Predicting Major Depressive Episodes and Substance Use Disorder in Youth

Nitharsan Sivakanthan

4/27/2022

Abstract:

Can we use a questionnaire to determine whether a child experiences a major depressive episode or substance use disorder? Knowing this could allow parents and schools to predict whether a child is experiencing these issues and promptly provide support to that child. We use data collected from the 2020 National Survey on Drug Use and Health to classify youth for two factors. Each of these factors tells us whether a particular child has experienced a major depressive episode and/or a substance use. Applying support vector classifier and support vector machine models, we will classify youth into these two factors using various other predictors from the data. The analysis reveals that due to the nature of the problem and the imbalance of youth in the classes of our factors, the accuracy measurements we used to choose the best model are not ideal. Further study is required for better results. However, we achieve promising results when we focus on the precision of predicting classes we deem as important. We can predict whether a youth has experienced a major depressive episode and/or substance use with 60% precision.

Problem Statement:

Using data from SAMHSA's National Survey on Drug Use and Health from 2020, we explore the factors of major depressive disorder and substance use disorders in youth. The survey collects individuals' answers to hundreds of questions. We use the answers to some of these questions to predict two different factors for youth. We predict a binary variable called YMDEORSUD5, which determines whether a youth has experienced a major depressive disorder and/or substance use disorder in the past year. We also predict a multi-class variable called YMDESUD5ONL, which determines whether a youth has experienced: a major depressive disorder alone, a substance use disorder alone, both a major depressive disorder and a substance use disorder, or neither a major depressive disorder nor a substance use disorder.

Background:

Support Vector Classifier (SVC)–

The support vector classifier creates a hyperplane to classify data into categories. This hyperplane is created by maximizing the distance of each data point within a margin of the hyperplane to the plane. To prevent overfitting, the support vector classifier takes on a tuning parameter C, which indirectly determines the number of data points that we allow the model to misclassify. Cross-validation on this parameter C is used to get the best model on the testing data.

Support Vector Machines (SVM)–

When data between classes cannot easily be separated by a hyperplane, support vector machines allow us to use non-linear boundaries to separate data into class categories. SVMs enlarge the feature space which allows us to make linear decision boundaries in higher dimensional space, but non-linear decision boundaries in the original dimensional space.

To compute the support vector classifier of the higher dimensional space, we use the inner product of the observation. We can generalize this inner product to a function called the kernel which computes the relationship between observations. In this paper, we will be using the polynomial and radial kernels.

Each of the two kernels require the same tuning parameter C used in the support vector classifier. In addition to that, the polynomial kernel requires a gamma parameter and a degree parameter, and the radial kernel requires a gamma parameter.

Multi-Class Classification with SVMs–

Because hyperplanes cannot divide more than two categories, we will utilize an approach to multi-class classification called one-versus-all classification. In this approach, we will focus on one class at a time and build a model that best determines that class between all other classes. This process is repeated for each class. In total we will build K total models, where K is the number of classes. An observation's class is determined by its distance to each of the hyperplanes of the K models.

Methodology:

Data-

The data comes from SAMHSA's National Survey on Drug Use and Health from 2020. The survey obtains information on the use of illicit drugs, prescription medication, alcohol, and tobacco. Participants are also asked to answer questions regarding substance use, major depressive episodes, and treatments they receive for these. The survey is created to estimate substance use and mental illness across the United States, track trends over time, and decide on policy for services to the public.

The survey only includes people over the age of 11 and there are some populations that are not included. This information and more on sampling methods can be found on samhsa.gov.

The survey collects answers to hundreds of questions from its participants. We choose 25 of these questions to predict an individual's answers to two questions.

The following table includes the names, descriptions, and class descriptions of the variables chosen to classify in our models:

Classification Variables

Name of Variable	Description	Classes
YMDEORSUD5	An individual has experienced a major depressive episode and/or substance use disorder	0 = no 1 = yes

YMDESUD5ONL	An individual has experienced a major depressive episode, a substance use disorder, both, or neither.	1 = substance use disorder only 2 = major depressive episode only 3 = both 4 = neither

The following table includes a sample of the names and descriptions of predictors used to classify the variables in the previous table:

Predictors

Name of Variable	Description
CADRLAST	The number of alcoholic drinks in the past 30 days.
EDUSKPMON	The number of days school was skipped.
YEPLMTTV	Do parents help with homework at home?
CIGAGLST	How old when last smoked a cigarette?
SDMJURGEED	Has felt a strong urge to use marijuana in the past year

Results:

We will compare SVC and SVM models for both the binary classification problem and the multi-class classification problem.

Binary Classification- YMDEORSUD5 variable

SVC -

Starting with the binary classifier, we will fit an SVC model to predict the classes of YMDEORSUD5.

Tuning for the C parameter on the training data, the model with the highest accuracy has a C = 5. The accuracy of this model is 79.54% on the testing data. Here is a confusion matrix of the results of the model on the testing data:

	truth	
predict	0	1
0	3126	703
1	165	249

The overall accuracy of this model is good. However, it is important to note the accuracy of predicting class 1 is 26.16% (we refer to this as Recall), and of those the model predicts to be in class 1, 60.14% are in class 1 (we refer to this as Precision). Due to the imbalance of youth in each class, we will continue to consider all these accuracy measures in our analysis.

SVM (polynomial kernel) -

Next, we fit an SVM model using a polynomial kernel to predict the classes of YMDEORSUD5.

This time we tune for three parameters on the training data: C, gamma, and degree. The model with the highest accuracy has C = 1, gamma = 2, and degree = 1. The accuracy of this model is 78.03% on the testing data. Here is a confusion matrix of the results of the model on the testing data:

	truth	
predict	0	1
0	3191	832
1	100	120

With this model, the recall is 12.6% and the precision is 54.54% for class 1.

SVM (radial kernel) –

Finally, we fit an SVM model using a radial kernel to predict the classes of YMDEORSUD5.

This time we tune for two parameters on the training data, C and gamma. The model with the highest accuracy has C = 1 and gamma = 2. The accuracy of this model is 78.03% on the testing data. Here is a confusion matrix of the results of the model on the testing data:

	truth	
predict	0	1
0	3191	832
1	100	120

With this model, the recall is 12.6% and the precision is 54.54% for class 1.

Multi-Class Classification- YMDESUD5ONL variable

SVC –

Now, we use the one-versus-all method to perform multi-class classification. We start with fitting an SVC model to predict the classes of YMDESUD5ONL.

Tuning for the C parameter on the training data, the model with the highest accuracy has a C = 1. The accuracy of this model is 77.80% on the testing data. Here is a confusion matrix of the results of the model on the testing data:

	truth			
predict	1	2	3	4
1	24	8	21	9

2	2	5	2	11
3	0	0	4	3
4	151	639	96	3268

Again, like with the binary classifier, we consider not just the overall accuracy of the model in our analysis. In addition, we will compare recall and precision for each class. The following table indicates the recall and precision values of each class.

Class	Recall	Precision
1	13.56%	38.71%
2	0.77%	25.00%
3	3.25%	57.14%
4	99.30%	78.67%

SVM (polynomial kernel) -

Next, we fit an SVM model using a polynomial kernel to predict the classes of YMDESUD5ONL.

We tune for three parameters on the training data: C, gamma, and degree. The model with the highest accuracy has C = .001, gamma = 1, and degree = 1. The accuracy of this model is 77.56% on the testing data. Here is a confusion matrix of the results of the model on the testing data:

	truth			
predict	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	177	652	123	3291

Notably, this model predicts each youth as belonging to class 4. This model has the highest accuracy of any model due to the imbalance of youth in this class. This model is clearly not ideal to use.

SVM (radial kernel) –

Next, we fit an SVM model using a radial kernel to predict the classes of YMDESUD5ONL.

We tune for two parameters on the training data, C and gamma. The model with the highest accuracy has C = .001 and gamma = 1. This is the same model produced using the polynomial kernel.

Again, this model is clearly not ideal to use.

Experimenting with Accuracy Measures –

Instead of using the model with the highest accuracy, we can choose to use models during tuning that have slightly lower accuracy to better our outcomes of predicting classes individually.

Fitting an SVC model with C = 100 on the training data, we can achieve better recall and precision on every class. The accuracy of this model is 77.59%. Here is the confusion matrix of the results of the model on the training data:

	truth			
predict	1	2	3	4
1	35	13	27	26
2	16	69	22	83
3	15	6	16	10
4	111	564	58	3172

Now, our model is predicting more individuals in each of the classes. The following table indicates the recall and precision values of each class.

Class	Recall	Precision
1	19.77%	34.65%
2	10.58%	36.32%
3	13.01%	34.04%
4	96.38%	81.23%

We found other models that had better precision and recall in certain classes, but this model comparably had higher values for each class overall.

Conclusion:

Practically, we would like to obtain high precision from a model so that parents, schools, counselors, etc. can use this model to determine youth that should go through a screening process for major depressive episodes or substance use.

We were able to achieve a promising result on the SVC for the binary classifier with class 1 precision of 60.14%. With the correct resources, this model can be beneficial. For example, a school/school counselors could have students answer the 25 questions. The model would classify them as either in class 0 or 1. If they are classified as class 1, they are 60% likely to have experienced a major depressive episode and/or a substance use disorder in the past year. A counselor could then flag these students for follow-up to work with that student individually and check on them.

In the future, we can improve on SVC and SVM models for both classifications by using other measures to choose the best model, such as the F1-score, which averages precision and recall.

Appendix:

```
#import libraries

library(tidyverse)

library(e1071)


#read in the data

data = read.delim("NSDUH_2020_Tab.txt", header = TRUE, stringsAsFactors = FALSE, quote = "", sep = "\t")


#filter the data and choose variables

data = data %>%

  filter(YEATNDYR == 1)

data = data %>%
select(CADRLAST,VAPNICREC,SDMJURGEED,SDCCURGEED,SDHAURGEED,EDUSKPMON,YESTSCIG,YESTSMJ,YEPLMTTV,YEPCHORE,YEPHLPHW,YEYARGUP,YEYFGTSW,YECOMACT,CIGAGLST,MRJAGLST,COCAGLST,CRKAGLST,HALLAGLST,LSDAGLST,PCPAGLST,ECSTMOAGL,INHLAGLST,YMDEORSUD5,YMDESUD5ONL)


#clean data

data = data %>%

  drop_na(YMDEORSUD5,YMDESUD5ONL)


#create train/test split

set.seed(1)

index = sample(1:nrow(data), nrow(data)*.2)

train = data[index,]

test = data[-index,]


# SVC or SVM model and tuning parameters

tune.out = tune(svm, as.factor(YMDEORSUD5) ~ .-YMDESUD5ONL, data = train, kernal = 'polynomial',

    ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100)
```

```
            ,gamma = c(1,2,3,4,5)
            ,degree = c(1,2,3,4,5)))
summary(tune.out)

# choose model with highest accuracy
bestmod = tune.out$best.model
summary(bestmod)

# Test accuracy of model
ypred = predict(bestmod, test,decision.values = TRUE)
table(predict = ypred, truth = test$YMDEORSUD5)
mean(ypred == test$YMDEORSUD5)
```