Jesse Loi

Diabetes Classification with Support Vector Machines

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

<u>Data Source</u>: IPUMS Health Survey extracts demographic and lifestyle data
<u>Goal</u>: 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
- Relevant Equation: the distance between the closest points
  - Determines boundary
  - SVM attempts to draw a boundary that would maximize the distance of the two closest points

Parameters
- <u>Cost</u>: How many points can be misclassified
- <u>Gamma</u>: Describes the influence of training examples

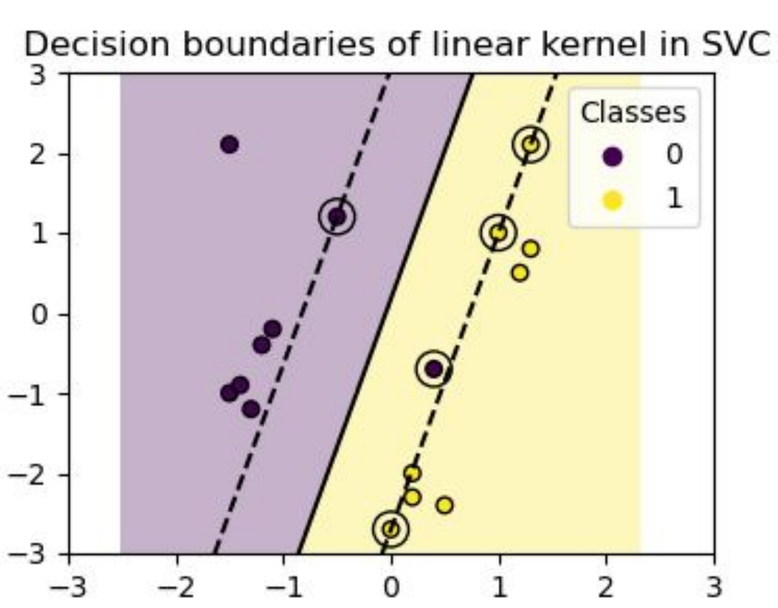
Advantages: intuitive understanding and presentability, many flexible kernels
Drawbacks: SVM requires no NA data, leading to difficult applications

Predictors

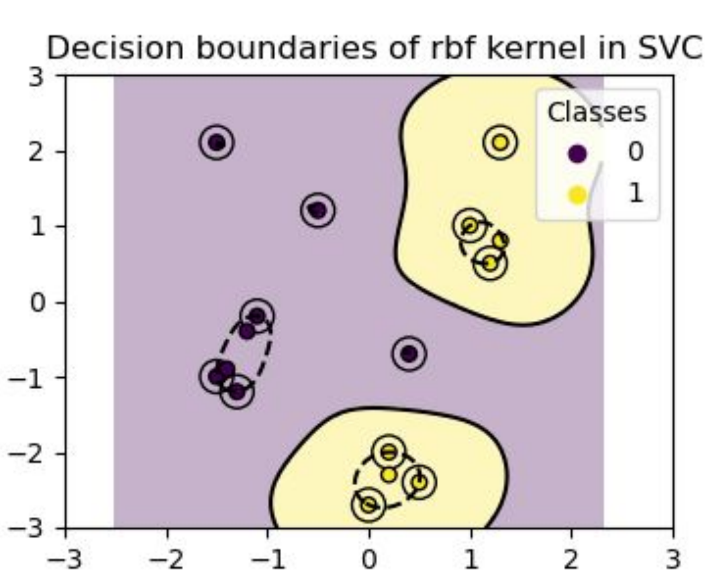
1. HINOTCOVE: health insurance coverage
2. EDUC: education level
3. VIG10DMIN: vigorous exercise per day
4. HOURSWRK: hours worked per week
5. SALADSNO: salads consumed

SVM Kernels

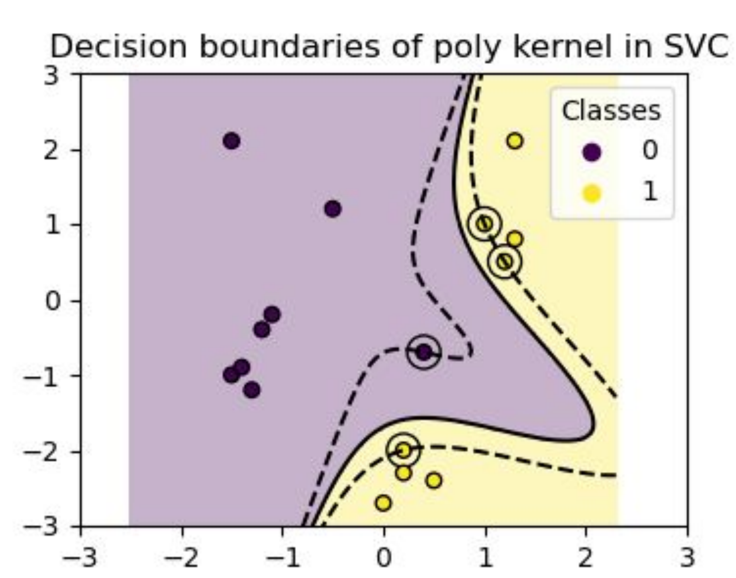
Linear: Creates a straight hyperplane



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

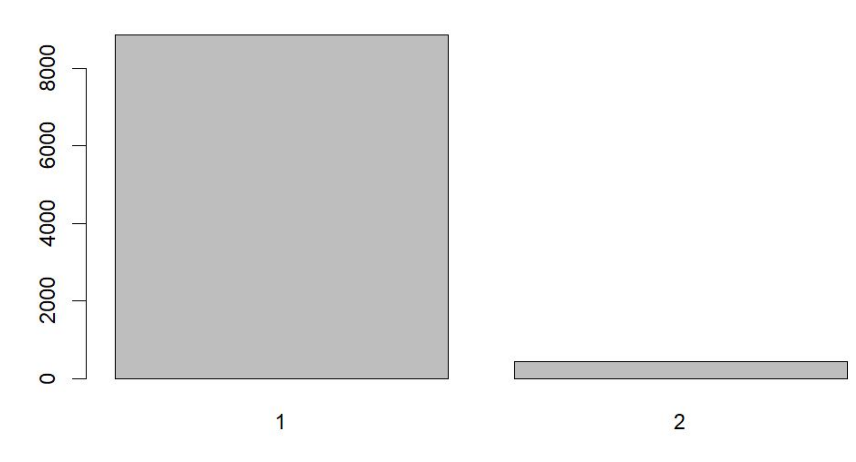


Polynomial: Creates a complex and curved hyperplane
- Relevant parameter: degree



<u>Methodology:</u>
- 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

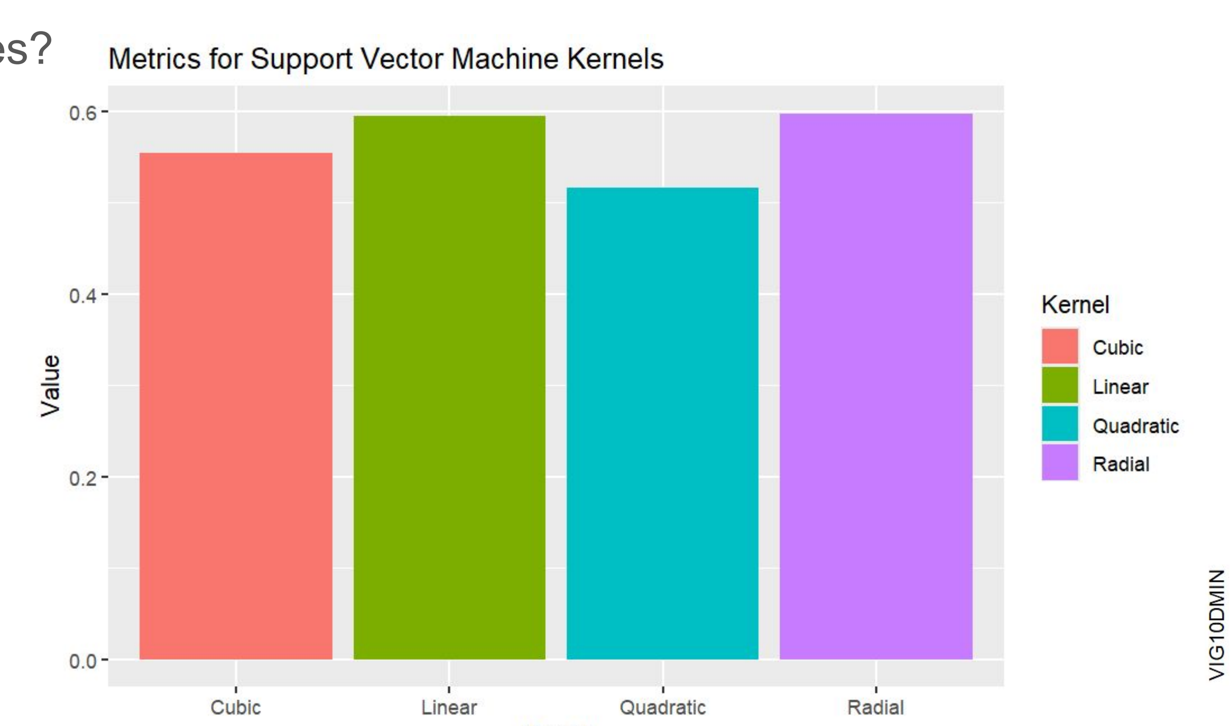
Results:
- Intuitively healthy habits seems to be associated with higher rates of diabetes classification
  - Question: why do high activity and high salad eating individuals have diabetes?
- Models including the 5 above predictors did not have very strong predictions
  - Radial SVMs perform the best, 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

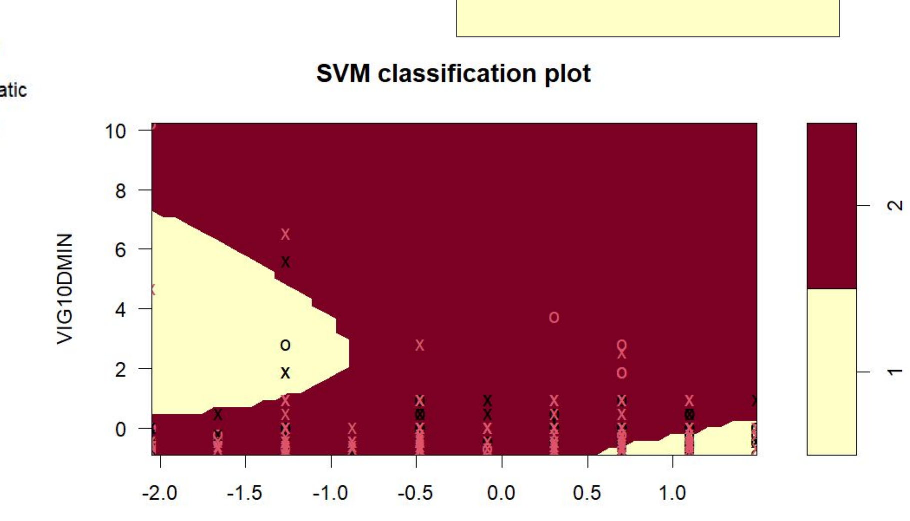
<u>Other options and methods</u>:
Re-sample and bootstrap to create another dataset
Collect more data from other sources
Use less variables to avoid deleting NA rows

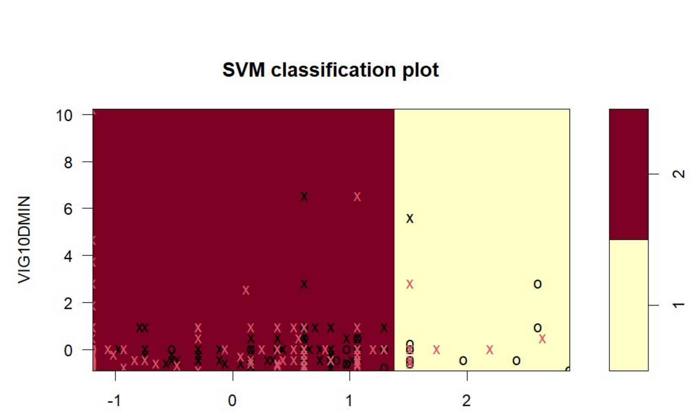
Kernel	Linear	Radial	Polynomial (Quadratic)	Polynomial (Cubic)
Accuracy	59.57%	59.78%	53.33%	57.63%

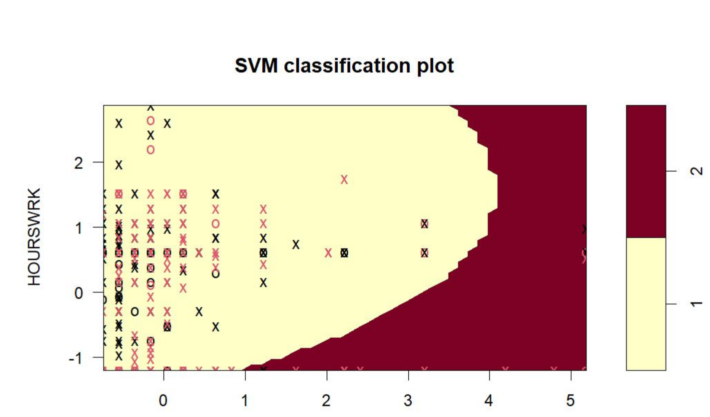


Diabetic

Nondiabetic







<u>Discussion</u>:

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.

<u>Conclusion and Key takeaways</u>:
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.

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
Scikit-learn developers. (n.d.). SVM: Separating hyperplane for different kernels. Scikit-learn. Retrieved from https://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html