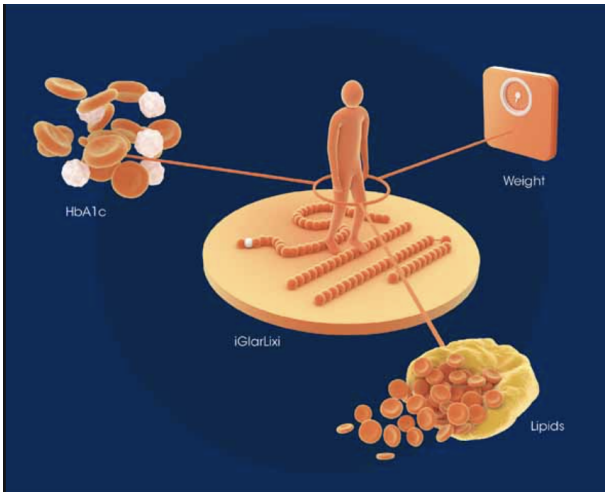
Project Report: Diabetes Prediction using Machine Learning

DONE BY:

JIL PATEL 1812046

KARAN SHAH 1812054

AKSHAT SHAH 1812059



Under the Guidance of

Prof. Akshata Prabhu

INTRODUCTION

Diabetes, is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. Acute complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death. Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes.

PROBLEM STATEMENT

We are trying to build a machine learning model to accurately predict whether the patients have diabetes or not.our objective is to prevent, cure and to improve the lives of all people affected by diabetes.

DETAILS ABOUT THE DATASET

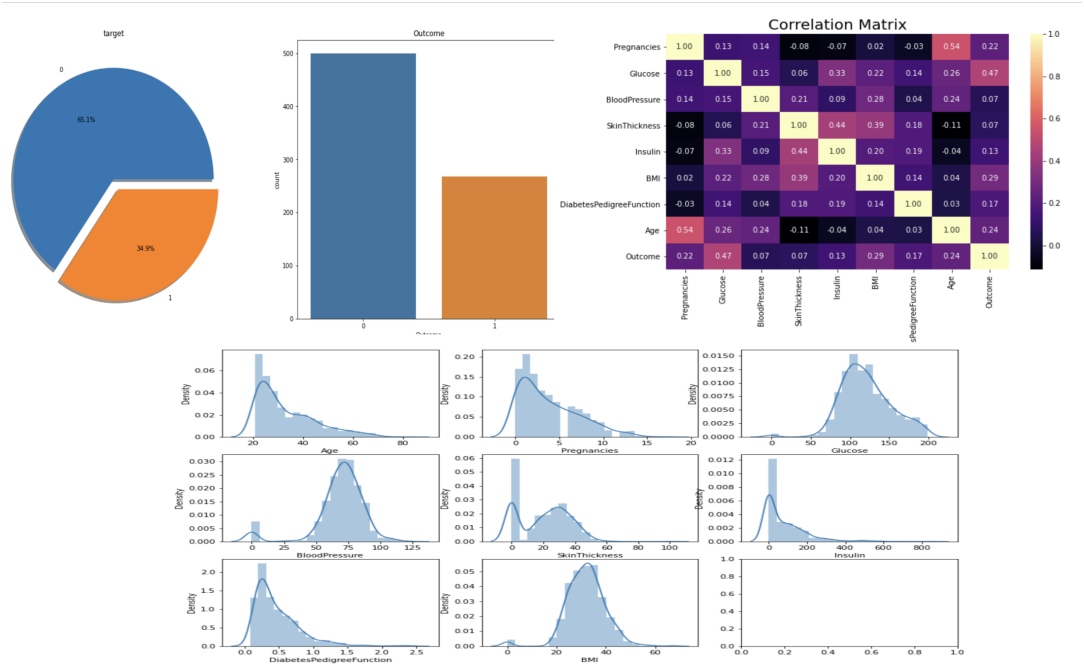
The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their **BMI, insulin level, age, and so on.**

1.**Pregnancies:** Number of times pregnant
2.**Glucose:** Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3.**BloodPressure:** Diastolic blood pressure (mm Hg)
4.**SkinThickness:** Triceps skin fold thickness (mm)
5.**Insulin:** 2-Hour serum insulin (mu U/ml)
6.**BMI:** Body mass index (weight in kg/(height in m)^2)
7.**DiabetesPedigreeFunction:**$Diabetespedigreefunction$
8.**Age:**$Age(years)$
9.**Outcome:**$Classvariable(0or1)$

Number of Observation Units: 768
Variable Number: 9

EXPLORATORY DATA ANALYSIS

The data set's structural data were checked. The types of variables in the dataset were examined. Size information of the dataset was accessed. The 0 values in the data set are missing values. Primarily these **0 values were replaced with NaN values.** Descriptive statistics of the data set were examined and **The distribution of the outcome variable were visualized.**



DATA PRE-PROCESSING

The NaN values missing observations were filled with the median values of each variable was sick or not.The outliers were determined by **LOF and dropped.** The X variables were standardized with the **rubost method.**

DESCRIPTION OF ALGORITHMS USED:

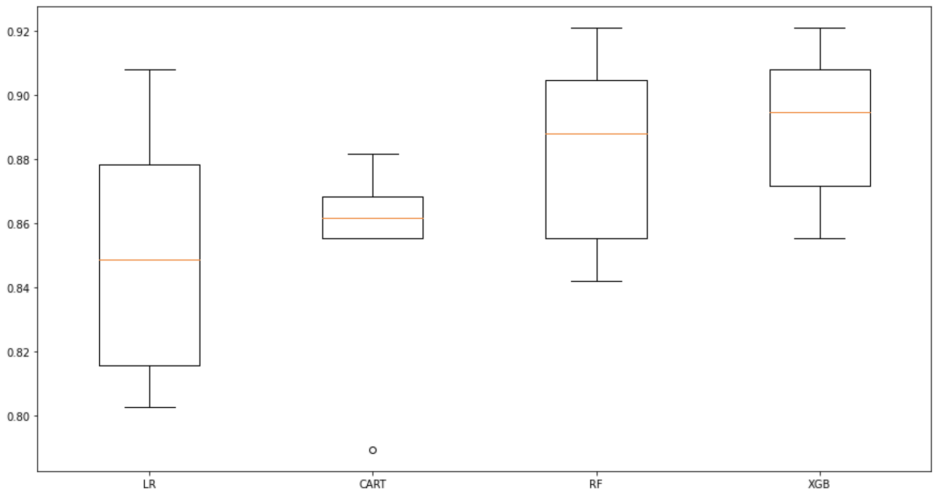
Logistic Regression is a classification method based on Linear Regression. Logistic Regression should be used for classification not for regression. The target variable can be a binary class or multi-class. In this project, we will apply some data exploration techniques to understand and explore the datasets. We will then apply the Logistic Regression classification algorithm.

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification.

Random Forest is supervised learning, used for both classification and Regression. The logic behind the random forest is bagging technique to create random sample features. The difference between the decision tree and the random forest is the process of finding the root node and splitting the feature node will run randomly.

Gradient Boosting Algorithm or GBM combines the predictions from multiple decision trees to generate the final predictions.the nodes in every decision tree take a different subset of features for selecting the best split. This means that the individual trees aren't all the same and hence they are able to capture different signals from the data.Additionally, each new tree takes into account the errors or mistakes made by the previous trees. So, every successive decision tree is built on the errors of the previous trees. This is how the trees in a gradient boosting machine algorithm are built sequentially.

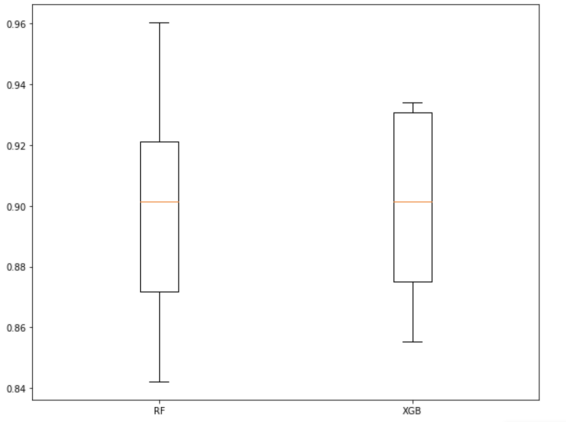
COMPARISION OF ALGORITHMS AND RESULTS



During Model Building: **Logistic Regression,CART, Random Forests, XGBoost,** using machine learning models Cross Validation Score and accuracy were calculated.

ALGORITHMS	ACCURACY (%)
LOGISTIC REGRESSION	84.8684%
DECISION TREE	85.7895%
RANDOM FOREST	88.1579%
GRADIENT BOOSTING	89.0789%

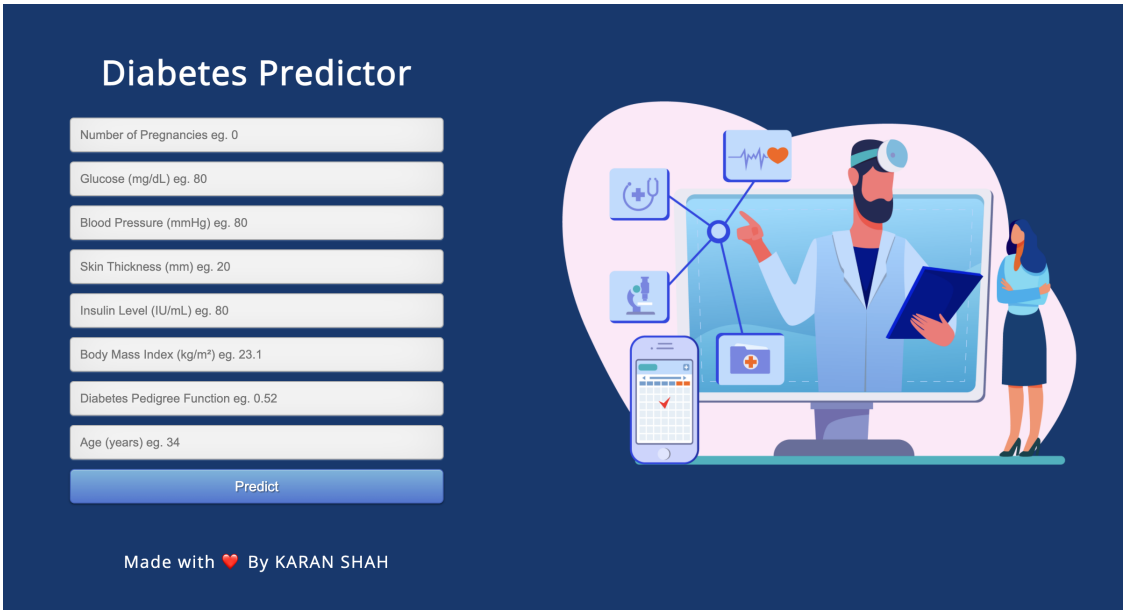
Later **Random Forests, XGBoost** hyperparameter optimizations optimized to **increase Cross Validation value.**



Result: The model created as a result of XGBoost hyperparameter optimization became the model with the lowest Cross Validation Score value(0.90)

MODEL DEPLOYMENT

we created a **web app using Flask which is a python micro framework**,so now people can fill the form and predict wheather they have diabetes or not.



CONCLUSION

Diabetes is a heterogeneous group of diseases. It's characterized by chronic elevation of glucose in the blood. The main motto of our project is **"To prevent and cure diabetes and to improve the lives of all people affected by diabetes".**Our proposed work also performs the analysis of the features in the dataset and selects the optimal features based on the correlation values.The model created as a result of **XGBoost hyperparameter optimization** became the model with the **lowest Cross Validation Score value(0.90)**.