AI융합비즈니스포럼

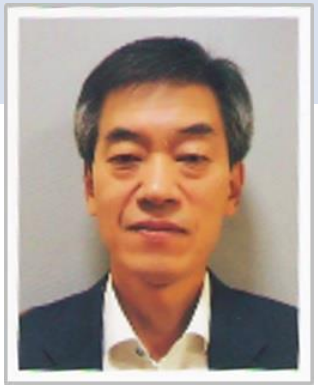
AI 융합전문가 제10기

데이터분석 II

2025.03.02.

강사 홍용기 컨설팅학 박사

현재 하는 일



https://landing.koex.kr

소속	업무 분야
㈜제타데이터 대표이사	데이터 분석, 전략 및 컨설팅, 데이터 가치평가 ODA 컨설팅 (Official Development Assistance)
㈜지구파트너스 감사	창업보육, 투자, 기업·기술가치평가, 사업타당성 분석
㈜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)

강사 홍용기 컨설팅학 박사

데이터 관련 비즈니스

◎ 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, 젖소) 및 돼지 발정행동 데이터
3차 06 인공지능 신기술 선도(자유 공모)

◎ 데이터 가치평가 컨설팅

23.09~23.11 중소벤처기업진흥공단

발급번호:78TP-0BH6-0048-FA6O-C525

스마트폰으로 QR코드를 스캔하시면 경력증명서 진위확인페이지로 이동됩니다.

소프트웨어기술자 경력증명서

성명	홍용기			생년월일	1964.08.25	
현 근무처	회사명			사업자등록번호		
	전화번호			업종		
	소재지					
근무경력	확인여부	근무기간	회사명	담당업무	부서/직위	
기술자격	종목 및 등급		등록번호	취득일	발급기관	
	빅데이터분석기사		BAE-002000023	2021.07.16	한국데이터산업진흥원	
	ADsP(데이터 분석 준전문가)		ADsP-028000961	2021.04.09	한국데이터산업진흥원	
학력	학교명		학과(전공)	수학기간	학위	
교육	기간		과정	수료번호	교육기관	
상훈	수여일		종류	상훈기관	근거	

기술경력	확인여부	참여사업명	참여기간	발주자	소속사	직위	담당업무
	확인	벤처기업 육성플랫폼 구축전략 수립	2023.08.14 ~ 2023.12.13	벤처기업 확인기관	(주)메타로직컨설팅	수석컨설턴트	IT컨설팅 > 정보기술기획
	확인	요르단 빅데이터 기반 치안정보 통합관리시스템 구축 ISMP 용역	2023.08.02 ~ 2024.05.31	경찰대학	(주)메타로직컨설팅	수석컨설턴트	IT컨설팅 > 빅데이터기획
	확인	국민 체감형 치안 안심 플랫폼 구축 정보화전략계획 사업	2022.08.04 ~ 2022.12.31	경찰대학교	(주)메타로직컨설팅	프리랜서	IT컨설팅 > 정보기술기획
	확인	서울시 차세대 지방세입정보시스템 통합구축 변화관리 컨설팅	2021.11.01 ~ 2022.06.30	서울특별시	(주)메타로직컨설팅	프리랜서	IT컨설팅 > 정보기술컨설팅
	확인	소상공인지원사업 디지털전환 정보화전략계획(ISP) 용역	2021.01.04 ~ 2021.05.03	소상공인시장진흥공단	(주)메타로직컨설팅	수석컨설턴트	IT컨설팅 > 정보기술컨설팅
	확인	한국연구재단 중장기 정보화전략계획(ISP) 수립	2020.10.05 ~ 2021.01.04	한국연구재단	(주)메타로직컨설팅	수석컨설턴트	IT컨설팅 > 정보기술컨설팅
	확인	창업기업확인시스템 구축을 위한 정보화전략계획(ISP) 수립	2020.05.18 ~ 2020.08.17	창업진흥원	(주)메타로직컨설팅	수석컨설턴트	IT컨설팅 > 정보기술컨설팅

강사 홍용기 컨설팅학 박사

데이터분석 관련 강의

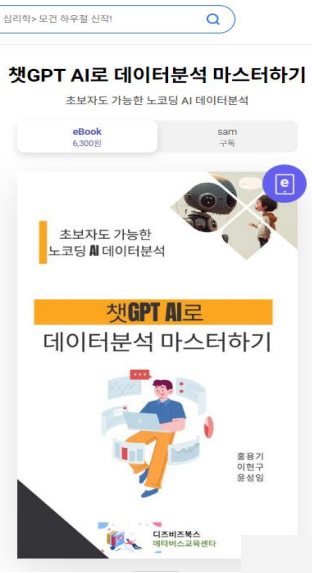
- 데이터분석 및 실전 R코딩 (경영기술지도사회, 빅데이터 분석기사 자격증 취득 과정)
- 데이터분석 Python 심화과정 (서울 여성능력개발원 강동 여성인력개발센터 / 용산 여성인력개발센터)
- 파이썬 코딩을 통한 크롤링 자동화 인텐시브 과정 (경영지도사 및 컨설턴트)
- AI & ChatGPT 활용 및 데이터분석 컨설팅 방법론 (경영기술지도사회, 국제공인컨설턴트 CMC 양성과정)
- AI & 데이터분석 (매경아카데미, 동북아 ICT 포럼)

AI융합비즈니스포럼

강사 홍용기 컨설팅학 박사

책 쓰기 프로젝트

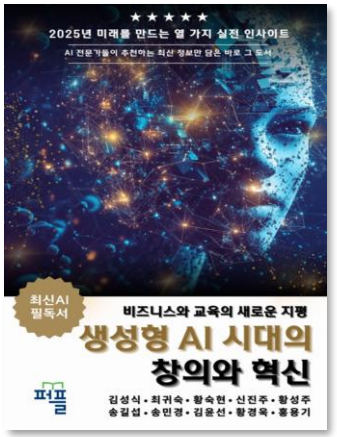
2018년 1월



2023년 5월

2024년 4월

2024년 06월



2024년 11월

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

Ⅰ. Understanding of Web & Internet

Ⅱ. Data Analysis Basic Theory

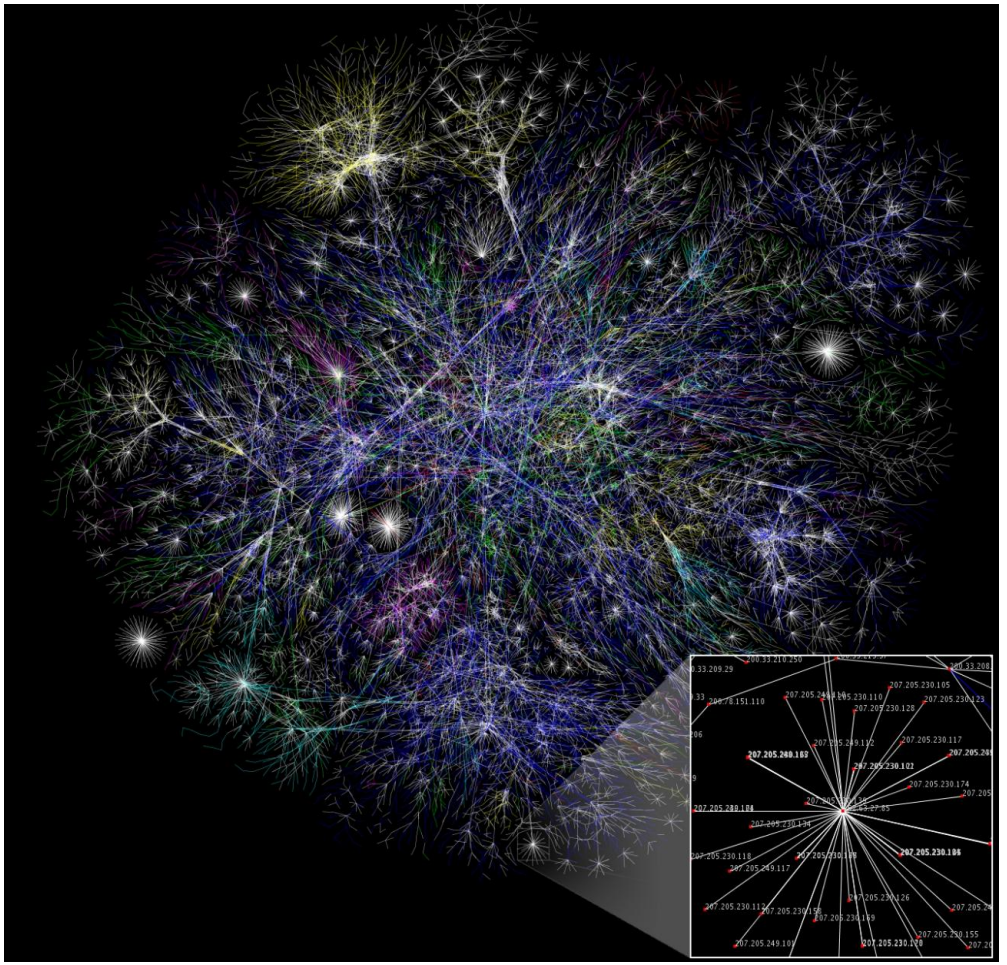
Ⅲ. Classification & Regression Analysis

Internet

인터넷은 인간이 발명해 놓고도 이해하지 못하는 최초의 발명품이며, 역사상 최대 규모의 무정부주의에 대한 실험이다.

The Internet is the first thing that humanity has built that humanity doesn't understand, the largest experiment in anarchy that we have ever had.

- Eric Emerson Schmidt



※라우터를 통해 연결된 인터넷을 시각화한 그림(위키백과)

Internet

인터넷(Internet)은
인터넷 프로토콜 스위트(TCP/IP)를 기반으로 하여 전 세계적으로
연결되어 있는 컴퓨터 네트워크 통신망을 일컫는 말이다.
그야말로 인류의 역사상 전례 없는 거대한 정보의 바다인 셈이다.

흔히 웹(WEB)이라고 줄여 부르는
월드 와이드 웹(World Wide Web; WWW)만 생각하기 쉽지만
인터넷은 월드 와이드 웹, 전자 메일, 파일 공유(토렌트, eMule 등),
웹캠, 동영상 스트리밍, 온라인 게임, VoIP, 모바일 앱 등
다양한 서비스들을 포함한다.

※출처: 나무위키(https://namu.wiki/인터넷)

WWW의 탄생

1989년 3월, CERN(유럽 입자 물리 연구소)의
소프트웨어 공학자 팀 버너스리는 CERN에서 인사 재배치 등으로
기존에 수행했던 실험 결과를 비롯한 각종 문서들이 유실되는 비율이
높은 것을 보고 이를 줄이기 위해
Information System: A Proposal을 제안하였다.

또한 여러 연구기관에 흩어져 있는 문서들을 체계화하여
전 세계의 대학 및 연구소들끼리 정보를 신속하게 교환할 수 있도록
해야 한다고 판단하여 문서 뿐만 아니라 소리, 동영상 등을 망라하는
데이터베이스를 구축하고 이를 전문 열람 소프트웨어로 열람하는
방식을 생각해 냈다.

인터넷과 WWW

워낙 WWW가 대세이기에 WWW를 인터넷으로 착각하는 경우가 많지만, 웹은 TCP/IP 기반 물리적 통신망인 인터넷을 활용한 서비스로 인터넷의 하위 개념으로 볼 수 있다.

위키백과에 따르면 WWW은 다음 세 가지의 기능으로 요약할 수 있음

첫 번째, 통일된 웹 자원의 위치 지정 방법 ➔ 예를 들면 URL(Uniform Resource Locator)

두 번째, 웹의 자원 이름에 접근하는 프로토콜(protocol) ➔ 예를 들면 HTTP

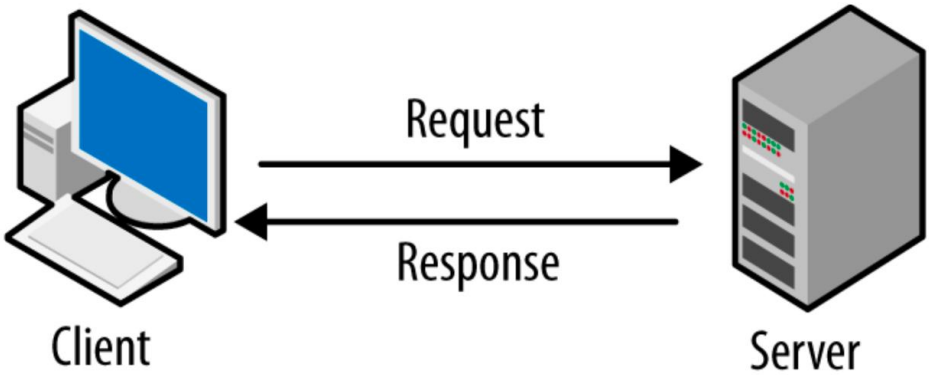
세 번째, 자원들 사이를 쉽게 항해할 수 있는 언어 ➔ 예를 들면 HTML

※ 세계 최초의 웹사이트: https://info.cern.ch/

HTTP

Hyper Text Transfer Protocol은 W3 상에서 정보를 주고받을 수 있는 프로토콜(규약)

- HTTP는 클라이언트와 서버 사이에 이루어지는 요청/응답 (request/response) 프로토콜(규약)
- 클라이언트인 웹 브라우저가 HTTP를 통하여 서버로부터 웹페이지(HTML)나 그림 정보를 요청하면, 서버는 이 요청에 응답하여 필요한 정보를 해당 사용자에게 전달
- 이 정보가 모니터와 같은 출력 장치를 통해 사용자에게 나타나는 것서 흔히 볼 수 있는 htm이나 html 확장자가 바로 이 언어로 작성된 문서



※그림 출처: https://velog.io/@seosu2000/Client-Server란 무엇인가

HTML

< > ••• </>

https://namu.wiki/w/HTML

HTML

웹사이트의 모습을 기술하기 위한 마크업 언어

- 프로그래밍 언어가 아니라 마크업 정보를 표현하는 마크업 언어로 문서의 내용 이외의 문서의 구조나 서식 같은 것을 포함
- HTML의 ML이 마크업 언어라는 뜻으로 웹사이트에서 흔히 볼 수 있는 htm이나 html 확장자가 바로 이 언어로 작성된 문서

```
<!DOCTYPE html>
<html>
	<head>
		<meta charset="utf-8">
	</head>
	<body>
		Hello, world!
	</body>
</html>
```

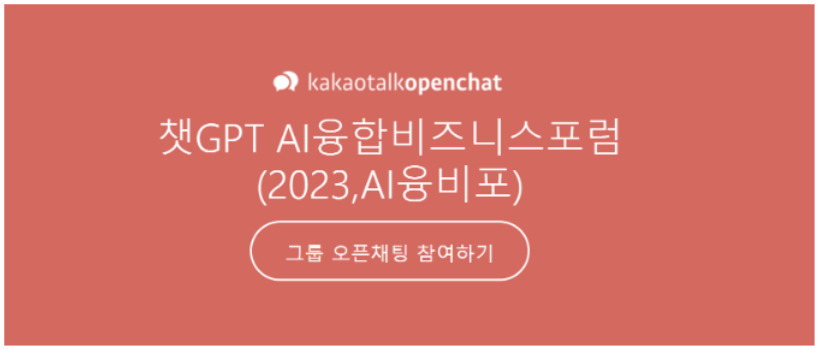
HTML

AI 융합전문가 제5기 기초과정

1. 챗GPT활용 컨설팅 방법론
2. 브랜드 전자책 쓰기
3. 이미지/동영상 AI콘텐츠 크리에이터 되기
4. 데이터분석 기초와 크롤링

AI 융합전문가 과정은 시시각각 변화하는 AI의 비즈니스 활용 역량을 향상시키고, 학습과 비즈니스를 같이 하며 서로 윈윈할 수 있는, 집단지성 커뮤니티 추구목적의 커리큘럼입니다(AI융합비즈니스포럼 연계).

강사진
- 이현구
- 윤성임
- 김성식
- 홍용기



```
<!DOCTYPE html>
<html lang="en>
<head>
    <meta charset="UTF-8">
    <title>Expert profile</title>
</head>
<body>
    <header>
        <h1> AI 융합전문가 제5기 기초과정 </h1>
    </header>
        <ol>
            <li><a href= '1강.html' >챗GPT활용 컨설팅 방법론 </li></a>
            <li><a href= '2강.html' >브랜드 전자책 쓰기 </li ></a>
            <li><a href= '3강.html' > 데이터분석 기초와 크롤링 </li ></a>
            <li><a href= '4강.html' > 이미지/동영상 AI콘텐츠 크리에이터 되기 </li ></a>
        </ol>
        <p> AI 융합전문가 과정은 시시각각 변화하는 AI의 비즈니스 활용 역량을 향상시키고, 학습과 비즈니스를 같이 하며 서로 윈윈할 수 있는, 집단지성 커뮤니티 추구 목적의 커리큘럼입니다 (<a href=https://open.kakao.com/o/gOS6mw8e target="_blank" title="AI 융합을 통해 집단지성을 추구합니다.">AI융합비즈니스포럼 연계</a>). </p>

        <h2> 강사진(가나다 순) </h2>
        <ul>
            <li> 김성식 </li>
            <li> 윤성임 </li>
            <li> 이현구 </li>
            <li> 홍용기 </li>
        </ul>
        <img src="forum.png", alt="포럼 대문 이미지">
    <footer>
    </footer>
</body>
</html>
```

```
<!DOCTYPE html>
<html>
    <head>
        <meta charset="utf-8">
    </head>
    <body>
        Hello, world!
    </body>
</html>
```

HTML은 '정보전달' 이 주목적

디자인 요소

UI

HTML은 '정보전달'에 충실

디자인 요소

CSS (Cascading Style Sheet)

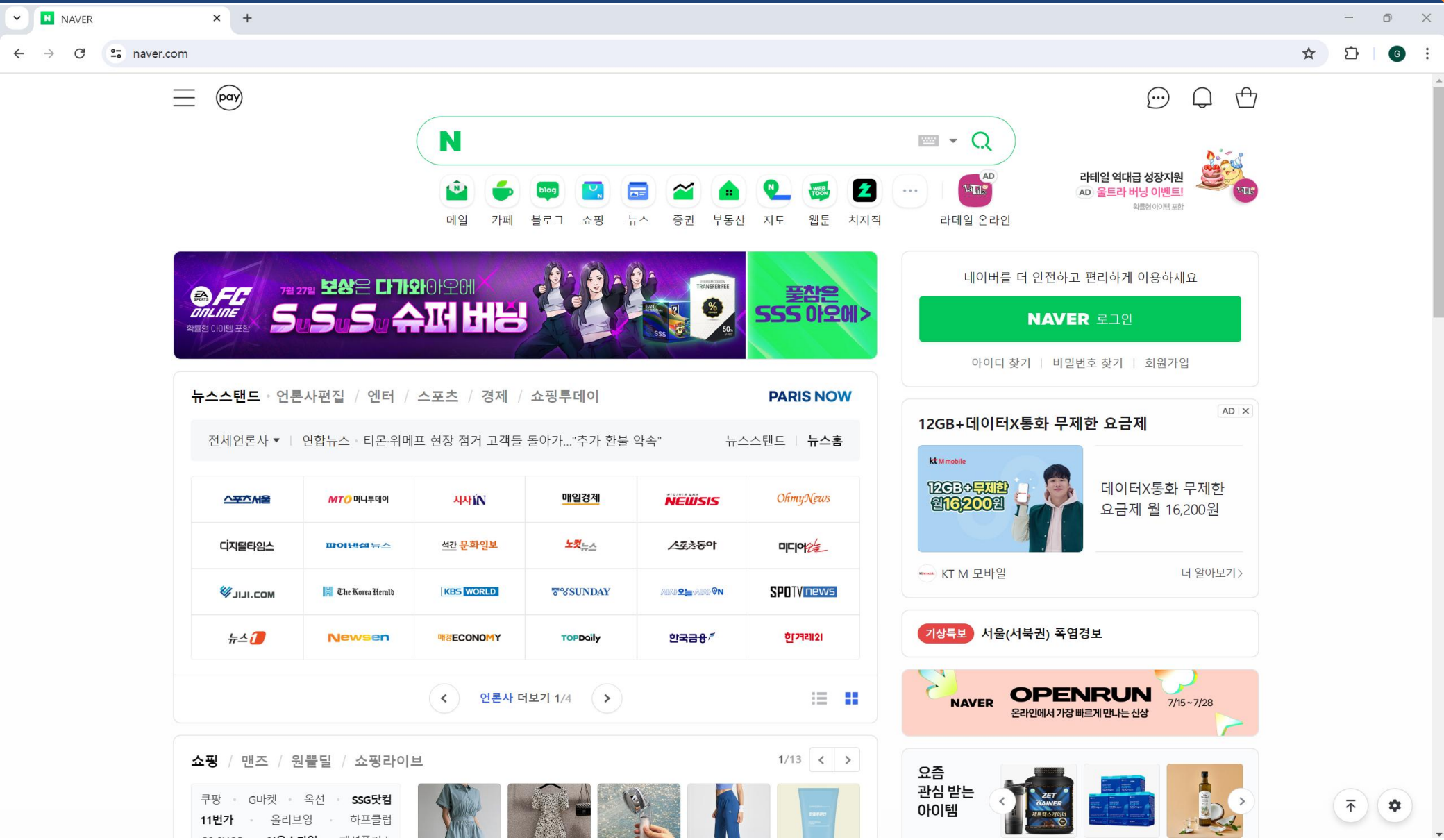
UI

JavaScript

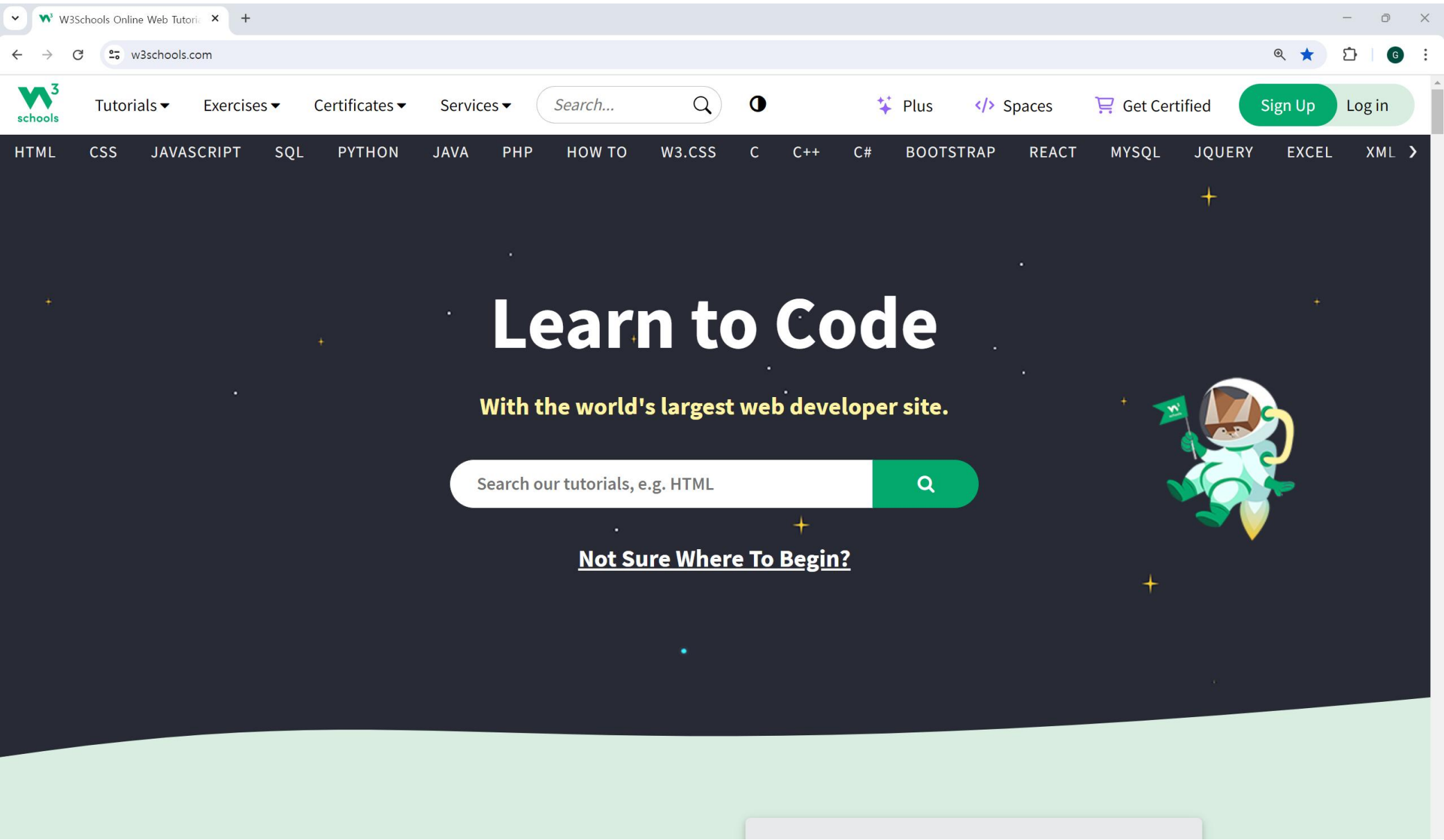
검색 엔진 최적화(SEO)

검색엔진최적화(Search Engine Optimization)는 검색엔진으로부터 웹사이트나 웹페이지에 대한 트래픽의 품질과 양을 개선하는 과정

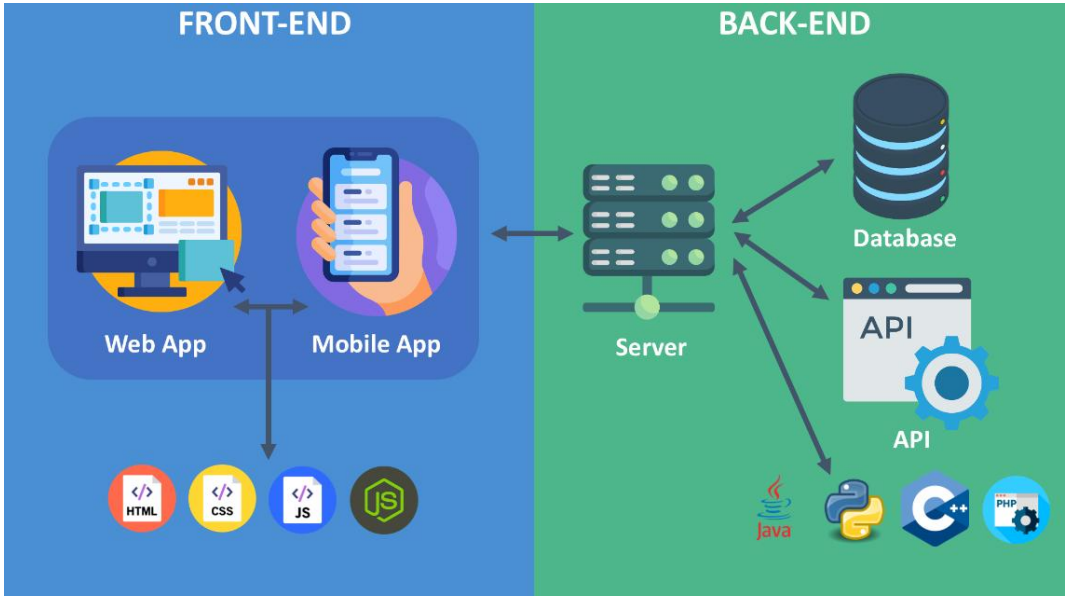
- 웹 페이지 검색엔진이 자료를 수집하고 순위를 매기는 방식에 맞게 웹 페이지를 구성해서 검색 결과의 상위에 나올 수 있게 함
- 웹 페이지와 관련된 검색어로 검색한 검색 결과 상위에 나오게 된다면 방문 트래픽이 늘어나기 때문에 효과적인 인터넷 마케팅 방법 중의 하나이며 비용처리 없는 마케팅이라고 할 수 있음
- 기본적인 작업 방식은 특정한 검색어를 웹 페이지에 적절하게 배치하고 다른 웹 페이지에서 링크가 많이 연결되도록 하는 것
- 구글 등장 이후 검색 엔진들이 콘텐츠의 신뢰도를 파악하는 기초적인 지표로 다른 웹사이트에 얼마나 인용되었나를 사용하기 때문에 타 사이트에 인용되는 횟수를 늘리는 방향으로 최적화함


NAVER
naver.com
pay
N
메일
카페
블로그
쇼핑
뉴스
증권
부동산
지도
웹툰
치지직
라테일 온라인
라테일 역대급 성장지원
AD 울트라 버닝 이벤트!
확률형 아이템 포함
7월 27일 보상은 다가와 아오에
SSS 슈퍼 버닝
풀참은 SSS 아오에>
네이버를 더 안전하고 편리하게 이용하세요
NAVER 로그인
아이디 찾기 | 비밀번호 찾기 | 회원가입
뉴스스탠드 · 언론사편집 / 엔터 / 스포츠 / 경제 / 쇼핑투데이
PARIS NOW
전체언론사 | 연합뉴스 · 티몬·위메프 현장 점거 고객들 돌아가..."추가 환불 약속"
뉴스스탠드 | 뉴스홈
언론사 더보기 1/4
12GB+데이터X통화 무제한 요금제
AD
데이터X통화 무제한
요금제 월 16,200원
KT M 모바일
더 알아보기>
기상특보 서울(서북권) 폭염경보
NAVER OPENRUN
온라인에서 가장 빠르게 만나는 신상
7/15~7/28
쇼핑 / 맨즈 / 원쁠딜 / 쇼핑라이브
1/13
쿠팡 · G마켓 · 옥션 · SSG닷컴
11번가 · 올리브영 · 하프클럽
요즘 관심 받는 아이템

WWW


W3Schools Online Web Tutori
w3schools.com
schools
Tutorials
Exercises
Certificates
Services
Search...
Plus
Spaces
Get Certified
Sign Up
Log in
HTML CSS JAVASCRIPT SQL PYTHON JAVA PHP HOW TO W3.CSS C C++ C# BOOTSTRAP REACT MYSQL JQUERY EXCEL XML
Learn to Code
With the world's largest web developer site.
Search our tutorials, e.g. HTML
Not Sure Where To Begin?

웹개발은 크게 기본, 프론트엔드, 백엔드 등 3가지 영역으로 구분



➔ 사례를 통한 이해 https://landing.koex.kr

➔ 깃허브 데이터 가져오기 https://github.com/GENEXIS-AI/prompt-gallery

※그림 출처: https://velog.io/@xenxxn/01, https://1.pearlvely.com/5

Break

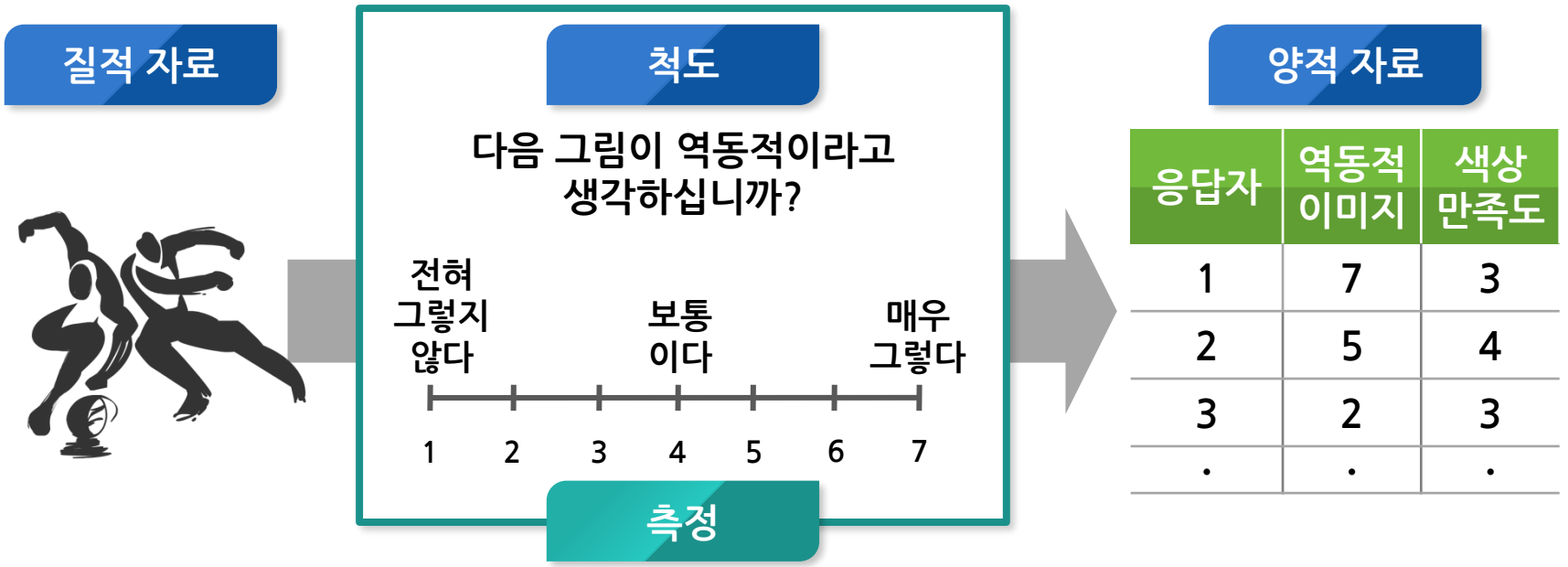
Like every great presentation, I've divided my talk into three subjects. Steve Jobs -

Ⅰ. Understanding of Web & Internet

Ⅱ. Data Analysis Basic Theory

Ⅲ. Classification & Regression 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 구분하기 위해 임의의 숫자를 부여하는 척도
서열척도	응답보기들을 구분하고, 구분한 응답보기들의 순서까지 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 구분하기 위해 임의의 숫자를 부여하는 척도
- 선택한 응답을 기준으로 응답자들을 특정 집단으로 분류하기 위해 사용(=분류정보)

> 예
>
> **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 구분하고, 구분한 응답보기들의 순서까지 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

> 예 **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 크기를 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 자체는 절대적 의미를 가지지 않음
- 절대 영점이 없기 때문에 숫자 간 비율개념 없음
- 간격 척도에서 숫자 간의 차이는 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

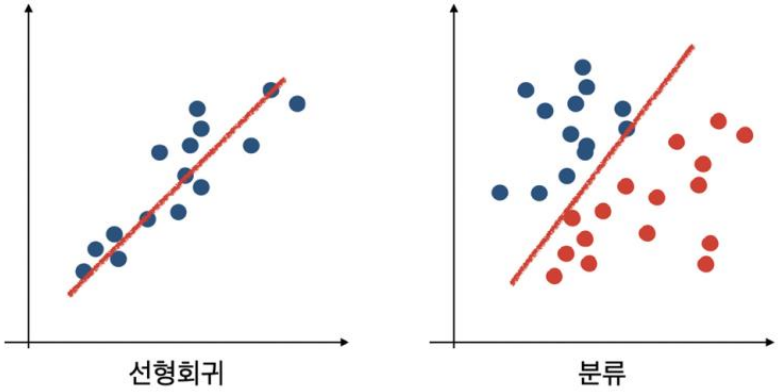
예 나이 ( )세, 근무기간( )년, 연봉( )원

예 하루에 보통 몇시간 정도 훈련이 가능하십니까? ( )시간

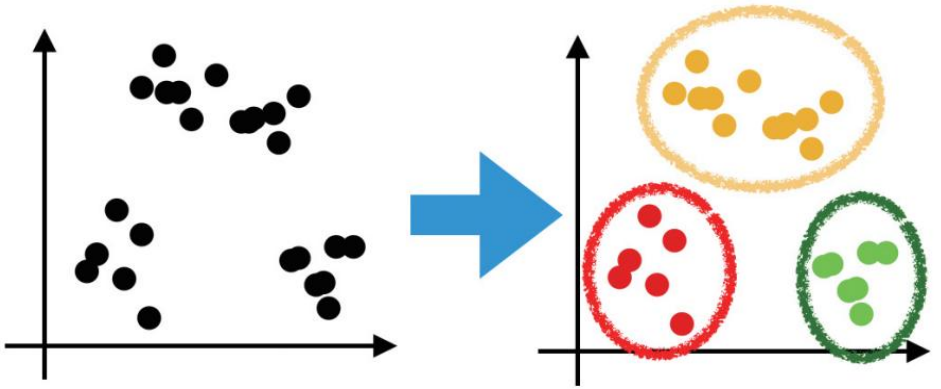
- 만족도, 선호도, 인지도 등 절대 영점이 존재하기 어려운 소비자의 사고나 인지수준에 대한 측정은 한계가 있음
- 직접 관찰할 수 있는 물리적 사건이나 현상을 측정하는데 주로 사용함
- 사칙연산이 가능하며, 대표치는 평균값
- 4가지 척도 중 가장 정보량이 많은 척도 : 분류정보 + 순서정보 + 상대적 크기 정보 + 절대적 크기 정보

Supervised learning VS Unsupervised learning

지도 학습supervised learning



비지도 학습unsupervised learning



※출처: 으뜸 데이터 분석과 머신러닝(박동규·강영민, 2021)

Data split

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	y

Data split

Training data

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	y
					X_train								y_train

y_predict

Test data

						X_test							y_test

Data split

Training data

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	y
					X_train								y_train

Validation data

y_predict

					X_val								y_val

Test data

					X_test								y_test

교차 검증의 의의

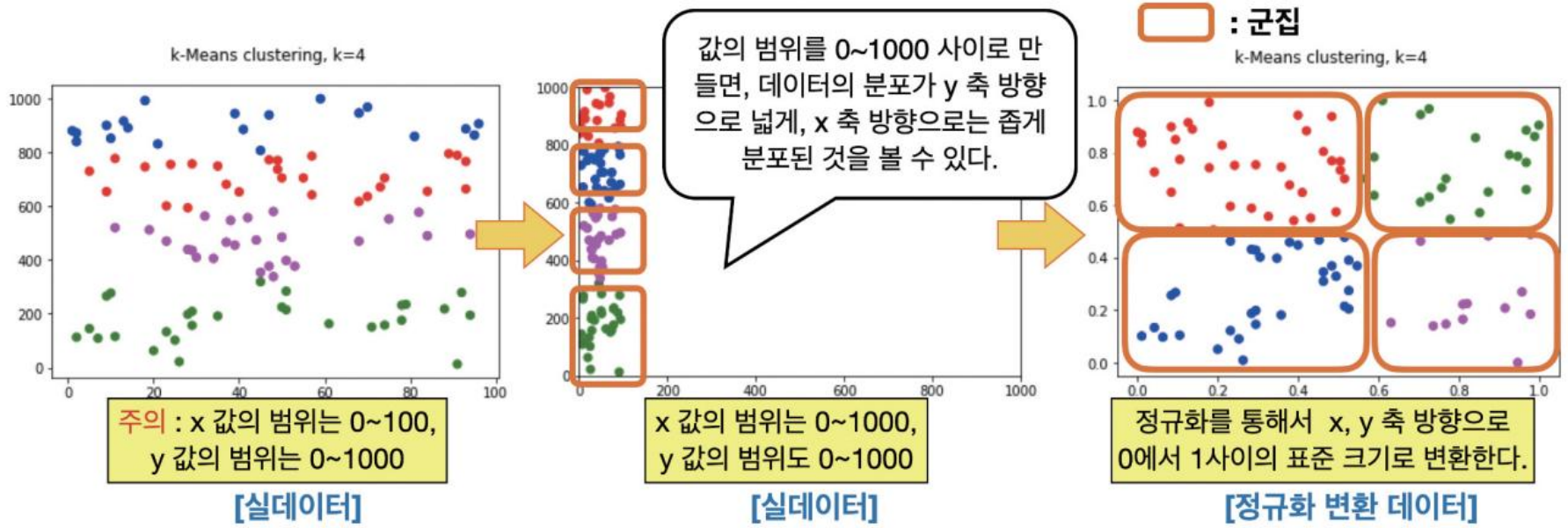
- 교차 검증은 일반화 성능을 재기 위해 훈련 세트와 검증 세트로 한 번 나누는 것보다 더 안정적이고 뛰어난 통계적 평가 방법
- 교차 검증에서는 데이터를 여러 번 반복해서 나누고 여러 모델을 학습
- 가장 널리 사용되는 교차 검증 방법은 k-겹 교차 검증(k-fold CV)으로 보통 5 또는 10을 사용
- 5-겹 교차 검증을 하면 데이터를 비슷한 크기의 부분 집합(5개의 폴드)으로 나누고, 일련의 모델을 만들어 훈련과 테스트를 반복

※출처: 파이썬 라이브러리를 활용한 머신러닝(번역개정판, 안드레아스뮐러 & 세라가이도, 2021.7)

Missing value : NA, NAN, Null

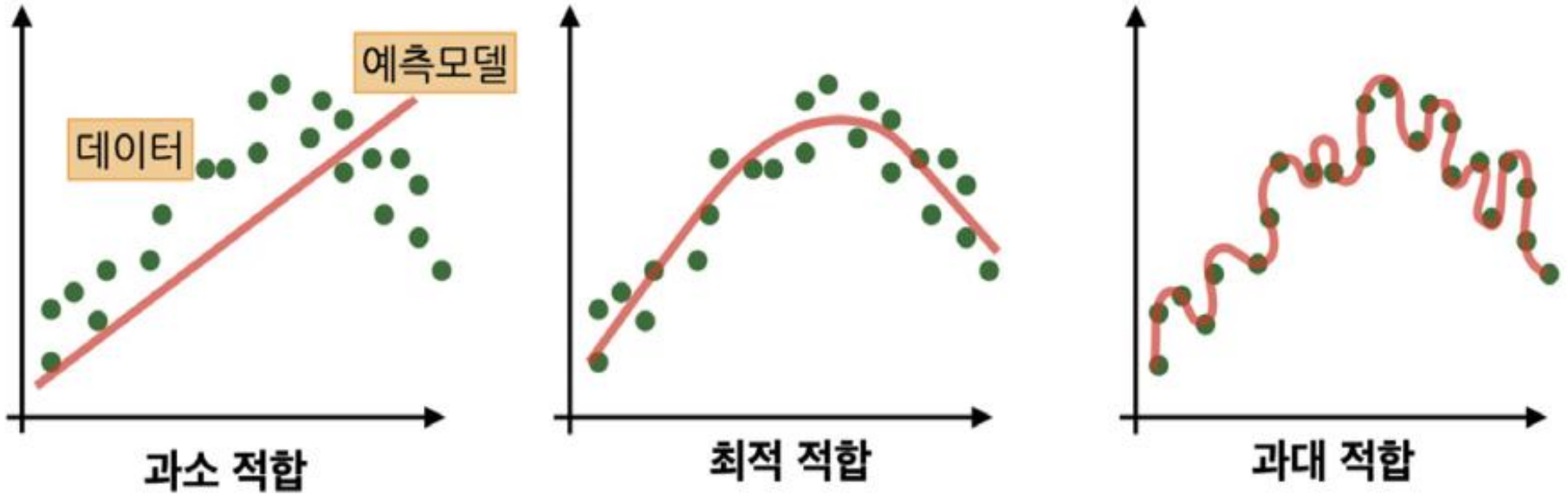
➔ 결측치 제거 또는 대체 (평균, 중위수, 최빈값)

정규화normalization(min-max scale), **표준화**standardization(평균=0, 분산=1로 만듦)



※출처: 으뜸 데이터 분석과 머신러닝(박동규·강영민, 2021)

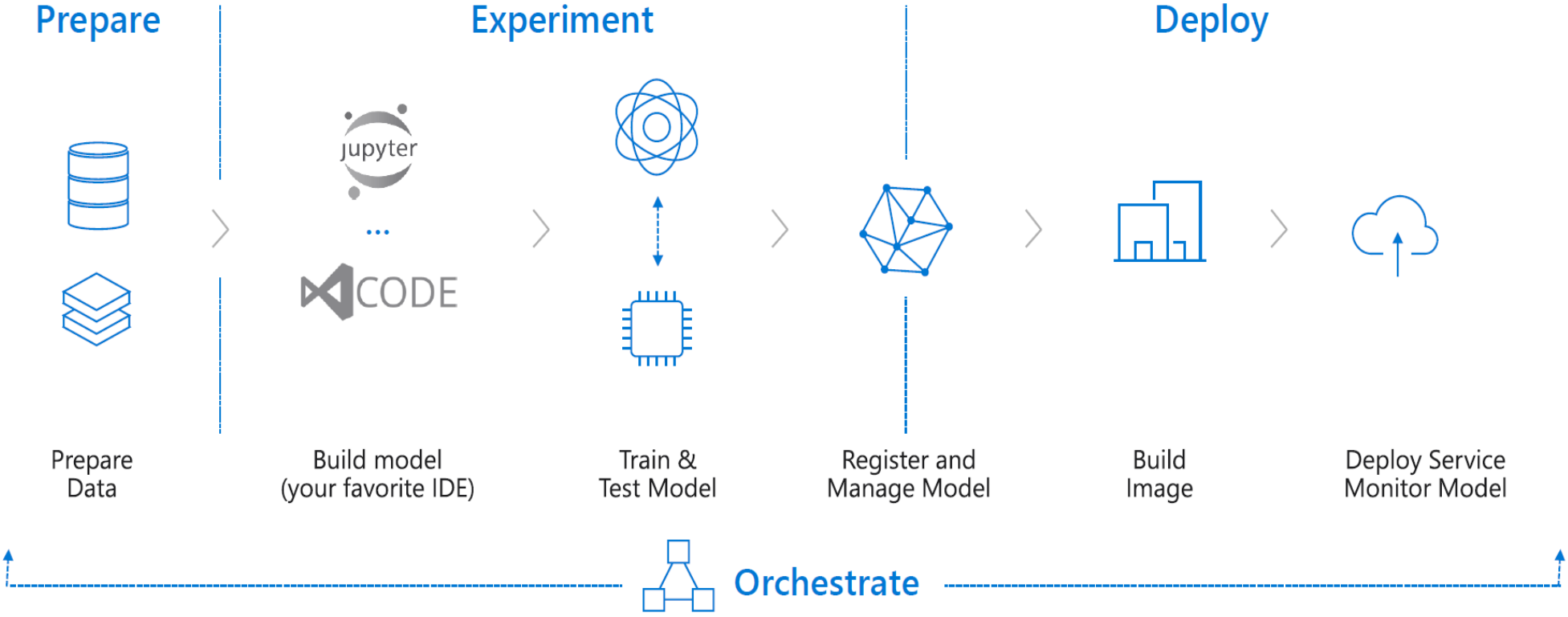
Underfitting, Optimal fitting, Overfitting



※출처: 으뜸 데이터 분석과 머신러닝(박동규·강영민, 2021)

Machine Learning

Typical E2E Process



※출처: Chat-GPT / Open AI가 제시하는 New Digital Experience

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

STEP 1 데이터 확인	STEP 2 데이터 분할	STEP 3 전처리	STEP 4 모델학습	STEP 5 성능평가
▪ 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	▪ 표준화(평균 0, 표준편차 1) ▪ 정규화(Min-Max Scaling) ▪ 범주형 독립변수 OHE ▪ 결측치 확인 후 처리 ▪ 이상치 확인 후 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

STEP 1 데이터 확인	STEP 2 데이터 분할	STEP 3 전처리	STEP 4 모델학습	STEP 5 성능평가
▪ 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	▪ 표준화(평균 0, 표준편차 1) ▪ 정규화(Min-Max Scaling) ▪ 범주형 독립변수 OHE ▪ 결측치 확인 후 처리 ▪ 이상치 확인 후 처리	▪ 머신러닝 알고리즘 적용 ▪ 회귀, 분류, 비지도학습 ▪ 하이퍼파라미터 탐색·조절 ▪ 최적의 하이퍼파라미터 결정	

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

STEP 1 데이터 확인	STEP 2 데이터 분할	STEP 3 전처리	STEP 4 모델학습	STEP 5 성능평가
▪ 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	▪ 표준화(평균 0, 표준편차 1) ▪ 정규화(Min-Max Scaling) ▪ 범주형 독립변수 OHE ▪ 결측치 확인 후 처리 ▪ 이상치 확인 후 처리	▪ 머신러닝 알고리즘 적용 ▪ 회귀, 분류, 비지도학습 ▪ 하이퍼파라미터 탐색·조절 ▪ 최적의 하이퍼파라미터 결정	▪ 평가데이터 최종모델 적용 ▪ 평가데이터에 대한 평가지표를 머신러닝 분석에 대한 성능으로 제시

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

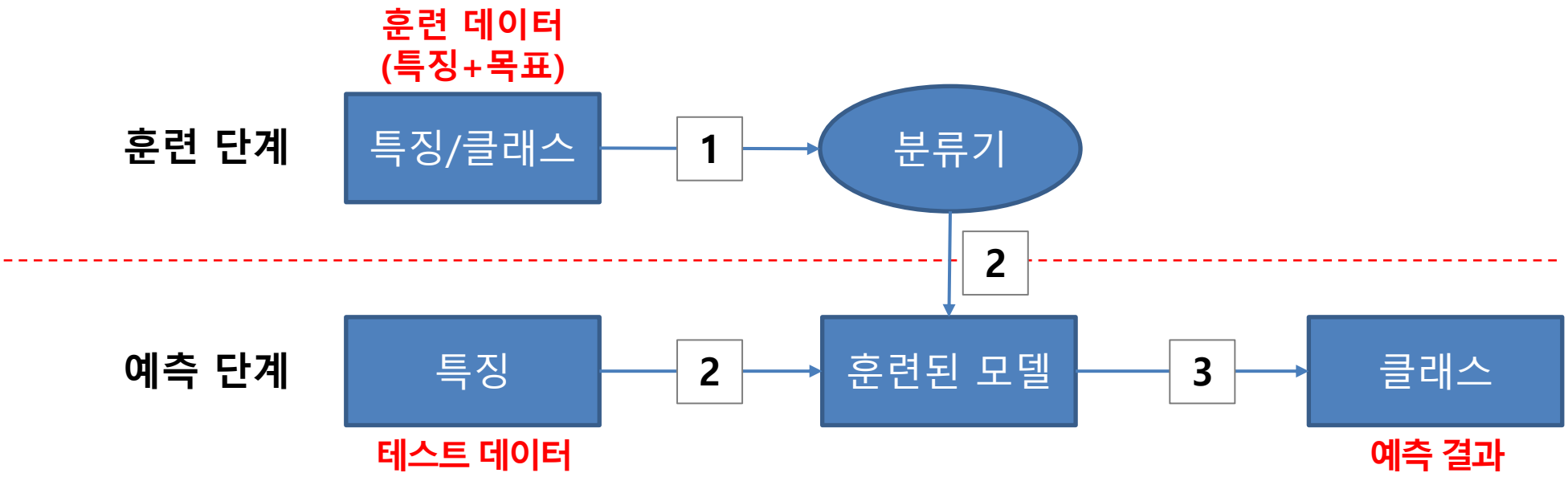
Ⅰ. Understanding of Web & Internet

Ⅱ. Data Analysis Basic Theory

Ⅲ. Classification & Regression 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 dassification)와 셋 이상의 클래스로 분류하는 다중 분류(multidass dassification)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

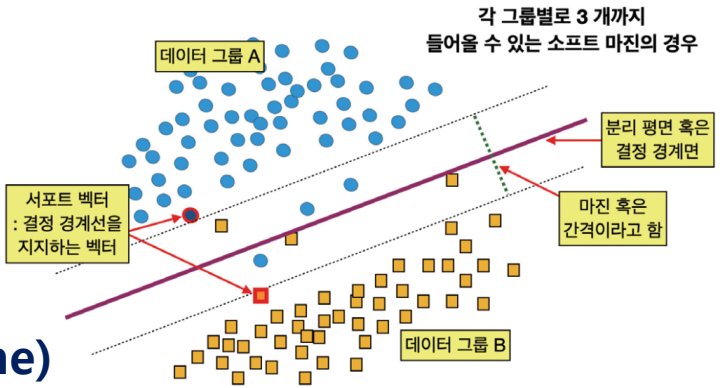
[일반화, 과대적합, 과소적합]

- ✓ 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

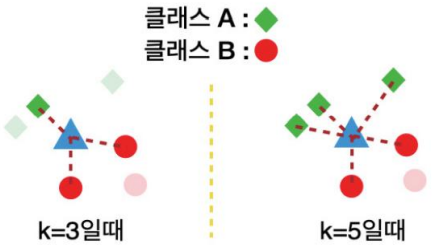
Algorithms ➔ 분류와 회귀 모두 가능한 알고리즘이 많이 있음

- **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**

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

Confusion matrix

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

Confusion matrix

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

Confusion matrix

		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 x 정밀도 x 재현률 / (정밀도 + 재현률)

Data : heart_disease.csv (미국 심장질환 데이터셋)

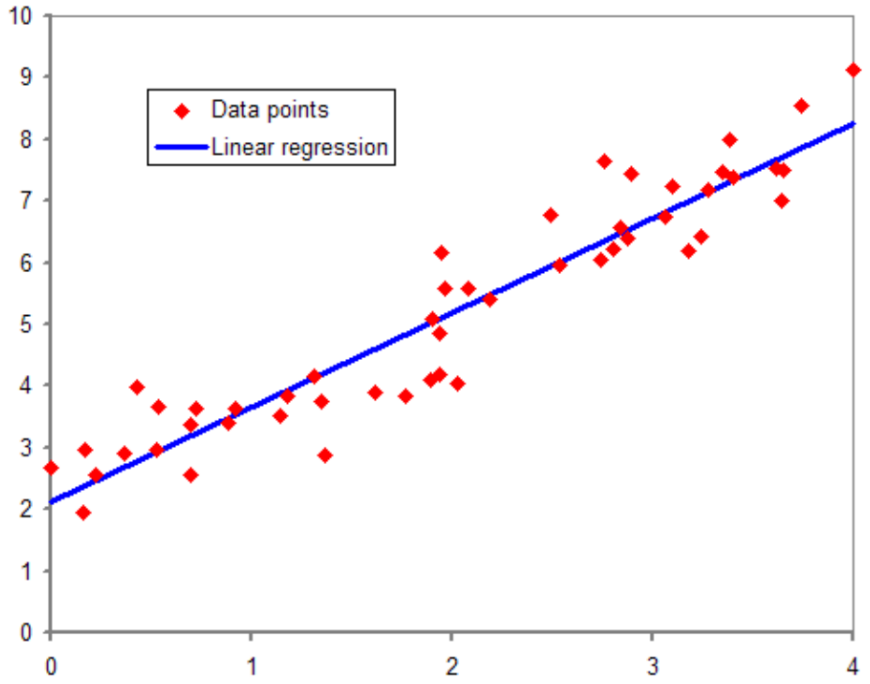
✓ 303행, 14개의 변수 (Target = target)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   age       303 non-null     int64
 1   sex       303 non-null     int64
 2   cp        303 non-null     int64
 3   trestbps  303 non-null     int64
 4   chol      303 non-null     int64
 5   fbs       303 non-null     int64
 6   restecg   303 non-null     int64
 7   thalach   303 non-null     int64
 8   exang     303 non-null     int64
 9   oldpeak   303 non-null     float64
 10  slope     303 non-null     int64
 11  ca        303 non-null     int64
 12  thal      303 non-null     int64
 13  target    303 non-null     int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

age	
sex	성별 (1=남성, 0=여성)
cp	가슴 통증 (1=안정형 협심증, 2=불안정형 협심증, 3=협심증 이외 통증, 4=무증상)
trestbps	휴식 시 혈압
chol	콜레스테롤 수치
fbs	공복 혈당
restecg	휴식 상태의 심전도
thalach	최대 심장 박동수
exang	운동 유발 협심증
oldpeak	운동에 의한 상대적 휴식 시 ST 하강
slope	최대 운동 ST 세그먼트의 기울기
ca	형광 투시로 착색된 주요 혈관 수
thal	탈라세미아 유형
target	심장질환의 존재 여부 (1=예, 0=아니오)

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.

Simple linear regression

- 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 = β_0 + β_1·(영업사원 수)**

Multiple linear regression

- 다중회귀분석 목적

1

2개 이상 변수(독립변수, 예측변수)를 이용해서 다른 변수(종속변수, 결과변수)를 예측함

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 설명함

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

- 다중회귀분석 회귀식

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

Y : 종속변수 β_i : X_i의 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 = β_0 + β_1(영업사원 수) + β_2(프로모션 횟수) + β_3(광고비)**

회귀식의 설명력 R^2

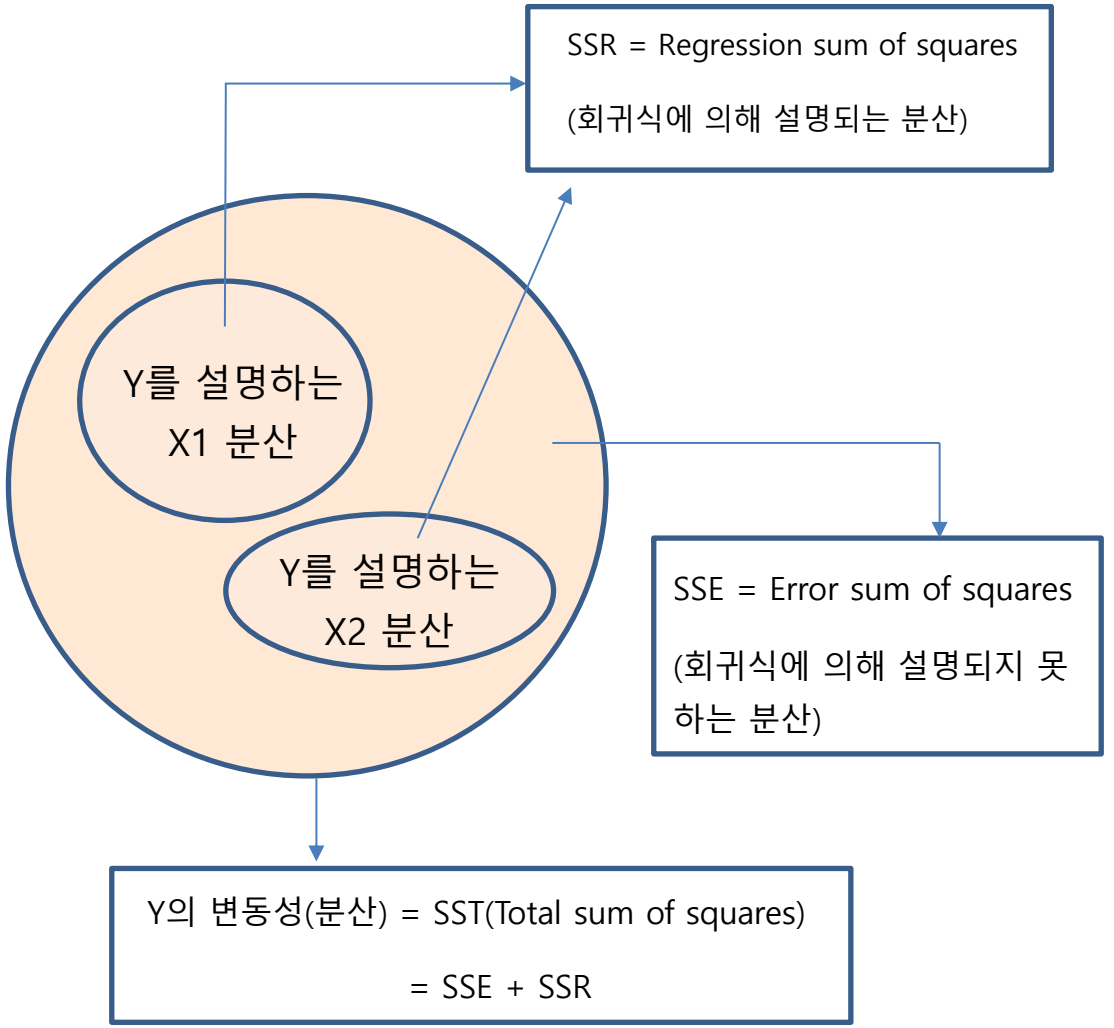
- 회귀식이 종속변수를 설명하고 예측하는데 유용한가를 판단
- 판단지표 : R^2 = (결정계수, 기여율, 설명력), $0 < R^2 < 1$
- R^2 은 종속변수의 분산 중 독립변수에 의해 설명되는 비율을 의미

예 R^2=0.76이라는 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^2에 변수의 수 만큼 penalty를 주는 지표인 $adjusted\ R^2$ 를 주로 활용



- 회귀 분석 결과 예시

모형		비표준화 계수		표준화 계수	t	유의확률	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

- $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$
- 소비자만족도 = -0.631 + 0.744 · 가격만족도 + 0.302 · 구매 횟수 + 0.011 · 연령

- 회귀 분석 결과 예시

모형		비표준화 계수		표준화 계수	t	유의확률	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-value < 0.05), 통계적으로 유의미한 변수로 판단
- 연령은 유의확률이 유의수준보다 크므로 (p-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	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Y_i - \hat{Y}_i\rvert$
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	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{Y_i - \hat{Y}_i}{Y_i}\right\rvert$

Data : insurance.csv (미국 건강보험료 데이터셋)

✓ 1338행, 7개의 변수 (Target = charges)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   age       1338 non-null    int64     연령
 1   sex       1338 non-null    object    성별 (male or female)
 2   bmi       1338 non-null    float64   체질량 지수(body mass index)
 3   children  1338 non-null    int64     자녀의 수(number of children)
 4   smoker    1338 non-null    object    흡연 여부(yes or no)
 5   region    1338 non-null    object    사는 지역(northeast, southeast, northwest, southwest)
 6   charges   1338 non-null    float64   건강보험에서 지출되는 개인별 의료비
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

고생 많으셨습니다. 감사합니다.

홍용기 컨설팅학박사
010-3366-9010 / 123biz@naver.com

크롤링 실행파일 증정

➔Youtube_Meister.exe