표지는 아직 미정

Chapter 0. Data and Information

Chapter 0의 Data Science에 있어 기반이 되는 지식이라 판단되는 내용을 작성함. 일부 내용은 다른 챕터에서 심도 있게 다룰 예정.

또한 이 모든 문서는 Pandas, Numpy, Scikit-Learn, Tensor_flow2, seaborn, matplots… 등의 사용방법과 문법 등을 알고 있다는 가정 하에 작성함.

List of contents

1. 데이터와 분석 기획

아래 내용과 표는 ADsP, TOPCIT 커리큘럼을 반영하여 자격증을 위한 상식 위주로 정리했습니다. 캡스톤 과제 구현에 있어서 크게 중요한 지식은 아니고 그냥 이런 것도 있구나 하고 넘어가면 되는 교양 지식입니다.

데이터의 유형과 특성

유형	내용
정성적 데이터	언어, 문자 등의 형태로, 저장과 분석에 많은 자원이 소모.
정량적 데이터	수치, 도형, 기호 등의 형태로 상대적으로 비용 소모가 적음

지식창조 메커니즘

특성	내용
Socialization	Tacit Knowledge (이하 암묵지)를 고차원의 암묵지로 전환하는 과정
Externalization	암묵지를 Explicit Knowledge (이하 형식지)로 전환하는 과정
Combination	분산된 형식지의 단편들을 수집, 분류, 통합하여 새로운 형식지를 창조하는 과정
Internalization	형식지를 암묵지로 전환하는 과정

지식 피라미드

지식 피라미드	내용
Wisdom	근본적인 원리를 이해하고, 이를 토대로 도출되는 창의적 아이디어
Knowledge	상호 연결된 정보패턴을 이해하고, 이를 토대로 예측한 결과물
Information	데이터의 가공, 상관관계 간 이해를 통해 패턴을 인식하고 의미를 부여한 데이터
Data	가공 전의 순수한 수치나 기호들을 의미함. Raw data.

데이터 베이스의 특징

구성	설명
Consolidated Data	데이터의 중복 없이 저장하여 데이터를 최소화한 데이터
Stored Data	접근 가능한 저장 매체에 저장되어 있는 데이터
Operational Data	목적에 맞게 운영할 수 있는 데이터
Public Data	기업, 조직 등이 공동으로 소유하고 활용하는 데이터

특징	설명
Real Time	다수의 사용자로부터 데이터 검색, 질의에 즉시 응답
Continuous Evaluation	데이터 입력, 수정, 삭제 등의 작업을 통해 최신화를 유지함
Concurrent Sharing	여러 사용자가 데이터에 접근하고 공유함
Contents of Reference	데이터의 참조는 데이터 내용에 의해 참조됨

장점 : 데이터 중복 최소화, 실시간 접근 가능, 데이터 보안 제공, 데이터 베이스의 논리적/물리적 독립성을 제공, 데이터 표준 및 데이터 공유, 데이터 일관성과 무결성 유지

단점 : 데이터 베이스 구축 비용, backup/Recovery가 복잡.

빅데이터의 특성

특성	내용
Volume	크다. (ㄹㅇ 적을게 이거밖에 없음)
Variety	데이터의 종류가 다양하다. (음성, 이미지, 텍스트 등)
Velocity	실시간 정보의 증가로 인해 분석/처리 속도가 중요해짐
Veracity	잘 정제된 데이터셋을 사용하자는 내용
Value	가치 있는 정보를 도출해야 한다는 내용

빅데이터 구축 및 운영을 위한 테크닉 (아래는 레퍼런스이며, 다음 챕터에서 자세히 설명합니다.)

테크닉	내용
Association Rule Learning	독립 변수간 관련성을 분석
Classification Tree Analysis	책 내용이 조금 이상해서 패스
Machine Learning	
Genetic Algorithm	
Regression Analysis	독립변수와 종속변수 간 관계를 분석
Sentiment Analysis	특정 주제에 대한 감정을 분석
Social Network Analysis	특정인과 다른 사람과의 관계를 분석

빅데이터 처리 프로세스 및 관련 기술에 대해서는 다음 챕터에서 서술합니다. 아래 내용부터는 ADP / ADsP 자격증을 위한 내용 중 일부를 요약하였습니다. (안 읽어도 됨)

빅데이터 위기요인과 통제

위기요인	통제방안	통제방법 예시
사생활 침입	특정 데이터가 본래 목적 외로 가공되어 활용될 가능성이 있다. → 개인정보를 사용, 분석하는 책임자를 지정. (이게 무슨 소리?)	말도 안되는 소리라 삭제.
책임원칙 훼손	분석 대상은 예측 알고리즘의 피해를 입을 수 있다. → 책임 원칙 강화	
데이터 오용	데이터에 대한 잘못된 인사이트 도출로 인한 부수적 피해 → 데이터 알고리즘에 대한 권한/인증을 통제하는 방법 도입	

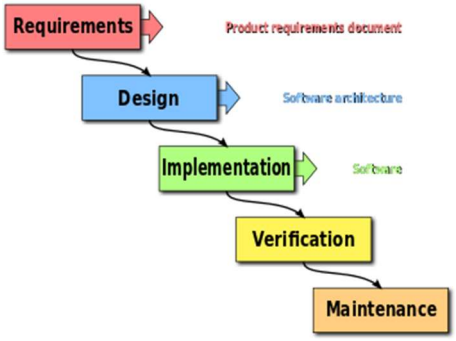
비식별화 (이건 괜찮은 내용)

비식별화 조치방법	설명 및 사례
Pseudonymisation	개인 정보 중 주요 식별요소를 다른 값으로 대체하여 식별을 어렵게 한다
Aggregation or Replacement	데이터의 총합 혹은 평균값으로 대체하여 개별 데이터의 값을 노출하지 않는다
Data Reduction	데이터셋에 구성된 내용 중 필요 없는 값이나 개인 식별 데이터를 삭제한다
Data Suppression	데이터의 값을 범주 값으로 변환하여 명확한 값을 노출하지 않는다
Data Masking	개인 식별에 기여할 확률이 높은 개인 식별자를 보이지 않게 처리한다.

2. 데이터 분석의 이해

빅데이터 분석 방법론은 분석 절차와 방법을 체계적으로 제시한 방법이다. 분석 절차 (Procedure), 방법 (Methods), 도구와 기법 (Tools & Techniques), 분석 단계별 Template과 Output을 정의한 것이다. 요약하면 방법론은 각 단계별 수행해야 할 활동 (Activity)과 작업 (Task), 산출물 (Artifact)를 정의한 것이다.

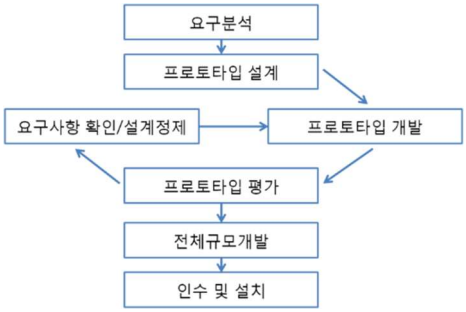
1. 폭포수 모델 (Waterfall Model)



고전적 모델로 순차적 단계를 제시함. 각 단계를 완료해야 다음 단계로 진행하며, 이전 단계로 되돌아 가기 어렵다는 단점이 있고 고객의 요구사항을 확인하기 어려운 문제가 있어 다음 모델인 프로토타이핑 모델이 등장했다.

특징	문제점
하향식 접근, 순차적 모형 표준화된 양식과 문서 중심 프로세스 정형화된 산출물을 중요시하는 모델	사용자 요구사항에 대한 반영과 확인이 어려움 단계별 완전성으로 인해 불필요한 문서 작업이 많음 개발 도중 중요 문제점이 생기면…

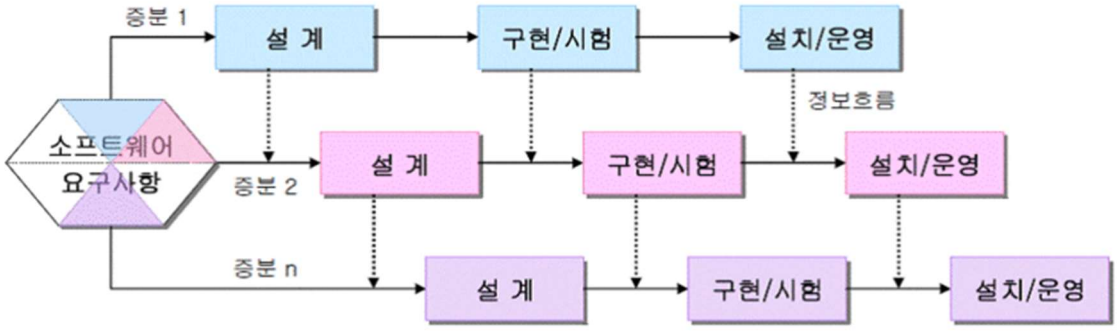
2. 프로토 타이핑 모델 (Prototyping Model)



개발하려는 모델의 주요 기능을 개발 초기에 간단한 모형을 만들어 평가 후 전체 시스템으로 확장하여 구현하는 모델. 폭포수 모델의 요구사항 반영과 확인이 어려운 점을 보완하기 위해 고안됨.

특징	문제점
요구사항 분석 효율적 진행 가능 시스템 전반의 이해와 품질 향상 사용자와의 의사소통이 비교적 원활	프로토 타입 결과를 최종 결과물로 오해하는 경우 중간단계 산출물에 대한 문서화 어려움 프로토타입이 잘못될 경우 오버헤드 발생

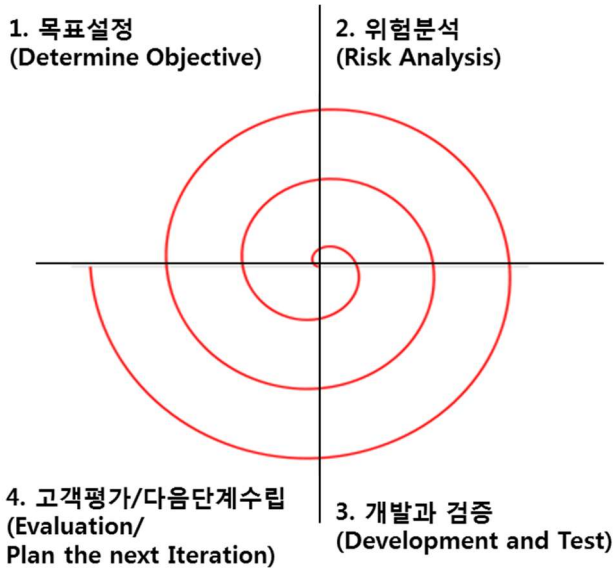
3. 반복 점증 진화 모델 (Iterative Incremental Model)



요구사항, 분석, 설계, 구현을 반복적으로 개발하는 모델이다. 각 파트를 모듈화 하여 여러 번 적용하고 그 결과들을 조합한다. 반복 수행 후에 만들어지는 소프트웨어에 대해 버전업을 하며 완성시켜 나가는 특징이 있다.

장점	단점
반복된 개발로 인해 높은 완성도를 가질 수 있음 위험이 높은 업무를 우선 개발하여 전체 모델에 대한 위험 부담을 낮출 수 있음 릴리즈 방식으로 요구사항 변화에 대응하기 용이함	현실적으로 각 기능과 단계를 적당히 나누기 어려움

4. 나선형 모델 (Spiral Model)



반복적인 위험 분석을 통해 위험성이 높은 개발 프로젝트에 적용하는 모델이다.

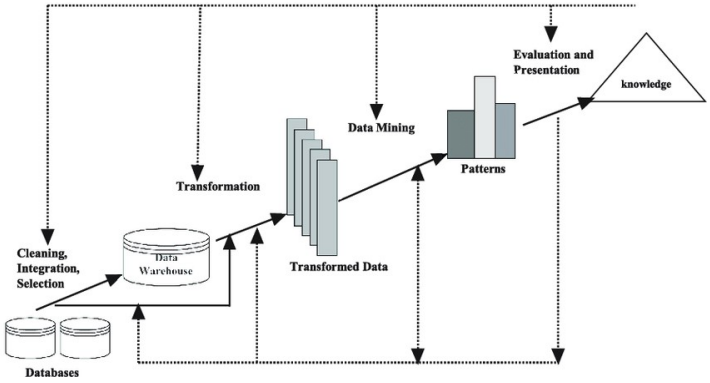
단계	내용
Determine Objective	타당성 검토, 요구사항 분석, 단계별 특정 목표 수립
Risk Analysis	위험 식별, 대응전략 수립, 위험 요인을 초기에 식별
Development and Test	프로젝트에 맞추어 개발 및 검증
Evaluation	결과물에 대한 평가를 수행 및 반복

3. 분석 방법론

KDD (Knowledge Discovery in Databases)

프로파일링 기술을 사용하여 데이터로부터 통계적 패턴이나 지식을 발견하기 위해 정리한 데이터 마이닝 프로세스

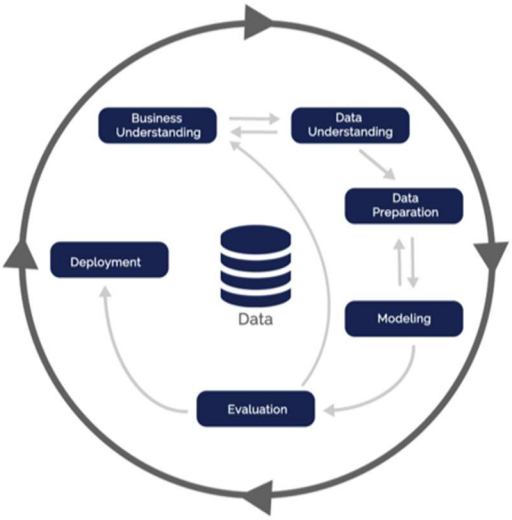
크게 데이터 선정, 전처리, 변환, 마이닝, 평가의 단계로 이루어져 있다.



단계	설명
Selection	OLTP (OnLine Transaction Processing)이나 Data Warehouse를 분석하여 필요한 raw data를 선정한다. (OLTP – 실시간 데이터 / Warehouse – 정적 데이터베이스)
Preprocessing	모델 학습을 하기 위해 데이터에 있는 노이즈, 결측치, 이상치를 정제하는 작업
Transformation	목적에 맞게 데이터를 변환하고 (정규화, 표준화 등등) 학습용 데이터와 검증용 데이터를 분리하는 과정

CRISP-DM (Cross-Industry Standard Process for Data Mining)

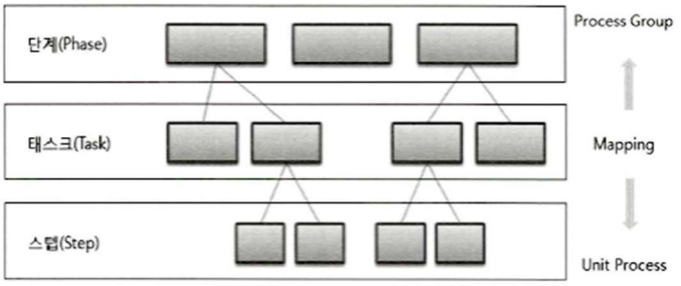
IBM에서 개발한 비즈니스 요구사항에 맞게 데이터 마이닝을 반복적으로 수행하는 사이클형 프로세스

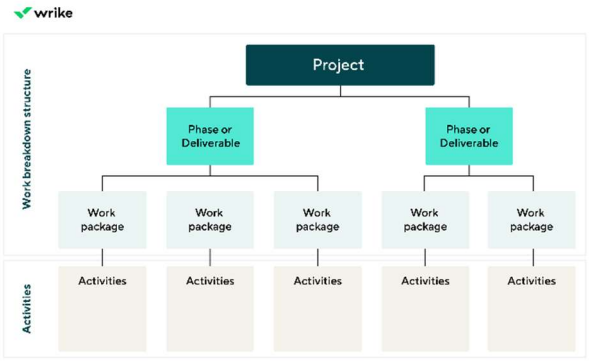


Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation의 단계로 구성

hierarchical analysis

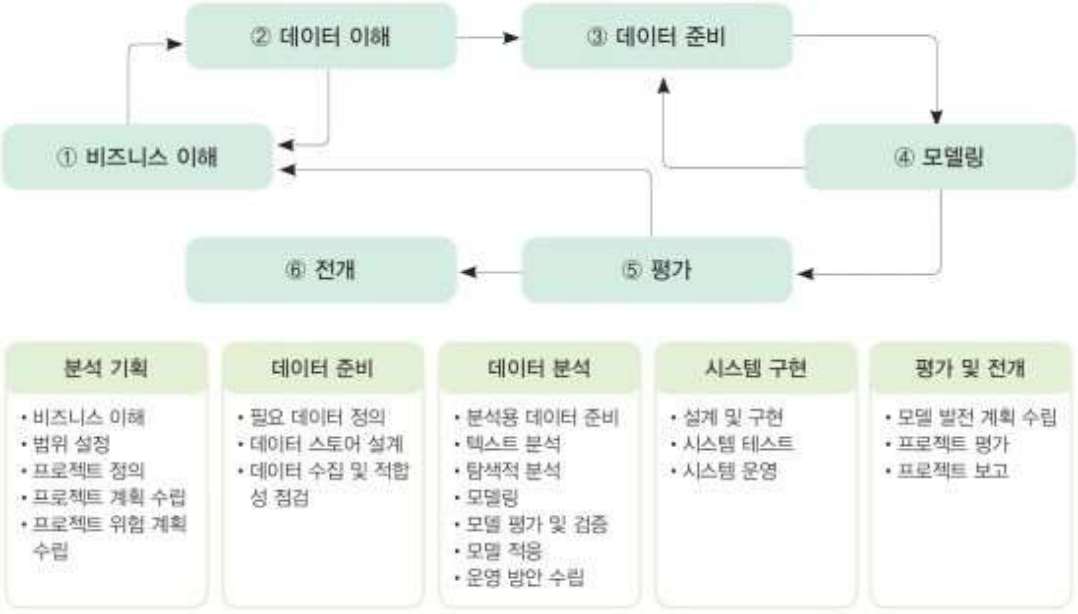
(3-계층 구조)





계층	내용
Phase	빅데이터 분석을 하기 위한 절차를 의미 각 단계는 기준을 설정하고 기준의 완성 여부와 품질을 관리
Task	각 단계별로 수행해야 하는 일을 의미 단계별 작업으로 작업에 대한 성과가 나오게 됨
Step	1~2주 내 완료 가능한 산출물을 의미 Input → Tools → Output으로 만들어지는 단위 프로세스

(5-계층 구조)

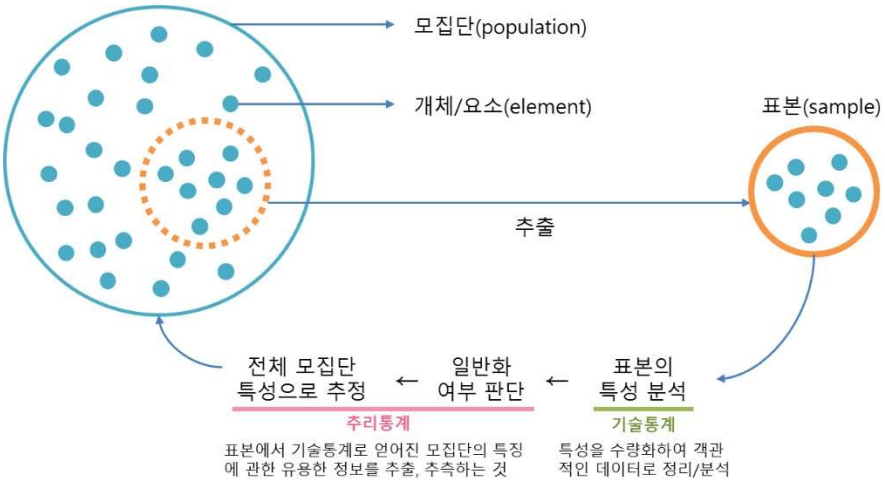


분석기획 → 데이터 준비 → 데이터 분석 → 시스템 구현 → 데이터 평가 및 전개로 이루어진 방법론

https://smartvision-me.com/the-crisp-dm-data-mining-methodology/

3. 통계 분석 - 기술적 통계 분석

통계분석 : 모집단에서 추출한 표본을 분석하여 표본과 모수간의 오차를 분석



Population → Element → Descriptive → Heuristic

변수

질적 변수 : 수치를 나타낼 수 없는 변수로, 빈도 분석을 수행

양적 변수 : 수치로 나타낼 수 있는 변수로, 평균/분산 분석을 수행

표본조사 : 모집단의 특성을 나타내는 자료를 수집하는 행위

1) 확률표집법

1. 단순 무작위 추출 : 모집단의 각각의 요소나 사례들이 표본으로 선택될 가능성이 동등한 표본 추출법
2. 층화 추출 : 모집단을 중복되지 않게 층으로 나눈 다음, 각 층에서 표본을 추출하는 방법
3. 군집 추출 : 모집단에서 집단을 일차 표집 후, 각 집단의 요소를 표본으로 추출하는 다단계 추출법

2) 비확률 표집법

1. 편의표본 추출 : 임의로 선정지역, 조사기간 등을 정의해 표본을 추출
2. 판단표본 추출 : 모집단의 의견을 반영할 수 있을 특정 집단을 표본으로 선정하는 방법
3. 할당표본 추출 : 정해진 분류기준에 따라 소집단으로 구분하고, 각 집단별로 샘플을 추출하는 방법

기술 통계 : 모집단으로부터 수집된 자료를 정리, 요약하여 자료의 특성을 분석.

Histplot, boxplot, Time Series 등을 통해 시각화 할 수 있다.

1) 중심 경향성 (Central Tendency)

중심적인 경향을 나타내는 기술 통계, 산술평균, 중위값, 최빈값 등이 있다. 중심 경향성 통계를 사용할 때는 중심 경향과 일부 데이터에 의해 왜곡되는 정보를 유의해야 한다.

2) 산포도 (Dispersion)

표본의 속성을 나타내는 데이터의 산포 정도를 설명하는 최댓값, 최솟값, 범위, 분산, 표준편차, 표준오차, 평균 표준오차 등이 있다. 평균의 표준오차는 표본 평균의 표본 추출 분포에 대한 표준편차로써 모집단으로부터 표본을 추출한 후, 각 표본에 대한 평균을 구하고 각 평균들에 대한 전체 평균을 다시 구하고, 각 평균들이 전체 평균으로부터 평균적으로 얼마나 떨어져 있는지를 나타낸 것이다. 표준오차가 적을수록 표본의 대표성이 높다.

3) 분포 (Distribution)

데이터 분포 형태와 대칭성을 확인하는 것으로, 첨도와 왜도가 있다. 왜도는 데이터가 대칭에 가까울수록 0에 근접하고, 치우침이 클수록 왜도 또한 증가한다. 첨도는 분포가 표준 분포와 어떻게 다른지를 나타낸다. Reference : https://url.kr/c9xeod

4) 백분위 (Percentile)

4. 통계분석 - 추론 통계

추론 통계 : 표본에서 얻은 어떤 특성의 통계치를 기초로, 오차를 고려하면서 모집단의 모수치를 확률적으로 추정하는 통계적 방법

원래 여기서 통계 용어를 정리하려고 했는데

https://ko.wikipedia.org/wiki/%ED%86%B5%EA%B3%84%ED%95%99#%EC%9A%A9%EC%96%B4

위키에 잘 나와 있습니다

통계적 추론의 종류

1) 모집단에 대한 가정 여부에 따른 통계적 추론의 분류

1. 모수적 추론 (Parametric Inference)

   모집단에 대한 특정 분포를 가정하고, 분포의 특성을 결정하는 모수에 대해 추론하는 방법

   모집단에 대한 가정의 적절함에 따라 최종 추론의 정확성에 영향을 준다.

2. 비모수적 추론 (Non-Parametric Inference)

   모집단에 대한 특정 분포를 가정하지 않고, 다양한 통계량을 고려하여 통계량의 성질을 유도하고 이를 기반으로 통계적 추론을 실행함.

2) 모수처리 방식에 따른 통계적 추론의 분류

1. Frequentist Inference

   $$P(H) = \theta \leq \lim_{x \to \infty} \frac{H}{n}$$

2. Bayesian inference

   이게 뭔소리야

   https://ko.wikipedia.org/wiki/%EB%B2%A0%EC%9D%B4%EC%A6%88_%EC%B6%94%EB%A1%A0

3) 추론 목적에 따른 통계적 추론의 분류

1. 점 추정 (point estimation) : 모수의 값을 추정함
2. 구간 추정 (Interval estimation) : 모수를 포함할 것으로 기대되는 구간을 확률적으로 구함
3. 가설 검정 (Testing Hypotheses) : 모수에 대한 가설을 세우고 그 가설을 확률적으로 판정하는 방법

가설 검정

1) 제 1종 오류

   귀무 가설이 참인데 기각하는 경우로 이 오류를 범할 확률은 alpha이며가설 검정에 대해 설정한 유의 수준이다. 예를 들어 alpha 값이 0.05라면 95%의 신뢰도를 설정한 것이다.

2) 제 2종 오류

   귀무 가설이 거짓인데 기각하지 않은 경우로 이 오류를 범할 확률은 beta이며 검정의 검정력에 따라 달라진다.

다음 내용 : 통계 분석 기법과 머신 러닝