

**Machine Learning (ML) Assignment Report**

**Created By:**

1. **Ansh Shah (22BCM007)**
2. **Aryana Ghughrawala (22BCM009)**

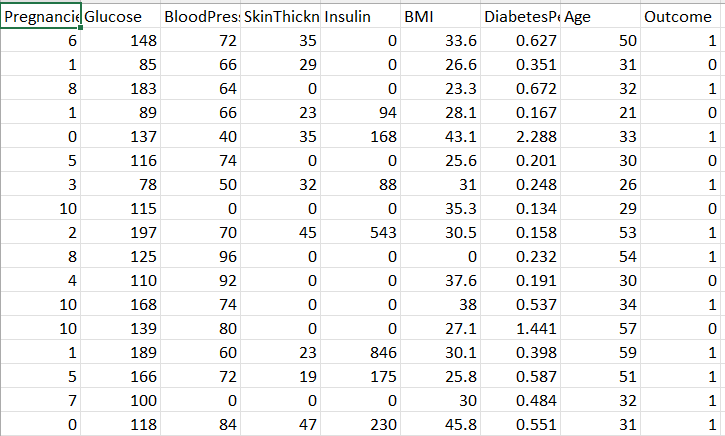
**Div : A**

**Batch: CSE – MBA**

**A] Dataset Description:**

**This dataset encompasses information pertaining to 769 women and includes a diverse range of health-related attributes. Please refer to the following table for a concise overview of the columns:**

1. **Pregnancies: The number of times the woman has been pregnant.**
2. **Glucose: Glucose concentration in the woman's plasma.**
3. **Blood Pressure: Blood pressure measurement.**
4. **Skin Thickness: Thickness of skinfold at the triceps.**
5. **Insulin: Insulin levels in the blood.**
6. **BMI (Body Mass Index): A measure of body fat based on height and weight.**
7. **Diabetes Pedigree Function: A function that scores the likelihood of diabetes based on family history.**
8. **Age: Age of the woman.**
9. **Outcome: The target variable indicating whether the woman has diabetes or not (1 for diabetic, 0 for non-diabetic).**

****

**The provided image represents a limited portion of the comprehensive dataset.**

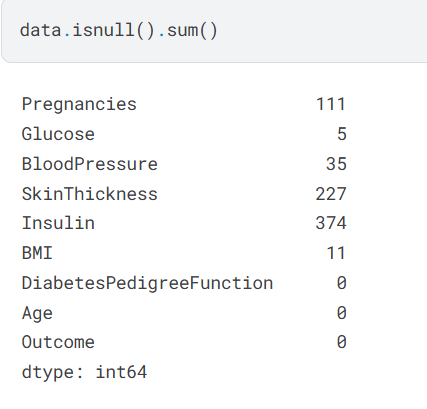
**Source:https://www.kaggle.com/datasets/pentakrishnakishore/diabetes-csv/data**

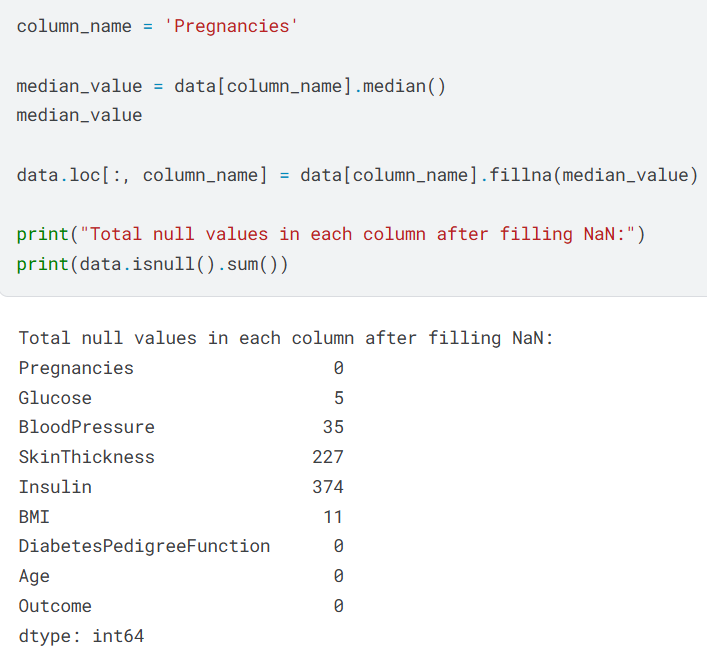
**B] Methodology Followed:**

1. ** Data Loading and Exploration:**

**The preceding lines are employed to describe the data and subsequently display a portion of it.**

1. **Data Cleaning :**

* **This method primarily addresses the handling of missing values within the data.**
* **From the provided image below, it is evident that which column in the dataset contains the number of missing values.**
* **So, if we proper data cleaning is not done, then the results obtained will not be accurate.**

****

* **Now, as seen we have carried out a very popular method called as “Handling missing values by mean”**
* **In this method, the mean of the entire dataset is calculated, and subsequently, all missing values are replaced with this mean.**
* **The preceding image serves as a simplified illustration, but in reality, this process is applied to all columns within the dataset that contain missing values.**

1. **Model Training:**

**In this section, we will present the models implemented by us and utilized in the paper.**

1. **Support Vector Classifier (SVC) :**

**The Support Vector Classifier (SVC) is a machine learning model employed for classification tasks. It belongs to the Support Vector Machine (SVM) family. In this code, the SVC model is trained and evaluated on a dataset to classify samples into distinct classes based on the patterns learned from the data.**

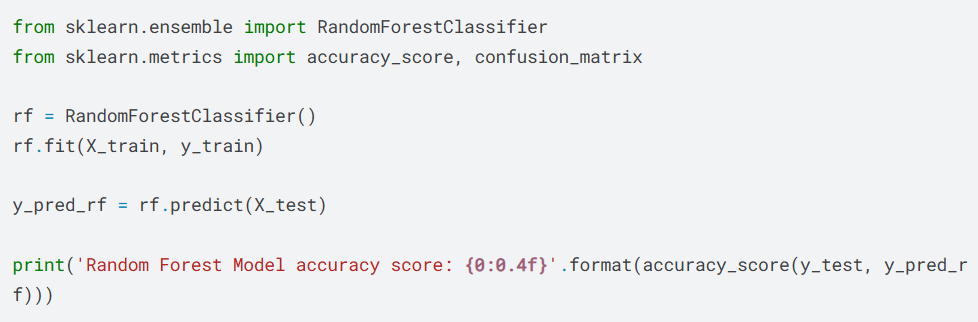
**Here's the code supporting this :**

****

**In conclusion, accuracy is computed as the ratio of the number of correct predictions to the total number of predictions made. It represents the percentage of instances that the model correctly classified out of all instances.**

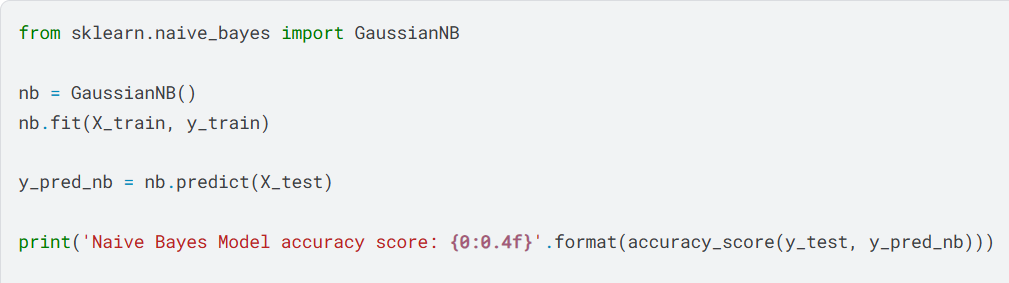
1. **Random Forest:**

**Random Forest is an ensemble learning method employed for both classification and regression tasks. During training, it constructs a multitude of decision trees and outputs the mode of their predictions (for classification) or the mean prediction (for regression). The method merges the predictions from multiple trees to enhance the overall accuracy and robustness of the model.**

****

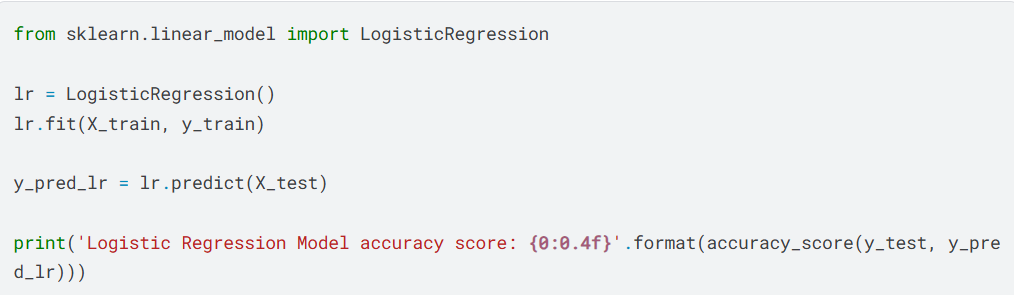
1. **Naïve Bayes:**

**Naive Bayes is a family of probabilistic algorithms that utilize Bayes’ theorem under the assumption of feature independence. It is commonly employed for classification tasks and exhibits exceptional efficacy in handling large datasets. The term “naive” stems from the assumption that the occurrence of a specific feature within a class is independent of any other feature, thereby simplifying the computational process.**

****

1. **Logistic Regression:**

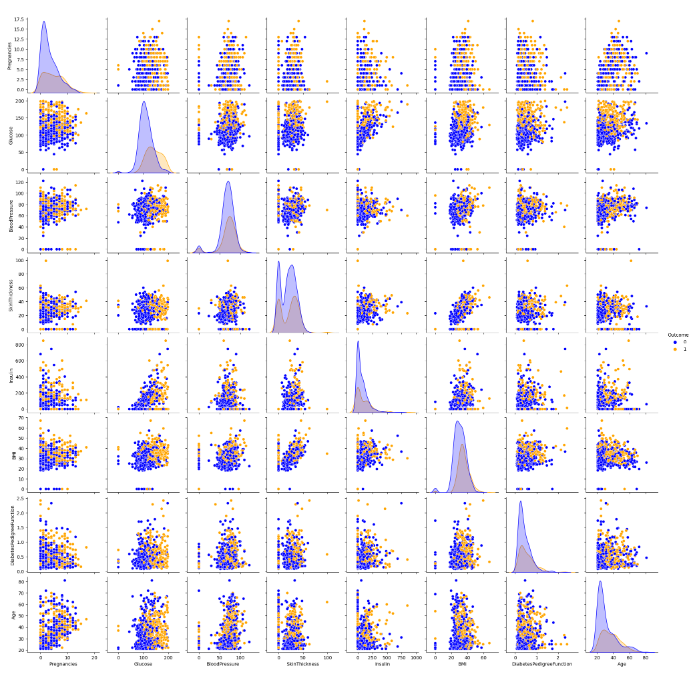
**Logistic Regression is a statistical method used for binary classification problems, where the outcome or dependent variable is categorical (often coded as 0 or 1). It predicts the probability that a given input point belongs to a particular category based on one or more independent variables (features).**

****

**D] Results:**

1. **Pairplot:**

* **A pairplot, a visualization tool in data analysis, especially useful in exploratory data analysis, creates a grid of scatterplots for each pair of variables in a dataset. It often displays the density spread of each variable along the diagonal.**



**The image’s blue and orange colors suggest a binary classification dataset with two classes.** **Some scatter plots show distinct clusters, indicating natural groupings in the data.** **Variables with good separation between classes might be crucial for classification.**

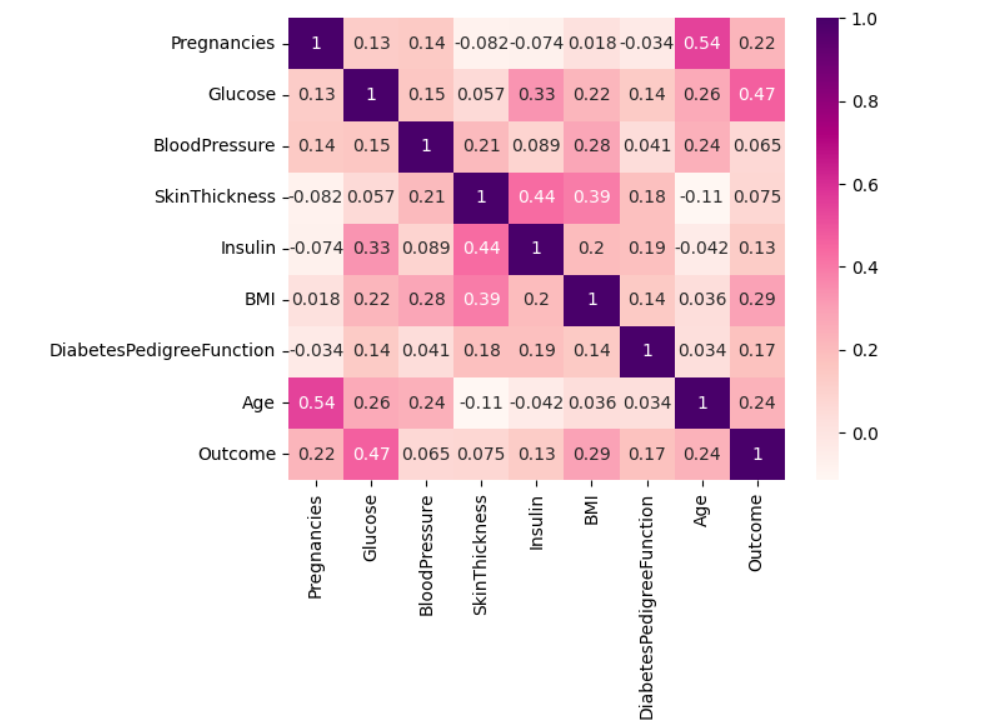
1. **Heatmap:**

* **A heatmap is a data visualization technique that employs color gradients to represent the magnitude of a phenomenon within two dimensions.**
* **Heatmaps utilize color scales to indicate the values of a matrix. Darker or brighter colors typically denote higher values, while lighter colors represent lower values, facilitating the identification of trends and concentrations of data.**
* **Frequently employed to visualize correlation matrices, heatmaps illustrate the relationships between various features. Values range from -1 to 1, where:**

**-1 signifies a perfect negative correlation,**

**1 indicates a perfect positive correlation, and**

**0 denotes no correlation.**



**Strongest Positive Correlations:**

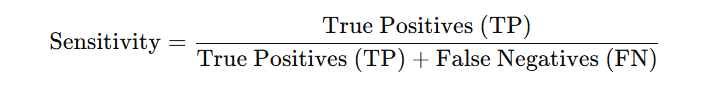
* **Pregnancies & Age (0.54): Older women tend to have more pregnancies.**
* **Glucose & Outcome (0.47): Higher glucose levels strongly indicate diabetes.**

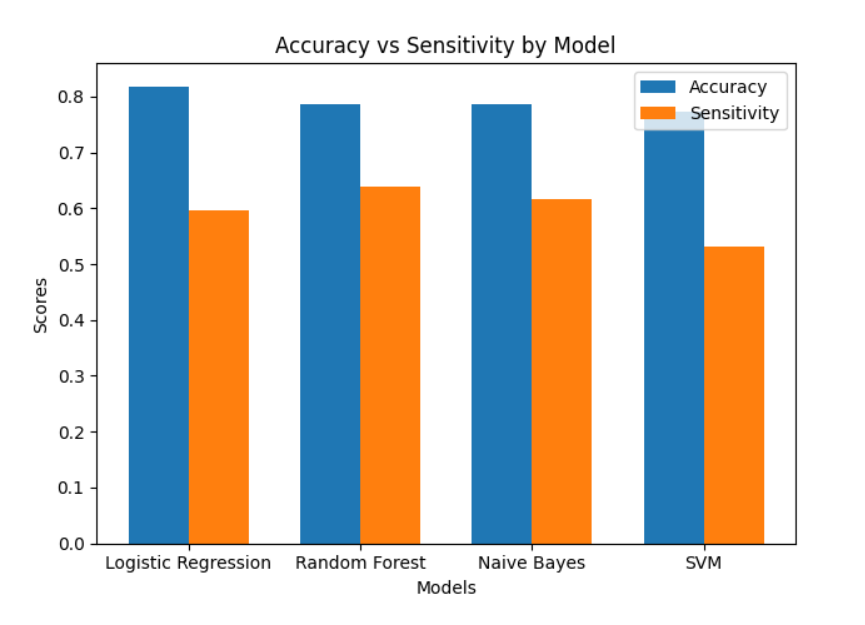
1. **Accuracy vs Sensitivity Comparison:**

**Accuracy : Accuracy is the ratio of the number of correct predictions to the total number of predictions made. It tells us the percentage of instances that the model classified correctly out of all instances.**

**Sensitivity: Sensitivity, also known as True Positive Rate or Recall, indicates the proportion of actual positive cases that are correctly identified by the model.**

**Sensitivity can be calculated using the following formula:**

****

****

1. **Accuracy and Sensitivity of all models:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** |
| **SVC** | **77.27%** | **53.19** |
| **Random Forest** | **79.87%** | **61.70** |
| **Naïve Bayes** | **78.57%** | **61.70** |
| **Logistic Regression** | **83.12%** | **59.57** |