Young People Survey

Explore the preferences, interests, habits, opinions, and fears of young people

計財所 碩二 108071601 賴冠維

Data Introduction

In 2013, students of the Statistics class at FSEV UK were asked to invite their friends to participate in this survey.

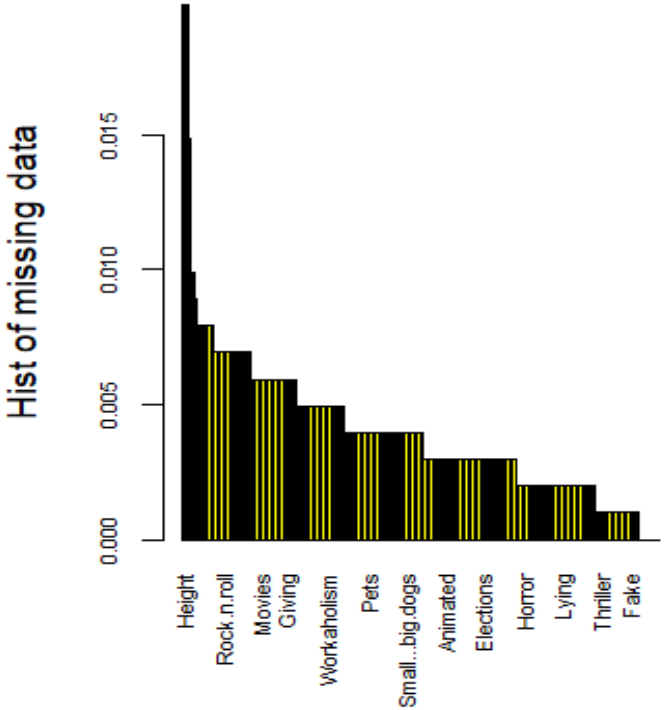
- Music preferences (19 items)
- Movie preferences (12 items)
- Hobbies & interests (32 items)
- Phobias (10 items)
- Health habits (3 items)
- Personality traits (57 items)
- Spending habits (7 items)
- Demographics (10 items)

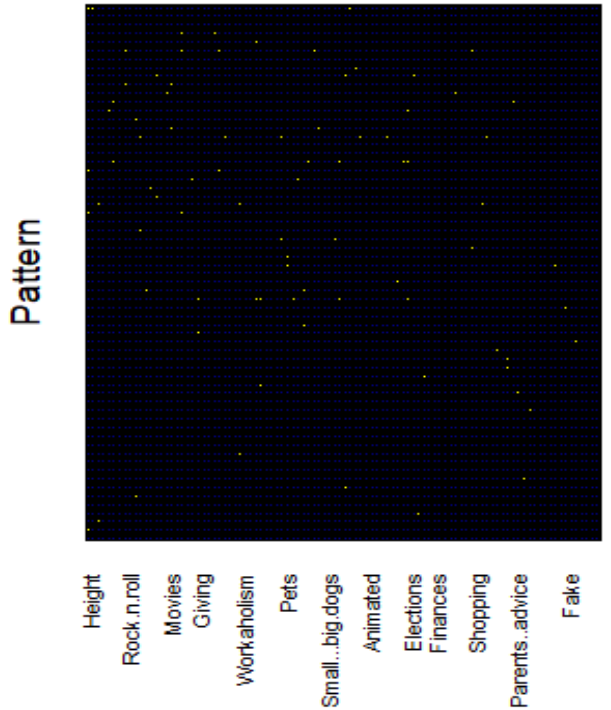
Outline

1. **Clustering** — Describe the composition of the participants
2. **Classification** — Use Xgboost to classifiy someone whether have alcohol addiction with other features.
3. **Relationship** — Use factor analysis to find out the relationship among the features that be selected in former.
4. **GMM** — Weighted the answer and classification by GMM

Missing Value Imputation

Use CART (Classification And Regression Tree) to imputate the missing value





Hierarchical Clustering

Describe the composition of the participants

- Ward's minimum variance method：

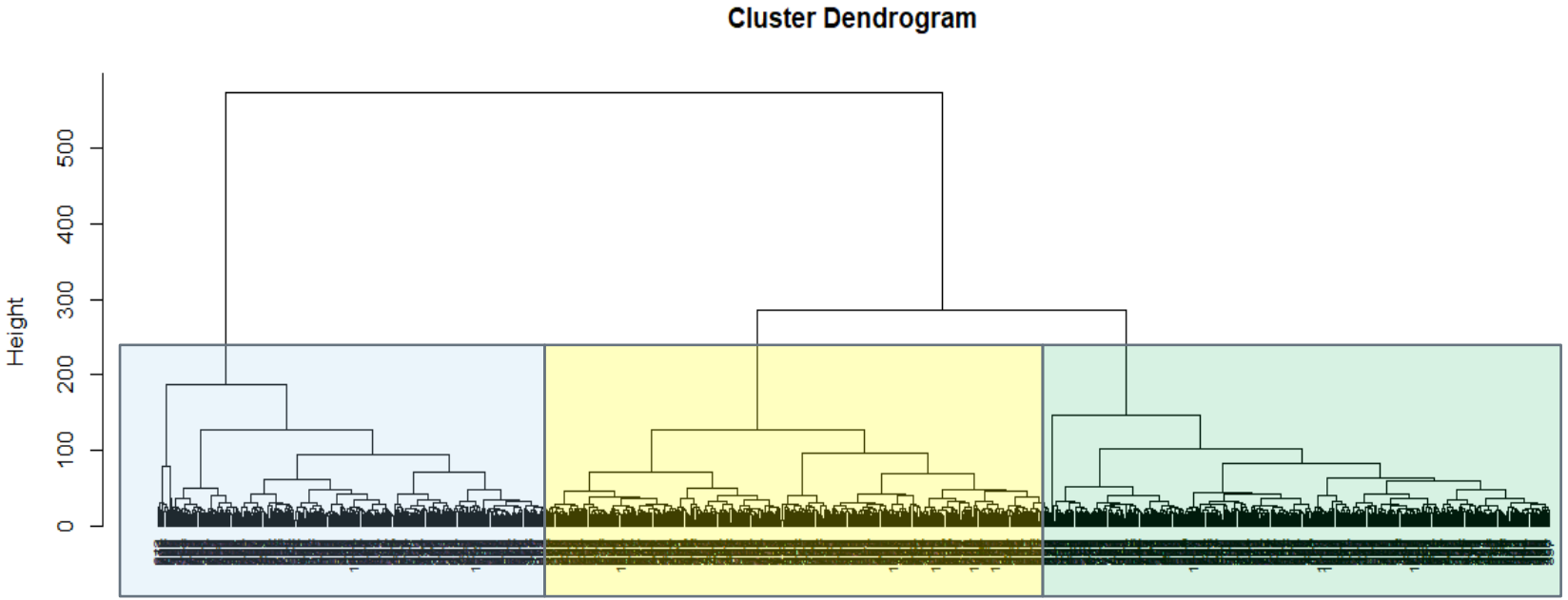
$$Total\ ESS = ESS_1 + ESS_2 + \cdots + ESS_k$$

$$ESS_k = \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^T \left(x_{ij} - \bar{x}_i\right)$$

- x_{ij} : $j^{th}\ number\ of component\ in\ i^{th}\ cluster$
- $\bar{x}_i$: $Mean\ of\ the\ i^{th}\ cluster$

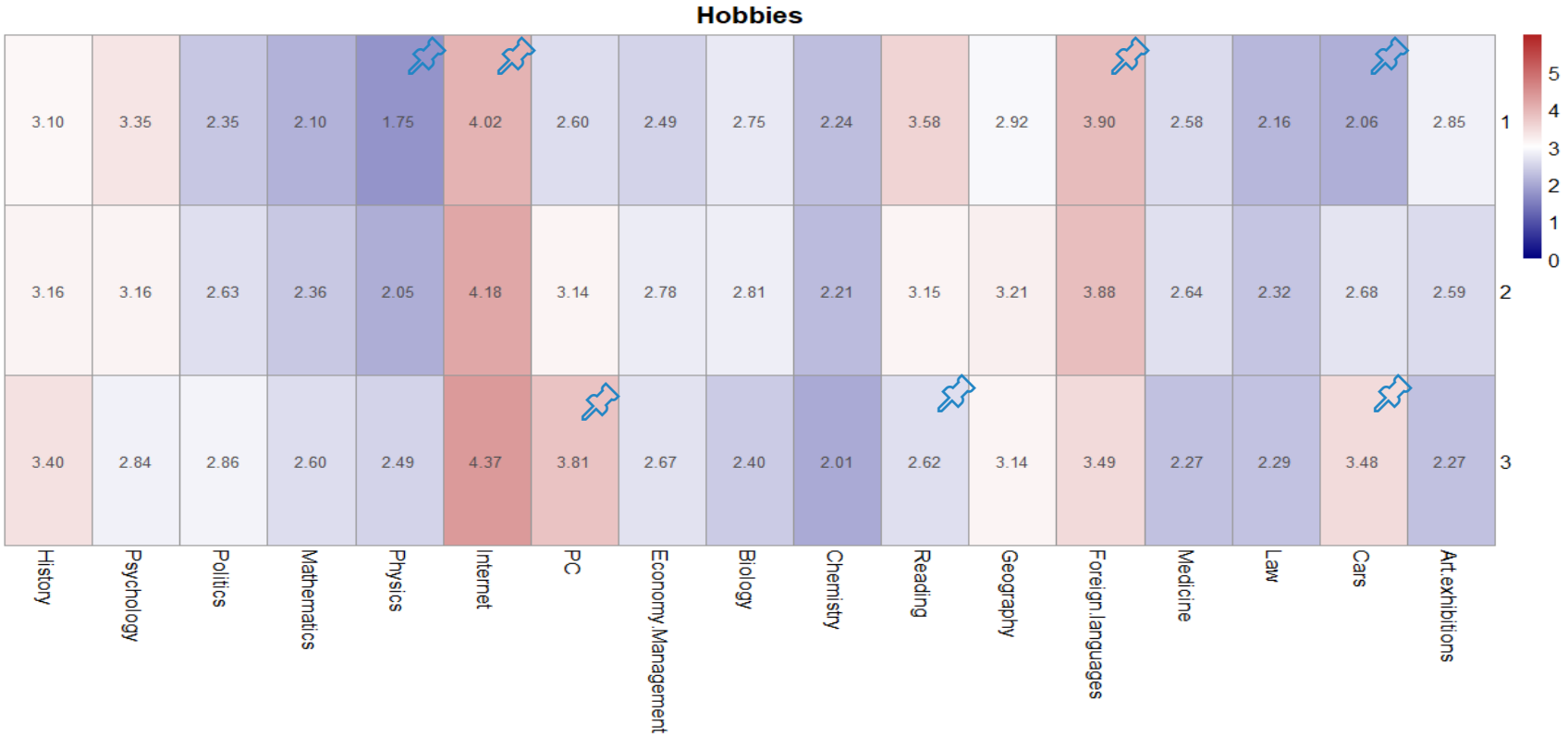
Clustering

Describe the composition of the participants



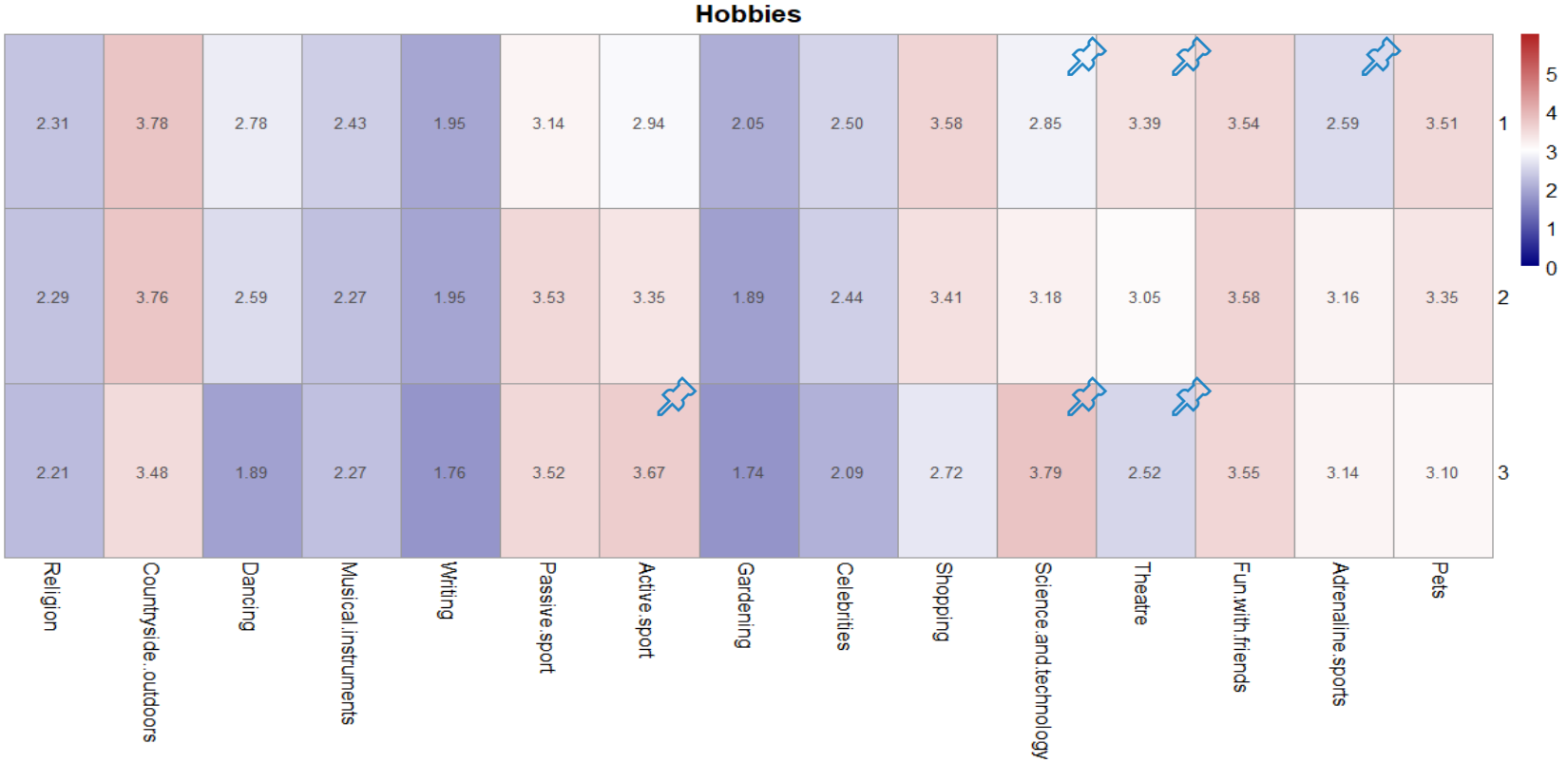
Clustering

Describe the composition of the participants



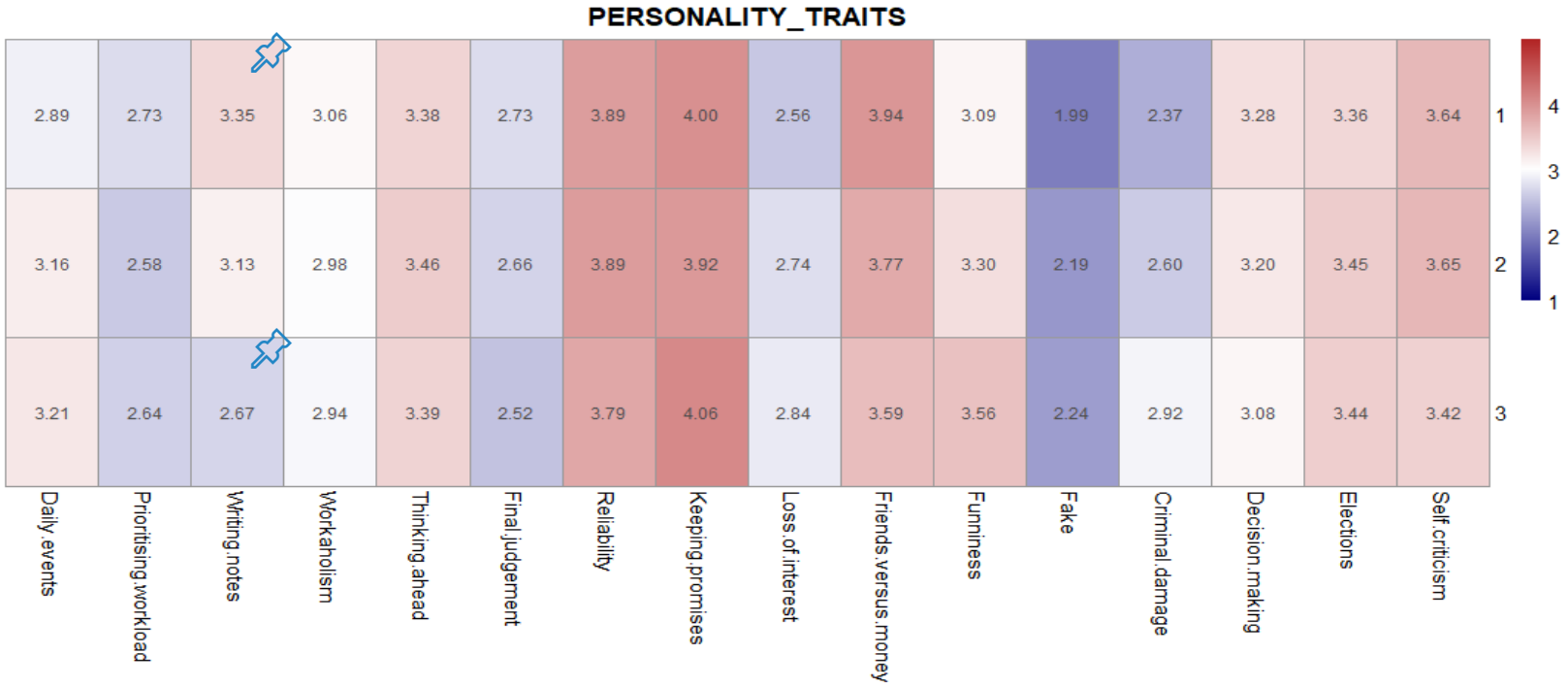
Clustering

Describe the composition of the participants



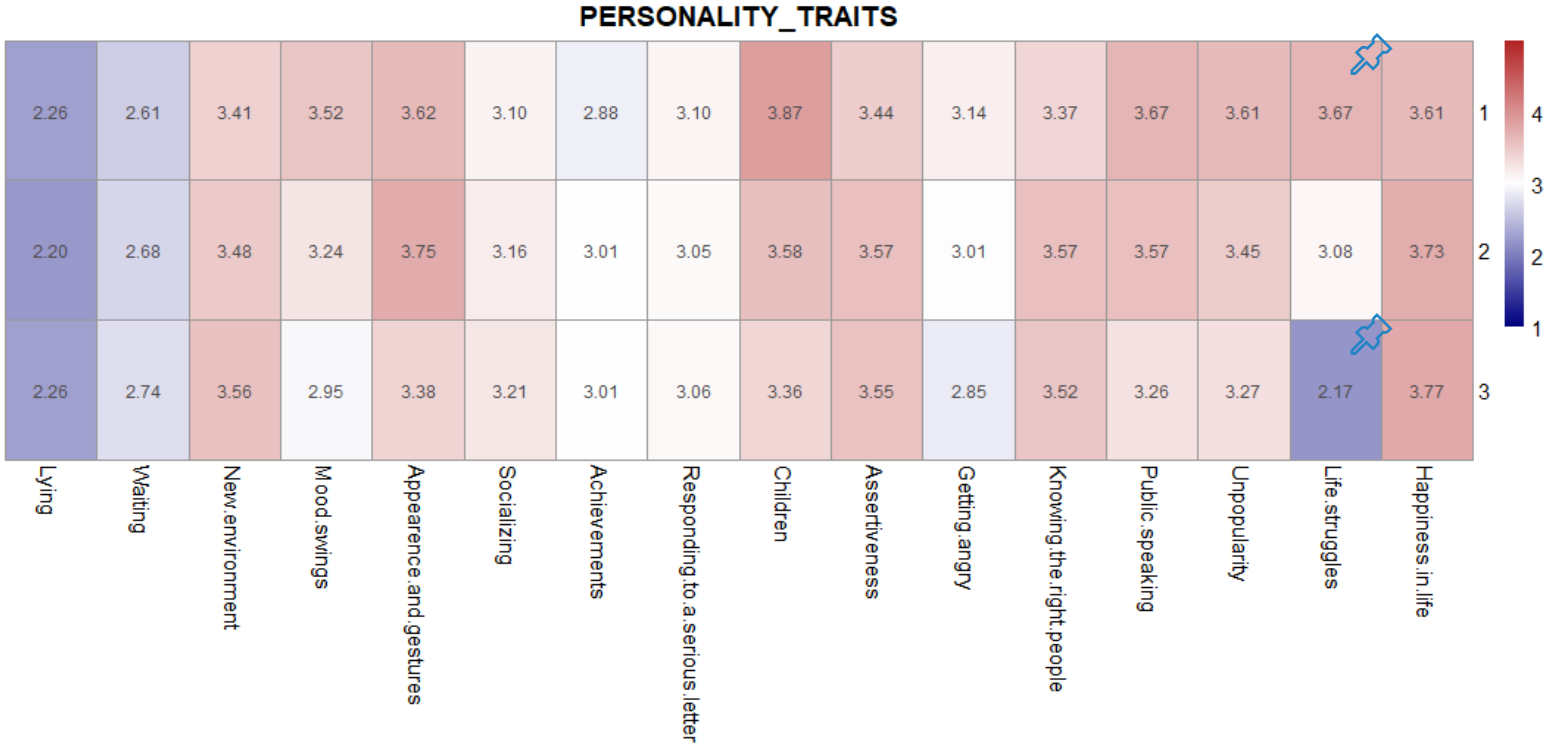
Clustering

Describe the composition of the participants



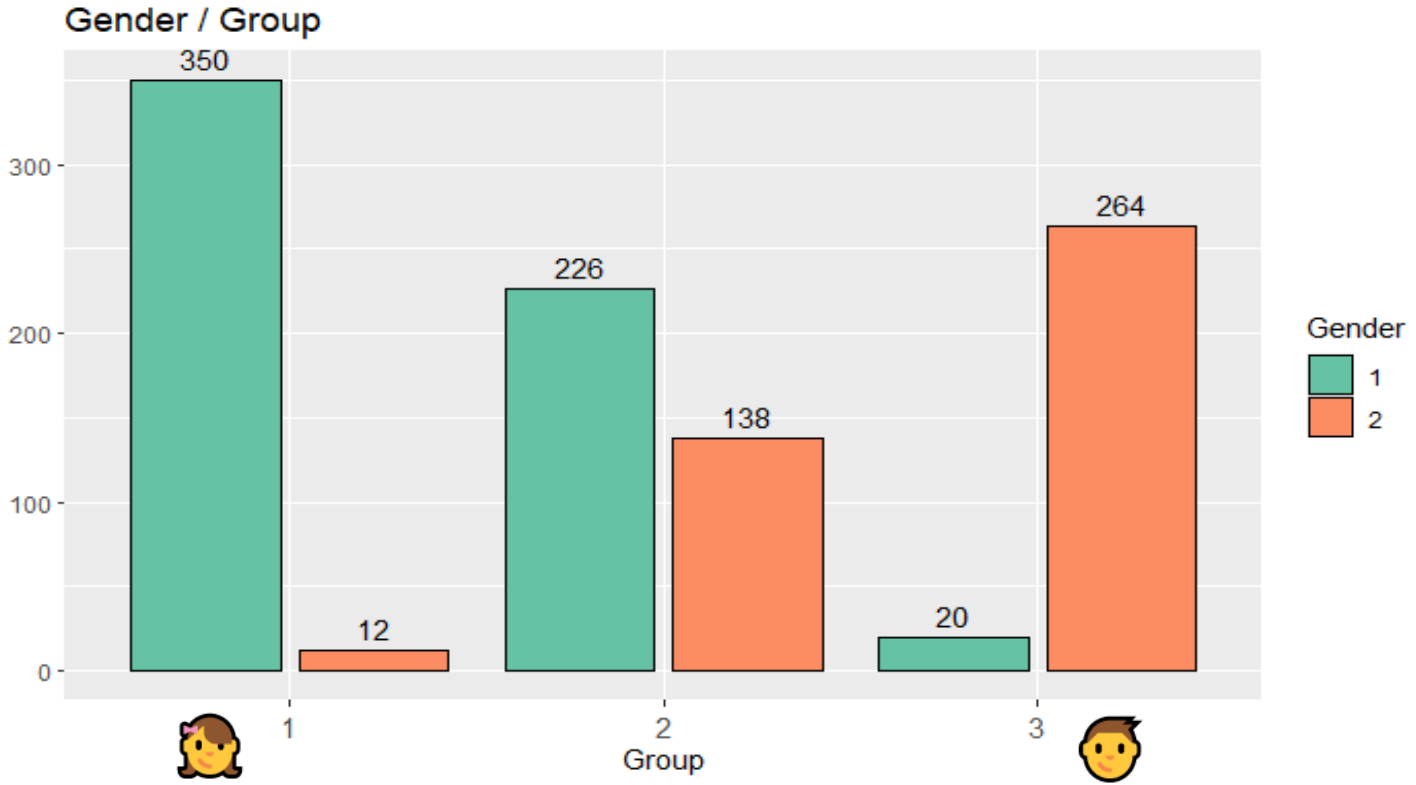
Clustering

Describe the composition of the participants



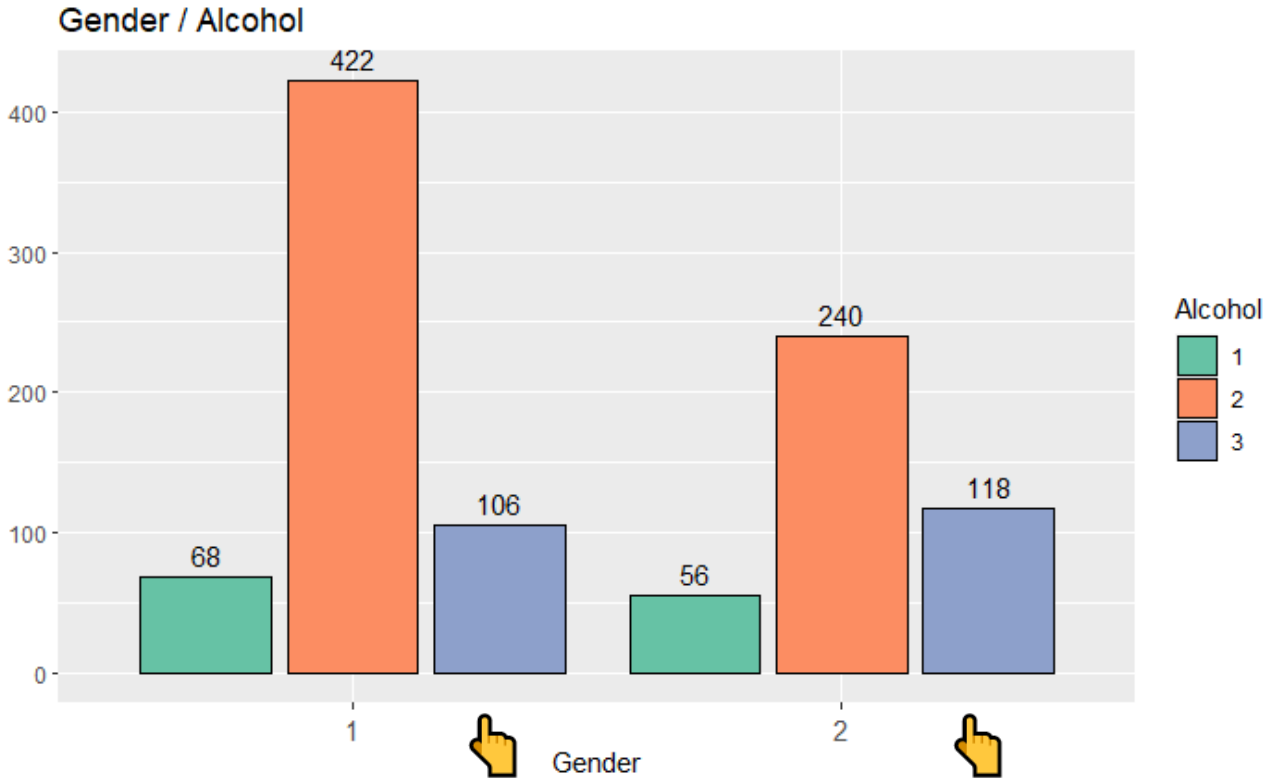
Clustering

Describe the composition of the participants



Predict

Predict someone have alcohol addiction with their other features



XGBoost

假設其分配為 Bernouli分配 如下，其Logistic的損失函數：

$$\prod_{1}^{n} \pi(x_i)^{yi}[1-\pi(x_i)]^{1-yi}, \pi(x) = \frac{exp(\beta_i X_i)}{1+exp(\beta_i X_i)}$$

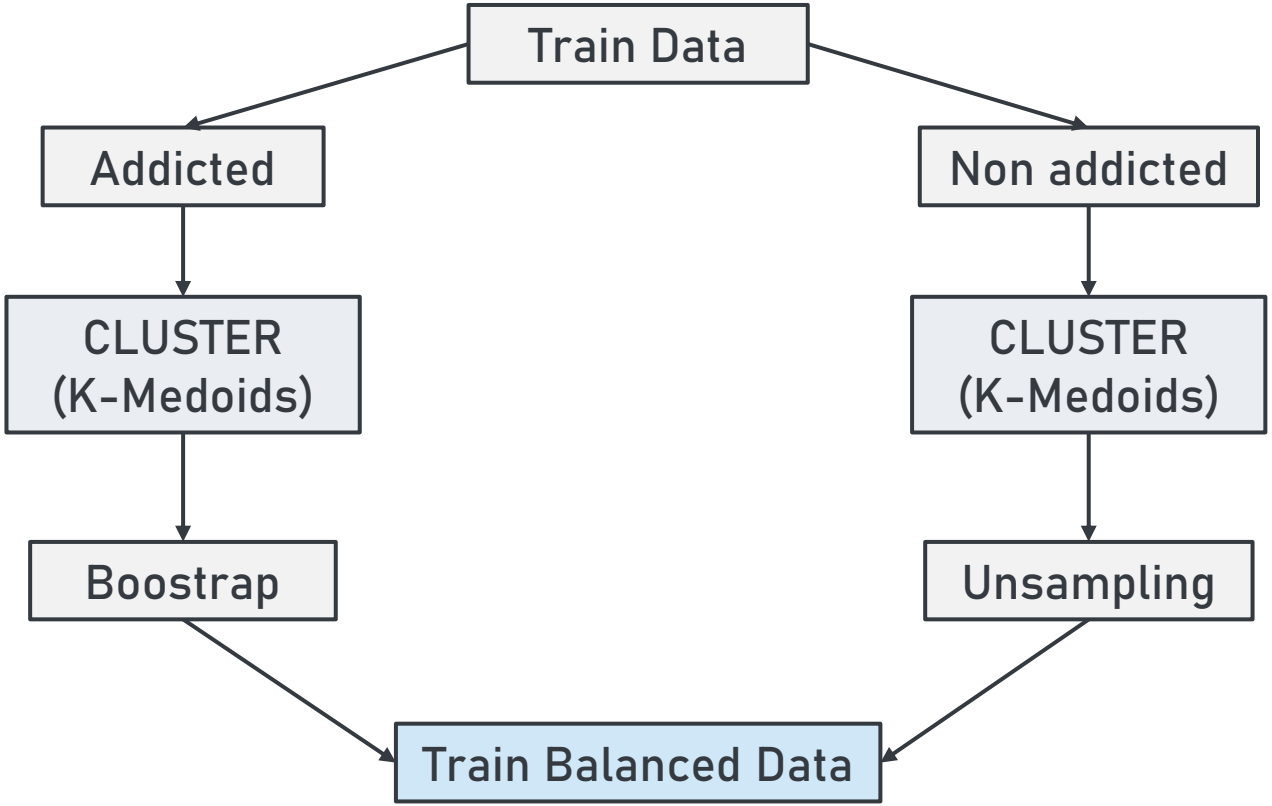
Boosting 中， $h_m(x_i)$代表經過m棵樹迭代後的估計值， 同Logistic裡的$\Sigma_{i=1}^{n}\beta_i Xi$

$$\hat{f}(x_i) = \Sigma_{m=1}^{M} h_m(x_i) = \Sigma_{i=1}^{n}\beta_i Xi$$

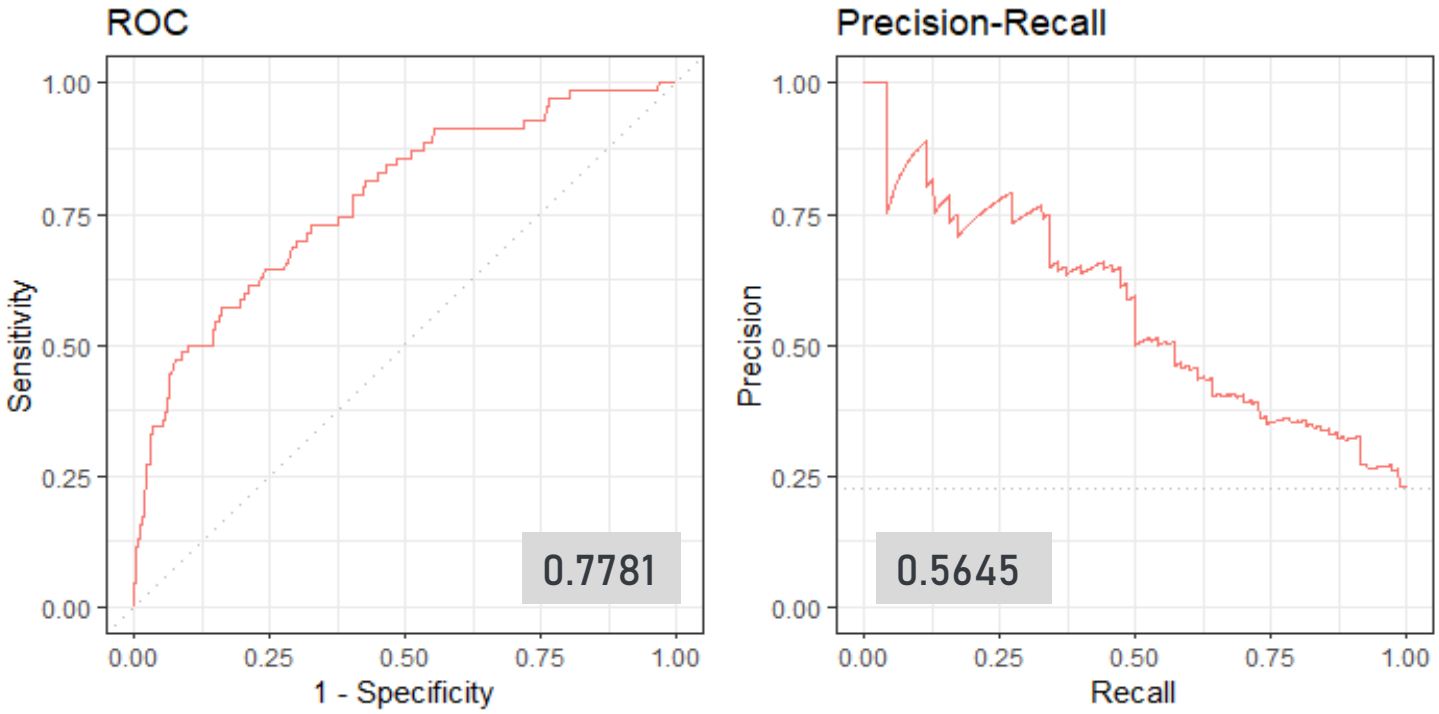
將其帶入上述損失函數後，可得下式：

$$L(y_i, f(x)) = y\ln(1+e^{-f(x)}) + (1-y)\ln(1+e^{f(x)})$$

Rebalance

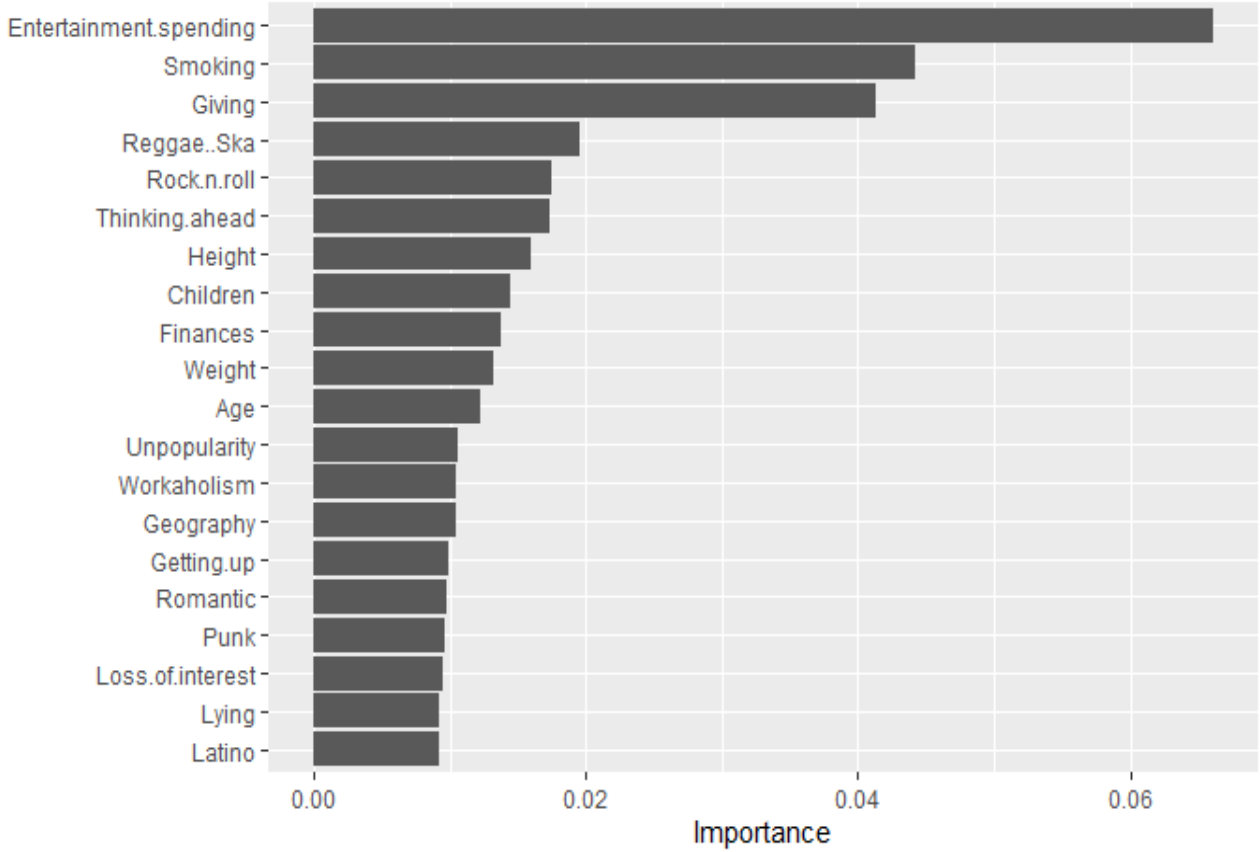

Train Data
Addicted
Non addicted
CLUSTER
(K-Medoids)
CLUSTER
(K-Medoids)
Boostrap
Unsampling
Train Balanced Data

ROC / Recall



Test Set

IMPORTANCE PLOT



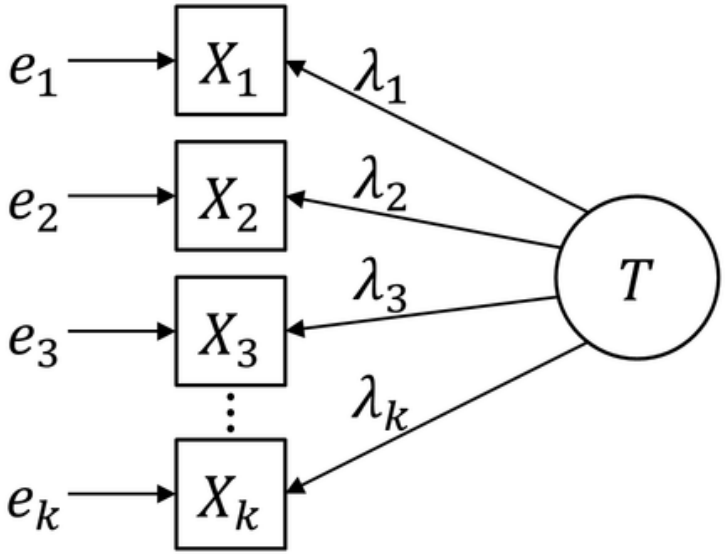
Relationship

Find out the relationship among the features that be selected.

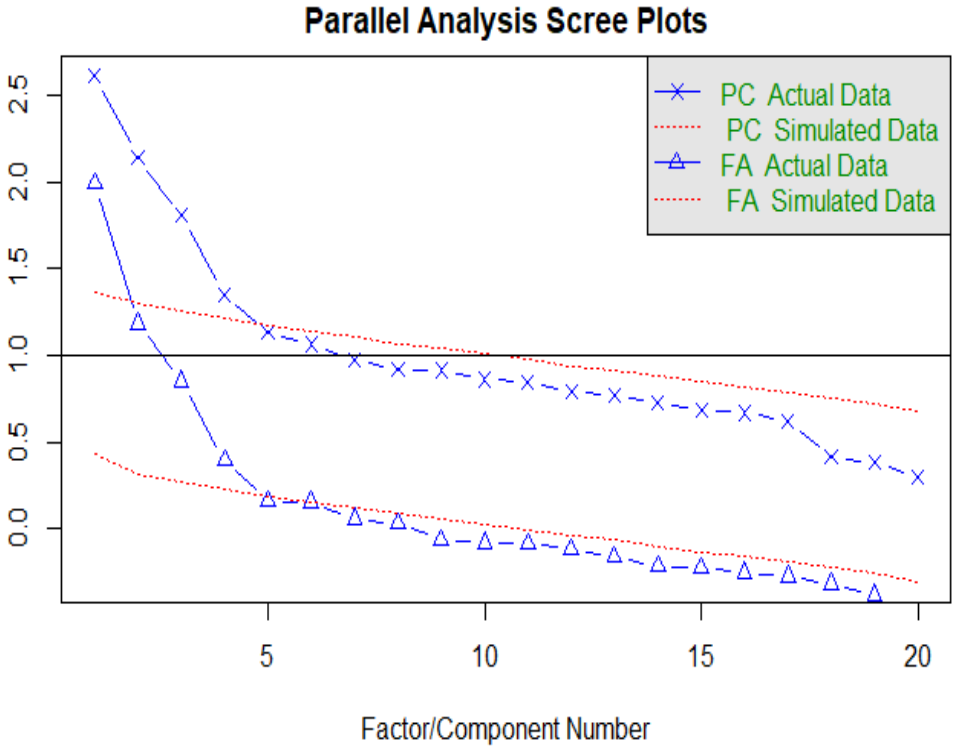
Factor Analysis：

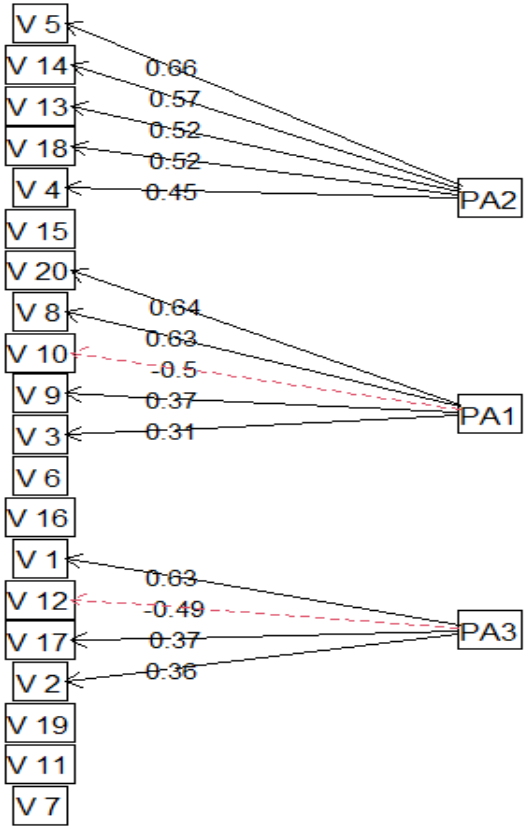
$$\boldsymbol{x = \mu + Fz + \epsilon}$$

- x: *Random Variables*
- μ: *Expection of* x
- F: *Factor Loading*
- z: *Hidden Facotr*
- ϵ: *Idiosyncratic factor*

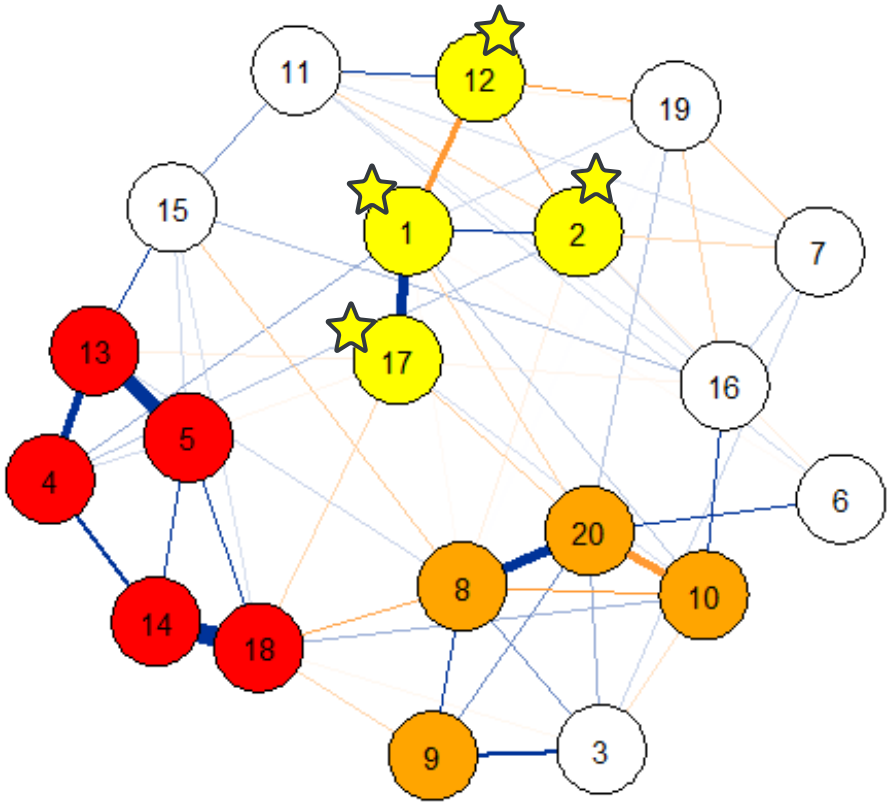


Factor Analysis





Factor Analysis

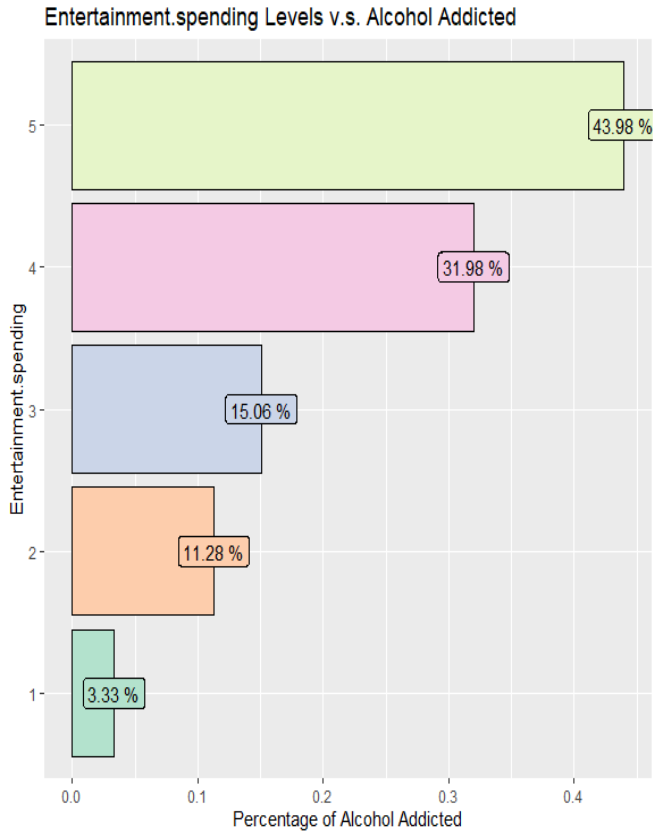


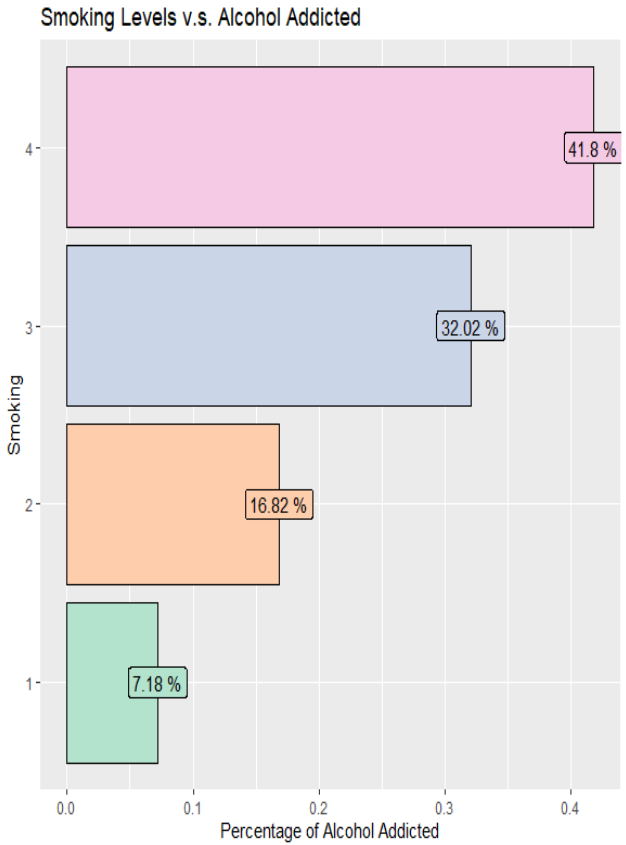
A
- 4: Reggae..Ska
- 5: Rock.n.roll
- 13: Swing..Jazz
- 14: Punk
- 18: Metal.or.Hardrock

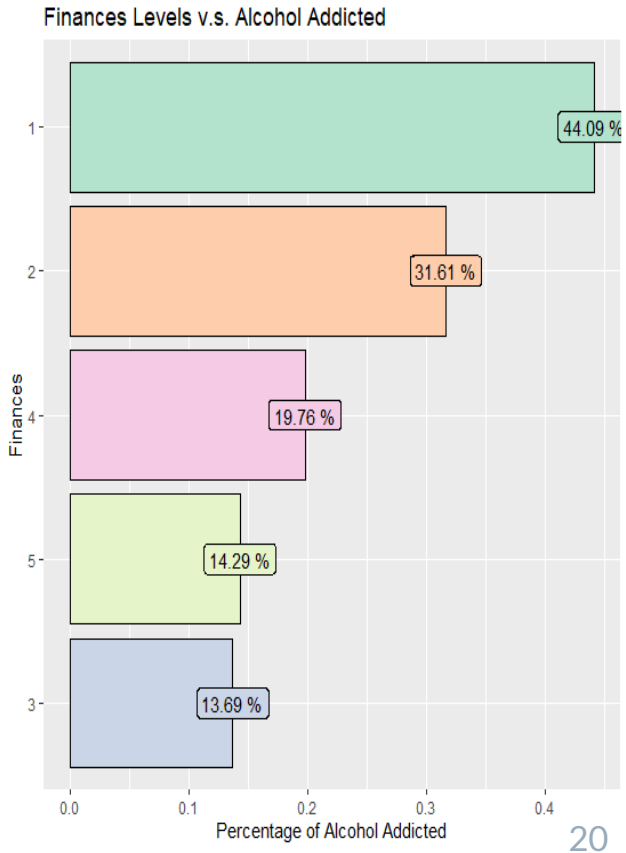
B
- 8: Romantic
- 9: Children
- 10: Weight
- 20: Life.struggles

C
- 1: Entertainment.spending
- 2: Smoking
- 12: Finances
- 17: Branded.clothing

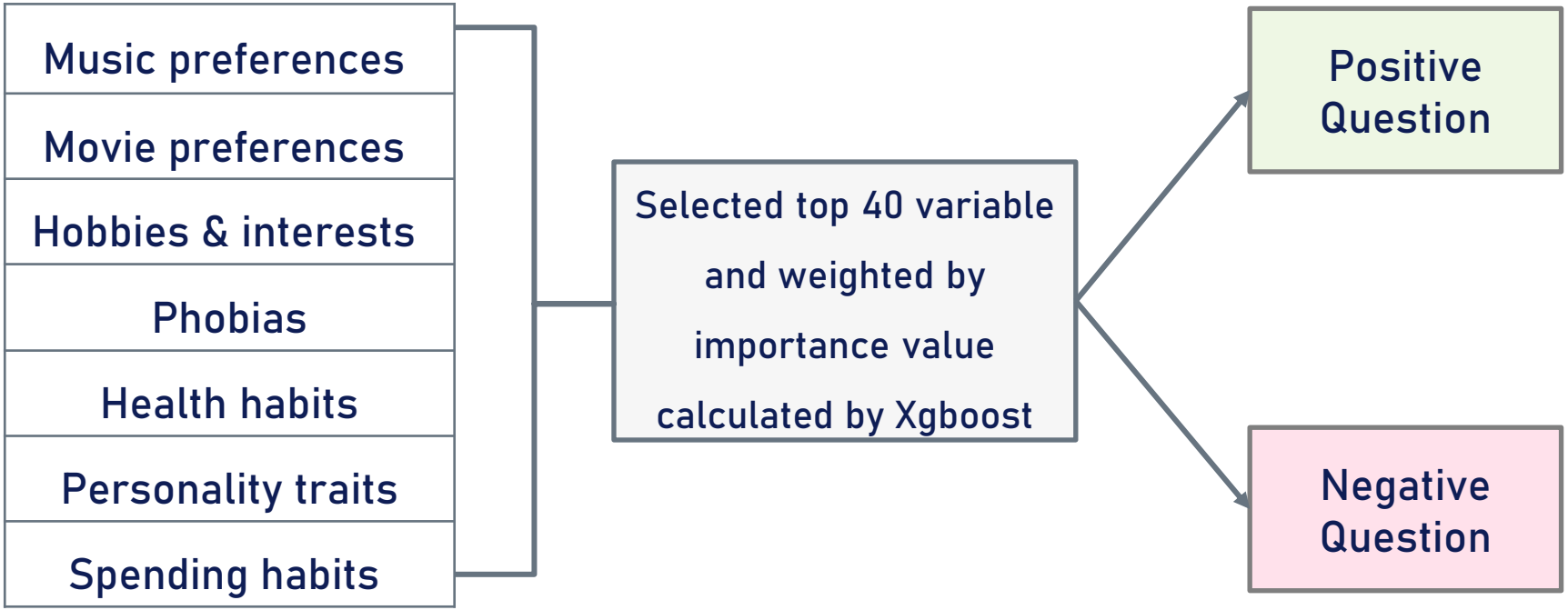
Visualization



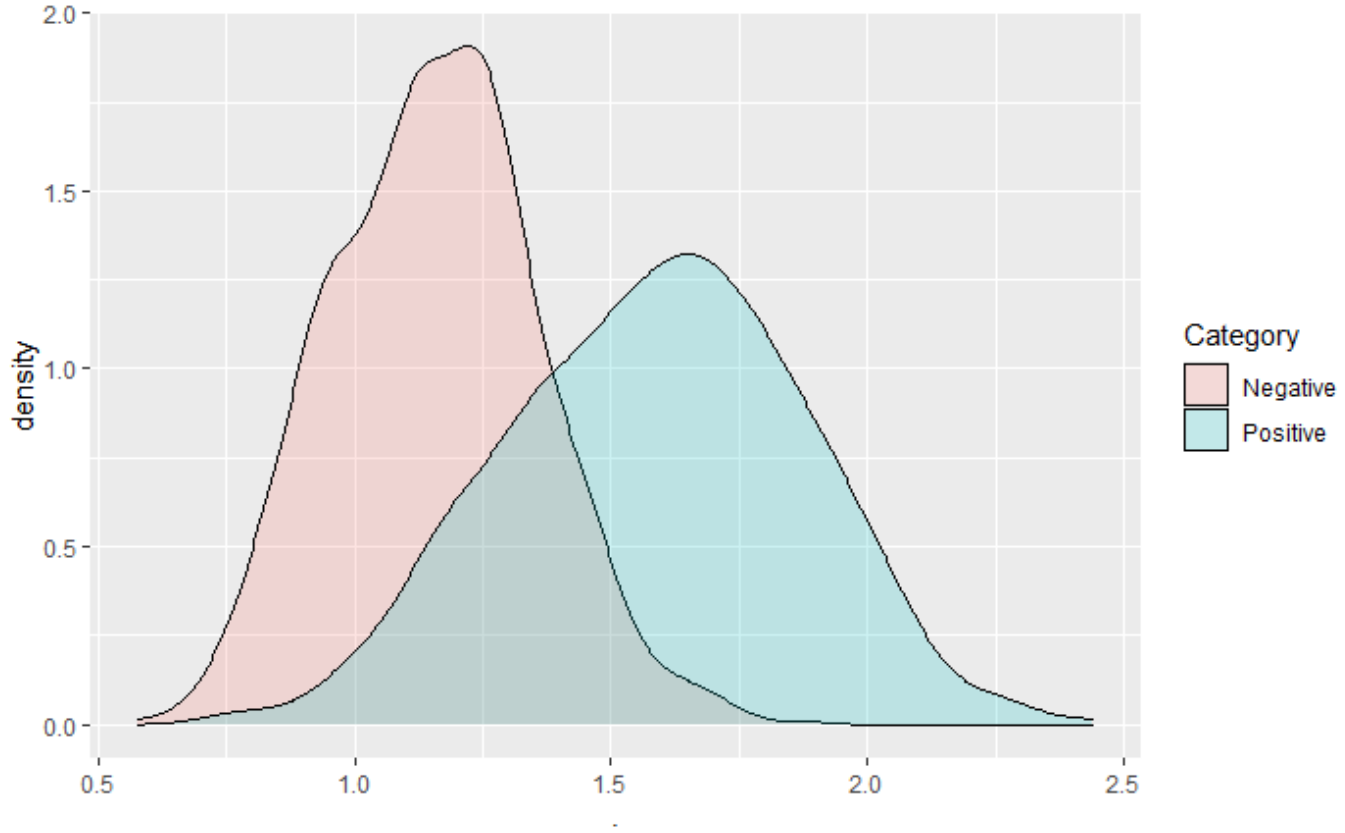




GMM

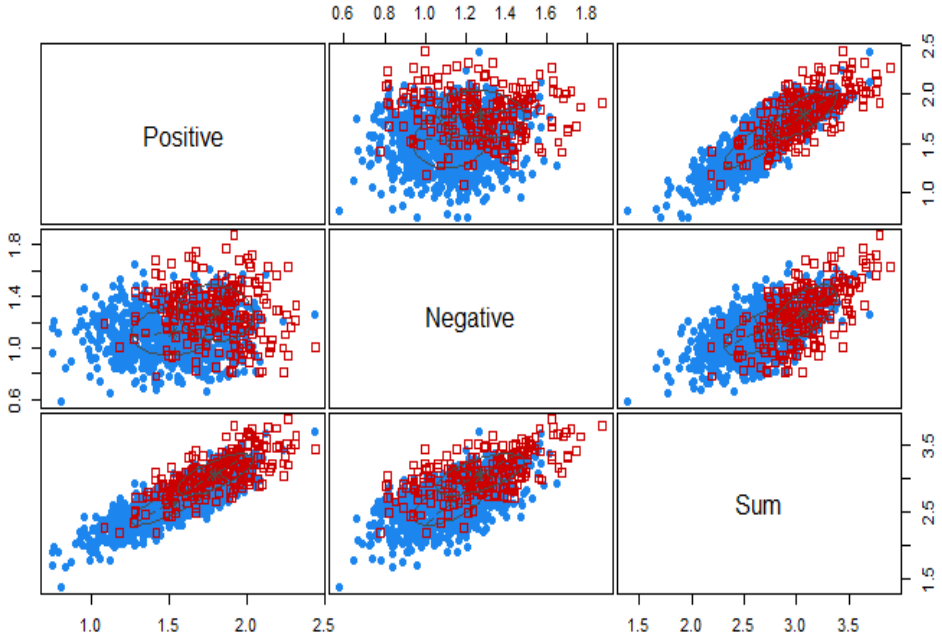

Music preferences
Movie preferences
Hobbies & interests
Phobias
Health habits
Personality traits
Spending habits
Selected top 40 variable and weighted by importance value calculated by Xgboost
Positive Question
Negative Question

GMM

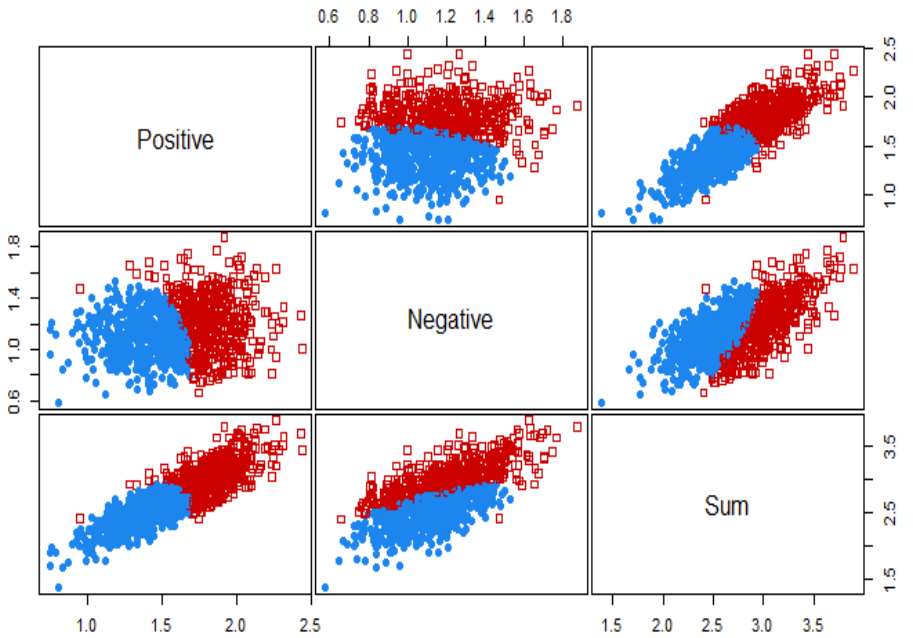

2.0
1.5
1.0
0.5
0.0
density
0.5
1.0
1.5
2.0
2.5
Category
Negative
Positive

GMM

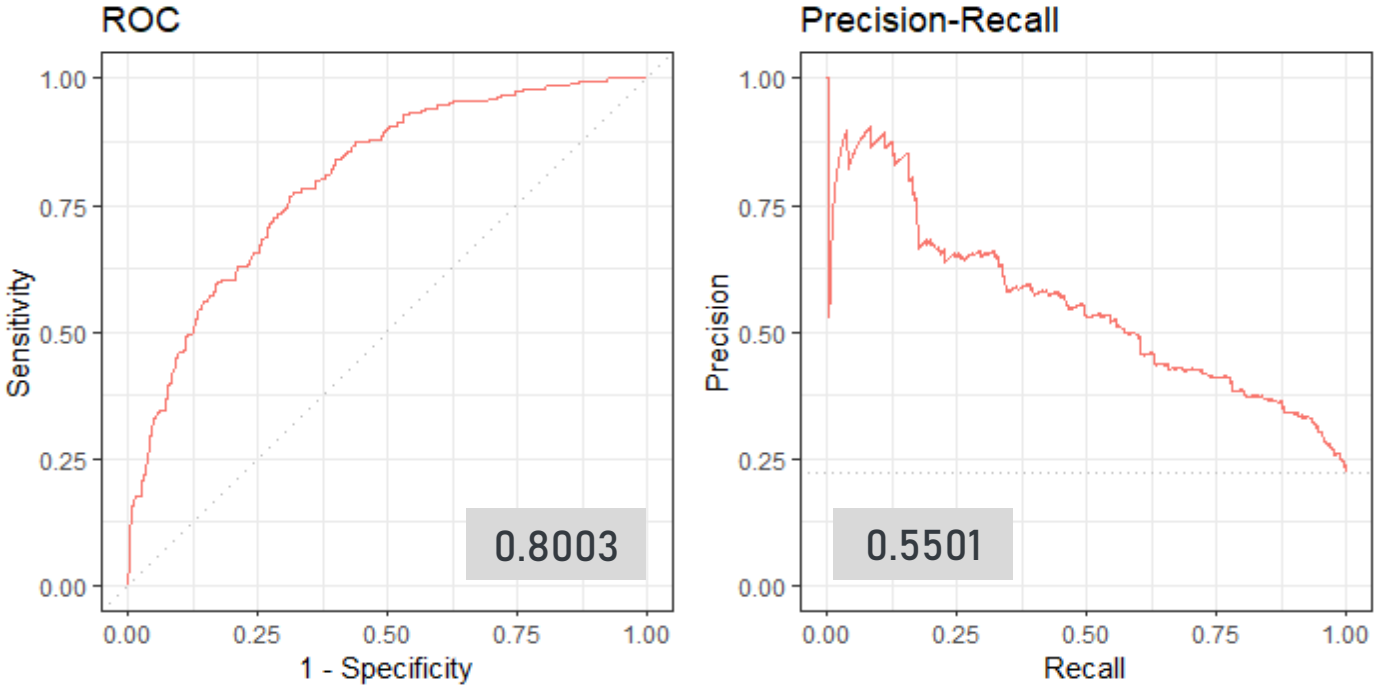
True Value Scattor Plot



Classification



GMM



Train Set