The Journal of the Convergence on Culture Technology (JCCT)
Vol. 9, No. 6, pp.131-136, November 30, 2023. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2023.9.6.131

JCCT 2023-11-16

머신러닝 기반 부산 청년인구 유출위험 요인 분석

Analysis of Risk Factors for Youth Population Outflow in Busan Based on Machine Learning

손서영*, 양혜성**, 박민서***

Seoyoung Sohn*, Hyeseong Yang, Minseo Park*****

요 약 지방을 떠나 수도권으로 이동하는 청년들이 나날이 늘고 있다. 청년 유출의 요인을 파악하기 위한 연구들이 다양하게 진행되고 있으나 각 지방별로 분석하는 연구는 부족한 실정이다. 이에 따라 본 연구는 부산의 청년 인구 유출 요인을 분석하고, 머신러닝 기법을 사용해 청년 인구 유출 위험 등급을 예측하고자 한다. 국가통계포털에서 수집한 부산의 구별 데이터를 활용하여 나이대(20대 초반, 20대 후반, 30대 초반)별로 세 그룹으로 나눈 뒤, 의사결정나무와 랜덤 포레스트 알고리즘을 이용해 청년 인구 유출 위험 등급을 분류하고 예측한다. 그 결과, 청년 인구 유출 위험 등급 예측 모델은 나이 그룹별 각각 최고 정확도 0.93, 0.75, 0.63을 갖는다.

주요어 : 머신러닝, 청년 유출, 지방 소멸

Abstract Local youth outmigration is increasingly growing. Various studies are being conducted to identify the factors contributing to this problem, but there is a lack of research analyzing each region individually. Therefore, this study aims to analyze the factors influencing youth outmigration in Busan and predict the risk levels of youth population outflow using machine learning techniques. By utilizing district-level data collected from the KOSIS, we divided the population into three groups based on age (the early 20s, late 20s, and early 30s) and employed Decision Tree and Random Forest algorithms to classify and predict the risk levels of youth population outmigration. The results indicate that the predictive model for youth outmigration risk levels achieves the highest accuracies of 0.93, 0.75, and 0.63 for each age group, respectively.

Key words : Machine Learning, Youth Outmigration, Local Extinction

Ⅰ. 서 론

지방 소멸은 대한민국에서 심각한 사회 이슈 중 하나이다. 저출생과 수도권 인구 집중 현상으로 인해, 2047년에는 전국 229곳이 소멸 위험 단계에 진입할 전망이다[1]. 수도권과 비수도권이 불균형하게 발전함에 따라 경제, 교육, 의료, 문화 등 대부분의 분야에서 비수도권이 경쟁력을 잃고 있는 것이다[2]. 여기에서 가장 주목해야 할 점은 청년들의 유출이다. 통계청의 인구 이동 데이터를 보았을 때, 현재 청년층은 지방에서 가

*준회원, 서울여자대학교 데이터사이언스학과 학부생 (제1저자)
**준회원, 서울여자대학교 데이터사이언스학과 학부생 (참여저자)
***정회원, 서울여자대학교 데이터사이언스학과 조교수 (교신저자)
접수일: 2023년 10월 3일, 수정완료일: 2023년 10월 19일
게재확정일: 2023년 11월 5일

Received: October 3, 2023 / Revised: October 19, 2023
Accepted: November 5, 2023
***Corresponding Author: mpark@swu.ac.kr
Dept. of Data Science, Seoul Women's Univ, Korea

장 많이 유출될 뿐만 아니라 출생률과 밀접한 관련이 있는 세대다. 또, 청년은 지역 경제 활성화 및 성장의 동력으로써 중요한 역할을 한다[3]. 따라서, 청년층의 유출은 지방도시의 존속에 큰 영향을 미치게 된다.

많은 지방도시 중에서 가장 큰 규모인 부산은, 제2의 도시라고 불릴 만큼 영향력 있는 도시이다. 그러나 부산은 2022년 기준 순유출 5974명으로 광역시 중에서 가장 많은 청년이 유출되었고, 청년 인구 비율이 27.3%였던 2013년에 비해 2022년에는 22.1%로 감소하며 전체 인구보다 더욱 가파른 하락세를 보였다[4]. 더불어 합계출산율은 0.72명으로 서울에 이어 두번째로 낮았으며[5], 국내 대도시 가운데 유일하게 초고령사회에 진입한 도시이기도 하다.

이러한 심각성을 인식하여 현재 부산 청년들의 유출 특성 요인을 분석하는 연구들이 활발히 이뤄지고 있다. 그러나, 선행 연구들은 권역 또는 지역 전체를 대상으로 하거나 모든 청년보다는 인재 유출(Brain Drain)을 중심으로 다루는 실정이다. 따라서, 본 연구는 청년 인구의 유출 위험 등급을 구별로 도출하고, 청년 유출에 영향을 미치는 요인을 분석하고자 한다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서 청년 인구 유출에 관한 선행 연구에 대해 살펴본다. 3장에서는 청년 인구 유출 위험 등급 예측에 효과적인 머신러닝 알고리즘을 사용하여 모델을 제안한다. 4장에서는 모델의 결과를 살펴보며, 5장에서 결론을 기술한다.

II. 청년 인구 유출 요인

청년 인구 유출과 관련된 선행연구를 조사하였다. 지방 주민들의 전출입과 관련된 지역 어메니티(amenity)에 대한 연구에서는 경제적 요소와 더불어 근린, 도시, 환경 어메니티가 주거 이전에 영향을 미친다고 나타났다[6]. 근린 어메니티로는 의료 서비스, 주택의 가격, 사회복지 서비스, 교육(초, 중, 고등학교)의 질 등이 있고, 도시 어메니티로는 백화점·대형마트, 대학의 질, 문화 공간 등이 있다. 더하여, 환경 어메니티로는 대기의 질, 공원 및 오픈 스페이스, 체육 시설 등이 있다. 또, 도시 어메니티의 지역 경제 활성화 효과에 관한 연구에서는 이러한 어메니티들을 중심으로 지역 경제가 활성화되며, 어메니티가 인구를 유입시키는데 긍정적인 작용을 한다고 밝혔다[7].

선행연구에 따라 본 연구에서는 청년들의 주거 이전에 영향을 미치는 다양한 요소들을 더 직관적으로 반영하기 위해 새로운 분류 체계를 적용하였다. 경제 및 근린, 환경, 도시 어메니티에 해당하는 의료 서비스, 교육, 사회복지 서비스, 문화 공간 등을 총 5가지 분야(경제, 보건, 공공행정, 환경 및 인프라, 교육)로 나누어 분류하였고, 해당 분야들을 머신러닝 시 특성 변수로 사용하였다([표 1] 참고).

III. 청년 인구 유출위험 등급 예측 모델

먼저, 부산의 청년 인구 유출 요인을 분석하기 위해 유출 위험 등급을 분류하였다. 나이대별로 세 그룹(20대 초반, 20대 후반, 30대 초반)으로 나누고, 각 그룹별로 총전출인구수 대비 총전입인구수를 5분위로 나누어 인구 유출 위험 등급을 도출하였다. 그리고 머신러닝의 대표적인 분류 모델인 의사결정나무(Decision Tree)와 랜덤 포레스트(Random Forest)를 적용하여 모델링을 수행하였다.

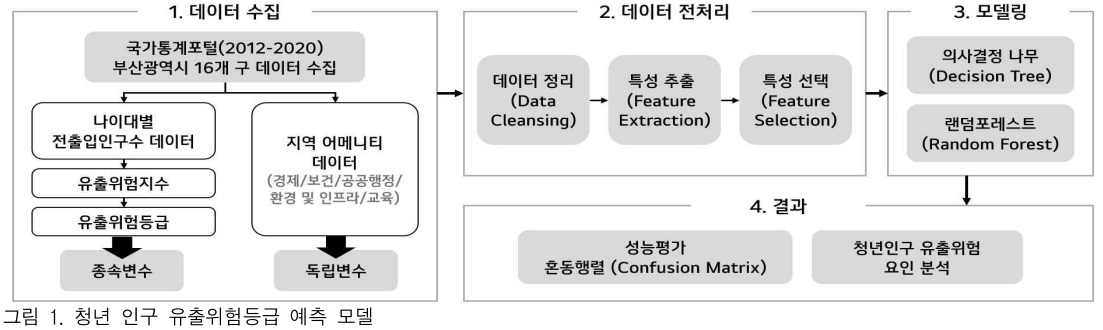

그림 1. 청년 인구 유출위험등급 예측 모델
Figure 1. Prediction Model for Risk Level of Youth Population Outflow

표 1. 특성에 따른 변수 구성
Table 1. Configuring Variables Based on Attributes

특성		변수
종속변수		유출위험지수로 계산된 나이대별 유출위험등급(안전~고위험)
독립변수	경제	산업별 지역 내 총생산, 산업별 사업체 수, 종사자 수 1000명 이상 산업체 수, 소상공인 개수, 소기업 개수, 중기업 개수, 중소기업 개수, 대기업 개수
	보건	병상 수, 종합병원 수, 요양병원 수, 치과 병/의원 수, 한방병원/한의원 수, 보건복지시설 수, 약국 수, 의원 수, 의료진 수
	공공 행정	혼인 건수, 인구수, 인구증가율, 주택보급률, 사회복지예산 비중
	환경·인프라	녹지개수, 녹지면적, 문화공간수, 주택 수
	교육	유치원 및 유치원 교직원 수, 초등학교 및 초등학교 교직원 수, 중학교 및 중학교 교직원 수, 고등학교 및 고등학교 교직원 수, 대학교 및 대학교 교직원 수, 사설학원 수
파생변수		산업별 비중, 노인인구 비중

또한, 모델의 성능을 높이기 위해 하이퍼파라미터 튜닝을 진행하였고, 최종적으로 가장 높은 성능을 보이는 모델의 변수 중요도를 시각화하였다. 전체적인 프로세스는 [그림 1]과 같다.

1. 데이터 수집

부산광역시의 청년인구 유출위험등급을 구하기 위해, 국가통계포털에서 제공하는 '시군구/연령(5세)별 이동자수' 데이터셋 중 2012년부터 2020년까지 부산광역시 16개 구의 총 전입인구수와 총 전출인구수 데이터를 수집하였다. 청년 기본법에 따라 청년 인구를 20세부터 34세로 한정했고, 나이대별로 특성이 다르기 때문에[8], 20대 초반(20~24세), 20대 후반(25~29세), 30대 초반(30~35세) 데이터를 각각 수집하였다. 이후, 총전입인구수를 총전출인구 수로 나눈 유출위험지수를 5분위로 나눈 유출위험 등급을 종속 변수로 활용하였다.

또한 앞서 지방 소멸에 영향을 주는 지역 어메니티에 대한 선행연구[7]를 기반으로 부산광역시 청년인구 유출에 영향을 미치는 5가지 분야의 특성 데이터를 수집하여 독립변수로 사용하였다. 5가지 분야는 경제, 보건, 공공행정, 환경 및 인프라, 교육이며 전체적으로 수집한 데이터셋은 [표 1]과 같다.

2. 데이터 전처리

1) 데이터 정리(Data Cleansing)

변수들의 결측치를 확인한 후 평균값이나 전년도 값으로 대체하였다. 병상 수와 같이 해마다 값이 변하는 변수의 결측치는 전년도와 후년도 평균값을 계산하여 대체하였고, 녹지 개수나 대기업 수와 같이 해마다 값의 변동이 거의 없는 변수들의 결측치는 전년도의 값으로 대체하였다. 결측치 처리가 끝난 후 데이터 부족으로 인한 정확도의 문제를 보완하기 위해 144개의 데이터셋을 100배로 복제하여 14,400개의 행을 가진 최종 데이터셋을 만들었다.

2) 특성 추출(Feature Extraction)

머신러닝 모델링 시 종속변수로 사용되는 위험지수는 총전입인구 수를 총전출인구 수로 나누어 계산하였다. 지수가 1 미만이면 전출 인구가 더 많아 순유출되고 있다는 뜻이고, 1 이상이면 전입 인구가 더 많아 순유입되고 있다는 의미이다. 따라서 [그림 2]와 같이 유출위험지수가 1 미만인 데이터와 1 이상인 데이터를 나눠, 각각 등급을 나누었다.

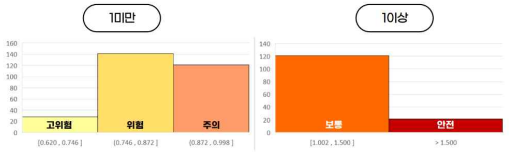

그림 2. 유출위험등급 히스토그램
Figure 2. Histogram of Outflow Risk Rating

그 결과 고위험, 위험, 주의, 보통, 안전 순으로 총 5개의 등급을 추출하였다. 이렇게 분류한 유

출 위험 등급을 종속변수로 활용하였다. 독립변수의 경우, 수집한 5가지 분야의 특성 데이터 중 사업체 수나 노인인구 수와 같이 개수보다는 지역에서 차지하는 비중이 중요한 경우, [표 1]에서 볼 수 있듯이 파생 변수를 생성하여 반영하였다.

3) 특성 선택(Feature Selection)

위의 과정을 거쳐서 수집한 총 75개의 독립변수 및 파생변수 타당성을 VIF(Variance Inflation Factor)를 통해 검증하였다. 높은 다중공선성을 보인 변수를 제거하였다. 그러나 혼인 지수의 경우는 VIF가 10.7로 10을 조금 넘지만 유의미하다고 추정되어 독립변수로 채택하였다. 최종적으로 선정된 변수들은 [표 2]와 같다.

표 2. 최종 선정된 독립변수들의 VIF 결과
Table 2. VIF results of the final selected independent variables

Features	VIF
혼인 건수	10.7
사업체 수_금융 및 보험업	8.7
지역내총생산_부동산업	8.7
대기업 개수	8.2
종합병원 수	6.8
전기, 가스, 증기 및 수도사업 비중	5.9
녹지 개수	5.2
종사자 수 1000명 이상 산업체 수	4.3
대학교 교직원 수	3.9
제조업 비중	3.7
문화공간 수	3.6
대학교 수	3.5
전문 과학 및 기술서비스업 비중	3.3
광업 비중	3.0
농업·임업 및 어업 비중	2.5
유치원 교직원 수	1.7
약국 수	1.5

3. 모델링

데이터 전처리 후 최종 선정한 데이터셋을 8:2의 비율로 훈련 데이터와 검정 데이터로 나누었고, 대표적인 두 가지 분류 모델(의사결정나무, 랜덤 포레스트)을 사용하여 학습하였다. 각 모델은 하이퍼파라미터(표 3 참조)를 튜닝하여 최적의 결괏값을 도출하였다. 성능 지표로는 혼동 행렬(Confusion Matrix)를 사용하였다. 랜덤 포레스트를 최종 모델로 선정하였다.

IV. 결 과

[표 3]은 연령대별 두 가지 분류 모델(의사결정나무, 랜덤 포레스트)의 하이퍼파라미터를, [표 4]는 성능지표를 비교한 표이다. 정확도(훈련 데이터의 정확도 Train_acc와 검정 데이터의 정확도 Test_acc) 및 F1-score 모두 랜덤 포레스트가 높은 성능을 보였으며, 학습 정확도와 검증 정확도의 차이도 적었다.

표 3. 연령대별 두 모델의 하이퍼파라미터
Table 3. Hyperparameters of Two Age Models

Age Group	Model	Hyperparameters	
		max_depth	n_estimators
20대 초반	Decision Tree	4	-
	Random Forest	6	35
20대 후반	Decision Tree	3	-
	Random Forest	4	35
30대 초반	Decision Tree	3	-
	Random Forest	4	15

표 4. 연령대별 두 모델의 결과
Table 4. Performance between Two Age Models

Age Group	Model	Train_acc	Test_acc	F1-score
20대 초반	Decision Tree	0.758	0.688	0.74
	Random Forest	**0.977**	**0.938**	**0.93**
20대 후반	Decision Tree	0.711	0.438	0.41
	Random Forest	**0.852**	**0.750**	**0.75**
30대 초반	Decision Tree	0.539	0.438	0.30
	Random Forest	**0.773**	**0.625**	**0.61**

20대 초반 모델에서는 학습 정확도 0.997, 검증 정확도 0.938, F1-score 0.93으로 우수한 성능을 보였다. 그러나 20대 후반과 30대 초반의 경우 랜덤 포레스트 모델에서도 각각 학습 정확도 0.852, 검증 정확도 0.750, F1-score 0.75의 성능과 학습 정확도 0.773, 검증 정확도 0.625, F1-score 0.61로 다소 낮은 정확도를 보였다.

취업 및 고용에 민감한 나이대인 만큼 해당 데이터의 부재가 영향을 미쳤을 것으로 판단된다. 하지만, 기존의 많은 연구들이 20대 후반과 30대 초반 청년 인구 유출을 고용과 취업 중점으로 본 것과 달리 삶의 질과 관련된 특성의 중요도를 살필 수 있었다는 점에서 의의가 있다. 따라서 모든 나이대에서 랜덤 포레스트를 최종 모델로 선정하였다.

각 나이대별 특성 중요도를 추출한 그래프를 나타내는 [그림 3]을 통해 부산 청년 인구 유출에 영향을 미치는 요인을 볼 수 있다. 먼저, 20대 초반 특성 중요도는 약국 수, 금융 및 보험업 사업체 수, 유치원 교직원 수, 혼인 건수, 문화공간 수, 대학교 교직원 수, 부동산업의 지역 내 총생산, 종합병원 수, 대기업 수, 그리고 대학교 수의 순으로 나타났다. 다른 연령대에 비해 교육적 특성이 많고 대학교 수라는 변수가 유일하게 상위 10개 안에 등장함으로써 20대 초반 청년들에게 교육, 경제, 의료, 문화적 특성 중에서 교육적 특성이 중요함을 알 수 있다.

다음으로 20대 후반 특성 중요도는 금융 및 보험업의 사업체 수, 유치원 교직원 수, 혼인 건수, 대기업 개수, 문화공간 수, 대학교 교직원 수, 부동산업의 지역 내 총생산, 공공행정, 국방 및 사회보장 행정의 지역 내 총생산, 종합병원 수, 그리고 사업서비스업의 지역 내 총생산의 순으로 나타났다. 20대 후반부터 일자리, 취업과 관련된 경제적 변수들이 증가하였고, 행정과 관련된 변수가 나오기 시작하며 복지에 대한 관심이 증가하였음을 알 수 있다.

30대 초반의 특성 중요도는 문화공간 수, 대학교 교직원 수, 약국 수, 제조업 사업체 수, 공공행정, 국방 및 사회보장 행정의 지역 내 총생산, 혼인 건수, 유치원 교직원 수, 부동산업의 지역 내 총생산, 금융 및 보험업의 사업체 수의 순으로 나타났다. 20대 후반에 이어 30대 초반 역시, 경제적 요소가 가장 많았다. 또, 행정과 관련된 변수의 중요도가 올라갔으며, 문화공간 수가 가장 높은 중요도를 차지하는 것을 보아 삶의 질과 관련된 관심이 증가하였음을 알 수 있다.

마지막으로 세 연령대에서 공통적으로 등장하는 변수는 부동산업의 지역 내 총생산, 금융 및 보험업의 사업체 수, 문화공간 수, 유치원 교직원 수, 대학교 교직원 수, 혼인건수였다. 이러한 변수들은 앞서 설정했던 5가지 지역 특성 분야들 중, 경제, 문화, 교육, 공공행정에 해당한다. 이는 해당 분야들이 모든 나이대의 청년 인구 유출에 영향을 미친다고 볼 수 있다.

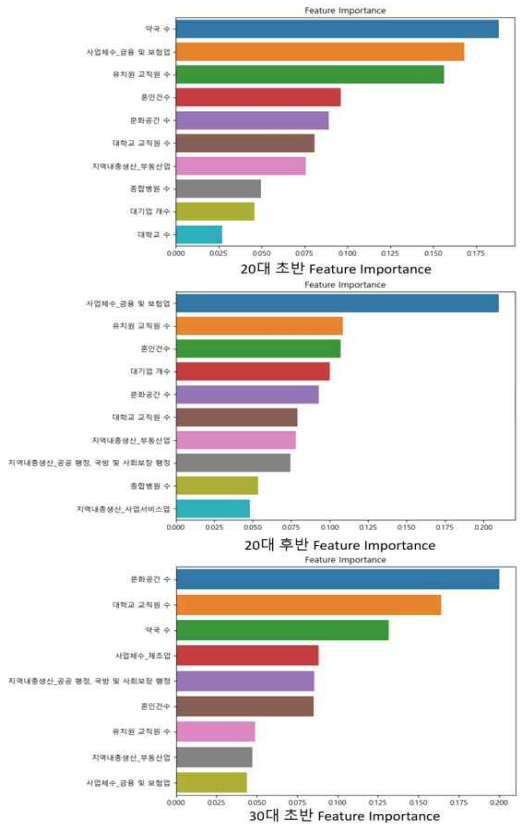

그림 3. 연령별 청년 유출 위험 예측 모델의 특성 중요도 그래프
Figure 3. Features Importance Graph of Youth Population Outflow Perspective Model

V. 결 론

본 연구는 머신러닝 기법을 활용하여 부산의 청년 인구 유출에 영향을 미치는 요인을 분석하고, 청년 인구 유출 위험 등급 예측 모델을 설계하였다. 나이대에 따라 20대 초반, 20대 후반, 30대 초반으로 데이터셋을 만들었고, 5가지(경제, 보건, 공공행정, 환경 및 인프라, 교육) 분야의 지역 특성을 독립변수로 사용하여 의사결정나무와 랜덤 포레스트 알고리즘으로 학습하였다. 하이퍼파라미터를 튜닝하고 학습 정확도, 검증 정확도, F1-score를 성능 지표로 활용하여 검증한 결과, 랜덤 포레스트 알고리즘이 모든 연령대에서 우수한 성능을

보였다.

20대 초반 모델에서는 학습 정확도 0.997, 검증 정확도 0.938, F1-score 0.93으로 가장 우수한 성능을 보였으며, 특성 중요도에서는 교육, 경제, 의료, 문화적 특성 중 교육적 특성이 두드러졌다. 20대 후반 모델에서는 학습 정확도 0.852, 검증 정확도 0.750, F1-score 0.75의 성능을 보였으며, 특성 중요도에서는 경제적, 행정적 요소가 두드려졌다. 30대 초반 모델에서는 학습 정확도 0.773, 검증 정확도 0.625, F1-score 0.61으로 다소 낮은 정확도를 보였다. 해당 모델의 특성 중요도에서는 경제보다 삶의 질과 관련된 문화, 행정적 요소가 두드러졌다. 20대 후반과 30대 초반 모델의 정확도가 아쉬우나, 이는 경제적 요소 중 취업과 고용에 민감한 세대인 만큼, 해당 데이터를 반영하지 못하여 나온 결과로 보인다. 추후 취업 및 고용 데이터를 추가적으로 반영한다면 정확도가 높아질 것이라 예상된다.

종합적으로 청년 인구 유출 요인을 분석하면, 경제 및 문화, 교육, 공공행정 시설은 청년 유출에 영향을 미치는 것으로 보인다. 따라서 모든 연령대에서 등장한 요소들과 각 연령별로 주요하게 여겨지는 요소들에 따라 해당하는 시설을 보완 및 확충할 필요가 있다. 이를 통해 모든 연령 및 구에 일률적으로 적용되었던 청년 정책을 해당하는 나이와 거주지에 따라 맞춤형으로 제공한다면, 기반 시설 확충 및 지원 사업에 필요한 예산을 적재적소에 활용할 수 있을 것이라 기대한다.

그러나 데이터의 크기가 크지 않아 증폭을 시도하였으며, 청년 유출의 주요 원인으로 꼽히는 고용 및 취업률 데이터를 반영하지 못하였다는 한계가 있다. 단순 사업체 수나 비중에서 나아가 고용 및 취업률 데이터를 반영하여 모델에 적용한다면, 비교적 낮은 정확도가 나온 30대 후반 모델의 정확도를 높일 수 있을 거라 예상한다.

References

[1] Audit Office, "Population Structure Changes and Measures Taken", 2021, https://www.bai.go.kr/bai/result/branch/detail?srno=2622

[2] Moon, Young-man, "Determinants of outflow of youth fromnon-metropolitan areas to the metropolitan area," Journal of Regional Studies, Vol. 30, No.4, pp 55-77, 2022, DOI:10.31324/JRS.2022.12.30.4.55

[3] Woo-Sik Park, Sang-Woo Park, Chang-Ok Um."An Analysis of the Deterioration of Economic Capacity by Local Brain Drain - A Case Study of Daegu-Gyeongbuk Area," Journal of Industrial Economics and Business, Vol. 24, No.4, pp 2247-2274, 2011, DOI: 10.17703/JCCT.2023.9.3.129

[4] Statistics Korea, Domestic Population Migration Statistics, 2022.

[5] Statistics Korea, Population Trend Survey, 2022

[6] Byung-Soo kang, "Research on the Relationship between Regional Amenities and Residential Relocation", Journal of The Korean UrbanManagement Association, pp 329-351, 2014

[7] Choi, Eugene, "Urban Amenities as Economic Engine: Empirical Research on Amenity Effects in Korean Municipality", The Korean Journal of Local Government Studies, Vol 20, No.4, pp 299-324, 2017, DOI:10.20484/klog.20.4.13

[8] Lee chanyoung, “An Analysis on the Determinants of Youth Population Movement across Regions and Prospects", Economic Research, Vol.34, no.4, pp 143-169, 2016.

※ 이 논문은 서울여자대학교 학술연구비의 지원에 의한 것임 (2023-0007).