데이터로 인사이트를 찾는
데이터 분석가 허주혁입니다.

지원자 허주혁

Contact

Github
https://github.com/hjuhyeok

Phone
010-2476-5021

Email
wngur1205@naver.com

긍정적인 연구원 허주혁입니다

허주혁 / Juhyeok Heo

1998.12.05 / 경기도 오산시

Tel. 010-2476-5021

Email. wngur1205@naver.com

SNS. https://github.com/hjuhyeok

GRADUATION

2017.02 운천고등학교 졸업

2023.08 충북대학교 정보통계학과 졸업

GPA : 3.98 / 4.5

2025.08 연세대학교 통계데이터사이언스학과 졸업(예정)

GPA : 4.19 / 4.5

PROGRAM SKILLS

PYTHON	95%
R	95%
TABLEAU	90%
QGIS	90%
SQL	85%

Experience

2022 이노포스트

- 보행안전지수 개발 및 시각화

2023 연세대학교 데이터사이언스연구소

- 석/박사 대상 통계 상담 및 의사결정 지원

2024 연세대학교 바른ICT연구소

- 악성댓글 비용 시뮬레이터 개발

PROJECT

2022 천안시 교통사고 취약지역 도출

2023 맞춤형 추천을 제공하는 개인화 추천 시스템 모델 구축

2024 Tabular Data Generation Using Generative Models

CERTIFICAITON

2020 워드프로세서

2020 사회 조사 분석사 2급

2021 데이터 분석 준전문가(ADsP)

2022 컴퓨터 활용능력 1급

2022 품질경영기사

2022 SQL 개발자(SQLD)

2022 빅데이터 분석기사

AWARDS

2022 천안시 데이터 기반 시각화 아이디어 공모전 수상

PAPER

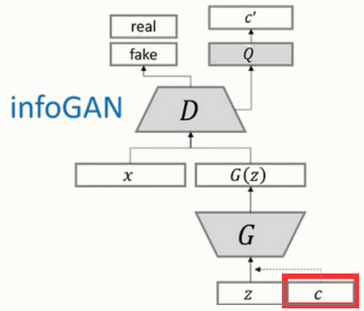
ADBoost: Tree-Based Boosting with Oversampling and Undersampling for Imbalanced Data

PROJECTS 1

Tabular Data Generation Using Generative Models

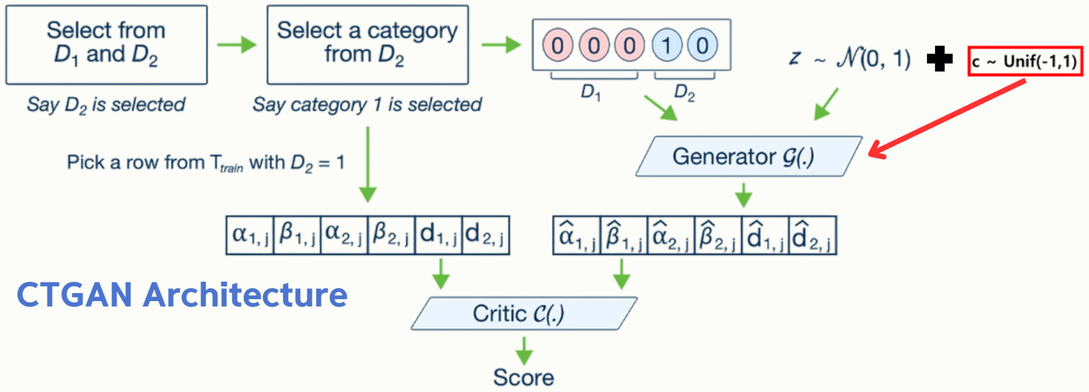
프로젝트 개요

CTGAN과 InfoGAN을 결합한 Tabular Data 생성 모델 설계 및 구현
데이터 품질 평가 및 비교를 통한 성능 개선



InfoGAN Architecture

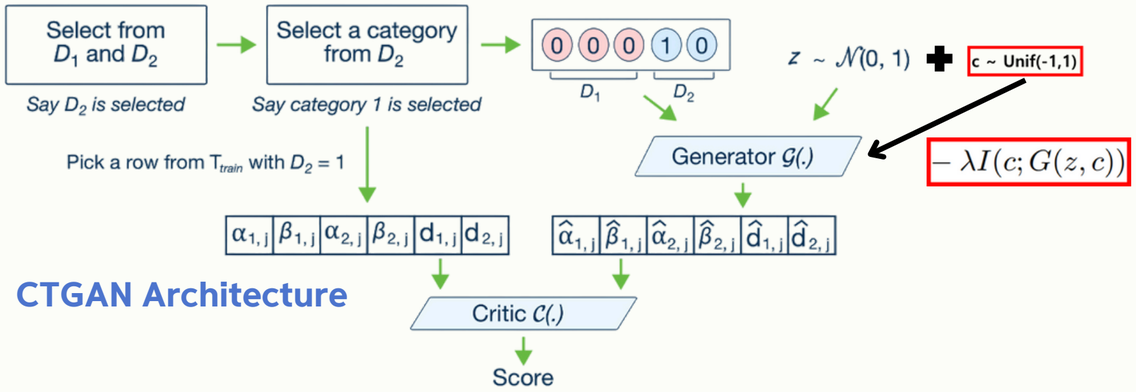
Step 1.



(1) Conditional vector, noise 외에 잠재 변수인 c를 input에 추가

⟶ 가정 : InfoGAN에서 c가 데이터의 의미적 특징을 학습하는 것처럼,
CTGAN에서도 c가 유의미한 특징을 학습할 것이라 가정하고 적용

Step 2.



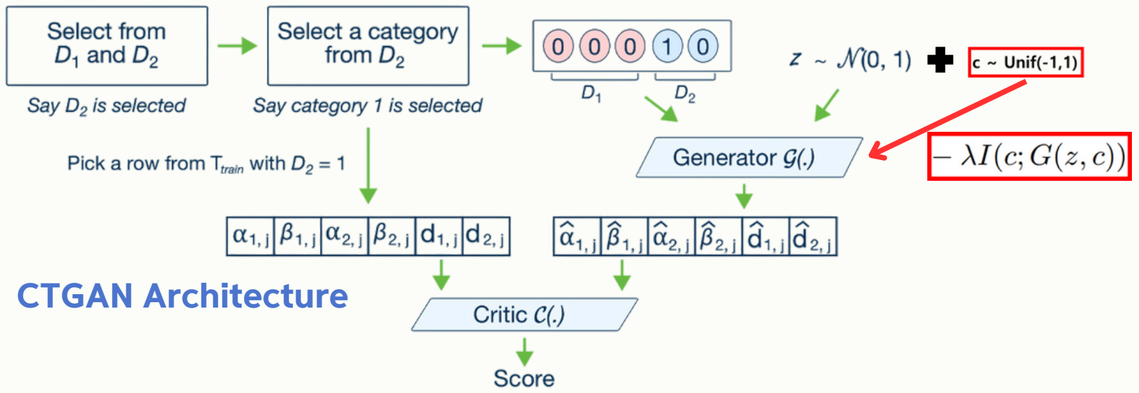
(2) Mutual information 항을 목적 함수에 추가

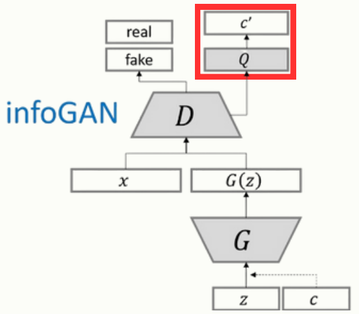
⟶ 생성된 데이터가 잠재 변수 c의 의미적 특징을 학습하도록 유도

PROJECTS 1

Tabular Data Generation Using Generative Models

Step 3.





InfoGAN Architecture

Mutual Information

$$
\begin{aligned}
I(c; G(z,c)) &= H(c) - H(c|G(z,c)) \\
&= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\
&= \mathbb{E}_{x\sim G(z,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot|x) \parallel Q(\cdot|x))}_{\geq 0} + \mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\
&\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c)
\end{aligned}
$$

Mutual Information의 직접 계산이 어려워 Variational Inference 적용

(3) Posterior distribution의 approximation을 위해 Auxiliary distribution $Q(c|x)$ 를 사용

$Q(c|x) = \mathcal{N}(\mu(x), \sigma^2(x))$로 가정하여 Mutual Information 최적화

결과 및 기대 효과

데이터 부족 문제 해결 및 AI 모델 성능 향상
제조, 금융, 의료 등 다양한 산업에 적용 가능

사용도구

python

기여도 ★★★★★

PROJECTS 2

보행안전지수 개발 및 시각화

프로젝트 개요

보행안전지수를 개발 및 시각화하여 보행약자 보호 정책 수립의 근거 마련
교통사고, 유동인구, 도로시설물 데이터를 활용해 위험 지역을 분석 및 지수화

진행 과정 및 역할

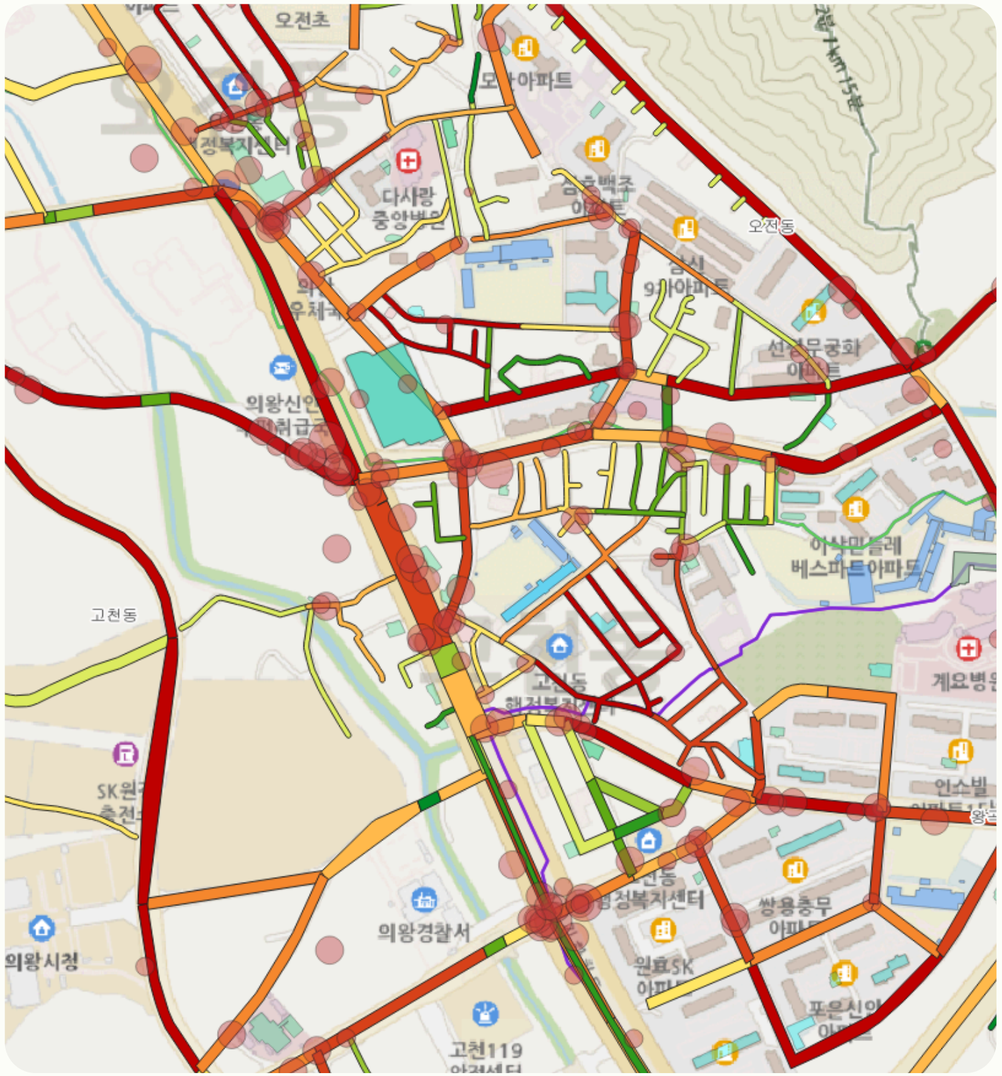
데이터 전처리: 보행 중 교통사고, 유동인구 데이터 정리 및 오류 수정
지수 산출: 변수 가중치 적용 후 보행안전지수 계산
시각화: QGIS 기반 공간 분석 및 지도 제작

결과 및 기대 효과

보행안전 취약 지역을 시각화(적색일수록 보행 안전이 취약한 도로를 의미함)
보행약자 보호 정책 수립을 위한 기초 자료로 활용 가능

사용도구

python



기여도 ★★★★☆

PROJECTS 3

천안시 교통사고 취약지역 도출

프로젝트 개요

천안시 교통사고 취약지역을 분석하여 우선적으로 해결해야 할 지역 도출
교통사고 데이터와 행정동 정보를 활용해 위험 지역을 군집화하고 시각화

진행 과정 및 역할

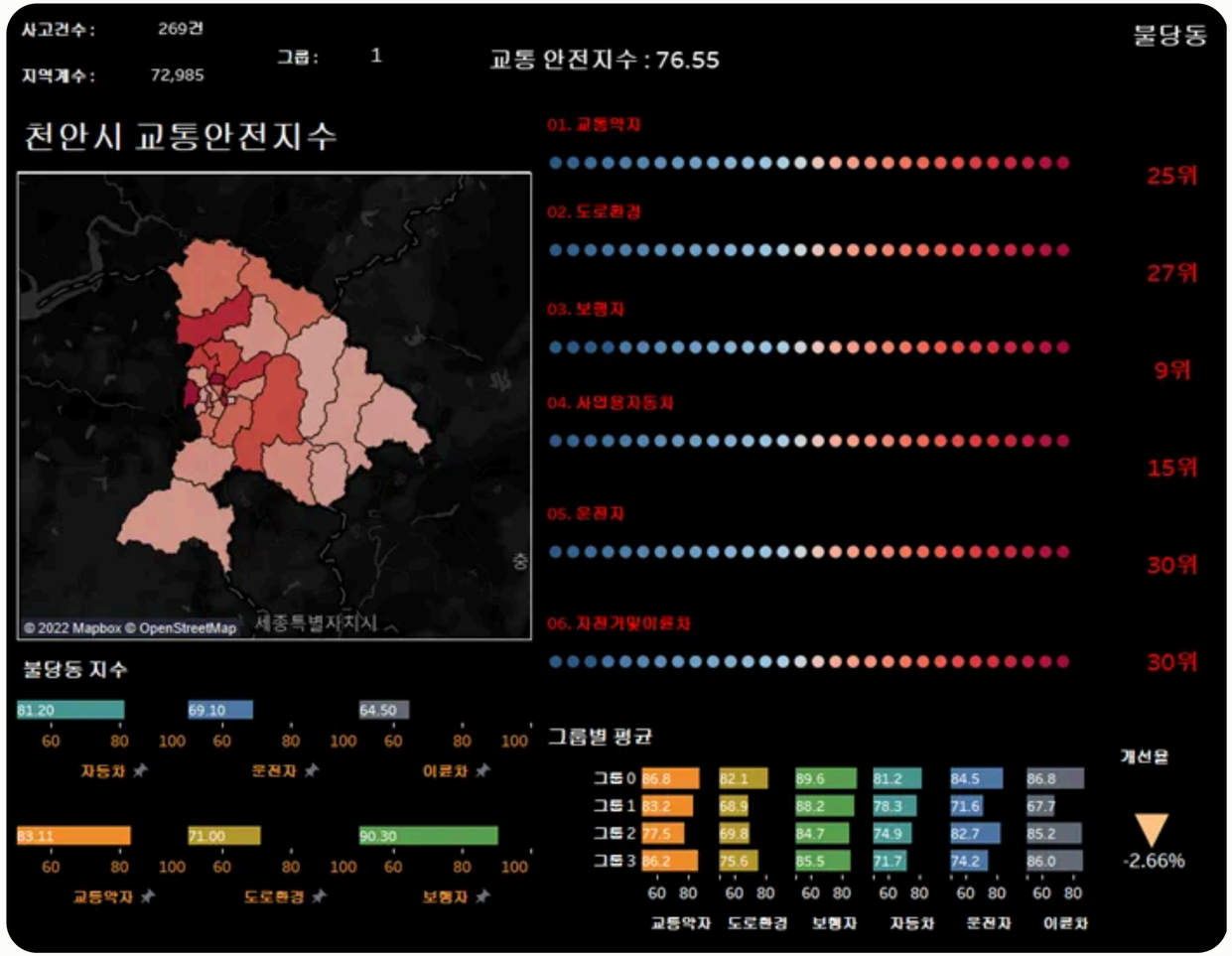
QGIS 활용: 천안시 행정동 구분 및 교통안전지수 산출 (6개 변수 적용)
데이터 분석: PCA 기반 차원 축소 및 군집 분석 수행 (4개 군집 분류)
시각화 및 대시보드 제작: Tableau를 활용한 취약 요소 분석 및 개선 방안 제안

결과 및 기대 효과

교통사고 위험 지역을 시각적으로 제시
행정적 개선 방향을 제안하였으며, 우수상 수상

사용도구

python QGIS tableau



기여도 ★★★★★

PAPER

ADBoost: Tree-Based Boosting with Oversampling and Undersampling for Imbalanced Data

논문 개요

1. 클래스 불균형 문제 해결을 위한 트리 기반 부스팅 프레임워크 ADBoost 제안
2. ARF를 활용해 소수 클래스 데이터를 생성하고, 원본 데이터와 결합하여 오버샘플링 수행
3. DRF를 이용해 다수 클래스 데이터만 포함된 리프 노드를 식별하고, 해당 데이터 일부 제거하여 언더샘플링 적용
4. 소수 클래스 오분류 데이터에 더 높은 가중치를 부여해 예측 성능 향상 유도

Adversarial Random Forest (ARF): 랜덤 포레스트의 학습 구조를 활용해 적대적 학습 방식으로 소수 클래스 데이터를 생성하는 기법

Double Random Forest (DRF): 이중 무작위성을 적용하여 데이터 샘플링과 특징 선택을 모두 랜덤화함으로써 일반화 성능을 향상시키는 기법.

현재 상태 6월 중 논문 투고 예정

기여도

연구 70%

논문 60%

사용도구

Algorithm

Algorithm 1 ADBoost Algorithm

Require: $\mathcal{S}_{\text{original}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, h = Base Classifier, T = number of iterations in boosting procedure, IR = Imbalance Ratio

Initialize the weight vector $D_1 = \{D_1(1), \ldots, D_1(n)\}$ such that $D_1(i) = \frac{1}{n}$ for $i = 1, \ldots, n$.
for $t = 1, \ldots, T$ **do**
  Use ARF to generate additional data for the minority class.
  Combine oversampled and original data, then perform DRF for undersampling to form $\mathcal{S}_{\text{temporary}}$.
  Train a Base Classifier $h_t : x \to y$ using $\mathcal{S}_{\text{temporary}}$.
  Compute $\epsilon_t = \sum_{i=1}^{n} D_t(i)\mathbb{I}(h_t(x_i) \neq y_i)$, where $\mathbb{I}$ is an indicator function and $(x_i, y_i) \in \mathcal{S}_{\text{original}}$.
  Compute $\beta_t = \alpha \log \frac{1-\epsilon_t}{\epsilon_t}$.
  Define imbalance adjustment function:

$$f(\text{IR}) = 2 - \frac{1}{\sqrt{\text{IR}}}$$

  Update weights:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \begin{cases} \exp(\beta_t) \cdot f(\text{IR}), & \text{if } h_t(x_i) \neq y_i \text{ and } y_i \text{ is minority class,} \\ \exp(\beta_t), & \text{if } h_t(x_i) \neq y_i \text{ and } y_i \text{ is majority class,} \\ 1, & \text{if } h_t(x_i) = y_i, \end{cases}$$

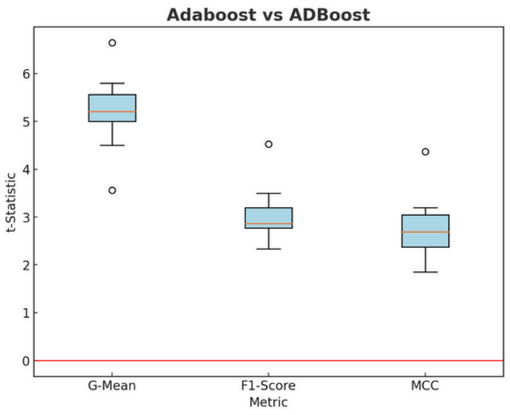
  where Z_t is the normalizing constant such that $\sum_{i=1}^{n} D_{t+1}(i) = 1$.
end for
Output: The final classifier:

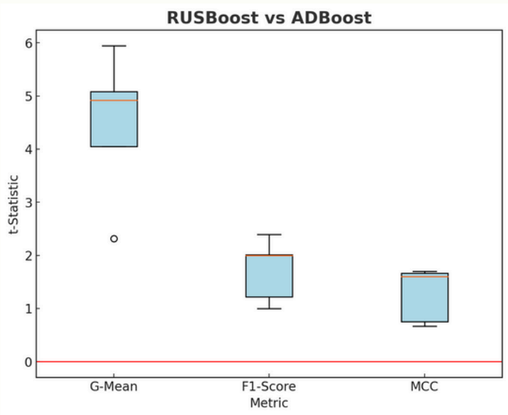
$$H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} \beta_t \mathbb{I}(h_t(x) = y).$$

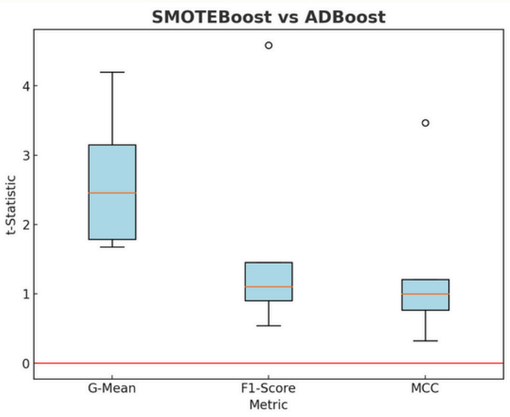
PAPER

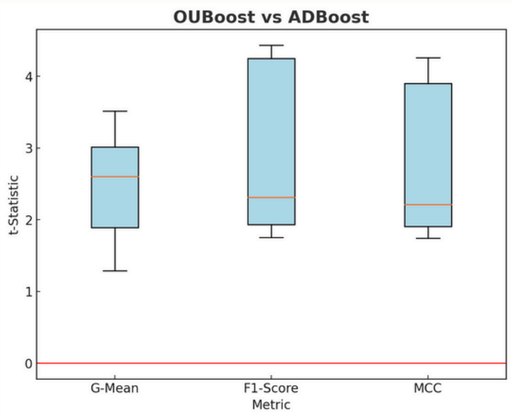
ADBoost: Tree-Based Boosting with Oversampling and Undersampling for Imbalanced Data

Experimental Results









다른 Boost 모델과 가지는 차별점

기존 부스팅 모델(SMOTEBoost, OUBoost 등)은 클래스 불균형을 해결하기 위해 단순한 오버샘플링 또는 언더샘플링 기법을 적용하지만, 선형적인 데이터 생성(SMOTE)이나 무작위 샘플 제거의 한계를 가진다.

ADBoost는 트리 기반 샘플링을 활용하여 ARF로 소수 클래스 데이터를 생성하고, DRF로 다수 클래스 데이터를 제거하는 방식으로 불균형을 조정한다. 이를 통해 기존 방식보다 **비선형적 데이터 분포를 효과적으로 반영**할 수 있다.

또한, 기존 부스팅 모델들이 오분류 데이터에 균일한 가중치를 부여하는 것과 달리, ADBoost는 소수 클래스 오분류 데이터에 더 높은 가중치를 부여하여 소수 클래스의 예측 성능을 강화한다.

실험 결과, ADBoost는 기존 부스팅 모델 대비 소수 클래스 예측 성능을 향상시키고, 보다 균형 잡힌 분류를 수행하는 효과적인 기법으로 확인되었다.

감사합니다.
잘 부탁드립니다!

지원자 허주혁

Contact

Github
https://github.com/hjuhyeok

Phone
010-2476-5021

Email
wngur1205@naver.com