Diabetic Classification of people over age 50 with SVM

Prasanth Gururaj

Introduction

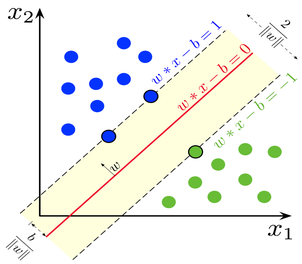
Using data from the 2022 National Health Interview Survey (NHIS), which collects health, demographic, and socioeconomic information from U.S. households, this project focuses on developing a predictive model to diagnose diabetes among individuals aged over 50.

I selected predictors such as BMI (BMICALC), daily fruit intake (FRUTNO), daily salad intake (SALADSNO), minutes of moderate physical activity per day (MOD10DMIN), average hours of sleep per night (HRSLEEP), weekly pizza consumption (PIZZANO), and weekly fries consumption (FRIESPNO), based on their relevance to lifestyle behaviors affecting diabetes risk.

To build the model, I applied a Support Vector Machine (SVM) approach, experimenting with different kernels (linear, radial, and polynomial) and tuning parameters like the cost function, gamma, and coefficients to optimize performance. The aim of this project is to better understand how lifestyle factors contribute to diabetes risk in older adults and to support efforts in early detection and prevention through reliable predictive modeling.

Theoretical Background

Support Vector Machines (SVMs) are supervised learning models used mainly for classification. They find the optimal hyperplane that separates classes while maximizing the margin between the closest points, called support vectors. A hyperplane is a flat boundary that divides the feature space into different classes.



The decision boundary is given by:
wTx + b =0
where w is the weight vector,
b is the bias term.

Key Parameters :
Cost (C)
Kernel Functions

Cost (C): Controls the trade-off between a wide margin and correct classification.
- •High C leads to a complex boundary and may overfit.
- •Low C allows a simpler, more generalizable model.

Kernel Functions: Allow SVMs to handle different types of data by transforming it into higher dimensions.
- **•Linear Kernel:** Used for data that is already linearly separable. It creates a straight-line (or flat) decision boundary.
- **•Polynomial Kernel:** Used for data where classes are separated by curved boundaries. It creates more flexible, curved decision boundaries based on the polynomial degree.
- **•Radial Kernel:** Used for complex, non-linear datasets. It creates circular or irregular decision boundaries by measuring distance between points with a gamma parameter.

degree (Polynomial Kernel): Controls the flexibility of the decision boundary.
- •Low degree (e.g., 2): Less complex, smoother curves.
- •High degree (e.g., 4): More complex, wavy decision boundaries.

gamma (Radial Basis Function and Polynomial Kernel): Controls how far the influence of a single training point reaches.
- •Low gamma (e.g., 0.1): Points have a wider influence, creating smoother decision boundaries.
- •High gamma (e.g., 10): Points have a narrow influence, leading to very tight and complex boundaries that may overfit.

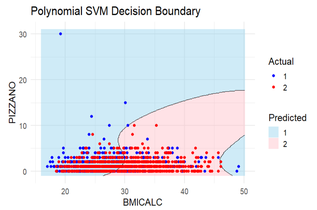
Methodology

Data Cleaning and Preprocessing:
The NHIS 2022 dataset was filtered to include adults (ASTATFLG = 1 or 6) and individuals aged 50 years and above. Selected predictors included BMICALC (BMI), FRUTNO (fruit intake), SALADSNO (salad intake), MOD10DMIN (minutes of moderate exercise), HRSLEEP (sleep duration), PIZZANO (pizza consumption), and FRIESPNO (fries consumption).

Invalid entries (e.g., special codes 996–999) were removed, and extreme values were clipped based on survey guidelines. After cleaning, the final dataset consisted of **9,242 individuals**: 87.28% without diabetes and 12.72% with diabetes.

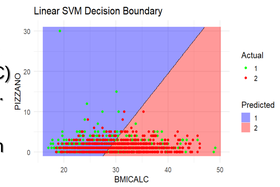
Data Splitting:
The cleaned data was randomly divided using an **80%-20% train-test split**.
- **Training Set:** Used to train and tune the models.
- **Testing Set:** Used for final model evaluation to measure real-world performance

Model Implementation:
Support Vector Machine (SVM) models were built using three different kernel types:
- **Linear Kernel**
- **Radial Basis Function (RBF) Kernel**
- **Polynomial Kernel**



A Polynomial SVM using BMI and Pizza Consumption was plotted to observe the separation of diabetic and non-diabetic individuals.

The Linear SVM model effectively separated diabetic and non-diabetic individuals based on BMI (BMICALC) and Pizza Consumption (PIZZANO).



The decision boundary is linear, with individuals having higher BMI and higher pizza intake more likely classified as diabetic.

Hyperparameter Tuning:
Grid search was performed across a range of hyperparameter values **only on the training data**:
- **Linear SVM:** Cost (C) values tested: 0.01, 0.05, 0.1, 0.5, 5
- **RBF SVM:** Cost (C) values 0.01–5; Gamma values 0.001–2
- **Polynomial SVM:** Cost values 0.01, 0.5, 5; Degrees 2, 3; Coef0 values 0, 2

The best model for each kernel was selected based on highest training accuracy.

Model Evaluation:
Models were evaluated **on the separate test set** using the following metrics:
- **Accuracy** (Correct classifications)
- **Error Rate** (Misclassification rate)
- **Precision** (Positive predictive value)
- **Recall** (Sensitivity)

Confusion matrices were created to further analyze model performance across the diabetic and non-diabetic classes.

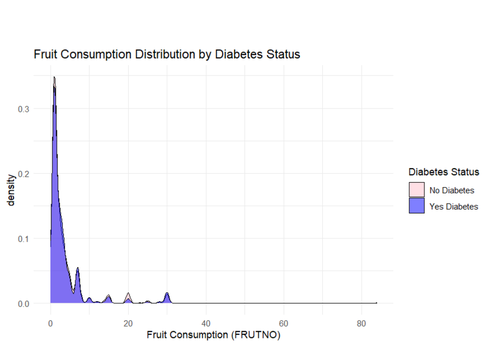
Model	Train_Accuracy	Test_Accuracy	Train_Error	Test_Error	Precision	Recall
Linear SVM	0.6073326	0.6731602	0.3926674	0.3268398	0.9207673	0.6844389
Radial SVM	0.6609989	0.6130952	0.3390011	0.3869048	0.9172862	0.6119033
Polynomial SVM	0.6370882	0.6271645	0.3629118	0.3728355	0.9200000	0.6274024

Discussion

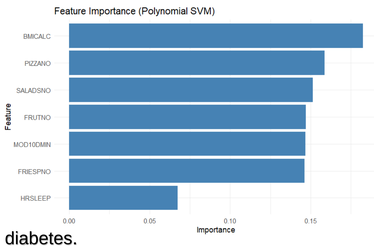
About Dataset
- The dataset closely resembles real-world healthcare data, with **87% non-diabetic** and**13% diabetic** individuals.
  This severe imbalance made it **challenging for SVM models to find a clear hyperplane** for separating the two classes effectively.
- In **initial experiments without down-sampling**, all SVM models (Linear, Radial, Polynomial) achieved **around 92% accuracy**.
  However, this **high accuracy was misleading**, as the models primarily predicted "No Diabetes," failing to learn meaningful patterns for diabetic cases.

Down-sampling
- Down-sampling was applied to balance the diabetic and non-diabetic classes in the training set.
  This adjustment helped models **focus more effectively** on correctly classifying diabetic individuals, leading to **more realistic evaluations**.

Issues faced:
- **Radial Kernel tuning** on the imbalanced dataset (about **17,000 data points**) took **significant time (~3844 seconds)**.
  Even after testing **a wide range of cost and gamma values**, the model's performance **remained stagnant (~92% accuracy)**.
- Additionally, during tuning, **iteration warnings were encountered**, indicating **optimization difficulties** due to poor class separation and high cost settings.
  This highlights the critical need to **properly choose a reasonable cost function range** and **apply balancing techniques** when training SVMs.



- **Fruit consumption distribution plots** showed **substantial overlap** between diabetic and non-diabetic groups, reflecting the **subtle real-world differences** in diet habits.
  Such overlaps further made **classification difficult**, even with advanced kernels.



Feature importance analysis revealed that **BMI (BMICALC)**, **Pizza Consumption (PIZZANO)**,and **Fruit Intake (FRUTNO)** were the **strongest predictors** of diabetes.
Individuals with **higher BMI** and **unhealthier eating patterns** were more likely to develop diabetes, aligning with known public health research.

Takeaways:
Proper **data balancing**, **controlled hyperparameter tuning**, and **careful interpretation of feature contributions** are essential steps to build reliable disease prediction models using SVM.

Conclusion

- After evaluating Support Vector Machine models with **Linear**, **Radial Basis Function (RBF)**, and **Polynomial kernels**, the **Linear SVM model** showed the **best generalization performance** on the test set.
- Across all models, **Linear SVM achieved the highest test accuracy**, with **lower error rates** compared to Radial and Polynomial models.

Metrics Summary (Test Set):
•Linear SVM: Highest Accuracy, Lowest Error Rate, Balanced Precision and Recall
•Radial SVM: Slightly lower accuracy, some improvement post down-sampling
•Polynomial SVM: Good flexibility but slightly higher test error compared to Linear SVM

Best Model Hyperparameters:
- **•Linear SVM:**
  - **Best Cost:** 0.05
- **•Radial SVM:**
  - **Best Cost:** 0.01
  - **Best Gamma:** 0.001
- **•Polynomial SVM:**
  - **Best Cost:** 0.5
  - **Best Degree:** 3
  - **Best Coef0:** 2

Key Takeaways:
•Simpler models like Linear SVM can outperform more complex kernels when features are moderately separable after balancing.
•Proper data preprocessing (down-sampling) and careful hyperparameter tuning are critical for real-world health prediction tasks.

This study shows that **unhealthy eating habits**, such as **high pizza consumption**, **higher BMI**, and **lower fruit intake**, are strong predictors of diabetes risk in adults over 50. **Feature importance analysis** confirmed that poor dietary choices and elevated body weight are major contributors to diabetes development.

References

- Down-sampling Documentation:
- Caret Package, down Sample Function. Retrieved from https://rdrr.io/cran/caret/man/downSample.html
- SVM Conceptual Image:
- Wikipedia Contributors. *Linear Support Vector Machine* image. Retrieved from Wikipedia - Linear SVM
- **SVM Hyperparameter Explanation:**
  Scikit-learn Developers. *Support Vector Machine Parameters: Cost, Gamma, Degree, and Coef0*. Retrieved from https://scikit-learn.org/stable/modules/svm.html