**EVALUATION OF ALCOHOL CONSUMPTION RISK**

ITM 6285 Data Mining

Project

by

Trung Phan

# Evaluation of alcohol consumption risk

## 1. Introduction

Alcohol use is related to adverse health and social outcomes. It is the third (after tobacco and poor diet/physical inactivity) leading cause of preventable cause of death in the U.S. and around the world, mainly due to injuries, violence, cardiovascular diseases, and increasing the risk of other health problems. Alcohol consumption, is also associated with massive social cost beyond the damage to the individual users, affecting health care, law enforcement, and legal systems.

Personality measurements are considered risk factors for alcohol consumption, and, in turn, liquor usage impacts individuals' traits. Furthermore, there is increasing interest in developing treatment approaches that match an individual's personality profile. The goal of present research is to study the association between personality factors and demographic differences of alcohol consumers and non-consumers to predict causes of alcohol addiction using Support Vector Machine (SVM).

Individual differences among substance abusers can play an important role in the choice of treatment options. Recently, more attention has been focused on personality trait effects on the efficacy of different treatment plans. This study suggests that all personality traits may be important targets in prevention and treatment heavy alcohol abuse. The findings contribute to an improvement of knowledge concerning the pathways leading to alcohol abuse.

## 2. Data Description

For this project we have selected Drug Consumption Data Set from UCI Machine Learning Repository. Our goal is to look at the relationship between demographic attributes/personality attributes and drug consumption. The database was collected by Elaine Fehrman between March 2011 and March 2012. An online survey tool from Survey Gizmo was employed to gather data.

The data set contains information on the consumption of 18 central nervous system psychoactive drugs.  An exhaustive search was performed to select the most effective subset of input features and data mining methods to classify the drug usage. In this project we have picked one drug(alcohol) to analyze the drug usage.

Database contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity.

All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real-valued Database contains 18 classification problems. Each of independent label variables contains seven classes: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day". For this analysis we are using seven class classifications for one drug(alcohol) using SVM classification.

Summary statistics*:*

Table 1 consists of five features of demographic attribute which belongs to 1885 surveyed participants. These categorical variables are converted and standardized into numerical format for analysis purpose. Table 2 displays personality traits under seven variables are presented in numerical data. As we can see both two tables are presenting Mean, Minimum, Maximum, Standard deviation and counts (frequency). Details of converting scale will be fully explained further in our project.

## 3. Research Methodology

Classification technique:

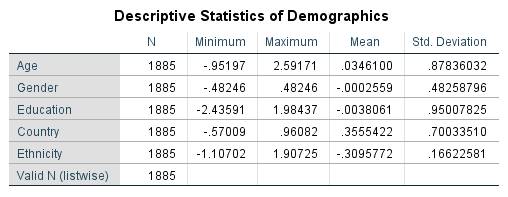
What we are looking for is any potential cause of alcoholic addiction from which we will establish precaution and awareness of alcoholic overuse ahead of time. We classify the intensity of Alcohol consumption under seven categories based on Demographic/Personality features in order to observe the strength of their relationships.

We choose SVM (Support Vector Machine) as a Machine Learning Classifier for this dataset. The reason we implement SVM because it works effectively with medium and large sample size. More than that, it also returns a very accurate result. This algorithm is run under R environment.

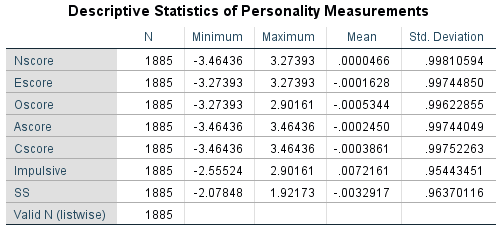
The implementation of SVM algorithm:

* First, we classify the data (Alcohol usage based on demographic and personality) using default setting parameters which has cost parameter of 1 (penalty to the misclassified points)
* Then, we check the model performance and calculate the accuracy of model.
* After that, we try to tune the model by changing the parameter values (using a ten-fold cross validation). The purpose of this is to search for the best model which has largest accuracy.
* Finally, we perform visualizations for the model using plot() function.

Descriptive Statistics:



**Table 1**



**Table 2**

**4. Empirical Analysis**

**4.1 Running model without modifying**

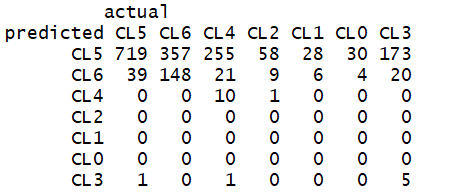
We are running model under svm method with the prediction and examine its performance. The accuracy of this prediction is reported at 46.79%

svm.model <- svm(Alcohol~.,data=drug)

pred<-predict(svm.model,drug)

performance<-table(predicted=pred,actual=drug$Alcohol)

performance



accuracy=sum(diag(performance)/sum(performance))

accuracy

[1] 0.4679045

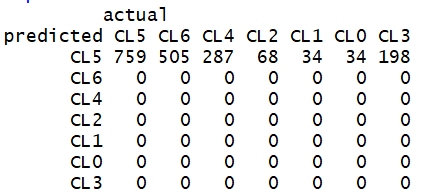
Since this figure is not a desirable number compare with what we are looking for, we will play around with changing parameters to find a better result. We next will change the kernel from radial (default) to linear.

svm.model1<-svm(Alcohol~., data=drug, kernel="linear")

pred<-predict(svm.model1,drug)

performance<-table(predicted=pred, actual=drug$Alcohol)

performance



accuracy=sum(diag(performance)/sum(performance))

accuracy

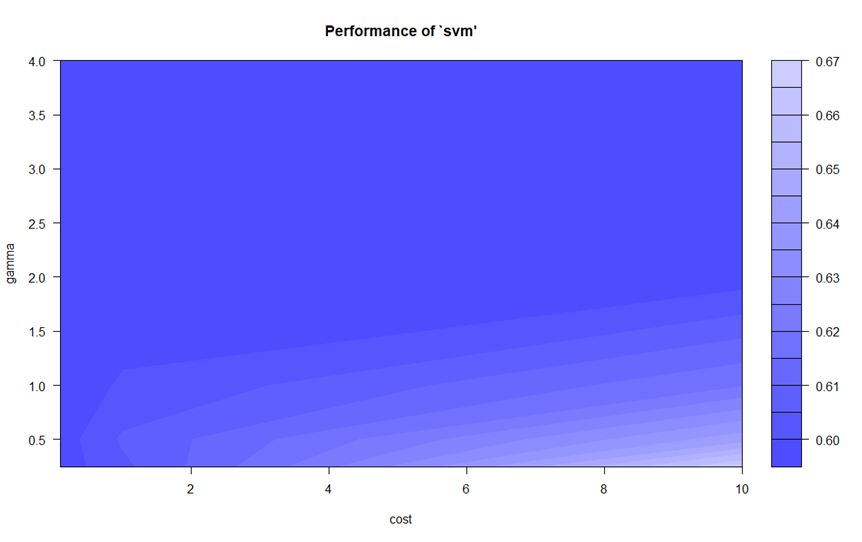
[1] 0.4026525

The result turns out even worse with a lower accuracy. This is what we are not looking for. So, we are now trying to tune the model by setting up the cost and gamma.

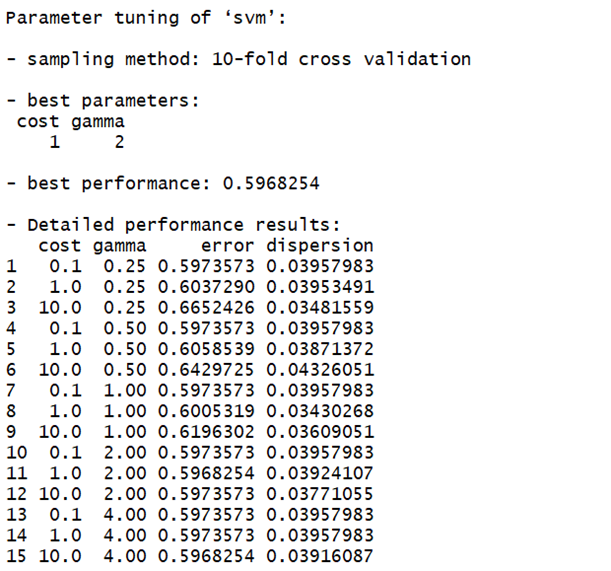
set.seed(123)

tunedmodel<- tune(svm,Alcohol~.,data=drug, ranges=list(cost=c(0.1,1,10),gamma=2^(-2:2)))

plot(tunedmodel)



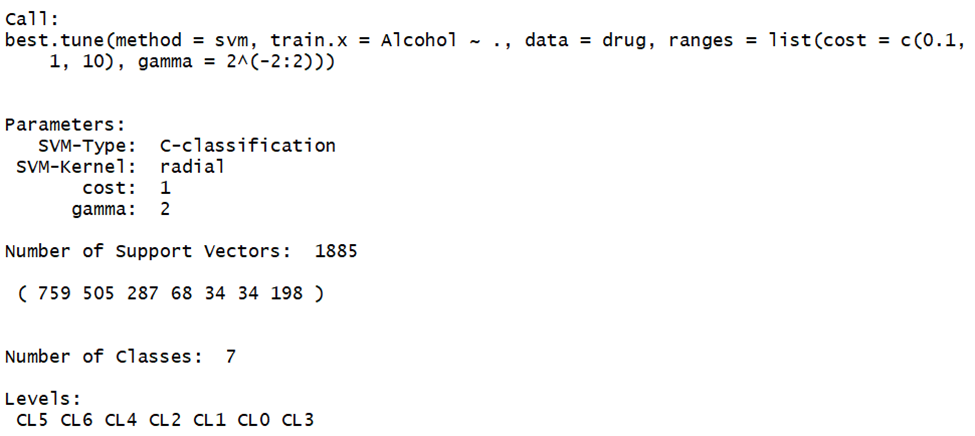
summary(tunedmodel)



It turns out the best performance is 0.5968 with the cost of 1 and gamma at 2

bestmodel<-tunedmodel$best.model

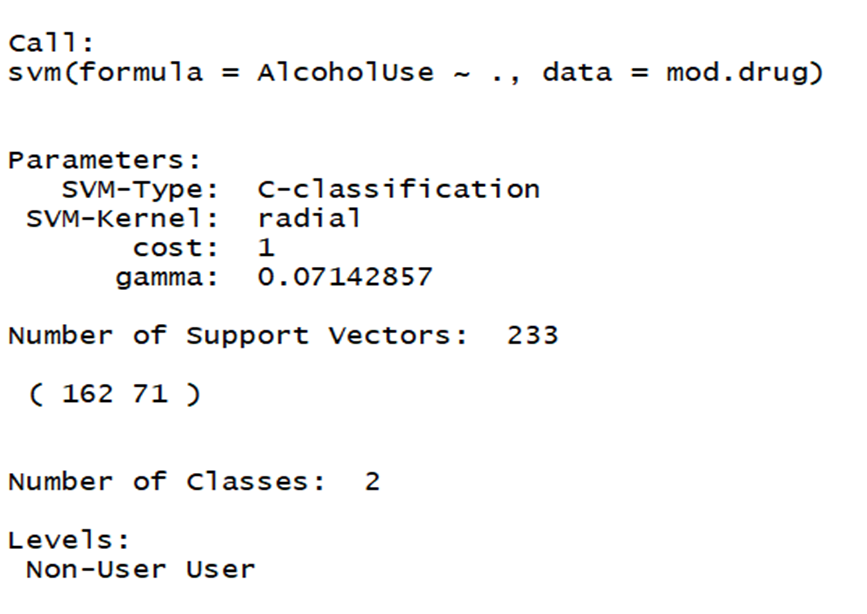
summary(bestmodel)



**4.2 Running modified model**

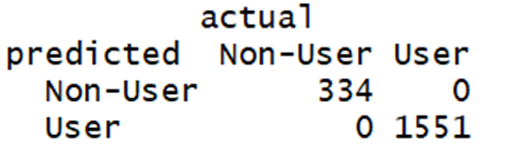
mod.svm.model <- svm(AlcoholUse~.,data=mod.drug)

summary(mod.svm.model)



mod.accuracy=sum(diag(mod.performance)/sum(mod.performance))

mod.accuracy

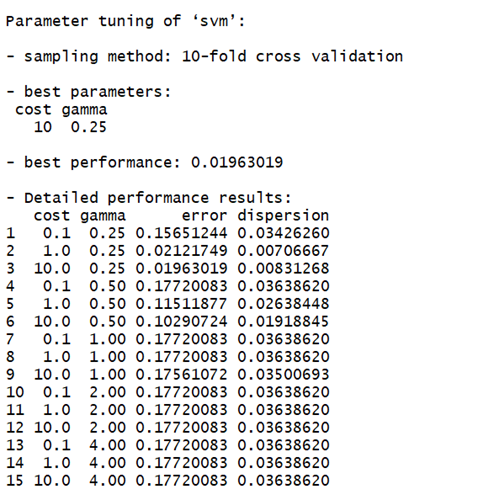


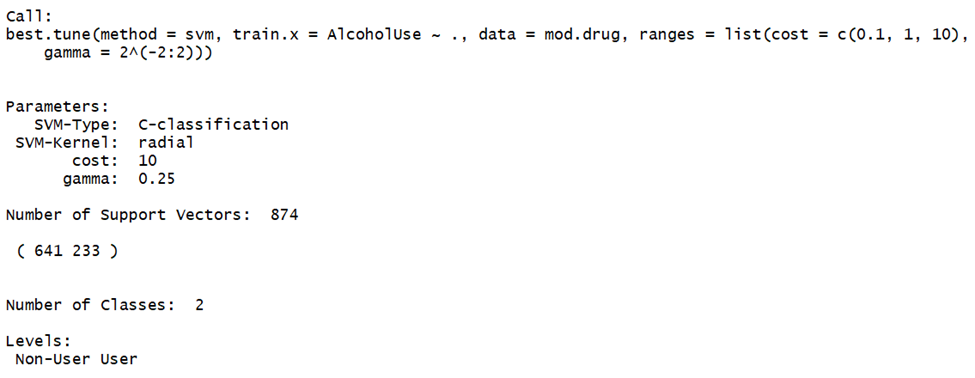
set.seed(123)

mod.tunedmodel<- tune(svm,AlcoholUse~.,data=mod.drug, ranges=list(cost=c(0.1,1,10),gamma=2^(-2:2)))

mod.bestmodel<-mod.tunedmodel$best.model

summary(mod.bestmodel)





The model after modifying performs the best at cost 10 and gamma 0.25. The accuracy of model is now 100% correct

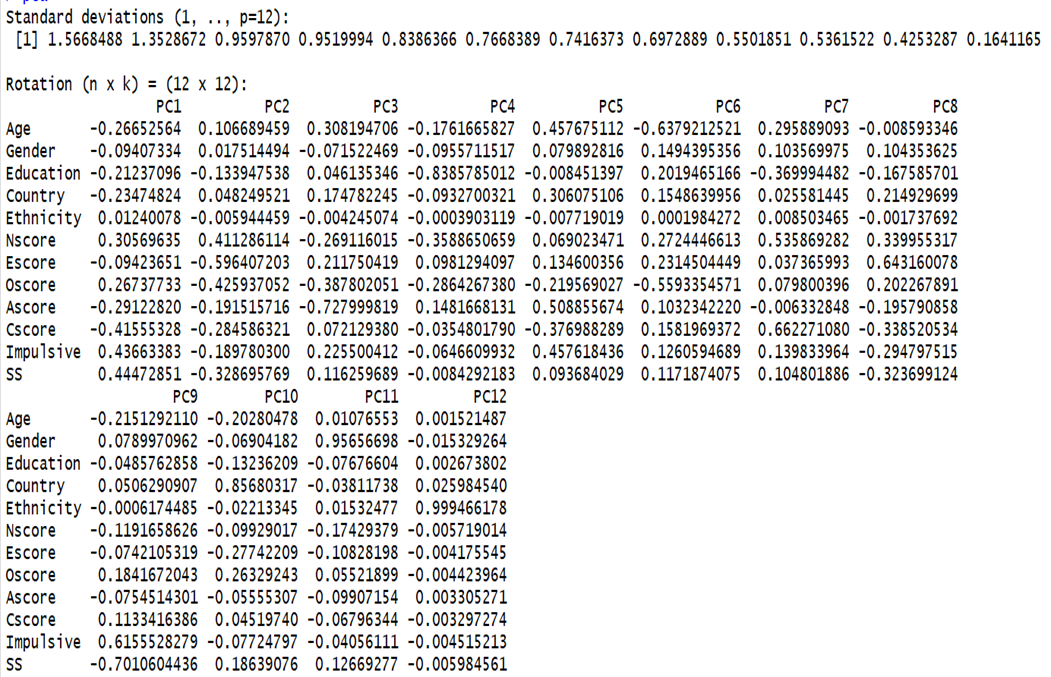
data=mod.drug[c(-1,-13,-14)]

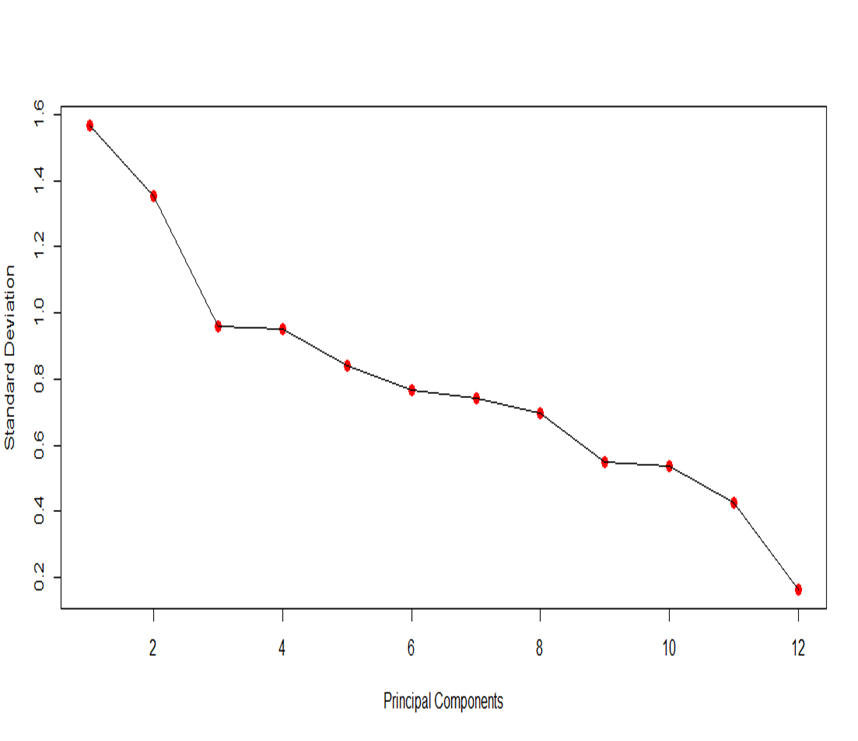
View(data)

pca<- prcomp(data,center = TRUE)

pca

plot(pca)





5. **Conclusion**

**5.1 Main cause of alcohol consumption:**

* Personality attributes lead people to alcohol abuse more than Demographics
* Big Five personality traits are cause for alcohol addiction

**5.2 Solutions:**

* Raise awareness of alcohol abuse, taking into account personality traits to prevent it in the future.
* Rehab center that helps alcoholics to overcome their alcohol dependence.
* Different treatments: alcoholism medication, group, one-to-one therapy, and counseling.

**Appendix:**

**R codes:**

# load packages

library(e1071)

# code without modifying dataset

summary(drug)

svm.model <- svm(Alcohol~.,data=drug)

summary(svm.model)

pred<-predict(svm.model,drug)

performance<-table(predicted=pred,actual=drug$Alcohol)

performance

accuracy=sum(diag(performance)/sum(performance))

accuracy

svm.model1<-svm(Alcohol~., data=drug, kernel="linear")

pred<-predict(svm.model1,drug)

performance<-table(predicted=pred, actual=drug$Alcohol)

performance

accuracy=sum(diag(performance)/sum(performance))

accuracy

set.seed(123)

tunedmodel<- tune(svm,Alcohol~.,data=drug, ranges=list(cost=c(0.1,1,10),gamma=2^(-2:2)))

plot(tunedmodel)

summary(tunedmodel)

bestmodel<-tunedmodel$best.model

summary(bestmodel)

**################################################################################**

#Remodify data:

#Divide class into two group. Cutoff is CL4

#First group is Alcoholic Nonuser which includes first four groups: CL0, CL1, CL2, CL3

#Second group is Alcoholic User includes last three groups: CL4, CL5, CL6

# code for modifying dataset

mod.drug <- drug

mod.drug$AlcoholUse=ifelse((drug$Alcohol=="CL0") | (drug$Alcohol == "CL1") | (drug$Alcohol=="CL2") | (drug$Alcohol=="CL3"),"Non-User","User")

mod.drug <- mod.drug[,-14]

mod.drug$AlcoholUse <- as.factor(mod.drug$AlcoholUse)

mod.drug$Usage=ifelse(mod.drug$AlcoholUse=="Non-User",0,1)

View(mod.drug)

mod.svm.model <- svm(AlcoholUse~.,data=mod.drug)

summary(mod.svm.model)

mod.pred<-predict(mod.svm.model,mod.drug)

mod.performance<-table(predicted=mod.pred,actual=mod.drug$AlcoholUse)

mod.performance

mod.accuracy=sum(diag(mod.performance)/sum(mod.performance))

mod.accuracy

set.seed(123)

mod.tunedmodel<- tune(svm,AlcoholUse~.,data=mod.drug, ranges=list(cost=c(0.1,1,10),gamma=2^(-2:2)))

plot(mod.tunedmodel)

summary(mod.tunedmodel)

mod.bestmodel<-mod.tunedmodel$best.model

summary(mod.bestmodel)

data=mod.drug[c(-1,-13,-14)]

View(data)

pca<- prcomp(data,center = TRUE)

pca

plot(pca)