Project Report

1. Executive Summary

This project successfully developed and evaluated an explainable machine learning model for predicting gallstone disease risk using a rich, non-imaging dataset. The XGBoost Classifier was identified as the top-performing model, demonstrating strong predictive capabilities. The core deliverable of this project is a comprehensive analysis that explains why the model makes its predictions. The report concludes that the model's decision-making is driven by a clear hierarchy of clinically relevant factors, including Body Mass Index (BMI), hepatic fat accumulation, and age, making it a trustworthy and potentially valuable tool for clinical screening.

2. Methodology & Model Performance

The project followed a robust pipeline from data preparation to model evaluation. A large dataset containing a wide range of demographic, comorbidity, bioimpedance, and laboratory data was used.

Model Selection: Three models—Logistic Regression, Random Forest, and XGBoost—were trained and compared. The XGBoost Classifier was selected for its superior or equivalent performance on the validation set, providing a powerful and robust foundation for explainability.

Final Performance: The chosen model was evaluated on a completely unseen test set to provide an unbiased measure of its real-world performance. The results confirmed its reliability.

Accuracy: 0.781

Recall (avoiding false negatives): 0.781

F1-Score: 0.781

The model’s high recall is a critical metric in this context, as it signifies a low rate of missed gallstone cases, which is crucial for a preliminary screening tool.

3. Model Explainability

This is the most significant contribution of this project. Using SHAP (SHapley Additive exPlanations), we dissected the model's internal logic.

Overall Feature Importance

The SHAP summary plot  below reveals the most influential features. The features are ranked by their overall impact on the model's output. A higher position on the plot and a greater spread of dots indicate higher importance.

The key insights are:

BMI is the single most important factor. High BMI values (red dots) strongly push the prediction towards a positive gallstone diagnosis.

Hepatic Fat Accumulation (HFA) is the second most important feature, with high values indicating a higher risk.

Age remains a critical factor, with older age correlating with increased risk.

Other important factors include Total Cholesterol (TC) and Obesity (%), which are also directly linked to health risks.

These findings align perfectly with established medical knowledge about risk factors for gallstones, validating the model’s reasoning.

Individual Prediction Breakdown

The SHAP waterfall plot  provides a personalized explanation for a single patient's prediction. The plot shows how each feature contributes to the final outcome, starting from the baseline prediction.

For the specific patient shown, the model predicted a very high risk (f(x)=0.94). This was primarily due to:

Their high BMI, which had the most significant positive impact.

Their high Hepatic Fat Accumulation.

Their high Age.

Conversely, their Triglyceride and Low Density Lipoprotein (LDL) levels had a negative influence on the prediction, but this was not enough to counteract the strong positive contributions from BMI, HFA, and age. This level of detail provides a clear, actionable explanation for a clinician.

4. Conclusion & Recommendations

This project has successfully created a robust, high-performing, and, most importantly, explainable model for non-invasive gallstone risk stratification. The model provides valuable, data-driven insights that can assist clinicians in preliminary patient screening and risk prioritization.

The final interactive clinical tool leverages these insights by providing both a risk score and a clear, personalized explanation. This ensures the model is not a "black box" but a transparent and trustworthy aid for medical professionals. Future work could focus on using a larger, more diverse dataset to further validate and generalize the model's findings.