Revolutionizing Liver Care: Predicting Liver Cirrhosis using Advanced Machine Learning Techniques

Team Name: APEX
Team Members: Kuriti Uma, K Hima Bindu, K Abhiram, Kamalsahebgari Kuljum

1 Phase 1: Brainstorming & Ideation

1.1 Objective

To develop a robust, intelligent system that can accurately predict the presence of liver cirrhosis in patients using machine learning techniques applied to structured medical data.

1.2 Problem Statement

Liver cirrhosis is a chronic, irreversible condition marked by the progressive destruction and scarring of liver tissue. Often asymptomatic in the early stages, cirrhosis is typically diagnosed only when substantial damage has occurred, limiting treatment options. This delay contributes to a high global mortality rate. The growing availability of clinical data creates an opportunity to leverage machine learning for earlier, more accurate diagnosis. Our project addresses this challenge by developing a predictive model using real patient data to classify individuals as either cirrhotic or non-cirrhotic. Such a system could enable timely medical interventions, improving patient outcomes and optimizing healthcare resource use.

1.3 Proposed Solution

We propose a supervised machine learning approach that utilizes classification algorithms trained on medical datasets. By analyzing various biochemical and demographic features, the model aims to detect patterns associated with liver cirrhosis and accurately predict the condition.

1.4 Target Users

medical professionals, hospital administrators, public health researchers, and medical technology companies who aim to improve liver disease diagnosis and management

1.5 Expected Outcome

A lightweight, interpretable, and scalable liver cirrhosis prediction model that provides real–time classification of patients and can be integrated into clinical decision support systems.

2 Phase 2: Requirement Analysis

2.1 Technical Requirements

- Python programming environment (Jupyter Notebook)
- Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- Data source: Structured CSV/XLSX dataset with liver profiles

2.2 Functional Requirements

The model must ingest patient records, handle missing values, engineer features, train on labeled data, and provide binary classification (Cirrhosis: Yes/No) with associated performance metrics.

2.3 Constraints & Challenges

Challenges included small sample size, significant class imbalance, and noisy real-world data. Mitigation strategies such as SMOTE, robust preprocessing, and careful model validation were applied.

3 Phase 3: Project Design

3.1 System Architecture

- **Step 1: Data Collection and Cleaning**
- **Step 2: Exploratory Data Analysis (EDA)**
- **Step 3: Feature Engineering and Selection**
- **Step 4: Model Training using classifiers**
- **Step 5: Model Evaluation (metrics and plots)**
- **Step 6: Output Risk Classification**

3.2 User Flow

User inputs patient data → Data preprocessed → Model prediction → Output displayed with risk category

3.3 UI/UX Considerations

The model is notebook-based for now but is designed with modular code for easy transition to a graphical or web-based interface in future deployments.

4 Phase 4: Project Planning (Agile)

4.1 Sprint Planning

- Sprint 1: Dataset collection and preprocessing
- Sprint 2: Model implementation and EDA
- Sprint 3: Evaluation, visualization, tuning
- Sprint 4: Report generation and video demo

4.2 Task Allocation

- **Kuriti Uma** – Data loading, initial analysis, demo video creation
- **K Hima Bindu** – Preprocessing pipeline and correlation mapping
- **K Abhiram** – Model development and accuracy tuning
- **Kamalsahebgari Kuljum** – Evaluation reporting, metrics interpretation, documentation

5 Phase 5: Project Development

5.1 Steps Followed

- **Loaded liver dataset using Pandas**
- **Preprocessed data: handled missing values, encoded labels**
- **Conducted EDA using histograms, boxplots, and heatmaps**
- **Trained ML models: Logistic Regression, KNN, Decision Trees, etc.**
- **Evaluated with accuracy, precision, recall, F1-score**

5.2 Technologies Used

Python, Scikit-learn, Matplotlib, Seaborn, Pandas, NumPy

5.3 Development Process

Applied various models including KNN, Logistic Regression, Random Forest, etc. Used EDA, feature selection, and cross-validation.

5.4 Challenges & Fixes

Handled missing data, class imbalance (using SMOTE), and overfitting

Model Performance Comparison

Performance Table: Below is the comparative performance of various machine learning algorithms applied to our dataset.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	35.79%	0.00	0.00	0.00
Random Forest	35.79%	0.00	0.00	0.00
Logistic Regression	81.58%	91.80	79.43	86.49
Ridge Classifier	84.21%	93.44	83.82	88.37
SVC	35.79%	0.00	0.00	0.00
Logistic Regression	79.47%	91.80	79.43	85.58
K-Nearest Neighbors	86.32%	94.26	85.82	89.84
XGBoost	35.79%	3.28	50.00	6.15

5.5 Charts and Visualizations

Visual summaries such as correlation heatmaps, boxplots for outliers, and confusion matrices were generated using Seaborn and Matplotlib .

- Feature correlation heatmap
- Outlier distribution boxplots
- Confusion matrix visualizations

6 Phase 6: Functional & Performance Testing

6.1 Testing Process

Split the dataset into training and testing sets. Evaluated with confusion matrices and classification reports.

6.2 Validation Results

Performance varied by algorithm. KNN and Ridge Classifier achieved the best F1-scores. Metrics were derived using Scikit-learn's evaluation functions.

6.3 Deployment

Not yet deployed; the notebook version is modular and ready for integration into a web or hospital backend interface.

7 Final Submission

Report – This document presents full workflow and findings.
GitHub Repository –`https://github.com/Uma-Kuriti/Liver-Cirrhosis-Prediction`
Demo Video – `https://drive.google.com/file/d/1910B2AiSoNyb9nOiQysh5CTJ5YWecuxP/view?usp=drivesdk`