Mathematics For Computer Science Engineers
UE23MA242A

Teaching Assistants : Archishman VB, Suchir M Velpanur, Neha Bhaskar

Heart disease Case Study

Introduction

The two sections below, Background and Case Study provide context for the data science hackathon. This exercise will allow you to test your skills in using the Python programming language to effectively explore the characteristics of a dataset and analyze the features using descriptive statistics such as summary statistics, tables, and graphs. Happy coding!

Background

Heart disease, or cardiovascular disease, is a leading cause of death globally, encompassing various conditions that impair heart function, such as coronary artery disease, heart failure, and arrhythmias. It often results from the buildup of fatty deposits in the arteries (atherosclerosis), leading to restricted blood flow, high blood pressure, and increased risk of heart attacks. Common risk factors include high blood pressure, high cholesterol, smoking, diabetes, obesity, and a sedentary lifestyle. By analyzing heart disease data, researchers aim to understand these risk factors, predict heart disease likelihood, and support targeted prevention and treatment strategies to improve public health outcomes.

Case Study

The dataset comprises information on 920 individuals, including their demographics, lifestyle factors, family medical history, and clinical measurements. The goal is to analyze this dataset to identify correlations between different variables and gain insights into the factors influencing the risk of developing heart disease. Using data analysis techniques, including machine learning algorithms, the researcher aims to predict potential risk factors and contribute to public health interventions for heart disease prevention.

Dataset Description

Dataset Features:

1. id (Unique id for each patient)

2. age (Age of the patient in years)
3. origin (place of study)
4. sex (Male/Female)
5. cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
6. trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
7. chol (serum cholesterol in mg/dl)
8. fbs (if fasting blood sugar > 120 mg/dl)
9. restecg (resting electrocardiographic results)
   -- Values: [normal, stt abnormality, lv hypertrophy]
10. thalach: maximum heart rate achieved
11. exang: exercise-induced angina (True/ False)
12. oldpeak: ST depression induced by exercise relative to rest
13. slope: the slope of the peak exercise ST segment
14. ca: number of major vessels (0-3) colored by fluoroscopy
15. thal: [normal; fixed defect; reversible defect]
16. num: the predicted attribute. Diagnosis of heart disease, often used as the target variable indicating heart disease presence (values may range from 0–4, with higher values indicating greater disease severity).

Problem Set

Unit-1

1. Classify the features in the Heart Disease dataset into their appropriate data types (ordinal, nominal, interval, or ratio). Provide a rationale for each classification.
2. A summary statistic provides a numerical summary of a specific feature within the dataset.There are two commonly used categories of summary statistics: those that indicate the central tendency and those that indicate the spread of the data. Identify the most appropriate measure of central tendency for each attribute in the dataset and state its corresponding value. Additionally, calculate the standard deviation and range of values for each column.
3. Identify and describe any data quality issues or inconsistencies within the Heart Disease dataset. What steps would you take to clean and preprocess the data to ensure its accuracy and reliability for further analysis?
4. Using a histogram and box plot, assess the presence of outliers in the 'age' and 'chol' variables. Describe the visualizations, identify any potential outliers, and explain how you determined their presence or absence.
5. What actions would you take to resolve the presence of outliers? Visualize the changes using box plots and histograms.

6. Examine the normal probability plot (Q-Q plot) for the 'chol' variable in the Heart Disease dataset. Based on the shape and trend of the plot, what conclusions can be drawn? Provide a rationale for your conclusions.
7. Calculate the correlation between 'Age' and all the other numerical variables (e.g., 'trestbps' or 'thalach''). Set a correlation threshold and create a heatmap to visualize the relationships.
8. Generate a pairplot that includes the variables 'age', 'chol', and 'thalach' while using 'num' as the hue in the dataset. What insights can be gained from the pairplot, and how does it help in visualizing the relationships between the given features and heart disease prevalence?

Unit-2

9. Is there a statistically significant difference in the mean cholesterol levels between individuals with heart disease (num > 0) and those without heart disease (num = 0)? Formulate the null and alternative hypotheses, then conduct a T-test to examine this relationship. Additionally, plot a histogram to visualize the cholesterol distribution for each group. Use a significance level of 0.05.

10. Calculate the margin of error to quantify the precision of the analysis done previously and infer what you can deduce from the results.

Unit-3

11. Perform a linear regression analysis to predict the variable chol (cholesterol level) using the features age, trestbps (resting blood pressure), thalch (maximum heart rate achieved), oldpeak (ST depression induced by exercise), ‘ca’ (number of major vessels), and ‘num’ (Diagnosis of heart disease). Validate the model using performance metrics like RMSE, MSE, and R-squared values to assess the model's effectiveness.

12. To improve the model’s predictive power in assessing heart disease risk, consider what additional features could be engineered from the existing data. Explain what new variables you would create and why they could enhance the predictive accuracy of the model.