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

This study explores the application of Support Vector Machines (SVM) utilizing linear, radial basis function (RBF), and polynomial kernels to predict the presence of diabetes based on demographic, health, and behavioral variables. Key predictors identified include hours worked per week, education level, poverty status, age, and engagement in vigorous physical activity. Baseline SVM models were developed and assessed using training error, test error, and key classification metrics—accuracy, precision, recall, and F1-score. To enhance predictive performance, hyperparameter tuning was conducted via grid search with cross-validation, leveraging parallel computing for efficiency. Tuned SVM models demonstrated superior generalization performance without evidence of overfitting. These results highlight the importance of systematic feature selection, careful hyperparameter optimization, and informed kernel selection in maximizing the effectiveness of SVM-based diabetes prediction frameworks.

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. Machine learning methods, particularly Support Vector Machines (SVM), offer powerful alternatives by capturing complex, non-linear relationships in high-dimensional data.

This study applies SVM with linear, radial basis function (RBF), and polynomial kernels to predict diabetes based on demographic, health, and behavioral factors. By systematically comparing kernel functions and applying hyperparameter tuning, the research aims to optimize predictive performance and better understand the variables influencing diabetes risk. These findings support the broader adoption of machine learning in public health analytics and personalized preventive care.

THEORETICAL BACKGROUND

Support Vector Machines (SVM) are powerful supervised learning algorithms commonly used for classification, especially with high-dimensional datasets. They work by identifying the optimal hyperplane that maximizes the margin between classes, enhancing generalization and reducing the risk of overfitting.

To model complex, non-linear patterns, SVMs employ kernel functions that map input data into higher-dimensional spaces. Common kernels include the linear kernel for linearly separable data, the radial basis function (RBF) kernel for capturing non-linear relationships based on distance, and the polynomial kernel for enabling curved decision boundaries to model intricate variable interactions. Hyperparameter tuning, particularly for parameters like cost (C), gamma (for RBF), and degree (for polynomial), is essential for controlling model flexibility and preventing overfitting. Grid search with cross-validation is a systematic method for selecting optimal hyperparameters.

Given the complex interplay of demographic, physiological, and behavioral factors in diabetes risk, SVMs are well-suited for this task. Utilizing multiple kernels and tuning strategies ensures that the models are accurately adapted to the underlying data structure, improving prediction performance.

METHODOLOGY

This study followed a systematic approach involving data preparation, feature selection, model training, hyperparameter tuning, and evaluation.

Data Preparation:
The dataset included demographic, health, and behavioral variables related to diabetes prediction. Only adults (age ≥ 18) were considered, and records with missing values were excluded to ensure data quality.

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

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. Class balancing methods were applied where needed.

Hyperparameter Tuning:
Grid search with cross-validation was used to optimize cost (C), gamma (for RBF), and degree (for polynomial). Parallel processing was employed to accelerate computation.

Model Evaluation:
Models were evaluated based on training error, test error, and classification metrics: accuracy, precision, recall, and F1-score. Comparative analysis was performed between baseline and tuned models to assess improvements.

COMPUTATIONAL RESULTS

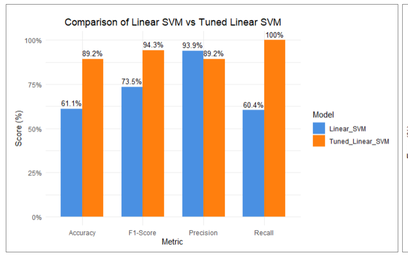

Figure 1. Performance Comparison of Baseline and Tuned Linear SVM Models

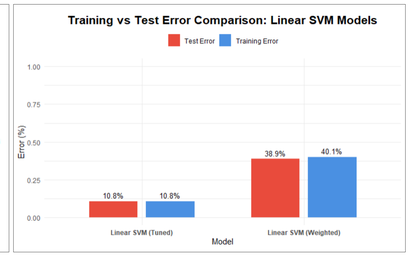

Figure 2. Training and Test Error Comparison for Linear SVM Models

Figures 1 and 2 demonstrate that tuning the Linear SVM significantly enhanced model performance and generalization. As shown in Figure 1, accuracy increased from 61.1% to 89.2%, and the F1-Score rose from 73.5% to 94.3%. Although precision slightly decreased from 93.9% to 89.2%, recall improved dramatically from 60.4% to 100%, making the tuned model particularly effective at detecting positive cases. Additionally, Figure 2 shows that the Tuned Linear SVM achieved much lower training and test errors (both 10.8%) compared to the Weighted Linear SVM (40.1% training error and 38.9% test error), indicating that tuning substantially reduced overfitting and improved the model's generalization capability.

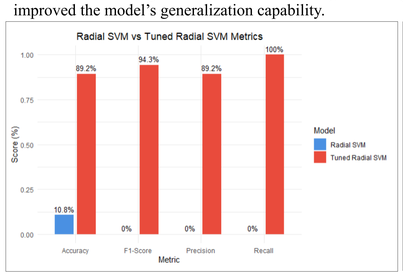

Figure 3. Performance Comparison of Baseline and Tuned Radial SVM Models

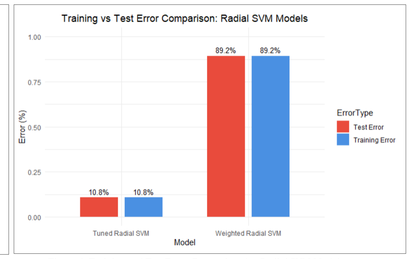

Figure 4. Training and Test Error Comparison for Radial SVM Models

Figures 3 and 4 show that tuning the Radial SVM significantly improved performance and reduced errors. The Tuned Radial SVM achieved 89.2% accuracy, 94.3% F1-Score, 89.2% precision, and 100% recall, while the original model performed poorly across all metrics. Additionally, training and test errors dropped to 10.8%, compared to 89.2% for the Weighted Radial SVM. Finally, tuning greatly enhanced predictive accuracy and generalization.

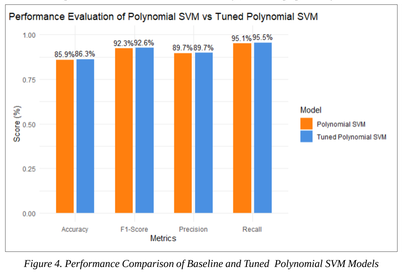

Figure 4. Performance Comparison of Baseline and Tuned Polynomial SVM Models

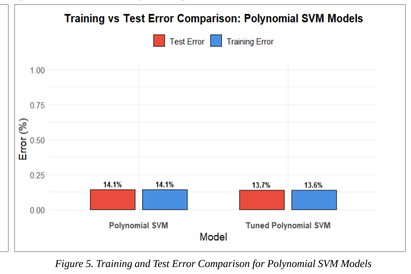

Figure 5. Training and Test Error Comparison for Polynomial SVM Models

Figures 3 and 4 show that tuning the Radial SVM significantly improved performance and reduced errors. The Tuned Radial SVM achieved 89.2% accuracy, 94.3% F1-Score, 89.2% precision, and 100% recall, while the original model performed poorly across all metrics. Additionally, training and test errors dropped to 10.8%, compared to 89.2% for the Weighted Radial SVM. Overall, tuning greatly enhanced predictive accuracy and generalization.

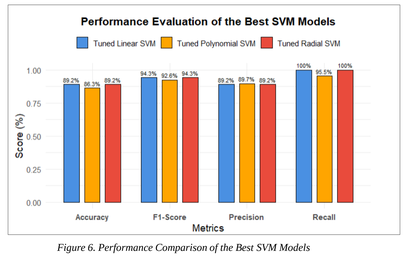

Figure 6. Performance Comparison of the Best SVM Models

Figure 7 presents a comparative evaluation of the best-tuned SVM models. Both the Tuned Linear SVM and Tuned Radial SVM achieved the highest performance, with **89.2% accuracy**, **94.3% F1-Score**, **89.2% precision**, and **100% recall**, demonstrating strong predictive power and perfect recall. The Tuned Polynomial SVM also performed competitively, with slightly lower accuracy (**86.3%**) and recall (**95.5%**) but comparable precision and F1-Score. Conclusion, the Tuned Linear and Radial SVM models showed the most balanced and robust performance across all metrics, making them the most reliable classifiers in this evaluation.

DISCUSSION

This study demonstrated that Support Vector Machines (SVM) with linear, radial basis function (RBF), and polynomial kernels are effective tools for predicting diabetes risk based on demographic, behavioral, and health-related variables. Baseline models exhibited reasonable predictive performance; however, hyperparameter tuning consistently improved model generalization and classification accuracy across all kernel types.

The Linear SVM achieved strong baseline precision, effectively identifying positive cases with minimal false positives. The tuned RBF SVM achieved a balanced trade-off between recall and precision, successfully capturing complex, non-linear relationships within the data. Polynomial SVM performance improved after tuning but remained highly sensitive to parameter adjustments, reflecting its increased model complexity and risk of overfitting.

Given the inherent class imbalance in the dataset, class weighting strategies were applied during model training to prevent bias toward the majority class, contributing to the improved recall and F1-scores observed across tuned models. Hyperparameter tuning, conducted via grid search and cross-validation, was computationally intensive, particularly for the polynomial and RBF kernels. To mitigate extended CPU runtimes, parallel processing techniques were employed, significantly expediting the model optimization process.

These findings highlight the importance of systematic feature selection, class imbalance handling, and model tuning in SVM-based health prediction tasks. High recall rates across tuned models are especially critical in public health contexts, where missing at-risk individuals can have significant consequences. Overall, while SVMs are robust classifiers, careful kernel selection and hyperparameter optimization are essential to fully realize their potential in complex biomedical applications.

CONCLUSION

Preventing chronic diseases like diabetes is increasingly vital in today's public health landscape. Early identification of at-risk individuals enables timely interventions that can reduce disease burden and improve quality of life. Machine learning methods, particularly Support Vector Machines (SVM), provide powerful tools for predicting diabetes risk based on demographic, behavioral, and health-related factors.

This study showed that SVM models with linear, radial basis function (RBF), and polynomial kernels effectively captured the complex relationships influencing diabetes risk. Systematic feature selection identified key predictors such as hours worked per week, education level, poverty status, age, and vigorous physical activity, emphasizing important lifestyle and socioeconomic factors.

Hyperparameter tuning and class imbalance handling further enhanced model performance, particularly improving recall—critical in minimizing missed high-risk cases. Although tuning required significant computational resources, parallel processing mitigated long runtimes and accelerated optimization.

In conclusion, the findings highlight the value of SVM-based predictive modeling for proactive disease prevention. This study offers a practical foundation for implementing data-driven approaches to support early diabetes detection and guide 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/