PREDICTING DIABETES RISK USING SUPPORT VECTOR MACHINES: A DEMOGRAPHIC AND BEHAVIORAL ANALYSIS

Erdenetuya Namsrai, Dr. Ariana Mendible
enamsrai@seattleu.edu, mendible@seattleu.edu

COLLEGE OF
SCIENCE AND ENGINEERING

ABSTRACT

Disease prevention is a critical priority for individuals and healthcare systems. Diabetes prevalence has risen sharply, with 37.3 million individuals diagnosed in the U.S. as of 2022—a number expected to grow annually. To support proactive intervention, this study applies machine learning to identify key factors associated with diabetes.

We investigated demographic, health metrics and habits variables using Support Vector Machine (SVM) models specifically Linear, Radial, and Polynomial kernels on the 2022 National Health Interview Survey (NHIS) dataset. The analysis identified age, body mass index, weight, weekly working hours, education level, alcohol consumption days and physical activity as significant predictors.

Model performance was evaluated using training/testing errors and classification metrics including precision, accuracy, recall, and F1-score. Hyperparameter tuning via cross-validated grid search and parallel computing improved both performance and efficiency. The tuned SVM models achieved strong predictive accuracy without overfitting.

INTRODUCTION

- Diabetes remains a major public health challenge globally, driven by demographic changes and lifestyle factors. Early identification of individuals at risk is crucial for effective intervention, yet traditional risk assessments often overlook key behavioral and demographic variables.
- This study investigates the use of **Support Vector Machine (SVM)** models to predict the presence of diabetes using data from the **2022 National Health Interview Survey (NHIS)**.
- The objective is to develop predictive models based on a selected set of **demographic**, **health metrics** and **habits** variables, such as **age**, **body mass index**, **alcohol use**, **physical activity**, and **work hours**.
- Machine learning methods, particularly Support Vector Machines (SVM), offer powerful alternatives by capturing complex, non-linear relationships in high-dimensional data. Three kernel based SVM models were applied and evaluated:
  - **- Linear SVM**
  - **- Radial Basis Function (RBF) SVM**
  - **- Polynomial SVM**
- Through systematic comparison and hyperparameter tuning, this study aims to identify the most effective SVM approach for diabetes classification and highlight the key variables influencing risk.

THEORETICAL BACKGROUND

- Support Vector Machines (SVM) are supervised learning algorithms ideal for classification tasks, especially with high-dimensional data. SVMs identify the optimal hyperplane that maximizes the margin between classes, improving generalization and reducing overfitting.
- To model complex patterns, SVMs use kernel functions to project input data into higher-dimensional spaces:
  - **- Linear kernel**: Best for linearly separable data.
  - **- Radial Basis Function (RBF)**: Captures non-linear relationships based on distance.
  - **- Polynomial kernel**: Allows curved decision boundaries and models variable interactions.
- **Hyperparameter tuning** (cost, gamma, degree) is essential to balance model flexibility and accuracy.
- **Grid search with cross-validation** is used to systematically identify optimal hyperparameters.
- SVMs are well-suited for **predicting diabetes** due to the complex interactions among demographic, health metrics, and habits factors.
- Combining multiple kernels with **tuning strategies** enhances the model's ability to capture diverse data structures and **improve prediction performance**.

METHODOLOGY

This study followed a structured machine learning process including data preparation, feature selection, model training, tuning, and evaluation.

Dataset:

- **Source**: National Health Interview Survey (NHIS) 2022.
- **Variables**: Included demographic, health metrics, and behavioral factors relevant to diabetes prediction.

Data Preparation:

- Included only adults (age ≥ 18) to focus on adult-onset diabetes.
- Removed records with missing values and cleaned invalid/special code values based on NHIS codebook:
  ALCANYNO ≥ 996, CIGDAYMO ≥ 996, EDUC ≥ 996, MOD10DMIN >= 996 …
- Created new binary target variable: **Diabetes**
- Converted categorical variables to factors: SEX, HINOTCOVE

Feature Selection:

Correlation analysis identified ten key predictors: AGE, BMICALC, WEIGHT, HOURSWRK, ALCDAYSYR, EDUC, VIG10DMIN, POVERTY, and MOD10DMIN representing socioeconomic, behavioral, and physiological factors.

Class Balancing:

The dataset was highly imbalanced. Downsampling was applied to match diabetic and non-diabetic class sizes 2165 each, improving fairness and model reliability.

Model Development:

Support Vector Machine (SVM) models with **linear**, **radial basis function (RBF)**, and **polynomial** kernels were developed. Baseline models were trained using default parameters, with a 70-30 train-test split.

Hyperparameter Tuning:

Grid search with cross-validation was used to optimize **cost** (C), **gamma** (for RBF), and **degree** (for polynomial).

Feature Scaling:

Predictor variables were standardized in both training and testing sets for consistent scaling.

Model Evaluation:

Evaluated models using:
- Training/Test Error
- Accuracy, Precision, Recall, F1-score, and Confusion Matrix
Comparative analysis was performed between baseline and tuned models to assess improvements.

COMPUTATIONAL RESULTS

Confusion Matrix of Tuned SVM Models

Tuned Linear SVM		
Diabetes	Predicted No Diabetes	Predicted Yes Diabetes
Actual No Diabetes	509 (TN)	214 (FP)
Actual Yes Diabetes	140 (FN)	435 (TP)

The Linear SVM model accurately identified 435 individuals with diabetes (true positives), while 140 diabetic cases were misclassified as non-diabetic (false negatives). These results demonstrate the model's strong classification performance, though the number of false negatives suggests room for improvement in recall (sensitivity) to enhance early detection.

Tuned Radial SVM		
Diabetes	Predicted No Diabetes	Predicted Yes Diabetes
Actual No Diabetes	511 (TN)	209 (FP)
Actual Yes Diabetes	138 (FN)	440 (TP)

The Radial SVM model successfully identified 440 individuals with diabetes (true positives) and misclassified 138 diabetic individuals as non-diabetic (false negatives). This reflects slightly improved recall compared to the linear model, indicating better sensitivity in detecting diabetes cases. However, further refinement may help reduce missed diagnoses.

Tuned Polynomial SVM		
Diabetes	Predicted No Diabetes	Predicted Yes Diabetes
Actual No Diabetes	571 (TN)	279 (FP)
Actual Yes Diabetes	78 (FN)	370 (TP)

The Polynomial SVM model correctly identified 370 individuals with diabetes (true positives) while misclassifying 78 diabetic individuals as non-diabetic (false negatives). This indicates a stronger ability to detect positive cases, achieving the best sensitivity among the three models evaluated. However, some diabetic cases are still missed, suggesting further tuning could enhance recall.

Performance Evalution of Linear SVM, Radial SVM, and Polynomial SVM

Model	Training Accuracy	Testing Accuracy	Training Error	Test Error
Linear SVM	72.1%	72.3%	27.9%	27.7%
Radial SVM	**75.6%**	**73.1%**	**25.4%**	**26.8%**
Polynomial SVM	72.9%	72.4%	27.2%	27.5%

Model Evaluation Using Precision, Recall, F1-Score, and Sensitivity

Model	Precision		Recall		F1-Score		Sensitivity
	Yes Diabetes	No Diabetes	Yes Diabetes	No Diabetes	Yes Diabetes	No Diabetes	
Linear SVM	70%	76%	78%	67%	74%	71%	78.4%
Radial SVM	70%	78%	**81%**	65%	75%	71%	81.2%
Polynomial SVM	68%	80%	85%	59%	76%	68%	85.3%

Among the tuned models, SVM with Radial kernel achieved the highest recall 81% for predicting Yes Diabetes cases, indicating it was the most effective at correctly identifying diabetic individuals. Meanwhile, Linear and Polynomial SVMs showed stable and balanced performance with minimal difference between training and test accuracy, suggesting good generalization.

Performance Evalution of Tuned Linear SVM, Tuned Radial SVM, and Tuned Polynomial SVM

Model	Training Accuracy	Testing Accuracy	Training Error	Test Error
Tuned Linear SVM	72.1%	72.7%	27.9%	27.2%
Tuned Radial SVM	**72.1%**	**73.2%**	**27.8%**	**26.7%**
Tuned Polynomial SVM	71.5%	72.5%	28.4%	27.5%

Model Evaluation Using Tuned Precision, Recall, F1-Score, and Sensitivity

Model	Precision		Recall		F1-Score		Sensitivity
	Yes Diabetes	No Diabetes	Yes Diabetes	No Diabetes	Yes Diabetes	No Diabetes	
Tuned Linear SVM	70%	76%	78%	67%	74%	71%	78.4%
Tuned Radial SVM	71%	76%	79%	68%	75%	72%	78.4%
Tuned Polynomial SVM	67%	83%	**88%**	57%	**76%**	67%	**87.9%**

- The Tuned Polynomial SVM demonstrated the best overall performance, achieving the highest recall (88%), F1-score (76%), and weighted accuracy (87.9%) for predicting diabetes cases—indicating strong sensitivity and balanced precision.
- The Tuned Linear and Tuned Radial SVMs showed slightly lower recall, they maintained stable training and test accuracies around 72%, with minimal overfitting, reflecting solid generalization.
- The Polynomial SVM is more effective at capturing nonlinear patterns critical to identifying diabetes, making it the most suitable model for this classification task.

COMPUTATIONAL RESULTS

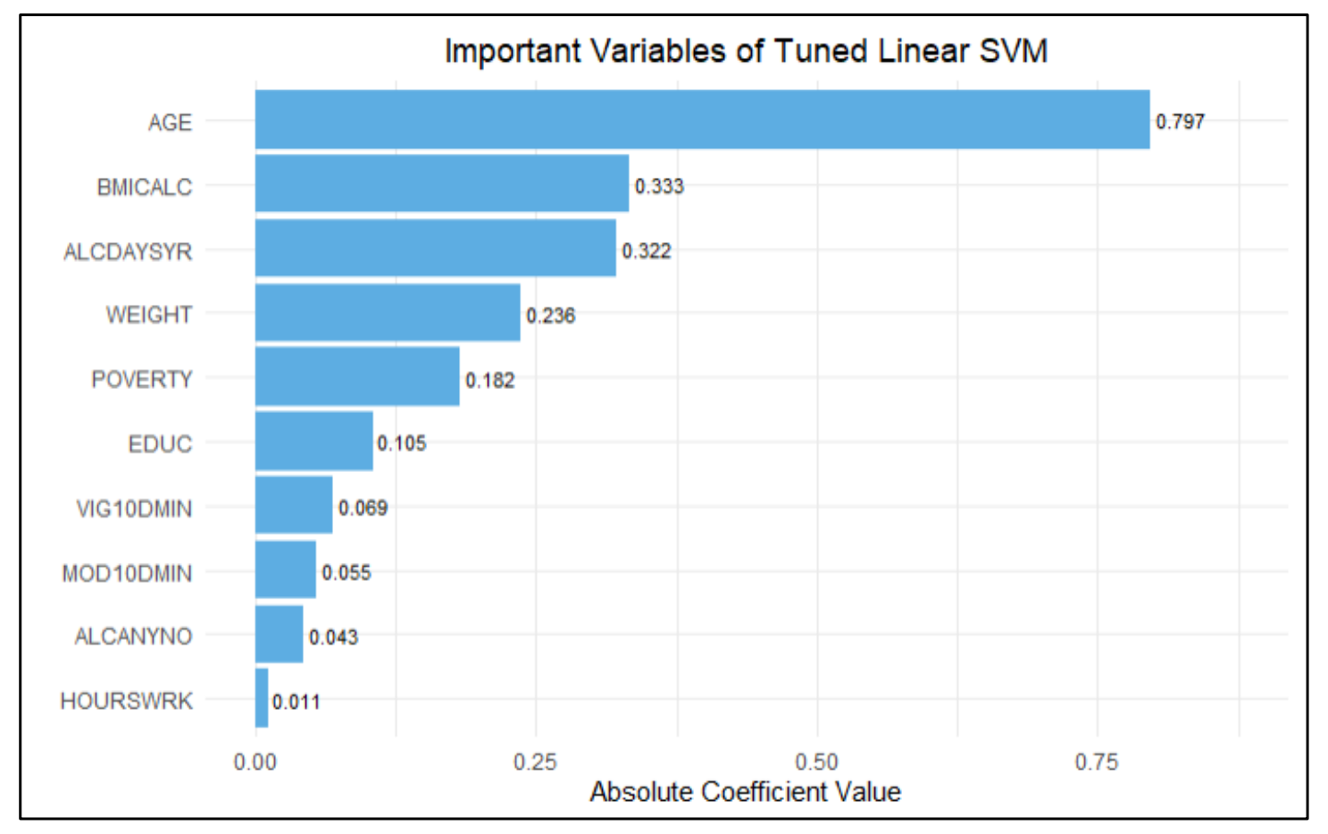
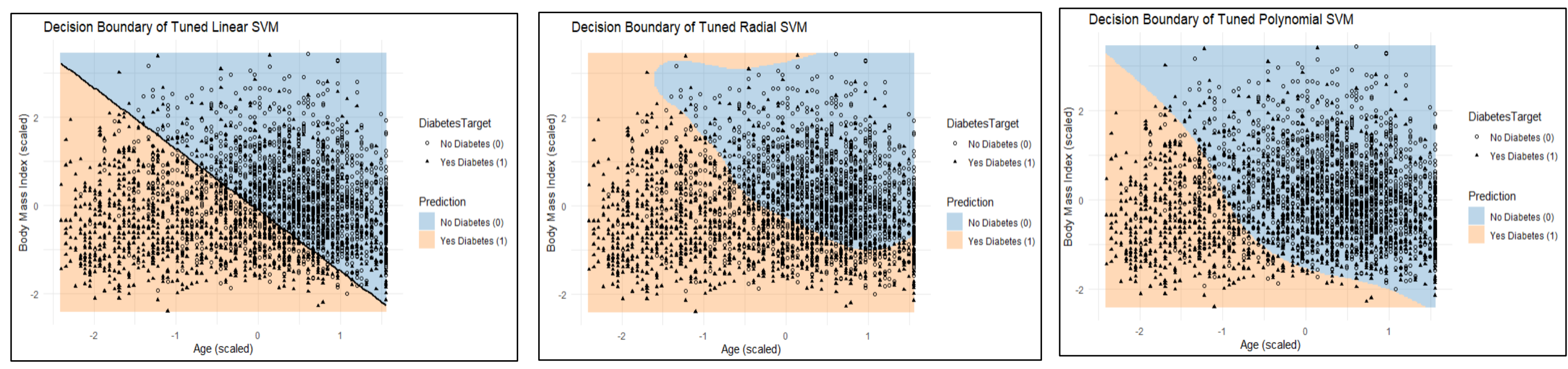


Figure 2. The most 10 important variables for SVM Models

The most influential predictor of diabetes:
- Age
- Body Mass Index
- Frequency drank alcohol in past year: Days in past year
- Weight
- Ratio of family income to poverty threshold
- Educational attainment
- Duration of moderate activity 10+ minutes
- Duration of vigorous activity 10+ minutes
- Frequency drank alcohol in past year: Number of units
- Total hours worked last week or usually

SVM model identified age as the most influential predictor of diabetes, followed by age, body mass index, and alcohol consumption frequency.

Decision Boundaries of Tuned SVM Models



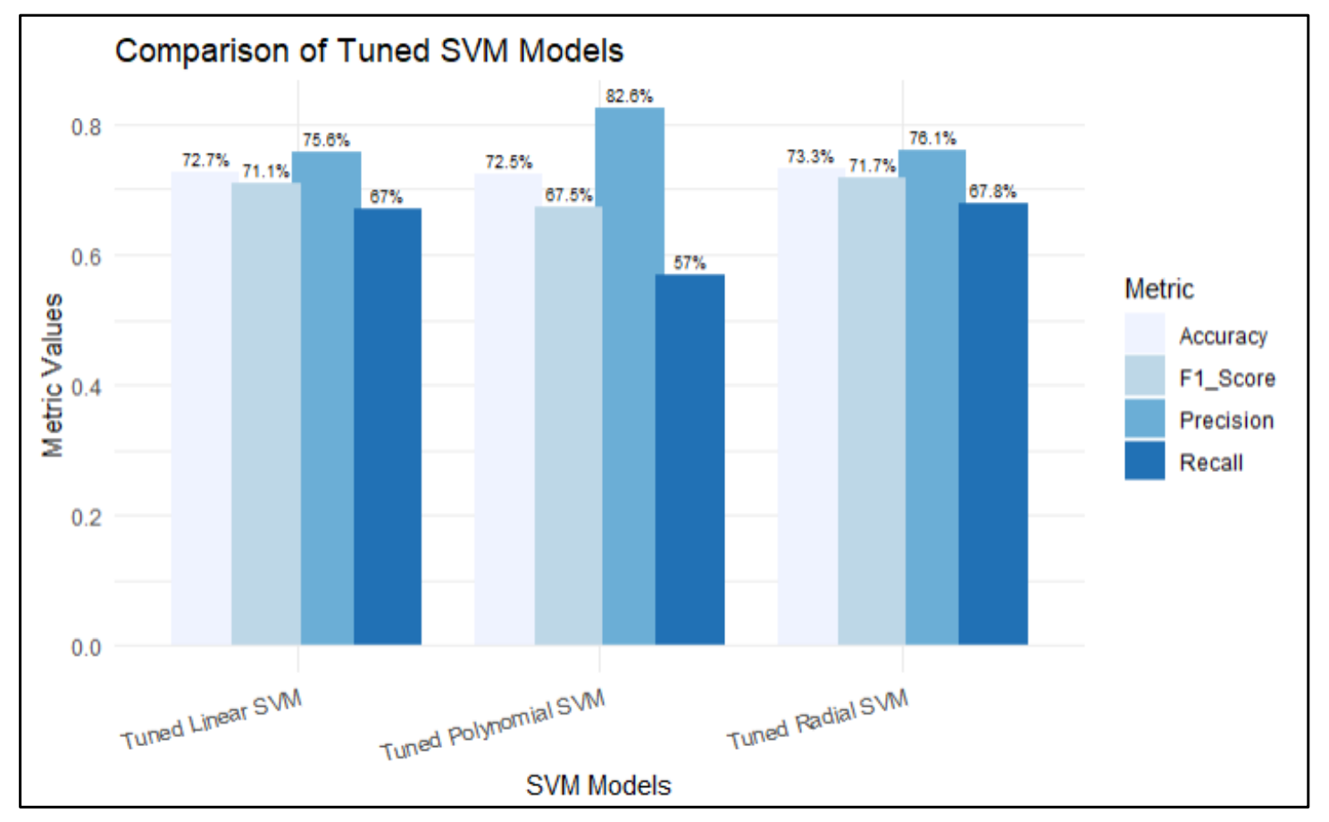


Figure 1. Performance Comparison of Baseline and Tuned Linear SVM Models

In the overall performance of the tuned SVM models:
- The **Radial SVM** achieved the best results, with the highest **accuracy (73.3%)**, **F1-score (71.7%)**, and strong **recall (67.8%)**, making it the most effective at identifying individuals with diabetes.
- The **Linear SVM** performed consistently across all metrics with good generalization.
- The **Polynomial SVM**, despite its high **precision (82.6%)**, showed the lowest **recall**, indicating a higher risk of missing diabetic cases.

These findings highlight **Radial SVM** as the most reliable model for accurate and early diabetes detection.

DISCUSSION

Class Balancing:

The dataset was highly imbalanced. Downsampling created a balanced 1:1 ratio (2165 each), improving fairness and model reliability.

Key Predictors:

SVM identified age, body mass index, and alcohol use as the strongest predictors, followed by weight, poverty, and education. Physical activity and work hours had minimal impact.

Model Comparison:

Radial SVM performed best with the highest accuracy (73.3%), F1-score (71.7%), and recall (67.8%).
Linear SVM showed stable, balanced performance.
Polynomial SVM had high precision (82.6%) but low recall (57%)

Insight:

Among U.S. adults aged 18 and older, age, body mass index, and alcohol consumption emerged as the strongest predictors of diabetes risk. These findings underscore the importance of targeting demographic and lifestyle factors in early detection strategies

CONCLUSION

- This study shows that Support Vector Machines (SVM) effectively predict diabetes risk using demographic, health metrics, and habits.
- The Radial SVM achieved the highest recall, making it the most reliable for identifying at-risk individuals.
- Model performance improved with class balancing and tuning, ensuring fairness and accuracy.
- Age and body mass index were the strongest predictors, aligning with public health evidence.
- The findings of this study have significant implications for the early detection and prevention of diabetes among American adults, and for informing targeted public health interventions

REFERENCES

1. James, G., Hastie, T., Witten, D., & Tibshirani, R. (2023). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.
2. Blewett, L. A., Rivera Drew, J. A., King, M. L., Williams, K. C. W., Backman, D., Chen, A., & Richards, S. (2024). *IPUMS health surveys: National Health Interview Survey, Version 7.4* [Data set]. IPUMS. https://doi.org/10.18128/D070.V7.4
3. Mendible. (2025). *nhis_2022.csv* [Data set]. GitHub. https://github.com/mendible/5322/blob/main/Homework%202/nhis_2022.csv
4. Mendible. (2025). *nhis_2022_codebook.pdf* [PDF]. GitHub. https://github.com/mendible/5322/blob/main/Homework%202/nhis_2022_codebook.pdf
5. R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/