Diabetes Prediction Using Machine Learning Algorithms

Alişan Köroğlu
2023900252

Agenda

- What is Diabetes?
- Type 2 Diabetes
- Pima Indian Diabetes Dataset
- Imputation
- Performance of Models

What is Diabetes?

- A condition when body cannot properly regulate blood sugar(glucose) levels.
- This happens because the body either doesn't produce enough insulin or cannot effectively use the insulin it produces.

Types of Diabetes:

- Type 1 Diabetes: The body’s immune systemattacks and destroy insülin cells in the pancreeas. Patients need lifelong insulin therapy.
- Type 2 Diabetes: The body becomes resistant to insülin or doesn’t produce enough. Often managed through diet, exercise, medication, and sometimes insulin.
- Gestationan Diabetes: Occurs during pregnancy and usualy resolves after childbirth. Increases the risk of developing Type 2 diabetes later in life.
- Prediabetes: Blood sugar levels are higher than normal but not high enough to be diagnosed as diabetes. A warning sign that Type 2 diabetes may develop without lifestyle changes.

Type 2 Diabetes

- Type 2 diabetes is a significant global health concern, with its prevalence increasing rapidly over the past few decades.
- **Prevalence**: As of 2024, approximately 537 million adults (20-79 years) worldwide are living with diabetes, with over 90% of these cases being type 2 diabetes. (IDF Diabetes Atlas)
- **Projected Increase**: The number of adults with diabetes is expected to rise to 783 million by 2045, indicating a significant global health challenge. (IDF Diabetes Atlas)
- **Mortality**: Diabetes was responsible for 6.7 million deaths in 2024, equating to 1 death every 5 seconds. (IDF Diabetes Atlas)

The Pima Indian Diabetes dataset

- It is primarily composed of medical data for the Pima Indian population in the U.S., a group with a notably high prevalence of type 2 diabetes. It was initially created by the National Institute of Diabetes and Digestive and Kidney Diseases. The subjects of the data set are female Pima Indians who are older than 21 years old.
- **Dataset Characteristics:**
  - **Number of Instances (Samples):** 768
  - **Number of Features (Columns):** 8 (plus 1 target variable)
  - **Target Variable (Outcome):** Binary (1 for diabetes-positive, 0 for diabetes-negative)
  - **Missing Values:** Some features have missing values represented by zeros (e.g., blood pressure, skin thickness).

The Pima Indian Diabetes dataset

Feature Name	Description	Range/Units
Pregnancies	Number of times pregnant	Integer
Glucose	Plasma glucose concentration (2-hour oral glucose test)	mg/dL
BloodPressure	Diastolic blood pressure	mm Hg
SkinThickness	Triceps skinfold thickness	mm
Insulin	2-hour serum insulin	mu U/ml
BMI	Body mass index (weight in kg/(height in m)^2)	kg/m^2
DiabetesPedigreeFunction	Diabetes pedigree function (genetic predisposition)	Continuous
Age	Age of the patient	Years
Outcome	Diabetes diagnosis (1: Positive, 0: Negative)	Binary

The Pima Indian Diabetes dataset

```
df.info() # structural information of the data set
```

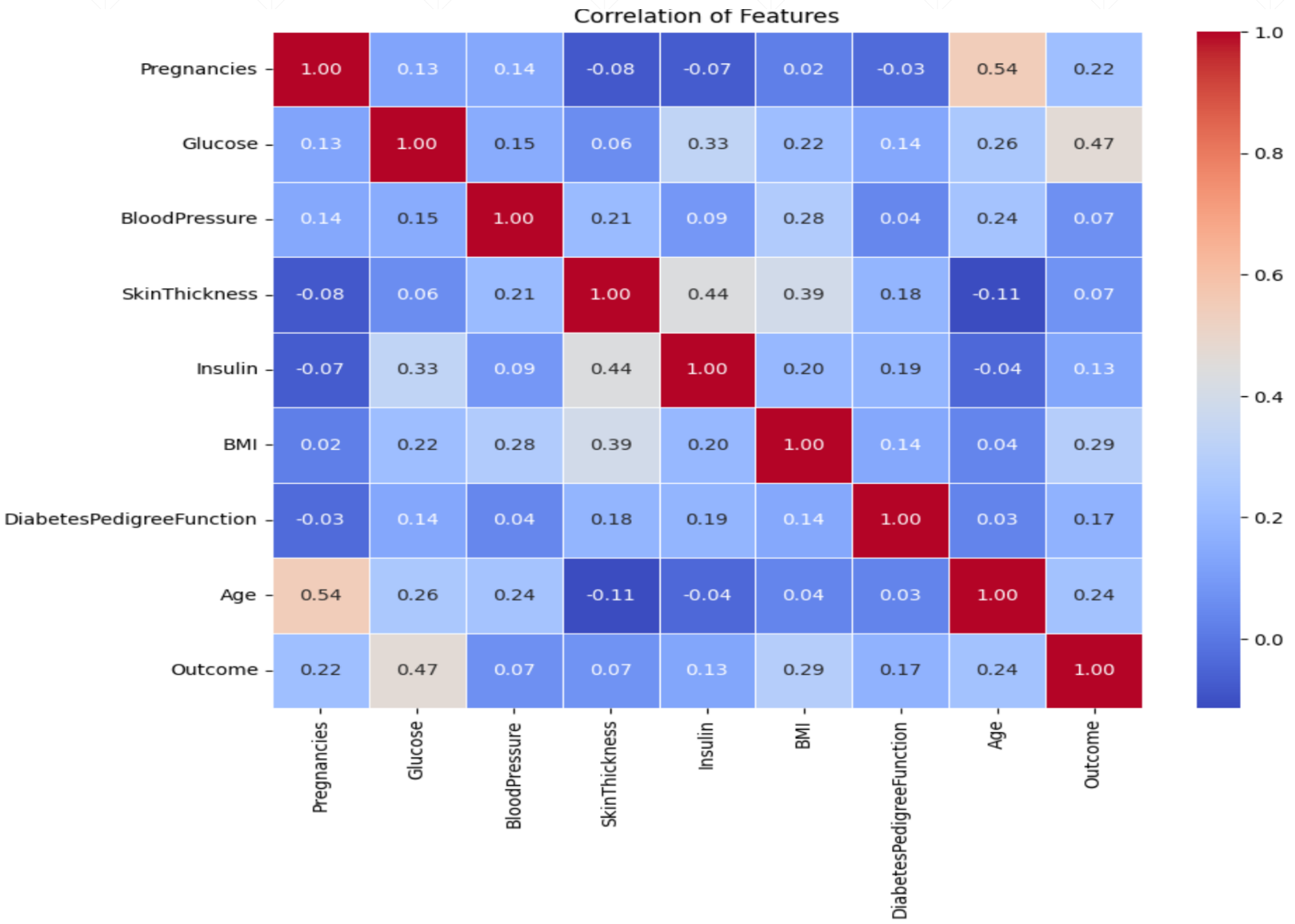
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

```
(df == 0).sum()
```

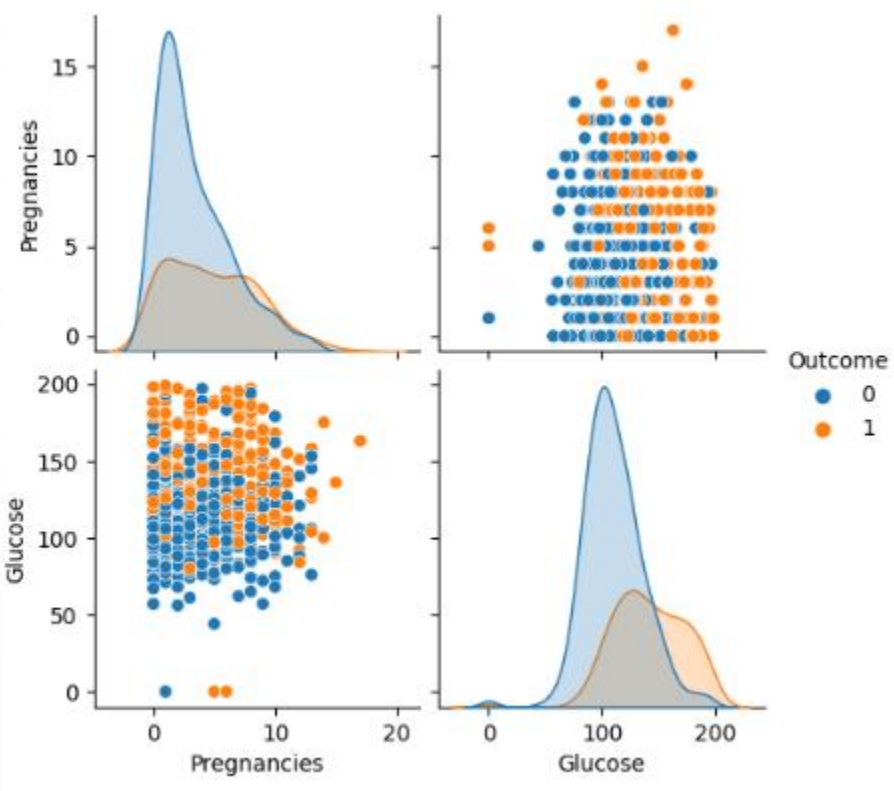
```
Pregnancies                 111
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64
```

```
df["Outcome"].value_counts()
```

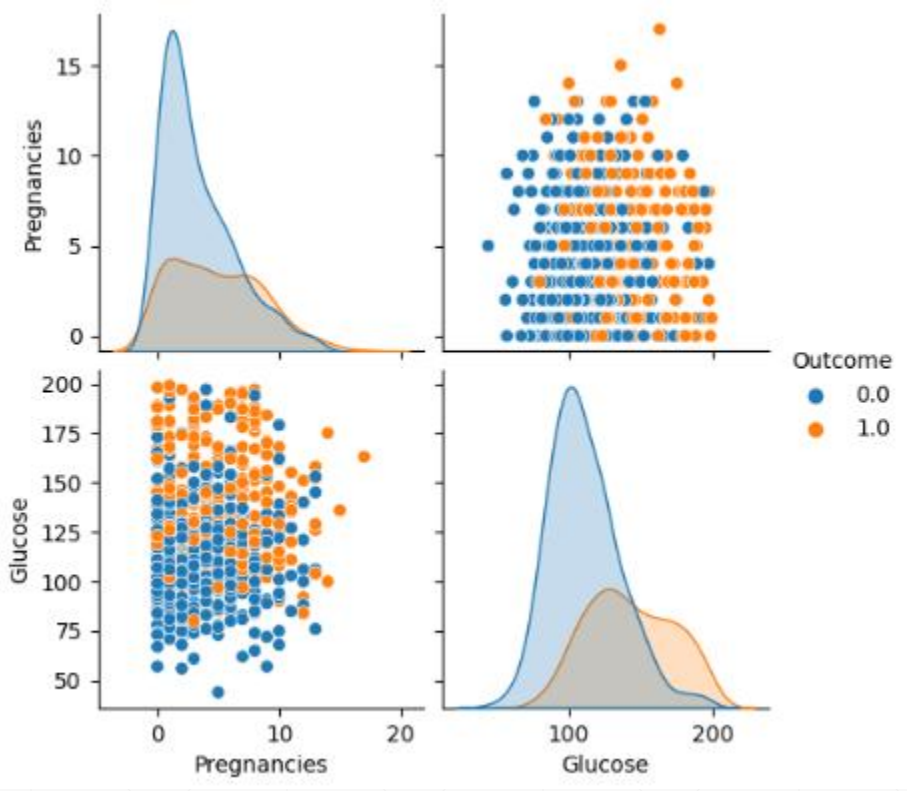
```
0    500
1    268
Name: Outcome, dtype: int64
```


Correlation of Features
Pregnancies 1.00 0.13 0.14 -0.08 -0.07 0.02 -0.03 0.54 0.22
Glucose 0.13 1.00 0.15 0.06 0.33 0.22 0.14 0.26 0.47
BloodPressure 0.14 0.15 1.00 0.21 0.09 0.28 0.04 0.24 0.07
SkinThickness -0.08 0.06 0.21 1.00 0.44 0.39 0.18 -0.11 0.07
Insulin -0.07 0.33 0.09 0.44 1.00 0.20 0.19 -0.04 0.13
BMI 0.02 0.22 0.28 0.39 0.20 1.00 0.14 0.04 0.29
DiabetesPedigreeFunction -0.03 0.14 0.04 0.18 0.19 0.14 1.00 0.03 0.17
Age 0.54 0.26 0.24 -0.11 -0.04 0.04 0.03 1.00 0.24
Outcome 0.22 0.47 0.07 0.07 0.13 0.29 0.17 0.24 1.00
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0

Imputation

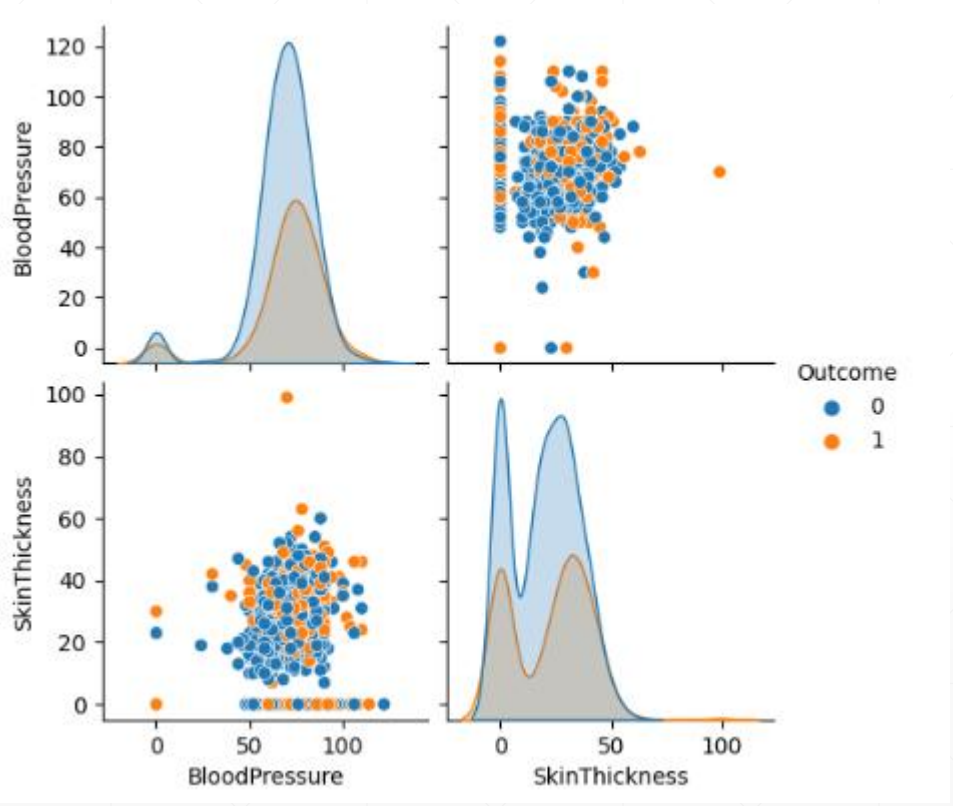


Orginal Dataset

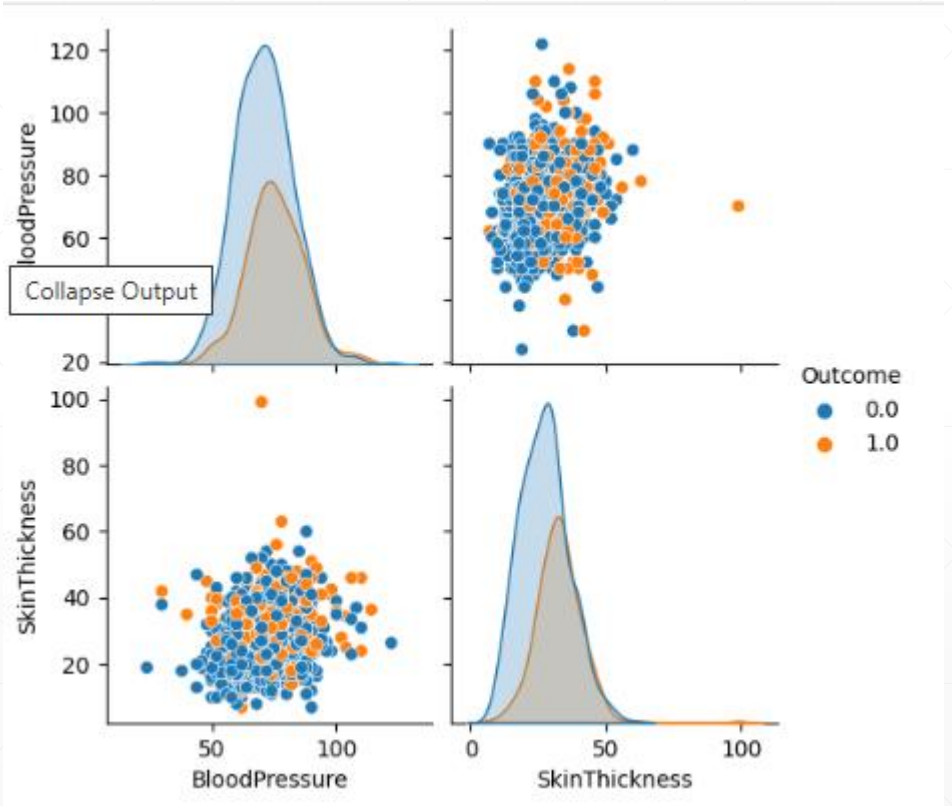


Imputed Dataset

Imputation

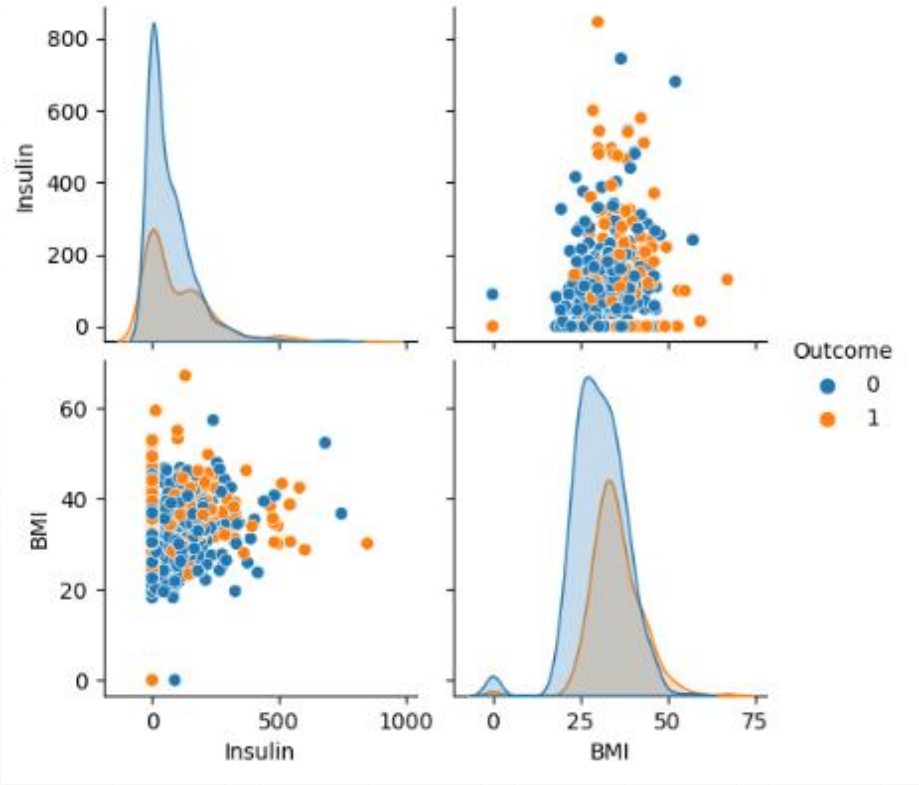


Orginal Dataset

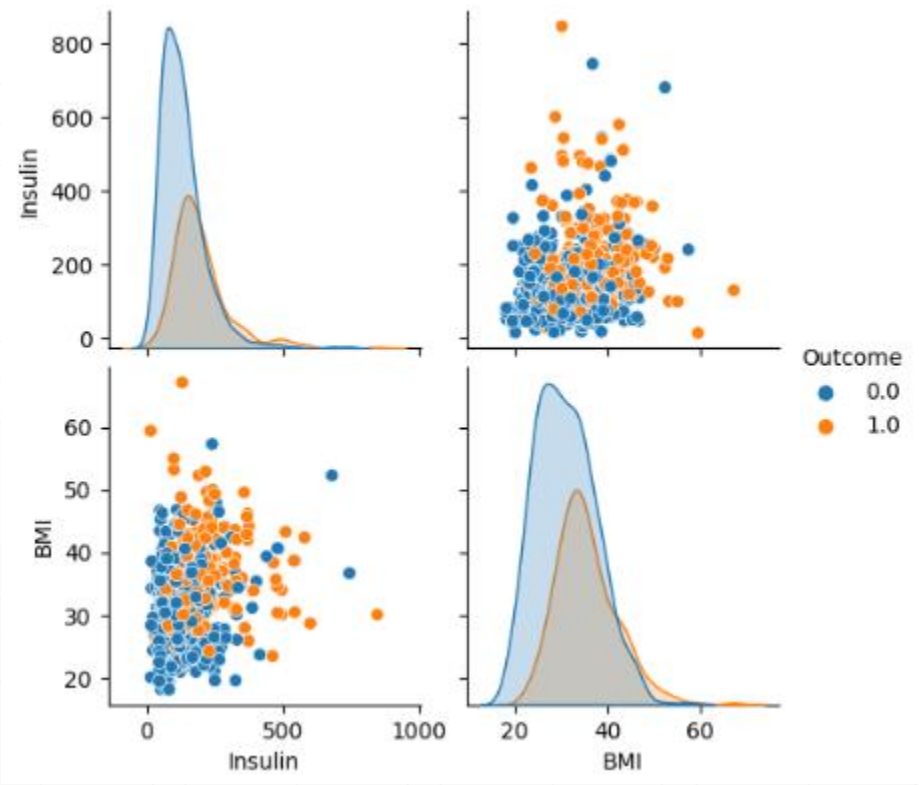


Imputed Dataset

Imputation

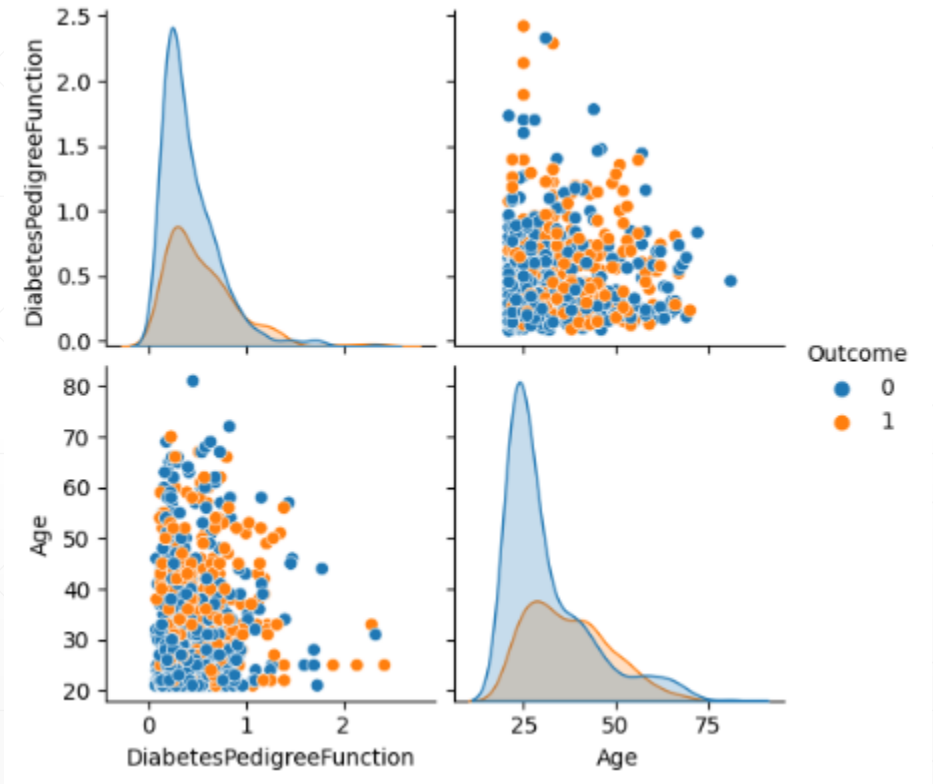


Orginal Dataset



Imputed Dataset

Imputation



Orginal Dataset

Performance of The Models