Support Vector Machines for Predicting Disease Outcomes

Goal: To investigate commonly agreed-upon demographic and lifestyle predictors for heart attack through the optimization of linear, radial, and polynomial support vector machines to determine if models trained using heart attack data from males can result in accurate predictions for females

Technical Background: What is a Support Vector Machine?

Support vector machines (**SVM**s) transform current data into higher dimensions to maximize the distance between points

Data transformations can be computationally costly as the size of the dataset increases. SVMs avoid this by running computations on a subset of data points: **support vectors**

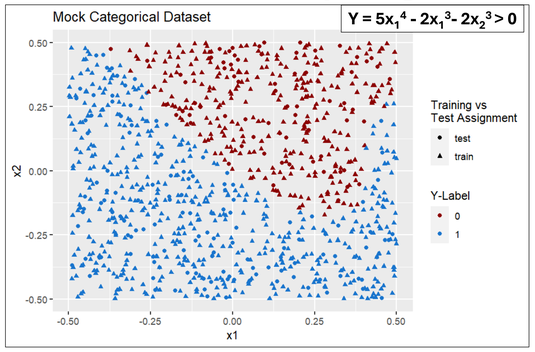
There are 3 types of **kernels** to run for SVMs:

- **Linear**: Line boundary
- **Polynomial**: Curved boundary
- **Radial**: Curved (potentially circular) boundary

The **parameters** that need to be tuned for these models are:

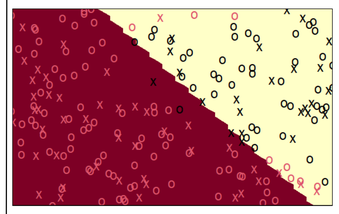
- **Cost** (Linear, Polynomial, and Radial): Determines the acceptable number of support vectors to include.
- **Degree** (Polynomial): Determines the polynomial relationship between points. Higher degrees allow for more inflections in decision boundary.
- **Coef0** (Polynomial): Constant term needed for polynomial equation
- **Gamma** (Radial): Inversely related to the sphere of influence for measuring distances between points

Results: Comparing SVM Kernels on Mock Data with Two Strong Predictor Variables with Classification Plots



Creating Mock Data: Mock data was created using a polynomial decision boundary for y based on variables x_1 and x_2. 1000 data points were simulated which were then randomly split into 80% for training data and 20% for test data for comparing SVM kernels.
SVM classification models to the right show the decision boundaries for the 20% test data.

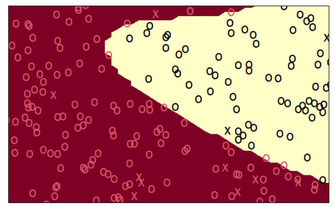
Linear Kernel



Parameters tested (and **optimal**):
Cost ← (0.001, 0.01, **0.1**, 1, 5, 10)
Training Support Vectors : 283
Time required to tune model: 0.45 sec
Training Misclassification Error: 12.50%
Test Misclassification Error: 8.5%

Pros: Fastest to tune; Cons: Highest errors

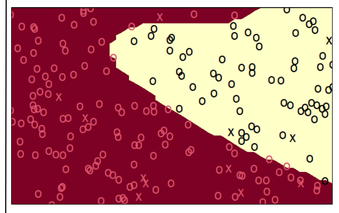
Polynomial Kernel



Parameters tested (and **optimal**):
Cost ← (0.001, 0.01, 0.1, 1, **5**, 10)
Degree ← (2, 3, **4**, 5, 8, 10)
Coef0 ← (0, 0.5, **1**, 2, 3, 5)
Training Support Vectors : 45
Time required to tune model: 33.7 sec
Training Misclassification Error: 0.50%
Test Misclassification Error: 1.50%

Pros: Low error; Cons: Slowest to tune, overfit data

Radial Kernel



Parameters tested (and **optimal**):
Cost ← (0.001, 0.01, 0.1, 1, 5, **10**)
γ ← (0.001, 0.01, 0.1, **1**, 5, 10)
Training Support Vectors : 10
Time required to tune model: 11.14 sec
Training Misclassification Error: 0.50%
Test Misclassification Error: 0.50%

Pros: Lowest test error; Cons: Slower than linear

Discussion: Mock Data

While both the radial and polynomial kernels had the same training error, the test error of the polynomial kernel increased, indicating it could have overfit this dataset more than the radial kernel. The linear kernel was by far the fastest model to tune, however it gave the highest errors.

Future datasets will need to balance time and accuracy for deciding which SVM to use. With a much larger dataset, the time required to tune could make the linear kernel ideal since it is the fastest. However, the linear kernel is limited to straight decision boundaries, and that could result in too low of an accuracy to be useful for some datasets, and so the time spent on a radial or polynomial SVM could be worth it.

Real-World Application: Heart Attack

Introduction: While men and women experience the same primary symptoms of heart attack[1], women globally are more likely to die from it[2]. SMVs will be applied to real-world data to investigate how heart attack events are influenced by lifestyle and demographic variables using men as the training data. The models will then be used to predict heart attacks in women to determine if the chosen predictors have the same influence for both sexes.

This data is from the 2022 National Health Interview Survey (NHIS) which collects health metrics from non-institutionalized Americans who are not on active military service through in-person interviews[3]. Data was harmonized by and accessed through IPUMS[4].

Methodology: Variable Clean-Up

Output Variable:

- **HEARTATTEV**: Binary; f a respondent was diagnosed with a heart attack (a.k.a. myocardial infarction). Only responses corresponding to "Yes" or "No" were kept.

Input Variables of Interest: Predictors were chosen based on what are considered generally as important factors in determining heart attack risk.

- **SEX:** Binary; sex of respondent. Entries without Female/Male as answers were removed.
- **AGE:** Numerical; age of the respondent. Entries without numerical age available were removed.
- **BMI**: Numerical; Body-mass index of participants. While the validity of BMI as a metric for health is debatable[5], it can overall provide insight into the respondent's weight while accounting for variations in height.
- **HRSLEEP**: Numerical; average number of hours respondent sleeps per day. Only responses with a provided number of hours were kept, and responses of less than one hour were rounded down to 0.
- **VIG10DMIN**: Numerical; Duration of vigorous activity of at least over 10 minutes. Entries without a real number of minutes were excluded, including those who had an "Extreme value" of exercise, since there is no way to differentiate when this was a true entry of >=12 hours or was due to calculation error.

After variable clean up, there were 8828 entries, including 27 and 115 heart attack events for females and males, respectively.

Methodology: SVM Creation from Real-World Data

4728 entries from male respondents were separated as the training data for creating SVMs, which accounted for roughly half of the dataset. It should be noted that only 115 heart attacks were reported out of these 4728 samples, which is a low number.

Test data used were the 27 entries that were female heart attacks, along with 270 randomly sampled entries from females without heart attacks to determine false positive rates.

The linear SVMs was tuned using the same range of cost values as what is listed above in the Mock Data section. For radial and polynomial SVMs, the ranges of cost, gamma, degree, and coeff0 parameter values tested were reduced due to concerns of runtime.

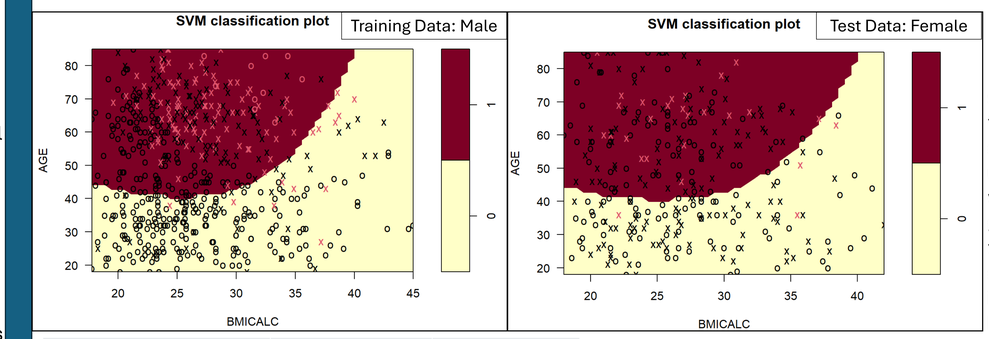
Results and Discussion: SVMs Trained on Male Data for Predicting Female Heart Attacks

Initial Attempt: Using the initial set-up described above in the "Methodology: SVM Creation from Data," all 3 SVM kernel types produced models that classified *all* entries as not having heart attacks. This is due to the fact that only ~2.5% of male responses were reported as having heart attacks, xo all model types determined it was best to ignore those and instead call all entries as having no heart attacks and accept a 2.5% error.

In an attempt to at least create models that can provide any insight into what variables are important, the dataset was reduced: similar to how the test data will include all women with heart attacks and some without, the training data was reduced to include all 115 male heart attacks, along with 460 randomly selected entries from males who did not have heart attacks. The hope is that if 20% of the dataset includes heart attacks versus the 2.5% before, there would be enough of a misclassification error if all were incorrectly categorized as non-heart attacks for the models to finally categorize some of them correctly.

For this second attempt, only linear and radial models were tuned. Even with a reduced range of parameters tested, tuning the polynomial SVM took over 15 minutes. While this second attempt includes significantly fewer data points and could take less time, the mock data above showed comparable performance between the polynomial and radial SVMs, so I will opt to skip the polynomial SVM.

Second Attempt: Both the linear and radial SVMs did finally predict some cases of heart attacks. Here are the radial results:



	Men (Training)	Women (Test)
# Heart Attacks	115	27
# W/O Heart Attacks	460	270
Misclassification Error	0.1443	0.1616
False Negative Rate	0.1252	0.0795
False Positive Rate	0.0435	0.1000

Radial SVM: Cost = 1, Gamma = 0.1

When plotting Age vs BMI, we see that for both males and females, the positive heart attack points cluster in the upper to middle left, which is where the decision boundary is generally located. This shows that these two variables together appear to have significant influence, although the true boundary is being dictated by other variables, too.

Overall, the error rates are comparable between males and females. The FNR decreased and FPR increased with the female dataset, but it could arguably be beneficial to have a model that is less likely to miss cases and more likely to raise alarms given the worse survival and treatment outcomes of women who suffer from heart attacks compared to men[2].

Discussion and Conclusions:

Technically, the radial SVM trained on male heart attack data was able to distinguish heart attacks among females to a comparable degree of accuracy. But one significant limitation within this dataset is the low frequency of heart attacks among it, which makes it highly doubtful that a thorough sampling of the population of those who have heart attacks was done.

A second significant limitation to this dataset is survivor's bias. Since this survey relies on in-person interviews, only people who suffered from and *survived* a heart attack could participate. This could explain why the plot of Age vs BMI below shows the clustering of heart attack cases in the upper left, on the lower side of BMI, because those with higher BMIs could have died from heart attacks. There are conflicting findings of patients with higher BMIs being less likely to die from sudden cardiac arrest[6], while women with a higher BMI were more likely to suffer from sudden cardiac death[7]. These conflicting findings highlight the need for further investigation into what variables determine the risk of a heart attack, and the need for special consideration into the role that sex plays in heart attack risk.

Given recent guidelines from the NIH dictating that grants will not be awarded for projects that include diversity, equity, and inclusion efforts[8], I would hope that the results presented on this poster (or more accurately the lack of clear results) prove that there is more work to be done for predicting disease outcomes. Just like how this dataset included too few heart attacks to draw significant conclusions, failure to include a diverse sampling of a population will result in a skewed dataset, and when attempting to model diseases, this will lead to worse health outcomes.

References:

[1] Salamon, Maureen. "Women's heart symptoms not so different after all." *Women's Health,* 1 June 2022, https://www.health.harvard.edu/womens-health/womens-heart-symptoms-not-so-different-after-all
[2] Lu H, Hatfield LA, Al-Azazi S, Bakx P, Banerjee A, Burrack N, Chen YC, Fu C, Gordon M, Heine R, Huang N, Ko DT, Lix LM, Novack V, Pasea L, Qiu F, Stukel TA, Uyl-de Groot CA, Weinreb G, Landon BE, Cram P. Sex-Based Disparities in Acute Myocardial Infarction Treatment Patterns and Outcomes in Older Adults Hospitalized Across 6 High-Income Countries: An Analysis From the International Health Systems Research Collaborative. Circ Cardiovasc Qual Outcomes. 2024 Mar;17(3):e010144. doi: 10.1161/CIRCOUTCOMES.123.010144. Epub 2024 Feb 8. PMID: 38328914.
[3] CDC, "About NHIS." *National Health Interview Survey,* 20 Nov. 2024, https://www.cdc.gov/nchs/nhis/about/index.html
[4] 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.org Links to an external site.
[5] AMA. "AMA adopts new policy clarifying role of BMI as a measure in medicine." *Press Releases*, 14 June 2023, https://www.ama-assn.org/press-center/ama-press-releases/ama-adopts-new-policy-clarifying-role-bmi-measure-medicine
[6] Matinrazm S, Ladejobi A, Pasupula DK, et al. Effect of body mass index on survival after sudden cardiac arrest. *Clin Cardiol*. 2018;41(1):46-50. doi:10.1002/clc.22847
[7] Chiuve, S, Sun, Q, Sandhu, R. et al. Adiposity Throughout Adulthood and Risk of Sudden Cardiac Death in Women. J Am Coll Cardiol EP. 2015 Dec, 1 (6) 520–528.https://doi.org/10.1016/j.jacep.2015.07.011
[8] NIH. "Notice of Civil Rights Term and Condition of Award." *NIH Grants & Funding,* 21 Apr. 2025, NOT-OD-25-090: Notice of Civil Rights Term and Condition of Award