

Bellevue University

Heart Attack Prediction | A Data-Driven Approach

MILESTONE 1 – PROJECT 2

Date: 09/27/2023

Aarti Ramani

DSC680-T302 Applied Data Science (2241-1)

Table of Contents

TOPIC	2
BUSINESS PROBLEM.....	2
DATASET	3
METHODS	4
ETHICAL CONSIDERATIONS	5
CHALLENGES & ISSUES.....	5
Challenges:.....	5
Issues:.....	6
REFERENCES:	6

TOPIC

In today's world of healthcare, it's crucial to have a clear understanding of heart attack risks. Heart attacks remain a major cause of loss of life globally, and the ability to predict who might be at higher risk can make a significant difference in preventing these life-threatening events. This project aims to take a data-driven approach to forecast heart attack risk, offering valuable insights into the field of cardiovascular health.

BUSINESS PROBLEM

The problem with heart attacks, also known as myocardial infarctions, is multifaceted and has significant implications for public health, individuals, and healthcare systems. The following is an overview of the key issues associated with heart attacks:

- **High Mortality Rate:** Heart attacks are a leading cause of death worldwide. They can be sudden and fatal, and survival often depends on prompt medical intervention. Despite advances in treatment, many heart attack victims do not survive or suffer severe consequences.
- **Lifestyle-Related Risk Factors:** Unhealthy lifestyle choices, including a poor diet, lack of physical activity, smoking, and excessive alcohol consumption, significantly increase the risk of heart attacks. These behaviors are prevalent in many societies and contribute to the problem.
- **Aging Population:** As the global population ages, the incidence of heart attacks is expected to rise. Aging is a non-modifiable risk factor, and older individuals are more susceptible to heart disease.

- **Healthcare Costs:** Heart attacks incur substantial healthcare costs, including hospitalizations, surgeries, medications, and long-term care. The financial burden on healthcare systems, insurers, and individuals can be overwhelming.
- **Treatment Gaps:** Access to timely and effective treatment is critical during a heart attack. Rural or underserved areas may lack adequate healthcare facilities, leading to treatment delays and poorer outcomes.

Addressing this problem requires a holistic approach that includes prevention, early detection, access to quality healthcare, and a focus on improving the overall cardiovascular health of individuals and populations.

DATASET

This dataset is sourced from Kaggle([Fahad Mehfooz.](#)). This dataset has around 14 features with around 300 rows. Following are the features available in this dataset.

- **Age :** Age of the patient
- **Sex :** Sex of the patient
- **exang:** exercise induced angina (1 = yes; 0 = no)
- **ca:** number of major vessels (0-3)
- **cp :** Chest Pain type chest pain type
 - Value 1: typical angina
 - Value 2: atypical angina
 - Value 3: non-anginal pain
 - Value 4: asymptomatic
- **trtbps :** resting blood pressure (in mm Hg)
- **chol :** cholestoral in mg/dl fetched via BMI sensor
- **fbs :** (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

- `rest_ecg` : resting electrocardiographic results
 - Value 0: normal
 - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
 - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- `thalach` : maximum heart rate achieved
- `target` : 0= less chance of heart attack 1= more chance of heart attack

Given the limited number of rows in the current dataset, should the model's outcomes exhibit inconsistency, I intend to explore an alternative dataset on Kaggle, one with a larger number of rows, approximately 900 ([Fedesoriano](#)).

METHODS

Before feeding data into the models, we will undertake several data preprocessing steps such as handling missing data, dropping duplicates, etc., to ensure the dataset's quality and relevance. With the cleansed dataset, I plan on building the following models for this project.

- **Logistic Regression:** Logistic regression is a simple and interpretable model that is often used for binary classification tasks like predicting heart attacks (0 or 1). It can provide insight into how each feature contributes to prediction.
- **Random Forest:** Random Forest is an ensemble model that is robust, handles non-linearity well, and works with both categorical and numerical features. It can capture complex relationships in the data and provide feature importance rankings.
- **Support Vector Machine (SVM):** SVMs work well for binary classification and can handle both linear and non-linear data. They are especially useful when dealing with high-dimensional data.

- **Naive Bayes:** Naive Bayes is a probabilistic classifier that works well with categorical features and can be computationally efficient. It's considered a good choice when dealing with discrete clinical data.

ETHICAL CONSIDERATIONS

Predictive modeling for heart attack prediction, like any healthcare-related project, comes with several ethical considerations that need to be carefully addressed to ensure responsible and ethical use of patient data and to prioritize patient well-being. Following are some key ethical considerations:

- i. It is imperative to ensure patient data used for training and testing the model are kept anonymous and comply with relevant privacy regulations (e.g., HIPAA in the United States). Informed consent from patients needs to be obtained if their data is used for research purposes, and clearly communicate how their data will be used.
- ii. Robust data security measures to protect patient data from unauthorized access, breaches, or cyberattacks should be implemented.
- iii. Collaboration with healthcare professionals and legal experts should be conducted to ensure that the project aligns with ethical standards and legal requirements.

CHALLENGES & ISSUES

Challenges:

- The availability of high-quality, and well-labeled healthcare data can be a significant challenge. Incomplete or inaccurate data can lead to unreliable predictions.

- Heart attack datasets may be imbalanced, with a significantly larger number of patients who haven't experienced heart attacks (target 0) compared to those who have (target 1). This can lead to biased models.
- Clinically validating the predictive model to ensure its safety, and efficacy in real-world healthcare settings is resource-intensive and can be time-consuming.

Issues:

Patient health statuses change over time, and models need to adapt to evolving data to remain accurate and up to date.

REFERENCES:

Fahad Mehfooz. *HeartAttack prediction with 91.8 % Accuracy, Kaggle* -

<https://www.kaggle.com/code/fahadmehfooz/heartattack-prediction-with-91-8-accuracy/input?select=heart.csv>

Fedesoriano - *Heart Failure Prediction Dataset, Kaggle* -

<https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data>