## 

CVD Prediction Using Machine Learning

Laxmi Sowjanya Doddi

A Project Submitted to

GRAND VALLEY STATE UNIVERSITY

In

Partial Fulfillment of the Requirements

For the Degree of

Master of Science in Applied Computer Science

School of Computing

April 2025



The signatures of the individuals below indicate that they have read and approved the project of Laxmi Sowjanya Doddi in partial fulfillment of the requirements for the degree of Master of Science in Applied Computer Science.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Yong Zhuang, Project Advisor 04/18/2025

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Robert Adams, Graduate Program Director 04/18/2025

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Dr. Marouane Kessentini 04/18/2025

# Abstract

The World Health Organization (WHO) reports that cardiovascular diseases (CVDs) are the leading global cause of death, accounting for approximately 17.9 million fatalities annually. These diseases include a variety of heart and blood vessel disorders such as coronary heart disease, cerebrovascular disease, and rheumatic heart disease. Critically, more than 80% of CVD-related deaths are caused by heart attacks and strokes, with a significant portion approximately one-third occurring prematurely in individuals under the age of 70. Early detection and prevention are essential to reducing this global health burden.

This project aims to contribute to the early identification of individuals at high risk of heart attacks by leveraging data-driven techniques. Using the Heart Attack Prediction dataset sourced from GitHub, which comprises nearly 9,000 patient records and 25 features, this study seeks to build a robust predictive model. The dataset includes a wide range of variables such as age, income level, stress indicators, physical activity, and other lifestyle and demographic factors.

By applying machine learning algorithms and data analysis techniques, the objective is to identify patterns and correlations that can accurately forecast the likelihood of a heart attack. The goal is to develop a reliable tool that could potentially aid healthcare professionals in risk assessment and early intervention, thereby improving patient outcomes and reducing preventable deaths.

# Introduction

This project is aimed at the robust and accurate prediction of cardiovascular disease (CVD) risk by analyzing not only traditional clinical indicators but also modern lifestyle and psychological parameters. The goal is to facilitate early intervention and preventive care that can benefit patients and reduce the global burden of heart disease. By leveraging data-driven approaches, this project aims to provide valuable insights that could assist healthcare professionals in risk assessment and early diagnosis.

The motivation for this research is that psychological stress, income level, family history, and geographic location (e.g., country of residence) are increasingly relevant in the context of modern lifestyles and may significantly influence cardiovascular outcomes. By incorporating these broader parameters, the project seeks to explore a more holistic and inclusive approach to heart attack risk prediction.

The primary goal of this project is to develop and evaluate machine learning models that can predict heart attack risk using both clinical and non-clinical factors. Specific objectives include:

* Preprocessing and analyzing a dataset that contains clinical, lifestyle, and psychological parameters.
* Identifying the role of modern lifestyle features such as family history, income, stress level, and country in predicting CVD risk.
* Applying and comparing several machines learning classifiers, including Decision Trees, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbors.
* Fine-tuning models using hyperparameter optimization and validating performance with cross-validation techniques.
* Interpreting the model outputs to determine which lifestyle-related features are strong predictors of heart attack risk.

# Background/Related Work

Numerous studies have explored the prediction of cardiovascular disease (CVD) using machine learning and statistical techniques. Traditional predictive models such as the Framingham Risk Score and other risk calculators rely primarily on clinical parameters like age, cholesterol levels, blood pressure, and smoking status. These models have been widely used due to their simplicity and interpretability, but they often fail to capture more complex and less quantifiable factors, such as psychological stress or socioeconomic conditions.

Recent research has introduced machine learning methods like Logistic Regression, Random Forests, Support Vector Machines (SVM), and Neural Networks to improve prediction accuracy. These models outperform traditional scoring systems in many cases due to their ability to detect nonlinear relationships and interactions between variables. For example, studies using the Cleveland Heart Disease dataset have demonstrated that Decision Trees and ensemble methods can yield better predictive performance than rule-based scoring systems.

However, many of these models are limited in scope:

* **Strengths**: They provide high accuracy with clinical data, leverage large datasets, and have been optimized using modern ML techniques.
* **Weaknesses**: They often exclude important lifestyle and psychological variables. Many models are built on narrowly defined datasets lacking diversity in geography, ethnicity, income levels, or mental health considerations. As a result, their applicability to broader, real-world populations can be limited.

This project differentiates itself by explicitly integrating modern lifestyle and psychological factors—such as stress level, income, family history, and geographic origin—into heart attack risk prediction models. By going beyond the conventional clinical focus, the project aims to evaluate the predictive power of these non-traditional variables in combination with standard clinical metrics. This approach acknowledges that heart disease is not only a medical condition but also a product of the broader social and psychological environment in which individuals live.

Additionally, this project contributes to the growing body of research advocating for more holistic, inclusive, and representative models in health informatics. The findings can potentially inform public health strategies, personalized medicine approaches, and early intervention programs that are better tailored to the realities of diverse patient populations.

This project aims to address these gaps by developing an ML model that integrates clinical indicators with modern lifestyle and psychological parameters. By doing so, it seeks to provide a more holistic and accurate prediction of heart attack risk, potentially leading to better-targeted interventions and improved patient outcomes.

# Methods

* 1. **Data Collection**

For this project, I utilized the **Heart Attack Prediction dataset** from GitHub. This dataset contains over 9,000 patient records which tell about the patient’s clinical and non-clinical data as the 25 input attributes (e.g., age, gender, cholesterol, income), and one binary target attribute (Heart Attack Risk) of values 0 and 1. 0 indicates “No Heart Attack Risk” and 1 indicates “Heart Attack Risk”. This dataset is well-suited for binary classification problems and provides a realistic scenario for building predictive models aimed at early diagnosis and risk stratification for heart disease.

**3.2 Data Preprocessing**

During preprocessing, one of the major challenges was handling the **"Blood Pressure"** feature, which was recorded in the combined **systolic/diastolic** format (e.g., 120/80). To make this feature more machine learning–friendly and enable numerical analysis, I **split this column into two separate features**:

* **BP\_systolic**: Extracted systolic pressure values.
* **BP\_diastolic**: Extracted diastolic pressure values.

After successfully creating these two new features, I **removed the original "Blood Pressure" column** from the dataset to avoid redundancy.

Additionally, certain features such as **Gender**, **Diet**, and **Smoking status** were **categorical** in nature and thus not directly usable by most machine learning algorithms, which require numerical inputs. To address this, I applied **Label Encoding** using LabelEncoder() to convert categorical string values into numeric labels. These encoded values were then integrated into the dataset to maintain consistency and ensure compatibility with downstream modeling algorithms.

Furthermore, I **checked for missing values**, outliers, and any inconsistent data entries. Missing values were handled through appropriate imputation techniques such as mean or mode substitution depending on the feature type. Outliers were examined using boxplots and z-score thresholds to assess whether they represented data entry errors or legitimate but extreme values.

A screenshot of a computer

AI-generated content may be incorrect.

Fig – 1: Dataset after the Data Preprocessing

* 1. **Exploratory Data Analysis**

To begin our exploratory data analysis, we first defined the numerical variables: age, cholesterol, systolic blood pressure, and diastolic blood pressure. We then calculated descriptive statistics for each variable, including the mean, median, maximum, minimum, and standard deviation.

A graph of a distribution of patient ages

AI-generated content may be incorrect.

Fig – 2: Example histogram for “age” feature

* 1. **Handling Class Imbalance**

An important observation during EDA was the presence of **class imbalance**—a disproportionately higher number of instances labeled **0 (No Risk)** compared to **1 (Risk)**. This imbalance could severely affect the model’s ability to learn minority class patterns and result in biased predictions favoring the majority class.

To address this, I employed **SMOTE (Synthetic Minority Oversampling Technique)**, which works by generating **synthetic examples** for the minority class rather than merely duplicating existing records. SMOTE synthesizes new samples by interpolating between existing samples in feature space, helping the model learn a more general decision boundary. This not only prevents overfitting but also improves the model’s **recall** for the minority class, which is crucial in medical applications where false negatives can have serious consequences.

A comparison of a graph

AI-generated content may be incorrect.

Fig – 3: Count values of Target feature before and using the SMOTE

* 1. **Modeling & Evaluation**

In this section, I built four supervised learning models— decision tree, random forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN)—to predict whether new samples would be classified as that there is Heart Attack Risk or not. I split the dataset using 70% as the training set and 30% as the test set. I used only the top predictor variables to train the model which got by using the heatmap.

Correlation plots were used to detect multicollinearity and uncover the relationships between variables and the target. Feature selection was guided by a **correlation heatmap**, which highlighted features that had the strongest linear relationships with the target variable.

**Key predictive features identified:**

* **Age**
* **Stress Level**
* **Cholesterol**
* **BP\_systolic and BP\_diastolic**
* **Previous Heart Problems**
* **Obesity**
* **Smoking**
* **Physical Activity per Week**

The correlation heatmap visualized the **Pearson correlation coefficients** between all pairs of variables:

* **Red tones** indicated strong positive correlations.
* **Blue tones** indicated strong negative correlations.
* Values close to **0** showed weak or no correlation.
* The **diagonal line** is all 1s because a variable is always perfectly correlated with itself.
* The **values inside the heatmap cells** are correlation coefficients, ranging from -1 to 1.

A blue and red squares

AI-generated content may be incorrect.

Fig – 4: Correlation heatmap

Once the models were built, I applied **GridSearchCV** for **hyperparameter tuning**, using **10-fold cross-validation** to systematically test various combinations of parameters. This ensured that the model selected was not only accurate but also generalized well to unseen data.

The best-performing model was then evaluated using a comprehensive set of metrics:

* **Accuracy**: Overall correctness of the model.
* **Precision**: Proportion of true positives among predicted positives.
* **Recall**: Proportion of true positives among actual positives.
* **F1 Score**: Harmonic mean of precision and recall.
* **ROC-AUC Curve**: Provided a visual representation of the trade-off between true positive rate and false positive rate at various thresholds.

A **confusion matrix** was also plotted to visualize true positives, false positives, false negatives, and true negatives, helping to assess model performance at a glance.

A close-up of a graph

AI-generated content may be incorrect.

Fig – 5: Confusion Matrices of SVM and Random Forest

# Results/Discussion

In the final stage of the project pipeline, I conducted a comprehensive comparison of the four machine learning models developed in the previous phase: **Support Vector Machine (SVM)**, **K-Nearest Neighbors (KNN)**, **Decision Tree**, and **Random Forest**. The objective was to identify the model that offered the best performance in accurately predicting heart attack risk from the patient data.

To evaluate the models objectively, I compiled their performance metrics into a single, easy-to-interpret **summary table** (refer to **Figure 5**). The evaluation metrics considered were:

* **Accuracy**: The overall proportion of correct predictions.
* **Precision**: The ability of the classifier to avoid false positives.
* **Recall**: The ability of the classifier to identify true positives.
* **F1 Score**: A balance between precision and recall.
* **AUC-ROC**: A measure of the model’s ability to distinguish between the two classes across different thresholds.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **AUC-ROC** |
| SVM | 61% | 58% | 60% | 59% | 0.64 |
| KNN | 59% | 56% | 57% | 56% | 0.61 |
| Decision Tree | 63% | 61% | 62% | 61% | 0.66 |
| Random Forest | **66%** | **64%** | **65%** | **64%** | **0.69** |

Fig – 5: Comparison table of the models with its performance evaluation metrics

Among the four machine learning models **Random Forest** outperformed all others with the highest scores across all metrics with 66% accuracy, indicating strong and balanced performance.

The **Decision Tree** model also performed reasonably well, with **63% accuracy**. However, it was slightly more prone to overfitting compared to Random Forest, which benefits from ensemble averaging and reduced variance.

The **SVM** model provided moderate performance, with a lower accuracy of **61%**, and slightly lower recall and precision. It appeared less effective in distinguishing between classes in this dataset, possibly due to the presence of non-linear relationships that were better captured by tree-based methods.

**KNN** had the **lowest performance overall**, particularly in terms of precision and recall. Its performance was likely impacted by the high dimensionality of the dataset and the presence of overlapping feature distributions between the classes.

# Conclusions

This project sets out to **predict the risk of heart attack** using supervised machine learning algorithms and to identify the most influential factors contributing to cardiovascular disease. Using the publicly available *Heart Attack Prediction* dataset from GitHub, the project involved a complete end-to-end pipeline, including data preprocessing, exploratory data analysis, feature engineering, model training, and performance evaluation.

Four machine learning models—**Support Vector Machine (SVM)**, **K-Nearest Neighbors (KNN)**, **Decision Tree**, and **Random Forest**—were trained on the processed dataset and evaluated using standard classification metrics: accuracy, precision, recall, F1-score, and AUC-ROC. While each model provided reasonable predictive performance, the **Random Forest algorithm outperformed all others**, achieving the highest scores across all metrics, including an accuracy of **66%**, indicating a moderately strong ability to classify patients into heart attack risk categories.

However, several **limitations** affected the outcome of the project:

* **Dataset limitations**: The original dataset consisted of over 9,000 records, but due to inconsistencies, missing values, and biologically implausible data (e.g., extremely low or high blood pressure values), extensive data cleaning was required. This reduced the usable dataset to approximately **5,000 patient records**, which limited the model’s ability to learn complex patterns.
* **Model complexity constraints**: Due to the relatively small dataset, I avoided using deep learning models, which typically require significantly more data to avoid overfitting and perform well.

Future work in this area is to build upon the foundation of this work, **future research** could explore the following:

* **Larger and higher-quality datasets**: Collaborating with healthcare institutions to obtain richer datasets with more clinical variables (e.g., family medical history, ECG results, lab values, medication history) could significantly enhance prediction quality.
* **Advanced machine learning models**: Gradient boosting algorithms (e.g., XGBoost, LightGBM) or neural networks could be applied if sufficient data and computational resources are available.
* **Deployment**: A user-friendly interface or mobile application could be developed to allow individuals to input their personal health data and receive real-time heart attack risk predictions. As more users interact with the model, the **training data would grow**, allowing the model to improve its accuracy over time through continuous learning or re-training on updated datasets.

In conclusion, while the **Random Forest model’s 66% accuracy** indicates a good starting point for heart attack risk prediction, it also highlights the importance of data quality and model complexity in achieving reliable clinical tools. With further development, integration of richer datasets, and continuous optimization, this work has the potential to evolve into a robust, real-world decision support system for cardiovascular health monitoring.