Problem Statement - Diabetes Risk Prediction

Case Study: Diabetes Risk Prediction

Context:

Diabetes is one of the most frequent diseases worldwide and the number of diabetic patients is 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.

Individuals with diabetes face a risk of developing some secondary health issues such as heart diseases and nerve damage. Thus, early detection and treatment of diabetes can prevent complications and assist in reducing the risk of severe health problems. Even though it's incurable, it can be managed by treatment and medication.

Researchers at the Bio-Solutions lab want to get a better understanding of this disease among women and are planning to use machine learning models that will help them to identify patients who are at risk of diabetes.

You as a data scientist at Bio-Solutions have to build a classification model using a dataset collected by the "National Institute of Diabetes and Digestive and Kidney Diseases" consisting of several attributes that would help to identify whether a person is at risk of diabetes or not.

Objective:

The data-set aims to answer the following key questions:

What are the different factors that can help in identifying whether a person is at risk of diabetes or not?

Can we build a model to identify persons who are at risk of diabetes? What should be the metric of choice to evaluate such a model?

Attribute Information:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)

- SkinThickness: Triceps skinfold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- Pedigree: Diabetes pedigree function - A function that scores likelihood of diabetes based on family history.
- Age: Age in years
- Class: Class variable (0: the person is not diabetic or 1: the person is diabetic)

Learning Outcomes:

- Exploratory Data Analysis
- Preparing the data to train a model
- Training and understanding of data using ensemble models
- Model evaluation

Steps and Tasks:

- Import Libraries and Load Dataset
- Overview of data
- Data Visualization
- Data preparation
- Choose Model, Train, and Evaluate
- Conclusion