Predicting Heart Attack Risk: A Machine Learning Approach

A Write-Up for the B. Tech Project

Project Supervisor:
Dr. Ashutosh Mani
Associate Professor, Department of Biotechnology

Submitted by:
Abhinav Aggarwal 20200003
Ratna Rathaur 20200041
Shivam Pandey 20200049

Biotechnology Department

Motilal Nehru National Institute of Technology Allahabad, Prayagraj

Uttar Pradesh, India, 211004

Predicting Heart Attack Risk: A Machine Learning Approach

1 Introduction

Cardiovascular diseases, including heart attacks, remains a significant global health concern, contributing to a substantial burden of morbidity and mortality. An estimated 17.9 million people died from cardiovascular diseases in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. Over three-quarters of heart attack-related deaths take place in low- and middle-income countries. Out of the 17 million deaths under the age of 70 which happened due to noncommunicable diseases, 38% were due to heart attacks.

The most prominent risk factors for heart attack are older age, active smoking, high blood pressure, diabetes mellitus, and total cholesterol and high-density lipoprotein levels. Heart attacks often strike suddenly and without warning, making early intervention and prevention paramount in mitigating the associated risks. Conventional risk assessment models, while informative, have limitations in their ability to account for the multifaceted nature of heart attack risk. Machine learning, on the other hand, has the capacity to consider a multitude of risk factors, some of which may be subtle or non-obvious, and thus holds promise in enhancing the accuracy of predictions and facilitating timely interventions.

This project explores the application of machine learning techniques in the prediction, early detection, and prevention of heart attacks. Machine learning, with its ability to analyze vast and complex datasets, uncover hidden patterns, and provide personalized risk assessments, stands at the forefront of a transformative paradigm shift in cardiovascular care. We aim to develop different machine learning models that can be used to predict heart attack by considering 13 potential risk factors for heart attack which have been outlined below. Through machine learning, we seek to unlock new insights and methodologies that hold the potential to save lives and enhance the quality of cardiovascular care.

2 About Myocardial Infarction

A myocardial infarction (MI), or heart attack, occurs when blood flow decreases or stops in one of the coronary arteries of the heart, causing infarction (tissue death) to the heart muscle. The most common symptom is chest pain or discomfort which may travel into the shoulder, arm, back, neck or jaw. Often it occurs in the center or left side of the chest and lasts for more than a few minutes. Other symptoms may include shortness of breath, nausea, feeling faint, a cold sweat, feeling tired, and decreased level of consciousness.

Risk factors:

The most prominent risk factors for myocardial infarction are older age, active smoking, high blood pressure, diabetes mellitus, and total cholesterol and high-density lipoprotein levels. Many risk factors of myocardial infarction are shared with coronary artery disease, the primary cause of myocardial infarction, with other risk factors including male sex, low levels of physical activity, a past family history, obesity, and alcohol use. Risk factors for myocardial disease are often included in risk factor stratification scores, such as the Framingham Risk Score. At any given age, men are more at risk than women for the development of cardiovascular disease. High levels of blood cholesterol is a known risk factor, particularly high low-density lipoprotein, low high-density lipoprotein, and high triglycerides.

Mechanism:

1. Atherosclerosis: The most common underlying cause of heart attacks is atherosclerosis. Atherosclerosis is a gradual process where fatty deposits, cholesterol, calcium, and other substances build up on the inner walls of the coronary arteries. These deposits form plaque, which narrows the arteries and reduces blood flow.
2. Plaque Rupture or Erosion: Within the coronary arteries, the plaque can become unstable and vulnerable to rupture or erosion. When this occurs, the contents of the plaque, including cholesterol and fatty substances, can spill into the bloodstream.
3. Thrombosis: When the plaque ruptures or erodes, it exposes the underlying tissue and triggers the body's natural clotting mechanism. Platelets in the blood rush to the site to form a blood clot (thrombus). This blood clot can rapidly grow and obstruct the coronary artery, further reducing blood flow to the heart muscle.
4. Ischemia: As the blood flow in the coronary artery becomes more restricted due to the blood clot, the heart muscle downstream from the blockage experiences ischemia, a condition characterized by insufficient oxygen supply. Ischemia can cause chest pain (angina), discomfort, or pressure in the chest area.
5. Infarction: If the blood clot is not quickly dissolved or bypassed, the lack of oxygen and nutrients eventually leads to the death of heart muscle cells. This is called an infarction. The size of the infarcted area depends on the location and severity of the blockage. The more extensive the damage, the more severe the heart attack.

3 Materials

3.1 Dataset

Sourced from UC Irvine Machine Learning Repository
(https://archive.ics.uci.edu/dataset/45/heart+disease)
300 Instances, 13 Features and 1 target variable

3.2 Tools and Technology Used

Google Colab, Python, Numpy, Pandas, Matplotlib, Seaborn, SkLearn

4 Method

Logistic regression, a type of generalized linear model, is suitable for predicting binary outcomes, making it ideal for classifying patients into "at risk" or "not at risk" categories for heart attacks. By modeling the relationship between patient characteristics and the likelihood of a heart attack, logistic regression can provide probabilities and is often used for classification tasks, offering valuable insights into risk assessment.

K-Nearest Neighbor is a simple and effective classification algorithm. In this project, KNN can be used to classify patients based on their similarity to other patients who have experienced heart attacks. The algorithm considers the attributes of the 'k' nearest neighbors to determine a patient's risk category, making it a useful tool for patient stratification and risk assessment.

Support Vector Machine (SVM) is a powerful algorithm for classification and regression tasks. For heart attack risk prediction, SVM can be employed to identify complex decision boundaries that separate patients at risk from those not at risk. SVM aims to maximize the margin between these groups, making it a robust method for risk classification based on multidimensional patient data.

Decision trees are a fundamental machine-learning technique for classification and regression tasks. In the project, decision trees can be applied to segment patients based on their attributes, creating a tree-like structure to predict heart attack risk. Decision trees offer interpretability and can help identify key risk factors, making them a valuable method for risk assessment.

A Random Forest classifier is an ensemble learning technique that combines multiple decision trees to enhance prediction accuracy. It can provide a robust and reliable model for heart attack risk prediction by aggregating the results of individual decision trees. It excels in handling noisy data and capturing complex relationships, offering a well-rounded approach to risk assessment.

XGBoost (Extreme Gradient Boosting) is a powerful ensemble learning algorithm known for its high predictive accuracy. An XGBoost classifier can be used to create a strong predictive model that combines multiple decision trees to assess the risk of heart attacks. It is particularly effective in handling complex, nonlinear relationships in the data, making it a valuable tool for accurate risk prediction.

5 Plan of Work

1. Data Collection and Preprocessing
   The Machine Learning dataset has been obtained from UC Irvine Machine Learning Repository. The dataset contains patient data obtained from 4 hospitals. We will be using the reliable Cleveland database. The database has 100% credibility, with actual data obtained from the Cleveland hospital. The dataset consists of 300 instances. There are 13 features which can be comprehensively used to predict heart attack. There are no missing values in the dataset.
2. Feature Selection and Engineering
   The most prominent risk factors for myocardial infarction are older age, active smoking, high blood pressure, diabetes mellitus, and total cholesterol and high-density lipoprotein levels. Thirteen such potential risk factors have been used as features in the dataset, which together can be used to make accurate predictions of heart attack, given below:

Attribute	**Description**
Age	Age of the person (years)
Sex	Sex of the person: 1 = male, 0 = female
Cp	Chest pain type 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic
Trestbps	Resting blood pressure (mm Hg)

Attribute	**Description**
Chol	Cholesterol level (mg/dL)
Fbs	Fasting blood sugar > 120 mg/dL: 1 = true 0 = false
Restecg	Resting electrocardiographic results 0 = normal 1 = having ST - T abnormality 2 = left ventricular hypertrophy
Thalachh	Maximum heart rate achieved (bpm)
Exng	Exercise Induced Angina: 1 = yes, 0 = no
Oldpeak	ST Depression induced by exercise relative to rest
Slp	Slope of the peak exercise ST segment
Ca	Number of major vessels colored by fluoroscopy that ranged between 0 to 3
Thal	Thalassemia (a type of blood disorder) results: 3 = normal 6 = fixed defect 7 = reversible defect
Diagnosis	Diagnosis Classes: 0 = healthy 1 = patient who is subject to possible heart disease

3. Exploratory Data Analysis
    a. General Feature Analysis
    b. Continuous Feature Distribution Analysis
    c. Categorical Feature Distribution Analysis
    d. Correlation Analysis
    e. Outlier Analysis

4. Model Development and Training
   We seek to implement the following machine-learning algorithms:
    a. Logistic regression
    b. K nearest Neighbor
    c. SVM
    d. Decision trees
    e. Random Forest Classifier
    f. XG Boost Classifier

   We will train and validate the models using a suitable metric like accuracy, precision, recall, etc.

5. Model Evaluation and Selection
   We will evaluate model performance using cross-validation and test dataset. We will compare the performance of different models on suitable metrics and select the best-performing one.

6. Interpretability and Insights
   We will employ model interpretability techniques to explain the predictions and identify the most influential features.

6 Expected Outcome

The primary goal of the project is to develop a machine learning model that accurately predicts the risk of heart attacks for individual patients. The expected outcome is a model that can provide reliable risk assessments based on patient data, which can assist healthcare professionals in early diagnosis.

This is done by validating and comparing different machine learning models to determine the most effective one for heart attack risk prediction. This comparative analysis will provide a clear understanding of the model's performance and its applicability in a clinical setting.