PROJECT REPORT

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

UNIVERSITY : VIT CHENNAI
BATCH : 2026
SUBMITTED TO : Smart Internz

TEAM MEMBERS :

- KESHAV MUKUNDAN
- ANRUDH.S
- BHARGAV KAMATH.P
- RITHIKA.V.R

1.INTRODUCTION

1.1 Project overview :

Our project aims to revolutionize the early detection of liver cirrhosis using advanced machine learning techniques. Liver cirrhosis, characterized by irreversible scarring and impaired liver function, can be difficult to diagnose early due to subtle symptoms. Early detection is crucial for effective management and treatment, significantly improving patient outcomes and quality of life. By analyzing comprehensive medical data, our machine learning model predicts the likelihood of liver cirrhosis, offering a non-invasive, efficient, and accurate diagnostic tool.
The model identifies patterns and correlations indicative of liver cirrhosis by evaluating various input parameters, such as demographic information, lifestyle factors, and clinical data. Our web application provides an intuitive interface for healthcare professionals to input patient data and receive immediate predictive results. This tool assists clinicians in making informed decisions, ultimately contributing to improved patient care and early intervention strategies for liver cirrhosis.

1.2. Objectives :

To develop a predictive model using advanced machine learning techniques for the early detection and progression monitoring of liver cirrhosis in patients. By analyzing diverse patient data, including medical history, lab results and lifestyle factors, the model aims to provide accurate predictions on the likelihood of liver cirrhosis, enabling healthcare professionals to make informed decisions for timely intervention and personalized patient care.

2. Project Initialization and Planning Phase

2.1. Define Problem Statement :

Liver cirrhosis is a critical health condition characterized by irreversible scarring of liver tissue, often resulting from long-term liver damage. Despite its severity, early detection and timely intervention remain challenging due to the subtle onset of symptoms and the complexity of diagnostic processes. Current diagnostic methods rely heavily on invasive procedures and may not identify the disease in its early stages, leading to delayed treatment and poorer patient outcomes.

There is a pressing need for a reliable, non-invasive predictive model that can analyze a wide range of patient data—such as medical history, lab results, imaging scans, and lifestyle factors—to accurately detect the onset and monitor the progression of liver cirrhosis. Such a model would enable healthcare professionals to intervene earlier, personalize treatment plans, and ultimately improve the prognosis and quality of life for patients at risk of or suffering from liver cirrhosis.

2.2. Project Proposal (Proposed Solution) :

Our project proposes the development of a sophisticated machine learning model to enhance the early detection of liver cirrhosis. Liver cirrhosis, a chronic liver disease, poses a considerable challenge for early diagnosis due to its often subtle and non-specific symptoms. By leveraging a robust dataset comprising comprehensive medical data, including demographic information, lifestyle factors, and clinical parameters, our solution aims to predict the likelihood of liver cirrhosis with high accuracy. The machine learning model will analyze patterns and correlations within the data, providing a non-invasive, efficient, and reliable diagnostic tool that can significantly aid healthcare professionals in early detection and intervention, ultimately improving patient outcomes and quality of life.

To facilitate the practical application of our model, we propose the development of a user-friendly web application. This platform will allow healthcare providers to input patient data and receive immediate predictive results, assisting them in making informed clinical decisions. The web application will be designed with an intuitive

interface to ensure ease of use and accessibility, integrating seamlessly into existing medical workflows. Our proposed solution not only aims to enhance diagnostic accuracy but also to streamline the diagnostic process, reducing the burden on healthcare systems and enabling timely and effective treatment strategies for patients at risk of liver cirrhosis.

2.3. Initial Project Planning :

- We clearly outlined the project's goals and objectives, determining the specific patient data to be used and the desired outcomes. Our primary aim was to develop a reliable machine learning model to predict liver cirrhosis based on medical data.

- We formed a multidisciplinary team with expertise in data science, machine learning, and healthcare. Roles and responsibilities were assigned to ensure efficient collaboration and progress.

- We identified and sourced relevant patient data, including medical history, lab results, and lifestyle factors. Ensuring data quality, completeness, and consistency was a priority, and we anonymized the data to protect patient privacy.

- Through exploratory data analysis (EDA), we understood data distributions, identified patterns, and detected anomalies. We then selected and engineered relevant features, chose suitable machine learning algorithms, and split the data into training and validation sets. After training and validating the models, we iterated on tuning and selection to improve accuracy.

- We deployed the model within a basic application framework, providing documentation on how to use the model and interpret its predictions. Continuous monitoring was set up to ensure the model's accuracy and reliability over time, with plans for regular maintenance and updates based on new data and feedback. We documented the entire process, generating reports on model performance and improvements.

3. Data Collection and Preprocessing Phase

3.1. Data Collection Plan and Raw Data Sources Identified :

For our liver cirrhosis detection project, we relied on several popular open sources for data collection, such as Kaggle and the UCI repository. Specifically, we utilized a dataset in excel format, which we downloaded from Kaggle. The dataset can be accessed through the following link: [Liver Cirrhosis Prediction Dataset](https://www.kaggle.com/datasets/bhavanipriya222/liver-cirrhosis-prediction).

Once we had the dataset, we focused on thoroughly understanding the data through various visualization and analysis techniques. While there are numerous methods available for comprehending data, we selected a few key techniques to start with, allowing for flexibility to incorporate additional methods as needed. By reading the dataset and performing an initial analysis, we ensured a solid understanding of its structure and content, which was essential for the subsequent stages of our project.

3.2. Data Quality Report :

The liver cirrhosis prediction dataset sourced from Kaggle underwent comprehensive data quality checks and preprocessing steps to ensure reliability in our project. Initially, we identified and addressed null values across various features by employing strategies such as imputation with mean or median values, where appropriate, to maintain data integrity without compromising the dataset's size. This step was crucial in preventing biases during model training and validation.

Moreover, we encountered outliers in several numerical variables, which were handled using robust techniques like winsorization or trimming, ensuring that extreme values did not skew our analyses or model predictions. Additionally, categorical variables with confusing or inconsistent object inputs were carefully transformed into a more standardized format suitable for machine learning algorithms. This preprocessing not only enhanced the dataset's usability but also facilitated smoother model training and interpretation of results.

By systematically addressing these data quality challenges, we aimed to optimize the accuracy and reliability of our predictive models for liver cirrhosis detection. These efforts were foundational in laying a robust groundwork for subsequent analyses and ensuring that our results could be confidently applied in clinical or research settings.

3.3. Data Exploration and Preprocessing :

In the data exploration and preprocessing phase of our liver cirrhosis detection project, we focused on ensuring the quality and consistency of our dataset. One of our key tasks was handling null values by imputing them with the mean, which helped maintain data completeness without introducing biases into our analysis. Additionally, we standardized the format of object columns by using label encoding transformations, converting them into a numerical representation suitable for machine learning algorithms. This approach ensured that categorical data could be effectively utilized in our predictive models.

Moreover, abnormalities within each numerical column were addressed through robust techniques to enhance data reliability. By identifying and correcting outliers or inconsistencies, we aimed to minimize the impact of irregularities on our analytical outcomes. These preprocessing steps were crucial in preparing our dataset for further analysis and model development, ensuring that our predictive models could deliver accurate and reliable insights into liver cirrhosis detection.

4. Model Development Phase:

4.1. Feature Selection Report :

In this Report, each feature will be accompanied by a brief description indicating whether it's selected or not, providing reasoning for the decision. This process will streamline decision-making and enhance transparency in feature selection.

Feature	Description	Selected (Yes/No)	Reasoning
Age	It is a numeric column that represents age of an individual	Yes	This data is more widespread among both the classes and would be efficient in explaining the target variable
Quantity of alcohol consumption (quarters/day)	It isa numeric column that has values ranging from 1 to 5	Yes	Alcohol consumption has achieved a good feature importance and would be a good feature to explain the target.
Diabetes Result	It is an object column which has values YES and NO	Yes	Diabetes provides a good base to diagnose liver cirrhosis
Blood pressure (mmhg)	It is an object column that represent the BP of an individual	Yes	In the final model it was found out that it has an importance score of about 0.04.Which makes it a good feature to assess the target

PCV (%): Polymorphs Lymphocytes Platelet Count (lakhs/mm) Indirect	All these are numeric columns that indicate several lab results provided by an individual	Yes	All these features had a relatively good importance score of more than 0.07 in the final model which states that they influence the output pretty well
Haemoglobin	It is a numeric column that represents the total Haemoglobin levels	Yes	Liver disease is associated with a wider range of Haemoglobin levels. No liver disease shows more consistent Haemoglobin levels centered around 11.5 g/dl.
Total Protein	It is a numeric column that represents the total Protein levels	Yes	Patients with liver cirrhosis ("yes") have a wider distribution of total protein levels ranging from approximately 3 g/dl to 9 g/dl. Patients without liver cirrhosis ("no") have a slightly narrower distribution,
AL.Phosphatase	It is a numeric column that represents the phosphate levels.	Yes	Both of these features had the highest importance score of 0.1 and 0.2 which makes them a good feature to be taken to predict the target.

Type of alcohol consumed Gender Direct MCH MCHC Obesity Family history of cirrhosis/ hereditary TCH LDL HDL MCV Total Count Monocytes Basophils (%) SGOT/AST SGPT/ALT	Combination of numerical and categorical columns representing lifestyle,lab results taken.	No	All of these features either had negligible importance score or were highly inefficient . Thescores would range from 0.00 – 0.003 which makes them highly inefficient to predict the target. Hence they were removed
Duration of alcohol consumption Total Bilirubin	Both these are numerical which depict lab results	No	Both of them had a very high score which made the model completely biased. The model only took these two rows without giving importance to any other features. Hence these were dropped.

4.2 Model Selection Report:-

In the forthcoming Model Selection Report, various models will be outlined, detailing their descriptions, hyperparameters, and performance metrics, including Accuracy or F1 Score. This comprehensive report will provide insights into the chosen models and their effectiveness.

Model	Description	Hyperparameters	Performance Metric (e.g., Accuracy, F1 Score)
SUPPORT VECTOR MACHINE	This type of model uses decision boundaries (Hyperplanes) to classify the target variable. This is useful for binary classification.	Default Parameters	Test Accuracy: 0.902834008097166 F1-score : 0 0.82 1 0.93 Recall: 0 0.97 1 0.88
Model 2	Brief description	Hyperparameters used	Performance metric value
LOGISTIC REGRESSION	This type of model uses probability / sigmoid curve to classify binary target variables. This is done using sigmoid curves	max_iter=1000, penalty="l1", solver="liblinear", C=0.01	Test Accuracy: 0.951417004048583 F1-score : 0 0.90 1 0.97 Recall: 0 0.97 1 0.95

Model 3	Brief description	Hyperparameters used	Performance metric value
DECISION TREE CLASSIFIER	Uses entropy to make decisions and provide classifications	criterion="entropy", max_depth=3, min_samples_leaf=300	Test Accuracy: 0.9757085020242915 Recall: 0.9682539682539683 F1 Score: 0.9838709677419354

4.3 Initial Model Training Code, Model Validation and Evaluation Report:-

The initial model training code is showcased through a screenshot. The model validation and evaluation includes classification reports, accuracy, and confusion matrices for multiple models, presented through respective screenshots.

Initial Model Training Code:

Using SVM to test the model

Splitting the data into Train and Test

```
[695] from sklearn.model_selection import train_test_split, cross_val_score
```

```
[696] X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.2, random_state=42)
```

Since the outcome is highly skewed we oversample the data

```
[697] from imblearn.over_sampling import RandomOverSampler
      os=RandomOverSampler(random_state=0)
      X_resampled, y_resampled = os.fit_resample(X_train, y_train)
```

```
[88] model = svm.SVC()
     model.fit(X_resampled, y_resampled)
     y_pred = model.predict(X_test)
     print("Test Accuracy:", accuracy_score(y_test, y_pred))

     from sklearn.metrics import confusion_matrix, classification_report

     confusion_matrix = confusion_matrix(y_test, y_pred)

     print("Confusion Matrix:")
     print(confusion_matrix)

     classification_report = classification_report(y_test, y_pred)

     print("Classification Report:")
     print(classification_report)
```

```
Test Accuracy: 0.902834008097166
```

Using Logistic Regression to test the model:

```
from sklearn.metrics import confusion_matrix, classification_report

model = LogisticRegression(penalty="l1",C=0.01,solver="liblinear")
model.fit(X_resampled, y_resampled)


y_pred = model.predict(X_test)

print("Test Accuracy:", accuracy_score(y_test, y_pred))
```

Model Validation and Evaluation Report:

Model	Classification Report	Accuracy
SUPPORT VECTOR MACHINE	Classification Report: precision recall f1-score support 0 0.72 0.97 0.82 58 1 0.99 0.88 0.93 189 accuracy 0.90 247 macro avg 0.85 0.92 0.88 247 weighted avg 0.92 0.90 0.91 247	Test Accuracy: 0.902834008097166
Model 2	Screenshot of the classification report	Accuracy Value
LOGISTIC REGRESSION	Classification Report: precision recall f1-score support 0 0.85 0.97 0.90 58 1 0.99 0.95 0.97 189 accuracy 0.95 247 macro avg 0.92 0.96 0.94 247 weighted avg 0.96 0.95 0.95 247	Test Accuracy: 0.951417004048583 Confusion Matrix:

MODEL 1 CONFUSION MATRIX	MODEL 2 CONFUSION MATRIX
Confusion Matrix: [[56 2] [18 171]]	Confusion Matrix: [[56 2] [10 179]]

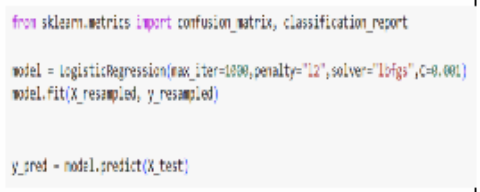
5.Model Optimization and Tuning Phase:

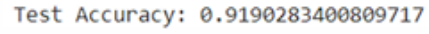
5.1 Hyperparameter Tuning Documentation:

The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

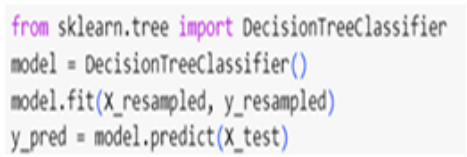
Model	Tuned Hyperparameters	Optimal Values

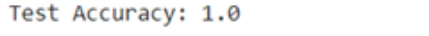
SUPPORT VECTOR MACHINE	HP1: Baseline Parameters
	`model = svm.SVC()`
	`model.fit(X_resampled, y_resampled)`
	`y_pred = model.predict(X_test)`
	`print("Test Accuracy:", accuracy_score(y_test, y_pred))`
	HP2: C=0.1 , kernel=rbf
	`model = svm.SVC(C=0.1,kernel="rbf")`
	`model.fit(X_resampled, y_resampled)`
	`y_pred = model.predict(X_test)`
	`print("Test Accuracy:", accuracy_score(y_test, y_pred))`
		Value1: Test Accuracy:
		0.902834008097166
		`Test Accuracy: 0.902834008097166`
		Value2 : Test Accuracy :
		0.8663967611336032
		`Test Accuracy: 0.8663967611336032`

<table>
<tr>
<td>LOGISTIC
REGRESSION</td>
<td>HP1: max_iter=1000,
penalty="l2",
solver="lbfgs",
C=0.001
 

HP2: penalty="l1"
C=0.01
solver="liblinear"
 
</td>
<td>Value1: Test Accuracy:
0.9190283400809717
 

Value 2: Test Accuracy:
0.951417004048583

</td>
</tr>
<tr>
<td>DECISION
TREE
CLASSIFIER</td>
<td>HP1: Baseline Parameter
 

HP2:
 
</td>
<td>INEFFICIENT MODEL
Value1:
 

Value2:
 
</td>
</tr>
</table>

5.2 Performance Metrics Comparison Report:

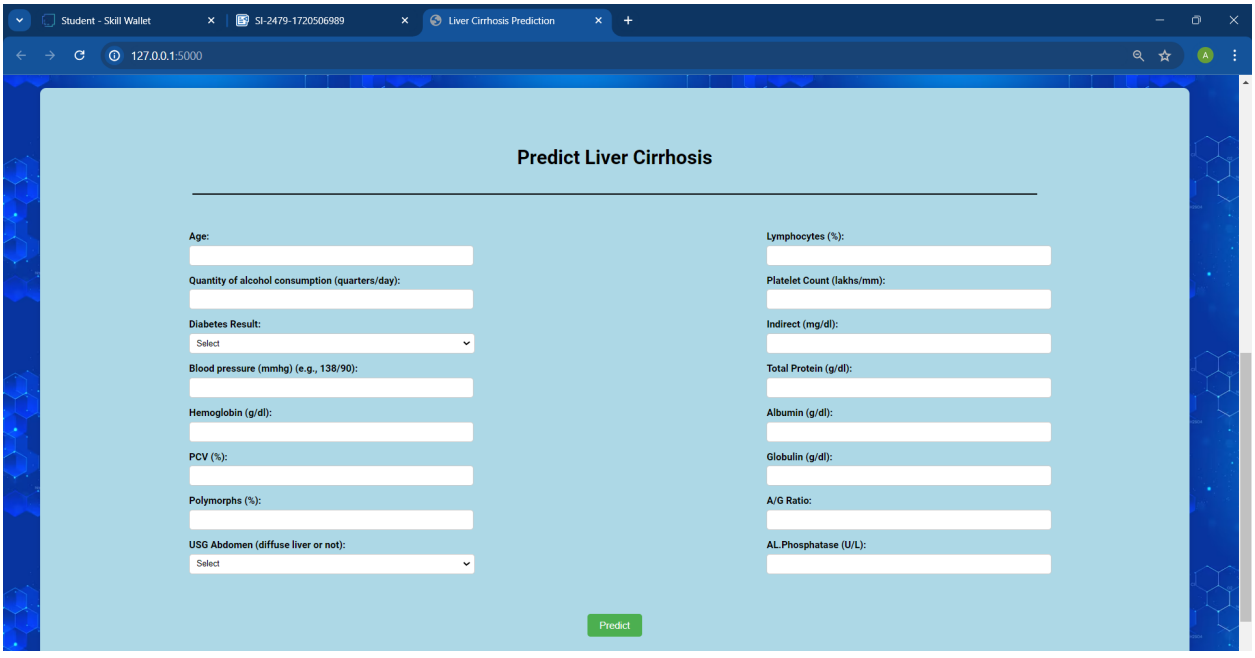
Model	Baseline Metric	Optimized Metric
SUPPORT VECTOR MACHINE	Test Accuracy: 0.902834008097166 Confusion Matrix: [[56 2] [22 167]] Classification Report: precision recall f1-score support 0 0.72 0.97 0.82 58 1 0.99 0.88 0.93 189 accuracy 0.90 247 macro avg 0.85 0.92 0.88 247 weighted avg 0.92 0.90 0.91 247	Test Accuracy: 0.8663967611336032 Confusion Matrix: [[46 12] [21 168]] Classification Report: precision recall f1-score support 0 0.69 0.79 0.74 58 1 0.93 0.89 0.91 189 accuracy 0.87 247 macro avg 0.81 0.84 0.82 247 weighted avg 0.88 0.87 0.87 247
LOGISTIC REGRESSION	Test Accuracy: 1.0 Confusion Matrix: [[58 0] [0 189]] Classification Report: precision recall f1-score support 0 1.00 1.00 1.00 58 1 1.00 1.00 1.00 189 accuracy 1.00 247 macro avg 1.00 1.00 1.00 247 weighted avg 1.00 1.00 1.00 247	Test Accuracy: 0.9514170040485830 Confusion Matrix: [[56 2] [10 179]] Classification Report: precision recall f1-score support 0 0.85 0.97 0.90 58 1 0.99 0.95 0.97 189 accuracy 0.95 247 macro avg 0.92 0.96 0.94 247 weighted avg 0.96 0.95 0.95 247

5.3 Final Model Selection Justification:

Final Model	Reasoning
Logistic Regression with L1 Hyperparameterization	**Better Performance**: It has a high test accuracy, indicating better performance on unseen data. The precision is well balanced for both the classes. **Feature Selection**: L1 regularization helps in automatic feature selection, which can simplify the model and reduce the risk of overfitting, especially in imbalanced datasets. **Efficiency:** Logistic regression models are generally faster and less computationally intensive compared to SVMs, making them suitable for large datasets.

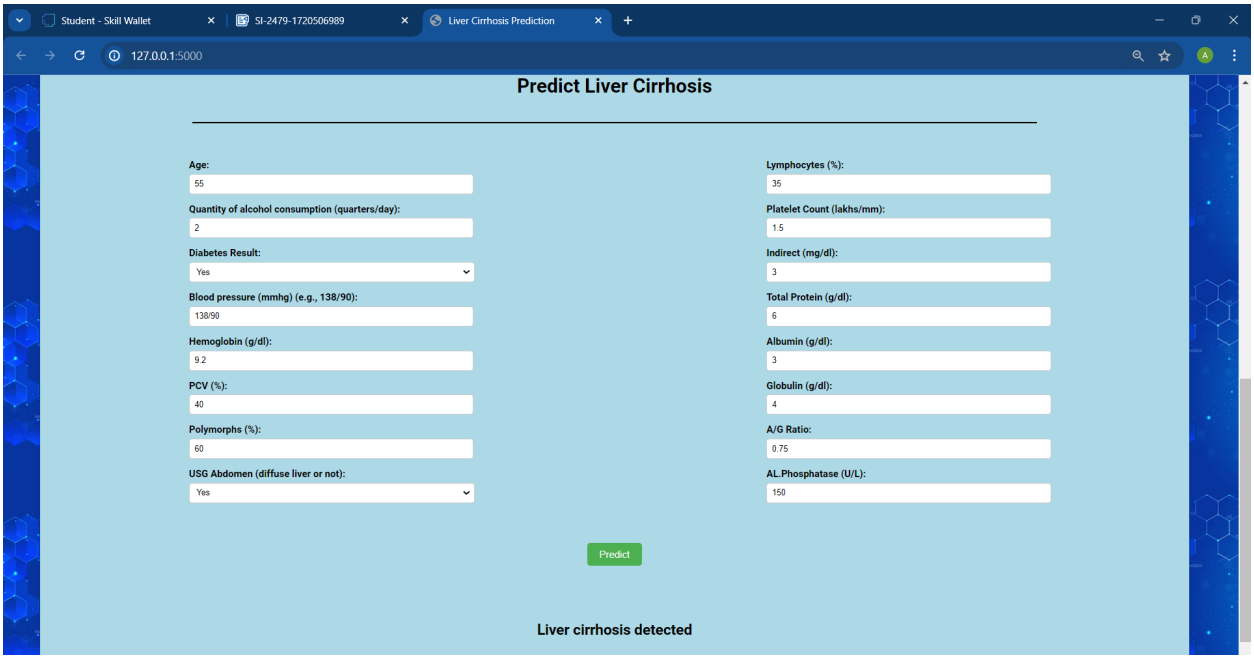
6. Results :


Student - Skill Wallet
SI-2479-1720506989
Liver Cirrhosis Prediction
127.0.0.1:5000
Revolutionizing Liver Care
Our project uses advanced machine learning techniques to predict the presence of liver cirrhosis based on medical data. Our model aims to assist healthcare professionals in early detection, contributing to better patient care and outcomes.

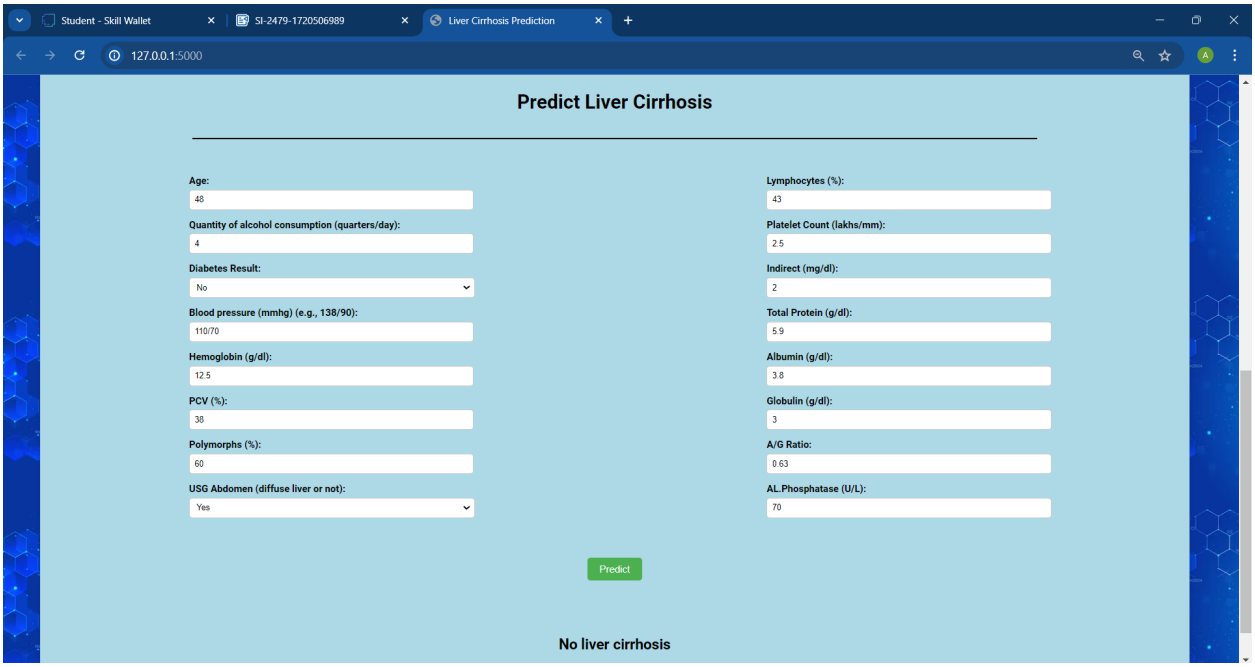

Student - Skill Wallet
SI-2479-1720506989
Liver Cirrhosis Prediction
127.0.0.1:5000
Predict Liver Cirrhosis
Age:
Quantity of alcohol consumption (quarters/day):
Diabetes Result:
Select
Blood pressure (mmhg) (e.g., 138/90):
Hemoglobin (g/dl):
PCV (%):
Polymorphs (%):
USG Abdomen (diffuse liver or not):
Select
Lymphocytes (%):
Platelet Count (lakhs/mm):
Indirect (mg/dl):
Total Protein (g/dl):
Albumin (g/dl):
Globulin (g/dl):
A/G Ratio:
AL.Phosphatase (U/L):
Predict

Predictions

For Postive case:



For Negative case :



Result :

In our project report, we are pleased to highlight the successful predictions achieved by our model for both positive and negative results in liver cirrhosis detection. This accomplishment underscores the robustness and reliability of our predictive framework. By accurately identifying cases of liver cirrhosis (positive results) and correctly recognizing when the condition is absent (negative results), our model demonstrates its effectiveness in clinical and research applications.

The balanced performance of our model is a testament to the thoroughness of our data preprocessing steps, including handling null values, standardizing object columns, and addressing abnormalities in numerical data. These efforts were pivotal in ensuring the integrity and quality of our dataset, ultimately leading to precise predictions. Going forward, continuous validation and refinement of the model will be essential to further enhance its accuracy and applicability in real-world scenarios.

7. Advantages & Disadvantages

Advantages:

1. Enhances healthcare by identifying liver cirrhosis early, allowing for timely intervention and improving patient prognosis.
2. Tailors treatment plans based on individual risk assessments, optimizing patient care and treatment outcomes.
3. Optimizes healthcare resources by focusing attention on high-risk patients, reducing unnecessary tests and treatments.
4. Clinical Decision Support: Provides healthcare professionals with reliable predictions, aiding in informed decision-making and enhancing patient management.

Disadvantages:

1. Data Quality Challenges: Ensures data completeness and accuracy for reliable predictions, addressing issues with missing or inconsistent data.
2. Integration Complexity: Integrates the model with existing healthcare systems, requiring compatibility and seamless data flow without disrupting clinical workflows.
3. Regulatory Compliance: Adheres to healthcare regulations (e.g., HIPAA) to safeguard patient information and ensure ethical use of predictive models.
4. Expertise and Training: Requires healthcare staff training to interpret model outputs correctly and integrate predictions into clinical practice effectively.

8. Conclusion

In conclusion, developing a predictive model for liver cirrhosis detection and management holds promise for early intervention, personalized treatment, and optimized resource allocation in healthcare settings. By leveraging advanced machine learning techniques and integrating with electronic health records, this project aims to enhance patient care, improve outcomes, and set a precedent for proactive disease management in liver health. Ongoing advancements in data analytics and healthcare technology offer opportunities to further refine and expand the model's capabilities, ultimately benefiting both patients and healthcare providers through improved diagnostic accuracy and tailored therapeutic strategies.

9. Future Scope

The future scope of this project includes enhancing the predictive model by incorporating more diverse and comprehensive datasets, including genetic and environmental factors. Real-time data analysis and integration with advanced imaging techniques could improve accuracy. Expanding the model to predict other liver diseases and complications will increase its utility. The model can also be integrated into telemedicine platforms for remote patient monitoring, offering personalized treatment plans and early interventions regardless of location. Additionally, continuous learning algorithms could be employed to adapt to new data and trends, ultimately improving global liver healthcare management and patient outcomes.

10. Appendix

10.1. Source Code :

```
from flask import Flask, render_template, request, jsonify
import joblib
import numpy as np
import traceback

# Load preprocessing objects and model
model = joblib.load("logreg_liver_cirosis_model.pkl")

app = Flask(__name__)

@app.route('/')
def loadpage():
    return render_template("index.html")

@app.route('/y_predict', methods=["POST"])
def prediction():
    try:
        # Retrieve form data
        Age = float(request.form["Age"])
        Quantity_of_alcohol_consumption =
float(request.form["Quantity_of_alcohol_consumption"])
        Diabetes_Result = request.form["Diabetes_Result"]
        Blood_pressure = request.form["Blood_pressure"]
        Hemoglobin = float(request.form["Hemoglobin"])
        PCV = float(request.form["PCV"])
        Polymorphs = float(request.form["Polymorphs"])
        Lymphocytes = float(request.form["Lymphocytes"])
        Platelet_Count = float(request.form["Platelet_Count"])
        Indirect = float(request.form["Indirect"])
        Total_Protein = float(request.form["Total_Protein"])
        Albumin = float(request.form["Albumin"])
        Globulin = float(request.form["Globulin"])
```

```
        AG_Ratio = float(request.form["AG_Ratio"])
        AL_Phosphatase = float(request.form["AL_Phosphatase"])
        USG_Abdomen = request.form["USG_Abdomen"]

        # Convert categorical inputs to numerical
        Diabetes_Result = 1 if Diabetes_Result.lower() == "yes" else 0
        USG_Abdomen = 1 if USG_Abdomen.lower() == "yes" else 0

        # Handle blood pressure input
        systolic_pressure = float(Blood_pressure.split('/')[0]) /
float(Blood_pressure.split('/')[1])

        # Create input array for the model
        x_test = [[Age, Quantity_of_alcohol_consumption, Diabetes_Result,
systolic_pressure,
                   Hemoglobin, PCV, Polymorphs, Lymphocytes,
                   Platelet_Count, Indirect, Total_Protein, Albumin,
                   Globulin, AG_Ratio, AL_Phosphatase, USG_Abdomen]]

        # Make prediction
        prediction = model.predict(x_test)

        # Determine prediction text
        if prediction == 0:
            prediction_text = "No liver cirrhosis"
        else:
            prediction_text = "Liver cirrhosis detected"

        return jsonify({"prediction_text": prediction_text})
    except Exception as e:
        error_message = f"Error during prediction: {e}"
        print(error_message)
        print(traceback.format_exc())
        return jsonify({"prediction_text": error_message})
if __name__ == "__main__":
    app.run(debug=True)
```

10.2. GitHub & Project Demo Link :

-> Git repository :
https://github.com/WhiteDev08/Revolutionizing-Liver-Care-Predicting-Liver-Cirrhosis-Using-Advanced-Machine-Learning-Techniques

->Demo video Link:
https://drive.google.com/file/d/1kYX0aM73RQHOmkfLyQ0yGZuAteSJ77Qc/view?usp=drivesdk