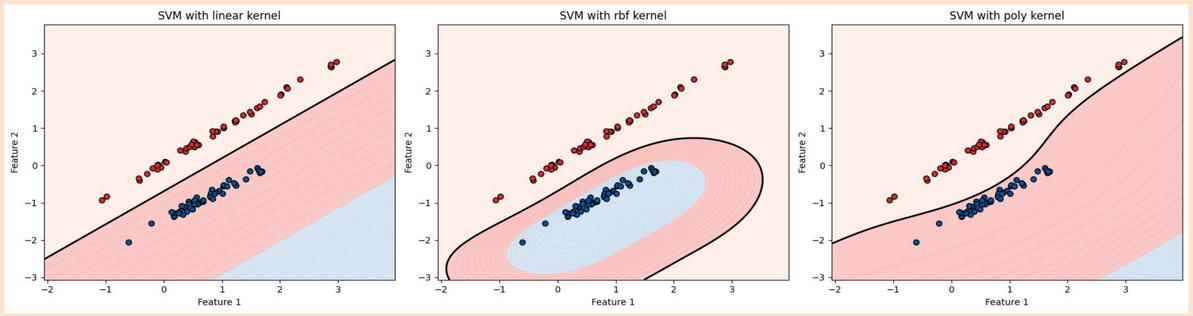
Margin of Survival: SVM for Cancer Detection

Christopher Yang | DATA5322

Technical Background

Support Vector Machine (SVM)

A supervised learning algorithm typically used for classification. The goal of a SVM is to find a hyperplane or decision boundary that maximizes the margin between different classes. An optimal hyperplane of decision boundary is one that provides the largest separation between the data points closest to the decision boundary known as support vectors. Linear SVM tries to separate data points using a straight line or flat hyperplane. Radial tries to separate data points using more complex boundaries. Polynomial tries to separate data points using higher order 'd' polynomial boundaries.



$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

Linear SVM
Simple dot product assuming data is linearly separable.

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$$

Radial SVM
Models similarity between points based on distance. Closer = more similar.

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

Polynomial SVM
Models interactions between features up to 'd' degrees.

Hyperparameters

(1) **C** (Linear, Radial, Poly) - Controls trade off between maximizing margin and minimizing classification errors. High C value gives priority to minimizing classification error, which can lead to overfitting.
(2) **Gamma** (Radial, Poly) - Controls shape of decision boundary. High gamma creates more complex decision boundaries.
(3) **Degree** (Poly) - Specifies degree of polynomial. Higher degree makes the model more complex and risk overfitting.
(4) **Coef0** (Poly) - Controls influence of higher degree terms in the polynomial kernels. Higher coef0 allows higher degree terms to have more influence in models.

Methodology

Model Flow

Load Data → Select Features → Clean Data → Grid Search Hyperparameters → Build & Test Models → Interpret Results

Features

AGE (DEMO)- Age of the individual
SEX (DEMO) - Sex of the individual
BMICALC (DEMO) - Body Mass Index of the individual
EDUC (DEMO) - Highest level of education an individual has completed
HINOTCOVE (DEMO) - Indicates whether the individual lacks health coverage
HRSLEEP (HABIT) - How many hours on average the individual sleeps per day
FRUTNO (HABIT) - How many times the individual ate fruit in a specified time period
VEGENO (HABIT) - How many times the individual ate vegetables in a specified time period
COFETEAMNO (HABIT) - How many times the individual drank coffee or tea sweetened with sugar or honey in a specified time period
CANCEREV (Target variable) - Indicates whether the individual has ever been diagnosed by a doctor or health professional as having cancer

Metrics

Precision - Of all the instances that the model classified as positive, how many were actually positive?

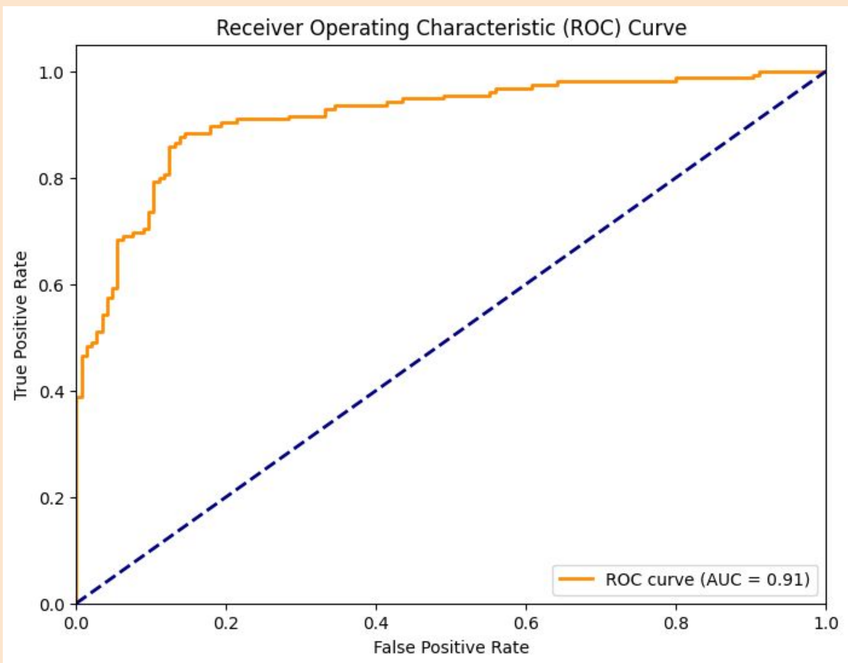
Recall - Of all the actual positive instances, how many did the model successfully identify?

ROC Curve - Curve that plots the true positive versus false positive rates at different thresholds. Curve that is closer to top left is better performing.

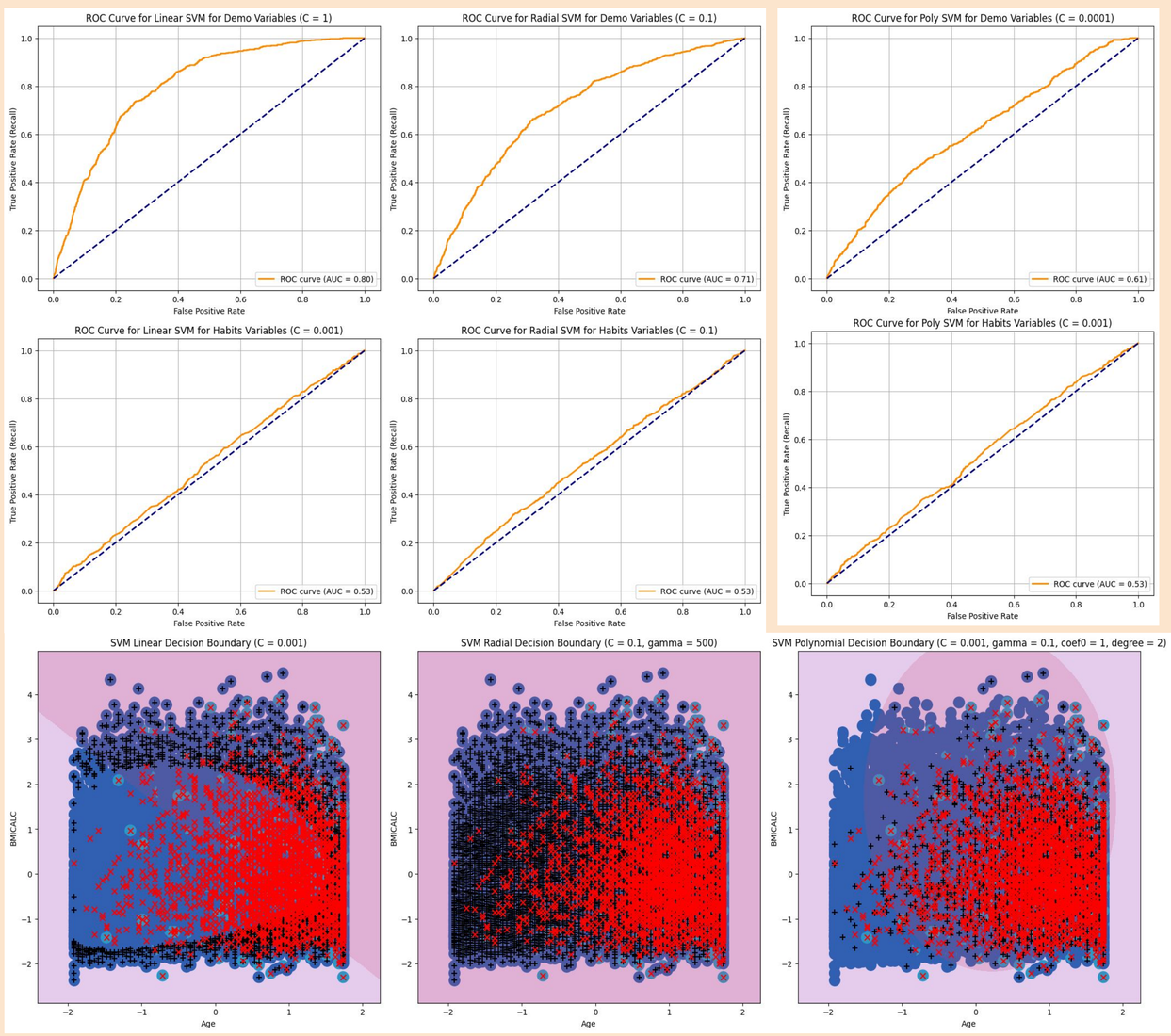
AUC - Area under the ROC curve that quantifies the overall performance of the model. Higher AUC indicates that the model has a higher probability of distinguishing between the two classes correctly.

Models

Linear SVM (Demo + Habit)
Radial SVM (Demo + Habit)
Polynomial SVM (Demo + Habit)
Features were filtered on invalid values like 999, 998, etc.



Results



Kernel Type	Variables	Class	AUC	Precision	Recall
Linear	Demographic	No Cancer	0.8	96%	63%
Linear	Demographic	Cancer	0.8	25%	83%
Linear	Habit	No Cancer	0.53	87%	67%
Linear	Habit	Cancer	0.53	14%	35%
Radial	Demographic	No Cancer	0.71	89%	90%
Radial	Demographic	Cancer	0.71	31%	29%
Radial	Habit	No Cancer	0.53	88%	71%
Radial	Habit	Cancer	0.53	15%	34%
Polynomial	Demographic	No Cancer	0.61	87%	100%
Polynomial	Demographic	Cancer	0.61	0%	0%
Polynomial	Habit	No Cancer	0.53	87%	96%
Polynomial	Habit	Cancer	0.53	17%	6%

The demographic features we selected do a decent job predicting the presence of cancer while the habitual features do not. The linear svm using demographic features did the best, which means factors like age, sex, education level, body mass index, and whether you have health insurance is effective in predicting cancer presence. One thing to note is that our dataset had class imbalance. Our test data had 4159 no cancer individuals and 630 cancer individuals (same ratio with train set), which is why our models do generally terrible when trying to classify cancer.
Even with some class imbalance and mediocre AUC scores, we can still somewhat see a trend that indicates habits like what you eat, how much you sleep or exercise, etc. is less predictive of cancer than demographic features. The results were also somewhat intuitive since older people likely get cancer at a higher rate than younger people, the healthier or more fit one is, the less likely they get cancer, and educational level and health insurance might indicate socioeconomic status so more privileged groups are less likely to get cancer than less privileged groups. Food and sleep habits might have a negative effect on individuals, but not enough to get cancer. Probably something like diabetes rather.
While demographic factors do impact the presence of cancer in individuals, they do not capture the entire scope of what determines the presence of cancer. If lawmakers were to make improvements, I would suggest creating policies that promote healthier and more active lifestyles, and more affordable healthcare. In our current state, healthcare is expensive and if you are poor, you basically die. Inflation and cost of living is rising while wages are remaining relatively the same so if people want to live a healthy lifestyle, they need help being able to afford healthier foods and gym memberships to stay fit.

[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..