**Name: Samar Wagih Elshedody**

**ID: 211002021**

**Diabetes Prediction Report**

[**GitHub Link**](https://github.com/samarwagih/Diabetes-Prediction-Using-ML/tree/main/Diabetes%20Prediction)

[**Presentation Link**](https://nileuniversity-my.sharepoint.com/:v:/r/personal/s_wagih2121_nu_edu_eg/Documents/Recordings/Meeting%20with%20Samar%20Wagih%20Elshedody-20240514_172151-Meeting%20Recording.mp4?csf=1&web=1&e=hpDtIc&nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJTdHJlYW1XZWJBcHAiLCJyZWZlcnJhbFZpZXciOiJTaGFyZURpYWxvZy1MaW5rIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXcifX0%3D)

[**App-Streamlit Link**](http://localhost:8501/)

**1 | Introduction**

**1.1 | About dataset**

The Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features pregnancies, glucose level, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, and age. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information. This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

**1.2 | Problem statement**

The objective of this dataset is to build a predictive model for diagnosing the model should predict whether a patient has diabetes (Outcome = 1) or does not have diabetes (Outcome = 0) based on several diagnostic measurements, including pregnancies, glucose level, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, and age.

**1.3 | Data description**

| **No** | **Column Name** | **Meaning** |
| --- | --- | --- |
| **1** | **Pregnancies** | Number of pregnancies |
| **2** | **Glucose** | Glucose level in blood |
| **3** | **Blood Pressure** | Blood pressure measurement |
| **4** | **Skin Thickness** | Thickness of the skin |
| **5** | **Insulin** | Insulin level in blood |
| **6** | **BMI** | Body mass index |
| **7** | **DiabetesPedigreeFunction** | Diabetes percentage |
| **8** | **Age** | Age |
| **9** | **Outcome** | Result (1: Yes, the individual has diabetes; 0: No, the individual does not have diabetes) |

**2 | Conclusion**

I have done the "Diabetes Prediction" task. Originally, the dataset had 769 records. Besides, creating 7 different models and tuning their parameters were very useful for evaluating their performance to decide which one is the most effective for predicting diabetes disease. As mentioned above, because of the unique characteristics of the medical industry, correctly predicting diseases such as diabetes is crucial so choosing which model provided the highest value of Recall should be on the top of priority, also in the feature engineering i create new columns "NewBMI\_Obesity1','NewBMI\_Obesity2','NewBMI\_Obesity'NewBMI\_Overweight','NewBMI\_Underweight','NewInsulinScore\_Normal','NewGlucose\_Low','NewGlucose\_Normal''NewGlucose\_Overweight', 'NewGlucose\_Secret'], enhance the predictive power of the dataset. Finally, as shown in the model score the Gradient Boosting Classifier provided the highest score (91.45%).

A screenshot of a computer

Description automatically generated

**2.2 | model performance**

**A graph of a performance evaluation

Description automatically generated with medium confidence**

**2.2 | Model Deployment:**

**A screenshot of a computer

Description automatically generated**

**2.3| Future Possible Work:**

- Betterment of results using different hyperparameters for tuning

- Implementing more models to gain better results

- Using the same method to predict response