Predicting Diabetes Using Demographic & Dietary Features Using NHIS DataSet

Devi Sowjanya Padala

Introduction

In this project, we aim to predict the presence of diabetes based on an individual's demographic characteristics and dietary habits by using the **2022 National Health Interview Survey (NHIS)** dataset.

The primary research questions we seek to answer are:
- Which demographic and dietary behavior variables are strong predictors of diabetes?
- How do different SVM kernels (linear, radial basis function (RBF), and polynomial) compare in performance for this classification task?

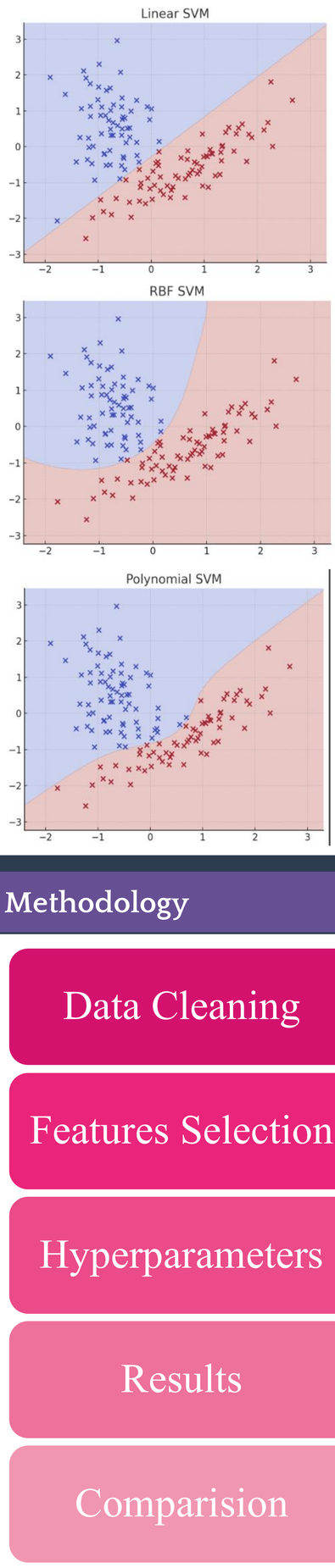
We carefully preprocess the dataset, perform exploratory data analysis, employ cross-validation and hyperparameter tuning to optimize model performance.

Theoretical Background

Support Vector Machines (SVM) are supervised learning models that find the optimal hyperplane to separate classes in a high-dimensional space.
- The goal of SVM is to maximize the margin between the support vectors, which are the critical elements of the training set that define the decision boundary.

SVM seeks to solve the optimization problem by minimizing $\frac{1}{2} \| \boldsymbol{w} \|^2$
subject to $\boldsymbol{y_i}(\boldsymbol{w}.\boldsymbol{x_i} + \boldsymbol{b}) \geq \mathbf{1}$ for all training examples (x_i, y_i).
Where:
- w is the weight vector. b is the bias. $y_i \in \{0, 1\}$ are the class labels.

When the data is not linearly separable, SVM uses the kernel to map inputs into a higher-dimensional space where a hyperplane can separate the classes. Common kernels include:



Linear Kernel: Suitable for linearly separable data. This is done by computing the dot product between the data points.

$$K(x, x') = x^T x'$$

- Simple, fast to train, easy to interpret; scales well to high-dimensional datasets.

Radial Basis Function (RBF) Kernel: This is based on distance between the data points. Maps data into an infinite-dimensional space, allowing nonlinear separation. The RBF kernel is defined by:

$$K(x, x') = \exp(-\gamma \| x - x' \|^2)$$

Where x and x' are input data points, γ is a hyperparameter that controls the width of the kernel and $\|.\|$ is the Euclidean distance between the points.

Polynomial Kernel: Allows the model to fit polynomial boundaries. It is defined as:

$$K(x, x') = (\gamma . x^T x' + r)^d$$

Where r is a coefficient term and d is the degree of the polynomial.
- These allows for more complex decision surfaces and also can control complexity by adjusting degree k.

Methodology

Data Cleaning
- Codes representing missing, unknown (996,997,998, 0,7,8,9) were removed.
- MARSTCUR and EDUC were recoded into meaningful categories and then one-hot encoding is used.
- The DIABETICEV is mapped from {1,2} -> {0,1} for better understanding.

Features Selection
- Only respondents from Region 4(South) were selected to focus the analysis geographically and reduce the size of dataset for faster processing.
- Demographic variables are selected(AGE, SEX, EDUC, MARSTCUR, BMI).
- Dietary features are selected (FRUTNO, VEGNO, PIZZANO, etc.)

Hyperparameters
- GridSearchCV was employed for hyperparameter tuning.
- Linear Kernel: Tuned **C** over **[0.01,0.1,1,5,10]**
- RBF Kernel: Tuned **C** over **[0.01,0.1,1,5,10]** and **gamma** over [**0.5, 1, 2, 3, 4]**
- Polynomial Kernel: Tuned **C** over **[0.01,0.1,1,5,10]** and **degree** over **[2,3,4,5].**
- 5-fold cross-validation was used for reliable model selection.

Results
- The dataset was split by sex to investigate how SVM models perform separately for males and females.
- Model performances were compared based on Precision, Recall, F1-Score and Accuracy.

Comparision
- Permutation importance was used to identify and rank feature importance for each model.
- The relationship between the target variable and the top 5 predictors was explored through count plot.
- To visually interpret model-decision making, the top 2 predictors were selected.

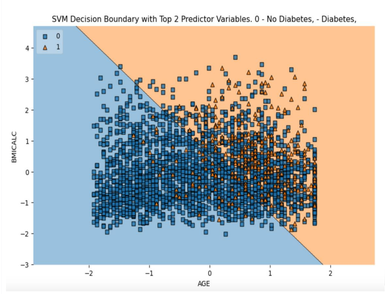
Results

The dataset was split by sex to investigate how Support Vector Machine (SVM) models perform separately for males and females. The dataset is split into 70% training and 30% testing.
- Before plotting SVM boundary the predictors AGE, BMI are standardized.

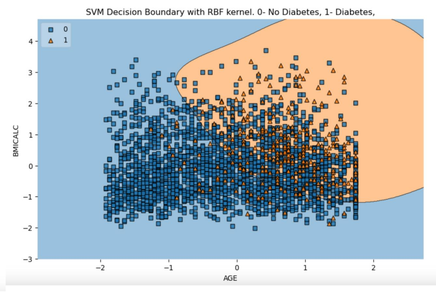
Model performances were evaluated based on:
- **Precision**: Proportion of positive predictions that were correct.
- **Recall**: Proportion of actual positives that were correctly identified.
- **Accuracy**: Overall proportion of correct predictions.

Kernel	Accuracy	Precision	Recall
Linear	Female – 0.66 Male – 0.64	Female – 0.96(No Diabetes) Male – 0.95	Female – 0.64 Male – 0.62
RBF	Female – 0.88 Male – 0.88	Female – 0.88 Male – 0.88	Female – 1 Male – 1
Polynomial	Female – 0.75 Male – 0.75	Female – 0.94 Male – 0.94	Female – 0.76 Male – 0.76

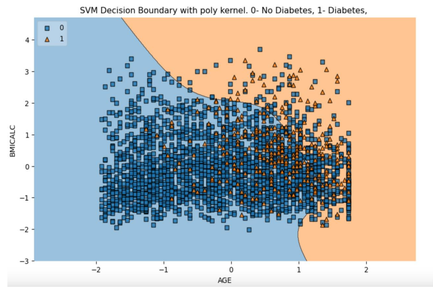
Linear



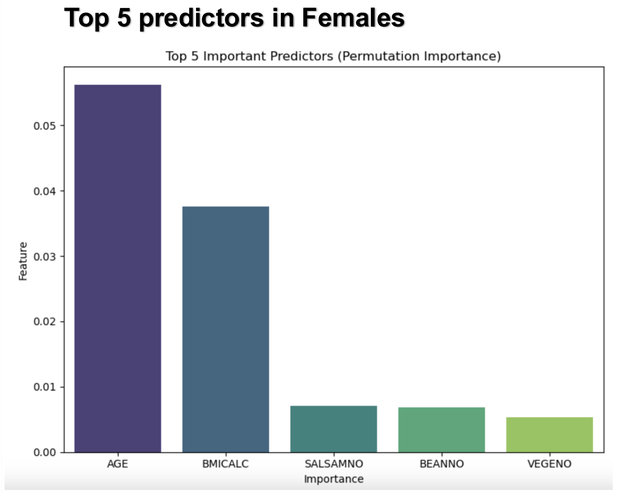
RBF



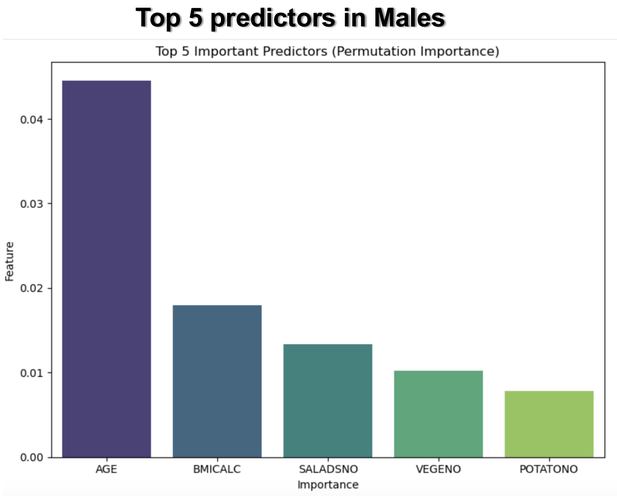
Poly



Top 5 predictors in Females



Top 5 predictors in Males



The above plots are the top 5 predictors for the SVM model with poly kernel.

- **AGE**
- **BMICAL** (Body mass Index)
- **SALSAMNO** – Weekly consumption of green salad.
- **BEANNO** – Weekly consumption of beans.
- **SALADSNO** – Weekly consumption of green salad.
- **VEGENO –** Weekly consumption of vegetables.
- **POTATONO –** Weekly consumption of non-fried potatoes.

Results Interpretation

- **The AGE** and **BMICALC** (Body Mass Index) consistently emerged as the strongest predictors of diabetes for both male and female groups.
- Among dietary habits, features like **SALADSNO** (salad consumption), **VEGENO** (vegetable intake), **POTATONO** (non-fried potato intake), and **SALSAMNO/BEANNO** (in females) were also relevant, though their predictive strength was comparatively lower. This supports the idea that while diet contributes to diabetes risk, **demographic and body composition factors dominate**.
- Result are slightly misleading due to class imbalance, with the majority class (no diabetes) far outnumbering the positive class (diabetes) by nearly **9 to 1**. This skew causes models to **favor predicting the majority class**, inflating metrics like accuracy while **failing to detect true positives**.
- This is especially evident in the **RBF kernel**, where the model achieved perfect recall (1.0), but this likely occurred because it predicted most observations as non-diabetic to increase sensitivity, potentially sacrificing precision.
- In this particular task, choosing poly SVM is better even though RBF achieved high accuracy, because it doesn't overfit as aggressively as RBF and provides **better precision and recall balance.**
- Also if we look at the SVM boundary plot for linear SVM, we can see the the top right orange region corresponds to the class 1(Diabetes). This suggests that higher age and higher BMI increase the likelihood of being classified as diabetic.

Suggestion to policy makers

Our analysis shows that diabetes prevention efforts must address both biological factors (age, obesity) and social/behavioral factors (diet quality and social support).
- Health interventions should focus on promoting healthy eating (increasing vegetable intake, reducing sugary beverage consumption
- Other policy change include improving economic conditions, such as combating food insecurity and ensuring affordable access to healthy food, which directly affects dietary choices and long-term health outcomes.

References

[1] 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. https://doi.org/10.18128/D070.V7.4. Links to an external site.http://www.nhis.ipums.orgLinks to an external site..

[2]https://www.geeksforgeeks.org/optimal-feature-selection-for-support-vector-machines/

[3]https://scikit-learn.org/stable/modules/svm.html

[4] https://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html