June 30, 2024

Authors: Hamoedi Al-Aeshi (2614214)
Hengyi Liu (2821015)
Otsile Senye (2733053)
Berk Yavas, (2836576)

Topic: How do certain lifestyles affect cardiovascular disease?

1 Introduction

1.1 Problem Context

Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide. The disease clogs up the arteries, which makes it harder for blood to flow through the body. There are no symptoms for CVD, thus understanding the lifestyle factors that contribute to individuals suffering from CVD is critical for developing effective prevention and intervention strategies. This study aims to analyze various lifestyle variables and their potential impact on cardiovascular health.

1.2 Research question

The primary research question addressed in this report is: "How do certain lifestyles affect cardiovascular disease?" The analysis focuses on how variables such as age, gender, body measurements, blood pressure, cholesterol levels, glucose levels, smoking habits, alcohol consumption, and physical activity correlate with the likelihood of falling sick with cardiovascular disease. After the analysis there should be a clearer understanding on which variables in the dataset are more pertinent in the occurrence of cardiovascular disease.

Organization. In Section 1.2, we present a detailed overview of the dataset [1] and preprocessing steps. In Section 2, we describe the methodology used to analyze the data, while the main results are outlined in Section 3. Finally, we conclude the report in Section 5 and give some future research directions. sectionExploratory Data Analysis

1.3 Dataset Description

The dataset used for this analysis includes the following variables:

- **id**: A unique identifier for individuals in the dataset.
- **age**: Age of the individual in days.
- **gender**: Gender of the individual (1 for female, 2 for male).
- **height**: Height of the individual in centimeters.
- **weight**: Weight of the individual in kilograms.
- **ap_hi**: Systolic blood pressure. (The blood pressure when the heart pumps blood into the arteries)
- **ap_lo**: Diastolic blood pressure. (The blood pressure when the heart rests between beats and fills with blood)
- **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**: Smoking habits (0: non-smoker, 1: smoker).
- **alco**: Alcohol intake (0: non-drinker, 1: drinker).
- **active**: Physical activity (0: not active, 1: active).
- **cardio**: Presence of cardiovascular disease (0: no, 1: yes).

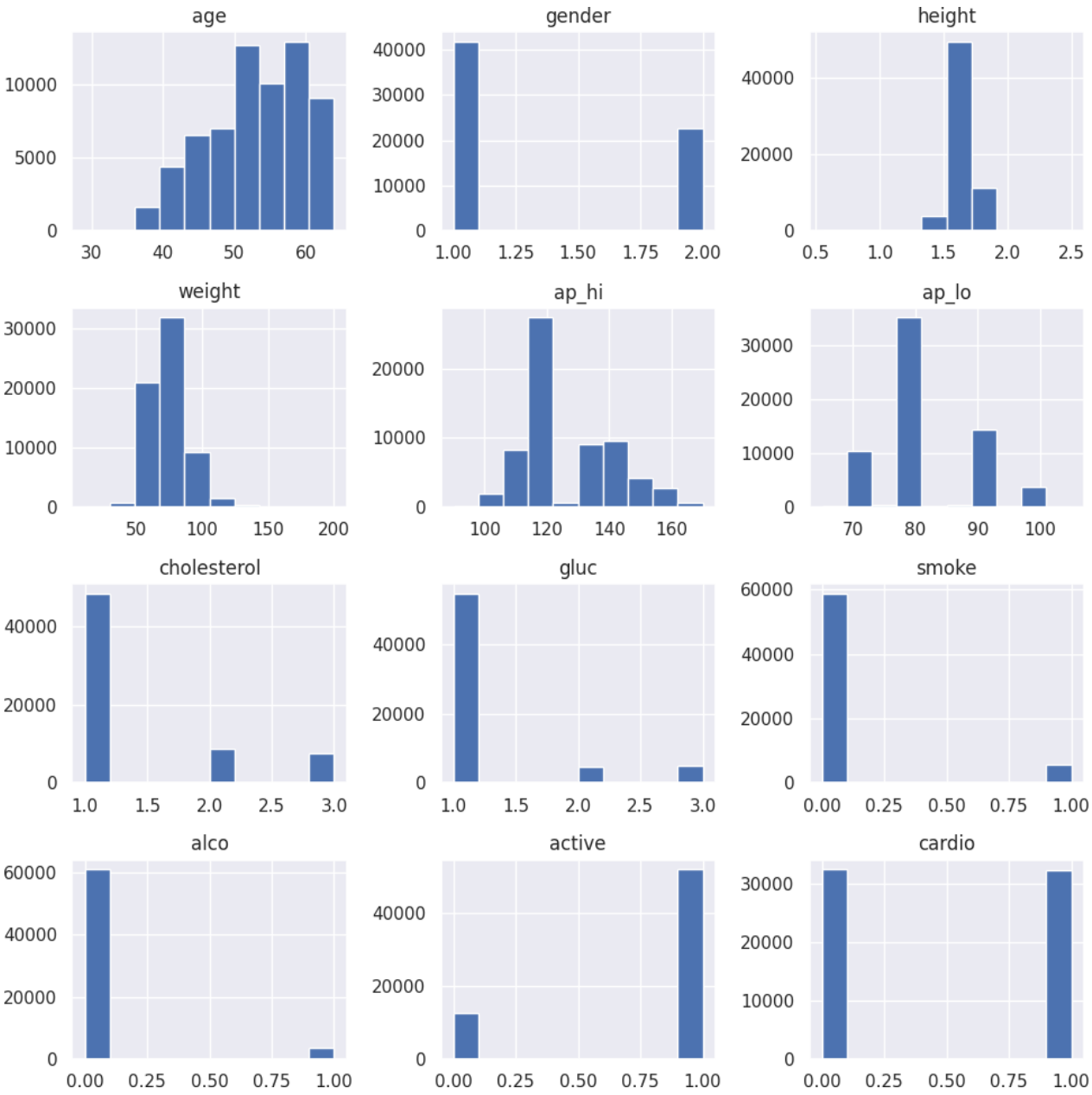

Figure 1: Distribution of the variables

Relevance of Variables Each variable provides insights into an individual's lifestyle and health status:

- *Age* and *gender* are fundamental factors in health analysis.
- *Height* and *weight* are essential for calculating Body Mass Index (BMI). BMI is a measure of body fat based on height and weight that applies to adult men and women. This can be used as an indicator for obesity, a CVD risk factor.
- *Blood pressure* (ap_hi and ap_lo) provides critical cardiovascular health indicators.
- *Cholesterol* and *glucose* levels indicate heart health.
- *Smoking* and *alcohol consumption* are lifestyle factors influencing CVD risk.
- *Physical activity* (active) is a positive lifestyle factor associated with reduced CVD risk.

1.4 Data Cleaning

The dataset underwent cleaning and adjustments as follows:

- Age was converted from days to years (dividing by 365.25).
- Gender, cholesterol levels, glucose levels, smoking, alcohol consumption, activity, and cardio variables were validated against predefined values.
- Height was converted from centimeters to meters for BMI calculation.
- Blood pressure values were validated for physical plausibility (non-negative values, systolic higher than diastolic).
- Outliers in blood pressure were identified and removed using the interquartile range method.

1.5 Data Analysis

After cleaning, the dataset was analyzed:

- The average *age* is approximately 53 years. With a youngest age of 29 years and an oldest age of 64 years.
- The average *systolic* blood pressure (ap_hi) is 126.6 and the average *diastolic* blood pressure (ap_lo) is 81.8.
- There were almost twice the number of females than males in the dataset. (41,803 females and 22,697 males).
- The percentage of people who have *cardiovascular disease* is 49.83%.
- The percentage of people who *smoke* is 8.76%, and the percentage of people who drink alcohol is 5.31%.
- The average *height* is 1.64 meters, and the average *weight* is 74 kilograms. Using the appropriate metric values for height and weight, we are able to calculate the BMI (Body Mass Index) using the formula:

$$BMI = \frac{weight}{height^2}$$

where:

- *weight* is the weight of the individual in kilograms,

- *height* is the height of the individual in meters.

- The average *BMI* is 27.46. Next, we wanted to see if this was impacted by the gender of the individual. The average male BMI is 26.73 and the average female BMI is 27.99. These results indicate that on average the female BMI is higher than the male BMI and the overall average BMI as well. The next step of the analysis was to see if these values would be impacted by age. In order to test this hypothesis the individuals in the dataset were split into three different age groups. Young (25-35 years), Middle-aged(35-55 years), and elderly(55-65 years). (Table 1) shows variations in BMI across the different ages.

Age Range	Men	Women
Young (25-35 years)	-	21.89
Middle-aged (35-55 years)	26.53	27.55
Elderly (55-65 years)	26.98	28.51

Table 1: Average BMI values for men and women across different age ranges.

From these results we are able to see that the BMI of individuals in the dataset increases as the age increases. We also see that even when individuals' ages increase the BMI of women still tends to be higher than the BMI of men. With the average BMI of men in each age group remaning less than the overall average BMI and the average BMI's of middle-aged and elderly women being higher than the overall average BMI.

- The impact of *cholesterol* levels on cardiovascular disease (CVD) was analyzed (Table 2 and Figure 2).

Cholesterol Level	With CVD	Without CVD
Normal	21316	27145
Above Normal	5142	3441
Well Above Normal	5688	1768

Table 2: Distribution of individuals with different cholesterol levels and their CVD status.

We can see from (Table 2) that for individuals with a normal cholesterol level there is a larger proportion of non-sufferers of cvd, although the number of sufferers and non-sufferers is relatively close. However, the table also shows that for individuals with above normal and well above normal cholesterol levels have a larger proportion of cvd sufferers with the difference between sufferers and non-sufferers being much larger for individuals with cholesterol levels well above normal.

- A similar analysis was conducted for *glucose* levels (Table 3 and Figure 3).

Glucose Level	With CVD	Without CVD
Normal	7139	16068
Above Normal	8358	7374
Well Above Normal	11649	4912

Table 3: Distribution of individuals with different glucose levels and their CVD status.

We can see from (Table 3) above that the pattern remains similar to that of the cholesterol levels. With individuals who have normal glucose levels have a larger proportion of non-sufferers. However, the proportion of sufferers increases as the glucose levels increase to above and well-above normal.

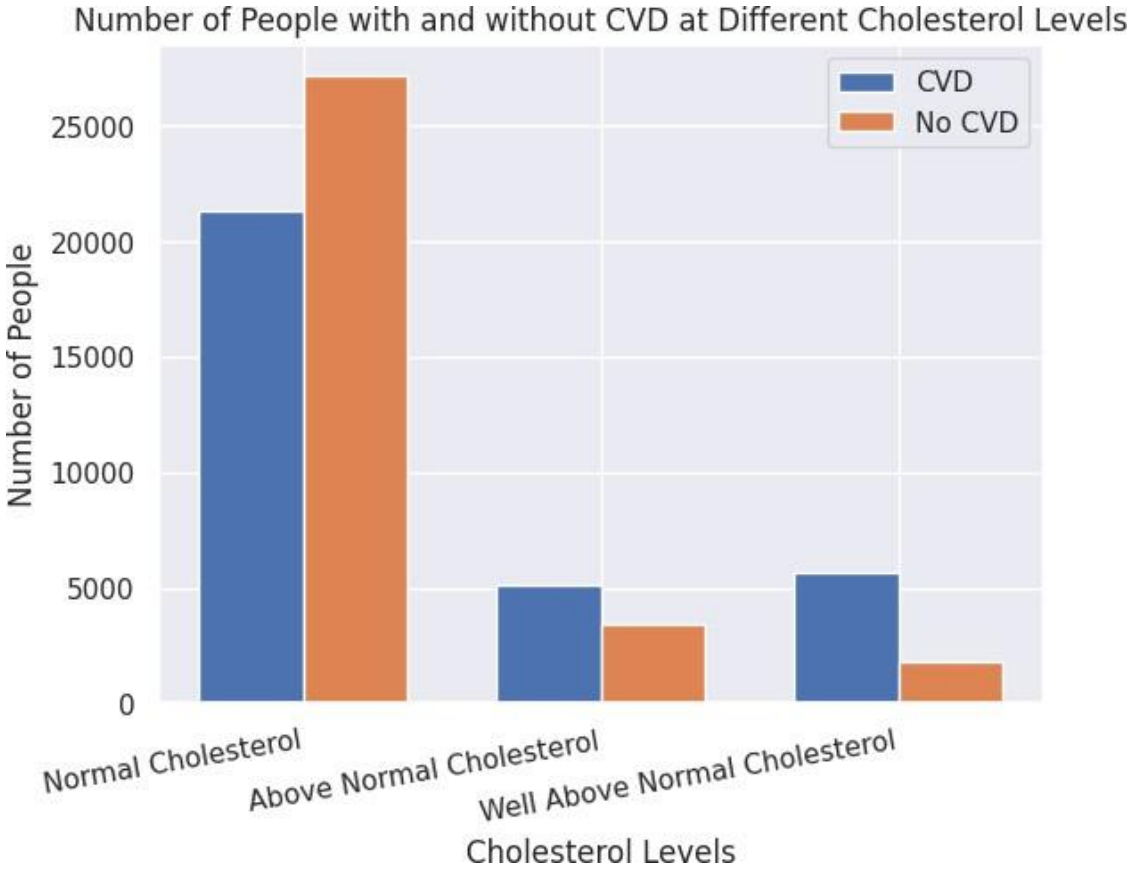

Figure 2: Plot illustrating the relationship between cholesterol levels and the incidence of CVD.

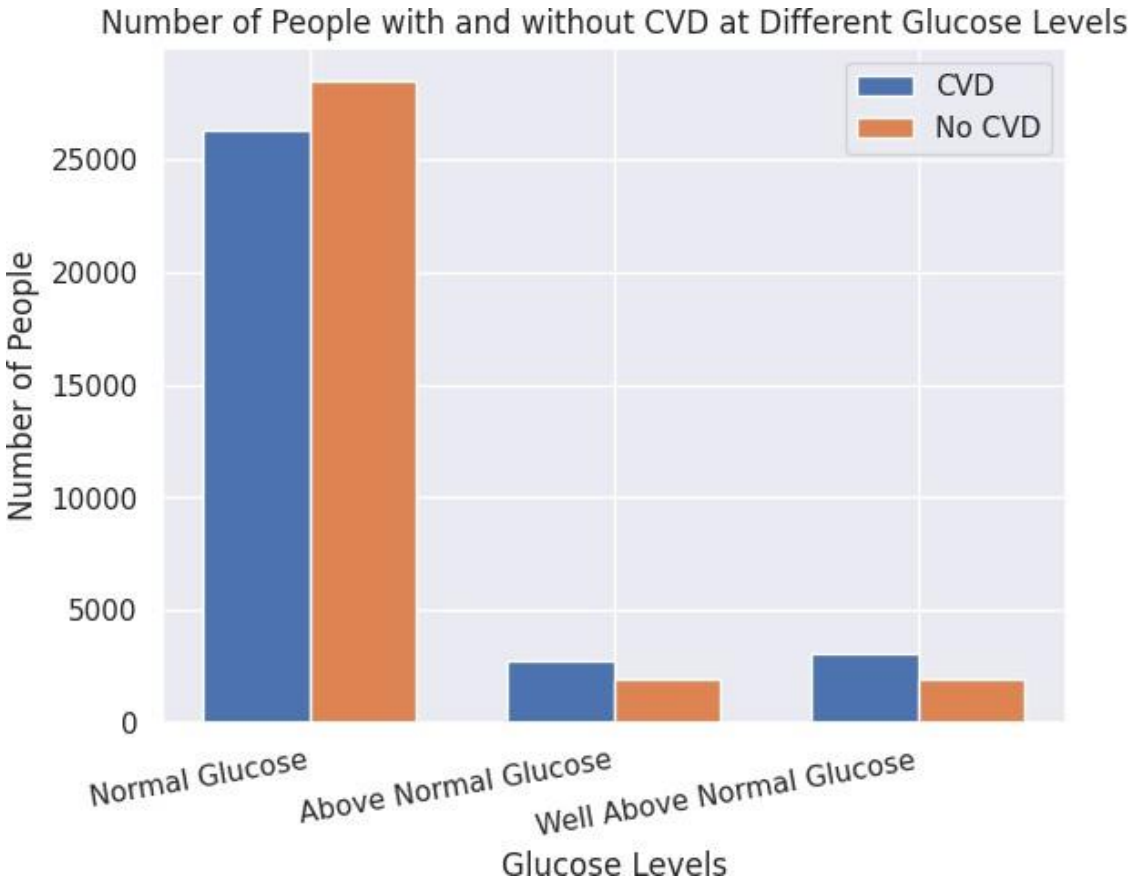

Figure 3: Plot illustrating the relationship between glucose levels and the incidence of CVD.

Cholesterol Level	Activity Status	Number of People	Proportion with CVD
Normal	Active	38,883	0.43
	Inactive	9,578	0.49
Above Normal	Active	6,845	0.59
	Inactive	1,738	0.64
Well Above Normal	Active	6,096	0.77
	Inactive	1,360	0.74

Table 4: Impact of Physical Activity on CVD Among Different Cholesterol Levels

We can see from (Table 4) that individuals with normal and above normal cholesterol levels have higher proportions of cvd sufferers for people who are inactive. Although, there is a larger number of individuals with a normal cholesterol level we are able to see that being active has a greater impact on the liki-hood of suffering from cvd than the cholesterol level for individuals in the dataset. We also see from the table that individuals with a cholesterol level well above normal have a higher proportion of cvd sufferers who are active. This is likely because individuals with a cholestrol leve well above normal already have a very high chnace of suffering from cvd and the number of active individuals is much higher than the number of inactive ondoviduals with a well above normal cholesterol level. This difference in proportion may be the reason for the larger proportion of suffererers despite them being active.

- The percentage of people who have an *active* lifestyle is 80.34%.

2 Methodology

In this section, we will utilize the dataset to build the models for cardiovascular disease prediction. We split the data into training (80%) and testing (20%) sets to train and evaluate the models. The numerical variables are scaled to a similar scale, including age, height, weight, ap_hi, ap_lo, cholesterol, and gluc.

2.1 KNN

First, we examined the dataset and found a balanced distribution of individuals with and without cardiovascular disease, which is suitable for applying the K-Nearest Neighbors algorithm (KNN). KNN, a non-parametric method for classification, finds the K-closest data points to a new data point and classifies it based on the majority class.

We built a KNN model with a chosen number of neighbors (K=9) and the Euclidean distance metric to train the initial model, which achieved an accuracy of approximately 69.8% on the testing set, which is considered acceptable. In Figure 4, the confusion matrix plot illustrated that our predictive model has a balanced performance across both classes (with or without cardiovascular disease). Accuracy can be calculated by $\frac{TP+TN}{TP+TP+FP+FN}$ to verify.

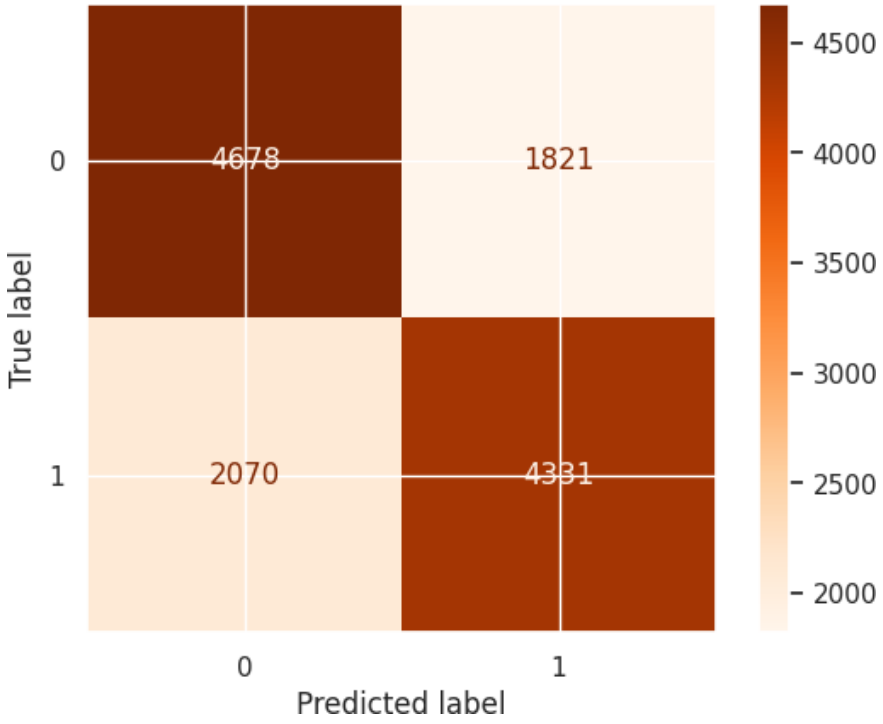

Figure 4: KNN

The performance of KNN is largely influenced by the number of neighbors (K). To fine-tune the model and identify the optimal K value, we employed cross-validation with 5 folds. This process was repeated with odd k values ranging from 3 to 27. The optimal number of neighbors identified was K=23, with a better corresponding accuracy of around 71.5%.

2.2 Logistic Regression

As the dependent variable 'cardio' is dichotomous, we applied Logistic Regression to classify. We train this logistic regression model with L2 regulation to handle possible outliers and avoid overfitting. It correctly predicts the existence of cardiovascular disease among the test set for approximately 72.16%.

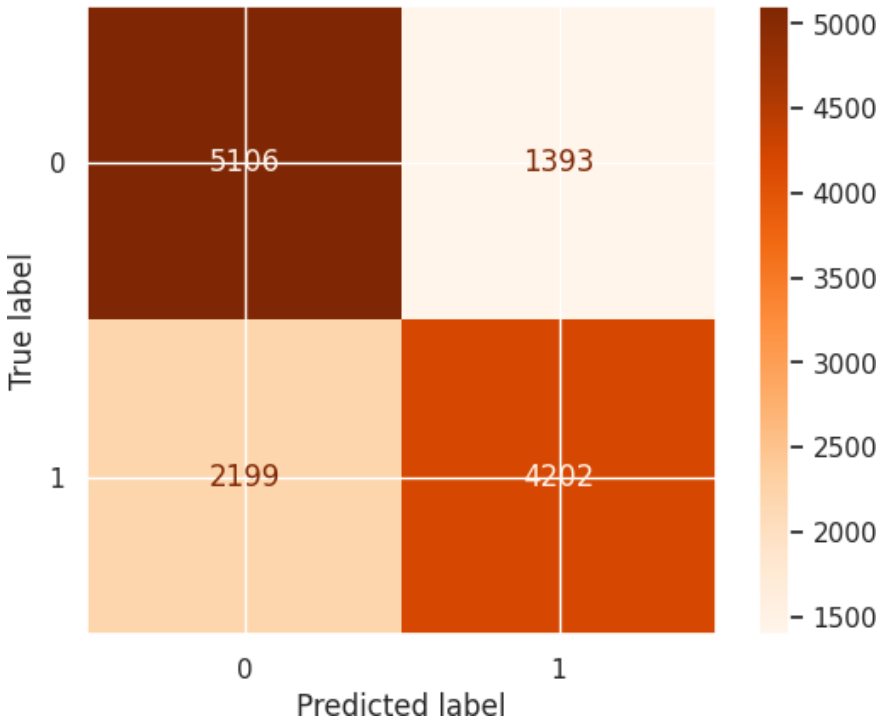

Figure 5: Logistic Regression

Moreover, we conducted feature importance to explore the potential irrelevant features. As depicted in the below Figure, the included features contribute to the predictive model varyingly. The systolic blood pressure (ap_hi) plays a crucial role in model prediction and its scores (0.876) are above twice higher than others, followed by age (0.352) and cholesterol (0.334). Though gluc (0.079) and height (0.039) have lower importance scores, they still contribute to the overall model prediction. Therefore, we did not reduce the dimension of the model to have a comprehensive use of the dataset. Reducing the number of involved variables may lead to over-simplification, as there are only 7 numerical features except for gender and the dependent variable - cardio.

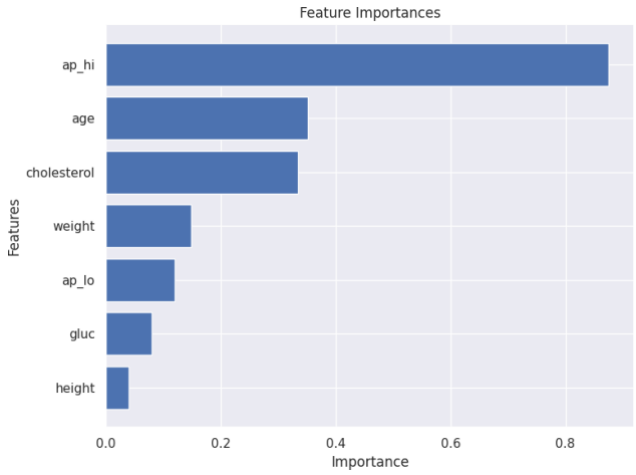

Figure 6: Feature Importance

2.3 Random Forest

Random Forest, a non-parametric model for regression and classification, is commonly used for medical diagnosis in healthcare fields. It combines many decision trees into a single model for better prediction, which can also avoid overfitting. The random forest prediction model is fitted on the training data

with default parameters setting, yielding an accuracy of 69.13%. To visualize its performance, the confusion matrix shows that the True Negative and True Positive are more intensely clustered than the previous two models, though their accuracy is slightly lower.

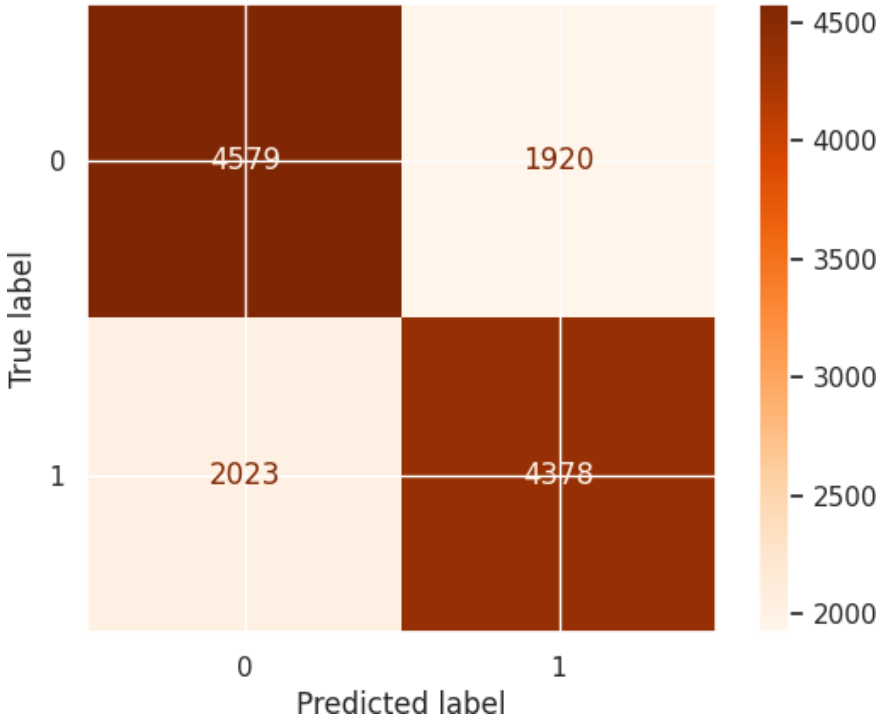

Figure 7: Random Forest

2.4 Decision Tree Classifier

Lastly, we train a Decision Tree Classifier, a tree-like model for clustering, though it is a relatively weak classifier but intuitively easy to interpret. Each node of the tree is a feature and the branches illustrate the ranges of the corresponding feature. This can be especially useful for making an informed decision and improving interpretability.

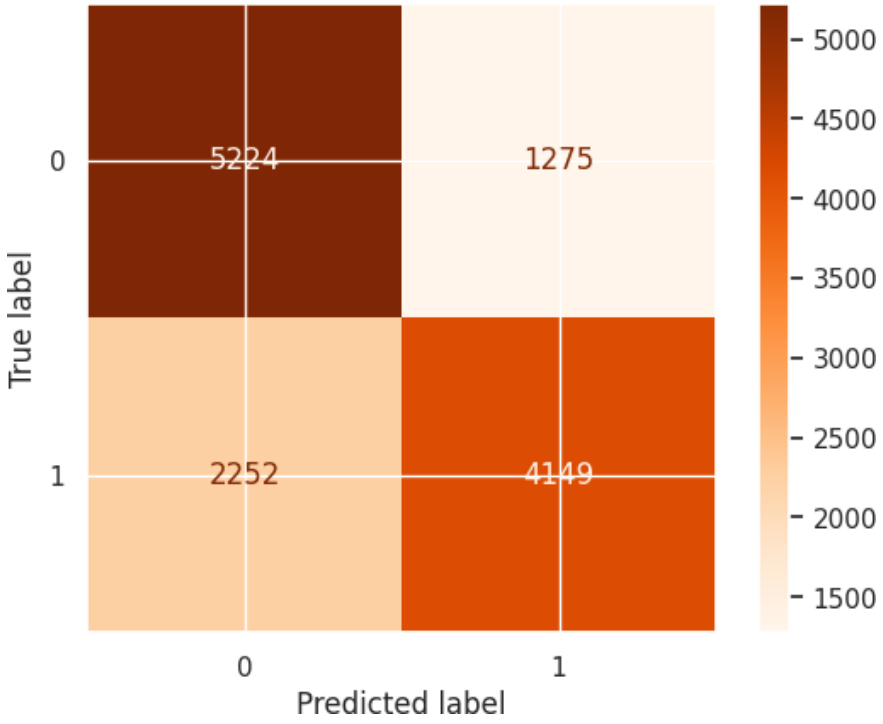

Figure 8: Decision Tree Classifier

Model	Correct classification percentage
KNN	71.58%
Logistic Regression	72.15%
Random Forest	69.54%
Decision Tree	72.66%

Table 5: Probability of correct classification for different estimation methods

Overall, the decision tree classifier method slightly outperforms the other models as it has a higher correct classification probability 72.66%. These 4 classification models have their strengths: KNN is easy to implement with little assumption; Logistic Regression is effective when the relationship is linear; Random Forest offers improved generalization [2].

3 Results and Discussion

One of the important questions to answer is, is it possible to lower the chances of getting cardiovascular disease while for example drinking, smoking, cholesterol levels, and blood pressure levels? To do this, we look at if people smoke for example, and divide the group of smokers into a group of people who are active and people who are not active. The result are as follows:

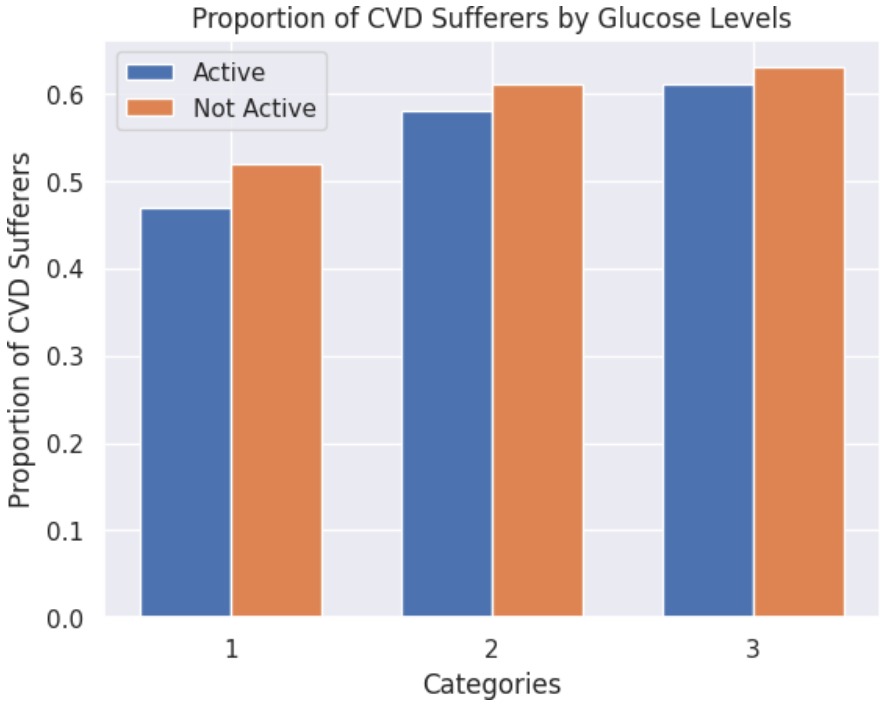

Figure 9: Plot showing the probability of having cardiovascular disease with different levels of glucose, while having an active or inactive lifestyle

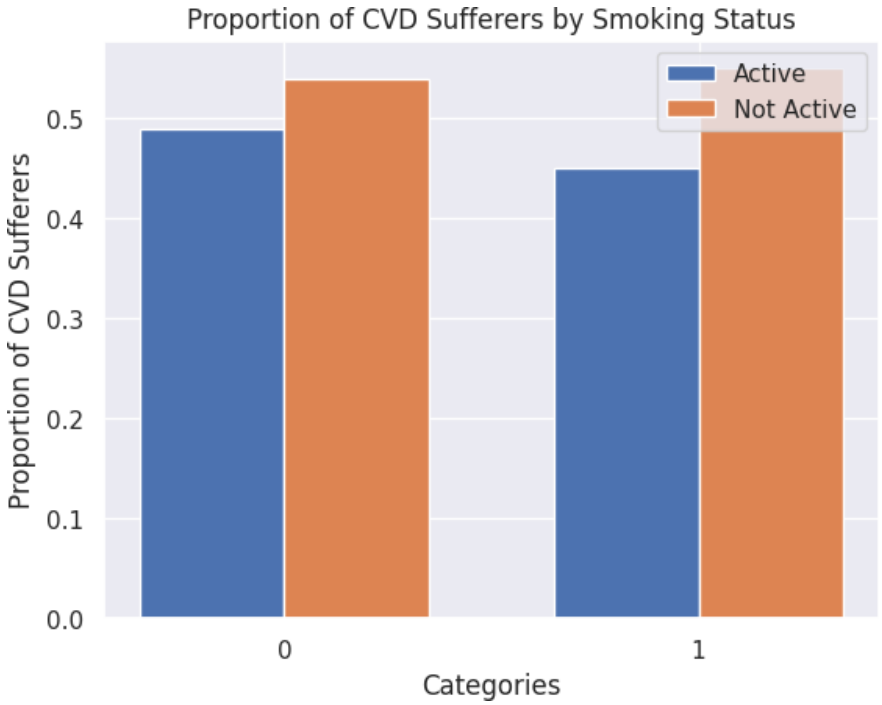

Figure 10: Plot showing the probability of having cardiovascular disease with people who smoke or do not smoke, while having an active or inactive lifestyle

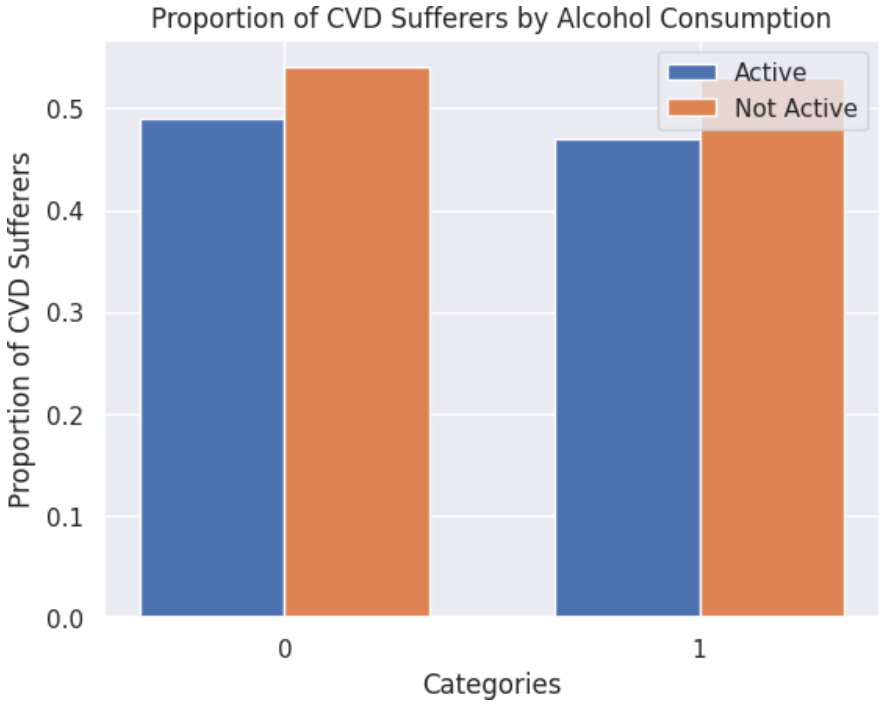

Figure 11: Plot showing the probability of having cardiovascular disease with people who drink or do not drink alcohol, while having an active or inactive lifestyle

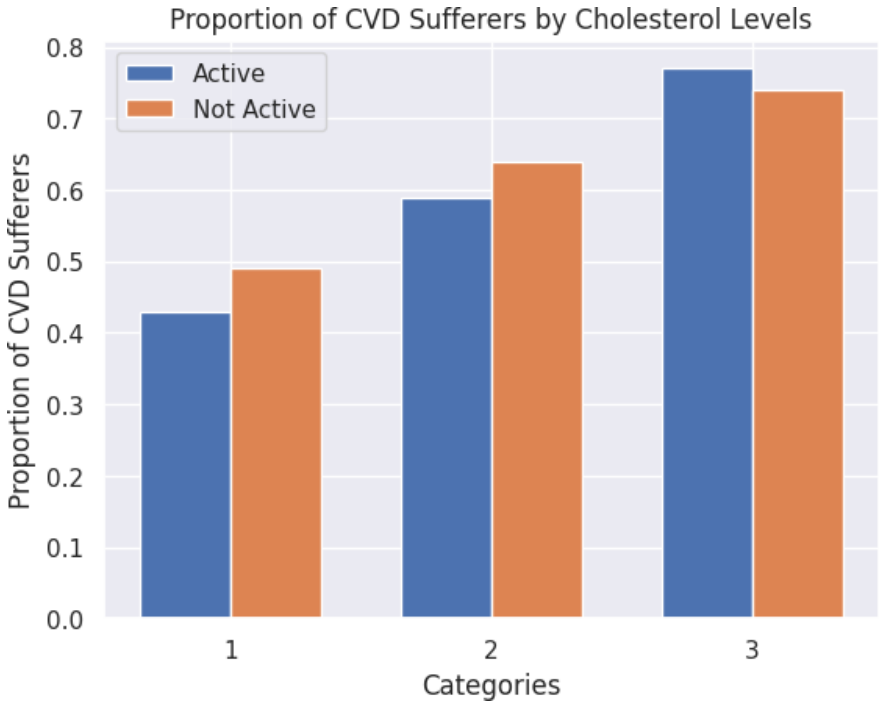

Figure 12: Plot showing the probability of having cardiovascular disease with different cholesterol levels, while having an active or inactive lifestyle

In figure 9 we see that as the glucose level increases, so does the proportion of people with cardiovascular disease. But, we also see that being active lowers the probability of having cardiovascular disease, compared to being not active. In figure 10 and 11 we do see something interesting. We see that the proportion of people who do smoke or drink, and are active is lower than people who do not smoke, but are active. In figure 12 we actually see the opposite happening. People who have well

above normal (3) cholesterol levels and are active, actually seem to have a higher chance of getting cardiovascular disease compared to people who are not active.

4 Limitations

There are limitations to every research, including this one. First, we will look at the dataset that is used. This dataset uses binary and ternary values. This may create an issue. That is because, for example, smoking one cigarette a day is very different from smoking one packet in a day. One could say that it is maybe more likely that someone who smokes more cigarettes per day, is more likely to have cardiovascular disease, compared to someone who barely smokes.

Next, we look at the distribution of the data. We see that there are twice as many females as there are males. Since the dataset is big enough, this might not be a problem. But since we are looking at for example males, who do smoke and are active and also have cardiovascular disease, this combination of number of males becomes very small. Another thing is the fact that we are looking at relatively old people, since the average age is about 53 years old.

Finally, There is also the issue with using BMI as an indicator of health and weight. Although generally it is true that as people have more fat, they are more likely to have cardiovascular disease and other health problems. But, BMI does not distinguish between the weight of fat and the weight of muscle. And we may assume that muscle weight is healthier to have than fat. Another thing BMI does not consider, is the fact that fat allocation is different per person. In fact, visceral fat is linked to cardiovascular disease [3]. One final thing to note is that since the average BMI is so high, within overweight territory, we are barely looking at what happens when someone is underweight.

5 Conclusion

In conclusion, we see that drinking, high glucose levels, high cholesterol levels, smoking, high blood pressure levels and being overweight may lead to a higher chance of having cardiovascular disease in general. However, we can mitigate these effects by exercising. In most cases, being active will lower the chances of having cardiovascular disease, even though the person has high glucose levels for example. Except for the case of having high cholesterol levels and being active. However, these results may be influenced by the fact that there is a higher proportion of active individuals in the higher cholesterol and glucose levels which would cause a skew in the results.

Furthermore a majority of methods has been used as approaches such KNN (K-Nearest Neighbours Algorithm, Logistic Regression, Random Forest and Decision Tree Classifier. KNN classifies based on proximity to data points; Logistic Regression models the probability of outcomes; Random Forest combines multiple decision trees to enhance prediction accuracy, while Decision Tree Classifier splits data into branches for classification. The data cleaning part proved to be crucial for normalizing the values as much as the sophisticated methods used in the evaluating and training.

Bibliography

[1] Cardiovascular Disease Dataset. Retrieved from:

https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[2] Shah, K., Patel, H., Sanghvi, D., Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1), 12.

[3] Romero-Corral, A., Somers, V. K., Sierra-Johnson, J., Thomas, R. J., Collazo-Clavell, M. L., Korinek, J., Allison, T. G., Batsis, J. A., Sert-Kuniyoshi, F. H., & Lopez-Jimenez, F. (2010). Accuracy of body mass index in diagnosing obesity in the adult general population. *International Journal of Obesity, 34* (6), 791–799. https://doi.org/10.1038/ijo.2010.5

A Additional figures

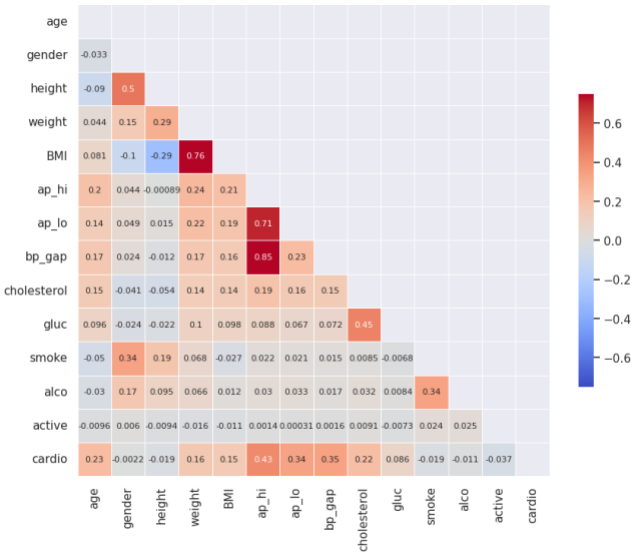

Figure 13: Heatmap for all the variables