The US Cardiovascular Diseases Lifestyle Risk Prediction Research

**1. Situation understanding**

**1.1 Identify the objectives of the situation**

a. background information

Cardiovascular Disease (CVD) refers to a group of diseases that affect the functioning of the heart and blood vessels, including coronary heart disease, heart attacks, strokes, high blood pressure, and heart failure. For many years, CVD has been a significant global public health concern, significantly impacting overall health and well-being.

Statistics indicate that CVD is the leading cause of death in the United States, resulting in 874,613 deaths in the United States in 2019(Connie W. Tsao et al., 2022). Despite advancements in medical technology and healthcare services that have helped reduce CVD-related mortality rates, its incidence continues to rise (Connie W. Tsao et al., 2022). This disease poses significant challenges to public health, exerting a profound impact on health, economics, and society.

Preventing CVD has become an urgent task, and early or preemptive interventions are crucial measures in reducing the risk of CVD. Among the modifiable risk factors contributing to the pathogenesis of CVD, lifestyle factors play a prominent role. In recent years, urbanization and modernization have led to significant changes in people's lifestyles, fostering unhealthy habits that contribute to the high prevalence of cardiovascular diseases.

Therefore, understanding the relationship between lifestyle and cardiovascular diseases and exploring how improving lifestyle can prevent and manage CVD have become critical topics in the field of public health.

b. objectives and success criteria

This research aims to investigate the association between cardiovascular diseases and lifestyle, seeking key connections between lifestyle choices and cardiovascular health through the collection and analysis of large sample data. It is hoped that this study will provide scientific evidence for the development of targeted public health interventions and personalized health recommendations, aiming to reduce the incidence and mortality rates of cardiovascular diseases and enhance people's quality of life and overall health.

* The objectives of the CVD Lifestyle Risk Prediction Research are:

1. To assess the impact of different lifestyle choices on the likelihood of developing CVD.

2. To predict the risk of CVD based on lifestyle factors.

* The success criteria of the CVD Lifestyle Risk Prediction Research are:

1. To provide actionable insights and recommendations for individuals to mitigate their CVD risk through lifestyle modifications.

2. To contribute valuable information for public health initiatives aimed at reducing the burden of CVD through targeted interventions and awareness campaigns.

3. To enhance the understanding of the relationship between lifestyle choices and CVD risk, enabling informed decision-making by healthcare professionals and policymakers.

**1.2 Assess the situation**

1.2.1 resource inventory

a. Hardware Resources:

Computer: High-performance computer or server capable of handling large-scale data analysis and model training.

Storage Devices: Sufficiently large hard drive space to store and manage a vast amount of data.

Database Server: Used to store and manage data in the database.

b. Data Sources:

Centers for Disease Control and Prevention (CDC): The CDC provides rich cardiovascular disease statistics and epidemiological information.

Medical Databases: Such as the National Health and Nutrition Examination Survey (NHANES), Medical Expenditure Panel Survey (MEPS), American Heart Association (AHA), etc.

Population Statistics Data: Demographic information, such as data from the U.S. Census.

Socioeconomic Data: For instance, employment and income data from the U.S. Bureau of Labor Statistics (BLS).

c. Knowledge Storage:

Literature Databases: Academic literature databases (e.g., PubMed, Google Scholar) can be used to search for relevant published research papers.

Data Storage and Backup: Ensuring data is securely stored and backed up to prevent loss of critical data.

d. Human Resources:

Data Scientists/Analysts: Responsible for data cleaning, analysis, and modeling.

Programmers: Responsible for developing and maintaining data analysis code and tools.

Subject Matter Experts: Familiar with cardiovascular diseases and lifestyle to provide insights and guidance on research direction.

Data Administrators: Responsible for data management and maintenance.

Project Managers: Coordinate team members, manage project progress, and allocate resources.

But currently, all the work is done by me. I will take all the job in this project.

1.2.2 Requirements, Assumptions, and Constraints

a. Requirements:

I need to ensure legal access and use of data, comply with data privacy and protection laws, and avoid disclosing sensitive information. The research results will be submitted to the University of Auckland through the Canvas student management system in the form of an assignment. The submission will include the report itself, raw data, and the data manipulation process used in the project.

b. Assumptions:

During the data collection process, I need to use free and open data sources to avoid incurring any costs. Regarding the data found, I will assume that its quality is reliable and accurate and it can be used to support my research objectives. As the project manager, my goal is to obtain the final conclusions through a robust model. If the model itself is flawed, I won't be able to arrive at the correct conclusions. To achieve accurate and meaningful results, a sound model is essential, but ultimately, the final conclusions are what matter the most.

c. Constraints

I need to ensure that I have access to the necessary data by having the required login credentials, such as using my student account credentials to access academic papers in the school library. For the data itself, I may also need to set up an opening password. However, at present, the data I have does not contain any personally sensitive information, so I do not need to set up a password for it. Additionally, the data I have is free and publicly available, and I will only be submitting it through the school's assignment submission system without uploading it to any other websites or platforms. Therefore, I will comply with all legal, ethical, and privacy constraints related to data usage and research.

1.2.3 Risks and Contingencies

a. Plan Risk:

Risk Description: The project may take longer than expected, leading to delays in progress.

Contingency Plan: Conduct a thorough evaluation of the project's complexity and time planning, regularly check the progress, and if necessary, adjust the research methods or scope to ensure the project is completed on time.

b. Data Risk:

Risk Description: Poor data quality or narrow scope may affect the credibility and effectiveness of research results.

Contingency Plan: Perform comprehensive data quality assessment and preprocessing before data collection, consider the reliability of data sources, increase the diversity of data sources, or reevaluate the research questions to adapt to the available data.

c. Result Risk:

Risk Description: The initial results may not meet expectations, potentially hindering the achievement of research objectives.

Contingency Plan: Promptly review and analyze the results, try different data analysis methods or models, seek advice from mentors or domain experts, reevaluate the research questions, and ensure meaningful results are obtained.

* 1. **Determine data mining objectives**

1. Data Mining Problem Type: Predictive Problem

Objective: To predict whether or not the given person may have cardiovascular diseases based on different lifestyle choices.

1. Technical Goal:

To find which lifestyle or lifestyle combination may be associated with the development of cardiovascular diseases.

1. Specific Numerical Goal:

To provide a relatively accurate risk assessment for the individual's likelihood of developing cardiovascular diseases, for example, predicting the probability of the individual developing cardiovascular diseases if the person continues his/her lifestyle.

**1.4 project plan**

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1: Iteration 1 Project Plan | | | |
| **Phase** | **Time** | **Job** | **Risks** |
| Business Understanding | 8.21-8.22 | Topic choosing | Hard to decide which topic to choose from. |
| Data Understanding | 8.23-8.27 | Search data | Hard to find free open data, or the data is irrelevant to research. |
| Data preparation | 8.28-9.1 | Clean data, preproduce data | Data is not enough after clean, or some of the data is not included. |
| Data transformation | 9.2 | reduction and projection | The use of software is not skilled and takes a long time |
| Data-mining method(s) selection | 8.3-9.6 | Method selection from IPSS current ones | It does not match the objective and needs to redo. Take longer time than expected. |
| Data-mining algorithm(s) selection | 9.7-9.10 | algorithm selection from IPSS's current ones | It does not match the objective and needs to redo. Take longer time than expected. |
| Data Mining | 9.11-9.14 | Search for pattern | Hard to find a pattern |
| Interpretation | 9.15-9.17 | Visualize, interpret, evaluate, and iterate | Take longer than expected |
| Report Evaluation | 9.18-9.22 | Report writing and evaluation | Report Template |

**2. Data understanding**

**2.1 Collect initial data**

The initial data for my research was obtained from the Kaggle.com website.

Initially, I tried to obtain the data from the official website of the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS). However, I found that the data on the website is in ASCII and SAS format which are not the best format to process.

Considering the substantial size of this dataset, I'd appreciate accessing the raw data in Excel or CSV format. These formats are compatible with a majority of data analysis tools, which streamlines the ensuing data analysis. Moreover, they offer a more direct and intuitive insight into the data's content. Consequently, I've sought alternative sources via various channels.

Eventually, I discovered the same data source onKaggle.com in Excel format. The website is <https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset>.

Consequently, I downloaded this preprocessed data from Kaggle.com to my local computer, facilitating my further analysis and operations.

**2.2 Describe the data**

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.

The dataset I got is based on the BRFSS Dataset in 2021. The dataset are 2 Excel documents. The part 1 Excel document is 12.4 MB, with 19 columns and 151471 rows. The part 2 document is 12.8 MB with 19 columns and 157383 rows. These 19 columns represent the content of lifestyle-related factors. They are General Health, Checkup, Exercise, Heart Disease, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Sex, Age Category, Height(cm), Weight(kg), BMI, Smoking History, Alcohol Consumption, Fruit Consumption, Green Vegetables Consumption, and Fried Potato Consumption. The meaning of each columns is shown below:

* General Health is an ordinal data which shows the sample’s general health condition. The categories include Excellent, Very Good, Good, Fair and Poor.
* Checkup is a nominal data which shows sample’s medical checkup frequency. The checkup includes never, 5 or more years ago, within the past 5 years, within the past 2 years, and within the past year.
* Exercise is a flag data which shows the sample doing workout regularly or not.
* Heart Disease is a flag data which shows the sample currently have heart disease or not.
* Skin Cancer is a flag data which shows the sample currently have skin cancer or not.
* Other Cancer is a flag data which shows the sample currently have other cancer or not.
* Depression is a flag data which shows the sample currently have depression or not.
* Diabetes is a flag data which shows the sample currently have diabetes or not.
* Arthritis is a flag data which shows the sample currently have Arthritis or not.
* Sex is a flag data which shows the sample is male or female.
* Age Category is a nominal data which shows the sample’s age category. The categories include 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, and 80+.
* Height(cm) is the continuous data shows the sample’s height in cm.
* Weight(kg) is the continuous data shows the sample’s weight in kg.
* BMI is the continuous data shows the sample’s BMI number.
* Smoking History is a flag data which shows the sample have smoking history or not.
* Alcohol Consumption shows the sample’s alcohol consumption level.
* Fruit Consumption shows the sample’s fruit consumption level.
* Green Vegetables Consumption shows the sample’s green vegetable consumption level.
* Fried Potato Consumption shows the sample’s fried potato consumption level.

The dataset is in tabular format. In iteration 2, the dataset is divided into 7 columns of continuous data, 9 columns of flag data, 2 column of nominal data, and 1 columns of ordinal data. But in iteration 3, as I am using python to process data, I have divided data into 2 categories: Numerical Variables and Categorial Variables. As shown below, the data type **float** and **integer** are treated as **Numerical Variables** and the **object** type data are treated as **Categorial Variables**.

电脑屏幕的照片上有文字

中度可信度描述已自动生成

*Figure 1: Infomation of the data*

电脑萤幕画面

描述已自动生成

*Figure 2: Dataset table screenshot part 1*

电脑萤幕画面

描述已自动生成

*Figure 3: Dataset table screenshot part 2*

The above screenshot is of my data table, displaying what each column in the table represents and the numbers shown in each row across the various columns. The column I have outlined in red is my target, indicating whether or not the person has heart disease.

**2.3 Explore the data**

In iteration 2, I've narrowed my focus solely on my hypothesis that Gender, Age, BMI, and Alcohol consumption might be the four key factors influencing the onset of heart disease. But in iteration 3, I plan to conduct a more in-depth data exploration by incorporating all columns into the analysis from the perspective of the two data types I've identified: Numerical Variables and Categorical Variables.

文本

描述已自动生成

*Figure 4: Definition of Numerical and Categorical Variables*

2.3.1 Numerical Variables Exploration

文本

描述已自动生成

*Figure 5: Coding in Numerical Data Visualization*

图表, 直方图

描述已自动生成

*Figure 6: Distribution of Height*

Height(cm): The height of the patients seems to follow a normal distribution, with the majority of patients having heights around 160 to 180 cm.

图表, 直方图

描述已自动生成 *Figure 7: Distribution of Weight*

Weight(kg): The weight of the patients also appears to be normally distributed, with most patients weighing between approximately 60 and 100 kg.

图表, 直方图

描述已自动生成

*Figure 8: Distribution of BMI*

BMI: The distribution of Body Mass Index is somewhat right-skewed. A large number of patients have a BMI between 20 and 30, which falls within the normal to overweight range. However, there are also a significant number of patients with a BMI in the obese range (>30).

图表, 直方图

描述已自动生成

*Figure 9: Distribution of AlcoholConsumption*

AlcoholConsumption: This feature is heavily right-skewed. Most patients have low alcohol consumption, but there are a few patients with high consumption.

图表

描述已自动生成 *Figure 10: Distribution of FruitConsumption*

FruitConsumption: This feature is also right-skewed. A lot of patients consume fruits regularly, but a significant number consume them less frequently.

图表, 直方图

描述已自动生成

*Figure 11: Distribution of GreenVegetablesConsumption*

GreenVegetablesConsumption: This feature appears to be normally distributed, with most patients consuming green vegetables moderately.

图表, 直方图

描述已自动生成 *Figure 12: Distribution of FriedPotatoConsumption*

FriedPotatoConsumption: This feature is right-skewed. Many patients consume fried potatoes less frequently, while a few consume them more often.

2.3.2 Categorical Variables Exploration

文本

描述已自动生成

*Figure 13: Coding in Categorical Data Visualization*

图表, 条形图

描述已自动生成

*Figure 14: Count of GeneralHealth*

GeneralHealth: Most patients describe their general health as "Good", with "Very Good" being the second most common response. Fewer patients rate their health as "Fair" or "Poor".

图表, 瀑布图

描述已自动生成

*Figure 15: Count of Checkup*

Checkup: The majority of patients had a checkup within the past year. Fewer patients had their last checkup 2 years ago or more than 5 years ago.

图表, 条形图

描述已自动生成

*Figure 16: Count of Exercise*

Exercise: More patients reported that they exercise compared to those who do not.图表, 条形图

描述已自动生成 *Figure 17: Count of GeneralHealth*

Heart\_Disease: Target variable. A significant majority of patients do not have heart disease. Only a small proportion of patients have heart disease.图表, 条形图

描述已自动生成

*Figure 18: Count of SkinCaner*

SkinCancer: The vast majority of patients do not have skin cancer.图表, 条形图

描述已自动生成

*Figure 19: Count of OtherCancer*

OtherCancer: Similar to skin cancer, most patients do not have other forms of cancer.

图表, 条形图

描述已自动生成

*Figure 20: Count of Depression*

Depression: Most patients do not suffer from depression. However, a non-trivial number of patients do report having depression.图表, 瀑布图

描述已自动生成

*Figure 21: Count of Diabetes*

Diabetes: Similar to the disease-related features above, most patients do not have diabetes. However, a small proportion do have diabetes.图表, 条形图

描述已自动生成

*Figure 22: Count of Arthritis*

Arthritis: Most patients do not have arthritis, but a significant number do.图表

描述已自动生成

*Figure 23: Count of Sex*

Sex: There are slightly more female patients than male patients in the dataset.图表, 条形图

描述已自动生成

*Figure 24: Count of AgeCategory*

AgeCategory: The dataset includes patients from a wide range of age categories. The 50-54 age category has the most patients, followed by the 55-59 and 60-64 categories.图表, 条形图

描述已自动生成

*Figure 25: Count of SmokingHistory*

SmokingHistory: The majority of patients do not have a history of smoking.

**2.4 Verify the data quality**

2.4.1 Null value

In the diagram shown below, the majority of the data in my database is complete, but there are still some issues with errors and missing values.

文本

描述已自动生成

*Figure 26: Data Null Amount of 19 fields*

In the diagram above, most of the fields have some empty or null values. The null values vary from 1 to 3.

2.4.2 Outliers

图表, 箱线图

描述已自动生成 *Figure 27: Data Box Plot of Numerical Variables*

The boxplots indicate that there are some potential outliers in our numerical data. These potential outliers and extreme values should be further investigated to determine their validity and possible impact on the analysis.

2.4.3 Duplicate data

文本

描述已自动生成*Figure 28: Data Duplicate Count*

As shown above, the number of duplicates is 52. These duplicate data need to be removed.

I will proceed to clean and enhance the data in the following steps.

**3. Data Preparation**

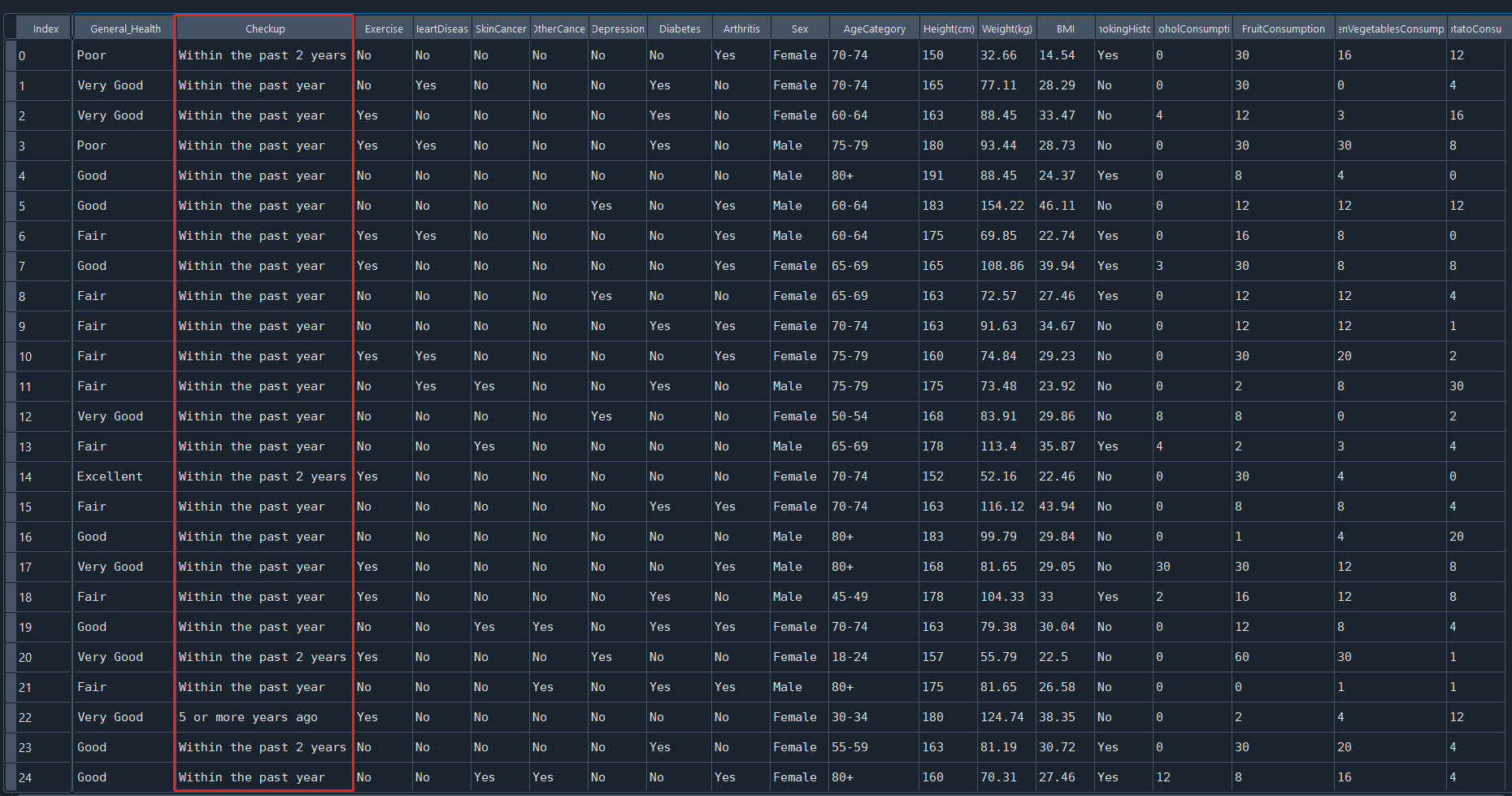
**3.1 Select the data**

Now, I have gained a relatively superficial understanding of my data. Next, I need to build on this by selecting from the existing raw data. I will make the selection based on the following factors.

1. Select data by business objectives

My business objectives are “to assess the impact of different lifestyle choices on the likelihood of developing CVD” and “to predict the risk of CVD based on lifestyle factors”.

In the raw data, some columns' content is irrelevant to lifestyle or cannot provide effective predictions and suggestions. The "Checkup" column records(figure29 below) the last time a physical examination was conducted, reflecting whether the subject undergoes regular checkups and the frequency of those checkups.



*Figure 29: The checkup column values in the dataset*

This factor does not align with the causes of heart disease. Patients can diagnose certain illnesses through physical examinations, but the act of having regular checkups itself has no direct correlation with acquiring diseases.

Furthermore, we often see that people who have regular physical examinations are more likely to be diagnosed with illnesses. It's not that those who don't have regular checkups are less prone to illness, but rather they find it difficult to diagnose diseases because they don't get checked. In other words, those who do not have regular checkups often remain unaware if they contract certain illnesses, and only when severe complications appear or the disease progresses to an unignorable stage do they seek medical attention and receive a diagnosis. Therefore, the data in this field need to be excluded

1. Select data by data quality
2. Null value

In section 2.4.1, it indicates that some columns have null data values, and I need to impute the missing values for these null entries.

1. Unclassified value

图表, 条形图

描述已自动生成

*Figure 30: The Count of AgeCategory*

图表, 瀑布图

描述已自动生成

*Figure 31: The Count of Diabetes*

Figure 30 displays the count of AgeCategory. There is a bar with value in 4 which is the unclassified value. In Figure 31, the third and fourth bar values need to be modified to "no" or "Yes”. Because the details are not needed in this analysis.

1. Duplicate value

In section 2.4.3 I found that there are some duplicate values in the table. These data also need to be imputed.

1. Outliers and extremes

In Section 2.4.2, I identified the presence of outliers in the table, which also require imputation. Specifically, I employed summary statistics to obtain insights and facilitate outlier detection, as illustrated in the subsequent figure. I then proceeded with steps to eliminate these outliers.

电脑萤幕画面

描述已自动生成

*Figure 32: The Statistical Summary of Numerical Data*

**3.2 Clean the data**

As I mentioned in Section 3.1 "Select the data," I need to remove the columns that are unrelated to my business objectives.

Additionally, I need to modify or remove any invalid data.

1. unrelated columns

电脑萤幕画面

描述已自动生成

*Figure 33: Drop the field that is irrelevant to business objectives*

The graph displayed above illustrates the code I employ to exclude the "Checkup" field, which needs to be removed as they are unrelated to the business objectives. By implementing the “drop” code, I can eliminate the “Checkup” data. The red circle highlights a change of removing a column.

1. Null value data

文本

描述已自动生成

*Figure 34: Code to remove null value*

As shown in the graph above, I execute “dropna” method to remove the null value in the dataset. The graph shown below shows the null value count results before and after the code is executed. The null value has been removed.

文本

描述已自动生成

*Figure 35: The null value count before remove*

文本

描述已自动生成

*Figure 36: The null value count before remove*

1. Unclassified value

文本

描述已自动生成

*Figure 37: Code to remove unclassified value in AgeCetegory*

图表, 条形图

描述已自动生成

*Figure 38:* *AgeCetegory Column before Reclassify*

图表, 条形图

描述已自动生成

*Figure 39: AgeCetegory Column after Reclassify*

The three charts above depict the process I undertook to handle the "AgeCategory" field. Using the code in Figure 37, I retained data in the AgeCategory that didn't have a value of 4. I then verified the results through data visualization. From Figures 38 and 39, it's evident that data with a value of 4 has been successfully removed. This confirms the success of this phase of my data cleansing.电脑游戏的屏幕

描述已自动生成

*Figure 40: Diabetes field before and after Reclassify*

In the "Diabetes" column, a binary outcome is required. As such, values labeled as "No, pre-diabetes or borderline diabetes" need to be modified to "No", and values "Yes, but female told only during pregnancy" should be changed to "Yes". By executing the code within the blue box shown in the above figure, I replaced the respective values in the diabetes column with "Yes" and "No". The successful outcome of this operation is demonstrated in the red box in the same figure.

1. Duplicate value

文本

描述已自动生成

*Figure 41: Code to remove Duplicate value*

文本

描述已自动生成

*Figure 42: Duplicate Value Count before and after Remove*

Figure 41 contains the code I used to remove duplicate data. Figure 42 shows that before removing duplicates, the count of duplicate values was 73. After the cleanup, the count was reduced to 0. This confirms the success of my data cleansing for this stage.

1. Outliers and extremes

文本

描述已自动生成

*Figure 43: Code of Outlier Define and Remove Outliers*

The image above shows the code is designed to remove outliers from the numerical features of my dataset. Specifically, for each numerical feature, it calculates the interquartile range (IQR) between the first (Q1) and third quartiles (Q3). Values falling below Q1−1.5×IQR or above Q3+1.5×IQR are considered outliers. The choice of 1.5 times the IQR is based on a common statistical rule, as values outside this range are deemed to be significantly distant from the central tendency of the data, thus classifying them as outliers. The dataset is then updated to exclude these outliers, and summary statistics are generated for the cleaned data.

电脑萤幕画面

描述已自动生成

*Figure 44: Summary* *stats after remove outliers*

From the image above, we can see that after processing, the value of the min and max have been changed. However, the handling of outliers and extremes can generate new ones. I will need to check in subsequent iterations whether I need to deal with outliers and extremes again or only focus on specific data.

**3.3 Construct the data**

In my data, there is a field for alcohol consumption. From the illustrative image circled in red below, we can see that a substantial portion of people have a data value of 0, indicating that they do not drink alcohol. Among those who do drink, there are significant differences in the amount consumed. Therefore, I want to categorize this lifestyle habit into two classes: drinking and not drinking. In subsequent steps, I can further subdivide among those who drink to see how the amount of alcohol consumed affects heart disease. As a result, I need to add a column for whether or not they drink alcohol. So, I generated a column named “DrinkingAlcohol” which shown in the graphs below.

图表

描述已自动生成

*Figure 45: Histogram plot circled Zero alcohol consumption*



*Figure 46: Code of Constructing drinking alcohol field*

电脑的屏幕截图

描述已自动生成

*Figure 47: the new generated field in the df variable*

**3.4 Integrate various data sources**

Due to the large volume of data, I have my 308,854 records stored in two different Excel spreadsheets. What I need to do now is merge these two files into one. The Excel spreadsheet for dataset part 1 contains 151,471 records, and the Excel spreadsheet for dataset part 2 contains 157,383 records.

图形用户界面, 文本, 徽标

描述已自动生成

*Figure 48: Code to append 2 excel data together*

图表, 树状图

描述已自动生成

*Figure 49: Data row numbers after integrating*

By executing the Figure 48 code, I have appended the two parts of the Excel file. As shown in the image above, the “df” data count is 308,854, which is the total data count.

**3.5** **Format the data as required**

In my previous section, I have already cleaned and processed the data. In iteration 2, I can use a type node to check whether my data is of the desired data type that I wish to obtain. But in order to use Python to obtain analysis I need to check the model I want to use.

In iteration 2, I found that I can use the logistic regression model. So, I need to change the data from object type to numerical data type.

电脑屏幕的截图

描述已自动生成

*Figure 50: Code of formatting data*

The code I used transforms all non-numeric columns into numeric formats. Besides converting 'yes' and 'no' values to 1 and 0 respectively, I also encoded the gender column, where males are represented as 0 and females as 1. Additionally, I numerically encoded other categorical data based on their distinct content. The age categories were ordered from youngest to oldest and labeled from 0 to 12. The 'GeneralHealth' column was encoded based on health levels, ranging from 1 to 5. After these numeric conversions, the dataset (as the red circle shown below) consists entirely of numeric types.

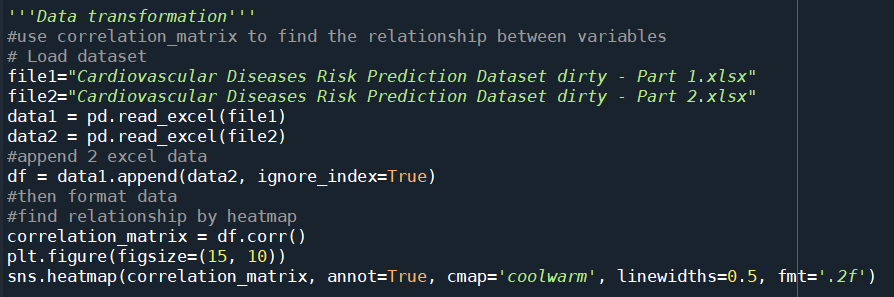
电脑屏幕的照片上有文字

描述已自动生成

*Figure 51: Data description after formatting*

**4. Data transformation**

**4.1 Reduce the data**

To uncover the relationships among the data, I employed the raw, uncleaned data to create a correlation matrix that describes the relationships between variables. The relevant code is as follows.

*Figure 52: Code to implement correlation matrix*

From the aforementioned code, I obtained the heatmap below. The reason for using the raw, unprocessed data is that the vast majority of my dataset consists of valid entries that are capable of establishing relationships. If I were to clean the data, such as by removing outliers, the results could vary based on my chosen boundaries. To minimize potential distortions caused by biases in my data cleaning approach, I opted to use the raw data for the correlation matrix analysis.

图表

描述已自动生成

*Figure 53: Heatmap of the dataset*

Based on the heatmap generated, I observed that 'Height', 'FruitConsumption', 'GreenVegetableConsumption', and 'FriedPotatoConsumption' have minimal correlation with 'HeartDisease' (values very close to 0). Consequently, these four columns can be removed.

文本

描述已自动生成

*Figure 54: Code to Remove less important fields*

**4.2 Project the data**

Currently, my target field is heart disease. The data is imbalanced at the moment. As shown in the image below. When I execute the code in the left, I can see that the number of "0" far exceeds the number of "1" in the data shown in the right side of the figure. I need to manipulate this data to make it essentially balanced.

屏幕上有字

描述已自动生成

*Figure 55: Current target field distribution*

文本

描述已自动生成

*Figure 56: Code of undersampling*

Given the substantial size of my dataset, I chose the undersampling approach to balance the data. The code is illustrated above. After applying undersampling, the counts for both 0 and 1 in my target column became consistent, each with 15,733 entries, as depicted in the following figure.

文本

描述已自动生成

*Figure 57: Comparison of the value count of Heart Disease before and after undersampling*

I have now adjusted the data, so I need to check again whether there is any column of insufficient importance to be removed. As shown in the image below, the heatmap result shows that all the current columns are essential, and I don't need to take any action.

图表

描述已自动生成*Figure 58: Checking fields’ importance by Heatmap*

**5. Data-mining method(s) selection**

**5.1 Match and discuss the objectives of data mining (1.1) to data mining methods**

5.1.1 Data mining methods introduction

There are primary data mining methods include **Classification, Regression, Clustering, Summarization, Dependency Modeling, and Change and Deviation Detection**(Fayyad et al., 1996).

Understanding various data mining techniques and aligning them with specific objectives is pivotal to ensuring our analysis methods resonate with the anticipated output. Within the context of predicting cardiovascular disease risks, each technique offers unique strengths and limitations.

Firstly, I delved into the classification method. Classification is akin to tagging an item, pinpointing which category it fits into. When I aim to predict an individual's predisposition to cardiovascular diseases, the classification method could tag them as "high risk" or "low risk" based on their lifestyle factors. However, the caveat with this method is that it chiefly provides a label without conveying the precise probability of risk.

Subsequently, I contemplated the regression method. Regression zeroes in on discovering relationships between variables to forecast a particular numeric value. For my purpose, it could estimate the probability of cardiovascular risk rooted in lifestyle factors. While it offers a precise numeric prediction, it might not adequately expose all potential risk factors.

Clustering was another method I considered. It revolves around grouping data based on certain similarities, but it wouldn't directly offer me a risk prediction. Nonetheless, by identifying which lifestyle combinations are most prevalent, it could shed light on certain insights.

Summarization enables me to extract core patterns or trends from a voluminous dataset. Although it offers a macro perspective, it may not be tailored for individualized risk predictions.

I also looked into dependency modeling, a technique focusing on relationships and interactions within the data. This could reveal which lifestyle factors are most correlated with elevated cardiovascular risks, adding depth to my predictions.

Lastly, the change and deviation detection method would offer me real-time alerts when data deviates from previous patterns. This is instrumental for tracking abrupt lifestyle changes and potential risk surges, but it leans more towards monitoring rather than predicting.

In conclusion, each data mining technique possesses its applicability and constraints. The crux lies in strategically combining these methods to furnish me with the most comprehensive and accurate assessment of cardiovascular disease risks.

**5.2 Select the appropriate data-mining method(s) based on discussion**

Predicting the onset of cardiovascular diseases based on diverse lifestyle choices is a complex endeavor. The interplay of data ranging from dietary habits to medical conditions like diabetes provides a mosaic of information. Deciphering this to forecast health outcomes demands a multifaceted approach. In this exploration, we delve into three pivotal modeling methodologies that not only offer predictive insights but also unravel the intricate relationship between lifestyle and heart health.

When confronted with the crucial task of predicting the risk of cardiovascular diseases based on different lifestyle choices, my initial inclination is to deploy the **classification** method. Now, I am holding a plethora of data on individuals, from the presence or absence of diabetes to their fruit intake. The challenge lies in determining, based on these data points, whether they are likely to suffer from cardiovascular diseases in the foreseeable future or not. This is where classification thrives. Through this approach, I can assign a distinct predictive label for each individual, enabling me to provide tailored advice and preventive measures for these individuals.

In iteration 2, I opted for the logistic regression algorithm from regression methods. However, after further study, I believe this algorithm should be categorized under classification methods. This is because the application scenarios and outputs of logistic regression align more closely with classification. Therefore, in this iteration, I will not be using regression approaches.

Yet, grounding all these predictions on a robust theoretical foundation is imperative. This calls for a profound comprehension of the genuine relationship between lifestyle choices and cardiovascular risk. And this is where **dependency** modeling comes into play. Through this modeling method, I can elucidate which factors bear the strongest correlation with cardiovascular risk, or if a particular factor amplifies the risk only under specific conditions. For instance, I might discover that a high sugar intake correlates with an increased risk of cardiovascular diseases only in conjunction with certain lifestyle combinations. Such profound insights not only fortify the foundation of my predictions but also empower me to dispense more targeted health advisories to the public.

The intricate dance of predicting cardiovascular risks based on lifestyle hinges on a multifaceted approach. Through the nuanced application of classification, regression, and dependency modeling, we achieve a holistic view that ranges from broad risk categorizations to precise probability estimations. Such a comprehensive approach ensures that individuals are not just informed of potential risks, but they are also empowered with detailed knowledge, leading to informed health decisions and targeted interventions.

**6.** **Data-mining algorithm(s) selection**

**6.1 Conduct exploratory analysis and discuss**

After analyzing the preliminary data mining methods, I've tailored my approach to two essential techniques pertinent to my research goals: Classification and Dependency Modeling. The crucial next step involves choosing data mining algorithms. This will, in turn, refine the accuracy and reliability of cardiovascular disease risk predictions. In iteration 2, due to software constraints, only a subset of algorithms can be utilized. In this iteration, I aim to explore other possibilities.

I'll first determine the training and test data percentages. In this step, I use the classic 80/20 method. The code is shown below.

文本

描述已自动生成

*Figure 59: Code of split train and test data*

1. Classification:

Firstly, I selected several models suitable for classification, as shown below. The reason I chose these particular algorithms for my classification models dictionary is twofold. On one hand, due to my current hardware constraints, my personal computer isn't capable of handling the extensive computations required by large and complex models. On the other hand, these algorithms are among the classic and commonly used methods for classification.文本

描述已自动生成

*Figure 60: Model dictionary definition*

This code shown below first trains each model and then evaluates them using X\_test. Performance metrics as per your specifications are computed for each model, and the results are displayed at the end.

文本

描述已自动生成

*Figure 61: Model raining and Evaluation*

In the code above, the following evaluation metrics were used to assess the performance of classification models:

* Accuracy:

Accuracy represents the proportion of predictions that the model gets right. It is the most direct metric in classification tasks. While useful when dealing with a relatively balanced dataset, it can be misleading if one class significantly outweighs the other. For instance, in a dataset with 95% negatives, a model that always predicts negative can achieve 95% accuracy, but such a model is not useful.

* Precision:

Precision denotes out of all the samples predicted as positive, how many actually are positive. In other words, when a model predicts a sample as positive, how often is it correct? It's a measure of the number of correct positive predictions made.

* Recall:

Recall indicates out of all the actual positives, how many were predicted as positive by our model. It is concerned with the model's capability to capture all the positive instances.

* F1-Score:

The F1-score is the harmonic mean of precision and recall. It provides a single, comprehensive metric especially useful when you need to balance between precision and recall. It's particularly handy when there's an imbalance between classes.

* ROC AUC:

The ROC AUC (Area Under the Receiver Operating Characteristic Curve) represents a measure of a model's ability to distinguish between positive and negative classes. An AUC of 1 implies a perfect classifier, while an AUC of 0.5 implies a random guess. It is especially useful when evaluating a model's performance across various threshold values.

* Confusion Matrix:

The confusion matrix is a table that lays out the performance of a model in terms of true and false positives and negatives. It offers a detailed view into the number of true positives, true negatives, false positives, and false negatives, providing a comprehensive overview of model performance.

文本

描述已自动生成 文本

描述已自动生成 文本

描述已自动生成 文本

描述已自动生成

*Figure 62: Results of Models*

As we can see above figure, the results of the classification models are shown. From the given results of various classification models on my dataset, I can make the following observations:

**Logistic Regression:**

With an accuracy of 76.34% and an F1-Score of 0.7705, this model performs reasonably well. The AUC of 0.8433 suggests that this model can distinguish between the positive and negative classes effectively.

The confusion matrix indicates that there are more false positives (857) than false negatives (632), but the model achieves a good balance between precision and recall.

**Decision Tree:**

This model has the lowest accuracy among all the models at 66.02%.

The F1-Score is also relatively low at 0.6604. The equal AUC and accuracy values suggest that the model's performance is average without any standout qualities.

**Random Forest:**

This ensemble method offers an improvement over the decision tree with an accuracy of 75.45%.

The AUC of 0.8305 is promising, and the F1-Score of 0.7624 indicates a balanced performance between precision and recall.

**SVM (RBF Kernel):**

The SVM model has a lower precision but a high recall, which means while it captures a large number of positive cases, it also misclassifies a significant number of negatives as positives.

The AUC of 0.8239 suggests decent discriminatory power.

**KNN:**

KNN's performance is comparable to the decision tree, with an accuracy of 66.67%. The AUC, however, is higher at 0.7173.

**Naive Bayes (Gaussian):**

The model shows balanced results in terms of precision and recall, as reflected in the F1-Score of 0.7273.

The AUC of 0.8049 indicates good discriminatory power.

AdaBoost:

AdaBoost achieves an accuracy similar to logistic regression and a competitive AUC of 0.8414.

**Gradient Boosting:**

Among all models, Gradient Boosting achieves the highest accuracy at 76.99%.

The AUC of 0.8434 is one of the top, and the F1-Score of 0.7793 suggests a good balance between precision and recall.

**XGBoost:**

XGBoost performs slightly less effectively than Gradient Boosting but still maintains an accuracy of 75.01% and an AUC of 0.8268.

**LightGBM:**

LightGBM's performance is quite similar to logistic regression, with a slightly better F1-Score of 0.7735. The AUC of 0.8401 is competitive.

1. Dependency Modeling:

In iteration 2, due to the model quantity constraints of the SPSS software, I had to choose from a limited set of model algorithms. Upon further study, I realized that the Apriori algorithm is typically not suitable for binary classification problems. This is because its primary goal is to discover frequently co-occurring item combinations rather than predicting a specific target variable. Unless the data is in the form of transactional or other sequential datasets and there's an interest in uncovering patterns or association rules, this algorithm might not be appropriate.

In this iteration, I opted for three commonly used algorithms for dependency analysis: LASSO Regression (L1 regularization), Recursive Feature Elimination (RFE), and Permutation Feature Importance. For a straightforward model evaluation, I employed logistic regression to discern the performance of these models. Here's the corresponding code.

文本

描述已自动生成

*Figure 63: Code of Dependency Models*

文本

描述已自动生成

*Figure 64: Evaluation Results of Dependency Models*

The graph above shows the evaluating results of the dependency models.

* LASSO Feature Selection:

Accuracy (76.34%): The LASSO model accurately predicted 76.34% of the samples. This is a fairly high accuracy rate, indicating that the LASSO selection method yielded meaningful features for the classification task.

Precision (74.47%): Of all instances predicted as the positive class, 74.47% actually belonged to the positive class. This shows that the LASSO model has a decent specificity.

Recall (79.82%): The model identified 79.82% of all actual positive instances. This means the model is sensitive to the positive class.

F1-Score (77.05%): This metric considers both Precision and Recall to give a balance between the two. An F1-score of 77.05% signifies that the model maintains a good trade-off between Precision and Recall.

ROC AUC (76.36%): The ROC AUC measures the model's ability to distinguish between the classes. An AUC of 76.36% indicates the model's strong discriminative power.

* RFE Feature Selection:

Accuracy (71.73%): The accuracy is lower than that of LASSO, suggesting that RFE might have eliminated some features that could have been useful for classification.

Precision (71.70%) and Recall (71.36%): Both these metrics are lower than the LASSO model, suggesting that the RFE model is both less specific and less sensitive.

F1-Score (71.53%): Correspondingly, the F1-score, which is the harmonic mean of Precision and Recall, is also lower than that of the LASSO model.

ROC AUC (71.73%): This confirms that the RFE model has a weaker discriminative ability compared to LASSO.

* Permutation Feature Importance Selection:

Accuracy (76.07%): The accuracy is very close to the LASSO model, suggesting that Permutation Feature Importance did a good job in retaining important features.

Precision (74.64%), Recall (78.64%): These metrics are also close to the LASSO model, indicating that the Permutation Feature Importance model has good specificity and sensitivity.

F1-Score (76.59%): The F1-score is also comparable to the LASSO model.

ROC AUC (76.08%): This shows that the model's discriminative power is almost at par with the LASSO model.

**6.2 Select data-mining algorithms based on discussion**

In step 6.1, I have obtained the evaluation results of my target models by using some evaluation criteria.

In the previous context where multiple **classification** algorithms were tested, the **Logistic Regression** model delivered superior performance with an accuracy of 76.34%, precision of 74.47%, recall of 79.82%, F1-Score of 77.05%, and ROC AUC of 76.36%. These metrics indicate that the model not only correctly classifies a high percentage of instances but also maintains a good balance between precision (minimizing false positives) and recall (minimizing false negatives). The ROC AUC value, which evaluates the model's ability to distinguish between the two classes, is also commendably high.

Compared to other algorithms like Decision Trees, SVM, KNN, etc., Logistic Regression demonstrated more consistent and balanced results across all evaluation metrics. Furthermore, Logistic Regression models are computationally efficient, easily interpretable, and require less tuning, making them an excellent choice for many classification tasks.

For **dependency** model algorithm selection,, upon detailed analysis in step 6.1, the **LASSO feature selection** seems to be the most appropriate for this dataset as it consistently performs the best across all evaluation metrics. Its inherent ability to shrink coefficients and effectively eliminate features that do not contribute to the predictive power might be the reason for its superior performance. While RFE lags in performance, Permutation Feature Importance comes close, but doesn't surpass LASSO. Given these insights, the LASSO model would be the recommended choice for this classification task.

To better harness the characteristics of the data, I considered integrating both strategies by adjusting the settings of Logistic Regression, specifically employing L1-regularized logistic regression. In the ‘sklearn’ library, this is effortlessly achieved by setting the penalty parameter of LogisticRegression to 'l1'. This approach not only retains the classification capabilities of logistic regression but also incorporates the feature selection benefits introduced by the L1 regularization in Lasso regression. Through this method, we can not only achieve accurate classification predictions but also ensure the model focuses solely on the truly impactful features, enhancing the model's generalization and interpretability.

**6.3 Build/Select appropriate model(s) and choose relevant parameter(s)**

Based on step 6.2, I have already chosen Logistic Regression. This step involves building and running a model based on my selection.

Firstly, I partitioned the prepared data into two categories: one being the training set, and the other the testing set, as shown in the figure below. I allocated 70% of the data for training and designated 30% as the testing set.



*Figure 65: Coding of Splitting dataset*

文本

描述已自动生成

*Figure 66: Coding of Modelling*

Afterward, I built the model.

* In Model Initialization:

logistic\_model is set with L1 regularization (penalty='l1').

The solver saga is specified because it can handle L1 penalty.

max\_iter=5000 specifies the maximum number of iterations for the solvers to converge. Setting max\_iter to a relatively large value ensures that the solver has enough iterations to converge towards an optimal solution without prematurely stopping.

random\_state=42 ensures reproducibility. Using the same random state means the random processes inside the model training will be the same every time it's run.

* In Model Training:

The .fit() method trains the model using the training data X\_train and y\_train.

* In Making Predictions:

After training, we use the .predict() method on the test data (X\_test) to get the predicted outcomes for the model.

* In Evaluation Function:

This is a helper function to print out the performance metrics for the logistic regression model. It takes the model's name, the true labels, and the predicted labels as input arguments.

Inside this function, various metrics are calculated:

Accuracy measures the percentage of correct predictions.

Precision calculates the ratio of correctly predicted positive observations to the total predicted positives.

Recall (Sensitivity) calculates the ratio of correctly predicted positive observations to the all actual positives.

F1-Score is the harmonic mean of precision and recall.

ROC AUC is the area under the receiver operating characteristic curve.

Confusion Matrix provides a matrix of true positives, false positives, true negatives, and false negatives.

* In Evaluate and Print Results:

This calls the above function for the model, displaying the respective performance metrics.

**7. Data Mining**

**7.1 Create and justify test designs**

Given my balanced dataset of 31466 entries, I've decided to employ a 70/30 split for segregating the training and testing sets. Here are my reasons:

1. Adequate Training Data: Using 70% of the data, approximately 22026 entries, allows me to ensure the model has sufficient data to understand the fundamental patterns and features. More data in machine learning typically allows for better generalization to unseen data.
2. Validating Model's Generalization: Allocating 30% of the data, around 9440 entries, as a test set lets me gauge the model's performance on unseen instances. This offers a sizable dataset for validation, ensuring the evaluation of the model is robust and dependable.
3. Preventing Overfitting: Over-relying on an extensive amount of data for training can lead the model to become overly complex and begin "memorizing" the training data, which in turn can affect its performance on the test set. A 70/30 split strikes a balance between training demands and the risk of overfitting.
4. Common Practice: The 70/30 split is a typical practice in data science. Experience has shown that it tends to yield solid results for a variety of datasets and tasks.
5. Balanced Data Consideration: As my dataset is balanced, I don't have to worry about class imbalances arising in the training or testing segments, which could potentially skew the model's evaluations. The 70/30 allocation ensures ample representation of both positive and negative cases in both the training and test sets.

In conclusion, given the size and characteristics of my dataset, the 70/30 split emerges as a logical and pragmatic choice, ensuring quality in model training and effective validation.

The graph showing below is the partition setting for my current dataset.



*Figure 67: Code of splitting dataset*

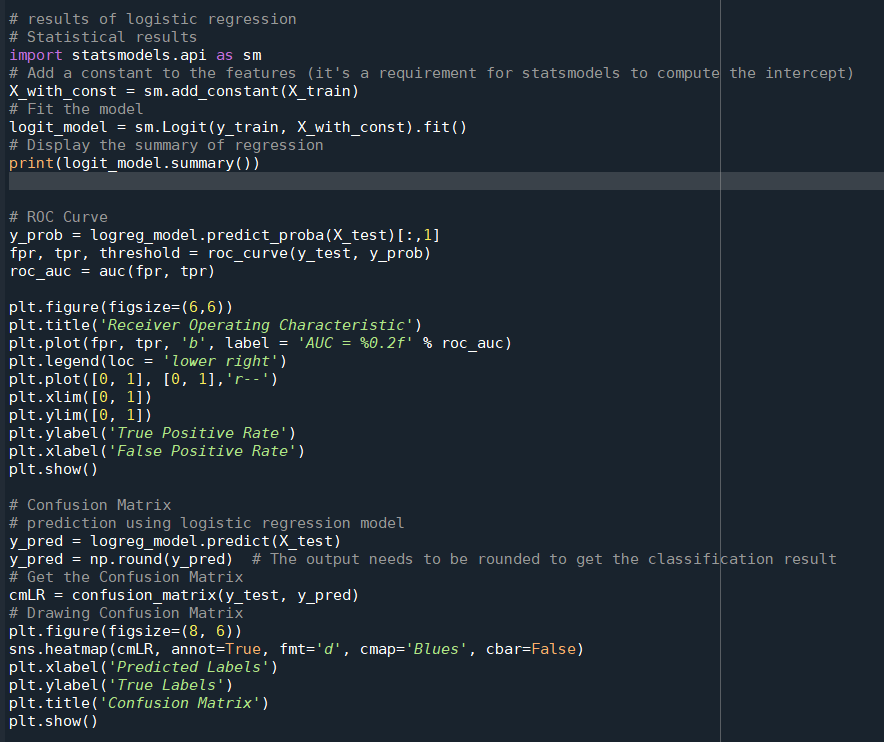
**7.2 Conduct data mining**

After setting up the testing and training sets, I use the parameters selected in step 6.3 for modeling. This is shown in the code below. I trained the logistic regression model using the training set data, and used the testing set data to predict the target column, then evaluated the predicted data using the evaluate\_model method, and finally printed the results in the console.

文本

描述已自动生成

*Figure 68: Code of building models*



*Figure 69: Code to produce Logistic Regression model results*

I will get the Logistic Regression models’ results by executing the code above.

The **first** section of the code focuses on obtaining statistical results for a logistic regression model. Using the statsmodels library, I first augment the training dataset with a constant term to account for the model's intercept. Then, I fit my logistic regression model to the training data. Finally, I display a comprehensive summary, which provides details like coefficients, standard errors, p-values, and other statistics, helping me understand the significance and influence of each predictor in the model.

The **second** section is for plotting the ROC (Receiver Operating Characteristic) curve, which is a tool for evaluating the performance of a binary classification model.

y\_prob stores the probabilities of the positive class predicted by the model.

fpr and tpr stand for false positive rate and true positive rate, respectively, and are used for plotting the ROC curve.

Next, the ROC curve is plotted using matplotlib. The area under the curve (AUC) is an indicator of the model's performance—the closer the AUC is to 1, the better the model.

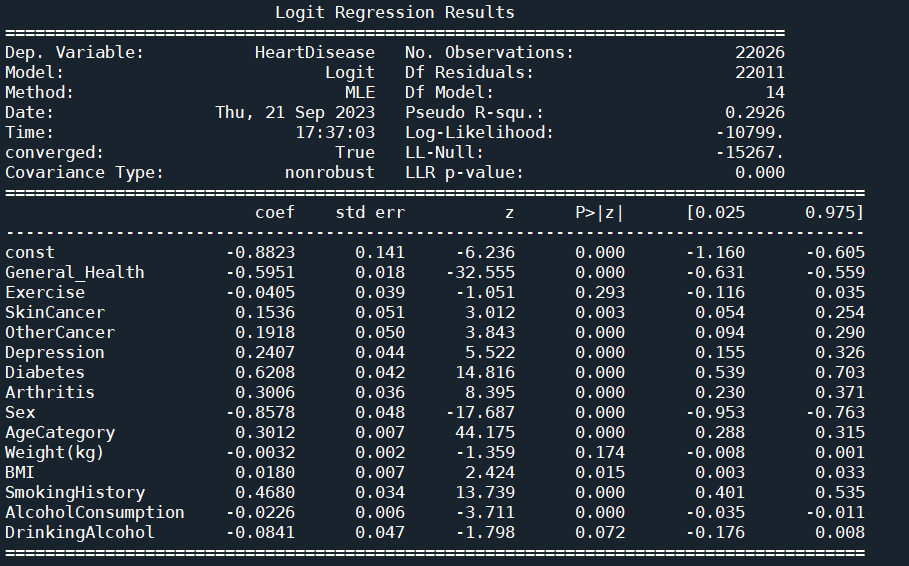
The **final** code block is for visualizing the confusion matrix. The confusion matrix is a table that shows the correct and incorrect classifications made by the model. The model's predictions are obtained using logreg\_model.predict(X\_test).The confusion\_matrix function is used to get the confusion matrix. The confusion matrix is then visualized using the sns.heatmap function.

By running the above code, I get the following results.

文本

描述已自动生成

*Figure 70: Evaluating results of Logistic Regression*



*Figure 71: Statistical results of Logistic Regression*

图表

描述已自动生成

*Figure 72: ROC curve of Logistic Regression*

图表, 树状图

描述已自动生成

*Figure 73: Confusion Matrix of Logistic Regression model*

**7.3 Search for patterns and document the model's output.**

In the context of data analysis and machine learning, the term "Pattern" usually refers to a certain regularity, association, or trend in the data, which helps unveil insights and information hidden behind the data. A "Pattern" can be evident or subtle, requiring specific algorithms and tools for identification. In this step, "pattern" can be interpreted as the key information or structure in the model output, providing deeper insights about the data or phenomenon.

Patterns in Logistic Regression:

1. Model Coefficients: Each feature's coefficient indicates the strength and direction of its relationship with the target variable. A positive coefficient means that as the feature increases, the log-odds of the target variable also increase, while a negative coefficient suggests the opposite. This offers me insights into which features influence the target variable the most.
2. Odds Ratios: Odds ratios further elaborate on the relationship between each feature and the target variable. Odds ratios greater than 1 suggest a positive correlation between the feature and the target variable, while those less than 1 indicate a negative correlation. This assists in understanding how a feature impacts the odds of the output.
3. p-values: The p-value informs me about the statistical significance of each feature. A small p-value implies that the relationship between the feature and the target variable is unlikely to be by chance, instilling greater confidence when considering it for decision-making.

Pattern 1:

In light of the information provided above, I determined the statistical significance of specific variables in the model based on their p-values. When P < 0.05, it is considered statistically significant. Therefore, in the following figure, I've highlighted the parameters that are statistically significant. They are General\_Health, SkinCancer, OtherCancer, Depression, Diabetes, Arthritis, Sex, AgeCategory, BMI, SmokingHistory, and AlcoholConsumption. This suggests that these variables have a significant predictive ability concerning the classification decision of having a heart disease or not.

屏幕上有字

描述已自动生成

*Figure 74: Statistical results in Logistic Regression model*

Pattern 2:

The coef column represents the coefficients or weights for each variable. Positive coefficient means when the eigenvalue increases by 1 unit, the logarithmic probability of an event will increase. Negative coefficient means when the eigenvalue increases by 1 unit, the logarithmic probability of an event will decrease. Therefore, the greater the absolute value, the greater the impact on the target result. Based on the Model Coefficient, I can determine that the Generalhealth, Diabetes, Sex and Smokinghistory are the top 4 influencer of heart disease.

图片包含 日历

描述已自动生成

*Figure 75: Coefficients in Logistic Regression model*

**8. Interpretation**

**8.1 Study and discuss the mined patterns**

* As shown in the figure below, while some lifestyle factors such as Exercise、Weight(kg), and DrinkingAlcohol appear to be statistically insignificant in relation to heart disease in my model, it's critical not to dismiss their importance outright. There may be other external or latent variables identified in different studies that influence these relationships. Hence, public health advisories might need to emphasize more on factors that have shown significant associations with heart disease in this research, rather than overemphasizing those that haven't.

电脑萤幕画面

描述已自动生成

*Figure 76: Variable not in the equation in Logistic Regression*

* From the logistic regression results, it's observed that while DrinkingAlcohol, a binary variable indicating if one drinks, isn't statistically significant in predicting heart disease, the variable AlcoholConsumption which quantifies the amount consumed, is significant. Several reasons could account for this phenomenon:

Quantitative Differences: Merely knowing if someone drinks might not be sufficient to predict their risk of heart disease. However, understanding the actual amount they consume might offer more insight. For instance, moderate drinking could be associated with a lower risk of heart disease, while excessive drinking might be linked to a higher risk.

Data Distribution: If a vast majority of participants in the dataset consume alcohol, then DrinkingAlcohol might not provide much additional information to the model due to its limited variability. In contrast, AlcoholConsumption may offer more variation and detail, giving the model richer information.

Interactions and Confounders: There might be other variables that correlate with both DrinkingAlcohol and AlcoholConsumption, affecting their relationship with heart disease. For example, if those who drink are also more likely to smoke, and smoking significantly correlates with heart disease risk, it could influence the outcomes.

In essence, while DrinkingAlcohol isn't a significant predictor statistically, it doesn't imply that alcohol consumption has no effect on heart disease risk. It might just not be a robust predictor within this specific dataset and model context.



*Figure 77: Comparison of 2 alcohol related variables*

* The **Sex** variable stands out prominently in the logistic regression results with a coefficient of -0.8578. Given the encoding where 0 represents males and 1 represents females, the negative coefficient can be interpreted in the following way:

The negative coefficient indicates that when transitioning from male (0) to female (1), there's a decrease in the log odds of the outcome, which in this context is having heart disease. In simpler terms, it suggests that females have a lower risk of heart disease compared to males when other factors are held constant.The magnitude of the coefficient, 0.8578, is relatively large in comparison to some of the other predictors. This emphasizes that gender plays a significant role in the model's predictions. Essentially, for each unit increase in the Sex variable (i.e., moving from male to female), the log odds of having heart disease decrease by approximately 0.8578, all else being equal.

This result aligns with various medical research findings that have indicated differences in heart disease risk between men and women. While the exact reasons can be multifaceted—ranging from biological, behavioral, to environmental factors—the model underscores gender as a crucial determinant. Healthcare practitioners might consider gender as a vital parameter when assessing heart disease risk.

* The relatively high positive coefficient for **Diabetes** implies that the presence of diabetes (i.e., when the Diabetes variable is 1) significantly increases the log-odds of having heart disease as compared to the absence of diabetes.

Diabetes has been consistently linked with an elevated risk of cardiovascular disease. High blood sugar levels, a characteristic of diabetes, can lead to damage over time to blood vessels and the nerves that control the heart. Moreover, people with diabetes often have other conditions that increase the risk, like high blood pressure, high cholesterol levels, and obesity. The interplay of insulin resistance, metabolic changes, and other factors make individuals with diabetes more prone to cardiovascular complications.

* The positive coefficient for **SmokingHistory** indicates that as the value of SmokingHistory increases, the log-odds of the outcome (having heart disease) also escalates. Essentially, this implies that those with a history of smoking have a higher likelihood or risk of developing heart disease compared to those without a smoking history.

Smoking is one of the leading risk factors for cardiovascular diseases. The harmful substances in tobacco can damage the heart and blood vessels, leading to the build-up of a fatty material called atheroma, which narrows the arteries. This can result in angina, a heart attack, or a stroke. Furthermore, the carbon monoxide in tobacco smoke reduces the amount of oxygen in the blood, which means the heart has to work harder to supply the body with the oxygen it needs. The nicotine in cigarettes stimulates the body to produce adrenaline, which makes the heart beat faster and raises blood pressure, further increasing the heart's workload.

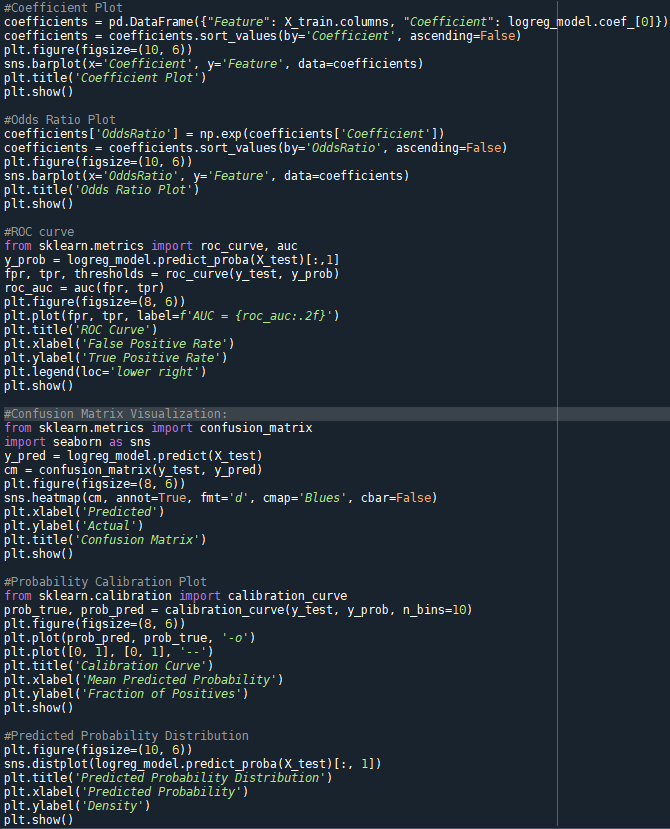
**8.2 Visualize the data, results, models, and patterns**

In this step, I mainly used 6 parts of the visualization results:

1. Coefficient Plot: To visualize the coefficients of the predictor variables, highlighting the relative importance and direction (positive/negative) of each predictor.
2. Odds Ratio Plot: To depict odds ratios for predictors, helping in understanding the change in odds of the outcome for a unit change in the predictor.
3. ROC Curve (Receiver Operating Characteristic): Demonstrates the true positive rate against the false positive rate, and the AUC provides a measure of the model's overall performance.
4. Confusion Matrix Visualization: Using heatmaps or color-coding to represent the true positive, true negative, false positive, and false negative rates.
5. Probability Calibration Plot: Shows the alignment of predicted probabilities with actual outcomes.
6. Predicted Probability Distribution: A histogram or density plot indicating the distribution of predicted probabilities, offering insights into the model's confidence.

Selecting from these visualizations can help provide a comprehensive overview of the logistic regression model's performance and characteristics.

The code is shown below.



*Figure 78: Code of visualization of Logistic Regression Model*

* Coefficient Plot of Logistic Regression Model

图表, 条形图

描述已自动生成

*Figure 79: Coefficient Plot of Logistic Regression Model*

* Odds Ratio Plot of Logistic Regression Model

图表

描述已自动生成

*Figure 80:* *Odds Ratio Plot of Logistic Regression Model*

* ROC curve of Logistic Regression Model

图表

描述已自动生成

*Figure 81: ROC curve of Logistic Regression Model*

* Confusion Matrix of Logistic Regression Model

图表, 树状图

描述已自动生成

*Figure 82: Confusion Matrix of Logistic Regression Model*

* Probability Calibration Plot of Logistic Regression Model

图表, 折线图

描述已自动生成

*Figure 83: Probability Calibration Plot of Logistic Regression Model*

* Predicted Probability Distribution of Logistic Regression Model

图表, 直方图

描述已自动生成

*Figure 84: Predicted Probability Distribution of Logistic Regression Model*

**8.3 Interpret the results, models, and patterns**

In my previous analysis, I utilized Logistic Regression to delve deeply into the relationship between lifestyle habits, age, and other health factors with heart disease.

Initially, while certain lifestyle habits such as exercise and food choices didn't have a significant effect in the model, this doesn't imply they're entirely unrelated to heart disease. There might be other external or hidden factors at play. This also reminds us that in public health promotion, emphasis should be placed more on factors that have been proven to have a significant association with heart disease.

Next, from the analysis, a distinctive observation emerges regarding alcohol consumption and its relation to heart disease. While the binary variable DrinkingAlcohol, which simply denotes whether an individual drinks alcohol or not, doesn't show significant association with heart disease, the continuous variable AlcoholConsumption, indicating the quantity of alcohol consumed, exhibits significance. This suggests that the mere act of drinking may not be a strong predictor for heart disease. Instead, the amount of alcohol consumed plays a pivotal role. It's plausible that moderate alcohol consumption might have neutral or even protective effects, but excessive consumption could pose cardiovascular risks. This underscores the importance of considering the quantity of alcohol intake when assessing its health implications, rather than a binary classification of drinkers and non-drinkers.

Moving on to the heavy influencer with heart disease, three factors—Sex, Diabetes, and SmokingHistory—stand out as particularly influential in predicting heart disease. The negative coefficient for Sex indicates that females are significantly less likely to have heart disease compared to males. On the other hand, the presence of diabetes significantly increases the likelihood of heart disease, as indicated by its positive coefficient. Similarly, a history of smoking elevates the risk of heart disease, emphasized by the positive coefficient for SmokingHistory. These findings highlight the paramount importance of considering gender, diabetic status, and smoking history when assessing cardiovascular risks.

In conclusion, although I couldn't explore all potential variables and associations in this study, these findings offer valuable insights for medical and health professionals regarding heart disease risks and possible prevention strategies. Future research might delve deeper into factors not considered here to enrich our understanding.

**8.4 Assess and evaluate results, models, and patterns**

1. Assess for model

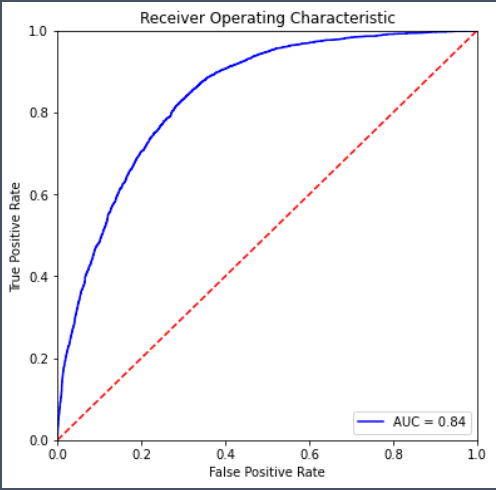
After evaluating my logistic regression model, I found it demonstrates quite commendable performance. Firstly, the model's accuracy rate reached 76.04%, implying that more than three-quarters of the predictions were correct across all test data. This instilled in me a substantial amount of confidence, leading me to believe that the model could maintain relative accuracy when dealing with unknown data.

文本

描述已自动生成

*Figure 85: Evaluation results for Logistic Regression*

However, what I place more emphasis on is the model's AUC value, as it offers a more comprehensive view of the model's ability to distinguish between positive and negative classes. My model's AUC stands at 0.84, which is typically considered "very good" in practical applications. An AUC value nearing 1 indicates that the model performs exceptionally well in differentiating between positive and negative categories, whereas a 0.5 AUC suggests the model lacks discriminative power, equivalent to a random guess. Given that 0.84 is significantly greater than 0.5 and approaches 1, it strongly signifies the model's excellent performance in the classification task.



*Figure 86: ROC curve for Logistic Regression*

2. Assess for decision making

In the intricate landscape of heart disease risk factors, I've identified distinct patterns through the logistic regression model that shed light on potential prevention strategies. Based on these discerned patterns, I present an assessment to guide decision-making in managing and mitigating heart disease risk.

Pattern 1: Importance of Lifestyle Factors.

Insight: Even though certain lifestyle factors don't exhibit a significant correlation with heart disease in this model, external studies may have established links.

Decision Recommendation: Rather than discarding the role of the stated lifestyle factors, public health campaigns should present a holistic view of heart health. They should strike a balance by emphasizing factors that showed significant associations in their study, while also acknowledging the potential roles of exercise, diet, and alcohol consumption as highlighted in external research. A nuanced approach, incorporating findings from multiple studies, would be most beneficial for public awareness and heart disease prevention.

Pattern 2: Sex as a strong influencer

I Insight: The analysis reveals a notable disparity between males and females when it comes to the risk of heart disease. Specifically, females are considerably less likely to develop heart disease compared to their male counterparts, as indicated by the highest negative coefficient for Sex.

Decision Recommendation: It's crucial for healthcare providers to recognize the higher vulnerability of males to heart disease. Male patients should be advised early on about lifestyle adjustments, regular screenings, and possible interventions to mitigate their heightened risk.

Pattern 3: Diabetes' Strong relation with Heart Disease.

nsight: The presence of diabetes is a significant predictor of heart disease. Individuals diagnosed with diabetes have a notably increased likelihood of developing heart disease, as evidenced by the relatively high positive coefficient for Diabetes.

Decision Recommendation: Given the established link between diabetes and heart disease, healthcare systems should prioritize proactive screenings for heart issues among diabetic patients. Moreover, diabetic patients should be offered educational programs highlighting the connection and emphasizing the importance of rigorous diabetic management and monitoring.

Pattern 4: Smoking and Heart Disease Association.

Insight: A history of smoking is strongly correlated with an elevated risk of heart disease. Individuals with a past marked by smoking have an increased propensity to develop heart conditions, underscored by the positive coefficient for SmokingHistory.

Decision Recommendation: Public health campaigns should continually emphasize the cardiovascular risks associated with smoking. For those with a smoking history, healthcare providers should offer specialized heart screenings and actively promote smoking cessation programs to help reduce the future risk of heart disease.

Overall Decision Implications:

The patterns deduced from the logistic regression model can be of immense value in shaping heart health strategies, public awareness campaigns, and medical interventions. By identifying and acting upon these specific risk factors and associations, policymakers, and medical professionals can enact targeted approaches that address the unique needs of different risk groups, potentially leading to improved heart health outcomes in the broader population.

**8.5 Iterate prior steps (1 – 7) as required**

1. Adjustment in Data Balancing Strategy:

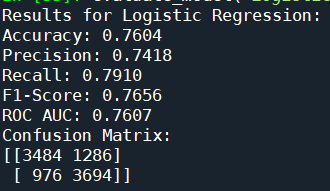
Initially, I utilized the Undersampling method, which meant downsizing data to achieve a balance between positive and negative samples. This approach, while effective, could lead to the potential loss of valuable information, especially when working with limited data. Recognizing this limitation, I shifted to the Oversampling method. Oversampling aims to enhance the balance by increasing the representation of the minority class, ensuring not only that the imbalance issue is addressed but also that I harness the full potential of the available data without unnecessary losses.The code is shown below.

文本

描述已自动生成

*Figure 87: Oversampling dataset to project data*

As illustrated in the following figure, we compared the accuracy of the Logistic Regression model after balancing the data using both the Undersampling and Oversampling methods. Employing the Oversampling method to balance the data resulted in a slight improvement in the model's accuracy and AUC.

 文本

描述已自动生成

*Figure 88: The evaluation results Comparison of 2 models*

2. Data split ratio Modification:

My preliminary dataset split was at a 70/30 ratio, a commonly accepted split that usually delivers reliable results, ensuring ample data for training while retaining a robust set for performance evaluation. However, considering the specific nature and big volume of my dataset, I believed that less training data might yield a more stable model. Consequently, I adjusted the ratio to 67/33. This would allow the model to learn from slightly less data set, potentially avoid overfit.

文本

描述已自动生成

*Figure 89: Changing data split ratio to 67/33*

As depicted in the graph below, after adjusting the partition, Logistic Regression model showed a subtle increase in accuracy.

文本

描述已自动生成文本

描述已自动生成

*Figure 90: Comparison of 2 models in analysis node*

Reference:

1. Connie W. Tsao, Aaron W. Aday, Zaid I. Almarzooq, Alvaro Alonso, Andrea Z. Beaton, Marcio S. Bittencourt, Amelia K. Boehme, Alfred E. Buxton, April P. Carson, Yvonne Commodore-Mensah, Mitchell S.V. Elkind, Kelly R. Evenson, Chete Eze-Nliam, Jane F. Ferguson, Giuliano Generoso, Jennifer E. Ho, Rizwan Kalani, Sadiya S. Khan, Brett M. Kissela, Kristen L. Knutson, ... Seth S. Martin. (2022). Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. *American Heart Association.* *145:e153–e639.* [https://www.ahajournals.org/doi/10.1161/CIR.0000000000001052?utm\_campaign=sciencenews21-22&utm\_source=science-news&utm\_medium=phd-link&utm\_content=phd-01-26-22#](https://www.ahajournals.org/doi/10.1161/CIR.0000000000001052?utm_campaign=sciencenews21-22&utm_source=science-news&utm_medium=phd-link&utm_content=phd-01-26-22)
2. Usama Fayyad, Gregory Piatetsky-Shapiro, & Padhraic Smyth. (1996). *Knowledge Discovery and Data Mining: Towards a Unifying Framework.* AAAI Press, <https://aaai.org/papers/014-knowledge-discovery-and-data-mining-towards-a-unifying-framework/>

I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain, then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data.