Predicting Health Risk from Sleep and Work Hours Using Support Vector Machines

Nikhil G. Ghugare, Seattle University
DATA 5322

Background & Objective

Maintaining good health is heavily influenced by behavioral factors, including sleep patterns and work-life balance.
Poor sleep and excessive work hours have been linked to a variety of negative health outcomes.

In this study, we aim to predict an individual's health risk using two behavioral variables:
•**HRSLEEP**: Average hours of sleep per night
•**HOURWRK**: Average hours worked per week
Dataset:
Data is sourced from the IPUMS Health Survey, which includes a broad range of demographic and behavioral health indicators.

Research Questions:
•Can behavioral factors like sleep and work hours predict health risk outcomes?
•Which SVM kernel (Linear, Polynomial, or RBF) provides the most effective classification?

Objective:
To model health risk categories using **Support Vector Machines (SVMs)**,
evaluate different kernel types, and recommend strategies based on model insights.

Introduction

Maintaining good health is strongly influenced by behavioral factors such as sleep and work-life balance.
In this study, we aim to predict health risk outcomes based on individuals' self-reported **hours of sleep** (HRSLEEP) and **hours worked per week** (HOURWRK).

Dataset:
Survey data collected from the IPUMS Health Survey database.

Variables used:
•**HRSLEEP**: Average hours of sleep per night
•**HOURWRK**: Average work hours per week

Goal:
To model and predict health risk categories using **Support Vector Machines (SVMs)**, examining the performance of different kernel types (Linear, Polynomial, and Radial Basis Function (RBF)).

Theoretical Background

Support Vector Machines (SVMs):
SVMs find the hyperplane that best separates classes by maximizing the margin.

Kernels:
•**Linear Kernel**: Best for data separable by a straight line.
•**Polynomial Kernel**: Captures complex, polynomial relationships.
•**RBF Kernel**: Captures highly nonlinear patterns using Gaussian functions.

Key Concepts:
•**Soft Margin SVM**: Allows some misclassifications to better generalize.

•**Tuning Parameters**:
- **C**: Controls margin versus classification error.
- **Gamma**: Controls boundary curvature (for RBF).
- **Degree**: Sets complexity (for Polynomial).

Interpretation:
Large margins with fewer support vectors suggest better generalization; kernel choice requires careful tuning to avoid overfitting.

Methodology

Preprocessing:
- Standardized features
- Addressed class imbalance using SMOTE (Synthetic Minority Oversampling)

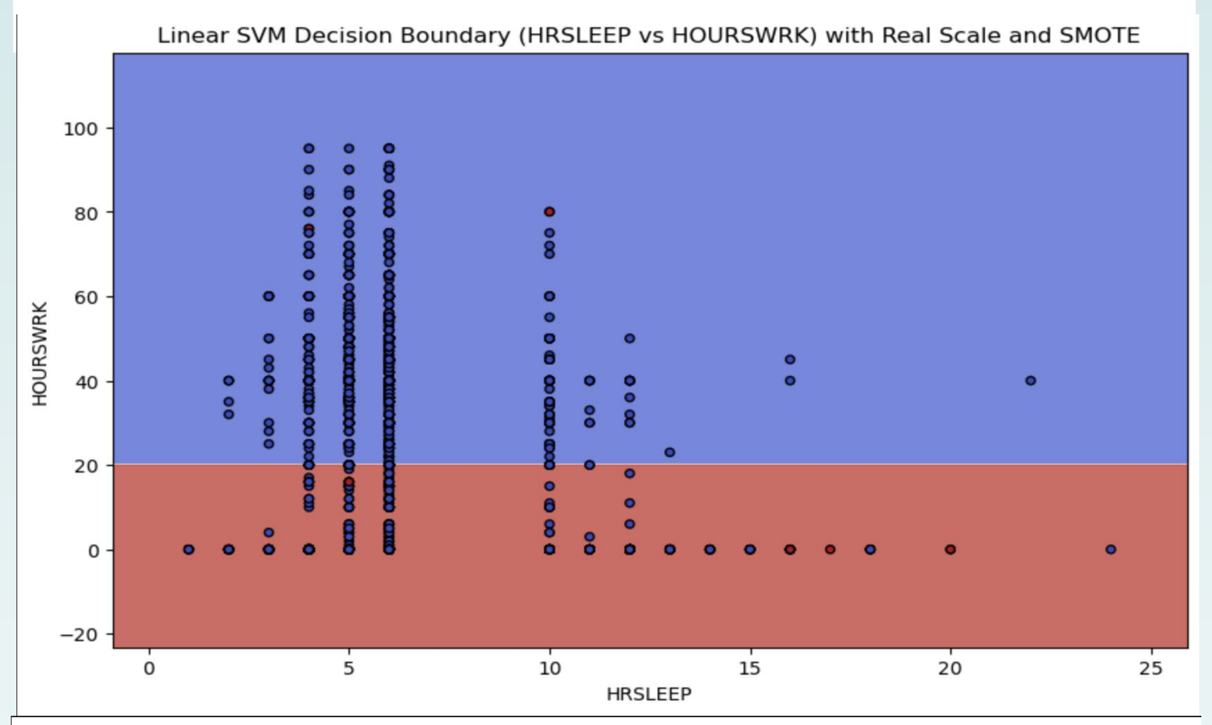
Modeling:
- SVMs trained with three kernels: Linear, Polynomial (degree 5), RBF
- Hyperparameter tuning for:
  - C (margin penalty)
  - degree (for polynomial kernel)
  - gamma (for RBF kernel)

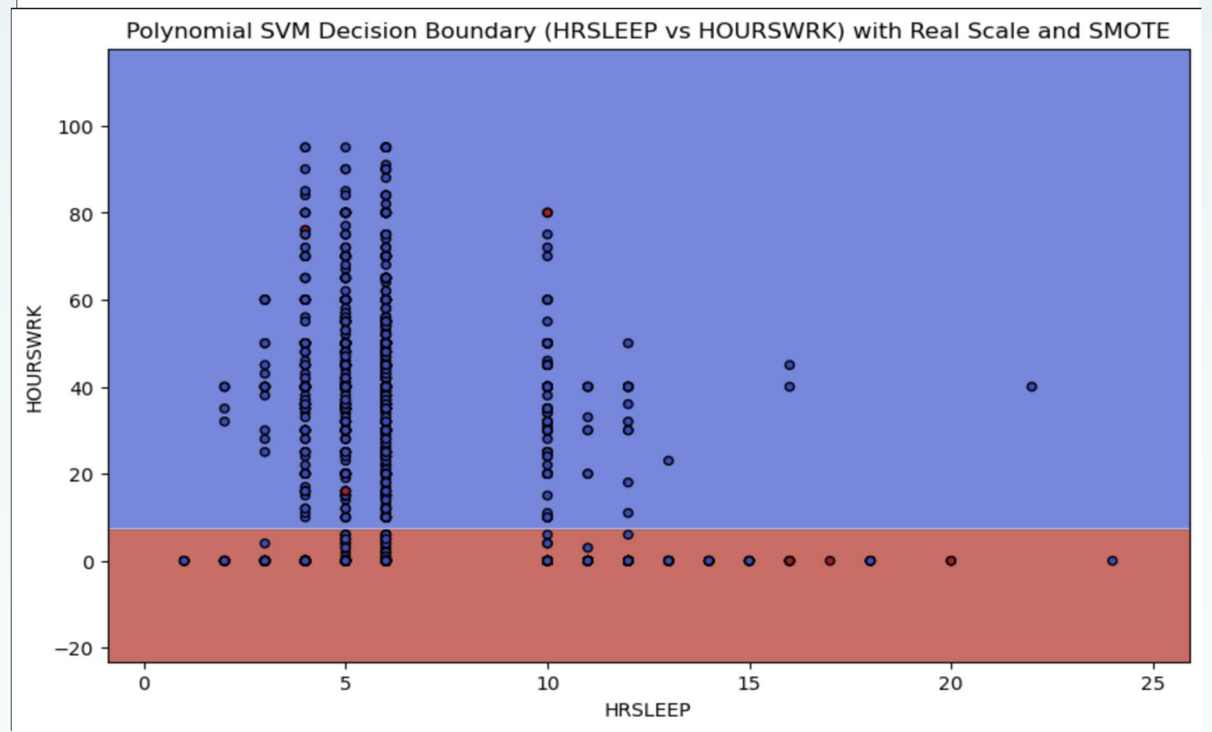
Evaluation Metrics:
- Accuracy
- Recall
- Precision
- AUC (Area Under Curve)

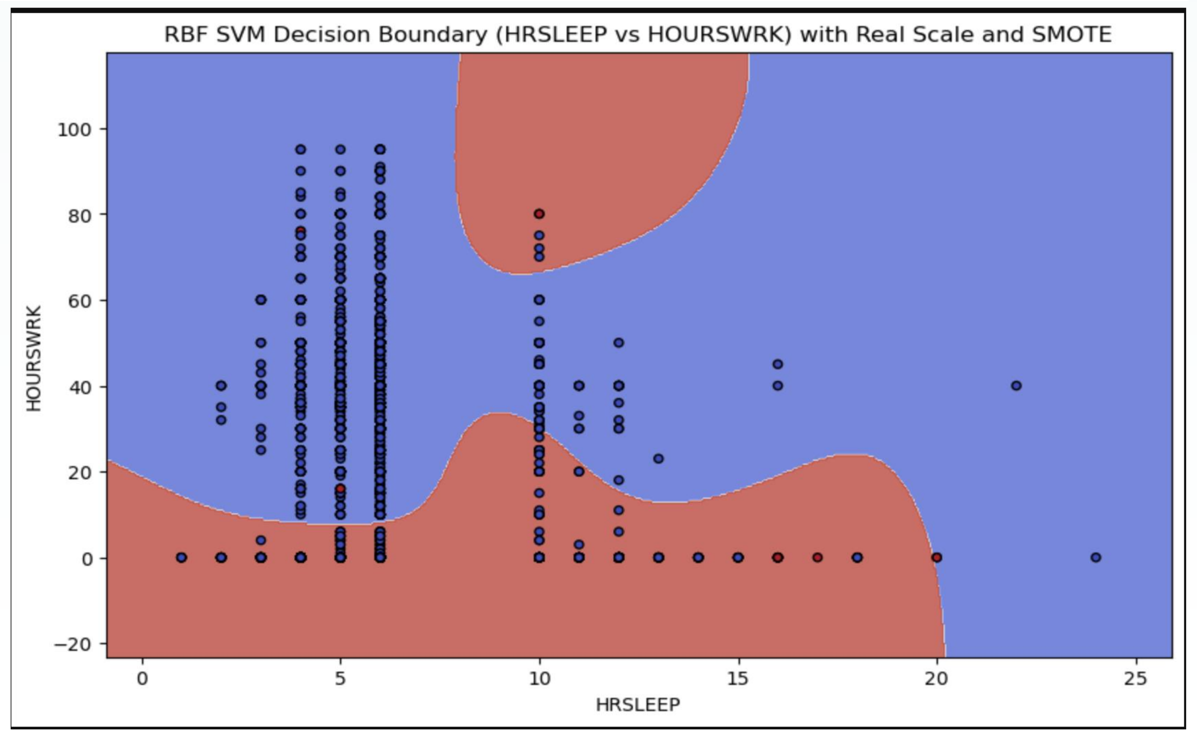
Results

Key Performance Table:

Metric	Linear SVM	Polynomial SVM	RBF SVM
Accuracy	68.0	65.0	68.0
Recall	84.0	89.0	84.0
Precision	16.0	15.0	16.0
F1 Score	0.26	0.26	0.26







Interpretation & Discussion

Findings:
•Linear and RBF SVM achieved high recall (84%), important for identifying high-risk individuals.
•Polynomial SVM had the highest recall (89%) but lower precision.
•Accuracy was moderate (65–68%), indicating a reasonable but improvable fit.

Interpretation:
•Sleep and work hours are important but insufficient alone for strong health risk classification.
•Polynomial and RBF kernels captured non-linear effects, though gains were limited with only two features.
•High recall suggests models are effective at flagging at-risk individuals, despite low precision.

Limitations:
•Limited features (HRSLEEP, HOURWRK) restrict model complexity.
•SMOTE improved recall but may introduce synthetic noise.

Potential Improvements:
•Add more variables (BMI, age, lifestyle factors).
•Explore ensemble methods (Random Forests, Boosting) for better accuracy.

Conclusion

Support Vector Machines, particularly with nonlinear kernels, provide effective tools for modeling behavioral health risks.
In contexts where missing a high-risk individual would have serious consequences, maximizing recall becomes critical.

Impact:

Encourages policies promoting better sleep and work-life balance.

Future predictive models could help allocate preventive healthcare resources more effectively.

References

1. 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. & http://www.nhis.ipums.org

2. IPUMS Health Survey Data. University of Minnesota IPUMS USA Project.

3. Blewett LA, Rivera Drew JA, King ML, Williams KCW. IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset].

4. Scikit-learn Developers. Scikit-learn: Machine Learning in Python. https://scikit-learn.org