# Healthcare

# **Case Study Final Report:**

The major aim of this project is to build a machine-learning model in order to predict whether a person is diabetic or not.

The major variables provided with the dataset are listed below:

* Pregnancies
* Glucose
* Blood Pressure
* Skin Thickness
* Insulin
* BMI
* Diabetes Pedigree Function
* Age and
* Outcome

Here the **independent variables** are Pregnancy, Glucose, Blood Pressure, Insulin, BMI, and Age and the **dependent variable** is Outcome.

The Outcome is based upon only two factors **yes** or **no**. Yes i.e., 1 which denotes that the person is diabetic, and no i.e., 0 which denotes the person is nondiabetic.

The problem statement clearly stratifies the rules of **binary classification** supervised machine learning algorithm.

The experiment was conducted on the same path following the rule and metrics of the classification algorithm to reach the precise result.

The following methods were adopted in this project.

## Part I: Data Exploration

As a part of data exploration various measures were taken in order to understand and study the structure of data, like

1. **Head**: top five records
2. **Tail**: Bottom five records
3. **Info**: concise information about a Data Frame
4. **Shape**: information on no. of Rows and columns
5. **Describe**: statistical summary information for each numeric column

As per the descriptive analysis, we saw that the minimum value is 0 for columns such as Pregnancies, Glucose, SkinThickness, Insulin, BMI, etc. which is practically impossible so we have to treat this as a null value (**NAN**) hence replaced 0 with its median value of every variable.

## Part II: Visualizations

As a part of visualizing for the various variables impacting the significance of diabetes, we have plotted the **histogram** for variables like Pregnancy, Glucose, Blood Pressure, Insulin, BMI, and Age.

Plotted **bar diagram** to study the value counts of the data type int and float and also diabetic and non-diabetic.

With the initial study, we found that there’s a **high imbalance** in the outcome and decided to use SMOTE to balance them moving forward.

Studied the depth of every variable plotting the scattered **pair-plot** of every variable.

We have also studied the correlation of every variable using **heatmap** with one another and tested the same with code as well.

### Part III: Feature Scaling

As a part of feature scaling, we segregated the variables into independent and dependent variables and used a **standard scaler** for the same.

**Standardization** is the process of scaling a feature to have a mean (average) of 0 and a standard deviation of 1.

Secondly, we have used SMOTE to balance the data that we found to be imbalanced earlier.

**SMOTE** stands for "Synthetic Minority Over-sampling Technique." It is a technique used in machine learning and data analysis to address the class imbalance problem in classification tasks

## Part IV: Data Modeling

As a part of data modeling conducted a detailed experiment with every ML Classification model like **KNN**, **logistic Regression**, **Decision Tree**, **Random Forest**, **SVM** & **NB** in order to test multiple algorithms to see which one works best for this specific problem.

Evaluated metrics like **test accuracy**, **training accuracy**, **AUC**, and **confusion matrices** for each model.

We observe an instance of **overfitting** in the data, which prompts us to **consider cross-validation**. In this case, we opt for **K-Fold cross-validation**.

In this scenario, it's worth noting that K-Nearest Neighbors (**KNN**) exhibit lower accuracy compared to Support Vector Machines (**SVM**) and **Random Forest**. As a result, we decided to implement **K-Fold cross-validation** for the Random Forest model.

Finally, also used the **XGBoost** library (xgboost) to build a binary classification model.

In summary, we have determined that **SVM** exhibits **superior accuracy** and **AUC**. Additionally, SVM stands out as the appropriate model for this dataset due to its absence of overfitting issues.

In this way, a **supervised classification machine learning model** was designed and implemented which is good at predicting whether a person is diabetic or not based on various variables like Pregnancy, Glucose, Blood Pressure, Insulin, BMI, Age etc.