AUDIT ANALYTICS – LIVER PATIENT RISK PREDICTION

Project Report

Submitted by:
Vaishnavi Mishra
B.Tech - Electronics and Communication Engineering
Vellore Institute of Technology
Submission Date: September 2024

Self-made Project

1. Objective

The goal of this project is to design a machine learning-based system that predicts whether a patient is likely to suffer from liver disease based on key medical indicators such as enzyme levels, bilirubin content, and protein measurements. This project assists doctors in data-driven decision-making, improving diagnostic accuracy and enabling early medical interventions.

2. Abstract

Liver disease is a major health concern worldwide. Early detection and timely treatment are crucial for survival. This project applies machine learning algorithms — Logistic Regression and Random Forest — to predict liver disease likelihood based on the Indian Liver Patient Dataset (ILPD). After preprocessing and training, both models were evaluated using metrics such as Accuracy, Recall, and AUC. Random Forest provided the best balance between accuracy and sensitivity.

3. System Overview

Input: Patient data (age, gender, bilirubin levels, enzymes, protein levels, etc.)
Process:
1. Data cleaning and preprocessing
2. Encoding categorical variables
3. Feature scaling and splitting
4. Model training and testing
5. Performance evaluation and visualization
Output: Predicted result – Liver Disease or No Disease.

4. Tools & Technologies Used

Programming Language: Python
IDE: Google Colab / Jupyter Notebook
Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn
Dataset: Indian Liver Patient Dataset (ILPD)
Algorithms: Logistic Regression, Random Forest

5. Methodology / Implementation Steps

Step 1: Data loading using pandas
Step 2: Data preprocessing (handling missing values, encoding gender)
Step 3: Train-test split (80-20 ratio)
Step 4: Model training with Logistic Regression & Random Forest
Step 5: Evaluation using Accuracy, Recall, AUC, Confusion Matrix
Step 6: Hyperparameter tuning with GridSearchCV for Random Forest

6. Results and Analysis

Both Logistic Regression and Random Forest initially achieved 71.55% accuracy. After tuning, Random Forest improved to around 78% accuracy with higher recall and AUC, indicating better ability to correctly identify liver disease cases.

7. Challenges and Solutions

Challenge	Solution
Missing or inconsistent data	Dropped null values and cleaned dataset
Unequal class distribution	Used stratified train-test split
Low initial accuracy	Applied feature scaling and hyperparameter tuning
Model interpretability	Compared Logistic Regression with Random Forest

8. Conclusion

This project successfully implemented a Liver Patient Risk Prediction System using machine learning. The Random Forest Classifier outperformed Logistic Regression, achieving ~78% accuracy and strong recall after optimization. The system can assist healthcare professionals in early diagnosis and reduce medical risks.

9. Future Scope

1. Apply deep learning models (ANNs) for higher accuracy.
2. Deploy as a web or mobile app for real-time prediction.
3. Incorporate larger and more diverse datasets.
4. Use SHAP or feature importance analysis for better interpretability.

10. References

1. Indian Liver Patient Dataset (UCI Repository)
2. Scikit-learn Documentation
3. Breiman, L. “Random Forests,” Machine Learning, 2001.
4. Research papers on liver disease prediction using ML.