Predicting Diabetes from IPUMS NHIS Data with Support Vector Machines

Elling Payne, Seattle University, April 2025

Introduction:

The data in this study comes from the IPUMS Health Surveys National Health Interview Survey conducted in 2022 [1]. The original data set contained information from over 35,000 sample adults and children representing their households. The survey includes information about several health-related behaviors, past disease diagnosis status, and basic demographic information. Given the nature of the sample and that properly weighting the samples given household and other demographics is beyond the scope of this study, no insights from this study can be reliably extrapolated beyond the population of sample individuals. Furthermore, this study only explores information about male sample adults. The data used in this study was obtained through Dr. Mendible of Seattle University [2]

Technical Background - Support Vector Machines:

Let M denote the magnitude of the margin selected on either side of the decision hyperplane in feature space of dimension p, $H := \{\vec{x} \in \mathbb{R}^p : \vec{\beta}^T\vec{x} = 0\}$. Let $\vec{x}_i$, y_i be the features and response class of the i^{th} observation. The i^{th} observation violates the margin if it falls on the side of its margin closer to the decision boundary, or on the wrong side of the decision boundary. Let B be the space of n-dimensional real-valued unit vectors. Let E denote the space of n-dimensional real-valued vectors with strictly non-negative components, and $\vec{\epsilon}$ a vector who's components measure the magnitude of margin violation for each observation, where $0 < \epsilon_i < 1$ means $\vec{x}_i$ is on the right side of the decision boundary but within the margin and $1 < \epsilon$ means $\vec{x}_i$ is misclassified. Then we may select and tune a total budget for margin violation, C, depending on the number of observations and amount of response class overlap. Then we may state the goal of a SVM.

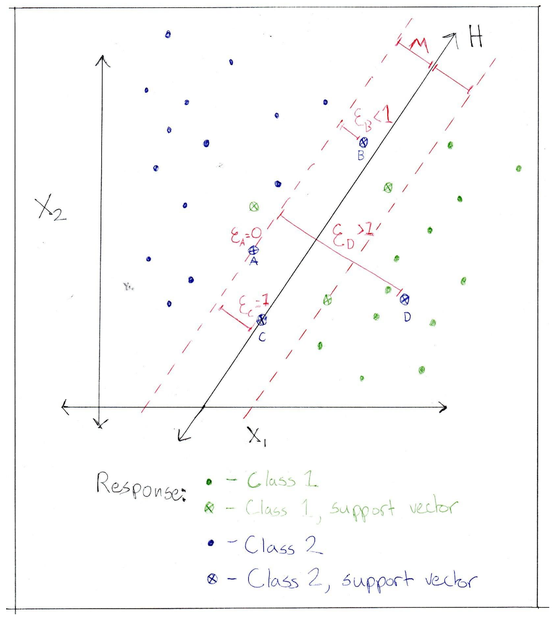
$$\text{Maximize:} \quad M \text{ for } \vec{\beta} \in B, \vec{\epsilon} \in E, M \in \mathbb{R}^+$$

$$\text{Subject to:} \quad \sum_{i=1}^{p} \beta_i^2 = 1$$

$$\forall i \in \{1, ..., n\}, y_i(\vec{\beta}^T\vec{x}_i) \geq M(1 - \epsilon_i)$$

$$\forall i \in \{1, ..., n\}, \epsilon_i \geq 0$$

$$\underline{n}$$



SVMs depend on a useful definition of distance in feature space. Therefore, normalization is necessary to equalize contributions of features. Alternatively, weighting may be done carefully in the case of known disparities in feature importance. Kernels allow for changing the notion of distance and therefore the types of decision boundaries that are possible. A SVM to classify $\vec{x}$ using n training observations can be rewritten as $\beta_0 + \sum_{i=1}^{n} \alpha_i K(\vec{x}, \vec{x_i})$, where $K(\cdot)$ is the Kernel, a function to measure distance between points in feature space. In the transformed, extended feature space, linear boundaries may project nonlinear boundaries onto the original feature space.

Linear Kernel:

$$K(\vec{u}, \vec{v}) = \vec{u}^T\vec{v}$$

Polynomial Kernel:

$$K(\vec{u}, \vec{v}) = (1 + \sum_{i=1}^{p} u_i v_i)^d$$

where d is the degree used for fitting the boundary

Radial (circular) Kernel:

$$K(\vec{u}, \vec{v}) = e^{-\gamma \sum_{i=1}^{p} (u_i - v_i)^2}$$

where $\gamma > 0$ shrinks further distances

While not applicable for this project, SVMs may be extended to multiple classification of K classes through two main ensemble methods, the "all-pairs" approach and the "one-versus-rest" approach.

All-Pairs: # of SVMs: $\binom{K}{2}$

Predicted class chosen by vote.

On-Versus-Rest: # of SVMs: K

Chosen by furthest distance from H

Methods:

Sample Subset: Male sample adults (10151 cleaned records)

Response: Has the participant ever been diagnosed with diabetes? (DIABETICEV)

1: No, 2: Yes (other values removed)

Features: (Codes not valid as numeric data removed)

- Participant age in years (AGE)
- Participant BMI calculated value (BMICALC)
- Calculated yearly days of alcohol consumption last year (ALCDAYSYR)
- Hours worked last week (HOURSWRK)
- Number of sodas consumed over a time period (SODAPNO)
- Number of fries consumed over a time period (FRIESPNO)

Model Types:

All six above features were used in the models. Models were tuned and compared using a linear, polynomial, and radial kernel.

Tuning and Training:

All numerical data was scaled by normalization for training. Inverse weighting was applied during tuning and training to help account for response class imbalance.

Parameters were tuned using five-fold cross validation with a balanced classification error as the metric. The models with minimum balanced CV error for each kernel were selected. Balanced error is calculated as the equally-weighted mean of classification error for each class. Training was performed on a random training sample comprising 80% of the total sample subset.

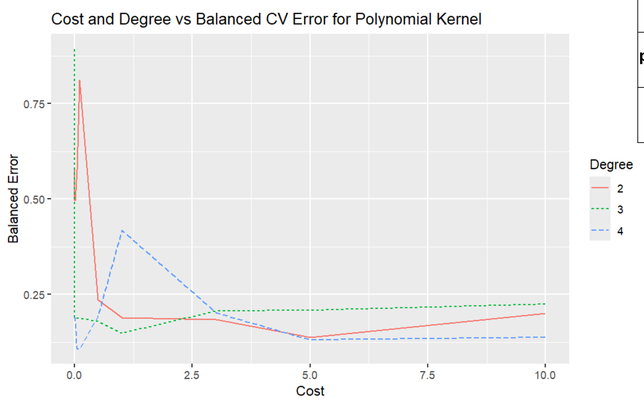
Final Evaluation and Kernel Evaluation:

Tuned models were compared primarily using test error with the balanced error metric, and secondarily with training error and training CV error with the balanced metric.

Results:

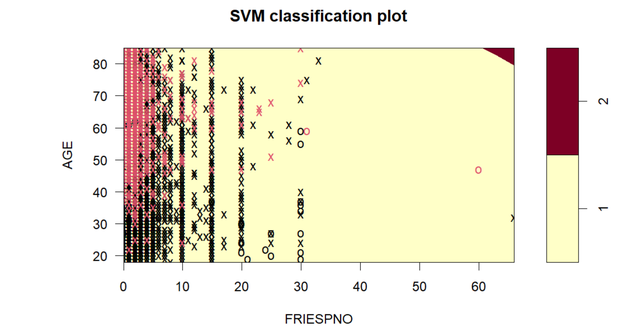
Highlight: The best model by balanced CV Error used all six features, a cost of 0.1 (C = 10), and a polynomial kernel of degree d=4.

Tuning:



Final Model Metrics:

kernel	cost	degree	gamma	CV Error	Training Error	Testing Error
linear	0.05	NA	NA	0.109961	0.109962	0.099507
polynomial	0.05	4	NA	0.110207	0.109962	0.1
radial	0.05	NA	0.001	0.109961	0.109962	0.099507



Discussion and Conclusions:

Based on initial plots of the relationship between potential features and the diabetes diagnosis flag variable, several variable stood out in terms of visible trends. The most obvious trend for all disease indicators in the dataset was age. This makes sense, as many of the diseases in the dataset, such as diabetes, cancer, and heart disease, can be the result of many small factors building up over time. Age also makes sense to include because it is a datapoint that is often readily available and may have interactions with other risk factors. Calculated BMI was another basic health metric that appeared to have a relationship visually with diabetes in particular, but not as much with other diseases. This makes sense given that diabetes is closely linked to metabolism, and severe diabetes often comes with difficulty in physical exertion. While higher BMI might not be the cause of diabetes, it may provide predictive power, especially in conjunction with other features. The last demographic variable used here was hours worked in the last week. Initially, my though was that overwork might be associated with greater sickness, but the opposite was true. In hindsight, this makes a lot of sense, and is likely caused by the participant being too sick to work consistently. This variable may have less predictive power that is specific to diabetes than some of the others, given that any serious disease could be the cause of missing work or working less. However, the variable was one of a few that seemed to have a relationship with diabetes. If the feature can give extra confidence that someone has any disease, and other variables can narrow it down to diabetes, it might still be useful to include. The other three features including were all behavioral health metrics for substances that seemed likely to be related to diabetes. Since diabetes is tied to metabolism and for some an inability to process sugars, consumption of three things that can spike sugar intake seemed relevant. The boxplots also showed a weak but potentially real relationship between alcohol consumption, soda consumption, and french fry consumption. Based on my findings, it seems a sensible approach to further explore the link between foods high in simple sugars or alcohol and diabetes. It also seems prudent to invest in prevention and education while people are still fairly young, since the risk increases with age.

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. https://www.nhis.ipums.org

[2] Mendible, Ariana. (2025). *5322* [source code]. GitHub. https://github.com/mendible/5322

[3] R Core Team (2025). *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

[4] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.

[5] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2024). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-16, https://CRAN.R-project.org/package=e1071.

[6] Lamport, Leslie (1986). *LaTeX: A Document Preparation System*. Addison-Wesley

[7] Overleaf (2025). *Overleaf, Online LaTeX Editor*. https://www.overleaf.com

[8] James, G. , Witten, D., Hastie, T., Tibshirani, R. (2023). An Introduction to Statistical Learning with Applications in R. Springer https://www.statlearning.com