Heart Disease Risk Assessment: A Machine Learning Approach

Mingju Kim, Ling Li, Lei Chen

October 12, 2023

1. Introduction

In the United States, heart disease stands as a predominant cause of both male and female mortality, accounting for 35% of total female deaths in 2019, as per Ikuta et al. (2022). Within the broad spectrum of heart disease, heart attacks, medically referred to as myocardial infarctions, emerge as a significant category. A heart attack transpires when a segment of the heart muscle experiences an insufficiency in blood flow, thereby making it a pivotal contributor to the overarching heart disease burden. Shockingly, the annual incidence of heart attacks surpasses 750,000 cases, as reported by Benjamin et al. (2017). Regrettably, more than one-fifth of individuals grappling with heart attacks succumb to this condition, with an even more disheartening statistic indicating that almost half of those fatalities transpire suddenly, often occurring prior to their arrival at a hospital, as documented by Go et al. (2013) and Ornato and Hand (2001).

Considering the alarming fatality rates and the narrow window for effective intervention in cases of heart attacks, early detection of such events emerges as a pivotal prerequisite for preventive therapeutic measures. Several well-documented risk factors, including but not limited to high blood pressure, diabetes, smoking, age, and gender, have been identified in prior studies, as elucidated in Becker (2005). However, the contemporary landscape presents a formidable challenge: the burgeoning volume of real-world data, replete with diverse features, makes traditional analytical approaches increasingly inadequate. The modern landscape poses a significant challenge, as the growing volume of diverse real-world data renders conventional analytical methods insufficient. Consequently, in this data-rich era, the rise of data mining techniques has empowered machine learning to play a central role in medicine. This intricate fusion of data-driven methodologies with the medical field holds the potential to revolutionize our approach to heart attack detection and prevention.

Compelling evidence underscores the significance of well-known risk factors like high blood pressure, high cholesterol, diabetes, obesity, unhealthy eating, a lack of physical activity, and smoking in the development of ischemic heart disease. Yet, there are equally important but often neglected risk factors, encompassing psychological, social, economic, and cultural factors, influenced by gender, that seem to contribute to cardiovascular disease, especially in women (Roth et al. (2017); Jenkins and Ofstedal (2014)). The Jackson Heart Study involving 5,301 African American participants has shown a stronger link between adult socioeconomic status and cardiovascular disease risk in women compared to men (Gebreab et al. (2015)). To advance the development of innovative strategies for early detection and precise management of heart attacks in women, it is crucial to prioritize the creation of new prediction models that are highly sensitive to the female population.

The primary aim of this study is to anticipate the occurrence of heart attacks in women as opposed to men, employing a range of distinct attributes. We focus on addressing two key inquiries:

1. Are there variations in heart attack frequency based on gender?

2. Is there differentiation in the heart attack prediction utilizing medical assessments and lifestyle factors between men and women?

2. Related Work

Data scientists have harnessed the power of machine learning to create numerous models for diagnosing, detecting, and predicting a range of diseases, including heart attacks. As an illustration, Commandeur et al. (2020) employed machine learning to integrate clinical data with quantitative imaging-based variables, resulting in a substantial enhancement in the accuracy of predicting myocardial infarctions and cardiac mortality. Additionally, Than et al. (2019) integrated age, sex, and paired high-sensitivity cardiac troponin I concentrations from patients, enabling the prediction of acute myocardial infarction likelihood.

Nonetheless, there remains a dearth of risk-assessment models that account for risk factors unique to females. Several conditions exclusive to women, such as obstetric and gynecological history encompassing gestational hypertension, gestational diabetes, and polycystic ovary syndrome (Vogel et al. (2021)), contribute to an elevated risk of cardiovascular disease. Therefore, to foster innovative advancements in early heart attack detection for women, greater emphasis must be placed on unraveling the influence of diverse risk factors on sex-specific disparities. Our project is dedicated to the prediction of heart attacks, particularly in women, employing a range of machine learning algorithms.

3. Preprocessing and Exploratory Data Analysis

3.1 Data source

We get the 'Cardiovascular Disease' dataset from Kaggle website [1]. Our choice of this particular dataset stems from several compelling reasons.

First and foremost, this dataset provides a comprehensive collection of features directly relevant to our research objectives, which primarily revolve around the prediction and analysis of cardiovascular disease. The dataset encompasses a wide array of essential attributes, including age, gender, blood pressure, smoking habits, and alcohol consumption, all of which are pivotal factors in the context of cardiovascular health. These features not only align closely with our research goals but also allow us to conduct a thorough and meaningful analysis of the problem at hand.

Furthermore, the dataset is a product of meticulous curation, consolidating information from two esteemed primary sources. It seamlessly merges data from the UCI Machine Learning Repository and the Kaggle Heart Disease Dataset by YasserH. These sources are recognized for their reliability and accuracy, providing a strong foundation for our research project. The convergence of data from such reputable origins instills confidence in the quality and integrity of the dataset.

Ethical considerations also played a crucial role in our choice. The dataset has been thoughtfully anonymized and follows established ethical guidelines. This anonymization ensures the privacy and confidentiality of individuals whose data is included, aligning with the highest standards of data ethics and safeguarding the sensitive nature of health-related infor-

[1] Cardiovascular Disease. https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease.

mation.

As illustrated in Table 1, the variables within our dataset can be logically categorized into distinct segments, each offering valuable insights into the complex landscape of cardiovascular health. This categorization serves as a foundational framework for our analysis, providing clarity and structure to the multifaceted data.

1. Cardiovascular Health Status: At the heart of our analysis lies the differentiation between patients who either exhibit signs of cardiovascular disease or remain free from such conditions. This pivotal classification is encapsulated by the target variable 'cardio,' which takes the binary values of 0 and 1, representing the absence or presence of cardiovascular disease, respectively.

2. Demographic Profile: The dataset also encompasses a set of demographic attributes that furnish essential context to each patient's profile. This includes age, gender, height, and weight, which collectively paint a comprehensive picture of the individuals under examination. These attributes play a crucial role in understanding how various demographic factors may influence cardiovascular health.

3. Medical Diagnostic Measures: To delve deeper into the health status of our subjects, the dataset incorporates medical test results, such as blood pressure, cholesterol levels, and glucose levels. These metrics provide quantitative insights into the physiological aspects of each patient's health, shedding light on the potential risk factors associated with cardiovascular disease.

4. Lifestyle and Behavioral Factors: Cardiovascular health is intricately linked to an individual's lifestyle choices and behaviors. The dataset captures pertinent information in this regard, encompassing variables such as smoking habits, alcohol intake, and physical activity. These lifestyle factors contribute to our understanding of how personal choices and habits may impact cardiovascular health outcomes.

Table 1: The meaning of each variable in the dataset

Variables	Description
id	Unique identifier for each patient
age	Age of the patient in days
age_years	Age of the patient in years
gender	Gender of the patient (1 for female and 2 for male)
height	Height of the patient in centimeters
weight	Weight of the patient in kilograms
ap_hi	Systolic blood pressure
ap_lo	Diastolic blood pressure
cholesterol	Cholesterol level (1: normal, 2: above normal, 3: well above normal)
gluc	Glucose level (1: normal, 2: above normal, 3: well above normal)
smoke	Whether the patient smokes (0 for no, 1 for yes)
alco	Whether the patient consumes alcohol (0 for no, 1 for yes)
active	Whether the patient is physically active (0 for no, 1 for yes)
bmi	Body Mass Index of the patient
bp_category	Blood pressure category based on the patient's systolic and diastolic values
bp_category_encoded	Encoded representation of the blood pressure category
cardio	Presence of cardiovascular disease (0 for no, 1 for yes)

3.2 Preprocessing

In our data preprocessing phase, we conducted a series of operations to ensure the quality and suitability of the dataset for our analysis. These steps were executed systematically to enhance the dataset's integrity and relevance. The key preprocessing steps undertaken are detailed as follows:

1. Handling Missing Values: We began by meticulously inspecting the dataset for any missing or null values. Fortunately, our dataset exhibited completeness and contained no instances of missing data, ensuring a robust foundation for analysis.

2. Feature Selection and Removal: In the pursuit of optimizing our dataset, we made judicious decisions regarding feature inclusion and exclusion. The following features

were identified for removal, accompanied by justifications for their exclusion:

   - 'bp_category': It was observed that two columns, 'bp_category' and 'bp_category_encoded,' essentially contained the same information. Given this redundancy, we opted to retain only one of these columns to streamline the dataset.
   - 'age': The 'age' column, representing a patient's age in days, was found to overlap in function with the 'age_years' column, which provides the age in years. The 'age_years' column aligns more conveniently with analytical purposes. Consequently, the 'age' column was removed to eliminate redundancy.
   - 'id': Upon scrutiny, it was confirmed that the 'id' column exhibited no duplicate values. Since this feature did not contribute to the prediction of the target variable 'cardio,' it was considered superfluous and subsequently removed.

3. Encoding of 'bp_category_encoded': The 'bp_category_encoded' column, which presumably represents various stages of blood pressure with a discernible order, was subjected to encoding. Ordinal encoding was chosen as the most suitable technique, as it retains the natural order of the categories and aligns well with numerous machine learning algorithms, particularly those based on decision trees.

After the above processing, we get the following description of data characteristics, as shown in Table 2.

Table 2: Data characteristics description

	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc
count	68205.0	68205.0	68205.0	68205.0	68205.0	68205.0	68205.0
mean	1.3	164.4	74.1	126.4	81.3	1.4	1.2
std	0.5	8.2	14.3	16.0	9.1	0.7	0.6
min	1.0	55.0	11.0	90.0	60.0	1.0	1.0
25%	1.0	159.0	65.0	120.0	80.0	1.0	1.0
50%	1.0	165.0	72.0	120.0	80.0	1.0	1.0
75%	2.0	170.0	82.0	140.0	90.0	1.0	1.0
max	2.0	250.0	200.0	180.0	120.0	3.0	3.0

	smoke	alco	active	cardio	age_years	bmi	bp_category_encoded
count	68205.0	68205.0	68205.0	68205.0	68205.0	68205.0	68205.0
mean	0.1	0.1	0.8	0.5	52.8	27.5	2.9
std	0.3	0.2	0.4	0.5	6.8	6.0	0.9
min	0.0	0.0	0.0	0.0	29.0	3.5	1.0
25%	0.0	0.0	1.0	0.0	48.0	23.9	3.0
50%	0.0	0.0	1.0	0.0	53.0	26.3	3.0
75%	0.0	0.0	1.0	1.0	58.0	30.1	3.0
max	1.0	1.0	1.0	1.0	64.0	298.7	4.0

3.3 Exploratory Data Analysis

We commence our analysis by focusing on the crucial target variable ’cardio,’ which serves as an indicator of the presence of cardiovascular disease within our dataset. In Figure 1, the distribution of this variable is visually depicted, employing a binary encoding where ’0’ signifies the absence of cardiovascular disease, while ’1’ signifies its presence. Notably, we observe that the dataset exhibits a near-even distribution of patients with and without cardiovascular disease. This balanced target distribution highlights the equitable representation of both positive and negative cases, setting the stage for a robust and meaningful analysis.

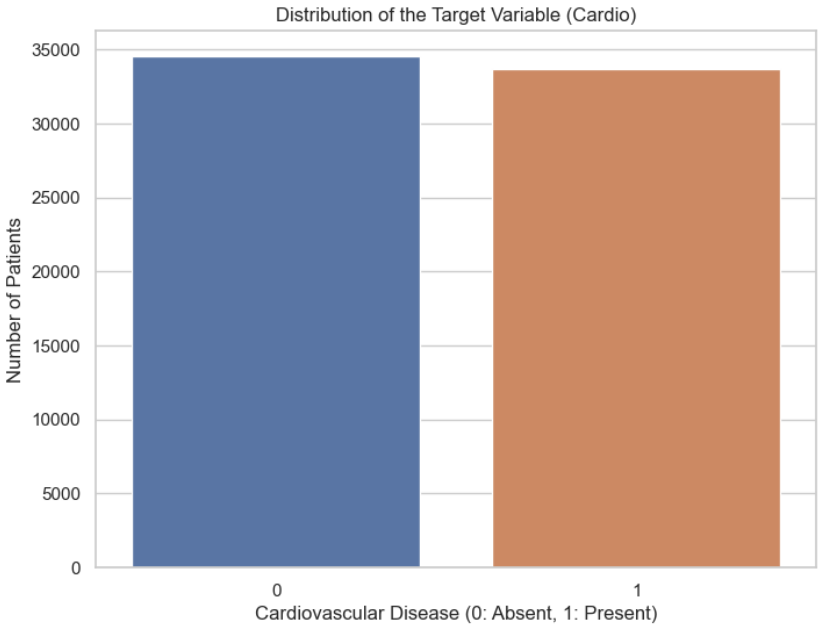


Figure 1: Distribution of the target variable

Then we explore the distribution of categorical features in Figure 2 including gender, cholesterol, gluc, smoke, alco, active, and bp_category_encoded.

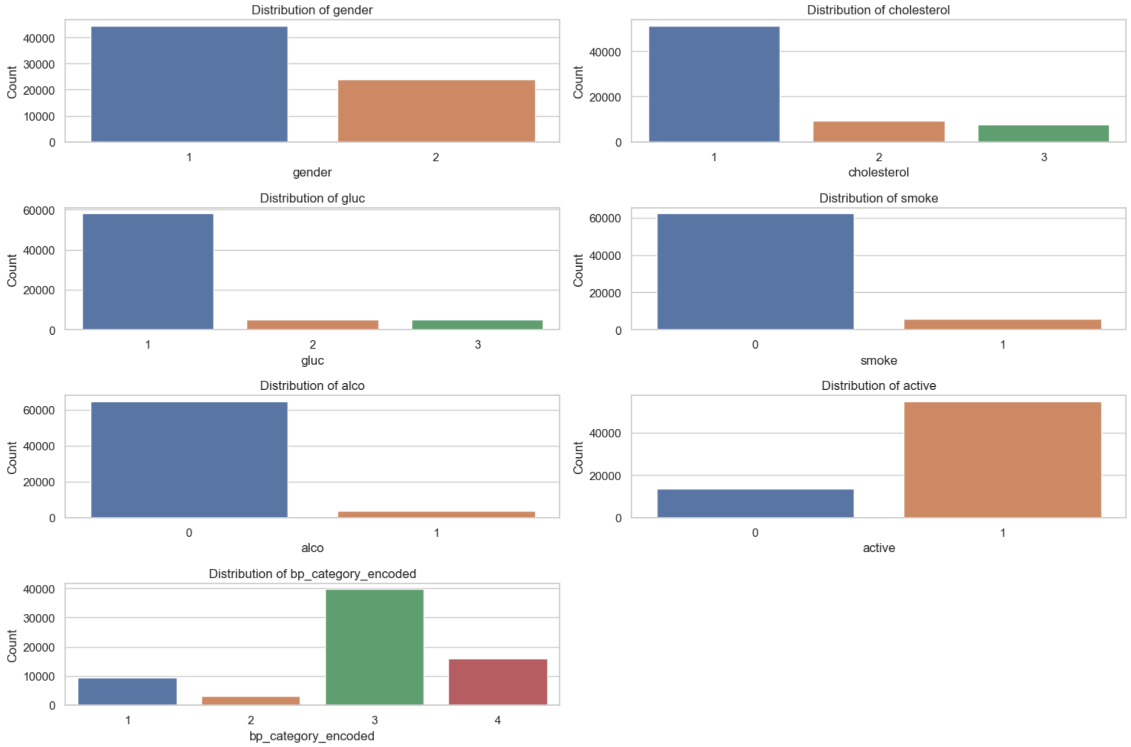


Figure 2: Distribution of categorical features

- Gender: Within the dataset, we observe an imbalance in gender representation. Female appears more frequently than male. However, the dataset does not provide additional context to clarify the specific gender assignment for each type, leaving us with a noticeable gender distribution discrepancy to consider in our analysis.

- Cholesterol: The data reveals that a substantial proportion of individuals exhibit Cholesterol level 1, which typically corresponds to a normal or healthy range. In contrast, Cholesterol levels 2 and 3 are less prevalent, with level 3 being slightly more frequent than level 2. This distribution underscores the predominance of individuals with normal cholesterol levels in the dataset.

- Glucose (gluc): The majority of individuals within the dataset exhibit a glucose level of 1, indicating that this level is the most prevalent, often associated with a "normal" range. Glucose levels 2 and 3 are less common, with level 3 being the least frequent.

This distribution suggests that most individuals in the dataset have glucose levels within the normal range.

- Smoke: The dataset depicts a significant disparity in smoking habits. The vast majority of individuals do not smoke, while only a comparatively smaller proportion are identified as smokers. This notable contrast in smoking behaviors emphasizes the predominance of non-smokers in the dataset.

- Alcohol (alco): Similar to the smoking distribution, the data illustrates that most individuals abstain from alcohol consumption, with only a smaller fraction categorized as alcohol consumers. This observation signifies a significant imbalance in alcohol consumption within the dataset.

- Physical Activity (active): The dataset categorizes individuals into two groups based on their level of physical activity. A substantial majority fall into the ”active” category, indicating a physically active lifestyle. A smaller proportion are categorized as ”inactive.” This notable contrast highlights the prevalence of individuals with active lifestyles in the dataset.

- Blood Pressure Category (bp_category_encoded): The distribution of individuals across blood pressure categories reveals distinct patterns. ”Hypertension Stage 1” is the most common category, followed by ”Normal,” and ”Hypertension Stage 2.” The remaining categories are represented by significantly fewer individuals. This distribution provides insights into the varying prevalence of different blood pressure stages within the dataset, with ”Hypertension Stage 1” being the most prevalent.

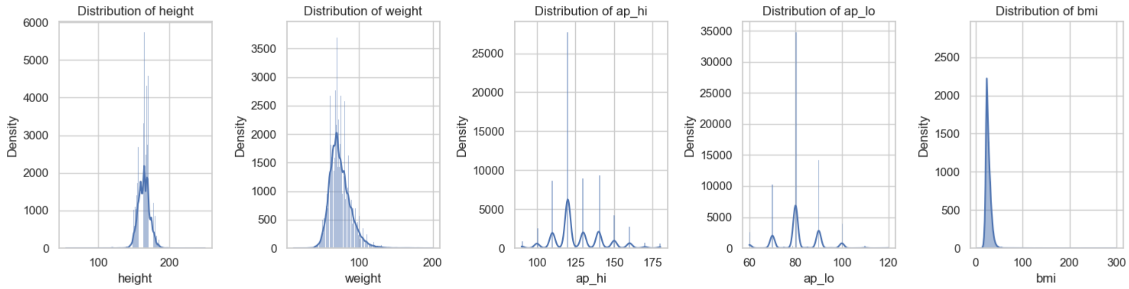


Figure 3: Distribution of continuous features

Now we differentiate the feature distribution based on the presence of cardiovascular disease, as shown in Figure 4.

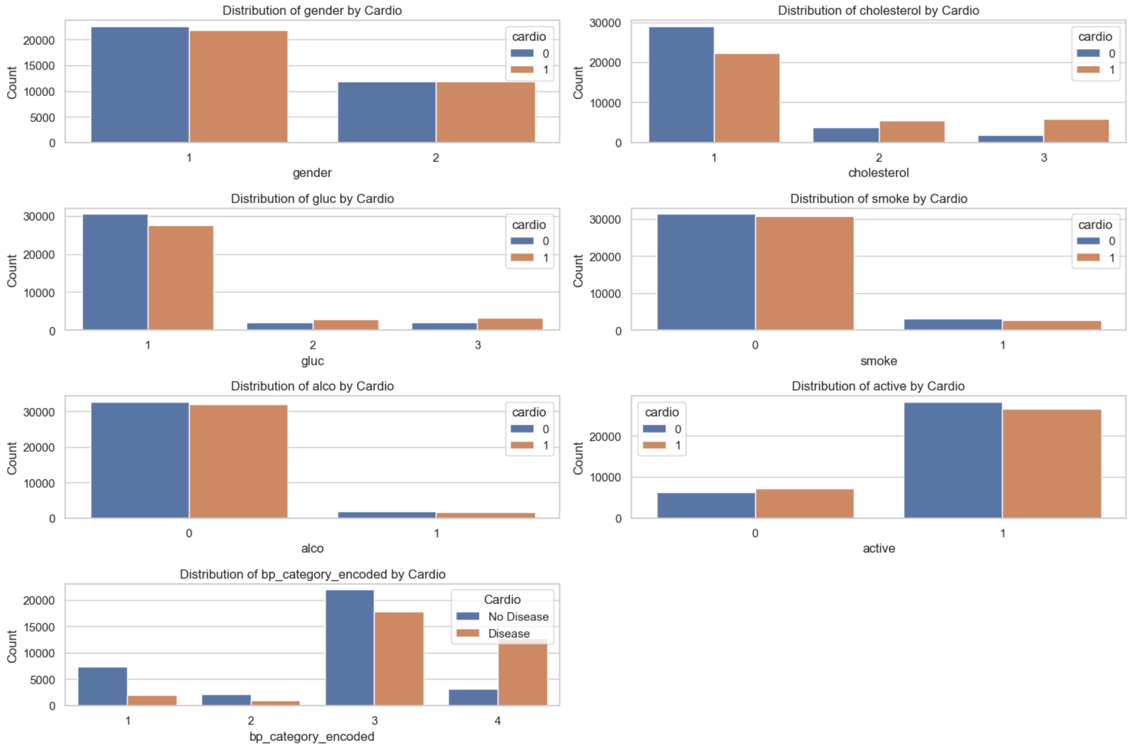


Figure 4: Distribution of categorical features by the presence of cardiovascular disease

In our dataset, various factors play a role in the presence of cardiovascular diseases. Gender types show a relatively balanced distribution, with both types having individuals both with and without cardiovascular diseases, albeit with a slightly higher prevalence in gender type 1. The relationship between cholesterol levels and cardiovascular diseases is intriguing. Individuals with cholesterol level 1 exhibit a fairly even distribution of cardiovascular disease presence. However, for cholesterol levels 2 and 3, there is a noticeable increase in the prevalence of cardiovascular disease. A similar pattern is observed with glucose levels; level 1 displays a balanced distribution, while levels 2 and 3 show a slightly higher prevalence of the disease. The balance between smokers and non-smokers in terms of cardiovascular disease prevalence is also noteworthy. Whether individuals smoke or not, there is an equitable distribution. Likewise, cardiovascular disease distribution is comparable for alcohol consumers and non-consumers. Active individuals have a marginally lower prevalence of cardiovascular diseases compared to non-active ones.

Furthermore, the data reveals intriguing insights when exploring blood pressure categories.

Individuals categorized under ”Hypertension Stage 1” exhibit an even distribution of cardiovascular disease presence. In contrast, those in the ”Normal” category display a lower prevalence of the disease. However, individuals in the ”Hypertension Stage 2” and other categories show a higher prevalence of cardiovascular diseases. These nuanced patterns within our dataset offer valuable insights into the interplay between various factors and the presence of cardiovascular diseases, forming a foundation for in-depth analysis and potential interventions.

The main insights are as follows. Cholesterol levels seem to be a significant indicator of cardiovascular disease risk. As the cholesterol level increases, the risk of having the disease also appears to rise. Activity level also seems to play a role. Being active might be associated with a lower risk of cardiovascular diseases. While smoking and alcohol consumption are often considered risk factors for many health conditions, in this dataset, they don’t show a pronounced difference in cardiovascular disease prevalence between their categories. The blood pressure category provides insights into the relationship between blood pressure stages and cardiovascular diseases. Those in the ”Normal” range seem to have a lower risk, while individuals in higher stages of hypertension show a higher risk.

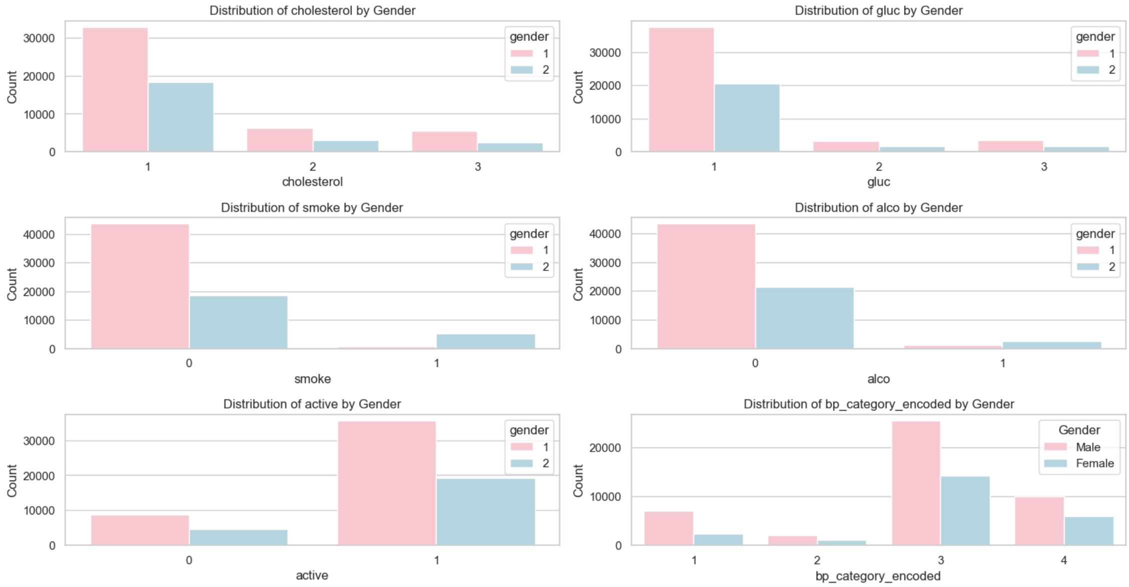


Figure 5: Distribution of categorical features by gender

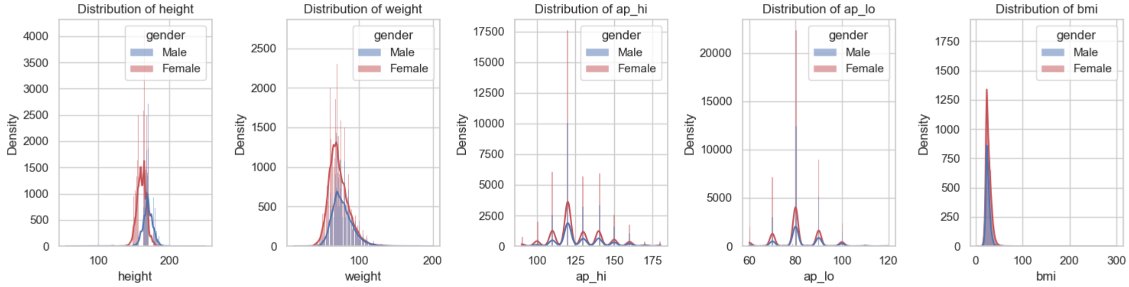


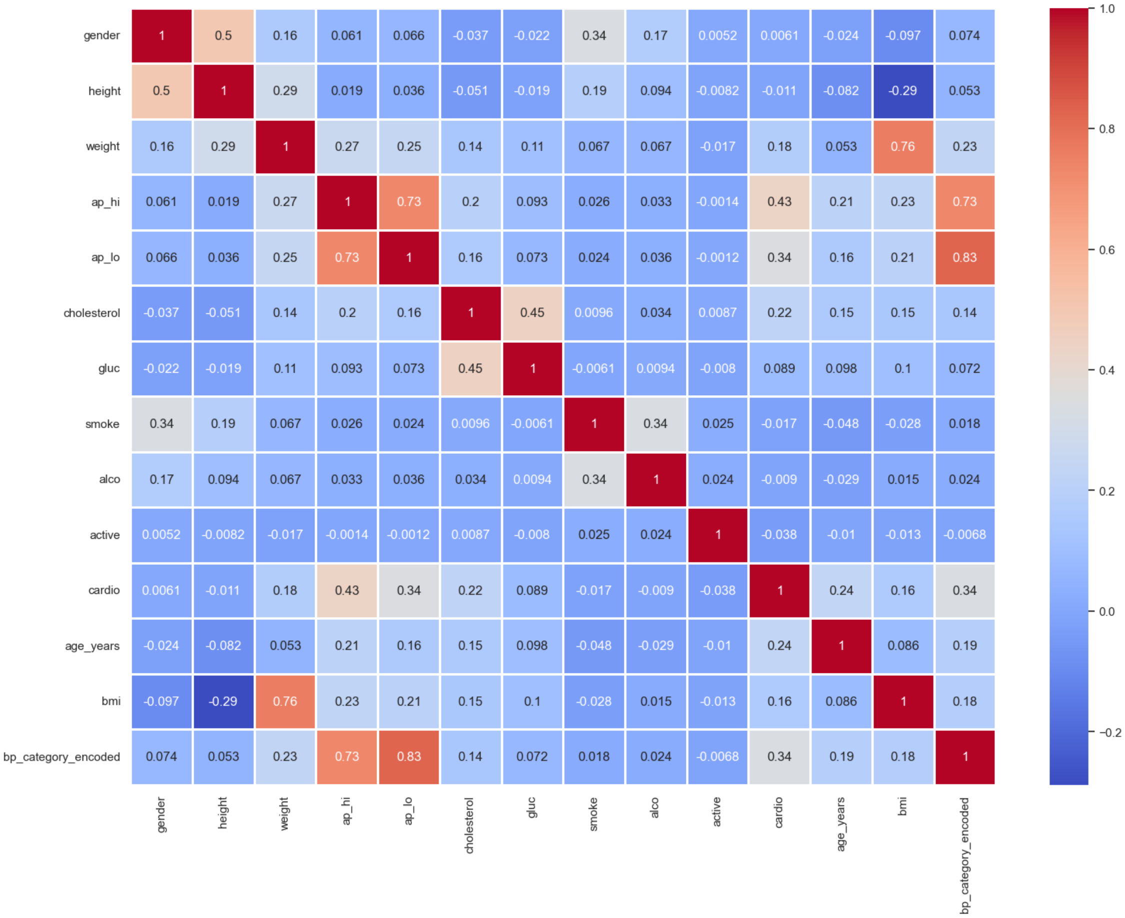
Figure 6: Distribution of continuous features by gender

We further Examine the correlation of variables, as represented in Figure 7. Key observations can be summarized as follows:

1. High Positive Correlations: Notably, 'age' and 'age_years' exhibit a perfect positive correlation with a coefficient of 1. This strong correlation is unsurprising, as 'age_years' is likely a derived value from 'age,' resulting from the conversion of days to years. Similarly, 'weight' and 'bmi' (Body Mass Index) share a robust positive correlation, indicating that as weight increases, BMI tends to rise as well. This correlation is as expected, considering that BMI is inherently a function of both weight and height.

2. Moderate Positive Correlations: Within the dataset, 'ap_hi' (systolic blood pressure) displays a moderate positive correlation with 'cardio,' suggesting that higher systolic blood pressure values may be associated with the presence of cardiovascular diseases. Additionally, 'cholesterol' and 'cardio' demonstrate a moderate positive correlation, implying that elevated cholesterol levels could be linked to cardiovascular diseases.

3. Low to Moderate Negative Correlations: A negative correlation is observed between 'active' (physical activity) and 'cardio,' though not particularly strong. This suggests that being physically active may be associated with a lower risk of cardiovascular diseases, aligning with earlier visual observations.



	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	age_years	bmi	bp_category_encoded
gender	1	0.5	0.16	0.061	0.066	-0.037	-0.022	0.34	0.17	0.0052	0.0061	-0.024	-0.097	0.074
height	0.5	1	0.29	0.019	0.036	-0.051	-0.019	0.19	0.094	-0.0082	-0.011	-0.082	-0.29	0.053
weight	0.16	0.29	1	0.27	0.25	0.14	0.11	0.067	0.067	-0.017	0.18	0.053	0.76	0.23
ap_hi	0.061	0.019	0.27	1	0.73	0.2	0.093	0.026	0.033	-0.0014	0.43	0.21	0.23	0.73
ap_lo	0.066	0.036	0.25	0.73	1	0.16	0.073	0.024	0.036	-0.0012	0.34	0.16	0.21	0.83
cholesterol	-0.037	-0.051	0.14	0.2	0.16	1	0.45	0.0096	0.034	0.0087	0.22	0.15	0.15	0.14
gluc	-0.022	-0.019	0.11	0.093	0.073	0.45	1	-0.0061	0.0094	-0.008	0.089	0.098	0.1	0.072
smoke	0.34	0.19	0.067	0.026	0.024	0.0096	-0.0061	1	0.34	0.025	-0.017	-0.048	-0.028	0.018
alco	0.17	0.094	0.067	0.033	0.036	0.034	0.0094	0.34	1	0.024	-0.009	-0.029	0.015	0.024
active	0.0052	-0.0082	-0.017	-0.0014	-0.0012	0.0087	-0.008	0.025	0.024	1	-0.038	-0.01	-0.013	-0.0068
cardio	0.0061	-0.011	0.18	0.43	0.34	0.22	0.089	-0.017	-0.009	-0.038	1	0.24	0.16	0.34
age_years	-0.024	-0.082	0.053	0.21	0.16	0.15	0.098	-0.048	-0.029	-0.01	0.24	1	0.086	0.19
bmi	-0.097	-0.29	0.76	0.23	0.21	0.15	0.1	-0.028	0.015	-0.013	0.16	0.086	1	0.18
bp_category_encoded	0.074	0.053	0.23	0.73	0.83	0.14	0.072	0.018	0.024	-0.0068	0.34	0.19	0.18	1

Figure 7: Correlation between variables

These correlation findings provide valuable guidance for further analysis. ’ap_hi’ and ’cholesterol’ are potential significant predictors for the target variable ’cardio,’ given their noticeable correlations. The perfect correlation between ’age’ and ’age_years’ implies redundancy, where one feature would suffice for modeling. While ’smoke’ and ’alco’ are often regarded as risk factors for cardiovascular diseases, their linear correlations with ’cardio’ are minimal in this dataset. However, their potential impacts might be better captured through non-linear relationships or interactions with other features. It’s essential to remember that correlation doesn’t imply causation, and further analyses, like regression or causal inference methods, are necessary to understand the nature and causality of these relationships thoroughly.

3.4 Normalization and Scaling

We choose to that Standard Scaling (Z-score normalization) for the following reasons. First, it centers the data by subtracting the mean from each feature, making it a preferred choice for machine learning algorithms, especially linear models like logistic regression and support vector machines. This centered data aids in achieving better model performance. Second, Standard Scaling imparts unit variance to the features by dividing each feature by its standard deviation, thereby standardizing the variance to 1. This ensures that all features contribute equally to distance-based computations, which is particularly valuable in applications like k-means clustering and k-nearest neighbors. Additionally, it offers robustness against outliers compared to other scaling methods, striking a balance between accommodating extreme values and preserving the overall dataset's scaling integrity. Furthermore, its versatility and general-purpose nature make it a reliable choice when specific data characteristics are not explicitly known, and it allows for meaningful feature coefficient comparisons in linear models, enhancing model interpretability.

Table 3 presents the description of data characteristics after preprocessing, normalization and scaling.

Table 3: Data characteristics description after preprocessing

	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc
count	68205.0	68205.0	68205.0	68205.0	68205.0	68205.0	68205.0
mean	1.3	-0.0	-0.0	-0.0	-0.0	0.0	0.0
std	0.5	1.0	1.0	1.0	1.0	1.0	1.0
min	1.0	-13.4	-4.4	-2.3	-2.3	-0.5	-0.4
25%	1.0	-0.7	-0.6	-0.4	-0.1	-0.5	-0.4
50%	1.0	0.1	-0.1	-0.4	-0.1	-0.5	-0.4
75%	2.0	0.7	0.6	0.8	1.0	-0.5	-0.4
max	2.0	10.5	8.8	3.4	4.2	2.4	3.1

	smoke	alco	active	cardio	age_years	bmi	bp_category_encoded
count	68205.0	68205.0	68205.0	68205.0	68205.0	68205.0	68205.0
mean	0.1	0.1	0.8	0.5	-0.0	0.0	2.9
std	0.3	0.2	0.4	0.5	1.0	1.0	0.9
min	0.0	0.0	0.0	0.0	-3.5	-4.0	1.0
25%	0.0	0.0	1.0	0.0	-0.7	-0.6	3.0
50%	0.0	0.0	1.0	0.0	0.0	-0.2	3.0
75%	0.0	0.0	1.0	1.0	0.8	0.4	3.0
max	1.0	1.0	1.0	1.0	1.7	45.0	4.0

4. Baseline Prediction Model

We use the Logistic model as our baseline model.

Table 4: Caption

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
Model_Overall	0.7250	0.7503	0.6570	0.7006	0.7236	[[5502 1460] [2291 4388]]
Model_Female	0.7366	0.7573	0.6845	0.7191	0.7358	[[3549 960] [1381 2996]]
Model_Male	0.7239	0.7658	0.6563	0.7069	0.7249	[[1860 484] [829 1583]]

Analyzing the logistic regression model's performance reveals several valuable insights:

- Overall Performance: The model exhibits a commendable level of consistency across the entire dataset, women's data, and men's data, with accuracies consistently hovering around 72-74%. This indicates that the model is effectively generalizing and not succumbing to overfitting when applied to specific gender subsets.
- Precision vs. Recall: A noteworthy trend emerges as precision consistently surpasses recall across all datasets. This implies that the model leans toward making cautious predictions about positive cases. In other words, when the model predicts a positive case, it's more likely to be correct. However, it also implies that the model is missing a

considerable number of actual positive cases, hinting at potential room for improvement in detecting such cases.

- Gender Differences: Further exploration indicates that the model performs slightly better for women in terms of accuracy and recall when compared to its performance for men. This observation might suggest that certain features within the dataset hold stronger predictive power for women or that the distribution of specific risk factors differs by gender.

- Potential Improvements: Given the discernible gap between precision and recall, there is a potential avenue for enhancing model performance. This could involve adjusting the decision threshold or employing various techniques such as oversampling, undersampling, or utilizing the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset, ultimately improving the model's ability to correctly identify high-risk individuals without significantly compromising precision.

In summary, the logistic regression model serves as a robust baseline for predicting cardiovascular diseases in this dataset. Its consistent performance across gender subsets showcases its adaptability. However, the model's performance might be further optimized by improving recall, especially if the objective is to ensure that a substantial number of high-risk cases are correctly identified without sacrificing precision.

5. Initial and Revised Plan

5.1 Initial plan

Our initial plan is as follows:

1. Data Collection: Gather a dataset containing patients' records and related information like age, gender, BMI, blood pressure, smoking habits etc.

2. Analyzing Categorical Columns Distribution

3. Data Preprocessing:

    - Clean the data.

  - Necessary transforming of existing features.
  - Encode categorical variables using techniques like ordinal encoding and one-hot encoding.
  - Normalize or standardize numerical features to ensure that they have similar scales.

4. Data Splitting: Split the dataset into training and testing sets.

5. Baseline Model Selection: Choose Logistic model for heart attack prediction.

6. Model Training: Train the selected models using the training dataset.

7. Model Evaluation: Assess the models' performance using evaluation metrics like accuracy, precision, recall, f1 score, ROC_AUC, etc.

8. Model improvement:
   - Adjust the decision threshold.
   - Employ techniques like oversampling, understanding, or using the SMOTE algorithm to balance the dataset.

5.2 Revised plan

Week 9: Baseline model

- Adjusting the decision threshold of baseline model.
- Employing oversampling, undersampling or utilizing the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset.

Week 10: Model selection and design

- Choose machine learning algorithms to use.
- Design the architecture of the model.

Week 11: Model development and training

- Develop and train the different machine learning models.
- Experiment with hyperparameter tuning.

Week 12: Model evaluation

- Evaluate model performance using appropriate metrics.
- Determine if further model adjustments are needed.

Week 13: Results analysis

- Analyze the results and draw insights.
- Compare findings with existing literature.

Week 14: Final adjustments and fine-tuning

- Make any necessary adjustments based on feedback and findings.
- Optimize the model and code for efficiency.

Week 15: Project review and validation

- Conduct a final review to ensure all objectives are met.
- Validate the project's reproductibility.

Week 16:

- Final report and documentation:
- Finalize the project report, incorporating feedback.
- Ensure all documentation is complete and organized.
- Project conclusion and future recommendations:
- Summarize the project's findings and contributions.
- Suggest future research directions and improvements.

References

Kevin S Ikuta, Lucien R Swetschinski, Gisela Robles Aguilar, Fablina Sharara, Tomislav Mestrovic, Authia P Gray, Nicole Davis Weaver, Eve E Wool, Chieh Han, Anna Gershberg Hayoon, et al. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 400(10369): 2221–2248, 2022.

Emelia J Benjamin, Michael J Blaha, Stephanie E Chiuve, Mary Cushman, Sandeep R Das, Rajat Deo, Sarah D De Ferranti, James Floyd, Myriam Fornage, Cathleen Gillespie, et al. Heart disease and stroke statistics—2017 update: a report from the american heart association. *circulation*, 135(10):e146–e603, 2017.

Alan S Go, Dariush Mozaffarian, Véronique L Roger, Emelia J Benjamin, Jarett D Berry, William B Borden, Dawn M Bravata, Shifan Dai, Earl S Ford, Caroline S Fox, et al. Heart disease and stroke statistics—2013 update: a report from the american heart association. *Circulation*, 127(1):e6–e245, 2013.

Joseph P Ornato and Mary M Hand. Warning signs of a heart attack. *Circulation*, 104(11): 1212–1213, 2001.

Richard C Becker. Heart attack and stroke prevention in women. *Circulation*, 112(17):e273–e275, 2005.

Gregory A Roth, Catherine Johnson, Amanuel Abajobir, Foad Abd-Allah, Semaw Ferede Abera, Gebre Abyu, Muktar Ahmed, Baran Aksut, Tahiya Alam, Khurshid Alam, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American college of cardiology*, 70(1):1–25, 2017.

Kristi Rahrig Jenkins and Mary Beth Ofstedal. The association between socioeconomic status and cardiovascular risk factors among middle-aged and older men and women. *Women & health*, 54(1):15–34, 2014.

Samson Y Gebreab, Ana V Diez Roux, Allison B Brenner, DeMarc A Hickson, Mario Sims, Malavika Subramanyam, Michael E Griswold, Sharon B Wyatt, and Sherman A James. The impact of lifecourse socioeconomic position on cardiovascular disease events in african americans: the jackson heart study. *Journal of the American Heart Association*, 4(6):

e001553, 2015.

Frederic Commandeur, Piotr J Slomka, Markus Goeller, Xi Chen, Sebastien Cadet, Aryabod Razipour, Priscilla McElhinney, Heidi Gransar, Stephanie Cantu, Robert JH Miller, et al. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study. *Cardiovascular research*, 116(14):2216–2225, 2020.

Martin P Than, John W Pickering, Yader Sandoval, Anoop SV Shah, Athanasios Tsanas, Fred S Apple, Stefan Blankenberg, Louise Cullen, Christian Mueller, Johannes T Neumann, et al. Machine learning to predict the likelihood of acute myocardial infarction. *Circulation*, 140(11):899–909, 2019.

Birgit Vogel, Monica Acevedo, Yolande Appelman, C Noel Bairey Merz, Alaide Chieffo, Gemma A Figtree, Mayra Guerrero, Vijay Kunadian, Carolyn SP Lam, Angela HEM Maas, et al. The lancet women and cardiovascular disease commission: reducing the global burden by 2030. *The Lancet*, 397(10292):2385–2438, 2021.