Data Mining, ITM, SeoulTech, 2025 Spring

Case Study: Obesity

20102105 Kim Seungjun
21102052 Lee Jeong-Yun
21102061 Hwang Hyunmin

Contents

1. Background
   - Problems
2. Case 1
   - Dataset
   - Methods
   - Results
   - Limitations
3. Case 2
   - Dataset
   - Methods
   - Results
   - Limitations
4. References
   - Citations

Background

Problems

Modern lifestyles shaped by industrialization and urbanization have led to reduced physical activity and unhealthy dietary patterns.
This has contributed to a global rise in obesity and related mental health issues.

The COVID-19 pandemic intensified isolation, stress, and depression.
These psychological effects have been linked to increased BMI, alcohol use, and unhealthy behaviors.

Obesity is not caused by a single factor, but by interacting variables: socioeconomic status, mental health, sleep, diet, and physical inactivity.

Traditional models are limited in capturing this complexity.
Data mining enables the discovery of hidden patterns and the prediction of high-risk groups.

Case 1

Exploring Factors for Predicting Depression Experiences Beyond BMI Overweight Using Data Mining Models in the Post-COVID-19 Era

by Kyung-Il Hwang

Case 1 (Hwang, 2024)

Dataset

Korea National Health and Nutrition Examination Survey (KNHANES), Phase 7 (2018) & Phase 8 (2021)

Preprocessing

Variable transformation and normalization, handling missing values
SMOTE for class imbalance
70:30 split for training/evaluation datasets

Population

Initial Subjects: 15,082
→ Adults aged 19–64 = 11,478
→ Overweight adults = 6,052
Final Sample: 2,216 respondents with valid depression experience responses

Variables Summary

Demographics: Gender, Age, Employment status, Education level, Marital status, BMI
Health Behavior: Sleep duration, Walking activity, Smoking/drinking experience
Psychosocial: Self-rated health, Stress awareness, Social isolation

Variable Significance Testing

Case 1 (Hwang, 2024)

- Before building the models, statistical tests were conducted to evaluate the relationship between each variable and depression status.
- Categorical variables were tested using Chi-square (χ^2) and Fisher's Exact Test.
- Continuous variables were tested using the Mann-Whitney U Test.
- The significance level was set at $p < 0.05$.

→ Statistically significant variables (e.g., gender, employment status, education level, BMI) helped identify key patterns, but all variables were included in the machine learning models to capture potential interactions.

Variable Significance Testing

Case 1 (Hwang, 2024)

<표 1> COVID-19 발생 전·후 신체적 요인 변화

구 분		COVID-19 발생 전 N	%	발생 후 N	%	X^2/Z	P
신장(cm)		165.13	8.88	165.51	8.97	-1.778	0.075
체중(kg)		65.39	13.16	66.32	13.89	-2.703	0.007
허리둘레(cm)		81.21	10.43	82.87	11.11	-6.792	0.000
체질량지수(BMI)		23.86	3.68	24.07	3.86	-2.540	0.011
체질량지수 범주 (BMI)	18.5 미만(저체중)	232	4.8	241	5.9		
	18.5~22.9(정상)	2,099	43.4	1,657	40.3		
	23.0~24.9(과체중)	1,048	21.7	887	21.5	17.700	0.001
	25.0~29.9(비만)	1,141	23.6	997	24.2		
	30.0 이상(고도비만)	316	6.5	332	8.1		
우울감 경험†	예	-	-	242	10.9		
	아니요	-	-	1,974	89.1		

N : frequency, % : percentage, P-value <0.05, X^2 : Chi-square test,

Z : mann-whitney test, † 2021년만 조사된 문항

<표 2> BMI 과체중 이상 우울감 경험 유무와 일반적 특성

구분		BMI 과체중 이상 우울감 경험 예 N	%	아니요 N	%	X^2/Z	p
성별	남자	105	8.5	1,124	91.5	16.027	<0.001
	여자	137	13.9	850	86.1		
나이		46.36	12.70	45.88	12.33	-0.703	0.482
경제활동 상태	예(취업자)	158	10.1	1,399	89.9	20.354	<0.001
	아니요(실업자, 비경제활동인구)	119	18.1	540	81.9		
교육수준	≤ Middle school	69	21.0	259	79.0	9.893	0.002
	> Middle or graduate school	208	11.0	1,680	89.0		
결혼 여부	기혼	174	10.3	1,513	89.7	2.671	0.102
	미혼	68	12.9	461	87.1		
결혼상태	유배우자, 동거	133	8.9	1,357	91.1	24.480	<0.001
	유배우자, 별거	1	8.3	11	91.7		
	사별	9	22.0	32	78.0		
	이혼	31	21.5	113	78.5		
평생 음주 경험	없음	16	14.2	97	85.8	1.284	0.257
	있음	226	10.7	1,877	89.3		
평생 일반담배 흡연 여부	5갑(100개비) 미만	6	9.5	57	90.5	0.469	0.791
	5갑(100개비) 이상	107	10.6	907	89.4		
	피운 적 없음	129	11.3	1,010	88.7		
신장(cm)		164.95	9.67	167.03	9.31	-3.192	0.001

구분		BMI 과체중 이상 우울감 경험 예 N	%	아니요 N	%	X^2/Z	p
체중(kg)		72.14	13.02	74.37	12.12	-3.328	<0.001
허리둘레(cm)		89.25	8.02	89.68	8.65	-0.622	0.534
체질량지수(BMI)		26.38	3.04	26.57	3.05	-0.747	0.455
체질량지수 범주	23.0~24.9(과체중)	90	10.8	740	89.2	1.835	0.399
	25.0~29.9(비만)	128	11.5	982	88.5		
	30.0 이상(고도비만)	24	8.7	252	91.3		
1주일간 걷기일 수	전혀 하지 않음	43	12.0	314	88.0	6.833	0.446
	1일	10	7.1	130	92.9		
	2일	20	7.6	243	92.4		
	3일	38	13.4	245	86.6		
	4일	19	10.6	160	89.4		
	5일	30	11.2	237	88.8		
	6일	16	11.9	119	88.1		
	7일(매일)	66	11.1	526	88.9		
1주일간 근력운동일 수	전혀 하지 않음	172	10.7	1,442	89.3	4.138	0.530
	1일	8	9.4	77	90.6		
	2일	9	7.8	107	92.2		
	3일	13	10.0	117	90.0		
	4일	9	14.3	54	85.7		
	5일 이상	29	13.9	179	86.1		
유산소 신체활동 실천	실천하지 않음	124	10.5	1,055	89.5	0.288	0.591
	실천	117	11.3	920	88.7		

N : frequency, % : percentage, P-value <0.05, X^2 : Chi-square test,

Z : mann-whitney test

Variable Significance Testing

Case 1 (Hwang, 2024)

<표 3> BMI 과체중 우울감 경험과 주관적 인식과 신체 요인

구분		BMI 과체중 이상 우울감 경험				X^2	p
		예		아니요			
		N	%	N	%		
주관적 건강인지	매우 나쁨	20	8.3	28	1.4	105.519	<0.001
	나쁨	71	29.3	274	13.9		
	보통	113	46.7	983	49.8		
	좋음	31	12.8	586	29.7		
	매우 좋음	7	2.9	103	5.2		
주관적 체형인식	매우 마른 편	1	0.4	0	0.0	11.004	0.023
	약간 마른 편	1	0.4	14	0.7		
	보통	45	18.6	469	23.8		
	약간 비만	128	52.9	1,065	54.0		
	매우 비만	67	27.7	426	21.6		
1년간 체중 변화 여부	변화 없음	79	32.6	977	49.5	25.232	<0.001
	체중 감소	43	17.8	236	12.0		
	체중 증가	120	49.6	761	38.6		
1년간 체중 감소량	3kg 이상 ~ 6kg 미만	24	55.8	162	68.6	10.196	0.006
	6kg 이상 ~ 10kg 미만	7	16.3	50	21.2		
	10kg 이상	12	27.9	24	10.2		
1년간 체중 증가량	3kg 이상 ~ 6kg 미만	74	61.7	517	67.9	13.111	0.001
	6kg 이상 ~ 10kg 미만	21	17.5	170	22.3		
	10kg 이상	25	20.8	74	9.7		
1년간 체중 조절 여부	체중 감소 노력	149	61.6	1,095	55.5	3.697	0.296
	체중 유지 노력	40	16.5	376	19.0		
	체중증가 노력	1	0.4	19	1.0		
	체중 조절 노력해본 적 없음	52	21.5	484	24.5		
평소 스트레스 인지 정도	대단히 많이 느낀다	8	3.3	261	13.2	250.509	<0.001
	많이 느끼는 편이다	70	28.9	1,237	62.7		
	조금 느끼는 편이다	111	45.9	407	20.6		
	거의 느끼지 않는다	53	21.9	69	3.5		

N : frequency, % : percentage, P-value <0.05, X^2 : Chi-square test, Z : mann-whitney test

<표 4> BMI 과체중 이상 우울감 경험과 주요 질환

구분		BMI 과체중 이상 우울감 경험				X^2	p
		예		아니요			
		N	%	N	%		
고혈압	없음	181	74.8	1,571	79.6	2.989	0.084
	있음	61	25.2	403	20.4		
이상지질혈증	없음	174	71.9	1,562	79.1	6.637	0.010
	있음	68	28.1	412	20.9		
뇌졸중	없음	248	90.2	1,900	97.9	2.497	0.173
	있음	27	9.8	41	2.1		
심근경색증 또는 협심증	없음	246	89.5	1,892	97.5	3.382	0.087
	있음	29	10.5	49	2.5		
당뇨병	없음	216	89.3	1,810	91.7	1.632	0.201
	있음	26	10.7	164	8.3		
빈혈	없음	223	86.4	1,833	93.6	6.492	0.011
	있음	35	13.6	125	6.4		

N : frequency, % : percentage, P-value <0.05, X^2 : Chi-square test, Z : mann-whitney test

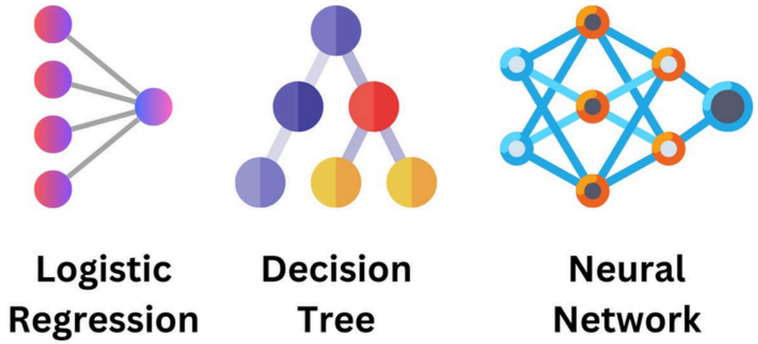
Methods

Case 1 (Hwang, 2024)

- Logistic Regression:
  - A statistical model that predicts the probability of depression based on a weighted combination of input variables.

- Neural Network:
  - A structure of interconnected layers (input → hidden → output) that processes information by passing signals through nodes.
  - It learns patterns in the data by adjusting the weights in the network during training.

- CHAID Decision Tree:
  - Similar to general decision trees, but CHAID splits based on chi-square test results and can produce more than two branches at once, making it suitable for analyzing categorical data and visualizing multiple outcome paths.



Results

Case 1 (Hwang, 2024)

<표 6> 로지스틱 회귀분석(Logistic Regression) 방정식 변수 결과표

구분		B	유의확률	Exp(B)	EXP(B)에 대한 95% 신뢰구간	
					하한	상한
변수	성별(여자)	0.333	0.480	1.396	0.554	3.519
	경제활동 상태(아니요)	0.337	0.267	1.401	0.772	2.541
	교육 수준(중학교 졸업 이상)	0.243	0.523	1.275	0.605	2.686
	결혼상태		0.054			
	결혼상태(유배우자, 별거)	-0.198	0.887	0.821	0.053	12.709
	결혼상태(사별)	1.407	0.048	4.084	1.014	16.456
	결혼상태(이혼)	0.866	0.038	2.378	1.051	5.380
	신장	0.016	0.623	1.016	0.954	1.082
	체중	-0.012	0.553	0.988	0.950	1.028
	주관적 건강인지		0.052			
	주관적 건강인지(나쁨)	-0.469	0.466	0.626	0.177	2.207
	주관적 건강인지(보통)	-0.872	0.174	0.418	0.119	1.470
	주관적 건강인지(좋음)	-1.861	0.013	0.156	0.036	0.673
	주관적 건강인지(매우 좋음)	-0.888	0.386	0.411	0.055	3.062
	주관적 체형 인식		0.719			
	주관적 체형 인식(보통)	-0.289	0.543	0.749	0.295	1.903
	주관적 체형 인식(약간비만)	-0.467	0.418	0.627	0.203	1.939
	1년간 체중 증가량		0.050			
	1년간 체중 증가량(감소)	-0.672	0.091	0.511	0.234	1.114
	1년간 체중 증가량(증가)	0.588	0.188	1.800	0.750	4.322
	평소 스트레스 인지 정도		0.000			
	평소 스트레스 인지 정도(많이 느낌)	0.214	0.740	1.239	0.349	4.404
	평소 스트레스 인지 정도(조금 느낌)	1.470	0.023	4.350	1.229	15.394
	평소 스트레스 인지 정도(거의 느끼지 않음)	2.383	0.001	10.838	2.712	43.310
	이상지질혈증(있음)	0.187	0.567	1.206	0.635	2.288
	빈혈(있음)	0.417	0.365	1.517	0.616	3.738
	상수항	-3.952	0.402	0.019		

변수 : 성별, 경제활동 상태, 교육수준, 결혼상태, 신장, 체중, 주관적 건강인지, 주관적 체형인식, 1년간 체중 증가량, 평소 스트레스 인지 정도, 이상지질혈증, 빈혈 유병여부 단계 입력

According to the logistic regression results, variables with a statistically significant effect ($p < 0.05$) on depression experience among overweight individuals included:

- Marital Status (Widowed): B = 1.407, p = 0.048
- Marital Status (Divorced): B = 0.866, p = 0.038
- Self-rated Health (Good): B = −1.861, p = 0.013
- Stress Awareness (Some stress): B = 1.470, p = 0.023
- Stress Awareness (Almost none): B = 2.383, p = 0.001

→ These variables had either a positive or negative effect on the probability of experiencing depression, as reflected by the sign of the coefficient (B value). – positive B value: higher likelihood of depression / negative B value: lower likelihood

Results

Case 1 (Hwang, 2024)

<표 5> 로지스틱 회귀분석(Logistic Regression) 본 분류표

관측		예 측		
		우울감 경험		분류정확 (%)
		예	아니요	
BMI 과체중 이상 우울감 경험	예	2.4	11.1	98.6
	아니요	1.2	85.3	17.5
정확도				87.7

a. 절단값은 .500

→ The results classified using Logistic Regression analysis showed an accuracy of 87.7%.

Results

Case 1 (Hwang, 2024)

<표 7> 신경망(Neural Networks) 분석 구조

입력층	요인	1	성별
		2	경제활동 상태
		3	교육 수준
		4	결혼상태
		5	신장
		6	체중
		7	주관적 건강인지
		8	주관적 체형 인식
		9	1년간 체중 변화 여부
		10	평소 스트레스 인지 정도
		11	이상지질혈증
		12	빈혈
	노드 수a		788
은닉층	은닉층 수		1
	은닉층 1에서 노드의 수a		2
	활성화 함수		쌍곡 탄젠트
출력층	종속변수	1	BMI 과체중 이상 우울감 경험
	노드 수		2
	활성화 함수		소프트맥스
	오차 함수		교차-엔트로피

a. 편향 단위 제외

- 12 input variables
- One hidden layer with two nodes
- Uses the hyperbolic tangent activation function in the hidden layer and the softmax function in the output layer
- optimized using cross-entropy error

Results

Case 1 (Hwang, 2024)

<표 8> 신경망(Neural Networks) 분류분석 결과

구 분	예측		
	예	아니요	정확도 퍼센트(%)
예	2.6	7.7	98.3
아니요	1.5	88.2	25.0
정확도(%)			90.8

종속변수 : BMI 과체중 이상 우울감 경험

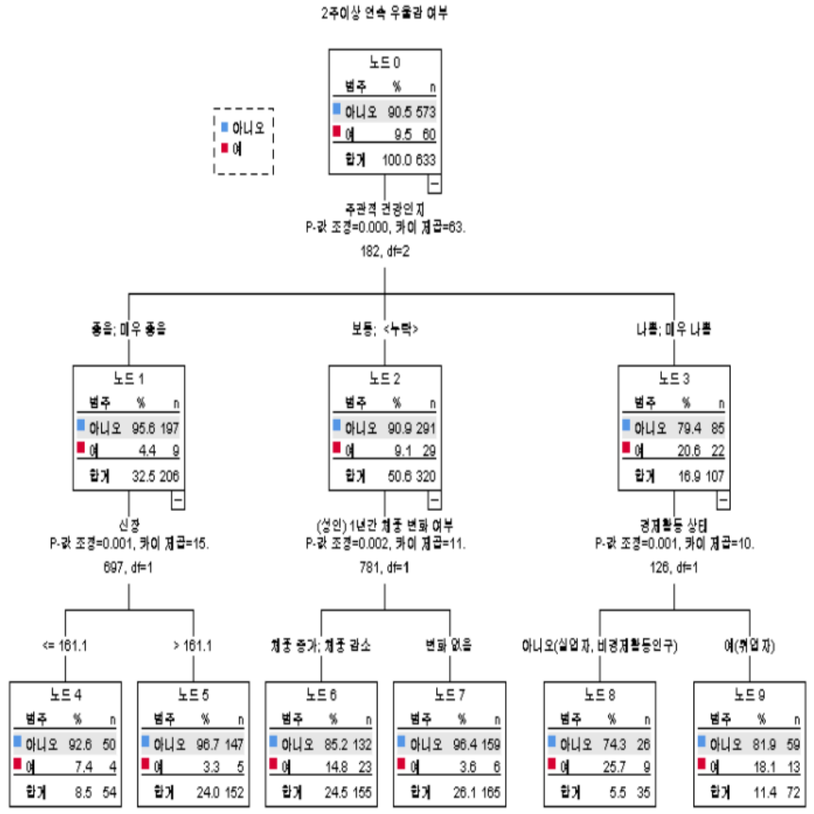
→ The results classified using Neural Networks analysis showed an accuracy of 90.8%.

Results

Case 1 (Hwang, 2024)

<표 9> BMI 과체중 이상 우울감 경험 예측 모형 요약

모형 요약		
지정 사항	성장방법	CHAID
	종속변수	우울감 경험
	독립변수	성별, 경제활동 상태, 교육 수준, 결혼상태, 신장, 체중, 주관적 건강인지, 주관적 체형 인식, 1년간 체중 변화 여부, 1년간 체중 감소량, 1년간 체중 증가량, 이상지질혈증, 빈혈
	검증	분할 표본
	최대 나무 깊이	3
	부모 노드의 최소 케이스	100
	자식 노드의 최소 케이스	50
결과	독립변수 포함	주관적 건강인지, 신장, 1년간 체중 변화 여부, 경제활동 상태
	노드 수	10
	터미널 노드 수	6
	깊이	2



<그림 4> BMI 과체중 이상 우울감 경험 의사결정트리(Decision Tree)

- maximum tree depth: 3, parent node size: 100, minimum child node size: 50
- depth: 2
- total nodes: 10 , terminal nodes: 6
- key variables: self-rated health, employment status, weight change over 1 year, height

Results

Case 1 (Hwang, 2024)

<표 10> BMI 과체중 이상 우울감 경험 예측 모형 분류 요약

구 분	예측		
	예	아니요	정확도(%)
예	90.5	0	100.0
아니요	9.2	0	0.0
정확도(%)			90.5

성장 방법 : CHAID
종속변수 : BMI 과체중 이상 우울감 경험

→ The results classified using CHAID Decision Tree analysis showed an accuracy of 90.5%.

Results

Case 1 (Hwang, 2024)

<표 11> 분석방법 및 정확도 비교

구 분	모형 분석결과		
	신경망 (Neural Network)	의사결정나무 (Decision Tree)	로지스틱 회귀분석 (Logistic Regression)
정확도	90.8%	90.5%	87.7%

- ✅ Neural Network achieved the highest accuracy; Decision Tree was best in interpretability, although Logistic Regression showed decent accuracy.
- ✅ Social isolation and self-rated health consistently emerged as key predictors.
- ✅ Psychosocial effects of obesity may outweigh the role of BMI itself.
- ✅ This study supports the use of basic public health survey data in mental health prediction.

Limitations

Cross-sectional data limits causal inference.

Self-reported survey responses may have bias.

Psychosocial variables like isolation/stress are hard to quantify precisely.

Case 2

Predicting Overweight and Obesity Status Among Malaysian Working Adults With Machine Learning or Logistic Regression: Retrospective Comparison Study

by Jyh Eiin Wong

Dataset

Case 2 (Wong et al., 2022)

Malaysia's Healthiest Workplace Survey 2019.

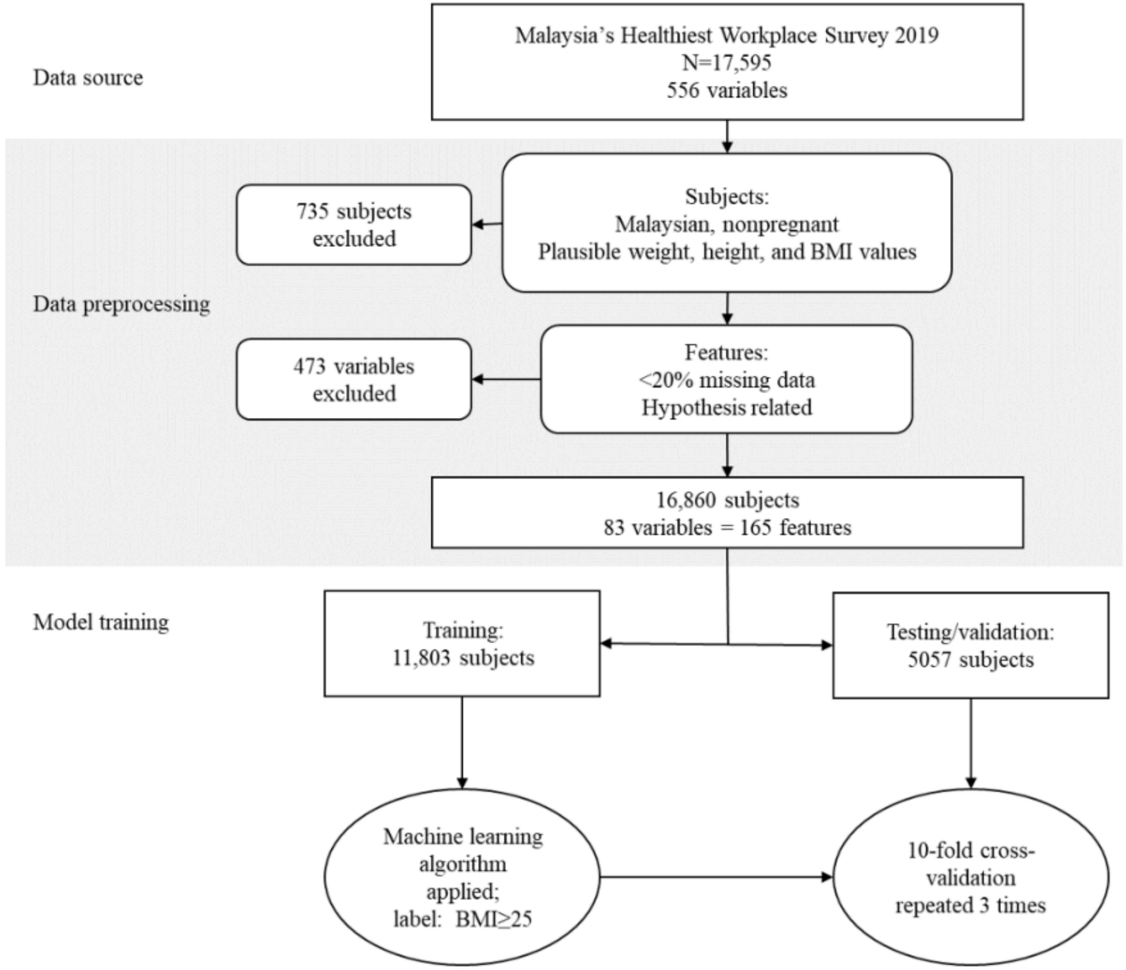
Preprocessing
735 subjects excluded (non–Malaysian, pregnant, or implausible BMI values)
473 variables excluded (irrelevant or >20% missing)
Categorical variables into binary format using one–hot encoding.

Population
16,860 working adults in Malaysia

Data Partitioning
70% (11,803 individuals) used for training
30% (5,057 individuals) used for validation

Overweight / Obesity Proportion
4,934 individuals (41.8%) in the training set were overweight or obese

Model Input
Total of 165 normalized features
Includes 120 binary variables



Methods

Case 2 (Wong et al., 2022)

Machine Learning

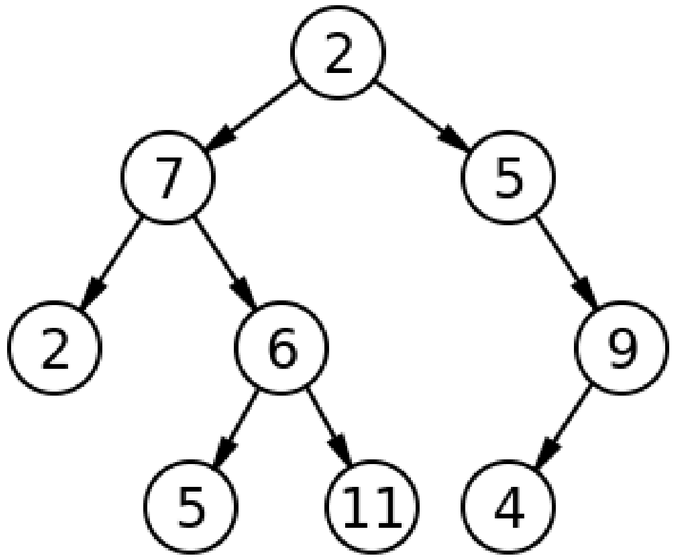
- XGBoost
- Random Forest(RF)
- SVM

Data Mining

- Logistic Regression

Methods

Case 2 (Wong et al., 2022)



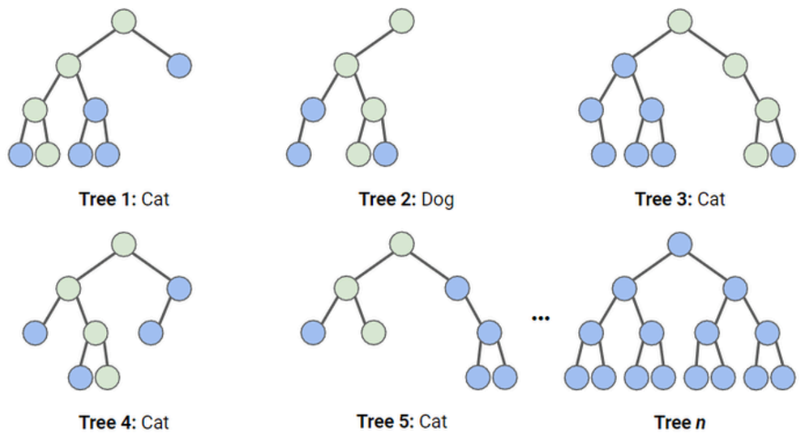
XGBoost (Extreme Gradient Boosting)

XGBoost builds decision trees sequentially to reduce prediction errors (residuals) step by step.

It combines gradient-based optimization, regularization, and parallel processing to create a highly accurate and robust model that is resistant to overfitting.

Methods

Case 2 (Wong et al., 2022)



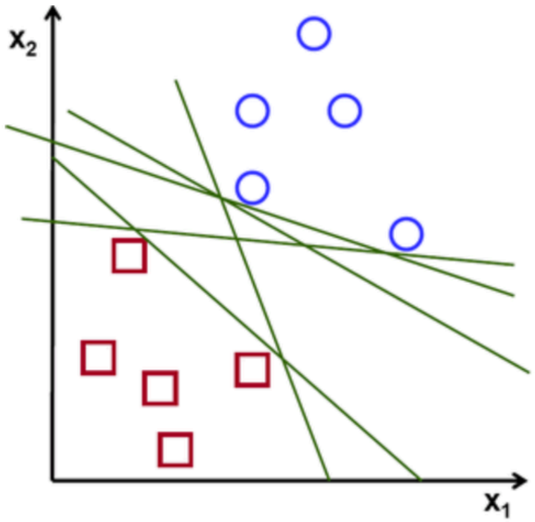
Random Forest

An ensemble learning method that builds multiple decision trees and aggregates their results to make more accurate and robust predictions.

Each tree is trained on a random sample of the data (bootstrapping).
At each split, a random subset of features is considered.
Final prediction = majority vote (classification) or average (regression)

Methods

Case 2 (Wong et al., 2022)



SVM (Support Vector Machine)

A supervised learning algorithm that finds the optimal boundary (hyperplane) to separate different classes with the maximum margin.

Identifies support vectors, the most critical data points near the margin

Maximizes the distance (margin) between classes

Can handle non-linear boundaries using kernel functions

Methods

Case 2 (Wong et al., 2022)

Table 1. Performance of machine-learning algorithms and logistic regression in obesity prediction.

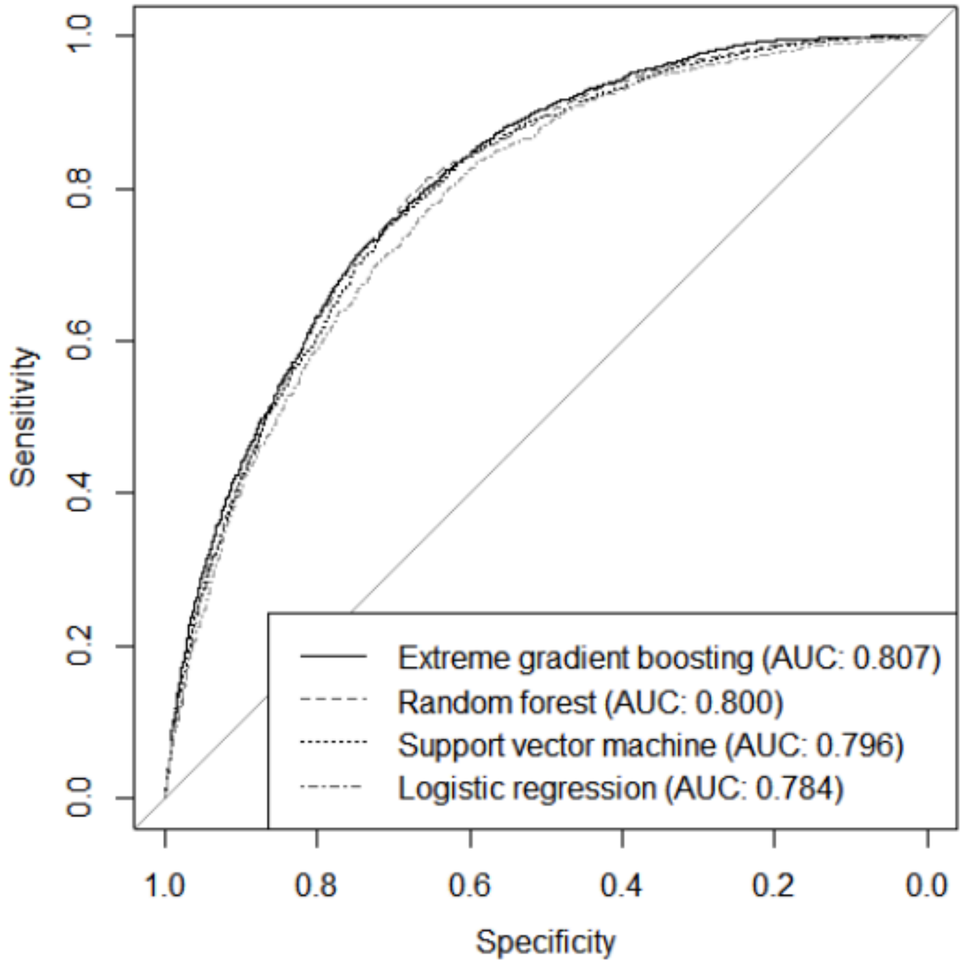
Metrics	Gradient boosting, mean (95% CI)	Random forest, mean (95% CI)	Support vector machine, mean (95% CI)	Logistic regression, mean (95% CI)
Accuracy[a]	0.73 (0.72-0.75)	0.73 (0.71-0.74)	0.72 (0.71-0.73)	0.71 (0.70-0.72)
Sensitivity[a]	0.67 (0.65-0.69)	0.60 (0.58-0.62)	0.65 (0.62-0.67)	0.56 (0.54-0.58)
Specificity[a]	0.78 (0.76-0.79)	0.82 (0.80-0.83)	0.77 (0.76-0.79)	0.82 (0.81-0.83)
Area under the curve[b]	0.81 (0.79-0.82)	0.80 (0.79-0.81)	0.80 (0.78-0.81)	0.78 (0.77-0.80)

XGBoost showed the best overall performance, especially in identifying obese individuals (high sensitivity and AUC).
While Logistic Regression is simpler, it tends to miss more obese individuals.
However, all models had AUCs above 0.78, indicating they are all practically usable for real-world screening.

Methods

Case 2 (Wong et al., 2022)

Figure 2. Receiver operating characteristic curves with corresponding AUC values; AUC values for each model are also presented in Table 2. AUC: area under the curve.



ROC(Receiver Operating Characteristic) Curve

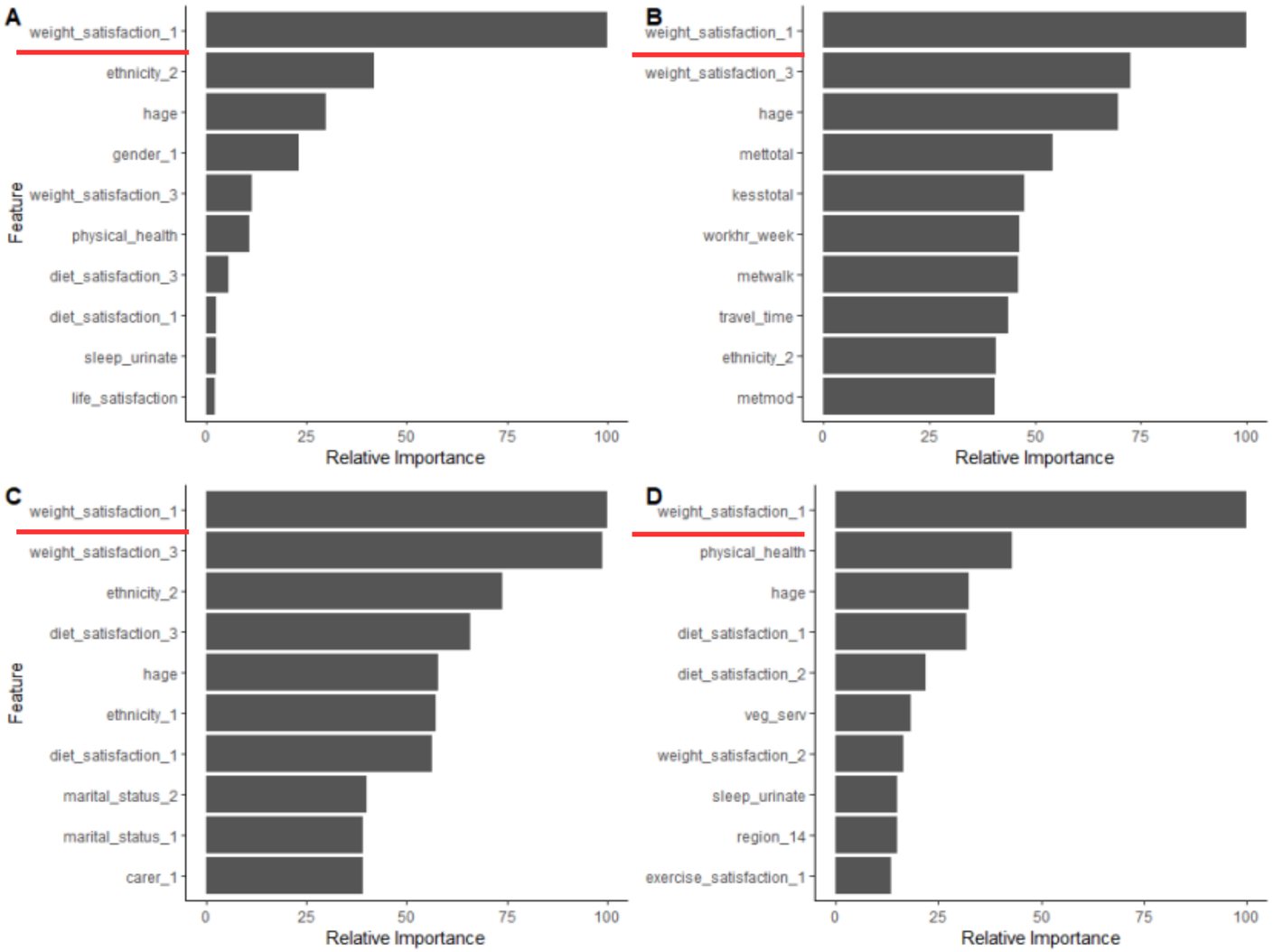
A graphical representation used to evaluate the performance of classification models

- The diagonal line represents a model that makes random guesses (AUC = 0.5). The further the curve is from the diagonal and toward the top-left corner, the better the model.
- XGBoost the highest, Logistic regression the lowest. Still they are all good models.

Methods

Case 2 (Wong et al., 2022)

Figure 3. Variable importance plots of obesity predictors for extreme gradient boosting (A), random forest (B), support vector machine (C) and logistic regression (D) models. The top 10 predictors are shown for all models.



A person's subjective perception of their weight (weight satisfaction) was the most powerful predictor of obesity status.

Those who are dissatisfied with their body weight were more likely to be overweight or obese.

Results

Case 2 (Wong et al., 2022)

- ✅ In conclusion, among the predictors, weight satisfaction was the most influential factor. This suggests that obesity interventions should not only focus on physical metrics, but also include psychological and perceptual components — such as body image, self-esteem, and health awareness — to promote sustainable behavior change.
- ✅ In the modeling perspective, among the three machine learning models (XGBoost, Random Forest, and SVM), XGBoost showed the highest accuracy and AUC.
- ✅ However, the logistic regression model was comparable to the overall performance of the ML models or even better in some parts.
- ✅ This suggests that even simple, interpretable models like logistic regression can perform competitively, and may be preferable when interpretability and practical implementation are important.

Limitations

Case 2 (Wong et al., 2022)

Self–reported weight and height may have led to misclassification and underestimation of obesity prevalence.

No external validation: Models were tested on the same dataset, which may limit generalizability.

High proportion of binary variables (73%) may have reduced the advantage of nonlinear ML algorithms.

References

Citations

📒 **Case 1:**

Hwang, K.-I. (2024). Exploring factors for predicting depression experiences beyond BMI overweight using data mining models in the post-COVID-19 era (Doctoral dissertation). Korea University, Department of Health Policy and Management. https://doi.org/10.23186/korea.000000277951.11009.0000395

📒 **Case 2:**

Wong, J., Yamaguchi, M., Nishi, N., Araki, M., & Wee, L. (2022). Predicting overweight and obesity status among Malaysian working adults with machine learning or logistic regression: Retrospective comparison study. JMIR Formative Research, 6(12), e40404. https://doi.org/10.2196/40404

thank you!

Data Mining, ITM, SeoulTech, 2025 Spring

20102105 Kim Seungjun
21102052 Lee Jeong-Yun
21102061 Hwang Hyunmin