Datathon: Diabetes-case-study

Data Information

About Dataset:

Pregnancies: To express the Number of pregnancies
Glucose: To express the Glucose level in blood
BloodPressure: To express the Blood pressure measurement
SkinThickness: To express the thickness of the skin
Insulin: To express the Insulin level in blood
BMI: To express the Body mass index
Diabetes Pedigree Function: To express the Diabetes percentage
Age: To express the age
: To express the final result 1 is Yes and 0 is No

Answer the questions given in different sections

Data Description Statistics

a. What is the structure (shape) of the dataset?

b. Show the min, max, and mean of Glucose, …?(Hint: Pandas function that shows for all the columns at once is available.)

Pre-processing

a. Check for NULLs/Duplicates. Drop attributes with more than 20% data missing.

b. Fill remaining NULLs with mode values

c. Are there categorical columns ?

Data Visualization

a. Make Histogram, and whisker plots to understand the meaning of the encoding.

Hypothesis Testing

a. Perform correlation Analysis.

Modelling

a. Build a Linear Regression Model.

1. MAE, MSE, and RMSE results.
2. Linear Regression R2 score.