Aim:

To apply the logistic regression algorithm for binary classification.

Objectives:

1. Explore and understand the structure and characteristics of the dataset.
2. Apply data pre-processing techniques and feature selection approaches.
3. Build a logistic regression model for classification.
4. Evaluate the model's performance using appropriate metrics and Interpret the results.

Question 1: What is data and where is data?

Dataset Overview

The Pima Indians Diabetes Database includes medical data from 768 adult women, all aged 21 or older, to predict diabetes. It was collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and available at UCI repository.

Attributes and Structure

The dataset includes 8 independent variables (features) and 1 dependent variable (target). The independent variables are numerical and represent various physiological and medical measurements that are potential indicators of diabetes. The dependent variable is a binary indicator representing the presence or absence of diabetes.

Reference:

<Feature Name>:<(Data Type)><Description about feature>

1. **Pregnancies**: (Numeric) The number of times the patient has been pregnant. This variable provides information on reproductive history, which can influence metabolic health.
2. **Glucose**: (Numeric) The plasma glucose concentration a few hours after an oral glucose tolerance test (OGTT). This measure is critical as it directly relates to the body's ability to metabolize glucose.
3. **BloodPressure**: (Numeric) Diastolic blood pressure (mm Hg). Blood pressure levels can be indicative of cardiovascular health, which is often associated with metabolic syndromes.
4. **SkinThickness**: (Numeric) The thickness of the triceps skinfold (mm), a measure used to estimate body fat. Although less commonly used in modern clinical settings, this measure provides an approximation of body fat distribution.
5. **Insulin**: (Numeric) The two-hour serum insulin level (mu U/ml) after the OGTT. Insulin levels are a direct measure of pancreatic function and insulin resistance.
6. **BMI**: (Numeric) Body Mass Index, calculated as weight in kilograms divided by the square of height in meters (kg/m^2). BMI is a widely used metric for obesity, which is a known risk factor for diabetes.
7. **DiabetesPedigreeFunction**: (Numeric) A function that scores the likelihood of diabetes based on family history and genetic predisposition. This factor combines genetic and familial factors.
8. **Age**: (Numeric) The age of the patient in years. Age is an important factor as the risk of diabetes increases with age.
9. **Outcome**: (Binary) The target variable, where 0 indicates non-diabetic and 1 indicates diabetic.

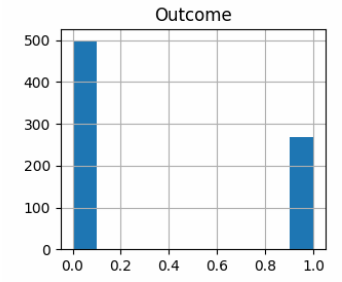
Target Column: Outcome with 2 classes **Yes/No.**

Question 2: Is DATA good ?

Since from the Dataset provided it can be said that the DATA is good and need not to be further processed as there are no null values.

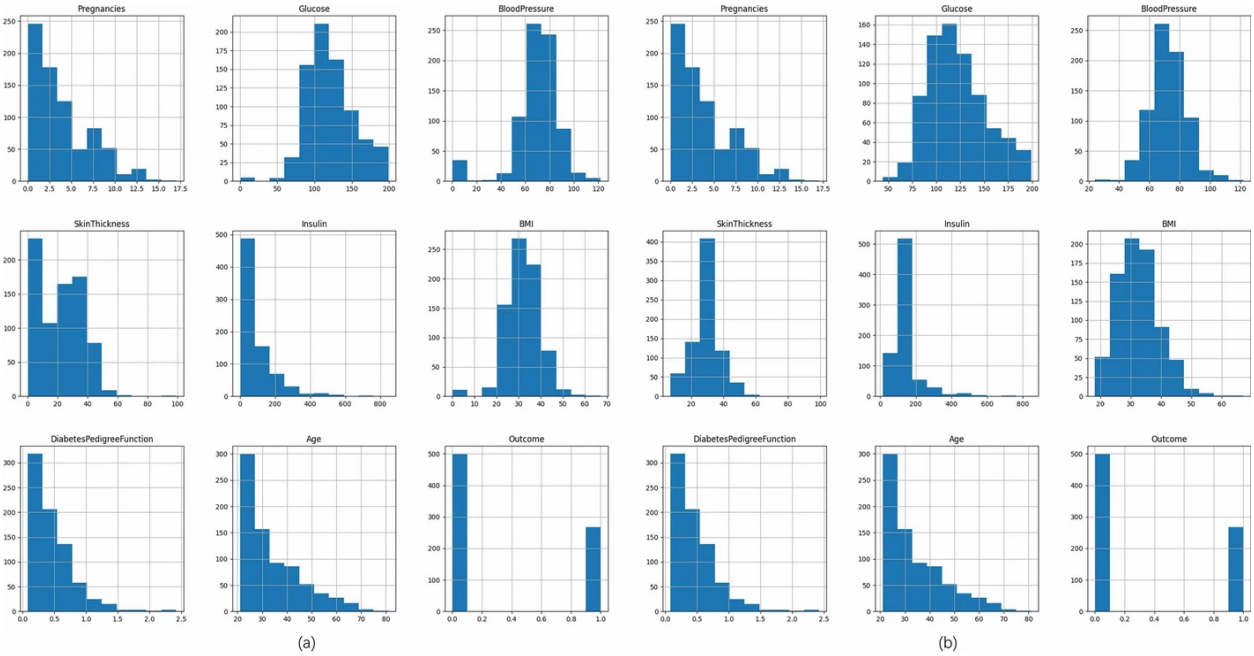
```
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             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
```

	0
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

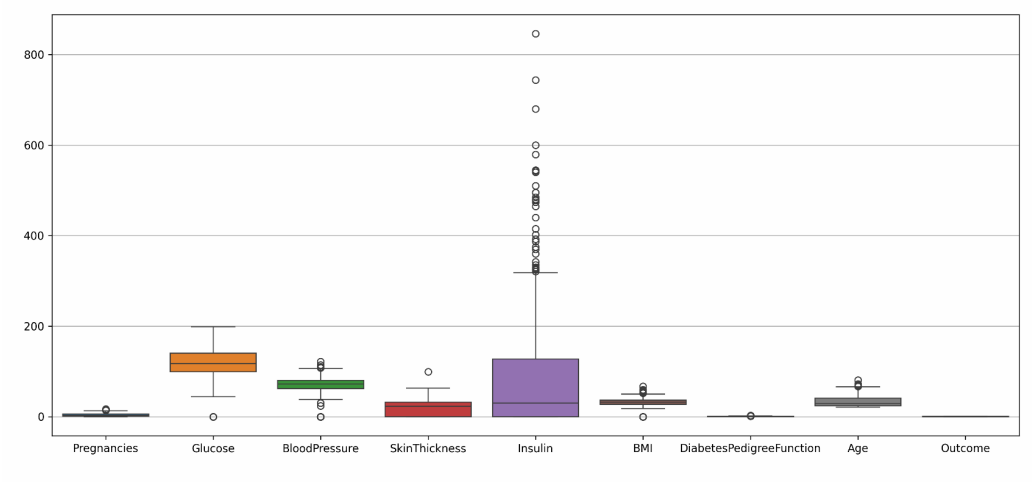


Class imbalance can be seen

- ***Outcome 0 or Non-diabetic = 500***
- ***Outcome 1 or Diabetic = 268***



Choosing the appropriate data source is important as (a) is the data before missing value imputation, (b) is the data after value imputation (According to Research Paper.) and this adhere with the performance of the algorithm

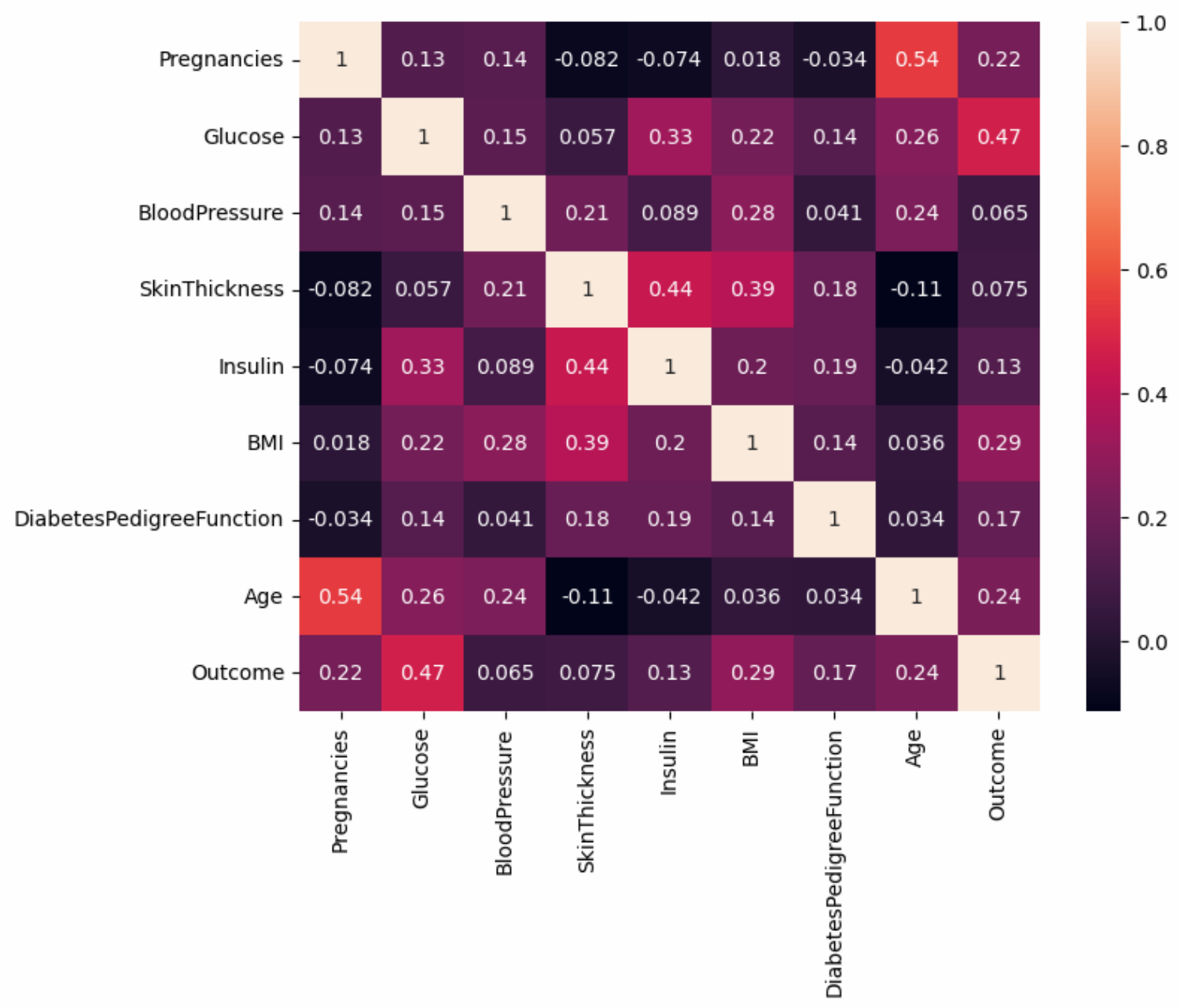


As per the boxplot it can be said that Attribute (Insulin) has a large number of outliers with high values so it should be avoided in the feature selection process.

Understanding Correlation Coefficients

- **+1** indicates a perfect positive linear relationship.
- **-1** indicates a perfect negative linear relationship.
- **0** indicates no linear relationship.

Interpreting the Correlation Matrix



After analyzing the correlation matrix following are the results:

1. **Strong Predictors**:
    - **Glucose**: Shows a high positive correlation with Outcome, indicating it is a significant predictor of diabetes.
    - **BMI**: Has a moderate positive correlation with Outcome, suggesting that individuals with higher BMI are at a greater risk of developing diabetes.
    - **Age**: Demonstrates a weaker but positive correlation with Outcome, implying that older individuals have a slightly higher risk.

2. **Weak Predictors**:
    - **Blood Pressure**: Lower correlation with Outcome, indicating it is not at all significant feature for prediction of outcome
    - **Skin Thickness:** Lower correlation with Outcome, suggesting that skin thickness has nothing to do with Diabetes presence.

Also from the correlation we can observe that though Insulin is said to be one of the crucial parameters for diabetes prediction but due to the large number of outliers even Pregnancies show higher correlation with Outcome than Insulin.

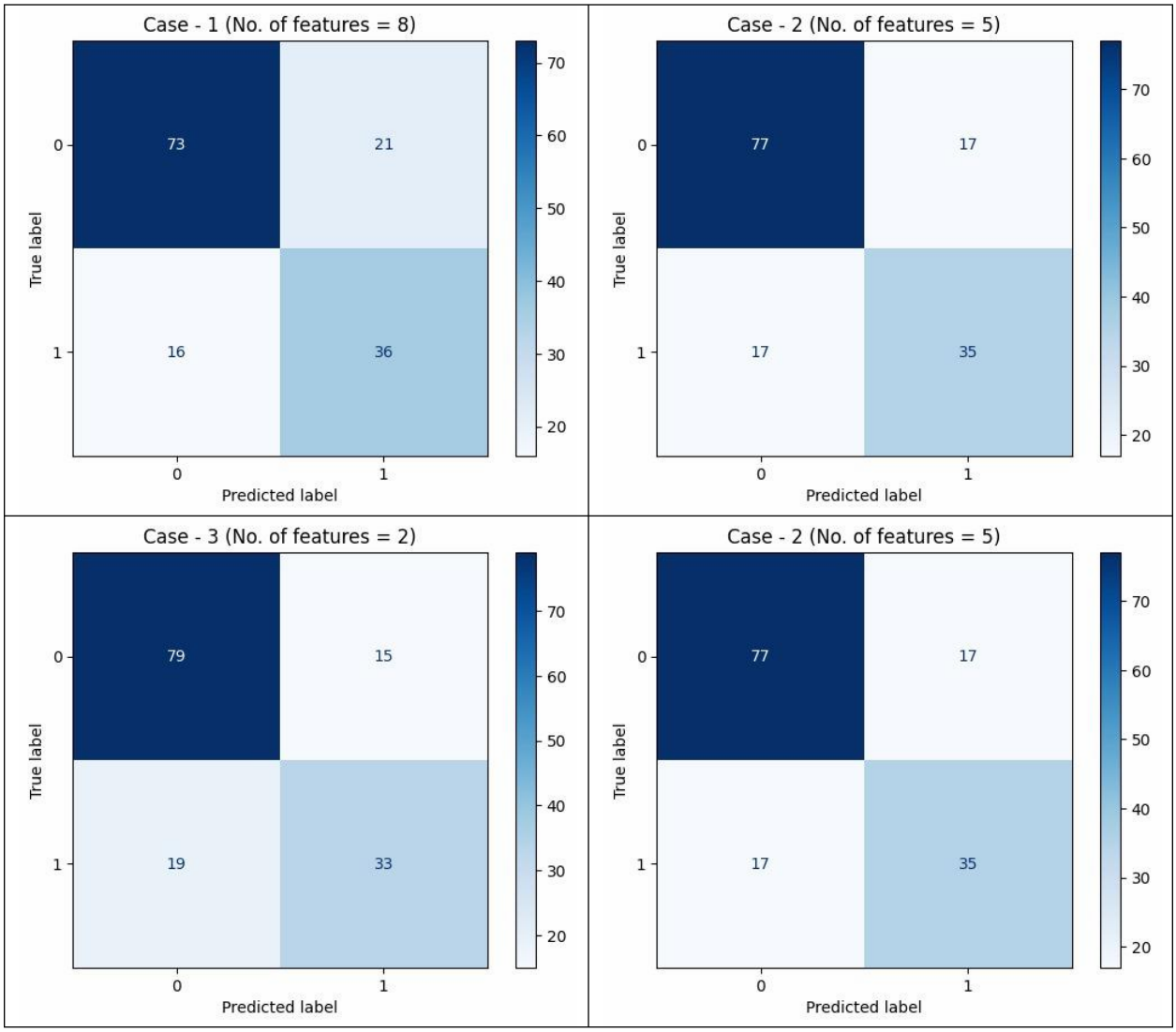
Feature Selection Approaches

Feature selection is a crucial step in the development of predictive models, particularly in the context of the Pima Indians Diabetes Dataset. The goal of feature selection is to identify the most relevant variables that contribute to the prediction of the target variable (Outcome). This process can enhance model performance, reduce overfitting, and improve interpretability. Below, we detail four distinct cases of feature selection:

Table 1: Feature Selection Table

Case No.	Selected Features	Reasons
1	Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age	Utilizes all available data, maximizes the information available for prediction
2	Glucose, Insulin, BMI, Diabetes Pedigree Function, Age	Suggested features by medical professionals for scientific and clinical relevance as per the literature survey (by referring existing research papers)
3	Glucose, BMI	Highly correlated features according to correlation matrix
4	Glucose	Single most significant predictor of diabetes and highest correlation

Confusion Matrices



Performance Metrics

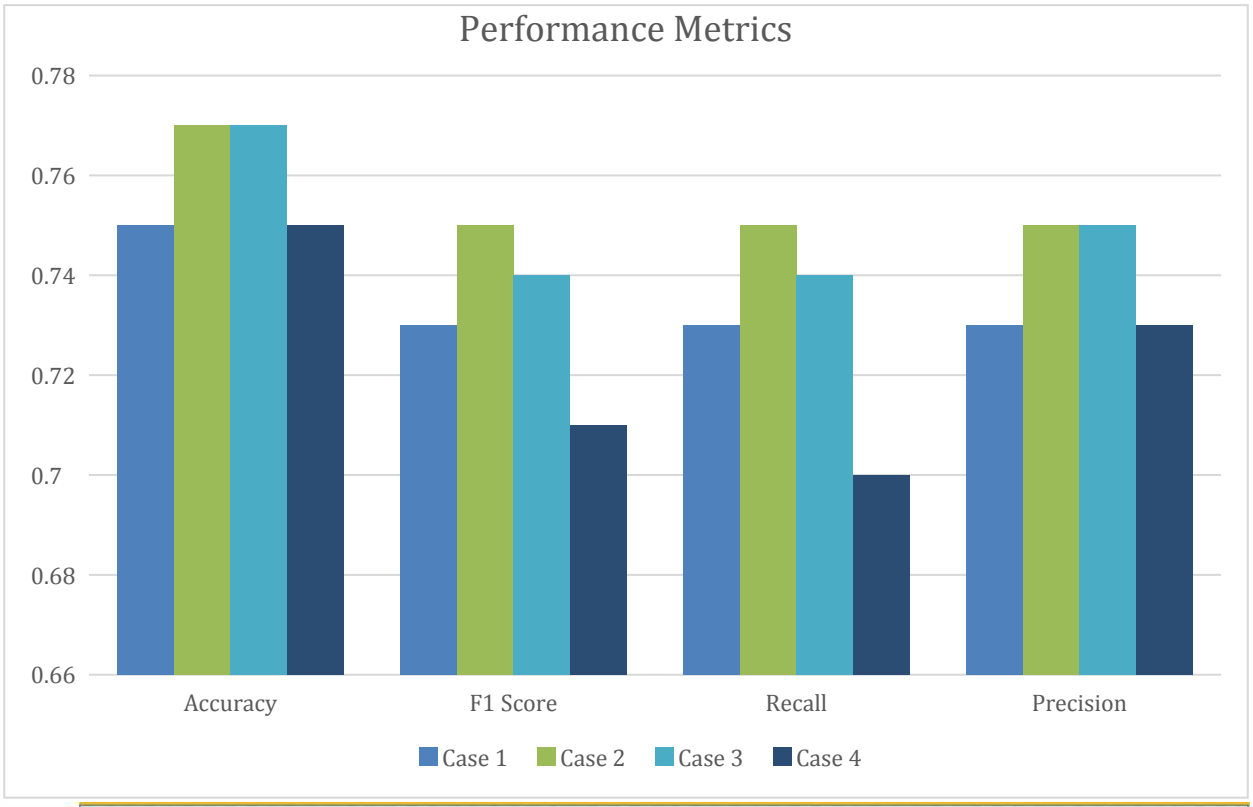
In this analysis, we evaluate the performance of a logistic regression model applied to the Pima Indians Diabetes Dataset across four different feature selection approaches. We assess each case using standard performance metrics: **accuracy**, **precision**, **recall**, **F1-score**, and **training time**. These metrics provide a comprehensive understanding of the model's effectiveness and efficiency.

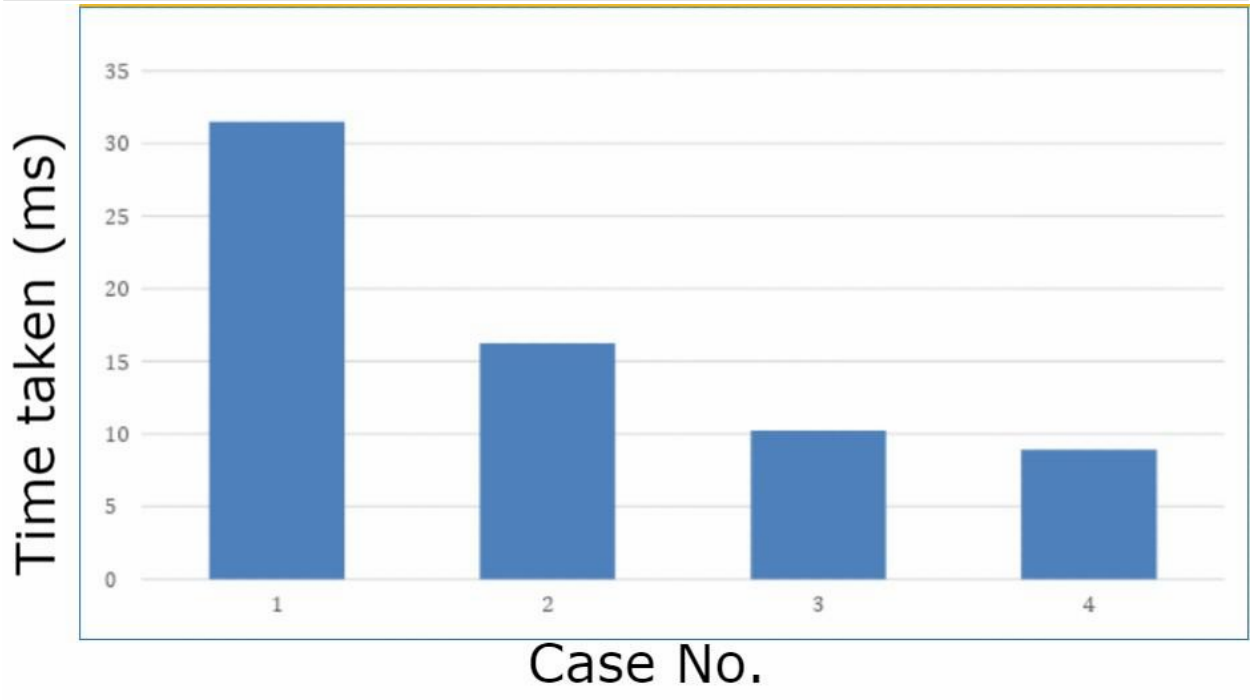
Performance Metrics Overview

1. **Accuracy**: The proportion of correctly classified instances (both true positives and true negatives) out of the total instances.
2. **Precision**: The ratio of true positives to the sum of true positives and false positives, indicating the model's accuracy in identifying positive cases.
3. **Recall (Sensitivity)**: The ratio of true positives to the sum of true positives and false negatives, measuring the model's ability to identify all actual positive cases.
4. **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two, especially useful when dealing with imbalanced datasets.
5. **Training Time**: The duration required to train the model, reflecting computational efficiency.

Table 2: Performance Table of various cases by feature selection

Case No.	Accuracy	F1 Score	Recall	Precision	Time (in ms)
1	75%	73%	73%	73%	31.48
2	77%	75%	75%	75%	16.22
3	77%	74%	74%	75%	10.24
4	75%	71%	70%	73%	8.89


Performance Metrics
0.78
0.76
0.74
0.72
0.7
0.68
0.66
Accuracy
F1 Score
Recall
Precision
Case 1
Case 2
Case 3
Case 4


Time taken (ms)
35
30
25
20
15
10
5
0
1
2
3
4
Case No.

Interpretation of results:

- Looking at the performance matrix it can be said that Case 2 and Case 3 outperforms other cases.
- Each evaluation measure of case 2 (5 features) is high compared to other cases where time taken is almost 50% less compared to 1st case (all 8 features).
- Looking at the time graph it can be seen that Time taken is directly proportional to No. of features selected as it requires less computational power.
  (Note: This proportionality equation differs if "max_iterations" is defined in logistic regression as no of iterations in cases with more features might exceed max iterations)
- While the overall performance rating of Case 2 is 75.5% and that of Case 4 is 72.25% but time taken by Case 4 is 50% less than that of Case 2. When there is a need to deploy a model on resource constrained hardware, we can use case 4 for computation efficiency by little compromising accuracy.
- Case 2 demonstrates superior results because it focuses on features that more effectively separate the diabetic and non-diabetic classes. The features excluded (Blood Pressure, Skin Thickness, and Pregnancies) contribute less to class discrimination, leading to overlapping data points that do not clearly distinguish between the two outcomes. By concentrating on features with better discriminatory power such as Glucose, Insulin, BMI, Diabetes Pedigree Function, and Age. Case 2 enhances the model's ability to differentiate between classes, resulting in improved accuracy and overall performance.
- By looking at case 1 and case 4, accuracy is same - it indicates that only 1 feature of case 4 can do better prediction compared to case 1(all features) - 8 features. As we can see from the above box plot of glucose vs Outcome, we can see that values of glucose below 110 always results in no diabetes and values above around 125 always result in diabetes and almost all the data points of glucose are concentrated around this range, hence it alone is a very good predictor of diabetes.