← Go Back to Foundations of Data Science

Course Content

Problem Statement - Pima Diabetes Analysis

Submission type	:	File Upload
Due Date	:	Mar 27, 7:30 PM
Total Marks	:	40
Available from	:	Mar 16, 3:30 PM

The due date for this assignment has passed.

Description

Welcome to the project on Foundations of Data Science. In this project, we aim to analyze diabetes data and address some important business problems/questions.

- This project is focused on Exploratory Data Analysis
- A solution notebook is shared for the analysis
- Many parts of the solution notebook are omitted and replaced with questions. You are expected to fill in the gaps as per the instructions/questions.

Problem Statement:

Diabetes is one of the most frequent diseases worldwide and the number of diabetic patients are growing over the years. The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes.

A few years ago research was done on a tribe in America which is called the Pima tribe (also known as the Pima Indians). In this tribe, it was found that the ladies are prone to diabetes very early. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients were females at least 21 years old of Pima Indian heritage. **Here, we are analyzing different aspects of Diabetes in the Pima Diabetes Analysis by doing Exploratory Data Analysis.**

Data Dictionary:

Below is the attribute information:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Blood pressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skinfold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml) test
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: A function that scores the likelihood of diabetes based on family history
- Age: Age in years
- Outcome: Class variable (0: the person is not diabetic or 1: the person is diabetic)

Best Practices for Notebook

- The notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
- The notebook should be run from start to finish in a sequential manner before submission.
- It is preferable to remove all warnings and errors before submission.
- The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb).

Submission Guidelines

1. The submission should be a well-commented Jupyter notebook [format - .html]
2. Any assignment found copied/ plagiarized with other groups will not be graded and awarded zero marks.
3. Please ensure timely submission as any submission post-deadline will not be accepted for evaluation.
4. Submission will not be evaluated if
    1. the code blocks are not executed,
    2. it is submitted post-deadline, or,
    3. more than 1 file is submitted.

Happy Learning!

Scoring guide (Rubric) - Pima Diabetes Analysis

Criteria	Points
Q 1. Import the necessary libraries and briefly explain the use of each library	3

Criteria	Points
Q 2. Read the given dataset	1
Q3. Show the last 10 records of the dataset. How many columns are there?	1
Q4. Show the first 10 records of the dataset	1
Q5. What do you understand by the dimension of the dataset? Find the dimension of the `pima` dataframe.	1
Q6. What do you understand by the size of the dataset? Find the size of the `pima` dataframe.	1
Q7. What are the data types of all the variables in the data set?	2
Q8. What do you mean by missing values? Are there any missing values in the `pima` dataframe?	2
Q9. What does summary statistics of data represents? Find the summary statistics for all variables except 'Outcome' in the `pima` data? Take one column/variable from the output table and explain all the statistical measures.	3
Q 10. Plot the distribution plot for the variable 'BloodPressure'. Write detailed observations from the plot.	2

Criteria	Points
Q 11. What is the 'BMI' for the person having the highest 'Glucose'?	1
Q 12. Q 12.1 What is the mean of the variable 'BMI'? 12.2 What is the median of the variable 'BMI'? 12.3 What is the mode of the variable 'BMI'? 12.4 Are the three measures of central tendency equal?	3
Q 13. How many women's 'Glucose' level is above the mean level of 'Glucose'?	1
Q 14. How many entries (women) have their 'BloodPressure' equal to the median of 'BloodPressure' and their 'BMI' less than the median of 'BMI'?	2
Q 15. Below is the pairplot of variables 'Glucose', 'SkinThickness' and 'DiabetesPedigreeFunction'. Write you observations from the plot.	4
Q 16. Plot the scatterplot between 'Glucose' and 'Insulin'. Write your observations from the plot.	2
Q 17. Plot the boxplot for the 'Age' variable. Are there outliers?	2
Q 18. Plot histograms for variable Age to understand the number of women in different Age groups given that they have diabetes or not. Explain both histograms and compare them.	3
Q 19. What is Inter Quartile Range of all the variables? Why is it used? Which plot visualizes the same?	2

Criteria	Points

Help