Using Support Vector Machines to Predict Health Conditions

Introduction

This research uses support vector machines (SVMs) to predict whether someone has had diabetes or cancer based on health variables. The data is from the IPUMS Health Survey [1]. It contains variables about the respondents' demographics, health habits, and whether they have had 5 different health conditions. I used the following models and predictors:

- An SVM with a linear kernel to predict diabetes presence based on BMI, weight, and age
- An SVM with a polynomial kernel to predict diabetes presence based on BMI, work hours, age, and height
- An SVM with a radial kernel to predict cancer presence based on the variables above and the days used alcohol yearly.

Theoretical Background

A support vector classifier (SVC) separates classes by fitting a linear boundary between them and maximizing the margin. The margin is the distance between the boundary and the closest points, or the support vectors. Some points can be misclassified or in the margin.

The equations that formulate an SVC are as follows:

$\max_{\beta_0,\beta_1,\dots,\beta_p,\epsilon_0,\dots,\epsilon_n,M} M$ subject to:

$\sum_{j=1}^{p} \beta_j^2 = 1$ (1)

$y_i(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) \geq M(1 - \epsilon_i)$ (2)

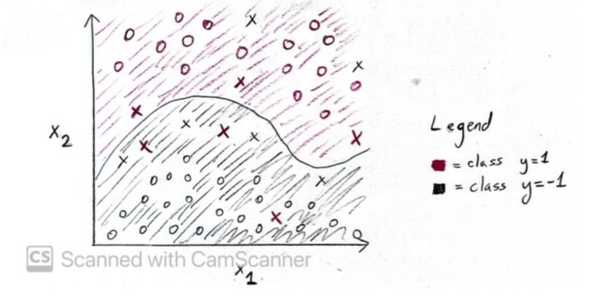
$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$ (3)

M is the margin size, and ϵ_i is the distance of a data point from the correct side of the margin. (2) allows some points to be on the wrong side, but, in (3), the tuning parameter C constrains how many points and how far away they can be.

SVMs can create nonlinear decision boundaries by combining SVCs with kernels. Kernels transform the data to a higher dimension, where it is now linearly separable.

A decision boundary with a kernel K is $f(x) = \beta_0 + \sum_{i \in \delta} \alpha_i K(x, x_i)$, where δ are the support vectors. A linear kernel is equivalent to an SVC. A polynomial kernel has the form $K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$. It creates polynomial relationships between points, allowing curved boundaries. *d* is the polynomial degree, a tuning parameter. Another nonlinear kernel is the radial kernel: $K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$. Gamma, a positive constant, is a tuning parameter.

The plot below shows an SVM with a nonlinear kernel. The points and the regions are color-coded by the classes. All the points above the decision boundary are classified as y = 1, and all the points below it are classified as y = -1. The x's are support vectors, while the round points are not support vectors. Notice that some points are on the wrong side of the decision boundary, but most are classified correctly.



Methodology

To pre-process the data, I selected the relevant variables, I dropped the rows where the response was "Unknown". I recoded the class labels to strings, and I performed standardization. A huge problem was class imbalance. About 93% of the respondents haven't had diabetes or cancer, and only about 7% have. Also, the classes were completely overlapped with no boundary. To mitigate the class imbalance, I used cost-sensitive SVMs. Cost-sensitive SVMs weight C differently for each class. Assigning a larger weight to the minority class forces minority points to be classified correctly.

I used 2-fold cross-validation for hyperparameter tuning. During cross-validation, I maximized the F1 score. Accuracy is misleading for imbalanced classes, and F1 score is a better metric because it balances both precision and recall.

For the linear kernel, I tried cost values of 0.01, 0.1, 1, and 10. The best was 1. For the polynomial kernel, I tried the same cost values, I tried degrees of 2, 3, and 4, and I tried coef0 values of 0, 1, and 2. The best model had cost = 10, degree = 2, and coef0 = 0. For the radial kernel, I tried cost = 0.01, 0.1, and 1; and gamma = 0.5, 1, and 2. To evaluate the models, I used the overall training and test accuracy, the training and test accuracy for each class, and ROC curves.

Discussion

Age, BMI, and weight highly impacted diabetes predictions. Older people, people with a higher BMI, and people with a higher weight were predicted to have diabetes. Also, people with a higher BMI or weight are at no risk of cancer. People with a low weight, however, are predicted to have cancer at younger ages. Interestingly, people with a moderate alcohol use are the least at risk for cancer. Finally, older people who work too few or too many hours are at risk of cancer.

Limitations were class imbalance and class overlap. The overlap was so severe that the minority class data points were located right in the middle of the majority class ones. So, a true, reasonable decision boundary doesn't exist. Also, the classes were so imbalanced that, even with cost-sensitive SVMs, most models had a lower accuracy for the minority class than the majority one. The only model that did not have this problem, overfit the minority class. An extension is to use neighborhood-based undersampling [2] as a data preprocessing step, to address both the imbalance and the overlap. This technique removes majority data points from the overlapping region while also minimizing information loss.

Older people are more likely to have diabetes, regardless of other health variables. So, I suggest that policymakers require insurance companies to cover insulin, making insulin easier to obtain for people with diabetes. I found that underweight people are more at risk of cancer, and a very common cause of being underweight is not having enough to eat due to poverty. So, I believe policymakers should invest more in food banks and food stamps, and that they should work towards addressing poverty.

Conclusion

Unfortunately, the findings linking weight and BMI to diabetes could be maliciously used to stigmatize people who have a higher weight or body size. However, using BMI to determine if someone is "healthy" can be misleading [3]. Firstly, BMI was originally developed by studying a population of only white men, so BMI may not have accurate results for people of all ethnicities and genders. Secondly, BMI should be used along with other metrics, like blood pressure, cholesterol, and glucose levels, to get an accurate picture of someone's health. Hopefully, my findings about BMI could motivate medical researchers to find better metrics for health status than BMI. Also, people could use the findings about age and diabetes to store emergency insulin in establishments where older people commonly spend time. The findings about working hours and cancer could cause employers to prioritize work-life balance more. Some researchers link mental stressors to a higher risk of physical health conditions. They could use my findings to support their research. The findings that underweight people are at risk of cancer could be beneficial for alleviating poverty. Hopefully,, United States lawmakers would work on addressing the problem of socioeconomic class inequality in America.

References

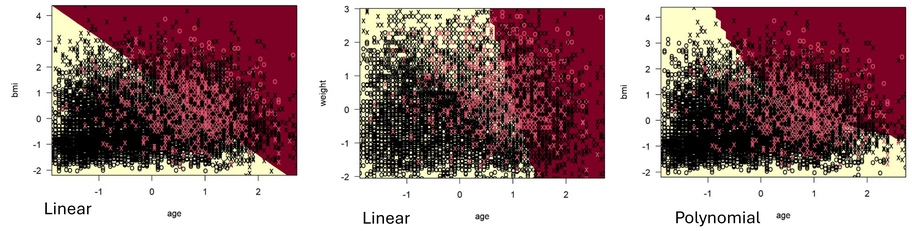
[1] Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Daniel Backman, Annie Chen, and Stephanie Richards. IPUMS Health Surveys: National Health Interview Survey, Version 7.4 [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D070.V7.4. Links to an external site.http://www.nhis.ipums.org.

[2] Vuttipittayamongkol, Pattaramon, and Eyad Elyan. "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data." *Information Sciences* 509 (2020): 47-70.

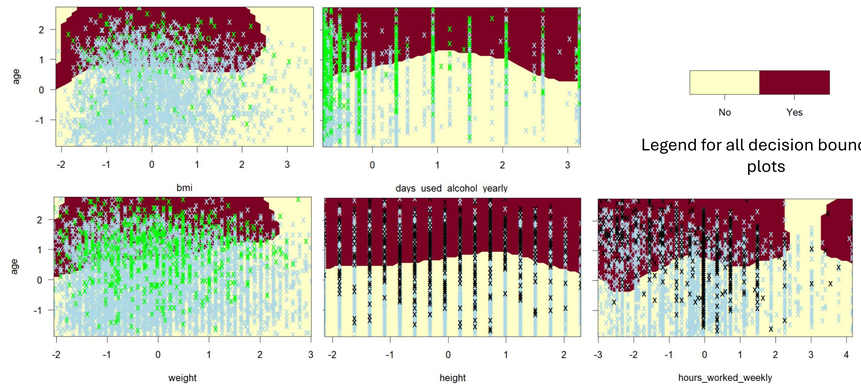
[3] Katella, Kathy. "Why You Shouldn't Rely on BMI Alone." *Yale Medicine*, 4 Aug. 2023, www.yalemedicine.org/news/why-you-shouldnt-rely-on-bmi-alone.

Results

Older people and people with a high BMI were predicted to have diabetes. Also, the linear model predicted people with a higher weight to have diabetes. The plots below are from both models that predict diabetes.



The radial kernel created interesting trends between the predictors. People with a higher BMI and a higher weight are at no risk of cancer. People with a low weight are more likely to have cancer at a younger age. There is a parabolic trend for both alcohol use and height. People whose alcohol use or height is ~ 1 standard deviation (SD) above the mean, can be older without being at risk for cancer. Furthermore, older people who work fewer hours are at risk of cancer. All people with working hours 2-3 SDs above the mean are at no risk. But older people whose working hours are 3-4 SDs above the mean are at risk. The plots below are for predicting whether someone has cancer (the y-axis is shared).



Relevant accuracy metrics are below:

Test Acc. - Overall	Test Acc. – Majority Class	Test Acc. – Minority Class
77.3%	78.5%	59.3%

Linear kernel to predict diabetes

Test Acc. - Overall	Test Acc. – Majority Class	Test Acc. – Minority Class
81.9%	83.9%	52.1%

Polynomial kernel to predict diabetes

Train Acc. - Overall	Train Acc. – Majority Class	Train Acc. – Minority Class	Test Acc. - Overall	Test Acc. – Majority Class	Test Acc. – Minority Class
70.3%	69.2%	85.9%	68.2%	68.0%	70.7%

Radial kernel to predict cancer

The linear model is better than the polynomial one for predicting people who *have* diabetes. The polynomial one is better for predicting people who *don't*, and it has the highest overall test accuracy. The radial model is overfitting the minority class because the test accuracy for that class is much lower than the training accuracy. However, I used the best class weights I could find. Despite the overfitting, the model for cancer predicts the minority class the best.

The ROC curves to the right show that the linear and polynomial models perform about equally well, considering all trade-offs. The radial model performs the worst.