Jesse Loi

Diabetes Classification with Support Vector Machines

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

Theoretical Background
Support Vector Machine
- Allows classification by drawing a boundary
- Relevant Equation: the distance between the closest points
  - Determines boundary

Parameters
- <u>Cost</u>: How many points can be misclassified

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

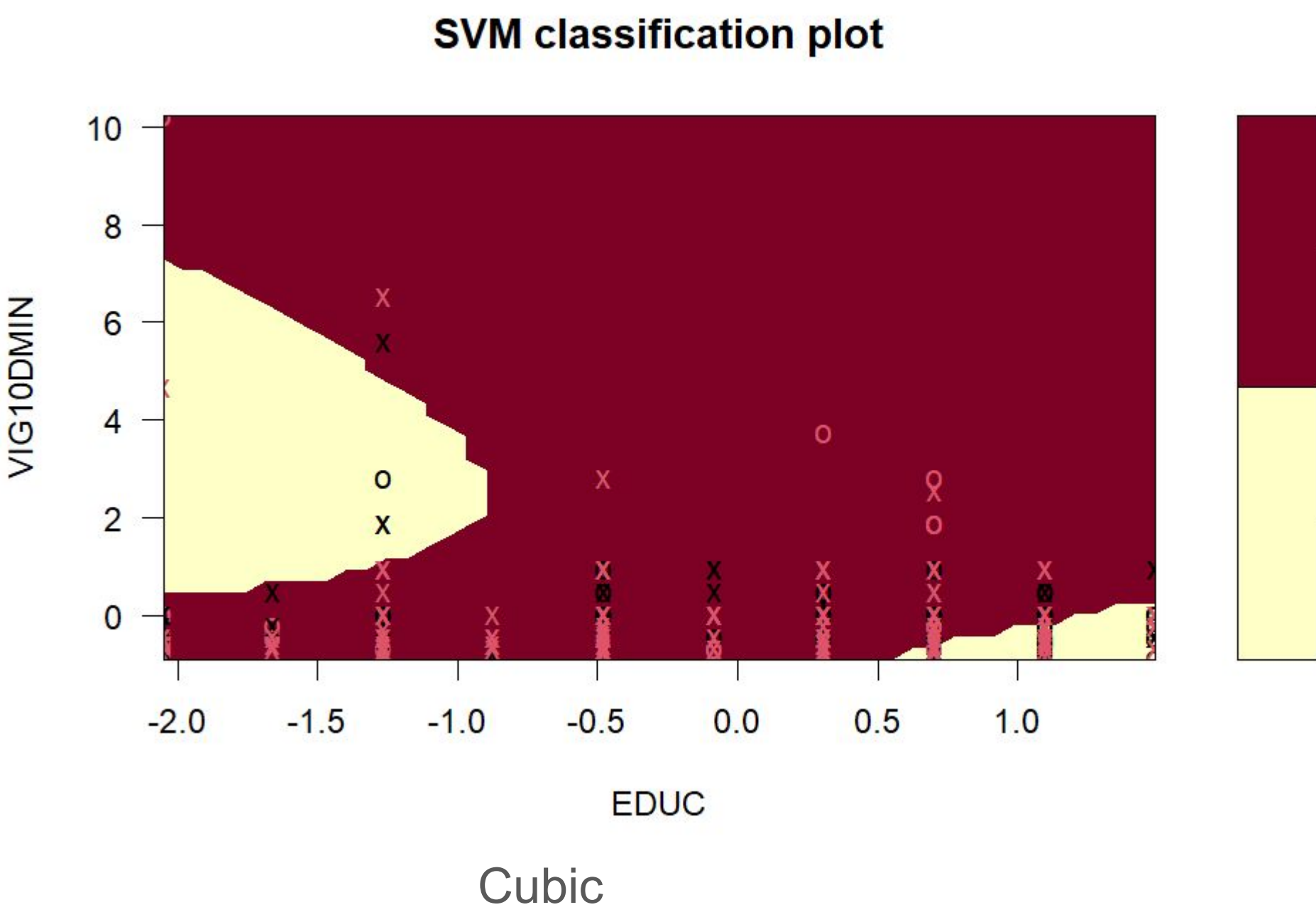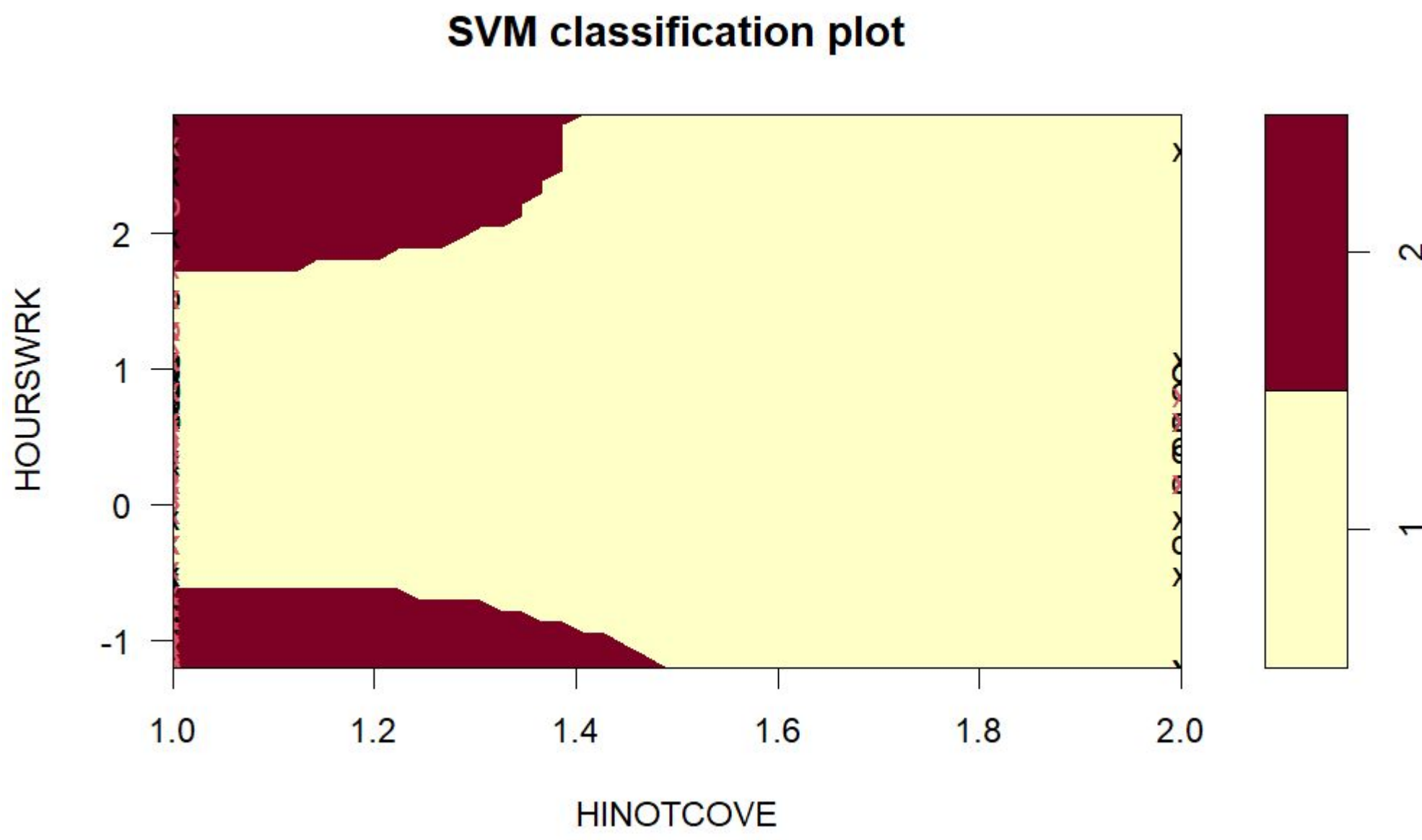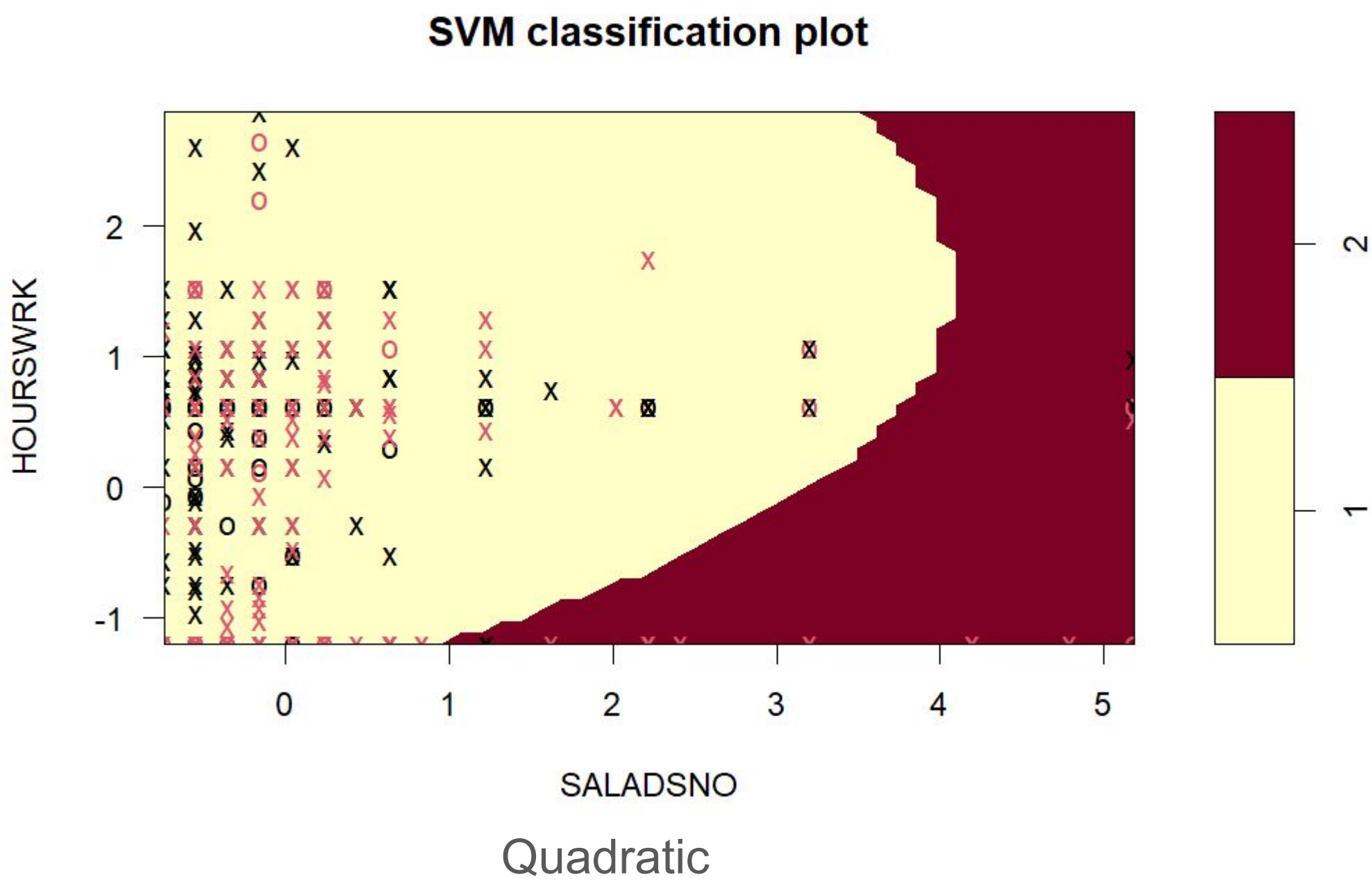## Predictors

VIG10DMIN: vigorous exercise per day

HOURSWRK: hours worked per week

SALADSNO: salads consumed

EDUC: education level

HINOTCOVE: health insurance coverage

Diabetic

Nondiabetic

SVM classification plot (SALADSNO vs HOURSWRK)

SVM classification plot (SALADSNO vs HOURSWRK)
Quadratic

SVM classification plot (EDUC vs VIG10DMIN)
Cubic

SVM classification plot (HINOTCOVE vs HOURSWRK)
Radial

SVM Kernels

**SVM classification plot** (HOURSWRK vs VIG10DMIN)
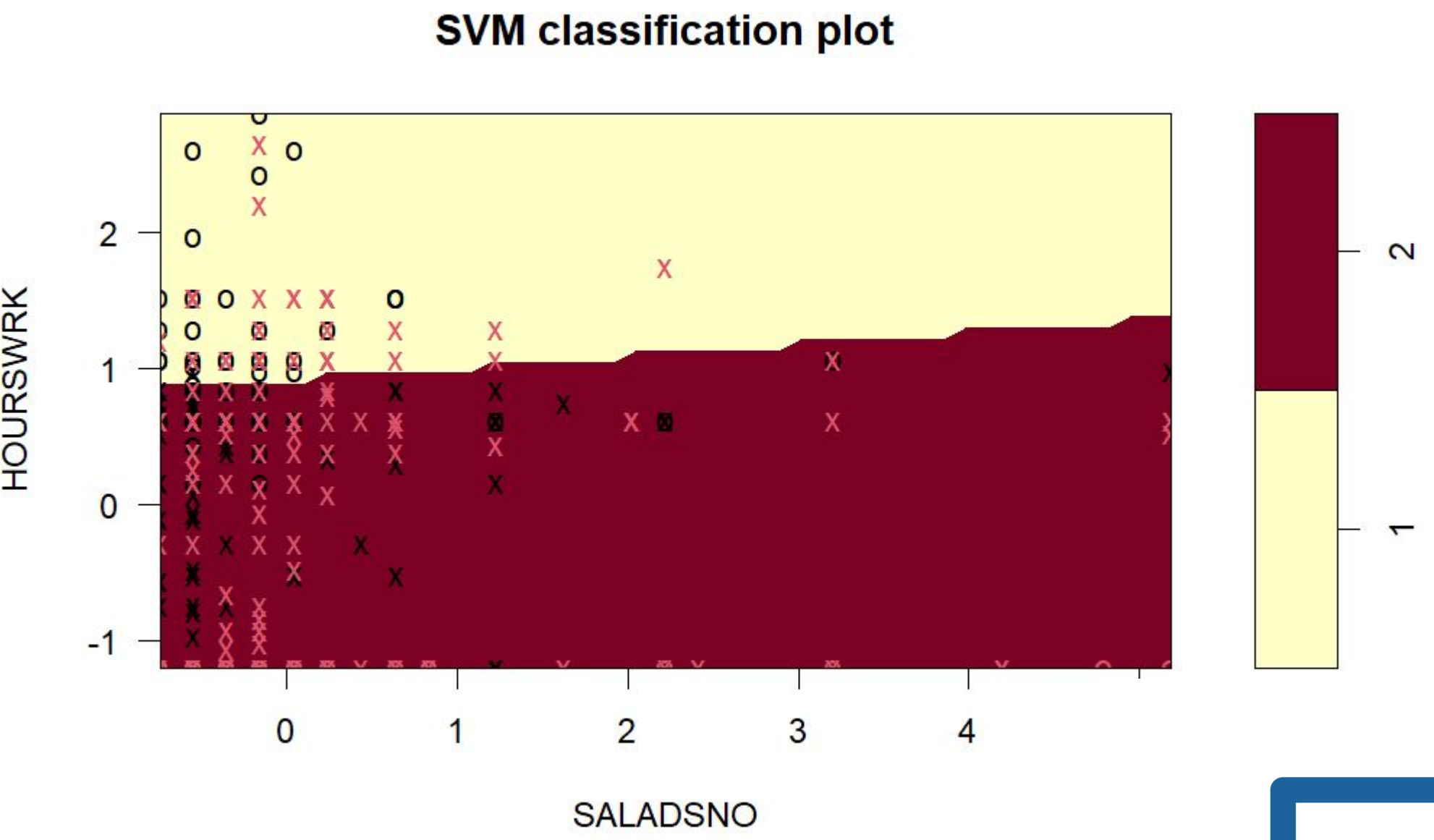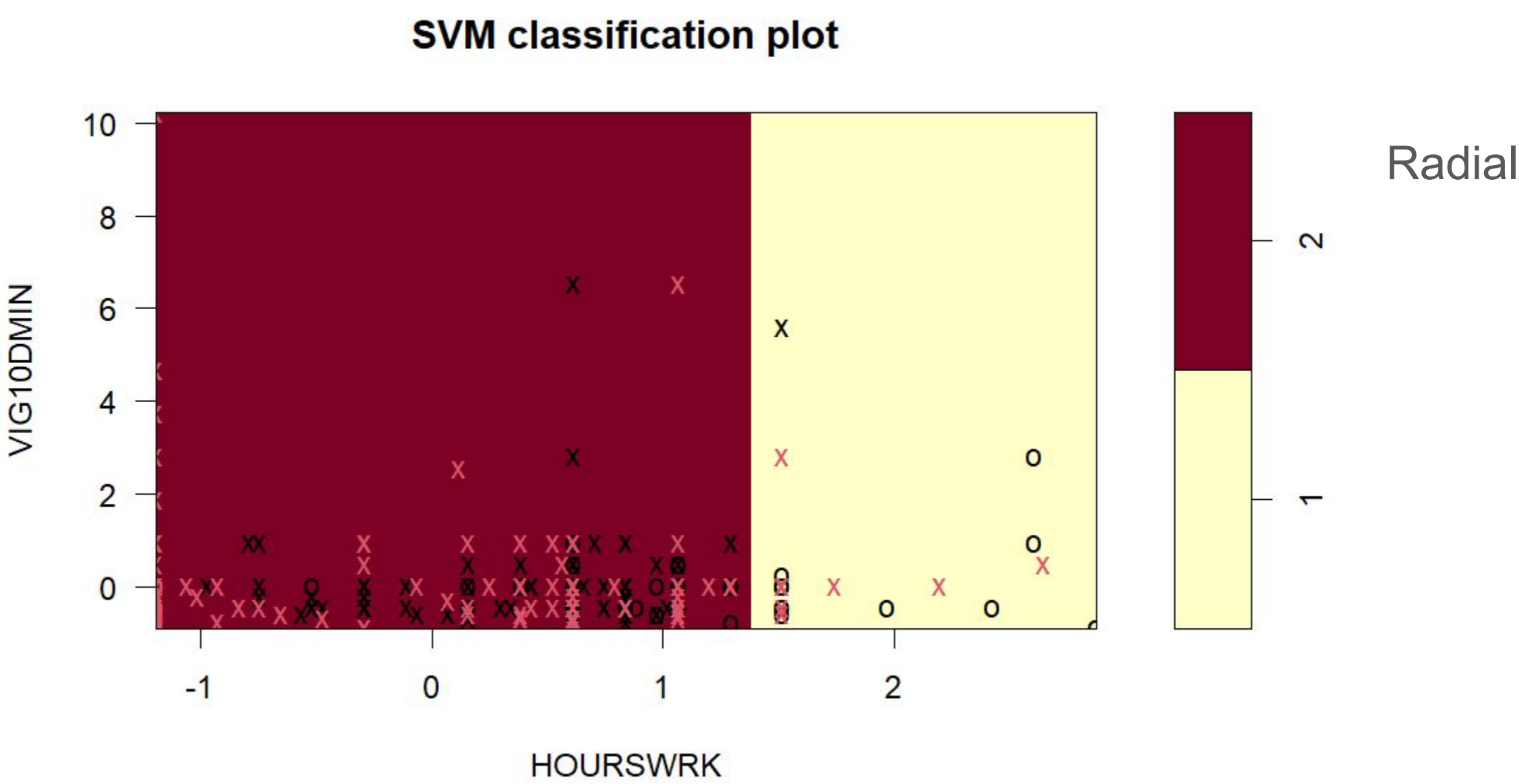
Linear

Methodology:
- Problem with unbalanced 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
  - Question: why do high activity and high salad eating individuals have diabetes
- Models including the 5 above predictors did not have very strong predictions
  - Linear SVMs perform the best, followed by radial, indicating a linear correlation

Diabetes Status (bar chart: 1, 2)
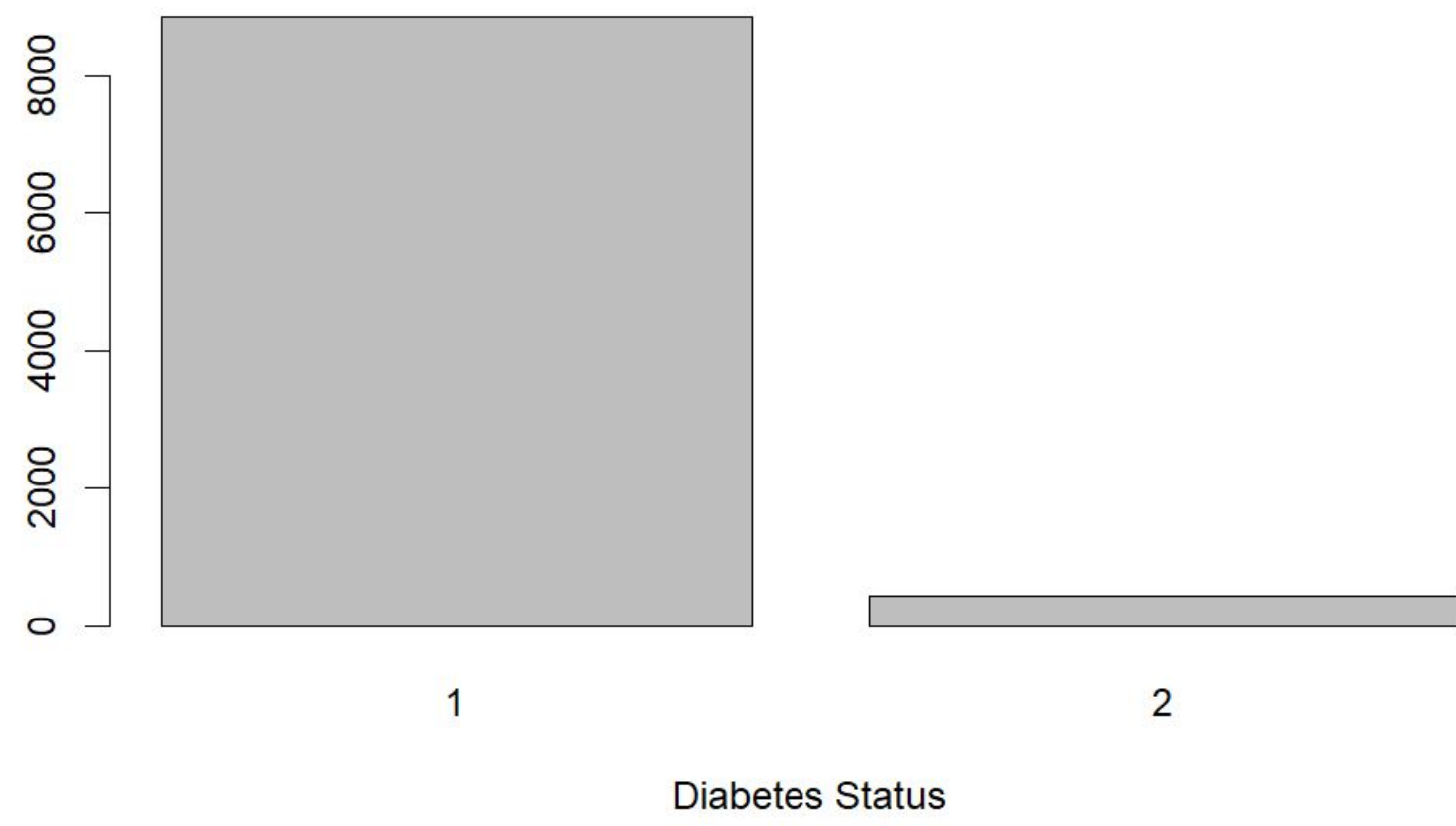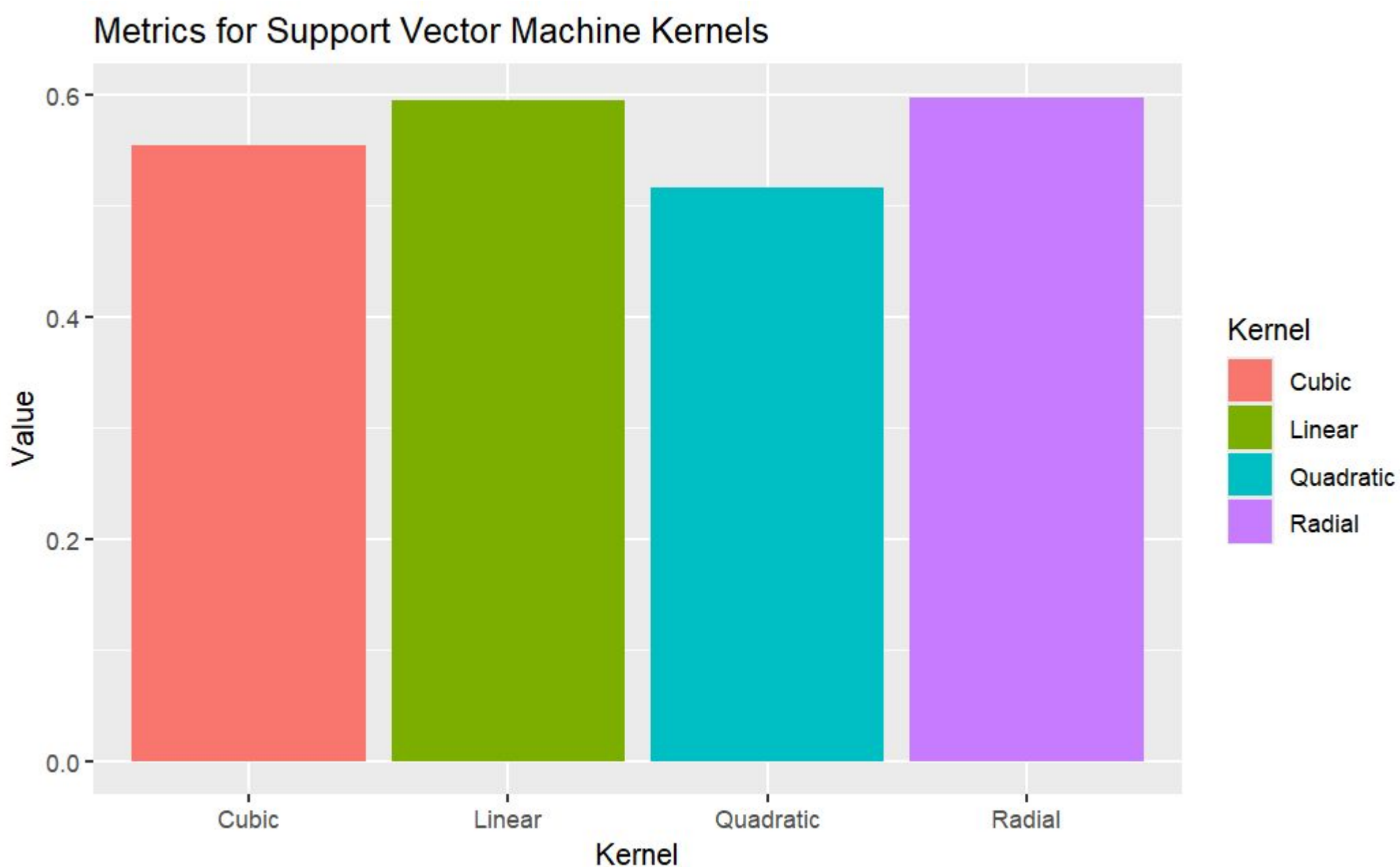
Issues with Shrinking Imbalanced Data:
- Unstable initial conditions with testing data
- Less quality in models
- High variance in data and model predictions

Discussion:
High levels of vigorous activity and healthy eating are often a response to diabetes, though not a cause of 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

Future study should try to attain more data to avoid tricky initial conditions.

Metrics for Support Vector Machine Kernels
(bar chart: Cubic, Linear, Quadratic, Radial)
Kernel: Cubic, Linear, Quadratic, Radial

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