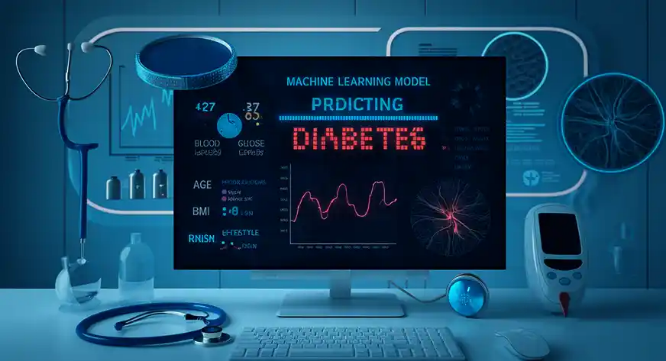
**Diabetes Prediction Project**

**Overview**

This project aims to develop a machine learning model for predicting the likelihood of diseases, such as diabetes, using patient data. By leveraging advanced machine learning algorithms, the project focuses on analyzing key health indicators, such as age, blood pressure, cholesterol levels, and other medical features. The goal is to build an effective prediction system that can assist healthcare professionals in early diagnosis and risk assessment.

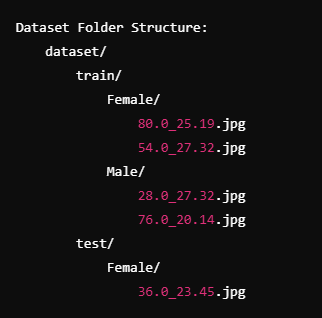
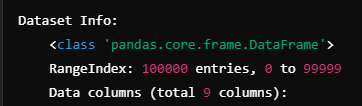
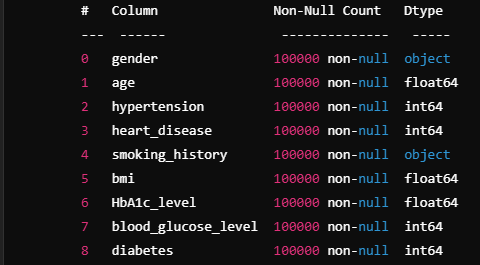
The process begins with data preprocessing, including handling missing values, normalization, and encoding categorical features. Exploratory Data Analysis (EDA) is conducted to understand feature distributions and their correlations. Various machine learning models, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), are trained and evaluated on the dataset. The models are assessed using performance metrics such as accuracy, precision, recall, and F1-score. The final objective is to compare these models, identify the most reliable one, and generate insights to inform decision-making in healthcare.



**Project Objectives**

1. **Train Predictive Models:**  
   Develop and train machine learning models to predict the likelihood of diseases such as diabetes, based on patient health data.
2. **Data Preprocessing:**  
   Handle missing data, normalize numerical features, and encode categorical variables to prepare the dataset for modeling.
3. **Exploratory Data Analysis (EDA):**  
   Analyze the distribution of features and identify correlations between them to better understand the dataset.
4. **Model Training:**  
   Implement and train various machine learning algorithms, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), to predict disease outcomes.
5. **Model Evaluation:**  
   Evaluate the performance of each model using metrics such as accuracy, precision, recall, and F1-score to assess prediction quality.
6. **Comparison of Models:**  
   Compare the performance of the models to identify the best-performing algorithm for disease prediction.
7. **Insight Generation:**  
   Provide actionable insights from the analysis and evaluation, supporting better disease risk assessment and decision-making.

**Dataset:**

**** **** 

**Libraries**

**What libraries that we used in this project actually.**

* Pandas Ipython
* Numpy
* Sklearn
* Matplotlib
* Seaborn

**Model Training**

We use some models and train them to predict the disease and find the accuracy.

1. **Logistic Regression**

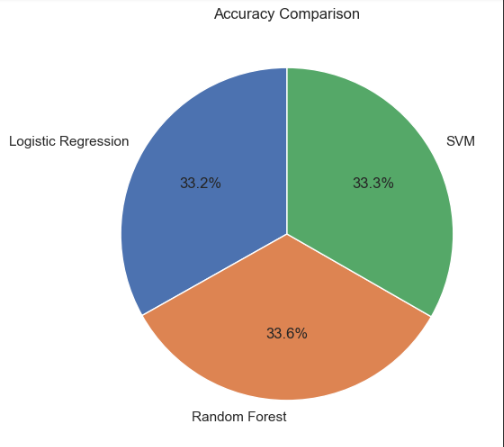
* F1-Score : 0.7167
* Accuracy : 0.9587
* Precision : 0.8637
* Recall : 0.6124

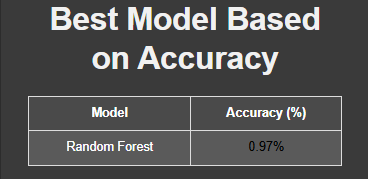
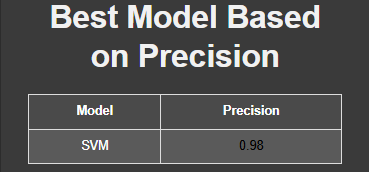
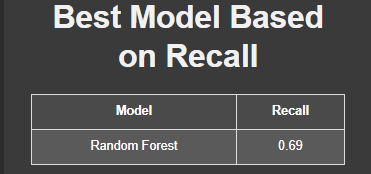
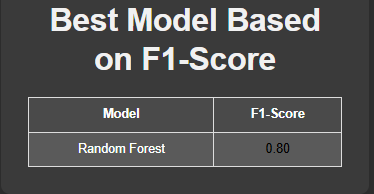
1. **Random Forest**

* F1-Score : 0.8009
* Accuracy : 0.9707
* Precision : 0.9517
* Recall : 0.6915

1. **SVM**

* F1-Score : 0.7222
* Accuracy : 0.9624
* Precision : 0.9761
* Recall : 0.5732



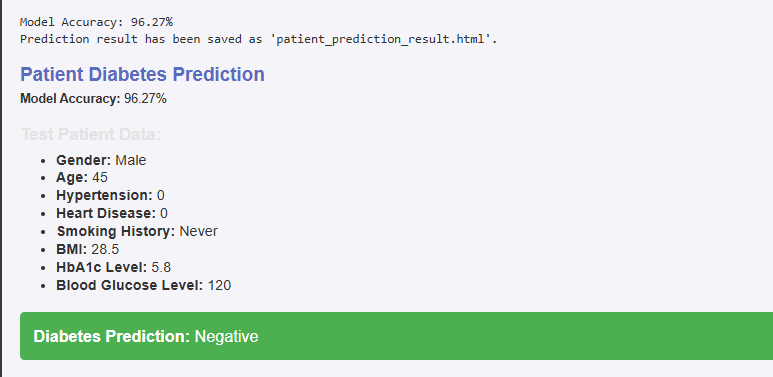
**Environment:**

Jupyter Lab (4.3.4)

From Anacona (24.11.3)

Python (3.12.3)

**Testing:**

****

**Project Access**

**Google colab link :** <https://colab.research.google.com/drive/1qJS4uv4gam7FSmQPSvoz9n99_Wtewfyh?usp=sharing>