Predicting Diabetes Risk from Health Behaviors Using Support Vector Machines

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

1. Background & Objective

Maintaining good health is influenced by behavioral and demographic factors such as BMI, age, sleep, and smoking.
Diabetes is a major public health concern, and early identification of at-risk individuals can improve prevention and intervention.
This study uses Support Vector Machines (SVMs) to predict diabetes risk from NHIS survey data.
Research Questions:

- Can demographic and behavioral factors predict diabetes?
- Which SVM kernel (Linear, Polynomial, RBF) provides the best classification for screening?

Objective:
Model diabetes risk using SVMs, evaluate different kernel types, and recommend strategies based on model insights.

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)
- BMICALC (BMI, kg/m²)
- HRSLEEP (average hours sleep/night)
- CIGDAYMO (cigarettes/month)
- BMI_AGE_INTERACTION (BMI × Age)
- Target: Doctor-diagnosed diabetes (0 = No, 1 = Yes)

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, Threshold Sweep

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.

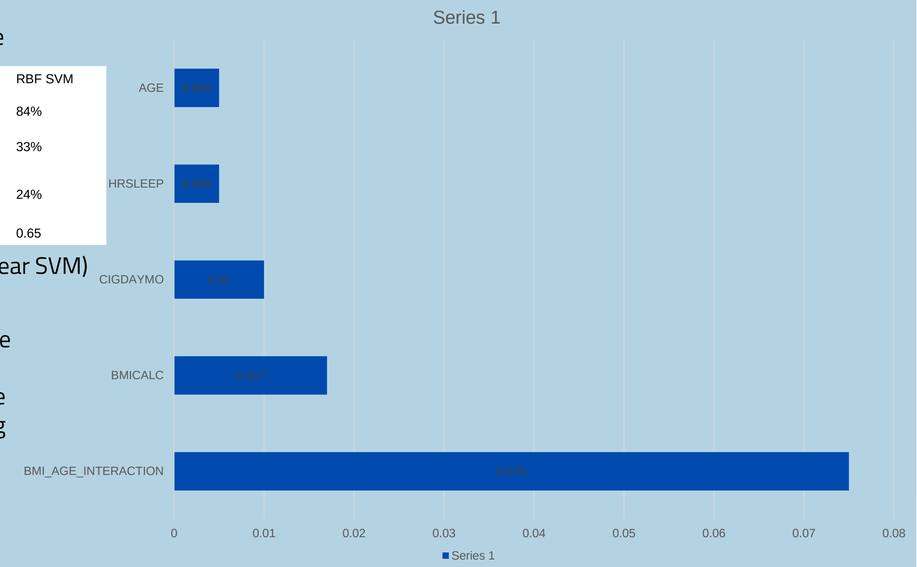
5. Results

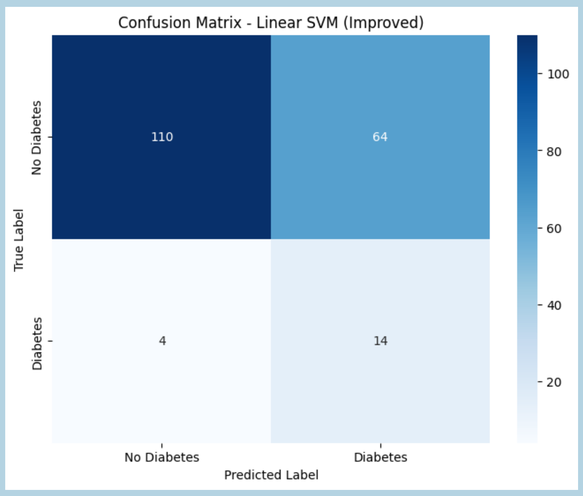
a. Key Performance Table

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

b. Feature Importance (Linear SVM) Barplot :

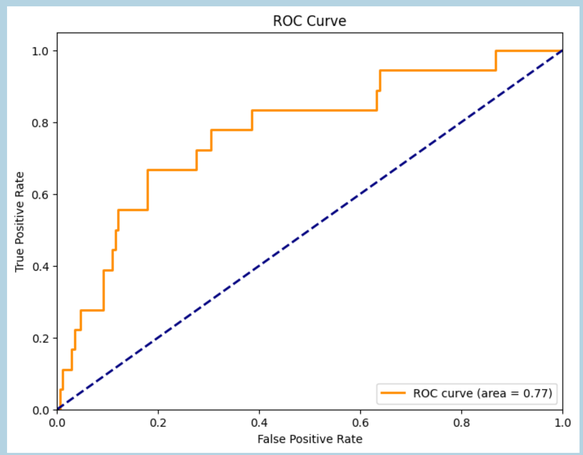
BMI × Age interaction is the most important predictor, followed by BMI. Age alone is least important, showing the value of engineered features.



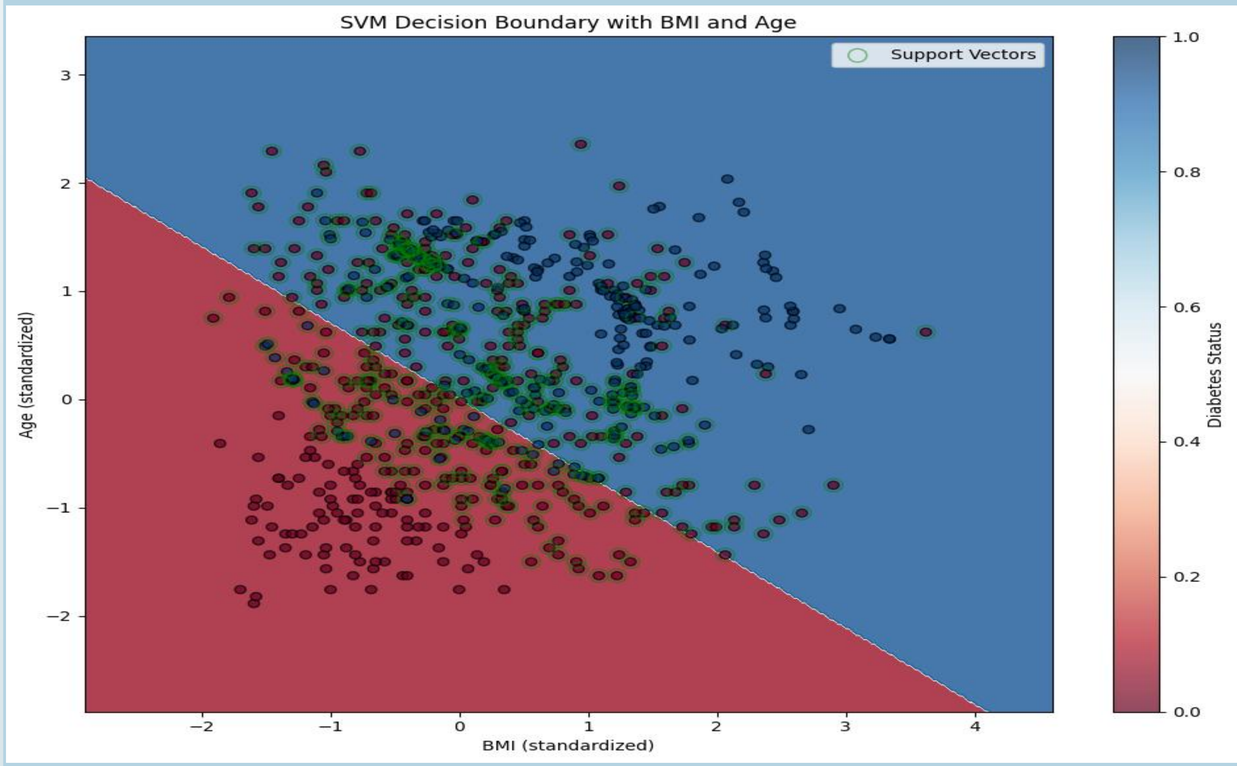


Decision Boundary Plot (BMI vs. Age, Linear SVM,)

c. Visualizations
Confusion Matrix



ROC Curve



6. Interpretation & Discussion

Findings:

- Linear SVM achieves the highest recall for diabetes (78%), making it best for screening.
- RBF and polynomial kernels offer higher accuracy and fewer false positives, but miss more diabetes cases.
- BMI × Age interaction is the strongest risk factor.

Interpretation:

- High recall is critical for screening and early intervention.
- Feature engineering (BMI × Age) greatly boosts predictive power.
- Nonlinear kernels (RBF, Polynomial) capture complex effects but do not improve recall for diabetes in this dataset.

7. Conclusion

- Support Vector Machines, especially with a linear kernel, are effective for diabetes risk screening due to high recall.
- Feature engineering, particularly BMI × Age interaction, is crucial.
- For screening, sensitivity (recall) is more important than overall accuracy or precision.
- Kernel choice affects the tradeoff between sensitivity and specificity.