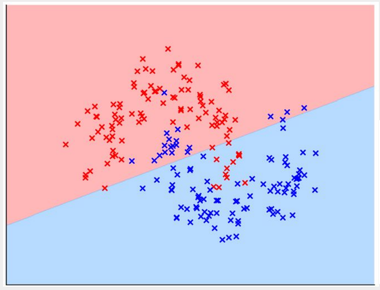
Predicting Diabetes Risk for People Above Age 30 (NHIS Data)

Sung Ahn | DATA 5322 | 4/28/2025

Technical Background

Support Vector Machines (SVMs) are supervised learning models used for classification tasks. They work by finding the optimal hyperplane that separates data points of different classes.

1. Linear SVMs

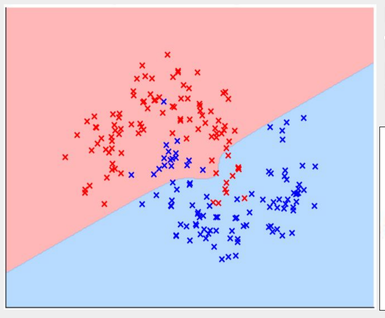


Linear SVMs are done by just the standard dot product. It assumes data is linearly separable in the input space.

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

TL;DR: Just separate with a straight line or flat plane

2. Polynomial SVMs

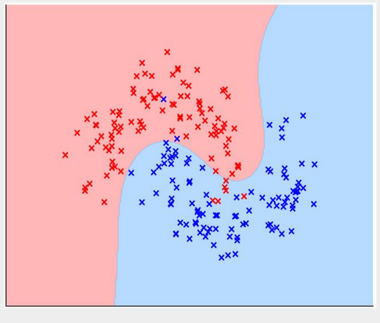


Models interactions between features up to degree "d". Takes dot product, shifts it by a constant "r" (1 in this case), and raise to degree "d.

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

TL;DR: Allow for curved boundaries & model feature interactions

3. Radial SVMs



Measures similarity based on distance. Points closer together have higher similarity. Good for highly non-linear boundaries.

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$$

TL;DR: If two points are close together, call them similar, doesn't care about a line at all. Very flexible but needs tuning

Note: The **kernel** computes similarity between two points, x_i and $x_{i'}$ without explicitly mapping them. Each kernel changes the notion of "similarity" differently.

Models & Methods

Target

DIABETICEV
(Told to have Diabetes)

Features

- BMICALC → Body Mass Index (BMI)
- MOD10DMIN → Moderate physical activity per day (≥10 minutes/session)
- PIZZANO → Weekly consumption of Pizza
- FRIESPNO → Weekly consumption of Fries
- VEGENO → Weekly consumption of vegetables

Model Flow

Raw Data

Data Cleaning
(Feature Selection)

Grid Search CV
(Hyperparameter Tuning)

Run Models & Evaluate

Mean Test Error,
AUC Score,
Precision-Recall

Models

Linear SVM: BMICALC, MOD10DMIN

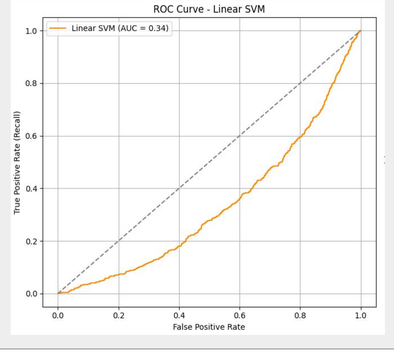
Poly SVM: PIZZANO, VEGENO, FRIESPNO, SALADSNO, SODAPNO

Radial SVM: BMICALC, PIZZANO, VEGENO, FRIESPNO, MOD10DMIN

Parameters

C (Penalty Parameter)	**Small C** → Wider margin, allow more misclassifications. **Large C** → Narrow margin, penalize misclassifications heavily.
γ (Gamma for RBF)	**Small γ** → Far influence (smooth decision boundary). **Large γ** → Close influence (tight decision boundary).
d (Degree for Polynomial)	**Higher** → More flexible, can fit more complex relationships, but risks overfitting.

Metrics



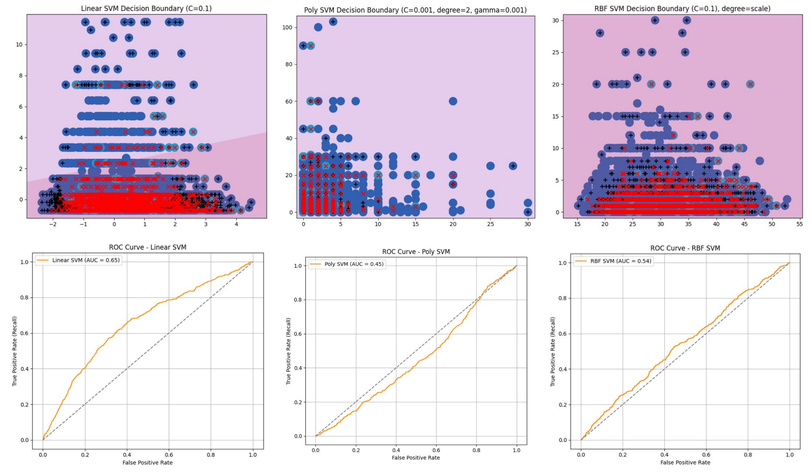
ROC AUC evaluates overall classification ability regardless of the threshold

Precision-Recall evaluates how well the model retrieves true positives without including too many false positives. Good for imbalanced datasets

Results & Discussion

SVM models default to always predicting 0. Precision/Recall for class 1 is 0.00. All 492 real patients with disease (class 1) are misclassified as healthy (class 0). This shows a class imbalance problem and is true as 18165 people reported "No" for diabetes while only 2458 reported "Yes"

Linear | **Poly** | **RBF**



Interpretation

Class 0 (No, disease) has 3632 samples
Class 1 (Yes, disease) has 492 samples

Kernel Type	Accuracy	AUC Score	Average Precision
Linear	88%	0.654	0.198
Poly	77%	0.45	0.105
RBF	88%	0.540	0.136

The current selected features do a terrible job separating diabetic individuals from healthy ones. It seems habits (BMI, exercise, diet) do matter as BMICALC and MOD10DMIN increase the model performance compared to models without them. However, they alone are not enough to predict whether a person would have diabetes.

Linear model with just BMI and exercise features surprisingly did the best. Poly model with food features did absolutely terrible (AUC below 0.5) and RBF model with a combination did okay, but not as great.

Demographic or Social Factors

As mentioned above, BMICALC and MOD10DMIN were used in the linear model and it seems that body mass index and physical activity are relatively important. Even with a major class imbalance, there is some positives shown through the AUC score and the precision score. Technically, food factors (like PIZZANO) could be a representation of lifestyle and often tied to socioeconomic background. Regardless, logically makes sense that these factors may affect diabetes.

Suggestions to Policy-Makers

Even though body metrics and habits like BMI, exercise, and diet matter, they aren't the full story. The system needs broader policies combining lifestyle improvements + social support to effectively reduce rate of diabetes in the population.

[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. Links to an external site.http://www.nhis.ipums.org