complex events, linear regression models may nevertheless be useful in some parts of HCV prediction. Linear regression models can be used for risk assessment, early detection, and the prediction of therapy response by looking at demographic, clinical, and laboratory data. They can assist identify high-risk patients and improve resource allocation while supporting approaches to tailored care.

But it's important to recognise that HCV is a complex illness that is influenced by a variety of genetic, environmental, and behavioural variables. The complex nonlinear interactions in the dynamics of HCV are not always captured by linear regression models, which presume a linear connection between predictors and outcomes. To improve prediction accuracy and comprehension, it is crucial to take into account the limitations of linear regression and investigate more sophisticated machine learning algorithms.

In conclusion, while linear regression models can offer insightful information about predicting HCV, they should be used in concert with other machine learning techniques and domain expertise to create thorough and precise prediction models. Any prediction model's success also depends on the availability of high-quality data, thorough model validation, and regular updates to take into account fresh information and developments in the field of HCV research.

VII.ABBREVIATIONS:

HCV: Hepatitis C virus; ML : Machine Learning; LFT: Liver Function Test; LR: Logistic Regression; RNA : Ribonucleic acid; ALB: albumin; ALP: Alkaline phosphatase; ALT: alanine aminotransferase; AST: aspartate aminotransferase; BIL: bilirubin; CHE: choline esterase; CHOL: cholesterol; CREA: creatinine blood test; GGT: γ-glutamyl-transferase; PROT: total protein test;

VII.REFERENCES:

[1].Hepatitis C virus data analysis and prediction using machine learning
Author links open overlay panelMete Yağanoğlu

[2].Artificial intelligence for hepatitis evaluation
Wei Liu, Xue Liu, Mei Peng, Gong-Quan Chen, Peng-Hua Liu, Xin-Wu Cui, Fan Jiang, and Christoph F Dietrich

[3].Applying data mining techniques to classify patients with suspected hepatitis C virus infection
Author links open overlay panelReza Safdari 1, Amir Deghatipour 2, Marsa Gholamzadeh 1, Keivan Maghooli 3

[4].Diagnosis of Hepatitis Disease with Logistic Regression and Artificial Neural Networks
1,2Alaa M. Elsayad, 1,3Ahmed M. Nassef and 4Mujahed Al-Dhaifallah

[5].World Health Organization (2021) Hepatitis C, WHO fact sheet No. 164, updated July 2021.

[6].Konerman MA, Beste LA, Van T et al (2019) Machine learning models to predict disease progression among veterans with Hepatitis C virus.

[7].Lindenmeyer CC (2021) Laboratory tests of the liver and gallbladder. Merck Manual Professional Edition. Updated December 2019. Accessed May 10, 2021

[8].American Liver Foundation (2017) [Internet]. New York: American Liver Foundation; c2017. Liver Function Tests; [updated 2016 Jan 25; cited 2017 Mar 13];

[9].Haga H, Sato H, Koseki A, Saito T, Okumoto K et al (2020) A machine learning-based treatment prediction model using whole genome variants of hepatitis C virus.

[10].Haga H, Sato H, Koseki A, Saito T, Okumoto K et al (2020) A machine learning-based treatment prediction model using whole genome variants of hepatitis C virus. PLoS One 15(11):e0242028.

[11].https://www.kaggle.com/code/ragishehab/hepatitis-c-prediction-achieving-94-accuracy

[12].Doyle OM, Leavitt N, Rigg JA (2020) Finding undiagnosed patients with Hepatitis C infection: An application of artificial intelligence to patient claims data. Sci Rep

[13].Yarasuri VK, Indukuri GK, Nair AK (2019) Third international conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) hepatitis diseases prediction using machine-learning technique (I-SMAC). IEEE

[14].Begg R (2009) Artificial intelligence techniques in medicine and health care. In: Sugumaran V (ed) Concepts, methodologies, tools, and application. ISBN: 9781599049410

[15].Davenport T, Kalakota R (2019) The potential for artificial intelligence in healthcare. Fut Healthcare J 6(2):94–98

[16].Barakat NH, Barakat SH, Ahmed N (2019) Prediction and staging of hepatic fibrosis in children with hepatitis C virus: a machine learning approach. Healthcare Informat Res 25(3):173–181

[17].Konerman MA, Beste LA, Van T et al (2019) Machine learning models to predict disease progression among veterans with Hepatitis C virus. PLoS ONE 14(1):14

[18].Krajden M (2001) Hepatitis. Canad J Infect Dis 12(6):329–31

[19].Gomaa A, Allam N, Elsharkway A et al (2017) Hepatitis C infection in Egypt: prevalence, impact and management strategies. Hepatic Med Evid Res 8:17–25

[20].Chicco D, Jurman G (2021) An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis. IEEE Access 9:24485–24498

[21].Nandipati SC, XinYing C, Wah KK (2020) Hepatitis C virus (HCV) prediction by machine learning techniques. Appl Model Simul 4:89–100

[22].Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt
Heba Mamdouh Farghaly, Mahmoud Y. Shams & Tarek Abd El-Hafeez.

By,
Nithya Santhoshini M (1NT21EC096)
Surepally Vishnnu Vardhini (1NT21EC156)
4th Sem,A Section