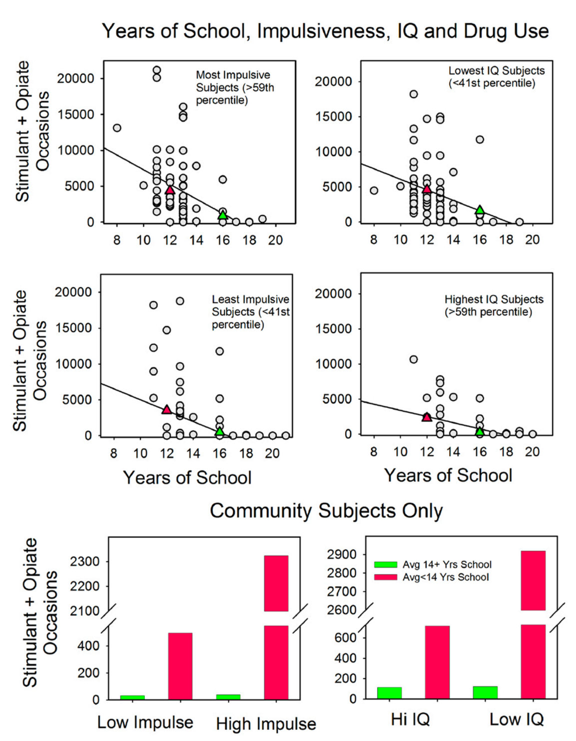
# Patterns and Correlates of Substance Abuse and Socio-Health Demographic Factors

Kun Woo Cho[1], Tyler Park[1], Minsung Kim[1], Sai Srikar Kasi[1]

## INTRODUCTION

- Substance abuse and mental health issues affect a large proportion of adolescents and adults in the United States every year and contributes heavily to the burden of disease.
- The National Survey on Drug Use and Health (NSDUH[1]) provides information on illicit drug use, and mental health issues for the civilian and non- institutionalized American population.
- The dataset consists of ~60k users with ~3k features each.



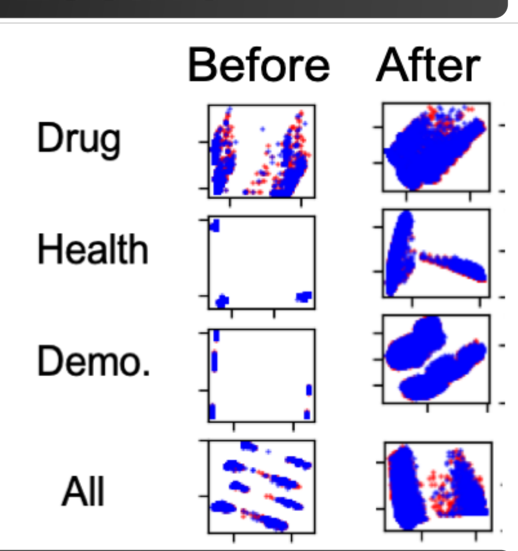

- Our project attempts to explore the underlying causes of substance abuse and its effect on mental health by finding correlations and patterns in the NSDUH dataset to help reduce medical disorders.

## PROJECT GOALS

- Finding and Analyzing : hidden patterns in the NSDUH dataset via several Unsupervised Machine Learning techniques, and interpreting a relatively better model observed among them.
- Predicting the cause : the substance usage patterns based on the individuals' demographics.
- Predicting the effect : the mental health disorder patterns of individuals' , raised due to such substance abuse.
- Investigating the relation : between the predictions made and the patterns observed in the original dataset by feature comparison.

## DATA PREPROCESSING

- Dataset are partitioned into health, substance, demography.
- Each subset is spitted into train/test ratio of 4:1
- Since most of missing data are NMAR, we use one-hot encoding rather than imputation.
- Features with over 80% of NaN are removed.



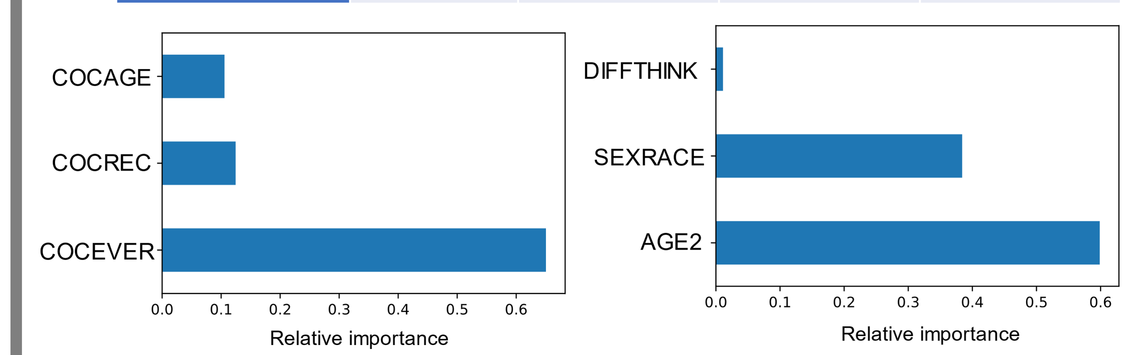

## ABBREVATIONS

- MSE – Mean Squared Error
- MAE – Mean Absolute Error
- $R^2$ – Coefficient of determination
- CV – Cross Validation
- RE – Reconstruction Error
- ALL – Average Log-Likelihood]
- SIL – Silhouette
- SCORE - the opposite of the value on the K-means objective.
- CIG – cigarette
- CRK – crack
- MJ – marijuana
- STIM – stimulant
- PNR – pain reliever,
- ALC – alcohol
- TOB – tobacco
- DRG – drug
- HALLUC – hallucination
- INHAL – inhalant
- OXC - oxycontin

## SUPERVISED LEARNING

Question 1: Which features can predict use of severe, rare substances? Hypothesis A: Use of gateway substances. Hypothesis B: Socio-economic status.

| Regressor | Hypothesis A | | Hypothesis B | |
|---|---|---|---|---|
| | MSE | $R^2$ | MSE | $R^2$ |
| LASSO | 0.030 | 0.056 | 0.029 | 0.056 |
| RIDGE | 0.023 | 0.149 | 0.028 | 0.073 |
| ELAS_NET | 0.029 | 0.056 | 0.029 | 0.056 |
| PLS | 0.025 | 0.112 | 0.028 | 0.066 |
| RND_FRST | 0.024 | 0.118 | 0.029 | 0.063 |




Question 2: Given an individual's substance usage data, how well can we predict the future potential mental disorders ?

- Models below are evaluated at their best hyperparameters which were exhaustively tuned through GridSearchCV.

| Regressor | Train_Test_Split | | 5-FOLD CV | | 10-FOLD CV | |
|---|---|---|---|---|---|---|
| | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ |
| **RIDGE** | **0.614** | **0.158** | **0.626** | **0.156** | **0.628** | **0.156** |
| LINEAR | 0.615 | 0.147 | 0.622 | 0.137 | 0.613 | 0.146 |
| ELAS_NET | 0.873 | 0.134 | 0.879 | 0.067 | 0.877 | 0.066 |
| LASSO | 1.015 | 0.004 | 1.018 | 0.0019 | 1.023 | 0.0037 |

AFTER BOOTSTRAPPING

| | $R^2$ (mean) | $R^2$ (std) |
|---|---|---|
| RIDGE | 0.0141 | 0.2635 |
| LINEAR | Poor | Poor |
| **ELAS_NET** | **0.0665** | **0.001360** |
| LASSO | 0.0388 | 0.000290 |

## UNSUPERVISED LEARNING

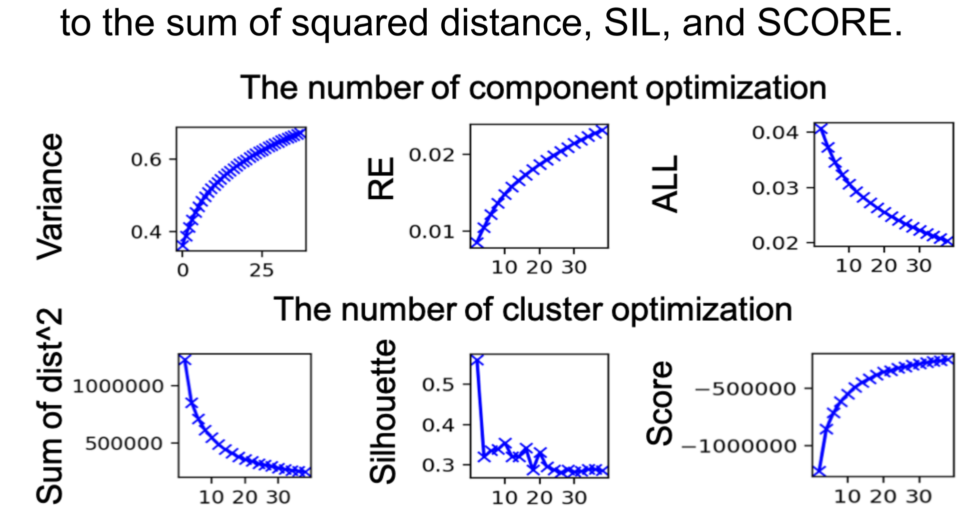
### Model performance comparison

- Some models are unable to generate RE and/or ALL, and thus written as NA.
- Overall, PCA has the smallest train time, RE and the second largest ALL value for train data.

| Model | Train Data | | | Test Data | |
|---|---|---|---|---|---|
| | Time | RE | ALL | RE | ALL |
| **PCA** | **4.36s** | **0.0323** | **0.0136** | **0.0323** | **0.0546** |
| LDA | 510.2s | NA | -2246 | NA | -2253 |
| FA | 52.62s | NA | 0.0076 | NA | -12071 |
| GMM | 2297s | NA | 0.2157 | NA | -0.3176 |
| NMF | 66.80s | 0.0346 | NA | 0.0346 | NA |

### Hyperparameter optimization for PCA and K-means

- The number of components for PCA is set to 8 based on the cumulative explained variance ratio, ALL, and RE.
- The number of clusters for K-means is set to 10 according to the sum of squared distance, SIL, and SCORE.



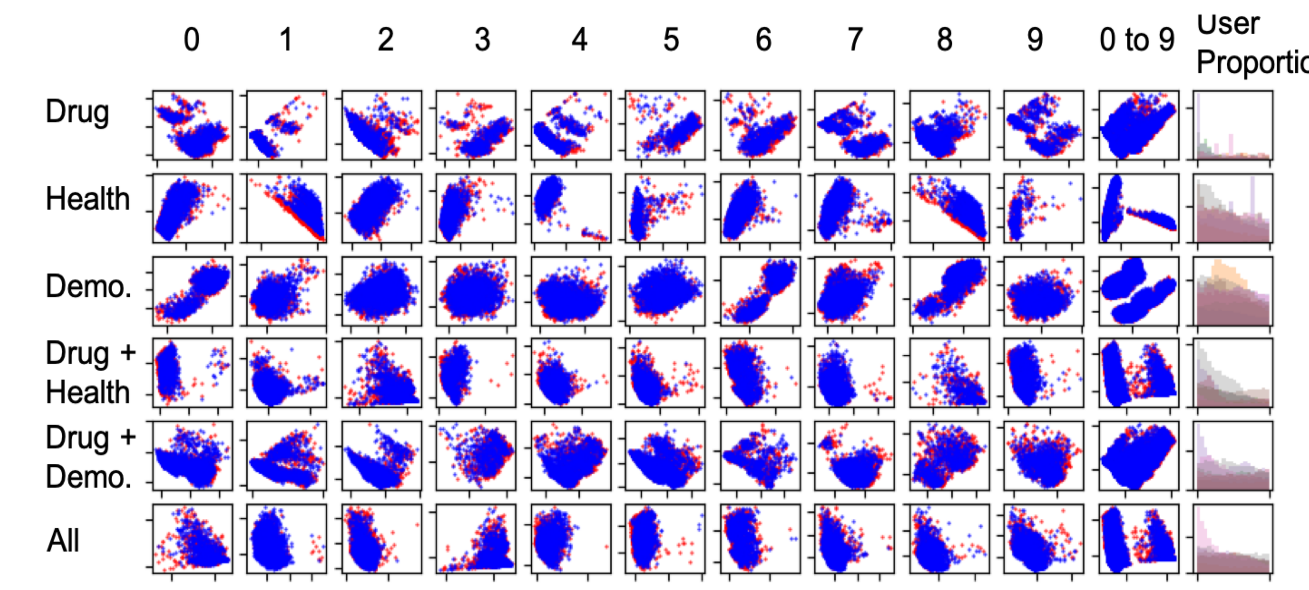

### Pattern analysis using PCA components

- We extract top 5 features with the highest pseudocount for each component and list the commonalities among 5 features
- For health and demo., the first half row shows its own top features while the second half row is the top features of drug.

| Data\Comp # | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Drug | Use CIG/MJ | Not use CG | N/A | N/A | Sometimes use STIM | Sometimes use PNR/ALC | Sometimes use TRQ, Not use PNR | Use PNR, Not use CIG |
| Health (Health & drug) | No youth mental service utilization | Has mental illness | No physical illness | No nervousness | Feel nervous | Has mental health treatment | No difficulty in daily routine | No difficulty in daily routine |
| | Use PNR<br>Refuse to answer METHA question | Not use INHAL | Use CIG | Not use ALC/TOB/DRG | Use ALC/TOB/DRG | Not use CIG | Not inhale marker<br>Sometimes inhale lighter gases | Sometimes inhale lighter gases |
| Demo (Demo & drug) | Less than $10k income | 26+ age | No health insurance<br>Part time job | No health insurance | 45+ pregnancy age<br>No test for ALC/DRG | Female, white | Male, white | Female<br>No children |
| | N/A | N/A | Not use CRK | Not use CRK | Not use MJ | N/A | N/A | N/A |
| ALL | 12-17 age | Use MJ/HALLUC<br>Not use PEYOTE | 18-25 age | Unemployed | Has mental illness | Not use CIG | Not use ETHER<br>Use INHAL | Use MJ/ALC/DRG<br>Unemployed |

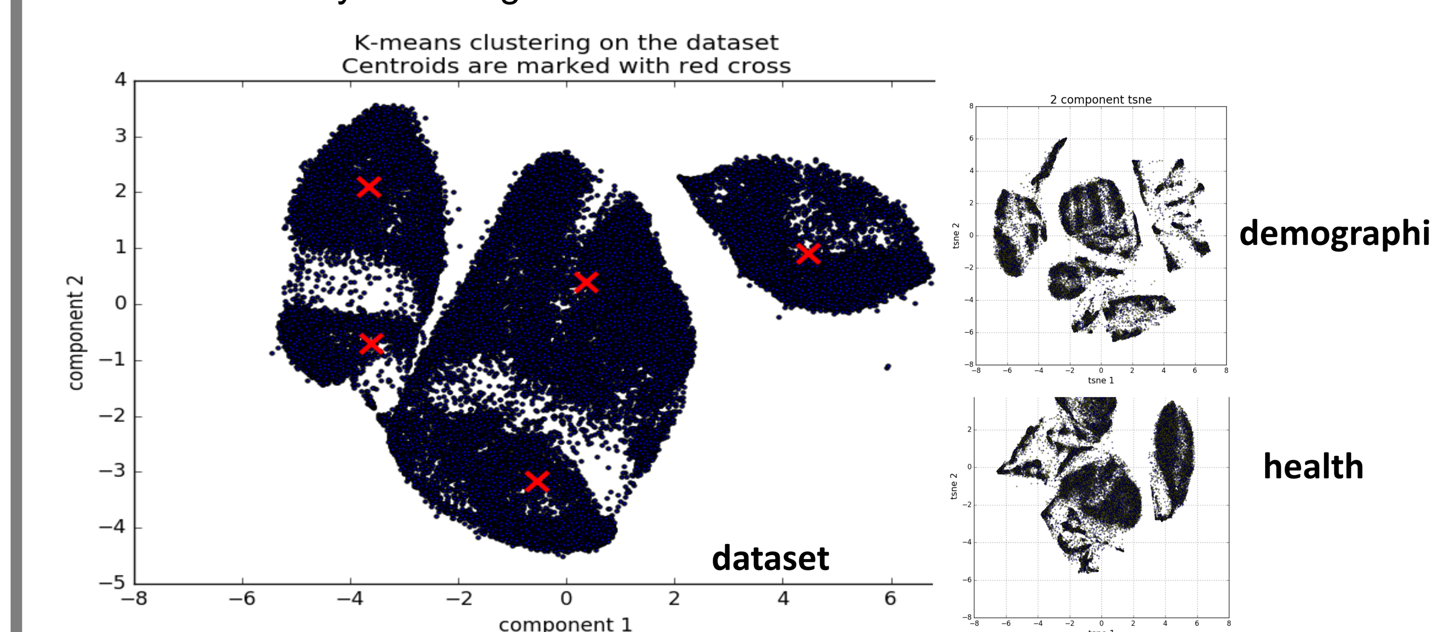
### Pattern analysis using K-means clusters

- Pattern in each data for cluster # 0 to 9 (red-male, blue-female)
- Clusters with drug data tend to be more separated because the data is largely divided into those who answered and those refused to answer the questions.




### T-SNE analysis

- T-SNE (component 2) scatter as visualization of dataset: 5 separate clusters and centroids for dataset and more clusters for subset.
- Most important features of each centroid among 50 nearest neighbors are : Experience of Heroin, sexual disease, pipe tobacco, difficulty of hearing, and difficulty of seeing.




## CONCLUSION

- We identified latent structure and correlation in the dataset using supervised and unsupervised learning.
- Supervised: low predictive power of the supervised learning
- Unsupervised: correlation of the substance use is greater towards health relative to demographic factors.

**BIBLIOGRAPHY**

[1] NSDUH. Website. https://nsduhweb.rti.org/respweb/homepage.cfm.

[2] Ryan, Heather, Angela Trosclair, and Joe Gfroerer. "Adult current smoking: differences in definitions and prevalence estimates—NHIS and NSDUH, 2008." *Journal of environmental and public health* 2012 (2012).

[3] Fix, Brian V., et al. "Patterns and correlates of polytobacco use in the United States over a decade: NSDUH 2002–2011." *Addictive behaviors* 39.4 (2014): 768-781.

*Computer Science Department, Princeton University*

PRINCETON UNIVERSITY