Predicting Diabetes Risk from Health Behaviors Using Support Vector Machines

Khaja Moinuddin Mohammed, Seattle University
DATA 5322 Statistical Machine Learning 2

1. Background & Objective

Background:

- Diabetes is a major public health concern affecting millions of adults worldwide.
- Early identification of at-risk individuals can help prevent complications and reduce healthcare costs.
- Survey data (e.g., NHIS) and machine learning methods offer new opportunities for risk prediction at the population level.
- This project applies Support Vector Machines (SVMs) with linear, polynomial, and radial (RBF) kernels to predict diabetes status using demographic and lifestyle variables.

Objectives:

- Evaluate how health behaviors and demographic factors relate to diabetes risk.
- Compare the predictive performance and clinical utility of different SVM kernels.
- Provide insights for public health screening and prevention strategies.

2. Data & Variables

- **Dataset: National Health Interview Survey (NHIS, IPUMS, 2022)**
  - Full Sample: 27,651 Sample Adults
  - Analytic Test Sample: n = 192
- **Variables Used:**
  - AGE (years): Uniformly distributed across adults 18–84, ensuring all age groups are represented.
  - BMICALC (BMI, kg/m²): Approximately normal, centered in the overweight/obese range.
  - HRSLEEP (average hours sleep/night): Roughly normal, most respondents sleep 6–8 hours per night.
  - CIGDAYMO (cigarettes/month): Right-skewed, most report low or no use.
  - BMI_AGE_INTERACTION (BMI × Age): Captures combined risk.
  - Target: Doctor-diagnosed diabetes (0 = No, 1 = Yes)
- **Exploratory Data Analysis:**
  - Diabetes prevalence in the analytic sample is 9.4%, closely matching national U.S. estimates.
  - No strong linear correlations between individual predictors and diabetes, highlighting the need for multivariate or non-linear modeling.
  - The BMI × Age interaction term is strongly right-skewed, emphasizing compounding risk for older adults with high BMI.

3. Methods

Preprocessing:

- Dropped missing values
- Standardized features
- Addressed class imbalance using SMOTE
- Feature engineering: BMI × Age interaction

Modeling:

- SVMs trained with three kernels: Linear, Polynomial (degree=5), RBF
- Hyperparameter tuning for C, gamma, degree, class_weight (GridSearchCV)

Evaluation Metrics:

- Accuracy, Recall, Precision, F1-score, AUC
- Confusion Matrix, ROC Curve

4. Technical Background

- **Support Vector Machines (SVMs):**
  - SVMs find the hyperplane that best separates classes by maximizing the margin.
  - The kernel trick enables SVMs to handle non-linear data by mapping it into higher-dimensional space.
- **Kernels:**
  - Linear: Best for data separable by a straight line.
  - Polynomial: Captures complex, polynomial relationships.
  - RBF (Radial): Captures highly nonlinear patterns using Gaussian functions.
- **Key Concepts:**
  - Soft Margin SVM: Allows some misclassifications for better generalization.
  - Tuning Parameters:
    - C: Controls margin vs. classification error
    - Gamma: Controls boundary curvature (RBF)
    - Degree: Sets complexity (Polynomial)

8. References & Acknowledgments

- Blewett LA, Rivera Drew JA, King ML, Williams KCW. IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset].
- Scikit-learn Developers. Scikit-learn: Machine Learning in Python.

5. Results

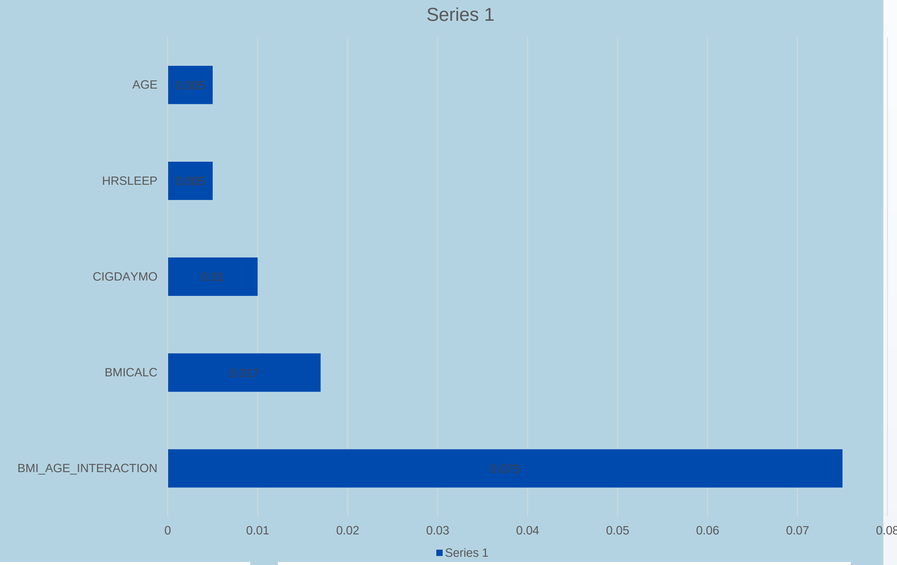
a. Key Performance Table

Metric	Linear SVM	oly SVM	RBF SVM
Accuracy	65%	80%	84%
Recall (diabetes)	78%	28%	33%
Precision (diabetes)	18%	16%	24%
AUC	0.77	0.65	0.65

- **Linear SVM:** Highest recall for diabetes (78%), best for screening but many false positives.
- **Polynomial SVM:** Higher accuracy, much lower recall for diabetes.
- **RBF SVM**: Highest accuracy, fewest false positives, but recall for diabetes remains low (33%).

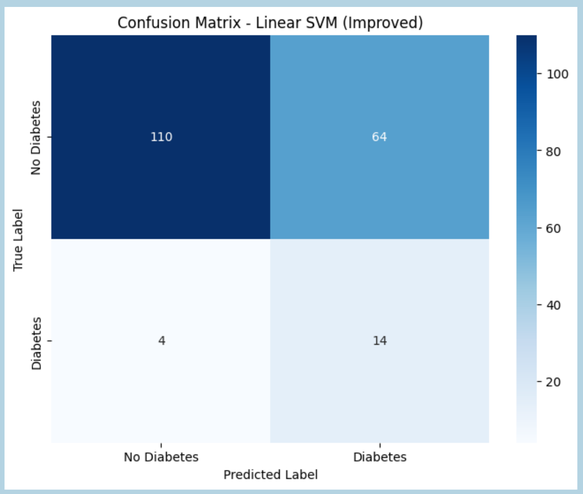
b. **Feature Importance (Linear SVM) Barplot :**

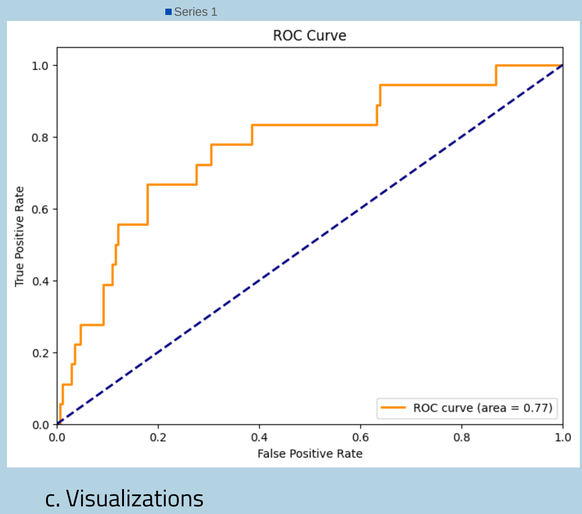
BMI × Age interaction is the most important predictor, followed by BMI. Age alone is least important, showing the value of engineered features.



Visualizations

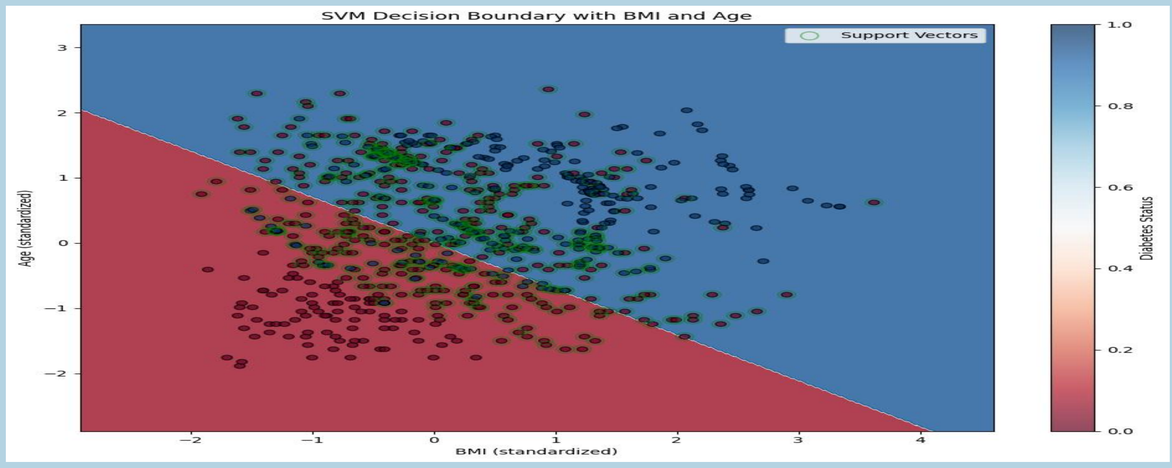
- Confusion Matrix (Linear SVM)
- ROC Curve (Linear SVM)
- Decision Boundary Plot (BMI vs. Age, Linear SVM)





c. Visualizations
Confusion Matrix

Decision Boundary Plot (BMI vs. Age, Linear SVM,)



6. Interpretation & Discussion

- **Key Predictors:**
  - The BMI × Age interaction is the strongest risk factor, indicating older adults with higher BMI are at greatest risk. This aligns with clinical knowledge and suggests interventions should focus on weight management, especially for middle-aged and older adults.
  - Demographic & Social Factors:
  - Age and BMI are well-established, easily measured risk factors. Smoking and sleep had less impact in this dataset, possibly due to NHIS variable limitations.
- **Policy Recommendations:**
  - Prioritize BMI screening and weight management programs for adults over 40.
  - Use SVM-based risk prediction to identify high-risk individuals for early intervention.
  - Encourage more detailed behavioral and social data collection (e.g., sleep quality, smoking history, food access) in future surveys.
- **Study Impact:**

Demonstrates that simple, interpretable models using basic survey data can effectively identify high-risk groups for diabetes. The approach can be adapted to other chronic diseases and inform policy and prevention strategies at the community level.

7. Conclusion

- Linear SVM is preferred for diabetes screening due to its high recall (78%), minimizing missed cases. This exceeds the sensitivity of traditional tools like HbA1c (50–59%), highlighting its value for public health screening.
- BMI × Age interaction is the most important predictor, emphasizing the need for age-specific weight management policies targeting older adults with high BMI.
- For screening, sensitivity (recall) is more critical than precision or overall accuracy.
- **Policy Suggestion:** Implement targeted screening and prevention programs for high-BMI adults over 40, and advocate for data modernization to capture granular behavioral metrics.
- **Limitations:** Limited feature set (lacks socioeconomic variables) and high false positives.
- **Future Work:** Incorporate socioeconomic factors (income, education) and test ensemble models.