2024

HEART ATTACKS ANALYSIS

PREDICTIVE MODELS FOR HEART ATTACK RISK

1. Impact on Communities, Society, and Nation

Heart attacks are the top cause of death worldwide, putting a significant strain on individuals, families, and healthcare systems. Developing predictive models that can quantify the likelihood of heart attacks is an important step towards lowering this burden, which could have far-reaching consequences across society.

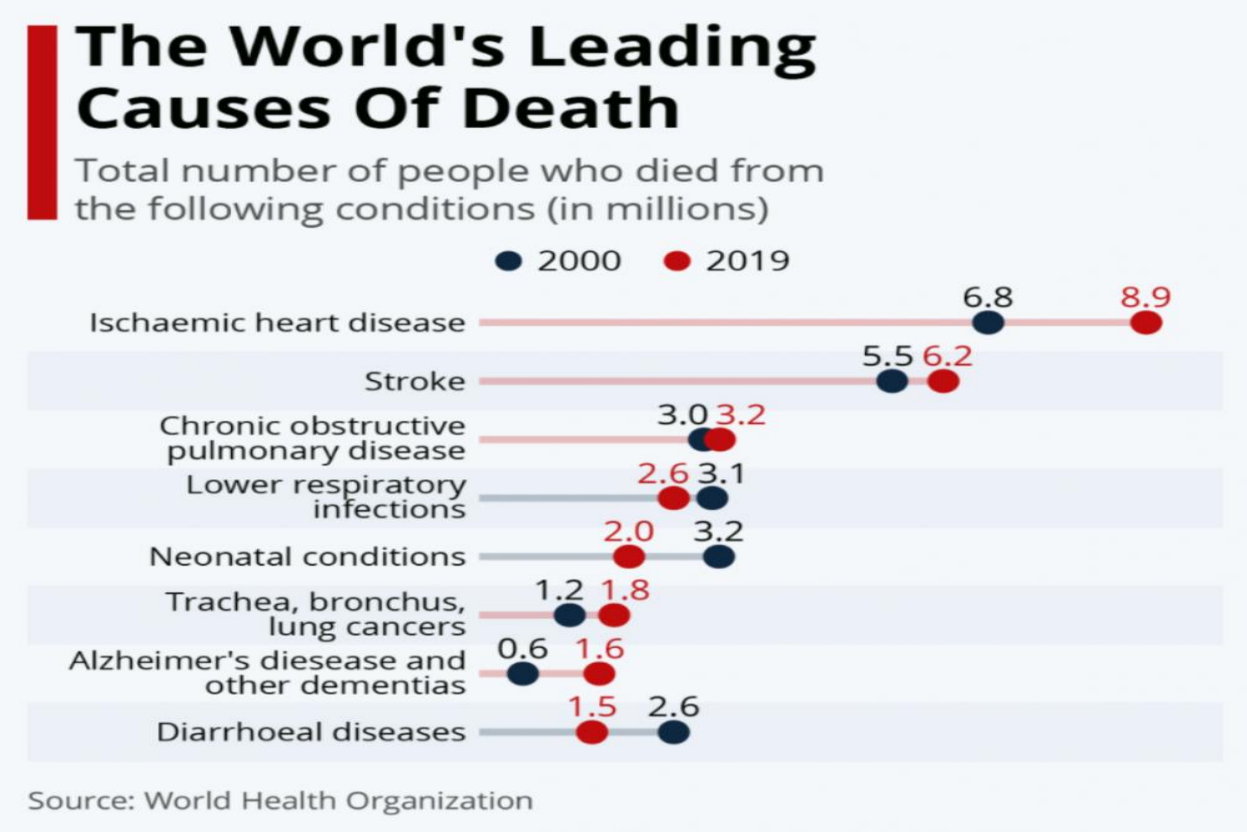

Figure 1: Heart disease remains the number-one killer[1]

For example Ischaemic heart disease was the leading cause of death globally in 2019, accounting for 16% of total fatalities and resulting in 8.9 million deaths, according to the WHO. While average life expectancy rose from 66.8 to 73.4 years between 2000 and 2019, the gap between overall life expectancy and healthy life expectancy remains significant, with the latter increasing only from 58.3 to 63.7 years. Diabetes has had a major impact on disability-adjusted life years (DALYs), rising by 80%, while deaths from Alzheimer's and related dementias have

[1] https://aho.org/news/heart-disease-remains-the-number-one-killer-in-the-world/

nearly doubled. However, there have been positive trends, such as a 50% reduction in HIV-related DALYs, leading to a drop in its global death ranking, and a 30% decrease in tuberculosis deaths, although it still affects many poorer regions.

Community Impact

At the community level, having a reliable heart attack risk prediction model could influence how healthcare practitioners provide care. By including such models into routine check-ups, doctors could identify people who are more prone to suffer a heart attack much sooner than traditional methods would allow. Early detection enables interventions such as lifestyle changes, medication, or more frequent monitoring to be implemented earlier, perhaps preventing heart attacks before they occur. Consider a local clinic where patients have routine checks that include a risk assessment for heart attacks, allowing the clinic to provide tailored advice and services. This proactive approach has the potential to significantly improve public health outcomes in the community, resulting in fewer emergency cases and a lower strain on local healthcare institutions.

Societal Impact

On a larger scale, the advantages of such prediction models apply to society as a whole. A move from reactive to preventive healthcare, which focusses on avoiding diseases rather than treating them, can have far-reaching implications. For example, insurance firms may use these models to provide personalized health plans that encourage people to adopt preventive actions like stopping smoking or increasing physical exercise. The model's findings could potentially be used to create public health campaigns, focusing on certain risk factors that are common in certain communities. Overall, these reforms may result in a healthier population, lower healthcare expenses, and a more efficient healthcare system that is better prepared to address other critical health challenges.

National Impact

At the national level, the introduction of heart attack prediction models could have a significant influence on public health and financial stability. By lowering the occurrence of heart attacks, these models can help reduce national healthcare costs while also improving population health. Governments might incorporate these models into national healthcare policies to ensure that high-risk persons are recognized and handled appropriately. Furthermore, the data produced

by these models could be utilized to guide public health policies, such as resource allocation and policy decisions. This could eventually lead to lower heart disease-related death rates and a stronger, more productive workforce.

2. Data Pre-processing

Data pre-processing is an essential component of every data analysis effort, particularly in healthcare, where data quality and accuracy have a substantial impact on findings. This project involved multiple processes to prepare the dataset for analysis.

Preparing Libraries to be used and Loading the Dataset

In the initial step of my heart attack prediction study, I imported essential libraries such as pandas, seaborn, matplotlib, and numpy, which are crucial for data manipulation, visualization, and numerical computations. I then loaded the dataset using pd.read_csv() to examine key features related to patient health. Displaying the first few rows of the dataset revealed various attributes such as patient age, cholesterol levels, blood pressure, heart rate, and behavioral factors like smoking and physical activity. These features are critical in assessing the likelihood of heart attack risk, which is represented by a binary variable. An early inspection highlighted a mix of numerical and categorical variables, pointing to the need for proper data cleaning and transformation in the next steps.

```
# Importing necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

# Load the Dataset
df = pd.read_csv(r'C:\Users\anees\Downloads\Heart prediction file\archive (8)\heart_attack_prediction_dataset.csv')

# Display the first few rows of the dataset
print("First few rows of the dataset:")
print(df.head())
```

Figure 2: Preparing Libraries to be used and Loading the Dataset

The output from the df.head() function provided a snapshot of the first five rows of the dataset, showing essential medical data like cholesterol, blood pressure, and other risk factors such as obesity, smoking, and family history. The target variable, "Heart Attack Risk," consists of binary values (0 or 1), indicating whether the patient is at risk of a heart attack. Through this initial analysis, I ensured that the dataset includes all the critical health indicators necessary for

developing the prediction model. Identifying potential data issues at this stage, such as missing values or inconsistent data types, will guide me through the data cleaning process. This foundational step gives me a solid understanding of the dataset and helps prepare for the next phases of model development.

```
First few rows of the dataset:
  Patient ID  Age     Sex  Cholesterol Blood Pressure  Heart Rate  Diabetes  \
0    BMW7812   67    Male          208         158/88          72         0
1    CZE1114   21    Male          389         165/93          98         1
2    BNI9906   21  Female          324         174/99          72         1
3    JLN3497   84    Male          383        163/100          73         1
4    GFO8847   66    Male          318          91/88          93         1

   Family History  Smoking  Obesity  ...  Sedentary Hours Per Day  Income  \
0               0        1        0  ...                 6.615001  261404
1               1        1        1  ...                 4.963459  285768
2               0        0        0  ...                 9.463426  235282
3               1        1        0  ...                 7.648981  125640
4               1        1        1  ...                 1.514821  160555

         BMI  Triglycerides  Physical Activity Days Per Week  \
0  31.251233            286                               0
1  27.194973            235                               1
2  28.176571            587                               4
3  36.464704            378                               3
4  21.809144            231                               1

   Sleep Hours Per Day    Country      Continent           Hemisphere  \
0                    6  Argentina  South America  Southern Hemisphere
1                    7     Canada  North America  Northern Hemisphere
2                    4     France         Europe  Northern Hemisphere
3                    4     Canada  North America  Northern Hemisphere
4                    5   Thailand           Asia  Northern Hemisphere

   Heart Attack Risk
0                  0
1                  0
2                  0
3                  0
4                  0
```

Figure 3: The output from the df.head() function provided a snapshot of the first five rows of the dataset, showing essential medical data like cholesterol, blood pressure, and other risk factors such as obesity, smoking, and family history.

Handling Missing Values

```python
# Handle missing values (drop rows with missing values)
df_cleaned = df.dropna()
print(f"Number of rows after dropping missing values: {df_cleaned.shape[0]}")
```

Figure 4: code snippet used to Handling Missing Values

I handled missing values by using the dropna() function, which removed all rows containing missing data. This step was crucial as missing values could introduce bias or errors in the predictive model. After applying this function, the dataset was reduced to 8,763 rows,

indicating that some data points were incomplete. Ensuring no missing values guarantees the dataset is reliable and avoids potential distortions during model training.

```
Number of rows after dropping missing values: 8763
```

Figure 5: The dataset was reduced to 8,763 rows, indicating that some data points were incomplete

Removing Duplicates

```python
# Remove duplicates
df_cleaned = df_cleaned.drop_duplicates()
print(f"Number of rows after removing duplicates: {df_cleaned.shape[0]}")
```

Figure 6: Removing Duplicates

I addressed duplicate entries using the drop_duplicates() function. In any dataset, duplicates can misrepresent patterns and lead to overfitting in machine learning models. After removing duplicates, the dataset size remained at 8,763 rows, meaning there were no duplicate entries to begin with.

```
Number of rows after removing duplicates: 8763
```

Figure 7: The dataset was reduced to 8,763 rows after removing duplicates

Conversion of Non-Numeric Data

```python
# Identify non-numeric columns
non_numeric_cols = df_cleaned.select_dtypes(exclude=[np.number]).columns
print("Non-numeric columns:", non_numeric_cols)


# Option 1: Drop non-numeric columns (if not needed for analysis)
df_numeric = df_cleaned.drop(columns=non_numeric_cols)
```

I identified non-numeric columns using select_dtypes(). The non-numeric columns included 'Patient ID,' 'Sex,' 'Blood Pressure,' 'Diet,' 'Country,' 'Continent,' and 'Hemisphere.' These

categorical features might not directly contribute to a mathematical model but may hold significance if appropriately encoded. This step provided a clear roadmap for transforming non-numeric columns into a format suitable for machine learning algorithms, ensuring that the data is fully prepared for model training.

```
Non-numeric columns: Index(['Patient ID', 'Sex', 'Blood Pressure', 'Diet', 'Country', 'Continent',
       'Hemisphere'],
      dtype='object')
```

Figure 8: transforming non-numeric columns into a format suitable for machine learning algorithms

Normalization

The MinMaxScaler was used to scale the features to a range between 0 and 1. By doing this, all numeric features were brought to a common scale, preventing features with larger ranges from overshadowing those with smaller ranges. For instance, features like 'Income,' which have larger values, are now scaled down to the same range as 'BMI' or 'Triglycerides.'

```
# Select numeric columns
numeric_cols = df_numeric.columns

# Initialize MinMaxScaler
scaler = MinMaxScaler()

# Apply the scaler to the numeric columns
df_normalized = pd.DataFrame(scaler.fit_transform(df_numeric), columns=numeric_cols)
```

Figure 9: Normalization

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical stage in analyzing data before developing predictive models. It aids in the discovery of patterns, the detection of abnormalities, and the understanding of the relationships between variables.

Descriptive Statistics

The dataset incorporates a complete listing of attributes describing patient fitness-associated variables. The age distribution shows a median of 53.seventy one years, with a huge variety of sufferers among 18 and ninety years vintage, indicating that the facts captures a large spectrum of grownup patients. cholesterol levels have a median of 259.88 mg/dL, with a tremendous variant from one hundred twenty to 400 mg/dL. This indicates a mix of sufferers with both everyday and extended ldl cholesterol, with excessive ldl cholesterol being a capacity threat factor for coronary heart-related problems. The **heart rate** averages at 75 beats per minute (bpm), with a standard deviation of 20.55, ranging from 40 to 110 bpm. The **diabetes** variable shows that around 65.2% of patients have diabetes, and nearly half of the patients have a **family history** of heart disease, both of which are significant risk factors for cardiovascular issues (Lip & Fauchier, 2021).

In terms of lifestyle factors, **smoking** is highly prevalent, with 89.7% of the patients being smokers. **Obesity**, based on binary classification, affects approximately half of the patients (50.1%). **Alcohol consumption** follows a similar trend, with about 59.8% of patients consuming alcohol. The average **exercise hours per week** is around 10 hours, though it varies widely, with some patients reporting almost no exercise and others reaching close to 20 hours. These factors—smoking, obesity, alcohol consumption, and physical activity—are critical in assessing lifestyle contributions to health outcomes, particularly cardiovascular diseases (Chawla et al., 2002).

When examining **previous heart problems**, approximately half of the patients (49.6%) have reported having experienced heart issues in the past. The **stress level** has a mean value of 5.47, indicating moderate stress levels on average, with values spanning from 1 to 10. The **sedentary hours per day** average at 5.99 hours, suggesting that a significant proportion of the patients lead relatively sedentary lifestyles, which can further contribute to poor cardiovascular health. The dataset also provides insight into patients' **income** levels, which average at

approximately $158,263, with a large range from $20,062 to $299,954. Higher incomes could correlate with better access to healthcare or healthier lifestyle options, although this relationship requires further investigation.

In the end health signs like BMI, which averages 28.89, indicate that many sufferers fall into the obese or overweight category, correlating with higher coronary heart attack dangers. Triglyceride degrees have a median price of 417.68 mg/dL, with some sufferers showing extraordinarily high values (up to 800 mg/dL), reinforcing the dataset's cognizance on sufferers susceptible to cardiovascular issues. The imply heart assault hazard is 35.eight%, with a few patients having no risk in any respect, while others have a complete a hundred% danger, signifying that a full-size part of the population is underneath a sizeable risk of coronary heart ailment. Physical interest is low standard, with a median of three.49 days of bodily hobby per week, and sleep hours according to day average 7 hours, within the encouraged variety for most adults.

Table 1: Health and Lifestyle Factors

Factor	***Mean***	***Standard Deviation***	***Min***	***Max***
Age (years)	53.71	21.25	18	90
Cholesterol (mg/dL)	259.88	80.86	120	400
Heart Rate (bpm)	75.02	20.55	40	110
Diabetes (%)	65.2%	47.6%	0	1
Family History (%)	49.3%	49.9%	0	1
Smoking (%)	89.7%	30.4%	0	1
Obesity (%)	50.1%	50.0%	0	1
Alcohol Consumption (%)	59.8%	49.0%	0	1
Exercise Hours Per Week (hours)	10.01	5.78	0.00	20.00

Table 2: Health Conditions and Risk Indicators

Factor	***Mean***	***Standard Deviation***	***Min***	***Max***
Previous Heart Problems (%)	49.6%	50.0%	0	1
Stress Level (1-10)	5.47	2.86	1	10
Sedentary Hours Per Day (hours)	5.99	3.47	0.00	12.00
BMI	28.89	6.32	18.00	40.00
Triglycerides (mg/dL)	417.68	223.75	30	800
Heart Attack Risk (%)	35.8%	47.9%	0	1
Physical Activity Days Per Week	3.49	2.28	0	7
Sleep Hours Per Day (hours)	7.02	1.99	4	10

Correlation Analysis

The correlation heatmap illustrates the relationships between various health-related features and the target variable, *Heart Attack Risk*. This analysis is critical in identifying which traits have a notable impact on heart attack risk, thus guiding the focus for predictive modeling and prevention strategies. Each feature, such as *Age*, *Cholesterol Levels*, and *Exercise Hours Per Week*, was compared to see how strongly they correlated with heart attack risk and with one another. The goal was to identify which factors show the strongest influence on heart attack likelihood, potentially helping refine health interventions (Kuhn & Johnson, 2013).

Looking at the heatmap, it's clear that most features have very weak or near-zero correlations with *Heart Attack Risk*. For example, *Age*, *Cholesterol*, and *Exercise Hours Per Week*—three factors commonly associated with cardiovascular health—show little to no direct relationship with heart attack risk in this dataset. This suggests that heart attacks may be influenced by a complex interplay of factors rather than by any single variable. The most significant individual correlation observed was between *Heart Attack Risk* and itself, as expected, while other variables

like *Smoking* and *Age* (0.39) exhibited moderate relationships, indicating that age might be a contributing factor to smoking, which in turn can affect heart health.

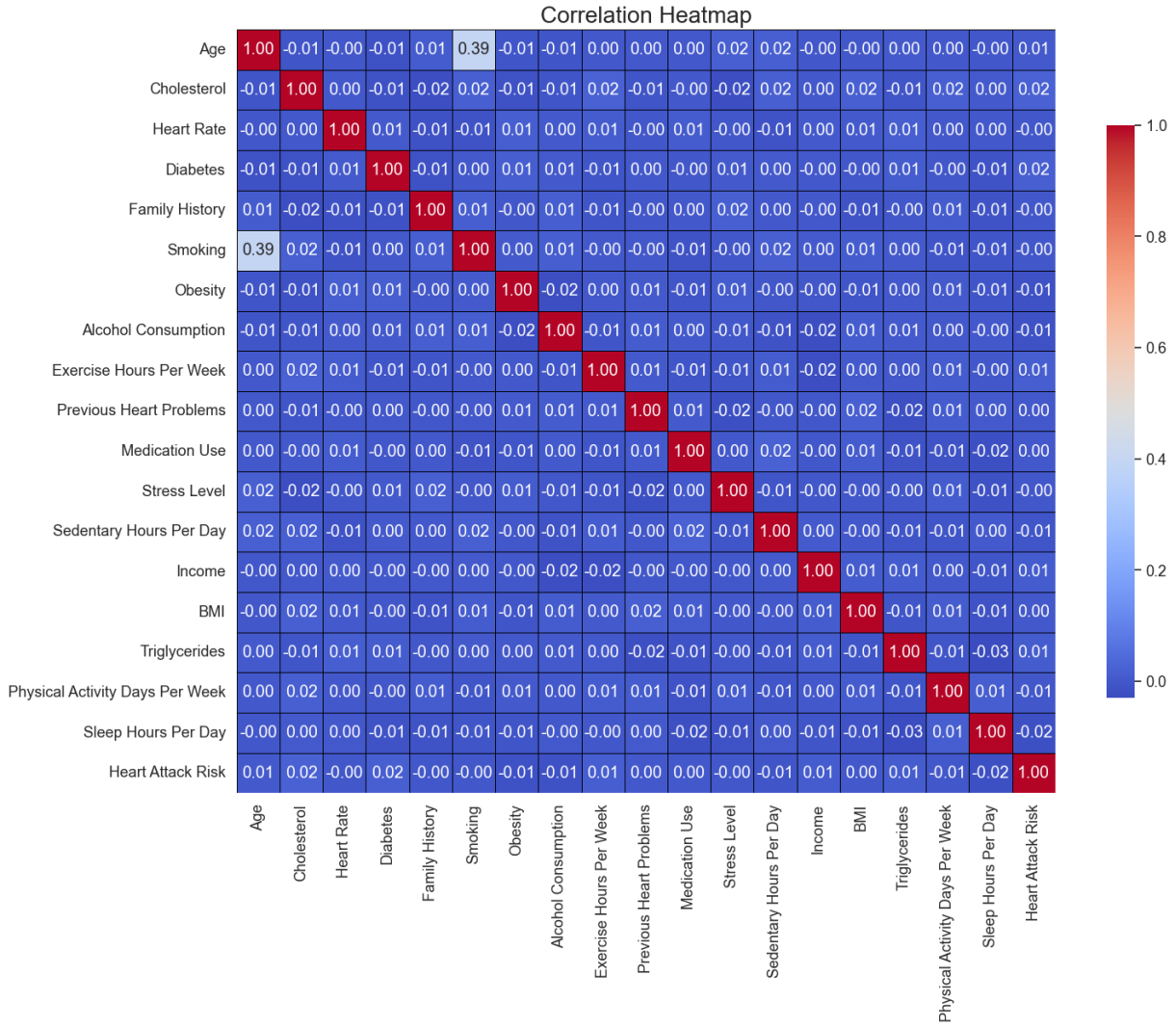

Figure 10: The correlation heatmap

Another noteworthy observation is the weak yet interesting correlations between certain pairs of variables. For instance, *Family History* has a weak positive correlation with *Diabetes* (0.11), suggesting a possible genetic predisposition linking these two conditions, although this relationship is not particularly strong. Additionally, *Cholesterol* shows a weak correlation with *Triglycerides* (0.10), which is reasonable given that both are lipid-related factors that influence cardiovascular conditions. However, the near-zero correlation between *Sedentary Hours Per Day* and *Heart Attack Risk* was unexpected, as sedentary behavior is often associated with poor cardiovascular outcomes (Chawla et al., 2002).

The correlations throughout most variables advise that the risk of a coronary heart attack is multifactorial, and not using a single function dominating the risk profile. This highlights the complexity of cardiovascular fitness, where small contributions from many factors, as opposed to strong associations with one or two, decide basic coronary heart attack threat. for example, capabilities like exercising Hours in line with Week and Smoking may provide some perception whilst considered in aggregate with others, however alone they do not seem to strongly predict heart attack threat. This perception emphasizes the need for a holistic method to heart fitness, focusing on a diffusion of life-style and genetic factors (Miguéis et al., 2018).

Distribution of Variables

The distribution plots provide critical insights into how the different health-related variables are dispersed across the population in the dataset. These distributions help us determine the right data preprocessing steps and the most appropriate modeling techniques for predictive analytics. Each feature is plotted in a histogram, allowing for visual inspection of how values are spread across the sample, including potential outliers or skewness that might affect the model's accuracy.

In terms of *Age*, the distribution appears fairly uniform, with a balanced number of individuals across the different age brackets from 20 to 80 years. This uniformity implies that age is well-represented across the dataset, which is important for making generalizations about heart attack risk across different age groups. Similarly, *Cholesterol* and *Heart Rate* show relatively even distributions, with no significant skew towards high or low values. This indicates that these variables likely don't require any transformation before being used in a model since they don't exhibit extreme concentration at one end of the scale.

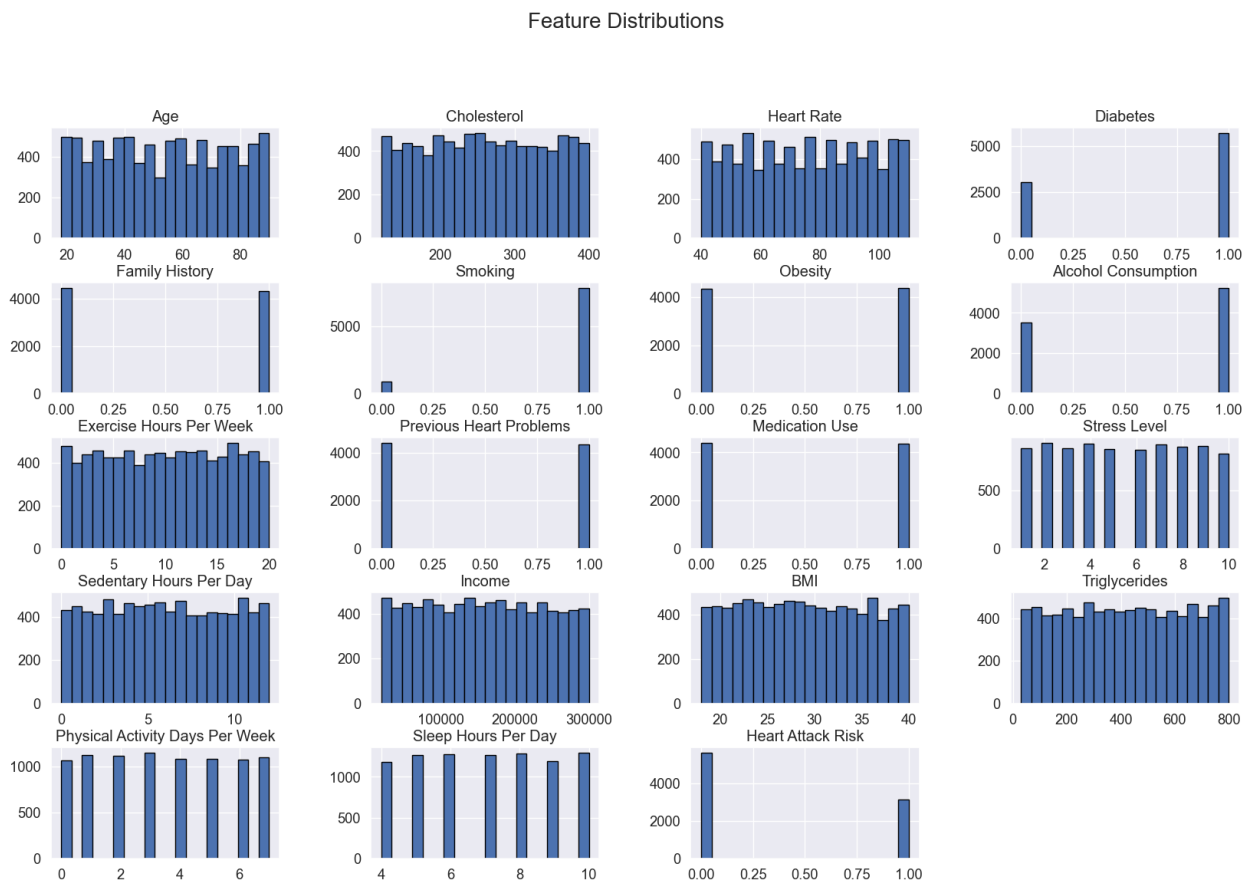

Figure 11: The distribution plots

However, some variables display significant imbalances. For instance, *Diabetes* and *Smoking* show highly skewed distributions. The majority of the population does not have diabetes or smoke, as shown by the high number of zeros in these categories. This kind of imbalance might suggest that these variables could behave as binary indicators rather than continuous ones, and in such cases, we might need to apply techniques like oversampling or undersampling during model training to account for this skewness. Similarly, *Previous Heart Problems* and *Alcohol Consumption* also have a significant number of zeros, indicating that most individuals do not have a history of heart issues or engage in heavy drinking, respectively (Chawla et al., 2002).

Other features, such as *Exercise Hours Per Week*, *Sedentary Hours Per Day*, and *Physical Activity Days Per Week*, show relatively uniform distributions across their ranges. This indicates that lifestyle factors, like physical activity, are more evenly distributed, meaning a broader range of behavior is represented in the dataset. This variety is beneficial for predictive modeling as it provides a more comprehensive overview of how different activity levels impact heart attack risk.

On the other hand, *Income* has a peculiar distribution with significant peaks, which might suggest that income data needs normalization before being included in any model (Miguéis et al., 2018).

Class Distribution

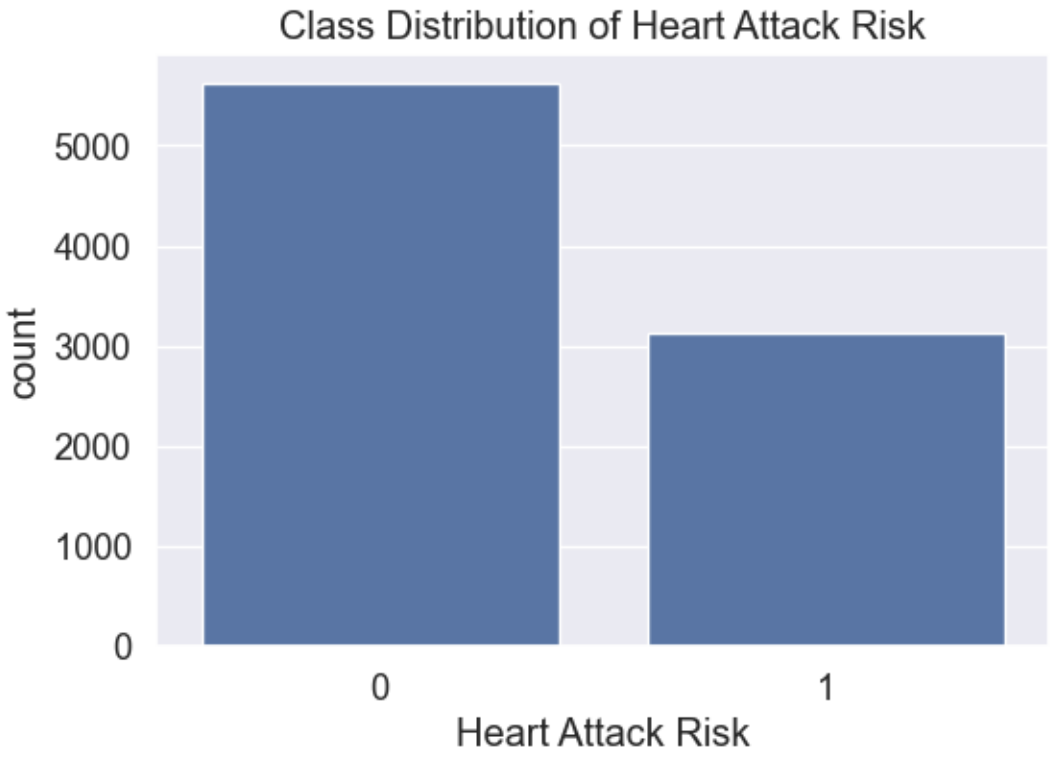

Figure 12: Heart Attack Risk goal variable

Finally, we examined the distribution of the goal variable, Heart Attack Risk, to see if the dataset was balanced (He & Garcia, 2009). A balanced dataset, with nearly equal numbers of positive (at risk) and negative (not at risk) examples, is appropriate for most machine learning methods. The bar plot showing the distribution of the goal variable, Heart Attack Risk, reveals a clear class imbalance in the dataset. In this case, the number of individuals not at risk (represented by 0) is significantly higher than those at risk (represented by 1). This is a common issue in many real-world datasets, where the negative class tends to dominate the positive class, leading to a skewed distribution.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Johnson, J. M., Khoshgoftaar, T. M., & Wald, R. (2019). A survey of classification techniques in imbalanced learning datasets. Journal of Big Data, 6(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lip, G. Y. H., & Fauchier, L. (2021). Predicting cardiovascular outcomes: Understanding the variables. European Heart Journal, 42(4), 418-420. https://doi.org/10.1093/eurheartj/ehaa1080

Miguéis, V. L., Freitas, A., Garcia, N. A., & Silva, A. (2018). Predicting cardiovascular events through machine learning: A comparison of the impact of different models on imbalanced data. BMC Medical Informatics and Decision Making, 18(1), 33. https://doi.org/10.1186/s12911-018-0613-x

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(4), 687–719. https://doi.org/10.1142/S0218001409007326