NAME : *BASMA NAGHMAN*

ROLL\_NO : *BSAIM – 048 – 3A*

**REPORT TITLE : Diabetes Prediction**

Contents

[OVERVIEW: 1](#_Toc184112556)

[Code Cells: 1](#_Toc184112557)

[Library Imports: 1](#_Toc184112558)

[Dataset Loading: 1](#_Toc184112559)

[Preprocessing Steps: 2](#_Toc184112560)

[Handling Missing Values : 2](#_Toc184112561)

[Data Encoding: 2](#_Toc184112562)

[Dataset Cleaning: 2](#_Toc184112563)

[Modeling Steps: 2](#_Toc184112564)

[Models trained include: 2](#_Toc184112565)

[Valuation Steps: 2](#_Toc184112566)

[Accuracy 2](#_Toc184112567)

[Prediction Example: 2](#_Toc184112568)

[Model Export: 3](#_Toc184112569)

# OVERVIEW:

To create a report for your file, I'll first need to examine its contents. I'll load the notebook and analyze its structure, code, and outputs. Let me do that .The notebook appears to be a project focused on diabetes prediction, likely using machine learning techniques.

## Code Cells:

The first code cell imports necessary libraries like pandas, numpy, matplotlib, seaborn , scikit-learn modules, and more.

Subsequent cells involve loading the dataset and displaying it.I will extract more detailed insights from the notebook, including preprocessing steps, model training, evaluation, and visualizations. Let me continue reviewing the contents.

# Library Imports:

Includes essential Python libraries for data analysis, visualization, and machine learning, such as pandas, numpy, seaborn, and scikit-learn classifiers (e.g., LogisticRegression, DecisionTreeClassifier, etc.)

# Dataset Loading:

- A dataset named "medical\_students\_dataset - Copy.csv" is loaded into a DataFrame.

- Dataset exploration includes:

-Displaying the dataset (df)

- Checking the value counts for a column named Diabetes

- Inspecting the dataset shape, column names, and null values

# Preprocessing Steps:

## Handling Missing Values :

- Numerical columns are filled with their mean values.

- Categorical columns are filled with their mode values.

## Data Encoding:

- Categorical columns (Gender, Blood Type, Diabetes, Smoking) are converted to integers using LabelEncoder.

## Dataset Cleaning:

- Rows with missing values in critical columns (Diabetes) are dropped.

- Features (X) and the target (y) are separated. Unnecessary columns like Student

ID are dropped**.**

# Modeling Steps:

## Models trained include:

Logistic Regression

Random Forest Classifier

Gradient Boosting Classifier

A sampling method (RandomUnderSampler) is used to handle class imbalance in the dataset.

# Valuation Steps:

Accuracy**:**

- The accuracy score is calculated on test predictions.

## Prediction Example:

- A single input sample is reshaped and passed to the model for prediction.

- The result is interpreted as "diabetes present" or "not present."

## Model Export:

- The trained model is serialized and saved using pickle for future use.