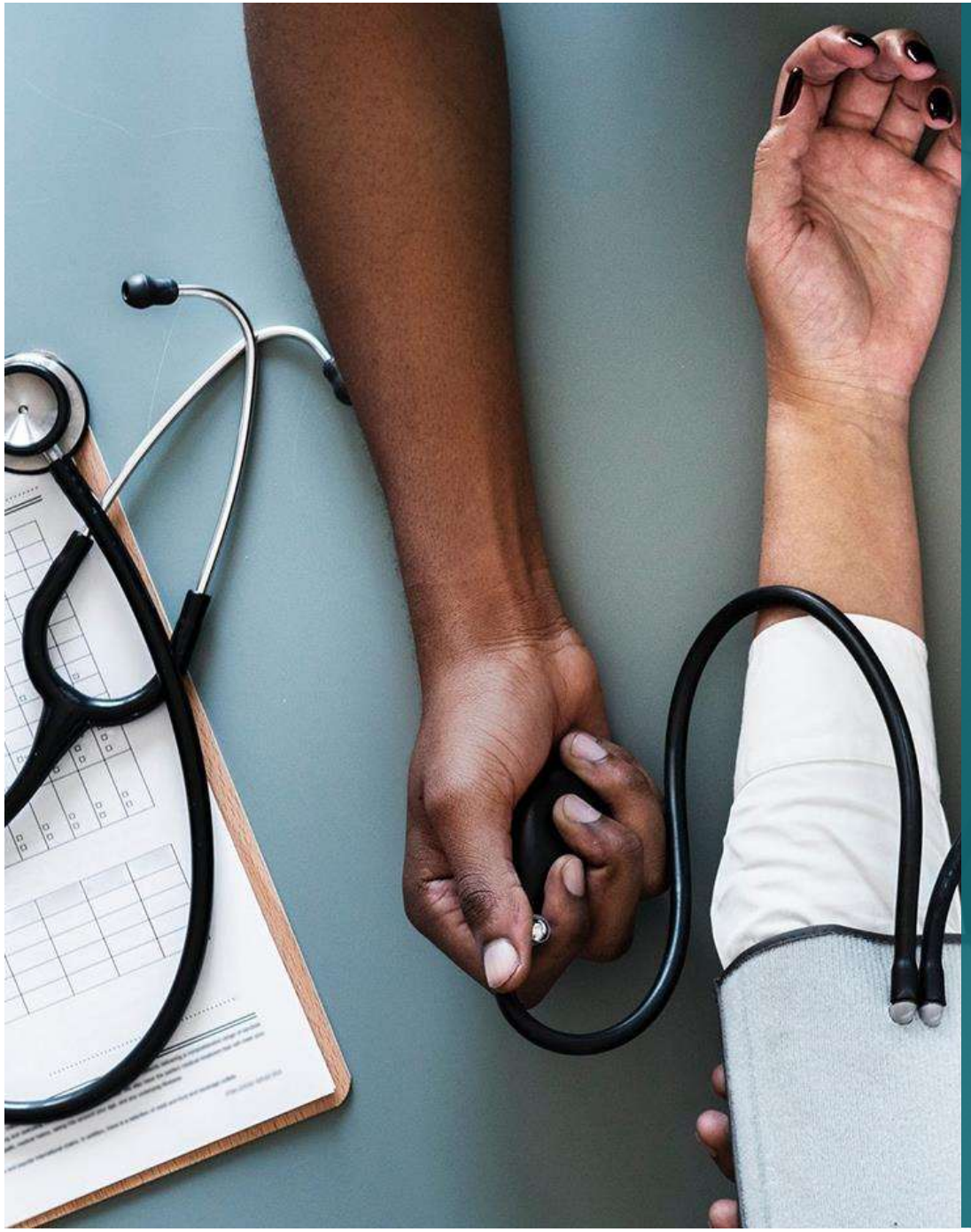


# Innovative Project

" Diabetes Prediction using Machine Learning "

# Project Description

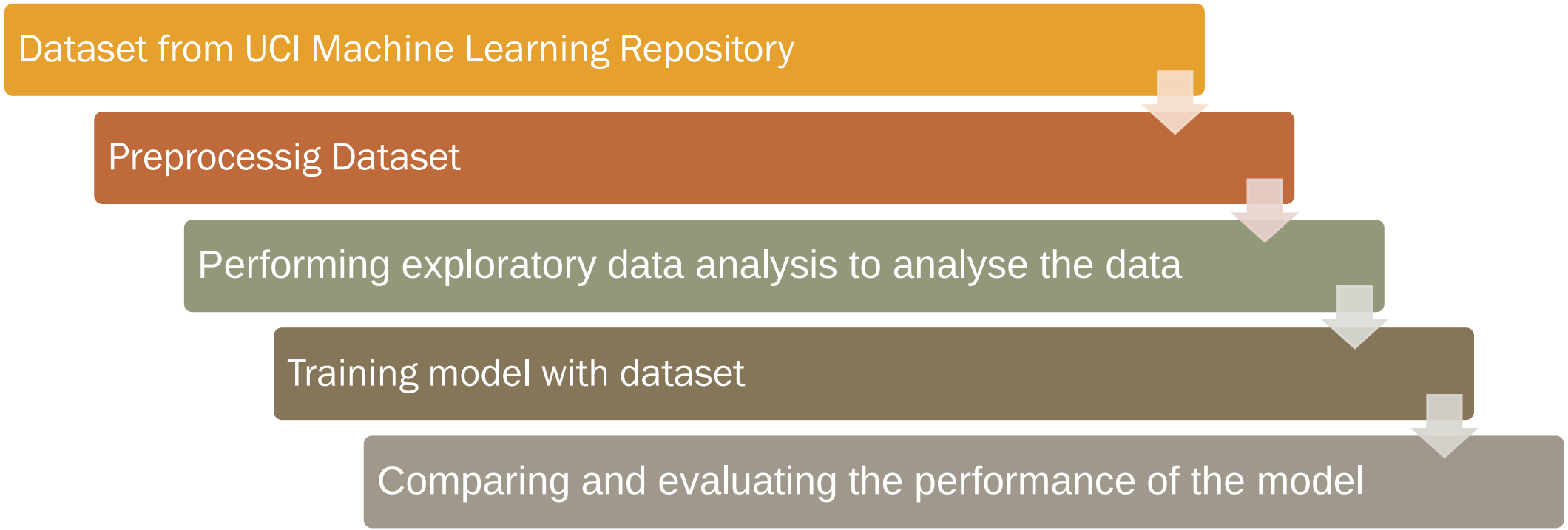
"Diabetes Prediction using Machine Learning"


**Project aims to develop a efficient machine learning model to predict TYPE 2 diabetes using Xgboost with hyperparameter optimization .**

*TYPE 2 is the noninsulin dependent diabetes. It is the most prevalent form of diabetes mellitus marked by Hyperglycemia and insulin glitch. Recognition of the cause of disease, precautionary measures should be inaugurated to downturn the dimensions of Diabetes Mellitus. Diabetes mellitus is a group of metabolic disorders where the blood sugar levels are higher than normal for prolonged periods of time . Diabetes is caused either due to the insufficient production of insulin in the body or due to improper response of the body's cells to Insulin.*

*Her we try to build a model which can predict type 2 diabetes using machine learning algorithms .*

# Project flow chart



# Dataset

**Context**

*This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.*

**Content**

*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.*

# Dataset


BloodPressure
Diastolic blood pressure (mm Hg)
61.00 - 73.20
Count: 261
0
122
Valid 768 100%
Mismatched 0 0%
Missing 0 0%
Mean 69.1
Std. Deviation 19.3
Quantiles
0 Min
62 25%
72 50%
80 75%
122 Max


Age
Age (years)
21
81
Valid 768 100%
Mismatched 0 0%
Missing 0 0%
Mean 33.2
Std. Deviation 11.8
Quantiles
21 Min
24 25%
29 50%
41 75%
81 Max


Outcome
Class variable (0 or 1) 268 of 768 are 1, the others are 0
0
1
Valid 768 100%
Mismatched 0 0%
Missing 0 0%
Mean 0.35
Std. Deviation 0.48
Quantiles
0 Min
0 25%
0 50%
1 75%
1 Max


Insulin
2-Hour serum insulin (mu U/ml)
0
846
Valid 768 100%
Mismatched 0 0%
Missing 0 0%
Mean 79.8
Std. Deviation 115
Quantiles
0 Min
0 25%
32 50%
128 75%
846 Max

# Dataset

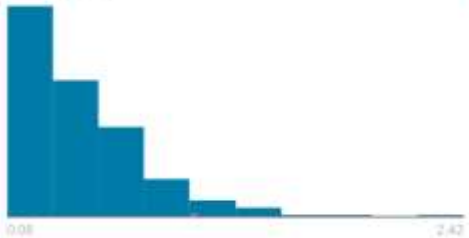
**# Pregnancies**

Number of times pregnant

| | | |
|---|---|---|
| Valid | 768 | 100% |
| Mismatched | 0 | 0% |
| Missing | 0 | 0% |
| Mean | 3.85 | |
| Std. Deviation | 3.37 | |
| Quantiles | 0 | Min |
| | 1 | 25% |
| | 3 | 50% |
| | 6 | 75% |
| | 17 | Max |

**# DiabetesPedigreeFunction**

Diabetes pedigree function



| | | |
|---|---|---|
| Valid | 768 | 100% |
| Mismatched | 0 | 0% |
| Missing | 0 | 0% |
| Mean | 0.47 | |
| Std. Deviation | 0.33 | |
| Quantiles | 0.08 | Min |
| | 0.24 | 25% |
| | 0.37 | 50% |
| | 0.63 | 75% |
| | 2.42 | Max |

**# Glucose**

Plasma glucose concentration a 2 hours in an oral glucose tolerance test

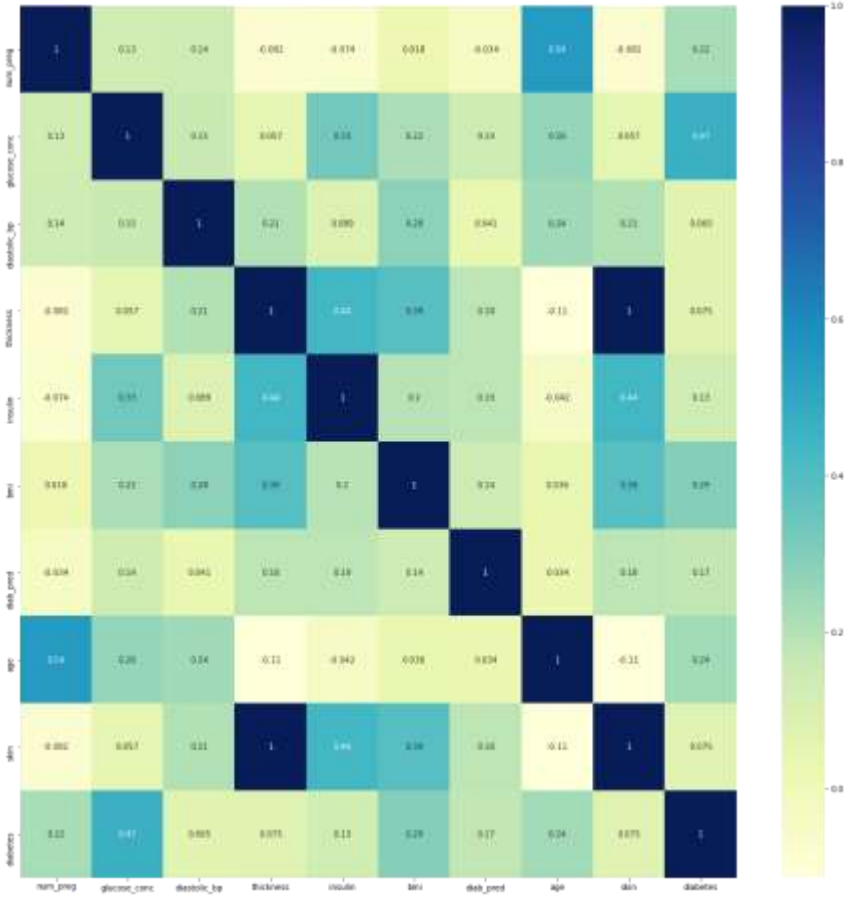
| | | |
|---|---|---|
| Valid | 768 | 100% |
| Mismatched | 0 | 0% |
| Missing | 0 | 0% |
| Mean | 121 | |
| Std. Deviation | 32 | |
| Quantiles | 0 | Min |
| | 99 | 25% |
| | 117 | 50% |
| | 141 | 75% |
| | 199 | Max |

**# SkinThickness**

Triceps skin fold thickness (mm)

| | | |
|---|---|---|
| Valid | 768 | 100% |
| Mismatched | 0 | 0% |
| Missing | 0 | 0% |
| Mean | 20.5 | |
| Std. Deviation | 15.9 | |
| Quantiles | 0 | Min |
| | 0 | 25% |
| | 23 | 50% |
| | 32 | 75% |
| | 99 | Max |

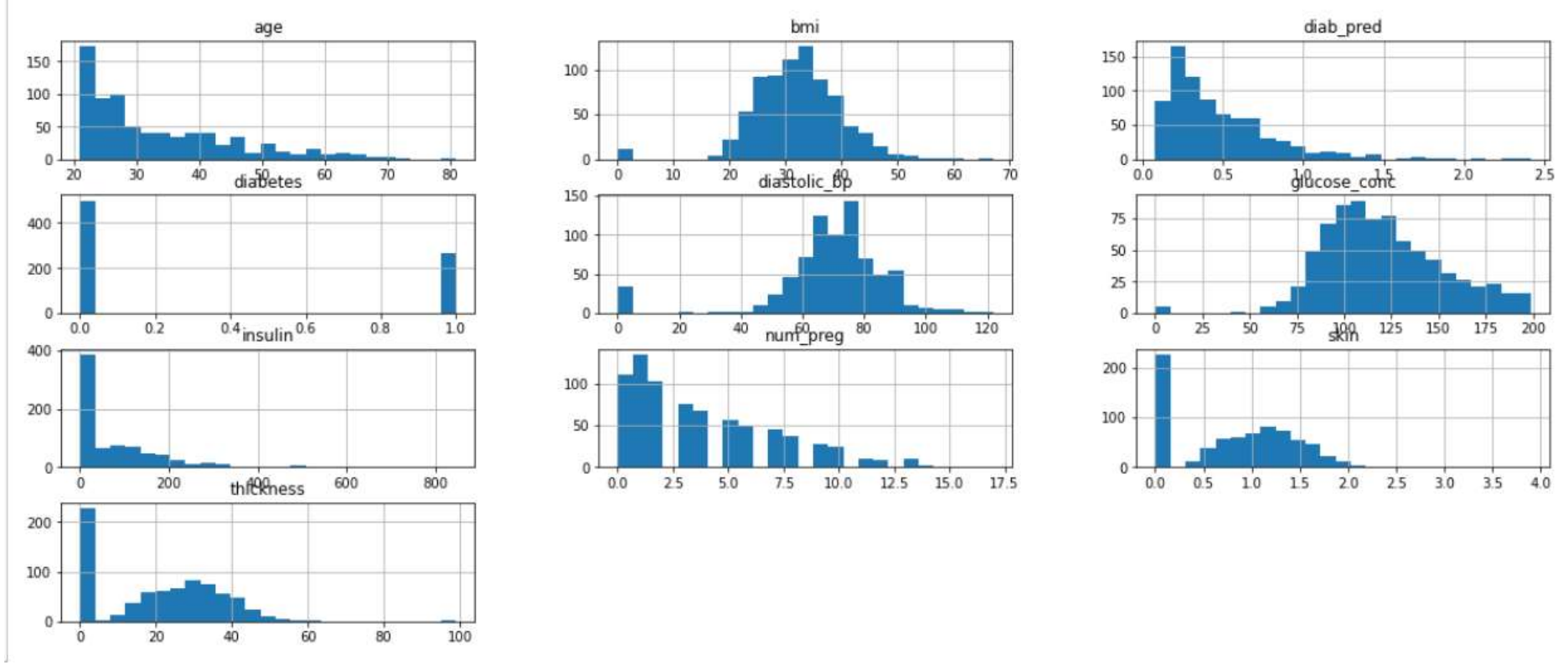
# Dataset Correlation



Here is a heat map to show the correlation of dataset . Correlation tells us how data is correlated .

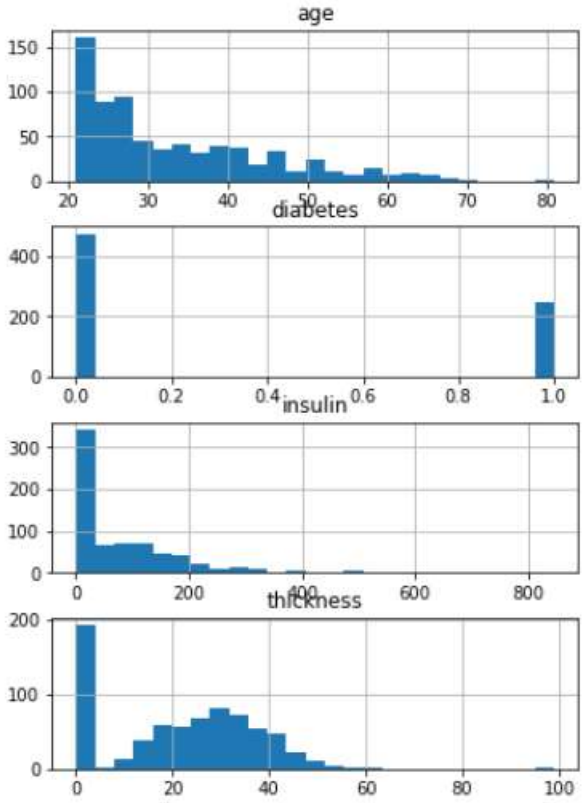
# Removing Outliers

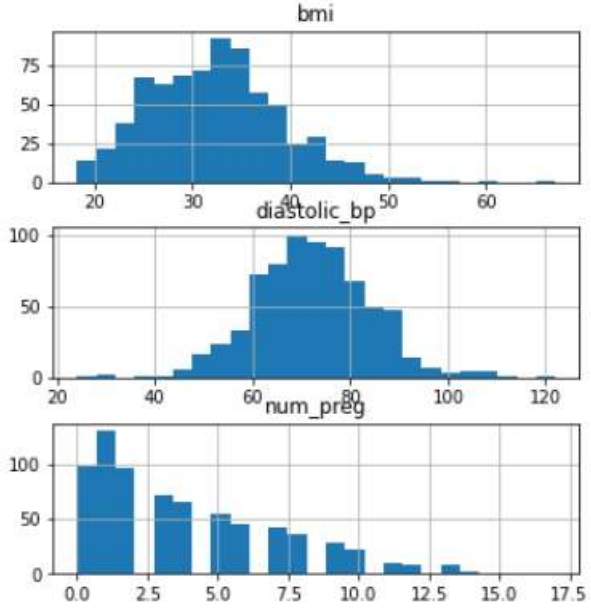
BEFORE

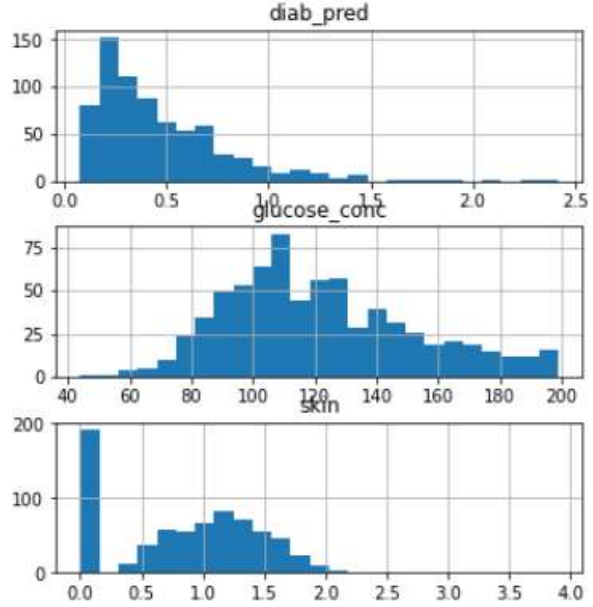


# Removing Outliers

AFTER



# XGBoost

- XGBoost is an advanced version of Gradient boosting method, it literally means eXtreme Gradient Boosting. XGBoost developed by Tianqi Chen, falls under the category of Distributed Machine Learning Community (DMLC).

- The main aim of this algorithm is to increase the speed and efficiency of computation. The Gradient Descent Boosting algorithm computes the output at a slower rate since they sequentially analyze the data set, therefore XGBoost is used to boost or extremely boost the performance of the model.

- XGBoost is designed to focus on computational speed and model efficiency. The main features provided by XGBoost are:
  - Parallelly creates decision trees.
  - Implementing distributed computing methods for evaluating large and complex models.
  - Using Out-of-Core Computing to analyze huge datasets.
  - Implementing cache optimization to make the best use of resources.

# Hyperparameter Optimisation

**Hyperparameters** *are all the parameters which can be arbitrarily set by the user before starting training (eg. number of estimators in Random Forest).*

**Random Search**

*In Random Search, we create a grid of hyperparameters and train/test our model on just some random combination of these hyperparameters. In this example, I additionally decided to perform Cross-Validation on the training set*.

# Result and Analysis

## XgBoost with hyperparameter optimisation

| | |
|---|---|
| Confusion matrix | [[128 12]<br>[ 38 40]] |
| Accuracy | 0.7706422018348624 |
| Precision | 0.91 |
| Recall | 0.77 |

## Random Forest Classifier

```
Accuracy = 0.752
```

Both the models were trained using the same dataset .Here we can see the accuracy of the XgBoost algorithm with hyperparameter optimisation is higher than in the case of Random Forest Classifier .

Accuracy can be further improved with more balanced dataset using same model .

# References

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (p. 261). American Medical Informatics Association.

Bhulakshmi, D., & Gandhi, G. (2020). *The Prediction of Diabetes in Pima Indian Women Mellitus Based on XGBOOST Ensemble Modeling Using Data Science* (No. 2864). EasyChair.