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.
Evaluated the model performance using accuracy, precision, recall score.

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 $\frac{1}{2}\parallel \boldsymbol{w} \parallel^2$
subject to $\boldsymbol{y_i}(\boldsymbol{w}.\boldsymbol{x_i}+\boldsymbol{b}) \geq \boldsymbol{1}$ 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:

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

Where x and x' are input data points, γ is a hyperparameter that controls the width of the kernel and $\parallel . \parallel$ 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 and gamma, Polynomial Kernel: Tuned C and degree.
- 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.

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**

Plots between target variable ('DIABTEIC') and top 3 predictors for the SVM with poly kernel

Top 5 predictors

- **AGE**
- **BMICAL** (Body mass Index)
- **SALADSNO** – Weekly consumption of green salad.
- **Married_SA** – Married but spouse absent.
- **FRTDRINKMNO** – Weekly consumption of fruit flavoured drinks

- Prediction accuracy for females is more when compared to males with linear kernel.
- Polynomial kernel better captured higher-order relationships, improving decision boundaries for complex feature interactions.
- RBF kernel achieved perfect recall (1.0) likely due to class imbalance.

Conclusion

Our findings show that **age** and **BMI** are strong biological predictors of diabetes, while **poor dietary habits** (low salad intake, high sugary drink consumption) and **social factors** (absence of a spouse) also contribute meaningfully.
Older adults, individuals with higher BMI, and those with poor diets are at elevated risk.
Social isolation may indirectly impact diabetes risk through stress and unhealthy behaviors.

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) and providing targeted support to socially vulnerable groups such as individuals living without a spouse to effectively prevent from diabetes.

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