Jesse Loi

Diabetes Classification with Support Vector Machines

Diabetic

Nondiabetic

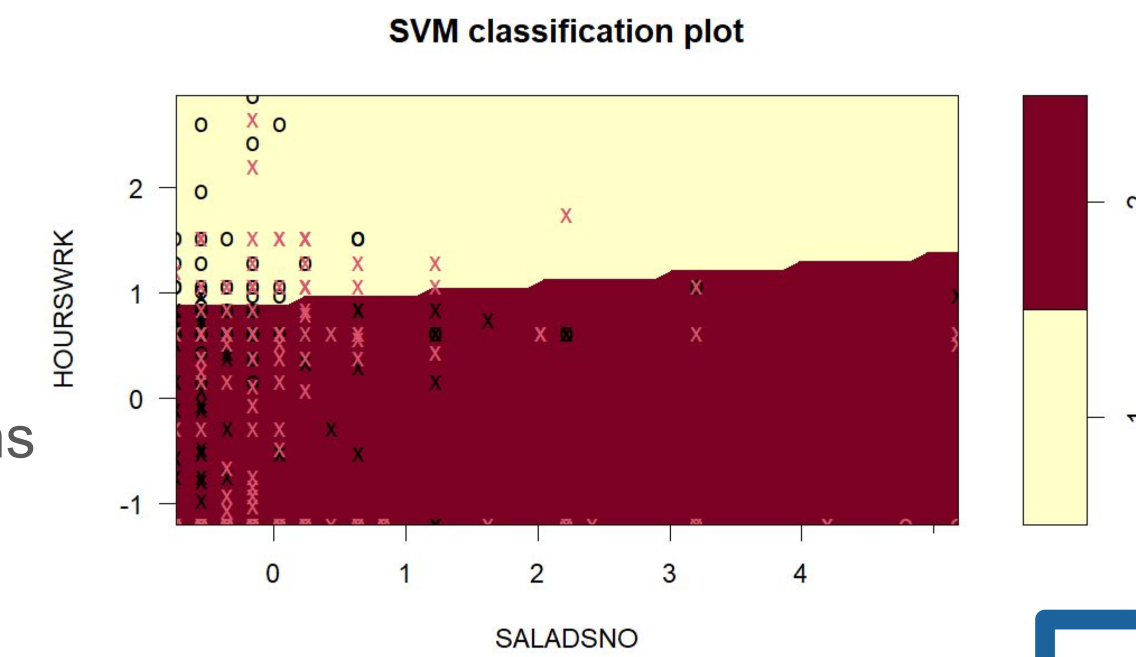
Data Source: IPUMS Health Survey
Goal: Classify instances of Diabetes

IPUMS HEALTH SURVEYS

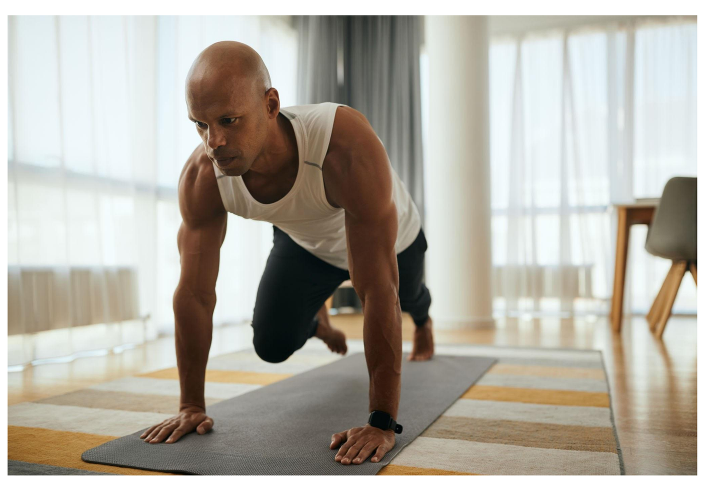
Theoretical Background
Support Vector Machine
- Allows classification by drawing a boundary

Relevant Equation: the distance between the closest points
Parameters
- Cost: How many points can be misclassified

Drawbacks: SVM requires no NA data, leading to difficult applications



Predictors



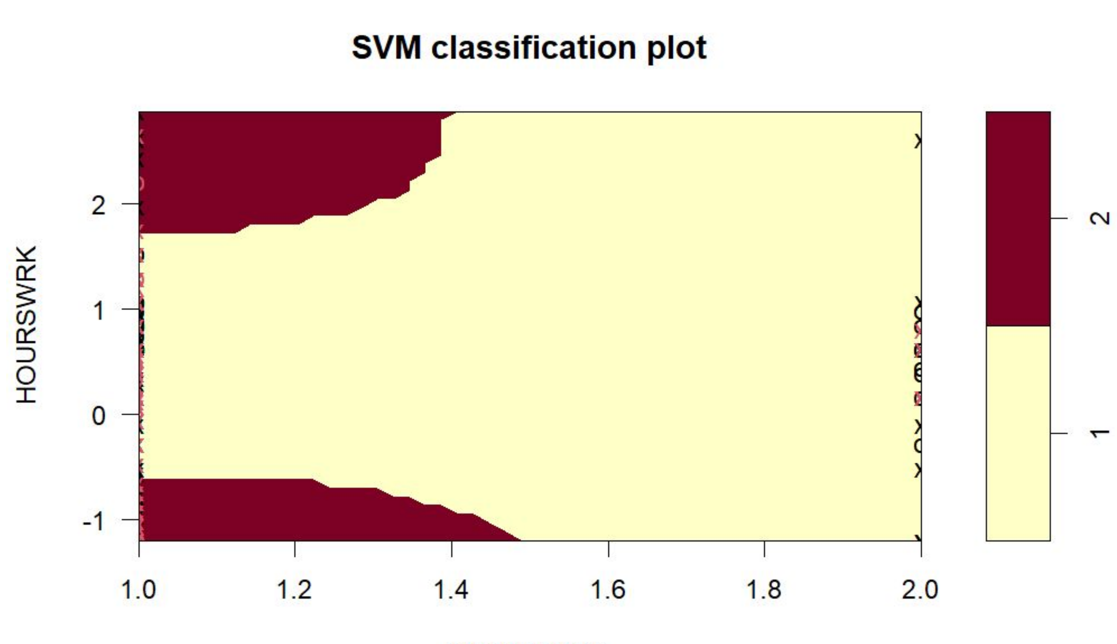
VIG10DMIN: vigorous exercise per day

SALADSNO: salads consumed

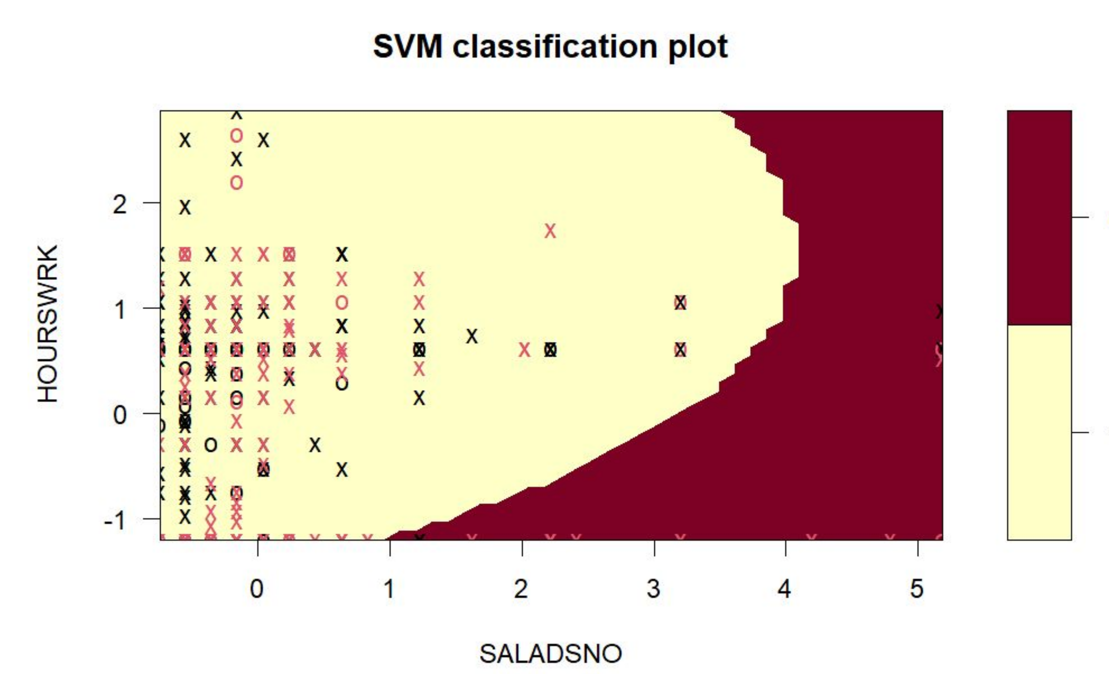
EDUC: education level

HINOTCOVE: health insurance coverage

HOURSWRK: hours worked per week

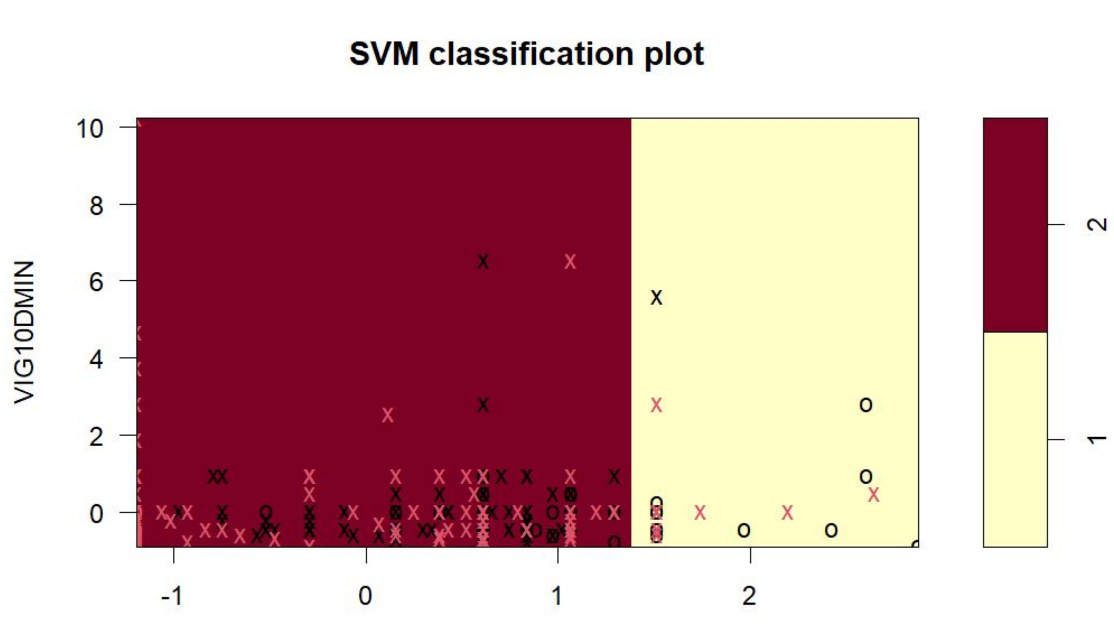


Radial

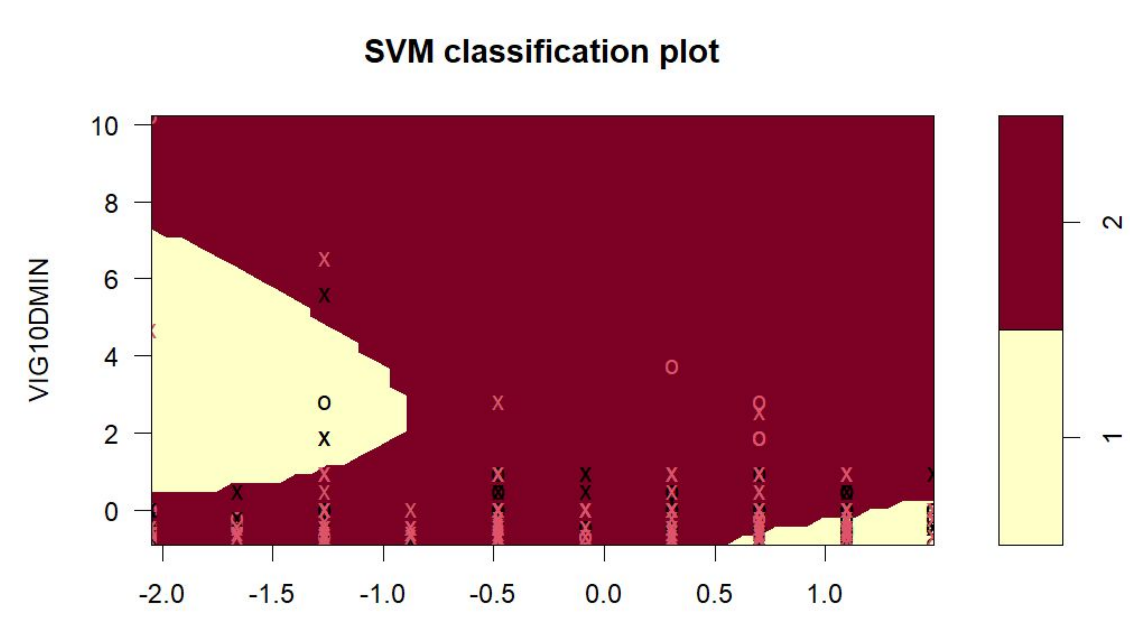


Quadratic

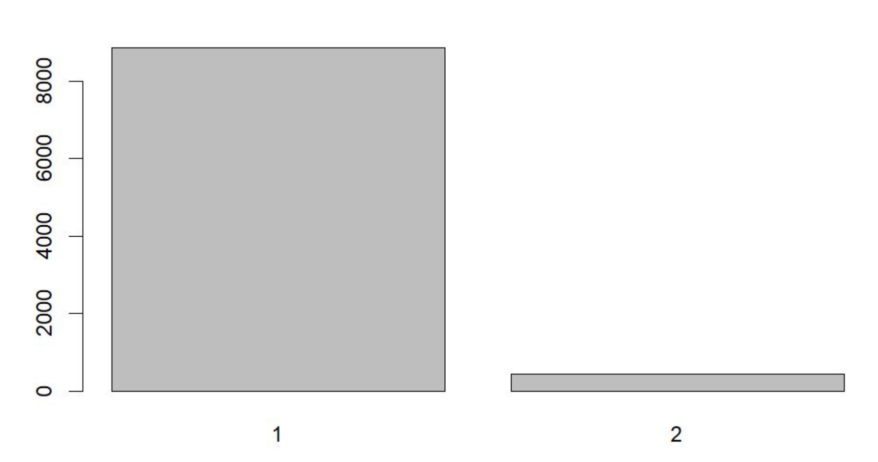
SVM Kernels



Linear

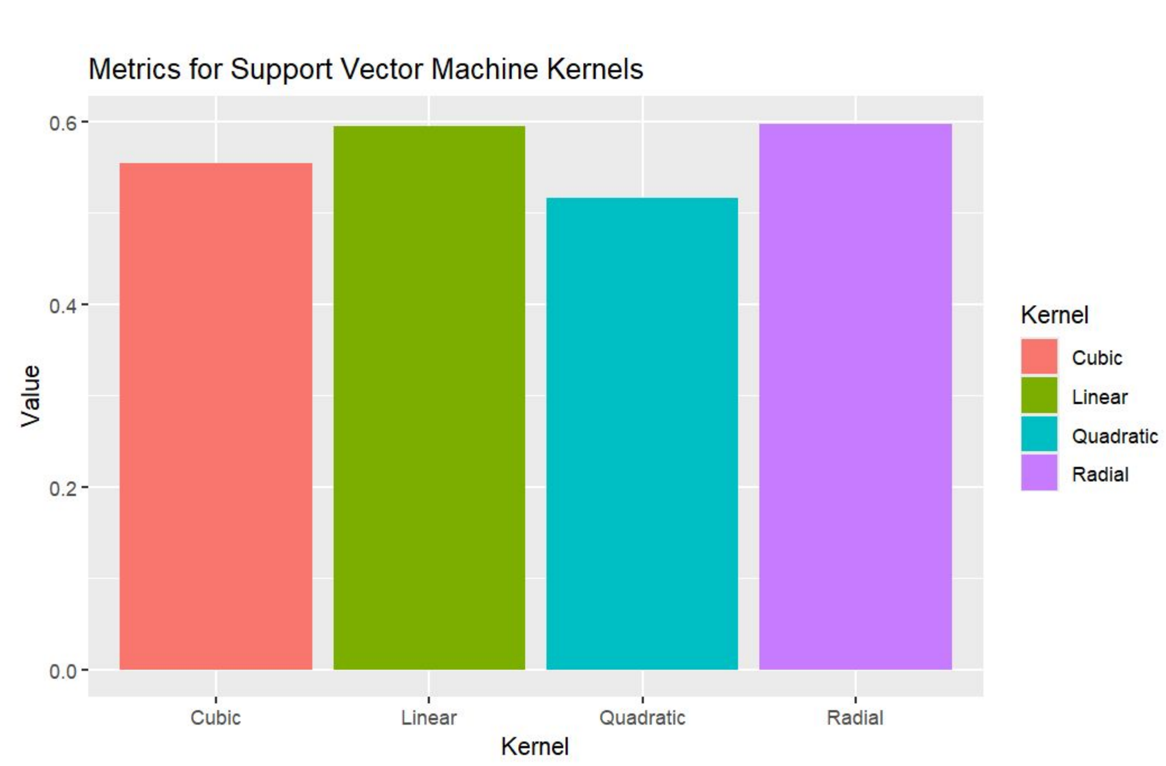


Cubic



Methodology:
- Problem with unbiased data.
  - Solve by decreasing the amount of non-diabetic samples to balance data

Results:
- Intuitively healthy habits seems to be associated with higher rates of diabetes classification
- Models including the 5 above predictors did not have very strong predictions



Conclusion:
High levels of vigorous activity and healthy eating are often a response to diabetes.
Low work hours are also associated with diabetes
- Perhaps diabetes prevents long hours of work
- Perhaps less work entails a more sedentary lifestyle (unlikely)

Likely explanation: diabetes causes the lifestyles, not the other way around

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
Drazen Zigic/Getty Images
rez-art/Getty Images
Jeffrey Hamilton/Getty Images
Dynamic Graphics/Getty Images