Prasad V Potluri Siddhartha Institute Of Technology

Python Project on

LIVER CIRRHOSIS PREDICTION

Presented by :

-VIMALA	22501a05i4
-ESWAR ADITYA	22501a05j8
-MEGHANA	22501a05e3
-DIVYA CHARITHA	22501a05i9

CONTENTS

INTRODUCTION

- **Liver cirrhosis** is a widespread problem especially in North America due to the **high intake of alcohol**.
- In this project, we will predict **liver cirrhosis** in a patient based on certain lifestyle and health conditions of a patient.
- Cirrhosis is a late stage of scarring (fibrosis) of the liver caused by many forms of liver diseases and conditions, such as **hepatitis** and **chronic alcoholism**.

OBJECTIVE

01.

Early Detection:
Develop a machine learning model capable of early detection of potential Liver Cirrhosis based on relevant clinical features.

02.

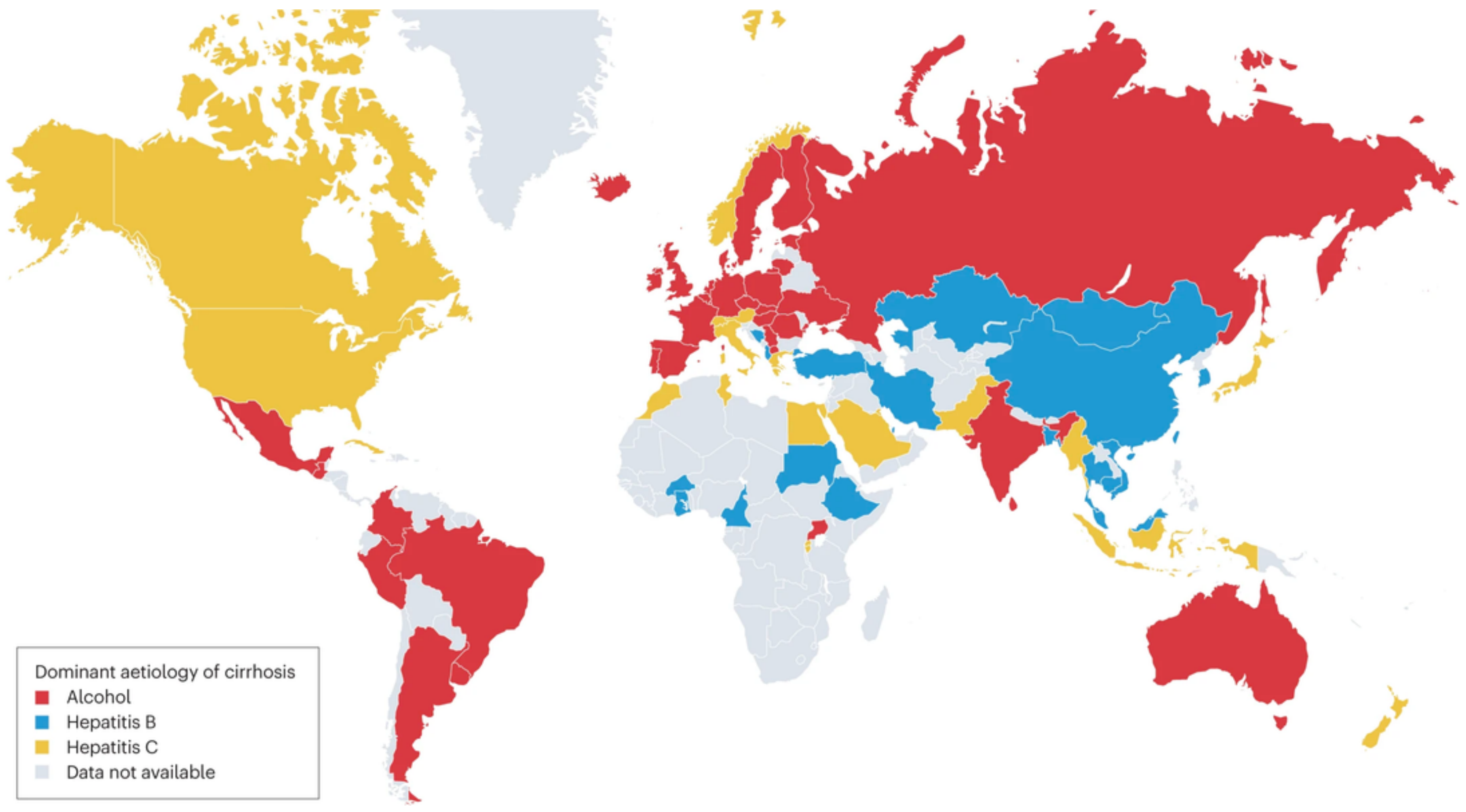
Accuracy Improvement:
Improve the accuracy and reliability of Liver Cirrhosis prediction compared to traditional risk assessment methods.

03.

Real-time Prediction:
Enable real-time prediction to provide immediate insights, allowing for prompt medical intervention and improving patient outcomes.

BACKGROUND

Dominant reported aetiology of cirrhosis from 1993 to 2021.



- Cirrhosis is an important cause of morbidity and mortality in people with chronic liver disease worldwide.
- In 2019, cirrhosis was associated with 2.4% of global deaths.
- Owing to the rising prevalence of obesity and increased alcohol consumption on the one hand, and improvements in the management of hepatitis B virus and hepatitis C virus infections on the other.

MOTIVATION

- This project on Liver Cirrhosis Prediction using Python ML is motivated by the urgent need for early detection solutions.
- The project aligns with a broader goal of making liver health assessments more accessible, particularly in regions with limited resources.
- Through data-driven insights, we strive to contribute to the understanding of cirrhosis dynamics.
- This endeavour showcases the impactful synergy between advanced technology and healthcare, addressing a critical need in public health.

PROBLEM STATEMENT

This project addresses the challenge of precise identification for individuals at risk of liver cirrhosis.

DATA COLLECTION

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 20 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   ID             418 non-null    int64  
 1   N_Days         418 non-null    int64  
 2   Status         418 non-null    object 
 3   Drug           312 non-null    object 
 4   Age            418 non-null    int64  
 5   Sex            418 non-null    object 
 6   Ascites        312 non-null    object 
 7   Hepatomegaly   312 non-null    object 
 8   Spiders        312 non-null    object 
 9   Edema          418 non-null    object 
 10  Bilirubin      418 non-null    float64
 11  Cholesterol    284 non-null    float64
 12  Albumin        418 non-null    float64
 13  Copper         310 non-null    float64
 14  Alk_Phos       312 non-null    float64
 15  SGOT           312 non-null    float64
 16  Tryglicerides  282 non-null    float64
 17  Platelets      407 non-null    float64
```

- The dataset employed for predicting is obtained from Kaggle [3], featuring 418 entries distributed across 20 columns.
- These columns encompass 'id,' 'N_Days,' 'age,' 'Sex,' 'Edema,' 'Albumin,' 'Copper,' 'Spiders,' 'Alk_Phos,'as the key attributes.

DATA PRE-PROCESSING

```
data.drop_duplicates() #removing the duplicates for dataframe
```

	ID	N_Days	Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders
0	1	400	D	D-penicillamine	21464	F	Y	Y	Y
1	2	4500	C	D-penicillamine	20617	F	N	Y	Y
2	3	1012	D	D-penicillamine	25594	M	N	N	N
3	4	1925	D	D-penicillamine	19994	F	N	Y	Y
4	5	1504	CL	Placebo	13918	F	N	Y	Y
...	...	...	...	...	...	...	...	...	...
413	414	681	D	NaN	24472	F	NaN	NaN	NaN
414	415	1103	C	NaN	14245	F	NaN	NaN	NaN
415	416	1055	C	NaN	20819	F	NaN	NaN	NaN
416	417	691	C	NaN	21185	F	NaN	NaN	NaN

- In the data preprocessing phase, we addressed missing values

- standardized numerical features, and encoded categorical variables for the cirrhosis prediction dataset.

DATA PRE-PROCESSING

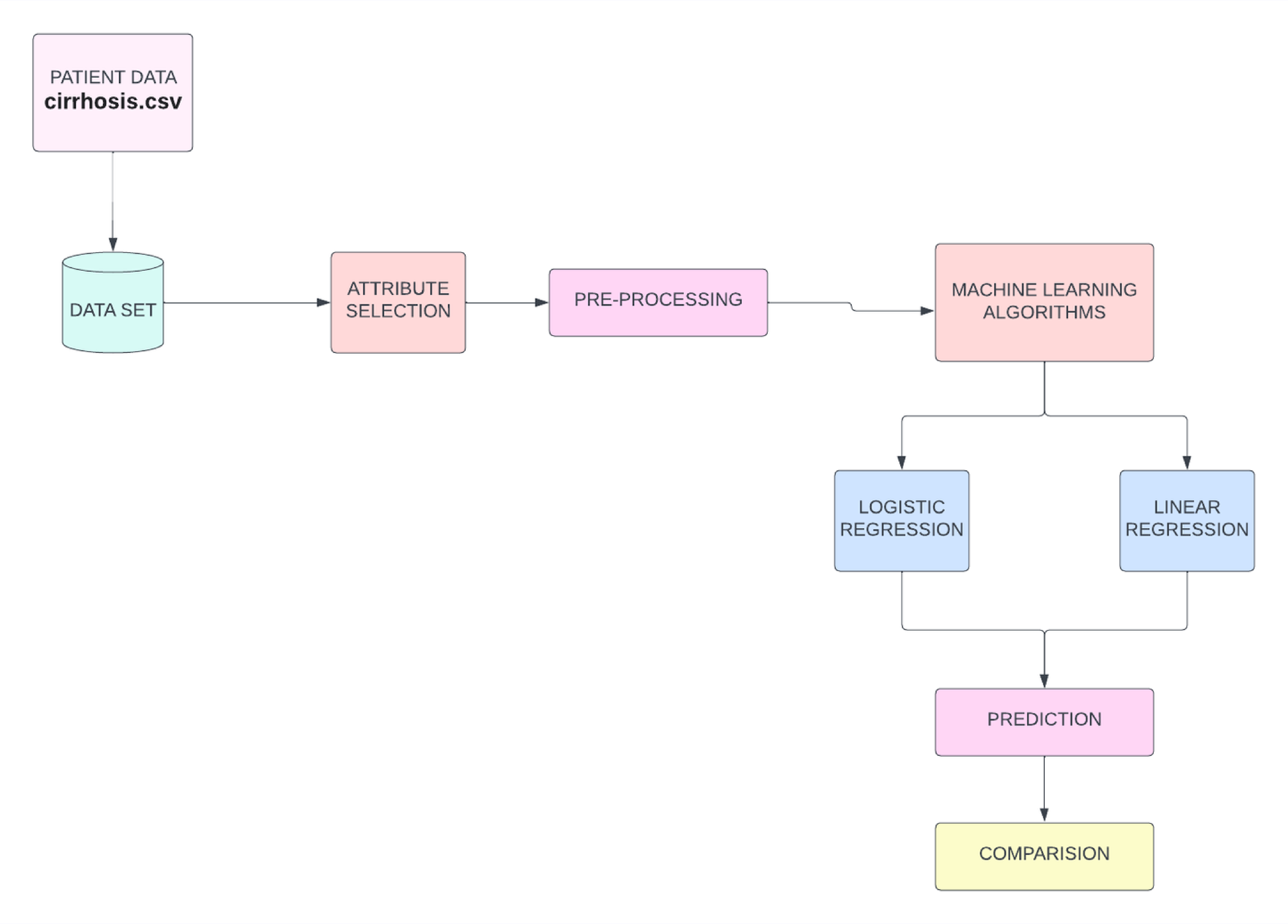
```
newdata=newdata.bfill() #filling null values with backward values
```

Feature Selection:

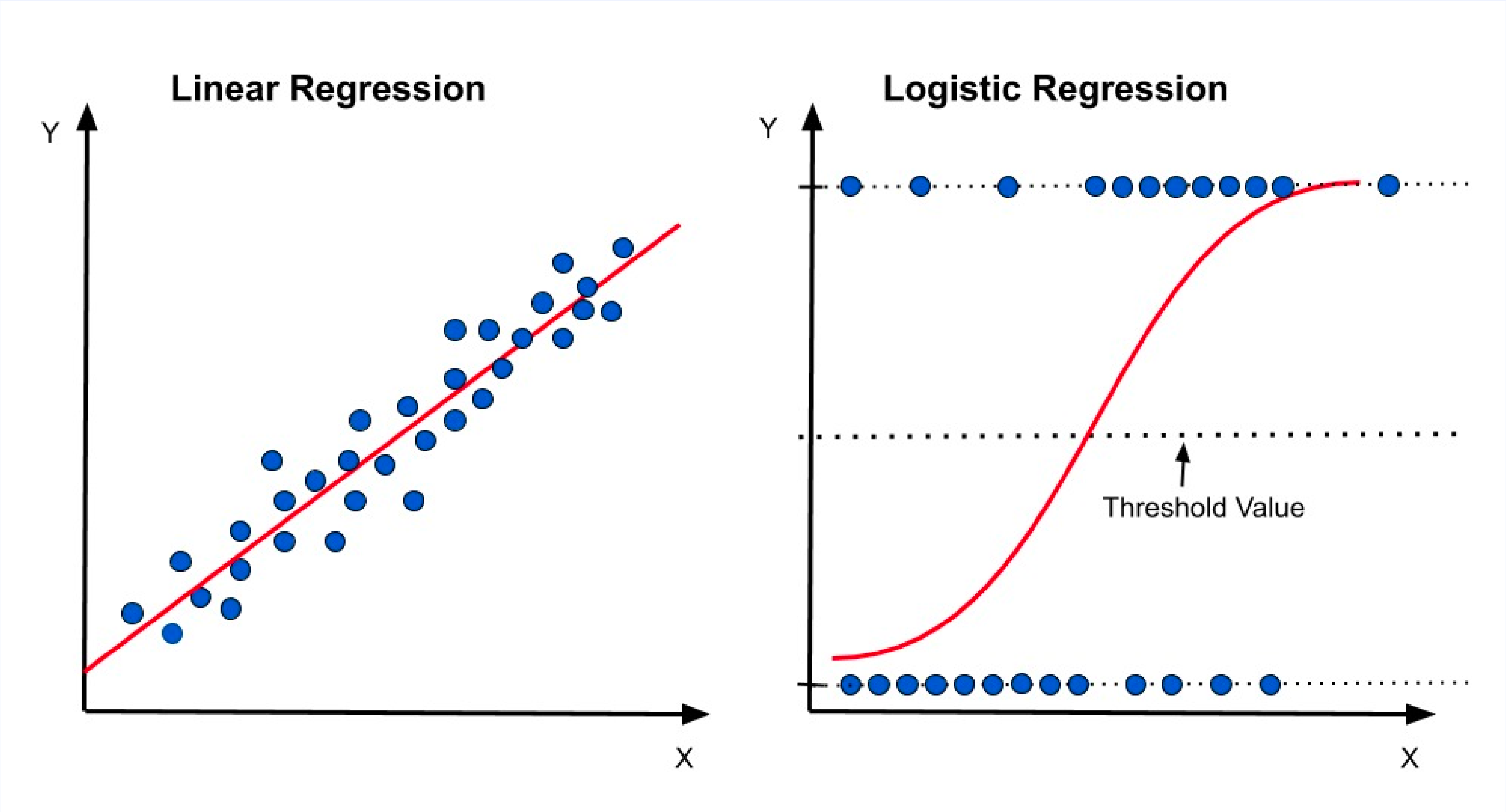
- This involved selecting the most relevant features from the dataset to improve the accuracy of the machine learning model.

- This was done using techniques such as correlation analysis and feature importance ranking.

DATA FLOW DIAGRAM



IMPLEMENTATION



Linear Regression:

- Prediction method
- Used for forecasting continuous outcome variables
- Assumes a linear relationship between predictor variables and the outcome

Logistic Regression:

- Classification method
- Predicts the probabilityof an instance belonging to a specific category
- Ideal for categorical outcomes

CODE DEVELOPMENT

```python
# replacing catagorical data with intigers.
newdata['Sex'] = newdata['Sex'].replace({'M':0, 'F':1})                       # Male : 0 , Female :1
newdata['Ascites'] = newdata['Ascites'].replace({'N':0, 'Y':1})               # N : 0, Y : 1
newdata['Drug'] = newdata['Drug'].replace({'D-penicillamine':0, 'Placebo':1}) # D-penicillamine : 0, Placebo : 1
newdata['Hepatomegaly'] = newdata['Hepatomegaly'].replace({'N':0, 'Y':1})     # N : 0, Y : 1
newdata['Spiders'] = newdata['Spiders'].replace({'N':0, 'Y':1})               # N : 0, Y : 1
newdata['Edema'] = newdata['Edema'].replace({'N':0, 'Y':1, 'S':-1})           # N : 0, Y : 1, S : -1
newdata['Status'] = newdata['Status'].replace({'C':0, 'CL':1, 'D':-1})        # 'C':0, 'CL':1, 'D':-1
```

```python
from sklearn import preprocessing
import pandas as pd
numeric_columns = ['Stage', 'Bilirubin', 'Cholesterol', 'Albumin']
non_numeric_columns = set(newdata.columns) - set(numeric_columns)
scaled_numeric = preprocessing.normalize(newdata[numeric_columns], axis=0)
scaled_df = pd.concat([pd.DataFrame(scaled_numeric, columns=numeric_columns), newdata[non_numeric_columns]], axis=1)
scaled_df.head()
```

	Stage	Bilirubin	Cholesterol	Albumin	Alk_Phos	Spiders	Copper	Age	Platelets	Drug	Edema	Ascites	N_Days	Hepatomegaly	Prothrombin	Sex	SGOT	Tryglicerides	ID	Status
0	0.071750	0.147295	0.035517	0.041524	1718.0	1	156.0	21464	190.0	0	2	1	400	0.0	12.2	1	137.95	172.0	1	-1
1	0.053812	0.011174	0.041096	0.066118	7394.8	1	54.0	20617	221.0	0	0	0	4500	0.0	10.6	1	113.52	88.0	2	0
2	0.071750	0.014222	0.023950	0.055578	516.0	0	210.0	25594	151.0	0	1	0	1012	1.0	12.0	0	96.1	55.0	3	-1
3	0.071750	0.018285	0.033204	0.040565	6121.8	1	64.0	19994	183.0	0	1	0	1925	0.0	10.3	1	60.63	92.0	4	-1
4	0.053812	0.034538	0.037966	0.056376	671.0	1	143.0	13918	136.0	1	0	0	1504	0.0	10.9	1	113.15	72.0	5	1

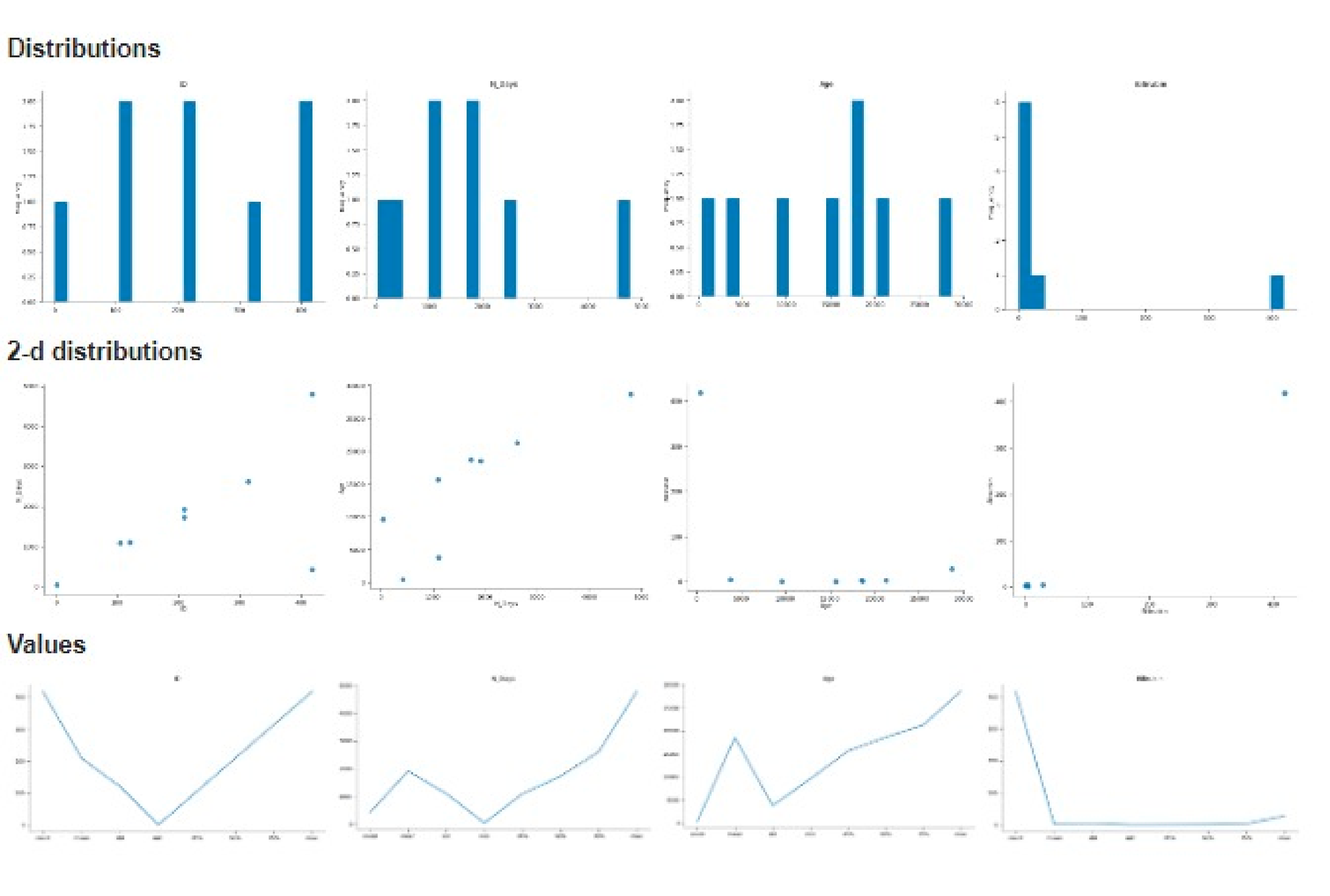
```python
from sklearn.model_selection import train_test_split #training and testing data split
from sklearn import metrics #accuracy measure
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score, classification_report #for confusion matrix
from sklearn.linear_model import LogisticRegression,LinearRegression #logistic regression
```

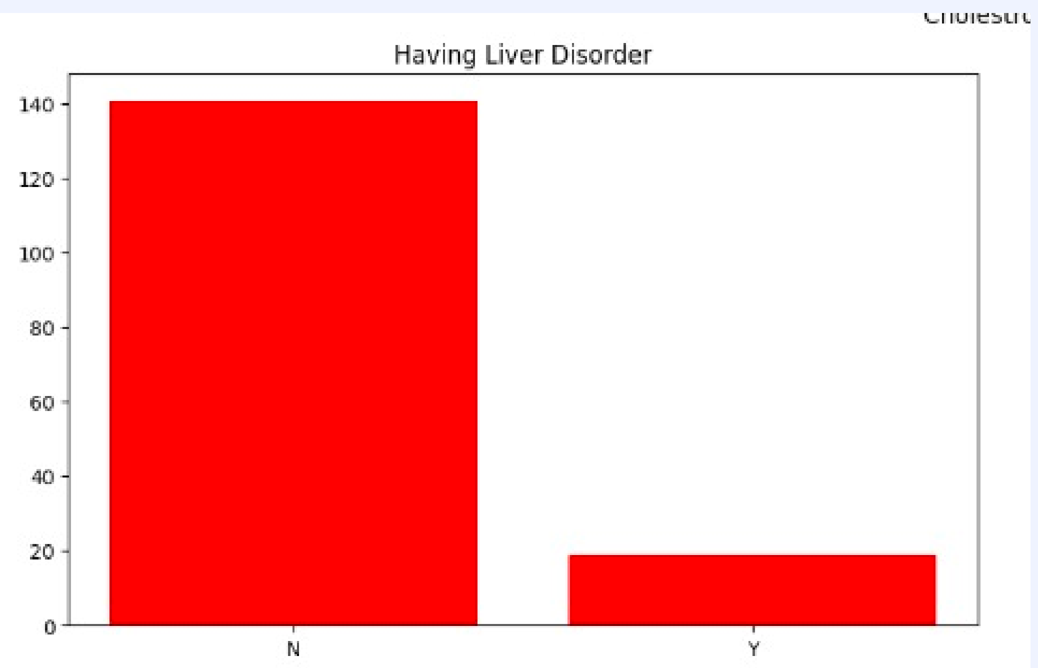
```python
from sklearn.preprocessing import LabelEncoder # converts gender into numbers
label_encoder = LabelEncoder()
categorical_cols_to_encode = ['Sex']
for col in categorical_cols_to_encode:
    newdata[col] = label_encoder.fit_transform(newdata[col])
newdata.head()
```

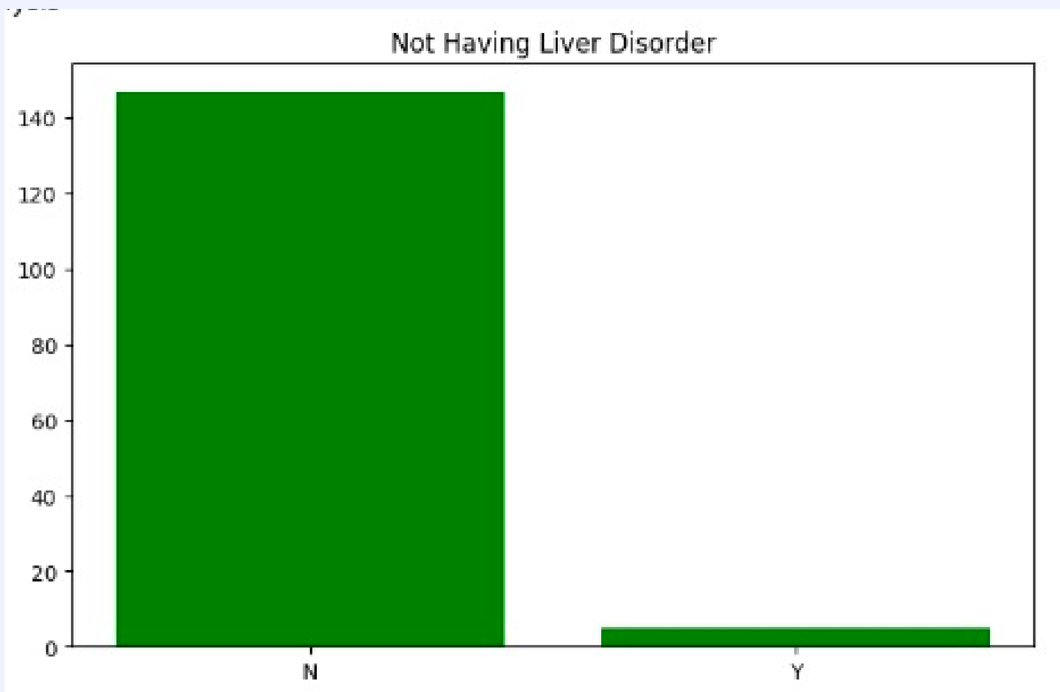
	ID	N_Days	Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage
0	1	400	-1	0	21464	1	1	0.0	1	2	14.5	261.0	2.60	156.0	1718.0	137.95	172.0	190.0	12.2	4.0
1	2	4500	0	0	20617	1	0	0.0	1	0	1.1	302.0	4.14	54.0	7394.8	113.52	88.0	221.0	10.6	3.0
2	3	1012	-1	0	25594	0	0	1.0	0	1	1.4	176.0	3.48	210.0	516.0	96.1	55.0	151.0	12.0	4.0
3	4	1925	-1	0	19994	1	0	0.0	1	1	1.8	244.0	2.54	64.0	6121.8	60.63	92.0	183.0	10.3	4.0
4	5	1504	1	1	13918	1	0	0.0	1	0	3.4	279.0	3.53	143.0	671.0	113.15	72.0	136.0	10.9	3.0

```python
train,test=train_test_split(newdata,test_size=0.3,random_state=0,stratify=newdata['Stage'])
```

VISUALIZATION THROUGH GRAPHS







RESULTS

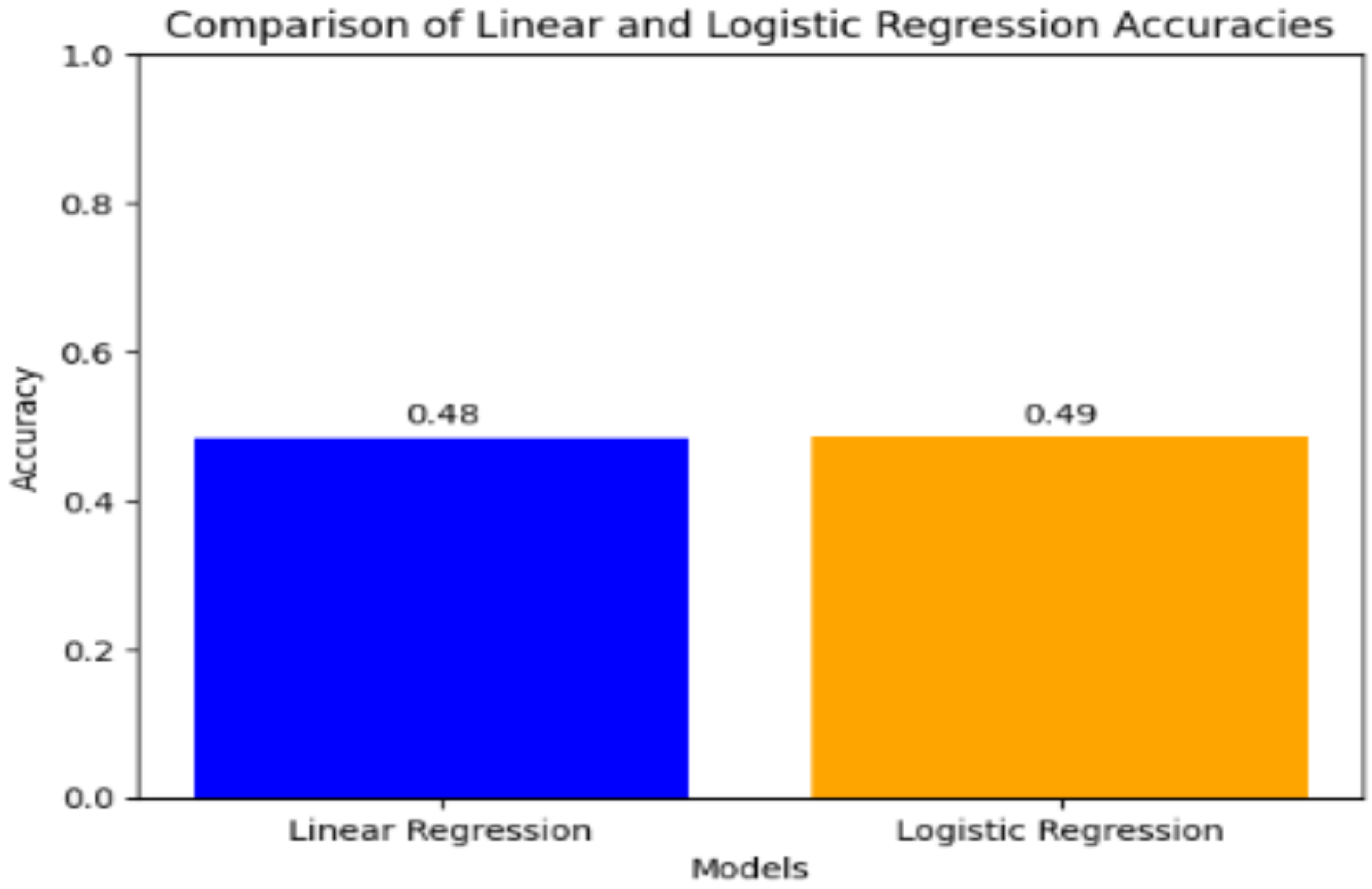
	PRECISION	RECALL	F1-SCORE	SUPPORT
1.0	0.00	0.00	0.00	4
2.0	0.33	0.14	0.19	22
3.0	0.48	0.77	0.59	39
4.0	0.55	0.44	0.49	36
ACCURACY			0.49	101
MACRO AVG	0.34	0.34	0.32	101
WEIGHTED AVG	0.45	0.49	0.44	101

Accuracy of Linear Regression is	0.4856183410909
Accuracy of Logistic Regression is	0.4851485148514851
Mean Squared Error is	0.4846183410909779
Root Mean Squared Error is	0.6961453448030648
Mean Absolute Error is	0.5610593793847464
R-Squared Error is	0.1718490640691145

- **The Linear Regression and Logistic Regression** has proven highly effective in predicting the presence of Liver Cirrhosis disease, boasting an impressive accuracy rate of **0.49%.**
- The application of machine learning techniques, including Linear and logistic Regressions holds promise for achieving precise predictions.
- However, it's essential to acknowledge that the obtained accuracy can vary based on factors such as the dataset, considered features, and intricacies of the model.

RESULTS

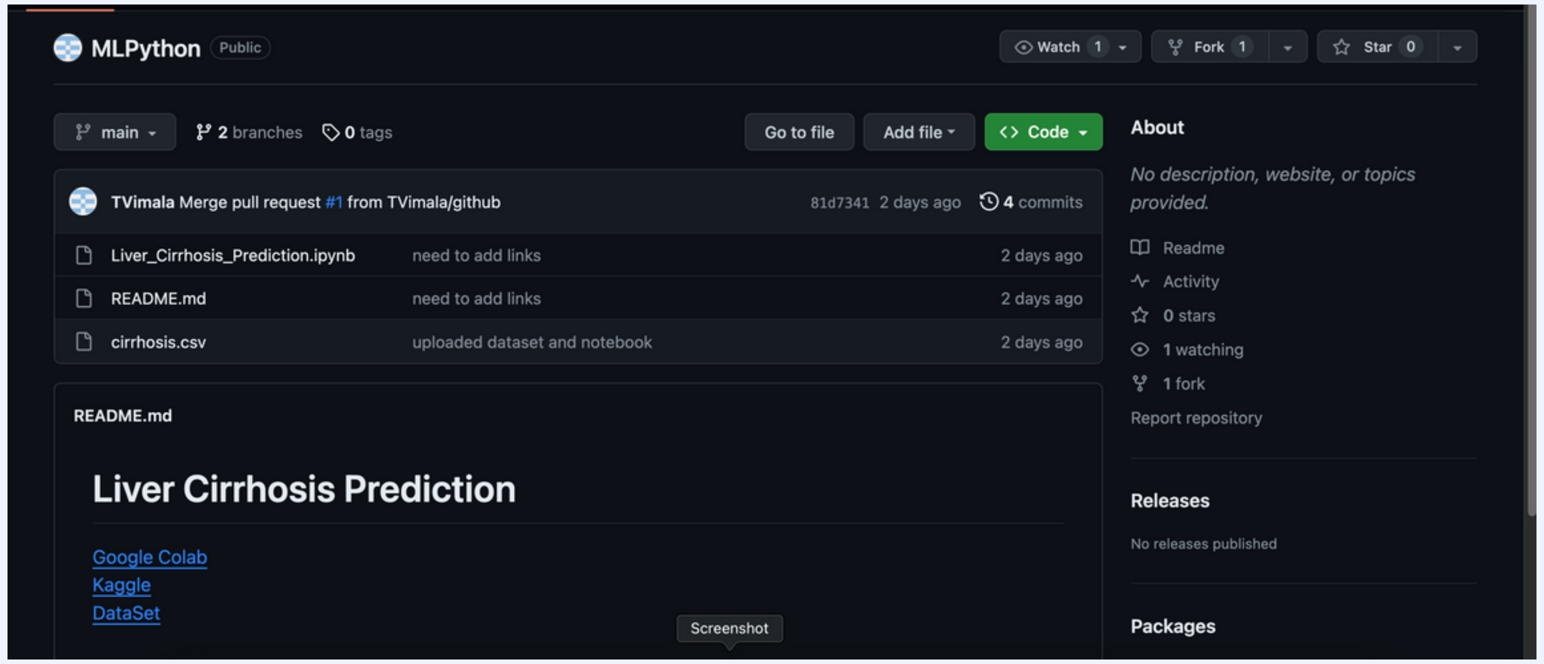
Accuracy of Linear Regression	0.4856183410909
Accuracy of Logistic Regression	0.4851485148514851

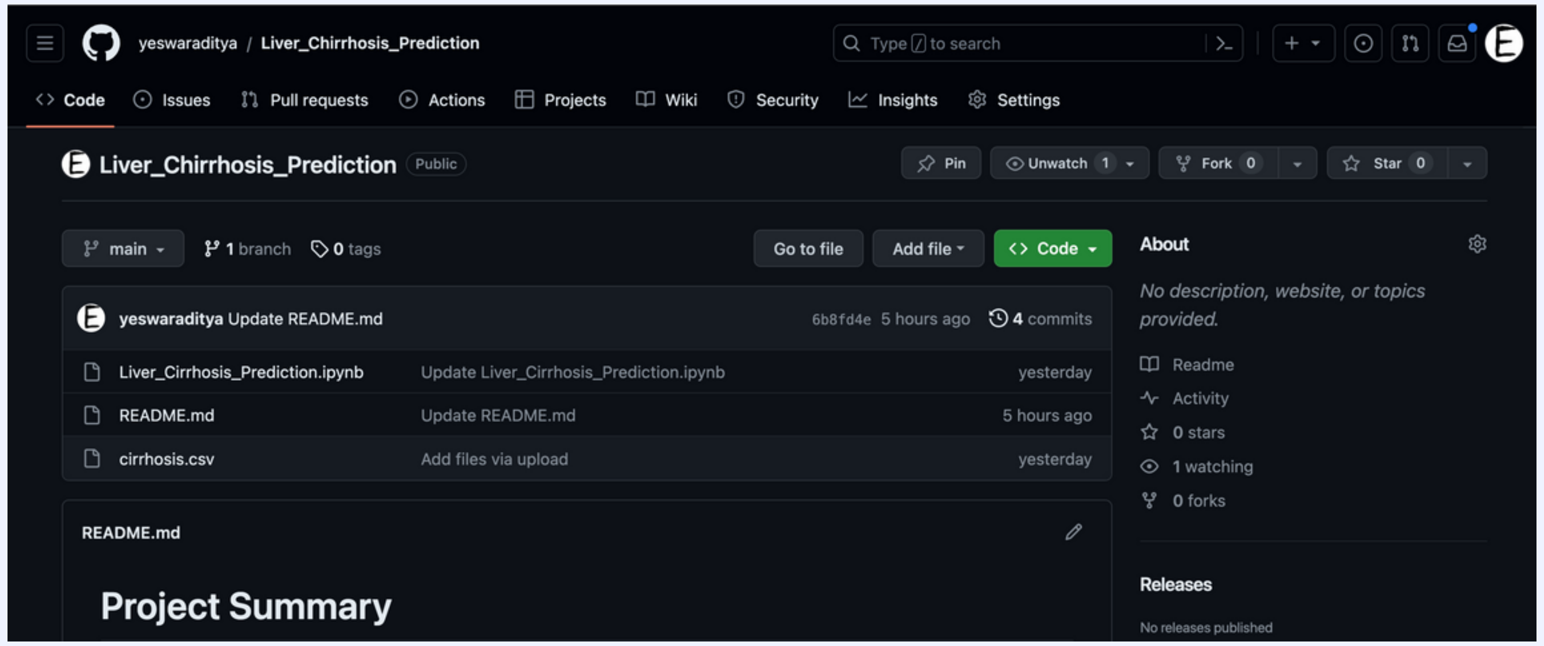


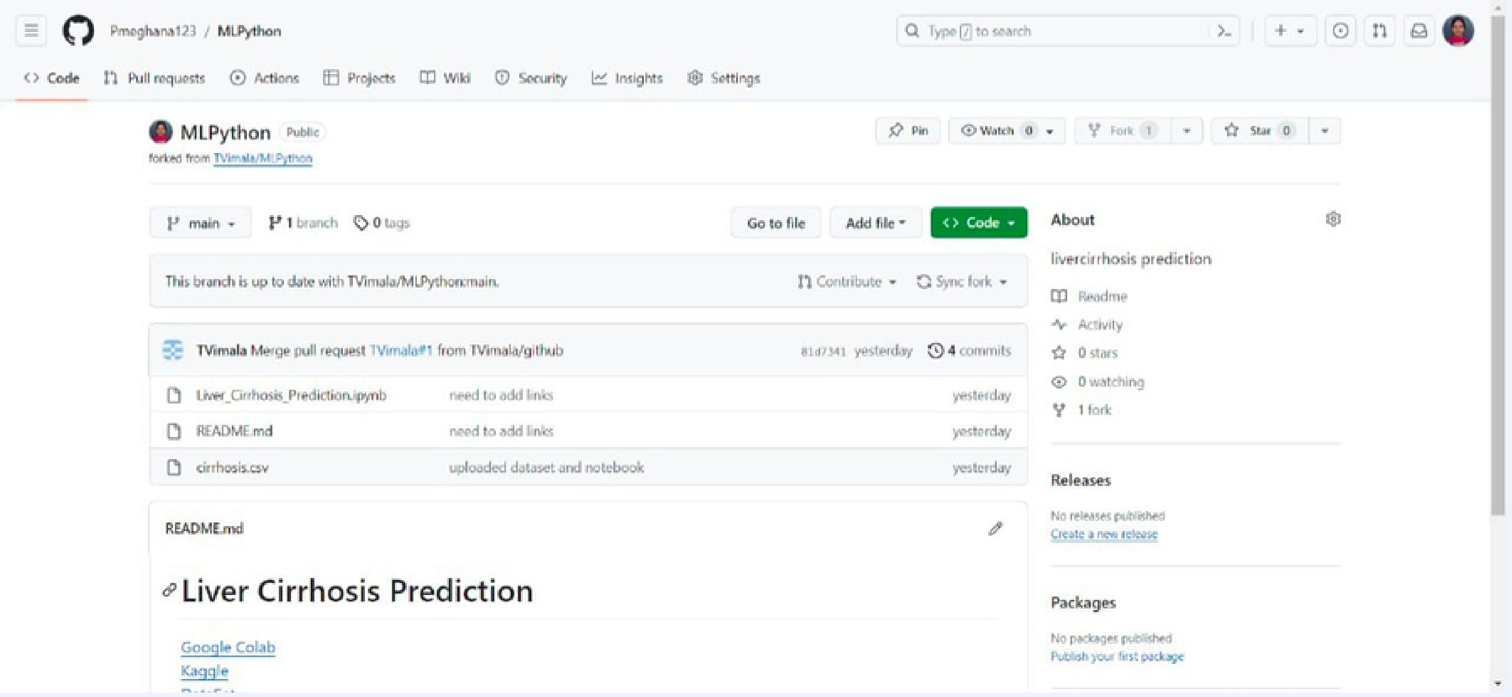
CONCLUSION

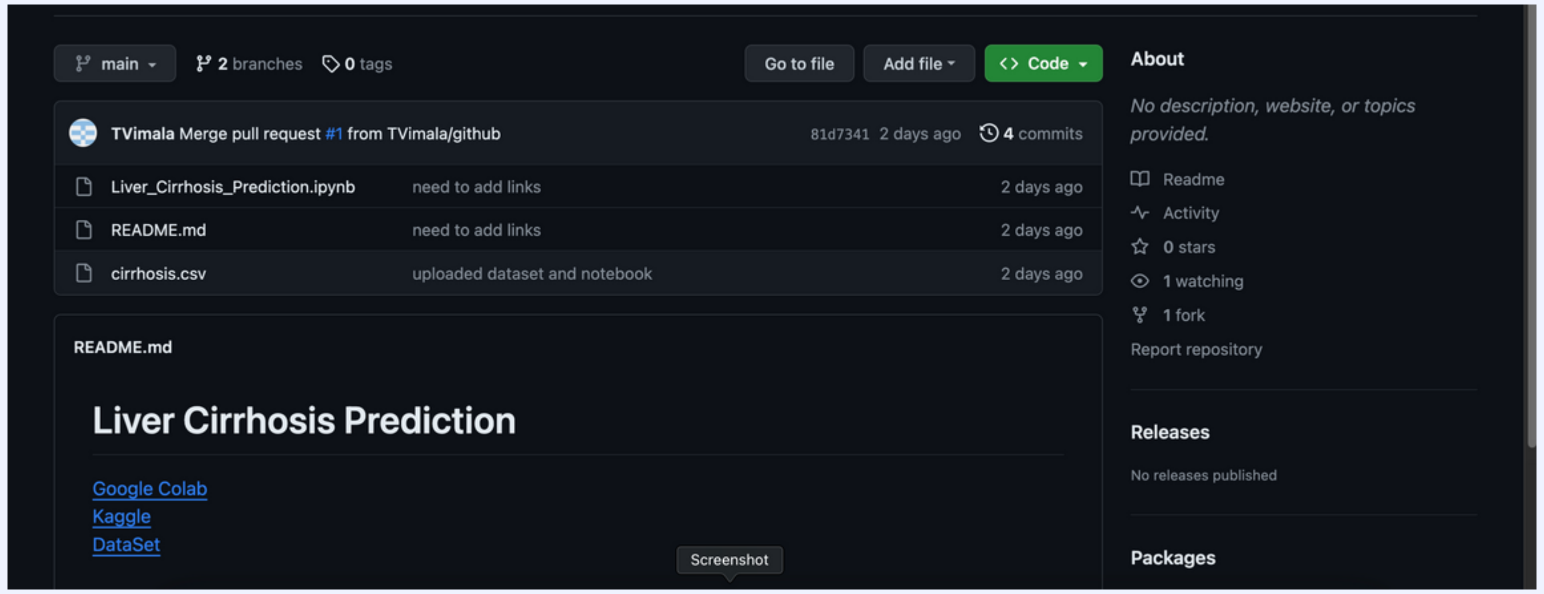
The results of this project suggest that there is considerable promise in utilizing machine learning to create precise and responsive approaches for classifying. Employing machine learning models in this scenario has the potential to facilitate earlier diagnosis and intervention, ultimately enhancing patient outcomes.

AVAILABILITY FOR OPEN SOURCING









THANK YOU