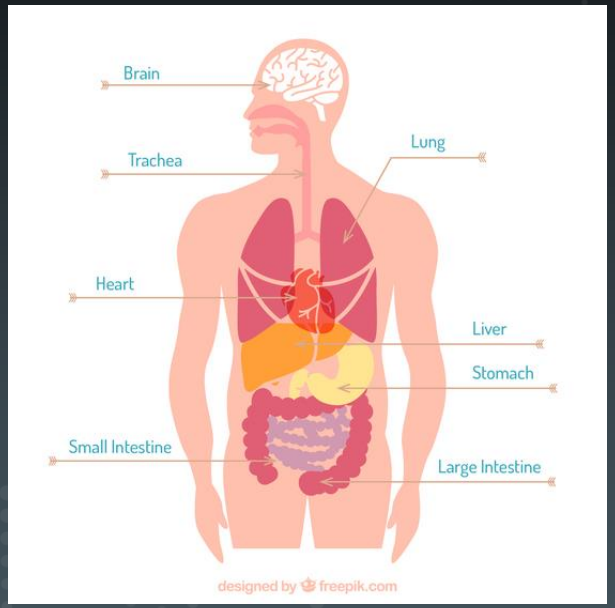
KOREA 2020

DISCOVERY SUMMIT ONLINE

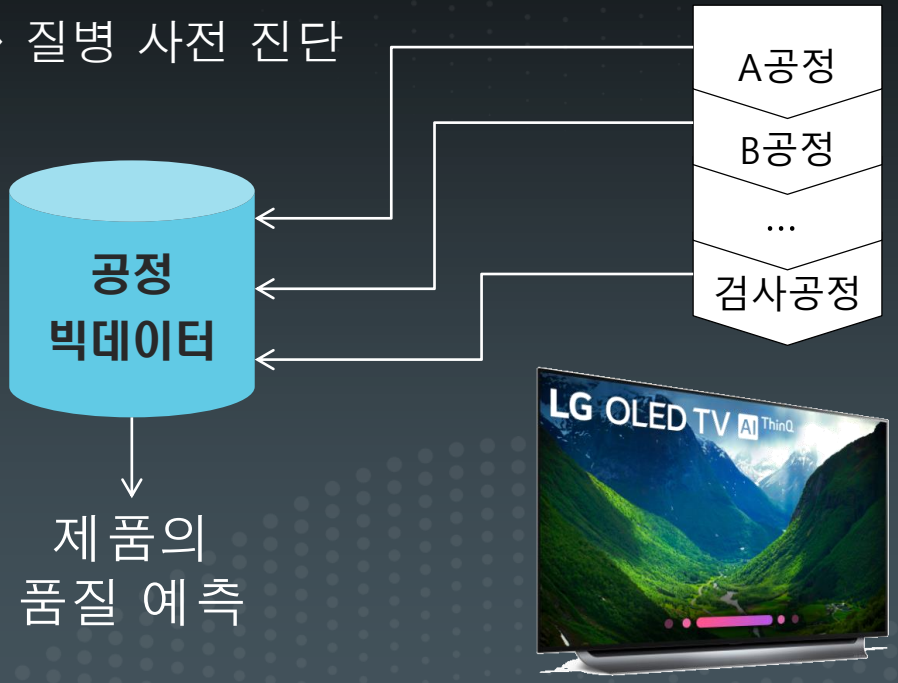
jmp

Various Statistical Analysis for Quality Prediction

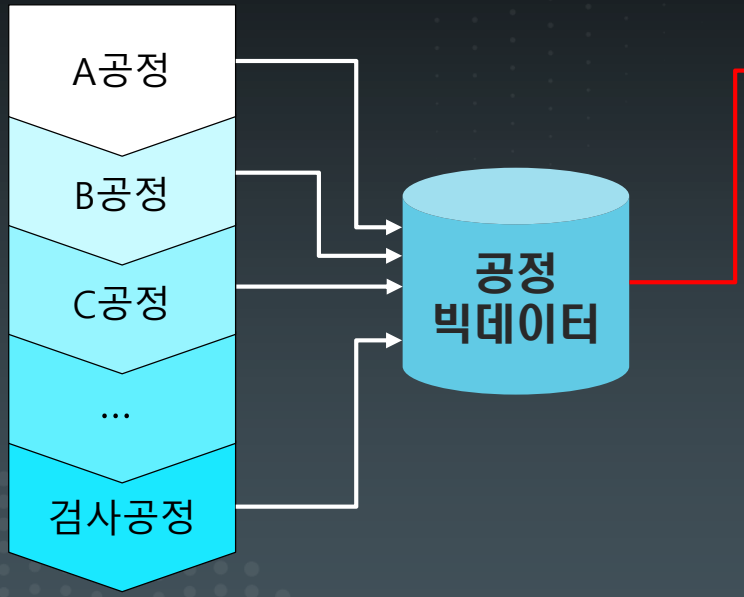
과제의 이미지



종합검진 > 질병 사전 진단



수집 데이터의 특징

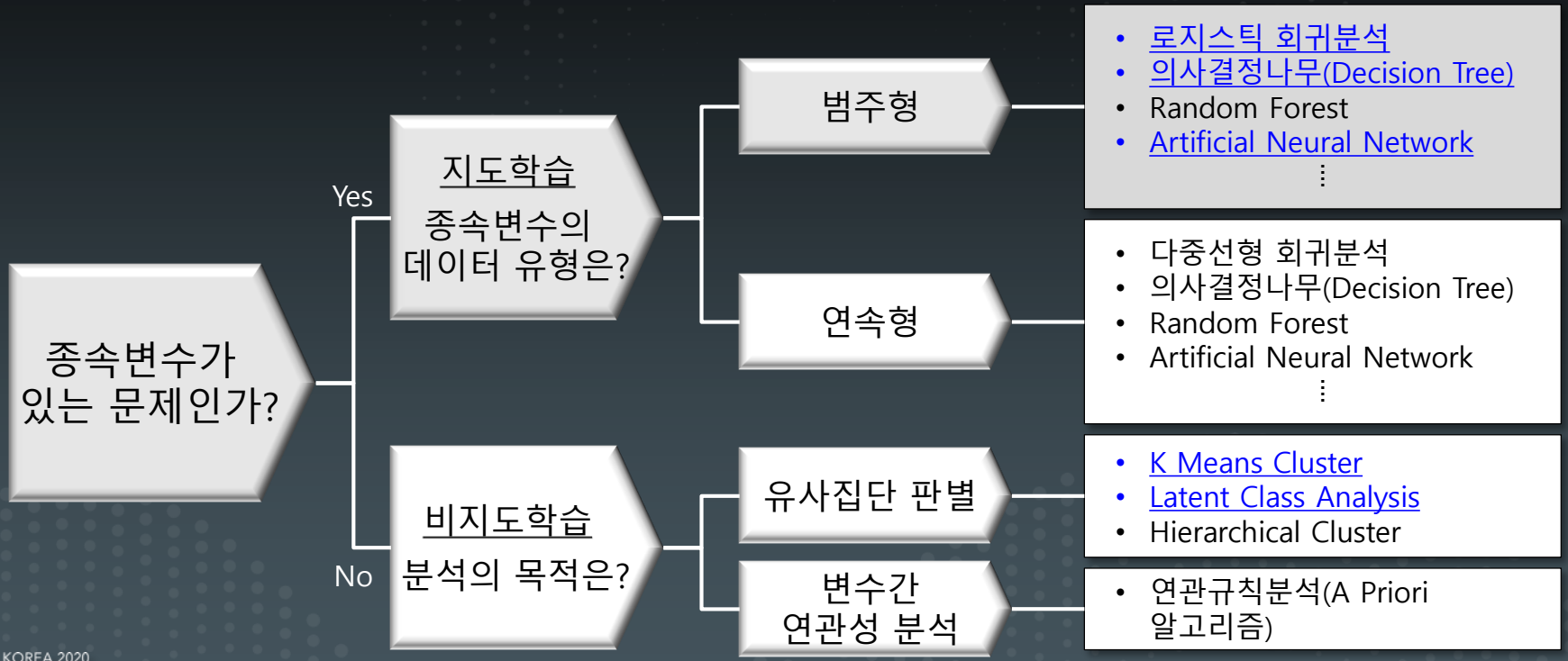


JMP Data Table

X1	X2	X3	…	X12	Y
26	34	0	…	0	OK
15	531	1	…	0	NG
30	20	0	…	1	OK
…	…	…	⋱	…	…
26	75	0	…	1	OK

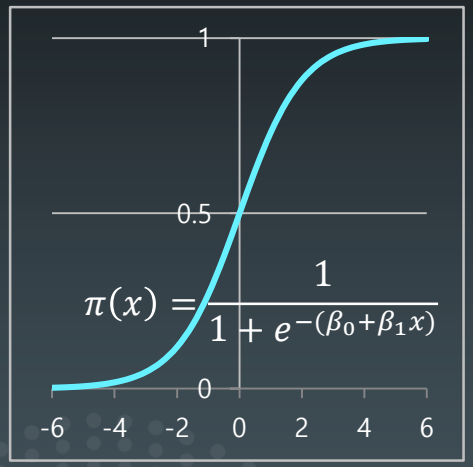
- X1, X5, X9, Y → Nominal Data
- X2, X6, X10 → Continuous Data
- X3, X4, X7, X8, X11, X12 → Ordinal Data

데이터 분석 계획



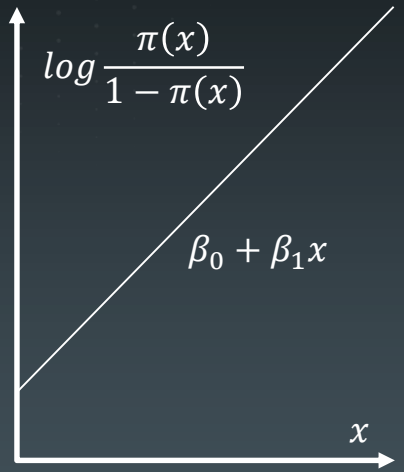
로지스틱 회귀분석

Sigmoid Function



- Sigmoid Function
- Large input→Small output(0~1)

Logistic Regression



- $Y = \beta_0 + \beta_1 x$ 형태의 회귀모형 도출

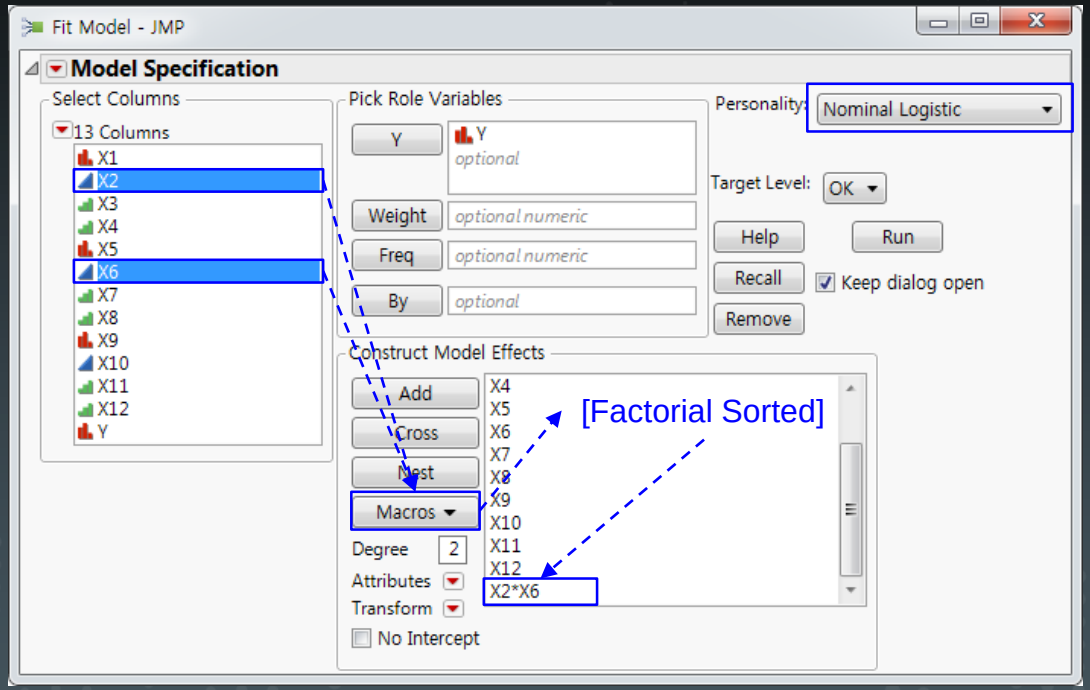
- 영국의 통계학자인 D. R. Cox가 1958년에 제안한 확률 모델로서 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 통계 기법
- 승산(odds)이란 p가 성공 확률일 때, 성공 대비 실패 확률 비율인데 확률 p(0~1)를 Sigmoid 함수로 대체함

$$odds = \frac{p}{1-p} = \frac{\pi(x)}{1-\pi(x)}$$

- Logit 함수는 승산(odds)에 로그를 취한 함수로서 입력 값 범위가 [0,1]일 때 출력 값 범위는 $[-\infty, +\infty]$

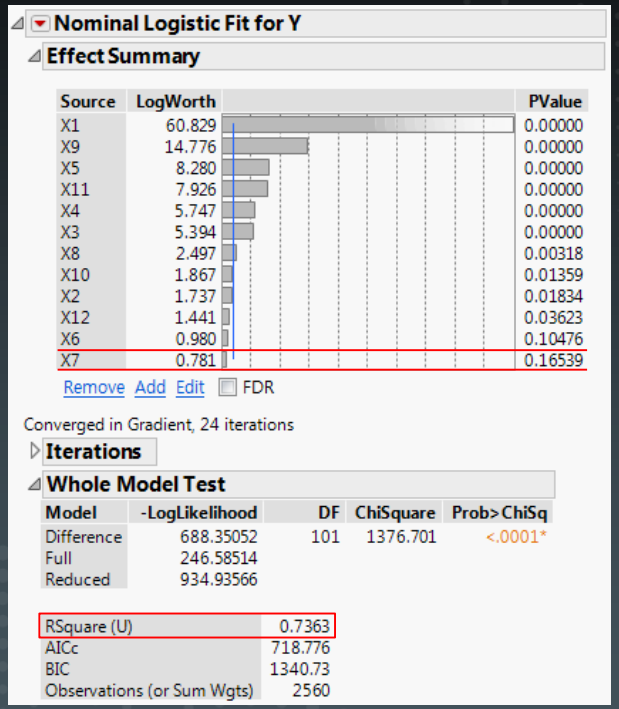
$$logit = log\frac{p}{1-p} = log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x$$

로지스틱 회귀분석



- Row Selection 메뉴 이용
  Training data: 80%
  Validation Data: 20%
- JMP: Analyze > Fit Model
  Y가 범주형 데이터이므로
  → Nominal Logistic
- [Macro]
  → 원하는 교호작용 항 생성

로지스틱 회귀분석 결과



Nominal Logistic Fit for Y

Effect Summary

Source	LogWorth	PValue
X1	60.829	0.00000
X9	14.776	0.00000
X5	8.280	0.00000
X11	7.926	0.00000
X4	5.747	0.00000
X3	5.394	0.00000
X8	2.497	0.00318
X10	1.867	0.01359
X2	1.737	0.01834
X12	1.441	0.03623
X6	0.980	0.10476
X7	0.781	0.16539

Remove Add Edit FDR

Converged in Gradient, 24 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	688.35052	101	1376.701	<.0001*
Full	246.58514			
Reduced	934.93566			

RSquare (U)	0.7363
AICc	718.776
BIC	1340.73
Observations (or Sum Wgts)	2560

Training Data

Confusion Matrix		예측	
		NG	OK
실제	NG	247	58
	OK	40	2215

Confusion Rates		예측	
		NG	OK
실제	NG	0.810	0.190
	OK	0.018	0.982

정분류율(accuracy)*: 96.2%
제2종 오류**: 19.0%

Validation Data

Confusion Matrix		예측	
		NG	OK
실제	NG	49	16
	OK	14	555

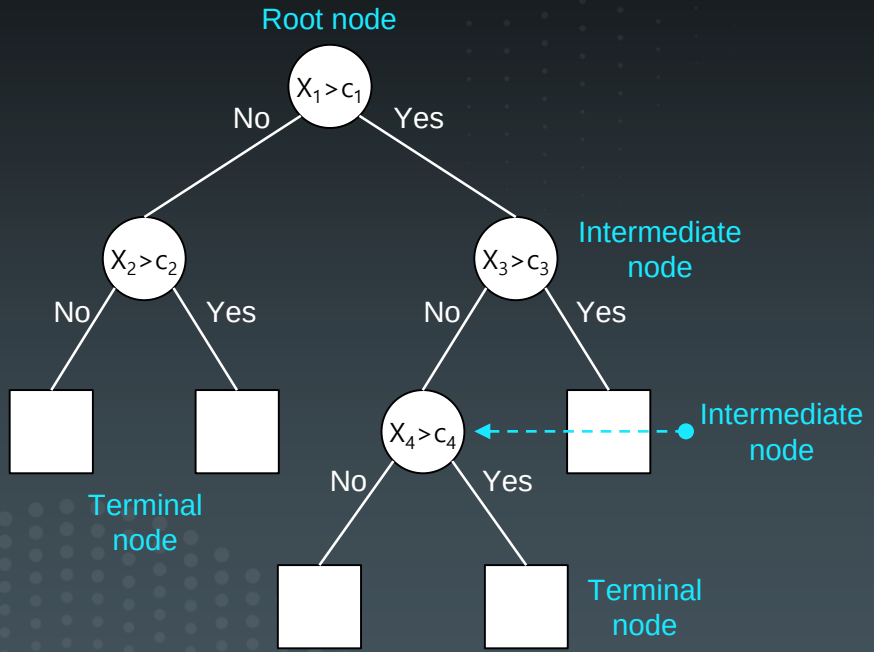
Confusion Rates		예측	
		NG	OK
실제	NG	0.754	0.246
	OK	0.025	0.975

정분류율(accuracy): 95.3%
제2종 오류: 24.6%

* 정분류율(accuracy): 'OK', 'NG'를 제대로 예측한 비율 / **제2종 오류: 실제 'NG' 중에서 'OK'로 오판한 비율

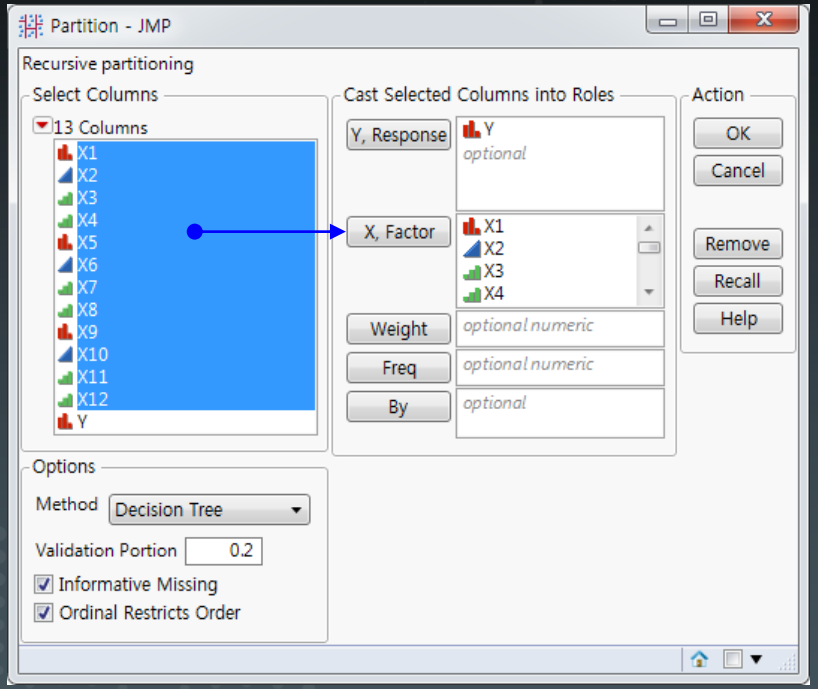
KOREA 2020 DISCOVERY SUMMIT ONLINE

Decision Tree



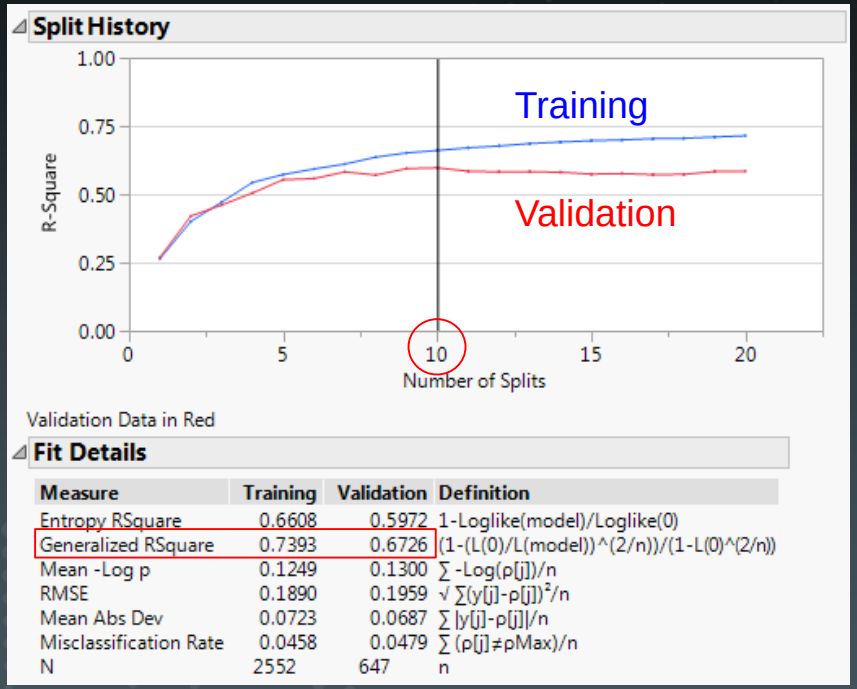
- 데이터를 규칙적 형태로 나누는 노드를 구성하고 그 노드에 속하는 학습 데이터의 반응변수 구성비를 이용하여 분류와 예측 모형을 구축하는 통계적 방법
- 데이터 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 나타내어 반응변수의 분류 또는 예측에 활용함
- X변수의 기준에 따라 비슷한 범주 혹은 값의 관측치끼리 구분하여 분류한다.
- 범주형 변수와 연속형 변수 모두 예측할 수 있어서 분류와 회귀 두 가지의 용도에 모두 사용할 수 있다.

Decision Tree 분석



- JMP: Analyze > Predictive Modeling > Partition
  [Y, Response] → Y 지정
- [Option]에서 Validation Portion 지정
  Training data: 80%
  Validation Data: 20%
- Decision Tree에서는 교호작용 항 생성 불가

Decision Tree 결과



Measure	Training	Validation	Definition
Entropy RSquare	0.6608	0.5972	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7393	0.6726	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1249	0.1300	∑ -Log(ρ[j])/n
RMSE	0.1890	0.1959	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0723	0.0687	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0458	0.0479	∑ (ρ[j]≠ρMax)/n
N	2552	647	n

Training Data

Confusion Matrix		예측	
		NG	OK
실제	NG	255	53
	OK	64	2180

Confusion Rates		예측	
		NG	OK
실제	NG	0.828	0.172
	OK	0.029	0.971

정분류율(accuracy): 95.4%
제2종 오류: 17.2%

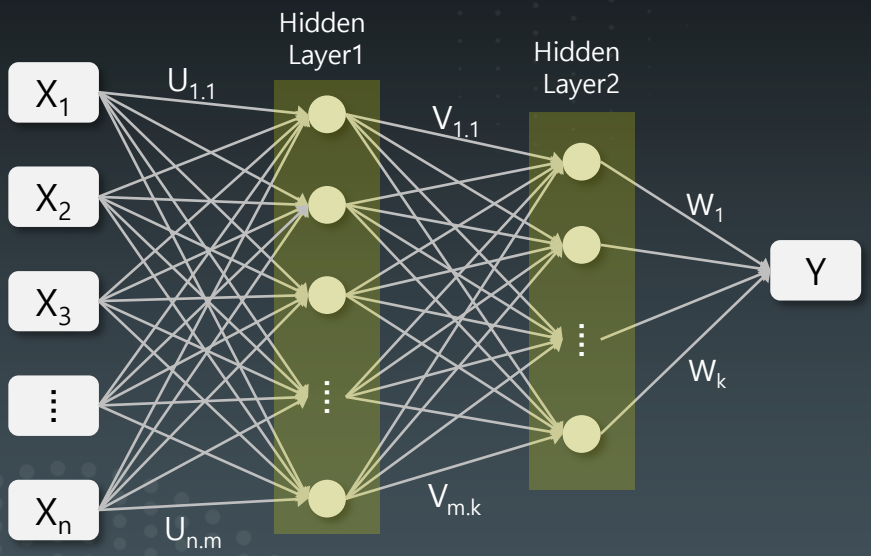
Validation Data

Confusion Matrix		예측	
		NG	OK
실제	NG	44	20
	OK	11	572

Confusion Rates		예측	
		NG	OK
실제	NG	0.688	0.313
	OK	0.019	0.981

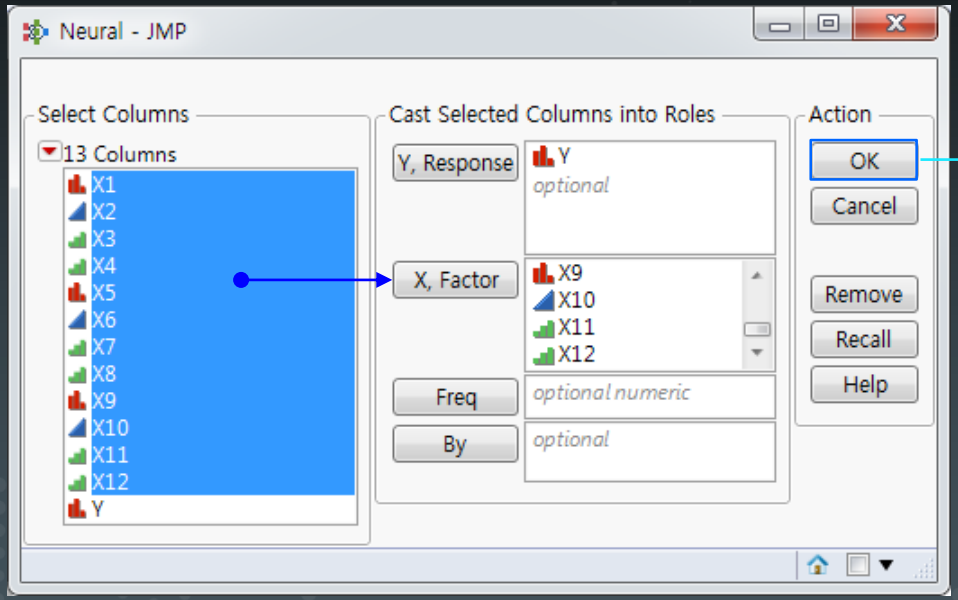
정분류율(accuracy): 95.2%
제2종 오류: 31.3%

Neural Network

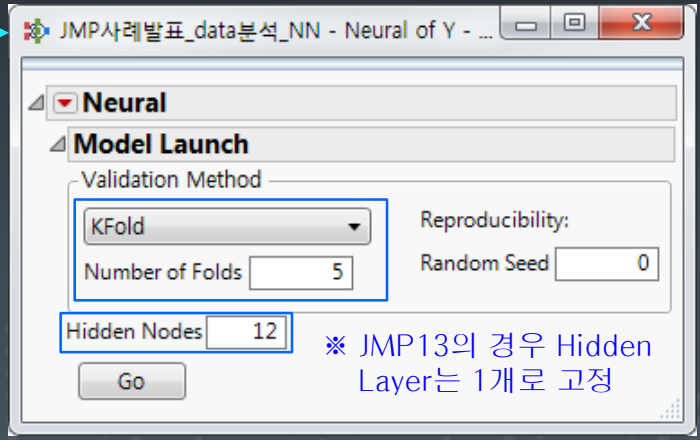


- 인간의 뇌 신경망 구조를 모방하여 X인자와 Y인자의 관계를 복잡한 네트워크 형태로 구조화하는 방법
- 인간의 뇌 신경망 구조를 차용하여 여러 인자들(X)로부터 종속변수(Y)의 관계를 구조화함
- 중간 은닉층(Hidden Layer), Node 개수가 결정되면 가중치(U, V, W)를 조절하면서 예측오차를 줄이는 모형을 찾음
- 예측력이 뛰어나지만 복잡한 구조를 해석할 수 없음(Black-box model)

Neural Network 분석

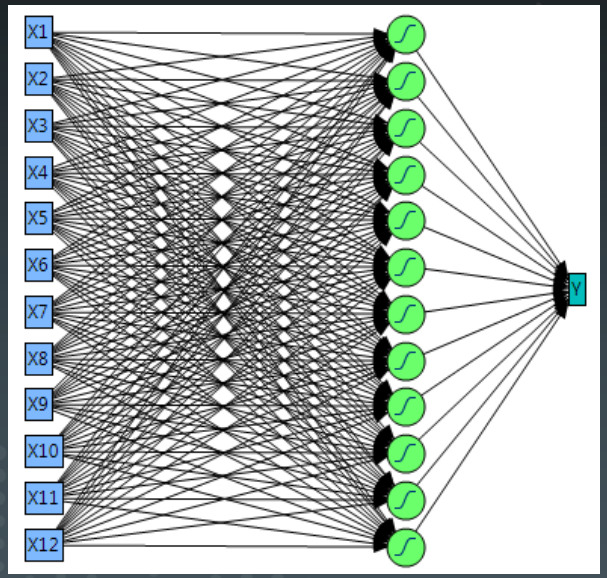


- JMP: Analyze > Predictive Modeling > Neural
- Validation Method: KFold [5]
  Training data 80%, Validation data 20%



Neural Network 결과

NN Diagram



Training Data

Measures	Value
Generalized RSquare	0.814043
RMSE	0.162429
Sum Freq	2560

Confusion Matrix		예측	
		NG	OK
실제	NG	239	59
	OK	35	2227

Confusion Matrix		예측	
		NG	OK
실제	NG	0.802	0.198
	OK	0.015	0.985

정분류율(accuracy): 96.3%
제2종 오류: 19.8%

Validation Data

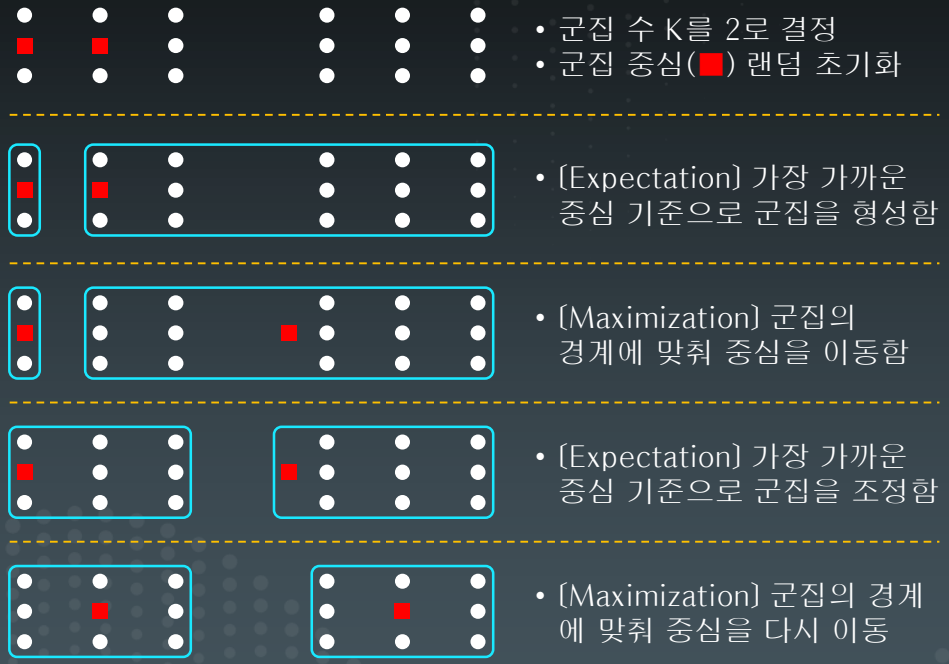
Measures	Value
Generalized RSquare	0.813215
RMSE	0.163617
Sum Freq	639

Confusion Matrix		예측	
		NG	OK
실제	NG	62	12
	OK	13	552

Confusion Matrix		예측	
		NG	OK
실제	NG	0.838	0.162
	OK	0.023	0.977

정분류율(accuracy): 96.1%
제2종 오류: 16.2%

K Means Cluster

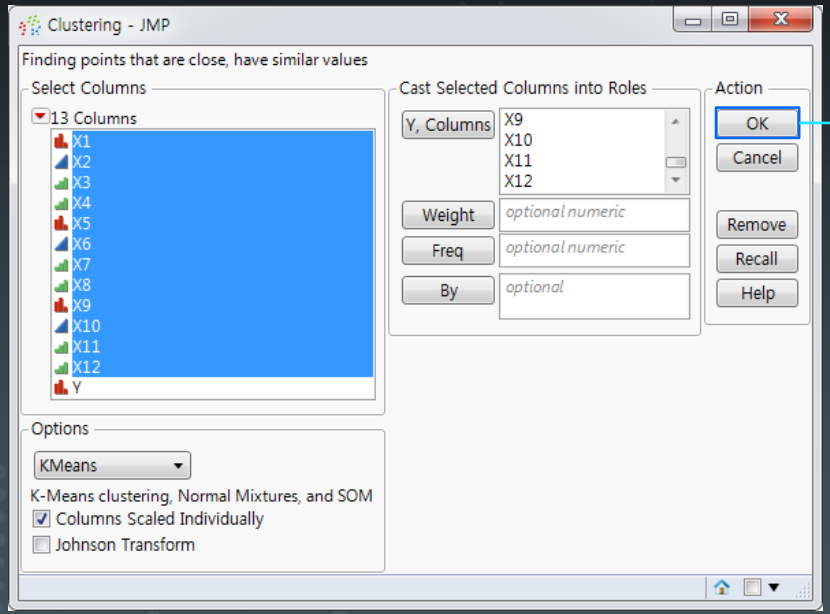


- 군집 수 K를 2로 결정
- 군집 중심(■) 랜덤 초기화
- (Expectation) 가장 가까운 중심 기준으로 군집을 형성함
- (Maximization) 군집의 경계에 맞춰 중심을 이동함
- (Expectation) 가장 가까운 중심 기준으로 군집을 조정함
- (Maximization) 군집의 경계에 맞춰 중심을 다시 이동

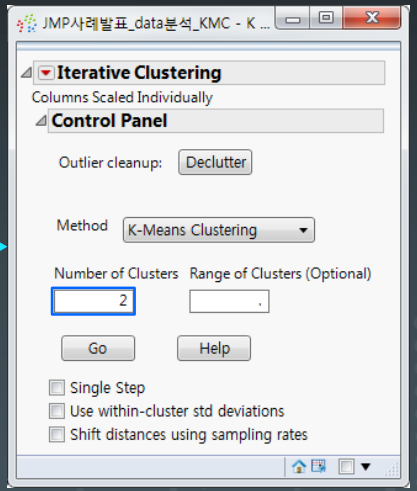
- K-평균 군집화(K Means Cluster)는 대표적인 분리형 군집화 알고리즘이며, 각 중심에 가장 가까운 데이터를 모아 군집을 형성하여 분류하는 통계적 방법임
- 종속변수 정보가 없는 데이터를 분석하여 유사한 군집으로 분류하는 모델에 사용함
- 각 군집의 중심 위치와 각 개체가 어떤 군집에 속하는지 두 가지 정보를 만족한 수준까지 찾기 위해 Expectation 스텝과 Maximization 스텝을 반복함
- 데이터 분포가 특이하면 군집이 잘 이루어지지 않음

K Means Cluster 분석

※ 본 과제의 종속변수 Y 정보를 반영하지 않고 K Means Cluster(비지도학습)을 적용하여 분석함



- JMP: Analyze > Clustering > K Means Cluster
- Number of Clusters: 2



KOREA 2020 DISCOVERY SUMMIT ONLINE

K Means Cluster 결과

JMP사례발표_data분석_KMC - K Me...

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	2	14.0007	Optimal CCC

Columns Scaled Individually

Control Panel

K Means NCluster=2

Columns Scaled Individually

Cluster Summary

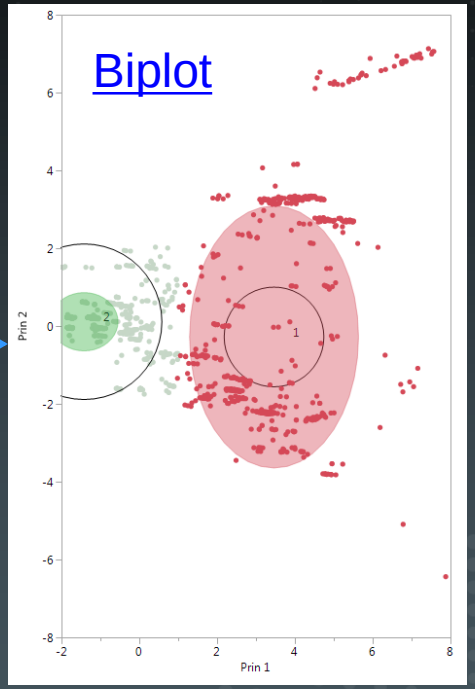
Cluster	Count	Step	Criterion
1	931	22	0
2	2268		

Cluster Means

Cluster	X1	X2	X3	
1	25.3888292	474.616541	0.94522019	1.26
2	28.7041446	55.4259259	0.05026455	0.03

Cluster Standard Deviations

Biplot



분류 성능 검토

Confusion Matrix		예측	
		NG	OK
실제	NG	266	106
	OK	665	2162

Confusion Rates		예측	
		NG	OK
실제	NG	0.715	0.285
	OK	0.235	0.765

정분류율(accuracy): 75.9%
제2종 오류: 28.5%

KOREA 2020 DISCOVERY SUMMIT ONLINE

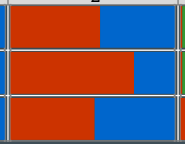
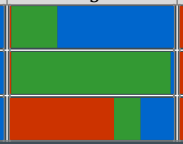
jmp

Latent Class Analysis

※ X인자(X_i)와 각 인자 별 범주(X_{ij})

Cluster	All	X_1		X_2		X_3			X_4		
		X_{11}	X_{12}	X_{21}	X_{22}	X_{31}	X_{32}	X_{33}	X_{41}	X_{42}	X_{43}
Cluster1	0.54	0.39	0.61	0.55	0.45	0.01	0.28	0.71	0.33	0.50	0.17
Cluster2	0.25	0.68	0.32	0.75	0.25	0.01	0.97	0.02	0.79	0.20	0.01
Cluster3	0.21	0.33	0.67	0.51	0.49	0.64	0.16	0.20	0.60	0.05	0.35

※ X인자의 수준(X_{ij}) 별 비율 패턴으로 분류된 Cluster

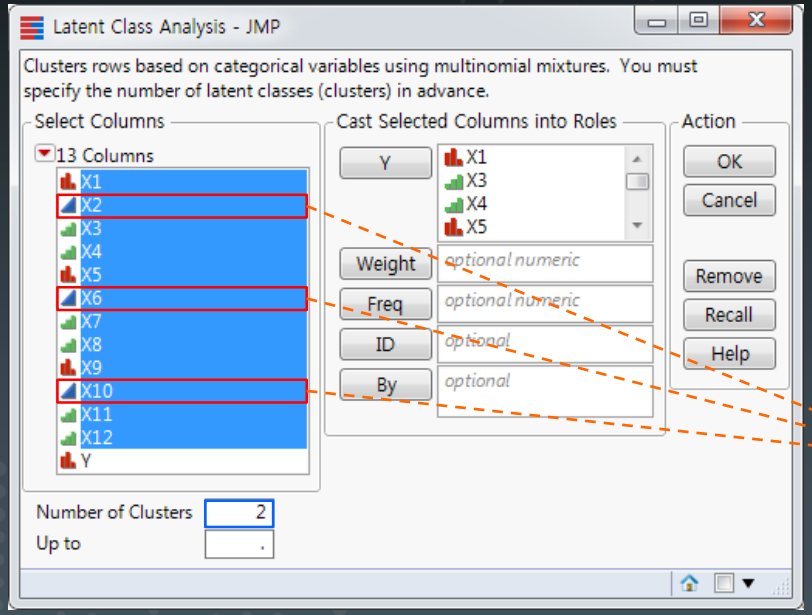
Cluster	All	X_1	X_2	X_3	X_4
Cluster1	0.54				
Cluster2	0.25				
Cluster3	0.21				

- 잠재 클래스 분석(LCA: Latent Class Analysis)은 다변량 범주형 변수를 기반으로 데이터를 분류하는 비지도 학습 알고리즘
- 종속변수에 대한 정보가 없는 범주형 데이터에 잠재되어 있는 군집을 분류함
- 잠재 클래스 모델은 관찰된 변수의 계층화된 교차 분류 테이블을 구하여 잠재적인 범주형 변수로 관찰된 변수들 사이의 유형을 설명한다.
- 각 관측 값을 확률적으로 "잠재 클래스"로 그룹화하여 각 관측 값 변수에 대한 관측치 반응의 기대치를 산출한다.

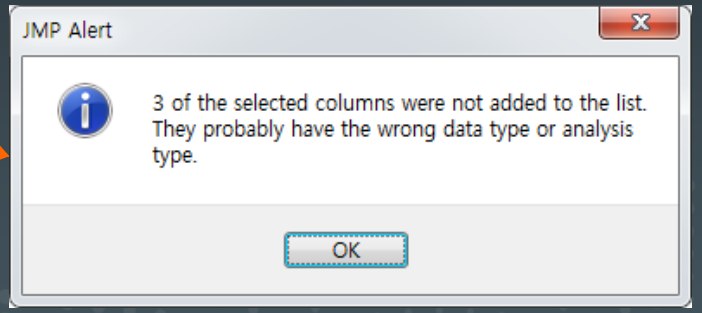
KOREA 2020 DISCOVERY SUMMIT ONLINE

Latent Class Analysis 분석

※ 본 과제의 종속변수 Y 정보를 반영하지 않고 Latent Class Analysis(비지도학습)을 적용하여 분석함

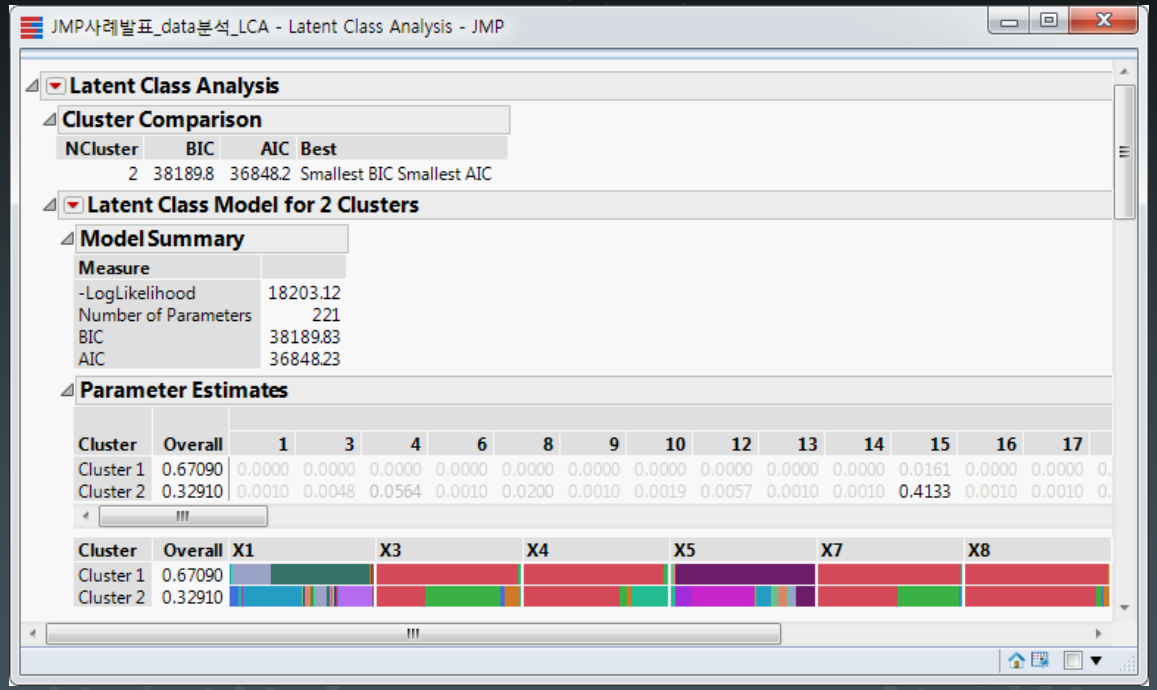


- JMP: Analyze > Clustering > Latent Class Analysis
- Categorical variables only
- Number of Clusters: 2



KOREA 2020 DISCOVERY SUMMIT ONLINE

Latent Class Analysis 결과



분류 성능 검토

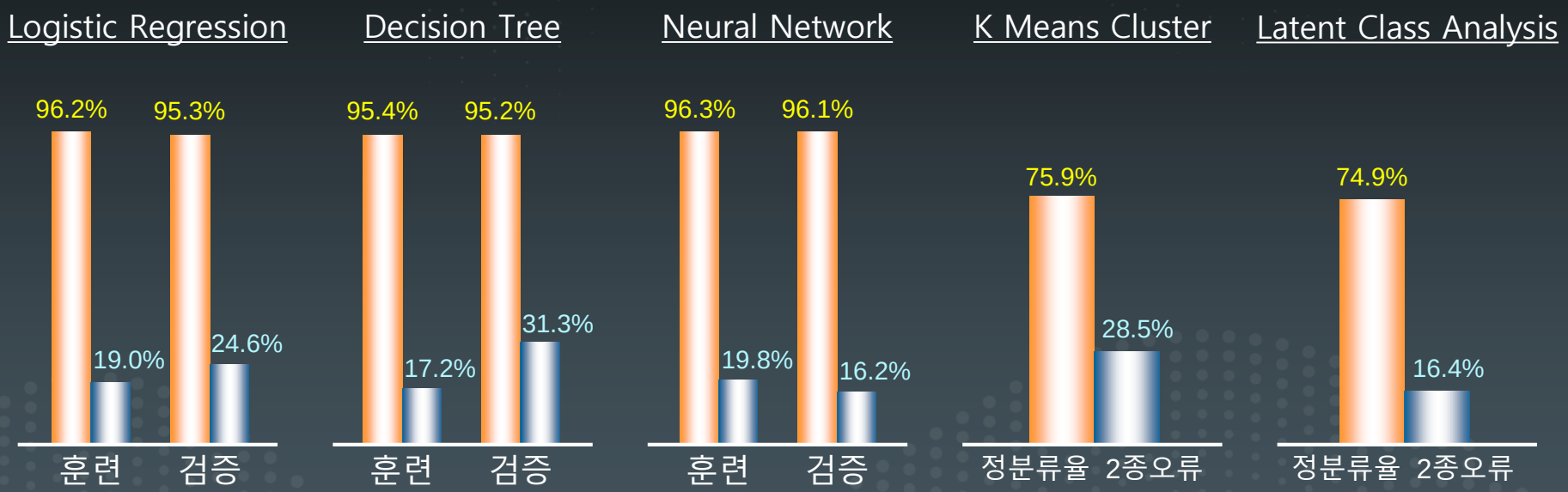
Confusion Matrix		예측	
		NG	OK
실제	NG	311	61
	OK	742	2085

Confusion Rates		예측	
		NG	OK
실제	NG	0.836	0.164
	OK	0.262	0.738

정분류율(accuracy): 74.9%
제2종 오류: 16.4%

분석 결과 종합

※ 지도학습은 데이터 셋을 Training/Validation 분리하였고, 비지도 학습은 전체 데이터를 분석함



* 정분류율(accuracy): 'OK', 'NG'를 제대로 예측한 비율 / **제2종 오류: 실제 'NG' 중에서 'OK'로 오판한 비율

KOREA 2020 DISCOVERY SUMMIT ONLINE

jmp

결론

- JMP를 이용하여 다양한 분석방법을 적용한 결과, 비지도 학습 대비 지도학습의 분류 성능이 우수함
- 지도학습의 분류 성능은 대체로 유사한 수준이었으나 Neural Network의 성능이 다소 우수한 결과를 보였음
- Neural Network을 반복하여 분석하면 매번 분석 결과에 다소 차이가 있으며 이는 범주형 변수의 범주 수가 크기 때문에 데이터 셋에 대한 의존성이 있는 것으로 보임

KOREA 2020

DISCOVERY SUMMIT ONLINE

Thank you!

jmp