EM – 623 Data Science and Knowledge Discovery

Final Project

Determining Risk Factors for Cardiovascular Heart Disease

Submitted by :

Sanhita Bhagwat

Sampada Chavan

Hrudaya Reddy

Rishi Deepak Savla

Guided by: Feng Liu

Table Of Content

Introduction:

Cardiovascular disease (CVD) is a significant health concern that affects millions of people worldwide. Various demographic elements, lifestyle choices, and biological indicators can influence an individual's risk of developing heart disease. Identifying these risk factors and understanding their relationships can help in developing effective preventive measures to reduce the incidence and impact of CVD.

The aim of this project is to determine how particular demographic elements, healthy lifestyle choices, and biological indicators influence the development of heart disease. We will focus on three key risk factors: smoking, high blood pressure, and high cholesterol, which are known to be significant contributors to the development of CVD.

The expected outcome of this project is to identify patterns of risk factors and their relationships with cardiovascular disease. By analyzing these patterns, we hope to develop personalized interventions that target individual risk factors, leading to improved preventive measures and better management of CVD.

To achieve this objective, we will utilize modern machine learning techniques to analyze a dataset containing information on various demographics, lifestyle, and biological indicators. We will use logistic regression and decision tree analysis to identify the most significant risk factors and their relationships with CVD. Additionally, we will employ clustering techniques to group individuals with similar risk factor profiles and explore potential differences in their CVD risk.

This project's importance lies in its potential to improve our understanding of the complex interplay between various risk factors and their impact on the development of CVD. By identifying patterns and relationships, we can develop targeted interventions that focus on addressing individual risk factors, leading to improved preventive measures and better management of the disease.

In the following sections, we will describe the data sources and methods used, present our findings, and discuss the implications of our results for future research and clinical practice.

Motivation:

Cardiovascular disease (CVD) is a significant global health issue, leading to high rates of morbidity and mortality. It is essential to gain a deeper understanding of the risk factors that contribute to the development of CVD.

The objective of this project is to explore potential relationships between various risk factors and the development of CVD by analyzing a large dataset using modern machine learning techniques.

The research potential of this project is significant. By analyzing the dataset, we can gain valuable insights into the relationships between risk factors and CVD, contributing to improved understanding, early detection, and the development of effective preventive measures.

The implications of this project's findings are also significant. They can contribute to the design of personalized interventions that target individual risk factors, improving preventive measures and disease management. Furthermore, the insights gained from this project can guide future research in this area, leading to improved understanding of CVD risk factors and the development of more effective treatments and diagnostic tools. Ultimately, this project has the potential to improve patient outcomes and reduce the overall burden of CVD globally.

Data source:

This dataset was taken from Kaggle. The Kaggle dataset collection is one of the most extensive public collections of datasets available, containing over 30,000 datasets in various formats. The datasets cover a broad range of topics, from health, finance, and economics to sports, politics, and climate change.

Kaggle datasets offer numerous benefits to data scientists, researchers, and analysts. They provide access to real-world data that is often difficult to obtain, allowing users to develop and test hypotheses, build models, and draw conclusions. Kaggle datasets are also well-documented, ensuring that users have access to the information they need to understand the data and perform their analyses effectively.

Dataset used:

We used heart_data.csv from Kaggle. This dataset contains detailed information on the risk factors for cardiovascular disease. It includes information on age, gender, height, weight, blood pressure values, cholesterol levels, glucose levels, smoking habits, and alcohol consumption of over 70 thousand individuals. Additionally, it outlines if the person is active or not and if he or she has any cardiovascular diseases. This dataset provides a great resource for researchers to apply modern machine learning techniques to explore the potential relations between risk factors and cardiovascular disease that can ultimately lead to improved understanding of this serious health issue and design better preventive measures.

Column name	Description
age	Age of the individual. (Integer)
gender	Gender of the individual. (String)
height	Height of the individual in centimeters. (Integer)
weight	Weight of the individual in kilograms. (Integer)
ap_hi	Systolic blood pressure reading. (Integer)
ap_lo	Diastolic blood pressure reading. (Integer)
cholesterol	Cholesterol level of the individual. (Integer)
gluc	Glucose level of the individual. (Integer)
smoke	Smoking status of the individual. (Boolean)

Column name	Description
alco	Alcohol consumption status of the individual. (Boolean)
active	Physical activity level of the individual. (Boolean)
cardio	Presence or absence of cardiovascular disease. (Boolean)

Data Cleaning:

Age Transformation:

To improve interpretability, the age variable was converted from days to years by dividing it by 365.This transformation allows for easier understanding of the age data, as it represents the age of individuals in years rather than days.

Duplicate Removal:

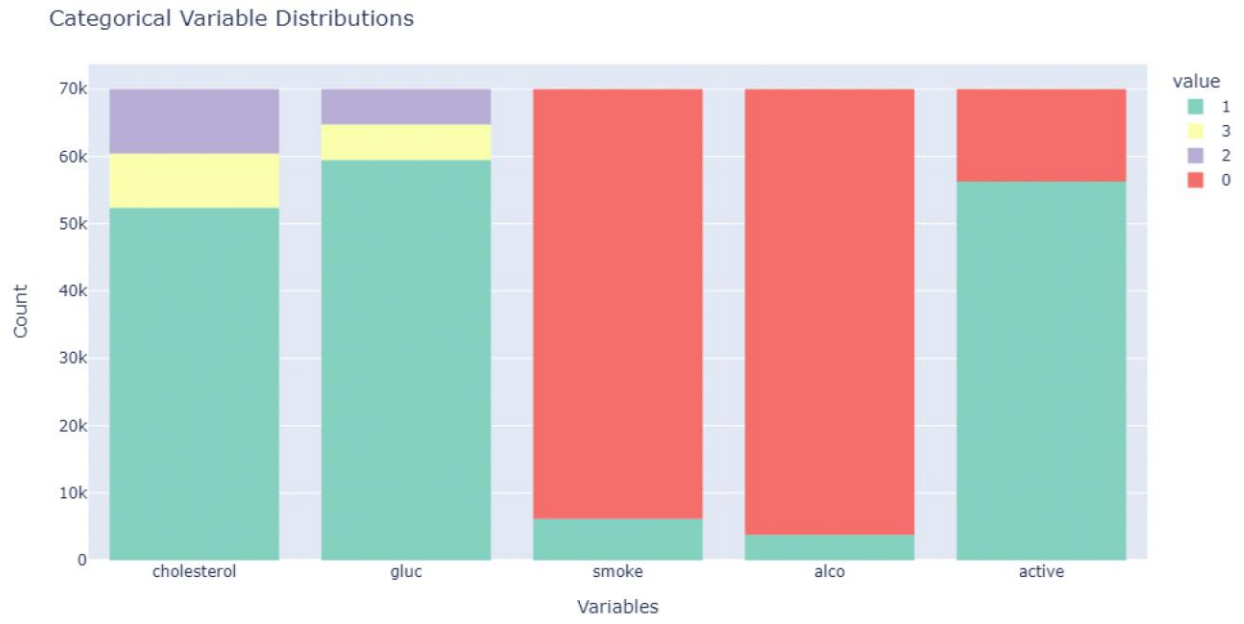
To ensure data integrity, any duplicate entries were removed. Duplicate entries can introduce biases or inaccuracies in the analysis, as they may artificially inflate the representation of certain individuals or introduce redundant information. By eliminating duplicates, the dataset is streamlined and each entry is unique, reducing the risk of skewed results.

Missing Value Handling:

To maintain the validity of the analysis, rows with missing values were dropped from the dataset. Missing values can create gaps in the data and hinder accurate analysis, as they may lead to biased or incomplete conclusions. By removing rows with missing values, the dataset becomes more robust and suitable for analysis, as it contains complete information for the variables of interest.

This way the data was brought to 69976 rows.

Data Visualization:



We are considering what can be the most important factors that can affect CVD. Here the graph gives us a better understanding of the data. The quantity of data is too much, so presenting it visually will help us to understand and interpret in an easy way.
From the above graph we can observe the following things:

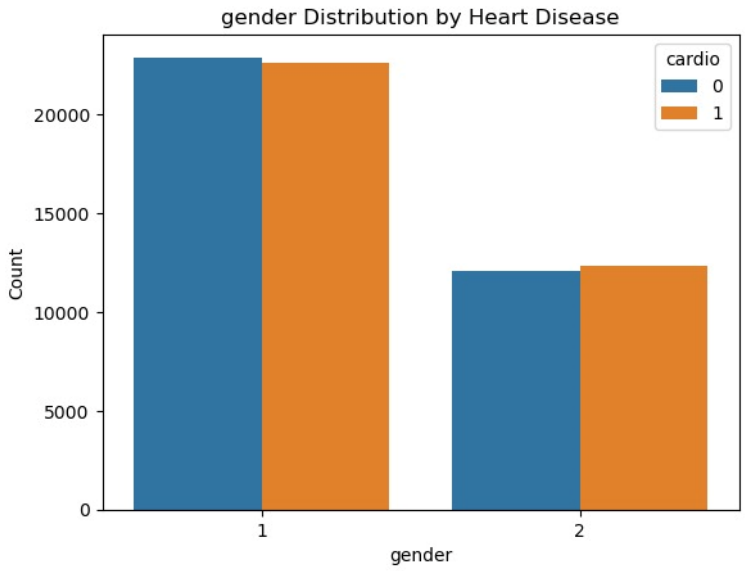
Cholesterol: Cholesterol levels are divided into three parts where 1 represents Normal Cholesterol Level, 2 represents Above Normal levels and 3 represents way above Normal level. Most of the people from the data have Normal levels of cholesterol, where as there are still significant number of people in part 2 and part 3.

Glucose: Glucose levels are divided into three parts where 1 represents Normal Glucose Level, 2 represents Above Normal levels and 3 represents way above Normal level. Again, we can observe that most of the population from the data have normal level of glucose level, there are still many people in the part 2 and part 3 but they are less than that were for Cholesterol Levels.

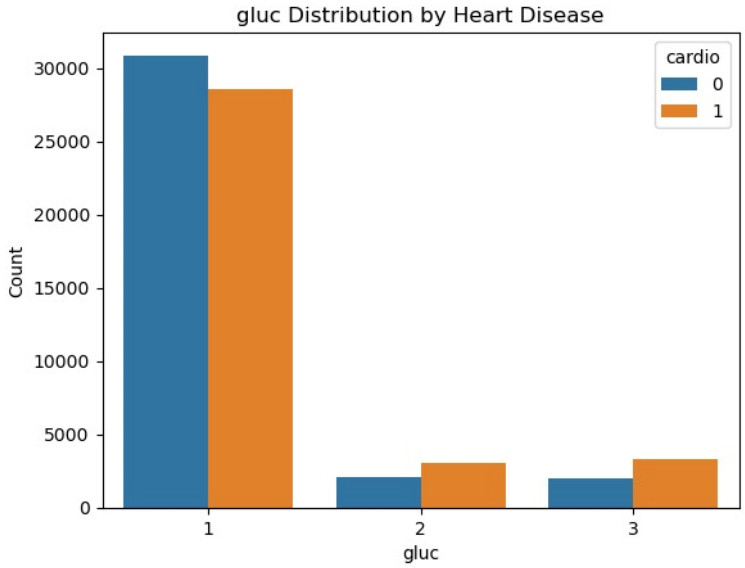
Smoking: Most of the graph is orange in color which represents the value 0, that is, the people there do not smoke and the people under the green section are smokers. Most of the population from the data is non-smoking.

Alcohol: the trend is similar to that of Smoking, where most of the population have normal or no alcohol levels.

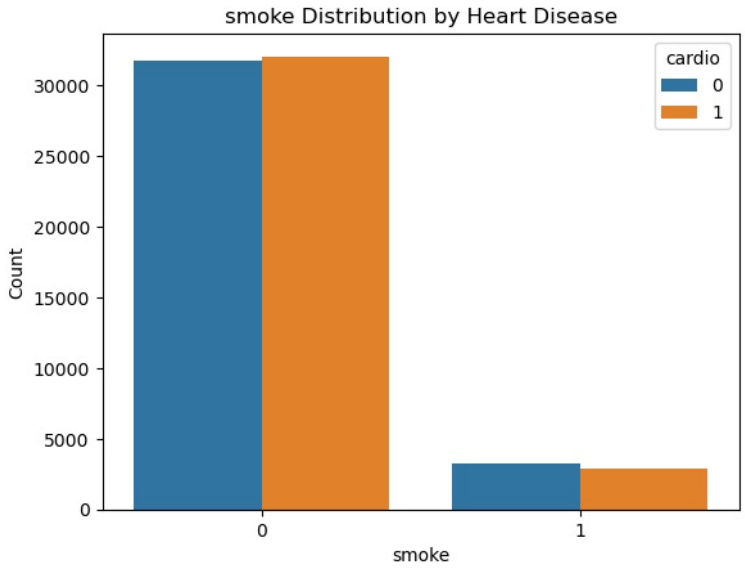
Physical Activity: The people in the green section are the ones that do some kind of physical activity and the others do not. Doing physical activity reduces the chances of having CVD.



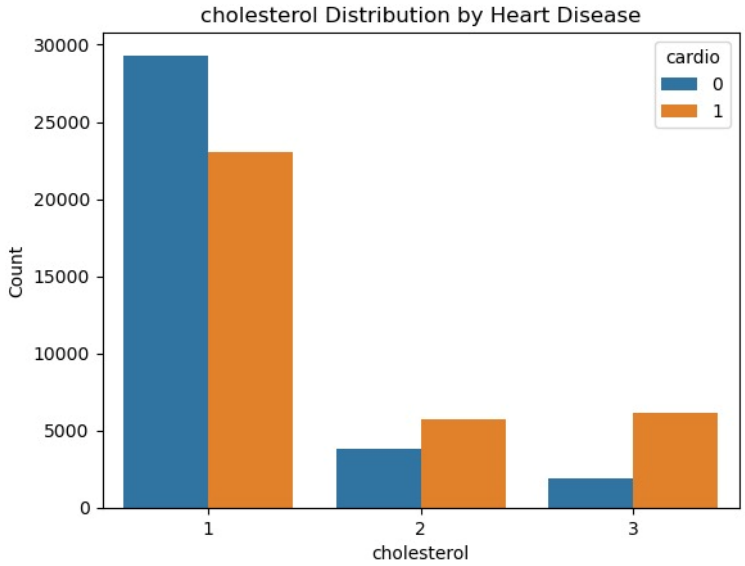
The graph shows distribution of CVD by gender. Form the graph above, it is observed that Women and Men both are equally distributed, that means the number of women or men having CVD is almost similar to the ones that don't have them. This indicates that gender might not be a strong determining factor for CVD.



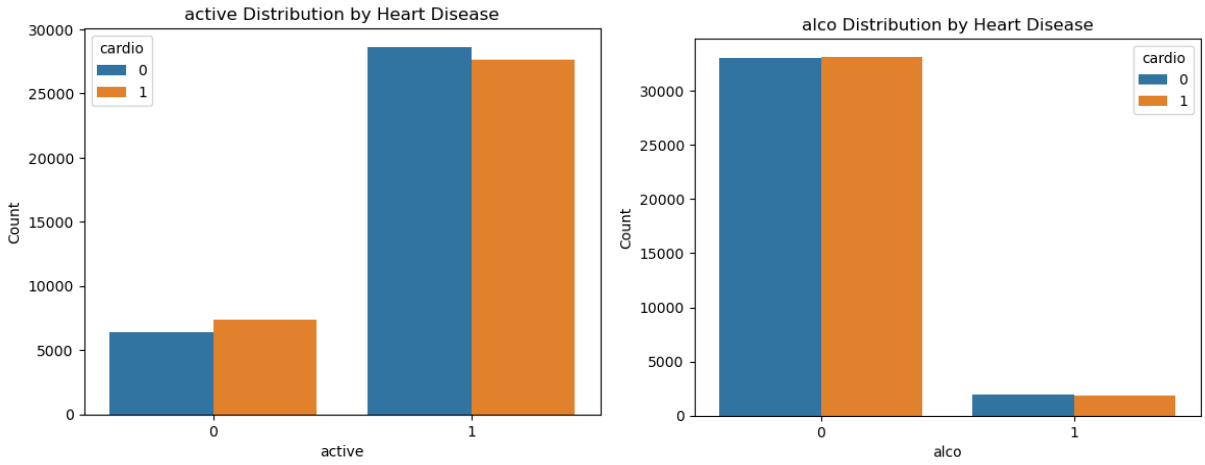
The graph shows the distribution of CVD by different amount of glucose levels. It is observed that people with higher glucose levels tend to have CVDs as compared to people with lower glucose levels.



This graph shows the distribution of CVD with respect to smoking. It is observed that the people that do not smoke are almost equally divided, blue represents the ones that do not have CVDs and orange are the ones that have CVDs. Again, for smoking people, they are almost divided equally, so we can say that smoking is not one of the major causes of CVD.

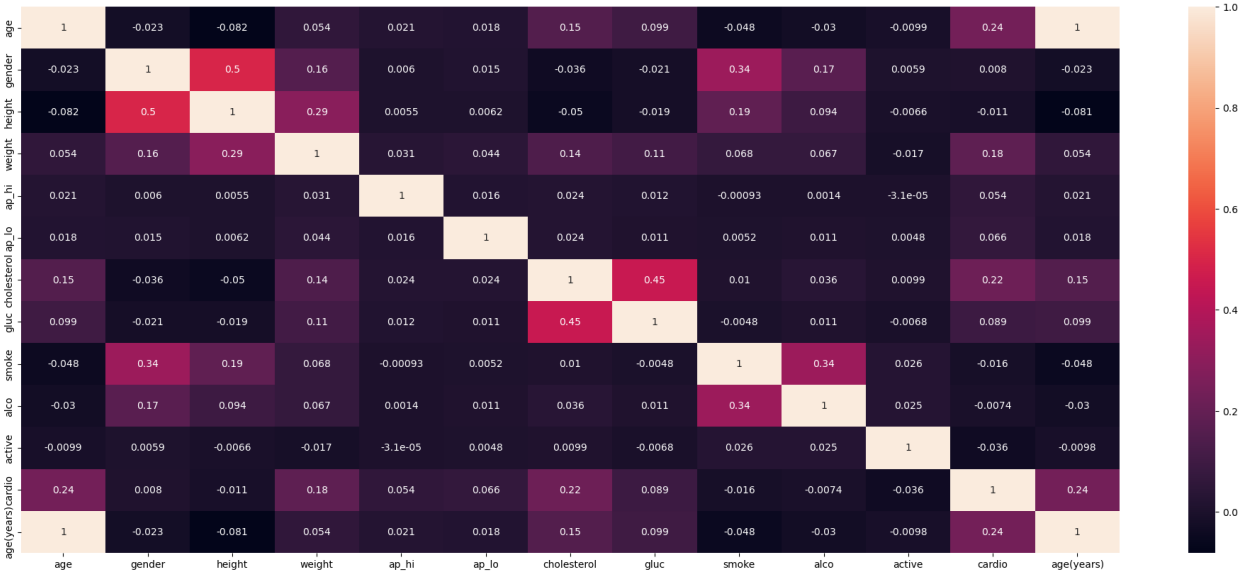


This graph shows the distribution of CVD among people with different cholesterol level. The people in the normal cholesterol section tend to have a less chance of having CVD. For Above normal and Way Above Normal cholesterol levels, it s clearly evident that having higher levels means the chances of having CVDs is way higher. Thus, it can be interpreted that cholesterol level is one of the major factor affecting CVDs.



The above two graphs show the distribution of CVD with respect to Physical Activity and Smoking. It is seen that the graphs are evenly distributed, so they don't come under major factors affecting CVDs.

A heat map is now plotted to better understand these causes



From the above heath map, we can see that the darker shades in the cardio column are the variables that affect more and are a major cause for having CVD. It is observed that Age, Cholesterol Level and Weight are three of the major factors, followed by Systolic Blood Pressure(ap_hi), Diastolic Blood Pressure(ap_lo) and glucose levels.

To better understand how weight affects Cardio health we have considered BMI (Body Mass Index). Body Mass Index (BMI) is a measure that relates a person's weight to their height and is commonly used to assess body composition and weight-related health risks. By calculating BMI, we can obtain a numerical value that provides an indication of body fatness and potential health implications associated with weight.

The formula to calculate BMI is as follows:

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

High BMI values, particularly those indicating overweight or obesity, are linked to an increased risk of developing cardiovascular conditions. By incorporating BMI into the analysis of risk factors for CVD, researchers and healthcare professionals can gain a more comprehensive understanding of the relationship between body composition and cardiovascular health.

Machine learning models Used:

We have used four models to better understand our data:

Decision Tree Classifier:

- It extracts the predictors (age, BMI, glucose level, cholesterol, systolic blood pressure, diastolic blood pressure) and the response variable (cardiovascular disease) from the dataset.
- The dataset is split into training and test sets using the train_test_split function.
- A decision tree classifier is created with a maximum depth of 3.
- The decision tree model is trained on the training data using the fit method.
- The model predicts the response variable for both the training and test data.
- The goodness of fit of the model is evaluated by calculating the classification accuracy for both the training and test datasets.
- The confusion matrix is plotted to visualize the model's performance on the training and test data.
- The decision tree itself is plotted using the plot_tree function, displaying the feature importance of each predictor.

Random Forest Classifier:

- It performs similar steps as the decision tree classifier, but this time using a random forest classifier.
- The predictors and response variable are defined, and the dataset is split into training and test sets.
- The random forest classifier is created and trained on the training data.
- The model predicts the response variable for both the training and test data.
- The classification accuracy and confusion matrix are calculated and plotted to evaluate the model's performance.

K-means Clustering:

- The predictors and response variable are defined, and the dataset is split into training and test sets.
- A K-means clustering model is created with 2 clusters.
- The model is trained on the training data.
- The model predicts the response variable for both the training and test data.
- The goodness of fit of the model is evaluated by calculating the classification accuracy for both the training and test datasets.

K-nearest Neighbors (KNN) Classifier:

- The K-nearest neighbors' classifier is created.
- The classifier is trained on the training data.
- The model predicts the response variable for the test data.

Out of all the four models, Decision tree classifier gives out the highest accuracy for test dataset at around 72.33% followed by Forest Random Classifier, which drops to almost 58%, we can see that the testing set accuracy is less than the training set accuracy because of overfitting and the other two are even lower than that.

```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.7265919743912198

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.7233908768720704
```

Accuracy for Decision Tree Classifier

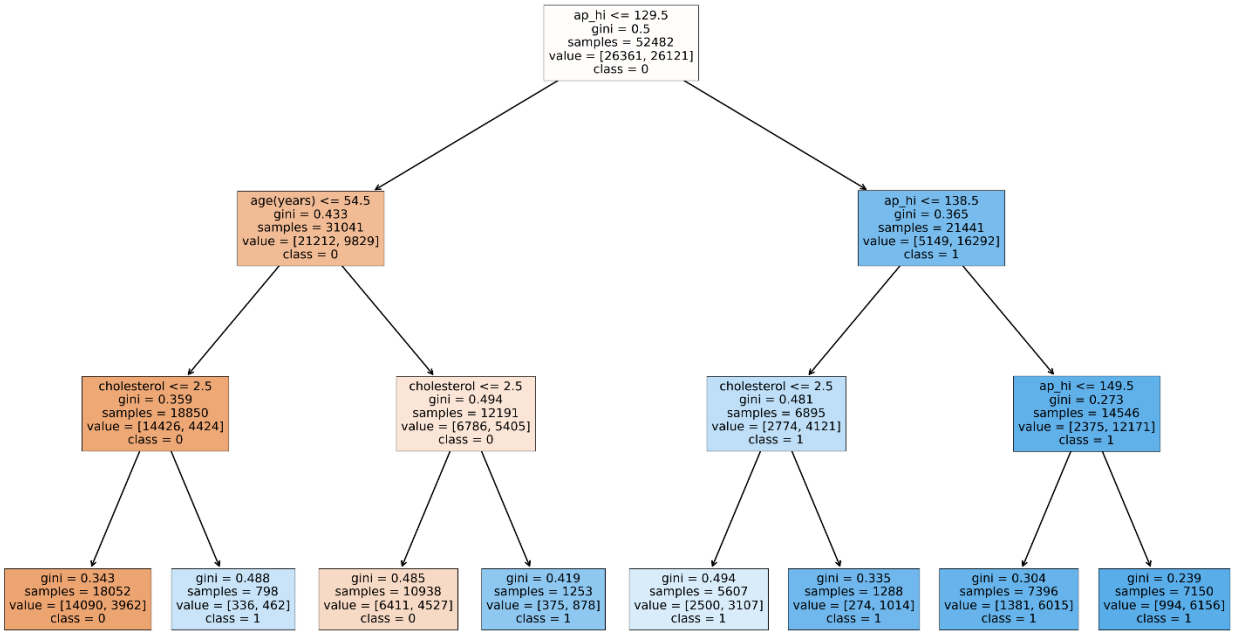
```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.6695057352997218

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.6342746084371784
```

Accuracy for Forest Tree Classifier

Decision Tree Classifier in Detail:

The Decision Tree

.

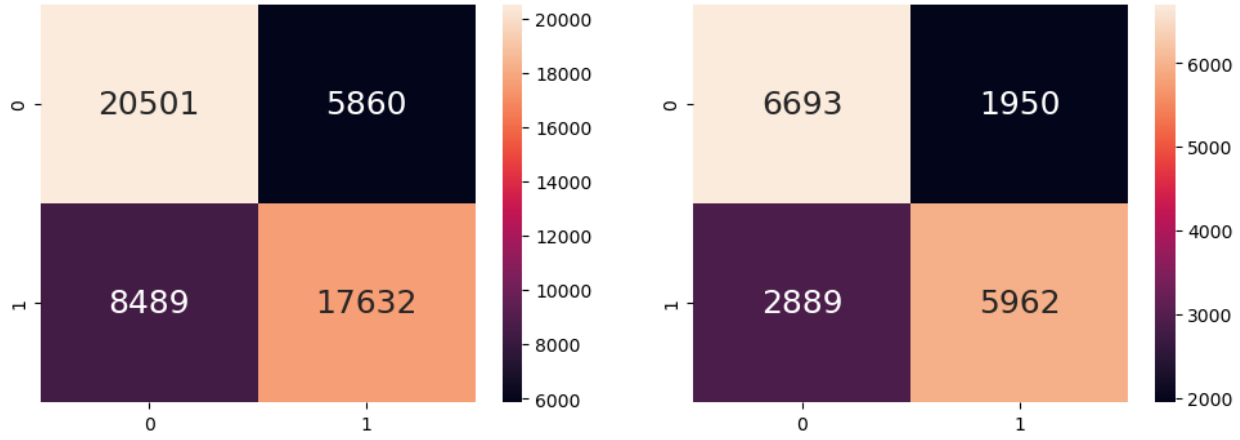


A decision tree is a graphical representation of a predictive model that bases choices or predictions on input information using a tree-like structure. It is made up of nodes and edges, where each node denotes a quality or characteristic and each edge a choice or result.

The decision tree displays the model's decision-making process from the root node, which is the highest node, all the way down to the leaf nodes, which are the bottommost nodes. The model follows the corresponding edge at each internal node depending on a decision that is made based on a particular feature or characteristic. The ultimate prediction or classification is represented by a leaf node, which is reached at the end of this procedure.

The decision tree provides a visual representation of the model's decision-making process, making it simpler to read and comprehend how the model generates its predictions. The information at each node in the tree includes the feature or attribute that is being taken into account, the decision rule or condition connected to that node, and any following branches or edges that connect that node to other nodes.

The Confusion Matrix



The elements of the confusion matrix represent the following:

True Positive (6693): The model correctly predicted CVD present when it was actually present.

True Negative (5962): The model correctly predicted CVD not present when it was actually not present.

False Positive (2889): The model incorrectly predicted CVD present when it was actually not present.

False Negative (1950): The model incorrectly predicted CVD not present when it was actually present.

After that we determined which numeric value is most important in determining CVD and we found out that Systolic Blood pressure, age and cholesterol were identified as the most important variables for this analysis.

```
ap_hi 0.8290399274670255
age(years) 0.09659152592208009
cholesterol 0.07436854661089443
ap_lo 0.0
gluc 0.0
Bmi 0.0
```

The value for Systolic Blood Pressure is the highest, therefore it is the most determining reason for CVD.

Conclusion:

The analysis aimed to identify the most important factors that can affect cardiovascular disease (CVD) using machine learning models and data visualization techniques. Here is a detailed conclusion based on the findings:

Factors Affecting CVD based on graph analysis of the data:

- Cholesterol Levels: The analysis revealed that higher cholesterol levels are associated with an increased risk of CVD. People with above-normal or way above-normal cholesterol levels showed a higher likelihood of having CVD.
- Glucose Levels: Elevated glucose levels were also found to be a factor influencing CVD. People with higher glucose levels had a higher prevalence of CVD compared to those with normal levels.
- Physical Activity: The analysis showed that engaging in physical activity has a positive impact on reducing the chances of having CVD. People who were involved in some form of physical activity had a lower risk of CVD.
- Systolic Blood Pressure (ap_hi): Higher systolic blood pressure was identified as an important variable affecting CVD. Increased blood pressure levels were associated with an increased likelihood of CVD.
- Age: Age was found to be a significant factor in CVD. As age increased, the risk of having CVD also increased.
- Gender: The analysis indicated that gender may not be a strong determining factor for CVD, as both men and women had a similar distribution of CVD cases.
- Smoking: The impact of smoking on CVD was relatively balanced, with a similar distribution of CVD cases observed among smokers and non-smokers. Thus, smoking may not be a major contributing factor for CVD in this dataset.

Machine Learning Model Performance:

- Decision Tree Classifier: The decision tree classifier demonstrated the highest accuracy among the models used, with an accuracy rate of approximately 72.33% for the test dataset. It provided valuable insights into the decision-making process and identified important features affecting CVD.
- Random Forest Classifier, K-means Clustering, and K-nearest Neighbors (KNN) Classifier: These models also provided insights but exhibited lower accuracy rates compared to the decision tree classifier.

Conclusion on Key Factors:

- Based on the analysis, age, cholesterol level, and systolic blood pressure were identified as the most important factors affecting the likelihood of CVD. These factors can be considered critical in assessing the risk of developing CVD.
- Higher cholesterol levels and elevated systolic blood pressure increase the risk of CVD, while engaging in physical activity helps in reducing the chances of CVD.
- Gender and smoking were not significant factors affecting CVD in this dataset, suggesting that other variables may play a more prominent role in CVD development.

Practical Implications:

- The findings can be used to inform preventive measures and lifestyle modifications aimed at reducing the risk of CVD. Emphasizing the importance of maintaining healthy cholesterol levels, managing blood pressure, and promoting physical activity can help in preventing CVD.
- Healthcare interventions can be targeted towards individuals with higher cholesterol levels, elevated blood pressure, and advancing age to effectively manage and reduce the risk of CVD.

Further Research:

- Further research and analysis can be conducted to explore additional factors that may influence CVD, such as genetic factors, dietary habits, and other lifestyle variables.
- Evaluating the impact of various treatments and interventions on reducing CVD risk can provide valuable insights for healthcare professionals and policymakers.

In conclusion, this analysis provides valuable insights into the important factors affecting CVD and can be used to guide public health strategies, promote healthy lifestyle choices, and improve cardiovascular health outcomes.

Summary

In the Data Science and Knowledge Discovery course, we focused on three major objectives. Our first objective was to provide us with a solid foundation in data mining and pattern recognition tasks and techniques. We learned various methods for processing data, explored the basics of classification, and delved into techniques such as Bayesian Decision Theory, Linear Discriminant Functions, Neural Networks Models, Support Vector Machines, Boosting and Bagging Methods, Variable Selection, Clustering Techniques, and Deep Learning Methods.

Our second objective was to develop our skills in reading and critically evaluating research papers in the field of data mining and analytics. We learned how to analyse and assess the validity and significance of research findings, enabling us to stay updated with the latest advancements in the field.

Our third objective focused on equipping us with practical implementation skills. We learned how to implement and utilize important data mining and pattern recognition models and algorithms. The emphasis was on solving interdisciplinary problems and making data-driven decisions in engineering and sciences. By gaining hands-on experience with implementing these techniques, we were prepared to apply them to real-world scenarios and address complex data-driven challenges.

Throughout the course, we engaged in theoretical and practical exercises, allowing us to understand the principles underlying data mining and machine learning techniques and gain practical experience through hands-on applications. By combining theory, research evaluation, and practical implementation, the course aimed to provide us with a comprehensive understanding of data science and knowledge discovery.

In the future, there are several additional topics and areas that could be explored in the Data Science and Knowledge Discovery course to further enhance our knowledge and skills:

Applications in the Real World and Case Studies: Including real-world case studies and hands-on projects will help us better grasp how data science and knowledge discovery approaches are used in various fields and sectors. These initiatives might entail using real datasets, tackling challenging issues, and presenting conclusions to mimic real-world situations.

Ethical Concerns and Data Privacy: It is crucial for data scientists to handle ethical issues and data privacy concerns. Discussions on ethical data usage, algorithmic fairness, bias reduction, privacy-preserving approaches, and laws like the General Data Protection Regulation (GDPR) and other pertinent regulations may be covered in later rounds of the course.

References

https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseases

https://ai.plainenglish.io/gradient-boosting-classifier-ml-model-in-python-1acedbc6cf5e

https://dataanalyticsbook.info/chapter-5.-learning-i-cross-validation-oob.html

https://data.world/kudem