
# SOUMYAJITA BOSE

IITP000180

2312RES655

DATA SCIENCE

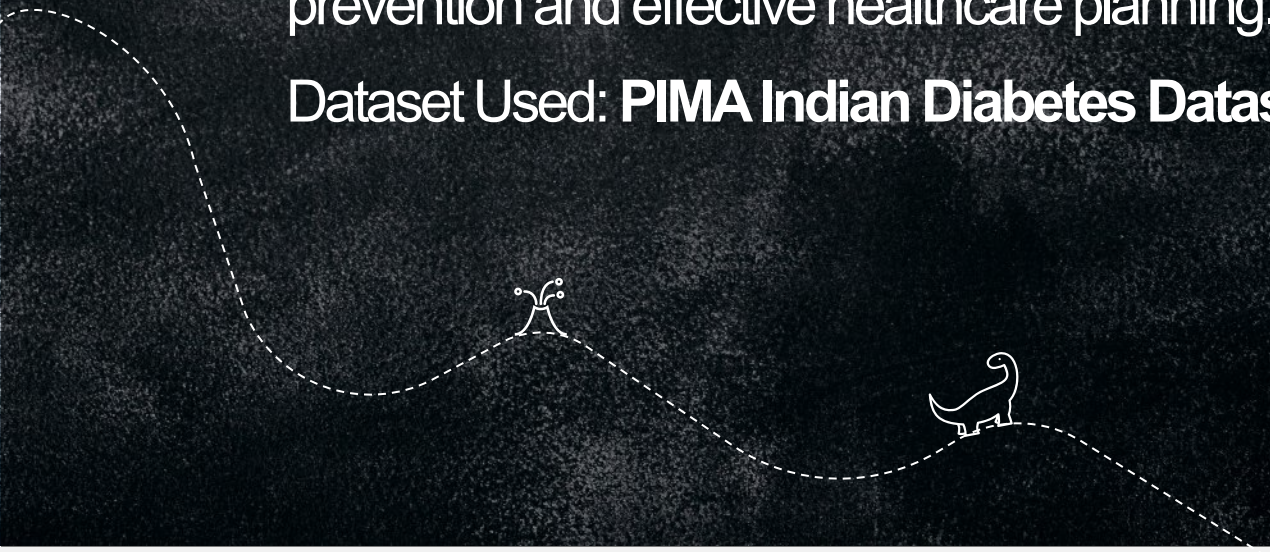
soumyajita_2312res655@iitp.ac.in


https://github.com/23f1000193/Diabetes-Risk-Prediction-from-Health-Records.git

# Diabetes Risk Prediction from Health Records

This project aims to predict the likelihood of diabetes in patients using health indicators such as **Glucose, BMI, Age, Blood Pressure, Insulin, and Skin Thickness**. By applying **data science techniques and machine learning models**, this project demonstrates how early detection can help in prevention and effective healthcare planning.

Dataset Used: **PIMA Indian Diabetes Dataset**



## 🎯 Goals & Motivation

- Build a **predictive model** to identify individuals at risk of diabetes.
- Explore and analyze medical data through **data visualization**.
- Apply **machine learning algorithms** to achieve reliable predictions.
- Provide **insights** into key health factors contributing to diabetes.
- Showcase an **end-to-end data science pipeline** from raw data to results.

- Models Implemented:
  - Logistic Regression
  - Random Forest Classifier
  - Gradient Boosting (XGBoost)
  - Support Vector Machine (SVM)
- **Train-Test Split**: 80:20
- **Cross-validation** used to ensure reliability

✅ Dataset loaded. Shape: (768, 9)

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | \ |
|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | |

| | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|
| 0 | 0.627 | 50 | 1 |
| 1 | 0.351 | 31 | 0 |
| 2 | 0.672 | 32 | 1 |
| 3 | 0.167 | 21 | 0 |
| 4 | 2.288 | 33 | 1 |

Cross-validation results:

| | Model | CV ROC-AUC |
|---|---|---|
| 0 | LogisticRegression | 0.849029 |
| 3 | SVC | 0.831892 |
| 1 | RandomForest | 0.830144 |
| 2 | GradientBoosting | 0.817014 |

✅ Best model: LogisticRegression

Test set evaluation:

| | Accuracy | Precision | Recall | F1 | ROC_AUC |
|---|---|---|---|---|---|
| 0 | 0.734375 | 0.637931 | 0.552239 | 0.592 | 0.83403 |

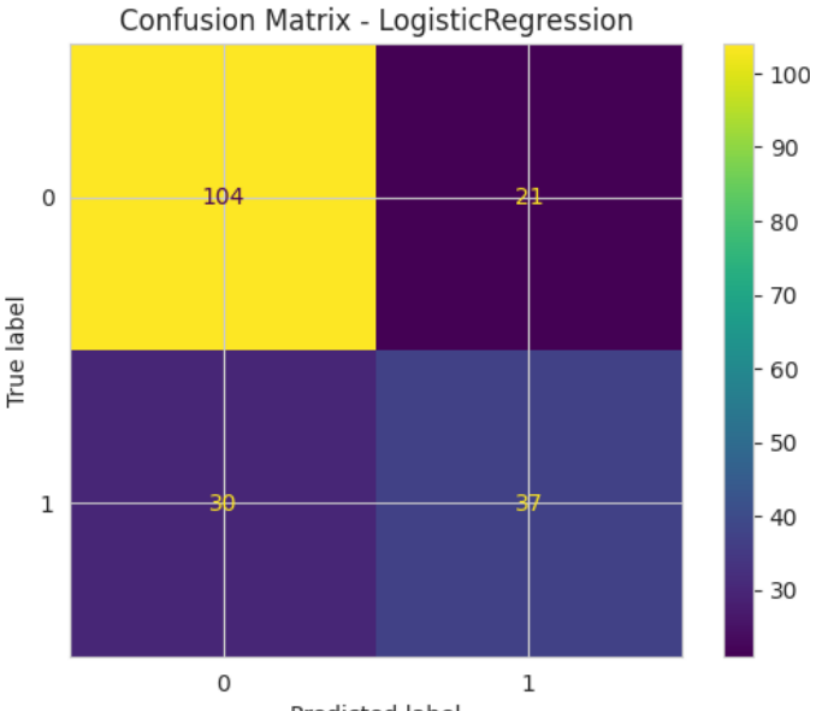
Classification Report:

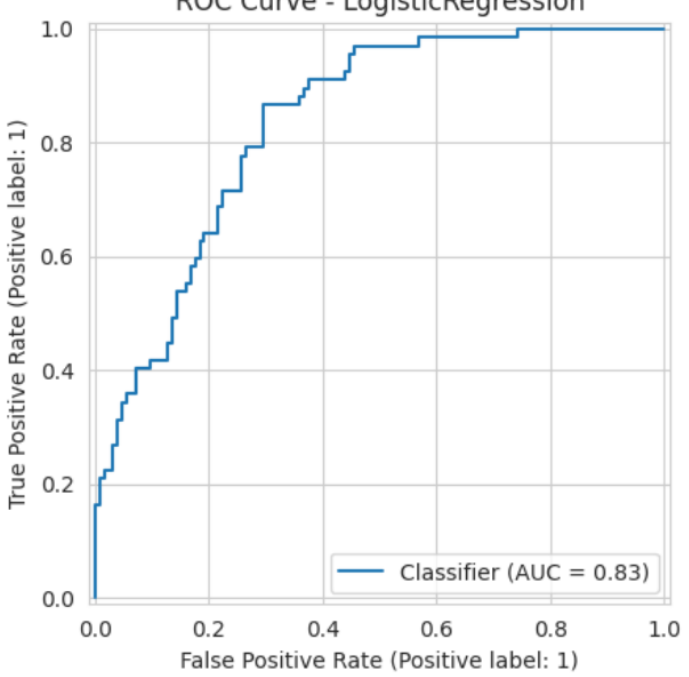
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.83 | 0.80 | 125 |
| 1 | 0.64 | 0.55 | 0.59 | 67 |
| accuracy | | | 0.73 | 192 |
| macro avg | 0.71 | 0.69 | 0.70 | 192 |
| weighted avg | 0.73 | 0.73 | 0.73 | 192 |

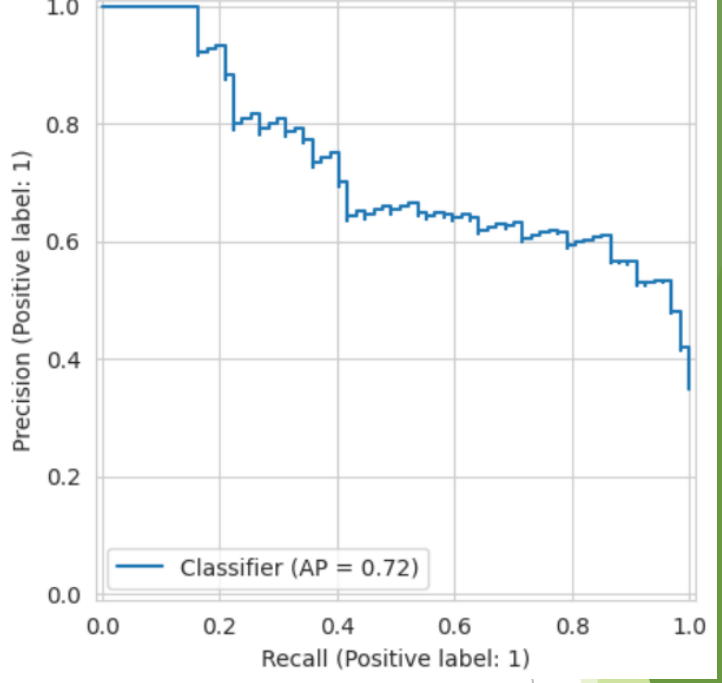
Summary statistics:

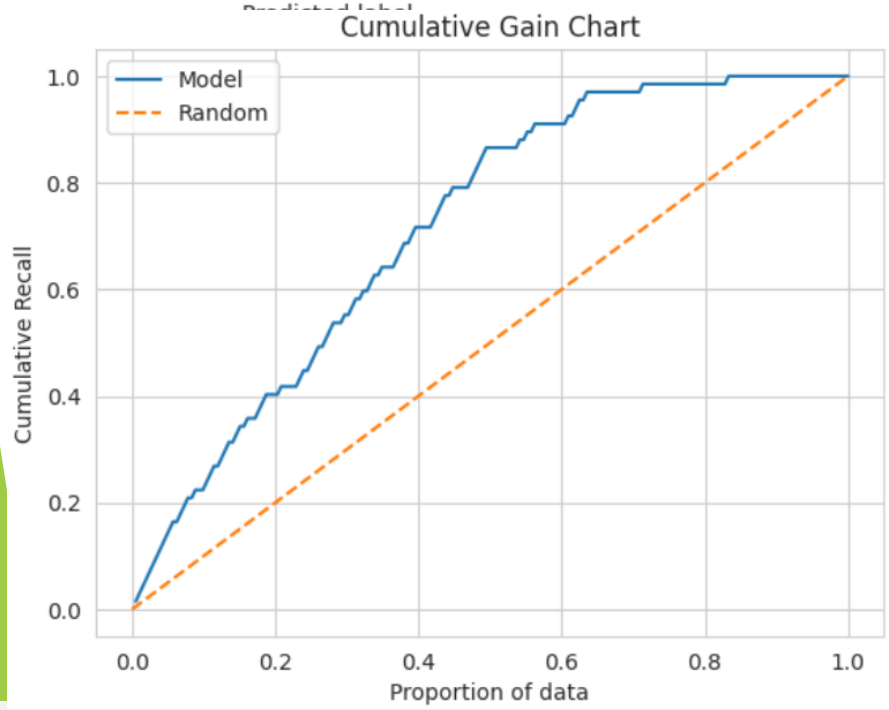
| | count | mean | std | min | 25% | \ |
|---|---|---|---|---|---|---|
| Pregnancies | 768.0 | 3.845052 | 3.369578 | 0.000 | 1.00000 | |
| Glucose | 768.0 | 121.656250 | 30.438286 | 44.000 | 99.75000 | |
| BloodPressure | 768.0 | 72.386719 | 12.096642 | 24.000 | 64.00000 | |
| SkinThickness | 768.0 | 27.334635 | 9.229014 | 7.000 | 23.00000 | |
| Insulin | 768.0 | 94.652344 | 105.547598 | 14.000 | 30.50000 | |
| BMI | 768.0 | 32.450911 | 6.875366 | 18.200 | 27.50000 | |
| DiabetesPedigreeFunction | 768.0 | 0.471876 | 0.331329 | 0.078 | 0.24375 | |
| Age | 768.0 | 33.240885 | 11.760232 | 21.000 | 24.00000 | |
| Outcome | 768.0 | 0.348958 | 0.476951 | 0.000 | 0.00000 | |

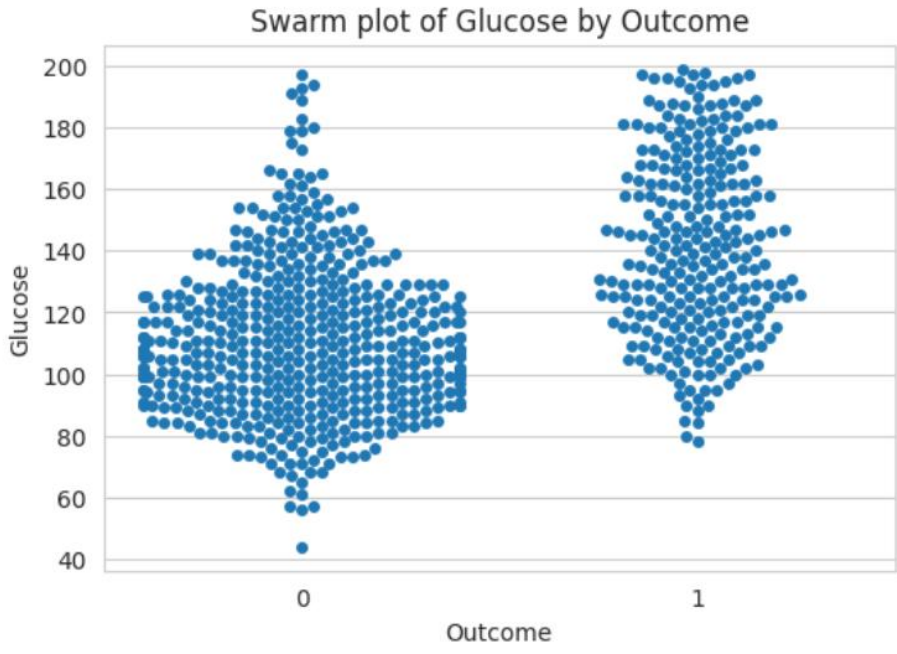
| | 50% | 75% | max |
|---|---|---|---|
| Pregnancies | 3.0000 | 6.00000 | 17.00 |
| Glucose | 117.0000 | 140.25000 | 199.00 |
| BloodPressure | 72.0000 | 80.00000 | 122.00 |
| SkinThickness | 23.0000 | 32.00000 | 99.00 |
| Insulin | 31.2500 | 127.25000 | 846.00 |
| BMI | 32.0000 | 36.60000 | 67.10 |
| DiabetesPedigreeFunction | 0.3725 | 0.62625 | 2.42 |
| Age | 29.0000 | 41.00000 | 81.00 |
| Outcome | 0.0000 | 1.00000 | 1.00 |

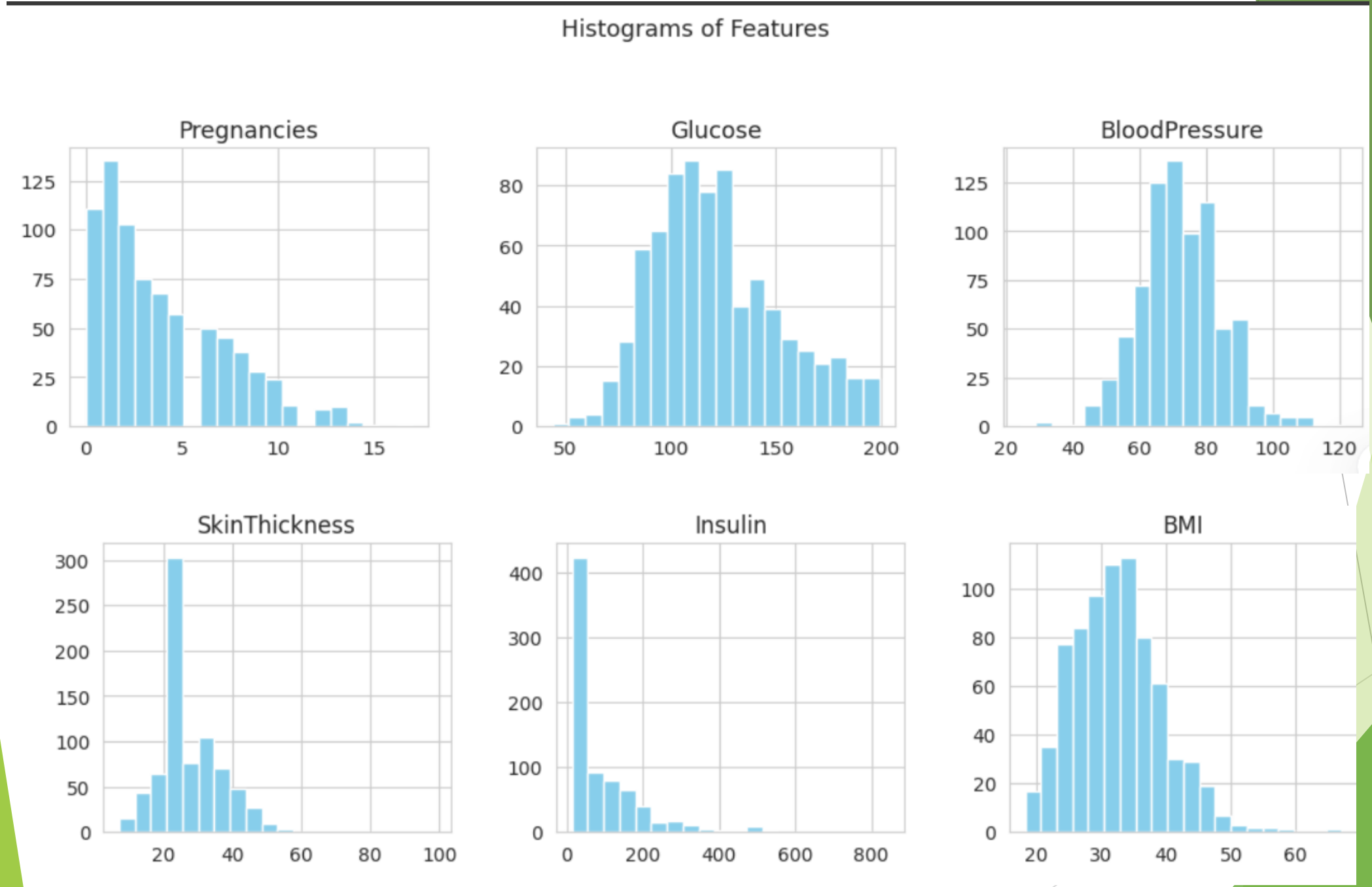

Confusion Matrix - LogisticRegression
True label
0
1
104
21
30
37
0
1
100
90
80
70
60
50
40
30

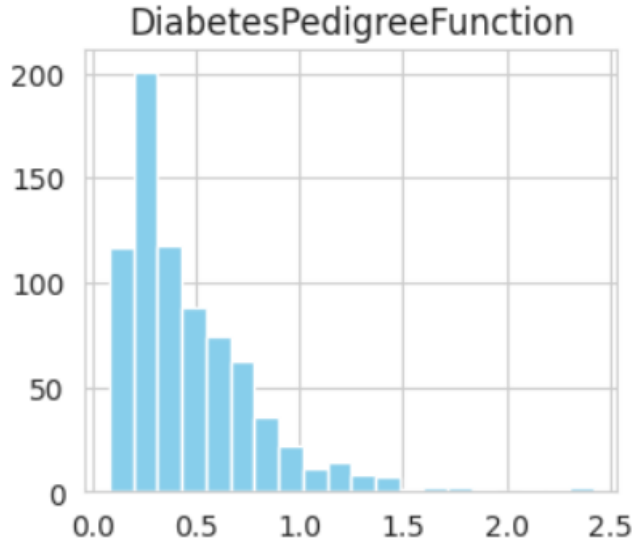

ROC Curve - LogisticRegression
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Classifier (AUC = 0.83)
0.0
0.2
0.4
0.6
0.8
1.0

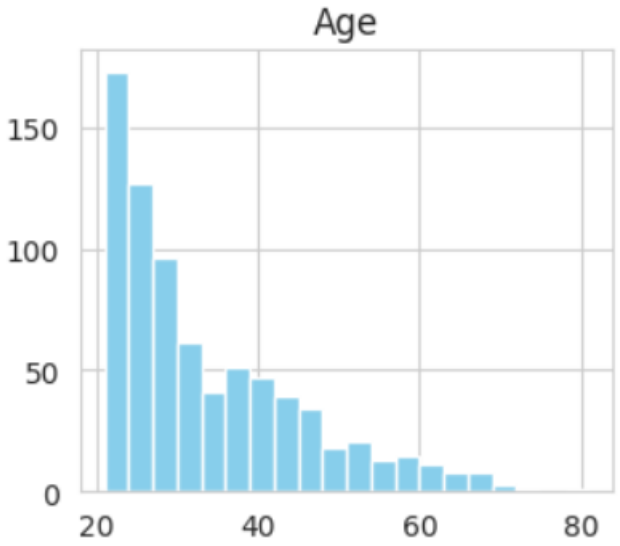

Precision (Positive label: 1)
Recall (Positive label: 1)
Classifier (AP = 0.72)
0.0
0.2
0.4
0.6
0.8
1.0

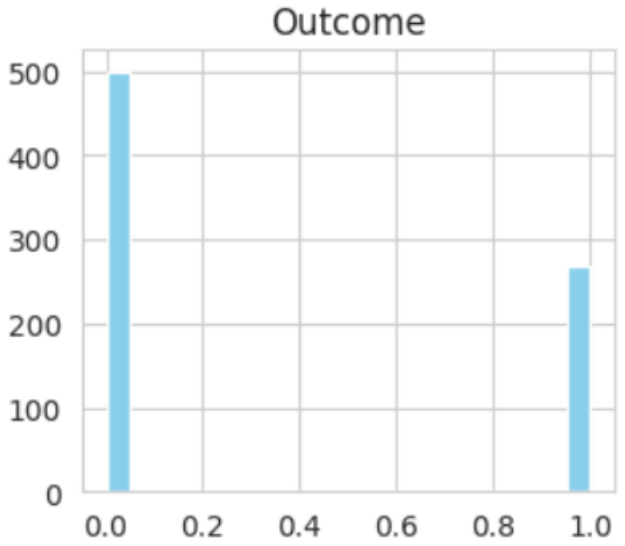

Cumulative Gain Chart
Model
Random
Cumulative Recall
Proportion of data
0.0
0.2
0.4
0.6
0.8
1.0

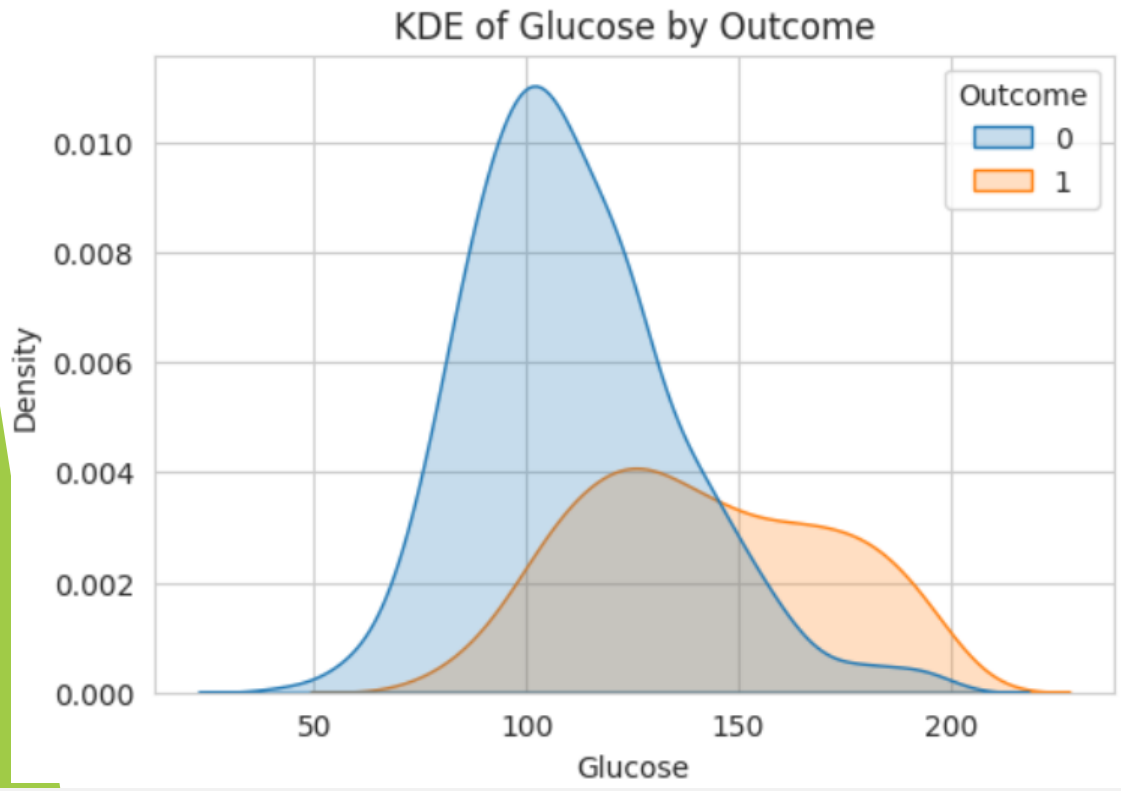

Swarm plot of Glucose by Outcome
Glucose
Outcome
200
180
160
140
120
100
80
60
40
0
1

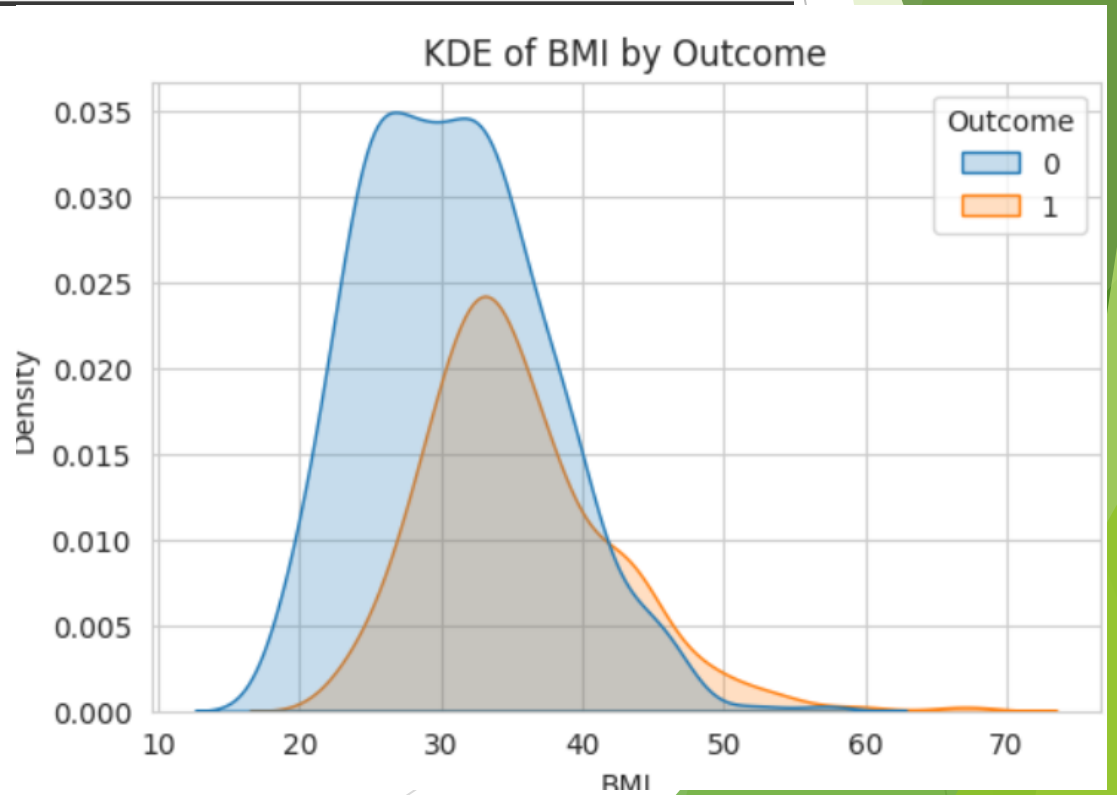

Histograms of Features
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI

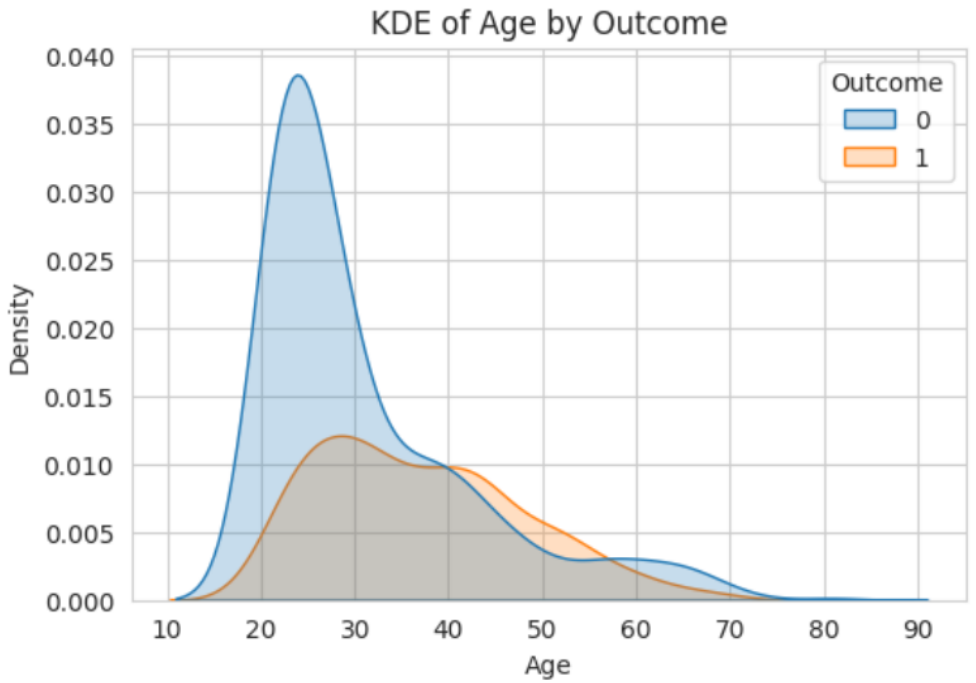

DiabetesPedigreeFunction
200
150
100
50
0
0.0
0.5
1.0
1.5
2.0
2.5

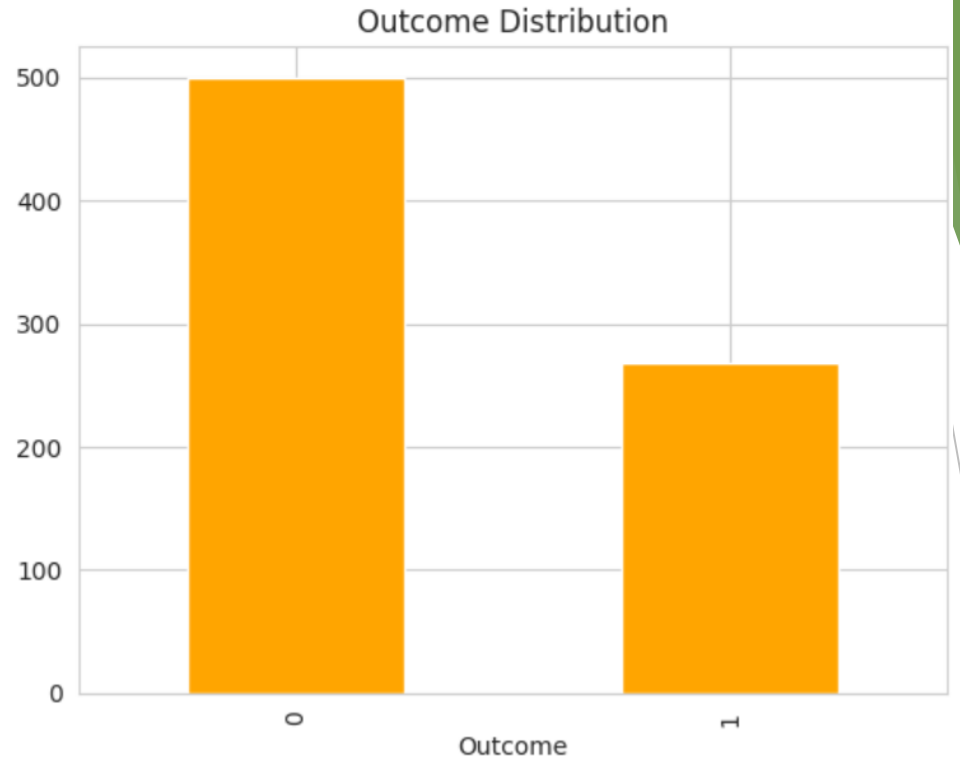

Age
150
100
50
0
20
40
60
80

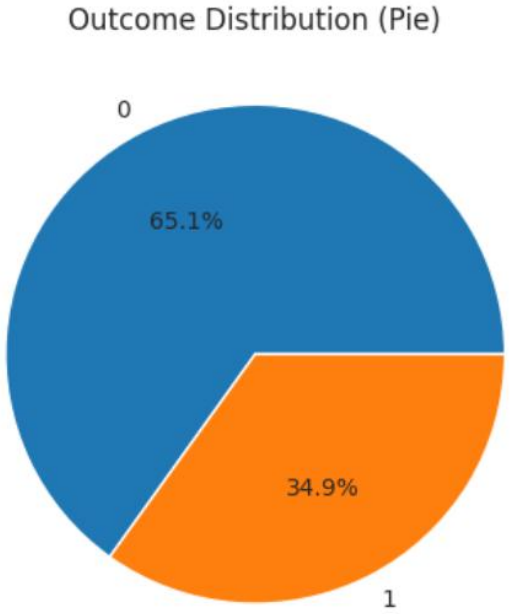

Outcome
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0

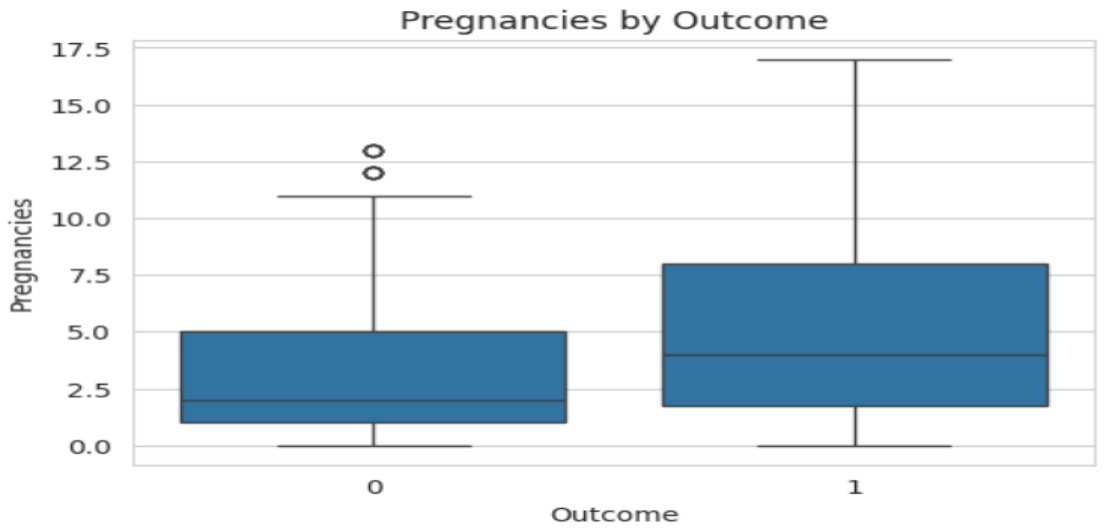

KDE of Glucose by Outcome
Outcome
0
1
Density
0.010
0.008
0.006
0.004
0.002
0.000
50
100
150
200
Glucose

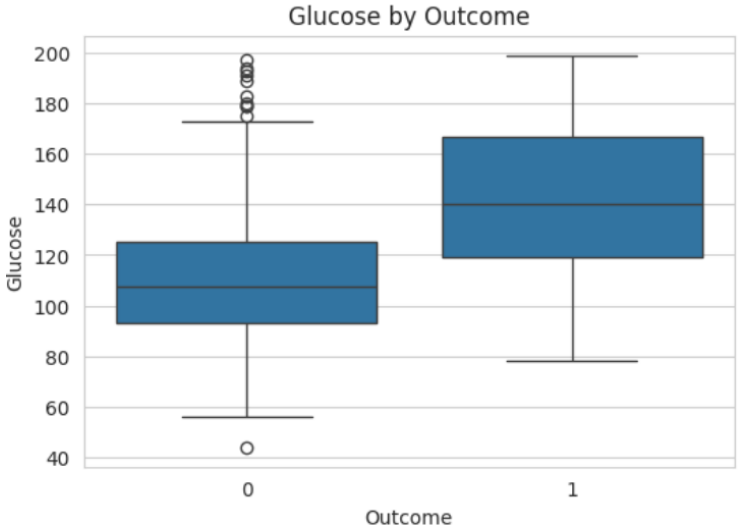

KDE of BMI by Outcome
Outcome
0
1
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
BMI

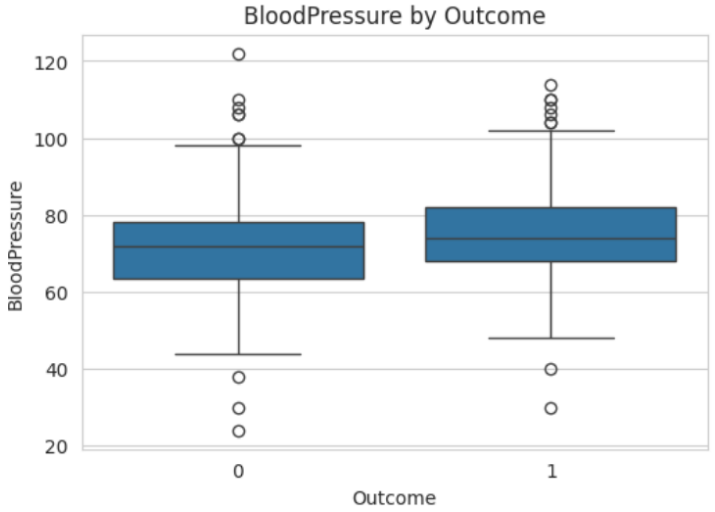

KDE of Age by Outcome
Outcome
0
1
Density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
80
90
Age

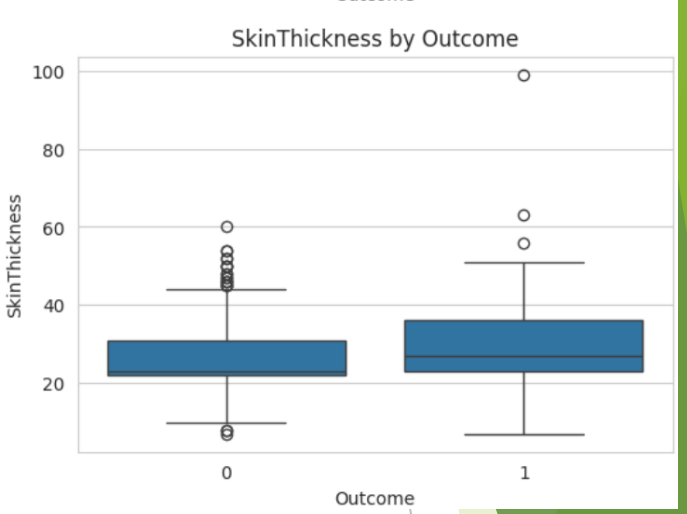

Outcome Distribution
500
400
300
200
100
0
0
1
Outcome

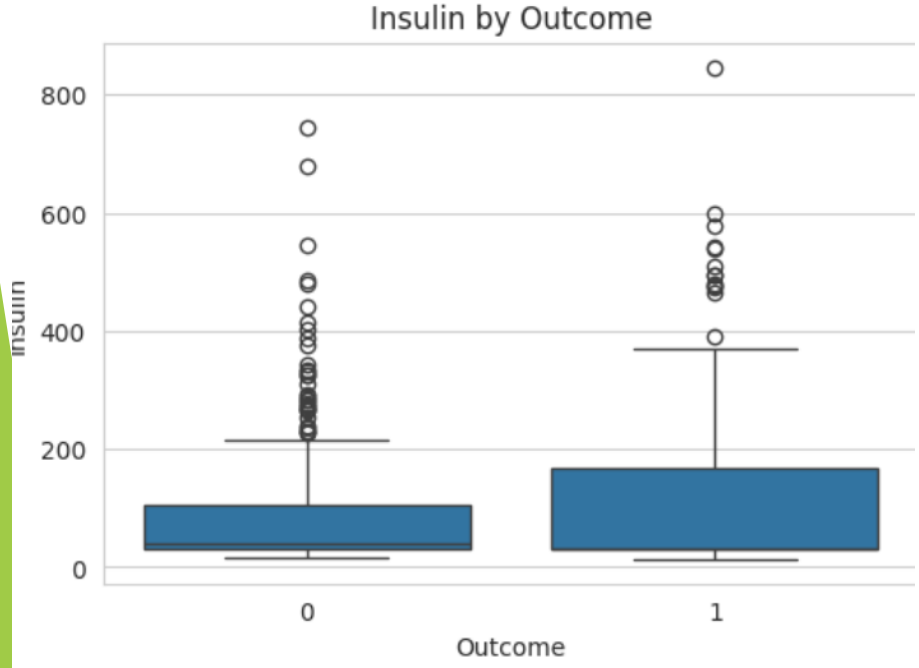

Outcome Distribution (Pie)
0
65.1%
34.9%
1

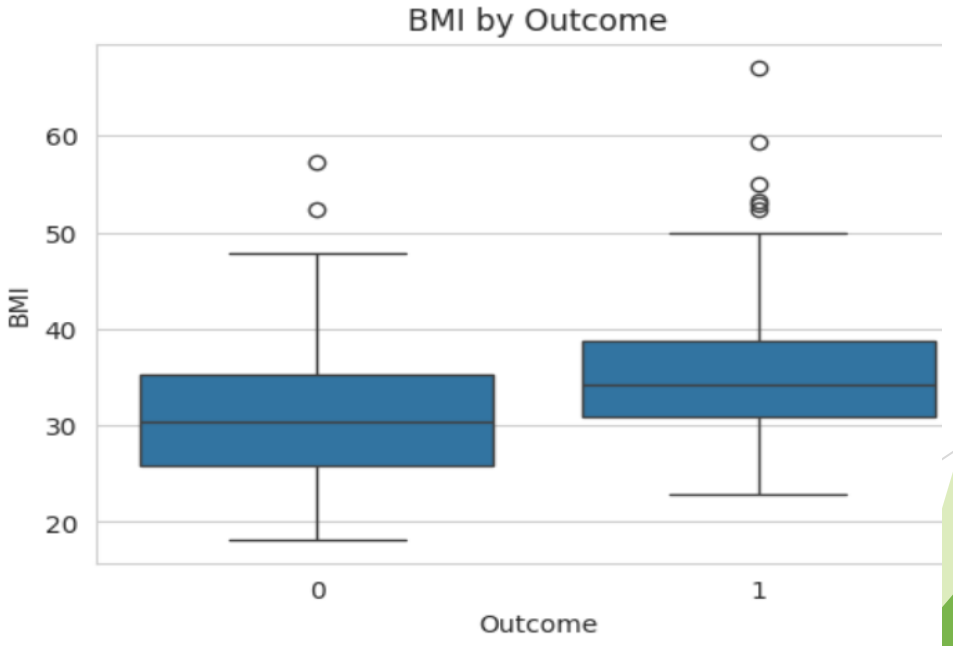

Pregnancies by Outcome
Pregnancies
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
1
Outcome

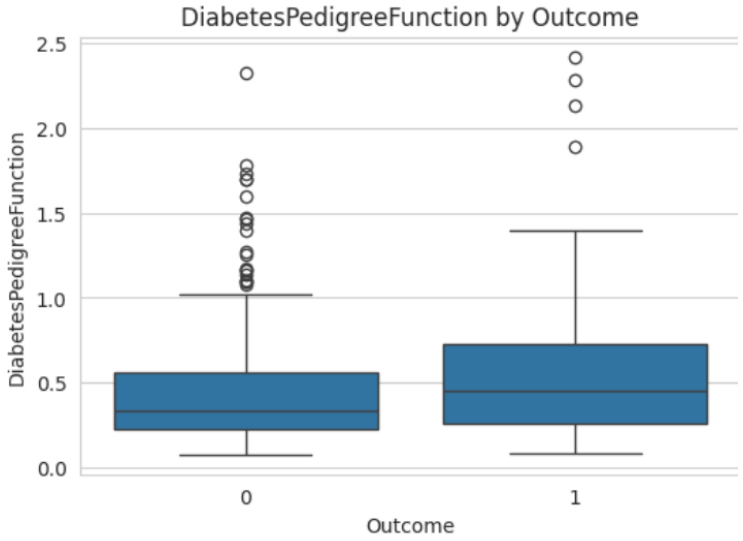

Glucose by Outcome
Glucose
Outcome
0
1

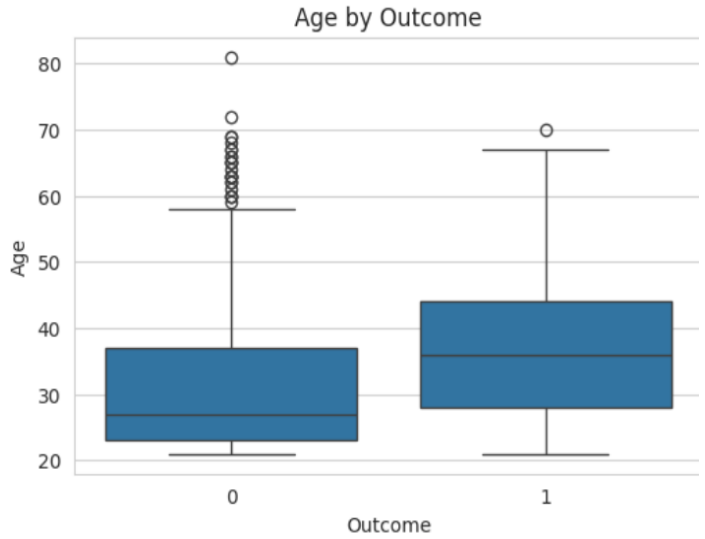

BloodPressure by Outcome
BloodPressure
Outcome
0
1


SkinThickness by Outcome
SkinThickness
Outcome
0
1

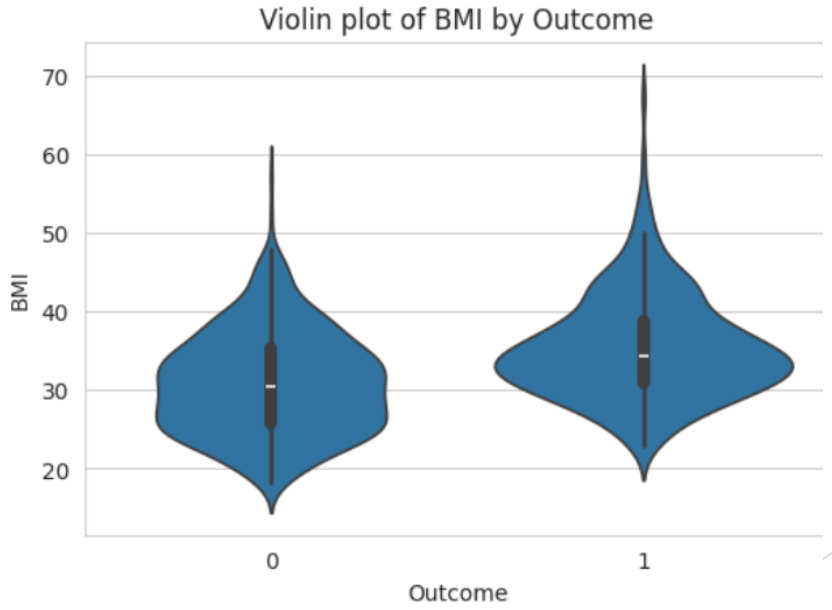

Insulin by Outcome
Insulin
Outcome
0
1


BMI by Outcome
BMI
Outcome
0
1

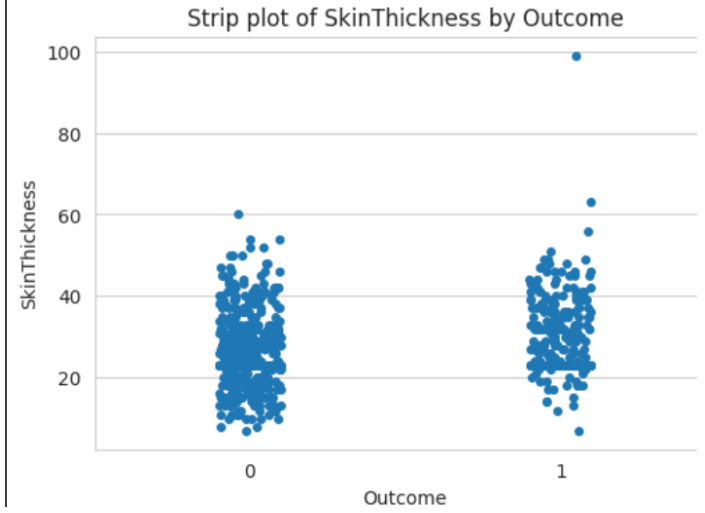

DiabetesPedigreeFunction by Outcome
DiabetesPedigreeFunction
2.5
2.0
1.5
1.0
0.5
0.0
0
1
Outcome

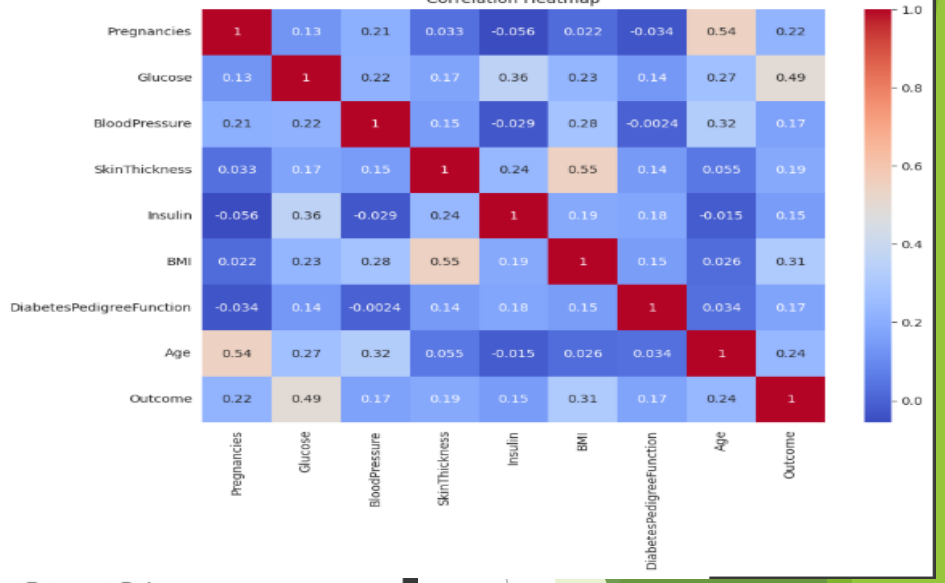

Age by Outcome
Age
80
70
60
50
40
30
20
0
1
Outcome

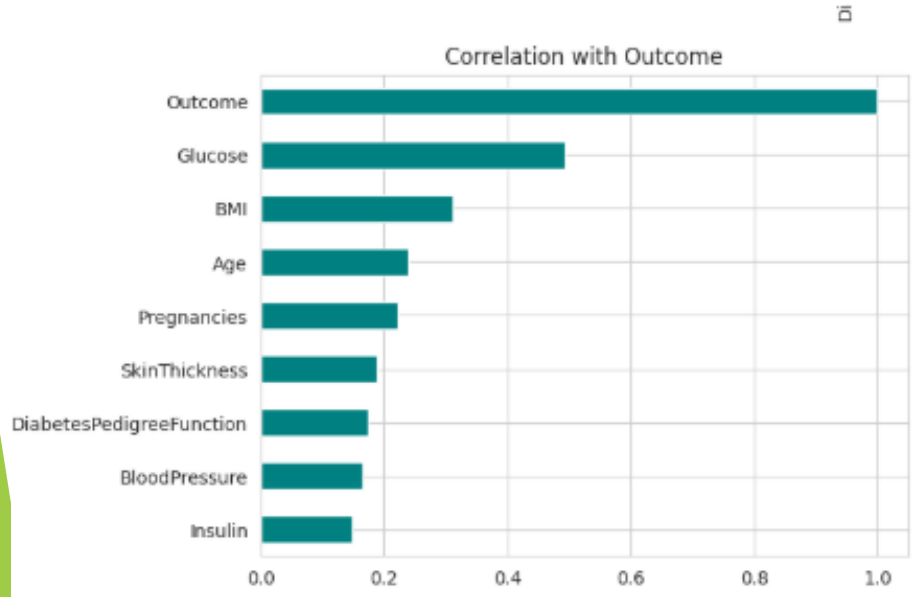

Violin plot of Glucose by Outcome
Glucose
225
200
175
150
125
100
75
50
25
0
1
Outcome

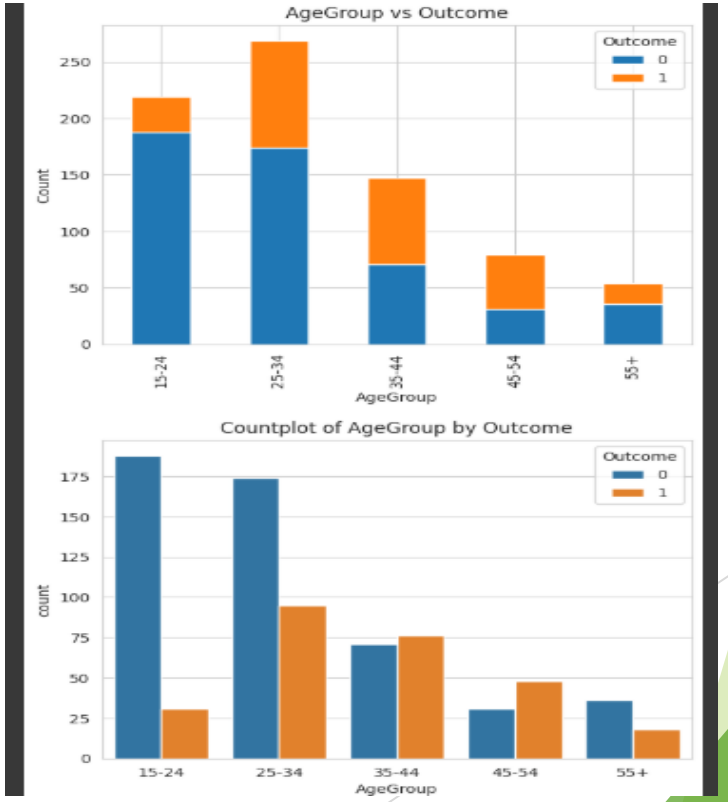

Violin plot of BMI by Outcome
BMI
70
60
50
40
30
20
0
1
Outcome


Strip plot of BloodPressure by Outcome
BloodPressure
Outcome
0
1
20
40
60
80
100
120


Strip plot of SkinThickness by Outcome
SkinThickness
Outcome
0
1
20
40
60
80
100


Pregnancies 1 0.13 0.21 0.033 -0.056 0.022 -0.034 0.54 0.22
Glucose 0.13 1 0.22 0.17 0.36 0.23 0.14 0.27 0.49
BloodPressure 0.21 0.22 1 0.15 -0.029 0.28 -0.0024 0.32 0.17
SkinThickness 0.033 0.17 0.15 1 0.24 0.55 0.14 0.055 0.19
Insulin -0.056 0.36 -0.029 0.24 1 0.19 0.18 -0.015 0.15
BMI 0.022 0.23 0.28 0.55 0.19 1 0.15 0.026 0.31
DiabetesPedigreeFunction -0.034 0.14 -0.0024 0.14 0.18 0.15 1 0.034 0.17
Age 0.54 0.27 0.32 0.055 -0.015 0.026 0.034 1 0.24
Outcome 0.22 0.49 0.17 0.19 0.15 0.31 0.17 0.24 1
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0


Correlation with Outcome
Outcome
Glucose
BMI
Age
Pregnancies
SkinThickness
DiabetesPedigreeFunction
BloodPressure
Insulin
0.0
0.2
0.4
0.6
0.8
1.0


AgeGroup vs Outcome
Outcome
0
1
Count
AgeGroup
15-24
25-34
35-44
45-54
55+
Countplot of AgeGroup by Outcome
count

# Thank You

SOUMYAJITA BOSE - IITP000180

soumyajita_2312res655@iitp.ac.in