Predicting Diabetes Using Support Vector Machines

(Data Source: National Health Interview Survey, IPUMS)

INTRODUCTION

- We predict diabetes occurrence using various health, demographic, and lifestyle features from the NHIS survey.
- Support Vector Machine (SVM) models with linear, RBF, and polynomial kernels are compared to evaluate performance.

THEORETICAL BACKGROUND

- A Support Vector Machine (SVM) is a supervised machine learning algorithm that finds an optimal hyperplane to classify data points into distinct classes.
- The SVM aims to maximize the margin — the distance between the hyperplane and the nearest data points (support vectors).

Key equations:

i. $f(x) = w^T x + b$

ii. $\min_{w,b} \frac{1}{2} || w ||^2$

iii. $\min_{w,b,\xi} \frac{1}{2} || w ||^2 + C \sum_{i=1}^{n} \xi_i$

Key tuning parameters:

- C: balances margin size vs classification error.
- Kernel: transforms data into higher dimensions (Linear, RBF, Polynomial).
- Gamma: controls curvature for RBF/Polynomial.
- Degree: specifies the degree of the polynomial function.

METHODOLOGY

- We selected different numeric features covering demographics, health metrics, and lifestyle behaviors.
- Split the data into 70% training and 30% testing sets.
- Three SVM models were trained using default and tuned hyperparameters where needed.
- Model performances were evaluated based on accuracy.
- A toy dataset was also used to visualize a sample SVM decision boundary with an RBF kernel.

RESULTS

	Model	Accuracy
0	Linear SVM	0.914191
1	Radial SVM (RBF)	0.914191
2	Polynomial SVM	0.913621

FIG 1: Computational Results

NOTE: All models performed similarly, suggesting that the predictors provide a strong linear separability.

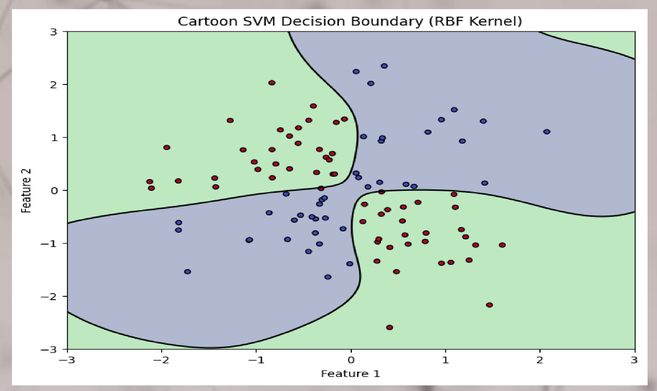

FIG 2: Cartoon SVM Plot

SVM creating a linear decision boundary between two classes.

DISCUSSION

- Key predictors for diabetes included age, BMI category, sleep duration, work hours, and physical activity levels.
- Older age, higher BMI, lower physical activity, and fewer sleep hours were associated with higher diabetes risk.
- Since all SVM models performed similarly with high accuracy, this suggests that these basic health and lifestyle indicators are powerful predictors.

CONCLUSION

- The results demonstrate that diabetes prediction can be effectively performed using readily available demographic, behavioral, and lifestyle variables.
- With all the SVM models achieving over 91% accuracy, the findings suggest that a simple linear classifier can be sufficient for real-world applications.
- These insights could be valuable for healthcare professionals and policymakers looking to design low-cost, scalable early screening tools based on basic health information.
- Efforts to monitor and promote improvements in physical activity, diet, sleep patterns, and work-life balance could significantly contribute to reducing diabetes risk at a population level.

REFERENCES

- Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Daniel Backman, Annie Chen, and Stephanie Richards. *IPUMS Health Surveys: NHIS, Version 7.4* [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D070.V7.4