Predicting Diabetes Using Support Vector Machines (SVM)

Pavan Arram | Data Science Practical Homework 2

SEATTLE UNIVERSITY

Introduction

Goal: Predict likelihood of diabetes based on lifestyle and health factors.

Data: NHIS 2022 Adult dataset (~35,000+ records). Focus: Variables related to health (BMI, exercise, diet) and demographics (age).

Model: Support Vector Machine (Linear, Radial, Polynomial kernels).

Theoretical Background

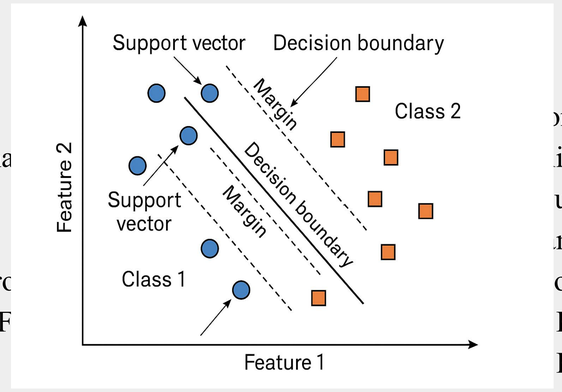
What is SVM?

A supervised learning method that finds an optimal separating hyperplane.

Important Equations:

Objective:

$(1/2) * ||w||^2$ subject to $y_i (w^T x_i + b) \geq 1$



Key oncepts:

Hyperplane, ma ization.

Support vectors define undary.

Tuning meters:

Cost (C): Contro olation.

Gamma (γ): F Kernel.

Degree: For Kernel.

Linear Kernel:

Projects data without transformation.

$K(x, y) = x . y$ $K(x,y)=x.y$

Works well when data is linearly separable.

Polynomial Kernel:

Projects into a higher dimension using polynomial terms.

$K(x,y)=(x.y+c)^d$

Suitable for moderate non-linearity.

RBF (Radial Basis Function) Kernel:

Projects into infinite-dimensional space to capture complex patterns.

$K(x,y)=e–(\gamma||x–y||2)$

Works well with highly non-linear data.

Methodology

Preprocessing:

Removed special codes (996–999).

Standardized all predictors.

Selected top predictors based on feature importance.

Model Building:

Trained Linear, Radial, and Polynomial SVMs.

Applied class weights to handle class imbalance (Diabetes Yes/No).

Tuned hyperparameters (Cost, Gamma, Degree).

Validation:

70% train / 30% test split.

Metrics: Accuracy, Precision, Recall, F1 Score.

Tuned Polynomial SVM (Mini Dataset)

Methodology:

A balanced mini dataset of 1000 samples was created (500 Yes, 500 No).

Top 6 predictive features were selected based on prior analysis.

Predictors were standardized using centering and scaling.

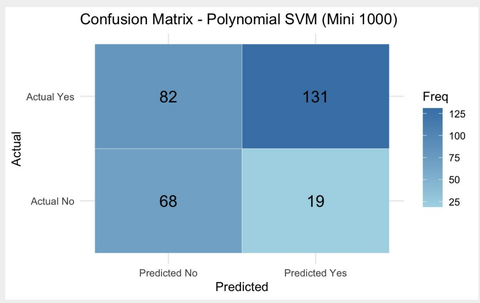
Tuning Parameters:

Kernel: Polynomial (degrees 2 and 3)

Cost values tested: 0.1, 1, 10

Class weights: No = 1, Yes = 1 (due to balanced sampling)

Confusion Matrix Summary:



The model correctly identified most positive cases, resulting in high recall.

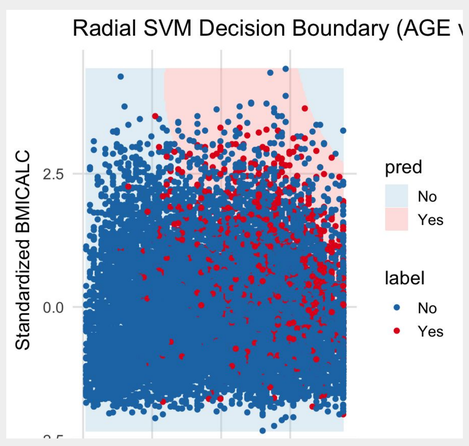
False positives were more frequent, reducing specificity. The confusion matrix confirmed the model's emphasis on detecting diabetes cases effectively.

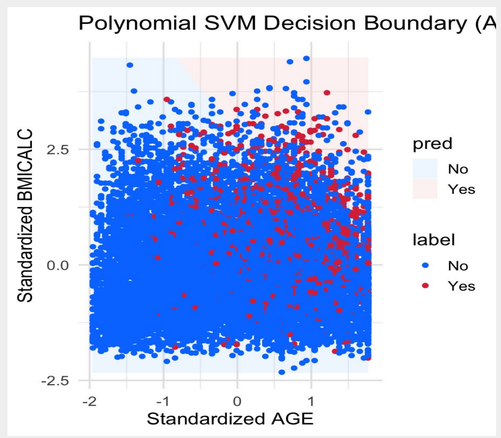
Results

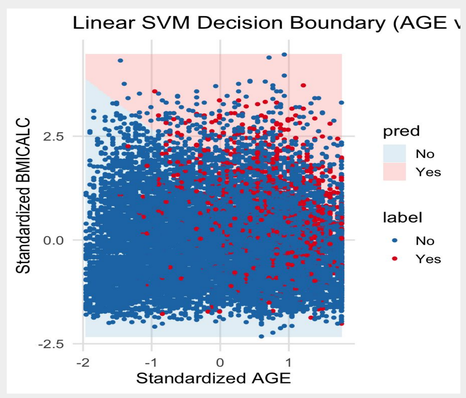
The radial SVM gave the highest accuracy among untuned models but struggled with recall. The untuned polynomial SVM performed poorly despite high accuracy. In contrast, the tuned polynomial SVM (on a balanced mini dataset) achieved the best overall balance, with strong recall and F1-score. Class weights helped, but tuning and dataset balancing made the biggest impact.

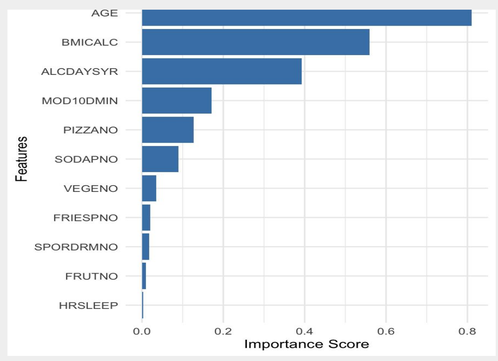
Model	Accuracy	Precision	Recall	F1 Score
Linear SVM	0.79	0.25	0.57	0.35
Radial SVM	0.80	0.26	0.50	0.35
Polynomial SVM	0.85	0.50	0.003	0.006
Tuned Poly SVM MINI	0.63	0.61	0.87	0.72

Plots









Discussion & Conclusion

Class-weighted SVMs helped improve the recall and F1-score, especially in detecting minority cases, but results still indicated limitations. Despite tuning and weighting, some minority class examples remained misclassified — likely due to overlapping features and limited data points. The RBF kernel captured non-linear relationships better than linear or polynomial models, yet even it struggled with recall, showing that class imbalance was not fully addressed. This highlights how SVMs, while powerful, require careful tuning and sometimes benefit from additional techniques like data resampling or ensemble models.

No matter which model we used, age, BMI, and alcohol consumption kept showing up as the strongest predictors of diabetes. This points to a clear takeaway: if we want to reduce risk early on, screening efforts and lifestyle support should especially focus on older individuals who have higher BMI and drink regularly.While nonlinear models like RBF provided more flexibility, the marginal performance gain reinforces the value of simple models — when tuned well and applied thoughtfully. Going forward, incorporating more balanced training data and exploring other algorithms could further enhance model reliability for real-world health applications.

References

- ChatGPT (2025). images created using OpenAI's DALL-E.
- U.S. National Health Interview Survey (NHIS) 2022. Dataset Access.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). Springer.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- American Diabetes Association. (2022). Standards of Medical Care in Diabetes—2022. *Diabetes Care*, 45(Supplement_1), S1–S264.