Predicting Diabetes Risk Using Demographics and Health Behaviors: A Support Vector Machine Approach

Exploring how lifestyle and socioeconomic factors shape diabetes risk through machine learning models

Ankit Bisleri

Introduction

Diabetes is a major chronic disease influenced by health behaviors and demographics. Machine learning offers tools to predict disease risk, potentially informing prevention efforts. In this study, we use Support Vector Machines (SVMs) to classify individuals diabetes status based on lifestyle factors and demographic information.

Methodology

- Data Preparation:
  - Selected 10 demographic and health behavior features from the 2022 NHIS survey dataset.
  - Missing and invalid entries were removed.
- Feature Scaling:
  - Applied StandardScaler to normalize predictors.
- Class Balancing:
  - Used SMOTE (Synthetic Minority Oversampling Technique) to balance diabetic and non-diabetic cases in the training set.
- Model Training:
  - Linear SVM (linear kernel)
  - RBF SVM (Radial Basis Function kernel)
  - Polynomial SVM (polynomial kernel)
- Hyperparameter Tuning:
  - Performed GridSearchCV with 5-fold cross-validation to optimize:
  - C (penalty parameter)
  - Gamma (for RBF kernel)
  - Degree (for polynomial kernel)
- Model Evaluation:
  - Accuracy
  - Recall
  - ROC Curve Analysis

Technical Background

Support Vector Machines (SVMs) are classification models that find the hyperplane maximizing separation between classes. When data is non-linearly separable, kernel methods (linear, RBF, polynomial) map inputs into higher dimensions.

The SVM optimization problem is:

$$\min_{w} \left(\frac{1}{2} \|w\|^2 + C \sum \xi_i \right)$$

$$\text{where} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0$$

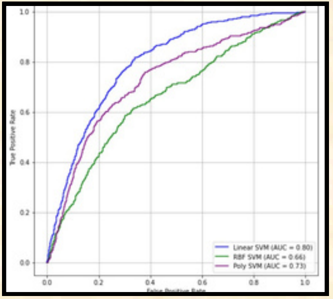
Key hyperparameters include:

- C: Tradeoff between margin size and misclassification penalty
- Gamma: Defines RBF kernel flexibility
- Degree: Controls complexity in polynomial kernels

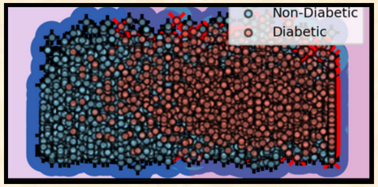
Results

Model	Accuracy	Recall	F1 Score	ROC-AUC
Linear SVM	~68.6%	~79%	~34%	~0.80
Polynomial SVM	~68.6%	~67%	~30%	~0.73
Radial SVM	~82.4%	~27%	~23%	~0.66

- Highest Accuracy: Radial SVM (~82.4%), but low recall for diabetic cases.
- Best Recall for Diabetics: Linear SVM (~79%) — critical for healthcare prediction.
- Best Overall Discrimination: Linear SVM had highest ROC-AUC (~0.80).



ROC curve comparison shows that the Linear SVM achieved the highest AUC (~0.80), indicating the best discrimination ability among tuned models.



The Linear SVM decision boundary shows separation between diabetic and non-diabetic individuals based on Age and BMI.

Discussion

- Age and BMI were the strongest predictors of diabetes, highlighting the impact of lifestyle and demographic factors on disease risk.

Conclusion

- Support Vector Machines effectively classified diabetes cases, suggesting that early interventions targeting weight and aging populations are critical.

Citation

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