

1. 데이터 이해

기출

“A 마트의 다른 상품들도 B 마트보다 쌀 것이라고 판단”	지혜
데이터 사이언티스트가 갖춰야 할 역량은 빅데이터의 처리 및 분석에 필요한 이론적 지식과 기술적 숙련에 관련된 능력인 (㉠) skill 과 데이터 속에 숨겨진 가치를 발견하고 새로운 발전 기회를 만들어 내기 위한 능력인 (㉡) skill 로 나누어진다.	(㉠) Hard (㉡) Soft
(㉠)는 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 그 의미를 부여한 것이며, 지식을 도출하기 위한 재료	(㉠) 정보
기업의 의사결정 과정을 지원하기 위한 주제 중심으로 통합적이며 시간성을 가지는 비휘발성 데이터의 집합을 (㉠)라고 한다.	(㉠) 데이터 웨어하우스
지난 몇 년간 여러 사일로 대신 하나의 데이터 소스를 추구하는 경향이 생겼다. 전사적으로 쉽게 인사이트를 공유하는 데 도움이 되기 때문이다. 다시 말해 별도로 정제되지 않은 자연스러운 상태의 아주 큰 데이터 세트인 (㉠)을 기업들이 구현하는 것은 2017 년 새롭게 등장한 트렌드가 아니다. 그러나 2017 년은 이를 적절히 관리해 운영하는 첫 해가 될 전망이다.	(㉠) 데이터 레이크
형태와 형식이 정해져 있지 않고 언어 또는 문자로 기술되는 데이터`	정성적 데이터
기가바이트(GB) < 테라바이트(TB) < (㉠) < 엑사바이트(EB)	(㉠) 페타바이트(PB)
(㉠)은 공장 내 설비와 기계 사물인터넷(IoT)이 설치되어 공정 데이터가 실시간으로 수집되고 데이터 기반한 의사결정이 일어짐으로써 생산성을 극대화 할 수 있는 기술	스마트 팩토리
1Gbps 는 1 초에 대략 1GB 의 데이터를 전달할 수 있는 속도를 나타낸다. 1Gbps 의 속도를 제공하는 통신망을 통해 1PB 크기의 데이터를 전송하는데 걸리는 시간은 대략 얼마인지 초 단위로	1TB = 1024GB 1PB = 1024TB 1,024 * 1,024 = 1,048,576 KB - MB - GB - TB - PB - EB
다양한 ICT 기술과 금융서비스의 결합은 새로운 금융분야의 변화로 나타나고 있으며, 그에 따른 정보보안이 더욱 중요하게 부각되고 있다. 초기 모바일 결제, 송금영역에서 시작하여 다양한 분야로 확대되고 있으며, 최근에는 빅데이터와 접목하려는 시도들이 잇따르고 있어 더욱 확장성이 기대되기도 한다. 이를 지칭하는 금융과 기술의 합성어	금융기술(FinTech)
이것은 데이터베이스의 구조와 제약조건에 관한 전반적인 명세를 의미하는 것으로서, 데이터베이스를 구성하는 데이터 개체(Entity), 속성(Attribute), 관계(Relationship) 및 데이터 조작 시 데이터 값들이 갖는 제약 조건 등에 관해 전반적으로 정의	스키마
"빅데이터 시대에는 다양한 사업자들이 각종 사용자 데이터나 M2M 센서 등에서 수집된 데이터를 가공처리 저장해 두고, 이 데이터에 접근할 수 있도록 API 를 공개하고, 다양한 서드파티 사업자들이 비즈니스에 필요한 정보를 추출해 활용하게 될 것이다."	플랫폼

서비스 사용자와 광고주를 연결하는 비즈니스에서 가장 중요한 것은 사용자의 특성을 보다 정교하게 파악해 광고주가 도달하고자 하는 정확한 고객군을 만들어 내는 것이다. 이 목표를 위해 활용되기 시작한 것은?	사용자 로그
분석 과제를 도출하기 위한 방식은 문제가 주어진 경우 해법을 찾기 위하여 절차적으로 수행하는 (ㄱ) 방식과 문제의 정의 자체가 어려운 경우 데이터를 기반으로 탐색하고 이를 지속적으로 개선해나가는 방식인 (ㄴ)로 분류된다.	(ㄱ) 하향식 접근방식 (ㄴ) 상향식 접근방식
분석기획은 단기적으로는 (ㄱ)를 도출하여 프로젝트 화 한 후 관리를 수행하여 분석결과를 도출하는 것이고, 중장기적으로는 (ㄴ)를 수행하여 지속적인 (ㄱ)수행을 지원할 수 있는 거버넌스 체계를 수립하는 것이다.	(ㄱ) 분석 과제 (ㄴ) 분석 마스터 플랜
분석 과제에 대한 포트폴리오 사분면 분석을 통해 과제의 1 차적 우선순위를 결정하고, 분석 과제별 적용범위 및 방식을 고려하여 최종적인 실행 우선순위를 결정한 후 실행하는 것으로 단계별로 추진하고자 하는 목표를 명확히 정의하고, 추진 과제별 선 후행 관계를 고려하여 단계별 추진내용을 정렬하는 과정	단계적 구현 로드맵
개인의 사생활 침해를 방지하고 통계 응답자의 비밀사항은 보호하면서 통계자료의 유용성을 최대한 확보 할 수 있는 데이터변환 방법	마스킹
인터넷상의 서버에서 데이터 저장, 처리, 네트워크, 콘텐츠 사용 등 서로 다른 물리적인 위치에 존재하는 컴퓨팅 자원을 가상화 기술을 통해 IT 관련 서비스를 한 번에 제공하는 혁신적인 컴퓨팅 기술	클라우드 컴퓨팅
데이터 사이언티스란 데이터로부터 의미 있는 정보를 추출하는 학문이다. 통계학이 정형화된 실험 데이터를 분석 대상으로 하는 것에 비해, 데이터 사이언스는 정형 또는 (ㄱ)을 막론하고 인터넷, 휴대전화, 감시용 카메라 등에서 생성되는 숫자와 문자, 영상 정보 등 다양한 유형의 데이터를 대상으로 한다.	(ㄱ) 비정형
인터넷 등 각종 경로로 정보를 수집하는 구글은 이미 지난 2010 년에 서비스 이용자가 1 시간 뒤 어떤 일을 할지 .. 예측할 수 있는 데이터와 분석 신뢰도를 확보하고 있다고 했다. 여행 사실을 트윗한 사람의 집에 강도가 노리는 고전적 사례도 발생. 이러한 사례를 통해 알 수 있는 빅데이터 시대의 위기 요인	사생활 침해
구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004 년 발표한 소프트웨어 프레임워크로 페타바이트 이상의 대용량 데이터를 신뢰도가 낮은 컴퓨터로 구성된 클러스터 환경에서 병렬 처리를 지원하기 위해서 개발된 것	맵 리듀스

2. 데이터 분석 기획

기출

분석 방법론의 “시스템 구현” 단계에서 시스템으로 구현된 모델은 검증을 위하여 단위 테스트, 통합 테스트, 시스템 테스트 등을 실시한다. 이중 (ㄱ) 테스트는 품질관리 차원에서 진행함으로써 적용된 시스템의 객관성과 안정성을 확보한다.	시스템
------------------------------------------------------------------------------------------------------------------------------------	-----

데이터 거버넌스 체계에서 데이터 저장소 관리란 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소로 구성된다. 저장소는 데이터 관리 체계 지원을 위한 (ㄱ) 및 관리용 응용소프트웨어를 지원하고 관리 대상 시스템과의 인터페이스를 통한 통제가 이루어져야 한다. 또한 데이터 구조 변경에 따른 (ㄴ)도 수행되어야 효율적인 활용이 가능하다.	(ㄱ) 워크플로우 (ㄴ) 사전영향 평가
문제 탐색을 통해서 식별된 비즈니스 문제를 변환하는 단계로써, 문제 탐색 단계가 무엇을 어떤 목적으로 수행해야 하는가에 대한 관점이었다면, (ㄱ) 단계는 이를 달성하기 위해서 필요한 데이터 및 기법(How)을 도출하기 위한 데이터 분석의 문제로의 변환을 수행하게 된다.	(ㄱ) 문제 정의
분석 모델을 가동중인 운영시스템에 적용하기 위해서는 모델에 대한 상세한 “알고리즘 설명서” 작성이 필요하다. “알고리즘 설명서”는 ‘시스템 구현’단계에서 중요한 입력 자료로 활용되므로 필요시 (ㄱ) 수준의 상세한 작성이 필요하다.	(ㄱ) 의사 코드
분석과제 관리 프로세스는 크게 과제 발굴과 (ㄱ)으로 나누어진다. 조직이나 개인이 도출한 분석 아이디어를 발굴하고 이를 과제하여 분석과제 풀(Pool)로 관리하면서 분석과제가 확정되는 분석과제 실행, 분석과제 진행 관리, 분석과제 결과 공유/개선의 분석관계 관리 프로세스를 수행하게 된다.	(ㄱ) 과제 수행
비즈니스 모델 캔버스는 9 가지 블록을 단순화하여 (ㄱ), (ㄴ), 고객단위로 문제를 발굴하고 이를 관리하는 규제와 감사, (ㄷ) 영역으로 나눠 분석 기회를 도출한다.	(ㄱ) 업무 (ㄴ) 제품 (ㄷ) 지원 인프라
기업 또는 기관의 전사 차원에서 식별된 다양한 분석과제를 대상으로 제한된 예산과 자원을 효과적으로 활용하기 위하여 우선순위를 평가하고, 평가 결과에 따른 단계별 구현 로드맵을 수립하는 실행 계획은?	분석 마스터 플랜
데이터 분석 기획을 위해서 데이터 분석 수준진단이 필요하다. 분석 준비도와 분석 성숙도를 통해 데이터 분석 수준을 진단하게 되는데, 분석 준비도 6 개의 영역 중 2 가지를 적으시오.	분석업무 분석 인력/조직 분석 기법 분석 데이터 분석 문화 분석 인프라
정보기술 또는 정보시스템을 전략적으로 활용하기 위하여 조직 내,외부 환경을 분석하여 기회나 문제점을 도출하고 시스템 구축 우선순위를 결정하는 등 중장비 마스터 플랜을 수립하는 절차	정보전략계획 (ISP : Information Strategy Planing)
ㄱ. 기업의 전사 또는 개별 업무별 주요 의사결정 포인트에 활용할 수 있는 분석의 후보들, ㄴ. Analytics 를 적용하였을 때 업무 흐름을 개념적으로 설명한 것으로 일반적으로 유즈케이스라고 표현	분석 유즈케이스 (Analytics Use Case)

<p>ㄷ. 비즈니스 모델을 구성하는 이론을 설명</p> <p>ㄹ. 하나 이상의 분석을 포함</p>	
(ㄱ)란 기업의 전사 또는 각 업무별 주요 의사결정 포인트에 활용할 수 있는 분석의 후보를 의미	분석 기회
빅데이터의 4V 크기(Volume), 다양성(Variety), 속도(Velocity), 가치(Value)를 분석 ROI 요소의 관점으로 살펴보면 투자비용(Investment) 측면의 요소와 비즈니스 효과(Return) 측면의 요소로 나누어 볼 수 있다. 이러한 빅데이터의 4V 중 비즈니스 효과(Return) 측면의 요소	<p><비즈니스 효과 측면> 가치 (Value) <투자비용 측면> 크기(Valume), 다양성(Variety), 속도(Velocity)</p>

1996 년 Fayyad 가 체계적으로 정리한 데이터 마이닝 프로세스로써 기계학습, 인공지능, 패턴인식, 데이터 시각화 등에서 응용될 수 있는 구조를 갖고 있는 분석	KDD (Knowlege Discovery in Database)
분석 과제를 도출하기 위한 방식은 문제가 주어진 경우 해법을 찾기 위하여 절차적으로 수행하는 (ㄱ)방식과 문제의 정의 자체가 어려운 경우 데이터를 기반으로 탐색하고 이를 지속적으로 개선해나가는 방식인 (ㄴ)로 분류된다.	(ㄱ) 하향식 접근 방식 (ㄴ) 상향식 접근 방식
분석 기획은 단기적으로는 (ㄱ)를 도출하여 프로젝트 화 한 후 관리를 수행하여 분석결과를 도출하는 것이고, 중장기적으로는 (ㄴ)을 수행하여 지속적인 (ㄱ)수행을 지원할 수 있는 거버넌스 체계를 수립하는 것이다.	(ㄱ) 분석 과제 (ㄴ) 분석 마스터 플랜
데이터 기반의 의사결정이 필요하지만, 기업의 합리적 의사결정을 가로막는 장애요소가 존재한다. 이 장애요소 3 가지는 무엇인가	고정 관념 편향된 생각 프레이밍 효과
분석 과제에 대한 포트폴리오 사분면 분석을 통해 과제의 1 차적 우선순위를 결정하고, 분석 과제별 적용범위 및 방식을 고려하여 최종적인 실행 우선순위를 결정한 후 실행하는 것으로 단계별로 추진하고자 하는 목표를 명확히 정의하고, 추진 과제별 선 후행 관계를 고려하여 단계별 추진내용을 정렬하는 과정을 무엇이라 하는가?	단계적 구현 로드맵
풀어야 할 문제에 대한 상세한 설명 및 해당 문제를 해결했을 때 발생하는 효과를 명시함으로써 향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용하도록 하는 것	분석 유즈 케이스

분석 활용 시나리오와 분석 체계를 보다 상세히 나타내는 방법으로서 분석별로 필요한 소스 데이터, 분석 방법, 데이터 입수 및 분석의 난이도, 분석수행 주기, 분석 결과에 대한 검증 오퍼쉽, 상세 분석 과정을 정의하는 방법	분석 정의서
사용자가 요구사항이나 데이터를 정확히 규정하기 어렵고 데이터 소스도 명확히 파악하기 어려운 상황에서 일단 분석을 시도해보고 그 결과를 확인해 가면서 반복적으로 개선해 나가는 방법	프로토타이핑 접근법 (prototyping)
관계형 데이터베이스나 다차원 데이터베이스를 이용하여 구축되면 대부분 데이터는 데이터 웨어하우스로부터 복제되지만, 자체적으로 수집될 수도 있는 데이터 웨어하우스와 사용자의 중간층의 데이터베이스	데이터 마트

3. 데이터 분석

기출

SQL 을 활용하거나 SAS 에서 porc sql 로 작업하던 사용자에게 R 프로그램에서 지원해주는 패키지	sqldf()
출력 결과 x <- 1:100 sum(x>50)	50
<p>y, x1, x2 사이의 적합 된 회귀식</p> <pre> Call : lm(formula = y ~ x1 + x2, data = datavar) Residuals : Min 1Q Median 3Q Max -0.0575279 -0.0163589 -0.0008483 0.0168662 0.0718922 Coefficients : Estimate Std. Error t value Pr(> t) (Intercept) 0.1570203 0.2324673 0.675 0.5058 x1 0.0028231 0.0004171 6.769 5.31e-07 *** x2 -0.2786003 0.1344397 -2.072 0.0491 * --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.03094 on 24 degrees of freedom Multiple R-squared: 0.7411, Adjusted R-squared: 0.698 F-statistic: 17.18 on 4 and 24 DF, p-value: 8.968e-07 </pre>	$y = 0.1570203 + 0.0028231x_1 - 0.2786003x_2$

다변량 회귀분석 결과 $u:1, v=2, w=0$ 때, y 값

```
> m<-lm(y~u+v+w)
> m
Call:
lm(formula = y ~ u + v + w)

coefficients:
(Intercept)      u          v          w
          3.8      -0.21      0.41      -0.16
```

$$y = 3.8 - 0.21u + 0.41v - 0.16w$$
$$y = 3.8 - 0.21 + 0.82$$

여러 대상 간의 객관적 또는 주관적 관계에 관한 수치적 자료를 이용해 유사성에 대한 측정치를 상대적 거리로 시각화하는 방법으로 설문지 응답자의 개개인의 유사성과 선호도 차이를 시각화하여 보고 설명하는 통계적방법론

다차원 척도법

평균으로부터 t standard deviation 이상 떨어져 있는 값들을 이상값(outlier)으로 판단하고 t 는 3으로 설정하는 이상값 검색 알고리즘은?

ESD
(Extreme Studentized Deviation)

최적방정식을 선택하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수를 하나씩 제거하면서 더 이상 유의하지 않는 변수가 없을 때까지 설명변수를 제거하는 방법

후진제거법

College 데이터의 Grad.Rate 변수의 기초통계량을 계산한 결과이다. College 데이터의 Grad.Rate 변수의 몇 %가 78보다 큰 값을 가지는가

```
>summary(College$Grad.Rate)
```

min.	1st Qu.	Median	Mean	3 rd Qu.	Max.
10.00	53.00	65.00	65.46	78.00	118.00

25%

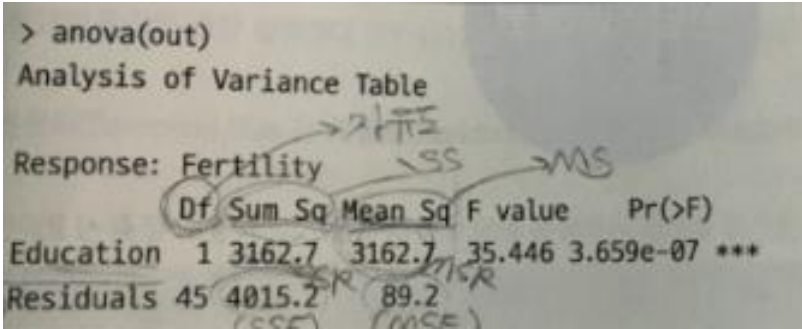
아래 주성분 분석의 결과에서 두 개의 주성분을 사용할 때 설명 가능한 전체 분산의 비율

```
> model<-princomp(Car)
> summary(model)
```

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.503	1.075	0.840	0.752	0.555
Proportion of Variance	0.453	0.231	0.141	0.113	0.061
Cumulative Proportion	0.453	0.684	0.825	0.938	1.000

68.4%

<p>아래 회귀분석 모형의 추정에 대한 설명에서 (ㄱ)은</p> <p>단순회귀분석 모형을 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$로 표현할 수 있다. 주어진 자료를 가장 잘 설명하는 회귀계수의 추정치는 보통 제곱오차 $\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$을 최소로 하는 값을 구한다. 이와 같이 구해진 회귀계수 추정량을 (ㄱ)이라고 한다.</p>	최소제곱
<p>번호를 부여한 샘플을 나열하여 k 개씩 n 개의 구간을 나누고, 첫 구간에서 하나를 임의로 선택한 후에 k 개씩 띄어서 표본을 선택하고 매번 k 번째 항목을 추출하는 표본 추출 방법</p>	계통추출방법
<p>귀무가설(H0)이 옳은데 귀무가설을 받아들이지 않고 기각하게 되는 오류</p>	제 1 종 오류
<p>조사하기 위해 추출한 모집단의 일부 원소</p>	표본(Sample)
<p>다차원척도법은 여러 대상간의 관계에 관한 수치적 자료를 이용해 유사성에 대한 측정치를 상대적으로 (ㄱ)로 시각화는 방법</p>	거리
<p>통계적 추론에서 (ㄱ)검정은 자료와 추출된 모집단의 분포에 대해 아무 제약을 가하지 않고 검정을 실시하는 검정방법으로, 관측된 자료가 특정 분포를 따른다고 가정할 수 없는 경우에 이용</p>	비모수
<p>상관분석은 데이터 안의 두 변수간의 관계를 알아보기 위해 상요한다. 두 변수간의 상관관계를 알아보기 위해 상관계수를 이용한다. 상관계수 중 서열척도인 변수간의 상관관계를 측정하는데 사용하는 상관계수</p>	스피어만 상관계수
<p>결정계수(R²) 계산</p>  <p>회귀식 변동량(R) : Education 잔차(오차)(E) : Residuals $R^2 = SSR / SST(SSR+SSE)$</p>	$3162.7 / (3162.7 + 4015.2) = 0.4406$
<p>y, x1, x2 사이의 적합 된 회귀식 작성</p>	$y = 0.15702 + 0.00282 \cdot x_1 + 0.27860 \cdot x_2$

```
Call:
lm(formula = y ~ x1+x2, data = datavar)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1570203  0.2324673   0.675   0.5058
          x1    0.0028231  0.0004171   6.769 5.31e-07 ***
          x2   -0.2786003  0.1344397  -2.072  0.0491  *

Residual standard error: 0.03094 on 24 degrees of freedom
Multiple R-squared:  0.7411,    Adjusted R-squared:  0.698
F-statistic: 17.18 on 4 and 24 DF,  p-value: 8.968e-07
```

시계열자료를 분석하는 목적 중 하나는 과거의 패턴이 유지된다는 가정 하에서, 현재까지 수집된 자료를 분석하여 미래에 대한 예측을 하는 것이다. 이를 위해 전체 자료를 이용하는 대신 최근 m 개의 관측값들만의 평균을 구하여 지엽적인 변동을 제거하여 장기적인 추세를 쉽게 파악할 수 있는 방법

자기회귀모형
(AR 모형 :
AutoRegression Model)

자료의 위치를 나타내는 척도의 하나로 관측치를 크기순으로 배열하였을 때 전체의 중앙에 위치한 수치이다. 평균에 비해 이상치에 의한 영향이 적기 때문에 자료의 분포가 심하게 비대칭인 경우 중심을 파악할 때 합리적인 방법

중앙값

이것은 인공지능망의 한계를 극복하기 위해 제안된 심화신경망을 활용한 기계학습 방법이다. 기존의 인공지능망은 높은 분해 정확도에 비해 속도가 느린 것이 단점이었다. 게다가 과적합도 웬만해선 해결되지 않는 과제였다. ...연구자르이 그에 대한 해법을 내놓으면서 다시 각광을 받기 시작했다. ...

딥러닝
(Deep Learning)

트랜잭션에서 추출된 연관규칙 중 하나인 "BC"의 신뢰도

Transaction #1 {A,B,C}
Transaction #2 {A,B,D}
Transaction #3 {A,B}
Transaction #4 {B,C}
Transaction #5 {A,B,C,D}
Transaction #6 {E}

신뢰도 : $P(A \cap B) / P(B)$

 $(3/6) / (5/6) = 3/5 = 0.6$

두 개체 A, B 사이의 유클리디안 거리	<table><tr><th>개체</th><th>변수1</th><th>변수2</th></tr><tr><td>A</td><td>3</td><td>4</td></tr><tr><td>B</td><td>6</td><td>8</td></tr></table>	개체	변수1	변수2	A	3	4	B	6	8	$\sqrt{(3-4)^2 + (6-8)^2}$ = $\sqrt{5}$
개체	변수1	변수2									
A	3	4									
B	6	8									
분류할 데이터와 주어진 데이터의 모든 거리를 계산하여 가까운 거리의 데이터를 K 개 만큼 찾은 후 그 중에서 가장 빈도수가 높은 클래스로 분류해주는 기법		K-NN									
최적화방법은 우리 생활과 밀접하게 연관되어 있다. 어떤 물건을 구입할 때 우리는 몇 가지 대안 중에서 재정적인 고려와 함께 구입 이유, 사용 기간, 가격 등 여러 조건을 비교 검토한 후 결정을 내린다. 이러한 결정을 내릴 때 최대 효과, 최소 비용, 최고의 선택과 같은 최적화의 개념을 인식하게 된다. 이러한 최적화 방법 중 가장 많이 사용되는 방법		선형 계획법									
"실제 상황을 수학적으로 모델화하고, 그 모델을 컴퓨터에 프로그램으로 최적화 후, 일어날 수 있는 가능 한 모든 상황을 입력함으로써 각각의 경우에 어떤 결과가 도출되는지 예측		시뮬레이션									
연관성 분석에서 "전체 거래 중 항목 A와 항목 B를 동시에 포함되는 거래의 비율"로 정의되는 척도		지지도 (Support)									
연관성 분석에서 "상품 A를 포함하는 거래 중 A와 B가 동시에 거래되는 비율"		신뢰도 (Confidence)									
연관성 분석에서 "상품 A가 주어지지 않았을 때 B의 확률 대비 A가 주어졌을 때 B의 확률 증가 비율"		향상도 (Lift)									
R에서 다음 명령의 결과 X <- c(1,2,3,NA) Mean(X)		NA									
분류분석의 모형평가 방법으로 랜덤모델과 비교하여 해당 모델의 성과가 얼마나 향상되었는지를 각 등급별로 파악하는 그래프		향상도 곡선									
오분류표를 활용하여 모형을 평가하는 지표 중 범주 불균형을 가지고 있는 데이터에 대한 중요한 범주만을 다루기 위해 사용되는 지표로 실제값이 False 인 관측치 중 예측치가 적중한 정도를 나타내는 지표		특이도									
코호넨에 의해 제시되었는데 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원 뉴런으로 정렬하여 지도의 형상화하는 클러스터링 방법		SOM (Self-Organizing Map)									
혼합분포군집은 모형 기반의 군집 방법으로서 데이터가 k 개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정 하에서 분석을 하는 방법이다. k 개의 각 모형은 군집을 의미하며 이 혼합모형의 모수와 가중치의 최대가능도추정에 사용되는 알고리즘		EM 알고리즘									
분류 모형의 성능을 평가하기 위하여 x축에는 (1-특이도), y축에는 민감도를 나타내어 이 두 평가값의 관계를 나타낸 그래프		ROC curve									
어떤 항목집합이 빈발하다면, 그 항목집합의 모든 부분집합도 빈발하다는 원리로 연관 규칙 알고리즘 중에서 가장 먼저, 많이 사용되고 있는 알고리즘		Apriori 알고리즘									

다수 모델의 예측을 관리하고 조합하는 기술을 메타 학습이라고 한다. 여러 분류기들의 예측을 조합함으로써 분류 정확성을 향상시키는 기법	앙상블 기법
----------------------------------------------------------------------------	--------

예상

공간적 차원과 관련된 속성들을 시각화에 추가하여 지도 위에 관련 속성들을 생성하고 크기, 모양, 선 굵기 등으로 구분하여 인사이트를 얻는 분석방법은 무엇인가?	공간분석
데이터 프레임명은 test, 경영학과 학생들의 데이터만 조회	<code>subset(test, subset=(학과==경영학과))</code>
A 반과 B 반 학생들이 동일한 과목을 들었다고 하자. A 반과 B 반 학생 모두를 대상으로 과목별 성적의 평균을 구하려고 할 때, A 반 학생 데이터와 B 반 학생 데이터를 class 라는 변수를 기준으로 합치려고 한다. R 프로그램으로 작성	<code>merge(A, B, by='class')</code>
학급 내 국어, 영어, 수학, 과학 석차로 구성된 데이터셋의 명이 test 라고 할 때 상관관계를 분석하고자 한다. R 을 활용하여 프로그래밍	<code>rcorr(as.matrix(test), type="spearman")</code>
R 명령의 결과 : 0/0	NaN (Not a Number)
상관관계가 있는 변수들을 결합해 상관관계가 없는 변수로 분산을 극대화하는 변수로, 선형결합을 해 변수를 축약하는데 사용되는 분석방법	주성분분석(PCA)
이것은 데이터 안의 두 변수 간의 관계를 알아보기 위해 사용하는 값이다. 두 변수간의 공분산으로는 음과 양의 관계를 파악할 수 있으나 관계 정도를 확인하기는 힘들다. 그래서 각 변수의 표준편차를 곱하여 공분산으로 나누어 -1 에서 1 사이 값으로 표준화하여 두 변수 간의 관계 정도를 확인 할 수 있도록 수치화 한 것	상관계수
어떤 객체가 불량인지 우량인지 또는 생존하느냐 못하느냐와 같이 0 과 1 로 구분하는데 활용되거나 A,B,C,D 또는 1 등급, 2 등급, 3 등급 중에 어느 등급에 속하는지와 같이 정해진 범주로 분류하는데 사용되는 데이터마이닝 분석방법	분류 (Classification)
빈도가 높고 핵심어 일수록 큰 글씨로 중심부에 표현되며, 어떤 말을 하고 있는지 한 눈에 볼 수 있도록 단어들이 구름처럼 만든 비주얼 분석도구	워드 클라우드 (Word Cloud)
통계분석 방법에는 크게 (ㄱ)와 (ㄴ)이 있는데 (ㄱ)은 수집된 자료를 이용해 대상 집단에 대한 특성값(모수)이 무엇인지를 추측하는 것을 의미하고 (ㄴ)은 수집된 자료를 정리, 요약하기 위해 평균, 표준편차, 중위수, 최빈값 등과 다양한 그래프를 통해 대상 집단을 분석하는 방법	(ㄱ) 통계적추론 (ㄴ) 기술통계
고객은 늘 구매하지 않는다. 경쟁사의 고객 빼앗기에 따른 고객의 변심 또는 고객의 니즈나 취향이 변해 더 이상 상품과 서비스를 사용하지 않고 경쟁사와 거래하는 고객	이탈고객

의사결정나무 중 연속형 타깃변수(또는 목표변수)를 예측하는 의사결정나무	CART															
데이터마이닝 모델링 분석 기업 중 random input 에 따른 forest of tree 를 이용한 분류방법으로 랜덤한 forest 에는 많은 트리들이 생성된다. 새로운 오브젝트를 분류하기 위해 forest 에 있는 트리에 각각 투입해 각각의 트리들이 voting 함으로써 분류하는 방식의 R 패키지	랜덤 포레스트 (Random forest)															
개인과 집단들 간의 관계를 노드와 링크로서 모델링해 그것의 위상구조와 확산 및 진화과정을 계량적으로 분석하는 방법론	사회연결망분석 (Social Network Analysis)															
텍스트 마이닝의 절차 중 데이터의 정제, 통합 선택, 변환의 과정을 거친 구조화된 단계로서 더 이상 추가적인 절차 없이 텍스트 마이닝 알고리즘 실험에서 활용될 수 있는 상태 (특정한 목적을 가지고 언어의 표본을 추출한 집합)	corpus															
<div>암 연구소에서 환자들을 대상으로 암을 예측하고자 하는 분류 문제를 해결하기 위해 학습데이터를 활용한 모델을 개발하였다. 테스트 데이터를 활용하여 모델의 성능을 평가하고자 할 때 분류표를 활용하여 모델의 정확도를 계산</div> <table><tr><td colspan="2"></td><th colspan="2">예측 결과</th></tr><tr><td colspan="2"></td><th>양성</th><th>음성</th></tr><tr><th rowspan="2">실제 결과</th><th>양성</th><td>55</td><td>10</td></tr><tr><th>음성</th><td>20</td><td>15</td></tr></table>			예측 결과				양성	음성	실제 결과	양성	55	10	음성	20	15	(55) / (55+20) = 73%
		예측 결과														
		양성	음성													
실제 결과	양성	55	10													
	음성	20	15													
데이터마이닝의 중심이 되는 학습 방법 중 자료가 입력변수와 출력변수로 주어지며 입력변수와 출력변수의 함수적 의존 관계를 자료로부터 추정함으로써 예측모형을 얻을 때 사용되는 학습방법	지도학습															
텍스트마이닝에서 문서에서 문장 내에 포함된 단어들을 어간과 어미로 분리하여 각 문서마다 사용된 단어의 어간들의 빈도를 표현하는 행렬을 만들 수 있다. R 프로그램을 통해 이러한 행렬을 만들고자 할 때 활용하는 함수	DocumentTermMatrix()															
문장에서 사용된 단어의 긍정과 부정여부에 따라 얼마나 긍정적인 추이가 증가하는지, 감소하는지를 분석하는 방법으로 Opinion mining 이라 언급되기도 하는 분석 방법	감성분석 (Sentiment Analysis)															
변수 X(연속형)와 변수 Y(연속형) 사이의 연관성을 살펴보고자 할 때, 제 3의 변수 Z(연속형)가 X와 Y에 연관되어있다고 가정하자. 이런 경우 Z에 조건화하여 X와 Y 간 상관계수를 산출할 필요가 있다. 이러한 상관계수를 무엇이라고 하는가	편상관															
여러 대상 간의 관계에 대한 수치적 자료를 이용해 유사성에 대한 측정치를 상대적 거리로 시각화하는 방법	다차원척도법															
분류모형의 평가에 사용되는 그래프로 x 축은 (1-특이도), y 축은 민감도로 그려지는 것은	ROC 그래프 (Receiver Operating Characteristic)															

흔히 기업의 데이터베이스에서 상품의 구매, 서비스 등 일련의 개체 또는 사건들 간의 규칙을 발견하기 위해 사용되는 대표적인 정형 데이터마이닝 기법	연관성분석(장바구니분석, 서열분석)
군집분석은 모집단에 대한 사전 정보가 없는 경우 주어진 관측값 사이의 유사성을 이용하여 전체를 몇 개의 집단으로 그룹화하여 각 집단의 성격을 파악하는 분석법이다. 군집을 나누는 방법 중 n 개의 관측값을 각각 하나의 군집으로 간주하고 관측값의 특성이 가까운 군집끼리 순차적으로 합해가는 방법	응집분석 병합적분석
연관성분석은 데이터 안에 존재하는 항목간의 연관규칙을 발견하는 과정이다. 연관성분석의 척도들 중 두 품목 A와 B의 지지도(Support)는 전체 거래 항목 중 항목 A와 항목 B가 동시에 포함되는 비율로 정의되며 전체 거래 중 항목 A와 항목 B를 동시에 포함하는거래가 어느 정도인지 나타내주어 이를 통해 전체 구매 경향을 파악할 수 있다. 그러나 지지도는 연관규칙 A->B와 B->A가 같은 지지도를 갖기 때문에 두 규칙의 차이를 알 수 없다. 이에 대한 평가 척도는	신뢰도 (Confidence)
데이터마이닝의 중심이 되는 학습 방법 중 자료가 출력변수 없이 입력변수만 주어진 경우, 입력변수들간의 상호관계나 입력 자료값들 간의 관계를 탐색적으로 분석할 때 사용되는 학습방법	비지도학습
데이터마이닝 모델링 분석 기법 중 CART와 유사한 트리를 생성하고 예측오차를 최소화할 수 있는 의사결정나무 기법의 패키지	rpart

출처: <https://data-make.tistory.com/145> [Data Makes Our Future]