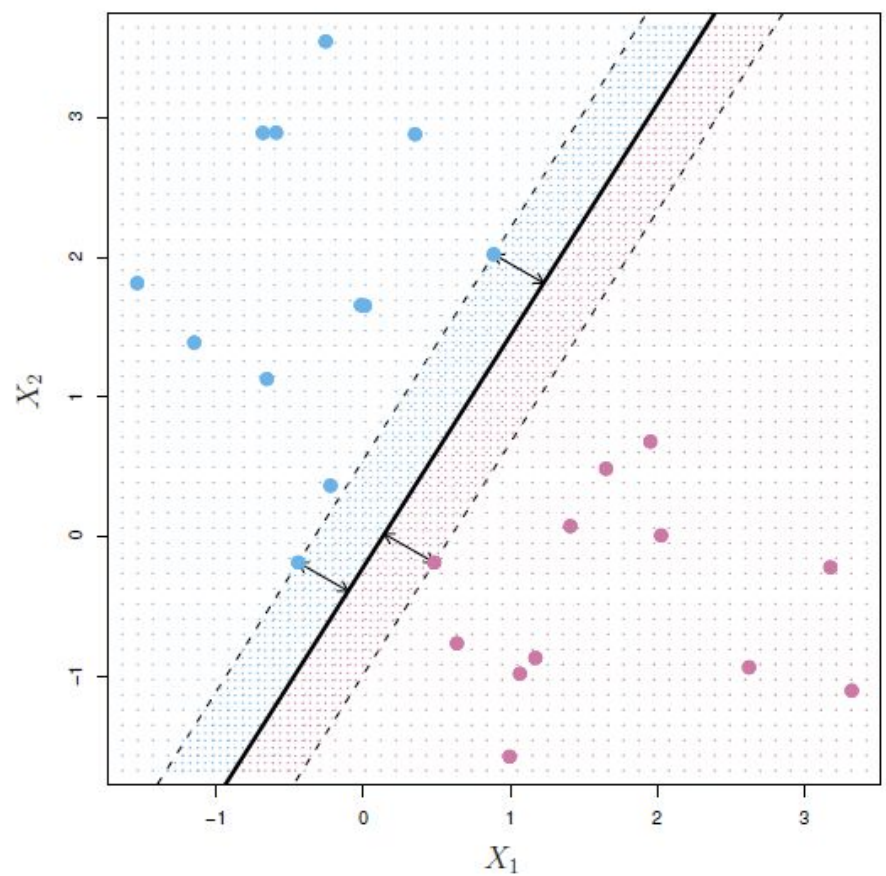


Diabetes Classification with Support Vector Machines



Problem: chronic diseases impact large proportions of the population, but there are many different lifestyles that may impact prevalence of diseases, like diabetes.

Data Source: IPUMS Health Survey extracts demographic and lifestyle data
Goal: Classify instances of Diabetes based on 5 predictors

Theoretical Background: Support Vector Machines

- Allows classification by drawing a hyperplane (a boundary)
 - In 2-D, a line; in 3-D, a plane
 - Question: aren't there infinitely many boundaries possibles? Yes, so use maximal margin
- Relevant metric: the distance between the closest points (the maximal margin)
 - Determines boundary
 - SVM attempts to draw a boundary that would maximize the distance of the two closest points
 - The closest points serve as our *support vectors*

Parameters

- Kernel: Our boundary does not have to be linear, as seen below
- Cost: How many points can be misclassified
 - Lost cost may lead to overfitting, but a high cost
- Gamma: Describes the influence of training examples

Advantages: intuitive understanding and presentability, many flexible kernels
Drawbacks: SVM requires no NA data, leading to difficult applications

Predictors



1.HINOTCOVE: health insurance coverage

2. EDUC: education level

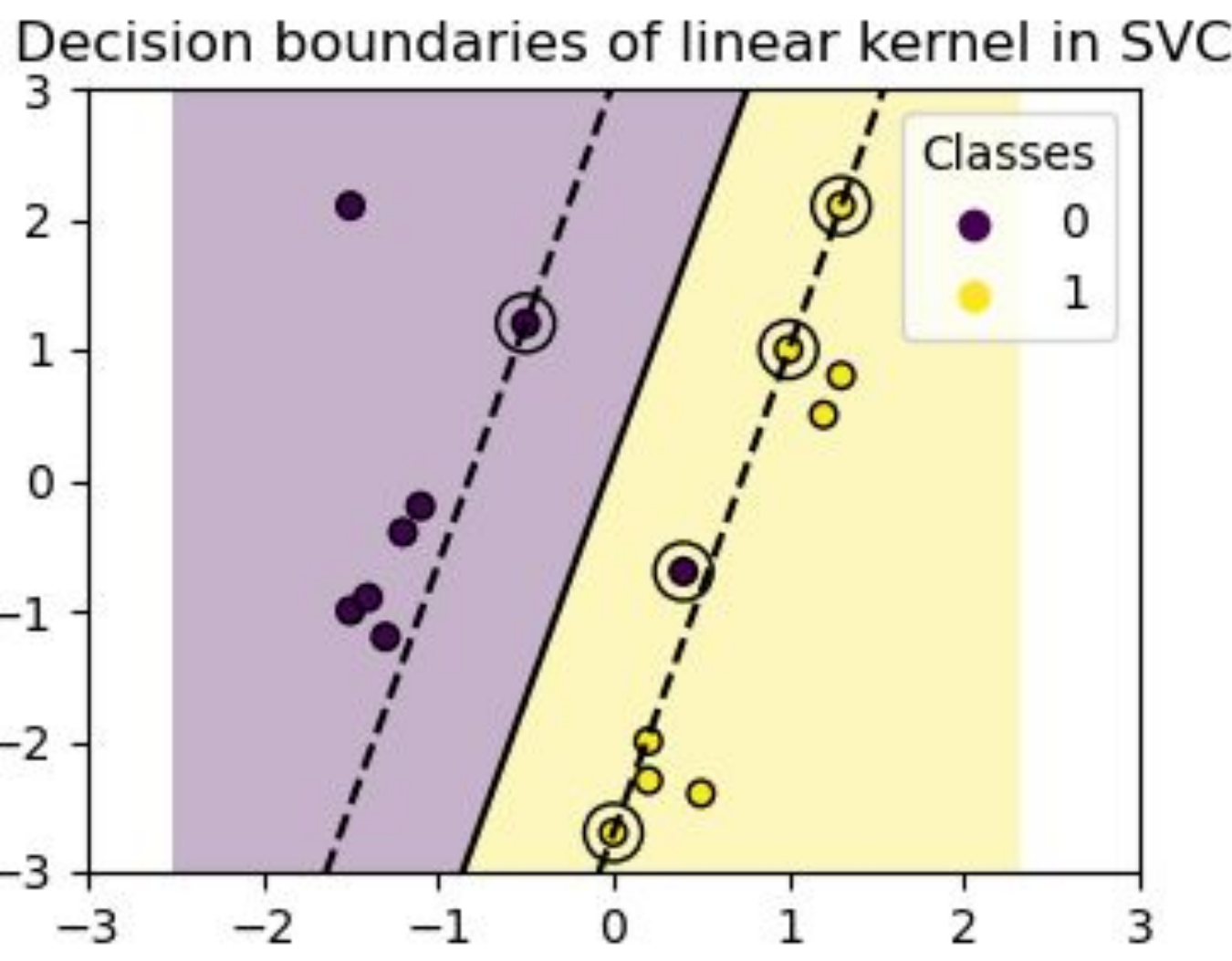
3. VIG10DMIN: vigorous exercise per day

4.HOURSWRK: hours worked per week

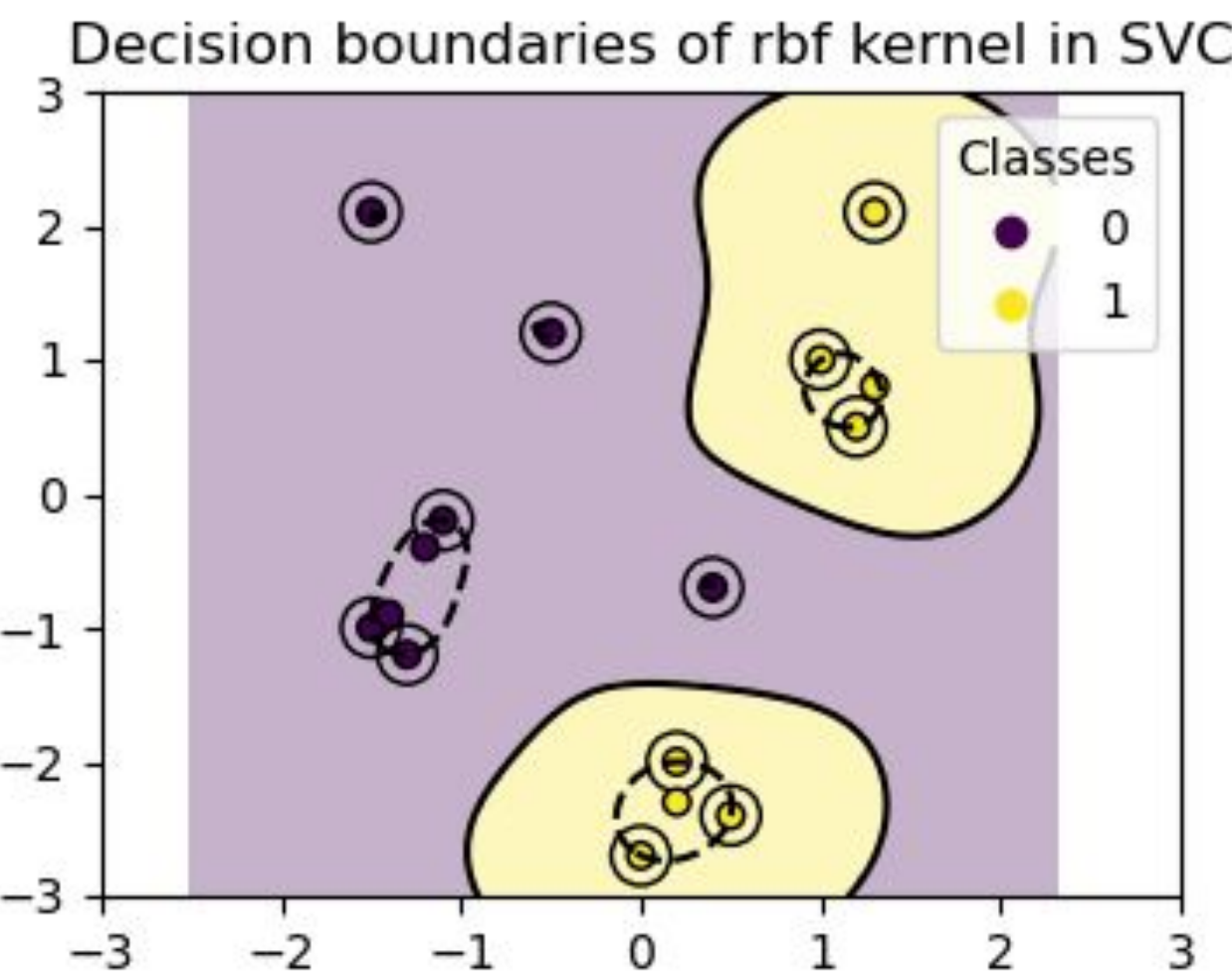
5.SALADSNO: salads consumed

SVM Kernels

Linear: Creates a straight hyperplane

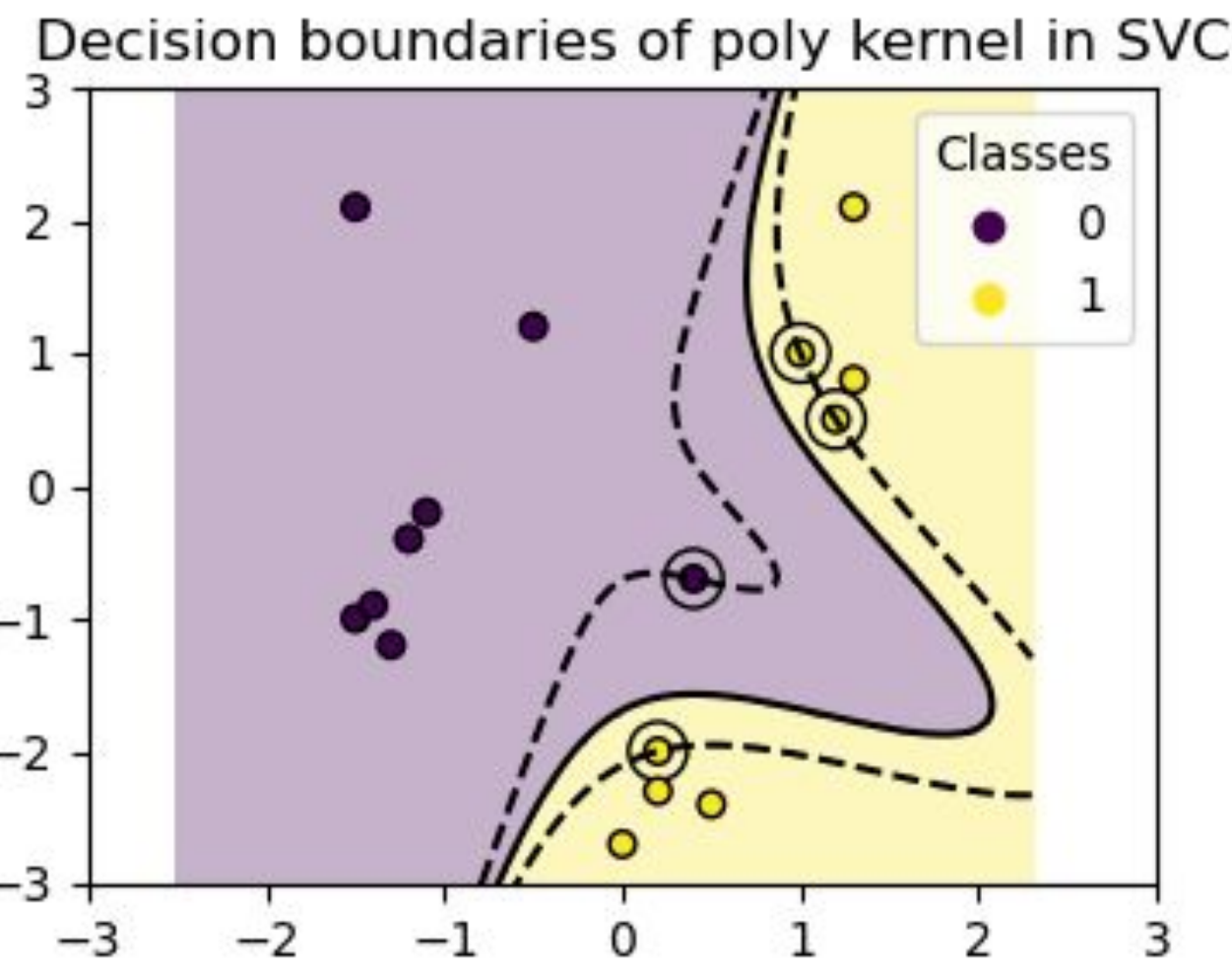


Radial: Creates a circular and rounded hyperplane, creating grouped boundaries



Polynomial: Creates a complex and curved hyperplane

- Relevant parameter: degree

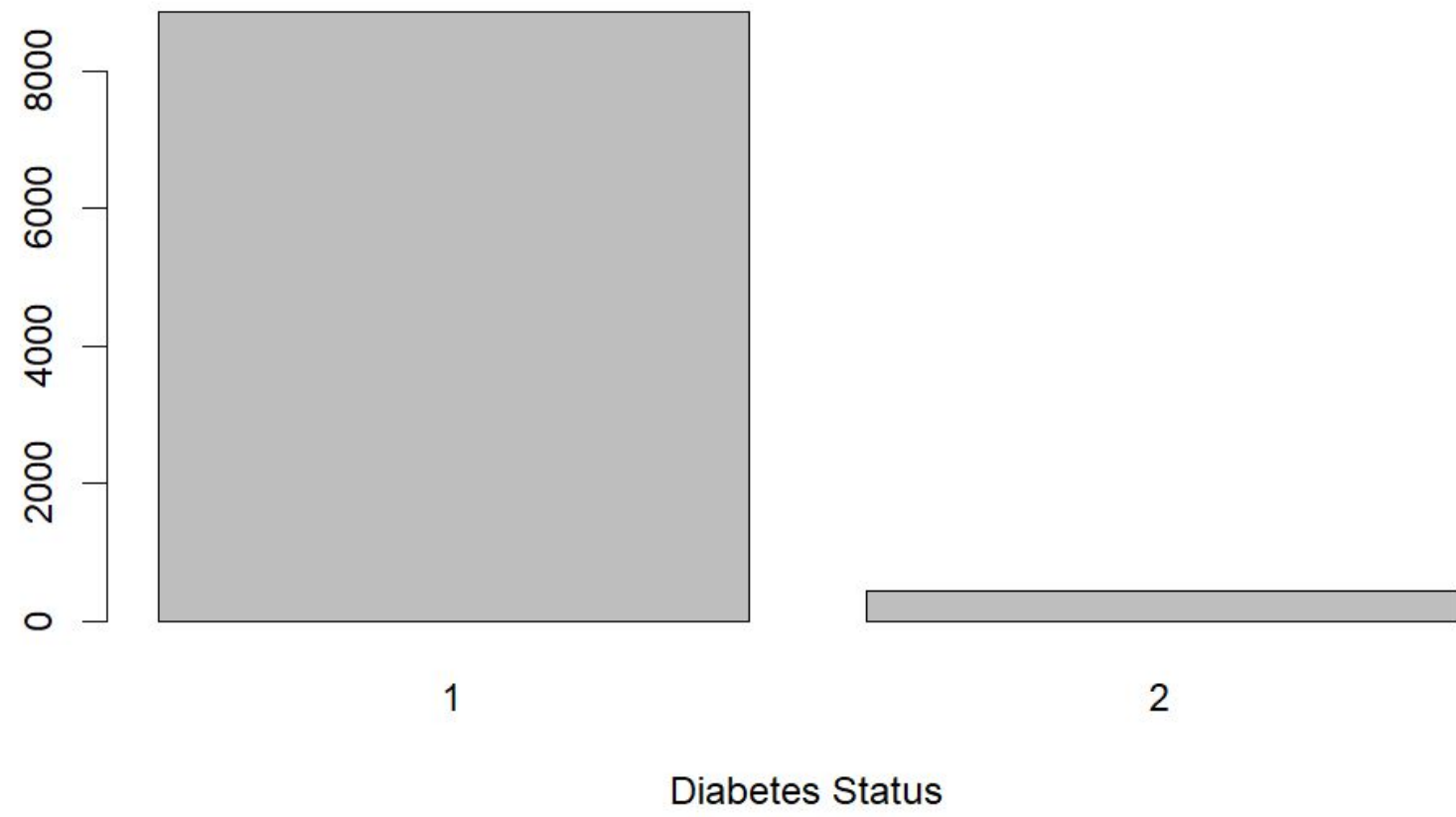


Methodology:

- Problem with unbalanced data.
 - Solve by decreasing the amount of non-diabetic samples to balance data
- Resolving difficult NA data
 - Many survey participants refused to answer many questions
 - We set those results to NA and had to remove them from the dataset, as SVM requires all data points to have a value

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
 - Radial SVMs perform the best, followed by linear



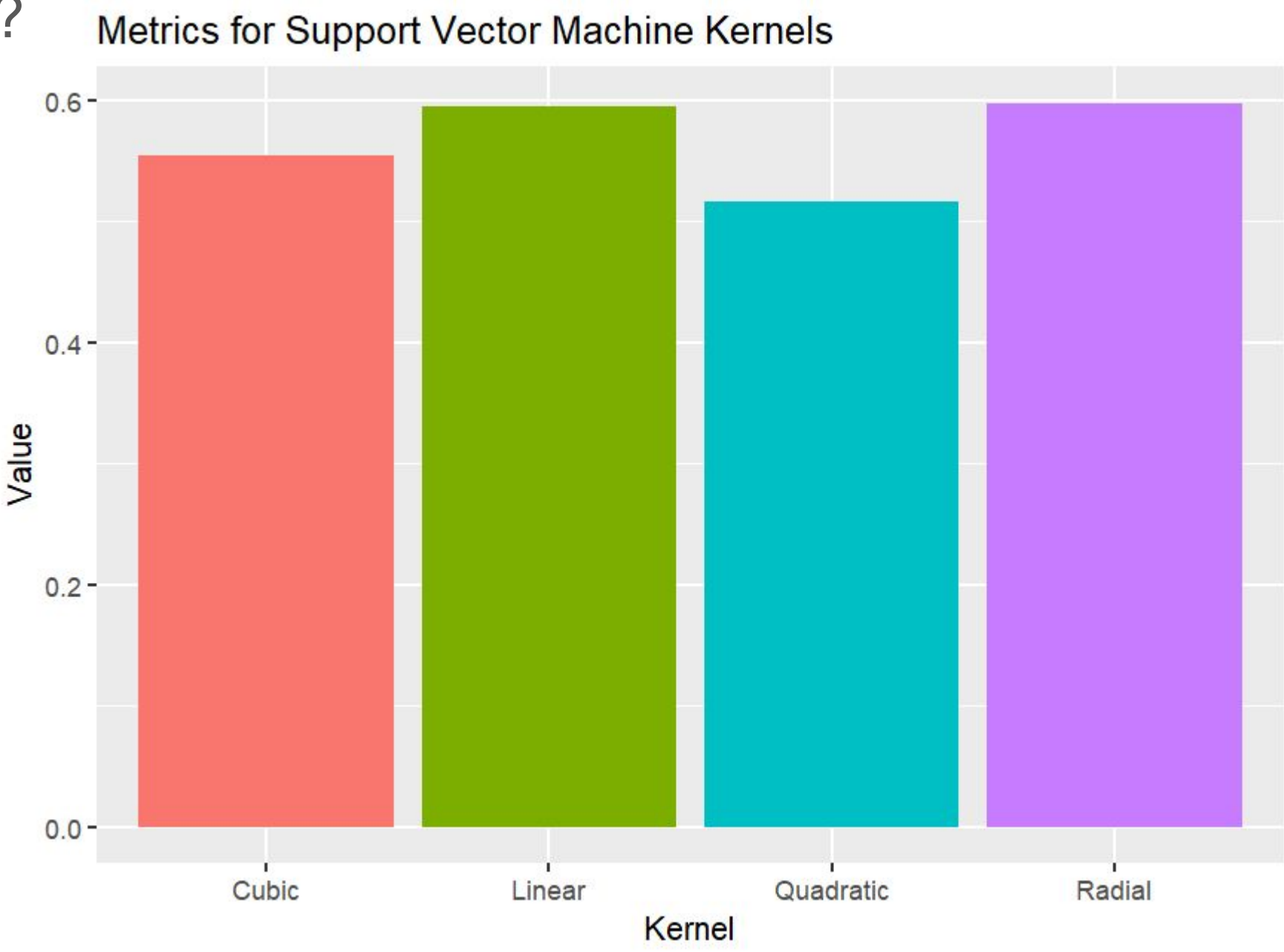
Issues with Shrinking Imbalanced Data:

- Unstable initial conditions with testing data
- Less quality in models
- High variance in data and model predictions

Other options and methods:

Re-sample and bootstrap to create another dataset
Collect more data from other sources
Use less variables to avoid deleting NA rows

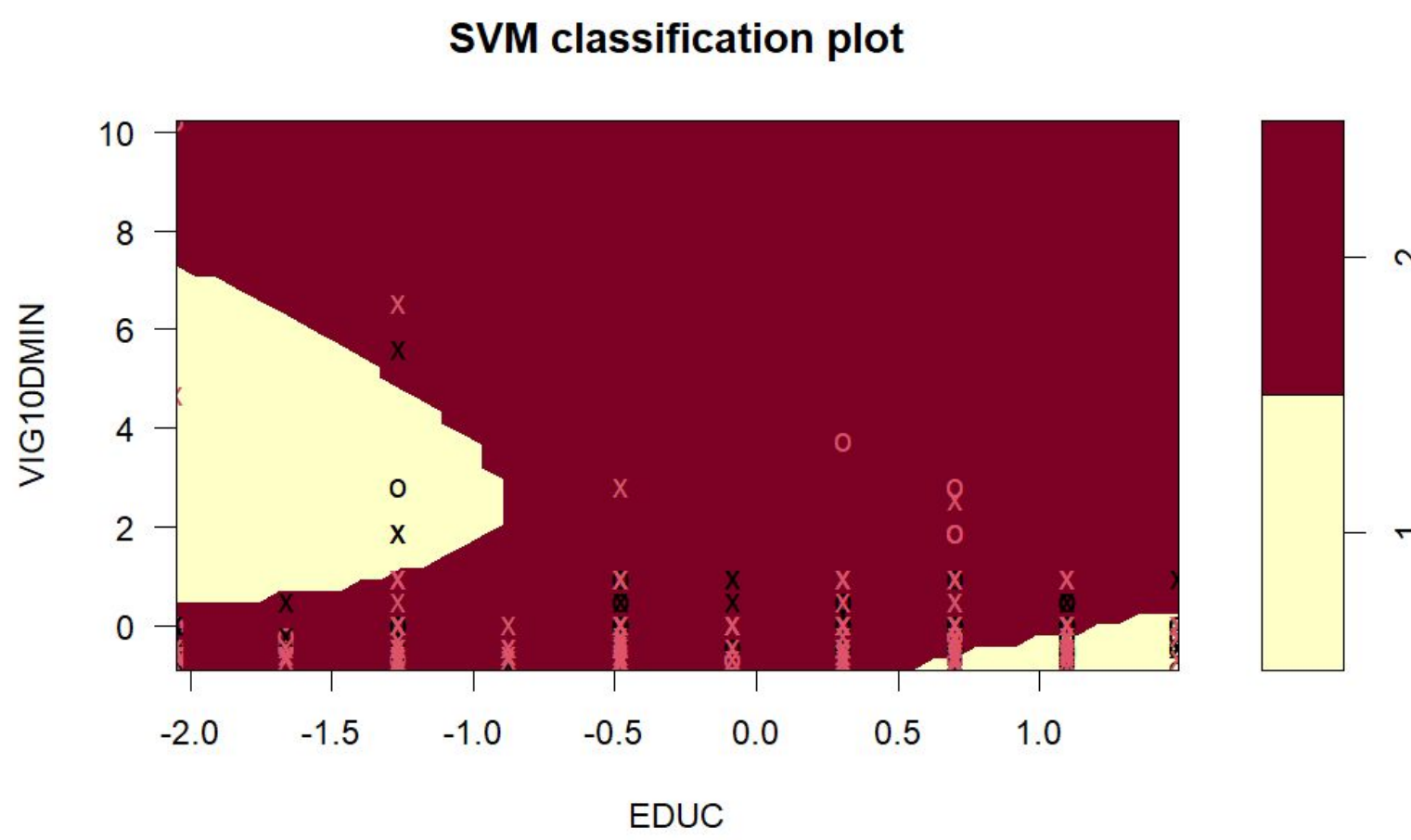
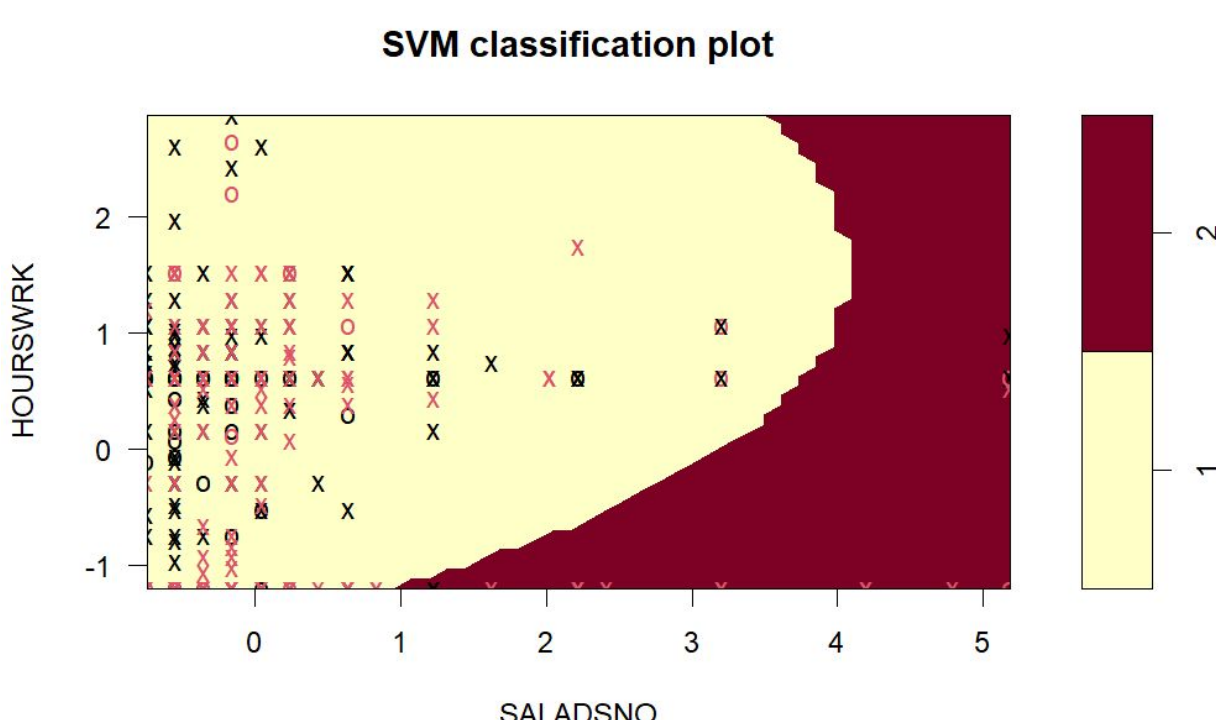
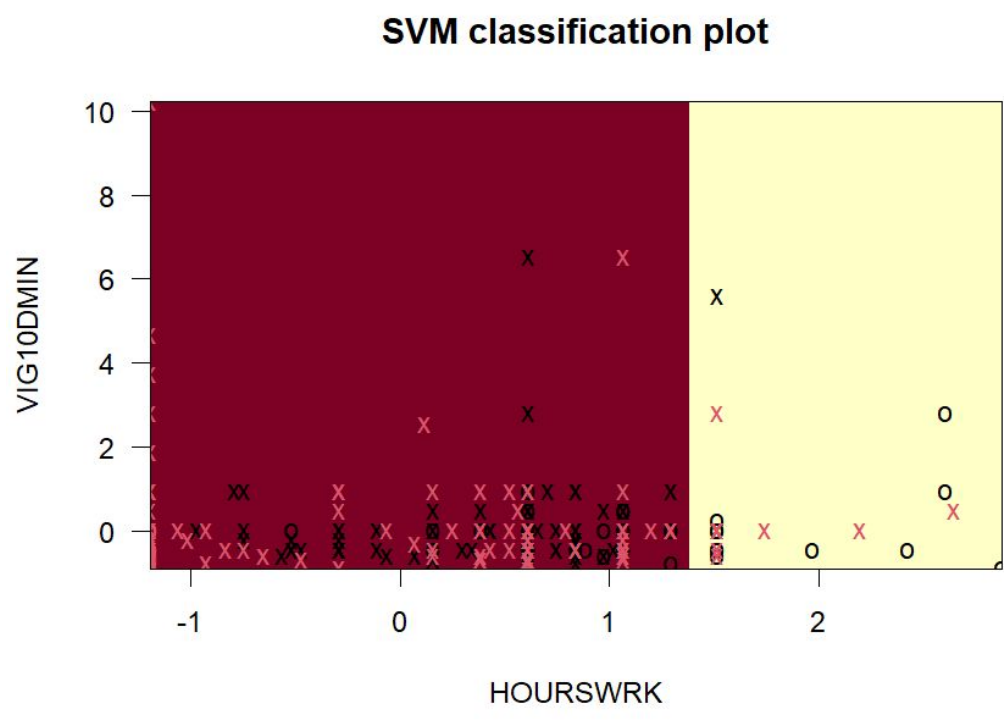
Kernel	Linear	Radial	Polynomial (Quadratic)	Polynomial (Cubic)
Accuracy	59.57%	59.78%	53.33%	57.63%



Diabetic



Nondiabetic



Discussion:

Model Performance: all models did about equally as well with the exception of the quadratic polynomial kernel.
Best model: Radial kernel

- This implies the data has a grouped structure, where ranges of values may have diabetes, but not exactly a strictly linear relationship where the increase of one status implies the increase of another.

Conclusion and Key takeaways:

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.