FINAL GROUP PROJECT REPORT

Cardio Disease Dataset Analysis

Aug 6, 2023

MBAN 6100
Prof. Delina Ivanova

Group 4
Pelumioluwa Abiola
Yuhui Gu
Dongchen Wu
Wanqiu Jiang
Mingcheng Xu
Baoqiang Zhang

ABSTRACT

This report provides a comprehensive analysis of a dataset focused on the prevalence of cardiovascular disease (CVD) in relation to specific determining factors. Utilizing data from 70,000 patient records, the study investigates the correlation between CVD and variables as well as predicts if a patient has CVD.

Contents

Introduction

Cardiovascular disease is a global major health concern. With the rise of unhealthy lifestyles and poor diets, the risk of cardiovascular disease is increasing. Therefore, detecting the possibility of cardiovascular disease early can make treatments more effective.

In our study, we analysed data from 70,000 patient records to predict the risk of cardiovascular disease. We considered factors like age, blood pressure, cholesterol, and lifestyle habits such as smoking and alcohol consumption. Importantly, we also introduced the body mass index (BMI) as an additional variable, as it also provides a clearer picture of an individual's health relative to their weight and height.

We used four different machine learning models to make these predictions: Logistic Regression, Random Forest, Support Vector Machines, and K-Nearest Neighbours. Our aim is to find out which model is the most accurate in predicting heart disease.

This report will detail our methods and findings. We hope our research can assist medical professionals in the early detection and treatment of cardiovascular disease.

Data Selection and Justification

The dataset under consideration consists of 70,000 observations. Each observation is characterized by 13 features, providing comprehensive data on individual patients. These features include:

- `id`: A unique identifier for each patient.
- `age`: The age of the patient in days.
- `gender`: Categorical data indicating the gender of the patient, represented as 1 for women and 2 for men.
- `height`: The height of the patient in centimetres.
- `weight`: The weight of the patient in kilograms.
- `ap_hi`: The systolic blood pressure measurement of the patient.
- `ap_lo`: The diastolic blood pressure measurement of the patient.
- `cholesterol`: A categorical variable indicating the cholesterol level of the patient. It is represented as 1 for normal, 2 for above normal, and 3 for well above normal levels.
- `gluc': A categorical variable indicating the glucose level of the patient. it is represented as 1 for normal, 2 for above normal, and 3 for well above normal levels.
- `smoke`: A binary variable indicating whether the patient is a smoker 1 or not 0.
- `alco`: A binary variable indicating whether the patient consumes alcohol 1 or not 0.
- `active`: A binary variable indicating whether the patient is physically active 1 or not 0.
- `cardio`: This is the target variable, It is a binary variable indicating 1 for the presence or 0 for the absence of cardiovascular disease in the patient.

Problem statement

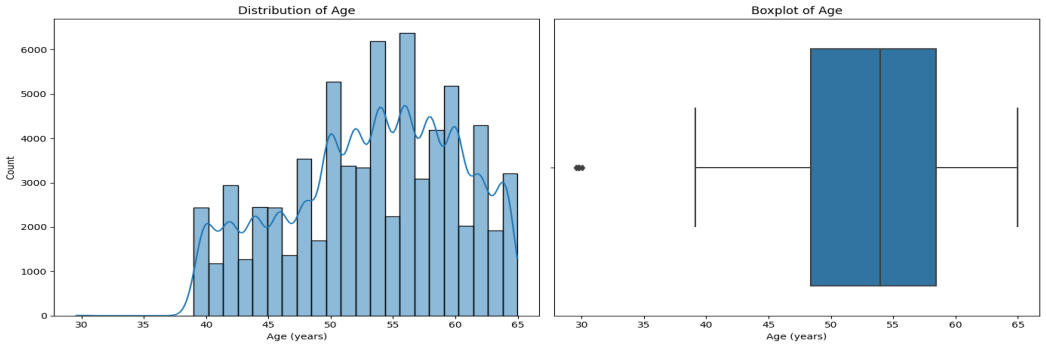
The primary objective of this project is to identify the key health parameters that significantly influence the presence of cardiovascular disease in individuals. Leveraging these parameters, we aim to develop an efficient predictive model capable of determining whether a patient is likely to have cardiovascular disease. Our goal is to compare the performance of different machine learning models and select the most efficient one for predicting the presence of cardiovascular disease in patients.

Hypothesis

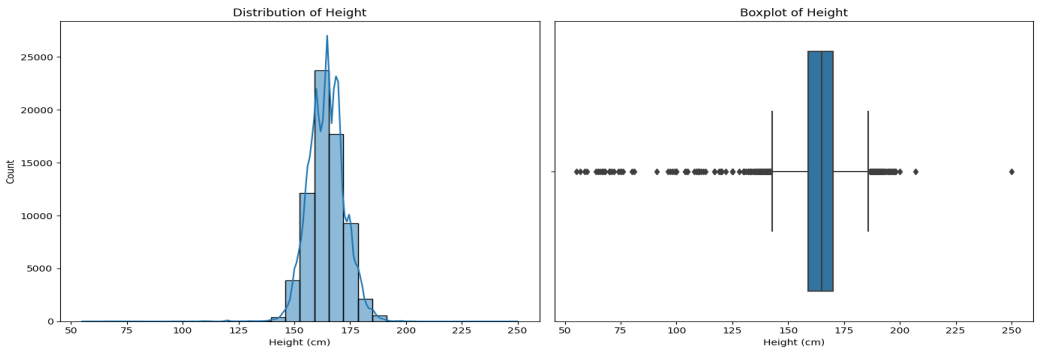
We hypothesize that age, blood pressure (both systolic and diastolic), and cholesterol levels are the most significant factors in determining the presence of cardiovascular disease in individuals. We postulate that older individuals, those with higher blood pressure, and those with elevated cholesterol levels are more likely to have cardiovascular disease. Furthermore, we anticipate that a predictive model utilizing these key health parameters will be able to accurately classify individuals as either having or not having cardiovascular disease. We expect that by comparing different machine learning models, we will be able to select the most efficient one for this task.

Numerical Variables

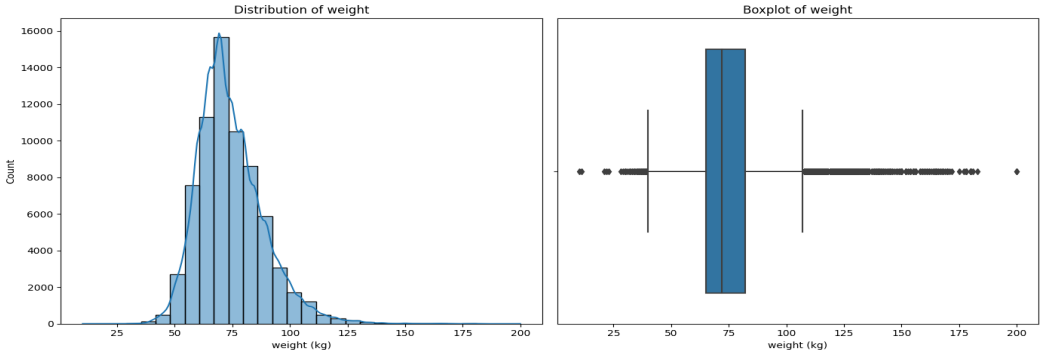
- **Age**: The average age of patients in the dataset is approximately 19,468 days, which translates to 53 years. The youngest patient is around 10,798 days old (30 years old), and the oldest is around 23,713 days old (65 years old). The age distribution is slightly left-skewed, indicating that there are more patients in the dataset who are younger than the average age. It's important to note that individuals with cardiovascular disease tend to be older than those without the disease.



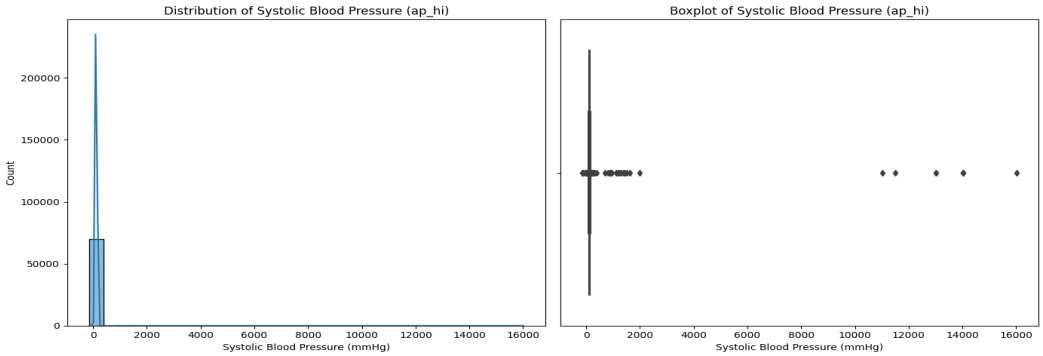
- **Height**: The patients' height ranges from 55 cm to 250 cm, with an average of approximately 164.36 cm. The distribution of height is roughly normal, but there are some extreme values, particularly at the lower end, that could be considered outliers. It's unlikely for an adult to be 55 cm tall, suggesting potential data entry errors. The median height is slightly lower for individuals with cardiovascular disease compared to those without the disease.



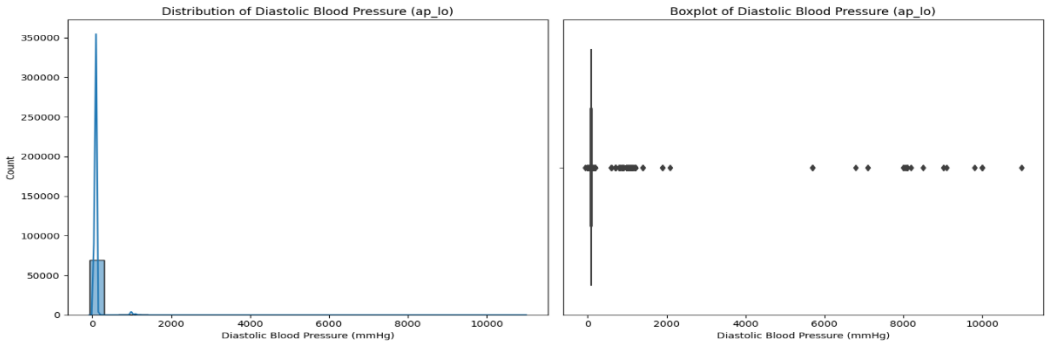
- **Weight**: The average weight of the patients is approximately 74.21 kg. The weight distribution is slightly right skewed, indicating that there are more patients with a weight lower than the average weight. The lightest individual weighs 10 kg, which is likely due to a data entry error, while the heaviest individual weighs 200 kg. Individuals with cardiovascular disease tend to have higher weights than those without the disease.



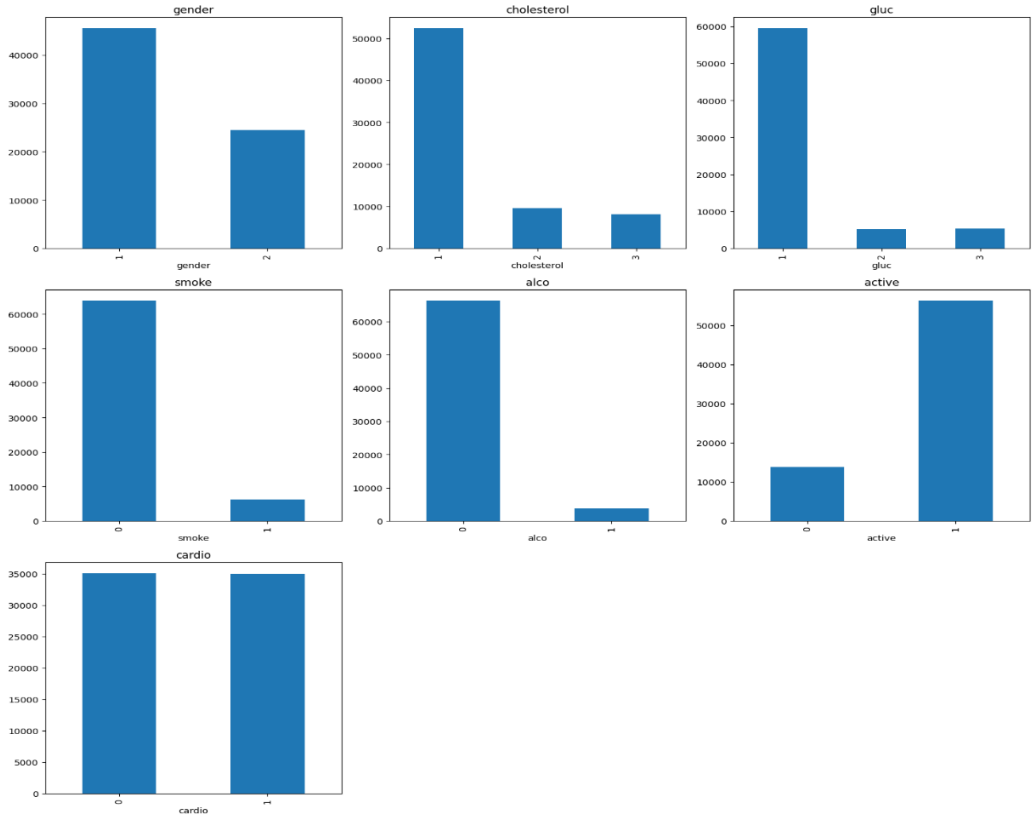
- **Ap_hi (Systolic blood pressure):** The systolic blood pressure is the blood pressure when the heart beats. The average systolic blood pressure is around 128.82, but there are some suspicious values such as -150 and 16,020, which are likely due to data entry errors. The distribution is heavily affected by outliers, obscuring the true distribution. However, it's clear that individuals with cardiovascular disease tend to have higher systolic blood pressure than those without the disease.



- **ap_lo (Diastolic blood pressure):** The diastolic blood pressure is the blood pressure in between heart beats. The average diastolic blood pressure is around 96.63. Similar to systolic blood pressure, there are some suspicious values such as -70 and 11,000, which are likely due to data entry errors. Despite the influence of outliers, it's clear that individuals with cardiovascular disease tend to have higher diastolic blood pressure than those without the disease.



Categorical Variables

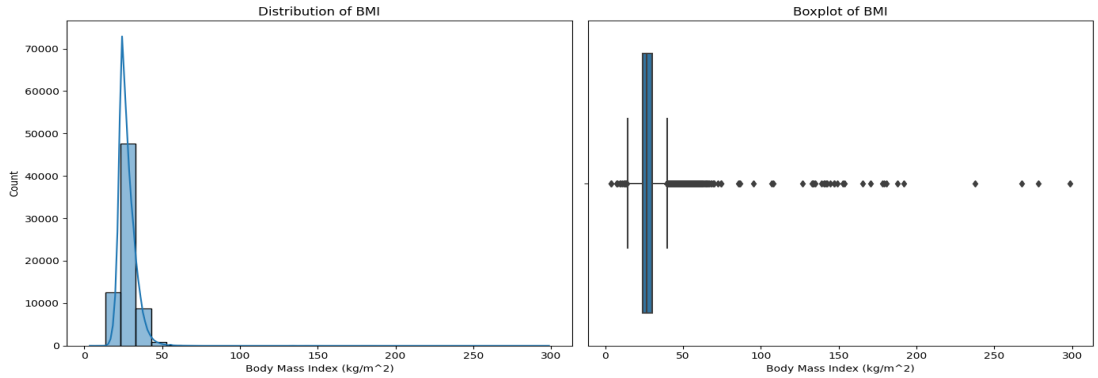


- **Gender**: The dataset contains more women (represented as 1) than men (represented as 2). However, there doesn't appear to be a significant difference in the prevalence of cardiovascular disease between the two genders.
- **Cholesterol:** Most individuals have normal cholesterol levels, but a significant number have levels that are above normal or well above normal. Individuals with above-normal and well above-normal cholesterol levels have a higher prevalence of cardiovascular disease than those with normal levels.
- **Gluc (Glucose):** Similar to cholesterol, most individuals have normal glucose levels, but a significant number have levels that are above normal or well above normal. Individuals with above-normal and well above-normal glucose levels have a higher prevalence of cardiovascular disease than those with normal levels.
- **Smoke:** Most individuals in the dataset do not smoke. There doesn't seem to be a significant difference in the prevalence of cardiovascular disease between smokers and non-smokers.
- **Alco (Alcohol):** Most individuals in the dataset do not consume alcohol. There doesn't seem to be a significant difference in the prevalence of cardiovascular disease between those who consume alcohol and those who do not.
- **Active (Physical activity):** Most individuals in the dataset are physically active. However, individuals who are not physically active seem to have a slightly higher prevalence of cardiovascular disease than those who are active.

Body Mass Index (BMI)

Weight is a crucial factor in cardiovascular health, but considering weight alone can be misleading as it doesn't account for the height of the individual. Therefore, we introduced a derived variable - BMI, calculated as 'weight' / ('height' / 100) ** 2. The average BMI is approximately 27.56, with a minimum of around 3.47 and a maximum of approximately 298.67.

This wide range suggests significant diversity in the body mass index of the individuals in the dataset. However, some extreme values, particularly on the higher end, might be due to outliers or data entry errors. These outliers require careful attention during the data cleaning process as they could influence the performance of the predictive models.



Exploratory Data Analysis

Our dataset consists of both numerical and categorical variables. Numerical variables include age, height, weight, ap_hi, and ap_lo, while categorical variables encompass gender, cholesterol, gluc, smoke, alco, active, and cardio. Fortunately, there are no missing values in the dataset.

Outlier Treatment

An essential part of our pre-processing was the identification and handling of outliers within our variables:

- **Age:** From the distribution of age, we set the age range between 40 and 65 years, excluding four outliers who were younger than 30 years.
- **Height**: From the distribution of height, values below 100 cm and above 220 cm were considered outliers due to the rarity of such values in adult human heights, leading to the removal of 30 outliers.
- **Weight:** from the distribution of weight, values below 40 kg or above 140 kg were flagged as outliers due to their uncommon occurrence in adult human weights. This process led to the removal of 150 outliers.
- **Ap_hi and Ap_lo:** For the systolic (ap_hi) and diastolic (ap_lo) blood pressure values, we established permissible ranges as 80-190 and 50-120, respectively. Although the Centers for Disease Control and Prevention (CDC) designates hypertension as a systolic reading of 120 mm Hg or higher or a diastolic reading of 80 mm Hg or higher. It is important to note that blood pressure readings outside the 'normal' range can still be clinically relevant, especially when studying their relationship with cardiovascular disease. Therefore, our chosen ranges are broader to ensure we capture all potentially significant data. Nonetheless, values beyond these thresholds were considered outliers, as they likely represent either data entry errors or extreme, atypical blood pressure readings that could distort our analysis. This led to the removal of 427 outliers in ap_hi and 1136 outliers in ap_lo.
- **Bmi**: From the distribution of BMI, we considered BMI values below 15 and above 45 as outliers. Typically, a BMI below 15 is classified as very severely underweight, while a BMI above 40 is considered very severely obese (Zierle-Ghosh A, 2022). However, in this dataset, there are at least 1,000 individuals with a BMI over 40. Therefore, to encompass this substantial portion of the data and provide a more comprehensive analysis, we opted to set the upper limit for BMI at 45. This led to the removal of 632 outliers in bmi.

Overall, we removed 2,122 outliers, which constituted about 3.0% of the total observations, leaving us with a dataset of 67,878 observations. This cleaning process had unnoticeable effects on the means of several variables.

- **Age**: The average age saw a minor decrease from 53.34 years to 53.32 years.
- Height: The average height experienced a slight increase from 164.36 cm to 164.48 cm.
- **Weight**: The average weight decreased from 74.21 kg to 73.34 kg, indicating that the removed observations were heavier than the average.
- **Ap_hi**: The average systolic blood pressure decreased from 128.82 to 126.39, implying that the removed observations had higher than average systolic blood pressure.
- **Ap_lo**: The average diastolic blood pressure decreased significantly from 96.63 to 81.22, suggesting that the removed outliers had considerably higher diastolic blood pressure than the average.
- **BMI**: The average BMI decreased from 27.56 to 27.28, indicating that the removed outliers or observations had higher than average BMI.

Below is the cleaned dataset.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol
count	67878.000000	67878.000000	67878.000000	67878.000000	67878.000000	67878.000000	67878.000000	67878.000000
mean	49977.712499	53.319455	1.350113	164.480318	73.739196	126.391438	81.218362	1.361826
std	28851.770832	6.767716	0.477009	7.816056	13.490588	16.225181	9.252232	0.676843
min	0.000000	39.000000	1.000000	120.000000	40.000000	80.000000	50.000000	1.000000
25%	24995.250000	48.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000
50%	50028.500000	54.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000
75%	74876.750000	58.000000	2.000000	170.000000	82.000000	140.000000	90.000000	1.000000
max	99999.000000	65.000000	2.000000	207.000000	140.000000	190.000000	120.000000	3.000000

gluc	smoke	alco	active	cardio	bmi
67878.000000	67878.000000	67878.000000	67878.000000	67878.000000	67878.000000
1.223268	0.088129	0.053272	0.803500	0.492590	27.282870
0.569122	0.283484	0.224577	0.397354	0.499949	4.839981
1.000000	0.000000	0.000000	0.000000	0.000000	15.035584
1.000000	0.000000	0.000000	1.000000	0.000000	23.875115
1.000000	0.000000	0.000000	1.000000	0.000000	26.291724
1.000000	0.000000	0.000000	1.000000	1.000000	30.083829
3.000000	1.000000	1.000000	1.000000	1.000000	44.997166

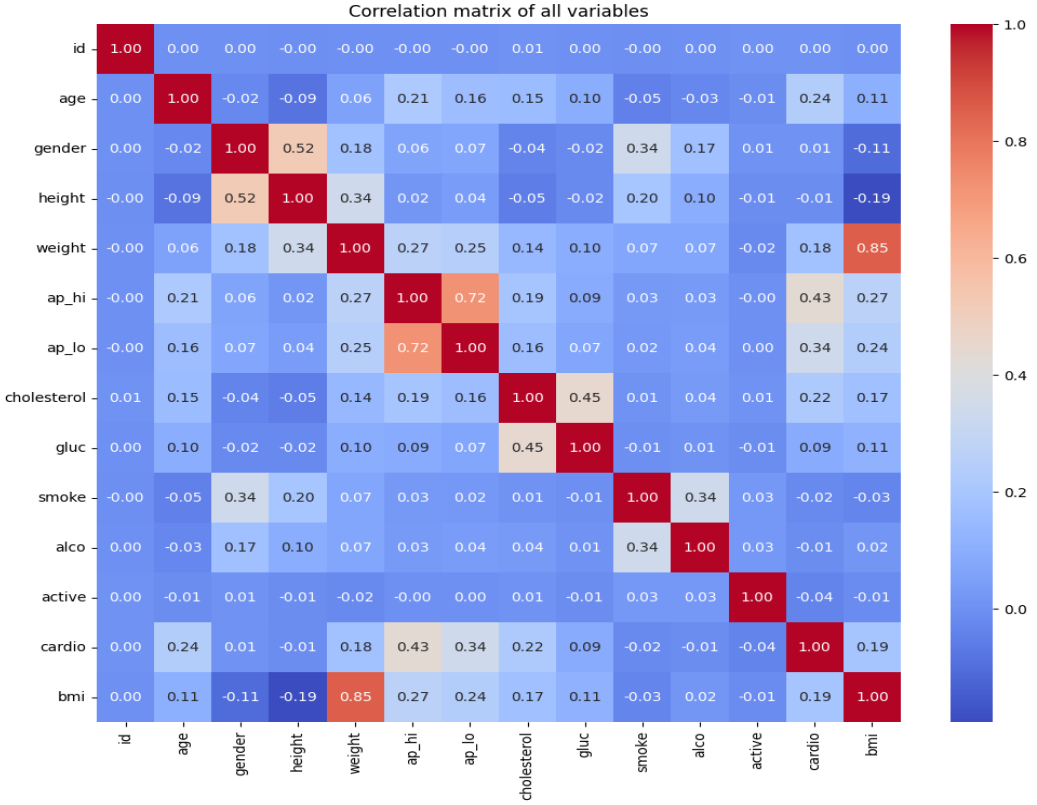
Correlation Matrix

Our correlation matrix provided valuable insights into the relationships between various factors and the presence of cardiovascular disease (cardio). Notably, age, weight, ap_hi, ap_lo, cholesterol, and BMI showed a moderate positive correlation with cardio, indicating that higher values of these factors are associated with a higher risk of cardiovascular disease.

- **Age**: There is a positive correlation with cardio (0.24), implying that the risk of cardiovascular disease tends to increase with age.
- **Weight and ap_hi/ap_lo:** Weight exhibits a positive correlation with both systolic (0.26) and diastolic (0.24) blood pressure, suggesting that higher weight may be linked to higher blood pressure.
- **Cholesterol**: There is a moderate positive correlation with cardio (0.22), suggesting that high cholesterol levels could be a risk factor for cardiovascular disease.

- **Ap_hi and ap_lo:** Both show a moderate positive correlation with cardio (0.43 and 0.34, respectively), indicating that higher blood pressure is associated with a higher risk of cardiovascular disease.
- **BMI**: A positive correlation with cardio (0.19) indicates that a higher BMI is associated with a higher risk of cardiovascular disease.

These observations align with our initial hypothesis that age, blood pressure, cholesterol level, and BMI are indeed the three most influential factors in adult patients aged between 40 to 65. Conversely, an active lifestyle is slightly negatively correlated (-0.03) with cardio, suggesting that active individuals have a slightly lower chance of having cardiovascular disease. The results of this analysis will be useful for developing predictive models in subsequent stages. However, further in-depth analysis and modelling are required to validate these findings and identify the most significant predictors of cardiovascular disease.



Feature Engineering

Once the EDA was concluded, we commenced the feature engineering in the following steps.

1. **High and low blood pressure columns**: Two columns ‘api-hi’ (blood pressure when the heart beats) and ‘api-low’ (blood pressure in between heart beats) in the dataframe were combined to determine if a patient has high blood pressure or low blood pressure. This information was used to create two new columns high blood pressure and low blood pressure.
2. **High glucose**: The column glucose which had categorical data of 1(normal glucose level), 2(above normal glucose) and 3 (highly above normal glucose) was used to create a new dummy column that categorised patients with high glucose level (1) or not (0).
3. **High cholesterol**: The column cholesterol which had categorical data of 1(normal cholesterol level), 2(above normal cholesterol) and 3 (highly above normal cholesterol) was used to create a new dummy column that categorised patients with high cholesterol level (1) or not (0).

4. **High BMI:** The BMI column was used to create a categorical variable to identify patients with high BMI and patients with medically recognized normal BMI. The threshold to determine if a BMI is high or not was retrieved from the centres of disease control and prevention (CDC).
5. **Glucose dummy**: The glucose column was converted to dummy data and added to the dataframe.
6. **Cholesterol dummy**: The cholesterol column was converted to dummy data and added to the dataframe.
7. **Alcohol, smoking and active dummy:** A new column was created to identify patients who drink alcohol, smoke and not active. These set of patients were categorized as 1, while patients who don't drink alcohol, don't smoke and are active were categorized as 0.

Further to the creation of the following columns, we needed to determine the variables that would be our X. With the correlation matrix, we received a numerical representation of the relationship of all variables to cardio variable. However, we need more information to determine which variables to select as our X variable. To gain more insights on the variable relationship with cardio, we did the following.

1. Creation of a pair plot: Because of the size of our dataset and limited computational power, we could not plot the pair plot for all the columns. Therefore, we plotted the pairplot for all numerical variables. This gave us a visual representation of the of how each column relates to each other and a distribution of patients with cardiovascular disease in the relationship.
2. Percentage distribution: To determine how our categorical data relates with the cardio column, we paired each categorical column with cardio column and determined the percentage of patients in the categories who had cardio or not.

With the correlation matrix, pairplot and percentage distribution of categories, we revised the hypothesis above to a new hypothesis.

Old hypothesis: Age, blood pressure (both systolic and diastolic), and cholesterol levels are the most significant factors in determining the presence of cardiovascular disease in individuals.

New hypothesis: Age, height, weight, blood pressure (systolic and diastolic - ap_hi and ap_lo), bmi, high glucose level, high cholesterol level, high bmi, high bp level, low bp level, normal glucose level, abnormal glucose level, very abnormal glucose level, normal cholesterol level, abnormal cholesterol level and very abnormal cholesterol level were significant determinants in determining the presence of cardiovascular disease in individuals. Therefore, we named these variables as our X variables.

Model Development

Upon conclusion of the feature engineering and the creation of our hypothesis, we chose the following four models to build.

1. **Logistic Regression**: We chose the logistic regression model, although we were aware that the model assumes linearity among the variables:
    I. Ease of use: It is an easy model to train and does not require much computational power.
    II. Overfitting: The model reduces overfitting.
    III. Nature of dataset: Our data is a large dataset with skewedness in multiple columns. This resulted in our reason for choosing logistic regression as it is a good model to deal with large, skewed data.

    The first step we took was to build a dummy model which included all the variables. Once we had done this, we reduced the variables in our X value to only variables we

had listed in our hypothesis. Once this was concluded, we separated the variables into X and Y variables into train and test and retrained the model. After this, we tested the trained model and began optimization of the model by calculating the geometric mean of the predicted value. With the use of the geometric mean, we determined a new probability threshold for the predicted value. A new probability threshold was set, and we re-categorised the predicted values with this new threshold.

2. **Random forest**: We chose this model for the following reasons, although we were aware that the model can be relatively slow and require more computational power:
    I. Nature of dataset: Our data is a large dataset with skewedness in multiple columns. This resulted in our reason for choosing random forest as it is a good model to deal with large, skewed data.
    II. Accuracy: The random forest is a generally good model with a good reputation for high accuracy
    III. Overfitting: random forest reduces overfitting that may occur as a result of a lot of columns

    The first step we took was to build a dummy model which included all the variables. Once we had done this, we reduced the variables in our X value to only variables we had listed in our hypothesis. Once this was concluded, we separated the variables into X and Y variables into train and test, tuned the model criterion to 'gini' and retrained the model. After this, we tested the trained model and began optimization of the model by calculating the geometric mean of the predicted value. With the use of the geometric mean, we determined a new probability threshold for the predicted value. A new probability threshold was set, and we re-categorised the predicted values with this new threshold.

3. **K-Nearest Neighbour**: We chose this model for the following reasons, although we were aware that the most efficient number of K can be difficult to determine, and the model can be relatively slow and require more computational power:
    I. No data assumption: It makes no previous assumption about the data set. Therefore, makes it compatible for non-linear relationships.
    II. Versatile model: The model is a versatile model and is very adaptable.
    III. Nature of dataset: Our data is a dataset with skewedness in multiple columns. This resulted in our reason for choosing KNN as it is a good model to deal with skewed data.

    The first step we took was to build a dummy model which included all the variables. Once we had done this, we reduced the variables in our X value to only variables we had listed in our hypothesis. Once this was concluded, we separated the variables into X and Y variables into train and test and retrained the model. After this, we tested the trained model and began optimization of the model by tuning the hyperparameters. We attempted to utilise random search and grid search cross validation to get the best parameters, however, due to limited computational power we could not utilize the random and grid search. Therefore, we manually tuned the parameters of the model and compared the performance of the model after each tuning. Once the best parameters were gotten, the model was retrained with the new parameters and new predictions were made.

4. **Support Vector Machine (SVM)**: This is a commonly used model for binary decision question. However, it could not deal with large number of features, we needed to select fewer features for SVC. The reason we chose this function is:
    I. Binary Decision Focus: This model is explicitly design for solve binary decision question. In addition, it also has a solid mathematics foundation supporting it.

II. Fewer Hyperparameters: Compare to other complex model SVC has less hyperparameters which means it is easier to be tuned, simplifying the optimization process.

The first step is because the SVC model is less efficient in dealing with a large number of features. The logistic regression was used to determine which feature were the most significant feature for the dataset. Second, we split the dataset into training and testing dataset. Then, we input the training dataset to train SVC. After SVC has been trained, we use the test dataset to generate the evaluation metrices we start to be tuning the SVC by using RandomizedSearchCV() to find the best hyperparameters for the SVC model.
Then we use the datasets to train and test the SVC with new hyperparameters to get the F1_score, precision score, accuracy score and recall score. However, the score all increased a little bit except the recall score, recall score has dropped 2%. This situation might case by minimum times of iteration which is due to the limited computational power.

Model Evaluation

We use Accuracy score, F1 score, recall score, precision score and training and prediction time of the model metrics to evaluate the performance of each of our models. We received the following metric for each of our models.

Metrics Selection Rationale:

1. **Precision Score**: This metric represents sure the true positive perdition out of all the positive predictions. It measures the accuracy of the positive prediction of the model. This is a crucial metric for this Cardiovascular dataset. Since the model is predicting if the person has Cardiovascular. A high Precision score model could have patients prevent unnecessary tests or treatment. It will also help hospitals save resources.
2. **Recall Score**: This metric represents the true positive prediction out of the actual positive sample, which measures the ability of the model to correctly identify the positive model. In this case, this metric will be a significant metric. Cardiovascular refers to a group of conditions that affect the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, and stroke. It is one of the leading causes of death worldwide. The best way to mitigate the risk of Cardiovascular is through prevention and early detection. A high Recall Score means the model could effectively predict the actual positive case which helps the patient detect the disease. After it is diagnosed by the doctor, the patient can effectively prevent the risk of Cardiovascular.
3. **Accuracy Score**: This provides a general measure of how well the model is we could use it as a side metric.
4. **F1_Score**: This is the harmonic mean of precision and recall, it balanced the precision and recall, and it is a good metric to evaluate the model if the dataset is unbalanced. According to our EDA, the dataset is balanced our model might not have bias. However, In order to prevent the imbalanced positive and negative samples in the training dataset, we put this as a support metric to mitigate the risk.
5. **Training and Prediction time**: It will not be the main evaluation metric in this case. However, when it comes to the big datasets or real-time prediction it could be a practical consideration.

	Accuracy Score	Precision Score	Recall score	F1_score	Training and prediction time
Logistic Regression	0.723	0.695	0.731	0.713	1.1 seconds
Random Forest	0.688	0.711	0.674	0.68	10.4 seconds
KNN	0.699	0.659	0.71	0.685	37.2 seconds
SVC	0.718	0.662	0.74	0.70	More than 1 hour

*The accuracy score, precision score, recall score, F1_score were estimated to 3 decimal places

SVC has the highest Recall score which means it has a better ability to correctly predict the positive model. This could help individuals predict if they have the potential to have cardiovascular, then distinguish the disease early. However, it is time-consuming to train and use the model to predict if the dataset is too large.

The Logistic regression model shows the highest evaluation score and most optimized time. It has the highest accuracy score, F1 score, second highest recall score, and relatively shorter training and predicting time. Thus, logistic regression is the best model among these models to predict if individuals have Cardiovascular or not.

Conclusions

In our endeavour to predict the risk of cardiovascular disease from a substantial dataset of 70,000 patient records, we embarked on a meticulous journey that encompassed initial data cleaning, in-depth analysis, feature modifications, and model development. Our analysis reaffirmed our hypothesis, highlighting age, blood pressure, cholesterol levels, and BMI as pivotal indicators in predicting the likelihood of cardiovascular diseases.

Addressing our problem statement, we employed four machine learning models: Logistic Regression, Random Forest, K-Nearest Neighbours, and Support Vector Machines. After testing, the Logistic Regression model stood out as the optimal choice. It effectively balanced accuracy with speed, and its impressive recall rate ensures that individuals at potential risk are promptly identified, facilitating early and potentially lifesaving interventions.

To sum up, the early identification and treatment of cardiovascular diseases are crucial. Our study showcases how machine learning can be a powerful tool in the healthcare sector, aiding professionals in identifying risks and making decisions based on data-driven insights. We believe that the methods and findings from this research will not only benefit medical practitioners but will also inspire more such studies, emphasizing the importance of technology in modern healthcare.

Reference

Understanding blood pressure readings. (2023, May 30). www.heart.org. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Zierle-Ghosh, A. (2022, September 11). *Physiology, body mass index*. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK535456/#:~:text=Centers%20for%20Disease%20Control%20and%20Prevention%20(CDC)%3A,to%2040.0%20kg%2Fm%5E2

Centres for disease control and prevention. (2023, June 23). https://www.cdc.gov/obesity/basics/adult-defining.html#:~:text=Adult%20Body%20Mass%20Index&text=If%20your%20BMI%20is%20less,falls%20within%20the%20obesity%20range.