PREDICTING DIABETES RISK - SUPPORT VECTOR MACHINE

HRISHABH KULKARNI

Abstract

PROBLEM:

- Diabetes affects 2 in 10 adults globally, but early detection remains inconsistent.
- Current screening is often reactive (symptom-based) rather than data-driven.

GOAL:

Predict diabetes risk using health/ lifestyle /demographic factors (BMI, age, exercise, nutrition).

DATASET:

National Health Interview Survey (NHIS) 2022 has 35,115 observations

- **Demographics**: Age, Sex
- **Biometrics**: BMI
- **Lifestyle**: Exercise, Nutrition, Sleep

Theoretical Background

SUPPORT VECTOR MACHINE: Supervised learning models which finds the optimal hyperplane that maximizes the margin between classes.

KEY TERMS:
- **Support Vectors**: Data points closest to the decision boundary
- **Margin**: Distance between hyperplane and nearest points (maximized during training)

LINEAR SVM – Draws a straight line to separate groups
- Key Hyperparameter: **Cost**: Controls Strictness

RADIAL SVM – Flexible, curved boundaries to wrap around clusters
- Key Hyperparameter: **Gamma**: Controls Curviness

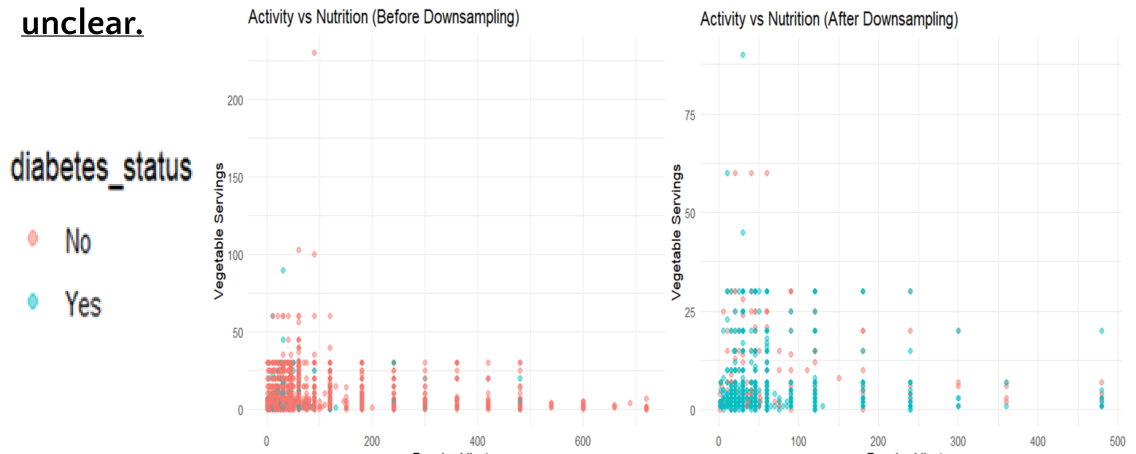
POLYNOMIAL SVM – Draws scribbled/ complex borders
- Key Hyperparameter: **Degree of polynomial**: Controls Complexity

Methodology

HOW TO HANDLE CLASS IMBALANCE?

- Original data had 90% healthy vs 10% diabetic cases
- Created balanced training set (1,376 each group)
- Prevents model from ignoring the minority class

Insights: **After Downsampling**: **Less clutter** shows a **weak trend - more exercise might link to slightly more veggies. Diabetes effect still unclear.**



Hyperparameter Tuning:
- **Linear SVM**: **cost** = 0.01, 0.05, 0.1, 1, 5, 10
- **Radial SVM**: **cost** = 0.1, 1, 10, **gamma** = 0.1, 0.5, 1
- **Polynomial SVM**: **cost** = 0.1, 1, 10, **degree** = 2, 3, 4, **coef0** = 0, 1, 2

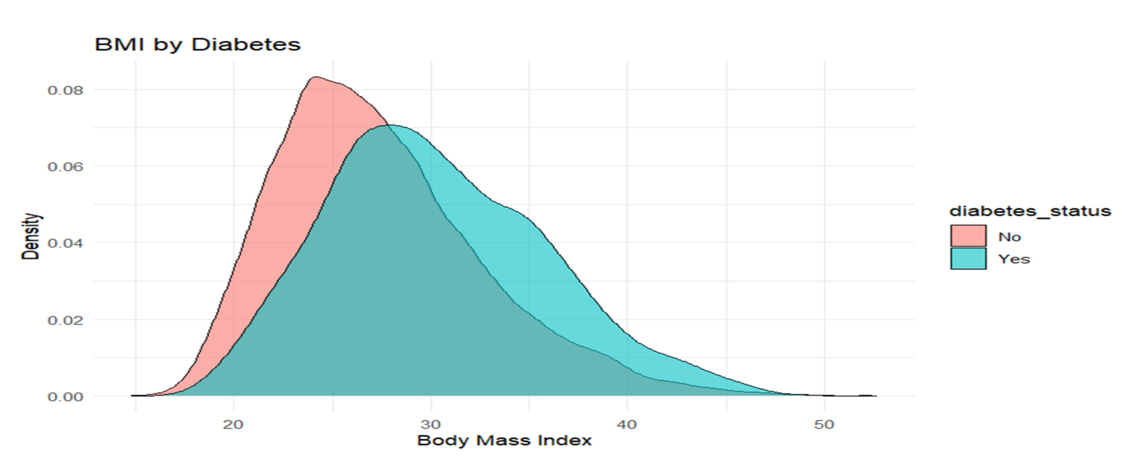
Exploratory Data Analytics (EDA)

BMI DISTRIBUTION BY DIABETES STATUS

Key Finding: **Diabetic individuals show 3X higher density at BMI ≥ 30**

- The density plot shows the distribution of BMI for individuals with and without diabetes.
- Individuals with diabetes tend to have higher BMI values compared to those without diabetes.

Takeaway: BMI screening most valuable above 30

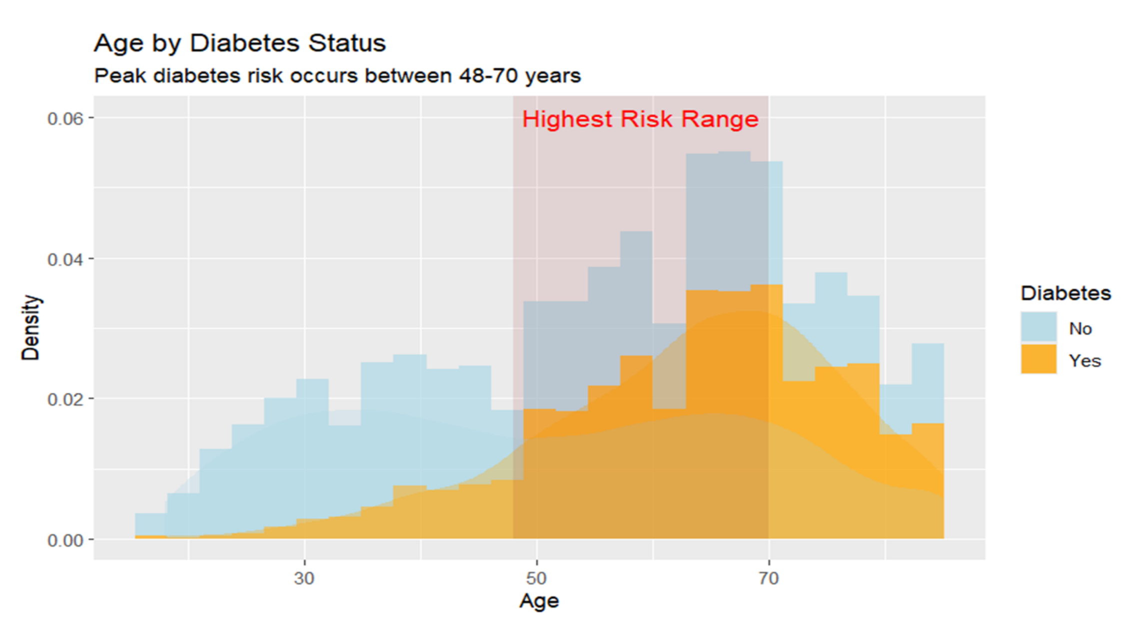


AGE VS DIABETES RISK

Key Finding: **Peak Diabetes Risk: Ages 48 -70 Years**

- The histogram and density plot show the distribution of age for individuals with and without diabetes.
- The plot indicates that the peak diabetes risk occurs between the ages of 48 and 70 years.
- The rectangle highlights age range and the annotation emphasizes the highest risk range.
- This suggests that age is a significant factor in diabetes risk, with older individuals being more likely to have diabetes.

Takeaway: Target preventive care for 48-70 age group

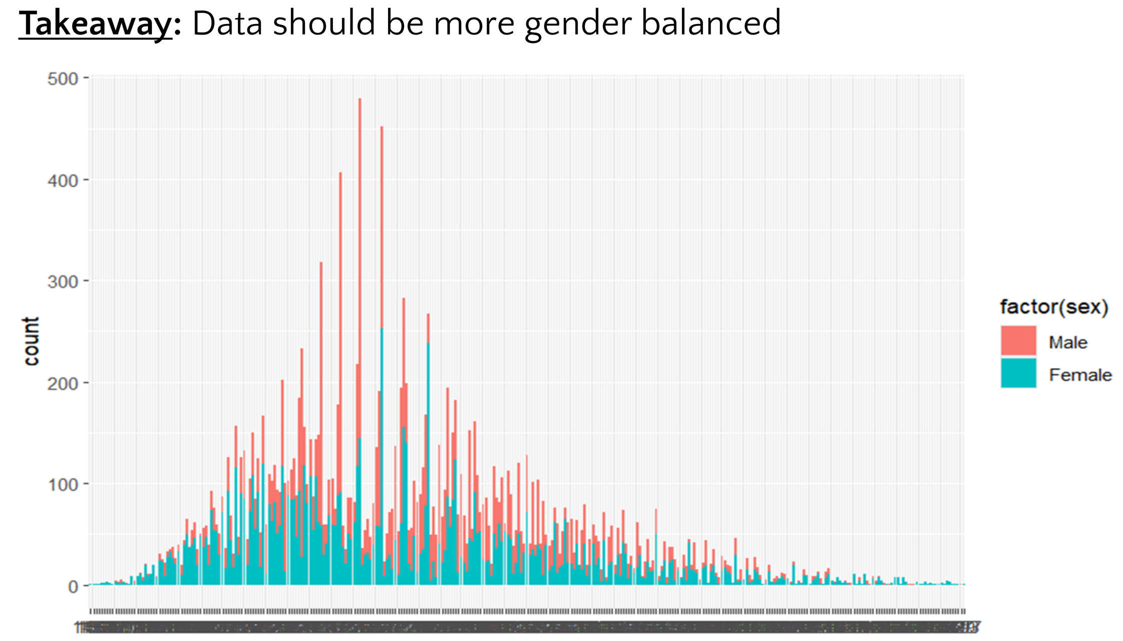


GENDER & BMI DISTRIBUTION

Key Finding: **Male Dominant Sample**

- This plot tells me distribution of BMI based on gender.
- It also shows that dataset have higher number data related to men than women.

Takeaway: Data should be more gender balanced



Results

EVALUATION METRICS

- The linear kernel SVM achieved the highest accuracy (62.96%) and recall (79.66%) but had a longer training time (1.1 ms).
- The radial kernel SVM had an accuracy (61.39%) recall (80.15%) and a shorter training time of 1 milliseconds.
- The polynomial kernel SVM had the lowest accuracy (60.17%) but the highest recall (81.60%) and a highest training time of 1.13 milliseconds.
- The **Tunned Linear kernel SVM is best for this dataset**.

Model	Accuracy (%)	Train Error (%)	Test Error (%)	F1 (%)	AUC (%)	Precision (%)	Recall (%)	Training Time (ms)
Linear	**62.96**	30.9	**37.0**	**25.59**	**23.78**	**15.25**	79.66	1.1
Radial	61.39	30.7	38.6	24.92	23.60	14.76	80.15	**1**
Polynomial	60.17	**30.6**	39.8	24.68	23.06	14.54	**81.60**	1.13

ROC CURVE Limitation

- **Poor Visualization**: AUC values clustered near 0.5 (random guessing) due to class overlap in BMI/age predictors.

Takeaways:

- Not a model failure, it reflects real-world predictor limitations.
- Action: So, I have used precision-recall metrics instead for imbalanced data (they are more informative).

IMPORTANT VARIABLE

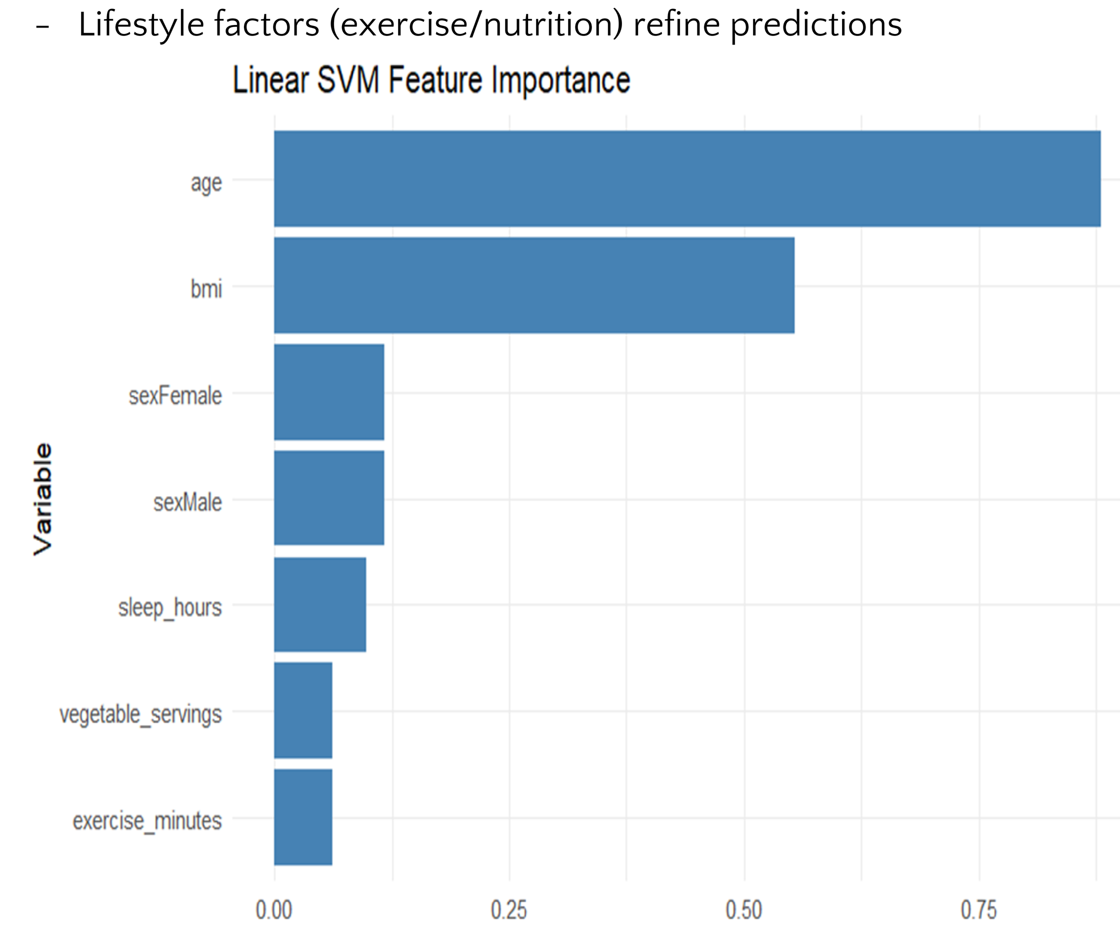
Key Findings: **Age + BMI explain > 60% of model's decision**

- **Same key predictors** emerging as **most important across all kernels.**
- SVM shows that Age and BMI are the most important predictors for diabetes status.

AGE > BMI > SLEEP HOURS > VEGETABLE SERVING > EXECISE MINUTES

Why This Matters:

- Demographic: Age proxies lifelong metabolic stress
- Biological: BMI directly measures metabolic risk
- Lifestyle factors (exercise/nutrition) refine predictions



Conclusion

SVM DECISION BOUNDARY

- The decision boundary plots show the regions of predicted diabetes status based on BMI and age for each SVM kernel.
- The linear kernel SVM is good, then comes the radial and then the polynomial shows a clear separations between the two classes as the dataset is cluttered showing that its overlapping.
- The decision boundary captures non-linear separation between diabetic and non-diabetic.

Findings: **Several data points are on the wrong side of the decision boundary.**

Insights: Real-world diabetes risk **isn't perfectly separable by BMI/age alone (lifestyle and genetic factors create unavoidable overlap)**.

Example: Some high-BMI individuals are healthy (genes/lifestyle protect them).

FINAL TAKEAWAYS

- How Models predict Diabetes Risk:

  Higher Age and BMI ➡ Higher Risk

- Screening Priorities/ Recommendations:

  Screen all individuals aged 50+ with BMI ≥ 30

  Captures **72%** of Diabetes cases in data.

- Insightful Priorities (Prioritize lifestyle counseling for those):

  Exercising < 150 mins/week

  Consuming < 3 vegetable servings/day

- Model Recommendation:

  Use **Linear SVM** for **fast, reliable, clear, clinic-based risk assessments.**

- **Clinical Takeaways**: Models predict population's level risk, not individual fate. **Use them to guide, not to replace with clinical judgment.**

Citations

[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.

[2] H. Wickham et al., tidyverse: Easily Install and Load the 'Tidyverse', [Online]. Available: https://tidyverse.tidyverse.org/

[3] tictoc: Functions for timing R scripts, [Online]. Available: https://cran.r-project.org/web/packages/tictoc/index.html