1주차

데이터사이언스 소개

Introduction to Data Science

> *"데이터로 세상을 바꾸는 기술의 시작"*

강의자: **노기섭 교수** | 소속: 소프트웨어융합학과

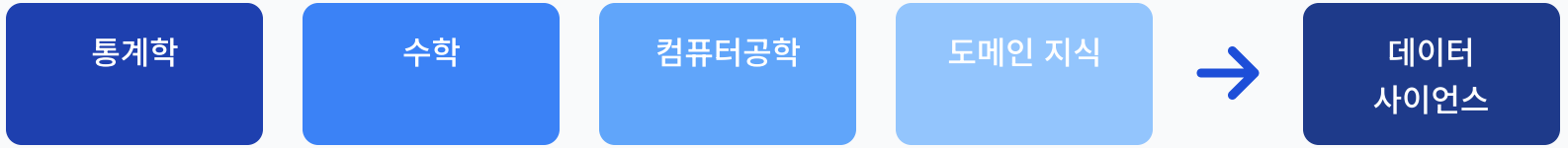
목차

학습목표

- 데이터사이언스의 개념과 필요성 이해
- 데이터 기반 문제 해결 전 과정 습득
- 데이터사이언티스트의 역할 파악
- 실제 산업계 응용 프로세스와 사례 학습

데이터사이언스란 무엇인가

- **다양한 데이터**로부터 의미 있는 정보를 추출하고, 이를 기반으로 **예측, 분류, 최적화** 문제를 해결하는 학문
- **통계학, 수학, 컴퓨터공학, 도메인 지식**이 융합된 형태
- 21세기 '**가장 매력적인 직업**'으로 불릴 정도로 산업 전반에 중요한 분야
- **실제 문제 해결과 의사결정 지원**에 핵심적인 가치 제공



예: 의료분야 환자 재입원 예측, 금융분야 부도 위험 탐지, 유통업 고객 구매 행동 분석

문제해결 프로세스 개요

데이터사이언스에서는 일관된 문제해결 프로세스를 따릅니다:

1. **문제 정의**: 분석 목표와 방향을 명확히 설정
2. **데이터 수집**: 다양한 소스에서 신뢰성 있는 데이터 확보
3. **데이터 전처리**: 결측치, 이상치, 중복값 처리
4. **탐색적 데이터 분석(EDA)**: 데이터 구조와 패턴 파악
5. **모델링**: 예측, 분류, 군집화 등 문제에 적합한 모델 구축
6. **평가 및 검증**: 모델 성능 측정 및 개선
7. **결과 해석 및 의사결정**: 분석 결과를 실제 적용

(다음 슬라이드부터 각 단계를 상세히 살펴봅니다)

1단계. 문제 정의

- 데이터 분석의 **첫 단계**는 해결하고자 하는 문제를 명확히 정의하는 것
- 문제 정의는 **분석 방향과 목표**를 결정하는 핵심 단계
- 잘못된 정의는 잘못된 결론으로 이어질 수 있음

성공 사례

통신사에서 '이탈 고객'을 명확히 정의하고 예측 모델을 구축하여 수익 감소 방지

실패 사례

한 온라인 쇼핑몰은 '구매율' 향상만 목표로 설정해 사이트 체류 시간은 늘었으나 실제 매출 증대 효과 없음

2단계. 데이터 수집

웹 데이터

데이터베이스

센서 데이터

외부 파일

- 데이터는 웹, 데이터베이스, 센서 등 다양한 소스에서 수집됨
- 수집 데이터의 품질과 신뢰성이 분석 성패를 좌우함

```
# 예시: 뉴스 기사 수집 (웹 크롤링)
import requests
from bs4 import BeautifulSoup

url = "https://news.ycombinator.com/"
res = requests.get(url)
soup = BeautifulSoup(res.text, "html.parser")
titles = [item.text for item in soup.select(".titleline > a")]
print(titles[:5])
```

- **성공 사례:** 금융권에서 고객 거래 데이터를 DB에서 추출해 신용평가 모델 구축
- **실패 사례:** 크롤링 데이터의 최신성 검증을 소홀히 하여, 오래된 정보로 잘못된 마케팅 전략 도출

3단계. 데이터 전처리

결측치, 이상치, 중복값 등 데이터 문제 처리

Pandas를 활용한 전처리 예시:

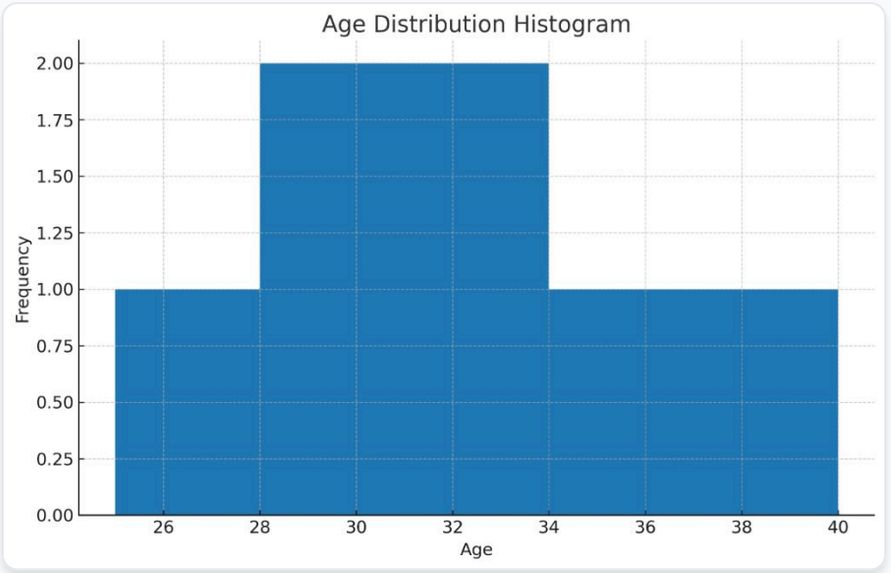
```
import pandas as pd
df = pd.DataFrame({'age': [25, None, 30, 120, 28]})
df['age'] = df['age'].fillna(df['age'].mean())  # 결측치 처리
df = df[df['age'] < 100]  # 이상치 제거
print(df)
```

```
    age
0   25.0
2   30.0
4   28.0
```

산업계 사례: 신용카드 거래 데이터에서 이상치를 탐지하여 사기 거래 조기 발견

4단계. 탐색적 데이터 분석(EDA)

- 데이터 구조와 특성 파악, 패턴 발견
- 히스토그램, 상관분석, 다양한 시각화 활용
- 산업계 활용: 보험사의 고객 연령별 사고율 시각화로 상품 가격 정책 수립



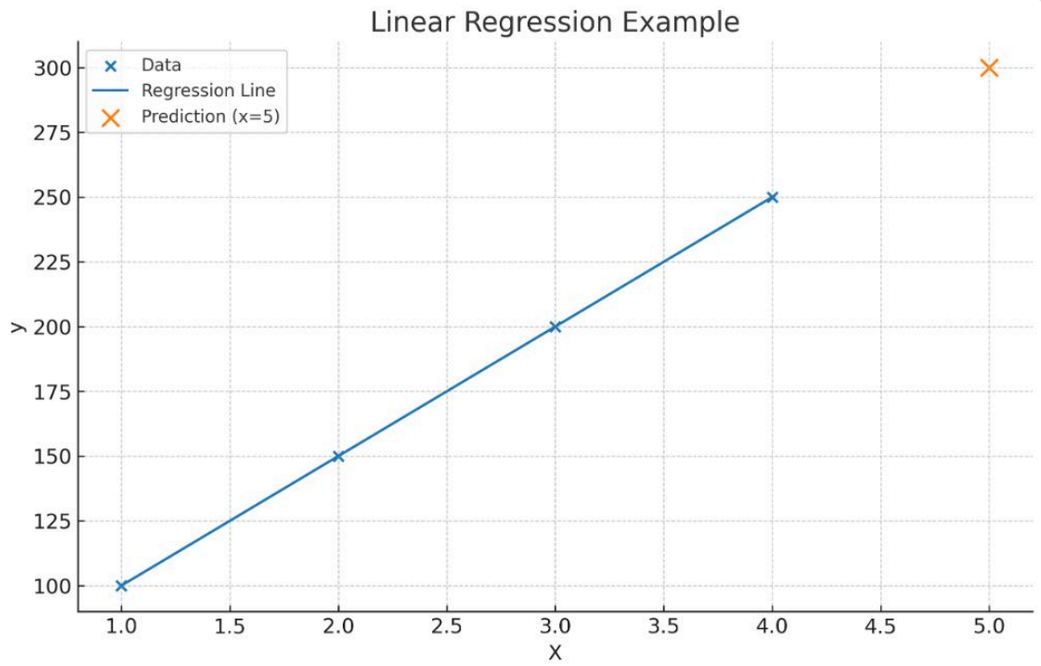
연령 분포 히스토그램 예시

EDA 핵심: 데이터를 분석하기 전에 시각화를 통해 분석 방향을 설정하고, 직관적으로 이해하는 과정입니다. '데이터를 알아가는 여정'으로 모든 고급 분석의 기초가 됩니다.

5~6단계. 모델링과 평가

모델링: 예측, 분류, 군집화 등 모델 구축

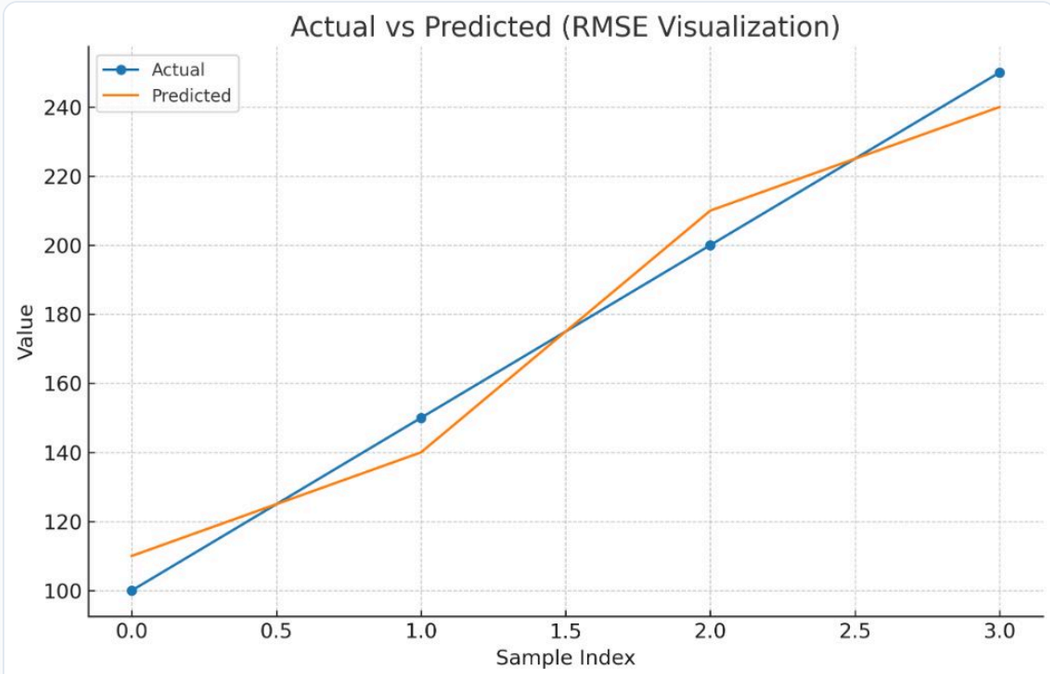
```
# 선형회귀 모델 예시
from sklearn.linear_model import LinearRegression
X = np.array([[1], [2], [3], [4]])
y = np.array([100, 150, 200, 250])
model = LinearRegression().fit(X, y)
print("예측값:", model.predict([[5]]))
```



선형 회귀 모델과 새 데이터 포인트 예측 (x=5일 때 y≈300)

평가 및 검증: 정확도, F1-score, RMSE 등 지표 활용

```
from sklearn.metrics import mean_squared_error
y_true = [100, 150, 200, 250]
y_pred = [110, 140, 210, 240]
rmse = mean_squared_error(y_true, y_pred, squared=Fals
print("RMSE:", rmse)
```



실제값과 예측값 비교 - 두 선 간 차이가 RMSE에 반영됨

핵심 포인트: 모델은 데이터 패턴을 학습하고, 평가는 모델 성능과 일반화 능력을 검증하는 과정

7단계. 결과 해석과 의사결정

- 모델 결과를 **실제 비즈니스 전략과 의사결정**으로 연결하는 단계
- 데이터 분석의 최종 목표는 **실질적인 가치 창출**

> 💡 **실제 활용 사례**
>
> 고객 이탈 예측 모델 → 이탈 가능성이 높은 고객 그룹 대상 집중 마케팅
>
> 제품 불량 예측 모델 → 생산 라인 최적화 및 품질 관리 전략 수립

💡 데이터사이언스는 **단순 분석을 넘어** 의사결정자의 행동을 이끌어내는 것이 핵심

데이터사이언티스트의 역할, 학습 가이드, 핵심 요약

데이터사이언티스트의 역할

- 데이터 수집 및 관리
- 데이터 분석 및 시각화
- 모델 개발 및 인사이트 도출
- 의사결정 지원 및 비즈니스 문제 해결

학습 가이드라인

- 수학·통계 기초 학습
- 프로그래밍 능력 강화 (Python, SQL, R)
- 작은 프로젝트 경험 축적 (Kaggle, 공공데이터)

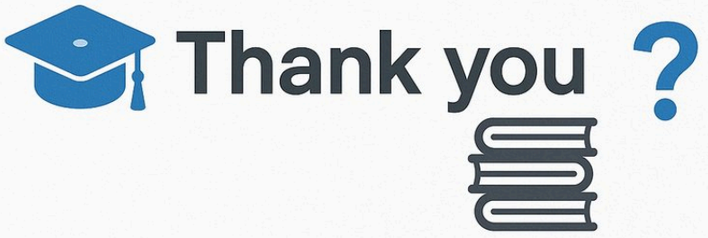
핵심 요약

- 데이터사이언스는 **데이터 기반 문제 해결**을 지원하는 융합 학문
- **7단계 프로세스:** 문제 정의 → 수집 → 전처리 → EDA → 모델링 → 평가 → 결과 해석
- 데이터사이언티스트는 **기술적 전문가**이자 **문제 해결자**로서 다양한 산업 분야에서 핵심 역할 수행
- 기술 역량 뿐 아니라 **비즈니스 이해력**과 **커뮤니케이션 능력**이 중요

다음 시간에는 Python 데이터 처리 기초를 다루겠습니다.

감사합니다.

감사합니다



수고하셨습니다

Q&A

질문이 있으신가요?

📧 교수 이메일: kafa46@hongik.ac.kr

🕒 상담 시간: 금 13:00-15:00, 연구실 D402호실 (사전 예약 필요)