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

Sowmya Polagoni

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 71% 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

Support Vector Machines: Support Vector Machine (SVM) is a supervised learning method used for used for classification and regression.
It aims to classify the data by finding the optimal separating hyperplane that maximizes the margin between different classes.

Hyperplane: Hyper plane is a decision boundary that separates two classes.
It is a line in 2D, a plane in 3D or a higher dimensional surface.

$$\vec{\omega} \cdot \vec{x} + b = 0$$

Where w = weight vector, b = bias , x = input data point

Support Vectors: Support vectors are the data points that are closest to the hyperplane important for determining the hyperplane and margin in SVM.

Margin: The margin is the distance between the hyperplane and the support vectors. SVM's objective is to maximize this margin while minimizing misclassification:

$$\textbf{Margin} = \frac{2}{||\vec{\omega}||}$$

Kernels in SVM: SVM applies the kernel trick when data is not linearly separable, transforming input data to a higher dimensional space where a linear hyperplane can be used.

Kernel types and Parameters:
Different kernel types are used in SVMs, the most common kernel types are linear, polynomial, radial basis function.
Linear Kernel: Linear kernel is used when the data is linearly separable. Computes the inner product of two vectors:

$$K(x_i, x_i') = \sum_{j=i}^{p}(x_i, x_i')$$

Parameter –
C : Controls the cost of misclassification
Small C : Wider margin, more tolerance for misclassification.
Large C : Less tolerance, risk of overfitting.
Radial Kernel: Used when data is complex and highly non-linear. Maps input space into infinite – dimensional space to find a decision boundary. The kernel function is:

$$K(x_i, x_i') = \exp\left(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{ij}')\text{^}2\right)$$

Parameters used –
C: Controls the cost of misclassification
gamma : Control influence of points
larger gamma: smaller sphere of influence
Small gamma: larger sphere of influence
Polynomial Kernel: Used when data has polynomial relationships.

$$K(x_i, x_i') = (1 + \sum_{j=1}^{P} x_{ij}x_i'_j)^d$$

Parameters used –
gamma: Control influence of points
degree: Degree of the polynomial (e.g. 2 = quadratic)
High degree: more inflections
Coef0: Bias term, affects curve shape
Large coef0: more curved and flexible decision surfaces

DISCUSSION

The results of the SVM models and decision boundary plots suggest that **AGE** and **BMI** are strong predictors of diabetes. People who were **older** or had a **higher BMI** were more likely to be classified as diabetic.

The decision boundary plots show that the **RBF SVM** was best at separating the two groups. It created a smooth, curved line that followed the shape of the data, helping it detect more complex patterns. This tells us that the relationship between AGE, BMI, and diabetes is **not simple or straight**, and models like RBF that can handle curves work better.

In comparison, the **Linear SVM** created a straight line that didn't match the shape of the data, so it missed many cases. The **Polynomial SVM** didn't separate the groups clearly and gave poor results, possibly because it was too sensitive to noise or badly tuned.

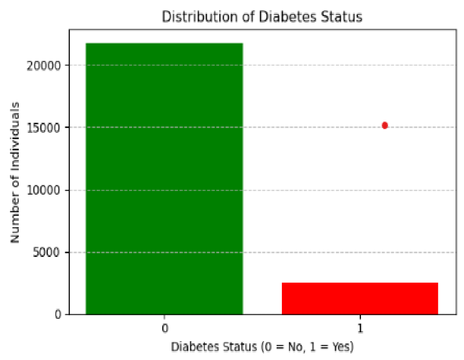
Hence **AGE and BMI together** carry important signals for predicting diabetes, but only the right kind of model can pick up on these patterns. These findings show how **lifestyle and age-related factors** are key in health prediction, and choosing the correct model is just as important as choosing the right features.

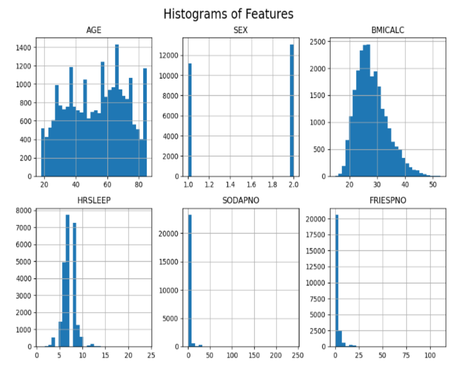
METHODOLOGY

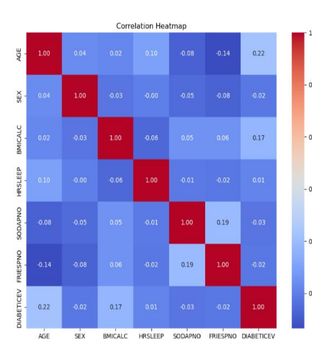
DATA PREPARATION

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 ANALYSIIS







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.

HYPERPARAMETER TUNING

Kernel Type	Hyperparameters searched	Best Parameters found
Linear	C =[0.01, 0.1, 1, 10]	C = 0.01
RBF	C = [0.1, 1, 10, 100] gamma = [0.001, 0.01, 0.1, 1]	C = 100 gamma = 1
Polynomial	C =[0.1, 1,10] Gamma = [0.001,0.01,0.1] degree = [2,3,4]	C = 0.1 degree = 2 gamma = 0.001

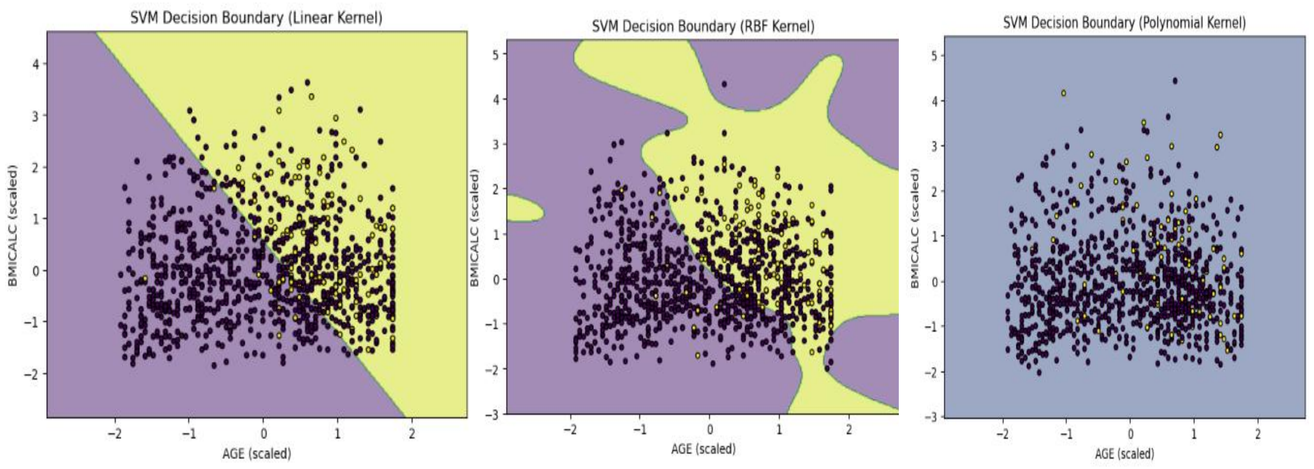
DECISION BOUNDARY VISUALIZATION

Plotted the decision boundaries for three different SVM models, Linear kernel, RBF kernel and Polynomial using two scaled features AGE and BMICALC.

MODEL EVALUATION

Model	Accuracy	Precision	Recall	F1 score
Linear	0.62	0.58	0.72	0.53
RBF	0.71	0.55	0.61	0.54
Polynomial	0.10	0.05	0.50	0.10

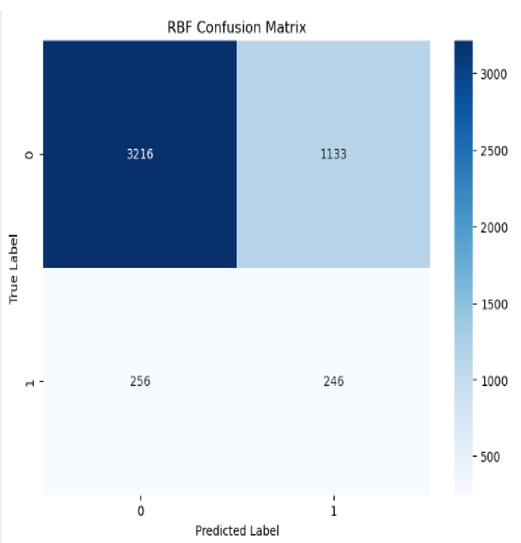
RESULTS



Three Support Vector Machine (SVM) models were evaluated using different kernels and Hyperparameter tuning techniques.
The **Linear kernel model** achieved an overall accuracy of 62.6%, with an average precision of 0.58, recall of 0.72, and an F1-score of 0.53. While it performed well on the dominant class, its overall balance was limited.
The **Tuned RBF kernel model** showed the best overall performance, achieving 71.4% accuracy. It had an average precision of 0.55, recall of 0.61, and an F1-score of 0.54, indicating a better trade-off between identifying positive cases and avoiding false alarms.
The **Polynomial kernel model** performed poorly, with only 10.3% accuracy, an average precision of 0.05, recall of 0.50, and an F1-score of 0.09. It failed to distinguish between the classes and misclassified most samples.

CONCLUSIONS

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



The RBF model performed best, as shown in its confusion matrix. It correctly identified **3,216 non-diabetic** and **246 diabetic** individuals, with fewer false negatives than other models. This shows that the RBF kernel handled the complex relationship between AGE, BMI, and diabetes more effectively than the Linear or Polynomial models.

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.