PROJECT WRITEUP: “Examining Factors Responsible for Heart Attacks”

DESCRIPTION:

Objective:

The objective of the project is to perform comprehensive Exploratory Data Analysis (EDA) on a healthcare dataset related to cardiovascular diseases. This includes data inspection, handling missing values and duplicates, exploring the relationships between various factors and the occurrence of heart attacks, building a baseline predictive model using logistic regression, and creating informative visualizations in Tableau to understand the attributes associated with diseased and healthy individuals. The goal is to gain insights into the factors influencing the occurrence of cardiovascular diseases and develop a predictive model for heart attack detection.

Problem Statement:

Read the information given below and refer to the data dictionary provided separately in an excel file to build your understanding.

Cardiovascular diseases are the leading cause of death globally. To identify the causes and to develop a system to predict heart attack in an effective manner is necessary. The presented data has all the information about all the relevant factors that might have an impact on heart health. The data needs to be explained in detail for any further analysis.

Domain: Healthcare

Content: Dataset: (“Heart Attacks Datasets” : Click here to download dataset)

1. Preliminary analysis:
    - Perform preliminary data inspection and report the findings as to the structure of the data, missing values, duplicates, etc.
    - Based on the findings from the previous question remove duplicates (if any), treat missing values using an appropriate strategy.
2. Prepare an informative report about the data explaining the distribution of the disease and the related factors. You could use the below approach to achieve the objective
    - Get a preliminary statistical summary of the data. Explore the measures of central tendencies and the spread of the data overall.
    - Identify the data variables which might be categorical in nature. Describe and explore these variables using appropriate tools e.g., count plot
    - Study the occurrence of CVD across Age.
    - Study the composition of overall patients w.r.t. Gender.
    - Can we detect a heart attack based on anomalies in the Resting Blood Pressure of the patient?
    - Describe the relationship between Cholesterol levels and our target variable.
    - What can be concluded about the relationship between peak exercising and the occurrence of a heart attack.
    - Is thalassemia a major cause of CVD?
    - How are the other factors determining the occurrence of CVD?
    - Use a pair plot to understand the relationship between all the given variables.

3. Build a baseline model to predict using a Logistic Regression and explore the results.

Project Task: Week 1

Importing, Understanding, and Inspecting Data:

1. Perform preliminary data inspection and report the findings as the structure of the data, missing values, duplicates, etc.
2. Based on the findings from the previous question, remove duplicates (if any) and treat missing values using an appropriate strategy.
3. Get a preliminary statistical summary of the data. Explore the measures of central tendencies and the spread of the data overall.

Performing EDA:

4. Identify the data variables which might be categorical in nature. Describe and explore these variables using appropriate tools. For example: count plot.
5. Study the occurrence of CVD across different ages.
6. Can we detect heart attack based on anomalies in resting blood pressure of the patient?
7. Study the composition of overall patients w.r.t. gender.

Project Task: Week 2

Performing EDA and Modeling:

1. Describe the relationship between cholesterol levels and our target variable.
2. What can be concluded about the relationship between peak exercising and occurrence of heart attack?
3. Is thalassemia a major cause of CVD? How are the other factors determining the occurrence of CVD?
4. Use a pair plot to understand the relationship between all the given variables.
5. Perform logistic regression, predict the outcome for test data, and validate the results by using the confusion matrix.

Dashboarding:

6. Visualize the variables using Tableau to create an understanding for attributes of a Diseased vs. a Healthy person.
7. Demonstrate the variables associated with each other and factors to build a dashboard.

Findings:

- There are no NaN values in the data.
- There are certain outliers in all the continuous features.
- The data consists of more than twice the number of people with `sex` = 1 than `sex` = 0.
- There is no apparent strong linear correlation between continuous variable according to the heatmap.
- It is intuitive that elder people might have higher chances of heart attack but according to the distribution of `age` group w.r.t. `output`, it is evident that this isn't the case.
- According to the distribution of `thalach` group w.r.t. `output`, people with higher maximum heart rate achieved have higher chances of heart attack.
- According to the distribution of `oldpeak group` w.r.t. `output`, people with lower oldpeak achieved have higher chances of heart attack.
- People with Non-Anginal chest pain, that is with `cp` = 2 have higher chances of heart attack.
- People with 0 major vessels, that is with `ca` = 0 have high chance of heart attack.
- People with `sex` = 1 have higher chance of heart attack.
- People with `thal` = 2 have much higher chance of heart attack.
- People with no exercise induced angina, that is with `exng` = 0 have higher chance of heart attack.

Thank you!