DIABETES PREDICTION USING MACHINE LEARNING

Kokoh Dwiko Listanto

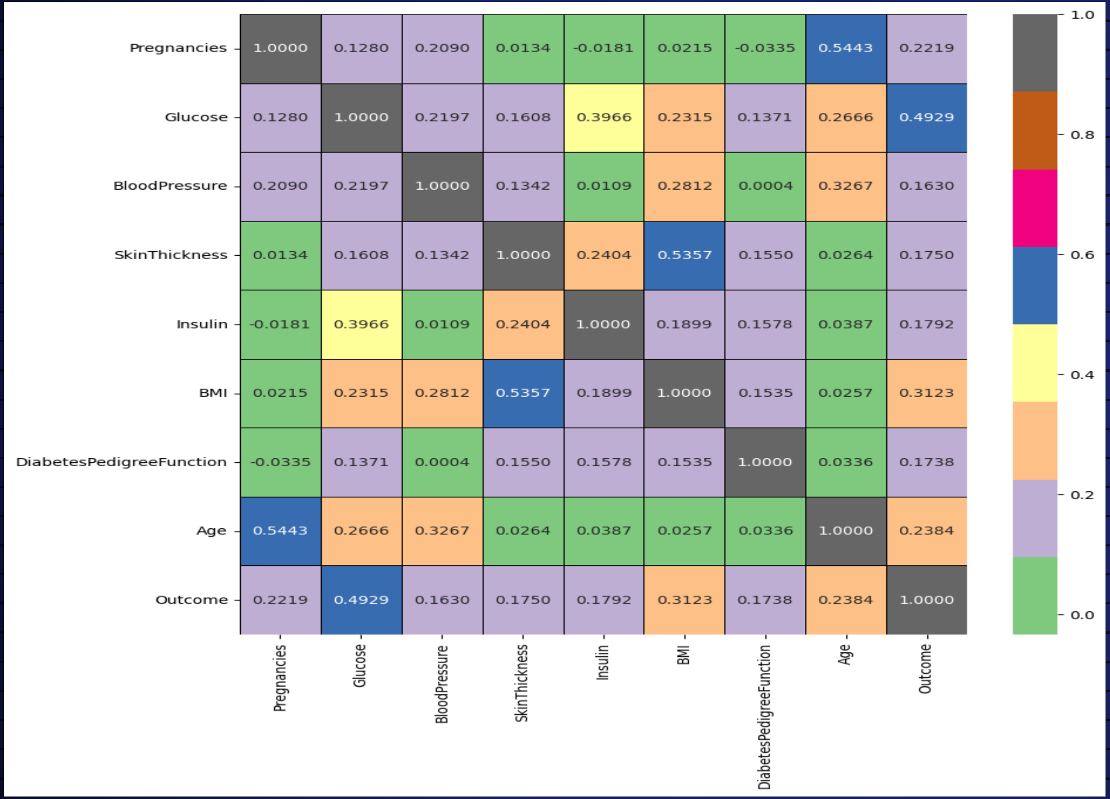
PROBLEM RESEARCH

The research objective is to develop a **predictive model** for diabetes diagnosis. This endeavor is motivated by the potential to **improve healthcare** by enabling early intervention and targeted care, contributing to public health insights, aiding clinical decision-making, and promoting preventative measures. Additionally, it provides an opportunity to **apply machine learning techniques** to a relevant healthcare challenge, ultimately benefiting both patients and the medical community.

DATASET INFORMATION

Pregnancies:	To express the Number of pregnancies
Glucose:	To express the Glucose level in blood
BloodPressure:	To express the Blood pressure measurement
SkinThickness:	To express the thickness of the skin
Insulin:	To express the Insulin level in blood
BMI:	To express the Body mass index
DiabetesPedigreeFunction:	To express the Diabetes percentage
Age:	To express the age
Outcome:	To express the final result 1 is Yes and 0 is No

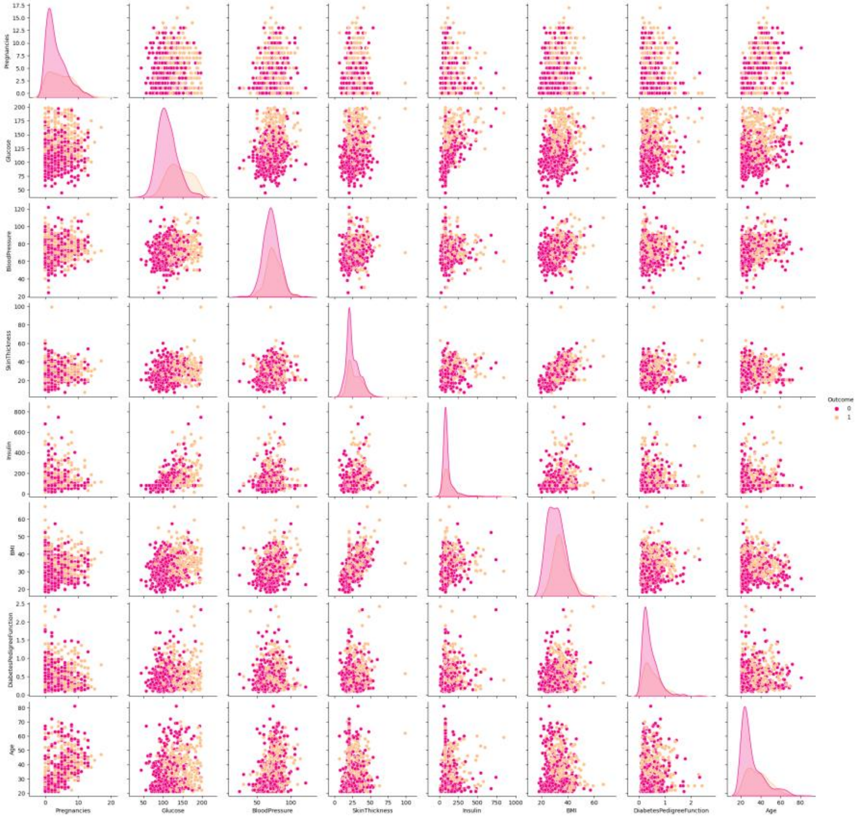
Data Visualization

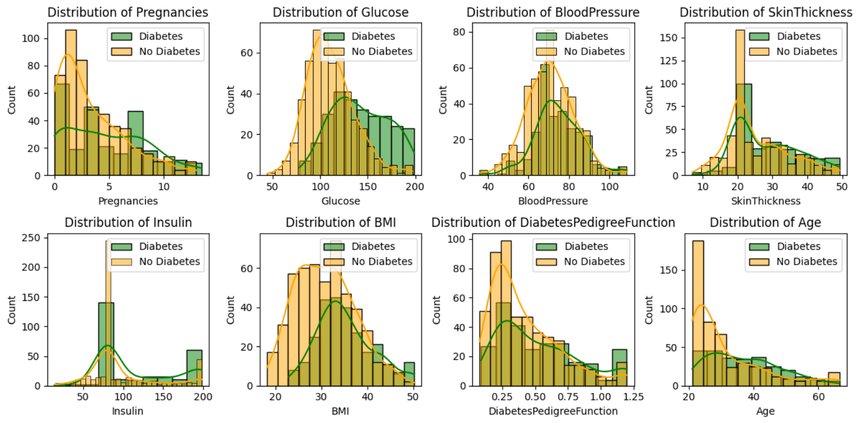


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.0000	0.1280	0.2090	0.0134	-0.0181	0.0215	-0.0335	0.5443	0.2219
Glucose	0.1280	1.0000	0.2197	0.1608	0.3966	0.2315	0.1371	0.2666	0.4929
BloodPressure	0.2090	0.2197	1.0000	0.1342	0.0109	0.2812	0.0004	0.3267	0.1630
SkinThickness	0.0134	0.1608	0.1342	1.0000	0.2404	0.5357	0.1550	0.0264	0.1750
Insulin	-0.0181	0.3966	0.0109	0.2404	1.0000	0.1899	0.1578	0.0387	0.1792
BMI	0.0215	0.2315	0.2812	0.5357	0.1899	1.0000	0.1535	0.0257	0.3123
DiabetesPedigreeFunction	-0.0335	0.1371	0.0004	0.1550	0.1578	0.1535	1.0000	0.0336	0.1738
Age	0.5443	0.2666	0.3267	0.0264	0.0387	0.0257	0.0336	1.0000	0.2384
Outcome	0.2219	0.4929	0.1630	0.1750	0.1792	0.3123	0.1738	0.2384	1.0000

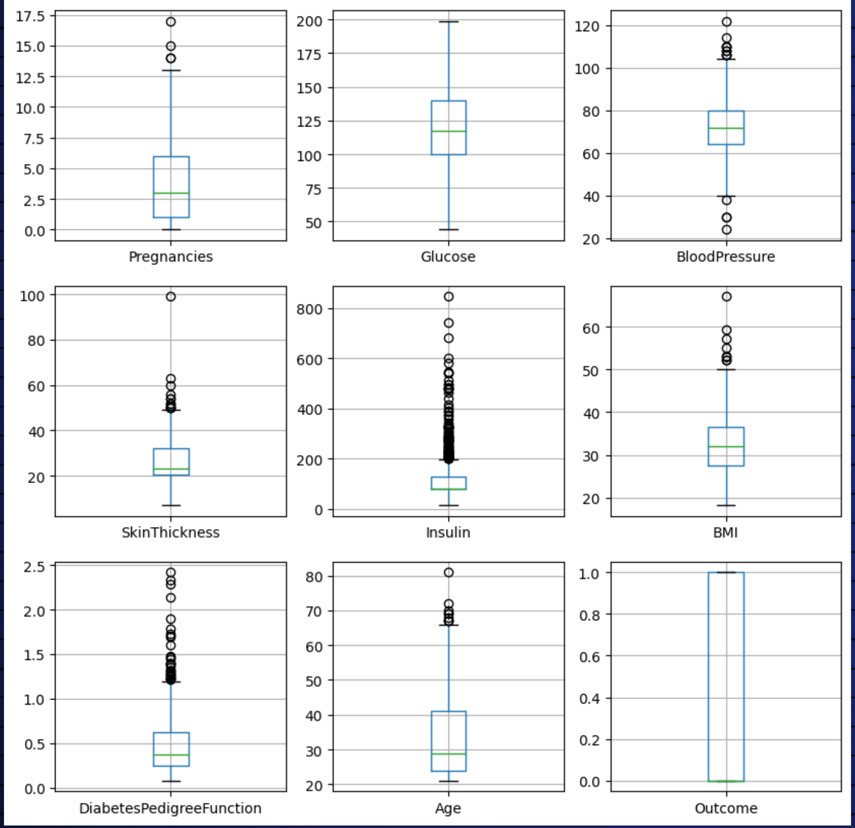
There are **no features that have a high positive correlation** with each other.

Data Visualization



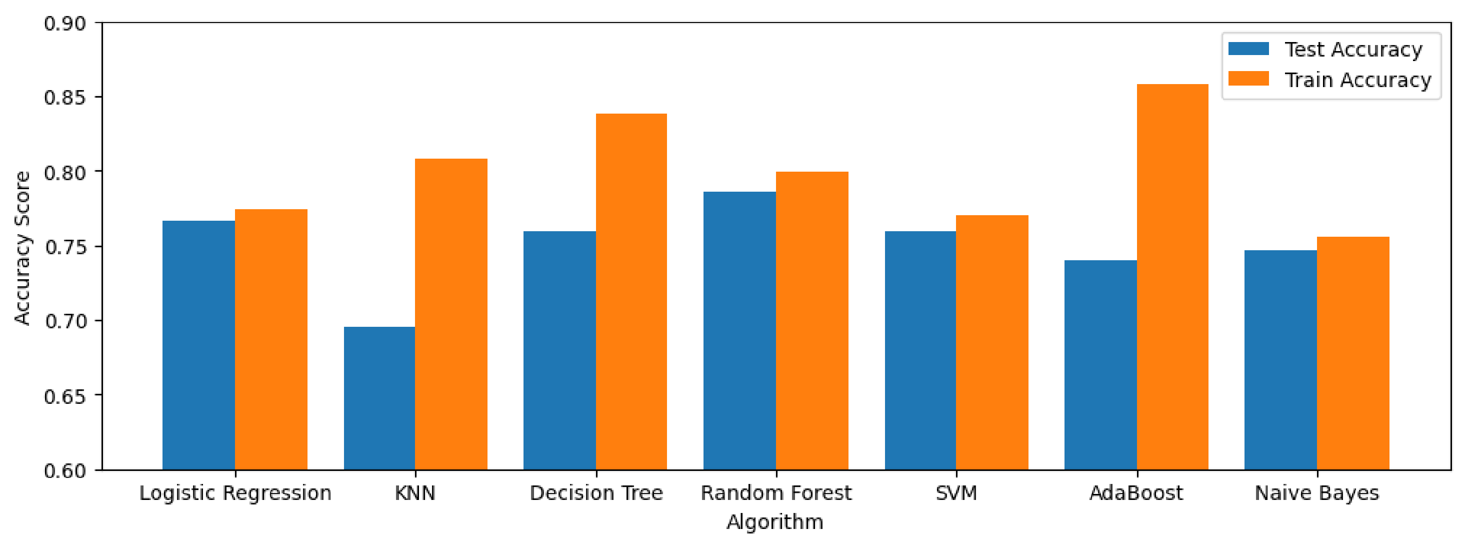


Data Visualization



Some data has many outliers, so these **outliers must be removed**

Machine Learning Output



Among the given machine learning algorithms, the **Random Forest algorithm** appears to perform the best with a test accuracy of 0.7857 and a relatively close train accuracy of 0.7997. Random Forest combines the strength of multiple decision trees and tends to be robust and accurate in a variety of scenarios, which **is reflected in its performance on the test data**. It strikes a **good balance** between generalization and overfitting, making it a strong candidate for the model of choice.