Warm Up

왜 머신러닝을 공부해야 하나요?

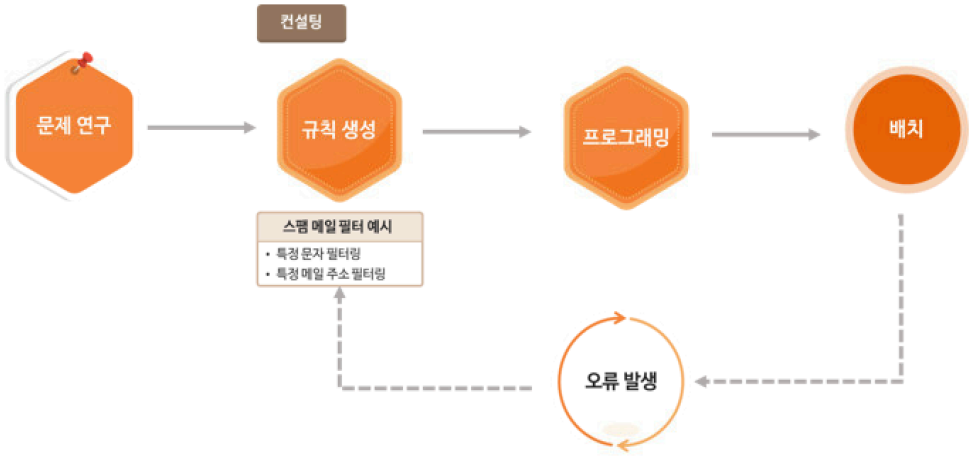
머신러닝을 통해 복잡한 문제에 대해 데이터 기반의 문제 해결을 수행할 수 있습니다.

1. **데이터 기반의 의사 결정**
    - 머신러닝은 데이터에서 패턴을 찾아내고 예측 모델을 구축하여 데이터 기반 의사 결정을 돕습니다.
2. **복잡한 문제 해결**
    - 머신러닝은 방대한 데이터를 분석하여 복잡한 문제를 해결하는데 탁월한 능력을 발휘합니다.

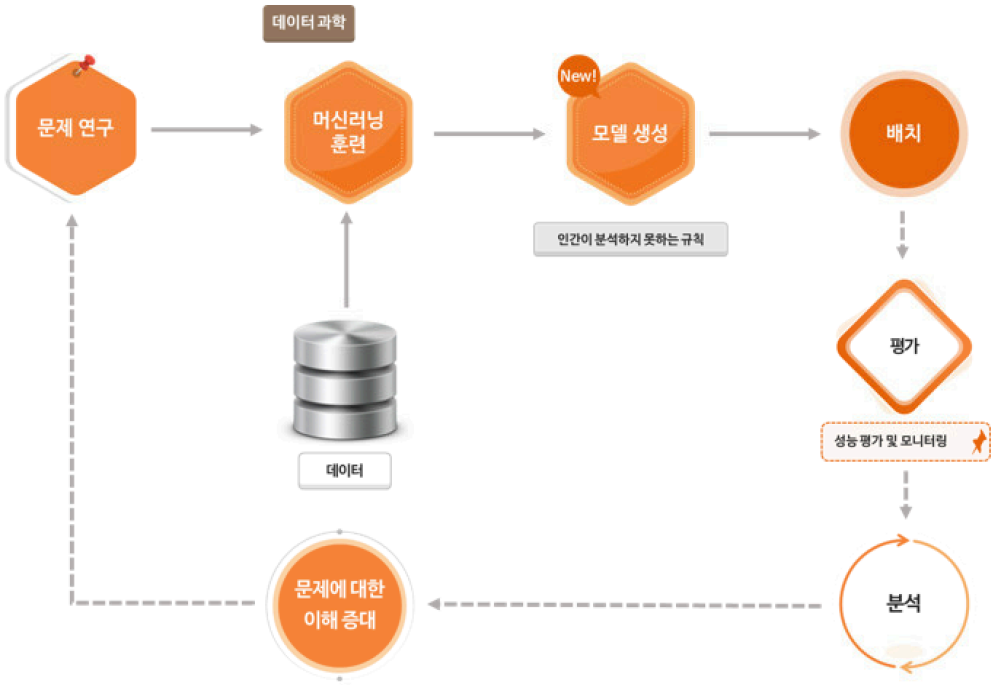
Example

> 스팸 메일 필터링에 대한 예시

1. **기존의 문제 해결 방법**



2. **머신러닝을 통한 문제 해결**



1. AI, Machine & Deep Learning

1.1 Terminology

1.1.1 Statistics

👀 **Definition**

> 데이터를 수집, 분석, 해석 및 시각화 하는 학문, 확률이론과 수학적 기법을 활용하여 데이터의 패턴을 분석하고 결론을 도출함

- 데이터의 특성과 관계를 파악함
- 학문적 이론을 바탕으로 예측을 수행함

🗣️ **Example**

- 기술 통계
  - 데이터를 요약하고 설명하는 통계 방법으로 평균, 중앙값, 표준편차와 같은 통계량으로 데이터를 이해
  - 회사의 매출 데이터를 분석하여 월별 평균 매출, 매출의 중앙값 등을 파악하고 이를 시각화하여 보고
  - 향후 EDA(탐색적 데이터 분석) 과정에서 주로 활용 됨
- 추론 통계
  - 표본 데이터를 통해 모집단에 대한 추론을 수행
  - 신제품 출시 전에 소규모 표본 고객의 반응 데이터를 수집하고 이를 바탕으로 시장 전체의 반응을 추정

1.1.2 Machine Learning

👀 **Definition**

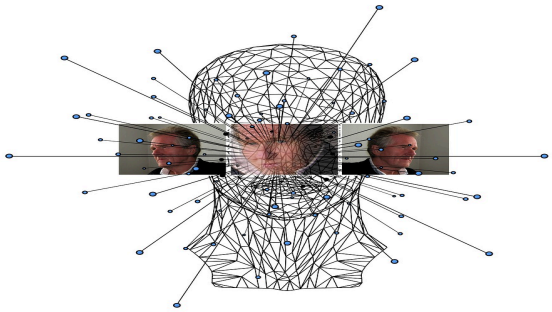
> 컴퓨터가 데이터로부터 패턴을 학습하고 학습한 패턴을 바탕으로 새로운 데이터를 예측하거나 분류하는 알고리즘을 개발하는 학문

- 데이터로부터 패턴을 학습함

🗣️ **Example**

- 지도 학습
    - 입력에 대한 답을 알려주고 패턴을 찾아냄
    - 키를 통해 몸무게를 예측 모델, 스펨 메일을 찾아내는 분류 모델
- 비지도 학습
    - 데이터 자체만을 보고 패턴을 찾아냄
    - 고객 구매 데이터에서 비슷한 구매 행동을 가진 고객을 그룹화

1.1.3 Deep Learning



👀 Definition

> 인공신경망을 기반으로 여러 층의 신경망을 통해 데이터를 학습하는 머신러닝의 하위 분야

- 인공신경망은 거대한 행렬 연산
- 머신러닝으로 찾기 힘든 복잡한 데이터의 패턴을 학습함

🗣️ Example

- 자연어 처리(NLP)
    - 텍스트 데이터에서 의미를 추출하고 이해하는 모델을 학습함
    - 주어진 언어를 인식하고 다른 언어로 변환하는 번역 시스템
- 이미지 처리(Vision)
    - 이미지 데이터를 분석하여 물체나 사람을 인식하고 분류함
    - 자율주행 기술에서 도로의 표시판, 보행자, 다른 차량 등을 인식하여 안전하게 주행하는데 도움을 줌

1.1.4 Artificial Intelligence(AI)

👀 Definition

> 컴퓨터 시스템이 인간의 지능적 활동을 모방하여 문제를 해결하거나 결정을 내리는 기술

- 딥러닝, 머신러닝 등의 기술을 활용해 인간을 보조함

🗣️ Example

- 거대 언어 모델(LLM)을 통해 업무를 도와주는 Chat GPT
- 이미지 인식 모델을 통해 객체를 인식하여 안전하게 운전을 하는 자율 주행 시스템

1.2 Relationship

1.2.1 Statistics vs Machine Learning

- 통계학은 머신러닝의 기초를 제공함
- 머신러닝 알고리즘의 많은 부분은 통계적 방법론과 이론에 기반하고 있음
- 통계학은 데이터의 이해와 해석에 중점을 두고 있으나 머신러닝은 데이터로부터 패턴을 학습하여 예측 결과를 도출하는 모델을 만드는데 초점을 맞춤

📍 More Detail

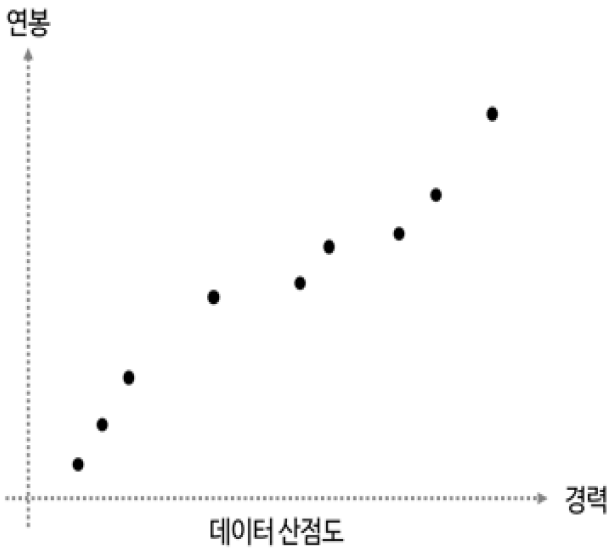
- **머신러닝**
  - 예측의 성공 확률을 높이는데 목적
  - 모델의 신뢰도나 정교한 가정은 상대적으로 중요성이 낮음
- **통계 분석**
  - 학술적인 이론을 통해 실패 확률을 줄이는데 목적 ☞ 많은 가정(제약)이 포함됨
  - 복잡한 것 보다는 단순성을 추구하며 모델의 신뢰도가 중요함

🗣️ Example

> 통계학과 머신러닝에서의 회귀 분석 관점의 차이

회귀 분석은 독립 변수와 종속 변수사이의관계를 찾는 방법입니다. 아래는 회귀 분석`을 설명하기 위한 간단한 데이터 집합의 예시 입니다.

독립변수	종속변수
경력	연봉
2	15
3	28
5	42
13	64
8	50
16	90
11	58
1	8
9	54

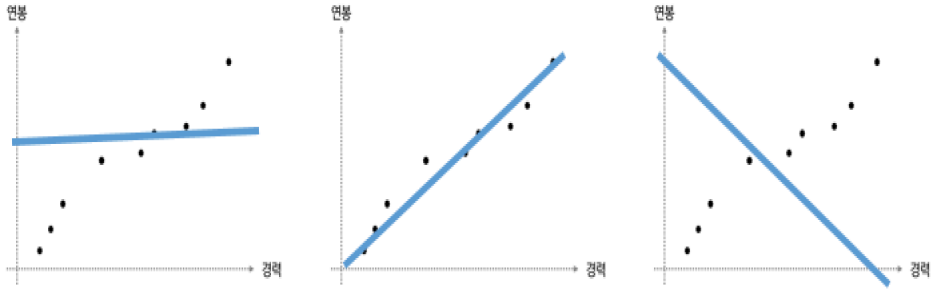


위 데이터에 회귀 분석을 사용하면, 독립 변수인 경력과 종속 변수인 연봉 사이의 관계를 찾을 수 있습니다. 관계를 찾게되면 향후 독립 변수인 경력에 대한 새로운 데이터가 들어왔을 시, 관계에 의거해 종속 변수인 연봉을 예측할 수 있게 됩니다. 위 산점도 그래프를 통해 관계를 예측할 수 있으신가요?

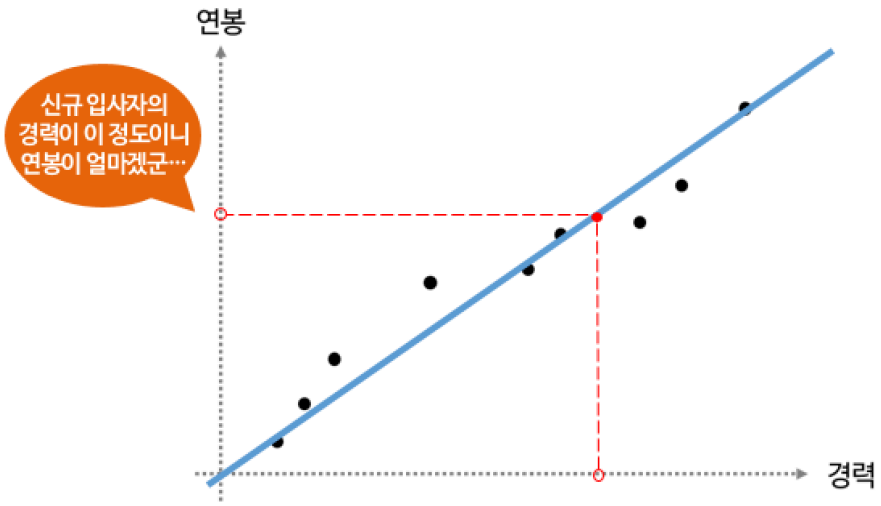
📍 독립변수 vs 종속변수

특징	독립 변수 (Independent Variable)	종속 변수 (Dependent Variable)
정의	실험이나 연구에서 조작하거나 통제하는 변수	독립 변수의 변화에 의해 영향을 받는 변수
역할	원인이나 입력 변수로, 결과에 영향을 미치는 변수	결과나 출력 변수로, 독립 변수의 영향을 받는 변수
예시	약물 실험에서 투약량, 교육 연구에서 교육 방법	약물 실험에서 환자의 혈압 변화, 교육 연구에서 학생의 성적
수학적 표현	모델에서 입력 변수로 사용 (X)	모델에서 예측 또는 설명하려는 출력 변수 (Y)
종속 관계	독립 변수는 다른 변수에 의해 영향을 받지 않음	종속 변수는 독립 변수에 의해 영향을 받음

회귀 분석은 관계를 수치적으로 명확히 알기 위해 관계를 잘 나타내는 선을 긋는 것(=line fitting)을 목표로 합니다. 아래와 같이 그어진 선 중에 어떤 것이 관계를 가장 잘 나태낸 것 처럼 보이시나요?

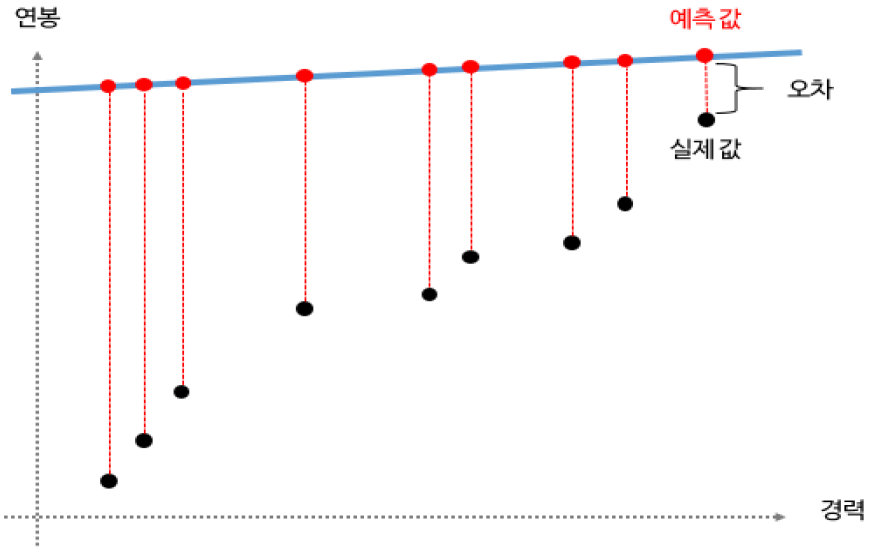


선과 데이터 간의 차이가 작을 수록 관계를 잘 나타내는 것으로 보입니다. 이때 선과 데이터의 차이는 오차라고 합니다. 이렇게 관계를 잘 나타내는 선은 새로운 데이터가 들어왔을때, 아래와 같이 관계에 의거해 아마도 이럴것이다 라는 답하는 예측의 역할을 할 수 있습니다.



📍 **오차를 수치화 하는 방법**

- 가장 대표적인 방법은 MSE(Mean Squared Error)
- MSE: 실제 값과 예측 값 사이의 차이(오차)의 제곱을 평균한 값



그렇다면 통계학과 머신러닝에서 회귀 분석을 바라보는 관점은 어떻게 다를까요? 회귀 분석은 관계를 잘 표현한 선을 긋는 것입니다. 이렇게 선을 잘 긋기 위해서 오차를 줄여야 함은 통계학과 머신러닝 둘다 같은 입장이지만, 오차를 줄이는 방식에 있어서 서로 다른 방식을 채택하여 사용합니다.

특징	통계학의 OLS	머신러닝의 Gradient Descent
목표	데이터의 선형 관계를 설명하고 변수 간의 관계 해석	선형 및 비선형 관계를 포함한 예측 성능 최적화
모델	선형 모델	선형 및 비선형 모델
오차 최소화	잔차 제곱합(SSR) 최소화	손실 함수(MSE) 최소화
계산 방법	수학적 명시적 해 계산	반복적인 최적화
모델의 해석 가능성	회귀 계수의 통계적 해석 용이	복잡한 모델에서는 해석이 어려울 수 있음
적용 가능성	작은 데이터셋과 선형 관계에 적합	대규모 데이터와 복잡한 관계에 적합

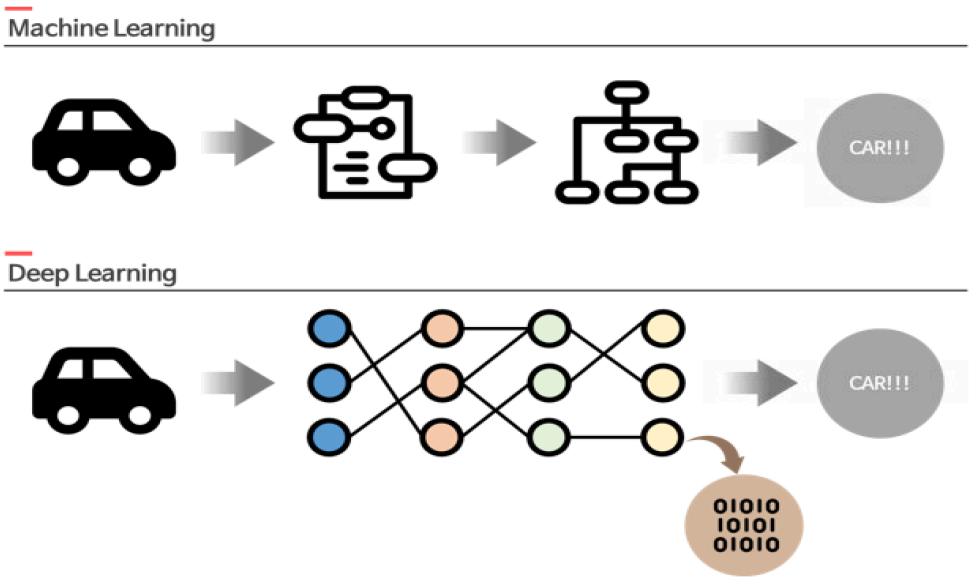
즉, 회귀 분석은 통계학을 통해 탄생했고, 오차를 최소화 하는 개념을 머신러닝이 활용하는 것으로 머신러닝의 기반은 통계로부터 파생되었음을 알 수 있습니다.

1.2.2 Machine Learning vs Deep Learning

- 딥러닝은 머신러닝의 하위 분야
- 딥러닝은 더 깊고 복잡한 신경망 구조를 사용하여 고차원 데이터와 복잡한 패턴을 학습

🗣️ **Example**

> 이미지 인식에서 머신러닝과 딥러닝의 수행 프로세스 차이



1.2.3 Machine Learning vs AI

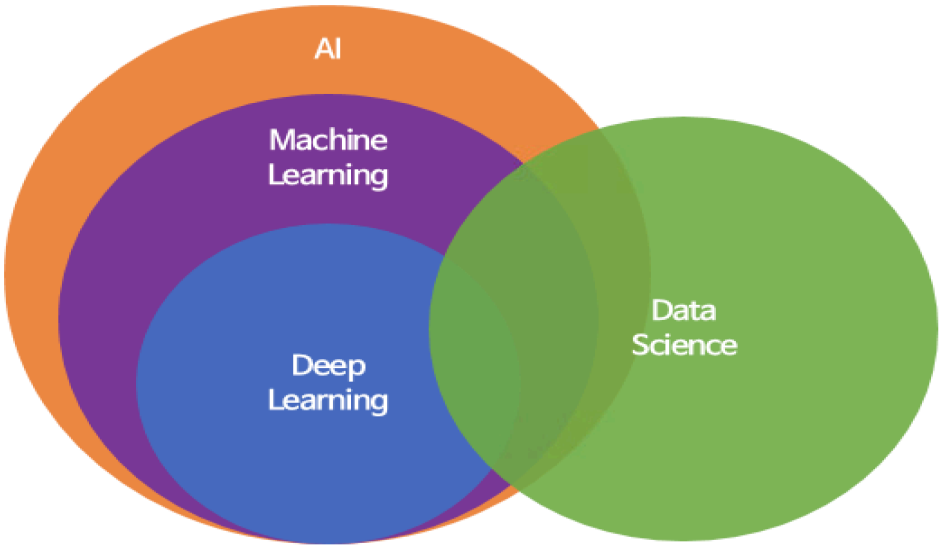
- 머신러닝은 AI의 하위 분야
- 머신러닝은 AI의 다양한 응용 중에서 데이터를 기반으로 학습하고 예측하는 기능을 담당
- AI는 머신러닝을 포함하여 다양한 기술과 방법을 사용하여 인간의 지능적 작업을 수행

1.2.4 Deep Learning vs AI

- 딥러닝은 AI의 하위 분야
- 복잡한 문제를 해결하기 위한 고차원 데이터의 학습에 특화된 기술을 제공
- 딥러닝의 발전을 통해 AI는 성장을 가속화하고 이미지, 음성, 자연어 등 다양한 분야에서 활용할 수 있게 함

📌 **Compare with Data Science**

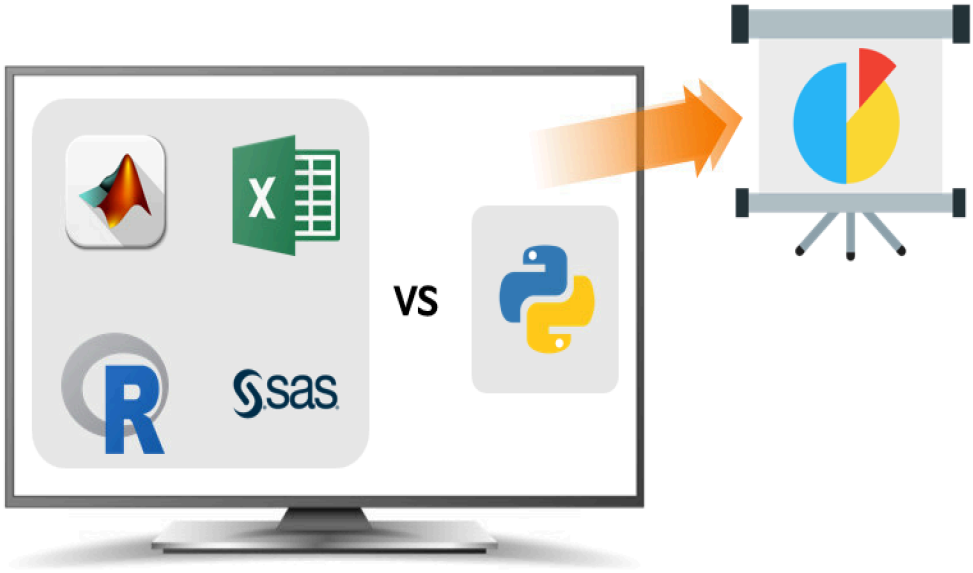
> AI, Machine Learning, Deep Learning 그리고 Data Science 간의 관계



- Data science와 Machine Learning은 매우 유사함
- Data science는 설명을 목적으로 함 (통계 분석)
- 머신러닝은 어플리케이션 활용을 목적으로 함 (모델링)

1.3 Python

1.3.1 Why use python for data analysis?



1. **가장 직관적이고 배우기 쉬움**
    - 코딩을 배워보지 않은 사람들도 거부감 없이 접근 가능

    📍 **타 언어와 비교**

    Python

    ```
    amount = 0
    print(amount)
    ```

    Java

    ```
    import java.lang.*;

    public class Main {
        public static void main(String[] args) {
            int amount = 0;
            System.out.println(amount);
        }
    }
    ```

2. **대용량 데이터 처리가 가능함**
    - 엑셀의 경우 대용량 데이터의 처리 속도가 느림
3. **다양한 라이브러리를 지원**
    - 기초 문법을 배우긴 어려우나 지원하는 라이브러리(툴)들이 다양함
    - 원하는 알고리즘을 가져다 쓰는 것이 용이함

4. **딥 러닝 프레임워크와의 연계**
    - 더 복잡한 분석들을 사용할 수 있음
5. **확장성이 좋음**
    - 다른 분석 툴과는 다르게 웹, 앱 등 에 배포하여 사용할 수 있음
6. **커뮤니티 활성화**
    - 질문과 답을 얻을 수 있는 공간이 다양함

📼 **Audiovisual Materials**