이상치 탐지를 이용한 쓰레기 매립지 입지선정 및 향후 토지 이용 방향

서희나(통계학과 2020314770)

윤경선(통계학과 2019312483)

이지윤 (통계학과 2020310130)

논문 초록:

코로나 19 바이러스 이후 쓰레기 배출량 증가 관련 이슈가 대두하는 한편, 인천시는 수도권 쓰레기 매립지 사용 종료를 선언하여 대체 쓰레기 매립지 조성이 필요한 상황이다. 하지만 혐오시설이라는 매립지의 특성을 고려할 때, 매립지 후보 지역을 선정하기는 쉽지 않고, 선정과정에서 환경, 기술, 사회, 경제, 관련 법 등 다양한 관점에서의 심도 있는 고려를 요구한다. 따라서 본 연구는 우선적으로 통계적 머신 러닝 기법을 활용하여 입지를 선정하는 것이 합리적이라 판단하였다. 기존 선행 연구와 대한민국 내 관련 법안, 수도권 권역의 특징을 고려하여 입지 선정 요인을 채택하였고, 관련 데이터를 수집하였다. 환경 지리적인 요인을 중심으로 매립지 입지선정을 진행하였던 기존 연구와는 달리 수도권 지역의 타 권역과의 차이점을 고려해 인문 지리적 요인을 주된 입지 선정 요인으로 채택하였다. 이때, 데이터의 결측값은 선형 회귀, 의사결정 트리 모델, 랜덤포레스트, SVM과 같은 회귀 모형을 통해 예측하였다. 이후 파생 변수를 생성 후 K-means, DBSCAN 과 같은 군집 형성 모델, 이상치 탐지 모형 Isolation Forest, QGIS를 이용해 최종 입지를 선정하였다. 입지 선정 이후에는 SHAP을 활용하여 입력 변수의 평균 영향을 통해 최종 지역 선정에 각 변수들이 가지는 영향력을 시각화하고 파악해 보았다. 또한, 입지 지역 선정에만 그치는 것이 아니라 국내외 지역 사례를 통해 향후 매립지 후보 지역의 토지 활용 방안과 그에 따른 경제적 가치를 분석해보았다.

1. 서론

1.1. 연구 배경 및 목적

코로나19 바이러스 발생으로 환경문제, 특히 쓰레기 배출량 관련 이슈가 대두되고 있다. 환경부 산하 한국환경공단에 따르면 2020년 하루 평균 폐기물 발생량은 54만 972톤으로 전년 48만 7238톤 대비 8.8% 늘어났다. 지속적으로 쓰레기가 증가하는 상황에서 인천시는 수도권 쓰레기 매립지 사용 종료를 선언하였다. 인천시는 서울시와 경기도에 대체 쓰레기 매립지 조성이 이루어져야 한다고 주장하였다. 반면, 서울시는 시내 매립지를 조성할 곳은 없다며 수도권 매립지를 계속 사용하겠다는 의견을 밝혔다. 수도권 쓰레기 매립지 문제는 서울시, 경기도, 인천시 주민들의 삶에 큰 영향을 주는 민감한 사안이다. 하지만 새로운 매립지를 찾기는 쉽지 않다. 작년 2차례에 걸쳐 대체 쓰레기 매립지를 찾고자 공모하였고 파격적인 인센티브까지 내걸었지만 단 한 곳의 지자체도 지원하지 않았다.

쓰레기 매립지에 관한 국내 연구로는 주로 지리적인 측면과 모델링 공간 분석을 주로 사용하여 진행됐고, 인문 사회적인 요소들과 통계적 기술을 사용한 연구는 많지 않다. 따라서 본 연구는 쓰레기 매립지와 연관된 특성들을 분석하고 통계적 머신러닝 기법을 활용해 알맞은 대책을 제시하고자 한다. 이러한 맥락에서 본 연구는 입지 선정에 있어 통계적인 기법을 사용하고 공공데이터 및 개방 데이터에 근거하여 의사결정을 진행했으며, 해당 지역에 매립지를 조성하였을 때의 경제적 효과까지 분석하였다는 점에서 의미가 있다.

1.2. 연구의 필요성

쓰레기 매립지를 조성하기 위해서는 부지면적, 실 매립 면적, 인구 밀집 지역 여부 등 여러 조건이 있다. 현재 수도권 대부분 지역이 개발되어 있으므로 위 조건을 만족하는 지역을 이른 시일 내에 찾아 수도권 쓰레기 매립지 사용 종료에 대비해야 한다.

첫째, 선행연구의 경우 하천 유무 등 단순한 공간분석으로 인문 사회적인 요소들을 거의 반영하지 못하였다. 하지만 본 연구에서는 상업 지적 특성을 나타내는 산업발전과 거주 인구와 유동인구를 함께 보여줄 수 있는 생활 인구 등 인문 사회적인 요소들을 입지 선정 요인으로 포함하였다. 또한, 사회 이슈를 반영한 요소들도 의사결정에 고려하였다. 최근 전력거래소에 따르면 전력 공급예비율은 떨어졌지만 전력 수요가 늘어나 전력 수급 비상경보가 발령될 수도 있다는 우려가 제기된다. 만약 쓰레기 매립지 조성한다면 매립 가스를 이용해 전기를 생산할 수 있다는 장점을 살려 전력소모량을 입지 선정 요인으로 활용하였다. 둘째, 쓰레기 매립지가 혐오시설이라는 시설적 특성을 고려할 때 우선으로 통계적인 방법론으로 입지를 선정하는 것이 합리적이라고 판단하였다. 셋째, 쓰레기 매립지를 혐오시설로만 취급할 뿐, 잠재적인 가치에 대한 사회적인 인식이 부족하므로 입지 선정을 했을 때의 경제적 효과까지 다루었다.

1.3. 연구의 범위 및 방법

광역폐기물처리시설로써 서울, 경기, 인천 지역에서 배출된 쓰레기를 매립할 지역을 모색하는 것을 반영하여 본 연구는 서울, 경기, 인천 지역을 대상으로 데이터를 수집하고 분석을 진행했다.

1.3.1. 데이터 및 변수

데이터 수집에 앞서 본 연구의 입지 선정 요인에 대한 채택이 필요했고 이를 위해 매립지 입지 선정에 관한 기존 연구를 검토하였다. Dawson과 Mercer(1986)은 환경지리적 관점에서의 매립지 입지 선정 배제 기준을 제시하였다. 이들은 우선 지질학적 측면에서는 기반암의 깊이가 깊지 못한 지역(10m 이하), 균열이 있는 석탄암 지역, 활단층과 근접한 지역(1.6km 이내)을 입지에서 배제해야 한다고 제시했다. 지질학적 측면에서는 범람원, 경사도가 25% 이상인 지역, 토양의 토심이 깊지 않은 지역(25cm 이하), 수문학적 측면에서는 백 년을 주기로 홍수가 발생하는 지역, 기후학적 측면에서는 강수량이 증발량보다 많은 지역, 태풍이나 토네이도의 경로 지역을 입지에서 배제해야 한다고 제시했다. Davis와 Lein(1991)은 환경 지리적 요인과 더불어 문화적 요인도 함께 고려한 입지 선정 배제 기준을 제시하였다. Dawson과 Mercer Jensen과 마찬가지로 범람원, 습지, 호수와 인접한 지역, 수원과 근접한 지역은 배제하는 것을 제시하였고, 문화적 요인으로는 국립공원, 관광 휴양 지역, 주거지에서 1000ft 이내의 지역은 배제되어야 한다고 하였다. Christensen(1986)은 기존 매립지나 공업 단지에서 최소한 200m 이상, 도로에서는 300m 이상 떨어진 지역, 해발고도가 높은 지역에 있어야 한다고 하였다. 미국 오하이오주의 경우 거주지, 학교, 병원, 공공시설로부터 2000ft 지역, 습지 홍수 위험 지역을 배제해야 한다고 지정하였다. (박순호,1997)

대한민국 현행 법령집 제34권, 도시계획 시설 기준에 관한 규칙 제127조는 쓰레기 매립장 입지와 관련하여 첫째, 인구 밀집 지역 및 공공기관 • 학교 • 연구시설 • 의료시설 • 종교시설 등과는 근접하지 말 것, 둘째, 풍향과 배수를 고려하여 시민의 보건 위생에 위해를 끼칠 우려가 없는 지역에 결정할 것, 셋째, 대기 및 수질 오염 등 각종 환경 오염 문제를 고려하여 결정할 것, 넷째, 용수와 동력의 확보가 용이하고 차량 접근이 쉬운 지역에 결정할 것, 다섯째, 매립 후의 토지 이용계획을 사전에 고려할 것을 명시하고 있다.

<표 1> 연구자/ 법령 별 입지 요인 및 배제 기준

연구자 / 법령	입지요인 및 배제 기준
Dawson & Mercer (1986)	지질학적 측면 ① 기반암의 깊이가 깊지 못한 지역 ② 균열이 있는 석탄암 지역 ③ 활단층과 근접한 지역 수문학적 측면 : 백 년을 주기로 홍수가 발생하는 지역 기후학적 측면 ① 강수량이 증발량보다 많은 지역 ② 태풍이나 토네이도의 경로 지역
Davis & Lein (1991)	환경적 요인 ① 범람원, 습지, 호수와 인접한 지역 ② 수원과 근접한 지역 문화적 요인 : 국립공원, 관광휴양 지역, 주거지에서 1000ft 이내의 지역
Christensen (1986)	①기존 매립지나 공업 단지에서 최소한 200m 이상인 지역 ② 도로에서는 300m 이상 떨어진 지역 ③ 해발고도가 높은 지역에 입지
미국 오하이오주	① 거주지, 학교, 병원, 공공 시설로부터 2000ft 지역 ② 습지, 홍수위험 지역
대한민국 현행 법령집 제34권 도시계획 시설 기준에 관한 규칙 제127조	① 인구 밀집지역 및 공공기관 • 학교 • 연구시설 • 의료시설 • 종교시설 등과는 근접하지 말 것 ② 풍향과 배수를 고려하여 시민의 보건 위생에 위해를 끼칠 우려가 없는 지역에 결정 ③ 대기 및 수질 오염 등 각종 환경 오염 문제를 고려하여 결정할 것 ④ 용수와 동력의 확보가 용이하고 차량 접근이 쉬운 지역에 결정 ⑤ 매립 후의 토지 이용계획을 사전에 고려

이를 모두 종합하여 입지 선정 분석을 위한 자료수집 기준을 수립했다. 우선, 대부분의 기존 연구가 홍수 및 범람, 호우 피해 발생 위험 지역을 입지에서 배제하고 있으므로 이를 반영하여 입지 분석을 진행하였다. 다만, 주로 환경 지리적 요인을 주요한 입지 요인으로 설정하는 기존 연구와는 달리, 본 연구에서는 문화적 요인, 산업적 요인과 같은 인문지리적 관점을 주요한 분석의 요인으로 설정하였고, 기존의 문화 ·관광 · 교육 · 산업시설에서 멀리 떨어진 곳에 입지를 선정하는 것을 목표로 삼았다. 이는 수도권 지역이 국내 타 권역에 비해 인구의 밀집 정도가 높고 산업시설의 집적 정도가 높음에서 기인한다.

설정한 기준에 따라 변수 생성을 위한 자료를 수집했다. 이때, 시군구 단위로 데이터를 모을 경우 66개의 row만 확보가 되어 학습이 제대로 이루어지지 않을 가능성을 우려하여 읍면동 단위로 자료를 수집하였다. 각 지역의 발전 및 생활 정도를 파악하기 위해 QGIS를 활용해 각 행

정동의 중심 좌표를 추출하고 카카오 API를 활용하여 중심 좌표 기준으로 각 읍면동의 반지름에 해당하는 거리 안에 있는 병원, 학교, 문화, 시설, 관광 명소, 교회 수를 추출했다. 단순 시설의 개수를 반영하는 것이 아닌 인구 대비 시설의 수를 반영하기 위해 공공데이터 포털에서 행정동 단위 주민등록인구 데이터를, 서울 열린 데이터 광장에서 시간대별 인구를 집계하는 생활 인구 자료를 수집했다. 생활 인구 데이터의 수집은 거주 인구뿐만이 아니라 지역의 위계적인 특성과 상업 지적 특성을 반영할 수 있게 하기 위함이다. 또한, 쓰레기 매립지 조성 시 전력생산이 가능하다는 장점을 활용할 수 있도록 전력 소모량 데이터를 한국전력공사 공공데이터로부터 수집하였다. 이외에도 지역의 경제적 지표를 반영할 수 있는 지역별 실거래가 데이터를 국토교통부 실거래가 시스템을 통해서 수집하였다.

<표 2 데이터와 출처, 제공 연도>

데이터 출처	데이터	제공 연도
GQIS	각 행정동의 중심 좌표	2022
카카오 API	병원, 학교, 학원, 문화시설, 관광명소, 교회	2022
통계청	행정동 단위 주민등록인구	2022
서울 열린 데이터 광장	서울시 생활인구	2022
국토 교통부 실거래가 공개시스템	토지 매매가	2020
국민 재난 안전 포털 자연 재난 상황 통계	10년치 강수 피해액	2010-2020
한국 전력공사 공공데이터	전력소모량	2020

1.3.2. 분석 방법

본 연구는 첫째, 기존 연구를 참고하여 쓰레기 매립지의 입지 선정을 위한 기준을 정립할 것이다. 둘째, 이러한 입지 선정을 준거로 지역 데이터들을 군집으로 묶어내고 QGIS를 활용하여 실제 쓰레기 매립지 후보 지역을 추릴 것이다. 셋째, 후보 지역들을 대상으로 환경적 특성, 인문지리적 특성을 고려하여 최종 쓰레기 매립지를 선정하는 부분으로 구성된다.

2. 쓰레기 매립지 선정에 대한 이론적 논의 및 선행연구

2.1. 혐오시설의 개념

국가적으로 매우 긴요하고 필수적인 시설이지만 그것이 해당 지역에 부정적인 효과를 일으켜 지역주민들이 건설을 반대하고 지역이기주의인 님비 현상(NIMBY, Not In My Back Yard)를 일으키는 공공시설을 혐오 시설이라고 지칭한다. 혐오 시설은 시설의 설치, 운영 관리가 제대로 이루어지지 못할 경우 환경, 경제, 사회, 정치적 문제 등 다양한 분야에서 부정적인 문제를 확산한다. 따라서, 혐오 시설의 입지를 선정해야 할 경우 환경, 기술, 사회, 경제, 관련법 등 다양한 관점에서의 심도 있는 고려를 요구하므로 시설의 입지 선정에는 다양한 제약이 뒤따른다. (오재식, 최준호, 2009)

2.2. 쓰레기 매립지의 개념

쓰레기 매립지는 쓰레기를 땅속에 매립하는 방식으로 운영되고 있으며 혐오시설로 분류된다. 따라서 매립장을 설치하고 운영하는 데 있어 많은 논란이 뒤따른다.

우선 최근에는 폐기물 감량화, 재활용, 소각 매립 등 폐기물 처리 문제가 우리나라 환경정책의 주요 관심사가 되면서 이런 혐오 시설물의 설치가 더 어려워지고 있다. 또한, 지방자치제 등 다양한 법률들이 실행되면서 주민들이 행정에 참여하기 시작했다. 이에 따라 공공의 이익보다 소집단의 이익이 더 중요시되어 님비 현상이 더욱 심해져 입지 후보지조차 쉽게 선정할 수 없게 되었다.

쓰레기 매립지는 버려지는 쓰레기가 증가할수록 악취, 소음, 오염물질이 더 많이 발생하는데, 지금과 같이 쓰레기 매립지가 한 장소에 밀집되어 있고 전국의 쓰레기가 한 매립장에서 모두 처리될 경우 넓은 면적의 땅이 필요하며 반대로 생각하면 쓰레기 매립지가 사용하는 땅만큼 발전이 이루어지지 못한다. 따라서 해당 구역의 발전은 끝났다고 취급 당해 주민들의 반발이 더욱 심해지는 것이다.

마지막으로 과거 쓰레기 매립지를 설치할 때 주민들의 동의, 투명한 과정 공개 등이 실행되지 않았기 때문에 더욱 반대하는 추세가 크다. 따라서 시설설치 계획 시 지역주민과 충분한 협의, 지역주민의 참여, 사전의 충분한 정보가 제공된다면 갈등이 어느 정도 해소될 것이다. (오재식, 최준호, 2009)

2.3. 쓰레기 매립지 입지 선정에 대한 선행연구

국내에는 쓰레기 매립지 관련 연구가 많이 이루어지지 않아서 공공시설의 입지 선정에 관한 선행연구도 함께 참고하고자 한다. 공공시설의 입지 선정과 쓰레기 매립지 연구의 동향은 지리정보체계를 이용한 연구, 쓰레기 매립지에 대한 갈등 연구, 도시 공공시설의 적정 입지 선정에 관한 연구 등으로 나타난다.

환용, 정일훈, 김철중(2010)은 파주시의 사례를 중심으로 도시 공공시설에 대한 우선 순위 분석을 진행하고, 도시 공공시설의 적정 입지를 선정하였다. 해당 연구에서는 동사무소, 경찰서, 소방서, 노인복지시설 등 생활에 필요한 20가지 공공시설의 입지를 선정했다. 또한, 연구의 객관성과 합리성을 높이기 위해 경제적, 사회적, 교통 지리적, 자연 환경적 요인으로 구분했으며 이를 토대로 GIS의 도면 중첩 방법을 이용하여 최종적인 위치를 선정하였다. 기존의 읍면동 단위에서 벗어나 건물이 지닌 특성을 고려한 분석을 했다는 점에는 의의가 있으나, 파주시 내에서만 분석이 진행되었고 공공시설의 설치기준이 모호했다는 점에서 지속적인 연구가 필요하다고 판단하였다.

김병철, 오상영, 류근호(2006)는 영향력을 고려한 적정 입지 선정 모델을 연구하였다. 밀도 기반 클러스터링 알고리즘인 DBSCAN과 밀도 기반 클러스터링에 가중치를 고려한 DBSCAN-W를 둘 다 고려하는 DBSCAN-1를 제안하며 공공시설의 입지 선정 문제, 상권분석 등 각 객체의 속성을 고려하며, 해당 기법을 스포츠 센터 입지 분석과정을 통해 분석 과정의 객관성과 합리성을 제공하고자 했다. 해당 논문에서는 기존의 공간적인 특성만을 고려한 연구에서 더 나아가 복잡한 문제 속에서 각

속성들이 갖는 영향력을 고려하여 종합적인 분석을 진행하였다.

배민기, 장병문(1998)은 입지 인자들의 상대적 중요도를 고려하여 경상북도 경산시를 대상으로 폐기물 매립지의 입지를 선정하였다. 먼저 매립지 건설비와 같은 경제적 인자, 토양과 지질과 같은 환경적 인자, 계곡까지의 거리와 같은 입지적 인자, 보호 지역 같은 제도적 인자의 중요도 평가를 계층적 분석 방법을 통하여 진행했다. 그 다음 지리정보체계를 이용하여 선형조합법과 요소 조합법을 적용하여 폐기물 매립후보지를 선정하였다. 해당 연구에서는 폐기물 매립지에 관한 입지 인자의 속성별 기준과 최종 후보지 선정을 위한 등급에 관한 제도적 기준이 필요하다는 시사점을 남겼다.

오재식, 최준호(2009)는 폐기물처리시설 설치로 인한 입지 갈등에 관한 이론적 고찰과 함께 대구시 사례 분석을 토대로 혐오시설의 갈등을 해소할 수 있는 효율적이고 바람직한 방안을 모색했다. 저자가 정의한 혐오시설 입지 갈등 유발원인은 행정에 대한 불신, 부동산 가치하락, 입지 거리, 환경적 영향 등이 있다. 해당 연구는 쓰레기 매립지의 입지를 선정할 시에는 다양한 요인과 이미 발생한 문제에는 투명성을 보장하는 것이 최선이라고 판단하였다.

이들 연구를 바탕으로 연구의 객관성과 합리성을 높이기 위해 다양한 분야를 고려해야 하며, 주민들의 관심과 환경적인 요소에 쓰레기 매립지 입지 선정이 밀접하게 관련되어 있다는 것을 확인할 수 있었다. 또한 공간분석 방식과 우선순위 분석 이 통계적 분석 방식과 환경 관련 데이터들을 주로 사용된다는 것을 확인했다. 따라서 본 연구는 환경적인 요인 대신 사회적인 항목들에 초점을 두었다는 점, 군집화 및 이상치 탐지와 같은 새로운 통계적 분석을 시도했다는 점, 매립지를 건설하였을 때 발생할 수 있는 사회적, 경제적, 환경적인 효과에 대한 분석을 진행할 것이다. 이러한 점들에서 선행연구와 차별성을 두고 있다.

3. 다양한 회귀 모델을 이용한 생활 인구 예측

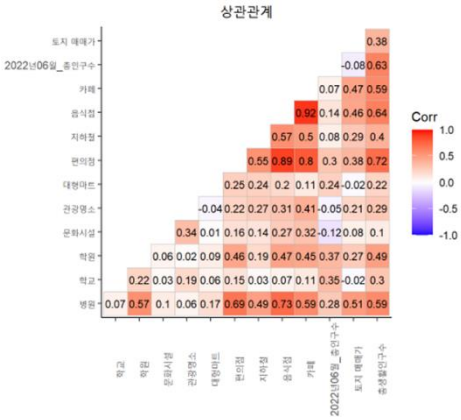
데이터 중 생활인구 데이터의 경우 서울 일부 지역에만 존재했기 때문에, 경기•인천 지역은 회귀 모델링을 통해 예측을 진행하였다. 회귀의 반응변수는 수도권 지역 내 읍• 면• 동 별 생활 인구, 설명 변수는 카카오 API를 통해 크롤링한 '병원', '학교', '학원', '문화시설', '관광명소', '대형마트', '편의점', '지하철', '음식점', '카페', '2022년 06월 총 인구수', '토지 매매가'로 설정하였다. 이때, 과적합을 방지하기 위해서 비교적 단순한 다중 선형 회귀 모델을 우선적으로 적합하고 이후 복잡한 모델을 적합하는 방식으로 예측을 진행해 보았다.

3.1. 선형 회귀

선형 회귀는 설명변수들의 선형 결합을 이용해 반응 변수를 예측하는 모형으로, 최소제곱법을 이용하여 관련 모수를 추정한다. 선형 회귀의 최소 제곱 추정과 그에 따른 통계 분석은 몇 가지 표준적인 가정에 근거한다. 첫째, 회귀 모형의 형태는 선형임을 가정한다. 둘째, 회귀 모델의 오차항들은 서로 독립이고 동일한 정규분포를 따르는 것을 가정한다. 셋째, 설명 변수는 확률변수가 아닌 미리 고정된 값으로, 선형 종속이 아니다. 이러한 가정들을 선형성, 오차항의 독립성·등분산성·정규성, 다

중공선성이라 한다.

<그림 1> 은 변수간 상관관계를 보여주는 그래프이다. 다중 선형 회귀의 경우, R 의 lmtest 패키지를 이용하여 모델을 적합하였다. 유의수준 0.05 하의 F 검정을 통해서 회귀 모델이 유의함을 확인했으나 관광명소, 대형마트, 카페, 2022 년 6 월 총 인구수, 토지 매매가를 제외한 변수들이 T 검정을 통과하지 못함을 확인하였다. 또한 상관관계 플롯을 통해서 변수들 간 높은 다중공선성을 확인해 AIC 를 기준으로 stepwise 변수 선택을 진행하였다.



<그림 1> 변수간 상관관계 플랏

변수 선택 결과 편의점, 2022 년 06 월 총 인구수, 토지 매매가, 관광명소, 카페가 설명 변수로 채택되었다. 이후 R 의 GVLMA 패키지와 lmtest 패키지를 활용하여 선형 회귀의 기본 가정 만족 여부를 확인했으나, 모델이 선형성 • 정규성 • 등분산성을 만족하지 못함을 발견했다. Min-Max 스케일링, 반응변수의 로그 변환, 제곱근 변환을 통해서 선형회귀 가정 불충족의 문제를 해결하고자 하였으나 변수 변환을 한 모델 역시 선형 회귀의 선형성과 정규성을 만족하지 못했다.

<표 3 선형 회귀의 기본 가정 충족 여부>

라이브러리	선형회귀의 기본 가정		가정 충족/ 불충족
gvlma	Global Stat	선형성	불충족
	Skewness	왜도	불총족
	Kurtosis	첨도	불총족
	Heteroscedasticity	등분산성	불총족
lmtest	shapiro.test	정규성	불총족
	ncvTest	등분산성	불총족
	dwtest	오차항의 독립성	불총족

Stepwise selection 이후 선형 회귀 가정 검정

라이브러리	선형회귀의 기본 가정		가정 충족/ 불충족
gvlma	Global Stat	선형성	불충족
	Skewness	왜도	총족
	Kurtosis	첨도	불총족
	Heteroscedasticity	등분산성	총족
lmtest	shapiro.test	정규성	불총족
	ncvTest	등분산성	불총족
	dwtest	오차항의 독립성	불총족

설명변수 min-max scaling, Y변수 로그 변환 이후 선형 회귀 가정 검정

라이브러리	선형회귀의 기본 가정		가정 충족/ 불충족
gvlma	Global Stat	선형성	불충족
	Skewness	왜도	총족
	Kurtosis	첨도	총족
	Heteroscedasticity	등분산성	총족
lmtest	shapiro.test	정규성	총족
	ncvTest	등분산성	불총족
	dwtest	오차항의 독립성	불총족

설명변수 min-max scaling, 예측변수 제곱근 변환 이후 선형 회귀 가정 검정

3.2. Random Forest와 SVM

다중선형회귀를 이용한 생활 인구 예측이 적합하지 않다고 판단하여 SVM, 비선형 트리 모델인 Random Forest와 의사결정 트리 모델 역시 적합해 보았다. 트리 모델과 SVM의 경우 PYTHON의 scikit-learn 모듈을 이용하였다.

의사결정 트리 모델은 분류 규칙을 나무 구조로 형성하여 데이터를 차례로 분류하고 특성을 예측하는 모형이다. 의사결정 트리 모델을 형성하기 위한 알고리즘은 CHAID, Exhaustive CHAID, CRT, QUEST 방법 등이 있다. 본 연구에 이용된 Python scikit-learn 모듈의 의사결정 나무는 CART 알고리즘 기반 모델이다. CART 는 이진 분할 방법을 이용하여 종속변수에 대해 가능한 많은 동질적인 그룹이 속하도록 노드를 생성하는 방법을 사용한다. 이때, 노드의 분할 규칙은 지니 지수, 엔트로피 지수, 편차 지수와 같은 불순도를 최소화하는 것을 선택한다. (손용정, 김현덕, 2012)

랜덤포레스트는 다수의 의사결정 트리를 구성하는 앙상블 학습을 통해 데이터 분류 및 회귀를 수행하는 모델이다. (김현일, 이연수, 김병현,2021) 모든 변수를 사용하여 가장 최적의 결과를 내는 분할로 각각의 노드를 나타낸 기존의 의사결정트리 모델과 달리, 랜덤 포레스트에서는 노드를 나타낼 때 무작위로 선택된 설명변수의 집합 중 최적의 결과를 내는 방법을 이용한다. (김판준,2019)

SVM (서포트 벡터 머신) 회귀 모델은 여러 클래스에 속하는 데이터들을 최대한의 margin 으로 분리하는 최적의 초평면을 찾는 방식으로 작동한다.

우선, 적합하는 세 모델 중 최적의 모델을 선정하기 위해 존재하는 서울시 생활 인구 데이터를 각각 학습 데이터, 검증데이터로 8:2 비율로 분할하여 모델의 성능을 평가하였다. 이때, 생활 인구 값이 여덟 자리수로 설명변수의 단위보다 훨씬 컸기 때문에, MSE (평균제곱오차) 값으로는 정확히 모델의 성능을 파악하기는 어렵다고 판단하였고, 잔차의 절댓값에 대한 평균을 비율로 나타내는 MAPE를 100%에서 뺀 값과 R^2값으로 모델의 성능을 평가하였고 MAPE와 R^2의 식은 다음과 같다.

$$MAPE = \frac{100\%}{n}\sum\left|\frac{y-\hat{y}}{y}\right|$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

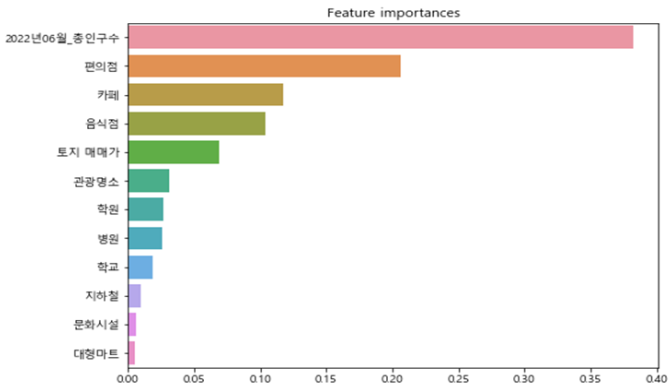
<표 4>은 랜덤포레스트, 의사결정트리, SVM 모델로 이용해서 학습데이터를 학습하고 검정 데이터를 이용해서 성능을 측정했을 때의 결과를 나타내고 있다. SVM 의 경우 R^2 값이 음수인데, 이는 회귀 모델의 성능이 평균값으로 예측하는 값보다 결과가 좋지 않음을 의미하기 때문에, 일차적으로 최종 모델 선정에서 배제하였다. 랜덤포레스트 회귀 모델이 (100-MAPE)(%) 값과 R^2 값이 제일 높았기 때문에 이를 생활 인구 예측을 위한 최종 모델로 선택하였다.

<표 4> 랜덤포레스트, 의사결정트리, SVM의 학습 결과

회귀 모델	100-MAPE (%)	R^2
RandomForestRegressor	79.132158263	0.68718228
DecisionTreeRegressor	73.7436045954	0.3974656714
SVM	59.13980930528	-0.075848099139

최종적으로 생활 인구를 예측하기에 앞서, 모델의 예측 값과 실제 값의 차이를 줄이는 방향, 즉 MAPE 를 최소화하는 방향으로 모델이 적합 될 수 있게 변수를 추출하고, 하이퍼파라미터 튜닝을 진행했다. 우선 변수 선택의 경우 싸이킷런 모듈에 내장된 feature_importances_함수를 통해 특성 중요도를 구하고, 이를 시각화해 보았다. 특정 변수를 기점으로 특성 중요도 값이 급격하게 낮아지는 것을 확인하였고, 중요도 상위 5 개 변수('2022 년 06 월 총인구수', '편의점, 카페', '음식점, 토지 매매가')로 예측 변수를 한정 지었다.

<그림 2> 랜덤 포레스트를 통해 도출해낸 변수 중요도



\

또한 GridSearchCV를 통해서 MAPE 값을 최소화하는 최적의 하이퍼파라미터 값은 결정 트리의 개수를 지정하는 n_estimators의 경우 300일 때, 최적의 분할을 위해 고려할 최대 feature 개수, max_features의 경우 3일 때임을 확인했다. 변수 추출과 하이퍼파라미터 튜닝을 마친 이후 경기•인천 지역의 생활인구의 예측을 최종적으로 진행했다. 특성 중요도에 따라 변수를 추출하고 GridSearchCV를 활용하여 하이퍼파리미터 튜닝을 완료했을 때 (100-MAPE)(%) 값이 상승한 것을 <표 5>를 통해 확인할 수 있다.

<표 5> 하이퍼파라미터 튜닝 전후 성능 비교

변수 추출/하이퍼 파라미터 최적화	100-MAPE(%) 값
이전	79.132158263
이후	80.67996688106766

3.3. 파생 변수

생활 인구 예측 완료 후 최종적인 군집 분석에 있어서 수집한 데이터를 그대로 사용하기 보다는 지역의 특성을 더 잘 반영할 수 있는 파생 변수를 형성하였다. 앞서 수집한 주민 등록 인구, 생활 인구, 시설의 수, 전력량, 강수 피해액 데이터를 이용하여 형성한 파생 변수는 <표 6>에 정리되어 있다.

<표 6> 파생 변수 수식

변수	수식
인구	$\frac{\text{주민등록인구}}{\text{(총 생활 인구 수 + 주민 등록 인구)}}$
토지 매매가	$\frac{\text{토지 매매가}}{\text{면적}}$
산업 발전	$\frac{\text{학교 + 문화시설 + 관광명소 + 음식점}}{\text{면적}}$
전력 소모량	$\frac{\text{전력량}}{\text{(총 생활 인구수 + 주민 등록 인구)}}$

4. 클러스터링 (Clustering)

우선 데이터에 클러스터링 기법을 적용하여 각 지역의 특징과 성격을 파악했다. 클러스터링이란 데이터 간의 유사성을 기준으로 데이터 들을 군집으로 묶는 기법이다.

4.1. K-means Clustering

K-means 클러스터링은 각 패턴과 클러스터의 중심값과의 거리 차이를 최소화하는 방법으로, (민준영,1995) 우선 패턴을 k개의 클러스터로 나눈 후 클러스터에 포함된 패턴들의 평균으로 클러스터의 중심값을 계산하고 이 중심값을 각 패턴과의 거리를 계산한 후 가장 거리가 가까운 클러스터에 패턴을 포함한다. (민준영,1995) 그 조건은 아래 식과 같다.

$$x_i \in c_j \ \ iif \ \ \|x_i - z_j\|^2 < \|x_i - z_k\|^2$$

K-means 클러스터링은 사전적으로 k 값을 정해주어야 하는데, 이는 엘보우(elbow)기법과 실루엣(silhouette) 기법을 통해 최적의 k 값을 구할 수 있다.

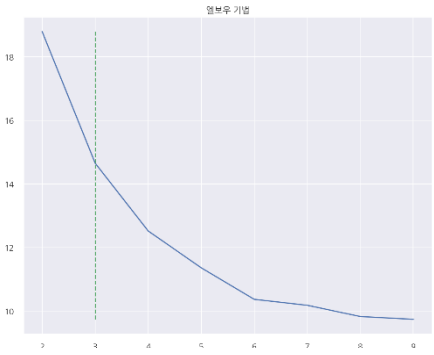
4.1.1. 엘보우(elbow) 기법

엘보우 기법은 군집 내 응집도의 오차제곱합(Sum of Squares Error, SSE)의 값이 최소가 되도록 클러스터의 중심을 결정해나가는 방법이며 (김현정, 2021) 오차제곱합을 계산하는 공식은 다음과 같다.

$$SSE(CL_i) = \sum_{x \in CL_i} Proximity(x, c_i)$$

엘보우 기법을 구현하기 위하여 오차제곱합을 그래프로 그리면, 클러스터 개수가 늘어나면서 centroid 간 평균 거리가 급격히 감소하는데, 적절한 k가 발견되면 실선의 기울기가 완만해진다. 그래프상 이 부분이 팔꿈치와 닮아서 엘보우 기법이라고 하며 완만해지는 첫 지점을 Elbow point라 한다. 이때, k가 최적의 클러스터 개수가 된다. (김현정, 2021)

<그림 3>은 엘보우 기법으로 최적의 k 값을 확인한 결과이다. 그래프의 표시 점선을 토대로 최적의 클러스터의 후보 개수는 3 개로 결정하였다.



<그림 3> 최적의 k값 확인

4.1.2. 실루엣(silhouette) 기법

최적의 k를 구하는 다른 방법인 실루엣 기법은 각 군집 간의 거리가 얼마나 효율적으로 분리되어 있는지 확인하는 간단한 방법으로 한 군집 안의 데이터들이 다른 군집과 비교해서 얼마나 (김현정, 2021)유사한지를 기준으로 k 값을 산정한다. 실루엣 계수의 수식은 다음과 같다.

$$(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad \Longrightarrow \quad \begin{cases} 1 - \dfrac{a(i)}{b(i)} & if \;\; a(i) < b(i) \\ 0 & if \;\; a(i) = b(i) \\ \dfrac{a(i)}{b(i)} - 1 & if \;\; a(i) > b(i) \end{cases}$$

실루엣 계수는 1 에 가까울수록 근처의 군집과 멀다는 의미로 군집 형성이 적절하며, -1 에 가까울수록 군집의 형성이 부적절하다고 판단한다. (김현정, 2021)

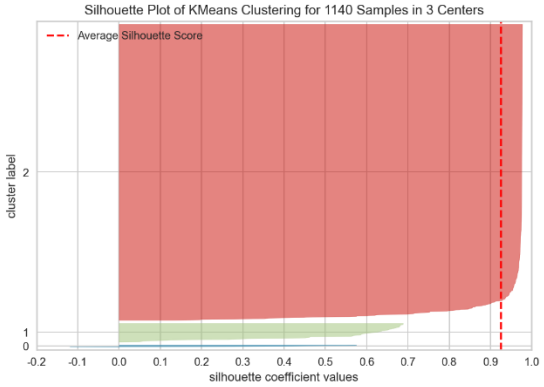
<표 7>은 클러스터 개수 별 실루엣 계수를 나타낸 테이블이다. 해당 테이블을 토대로 실루엣 기법에 따른 최적의 K 는 엘보우 기법과 동일하게 3 임을 확인하였다. 하지만, 실루엣 계수가 높다고 반드시 클러스터링이 올바르게 되었다는 보장이 없으므로 반드시 군집화 된 데이터를 시각화하여 각 군집의 실루엣 그래프 두께가 균등하게 분포되어 있는지 확인해보아야 한다.

<표 7> 실루엣 계수로 확인한 최적의 K 개수

	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9
silhouette_score	0.9250	0.9262	0.9170	0.9162	0.9132	0.8942	0.8943	0.8578
Best = 3								

<그림 4>는 k 가 3 일 때의 실루엣 계수를 시각화한 그래프이다. 군집 2 의 실루엣 계수 그래프가 다른 두 군집에 비해 훨씬 두꺼워 각 군집의 실루엣의 편차가 심하다는 것을 확인할 수 있다.

<그림 4> k=3일 때 실루엣 계수 시각화



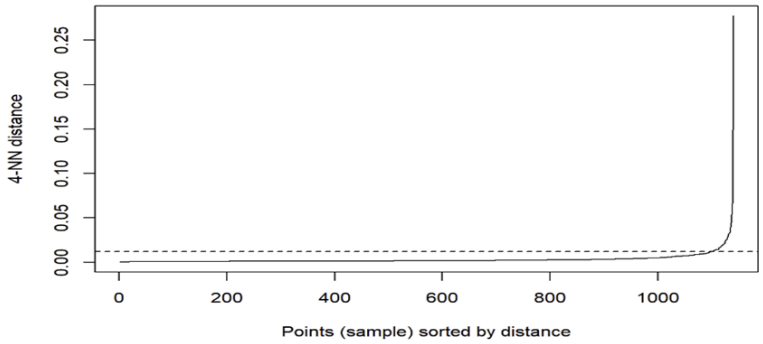
이상적인 군집 분포는 실루엣 계수의 분포 그래프가 균등한 두께를 가지면서 큰 실루엣 계수를 가져야 한다(김현정, 2021)는 논거에 의하여 K-means를 통한 데이터 군집이 잘 이루어지지 않았다고 판단하였다. 따라서 중심 기반의 K-means 방식이 아닌 밀도 기반 군집화 방식인 DBSCAN을 통해 데이터 군집을 다시 진행했다.

4.2. DBSCAN

DBSCAN 은 밀도 기반 군집화 방식으로, 주어진 데이터 집합에서 클러스터와 클러스터에 속하지 않은 노이즈를 식별한다. K-means 클러스터링과 계층적 클러스터링의 경우 군집 간의 거리를 바탕으로 클러스터링이 작동하고, 원의 형태로 클래스가 분포한다. 반면, 밀도 기반 군집화는 이웃 개체와의 밀도를 계산하여 클러스터링을 해 비선형 클러스터나 다양한 크기를 갖는 공간데이터를 더욱 효과적으로 군집할 수 있다. 또한, 클러스터 내외 간 데이터들의 밀도의 차이는 상당이 크기 때문에 잡음을 인식할 수 있다. 해당 알고리즘은 기준점으로부터 엡실론(Epsilon) 거리 내에 최소 점의 개수를 설정하고 이를 만족하면 하나의 군집으로 판단하기에, 이들을 각각 입력 모수 Eps, MinPts 로 설정해야 한다.

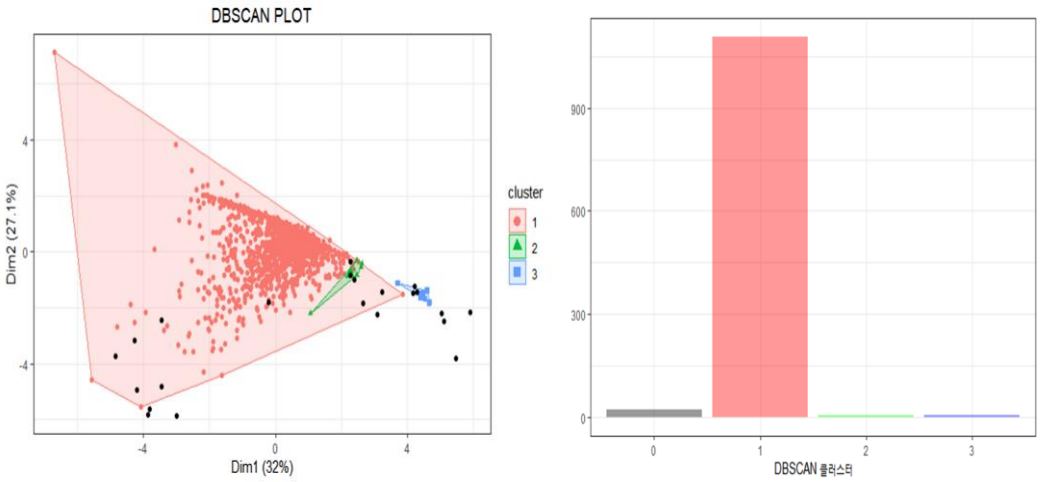
DBSCAN 을 이용해 매립지 입지에 적합한 클러스터를 찾기에 앞서 Marin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu 가 제안한 Heuristic 한 방법으로 모델의 입력 모수를 결정하였다. 이들은 MinPts 를 k 개라고 할 때 K-NN 의 거리 K-dist 를 구하고, 이를 정렬하여 구하는 그래프를 그릴 것을 제안한다. 이때 K-dist 를 내림차순으로 정렬한다면, 그래프의 Elbow point 의 왼쪽은 노이즈, 오른쪽은 군집으로 구분된다. 논문에서는 이 첫 번째 Elbow point 를 Eps 로 결정지을 수 있다고 소개한다. 또한 K-dist 그래프를 그릴 때 k>4 일 때, K-dist 그래프는 4-dist 그래프는 유의한 변화가 없을뿐더러, K 값이 커짐에 따라 연산량이 상당이 커지기 때문에 2 차원 데이터에서는 MinPts 를 4 개로 하는 것을 권장한다. (Marin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu,1996)

<그림 5> MinPts =4일 때의 k-dist의 시각화



<그림 5>는 수도권 지리 데이터에 Min-Max 스케일링을 진행하고 MinPts=4일 때의 k-dist 그래프이다. 4-NN의 distplot을 그렸을 때 Elbow point가 0.012에서 형성되었기 때문에 Eps를 0.012, MinPts를 4로 설정하고 클러스터링을 진행했다. 다만, 중심 기반 클러스터링과 마찬가지로 <그림 6>에 나타나 있다시피 하나의 클러스터와 소수의 노이즈로 인식되는 심각한 클래스 불균형이 발생했다.

<그림 6> DBSCAN 클러스터링 결과 시각화



4.3. Isolation Forest를 이용한 수도권 지역 내 이상치 탐지

중심기반의 K-means 방식과 밀도기반의 DBSCAN 방식을 통해 데이터 군집을 시도하였으나 모두 클래스 불균형이 나타났다. 앞서 데이터 분석의 목적을 재고해 보았을 때, 혐오시설이라는 매립지의 시설적 특징을 고려할 때, 일반적인 군집 분석 방식보다는 이상치 탐지로 대체 매립지 후보 지역을 찾는 것이 더 타당하다고 판단하였다.

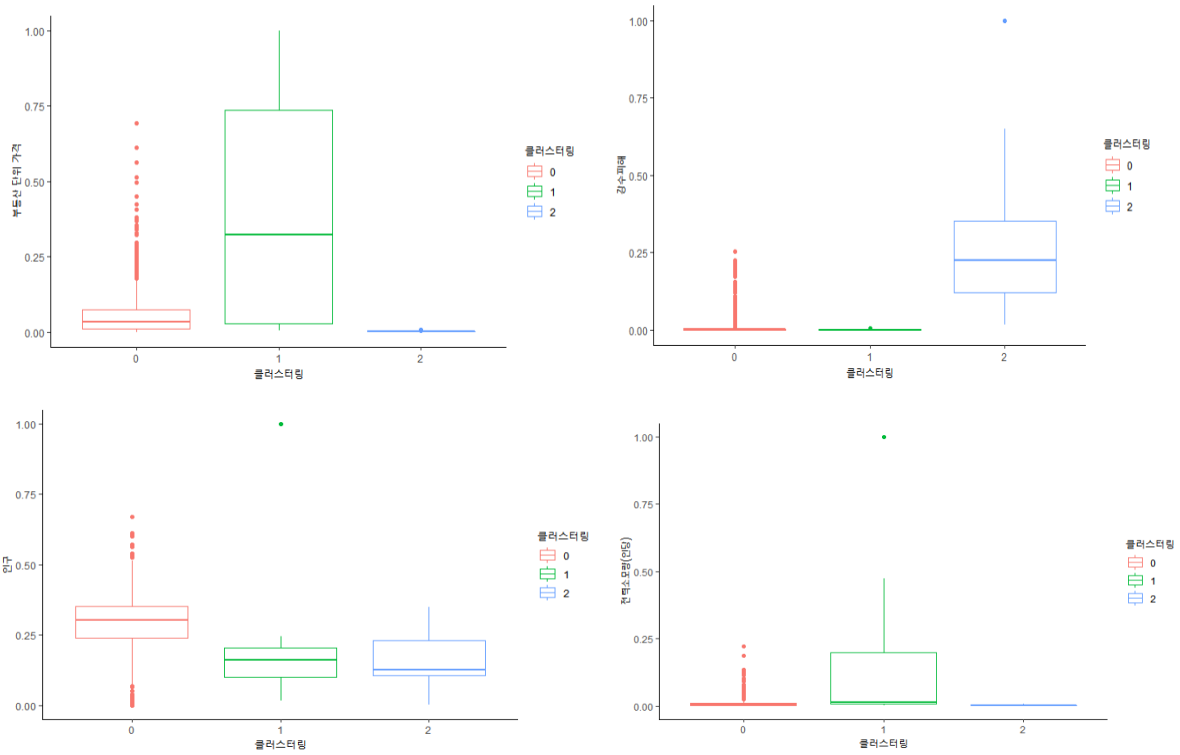
Isolation Forest 는 트리 기반 비지도 이상치 탐지 모델로, 자료의 거리나 밀도에 의존하지 않고 무작위로 데이터를 분할하며 관측치를 고립시킨다. 이때, 관측치를 고립시키는 것은 이상치가 정상

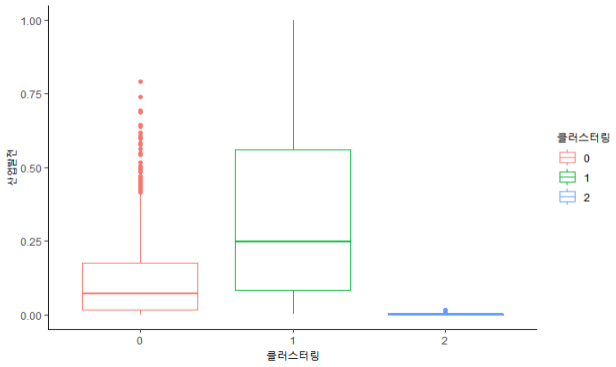
데이터보다 쉽다는 개념 하에 모델이 작동한다. 이상치의 고립이 더 쉬운 것은 이상치를 적은 횟수의 분할로 고립시킬 수 있음을 의미한다.

이상치는 특성상 고립에 취약하여 트리모델의 뿌리에 가깝게 고립될 가능성이 높아 상단부에, 정상치는 트리모델의 하단부에 깊게 위치한다. 데이터의 이상 유무에 따라 경로의 길이에 차이가 발생하며, 이를 기준으로 이상치와 정상치의 분류가 이루어진다. 이때, 이상치 탐색을 위한 공간분할은 무작위로 생성되며, 예상 경로 길이를 계산하기 위해 n 회의 반복학습 후의 평균 경로 길이를 계산한다. (진승종, 유상철, 김남기, 하윤우, 왕지남, 2022)

수도권 데이터를 DBSCAN 클러스터링과 마찬가지로 Min-Max 스케일링을 진행한 후 R 의 isotree 패키지를 활용하여 Isolation Forest 에 적합하였다. 그 결과 이상치의 개수는 26 개로 나타났다. 이 중 매립지 선정에 있어서 실매립 면적인 100 만 제곱을 수용할 수 없는 지역을 후보 지역에서 제거하였다. 또한, 이상치로 분류된 지역을 (주민 등록 인구 + 생활 인구) 값을 기준으로 내림차순 정렬하고 상위 50%와 하위 50%으로 나누어 분포를 확인해보았다. 이상치의 상위 값과 하위 값의 분포 차이는 각각 부동산 단위 가격, 전력 소모량, 산업 발전, 강수 피해에서 뚜렷하게 나타났다. 대체로 (주민 등록 인구 + 생활 인구) 값이 큰 지역의 부동산 단위 가격, 전력 소모량 분포 값이 크게 나타나고 반대로 강수 피해 값은 작게 나타났는데, 이는 파생 변수를 형성할 때 면적 당 강수피해액에 (주민 등록 인구 + 생활 인구)를 나눠줬기 때문이다.

<그림 7> DBSCAN 클러스터링 결과 시각화





5. 결과 해석

5.1. SHAP (Shapley Additive exPlanations)

5.1.1. XAI와 SHAP

머신 러닝은 연산 과정이 많을수록, 절차가 깊어질수록 처리해야 하는 매개변수가 많아지기 때문에 학습 과정이 복잡해진다. 다수의 모델은 사람이 이해하기는 너무나 복잡한 매개변수와 의사 결정 과정을 가지고 있다. 이처럼 머신 러닝 모델의 의사 결정 과정을 인간이 직접 이해할 수 없을 경우의 모델을 블랙박스 모델이라고 부른다. 머신 러닝 기법들은, 특히 딥러닝, 대부분 블랙박스 성질을 지니고 있다.

XAI란 인공지능 모델이 특정 결론을 내리기까지 어떤 근거로 의사 결정을 내렸는지 알 수 있도록 설명 능력을 부여하는 기법으로 인간과 기계의 상호작용에서의 합리성을 확보해준다. 즉, XAI란 기존 인공지능, 머신 러닝 모델 위에 설명성을 부여해 블랙박스를 들여 볼 수 있게 해준다. (안재현, 2020)

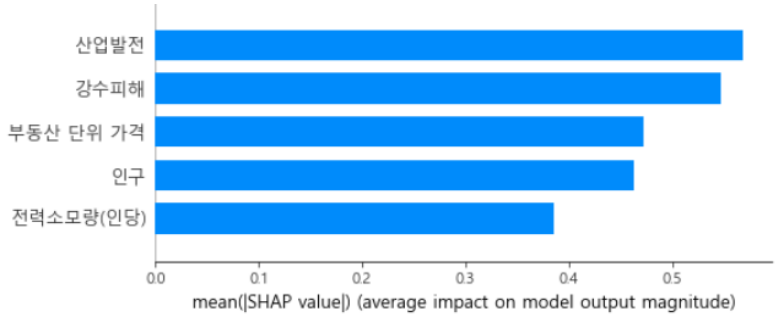
SHAP은 로이드 새플리가 만든 이론 위에 피처 간 독립성을 근거로 덧셈이 가능하게 활용도를 넓힌 논문이다. SHAP은 새플리 값과 피처 간 독립성을 핵심 아이디어로 사용하는 XAI 기법 중 하나로 각각의 입력 변수에 대한 새플리 값을 계산함으로써 입력 변수와 모델의 결과값 사이의 관계를 분석한다. 새플리 값은 게임이론에서 게임의 참여자 간 협조로부터 얻어진 전체가 성과에서 각 참여자가 얼마나 공헌 했는지를 나타내는 값을 의미한다. 즉, 기계 학습 모델에서는 예측 모델의 결과에 대한 변수 기여도를 모든 가능한 변수 조합에 대하여 종합적으로 구하며 피처 간 의존성까지 고려해서 모델 영향력을 계산하는 것이다. SHAP은 각각의 입력 변수의 중요도뿐만 아니라 결과에 미치는 영향력의 방향성과 크기까지 확인할 수 있어 SHAP을 활용하면 각각의 입력 변수가 예측값에 어떠한 효과를 미치는지 확인할 수 있다. (박성우, 노윤아, 정승민, 황인준, 2021)

5.1.2. 변수 중요도

본 연구에서는 비지도 학습 중 하나인 Isolation forest를 사용하여 최종 지역을 선정했

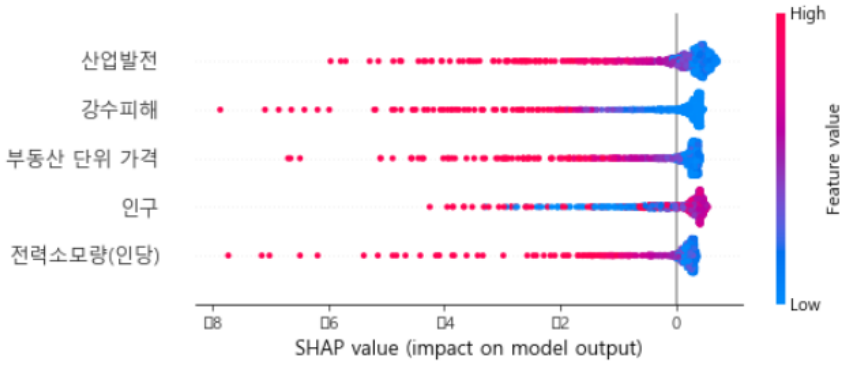
다. Isolation forest 또한 블랙박스 모델이기 때문에 SHAP을 사용하여 각 변수의 중요도와 결과에 미치는 영향력과 방향성을 알아보고자 한다.

<그림 8> 각 변수의 평균 SHAP 값 시각화



<그림 8>은 입력변수들의 평균 SHAP값을 시각화 한 것으로, 각 변수의 평균 영향을 나타내고 있다. IsolationForest의 SHAP을 적용한 결과를 토대로 분석하자면, '산업발전', '강수 피해', '부동산 단위 가격', '인구', '인당 전력소모량' 순으로 지역 산발 여부에 중요하게 작용함을 알 수 있다. 각 변수의 영향력을 산업 발전과 강수피해, 부동산 단위 가격과 인구와 같이 균등하게 감소하는 것을 볼 수 있다.

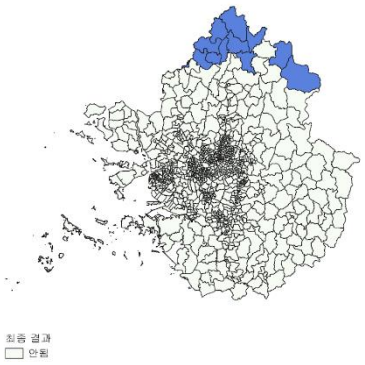
<그림 9> 입력변수의 SHAP 값의 분포도 및 영향도



<그림 9>는 입력변수들의 SHAP value 값이 표시된 분포도이며, SHAP value의 절댓값이 높은 순으로 표시되어있다. 중앙선을 기준으로 왼쪽은 예측 값이 낮아지는데, 오른쪽은 예측 값이 높아지는 데 영향을 준다는 것을 의미한다. 이를 토대로 해석해보면, 모든 변수의 데이터는 중앙선 기준 오른쪽이 집중되어있음을 알 수 있다. 따라서, 산업발전, 강수 피해, 인당 전력소모량은 낮을수록 인구는 높을수록 최종 지역 선정에 높은 영향을 가진다는 것을 확인할 수 있었다. 이때, 인구 변수가 높아야 하는 이유는 인구 변수는 주민등록인구를 총 생활 인구 수와 주민등록인구 수로 나누었기 때문에 그 수가 높다는 것은 실제로 그 지역에 거주하고 있는 사람이 많지 않다는 것을 의미한다.

5.2. 최종 결과

<그림 10>은 최종 지역의 후보를 시각화 한 것이다. 파란색으로 표시된 지역들이 최종 후보지로 선정된 지역으로 모두 경기도 북쪽에 위치하고 있음을 확인할 수 있다. 해당 지역에 관하여 자세한 결과는 <표 8>를 통해 확인할 수 있다. SHAP의 변수 중요도를 토대로 <그림 11>을 통해 포천시와 연천군의 산업 발전, 강수 피해, 부동산 단위 가격, 인당 전력소모량은 낮다는 것을 확인할 수 있었다. 하지만, 인구 변수는 높지 않고 낮았는데, 이는 지역 특성상 주말 등 특정 날짜에만 생활 인구가 많이 방문했기 때문이다.



<그림 10> 최종 선택 지역 시각화

반대로 하얀색으로 칠해진 지역 같은 경우, 선택 여부에 큰 영향을 끼치는 산업발전 값이 높으며 부동산 단위 가격 또한 높다는 것을 알 수 있다. 해당 변수들의 값이 크다는 것은 개발할 땅과 많은 자본이 투자되어야 한다는 것을 의미하기 때문에 최종 지역으로는 선정되지 못하였다.

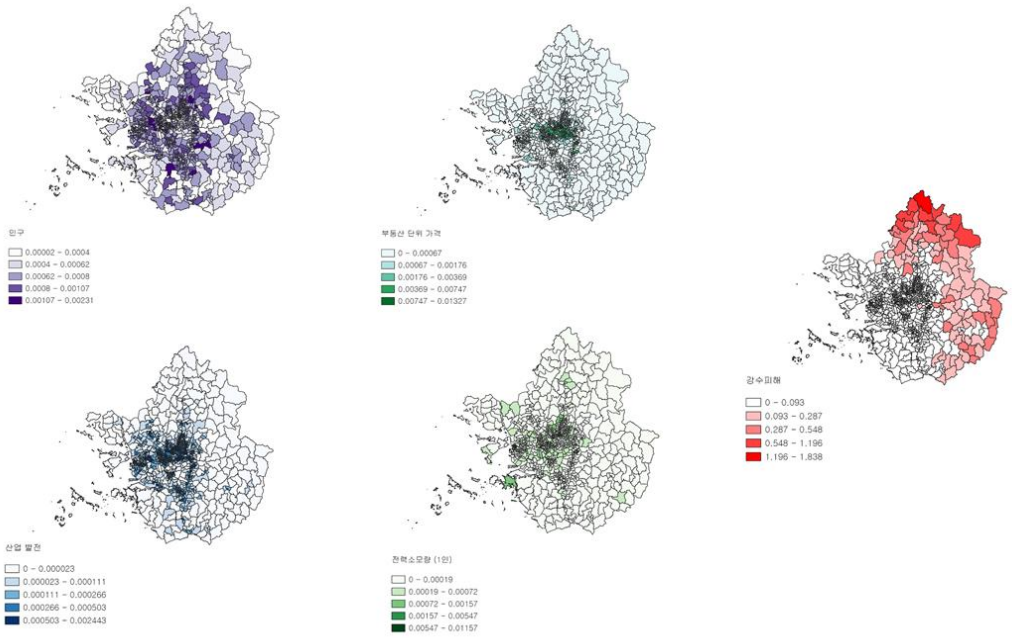
포천시와 연천군이 최종적인 후보지로 선정되었지만 연천군은 모든 지역이 군사시설 보호 구역으로 지정되어 있어 건축, 산림훼손, 농지 개간 등 인허가가 필요할 경우 관할 군부대를 통해 군 협의를 득의 하여야 한다. 즉, 더 큰 비용과 정책이 필요하겠다고 판단하여 연천군은 최종 후보지에서 제거하였다. (Yes, 연천!, 2022)

포천시는 경기도 동북단에 위치한 시로 1 개 읍, 11 개면, 2 개 행정동, 251 개의 행정리로 이루어져 있다. 포천시의 전 토지의 69.1%는 임야이고 경지는 17.6%이며 광주산맥의 지맥이 지나고 있어 대체로 산지가 많다. 주요 수계는 시 중앙에 흐르고 있는 포천천이며 이 하천은 한탄강으로 흘러 들어간다. 타 지역에 비하여 산이 많으므로 문화적 관광지보다는 폭포, 산악관광지, 골프장, 온천 등 자연 관련 관광지가 발전된 편이다. (한국민족문화대백과사전, 2022)

<표 8> 최종 선택 지역의 데이터

시도	시군구	읍면동	부동산 단위 가격	강수피해액	인구	전력소모량(인당)	산업발전
경기	포천시	창수면	0.000010	0.654538	0.000226	0.000072	6.013328e-07
경기	포천시	이동면	0.000010	0.682395	0.000442	0.000075	2.385020e-06
경기	포천시	관인면	0.000005	0.628322	0.000281	0.000069	9.991348e-07
경기	연천군	연천읍	0.000013	0.764389	0.000535	0.000014	1.970617e-06
경기	연천군	군남면	0.000009	0.548451	0.000307	0.000010	2.129152e-06
경기	연천군	백학면	0.000011	0.862142	0.000258	0.000016	6.518324e-07
경기	연천군	미산면	0.000012	0.543631	0.000184	0.000010	3.615839e-07

<그림 11> 각 변수 별 데이터 분포



6. 쓰레기 매립지 설치 시 갈등 해소 및 경제적 효과

쓰레기 매립지를 조성한다면 혐오시설로 여겨지는 것이 아닌 주민 친화적이고 환경친화적인 또 하나의 문화공간이 될 수 있다. 국내 사례를 살펴본다면 남양주시와 구리시에서 발생하는 폐기물을 처리하는 시설이자 지역의 명소로 자리 잡은 '에코-랜드'가 있다. 남양주도시공사는 에코-랜드를 단순한 폐기물 매립지에서 벗어나 부지 면적 일부분을 축구장, 실내 체육 시설 등 주민 친화 공간으로 활용하고 있다. 또한 매립지 외곽으로는 산책로를 조성하여 주민들이 운동도 즐기고 폐기물 매립 작업도 볼 수 있도록 관리하고 있다. 더 나아가 환경 교실도 운영하면서 쓰레기 매립지에 대한 이미지를 개선하고자 노력하고 있다. 환경적 측면에서도 정기적으로 매립장 주변의 수질 및 가스 확산 여부 등을 조사하고 있다. 매립장으로 인한 오염이 확산하지 않도록 심혈을 기울이고 있으며 그 결과 다양한 생물종이 서식하고 있다.

더 나아가 미개발 지역을 새로운 관광지로 탈바꿈하며 추가적인 경제적 효과를 볼 수 있다. 폐기물 처리시설은 크게 쓰레기 매립장과 소각장으로 분류되는데 우리나라에서는 주로 매장하는 방식으로 쓰레기가 처리되지만, 해외에서는 소각을 통해 주로 쓰레기를 처리하고 있다. 해외 사례의 경우 일본은 지진과 쓰나미 같은 자연재해가 자주 발생하는 만큼 침수 관리와 같은 여파 관리가 어렵기 때문에 자원 재활용과 소각로를 결합한 에너지 회수시설로 문제를 해결한다. 오사카 마이시마 소각장의 경우 하루 900 톤을 처리하는 대형 처리장이지만 소각장 주변을 공원화하고 스포츠 시설을 조성함으로써 오사카 관광지 114 곳 중 27 위로 선정될 만큼 많은 사람이 방문하고 있다. 이와

유사하게 영국 옥스퍼드셔의 아들리 소각장도 혐오시설로 여겨지는 소각장이 주민 친화적인 외관을 가지고 전력을 공급함으로써 주민들과의 혼란을 해소하였다. 아들리 소각장에서는 하루 900 톤의 쓰레기가 처리되며 총 24MW 의 전기를 생산하고 있으며, 이는 옥스퍼드셔 지역 3 만 가구에 공급하는 전력량에 해당한다. 마이시마 소각장과 아들리 소각장에서 착안하여, 새로운 쓰레기 매립지를 관광지화 할 때 해당 지역의 경제와 토지 이용의 능률을 소생시킬 수 있을 것이다.

난지도 사례를 참고한다면 쓰레기 매립지를 폐쇄 후에도 생태 복원 및 인근 환경 개선 효과를 볼 수 있을 것으로 기대된다. 난지도는 1978 년 쓰레기 매립지로 지정된 후 쓰레기 수용 한계량에 도달하여 1993 년에 매립지 사용을 종료하였다. 서울시는 난지도를 친환경적인 공원으로 탈바꿈하고자 침출수 처리, 상부 복토화 작업 등 안정화 공사에 착수하였다. 매립지 복원 사례로 국제적으로 인정받았을 만큼 오염된 환경을 되살리고자 여러 노력을 기울였고, 현재는 많은 시민이 찾는 공원이 되었다. 이를 통해 난지도의 생태계는 멸종위기종이 살고 있을 만큼 다양한 동식물이 살고 있다. 그리고 경제적인 측면에서 매립가스 회수를 통해 자원을 절약하였고 미개발 지역이었던 난지도에 택지 개발로 주변 환경이 크게 개선되었다. 또한 난지도 생태공원은 서울시의 환경 친화 사업으로 자리 잡아 여러 관광객이 몰려 추가적인 사회 경제적 효과도 보고 있다. 난지도의 사례처럼 포천시도 쓰레기 매립지 사용을 종료한 후에도 복원을 통해 새로운 기회를 엿볼 수 있을 것으로 기대된다.

7. 결론

본 논문에서는 최적의 쓰레기 매립지 입지를 파악하기 위해 K-means 클러스터링, DBSCAN 과 이상치 탐지 기법을 이용하여 최종 지역을 선발했다. 먼저, 데이터를 확보하는 데 있어 생활 인구 관련 데이터가 부족하다는 것을 확인하여 랜덤포레스트와 같이 다양한 방법론을 통해 생활 인구를 예측하였다. 그리고, 클러스터링을 통해 클래스 불균형이 심하다는 것을 확인하여 이상치 탐지를 통해 최종 후보 지역을 선정하였다. 해당 결과를 토대로 SHAP 을 통해 변수 중요도를 구하여 각 변수가 의사결정에 얼마만큼의 효과가 있었는지 확인하였다. 그 후 정부가 내세운 쓰레기 매립지 입지 기준에 맞추어 최종 지역을 선정하였다.

구하고자 하는 시설의 특성상 다양한 모델을 사용하는 것이 힘들어 Isolation Forest 만 최종적으로 사용하였다는 것이 본 연구의 한계점이다. 다양한 모델들을 활용하여 결과를 비교하며 분석을 진행했다면 더 풍부한 분석이 되었을 것이다. 또한, 사용된 데이터의 수집 시점이 모두 다르다는 것 역시 본 연구의 한계점이다. 다양한 기관에서 수집한 데이터를 이용하다 보니 수집 시점이 모두 달랐는데, 만약 데이터의 시점이 일치했다면 더욱 정확한 분석이 되었을 것이다.

그럼에도 본 연구는 몇몇 시사점을 갖는다. 먼저, 공간 분석과 환경 데이터 위주로 분석했던 기존의 연구들과 달리 사회적인 관점에서 접근한 것과 통계학적으로 문제에 접근했다는 점에서 의미가 있다. 수도권 매립지의 종료 선언과 함께 쓰레기 매립지의 입지 선정에 대한 구체적인 해결책이 필요하다. 하지만, 기존의 공간 분석이나 환경적 데이터로만으로는 합리적인 선정을 할 수 없다. 본 연구에서 입지 선정하고자 했던 시설이 선호 시설이 아니었기 때문에 일반적인

클러스터링을 진행할 수 없어 이상치 탐지 기법으로 접근했으며, SHAP 을 통해 각 변수의 중요도 또한 판단했다. 만약 쓰레기 매립지가 아닌 다른 혐오 시설의 입지를 선정해야 할 경우에도 본 연구에서 사용한 모형과 지표를 참고할 수 있을 것이다.

또한 본 연구는 혐오 시설의 입지 선정에 있어 공공데이터의 활용 및 데이터 기반의 의사결정의 필요성과 접근 방법을 보여준다. 본 연구의 목적은 선호 시설이 아닌 혐오 시설의 입지 선정이었기 때문에 아무리 좋은 품질의 데이터가 있더라도 가공하지 않으면 분석을 시도하기 어렵다. 따라서 본 연구에서는 사회, 환경과 같이 다양한 분야의 데이터를 통합하고 가공하여 더 넓은 시각에서의 분석을 진행하였고, 이를 통해 구체적인 입지를 추천할 수 있었다. 하지만, 앞서 언급했듯이 본 연구에 사용된 데이터의 시점이 모두 다르다는 한계점을 지닌다. 만약 양질의 데이터를 추가로 수집할 수 있고 다른 나라의 데이터를 Auxiliary Data 의 형태로 참고해서 사용한다면 분석의 한계를 극복하고 더 좋은 결과를 도출할 수 있을 것이라고 기대된다.

참고 문헌

김병철, 오상영, 류근호, 2006, 영향력을 고려한 적정입지선정 모델 연구, 한국산학기술학회논문지, p895-900

김판준, 2019, 랜덤포레스트를 이용한 국내 학술지 논문의 자동분류에 관한 연구, 정보관리학회지, p59

김현일, 이연수, 김병현, 2021, 랜덤포레스트 회귀모형을 적용한 도시지역에서의 실시간 침수 예측, 한국수자원학회 논문집, p1121

김현정, 2021, 소하천 재난관리체계 마련을 위한 k-means 기반의 군집화 평가, 부산대학교,p 11

민준영, 1995, 신경망 클러스터리의 성능평가: GLVQ와 k-means알고리즘을 중심으로, 성균관대학교 대학원, p25-26

박성우, 노윤아, 정승민, 황인준, 2021, LSTM을 사용한 SHAP 기반의 설명 가능한 태양광 발전량 예측 기법, ACK 2021 학술대회 논문집, p845-848

박순호, 1997, 농촌지역 쓰레기 매립장 입지 선정에 관한 연구-경상북도 영양군을 사례로, 한국지역지리학회지, p66-80

박환용, 정일훈, 김철중, 2010, 도시공공시설의 적정입지 선정에 관한 연구: 파주시를 중심으로, 국토연구 제 66권, p148-168

배민기, 장병문, 1998, 지리정보체계를 이용한 일반폐기물 매립후보지의 입지선정에 관한 연구, 한국지리정보학화지 1권 2호, p14-25

송용정, 김현덕, 2012, 의사결정나무분석을 이용한 컨테이너 수출입 물동량 예측, 한국항만경제학회지, p198

오재식, 최준호, 2009, 매립장 입지선정의 갈등과 해소방안 - 대구시 쓰레기 매립장 사례를 중심으로-, 영남대학교, p1-112

정상용, 2001, 비위생 생활폐기물 매립지의 문제점과 대책, 한국농공학회지 제 43권 제6호, p.44-53

진승종, 유상철, 김남기, 하윤우, 왕지남, 2022, AutoEncoder/Isolation Forest 알고리즘을 이용한 용접 공정 시계열 데이터 이상탐지 , 한국경영과학회 학술대회논문집, p3

Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu, 1996 ,A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Institute for Computer Science, University of Munich,p230

안재현. 『XAI 설명가능한 인공지능, 인공지능을 해부하다』. 위키북스, 2020

군사시설보호구역협의, Yes, 연천!, <https://www.yeoncheon.go.kr/www/contents.do?key=3284>, 2022.08

포천시(抱川市), 한국민족문화대백과사전,

<http://encykorea.aks.ac.kr/Contents/Item/E0060191>, 2022.08