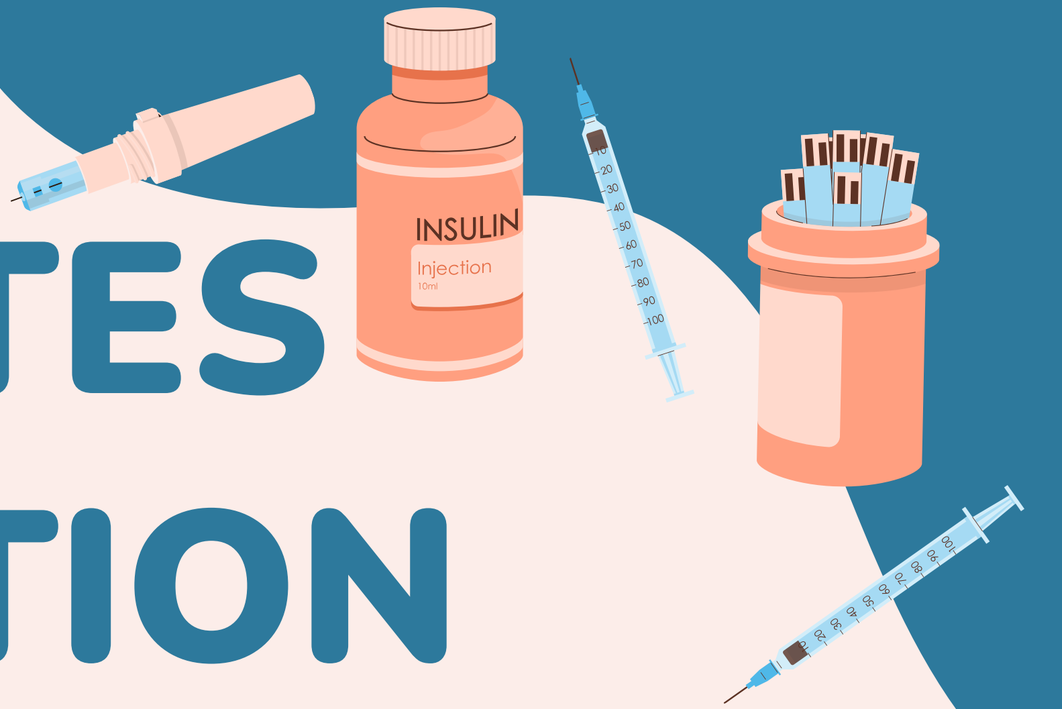
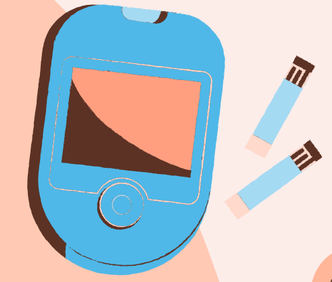
DIABETES PREDICTION

MACHINE LEARNING - WEBAPP





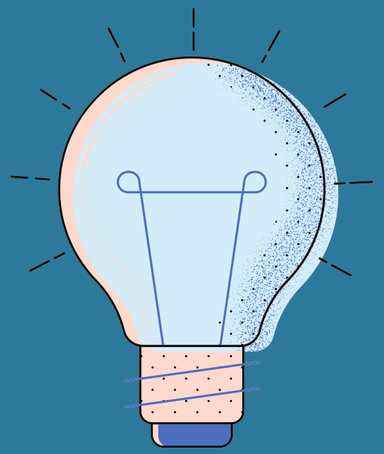
Bhavya Marupura - 50
Sai Krishna Vaibhav Mogili - 56

TODAY'S AGENDA

- Objectives
- Introduction
- Research Gaps
- Methodology
- Results
- Conclusions
- Future Works

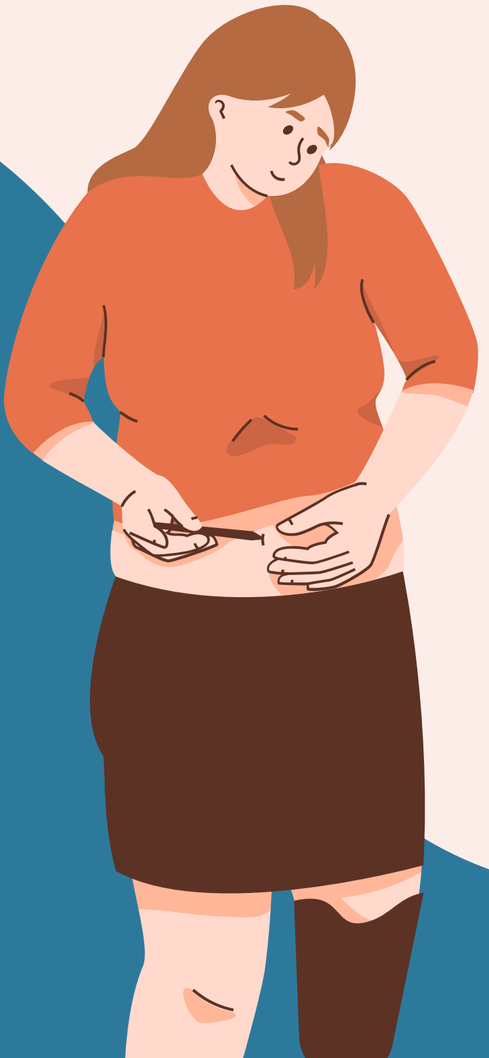
OBJECTIVES

- To enhance accessibility through a user-friendly interface, ensuring easy input of health parameters for diabetes prediction.
- Implementing strategies for continuous training of the model to improve the model’s efficacy in predicting diabetes across diverse demographics.

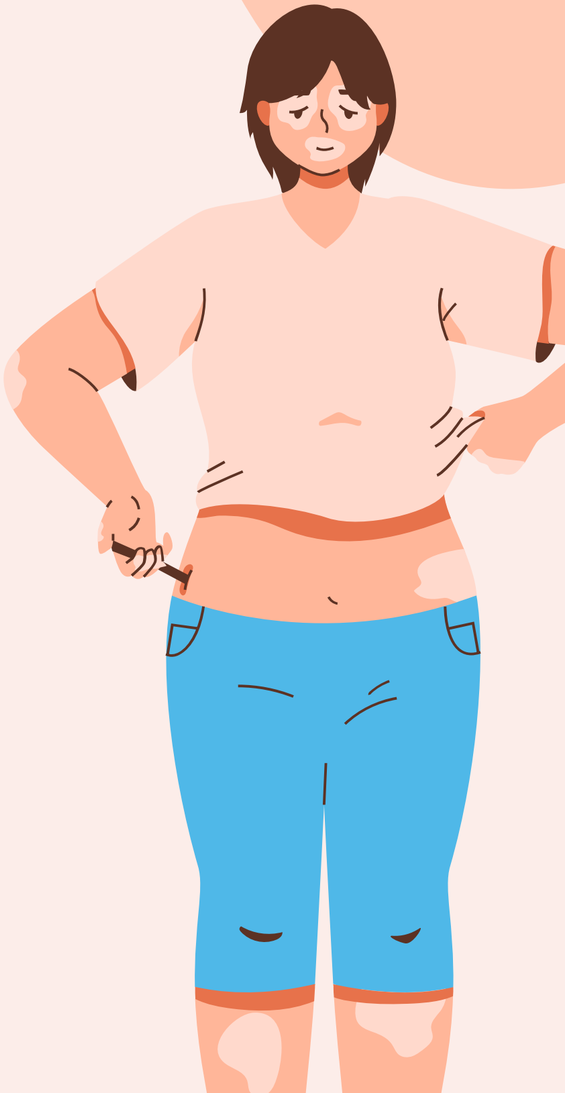


INTRODUCTION

WHAT IS DIABETES?



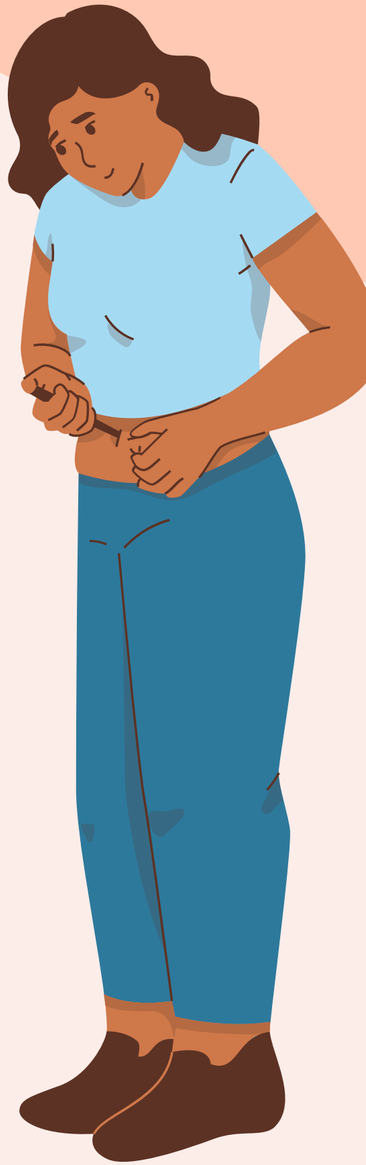
- Diabetes puts a person at risk of heart disease, kidney disease, stroke, eye problems, blood vessel damage, nerve damage, and other conditions making the body incapable of producing insulin.
- Proactive management and early detection are essential to reducing its negative effects on people's health.



INTRODUCTION

WHAT IS OUR PROJECT?

- This project presents a novel method of predicting diabetes by utilizing machine learning, specifically Support Vector Machine (SVM) algorithms with four types of kernels: polynomial, sigmoid, RBF, and linear.
- The model is trained using data from both diabetic and nondiabetic instances (PIMA Indian Dataset) and is integrated into an easy-to-use web application interface using the Streamlit library in Python.

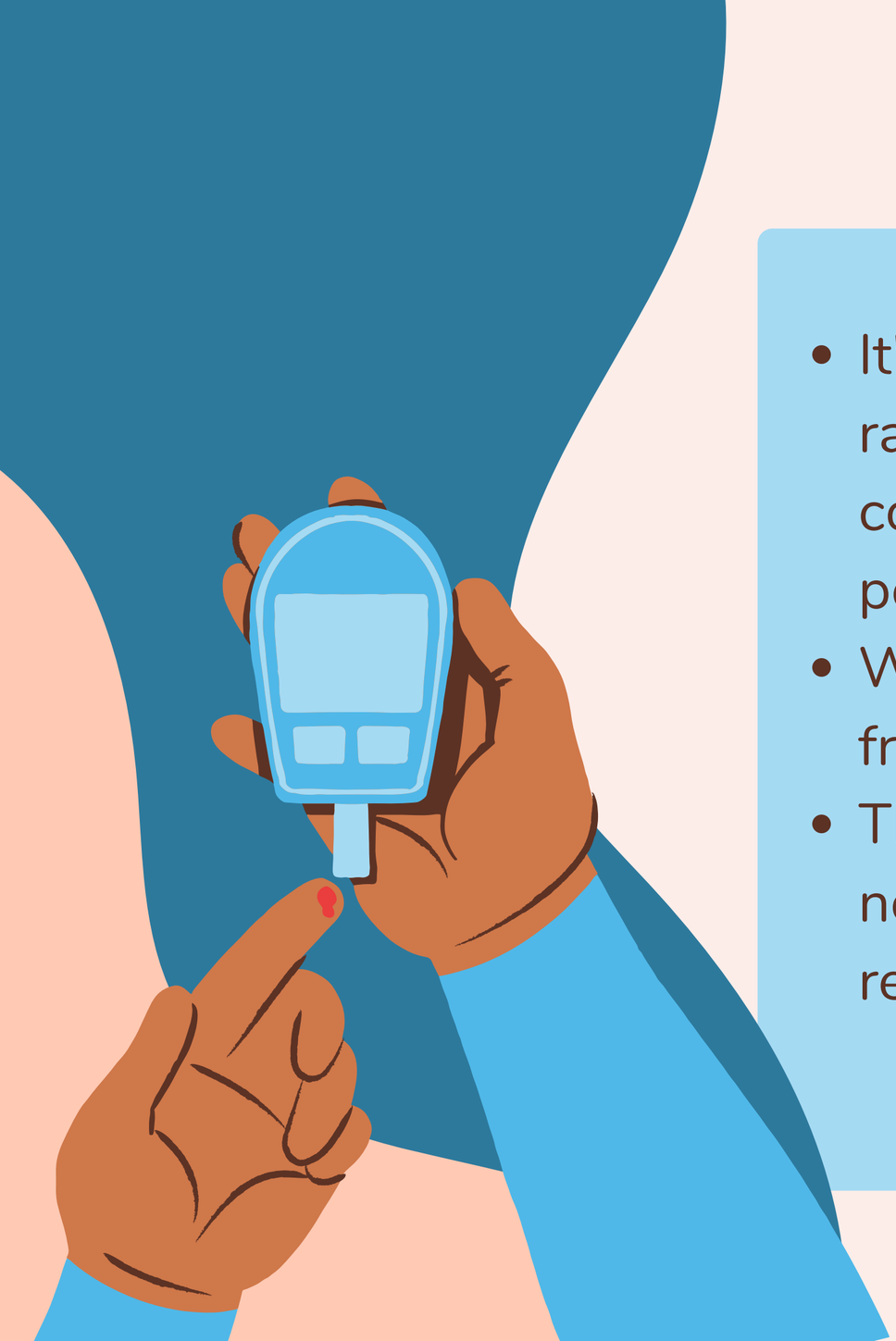


PIMA INDIAN DATASET

diabetes csv

File Edit View Help

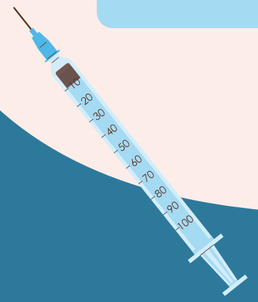
	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1



RESEARCH GAPS

- It's possible that the model's applicability to a broader range of demographics than just the Pima Indian community was limited by the dataset's lack of diversity in population representation.
- Widespread adoption may be impeded by the lack of user-friendly interfaces in many of the current models.
- The model may lose its efficacy over time if fresh data are not introduced to it regularly or if it is not adjusted to reflect the evolving health trends.

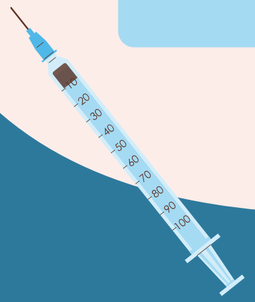
METHODOLOGY



1) Importing Libraries and Data Collection

Utilize Python libraries like Pandas for data alteration, NumPy for computational operations, and Scikit-learn for machine learning tools and access the PIMA Indian Diabetes dataset

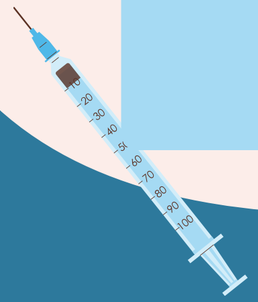
METHODOLOGY



2) Data Preprocessing and Standardizing

Handle missing values, outliers, and inconsistencies within the dataset and scale features to ensure all parameters have a similar impact during modeling.

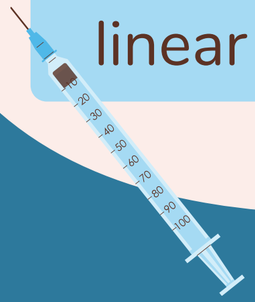
METHODOLOGY



3) Data Splitting

Using an 80:20 ratio, split the pre-processed dataset into training and testing sets. Model training will take place on the training set (80%), and model performance evaluation will take place on the testing set (20%).

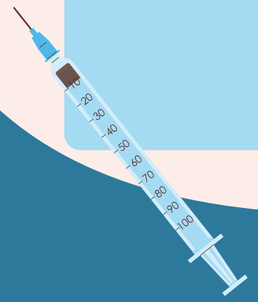
METHODOLOGY



4) Training Predictive Model

Machine Learning models are trained using Support Vector Machines (SVM - supervised learning algorithm).
SVM aims to find a hyperplane that maximizes the margin between two classes (either diabetic or non-diabetic).
In this paper, we employ support vector machines (SVM) with four different kernel types: sigmoid, polynomial, RBF, and linear to identify diabetes and assess each case's accuracy.

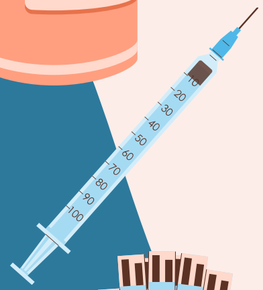
METHODOLOGY

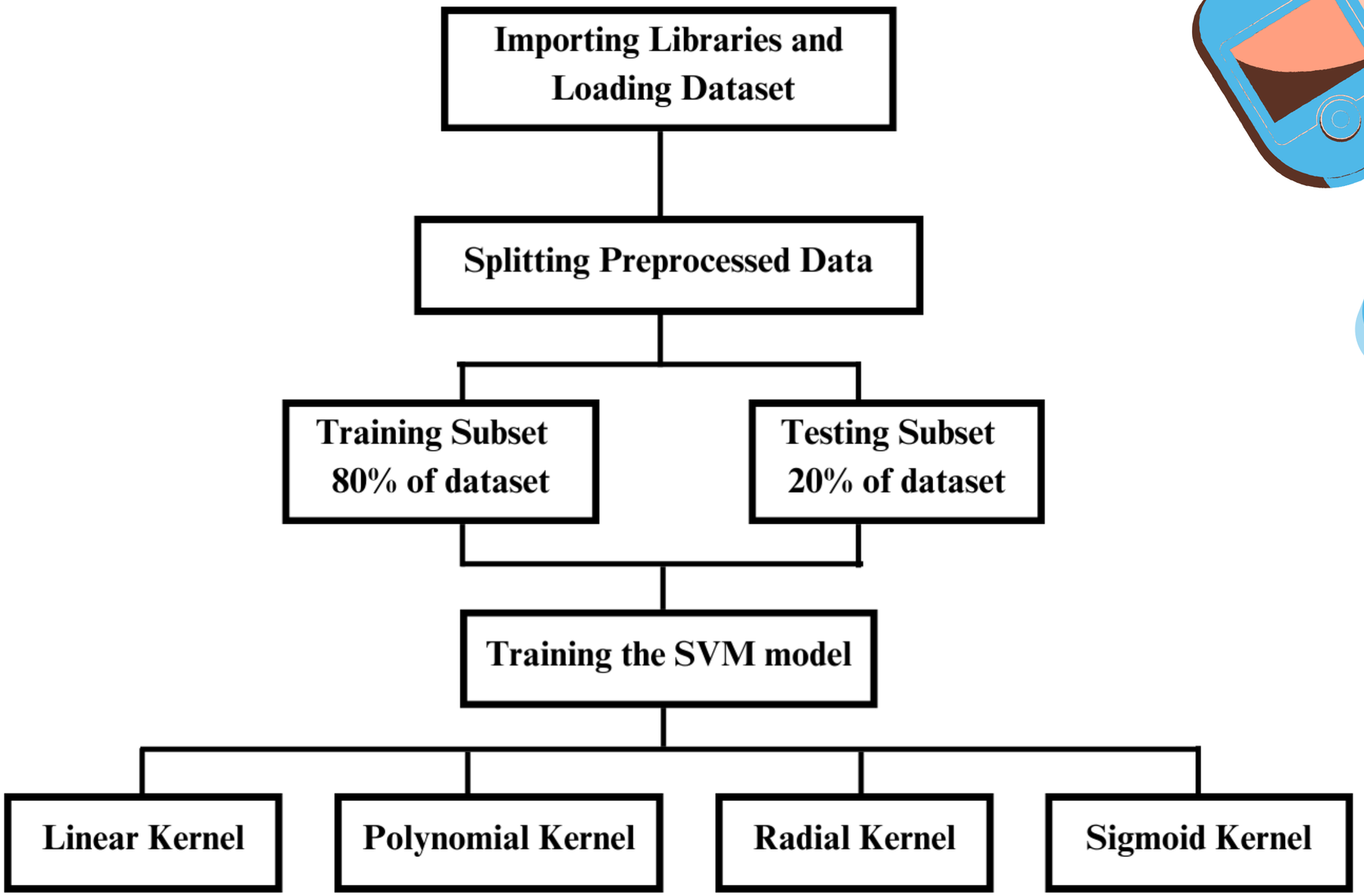


5) Web Application – Streamlit Library

By importing the pickle library from Python, we load our trained SVM models in binary format and use the Streamlit library to create an intuitive web interface.
We can now incorporate input fields in the web application to collect user medical data, passing it through the loaded SVM model will display the predicted diabetic status of the user.

METHODOLOGY





↓RESULTS

Features	Standard Values
Number of Pregnancies	4
Glucose Level	120.89
Blood Pressure	69.10
Skin Thickness	20.53
Insulin Level	79.79
BMI	31.99
Diabetes Pedigree Function	0.47
Age	34

The table below illustrates a trend that reveals the standard range of parameters that indicate the probability that a user has diabetes based on the values of the 789 instances of the dataset that are present. A user is more likely to have diabetes if their records are higher than the standard values, as shown in the table below.

SEM 5/ML LAB/ | Diabetes Prediction-SVM-KERNE | diabetes_prediction_webapp · Str

localhost:8501

Deploy

Diabetes Prediction WebApp

Number of Pregnancies

2

Glucose Level

197

Blood Pressure Value

70

Skin Thickness Value

45

Insulin Level

543

Body Mass Index (BMI) Value

30.5

Diabetes Pedigree Function Value

0.158

Age of the User

Blood Pressure Value

70

Skin Thickness Value

45

Insulin Level

543

Body Mass Index (BMI) Value

30.5

Diabetes Pedigree Function Value

0.158

Age of the User

53

Diabetes Test Result

Person is Diabetic

SEM 5/ML LAB/ | Diabetes Prediction-SVM-KERNE | diabetes_prediction_webapp · Str

localhost:8501

Deploy

Diabetes Prediction WebApp

Number of Pregnancies

1

Glucose Level

103

Blood Pressure Value

30

Skin Thickness Value

38

Insulin Level

83

Body Mass Index (BMI) Value

43.5

Diabetes Pedigree Function Value

0.183

30

Skin Thickness Value

38

Insulin Level

83

Body Mass Index (BMI) Value

43.5

Diabetes Pedigree Function Value

0.183

Age of the User

33

Diabetes Test Result

Person is Not-Diabetic

↓CONCLUSIONS

Kernel Type	Train Accuracy	Test Accuracy
Linear Kernel	0.7654	0.8246
Polynomial Kernel	0.7850	0.7792
RBF Kernel	0.8469	0.7922
Sigmoid Kernel	0.6840	0.7402

Considering these results, the RBF Kernel emerges as the most suitable choice for the model. Despite a slightly lower test accuracy compared to the Linear Kernel, its robust performance on both training and test datasets signifies a good balance between complexity capture and generalization.

FUTURE WORKS

- Expanding and updating the dataset to allow accurate prediction of a wider diversity.
- Enhancing the web application's by integrating features such as personalized health recommendations based on predictions and incorporating additional health parameters could provide a more comprehensive health assessment for users.
- Collaborations with healthcare professionals to validate the model's predictions and ensure alignment with clinical diagnoses would bolster its reliability and applicability in real-world healthcare settings.

THANK YOU!