Drawing the line: Predicting Cancer from Lifestyle Factors Using SVMs

Eloho Okoloko
College of Science and Engineering, Seattle University
Professor: Ariana Mendible, Ph.D

Introduction

This project investigates how health behaviors and demographics influence the presence of cancer. Using data from the 2022 National Health Interview Survey (NHIS), we focus on predicting cancer diagnoses based on five predictors: hours of sleep, physical activity, body mass index (BMI), work hours, and smoking status. Support Vector Machines (SVMs) with different kernels (linear, radial, polynomial) are implemented and compared to identify the most effective modeling approach.

Theoretical Background

Support Vector Machines are supervised learning models that construct hyperplanes to separate data classes with maximal margins. In cases where the data is not linearly separable, kernel functions project the features into higher-dimensional spaces. The models we consider include:

•**Linear Kernel**: Separates data with a straight hyperplane.
•**Radial Basis Function (RBF) Kernel**: Uses similarity scores between points for flexible boundaries.
•**Polynomial Kernel**: Captures curved decision boundaries based on polynomial transformations.
Key Parameters:
•**Cost (C)**: Controls tradeoff between margin width and classification error.
•**Gamma (γ)**: In RBF, controls the reach of a single training point.
•**Degree**: In polynomial kernels, controls the complexity of the transformation

Methodology

•**Dataset**: Adults aged 18+ from NHIS 2022 (n ≈ 24,000).
•**Target Variable**: CANCEREV — whether the respondent was ever diagnosed with cancer.
•**Predictors**:
•HRSLEEP - hours of sleep
•VIG10DMIN -daily minutes of vigorous physical activity
•BMICALC -calculated BMI
•HOURSWRK - weekly hours worked
•VEGENO -frequency of vegetable consumption
•**Preprocessing**:
•Removed invalid responses (e.g., 996–999).
•Standardized predictors using z-scores.
•Train/test split (80/20).
•Set "Yes" as the positive class in the factorization of CANCEREV.
•Applied class.weights = c(\"No\"=1, \"Yes\"=5) to handle imbalance.

Results

Model	Parameters	Test Accuracy	Balanced Accuracy	Sensitivity (Recall)	Precision
Linear SVM	Cost = 1 (weighted)	**58.8%**	**65.0%**	**73.3%**	**19.6%**
Radial SVM (tuned)	Cost = 1, Gamma tuned (~0.01)	**87.4%**	**50.1%**	**0.33%**	**50.0%**
Polynomial SVM	Cost = 1, Degree = 3 (no tuning)	**78.6%**	**57.9%**	**30.3%**	**23.2%**

Fig 1. Linear SVM achieved the best sensitivity (73%) and highest balanced accuracy (65%) at the cost of lower overall accuracy (58.8%). **Radial SVM**, even after hyperparameter tuning, achieved high overall accuracy (87.4%) but failed to correctly detect positive cancer cases, with sensitivity near zero (~0.3%). **Polynomial SVM**, trained without hyperparameter tuning due to time constraints, balanced trade-offs better, achieving a sensitivity of 30.3% and a balanced accuracy of 57.9%.

Discussion

•The Linear SVM, despite lower overall accuracy, proved more effective in identifying individuals with a history of cancer due to class weighting. This is critical in health applications where false negatives (missed cancer cases) are costly.
•The Radial SVM, while accurate overall, failed to handle the minority class, emphasizing that high accurate accuracy alone is not sufficient for evaluating performance in imbalanced datasets.
•The Polynomial SVM offered a middle ground, catching more positives than Radial SVM, while maintaining reasonable overall performance.
•One limitation is that even with class weighting, sensitivity remained a challenge for some models, suggesting that more advanced techniques could improve performance.

Conclusion

Support Vector Machines, particularly when combined with class weighting, can meaningfully model cancer risk based on health behaviors. However, high overall accuracy does not guarantee success at detecting critical minority outcomes like cancer diagnosis. All three models performed poorly and can be improved.

References

•James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

•Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Daniel Backman, Annie Chen, and Stephanie Richards. IPUMS Health Surveys: National Health Interview Survey, Version 7.4 [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D070.V7.4. Links to an external site.http://www.nhis.ipums.orgLinks to an external site..