J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr


# 빅데이터 표준분석모델을 활용한 초등돌봄 수요예측 사례연구☆

# The Case Study for Childcare Service Demand Forecasting Using Bigdata Reference Analysis Model

윤 충 식[1] 정 승 렬[1*]
Chung-Sik Yun Seung Ryul Jeong

**요 약**

행정이 고도의 전문성과 설득력을 갖추기 위해서는 행정 영역에서 '빅데이터'를 활용하고 이러한 과학적 근거에 기반하여 정책의 수립·집행·평가가 이뤄져야 한다는 관점에서 데이터기반 행정에 대한 시대적 요구가 높아지고 있다. 본 연구는 신규 공동주택단지의 초등돌봄 수요예측을 위해 지역의 특성을 기계학습 기반으로 분석·예측하였다. 이를 위해 전용면적, 세대당 주차대수, 건폐율 등 아파트의 구조와 관련된 데이터, 초등학교까지의 거리 등 아파트 주변의 환경 데이터 및 행정구역의 인구 데이터 등 총 292종의 변수가 활용되었다. 다양한 변수의 활용에 큰 의의가 있으며 복합적인 분석에도 의미가 있다. 또한 실제 기초 지방자치단체의 실제값과 비교를 통해서 모델의 신뢰성을 높인 실증기반 사례연구이다.

☞ 주제어 : 빅데이터, 데이터기반 행정, 표준분석모델, 초등돌봄 수요예측, 행정 효율성

**ABSTRACT**

This paper is an empirical analysis as a reference model that can predict up to the maximum number of elementary school student care needs in local governments across the country. This study analyzed and predicted the characteristics of the region based on machine learning to predict the demand for elementary care in a new apartment complex. For this purpose, a total of 292 variables were used, including data related to apartment structure, such as number of parking spaces per household, and building-to-land ratio, environmental data around apartments such as distance to elementary schools, and population data of administrative districts. The use of various variables is of great significance, and it is meaningful in complex analysis. It is also an empirical case study that increased the reliability of the model through comparison with the actual value of the basic local government.

☞ keyword : Bigdata, Reference Analysis Model, effectiveness

## 1. 서 론

디지털 전환은 우리 사회 전반과 인류의 삶에 중대한 변화를 일으킨다. 행정 영역에서도 지능정보기술을 활용하려는 시도가 활발하다. 행정에 지능정보기술을 활용할 경우, 급부행정에서의 사각지대를 개선하고 복지 수혜의 폭을 넓히기 위해 데이터를 활용한 적극 행정을 도모할 수 있다[1]. 이에 따라, 공공행정 분야에서도 경험과 직관에 따라 정책을 수립하는 방식에서 데이터를 근거로 객관적이고 과학적인 정책 수립과 의사 결정을 추진하는 '데이터기반 행정'으로의 변화에 대한 시대적 요구가 높아지고 있다.

정부에서도 이러한 시대적 요구에 부응하기 위해 2020년 데이터기반 행정법을 제정·시행하고 데이터기반 과학적 행정이 모든 행정·공공기관에 정착될 수 있도록 다각적인 노력을 기울이고 있다. 특히, 행정안전부에서는 공공기관의 우수한 빅데이터 분석과제를 지속해서 발굴하고, 우수한 빅데이터 분석모델을 표준화함으로써 유사한 분석모델 개발을 최소화하고, 분석예산을 절감하기 위한 목적으로 2016년부터 분야별 '표준분석모델 정립 사업'을 추진하고 있다[2].

특히, 빅데이터 분석 결과를 바탕으로 전기차 충전소·공공 와이파이 설치지역·국공립 어린이집 위치를 선정하고[3] 무인민원발급기 설치지역 선정, 지진 대피소 위치

1 Graduate School of Business IT, Kookmin University, Seoul, 136-702, Korea.
* Corresponding author (srjeong@kookmin.ac.kr)
[Received 30 October 2022, Reviewed 2 November 2022(R2 2 December 2022), Accepted 5 December 2022]
☆ A preliminary version of this paper was presented at ICONI 2022
☆ 본 논문은 교육부 및 한국연구재단의 4단계 두뇌한국21사업(4단계 BK21 사업)으로 지원된 연구임

평가[4] 및 공공시설물 최적입지 선정[5] 등 데이터 기반 과학적 의사 결정에 활용하고 있다.

한편, 국토교통부는 2020년 7월, 신규아파트 단지의 주민공동시설에 온종일 돌봄 시설인 '다함께돌봄센터'* 설치를 의무화하겠다는 입법예고를 발표하였다[6]. 그러나 돌봄이 필요한 초등학생 수를 사전에 정확히 예측해 적정 규모의 돌봄 시설을 만들기는 쉽지 않은 일이었다.

행정안전부는 2020년 신규 공동주택의 초등돌봄 수요를 예측하기 위한 분석모델을 개발하기 위해 우선 시범지역을 대상으로 연구를 수행한 후, 표준분석모델 정립 사업을 통해 기초단체에서 활용하기 위한 모델의 정립 및 확산을 추진하였다. 본 논문은 시범지역인 기초단체 C시에 적용된 예측모델의 실증사례이며 그 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고, 3장에서는 초등돌봄 수요예측 모델의 개발 및 검증과정을 기술하였다. 4장에서는 표준분석모델 적용사례를 살펴보고, 끝으로 마지막 장에서는 본 논문의 공헌과 한계, 그리고 미래의 연구를 위한 방향을 제시한다.

## 2. 관련 연구

본 연구에서는 공공시설물 중의 하나인 초등돌봄센터의 규모 산정과 관련된 돌봄수요를 예측하는 데 인구, 공동주택, 공공시설물과 관련된 지역의 특성을 공동주택단지를 중심으로 분석하고자 하는 것이다. 선행연구를 통해 분석에 필요한 내용에 대해서 살펴보면 다음과 같다.

행정서비스 수요와 거주인구의 변화와의 관계를 분석한 사례[7]에 따르면, 지역주민 구성 변화에 영향을 미치는 대표요인은 거주민들의 전·출입에 따른 인구이동이며, 실제 민원 데이터를 분석한 결과 차이가 있었다. 그리고 노인요양시설과 복지시설의 수요를 예측한 사례에서는 미래시점 인구를 예측하기 위하여 인구총조사를 바탕으로 조성법(Cohort-Component Method)을 활용하여 미래시점의 성별, 연령별, 지역별 예측을 시도하였다. 그리고 공공자전거의 수요예측에 관한 연구는 다수가 있으나, 최근 사례로 서울시 공공자전거 수요예측 모형의 비교 분석사례가 있다. 과거 1년간 대여소별 데이터를 바탕으로 상관성을 가정한 벡터 자기 회귀모형(VAR)과 서포트 벡터 회귀모형(SVR), 딥러닝 기법인 LSTM(Long short-term memory networks) 모형 등을 비교 분석한 결과, LSTM이 다른 기법에 비해 뛰어난 성능을 보였으며 SVR 모형에서는 시계열 군집을 기반으로 최적의 모형을 만들면 예측력이 높게 나올 수 있음을 보였다[8].

그리고 보육 및 아이돌봄서비스 수요를 예측한 연구로는 유치원과 어린이집 수요 예측하기 위해서 지역별 연령별 유치원과 어린이집 이용률 추이를 분석하여 시계열 모형을 제시한 사례[9]와 전국단위의 읍·면·동별 돌봄수요 예측을 위하여 인구주택총조사의 2% 마이크로데이터에서 초등학생 자녀가 있는 가구를 추출하고, 다시 맞벌이 가구를 추출한 후 돌봄 유형을 분석한 사례가 있다[10].

한편, 돌봄수요는 맞벌이 가정이나 편부모 가정의 초등학생 자녀의 수로 결정할 수 있으나, 이에 대한 정확한 통계를 대신하여, 통계청(2018) 「2018 일·가정 양립지표」에서 제시한 초등학생 자녀를 둔 모(母)의 고용율을 전체 초등학생 수에 곱하여 산정하기도 하였다[11].

선행연구를 통한 시사점으로 공공시설물 수요는 지역의 인구요인이 무엇보다 중요한 요소라는 것과 과거 시설물 이용자 수, 이용률 기반의 시계열 데이터가 필요하다는 것을 알 수 있었다. 다만, 공공시설물 수요자 관련 거주지역의 인구 특성, 생활편의 특성, 교육 특성 등을 분석하여 미래 수요를 예측한 사례가 부족하여 이에 관한 연구의 필요성도 인식할 수 있었다.

## 3. 초등돌봄 수요예측 모델

신규 공동주택 단지에 필요한 초등돌봄 수요를 예측하기 위하여 선행연구에서 파악한 요인과 수집 가능한 공공데이터를 수집하여 전처리 후 입주 시점의 초등학생 수를 예측하기 위한 모델을 생성하였다. 해당 단지의 초등돌봄 수요는 예측한 초등학생 수에 통계청에서 생산한 해당 지역 맞벌이 비율을 곱하여 산정하였다.

### 3.1 데이터 수집

모델에 필요한 데이터는 통계청, 행정안전부, K-APT 공동주택관리 시스템, 교육통계 및 국가도서관 통계 등에서 공공데이터 수집하였다. 공간적 범위는 기초단체인 C시, 시간적 범위는 2019년 12월 말을 기준으로 과거 10년간 그리고 공동주택 관련 데이터는 준공년도를 기준으로 준공 전 과거 3년간 자료를 수집하였다. 수집 데이터의 목록 및 주요 속성은 Table 1과 같다.

* 초등학교의 정규교육 이외의 시간 동안 돌봄서비스를 제공하기 위해 시·도지사 및 시장·군수·구청장이 설치·운영하는 시설 「아동복지법」 제44조2의 1항)로 초등돌봄 사각지대 해소 및 육아 부담 경감을 위해 보편적 돌봄서비스 제공을 위해 설치한 센터

(Table 1) Collection Data List and Key Attribute

| 구분 | 데이터 셋 | 주요속성 | 출처 |
|---|---|---|---|
| 인구 | 국내인구 이동통계 | 전입·전출일, 연령·세대별 전입·전출자 등 | 통계청 |
| | 행정동 인구통계 | 인구수 및 초등학생 수 | 행안부 |
| | 우편번호 코드 | 전국 시군구/읍면동 우편번호 코드 | 행안부 |
| 공동주택 | 공동주택 기본정보 | 공동주택명, 주소, 준공년도, 용적율, 건폐율, 총 세대수 등 | K-apt 네이버 부동산 |
| | 편의정보 | 단지 내 독서실, 어린이집 설치여부 등 | |
| 공공시설물 | 초등학교 | 학교명, 주소 등 | 교육통계 서비스 |
| | 도서관 | 도서관 명, 주소, 운영시간 | 국가도서관 통계시스템 |
| | 병원 | 병원 주소 | 네이버 |
| 기타 | 가계동향조사 | 맞벌이 비율 | 통계청 |

## 3.2 데이터 전처리

전처리의 첫 번째는 공동주택 단지의 소재지 기준 행정동별 집계와 단지별 집계 처리 단계이다. 26개 행정동을 기준으로 동별인구, 동별 초등학생 인구, 시내·시외·연령별 전입·전출자 수를 집계하였다. 그리고 79개 공동주택 단지의 용적률, 건폐율, 총세대수 등의 기본정보를 바탕으로 평균면적, 평균방수, 평균 욕실 수 등의 부가 정보를 생성한다. 두 번째 단계는 분석을 위한 추가 파생변수를 생성하는 단계로 공동주택 준공년도를 기준으로 전년도와 전전년도의 인구수, 전입·전출자 수와 각 년도별 인구수 차이, 전입·전출자 수 차이 등을 계산하여 변수로 생성하였다. 공동주택 단지로부터 공공시설물까지의 거리도 산정하고 단지 내 편의점, 헬스장 등 각종 편의시설 설치 여부 등의 파생 데이터를 생성하였다. 세 번째 단계는 공동주택 단지별로 사분위 수 면적($m^2$)과 각 사분위 수 면적의 비율을 산정하여 변수를 생성하였다.

최종적으로, Table 4와 같이 초등학생 수를 종속변수로 ① 행정동별 시내·외 전체 전출입 세대수 및 초등학생 나이별 전출입 학생 수, ② 행정동별 인구정보, ③ 공동주택(아파트단지)의 아파트 정보(79개 아파트단지) 등 292개 변수를 독립변수로 하는 분석용 데이터셋을 생성하였다.

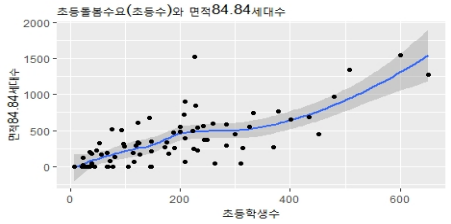
## 3.3 수요예측 모델생성

수요예측 모델 생성은 ①수집된 데이터에 대한 탐색적 분석, ② 독립변수와 종속변수 간의 상관관계 분석, 선형 회귀분석 그리고 ③ 랜덤포레스트 분석과 서포트벡터머신 분석 ④알고리즘 성능 비교의 순서로 진행하였다.

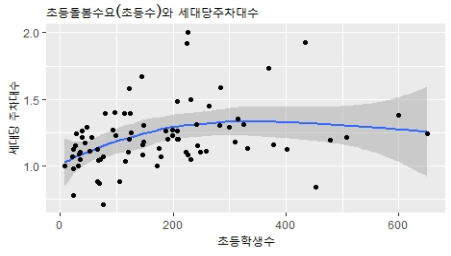
탐색적 데이터 분석 결과 기초단체 C시의 전체인구는 2020년 12월 말 기준으로 103만 명, 초등학생수는 5만9천여 명, 세대수는 422,097이고, 세대당 인구는 2.46명이다. 동별 평균 인구는 18,745명, 동별 평균 초등학생수는 1,266명이다. 준공년도를 기준으로 2010년부터 2020년 12월 말까지 수집된 공동주택 단지는 총 79개로 매년 평균 8.8개가 준공되었다.

### 3.3.1 상관관계 및 회귀분석

초등학생수와 아파트 면적이 84.84m2인 세대수의 대략적인 관계는 Figure 1과 같이 두 변수 간에는 정(+)의 상관관계가 존재하고 초등학생수와 세대당 주차대수의 관계는 Figure 2와 같이 정(+)의 상관관계가 존재한다.




(Figure 1) Scatter plot of the number of elementary school students and apartment area 84.84 households




(Figure 2) Scatter plot of the number of elementary school students and the number of cars parked per household

종속변수인 초등학생 수와 독립변수 292개 중 숫자 변수인 279개 변수 대상으로 상관관계 분석 결과는 Table 2와 같다. 95%의 신뢰구간에서 14개 변수가 유의한 결과

를 보였다. (P-Value < 0.05)

상관관계 분석 결과, 전용면적이 크고, 세대당 주차대수가 많을수록 초등학생 수가 증가하며, 아파트 건폐율이 클수록 초등학생 수가 감소하는 것으로 나타났다.

(Table 2) Correlation Analysis Results Table

| 번호 | 변수명 | 상관계수 | P-Value |
|---|---|---|---|
| 1 | 총세대수 | 0.785 | 0.000 |
| 2 | 면적84.84세대수* | 0.734 | 0.000 |
| 3 | 동수 | 0.516 | 0.000 |
| 4 | 최고층 | 0.455 | 0.000 |
| 5 | 면적84.84세대수_비율 | 0.34 | 0.002 |
| 6 | 평균면적 | 0.335 | 0.003 |
| 7 | 평균욕실수 | 0.329 | 0.003 |
| 8 | 면적73.2세대수_비율 | -0.327 | 0.003 |
| 9 | 세대당주차대수* | 0.307 | 0.006 |
| 10 | 평균방수 | 0.3 | 0.007 |
| 11 | 면적73.2_80.46세대수 | 0.28 | 0.012 |
| 12 | 최저층 | 0.252 | 0.025 |
| 13 | 건폐율* | -0.245 | 0.03 |
| 14 | 전출_시외_11_차이_1 | -0.23 | 0.041 |

* 상관계수 및 P-Value(0.05 미만)

종속변수인 초등학생 수와 상관관계 상 유의한 결과가 도출된 14개 변수를 기준으로 다중 선형 회귀분석 결과, Table 3과 같이 “총 세대수”와 “면적84.84세대수”는 “초등학생 수”와 선형적 관련성을 갖지만, 나머지 12개의 회귀계수는 통계적으로 유의하지 않았다. 먼저, F값을 보면 대응되는 p-값이 유의수준 0.05에 비해 매우 작으므로 다중회귀모델의 회귀식은 통계적으로 유의하다. 개별 회귀계수의 유의성을 살펴보면, “총세대수”와 “면적84.84세대수”의 회귀계수는 유의수준 0.05에서 통계적으로 유의하며, 따라서 이 두 변수의 회귀계수는 0이라고 볼 수 없다, 그리고 회귀모델의 예측변수들은 초등학생 수의 변동성의 74%를 설명한다. (R2 = 0.74, 수정된 R2 = 0.69)

C시의 아파트단지별 초등학생 수(Y)와 단지별 총세대수(X1), 면적84.84세대수(X2)와의 회귀식은 아래와 같다.

$$Y = 0.433849729x\ X_1 + 0.472333043x\ X_2$$

(Table 3) Regression Results Table

| 번호 | 변수명 | Estimate | Std. |
|---|---|---|---|
| 1 | 총세대수 | 0.1** | 0.1 |
| 2 | 면적84.84세대수 | 0.2** | 0.1 |
| 3 | 동수 | 1.8 | 4.2 |
| 4 | 최고층 | 1.4 | 2.8 |
| 5 | 면적84.84세대수_비율 | -0.4 | 0.6 |
| 6 | 평균면적 | -2.5 | 2.0 |
| 7 | 면적73.2세대수_비율 | -1 | 0.6 |
| 8 | 평균욕실수 | 129.0* | 75.6 |
| 9 | 세대당주차대수 | -28 | 67.6 |
| 10 | 평균방수 | -27.6 | 37.7 |
| 11 | 면적73.2_80.46세대수 | 0.04 | 0.1 |
| 12 | 최저층 | 0.4 | 2.3 |
| 13 | 건폐율 | 0.3 | 0.8 |
| 14 | 전출_시외_11_차이_1 | 1 | 2.3 |
| Constant | | 71.7 | 128.0 |
| Observations | | 79 | |
| R2 | | 0.74 | |
| Adjusted R2 | | 0.69 | |
| Residual Std. Error | | 78.3 (df=64) | |
| F Statistic | | 13.3***(df=14, 64) | |
| Note: *p < 0.1: **p < 0.05 ; ***p < 0.01 | | | |

### 3.3.2 서포트벡터머신 분석

서포트벡터머신(Support Vector Machine) 분석을 위해 먼저 준공년도를 기준으로 2019년 이전 데이터(74건)를 모델생성을 위한 데이터로 2019년 이후 데이터(5건)를 테스트용 데이터 분리하였다. 그리고 모델 생성용 데이터는 8:2의 비율로 학습용과 검증용으로 분리하였다.

그리고 예측모델의 해석을 용이하게 하도록 292개 독립변수를 변수별 특성을 기준으로 Table 5와 같이 5개의 변수 군으로 분류하고, 2개 이상의 변수 군으로 생성할 수 있는 변수 군 조합 15가지(Feature Set)를 Table 6과 같이 정의하였다.

(Table 5) Variable group List

| 변수군 | 변수 | 변수 수 |
|---|---|---|
| Ⓐ | 전출입 세대수 정보 | 30 |
| Ⓑ | 전출입 초등학생 나이별 인구정보 | 180 |
| Ⓒ | 전출입 초등학생 인구정보 | 30 |
| Ⓓ | 동별 인구정보 | 15 |
| Ⓔ | 공동주택단지 정보 | 37 |
| 계 | | 292 |

그리고 종속변수를 초등학생 수, 독립변수는 1번~15번 Feature Set 별로 분류한 후 서포트백터머신 모델을 생성하였다.

(Table 4) Master data set for analysis

| 변수명 | | | | | No | 변수 설명 |
|---|---|---|---|---|---|---|
| Key | | ID | | | | 공동주택 고유번호 |
| | | 건물명 | | | | 공동주택 건물명 |
| | | 주소 | | | | 주소 |
| | | 준공연도 | | | | 준공연도 |
| | | 초등학교명 | | | | 공동주택 지정 초등학교명 |
| Y* | | 초등학생수 | | | 1 | 공동주택 단지 내 초등학생 수(6세~11세 인구수) |
| X** | 공동주택단지 정보Ⓔ | 거리 | | | 2 | 단지 중심으로부터 지정 초등학교까지의 거리(단위:m) |
| | | 세대당주차대수 | | | 3 | 단지 세대별 주차대수 |
| | | 용적률/ 건폐율 | | | 4 | 공동주택 용적률, 건폐율 |
| | | 평균면적 | | | 6 | 공동주택 평균면적 |
| | | 총세대수 | | | 7 | 공동주택 단지별 세대수 |
| | | 면적73.2세대수 | | | 8 | 단지 내 면적 73.2세대수(4분위 수) |
| | | 면적73.2_80.46세대수 | | | 9 | 단지 내 면적 73.2_80.46세대수(4분위 수) |
| | | 면적80.46_84.84세대수 | | | 10 | 단지 내 면적 80.46_84.84세대수(4분위 수) |
| | | 면적84.84세대수 | | | 11 | 단지 내 면적 84.84세대수(4분위 수) |
| | | 면적73.2세대수_비율 | | | 12 | 단지 내 면적 73.2세대수_비율(4분위 수) |
| | | 면적73.2_80.46세대수_비율 | | | 13 | 단지 내 면적 73.2_80.46세대수_비율(4분위 수) |
| | | 면적80.46_84.84세대수_비율 | | | 14 | 단지 내 면적 80.46_84.84세대수_비율(4분위 수) |
| | | 면적84.84세대수_비율 | | | 15 | 단지 내 면적 84.84세대수_비율(4분위 수) |
| | | 동수/최저층/최고층 | | | | 단지 전체 동수, 최고층 수, 최저층 수 |
| | | 평균방수/ 평균욕실수 | | | | 단지 평균 방수, 욕실 수 |
| | | 종합병원까지의거리 | | | 21 | 단지 중심으로부터 종합병원까지의 거리(m) |
| | | 대학병원까지의거리 | | | 22 | 단지 중심으로부터 대학병원까지의 거리(m) |
| | | 대형마트까지의거리 | | | 23 | 단지 중심으로부터 대형마트까지의 거리(m) |
| | | 도서관까지의거리 | | | 24 | 단지 중심으로부터 도서관까지의 거리(m) |
| | | 중학교까지의거리 | | | 25 | 단지 중심으로부터 중학교까지의 거리(m) |
| | | 난방방식/ 난방종류 | | | | 단지의 난방방식(도시가스·열병합), 종류(지역·개별난방 등) |
| | | 현관구조 | | | 28 | 단지의 현관 구조(계단식, 복도식 등) |
| | | 마트/ 편의점 | | | 29 | 단지 내 마트·편의점 유무 (1:있음, 0 없음) |
| | | 어린이집 /헬스장 | | | 31 | 단지 내 어린이집·헬스장 유무 (1:있음, 0 없음) |
| | | 실내골프장 | | | 33 | 단지 내 실내골프장 유무 (1:있음, 0 없음) |
| | | 배드민턴장 | | | 34 | 단지 내 배드민턴장 유무 (1:있음, 0 없음) |
| | | 무인택배함 | | | 35 | 단지 내 무인택배함 유무 (1:있음, 0 없음) |
| | | 독서실 / 카페 | | | 36 | 단지 내 독서실·카페 유무 (1:있음, 0 없음) |
| | | 농구장 | | | 38 | 단지 내 농구장 유무 (1:있음, 0 없음) |
| | 전출입 세대수 정보Ⓐ | 기준연도 | 전입 | 전체수 | 39 | 읍면동 전체 전입자 수(6세~11세) |
| | | | | 시내수 | 40 | 동일시도 내 읍면동 경계를 넘어선 전입자 수(6세~11세) |
| | | | | 시외수 | 41 | 시도 경계를 넘어선 읍면동 전입자 수(6세~11세) |
| | | | 전출(전체수, 시내수, 시외수) | | | |
| | | 1년 전 전입(전체/시내/시외), 전출(전체/시내/시외) | | | | |
| | | 2년 전 전입(전체/시내/시외), 전출(전체/시내/시외) | | | | |
| | | 기준연도 | 전입 | 전체수차이 | 57 | 기준연도와 1년전 읍면동 전체 전입자 수 차이(6세~11세) |
| | | | | 시내수차이 | 58 | 동일 시도내 읍면동 경계를 넘어선 전입자 수 차이(6세~11세) |
| | | | | 시외수차이 | 59 | 시도 경계를 넘어선 읍면동 전입자 수 차이(6세~11세) |
| | | | 전출(전체수 차이, 시내수 차이, 시외수 차이) | | | |

| 변수명 | | | | | No | 변수 설명 |
|---|---|---|---|---|---|---|
| | | 1년 전 전입(전체/시내/시외), 전출(전체/시내/시외) 차이(6세~11세) | | | | |
| | 전출입 초등학생 나이별 인구정보Ⓑ | 기준년도 | 전입 | 6세 전체수 | 69 | 기준연도 읍면동 기준 6세 전입자 수 |
| | | | | 6세 시내수 | 70 | 기준연도 동일시도 내 읍면동 경계를 넘어선 6세 전입자 수 |
| | | | | 6세 시외수 | 71 | 기준연도 시도 경계를 넘어선 6세 전입자 수 |
| | | | 전출(6세 전체수, 6세 시내수, 6세 시외수) | | | |
| | | | 7세~10세 전입(전체/시내/시외), 전출(전체/시내/시외) | | | |
| | | | 전입 | 11세 전체수 | 99 | 기준연도 읍면동 기준 11세 전입자 수 |
| | | | | 11세 시내수 | 100 | 기준연도 동일시도 내 읍면동 경계를 넘어선 11세 전입자 수 |
| | | | | 11세 시외수 | 101 | 기준연도 시도 경계를 넘어선 11세 전입자 수 |
| | | | 전출(11세 전체수, 11세 시내수, 11세 시외수) | | | |
| | | 1년 전 전입(전체/시내/시외), 전출(전체/시내/시외) (6세~11세) | | | | |
| | | 2년 전 전입(전체/시내/시외), 전출(전체/시내/시외) (6세~11세) | | | | |
| | | 기준연도 | 전입 | 6세 전체 차이 | 177 | 기준연도와 1년전 읍면동 전체 6세 전입자 수 차이 |
| | | | | 6세 시내 차이 | 178 | 동일시도 내 읍면동 경계를 넘어선 6세 전입자 수 차이 |
| | | | | 6세 시외 차이 | 179 | 시도 경계를 넘어선 읍면동 6세 전입자 수 차이 |
| | | | 전출(6세 전체수 차이, 6세 시내수 차이, 6세 시외수 차이) | | | |
| | | | 7세~10세 전입(전체/시내/시외), 전출(전체/시내/시외) 차이 | | | |
| | | | 전입 | 11세 전체 차이 | 207 | 기준연도와 1년전 읍면동 전체 11세 전입자 수 차이 |
| | | | | 11세 시내 차이 | 208 | 동일시도 내 읍면동 경계를 넘어선 11세 전입자 수 차이 |
| | | | | 11세 시외 차이 | 209 | 시도 경계를 넘어선 읍면동 11세 전입자 수 차이 |
| | | | 전출(11세 전체수 차이, 11세 시내수 차이, 11세 시외수 차이) | | | |
| | | 1년 전 전입(전체/시내/시외), 전출(전체/시내/시외) 차이(6세~11세) | | | | |
| | 전출입초등학생 인구정보Ⓒ | 기준연도 | 전입 | 전체수 | 249 | 기준연도 읍면동 기준 6세~11세 전입자 수 |
| | | | | 시내수 | 250 | 동일시도 내 읍면동 경계를 넘어선 6세~11세 전입자 수 |
| | | | | 시외수 | 251 | 기준연도 시도 경계를 넘어선 6세~11세 전입자 수 |
| | | | 전출(전체수, 시내수, 시외수) | | | |
| | | 1년 전 | | | | |
| | | 2년 전 | | | | |
| | | 기준연도 | 전입 | 전체수차이 | 267 | 기준연도와 1년전 읍면동 기준 전입자 수 차이(6세~11세) |
| | | | | 시내수차이 | 268 | 동일시도 내 읍면동 경계를 넘어선 전입자 수 차이(6세~11세) |
| | | | | 시외수차이 | 269 | 시도 경계를 넘어선 전입자 수 차이(6세~11세) |
| | | | 전출(전체수 차이, 시내수 차이, 시외수 차이) | | | |
| | | 1년 전 | | | | |
| | 동별인구 정보Ⓓ | 기준연도 | 전체 인구수 | | 279 | 기준연도 행정동 인구수 |
| | | | 초등학생 인구수 | | 280 | 기준연도 행정동 인구수(6세~11세) |
| | | | 초등학생 인구비율 | | 281 | 기준연도 행정동 인구 비율(6세~11세) |
| | | 1년 전 전체 인구수, 초등학생 인구수, 초등학생 인구비율 | | | | |
| | | 2년 전 전체 인구수, 초등학생 인구수, 초등학생 인구비율 | | | | |
| | | 기준연도 | 전체 인구수 차이 | | 288 | 기준연도와 1년전 행정동 인구수 차이 |
| | | 1년 전 | 인구수 차이 | | 289 | 기준연도와 1년전 행정동 6세~11세 인구수 차이 |
| | | | 인구수 차이 비율 | | 290 | 기준연도와 1년전 행정동 6세~11세 인구수 차이 비율 |
| | | 2년 전 | 전체 인구수 차이 | | 291 | 1년전과 2년전 행정동 인구수 차이 |
| | | | 인구수 차이 | | 292 | 1년전과 2년전 행정동 6세~11세 인구수 차이 |
| | | | 인구수 차이 비율 | | 293 | 1년전과 2년전 행정동 6세~11세 인구수 비율 |

*Y는 종속변수, ** X는 독립변수

(Table 6) Feature Set List by Combination by Variable Group

| 구분 | 변수군 조합 | 변수 수 |
|---|---|---|
| 1번 Feature Set | Ⓐ+Ⓑ+Ⓒ+Ⓓ+Ⓔ | 292 |
| 2번 Feature Set | Ⓐ +Ⓒ+Ⓓ+Ⓔ | 112 |
| 3번 Feature Set | Ⓐ+Ⓑ +Ⓓ+Ⓔ | 262 |
| 4번 Feature Set | Ⓐ+Ⓑ+Ⓒ +Ⓔ | 277 |
| 5번 Feature Set | Ⓑ+Ⓒ+Ⓓ+Ⓔ | 262 |
| 6번 Feature Set | Ⓐ+Ⓑ +Ⓔ | 247 |
| 7번 Feature Set | Ⓐ +Ⓒ +Ⓔ | 97 |
| 8번 Feature Set | Ⓐ +Ⓓ+Ⓔ | 82 |
| 9번 Feature Set | Ⓑ+Ⓒ +Ⓔ | 247 |
| 10번 Feature Set | Ⓑ +Ⓓ+Ⓔ | 232 |
| 11번 Feature Set | Ⓒ+Ⓓ+Ⓔ | 82 |
| 12번 Feature Set | Ⓐ +Ⓔ | 67 |
| 13번 Feature Set | Ⓑ +Ⓔ | 217 |
| 14번 Feature Set | Ⓒ +Ⓔ | 67 |
| 15번 Feature Set | Ⓓ+Ⓔ | 52 |

모델 생성 시 옵션으로 Epsilon= 0.001, 0.01. 0.1, 1의 4가지와 Cost= 0.01, 0.1. 1.10. 100의 5가지 경우를 각각 적용하되 임의 복원추출을 100번 반복한 후 평균 제곱근의 오차(RMSE*) 평균을 바탕으로 총 300개의 RMSE 값을 비교하였다. 비교 결과, Table 7에서 보는 바와 같이 동별 인구변수(15개)와 공동주택 단지 정보 변수(37개)를 독립변수로 하고 Epsilon = 0.001, Cost = 100의 경우 RMSE 값이 가장 적은 결과를 보였다.

(Table 7) Result of SVM Model

| 번호 | Feature Set | Epsilon | Cost | 평균 RMSE |
|---|---|---|---|---|
| 1 | 15번 | 0.001 | 100 | 75.02862184 |
| 2 | 15번 | 0.001 | 10 | 75.04473021 |
| 3 | 15번 | 0.01 | 100 | 75.09482541 |
| .. | ... | ... | ... | ... |
| 300 | 1번 | 0.001 | 0.01 | 141.7650714 |

### 3.3.3 랜덤포레스트(Random Forest) 분석

서포트벡터머신 분석 절차와 같이 준공년도를 기준으로 모델 생성과 테스트데이터로 분리하고, 테스트데이터를 8:2의 비율로 학습용과 검증용으로 분리한 후 앞에서 정의한 5개 변수 군의 변수 군별 조합 15가지 Feature Set을 다시 활용하였다. 모델은 Table 8에서와 같이 mtree는 5가지, ntree는 10가지 옵션을 각각 적용하여 랜덤포레스트 모델을 생성하였다. 임의 복원추출을 100번 반복한 후 평균 제곱근의 오차(RMSE) 평균을 바탕으로 총 750개의 RMSE 값을 비교하였다.

(Table 8) Random Forest Model's Parameter

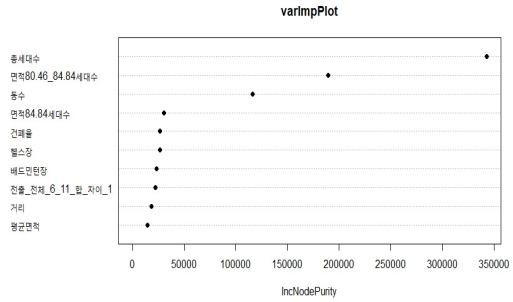
| 번호 | 구분 | 내용 |
|---|---|---|
| 1 | 변수 조합 (Feature Set) | 1번 ~15번 |
| 2 | 임의 복원추출 | 100번 |
| 3 | mtree 옵션 (5가지) | 학습용 데이터 셋의 변수 개수 -1) / i 단, i = 1, 2, 3, 4, 5 |
| 4 | ntree 옵션 (10가지) | 100, 200, ~ 900, 1000 |

비교 결과, Table 9에서 보는 바와 같이 전출입 초등학생 인구변수(30개)와 공동주택 단지 정보 변수(37개)를 독립변수로 하고 옵션으로 mtree = 34, ntreet = 900을 적용한 결과가 가장 높게 평가되었다.

(Table 9) Result of RF Model

| 번호 | Feature Set | mtree | ntree | 평균 RMSE |
|---|---|---|---|---|
| 1 | 14번 | 34 | 900 | 86.53589 |
| 2 | 14번 | 34 | 800 | 86.5537 |
| .. | ... | ... | ... | ... |
| 750 | 1번 | 58 | 100 | 100.2811 |

변수중요도 분석 결과 Figure 3과 같이 단지의 총세대수, 단지 내 면적 80.46_84.84세대수, 단지 전체 동수, 면적 84.84세대수, 건폐율, 헬스장 등이 주요 변수로 나타났다.




(Figure 3) Evaluate the importance of variables

## 3.4 모델 비교 및 검증

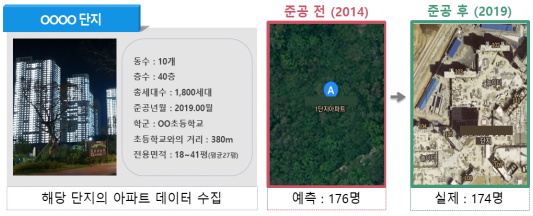
초등돌봄 수요 예측분석은 해당 아파트가 위치한 지역의 특성을 활용하고 관련 데이터에 기반한 알고리즘으로 분석하여 돌봄이 필요한 초등학생 수를 예측하는 것이다.

* 평균 제곱근 편차 또는 평균 제곱근 오차는 추정 값 또는 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 흔히 사용하는 측도

이를 위해 전용면적, 세대당 주차대수, 건폐율 등 아파트의 구조와 관련된 데이터, 초등학교까지의 거리 등 아파트 주변의 환경 데이터 및 행정구역의 인구 데이터 등 총 292개의 변수가 활용되었다.

Random Forest, SVM(Support Vector Machine) 등의 알고리즘을 적용하고 분석한 결과, 알고리즘은 초등학교와의 거리 등 초등학생 아이의 보육환경에 중요한 특성을 자동으로 파악하여 아파트단지의 초등돌봄 수요를 예측할 수 있었으며, SVM이 Random Forest보다 우수한 성능을 보였다.

그리고, C시 OOO 아파트단지의 준공 전 예측 초등학생 수와 준공 후 초등학생 수를 비교해본 결과 Figure 4에서와 보는 바와 같이 예측은 176명 실제값은 174명으로 의미 있는 검증 결과를 보였다.




(Figure 4) Example of comparing predicted and actual values

## 4. 표준분석모델 정립 및 활용사례

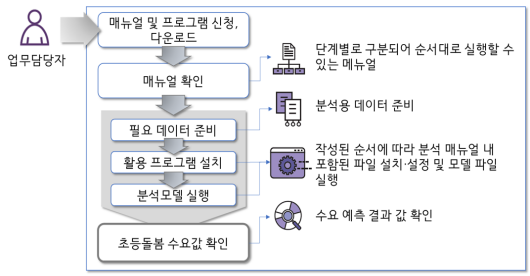
빅데이터 표준분석모델이란, 공공분야 유사 중복분석을 방지하기 위해 공공기관에서 수행한 우수한 빅데이터 분석모델의 활용 데이터, 절차, 기법 등을 표준화해 정립한 참조모델을 의미한다.*

행정안전부에서는 표준분석모델을 통해 수집 데이터 목록, 데이터 형식, 수급 절차 등을 표준화하여 자치단체별로 서로 다른 데이터와 분석모델 및 분석 결과의 차이점을 극복하고 분석의 효율성과 정확성을 향상할 수 있었다[12].

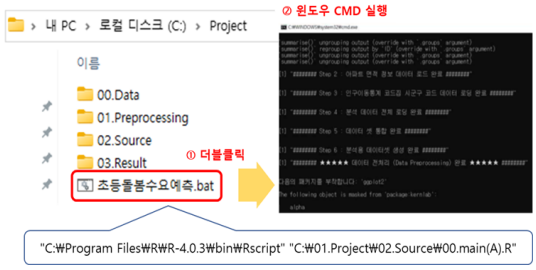
행정안전부는 초등돌봄 수요예측 표준분석모델은 표준화 과정을 통해서 참조모델로 구축한 후 매뉴얼, 분석코드, 샘플 데이터 등을 지자체에서 쉽게 활용할 수 있는 형태로 패키징하여 빅데이터 공통기반인 혜안을 통해서 배포하였다.

혜안은 공공·민간의 데이터를 연계·수집·저장 분석하고 분석 결과를 공유·활용하는 범정부 빅데이터 분석 시스템으로 민원, 복지, 환경, 교통, 보건의료, 문화관광 등 다양한 업무 분야에서 빅데이터를 활용하여 공공서비스 개선과 행정 효율성 향상의 효과를 거둘 수 있도록 다양한 서비스를 제공하고 있다.

수요예측 모델을 활용하고자 하는 업무담당자는 Table 1에 필요한 데이터를 준비하고 매뉴얼에 따라 정의된 폴더에 분석 도구인 오픈소스 R 프로그램을 설치한 후, 준비된 파일을 저장한다. 그리고, 혜안에서 다운받은 패키징 파일을 매뉴얼에 따라 압축을 해제하여 저장한 후, 해제된 매크로로 파일을 실행하면, 최종 결과 파일에 지정된 폴더에 엑셀로 저장된다. Figure 5는 모델의 배포 및 활용 절차를 Figure 6은 저장된 매크로 파일 실행 결과를 예시로 표현한 것이다.




(Figure 5) Examples of reference model deployment and utilization procedures




(Figure 6) Macro File Execution Example

초등돌봄 수요예측 표준분석모델은 분석 도구인 R을 기본으로 활용하되 매크로를 통해서 자동으로 결과물이 엑셀 파일로 생성되게 함으로써 사용자는 R의 분석환경에 대한 이해 없이도 결과물을 활용할 수 있게 되었다.

## 5. 결 론

본 연구에서는 C시 신규 공동주택단지의 초등돌봄 수

* 조명의, "공공 빅데이터 표준분석모델 정립 추진", 테크월드뉴스, 2020년 9월 3일

요예측을 위해 해당 아파트가 위치한 지역의 특성을 활용하고 관련 데이터에 기반한 알고리즘으로 분석하여 돌봄이 필요한 초등학생 수를 예측하고자 하였다. 이를 위해 전용면적, 세대당 주차대수, 건폐율 등 아파트의 구조와 관련된 데이터, 초등학교까지의 거리 등 아파트 주변의 환경 데이터 및 행정구역의 인구 데이터 등 총 292종의 변수가 활용되었다.

C시 사례분석 결과, 전용면적이 크고 세대당 주차대수가 많을수록, 주변 종합병원 등 인프라가 갖추어진 환경일수록 아파트 거주 초등학생 수가 증가하며, 초등학교가 멀거나 아파트의 건폐율이 클수록 초등학생 수가 감소하는 것으로 나타났다. 또한, 초등학생 수 예측에 유의미한 변수를 활용하여 머신러닝 알고리즘을 적용하고 초등학교와의 거리 등 초등학생 아이의 보육환경에 중요한 특성을 자동으로 파악하여 아파트단지의 초등돌봄 수요를 예측하기 위한 참조모델인 표준분석모델에 대해 살펴보았다.

본 연구의 결과는 다음과 같은 시사점을 제시한다.

첫째, C시 지역의 입주 전 아파트·주변 환경 데이터, 인구 데이터를 통해서 예측한 초등학생 수와 입주 후 실제 초등학생 수와 상당 부분 일치성을 나타내서 표준분석모델의 정확성을 확인하였다. 둘째, 신속한 의사 결정이 가능해졌다. 지자체 공무원의 경우 표준분석모델을 활용함으로써 필요데이터 수집 후 입주 전 신규 공동주택단지의 초등학생 수를 신속하게 예측할 수 있게 됨으로써 신규 주택단지 내 돌봄센터 규모 산정이라든지 돌봄수요에 따른 돌봄 사각지대 파악이 쉬워졌다. 셋째, 빅데이터 표준분석모델을 활용함으로써 지자체는 데이터 획득 및 빅데이터 분석 도구와 분석과정에서 발생하는 인력, 재정, 시간 등의 기회비용을 대폭 절감할 수 있게 되었다.

본 연구는 데이터 기반 지능형 행정 업무 수행을 위한 융합형 연구과제로 실제 공공데이터를 활용하여 머신러닝 기반 초등돌봄 수요인 초등학생 수를 예측한 것이다. 그리고 실제 해당 지역의 실제 값과 비교를 통해서 모델의 신뢰성을 높인 실증기반 사례연구이다.

향후 공공행정 분야에서 행정서비스와 관련된 불확실한 미래 수요를 예측하는 등 주요 정책 의사 결정 과정에 참조모델로 활용된다면 정책 의사결정과정의 투명성을 높일 수 있을 것으로 보인다.

## 참고문헌(Reference)


[1] JW Seon, “Legal issues of Data-driven Administration”, Korea Administrative Law and Practice Association Administrative Law Journal Vol. 66, pp.107, 2021. https://doi.org/10.35979/alj.2021.11.65.107

[2] Ministry of the interior and safety, “2019 White Paper”, pp.522, 2020. https://www.mois.go.kr/

[3] Ministry of the Interior and Safety, ″Solving people’s life and local administration challenges with Big data.″,Korea Policy Briefing, 2018. https://www.korea.kr

[4] Ministry of the Interior and Safety ″Data supports forest fire prevention and revitalization of the local economy″, Korea Policy Briefing, 2019. https://www.korea.kr

[5] Ministry of the Interior and Safety, ″Selection of public facility locations through big data analysis″, Korea Policy Briefing, 2022. https://www.korea.kr

[6] Ministry of Health and Welfare, ″Partial revision of childcare business guide″, Korea Policy Briefing, 2022. https://www.korea.kr

[7] DS Jun, “Public service forecasting based on Public Data”, The Korean Association for Regional Information Society’s conference paper 2018, pp.89~100, 2018. https://snu.ac.kr

[8] SA Min, YS Jung, “Comparative study of prediction models for public bicycle demand in Seoul, Korean Data and Information Science Society”, Journal of the Korean Data & Information Science Society 2021, Vol.32, pp.585–592, 2021. https://doi.org/10.7465/jkdi.2021.32.3.585

[9] JA Park, CH Park, JW Eum, “Mid-to-Long Term Forecasts of the Demand and Fiscal Spending for Early Childhood Education and Childcare”, Korea Institute of Child Care and Education, pp.51~56, 2015. https://kicce.re.kr

[10] YR Kim, SJ Cho, BY Bae, JS Kim, YM Jung, “A Study on the Actual Condition of Elementary School Care and Demand Analysis”, Korean Women’s Development Institute, pp121~122, 2018 https://kicce.re.kr

[11] JW Lee, JY Kim, KW Yoo, SC Yang, “Optimal Location Modeling for Elementary Student’s Care facility using Public Data”, Journal of Cadastre & Land InformatiX 2019 Vol.49, pp109~122, 2019. https://doi.org/10.22640/lxsiri.2019.49.2.109

[12] CS Sung, JY Park, HK Ka, “The Case Study of CCTV Priority Installation Using BigData Standard Analysis Model”, Journal of Digital Convergence 2017 May, pp.62, 2017.
https://doi.org/10.14400/JDC.2017.15.5.61

## ◐ 저 자 소 개 ◑

**윤 충 식(Chung-sik Yun)**
1993년 성균관대학교 경제학과(경제학사))
2001년 성균관학교 경영대학원 경영정보학과(경영학석사)
2020년 국민대학교 비즈니스IT전문대학원 비즈니스IT전공 박사과정 수료
2017년 행정안전부 공공데이터정책과
2020년 행정안전부 빅데이터분석활용과
2021년~현재 환경부 정보화담당관
관심분야 : 빅데이터, 인공지능, 프로세스 관리, ISP, CRM, etc.
E-mail : ycs0765@naver.com

**정 승 렬(Seung Ryul Jeong)**
1985년 서강대학교 경제학과(경제학사)
1989년 미국 위스컨신대학교 대학원 (이학석사)
1995년 미국 사우스 케롤라이나 대학교 대학원 (경영정보학박사)
1997년~현재 국민대학교 경영정보학부 및 비즈니스IT전문대학원 교수
관심분야 : 데이터 시각화, 기계학습, 빅데이터, 시스템 구현, 프로세스 관리, ISP etc.
E-mail : srjeong@kookmin.ac.kr