**Final Project Report**

|  |  |
| --- | --- |
| Date | 24 June 2025 |
| Team ID | SWTID1749708868 |
| Project Title | Revolutionizing Liver Care : Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques |
| Maximum Marks | 2 Marks |
| Team Members | Dhawan Singh Saini  Dhruv Diguvapalli  Abhyuday Sajlani  Prathamshree Kumar |

**1. Introduction**

**1.1. Project Overview**

This project proposes the development of a machine learning model to support the early identification of liver cirrhosis risk based on patient data, including clinical history, lab results, and lifestyle factors. The system is designed to assist healthcare professionals in making more informed, timely decisions by providing predictive insights into disease progression.

**1.2. Objectives**

The primary objective is to build a machine learning model that predicts the risk of liver cirrhosis based on patient health data. The project's scope includes data preprocessing, model training, evaluation, and deployment as a web-based tool for clinical use. The application aims to assist in screening, monitoring disease progression, and optimizing treatment strategies for more proactive and personalized liver care.

**2. Project Initialization and Planning Phase**

**2.1. Define Problem Statement**

Patients and healthcare providers often struggle with the early detection of liver cirrhosis due to vague symptoms and limited diagnostic access, leading to treatment delays and worsening outcomes. This project addresses the critical need for a predictive machine learning model that leverages patient data (medical history, lab results) to assess cirrhosis risk early, enabling timely intervention, personalized care planning, and better resource allocation in healthcare settings.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/1.%20Project%20Initialization%20and%20Planning%20Phase/Problem%20Statement.pdf)

**2.2. Project Proposal (Proposed Solution)**

The proposed solution involves developing a machine learning model that analyzes patient records (lab values, history, lifestyle inputs) to identify cirrhosis risk through regression algorithms. This system enables earlier diagnosis, supports personalized treatment planning, and helps optimize healthcare resources. The project follows a stepwise pipeline including data cleaning, exploratory analysis, model selection, optimization, and Flask web application deployment. Key features include an interactive web interface, real-time prediction output, and regression-based modeling.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/1.%20Project%20Initialization%20and%20Planning%20Phase/Project%20Proposal.pdf)

**2.3. Initial Project Planning**

The project follows a structured sprint schedule:

* **Sprint 1: Data Collection and Preprocessing** (June 17, 2025 - June 19, 2025): Load, explore, handle missing values and outliers, visualize, and analyze data.
* **Sprint 2: Model Development** (June 19, 2025 - June 21, 2025): Train ML regression models, evaluate performance, and tune hyperparameters.
* **Sprint 3: Model Tuning and Testing** (Planned): Further hyperparameter tuning and performance evaluation, justifying final model selection.
* **Sprint 4: Web Integration and Deployment** (Planned): Develop Flask web application, integrate model, and deploy/test.
* **Sprint 5: Project Report** (Planned): Compile the project report.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/1.%20Project%20Initialization%20and%20Planning%20Phase/Project%20Planning.pdf)

**3. Data Collection and Preprocessing Phase**

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

The dataset was sourced from Kaggle, titled "liver-cirrhosis-prediction" by bhavanipriya222. It includes patient medical records with liver health indicators in Excel format, approximately 214 KB, and is publicly accessible. The data collection plan involved acquiring this dataset and validating its credibility through metadata review, assessing completeness, and performing initial cleaning.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/2.%20Data%20Collection%20and%20Preprocessing%20Phase/Raw%20Data%20Sources%20And%20Data%20Quality%20Report.pdf)

**3.2. Data Quality Report**

The primary data quality issue was the presence of null values (High severity). This was resolved by substituting null values with the median or mode, depending on the variable type.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/2.%20Data%20Collection%20and%20Preprocessing%20Phase/Data%20Quality%20Report.pdf)

**3.3. Data Exploration and Preprocessing**

The dataset has a dimension of 950 rows and 13 columns. Preprocessing involved statistical analysis, visualization, and cleaning using Python.

* **Handling Missing Data:** Columns with more than 50% missing values were dropped. Categorical missing values were filled with "Unknown," and numerical missing values with the median. KNN imputation was also used.
* **Feature Mapping:**
  + Gender: Male (0), Female (1), Transgender (2)
  + Place (location where the patient lives): Rural (0), Urban (1), Unknown (2)
  + Type of alcohol consumed: Branded (0), Both (1), Country (2)
  + Test results (Hepatitis B, Hepatitis C, Diabetes, Obesity): Negative (0), Positive (1)

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/2.%20Data%20Collection%20and%20Preprocessing%20Phase/Data%20Exploration%20and%20Preprocessing.pdf)

**4. Model Development Phase**

**4.1. Feature Selection Report**

Features were selected based on clinical relevance, correlation with the target variable, and impact on model performance.

* **Retained Features:** Age, Gender, Alcohol History, Hepatitis B/C Infection, Jaundice, Ascites, Bilirubin levels, Liver Enzymes (ALP, SGOT/AST, SGPT/ALT), Albumin, Total Protein, A/G Ratio, Total Leucocyte Count, Platelet Count, Serum Bilirubin (Total, Direct, Indirect). These are strongly associated with liver function.
* **Excluded Features:** "S.NO" (record index) and "Place (location where the patient lives)" (geographical information) were excluded to reduce noise and overfitting, as they were not directly relevant identifiers.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/3.%20Model%20Development%20Phase/Feature%20Selection%20Report.pdf)

**4.2. Model Selection Report**

The following models were considered:

* **Logistic Regression:** Simple, interpretable, suitable for linearly separable data. Hyperparameters: max\_iter=1000, random\_state=42. Performance: Accuracy: 0.909, F1 Score: 0.896.
* **SVM (RBF):** Effective in high-dimensional spaces, captures non-linear relationships. Hyperparameters: gamma='auto', probability=True, random\_state=42. Performance: Accuracy: 0.941, F1 Score: 0.913.
* **K-Nearest Neighbors:** Instance-based learner, works well with clear local patterns. Hyperparameters: n\_neighbors=5. Performance: Accuracy: 0.941, F1 Score: 0.913.
* **Decision Tree:** Handles non-linear data and interactions. Hyperparameters: random\_state=42. Performance: Accuracy: 0.936, F1 Score: 0.910.
* **Random Forest:** Ensemble method, reduces overfitting, handles complex data. Hyperparameters: random\_state=42. Performance: Accuracy: 0.947, F1 Score: 0.923.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/3.%20Model%20Development%20Phase/Initial%20Model%20Training%20Code%2C%20Model%20Validation%20and%20Evaluation.pdf)

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

Multiple classification models were trained. Feature scaling was applied using Standard Scaler. Models were evaluated using accuracy, weighted F1-score, confusion matrix, and classification report to compare performance and ensure robust validation.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/3.%20Model%20Development%20Phase/Model%20Selection%20Report.pdf)

**5. Model Optimization and Tuning Phase**

**5.1. Hyperparameter Tuning Documentation**

This section would typically detail the specific hyperparameter tuning methods used (e.g., GridSearchCV, RandomizedSearchCV) and the best parameters found for each model.

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/blob/main/4.%20Model%20Optimization%20and%20Tuning%20Phase/Model%20Optimization%20and%20Tuning%20Phase.pdf)

**5.2. Performance Metrics Comparison Report**

| Model | Baseline Accuracy | Optimized Accuracy |
| --- | --- | --- |
| Logistic Regression | 0.91 | 0.95 |
| SVM (RBF) | 0.94 | 0.95 |
| K-Nearest Neighbors | 0.94 | 0.95 |
| Decision Tree | 0.94 | 0.95 |
| Random Forest | 0.95 | 0.95 |

**5.3. Final Model Selection Justification**

The optimization and tuning phase improved the performance of most models, with Logistic Regression showing a notable increase in accuracy. The other models maintained or slightly improved their high accuracy after optimization. Based on the

provided performance metrics, all optimized models achieved a high accuracy of 0.95, indicating strong predictive capabilities for liver cirrhosis.

**6. Results**

**6.1. Output Screenshots**

****6.2. Advantages & Disadvantages**

* **Advantages:**
  + Early detection of liver cirrhosis, enabling timely intervention.
  + Supports personalized care planning.
  + Optimizes healthcare resource allocation.
  + User-friendly web interface for clinical use.
  + Robust model performance after optimization.
* **Disadvantages:**
  + Reliance on quality and completeness of input data.
  + Generalizability might be limited by dataset diversity.
  + Interpretation of complex models like SVM and Random Forest can be challenging for direct clinical insights without further explanation.

**7. Conclusion**

This project successfully developed and optimized machine learning models for early prediction of liver cirrhosis using patient health data. By addressing data quality issues, performing comprehensive feature selection, and fine-tuning model hyperparameters, high predictive accuracy was achieved. The proposed system has the potential to significantly improve patient outcomes and optimize healthcare efficiency by enabling earlier diagnosis and more informed clinical decisions.

**8. Future Scope**

Future work could include:

* Integrating real-time data streams from electronic health records (EHRs).
* Exploring more advanced deep learning architectures for improved prediction accuracy.
* Conducting external validation with diverse patient cohorts to ensure generalizability.
* Developing explainable AI (XAI) features to provide more transparent and trustworthy predictions to healthcare professionals.
* Incorporating additional data modalities, such as imaging data or genetic markers, for a more holistic risk assessment.

**9. Appendix**

**9.1. Source Code**

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/tree/main/5.%20Project%20Executable%20File)

**9.2. GitHub & Project Demo Link**

Link : [Github](https://github.com/DhawanSinghSaini/Liver-Cirrhosis-Prediction/tree/main/5.%20Project%20Executable%20File).