**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.

**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.

**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.

## 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| 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.

|  |  |
| --- | --- |
| **Section** | **Description** |
| 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) |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |  |
| Handling Missing Data |  |
| Data Transformation |  |
| Feature Engineering |  |
| Save Processed Data |  |

## 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  percentage (%) |  |  | | --- | |  | |  |  | | --- | |  | |  |  | | --- | |  | | No | |  | | --- | | Initial analysis showed no significant  correlation with liver cirrhosis. |  |  | | --- | |  | |
| Basophils | |  |  |  | | --- | --- | --- | | |  | | --- | | Basophil  percentage (%) |  |  | | --- | |  | |  |  | | --- | |  | | No | |  |  |  |  |  | | --- | --- | --- | --- | --- | | |  |  |  | | --- | --- | --- | | |  | | --- | | Initial analysis showed no significan  t 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. | Globulin levels can reflect immune function and protein synthesis. |  |  | | --- | |  | |
| |  | | --- | | 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 performs well with both structured and unstructured data. | n\_estimators, learning\_rate, max\_depth | Accuracy : 0.997738  f1\_score: 0.941176 |
| 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 improves accuracy. | n\_estimators, max\_depth | Accuracy : 1.000000  f1\_score: 1.000000 |
| 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 regression | Screenshot of the classification report | 0.996606 |  |
| logistic regression CV | Screenshot of the classification report | 0.996606 |  |
| XGBoost | … | 0.997738 |  |
| Ridge classifier |  | 0.977376 |  |
| KNN |  | 0.935520 |  |
| Random Forest |  | 1.000000 |  |
| Support Vector 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

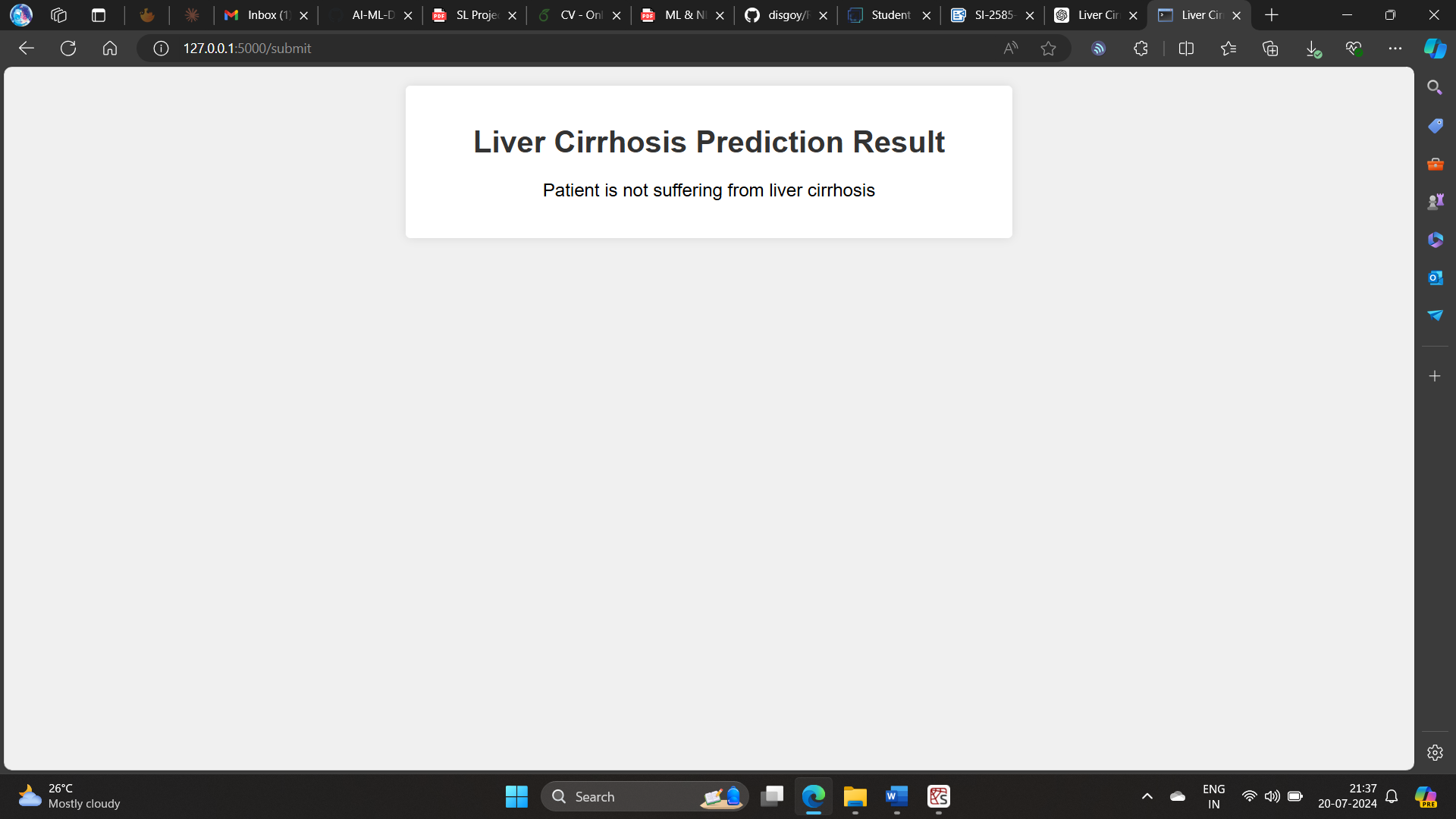
|  |  |  |
| --- | --- | --- |
| **Model** | **Baseline Metric** | **Optimized Metric** |
| Logistic Regression | 0.996606 | 0.996606 |
| Logistic Regression CV | 0.996606 | 0.996606 |
| XGBoost | 0.997738 | 0.997738 |
| Ridge Classifier | 0.977376 | 0.977376 |
| KNN | 0.935520 | 0.935520 |
| Random 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)](https://github.com/disgoy/Revolutionizing-Liver-Care-Predicting-Liver-Cirrhosis-Using-Advanced-Machine-Learning-Techniques/tree/main/5.%20Project%20executable%20files)

### 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)](https://github.com/disgoy/Revolutionizing-Liver-Care-Predicting-Liver-Cirrhosis-Using-Advanced-Machine-Learning-Techniques/tree/main)