Islamic University Faculty of Engineering		الجامعة الإسلامية كلية الهندسة

ENGR 3701
ARTIFICIAL INTELLIGENCE (SPECIAL TOPICS)

[CO_6, PI_3_1, SO_3]

Explainable AI for Cardiovascular Disease Risk Prediction using SHAP

Submitted To

Prof. Adel Abdennour

Submitted by

Mahmoud Sameh, 421018033

Section No: 6063

Islamic University Faculty of Engineering		الجامعة الإسلامية كلية الهندسة

EE 3701 – Artificial Intelligence (Special Topic)

DECLARATION STATEMENT FOR THE ORIGINALITY OF THE DESIGN PROJECT

We hereby declare that this design project comprises our original work. No material in this design project has been previously published and written by another person, except where due reference is made in the text of the problem definition statement.

Name, Family Name : Mahmoud Sameh **Signature :** 

Table of Contents

Abstract

Cardiovascular Disease (CVD) remains a leading cause of mortality worldwide, underscoring the importance of accurate early risk prediction. While machine learning models offer promising predictive capabilities, their "black box" nature often limits clinical adoption due to a lack of transparency. This project addresses this challenge by first developing and tuning a Logistic Regression model to predict CVD risk using a publicly available dataset featuring demographic, objective, examination, and subjective patient features. Following model evaluation, the SHAP (SHapley Additive exPlanations) framework was comprehensively employed to interpret the model's behavior. Global SHAP explanations successfully identified key risk factors such as systolic blood pressure, patient age, and high cholesterol levels as primary drivers of the model's predictions, aligning with established medical knowledge. Furthermore, local SHAP explanations, visualized through detailed waterfall and force plots, provided granular insights into how individual patient characteristics quantitatively contribute to their specific risk scores, even illuminating the model's reasoning for misclassified instances.

1. Introduction

Cardiovascular Diseases (CVDs) represent a formidable global health challenge, consistently ranking as a leading cause of morbidity and mortality worldwide, accounting for an estimated 17.9 million lives each year [1]. The significant burden imposed by CVDs on individuals and healthcare systems underscores the critical need for effective preventative strategies, chief among them being the early and accurate identification of at-risk individuals [2]. Advances in Artificial Intelligence (AI) and Machine Learning (ML) have shown considerable promise in developing predictive models capable of assessing CVD risk by learning complex patterns from diverse patient data [3]. These models can analyze demographic information, objective clinical measurements, results from medical examinations, and subjective lifestyle factors to estimate an individual's likelihood of developing CVD.

However, a significant hurdle to the widespread clinical adoption of many high-performing AI models is their inherent "black-box" nature [4]. While models like deep neural networks or ensemble methods might achieve high predictive accuracy, their internal decision-making processes often remain opaque to users. This lack of transparency can erode trust among clinicians, who require a clear understanding of *why* a model arrives at a particular risk assessment before incorporating it into critical patient care decisions [5]. Without interpretability, it is difficult to validate the model's reasoning, identify potential biases, or confidently explain predictions to patients. The field of Explainable AI (XAI) has emerged to address these concerns, aiming to make AI systems more transparent and understandable [6].

This project aims to address this critical gap by focusing on XAI techniques for CVD risk prediction. The primary goal is to develop a robust CVD risk prediction model and subsequently employ the SHAP (SHapley Additive exPlanations) framework [7] to comprehensively interpret its predictions. Specifically, this project will involve:

1. Building and evaluating a Logistic Regression model, a common and relatively interpretable baseline for risk prediction tasks.
2. Applying SHAP to elucidate global feature importance, identifying which patient characteristics most significantly drive the model(s) predictions overall.
3. Utilizing SHAP to generate local explanations, detailing how specific features contribute to the risk score for individual patient instances.

The basic approach involves utilizing a publicly available cardiovascular disease dataset [8]. After appropriate data preprocessing, including outlier handling, feature engineering (such as BMI calculation and age conversion), scaling, and encoding, the predictive model(s) will be

trained and tuned. SHAP analysis will then be performed to generate the necessary explanations. While the development of novel AI algorithms is beyond the scope of this project, the application and interpretation of advanced XAI methods like SHAP to a clinically relevant problem like CVD risk prediction constitute its primary contribution. Related work in the field has increasingly emphasized the need for XAI in healthcare, with studies demonstrating its utility in various medical applications [9], [10]. This project seeks to demonstrate the practical utility of SHAP in making AI-driven risk assessment tools more transparent, understandable, and ultimately, more trustworthy. Table 1.1 outlines the key design specifications for this project.

Table 1.1: Design Specifications

Specification Category	Detail
Project Objective	Predict presence/absence of Cardiovascular Disease (CVD) and provide interpretable explanations for model predictions.
Input Data Source	Cardiovascular Disease Dataset [8]
Key Input Features	Demographic (Age, Gender), Objective (Height, Weight, BMI), Examination (Systolic/Diastolic BP, Cholesterol, Glucose), Subjective (Smoking, Alcohol, Physical Activity).
Predictive Model(s)	1. Logistic Regression
Explainability Method	SHAP (SHapley Additive exPlanations) using appropriate explainers
Key Output	1. CVD risk predictions (binary). 2. Global feature importance rankings (SHAP summary plots). 3. Local patient-specific explanations (SHAP waterfall/force plots).
Performance Metrics	ROC AUC, Accuracy, Precision, Recall, F1-Score.
Software/Tools	Python, Pandas, NumPy, Scikit-learn, Keras, SHAP, Matplotlib.

The remainder of this report is organized as follows: Section 2 details the theoretical background, including the dataset, preprocessing steps, the predictive models employed, and an overview of the SHAP methodology. Section 3 presents the simulation results, including model performance and the global and local SHAP explanations. Section 4 provides concluding remarks. Finally, references and an appendix containing the project code are provided.

2. System Modeling

2.1 Dataset Description

The data utilized for this project was sourced from the publicly available "Kaggle Cardiovascular Disease Dataset," [8]. This dataset is commonly used for predictive modeling

tasks related to cardiovascular health and contains a variety of features collected from patients. The dataset comprises approximately 70,000 patient records and 11 features and 1 target before preprocessing and feature engineering.

The features can be broadly categorized into objective (factual information), examination (results from medical examination), and subjective (information provided by the patient) types. These features include demographic details, anthropometric measurements, blood pressure readings, blood test results for cholesterol and glucose, and lifestyle habits. The target variable for prediction is cardio, a binary indicator representing the presence (1) or absence (0) of cardiovascular disease in the patient at the time of examination. A detailed breakdown of the relevant features used in this study is provided in Table 2.1.

Table 2.1: Dataset Feature Description

Feature Name	**Original Category**	**Type**	**Description / Units / Coding**
age	Objective	Integer	Age in days
gender	Objective	Categorical	1: Women, 2: Men
height	Objective	Integer	Height in cm
weight	Objective	Float	Weight in kg
ap_hi	Examination	Integer	Systolic blood pressure (mmHg)
ap_lo	Examination	Integer	Diastolic blood pressure (mmHg)
cholesterol	Examination	Categorical	1: Normal, 2: Above normal, 3: Well above normal
gluc	Examination	Categorical	1: Normal, 2: Above normal, 3: Well above normal
smoke	Subjective	Binary	0: Does not smoke, 1: Smokes
alco	Subjective	Binary	0: Does not consume alcohol, 1: Consumes alcohol

active	Subjective	Binary	0: Not physically active, 1: Physically active
cardio (Target)	-	Binary	0: No Cardiovascular Disease, 1: Presence of Cardiovascular Disease
(Derived) age_years	Objective	Float	Age in years
(Derived) bmi	Objective	Float	Body Mass Index (kg/m²)

2.2 Data Preprocessing

Prior to model development, the raw dataset underwent several preprocessing steps to ensure data quality and suitability for machine learning algorithms. Initially, any non-predictive identifier columns were removed. A key feature engineering step involved converting the patient's age from days, as provided in the dataset, to years for better interpretability and more conventional use in risk modeling. Furthermore, Body Mass Index (BMI), a widely recognized indicator of body fat based on height and weight, was calculated for each patient using the formula: BMI = weight (kg) / (height (m))^2. This derived feature often provides more predictive power than height and weight as individual inputs. Significant attention was given to outlier detection and removal to mitigate their potential adverse effects on model performance and interpretation. This involved programmatic filtering based on physiological plausibility, such as ensuring systolic blood pressure (ap_hi) was greater than or equal to diastolic blood pressure (ap_lo), and that both blood pressure readings fell within predefined acceptable ranges (e.g., 80-240 mmHg for systolic, 50-180 mmHg for diastolic). Extreme values for height, weight, and the newly engineered BMI were also addressed by removing records falling outside the 2.5th and 97.5th percentiles for these respective features, thereby retaining the central 95% of the data distribution for these anthropometric measures.

Following feature engineering and outlier mitigation, the data was prepared for model input. Categorical features, namely gender, cholesterol, and gluc, were transformed using one-hot encoding. This process converts each category within these features into a new binary (0 or 1) column, enabling the models to process these nominal data types without imposing an artificial ordinal relationship. Numerical features, including the derived age_years and bmi, along with original measurements like height, weight, ap_hi, ap_lo, and binary lifestyle indicators (smoke,

alco, active), were standardized using Standard Scaling. This technique transforms the data to have a mean of zero and a standard deviation of one, which is crucial for distance-based algorithms and gradient-based optimization methods (like those in Logistic Regression and Neural Networks) to ensure that features with larger value ranges do not disproportionately influence the model. After these steps, the dataset was split into training and testing sets to facilitate model training and unbiased evaluation.

2.3 Predictive Model

Logistic Regression is a widely used statistical model for binary classification tasks, valued for its simplicity and inherent interpretability through its coefficients. It models the probability of a binary outcome by applying a sigmoid function to a linear combination of input features. The model equation can be expressed as $\log(P(Y=1)/P(Y=0)) = \beta_0 + \Sigma(\beta_i X_i)$, where $P(Y=1)$ is the probability of the positive class (presence of CVD), X_i are the input features, and β_i are their corresponding coefficients.

For this project, the Logistic Regression model was implemented within a scikit-learn pipeline that included the preprocessing steps (scaling and one-hot encoding) described in Section 2.2. To optimize its performance, hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation. This process systematically explored various model configurations. The optimal configuration identified involved applying L1 regularization (Lasso), which aids in preventing overfitting and can also perform implicit feature selection by shrinking some feature coefficients to zero. The regularization strength was set to a moderate level (inverse of regularization strength $C = 0.1$). The model was configured to fit an intercept term, and class weights were adjusted to be 'balanced' to account for any potential imbalance in the prevalence of cardiovascular disease within the training data. This optimized configuration achieved a cross-validated Area Under the ROC Curve (AUC) score of approximately 0.787 during the tuning phase. The performance of this finalized model on the independent test set is further detailed in Section 3.1.

2.4 Predictive Models

To quantitatively assess the performance of the developed predictive model (Logistic Regression), a standard set of evaluation metrics for binary classification tasks was employed. These metrics provide a comprehensive view of the models' ability to correctly identify

individuals with and without cardiovascular disease, as well as their overall discriminative power. The primary metrics used are described below:

- **Accuracy:** This is the most intuitive metric, representing the proportion of total predictions that the model classified correctly. It is calculated as: Accuracy = (True Positives + True Negatives) / (Total Number of Samples)
- **Precision (Positive Predictive Value):** It answers the question: "Of all patients predicted to have CVD, how many actually do?" High precision indicates a low false positive rate. It is calculated for the positive class as: Precision = True Positives / (True Positives + False Positives)
- **Recall (Sensitivity or True Positive Rate):** It answers the question: "Of all patients who actually have CVD, how many did the model correctly identify?" High recall indicates a low false negative rate, which is often critical in medical diagnoses. It is calculated as: Recall = True Positives / (True Positives + False Negatives)
- **F1-Score:** The F1-score is the harmonic mean of Precision and Recall, providing a single metric that balances both

2.5 Explainable AI - SHAP

While robust predictive performance is essential, understanding why a model makes certain predictions is equally crucial, particularly in high-stakes domains such as healthcare. To address the "black-box" nature of the developed predictive models, this project employs SHAP (SHapley Additive exPlanations) [7], a state-of-the-art Explainable AI (XAI) technique. SHAP provides a unified and theoretically grounded framework for interpreting the output of any machine learning model by assigning each feature an importance value – a "SHAP value" – for a particular prediction.

The foundation of SHAP lies in Shapley values, a concept originating from cooperative game theory. In this context, features are treated as "players" in a game, and the prediction is the "payout." Shapley values fairly distribute this payout among the features, reflecting each feature's marginal contribution to the prediction, considering all possible combinations (coalitions) of features. SHAP values offer desirable properties such as local accuracy (the sum of SHAP values for all features for a given instance equals the difference between the model's prediction for that instance and a baseline or expected model output) and consistency (a

feature's SHAP value should not decrease if the model changes such that the feature's marginal contribution increases or stays the same).

3. Simulation and Design Methods

3.1 Model Performance

The optimized Logistic Regression model, after hyperparameter tuning using GridSearchCV (with best parameters: L1 penalty, C=0.1, balanced class weights, and intercept fitting), yielded the following performance on the independent test set, as summarized in Table 3.1.

Table 3.1:Performance Metrics for Logistic Regression

Metric	**Value**
Accuracy	0.7244
ROC AUC Score	0.7861
Precision (Class 0: No CVD)	0.71
Recall (Class 0: No CVD)	0.78
F1-Score (Class 0: No CVD)	0.74
Precision (Class 1: CVD)	0.74
Recall (Class 1: CVD)	0.67
F1-Score (Class 1: CVD)	0.71
Macro Average F1-Score	0.72

The Logistic Regression model achieved an ROC AUC of 0.7861, indicating a good level of discriminative ability. The accuracy was 72.44%. The F1-score for predicting the presence of CVD (Class 1) was 0.71. While these metrics indicate reasonable performance, the focus remains on understanding how the model arrived at these predictions.

This level of performance, while not perfect, is understandable given that the available input features, though relevant, represent only a subset of the complex factors contributing to cardiovascular disease. The primary goal of this project, however, is not to achieve cutting-edge accuracy but to thoroughly explore and contrast the explainability of these different modeling approaches. The subsequent sections will therefore focus on using SHAP to understand the decision-making processes of both models.

3.2 Global Explanations (SHAP)

To understand the overall behavior of the predictive models, SHAP summary plots were generated. These plots illustrate the global importance of each feature by aggregating their SHAP values across all instances in the test set. The global feature importance for the tuned Logistic Regression model is depicted in Figure 3.1 (Bar Plot) and Figure 3.2 (Beeswarm Plot).

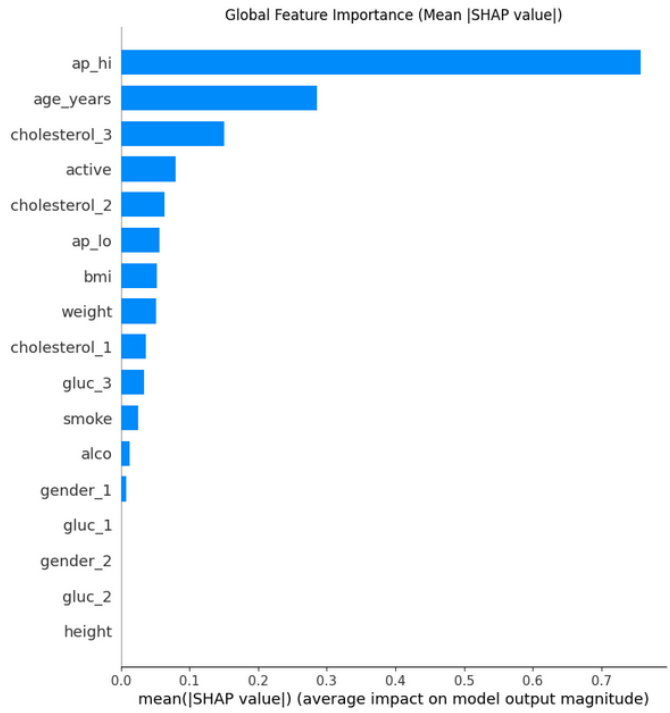

Figure 3.1: Global feature importance for the Logistic Regression model, ranked by the mean absolute SHAP value. Higher values indicate greater average impact on model predictions.

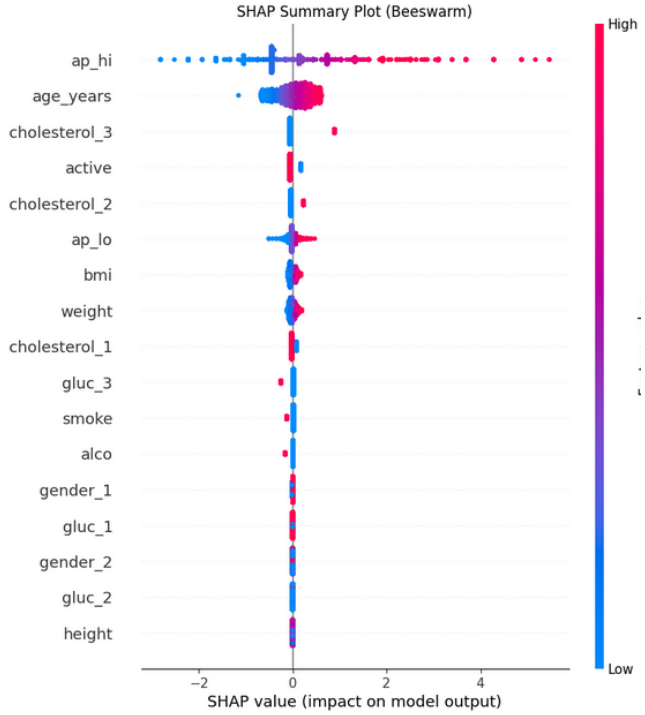

Figure 3.2: SHAP beeswarm summary plot for the Logistic Regression model. Each point represents a feature's SHAP value for a specific instance. The color indicates the feature's value (red for high, blue for low), and the x-axis shows the SHAP value's impact on the model output.

The bar plot in Figure 3.1 clearly ranks features by their mean absolute SHAP value, indicating their overall influence on the model's predictions. The top three most influential features are ap_hi (systolic blood pressure), age_years, and cholesterol_3 (representing well above normal

cholesterol). This aligns with established medical knowledge regarding significant risk factors for cardiovascular disease. Features such as active (physical activity), cholesterol_2 (above normal cholesterol), and ap_lo (diastolic blood pressure) also demonstrate notable contributions. Conversely, features like height, gluc_2 (above normal glucose), and gender_2 (representing men, based on the coding 1:women, 2:men) exhibit minimal global impact on this model's predictions.

The beeswarm plot in Figure 3.2 provides a more nuanced view. For ap_hi, high values (red dots) are predominantly associated with positive SHAP values, indicating they increase the predicted log-odds of CVD, while lower values (blue dots) have negative SHAP values, decreasing the predicted risk. A similar pattern is observed for age_years, where older ages (redder dots) contribute positively to risk. For cholesterol_3, its presence (which would be a high value in its one-hot encoded form, though color here might represent the original categorical value if SHAP was given that for coloring) strongly pushes predictions towards higher risk. The plot also reveals the distribution of impact for each feature; for instance, while active has a moderate overall importance, its impact is primarily negative (protective) when present (feature value is 1, likely shown as red). Features like bmi and weight show a mix of positive and negative SHAP values, with higher values generally increasing risk. The minimal impact of height is further confirmed, with most of its SHAP values clustered tightly around zero.

3.3 Local Explanations (SHAP) - Case Study:

Beyond understanding global feature importance, SHAP provides powerful local explanations, detailing how individual features contribute to the prediction for a specific instance. This section presents a case study for one patient from the test set to illustrate how SHAP elucidates the decision-making process of the Logistic Regression model.

- **Original Feature Values (Selected):**
    - age_years: 55.33
    - gender: 1.0 (Woman)
    - ap_hi (Systolic BP): 130.0 mmHg
    - ap_lo (Diastolic BP): 90.0 mmHg
    - cholesterol: 1.0 (Normal)
    - gluc: 1.0 (Normal)

- smoke: 0.0 (Non-smoker)
- active: 1.0 (Physically active)
- bmi: 26.40

- **Actual CVD Status:** 1 (Presence of Cardiovascular Disease)
- **Model's Predicted Likelihood (CVD=1):** 24.6% (corresponding to a log-odds output of approximately 0.246)
- **Model's Predicted Label:** 0 (Absence of Cardiovascular Disease)
- **Prediction Correctness:** INCORRECT (This is a False Negative)

The SHAP force plot (Figure 3.3) and waterfall plot (Figure 3.4) provide a detailed breakdown of the feature contributions leading to this prediction.

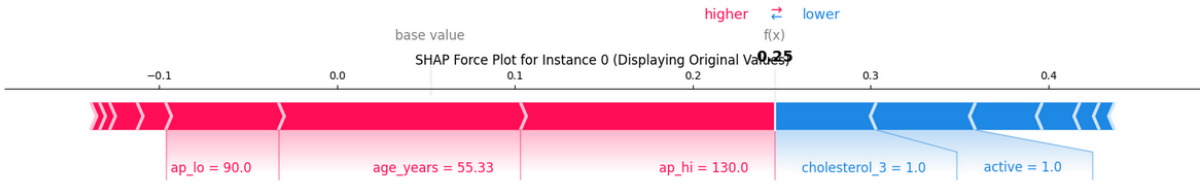

Figure 3.3: SHAP force plot for Instance 0 using the Logistic Regression model. Red features push the prediction towards higher risk (positive class), while blue features push it towards lower risk. The final output f(x) is 0.246 (log-odds scale).

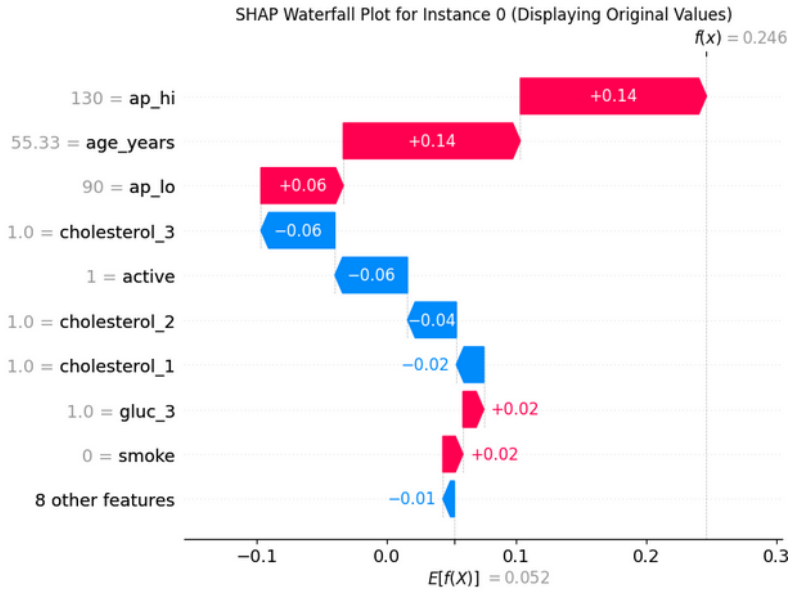

Figure 3.4:SHAP waterfall plot for Instance 0 using the Logistic Regression model. It illustrates how each feature's SHAP value contributes to shifting the prediction from the base value E[f(x)] = 0.052 to the final output f(x) = 0.246 (log-odds scale).

The SHAP waterfall plot (Figure 3.4) provides a granular deconstruction of the feature contributions for this specific patient instance, illustrating how the model's prediction deviates from a baseline (expected) log-odds value E[f(x)] of approximately 0.052 to a final log-odds

output of 0.246. Several features were identified as key drivers influencing this prediction. Notably, the patient's systolic blood pressure (ap_hi) of 130.0 mmHg and their age (age_years) of 55.33 years were the two most significant factors increasing the predicted risk, each contributing +0.14 to the log-odds. The diastolic blood pressure (ap_lo) of 90.0 mmHg also moderately increased the risk with a SHAP value of +0.06. Interestingly, for this instance, the patient being a non-smoker (smoke=0) and having normal glucose levels (implying the processed feature gluc_3 for "well above normal glucose" is 0) both registered small positive SHAP contributions (+0.02 each). This suggests that, relative to the model's baseline expectation for these features, these specific states (non-smoking, normal glucose) in combination with the patient's other characteristics slightly nudged the prediction towards higher risk, or that their protective effect was less than the average expectation.

Conversely, several factors acted to decrease the predicted risk for this individual. The patient's normal cholesterol level (original cholesterol=1.0, meaning the processed feature cholesterol_3 representing "well above normal cholesterol" is 0) was the most significant protective contributor, with a SHAP value of -0.06. Similarly, being physically active (active=1.0) also provided a protective effect, reducing the log-odds by -0.06. Other cholesterol-related features, reflecting normal or only moderately elevated levels (e.g., cholesterol_2=0, cholesterol_1=1), further contributed to lowering the predicted risk, albeit with smaller magnitudes (-0.04 and -0.02, respectively). The SHAP force plot (Figure 3.3) offers a complementary visualization, clearly depicting systolic blood pressure, age, and diastolic blood pressure as the primary factors increasing risk (red segments), while the absence of very high cholesterol and the presence of physical activity act as the main counteracting forces (blue segments).

This particular instance represents a false negative: the patient's actual cardiovascular disease status was positive (1), yet the model predicted a negative status (0) with a corresponding low likelihood of 24.6% for having CVD. The SHAP explanations illuminate the model's reasoning for this misclassification. While the patient's age and elevated blood pressure readings correctly pushed the predicted risk upwards, the model assigned substantial negative (protective) weight to their normal cholesterol profile and physically active lifestyle. In this specific case, the cumulative impact of these protective factors, as interpreted by the model, was sufficient to counterbalance the risk-indicating features, leading to an overall low predicted probability of CVD. Such instance-level explanations are invaluable for diagnosing model behavior, understanding potential failure modes, and identifying patient profiles where the model might

struggle, which can guide further model refinement or highlight the need for additional discriminative features.

4. Conclusions

This project successfully demonstrated the development of predictive models for cardiovascular disease risk and, more critically, showcased the profound utility of the SHAP framework in demystifying their decision-making processes. By applying SHAP to a Logistic Regression model, global feature importances were identified—highlighting key risk factors such as blood pressure, age, and cholesterol levels—and local, instance-specific explanations were generated, offering granular insights into why individual predictions were made. The ability to dissect model behavior, understand feature contributions (even for misclassifications), and compare the "reasoning" of different model architectures underscores the indispensable role of Explainable AI in fostering trust, enabling model diagnostics, and paving the way for the responsible application of machine learning in critical domains like healthcare.

References

[1] "Cardiovascular diseases (CVDs)." Accessed: May 16, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] S. P. Karunathilake and G. U. Ganegoda, "Secondary Prevention of Cardiovascular Diseases and Application of Technology for Early Diagnosis," *BioMed Res. Int.*, vol. 2018, no. 1, p. 5767864, 2018, doi: 10.1155/2018/5767864.

[3] D.-I. Kasartzian and T. Tsiampalis, "Transforming Cardiovascular Risk Prediction: A Review of Machine Learning and Artificial Intelligence Innovations," *Life*, vol. 15, no. 1, Art. no. 1, Jan. 2025, doi: 10.3390/life15010094.

[4] W. Yang *et al.*, "Survey on Explainable AI: From Approaches, Limitations and Applications Aspects," *Hum.-Centric Intell. Syst.*, vol. 3, no. 3, pp. 161–188, Sep. 2023, doi: 10.1007/s44230-023-00038-y.

[5] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Comput. Appl.*, vol. 32, no. 24, pp. 18069–18083, Dec. 2020, doi: 10.1007/s00521-019-04051-w.

[6] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Mag.*, vol. 40, no. 2, Art. no. 2, Jun. 2019, doi: 10.1609/aimag.v40i2.2850.

[7] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.

[8] "Cardiovascular Disease dataset." Accessed: May 16, 2025. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[9] S. S Band *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Inform. Med. Unlocked*, vol. 40, p. 101286, Jan. 2023, doi: 10.1016/j.imu.2023.101286.

[10] J. Caterson, A. Lewin, and E. Williamson, "The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review," *Digit. Health*, vol. 10, p. 20552076241272657, Sep. 2024, doi: 10.1177/20552076241272657.

Appendix

The code and output can be found at https://www.kaggle.com/code/mah01sam/cardio-project

Or conveniently scan this QR code: