**26-10. 마스터 플랜** 수립 시 **우선 순위 고려요소**

· 전략적 중요도, 비즈니스 성과/ROI, 실행 용이성

26-12.

· **데이터 표준화** : 데이터 표준 용어 설정, 명명 규칙 수립, 메타 데이터 구축, 데이터 사전 구축 등의 업무로 구성됨

· **데이터 관리 체계** : 데이터 정합성 및 활용의 효율성을 위해 표준 데이터를 포함한 메타 데이터와 데이터 사전의 관리 원칙을 수립하는 것

· **데이터 저장소 관리** : 메타 데이터 및 표준데이터를 관리하기 위한 전사 차원의 저장소를 구축하는 것

· **표준화 활동** : 데이터 거버넌스 체계를 구축한 후 표준 준수 여부를 주기적으로 점검하고 모니터링 하는 것

26-15.

**· 하향식 접근방식**

* 분석과제가 주어지고 이에 대한 해법을 찾기 위해 각 과정이 체계적으로 단계화 되어 수행하는 방식
* 단계 : 문제 탐색(Problem Discovery) -> 문제 정의(Problem Definition) -> 해결방안 탐색(Solution Search) -> 타당성 평가(Feasibility Study)

**· 상향식 접근방식**

* 문제의 정의 자체가 어려운 경우 데이터를 기반으로 문제를 지속적으로 개선하는 방식
* 다양한 데이터를 대상으로 가치 있는 문제 도출

26-16. 빅데이터 분석 방법론 중 **시스템 구현** 설명

· 정보보안영역, 코딩은 주요 고려사항 X·

· 시스템 설계 및 구현, 테스트 및 운영이 주요 고려사항임

26-23, 28-39

· **지지도** : 전체 거래 중 A와 B를 동시에 **포함**하는 거래의 비율로 정의

· **신뢰도** : A를 **포함**한 거래 중에서 A와 B가 같이 **포함**될 확률, 연관성의 정도 파악 가능

= 지지도 / P(A)

· **향상도**

* A가 구매되지 않았을 때 B의 구매확률에 비해 A가 구매됐을 때 B의 구매확률의 증가 비율
* A와 B의 구매가 서로 관련 X -> 향상도=1
* 향상도가 1보다 크면 해당 규칙이 결과를 예측하는데 있어 우수함
* 향상도 = 신뢰도 / P(B)

26-26

· **최적회귀방정식**

* Step 함수를 사용하여 변수 선택법 적용함
  + direction : forward(전진선택법), backward(후진제거법), stepwise & both(단계적선택법)

26-29

· **비계층적 군집 방법**

* 군집 수에 따른 **집단 내 제곱합의 그래프**를 그려 군집의 수를 정함

26-31

· **로지스틱 회귀모형**

* **오즈(odds)** : 성공할 확률이 실패할 확률의 몇배인지 나타내는 확률 = p/(1-p)
* **오즈비** : 오즈의 비율
* **로짓(logit) 변환**
  + log + odds로 오즈에 자연로그를 씌운 것
  + [0, 1]의 범위를 [로 바꿔주는 변환

26-32

· **판별분석**

* 거리 측정을 위해 유클리드, 마할라노비스, 표준화, 맨하탄 거리 사용
* 로지스틱 회귀분석과 비슷하며 프로파일링에 사용
* 고객그룹을 나누거나 자연과학, 금융, 마케팅에 활용
* 선형 분류함수 적용 (예 : 피셔의 선형 분류함수)

26-33

· **주성분분석**

* 분산을 최대화하는 차원을 찾는 방법
* **평균 고유값 방법** : 고유값들의 평균을 구한 후 고유값이 평균값 이상이 되는 주성분을 설정하는 것

26-40

· **시계열분석**

* 시계열 자료는 대부분이 비정상 자료임
* 비정상 자료의 **정상성 조건**을 만족시켜 정상 시계열로 만든 후 시계열 분석 실시
* 폭발적인 추세를 보이거나 시간에 따라 분산의 변화 여부로 정상/비정상 시계열 판단함
* 시계열자료는 시간의 흐름에 따라 관측되기 때문에 대체로 독립적이지 않음
* 자료가 추세를 보이는 경우 **차분**을 통해 비정상시계열을 정상시계열로 바꿔줌
* 계절 요인
  + 일정한 주기를 가지고 반복적으로 같은 패턴을 보이는 변화를 나타내는 요인
  + 요일, 월, 사계절 각 분기에 의한 변화 등 고정된 주기에 따라 자료가 변하는 경우를 말함

26-단1

· **플랫폼**

* 용도에 따라 다양한 형태로 활용될 수 있는 공간
* 4차 산업혁명시대의 소프트웨어, 서비스로 확장되어 활용되는 개념

26-단2

· **정보**

* 데이터의 가공, 처리와 데이터간 연관관계 속에서 의미가 도출된 것

26-단3

· **메타 데이터**

* 데이터의 데이터
* 데이터에 관한 구조화된 데이터로 다른 데이터를 설명해주는 데이터

26-단4

· 데이터 거버넌스

* 전사 차원의 모든 데이터에 대해 정책 및 지침, 표준화, 운영조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축하는 것
* 마스터데이터, 메타데이터 데이터사전을 관리함

26-단6

· **코사인 유사도**

* 내적공간의 두 벡터간 각도의 코사인 값을 이용하여 측정된 벡터 간의 유사한 정도를 의미함

26-단7

· **SOM**

* 코호넨에 의해 제시, 개발됨
* **비지도** 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화함
* 경쟁층 : 입력벡터의 특성에 따라 벡터가 한 점으로 클러스터링 되는 층

26-단8

· ROC Curve

* 가로축을 FPR(1-특이도), 세로축을 TPR(민감도)값

26-단10

· **AR 모형**

* 시계열 모델 중 자기 자신의 과거 값을 사용하여 설명하는 모형
* 백색 잡음의 현재 값과 자기 자신의 과거 값의 선형 가중합으로 이루어진 정상확률 모형
* 모형에 사용하는 시계열 자료의 시점에 따라 1차, 2차, …, p차 등을 사용
* 정상시계열 모형에서는 주로 1, 2차 사용

27-2

· 데이터 유형 : 정성적 / 정량적

27-8

· 신경망 모형 : 인간의 뇌를 형상화한 기법으로 딥러닝에 기반을 두고 있는 모형

27-9

· 채널

* 기업이 고객 세그먼트에게 가치를 제안하기 위해 커뮤니케이션을 하고 상품이나 서비스를 전달하는 방법
* 기업과 고객의 인터페이스 전반 (커뮤니케이션, 물류, 판매채널)
* 유통 채널을 공급하는 것 X

27-10,13

· CRISP-DM 분석 방법론

|  |  |  |
| --- | --- | --- |
| **단계** | **내용** | **수행업무** |
| 업무이해  Business Understanding | · 비즈니스 관점에서 프로젝트의 목적과 요구사항 이해  · 데이터 분석을 위한 문제 정의 | 업무 목적 파악, 목표 설정, 프로젝트 계획 수립 |
| 데이터 이해  Data  Understanding | · 데이터 수집 및 데이터 속성을 이해하는 단계  · 데이터 품질에 대한 문제점을 식별하고 숨겨져 있는 인사이트를 발견하는 단계 | 초기 데이터 수집, 데이터 탐색 |
| 데이터 준비  Data Preparation | · 분석 기법에 적합한 데이터 편성 | 데이터 셋 선택, 데이터 정제, 데이터 포맷팅 |
| 모델링  Modeling | · 모델링 기법, 알고리즘 선택 및 파라미터 최적화  · 과적합 여부 확인 | 모델링 기법 선택, 모델 평가 |
| 평가  Evaluation | · 목적에 부합하는지 평가 | 분석결과 평가, 모델링 과정평가, 모델 적용성 평가 |

\*위대한 실패

**평가** 단계에서 모델에 대한 평가 등을 파악하여 Business에 대한 이해가 부족해 모형 개발이 잘못되었을 때 다시 **업무이해(Business understanding)**단계로 돌아감

27-11

· 분석과제 기획

* 데이터 분석을 위해서는 데이터정형화 필수X
* 분석을 위한 데이터 확보가 우선적
* 데이터 유형에 따라 적용 가능한 솔루션 및 분석방법이 다르기 때문에 유형에 대한 분석이 선행적으로 이루어져야 함

27-17

· 의사결정나무 불순도 측도 (p.406 참고)

* **카이제곱 통계량**
  + ((실제도수-기대도수)의 제곱 / 기대도수)의 합
  + 기대도수 = 열의합계 \* 합의 합계 / 전체합계
* **지니지수**
  + 노드의 불순도를 나타내는 값
  + 지니지수의 값이 클수록 이질적이며 순수도가 낮음
* 엔트로피 지수
  + 열역학에서 쓰는 개념으로 무질서 정도에 대한 측도
  + 엔트로피 지수가 클수록 순수도가 낮음
  + )

27-24

· **계층적 군집분석**

* 최단연결법
  + n\*n 거리행렬에서 거리가 가장 가까운 데이터를 묶어 군집 형성
  + 군집 간 또는 데이터와의 거리 계산 시 **최단**거리를 거리로 계산하여 거리행렬 수정 진행
  + 수정된 거리행렬에서 거리가 가까운 데이터 또는 군집을 새로운 군집으로 형성
* 최장연결법
  + 군집 간 또는 데이터와의 거리 계산 시 **최장**거리로 계산하여 거리행렬을 수정하는 방법
* 평균연결법
  + 군집 간 또는 데이터와의 거리 계산 시 **평균** 거리로 계산하여 거리행렬을 수정하는 방법
* 와드연결법
  + 군집내 편차들의 제곱합을 고려한 방법
  + 군집 간 정보의 손실을 최소화하기 위해 군집화 진행

27-31

· 회귀분석의 가정

* 선형성 : 독립변수와 종속변수가 선형적이어야 함, 산점도로 확인 가능
* 독립성 : 서로 독립적이어야 함
* 등분산성 : 분산이 일정해야 함
* 비상관성 : 오차들끼리 상관이 없음

· **다중 공선성**

* 3개 이상의 독립변수 간에 상관관계를 띄면 안됨
* 다중 공선성이 나타나면 설명력이 약해지고 표준오차가 증가함
* 중요하지 않으면서 다른 변수와 상관성이 높은 변수를 제거하여 해결
* 표본수가 증가해도 VIF에서 일반 결정계수는 크게 변하지 않음
* 구조적 다중 공선성의 문제가 있는 경우, 데이터의 평균 중심을 변화함
* 변수 k의 분산팽창요인 VIF=1/(1-R2)으로 결정계수에 영향받음

27-33

· **중심극한정리**

* 표본평균의 분포는 표본의 크기가 커짐에 따라 정규분포로 근사함
* 동일한 확률분포를 가진 독립 확률 변수의 분포는 n이 적당히 크면 (n이 30 이상) 정규분포에 가까워진다는 정리
* 여러 통계적 방법론에는 정규데이터가 필요하지만 중심극한정리를 사용하면 비정규적인 데이터에도 이와 유사한 절차 적용 가능
* 모집단의 분포가 대칭이면 표본의 크기가 작아도 되지만 모집단의 분포가 비대칭이면 표본의 크기가 30 이상이 되어야 함

**· 부스팅**

* 예측력이 약한 모형들을 결합하여 강한 예측 모형을 만드는 방법
* 약한 학습기의 오류 데이터에 가중치를 부여하면서 최종 모형을 만들어가는 방법
* 재표본 과정에서 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법

**· 배깅**

* 주어진 자료에서 여러 개의 부트스트랩 자료를 생성하고 각 부트스트랩 자료에 예측 모형을 만든 후 결합
* 보팅을 통해 최종 모형을 만드는 방법
* 데이터 집합으로부터 크기가 같은 표본을 여러 번 단순 임의 복원 추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블 하는 방법

**· 랜덤 포레스트**

* 결정트리의 단점인 훈련 데이터에 오버피팅되는 경향을 보완한 방법
* 의사결정나무 모형의 특징인 분산이 크다는 점을 고려하여 배깅보다 더 많은 무작위성을 추가한 방법
* 약한 학습기들을 생성하고 이를 선형 결합해 최종 학습기를 만드는 방법

**· 앙상블 모형**

* 훈련을 한 뒤 예측을 하는데 사용 (지도학습법)
* 주어진 데이터 집합을 바탕으로 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모델을 만드는 방법
* 각 모형의 상호 연관성이 높을수록 정확도는 떨어짐
* 이상값에 대한 대응력이 높음
* 전체적인 예측값의 분산을 감소시켜 정확도를 높일 수 있음
* 모형의 투명성이 떨어져 원인 분석에는 적합하지 않음

28-01

· 데이터 양이 늘어난다고 더 많은 분석 필요 X

· 데이터 양이 늘어날수록 데이터를 처리하는 저장 장치 필요

28-02

· **데이터 단위**

* 바이트(1byte)<킬로바이트(1024B)<메가바이트(1024KB)<기가바이트(1024MB)<테라바이트(1024GB)<페타바이트(1024TB)<엑사바이트(1024PB)<제타바이트(1024EB)<요타바이트(1024ZB)

28-09

· 분석과제 관리 프로세스

* 과제발굴단계 : 분석 아이디어 발굴, 분석과제 후보제안, 분석과제 확정 프로세스
* 과제수행단계 : 팀 구성, 분석과제 실행, 분석과제 진행관리, 결과 공유 프로세스
* 분석과제 중에 발생된 시사점과 분석 결과물을 풀(Pool)로 관리하고 공유됨
* 분석과제로 확정된 분석과제는 풀(Pool)로 관리하지 않음

28-12

· DML(Data Manipulation Language)

* 데이터 조작어로 정의된 DB에 입력된 코드를 조회하거나 수정, 삭제하는 언어
* SELECT, INSERT, UPDATE, DELETE

· DDL(Data Definition Language)

* DB를 정의하는 언어
* 데이터를 생성, 수정, 삭제하는 등 데이터 전체의 골격을 결정하는 언어
* CREATE, ALTER, DROP, TRUNCATE

28-18

· 플랫폼

* **광의**의 분석 플랫폼 : 분석 서비스 엔진, 분석 어플리케이션, 분석 서비스 제공 API
* **협의**의 분석 플랫폼 : 데이터 처리 Framework, 분석 엔진, 분석 라이브러리

28-21

· **순서척도** : 측정 대상의 서열관계를 관측하는 척도

ex) 만족도, 선호도, 학년, 신용등급

· **명목척도** : 측정 대상이 어느 집단에 속하는지 분류할 때 사용 ex) 성별, 출생지

· **구간척도** : 측정 대상이 갖고 있는 속성의 양을 측정하는 것, 구간이나 구간 사이의 간격이 의미 있는 자료 ex) 온도, 지수

· **비율척도** : 간격에 대한 비율이 의미를 가지는 자료, 절대적 기준인 0이 존재하고 사칙연산이 가능하며 제일 많은 정보를 가지는 척도 ex) 무게, 나이, 시간, 거리

28-33

· **의사결정나무**

* **가지치기** : 오차를 높일 위험이 크거나 부적절한 추론 규칙을 가지고 있는 가치 또는 불필요한 가지를 제거하는 단계
* **정지 규칙** : 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 정하는 규칙

28-35

· 밀도기반 군집분석

* 어느 점을 기준으로 주어진 반경 내에 최소 개수만큼의 데이터들을 가질 수 있도록 함으로써 특정 밀도함수 혹은 밀도에 의해 군집을 형성해 나가는 기법
* ex) DBSCAN, OPTICS, DENCLUE

28-37

· FP-Growth 알고리즘

* 후보 빈발항목집합을 생성하지 않음
* FP-Tree를 만든 후 분할정복 방식을 통해 Apriori 알고리즘보다 더 빠르게 빈발항목집합을 추출할 수 있는 방법
* 데이터베이스 스캔 횟수가 적음

28-38

· 정규화

* MinMax 정규화 : 원데이터의 분포를 유지하면서 [0,1]의 값을 가지도록 정규화 하는 방법

28-단1

· BI (Business Intelligence)

* 기업이 보유하고 있는 수많은 데이터를 정리하고 분석해 기업의 의사결정에 활용하는 일련의 프로세스
* 리포트 중심의 도구

· OLAP

* 다차원 데이터를 대화식으로 분석하기 위한 소프트웨어

· Analytics

* 의사결정을 위한 통계적이고 수학적인 분석에 초점을 둔 기법

28-단2

· 데이터 사이언티스트 요구역량

* 하드 스킬 : 빅데이터 처리 및 분석에 필요한 이론적 지식과 기술적 숙련
* 소프트 스킬 : 데이터 속에 숨겨진 가치를 발견하고 새로운 발전 기회를 만들어 내기 위한 능력

28-단3

· 비즈니스 모델 캔버스

* 9가지 블록을 단순화하여 **업무, 제품, 고객** 단위로 문제를 발굴함
* 이를 관리하는 두가지 영역인 **규제 및 감사**영역과 **지원 인프라** 영역에 대한 기회를 추가로 도출하는 작업 수행

28-단4

· 능력 성숙도 통합 모델 (CMMI)

* 소프트웨어 개발 및 전산장비 운영 업체들의 업무 능력 및 조직 성숙도를 평가하기 위한 모델
* 1~5단계로 구성된 성숙도 모델

28-단5

· **결정계수 (R 제곱)**

* SSR(Sum of Square Regression)/SST(Sum of Square Total)를 통해 산출 가능
* 회귀모델에서 독립변수가 종속변수를 얼마나 잘 설명하는지 나타내는 것

28-단10

· **비모수 모형**

* 통계학에서 모수에 대한 가정을 전제로 하지 않고 모집단의 형태에 관계없이 주어진 데이터에서 직접 확률을 계산하여 통계학적 검정을 하는 분석법
* 의사결정나무와 같이 선형성, 정규성, 등분산성 등의 가정을 필요로 하지 않는 모형

29-06

· 개인정보 비식별화 기법

* **가명처리**
  + 직접 식별할 수 없는 다른 값으로 대체함
  + 완전 비식별화가 가능하며 데이터의 변형, 변질이 낮음
* **총계처리**
  + 개인정보에 통계값을 적용함
  + 다양한 통계 분석용 데이터 셋 작성에 유리함
* **데이터값 삭제**
  + 개인정보 식별이 가능한 특정 데이터값 삭제 처리 기법
  + 민감한 개인 식별 정보에 대해 완전한 삭제 처리가 가능하여 예측, 추론이 어렵도록 함
* **범주화**
  + 단일 식별 정보를 해당 그룹의 대푯값이나 구간값으로 변환
  + 고유 정보 추적 및 식별 방지 기법
  + 범주나 범위는 통계형 데이터 형식이므로 다양한 분석 및 가공 가능
* **데이터 마스킹**
  + 개인 식별 정보의 전체 또는 부분을 대체값, 공백, \*로 변환하는 기법
  + 완전 비식별화 가능하며 원시 데이터의 구조에 대한 변형이 적음

29-11

· 빅데이터 분석방법론 – 분석 기획 단계

* 프로젝트 위험 대응 계획 수립 시 예상되는 위험 대응 방법
  + 회피(Avoid), 전이(Transfer), 완화(Mitigate), 수용(Accept)

29-12

· 성숙도 모델

* 도입단계 : 데이터 웨어하우스, 데이터 마트, ETL/EAI, OLAP
* 활용단계 : 실시간 대시보드, 통계분석 환경
* 확산단계 : 빅데이터 관리 환경, 시뮬레이션·최적화, 비주얼 분석, 분석 전용 서버
* 최적화단계 : 분석 협업환경, 분석 Sandbox, 프로세스 내재화, 빅데이터 분석

· 빅데이터 3V : Volume(크기), Variety(다양성), Veloctiy(속도)

+ 4V : Veracity(정확성), Validity(타당성), Volatility(휘발성), Value(가치)

29-18

· 상관계수

* **스피어만 상관계수**
  + 서열자료인 경우 이용하는 상관계수
  + 순위를 기준으로 상관관계를 측정하는 비모수적 방법
  + 두 변수의 차이가 클수록 상관계수 값 커짐
  + 1에 가까울수록 변수 간에 단조적 상관성을 가짐
  + 한 변수를 단조 증가 함수로 변환하여 다른 변수를 나타낼 수 있는 정도
  + 두 변수의 선형 관계 크기 뿐만 아니라 비선형적 관계도 나타냄
  + 0에 가까울수록 상관성 없음
* **피어슨 상관계수**
  + -1 ~ +1 사이의 값
  + +1은 양의 상관관계
  + 0은 상관관계 없음
  + -1은 음의 상관관계

29-19

· 베이즈 정보기준(BIC)

* AIC가 과적합되는 경향을 제어하기 위한 방법 X
* 패널티 기간에 대한 차이로 인해 다른 방법으로 사용됨

29-21

· 공분산

* 확률 변수 2개의 선형관계를 나타내는 값
* 측정 단위에 영향 받음

29-39

· 제 1종 오류 : 귀무가설이 옳지만 대립가설을 채택

· 제 2종 오류 : 대립가설이 옳지만 귀무가설을 채택

29-단1

· 맵리듀스

* 하둡분산파일시스템에 저장된 대용량의 데이터들을 대상으로 SQL을 이용해 사용자의 질의를 실시간으로 처리하는 기술
* 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위해 구글에서 제작한 프레임워크

29-단2

· 블록체인

* 거래 정보를 하나의 덩어리로 보고 이를 차례로 연결한 거래장부
* 중앙 집중형 서버에 거래 기록을 보관하는 기존과 달리 거래에 참여하는 모든 사용자에게 거래 내역을 보내주며 거래마다 이를 대조해 데이터 위조를 막는 방식
* P2P 방식을 기반으로 생성된 체인형태의 연결고리 기반 분산 데이터 저장 환경에 저장하여 누구라도 임의로 수정할 수 없고 누구나 변경의 결과를 열람할 수 있는 분산 컴퓨팅 기술 기반의 원장관리 기술

29-단4

· ISP

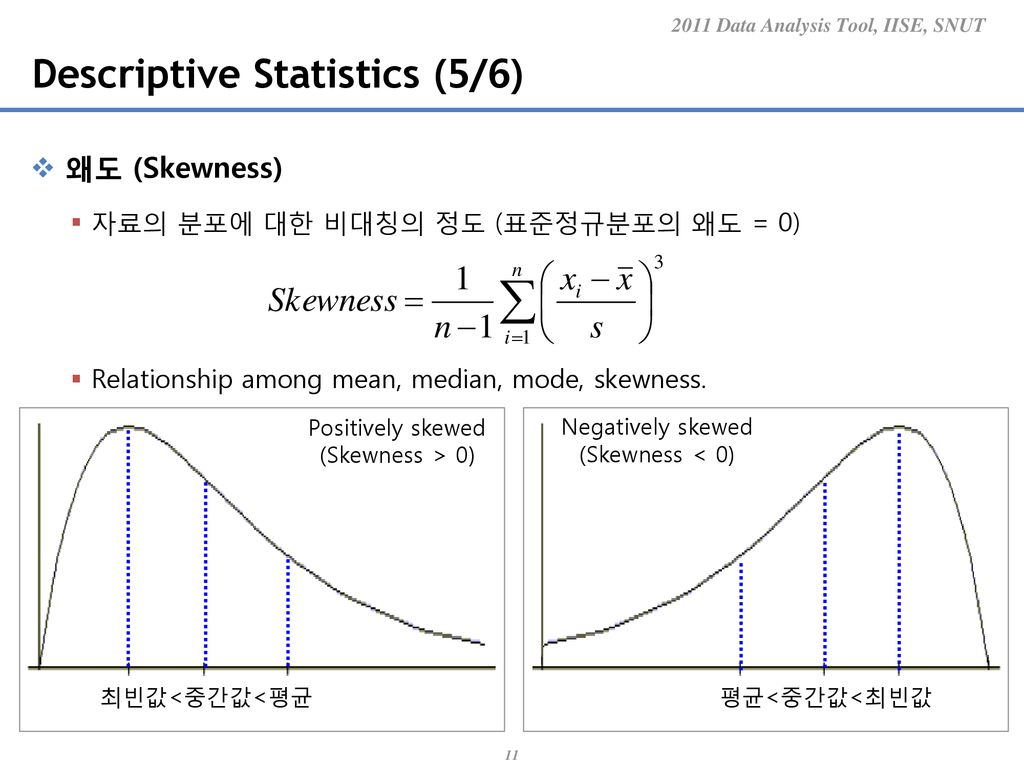
* 정보기술 또는 정보시스템을 전략적으로 활용하기 위해 조직 내외부 환경을 분석하여 기회나 문제점 도출
* 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정 (중장기 마스터 플랜 수립 절차)

29-단5

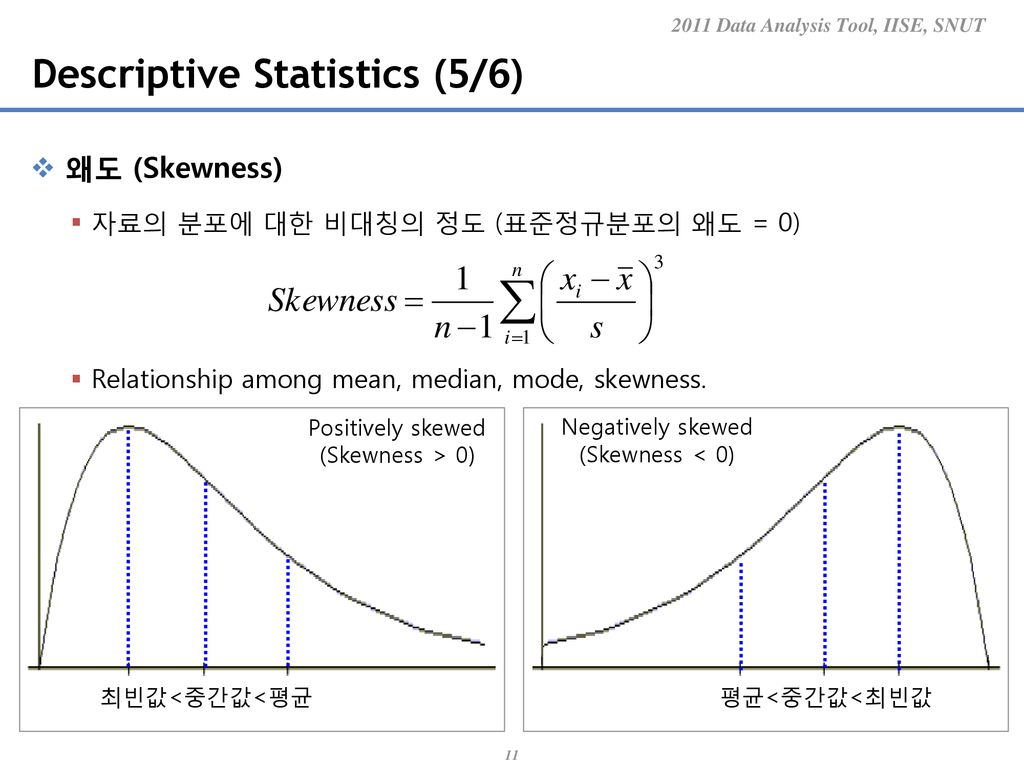
· 검증용 데이터

* 데이터 마이닝 적용 후 결과의 신빙성을 검증하기 위해 분할한 데이터
* 구축된 모델의 과대적합, 과소적합 등에 대한 미세 조정 절차를 위해 사용되는 데이터

오른쪽으로 긴 꼬리



왼쪽으로 긴 꼬리



· **사분위범위 (IQR)**

* 1사분위수 (Q1) : 자료의 1/4
* 3사분위수 (Q3) : 자료의 3/4
* 사분위범위 (IQR) : Q3-Q1 (자료의 50% 범위 내에 위치함)
* 최대값 = Q3 + 1.5 \* IQR
* 최소값 = Q1 – 1.5 \* IQR

· **확률 표본추출 기법**

**· 단순무작위 추출**

* 편향성 제거
* 모집단에 대한 사전지식이 많지 않은 경우 시행함

**· 계통추출**

* 추출간격을 설정하여 간격 사이에서 무작위로 추출
* N개인 집단에서 K라는 추출 간격으로 뽑는다면 N/K 수만큼 표본이 선택됨

**· 층화추출**

* 모집단을 서로 겹치지 않게 동질적인 여러 층으로 나누어 분할된 층별로 표본을 단순임의 추출에 따라 추출
* 모집단을 특정한 기준에 따라 서로 상이한 소집단으로 나누고 각각의 소집단으로부터 일정한 표본을 무작위로 추출
* 특성치의 효율적인 추정이 필요한 경우 시행
* 단순임의 추출에 비해 추정의 정도를 높일 수 있음
* ex) 모집단 남녀 성비 3:2일 때 표본 성비도 3:2로 추출할 경우

**· 군집추출 (집락추출)**

* 서로 인접한 데이터를 묶어 군집으로 구분하여 일부 군집을 추출
* 단순 임의추출에 비해 표본 오차가 클 수 있음

30-05

**·** 데이터 거버넌스 체계 – 데이터 관리 체계

* 데이터 정합성 및 활용의 효율성을 위하여 표준데이터를 포함한 메타 데이터와 데이터 사전의 관리 원칙을 수립함.
* 빅데이터의 경우 데이터 양의 급증으로 데이터의 생명 주기 관리 방안을 수립하지 않으면 데이터 가용성 및 관리비용 증대 문제에 직면할 수 있음
* 메타데이터 관리, 데이터 사전 관리, 데이터 생명주기 관리

30-12

**· 과제 중심적 접근 방식**

* Speed & Test, Quick-Win, Problem Solving

**· 장기적인 마스터 플랜 방식**

* Accuracy & Deploy

30-16

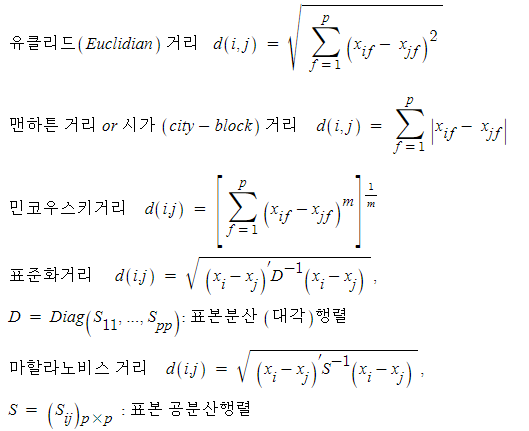
**· 분석 준비도**

* 기업의 데이터 분석 도입의 수준을 파악하기 위한 진단 방법
* 분석업무파악, 인력 및 조직, 분석기법, 분석 데이터, 분석문화, IT 인프라로 구성됨
* 분석기법 : 업무별 적합한 분석기법 사용, 분석업무 도입 방법론, 분석기법 라이브러리, 분석기법 효과성 평가, 분석기법 정기적 개선

30-17

**· 민코우스키 거리**

* 맨하탄 거리와 유클리디안 거리를 한번에 표현한 공식으로 L1 거리 (맨하탄거리), L2 거리 (유클리디안 거리)로 불림



**· 마할라노비스 거리**

* 변수의 표준화와 변수 간의 상관성을 동시에 고려한 통계적 거리
* 통계적 개념이 포함된 거리
* 변수들의 산포를 고려하여 이를 표준화한 거리

30-23

**· 로지스틱 회귀분석 검정방법**

* 카이제곱 검정

**· 선형회귀분석 검정방법**

* F-검정, T-검정

30-31

**· 영향력 진단**

* 적합된 회귀모형의 안정성을 평가하기 위한 통계적 방법
* **쿡의 거리** : 관측 개체 하나가 제외되었을 때, 최소제곱추정치 벡터의 변화를 표준화한 측도
* **영향점** : 비교할 대상이 있어 그 값들에 비해 값이 매우 크거나 작아 회귀 계수 추정값을 변화 시키는 관측개체
* **DFBETAS**의 절대값이 유난히 큰 관측개체는 해당 회귀계수의 추정에 대해 큰 영향력을 행사하는 것으로 간주함
* **DFFITS**는 절대값이 공식에 대입한 값보다 큰 값이 나타나야 높은 영향력으로 간주함

**· 데이터 정규성 확인**

* 히스토그램, Q-Q plot, Shaprio-Wilk 검정

**· 회귀 모형 오차항의 자기상관 유무 검정**

* Durbin Watson 검정

30-33

**· 1종 오류** : 귀무가설이 참이지만 귀무가설을 기각하는 오류

**· 2종 오류** : 귀무가설이 거짓이지만 귀무가설을 채택하는 오류

30-36

**· 사회연결망 분석**

* 밀도 : 연결망 내에서 전체 구성원들이 서로 얼마나 많은 관계를 맺고 있는가를 나타냄

30-39

**· 안정적 시계열**

* 현재의 충격은 미래의 y값에 관한 예측치에 아무런 영향을 미치지 못함
* 어느 시기에 충격이 발생하여 y값이 평균 이하로 감소하면 미래의 어느 기간에 걸쳐서 y의 증가율이 일시적으로 평균 수준보다 더 높아야 y가 평균 수준을 회복하여 현재의 충격이 무한 미래의 y에 미치는 영향이 소멸됨
* 시간의 추이와 관계없이 평균, 분산이 불변하여 변화했다고 해도 다시 평균으로 회귀하는 경향을 보임

31-01

**· 사물인터넷**

* 인터넷을 기반으로 모든 사물을 연결해 사람과 사물, 사물과 사물 간의 정보를 상호 소통하는 지능형 기술 및 서비스
* 사물에서 생성되는 데이터를 활용한 분석을 통해 마케팅 등에 활용 가능
* 데이터화(Datafication)

31-02

**· 데이터 분석 조직의 유형**

* **기능 구조**
  + 별도의 분석 조직 없이 해당 업무부서에서 분석을 수행함
  + 국한된 분석 수행 이슈 존재
* **분산 구조 :** 별도 분석 전담 조직, 분석조직 인력을 현업부서로 직접 배치함
* **집중구조**
  + 조직 내 별도의 분석 전담조직을 독립적으로 구성함
  + 분석 업무의 중복, 이원화의 이슈 존재

31-07

**· 데이터베이스**

* 종속성과 중복성 배제함
* 데이터 종속성 : 응용프로그램별로 데이터를 별도 관리함

31-10

**· 솔루션**

* 조직 내 분석 대상이 무엇인지 인지하고 있으나 데이터 분석 방법과 다양한 분석 구조를 이해하지 못하는 유형

31-11

**· 분석 유즈 케이스**

* 현재의 비즈니스 모델 및 유사/동종사례 탐색을 통해 빠짐없이 도출한 분석 기회들을 구체적인 과제로 만들기 전에 표기하는 것
* 풀어야 할 문제에 대한 상세 설명 및 해당 문제를 해결했을 때 발생하는 효과를 명시함으로써 향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용함
* 기업의 전사나 개별 업무별 주요 의사결정 포인트에 활용할 수 있는 분석의 후보들을 의미함

31-12

**· 프로토타이핑 방법론의 프로세스**

* 가설 생성
* 디자인에 대한 실험
* 실제 환경 테스트 결과에서 통찰 도출 및 가설 확인

**· 나선형 모델**

* 반복적으로 위험분석을 수행하여 위험을 관리하며 순환적으로 개선하는 것

31-14

**· 빅데이터 거버넌스**

* 단순히 대용량 데이터를 수집·축적하는 것보다는 어떤 목적으로 어떤 데이터를 어떻게 분석에 활용할 것인가가 중요함
* 빅데이터의 경우 데이터양의 급증으로 데이터의 생명 주기 관리방안을 수립하지 않으면 데이터 가용성 및 관리비용 증대 문제에 직면할 수 있음
* ERD는 운영중인 데이터베이스와 일치하기 위해 철저한 변경관리가 필요함
* 빅데이터 거버넌스 산업분야별, 데이터 유형별, 정보 거버넌스 요소별로 구분하여 작성함

31-15

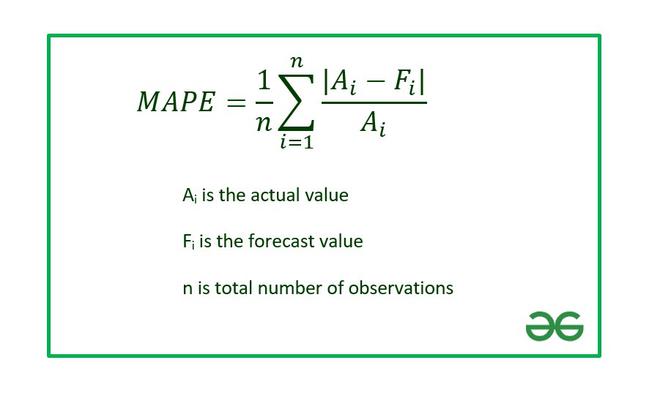
**· 분석을 사용한 전략적 통찰력을 얻는 방법**

* 경영의 본질을 제대로 바라볼 수 있도록 분석함
* 경영진은 직관적 결정을 지양하고 데이터 기반의 객관적으로 의사결정 함
* 사업 상황을 확인하기 위해 사업 내부뿐만 아니라 외부 환경을 같이 분석해야 함
* 비즈니스의 핵심가치와 관련된 분석 프레임워크와 평가지표를 개발함

31-21

**· MAPE**

* MAE를 퍼센트로 변환한 것



A : 실제값

F : 예측값

31-22

**· k-means 클러스터링**

* 이상값에 민감하여 군집 경계 설정이 어려움
* k-medoids 군집화로 극복 가능

\*k-medoids 군집화 : 군집의 가장 중심에 있는 값(medoid)을 사용하여 군집을 찾음

31-24

· 학습모형의 유연성이 클수록 분산은 높고 편향은 낮다.

31-28

**· 군집분석** : k-means 군집화, Single linkage method, DBSCAN