CARDIOVASCULAR RISK PREDICTION

PROBLEM STATEMENT

- Cardiovascular disease is one of the leading causes of death globally. Early detection using machine learning can help identify high-risk individuals and take preventative actions. This project aims to build predictive models using health and lifestyle data to classify whether an individual is at risk for cardiovascular disease.

OBJECTIVES

- Analyze the distribution and relationships between key health features.
- Build classification models to predict cardiovascular disease.
- Evaluate and compare model performance using various metrics.
- Provide data-driven recommendations and highlight potential limitations.

TECHNOLOGIES AND LIBRARIES

- **Python**: Programming language
- **Jupyter Notebook**: For analysis and documentation
- **Pandas / NumPy**: Data manipulation
- **Matplotlib / Seaborn**: Data visualization
- **Scikit-learn**: ML models and metrics
- **XGBoost**: Gradient boosting algorithm
- **LogisticRegressionCV**: Regularized logistic regression with built-in CV

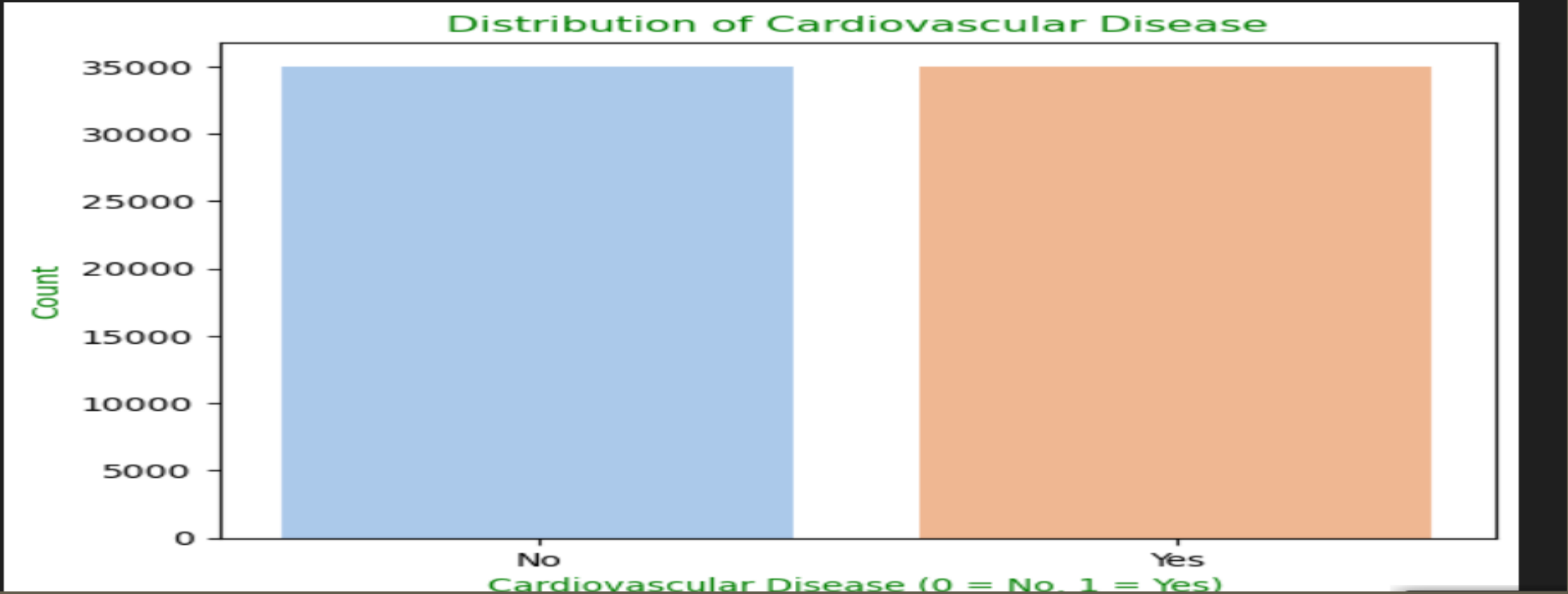
DATASET OVERVIEW

The dataset has been sourced from Kaggle having 70000 records and 12 important features

The dataset contains information on several health-related attributes for individuals, including:

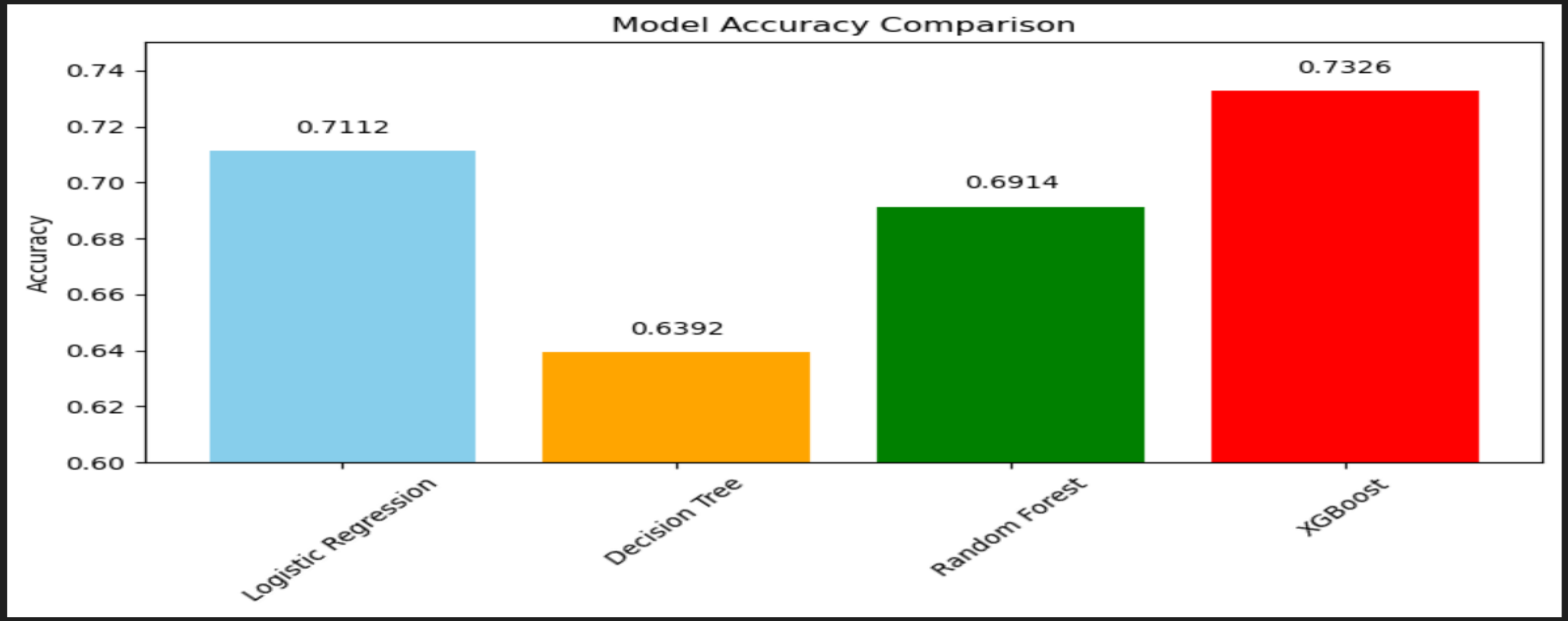
- Presence of cardiovascular disease (target variable)
- Age (in days)
- Diastolic blood pressure
- Systollic blood pressure
- Gender
- Height and Weight
- Body Mass Index (BMI)
- Cholesterol level
- Glucose level
- Smoking status
- Alcohol intake
- Physical activity

TARGET CLASS DISTRIBUTION



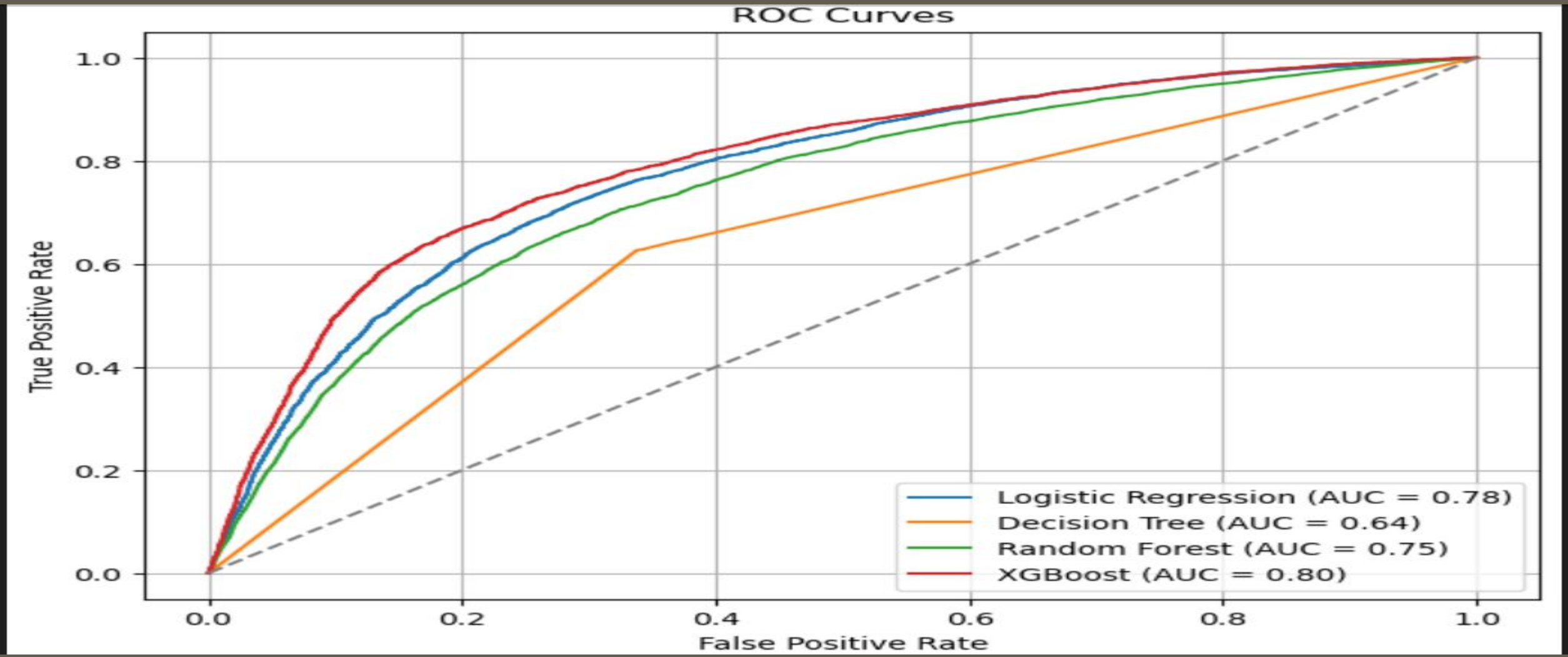
The dataset contains an approximately **equal number of individuals with and without cardiovascular disease**, indicating that the target variable is **balanced**. This is beneficial for training machine learning models, as it reduces the risk of bias toward one class.

ACCURACY COMPARISONS



Based on the cross-validation results, **Logistic Regression** and **XGBoost** emerged as the top-performing models in terms of **accuracy**

AUC COMPARISONS



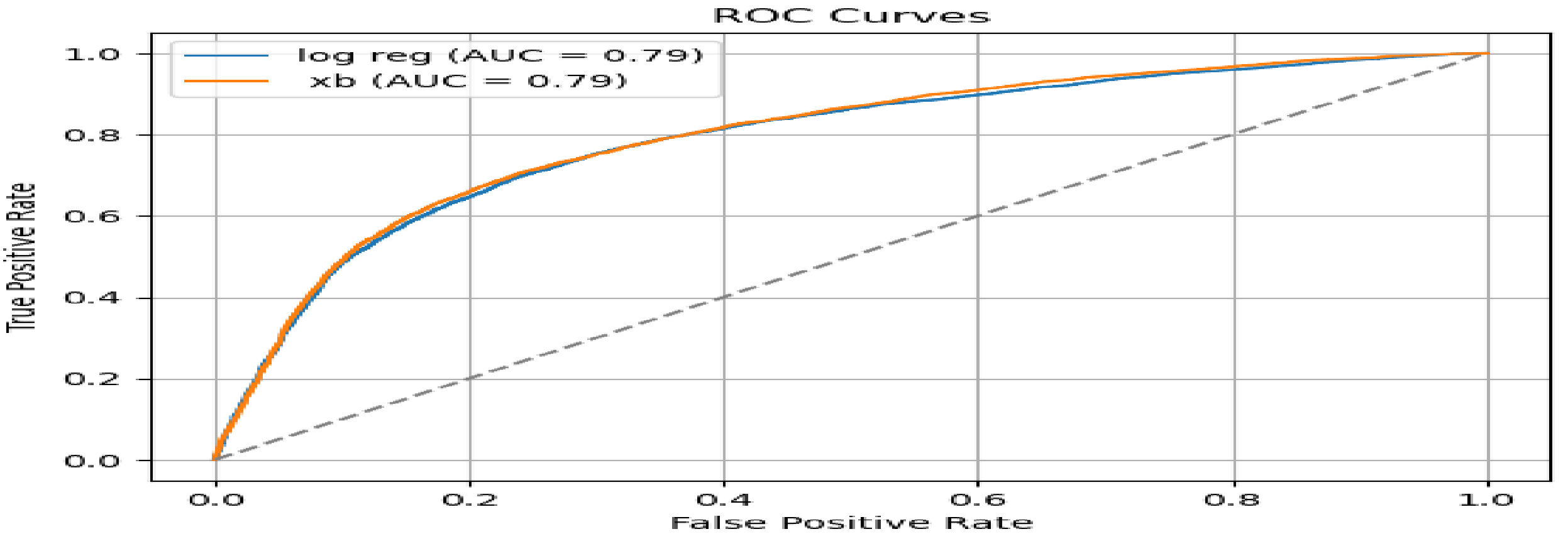
Based on the cross-validation results, **Logistic Regression** and **XGBoost** emerged as the top-performing models in terms of **AUC**

LOGISTIC REG & XGBOOST COMPARISON

Model	Dataset	Accuracy	Class	Precision	Recall	F1-Score	Macro Avg	Weighted Avg
Logistic Regression	Training	0.73	0	0.71	0.78	0.74	0.73	0.73
			1	0.75	0.67	0.71		
	Test	0.73	0	0.71	0.78	0.74	0.73	0.72
			1	0.75	0.67	0.71		
XGBoost	Training	0.76	0	0.74	0.81	0.77	0.76	0.76
			1	0.78	0.72	0.75		
	Test	0.73	0	0.71	0.78	0.75	0.73	0.73
			1	0.75	0.68	0.72		

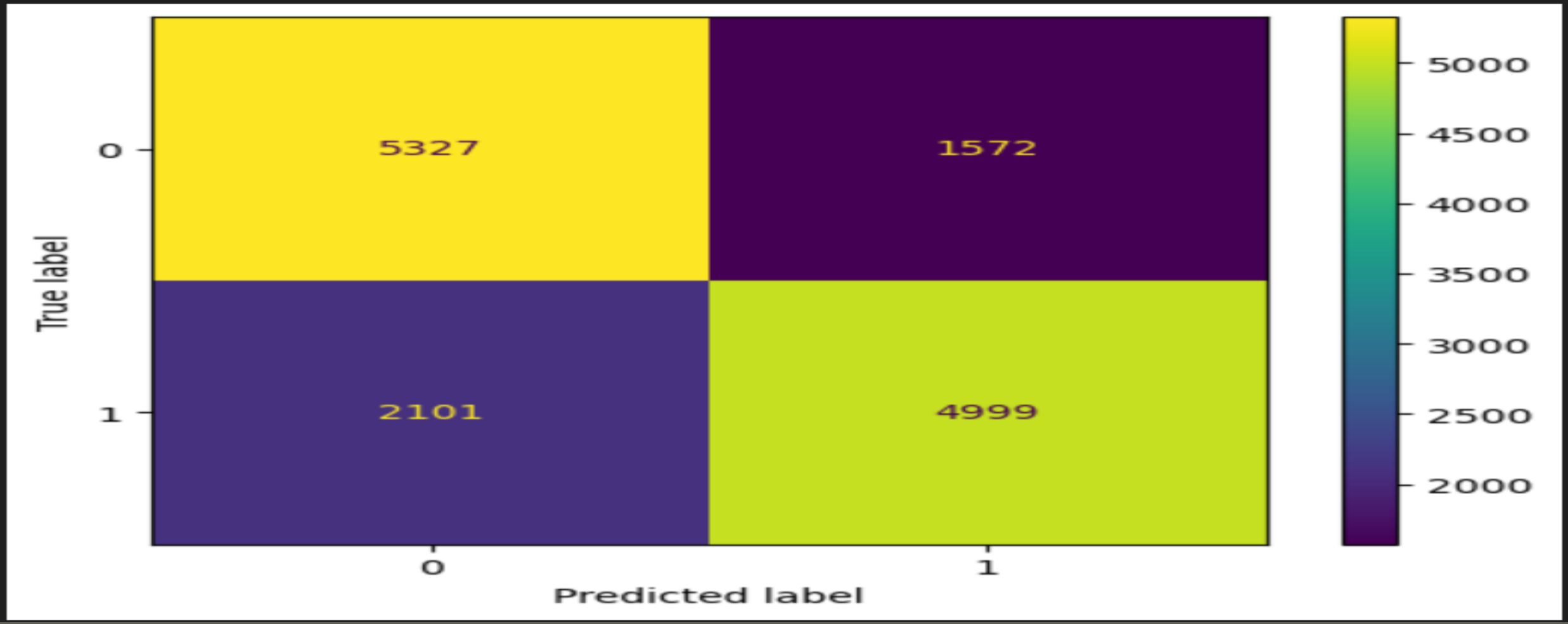
Logistic Regression shows consistent performance across training and test sets, suggesting strong generalization. XGBoost achieves higher training accuracy and class-wise metrics, but its test performance aligns closely with Logistic Regression—indicating potential overfitting. Despite these differences, both models perform comparably on unseen data, with XGBoost offering a slight edge on the training set

LOGISTIC REG VS XGBOOST AUC



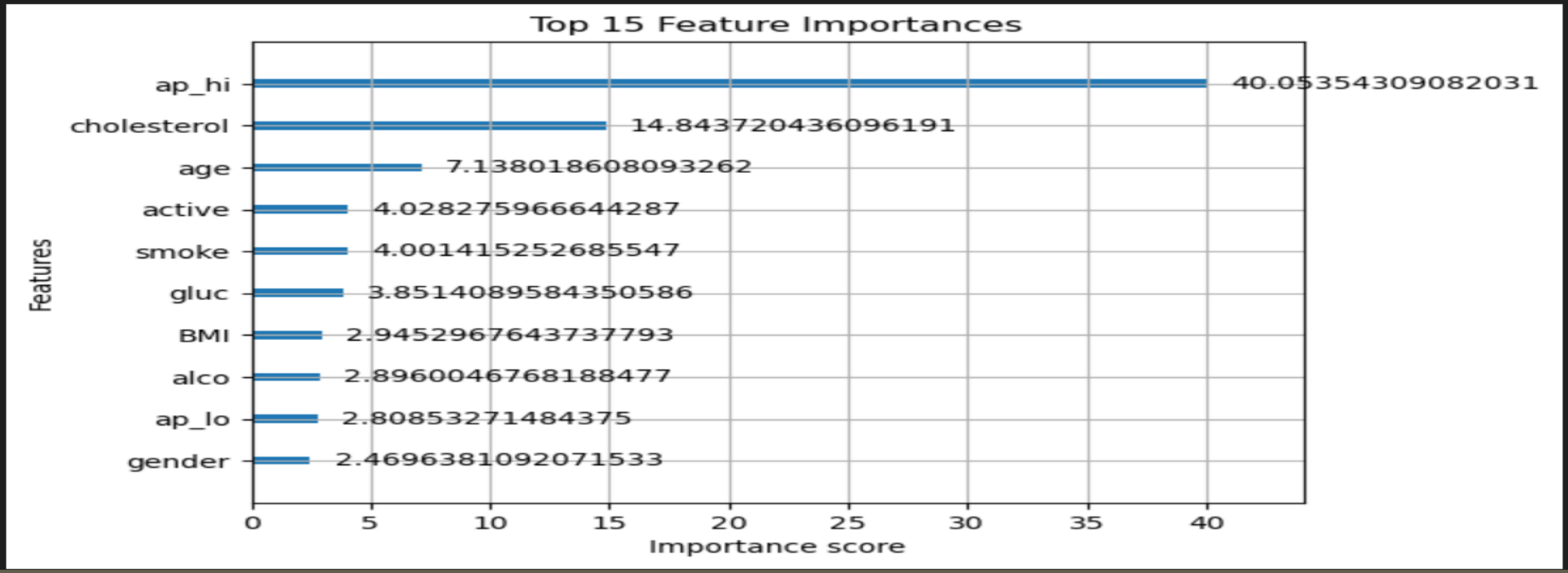
Both models achieve the same AUC, indicating equal overall discriminative power. However, XGBoost's slightly more curved ROC suggests better true positive rates at certain thresholds, offering improved sensitivity in specific cases. Logistic regression provides more uniform performance across all thresholds.

XGBOOST CONFUSION MATRIX



These results show that the model performs reasonably well in detecting both classes. However, the number of false negatives is still notable, meaning some true positive cases are being missed. Overall, the confusion matrix reflects a balanced model, with slightly stronger performance in identifying negative cases.

XGBOOST FEATURE IMPORTANCE



For XGBoost, systolic blood pressure and cholesterol emerge as the most influential features in the predictive task. Age also plays a crucial role in prediction. These features are strongly associated with cardiovascular risk and thus provide the model with high predictive power. However, the remaining features—such as BMI, glucose, and lifestyle indicators—also contribute meaningfully to the model's performance, helping it capture additional variance in the data.

RECOMMENDATIONS

- **Use Ensemble Models for Production**
  XGBoost should be deployed when accuracy is a priority.

- **Consider Logistic Regression for Clinical Use**
  It offers easier interpretation for healthcare professionals.

- **Retrain Periodically**
  Update models as new data becomes available to maintain performance.

- **Collect Additional Data**
  Include features like family history, blood pressure variability, and medication history.

- **Deploy as a Web App**
  Use Flask or Streamlit to provide predictions as a web service.

LIMITATIONS

- **Data Bias**
  More females than males; could affect model fairness.

- **Limited Clinical Variables**
  Dataset lacks some critical risk factors like medical history or genetics.

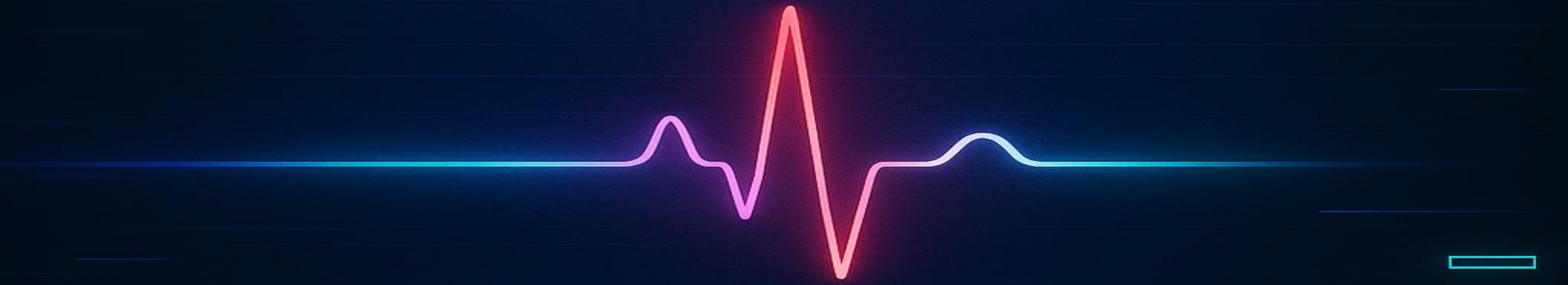
- **Interpretability**
  Complex models (like XGBoost) can be hard to interpret without SHAP/LIME.

- **Static Dataset**
  Not integrated with a real-time data pipeline.

FUTURE WORK

- Visualize model explanations using **SHAP** or **LIME**.
- Deploy the model as an interactive dashboard.
- Add real-time inference or REST API integration.
- Experiment with deep learning models (if more data becomes available).

THANK YOU



ANY QUESTIONS?

kibet8413@gmail.com