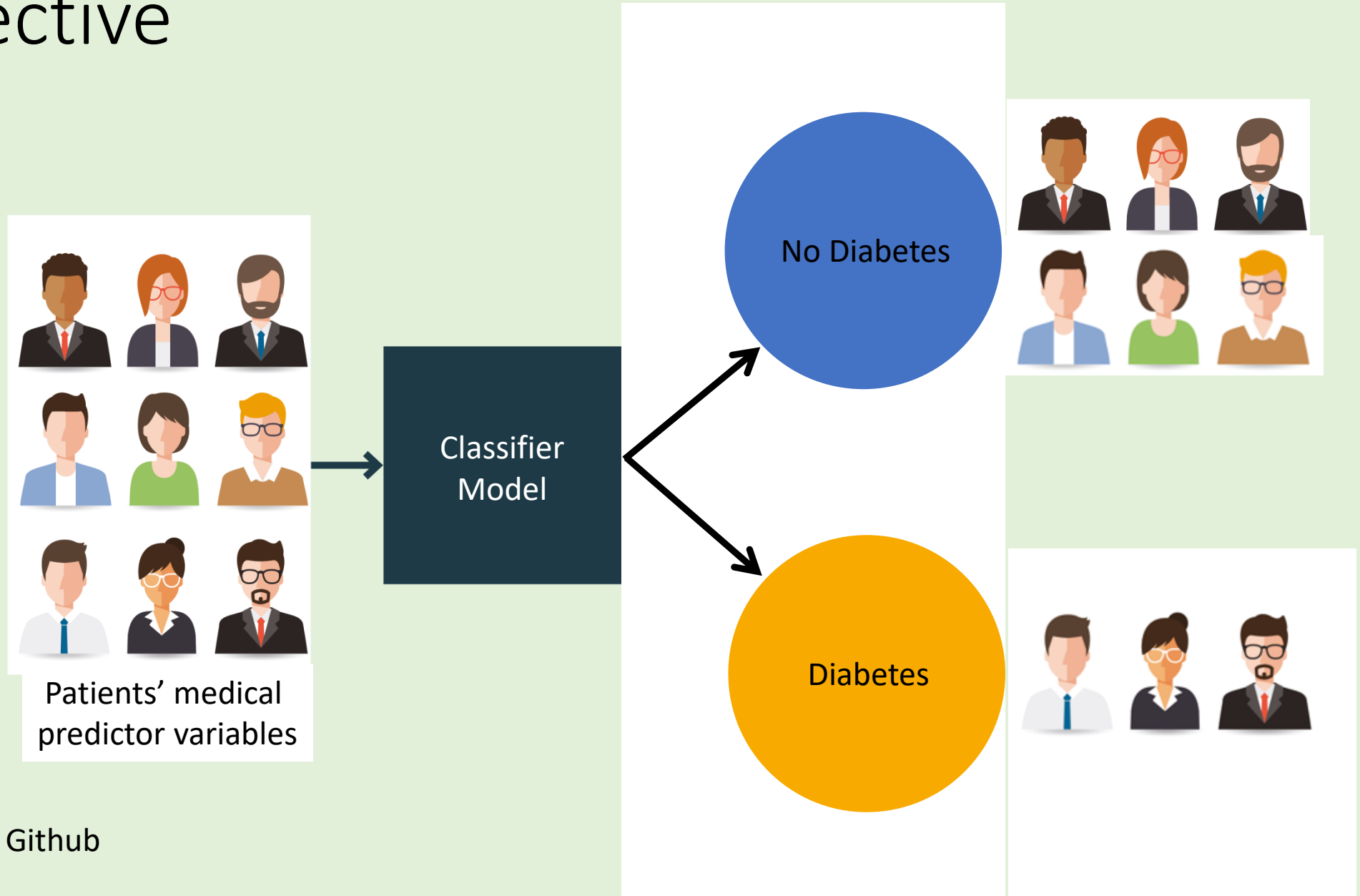
# Predicting Diabetes using ML

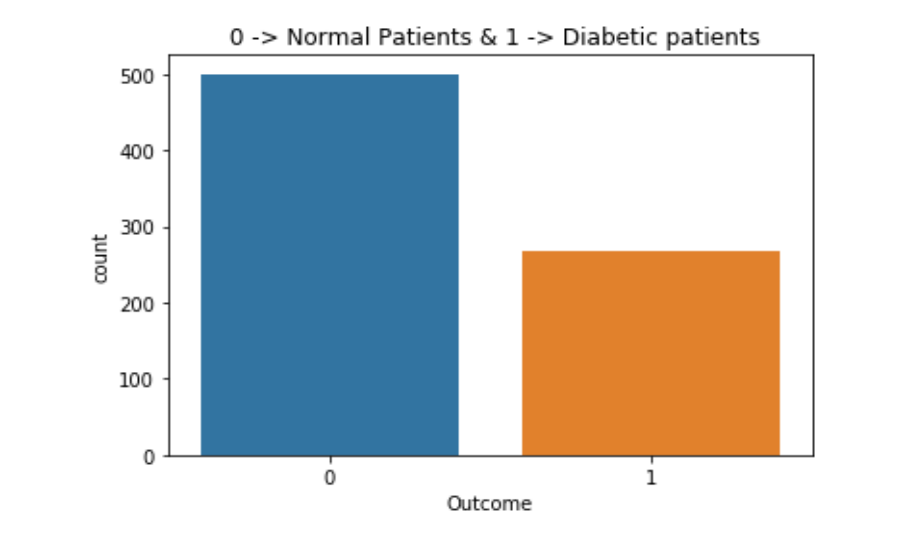
**Developing ML model based on medical records of the patients**

# Objective




Picture source Github

# Dataset



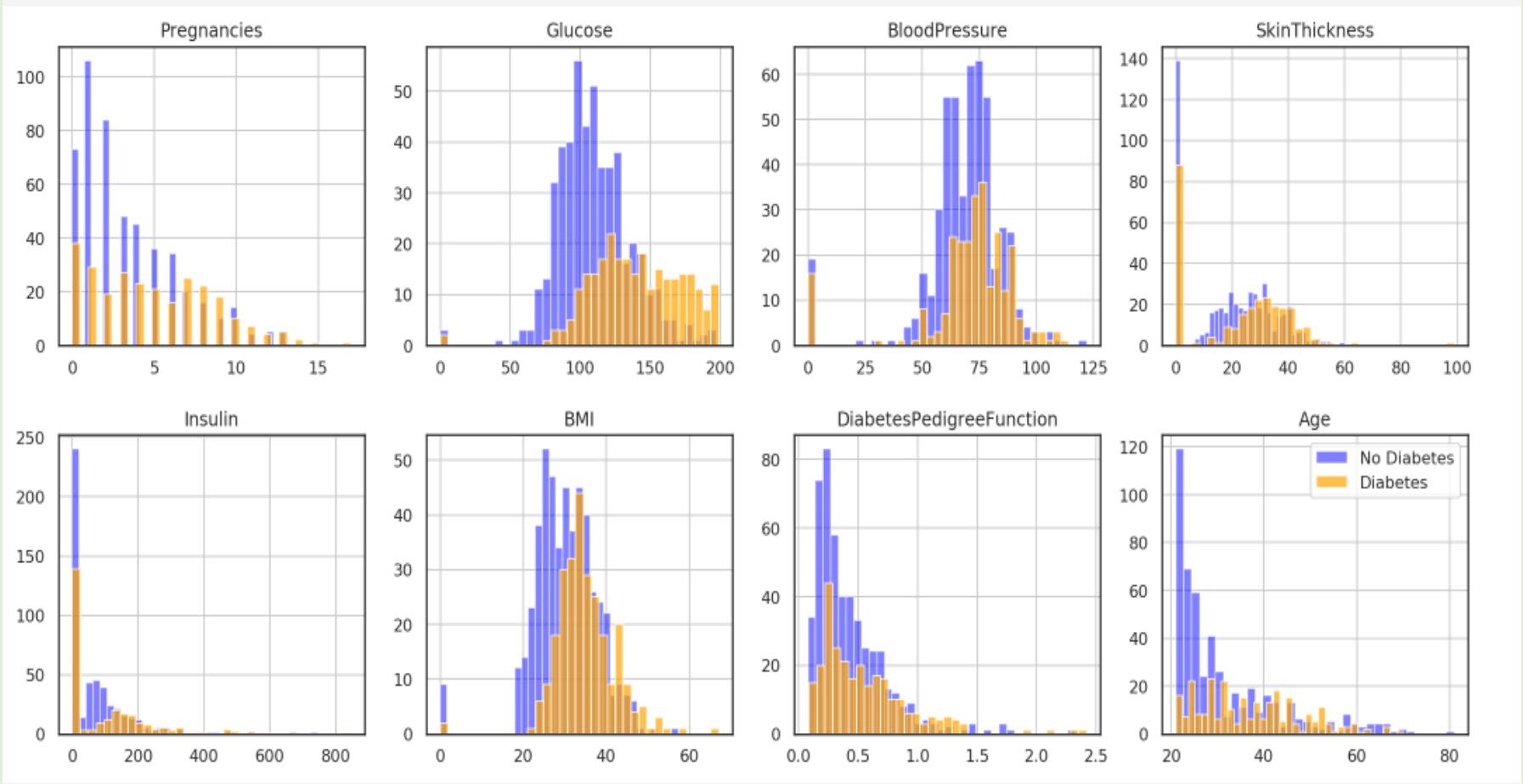

Data set downloaded from Kaggle

There are 768 observations with 8 medical predictor features (input) and 1 target variable

The 8 medical predictor features are:

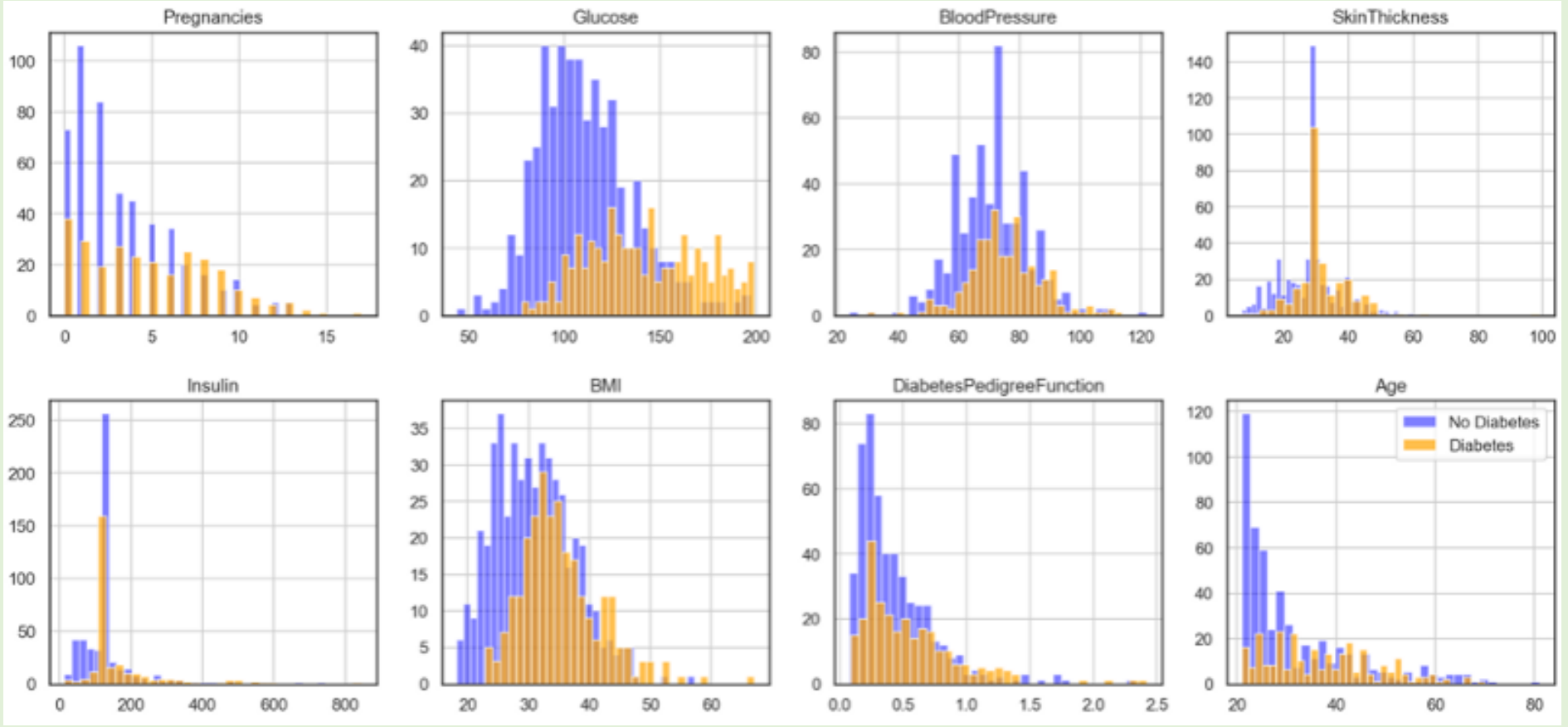
· **Pregnancies**: Number of times pregnant
· **Glucose**: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
· **Blood Pressure**: Diastolic blood pressure (mm Hg)
· **Skin Thickness**: Triceps skin fold thickness (mm)
· **Insulin**: 2-Hour serum insulin (mu U/ml)
· **BMI**: Body mass index (weight in kg/(height in m)$^2$)
· **Diabetes PedigreeFunction**: Diabetes pedigree function
· **Age**: Age (years)

# Exploratory Data Analysis (EDA)



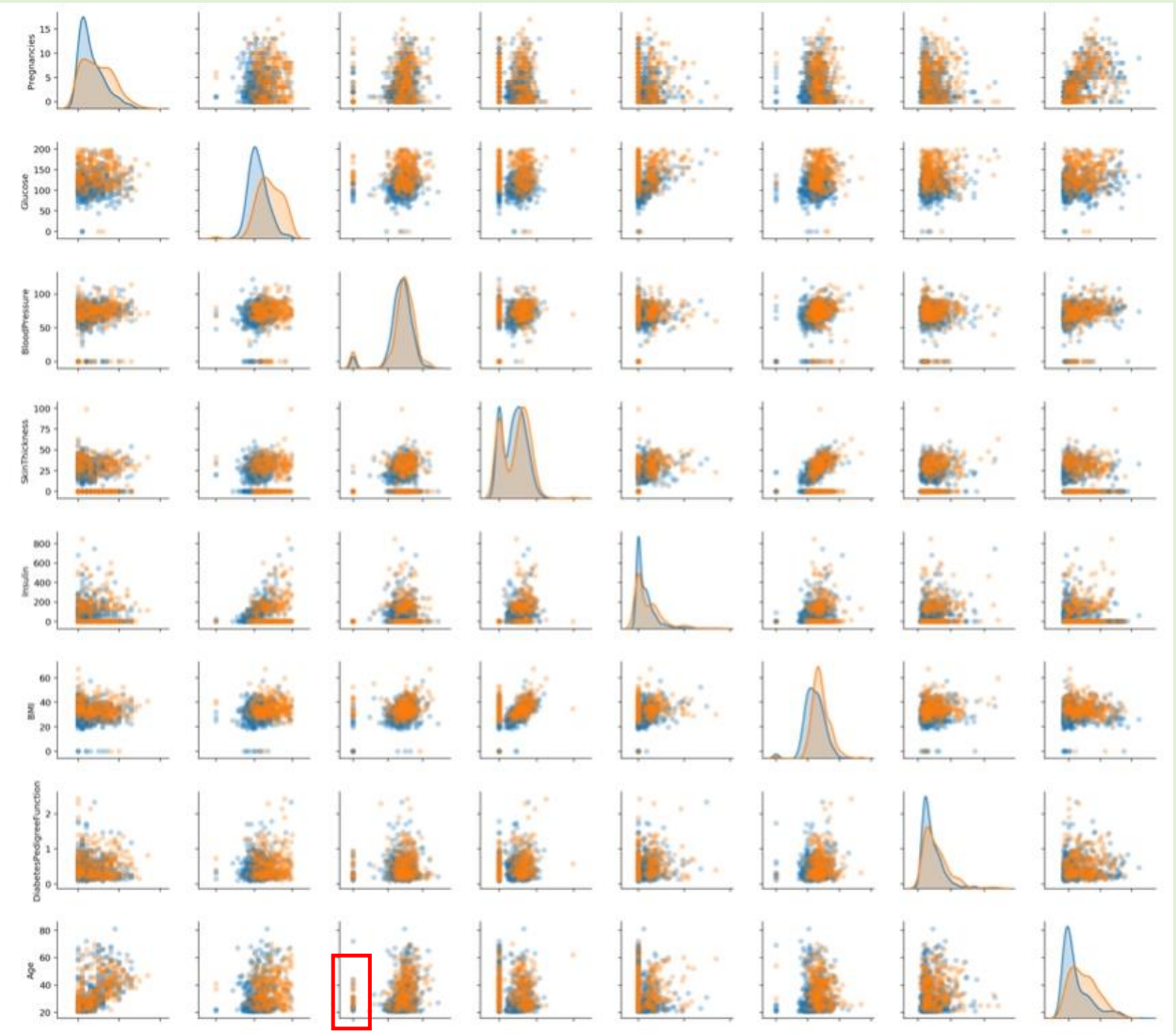

feature-outcome distribution before removing zeros

# Exploratory Data Analysis (EDA)

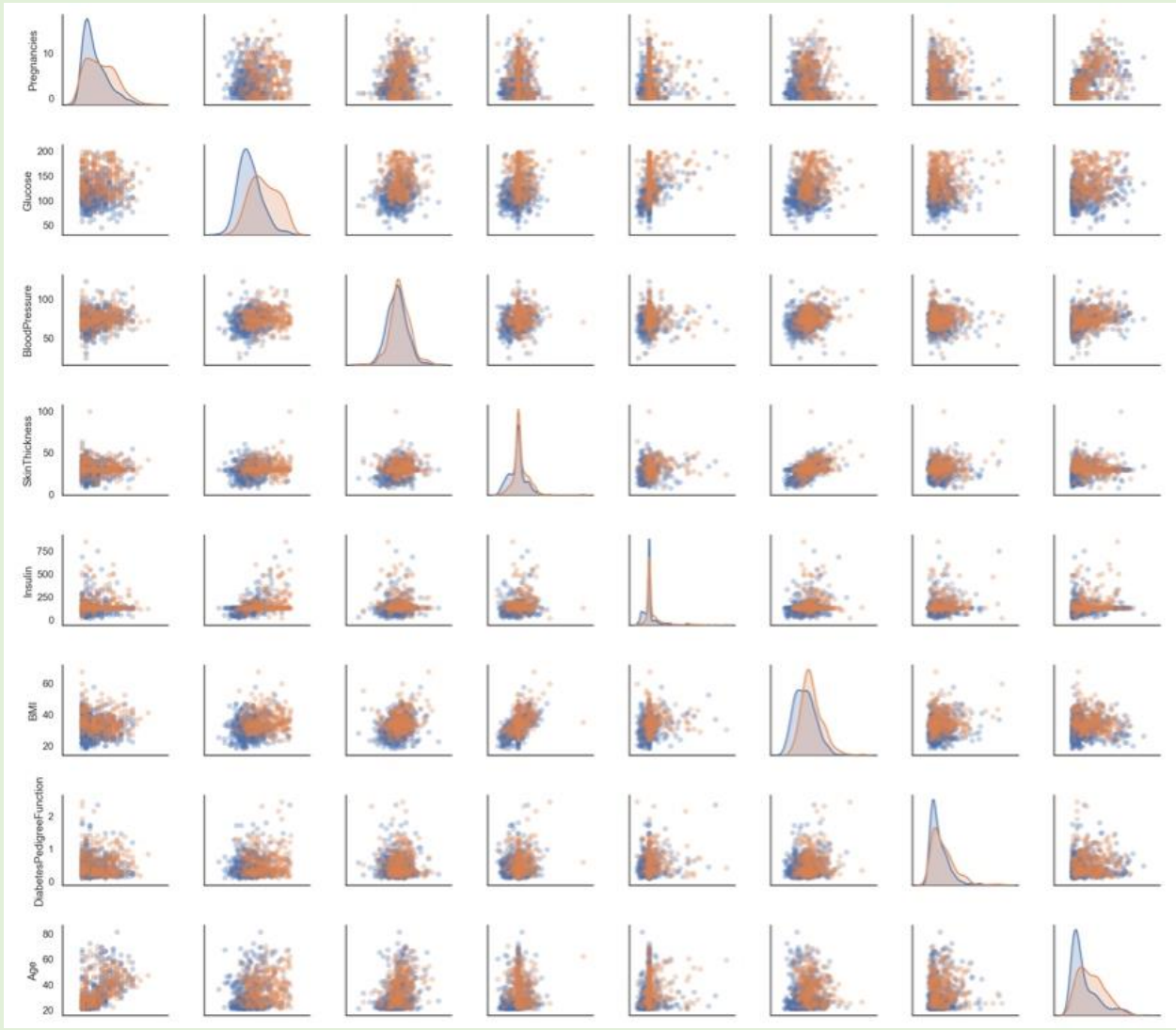



feature-outcome distribution after removing zero and missing values

# Pair Plots

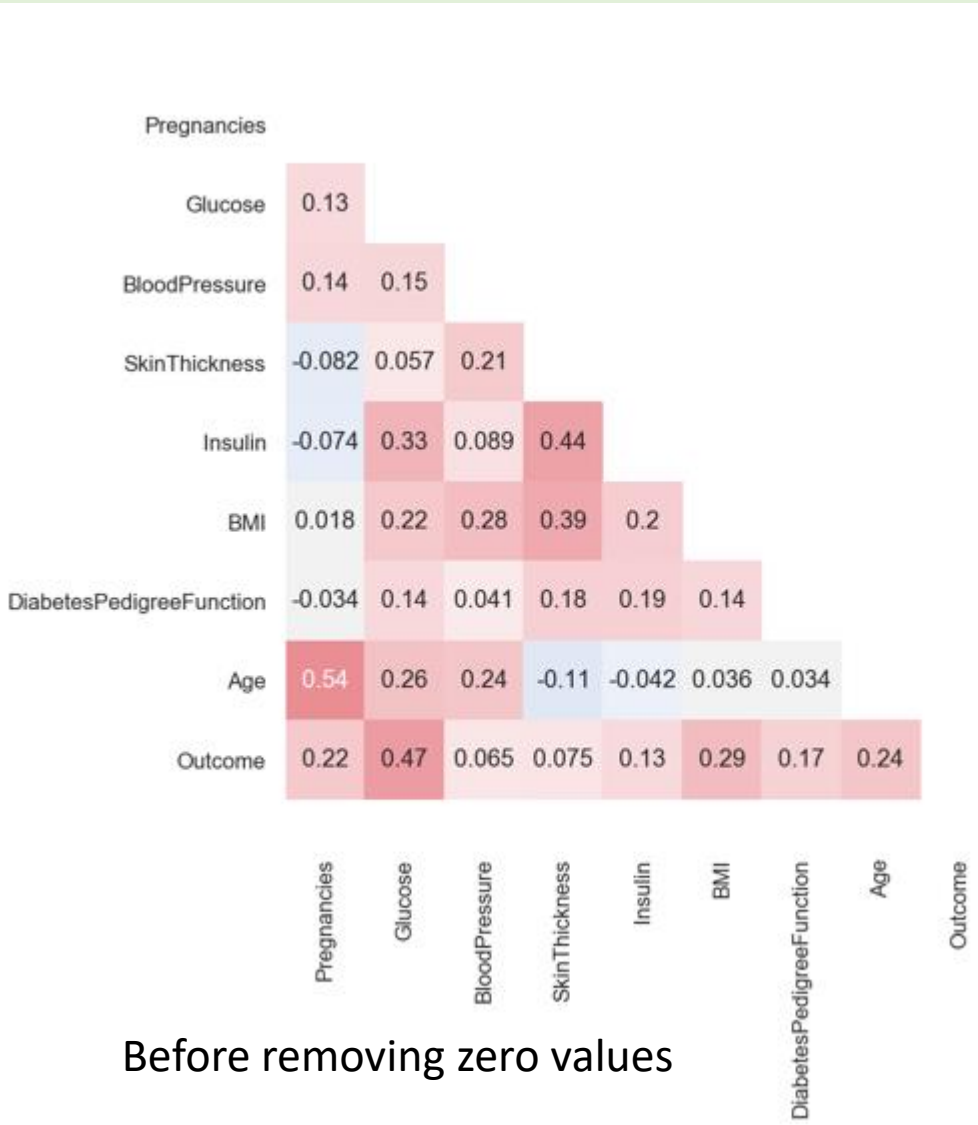



Before removing zero values

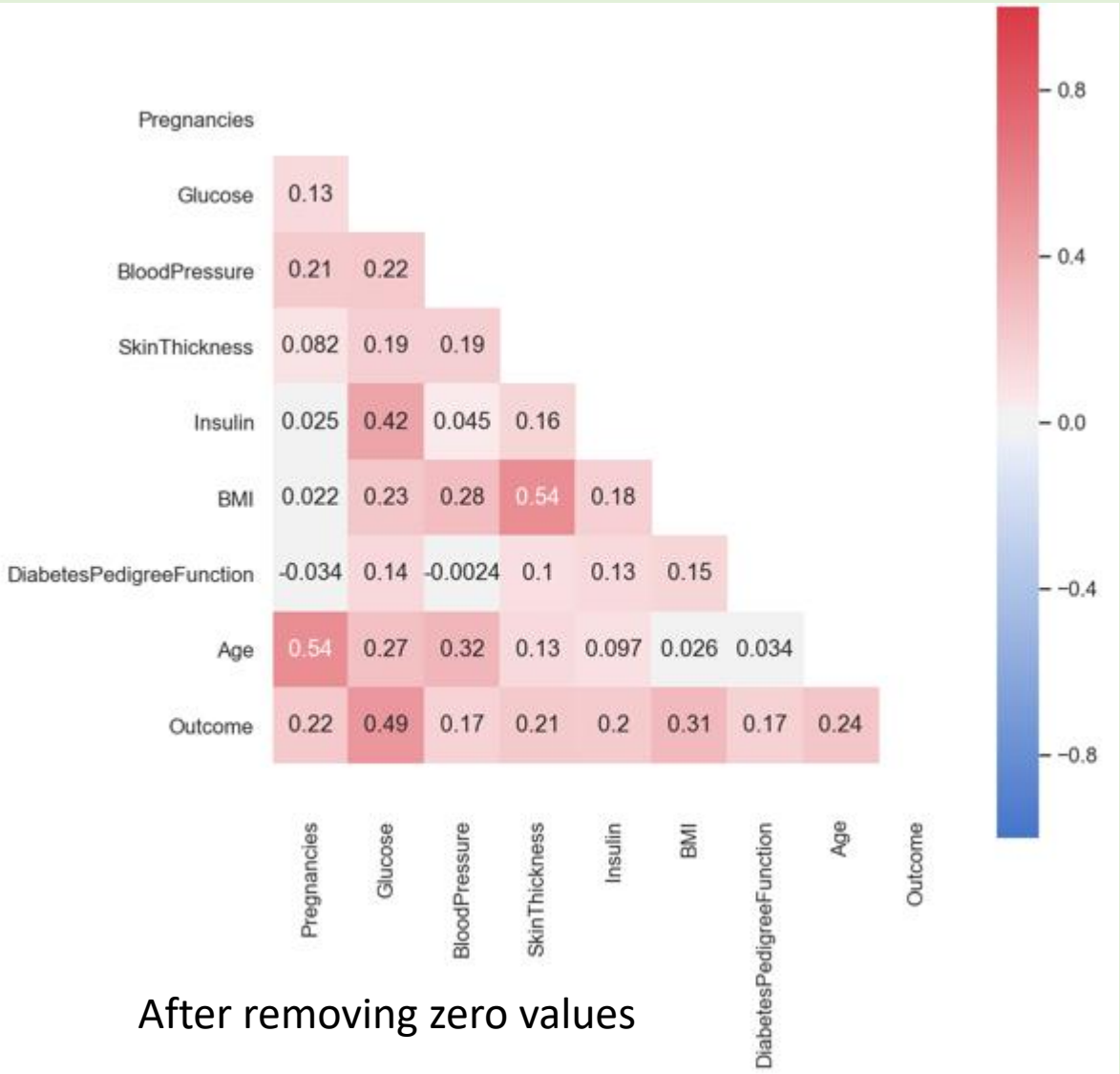



After removing zero values

# Correlation Between Features & Outcome

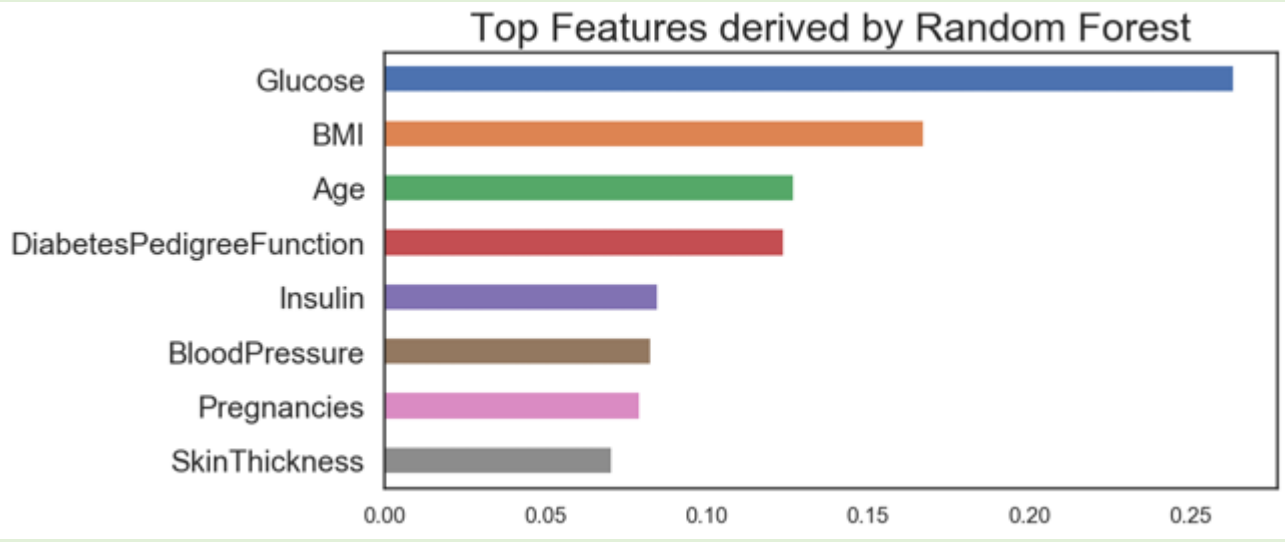


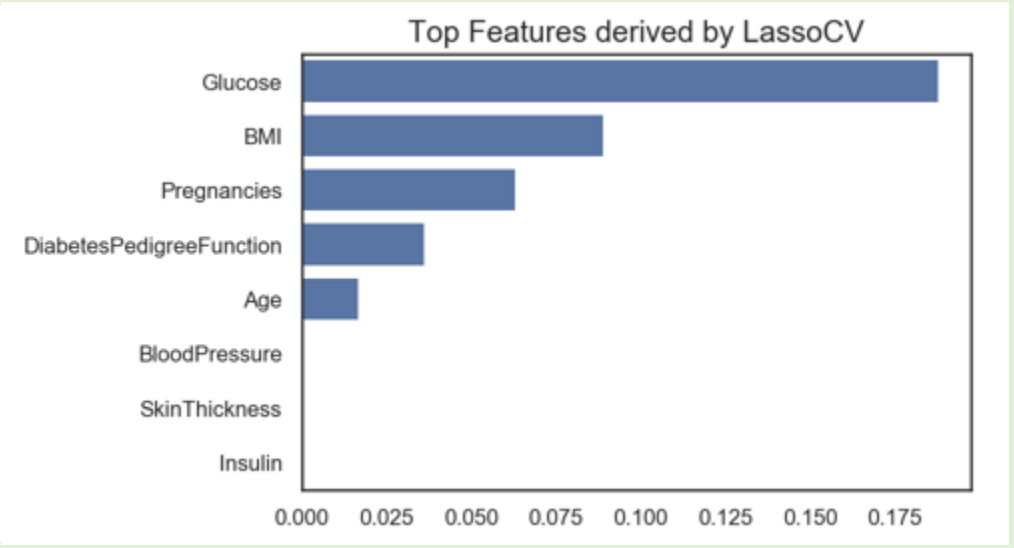

Before removing zero values




After removing zero values

# Feature Importance




‘Glucose’ and ‘BMI’ are the most important medical predictor features.

# Model Evaluation

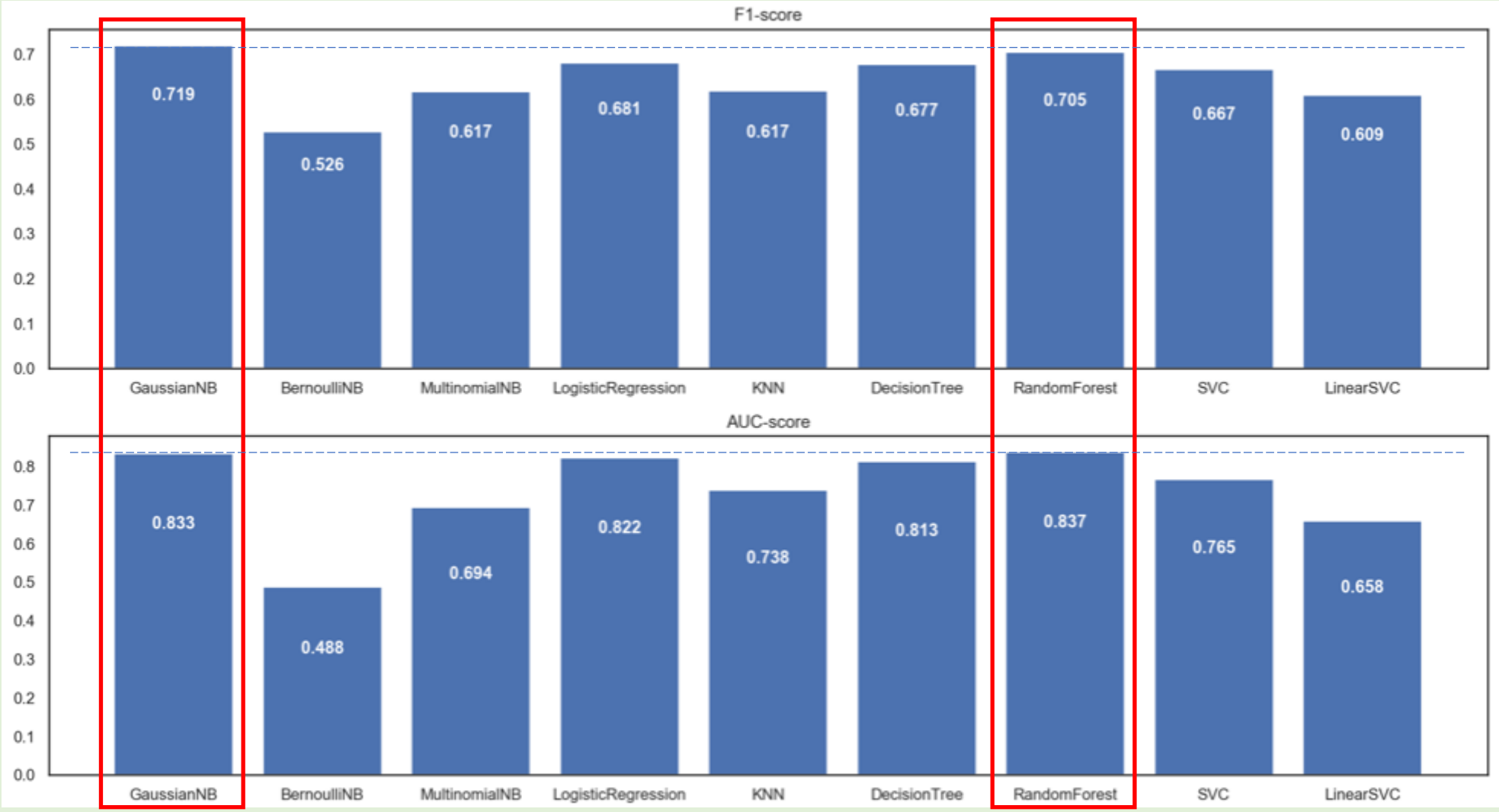
Following 8 models have been evaluated:

· Gaussian Naive Bayes
· Deep Learning MLP
· Multinomial Naive Bayes
· Logistic Regression
· K Nearest Neighbour
· Decision Tree Classifier
· Random Forest Classifier
· Support Vector Classification (SVC)

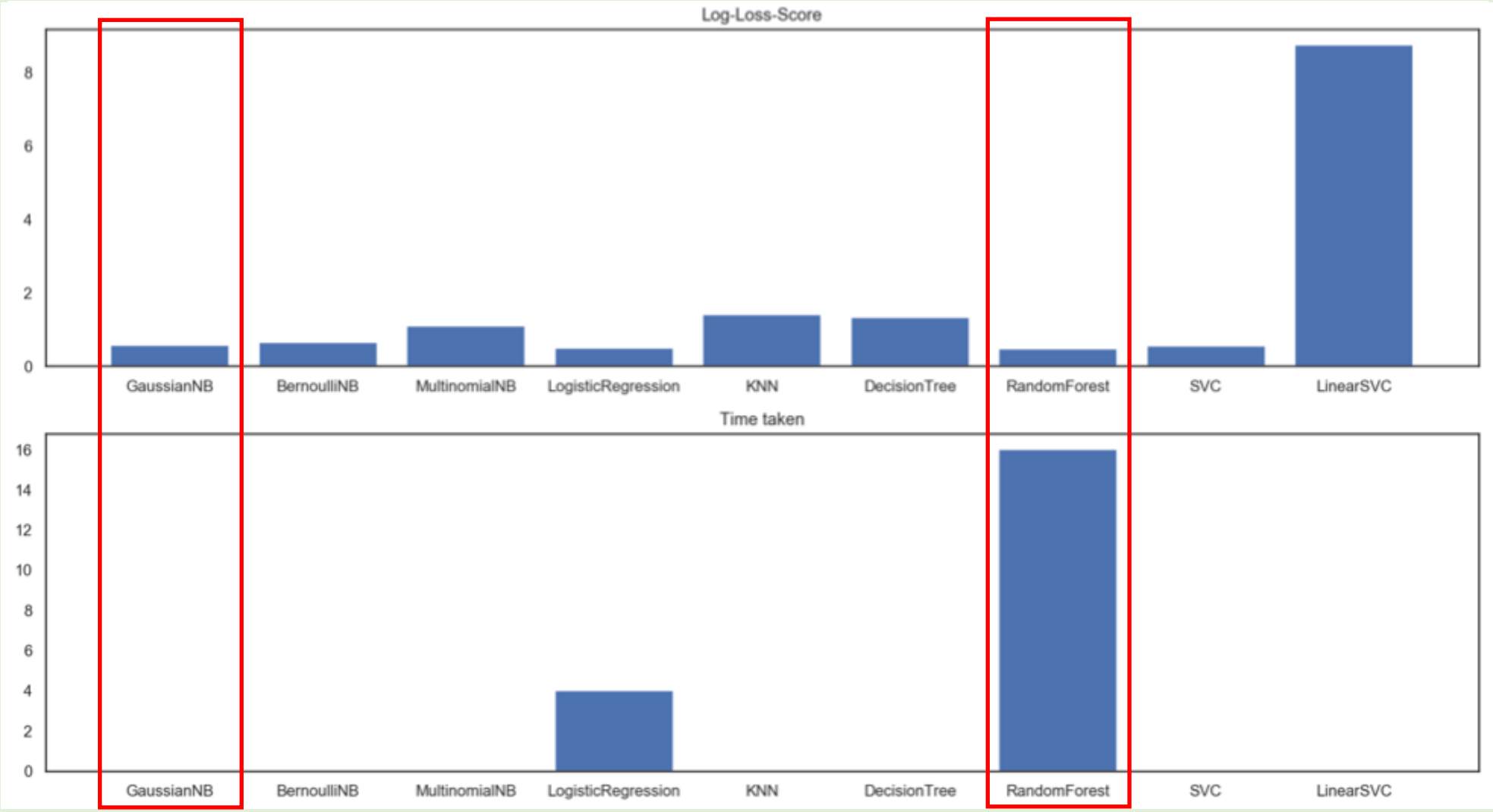
The performance metrics used in the evaluation are:

· Accuracy Score: proportion of correct predictions out of the whole dataset.
· Precision Score: proportion of correct predictions out of all predicted diabetic cases.
· Recall Score: proportion of correct predictions out of all actual diabetic cases.
· F1 Score: optimised balance between Precision and Recall for binary targets.
· Log Loss: aka logistic loss or cross-entropy loss, defined as the negative log-likelihood of the true labels given a probabilistic classifier's predictions, and has to be as low as possible.

# Model Performance



# Model Performance



# Conclusion

In this project, the **Gaussian Naive Bayes** model has achieved prediction (Recall) score of **76**

Out of all diabetic patients, 76% of them will be classified correctly using medical diagnostic measurements

Similarly its predicting non diabetic as non-diabetic of 84%.

Glucose and BME are the most contributing features for Diabetes.