IDENTIFYING THE RISK OF DIABETES FROM BEHAVIORAL AND DEMOGRAPHIC FACTORS USING SUPPORT VECTOR MACHINES

LAVANYA BUNADRI

INTRODUCTION

- This study delves into the Support Vector Machine models to predict the presence of diseases using data from the National Health Interview Survey (NHIS) 2022, accessed through IPUMS Health Surveys.
- Specifically, this analysis aims to determine SVM models to predict the presence of individuals as having ever been diagnosed with diabetes based on a selected set of demographic and health behavior variables, including age, Body Mass Index (BMI), hours of sleep, sex, and frequency of soda consumption.
- Three types of kernel functions were employed- linear, radial, and polynomial to build support vector models, and their performance is assessed using a subset of the data, this aims to uncover meaningful insights and predictive patterns.
- The methodology involves cleaning and structuring the data for optimal analysis, followed by employing SVM algorithms to explore the most effective method for predicting diabetes status.

THEORETICAL BACKGROUND

- The Support Vector Machine (SVM) is a versatile machine learning tool primarily used for classification tasks, although it can also handle regression.
- It helps to find the optimal hyperplane that best separates data into distinct classes. This hyperplane maximizes the margin between the classes, improving the model's ability to generalize to new data.

KERNALS: Kernels are used to transform input data into a higher-dimensional space where a linear decision boundary can be found to separate different classes of data points.

➢ **LINEAR KERNAL:** It's a simple and efficient approach in SVMs that works best when data is linearly separable. It calculates the similarity between points using the dot product and is ideal for large datasets due to its low computational cost.

$$K(X,Y)=(X^{T})Y$$

➢ **RADIAL KERNAL:** The Radial Basis Function (RBF) kernel transforms data into a higher-dimensional space, enabling SVMs to classify non-linear patterns effectively. It is highly flexible and performs well with a wide range of complex datasets.

$$K(X,Y)=\exp((\|X-Y\|^{2})/2(\sigma^{2}))\text{ i.e. }\exp(-\gamma\cdot(\|X-Y\|^{2})),\gamma>0$$

➢ **POLYNOMIAL KERNAL:** The Polynomial kernel allows SVMs to handle data that isn't linearly separable by mapping it into a higher-dimensional space using polynomial functions. It captures complex relationships between features, enabling more flexible decision boundaries for accurate classification.

$$K(X,Y)=(\gamma.(X^{T})Y+r)^{d},\gamma>0$$

SUPPORT VECTOR CLASSIFIER

The Support Vector Classifier (SVC) is a binary classification algorithm that finds the best line or hyperplane to separate data into two classes. It maximizes the margin between classes for better generalization and uses kernel functions to handle complex data patterns.

MULTI-CLASS CLASSIFICATION

Support Vector Machine (SVM) for Multiclass Classification: SVM can be extended to handle multiclass classification tasks through techniques like One-vs-One or One-vs-All

• **One-vs-One**: SVM creates binary classifiers for each pair of classes and combines their votes to determine the final prediction.

• **One-vs-All**: SVM trains binary classifiers for each class against the rest and selects the class with the highest confidence score.

TUNING PARAMETERS: Tuning parameters in machine learning, particularly in algorithms like Support Vector Machines (SVM), involve adjusting the model's settings to optimize its performance.

- **COST**

  The cost parameter (C) in SVM controls the trade-off between maximizing the margin and minimizing classification errors. A high C creates a tighter fit to the training data, while a low C allows for a simpler boundary with some misclassifications.
- **GAMMA**

  In SVM with an RBF kernel, gamma controls how far the influence of a single training point reaches. High gamma values create more complex, flexible boundaries, while low gamma values produce smoother boundaries that may underfit the data.
- **DEGREE**

  In SVM with a polynomial kernel, the degree parameter determines the complexity of the polynomial function used to separate classes. Higher degrees capture more complex patterns but risk overfitting, while lower degree values result in simpler decision boundaries but may struggle to capture intricate patterns in the data.

METHODOLOGY

➢ **Data Preparation:**
- **Data Source:** 2022 National Health Interview Survey (IPUMS).
- **Target Variable:** Diabetes status (DIABETICEV).
- **Selected Predictors: Demographics:** Age (AGE), Sex (SEX),Health Behaviors: Body Mass Index (BMICALC), Hours of Sleep (HRSLEEP), Soda Consumption Frequency (SODAPNO)

➢ **Data Cleaning:**
- Removed invalid entries: Dropped values ≥996 used for "Don't know and refused columns in survey AGE, BMICALC, HRSLEEP, and SODAPNO
- Removed specific codes 97, 98, 99 from HRSLEEP
- Re-coded categorical variables: SEX: 1 = Male → 0, 2 = Female → 1 and DIABETICEV: 1 = No Diabetes → 0, 2 = Yes Diabetes → 1
- Removed all rows with missing values using dropna()

➢ **Class Balancing:** The DIABETICEV target variable was imbalanced so performed under sampling to the majority class with no diabetes to match the number of instances in the minority class yes diabetes.

➢ **Data Splitting:** The balanced data was split into training 70% and testing 30% sets using random sampling.

➢ **Feature Scaling:** StandardScaler was used to scale the predictor variables in the training and testing set.

➢ **Model Implementation and Tuning:**The best hyperparameter combination for each kernel was selected based on the accuracy achieved on the test set after training with those parameters on the training set.

Model	Kernel Type	Hyperparameters Tried
SVM Linear	linear	C = [0.01, 0.1, 1, 10, 100]
SVM RBF	rbf	C = [0.01, 0.1, 1, 10, 100]; Gamma = 'scale'
SVM Polynomial	poly	C = [0.1, 1, 10]; Gamma = [0.1, 1]; Degree = [2, 3]

➢ **Evaluation & Comparison**
- **Metrics Used:**Accuracy,Precision,Recall,F1-Score and Confusion Matrix were used for evaluation.
- **Comparison Strategy:** Model performance across different SVM kernels (Linear, RBF, Polynomial) was directly compared using these evaluation metrics.Best models were selected based on achieving higher test accuracy while balancing precision and recall.

RESULTS

➢ **Performance Evaluation**

Before Tuning

Model	Training Accuracy	Test Accuracy	Training Error	Test Error
SVM - Linear	0.7069	0.6886	0.2931	0.3114
SVM - Radial	0.7180	0.7013	0.2820	0.2987
SVM- Polynomial	0.7029	0.6899	0.2971	0.3100

After Tuning

Model	Training Accuracy	Test Accuracy	Training Error	Test Error
SVM - Linear	0.7060	0.6886	0.2940	0.3114
SVM - Radial	0.7111	0.7019	0.2889	0.2981
SVM - Polynomial	0.7043	0.6899	0.2957	0.3100

Detailed Model Evaluation Using Precision, Recall, and F1-Score (After-Tuning)

Model(Tuned)	Precision (Yes Diabetes)	Recall (Yes Diabetes)	F1-Score (Yes Diabetes)
SVM Linear	0.66	0.76	0.71
SVM Radial	0.66	0.80	0.72
SVM Polynomial	0.66	0.74	0.70

- Among the tuned models, SVM with Radial kernel achieved the highest recall 80% for predicting yes Diabetes cases, indicating it was best at correctly identifying diabetic individuals.

➢ **Confusion Matrix:** **Linear Kernel -Test Set After Tuning**

	Predicted No Diabetes	Predicted Yes Diabetes
Actual No Diabetes	475 (TN)	294 (FP)
Actual Yes Diabetes	174 (FN)	560 (TP)

- The linear kernel model achieved a good balance between identifying diabetic and non-diabetic individuals.
- It correctly predicted 560 individuals with diabetes and 475 without. However, it also missed 174 diabetics (false negatives) and incorrectly labeled 294 non-diabetics as diabetic (false positives).

Radial Kernel -Test Set After Tuning

	Predicted No Diabetes	Predicted Yes Diabetes
Actual No Diabetes	469 (TN)	300 (FP)
Actual Yes Diabetes	148 (FN)	586 (TP)

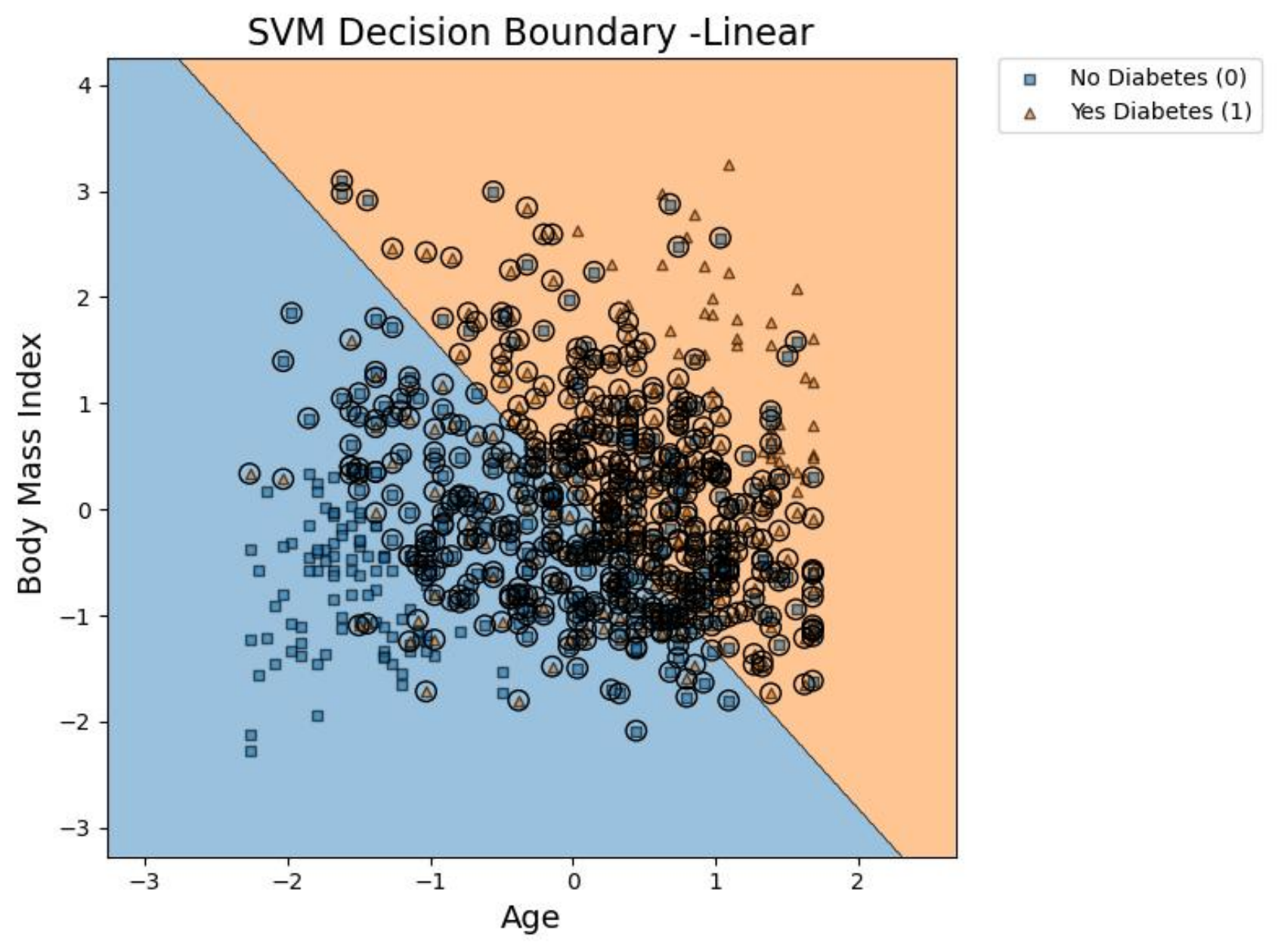
- The RBF kernel showed the strongest performance in correctly identifying diabetic individuals, with 586 true positives and only 148 false negatives the lowest among all models.
- This improved sensitivity came at the cost of slightly more false positives 300 non-diabetics misclassified as diabetic. This model is highly effective at detecting diabetes.

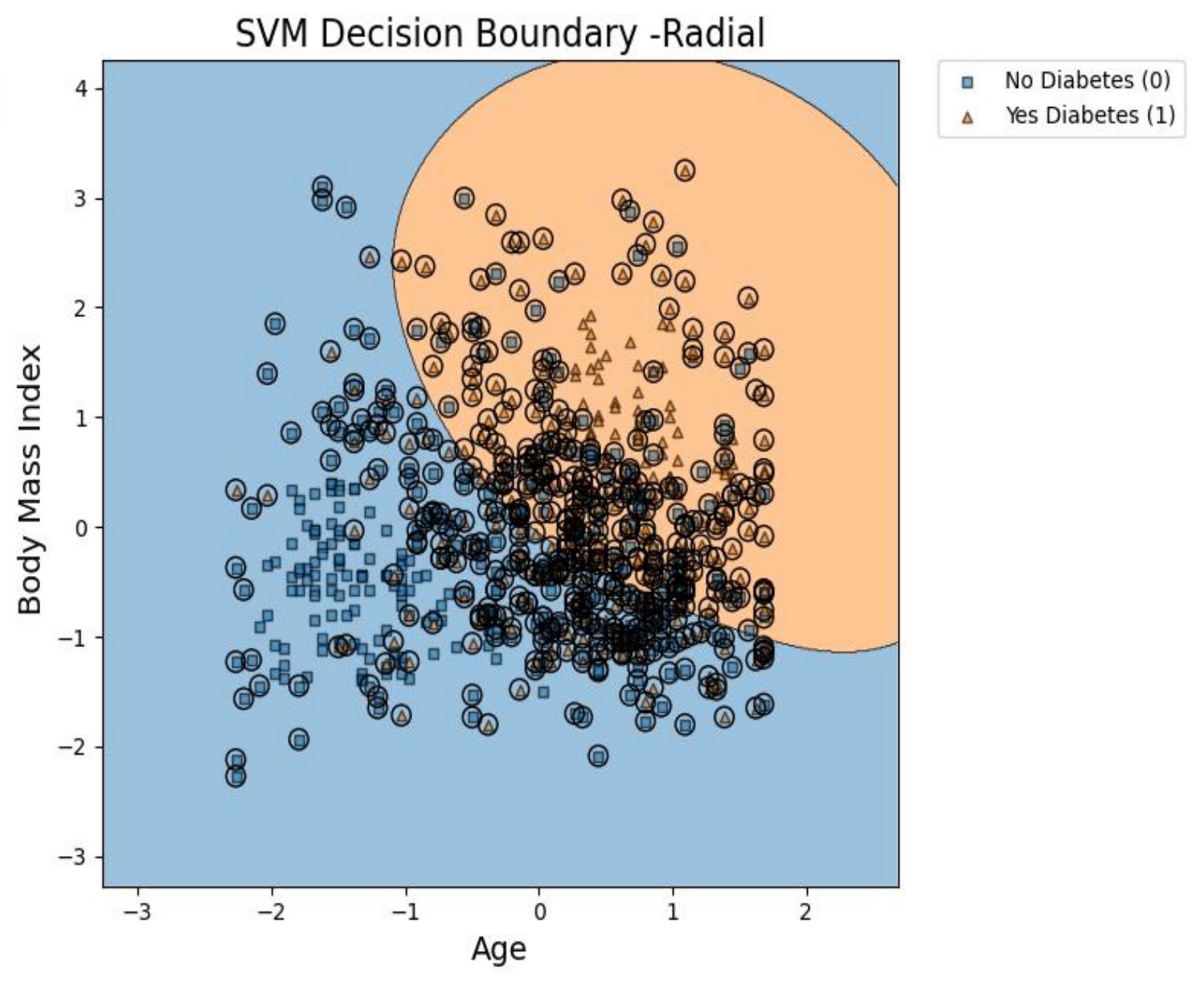
Polynomial Kernel -Test Set After Tuning

	Predicted No Diabetes	Predicted Yes Diabetes
Actual No Diabetes	494 (TN)	275 (FP)
Actual Yes Diabetes	191 (FN)	543 (TP)

- The polynomial kernel it had the fewest false positives 275, means it was not better at misclassifying non-diabetic individuals it missed 191 actual diabetic cases, the highest false negatives of the three models. While its precision was strong, its lower recall shows it may under-detect diabetes.

➢ **Visualizing SVM decision boundary on two strong predictor variables.**



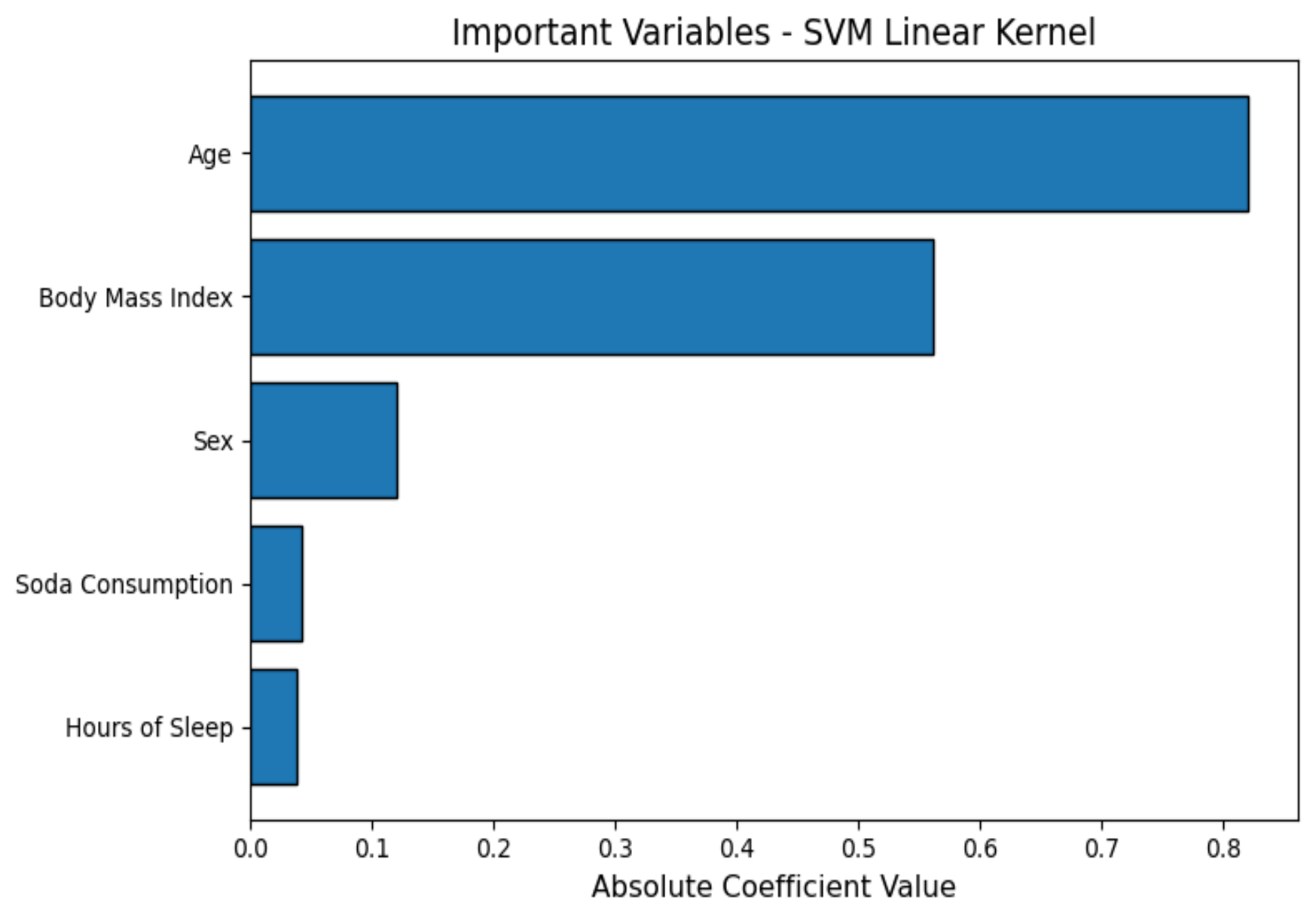


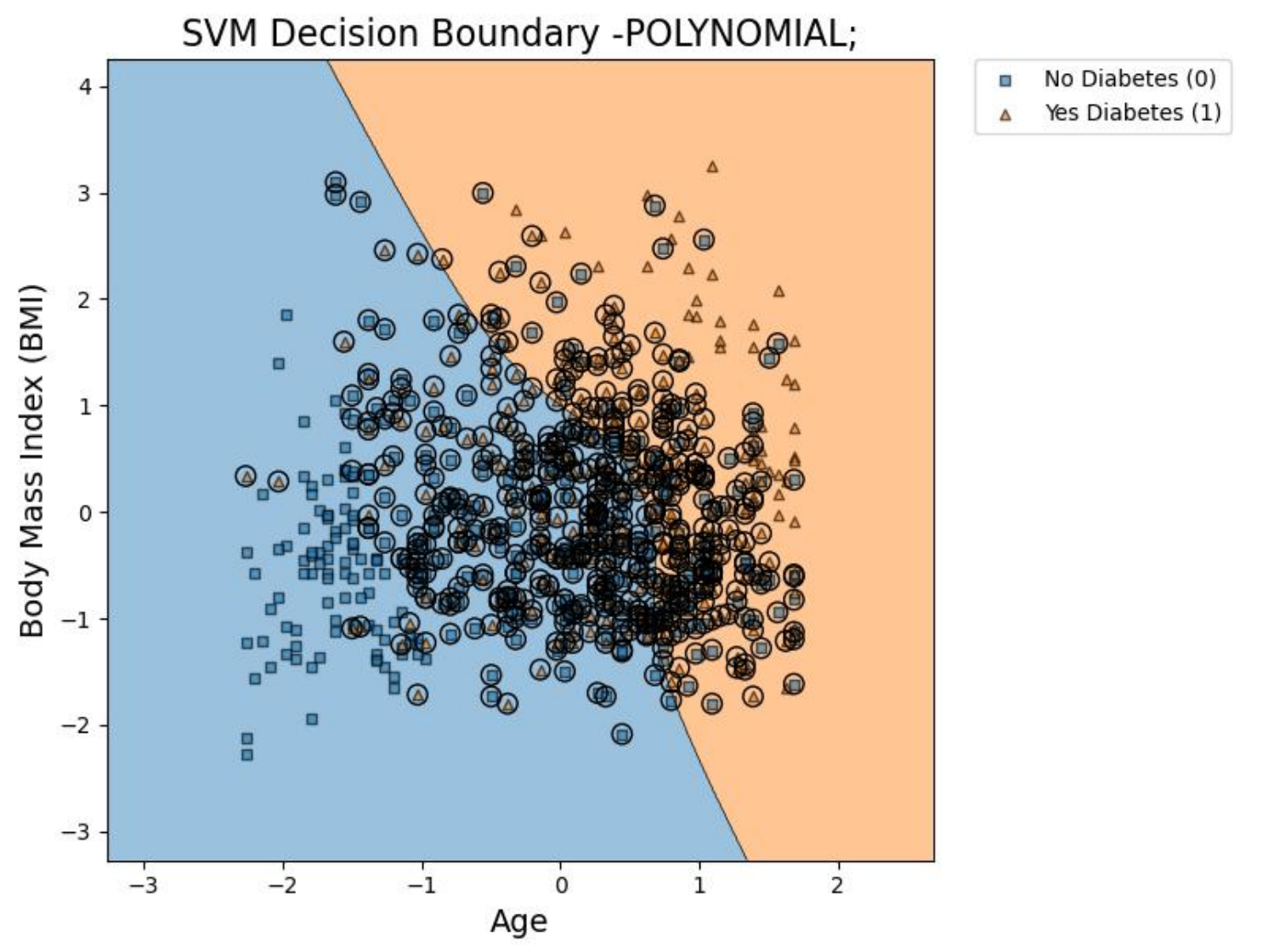
DISCUSSION

➢ **Model Comparison and Performance:**

Model	Test Accuracy	Test Error	Recall (Yes Diabetes)	Precision (Yes Diabetes)	F1-Score (Yes Diabetes)	False Positives	False Negatives
SVM - Linear	68.9%	31.1%	76%	66%	71%	294	174
SVM - Radial (RBF)	70.2%	29.8%	80%	66%	72%	300	148
SVM - Polynomial	69.0%	31.0%	74%	66%	70%	275	191

➢ **variables seemed to be strong predictors of disease:**





➢ **How Behavior and Demographics Impact Diabetes Prediction**

- BMI and Age were the most impactful predictors across all SVM models. Higher age and BMI values were strongly associated with the diabetic class region. This supports medical understanding that aging and obesity increase diabetes risk.
- Soda Consumption showed limited but visible influence. It indicates that dietary habits may play a role in metabolic health and chronic disease.
- Hours of Sleep had a low influence on the model's decision boundary. May reflect inconsistent effects of sleep on health or measurement noise.
- Sex (a demographic factor) had moderate predictive power. Aligns with known biological and behavioral differences in diabetes risk across genders.

Overall, the plots suggest that clinical metrics like Age, BMI and demographics like Sex are more informative than behavioral habits like soda use or sleep in predicting diabetes from this dataset.

CONCLUSION

- Body Mass Index and age were identified as significant factors associated with a history of diabetes, consistent with established health knowledge.
- Models achieved about 70% accuracy, showing that machine learning can help identify diabetes risk early
- The Radial SVM model performed best overall, achieving the highest recall and strongest ability to correctly detect diabetic cases.
- Policymakers could consider strengthening initiatives that support healthy weight management through education and access to resources.
- Future work should explore including more variables and advanced modeling techniques to improve prediction accuracy.
- These findings can contribute to improved prevention and screening strategies for diabetes