Project Report: Heart Disease Prediction

Fadi Abbara (Coding)
Baraa Alkilany (Presentation)
Anas Zahran (Documentation)
Hayan Azzam (Multinomial regression)

Course: Artificial Intelligence (SS 2025)
Instructor: Dr. Harald Stein

July 10, 2025

1 Problem Statement

The objective of this project is to develop and evaluate machine learning models to predict heart attack risk. Using a patient dataset, we implement and compare three models: **Logistic Regression**, **Random Forest**, and **XGBoost** to classify patients into high-risk and low-risk categories. The goal is to identify the most effective model for this binary classification task based on performance metrics like accuracy and F1-score. Additionally, a **Softmax Regression** model is built from scratch as an add-on to demonstrate multi-class classification.

2 Methodology

2.1 Dataset Exploration

- **What**: We began by loading the dataset and performing an initial exploratory data analysis (EDA).
- **Why**: This was to understand the dataset's structure, check for missing values, analyze feature distributions, and identify correlations between variables before modeling.
- **How**: The dataset was loaded into a `pandas` DataFrame. We used `.info()` to check data types and for null values, `.describe()` for statistical summaries, and `seaborn` to visualize the data distribution and correlation matrix.
- **Result**: The dataset contains **8763 patient records** and **24 features** with no missing values. The distribution of heart attack risk is shown in Figure 1. The correlation matrix (Figure 2) showed generally weak correlations between most features.

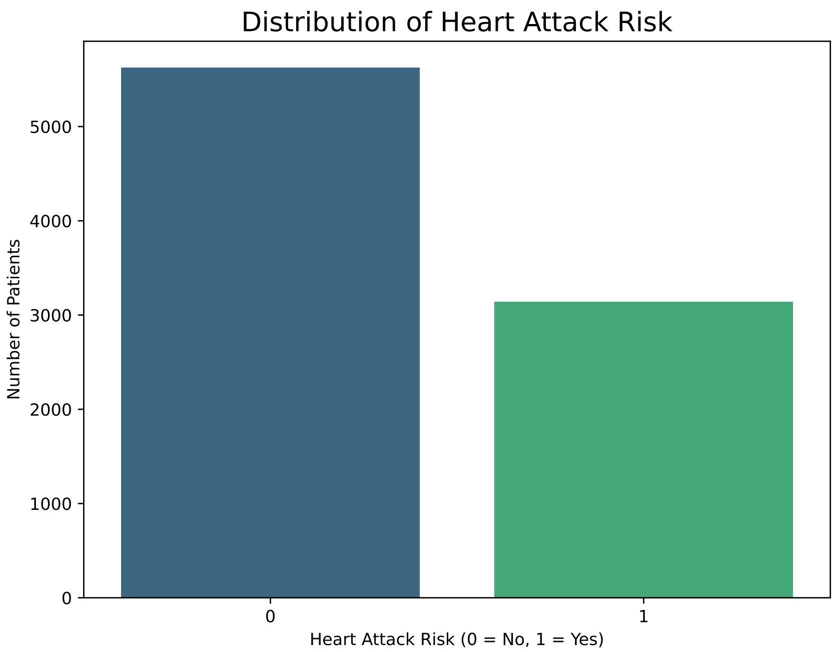

Figure 1: Distribution of Heart Attack Risk.

2.2 Model Training and Evaluation

- **What**: We trained, tuned, and evaluated three primary classification models.
- **Why**: To determine which algorithm performs best for predicting heart attack risk on this dataset.

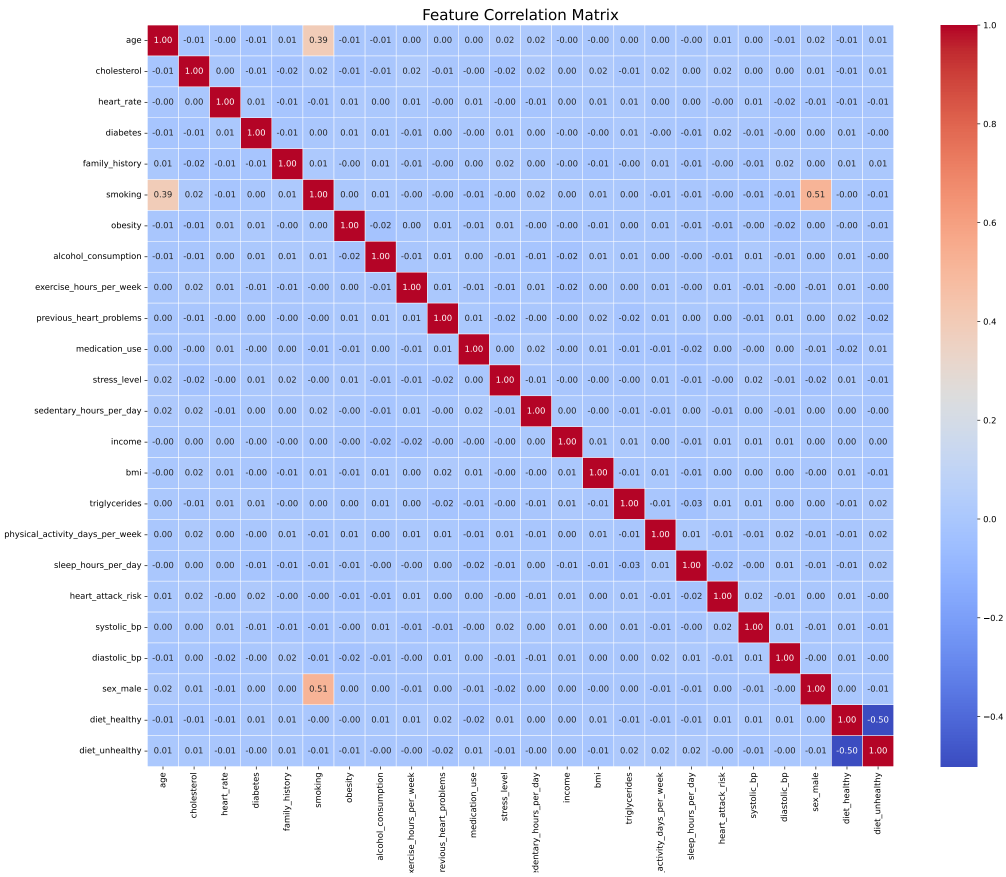

Figure 2: Correlation Matrix of Features.

- **How**: For each model (Logistic Regression, Random Forest, and XGBoost), we trained it on the preprocessed data. For Random Forest and XGBoost, we performed hyperparameter tuning using `GridSearchCV` to find the optimal settings. Model performance was evaluated using accuracy and F1-score, with detailed classification reports and confusion matrices generated for each.
- **Result**: Each model produced distinct performance metrics. The best-performing model was identified based on the weighted F1-score.

3 Results and Evaluation

3.1 Performance Metrics

The performance of the models was evaluated on the test set. The **best overall model was XGBoost**, with the highest weighted F1-score of 0.5514.

Table 1: Model Performance Metrics

Model	**Accuracy**	**F1-Score (Weighted)**	**Key Finding from Report**
Logistic Regression	0.6418	0.5017	Fails to identify any high-risk patients (Recall=0.00).
Random Forest	0.6372	0.5114	Identifies very few high-risk patients correctly (Recall=0.02).
XGBoost	0.6007	0.5514	Best F1-score, but still struggles with high-risk recall (0.17).

3.2 Confusion Matrices

The confusion matrices below visualize the performance of each model.

- **Logistic Regression**: Classifies all patients as "Low Risk," resulting in 1125 correct predictions but 628 false negatives.

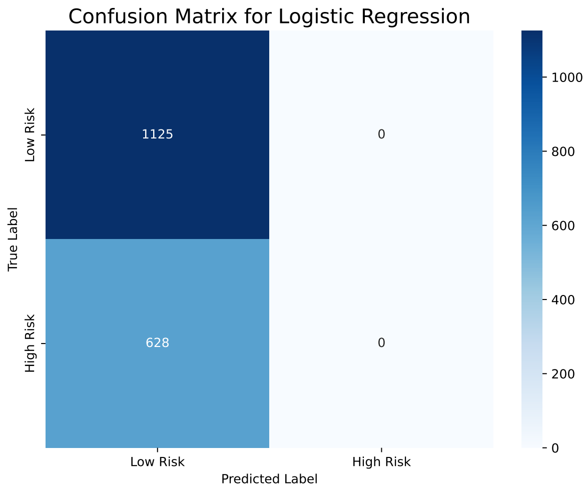

Figure 3: Confusion Matrix for Logistic Regression.

- **Random Forest**: Correctly identifies only 12 high-risk patients while misclassifying 616.

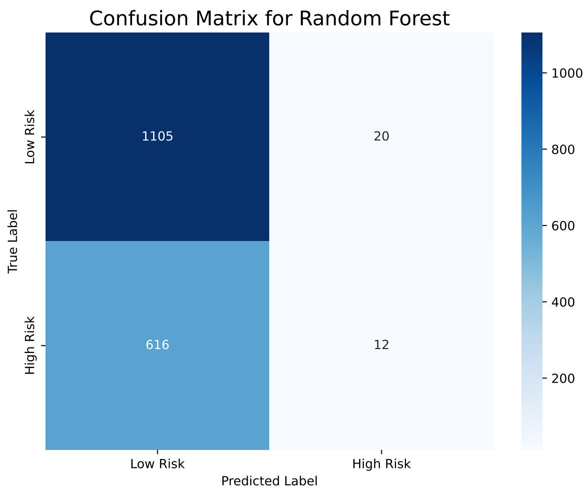

Figure 4: Confusion Matrix for Random Forest.

- **XGBoost**: Correctly identifies 105 high-risk patients but still misses 523.

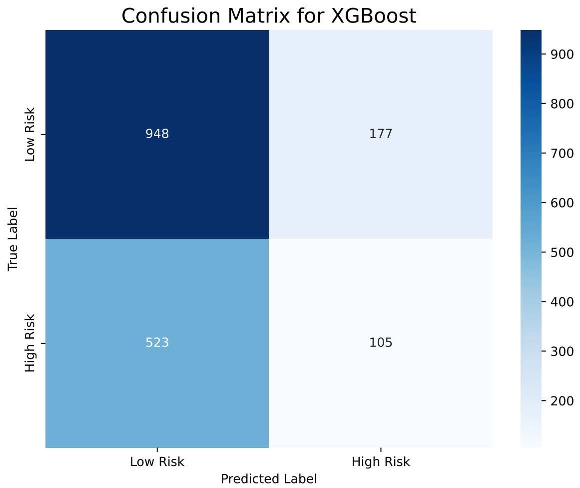

Figure 5: Confusion Matrix for XGBoost.

4 Add-on: Multinomial (Softmax) Regression

As an additional task, a Softmax Regression model was built from scratch to handle multi-class classification.

- **Process**: A synthetic dataset with 1000 samples and 3 classes was generated. The model was trained and evaluated on this data.
- **Performance**: The model achieved very high performance with an **accuracy of 0.9550** and a macro F1-score of 0.9550. The confusion matrix (Figure 6) shows strong performance across all three classes.

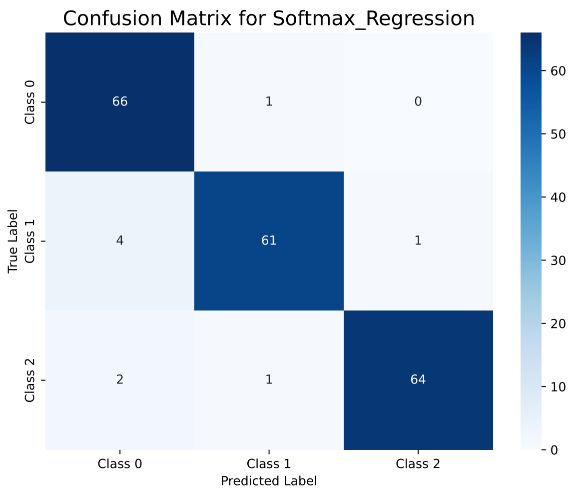

Figure 6: Confusion Matrix for Softmax Regression.

5 Discussion and Conclusion

While we successfully built and evaluated three machine learning models, their performance in predicting heart attack risk on this dataset was poor. The best model, **XGBoost**, achieved an F1-score of only 0.5514 and failed to reliably identify high-risk patients. The Logistic Regression model was completely non-functional for this task, predicting only the majority class.

The models' poor performance suggests issues with the dataset itself; the features may not be sufficiently predictive of heart attack risk. For a real-world medical application, none of these models would be suitable due to the high number of false negatives.

In contrast, the from-scratch **Softmax Regression** model performed exceptionally well on a synthetic multi-class dataset. This highlights the difference between implementing an algorithm and applying it effectively to a real-world problem, where data quality is paramount.