신용카드 사기 거래 탐지 AI 경진 대회 PBL 기획서

기간 : 23/01/16 ~ 23/02/24

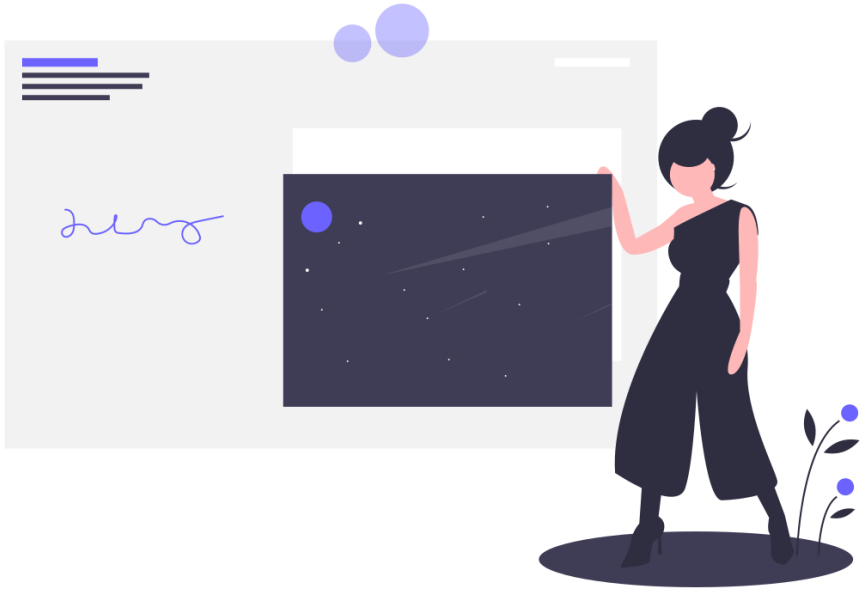
목차

1 PBL **주제**

2 Stage **별 기획**

3 **팀 멤버 및 역할**

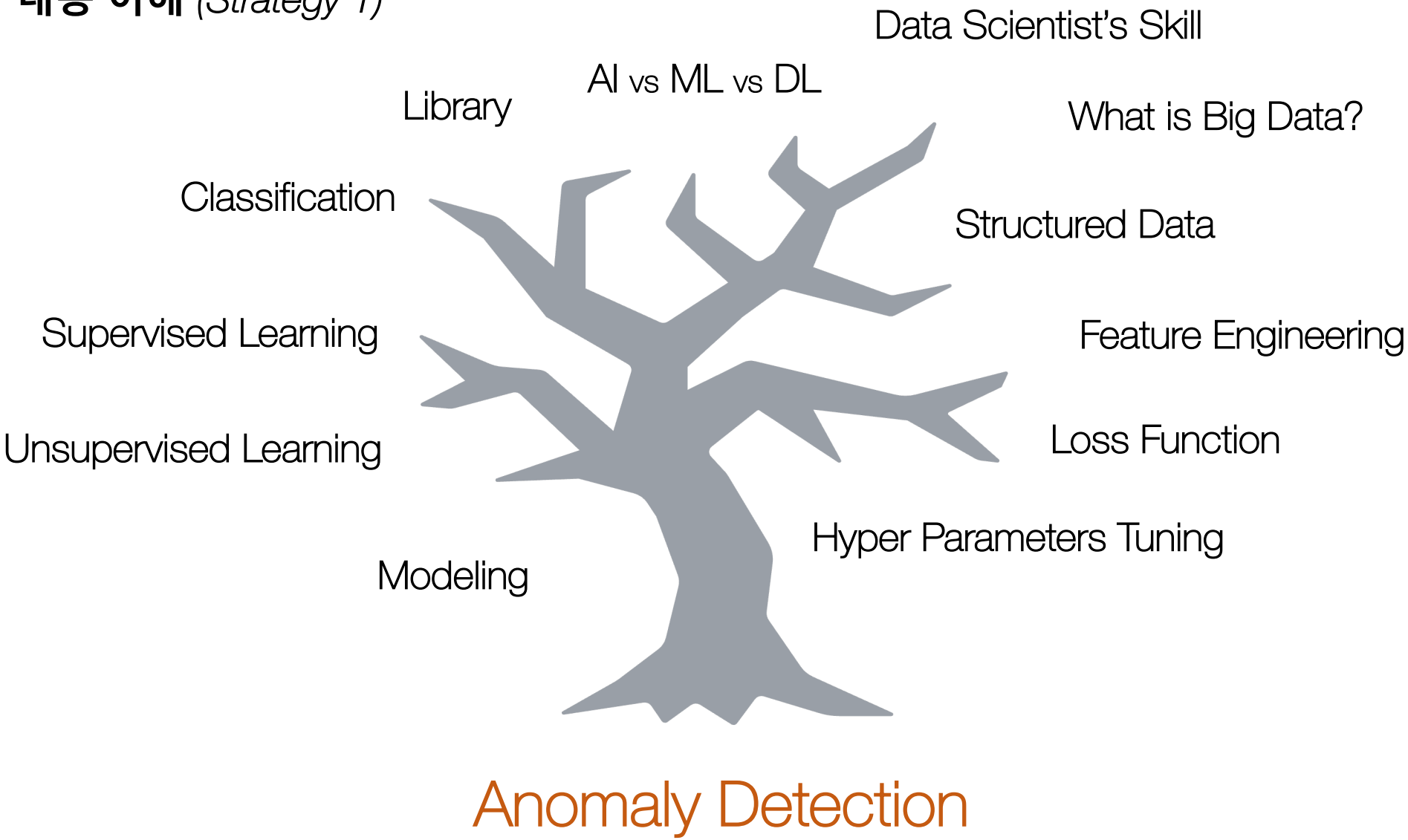
1. PBL 주제



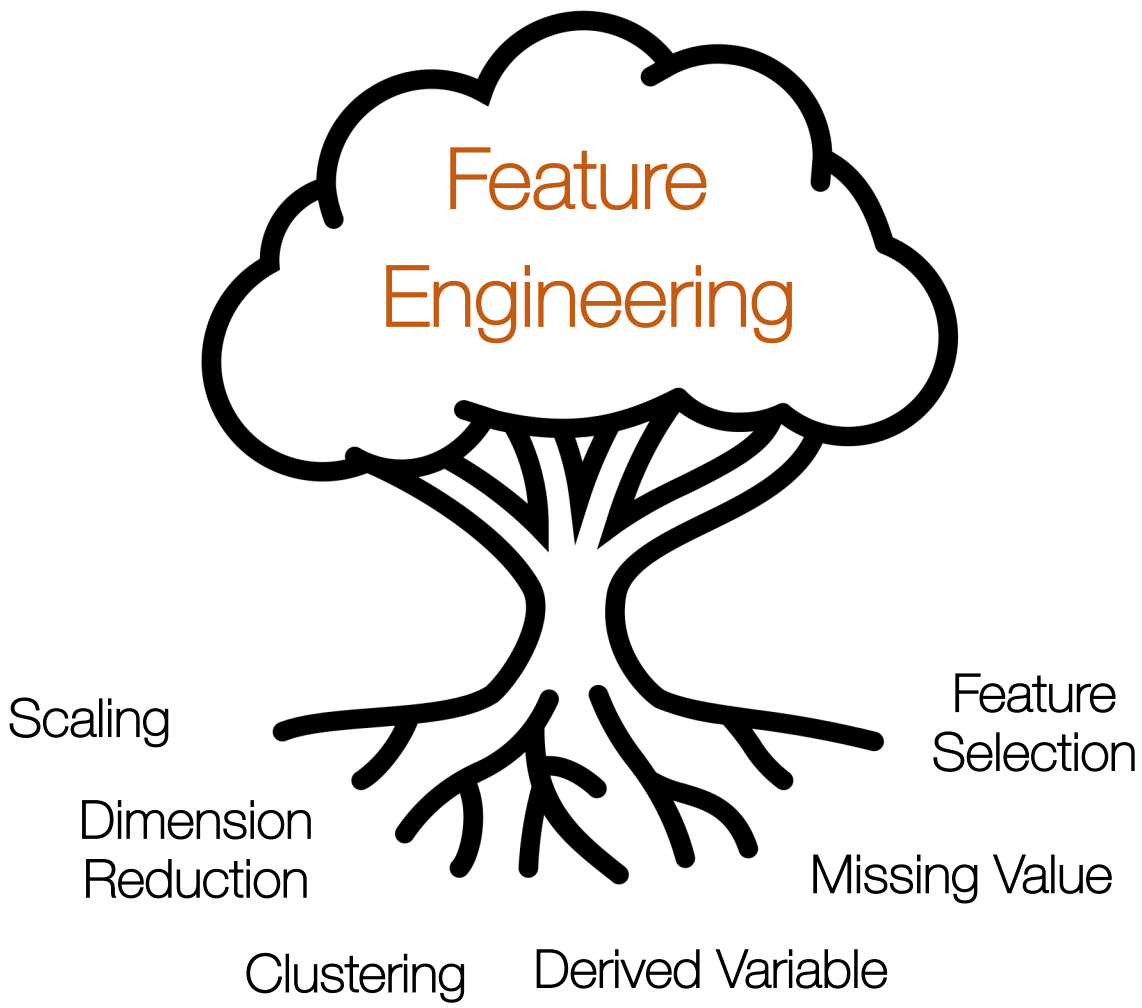
주제 *(Mission)* : Anomaly Detection**에서 머신러닝 엔지니어링 기법 정확히 이해하기**

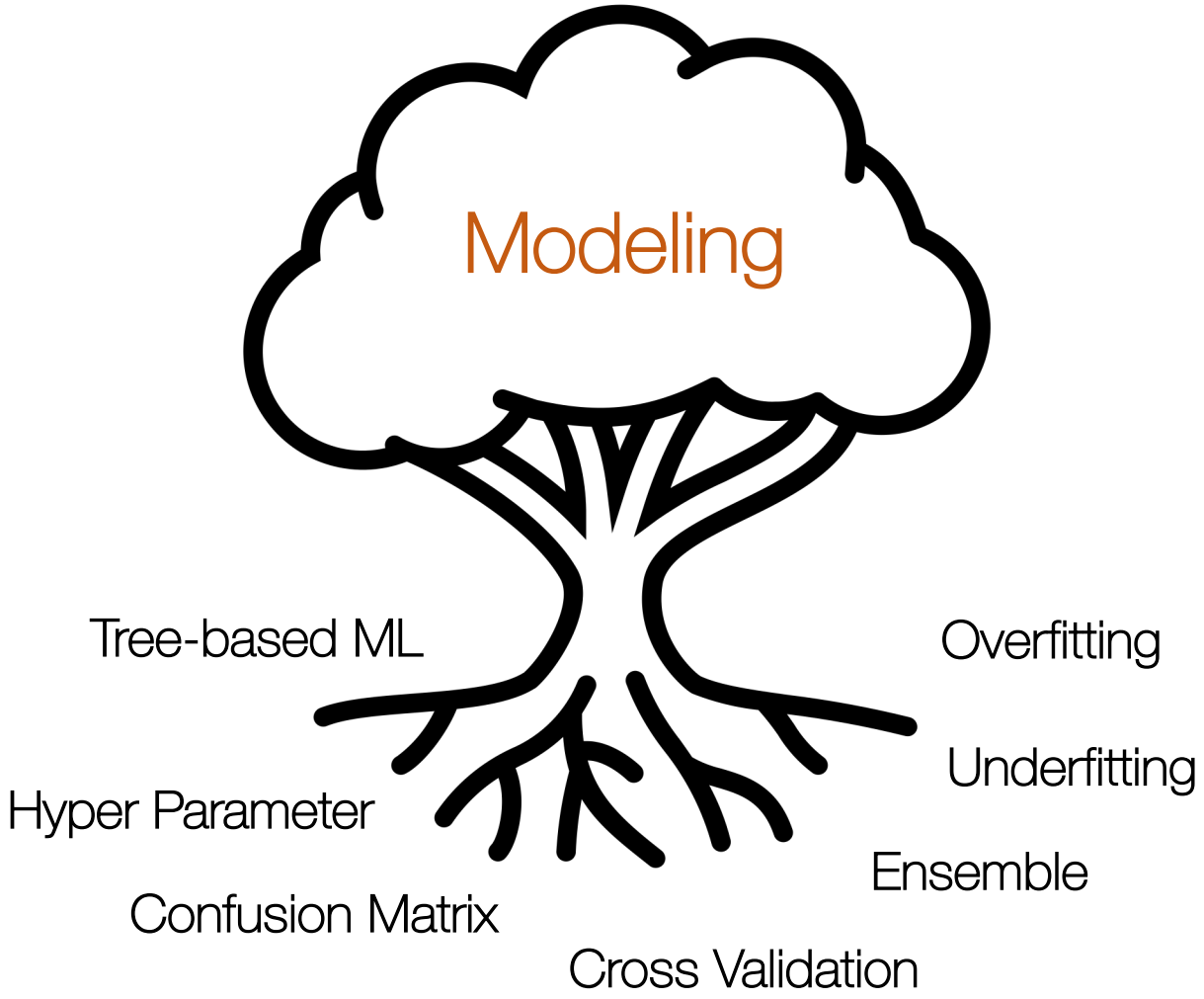
부제 *(Vision)* : PBL**만 열심히 하면 다른 곳에서도 좋은 성적을 낼 수 있다**

전반적인 내용 이해 *(Strategy 1)*



세부적인 내용 이해 *(Strategy 2)*





퀴즈 풀기 *(Strategy 3)*

Example Quiz

1. **해당 데이터로** Box Plot**를 그려보았습니다. 다수의 이상치가 보입니다. 해당 데이터의 평균값, 중앙값, 표준편차를 구해보시오. 그리고 다음의 어떤** Scaling **기법을 적용하면 좋을 지 생각해보시오.**

   Hint : **이상치는 정규화 과정에서 평균값과 표준편차의 값에 큰 영향을 줍니다**

2. **사기거래 식별 문제를 머신러닝으로 해결하고자 합니다. 머신러닝의 성능 지표로 ' Accuracy'와** 'F1 Score' **중 어떤 것을 선택하면 좋을 지 생각해보시오.**

   Hint : **사기거래가 발생할 가능성은 극히 적습니다.**

3. Train Score**와** Val Score**의 차이가 커 보입니다. 이들 간의 격차를 줄이기 위해서는** Random Forest**의 무슨** Hyper Parameter**를 어떻게 설정하면 좋을 지 생각해보시오?**

   Hint : 'max_depth'**는 뿌리의 깊이,** 'min_samples_leaf'**는 마지막 뿌리가 되기 위한 최소한 데이터 수,** 'min_samples_split'**는 가지를 나누기 위한 최소한의 데이터 수와 관련이 있습니다.'**

2. Stage 별 기획

문제 기반 실무 학습의 이해

학습 목표

- 데이터 사이언티스트의 역할과 역량을 이해한다
- 머신러닝 엔지니어링은 무엇이고 프로세스는 어떤 지 이해한다
- AI & ML & DL가 무엇이고 차이점을 이해한다
- 지도 학습과 비지도 학습을 이해한다
- 빅데이터가 무엇인지 이해한다
- 데이터의 종류를 구분하고 이해한다
- Anomaly Detection이 무엇인지 이해한다

참여 실습

- 파이썬 라이브러리가 무엇이고 불러올 수 있다
- CSV 파일을 Data Frame으로 불러온다
- 데이터의 Row와 Column을 이해하고 개수를 셀 수 있다
- 데이터의 통계값을 확인할 수 있다
- 퀴즈 : 각각 Column의 Type을 확인할 수 있다

데이터 사이언티스트가 되기 위해 무엇이 중요한지 이해한다

문제 기반 실무 학습의 기초

학습 목표

- 데이터 전처리가 왜 필요한 지 이해한다
- 결측치가 무엇이고 어떻게 처리하는 지 이해한다
- 파생변수, 스케이링, 클러스터링, 차원 축소 등 다양한 기법을 이해한다
- 언더 샘플링과 오버 샘플링을 이해한다
- 변수 선택법을 이해한다

참여 실습

- IQR을 기준으로 이상치를 이해한다
- 파생변수를 생성해본다
- 스케이링을 적용한다
- 클러스터링을 적용한다
- 테이블 핸들링을 통한 변수 선택을 해본다
- 퀴즈 : 여러 조건에 부합하는 데이터만 추출해본다

데이터를 내가 원하는 데로 자유롭게 핸들링 할 수 있다

문제 및 모델링의 이해

학습 목표

- 지도 학습과 비지도 학습의 문제 접근 방식을 이해한다
- 다양한 머신러닝 기법을 이해한다
- 딥러닝은 무엇인지 이해한다
- 각 모델의 하이퍼 파라미터를 이해한다
- 성능을 높일 수 있는 기법들을 이해한다 (앙상블, 교차검증)
- 손실 함수를 이해한다 (혼동 행렬)
- 과소 적합과 과대 적합을 이해한다

참여 실습

- 데이터를 학습 데이터와 검증 데이터로 분할한다
- 머신러닝을 수행해본다 (트리 기반 머신러닝)
- 머신러닝을 하이퍼 파라미터를 변경해본다
- 혼동 행렬 지표를 구해본다
- 성능 지표를 비교 및 확인한다
- 과소 적합과 과대 적합 문제를 겪어본다
- 퀴즈 : 성능 개선 방안을 찾아본다

모델의 작동 방식을 이해하고 성능 향상시키는 방법을 이해한다

기초 EDA 및 모델링의 활용

학습 목표

- 데이터 타입별로 어떤 그래프를 그릴 수 있는 지 이해한다
- 각각의 그래프가 줄 수 있는 의미를 이해한다
- 그래프를 통해 데이터의 분포와 통계값을 추측해본다
- 성능 지표를 그래프로 표현한다

참여 실습

- 데이터 타입에 맞는 그래프를 그릴 수 있다
- 그래프를 자기가 원하는데로 수정할 수 있다 (X & Y축, 크기 등등)
- 그래프를 'PNG' 파일로 저장할 수 있다
- 하나의 창에 여러 그래프를 동시에 나타낼 수 있다
- Target의 분포를 확인할 수 있다
- 데이터가 불균형 상태인지 확인할 수 있다
- 퀴즈 : 시각화를 통해 데이터 분포를 확인하고 그에 맞는 전처리 기법을 선택한다

시각화를 자유롭게 다루어 Insight를 찾을 수 있는 방안을 모색할 수 있다

데이터 가공 및 EDA 기반 Insight 시각화의 이해와 활용

학습 목표

- 전처리로 인한 데이터 변화를 시각화로 확인한다
- 히스토그램과 박스 플랏으로 데이터 분포와 이상치 여부를 확인한다
- 산점도를 통해 상관관계를 확인한다
- 전처리를 통해 데이터 분포가 어떻게 변화하였는 지 시각화로 확인한다

참여 실습

- 박스 플랏을 통해 이상치를 확인하고 처리할 수 있다
- 산점도를 통해 파생변수를 생성할 수 있다
- 전처리된 데이터를 시각화해보고 모델의 비교한다
- 퀴즈 : 모델링 수정에 따른 성능 지표 결과 전 후를 시각화로 비교한다

Insight를 통해 모델링을 수정하여 성능을 향상시킬 수 있다

고급 데이터 사이언스 기법의 이해와 활용

Insight 도출 및 증명의 이해와 활용

학습 목표

- 상위권 코드를 이해하고 구현할 수 있다
- 가설 검증에 대해 이해한다
- 시각화를 바탕으로 한 가설을 검증할 수 있다
- A/B 테스트를 수행할 수 있다
- 학습한 내용을 바탕으로 사기거래 식별 모델링 성능을 올려본다

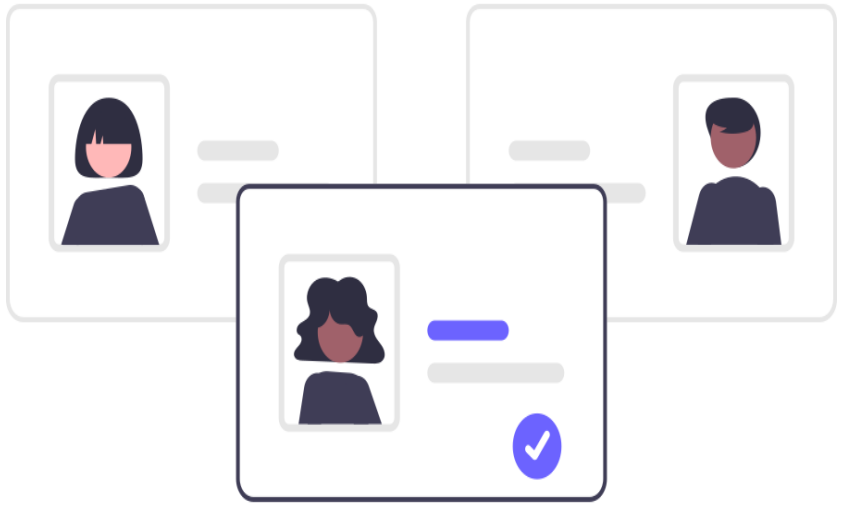
참여 실습

- 상위권에서 사용한 기법 중 사용할 수 있는 것을 구현해본다
- 가설을 설정하고 검증해본다
- 모델링을 수행하고 문제점을 진단하고 개선방안을 찾는다
- 퀴즈 : 최종 결과물에 대해 논리적으로 설명해본다

수행한 머신러닝 엔지니어링에 대해서 논리적으로 설명할 수 있다

'무엇'을 '어떻게'하여 '왜' 수행하였는지

3. 팀 멤버 및 역할



팀 멤버 및 역할

권남우

- **팀장**
- PBL **총괄**
- PDF **제작**
- **코드 작성** (IPYNB)
- **으쌰으쌰 다같이 공부**

빵일이삼

- **팀원**
- **자료 조사**
- **코드 작성** (IPYNB)
- **으쌰으쌰 다같이 공부**

감사합니다

일러스트 : unDraw (https://undraw.co/illustrations)