Diabetes Prediction – Basic EDA & KNN Model

By Mushfiqur Rahman

About the Dataset:

The dataset has 768 observations and 9 variables (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome). All the variables are integer or float type. The dataset has no null values or duplicated values.

Objective:

Our objective is to predict whether the respondent has diabetes or not using KNN Model. If the respondent has diabetes, the outcome variable will be 1, otherwise it will be 0.

Exploratory Data Analysis (EDA):

We can see from the graphs that, the number of respondents not having diabetes (Outcome=0) is larger than that of those who have diabetes. More respondents were not pregnant or got pregnant only once or twice. In most cases, their blood pressure is within 55-80. Some scatter plot along with a heatmap have been showcased to determine the relationship between several variables.

KNN Model:

Here, outcome variable is the dependent variable and remaining variables belong to the dependent variable. Test data size is 20% of the whole dataset. From evaluation metrics, the mean absolute error and mean squared error are 0.27. From the classification report, we see that the accuracy is 73%.