**1. Statement of Purpose**

**1.1 Background and Relevance**

Heart disease is a major global health challenge, affecting millions of people worldwide.

According to the Global Burden of Disease Study, cardiovascular diseases, including heart disease, are the leading cause of death globally, accounting for an estimated 17.9 million deaths annually (World Health Organization [WHO], 2021). Cardiovascular diseases contributed to 32.1% of all deaths in 2015, increasing from 12.3 million deaths (25.8%) in 1990 (Roth et al., 2018). In 2021, over 64 million people were living with cardiovascular conditions globally, and this number is expected to rise due to aging populations and the increasing prevalence of risk factors such as hypertension, diabetes, and obesity (GBD 2019 Diseases and Injuries Collaborators, 2020).

In Canada, heart disease affects approximately 750,000 individuals, with thousands of new cases diagnosed each year (Heart & Stroke Foundation of Canada). The economic burden is substantial, as heart disease is one of the leading causes of hospitalization, costing the Canadian healthcare system billions of dollars annually in direct and indirect costs.

Early detection and risk prediction are crucial to improving patient outcomes and reducing healthcare costs. Traditional clinical risk scoring methods, such as the Framingham Risk Score, have been widely used but may not fully capture the complexities of heart disease progression. Research indicates that the Framingham Risk Score may underestimate cardiovascular disease mortality risk in socioeconomically deprived populations and overestimate risk in lower-risk populations (Brindle et al., 2006; Hign Institute, n.d.). Additionally, it was developed primarily from a Caucasian population, potentially limiting its applicability to ethnically diverse groups (D’Agostino et al., 2001). Moreover, the score does not account for the effects of medical interventions and preventive treatments, which can significantly alter an individual's risk profile (Cooney et al., 2010). Due to these limitations, newer predictive models incorporating machine learning and advanced biomarkers have been proposed to improve accuracy (Krittanawong et al., 2020). Recent advancements in machine learning (ML) provide an opportunity to enhance heart disease prediction by analyzing large-scale clinical datasets and identifying hidden patterns that may not be apparent through conventional statistical methods. Studies have shown that ML models can achieve higher accuracy in predicting heart disease risk, potentially improving early diagnosis and enabling more effective interventions.

**2. Scope of the Project**

This project focuses on predicting the risk of heart disease using machine learning techniques.

The dataset used is the Heart Disease Predictor XM Dataset from Kaggle competitions, which contains 952 records and 12 features. These features include age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, as well as two specific ECG measurements: "oldpeak" and "ST slope." The "oldpeak" feature represents the ST depression induced by exercise relative to rest, while the "ST slope" refers to the slope of the peak exercise ST segment, both of which are critical indicators measured during an electrocardiogram (ECG), particularly during a stress test. These features collectively provide valuable insights into physiological and diagnostic markers associated with heart disease, enabling the development of robust predictive models.

**2.1 Data Preprocessing & Exploration**

1. Standardizing data types to ensure consistency in numerical variables (e.g., converting age to integer).
2. Handling missing values and duplicate records to maintain data integrity.
3. Detecting and removing outliers using the Interquartile Range (IQR) method to improve data quality.
4. Encoding categorical variables if required for seamless model integration.
5. Performing exploratory data analysis (EDA) to understand feature distributions and identify trends.

**2.2 Model Development, Evaluation & Deployment**

1. Implementing multiple machine learning models (Logistic Regression, Decision Trees, Random Forest).
2. Optimizing hyperparameters for better model performance.
3. Evaluating models using cross-validation and performance metrics.
4. Deploying the model to cloud based environment for user interaction.

**2.3 Visualization & Interpretation**

1. Creating visual representatives of key findings using Python.
2. Presenting feature importance to identify critical risk factors for heart disease prediction.

**2.4 Final Report & Business Implications**

1. A structured document summarizing the study’s methodology, findings, and business relevance.
2. Recommendations for applying the model in real-world healthcare settings.

**3. Background Research and Literature**

Heart disease is a chronic and widespread condition that significantly impacts global and national healthcare systems. The increasing prevalence of heart disease, along with its associated hospitalization rates and mortality risks, has driven researchers to explore advanced analytical methods, including machine learning (ML), to enhance early detection and risk assessment.

**3.1 Heart Failure and its Global Impact**

Heart disease affects millions of people worldwide and is one of the leading causes of death globally. Its prevalence is expected to rise due to aging populations and increasing risk factors such as diabetes, hypertension, and obesity. The condition accounts for a significant portion of hospitalizations, particularly among individuals over 65 years old, and contributes to high healthcare costs. Given these challenges, predictive models that assess heart disease risk can play a critical role in early intervention, treatment planning, and reducing the burden on healthcare systems (O'Meara & Ezekowitz, 2022).

**3.2 Machine Learning for Heart Disease Prediction**

Machine learning has become increasingly prominent in medical diagnostics, offering significant

improvements in the accuracy of prognosis models for cardiovascular diseases. Traditional statistical methods, such as logistic regression and Cox proportional hazards models, have been widely used for risk prediction but often struggle to capture complex, nonlinear relationships within clinical datasets (Chicco & Jurman, 2020). In contrast, machine learning algorithms, such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting, excel at handling large datasets, identifying intricate patterns, and enhancing predictive performance.

A study by Chicco and Jurman (2020) explored the application of machine learning techniques

for predicting heart failure outcomes using clinical datasets. Their findings demonstrated that ML models, particularly Support Vector Machines (SVM), outperformed traditional logistic regression models in classifying patient outcomes. While SVM showed strong performance in their study, Random Forest has also been widely recognized for its robustness and ability to handle high-dimensional data, making it a popular choice for cardiovascular disease prediction tasks. In our study, we employed the Random Forest

model, which demonstrated superior performance in predicting heart disease risk, achieving the highest accuracy and AUC-ROC scores among the algorithms tested.

Recent research by Zhang et al. (2023) further supports the use of ensemble learning methods,

such as Random Forest and XGBoost, for cardiovascular disease prediction. Their findings highlighted that these methods consistently achieved high performance in terms of risk stratification and predictive accuracy. This aligns with our results, where Random Forest emerged as the best-performing model, underscoring its effectiveness in capturing complex relationships within clinical data.

In this project, we leverage the Heart Disease Predictor XM Dataset from Kaggle competitions, which includes 12 clinical features such as age, sex, chest pain type, resting blood pressure, cholesterol levels, and ECG measurements like "oldpeak" and "ST slope." By utilizing the Random Forest algorithm, we aim to identify key risk factors and optimize model performance to improve early detection and risk prediction for heart disease. This approach aligns with the growing body of evidence supporting the use of advanced ML techniques to enhance diagnostic accuracy and patient outcomes in cardiovascular care.

**3.3 Clinical Features and Their Predictive Relevance**

The dataset used in this study comprises 13 features, including key physiological indicators such as resting blood pressure, cholesterol levels, fasting blood sugar, and maximum heart rate, as well as patient demographics like age, sex, and chest pain type. Several studies have highlighted the predictive power of these features in assessing cardiovascular health. For instance, elevated resting blood pressure and cholesterol levels have been strongly linked to increased heart disease risk, while reduced maximum heart rate recovery is indicative of poor cardiac function (Smith et al., 2022). By integrating these variables into a robust machine learning model, this study aims to provide clinically relevant predictions that can assist healthcare professionals in early diagnosis and risk stratification for heart disease.

**4. Design and Data Collection Methods**

**4.1. Data Source and Collection**

The dataset used in this project is sourced from the Heart Disease Predictor XM competition on

Kaggle, which aims to develop machine learning models for predicting heart disease based on clinical and demographic data. The dataset is publicly available at Heart Disease Predictor XM. This dataset includes real-world medical data collected from patients diagnosed with heart disease. It contains various physiological and demographic attributes, such as age, blood pressure, cholesterol levels, and other health indicators, which are critical in assessing heart disease risk. The competition encourages participants to explore machine learning and deep learning techniques to improve heart disease prediction, contributing to advancements in medical data science.

**4.2 Dataset Description**

The dataset consists of 952 observations (patient records). The features include numerical, categorical, and binary variables, capturing essential medical factors.

A table of medical information

AI-generated content may be incorrect.

**4.3 Analytical Approach**

The analysis follows a structured data science methodology, ensuring rigorous data preprocessing, exploratory data analysis, feature selection, and predictive modeling.

**4.3.1 *Data Preprocessing***

Before the preprocessing phase, the dataset had the following structure:

A screen shot of a computer

AI-generated content may be incorrect.

Figure 4.1: Dataset Features and Data Types Before Preprocessing

This table outlines the dataset’s features, data types, and completeness before any modifications were made.

Before proceeding with data analysis and model development, the dataset was examined to

understand its structure, identify potential issues, and ensure data consistency. The dataset initially contained numerical and categorical variables with varying scales, some extreme values, and floating-point values in the age column. The preprocessing steps aimed to enhance data quality and reliability for further analysis.

**4.3.1.1 Data Type Standardization**

The dataset’s structure was well organized, with all data types appropriately defined and consistent. No changes were required for the age column or other features, as they were already in suitable formats, ensuring uniformity and readiness for analysis.