INTRODUCTION

The Pima Indians diabetes dataset is a well-known dataset used for analysing and researching the prediction of diabetes. It contains medical records for patients from the Pima Indian population and includes various measures of patients' health, such as

- Past pregnancies
- Blood pressure
- Glucose level
- The thickness of the skin fold of the triceps
- Insulin level
- BMI Index (Body Mass Index)
- Family history of diabetes
- Age

The response variable of interest is binary and indicates whether a patient has diabetes or not. Statistical and data analysis can be used to identify patterns and relationships in the Pima Indians diabetes dataset that can be used to predict the likelihood of a patient having diabetes. This can be accomplished using various statistical and machine learning methods, such as **logistic regression, Classification trees .**

In this project we use the machine learning methods as

- Classification trees
- Logistic regression

Statistical methods as ,

- Data exploration like,
  - Data distribution inferences.
  - Univariate Data analysis.
  - Two-sample t-test
- Data Correlation Analysis
- Basic General Model

One common approach is to fit a logistic regression model to the data, where the predictor variables are the patient characteristics (such as age, blood pressure, insulin levels, and BMI) and the response variable is whether or not the patient has diabetes.

The coefficients from the logistic regression model can then be used to make predictions about new patients and to understand the importance of each predictor variable in predicting diabetes.

Another approach is to use machine learning techniques, such as decision trees , which are able to handle non-linear relationships and interactions between predictor variables. These methods can also provide insights into the important predictor variables, but they do so in a different way than logistic regression.

Regardless of the approach used, it is important to perform model diagnostics to ensure that the results are valid. This can include checking the residuals for **normality,** checking the model's ability to predict the response variable using techniques such as the **receiver operating characteristic curve**, and assessing the performance of the model on an independent validation dataset.

The Dataset Contains following variables and their description,

Variable	Description	Norm
Pregnancies	Number of pregnancies	
Glucose	Plasma glucose concentration after 2 hours in the oral glucose tolerance test	70-140 mg/d l
BloodPressure	Diastolic blood pressure	50-90 mm Hg
SkinThickness	Triceps skinfold thickness	
Insulin	2-hour serum insulin (µU / ml)	do 140 mU/ml, 140-199 it is pre-diabetes
BMI	Body mass index (weight / (height ^ 2))	18,5–24,9
Diab.Pedigree.Fun	Diabetes prevalence rate in relatives	
Age	Age	
Outcome	Does the person have diabetes?	

Overall, the Pima Indians diabetes dataset provides a valuable opportunity for researchers and practitioners to develop and test predictive models for diabetes, and to better understand the relationships between predictor variables and the risk of developing diabetes. .