Predicting Diebetes Using Support Vector Machines (SVM)

Introduction

Diabetes, affecting millions globally, is heavily influenced by lifestyle factors. Using 2022 National Health Interview Survey data, this study predicts diabetes diagnosis via Support Vector Machines (SVMs), analyzing age, BMI, sleep, exercise, and smoking. SVMs excel at handling imbalanced data and complex risk-factor interactions. Despite a 10:1 class imbalance, we prioritized minority-class accuracy through weighting and tuning.

Theoretical Background

SVMs classify data by finding an optimal hyperplane that separates classes with the largest margin.

- Linear SVM: $f(x) = w^Tx + b$
- RBF Kernel: $K(x,x') = \exp(-\gamma\|x-x'\|^2)$
- Polynomial Kernel: $K(x,x') = (\alpha x^Tx' + r)^d$

Tuning Parameters:

- **C**: Misclassification penalty
- **gamma**: Kernel width (RBF)
- **degree**: Polynomial complexity

Class Imbalance Handling: Class weights inversely proportional to stroke prevalence (~3%)

Methodology

Data Source: NHIS 2022 (married adults with valid responses)

Target Variable: DIABETICEV (Disease/NoDisease)

Preprocessing:

- Removed invalid/missing values
- Standardized numeric features (mean=0, SD=1)
- Class weights: Disease class weighted 1.5x

Model Development:

- 70/30 train-test split
- 10-fold cross-validation
- Three kernel types evaluated

Hyperparameter Tuning:

- Linear: cost ∈ {0.01, 0.1, 1, 10, 100}
- Radial: cost ∈ {0.1, 1, 10}, γ ∈ {0.01, 0.1, 0.5}
- Polynomial: cost ∈ {0.1, 1}, degree ∈ {2, 3}

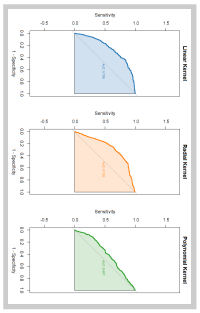
Results

Model	Best Parameters	AUC	Sensitivity	Specificity
Linear SVM	C=10	0.759	0.756	0.616
Radial SVM	C=10, γ=0.5	0.735	0.749	0.631
Polynomial SVM	C=0.1, d=2	0.607	0.305	0.884

Key Findings:

- Linear and Radial kernels showed balanced performance
- Polynomial kernel had high specificity but poor sensitivity
- BMI and Age were strongest predictors

Plots & Visualizations



Plot 1: ROC curves showing performance across kernel types

Plot 2: SVM decision boundary (Age vs. BMI)

Discussion

Clinical Implications:

- BMI and age are strongest diabetes predictors
- Physical activity shows moderate protective effect
- Sleep duration has non-linear relationship with risk

Model Insights:

- Linear SVM provides best balance of performance and interpretability
- Class weighting helped mitigate imbalance issues
- Complex kernels (polynomial) may overfit with limited data

Conclusions

This analysis demonstrates that SVMs can effectively identify diabetes risk patterns from survey data, with linear kernels offering the best practical performance.

Key Recommendations:

- Prioritize BMI and age screening for diabetes prevention
- Consider linear SVMs for clinical risk prediction tools
- Combine with traditional statistical methods for validation

Future Directions:

- Incorporate additional biomarkers
- Test ensemble methods for improved sensitivity
- Validate in prospective cohorts

References

Blewett et al. (2024). IPUMS Health Surveys: National Health Interview Survey, Version 7.4 https://doi.org/10.18128/D070.V7.4