2024년 영주시 데이터분석 활용 공모전

영주시 빈집활용 최적의 청년시설 선정

목차

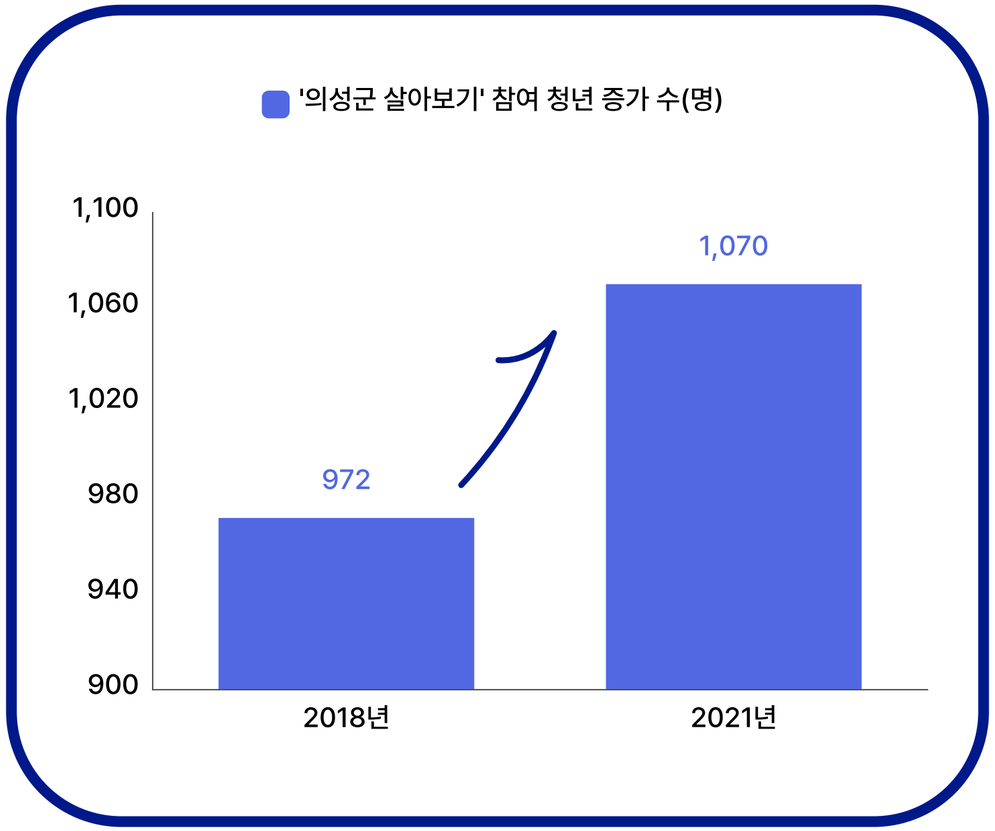
연구요약 & 목적

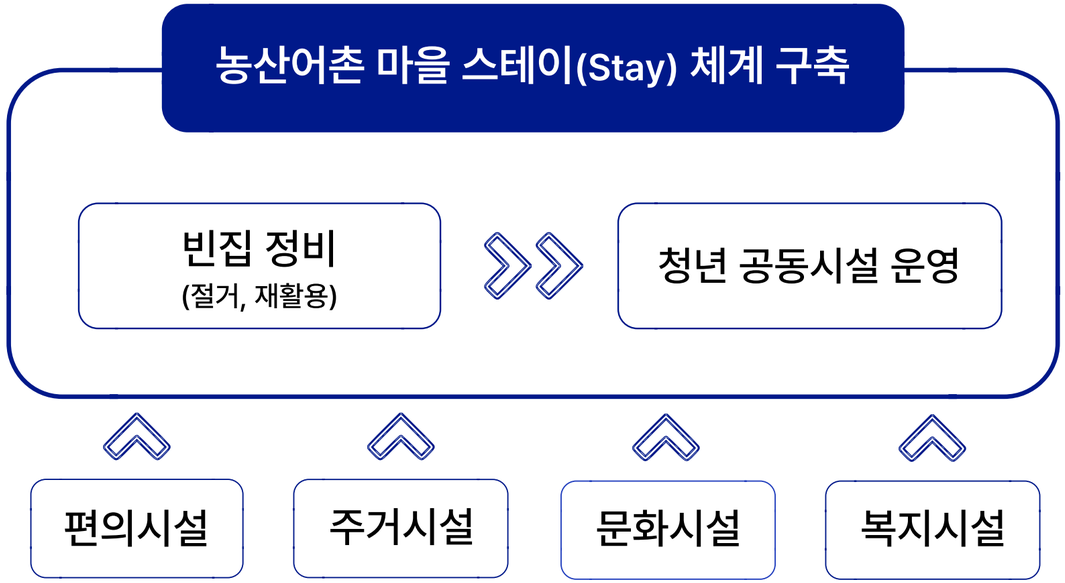
2011년부터 2020년까지 영주시 청년 인구가 41.4%까지 감소한 상황이다. 2024년을 시작으로 사회와 환경적으로 문제가 되고 있는 빈집을 매입해 주민 공유시설로 조성하는 빈집 정비활용 사업을 추진하고 있다. 이에 따라, 본 연구에서 청년 소멸지역으로 손꼽히는 영주시에 어떠한 청년시설이 어느 지역에 들어가야 청년 인구 증가에 효과적인지 알아보았다. 청년시설이란 청년의 활동을 지원하고 청년의 자발적인 참여를 이끌어 냄으로써 청년 정책의 목적을 실현할 수 있도록 조성된 시설을 말한다. 2018년을 기준으로 2020년, 의성군은 청년 시설과 프로그램의 증가로 청년 인구 감소를 극복한 사례가 있다. 이에 따라 영주시의 청년 시설을 강조하고 다중회귀, 결정트리, 랜덤포레스트 알고리즘을 활용하여 위에 언급된 문제점을 다각적 측면에서 고려하여 후보지역을 선정하고, 각 후보지에 타당한 청년시설을 함께 제안하였다.

2. 연구배경

청년시설 필요성 증가

> 의성군 농어촌 마을 스테이(Stay) 체계 구축 사례





[출처] 의성군의 청년 인구 증가 사례 기사 : https://www.hani.co.kr/arti/economy/heri_review/966936.html, https://www.mk.co.kr/news/society/10134504

2. 연구배경

필요성이 높은 청년시설

> 영주시 미래전략실 설문조사에 따르면

01
주거 문제

- 주거 비용 부담
- 특정 지역 쏠림현상

02
교육 문제

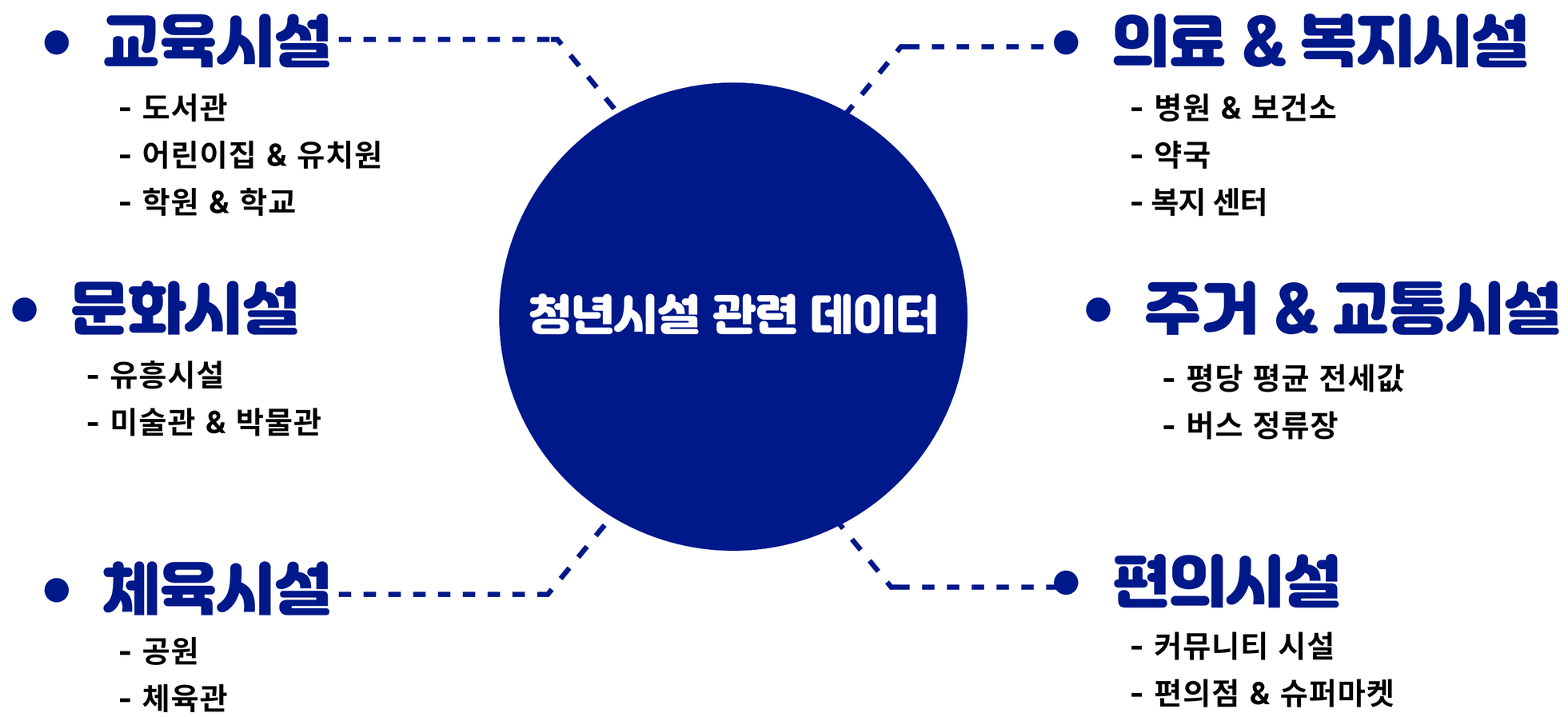
- 학습공간 부진
- 지역별 교육 격차

03
문화생활

- 문화시설 부진
- 문화시설 불균형

3. 분석 방법

> 데이터 활용


청년시설 관련 데이터
교육시설
- 도서관
- 어린이집 & 유치원
- 학원 & 학교
의료 & 복지시설
- 병원 & 보건소
- 약국
- 복지 센터
문화시설
- 유흥시설
- 미술관 & 박물관
주거 & 교통시설
- 평당 평균 전세값
- 버스 정류장
체육시설
- 공원
- 체육관
편의시설
- 커뮤니티 시설
- 편의점 & 슈퍼마켓

3. 분석 방법

> 데이터 출처

교육시설

- 경상북도 빅데이터 통합플랫폼 GB모아

의료 & 복지시설

- 중소벤처기업부, 상권정보 누리집
- 보건 의료빅데이터 개방 시스템 건강보험심사평가원

문화시설

- 문화체육관광부, 문화샘터 누리집

주거 & 교통시설

- 국토교통부, 국토교통부 실거래 공개시스템

체육시설

- 문화체육관광부, 문화샘터 누리집

편의시설

- 중소벤처기업부, 상권정보 누리집

인구 데이터

- 행정안전부, 주민등록 인구통계
- 경상북도 빅데이터 통합플랫폼 GB모아, 경상북도 시군구연령별 취업자 및 고용률

3. 분석 방법

> 데이터 수집 지역

데이터 수집 지역

영주 내 지역	영주 내 지역	영주 외 지역	영주 외 지역
가흥 1, 2동	영주 1,2동	완주군_삼례읍	완주군_봉동읍 & 용진읍
휴천 1,2,3동	상망동	예천군_예천읍 & 호명읍	의성군_의성읍
하망동	풍기읍	봉화군_봉화읍	증평군_증평읍
		삼척시_교동	삼척시_남양동 &정라동

3. 분석 방법

> 데이터셋 전반적인 내용

데이터 셋

영주1,2동

- 유흥시설 14개, 문화시설 1개
- 교육시설 10개, 어린이 시설 7개
- 의료시설 68개, 체육시설 7개

풍기읍

- 유흥시설 8개, 문화시설 2개
- 교육시설 13개, 어린이 시설 7개
- 의료시설 11개, 체육시설 3개

교육시설

- 학교
- 도서관
- 스터디 카페
- 학원

체육시설

- 운동장
- 체육관

3. 분석 방법

> 분석툴

R

데이터 전처리
회귀 분석
다중 공선성 확인

Python

데이터 전처리
결정트리 알고리즘
그리드 서치 알고리즘
랜덤 포레스트 알고리즘

Excel

데이터 수집 및 정렬
데이터 전처리

3. 분석 방법

> 전처리

청년 경제 지수

> 경제 참여율의 개념을 활용

$$\text{청년경제지수} = \frac{\text{지역의 청년인구 수} \times \text{청년 고용률}}{\text{지역의 전체인구 수}} \times 100$$

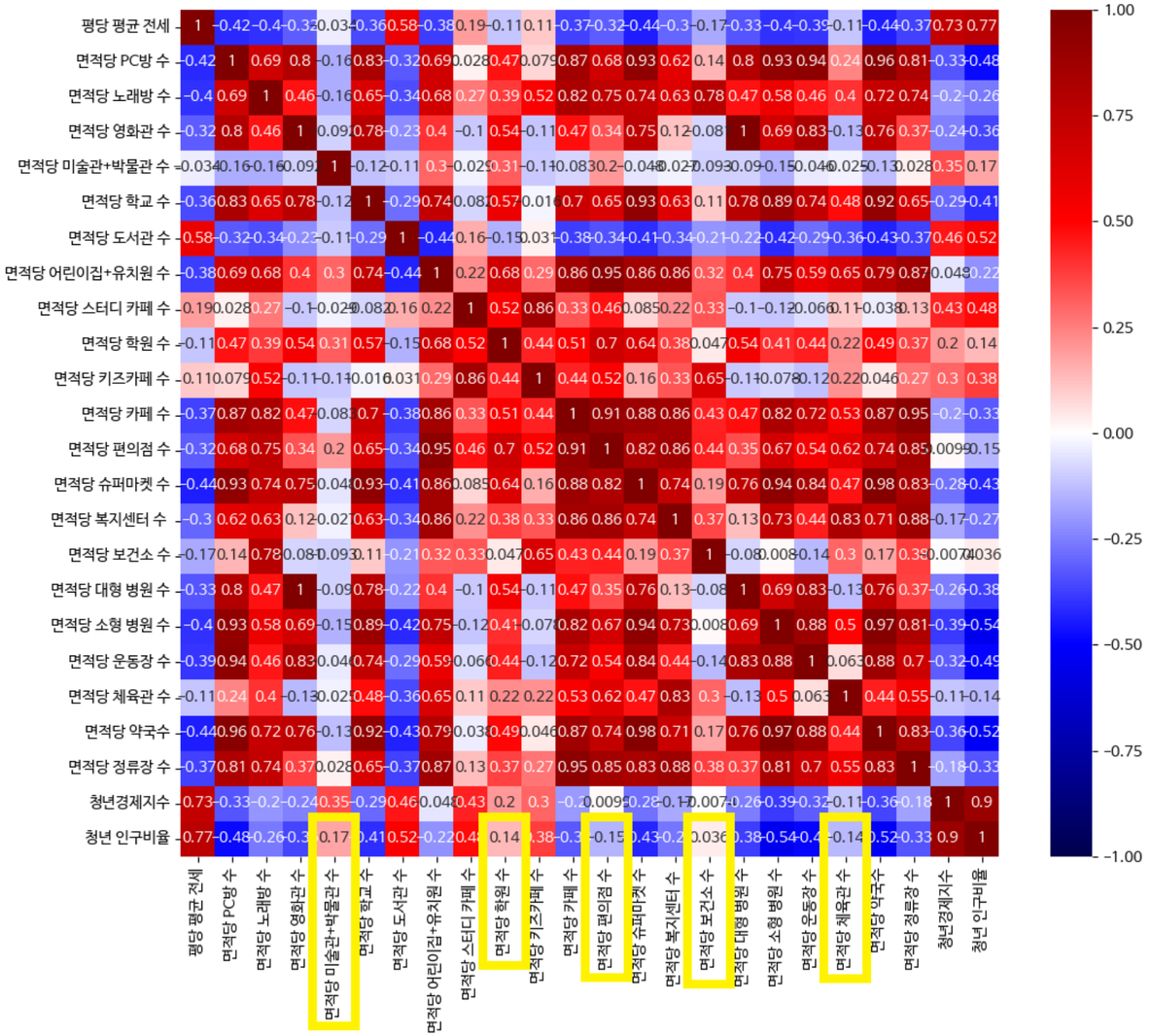
각 지역별 청년인구 수를 고려하여 전체 인구 수 대비 현재 영주시의 청년 고용률의 규모가 적정한지 판단하기 위한 계량적 척도

[참고 논문] : J. T. Park and H. S. Jang, "A Regional Comparison Study for the Variability of Employment Statistics in Korean Young Man: Focus on Economically Active Population Rate, Employment Population Rate, Unemployment Rate," Journal of Service Research and Studies, vol. 5, no. 1, pp. 35–43, Mar. 2015.

3. 분석 방법

> 전처리

상관 관계 히트맵



상관관계 : 변수들 간의 관계를 계산해서 수로 표현

상관관계가 0에 가까울 수록
두 변수는 서로 관계가 없음

청년 인구비율과 관계가 없는 변수들이 존재한다는 결과는
이 변수들은 청년 인구비율에 전혀 영향을 주지 못한다는 의미

3. 분석 방법

> 전처리

다중공선성 확인

```
> vif(lm_result)
          면적.km.2.           평당.평균.전세       면적당.노래방.수       면적당.영화관.수
            2.536458            4.483478            4.861760            6.680950
면적당.미술관.박물관.수       면적당.학교.수       면적당.도서관.수       면적당.학원.수
            2.838861            3.104109            3.761447            4.347157
   면적당.키즈카페.수       면적당.카페.수       면적당.편의점.수       면적당.보건소.수
            4.080733           15.215240            9.418692            5.094359
   면적당.대형.병원.수   면적당.소형.병원.수       면적당.운동장.수       면적당.체육관.수
            2.242865           17.921759            4.650767            4.131905
      청년경제지수       면적당.슈퍼마켓.수       면적당.정류장.수       면적당.약국수
            5.728081           15.593335           11.977410           17.770856
  면적당.PC방.수 면적당.어린이집.유치원.수       면적당.복지센터.수
            5.448441            8.739269            9.675137
```

독립 변수들 간의 선형 관계를 나타내는 상관 행렬

(독립 변수들 간에 강한 상관 관계가 있을 때 발생)

$$VIF() < 5 \quad : \text{안전}$$

$$5 \leq VIF() < 10 \quad : \text{다중공선성 유의}$$

$$10 < VIF() \quad : \text{다중공선성 상태}$$

다중선형회귀 분석

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           11.54191    2.31068   4.995  3.4e-05 ***
면적.km.2.            -0.01935    0.02338  -0.828  0.41548
평당.평균.전세         0.01125    0.00463   2.431  0.02228 *
면적당.노래방.수       -0.29105    0.29019  -1.003  0.32569
면적당.미술관.박물관.수  0.83645    2.33443   0.358  0.72300
면적당.학교.수          0.20592    0.44999   0.458  0.65104
면적당.도서관.수       -0.91865    1.80573  -0.509  0.61522
면적당.학원.수          0.36391    0.28063   1.297  0.20610
면적당.키즈카페.수      0.26750    1.45928   0.183  0.85598
면적당.카페.수         -0.05548    0.14394  -0.385  0.70306
면적당.보건소.수        1.98940    1.87915   1.059  0.29949
면적당.대형.병원.수    -1.48924    0.65719  -2.266  0.03200 *
면적당.운동장.수       -0.49972    0.41598  -1.201  0.24046
면적당.체육관.수        0.02204    0.55737   0.040  0.96876
청년경제지수            0.55561    0.19739   2.815  0.00918 **
면적당.PC방.수          0.34635    0.54945   0.630  0.53396
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.096 on 26 degrees of freedom
Multiple R-squared:  0.7698,    Adjusted R-squared:  0.637
F-statistic: 5.797 on 15 and 26 DF,  p-value: 4.987e-05
```

귀무가설 기각 : p-value 0.00004

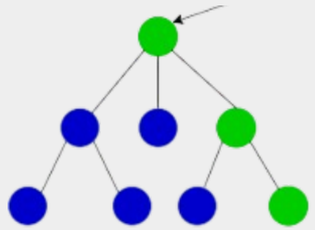
영주시 청년인구 증가에 유의미한 청년시설이 아닐 것이다 라는 귀무가설을 기각

→ 결론으로, p-value 값이 0.05보다 작기 때문에 해당 분석은 통계적으로 유의하다고 볼 수 있음

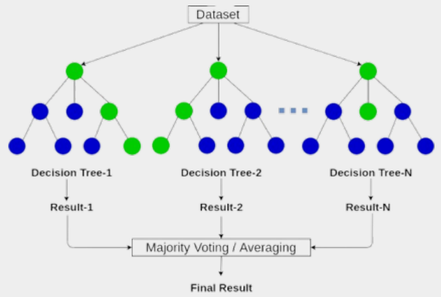
3. 분석 방법

> 분석 프로세스

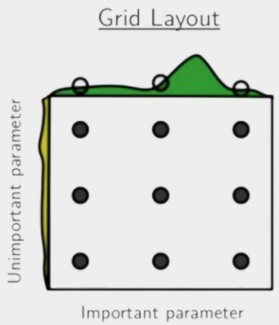
청년시설 최적의 입지 선정 프로세스



Decision Tree



Random forest



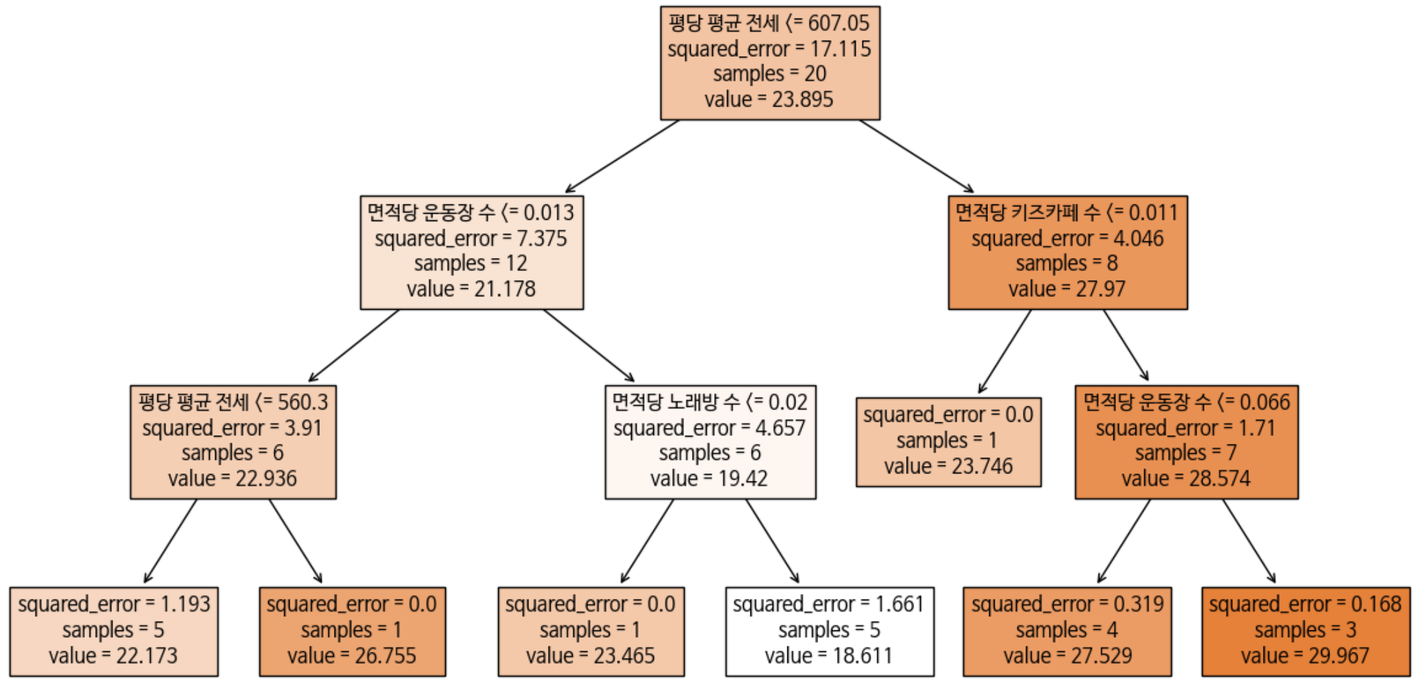
Grid search

Result

3. 분석 방법

> 분석 프로세스

Decision Tree



결정 트리 알고리즘 :

변수의 특정값을 기준으로
데이터를 분류해서
결과를 예측하는 알고리즘

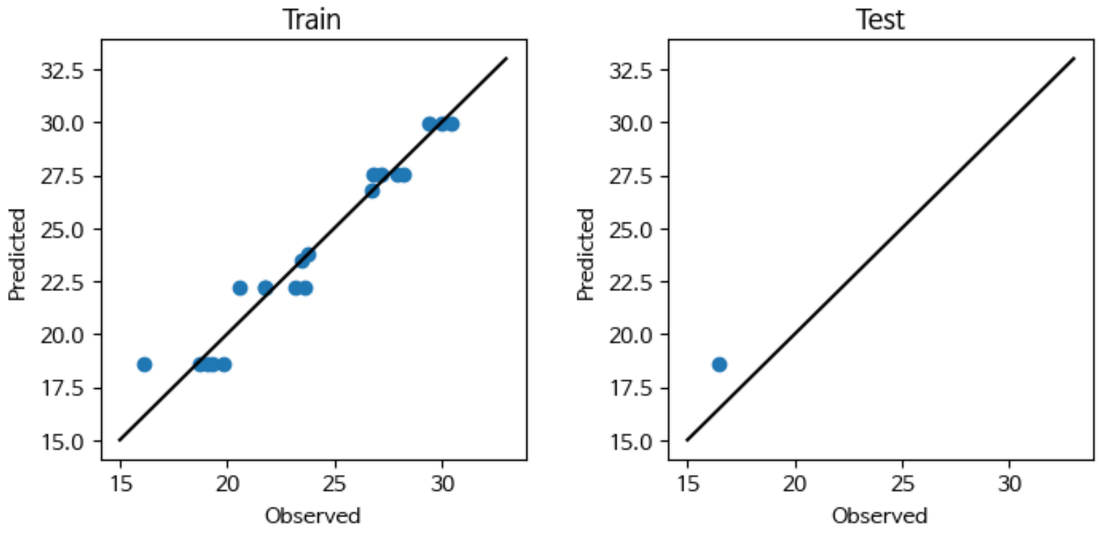
영주시를 제외한 20개 지역의
데이터로 모델을 학습시킨 결과,
키즈카페, 운동장의 수가 높을 수록
청년 인구 비율을 높게 예측

3. 분석 방법

> 분석 프로세스

Decision Tree

KEEP




```
[22]  RMSE(y_train, Tree_pred_train)
```

0.895809099225878 ——— 학습 데이터의 오차 평균

```
[23]  RMSE(y_test, Tree_pred_test)
```

2.1181723419999976 ——— 영주 1동의 실제 값과 머신의 예측값의 차이

RMSE (평균 제곱근 오차) :

$$\text{평균 제곱근 오차} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

실제 값과 예측 값의 차이를 제곱한 후
평균을 구해 제곱근을 취한 값

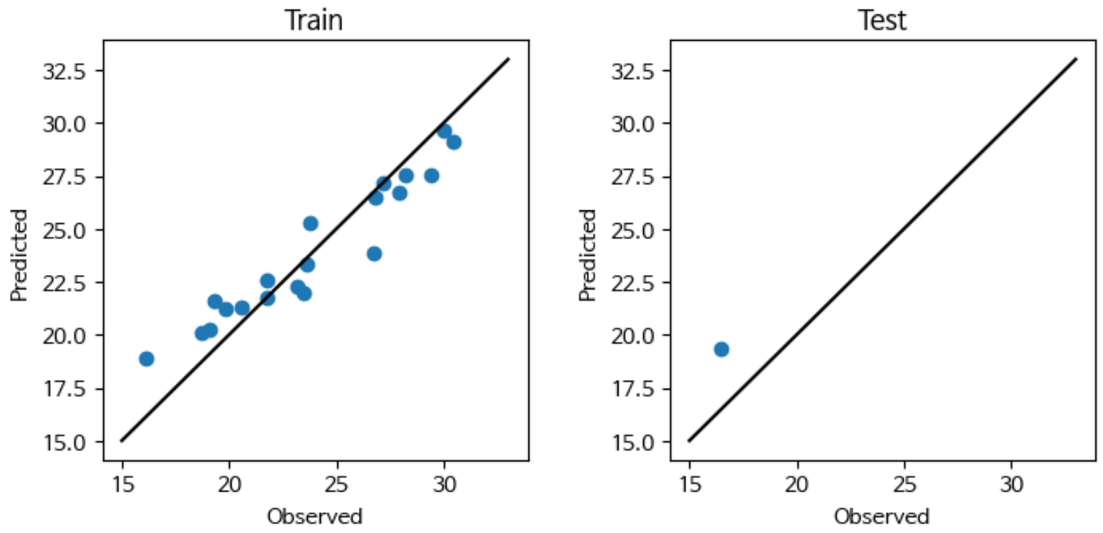
20개 지역의 평균 제곱근 오차는 0.89 정도로
결과가 작게 나왔으나 영주 1동은 예측 값과
실제 값의 차이가 2.11 정도로 크게 나옴
따라서 이 모델을 보류

3. 분석 방법

> 분석 프로세스

Random Forest

REJECTED




```
[30]  RMSE(np.array(y_train), RF_pred_train)
      1.4312692391594224
```

학습 데이터의 오차 평균

```
[31]  RMSE(np.array(y_test), RF_pred_test)
      2.8439502991722208
```

영주 1동의 실제 값과
머신의 예측값의 차이

랜덤 포레스트 알고리즘 :

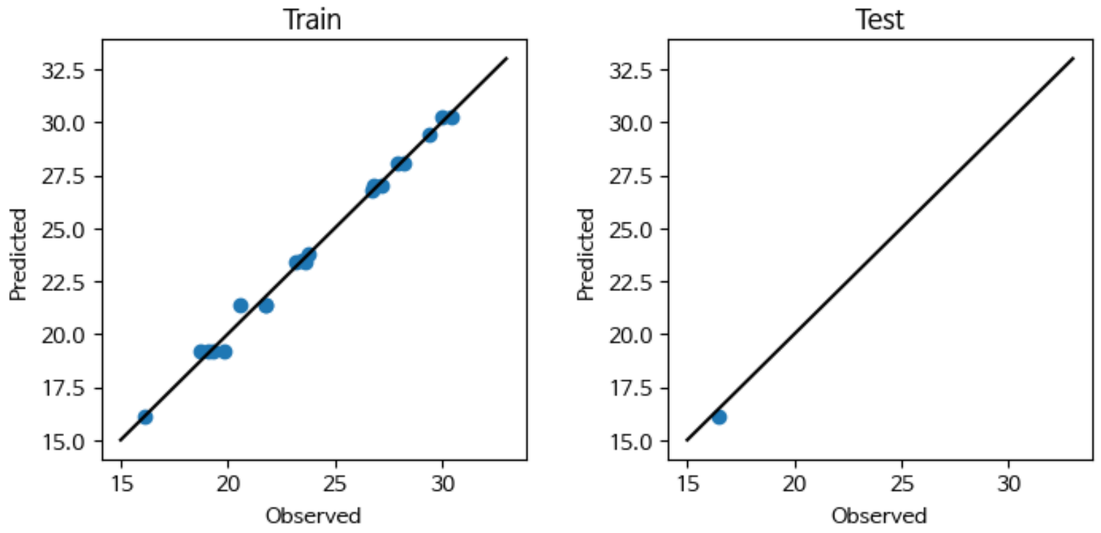
여러 개의 결정 트리를 만들고
그 중에서 가장 나은 트리를 하나 뽑는
알고리즘으로 결정트리와 비슷하지만
더 나은 예측 성능을 보일 수 있음

그러나 20개 지역의 평균 제곱근 오차가 1.43,
영주 1동의 평균 제곱근 오차는 2.84로
결정 트리 모델보다 오히려 예측 성능이 떨어짐
따라서 이 모델을 기각

3. 분석 방법

> 분석 프로세스

Grid Search




```
RMSE(np.array(y_train), GS_DT_pred_train)
```

0.31013910666382216 — 학습 데이터의 오차 평균

```
[50] RMSE(np.array(y_test), GS_DT_pred_test)
```

0.3565592800000026 — 영주 1동의 실제 값과 머신의 예측값의 차이

그리드 서치 알고리즘 :

알고리즘의 최적 파라미터를 찾는 알고리즘
랜덤 포레스트를 기각하였으므로
결정트리의 파라미터를 최적화하기로 결정

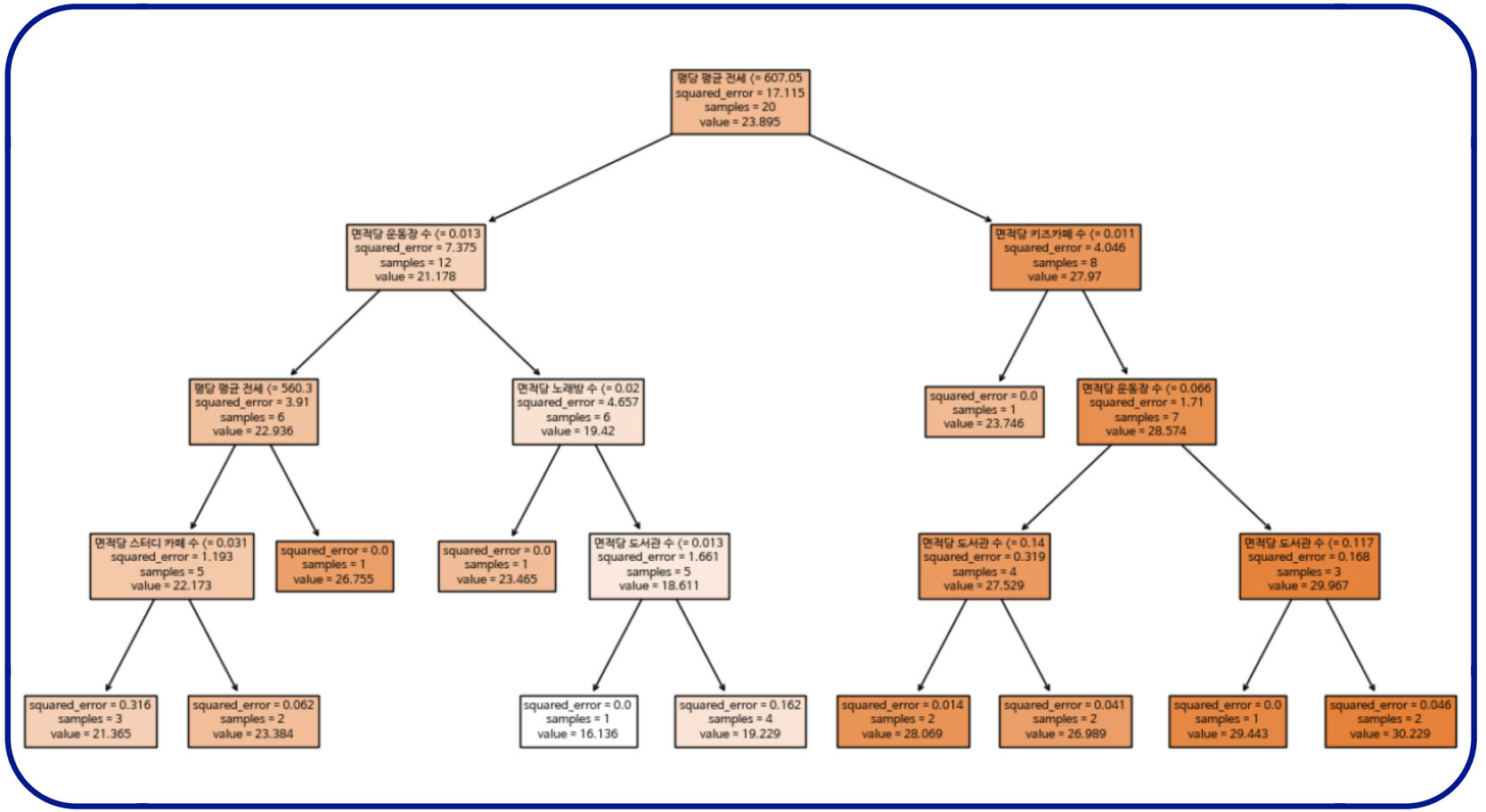
20개 지역의 평균 제곱근 오차가 0.31,
영주 1동의 평균 제곱근 오차는 0.35로
랜덤 포레스트는 물론 기존의 결정 트리보다
예측 성능이 훨씬 좋다고 판정
따라서 이 모델을 **채택**

ACCEPTED

3. 분석 방법

> 분석 프로세스

Grid Search




```
[47]  GS_DT_model.best_params_

      {'max_depth': 4, 'random_state': 40}
```

그리드 서치 결과
최대 깊이 4가 최적

그리드 서치로 파라미터를 최적화 시킨 결정트리의 트리구조에서 운동장, 도서관, 키즈카페, 스터디카페가 많을 수록 청년 인구 비율이 높게 예측

4. 분석결과

> 최종지역 선정

	면적(km^2)	유흥시설	교육시설	보육시설	체육시설	청년경제지수	청년인구비율
영주 1동	1.02	10	9	4	4	9.1	16.5
영주 2동	0.8	4	1	3	3	8.7	16.1
풍기읍	76	8	13	7	3	10.5	20.6
평균	28	4	5	6	2	–	–

지역 선정 근거 :
영주시에서 청년경제지수, 청년인구비율 낮은 지역 3개 선정

영주 1동 : 보육시설
영주 2동 : 교육시설, 보육시설
풍기읍 : 체육시설

4. 분석결과

> 분석결과

영주 1, 2 동의 청년 인구 비율 증가 예시

기존 데이터

```
Tree_model.predict(X_test)
```

```
array([16.13601906])
```

```
Tree_model.predict(X_yj2.values.reshape(1,-1))
```

```
array([16.13601906])
```

영주 1동과 영주 2동의 도서관 수를 늘리자
청년 인구 비율이 기존보다 높게 예측됨

도서관 1개를 추가할 경우

```
Tree_model.predict(X_yj1)
```

```
array([19.22943359])
```

청년인구 비율
19.17% 증가
+ 청년경제지수 증가

```
Tree_model.predict(X_yj2.values.reshape(1,-1))
```

```
array([19.22943359])
```

청년인구 비율 19.17% 증가

영주 1동과 영주 2동의 면적이 넓지 않음을 고려하여 영주 1동과 영주 2동의 중간 지점 인근에 도서관을 세우는 것이 필요하다고 예측됨

4. 분석결과

> 분석결과

풍기읍의 청년 인구 비율 증가 예시

기존 데이터

```
Tree_model.predict(X_pg.values.reshape(1,-1))
```

```
array([21.36499665])
```

스터디카페 3개를 추가할 경우

```
Tree_model.predict(X_pg.values.reshape(1,-1))
```

```
array([23.38432496])
```

청년인구 비율 9.45% 증가

+ 청년경제지수 증가

풍기읍의 스터디 카페 수를 늘리자
청년 인구 비율이 기존보다 높게 예측됨

풍기읍의 교육 시설 수가 많음을 고려해
빈 집을 활용해 학생들이 자유롭게 자습할 수 있는
공간을 만들어주는 것이 필요하다고 예측됨

기대 효과

프로젝트 기대효과

- 독립변수 데이터가 확보된다면 예측의 정확도가 높아지며 모델의 안정성이 더욱 증가될 것이라 기대된다.
- 빈 집의 위치 데이터나 면적을 파악할 수 있는 데이터가 있다면, 특정 위치에 어떤 청년시설이 필요로 하고 들어가면 좋을지 기대된다.

문화

지역 문화 활동 촉진

사회

지역 커뮤니티 활성화

경제

지역 경제 발전

4. 분석결과

> 참고문헌

참고문헌

[1] 이유재(1994). 상호작용효과를 포함한 다중회귀분석에서 주효과의 검증에 대한 연구, 경영학연구.

[2] 고정우, 이주림, 구자훈. (2023). 대도시 및 중소도시 입지별 혁신도시의 청년인구 유입효과 비교분석. 한국지역개발학회지, 35(2), 45-68.

[3] 김재환, 이민구(2018). 의사 결정 트리와 랜덤 포레스트 모델 분석, 한국정보기술학회, 한국디지털콘텐츠학회

[4] 박태일. "인공지능 알고리즘과 최적화 기법의 결합을 통한 비선형 회귀 알고리즘 개발." 국내석사학위논문 한국항공대학교 일반대학원, 2023. 경기도

[5] 이광원. (2023). 위계적 다중 회귀분석을 활용한 지방소멸 요인이 핵심생산가능인구의 순이동에 미치는 영향. 한국정책과학학회보, 27(3), 131-155, 10.31553/kpsr.2023.9.27.3.131

[6] 손창희, 장한두. (2019). 지방도시의 노후주거지 정비와 관리 정책에 관한 연구 - 주거환경개선사업의 도시규모별 대안을 중심으로 -. 주거환경, 17(2), 197-214.

[7] 조현영, 이상복, 김홍렬, 외 2인, (2012). 전국 도서관 통계 조사지표 개선 연구, 문화체육관광부.

[8]J. T. Park and H. S. Jang, “A Regional Comparison Study for the Variability of Employment Statistics in Korean Young Man: Focus on Economically Active Population Rate, Employment Population Rate, Unemployment Rate,” Journal of Service Research and Studies, vol. 5, no. 1, pp. 35–43, Mar. 2015.

감사합니다