Mathematics For Computer Science Engineers UE23MA242A

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

Diabetes 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

Diabetes mellitus, commonly referred to as diabetes, is a chronic metabolic disorder characterized by elevated blood sugar levels over a prolonged period. The condition results from either insufficient insulin production by the pancreas or the body's inability to effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar (glucose) and facilitates its absorption into cells for energy

Case Study

Consider a scenario where a researcher is studying the trends and patterns in diabetes prevalence and its influencing factors within a particular population. They gather a detailed dataset, encompassing individual demographics, lifestyle habits, genetic tendencies, medical histories, and metrics such as glucose levels and insulin production. The objective is to analyze this extensive data to uncover correlations among variables and understand the factors contributing to diabetes onset. By employing analytical methods, including machine learning algorithms, the researcher aims to identify potential risk factors, evaluate the success of current prevention strategies, and provide valuable insights to support the development of targeted interventions, ultimately enhancing public health efforts to combat diabetes.

Dataset Description

The dataset comprises information on 768 individuals, including the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), a diabetes pedigree function reflecting genetic influence, age, and an outcome variable indicating the presence (1) or absence (0) of diabetes

1. Pregnancies: Number of times pregnant.
2. Glucose: Plasma glucose concentration after 2 hours in an oral glucose tolerance test.
3. BloodPressure: Diastolic blood pressure (mm Hg).
4. SkinThickness: Triceps skinfold thickness (mm).
5. Insulin: 2-Hour serum insulin (mu U/ml).
6. BMI: Body mass index (weight in kg/(height in m)^2).
7. DiabetesPedigreeFunction: Diabetes pedigree function, a measure of the diabetes genetic influence.
8. Age: Age in years.
9. Outcome: Binary variable indicating whether a person has diabetes (1) or not (0).

Problem Set

Unit-1

1. Classify the features in the Diabetes 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 Diabetes 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 'DiabetesPedigreeFunction' 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.
Hint: Use boxplot and histogram
6. Examine the normal probability plot (Q-Q plot) for the 'DiabetesPedigreeFunction' variable in the Diabetes 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., Pregnancies or BloodPressure). Set a correlation threshold and create a heatmap to visualize the relationships.
8. Generate a pairplot that includes the variables 'Age’, ‘DiabetesPedigreeFunction', and 'Pregnancies' while using 'Outcome' 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 if someone is diabetic or not?

Unit-2

9. Is there a significant difference in the mean glucose levels between individuals with diabetes (Outcome = 1) and those without diabetes (Outcome = 0)?Formulate null and alternative hypotheses and employ a T-test to examine this relationship.Plot a histogram to visualize the results. Assume significance level as 0.05.

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

Unit-3

11. Perform a linear regression to predict 'DiabetesPedigreeFunction' using 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'Age'. Validate the model using performance metrics like RMSE, MSE and R-squared value.
12. What additional features could be engineered from the existing data in the diabetes dataset to improve the prediction of diabetes outcomes?