

# Data Science Case Studies

노기섭 교수

([kafa46@hongik.ac.kr](mailto:kafa46@hongik.ac.kr))

# Lecture Goals

- 다양한 산업 문제를 어떻게 해결하는지 개요
- 금융 도메인 사례(사기 탐지, 신용평가)
- 의료 도메인 사례(질병 예측, 의료 영상)
- 산업 도메인 사례(예지보전, 비전 기반 품질검사)
- 미니 캡스톤 (팀 프로젝트)

# 개 요

# Introduction

## ■ 데이터 사이언스의 확장

- 더 이상 특정 기술 기업만의 전문 분야에 국한되지 않음
- 금융, 의료, 제조, 유통, 공공 서비스 등 다양한 산업에서 핵심 경쟁력을 만들어내는 필수 역량

## ■ 데이터 사이언스의 기여

- 데이터 기반 분석과 예측 모델은 기업의 운영 비용을 절감
- 새로운 매출 창출 기회를 발굴
- 의료·안전 분야에서는 사람의 생명을 보호하는 데까지 발전

## ■ 데이터 사이언스를 제대로 바라보는 관점

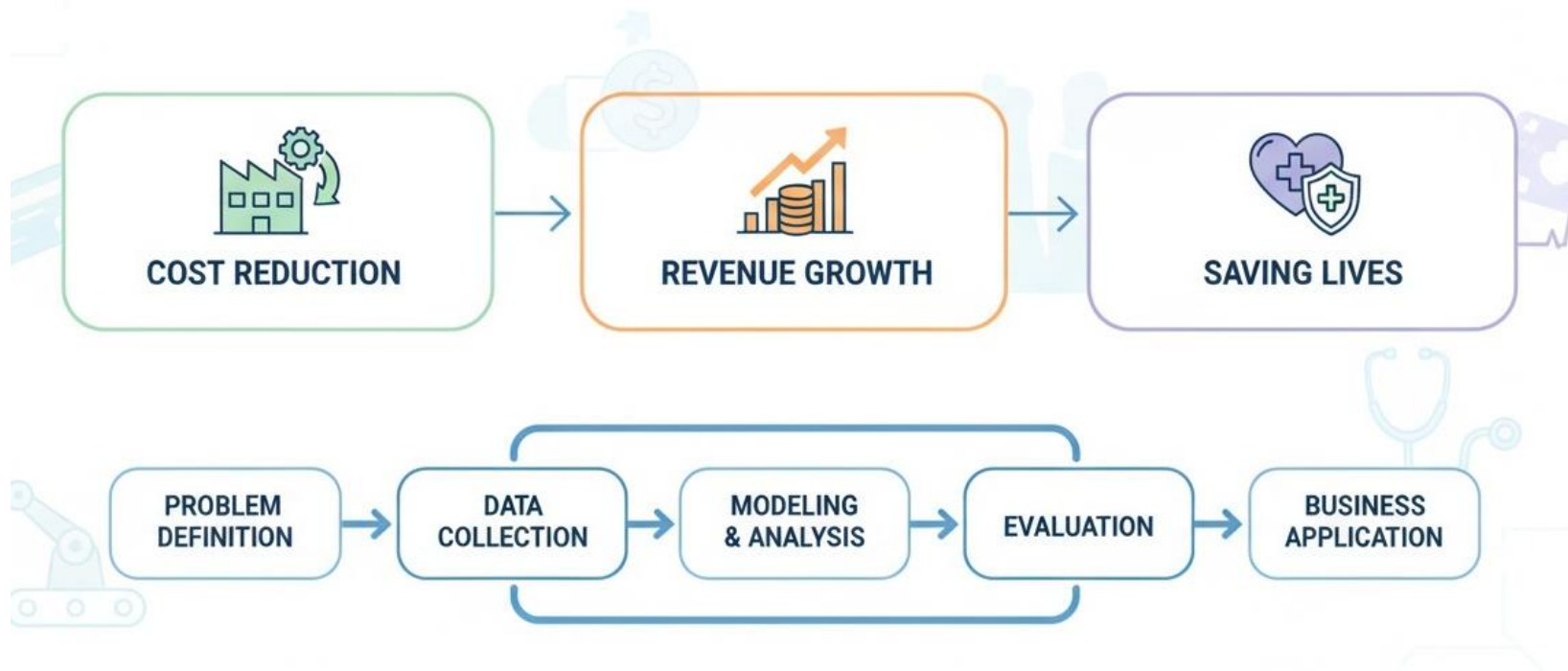
- 현업의 실제 문제를 어떻게 해결하고, 어떤 가치를 창출하는지를 파악하는 것이 중요
- 현장에서의 문제 정의, 데이터 수집, 모델링, 평가, 그리고 비즈니스 적용까지의

**전 과정 전체 흐름의 이해하는 것**

# 데이터 사이언스의 흐름

- 문제 정의 → 데이터 수집 → 모델링·분석 → 평가 → 비즈니스 적용
- 현실 세계의 문제 해결 과정에서 데이터가 어떻게 활용되는지 파악

## WHY DATA SCIENCE MATTERS



# Finance Case Studies



# Credit Card Fraud Detection

## Finance Case Studies (1/2)

# Credit Card Fraud Detection



# Credit Card Fraud Detection

## ■ 신용카드 사기 탐지

- 신용카드 사기 거래는 매년 증가
- 금융기관은 실시간 탐지를 통해 고객 피해와 금융 손실을 최소화해야 할 책임이 있음.

## ■ 데이터 사이언스

- 거래 데이터를 분석하여 정상 패턴과 비정상 패턴을 구분
- 사기 거래를 조기에 발견하는 데 중요한 역할
- 이상 거래 탐지 기법과 분류 모델을 적용해 실제 금융 서비스 환경에서 활용
- 목표
  - 신용카드 거래 내역 중 사기 거래를 실시간으로 탐지하여 피해를 최소화
- 고려사항
  - 사기 거래는 전체 거래 대비 매우 적은 비율로 발생하는 불균형 데이터
  - 일반적인 정확도(Accuracy)만으로는 모델 성능을 적절히 평가하기 어려움



# 데이터 예시 1. 거래 로그 (결제 시간, 결제 방식, 거래 지역)

## CREDIT CARD TRANSACTION LOGS: A DATA SCIENCE PERSPECTIVE



WHEN?



Transaction Frequency by Hour



Timestepping each event

HOW?



Method of Payment

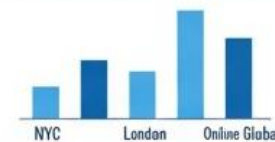
Payment Method Breakdown:



WHERE



Geographic Coordinates/Merchant Location

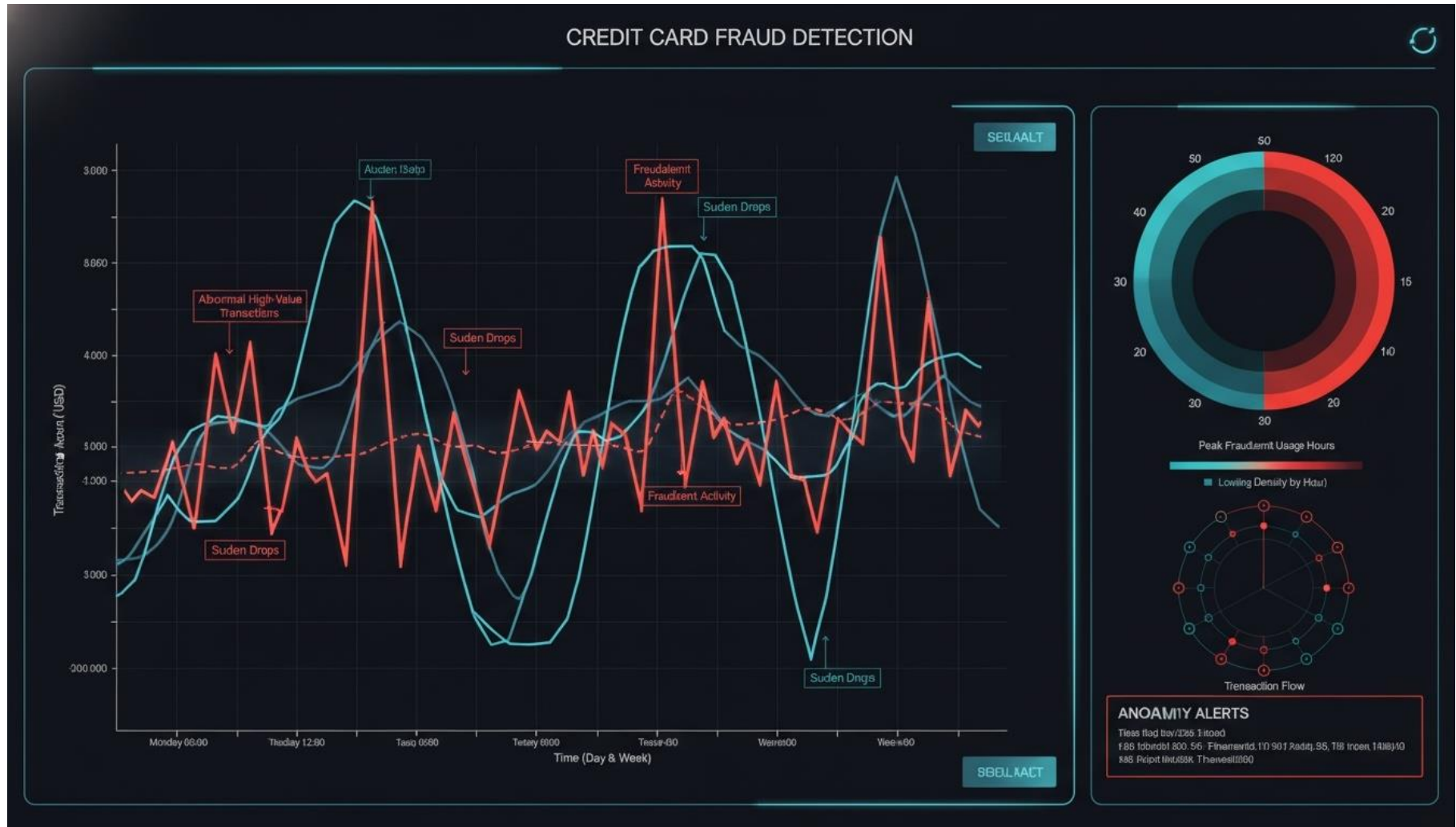


Top Locations of Volume:

**DATA SCIENCE CONTEXT:** Raw transaction data forms a chronological, multi-dimensional log for fraud detection, behavioral analytics & market Basket analysis



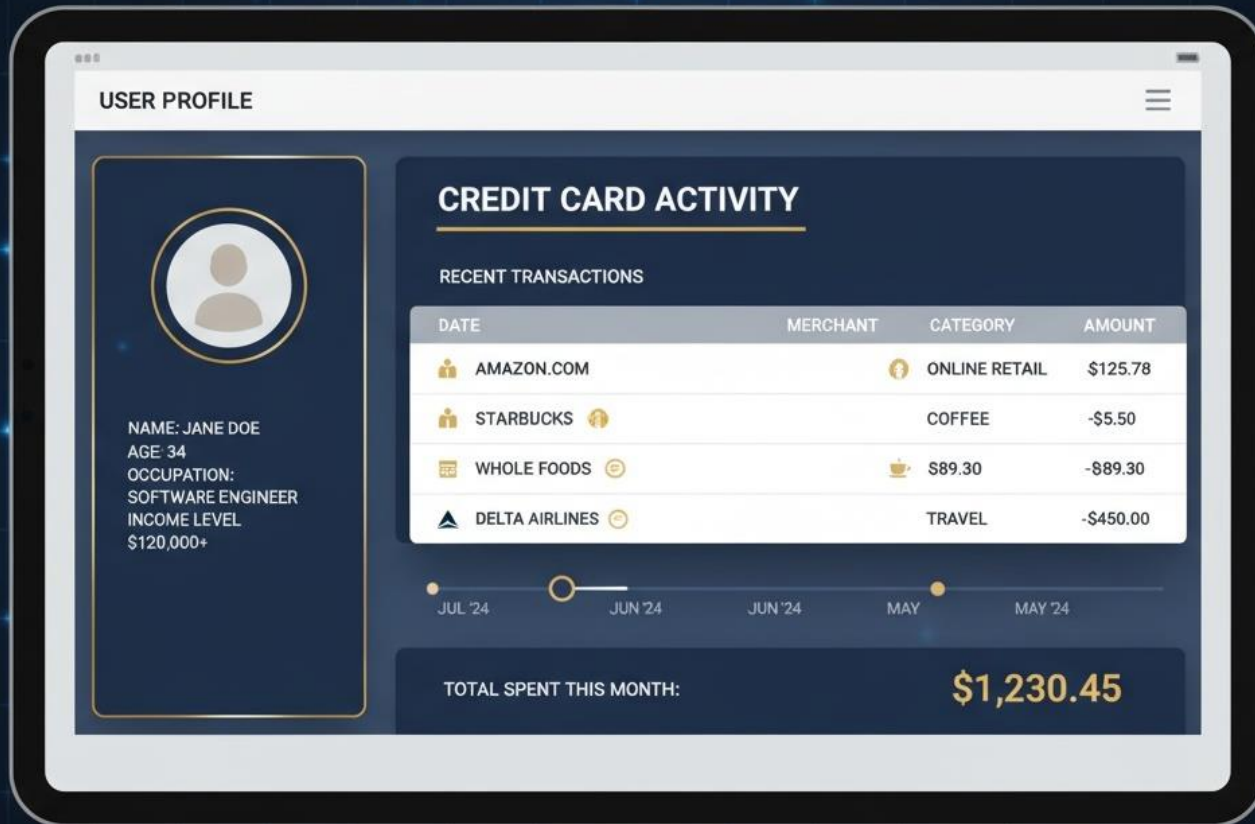
## 데이터 예시 2. 결제 금액 및 사용 패턴



# 데이터 예시 3. 디바이스 정보 (사용 기기, IP, 위치 변화)



# 데이터 예시 4. 사용자 프로파일 및 과거 거래 이력



# 신용 카드 사기 - 분석 기법

## ■ 이상치 탐지 기법

- 정상 패턴에서 벗어난 거래를 탐지
  - [Isolation Forest](#)
  - [Local Outlier Factor\(LOF\)](#)

## ■ 분류 모델 기반 탐지 (지도학습)

- 정상 또는 사기로 분류
  - [Logistic Regression](#)
  - [Random Forest](#)
  - [XGBoost](#)

# 성능 평가 및 분석 방법

## ■ 평가 지표

- Accuracy보다 Precision과 Recall의 균형이 중요
- 특히 사기 거래를 놓치지 않는 Recall이 핵심
- 다만, Recall만 높이고 Precision이 낮아지면 정상 거래 차단이 불필요하게 증가
  - 두 지표 간의 적절한 균형 조정이 필요
- 불균형 데이터에서는 Precision·Recall을 함께 보는 F1이 유용하지만, 사기 탐지처럼 Recall을 더 중시할 때는  $\beta > 1$ 로 가중하는  [\$F\_\beta\$ -score](#)나 [PR-AUC](#) 등을 함께 검토한다.
  - [\$F\_\beta\$ -score](#)나 [PR-AUC](#) : 조금 더 구체적으로 알아보기 → 다음 슬라이드
  - $F_\beta$ -score: Precision과 Recall을  $\beta$  가중 조화평균으로 묶어, Recall 중요도를 반영한 단일 지표이다.  $\beta = 1$  이면 F1과 같고,  $\beta > 1$ 이면 Recall을 더 강조하고  $\beta < 1$ 이면 Precision을 더 강조
  - PR-AUC

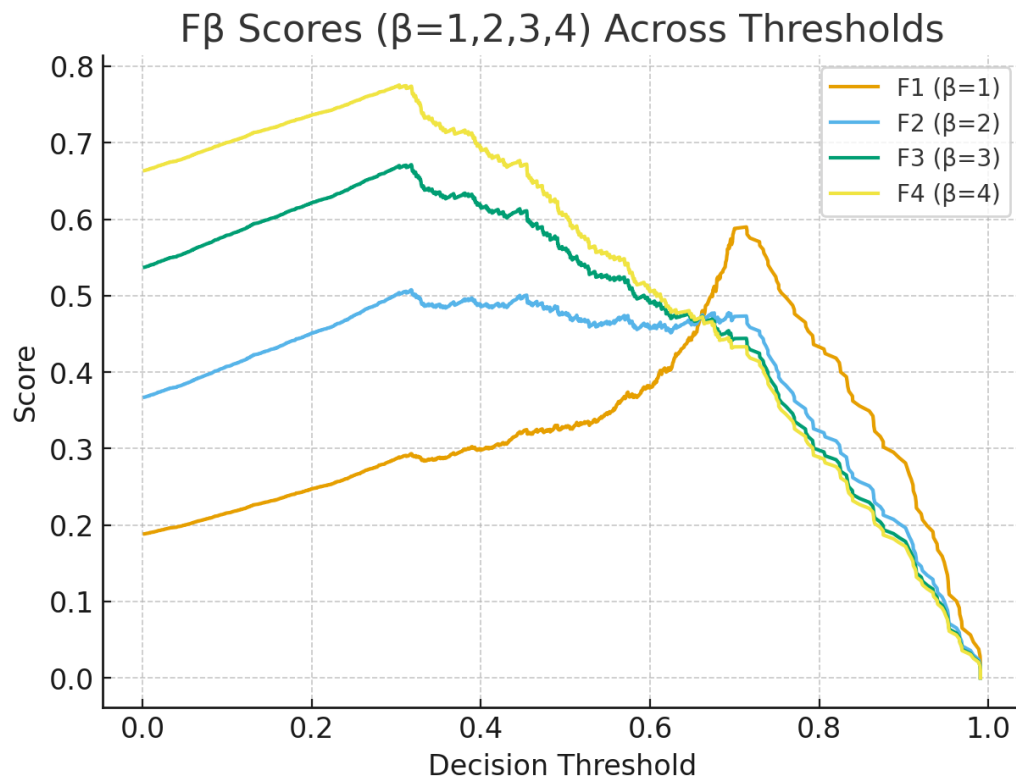
# $F_\beta$ - score

## ■ $F_\beta$ - score

Precision과 Recall을  $\beta$  가중 조화평균으로 묶어,

Recall 중요도를 반영한 단일 지표이다.

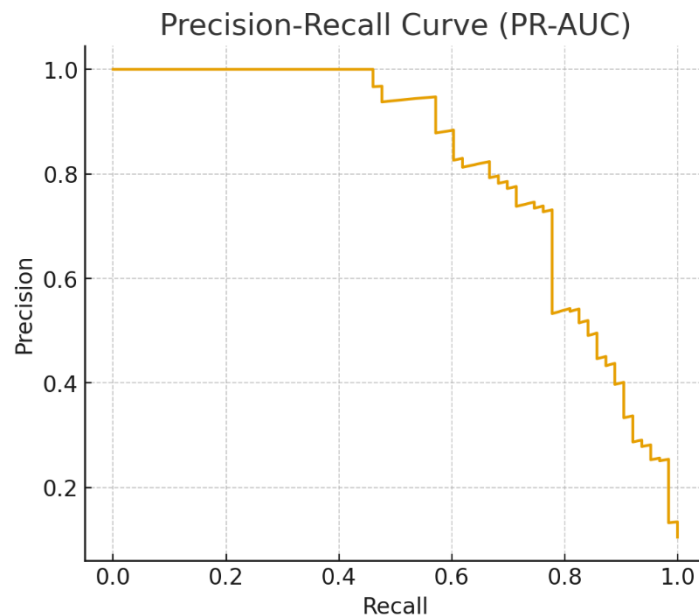
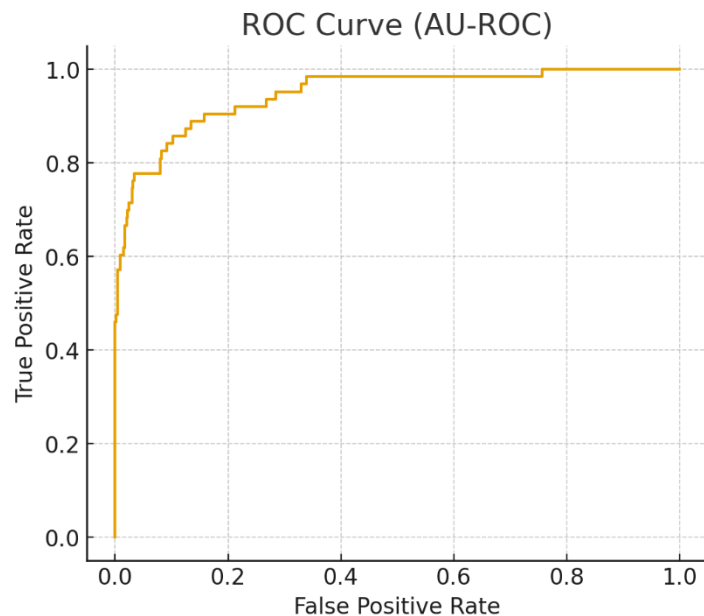
- $\beta = 1$  이면 F1과 같다.
- $\beta > 1$ 이면 Recall을 더 강조
- $\beta < 1$ 이면 Precision을 더 강조



# PR-AUC

## ■ PR-AUC

- PR-AUC (Precision-Recall Curve): Precision-Recall 곡선 아래 면적
  - 클래스 불균형 상황에서 전 점수 구간의 분류 성능을 요약하며,
  - AU-ROC(예측 순위 기반 TPR/FPR)보다 양성 클래스 품질에 민감





# 성능 평가 및 분석 방법

## ■ 기대 효과 및 분석 결과

- 적절한 모델 선택과 임계값 조정을 통해 사기 거래 탐지율을 높이면서 정상 거래 오탐률을 효과적으로 낮출 수 있는 전략을 수립
- Recall을 높여 사기 거래 탐지 능력을 향상시키되, Precision을 유지하여 정상 거래가 불필요하게 차단되는 상황을 최소화하는 것이 금융 서비스 품질 유지의 핵심.

# Mini Project

## ■ Kaggle의 Credit Card Fraud Detection 데이터셋 활용

- [데이터셋 다운로드 \(Kaggle\)](#)

## ■ 데이터 불균형 문제 해결 기법 적용

- [언더샘플링 \(RandomUnderSampler\)](#)
- [오버샘플링 \(RandomOverSampler\)](#)
- [SMOTE](#)

## ■ [Logistic Regression](#), [Random Forest](#), [XGBoost](#) 성능 비교

- [Random Forest](#), [XGBoost](#) 적용은 선택사항

## ■ F1 Score 및 AUC 기반 성능 평가

## ■ 임계값 변경 시 Precision과 Recall 변화 시각화 분석



# Credit Scoring

## Finance Case Studies (2/2)

# Credit Scoring



# Credit Scoring

## ■ Credit Scoring

- 대출 심사 과정에서 금융기관은 고객의 신용도를 평가하여 대출 승인 여부를 결정
- 데이터사이언스 기반의 신용평가 모델
  - 기존의 단순 점수 기반 방식보다 더 정확하고 일관성 있는 의사결정을 내릴 수 있도록 지원
  - 모델이 내린 판단이 공정하며 설명 가능해야 함.
  - 성능과 함께 해석력 역시 중요한 요소



# 문제 정의

## ■ 목표

- 고객의 대출 신청에 대해 승인 또는 거절 여부를 정확하고 공정하게 예측하는 것

## ■ 필수 고려사항

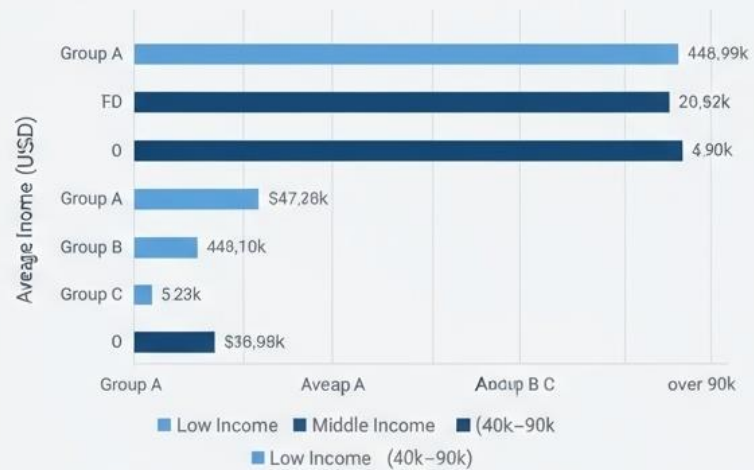
- 특정 개인에게 불리하지 않도록 판단해야 함.
- 일부 집단에게 불리하게 작동하는 편향된 모델이 만들어지지 않도록
  - 데이터 선정, 모델 학습, 평가 기준에서 공정성이 고려해야 함.

# 데이터 예시 1. 소득 수준

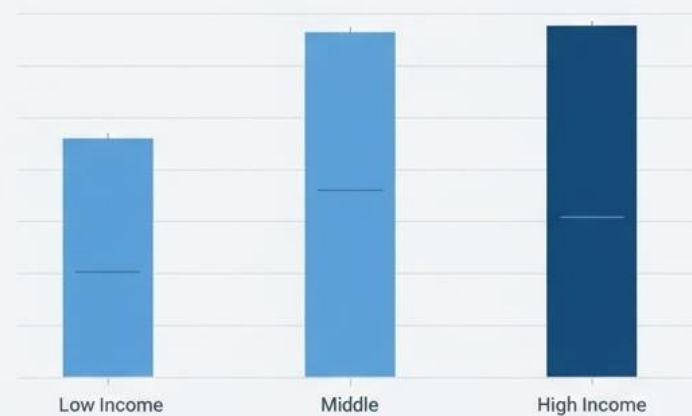
## CREDIT SCORING INCOME ANALYSIS

Applicant Income Levels & Distribution

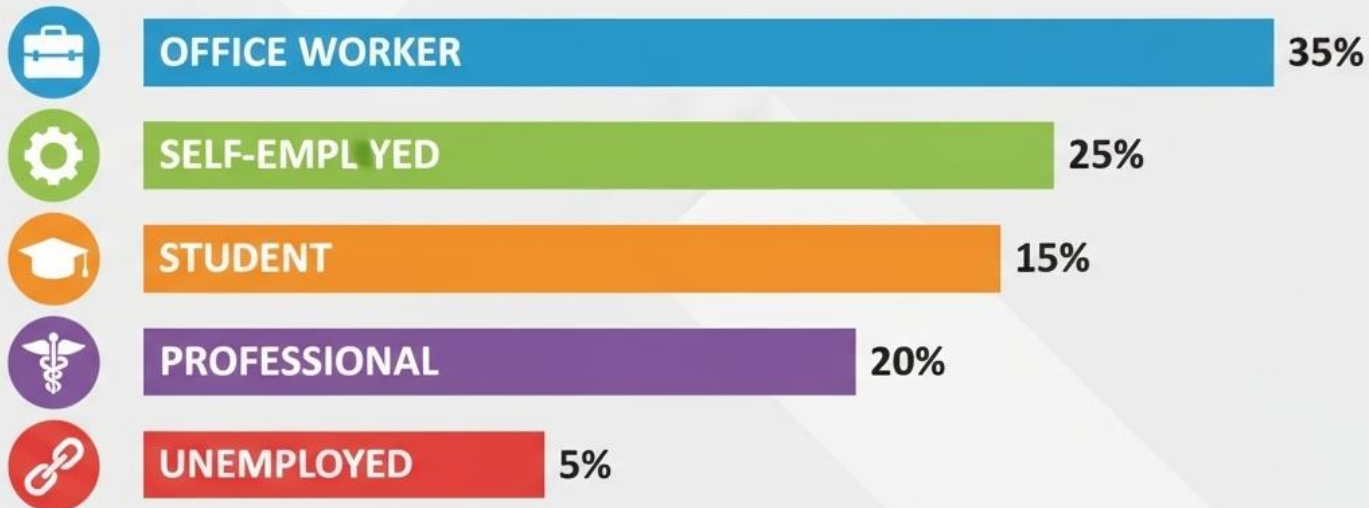
Applicant Groups



Income Distribution by Level



### APPLICANT OCCUPATION FOR CREDIT SCORING



#### IMPACT ON CREDIT SCORE:

↑ Higher income stability generally leads to better scores. ✓

# 데이터 예시 3. 연체 기록

## LOAN & CREDIT PAYMENT DELINQUENCY HISTORY



## Late Payments by Applicant Category





# 데이터 예시 4. 부채 비율

## DEBT-TO-INCOME (DTI) RATIO



## APPLICANT FINANCIAL COMPARISON

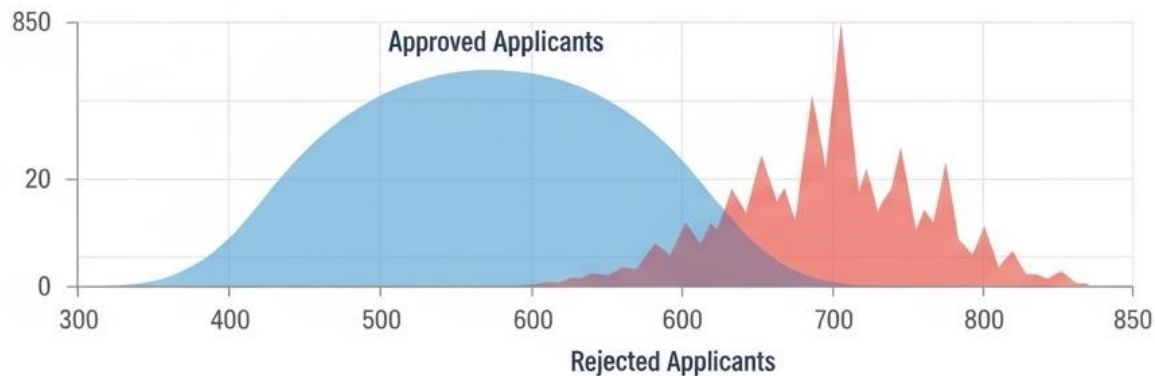
$$DTI = (\text{Monthly Debt} / \text{Monthly Income}) \times 100$$



**38%** Applicant's DTI Ratio

# 데이터 예시 5. 신용 등급

## CREDIT SCORE DISTRIBUTION & LOAN APPLICANT ANALYSIS



# 분석 기법

## ■ 분석 접근법 및 전략

- [Logistic Regression](#)

비교적 단순한 모델이지만 해석력이 뛰어나며, 어떤 변수들이 대출 승인 결정에 영향을 미치는지 명확히 확인할 수 있다.

- [LightGBM](#)

대용량 데이터 학습에 강점이 있으며 높은 예측 성능을 보이는 부스팅 모델이다. 다만, 모델 구조가 복잡하여 해석이 어려울 수 있다.

- 예측 결과에 "설명 가능성(XAI) 기법"을 적용하여 모델의 판단 근거를 확인

- 편리하게 사용할 수 있는 도구: [SHAP](#)

- 고객 속성에 따른 차별적 영향 여부를 분석하며, 금융 공정성(Fairness) 관점에서 모델 평가

# Mini Project (팀 수행)

## ■ 2명 이상이 팀으로 진행

## ■ 토론 결과는 보고서로 작성하여 제출할 것

- 제출 파일: .hwp, .docx

## ■ 토론 주제

- 해석력이 높은 모델과 성능이 높은 모델 중 금융권에서는 어떤 모델을 선택해야 하는가?
- 특정 연령, 직업, 성별 집단에 대해 모델이 불리하게 작동한다면 어떻게 개선할 수 있는가?
- XAI가 실제 금융 의사결정 과정에서 어떤 역할을 할 수 있는가?

# Healthcare Case Studies



## Disease Prediction

# Healthcare Case Studies (1/2)

## Disease Prediction



# Disease Prediction

## ■ Disease Prediction

- 의료 분야에서 데이터 사이언스
  - 질병 발생 가능성을 조기에 예측
  - 예방적 치료와 맞춤형 관리가 이루어질 수 있도록 지원
- 특징
  - 당뇨병, 심장 질환과 같은 만성 질환은 조기 발견 여부가 향후 치료 효과와 생존율에 큰 영향
  - 예측 모델의 정확성과 의료진 신뢰 확보가 중요

# 문제 정의

## ■ 문제 정의

- 환자의 건강 데이터를 기반으로 질병 발생 가능성을 사전에 예측하는 것이 목표
- 의료 분야에서는 질병 여부를 놓치지 않는 것이 매우 중요
- 정확도(Accuracy)보다 민감도(Recall)가 더 중요한 평가 기준으로 고려



# 데이터 예시 1. 건강검진 수치 (혈압, 혈당, 콜레스테롤 등)

## + HEALTH METRICS FOR DISEASE RISK ASSESSMENT

### BLOOD PRESSURE (mmHg)



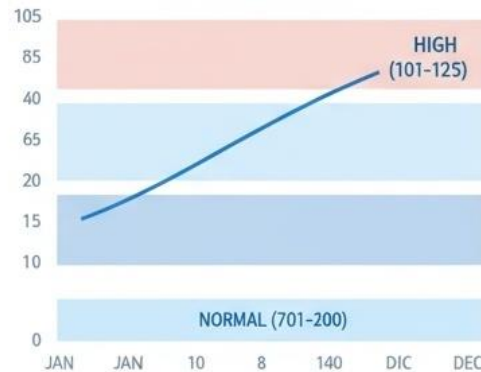
DYSTICAL / >160/100)



### BLOOD GLUCOSE (mg/dL)



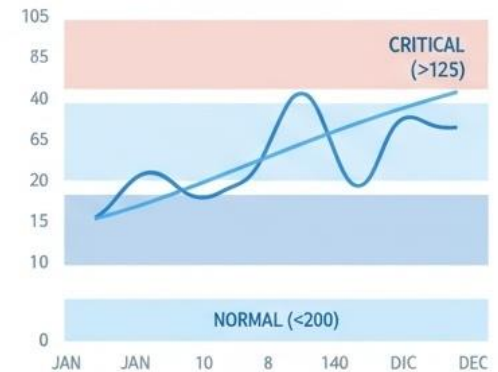
HIGH / COMPARSION



### CHOLESTROL (mg/dL)



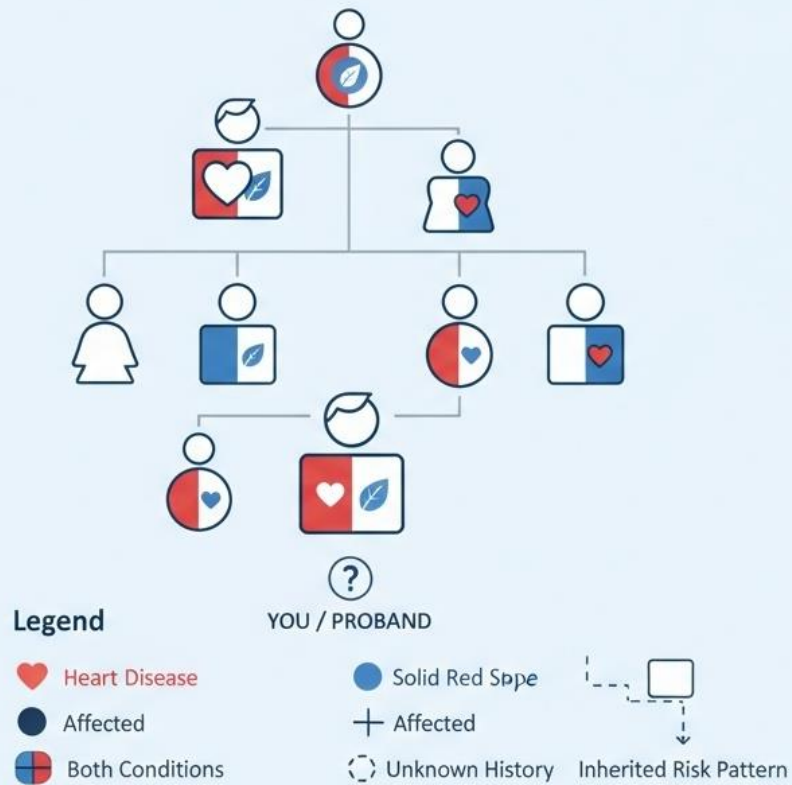
TREND / COMPARSION



## 데이터 예시 2. 가족력 정보

### FAMILY HEALTH COMPASS

Disease Risk Assessment & Inherited Patterns

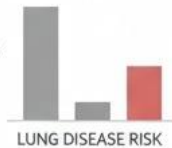
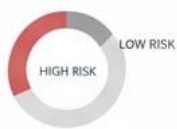


# 데이터 예시 3. 라이프스타일 정보 (흡연, 음주, 운동량 등)

## LIFESTYLE & DISEASE RISK

### HABITS

#### SMOKING HABITS



### CONSUMPTION

#### ALCOHOL CONSUMPTION



### ACTIVITY

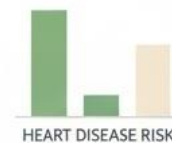
#### EXERCISE FREQUENCY



#### HEART DISEASE RISK



#### ACTIVE



#### UNHEALTHY LIFESTYLE



INCREASED DISEASE RISK  
(Heart, Diabetes, Cancer)



#### HEALTHY LIFESTYLE



DECREASED DISEASE RISK  
(Wellness, Longevity)

# 분석 기법

## ■ 분류 모델 비교

- [SVM](#)과 [Random Forest](#) 모델의 성능을 비교해 본다.
  - [SVM](#): 여러 결정트리를 앙상블해 과적합을 완화하고 안정적인 예측을 제공한다.
  - [Random Forest](#): 초평면으로 데이터 경계를 학습해 고차원에서 분류 성능을 높인다.
- Neural Network 모델을 적용하여 예측 성능을 비교

## ■ 모델 해석 기반 신뢰 확보

- [SHAP](#)을 활용해 모델이 어떤 요인을 근거로 질병 위험을 판단했는지 설명
- 의료진과 환자 모두가 결과를 이해하고 신뢰할 수 있도록 조치

# 기대 효과 및 분석 인사이트

## ■ 기대 효과

- 질병 예측 모델은 의료진이 환자의 위험 요인을 사전에 파악하여 예방적 처치를 제공
- 의료 환경에서는 질병을 놓치는 경우(False Negative)의 비용이 매우 크다.
- 정확도보다 민감도를 우선 고려하는 것이 중요

## ■ 예시: 심장 질환 위험 예측

- Recall이 낮아 고위험 환자를 정상으로 판단
  - 치료 시기를 놓쳐 심각한 결과로 이어질 수 있음
  - 의료 모델 평가에서는 Recall을 핵심 지표로 반영하는 것이 타당

# Mini Project (팀 수행)

## ■ 2명 이상이 팀으로 진행

## ■ 토론 결과는 보고서로 작성하여 제출할 것

- 제출 파일: **.hwp**, **.docx**

## ■ 토론 주제

- 질병 예측 모델이 높은 정확도를 보이지만 Recall이 낮다면 어떤 위험이 발생하는가?
- 의료 모델에서 Precision과 Recall의 균형을 맞추기 위한 전략에는 어떤 것이 있을까?
- (선택) SHAP 기반 설명 가능성이 의료 AI 도입 과정에서 중요한 이유는 무엇인가?



# Medical Imaging Diagnosis

## Healthcare Case Studies (2/2)

# Medical Imaging Diagnosis



# Medical Imaging Diagnosis

## ■ Medical Imaging Diagnosis

- 의료 영상 진단 분야
  - X-ray, CT, MRI 이미지 분석을 통해 질병을 자동으로 판별하는 기술이 빠르게 발전
  - 딥러닝 기반 의료 영상 분석 모델은 진단 속도 향상과 정확도 제고에 기여함
  - 의료진의 의사결정을 보조하는 핵심 기술로 활용





# 핵심 고려 요소 (1/3)

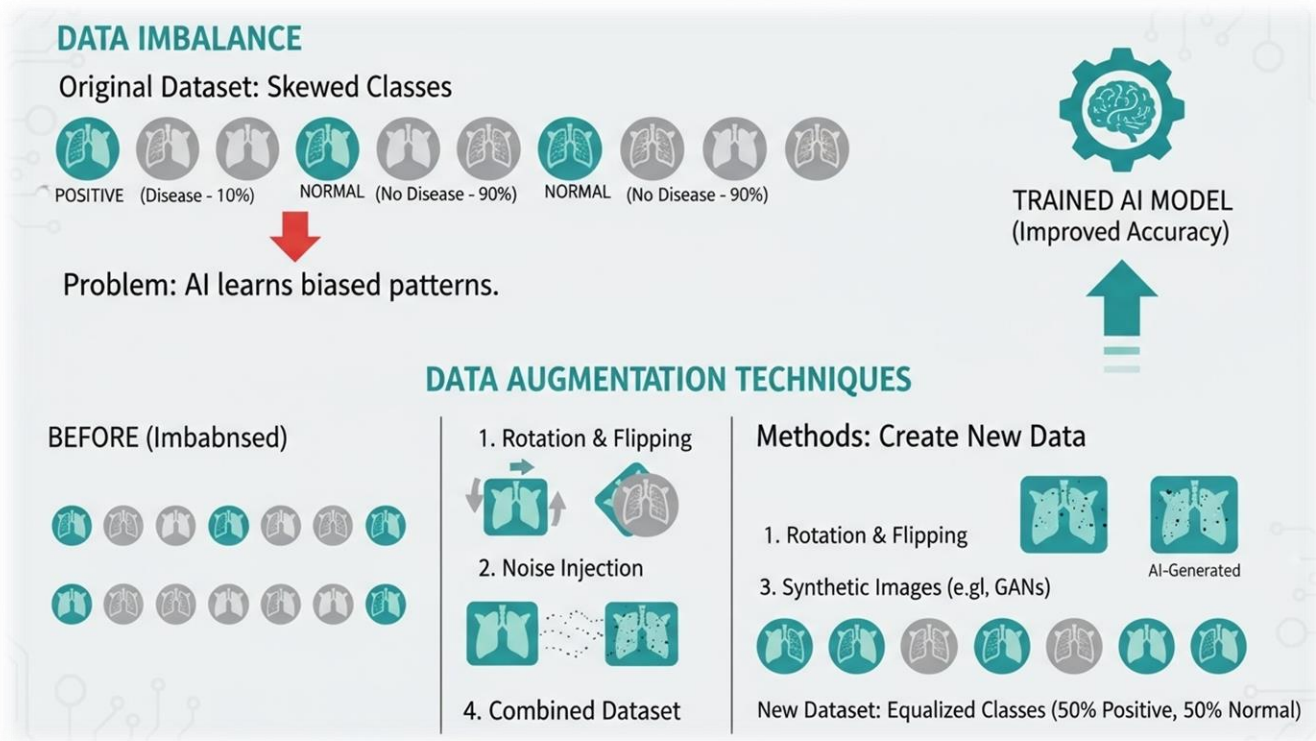
## ■ 개인정보보호

- 의료 데이터 윤리 및 개인정보 보호 의료 영상 데이터는 민감정보에 해당
- [HIPAA](#), [한국 의료법](#) 등 관련 규정 준수
- 데이터 저장, 처리, 공유 과정에서 개인정보 비식별화와 보안 체계가 확보

# 핵심 고려 요소 (2/3)

## ■ 데이터 불균형 해결

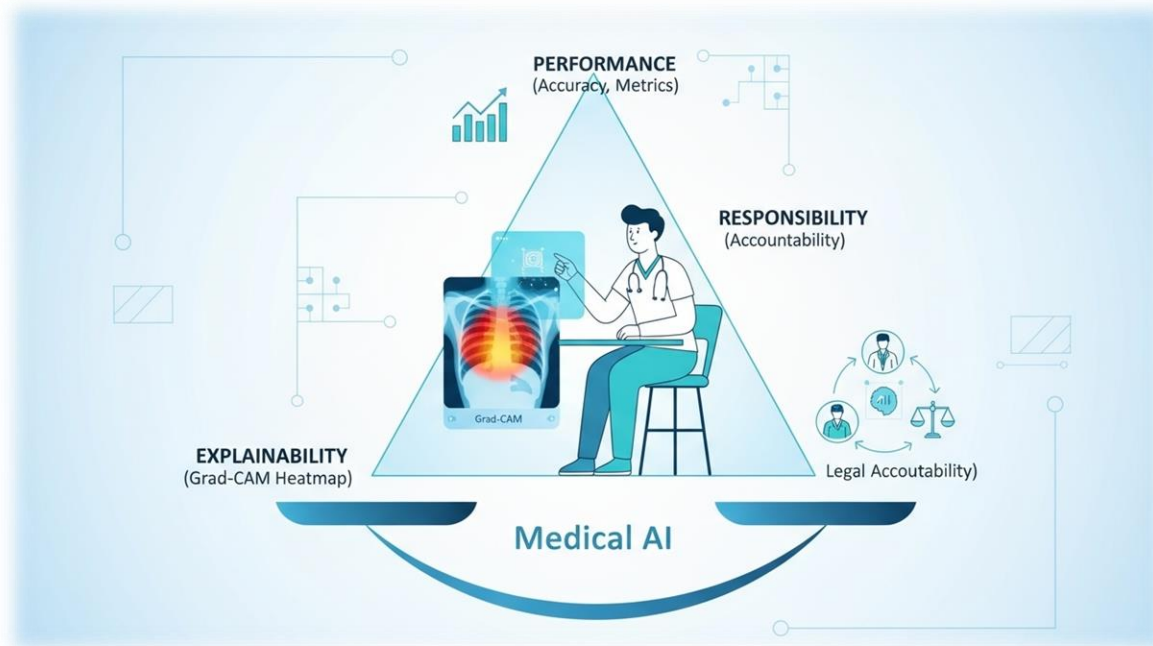
- 질병 보유 환자 수가 적은 경우가 많아 데이터 불균형 문제가 발생
- 데이터 증강, [SMOTE](#), 클래스 가중치 등 다양한 기법을 적용하여 모델의 일반화 성능



# 핵심 고려 요소 (3/3)

## ■ 의료진의 신뢰 확보

- 모델 성능, 해석력, 책임 소재의 균형 모델의 성능이 높더라도
- 진단 근거가 불분명하면 의료진이 결과를 신뢰하기 어렵다.
- 시각화 기법을 활용해 모델의 판단 근거를 제시하고, 오진 발생 시 책임 소재 규정 논의 진행



# 기대 효과 및 분석 인사이트

## ■ 기대 효과

- 의료 영상 진단 모델은 대량의 영상 데이터를 신속하게 분석
- 진단 효율성을 높이고 의료진의 업무 부담을 줄이는 데 기여
- 특히 의료 영상 AI는 전문의 부족 지역에서 원격 진단 지원 기술로 활용
- 신속한 1차 판독 기능을 제공하여 치료 골든 타임 확보에 도움

## ■ 주의 사항

- 다만 의료 AI가 독립적 진단 도구로 사용되기 위해서는 정확도뿐 아니라 설명 가능성과 법·윤리적 책임 영역이 명확히 정립되어야 한다.

# Mini Project (팀 수행)

## ■ 2명 이상이 팀으로 진행

## ■ 토론 결과는 보고서로 작성하여 제출할 것

- 제출 파일: **.hwp**, **.docx**

## ■ 토론 주제

- 의료 영상 진단 모델이 높은 정확도를 보이더라도 해석력이 부족하면 어떤 문제가 발생할까?
- 데이터 불균형이 의료 AI의 편향된 진단 결과로 이어질 수 있는 이유는 무엇인가?
- 오진 발생 시, 책임은 AI 개발자, 의료진, 병원 중 누구에게 있는가?

# Industry Case Studies



# Predictive Maintenance

## Industry Case Studies (1/2)

# Predictive Maintenance



# Predictive Maintenance

## ■ Predictive Maintenance (예지 보전)

- 산업 현장에서 설비의 갑작스러운 고장은 생산 중단, 비용 증가, 납기 지연 등 큰 손실로 이어짐.
- Predictive Maintenance는
  - 센서 데이터를 기반으로 설비의 고장 시점을 미리 예측
  - 사전 점검과 유지보수 계획을 가능하도록 유도
  - 장비 운영 안정성과 비용 절감을 동시에 달성하는 접근 방식

## ■ 문제 정의

- 목표: 설비의 고장 가능성을 조기에 감지하여 계획되지 않은 다운타임을 최소화하는 것
- 고장을 미리 예측할 수 있다면 불필요한 정비를 줄이고, 필요한 시점에 최적의 유지보수 조치 가능



# 데이터 예시 1. 센서 시계열 데이터 (온도, 진동, 전류 등)



# 데이터 예시 2. 설비 사용 이력 및 유지보수 기록

## DASHBOARD OVERVIEW

TOTAL OPERATING HOURS:

18,500

EQUIPMENT HEALTH INDEX

MAINTENANCE EVENTS

12

PREDICTIVE ALERTS

MAINTENANCE (LAST MONTH)

12

PREDICTIVE ALERTS

## MACHINE USAGE & MAINTENANCE HISTORY



## MAINTENANCE LOGS

NOV 01: BEARING REPLACEMENT (UNIT 3)  
4500 HRS

OCT 26: SENSOR CALIBRATION (UNIT 1)  
POST REPAIR

OCT 28: ROUTINE INSPECTIVE  
CHANGING (ALL UNITS)

## UPCOMING EVENTS

NOV 15: AI-DRIVEN PREDICTIVE ANALYTICS RUN

NOV 15: AI-DRIVEN PREDICTIVE ANALYTICS RUN

## PREDICTIVE MAINTENANCE SCHEDULE

MONTH	DATE	TIME	TYPE	STATUS	LOCATION	ASSIGNED	START TIME	END TIME	TO DO	STATUS	LOCATION	ASSIGNED	START TIME	END TIME
1	8	3	8	9	9	10	10	12	13	13	14	13	16	16
15	15	14	15	16	16	10	17	18	20	20	21	22	23	23
22	23	23	24	23	24	30	24	25	26	27	26	27	38	30

# 분석 기법

## ■ 분석 기법

- 수집된 데이터에 시계열 예측 모델 적용
  - [ARIMA](#): 계절성과 추세를 분해하여 선형 시계열 패턴을 예측
  - [LSTM](#): 장·단기 의존성을 학습해 복잡한 비선형 시계열 변화까지 포착

## ■ 기대 효과 및 분석 인사이트

- 갑작스러운 설비 고장을 예방하여 생산 중단을 최소화
- 유지보수 비용을 절감
- 고장 직전의 패턴을 시각화 → 관리자가 이상 징후를 직관적으로 이해할 수 있도록 지원

# Mini Project (팀 수행)

## ■ 2명 이상이 팀으로 진행

## ■ 토론 결과는 보고서로 작성하여 제출할 것

- 제출 파일: **.hwp**, **.docx**

## ■ 토론 주제

- (선택) Remaining Useful Life (RUL) 예측이 가능할 때 유지보수 정책은 어떻게 달라질 수 있는가?
- 시계열 기반 모델과 이상 탐지 기반 모델을 결합하는 이유는 무엇인가?
- 예측 정확도 외에 현장에서 중요한 성과 지표에는 무엇이 있을까?



# Vision-Based Quality Control

## Industry Case Studies (2/2)

# Vision-Based Quality Control



# Vision-Based Quality Control

## ■ Vision-Based Quality Control

- 제조 공정에서는 제품 불량을 빠르게 식별하고 제거하는 것이 품질 관리의 핵심
- Vision-Based Quality Control
  - 카메라와 AI 기반 영상 분석 기술을 활용하여 불량품을 실시간으로 자동 검출하는 시스템
- 검사 정확도를 높이고 사람의 수작업 검사 과정에서 발생하는 누락과 비용 절감 가능

## ■ 문제 정의

- 목표: 생산 공정 중 촬영된 영상 데이터를 분석하여 정상 제품과 불량 제품을 자동으로 구분
- 검사 속도, 정확성, 일관성이 중요하며 실시간 처리 능력이 요구 됨

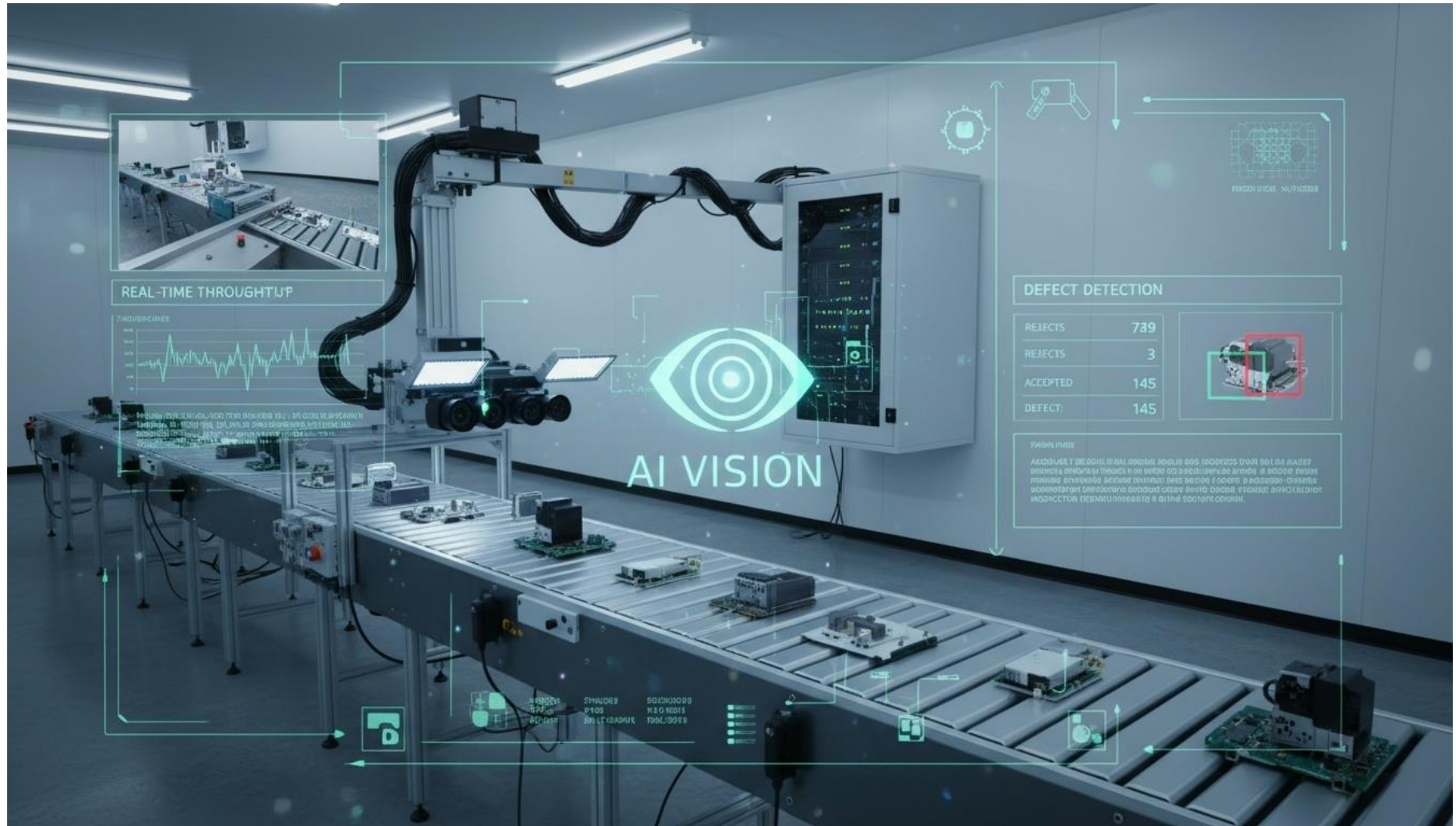
# 데이터 예시 1. 공정 중 촬영된 제품 이미지

## MANUFACTURING QUALITY INSPECTION - PRODUCT IMAGES





# 데이터 예시 2. 생산 라인 영상 데이터





# 데이터 예시 3. 정상 및 불량 샘플 데이터셋

## PRODUCT DEFECT DETECTION FOR AI

### NORMAL SAMPLES



CLASS: APPLE\_GOOD



CLASS: APPLE\_BAD



CLASS: APPLE\_GOOD

### DEFECTIVE SAMPLES



CLASS: APPLE\_BAD



CLASS: APPLE\_BAD



CLASS: APPLE\_BAD

AI MODEL TRAINING



COMPUTER  
VISION SYSTEM



CORRECT CLASSIFICATION



INCORRECT  
CLASSIFICATION

# 분석 기법

## ■ 분석 기법

- 딥러닝 기반 영상 분류 적용
  - [CNN](#)이나 [Vision Transformer\(ViT\)](#)를 활용 → 제품 이미지를 분석하고 정상 여부를 분류

## ■ 기대 효과 및 분석 인사이트

- 검사 자동화를 통해 검사 정확도 향상과 인건비 절감 효과를 동시에 얻을 수 있다.
- 일관된 품질 관리를 통해 불량률을 낮추고 고객 만족도를 높일 수 있다.
- 성과 측정 지표는
  - 불량 검출률 개선,
  - 검사 비용 절감률,
  - 자동화 도입에 따른 ROI(Return on Investment) 등

# Mini Project (팀 수행)

## ■ 2명 이상이 팀으로 진행

## ■ 토론 결과는 보고서로 작성하여 제출할 것

- 제출 파일: **.hwp**, **.docx**

## ■ 토론 주제

- 영상 기반 검사 자동화가 품질 관리 프로세스에 미치는 가장 큰 변화는 무엇인가?
- (선택) CNN과 ViT 중 제조 현장에서 어떤 모델이 더 적합하다고 판단할 수 있는가?
- 데이터 기반 이상 탐지가 사람의 수작업 검사와 비교해 가지는 강점은 무엇인가?

# Mini Capstone

# 캡스톤(Capstone)이란 무엇인가?

## ■ 캡스톤(Capstone)?

- 본래 건축에서 건물의 맨 위를 완성하는 마감석을 뜻함
- 대학 교육
  - 개념을 차용해, 학생들이 그동안 배운 지식과 기술을 하나의 완성된 프로젝트 형태로 통합하는 학습 활동을 의미
  - 캡스톤 프로젝트는 학기나 전공 과정의 마지막 단계에서 수행하는 종합 실무형 과제

## ■ 실제 운영은?

- 실제 데이터를 다루고 문제를 정의하며 결과를 스스로 도출하는 창의적 문제 해결 경험을 중심
- 일부 대학에서는 졸업을 위한 최고 학년의 필수 과목
  - 종합설계 또는 졸업 프로젝트 등과 같은 과목으로 운영

# 대학생이 캡스톤 프로젝트를 수행하는 이유

이유	설명
이론의 실제 적용	교과서에서 배운 분석 기법과 모델링 방법을 실제 데이터에 적용해본다.
창의적 문제 해결 경험	스스로 주제를 선택하고 방법을 설계하며 문제를 해결하는 과정을 통해 논리적 사고력을 기른다.
팀워크와 협업 능력 향상	여러 전공과 역할이 함께 어우러진 팀 단위 프로젝트를 통해 실무형 협업 능력을 배운다.
포트폴리오 구축	프로젝트 결과물을 향후 취업, 대학원 진학, 연구 활동의 대표 사례로 활용할 수 있다.
비판적 사고와 자기 주도성 강화	단순한 과제 수행이 아닌, 스스로 계획·분석·평가하는 자기 주도적 학습 과정이 된다.

# 캡스톤 프로젝트의 성과 목표

목표 영역	구체적 성과
기술적 성취	데이터 수집·분석·모델링 능력 향상
논리적 사고력	문제를 정의하고, 가설을 세우며, 결과를 해석하는 과정의 체계화
창의성	기존 방법을 개선하거나 새로운 아이디어를 도입
협업 및 커뮤니케이션	역할 분담, 회의, 발표 등 협업 과정 경험
실무 감각	실제 현업 데이터 분석 흐름을 체험하고 보고서·대시보드 형태로 전달

# 데이터사이언스 과목에서 캡스톤을 해야 하는 이유

## ■ 데이터사이언스는

- 단순히 코드를 배우는 학문이 아니라, 데이터를 통해 **세상의 문제를 정의하고 해결**하는 학문
- 이론과 실습만으로는 한계가 있으며, **실제 프로젝트 수행을 통해서만 진정한 실력**을 체득

통합적 학습	데이터 전처리, 시각화, 모델링, 해석 등 과목 전체 내용을 하나로 연결
실제 문제 해결 경험	실제 사회·산업 데이터를 분석하며 현장감 있는 학습 경험
성과 기반 평가	단순 시험 점수가 아닌, 분석 과정과 결과의 완성도로 성과 평가 가능
윤리·공정성 학습의 전 단계	14주차의 '데이터 윤리와 편향' 토론으로 이어지는 준비 과정 역할 수행



# 미니 캡스톤 안내

본인이

“데이터로 문제를 정의하고, 분석하고,  
스토리로 전달할 수 있는가”를

직접 증명 해보세요!

[https://www.deepshark.org/courses/  
data\\_science/w/13\\_case\\_studies#mini-  
capstone](https://www.deepshark.org/courses/data_science/w/13_case_studies#mini-capstone)



수고하셨습니다 ..^^..