PROJECT ABSTRACT

Liver Disease Prediction using Machine Learning

Liver-disease prediction on the Indian Liver Patient Records dataset (UCI, 583 records): imbalance-aware pipeline, five classifiers, confusion-matrix-led evaluation. Built-to-order with code, training notebook, report, PPT and viva kit.

01. Title

Liver Disease Prediction using Machine Learning

02. Introduction / Background

Liver disease is among the leading causes of adult mortality in India, yet its early biochemical signals — rising bilirubin, climbing SGPT/SGOT transaminases, falling albumin — sit in routine blood panels long before symptoms. Reading those panels is pattern recognition over ten correlated variables. The Indian Liver Patient Records dataset (UCI ML Repository, 583 records from north-east Andhra Pradesh) captures the problem publicly: 416 liver-disease and 167 healthy records across 10 clinical variables. It is a genuinely hard task — class lab ranges overlap heavily — so this project is built around honest evaluation: five classifiers compared on the same stratified split, SMOTE on the training fold only, and the confusion matrix and per-class recall leading the results.

03. Problem Statement

The ILPD task punishes lazy evaluation: a single accuracy number hides the minority-class story, and naive oversampling leaks synthetic points into validation. The project addresses this with an imbalance-aware pipeline — stratified splits, training-fold-only SMOTE, per-class metrics — comparing XGBoost, random forest, logistic regression, SVM and KNN, plus a gender-stratified analysis (published work shows markers behave differently across sexes here) and a live risk-estimator interface over the 10 clinical variables.

04. Objectives

- Compare five classifiers on the same stratified 80/20 split with a design target of ~0.72 test accuracy — reported honestly with confusion matrix and per-class recall.
- Handle the 71/29 imbalance with stratified splits and SMOTE on the training fold only.

- Run a gender-stratified performance breakdown in the notebook.
- Deliver the risk-estimator interface (reference ranges shown) plus code, notebook, report, PPT and viva Q&A.

05. Proposed Methodology

Four stages. (1) Profiling: class balance, gender skew (441 male / 142 female) and feature distributions. (2) Preprocessing: StandardScaler, gender encoding, stratified 80/20 split; SMOTE applied to the training fold only. (3) Modeling: five classifiers with grid-searched hyperparameters under 5-fold stratified CV. (4) Evaluation: confusion matrix, per-class precision/recall/F1, ROC-AUC, gender-stratified breakdown, and an error analysis of which lab patterns confuse the models.

06. System Architecture / Working

Data layer loads the ILPD records. Pipeline layer chains scaling, encoding, SMOTE (training fold only) and the classifier. Experiment layer: the notebook trains five models, plots confusion matrices and ROC curves, and tabulates per-class and gender-stratified metrics. Serving layer: the serialized best model powers the estimator interface with per-variable contribution bars and reference ranges.

07. Hardware / Software Requirements

- Any 64-bit PC (8 GB RAM); CPU-only — no GPU needed
- Python 3.10, scikit-learn, XGBoost, imbalanced-learn, pandas, NumPy
- Matplotlib, Seaborn (plots), Jupyter (notebook)
- joblib (model serialization)
- A modern browser for the estimator interface

08. Expected Outcomes

An imbalance-aware pipeline, a best model near the ~0.72 design target with the confusion matrix front and center, gender-stratified results, an error analysis of misclassified lab patterns, and a working estimator interface. This project is an academic research prototype and is not a medical device; it must not be used for diagnosis or clinical decisions.

09. Applications

- Teaching class-imbalance handling and SMOTE discipline
- Template for hard tabular clinical tasks
- Demonstration of confusion-matrix-led reporting
- Baseline for cost-sensitive learning experiments

10. Future Scope

- Cost-sensitive learning vs SMOTE comparison
- Calibration analysis on the minority class
- Multi-cohort validation study design
- Integration of imaging features alongside lab values