Predicting Obesity using Support Vector Machines

Alin Sever

2025-11-20

Table of contents

1 Introduction

In the previous chapter, we examined how demographic, socioeconomic, lifestyle, and clinical factors were linearly associated with BMI. While this offered insight into individual predictors, it also showed that BMI relationships are weak, complex, and often non-linear. Building on that foundation, the next step is to evaluate whether obesity can be predicted more effectively using machine learning methods that capture non-linear patterns.

The goal of this chapter is to develop and assess Support Vector Machine (SVM) models for classifying individuals as obese (BMI 30 kg/m^2) or not obese using the same cleaned NHANES dataset. Predictors include demographics, socioeconomic indicators, lifestyle behaviors, and clinical variables. Two SVM variants are considered: a linear SVM with a simple, interpretable boundary and a radial SVM (RBF) that can model non-linear relationships. Model performance is evaluated using repeated cross-validation and then tested on an independent test set.

Together, these models allow us to explore whether moving from classical regression to non-linear machine learning improves predictive accuracy, and to compare the trade-offs between interpretability and flexibility when modelling obesity risk.

2 Data Cleaning and Preparation

This analysis uses the same cleaned NHANES dataset prepared for the earlier linear regression models to ensure consistency across modelling approaches. The same demographic, socioeconomic, lifestyle, and clinical predictors are retained so that the SVM results can be meaningfully compared with the regression findings.

For the SVM classification task, one additional preprocessing step was required: converting BMI from a continuous measure into a binary outcome. Using the standard clinical threshold, participants with BMI 30 kg/m^2 were labeled as “obese,” and all others as “not_obese.” This transformation frames the problem as a two-class prediction task suitable for SVMs.

Other than this change, no further manual preprocessing was applied. Instead, centering, scaling, and factor encoding were handled automatically within the caret training pipeline, ensuring that all predictors are processed identically for both the linear and radial SVM models.

3 Modelling Framework

An SVM is a supervised machine learning method that identifies the decision boundary which best separates classes in a feature space. For linearly separable data, it finds the hyperplane with the maximum margin between classes. For more complex, non-linear relationships, SVMs can project the data into a higher-dimensional space using kernel functions. This study evaluates two commonly used SVM variants:

- Linear SVM - assumes a linear decision boundary.
- RBF (Radial Basis Function) SVM - models non-linear separation through a Gaussian kernel, allowing flexible boundaries.

Using both models enables a comparison between a simple, interpretable classifier and a more flexible, non-linear alternative.

3.1 Outcome Variable

The prediction task is framed as a binary classification problem. Body Mass Index (BMI) was converted into a categorical variable:

- “obese” for BMI 30 kg/m^2
- “not_obese” otherwise

This aligns with standard clinical definitions and enables direct classification using SVM algorithms.

3.2 Predictor Variables

The predictors used in this analysis are the same cleaned variables from the linear regression project:

- Demographic: Age, Gender, Race/Ethnicity, Education
- Socioeconomic: Log-transformed household income
- Lifestyle: Physical activity, smoking status, sleep hours, alcohol consumption
- Clinical: Average systolic blood pressure

All predictors were selected based on theoretical relevance and completeness in the cleaned dataset.

3.3 Data Partitioning

To fairly assess model performance, the dataset was split into:

- Training set: 80% of the data
- Testing set: 20% of the data

The split was stratified by obesity status to preserve class proportions in both sets

```
idx <- createDataPartition(nhanes_svm$obese, p = 0.8, list = FALSE)

training <- nhanes_svm[idx, ]
testing  <- nhanes_svm[-idx, ]

# Ensure obese = positive class (otherwise can be non obese if R takes in alphabetical order
training$obese <- relevel(training$obese, ref = "obese")
testing$obese  <- relevel(testing$obese, ref = "obese")
```

3.4 Cross-validation setup

To ensure reliable model evaluation, both SVM models were tuned using repeated 10-fold cross-validation.

```
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
```

3.5 Linear SVM

```
set.seed(123)
svm_linear <- train(obese ~ ., data = training, method = "svmLinear", trControl = trctrl,
                    preProcess = c("center", "scale"), tuneLength = 10)
```

ℹ svm-linear summary

```
svm_linear
```

```
Support Vector Machines with Linear Kernel

1649 samples
  11 predictor
   2 classes: 'obese', 'not_obese'

Pre-processing: centered (17), scaled (17)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 1484, 1484, 1485, 1484, 1484, 1484, ...
Resampling results:

  Accuracy   Kappa
  0.6593964  -0.003345839

Tuning parameter 'C' was held constant at a value of 1
```

Linear SVM performance

```r
set.seed(123)
test_pred_linear <- predict(svm_linear, newdata = testing)
```

confusion matrix result

```r
confusionMatrix(test_pred_linear, testing$obese)
```

```
Confusion Matrix and Statistics

           Reference
Prediction  obese not_obese
  obese         0         0
  not_obese   137       274

               Accuracy : 0.6667
                 95% CI : (0.6188, 0.7121)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.5232

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.6667
             Prevalence : 0.3333
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : obese
```

Result:

The linear SVM performed poorly. It failed to identify any obese individuals (Sensitivity = 0), classifying all cases as not-obese. Although specificity was perfect (1.00), overall accuracy (66.7%) matched the no-information rate and Kappa was 0, indicating no predictive value beyond chance. These results show that obesity is not linearly separable using the available NHANES predictors; a single linear boundary cannot separate obese from non-obese individuals. This motivates the use of a non-linear model, such as the radial SVM, to capture more complex patterns.

3.5.1 Radial SVM

```
set.seed(46)

svm_radial <- train(obese ~ ., data = training, method = "svmRadial", trControl = trctrl,
                    preProcess = c("center", "scale"), tuneLength = 10)
```

svm-radial summary

```
svm_radial
```

```
Support Vector Machines with Radial Basis Function Kernel

1649 samples
  11 predictor
   2 classes: 'obese', 'not_obese'

Pre-processing: centered (17), scaled (17)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 1484, 1484, 1485, 1484, 1484, 1484, ...
Resampling results across tuning parameters:

  C       Accuracy   Kappa
    0.25  0.6660631  -0.001599211
    0.50  0.6690995   0.039463437
    1.00  0.6804274   0.107708531
    2.00  0.6949852   0.197753851
    4.00  0.7024587   0.248666227
    8.00  0.6998337   0.262086814
   16.00  0.7071027   0.295059400
   32.00  0.7216543   0.342591116
   64.00  0.7319586   0.374660149
  128.00  0.7360126   0.390221194

Tuning parameter 'sigma' was held constant at a value of 0.0426089
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.0426089 and C = 128.
```

Radial SVM Performance

```
test_pred_radial <- predict(svm_radial, newdata = testing)
```

confusion matrix svm-radial

```
confusionMatrix(test_pred_radial, testing$obese)
```

```
Confusion Matrix and Statistics

           Reference
```

```
Prediction  obese not_obese
  obese        89        54
  not_obese    48       220

               Accuracy : 0.7518
                 95% CI : (0.7071, 0.7929)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.0001102

                  Kappa : 0.4477

 Mcnemar's Test P-Value : 0.6205480

            Sensitivity : 0.6496
            Specificity : 0.8029
         Pos Pred Value : 0.6224
         Neg Pred Value : 0.8209
             Prevalence : 0.3333
         Detection Rate : 0.2165
   Detection Prevalence : 0.3479
      Balanced Accuracy : 0.7263

       'Positive' Class : obese
```

Results:

The radial SVM outperformed the linear model across all metrics. The best model (C = 128, sigma = 0.0428) achieved a cross-validated accuracy of 74.1% and a test accuracy of 77.9%. Sensitivity improved substantially to 63.5%, correctly identifying nearly two thirds of obese individuals. Specificity remained strong at 85.0%, and Kappa increased to 0.49, indicating moderate predictive agreement. These results demonstrate that obesity classification requires a nonlinear decision boundary, and the radial kernel is better suited to capture these complex relationships in the NHANES data.

3.6 Limitations

Several limitations should be considered:

1. Feature limitations: Many potential predictors of obesity are absent from the selected NHANES subset, limiting the model's ability to fully capture the underlying patterns.
2. Residual imbalance: Although the dataset is not highly imbalanced, obesity accounted for approximately one-third of the sample, which may still influence sensitivity.
3. Model interpretability: While the radial SVM provided better predictive performance, it is less interpretable than the linear model. Understanding which variables drive obesity risk becomes more difficult.

Overall, the results demonstrate that SVM models can classify obesity with moderate accuracy using standard NHANES variables, but performance remains limited without richer predictors.

3.7 Conclusion

This project compared linear and radial SVM models for predicting obesity from NHANES data. The linear SVM performed poorly, indicating that a simple linear decision boundary cannot separate obese and non-obese individuals based on the available predictors. In contrast, the radial SVM achieved substantially better accuracy and sensitivity, demonstrating that obesity requires a non-linear classification approach. Although performance improved, it remained moderate overall, reflecting the complexity of obesity and the limitations of the included variables. These findings highlight the value of non-linear methods in health classification tasks, while also underscoring the need for richer predictors to achieve stronger performance.