Predicting Diabetes Outcomes Using Support Vector Machines (SVMs) Evidence from NHIS 2022 Health Survey

By Surya Kailash Ramesh

Introduction

This project is to investigate whether it's possible to predict the presence of chronic diseases such as diabetes, heart disease, and cancer, using demographic and lifestyle data from U.S. adults. My goal was not only to build accurate predictive models but also to understand how behaviors like diet, physical activity, sleep, and substance use relate to chronic health outcomes. particularly interested in whether these variables, when combined, could effectively separate individuals with and without a disease diagnosis.

To explore this, this project uses the 2022 National Health Interview Survey (NHIS) data, accessed through the IPUMS Health Surveys platform. The dataset includes a wide range of features covering demographics, socioeconomic status, and health behaviors. I limited this analysis to adult respondents flagged as "sample adults" to ensure I had complete information for everyone. To answer this research question, three types of Support Vector Machines (SVMs): a linear SVM, a polynomial SVM, and a radial basis function (RBF) SVM. Each of these models offers a different perspective on the data. The linear SVM helps determine whether a simple, linear combination of health and lifestyle variables is enough to distinguish between people with and without disease. The polynomial SVM allows me to explore interactions between variables and non-linear relationships, which can reveal more complex patterns in health behavior. The RBF SVM, known for its flexibility, tests whether the boundary between disease and no-disease groups is highly non-linear and requires a model that adapts to complex structures in the data.

By comparing these models, I aimed to identify not only the most accurate approach but also the one that best captures the underlying patterns in how lifestyle and demographics contribute to chronic illness.

Technical Background

Support Vector Machines (SVMs) for binary classification using three kernel types: **linear, polynomial, and radial basis function (RBF)**. SVM is a supervised learning model that finds an optimal hyperplane to separate data points of different classes with the maximum margin.

The basic formulation for a linear SVM seeks to minimize the term $\frac{1}{2} \| \boldsymbol{w} \|^2$, subject to the constraint that each training instance satisfies $\boldsymbol{yi}(\boldsymbol{w}^T \boldsymbol{x_i} + \boldsymbol{b}) \geq \mathbf{1}$, where xi is the feature vector, $\boldsymbol{y_i} \in \{-\mathbf{1}, \mathbf{1}\}$ is the class label, w is the weight vector, and b is the bias.

For non-linearly separable data, slack variables $\boldsymbol{\xi_i}$ are introduced to allow for misclassification. The soft-margin SVM objective becomes minimizing $\frac{1}{2} \| \boldsymbol{w} \|^2 + \boldsymbol{C} \sum \boldsymbol{\xi_i}$, where $\boldsymbol{C} > \mathbf{0}$ is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error. To model non-linear relationships, kernel functions are used to map the input space to a higher-dimensional feature space. The three kernels used are:

(1)The **linear kernel**, defined as

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{x_i}^T \boldsymbol{x_j}$$

(2)The **polynomial kernel**, defined as

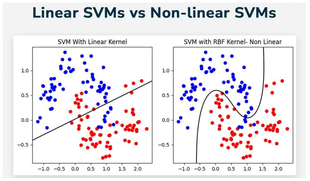
$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\boldsymbol{\gamma} \boldsymbol{x_i}^T \boldsymbol{x_j} + \boldsymbol{r})^{\boldsymbol{d}}$$

where γ is a scaling factor,
r is a coefficient,
d is the degree of the polynomial;

(3) The **RBF kernel**, defined as

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{exp}(-\boldsymbol{\gamma} \| \boldsymbol{xi} - \boldsymbol{xj} \|^2)$$

where γ controls the influence of each training point.



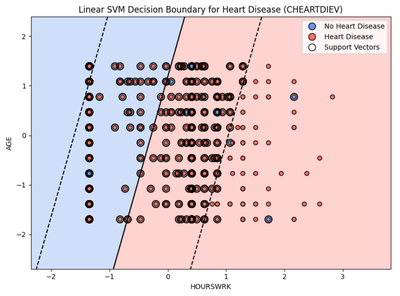
Hyperparameter tuning was performed using grid search with 5-fold cross-validation. The parameters searched included C (regularization), γ (for RBF and polynomial kernels), and d (for polynomial kernels), using ranges such as C∈{0.1,1}, γ∈{0.01,0.1,1}, and d∈{2,3,4}. Model performance was evaluated using metrics including **accuracy, precision, recall, F1-score, and AUC-ROC.**

Confusion matrices were used to interpret classification errors, and ROC curves helped visualize the model's ability to distinguish between classes. The linear SVM offers a simple decision boundary, making it useful when the data is linearly separable. The polynomial SVM introduces flexibility but may lead to overfitting if the degree is too high. The RBF kernel proved most effective in modeling complex, non-linear decision boundaries. A visual diagram of the decision boundaries for each kernel type enhances understanding of how each model behaves in feature space.

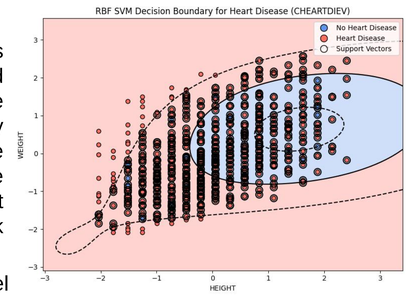
Results

Each figure displays the decision surface in color (red and blue regions), where the background color indicates the predicted class: light red for "positive" (disease) and light blue for "negative" (no disease). The dashed red lines represent the decision margins, and the solid boundary between color regions is the SVM's decision boundary. Circles outlined in black indicate support vectors.

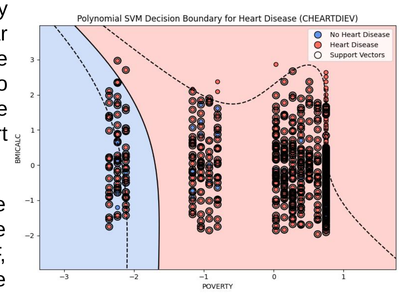
This plot shows how a **Linear SVM** separates individuals into heart disease and no heart disease groups using age and hours worked. The solid black line is the decision boundary, and the dashed lines mark the margins. Red and blue regions show the model's predictions, while black-circled points are support vectors that define the boundary.



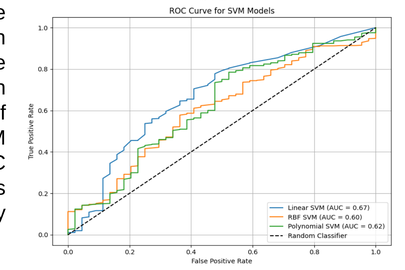
RBF SVM classifies individuals as having heart disease or not based on their height and weight. The flexible, curved decision boundary adjusts to the distribution of the data. Red and blue areas show the model's predictions, with support vectors highlighted by black outlines.



Polynomial SVM uses poverty level and BMI to classify individuals by heart disease status. The nonlinear decision boundary captures the interaction between the two features. Predicted outcomes are shown in red and blue, with support vectors marked by black circles.



The ROC curve in the image compares the performance of three different SVM models: Linear, RBF, and Polynomial. The curve plots the true positive rate against the false positive rate at various classification thresholds. Among the three, the Linear SVM performs the best with an AUC (Area Under the Curve) of 0.67, followed by Polynomial SVM (AUC = 0.62), and RBF SVM (AUC = 0.60). An AUC of 0.5 corresponds to a random classifier, as shown by the diagonal black dashed line.



Since all three models perform better than random guessing, but only modestly, this indicates the features used provide some predictive value but may not be strongly separable or informative. The relatively higher performance of the Linear SVM suggests that the data may be more linearly separable than non-linear, or that the complexity introduced by non-linear kernels like RBF and Polynomial may not be justified by the data distribution.

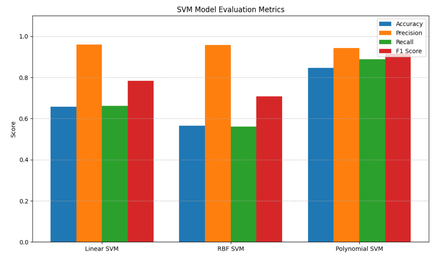
Methodology

Many variables in the NHIS use special codes for missing or invalid responses, which I filtered out to avoid bias. I scaled all numeric variables to standardize their influence on the SVM's decision boundary. To narrow the analysis, I filtered the dataset to respondents between 50 and 60 years old. This wasn't arbitrary, targeting this age group allowed me to focus on a high-risk population and avoid confounding effects from much younger or older individuals. To ensure model interpretability and prevent scale bias, all numeric predictors were standardized using **StandardScaler**, and only complete cases were retained for analysis.

For **feature selection**, I took slightly different routes depending on the kernel. With the linear SVM, I used **SelectKBest** with a univariate regression test to find the five most predictive features. For the RBF and Poly SVM, I used Random Forest feature importances, which gave me a more nonlinear perspective on variable influence.

Hyperparameter tuning was done using grid search and 5-fold cross-validation. I tried different values of C for all models, and for RBF and Polynomial SVMs, I also tuned gamma and degree as needed. Each model was trained on different subsets of features to match its strengths.

Model performance was evaluated using four primary metrics: **accuracy, precision, recall, and F1-score**. Results shows that the **Polynomial SVM** outperforms both the Linear and RBF kernels across all metrics, achieving an accuracy of 0.847, precision of 0.944, recall of 0.890, and F1-score of 0.916. The linear model, while less flexible, still performed moderately well (F1-score: 0.783), while the RBF model had the lowest accuracy (0.566) and recall (0.561), despite high precision (0.959).



Confusion matrices and decision boundary visualizations were also generated to help understand the classification behavior and highlight the role of support vectors in shaping each boundary.

Conclusions

This project gave me a deeper idea of how even simple demographic and lifestyle variables can reveal patterns in chronic disease risk. Of the models tested, the Polynomial SVM stood out. It achieved the best performance across accuracy, precision, recall, and F1-score.

More importantly, it captured the kind of nuanced, nonlinear relationships suspected might be present, particularly between BMI, poverty, and heart health. Working with this dataset also reminded me how important thoughtful data preparation is. Filtering the age group, handling missing codes properly, and selecting the right features all had a big impact on model performance. And also learned that high precision doesn't always mean the model is doing well overall, RBF SVM had great precision but suffered in recall, which is made to reconsider how evaluate performance.

In a broader sense, this project reinforced for me how valuable machine learning can be in public health, especially when it's grounded in clean data and clear questions. Tools like SVMs aren't just math, they're decision aids. They can help flag at-risk individuals, support early intervention, and ultimately contribute to better health outcomes at scale. Going forward, I'd be interested in testing this approach on other chronic conditions and integrating behavioral data in more detail.

References

- "sklearn.svm.LinearSVC." scikit-learn: Machine Learning in Python, version 1.6.1, 2025, https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html.
- GeeksforGeeks. "Feature Selection Using Random Forest." GeeksforGeeks, 2025, https://www.geeksforgeeks.org/feature-selection-using-random-forest/.
- 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 Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D070.V7.4. Links to an external site.http://www.nhis.ipums.orgLinks to an external site.
- GeeksforGeeks. "Linear vs Non-Linear Classification: Analyzing Differences Using the Kernel Trick." GeeksforGeeks, 4 Mar. 2024, www.geeksforgeeks.org/linear-vs-non-linear-classification-analyzing differences-using-the-kernel-trick/. Accessed 8 May 2025.