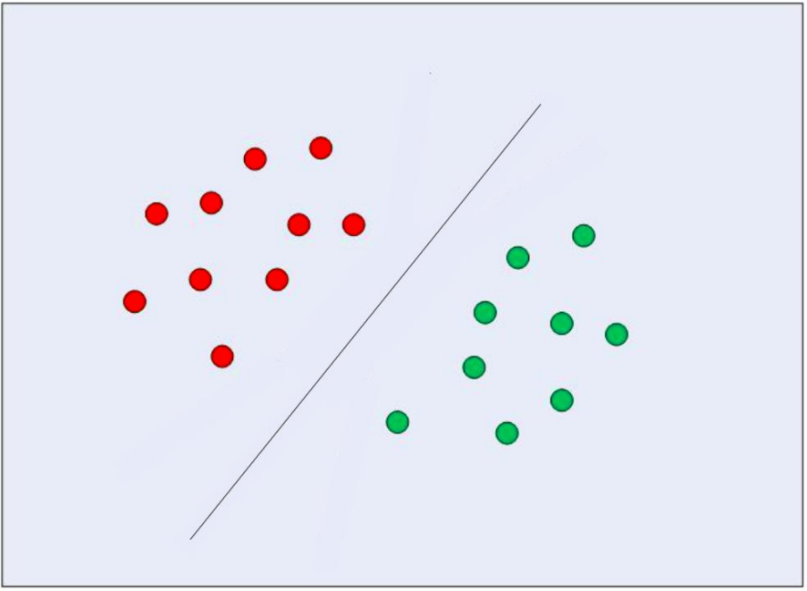


SVM- Based Classification of Diabetes Risk Using Demographic and Behavioral Features

ABSTRACT

The aim of the project is to classify individuals as diabetic and non- diabetic based on selected features using SVM models. After data cleaning and preprocessing, three types of SVM classifiers were trained using linear, radial basis function and polynomial kernel. Hyperparameter tuning was performed using GridSearchCV with cross- validation to optimize model performance. The models were evaluated using accuracy, confusion matrices and classification reports. Decision boundaries were visualized using two features for better understanding of model behavior. Results indicated that RBF performed well with approximately 81% test accuracy. These findings suggest that non- linear svm classifiers are used for early diabetes prediction based on basic demographic and lifestyle features.

INTRODUCTION

Diabetes is one of the major health issues affecting millions of people worldwide. It is a chronic condition that can lead to serious health complications if not detected and managed early. So early prediction and prevention is very important.

Machine learning, especially classification algorithms, offers powerful tools to identify individuals at risk of developing diabetes using demographic and lifestyle data. This study applies Support Vector Machine (SVM) models with linear, RBF, and polynomial kernels to predict diabetes status based on selected features from the 2022 NHIS dataset.

THEORETICAL BACKGROUND

SVM: SVM are supervised learning methods used for classification and regression.

Tries to find the best boundary that separates individual with diabetics and individual with no diabetics with the maximum margin between them.

Hyperplane: Separates the data points of different classes it is a line in 2D, a plane in 3D or a higher dimensional surface.

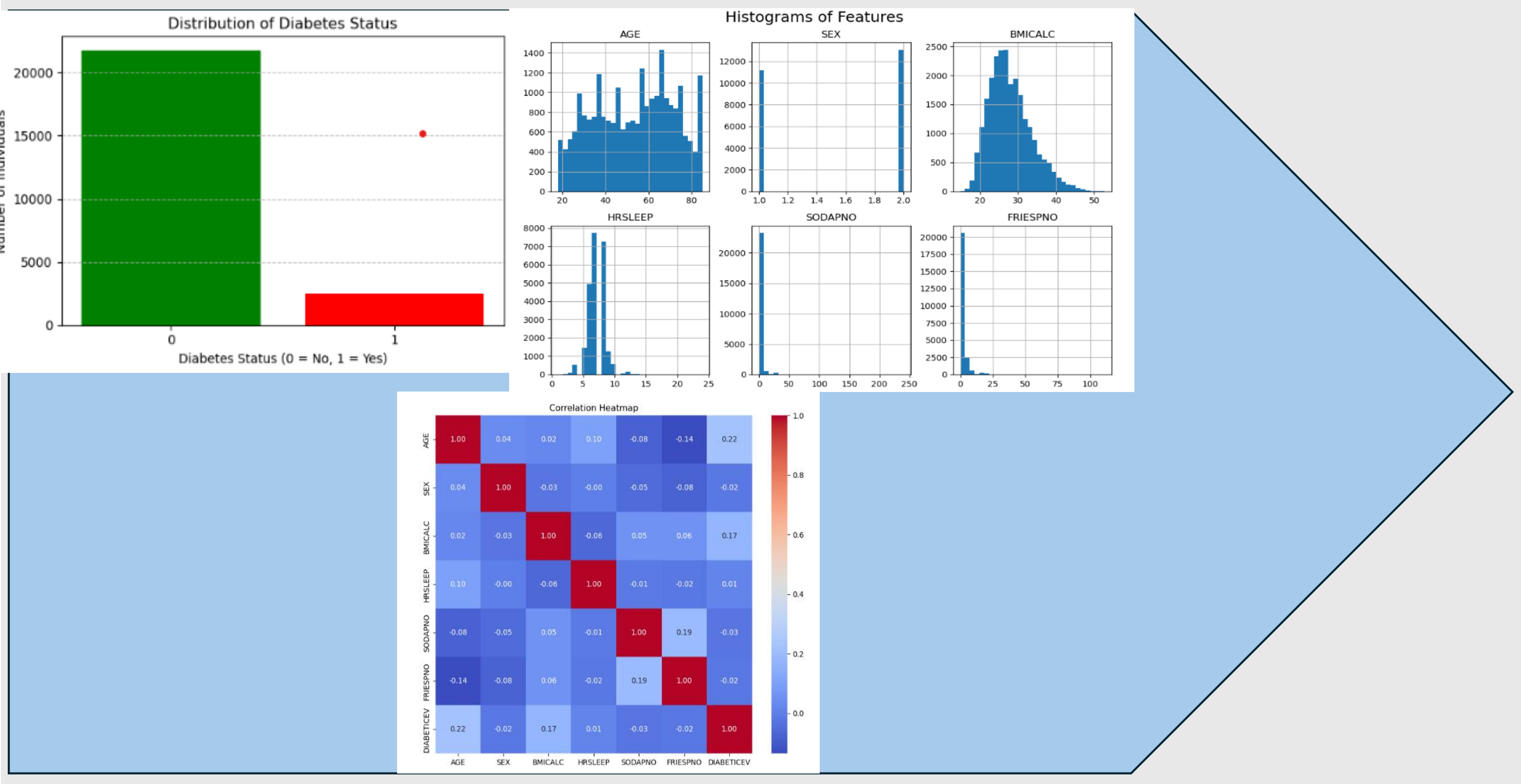
Support Vector Machine (SVM) classifiers were fitted using three different kernels: linear, radial basis function (RBF), and polynomial

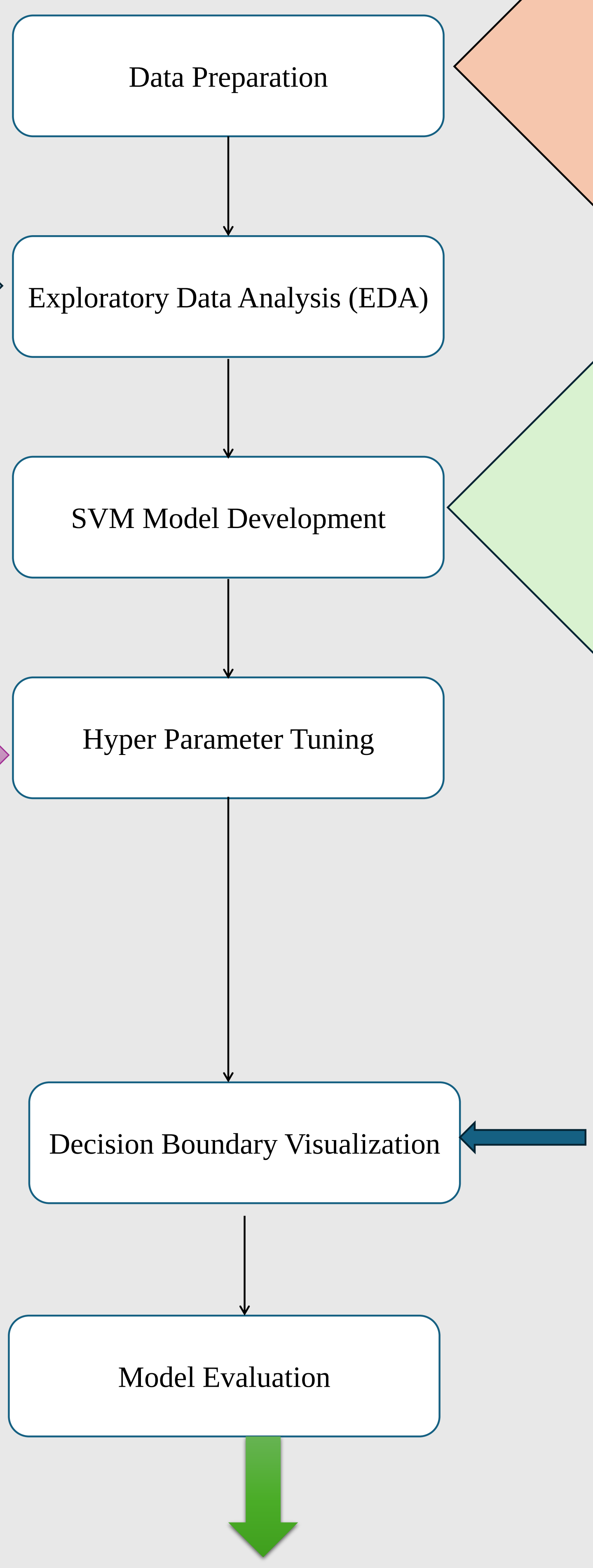
Kernel	Decision boundary shape	Hyperparameters	Values used
Linear	Straight line	Cost (C)	C= 0.01, 0.1,1

RBF	Smooth Curve	Cost(C)
		Gamma
			C = 0.01, 0.1, 1
			Gamma = 10, 1, 0.1

Polynomial	Polynomial curves	Cost
		Degree
			C = 0.01, 0.1, 1
			Degree = 2,3,4

METHODOLOGY





Data Preparation

1. Dropped survey-related and irrelevant fields.
2. Replaced special codes with NaN values.
3. Imputed missing values using median (for continuous) and mode (for categorical).
4. Dropped remaining rows with missing values.
5. Selected features: AGE, BMI, SEX, hours of sleep, soda and fries consumption.
6. Encoded target variable.

Exploratory Data Analysis (EDA)

SVM Model Development

1. Developed three SVM models: Linear Kernel, RBF Kernel, and Polynomial Kernel.
2. 80% of data is split to training set and 20% as test set.
3. Standardized features using StandardScaler.
4. Models fitted on training set and evaluated on the test set.

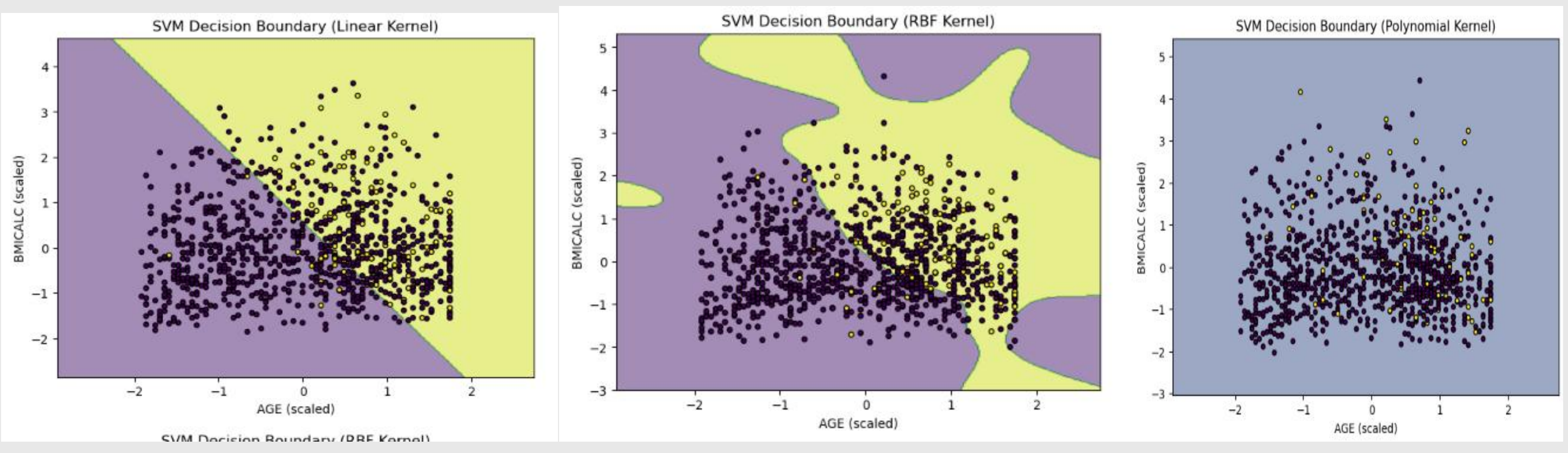
Hyper Parameter Tuning

Model	Hyperparameters Searched	Best Parameters Found
Linear SVM	C =[0.01, 0.1, 1, 10, 100]	C =1

RBF SVM	C = [0.1, 1, 10, 100]
	gamma = [0.001, 0.01, 0.1, 1]
		C = 100
		gamma = 1

Polynomial SVM	C =[0.1, 1,10]
	Gamma = [0.001,0.01,0.1]
	degree = [2,3,4]
		C = 0.1
		gamma = 0.001
		degree = 2

Decision Boundary Visualization



Model Evaluation

	Before Tuning				After Tuning			
Model	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 score
Linear	**63%**	**57%**	**69%**	**53%**	**63%**	**58%**	**72%**	**53%**
RBF	**64%**	**57%**	**69%**	**53%**	**71%**	**55%**	**61%**	**54%**
Polynomial	**62%**	**57%**	**69%**	**52%**	**10%**	**5%**	**50%**	**9%**

RESULTS

1. Before tuning all three models performed similarly.
2. After tuning RBF achieved highest accuracy 71% followed by linear svm.
3. Performed poorly, has low accuracy 10%. Predicting all as diabetic

CONCLUSION

SVM models identified individual at risk for diabetes with high recall but low precision. In real world application , such models can be served as screening tools to flag individual for further clinical evaluation

To strengthen the analysis additional medical, lifestyle and laboratory data would be needed to improve prediction precision and overall model reliability.

REFERENCES

[1] Scikit-learn developers. (2024). **Scikit-learn: Machine Learning in Python**. Retrieved from https://scikit-learn.org/stable/
[2] NIH National Institute of Diabetes and Digestive and Kidney Diseases. (2023). **Diabetes Overview**. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview
[3] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). **An Introduction to Statistical Learning: With Applications in R**. Springer.