AI 융합전문가

(마스터과정-4차시)

데이터분석

2024. 03. 02.

Master Course

Ⅰ.	Ⅱ.	Ⅲ.	Ⅳ.	Ⅴ.
Intro	Web & Web scraping	Data Analysis Theory	Classification Analysis	Regression Analysis

강사 소개_ 홍용기 컨설팅학 박사

소속	주요 업무
㈜제타데이터 대표이사	데이터 전략 및 컨설팅 ODA 컨설팅 (Official Development Assistance)
㈜지구파트너스 감사	창업보육, 투자, 기업·기술가치평가
㈜메타로직 컨설팅 수석	ISP 컨설팅 (Information Strategy Planning) ISMP 컨설팅 (Information System Master Plan)
㈜이너피플 이사	데이터분석, 데이터 가치평가 컨설팅, 데이터바우처 사업수행

◎ 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)**

Intro (https://padlet.com/topmind1472_/ai-3-o7wmxktnk3214u39)

:Padlet

Gerald Hong • 1시간

AI융합전문가 3기 마스터과정

4차시 - 데이터분석(2024. 3. 2)

Intro

Personal Webpage - 1

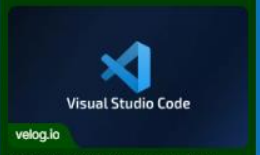
HYK

파이썬(Python) 설치하세요~

[python] 파이썬 설치 하는 방법 (Windows 윈도우)

Web & Crawling

VS code 설치하세요~



[VScode] 비주얼 스튜디오 코드 설치 (Visual Studio Code Install)

Web

Internet
WWW(World Wide Web)
HTTP
HTML(Tag, ID & Class)

daum 크롤링 실습

파이썬과 VS code가 설치되어야 합니다.

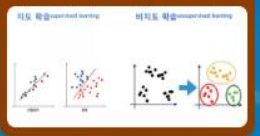
금융 용어 크롤링 실습

파이썬과 VS code가 설치되어야 합니다.

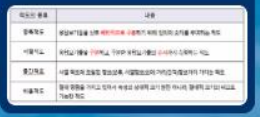
Basic Theory

Data Type

지도학습 vs 비지도학습



척도



머신러닝 프로세스

결측치, 정규화, 표준화

NA, NAN
Null

Classification

분류 알고리즘

의사결정 나무 (Decision Tree)
앙상블 모형 (Ensemble)
1. Bagging
2. Boosting
- AdaBoost (Adaptive Boosting)
- GBM
- XGBoost
- LightGBM
- CatBoost
3. Random Forest
서포트 벡터 머신
(SVM; Support Vector Machine)
K 최근접 이웃 (K-Nearest Neighbor)
소프트맥스 (Softmax) 회귀

Confusion Matrix

ChatGPT 활용 실습

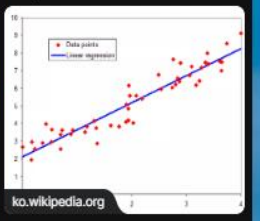
CSV

heart_disease

심장질환 데이터 csv 파일

Regression

회귀분석



단순회귀분석
다중회귀분석
로지스틱회귀분석

회귀분석 결과 해석

회귀식
t-value, p-value

회귀모델의 성능지표

MAE, MSE, RMSE, MAPE

ChatGPT 활용 실습

CSV

CarPrice_Assignment

자동차가격 데이터 csv 파일

Etc.

GitHub.com

Personal Webpage - 2

Expert

Power BI 소개

Power BI - 데이터 시각화 | Microsoft Power Platform

파워 BI 보고서 전체 현황

시각화 측면에서 최강

섹션 추가

What to do?

데이터 전문가 자격증

데이터분석 전문가 ADP (Advanced Data Professional)	필기 실기	국가 공인자격	K-Data
빅데이터 분석기사 BAE (Bigdata Analysis Engineer)	**필기 실기**	**국가 기술자격**	**과기정통부 & 통계청**
데이터분석 준전문가 ADsP (Advanced Data Semi Professional)	필기	국가 공인자격	K-Data

경영정보 시각화능력(BI specialist) 소개

빅데이터시대 취업 필수 국가기술자격증

경영정보시각화능력
BI Specialist

대한상공회의소

경영정보시각화능력(Business Intelligence Specialist)

기업의 내외부 정보를 시각화 요소들을 사용하여 효과적으로 표현하고 전달하는 역량을 평가하는 국가기술자격입니다.

경영정보시각화능력(BI Specialist)을 취득해야하는 이유

취업 활용도 높은 자격

취업에 도움되는 자격 - 전세계적으로 기업, 공공기관 BI 도입으로 인력수요 증가

빅데이터시대 필수역량

데이터분석, 보고서 작성의 핵심역량 - 시각화능력

미래 소프트웨어

엑셀, 파워포인트에 이은 미래 필수 사무 소프트웨어

국가기술자격증

국가에서 시각화분야 전문성을 인정하는 자격증

어느 직무에나 필요

컴퓨터활용능력과 같이 취업 및 사무에 요구되는 필수 자격

비즈니스 인사이트

BI 툴을 사용하여 직접 비즈니스 인사이트 도출

컴활에 이은 사무 필수 자격증

Business Intelligence Specialist

경영정보 시각화능력(BI specialist) 시험일정

경영정보시각화능력 국가기술자격

종목소개

시험일정

사무정보

컴퓨터활용능력
경영정보시각화능력
워드프로세서
비서
엑셀플러스
파워포인트플러스
한글플러스
IT Plus

유통/마케팅

유통관리사
전자상거래관리사
전자상거래운용사

시험안내 | 시험문제 | 시험일정 | 관련자료 | FAQ

경영정보시각화능력 시험일정

- 인터넷접수 여건이 안되는 수험자께서는 인터넷접수 기간중 해당상의를 방문해 주시면 접수에 필요한 시스템(사진 스캔 등)을 제공해 드립니다.
- 접수기간 중이라도 수험자가 많을 경우 시험장은 조기 마감될 수 있습니다.
- 정기 검정 원서접수 마지막 날의 접수마감은 18:00 까지 입니다.(수험료 결제까지 완료하고 접수증이 확인되어야 원서접수가 된 것입니다.)

정기시험 일정안내

2024 조회

종목	회별	구분	등급	인터넷접수	시험일자	발표일자
경영정보시각화능력	1	필기	단일등급	(1차) 2024.03.18 ~ 2024.03.24 (2차) 2024.04.17 ~ 2024.04.23	2024.05.18	2024.06.18
경영정보시각화능력	1	실기	단일등급	2024.08.28 ~ 2024.09.03	2024.09.28	2024.11.18
경영정보시각화능력	2	필기	단일등급	(1차) 2024.09.30 ~ 2024.10.06 (2차) 2024.10.30 ~ 2024.11.05	2024.11.30	2024.12.31

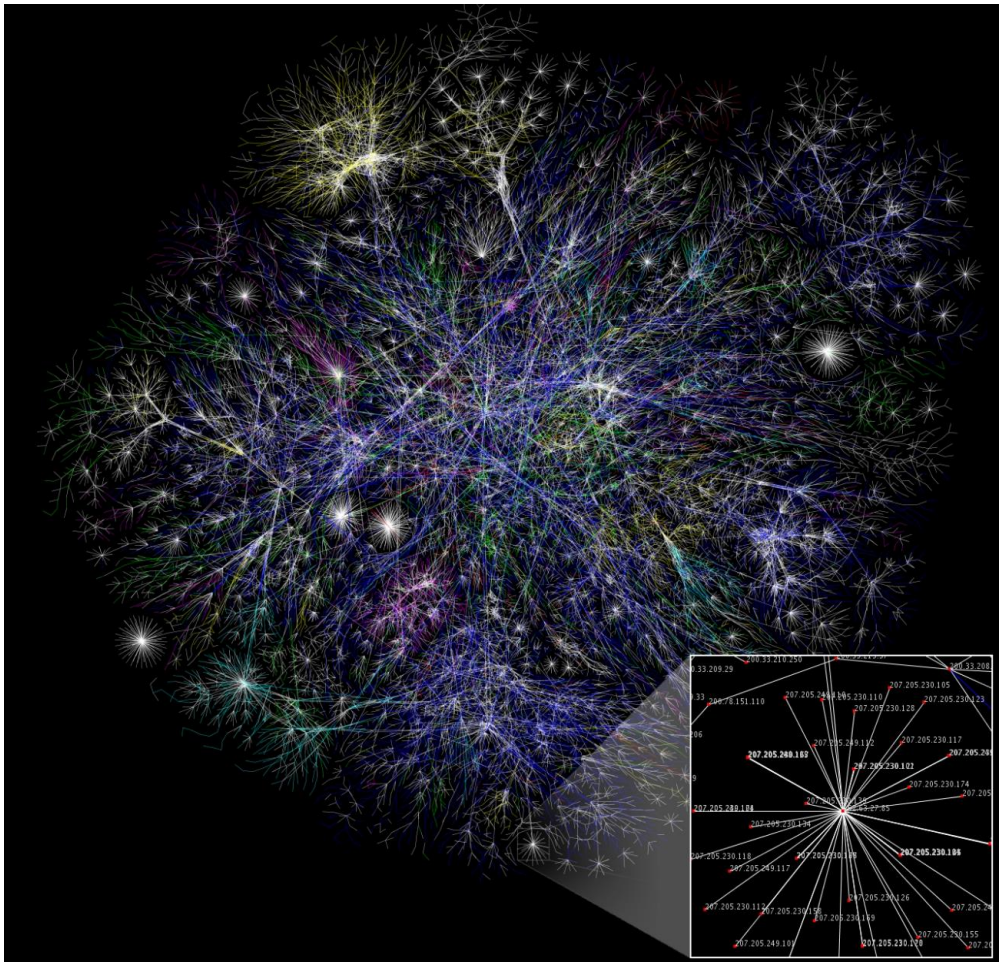
Ⅰ.	Ⅱ.	Ⅲ.	Ⅳ.	Ⅴ.
Intro	**Web & Web scraping**	Data Analysis Theory	Classification Analysis	Regression Analysis

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)

World Wide Web

- 1989년 3월, CERN(유럽 입자 물리 연구소)의 소프트웨어 공학자 팀 버너스리 개발

- WWW은 다음 세 가지의 기능으로 요약할 수 있음

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

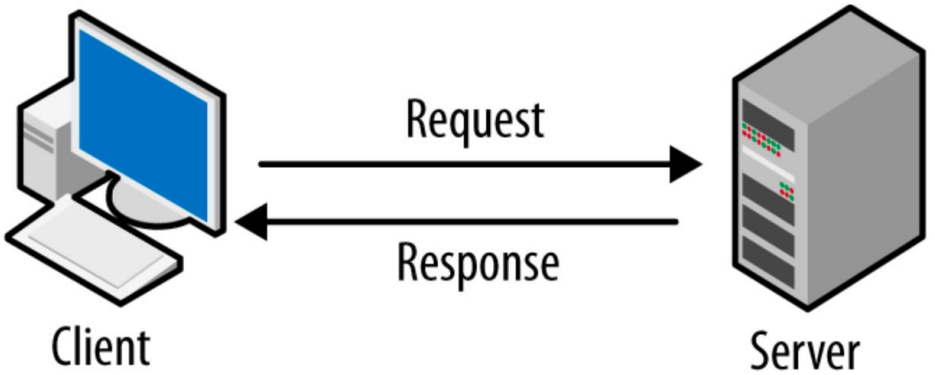
  두 번째, 웹의 자원 이름에 접근하는 프로토콜(protocol) ➔ ex. HTTP

  세 번째, 자원들 사이를 쉽게 항해할 수 있는 언어 ➔ ex. HTML

HTTP

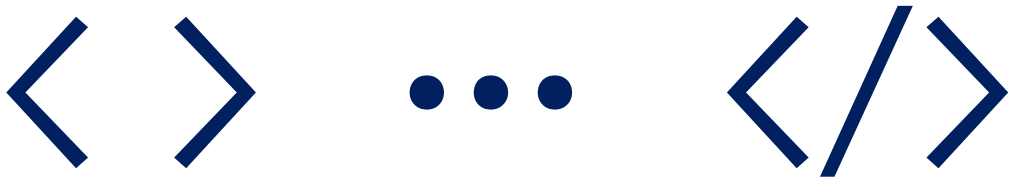
Hyper Text Transfer Protocol은 W3 상에서 정보를 주고받을 수 있는 프로토콜(규약)

- 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



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>
```

HTML은 '정보전달' 이 주목적

디자인 요소

UI

HTML은 ‘정보전달’에 충실

디자인 요소

CSS(Cascading Style Sheet)

UI

JavaScript

정적 크롤링 vs 동적 크롤링

	정적 크롤링	동적 크롤링
연속성	주소를 통해 단발적으로 접근	브라우저를 사용하여 연속적으로 접근
수집 능력	수집 데이터의 한계가 존재	수집 데이터의 한계가 없음
속도	빠름	느림
라이브러리	requests, BeautifulSoup	selenium, chromedriver

출처: https://jaaamj.tistory.com/101

정적 크롤링: 다음 뉴스 크롤링해서 텍스트 파일로 저장(실습)

```
import datetime
import urllib.request as ur
from bs4 import BeautifulSoup as bs

url = 'https://news.daum.net/'
f = open(str(datetime.date.today()) + 'articles.txt', 'w')
soup = bs(ur.urlopen(url).read(), 'html.parser')

for i in soup.find_all('div', {'class':'cont_thumb'})[:20]:
    try:
        print(i)
        f.write(i.text)
        f.write(i.find_all('a')[0].get('href') + '\n')
        soup2 = bs(ur.urlopen(i.find_all('a')[0].get('href')).read(), 'html.parser')
        for j in soup2.find_all('p'):
            print(j)
            f.write(j.text)
    except:
        pass

f.close()
```

정적 크롤링: 금융 용어사전 크롤링해서 엑셀 파일로 저장(실습)

```
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd

df = pd.DataFrame(columns=['term', 'content'])
for i in range(1, 55):
    url = "https://fine.fss.or.kr/fine/fnctip/fncDicary/list.do?menuNo=900021&pageIndex="+str(i)+"&src=&kind=&searchCnd=1&searchStr="
    print(url)
    response = requests.get(url)
    html = response.text
    soup = bs(html, 'html.parser')
    results = soup.select('#content > div.bd-list.result-list > dl')

    for result in results:
        title_split = result.find('dt').text.split('.')
        index = '{0:03d}'.format(int(title_split[0].replace('\r','').replace('\n','').replace('\t','')))
        term = title_split[1].replace('\r','').replace('\n','').replace('\t','')
        content = result.find('dd').text
        df.loc[index] = [term, content]

df.to_excel("financial terms.xlsx", index=False)
```

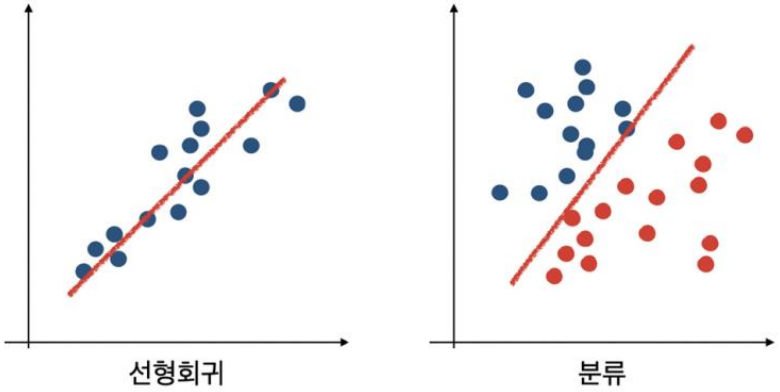
동적 크롤링(시연)

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.select import Select
import pandas as pd
from datetime import datetime
import time
import math

df = pd.DataFrame(columns=['Date', 'Source', 'Title', 'Summary', 'Link'])
url = "https://dream.kotra.or.kr/kotranews/cms/com/index.do?MENU_ID=170#"
options = webdriver.ChromeOptions()
options.add_argument('--start-maximized')
options.add_argument('--disable-gpu')
driver = webdriver.Chrome(options=options)
driver.get(url)
time.sleep(1)

cal = driver.find_element(By.XPATH, '//*[@id="fd-but-pStartDt"]/span[1]').click()
year = driver.find_element(By.XPATH, '//*[@id="pStartDt-prev-year-but"]')
year.click()
time.sleep(1) … (생략)
```

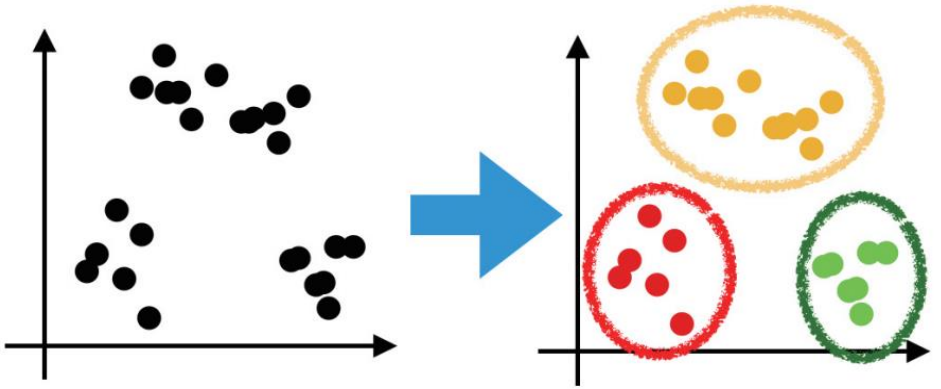
Ⅰ.	Ⅱ.	Ⅲ.	Ⅳ.	Ⅴ.
Intro	Web & Web scraping	**Data Analysis Theory**	Classification Analysis	Regression Analysis

Supervised Learning VS Unsupervised Learning

지도 학습supervised learning

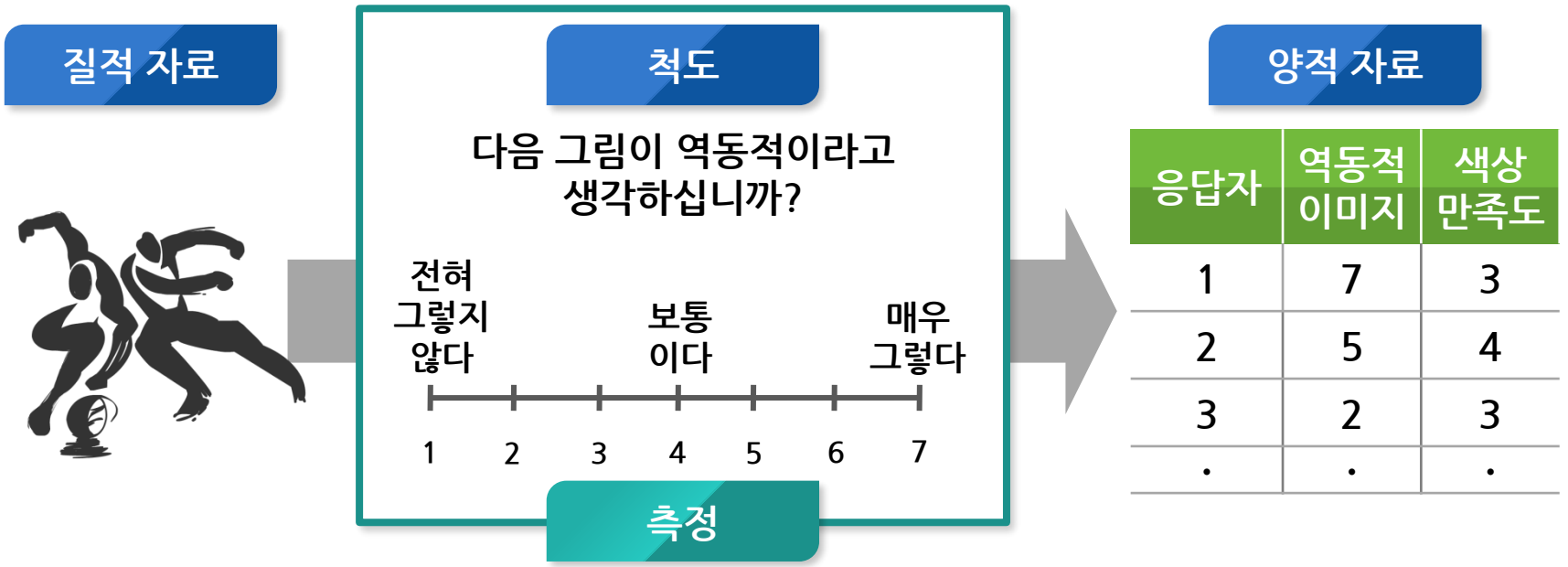


비지도 학습unsupervised learning



※출처: 으뜸 데이터 분석과 머신러닝(박동규·강영민, 2021)

- 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)

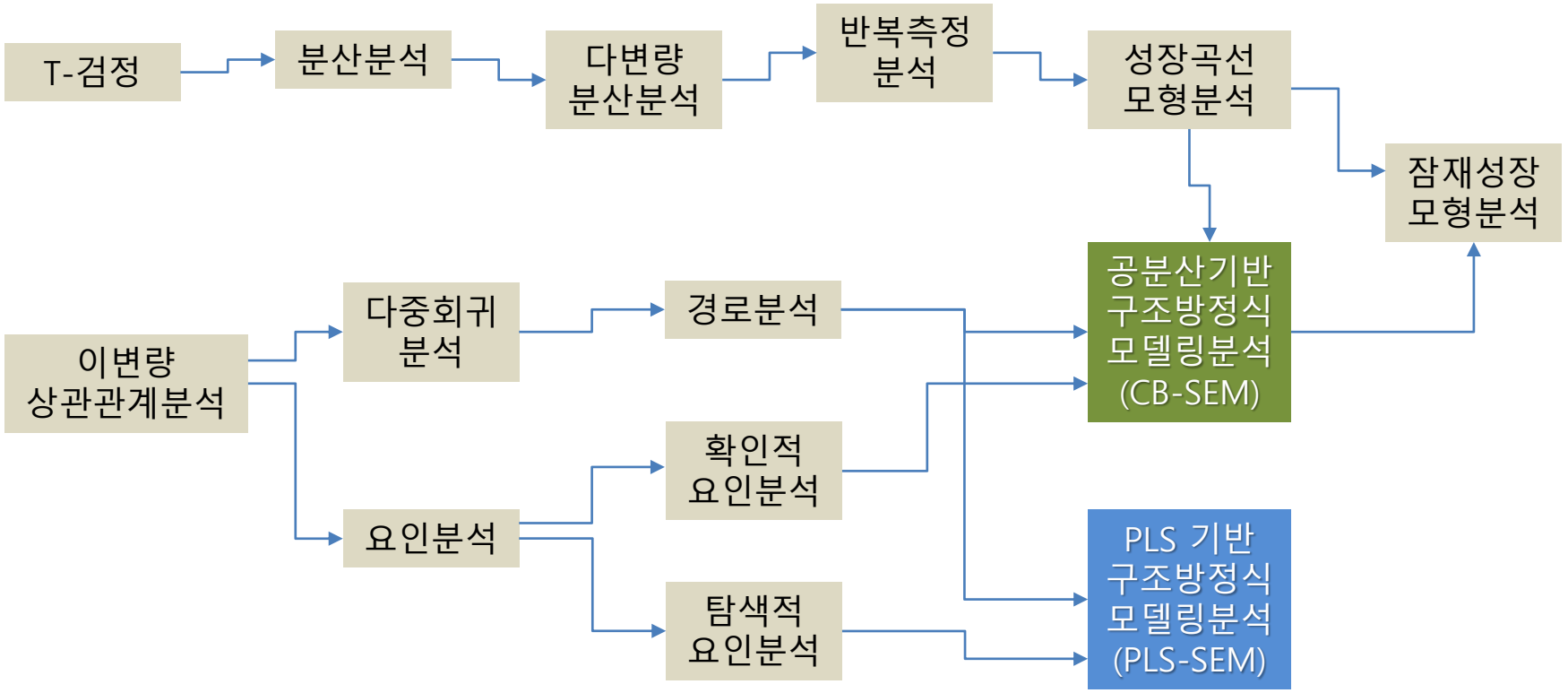
- 절대 영점을 가지고 있어서 속성의 상대적 크기 뿐만 아니라, 절대적 크기의 비교도 가능한 척도

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

통계분석 기법의 발전



- **구조방정식모델링(Structural Equation Modeling: SEM)은 통계학이 발전함에 따라 회귀분석(regression analysis), 요인분석(factor analysis) 및 경로분석(path analysis)이 결합되어 발전된 다변량통계기법(multivariate analysis)의 하나임**

※출처: SmartPLS 3.0 구조방정식모델링(신건권, 2018)

데이터 분석의 목적

1. 데이터 분석의 결과를 활용하여 현재의 문제 해결 내지 현상을 설명하기 위한 목적
2. 미래에 대한 예측 및 판단과 같은 의사결정에 활용하기 위한 목적

- 전자를 목표로 주요하게 활용된 기법은 주로 통계학 분야 내의 방법론들이 다수이며, 후자를 주요 목표로 통계, 수학, 컴퓨터과학 등 가능한 모든 기술을 동원하는 것은 최근의 데이터 마이닝이라 일컬어지는 분야

- 예컨대 분석과정에 통계기법의 하나인 회귀분석(Regression)이 사용되었다고 가정할 때, 통계적 관점에서는 종속변수에 대한 독립변수들의 설명력을 도출하고 검증하는 것만으로도 충분

- 데이터 분석은 도출된 회귀식을 근거로 하여 (미래시점에 등장할) 종속변수의 추정 값이나 분류결과 등을 미리 제시해 주는 단계까지 진행한다는 점에서 차이점이 있음

현상에 대한 정확하고 근거 있는 설명을 목표로 하는 통계는, 필연적으로 데이터에 대한 엄격한 통제를 중시

- ✓ 표본 수집
- ✓ 데이터의 순수성
- ✓ 엄격한 포맷
- ✓ 결과의 엄격한 해석을 위한 유의성 기준 등

반면, 미래 시점에서의 다양한 목적 달성을 지향하는 데이터 마이닝의 경우 각 단계 별로 전략적 판단 하에 다소 느슨한 형태의 의사결정 및 적용이 허용됨

- ✓ 전수 조사
- ✓ 현행 서비스 고도화 등의 비즈니스 측면
- ✓ 경제적 측면
- ✓ 마케팅적 측면 등

- 목표 달성을 위하여 필요한 분석 기법에 있어서도 오랜 기간 동안 전통적인 통계기법을 포함, 컴퓨터과학의 인공지능(딥러닝과 머신러닝), 데이터베이스, 패턴(음성인식)과 같은 분야의 주요 기술들을 데이터 분석의 구체적인 기술적 수단으로 활용

- 이에 더해 4차 산업 혁명 이후 데이터의 빅뱅, 관련 S/W 및 H/W의 비약적 발전 등으로 인해 공학/수학/통계학 분야 외에도 광범위한 분야의 다양한 기술들이 데이터 분석에 접목되어 다양하게 활용되고 있는 상황

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

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

Data split

Training data

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

Validation data

					X_val								y_val

Test data

y_predict

					X_test								y_test

교차 검증의 의의

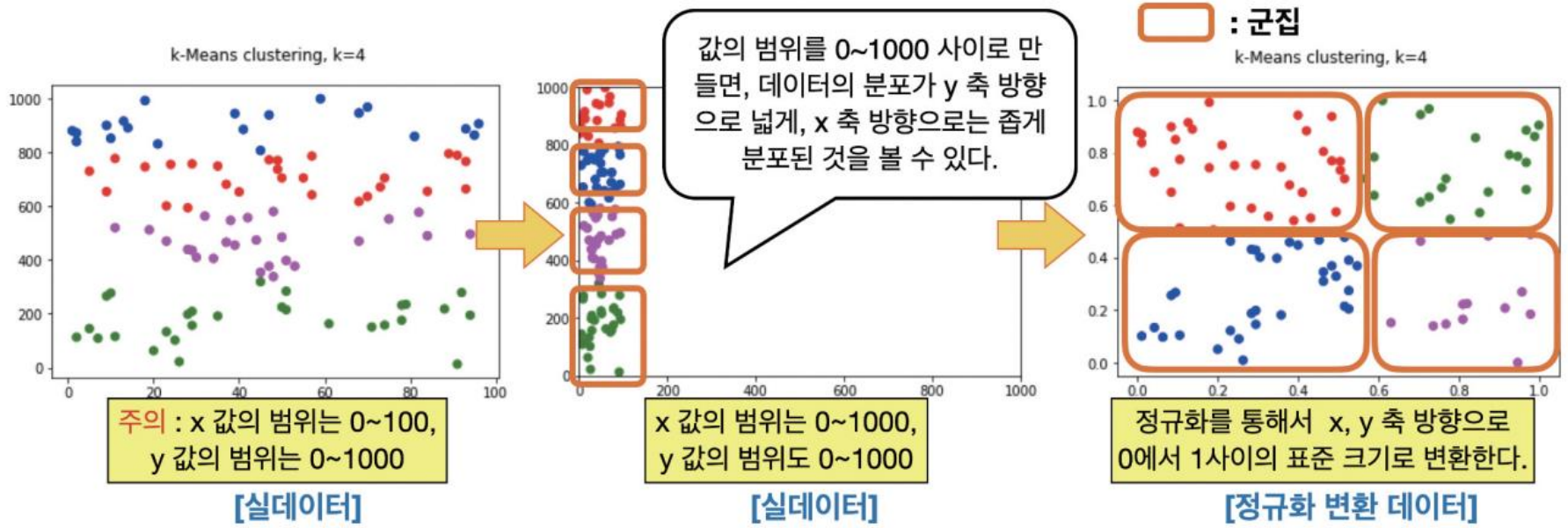
- 교차 검증은 일반화 성능을 재기 위해 훈련 세트와 검증 세트로 한 번 나누는 것보다 더 안정적이고 뛰어난 통계적 평가 방법
- 교차 검증에서는 데이터를 여러 번 반복해서 나누고 여러 모델을 학습
- 가장 널리 사용되는 교차 검증 방법은 k-겹 교차 검증(k-fold CV)으로 보통 5 또는 10을 사용
- 5-겹 교차 검증을 하면 데이터를 비슷한 크기의 부분 집합(5개의 폴드)으로 나누고, 일련의 모델을 만들어 훈련과 테스트를 반복

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

Missing value : NA, NAN, Null

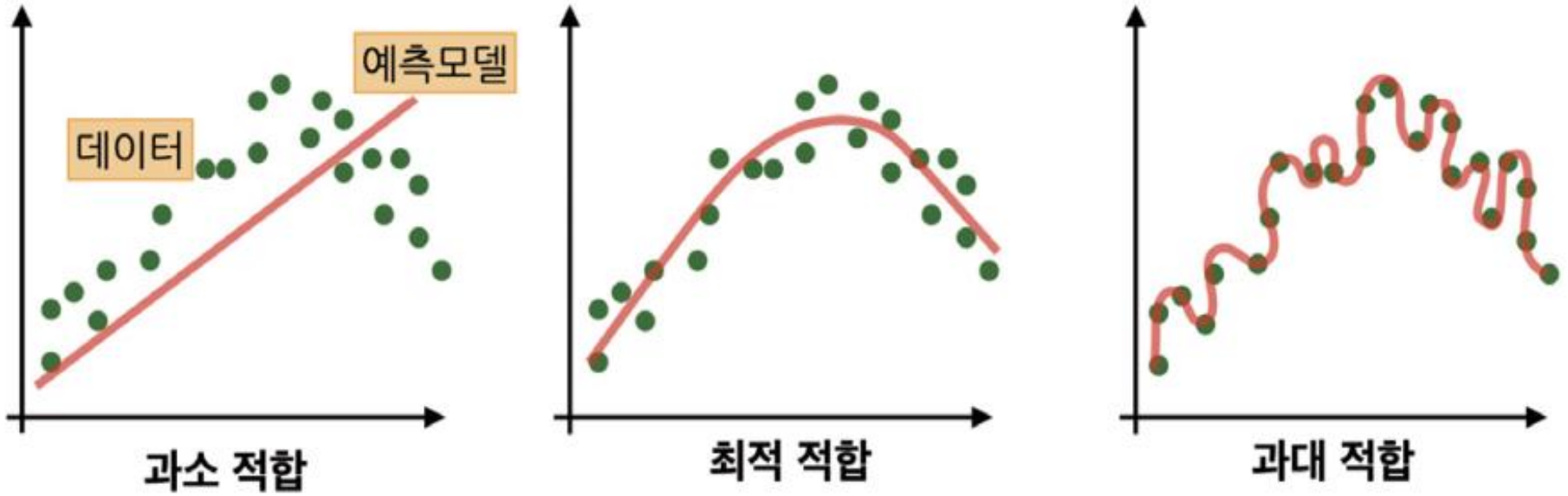
➔ 결측치 제거 또는 대체 (평균, 중위수, 최빈값)

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)

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

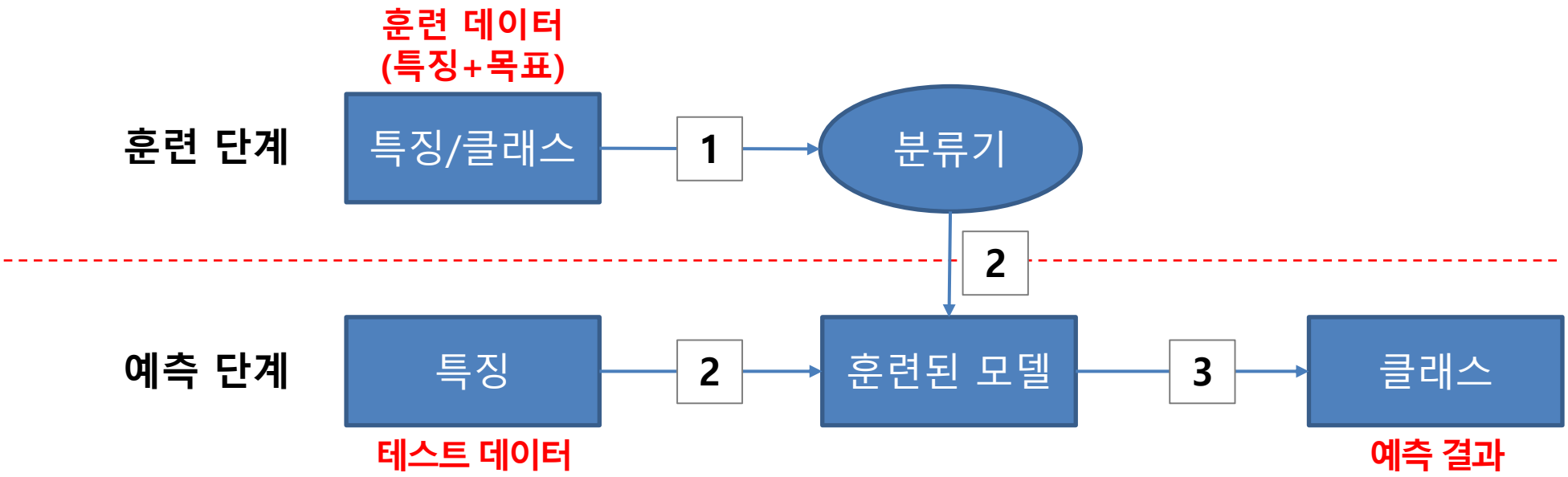
- 최적의 하이퍼파라미터 및 최종모델 결정 단계
- 최종모델에 평가데이터를 적용하여 머신러닝 알고리즘의 예측성능을 평가
- 평가데이터는 반드시 학습 과정이나 검증 과정에서 사용되지 않은 데이터로 사용해야 함
- 평가데이터에 대한 평가지표를 머신러닝 분석에 대한 최종성능으로 제시

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

Ⅰ.	Ⅱ.	Ⅲ.	Ⅳ.	Ⅴ.
Intro	Web & Web scraping	Data Analysis Theory	**Classification Analysis**	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

단순회귀분석 (Simple Linear Regression)

- **하나의 독립변수와 하나의 종속변수 간의** 선형적인 관계를 모델링하는 회귀분석 방법
- 독립변수와 종속변수 간의 관계를 파악하고, 독립변수의 값을 통해 종속변수 값을 예측하거나 설명

다중회귀분석(Multiple Linear Regression)

- **둘 이상의 독립변수와 하나의 종속변수 간의** 선형적인 관계를 모델링하는 회귀분석 방법

로지스틱 회귀분석(Logistic Regression)

- **종속변수가 이항형(binary)**일 때, 독립변수와 종속변수 간의 선형적인 관계를 모델링하는 회귀분석 방법
- 로지스틱 회귀분석은 이진 분류(binary classification, 0과 1로 분류)에 널리 사용되며, 예측하려는 결과가 두 가지 중 하나인 경우에 사용

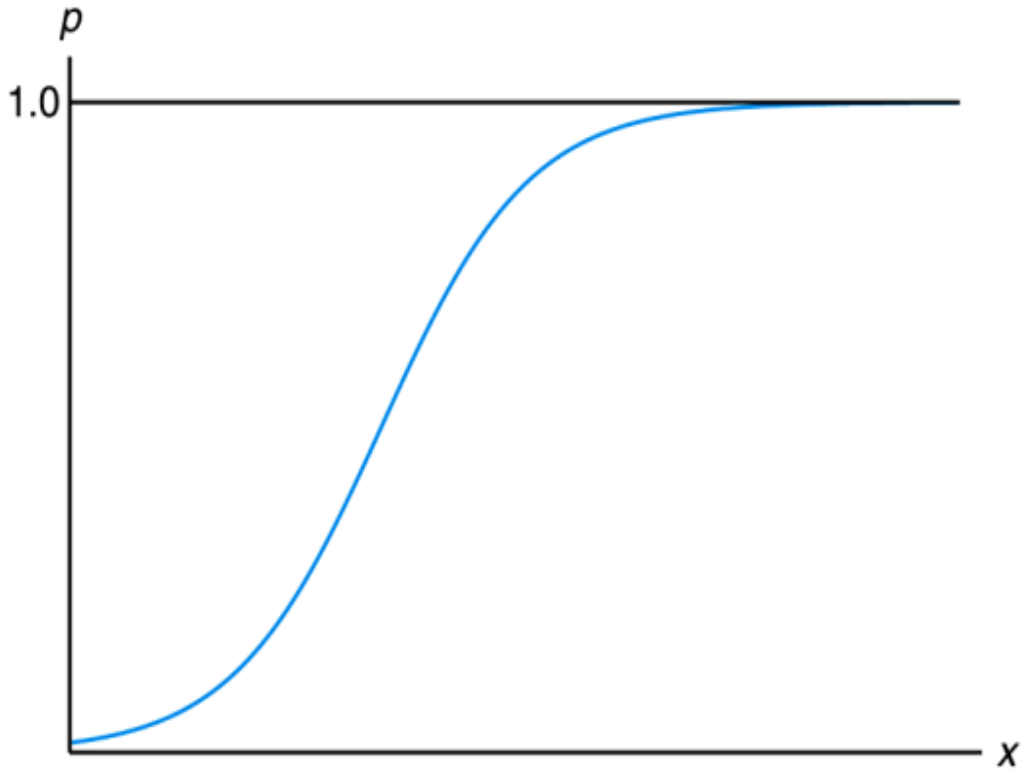
Logistic regression

- **결과변수는 범주형 범주로서 1(사건발생), 0(사건 미 발생)의 값을 갖기 때문에 결과변수의 기대 값은 항상 0과 1 사이의 값을 가짐**
- **결과변수 예측 값은 사건이 발생할 확률을 나타냄**
  - 특정 고객의 카드연체 가능성 예측 모델 결과값이 0.78이라면, 카드를 연체할 확률이 78%라는 의미임
- **분류 문제에서는 0.5를 기준으로 0과 1을 분류함**

예1 카드회사에서 신규 카드 발급 시 고객정보를 기반으로 연체 가능성 예측

예2 환자의 연령, 성별, 혈액 검사 결과를 기반으로 암 진단

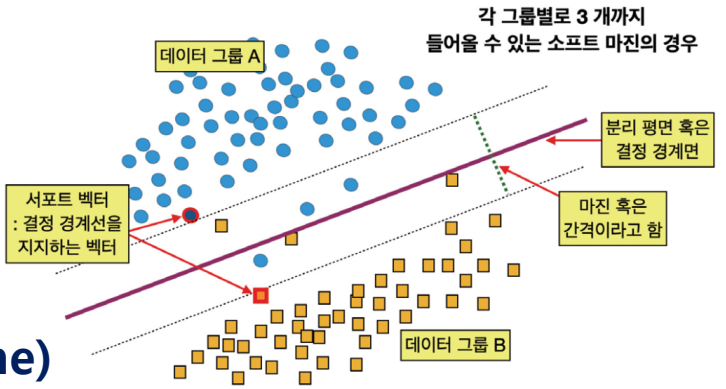
- 로지스틱 회귀곡선
  - 독립변수과 결과변수의 관계는 비선형 S자의 형태
  - 독립변수의 일정 수준까지 서서히 증가하다가 일정 수준이 지나면 결과변수 값이 급격하게 상승



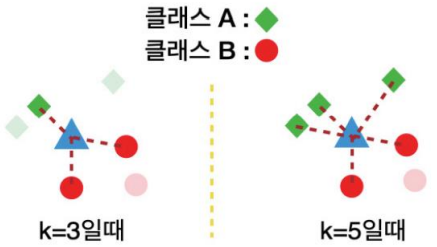
Algorithms ➔ 분류와 회귀 모두 가능한 알고리즘이 많이 있음

- **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)

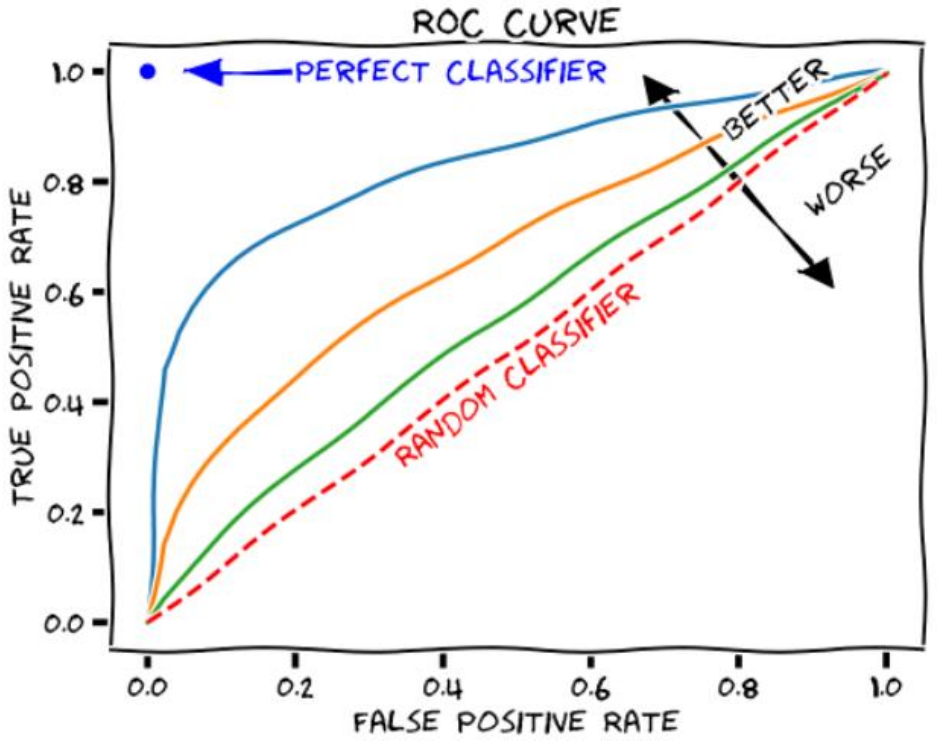
분류 모델의 성능 지표 : ROC 곡선, AUC

- **ROC(Receiver Operating Characteristic) 곡선**

분류 모델의 임계값(threshold)을 변화시켰을 때, 모델의 TPR(True Positive Rate)과 FPR(False Positive Rate)이 어떻게 변화하는지를 나타내는 그래프

- **AUC(Area Under the Curve)**

ROC 커브 아래 면적을 나타내는 지표로 1에 가까울 수록 성능이 우수한 것으로 판단



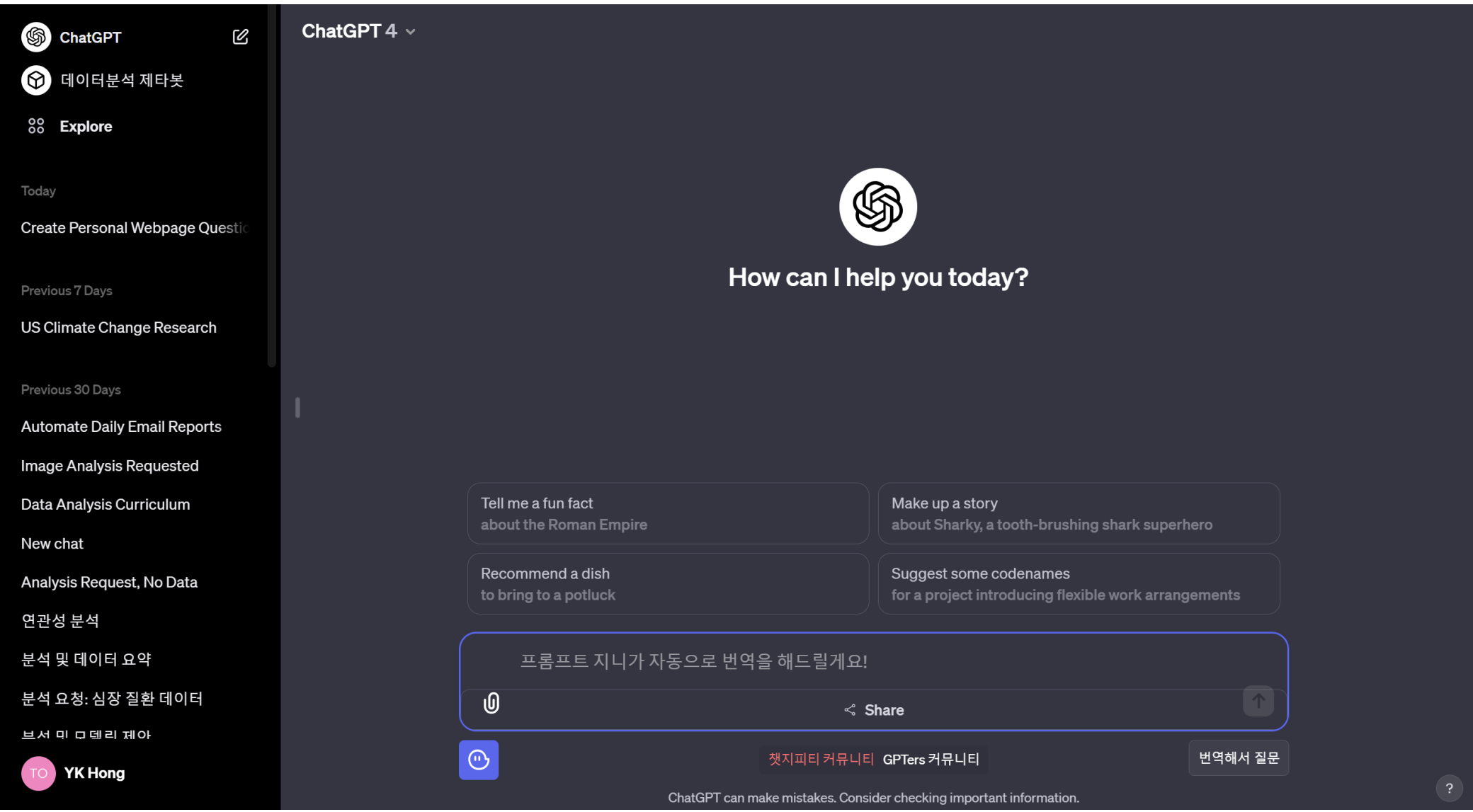
Data : heart_disease.csv (Kaggle의 심장질환 데이터셋)

✓ 303행, 14개의 변수(원래 76개의 속성을 가지고 있었지만 14개로 축소하여 배포)

```
<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
```

0	age	303 non-null	int64	
1	sex	303 non-null	int64	성별 (1=남성, 0=여성)
2	cp	303 non-null	int64	가슴 통증 (1=안정형 협심증, 2=불안정형 협심증, 3=협심증 이외 통증, 4=무증상)
3	trestbps	303 non-null	int64	휴식 시 혈압
4	chol	303 non-null	int64	콜레스테롤 수치
5	fbs	303 non-null	int64	공복 혈당
6	restecg	303 non-null	int64	휴식 상태의 심전도
7	thalach	303 non-null	int64	최대 심장 박동수
8	exang	303 non-null	int64	운동 유발 협심증
9	oldpeak	303 non-null	float64	운동에 의한 상대적 휴식 시 ST 하강
10	slope	303 non-null	int64	최대 운동 ST 세그먼트의 기울기
11	ca	303 non-null	int64	형광 투시로 착색된 주요 혈관 수
12	thal	303 non-null	int64	탈라세미아 유형
13	target	303 non-null	int64	심장질환의 존재 여부 (1=예, 0=아니오)

chat.openai.com

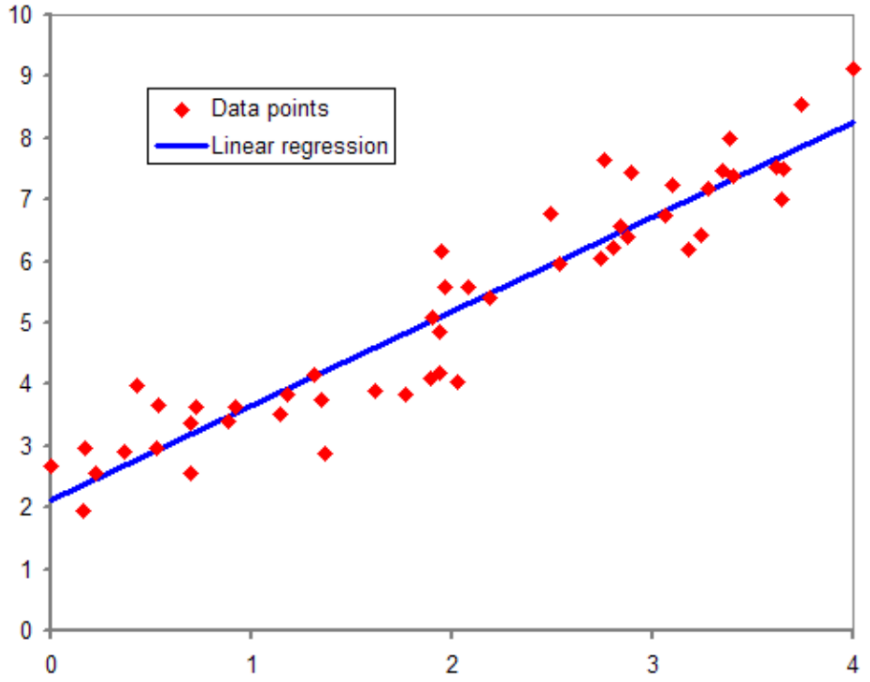


※출처: ChatGPT 홈페이지(https://chat.openai.com/ 2023.11.22. 캡처)

Ⅰ.	Ⅱ.	Ⅲ.	Ⅳ.	Ⅴ.
Intro	Web & Web scraping	Data Analysis Theory	Classification Analysis	**Regression 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 = 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)**

OLS Regression Results						
Dep. Variable:	불량률	R-squared:	0.281			
Model:	OLS	Adj. R-squared:	0.228			
Method:	Least Squares	F-statistic:	5.273			
Date:	Tue, 12 Sep 2023	Prob (F-statistic):	0.00117			
Time:	20:12:00	Log-Likelihood:	-102.95			
No. Observations:	59	AIC:	215.9			
Df Residuals:	54	BIC:	226.3			
Df Model:	4					
Covariance Type:	nonrobust					
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	28.1944	6.858	4.111	0.000	14.446	41.943
함수율	-0.4248	0.118	-3.587	0.001	-0.662	-0.187
온도	-0.0912	0.038	-2.430	0.018	-0.166	-0.016
습도	0.0053	0.019	0.280	0.781	-0.033	0.043
미세먼지	-0.0145	0.014	-1.032	0.307	-0.043	0.014
Omnibus:	3.587	Durbin-Watson:	1.887			
Prob(Omnibus):	0.166	Jarque-Bera (JB):	2.774			
Skew:	0.512	Prob(JB):	0.25			
Kurtosis:	3.284	Cond. No.	3.09E+03			

회귀식의 설명력 R^2

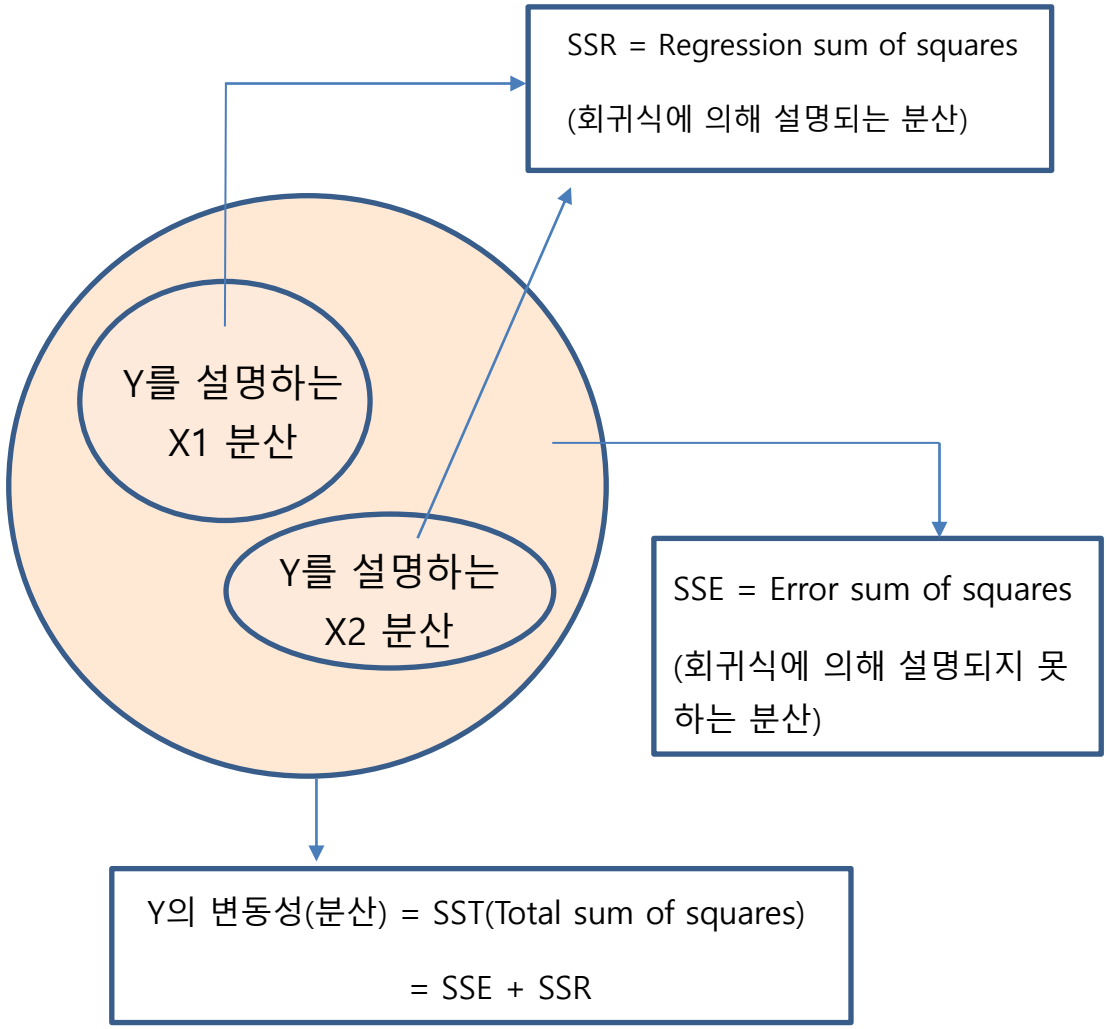
- 회귀식이 종속변수를 설명하고 예측하는데 유용한가를 판단
- 판단지표 : R^2 = (결정계수, 기여율, 설명력), $0 < R^2 < 1$
- R^2 은 종속변수의 분산 중 독립변수에 의해 설명되는 비율을 의미

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	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

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

- 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

독립변수의 상대적 영향력 크기

- 표준화 회귀계수(베타)의 크기는 가격만족도, 구매횟수, 연령순으로 나타나 가격만족도가 종속변수 (소비자만족도)에 가장 큰 영향을 미치는 변수임을 알 수 있음

- 변수선택법 : 변수가 여러 개일 때 최적의 변수 조합을 찾는 방법

구 분	개 요
전진선택법 (Forward Selection)	- 가장 중요한 변수부터 하나씩 추가해가면서 최적의 모델을 찾는 방법 - 먼저 한 개의 변수를 선택하고, 이 변수에 대한 회귀 모델을 돌려본 후, 다른 변수를 하나씩 추가하면서 회귀 모델의 성능을 측정
후진제거법 (Backward Elimination)	- 모든 변수를 포함한 회귀 모델에서 가장 중요하지 않은 변수부터 하나씩 제거하면서 최적의 모델을 찾는 방법
단계적 선택법 (Stepwise Selection)	- 전진 선택법과 후진 제거법의 조합으로, 새로운 변수를 추가하거나 기존 변수를 제거하는 과정을 반복하여 최적의 모델을 찾는 방법 - 전진 선택법에서 선택된 변수도 중요도를 다시 평가하여 제가 할 수 있음

- 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	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{Y_i - \hat{Y}_i}{Y_i}\right\rvert$

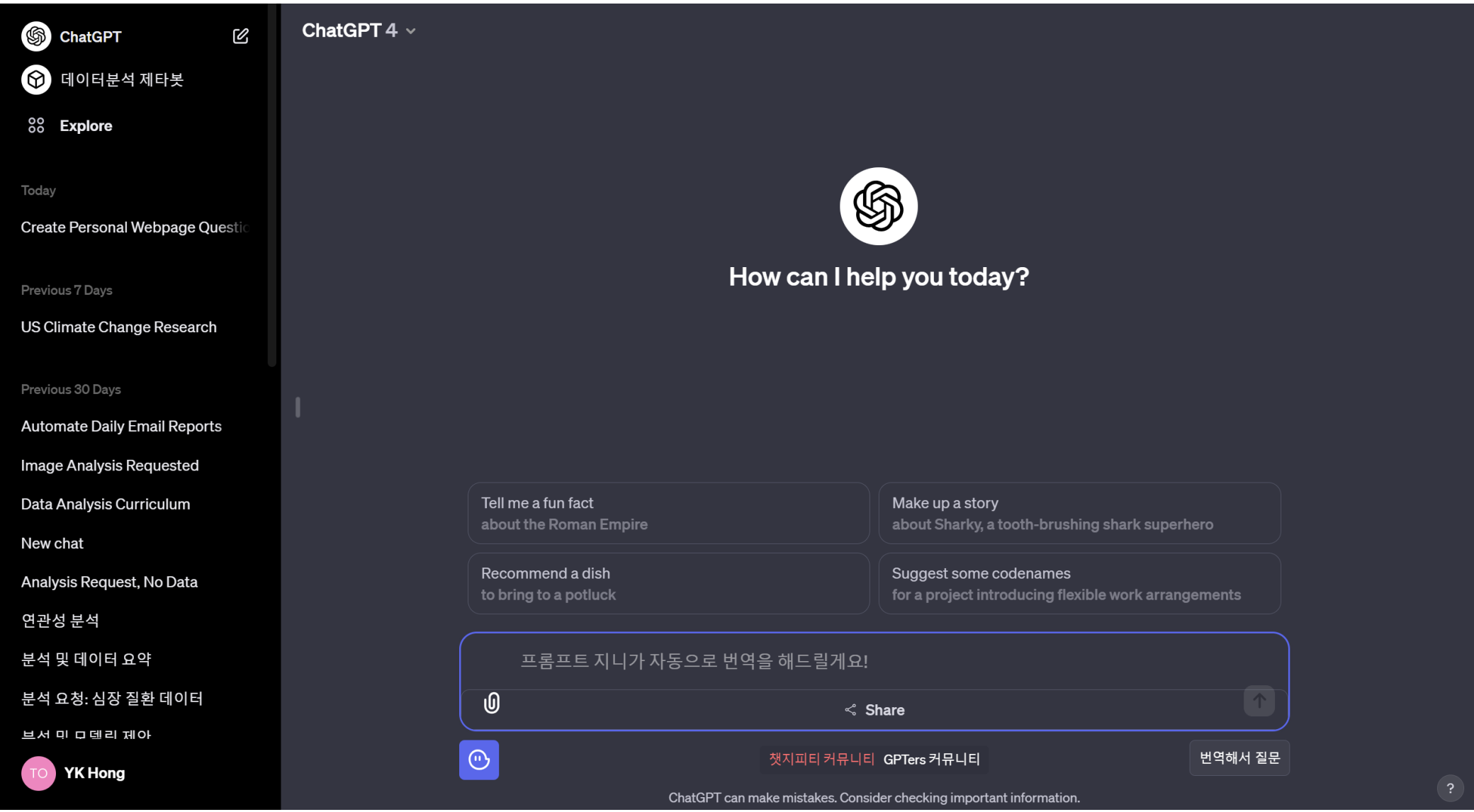
Data : CarPrice_Assignment.csv (Kaggle의 자동차 데이터셋)

✓ 205행, 26개의 변수 (Target = price)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   car_ID            205 non-null    int64
 1   symboling         205 non-null    int64
 2   CarName           205 non-null    object
 3   fueltype          205 non-null    object
 4   aspiration        205 non-null    object
 5   doornumber        205 non-null    object
 6   carbody           205 non-null    object
 7   drivewheel        205 non-null    object
 8   enginelocation    205 non-null    object
 9   wheelbase         205 non-null    float64
 10  carlength         205 non-null    float64
 11  carwidth          205 non-null    float64
 12  carheight         205 non-null    float64
 13  curbweight        205 non-null    int64
```

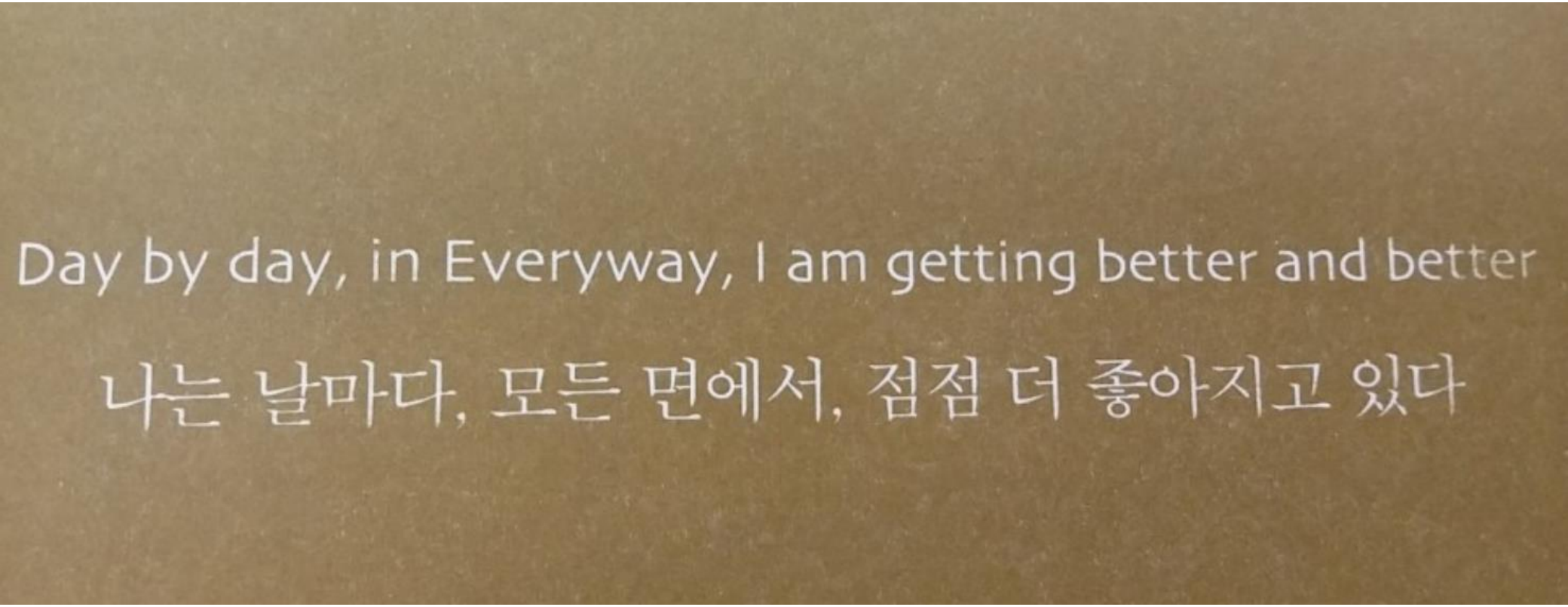
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 14  enginetype        205 non-null    object
 15  cylindernumber    205 non-null    object
 16  enginesize        205 non-null    int64
 17  fuelsystem        205 non-null    object
 18  boreratio         205 non-null    float64
 19  stroke            205 non-null    float64
 20  compressionratio  205 non-null    float64
 21  horsepower        205 non-null    int64
 22  peakrpm           205 non-null    int64
 23  citympg           205 non-null    int64
 24  highwaympg        205 non-null    int64
 25  price             205 non-null    float64
dtypes: float64(8), int64(8), object(10)
memory usage: 41.8+ KB
```

chat.openai.com



※출처: ChatGPT 홈페이지(https://chat.openai.com/ 2023.11.22. 캡처)

“꾸준한 연습과 반복을 권합니다.”



도서「자기암시」 -에밀 쿠에-

Q & A