Mathematics For Computer Science Engineers

UE23MA242A

Teaching Assistant: **SHRUTI C and SRAGVI ANIL SHETTY**

Problem Set: Hepatitis C Analysis Case Study

Hepatitis C Dataset Description

The Hepatitis C dataset contains detailed medical and demographic information on patients with Hepatitis C or at risk for the disease. This dataset captures various attributes, including patient age, biochemical test results, and diagnostic information, allowing for in-depth analysis of disease progression, potential risk factors, and diagnostic patterns.

Background

Hepatitis C is a liver infection caused by the Hepatitis C virus (HCV), which can lead to severe liver damage if untreated. The disease is typically spread through exposure to infected blood. This dataset can be instrumental for healthcare researchers and public health analysts to explore patterns and correlations in Hepatitis C diagnosis and severity, enabling a better understanding of disease progression and risk factors.

Dataset Overview

The dataset includes the following primary variables:

- **Age**: The age of the patient.
- **Sex**: The gender of the patient, indicated as Male or Female.
- **ALB**: Albumin levels in the blood, an important liver function marker.
- **ALP**: Alkaline phosphatase levels, another indicator of liver function.
- **ALT**: Alanine aminotransferase levels, elevated in liver disease.
- **AST**: Aspartate aminotransferase levels, also a key marker in liver health.
- **BIL**: Bilirubin levels, with elevated levels often seen in liver disease.
- **CHE**: Cholinesterase levels, which can be affected by liver conditions.
- **CHOL**: Cholesterol levels, measured to monitor general health.
- **CREA**: Creatinine levels, a kidney function marker but also monitored in liver disease.
- **GGT**: Gamma-glutamyl transferase levels, elevated in liver conditions.
- **PROT**: Total protein levels, important in assessing overall health.

Objective

This dataset provides an opportunity to perform a thorough statistical and exploratory analysis, focusing on understanding the factors contributing to Hepatitis C diagnosis, progression, and severity. Through analysis, healthcare practitioners can gain valuable insights into patient characteristics, potential risk factors, and key biochemical indicators associated with Hepatitis C, aiding in early detection and treatment planning.

Unit 1: Exploratory Data Analysis & Preprocessing

1. **Uncovering Feature Types (Exploring Data Types and Value Ranges)**
   Identify the data types of each feature and discuss whether these data types are appropriate. Identify the first five and the last five records in the dataset.
2. **Data Quality Insights**
   Identify and discuss any inconsistencies or quality concerns in the dataset. Outline a strategy to clean and refine the data for dependable analysis.
3. **Statistical Summary of Numerical Attributes**
   Calculate key summary statistics (mean, median, standard deviation, and range) for each numerical feature. Explain which measures of central tendency and variability best represent each feature.
4. **Visual Data Exploration**
   Generate histograms and box plots for variables like Age and AST levels.
   - i) Describe the shapes and trends you observe.
   - ii) Identify potential outliers.
   - iii) Bonus: Make adjustments based on observations and re-plot for comparison.
5. **Outlier Management and Comparison**
   Describe your approach to handling outliers, with visualizations illustrating the data before and after adjustments.
6. **Normality Check via Q-Q Plot**
   Construct a Q-Q plot for ALT levels. Analyze the plot’s shape and discuss any insights about the distribution.
7. **Correlation Mapping**
   Examine correlations between Age and other numerical attributes. Identify which feature has the strongest relationship with *ALT*.

Unit 2: Hypothesis Testing & Inference

8. **Testing for Differences in Viral Load Across Hepatitis Stages**
   Define a null and alternative hypothesis to evaluate if there’s a statistically significant difference in *ALT* across hepatitis category. Select an appropriate test to assess this relationship.
9. **Calculating Precision: Margin of Error**
   Determine the margin of error for your hypothesis test results. Discuss what this margin reveals about the precision and reliability of your findings.

Unit 3: Predictive Modeling & Advanced Analysis

10. **Testing for Differences in ALT Levels Across Hepatitis Stages**
    Define a null and alternative hypothesis to evaluate if there’s a statistically significant difference in ALT (Alanine Aminotransferase) levels across different categories in the dataset (e.g., blood donors, hepatitis patients). Choose an appropriate statistical test to assess this relationship.
11. **Exploring Categorical Relationships with Chi-Square Testing**
    Perform a Chi-Square test to assess the relationship between Category (e.g., blood donor, hepatitis status) and Sex. Interpret the test results and discuss whether there is a statistically significant association between these variables.