Margin of Survival: SVM for Cancer Detection

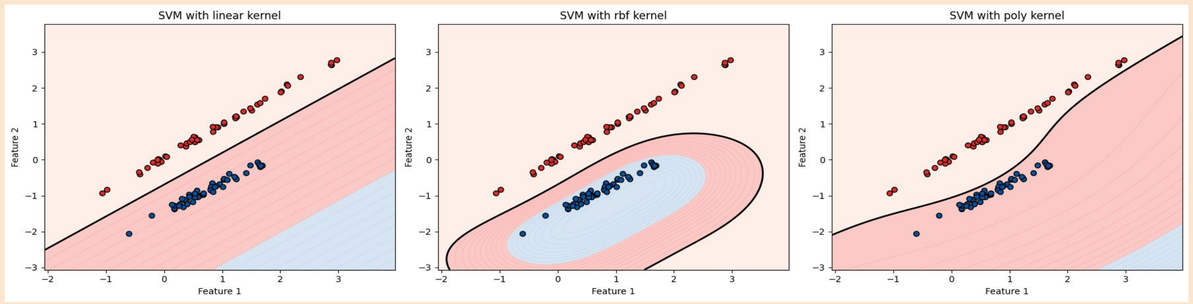
Christopher Yang | DATA5322

In this project, we explore how demographic and lifestyle factors can be predictive of the presence of cancer in an individual using support vector machine classifiers. Using data from the National Health Interview Survey, we built and compared SVM models with linear, radial, and polynomial kernels based on demographic variables (age, sex, education, and health insurance) and lifestyle habits (diet, sleep, physical activity). Our goal was to identify which factors could serve as strong predictors of cancer and reflect on the implications for public health policy.

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.

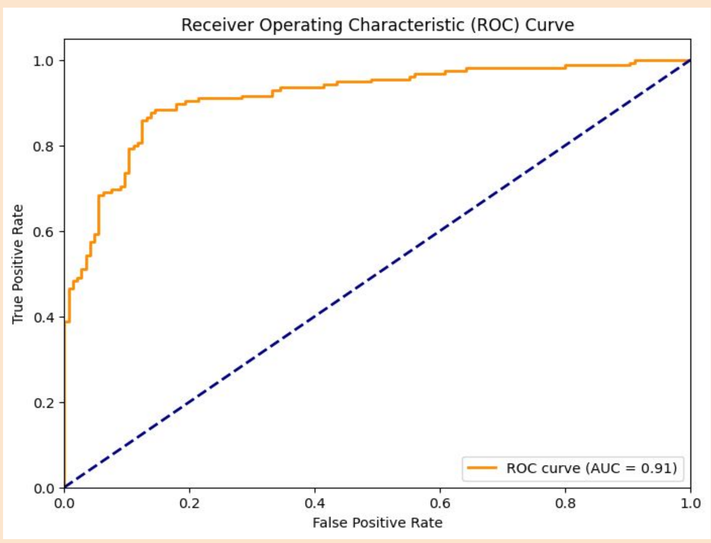
Models

Linear SVM (Demo + Habit)

Radial SVM (Demo + Habit)

Polynomial SVM (Demo + Habit)

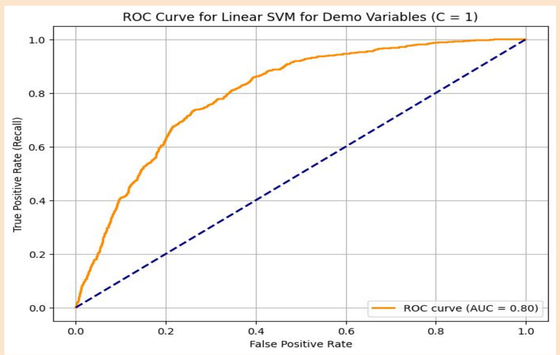
Features were filtered on invalid values like 999, 998, etc.

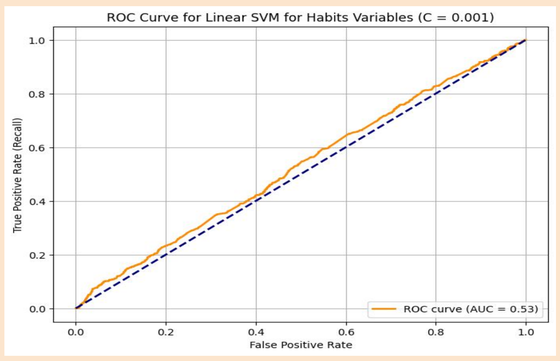


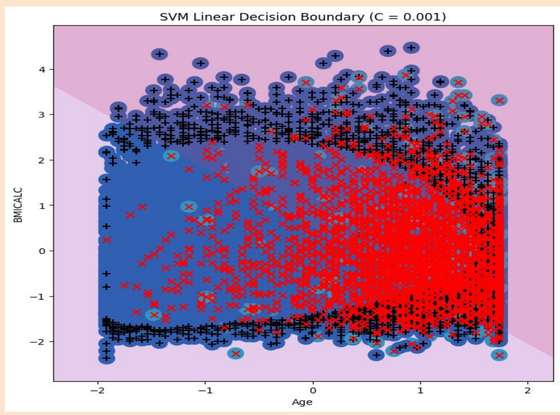
Data

- National Health Interview Survey (NHIS)
- ~ 35,000 individuals (~23,000 after cleaning)
- 48 variables (10 used: 9 feature, 1 target)

Results







Kernel Type	Variables	Class	AUC	Precision	Recall	Samples/Supports
Linear	Demographic	No Cancer	0.8	96%	63%	4159
Linear	Demographic	Cancer	0.8	25%	83%	630
Linear	Habit	No Cancer	0.53	87%	67%	4159
Linear	Habit	Cancer	0.53	14%	35%	630
Radial	Demographic	No Cancer	0.71	89%	90%	4159
Radial	Demographic	Cancer	0.71	31%	29%	630
Radial	Habit	No Cancer	0.53	88%	71%	4159
Radial	Habit	Cancer	0.53	15%	34%	630
Polynomial	Demographic	No Cancer	0.61	87%	100%	4159
Polynomial	Demographic	Cancer	0.61	0%	0%	630
Polynomial	Habit	No Cancer	0.53	87%	96%	4159
Polynomial	Habit	Cancer	0.53	17%	6%	630

Interpretation

- Demographic features we select did a decent job of predicting the presence of cancer while habitual features did not
- Linear SVM using demographic features performed the best, suggesting factors like age, sex, education level, body mass index, and whether an individual has health insurance are somewhat effective in predicting the presence of cancer
- Results are a little misleading since our dataset had class imbalance with 4159 “no cancer” individuals and 630 “cancer” individuals in the test set (same ratio in train set), which likely caused the models to struggle overall when trying to classify the two cases (does better classifying “non cancer” than “cancer”)
  - There is a trend showing habitual features like eating, sleeping, and exercising are less predictive of cancer than demographic features
  - This is somewhat intuitive since older people are more likely to get cancer than young people, healthier or more fit individuals are likely to get cancer, and education and health insurance may suggest socioeconomic status (with more privileged groups being less likely to get cancer)
- Food and sleep habits might negatively affect individuals, but not necessarily enough to cause cancer. These might be linked more to diabetes

Suggestions to Policy Makers

- While demographic factors do impact the presence of cancer, they do not capture the entire picture of what exactly determines whether someone gets cancer
- If lawmakers wanted to make improvements, suggest policy changes include:
  - Promoting healthier, more active lifestyles
  - Making healthcare more affordable and accessible
  - Addressing rising inflation and stagnant wages, which prevent people from affording healthy food options, gym memberships, and preventive healthcare services

[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..