Symptom-Based Disease Diagnosis

Introduction

Welcome to the project progress report of symptom-based disease diagnosis. In this session, we will review our progress on the project.

The Need for Predictive Diagnosis

Traditional diagnostic methods often rely on subjective observations and can lead to misdiagnosis. **Advanced predictive models** offer a data-driven approach to accurately identify diseases based on *symptoms*. By leveraging machine learning algorithms and big data, we can achieve faster and more accurate diagnoses, leading to improved patient outcomes.

Machine Learning in Disease Diagnosis

Machine learning algorithms play a crucial role in **building predictive models** for disease diagnosis. By analyzing large datasets of patient symptoms and outcomes, these models can learn patterns and make accurate predictions. This enables **early detection** of diseases, personalized treatment plans, and improved healthcare delivery.

Benefits of Predictive Models

Predictive models offer several benefits in disease diagnosis. They enable **early intervention** by identifying diseases at their earliest stages. These models provide **personalized treatment plans** tailored to individual patients, improving patient care and outcomes. Additionally, they assist healthcare professionals in **decision-making** by providing evidence-based recommendations.

Challenges and Ethical Considerations

While predictive models hold immense potential, they also come with challenges and ethical considerations. Ensuring **data privacy** and **security** is crucial to protect patient information. Additionally, **algorithm bias** and **interpretability** of results need to be addressed to ensure fair and transparent diagnosis. Overcoming these hurdles will pave the way for widespread adoption of predictive models in healthcare.

Testing of Model Code

We have used a data set of patient with diabetes for testing various machine learning algorithms which can be applied to solve our model of symptom based disease detection.
The Dataset consists of various factors such as BMI, blood pressure, insulin etc. The Data is trained in a few ML algorithms using a 70:30 split .

About the dataset

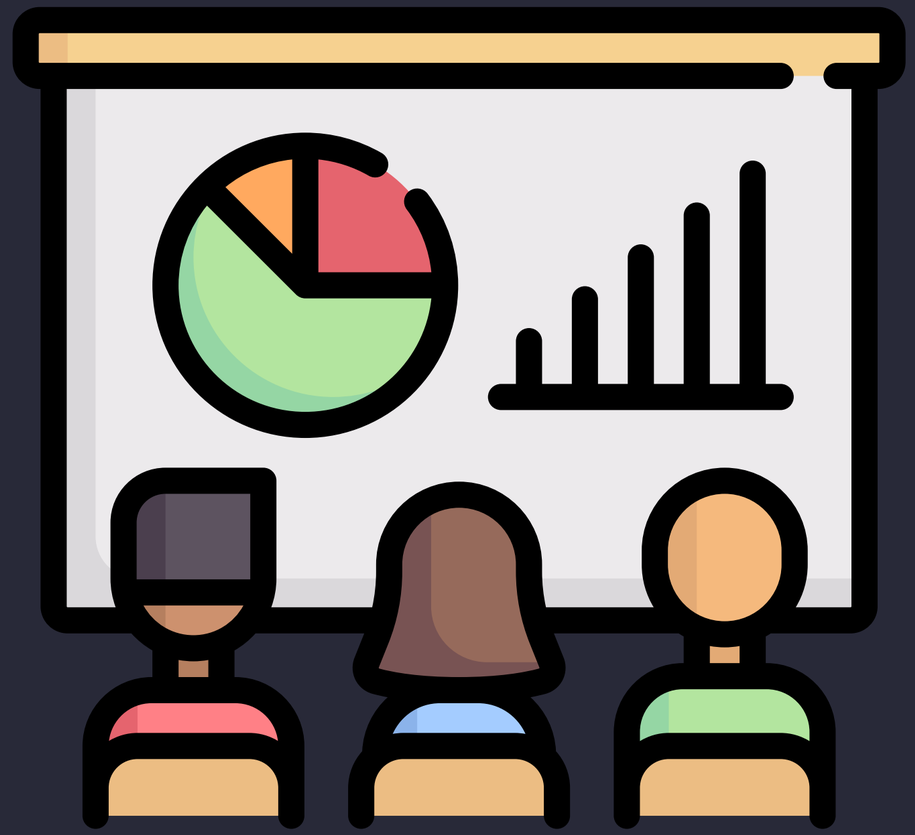
A premade dataset was available for us
The dataset has been taken from kaggle

About the test

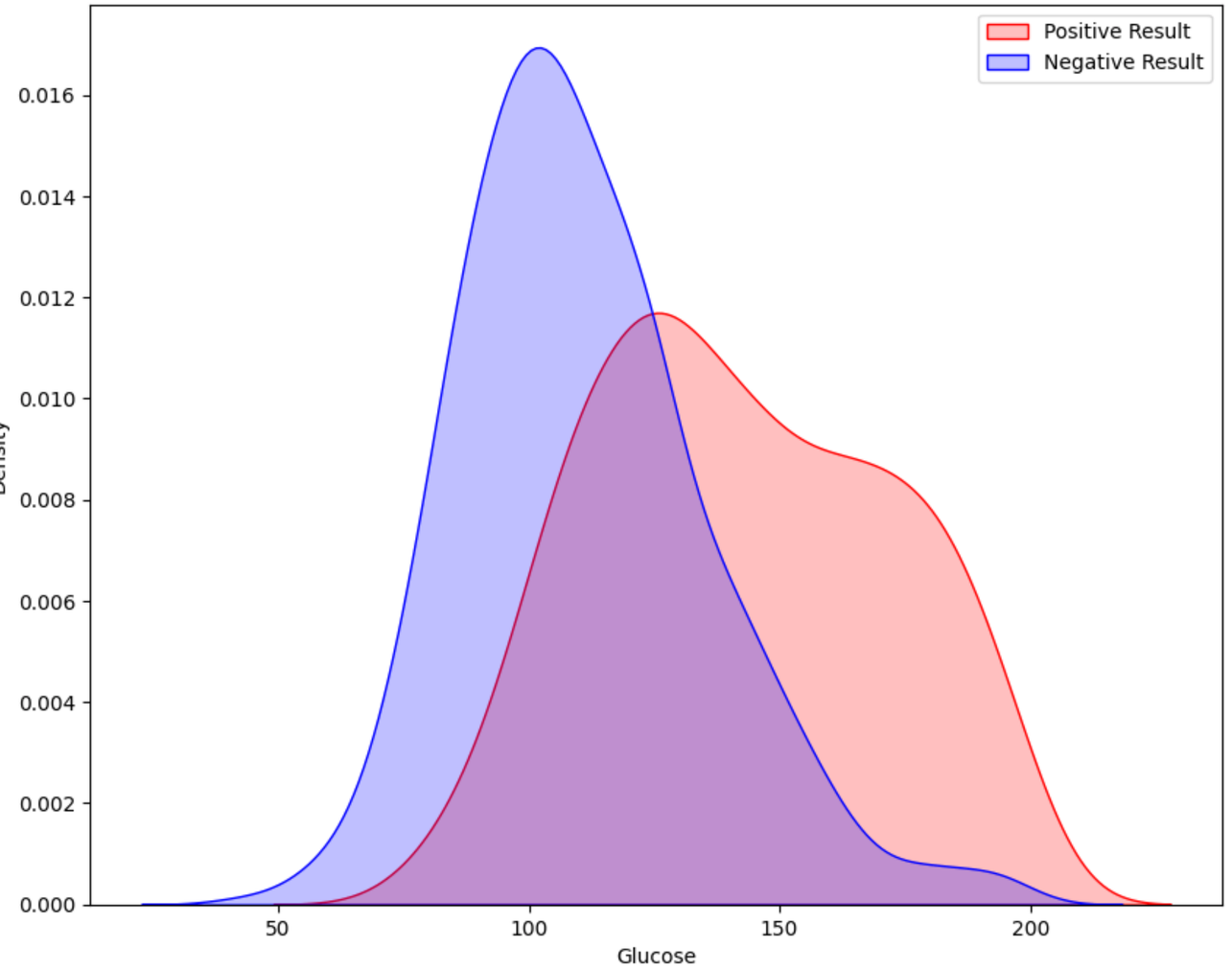
The obtained dataset was trained using the following ML models : KNN, Logistic Regression, Random-forest, SVM and Decision Tree

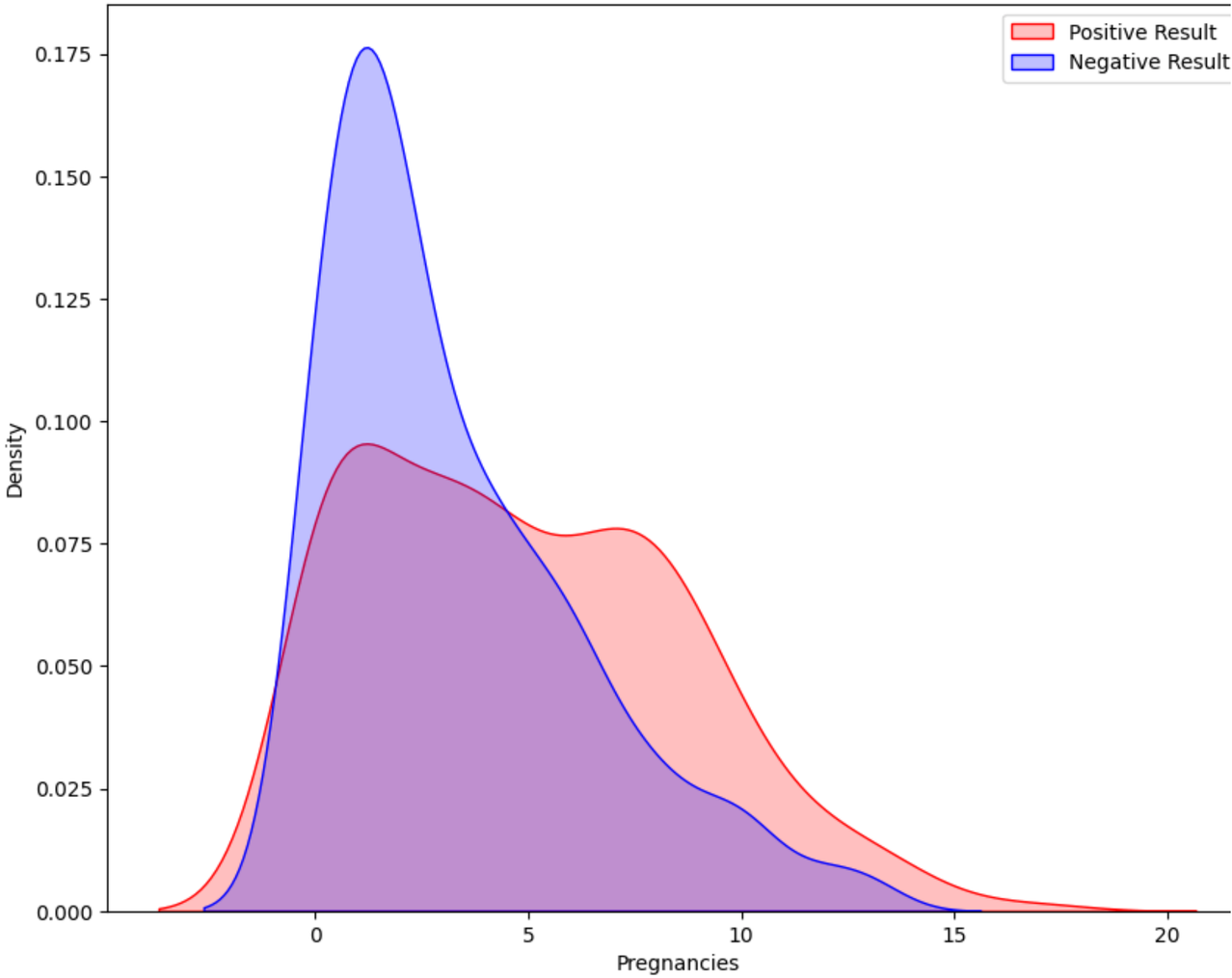
The dataset was cleaned so that it can be used for our work and the data was plotted on a correlation matrix for easy understanding of the relation between two elements.

The following results were obtained



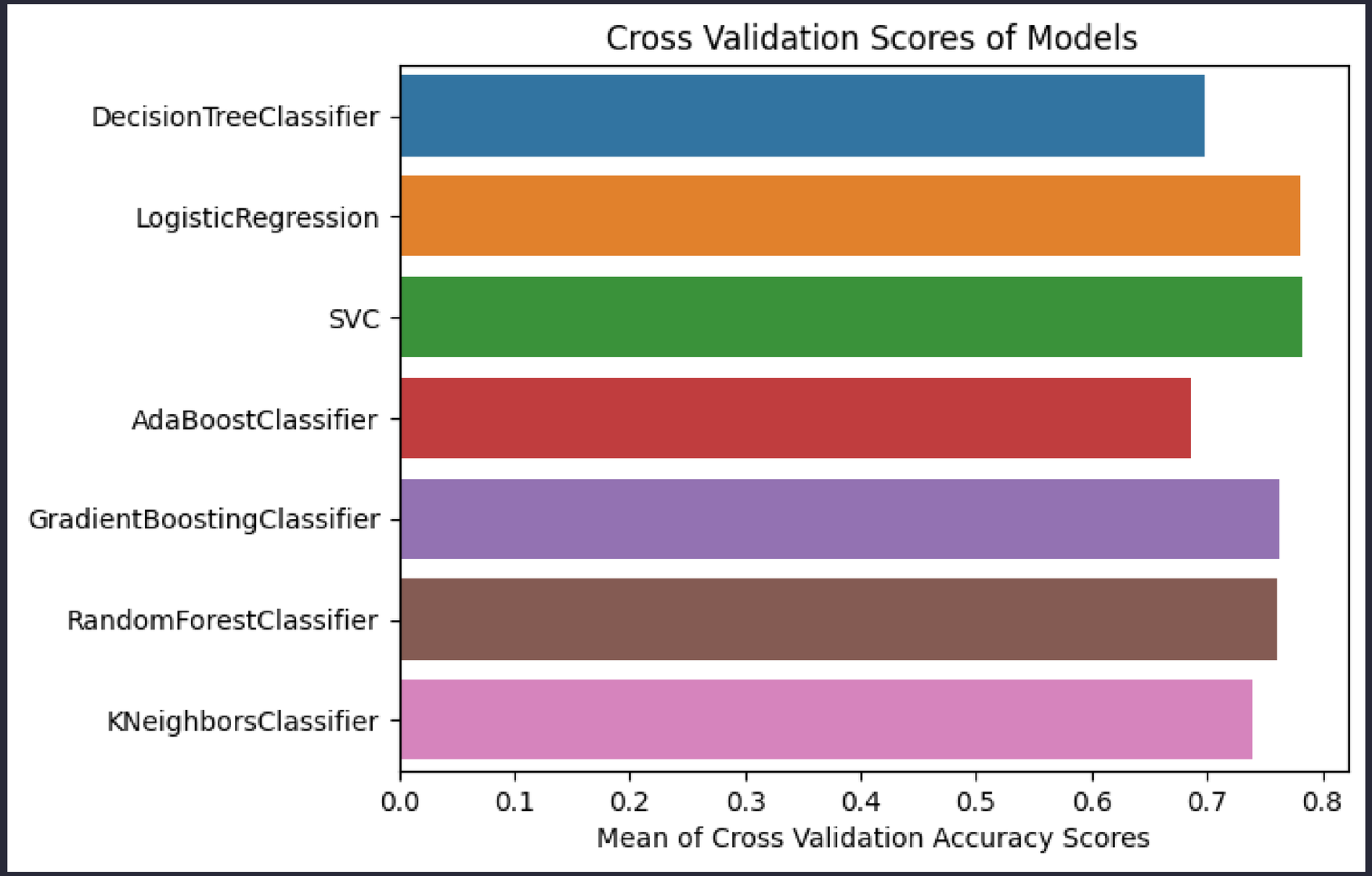
Observations show that characteristics like pregnancy, glucose, BMI, and age are more closely associated with outcomes. I demonstrated a detailed illustration of these aspects in the following phases.


Positive Result
Negative Result
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
50
100
150
200
Glucose


Positive Result
Negative Result
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Density
0
5
10
15
20
Pregnancies

The below acurracies were obtained after training

The acurracies were compared using a bar graph



Random-forest, Logistic regression and Support vector machines had higher accuracy so we further proceeded with tuning the hyperparameters for these 3 models and the accuracy scores were obtained

Outcome

Models	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity	Cohen's Kappa Value
KNN	82.69	0.79	0.76	0.78	0.82	0.83	0.632
LR	88.46	0.92	0.92	0.92	0.91	0.83	0.745
DT	78.84	0.64	0.89	0.74	0.73	0.88	0.571
RF	96.15	0.92	0.96	0.94	0.97	0.94	0.915
GBM	65.38	0.65	0.62	0.64	1	0	0
XGBoost	90.38	0.88	0.83	0.86	0.94	0.83	0.784
Ada-Boost	78.84	0.64	0.79	0.84	0.73	0.88	0.571
SVM	65.38	0.65	0.71	0.69	1	0	0.554
GNB	92.3	0.94	0.83	0.88	0.97	0.83	0.825

Outcome

Models	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity	Cohen's Kappa Value
KNN	84.9	0.86	0.98	0.92	0.84	0.86	0.185
LR	85.2	0.87	0.98	0.92	0	1	0.175
DT	84.1	0.84	1	0.91	0	1	0.198
RF	90.1	0.9	0.94	0.91	0.0054	0.997	0.197
GBM	84.1	0.85	1	0.9	0	1	0
XGBoost	96.1	0.96	0.95	0.98	0.78	0.94	0.278
Ada-Boost	84.1	0.84	1	0.91	0	1	0
SVM	83.8	0.84	0.98	0.91	0.055	0.965	0.187
GNB	75.8	0.91	0.81	0.85	0.0426	0.9891	0.279

Observation

It's observed that the ML algorithms Knearest Neighbour, Logistic Regression, Random Forest Classifier and Decision Tree are giving the best accuracies among others.

Thank you