2. Literature survey

We were able to find a research paper “A Machine Learning Approach to College Drinking Prediction and Risk Factor Identification” in the Google Scholar that is related to our final project about machine learning of student alcohol consumption. According to the authors, parametric inference models like logistic regression and linear regression were commonly used for validating hypothesized risk factors at the time. If datasets have complex structure, then elaborated methods such as negative binomial regression and multinomial logistic regression would be used. However, there were two drawbacks of these methods: first, they were not able to determine the correlation between the identified predictors; second, these methods were not generalized enough for longitudinal datasets.

In this paper, the authors were aiming to design a new machine learning algorithms with improved generalizability for longitudinal datasets in order to help secondary analysis of alcohol misuse and to provide insights for alcohol intervention programs. In the paper, a longitudinal college drinking dataset that has total 15120 number of records and 10 features (Table 1) was used, and authors proposed two methods which were temporally-correlated support vector machine and a combination of cluster analysis to accomplish their goals(Bi, Sun, Wu, Tennen, & Armeli, 2013).

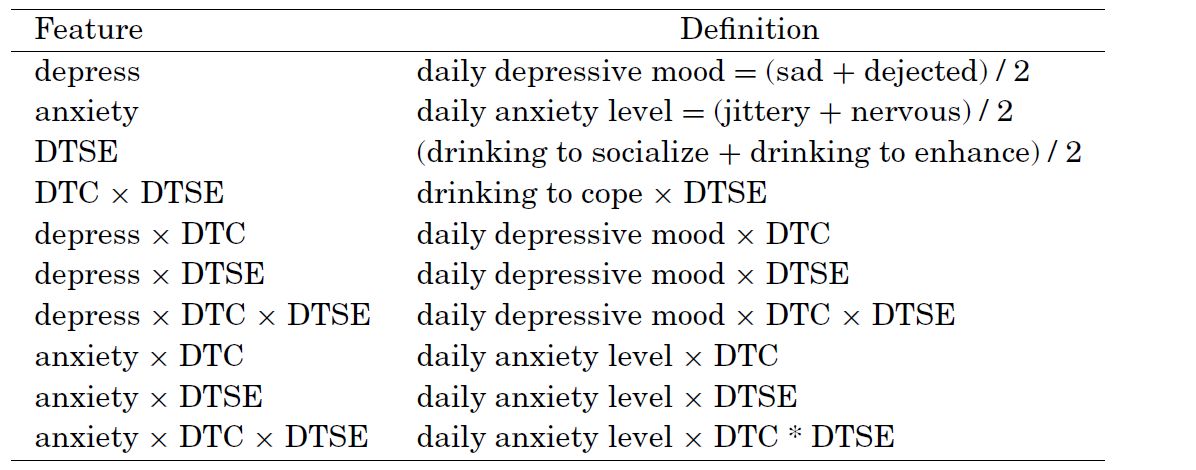
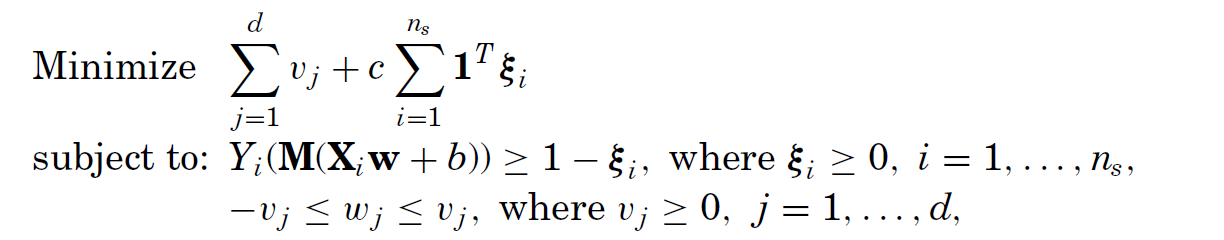


Table 1. Variables of the dataset

In order to construct a classifier to identify risk factors for a longitudinal dataset, the authors used the variable selection method to measure the goodness of each subset variables in order to identify optimal variable subsets. In detail, they used a wrapper method to wrap a temporally-correlated support vector machine that was derived based on the 1-norm SVM to deal with correlated data. In order to find out the best within-day effect model f(x)=wTx+b

for each training set, authors derived an optimization algorithm(Table 2) to solve the optimization problem

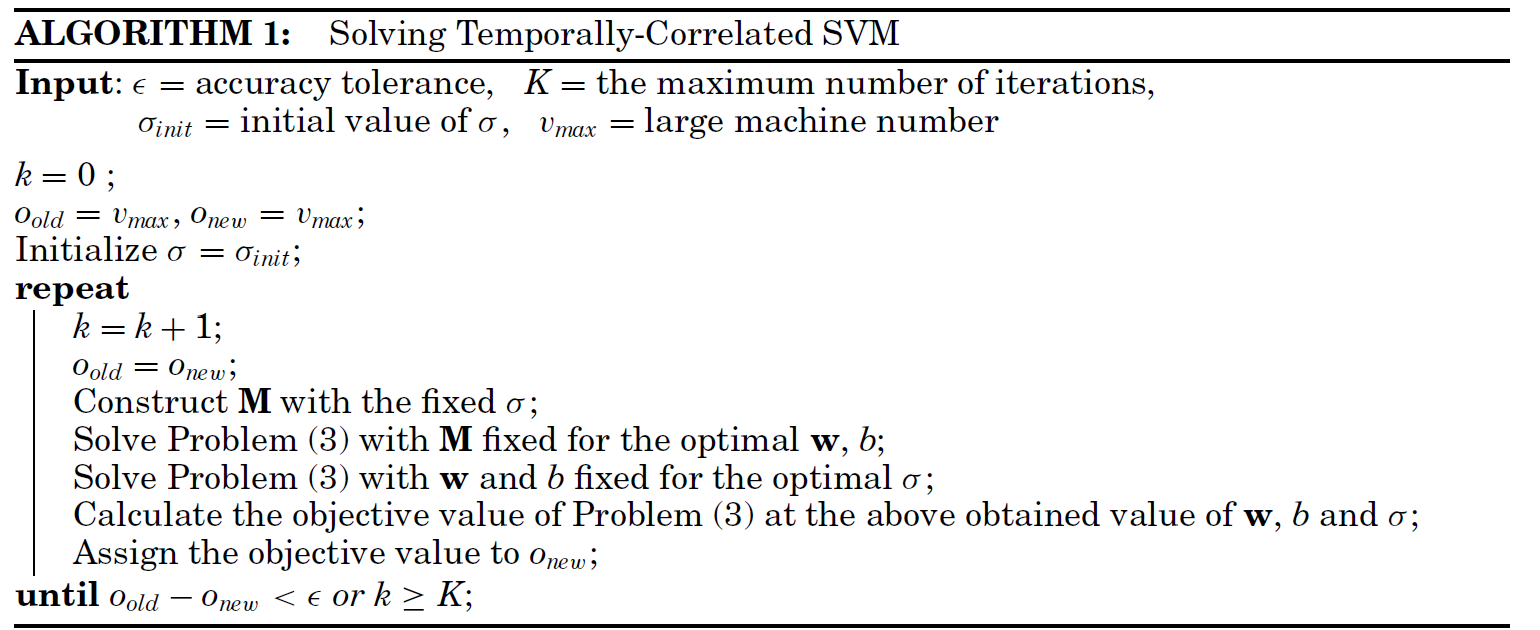


Table 2. Optimization algorithm 1

Authors used ROC curve to compare temporally-correlated SVM, 1-norm SVM, logistic regression on a 30 days daily dairies dataset, and the result showed that temporally-correlated SVM has the highest true positive rates. Besides that, AUC comparison also demonstrated that temporally-correlated SVM has the best performance among the three methods. By proposed algorithm, authors got top 10 daily factors averaged weight values for male(Table 3) and female (Table 4).

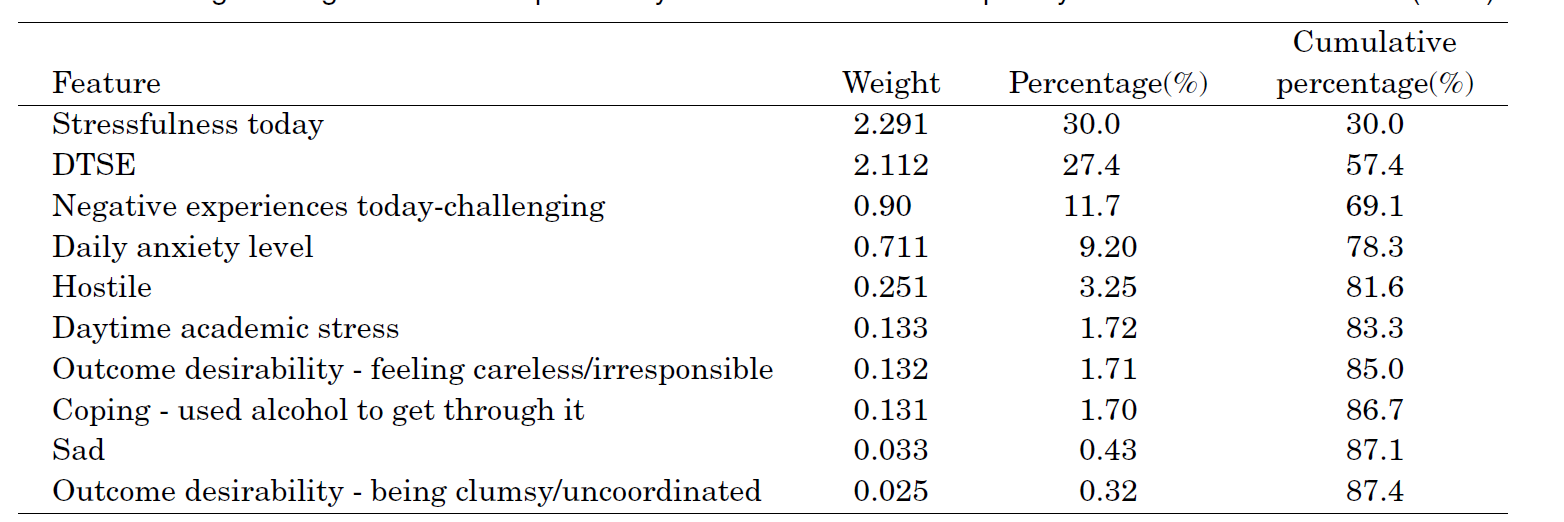


Table 3. Daily factors averaged weight values for male

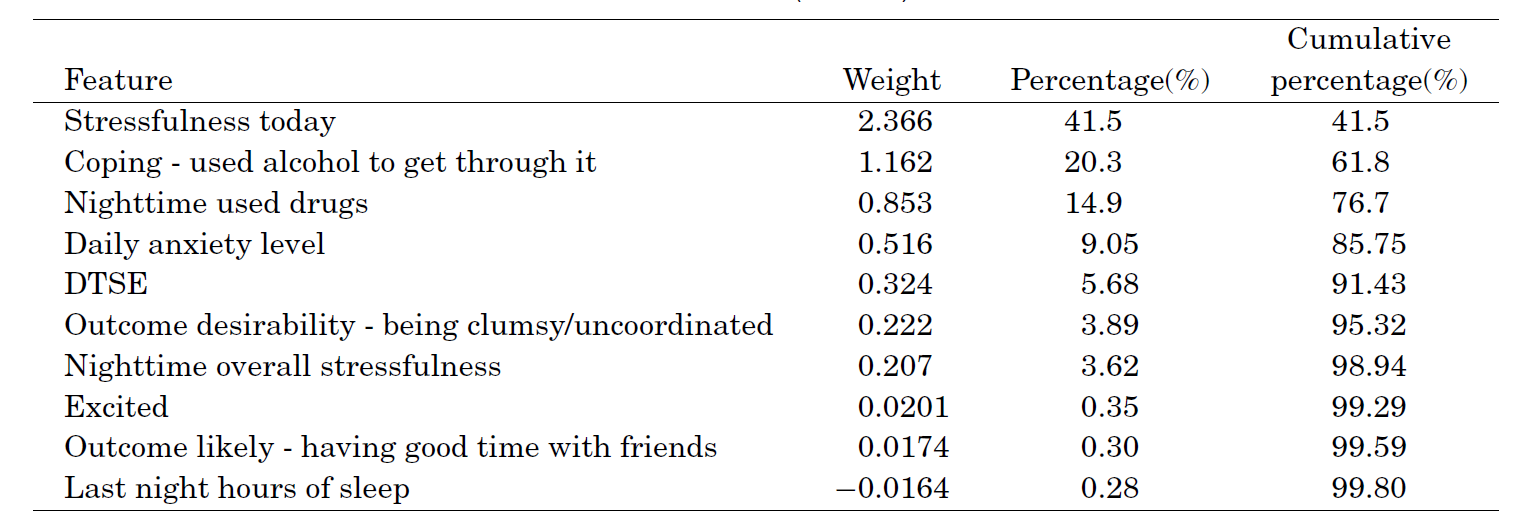


Table 4. Daily factors averaged weight values for female

Since the authors tried to construct a model to identify person-level risk factors for heavy drinking, they applied multiple correspondence analysis (MCA) to compacts categorical variables into lower-dimensional space for continuous variables. To deal with different scales that were used in the dataset, standard normalization was used for each risk factor. K-Medoids clustering algorithm was used to group 4 objects(DrinkFreq, DrinkAmnt, DrunkFreq, AlcDep) into 3 clusters. Then 3 classifiers were built to separate each drinker group, and all the labeled data was divided into male and female then split into 2/3 training set and 1/3 test set. Authors used ROC Curve to tested accuracy for each classifier and the results were shown that for female, non-risk drinking classifier had the highest accuracy and heavy drinking classifier had high accuracy. For male, both non-risk drinking classifier and heavy drinking classifier had moderate accuracy. But moderate drinker classifier performs poorly in both male and female dataset.

References

Bi, J., Sun, J., Wu, Y., Tennen, H., & Armeli, S. (2013). A machine learning approach to college

drinking prediction and risk factor identification. *ACM Transactions on Intelligent*

*Systems and Technology,4*(4), 1-24. doi:10.1145/2508037.2508053