PREDICTING ADULTS STROKE RISK USING SUPPORT VECTOR MACHINES

BY ALEKHYA DABBBIRU

INTRODUCTION

This study uses Support Vector Machines to predict disease presence based on demographics and health behaviors. Using data from the 2022 National Health Interview Survey (NHIS), with 48 variables on age, sex, sleep patterns, alcohol and cigarette use, and physical activity from over 35,000 adults, SVMs like linear, polynomial, and radial classifiers to model disease risk and assess their effectiveness in capturing complex health outcomes.

TECHNICAL BACKGROUND

Support Vector Classifiers find an optimal hyperplane that maximizes the margin between two classes.

The margin is the distance between the boundary and the nearest points.

Kernels measure similarity between observations and come in three types:

- Linear (cost)
- Polynomial (degree, cost)
- Radial (gamma, cost**)**

$$y_i\left(\beta_0+\beta_1X_1+\beta_2X_2\right)\ge M(1-\epsilon_i)$$

This equation shows that Support Vector Machines maximize the margin while allowing some classification errors, controlled by slack variables. The cost parameter C balances margin size and errors, where a larger C penalizes errors more heavily and results in a narrower margin.

Advantages: Handles nonlinear patterns, tolerates outliers, performs well in high dimensions.

Disadvantages: Slower training, noise sensitivity, requires careful tuning.

Applications: Face recognition, geospatial analysis, disease prediction, and medical diagnostics.

Performance Metrics: Accuracy, precision, recall, F1-score, confusion matrix. Interpretation is easier for linear SVCs through feature coefficients but less direct for nonlinear models.

METHODOLOGY

DATA CLEANING: The data was cleaned by converting coded missing values to NA and dropping variables with high missingness. Categorical variables were factorized, and continuous predictors were scaled to ensure uniformity.

TRAIN-TEST SPLIT:A random 70% of the data was used for training and 30% for testing.

CLASS IMBALANCE: To address class imbalance, class weights were applied to give more importance to positive cases during model training.

MODEL IMPLEMENTATION: Three SVM models were implemented: linear (cost tuned), polynomial (degree and cost tuned), and radial (gamma and cost tuned)

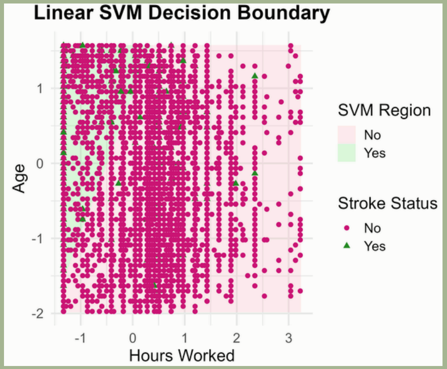
HYPER PARAMETER TUNING: Linear SVM: Tuned cost (0.01, 0.1, 1); Polynomial SVM: Tuned cost, degree (3, 4), coef0 (0.5, 1); Radial SVM: Tuned cost, gamma (0.01, 0.1);Best models were selected based on validation accuracy.

MODEL EVALUATION: models were evaluated using accuracy, precision, recall, F1-score, and confusion matrices.

RESULTS

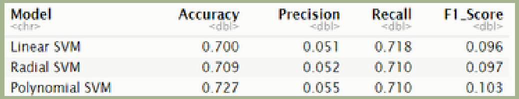
The linear SVM achieved the highest recall but suffered from low precision, resulting in many false positives. The polynomial SVM improved overall accuracy but still had moderate precision and recall.

The radial SVM produced the highest accuracy but struggled with identifying positive stroke cases.



The above plot demonstrates the decision boundary created by a Linear SVM model (using real data) on Hours Worked and Age, highlighting that while a boundary exists, substantial overlap poses challenges in accurately separating stroke and non-stroke cases.

The table below summarizes SVM models metrics, showing that although Polynomial SVM achieved the highest accuracy, all models struggled with precision in detecting stroke cases.

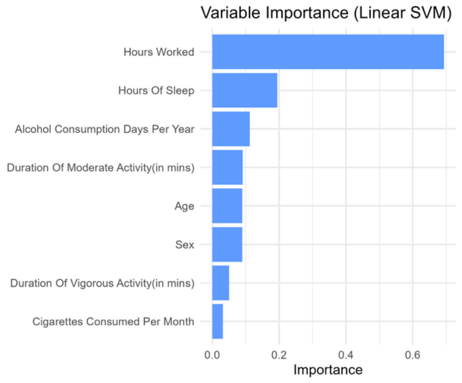


Model <chr>	Accuracy <dbl>	Precision <dbl>	Recall <dbl>	F1_Score <dbl>
Linear SVM	0.700	0.051	0.718	0.096
Radial SVM	0.709	0.052	0.710	0.097
Polynomial SVM	0.727	0.055	0.710	0.103

DISCUSSION

'Age', 'Hours Worked', and 'Hours of Sleep' emerged as strong predictors of stroke risk, but the SVM decision boundary showed that these factors alone could not clearly separate stroke from non-stroke cases, highlighting the complexity of disease prediction.

‘Age’ was a key demographic factor, aligning with the increased stroke risk observed with aging. Variable importance from the Linear SVM highlighted 'Hours Worked', 'Hours of Sleep', and 'Alcohol Consumption Days Per Year' as influential predictors. Based on these findings, policy-makers should promote healthier sleep, balanced work schedules, and early screenings for older adults.



CONCLUSION

'Age', 'Hours Worked', and 'Hours of Sleep' were key predictors of stroke risk. Although SVM models struggled to perfectly classify stroke and non-stroke cases, the results highlight the need for healthier sleep habits, balanced work schedules, and early screening to reduce stroke incidence.

BIBLIOGRAPHY

- Blewett, L. A., Rivera Drew, J. A., King, M. L., Williams, K. C. W., Backman, D., Chen, A., & Richards, S. (2024). IPUMS Health Surveys: National Health Interview Survey, Version 7.4 [dataset]. Minneapolis, MN: IPUMS.
- Dr. Mendible. (2025). Support Vector Machines (Combined) [Lecture Slides]. Seattle University.