Predicting Heart Attack History in Men Through The Usage of Support Vector Machines

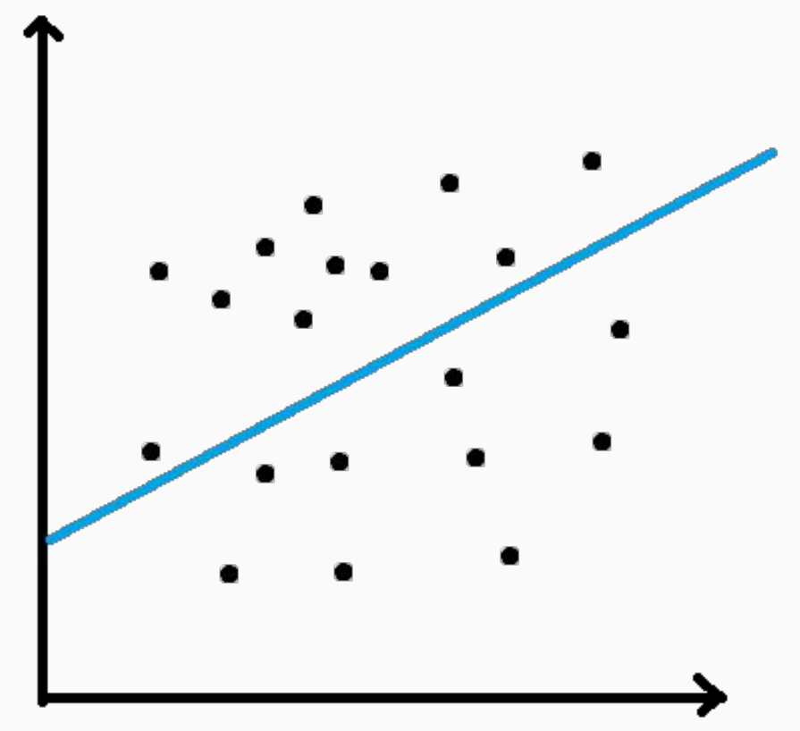
Aakash Krishna | 05/09/2025

For this project, we hope to be able to predict the presence of heart attack history in male respondents based on their demographics and habits using data obtained from the National Health Interview Survey. We will be using support vector machine learning models in order to do so.

Theoretical Background

Support Vector Machines (SVMs for short) are supervised machine learning models that compute an optimal "hyperplane", which you can think of as a line that separates individual data points that are in different classes. There are three kinds of SVMs we will concern ourselves with:

Linear

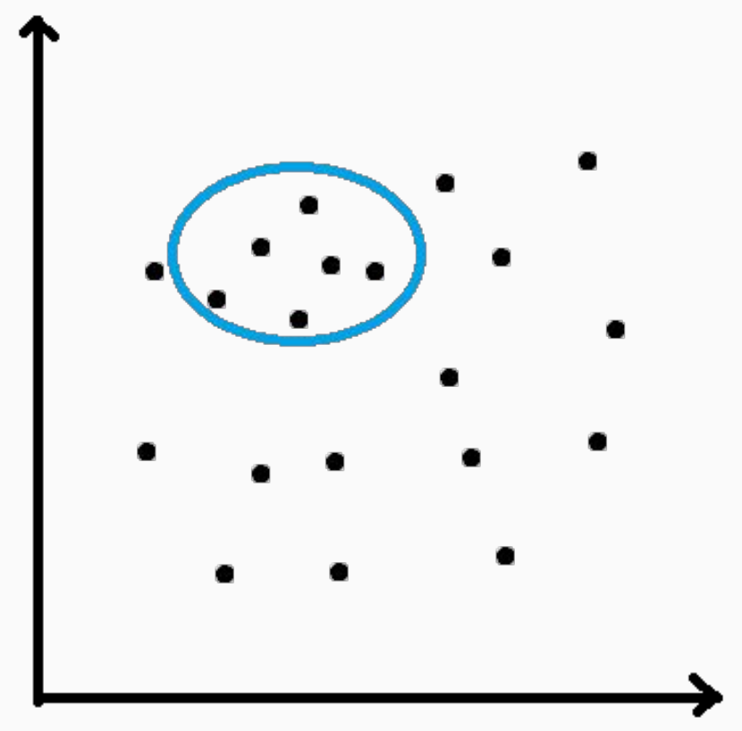


Classes are separated by a straight line/flat plane.

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

They are best when boundaries are linearly separable.

Radial

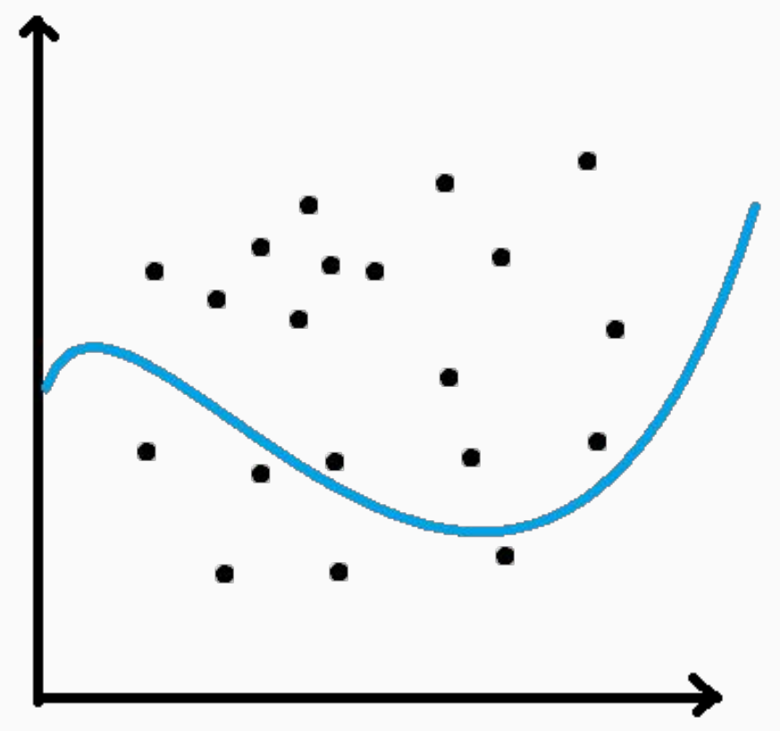


Classes are separated based on computed distance between surrounding points.

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

They are best when boundaries are highly non-linear.

Polynomial



Classes are separated by curved boundaries up to a specified degree.

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

They are best when boundaries are slightly non-linear.

Methods

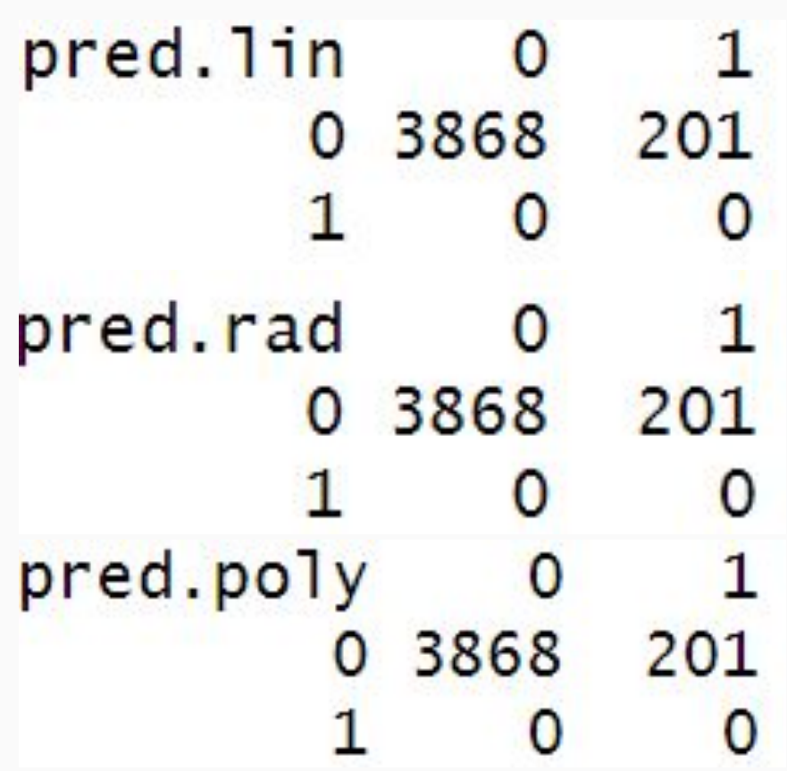
Our target variable is HEARTATTEV, a measure of if the patient has ever reported a heart attack. In order to predict this, we use the following features (orange variable name means said feature was excluded from radial model, blue variable name means an exclusion from the polynomial model, normal text color indicates it was used in all three models):

Tuning Parameters	
Cost (c)	Inversely proportional to margin size, "budget" for points on wrong margin side
Gamma (γ)	Inversely proportional to sphere of influence where similarity of points matters
Degree (d)	Number of curves in line

Variable Name	Measure
HEIGHT	Height in inches without shoes
BMICALC	Calculated Body Mass Index
MOD10DMIN	Average duration of moderate activity per day
JUICEMNO	Number of times consuming 100% fruit juice in last month
SALSAMNO	Number of times consuming salsa in last month
TOMSAUCEMNO	Number of times consuming tomato sauce in last month
SPORDRMNO	Number of times consuming sports drinks in last month
FRTDRINKMNO	Number of times consuming sugary fruit drinks in last month
COFETEAMNO	Number of times consuming coffee/tea in last month
HRSLEEP	Usual hours of sleep per day
ALCDAYSM	A variable created from ALCDAYSYR in the dataset, to record number of days alcohol is consumed per month on average.
HTATK	Factor created from HEARTATTEV to make predicting HEARTATTEV easier.

Results

Confusion Matrices



```
pred.lin     0    1
       0  3868  201
       1     0    0
```

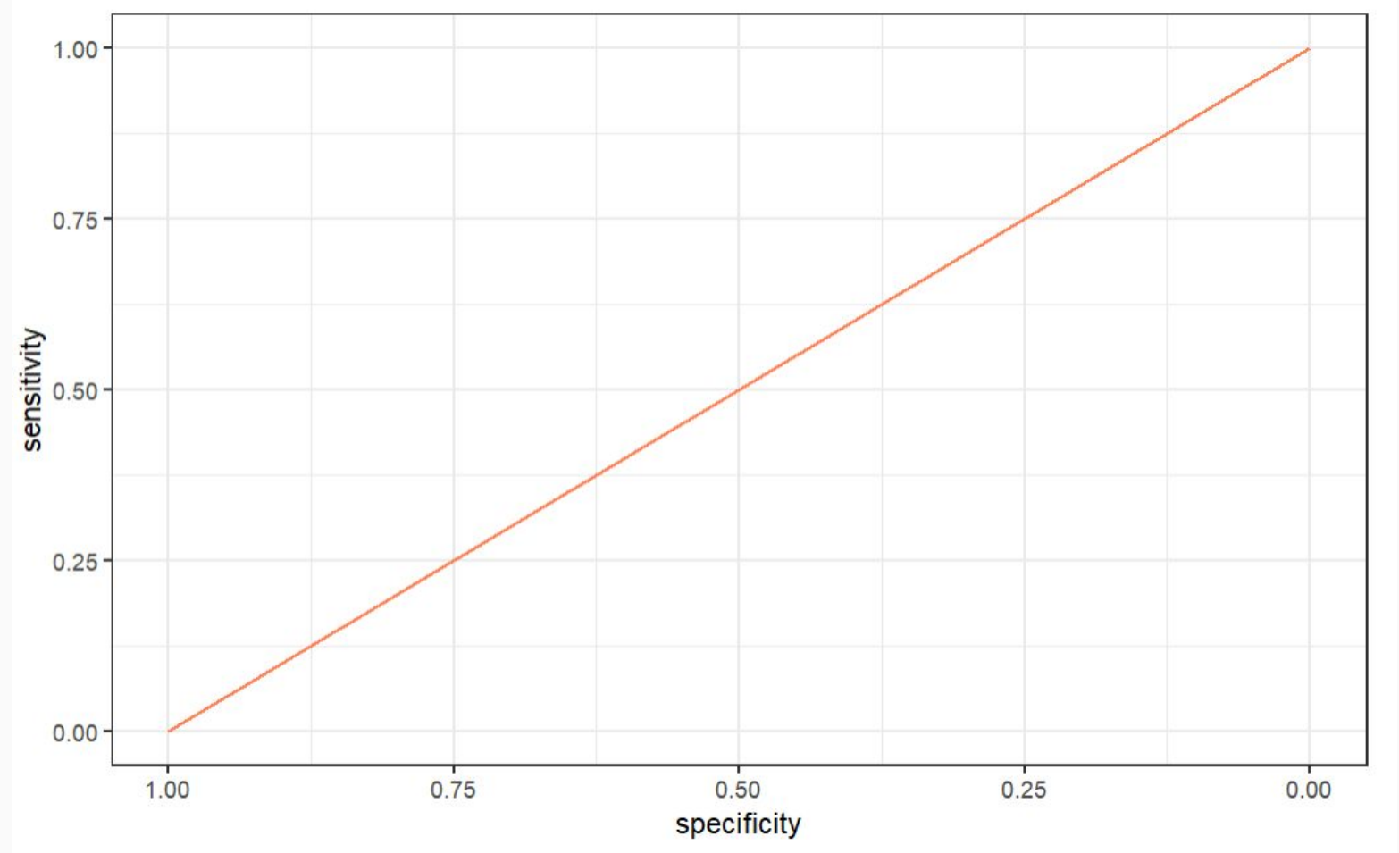
(Linear)

```
pred.rad     0    1
       0  3868  201
       1     0    0
```

(Radial)

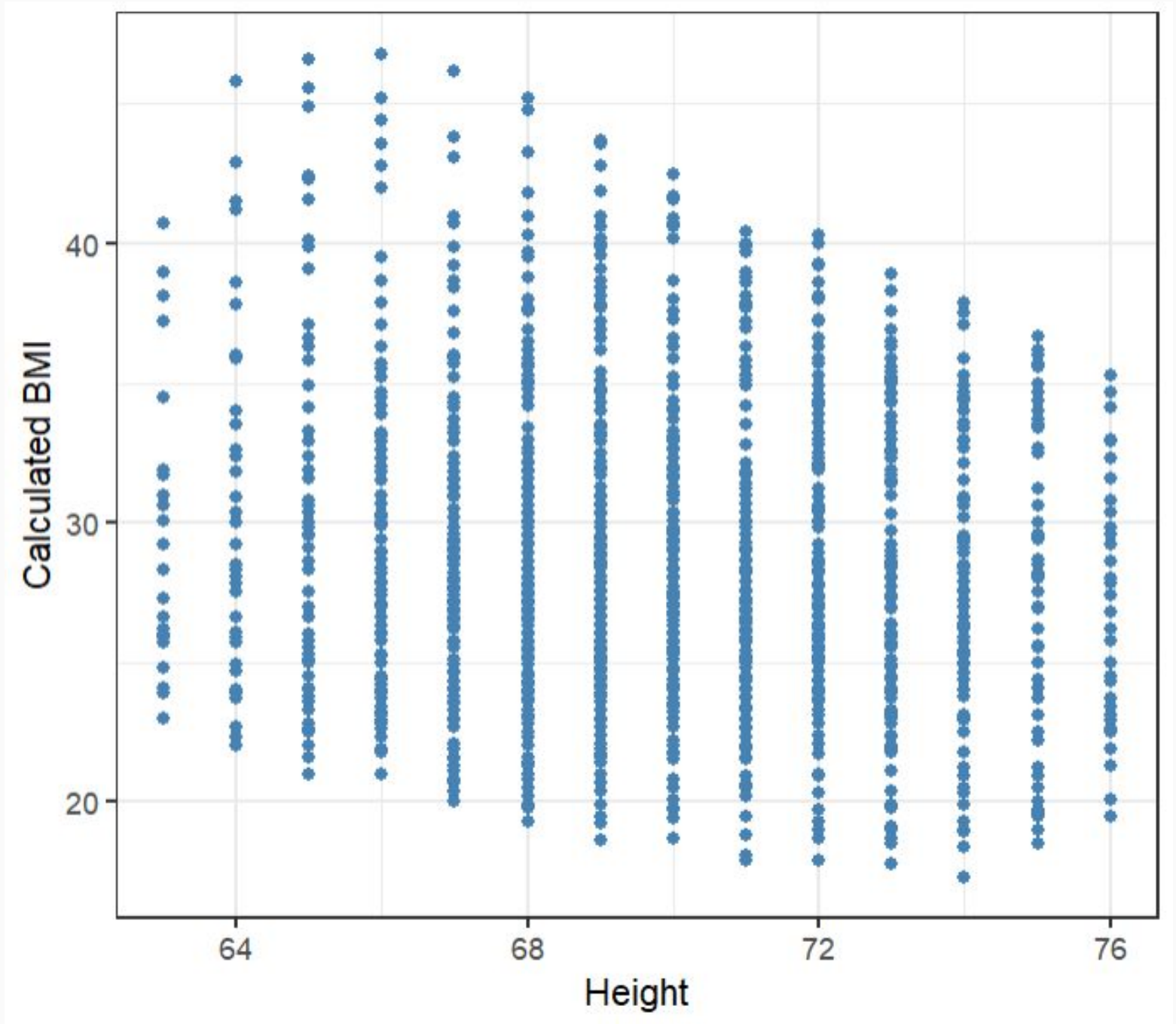
```
pred.poly     0    1
        0  3868  201
        1     0    0
```

(Polynomial)

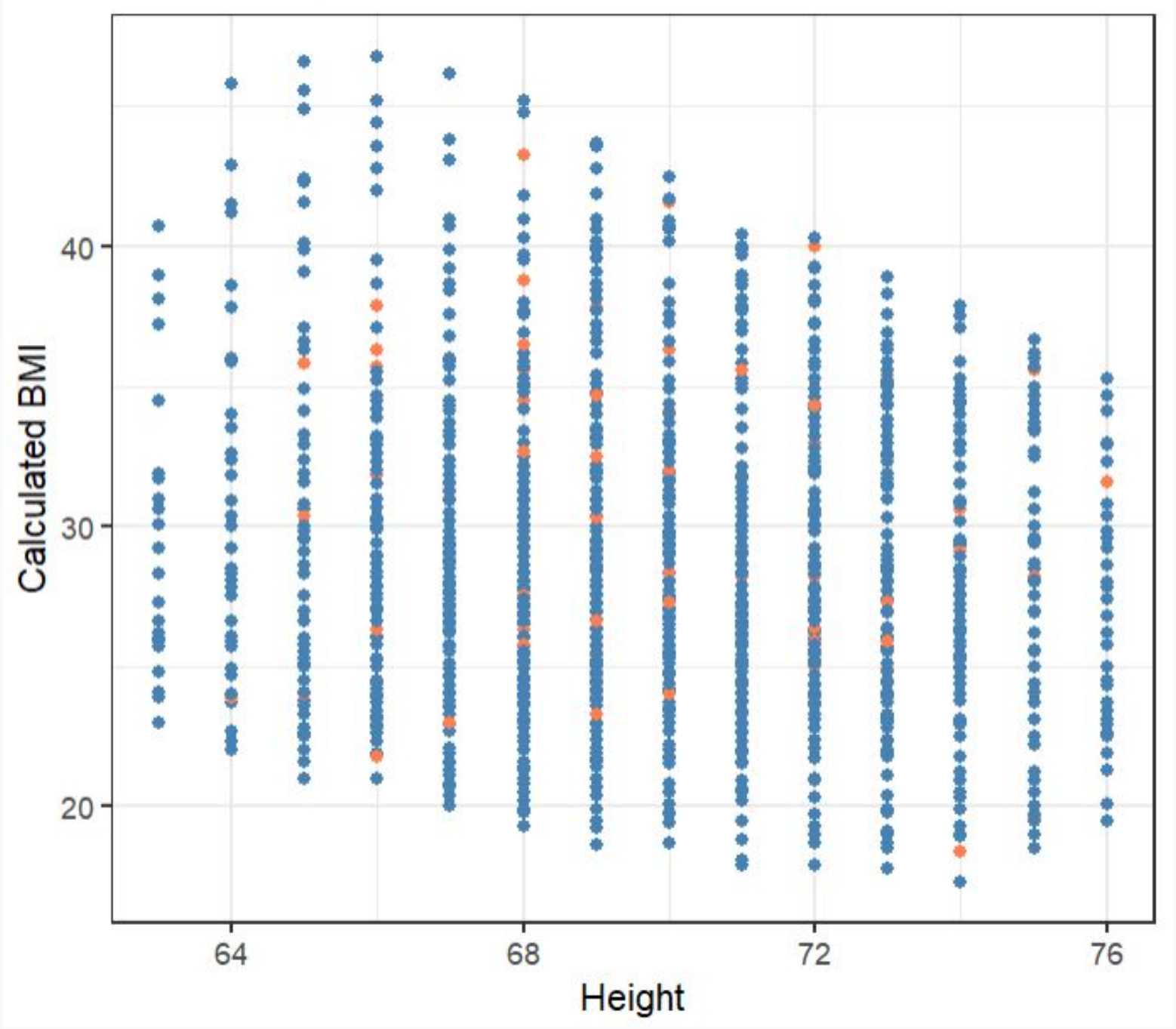


Area Under Curve (All Three Models)

Prediction Graph (All Models)



Actual Graph (for comparison)



Dot Color Significance

No Heart Attack History

Heart Attack History

Discussion

All SVM models default to predicting 0 (no heart attack history), with all 201 patients with heart attack history in the test data set being incorrectly classified as not having said history. This is most likely due to severe class imbalance, which is difficult to fix without significantly increasing the runtime of the tuning process, which is already quite large.

Kernel Type	Accuracy	MSE Score	AUC Score
Linear	95%	0.049	0.5
Radial	95%	0.049	0.5
Polynomial	95%	0.049	0.5

All models technically perform equally well, but are still unusable predictors of heart attack rates. No combination of factors seems to impact this.

Conclusions

The biggest issue that was faced by this project was the issue of class imbalance that was primarily due to the small proportion of patients who had reported suffering from a heart attack in the data set, making any conclusion from our models unreliable. It is difficult to make suggestions based on the results received, but it seems fair to conclude that factors such as sleep per day, body mass index, and regularity of alcohol consumption do affect heart attack occurrence to some extent. If they did not, we might have seen models that entirely scaled off of one or two of the provided factors.

In conclusion, it is difficult to find accurate measures of specific factors that lead to heart attacks. Policy makers will have to focus on overall lifestyle change along with robust medical and nutritional support in order to better combat the rise of heart attack occurrences in the population they represent.

Citation

[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