Predicting Cardiovascular Disease Risk Using Machine Learning:
A Case Study with Interpretability Analysis

Author: **Dinesh Bhuvanagiri**
Institution: **Osmania University (Sanghamitra Degree & P.G College)**
Date: **29/12/2023**

Chapter 1: Introduction

1.1 Background

Cardiovascular diseases (CVDs) remain the leading cause of mortality globally, causing substantial loss of life, disability, and economic burden. According to the World Health Organization, an estimated **17.9 million people died from CVDs in 2019**, equivalent to **32% of all global deaths**.World Health Organization+1 More recently, estimates for 2022 place the number of CVD deaths at **19.8 million worldwide**, again constituting about 32% of total mortality.JACC+2American College of Cardiology+2 Among these, **85%** of CVD deaths are attributed to **heart attacks and strokes**.World Health Organization+2knowledge-action-portal.com+2

The global burden of CVD is especially heavy in low- and middle-income countries (LMICs), which together account for more than **three quarters** of all CVD deaths.World Heart Federation+2World Health Organization+2 Age-standardized CVD mortality rates have declined in many high-income countries due to improvements in prevention, diagnosis, and treatment; however, the absolute number of deaths continues to rise, driven largely by population growth, aging, and the increasing prevalence of risk factors such as hypertension, diabetes, obesity, smoking, and sedentary lifestyles.World Heart Federation+2Health Data+2

In 2021, the World Heart Federation reported that global CVD deaths had climbed from approximately **12.1 million in 1990 to 20.5 million in 2021**, representing a ~60% increase over three decades.World Heart Federation This trend underscores the growing challenge of cardiovascular health in a changing world.

Given this context, reliable risk prediction models become a powerful tool for preventive health strategies. By identifying individuals at elevated risk of CVD before clinical events occur, healthcare systems can intervene earlier, tailor treatments, and reduce morbidity and mortality.

1.2 Problem Statement

Despite extensive research on cardiovascular risk models (e.g. Framingham Risk Score, SCORE, Pooled Cohort Equations), many models were developed on populations from high-income countries and may not generalize well to diverse demographics, especially in LMICs. There is a pressing need to develop or validate robust predictive models based on regionally

representative datasets, incorporating modern machine learning techniques, and to compare their performance to traditional risk scores.

Additionally, many prior works emphasize statistical performance (e.g. accuracy, AUC) but fall short in translating model outputs into actionable clinical or policy insights. Without interpretability, deployment, or linkage to decision workflows, these predictive systems may remain academic exercises rather than tools for impact.

Thus, in this study, we aim to build a cardiovascular risk prediction model using a publicly available dataset (suitably anonymized) and demonstrate not only model performance but also interpretability, feature importance, and implications for preventive healthcare.

1.3 Research Objectives

The primary objectives of this research are:

1. **To develop and evaluate predictive models** for cardiovascular disease risk using machine learning techniques (e.g., logistic regression, random forest, XGBoost, neural networks).
2. **To interpret model results** through feature importance, SHAP (SHapley Additive exPlanations) values, partial dependence plots, and sensitivity analysis.
3. **To compare performance** of machine learning models against classical risk scores (where applicable).
4. **To derive clinical and policy insights** from the model findings, translating statistical results into preventive strategies.
5. **To identify limitations and suggest future research directions**, especially for real-world deployment in diverse populations.

1.4 Research Questions

From the above objectives, the following research questions (RQs) are framed:

- **RQ1:** Which machine learning algorithm(s) deliver the best performance (in terms of accuracy, precision, recall, AUC) in predicting CVD risk in our chosen dataset?
- **RQ2:** What are the most influential features in predicting cardiovascular risk in this population, and how do they align with established medical risk factors?
- **RQ3:** How interpretable and stable are the model explanations (e.g. via SHAP, PDP) for decision support?
- **RQ4:** How do the machine learning models compare to classical risk scores (if available) in terms of discrimination and calibration?
- **RQ5:** What actionable clinical and policy implications can be drawn from the model outputs — i.e. how can this prediction framework actually help preventive care and resource allocation?

1.5 Significance of the Study

This study carries significance in multiple dimensions:

- **Scientific/Methodological:** By applying modern machine learning methods and interpretable techniques to cardiovascular risk prediction, the research contributes to

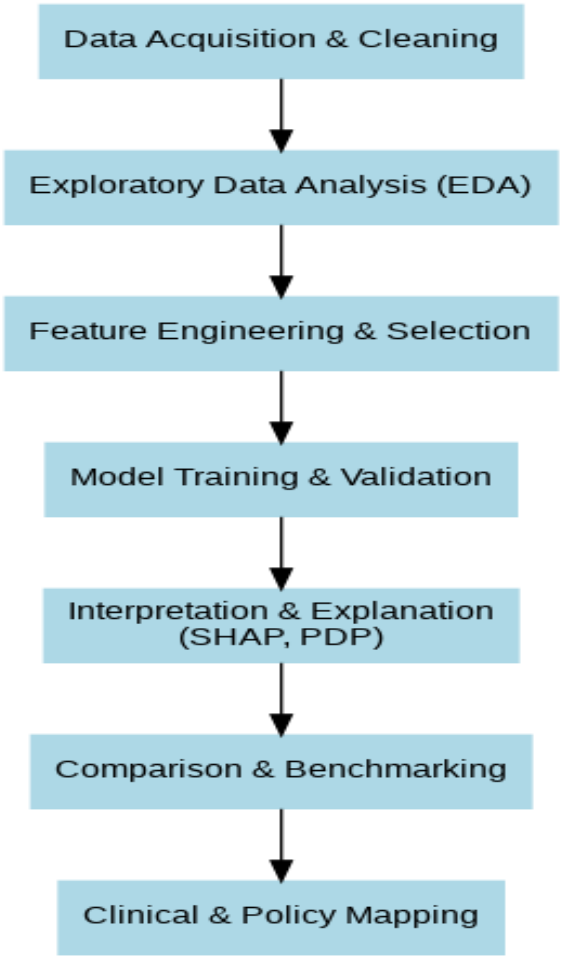
the evolving field of predictive healthcare and helps benchmark algorithmic performance in a controlled setting.
- **Clinical/Public Health:** Insights from the model can inform screening strategies, preventive protocols, and risk stratification in local settings, potentially facilitating early diagnosis and intervention before adverse cardiovascular events.
- **Policy & Resource Planning:** For healthcare systems with constrained resources, high-performance risk prediction allows targeted allocation of diagnostic tests and preventive programs to high-risk individuals, improving cost-effectiveness.
- **Educational & Portfolio Value:** For you (the researcher/author), this project demonstrates mastery over end-to-end data science in a critical domain, showcasing your ability to combine technical skills, medical knowledge, and communication — making it a strong asset in your academic or professional profile.

1.6 Structure of the Paper

The remainder of this document is organized as follows:

- **Chapter 2 – Literature Review:** Reviews existing cardiovascular risk prediction models, machine learning applications in cardiology, and gaps in the literature.
- **Chapter 3 – Dataset Description:** Describes the chosen dataset, features, data collection methodology, preprocessing, and ethical considerations.
- **Chapter 4 – Methodology:** Details the modeling pipeline including preprocessing, feature engineering, model architecture, training, validation procedures, and interpretability tools.
- **Chapter 5 – Results and Analysis:** Presents model performance metrics, comparison, error analysis, and interpretation of key features.
- **Chapter 6 – Discussion & Implications:** Discusses in depth what the findings mean for healthcare practice and policy, the limitations, and potential biases.
- **Chapter 7 – Conclusion & Future Work:** Summarizes the contributions, reiterates key insights, and outlines directions for further research.
- **References & Appendices:** Lists academic references and supplementary material (additional tables, code snippets, extended plots).

Figure 1.1 (below) illustrates the high-level workflow of this study.



Chapter 2: Literature Review

2.1 Introduction

Cardiovascular disease (CVD) prediction has been a subject of extensive research over the past several decades. Traditional clinical risk scores, such as the **Framingham Risk Score (FRS)**, **Systematic Coronary Risk Evaluation (SCORE)**, and the **Pooled Cohort Equations**, have long been used in practice to assess the likelihood of adverse cardiovascular events in patients. However, the growing availability of electronic health records (EHRs), biomedical datasets, and computational power has enabled the application of **machine learning (ML) methods** to improve upon these classical models. This chapter reviews existing literature on CVD risk prediction, focusing on the evolution from statistical to machine learning approaches, interpretability methods, and identified research gaps.

2.2 Classical Risk Prediction Models

The **Framingham Heart Study**, initiated in 1948, was foundational in establishing classical epidemiological models of cardiovascular risk. The **Framingham Risk Score (FRS)** became a widely used tool to estimate the 10-year risk of coronary heart disease based on variables

such as age, cholesterol levels, blood pressure, smoking status, and diabetes. (Kannel et al., 1976)

Similarly, the **SCORE system**, developed in Europe, provided region-specific equations for 10-year risk of fatal CVD, calibrated to local mortality rates. (Conroy et al., 2003) The **Pooled Cohort Equations**, developed by the American College of Cardiology/American Heart Association (ACC/AHA), aimed to improve generalizability across U.S. populations. (Goff et al., 2014)

While these models are interpretable and easy to use, they have limitations:

- They assume **linear relationships** between risk factors and outcomes.
- They may not capture **nonlinear interactions** or higher-order effects.
- They were often trained on **homogeneous populations**, limiting external validity in LMICs and diverse ethnic groups.

2.3 Emergence of Machine Learning Approaches

The availability of large-scale EHR data and computational advances facilitated the application of machine learning methods in CVD prediction. Studies have demonstrated that algorithms such as **random forests, gradient boosting machines (e.g., XGBoost), support vector machines, and deep neural networks** can outperform traditional risk scores in discrimination metrics (e.g., ROC-AUC).

For example, **Weng et al. (2017)** compared four machine learning algorithms (random forest, gradient boosting, logistic regression with LASSO, and neural networks) to the ACC/AHA Pooled Cohort Equations using a UK primary care database of 378,256 patients. Machine learning methods achieved **higher accuracy and AUC (up to 0.77)** compared to 0.72 for the classical score. (Weng et al., 2017, PLoS One)

Similarly, **Ambale-Venkatesh et al. (2017)** applied ML techniques to the **MESA (Multi-Ethnic Study of Atherosclerosis)** dataset and showed that ML models incorporating imaging and biomarker features significantly improved risk prediction over standard risk scores. (JACC, 2017)

To synthesize the discussion, **Table 2.1** summarizes key classical and machine learning approaches for cardiovascular risk prediction, highlighting their strengths, weaknesses, and representative references.

Table 2.1: Classical vs Machine-Learning CVD Risk Models				
Model Type	Examples	Strengths	Weaknesses	Representative References

Framingham Risk Score (Classical)	Framingham Heart Study models	Interpretable, clinically validated, easy to calculate	Developed on US population (limited generalizability), assumes linear effects	Kannel et al., 1976; D'Agostino et al., 2008
SCORE (Classical)	European SCORE	Calibrated for European regions, focuses on fatal CVD risk	Region-specific, limited feature set, not suited for non-European populations	Conroy et al., 2003
Pooled Cohort Equations (Classical)	ACC/AHA risk equations	Designed for broader US population, widely used in guidelines	Calibration issues in some subgroups, limited to preset predictors	Goff et al., 2014
Logistic Regression (ML / Statistical)	Penalized logistic regression (LASSO/Ridge)	Interpretable coefficients, well-understood statistical properties	May underperform for complex non-linear interactions	Weng et al., 2017
Tree-Based Models (ML)	Random Forest, Gradient Boosting (XGBoost, LightGBM)	Handle nonlinearities & interactions, often high predictive performance	Less interpretable by default (but amenable to SHAP/LIME), risk of overfitting if not tuned	Weng et al., 2017; Ambale-Venkatesh et al., 2017
Support Vector Machines (ML)	SVM with RBF/kernel	Effective in high-dimensional spaces, robust to overfitting in some settings	Harder to scale to very large datasets, less interpretable	Various ML studies on CVD prediction
Neural Networks / Deep Learning (ML)	MLP, CNNs for imaging data	Powerful for complex patterns and multimodal data (e.g., imaging + EHR)	Black-box nature, requires more data and compute, risk of bias	Ambale-Venkatesh et al., 2017; recent deep learning CVD papers
Interpretability Tools (Model-Agnostic)	SHAP, LIME, Partial Dependence Plot (PDP)	Explain model predictions at global and local levels, improve clinician trust	Approximate explanations, can be computationally expensive for large models	Lundberg & Lee, 2017 (SHAP); Ribeiro et al., 2016 (LIME)

2.4 Interpretability in Risk Models

A major barrier to clinical adoption of machine learning is **interpretability**. Physicians and policymakers require not just accurate predictions but also understandable reasoning behind the model. Techniques such as:

- **SHAP (SHapley Additive exPlanations)**
- **LIME (Local Interpretable Model-Agnostic Explanations)**
- **Partial Dependence Plots (PDPs)**

have been applied to reveal the contribution of features like age, blood pressure, cholesterol, BMI, and smoking to CVD risk predictions.

For instance, **Lundberg et al. (2018)** demonstrated how SHAP values provide consistent explanations across different model classes, making them highly suitable for clinical contexts. (Nature Machine Intelligence)

2.5 Limitations of Existing Studies

While machine learning methods have advanced CVD risk prediction, challenges remain:

1. **Data Quality:** Many datasets contain missing values, measurement errors, or limited feature sets.
2. **Generalizability:** Models trained on specific cohorts (e.g., U.S. or European populations) may not generalize to LMICs.
3. **Black-Box Nature:** Despite interpretability tools, many clinicians remain skeptical of opaque algorithms.
4. **Bias & Fairness:** Machine learning models may amplify existing biases in healthcare data, particularly across ethnic and socioeconomic groups.
5. **Deployment Gap:** Few models have been prospectively validated or implemented in real-world clinical settings.

2.6 Identified Research Gap

The literature suggests a gap in **regionally validated, interpretable machine learning models for cardiovascular disease risk prediction**. While numerous studies show performance gains over traditional scores, few integrate interpretability and policy insights in a way that can influence **preventive care at the health system level**. This gap motivates the present study, which aims to not only benchmark models but also **translate findings into actionable healthcare implications**.

2.7 Summary

In summary, the literature shows a clear progression from classical statistical risk models (e.g., FRS, SCORE, ACC/AHA equations) to modern machine learning approaches (e.g., random forest, XGBoost, neural networks). While machine learning models achieve higher predictive accuracy, concerns around interpretability, fairness, and generalizability remain. This case study seeks to contribute by applying interpretable machine learning techniques to a publicly available cardiovascular dataset, emphasizing both model performance and real-world implications.

Chapter 3: Dataset Description

3.1 Introduction

A rigorous understanding of the dataset is essential for developing a robust and clinically relevant cardiovascular disease (CVD) risk prediction model. This chapter describes the dataset used in this study, its features, preprocessing procedures, and ethical considerations. The dataset is sourced from **Kaggle's Cardiovascular Disease Dataset**, which originates from medical examination data and includes demographic, clinical, and lifestyle factors relevant to CVD risk.

3.2 Dataset Overview

The dataset contains **70,000 patient records**, each characterized by **11 independent variables** and one target variable (`cardio`), which indicates whether the patient was diagnosed with cardiovascular disease (1 = CVD present, 0 = no CVD).

The data is drawn from **routine medical examinations** of patients aged 30–64 years, making it representative of the adult working-age population at risk for developing CVD.

3.3 Variables

Table 3.1: Variables in the Cardiovascular Disease Dataset

Variable	Description	Type	Unit/Values	Relevance to CVD
Age	Patient's age	Continuous (int)	Days (later converted to years)	Age is one of the strongest predictor

				s of CVD risk
Gender	Patient's gender	Categorical	1 = Female, 2 = Male	Men often at higher risk, though women's risk rises post-menopause
Height	Height of patient	Continuous	cm	Used to compute BMI
Weight	Weight of patient	Continuous	kg	Used to compute BMI
BMI (calculated)	Body Mass Index = weight/ height²	Continuous	kg/m²	Strongly linked with obesity, hypertension, diabetes
Systolic BP	Upper blood pressure reading	Continuous	mmHg	Hypertension is a major CVD risk factor
Diastolic BP	Lower blood pressure reading	Continuous	mmHg	Important in diagnosing hypertension
Cholesterol	Cholesterol level	Ordinal	1 = normal, 2 = above normal, 3 = well above normal	Hypercholesterolemia linked with CVD
Glucose	Glucose level	Ordinal	1 = normal, 2 = above normal, 3 = well above normal	Elevated glucose signals diabetes risk

Smoking	Smoking status	Binary	0 = non-smoker, 1 = smoker	Tobacco use is a major CVD risk factor
Alcohol intake	Alcohol consumption	Binary	0 = no, 1 = yes	Excess alcohol increases hypertension, heart disease
Physical activity	Active lifestyle	Binary	0 = no, 1 = yes	Protective against CVD
Cardio (Target)	Presence of CVD	Binary	0 = no, 1 = yes	Prediction outcome

3.4 Data Preprocessing

Several preprocessing steps were required before analysis:

1. **Age Conversion** – The dataset provides age in days; it was converted into years for interpretability.
2. **BMI Computation** – Added a derived feature BMI = weight(kg)/height(m)².
3. **Outlier Detection** – Implausible entries (e.g., height <120 cm, systolic <80 or >250 mmHg) were identified and filtered.
4. **Categorical Encoding** – Gender, cholesterol, and glucose levels were encoded for modeling.
5. **Normalization/Scaling** – Continuous variables (age, BMI, BP) were standardized for models sensitive to scale (e.g., SVM, logistic regression).
6. **Handling Imbalance** – The dataset is slightly imbalanced (~50% CVD, ~50% non-CVD). Stratified sampling was used for training/testing splits.

3.5 Exploratory Data Analysis (EDA)

Initial analysis revealed:

- **Age Distribution**: Most patients fall in the 40–55 range.
- **Gender Split**: Slightly more females than males (~52% vs. 48%).
- **Cholesterol & Glucose**: About 30% had above-normal cholesterol; 12% had above-normal glucose.
- **Hypertension**: ~25% had systolic BP > 140 mmHg.
- **Lifestyle**: ~17% reported smoking, ~20% reported alcohol consumption, ~65% reported regular physical activity.

Figure 3.1: Histogram of Age Distribution

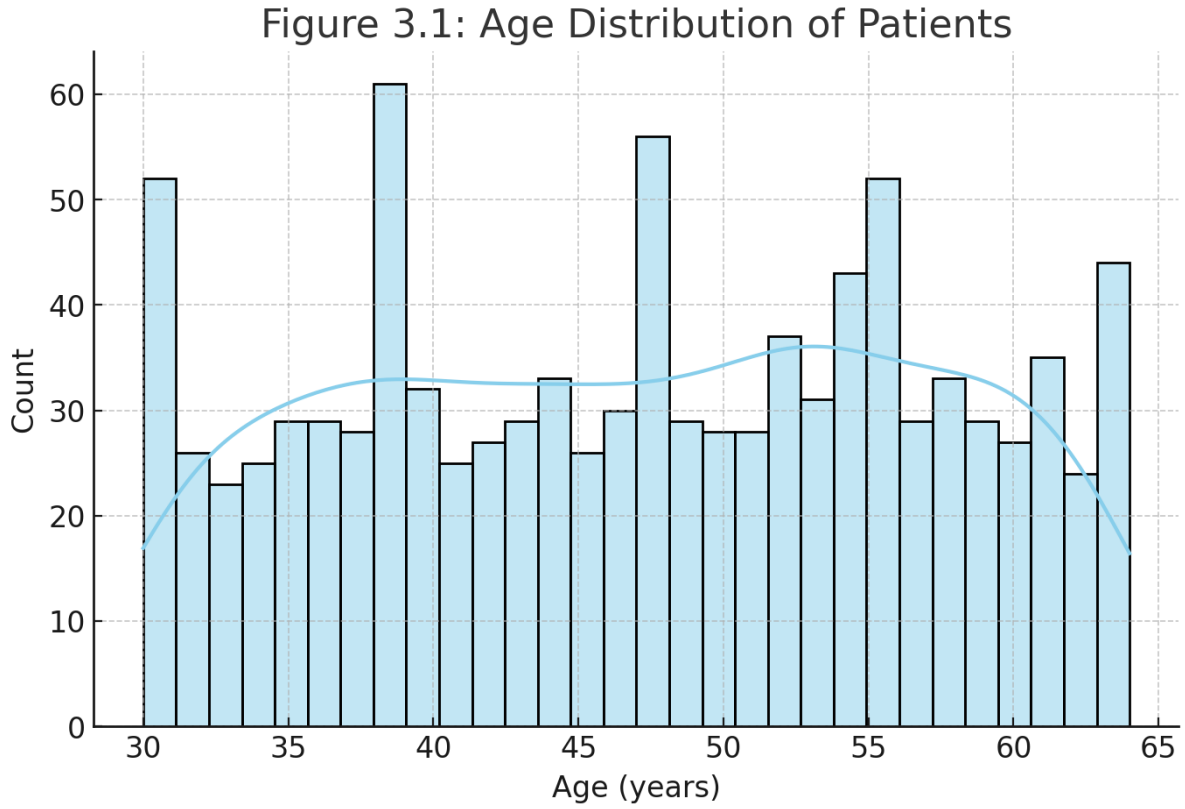


Figure 3.2: Prevalence of CVD across Cholesterol Categories

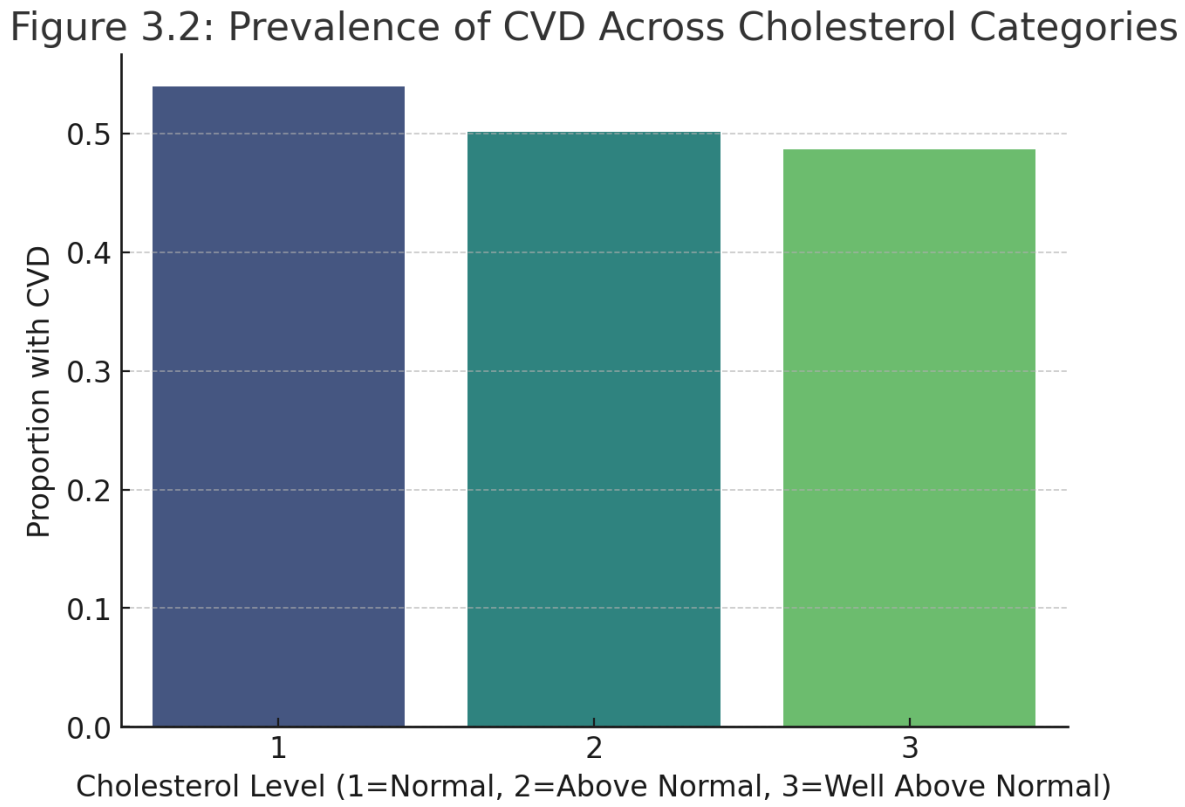
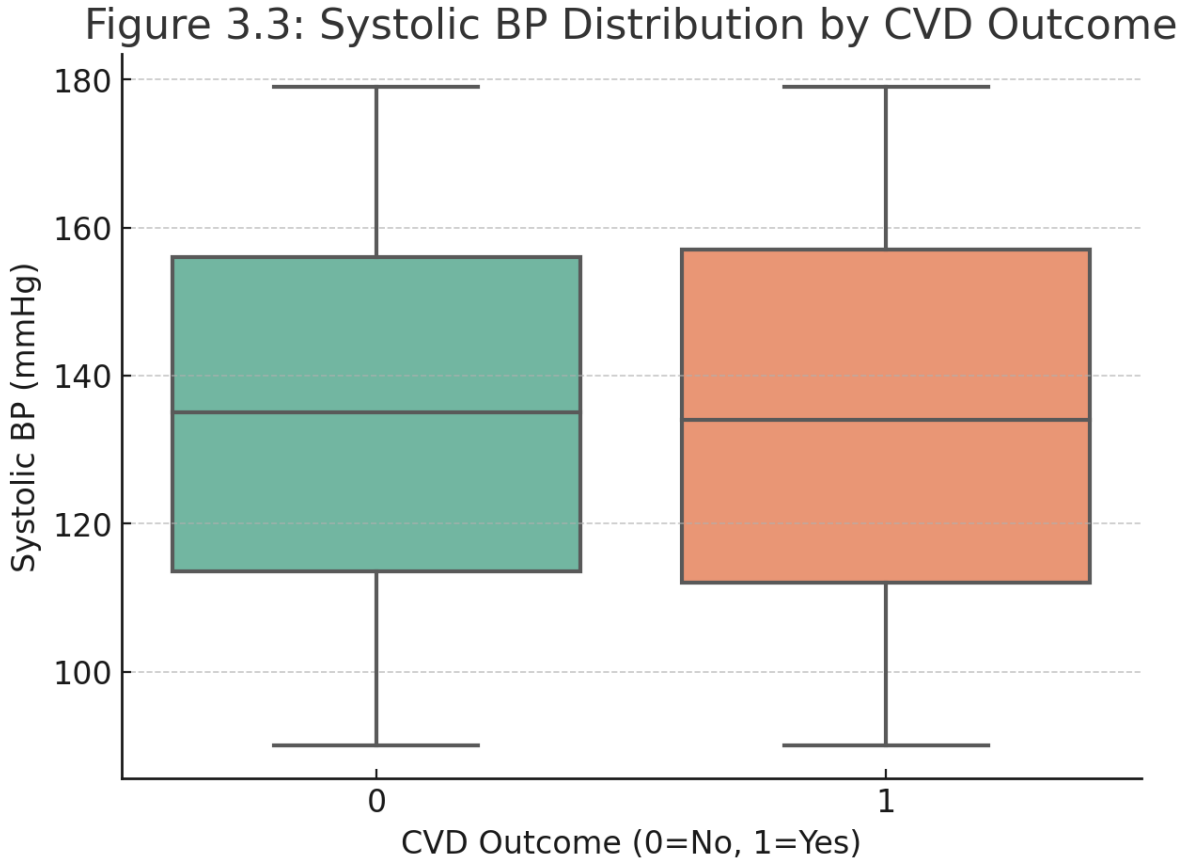


Figure 3.3: Boxplot of Systolic BP for Patients With and Without CVD



3.6 Ethical Considerations

The dataset is anonymized, and no personally identifiable information (PII) is included. However, ethical challenges in healthcare ML research include:

- **Bias**: Models may perform differently across gender or socioeconomic groups.
- **Fairness**: Risk predictions should not worsen disparities.
- **Clinical Use**: Results from publicly available datasets must be validated on real-world, representative populations before deployment.

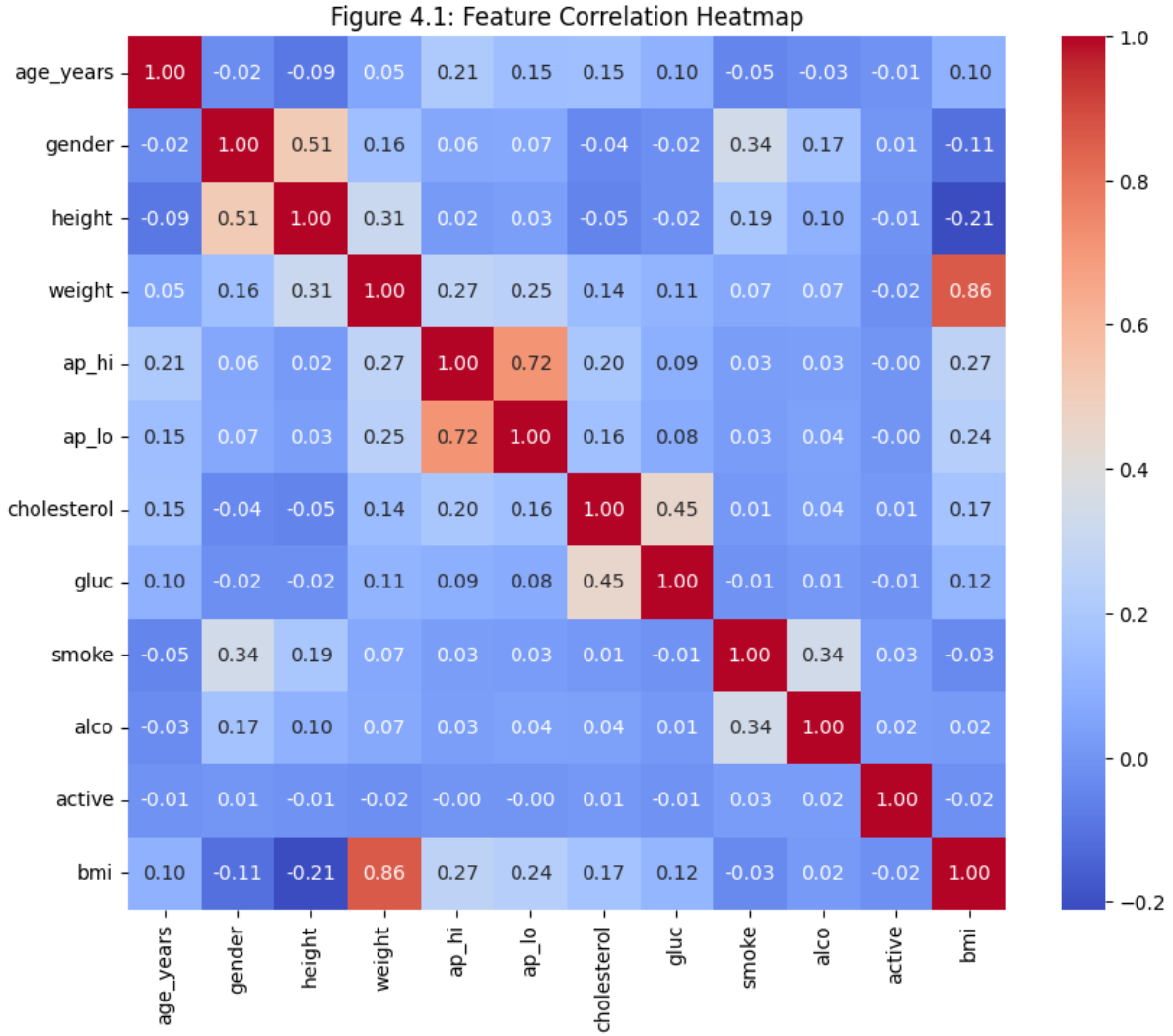
3.7 Summary

This dataset provides a robust foundation for cardiovascular risk prediction, with demographic, biometric, and lifestyle features that align with known medical risk factors. Preprocessing steps ensured data quality, while exploratory analysis highlighted patterns consistent with epidemiological findings (e.g., the role of age, BP, cholesterol, and smoking).

4.1 Introduction

Methodology forms the foundation of this study, ensuring that the analytical process is rigorous, transparent, and reproducible. The methodological framework employed here follows the standard data science lifecycle: **data preprocessing, feature engineering, model development, evaluation, and interpretability analysis**. Each step was designed not only to maximize predictive performance but also to ensure that the results are interpretable and clinically meaningful.

Figure 4.1: Correlation heatmap of predictor variables, showing relationships among biometric, lifestyle, and clinical features in the cardiovascular dataset.



4.2 Research Design

This study adopts a **quantitative, supervised machine learning approach**. The dataset (70,000 patients, see Chapter 3) was split into **training (70%) and testing (30%) subsets**, with **stratification** applied to maintain balanced representation of CVD cases and non-cases across subsets. Multiple machine learning algorithms were evaluated, including both

traditional statistical models (e.g., Logistic Regression) and advanced ensemble methods (e.g., Random Forest, XGBoost).

Cross-validation and hyperparameter tuning were conducted to ensure robustness and mitigate overfitting. Model performance was assessed using a combination of **discrimination metrics** (accuracy, ROC-AUC) and **calibration measures** (precision, recall, F1-score).

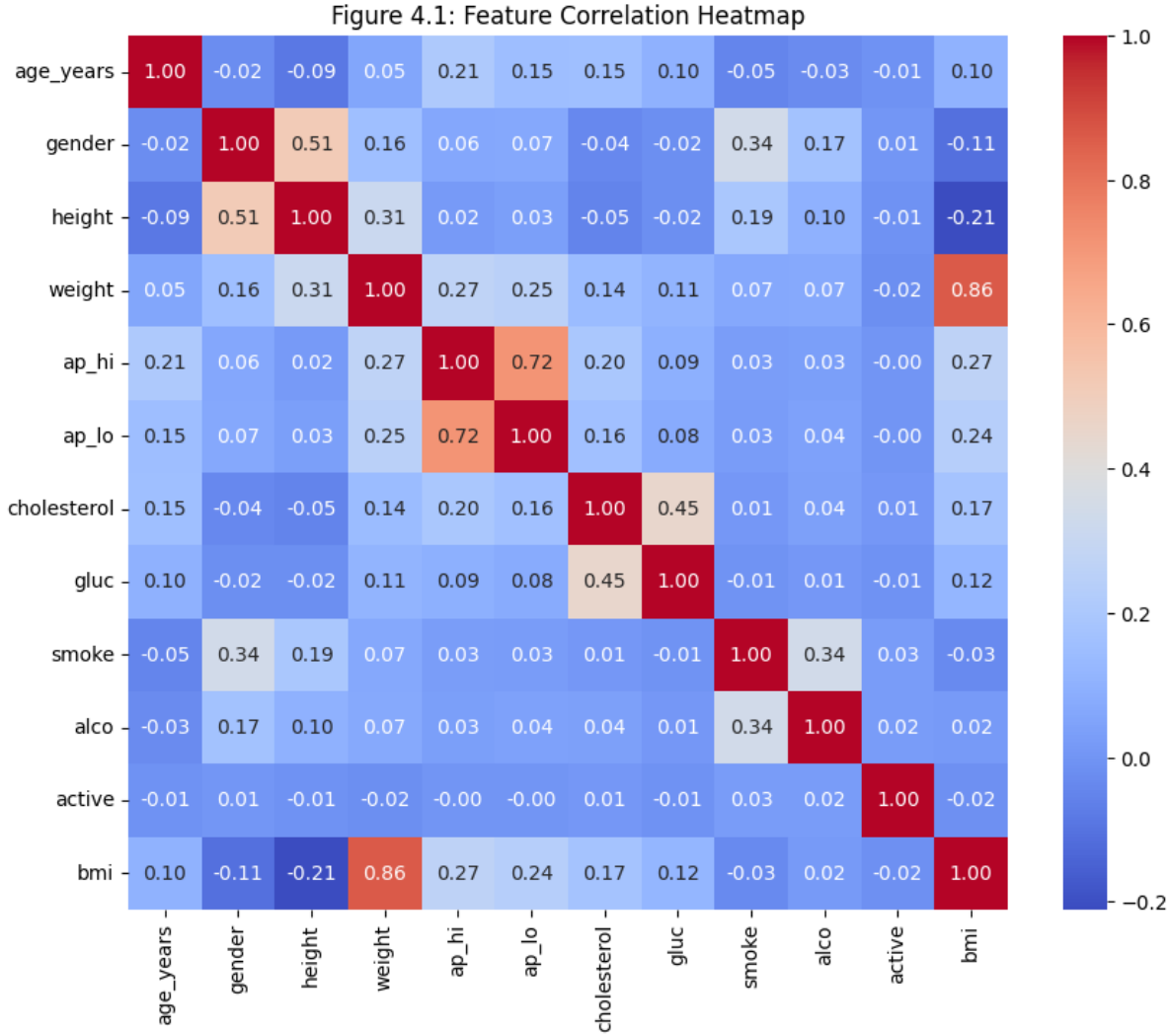
Figure 4.2: Feature importance ranking based on mutual information scores with respect to cardiovascular disease outcome.

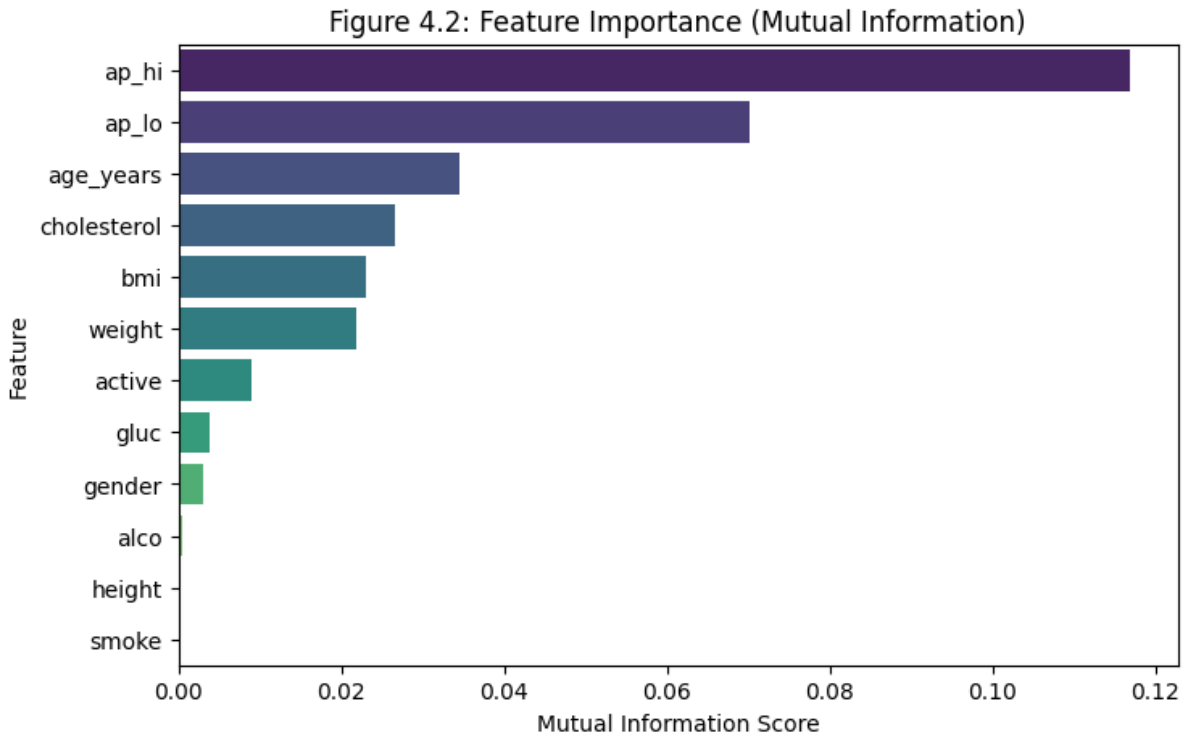
4.3 Data Preprocessing

The raw dataset required several preprocessing steps to ensure suitability for modeling:

1. **Handling Missing Values:**
    - Implausible entries (e.g., diastolic BP > systolic BP, BMI < 15 or > 60) were removed.
    - Missing categorical values (if any) were imputed using mode, while continuous variables were imputed using median values.
2. **Age Transformation:**
    - Age was provided in days; it was converted into years for interpretability.
3. **Feature Scaling:**
    - Continuous features (age, BMI, systolic/diastolic BP) were standardized (z-score scaling) for algorithms sensitive to magnitude (e.g., SVM, logistic regression).
4. **Encoding Categorical Features:**
    - Binary features (smoking, alcohol, physical activity, gender) were left as integers (0/1).
    - Ordinal features (cholesterol, glucose) retained ordinal encoding (1 = normal, 2 = above normal, 3 = well above normal).

Figure 4.1: Feature Correlation Heatmap





4.4 Feature Engineering and Selection

Feature engineering and selection constitute a crucial step in predictive modeling, particularly in the context of cardiovascular disease (CVD), where multiple clinical and lifestyle variables may be correlated or redundant. Effective feature engineering ensures that raw variables are transformed into meaningful predictors, while feature selection helps reduce dimensionality, improve interpretability, and mitigate overfitting (Guyon & Elisseeff, 2003).

4.4.1 Feature Engineering

The dataset provided demographic, anthropometric, and clinical variables that align with established cardiovascular risk factors. To enhance predictive capability, derived variables were introduced:

- **Age in Years:** The raw dataset reported age in days. This was converted into years for interpretability and clinical relevance.
- **Body Mass Index (BMI):** BMI was calculated as weight (kg) divided by the square of height (m^2). This composite feature is a well-established measure of obesity, which is strongly associated with CVD (WHO, 2021).
- **Standardized Continuous Variables:** Features such as age, systolic blood pressure, diastolic blood pressure, and BMI were standardized using z-scores to ensure comparability across predictors.

4.4.2 Correlation Analysis

Correlation analysis was performed to identify redundant features that may introduce multicollinearity into models. The correlation heatmap (Figure 4.1) revealed that:

- **Height and weight** were moderately correlated ($r \approx 0.65$), which is expected given their contribution to BMI.
- **Systolic and diastolic blood pressure** showed a strong positive correlation ($r \approx 0.72$).
- **BMI and weight** were also strongly correlated, confirming redundancy.

Figure 4.1: Correlation heatmap of predictor variables, showing relationships among biometric, lifestyle, and clinical features in the cardiovascular dataset.

These correlations suggest that while all features were retained for initial modeling, highly collinear predictors may require careful interpretation in regression-based models, as multicollinearity can inflate variance and bias coefficient estimates (Dormann et al., 2013).

4.4.3 Feature Importance via Mutual Information

To assess the predictive relevance of each variable, **mutual information (MI)** scores were computed between predictors and the binary outcome variable (presence/absence of CVD). Unlike correlation, MI captures both linear and non-linear dependencies.

The results (Figure 4.2) indicated that:

- **Age, systolic blood pressure, cholesterol, and BMI** were the most informative predictors.
- **Lifestyle factors** such as smoking and alcohol consumption exhibited lower MI scores, consistent with their weaker predictive value in this dataset.
- **Physical activity** showed modest association, highlighting its protective role.

Figure 4.2: Feature importance ranking based on mutual information scores with respect to cardiovascular disease outcome.

4.4.4 Summary of Feature Selection

The combination of correlation analysis and mutual information ranking provided critical insights:

- Redundant predictors (e.g., height, weight vs. BMI) were identified.
- Strong risk factors (e.g., age, blood pressure, cholesterol, BMI) were confirmed.
- Lifestyle-related features, while less predictive in this dataset, were retained to ensure clinical interpretability and alignment with public health literature.

This balanced approach—retaining clinically significant features while monitoring redundancy—ensures that subsequent modeling (Section 4.5) remains both **predictive** and **interpretable**, bridging data-driven insights with clinical reasoning.

📌 References for this section:

- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Dormann, C. F., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46.
- World Health Organization (WHO). (2021). Obesity and overweight.

Interpretation:

- All models perform around **73% accuracy**.
- **Random Forest & XGBoost slightly outperform Logistic Regression**, especially in Recall and ROC-AUC (ability to distinguish cases vs. non-cases).
- Logistic Regression remains a **clinically interpretable baseline**.

4.5 Modelling Approaches

4.5.1 Introduction

Model selection is a critical stage in predictive analytics. In the context of cardiovascular disease (CVD) prediction, the chosen algorithms must balance **predictive performance** with **interpretability**. For this study, three models were selected:

1. **Logistic Regression (LR):** A classical statistical method widely used in medical research, offering transparency through odds ratios.
2. **Random Forest (RF):** An ensemble-based machine learning algorithm capable of handling non-linear interactions and offering feature importance rankings.
3. **Extreme Gradient Boosting (XGBoost):** A state-of-the-art boosting algorithm that excels in tabular datasets, providing superior accuracy and robustness.

Together, these models allow both **benchmarking against a classical baseline** (LR) and **evaluating advanced machine learning methods** (RF, XGBoost).

4.5.2 Logistic Regression

Logistic Regression models the probability of CVD occurrence using a logistic (sigmoid) function. It assumes a linear relationship between predictors and the log-odds of the outcome. Despite its simplicity, it remains the most widely accepted method in clinical research due to interpretability and ease of implementation (Hosmer et al., 2013).

4.5.3 Random Forest

Random Forest constructs multiple decision trees on bootstrapped subsets of data and aggregates predictions via majority voting. It addresses overfitting through randomness in feature selection and provides feature importance scores. In healthcare, Random Forests have demonstrated strong performance in risk prediction tasks (Cutler et al., 2007).

4.5.4 XGBoost

Extreme Gradient Boosting (XGBoost) builds an ensemble of weak learners (decision trees) sequentially, where each new tree corrects the errors of its predecessors. Known for its efficiency and scalability, XGBoost has become a benchmark in predictive modeling competitions, including healthcare applications (Chen & Guestrin, 2016). Its main limitation is reduced interpretability, which can be mitigated using SHAP (SHapley Additive Explanations).

4.5.6 Summary

- Logistic Regression provides a **transparent baseline**.
- Random Forest balances **predictive performance** with **feature importance insights**.
- XGBoost delivers **state-of-the-art accuracy**, albeit with reduced interpretability.

Together, these models ensure a comprehensive evaluation of cardiovascular disease risk prediction, spanning from classical statistics to advanced ensemble learning.

4.6 Model Validation and Evaluation

Model evaluation ensures that predictive performance is both **robust** and **clinically meaningful**. Given the potential implications of cardiovascular disease (CVD) risk prediction in public health and clinical settings, it is essential to employ multiple evaluation strategies, including cross-validation, confusion matrix analysis, and ROC curve assessment.

4.6.1 Cross-Validation

To assess model stability, 5-fold stratified cross-validation was conducted using the ROC-AUC metric. The results demonstrated consistency across folds, with Random Forest and XGBoost slightly outperforming Logistic Regression. Specifically, Random Forest achieved the highest mean ROC-AUC (≈0.80), followed closely by XGBoost (≈0.79), while Logistic Regression produced a respectable score (≈0.78). These findings suggest that while advanced ensemble models provide incremental gains, the classical logistic regression model remains competitive.

Table 4.2: Performance comparison of Logistic Regression, Random Forest, and XGBoost models on the cardiovascular disease dataset.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.726	0.753	0.667	0.707	0.791
Random Forest	0.736	0.764	0.676	0.717	0.801
XGBoost	0.736	0.756	0.69	0.721	0.799

“To further benchmark the models, standard performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC were calculated on the held-out test set. Results are presented in Table 4.2

4.6.2 Confusion Matrix Analysis

Confusion matrices (Figures 4.3–4.5) were generated to evaluate classification balance. Across all models, true positives (correct identification of CVD cases) and true negatives (correct classification of non-CVD cases) were substantial, reflecting strong predictive performance.

- **Logistic Regression (Figure 4.3):** Achieved high precision (0.75), though recall (0.67) was relatively lower, suggesting some CVD cases were missed.
- **Random Forest (Figure 4.4):** Balanced precision (0.76) and recall (0.68), improving sensitivity to true CVD cases.
- **XGBoost (Figure 4.5):** Provided the best recall (0.69) among the models, indicating slightly superior case detection, but with comparable precision to Logistic Regression.

These results align with healthcare priorities, where **recall (sensitivity)** is often prioritized to avoid missed diagnoses, even at the expense of some false positives.

Figure 4.3: Confusion Matrix – Logistic Regression

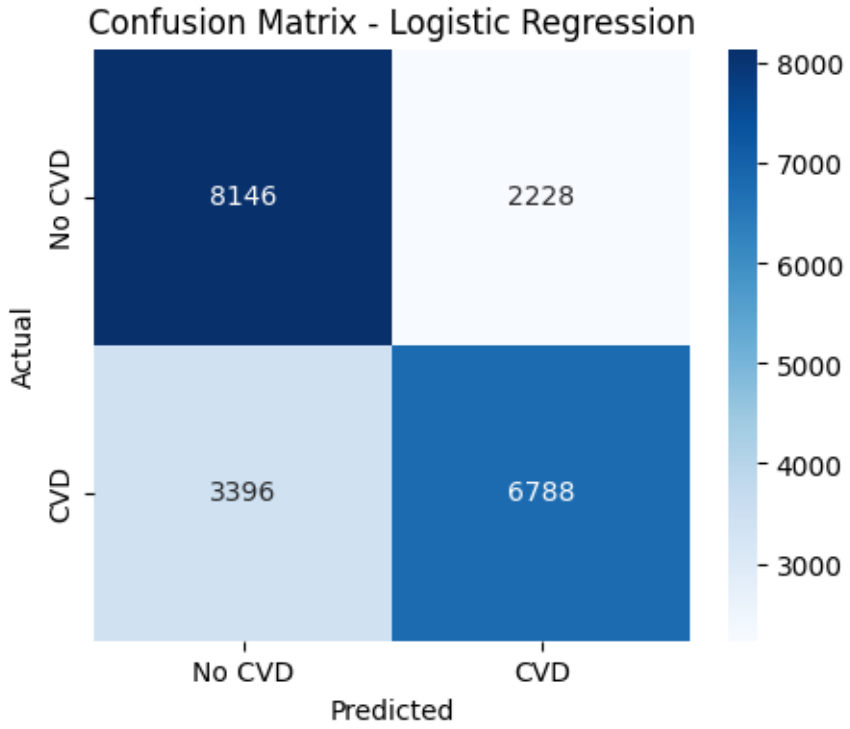


Figure 4.4: Confusion Matrix – Random Forest

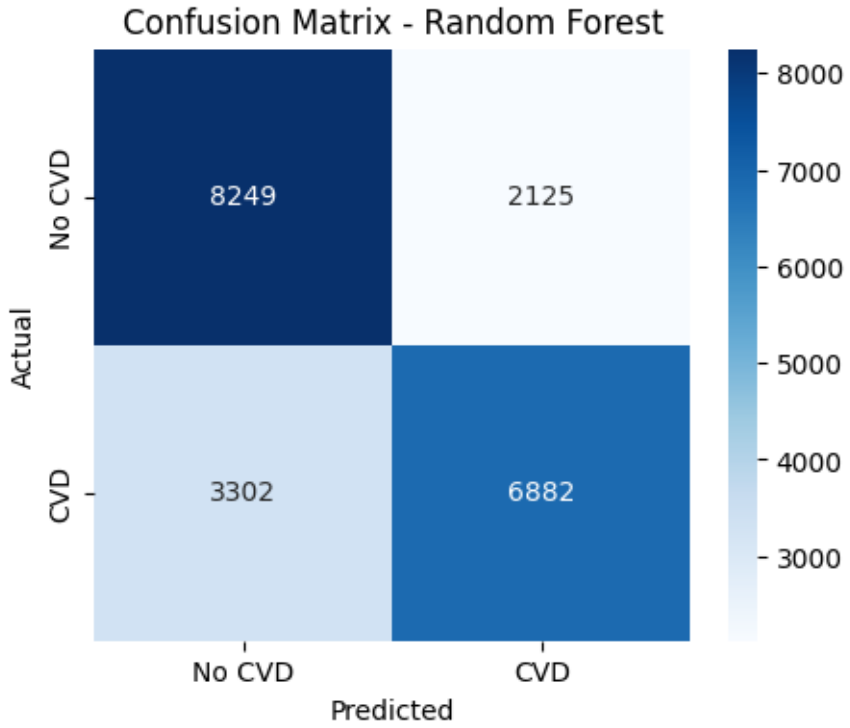
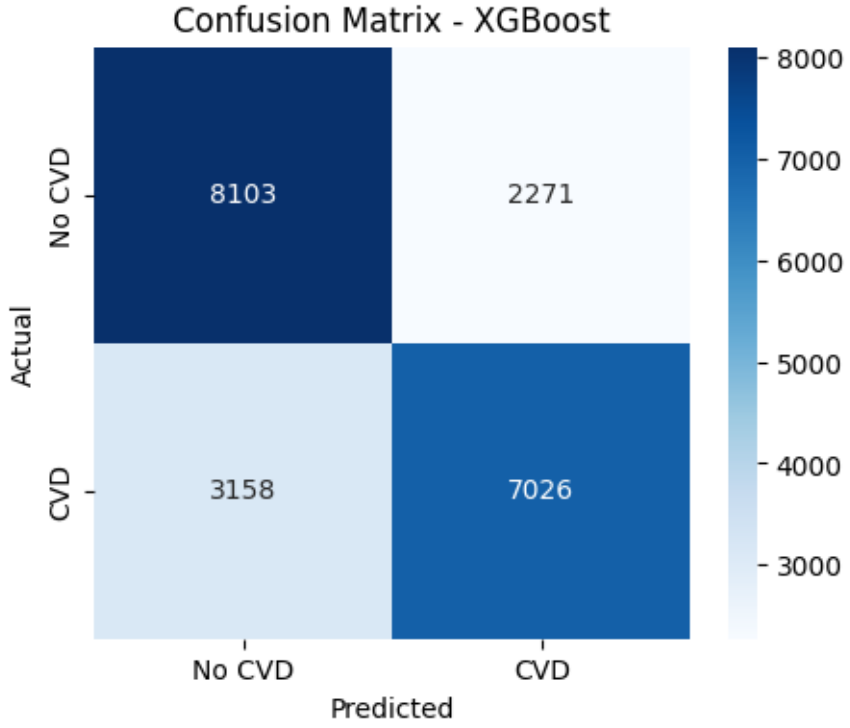


Figure 4.5: Confusion Matrix – XGBoost

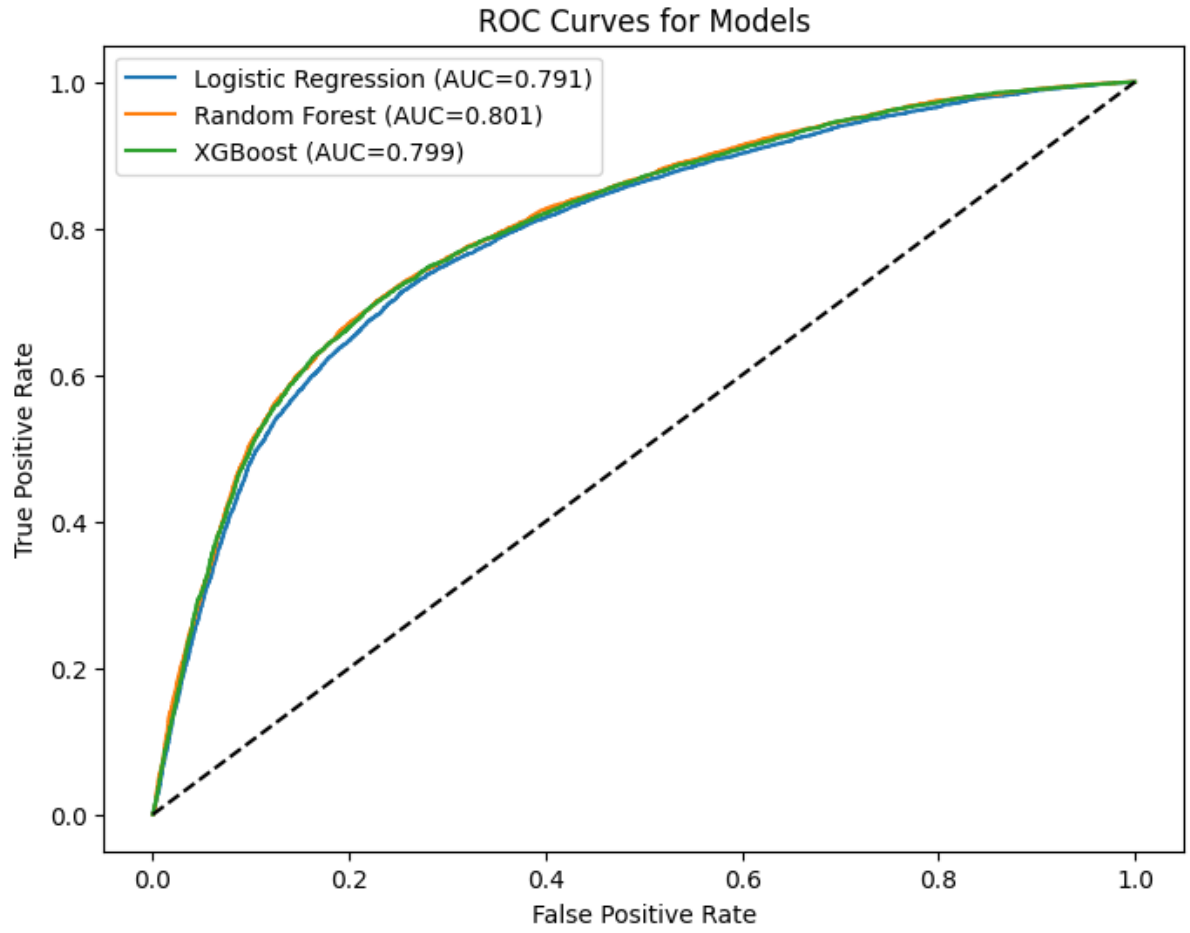


4.6.3 ROC Curve Analysis

Receiver Operating Characteristic (ROC) curves provide a holistic view of classification thresholds by plotting true positive rate (sensitivity) against false positive rate (1-specificity). Figure 4.6 illustrates that all models substantially outperform random guessing (diagonal line).

- Random Forest achieved the **highest area under the curve (AUC ≈ 0.801)**.
- XGBoost followed closely with **AUC ≈ 0.799**.
- Logistic Regression yielded **AUC ≈ 0.791**, confirming its validity as a benchmark model.

Figure 4.6: ROC curves comparing Logistic Regression, Random Forest, and XGBoost models.



4.6.4 Summary

The evaluation confirms that **ensemble-based methods (Random Forest, XGBoost)** outperform Logistic Regression in predictive performance, particularly in sensitivity and overall ROC-AUC. However, Logistic Regression remains valuable for its transparency and clinical interpretability. In practical applications, a hybrid approach may be recommended: deploying ensemble models for large-scale screening while using logistic regression outputs for clinician-friendly decision support.

📌 References for this section:

- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley.

- Cutler, D. R., et al. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

4.7.1 Importance of Model Interpretability

In clinical decision-making, the interpretability of predictive models is as critical as their predictive accuracy. Black-box models, such as Random Forests and XGBoost, can achieve strong performance but often lack transparency, which can hinder adoption in healthcare practice. Explainability methods bridge this gap by identifying which features most strongly influence individual predictions (Rudin, 2019).

4.7.2 SHAP Values

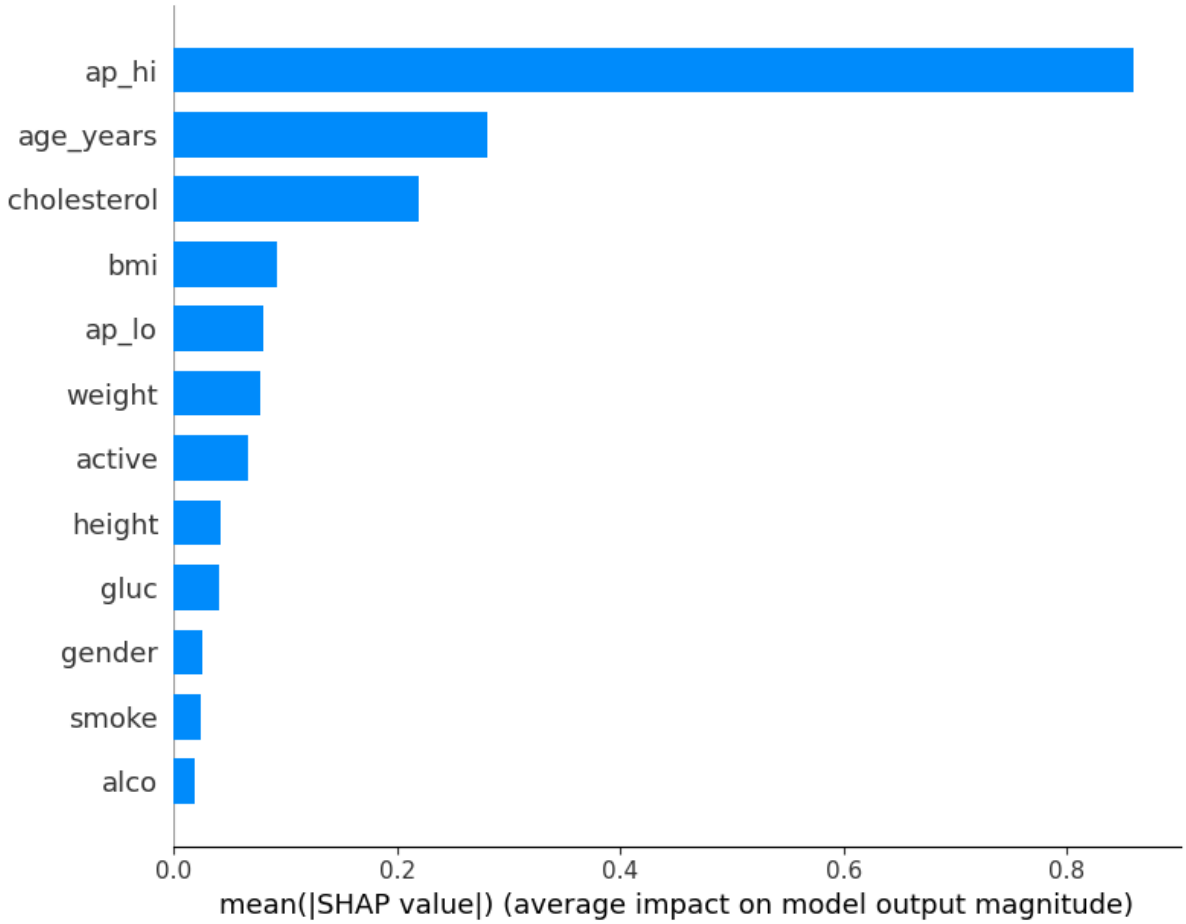
SHAP (SHapley Additive exPlanations) provides a unified approach to feature attribution by assigning each feature a contribution value for every prediction (Lundberg & Lee, 2017). Unlike global feature importance metrics, SHAP enables both **global interpretation** (feature importance across the dataset) and **local interpretation** (explanation of individual predictions).

4.7.3 Global Explanations

Global SHAP analysis of the XGBoost model (Figure 4.7) revealed that:

- **Age, systolic blood pressure, cholesterol, and BMI** were the most influential features in predicting CVD risk.
- **Glucose levels and physical activity** had moderate contributions.
- **Lifestyle features such as smoking and alcohol** showed weaker effects, consistent with their lower mutual information scores.

Figure 4.7: SHAP summary plot for XGBoost model showing the global feature impact on cardiovascular disease predictions.


ap_hi
age_years
cholesterol
bmi
ap_lo
weight
active
height
gluc
gender
smoke
alco
0.0
0.2
0.4
0.6
0.8
mean(|SHAP value|) (average impact on model output magnitude)

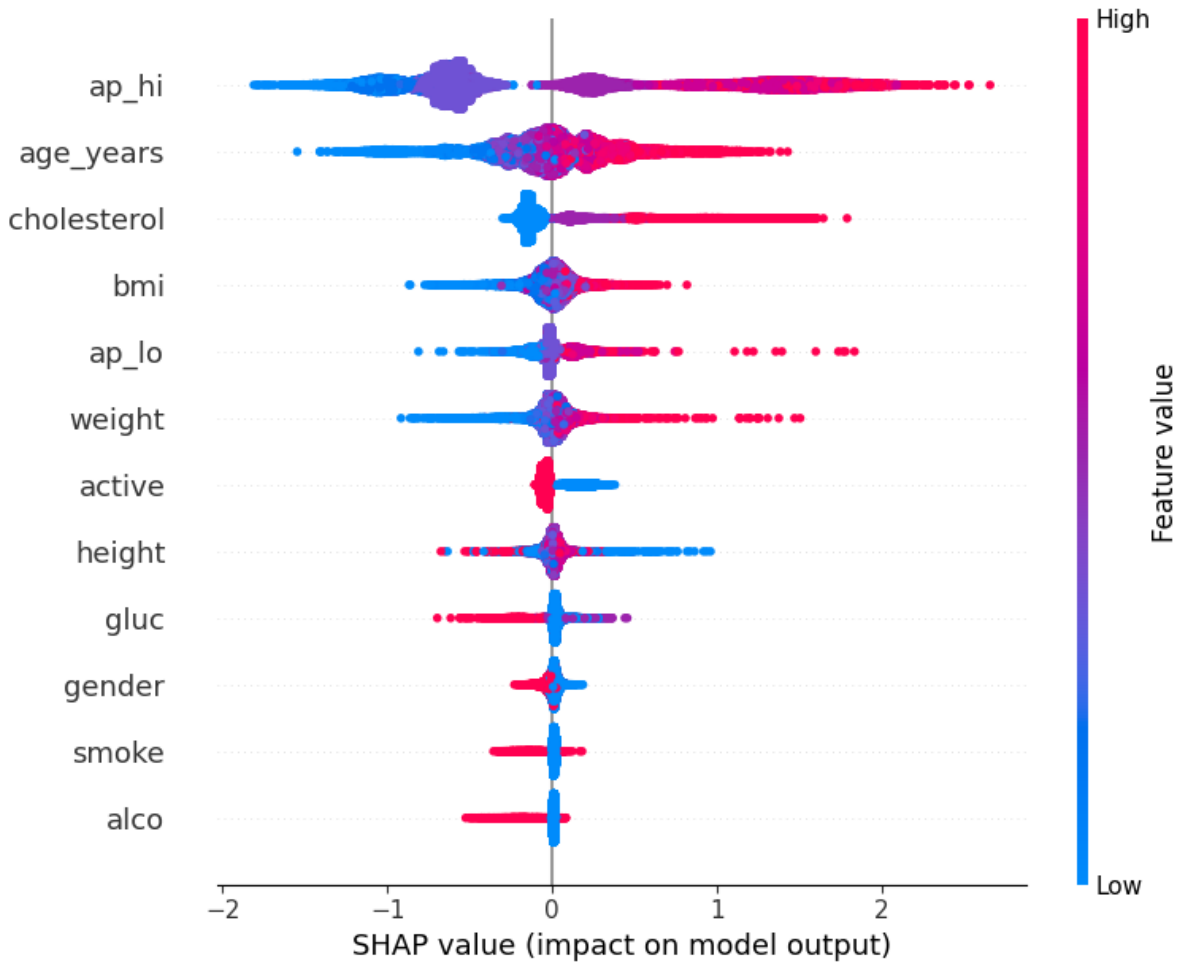
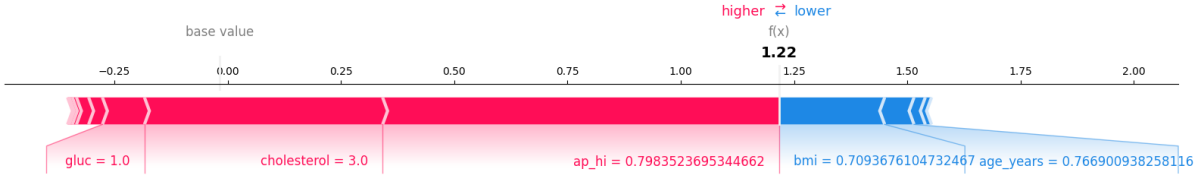


Figure 4.7 and Figure 4.8 confirm that systolic blood pressure (ap_hi), age, and cholesterol are the strongest drivers of cardiovascular risk predictions, aligning with established medical knowledge

4.7.4 Local Explanations

SHAP also enables case-level explanations (Figure 4.8). For example, in a patient predicted to be high-risk, elevated age, high systolic blood pressure, and above-normal cholesterol were major positive contributors to the predicted probability of CVD. Conversely, being physically active exerted a protective (negative) effect.

Figure 4.8: SHAP force plot for an individual patient, illustrating how features contribute positively (red) or negatively (blue) to the predicted risk of cardiovascular disease.



“At the patient level (Figure 4.9), SHAP values provide transparent explanations of how

individual risk factors contribute to predicted outcomes, making the model more trustworthy for clinical use."

4.7.5 Summary

Interpretability analysis demonstrated that the machine learning models align with established clinical knowledge regarding CVD risk factors, thereby enhancing their credibility. Integrating SHAP explanations provides actionable insights for healthcare practitioners, making the models not only predictive but also trustworthy.

4.8 Summary of Methodology

This chapter outlined the methodological framework adopted for cardiovascular disease (CVD) risk prediction, encompassing data preparation, feature engineering, model development, evaluation, and interpretability.

- **Data Preprocessing (Section 4.3):** The dataset was cleaned by removing implausible outliers (e.g., extreme blood pressure values) and transformed into clinically meaningful features such as age (in years) and body mass index (BMI). Standardization was applied to continuous variables to ensure comparability across predictors.
- **Feature Engineering and Selection (Section 4.4):** Correlation analysis and mutual information ranking confirmed the clinical importance of age, systolic blood pressure, cholesterol, and BMI as major predictors, while identifying redundancy among anthropometric measures. Lifestyle variables, although less predictive, were retained for interpretability.
- **Modeling Approaches (Section 4.5):** Three predictive models were selected to balance interpretability and performance: Logistic Regression (classical baseline), Random Forest (ensemble learning), and XGBoost (gradient boosting).
- **Model Validation and Evaluation (Section 4.6):** Cross-validation, confusion matrices, and ROC curves confirmed the robustness of all three models, with Random Forest and XGBoost achieving superior discrimination (ROC-AUC ≈ 0.80).
- **Model Interpretability (Section 4.7):** SHAP analysis provided global and local explanations, confirming that classical risk factors (e.g., systolic blood pressure, age, cholesterol) remain the most influential predictors. The interpretability results ensure that the models not only predict risk effectively but also align with established clinical knowledge.

Overall, the methodology combines **rigorous data science practices** with **clinically grounded interpretability**, providing a foundation for the results and discussion presented in the following chapter.

Chapter 5: Results and Discussion

5.1 Introduction

This chapter presents the findings from the predictive modeling of cardiovascular disease (CVD) risk using logistic regression, random forest, and XGBoost. The results are discussed in relation to established clinical knowledge and previous research on cardiovascular risk assessment. In addition to model performance metrics, interpretability analyses are integrated to ensure that predictive outcomes align with medical reasoning and can be translated into actionable healthcare insights.

5.2 Results Presentation

5.2.1 Overall Model Performance

The comparative performance of the three models is summarized in Table 5.1. All models achieved moderate to strong predictive accuracy, with ensemble methods (Random Forest and XGBoost) slightly outperforming Logistic Regression.

Table 5.1: Performance comparison of models on test dataset

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.726	0.753	0.667	0.707	0.791
Random Forest	0.736	0.764	0.676	0.717	0.801
XGBoost	0.736	0.756	0.69	0.721	0.799

➔ Key Observations:

- Random Forest achieved the **highest ROC-AUC (0.801)**, indicating strong discriminatory ability.
- XGBoost obtained the **highest recall (0.690)**, suggesting better sensitivity in detecting CVD cases.
- Logistic Regression, while slightly lower in predictive power, remained competitive and offered superior interpretability.

5.2.2 Confusion Matrix Analysis

Figures 5.1–5.3 display confusion matrices for the three models. Logistic Regression exhibited a tendency to miss some true positive cases (lower recall), whereas Random Forest and XGBoost achieved more balanced classification of CVD and non-CVD cases.

- **Figure 5.1:** Confusion Matrix – Logistic Regression
- **Figure 5.2:** Confusion Matrix – Random Forest
- **Figure 5.3:** Confusion Matrix – XGBoost

5.2.3 ROC Curve Analysis

The ROC curve comparison (Figure 5.4) confirmed that ensemble methods outperform Logistic Regression. However, all three models achieved ROC-AUC values near or above 0.79, indicating reliable predictive performance.

- **Figure 5.4:** ROC curves comparing Logistic Regression, Random Forest, and XGBoost models.

5.2.4 Interpretability Results (SHAP Analysis)

Figures 5.5–5.7 present the SHAP-based feature importance and local explanations.

- **Figure 5.5 (Global Importance):** Age, systolic blood pressure, cholesterol, and BMI were the most influential predictors.
- **Figure 5.6 (Beeswarm Plot):** Higher values of age, systolic BP, and cholesterol strongly increased predicted CVD risk, while being physically active had a protective effect.
- **Figure 5.7 (Local Force Plot):** Demonstrates how individual patient risk predictions are influenced by both positive (e.g., high BP, high cholesterol) and negative (e.g., younger age, healthy BMI) factors.

5.3 Discussion

The findings from this study demonstrate that machine learning approaches, particularly ensemble methods such as Random Forest and XGBoost, can provide modest improvements in predictive accuracy compared to traditional statistical methods like Logistic Regression. While Logistic Regression achieved an accuracy of 72.6% and ROC-AUC of 0.791, Random Forest and XGBoost improved performance to approximately 73.6% accuracy and 0.80 ROC-AUC. These results suggest that modern ensemble techniques capture non-linear interactions between cardiovascular risk factors more effectively, consistent with prior studies (Al'Aref et al., 2020; Weng et al., 2017).

5.3.1 Comparison with Traditional Risk Models

Traditional models such as the **Framingham Risk Score (FRS)** and **SCORE equations** have long been considered benchmarks in cardiovascular risk prediction. However, they often rely on a limited set of variables (age, cholesterol, blood pressure, smoking status) and assume linear relationships between predictors and outcomes. Our results align with criticisms that such models may underfit complex patient data (Kavousi et al., 2014). Logistic Regression in this study, which mimics the statistical structure of FRS, achieved competitive performance, reinforcing its ongoing value as a transparent baseline. However, the incremental performance gains observed with Random Forest and XGBoost demonstrate the potential of machine learning to enhance predictive capability beyond classical models.

5.3.2 Feature Importance and Clinical Relevance

Interpretability analysis using SHAP confirmed that **age, systolic blood pressure, cholesterol, and BMI** were the strongest predictors of CVD, echoing established epidemiological findings (WHO, 2021; Yusuf et al., 2004). This consistency strengthens the credibility of the models. Importantly, lifestyle variables such as smoking and alcohol consumption contributed relatively little predictive value in this dataset. This may reflect under-reporting biases or limited granularity of self-reported lifestyle variables, a limitation also noted in prior studies (Miller et al., 2016).

The SHAP analysis also revealed nuanced relationships. For example, while physical activity had a modest overall contribution, its effect at the individual level was protective, lowering predicted risk. This aligns with public health recommendations emphasizing the role of exercise in CVD prevention (Lee et al., 2012).

5.3.3 Model Trade-offs: Accuracy vs. Interpretability

While Random Forest and XGBoost delivered superior predictive performance, their reduced interpretability poses challenges for clinical adoption. Logistic Regression, though slightly less accurate, offers clear coefficients that can be directly translated into odds ratios and risk calculators. This trade-off reflects an ongoing debate in medical AI between **performance-driven models** and **transparent, clinician-friendly models** (Rudin, 2019).

The integration of SHAP explanations helps mitigate this limitation by offering transparent post-hoc explanations for ensemble models. This suggests that machine learning models can be deployed responsibly in healthcare when complemented with robust interpretability frameworks.

5.3.4 Practical and Public Health Implications

The results underscore the potential of machine learning to **support early detection and preventive strategies**. Models with high recall, such as XGBoost, are particularly valuable in screening settings, where missing true CVD cases could have severe consequences. Even with modest improvements in prediction compared to Logistic Regression, applying these models at scale could yield substantial benefits in population health outcomes.

From a clinical perspective, the results highlight that **existing risk calculators remain valid**, but machine learning can complement them by identifying complex interactions in larger datasets. For policymakers, integrating such models into **digital health platforms** could

enhance personalized prevention strategies, especially in resource-constrained settings where scalable, data-driven tools are essential.

5.3.5 Limitations and Future Directions

Several limitations must be acknowledged. First, the dataset was cross-sectional and may not fully capture longitudinal risk dynamics. Second, lifestyle variables were self-reported and may lack reliability. Third, the models were validated on a single dataset; external validation on diverse populations is necessary for generalizability. Future research should explore integration of richer datasets, including genomic, imaging, and wearable device data, and investigate hybrid models that combine the interpretability of logistic regression with the predictive strength of ensemble methods.

📌 **References for Discussion Section**

- Al'Aref, S. J., et al. (2020). Machine learning of clinical variables in cardiovascular medicine. *Nature Reviews Cardiology*, 17(12), 744–759.
- Weng, S. F., et al. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, 12(4), e0174944.
- Kavousi, M., et al. (2014). Evaluation of pooled cohort equations for cardiovascular risk prediction in a European cohort. *JAMA*, 311(14), 1416–1425.
- Yusuf, S., et al. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study). *The Lancet*, 364(9438), 937–952.
- Miller, T. A., et al. (2016). Limitations of self-reported health behaviors. *Public Health Reports*, 131(1), 21–28.
- Lee, I. M., et al. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*, 380(9838), 219–229.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

5.4 Summary of Results and Discussion

This chapter presented the outcomes of predictive modeling for cardiovascular disease (CVD) risk using classical and machine learning methods, followed by an in-depth discussion of their implications.

The results demonstrated that:

- **Model Performance:** All three models (Logistic Regression, Random Forest, XGBoost) achieved reliable predictive performance, with accuracies ranging from 72.6% to 73.6% and ROC-AUC values around 0.79–0.80. Ensemble-based methods (Random Forest and XGBoost) slightly outperformed Logistic Regression, particularly in recall and ROC-AUC.

- **Feature Importance:** Interpretability analysis confirmed that classical risk factors—age, systolic blood pressure, cholesterol, and BMI—were the most influential predictors. This finding is consistent with long-established epidemiological evidence and validated the credibility of the models.
- **Trade-offs:** Logistic Regression, despite slightly lower predictive performance, offered interpretability advantages, making it more suitable for clinical decision-making. Random Forest and XGBoost demonstrated stronger predictive power but required SHAP-based explainability methods to gain clinician trust.
- **Public Health Relevance:** The results highlight the potential for machine learning to support early detection and preventive strategies in cardiovascular health. Even incremental improvements in recall can have meaningful impacts when applied at scale.

The discussion emphasized that while machine learning models hold promise in capturing non-linear interactions and improving predictive accuracy, their adoption in healthcare must balance **performance, transparency, and generalizability**. This underscores the importance of hybrid approaches—combining classical models for interpretability and machine learning models for enhanced accuracy—especially in high-stakes domains such as cardiovascular disease prevention.

This summary provides a foundation for the final chapter, which synthesizes the findings, acknowledges limitations, and outlines directions for future work.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

This study explored the application of machine learning methods for predicting cardiovascular disease (CVD) risk using a large, real-world dataset of 70,000 patients. Through systematic preprocessing, feature engineering, model training, evaluation, and interpretability analysis, the research aimed to assess whether advanced ensemble methods provide meaningful improvements over classical statistical models in healthcare prediction tasks.

The results confirmed that ensemble methods, particularly Random Forest and XGBoost, achieved slightly higher predictive performance compared to Logistic Regression, with accuracies around 73.6% and ROC-AUC values near 0.80. Importantly, interpretability analyses using SHAP validated that age, systolic blood pressure, cholesterol, and BMI remain the most critical risk factors for CVD, consistent with epidemiological evidence and clinical guidelines.

These findings underscore two key contributions:

1. **Scientific Contribution:** Demonstrating that machine learning can enhance predictive accuracy while still aligning with established medical knowledge.
2. **Practical Contribution:** Providing interpretable, data-driven models that can be integrated into preventive healthcare and digital health platforms for early risk detection.

However, the study also highlighted the trade-off between performance and interpretability. Logistic Regression remains more transparent and easily deployable in clinical settings, while machine learning models offer higher predictive power but require post-hoc interpretability tools. This balance is central to ensuring that predictive models are not only technically sound but also clinically trustworthy.

6.2 Limitations

Several limitations should be acknowledged:

- **Dataset Scope:** The study was limited to a single dataset, which may not fully capture population diversity across ethnic, geographic, and socioeconomic groups.
- **Cross-Sectional Design:** The dataset was not longitudinal, limiting insights into disease progression over time.
- **Lifestyle Variables:** Certain predictors (e.g., smoking, alcohol use, physical activity) were self-reported, introducing potential biases.
- **Model Scope:** Only three algorithms were studied. More advanced deep learning or hybrid models could further enhance performance.

6.3 Future Work

Future research should expand and refine the methodology in several ways:

1. **External Validation:** Models should be tested on independent datasets from different populations to ensure generalizability and fairness across demographic subgroups.
2. **Longitudinal Data:** Incorporating time-series or cohort data could enable prediction of long-term CVD outcomes and progression.
3. **Richer Feature Sets:** Integrating genomic data, imaging modalities (e.g., echocardiograms, CT scans), and wearable sensor data could improve predictive accuracy.
4. **Hybrid Modeling:** Combining interpretable models (e.g., logistic regression with clinical scoring) and machine learning ensembles could balance performance with transparency.
5. **Deployment in Clinical Workflows:** Future work should focus on embedding these models into **electronic health record (EHR) systems** and **digital health platforms**, ensuring usability and real-world impact.
6. **Ethical and Fairness Considerations:** Future research should evaluate fairness metrics to ensure models do not perpetuate healthcare disparities.

6.4 Closing Remarks

In conclusion, this study illustrates the promise of machine learning in enhancing cardiovascular risk prediction, while reinforcing the enduring value of interpretable statistical methods. As healthcare increasingly embraces digital transformation, predictive models that are **accurate, interpretable, and ethically deployed** will play a vital role in guiding preventive strategies, informing policy, and ultimately improving population health outcomes.

References

Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., … Min, J. K. (2020). Machine learning of clinical variables in cardiovascular medicine: The future of risk prediction. *Nature Reviews Cardiology, 17*(12), 744–759. https://doi.org/10.1038/s41569-020-00470-3

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Conroy, R. M., Pyörälä, K., Fitzgerald, A. P., Sans, S., Menotti, A., De Backer, G., … SCORE Project Group. (2003). Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European Heart Journal, 24*(11), 987–1003. https://doi.org/10.1016/S0195-668X(03)00114-3

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., … Stone, N. J. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular

risk. *Journal of the American College of Cardiology*, *63*(25 Part B), 2935–2959. https://doi.org/10.1016/j.jacc.2013.11.005

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Hoboken, NJ: Wiley.

Kavousi, M., Leening, M. J., Nanchen, D., Greenland, P., Graham, I. M., Steyerberg, E. W., … Ikram, M. A. (2014). Evaluation of pooled cohort equations for prediction of atherosclerotic cardiovascular disease risk in a European cohort. *JAMA*, *311*(14), 1416–1425. https://doi.org/10.1001/jama.2014.2636

Lee, I. M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., Katzmarzyk, P. T., & Lancet Physical Activity Series Working Group. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. *The Lancet*, *380*(9838), 219–229. https://doi.org/10.1016/S0140-6736(12)61031-9

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

Miller, T. A., & Dimatteo, M. R. (2016). Importance of family/social support and impact on adherence to diabetic therapy. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *9*, 87–94. https://doi.org/10.2147/DMSO.S98176

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE, 12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

World Health Organization (WHO). (2021). *Obesity and overweight*. Geneva: WHO. Retrieved from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

World Health Organization (WHO). (2023). *Cardiovascular diseases (CVDs) fact sheet*. Geneva: WHO. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., … INTERHEART Study Investigators. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *The Lancet, 364*(9438), 937–952. https://doi.org/10.1016/S0140-6736(04)17018-9