

**Problem:** Diabetes is a chronic medical condition that affects how the body processes glucose. The purpose of this project is to come up with an accurate model for the diagnostic prediction of diabetes in patients, based on certain diagnostic measurements. The dataset focuses on females aged 21 and older of Pima Indian heritage and will leverage machine learning techniques to make these predictions.

The dataset used in this analysis is a health dataset containing information related to factors that may influence the likelihood of a person developing diabetes. This dataset consists of 768 observations with 8 independent variables and 1 dependent variable, which indicates whether a person has been diagnosed with diabetes or not.

**Client:** Biotech Companies, e.g. Atrogi, DiogenX, Fractyl Health, etc.

**Sample Data:** <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

**Methodology:**

- Obtain and wrangle the data from the Kaggle
  - Clean data sources; the dataset contains various medical predictor variables along with a single target variable, known as Outcome. The predictor variables include factors such as the patient's number of pregnancies, BMI, insulin levels, age, and more.
- Exploratory Data Analysis:
  - How do the predictor variables given in the dataset relate to the occurrence of diabetes in patients?
  - What are current predictor models? Can they be improved?
  - What machine learning method will best predict diabetes in a patient?
- Machine Learning
  - Perform several machine learning techniques, such as
  - Am I able to create a machine learning model to accurately predict diabetes based on the given predictor variables?
- Document, interpret and summarize results

**Deliverables:**

- Jupyter Notebook for detailed explanation and analysis
- PPT
- Github repository