The Impact on Body Signals of Smoking and Drinking

FALL2023 CSC240 Final Presentation

Yuesong Huang, Junhua Huang, Yuyang Wang, Boyang Wang

University of Rochester

MOTIVATION

"Personal Experiences Inspire Research "

Team of smokers, drinkers, and non-users 🚬 🍷 🚫

How do these habits affect health? 💭?

Combine the personal experience with the courses technique.

Expanding from personal to universal insights 🌍.

Literature review

- *The impact of current smoking, regular drinking, and physical inactivity on health care-seeking behavior in China*
  - Adults who are current smokers are 0.65 times less likely to seek health care than former smokers.
  - Adults who regularly drink alcohol are less likely to seek health care than non-drinkers.
- *Predicting Tobacco and Alcohol Consumption Based on Physical Activity Level and Demographic Characteristics in Romanian Students*
  - Tobacco and Alcohol Consumption can be predicted
  - 'Results showed that moderate consumption of tobacco and harmful consumption of alcohol had high prevalence among age, gender, year of study and PA(Physical activity) level categories.'

DATASET

sex	age	height	weight	waistline	sight_left	sight_right	hear_left	hear_right	SBP	DBP	BLDS
Male	35	170	75	90	1	1	1	1	120	80	99
Male	30	180	80	89	0.9	1.2	1	1	130	82	106
Male	40	165	75	91	1.2	1.5	1	1	120	70	98
Male	50	175	80	91	1.5	1.2	1	1	145	87	95
Male	50	165	60	80	1	1.2	1	1	138	82	101
Male	50	165	55	75	1.2	1.5	1	1	142	92	99
Female	45	150	55	69	0.5	0.4	1	1	101	58	89
Male	35	175	65	84.2	1.2	1	1	1	132	80	94
Male	55	170	75	84	1.2	0.9	1	1	145	85	104
Male	40	175	75	82	1.5	1.5	1	1	132	105	100
Male	45	155	55	79.2	1	1	1	1	118	70	90
Male	65	155	75	98	1.2	9.9	1	1	109	69	137
Female	55	150	55	72.3	1.2	0.9	1	1	130	80	106
Male	30	175	75	88	1.2	1.2	1	1	118	72	82
Female	30	160	50	76	0.9	1	1	1	129	77	79
Male	40	170	65	80	1	1	1	1	113	72	104
Female	25	160	65	73	1.2	0.9	1	1	126	78	96
Male	25	170	65	78	1.2	1.2	1	1	119	67	100
Male	50	170	85	99	0.7	0.8	1	1	121	74	99
Male	60	165	60	85	0.3	0.7	1	1	120	85	105
Female	35	170	50	67	1	0.8	1	1	111	65	88
Male	25	175	65	82	1.5	1.5	1	1	130	76	95
Female	45	155	50	62	0.5	0.7	1	1	109	64	111
Male	40	165	75	92	1	1.5	1	1	110	70	102
Female	20	160	55	79	1.2	1.5	1	1	110	70	87

SOO.Y · UPDATED 3 MONTHS AGO

▲ 104 | New Notebook | Download (29 MB)

Smoking and Drinking Dataset with body signal

Predict smokers and drinkers using body signal data.

1. The Smoking and Drinking Dataset with Body Signal
2. 991,346 observations
3. Includes crucial data points like age, height, weight, blood pressure, cholesterol, hemoglobin, and smoking/drinking status."

- [1] Soo.Y, “Smoking and drinking dataset with body signal,” Kaggle, https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset (accessed Dec. 8, 2023).

Attributes introduction:

- **SBP :** Systolic blood pressure [mmHg]
- **DBP :**Diastolic blood pressure [mmHg]
- **SGOT_AST :** SGOT (Glutamate-oxaloacetate transaminase) AST (Aspartate transaminase) [IU/L]
- **SGOT_ALT :** ALT (Alanine transaminase) [IU/L]
- **Tot_chole :** total cholesterol [mg/dL]
- **Gamma_GTP :** y-glutamyl transpeptidase [IU/L]

Correlation Observations

Some strong correlation observations:

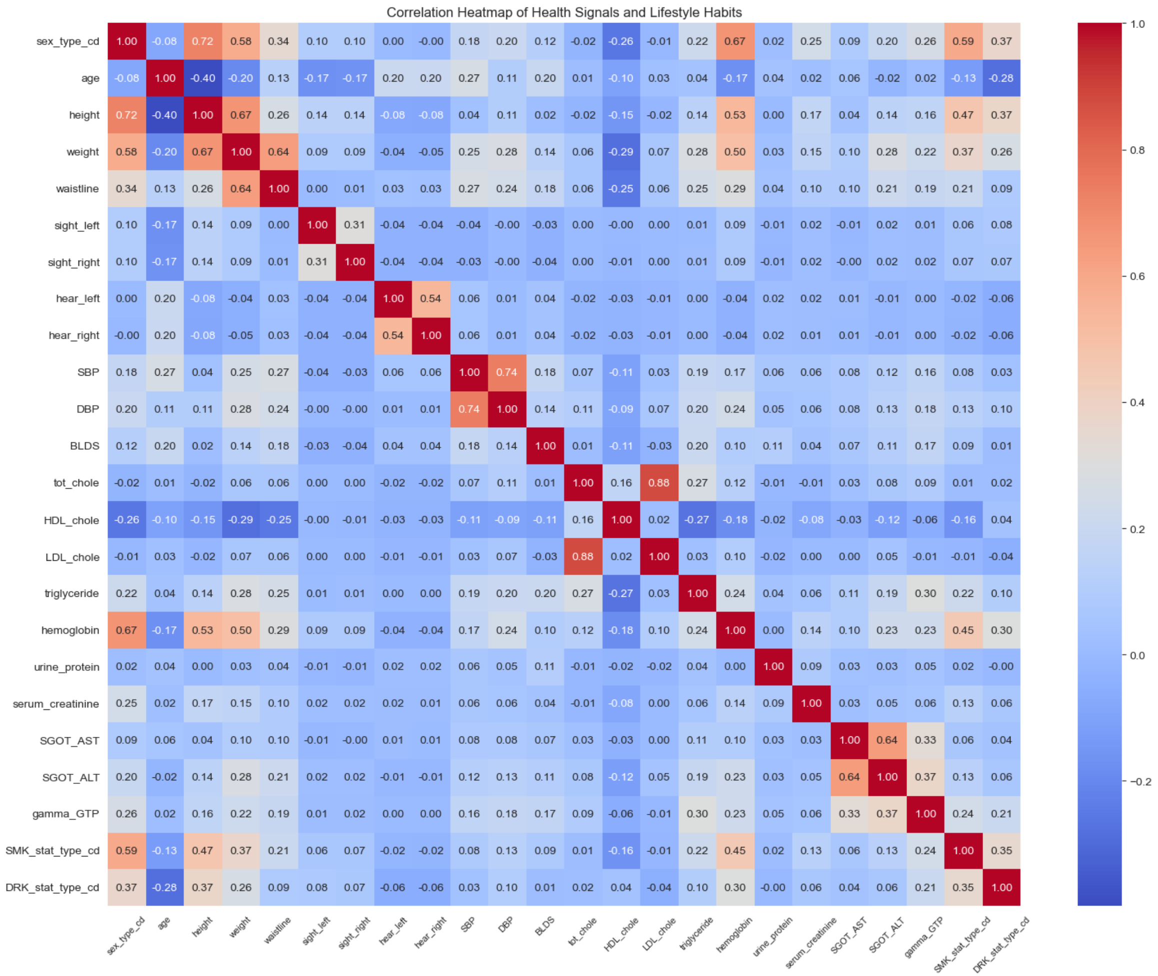
The correlation between LDL_chole and tot_chole is Strong (0.877367).

Some medium correlation observations:

The correlation between height and sex_type_cd is Medium (0.722774).

The correlation between weight and sex_type_cd is Medium (0.581707).

The correlation between hemoglobin and height is Medium (0.531898).



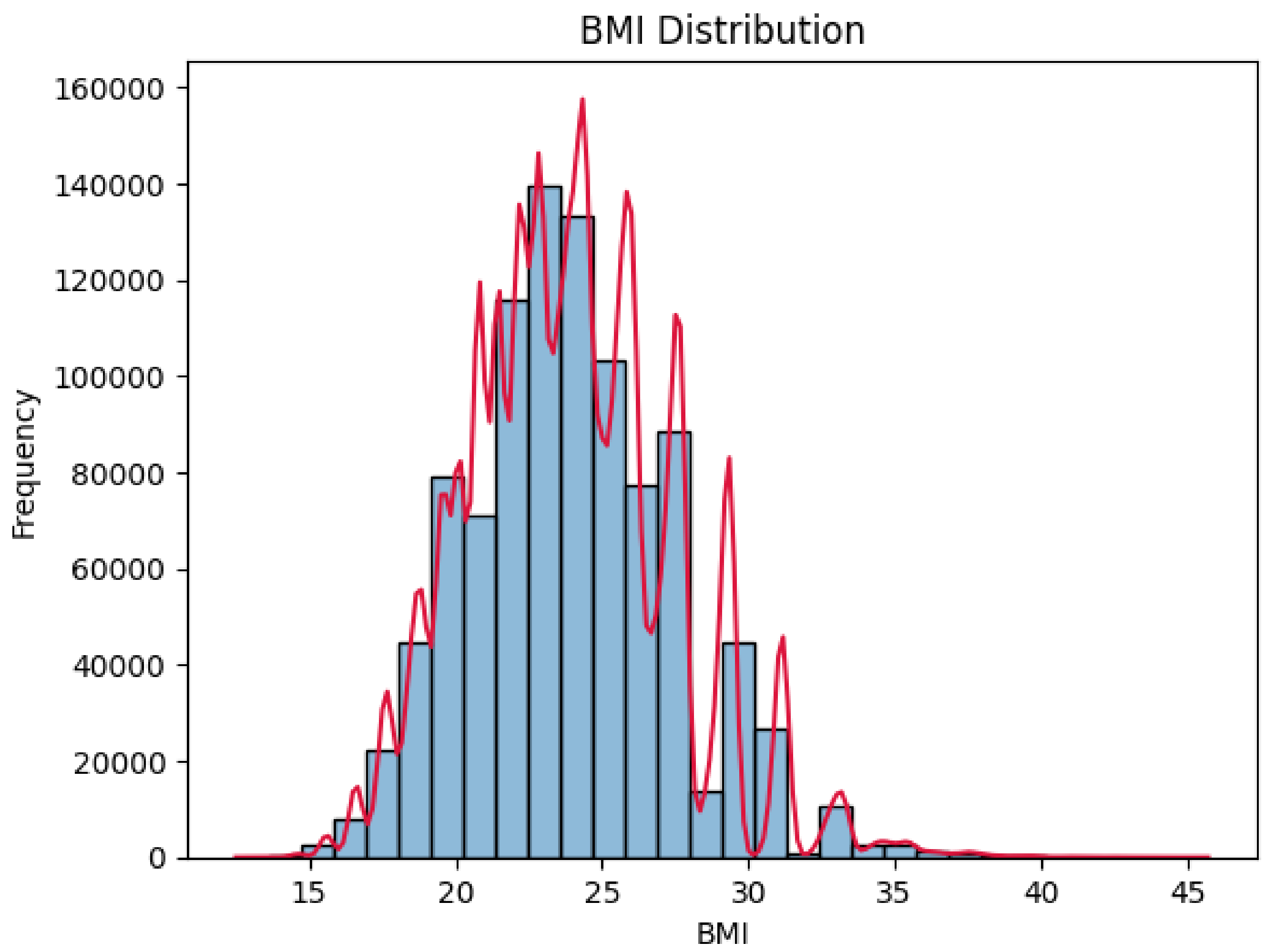
Hypotheses & Goals

- Find the correlation between lifestyle and health condition
- Infer somebody's drinking and smoking status based on their health condition
- Assess the risk levels associated with smoking and drinking patterns
- Improve public health using advanced machine learning mechanisms.

Data Preprocessing & Feature Engineering

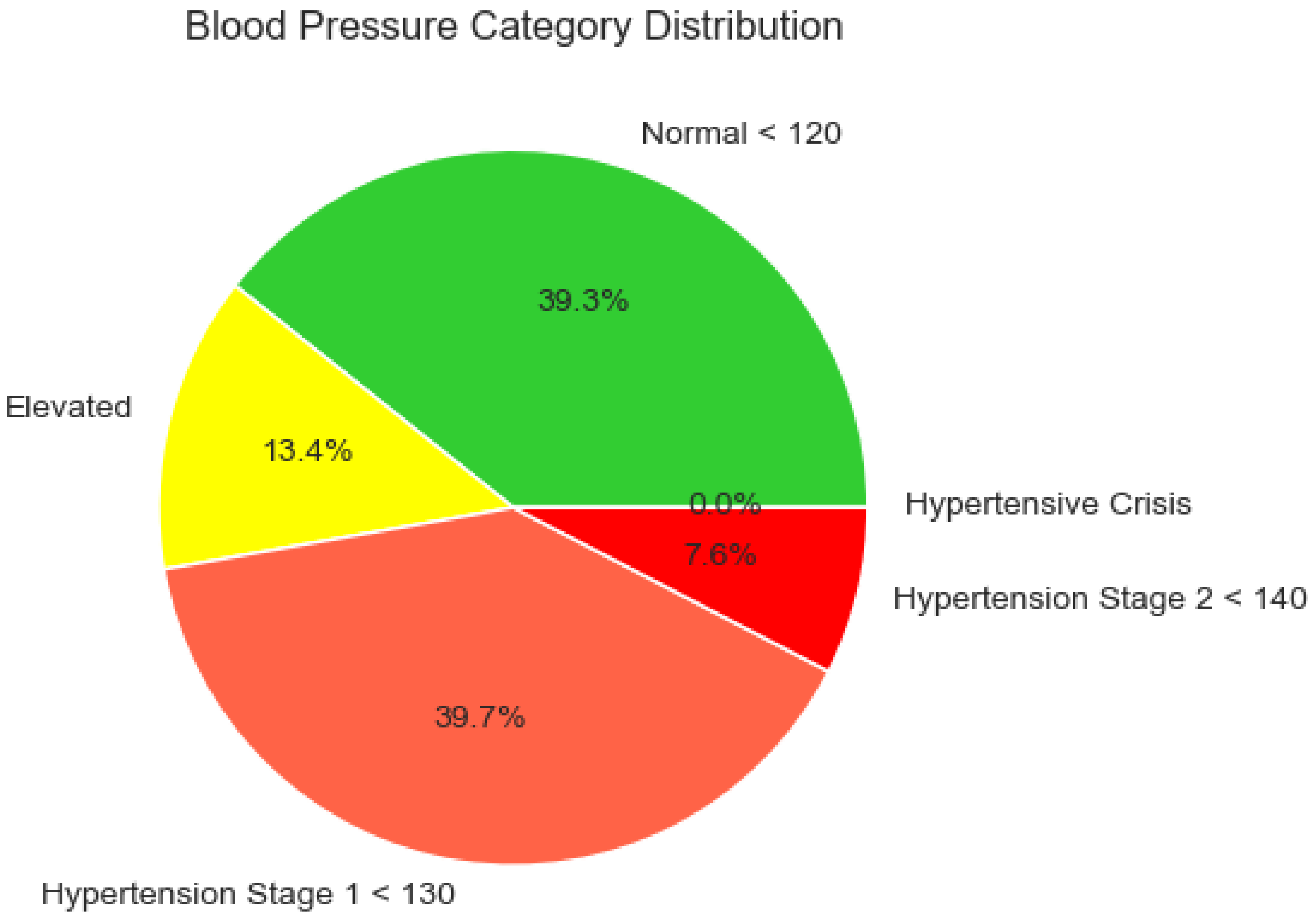
Body Mass Index

- BMI = Weight (kg) / ((Height (m))^2
- A insightful attributes that links the weight and height to the smoking and Drinking
- Gain the value from “weight” &”Height”
- **Why BMI Matters:**
  - Assesses weight categories.
  - Predicts health risks.
  - Relevant to lifestyle habits.



Blood pressure

- **A key indicator comprising systolic and diastolic pressures.**
- we categorize blood pressure into stages.
- Blood Pressure Categorizing:
- Elevated, stage1, stage2…
- Helps in accurately identifying individuals at different levels of health risk.



Apriori

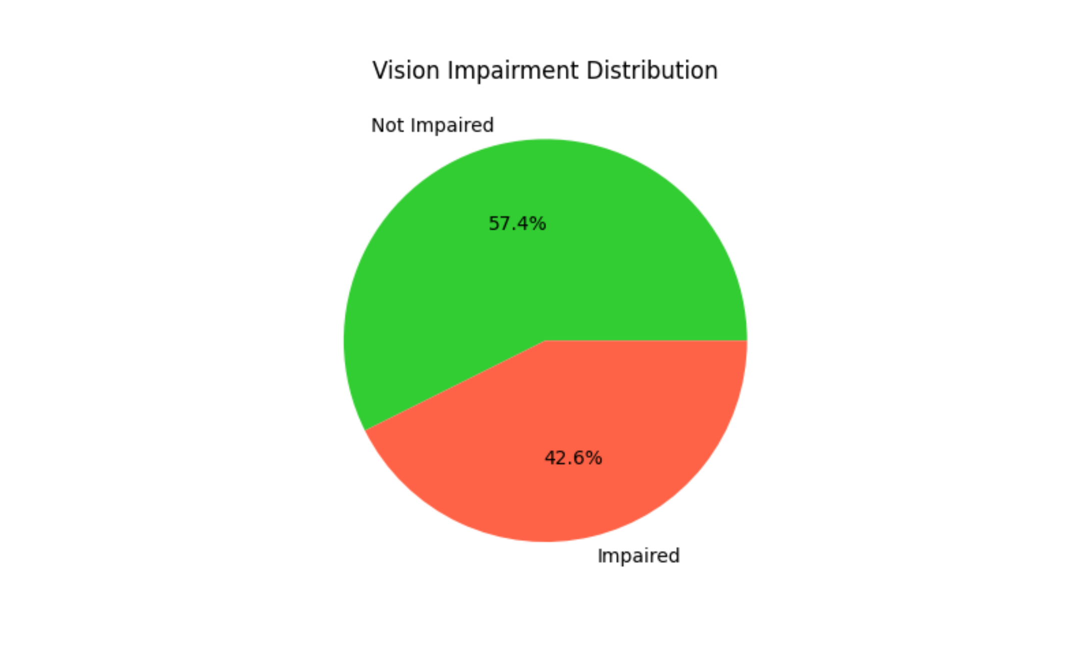
minSup=0.2

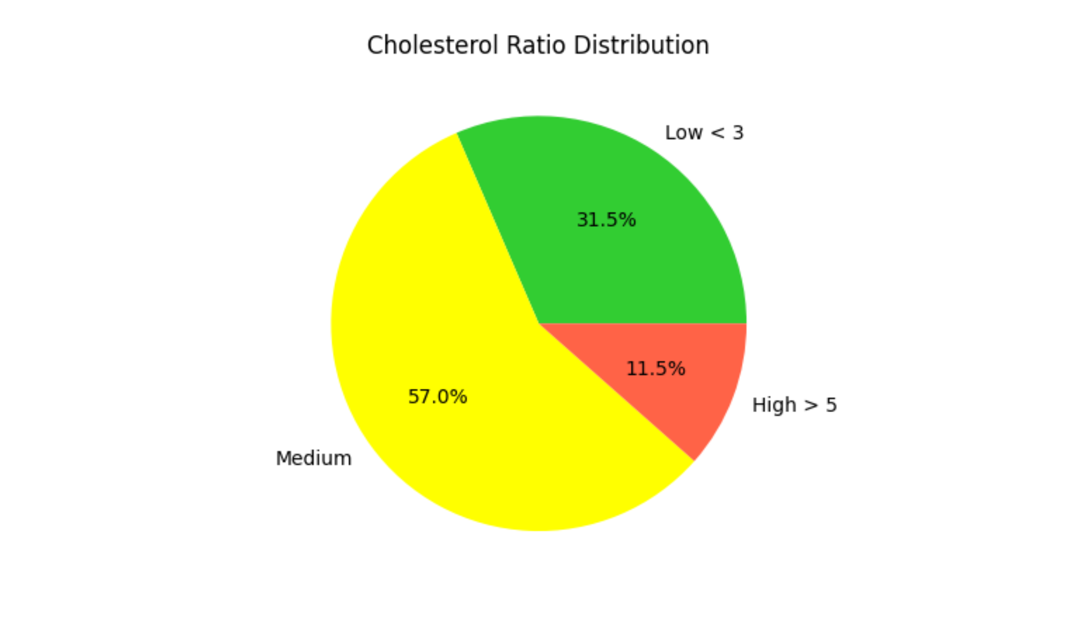
index	support	itemsets
8	0.943338652	frozenset({'urine_protein=1.0'})
2	0.607700036	frozenset({'SMK_stat_type_cd=0.0'})
16	0.575223988	frozenset({'SMK_stat_type_cd=0.0', 'urine_protein=1.0'})
5	0.531010363	frozenset({'sex_type_cd=1.0'})
0	0.50018661	frozenset({'DRK_stat_type_cd=0.0'})
1	0.499813385	frozenset({'DRK_stat_type_cd=1.0'})
19	0.498955964	frozenset({'urine_protein=1.0', 'sex_type_cd=1.0'})
11	0.472471770	frozenset({'urine_protein=1.0', 'DRK_stat_type_cd=0.0'})
14	0.470866881	frozenset({'DRK_stat_type_cd=1.0', 'urine_protein=1.0'})
4	0.468989636	frozenset({'sex_type_cd=0.0'})

index	support	itemsets
10	0.32658628	frozenset({'sex_type_cd=0.0', 'DRK_stat_type_cd=0.0'})
20	0.315913919	frozenset({'SMK_stat_type_cd=0.0', 'sex_type_cd=0.0', 'DRK_stat_type_cd=0.0'})
22	0.309761677	frozenset({'sex_type_cd=0.0', 'urine_protein=1.0', 'DRK_stat_type_cd=0.0'})
26	0.299769202	frozenset({'SMK_stat_type_cd=0.0', 'sex_type_cd=0.0', 'urine_protein=1.0', 'DRK_stat_type_cd=0.0'})
3	0.215821721	frozenset({'SMK_stat_type_cd=2.0'})
12	0.215294155	frozenset({'SMK_stat_type_cd=0.0', 'DRK_stat_type_cd=1.0'})
7	0.206278130	frozenset({'sight_right=1.0'})
23	0.203424435	frozenset({'SMK_stat_type_cd=0.0', 'DRK_stat_type_cd=1.0', 'urine_protein=1.0'})
6	0.203176287	frozenset({'sight_left=1.0'})
17	0.202861563	frozenset({'SMK_stat_type_cd=2.0', 'urine_protein=1.0'})

WHAT’S MORE

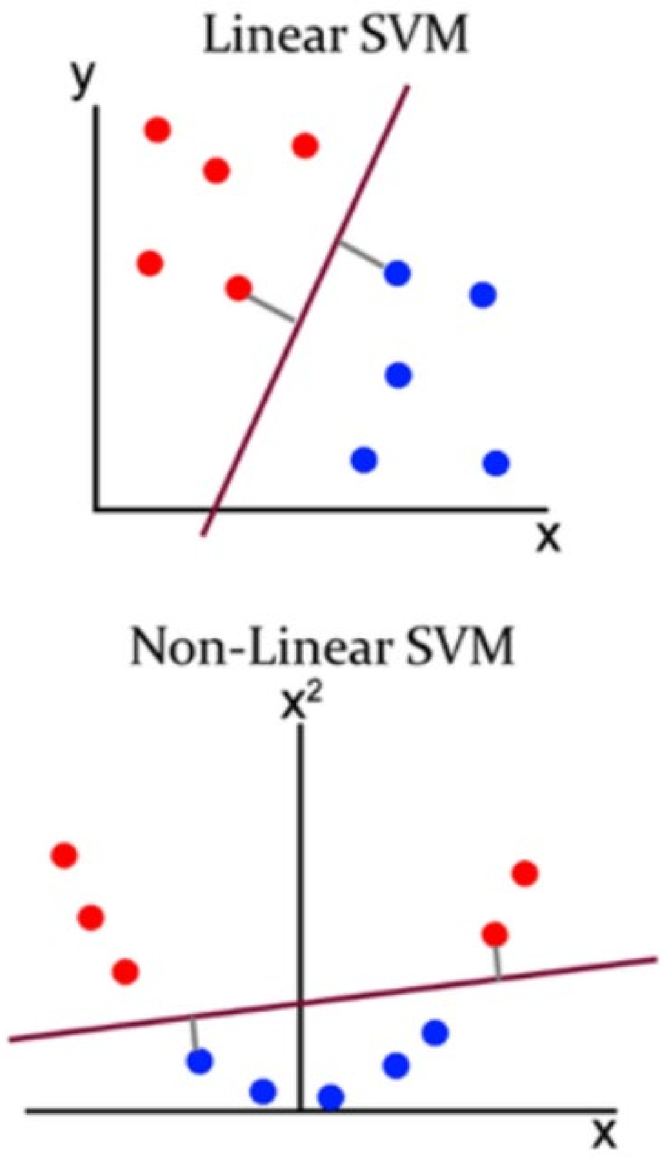
- We also tried:
  - Categorizing Cholesterol Ratio
  - Vision Impairment
  - Spectrum & Kmean
  - Minimax & Standard Scalar
  - PCA
  - ......





Model

Construction & Evaluation



Model Building

- Linear & Kernel Support Vector Machine (SVM)
- SVM is a powerful Classifier
- Cons: not very time efficient
- `GridSearchCV(xgb_DRK, param_grid=param_grid, cv=5)`

```
param_grid = [
    {
        'scaler': [StandardScaler(), MinMaxScaler()],
        'classifier__C': [0.001, 0.01, 0.1, 1.0, 5.0],
        'classifier__kernel': ['linear']
    },
    {
        'classifier__C': [0.001, 0.01, 0.1, 1.0, 5.0],
        'classifier__gamma': [0.001, 0.01, 0.1, 1.0, 'scale', 'auto'],
        'classifier__kernel': ['rbf']
    }
]
```

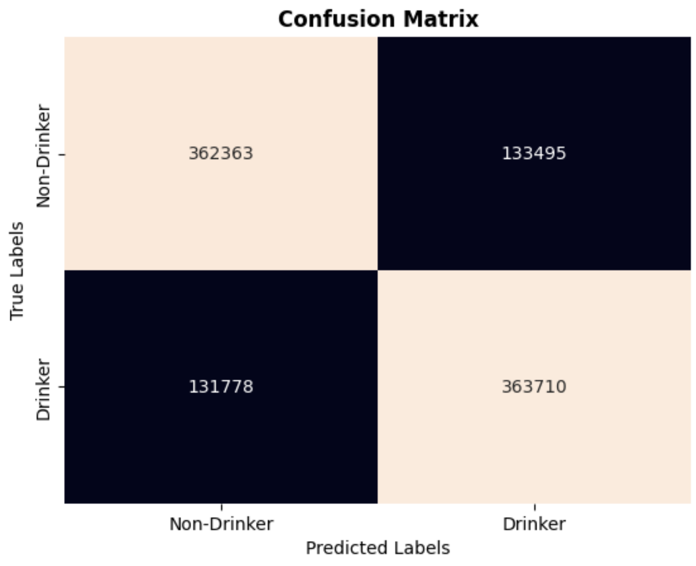
Model Building

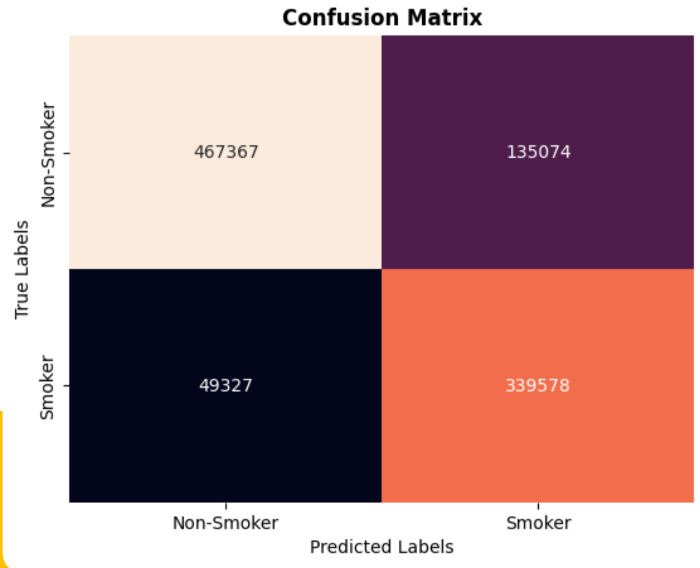
Linear & Kernel Support Vector Machine (SVM)

- SVM is a powerful Classifier
- Cons: not very time efficient
- `GridSearchCV(svm, param_grid=param_grid, cv=5)`

Executed at 2023.12.08 08:13:30 in 2h 17m 52s

```
Fitting 5 folds for each of 40 candidates, totalling 200 fits
[CV] END classifier__C=0.001, classifier__kernel=linear, scaler=StandardScaler();
[CV] END classifier__C=0.001, classifier__kernel=linear, scaler=StandardScaler();
[CV] END classifier__C=0.001, classifier__kernel=linear, scaler=StandardScaler();
[CV] END classifier__C=0.001, classifier__kernel=linear, scaler=StandardScaler();
[CV] END classifier__C=0.001, classifier__kernel=linear, scaler=StandardScaler();
```





Model Building

Sample Size: 100,000

- Smoking Status Prediction (svm_SMK)
  - Pipeline: StandardScaler and SVC with linear kernel, C=0.001.
  - Accuracy for Smoker prediction: 0.8542
  - F1 score for Smoker prediction: 0.8197

- Drinking Status Prediction (svm_DRK)
  - Similar setup as svm_SMK .
  - Accuracy for Drinker prediction: 0.8162
  - F1 score for Drinker prediction: 0.7775

- Accuracy for overall prediction: 0.8351597323235278
- F1 score for overall prediction: 0.7985889022361989

Model Building

- We start looking for other classifiers such as
  - Random Forest
  - Naïve Bayesian Network.
  - XGBoost
  - MLP (Impractical in a reasonable time)
  - etc.

Model Building

XGBoost

- XGBoost is a highly efficient Classifier
- `GridSearchCV(xgb, param_grid=param_grid, cv=5)`

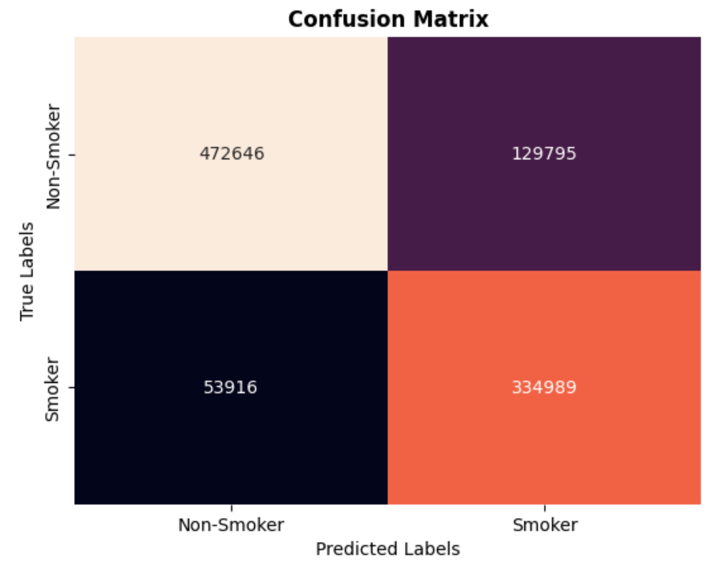
```
param_grid = {
    'scaler': [StandardScaler(), MinMaxScaler()],
    'classifier__n_estimators': [100, 200],
    'classifier__learning_rate': [0.01, 0.1],
    'classifier__subsample': [0.8, 1.0],
    'classifier__colsample_bytree': [0.8, 1.0]
}
```

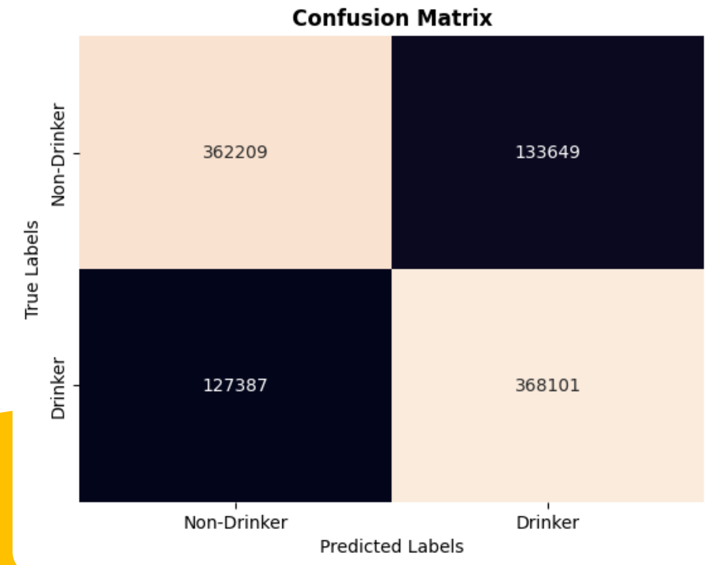
Model Building

XGBoost

- XGBoost is a highly efficient Classifier
- `GridSearchCV(xgb, param_grid=param_grid, cv=5)`

```
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=StandardScaler(); total time=   8.3s
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=StandardScaler(); total time=   5.3s
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=StandardScaler(); total time=   8.4s
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=StandardScaler(); total time=   5.3s
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=MinMaxScaler(); total time=   8.0s
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=MinMaxScaler(); total time=   5.2s
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=MinMaxScaler(); total time=   8.1s
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=MinMaxScaler(); total time=  10.2s
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.1, classifier__n_estimators=200, classifier__subsample=1.0, scaler=MinMaxScaler(); total time=   5.3s
The best parameters are {'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.1, 'classifier__n_estimators': 200, 'classifier__subsample': 1.0, 'scaler': StandardScaler()} with a score of 0.8148194
```





Model Building

Full dataset

- Smoking Status Prediction (`xgb_SMK`)
  - Pipeline: XGBClassifier with 'gbtree' booster.
  - Grid Search CV for hyperparameter tuning for the best model.
  - Accuracy for Smoker prediction: 0.8142
  - F1 score for Smoker prediction: 0.7837
  - [Classification Report]
- Drinking Status Prediction (xgb_DRK)
  - Similar pipeline and process as `xgb_SMK`.
  - Accuracy for Drinker prediction: 0.7362
  - F1 score for Drinker prediction: 0.7375
  - [Classification Report]

- Accuracy for overall prediction: 0.7751597323235278
- Accuracy for overall prediction: 0.7605889022361989

Classification Report for Smoking Status:

	precision	recall	f1-score	support
Non-Smoker	0.90	0.78	0.84	602441
Smoker	0.72	0.86	0.78	388905
accuracy			0.81	991346
macro avg	0.81	0.82	0.81	991346
weighted avg	0.83	0.81	0.82	991346

Accuracy for Smoker prediction: 0.8146852864691037
F1 score for Smoker prediction: 0.7848033651599119

Classification Report for Drinking Status:

	precision	recall	f1-score	support
Non-Drinker	0.74	0.73	0.74	495858
Drinker	0.73	0.74	0.74	495488
accuracy			0.74	991346
macro avg	0.74	0.74	0.74	991346
weighted avg	0.74	0.74	0.74	991346

Accuracy for Drinker prediction: 0.736685274364349
F1 score for Drinker prediction: 0.7382410217019408

Result & Model Evaluation

- Performance: SVM
- Time Efficiency: XGBoost
- Trade-off: between accuracy and time efficiency

Conclusion

Key Findings

- Interesting correlation between features
- Archived most Goals
- XGBoost: Efficiently processed the full dataset, balancing accuracy and speed.
- SVM: Delivered higher accuracy but at the cost of time efficiency.

Future Prospect

- Next Step:
  - Fine Tuning, Complex Model → Better Accuracy!
  - Pack up the classifier into an APP → Better Medical Service!
  - If you want to contribute to this project, please fill out the following Questionnaire: https://forms.gle/PCvtiMzJW1nq3Nce8
  - (Your data will be collected anonymously, and you will receive detailed prediction result!)
- Feedbacks & Questions goes to bwang55@u.rochester.edu

Reference

- [1] H. J. Little, “Behavioral mechanisms underlying the link between smoking and drinking,” Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6709747/ (accessed Dec. 8, 2023).
- [2] A. Mandil et al., “Smoking among university students: A gender analysis,” Journal of Infection and Public Health, vol. 3, no. 4, pp. 179–187, 2010. doi:10.1016/j.jiph.2010.10.003
- [3] M. S. Abirami, B. Vennila, E. L. Chilukalapalli, and R. Kuriyedath, “Retracted article: A classification model to predict onset of smoking and drinking habits based on socio-economic and sociocultural factors,” Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 3, pp. 4171–4179, 2020. doi:10.1007/s12652-020-01796-4
- [4] G. Badicu, S. H. Zamani Sani, and Z. Fathirezaie, “Predicting tobacco and alcohol consumption based on physical activity level and demographic characteristics in Romanian students,” Children, vol. 7, no. 7, p. 71, 2020. doi:10.3390/children7070071
- [5] X. Dai et al., “Health effects associated with smoking: A burden of proof study,” Nature Medicine, vol. 28, no. 10, pp. 2045–2055, 2022. doi:10.1038/s41591-022-01978-x
- [6] J. Tan et al., “Smoking, blood pressure, and cardiovascular disease mortality in a large cohort of Chinese men with 15 years follow-up,” International Journal of Environmental Research and Public Health, vol. 15, no. 5, p. 1026, 2018. doi:10.3390/ijerph15051026
- [7] C. Stanley, “How smoking and drinking affect the body,” MEH, https://www.mountelizabeth.com.sg/health-plus/article/how-smoking-and-drinking-affects-the-body (accessed Dec. 8, 2023).
- [8] K. J. Mukamal, “The effects of smoking and drinking on cardiovascular disease and risk factors,” Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6527044/ (accessed Dec. 8, 2023).
- [9] C. Li and J. Sun, “The impact of current smoking, regular drinking, and physical inactivity on health care-seeking behavior in China,” BMC Health Services Research, vol. 22, no. 1, 2022. doi:10.1186/s12913-022-07462-z
- [10] Soo.Y, “Smoking and drinking dataset with body signal,” Kaggle, https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset (accessed Dec. 8, 2023).