Diabetes Detection
ML Classification

Business & Decision
Life Sciences

CONTENT

- Business Discovery
- Data Preparation
- Data Preprocessing
- Model Building
- Model’s performances
- User Experience
- Openings

Business Discovery

Business Domain

AI in medicine market value: from USD 4490.3 million in 2020
Expectations by 2026: and is expected to reach USD 34882.58 million.
Challenges : Data repositories, Regulations and Integration

Frame the business problem as an analytics challenge

As an AI Consultant working for Business&Decision, we can applicate big data solutions to improve medical diagnostics. In fact, based on ML classification, we want to incorporate data from diabetes in order to create a fuller picture of the user preferences and potential needs about all our solutions, through a Diabetes Detection API.

Data Preparation

- **Assess the resources available to support the project**

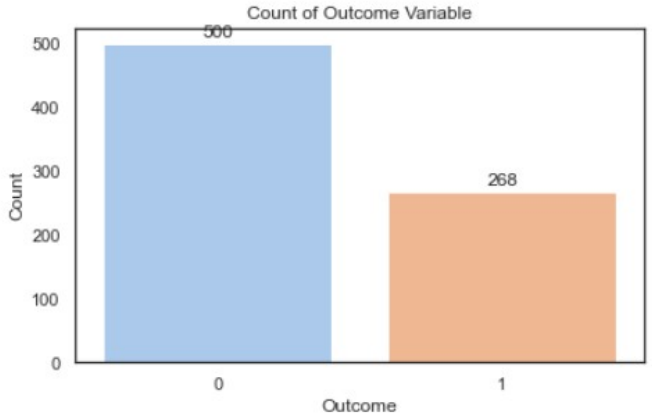
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

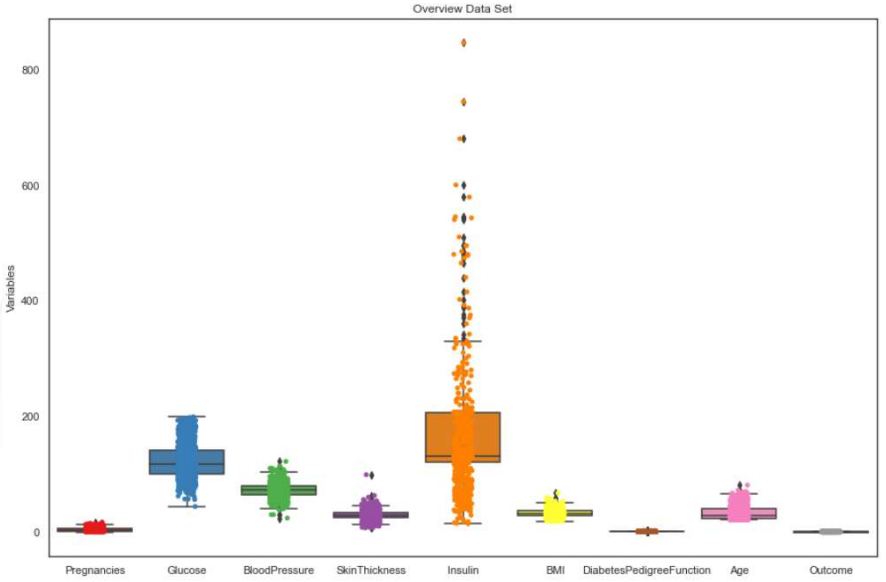
Data Preprocessing (& Visualization)

- **Replacing missing values**
- **Overview of the dataset (Boxplots Summary)**
- **Viewing the distribution of the target variable**

```
df.isnull().sum()
```

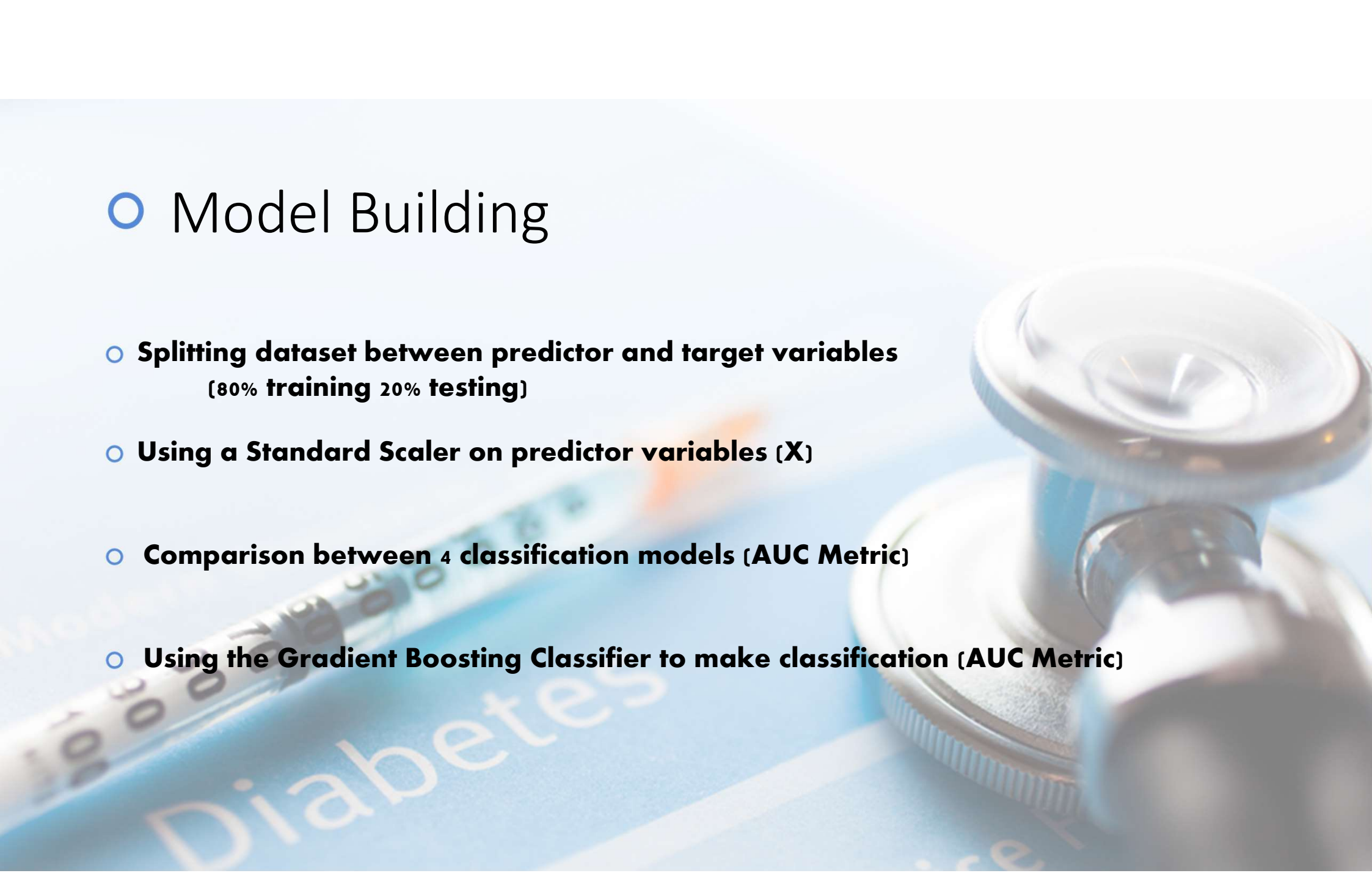
Pregnancies	0
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	0
dtype: int64	





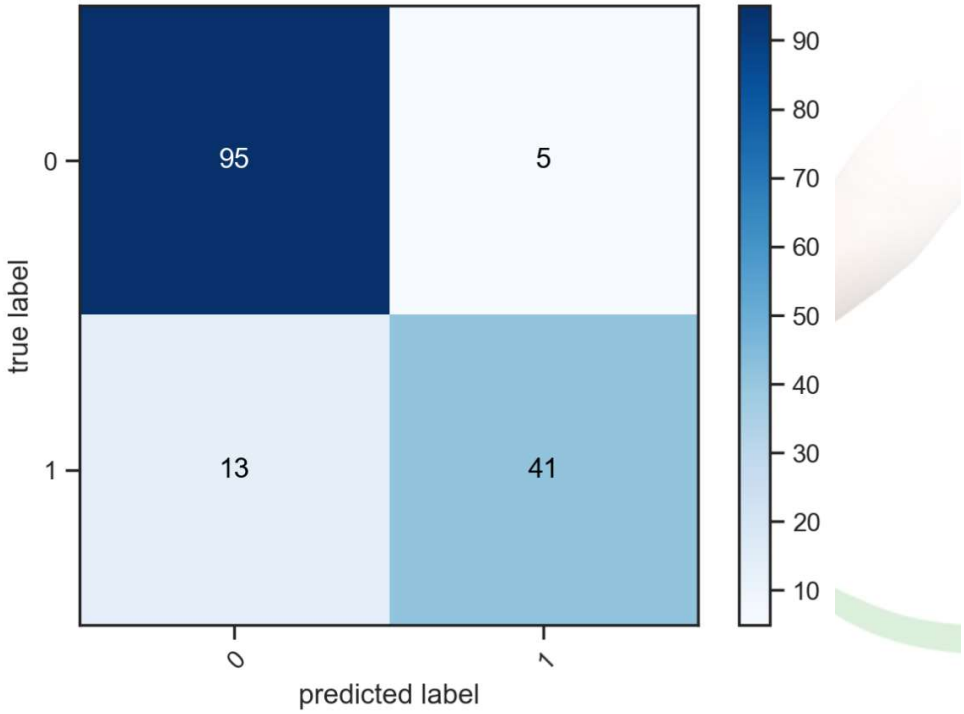
Model Building

- **Splitting dataset between predictor and target variables (80% training 20% testing)**
- **Using a Standard Scaler on predictor variables (X)**
- **Comparison between 4 classification models (AUC Metric)**
- **Using the Gradient Boosting Classifier to make classification (AUC Metric)**

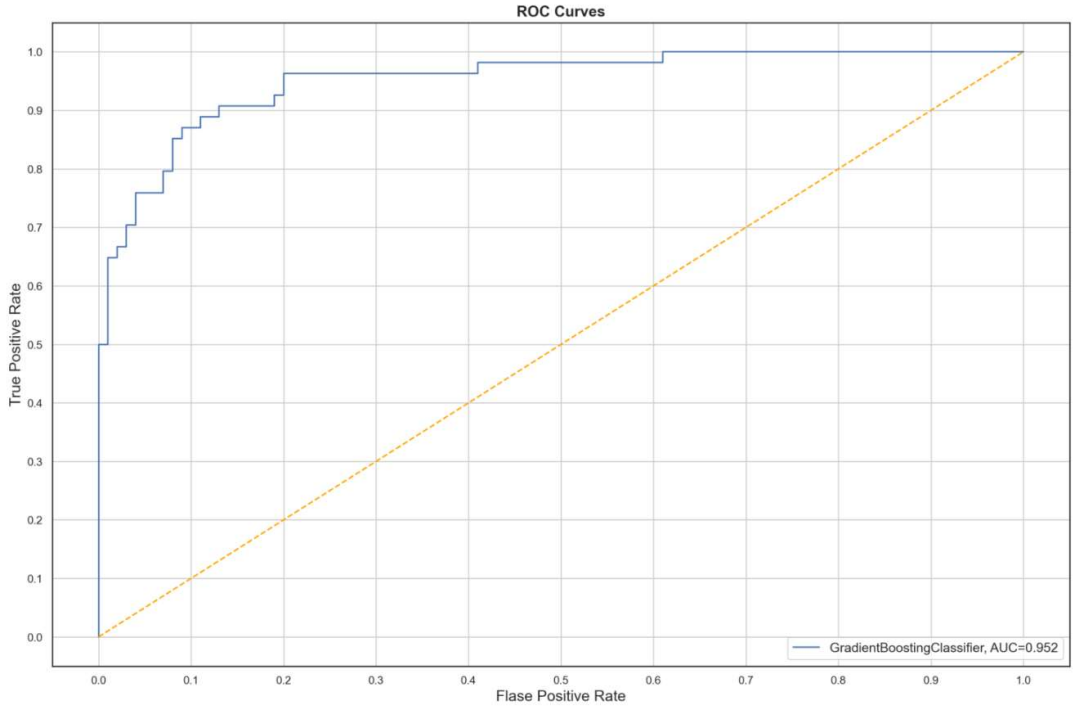


Model's Performances

- **Confusion Matrix**



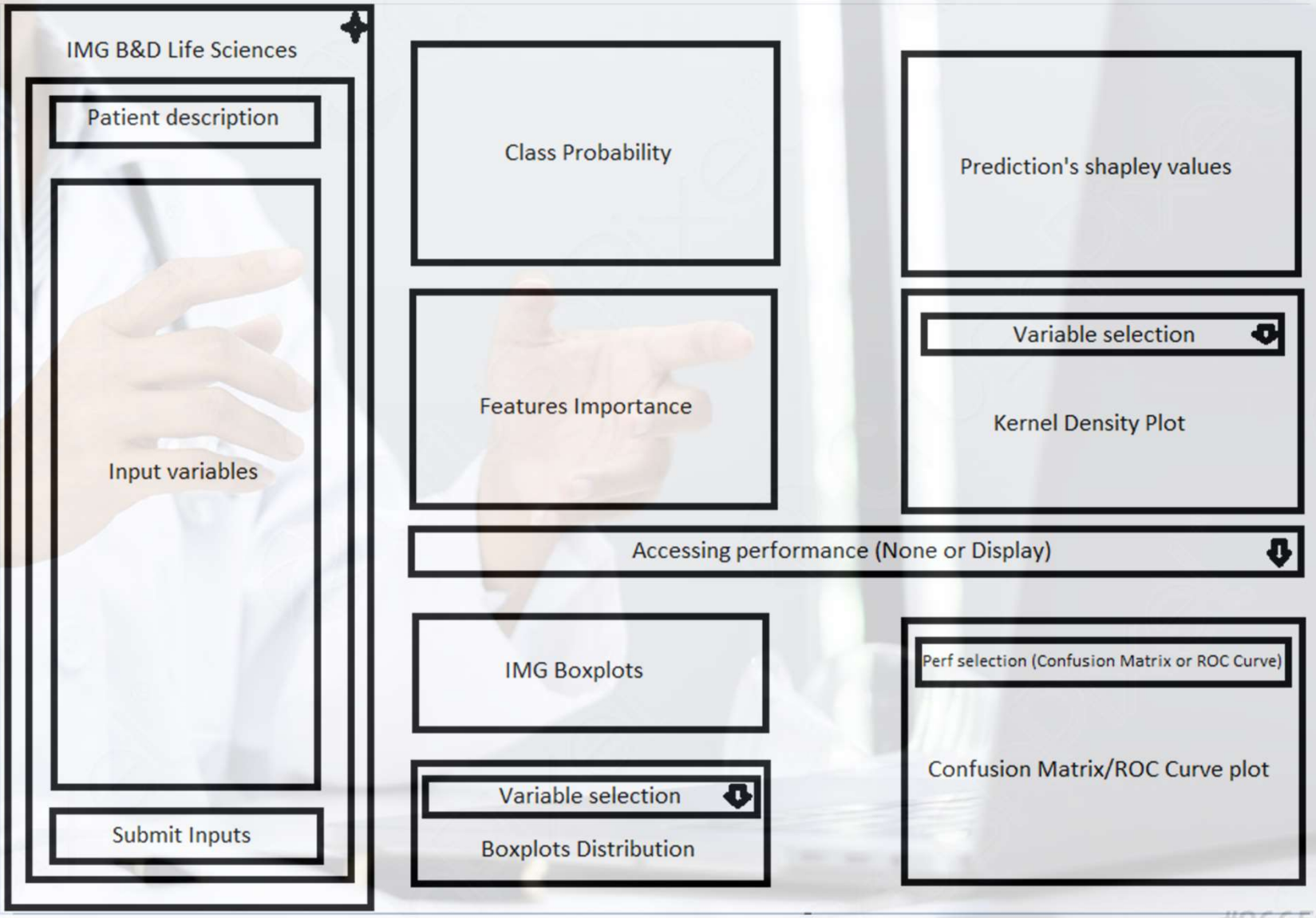
- **ROC Curve on Test Set (AUC=0.952)**



User Experience / User Interface

- **Using streamlit to build interactive data app**
- **Unique layout design**
- **Batch inference**

App Mock-Up


IMG B&D Life Sciences
Patient description
Input variables
Submit Inputs
Class Probability
Prediction's shapley values
Features Importance
Variable selection
Kernel Density Plot
Accessing performance (None or Display)
IMG Boxplots
Perf selection (Confusion Matrix or ROC Curve)
Variable selection
Boxplots Distribution
Confusion Matrix/ROC Curve plot

Sample View

Business & Decision Life Sciences

PLEASE INSERT PATIENT'S INFORMATION

pregnancies

0

glucose

121.69

blood_pressure

72.42

skin_thickness

DIABETES DETECTION

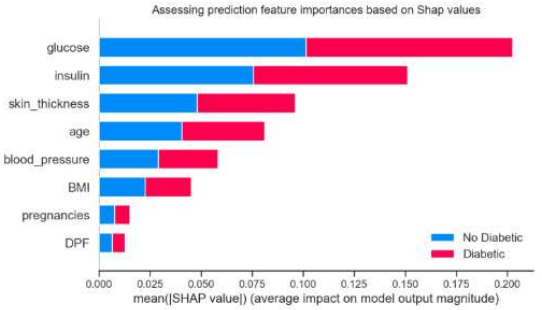
Class PROBABILITY in %

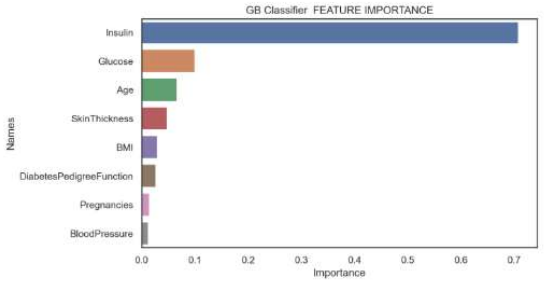
Diabetic

50.91

No Diabetic

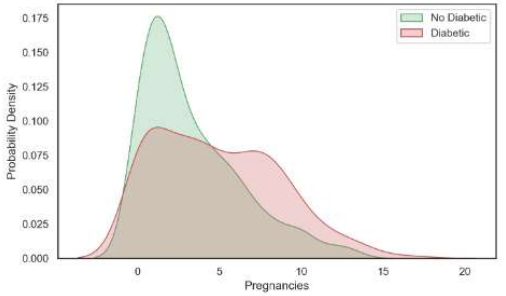
49.09





VARIABLE SELECTION

Pregnancies



Select Variable

Sample View

0

glucose

121.69

blood_pressure

72.42

skin_thickness

29.24

insulin

156.99

BMI

32.44

DPF

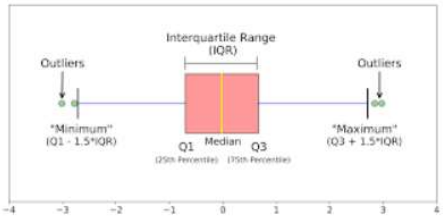
0.47

age

33

Submit Inputs

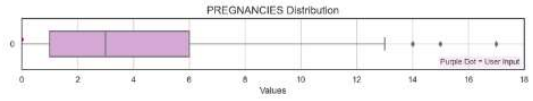
Display



Read a Boxplot

VARIABLE SELECTION

Pregnancies

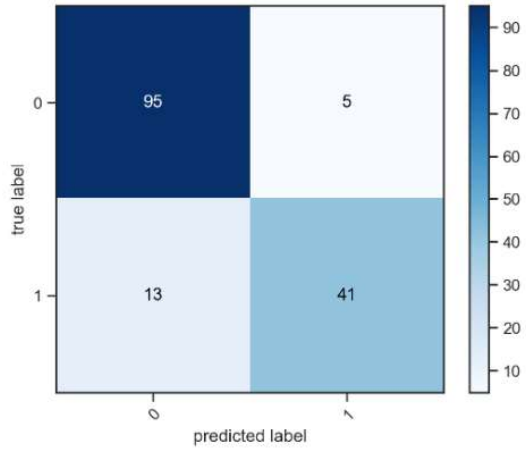


Select Variable

PERFORMANCE SELECTION

Confusion Matrix

CONFUSION MATRIX



Sensitivity: 0.7592592592592593

Specificity: 0.95

Select Performance

Openings

Code optimization

Need to implement more functions to lighten the code and do the pep8 check (python coding convention)

Model Optimization

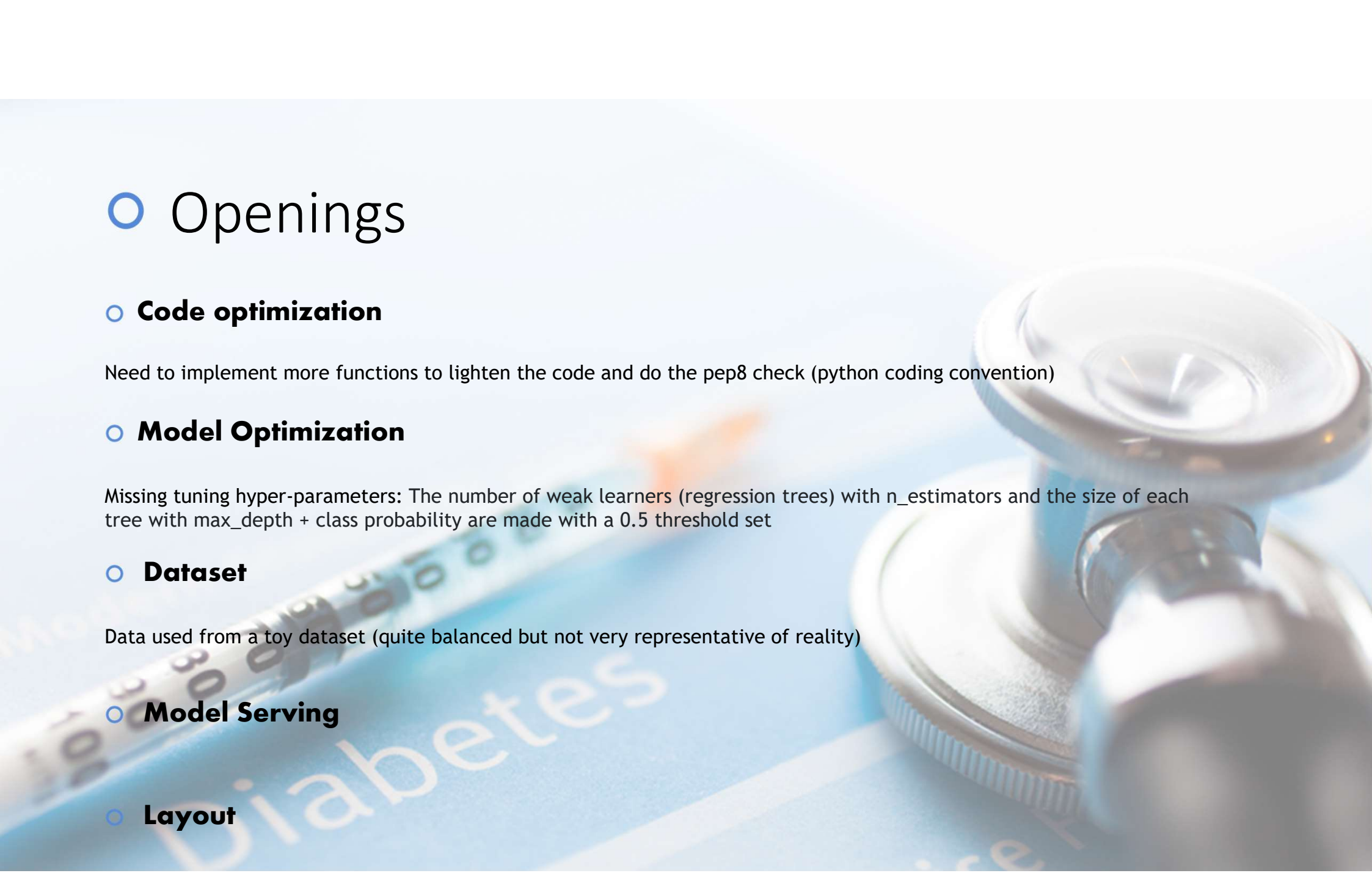
Missing tuning hyper-parameters: The number of weak learners (regression trees) with n_estimators and the size of each tree with max_depth + class probability are made with a 0.5 threshold set

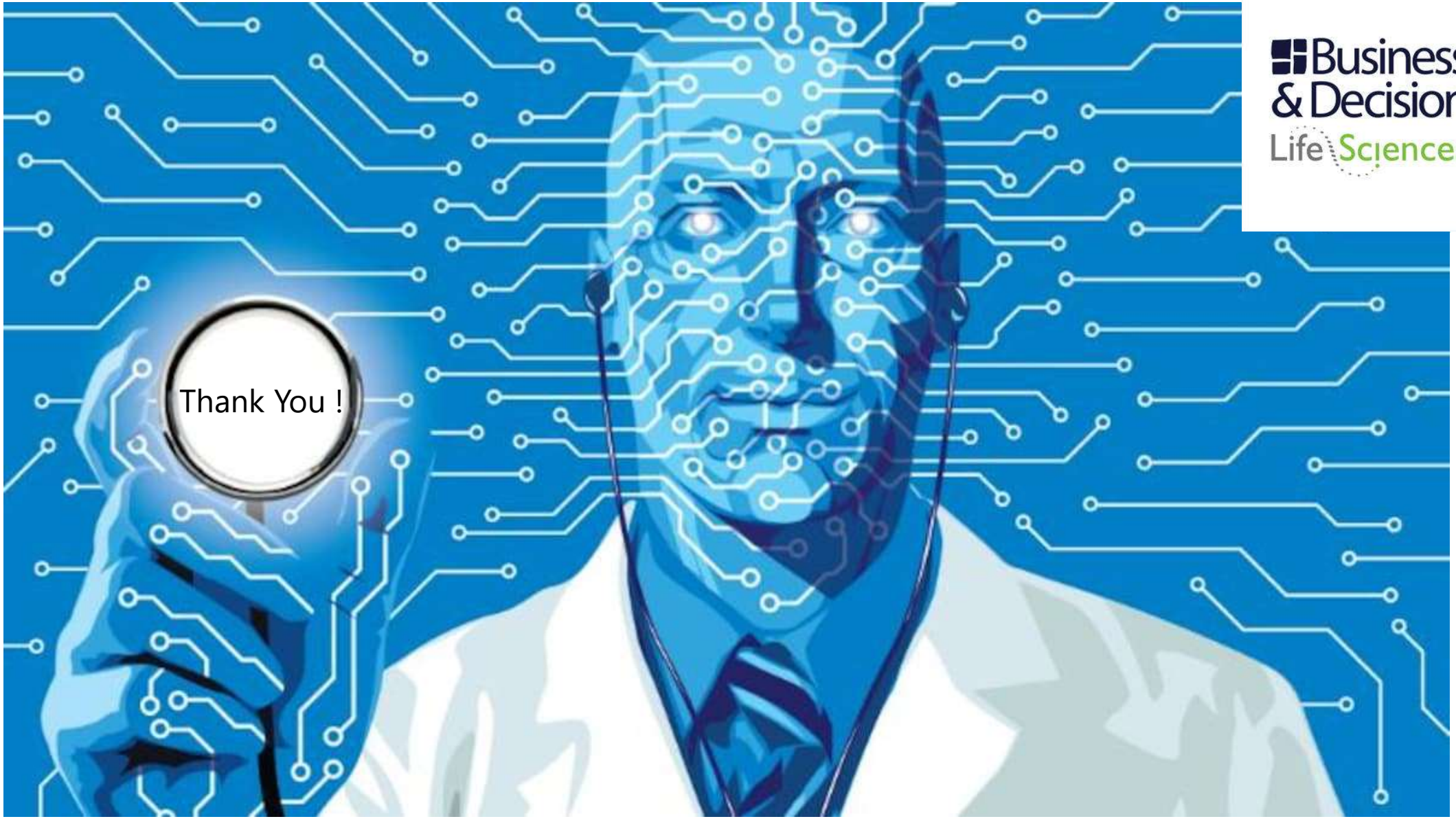
Dataset

Data used from a toy dataset (quite balanced but not very representative of reality)

Model Serving

Layout





Business & Decision
Life Sciences

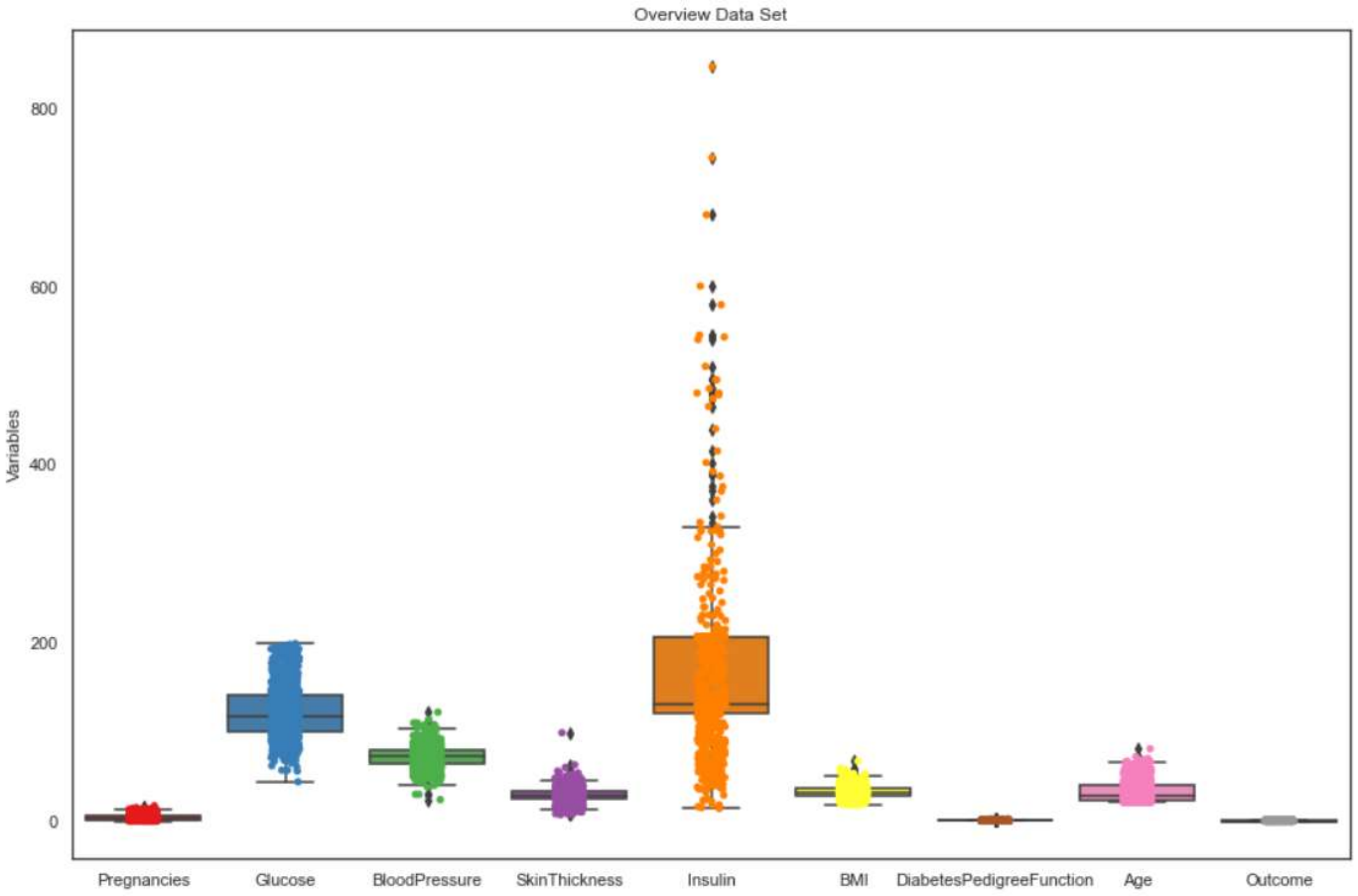
Appendix

Pima Indians Dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

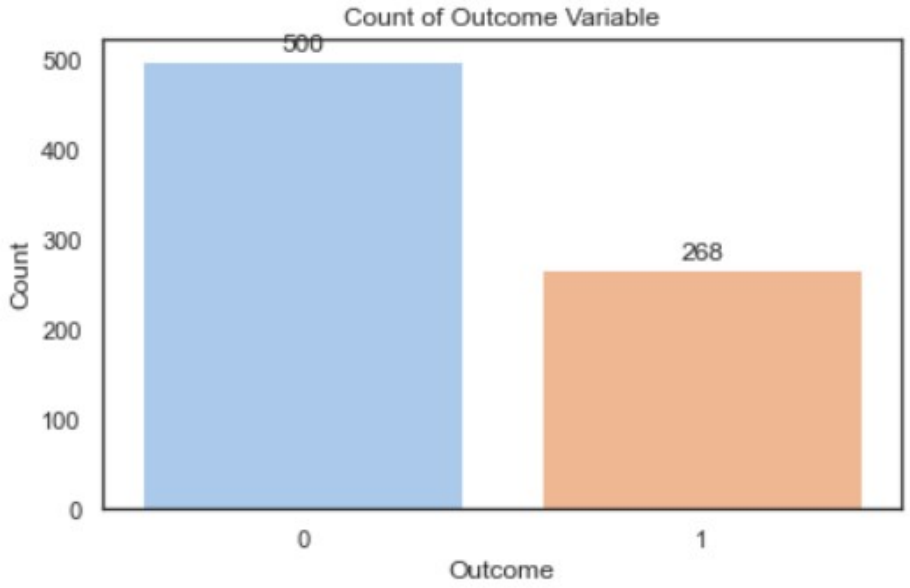
Appendix

Dataset Overview



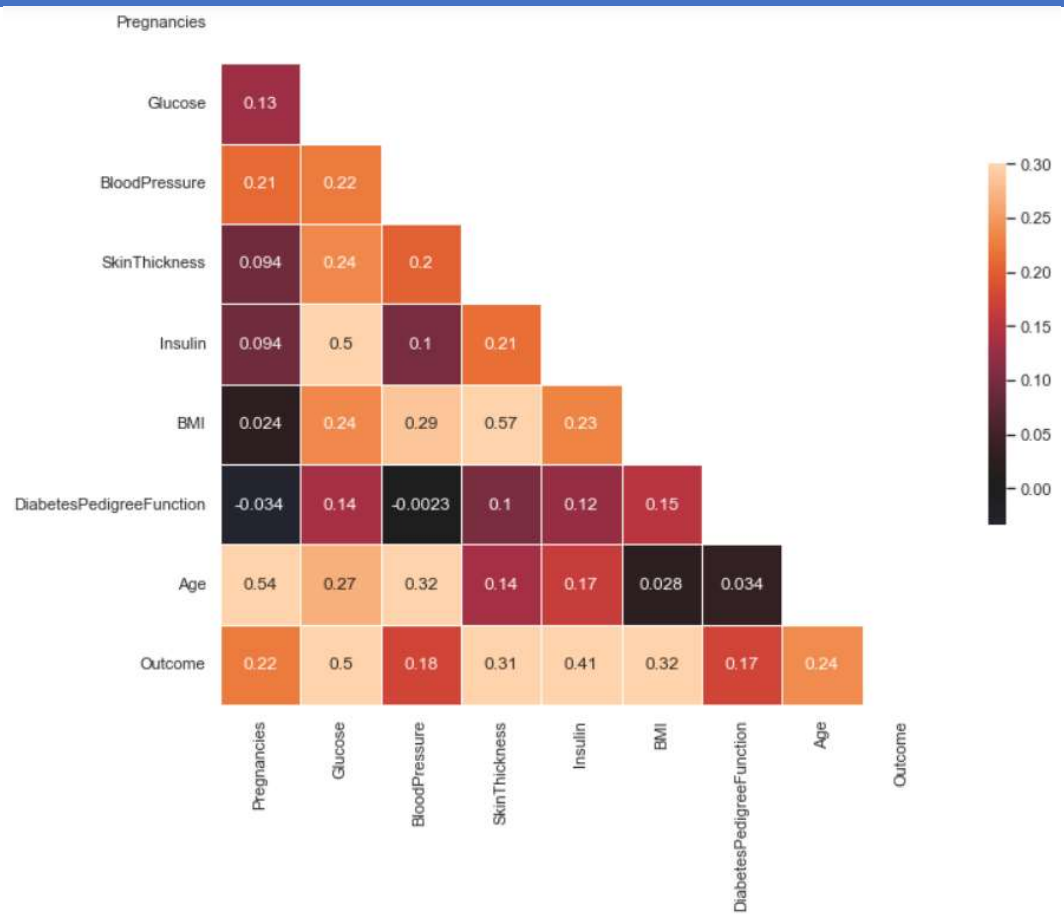
Appendix

Data Imbalance



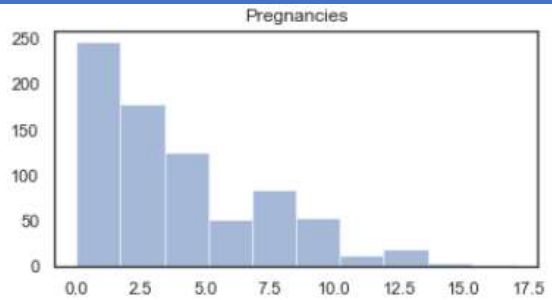
Appendix

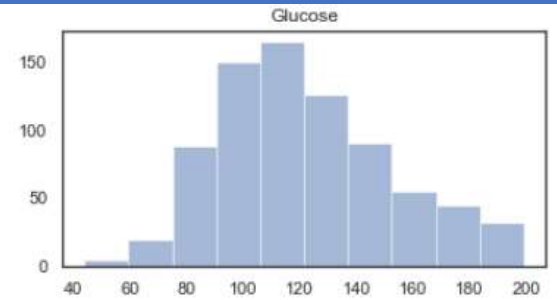
Correlation plot

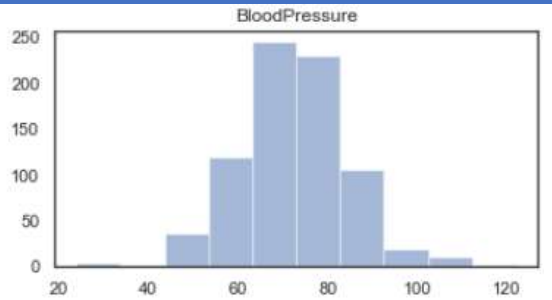


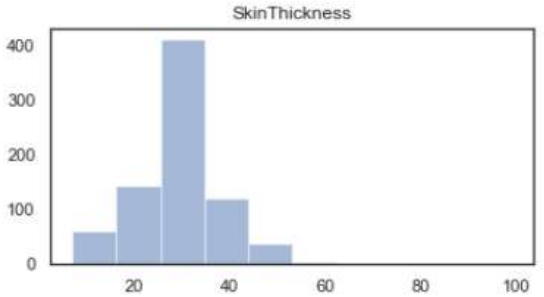
Appendix

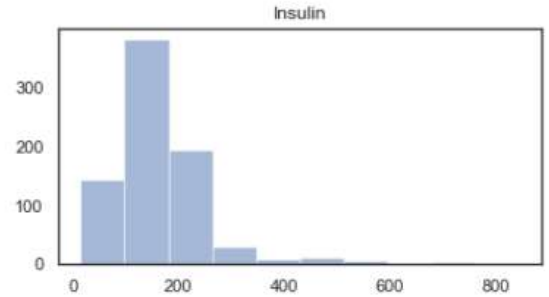
Distribution/Variable

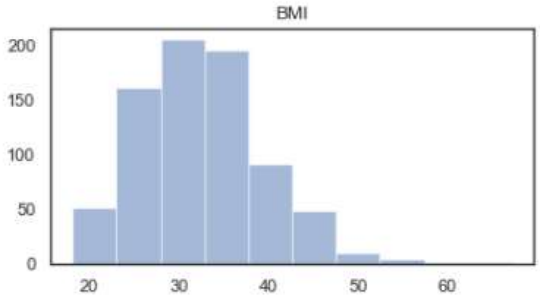


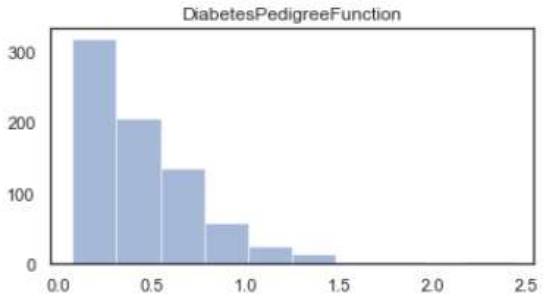


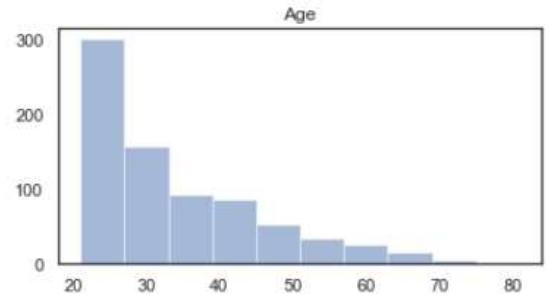


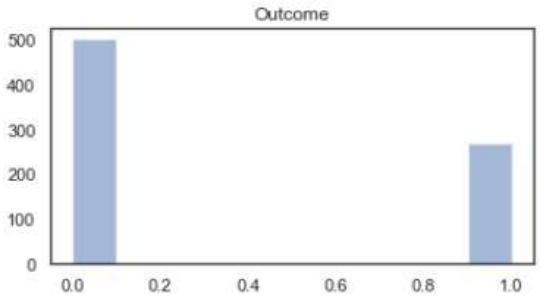






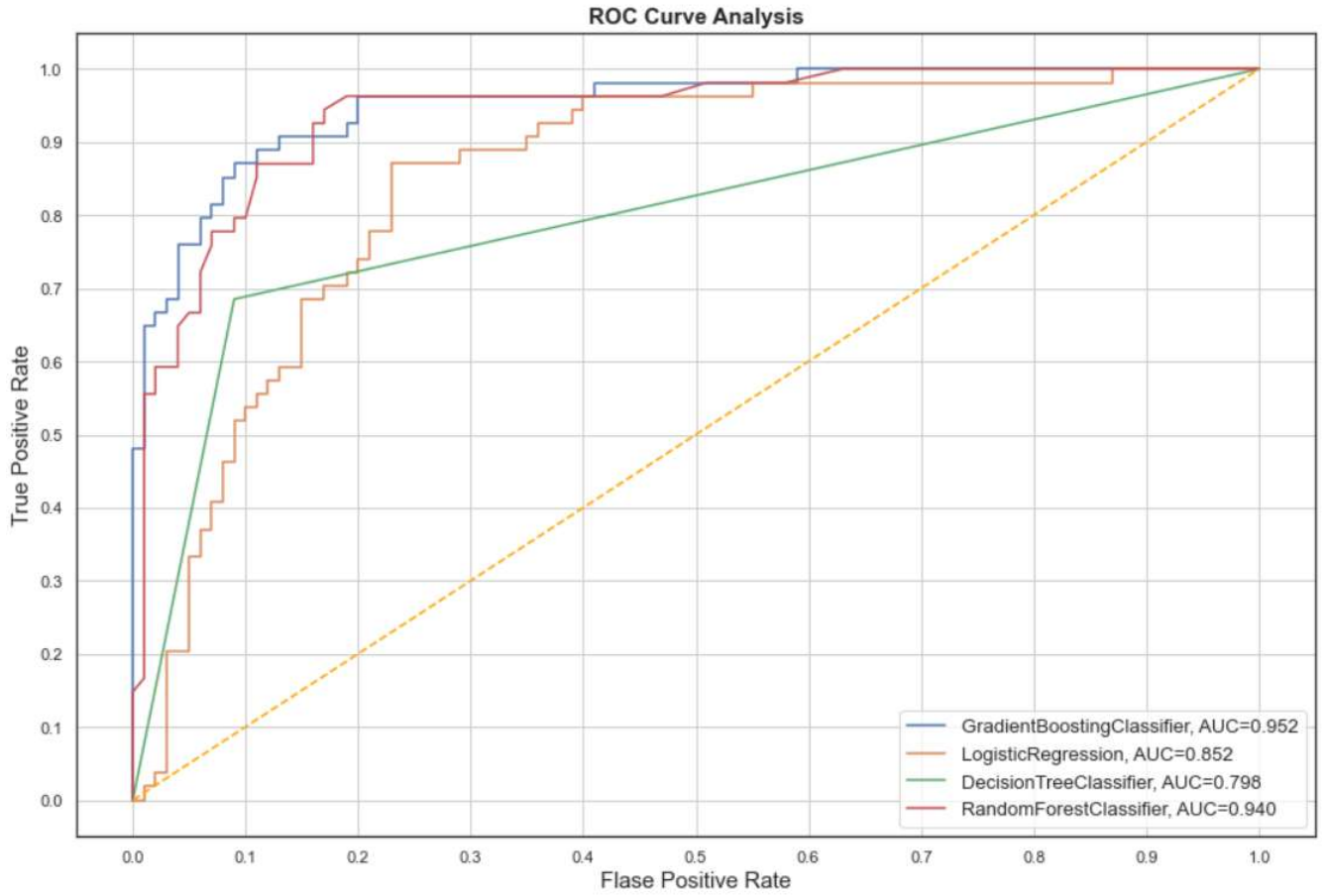






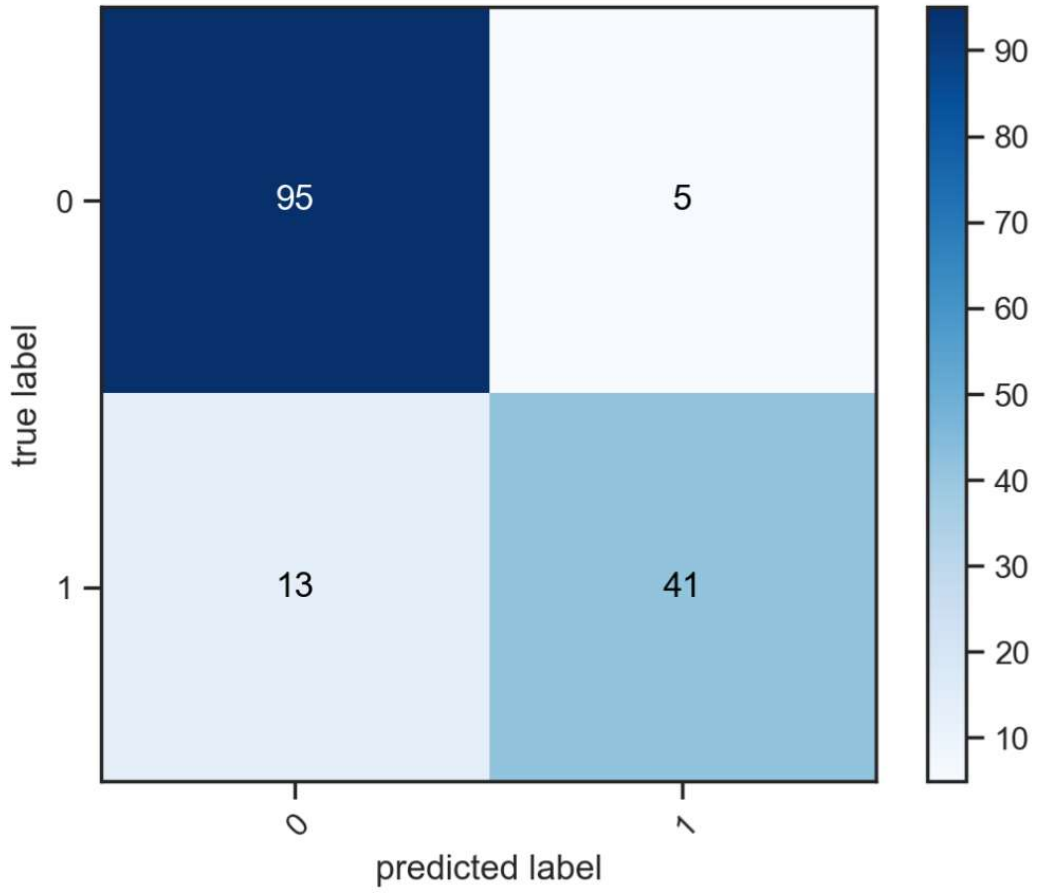
Appendix

ROC Curve Model Comparison



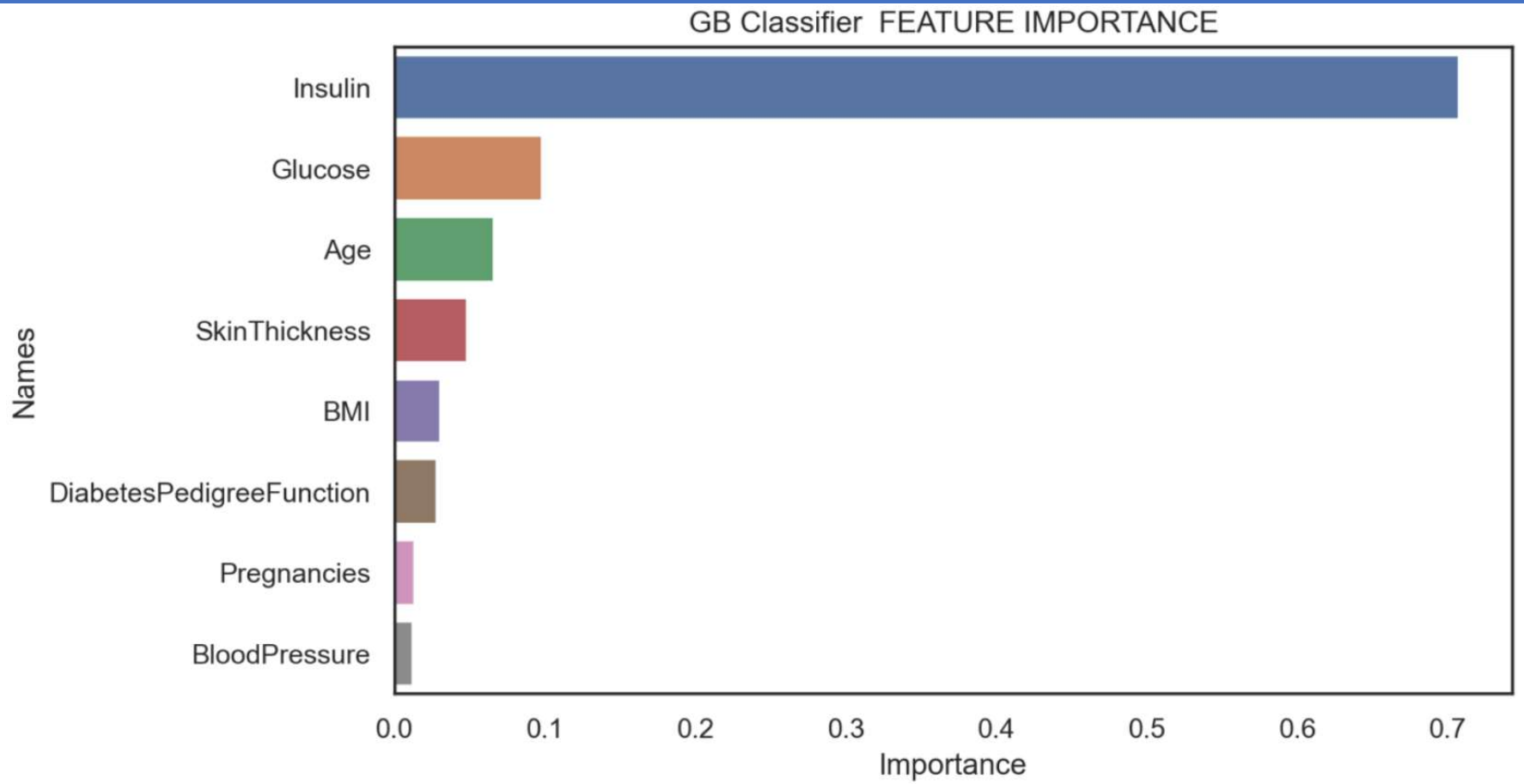
Appendix

GB Classifier Confusion Matrix



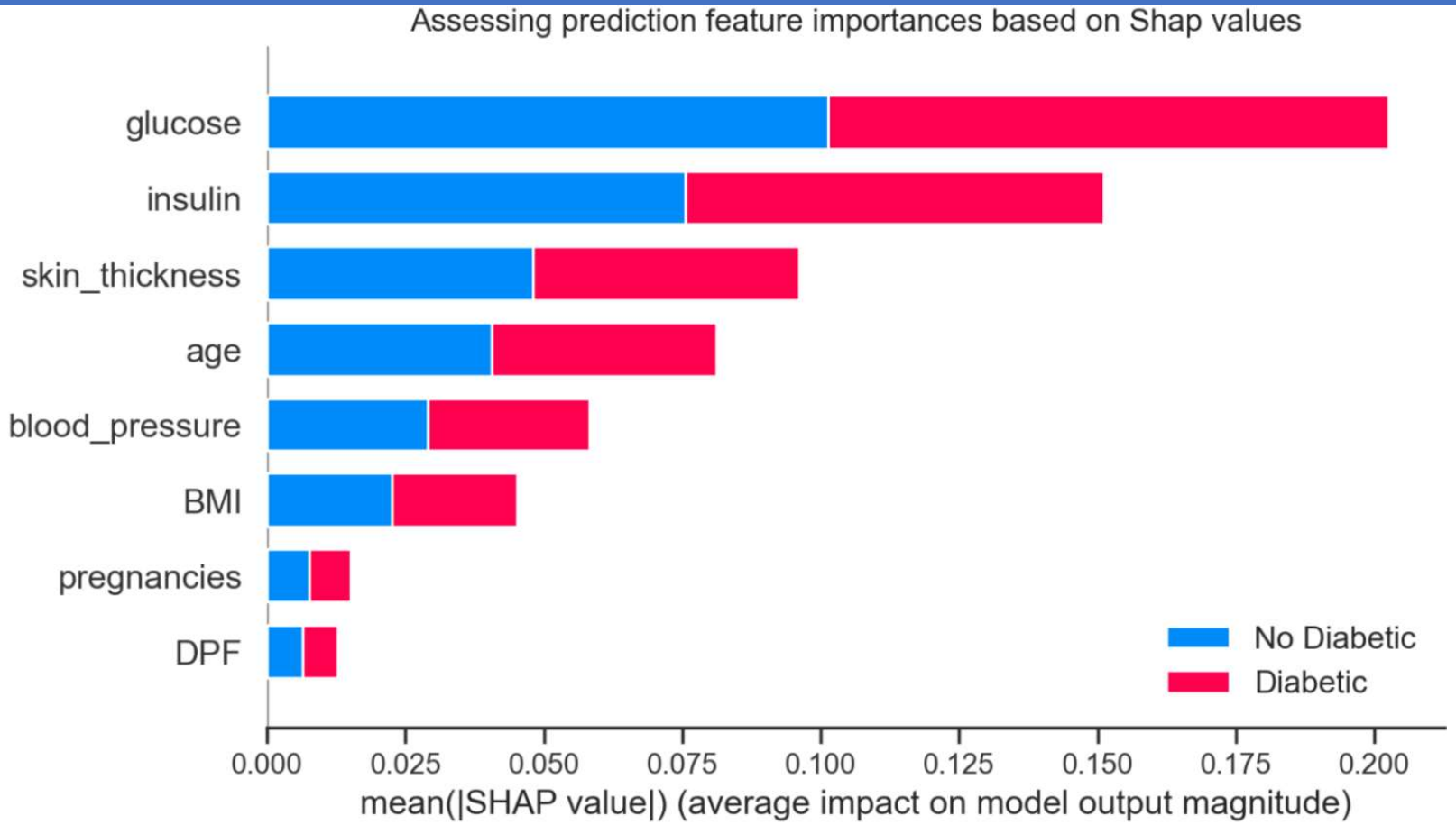
Appendix

GB Classifier Feature Importance



Appendix

Predicition’s Shap Values



Appendix

Kernel Density Plot (Pregnancies Variable)