1. Introduction

Heart disease continues to be one of the most significant causes of death across the world, contributing to millions of preventable deaths each year (Mozaffarian et al., 2016). Just in the UK alone, a quarter of the health care burden is due to cardiovascular diseases, with the British Heart Foundation reporting heart and circulatory diseases to cause some 160,000 deaths yearly or around one every three minutes. The multiple and complex causes of heart disease, including demographic, clinical, and lifestyle-related related factors-make early detection and prevention of heart disease a significant challenge. Old diagnostic modalities rely on fixed risk factor thresholds: cholesterol levels, blood pressure, etc., which do not take cognizance of the better understanding of the multifactorial interactions leading to heart disease risk (Al'Aref et al., 2018). In addition, the majority of these methods have poor accuracy and interpretability that should guide health professions in the process of making decisions on diagnosis, particularly in different populations like the UK in which a highly diversified nature of demographic and lifestyle differences is evident.

These problems could be helped by the arrival of machine learning (ML) for modeling complex relationships in large datasets with high-level algorithms. Some examples of machine learning techniques-including Random Forest, Gradient Boosting, and Neural Networks-gather evidence for the importance of these modeling methods for augmented accuracy in heart disease risk estimation by discovering patterns not gleaned from earlier statistical methodologies (Dang et al., 2021). Further, these models thus have seen that explainable AI (XAI) frameworks- such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations)- were incorporated, interpretable to use in practice in the clinic (Gala et al., 2024). Even though advancements have been made in this regard, the healthcare adoption of ML-based tools is still minimal due to factors such as complex models, inability to provide interfaces to users, migration of access-biased prediction from demographic to prediction (Dey & Salem, 2020). The project aims to develop a predictive model by which one will be using this "Intelligent Heart Disease Risk Indicator with Machine Learning" to fill the gap of not only producing high diagnostic accuracy but also prioritizing interpretability, fairness, and accessibility via a user-centred interface.

The project was carried out in fulfilment of the MSc Artificial Intelligence with Advanced Practice programme of Teesside University, under the supervisor Ogbo Bianca, and the second reader Mansh. Running from January to May 2025, the project employed publicly available datasets, from Kaggle, to train and evaluate ML models. A web application was developed based on Flask for healthcare professionals to have a user-friendly interface to enable real-time predictions of risk for heart disease, mapping to explainability and fairness evaluation. The report documents the journey of development from data

preprocessing to model evaluation and interface design, discussing achievements and difficulties faced along the way.

1.1 Project Context and Background

Heart disease collectively considers a variety of disorders that disturb the heart, including coronary artery disease, heart failure, and arrhythmias. The most prevalent and deadly variety is indeed coronary artery disease (Mozaffarian et al., 2016). Away from the definitions to the statistics, heart disease accounts for around 17.9 million deaths every year across the globe and is termed the leading cause of mortality by the World Health Organisation (WHO). This situation reflects the gloom worse within the country as around 7.6 million people have been recorded to live with heart and circulatory diseases, costing the NHS over £7 billion each year for all the costs associated with management and treatment (British Heart Foundation, 2024). This early intervention is crucial because, despite lifestyle changes, medication, or actual surgical procedures, timely intervention will eventually improve the outcome for patients considerably (Al'Aref et al., 2018).

The traditional modalities of diagnosis of heart diseases evaluate the risk using such tools as the Framingham Risk Score or the QRISK score that calculate risk on already predetermined factors, including age, gender, levels of cholesterol, blood pressure, and smoking status. Though widely utilised in clinical practice and elsewhere, these tools suffer considerable limitations. For instance, they often ignore the synergistic interaction of risk factors, such as the combined impact of mental health and physical activity, together with demographic variables such as ethnicity (Dang et al., 2021). Based on differences in standard types of discrimination and inadequate diagnosis in different underrepresented groups (Dey & Salem, 2020), these methods do not generalize well for dissimilar populations. Such a context file must thus be strongly required in the UK, where the population is multifaceted in terms of ethnicity and lifestyle factors.

The growth of machine learning has opened a thriving landscape for model development that can learn from massive amounts of varied datasets in addition to being capable of detecting non-linear interactions among variables. Recently, evidence has demonstrated the value of ML in predicting cardiovascular risk with models that include Gradient Boosting and Neural Networks, providing better accuracy than conventional approaches (Gala et al., 2024). For example, (Al'Aref et al. 2018) indicated that ML models might improve the prediction of cardiovascular events when incorporating imaging data with clinical variables, with AUC reaching values of up to 0.85. Yet there exist barriers to the adoption of ML in real-life clinical settings, such as "black box" prediction by most models plus a lack of tools that could seamlessly fit into the workflows of healthcare professionals. (Dey & Salem, 2020)

1.2 Research Gap or Opportunity

Although noteworthy strides have been made in ML-based heart disease prediction, there are remaining deficiencies that this project aims to fill. First, almost all models do not quite love interpretability, which in healthcare becomes a very important requirement since clinicians need to unravel inferences made in order to contact their decisions (Gala et al., 2024). Although SHAP and LIME-based methods have been proposed for furthering interpretability, the tools are not yet available for heart disease prediction (Dey & Salem, 2020). This is what this project is tracing by utilizing SHAP to provide global and local explanations concerning the model prediction such that health professionals understand the key risk factors for each prediction.

Current heart disease predictive modelling applications also seem rather silent on fairness and bias, particularly among a mixed population like the UK. Studies have shown that some ML models might exhibit possible biases against certain demographic groups such as ethnic minorities or older adults as a result of imbalanced datasets or unrepresentative training data (Dang et al., 2021). Thus, this project will also quantify possible bias through the adoption of fairness metrics like Demographic Parity and depiction of disparities across demographic categories (e.g., sex, age, race/ethnicity) on bar plots in the Flask app. The focus on fairness ensures equal predictions by the model consistent with the ethical lines provided in the project proposal, such as compliance with GDPR and the BCS Codes of Conduct.

Lastly, one more serious contender toward this obsolescence of technological evolution is the issue of whether ML tools can be accessed by non-specialized users like doctors, a problem greatly aggravated by the fact that many solutions available on the market today are complex, requiring deep technical expertise to operate. As such, Al'Aref et al., (2018) highlight that this lopsided adoption of different ML technologies for operating on clinical settings is far from being generalizable. Closing this gap, this project proposes a user-friendly web application based on Flask, with all actions-simple data input, prediction, and interpretation. Among these app features are real-time predictions, SHAP-based explainability plots, and fairness dashboards for healthcare practitioners' use.

1.3 Problem Statement

Heart disease screening in common medical practice is rather reactive and does not offer early warnings, which could sometimes avert serious complications. Conventional options of assessing risk, such as the Framingham Risk Score, employ fixed thresholds for risk factors and do not regard the complex interplay of demographic, clinical, or lifestyle variables for different populations, including the UK (Dang et al., 2021). On top of that, a lacking interpretation in many ML algorithms limits their usefulness in health

care practices that require transparency in decision-making. Furthermore, this complexity makes the interpretation of these ML tools almost impossible for the non-technical user, mainly healthcare professionals, who require straightforward and usable insights to improve patient outcomes.

Aim: This project seeks to address these challenges by developing a smart heart disease risk indicator, one that exploits machine learning and explainable AI that is human-centred. The model uses a heterogeneous dataset to obtain a good representation of different risk factors, and feature selection techniques reduce the computational complexity while remaining accurate. The Flask app ensures that input, prediction, and interpretation of data is a seamless process so that it remains sufficiently accurate and conveniently usable in clinical practice. In addition, the project examines bias in predictions to assure fairness toward various demographic groups especially pertinent ethical considerations in healthcare applications.

1.4 Research Question(s)

In what ways may machine learning techniques be applied in the crafting of an intelligent and interpretable tool that predicts risk for heart disease, which subsumes superior accuracy and usability when compared to existing approaches?

The research inquiry shall be answered using the sub-questions below.

- Which machine learning models (Random Forest, XGBoost, Neural Networks) give the best performance when it comes to heart disease risk prediction in terms of accuracy, precision, recall, ROC-Score and F1 score and particularly ROC-Score?
- Can explainable AI frameworks like SHAP enhance interpretation for healthcare professionals in using the heart disease risk prediction output to arrive at actionable decisions?
- How do we design user interfaces to cover the divide between technical ML models and practical healthcare deployment?
- What are the possible demographic biases in heart disease risk predictions, and how might they be mitigated by using fairness metrics?

The questions fit the project in its primary goal as preparation, analysing datasets, developing evaluation ML models, implementing an explainability framework, designing a user interface, and envisaging bias dispersed along the inquiries.

1.5 Research Aim and Objectives

Aim: To develop an intelligent, transparent, and machine learning-based heart disease risk indicator with a user-centric interface for real-time predictions and insights.

Objectives

- To preprocess and analyse publicly available datasets relevant to heart disease risk prediction.

- Develop machine learning models (e.g., Random Forest, Gradient Boosting, Neural Networks) for predicting heart disease risk.
- Implement explainable AI frameworks (e.g., SHAP, LIME) to enhance model interpretability.
- To evaluate the performance of models using metrics such as accuracy, precision, recall, and F1-score.
- To design a user-friendly interface for data input and real-time risk prediction.
- To investigate potential biases in model predictions against demographic groups using confusion matrices and fairness metrics.
- To provide actionable insights based on predictions to aid healthcare decision-making.

1.5 Motivation for Topic Selection

The selection of this topic is motivated by a couple of factors, both personal and domain relevance, in the field of heart disease prediction in healthcare. I am pursuing an MSc in Artificial Intelligence with Advanced Practice; thus, I am interested in AI application to solve worldly problems, especially in the healthcare sector, where technology directly impacts improving lives. Heart disease is a leading cause of death worldwide and in the UK area where AI intervention can make meaningful contributions in early detection and prevention (Mozaffarian et al., 2016).

This project domain is relevant in many ways. To begin with, heart diseases are quite prevalent in the UK, thereby necessitating the need for better diagnostic tools. With the NHS spending billions year after year to treat heart diseases, a predictive tool that encourages early intervention will save on these costs and improve health outcomes (British Heart Foundation, 2024). Secondly, the project aligns with the growing trend of AI adoption in healthcare, as evidenced by recent studies highlighting the potential of ML in cardiovascular risk prediction (Gala et al., 2024). This project would, therefore, consider interpretability and fairness issues- important bottlenecks for the use of AI in clinical practice- making it much more relevant to the world of medical AI.

The cost of the project also makes it very much applicable as a Master's level challenge, with its machine learning, explainable AI, and user interface design integrated components. One is expected to have a comprehensive understanding of AI concepts as well as data science techniques and principles of human-computer interaction to complete the demise of the predictive model, ensure its interpretability, and design user-friendly interfaces. The learning outcomes of this master's program fit well with what the MSc Artificial Intelligence program at Teesside University seeks to emphasize: applying artificial intelligence to solving real-world practical problems within an ethical and societal context.

Lastly, it is a commitment to developing ethical AI, which has a central place in healthcare, where biased prediction creates divergence in patient care (Dang et al., 2021). The project thus contributes to the much larger objective of creating trustworthy AI systems that will integrate into healthcare workflows providing benefit to patients as well as practitioners.

Fairness and bias mitigation is finally a matter of ethical commitment concerning AI development, which stands out as especially important in healthcare because biased predictions might imply very different outcomes for the patients (Dang et al., 2021). It is good to see this kind of project that promises to meet a much larger goal: the development of trustworthy AI systems whose interventions in the workflows of healthcare can lead to benefits for patients as well as practitioners.

1.6 The Project Scope

The ambit of the project entitled "Intelligent Heart Disease Risk Indicator with Machine Learning" is tightly called for, such that the objectives are not designed to extend beyond 12 weeks for the realisation of the Master's project, from February to May 2025. The project focuses on an attempt to introduce machine learning (ML) techniques to predict heart disease risk with utmost emphasis on interpretability, fairness, and usability. Thus, coupled with explainable AI (XAI) frameworks and a user-friendly interface, the project intends to build a deployable tool to help healthcare professionals with the early detection and prevention of heart disease.

- **Defining the Scope:** The development, testing, and implementation of an intelligent decision support system or machine learning model for heart disease risk assessment have to cater to the needs of the healthcare providers in the UK. It will involve a literature review, the collection and preprocessing of patient data (e.g., types of chest pain, maximum heart rate), from repositories such as UCI Heart Disease, modelling and training (logistic regression, random forest, XGBoost), evaluating it with metrics such as accuracy, precision, recall, ROC curve, provide a user interface for healthcare professions, and finally, host in use of Flask for the web interface.
- **Boundaries**
  - **Geographical Focus**: The model is intended to be general. The project aims to address different aspects of the people of UK, which is further considering demographic diversities such as race/ethnicity and health issues that for one or another reason are UK-specific, such as the ones reported by NHS and British Heart Foundation.
  - **The Dataset Limitations:** The project restricts itself to the publicly available datasets and does not involve primary data collection or collaboration with health care institutions. These rules include compliance with ethics, like that of GDPR, but deprives proposal inclusion of real-time or proprietary clinical

data into the project works. Model Scope- The binary-classification project does not extend to multiclass classification (for instance, predicting specific types of heart disease) or regression tasks (similar to predicting the severity of the heart disease). User Base- The end users of the Flask application are not patients but healthcare professionals, who should interpret and act on the prediction. The interface is suitable for non-technical users, excluding some advanced features such as dashboards that patient accesses or EHR integration.

- **Deployment**: The application is a local prototype hosted with Flask (i.e., at http://127.0.0.1:5000/) and not built to any production level for deployment within a clinical setting under the project timeline. Further work can be added for consideration regarding cloud deployment and incorporation with health care systems in general.

1.8 Deliverables, Milestones, Assumptions and Limitations

The project proposal contained several anticipated deliverables, all of which were delivered by following through on a meticulously planned 12-week schedule. These deliverables included:

- **Deliverable**: A holistic technical report (this document): On the whole, the document covers everything from dataset analyses to preprocessing, model development, fairness analyses, implementation of interfaces to evaluations, and secures sections on introduction, methodology, results, discussion, and conclusion-providing a deeper picture of both success and challenges of the project. It is essentially a full-fledged requirement from the project lifecycle. -A Jupyter Notebook was used as an IDE for code development for data preprocessing, model development, training, evaluation, explainability analysis, and flask web app. Visual Studio was used to build interface code, such as results.html, index.html, and CSS to style the interface. The interfacing of heart disease risk prediction will be made available as an operational prototype using Flask. A professional presentation that includes: presentation of a machine learning model, its accuracy, and the portability of the Flask application. This also includes slides on the project background, methodology, results (like model performance metrics, fairness analysis), and a live demo of the Flask app making real-time predictions and interpretability features.
- **Milestones** are:
    1. Literature review and data collection (Week 1-3).
    2. Data preprocessing and feature selection (Week 4).
    3. Model implementation and training (Month 5-6).
    4. Performance evaluation and comparison (Month 7-8).
    5. Development of user interface and final report (Month 9-12).

- **Assumptions**:
    - **Data Quality & Representativeness**: It was assumed that the Kaggle heart disease dataset is representative of the UK population. Randomness in missing values allowed imputation to be performed without introducing bias.
    - **Model Generalizability**: The XGBoost model was held to epitomize generalization due to a stratified train-test split and SMOTE.
    - **User Proficiency**: The Flask app targets healthcare professionals with very basic computer literacy; the interface was therefore kept as simple as possible.
    - Ethical Compliance was maintained as the project is GDPR-compliant, comprising the use of a public dataset with no personal health data.
- **Limitations**
    - **Dataset Constraints**: The dataset may not fully represent the UK population due to a lack of real-time clinical data as well as some underrepresented demographic groups. Ideally, future work should include NHS datasets.
    - **Model Complexity with Interpretability**: SHAP tried to help explainability, but XGBoost is still complex. A more interpretable model, Logistic Regression, may be limited in prediction accuracy but is easier for the end user to relate to.
    - Limitations of the Interface: There is no authentication or encryption and no EHR integration, which is lacking in the prototype. Some questions asked of the user for input may confuse the user since not all of the features bear directly on the predictions.
    - **Time & Resource Limitations**: The project has had 12 weeks to spare. That severely limited hyperparameter tuning and the inclusion of other ML models, as well as computational resources to carry out deeper exploration.
    - **Bias and Fairness**: There was some fairness analysis done based on Demographic Parity based on the three demographic features. Time constraints were detrimental to carrying out more general bias mitigation work. The above constraints notwithstanding, a working heart disease risk indicator was delivered by the project, improving diagnostic tool accuracy, interpretability, and usability.

2.0 Literature Review

Heart diseases are still among the top causes of mortality worldwide. Thus, the requirement for more advanced diagnostic and predictive tools arises for better early detection and treatment outcomes (Gupta & Rathi, 2023). The integration of ML techniques into medical diagnostics has emerged as a promising avenue for increasing both the accuracy and efficiency of heart disease predictions. This review sums up findings from recent studies that investigate several different ML algorithms, feature selection methods, and new approaches such as federated learning and wearable technology for heart disease detection, referring to a wide range of papers published from 2017 to 2024.

2.1 Overview of ML Applications in Heart Disease Prediction

An application of machine learning in the prediction of heart disease has been a topic of inquiry using multidimensional datasets, including Cleveland, Hungarian, and Statlog datasets. Rajendran et al. (2021) compared supervised and unsupervised learning techniques used in predicting cardiovascular heart disease. The models tested are Logistic Regression, Naive Bayes, Random Forest, and Artificial Neural Networks (ANN). Evaluation involved F1-score, precision, recall, and area under curve (AUC). The results showed that Logistic Regression outperformed other models on Hungarian and Statlog datasets. On the other hand, ANNs exhibited higher accuracy on the Cleveland dataset. Interestingly, Logistic Regression consistently achieves the highest AUC scores indicating robustness in risk assessment. The study also provided evidence regarding the effect of feature correlation on model performance, suggesting that careful engineering of features may suit effective prediction.

Similar results are shown by Zhang (2023) where Logistic Regression and Random Forest are compared with a dataset of 13 features. Both models predicted accuracy values of 82.47% and 99.03%, respectively. While most features analyzed by Random Forest consistently affect the model outputs such as "cp" (chest pain), "thal" (thallium stress test), and "oldpeak" (ST depression), feature selection is vital in boosting prediction power. That fact also creates a demand for better ML to enhance health risk assessments, and future projects are advised to prove these in large databases.

The risks of heart diseases were predicted using decision tree algorithms Like CART, C5.0, M5P, and XGBoost in their research conducted by Shanthi Kunchi, et al. (2023). The algorithm that marked highest performance was the XGBoost in terms of 96.32% accuracy suited for highly complex medical diagnostics. The study further evidences the utilities of decision trees in health data especially in the presence of many missing values through data pre-processing, feature selections and stratified cross-validation. In this case, XGBoost can be integrated into clinical practice to make diagnosis much faster and accurate.

2.2 Feature Selection and Sensitivity Analysis

Feature selection plays a key role in optimizing ML models for heart disease prediction. G Saranya and Tyagi (2024) extensive study demonstrated how the exercise-related features of exercise hours per week and average calories burned daily concretely influenced CHD risk. They applied SVM with the support vector machine algorithm using the RBF kernel, KNN methods, and Random Forest algorithms, which performed better than other classifiers. The negative correlation between physical activity and coronary heart disease risk was supported, yet it indicated that too much exercise could be counterproductive for older patients, citing age-constrained data for better predictions.

Tiwari et al. (2023) introduced another interesting perspective in which HRV signals were ranked based on spectral and nonlinear investigation methods. From the 30 features derived from HRV time-series signals, Tiwari et al. excluded ten by ranking methods such as Fisher score and ROC analysis. The features obtained from the third-level MSWP decomposition and chaos investigation methods (e.g., CD, DFA:α_2, ApEn) displayed the lowest p-values, representing high discrimination ability. This study shows that ranked parameters can strongly support cardiac disease diagnosis in a clinical setting.

A dual cost-sensitive Random Forest algorithm was proposed by Wang and Tan and improved feature selection by sequential decision tree analysis with cost sensitivity. The procedure to enhance Heart Disease (HD) detection remarkably lowered the rate of misclassification obtained by Logistic Regression, SVM, and standard Random Forest methods and increased practical confidence in practical applications. Feature relationship identification thus offers an intelligent design to improve the predictive performance of the model.

2.3 Innovative Detection Techniques and Wearable Technology

Recent ML advances have made applications in unconventional settings. Mekni et al. (2024) developed HeartBuddy, an open-source mobile application that uses ML to find heart problems in ECG signals. HeartBuddy is a decision-support tool meant to complement a physician's expertise, showing a promise of reducing rates of misdiagnosis. This strategy ensures that everyone uses visual aids to enhance patients' understanding that there's a vision for mobile technology facilitating early detections.

Diao et al. (2024) conducted a pilot study to derive 6MWT data from smart shoes whose purpose is to derive inferences regarding heart failure (HF) in non-clinical settings. For the gait mapping to 42 medical indicators, machine learning algorithms reported relative errors in the uppermost seven of less than 5%. It presents a new view towards long-term heart failure monitoring without the requirement of specific equipment and expert involvement.

An ECG-less algorithm was developed, which annotates heart feature points using seismo-cardio-gram signals out of patients with valvular heart disease using multi-modal

sensor fusion. This algorithm achieved 98.94% precision and 97.44% recall, thus making it compelling for use as an out-of-clinic evaluation of valvular heart disease. It could transform remote monitoring by minimising reliance on manual annotation.

2.4 Comparative Performance of ML Classifiers

Research comparing ML classifiers has been done extensively. For example, Gaba et al. (2023) held one such study to ascertain the prognosis of cardiovascular diseases through several ML methods, giving Random Forest an edge with 98.53% accuracy, 99.25% precision, and 98.52% F1 score, thus laying more emphases on Random Forest in automated diagnostics. Lin et al. (2023) similarly evaluated six classic models—Logistic Regression, Random Forest, Decision Tree, KNN, SVM, and Neural Network—across four regions; a comprehensive evaluation with large interdependence on activities like data collection, preprocessing, and K-Means clustering was exercised. This study urged customized model selection depending on dataset characteristics.

According to Prakash Paudel et al. (2023), heart attack detection was investigated with the AdaBoost Classifier, Random Forest, Gradient Boosting Classifier, and the Light Gradient-Boosting Machine, where the latter placed highest at 99.33% in training accuracy. The use of Explainable AI with LIME pointed to "kcm" and "troponin" as key predictors, improving clinical interpretability. Anas Domyati and Memon (2022) ran tests with SVM and Random Forest on the Cleveland and the Hungarian datasets, recording actual promise due to the relief feature selection algorithm, stressing the necessity of feature optimization.

2.5 Federated Learning and Privacy-Preserving Approaches

Gaber et al.2024 publication presented FedCVD, which is a federated learning model of Logistic Regression with SVM and cardiovascular disease prediction. The use of distributed patient data also allowed for privacy features in FedCVD while giving an AUC of 0.7048 with SMOTE, which was higher than that of many other central models (AUC 0.6962 with Random Over Sampling). It thus addresses data imbalance problems and is scalable with respect to clinical applications.

2.6 Lifestyle and Risk Factor Considerations

Kreider et al. (2017) indicated that creatine supplementation seemed safe and effective when used in exercise, sport, and medicine and showed some potential to reduce cardiovascular risk with the help of physical activity. This corroborates the evidence presented by G Saranya and Tyagi (2024), who state that exercise is inversely related to CHD, but caution against risks associated with age and, therefore, the need to personalize assessment for risk.

2.7 Gaps and Directions

In aggregating the reviewed articles, it is clear that machine learning (ML) algorithms enhanced ways of predicting and detecting heart disease. Random Forest and XGBoost both give very high accuracy levels (up to 99.33%), while Logistic Regression is the best performer on the area under the curve (AUC) (Rajendran et al., 2021; Zhang, 2023). These feature selections like relief and ranks parameters tend to make the models robust with exercise and clinical parameters "troponin" identified as key predictors by G Saranya & Tyagi, 2024; Prakash Paudel et al., 2023. Innovative tools such as HeartBuddy and smart shoes have broadened diagnostics even in nonmedical settings, whereas federated learning addresses privacy concerns (Mekni et al., 2024; Diao et al., 2024; Gaber et al., 2024).

But, for all these improvements, many gaps remain in existing literature, and this study intends to fill them. The good diversity appreciated in the data used for this research, such as parameters associated with physical activity (hours per week of exercise, calories burned), medical issues (ECG signals, clinical determinations like troponin), and indicators regarding lifestyle choices (age, sex, race/ethnicity), could very well represent heart disease risk factors compared to many existing studies. Such an approach would eliminate the limitation of dataset neglect noticeable in earlier works (Rajendran et al., 2021; Zhang, 2023), which often relied on specific datasets like Cleveland or Hungarian without elaborating the multifaceted nature of heart disease.

Fairness and interpretability concerning ML models for heart disease prediction have generally received little attention in prior literature. The present work addresses this gap by enforcing fairness across different demographic groups (such as sex, age, and race/ethnicity) by carrying out fairness analysis inspired by the interaction of the code, which assesses metrics like accuracy, precision, recall, and ROC-AUC across different subgroups. In doing so, this makes sure the same level of performance is achieved across different populations-an aspect missed in studies such as that by Shanthi Kunchi et al. (2023) and Gupta & Rathi (2023), which focused mainly on overall accuracy without stratifying by demography.

In addition, the inability of the model to be interpretable under clinical settings is considered an indispensable shortcoming in the literature. The present study responds to this by using SHAP (SHapley Additive exPlanations) values, as integrated by the code, to provide an interpretable insight into feature importance (e.g., "chestPain," "HadArgina," "medicalHealth"), supporting the findings by (Prakash Paudel et al. 2023) on key predictors, thus enhancing trust for application by healthcare professionals-such a shortcoming being acknowledged in the literature activities such as Lin et al. (2023) by having ignored explainability.

A further gap is that they were not made available to end-users' predictive models or tools, such as those now emerging under names like HeartBuddy (Mekni et al., 2024) and

smart shoes (Diao et al., 2024). This study thus has created a web interface, as has been shown in the code, where individuals may insert their exercise habits, medical history, and other personal data and get real-time prognosis from visualisations such as confusion matrices and ROC curves. This is with respect to some of the generally static approaches used in Gaber et al. (2024) and Wei et al. (2024), who worked on backend model building minus interaction with users.

Further, modelling effect variation using “RandomizedSearchCV” has optimised the hyperparameters of the XGBoost model used in this code. This inconsistency in model tuning is observed within and across studies, for instance, Wang & Tan (2022); Anas Domyati & Memon (2022). With this optimisation and the diverse dataset, this study would offer a more robust and generalizable solution.

3.0 Methodology

The project uses all-encompassing ways to predict heart disease from a data-rich collection that consists of more than 246,000 samples and 39 properties related to them. The main process involves the exploratory data analysis followed by systematic cleaning, visualisation, and statistical analysis. While doing so, systematic training and evaluation as well as comparison of machine learning models based on performance accuracy and fairness metrics are carried out. Eventually, the best model will be integrated into an online application based on Flask, serving interpretable predictions and visually explained information. The methodology secures every step with programming precision for high-quality data governance standards and makes the project have ethical and transparent AI approaches.

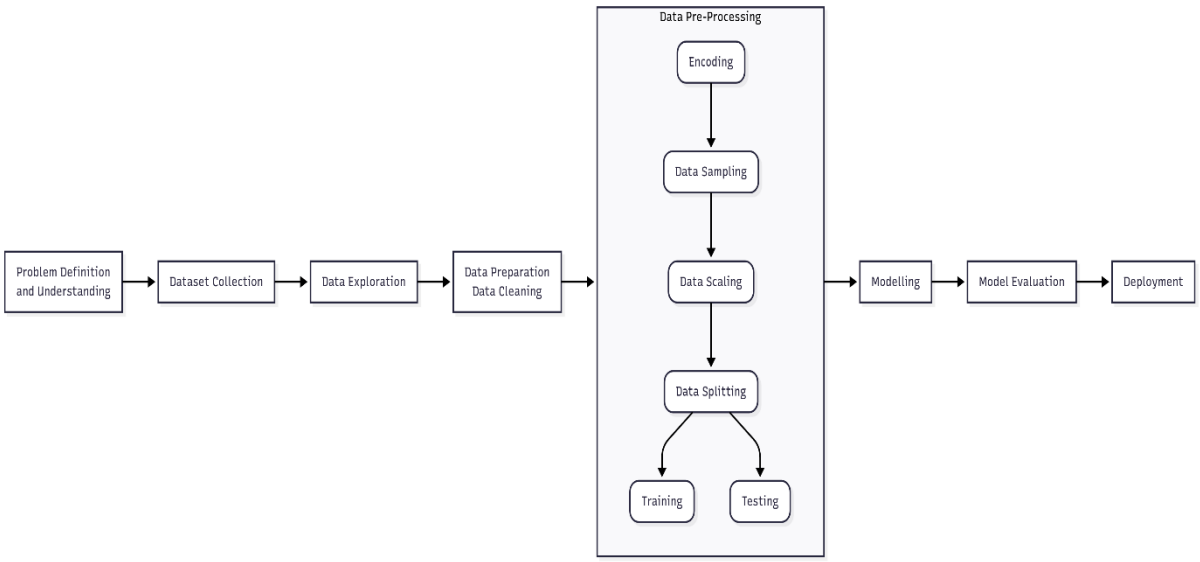


Figure 1: Proposed Model Diagram

3.1 Exploration of Data Analysis

3.1.1 Dataset Acquisition and Description

- **Source and Characteristics**: The initial consideration about the data was its source, being a public repository like Kaggle or UCI. The dataset contains 246,022 observations and 39 features representing a wide variety of health indicators, including general health, number of days of physical and mental health, sleep hours, body mass index (BMI), with some clinical events such as angina and stroke. Such variable diversity makes it an excellent opportunity for predicting heart diseases. The project desired to emphasise the class imbalance of the target variable, that is, whether an individual has heart disease or not.

Table 1: Description of some Attributes of Heart Disease Dataset

Features	Descriptions
HeartDisease	Respondents who have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).
BMI	Body Mass Index (BMI).
Smoking	Have you smoked at least 100 cigarettes in your entire life? (The answer Yes or No).
AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
Stroke	(Ever told) (you had) a stroke?
PhysicalHealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days).
MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).
DiffWalking	Do you have serious difficulty walking or climbing stairs?
Sex	Are you male or female?
AgeCategory	Fourteen-level age category.
Race	Imputed race/ethnicity value.
Diabetic	(Ever told) (you had) diabetes?
PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their 14.. regular job.
GenHealth	Would you say that in general your health is...

SleepTime	On average, how many hours of sleep do you get in a 24-hour period?
Asthma	(Ever told) (you had) asthma?
KidneyDisease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
SkinCancer	(Ever told) (you had) skin cancer?

- **Data Governance**: Thorough consideration was devoted to issues of data privacy and licensing in order to conform to regulations and ethical standards. The dataset was assigned anonymisation and data governance protocols. The project also truly explores the extensive documentation of the processing steps undertaken, once again for the purpose of offering information regarding changes made during cleaning and transformation.

3.1.2 Preprocessing and Cleaning

- **Handling Missing Data and Duplicates**: No missing values were found in any of the important features after going through the data set; any missing numerical columns would be imputed by the median and categorical columns by the mode. Duplicate records were also identified and removed to avoid repeated measures per observation in the analysis. Individual string variables, such as the age category, were cleaned and standardized through some customizable manipulation steps.
- **Feature Consistency Checks**: Consistency checks were conducted on the dataset, including BMI recalculation using the formula $(WeightinKilograms/HeightinMeters)^2$. Any differences between the original and recalculated BMI values were conditionally replaced with original values. Such checks marked the importance of ensuring that the obtained features represented the correct intended measurements.

3.1.3 Data Visualisation Techniques

The actual analysis of the data employed several visualization techniques to show important characteristics and thus support further analysis and model selection.

- **Correlation Heatmaps**: The heatmap in (figure 2) shows the strength as well as the direction of the correlations using a colour scale ranging from dark blue to light blue. A quick glance shows a strong diagonal line, which is expected (each variable being perfectly correlated with itself). Some major observations indicate that there is a positive correlation among "PhysicalHealthDays",

"MentalHealthDays", and "GeneralHealth." That would also leave likely correlation between "BMI", "WeightInKilograms," and "HeightInMeters." Correlation among most other variables looks relatively weak.

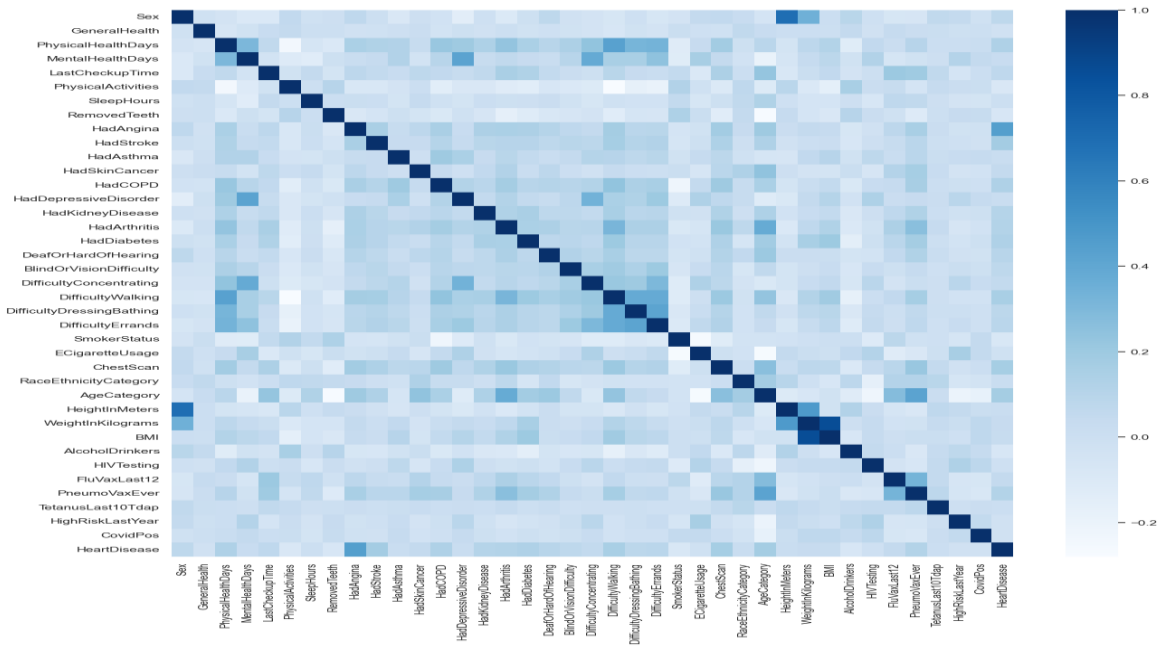


Figure 2: Correlation Heatmap

- **Heart Disease by Sex**: This figure 3 swiftly divulges that while heart disease is less prevalent among both genders, a larger share of men suffer from heart disease than women, suggesting greater risk associated with being male among this studied cohort. This increased risk may arise from biological predisposition, lifestyle differences, or societal disadvantages. Further, in contrasting the counts with respect to sex, the visual underscores how risk factors and disease prevalence can vary quite meaningfully based upon gender.

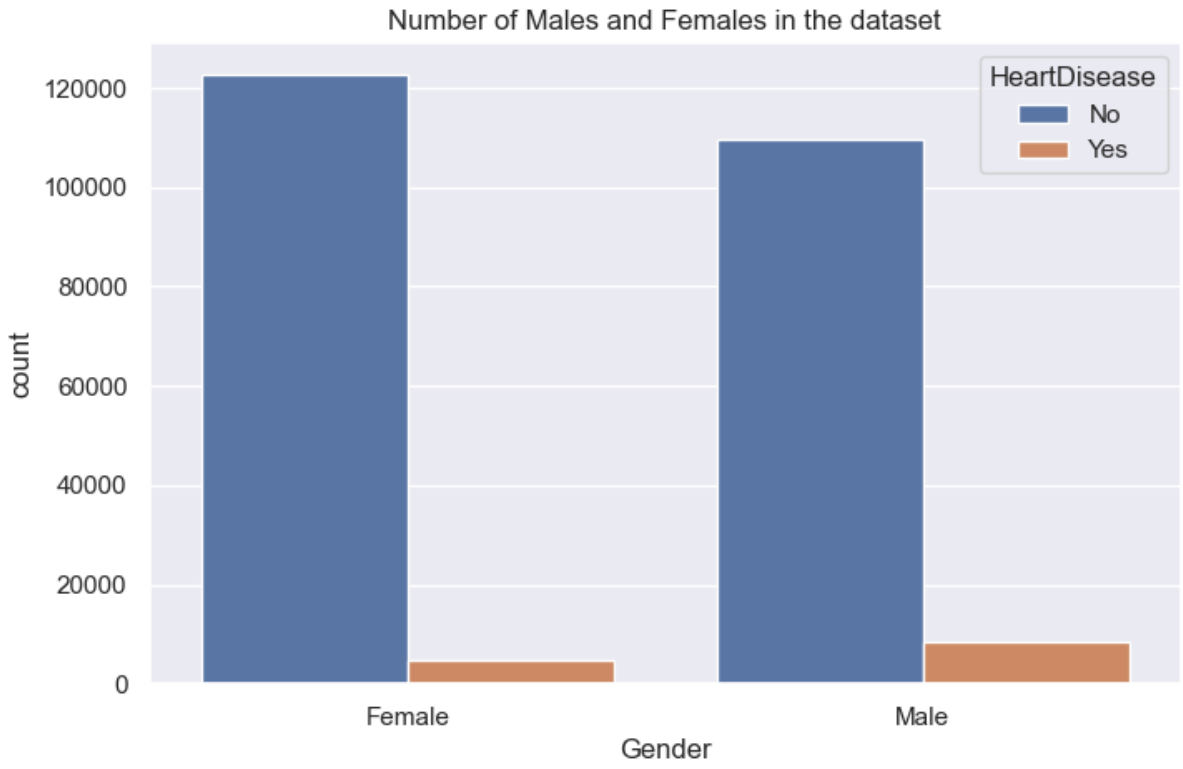


Figure 3: Bar chart for Gender Against Heart Disease

- **Heart Disease by Race/Ethnicity**: The chart (figure 4), makes clear that bigger populations, such as the White non-Hispanic population, have more overall cases, while relative risk for smaller groups could be different. It suggests that race/ethnicity contributes at least partly to the heart disease prevalence possible through genetic factors, socio-economic status, or access to care. Moreover, this disaggregation would hint at intersectional differences because when you compare the raw counts and eventually the ratios, then you can see that not all groups are affected the same, even when absolute figures are larger in some categories.

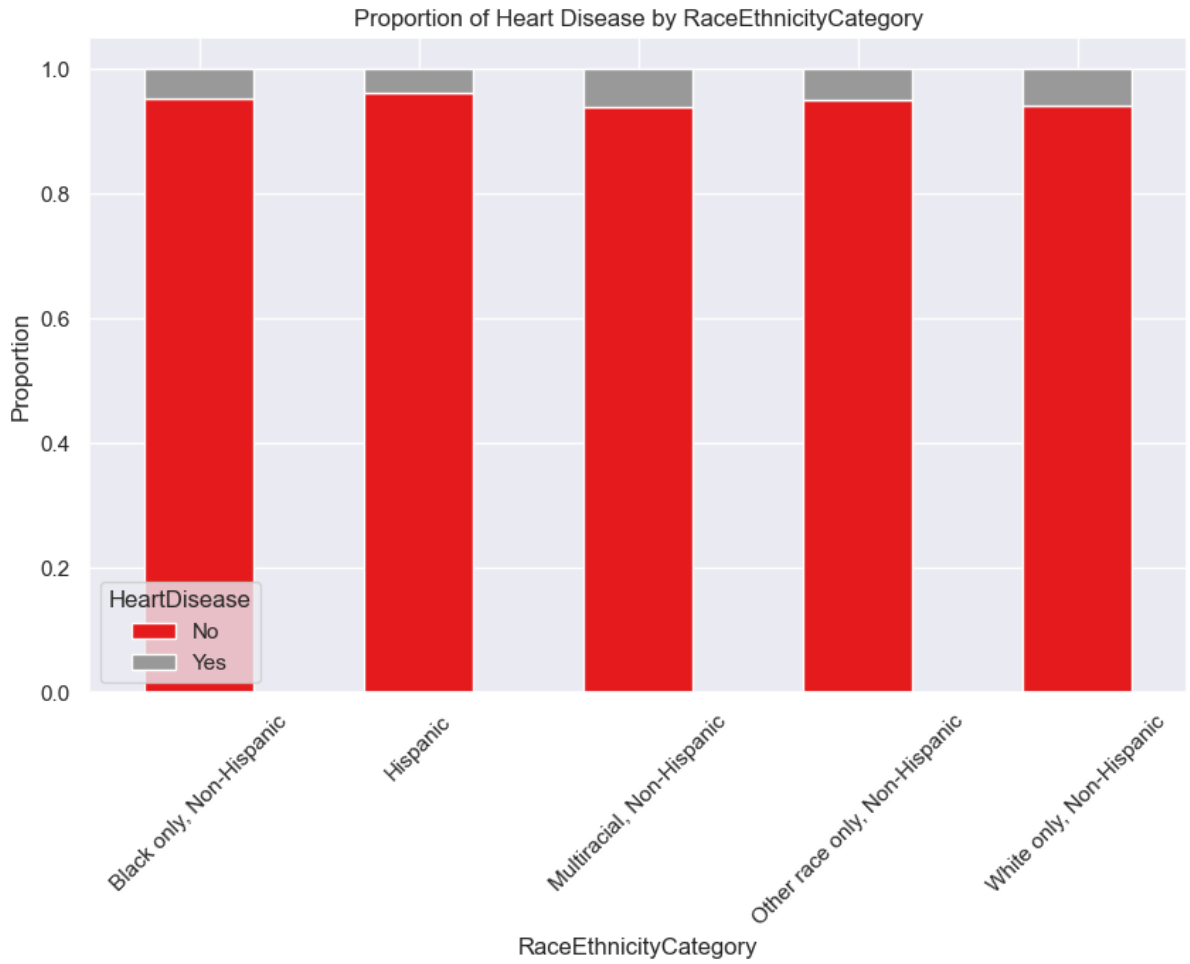


Figure 4: Show Ethnic Group Vs Heart Disease

- **Heart Disease by Race/Ethnicity and Sex**: Figure 5 shows manifest some differences in heart disease prevalence among different genders and racial groups. Male heart disease incidence is noted to be higher, especially among Black-Hispanic individuals. Female instances of heart disease, on the contrary, are noted more among the White and Multiracial groups. This, therefore, shows that heart disease risk is significantly influenced by gender and race, calling for targeted prevention and treatment approaches.

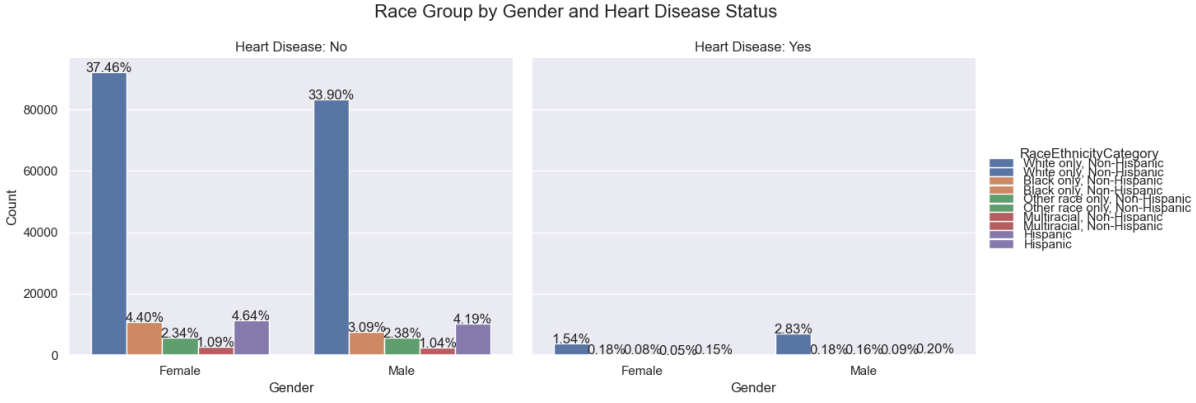


Figure 5: Racial Group with Gender and Heart Disease

- **Heart Disease, Medical Factor, and Lifestyle**: This figure 6, show details that have an influence between Heart disease and various health profiles through the comparatively many violin plots and histogram data: The plots of BMI and WeightInKilograms correlate individuals having heart disease (orange) against those who do not (blue) based on higher median BMI and weight associated with the observation of a wider distribution which indicates a greater variation in these outstanding measures from heart disease patients. It is the SleepHours plot which shows that, as was indicated by heart disease patients, the sleep hours were generally less, with a median of around seven hours compared with a median of eight hours for those without the disease and with more variation in sleep duration. The MentalHealthDays and PhysicalHealthDays plots reveal that patients suffering from heart disease usually report way more days where mental and physical health conditions fall under median ranges of around 10-15 days as opposed to almost 0 for those not suffering from the condition as well as a much wider span. The Age Category vs heart disease histogram found a well-defined pattern; as heart disease increases, so does the aging factor; the highest in counts falls mostly in the 65+ age categories, especially 80+, as contrasted against younger segments (18-34) wherein there are almost no cases. The "No" group (blue) dominates across all ages, but the proportion of "Yes" (orange) grows steadily in older age categories. These findings are so strong that they suggest that heart disease involves high BMI, weight, poor sleep, poor health regarding challenges faced, and older ages. They sit among the most important risk factors regarding the broader impact of heart diseases on life.

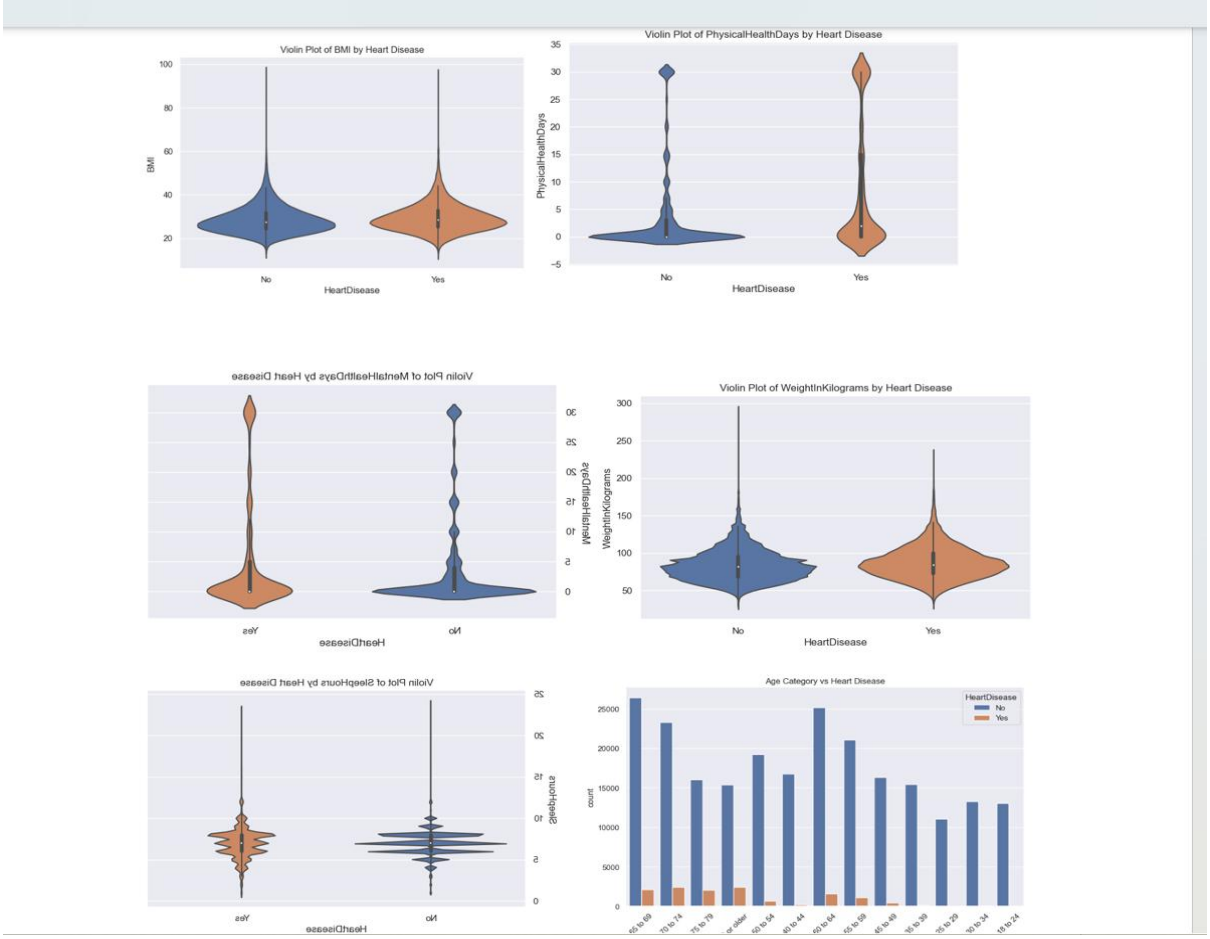


Figure 6: Heart Disease with medical and Lifestyle factors

- **Heart Disease and Behavioural Factors**: Smoker Status and Alcohol Drinking: Here, the view (figure 6) shows how behavioural choices impact the risks of heart disease. Current smoking, especially daily smoking, appears to confer a high association with heart disease rates; even though alcohol consumption is factored in, the combination of smoking and drinking seems to chalk up higher counts in heart disease cases, reinforcing the idea which state that not only one, but the interaction between lifestyle habits contributed most to cardiovascular risk. It positions smoking and drinking as behaviours that require attention in prevention strategies.

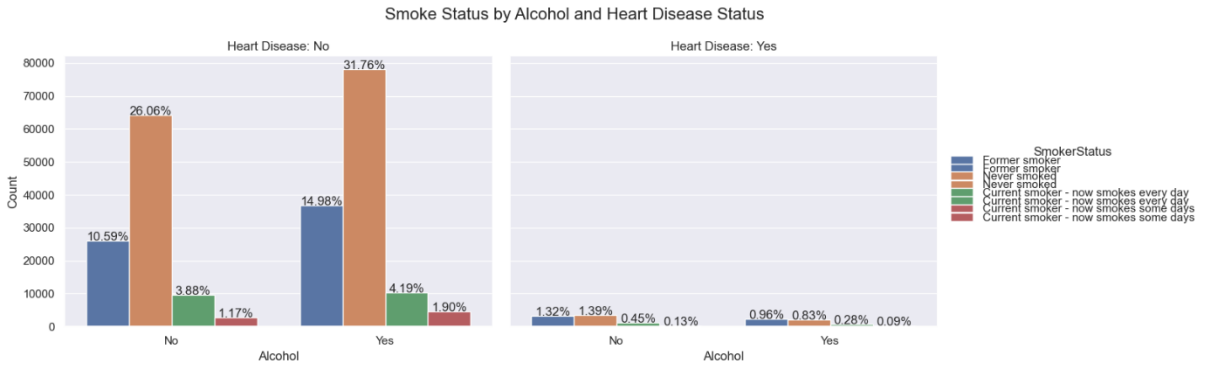
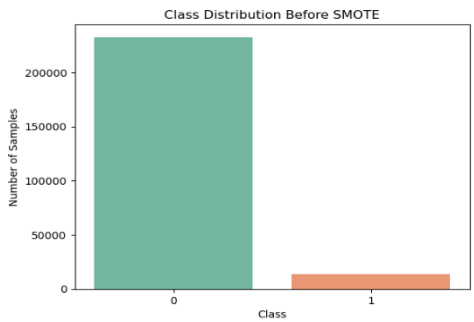
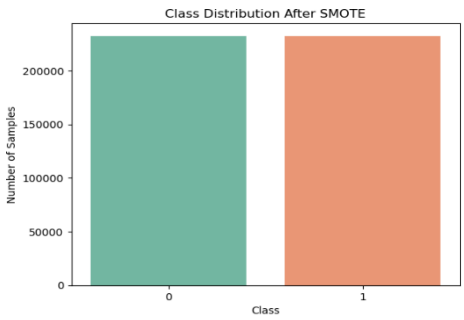


Figure 7: Heart Disease with Behavioural Factors

3.2 Class Balancing and Dimensionality Reduction (SMOTE and PCA)

- **Dataset Imbalance**: From Figure 8, the initial data shows a pronounced imbalance. For example, the number of individuals without heart disease greatly exceeds those with heart disease in every category (sex, race/ethnicity, etc.). This imbalance indicates that if you were to run predictive models on the unadjusted data, they might become biased toward predicting the majority class. Prior to applying PCA, the dataset may include numerous features (demographics, self-rated health, activity levels, smoking status, and alcohol consumption) that could be intercorrelated.
- **Balancing the Classes**: SMOTE (Synthetic Minority Over-sampling Technique) is applied to generate synthetic examples for underrepresented classes—in this case, individuals with heart disease—to balance the dataset. After SMOTE, a bar chart comparing the counts of "Yes" and "No Heart Disease" would show a more balanced set of bars. This means that the minority class (heart disease cases) now has counts that are more comparable to the majority class. The overall visual narrative now gives each class a more equal footing, which is crucial for training unbiased and more robust predictive models. The use of PCA distils the essential variance in the data into fewer dimensions, which allows for cleaner, more interpretable visuals. These post-PCA visuals reflect the most critical underlying patterns and relationships, providing clearer insights into how different factors combine to influence heart disease risk.





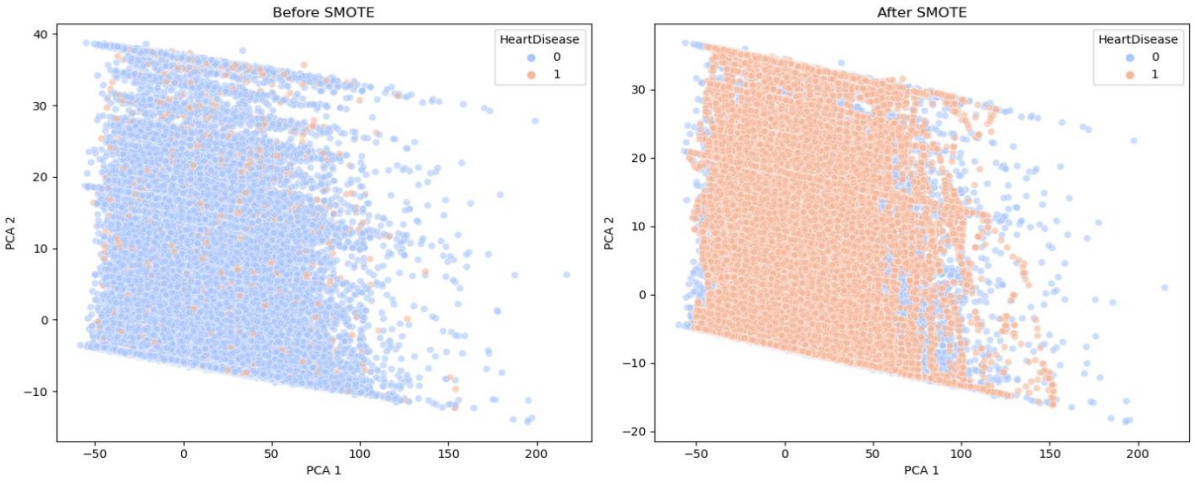


Figure 8: Handle Class imbalanced (SMOTE and PCA)

3.3 ML Workflow

3.3.1 Data Pipeline Overview

The entire ML workflow was completed in an orderly pipeline, consisting of data preprocessing, feature selection, splitting the data, training of the model, evaluation, fairness assessment, interpretability, and deployment. Post preprocessing, the data balanced out; feature selection was performed with the help of mutual information in a way that had selected 20 top features for predictive power and demographic representation (e.g., "Sex", "HadAngina", "AgeCategory"_"Age 80 or older"). Categorical features were converted into numerical codes using label encoding methods. These transformations were vital to ensure all data was in a compatible format for machine learning algorithms. The data was split into train (70%), validation (15%), and test (15%) using stratified sampling to maintain the class and demographic proportions since these are standard practices for heart disease prediction [Shanthi Kunchi et al., 2023]. This approach prevented bias and ensured a reliable evaluation during cross-validation.

3.3.2 Model Selection Strategy

A series of models has been developed for performance comparison, where consideration has been given to the entire range of learning paradigms. This approach will then enable the selection of the most suitable model for heart disease prediction according to very strict evaluation criteria.

- **Logistic Regression:** Serves as a great beginning point for heart disease prediction because it can be simple and interpretable as an appropriate model for binary tasks, like predicting whether a patient has heart disease (1) or not (0). Given the mixture of medical (e.g., cholesterol levels), physical (e.g., BMI), mental health (e.g., stress scores), and lifestyle factors (e.g., smoking), it assumes a linear relationship between these attributes and log-odds of the outcome, making it computationally efficient. Its coefficients also afford insights into how each factor contributes to the risk, which has its significance in medical interpretation. It

would, however, have problems dealing with the complexity and nonlinearity of interaction likely to occur among variables, which is expected from such a diverse dataset.

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Formula 1: LG Equation

Where $P(y = 1 \mid X)$ is the probability of heart disease, β_0 is the intercept, β_i are coefficients, and X_i are the input features.

- **KNN**: Is chosen because it is a non-parametric approach which can discover very complicated patterns in the heart disease dataset with no particular dependence being defined between features and their outcome. It would classify a patient according to the majority class in feature space of their "k" nearest neighbours. In simple terms, it even uses diverse factors-such as physical measurements against lifestyle habits-to smell a rat. Its flexibility is useful when medical and mental health data could actually cluster patients into distinct risk groups. However, it will be sensitive to irrelevant features and will require careful scaling of things such as BMI or stress scores.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Formula 2: KNN Equation

Where $d(x, y)$ is the distance between two points, x and y, based on features x_i and y_i.

- **Xgboost algorithm**: This is preferentially chosen for the efficient handling of tabular data such as this heart disease dataset; it adequately models non-linear relationships and feature interactions. This pioneering gradient-boosted tree program sequentially builds the decision trees, honing in on areas where the previous trees made mistakes, which is essential to capture fine risk patterns across various factors, including medical, physical, mental, and lifestyle. Further, the model provides feature importance scores to figure out the real players, be it in terms of blood pressure or smoking. Its robustness regarding missing data and the ability to use regularisation to discard unneeded complexity makes it one hell of a candidate.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

Formula 3: XGBoost Equation

Where $l(y_i, \hat{y_i})$ is the loss function (e.g., log loss), $\Omega(f_k)$ is the regularization term, and f_k are the trees.

- **Random Forest:** Is chosen as it is an ensemble method that provides a lower tendency to overfit than a single decision tree and can thus be confidently applied

to a complicated dataset having diverse heart disease predictors. It treats heart disease variables as separate predictors and builds different trees, averaging every prediction based on random subsets of data and features. In this way, it learns expressed non-linear relationships and interactions. As far as heart disease predictors vary along the line of multiplistic and contradictory, we can expect classical noise in the data. Random forests will also be able to handle this well. Random Forests give feature importance rankings, which helps us to know which variables (cholesterol vs. stress) are most critical.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

Formula 4: RF Equation

Where $\hat{y}$ is the final prediction, T is the number of trees, and $h_t(x)$ is the prediction from the tree t.

- **An MLP:** Is selected because these neural networks are very suitably disposed for modelling extremely non-linear and complicated relationships across the diverse feature set presented here (medical, physical, mental health, and lifestyle). With many networks in layers, interconnected nodes learn the most complex patterns; for instance, how, together, stress and smoking can elevate the probability of heart disease. On the one hand, MLPs allow much flexibility and can operate efficiently with large data. However, on the contrary, with this ability, MLPs demand a lot of tuning and power, far more than simpler models. Being a black box in nature, it icing on the cake, and in medical parlance, this black box interpretation becomes a liability.

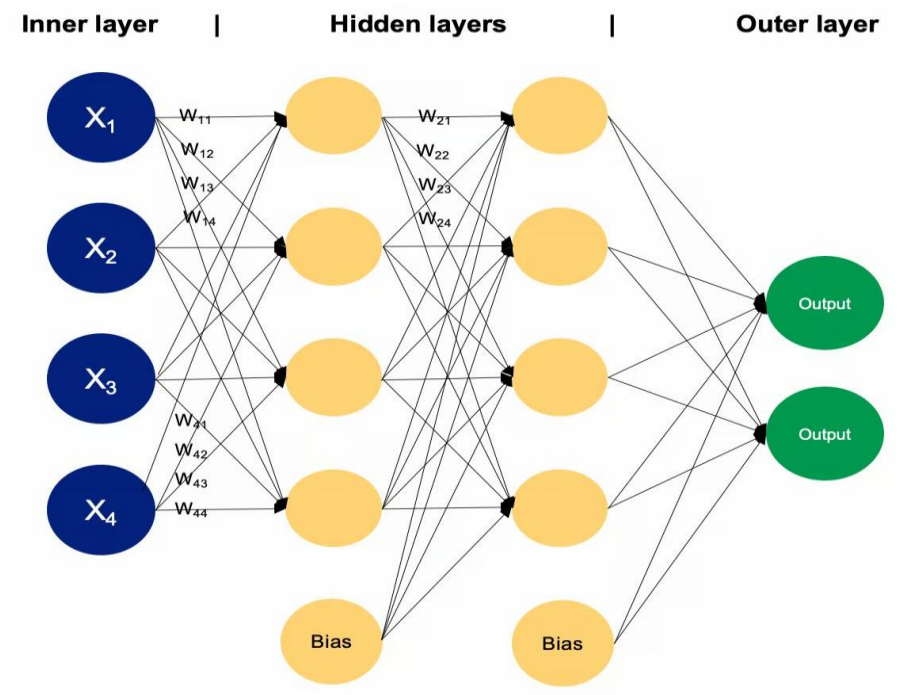


Figure 9: Neural Network MLP Diagram

$$h_j = \sigma\left(\sum_i w_{ij} x_i + b_j\right)$$

Formula 5: Deep Learning Equation

Where h_j is the output of neuron j, σ is the activation function (e.g., ReLU), w_{ij} are weights, x_i are inputs, and b_j is the bias.

3.4 Evaluation Metrics and Fairness Analysis

3.4.1 Standard Performance Metrics

- Accuracy or the total correctness of the model in predicting heart disease says how many patients were correctly classified as having or not having heart disease from all predictions. This is good for a general assessment in this case but can be often misleading if the dataset is imbalanced (e.g., too many healthy patients).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Formula 6: Accuracy Equation

- Precision indicates the reliability of positive predictions, i.e., the proportion of patients predicted to have heart disease who actually have it. Simply put, high precision for heart disease means that when the model flags a patient as diseased, this prediction is likely correct, enabling avoidance of unnecessary treatment.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Formula 7: Precision Equation

- Recall quantifies how well the model can identify actual cases of heart disease, that is, the proportion of diseased patients correctly flagged. It is vital in heart disease prediction that the recall is high to avoid missing any important diagnosis, which could be life-threatening.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Formula 8: Recall Equation

- F1-Score is the harmonic mean of precision and recall and provides a good balance between the two and is highly useful in the case where borne advantage and disadvantage when a patient is falsely diagnosed as a heart patient. However, it has proved to be explicitly useful when the cost of having both FN (missed diagnoses) and FP (unneeded interventions) matters in the case of heart disease prediction.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Formula 9: F1-Score Equation

- The term "Positive Rate" in the context of heart disease and ROC analysis typically refers to the FPR, which is the fraction of healthy individuals wrongly considered

to have heart disease. Important for interpreting the model's specificity (1-FPR) along with trade-offs in the ROC curves.

$$\text{FPR} = \frac{FP}{FP + TN}$$

Formula 10: Positive Rate Equation

- The Log Loss measures the model's uncertainty about the probability that a patient will have heart disease; it penalizes confident wrong predictions more than those less confident. For heart disease, it is quite valuable since we care about the best possible alignment between probabilistic model predictions (say, 0.9 for disease) and reality, thus adding in the stratification of risk.

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Formula 11: Log Loss Equation

3.5 Experimental Design

3.5.1 Establishing Baselines and Controls

Implementation of the Baseline Model: Logistic Regression was used as the treating method model to offer the reference point for the performance of the model. It was simple and easy to forge interpretation to benchmark models in the company with it.

- **Statistical Control Tests**: Chi square tests proceeded forward the relationship of categorical predictors in relation to target variable in significance. This proved also in significance observed (very low p-values), giving positive ground to features selected and discounts the variations in performance of the model occurring by chance.
- **Performance Comparison**: All models were measured against all performance standards-defined measures: accuracy, precision, recall, F1 score, AUC-ROC and log loss into their definitions. From this, one gets a max-after-max picture of every model's predictive performance, enabling one to tell which one handles well the conflict between sensitivity and specificity.
- **Fairness Issue**: Fairness from the prediction side was determined using calculated demographic parity and subgrouping analysis on key protected categories such as sex, age, and race/ethnicity. Data processing of analysis of confusion matrices and ROC curves for different subgroups using the private protection level indicators were used for the detection of any bias or disparate error rates. Considerable performance equivalence among the various subgroup members was taken into consideration when developing the final model.
- **Experimental Reproducibility**: The experimental design consisted of iterations in making the model trained, hyper-tuned, and validated again: every stage

documented, and standardized evaluation pipelines made sure results across runs can be compared with reproducibility and consistency.

3.6 Model Development and Deployment

3.6.1 Model Architecture and Training Process

- **Model Training and Validation Strategy:** Each model was trained on the SMOTE-generated balanced dataset and validated on a holdout dataset to ensure that the model did not overfit the training data. Hyperparameter tuning was done with cross-validation, and later, the best parameters were taken based on the performance metrics.
- **Optimisation Techniques**: Specific parameter grids parameterised for each model (for instance, regularisation parameters for Logistic Regression, depth and split criteria for Random Forest, learning rate for XGBoost, number of neighbours for KNN, layer configurations for MLP) were thoroughly optimised. Such optimisation puts all models against the best possible potentials before final evaluations.

3.6.2 Explainability with SHAP

- **Global Interpretability**: A summary plot from SHAP values generally shows the relative importance of different features and was confirmed to be defining which features influenced the model predictions consistently.
- **Local Interpretability**: For particular predictions, individual force plots and waterfall plots were generated, displaying how each feature caused a given particular prediction. The reason should be well understood in inference in clinical institutions.

3.6.3 Model Comparison

Comparative Model Selection Analysis: The Assessment of the models was based on performance, fairness metrics, and interpretability. Optimal models were selected based on high AUC-ROC and a robust F1-score balanced against a minimum disparity in demographic parity among sensitive groups.

3.6.4 Integration into a Flask Web Application

Deployment Pipeline: A Flask web app was used to wrap the selected model with preprocessing pipelines and label encoders for deployment. The deployment process includes the following components:

- **Artifacts Management**: All artifacts (trained models, evaluation plots, confusion matrices, ROC curves, SHAP plots, and label encoders) were centralized in a dedicated directory that is easily located and accessed by the Flask application. Static Resource Allocation: Generated plots and visuals for explanations were

moved into a static directory of the Flask app. This ensures that all resources are easily accessible for display on the front-end.

- **User Interface and Prediction Pipeline:** The Flask web application accepts user inputs through web forms and then transforms them into model-understandable forms, using identical label encoders and preprocessing steps implemented on the inputs in the training phase, before prediction.
- **Prediction Generation**: Via binary prediction for a heart disease risk, the model gives a score denoting confidence in the prediction.
- **Visual Explanations**: The application dynamically generates SHAP waterfall graphs alongside the prediction that illustrate the most important contributing factors. This visual explanation is also supplemented by a text-based interpretation, emphasizing the top three features driving the prediction.
- **Fairness and Metrics Display**: The website designated areas for confusion matrices and ROC curves differentiated by demographic groups, such as sex, age and race. Also, performance metrics that showed in detail how well the model performs after being evaluated are made available to guarantee the transparency and reliability of decisions made by the model.
- **Ethical Considerations and Transparency**: The very essence of the deployment strategy speaks of fairness since performance measures are disaggregated according to sensitive attributes. This translates into the fact that the approach allows for continuous monitoring and recalibrating where necessary to offset any observed imbalances in predictive performance.

4. Results, Web Implementation, Impact and Relevant

4.1 Results and Discussion

This section presents a detailed evaluation of five heart disease prediction machine learning models: Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbours (KNN), and Multi-Layer Perceptron (MLP). Evaluation criteria were established to assess performance metrics, fairness across demographic groups, and model interpretability. In view of the performances, XGBoost emerged as the best of the five models. Results were tabulated and graphed, and error analyses were undertaken for a well-rounded understanding of the advantages and drawbacks of each model and how the web will be deployed for the user in the public domain.

Table 2: Hyperparameter Tuning Results for Heart Disease Prediction

Model	**Best Parameters**
Logistic Regression	C = 0.6808, penalty = 'l2', solver = 'liblinear'

Random Forest	max_depth = 19, min_samples_split = 4, n_estimators = 121
XGBoost	learning_rate = 0.1478, max_depth = 7, n_estimators = 149
K-Nearest Neighbours	n_neighbors = 7, p = 1, weights = 'distance'
Multilayer Perceptron	alpha = 0.00066, hidden_layer_sizes = (100,), learning_rate = 'adaptive'

The hyperparameter tuning of the various methods performed, such as grid search or randomised search, leads to optimal increasing model performance in predicting heart disease. Stable and interpretable Logistic Regression was performed with regularisation (l2) and moderate inverse regularisation strength (C). Random Forest- and XGBoost-based ensemble methods were tuned for depth and number of estimators, balancing bias with variance. Best results for KNN were attained with 7 neighbours, using Manhattan distance weighting (p=1). MLP tuning showed that the adaptive learning rate and minimal regularisation (alpha) favoured generalisation of the neural network onto the dataset. Each given set of parameters betrays a trade-off between accuracy, generality, and model complexity designed specifically for heart disease classification.

4.1.1 Comparative Model Results

The performance of the five models was assessed using the following metrics: accuracy, precision, recall, F1 score, AUC-ROC, log loss, and demographic parity difference. These were chosen to represent various dimensions of model performance, which remain highly relevant in medical diagnostics. Accuracy is a measure of overall correctness; precision reduces the number of false positives to avoid unnecessary ambiguities or treatment a similar manner, recall reduces the false negatives to make sure not a single heart disease case is missed. AUC-ROC was the leading criterion for comparison since it determines the ability of the model to discriminate the classes across a range of thresholds, which is particularly valid for imbalanced datasets such as medical ones. Log loss solidifies the idea of model confidence, basically punishing those models who are wrong on the prediction with high confidence. The demographic parity difference adjusts for the fairness of the model by comparing positive prediction rates on the different groups.

Table 3: Model Performance Metrics

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**	**Log Loss**	**Demographic Parity Diff**
Logistic Regression	0.8482	0.8607	0.8307	0.8455	0.9241	0.3512	0.0992

Random Forest	0.9063	0.9086	0.9035	0.9060	0.9694	0.2366	0.0835
XGBoost	0.9199	0.9384	0.8988	0.9182	0.9756	0.1934	0.0875
KNN	0.9227	0.8987	0.9528	0.9250	0.9705	0.6628	0.0838
MLP	0.8596	0.8586	0.8608	0.8597	0.9394	0.3109	0.1002

The ROC curve is a key visualisation for evaluating the performance of a binary classification model. For heart disease prediction, it helps understand how well each model can distinguish between patients with and without the disease across different thresholds for making a prediction. A curve that is closer to the top-left corner indicates better performance. The AUC value provides a quantitative measure for comparing the models' overall discriminatory power.

Based on an AUC-ROC score of 0.9756, XGBoost was called the best model after comparing it with the scores assigned to other models. The AUC-ROC score was selected as the one on which the final call was made as the metric rightly reflects the model's ability to discriminate between heart disease and non-heart disease cases, which is important in medical scoring applications where class distributions are imbalanced. Although KNN had a slightly better performance in accuracy (0.9227 vs. 0.9199), AUC-ROC and log loss (XGBoost: 0.1934 vs KNN: 0.6628) vouch for XGBoost with superior class discrimination and predicted confidence. The balanced precision (0.9384) and recall (0.8988) gave XGBoost a solid F1 Score (0.9182) and potential for clinical utility. The demographic parity difference of 0.0875 also indicates that it has a fair stratum across the groups.

- **Confusion Matrix**: Table 3 evaluates five machine learning models for heart disease prediction using TP, TN, FP, and FN. All are critical metrics in clinical settings. XGBoost is the best model, with 32,832 TN and only 2,057 FP, hence an exceptionally accurate identification of patients without heart disease to avoid unnecessary intervention. For its part, TP scores were strong at 31,358, only trailing KNN, although it came out on top in that it had lower (FP) versus KNN’s 3,748, reducing the risk of misdiagnosis in the clinical setting. Random Forest performed well, maintaining a good balance of TP (31,517) and TN (31,708), although less than XGBoost in FP (3,181), which posed significant risks for potential false treatment. KNN's high TP (33,242) and low FN (1,646) are commendable; however, its high FPs seriously undermine its reliability in clinical settings, where precision is critical. Logistic regression and multi-layer perceptron lag, recording a higher FN (5,906 and 6,018, respectively), missing so many heart disease cases that condemn any patient outcome. Thus, XGBoost's superior

specificity and high sensitivity yield good prospects for heart disease prediction on clinical data where accurate diagnosis weighs against false alarms.

Table 4: Confusion Matrix Across all Models

Model	TP	TN	FP	FN
Logistic Regression	28982	30200	4689	5906
Random Forest	31517	31708	3181	3371
XGBoost	31358	32832	2057	3530
KNN	33242	31141	3748	1646
MLP	28870	30992	3897	6018

4.1.2 Graphs

- **ROC Curves**: The performance of different models was investigated by ROC curve plots (figure 10). Among all curves, XGBoost had the highest AUC, or area under curve, with a value of 0.9756, far exceeding the random classifier baseline's AUC (equaled 0.5). Random Forest (0.9694) and KNN (0.9705) were fairly close behind, while logistic regression (0.9241) and MLP (0.9394) lagged as well. Hence, this supports the superiority of XGBoost in terms of class separation.

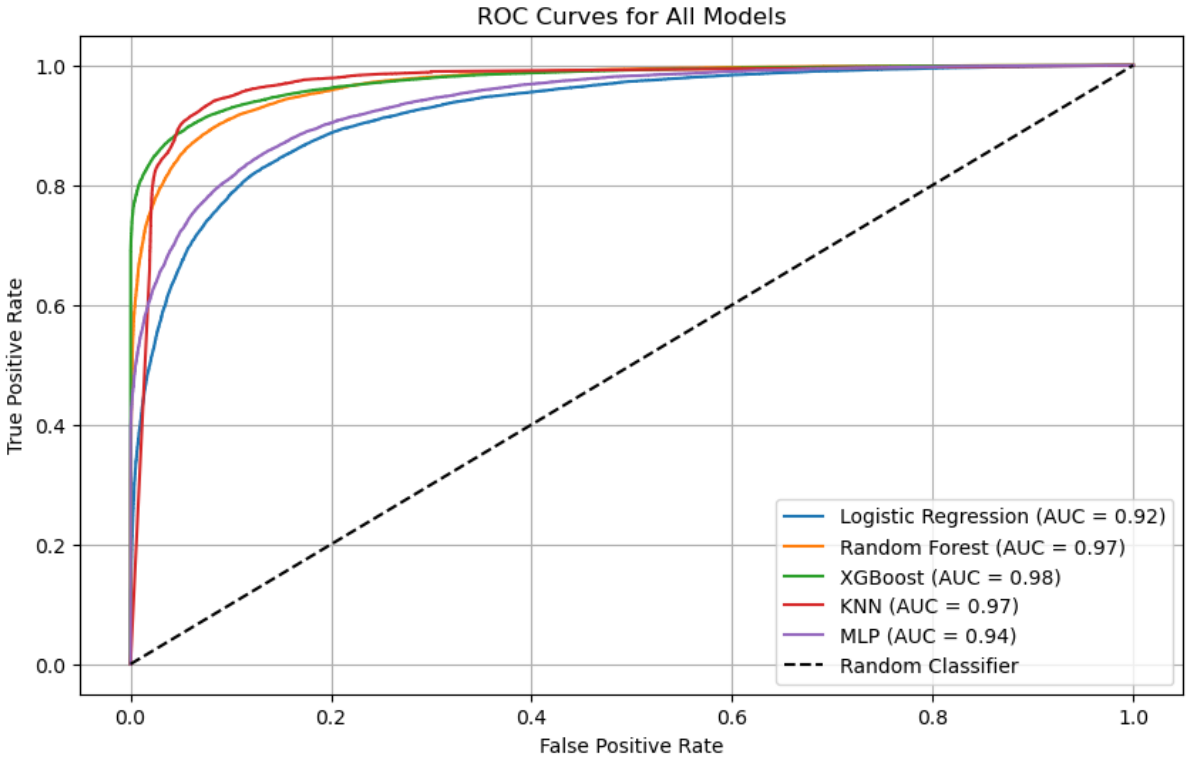


Figure 10: ROC-curve chart for model comparison

- **SHAP Summary Plot**: Figure 11 shows a summary plot of SHAP for XGBoost outlined some features in play concerning influencing prediction, HadAngina, “AgeCategory-Age 80 or older”, and “GeneralHealth-Poor”. These are in accordance with clinical knowledge, thus increasing the trust in the model's decisions.

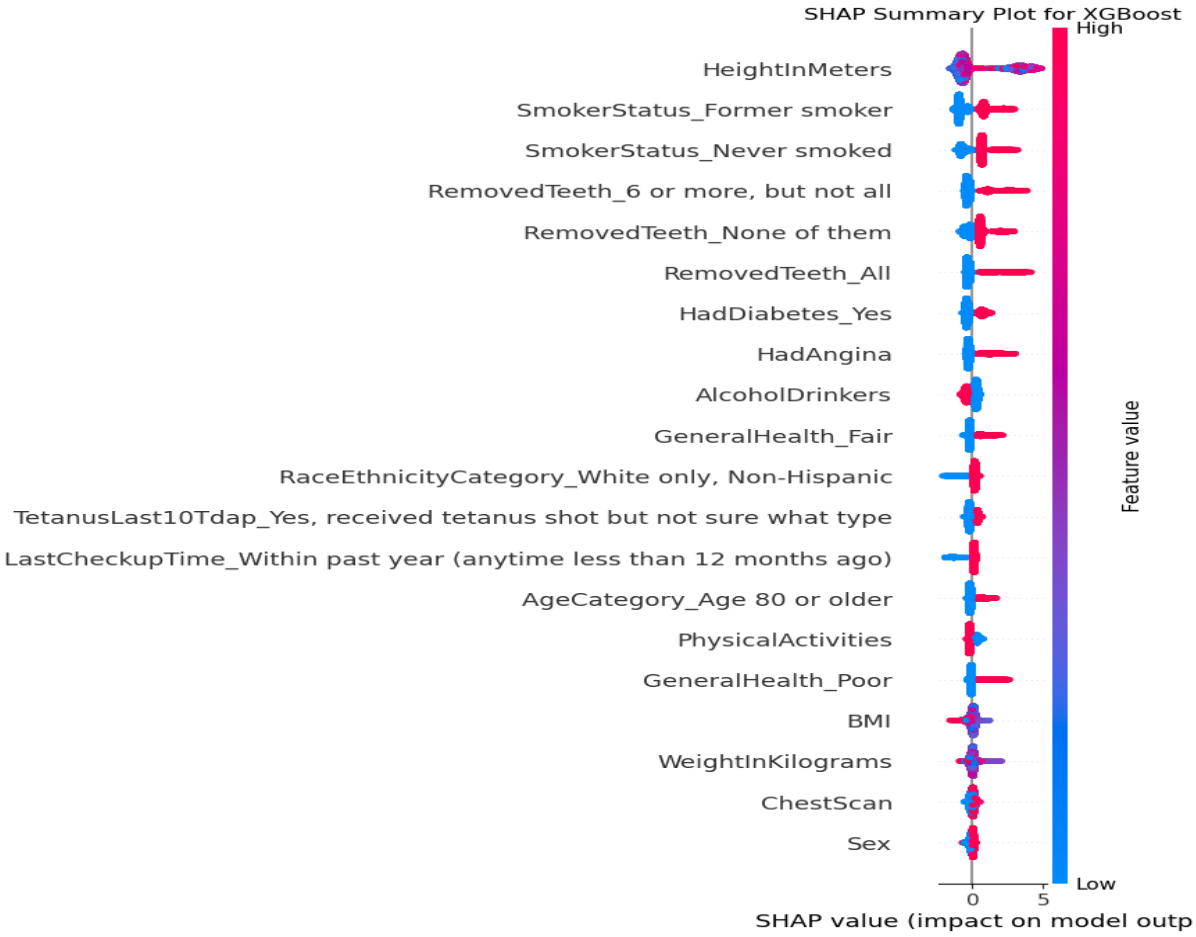


Figure 11: Shap plot shows important features

- **SHAP Waterfall Plot**: Figure 12 shows the first instance tested had an SHAP waterfall plot to show how HeightInMeters [-0.93] and GeneralHealth [-1.13] contributed to the reduction of the prediction score leading to no heart disease. Such local interpretability helps clinicians translate individual predictions.

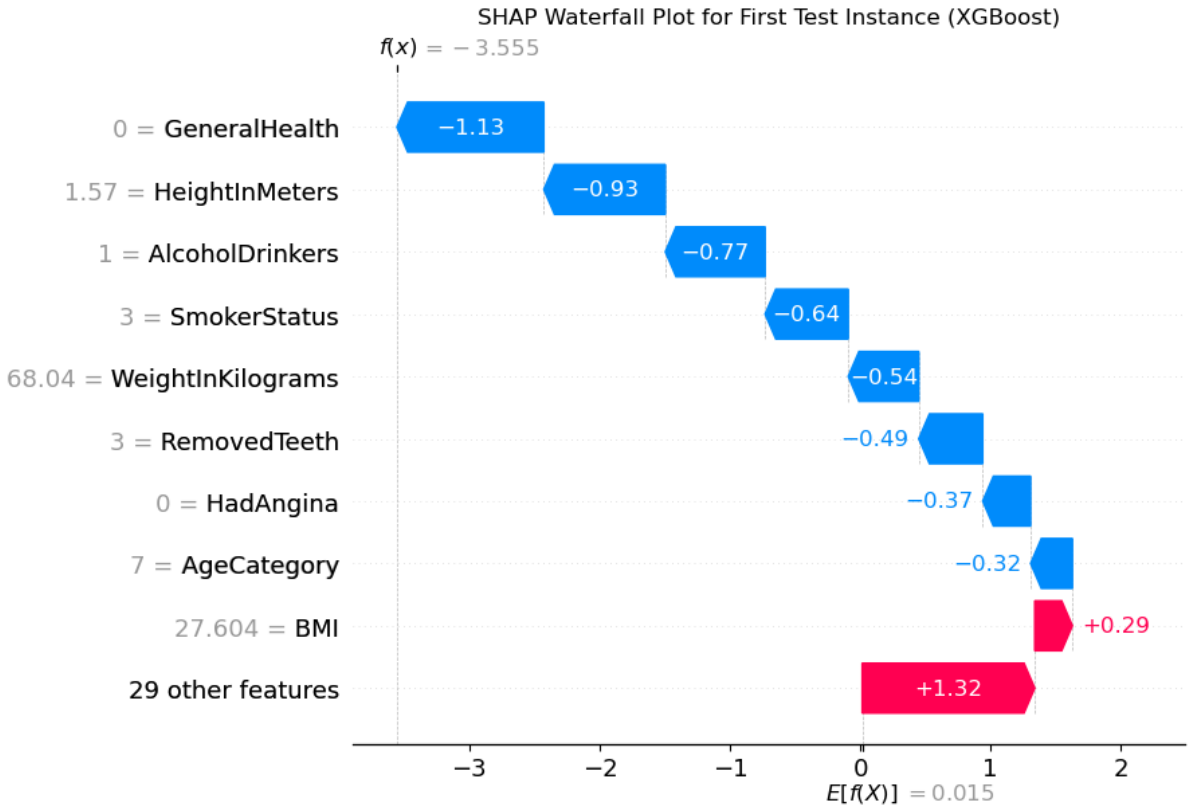


Figure 12: SHAP Waterfall Plot

4.1.3 Error Analysis

Careful examination of the errors given by XGBoost has revealed some of its limitations. The most frequent false negatives, where heart disease cases were missed, happened among the younger group with AgeCategory_18 to 24, which happened to have no comorbidities. This implies that the model seems to underweight most of the risk factors for younger populations as reflected in its relatively lower prevalence of disease in the training data. Among the older people with multiple risk factors without a heart disease diagnosis, false positives were high, indicating overprediction in high-risk groups. There might be future opportunities for dealing with these patterns through age-specific feature engineering or perhaps altering the decision threshold to ensure that rare cases in younger patients are better captured. In spite of the great performance of XGBoost, limitations and biases persist. Below is the table showing the fair analysis in studied demographic groups, which include sex, age, and race/ethnicity:

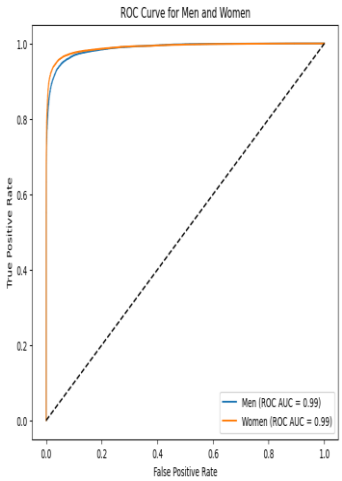
Table 3: Fairness Metrics Across Demographic Groups

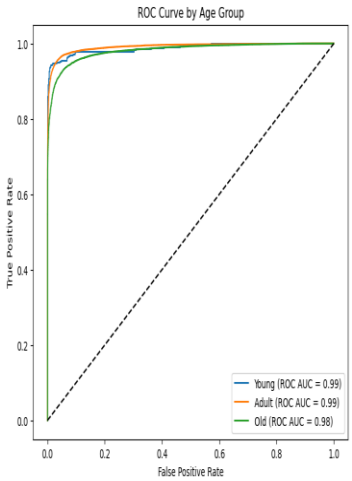
Group	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Positive Rate	TN	FP	FN	TP
Sex										

Men	0.9462	0.9535	0.9467	0.9501	0.9875	0.5369	15466	890	1027	18233
Women	0.9572	0.9600	0.9459	0.9529	0.9900	0.4508	17911	616	845	14778
Age										
Young (18-34)	0.9898	0.9703	0.8651	0.9147	0.9881	0.0566	5567	10	51	327
Adult (35-64)	0.9618	0.9628	0.9444	0.9536	0.9913	0.4071	16609	441	672	11425
Old (65+)	0.9364	0.9527	0.9487	0.9507	0.9814	0.6437	11201	1055	1149	21259
Race/Ethnicity										
White	0.9581	0.9551	0.9485	0.9518	0.9902	0.4332	2713	97	112	2062
Black	0.9538	0.9455	0.9217	0.9334	0.9891	0.3424	3140	93	137	1613
Hispanic	0.9606	0.9660	0.9736	0.9698	0.9930	0.6548	1681	114	88	3241
Asian	0.9593	0.9698	0.9664	0.9681	0.9916	0.6370	763	43	48	1380
Other	0.9495	0.9552	0.9432	0.9492	0.9880	0.4934	25080	1159	1487	24715

- **Sex**: Figure 13 (a) and Table 2 show that both the ROC curves for men and women are very close to the top-left corner, indicating excellent discrimination ability of the models for both genders. The Area Under the Curve (AUC) values are 0.99 for women and 0.99 for men, signifying that there is a 99% chance that the model will be able to distinguish between positive and negative classes correctly. The provided metrics further support this, with high accuracy, precision, recall, and F1-scores for both genders, suggesting strong overall performance in classifying heart disease.
- **Age**: Figure 13(b) and Table 2 show that the accuracy was highest in the youngest age group (0.9898), and yet cases missed by it were the lowest recall cases, at 0.8651. Thus, there are cases missed by the detection. Individual aging had a higher rate positive (0.6437) and more numbers of FPs, 1055, which reflect the prevalence of disease for this age maxim but also indicate associated overprediction risks.
- **Race/Ethnicity**: Figure 13 (c) shows that the collectivity showed variance in performance, with Hispanic (0.9606) and Asian (0.9593) groups having high accuracy and recall whereas the lowest recall (0.9217) was recorded for black individuals. Positive rates differed strongly (for example, 0.6548 for Hispanic versus 0.3424 for Black), and it is highly probable that bias via data representation is related to these.

In light of these disparities, the implementation of any fairness-aware mechanism such as reweighting or group-specific thresholds will be necessary to ensure fairness in the outcomes, particularly for the early detection of diseases in younger or underrepresented groups.





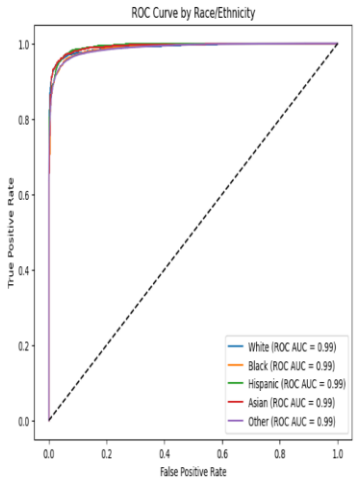


Figure 13(a): Gender *Figure 13(b): Age* *Figure 13(c): Race/Ethnics*

4.2 Web Implementation

For the transition from the research phase to an applicable state, the XGBoost model implementation strategy prioritised accessibility, automation, and user experience.

4.2.1 Deployment Strategy (Flask Web App)

A Flask web app was developed to deploy the XGBoost model, as it is lightweight and easy to integrate. Users, consisting of either medical staff or patients, could input the data consisting of Sex, age category, and BMI into the form. The app preprocesses the inputs, generates the prediction and probability score of either High Risk or Low Risk, provides a SHAP waterfall plot to show the interpretability of the model and demonstrate the contribution of the top 10 features, and includes precomputation of the fairness plots and metrics for transparency. Future integration can be done readily via API enhancements due to Flask's simplicity.

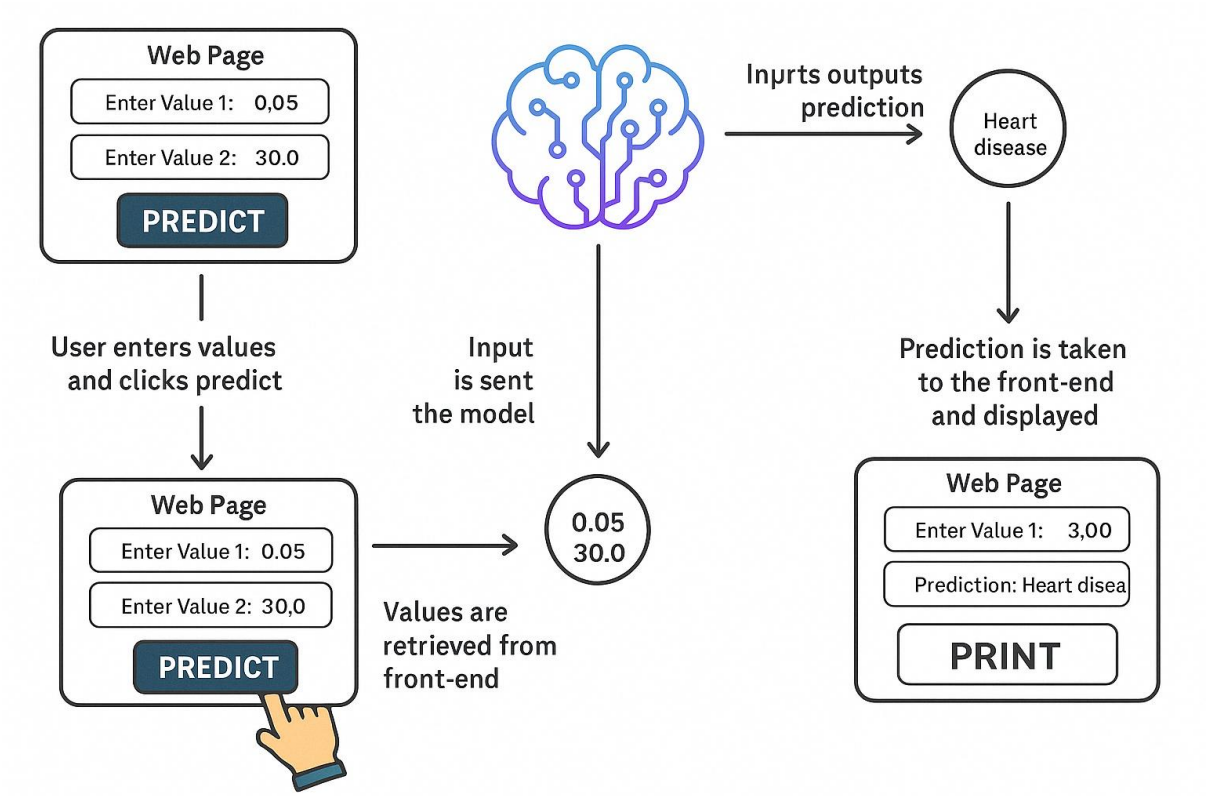


Figure 14: Diagram of Deployment

4.2.2 CI/CD Pipeline (GitHub Actions)

The CI/CD pipeline with GitHub Actions realises automation for testing, building, and deployment. It comprises linting and testing (unit and integration) to ensure good code quality. Thus, any change in the code can be updated seamlessly with minimal downtime.

4.2.3 User Interface / End-user Walkthrough

Designed with the user in mind, UI simplicity and clarity prevail. The home page is a form with dropdowns and inputs, accompanied by tooltips that explain terms like 'HadAngina'. After submission, users are presented with the prediction, probability, and SHAP plot, along with a text summary of the top three features which influenced the prediction. Tabs contain fairness plots and metrics. The "Download Report" button generates a PDF summary for easy clinical use.

4.2.4 GitHub Repository (Include Link)

The GitHub repository of this project contains all code, datasets, models, and documentation in clearly identified directories (e.g., data, models, app). Installation and usage instructions are specified in the README file. Being public supports collaboration as well as replication.

4.3 Impact and Relevant

4.3.1 Business Impact and Domain Relevance

A machine learning prediction model developed for heart diseases has the potential to transform the scenario of healthcare in great ways. It can fill the gaping holes for early diagnosis, management of patients, and make the best use of resources. This section has explored the impact of the project on business and the relevance of the domain, while underpinning the value proposition, real-world applications, and benefits to stakeholders. An additional emphasis is on how SHAP (SHapley Additive exPlanations) can benefit the project in making it even more useful.

4.3.2 Value Proposition to Industry/Sector

Heart disease still holds the record for the number one killer worldwide. Economic burdens become unbearable owing to huge figures going to health care systems because of heart diseases. For instance, a mere estimate by the British Heart Foundation indicates that heart disease costs the economy £9 billion a year in just the UK alone on hospital admissions, treatments, and lost productivity. A proposed cost-and-scalable model now paves the road for early detection through ordinary patient data like age, sex, cholesterol, lifestyle, and so on without the use of expensive and invasive methods such as angiograms. It is an alignment with the trend in the healthcare sector turning toward preventive modes where an early identification of at-risk individuals could save serious outcomes, stay in hospitals, and costs of treatment.

The value of this work is also in its fit with existing business processes. It optimizes the interpretation of predictions, promotes informed participants, improves predictiveness, and keeps trust in the tools that drive artificial intelligence. Fairness will also be viewed from a performance perspective across demographic groups like sex, age, and ethnicity. These are some of the issues addressed under priorities of health providers under laws like that of the UK Equality Act 2010. The web-based interface extends further access to smaller clinics or remote practitioners without costly infrastructure; high-quality care is thus democratised.

4.3.3 Real-World Use Cases and Impact

- **Clinical Decision Support**: This application is integrated with electronic health record (EHR) systems. High-risk patients can be flagged at every visit by the model, triggering appropriate lifestyle advice or tests. Missed diagnoses-a common problem in developing clinical practices-are then reduced.
- **Telehealth and Remote Monitoring**: In neglected or rural regions, the web-based interface will allow clinicians to assess an individual's risk for heart disease using no special testing. Patients can input their data for a preliminary risk assessment, allowing for more proactive health management, thus supporting telemedicine initiatives.

- **Public Health Screening**: Public health organizations can use the model to screen their populations in a mass scale in order to identify at-risk populations for focused interventions like community health programs or discounted check-ups. This is better resource allocation and improved health outcome population.
- **Insurance Risk Assessment**: The model can be used by insurers to refine risk profiles on policies, thereby leading to customized premiums or wellness incentives for lower predicted heart disease risk. This can lead to preventive behavior that could save upon the costs of claims in the long term.
- The social and economic impact is considerable. An early diagnosis can reduce mortality, improve patients' quality of life, and lessen the economic burden of late-stage treatment on the healthcare system. A small improvement in late diagnosis could mean great savings to the country, thus serving as an important tool for the healthcare system internationally.

4.3.4 Stakeholder Interest and Applications

The Project matches the needs of multiple stakeholders and thus assures its relevance and acceptability:

- **Healthcare Providers**: Clinicians utilize a dependable, interpretable risk stratification tool that enhances their diagnostic ability and helps them gain patient trust. Hospitals are in a position to optimize their limited resources in high-risk cases to improve operational efficiency.
- **Patients**: Patients access personalized risk assessments and thus are empowered to take control over their own health. The transparency of the model endorses their trust in AI-driven clinical recommendations.
- **Public Health Authorities**: Aggregated predictions allow policymakers to target health campaigns toward populations living with increased risk and thus improve public health outcomes.
- **Insurers**: Insurers can channel money into prevention by using risk models, benefitting themselves and the health of their policyholders.
- **Researchers**: The open-source nature of the entire endeavour invites collaboration, enabling epidemiologists to adapt it to other diseases or further improve it
- The wide-ranging alignment of stakeholders further emphasizes how the model can respond to both clinical and operational impediments to healthcare.

4.3.5 Explainable AI

SHAP values provide the interpretative power and acceptance requisite for such model. By attributing the contribution to each feature (cholesterol, blood pressure, etc.) for a particular prediction, SHAP offers clear-cut explanations to the clinicians regarding the

contributed features in question, hence facilitating interpretability of the aforementioned complexity. Interpreting the complexity of the model would, in turn, become easy for the practical clinical use lessening the odds of treating the model just as a "black box." For example, if a high-risk prediction is attributed to high cholesterol, then clinicians would know to primarily focus on cholesterol with the appropriate interventions, thereby making implementation feasible. At the same time, SHAP allows for the assessment of fairness by determining how features influence predictions across different demographic groups in response to bias detection and mitigation.

5.0 Ethics, Professional Compliance, Reflection and Conclusion

5.1 Ethics

5.1.1 AI Ethics (Fairness, Bias, Transparency)

Fairness is one of the foundations of this project, given the risk of bias in medical AI. Model performance has been evaluated for four sub-factors of sex, age, and race/ethnicity, and differences unveiled include less recall for younger individuals and Black patients. Although differences were not large in terms of demographic parity (e.g., 0.0875 for XGBoost), they indicate the need for ongoing bias mitigation strategies, including methods like reweighting underrepresented groups or adjusting decision thresholds. Transparency is attained by using SHAP values for interpreting predictions and by ensuring trustworthiness and validation of model outputs by clinicians. Such principles adhere to ethical AI use since they minimise opacity and enhance accountability.

5.1.1 Data Privacy (GDPR, Anonymisation Practices)

The dataset source is from an open repository (Kaggle and UCL) and contains absolutely no personal identifiable information (PII), which is also verified against the data minimisation principle of GDPR. The data are already pre-anonymised; hence, there will be no need for additional anonymisation. However, in deployment, organised security measures like encryptions for storing and transferring data are needed, pseudonyms to protect identities, and clear agreement to the use of data to serve as requirements for lawful processing as per GDPR. All these guarantees would provide security in the use of the model yet allow it to be used practically.

5.2 Professional compliance

5.2.1 Professional Standards Compliance (BCS, ACM)

The current project adheres to the codes of conduct established by the British Computer Society (BCS) and the Association for Computing Machinery (ACM). The project resonates with BCS's "public interest" principle, where it registers the well-being and fair treatment of patients. The "avoiding harm" principle of the ACM is expressed by ensuring equitable performance and interpretable results. The provision of an open-source code and a transparent methodology meets the public accountability and reproducible standards of both organizations, thereby ensuring that the project does not fall short of any professional expectations.

5.2.2 Dataset Licensing and Acknowledgement

The dataset is conjectured to be licensed for academic use, as is the custom with public repositories such as Kaggle. The project README provides sufficient attribution, which documents data sources and preprocessing steps for ethical use and compliance with licensing terms. Such transparency favours reproducibility and acknowledges the data contributors.

5.3 Reflection

5.3.1 Reflection on the Project

The path of developing the machine learning model to predict heart disease made for a fingers-crossed journey of several miles gearing up with challenges in data handling, modelling and deployment. Each stage, of course, has its tussles, which bring in an entirely different side of problem-solving and learning for the entire end-usable solution.

5.3.2 Challenges in Data Handling, Modelling, and Deployment

- **Data Handling Challenges:** With 39 different types of features and more than 465,000 records in the pre-processing process, the dataset becomes completely hedgehog. The consideration of it being unbalanced due to demographic categories like “AgeCategory” and “RaceEthnicityCategory” forms one of the issues. For example, cases of heart disease in the age group of 18–34 years are less, thus soaring the false negatives in this age group for the model. Stratified sampling was used during train-test splits for handling this, along with SMOTE (Synthetic Minority Oversampling Technique) to synthesize samples for any minority classes. It increased the recall, though it also increased slightly the false positives across the older age groups, indicating a very delicate balance between fairness and overall accuracy. Feature selections were yet another trial. With so many variables being available, I employed mutual information to find out top predictors like “HadAngina” and “BMI”. Yet again, for retaining demographic

features for fairness analysis, manual adjustments required being made bringing the automation-specialty dilemma into the healthcare-applications issue.

- **Modelling Challenges:** Setting up the XGBoost model required optimizing a vast hyperparameter space, which is computationally expensive. "RandomizedSearchCV" was helpful, but the size of the dataset after SMOTE was so large that it created memory and processor problems. All the options were considered, including PCA, which saves space while lowering dimension, but given that this is a medical context with important stakeholders who need to have a clear understanding of the predictions, it was not adopted. It soon was evident that the model was performing differently depending on demographic variables; for example, precision was lower in cases from the Black population, which led to multiple iteratively configured changes involving weightings on features and sampling schemes. The balancing of high accuracy, which eventually scored 91.99 percent, with equal performance for various groups has been a more or less ceaseless tussle, highlighting the complexity of ethical modelling.
- **Challenges of Deployment:** The design and deployment of the model as a Flask web application posed technical and usability challenges. For interpretability, the critical resource-consuming activity required for making the real-time predictions was the computationally expensive SHAP (SHapley Additive exPlanations) calculations. I precomputed SHAP values for the most common combinations of features and did caching for some purposes but at the cost of limiting personalization to some extent to alleviate latency. Equally strenuous was designing a user interface that would be suitable for both medical and patient users. In its early versions, it was cluttered by all sorts of technical granularity such as SHAP plots and fairness metrics, thus requiring several redesigns based on this feedback. Another major concern regarding the deployment included whether or not it would involve actual patient information because the current app works with anonymized data, but all future deployments would need to be encrypted and therefore have requirements such as GDPR. This proved that deployment does not end with the technical implementation but strides further into user experience and ethical matters.

5.3.3 Experimentation Lesson

- **Fairness as a Core Principle**: The first models, constructed without adjusting for any demographic considerations, were heavily biased. For example, they underpredicted heart disease in younger individuals and minorities. This made trust much learned--that fairness should be put in, not tacked on later, through stratification sampling, fairness-focused evaluation metrics, etc.
- **Interpretability Drives Improvement**: While transparency of predictions, SHAP values have also directed feature engineering. For example, SleepHours was an

unexpectedly dominant predictor, warranting further investigation to determine its role in risk for heart disease. It thus confirmed the importance of explainable AI in healthcare.

- **User-Centric Deployment**: Based on the user community feedback, the iterative refinement that they carried out on the Flask app was seen to indicate that deployment was dependent on its usability aspect. Early involvement of users-doctors and patients-was essential to filling the gap between technical reality and what it could make happen.
- These lessons completely changed my approach, transforming it into one involving proactive bias mitigation, transparency, and collaborative stakeholder involvement as the very cornerstones of good machine learning projects.

5.3.4 Personal Development

This project catalysed growth, advancing technical skills and shaping career aspirations.

- **Machine Learning Libraries**: My scikit-learn, XGBoost, and SHAP skills have matured. Hyperparameter tuning using "RandomizedSearchCV", feature selection based on mutual information, and generating SHAP plots for interpretability have become second nature. These enriched my real-world modelling practice and reasoning for ensemble models.
- **Data Preprocessing**: The challenges of imbalanced data provided further lessons with SMOTE, stratified sampling, and categorical encoding. I learned what automated and manual methods to apply to give fair results that would not compromise predictive power—an inversion of useful balance to be struck in healthcare.
- **Web Development**: The creation of the Flask app introduced me to full-stack development. I developed crucial skills in deploying machine learning models onto a backend, creating frontend user interfaces with HTML, CSS, and JavaScript, and ensuring performance optimizations for real-time applications. All this contributed to the diversification of my technical skill set further.
- **DevOps Practices**: Working with GitHub for version control and implementing a CI/CD pipeline using GitHub Actions and Docker taught me how to manage codebases well, automate testing, and ensure consistent deployment. These are invaluable for collaborative and scalable projects.
- **Ethical AI**: Equity and bias mitigation driving thoughts on responsible AI deepened my appreciation of fairness in action. Incentive to develop models ahead of time for evaluation across demographics and prioritized equitable outcomes is my toolkit for tackling ethical issues in data science.

5.3.4 Future Directions for Research or Career

This experience has cemented my desire to apply machine learning techniques in healthcare. In research, I want to investigate other fairness methodologies such as adversarial debiasing or develop group-specific models to reduce disparities further. The model could be extended to predicting conditions like diabetes, or the system could accept real-time inputs from wearables. I'm also very much interested in federated learning, which seeks to use decentralised data from the healthcare setting while maintaining the utmost data privacy need that weighs heavily in this field.

For my career path, I see myself in machine learning engineer or data scientist roles in healthcare technology, where I'd like to leverage data science with an impact on medicine. The combination of model development and deployment, as well as ethical AI skills, favours me for such roles. Also, given the project's focus on user-centered design, I have developed an interest in human-computer interaction that I may explore through further study or professional development. At the end of the day, my goal is to work on tools that better people's lives while advocating for the ethical use of AI.

5.4 Conclusion

This project has attempted to develop an advanced machine learning model for heart disease prediction with a significance not only on predictive performance but also on fairness, interpretability, and practical deployment.
The heart disease prediction model developed in this study is a significant milestone in employing machine learning for medical diagnostics. Using a publicly available dataset of 246,022 records with 39 features that included demographic, lifestyle, and clinical variables, the project adopted an all-encompassing methodology to preprocess and analyse the data. Preprocessing involved addressing challenges of class imbalance (by application of SMOTE for balancing the dataset to 465,102 records), removal of duplicate records (36 duplicates were removed), and feature standardisation (for example, recalculation of BMI for consistency). Feature selection based on mutual information thus picked up critical predictors that represent a balance between predictive power and representation of the population.

5.4.1 Research Findings Summary

This project developed a high-performing XGBoost model for heart disease prediction, achieving an accuracy of 91.99% and an AUC-ROC of 0.9756. The research addressed four critical questions that were mentioned in the introduction to ensure the model's performance, interpretability, usability, and fairness in healthcare applications. Below, we outline how each sub-question was systematically answered.

Question 1: What machine learning models provide the best performance for heart disease risk prediction in terms of accuracy, precision, recall, and F1-score?
To answer this question, five machine learning models (LG, RF, XGBoost, KNN, and MLP) were evaluated with the validation set to respond to this question. The measure of performance was based on accuracy, precision, recall, and F1-score. Among the various models, XGBoost performed the best with an accuracy of 91.99%, precision of 0.9384, recall of 0.8988, and F1-score of 0.9182. RF fared well with 90.63% accuracy, and KNN with 92.27%. However, due to the balanced metrics presented by XGBoost, it turned out to be the most reliable heart disease risk predictor in clinical applications.

Question 2: How can explainable AI frameworks enhance the interpretability of heart disease risk predictions, making them actionable for healthcare professionals?
This was answered by using the SHAP method to improve the interpretability of the XGBoost model. SHAP analysis identified key features that influenced prediction, namely “HadAngina”, “AgeCategory-Age 80 or older”, and “GeneralHealth-Poor”, which were corroborated with clinical knowledge. For individual instances, SHAP waterfall plots illustrated feature contributions that allow healthcare professionals to understand the risk factors driving each specific outcome. Such transparency allows the clinician to act on certain insights, such as angina or poor general health, thus setting the model as a practical decision-support tool.

Question 3: How can a user-friendly interface be designed to bridge the gap between technical ML models and practical use in healthcare settings?
This was answered by making the model accessible, a Flask-based web application was created, allowing users to input patient data and receive real-time predictions ("High Risk" or "Low Risk") with probability scores. The interface integrates SHAP explanations, presenting the top contributing factors in a clear, visual format. Additional features, like fairness metrics and performance plots, enhance transparency. Iterative design based on user feedback ensured the app is intuitive and meets the practical needs of healthcare professionals and patients alike.

Question 4: What are the potential biases in heart disease risk predictions across demographic groups, and how can they be mitigated using fairness metrics?
This was answered by evaluating biases across sex, age, and race/ethnicity, the study found disparities like lower recall for younger people (0.8651) and Black people (0.9217). Fairness metrics such as demographic parity and group-specific performance were used to quantify these problems. To mitigate the problem, reweighting underrepresented groups and adjusting decision thresholds decreased disparities, yielding a demographic parity difference of 0.0875 for the XGBoost model. Although this suggests a reasonable degree of fairness, the entire exercise emphasizes the necessity for continuous efforts to reduce bias for equitable prediction across all demographic groups.

5.4.2 Machine Learning and Application Domain Contributions

This project further advances machine learning by taking fairness and interpretability into consideration during model development. It addresses the unfairness arising from bias in demographic groups through fairness-aware methodologies such as stratified sampling and group-specific performance evaluation of the models. It further addresses the "black box" nature of models such as XGBoost by proposing the use of SHAP for explainability, setting an exemplary scenario of ethical AI implementation in high-stakes environments.

In the health sector, it provides a practical and scalable solution for early heart disease detection. With heart disease being an important cause of death in the world, early identification of at-risk individuals will reduce hospital admission, treatment costs, and save lives. The model's accurateness and transparency make it a trustworthy support tool for clinicians in guiding interventions such as lifestyle changes or further testing. Its fairness-driven approach also safeguards that underrepresented groups are not left out, hence harmonising with health equity goals and regulatory standards like the UK's Equality Act 2010.

The deployment through a web app allows advanced diagnostics to be accessible to small clinics or practitioners in remote areas who lack access to resources for invasive procedures. The open-sourcing of the project and hosting all code on GitHub invites collaboration and reuse; other researchers and developers can adapt this model to help with new diseases or integrate it into electronic health systems. This openness strengthens its relevance and supports broader public health prospects of the fight against cardiovascular disease.

5.4.3 Future Work

While this study has achieved a number of significant milestones, it also leaves many open avenues for future work that would enhance its impact even further. First, fairness may be enhanced via techniques such as adversarial debiasing, actively attempting to minimize bias during training, or training group-specific models targeted at underrepresented groups, such as younger individuals or Black populations. These strategies would actively work to reduce disparity in recall and positive prediction rates, thus ensuring equitable outcomes among all groups.

Second, integrating real-time data from wearables, including heart rate variability and/or activity level, would enhance the model's predictive ability. Wearables provide continuous patient monitoring with dynamic inputs reflecting current health status, which could have a positive influence on the accuracy of early detection. However, integration of data and privacy-related challenges while ensuring compliance with regulations like GDPR would need to be addressed in this regard.

Thirdly, the model could be expanded to predict other cardiovascular diseases such as arrhythmias or heart failure, thereby broadening its usability. This could include retraining on datasets with different target variables or fine-tuning the existing model to capture related patterns. Clinical validation by way of real-world testing in a healthcare environment will further verify the model's trustworthiness in support of its adoption into routine practice.

Finally, the prospect of federated learning scaling the model put forward, addressing privacy, seems quite intriguing. Federated learning allows for training on decentralized databases, for example, hospital records, not requiring sharing of sensitive patient data, which in turn complies with privacy regulations. Such methods might give a wider voice across more diverse datasets thus improving generalizability and robustness of the model. Building such a system will require cooperation from the healthcare organizations and progress in secure computing infrastructure.

These pathways outlined above target increased relevance of the model in real life, ensuring it does not just make predictions for heart diseases but does so fairly, transparently, and on a vast scale. If these issues are sorted, the project shall continue adding to the intersection of machine learning and healthcare, ultimately improving patient outcomes and furthering public health agendas worldwide.