

# 1. Introduction

## 1.1. Project Overview

This project focuses on developing a machine learning model to predict liver cirrhosis based on patient health data. The goal is to provide a tool for early detection and diagnosis to aid in timely medical intervention.

## 1.2. Objectives

- To collect and preprocess patient health data.
- To explore and analyze the data to understand key features.
- To develop and evaluate different machine learning models.
- To optimize and tune the selected model for better performance.
- To integrate the model into a web application for easy accessibility.

# 2. Project Initialization and Planning Phase

## 2.1. Define Problem Statement

The problem is to predict whether a patient is suffering from liver cirrhosis based on various health metrics and historical data.

Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	Researcher studying liver diseases	identify patterns and factors contributing to cirrhosis	The sheer volume of medical data, including patient histories, lab results, imaging studies, and genetic information, can be overwhelming.	Traditional statistical methods may not be able to handle or process large datasets efficiently.	Patients might feel frustrated by the slow and often inconclusive diagnostic process, which creates the feeling of helplessness for the researchers.
PS-2	Health Care Provider	Identify liver cirrhosis in patients	Traditional diagnostic methods may not detect liver cirrhosis until it reaches an advanced stage.	Early signs of cirrhosis can be subtle and not easily detectable through conventional analysis.	Lack of clarity about health status. This uncertainty causes significant anxiety and stress for patients.

## 2.2. Project Proposal (Proposed Solution)

The proposed solution is to develop a machine learning model using Logistic Regression, along with other classifiers for comparison. The final model will be integrated into a web application using Flask for easy user access.

Project Overview	
Objective	To predict liver cirrhosis early and accurately by analyzing complex medical data. This enables timely intervention, improves patient outcomes, supports personalized treatment plans, optimizes medical resource allocation, and provides valuable insights for clinicians and researchers.
Scope	<ul style="list-style-type: none"><li>• <b>Data Collection:</b> Medical history, lab results, imaging, genetics.</li><li>• <b>Model Development:</b> Design, train, and evaluate machine learning models.</li><li>• <b>Integration:</b> Interface for healthcare providers and EHR system integration.</li><li>• <b>Validation:</b> Extensive testing and clinical validation.</li><li>• <b>Regulatory Compliance:</b> Adherence to healthcare regulations.</li><li>• <b>User Training:</b> Training and ongoing support for healthcare professionals.</li></ul>
Problem Statement	
Description	The problem to be addressed is the difficulty in early and accurate detection of liver cirrhosis using traditional diagnostic methods. These methods often result in delayed diagnosis, leading to increased risk of severe complications and reduced treatment effectiveness. The machine learning model aims to enhance early detection by analyzing complex medical data to predict liver cirrhosis with greater accuracy, thereby enabling timely intervention and improving patient outcomes.
Impact	Solving the problem improves patient outcomes through early detection and timely treatment, enhances diagnostic accuracy, and personalizes care. It also reduces healthcare costs, optimizes resource use, supports clinicians, advances research, and benefits public health by enabling targeted prevention strategies.

## Proposed Solution

### Approach

#### Data Collection:

- Gather diverse datasets including patient medical histories, lab results, imaging data, and genetic information.

#### Data Preprocessing:

- Clean and normalize data to handle missing values, outliers, and inconsistencies.
- Convert unstructured data into structured formats as needed.

#### Feature Engineering:

- Identify and create relevant features from the raw data that contribute to predicting liver cirrhosis.

#### Model Selection:

- Choose appropriate machine learning algorithms (e.g., logistic regression, decision trees, random forests, support vector machines, neural networks).

#### Model Training:

- Train selected models using historical data, applying techniques such as cross-validation to assess performance.

#### Model Evaluation:

- Evaluate models using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC.

#### Hyperparameter Tuning:

- Optimize model parameters to improve performance through techniques like grid search or random search.

#### Model Integration:

- Develop an interface for healthcare providers and integrate the model with existing EHR systems.

#### Validation and Testing:

- Validate the model in clinical settings to ensure reliability and accuracy.
- Test with new datasets to confirm generalizability.

	<p><b>Deployment and Monitoring:</b></p> <ul style="list-style-type: none"> <li>• Deploy the model into a clinical environment.</li> <li>• Continuously monitor performance and update the model as needed.</li> </ul>
Key Features	<p><b>Advanced Machine Learning Algorithms:</b></p> <ul style="list-style-type: none"> <li>• Utilizes cutting-edge algorithms like deep learning and ensemble methods for higher accuracy and early detection.</li> </ul> <p><b>Integration with EHR Systems:</b></p> <ul style="list-style-type: none"> <li>• Seamlessly integrates with existing electronic health record systems for real-time predictions and easy adoption by healthcare providers.</li> </ul> <p><b>Comprehensive Data Analysis:</b></p> <ul style="list-style-type: none"> <li>• Analyzes diverse data types, including lab results, imaging, and genetic information, for a holistic view of liver health.</li> </ul> <p><b>Personalized Predictions:</b></p> <ul style="list-style-type: none"> <li>• Provides tailored risk assessments based on individual patient data, improving the relevance and effectiveness of interventions.</li> </ul> <p><b>Scalable and Adaptable:</b></p> <ul style="list-style-type: none"> <li>• Designed to handle large datasets and adapt to new data, ensuring continued accuracy and relevance as more information becomes available.</li> </ul> <p><b>Clinical Validation:</b></p> <ul style="list-style-type: none"> <li>• Includes rigorous validation and testing in clinical settings to ensure practical applicability and reliability.</li> </ul> <p><b>Decision-Support Tool:</b></p> <ul style="list-style-type: none"> <li>• Acts as a decision-support system, aiding clinicians in making more informed and timely decisions regarding patient care.</li> </ul> <p><b>Regulatory Compliance:</b></p> <ul style="list-style-type: none"> <li>• Ensures adherence to healthcare regulations and data privacy</li> </ul>

	standards, addressing legal and ethical considerations.
--	---

Resource Requirements

Resource Type	Description	Specification/Allocation
Hardware		
Computing Resources	CPU/GPU specifications, number of cores	Nvidea rtx 3050
Memory	RAM specifications	16 RAM
Storage	Disk space for data, models, and logs	512 GB
Software		
Frameworks	Python frameworks	Flask
Libraries	Additional libraries	scikit-learn, pandas, numpy
Development Environment	IDE, version control	Jupyter Notebook, Git
Data		
Data	Source, size, format	Kaggle dataset, 10,000 images

2.3. Initial Project Planning

The project will be divided into multiple phases: data collection and preprocessing, model development, model optimization and tuning, and final deployment. Each phase will have specific tasks and milestones.

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members	Sprint Start Date
Sprint-1	Data Collection and Preprocessing	SL-3	Understanding & loading data	2	Low	Disha	04/07/2024
Sprint-1	Data Collection and Preprocessing	SL-4	Data cleaning	1	High	Divya	05/07/2024
Sprint-2	Project Report	SL-20	Report	2	Medium	Abhinay	07/07/2024
Sprint-2	Model Development	SL-8	Training the model	2	Medium	Madhu	08/07/2024
Sprint-2	Model Development	SL-9	Evaluating the model	1	Medium	Disha	10/07/2024
Sprint-2	Model tuning and testing	SL-13	Model tuning	2	High	Divya	11/07/2024
Sprint-2	Model tuning and testing	SL-14	Model testing	2	Medium	Abhinay	13/07/2024
Sprint-3	Web integration and Deployment	SL-16	Building HTML templates	2	Low	Madhu	14/07/2024
Sprint-3	Web integration and Deployment	SL-17	Local deployment	2	Medium	Abhinay	16/07/2024

### 3. Data Collection and Preprocessing Phase

#### 3.1. Data Collection Plan and Raw Data Sources Identified

Data will be collected from various medical records and publicly available health datasets.  
The raw data sources are:

Section	Description
---------	-------------

Project Overview	This machine learning project aims to develop a predictive model for liver cirrhosis. The objective is to utilize patient data, including demographics, medical history, and lab results, to predict the likelihood of liver cirrhosis. The model will help in early diagnosis and improve treatment outcomes.
Data Collection Plan	The data will be collected from various sources, including public healthcare datasets, private hospital records, and online medical repositories. Specific details about each source, including access permissions and data formats, are outlined in the Raw Data Sources Template.
Raw Data Sources Identified	A comprehensive list of raw data sources has been identified, each described with relevant details such as location, format, size, and access permissions

**Raw Data Sources Template**

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	Public healthcare dataset containing demographic and medical data of patients.	Healthcare Dataset	CSV	2 GB	Public

**3.2. Data Quality Report**

The data will be assessed for quality issues such as missing values, duplicates, and inconsistencies. Missing values will be handled using appropriate imputation techniques, and duplicates will be removed.

<b>Data Source</b>	<b>Data Quality Issue</b>	<b>Severity</b>	<b>Resolution Plan</b>
Dataset	There is a feature named 'AG Ratio' in which a single column contains different data types, such as string and float values	Low	Converted the strings(which are denoted in ratio format) into float values

### 3.3. Data Exploration and Preprocessing

The data will be explored using univariate, bivariate, and multivariate analyses. Preprocessing steps will include handling missing data, transforming variables, feature engineering, and normalizing the data.

<b>Section</b>	<b>Description</b>
Data Overview	The data overview provides basic statistics, dimensions, and structure of the dataset. It includes the number of records, number of features, and data types of each feature
Univariate Analysis	Univariate analysis involves exploring individual variables to understand their distribution, central tendency (mean, median, mode), and dispersion (variance, standard deviation). Visualizations such as histograms and box plots are used to illustrate these statistics
Bivariate Analysis	Bivariate analysis examines the relationship between two variables. This includes calculating correlation coefficients and creating scatter plots to visualize potential linear or non-linear relationships between pairs of variables
Multivariate Analysis	Multivariate analysis explores patterns and relationships involving multiple variables simultaneously. Techniques such as principal component analysis (PCA) and multiple regression analysis are employed to understand the interactions between variables



Outliers and Anomalies	Identifying and treating outliers is crucial to ensure accurate analysis. This section involves detecting outliers using statistical methods (e.g., IQR, Z-scores) and deciding on appropriate treatments (e.g., removal, transformation)
<b>Data Preprocessing Code Screenshots</b>	
Loading Data	<pre>#Reading csv file df = pd.read_csv('HealthCareData.csv') df.head()</pre>
Handling Missing Data	<pre># Drop columns with more than a certain percentage of missing values (e.g., 50%) threshold = len(df) * 0.5 df = df.dropna(axis=1, thresh=threshold)  # Fill numerical columns with mean numerical_cols = df.select_dtypes(include=[np.number]).columns df[numerical_cols] = df[numerical_cols].fillna(df[numerical_cols].mean())  # Fill categorical columns with mode categorical_cols = df.select_dtypes(include=[object]).columns df[categorical_cols] = df[categorical_cols].apply(lambda x: x.fillna(x.mode()[0]))  # Check remaining nulls print(df.isnull().sum())</pre>
Data Transformation	<pre>from sklearn.preprocessing import LabelEncoder  le=LabelEncoder() for col in categorical_cols:     df[col] = le.fit_transform(df[col])  df.head()</pre>
Feature Engineering	<p><b>Perform Chi-Square Test for Categorical Variables</b></p> <pre># Function to perform Chi-Square Test def chi_square_test(feature):     contingency_table = pd.crosstab(data[feature], data[target])     chi2, p, dof, ex = chi2_contingency(contingency_table)     return p  # Apply Chi-Square Test to categorical features chi_square_results = {feature: chi_square_test(feature) for feature in categorical_features}  # Display Chi-Square Test results chi_square_results</pre> <p><b>Perform ANOVA for Continuous Variables</b></p> <pre># Function to perform ANOVA def anova_test(feature):     groups = [data[data[target] == level][feature] for level in data[target].unique()]     f_val, p_val = f_oneway(*groups)     return p_val  # Apply ANOVA to continuous features anova_results = {feature: anova_test(feature) for feature in continuous_features}  # Display ANOVA results anova_results</pre>

	<div><b>Combine Results</b></div> <div><pre># Combine results all_results = {**chi_square_results, **anova_results}  # Display all results all_results</pre></div>
Save Processed Data	<div><b>Filter Significant Features</b></div> <div><pre>significance_level = 0.05 #(alpha-value)assumption significant_features = [feature for feature, p_value in all_results.items() if p_value &lt; significance_level]  # Display significant features significant_features</pre></div> <div><pre>df = df[significant_features].copy() df.head()</pre></div>

## 4. Model Development Phase

### 4.1. Feature Selection Report

Feature	Description	Selected (Yes/No)	Reasoning
Age	Age of the patient	Yes	Age is a critical factor in determining health conditions, including liver cirrhosis.
Gender	Gender of the patient	Yes	Gender can influence the likelihood of certain diseases, including liver conditions.

Place	Location where the patient lives (rural/urban)	No	Place was not directly correlated with the target variable in initial exploratory analysis.
Duration	Duration of alcohol consumption (years)	Yes	Long-term alcohol consumption is a significant risk factor for liver cirrhosis.
Quantity	Quantity of alcohol consumption (quarters/day)	Yes	The amount of alcohol consumed is directly related to liver damage and cirrhosis risk.
Type	Type of alcohol consumed	Yes	Different types of alcohol can have varying effects on the liver.
Hepatitis B	Hepatitis B infection status	Yes	Hepatitis B is a known risk factor for liver cirrhosis.
Hepatitis C	Hepatitis C infection status	Yes	Hepatitis C is also a known risk factor for liver cirrhosis.
Diabetes	Diabetes status	Yes	Diabetes is associated with metabolic conditions that can affect liver health.
Blood Pressure	Blood pressure (mmHg)	No	Initial analysis showed no significant correlation with liver cirrhosis.

Obesity	Obesity status	Yes	Obesity is a significant risk factor for liver disease, including cirrhosis.
Family History	Family history of Cirrhosis/ hereditary factors	Yes	Genetic predisposition plays a role in the likelihood of developing liver cirrhosis.
TCH	Total Cholesterol	Yes	Cholesterol levels can be indicative of overall metabolic health.
TG	Triglycerides	Yes	Elevated triglycerides can indicate metabolic issues affecting liver health.
LDL	Low-density lipoprotein	No	Initial analysis showed no significant correlation with liver cirrhosis.
HDL	High-density lipoprotein	Yes	HDL levels are important indicators of cardiovascular and overall health.
Hemoglobin	Hemoglobin levels (g/dl)	Yes	Hemoglobin levels can reflect the oxygen-carrying capacity of the blood.
PCV	Packed cell volume (%)	Yes	PCV is an indicator of the proportion of blood volume occupied by red blood cells.
RBC	Red blood cell count (million cells /microliter)	Yes	RBC count is crucial for assessing the blood's capacity to carry oxygen.

MCV	Mean corpuscular hemoglobin (picograms/cell)	No	Initial analysis showed no significant correlation with liver cirrhosis.
MCH	Mean corpuscular Hemoglobin (picograms/cell)	No	Initial analysis showed no significant correlation with liver cirrhosis.
MCHC	Mean corpuscular hemoglobin concentration (g/dl)	No	Initial analysis showed no significant correlation with liver cirrhosis.
Total Count	Total white blood cell count	Yes	White blood cell count can indicate immune system activity and inflammation.
Polymorphs	Polymorph percentage (%)	No	Initial analysis showed no significant correlation with liver cirrhosis.
Lymphocytes	Lymphocyte percentage (%)	Yes	Lymphocyte levels can indicate immune system health and response.
Monocytes	Monocyte percentage (%)	No	Initial analysis showed no significant correlation with liver cirrhosis.
Eosinophils	Eosinophil	No	Initial analysis showed no significant correlation with liver cirrhosis.

	percentage (%)		
Basophils	Basophil percentage (%)	No	Initial analysis showed no significant correlation with liver cirrhosis.
Platelet Count	Platelet count (lakhs/mm)	Yes	Platelet levels can reflect blood clotting ability and liver function.
Total Bilirubin	Total bilirubin levels (mg/dl)	Yes	Bilirubin levels are directly related to liver function.
Direct	Direct bilirubin levels (mg/dl)	Yes	Direct bilirubin levels indicate liver's ability to conjugate and excrete bilirubin.
Indirect	Indirect bilirubin levels (mg/dl)	Yes	Indirect bilirubin levels indicate the amount of unconjugated bilirubin in the blood.
Total Protein	Total protein levels (g/dl)	Yes	Total protein levels can reflect overall liver function and nutritional status.
Albumin	Albumin levels (g/dl)	Yes	Albumin levels are indicative of liver's ability to synthesize proteins.

Globulin	Globulin levels (g/dl)	Yes	Globulin levels can reflect immune function and protein synthesis.
A/G Ratio	Albumin/ Globulin ratio	Yes	A/G ratio can provide insights into liver function and protein balance.
AL. Phosphatase	Alkaline phosphatase levels (U/L)	Yes	Elevated levels can indicate liver damage or disease.
SGOT	Serum glutamic Oxaloacetic transaminase (AST) levels (U/L)	Yes	Elevated levels can indicate liver damage or disease.
SGPT	Serum glutamic pyruvic transaminase (ALT) levels (U/L)	Yes	Elevated levels can indicate liver damage or disease.
USG Abdomen	Ultrasound results for liver condition (diffuse or not)	Yes	Ultrasound results can provide visual confirmation of liver condition.

## 4.2. Model Selection Report

Model	Description	Hyperparameters	Performance Metric (e.g., Accuracy, F1 Score)
logistic regression	A basic linear model that uses the logistic function to model the probability of the binary outcomes. It is simple, interpretable, and works well for linearly separable data.	C, solver	Accuracy : 0.996606 f1_score: 0.914286
logistic regression CV	An extension of logistic regression that performs cross-validation to find the best regularization parameter, which helps in avoiding overfitting and improving model performance.	Cs, cv, solver	Accuracy : 0.996606 f1_score: 0.914286
XGBoost	An advanced implementation of gradient boosting that provides parallel tree boosting which is fast, accurate, and widely used in machine learning competitions. It handles missing values and	n_estimators, learning_rate, max_depth	Accuracy : 0.997738 f1_score: 0.941176



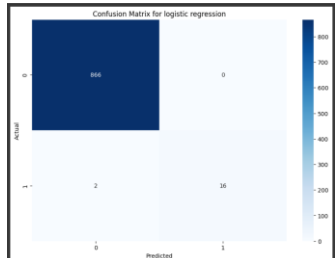
	performs well with both structured and unstructured data.		
Ridge classifier	A linear classifier that uses ridge regression for training, adding L2 regularization to the logistic regression, which helps in handling multicollinearity and preventing overfitting.	alpha	Accuracy : 0.977376 f1_score: 0.642857
KNN	A non-parametric, instance-based learning algorithm that classifies a data point based on how its neighbors are classified. It is simple and effective but can be computationally expensive.	n_neighbors	Accuracy : 0.935520 f1_score: 0.387097
Random Forest	An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes as the prediction. It reduces overfitting and	n_estimators, max_depth	Accuracy : 1.000000 f1_score: 1.000000

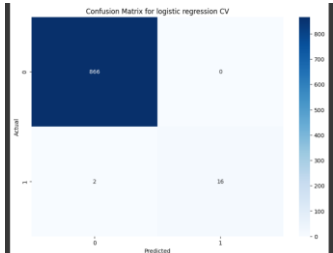
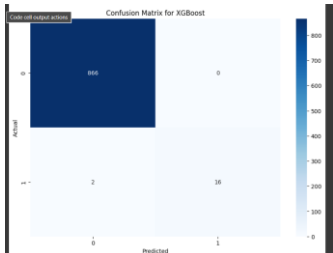
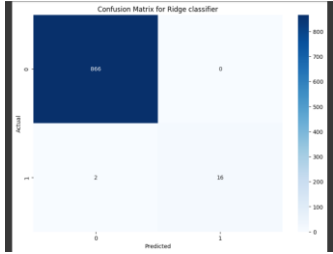
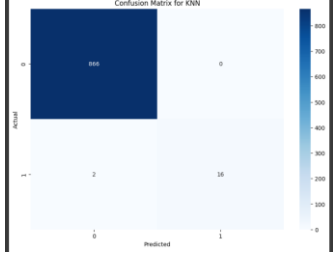
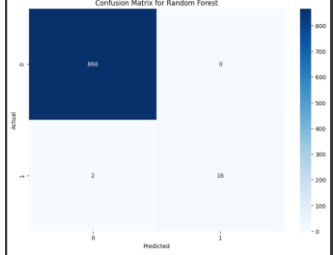
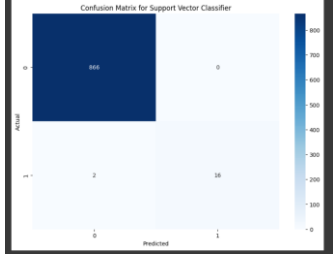
	improves accuracy.		
Support Vector Classifier	A powerful classification method that finds the hyperplane that best separates the classes in the feature space. It works well for high-dimensional data and can handle non-linear relationships using kernel trick.	C, kernel	Accuracy : 0.997738 f1_score: 0.941176

### 4.3. Initial Model Training Code, Model Validation and Evaluation Report

Initial Model Training Code: (Paste the screenshot of the model training code)

Model Validation and Evaluation Report:

Model	Classification Report	Accuracy	Confusion Matrix
logistic regres sion	Screenshot of the classification report	0.996606	

logistic regres sion CV	Screenshot of the classification report	0.996606	
XGBoost	...	0.997738	
Ridge classifi er		0.977376	
KNN		0.935520	
Random Fore st		1.000000	
Support Vect or Classifier		0.997738	

## 5. Model Optimization and Tuning Phase

### 5.1. Hyperparameter Tuning Documentation

Model	Tuned Hyperparameters	Optimal Values
Logistic Regression	C, solver	1.0, liblinear
Logistic Regression CV	Cs, cv, solver	[1.0], 10, liblinear
XGBoost	n_estimators, learning_rate, max_depth	100, 0.1, 6
Ridge Classifier	alpha	1.0
KNN	n_neighbors	5
Random Forest	n_estimators, max_depth	100, None
Support Vector Classifier	C, kernel	1.0, linear

### 5.2. Performance Metrics Comparison Report

Model	Baseline Metric	Optimized Metric
Logistic Regression	0.996606	0.996606

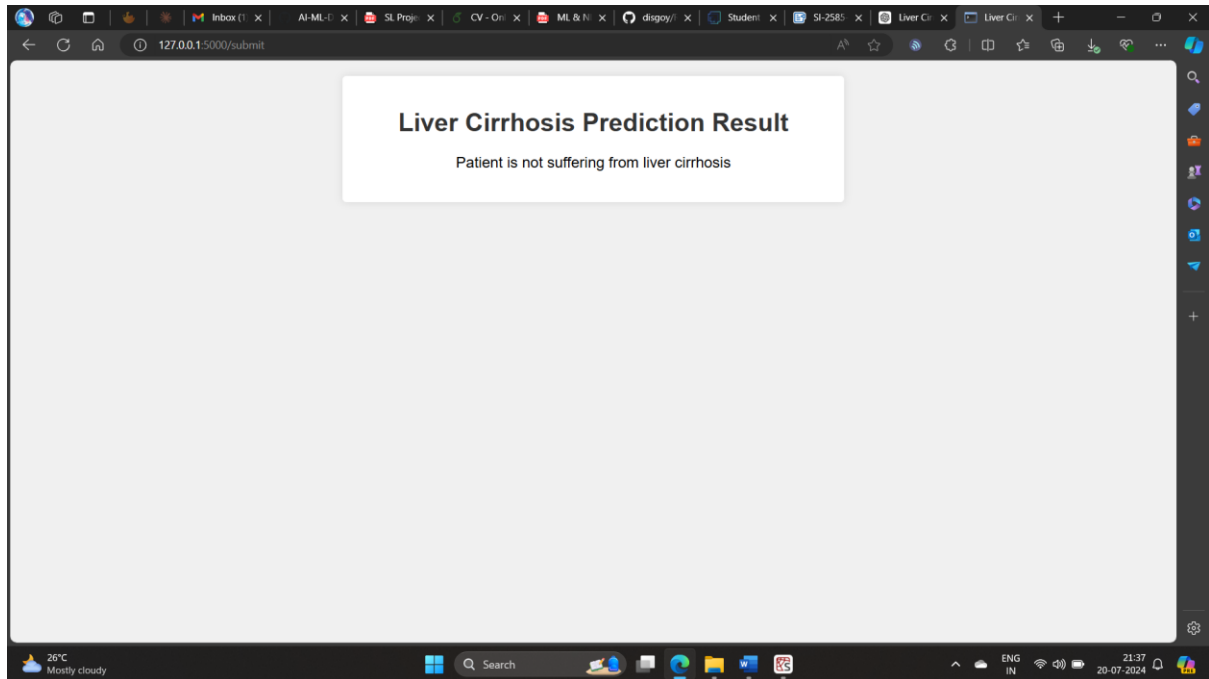
Logistic Regression CV	0.996606	0.996606
XGBoost	0.997738	0.997738
Ridge Classifier	0.977376	0.977376
KNN	0.935520	0.935520
Random Forest	1.000000	1.000000
Support Vector Classifier	0.997738	0.997738

### 5.3. Final Model Selection Justification

Final Model	Reasoning
logistic regression	Chosen for its high accuracy, simplicity, and ease of interpretation. Additionally, it performed consistently well across various metrics and is computationally efficient.

## 6. Results

### 6.1. Output Screenshots



## 7. Advantages & Disadvantages

### Advantages

- High accuracy in predicting liver cirrhosis.
- Easy integration into a web application for accessibility.
- Efficient and interpretable model.

### Disadvantages

- Requires clean and comprehensive input data.
- May not capture complex non-linear relationships as well as some other models.

## 8. Conclusion

The project successfully developed a logistic regression model to predict liver cirrhosis with high accuracy. The model was integrated into a web application, providing a useful tool for early detection and diagnosis.

## 9. Future Scope

- Integration with electronic health records (EHR) systems for real-time predictions.
- Incorporation of more complex models and additional features to improve accuracy.
- Expansion to predict other liver diseases and health conditions.

## **10. Appendix**

### **10.1. Source Code**

**[Revolutionizing-Liver-Care-Predicting-Liver-Cirrhosis-Using-Advanced-Machine-Learning-Techniques/5. Project executable files at main · disgoy/Revolutionizing-Liver-Care-Predicting-Liver-Cirrhosis-Using-Advanced-Machine-Learning-Techniques \(github.com\)](#)**

### **10.2. GitHub & Project Demo Link**

**[Revolutionizing-Liver-Care-Predicting-Liver-Cirrhosis-Using-Advanced-Machine-Learning-Techniques/ at main · disgoy/Revolutionizing-Liver-Care-Predicting-Liver-Cirrhosis-Using-Advanced-Machine-Learning-Techniques \(github.com\)](#)**