The US Cardiovascular Diseases Lifestyle Risk Prediction Research

https://github.com/XLUO937/BDAS-For-722.git

**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 | 9.23-9.24 | Topic choosing | Hard to decide which topic to choose from. |
| Data Understanding | 9.25-9.26 | Search data | Hard to find free open data, or the data is irrelevant to research. |
| Data preparation | 9.27-9.28 | Clean data, preproduce data | Data is not enough after clean, or some of the data is not included. |
| Data transformation | 9.29 | reduction and projection | The use of software is not skilled and takes a long time |
| Data-mining method(s) selection | 9.30-10.2 | 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 | 10.2-10.3 | algorithm selection from IPSS's current ones | It does not match the objective and needs to redo. Take longer time than expected. |
| Data Mining | 10.4-10.6 | Search for pattern | Hard to find a pattern |
| Interpretation | 10.7-10.9 | Visualize, interpret, evaluate, and iterate | Take longer than expected |
| Report Evaluation | 10.10-10.13 | 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 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 CSV 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 CSV documents. The part 1 CSV 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:

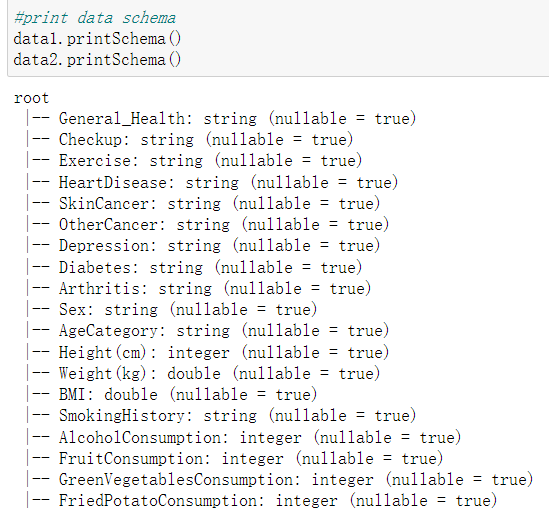
* General Health is an ordinal data which shows the sample’s general health condition. The categories include Excellent, Very Good, Good, Fair and Poor.
* Checkup is a nominal data which shows sample’s medical checkup frequency. The checkup includes never, 5 or more years ago, within the past 5 years, 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.

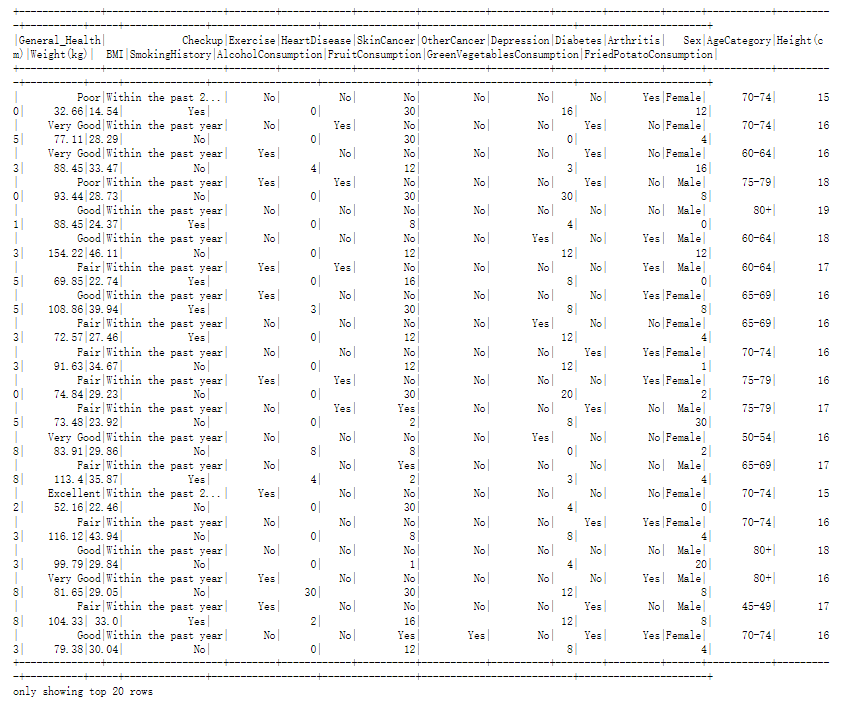
In iteration 2, the dataset is divided into 7 columns of continuous data, 9 columns of flag data, 2 column of nominal data, and 1 columns of ordinal data.

In iteration 3, as I used python to process data, I 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**.

In this iteration, similar to iteration 3, I will segregate the data into the two categories: Numerical Variables and Categorial Variables. However, this time, I will be utilizing the PySpark dataframe format.



*Figure 1: Infomation of the data*



*Figure 2: Dataset table screenshot part 1*

表格

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.

In iteration 3, I conducted 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.

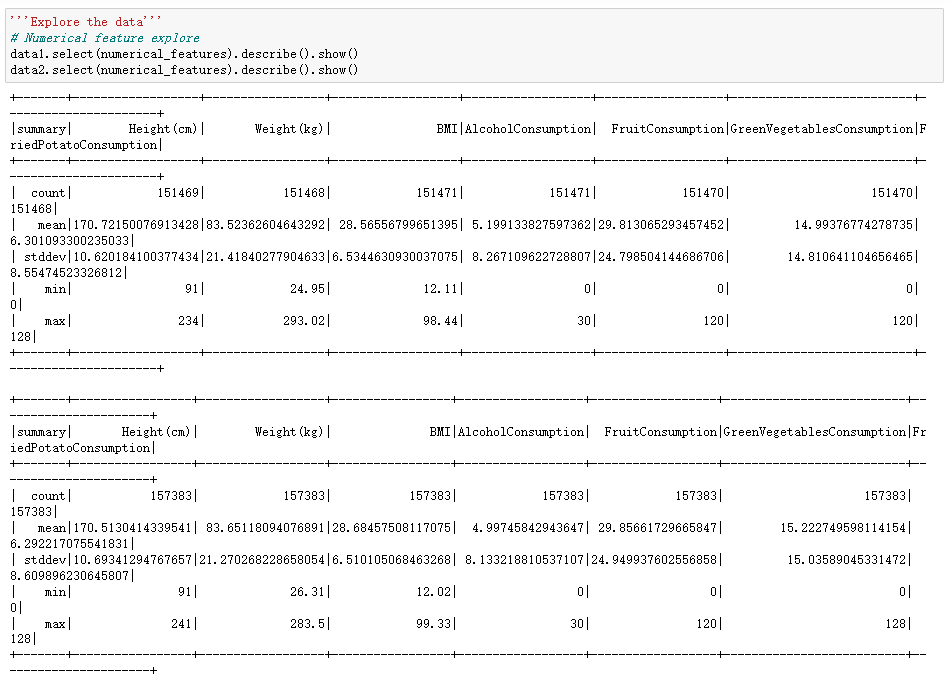
In this iteration, we will continue to follow the data exploration direction from the third iteration, but with a different data exploration tool. This time, I have chosen Jupyter Notebook as the tool for data exploration, processing, and mining. Given that PySpark's DataFrame does not support direct data visualization, in this round of data exploration, I can only conduct more basic analyses compared to previous iterations.

文本

描述已自动生成

*Figure 4: Definition of Numerical and Categorical Variables*

2.3.1 Numerical Variables Exploration

**Let's start with** some basic data exploration. By running the code below, I can obtain basic statistical values for the numerical variables in my dataset

*Figure 5: Basic Description of Numerical Variables in Dataset*

**On this basis**, I aim to obtain more in-depth statistical data. By executing the code below, I can retrieve statistical results for the numerical variables, including Average (avg), Standard Deviation (stddev), Skewness (skewness), Kurtosis (kurtosis), and Variance (variance).

图形用户界面, 应用程序

描述已自动生成

*Figure 6: Code to Get Statistical Performance of Numerical Variables*

表格

描述已自动生成

表格

描述已自动生成

*Figure 7: Statistical Performance of Numerical Variables*

**Additionally**, I would like to explore the correlation with the target column 'HeartDisease.' To do so, I need to convert the 'HeartDisease' column from string values ('yes' and 'no') to numeric values (1 and 0) before conducting the correlation analysis.

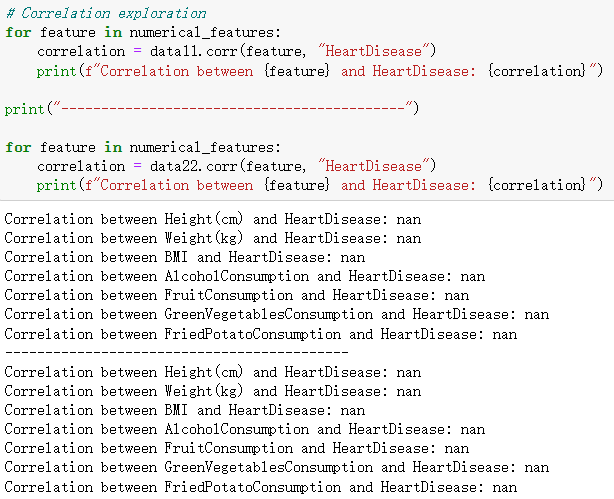
We can observe that after the conversion, the data type of the 'HeartDisease' column has been changed to integer, confirming the success of our data transformation.

文本

描述已自动生成

*Figure 8: Data Type Transformation*

Next, I conducted an exploratory analysis to assess the correlation between all numerical variables and the target column 'HeartDisease.' The results are as follows. At the moment, it appears that there is not a strong relationship between the numerical variables and the target column 'HeartDisease.



*Figure 9: Correlation Exploration* between numerical variables and 'HeartDisease..

2.3.2 Categorical Variables Exploration

文本

描述已自动生成

*Figure 10: Coding in Categorical Data Count*

表格

描述已自动生成

*Figure 11: Count of GeneralHealth*

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

表格

描述已自动生成

*Figure 12: 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.

表格

描述已自动生成

*Figure 13: Count of Exercise*

Exercise: More patients reported that they exercise compared to those who do not.

表格

描述已自动生成

*Figure 14: 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.

表格

描述已自动生成

*Figure 15: Count of SkinCaner*

SkinCancer: The vast majority of patients do not have skin cancer.

表格

描述已自动生成

*Figure 16: Count of OtherCancer*

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

表格

描述已自动生成

*Figure 17: Count of Depression*

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

表格

描述已自动生成

*Figure 18: Count of Diabetes*

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

表格

描述已自动生成

*Figure 19: Count of Arthritis*

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

文本, 表格

描述已自动生成

*Figure 20: Count of Sex*

Sex: There are slightly more female patients than male patients in the dataset.

表格

描述已自动生成表格

描述已自动生成

*Figure 21: 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.

表格

描述已自动生成

*Figure 22: 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. Only data1 has missing value and data2 has complete database.

表格

描述已自动生成

*Figure 23: 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

Now, I use the IQR method to define a outlier.

First, for each numerical variable, calculate the first quartile (Q1) and the third quartile (Q3).

Next, compute the IQR, which is Q3 minus Q1.

Using the IQR, we set boundaries to define outliers. Typically, any value below Q1 - 1.5 IQR or above Q3 + 1.5 IQR is considered an outlier.

文本

描述已自动生成

*Figure 24: Data Box Plot of Numerical Variables*

The outliers numbers of numerical variables above 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

图片包含 文本

描述已自动生成 *Figure 25: Data Duplicate Count*

As shown above, the number of duplicates is 24+24=48. 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(figure26 below) the last time a physical examination was conducted, reflecting whether the subject undergoes regular checkups and the frequency of those checkups.

表格

描述已自动生成

*Figure 26: 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

表格

描述已自动生成

*Figure 27: The Count of AgeCategory*

图片包含 图表

描述已自动生成

*Figure 28: The Count of Diabetes*

Figure 27 displays the count of AgeCategory. There is a bar with value in 4 which is the unclassified value. In Figure 28, 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.

文本

描述已自动生成

*Figure 29: Code to Get Statistical Summary of Numerical Data*

文本

描述已自动生成

*Figure 30: 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

文本

描述已自动生成文本

描述已自动生成

*Figure 31: 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

图形用户界面

中度可信度描述已自动生成

*Figure 32: Code to remove null value*

As shown in the graph above, I execute “.na.drop” 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.

表格

描述已自动生成

*Figure 33: The null value count before remove*

手机屏幕的截图

描述已自动生成

*Figure 34: The null value count after remove*

1. Unclassified value

文本

描述已自动生成

*Figure 35: Code to remove unclassified value in AgeCetegory*

表格

描述已自动生成

*Figure 36:* *AgeCetegory Column before Reclassify*

表格

描述已自动生成

*Figure 37: AgeCetegory Column after Reclassify*

The three charts above depict the process I undertook to handle the "AgeCategory" field. Using the code in Figure 35, I retained data in the AgeCategory that didn't have a value of 4. I then verified the results through data visualization. From Figures 36 and 37, 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.

文本, 日程表

描述已自动生成

*Figure 38: Code to remove unclassified value in Diabetes*

表格

描述已自动生成

*Figure 39: Diabetes field before Reclassify*

图示

描述已自动生成

*Figure 40: Diabetes field 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 shown in above figure 38, I replaced the respective values in the diabetes column with "Yes" and "No". The successful outcome of this operation is demonstrated in figure 40 compared with figure 39.

1. Duplicate value

文本, 电子邮件

描述已自动生成

*Figure 41: Code to remove Duplicate value*

图形用户界面, 文本, 应用程序

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 74. 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 the outlier count before and after removing outliers is shown below.

表格

描述已自动生成

*Figure 44: Outlier count before and after removing outliers*

From the image above, we can see that after processing, the counts of outliers in the same definition have been changed to zero after removing outliers.

**3.3 Construct the data**

In my data, there is a field for alcohol consumption. From the figure 45, 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.

表格

描述已自动生成

*Figure 45: AlcoholConsumption count circled Zero alcohol consumption*

图片包含 图形用户界面

描述已自动生成

*Figure 46: Code of Constructing drinking alcohol field*

文本

描述已自动生成文本

描述已自动生成

*Figure 47: the new generated field in the data schema*

**3.4 Integrate various data sources**

Due to the large volume of data, I have my 308,758 records stored in two different CSV spreadsheets. What I need to do now is merge these two files into one. The CSV spreadsheet for dataset part 1 contains 151,413 records, and the CSV spreadsheet for dataset part 2 contains 157,345 records.

文本, 信件

描述已自动生成

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

文本, 表格

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 CSV file. As shown in the image above, the “df” data count is 308,758, 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.

In iteration 3 I changed the data from object type to numerical data type because I used the logistic regression model.

In this iteration, by using PySpark DataFrames, machine learning models require numerical input features. So, I need to format the data into features that are suitable format and representation for the logistic regression model in PySpark's machine learning library.

图片包含 文本

描述已自动生成

*Figure 50: Code of formatting data*

The code above shows the process of formatting data in my dataset.

I use StringIndexer to convert each string categorical column into a column of numerical indices. By converting string categorical values into a distinct index, I create a numerical representation of the categories. Each distinct string value in the original column will get an index, starting from 0.

报纸上的文字

描述已自动生成

*Figure 51: Data schema after formatting*

**4. Data transformation**

**4.1 Reduce the data**

In iteration 3 I used the raw, uncleaned data to create a correlation matrix that describes the relationships between variables.

In this iteration, I will compute correlations to find the features with the least correlation.

图片包含 文本

描述已自动生成

*Figure 52: Code to compute correlations*

In the code above, I first list out all the numerical columns present in the DataFrame, as the correlation function (Pearson's correlation) works with numerical data.

Then, for each numerical feature, I calculate its correlation with the HeartDisease\_indexed column using the corr function available in the stat methods of the DataFrame. The result is a dictionary (correlations) with features as keys and their corresponding correlation values as the dictionary values.

I want to find the features with the least correlation. So, finally, we sort the features based on the absolute values of their correlations. This way, we capture both positive and negative weak correlations. I then pick the first four features from the sorted list, which represent the least correlated features.



*Figure 53: Results of the 4 features least correlated with HeartDisease*

Based on the results above, I observed that 'FruitConsumption', 'Height(cm)', 'GreenVegetableConsumption', and 'FriedPotatoConsumption' have minimal correlation with 'HeartDisease' (the top 4 features least correlated with HeartDisease). Consequently, these four columns can be removed. This result is the same with iteration 3.

文本

描述已自动生成

*Figure 54: Code to Remove least correlated fields*

文本

描述已自动生成

*Figure 55: Data fields after removing least correlated 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 run the code at the top, I can see that the number of "No" far exceeds the number of "Yes" in the data shown on the bottom side of the figure. I need to manipulate this data to make it essentially balanced.

表格

低可信度描述已自动生成

*Figure 56: Current target field distribution*

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

*Figure 57: 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 “No” and “Yes” in my target column became consistent, each with 15,571 entries, as depicted in the following figure.

文本

描述已自动生成

*Figure 58: Comparison of the value count of Heart Disease before and after undersampling*

**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 iteration 3, I didn’t use regression approaches.

After careful consideration, I've decided to forgo the use of the Dependency Modeling method. Based on results from my previous iterations (iteration2 and iteration3), logistic regression has already provided valuable insights into which columns have a stronger or weaker relationship with the target column, HeartDisease. This approach appears more straightforward and efficient for my current objectives, making Dependency Modeling methods less relevant in this context. So, in this iteration 4, I only use the classification method.

**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 one essential technique pertinent to my research goals: Classification 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 iteration 3, I explored a lot of possibilities using python. In this iteration 4, I will explore the algorithms available in PySpark's ML library, as I am working with PySpark DataFrames.

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 and Results of split train and test data*

The code splits a balanced DataFrame into training (80%) and test (20%) sets. Then I define a list of input features which includes columns ending with "\_indexed" and specific numeric columns but exclude the target column. I use the VectorAssembler to assemble these features into a single feature vector for both the training and test data. And finally, I print out the number of records in each dataset.

**Classification algorithm selection**

Firstly, I selected several models suitable for classification, as shown below. The reason I chose these particular algorithms for my classification models dictionary is that these algorithms are among the classic and commonly used methods for classification in the PySpark ML library.

图表, 散点图

描述已自动生成

*Figure 60: Model dictionary definition*

This code shown below first defines the evaluation method with selected metrics then trains and with the training dataset. After that, I evaluate the data with the testing dataset. The final model performance metrics as per your specifications are computed for each model, and the results are displayed at the end.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图片包含 图形用户界面

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:

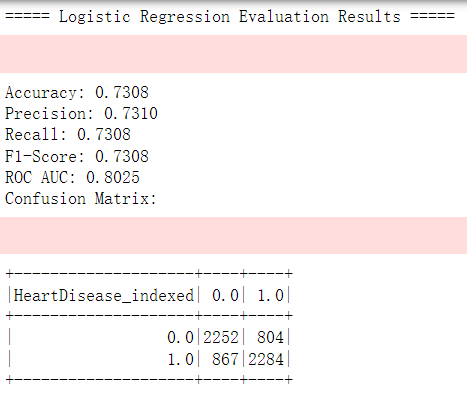
The F1-score is the harmonic mean of precision and recall. It provides a single, comprehensive metric especially useful when you need to balance between precision and recall. It's particularly handy when there's an imbalance between classes.

* 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.

表格

低可信度描述已自动生成

表格

中度可信度描述已自动生成 表格

描述已自动生成 表格

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 73.08% and an F1-Score of 0.7308, this model performs reasonably well. The AUC of 0.8025 suggests that the model can distinguish between the positive and negative classes effectively. The confusion matrix shows that there are 804 false positives and 867 false negatives. The precision of 73.10% and recall of 73.08% suggest a balanced performance between identifying true positives and minimizing false positives.

**Decision Tree:**

Achieving an accuracy of 74.96% and an F1-Score of 0.7461, the Decision Tree model delivers a commendable performance. The AUC score of 0.7178 demonstrates the model's effective differentiation between the two classes. The confusion matrix indicates 1120 false positives and 434 false negatives. With a precision of 76.17% compared to a recall of 74.96%, the model strikes a balance between the prediction of true positives and the reduction of false positives.

**Random Forest:**

Random Forest displays an accuracy of 75.33% and an F1-Score of 0.7514, showcasing its solid performance. Its AUC value of 0.8204 indicates efficient class differentiation. The confusion matrix presents 1019 false positives and 512 false negatives. A precision of 75.97% against a recall of 75.33% signifies a balanced trade-off between true positive prediction and false positive minimization.

**GBTClassifier:**

The GBTClassifier reports an accuracy of 75.62% and an F1-Score of 0.7556, denoting a notable performance. Its AUC score stands at 0.8353, suggesting effective class discrimination. The confusion matrix points out 893 false positives and 620 false negatives. With a precision of 75.79% in comparison to a recall of 75.62%, the model achieves a balanced result between the detection of true positives and the limitation of false positives.

**NaiveBayes:**

With an accuracy of 69.12% and an F1-Score of 0.6899, the NaiveBayes model exhibits a moderate performance. The AUC score is at 0.4871, which might raise concerns regarding its ability to differentiate between classes effectively. The confusion matrix identifies 1135 false positives and 782 false negatives. Although the precision stands at 69.31% and the recall at 69.12%, indicating a balanced act, the model's overall performance may require further enhancement.

**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.

**Logistic Regression** stands out in terms of its balanced performance metrics, its interpretability, and its computational efficiency. Here's a breakdown of the decision:

Accuracy and F1-Score: While Logistic Regression doesn't have the highest accuracy among the models (it's 73.08%, compared to the highest, which is GBTClassifier at 75.62%), its performance is still competitive. The F1-Score of the Logistic Regression model (73.08%) is close to the highest (GBTClassifier at 75.56%). F1-Score gives a balanced measure of a model's performance, especially when classes might be imbalanced.

ROC AUC: The AUC for Logistic Regression is 0.8025, which is commendable. While the GBTClassifier and Random Forest models have slightly higher AUCs, the difference is not substantial enough to disregard Logistic Regression, especially when other factors come into play.

Precision and Recall: Logistic Regression offers balanced precision and recall metrics (both at 73.08%), which is crucial in medical scenarios to ensure that both false positives and false negatives are minimized.

Interpretability: One of the significant advantages of Logistic Regression over models like Random Forest or GBTClassifier is its simplicity and interpretability. With Logistic Regression, it's straightforward to understand the impact of each feature on the prediction, making it easier to explain to non-technical stakeholders. This is especially crucial in the medical domain, where explaining the decision-making process can be as important as the decision itself.

Computational Efficiency: Logistic Regression models tend to be computationally less intensive than ensemble methods like Random Forest or GBTClassifier. This makes them faster to train and predict, especially when the dataset grows.

Consistency with Previous Version: Consistency in model selection, when backed by valid reasons, is essential to maintain the credibility of analyses. Since Logistic Regression was previously chosen based on its merits, there's continuity in choosing it again.

In conclusion, while ensemble methods like GBTClassifier and Random Forest have slightly higher performance metrics in certain areas, the overall balanced performance, interpretability, and computational efficiency of Logistic Regression make it the optimal choice for this classification task, reaffirming the selection made in the previous version.

**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.

文本

描述已自动生成

*Figure 63: Coding of Splitting dataset*

文本

描述已自动生成

*Figure 64: Coding of Modelling*

Afterward, I built the model.

* In Model Initialization:

Here, I am initializing the logistic regression model. The labelCol parameter indicates the column in the dataset that contains the target variable or labels, in this case, "HeartDisease\_indexed". The featuresCol parameter indicates the column that contains the feature vectors, named "features".

* In Model Training:

The fit method trains the logistic regression model on the training data provided (train\_data).

* In Evaluation Function:

Here, the trained model (lr\_model) is evaluated on the test data (test\_data). This is done using the evaluate\_model function, which is defined to assess the model's performance using various metrics.

The evaluate function has a loop iterates through the results returned by the evaluate\_model function. For each metric (like Accuracy, Precision, etc.), the value is printed out in a formatted manner. If the metric is the "Confusion Matrix", it displays the matrix in a tabulated form using the show method.

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.

**7. Data Mining**

**7.1 Create and justify test designs**

Given my balanced dataset of 31142 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 21846 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 9296 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.

文本

描述已自动生成

*Figure 65: Code of splitting dataset*

**7.2 Conduct data mining**

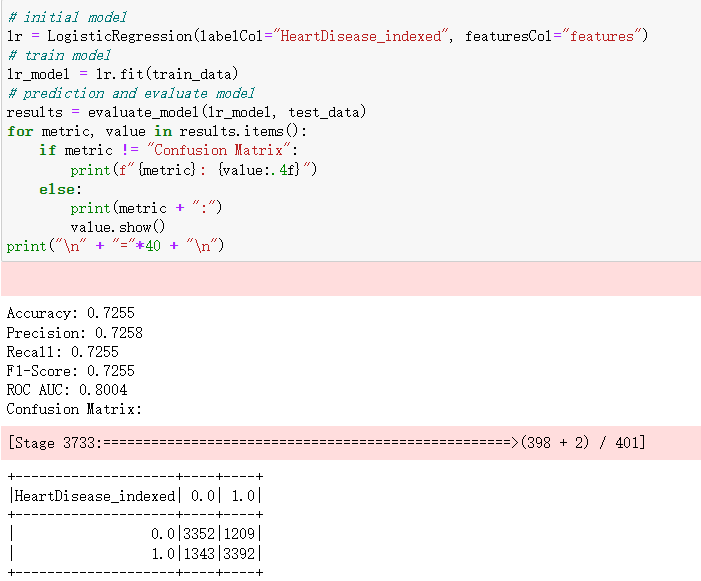
After setting up the testing and training sets, I use the parameters selected in step 6.3 for modeling. This is shown in the code below. I trained the logistic regression model using the training set data, and used the testing set data to predict the target column, then evaluated the predicted data using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

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

*Figure 66: Code of defining evaluation method*

The code in Figure 66 shows the evaluate\_model function assesses the performance of a given machine learning model on provided test data. Initially, it applies the model to predict outcomes on the test dataset. Subsequently, it computes various evaluation metrics, including accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve (AUC). The function employs two evaluators: a multiclass\_evaluator to measure accuracy, precision, recall, and F1-score, and a binary\_evaluator to determine the ROC AUC value for binary classification tasks. Moreover, the function constructs a confusion matrix by grouping the actual labels (HeartDisease\_indexed) and comparing them with the model's predictions, filling any missing values with zeros. The function then returns a dictionary containing all these evaluation metrics and the confusion matrix.



*Figure 67: Code of producing evaluation results*

The given code in Figure 67 snippet begins by initializing a Logistic Regression model with specified columns for labels (HeartDisease\_indexed) and features (features). Once initialized, the model is trained using the train\_data dataset. After training, the evaluate\_model function is called to assess the model's performance on the test\_data. The evaluation results, which include various performance metrics and a confusion matrix, are then iterated over and printed out. For each metric, the value is displayed in a formatted manner, and if the metric pertains to the confusion matrix, the matrix is explicitly shown. Finally, a series of equals signs are printed to provide a visual separation in the output.

图形用户界面, 文本, 应用程序

描述已自动生成

*Figure 68: Code to produce Logistic Regression model coefficients results*

This code segment in Figure 68 is dedicated to visualizing the importance of features used in the Logistic Regression model. The coefficients from the trained model, which reflect the significance of each feature, are extracted and converted into an array. Using the matplotlib library, a bar graph is plotted where each bar represents a feature's coefficient, effectively illustrating its importance. The bars are labeled with the corresponding feature names, rotated at a 45-degree angle for clarity. The graph is adorned with a title, and the y-axis is labeled to indicate the coefficient values. After setting up the visual properties, the plot is rendered using the show() method.

文本

描述已自动生成

*Figure 69: Code to produce ROC curve*

This piece of code in Figure 69 aims to visualize the Receiver Operating Characteristic (ROC) curve for the trained Logistic Regression model. Initially, a Binary Classification Evaluator is imported from PySpark's ML library. The training summary of the logistic regression model is then utilized to extract the ROC curve data, which is converted to a pandas DataFrame. Using matplotlib, a plot is created where the False Positive Rate (FPR) is plotted against the True Positive Rate (TPR), producing the ROC curve. The x-axis and y-axis are aptly labeled, a title is provided, and a legend is added to annotate the curve. The visualization is finally displayed using the show() method, offering a graphical representation of the model's ability to discriminate between the positive and negative classes.

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

表格

描述已自动生成

*Figure 70: Evaluating results of Logistic Regression*

图表, 条形图

描述已自动生成

*Figure 71: Feature importance from Logistic Regression*

图表

描述已自动生成

*Figure 72: ROC curve of Logistic Regression*

**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: In this iteration 4, as I utilized PySpark data frame, the Logistic Regression Model does offer coefficients for the logistic regression model, but it doesn't directly provide the p-values for the coefficients.

**Pattern for Coefficients:**

The `Coefficients` column signifies the weight or magnitude each feature has in influencing the prediction outcome. A positive coefficient suggests that with a unit increase in the feature's value, the log-odds of the event (in this context, the heart disease occurrence) increases, pushing the prediction more towards a positive outcome. Conversely, a negative coefficient indicates that a unit increment in the feature's value reduces the log-odds, making the event less likely. Assessing the magnitude of these coefficients, it's discernible that `Sex\_indexed`, `Diabetes\_indexed`, `Arthritis\_indexed`, and `SkinCancer\_indexed` have the most profound influence on heart disease, given their relatively high absolute values.

**Pattern for Odds Ratios:**

The `Odds Ratios` delineate the odds of an event occurring concerning a one-unit increase in the feature. An odds ratio greater than 1 indicates that the feature increases the likelihood of the event, while an odds ratio less than 1 suggests the opposite. Interpreting the obtained ratios, we can surmise that factors like `Diabetes\_indexed`, `Sex\_indexed`, and `Arthritis\_indexed`, which have odds ratios significantly greater than 1, tend to substantially elevate the chances of heart disease. On the other end, features such as `AgeCategory\_indexed` and `DrinkingAlcohol\_indexed` with odds ratios less than 1 suggest a diminishing effect on the probability of heart disease onset.

文本

描述已自动生成

*Figure 73: Statistical results in Logistic Regression model*

**8. Interpretation**

**8.1 Study and discuss the mined patterns**.

1. Sex\_indexed:

The positive coefficient of 1.00157179 for `Sex\_indexed` indicates that a change in this feature has a pronounced effect on the log-odds of developing heart disease. Specifically, the feature value corresponding to a particular gender (e.g., male) significantly boosts the log-odds of having heart disease compared to the other gender (e.g., female).

Studies have consistently shown gender disparities in heart disease risks. Men, especially at a younger age, are often more susceptible to cardiovascular diseases. This could be attributed to hormonal differences, lifestyle factors, or even how stress and hypertension manifest differently in different genders.

2. Diabetes\_indexed:

The coefficient for `Diabetes\_indexed` is a significant 0.84045997. This suggests that the presence of diabetes can drastically elevate the log-odds of a person developing heart disease. The high odds ratio of 2.31743268 reinforces this claim, as it indicates that individuals with diabetes have over twice the risk of heart disease compared to those without.

Diabetes has been universally recognized as a major risk factor for cardiovascular diseases. The prolonged presence of high blood sugar levels can cause damage to various organs, including the heart. Additionally, diabetes often comes accompanied by other risk factors such as obesity, high blood pressure, and dyslipidemia, which further increase heart disease susceptibility.

3. Arthritis\_indexed:

The coefficient for `Arthritis\_indexed` stands at 0.69488567, suggesting that having arthritis also contributes positively to the log-odds of developing heart disease. The corresponding odds ratio of 2.00348000 indicates that those with arthritis have a doubled risk.

There's an emerging body of evidence connecting arthritis, especially rheumatoid arthritis, to an increased risk of heart disease. The inflammation characterizing arthritis could have systemic effects, including on the heart and blood vessels, leading to a higher cardiovascular risk.

4. SkinCancer\_indexed:

The `SkinCancer\_indexed` feature has a coefficient of 0.59793052. This positive value implies that individuals diagnosed with skin cancer may have a higher log-odds of also having heart disease. With an odds ratio of 1.81835187, their risk is substantially elevated.

While skin cancer is primarily linked to UV radiation and doesn't directly cause heart disease, individuals with a history of skin cancer might have lifestyles or exposures that increase their cardiovascular risk. There's also the possibility of shared genetic vulnerabilities or the effects of treatments.

5. AgeCategory\_indexed and DrinkingAlcohol\_indexed:

The negative coefficients for both `AgeCategory\_indexed` and `DrinkingAlcohol\_indexed` indicate that as these variables increase, the log-odds of developing heart disease decrease. Specifically, the odds ratios of 0.87493434 and 0.84147043, respectively, suggest a protective effect against heart disease.

Age can be a bit counterintuitive, as older individuals generally have a higher risk. However, the way age categories are indexed might show that certain age groups (like middle-aged individuals) could be at lower risk than the youngest age bracket, depending on the dataset's structure. As for alcohol, moderate alcohol consumption, especially of red wine, has been suggested to have cardiovascular benefits, but excessive drinking has detrimental effects. It's crucial to interpret this in the context of the dataset and the indexed values for alcohol consumption.

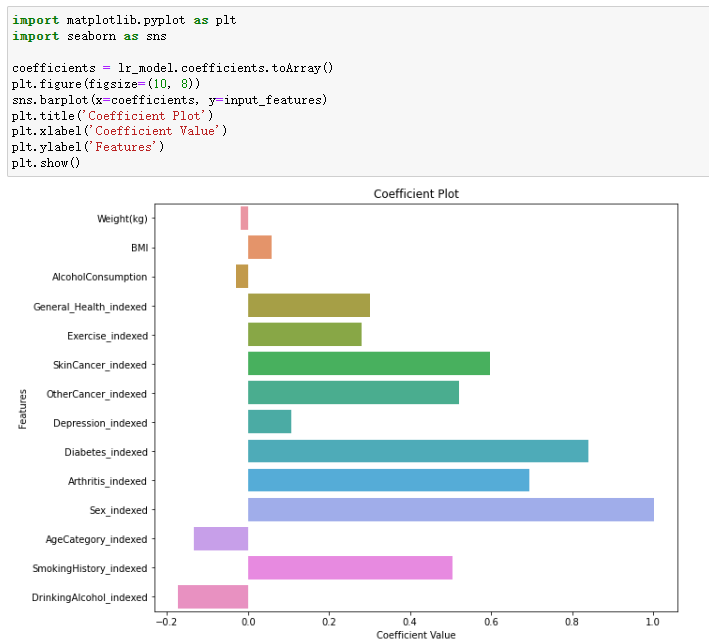
**8.2 Visualize the data, results, models, and patterns**

In this step, I mainly used 5 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. 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.

* Coefficient Plot of Logistic Regression Model



*Figure 74: Coefficient Plot of Logistic Regression Model with Code*

* Odds Ratio Plot of Logistic Regression Model

图表, 条形图

描述已自动生成

*Figure 75:* *Odds Ratio Plot of Logistic Regression Model with Code*

* ROC curve of Logistic Regression Model

图表

低可信度描述已自动生成

*Figure 76: ROC curve of Logistic Regression Model with Code*

* Confusion Matrix of Logistic Regression Model

图表

描述已自动生成

*Figure 77: Confusion Matrix of Logistic Regression Model with Code*

* Predicted Probability Distribution of Logistic Regression Model

图表

描述已自动生成

*Figure 78: Predicted Probability Distribution of Logistic Regression Model with Code*

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

Logistic Regression was harnessed to gain insights into the complex interplay between demographics, health conditions, and lifestyle factors vis-a-vis heart disease prevalence.

1. Significance of Lifestyle Factors: Exercise and dietary habits, while contributing in some manner, weren't the most defining variables in the model. This doesn't render them inconsequential for heart health, but highlights the interwoven complexity of lifestyle and genetic factors. Such results accentuate the need for broader health awareness campaigns to inform the public about diverse risk factors.

2. Alcohol’s Dual Role: A nuanced discovery was the differentiation between the binary variable `DrinkingAlcohol\_indexed` and the continuous one `AlcoholConsumption`. The former, indicating the presence or absence of alcohol consumption, did not significantly correlate with heart disease. However, the magnitude of alcohol intake showed importance. This suggests that the simple act of drinking isn't a determining factor, but the quantity is. This finding underscores the age-old wisdom of moderation. It's imperative for health guidelines to differentiate between occasional drinkers and heavy consumers, as lumping them together might obfuscate true health impacts.

3. The Prime Predictors: Three features - `Sex\_indexed`, `Diabetes\_indexed`, and `Arthritis\_indexed` - prominently dominate the prediction landscape. Notably, the high positive coefficient of `Sex\_indexed` implies that males, when compared to females, are at a higher risk of heart disease. Meanwhile, diabetes, true to its notorious reputation, emerged as a significant heart disease precursor, marked by its high coefficient. Another substantial influencer is arthritis, which often doesn’t get as much limelight in cardiovascular discussions but clearly plays a noteworthy role.

4. Revisiting Established Norms: The role of `SkinCancer\_indexed` in predicting heart disease is fascinating. With a high coefficient, its presence amplifies the chances of heart disease, pointing towards the need for integrated health check-ups, especially for patients diagnosed with certain types of cancers.

5. Subtle yet Vital Players: The negative coefficient for `AgeCategory\_indexed` and `DrinkingAlcohol\_indexed` suggests a counter-intuitive association where increasing age or alcohol consumption could potentially reduce heart disease risk. These are areas ripe for deeper exploration in subsequent studies.

In wrapping up, this analysis has thrown light on the multifaceted factors influencing heart disease. While it doesn't paint the full picture, it certainly adds nuanced strokes, urging medical practitioners and policymakers to adopt a more holistic lens when devising health strategies. Further probing into unexplored variables will undoubtedly refine our understanding in the future.

**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 72.55%, implying that about 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.

表格

描述已自动生成

*Figure 79: 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 0.80, 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 0.80 is significantly greater than 0.5 and approaches 1, it strongly signifies the model's excellent performance in the classification task.

图表, 折线图

描述已自动生成

*Figure 80: ROC curve for Logistic Regression*

2. Assess for decision making

In the complex arena of heart disease determinants, the logistic regression model I employed unveiled pronounced patterns that offer pivotal insights into potential preventative avenues. To facilitate astute decision-making in managing heart disease susceptibility, I hereby detail an assessment derived from these patterns:

Pattern 1: The Ambiguous Role of Lifestyle Factors.

Insight: While certain lifestyle elements like exercise and diet didn't emerge as dominant predictors in this model, they shouldn't be relegated to the sidelines in heart health discussions.

Decision Recommendation: Health campaigns shouldn't solely focus on factors with pronounced coefficients from this singular study. A more holistic approach is required, emphasizing both significant contributors from this analysis and potential influencers like exercise and alcohol consumption found in wider literature. This integrated perspective ensures a comprehensive heart health narrative.

Pattern 2: The Gender Disparity.

Insight: Males face a palpably higher heart disease risk compared to females, evidenced by the positive coefficient of `Sex\_indexed`.

Decision Recommendation: Medical practitioners must be acutely aware of this gender-based risk gradient. Early interventions, lifestyle counselings, and frequent screenings should be emphasized more for male patients, given their inherent susceptibility.

Pattern 3: Diabetes - A Notorious Contributor.

Insight: Diabetes, given its prominent coefficient, stands out as a major precursor to heart disease, reinforcing its notoriety.

Decision Recommendation: In light of this, healthcare frameworks should instigate heart screenings as an integral component of diabetic care. Additionally, equipping diabetic patients with knowledge about this interrelation and promoting diligent diabetes management is paramount.

Pattern 4: The Underestimated Role of Arthritis.

Insight: Arthritis, despite often being overshadowed in cardiovascular discourses, emerged as a substantial influencer, underscoring a potential link between inflammation and heart health.

Decision Recommendation: Given this newfound correlation, individuals diagnosed with arthritis should be considered for regular heart screenings. There's a need to sensitize both patients and medical professionals about this less-discussed association, ensuring arthritis management encompasses potential heart risks.

Pattern 5: Cancer's Surprising Relevance.

Insight: The role of `SkinCancer\_indexed` in heart disease prediction is intriguing, suggesting an interconnected health landscape.

Decision Recommendation: Medical screenings should adopt a more interconnected stance. Those diagnosed with specific cancers, like skin cancer, should also undergo cardiovascular assessments to ascertain any latent risks, fostering a more integrative health surveillance.

Concluding Decision Implications:

The delineated patterns, mined from the depths of logistic regression, are invaluable compasses for heart health strategies, awareness drives, and tailored medical interventions. By pinpointing and actively engaging with these identified risk paradigms, medical professionals and policymakers can pioneer proactive approaches, catering to the distinct needs of various risk cohorts, fortifying the heart health of the wider populace.

**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.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

*Figure 81: Oversampling dataset to project data*

As illustrated in the following figure, I compared the accuracy of the Logistic Regression model after balancing the data using both the Undersampling and Oversampling methods. Employing the Undersampling method to balance the data resulted in a better peformance in the model's accuracy and AUC. So, I keep the undersampling method in balancing data.

表格

描述已自动生成表格

描述已自动生成

*Figure 82: 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 80/20. 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 83: Changing data split ratio to 80/20*

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

表格

描述已自动生成表格

描述已自动生成

*Figure 84: 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.