Predicting Cancer Diagnosis Using Support Vector Machines On IPUMS Health Survey Data

By Prateek Pagare

DATA 5322 – Machine learning 2

Introduction

- In this project, I investigate cancer prediction using survey data from the 2022 National Health Interview Survey (NHIS).
- I used support vector machines (SVM) with different kernels — linear, radial, and polynomial — to explore relationships between demographic, health, and behavioral factors and cancer history.
- Variables studied included age, sex, BMI, poverty level, and alcohol use
- I focus on building predictive models while improving interpretability using plots and performance comparisons.
- Goal: Predict history of cancer diagnosis based on key health and demographic predictors.

Theoretical Background

Support Vector Machines (SVM) are supervised learning algorithms that find a hyperplane to best separate classes.

- **Linear SVM:** Fits a straight boundary. Tuning parameter: Cost (C) — controls margin width vs misclassification.
- **Radial SVM (RBF):** Allows flexible, non-linear boundaries. Tuning parameters: Cost and Gamma (γ) — controls curvature.
- **Polynomial SVM:** Uses polynomial features to separate data. Tuning: Cost and Degree of the polynomial.

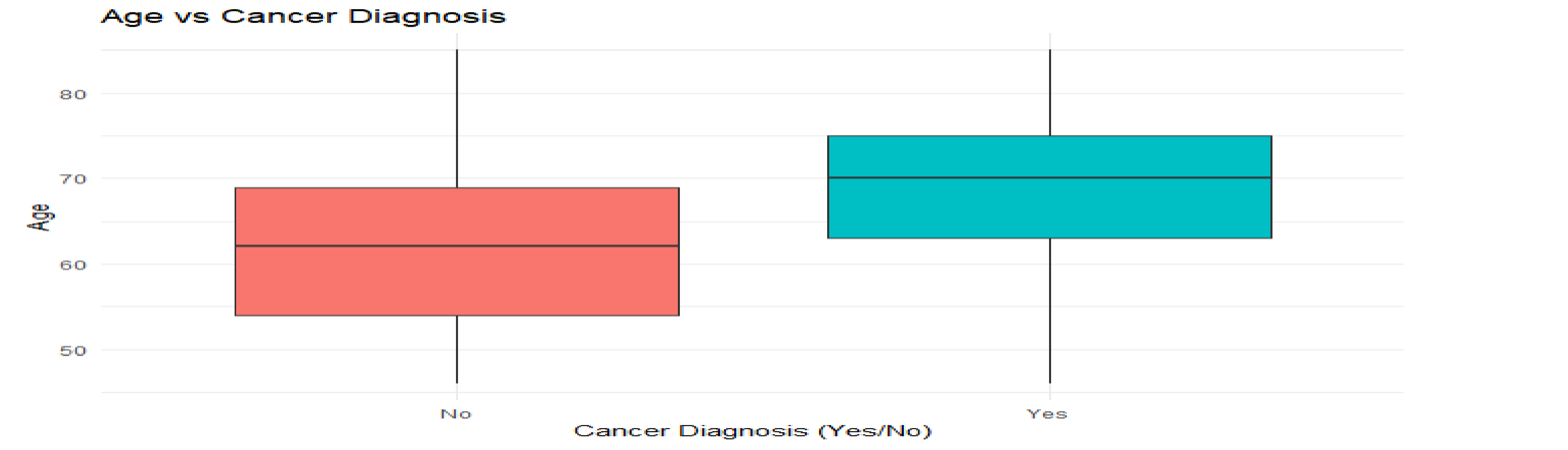
Tuning Method:
I used 10-fold cross-validation to select optimal Cost and Gamma/Degree values where appropriate.

Performance Metrics:
Accuracy, Sensitivity, Specificity, F1-Score.

Methodology

- **Data Cleaning:**
  Removed missing values in CANCEREV, SEX, BMICALC, ALCDAYSYR, AGE, POVERTY.
- **Subsetting for Focus:**
  Selected individuals:
  - Age > 45 years
  - Married (MARSTCUR == 1)
  - Alcohol consumption >50 days/year
- **Modeling Steps:**
  - Train/test split (70%/30%)
  - Fit SVM models (Linear, Radial, Polynomial)
  - Tune hyperparameters with cross-validation

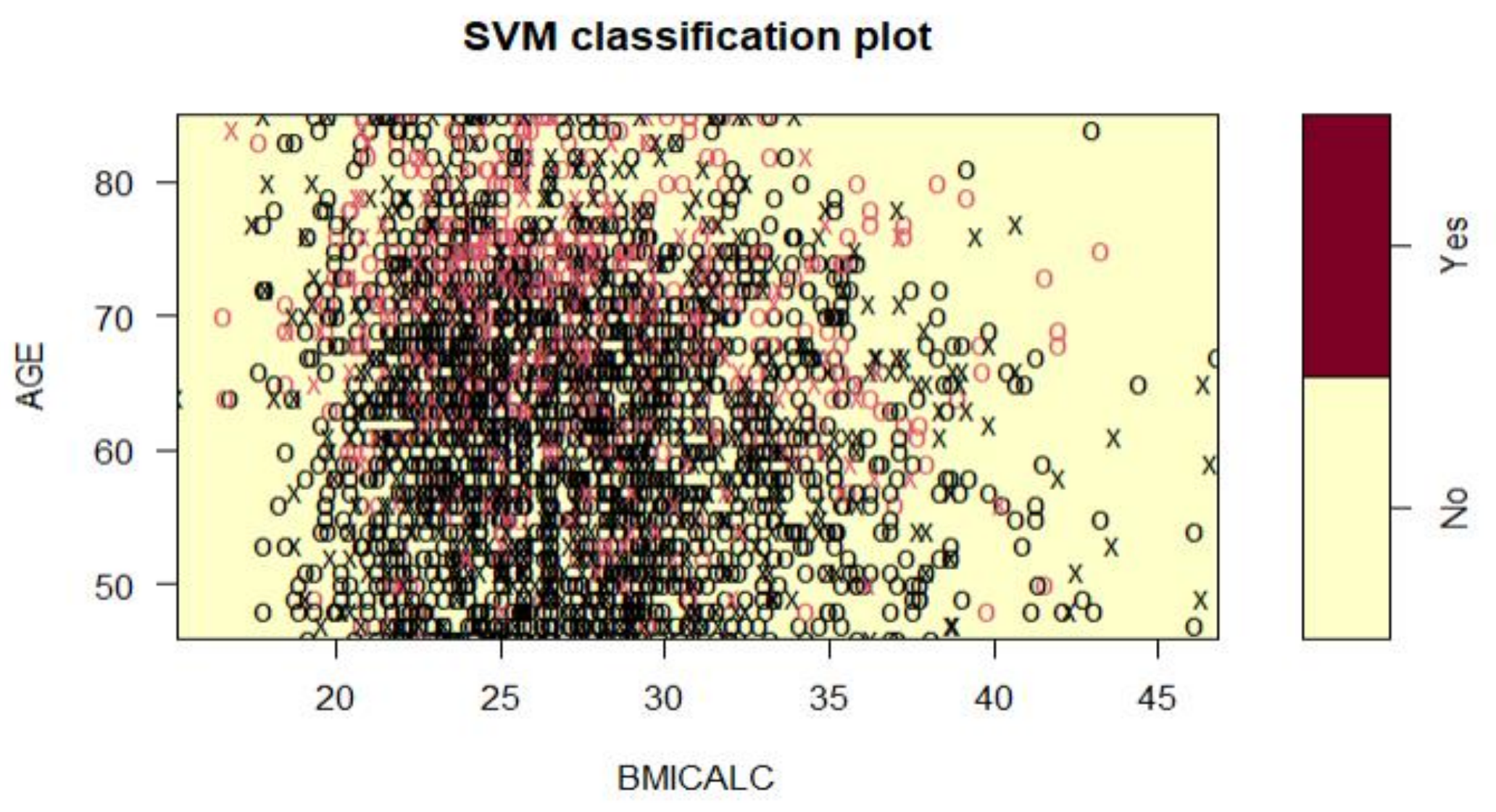
Reason for Subsetting:
Real-world data was highly imbalanced; subsetting improved balance and allowed models to better learn patterns.



Results

Linear SVM

- Predictors: AGE, SEX, POVERTY, BMICALC, ALCDAYSYR
- Accuracy: 81% | F1: NA
- Plot:

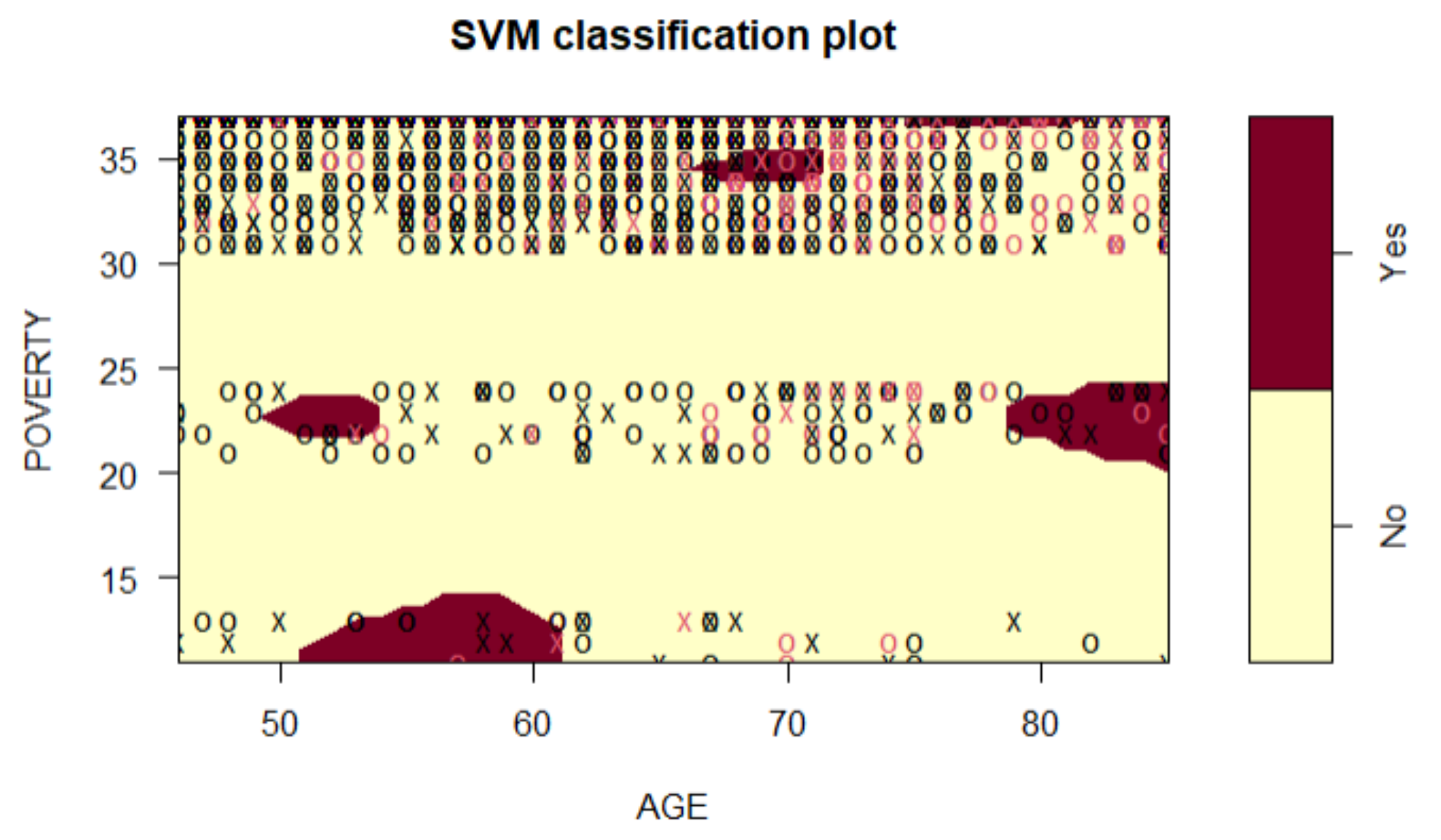


Linear SVM achieved 81% accuracy but failed to classify positive cases. The boundary was too rigid for real-world data. Linear SVM underperformed for detecting "Yes" class.

- Confusion Matrix :
  - Model predicted only "No" — unable to detect positive cancer cases.

Radial SVM

- Predictors: AGE, POVERTY, ALCDAYSYR
- Accuracy: 78.6% | F1: 0.82
- Plot:

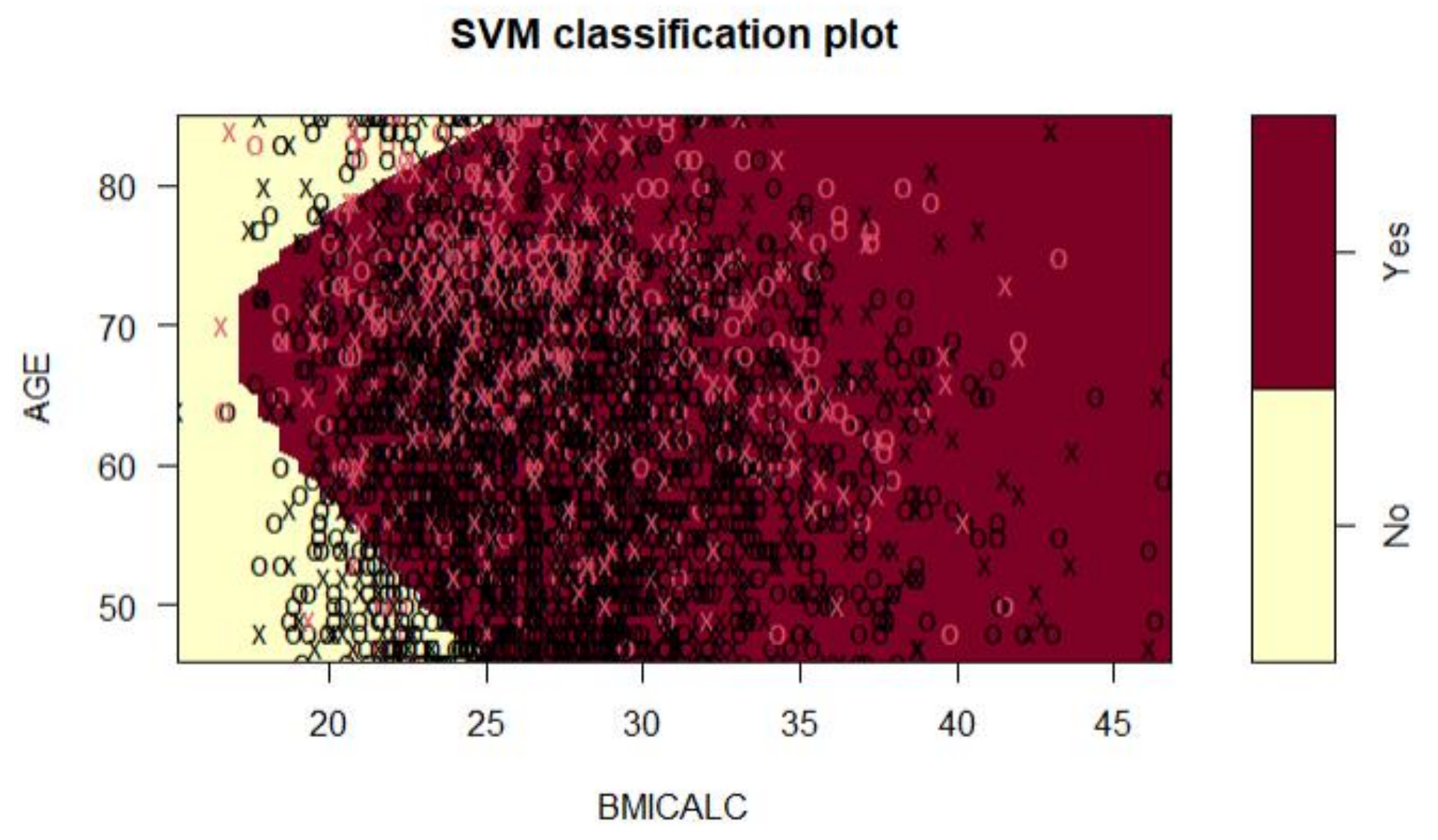


Linear SVM achieved 81% accuracy but failed to classify positive cases. The boundary was too rigid for real-world data. Linear SVM underperformed for detecting "Yes" class.

- Confusion Matrix :
  - Much better at catching "Yes" cases, lower specificity.

Polynomial SVM

- Predictors: AGE, SEX, POVERTY, BMICALC, ALCDAYSYR
- Accuracy: 80.6% | F1: 0.81
- Plot:



Polynomial SVM fitted complex boundaries but overfit the "No" class, leading to poor generalization despite decent overall accuracy.

- Confusion Matrix :
  - Similar to radial but slightly worse balance.

Table 1. Summary Table.

Model	Predictors	Accuracy	F1-Score
Linear SVM	5((AGE, SEX, POVERTY, BMICALC, ALCDAYSYR))	81%	NA
Radial SVM	3 (AGE, POVERTY, ALCDAYSYR)	78.6%	0.82
Polynomial	5((AGE, SEX, POVERTY, BMICALC, ALCDAYSYR))	80.6%	0.81

Discussion

Age, BMI, poverty ratio, and alcohol use emerged as strong predictors of cancer history.SVM plots show complex relationships between health habits and disease likelihood.Demographic factors like older age and lower socioeconomic status are associated with greater cancer prevalence.

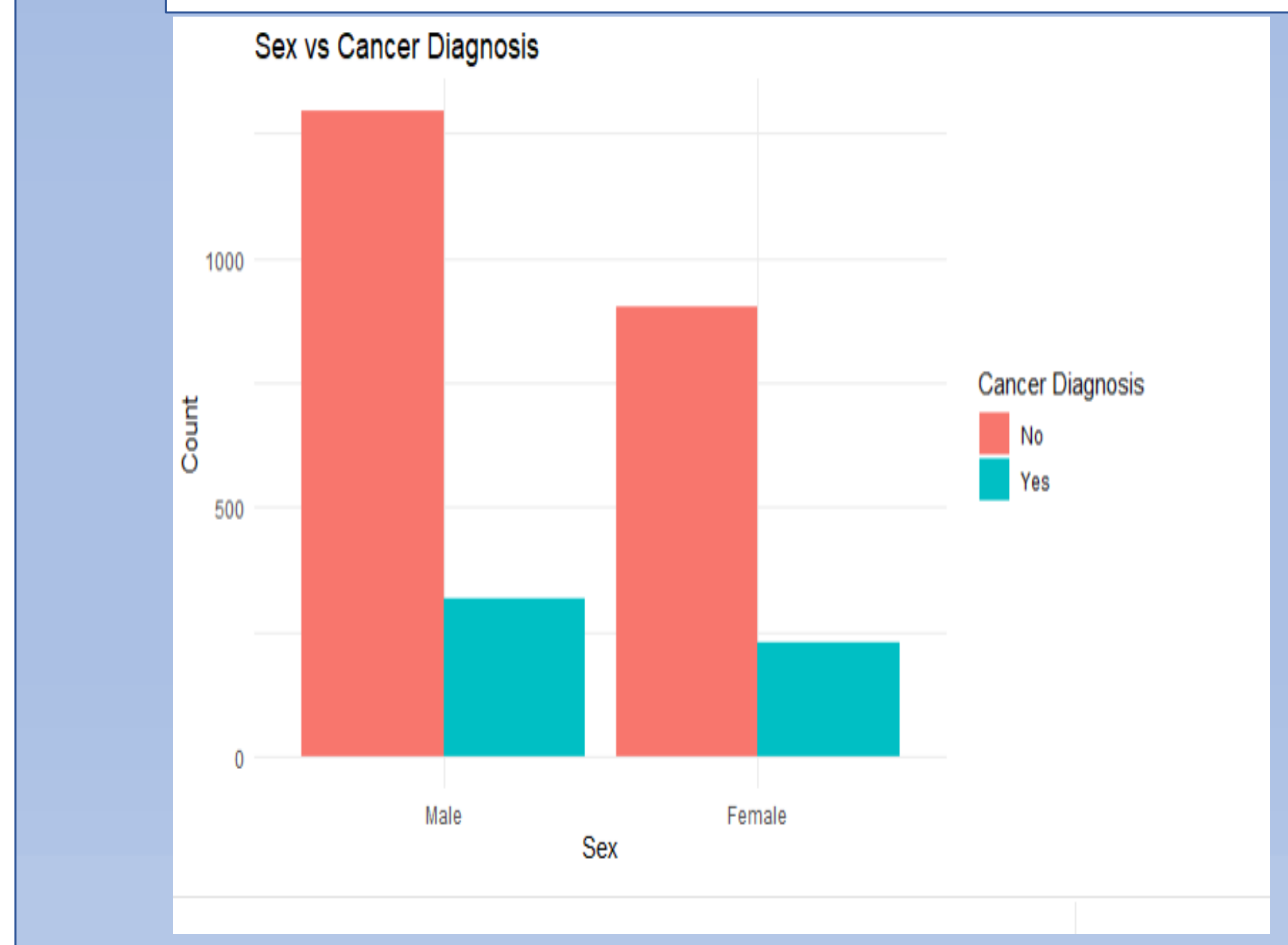
Radial SVM was best suited to capture non-linear patterns, while linear and polynomial models struggled to balance sensitivity and specificity.

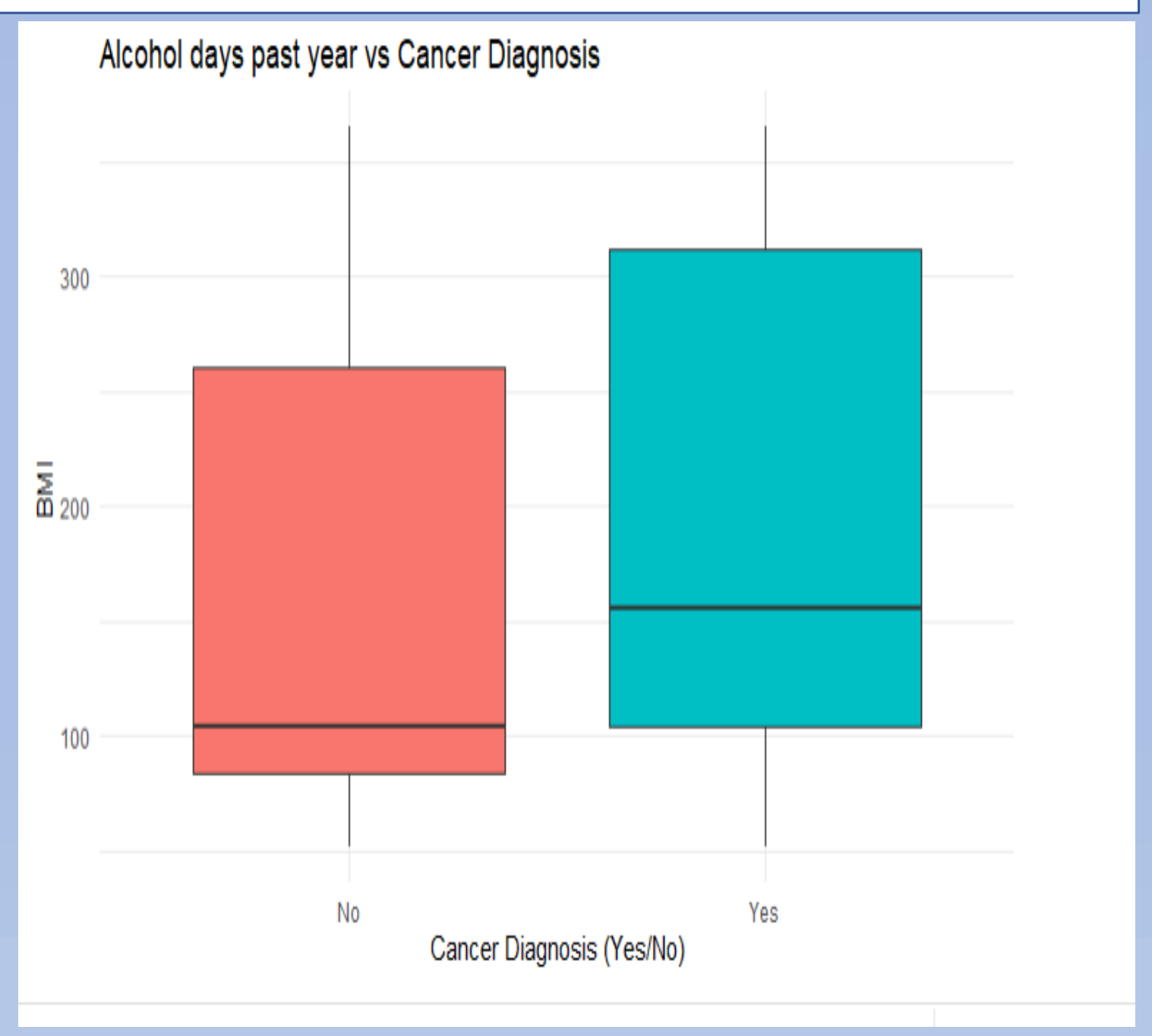
Conclusions

This project showed that Radial SVM performed better than Linear and Polynomial SVMs in predicting cancer history based on real-world health data. Age, BMI, poverty level, and alcohol use were key predictors, with older individuals and higher alcohol consumption linked to higher cancer prevalence.

Radial SVM captured the complex, non-linear patterns in the data more effectively, though challenges remained due to class imbalance.

These results suggest that public health strategies should focus on preventative care for older adults, especially those facing economic hardship or heavy alcohol use. Future work could explore better handling of imbalanced data to improve predictions for high-risk groups.





References

1. ISLR - Introduction to Statistical Learning (James et al.) https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.download.html
2. Caret R Package Documentation https://cran.r-project.org/web/packages/caret/vignettes/caret.html
3. fourfoldplot: Fourfold Plots https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/fourfoldplot
4. National Health Interview Survey (NHIS 2022) https://www.cdc.gov/nchs/nhis/documentation/2022-nhis.html