Report

|  |  |  |
| --- | --- | --- |
| CWID | Name | Percent rate |
| A20563427 | gengxiangChen | 100% |

1. Dataset Description

Data Source: Kaggle (heart\_disease\_risk\_dataset\_earlymed.csv)

Sample Size: The total number of rows is not specified. It includes the following features for each patient:

Symptoms (binary variables, 0/1): chest pain (Chest\_Pain), shortness of breath (Shortness\_of\_Breath), fatigue (Fatigue), etc.

Risk factors (binary variables, 0/1): hypertension (High\_BP), hypercholesterolemia (High\_Cholesterol), diabetes (Diabetes), etc.

Demographic statistics: Gender (0/1), Age (numeric).

Target variable: Heart disease risk (Heart\_Risk, 0 = low risk, 1 = high risk).

Key observations:

Age range: 20 - 84 years old. Age is positively correlated with the risk of heart disease.

Among high-risk groups, risk factors such as hypertension, smoking and obesity are widespread.

2. Visualization Presentation and Method Description

2.1 Bar Chart: The Relationship between Symptom Frequency and Heart Disease Risk

Code File: 2\_bar\_chart.py

Method:

Use pandas to group by Heart\_Risk and calculate the average frequency of each symptom.

Draw a stacked bar chart through matplotlib to compare the symptom distribution between high-risk and low-risk groups.

Result:

Among high-risk groups, the frequency of symptoms such as chest pain, shortness of breath, and fatigue is significantly higher (Figure 1).

Bar chart

Figure 1: The Relationship between Symptom Frequency and Heart Disease Risk

2.2 Pie Chart: The Impact of Unhealthy Habits on Heart Disease

Code File: Project report.docx Embedded Code

Method:

Use Plotly to generate a pie chart to show the proportion of smoking, drinking, and lack of exercise in the risk of heart disease.

Result:

The contributions of the three types of bad habits to the risk of heart disease are equal (each accounting for 33.3%), indicating the need for comprehensive intervention (Figure 2).

Pie chart

Figure 2: Proportion analysis of bad living habits

2.3 Heatmap: Correlation between Symptoms and Risk Factors

Code file: 3\_heatmap.py

Method:

Calculate the feature correlation matrix using seaborn and generate a heatmap.

The color mapping (coolwarm) represents the strength of correlation, with red indicating a strong positive correlation and blue a strong negative correlation.

Age (r = 0.61), hypertension (r = 0.41), and smoking (r = 0.40) are highly correlated with the risk of heart disease (Figure 3).

There are correlations among symptoms (such as chest pain and fatigue, r = 0.36).

Heatmap

Figure 3: Heatmap of the Correlation between Symptoms and Risk Factors

2.4 Three-dimensional Scatter Plot: The Relationship between Age, Number of Symptoms and Heart Disease Risk

Code File: Project report.docx Embedded Code

Method:

Use Plotly to create a 3D graph, with age on the X-axis, heart disease risk on the Y-axis, and the number of symptoms on the Z-axis.

The color gradient indicates the risk level (blue = low risk, yellow = high risk).

Individuals with advanced age and a large number of symptoms have the highest risk, while younger groups have a lower risk (Figure 4).

Three-dimensional graph

Figure 4: Three-dimensional scatter plot analysis

3. Tools and Libraries Used

Data processing: pandas (data loading, group calculation).

Visualization:

Matplotlib & Seaborn: Static Charts (Bar Charts, Heatmaps).

Plotly: Interactive charts (pie charts, 3D charts).

Dependency Management: Integrating code and visualization output through Jupyter Notebook.

4. Results Analysis and Conclusions

Key Risk Factors: Age, hypertension, and smoking are the main risk factors and should be prioritized for intervention.

Symptom correlation: The concurrent occurrence of chest pain and fatigue may indicate a high risk. It is recommended that clinical monitoring be conducted in combination.

Multidimensional analysis: A three-dimensional graph shows that advanced age and the superposition of multiple symptoms significantly increase the risk, supporting a stratified medical treatment strategy.

5. Limitations and Directions for Improvement

Data Limitations: Binary classification variables cannot reflect the severity of risk factors (such as obesity levels).

Technical improvement: Machine learning models (such as logistic regression) can be introduced for risk prediction.