Early Prediction of Gallstone using Statistical and Machine Learning Techniques

Submitted to **Prof. Sabyasachi Mukhopadhyay**

GROUP NO.- 12

Source of Dataset: https://archive.ics.uci.edu/dataset/1150/gallstone-1

Name	Roll No.	Email Id	Phone No.
Arunava Mukherjee	24BM6JP11	arunavamba2026@email.iimcal.ac.in	8017729727
Prabudha Durge	24BM6JP17	prabuddhavdba2026@email.iimcal.ac.in	7385658416
Harshal Ugalmugle	24BM6JP20	harshaluba2026@email.iimcal.ac.in	9082431536
Hemraj Chakravarti	24BM6JP22	hemrajcba2026@email.iimcal.ac.in	7049729528

1. Introduction

This project focuses on predicting the **gallstone status** of patients using clinical and biometric data. The dataset contains **319 patient records** and **39 features**, out of which **7 are categorical** (e.g., Gender, Comorbidity, CAD, Hypothyroidism etc), and the rest are numerical measurements such as age, weight, cholesterol, glucose levels, and body composition metrics.

The **target variable** is Gallstone, indicating whether the patient has gallstones (1) or not (0).

Business Use Case

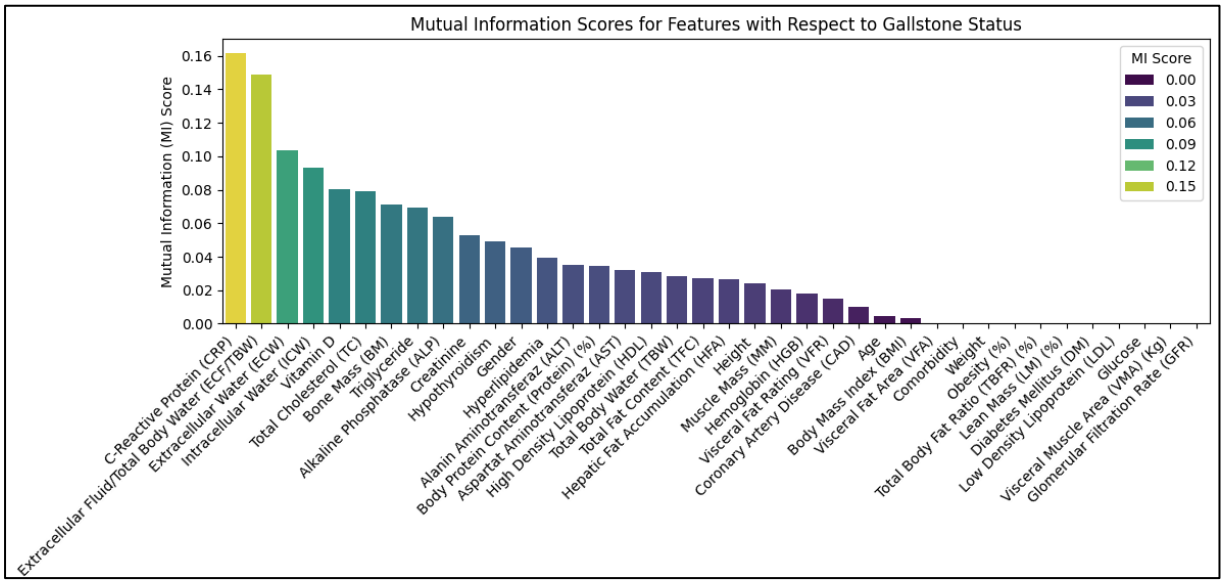
Gallstones are a common but often undiagnosed condition that can lead to severe health issues if not detected early. This classification model can help healthcare providers:

- Identify high-risk patients during routine exams,
- Optimize use of diagnostic imaging,
- Offer preventive care and personalized interventions.

By predicting gallstone risk accurately, this solution supports early diagnosis, reduces healthcare costs, and improves patient outcomes.

2. Feature Selection Using Mutual Information

To identify the most relevant features for predicting gallstone status, we computed the **Mutual Information (MI)** scores between each feature and the target variable (Gallstone_Status). MI measures the amount of information one variable contains about another, making it well-suited for capturing both linear and non-linear dependencies.
Using `mutual_info_classif` from scikit-learn, we calculated MI scores for all 38 predictor variables. Features with higher MI scores were considered more informative for the classification task.
A bar plot of the MI scores was generated to visualize the relative importance of each feature. This helped in selecting the top predictors and reducing noise from less relevant variables, ultimately aiding in building a more efficient and accurate model.



3. Statistical Significance Testing and Feature Elimination

To validate feature relevance beyond Mutual Information scores, we performed:

- **Chi-Square Tests** for categorical features, and
- **Mann-Whitney U Tests** for continuous variables.

This helped us shortlist **five features** as statistically insignificant. To rigorously assess their combined predictive power, we fit a **logistic regression model** using only these five features and evaluated it using a **Likelihood Ratio Test (LRT)** against an intercept-only model.

Logistic Regression Summary (Selected Features)	**Key Coefficients (with p-values):**
Log-Likelihood (Full Model): -164.11	**Visceral Fat Rating (VFR):** Positive & significant ($p < 0.001$)
Log-Likelihood (Null Model): -176.75	**Age:** Negative & significant ($p < 0.001$)
Pseudo R-squared: 0.071	**Visceral Fat Area (VFA):** Negative & significant ($p < 0.001$)
LLR p-value: 0.00012 (Significant)	**Coronary Artery Disease & BMI:** Not significant

While the initial assumption was that these five features might be uninformative, the **Likelihood Ratio Test showed statistical significance (p = 0.00012)**, indicating the model with these features fits significantly better than the null model. However, within the group, only **three variables (VFR, Age, VFA)** showed individual significance, suggesting they may still hold predictive value and shouldn't be discarded outright.

5. Model Building and Evaluation

After refining the feature set through mutual information, hypothesis testing, and logistic regression diagnostics, we trained and evaluated multiple classification models to predict **Gallstone Status**. The goal was to identify the best-performing algorithm in terms of accuracy, interpretability, and generalizability.

Models	Grid search best parameters	Performance metric
Logistic Regression	baseline model	Nagelkerke R^2: 0.6097 Test ROC-AUC: 0.8524 Test Accuracy: 0.7969
Decision Tree	**max_depth:** 2	Test Accuracy: 0.7500 Test Precision: 0.7576 Test Recall: 0.7576 Test F1 Score: 0.7576 Test ROC-AUC: 0.8016
Random Forest	max_features=sqrt, max_samples=0.75, n_estimators=1200, min_samples_split=2	Test Accuracy: 0.8125 Test Precision: 0.7838 Test Recall: 0.8788 Test F1 Score: 0.8286 Test ROC-AUC: 0.8895
XGBoost	booster=dart, eta=0.4, alpha=0, n_estimators=400	Test Accuracy: 0.8594 Test Precision: 0.8750 Test Recall: 0.8485 Test F1 Score: 0.8615 Test ROC-AUC: 0.8817

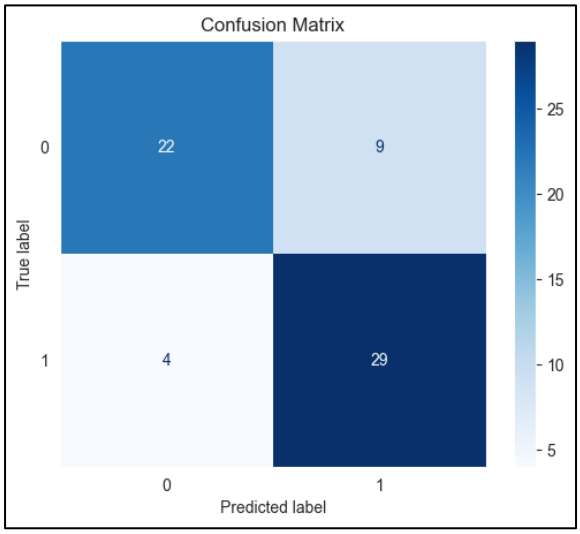
XGBoost delivered the **best overall performance**, achieving the highest test accuracy and strong generalization, making it the most promising model for deployment.

6. Model Interpretability and Evaluation Visuals

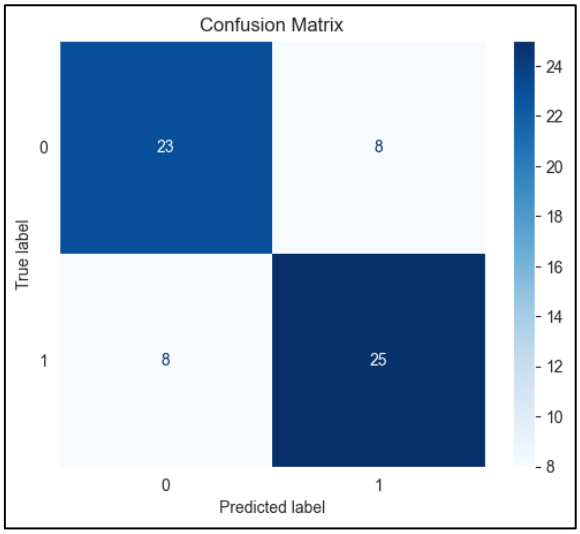
To gain deeper insights into model behavior and performance, several diagnostic and interpretability plots were generated:

6.1 Confusion Matrices

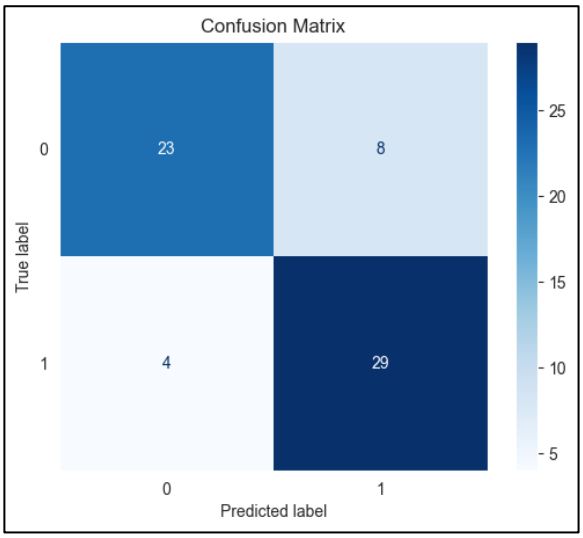
Confusion matrices for each model were plotted to visualize the distribution of true positives, true negatives, false positives, and false negatives.



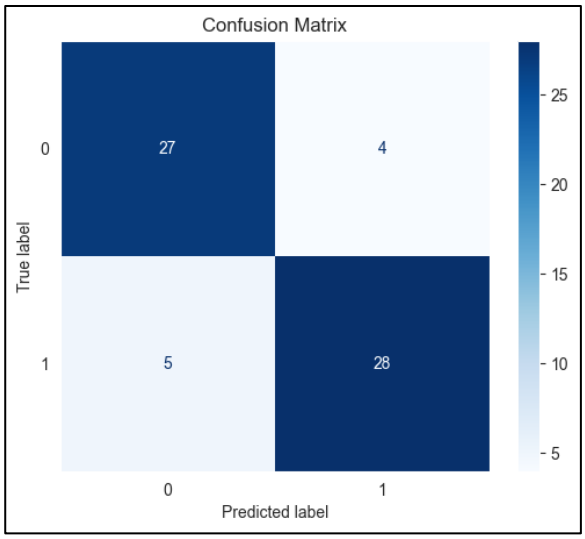
Logistic Regression



Decision Tree



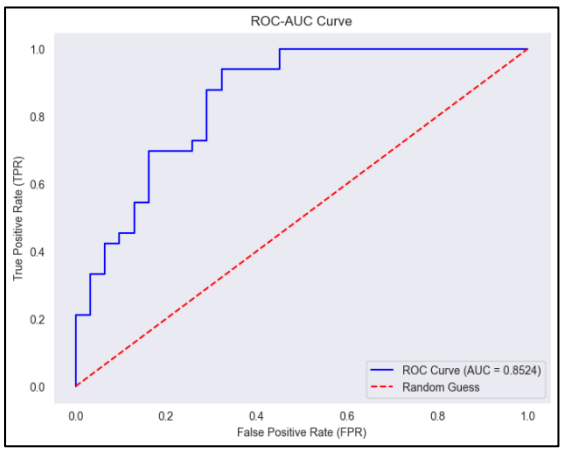
Random Forest

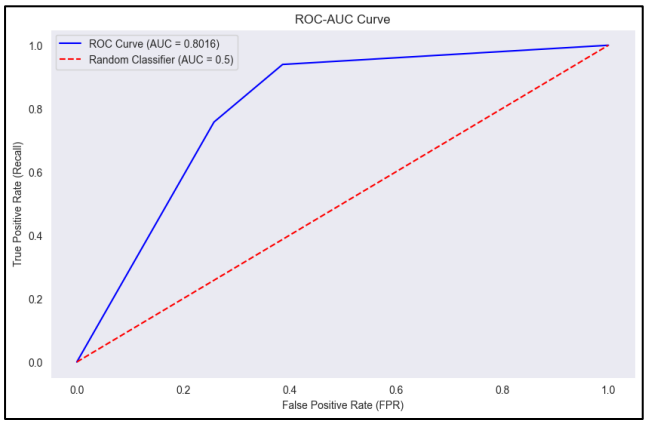


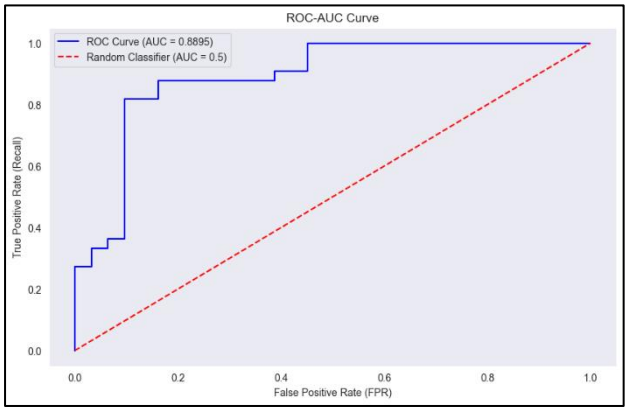
XGBoost

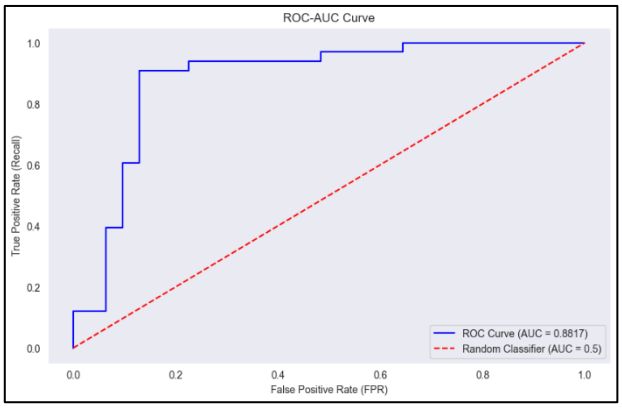
6.2 ROC Curves

ROC curves for all models were plotted to assess the trade-off between sensitivity and specificity across thresholds.



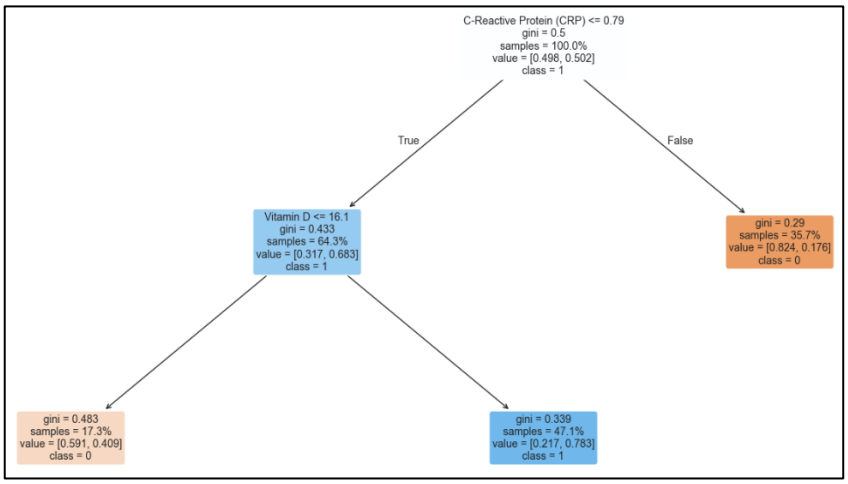






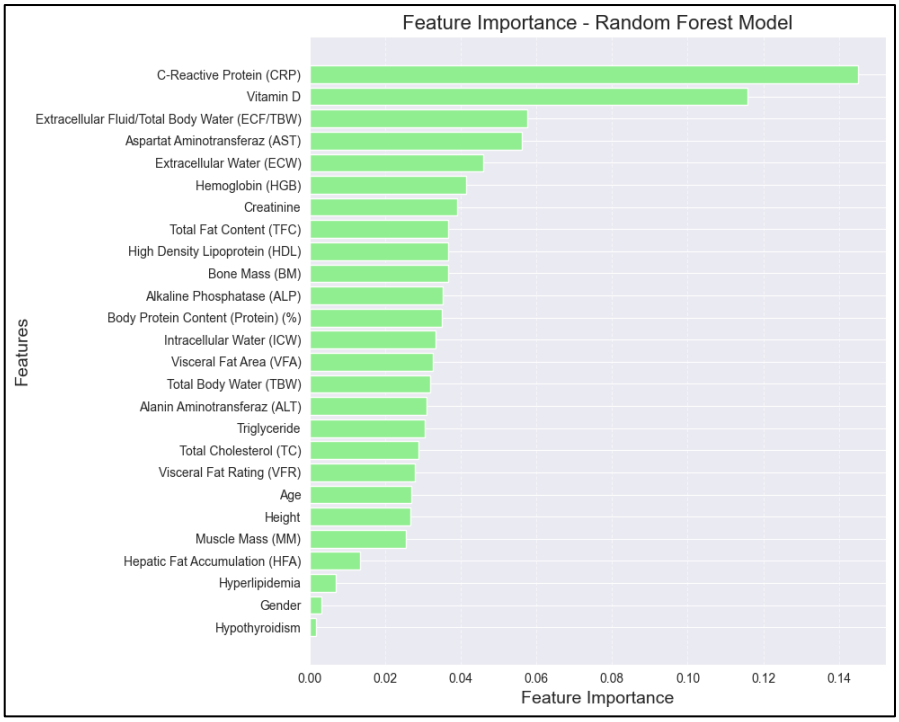
6.3 Decision Tree Visualization

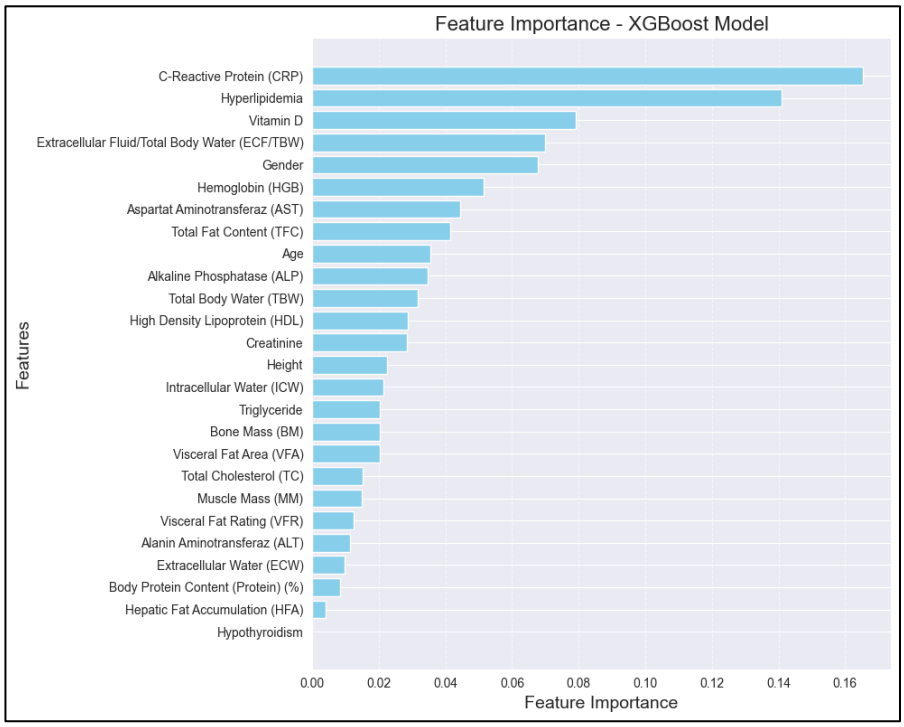
The optimal decision tree (depth = 2) was visualized using a tree plot. Despite its simplicity, the tree captures some key splits that can be easily interpreted by clinicians.



6.4 Feature Importance

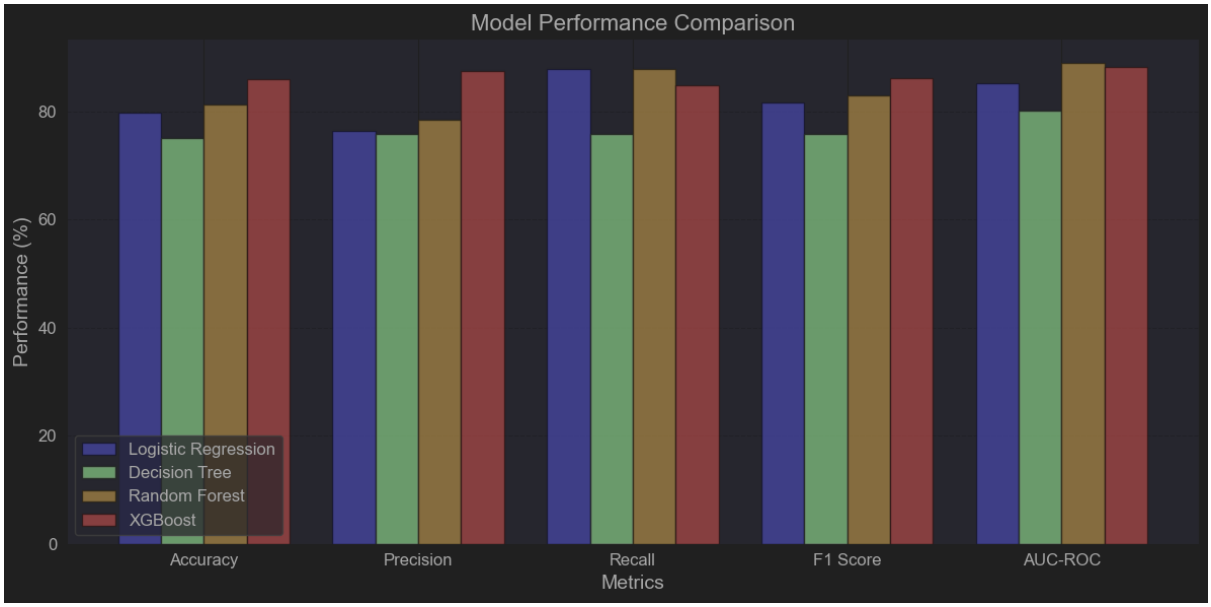
- **Random Forest** and **XGBoost** feature importance plots highlighted consistent top predictors:





7. Conclusion

In this project we performed feature selection using Mutual-Information, Chi-Sq Tests, Mann-Whitney U Tests and Likelihood Ratio Tests and got 28 Features for model building. We implemented various models like Logistic Regression, Decision Trees, Random Forests, and Xgboost with the best overall performance from Xgboost as seen from the diagram below:



We also analysed features and found that C-Reactive Protein, Vitamin-D, Extracellular Fluid, and Hyperlipidaemia are the most significant factors for early detection of Gallbladder stones.

This model was developed to help doctors and medical practitioners a quick and cheap view of the risk of Gallbladder stones for patient prognosis.