



AI 융합전문가 8차시 데이터분석 II

2024.10.27.





현재 하는 일



소속	주요 업무
(주)제타데이터 대표이사	데이터 분석, 전략 및 컨설팅, 데이터 가치평가 ODA 컨설팅 (Official Development Assistance)
(주)지구파트너스 감사	창업보육, 투자, 기업·기술가치평가, 사업타당성 분석
(주)메타로직 컨설팅 수석	ISP 컨설팅 (Information Strategy Planning) ISMP 컨설팅 (Information System Master Plan)

◎ 자격증

1. 경영지도사31기 (인적자원, 2016)
2. 창업보육매니저 (BI협회, 2018)
3. 기업·기술가치평가사 (기업·기술가치평가협회, 2018)
4. 기업재난관리사 실무과정 (행정안전부, 2019)
5. 데이터분석 준전문가 ADsP (데이터산업진흥원 K-Data, 2021)
6. 빅데이터 분석기사 (과학기술정보통신부 · 통계청, 2021)
7. 국제공인컨설턴트 CMC (ICMCI, 2023)
8. 인공지능(AI) 활용마스터1급 (뉴미디어교육연구소, 2024)



AI & 데이터 관련 비즈니스

한국경제

Hankyung.com

가업을 잇는 생각

알고 보면 재미있는 경영이야기 - 기업 경영과 자본



▲기업철학 홍용기 전문위원

누구나 알고 있는 것처럼 기업은 경영하려면 여러 가지의 자원이 필요하다. 크게 사물로 대표되는 인적·자본·물적 자원으로 나눌 수 있고, 물적 자원의 가장 대표적인 주자가 바로 자본이다.

자본은 무언가 흔히 부채라고 부르는 타인자본과 자기자본으로 나눌 수 있다. 그런데 기업은 경영하는 입장에서 타인자본은 이자를 지급해야만 하고 또한 복귀 상환의 의무를 진다. 반면에 자기 자본은 이익이 있으면 배당할 하고 그렇지 않으면 배당할 하지 않아도 된다. 나아가 회사 자산이 부채로 넘어서는 위험이 없다. 다만 자산이 부채보다 작아지면 회사 부채가 위험을 가질 뿐이다.

따라서 기업의 경영 상황을 나타내는 재무정보의 표기방식에 있어서도 이런 논리가 적용된다. 기업의 재무상태를 나타내는 '재무상태표'에서도 부채와 자본에 표시되는 금액의 합이 총액에 있어서 타인 자본과 자기 자본이 합쳐져서 총액에 표시된다.

일정 기업의 경영실적을 나타내는 '손익계산서'의 경우에도 마찬가지다.

대중에서 이익을 창출하는 순서가 있어서도 영업이익에서 타인자본에 대한 이자비용을 먼저 지출한 이후에 최종적으로 당기순이익을 산출한다.

그런데 중소기업은 경영하는 대표자 혹은 경영자는 타인자본의 부담에 대해서는 당연히 이자를 지급하지만, 자기자본에 대해서는 그 비용이 발생하지 않는 것으로 생각해서는 곤란이 의외로 많다. 하지만 그렇지가 않다. 기업은 경영하는 과정에서 자기자본에 대해서도 비용을 발생하고 그에 대한 배분, 개수, 분할 수급에 대한 책임 경영을 할 수 있다.

당장을 반박해 자본비용을 출자하는 주주, 즉 투자자 입장에서 생각해도, 출자한 금액에 대하여 타인자본처럼 경영실적과 무관하게 이자를 받을 수 있고, 경영실적이 좋지 않은 경우는 배당도 받지 못하게 되며, 나아가 회사 부채를 앞둔 경우 투자 비용 회수에 대한 보장이 되지 않으므로 자기자본에 투자하는 위험들은 타인자본보다 높다고 할 수 있다.

위험률이 높다고 하는 것은 위험이 높은 만큼 요구하는 수익률이 높게 마련이고, 따라서 회사의 자본으로 투자하는 투자자는 타인자본보다 더 높은 수익률을 요구하는 것이 자국의 당연하며, 그에 따라 자기자본 비용은 타인자본 비용보다 더 높을 것이 명상이다.

자본비용은 타인자본 비용과 자기자본 비용을 구분하는 방법으로서 산출하게 된다. 구분공과자본비용 WACC를 간단히 설명하면, 현재자본 중 타인자본과 자기자본 비중이 각각 4/10과 6/10이고, 타인자본 비용과 자기자본 비용이 각각 5%, 10%라고 가정할 때,

* 구분공과 자본비용 WACC
= 0.4 × 5% + 0.6 × 10% = 8%

WACC를 재무담당자가 산출하면 되니 경영자는 몰라도 된다. 다만 WACC가 구분공과 자본비용이라는 뜻만 이해하면 된다. 나아가 자급액으로 더 중요한 것이 차입금에 대한 '책표이자보상비율'이다.

책표이자보상비율은 영업이익으로 자본비용을 감당하는 정도를 말하는데, 타인자본 뿐만 아니라 자기자본까지 포함한 전체 자본비용을 감당할 수 있도록 충분한 개수를 말한다.

예컨대 위와서처럼 자기자본 비중이 타인자본 비중의 두 배라고 가정할 때(실제로도 있다) 그중 자기자본 비중을 1로 보고 자기자본의 비중을 소수점 한자리 자릿수로 환산한 수(자기자본 비중이 40%일 때는 1.4, 60%일 때는 1.6)를 회사운영의 최소한의 책표이자보상비율로 정하면 된다.

책표이자보상비율이 1.60이면, 영업이익이 최소한 이자비용의 1.6배 이상이어야 안정적인 경영이 가능하다는 의미이다. 또 영업이익을 책표이자보상비율로 나눈 값을 타인자본 비중의 곱으로 나눈 값은 회사에서 차입할 수 있는 차입금의 한도액'을 산출할 수 있다. 예컨대 영업이익이 3.2억이라면 3.2억을 1.6으로 나누어 2억을 산출하고, 평균 차입금회전 5%라면 2억을 0.05로 나누어 40억을 산출해 내게 된다.

또한 영업이익 3.2억을 WACC인 0.08로 나누어 위에서 산출한 금액과 같은 수(40억)를 산출할 수 있는데, 영업이익이 영구히 발생한다는 전제를 전제하면 그 금액이 곧 그 기업의 가치를 나타내게 되고, 일부 신용평가 기관에서는 이런 방식으로 산출한 금액을 그 기업의 가치로 평가하기도 한다.

물론 실제 당국에서는 부의 편입 용어나 상식을 탈피하고 정서 복귀를 목적으로 발행하기 위한 비상장기업의 주식을 평가하는 방법론과 관련해서도 명시해 놓았지만, 그것과는 별도로 일부 신용평가 기관에서는 이런 식으로 기업가치를 평가해서 추가적인 대출여부를 판단한다는 사실, 중소기업 대표님들은 명심하시라.

한경 경영지원단
<http://clean.hankyung.com>

한국경제

한경경영지원단

◎ 정보화전략계획수립(ISP) 컨설팅 수행

- 20.05~20.08. 창업진흥원
- 20.10~20.12. 한국연구재단
- 21.01~21.04. 소상공인시장진흥공단
- 21.11~22.06. 서울특별시
- 22.08~22.12. 경찰대학교
- 23.08~24.05 ODA (요르단 경찰청 PSD)

◎ 2022년 AI학습용데이터 구축사업 평가

- 1차 08 방송 콘텐츠 대화체 음성인식 데이터
- 09 방송 콘텐츠 한국어·영어 통번역 데이터
- 43 갑각류 종자생산 데이터
- 48 식생 탄소 포집량 식별 데이터
- 2차 74 축산 기자재(소, 돼지) 3D 데이터
- 75 소(한우, 젃소) 및 돼지 발정행동 데이터
- 3차 06 인공지능 신기술 선도(자유 공모)

◎ 데이터 가치평가 컨설팅

- 23.09~23.11 중소벤처기업진흥공단



AI & 데이터분석 관련 강의

- 데이터분석 및 실전 R코딩 (경영기술지도사회, 빅데이터 분석기사 자격증 취득 과정)
- 데이터분석 Python 심화과정 (서울 여성능력개발원 강동 여성인력개발센터 / 용산 여성인력개발센터)
- 파이썬 코딩을 통한 크롤링 자동화 인텐시브 과정 (경영지도사 및 컨설턴트)
- AI & ChatGPT 활용 및 데이터분석 컨설팅 방법론 (경영기술지도사회, 국제공인컨설턴트 CMC 양성과정)
- AI & 데이터분석 (AI융합비즈니스포럼, 매경아카데미, 동북아 ICT 포럼)





책 쓰기 프로젝트

KYOBObEBook eBook <본의 심리학> 모건 하우절 신작!

2018년 1월

KYOBObEBook 통합검색 > 대입을 결정하는 초등 영어 공부법

대학교재 > 어린이 > 베스트 > 신상품 > 이벤트 > 사운품 > PICKS > CASTing > 교보ONLY

국내도서 > 경제/경영 > 경영전략 > 경영전략일반 >

성장하는 기업의 5가지 조건

한치호, 홍용기, 하현식, 최충철 저(공)

한국경제신문 > 2018년 01월 29일

10.0 (1개의 리뷰) 도움돼요 (100%의 구매자)

성장하는 기업의 5가지 조건

"어떻게 지속적으로 성장할 것인가?"

같이 보아도 없는 세상에서 리더의 위치를 움켜쥐고
무한 경쟁 속에서 살아남는 전략을 찾아라!

한국경제신문!

챗GPT AI로 데이터분석 마스터하기

초보자도 가능한 노코딩 AI 데이터분석

eBook 6,300원

SAIT 구독

초보자도 가능한 노코딩 AI 데이터분석

챗GPT AI로 데이터분석 마스터하기

홍용기
이원구
홍성일

다즈베즈북스
데이터베이스교육센터

2023년 5월

이 모든 것을 담은 All-In-One

생성형 AI 활용 전략 전자책

ChatGPT 활용 전략

ChatGPT 활용 전략

ChatGPT 활용 전략

2024년 4월

AI와 데이터분석은 전문가들의 전유물이 아니다.
AI & Data Literacy for Everyone.

일반인을 위한 인공지능과 데이터 리터러시

챗GPT-4o(omni) 활용 데이터분석

컨설팅학박사 홍용기 저

일반인들 AI와 데이터의 세계로 안내하는 입문서

인공지능과 데이터분석의 기본개념부터 응용까지

마케팅, 인사, 생산관리, 재무 등 실제 응용 가능한 7가지 분석사례

퍼플

2024년 6월





Like every great presentation, I've divided my talk into three subjects. Steve Jobs -

I .

**Data Analysis
Basic Theory**

II .

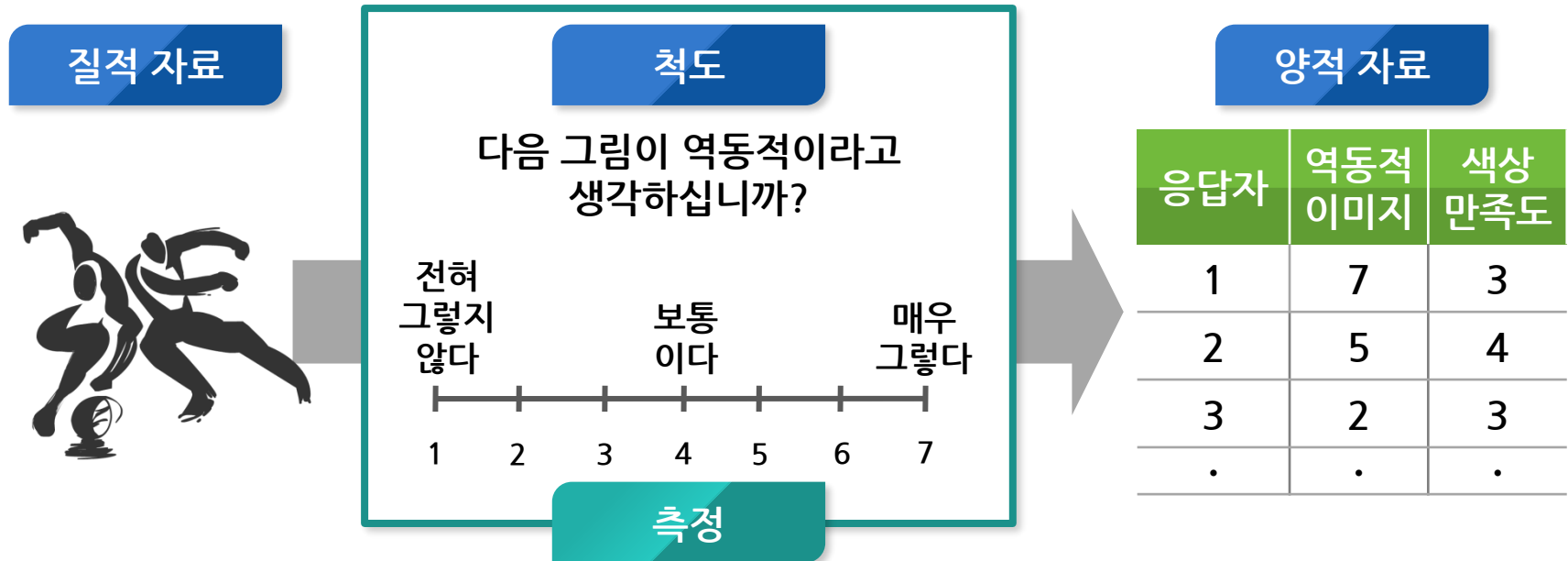
**Regression
Analysis**

III .

**Classification
Analysis**

• 척도의 개념

- 특정 속성을 측정하여 그 정도를 숫자로 나타내는 규칙
- 질적 자료를 양적 자료로 전환시켜 주는 도구



- 척도의 종류
- 어떤 척도를 사용하는지에 따라 측정된 숫자에 내재된 정보량이 달라지며, 적용 가능한 통계분석 기법이 달라짐

척도의 종류	내용
명목척도	응답보기들을 상호 배타적으로 구분 하기 위해 임의의 숫자를 부여하는 척도
서열척도	응답보기들을 구분 하고, 구분한 응답보기들의 순서 까지 측정하는 척도
등간척도	서열 척도에 포함된 정보(분류, 서열정보)외에 거리(간격)정보까지 가지는 척도
비율척도	절대 영점을 가지고 있어서 속성의 상대적 크기 뿐만 아니라, 절대적 크기의 비교도 가능한 척도

1. 명목 척도(Nominal scale)

- 응답보기들을 상호 **배타적으로 구분**하기 위해 임의의 숫자를 부여하는 척도
- 선택한 응답을 기준으로 응답자들을 특정 집단으로 분류하기 위해 사용(=**분류정보**)



귀하는 다음 중 어떤 훈련과정에 입학을 원하십니까?

1) A과정 2) B과정 3) C 과정 4) D 과정 5) 기타

- 숫자는 '크기'의 의미가 없는 명칭에 해당하기 때문에 **사칙연산은 무의미함**
- 대표치는 **최빈치(Mode)** : 응답보기 중 가장 많이 선택된 응답보기의 선택된 수
- 4가지 척도 중 정보량이 가장 적은 척도 : **분류 정보만 보유**

2. 서열 척도(Ordinal scale)

- 응답보기들을 **구분**하고, 구분한 응답보기들의 **순서**까지 측정하는 척도
- 응답보기들의 속성을 서열로 나타내는 척도(=**분류정보** + **순서정보**)
- 응답보기 간의 **간격은 측정하지 않고** 순서만 측정함
 - 응답 보기들 간의 순위만 나타낼 뿐, 얼마나 더 선호되는지는 측정이 불가능함



다음 교육과정 중 귀사에서 가장 중요하다고 생각하는 대로 순서를 기입해 주십시오.
A과정 (), B 과정 (), C 과정 (), D 과정 ()

- 사칙연산은 무의미
 - 순위 간 간격이 서로 달라 숫자 차이에 절대적 의미가 없기 때문
 - 1, 2순위의 차이보다 3, 6순위의 차이가 3배 크다고 할 수 없음
- 대표치로서 중앙값(Median)를 사용함
- 명목 척도 다음으로 적은 정보를 보유함 : 분류 정보 + 순서 정보

3. 등간 척도(Interval scale)

- 서열 척도에 포함된 정보(분류, 서열정보)외에 거리(간격)정보까지 가지는 척도
- **간격이 동일한 서열척도**
- 속성의 **상대적 크기**를 측정하기 위해 균일한 간격으로 분할한 길이를 이용하여 측정
 - 예) 온도계, IQ 등
 - 온도계로 측정한 1도와 2도 간의 차이는 2도와 3도 간의 차이와 동일함
- 간격 척도의 **숫자 자체**는 절대적 의미를 가지지 않음
- 절대 영점이 없기 때문에 **숫자 간 비율개념 없음**
- 간격 척도에서 **숫자 간의 차이**는 절대적 의미를 가짐(**차이 값 간 비율개념있음**)
- 대표치로서 산술평균을 사용
- 정보량 : 분류 정보 + 순서 정보 + 상대적 크기 정보

3. 등간 척도(Interval scale)

예

지난 6개월간 참여하신 교육과정이 취업역량 확보에 도움이 되셨습니까?



- 5점 응답자와 3점 응답자의 만족도 차이가 5점과 4점 응답자의 만족도 차이 보다 2배 크다고 할 수 없음(응답보기(척도점) 간 간격이 동일하다고 볼 수 없기 때문)
- 따라서 **간격 척도라기 보다 서열척도에 가까움**
- 하지만, 사회과학연구의 특성을 고려하여 척도점 간 간격이 동일하고, 각 척도점의 의미를 응답자들이 동일하게 이해한다는 전제 하에 간격 척도로 인정함

4. 비율 척도(Ratio scale)

- 절대 영점을 가지고 있어서 속성의 상대적 크기 뿐만 아니라, 절대적 크기의 비교도 가능한 척도

예

나이 ()세, 근무기간()년, 연봉()원

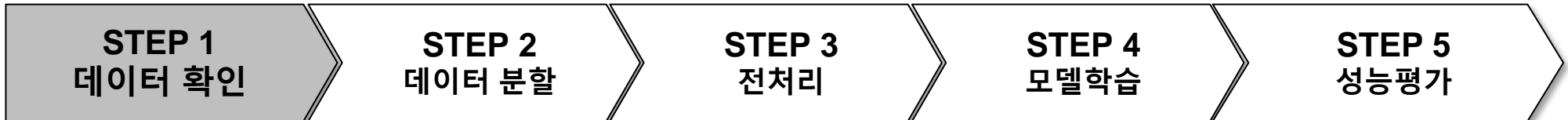
예

하루에 보통 몇시간 정도 훈련이 가능하십니까? ()시간

- 만족도, 선호도, 인지도 등 절대 영점이 존재하기 어려운 소비자의 사고나 인지수준에 대한 측정은 한계가 있음
- 직접 관찰할 수 있는 물리적 사건이나 현상을 측정하는데 주로 사용함
- 사칙연산이 가능하며, 대표치는 평균값
- 4가지 척도 중 가장 정보량이 많은 척도 : 분류정보 + 순서정보 + 상대적 크기 정보 + 절대적 크기 정보

1단계 : 데이터 확인

- 분석할 데이터의 특성을 확인하는 단계
- 변수의 특성(독립변수/입력변수)과 타겟(종속변수/반응변수)의 존재 여부 파악
- 적용가능한 분석모델 확인(ex. 타겟 연속된 수치형이라면 회귀분석, 범주형이라면 분류분석)
- 타겟이 없는 데이터라면 비지도학습 적용



- 독립변수, 종속변수 확인
- 연속형 vs 범주형 확인
- 범주형 독립변수 여부확인
- 적용가능한 분석모델 확인
(회귀, 분류, 비지도 학습)

2단계 : 데이터 분할

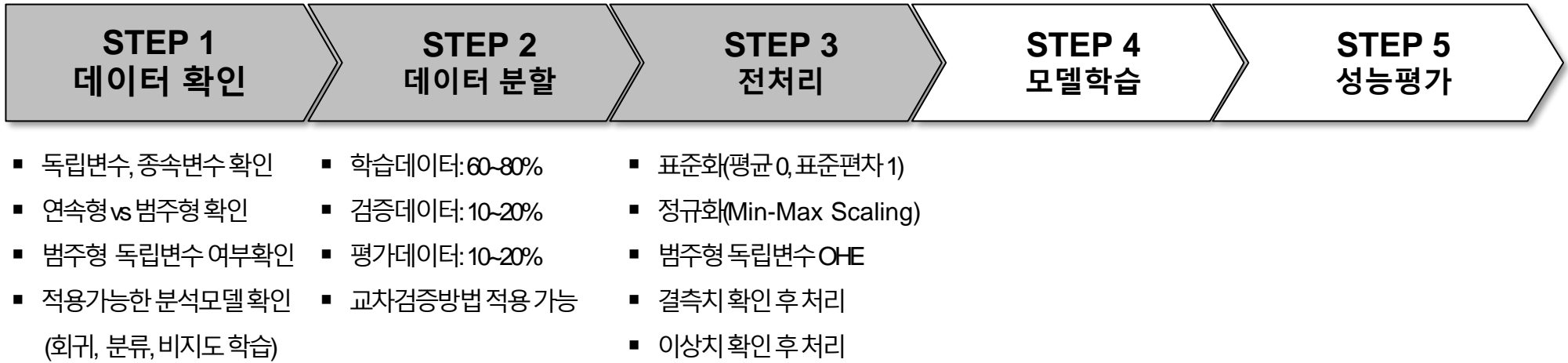
- 학습용 데이터와 평가용 데이터를 분할하는 단계
- 데이터는 학습데이터(60~80%), 검증데이터(10~20%), 평가데이터(10~20%)로 분할
- 예측을 수행하는 데이터 세트는 학습용 데이터 세트가 아니라 평가 전용 데이터세트여야 함
- 단순 학습데이터 + 복잡한 평가데이터의 경우 평가데이터의 특징을 반영하지 못할 수 있음
- 데이터 크기가 작은 경우나, 검증 결과를 일반화하기 위해 교차검증방법을 적용



- 독립변수, 종속변수 확인
- 연속형 vs 범주형 확인
- 범주형 독립변수 여부확인
- 적용가능한 분석모델 확인
(회귀, 분류, 비지도 학습)
- 학습데이터: 60~80%
- 검증데이터: 10~20%
- 평가데이터: 10~20%
- 교차검증방법 적용 가능

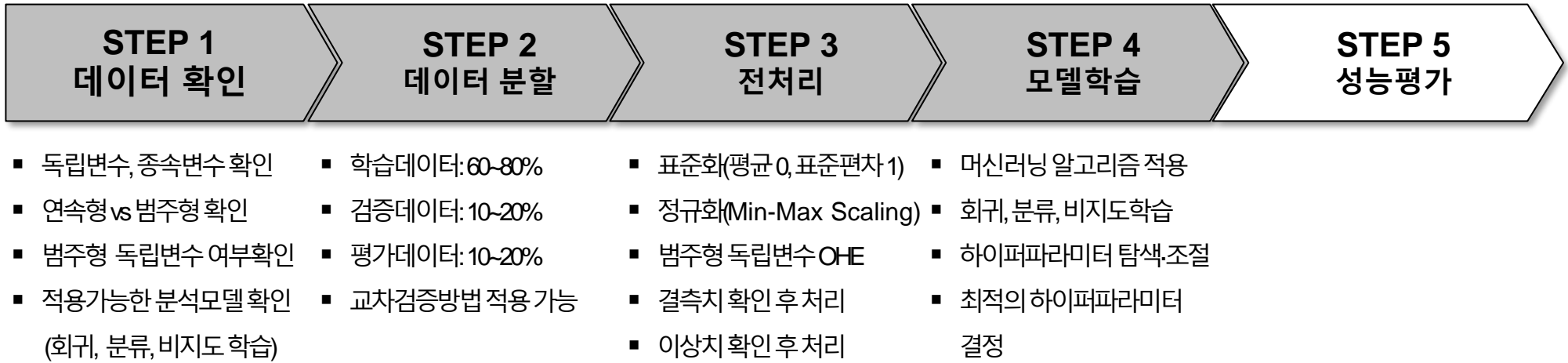
3단계 : 전처리

- 데이터의 특성에 따라 분석이 가능한 형태로 변형하는 단계
- 독립변수에 범주형 변수가 있을 경우 데이터 분할 전 One-hot Encoding으로 데이터를 변형
- 변수마다 단위 특성에 차이가 클 때 분석결과에 영향을 줄 수 있으므로, 정규화나 표준화 실시
- 결측치와 이상치는 분석가의 판단과 도메인 상황에 따라 적절한 방법으로 처리



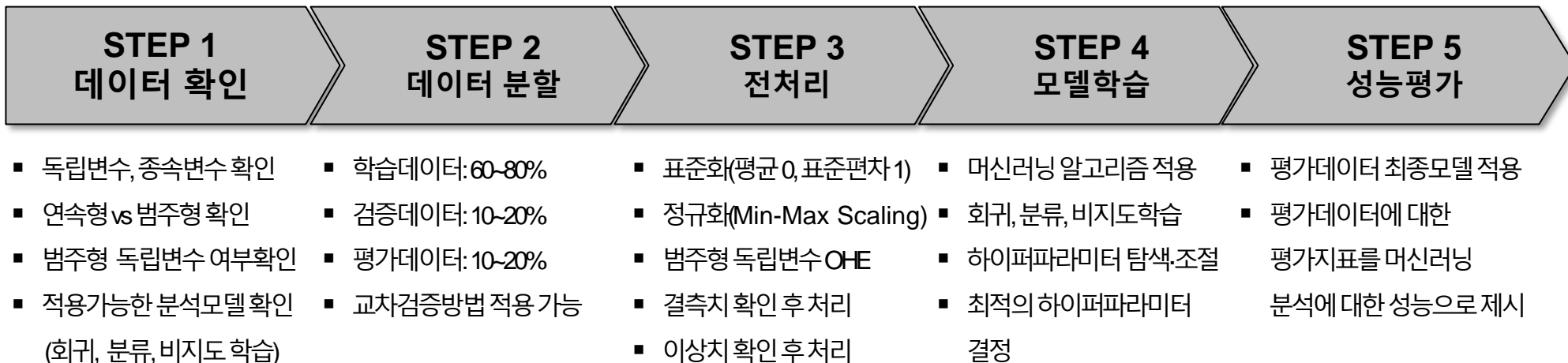
4단계 : 모델학습

- 머신러닝 알고리즘을 학습데이터에 적용하는 단계
- 1단계에서 파악한 분석방법에 따라 적합한 라이브러리를 사용해 머신러닝 알고리즘을 적용
- 머신러닝 분석방법은 지도학습과 비지도학습으로 구분되며, 지도학습은 회귀와 분류로 나뉨
- 학습데이터로 학습을 수행, 검증데이터로 학습결과 확인 후 하이퍼파라미터 탐색 및 조절



5단계 : 성능평가

- 최적의 하이퍼파라미터 및 최종모델 결정 단계
- 최종모델에 평가데이터를 적용하여 머신러닝 알고리즘의 예측성능을 평가
- 평가데이터는 반드시 학습 과정이나 검증 과정에서 사용되지 않은 데이터로 사용해야 함
- 평가데이터에 대한 평가지표를 머신러닝 분석에 대한 최종성능으로 제시





Like every great presentation, I've divided my talk into three subjects. Steve Jobs -

I .

Data Analysis
Basic Theory

II .

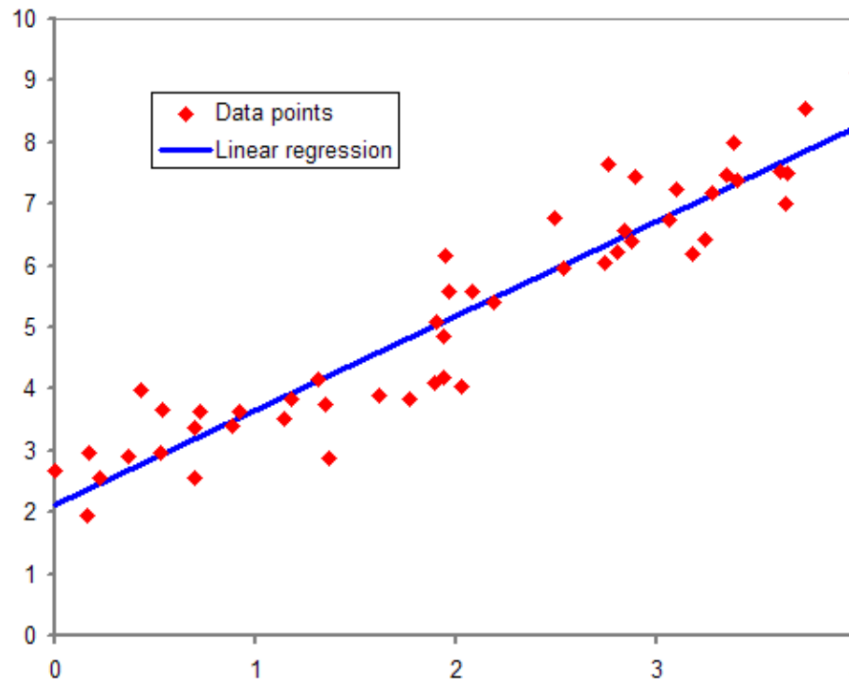
Regression
Analysis

III .

Classification
Analysis

위키백과 : '회귀분석'

- 회귀(regress)의 원래 의미는 옛날 상태로 돌아가는 것을 의미. 영국의 유전학자 프랜시스 골턴은 부모의 키와 아이들의 키 사이의 연관관계를 연구하면서 부모와 자녀의 키 사이에는 선형적인 관계가 있고, 키가 커지거나 작아지는 것보다는 전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세웠으며, 이를 분석하는 방법을 '회귀분석'이라고 함
- 이후 칼 피어슨은 아버지와 아들의 키를 조사한 결과를 바탕으로 함수 관계를 도출하여 회귀분석 이론을 수학적으로 정립



Regression line for 50 random points in a [en:Gaussian distribution](#) around the line $y=1.5x+2$ (not shown). The regression line (shown) that best fits these points is actually $y=1.533858x+2.129333$.

- 단순회귀분석 목적

1

하나의 변수(독립변수, 예측변수)를 이용해서 다른 변수(종속변수, 결과변수)를
예측함

예

영업사원의 수나 판촉행사 횟수, 매장의 면적 등 어떤 특정한 하나의 변수를 이용해서 매출액을 예측함

2

하나의 변수(독립변수, 설명변수)를 이용해서 다른 변수(종속변수, 결과변수)를
설명함

예

가격만족도, 품질만족도 등 어떤 특정한 하나의 변수를 이용해서 전반적인 만족도를 설명함

- 단순회귀분석 회귀식

$$Y = \beta_0 + \beta_1 \cdot X$$

Y : 종속변수 X : 독립변수 β_1 : 회귀계수 β_0 : 상수

예 우리회사 내년도 매출액 규모(Y)를 영업사원 수(X)로 예측

➔ 매출액 = $\beta_0 + \beta_1 \cdot (\text{영업사원 수})$

- 다중회귀분석 목적

1

2개 이상 변수(독립변수, 예측변수)를 이용해서 다른 변수(종속변수, 결과변수)를 **예측**함

예

영업사원의 수, 판촉행사 횟수, 매장의 면적 등 3가지 변수를 이용해서 매출액을 예측함

2

2개 이상 변수(독립변수, 예측변수)를 이용해서 다른 변수(종속변수, 결과변수)를 **설명**함

예

가격만족도, 품질만족도, 디자인만족도, 무게만족도 등 4가지 변수를 이용해서 전반적인 만족도를 설명함

Multiple linear regression

- 다중회귀분석 회귀식

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \dots + \beta_i \cdot X_i$$

Y : 종속변수

β_i : X_i 의 회귀계수

X_i : 독립변수

β_0 : 상수


예

우리회사 내년도 매출액 규모(Y)를 '영업사원 수(X_1), 프로모션 횟수(X_2), 광고비 규모(X_3)'를 이용해 예측하는 다중 회귀식

➡ 매출액 = $\beta_0 + \beta_1$ (영업사원 수) + β_2 (프로모션 횟수) + β_3 (광고비)

회귀식의 설명력 R^2

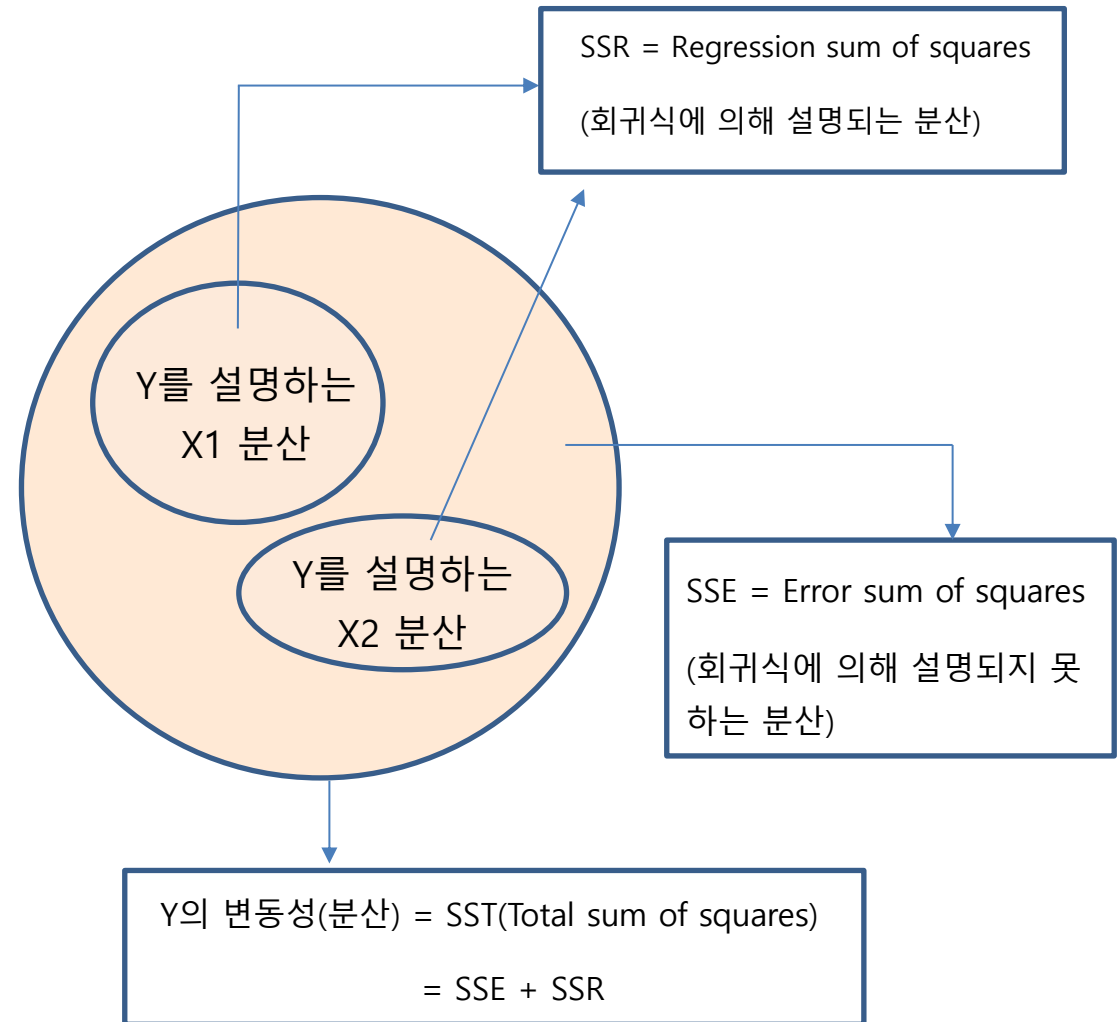
- 회귀식이 종속변수를 설명하고 예측하는데 유용한가를 판단
- 판단지표 : $R^2 = (\text{결정계수, 기여율, 설명력}), 0 < R^2 < 1$
- R^2 은 종속변수의 분산 중 독립변수에 의해 설명되는 비율을 의미

 예 $R^2=0.76$ 이라는 것은 종속변수가 가지는 정보 중에서 76%를 독립변수가 설명할 수 있다는 의미

회귀식의 설명력 R^2

$$R^2 = \frac{SSR}{SST}$$

- 그러나, 변수의 수가 증가하면 SSR이 증가하면서 R^2 도 증가하는 하는 문제가 있음
- R^2 에 변수의 수 만큼 penalty를 주는 지표인 *adjusted* R^2 를 주로 활용



회귀 분석 결과 예시

모형		비표준화 계수		표준화 계수	t	유의 확률	공선성 통계량	
		B	표준오차	베타			공차	VIF
1	(상수)	-.631	.519		-1.215	.235		
	가격만족도	.744	.114	.668	6.528	.000	.298	3.356
	구매 횟수	.302	.094	.331	3.223	.003	.295	3.387
	연령	.011	.011	.054	.962	.345	.983	1.017

a. 종속변수 : 소비자만족도

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

$$\text{소비자만족도} = -0.631 + 0.744 \cdot \text{가격만족도} + 0.302 \cdot \text{구매 횟수} + 0.011 \cdot \text{연령}$$

회귀 분석 결과 예시

모형		비표준화 계수		표준화 계수	t	유의 확률	공선성 통계량	
		B	표준오차	베타			공차	VIF
1	(상수)	-.631	.519		-1.215	.235		
	가격만족도	.744	.114	.668	6.528	.000	.298	3.356
	구매 횟수	.302	.094	.331	3.223	.003	.295	3.387
	연령	.011	.011	.054	.962	.345	.983	1.017

a. 종속변수 : 소비자만족도

- 가격만족도와 구매횟수의 유의확률이 유의수준보다 작으므로($p\text{-value} < 0.05$), 통계적으로 유의미한 변수로 판단
- 연령은 유의확률이 유의수준보다 크므로 ($p\text{-value} > 0.05$), 통계적으로 유의하지 않으며 소비자만족도에는 영향을 미치지 않는 변수로 판단

- 회귀모델의 성능 지표

구 분	개 요	수식
평균절대오차 MAE (Mean Absolute Error)	실제 값과 예측한 값의 차이를 절댓값으로 변환해 평균한 값	$MAE = \frac{1}{n} \sum_{i=1}^n Y_i - \hat{Y}_i $
평균제곱오차 MSE (Mean Squared Error)	실제 값과 예측한 값의 차이를 제곱한 후 평균한 값	$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
평균제곱근오차 RMSE (Root Mean Squared Error)	실제 값과 예측한 값의 차이를 제곱한 후 평균한 값의 제곱근	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$
평균절대비율오차 MAPE (Mean Absolute Percentage Error)	실제 값과 예측한 값의 차이를 백분율로 표현	$MAPE = \frac{100}{n} \sum_{i=1}^n \left \frac{Y_i - \hat{Y}_i}{Y_i} \right $



Like every great presentation, I've divided my talk into three subjects. Steve Jobs -

I .

Data Analysis
Basic Theory

II .

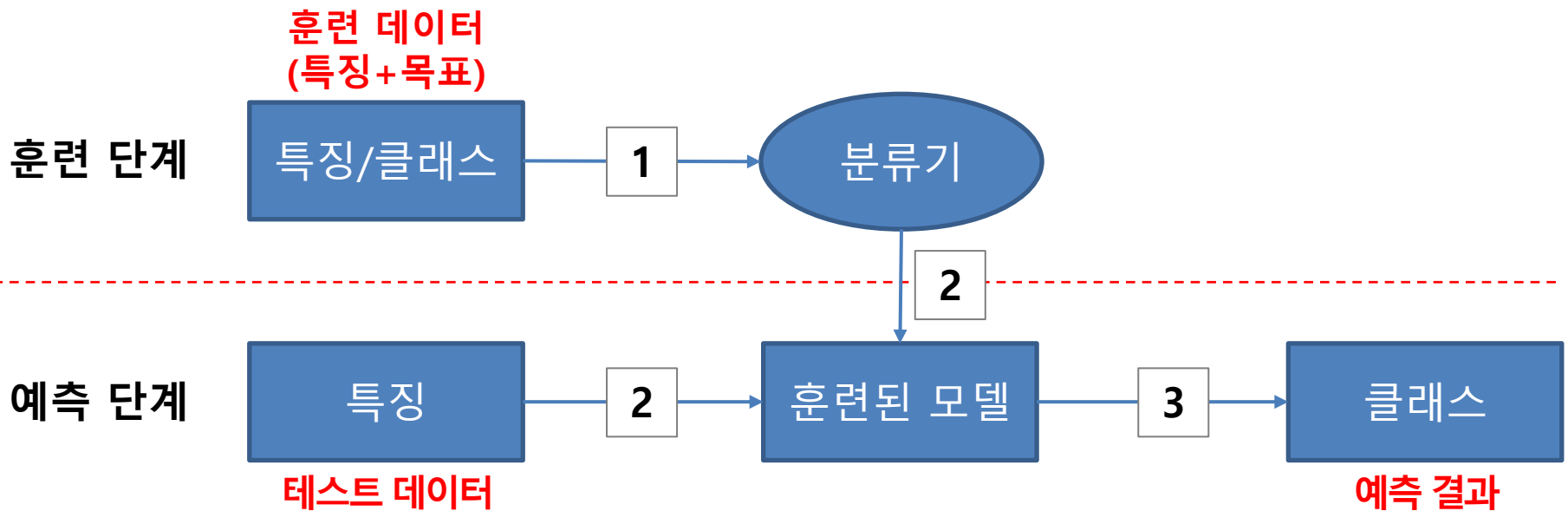
Regression
Analysis

III .

Classification
Analysis

분류의 개념

- 분류(classification)는 지도학습의 하나로 관측값과 해당 관측값에 대한 범주형 출력을 포함하는 훈련데이터셋이 주어졌을 때 관측값을 목표 범주에 올바르게 매핑하는 규칙을 학습하는 것
- 관측값(observation)은 특징(feature) 또는 예측변수라고도 하며, 목표 범주(category)는 레이블(label), 클래스(class) 또는 타겟(target)이라고도 한다.



분류의 종류와 클래스

- 일반적으로 분류는 두개의 클래스로 분류하는 이진 분류(binary classification)와 셋 이상의 클래스로 분류하는 다중 분류(multiclass classification)로 나눌 수 있음
- 이진 분류에서 한 클래스를 양성(positive) 클래스, 다른 하나를 음성(negative) 클래스라 함
- 양성 클래스라고 해서 좋은 값이나 장점을 나타내는 것이 아니고 학습하고자 하는 대상을 의미
- 일반 메일에서 스팸 메일을 골라내는 분석의 경우 스팸메일이 양성 클래스가 되고, 양성 종양과 악성 종양을 분별하는 분석에서는 악성 종양이 양성 클래스가 됨

[일반화, 과대적합, 과소적합]

- ✓ 지도학습에서는 훈련데이터로 학습한 모델이 훈련데이터와 특성이 같다면 새로운 데이터가 주어져도 정확히 예측할 거라 기대
- ✓ 모델이 처음 보는 데이터에 대해 정확하게 예측할 수 있으면 이를 "훈련세트에서 데이터 세트로 일반화" 되었다고 함
- ✓ 과대적합은 모델이 훈련세트의 각 데이터에 너무 맞춰져서 새로운 데이터에 일반화되기 어려움
- ✓ 과소적합은 모델이 너무 간단하여 데이터의 면면과 다양성을 잡아내지 못하고 훈련세트에도 잘 맞지 않음

Confusion matrix

		예측(prediction)	
		양성(Positive)	음성(Negative)
		Positive	Negative

Confusion matrix

		예측(prediction)	
		양성(Positive)	음성(Negative)
실제 (real)	양성	Positive	Negative
	음성	Positive	Negative

		예측(prediction)	
		양성(Positive)	음성(Negative)
실제 (real)	양성	True Positive	False Negative
	음성	False Positive	True Negative

		예측(prediction)	
		양성(Positive)	음성(Negative)
실제 (real)	양성	TP (True Positive)	FN (False Negative)
	음성	FP (False Positive)	TN (True Negative)

		예측(prediction)	
		양성(Positive)	음성(Negative)
실제 (real)	양성	TP	FN
	음성	FP	TN

- 정확도(Accuracy) = (제대로 예측)/(전체) = $(TP+TN)/(TP+FN+FP+TN)$
- 정밀도(Precision) = (실제 양성)/(양성으로 예측) = $TP/(TP+FP)$
- 재현률(Recall) = (양성으로 예측)/(실제 양성) = $TP/(TP+FN)$ = 민감도(Sensitivity)
- 특이도(Specificity) = (음성으로 예측)/(실제 음성) = $TN/(TN+FP)$
- 거짓양성율(FPR) = 1 - 특이도
- F1 score = $2 \times \text{정밀도} \times \text{재현률} / (\text{정밀도} + \text{재현률})$

Algorithms → 분류와 회귀 모두 가능한 알고리즘이 많이 있음

● 의사결정 나무 (Decision Tree)

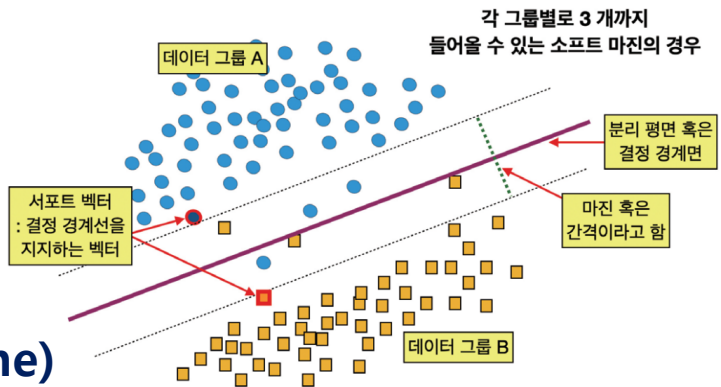
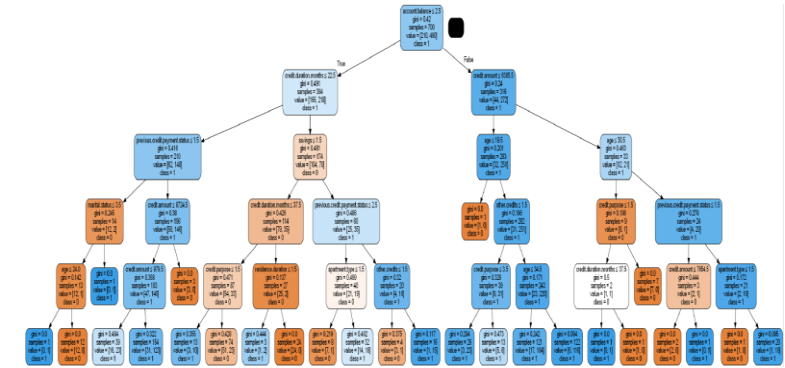
● 앙상블 모형 (Ensemble)

1. Bagging

2. Boosting

- AdaBoost (Adaptive Boosting)
- GBM (Gradient Boosting Machine)
- XGBoost
- LightGBM
- CatBoost

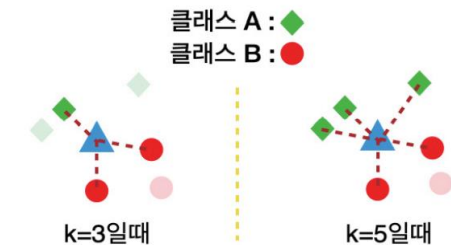
3. Random Forest



● 서포트 벡터 머신 (SVM; Support Vector Machine)

● K 최근접 이웃 (K-Nearest Neighbor)

● 소프트맥스 (Softmax) 회귀 → 다항 로지스틱 회귀라고도 함





MS Excel vs Google Sheets



Google Sheets