Georgia State University

CSC4780/6780&DSCI4780 – Fundamentals of Data Science

Fall-2024

Project Progress Report

Heart Attack Prediction

The Prediction Wizards

Nikitha Rajendran

Venkata Dheeraj Bhogi

Benedict Antonio Mervyn

Table of Contents:

1. Business Understanding

1.1 Business Problem

Cardiovascular diseases (CVDs), including heart attacks and strokes, are the leading causes of mortality worldwide, accounting for approximately 17.9 million deaths each year, with a significant portion of these fatalities occurring in individuals under the age of 70. Given the widespread nature of this problem, early detection and timely intervention are crucial to improving patient outcomes. Heart attacks, in particular, can be life-threatening if not detected early, making it imperative for medical practitioners to identify individuals at high risk before an incident occurs. Current diagnostic methods, although effective, can often be time-consuming and may not always provide immediate insights for timely interventions.

The business problem we aim to tackle is to predict the likelihood of a heart attack in patients using clinical and demographic data. A predictive model will allow healthcare providers to identify individuals at high risk, enabling early intervention, personalized treatment plans, and ultimately saving lives. By implementing a data-driven solution, we aim to support healthcare professionals in their decision-making processes, potentially transforming the way heart attacks are predicted and managed in clinical settings.

1.2 Dataset

The dataset consists of 1,319 patient samples, each containing a combination of clinical and demographic features related to cardiovascular health. The below mentioned eight features provide detailed insights into the physiological state of each patient, including vital signs, biochemical markers, and demographic factors. Together, they allow for the analysis and prediction of heart attack risk, making this dataset a valuable resource for understanding and modeling cardiovascular conditions.

Features:

- age: The age of the patient (integer).
- gender: Gender of the patient (boolean, 0 for female & 1 for male).
- impulse: The heart rate of the patient (integer).
- pressurehight: It is the pressure in blood vessels when the heart beats known as systolic blood pressure (integer).
- pressurelow: It is the pressure in blood vessels when the heart is at rest between beats known as diastolic blood pressure (integer).
- glucose: A measure of glucose in the blood (float).
- kcm: A marker enzyme found in heart muscle cells that can indicate heart injury known as Creatine Kinase-MB (float).
- troponin: It is a protein complex in heart muscle cells that regulates contraction and is released into the bloodstream when the heart muscle is damaged.

Label:

- class: The label feature indicating whether the patient has experienced a heart attack or not (boolean, positive and negative).

1.3 Proposed Analytics Solution

The goal of this project is to develop a machine learning model to classify patients as "at risk" or "not at risk" for a heart attack. We will be implementing an array of models which include but not limited to the following:

- **Logistic Regression**: A fundamental model identifying linear relationships.
- **Decision Tree**: An interpretable, rule-based model for decision-making.
- **Random Forest**: An ensemble method enhancing accuracy and reducing overfitting.
- **Support Vector Machine (SVM)**: A model that maximizes class separation with optimal hyperplanes.

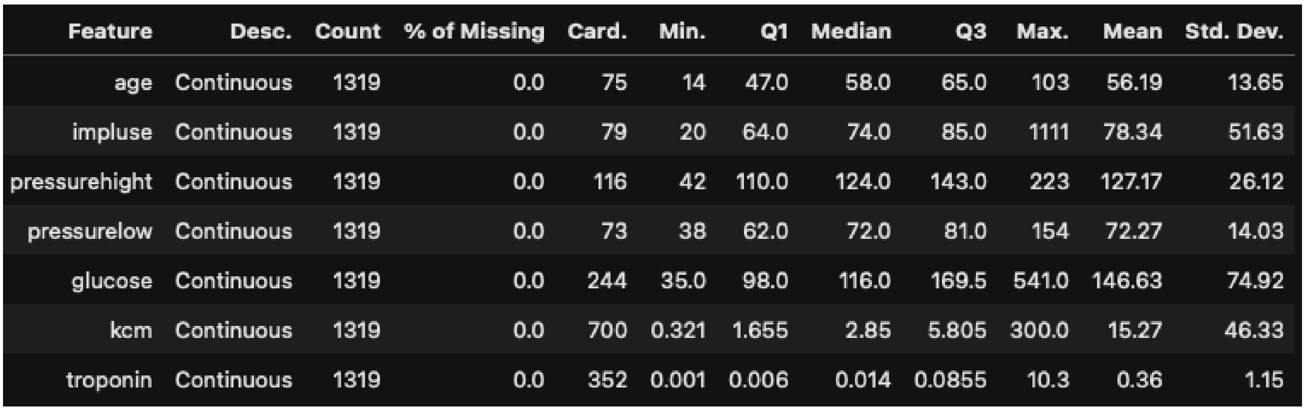
The models will be evaluated using metrics like accuracy, precision, recall, F1-score, and ROC-AUC to identify the best-performing one. The final model will help healthcare providers assess heart attack risk efficiently, enabling early intervention and personalized care, with potential integration into clinical decision-making tools.

2. Data Exploration and Preprocessing

2.1 Data Quality Report

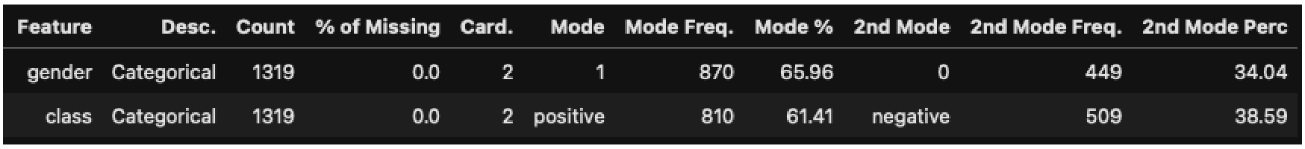
The dataset contains in total 8 descriptive features which are continuous except for gender feature which is encoded to 0 and 1 for female and male respectively. Please find below data quality reports for continuous and categorical features.

Continuous features

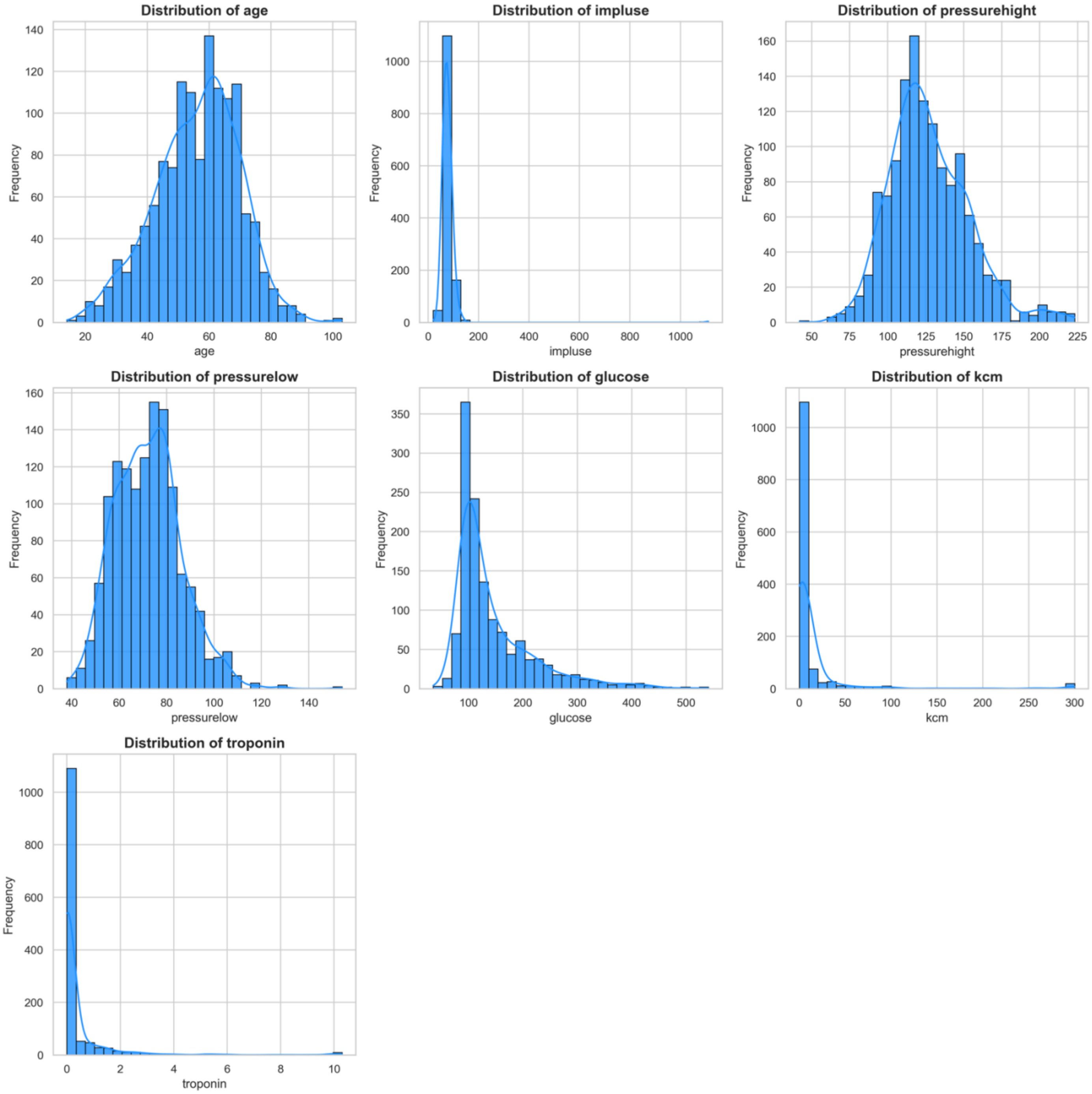


Feature	Desc.	Count	% of Missing	Card.	Min.	Q1	Median	Q3	Max.	Mean	Std. Dev.
age	Continuous	1319	0.0	75	14	47.0	58.0	65.0	103	56.19	13.65
impluse	Continuous	1319	0.0	79	20	64.0	74.0	85.0	1111	78.34	51.63
pressurehight	Continuous	1319	0.0	116	42	110.0	124.0	143.0	223	127.17	26.12
pressurelow	Continuous	1319	0.0	73	38	62.0	72.0	81.0	154	72.27	14.03
glucose	Continuous	1319	0.0	244	35.0	98.0	116.0	169.5	541.0	146.63	74.92
kcm	Continuous	1319	0.0	700	0.321	1.655	2.85	5.805	300.0	15.27	46.33
troponin	Continuous	1319	0.0	352	0.001	0.006	0.014	0.0855	10.3	0.36	1.15

Categorical features:



Feature	Desc.	Count	% of Missing	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode Perc
gender	Categorical	1319	0.0	2	1	870	65.96	0	449	34.04
class	Categorical	1319	0.0	2	positive	810	61.41	negative	509	38.59


Distribution of age

Distribution of impluse

Distribution of pressurehight

Distribution of pressurelow

Distribution of glucose

Distribution of kcm

Distribution of troponin

Figure: Continuous features distribution.

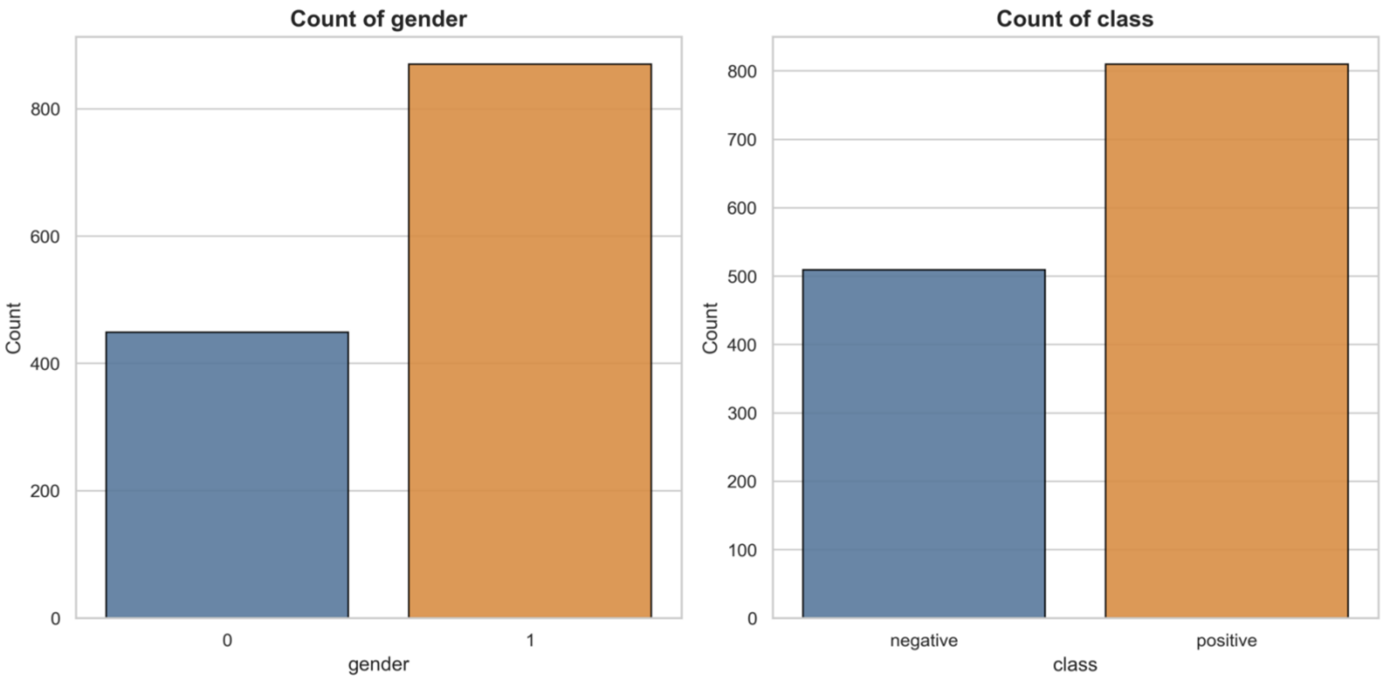


Figure: Categorical features distribution

2.2 Missing Values and Outliers

The dataset was confirmed to have no missing values based on the data quality report. However, several features contained extreme outliers that required handling:

- **Impluse**: Values exceeding 200 bpm, considered rare and likely measurement errors, were capped to maintain data consistency while retaining valid observations.
- **Pressurehight and Pressurelow**: Outliers were capped to address potential errors while preserving significant values indicative of high-risk cardiovascular conditions.
- **Glucose**: High values, likely from diabetic patients, were capped to manage extreme outliers while retaining critical data points related to elevated risk.
- **KCM and Troponin**: Significant outliers were retained, as they provide essential information for heart attack prediction. Log transformations were applied to normalize their distributions.

To address these outliers, the Interquartile Range (IQR) method was employed, and values outside the acceptable range were capped. This approach effectively managed outliers while preserving valuable information crucial for diagnosing heart conditions.

2.3 Normalization

Following normalization was applied to prepare the dataset for modeling:

- Scaling was performed on age, impluse, pressurehight, and pressurelow using StandardScaler to standardize them to a mean of 0 and a standard deviation of 1.
- Log normalization was applied to glucose, kcm, and troponin to reduce skewness and normalize their distributions.

These normalization techniques ensured that the features were on comparable scales, improving consistency and model performance.

2.4 Transformation

Following feature engineering was performed to enhance the predictive capabilities:

- A new feature bp_ratio was created by calculating the ratio of systolic to diastolic blood pressure (pressurehight / pressurelow).
- The impluse feature was categorized into Low, Normal, and High to classify heart rates into interpretable ranges.
- the age feature was grouped into Young, Middle-Aged, and Elderly categories based on predefined age ranges. These transformations enriched the dataset by adding derived features to improve model performance.