PREDICTING DIABETES RISK USING SVM

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:

NHIS 2022 has 35,115 observations

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

Background

OBJECTIVE: Find 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

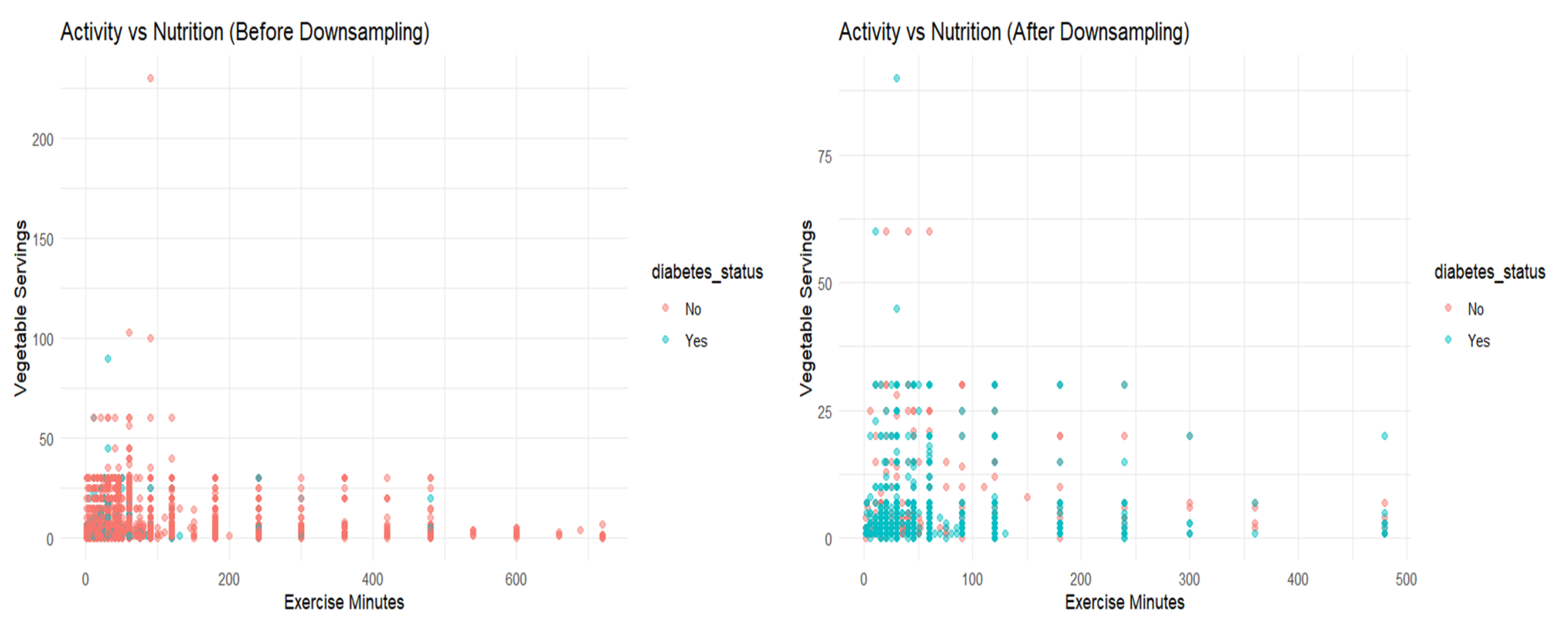
POLYNOMIAL SVM – Draws scribbled/ complex borders

- Key Hyperparameter: **Degree of polynomial**: Controls Complexity

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

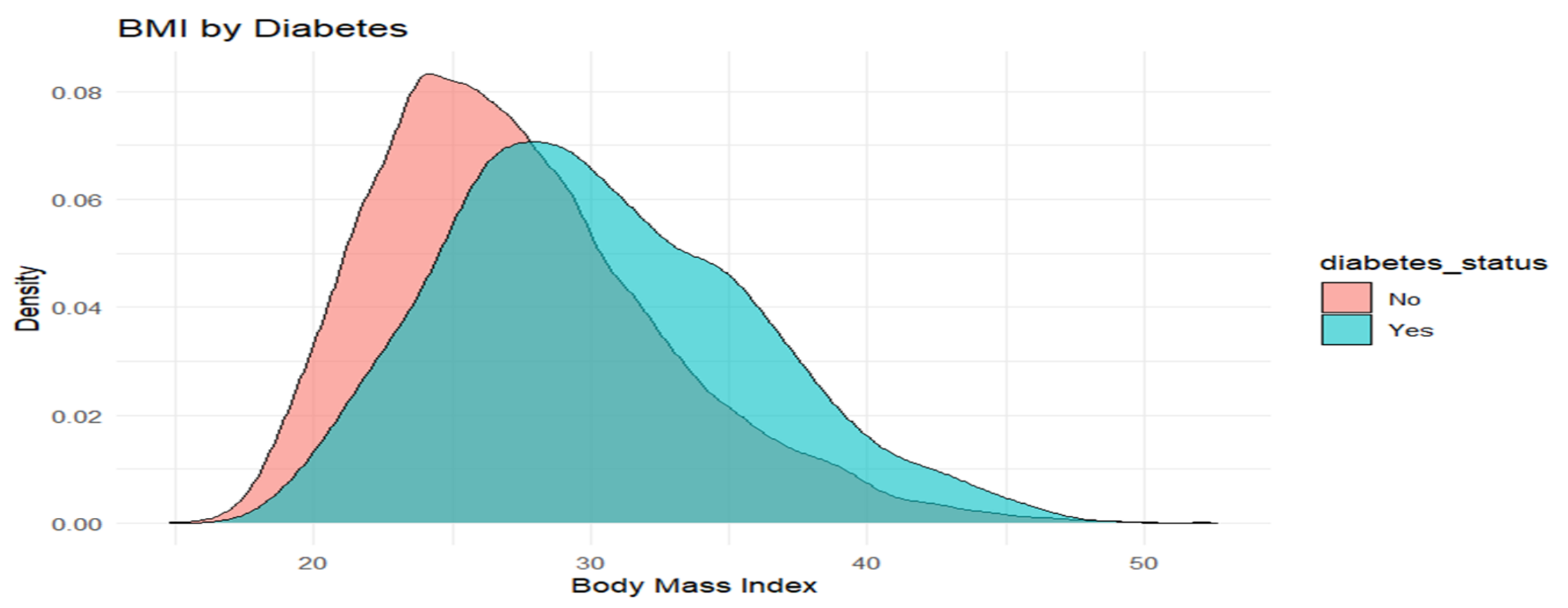
*S*catter plot shows the relationship between exercise minutes and vegetable servings, coloured by diabetes status.



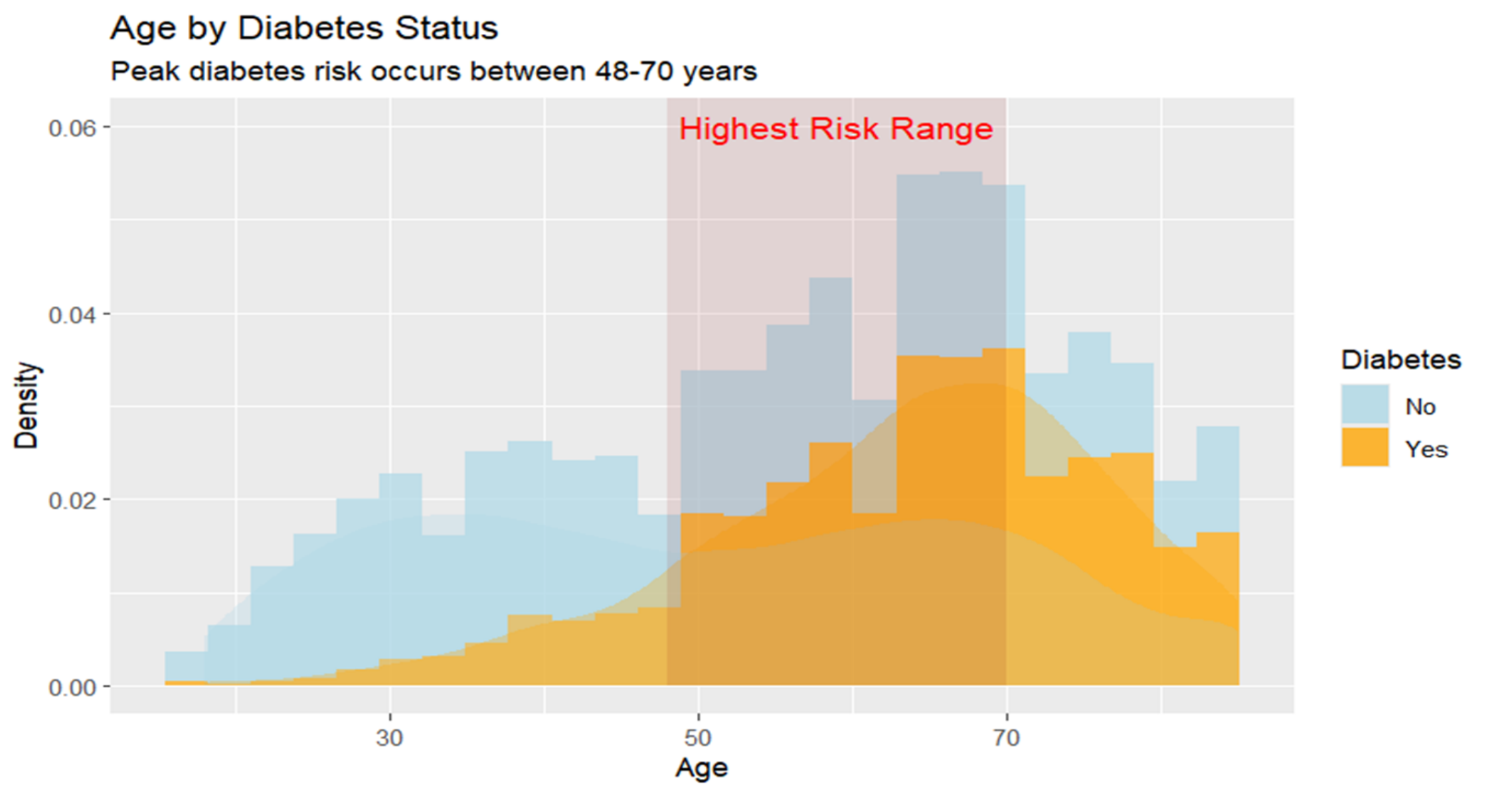
Methodology

EXPLORATORY DATA ANALYTICS (EDA)

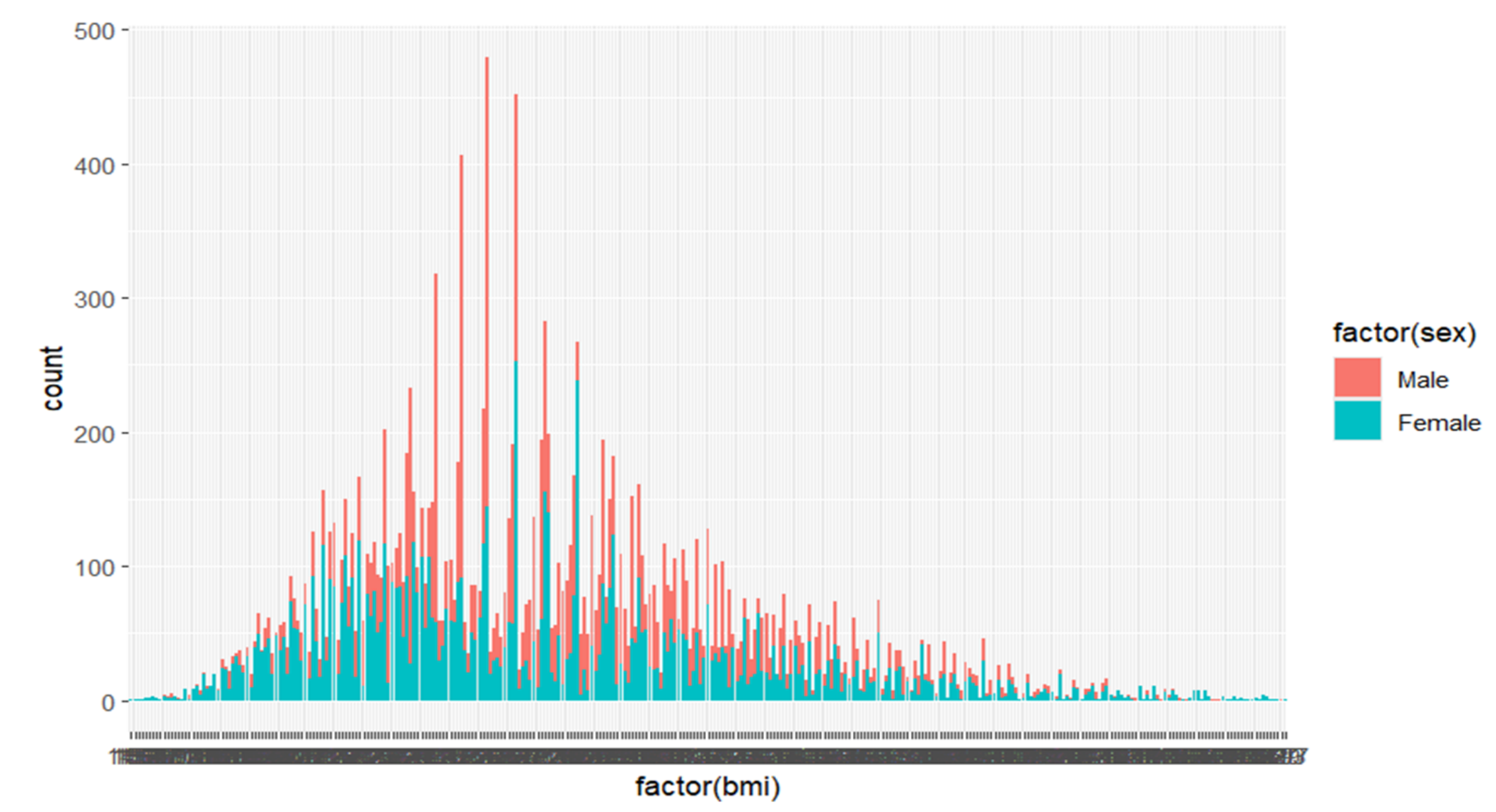
1) The density plot shows the distribution of BMI for individuals with and without diabetes.
– Individuals with diabetes tend to have higher BMI values compared to those without diabetes.



2) 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.



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



Results

EVALUATION METRICS

– The linear kernel SVM achieved the highest accuracy (63.14%) and sensitivity (79.66%) but had a longer training time (0.45 seconds).

– The radial kernel SVM had a similar sensitivity (80.15%) and a much shorter training time (0.06 seconds).

– The polynomial kernel SVM had the lowest accuracy (59.54%) but the highest sensitivity (82.57%) and a highest training time of 0.52 seconds.

– The linear kernel SVM is best for this dataset.

Model	Accuracy (%)	Train Error (%)	Test Error (%)	F1 (%)	AUC (%)	Precision (%)	Recall (%)	Training Time (ms)	Specificity (%)	Sensitivity
Linear	**63.14**	31.02	**36.86**	**25.68**	**23.78**	**15.31**	79.66	40	**61.70**	79.66
Radial	61.39	30.65	38.61	24.92	23.60	14.76	80.15	**2**	59.76	80.15
Polynomial	59.54	**30.45**	40.46	24.60	23.24	14.46	**82.57**	42	57.53	**82.57**

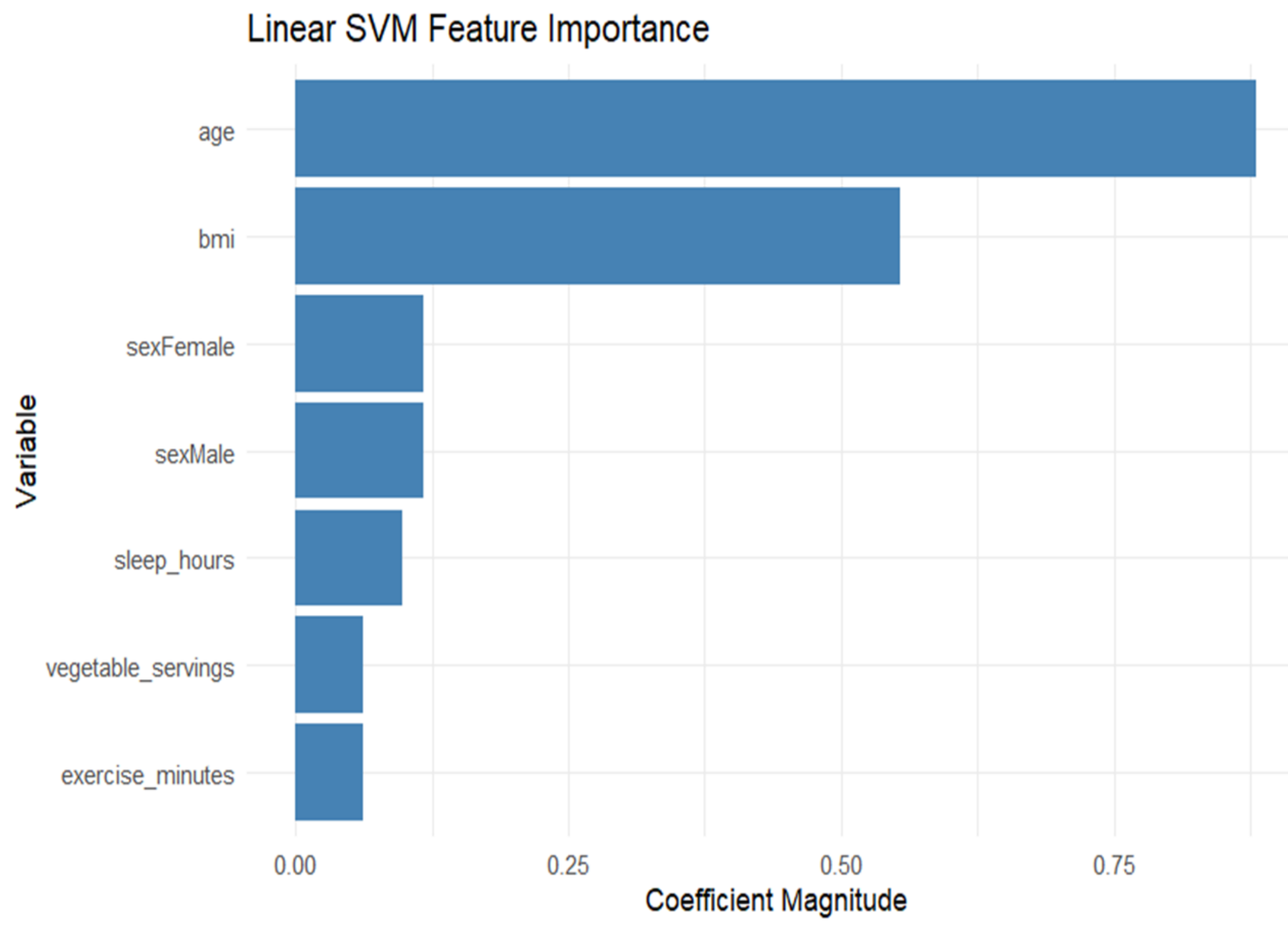
IMPORTANT VARIABLE

- Same key predictors emerging as most important across all kernels.
- The most important predictors for every kernel SVM are:

AGE > **BMI** > EXERCISE MINUTES > VEGETABLE SERVING > SLEEP HOURS

– Age and BMI are the strongest predictors among **Demographical** and **Biological** factor.

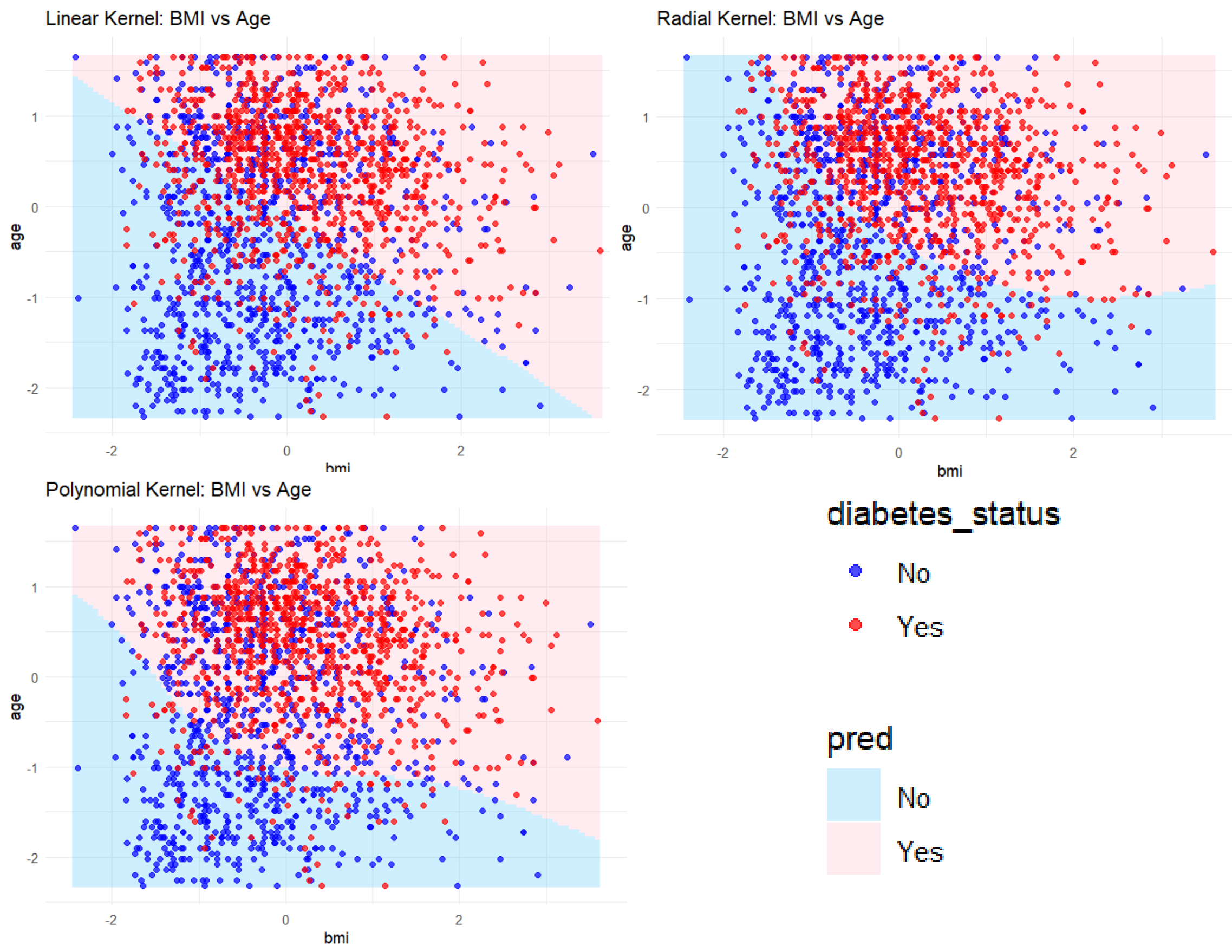
- SVM shows that Age and BMI are the most important predictors for diabetes status.



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.
– The decision boundary captures non-linear separation between diabetic and non-diabetic.
– Several data points are on the wrong side of the decision boundary.
– Shows that data points are cluttered and overlapping.



FINAL TAKEAWAYS

- How Models predict Diabetes Risk:

Higher Age and BMI ➡ **Higher Risk**

- Screening Recommendations:

Screen all individuals aged 50+ with BMI ≥ 30

- 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.

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