Predicting Cardiovascular Risk: Analyzing Lifestyle, Socioeconomic, and Medical Factors

(COMP3125 Individual Project)

Yusuf Ali
Wentworth Institute of Technology

***Abstract*—This project investigates the key factors contributing to cardiovascular issues using data from the BRFSS dataset and Cardiovascular Disease Dataset. A Random Forest classifier was applied to identify the most significant predictors of heart attacks. Results showed that smoking, alcohol consumption, lack of physical activity, and low socioeconomic status are correlated with increased risk of cardiovascular issues. The Random Forest model achieved a accuracy of 94.81% and provided insights into what were the key predictors including angina, BMI, and diagnostic scan history. These findings offer insights that can inform public health strategies through targeted prevention and intervention.**

Keywords—cardiovascular disease, Random Forest classifier, heart attack prediction, BRFSS dataset, risk factors

I. Introduction (*Heading 1*)

According to the World Health Organization (WHO) heart disease is the leading cause of death worldwide. Contributing to 16% of all deaths globally [1]. This widespread prevalence underscores a critical public health crisis that affects billions of lives and demands urgent attention. For many, including myself, the issue is also deeply personal, as heart disease runs in my family.

Over the past few years, technology has rapidly advanced, opening new possibilities in the field of healthcare. Data science and machine learning has transformed the way we analyze, interpret and understand medical information. These tools offer powerful methods for identifying patterns, uncovering hidden relationships, and predicting outcomes based on a collection of patient data.

Using these tools, we can explore the key factors that contribute to heart disease including biological and medical factors such as age, cholesterol, and blood pressure, as well as behavioral and lifestyle factors like smoking, physical activity, and diet. Additionally, socioeconomic factors such as income level, education, and access to healthcare could be critical predictors of heart problems. By analyzing these diverse variables through machine learning models we can move beyond surface-level associations and gain a deeper, more holistic understanding of what puts individuals at risk for heart disease.

A lot of research has been done to try to understand what factors cause heart disease. Traditional risk factors include high blood pressure, elevated cholesterol, obesity, diabetes, smoking, sedentary behavior, and poor diet [2]. More recently, studies have emphasized the impact of stress and socioeconomic factors such as income level, education, and access to healthcare [3] .

Advances in machine learning have enabled researchers to build predictive models that can identify individuals at risk of heart disease with greater accuracy than conventional statistical approaches. For instance, one study published in the National Library of Medicine by Ahmad Ayid Ahmad and Huseyin Polat from Gazi University in Turkey created a machine learning model to proactively predict heart disease using jellyfish optimization and SVM classifier [4]. This model had sensitivity, specificity and accuracy all over 98% one of the highest performance in heart disease prediction.

Through this project, I aim to bridge data science and healthcare, to build a predictive model and to gain and understanding on the broader factors contributing to heart disease. In doing so, the project will support data-driven, proactive, and equitable approaches in health diagnostics and prevention.

A. Source of dataset (Heading 2)

For this report 2 different datasets were utilized.

The first dataset used in this project originates from the Centers for Disease Control and Prevention (CDC) and is part of the Behavioral Risk Factor Surveillance System (BRFSS) [5]. This dataset was collected through an yearly telephone survey conducted by the CDC. This survey attempts to assess the health, behaviors, and risk factors of adults living in the United States. The survey covers questions related to demographics, current health status, and specific health conditions including heart disease.

The BRFSS, was established in 1984 in only 15 states, since then BRFSS has grown significantly over the decades. It now collects health-related data from all across the United States, conducting over 400,000 interviews each year. This makes it the largest continuously conducted health survey system in the world. The dataset used here is based on the 2022 BRFSS annual survey. The dataset has survey results from 246,026 adults in the United States.

To allow for analysis, two versions of the dataset were provided by the CDC one containing missing values and one where these were cleaned or removed. Further data cleaning and preprocessing were conducted as part of this project to ensure the data was suitable for analytical and predictive modeling purposes.

The CDC is a United States federal agency and part of the Department of Health and Human Services. The CDC is a highly credible source of public health information. In

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

specific, the BRFSS survey data is widely used in academic, public policy, and healthcare research settings.

The 2nd dataset used in the report is the Cardiovascular Disease Dataset. This dataset was made publicly available on April 16, 2021. It was published on Mendeley Data a cloud based database repository. This dataset was contributed by Bhanu Prakash Doppala and Debnath Bhattacharyya, and originates from a multispecialty hospital in India [6].

The dataset includes data for 1,000 individuals and features 12 medical attributes that are commonly associated with heart disease. These include factors such as age, sex, blood pressure, cholesterol levels, fasting blood sugar, electrocardiographic results, and more. The goal of this dataset is to help support early-stage detection of heart disease.

Compared to large-scale population surveys like the BRFSS, this dataset provides a more clinical, single area focus. As it was collected at a single hospital in India.

While the Cardiovascular Disease Dataset does not come from a government institution like the CDC, it is still credible. It was published through Mendeley Data, a well-known academic database repository that peer reviews datasets. Additionally, the dataset was contributed by researchers who are affiliated with well recognized institutions.

Both datasets used in this report are credible and serve different purposes. The CDC's BRFSS dataset offers a broad, insights, while the Cardiovascular Disease Dataset provides detailed specific clinical data. Together, these databases provide a well-rounded perspective for studying heart disease.

B. Character of the datasets

The two datasets used in this report, the BRFSS 2022 dataset and the Cardiovascular Disease Dataset, are structured differently and served different purposes for this report.

Dataset name	**Format**	**Size (Rows x Columns)**	**Source**	**Type**
BRFSS 2022 Dataset	CSV	246,023 × 40	CDC (BRFSS Survey)	Population Survey
Cardiovascular Disease Dataset	CSV	1,000 × 14	Mendeley Data	Clinical Dataset

The BRFSS dataset includes a mix of demographic, behavioral, and health-related attributes, collected via self-reported surveys. These variables are important to gain insight risk factors across the population.

Column Name	**Description**	**Type**	**Units/Value**
HadHeartAttack	Had a Heart attack	Binary	Yes/No
BMI	Body Mass Index	Numeric	kg/m²
Smoking	Whether the individual smokes	Binary	Former, Daily, Occasional, Non Smoker
AlcoholDrinking	Drinks Alcohol	Binary	Yes/No
PIR	Income level of the individual	Categorical	High, Middle, Low
PhysicalHealth	Number of unhealthy physical health days (past month)	Numeric	Days (0-30)
MentalHealth	Number of unhealthy mental health days (past month)	Numeric	Days (0-30)
Diabetic	Has diabetes	Categorical	No, Yes, Borderline
Sex	Gender	Categorical	Male/Female
AgeCategory	Age range	Categorical	(18-24, 25-34, etc.)
Race	Self-reported race/ethnicity	Categorical	White, Black, Asian, etc.
PhysicalActivity	Engages in physical activity	Binary	Yes/No
GenHealth	General health status	Categorical	Excellent, Good, Fair, Poor
SleepTime	Average hours of sleep per night	Numeric	Hours
RemovedTeeth	# of teeth removed	Categorical	None, 1-5,6+, All
ECigaretteUsage	Usage of E-Cigarette	Categorical	Never, Former, Daily, Occasioanly

The Cardiovascular Disease Dataset is clinical in nature, containing medical attributes collected in a hospital setting. These variables are critical for early-stage heart disease detection.

Column Name	Description	Type	Units/Values
Age	Age of the patient	Numeric	Years
Sex	Gender of the patient	Binary	1 = Male, 0 = Female
Height	Patient's height	Numeric	cm
Weight	Patient's weight	Numeric	kg
BMI	Body Mass Index (calculated)	Numeric	kg/m²

Cholesterol	Cholesterol level category	Categorical	Normal, Above Normal, High
Glucose	Glucose level category	Categorical	Normal, Above Normal, High
Smoking	Whether the patient smokes	Binary	Yes/No
Alcohol Intake	Whether the patient consumes alcohol	Binary	Yes/No
Physical Activity	Whether the patient is physically active	Binary	Yes/No
Cardio	Whether the patient has cardiovascular disease	Binary	Yes/No

II. Methodology

To analyze and understand the factors three different methodologies were used. Random Forest Classification, data visualization and statistical testing. Each method offered new insights and allowed for a holistic understanding of the dataset.

A. Random Forest Classifer

The random forest classifier works by constructing a group of decision trees. Each tree is trained on a different subset of the data. Their outputs are then combined to produce accurate and stable predictions. One of the key strengths of Random Forest is its ability to manage large, complex datasets like in this case with the BRFSS dataset it was trained on. The assumption of the random forest classifier is that training sample data is representative of the general population.

The classifier was implemented using the RandomForestClassifier from Python's scikit-learn library. Some hyperparameter tuning such as adjusting the number of estimators and maximum tree depth was done in an attempt to optimize performance.

B. Data Visualization

Visualization of the data was done through Matplotlib and Seaborn, two of the most popular Python libraries for data visualization. These tools were used to create bar graphs, histograms, and box plots to clearly illustrate trends and differences between groups. Visualizations were used to explore variables such as age, smoking habits, blood pressure, heart rate etc. Graphs were carefully labeled and styled for clarity, with legends, axis titles, and value annotations added to ensure the insights were accessible to a broader audience.

C. Statisticial Methods

In cases where visualizations did not provide a clear result or correlation statistical methods were used. The point-biserial correlation coefficient was used to measure the relationship between a binary variable and a continuous one. This method was suitable for evaluating measurements such as serum cholesterol, resting heart rate or average hours slept a night related to the binary presence or absence of cardiovascular issues. This method has a couple of assumptions that the continuous variable is normally distributed within each binary group and the point-biserial correlation coefficient is limited to identifying linear relationships. This method was implemented using the pointbiserialr function from the SciPy. Stats module. Variables were carefully selected and checked before being used.

III. Results

The results provide valuable insights into which factors biological, behavioral, or socioeconomic play a dominant role in predicting cardiovascular disease risk.

A. Lifestyle Results

Analysis of lifestyle variables revealed several strong correlations with heart disease risk. These findings suggest that individuals can reduce their chances of developing cardiovascular conditions by making conscious changes in their daily habits.

The first lifestyle habit that was examined was smoking.

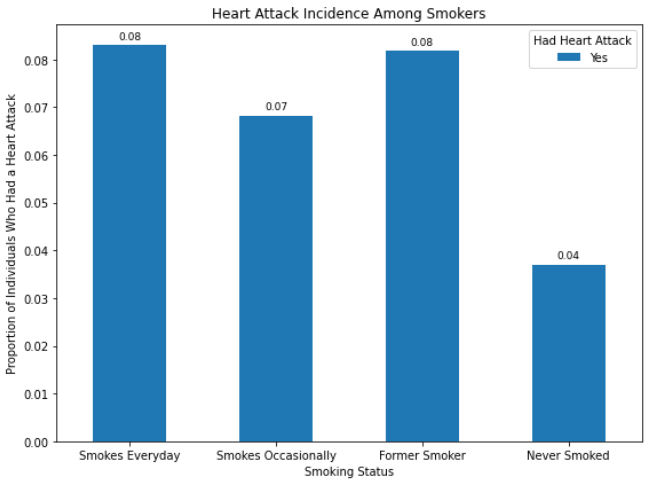

Fig. 1. Proportion of heart attacks among smokers vs. non-smokers.

The data show that individuals who smoke or have a history of smoking have a noticeably higher incidence of heart attacks. As illustrated in the first bar chart, the proportion of heart attacks among smokers is significantly higher compared to non-smokers. This correlation emphasizes the well-established link between tobacco use and cardiovascular disease.

Similarly, alcohol consumption also demonstrates a notable correlation with heart attack incidence.

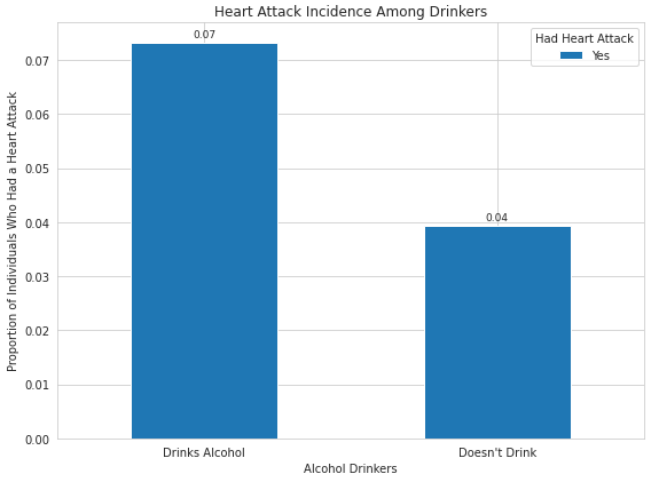

Fig. 2. Comparison of heart attack rates based on alcohol consumption.

As shown in figure 2, individuals who consume alcohol have a higher proportion of heart attacks (around 7%) compared to those who do not drink (approximately 4%). This correlation suggests that alcohol use, especially when excessive or frequent can be a contributing factor to cardiovascular issues. Therefore, limiting or abstaining from alcohol may be a proactive step toward better heart health.

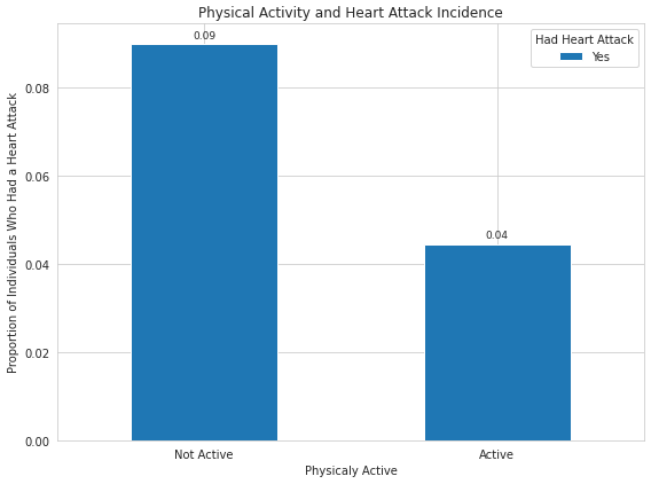

Fig. 3. Comparison between physical activity and heart attack incidence.

As seen in figure 3 being physically active makes a great difference in cardiovascular issues. The data indicates that those individuals who engage in regular physical activity experience a noticeably lower rate of heart attacks compared to those with inactive and sedentary lifestyles. These results reinforces the widely accepted understanding that exercise helps strengthen the cardiovascular system thus reducing the risk of cardiovascular issues. These findings highlight the importance of incorporating consistent physical activity as a preventative measure against heart-related conditions.

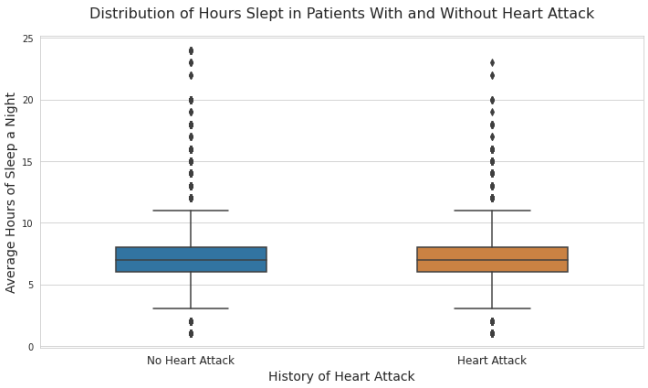

Fig. 4. Box plot comparing average hours of sleep and heart attack occurrence.

The next lifestyle factor that was analyzed was the average hours slept per night. As illustrated in Figure 4, the box plots appear to be similar. The median sleep duration for both groups was approximately seven hours. The interquartile range (IQR), representing the middle 50% of the data, was also roughly consistent (around 6 to 8 hours) for both groups. To double check that there was no correlation between average hours slept a night and heart attacks the Point-biserial correlation coefficient was calculated. The Point-biserial correlation coefficient looked to see if there was a linear correlation between the two variables. The Point-biserial correlation coefficient is 0.0036 and a P-value of 0.0717. The Point-biserial correlation coefficient is very close to 0, telling us that there is virtually no relationship between sleep hours and whether someone had a heart attack, and the P-value tells us that the result is not statistically significant. These supports the visual evidence from the box plots and suggest that average hours of sleep is not correlated with the likelihood of experiencing a heart attack.

B. Socioecnomic Factors

This section explores how socioeconomic factors , such as income, education, and race impact the likelihood of heart attacks. This will help highlight the influence of social inequality on cardiovascular health.

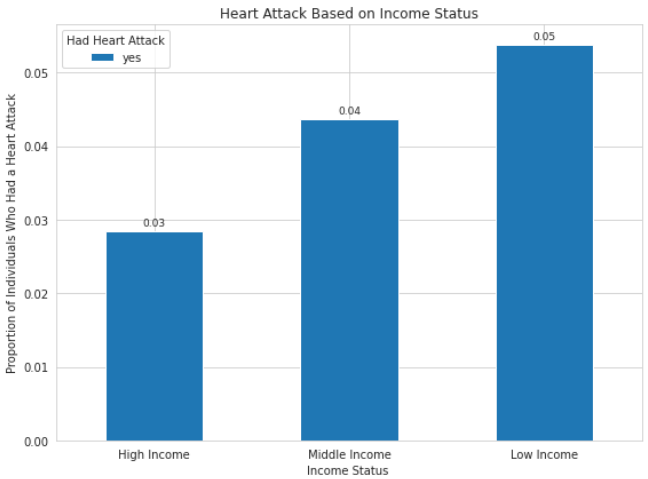

Fig. 5. Heart attack risk across different income levels.

As we can see in figure 5 there appears to be a correlation between income status and the risk of a heart attack. Those with lower incomes have a higher risk of suffering from an

heart attack. This suggests that individuals with lower income levels may face greater vulnerability to heart disease, potentially due to reduced access to healthcare, nutritious food, and health education.

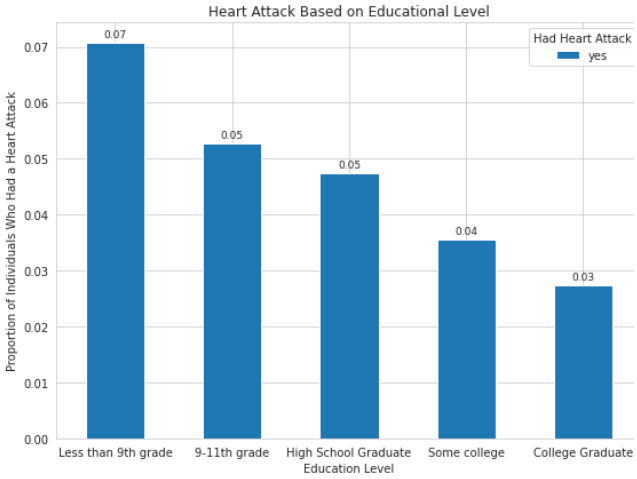

Fig. 6. Correlation between education level and heart attack risk

Another socioeconomic factor that was analyzed was the level of education. Looking at figure 6, the results show a clear trend. Individuals with lower levels of education reported higher rates of heart attacks.

The highest rate (7%) is seen in the group with less than a 9th-grade education, while the lowest rate (3%) is found among college graduates. This suggests a potential link between education level and cardiovascular health. This could be due to the fact that those with better educational status typically are higher income earners.

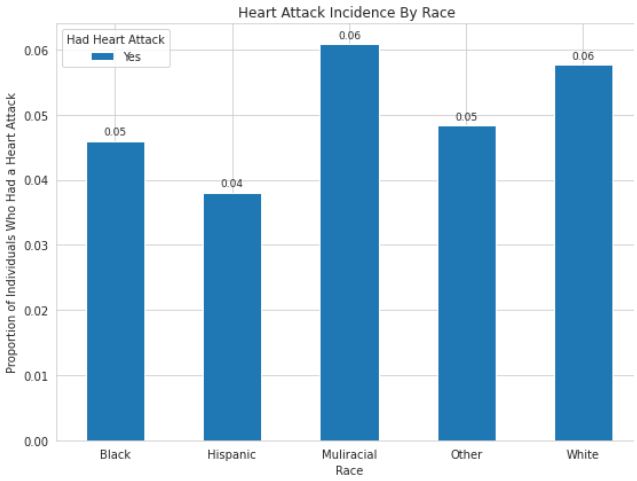

Fig. 7. Heart attack distribution across different racial groups.

Race was another key variable analyzed to check for heart attack incidence. Figure 7 shows that the highest proportions were observed among individuals identifying as Multiracial and White (6%). In contrast, the lowest proportion was reported by individuals identifying as Hispanic (4%). These findings show disparities across the racial groups.

C. Medical Factors

To explore the correlation between medical factors, visualizations were created to highlight most common risk factors for cardiovascular issues. These factors inclued, age, resting blood presure, maxium heart rate as well as serum cholesterol.

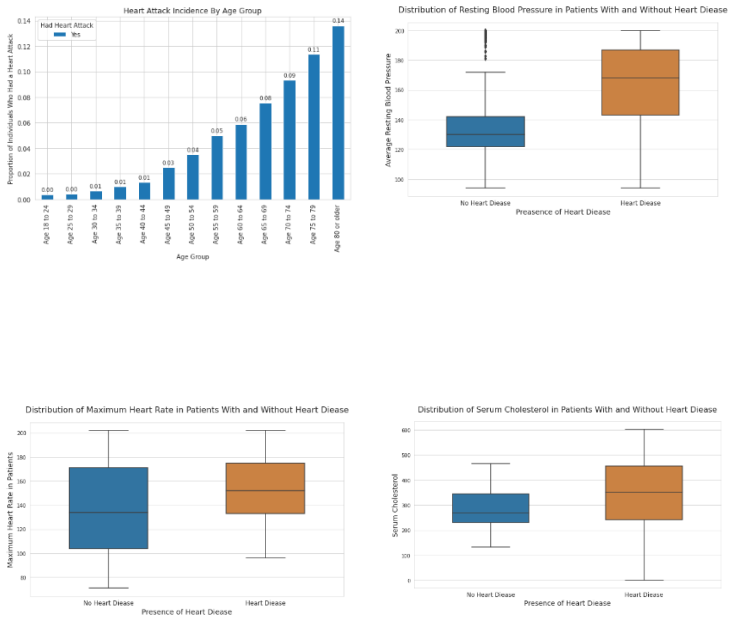

Fig. 8. Relationship between medical factors (age, cholesterol, BP, max heart rate) and heart attack risk.

Looking at Figure 8 we can see that age, resting blood pressure, maximum heart rate and serum cholesterol all have correlate to increased risk of heart attacks. Being older, increased resting blood pressure, increased maximum heart rate and increased serum cholesterol all correlate to increased risks of heart attacks.

To dive in further to what factors play a role and to be able to predict heart attacks based on features a random forest classifier was implemented.

```
Accuracy: 0.9480743826846865
Confusion Matrix:
 [[46366   207]
 [ 2348   284]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      1.00      0.97     46573
           1       0.58      0.11      0.18      2632

    accuracy                           0.95     49205
   macro avg       0.77      0.55      0.58     49205
weighted avg       0.93      0.95      0.93     49205
```

The Random Forest model trained to predict heart attack occurrence achieved an overall accuracy of 94.81%. The confusion matrix indicates that the model is highly effective at identifying individuals who did not have a heart attack. Out of 46,573 actual non-heart attack cases, it correctly predicted 46,366, misclassifying only 207 of them.

However, when it comes to identifying individuals who did have a heart attack, the model correctly predicted 284 out of 2,632 cases. These results are understandable given the imbalance in the BRFSS dataset. As there are significantly more records of individuals who did not report a heart attack than those who did. This imbalance makes it challenging for the model to learn and detect patterns related to the less frequent group.

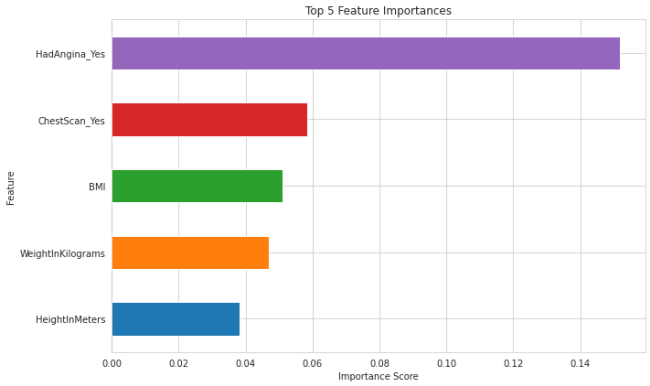

Fig. 9. Top five most important features from the Random Forest model.

The random forest classifier model also gave us the most important predictors to predict heart attacks. The top 5 features are shown in figure 9. The most important predictor is an individual who has Angina (also called ischemic chest pain, heart pain or discomfort that occurs when the heart doesn't get enough oxygen). Other important predictors include individuals who have had a chest scan, BMI, weight and height.

IV. Discussion

The methods applied in this yielded valuable insights. One of the key challenges was the imbalance and quality of the dataset. For example, certain categories, such as individuals who had a heart attack, were underrepresented, which may have skewed the classifier's ability to generalize and reduced its predictive power. Additionally, although the Random Forest model performed reasonably well, its predictions may have been influenced by multicollinearity among features or missing information not captured in the dataset (such as genetic factors or in-depth data on dietary habits). Another limitation was using statical tests such as the point-biserial correlation coefficient where only linear relationship were captured .

To address these issues, future work could involve acquiring larger and more balanced datasets. Additionally applying advanced preprocessing techniques such as SMOTE (Synthetic Minority Over-sampling Technique) could help balance the data. Furthermore, exploring additional machine learning models such as neural networks could offer improvements in accuracy.

V. Conclusion

The results of this project offer significant insights into the key factors that contribute to cardiovascular disease risk.

In terms of lifestyle, the analysis revealed that smoking and alcohol consumption were strongly correlated with a higher incidence of heart attacks. This finding follows established medical research that links tobacco use and excessive alcohol consumption to cardiovascular disease. Additionally, physical activity emerged as a significant factor, with those engaging in regular exercise displaying a considerably lower incidence of heart attacks. This highlights the well-known benefits of maintaining an active lifestyle as a preventive measure against heart disease. On the other hand, the analysis found no correlation between average hours slept and risk of heart attack.

Socioeconomic factors also played a prominent role in predicting heart attack risk. The results demonstrated that individuals with lower income and education levels had a higher incidence of heart attacks. Additionally, race-based differences were observed, with Multiracial and White individuals experiencing higher rates of heart attacks compared to Hispanic and Black individuals.

Medical factors, such as age, resting blood pressure, maximum heart rate, and serum cholesterol, were found to be strong predictors of heart attack risk.

The Random Forest model, trained to predict heart attack occurrence, achieved a high accuracy of 94.81%. The model performed well in predicting non-heart attack cases but struggled with heart attack cases due to the imbalance in the dataset. The model provided valuable insights into the top predictors of heart attacks, with Angina, BMI, weight, and height being the most impactful .

Overall, these results have practical implications for public health. Telling us that making changes in lifestyle such as not smoking or drinking can help reduce the risk of heart disease. Socioeconomically those with lower incomes and education levels By leveraging the Random Forest Classifier, we gained insight into the key predictors of cardiovascular issues. By using these insights to better public health ultimately improves health outcomes on a global scale.

References

[1] World Health Organization, “The top 10 causes of death,” WHO, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] Centers for Disease Control and Prevention, “Know Your Risk for Heart Disease,” CDC, 2023. [Online]. Available: https://www.cdc.gov/heart-disease/risk-factors/index.html

[3] S. Greenland, “Modeling and variable selection in epidemiologic analysis,” *American Journal of Public Health*, vol. 89, no. 5, pp. 708–716, 1999. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3863696/

[4] J. Zhao et al., “Predicting heart disease using machine learning algorithms,” *Journal of Healthcare Engineering*, vol. 2023, Article ID 1129483, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10378171/

[5] Centers for Disease Control and Prevention, “Behavioral Risk Factor Surveillance System (BRFSS) 2022 Summary Data Quality Report,” CDC, 2023. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2022.html

[6] N. Nahiduzzaman, “Cardiovascular Disease Dataset,” Mendeley Data, v1, 2022. [Online]. Available: https://data.mendeley.com/datasets/dzz48mvjht/1