석사학위논문

빅데이터 분석을 통한 상수도 누수 위험도 예측모델 구축 및 검증에 관한 연구

2021년

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합컨설팅전공

박 진 우

석 사 학 위 논 문
지도교수 정수환

빅데이터 분석을 통한 상수도 누수 위험도 예측모델 구축 및 검증에 관한 연구

2020년 12월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합설팅학과

스마트융합컨설팅전공

박 진 우

석 사 학 위 논 문
지도교수 정수환

빅데이터 분석을 통한 상수도 누수 위험도 예측모델 구축 및 검증에 관한 연구

위 논문을 컨설팅학 석사학위논문으로 제출함

2020년 12월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합설팅학과

스마트융합컨설팅전공

박 진 우

박진우의 컨설팅학 석사학위논문을 인준함

2020년 12월 일

심사위원장 _____(인)

심 사 위 원 _____(인)

심 사 위 원 _____(인)

국 문 초 록

빅데이터 분석을 통한 상수도 누수 위험도 예측모델 구축 및 검증에 관한 연구

한성대학교 지식서비스컨설팅대학원
스 마 트 융 합 컨 설 팅 학 과
스 마 트 융 합 컨 설 팅 전 공
박 진 우

빅데이터 예측 분석 모델이 적용되고 있는 공공 분야 및 민간 분야의 다양한 사례들을 살펴보면 머신러닝 예측 분석 모델의 모형 평가에 의한 성능 검증은 기본적으로 포함하고 있으나 머신러닝 기법에 의해 예측한 분석 결과가 이후 실제로 발생되는 결과 데이터와 잘 맞는지 확인하는 현실적 검증을 한 사례는 생각보다 많지 않다.

따라서 이와 관련하여 범죄율, 요금체납, 영화흥행, 온라인판매, 유동인구수 등을 예측하는 그 분야별 사례들을 살펴보고 현실적 검증을 효과적으로 할 수 있는 방안에 대해서 하나의 〈예측 모델의 현실적 검증 평가방법론〉으로 정립할 수 있도록 시사점 및 발전 방향을 찾아보고자 한다.

【주요어】 빅데이터 분석, 누수 위험도, 머신러닝, 예측 모델, 공공 빅데이터

목 차

표 목 차

그 림 목 차

Ⅰ. 서론

1.1 연구의 배경

상수도 누수 위험도 예측 모델은 유사사업 결과를 바탕으로 표준화된 표준 분석모델을 정립하는데 목적이 있으며 상수관로 및 누수발생 패턴 분석과 함께 상수관로 누수 위험도를 예측하는 분석 기능을 제공한다.

정립된 분석 모델을 향후 다른 기관들에서도 활용할 수 있도록 개방하고 공유하여 지자체 상수관로 데이터 및 누수발생이력 데이터를 기반으로 상수관로 누수 위험도를 분석하여 상수관로의 신설 및 교체 우선순위 대상을 데이터 기반의 의사 결정 체계로 전환시키는 동시에 상수도 관련 누수발생 민원이나 타 기관들의 요구에 객관적인 근거자료로 활용하는데 그 목적이 있다고 할 수 있다.

1.2 연구의 목적

빅데이터 분석 모델 측면에서 여러 분야별로 사례들을 살펴보면 분석모델의 모형 평가에 의한 성능 검증은 기본적으로 하고 있으나 머신러닝 기법에 의해 예측한 분석 결과가 발생 결과와 잘 맞는지 확인하는 현실적 검증을 한 사례는 생각보다 그렇게 많지 않아 이에 대해 그 분야별 사례들까지 한번 살펴보고 시사점 및 향후 발전 방향에 대해 짚어보고자 한다.

빅데이터 분석 모델에 있어서 모형 평가에 의한 성능 검증과 머신러닝 예측 결과에 대한 사후 현실적 검증 사례들을 살펴보고 실제 이 둘 간의 관계에 있어서 어떤 시사점 및 발전방향을 도출할 수 있는지 알아보고자 한다.

Ⅱ. 선행연구

2.1 빅데이터 분석 조사

2.1.1 빅데이터 개념

빅데이터는 기존의 데이터들보다 너무 양이 많고 범위가 넓어서 기존의 방법이나 도구들로는 수집, 저장, 분석이 어려운 정형 데이터 또는 비정형 데이터들을 의미한다.

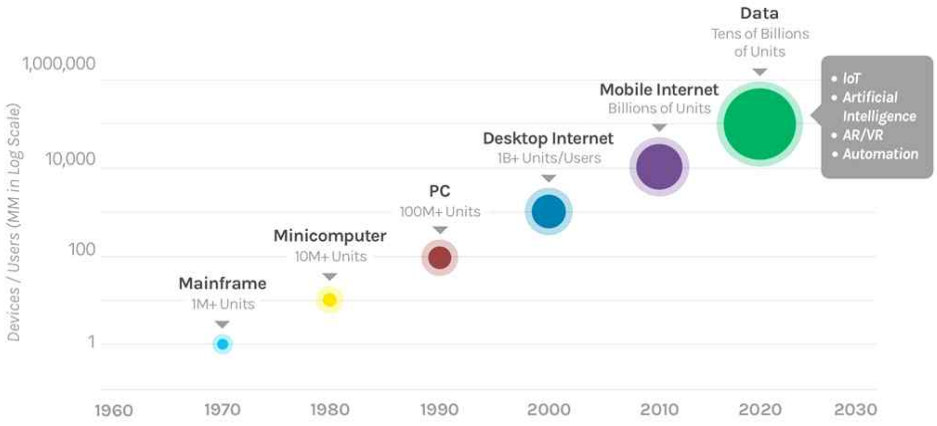
빅데이터의 특징으로는 대규모(Volume), 속도(Velocity), 다양성(Variety) 및 분석가치(Value)를 들 수 있다. 데이터의 크기는 수십 테라 또는 수십 페타바이트 이상의 데이터 속성을 의미한다.

[그림 2-1] 빅데이터의 4가지 요소

앞으로의 시대는 디지털의 시대(2000년대)와 연결의 시대(2010년대)를 넘어서 데이터의 시대(2020년대), 즉 빅데이터, 인공지능, IOT 융합의 시대에 진입했다고 할 수 있다.

[그림 2-2] 10년 주기 시대의 변화

다양한 기기를 인터넷에 연결 작업이 완료되고 연결된 장치가 생성하는 데이터 수집과 활용도로 초점이 옮겨졌다. 즉, 요즈음 시대에 많은 소비자의 선택이나 비즈니스 의사결정들이 데이터 기반으로 진행되고 있음을 말한다.



[그림 2-3] Data Decade defined by Morgan Stanley

2020년 현재 '데이터는 잠을 자지 않는다'(Data Never Sleeps 8.0, 출처: https://www.domo.com/learn/data-never-sleeps-8)라는 이름으로 정리된 이 인포그래픽에는 2020년 기준으로 매 1분간 인터넷의 주요 서비스에서 일어나는 일들을 보여준다.

대략 지난 1분 사이 인스타그램에서는 347,222개의 스토리가 만들어졌고 유튜브에는 500시간 분량의 동영상이 업로드되고, 트위터에는 319명의 새 사용자가 생겼고 페이스북에서는 15만 개 이상의 메시지가 공유됐다. 틱톡 앱은 2,700여 번 설치됐고, 스포티파이에는 28곡이 새로 올라갔으며 넷플릭스에선 40만 시간이 넘는 분량의 동영상이 재생됐다고 한다.

[그림 2-4] 1분간 인터넷의 주요 서비스에서 일어나는 일들

이렇게 우리 주변에는 데이터가 넘쳐나는데 표현그대로 빅데이터 시대에 머신러닝이 본격적으로 주목받기 시작했다. 데이터를 보유한 기업이나 공공기관에서 추진해볼 수 있는 가장 효과적인 접근 방식 중 하나이기도 하며 이제 모든 기업과 공공기관은 비즈니스 혁신을 위해 머신러닝이 필수적으로 필요하게 되었다.

2.1.2 빅데이터 분석 기법

빅데이터 분석 컨설팅 시장에서는 다양한 분석 기법들이 사용되고 있는데 그 가운데 머신러닝(Machine Learning)은 인공지능(Artificial Intelligence)의 하나의 분야로써 기계가 대용량의 데이터로부터 스스로 지식이나 패턴을 찾아 학습하고 이를 통해 예측 활동을 수행하는 것이다.

머신러닝 예측기법을 적용하기 위해 일반적인 프로세스는 다음과 같다.

[그림 2-5] 머신러닝 일반적인 프로세스

첫째, 문제 정의 및 분류 단계에서는 비즈니스 요청사항에 대해 문제를 해결하거나, 분류할 수 있도록 정의하는 것이 필요하다.

둘째, 데이터 병합 단계에서는 데이터가 어디 있는지 식별하고, 필요하다면 병합 또는 분리한다. 머신러닝에 사용하는 데이터들은 다양한 곳에 존재하며

직접 고객사에게 요청해서 수급 받거나 직접 개방 공공 데이터들을 수급할 수 있으며, 민간 판매용 데이터를 구매해야 될 수도 있다.

셋째, 데이터 처리 단계에서는 훈련 및 검증 데이터 세트를 만들어 데이터 변환, 정규화 및 클랜징과 같은 데이터 전처리 프로세싱을 하는 단계이다.

넷째, 머신러닝 모델링 단계에서는 데이터 분류를 위해 클러스터링을 할 것인지, 정확한 예측 값을 얻기 위해 훈련할 것인지 머신러닝 알고리즘을 선택하는 단계이다.

참고로 머신러닝에는 수많은 알고리즘이 있으며, 다양한 환경에서 다양한 목적으로 결과값을 얻도록 한다.

다섯째, 모델 검증 및 실행 단계어서는 예측한 결과 값이 맞는지 검증하고, 머신러닝 모델들과 알고리즘들을 실행하는 플랫폼을 판별하는 단계이다.

또한 머신러닝 루틴을 구성하고, 그 결과값을 조율하며 정제한다.

여섯째, 모델 배포 단계에서는 머신러닝으로 처리한 데이터 값을 비즈니스 의사결정에 활용할 수 있도록 확인하거나, 또 다른 어플리케이션, 시스템에 다시 입력 데이터로 활용하도록 한다.

마지막으로 유지보수 및 운영관리 단계어서는 데이터 사이언티스트, IT 운영자 등과 함께 배포 및 유지보수를 자동화할 수 있는데 이러한 운영 관리를 MLOps (Machine Learning Operations)라고 부르며 요즘 이를 위해 머신러닝 전담팀을 별도로 꾸리는 IT 부서가 많이 생기고 있는 추세이다.

머신러닝 개발에 많이 사용되고 있는 언어는 R과 Python이다.

R은 1만개 이상이 패키지로 구성된 강력한 오픈소스 통계분석 언어이다.

[그림 2-6] R언어의 소개 및 특징

또한, Python은 쉽고 가장 강력한 머신러닝 개발 언어이다.

[그림 2-7] Python 언어의 소개 및 특징

머신러닝 알고리즘은 다양한 환경에서 매우 다양한 알고리즘들이 사용되고 있는데 여기서는 본 연구와 연관된 알고리즘 위주로 살펴보기로 한다.

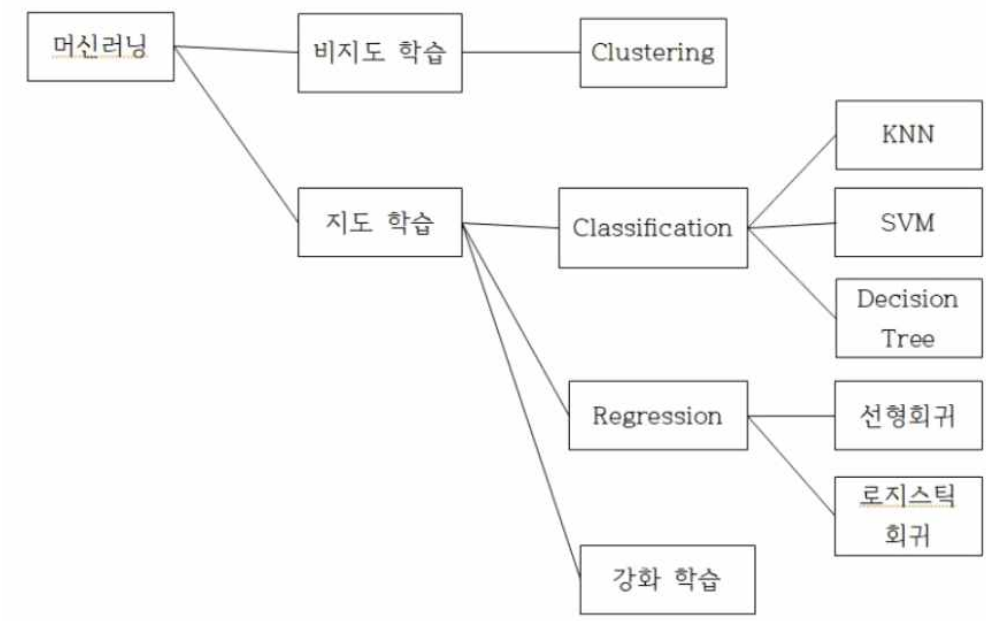
머신러닝은 인공지능의 한 분야로 기존의 데이터를 기반으로 학습을 통해 예측을 수행하는 시스템 및 그 알고리즘을 연구하는 기술이다. 머신러닝 분야의 알고리즘들은 입력 데이터들을 기반으로 예측이나 의사결정을 도출하기 위해 맞춤형 모델을 구축하는 방식을 취한다.

머신러닝 알고리즘은 아래와 같이 크게 세 가지로 분류하고 있다.

[표 2-1] 머신러닝 알고리즘 유형

구 분	예측 기법
지도 학습 (supervised)	입력과 이에 대응하는 출력(인간 제공)을 매핑(mapping)하는 함수를 학습하는 과정
비지도 학습 unsupervised)	출력 없이 입력만으로 모델을 구축하여 학습. 데이터마이닝의 대부분의 기법이 이에 해당
강화 학습 (reinforcement)	컴퓨터가 주어진 여건에서 최적의 판단을 통한 행동을 하도록 선택하는 학습 방법

앞에서 말한 지도 학습과 비지도 학습에는 다음과 같은 종류가 있다.

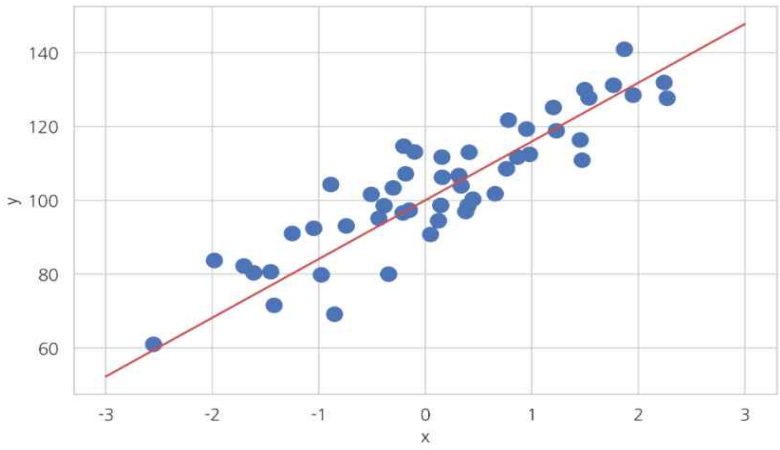


[그림 2-8] 지도 학습과 비지도 학습

먼저 회귀 분석(regression analysis) 방법론에 대한 설명이다.

□ 회귀 분석은 데이터를 바탕으로 어떤 현상의 원인이 되는 변수들을 독립 변수로 보고, 그 결과를 종속 변수로 보며 독립 변수와 종속 변수사이의 모델을 구한 뒤 모형의 적합도를 측정하여 최적의 모형을 구해내는 분석 방법이다.

□ 종속 변수가 하나이고 독립변수도 하나인 경우를 단순 회귀 분석(simple regression analysis), 종속변수가 하나이고 독립변수가 여러개인 경우를 다중 회귀 분석(multiple regression analysis)이라고 한다.

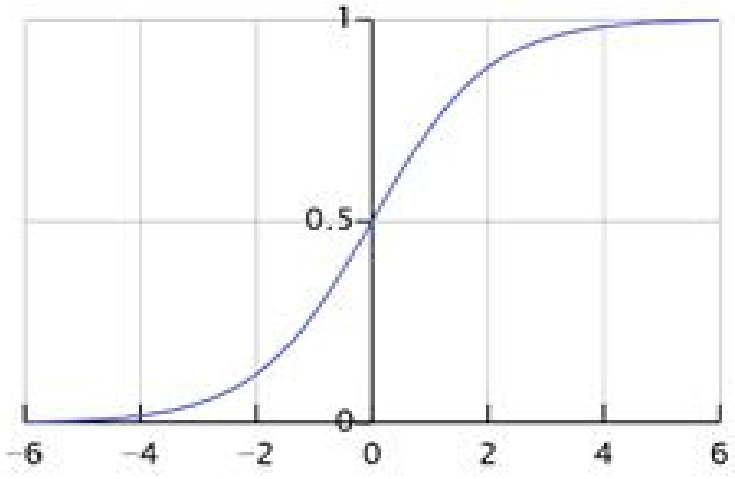


[그림 2-9] 회귀(Regression) 함수

다음은 로지스틱 회귀(Logistic Regression) 방법론에 대한 설명이다.

□ 로지스틱 회귀는 독립 변수의 선형결합으로 종속 변수를 설명하는 선형 회귀와 유사한 모델이며 선형회귀와는 다르게 종속변수가 범주형 변수일 때 사용된다.

□ 특징으로는 계산비용이 적고 구현하기 쉽다는 장점이 있으며 아래의 그림과 같이 연속이고 증가함수이다.



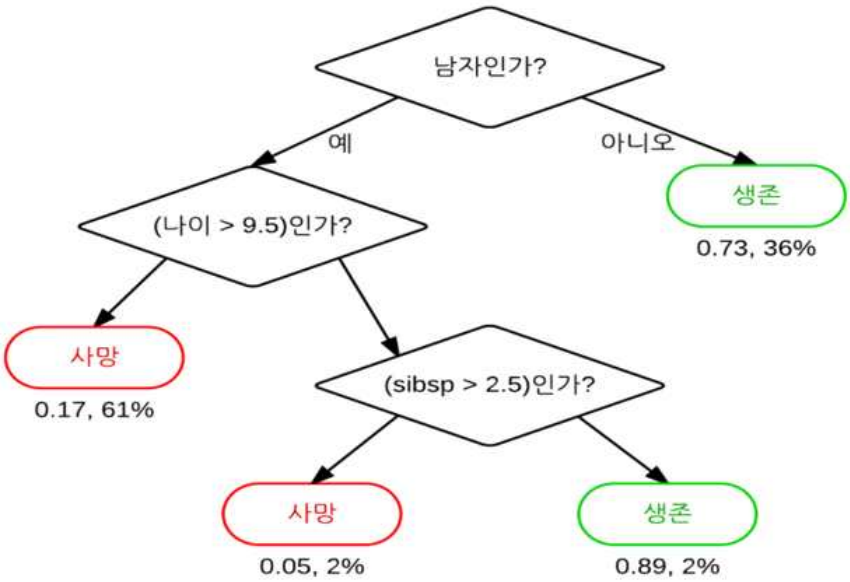
[그림 2-10] 로지스틱 회귀(Logistic Regression) 함수

다음은 의사결정나무(Decision Tree) 방법론에 대한 설명이다.

□ 의사결정나무(Decision Tree)는 의사결정규칙(decision rule)을 트리 방식으로 구조화하여 관심대상 집단을 여러개 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법의 하나이다.

□ 의사결정나무는 주어진 데이터를 바탕으로 종속변수와 독립변수와의 관계를 통해 나무를 생성하며, 각 노드는 각각의 독립 변수 및 조건을 가지며, 마지막의 잎 노드는 종속 변수의 클래스와 확률 값을 가진다.

□ 출력변수가 연속형 회귀모형에서는 예측력이 낮아지며, 복잡한 나무모형일수록 예측력이 저하된다는 단점이 있다.



[그림 2-11] 의사결정나무(Decision Tree) 방법론

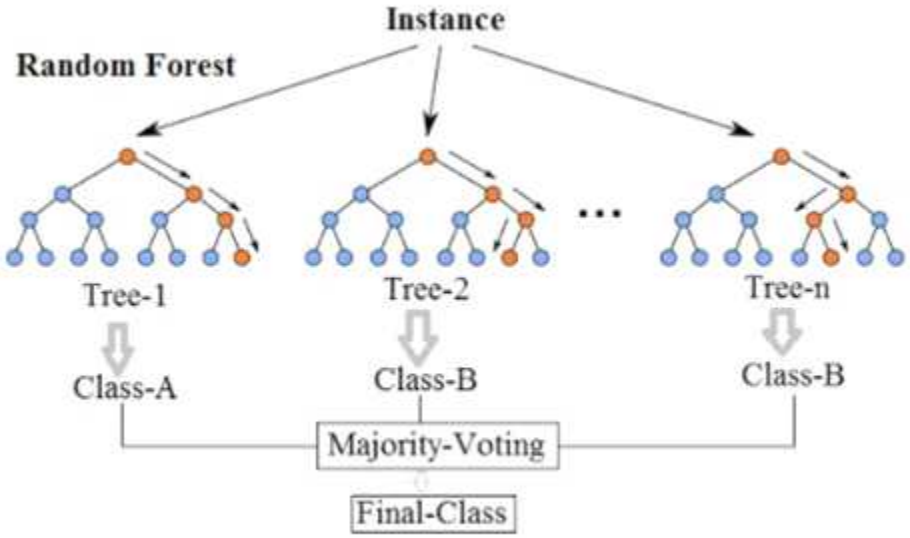
다음은 랜덤포레스트(Random Forest) 방법론에 대한 설명이다.

□ 랜덤포레스트(Random Forest)는 분류 분석의 한 방법론으로 분석하고자 하는 대상들이 두개의 집단으로 나누어진 경우에 개별 관측 데이터들이 어느 집단으로 분류될 수 있는가를 분석하고 예측하는 분석기법이다.

□ 또한 랜덤 포레스트는 앙상블(Ensemble) 학습 기반의 모델로써

앙상블 학습은 데이터들로부터 여러 개의 모델을 동시에 학습한 다음, 예측 단계에서 여러 모델의 예측 결과들을 종합해 정확도를 높이는 기법이다.

□ 일반적으로 랜덤 포레스트(Random Forest)는 성능이 뛰어나며 의사결정 나무 한개가 아니라 여러 개의 나무를 사용해 과적합 문제를 피한다.



[그림 2-12] 랜덤포레스트(Random Forest) 프로세스

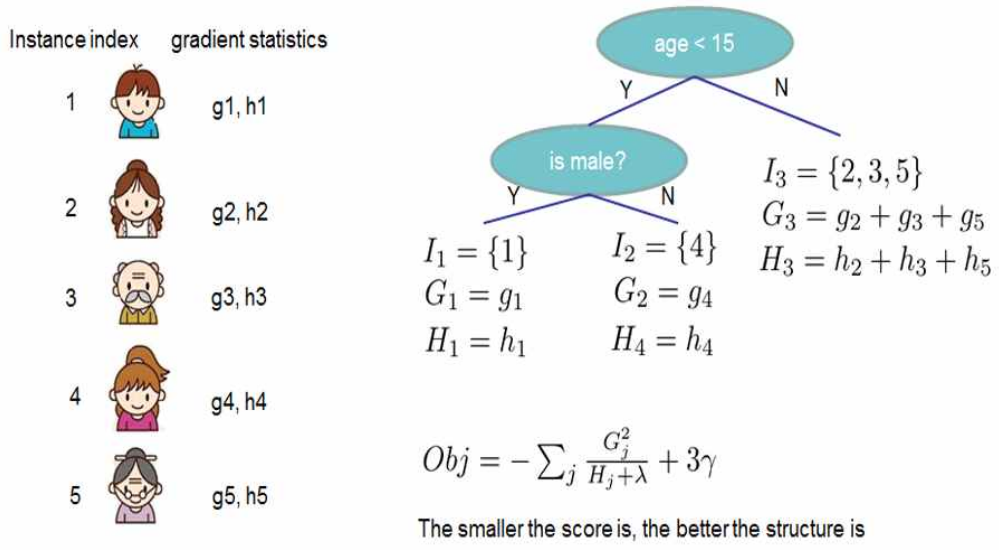
다음은 XGBOOST(Extreme Gradient Boosting) 방법론에 대한 설명이다.

□ 여러 개의 의사결정나무를 조합한 앙상블(Ensemble) 모델로 이전 의사결정나무의 오차를 보완하는 방식으로 순차적 의사결정나무를 만든다.

□ 서브샘플링을 통해 주어진 데이터를 분할하여 반복학습을 통해 과적합의 위험도를 낮추는 특징을 가진다.

□ 병렬 처리를 하기 때문에 타 부스팅 모델에 비해 수행시간이 짧다는 장점이 있다.

□ Greedy-algorithm(탐욕 알고리즘 - 현 상태에서의 최선의 방법 선택)을 사용한 자동 가지치기로 인해 과적합이 잘 일어나지 않는다.



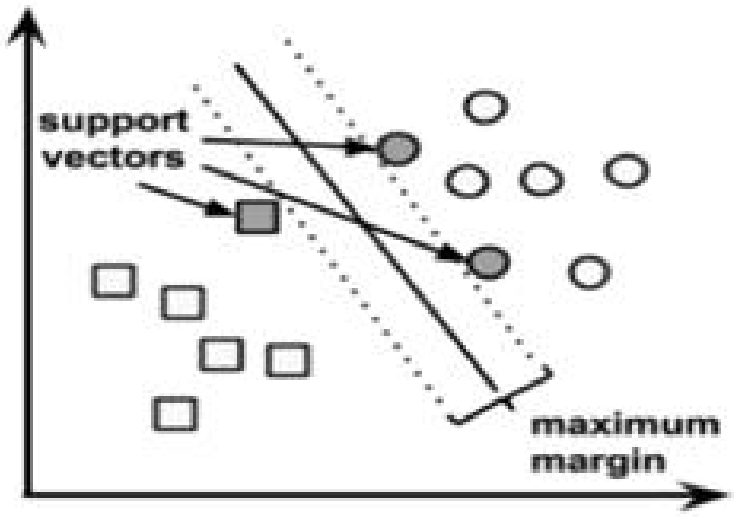
[그림 2-13] XGBOOST(Extreme Gradient Boosting) 프로세스

다음은 SVM(Support Vector Machine) 방법론에 대한 설명이다.

□ SVM은 분류 과제에 사용할 수 있는 지도학습 모델이고 써포트 벡터(support vectors)를 사용하여 결정 경계(Decision Boundary)를 정의하고, 분류되지 않은 점들은 해당 결정 경계와 비교해서 분류한다.

□ 분류문제나 예측문제에 동시에 사용하는 것이 가능하며 신경망기법에 비해 과적합도가 낮으며 예측의 정확도가 높은 반면 사용하기 용이하다.

□ 단점은 Kernel 모델과 파라미터를 조절하기 위한 테스트 과정을 여러 번 수행해야 최적화된 모형을 개발할 수 있고 모형 구축 기간이 오래 소요되며 결과에 대한 설명력이 낮다.



[그림 2-14] SVM(Support Vector Machine) 개념

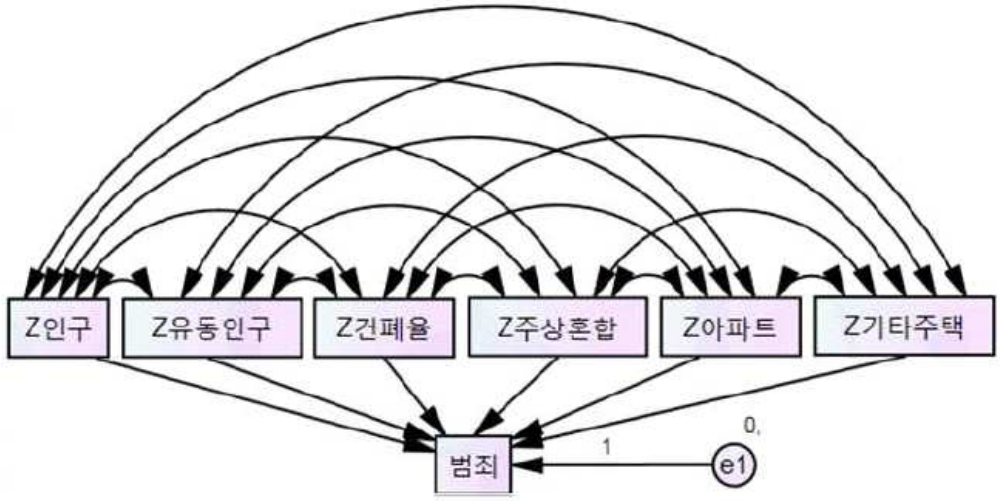
이상으로 실제 필드에서 많이 사용하고 있는 다양한 머신러닝 알고리즘들을 간단히 살펴 보았다.

2.2 선행 연구 조사

2.2.1 사례1: 범죄율 예측분야

〈공간 빅데이터를 활용한 범죄 발생 위험지역 예측 모형 구축〉 사례이다.

사례1에서 보면 범죄에 관한 연구들은 범죄 발생 영향요인 분석, 범죄자 예측, 범죄 위험지역 예측, 범죄 발생 분석 등 다양한 분야에서 진행되어왔다. 여기서는 범죄 발생 영향요인 및 범죄발생에 대한 연구가 진행되었다.



[그림 2-15] 범죄율 예측 연구모형

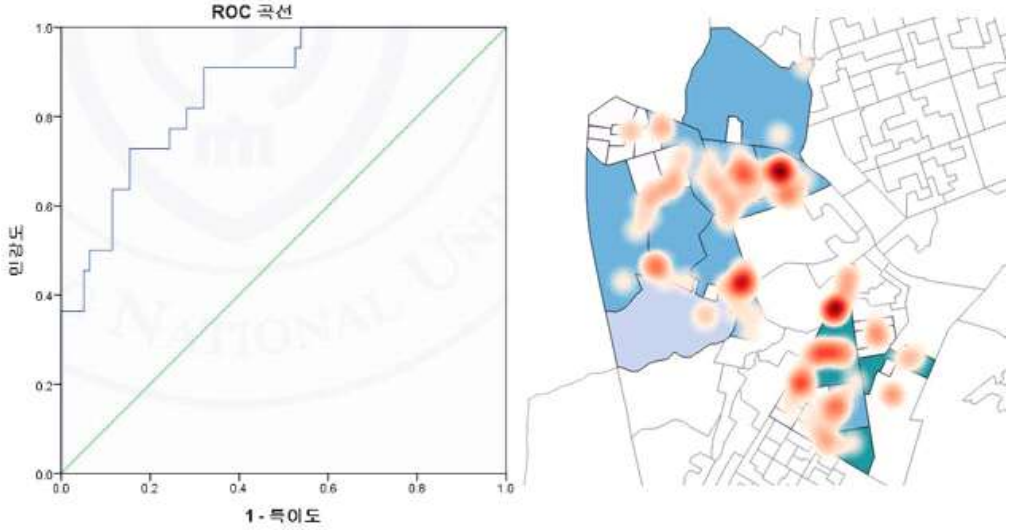
다중회귀분석을 통하여 도출된 6개의 변수 즉, (주상복합, 건폐율, 아파트, 인구, 유동인구, 기타주택)을 기반으로 AMOS 소프트웨어를 활용하여 베이지안 추론 수행, 블록당 발생범죄건수 평균값 기준으로 적음과 높음으로 구분하였다.

$$Y_i = 0.149X_1 + 0.042X_2 + 0.089X_3 + 0.091X_4 - 0.195X_5 - 0.027X_6 + e$$

Y_i : 범죄 발생 위험성
X_1 : 평균 주거 인구
X_2 : 평균 유동인구
X_3 : 평균 건폐율
X_4 : 주상 혼합 주택 수
X_5 : 아파트 수
X_6 : 기타 주택(주택 이외 거처) 수
e : 오차항

[그림 2-16] 범죄율 예측 다중회귀식

예측된 결과의 검증을 위하여 실제로 범죄가 발생한 지점의 밀도 분석결과와 중첩해서 비교하였으며, 정량적인 검증을 위하여 ROC곡선을 이용하였다. 그 결과 아래 그림과 같이 AUC값이 86.1%의 정확도를 나타냈다.



[그림 2-17] 범죄율 예측모형 검증 및 비교

2.2.2 사례2: 상수도 체납 예측

〈빅데이터 기반 체납 수용가 예측 모델 개발〉 연구사례이다.

본 연구에서는 수용가의 체납 발생에 영향을 줄 것으로 예상되는 71종의 데이터들에 대해서는 데이터 전처리를 통해 체납예측 모델을 만드는데 사용하고, 체납예측 모델 검증을 위해 K사가 N지자체 대상으로 시범운영을 통해 체납 예측 정도를 검증하였다.

N지자체 데이터를 기반으로 성능 비교를 해보니 아래 표와 같이 의사결정트리가 로지스틱회귀보다 조금 더 정확한 수용가 납부 패턴을 찾아 낼 수 있음을 확인하였다.

논산시	의사결정트리	로지스틱회귀
실험데이터 1	76.29%	75.65%
실험데이터 2	76.32%	75.61%
실험데이터 3	76.29%	75.63%
실험데이터 4	76.18%	75.69%
실험데이터 5	76.39%	75.69%
평균(±표준편차)	76.294%(±0.075)	75.654%(±0.035)

[표 2-2] 알고리즘 성능 비교 결과

규칙	규칙 조건	가능성 (%)	실험 번호
체납-규칙1	(1) 전월연속체납건수 > 0 AND (2) 고지방법 = 일반지로 AND (3) 전월수납방법건수_인터넷지로 <=0 AND (4) 최근체납비율 > 0.67	92.00	1
		-	2
		-	3
		92.13	4
		-	5

[그림 2-18] 당월 납부유형 예측 규칙

이번 연구에서 개발한 요금 체납 수용가 사전 예측 결과와 실제 납부 현황의 비교 결과는 아래 표와 같다.

[표 2-3] 예측 결과와 실제 납부 결과 비교

고지년월	체납 예측(A)	실제 체납(B)	정확도(B/A)
2017년 7월	2,685	1,719	64.02%
2017년 8월	2,683	1,713	63.85%
2017년 9월	3,267	1,872	57.30%

N지자체를 대상으로 하여 실제 업무에 적용해 운영해 본 결과 기대한 예측치 보다는 낮게 나왔는데 본 서비스를 운영하는 과정에서 체납에 영향을 주는 변수들이 새로 도출 가능하고 모델의 업그레이드도 필요하다고 하겠다.

2.2.3 사례3: 영화 흥행 예측

본 연구는 영화 흥행 예측을 위하여 2011년에서 2017년 11월까지 국내 극장에서 개봉한 한국 국적 일반 상업영화들의 데이터를 수집, 정제하였다. 회귀분석을 통해 흥행에 유의미한 영향을 미치는 요인을 분석하고, 그를 바탕으로 나이브 베이즈 분류(Naïve Bayes Classification)를 통해 2017년 12월~2018년 4월에 개봉한 영화의 흥행을 예측, 실제 결과와 비교하였다.

```
Call:
lm(formula = total_people ~ grade)

Residuals:
     Min       1Q   Median       3Q      Max
-2093962  -164568  -164437  -163120 15519719

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1840814     185384   9.930  < 2e-16 ***
grade15        253149     218853   1.157    0.248
gradeAll     -1639396     301300  -5.441 6.69e-08 ***
grade19      -1676245     197944  -8.468  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1729000 on 978 degrees of freedom
Multiple R-squared:  0.2001,    Adjusted R-squared:  0.1976
F-statistic: 81.55 on 3 and 978 DF,  p-value: < 2.2e-16
```

$$\text{total\_people} = 1840814 + (253149 \times \text{grade15}) + (-1639396 \times \text{gradeAll}) + (-1676245 \times \text{grade19})$$

[그림 2-19] 회귀분석 결과 및 추정 회귀식

나이브 베이즈 분류는 주로 스팸 메일 분류에 사용되어 왔으며 영화 흥행 예측 분야에서도 많이 사용되는 방법으로 분류 학습에서 매우 정확한 결과를 보여준다.

$$n = \arg\max_n P(A_n) \prod_{i=1}^{k} p(x_i|c_i)$$

[그림 2-20] 나이브 베이즈 분류식

분석 대상은 2011년에서 2017년 11월 사이 개봉한 1,118편 한국영화이다. 영화관입장권통합전산망 대상 영화의 관객 수, 스크린 수, 배우, 감독, 배급사 등의 데이터를 수집하고 정제하였고 흥행에 영향을 미치는 요인들을 분석한 후 2017년 12월 1일에서 2018년 4월 30일까지 국내 개봉한 167편의 한국 상업영화 배우와 감독 등의 정보가 정확한 영화 중에서 영화 56편을 선정하여 한국영화의 흥행을 예측하였다.

[표 2-4] 흥행 예측 결과

제목	실제 흥행 여부	예측된 흥행 여부
신과함께-죄와 벌	Y	Y
1987	Y	N
그것만이 내 세상	Y	N
곤지암	Y	N
지금 만나러 갑니다	N	N
조선명탐정: 흡혈괴마의 비밀	N	N
리틀 포레스트	N	N
골든슬럼버	N	N
공조	N	Y
사라진 밤	N	N
염력	N	N
7년의 밤	N	N
강철비	N	N
흥부: 금로 세상을 바꾼 자	N	N
치즈인더트랩	N	N

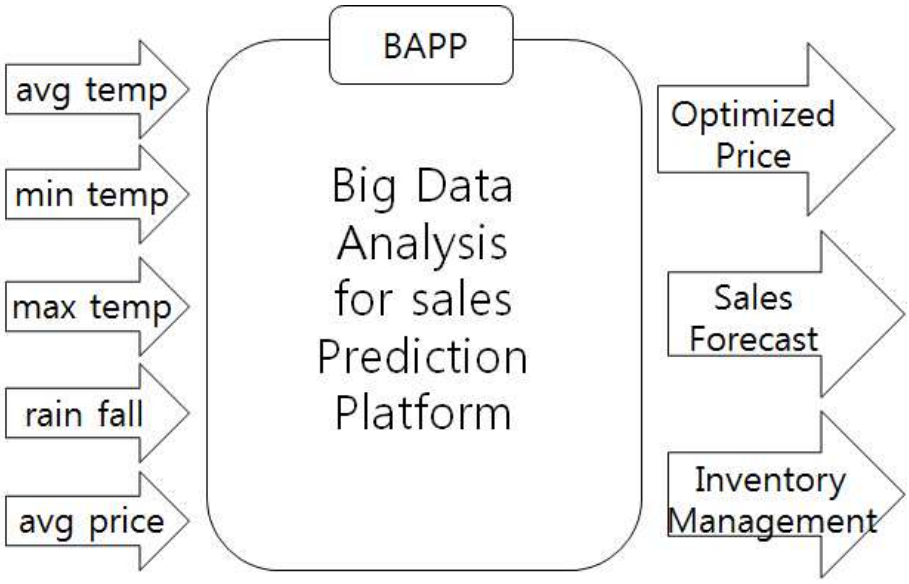
본 논문에서는 총 관객수가 2,650,000명 이상일 경우 흥행이라고 판단하였는데 분석 결과 56편 영화 가운데 51편이 실제로 흥행 여부와 동일하게 나타났는데 분류 정확도(Accuracy)는 91% 정도로 나타났다.

2.2.4 사례4: 온라인 판매 예측

〈빅데이터 분석을 활용한 온라인 판매 수요 예측〉 연구사례이다.

본 논문에선, 2014년부터 2018년까지 온라인쇼핑몰 'A'에서 판매된 약 18만개 반팔 티셔츠의 판매량과 약 4.1만 아우터 판매량 그리고 기상청의 하루 평균 기온 데이터를 수집하여 기온변화에 따른 반팔 티셔츠와 아우터웨어의 판매량을 분석하였다.

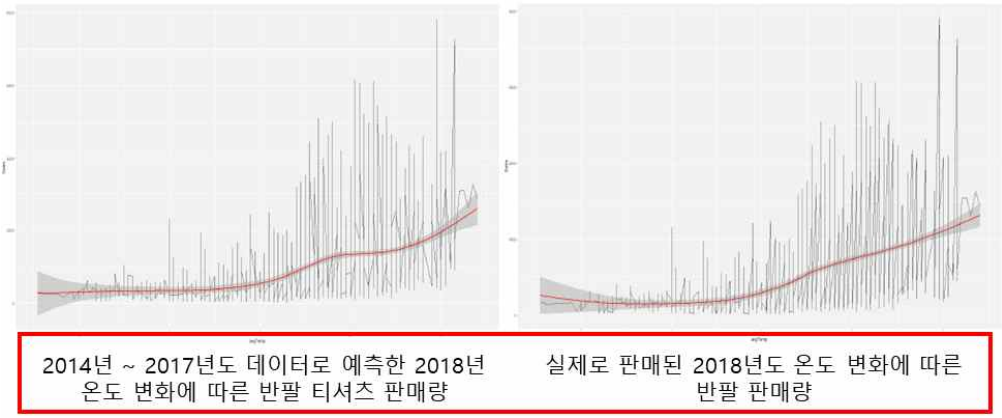
평균 온도, 평균 가격 변화에 따른 판매량에 대한 분석이며 반팔 티셔츠, 아우터 웨어(Outer Wear) 2가지 제품에 대한 분석을 다뤘다.



[그림 2-21] BAPP(Big Data Analysis for sales Prediction Platform) 시스템 개요

BAPP는 Date:날짜, avg Temp:평균 온도, min Temp:최소 온도, max Temp:최고 온도, rain Fall:강수량, avg Price:평균 가격, Sales:판매량의 데이터를 통해 수요 예측 시스템을 구성한다.

2014년부터 2017년도까지의 데이터를 분석하여 2018년도 판매량을 예측하였으며 반팔 티셔츠와 아우터웨어의 예측값과 실제 판매량의 오차율은 각각 ±1.5%, ±8%를 나타냈다.



[그림 2-22] 예측 판매량과 실제 판매량 비교

본 연구를 통해 장기간에 걸쳐 데이터가 누적될수록 기간적인 더 정확한 예측이 가능할 것으로 기대되며 이를 활용하여 판매량을 예측하고 효율적인 재고관리가 가능할 것으로 기대된다.

2.2.5 사례5: 유동 인구 예측

몇 가지 머신러닝 기법을 활용하여 기후, 날씨, 시간 정보 등 다양한 변수를 활용하여 지역별 유동 인구수를 예측하는 모델을 만들고자 하였다.

예측 모델은 다중회귀분석, 의사결정나무 등 머신러닝 기법을 활용하여 모델링 한 후 성능 비교를 진행하였다.

학습용 데이터는 2017년 1월에서 2018년 8월까지의 데이터를 활용하였고, 테스트용 데이터는 2018년 9월 데이터를 활용하였다.

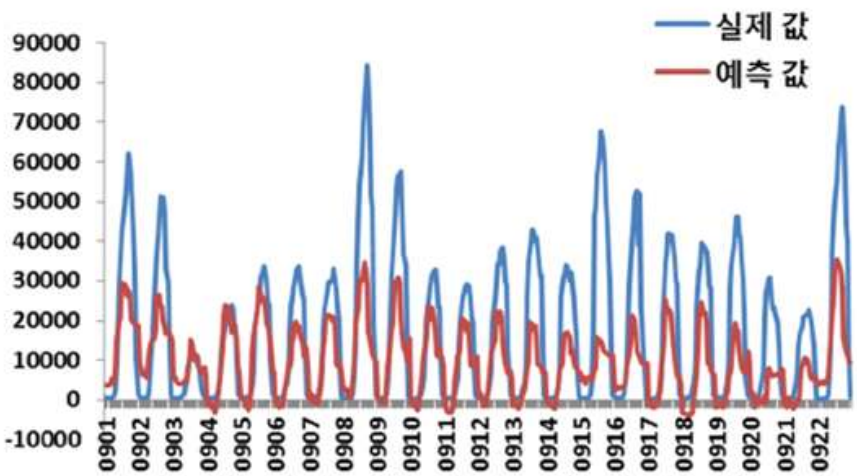
예측 모델을 학습시키기 위한 종속 변수는 다음과 같다.

[표 2-5] 모델에 사용된 독립 변수

변수명	변수 설명	type	비고
mm	월	num	
hh	시간	num	
rain	강수	num	
rain_type	강수형태	factor	없음, 비, 비/눈, 눈
temp	기온	num	
lighting	뇌전	num	
humidity	습도	num	
wind	풍속	num	
wind_direction	풍향	num	
sky_type	하늘상태	factor	맑음, 구름조금, 구름많음, 흐림
mise	미세먼지	num	
week_nm	요일	factor	월, 화, 수, 목, 금, 토, 일
hday_yn	휴일/평일	factor	휴일, 평일

회귀분석 모델링 시행 결과, Adjusted R-squared 이 0.5477 되어 54.8%의 설명력을 가지고, p-value 0.05 이하로 유의미한 모델이 생성 되었다.

의사결정나무 모델의 경우 약 75%의 정확률을 보였고 시계열 예측 모델 경우는 약 67%의 정확률을 나타냈다.



[그림 2-23] 다중회귀분석 예측 수치와 실제 수치 비교

인구수 예측 성능은 다음과 같이 의사결정나무가 가장 좋은 성능을 보였다.

[표 2-6] 모델별 예측 정확률 비교

알고리즘	정확률
회귀 분석	60%
의사 결정 나무	75%
시계열 분석	67%

인구수 예측 모델에 큰 영향을 주는 변수는 요일, 시간, 휴일여부 등으로 나타났고 이 패턴은 결과에 잘 반영하였지만, 날씨나 기후 변수만으로 인구수 예측을 하기에는 분명 한계점도 있었다. 따라서 추가적인 변수들을 확보하고, 최적의 변수 생성 및 분석 기법을 통해 모델의 성능 향상을 가져온다면 활용 분야는 매우 많을 것으로 기대한다.

2.3 선행 연구 요약

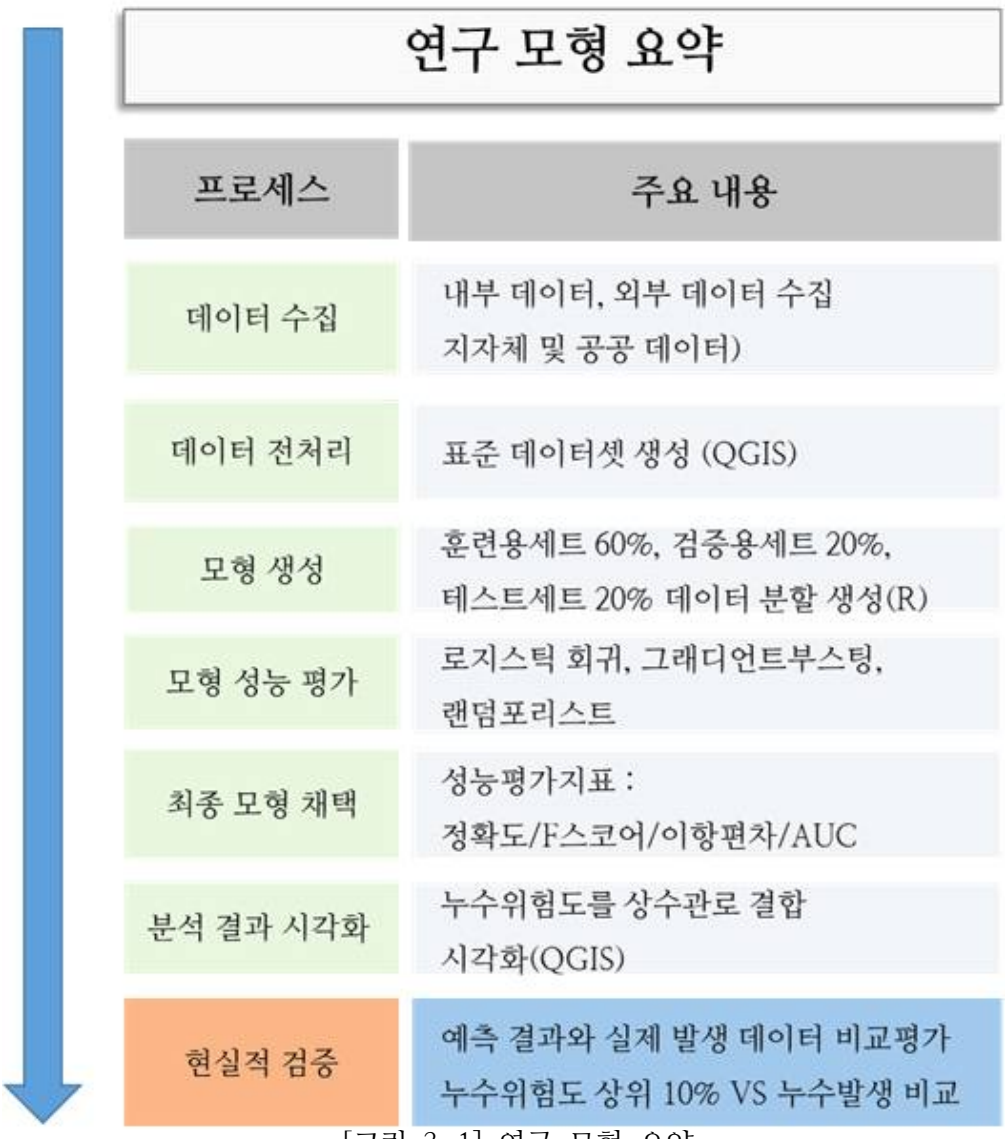
앞에서 살펴본 선행연구들을 요약하면 다음과 같다.

[표 2-7] 기존 선행연구 요약

선행 연구	적용 분야	예측 기법	통계적 성능 지표	현실적 검증 항목
사례1	범죄율 예측	다중회귀분석(6개의 주요 변수 도출) 기반으로 AMOS 베이지안 추론 수행	ROC 곡선 AUC값이 0.861	집계구당 범죄발생 예측건수 VS 블록당 범죄 발생 건수(관할 경찰서 범죄관리대장) 비교
사례2	체납 예측	의사결정트리 및 로지스틱회귀	의사결정트리(76.294%, ±0.075, 로지스틱회귀(75.654%, ±0.035)	체납 예측 결과와 실제 납부 현황 데이터 비교
사례3	영화흥행 예측	회귀분석(유의미한 변수) 및 나이브 베이즈 분류	나이브 베이즈 분류의 정확도(Accuracy)는 91%	총 관객 수가 2,650,000명 이상 흥행 판단, 56편 중 51편의 영화가 실제 흥행 여부와 동일
사례4	온라인 판매 예측	시계열 예측방법 ARIMA 모형	반팔 티셔츠와 아우터웨어의 예측값과 실제 판매량의 오차율은 각각 ±1.5%, ±8%	2018년도 판매량 예측 VS 실제 2018년 판매량 비교
사례5	유동인구수 예측	다중 회귀 분석, 의사결정나무, 시계열 분석	회귀분석 60%, 의사결정나무 75%, 시계열 예측 67%	인구수 예측값 VS 유원지 시간대별 인구 수

Ⅲ. 연구방법

본 연구에서 제시하는 연구모형을 도식화하면 다음과 같다.



[그림 3-1] 연구 모형 요약

3.1 상수도 누수 위험도 예측모델 개요

3.1.1 상수도 분석 목적 및 배경

상수도 누수는 도로 대형 포트홀, 교통마비, 상가영업 중지, 주민생활 불편 등으로 이어지고 막대한 복구 예산이 소요된다.

광주 도심에서 상수도관이 터져 인근 상가 20여 곳이 단수된 사례
광주시 상수도사업본부는 20년 넘은 수도관이 노후화돼 파손된 것으로 보고 정확한 누수 경위를 조사

경남 창원시내 대형 상수도관의 노후화로 인해 800mm 상수도관 파열후 아스팔트 도로가 내려 앉고 주변 100m가 물에 잠겨 복구 전까지 교통이 마비되어 버린 사례

[그림 3-2] 상수도 누수 피해복구 기사

또한, 매년 상수도 누수로 인해 낭비되는 예산은 매년 증가하고 있다.

전국 시도별 수돗물 누수율 현황

전국 지자체 누수율 심각 사례

[그림 3-3] 상수도 전국 누수율

3.1.2 상수도 표준분석모델 개요

상수도 누수 위험도 예측 모델은 상수관로 및 누수 패턴 분석, 상수관로 누수 위험도 분석 같은 지자체 상수도 관련하여 핵심 기능을 제공한다.

[그림 3-4] 상수도 표준분석모델 개요

상수도관 노후 문제와 교체 예산 부족으로 인한 효율적인 누수관리 및 개선방안의 해결책으로 노후 수도관 교체 우선 지역 선정을 위한 누수 예측 모델을 개발하여 보다 나은 의사결정 및 정책에 반영할 수 있도록 하는데 그 목적이 있다.

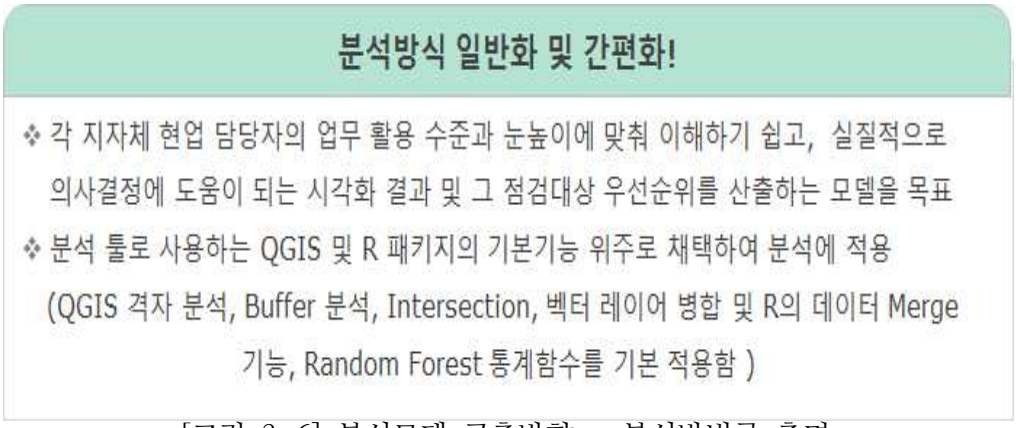
3.1.3 상수도 표준분석모델 방향

상수도 누수 위험도 표준분석모델은 데이터 수급측면, 분석방법론 측면 및 모델 확산화 측면에서 각각 다음의 특장점을 가지고 있다.

상수도 누수와 관련된 가장 핵심적인 요인 위주로 핵심 변수화를 통해 수급데이터를 최소화하고, 개방 공공 데이터들을 최대한 활용하였다.

[그림 3-5] 분석모델 구축방향 - 데이터수급 측면

모델을 사용하는 분석 사용자의 현실에 맞춰 사용되는 분석 기능과 함께 분석 프로세스, 분석 결과, 시각화 난이도 등을 설계하였다.



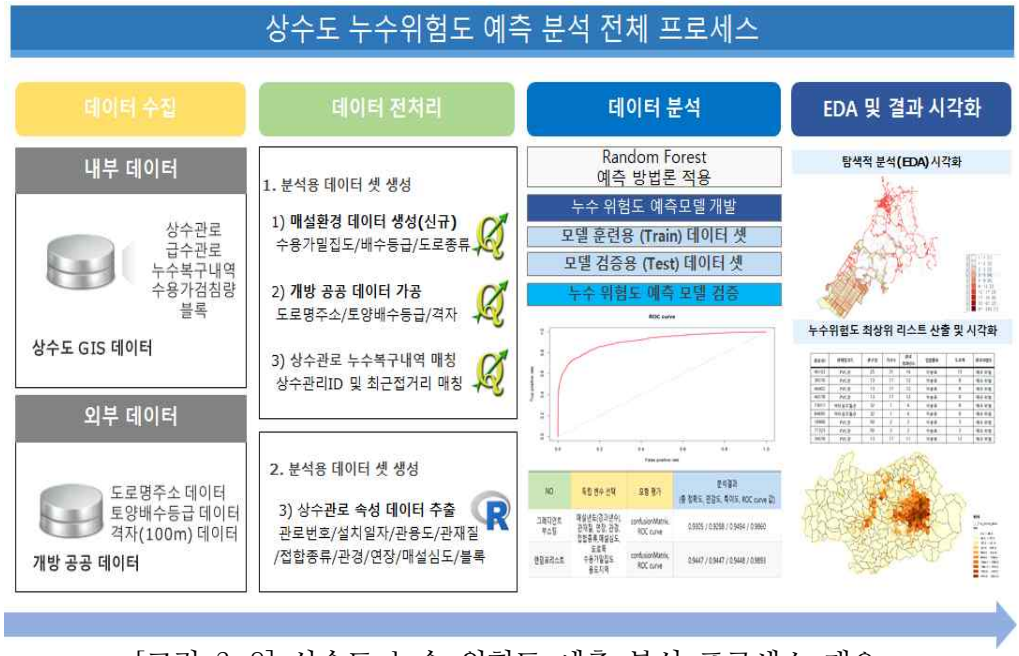
[그림 3-6] 분석모델 구축방향 - 분석방법론 측면

분석 Tool의 비용, 확장성 등을 고려하여 오픈소스 기반의 공간분석 Tool인 QGIS 및 오픈소스 통계분석 패키지인 R스튜디오를 채택하였다.

[그림 3-6] 분석모델 구축방향 - 모델 확산화 측면

3.2 상수도 누수 위험도 분석 프로세스

상수도 누수 위험도 예측 모델의 전체 분석 프로세스는 다음과 같다.



[그림 3-8] 상수도 누수 위험도 예측 분석 프로세스 개요

프로세스는 데이터 수집, 데이터 전처리, 데이터 분석 및 분석결과 시각화로 구성되며 세부적인 내용은 다음 페이지 이후부터 설명하도록 한다.

3.3 데이터 수집 및 전처리

3.3.1 데이터 수집

상수도 누수 위험도 분석을 위해 수급 표준 데이터들은 다음과 같다.

아래의 표에서 보면 상수관로 데이터, 누수복구내역 데이터, 수용가 검침 데이터, 계량기 데이터, 블록 데이터들은 지자체 담당부서로부터 수급이 가능

하고 개방 공공 데이터는 각각의 보유기관 사이트에서 다운로드가 가능하다.

[표 3-1] 상수도 누수 위험도 모델 표준 데이터 리스트

분석 모델	수집 데이터	보유 기관	수집 유형	수집 방식
상수도 관로 누수위험도 예측	상수관로GIS 급수관로GIS	상수도과/ (누수방지과)	SHP	오프라인
	누수민원지점 복구내역		SHP	
옥내 누수 판별	수용가 검침량	요금과	Excel	
	급수전 계량기	상수도과	SHP	
	블록	상수도과	SHP	
개방 공공 데이터	도로명주소	행정안전부	SHP	온라인
	토양종류	농촌진흥청	SHP	
	격자 100m	국토교통부	SHP	

분석 모델링을 위해 표준화 작업한 데이터셋은 다음과 같다.

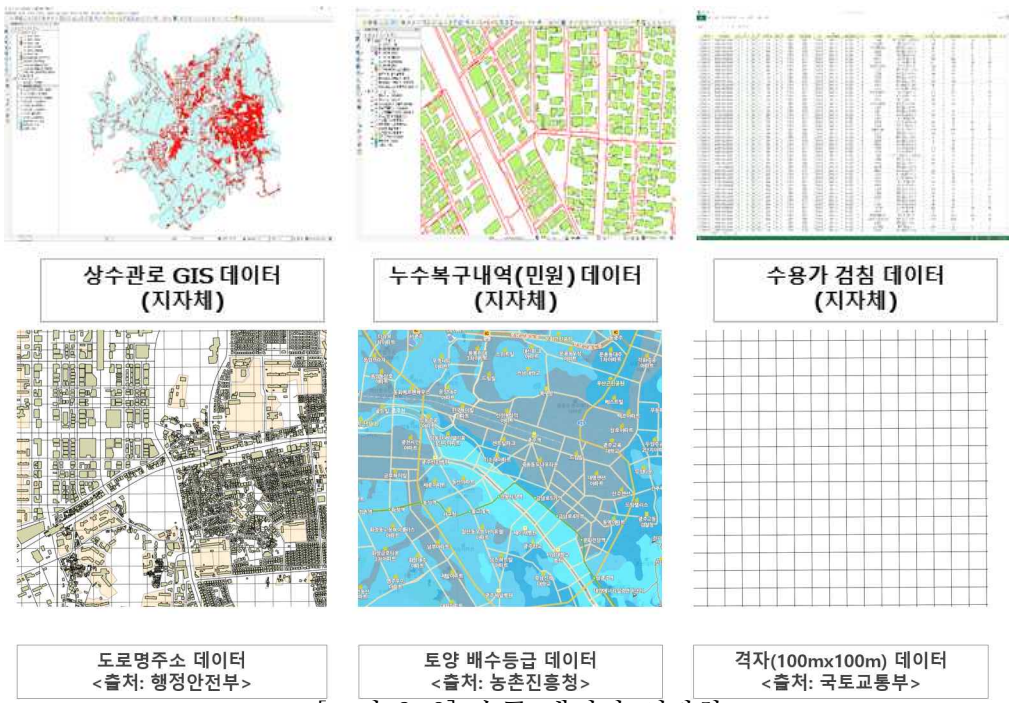
[표 3-2] 상수관로 GIS 데이터 표준화 예시

데이터 영문명	WTL_PIPE_LM	데이터 한글명	상수도 관로 GIS		
컬럼 명	컬럼 한글명	데이터 타입	길이	필수(Y/N)	기본키(PK)
FTR_IDN	상수관리번호	숫자	10	Y	PK
IST_YMD	설치일자	문자	8	Y	
SAA_CDE	관용도	문자	6	Y	
MOP_CDE	관재질	문자	6	Y	
STD_DIP	관경	숫자	6, 2	Y	
PIP_LEN	연장	숫자	11, 2	Y	
JHT_CDE	접합종류	문자	6	Y	
LOW_DEP	최저깊이	숫자	5, 2	Y	
HGH_DEP	최고깊이	숫자	5, 2	Y	
AVG_DH	매설심도	숫자	5, 2	Y	
CNT_NUM	공사번호	문자	50	Y	FK
PIP_LBL	관라벨	문자	50	Y	

[표 3-3] 누수지점복구내역 데이터 표준화 예시

데이터 영문명	WTL_LEAK_PS	데이터 한글명	누수민원지점 복구내역		
컬럼 명	컬럼 한글명	데이터 타입	길이	필수(Y/N)	기본키(PK)
PIP_IDN	관리번호	숫자	10	Y	PK
RCV_NUM	민원접수번호	문자	50	Y	FK
LEK_YMD	누수일자	문자	8	Y	
LEK_LOC	누수위치설명	문자	100	Y	
LRS_CDE	누수원인	문자	6	Y	
LEP_CDE	누수부위	문자	6	Y	
LEK_EXP	누수현황	문자	50	Y	
REP_EXP	누수복구내용	문자	50	Y	
REP_NAM	누수복구자명	문자	100	Y	

위에서 수급한 데이터들의 시각화한 화면은 다음과 같다.



[그림 3-9] 수급 데이터 시각화

3.3.2 데이터 전처리

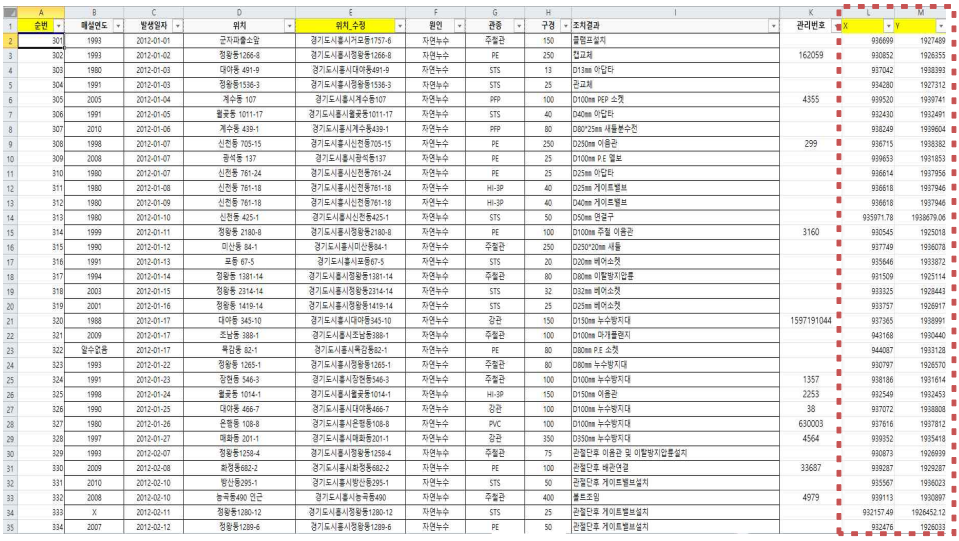
◇ 지오코딩(공간정보화)

상수도 분석모델을 적용하기 위해 관련 데이터들을 공간정보로 변환해야 될 필요가 있는데 누수지점복구내역 데이터는 경우에 따라 GIS 세이프 파일(shp) 형태가 아닌, 관리DB 형태인 엑셀파일(xls) 형태로 존재하는 경우가 있다.

지오코딩(Geocoding)이란 텍스트 형태의 주소 데이터를 정제하여 표준행정구역과 주소체계로 표준화하고 지도상의 좌표(위도, 경도)로 변환하는 과정을 지오코딩(Geocoding)이라 하는데 여기서는 개념만 설명하고 구체적인 내용은 생략하기로 한다.

[그림 3-10] 지오코딩 개념

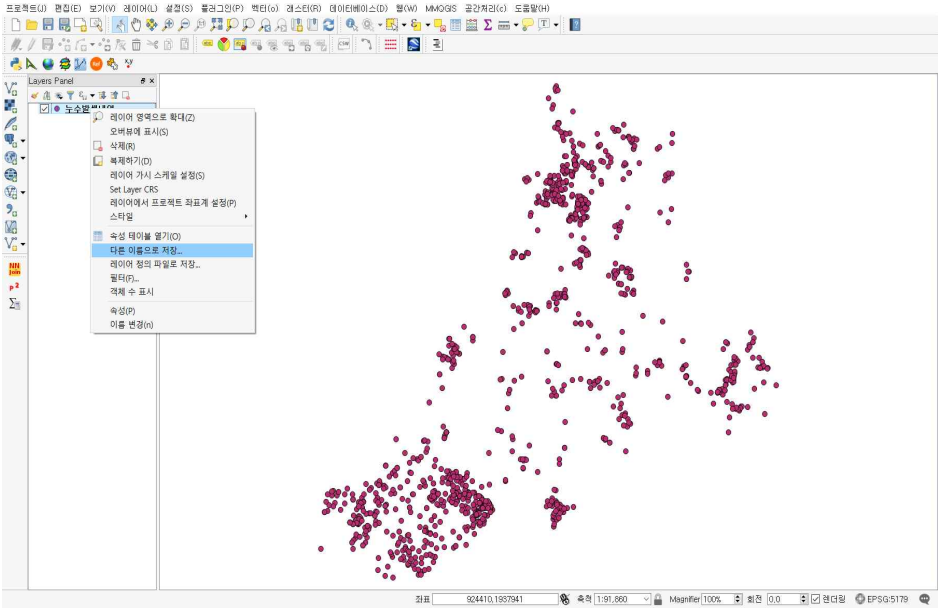
다음은 지오코딩 과정을 통해 컬럼별 X,Y 좌표를 생성한 화면이다.



	A	B	C	D	E	F	G	H	I	K	L	M
1	순번	매설연도	발생일자	위치	위치_수정	원인	관종	구경	조치결과	관리번호	X	Y
2	301	1993	2012-01-01	군자파출소앞	경기도시흥시저오동1757-6	자연누수	주철관	150	클램프설치		936699	1927489
3	302	1993	2012-01-02	정왕동1266-8	경기도시흥시정왕동1266-8	자연누수	PE	250	캡교체	162059	930852	1926355
4	303	1980	2012-01-03	대야동 491-9	경기도시흥시대야동491-9	자연누수	STS	13	D13mm 어답타		937042	1938393
5	304	1991	2012-01-03	정왕동1536-3	경기도시흥시정왕동1536-3	자연누수	STS	25	관교체		934280	1927312
6	305	2005	2012-01-04	계수동 107	경기도시흥시계수동107	자연누수	PPP	100	D100mm PEP 소켓	4355	939520	1939741
7	306	1991	2012-01-05	월곶동 1011-17	경기도시흥시월곶동1011-17	자연누수	STS	40	D40mm 어답타		932430	1932491
8	307	2010	2012-01-06	계수동 439-1	경기도시흥시계수동439-1	자연누수	PPP	80	D80*25mm 새들분수전		938249	1939604
9	308	1998	2012-01-07	신천동 705-15	경기도시흥시신천동705-15	자연누수	PE	250	D250mm 이음관	299	936715	1938382
10	309	2008	2012-01-07	광석동 137	경기도시흥시광석동137	자연누수	PE	25	D100mm P.E 엘보		939653	1931853
11	310	1980	2012-01-07	신천동 761-24	경기도시흥시신천동761-24	자연누수	PE	25	D25mm 어답타		936614	1937956
12	311	1980	2012-01-08	신천동 761-18	경기도시흥시신천동761-18	자연누수	HI-3P	40	D25mm 게이트밸브		936618	1937946
13	312	1980	2012-01-09	신천동 761-18	경기도시흥시신천동761-18	자연누수	HI-3P	40	D40mm 게이트밸브		936618	1937946
14	313	1980	2012-01-10	신천동 425-1	경기도시흥시신천동425-1	자연누수	STS	50	D50mm 연결구		935971.78	1938670.06
15	314	1999	2012-01-11	정왕동 2180-8	경기도시흥시정왕동2180-8	자연누수	PE	100	D100mm 주철 이음관	3160	930545	1925018
16	315	1990	2012-01-12	미산동 84-1	경기도시흥시미산동84-1	자연누수	주철관	250	D250*20mm 새들		937749	1936078
17	316	1991	2012-01-13	포동 67-5	경기도시흥시포동67-5	자연누수	STS	20	D20mm 베어소켓		935646	1933872
18	317	1994	2012-01-14	정왕동 1381-14	경기도시흥시정왕동1381-14	자연누수	주철관	80	D80mm 이형직압콘		931509	1925114
19	318	2003	2012-01-15	정왕동 2314-14	경기도시흥시정왕동2314-14	자연누수	STS	32	D32mm 베어소켓		933325	1928443
20	319	2001	2012-01-16	정왕동 1419-14	경기도시흥시정왕동1419-14	자연누수	STS	25	D25mm 베어소켓		933757	1926917
21	320	1988	2012-01-17	대야동 345-10	경기도시흥시대야동345-10	자연누수	강관	150	D150mm 누수방지대	1597191044	937365	1938991
22	321	2009	2012-01-17	조남동 388-1	경기도시흥시조남동388-1	자연누수	주철관	100	D100mm 마개클랜지		943168	1930440
23	322	일수없음	2012-01-17	목감동 82-1	경기도시흥시목감동82-1	자연누수	PE	80	D80mm P.E 소켓		944087	1933128
24	323	1993	2012-01-22	정왕동 1265-1	경기도시흥시정왕동1265-1	자연누수	주철관	80	D80mm 누수방지대		930797	1926570
25	324	1991	2012-01-23	장현동 546-3	경기도시흥시장현동546-3	자연누수	주철관	100	D100mm 누수방지대	1357	938186	1931614
26	325	1998	2012-01-24	월곶동 1014-1	경기도시흥시월곶동1014-1	자연누수	HI-3P	150	D150mm 이음관	2253	932549	1932453
27	326	1990	2012-01-25	대야동 466-7	경기도시흥시대야동466-7	자연누수	강관	100	D100mm 누수방지대	38	937072	1938808
28	327	1980	2012-01-26	은행동 108-8	경기도시흥시은행동108-8	자연누수	PVC	100	D100mm 누수방지대	630003	937616	1937812
29	328	1997	2012-01-27	매화동 201-1	경기도시흥시매화동201-1	자연누수	강관	350	D350mm 누수방지대	4564	939352	1935418
30	329	1993	2012-02-07	정왕동1258-4	경기도시흥시정왕동1258-4	자연누수	주철관	75	관절단후 이음관 및 이탈방지압륜설치		930873	1926939
31	330	2009	2012-02-08	화정동682-2	경기도시흥시화정동682-2	자연누수	PE	100	관절단후 배관연결	33687	939287	1929287
32	331	2010	2012-02-10	방산동295-1	경기도시흥시방산동295-1	자연누수	STS	50	관절단후 게이트밸브설치		935567	1936023
33	332	2008	2012-02-10	능곡동490 인근	경기도시흥시능곡동490	자연누수	주철관	400	볼트조임	4979	939113	1930897
34	333	X	2012-02-11	정왕동1280-12	경기도시흥시정왕동1280-12	자연누수	STS	25	관절단후 게이트밸브설치		932157.49	1926452.12
35	334	2007	2012-02-12	정왕동1289-6	경기도시흥시정왕동1289-6	자연누수	PE	50	관절단후 게이트밸브설치		932476	1926033

[표 3-4] 지오코딩 변환 예시

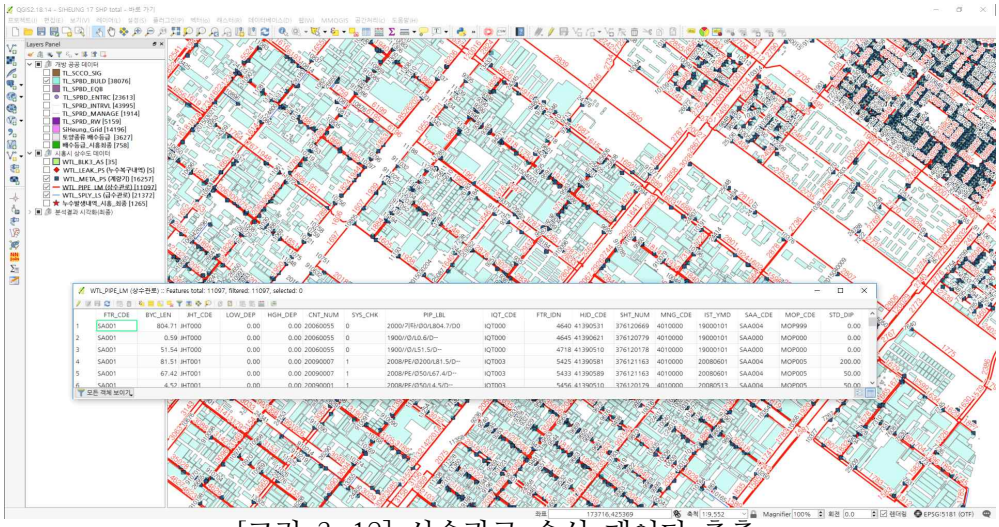
다음은 GIS프로그램을 이용하여 누수복구내역 데이터를 불어온 화면이다.



[그림 3-11] 누수복구내역 업로드 화면

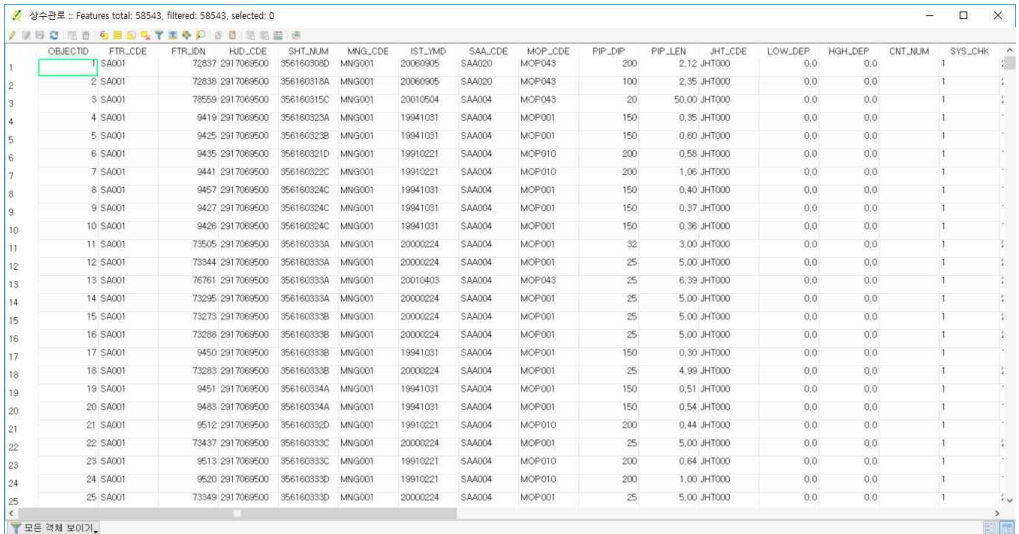
◇ 속성 데이터 추출

여기에서는 오픈소스인 QGIS 프로그램을 활용하여 상수관로에 포함되어 있는 설치일자, 관용도, 관재질, 관경, 연장, 접합종루, 매설심도, 공사번호, 관라벨 등의 상수관로 관련 데이터들을 추출하였다.



[그림 3-12] 상수관로 속성 데이터 추출

다음은 상수관로 레이어가 포함하고 있는 속성정보들을 추출한 화면이다.



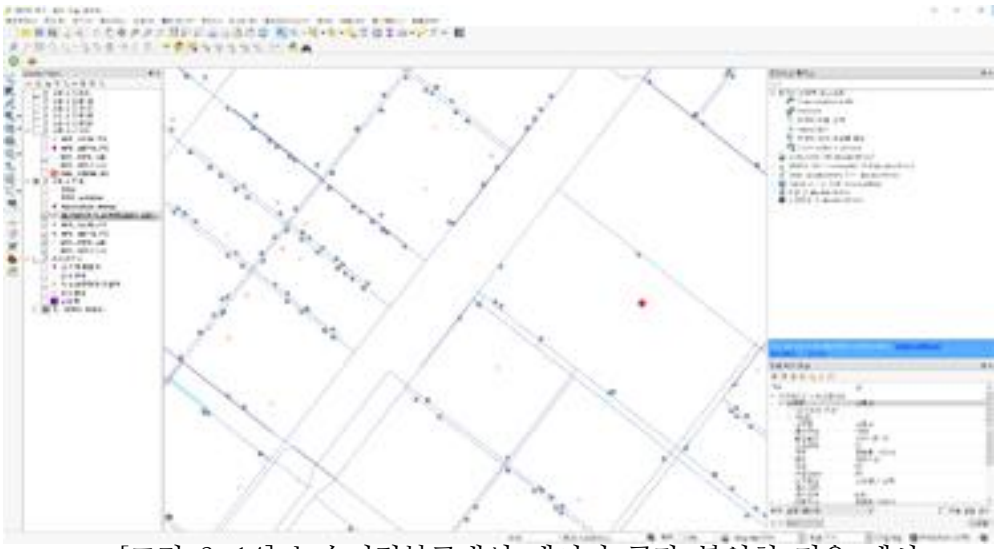
	OBJECTID	FTR_CDE	FTR_IDN	HID_CDE	SHT_NUM	MNG_CDE	IST_YMD	SAA_CDE	MOP_CDE	PIP_DIP	PIP_LEN	JHT_CDE	LOW_DEP	HGH_DEP	CNT_NUM	SYS_CHK
1	1	SA001	72837	2917069500	356160308D	MNG001	20060905	SAA020	MOP043	200	2,12	JHT000	0,0	0,0		1
2	2	SA001	72838	2917069500	356160318A	MNG001	20060905	SAA020	MOP043	100	2,35	JHT000	0,0	0,0		1
3	3	SA001	78559	2917069500	356160315C	MNG001	20010504	SAA004	MOP043	20	50,00	JHT000	0,0	0,0		1
4	4	SA001	9419	2917069500	356160323A	MNG001	19941031	SAA004	MOP001	150	0,35	JHT000	0,0	0,0		1
5	5	SA001	9425	2917069500	356160323B	MNG001	19941031	SAA004	MOP001	150	0,60	JHT000	0,0	0,0		1
6	6	SA001	9435	2917069500	356160321D	MNG001	19910221	SAA004	MOP010	200	0,58	JHT000	0,0	0,0		1
7	7	SA001	9441	2917069500	356160322C	MNG001	19910221	SAA004	MOP010	200	1,06	JHT000	0,0	0,0		1
8	8	SA001	9457	2917069500	356160324C	MNG001	19941031	SAA004	MOP001	150	0,40	JHT000	0,0	0,0		1
9	9	SA001	9427	2917069500	356160324C	MNG001	19941031	SAA004	MOP001	150	0,37	JHT000	0,0	0,0		1
10	10	SA001	9426	2917069500	356160324C	MNG001	19941031	SAA004	MOP001	150	0,36	JHT000	0,0	0,0		1
11	11	SA001	73505	2917069500	356160333A	MNG001	20000224	SAA004	MOP001	32	3,00	JHT000	0,0	0,0		1
12	12	SA001	73344	2917069500	356160333A	MNG001	20000224	SAA004	MOP001	25	5,00	JHT000	0,0	0,0		1
13	13	SA001	76761	2917069500	356160333A	MNG001	20010403	SAA004	MOP043	25	6,39	JHT000	0,0	0,0		1
14	14	SA001	73295	2917069500	356160333A	MNG001	20000224	SAA004	MOP001	25	5,00	JHT000	0,0	0,0		1
15	15	SA001	73273	2917069500	356160333B	MNG001	20000224	SAA004	MOP001	25	5,00	JHT000	0,0	0,0		1
16	16	SA001	73288	2917069500	356160333B	MNG001	20000224	SAA004	MOP001	25	5,00	JHT000	0,0	0,0		1
17	17	SA001	9450	2917069500	356160333B	MNG001	19941031	SAA004	MOP001	150	0,30	JHT000	0,0	0,0		1
18	18	SA001	73283	2917069500	356160333B	MNG001	20000224	SAA004	MOP001	25	4,99	JHT000	0,0	0,0		1
19	19	SA001	9451	2917069500	356160334A	MNG001	19941031	SAA004	MOP001	150	0,51	JHT000	0,0	0,0		1
20	20	SA001	9483	2917069500	356160334A	MNG001	19941031	SAA004	MOP001	150	0,54	JHT000	0,0	0,0		1
21	21	SA001	9512	2917069500	356160332D	MNG001	19910221	SAA004	MOP010	200	0,44	JHT000	0,0	0,0		1
22	22	SA001	73437	2917069500	356160333C	MNG001	20000224	SAA004	MOP001	25	5,00	JHT000	0,0	0,0		1
23	23	SA001	9513	2917069500	356160333C	MNG001	19910221	SAA004	MOP010	200	0,64	JHT000	0,0	0,0		1
24	24	SA001	9520	2917069500	356160333D	MNG001	19910221	SAA004	MOP010	200	1,00	JHT000	0,0	0,0		1
25	25	SA001	73349	2917069500	356160333D	MNG001	20000224	SAA004	MOP001	25	5,00	JHT000	0,0	0,0		1

[그림 3-13] 상수관로 레이어 속성정보

◇ 상수관로와 누수지점 및 복구내역 매칭

누수지점복구내역 데이터는 QGIS에서 업로드 가능한 GIS 세이프 파일(shp) 또는 엑셀파일(xls) 형태로 존재한다. 후자의 경우 문제는 GIS시스템 상에서 직접 상수관로를 직접 지정하는 작업을 저장해 놓지 않아서 어느 상수관로에서 누수복구 문제가 발생했었는지 이력을 알 수 없는 문제가 발생한다.

이 경우 상수관로 데이터와 누수지점 및 복구내역 데이터를 매칭해야 되는데 GIS 공간정보 시스템 상에서 최단거리에 매칭되는 상수관로임을 확인함과 동시에 누수복구내역의 속성정보와 비교하여 판단하게 되는 과정을 거친다.



[그림 3-14] 누수지점복구내역 데이터 공간 불일치 경우 예시

3.3.3 데이터 셋 생성

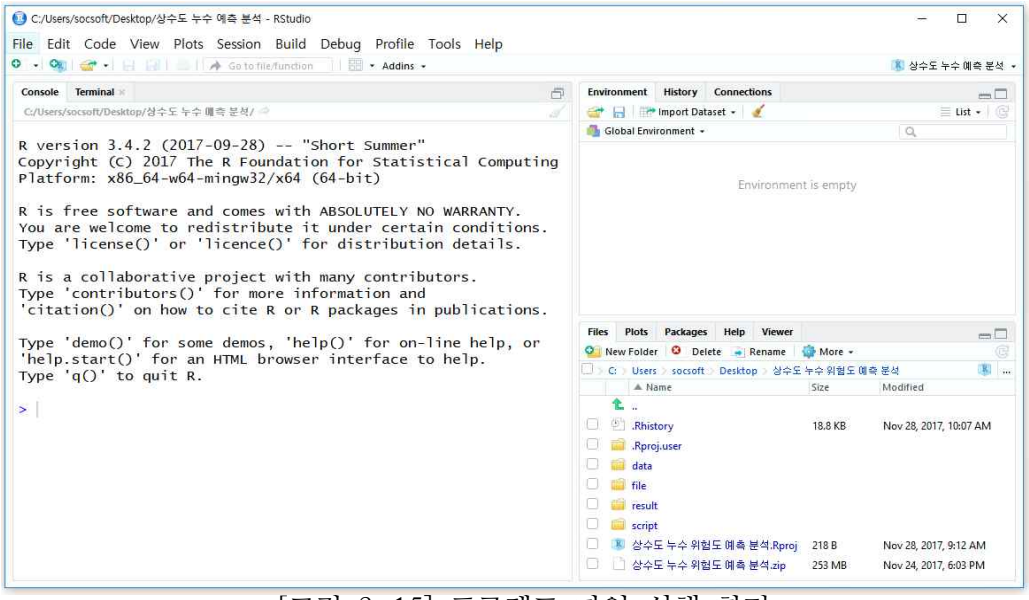
다음은 상수도 누수 위험도 분석 모델 변수 표준화 결과이다.

[표 3-5] 표준화 변수 리스트

필드명	필드 설명
FTR_IDN	상수관로관리번호
SAA_CDE	관용도
MOP_CDE	관재질
PIP_DIP	관경
PIP_LEN	관연장
JHT_CDE	접합종류
PIP_DEP	매설심도
PIP_LBL	관라벨
LEK_YMD	누수발생일자
WTLKG_CHK	누수여부체크
WTLKG_CNT	누수발생건수
CNSMR_CNT	수용가수
ROAD_CD	도로종류
DRNG_CD	토양배수등급

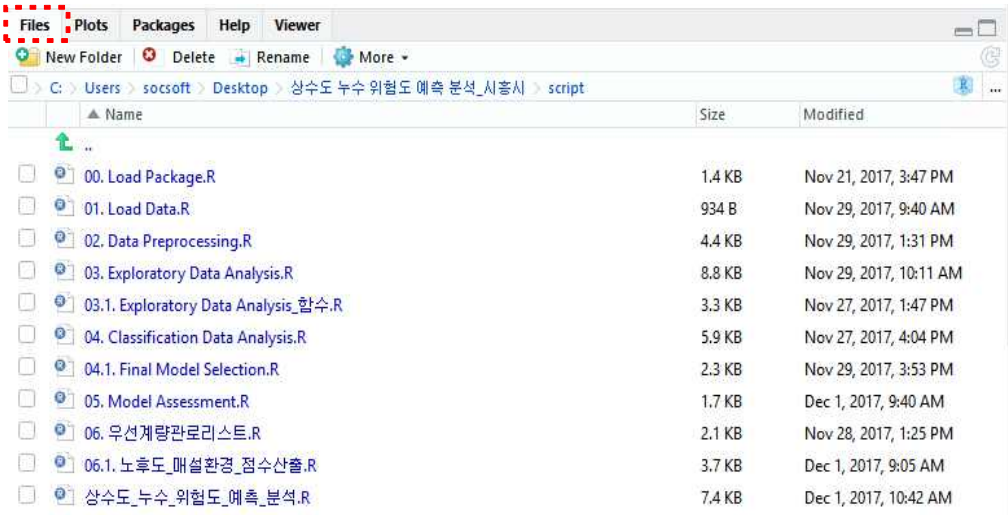
데이터 전처리 단계인 지오코딩, 속성 데이터 추출 및 상수관로와 누수지점 및 복구내역 매칭 등의 전처리 작업을 통해 위와 같이 분석 모델링용 최종 표준 데이터 셋을 생성할 수 있다.

다음은 상수도 누수 위험도 예측분석을 위한 R스튜디오 실행화면이다.



[그림 3-15] 프로젝트 파일 실행 화면

아래는 분석모형 프로세스 단계별로 R소스코드 리스트이다.



[그림 3-16] R스튜디오 분석모델 코드 리스트

데이터 전처리 단계에서는 각종 변수 생성과 함께 결측치 확인 및 제거 등을 통해 분석용 데이터 셋 생성하며 그 다음 데이터 분할(train data, validation data, test data), 훈련용 데이터세트의 over sampling을 포함하고 있다.

다음은 R코드에서 전처리한 데이터 셋을 업로드하는 작업 단계이다.

```
wtl_pipe_leak_gwangju <- data.table(wtl_pipe_leak_gwangju)

# 결측치 확인
colSums(is.na(wtl_pipe_leak_gwangju))
```

[그림 3-17] R스튜디오에서 데이터 업로드 단계

다음은 위의 명령에 따라 결측치 개수를 출력하는 화면이다.

```
##  OBJECTID   FTR_CDE   FTR_IDN   HJD_CDE   SHT_NUM   MNG_CDE   IST_YMD
##         0         0         0         1         0         0         2
##   SAA_CDE   MOP_CDE   PIP_DIP   PIP_LEN   JHT_CDE   LOW_DEP   HGH_DEP
##         0         0         0         0         0         0         0
##   CNT_NUM   SYS_CHK   PIP_LBL   PIP_NUM   PIP_DEP   PIP_STT   PIP_END
##         0         0         0     10408         0         0         0
##   MND_CDE   COM_NAM   RSN_CDE   TAG_NUM   FTR_USR   FTR_TIM   PRS_ARE
##         0         0         0     58543         0         0         0
##   BLK_NAM   INQ_CDE SHAPE_LEN   ROAD_CD CNSMR_CNT WTLKG_CNT WTLKG_CHK
##         0         0         0         0         0     54478     54478
##   LEK_YMD
##     54554
```

[그림 3-18] 결측치 결과값 확인

다음은 표준 데이터 셋에서 파생변수를 생성하는 단계이다.

```
# 관로 경과년수 생성
date <- now() # 현재 날짜 추출
now.year <- year(date) # 현재 년도 추출
wtl_pipe_leak_gwangju$LEK_Y <- as.numeric(substr(wtl_pipe_leak_gwangju$LEK_YMD, 1, 4)) # 누수
wtl_pipe_leak_gwangju$PIP_Y <- as.numeric(substr(wtl_pipe_leak_gwangju$IST_YMD, 1, 4)) # 매설
wtl_pipe_leak_gwangju[is.na(LEK_Y) == T, PIP.Y := now.year - PIP_Y] # 누수되지 않은 관로에 대한
wtl_pipe_leak_gwangju[is.na(LEK_Y) == F, PIP.Y := LEK_Y - PIP_Y]  # 누수시점에서의 관로 경과년
wtl_pipe_leak_gwangju[LEK_Y < PIP_Y, PIP.Y := now.year - PIP_Y] # 누수시점 이후에 새로 매설된 관


# 누수관로 여부 변수 생성
wtl_pipe_leak_gwangju[, LEAK := "N"] # 매칭이 되지 않은 관로는 정상관로
wtl_pipe_leak_gwangju[WTLKG_CHK > 0, LEAK := "Y"] # 매칭이 된 관로는 누수관로
wtl_pipe_leak_gwangju[LEK_Y < PIP_Y, LEAK := "N"]
wtl_pipe_leak_gwangju$LEAK <- as.factor(wtl_pipe_leak_gwangju$LEAK)
table(wtl_pipe_leak_gwangju$LEAK)
```

[그림 3-19] R스튜디오에서 파생변수 생성 단계

다음은 위의 명령에 따라 결측치 개수를 출력하는 화면이다.

```
##
##     N     Y
## 54576  3967
```

[그림 3-20] 결측치 결과값 확인

관로 경과년수 변수는 관 라벨 변수의 매설년도 기준으로 산출 하였다.

[표 3-6] 상수관로 누수 경과년수 생성 기준

종 류	경과년수
누수이력이 없는 관로	현재년도에서 매설시점의 년도를 뺀 년수
누수이력이 있는 관로	누수시점에서 매설시점의 년도를 뺀 년수
	누수시점 이후에 새로 매설된 관로 경과년수

아래는 전처리 과정을 통해 분석용 데이터 셋을 최종 생성하는 단계이다.

```
# 관로별 누수개수 결측치 0으로 대체
wtl_pipe_leak_gwangju$WTLKG_CNT <- ifelse(is.na(wtl_pipe_leak_gwangju$WTLKG_CNT) == T, 0, wtl_pipe_leak_gwangju$WTLKG_CNT)

# 도로종류 변수 범주화
wtl_pipe_leak_gwangju$ROAD_CD <- as.factor(wtl_pipe_leak_gwangju$ROAD_CD)
levels(wtl_pipe_leak_gwangju$ROAD_CD) <- c("고속도로", "4차선이상도로", "2차선이상도로", "이면도로", "보도및노지")


# 분석용 데이터셋 생성
wtl_pipe_leak_gwangju_ant <- wtl_pipe_leak_gwangju[ ,c("SAA_CDE", "MOP_CDE", "PIP_DIP", "PIP_LEN",
                                                       "JHT_CDE", "PIP_DEP", "FTR_IDN", "ROAD_CD",
                                                       "CNSMR_CNT", "WTLKG_CNT", "LEAK", "PIP.Y")]
wtl_pipe_leak_gwangju_ant$MOP_CDE <- as.character(wtl_pipe_leak_gwangju_ant$MOP_CDE)
wtl_pipe_leak_gwangju_ant <- wtl_pipe_leak_gwangju_ant[MOP_CDE == "MOP006" | MOP_CDE == "MOP001" |
                                                         MOP_CDE == "MOP019" | MOP_CDE == "MOP013" |
                                                         MOP_CDE == "MOP004" | MOP_CDE == "MOP005" |
                                                         MOP_CDE == "MOP043" | MOP_CDE == "MOP044" |
                                                         MOP_CDE == "MOP050" | MOP_CDE == "MOP028" |
                                                         MOP_CDE == "MOP010" | MOP_CDE == "MOP003" |
                                                         MOP_CDE == "MOP011",]
wtl_pipe_leak_gwangju_ant$MOP_CDE <- as.factor(wtl_pipe_leak_gwangju_ant$MOP_CDE)


# 결측치 확인 후 제거
colSums(is.na(wtl_pipe_leak_gwangju_ant))
```

[그림 3-21] R스튜디오에서 모델링 데이터셋 생성 단계

다음은 위의 명령에 따라 결측치 개수를 출력하는 화면이다.

```
##   SAA_CDE   MOP_CDE   PIP_DIP   PIP_LEN   JHT_CDE   PIP_DEP   FTR_IDN
##         0         0         0         0         0         0         0
##   ROAD_CD CNSMR_CNT WTLKG_CNT      LEAK     PIP.Y
##         0         0         0         0         1
```

[그림 3-22] 결측치 결과값 확인

3.3.4 데이터 분할

모형 평가와 모형 선택 단계에서는 정확한 모형 평가를 위해서는 데이터를 개념적으로 훈련용(Train data), 검증용(Validation data), 테스트용(Test data) 형태로 구분해야 한다.

훈련용 데이터세트(train dataset)는 모형의 적합과 모수의 추정에 사용되고 검증용 데이터세트(validation dataset)는 파라미터 튜닝과 변수 선택, 모형 선택에 사용되고 테스트용 데이터세트(test dataset)는 모형 적합과 모형 선택이 끝난 후 최종 모형의 오류 확률을 추정하기 위해 사용된다.

데이터를 훈련용 60%, 검증용 20%, 테스트용 20%로 나눈다.

다음은 분석용 데이터에서 모델 훈련용(Train) 데이터과 모델 테스트용(Test) 데이터를 생성하는 단계이다.

```
# testdata & traindata 생성 (holdout 방법)
set.seed(123)
parts_gwangju <- createDataPartition(wtl_pipe_leak_gwangju_ant$LEAK, p = 0.6)
trainData_gwangju <- wtl_pipe_leak_gwangju_ant[parts_gwangju$Resample1, ]; table(trainData_gwangju$LEAK)
```

```
##
##     N     Y
## 32494  2376
```

```
testData_gwangju <- wtl_pipe_leak_gwangju_ant[-parts1_gwangju$Resample1, ]; table(validationData_gwangju$LEAK)
```

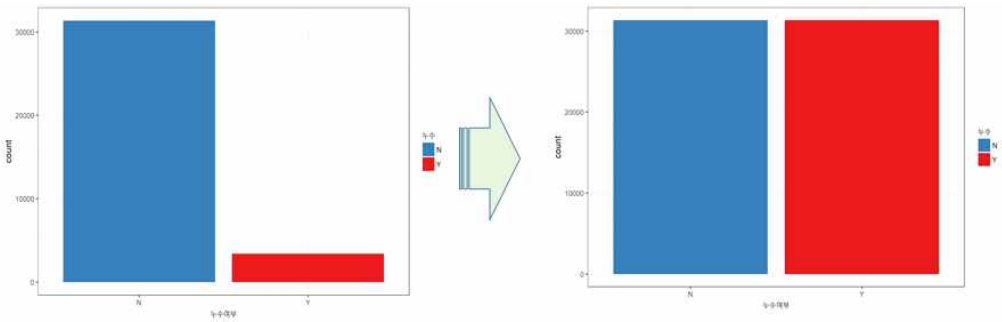
```
##
##     N     Y
## 11239   384
```

[그림 3-23] Train data & Test data 생성

□ 훈련용 데이터세트의 오버샘플링(over-sampling) 문제

앞에서도 보았듯이 훈련용 데이터 셋의 누수여부 분포를 살펴 볼 때, 누수가 없는 정상관로(N으로 표기)보다 누수발생 이력이 있는 누수관로(Y로 표기)가 매우 작음을 확인 할 수 있는데 이러한 자료를 불균형자료라고 정의한다.

반응변수(여기서는 누수여부)의 클래스 비율이 상대적으로 불균형한 경우, 다시 말해 대부분의 데이터가 누수가 없는 정상관로(N으로 표기)로 구성 되었을 때 이러한 데이터의 불균형 상태는 분류 모형의 성능에 문제가 될 수 있다. 그래서 이 부분에서의 문제를 해결하기 위해 우리는 오버샘플링 (over-sampling) 방법을 적용하였다.



[그림 3-24] 오버샘플링 전과 후의 데이터 개수 변화

위의 좌측 그림은 원래의 훈련용 데이터세트이고 우측 그림은 오버샘플링 (over-sampling)을 통해 정상관로(N으로 표기)와 누수관로(Y로 표기)의 개수를 균형적으로 맞춘 결과이다.

3.4 예측모델 생성 및 성능평가

3.4.1 예측 모형 생성

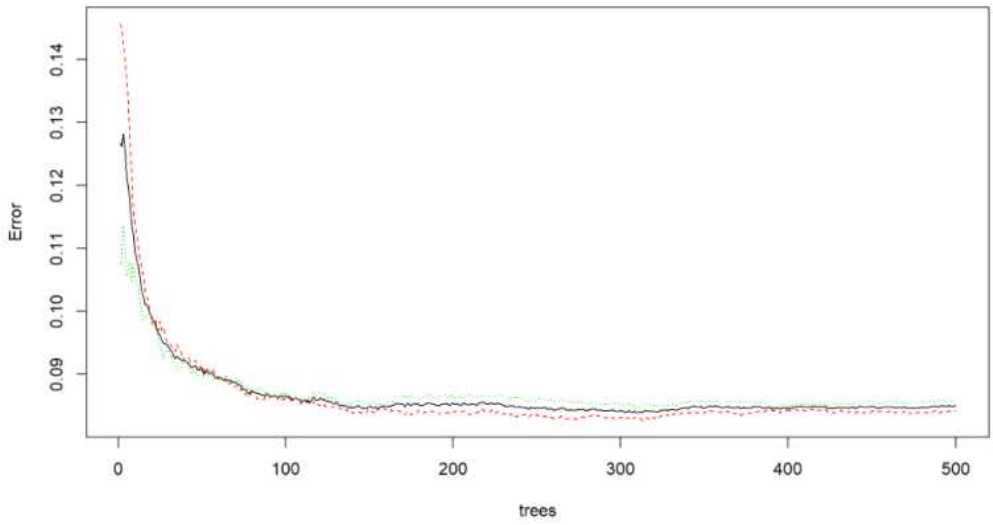
상수관로 누수 위험도 예측분석은 분류분석 모형인 로지스틱 회귀분석, 그레디언트 부스팅, 랜덤 포레스트 3가지를 기본적으로 생성하였으며, 여기서는 모형의 성능이 가장 우수했던 랜덤포레스트 모형을 중심으로 설명하겠다.

다음은 랜덤포레스트(Random Forest) 모형을 생성하는 단계이다.

```
rf_gwangju <- randomForest(Class ~ MOP_CDE + PIP_DIP + PIP_DEP + CNSMR_CNT + PIP.Y + ROAD_CD,
                           data = trainData_gwangju_up, importance = TRUE,
                           ntree = 500, mtry = 2)
print(rf_gwangju)

##
## Call:
##  randomForest(formula = Class ~ MOP_CDE + PIP_DIP + PIP_DEP +
##  CNSMR_CNT + PIP.Y + ROAD_CD, data = trainData_gwangju_up,
##   importance = TRUE, ntree = 500, mtry = 2)
##
##               Type of random forest: classification
##                     Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 8.5%
## Confusion matrix:
##       N     Y class.error
## N 29753  2741  0.08435403
## Y  2786 29708  0.08573891

plot(rf_gwangju) # 나무개수에 따른 에러율

varImpPlot(rf_gwangju) # 변수별 중요도
```

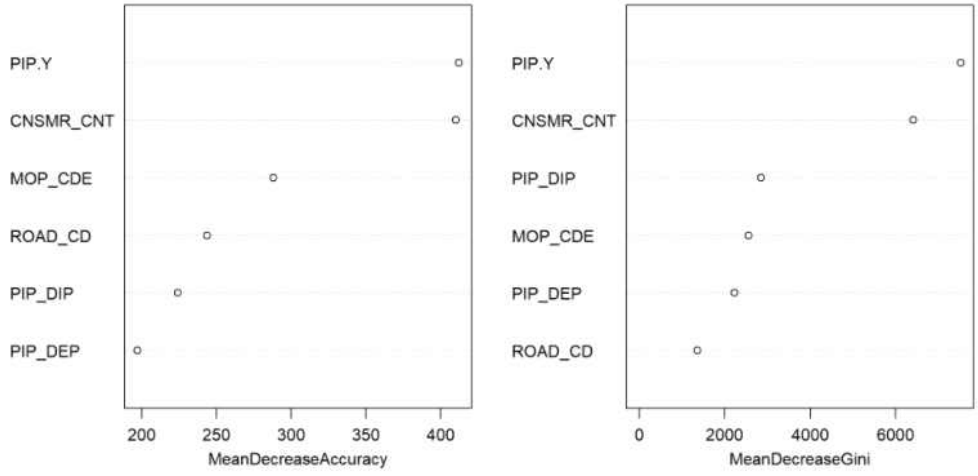
[그림 3-25] 랜덤포레스트 모형의 생성

아래 그림은 나무개수에 따른 에러율 변화를 출력한 화면이다.



[그림 3-26] 모형의 나무갯수에 따른 에러율

아래 그림은 모형에 사용된 변수별 중요도를 출력한 화면이다.



[그림 3-27] 모형에 사용된 변수별 중요도

다음은 랜덤포레스트(Random Fores) 모형을 평가하는 단계이다.

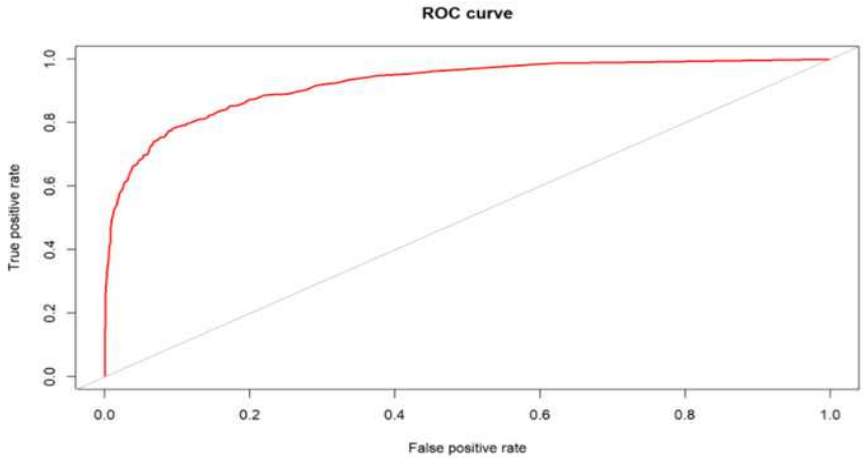
```
## confustion matrix
rf_pred_gwangju <- predict(rf_gwangju, newdata = validationData_gwangju)
print(confusionMatrix(rf_pred_gwangju, validationData_gwangju$LEAK))

## Confusion Matrix and Statistics
##
##          Reference
## Prediction     N     Y
##          N 10511   105
##          Y   728   279
##
##                Accuracy : 0.9283
##                  95% CI : (0.9235, 0.933)
##     No Information Rate : 0.967
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3711
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9352
##             Specificity : 0.7266
##          Pos Pred Value : 0.9901
##          Neg Pred Value : 0.2771
##              Prevalence : 0.9670
##          Detection Rate : 0.9043
##    Detection Prevalence : 0.9134
##       Balanced Accuracy : 0.8309
##
##        'Positive' Class : N
##
```

[그림 3-28] 랜덤포레스트 모형 평가

다음은 ROC Curve를 생성하는 R소스코드 화면이다.

```
## ROC curve
rf_prob_gwangju <- predict(rf_gwangju, newdata = validationData_gwangju, type = "prob")[,2]
rf_roc <- roc.curve(validationData_gwangju$LEAK, rf_prob_gwangju, col = 2, lwd = 2, lty = 1)$auc
```

[그림 3-29] ROC Curve 생성 코드



[그림 3-30] ROC Curve 생성 결과

다음은 모형의 성능 평가 지표로 사용된 결과를 출력한 화면이다.

```
print(paste("Area under the Curve", ":", round(rf_roc, 4)))
```

```
## [1] "Area under the Curve : 0.922"
```

```
## F1-score
rf_precision <- posPredValue(rf_pred_gwangju, validationData_gwangju$LEAK)
rf_recall <- sensitivity(rf_pred_gwangju, validationData_gwangju$LEAK)
rf_F1 <- (2 * rf_precision * rf_recall) / (rf_precision + rf_recall)
print(paste("F1-score", ":", round(rf_F1, 4)))
```

```
## [1] "F1-score : 0.9619"
```

```
## 이항편차
yhat_rf <- ifelse(rf_pred_gwangju == "Y", 1, 0)
(bin_dev_rf <- binomial_deviance(y_obs, yhat_rf))
```

```
## [1] 15346.59
```

[그림 3-31] 모형 성능 출력 코드

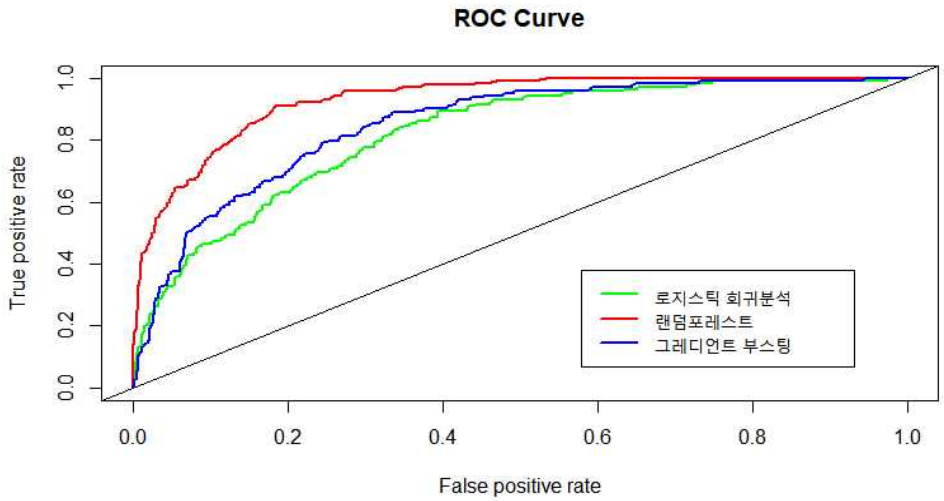
3.4.2 모형별 성능 평가

상수관로 누수 위험도 예측분석은 분류분석(로지스틱 회귀분석, 그레디언트 부스팅, 랜덤 포레스트)모형 3가지를 성능 비교 및 평가하였다.

[표 3-7] 모델별 성능 비교 및 평가

평가지표	Logistic Regression	Gredient Boosting	Random Forest
정확도	78.87%	79.24%	93.99%
F-score	0.8563	0.8688	0.9361
이항편차	50087.60	46164.04	24243.67
AUC	0.8016	0.8084	0.8568

아래는 적용했던 모형별로 비교한 ROC Curve이다.



[그림 3-32] 분석모형별 ROC Curve

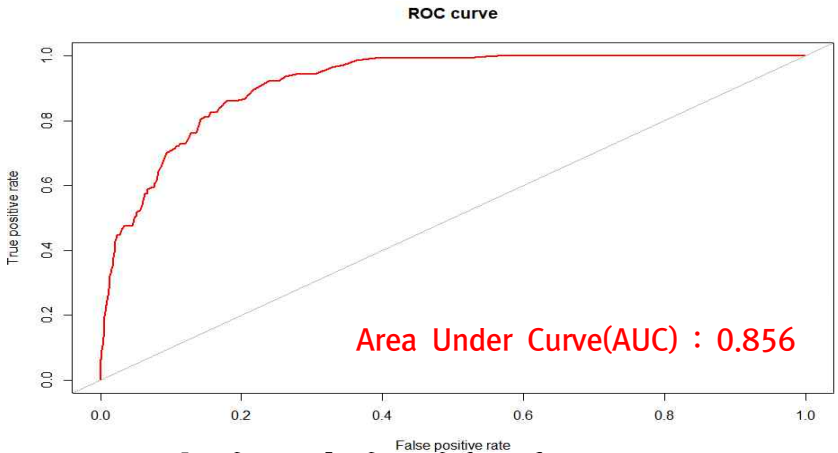
3.4.3 최종 모형 채택

K시, 상수관로 데이터셋 분석결과, 로지스틱 회귀 (Logistic Regression), 그라디언트 부스팅 (Gredient Boosting), 랜덤 포레스트 (Random Forest) 중에서 예측 성능이 가장 우수한 랜덤 포레스트 예측 방법론을 채택하였다.

이는 정확도(Accuracy)가 높고 F-score값이 높으며, 이항편차가 작은 결과가 나타났으며 최종 평가지표 수치는 다음과 같다.

[표 3-8] 최종 모형 최종 평가지표 수치

평가지표	Random Forest
정확도	86.74%
F-score	0.9241
이항편차	28296.176
AUC	0.856



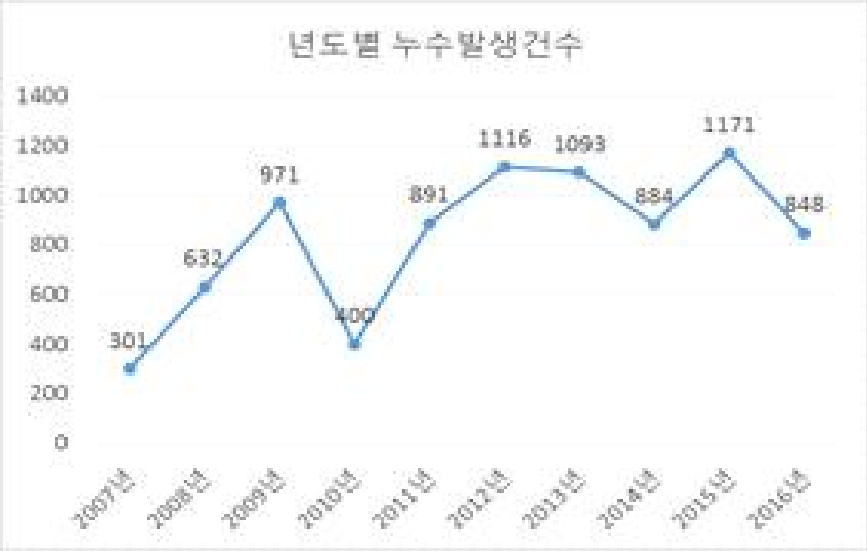
[그림 3-33] 최종 채택 모형 ROC Curve

정확도가 86.74%, F-score는 0.9241로 높은 값을 가지므로 모형의 정확도가 높은 수준이며, AUC은 0.856로 평가 등급(Good) 역시 높은 수준이다.

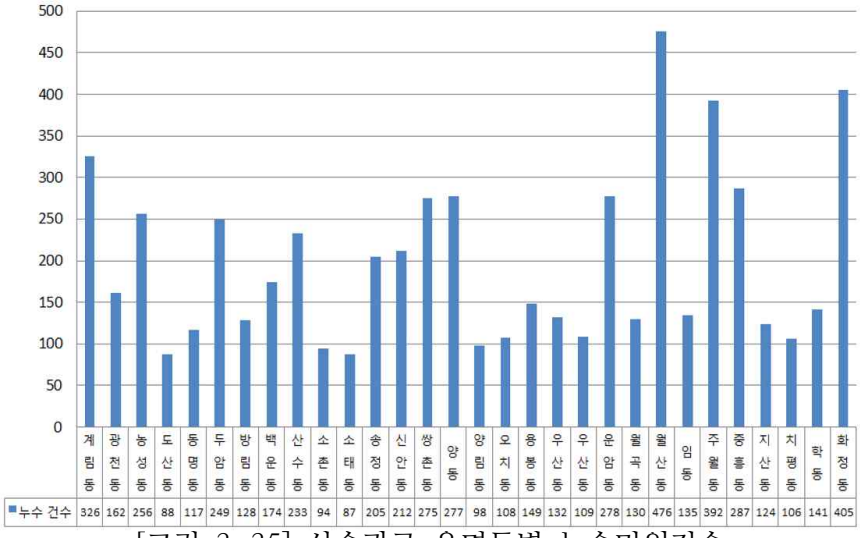
3.5 상수도 누수 위험도 분석결과 시각화

3.5.1 탐색적 분석(EDA)

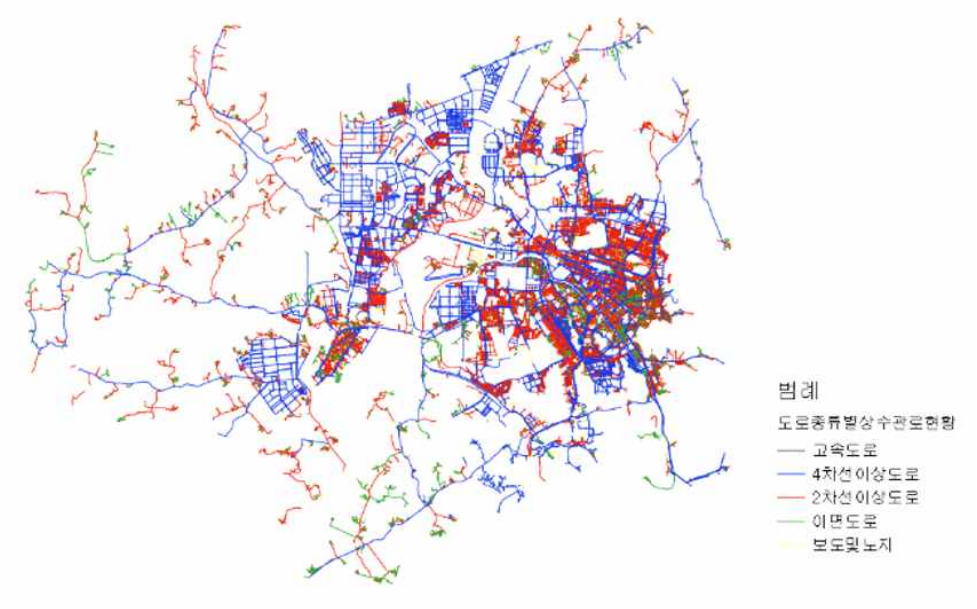
K시 상수관로 데이터 탐색적 현황 분석결과는 아래 그림들과 같다.



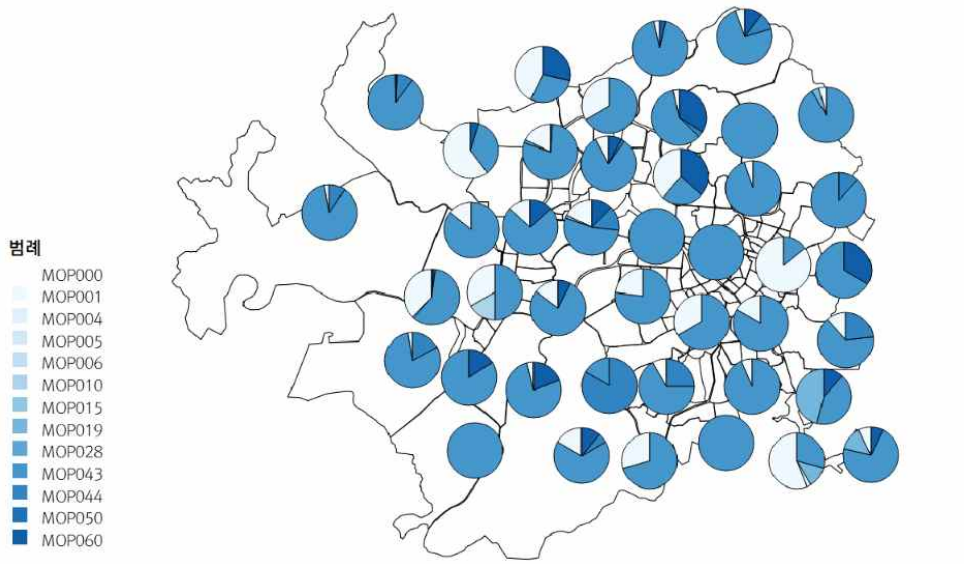
[그림 3-34] 상수관로 년도별 누수발생건수



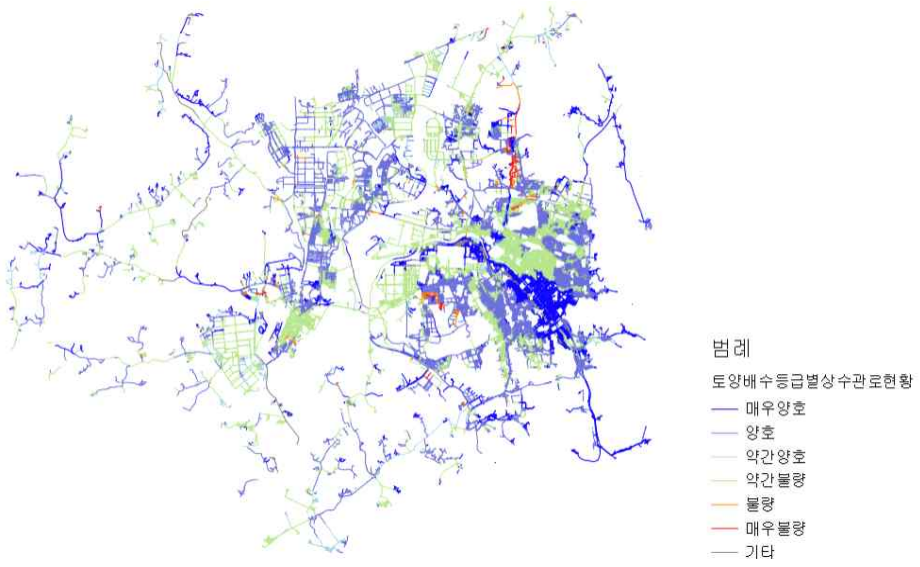
[그림 3-35] 상수관로 읍면동별 누수민원건수



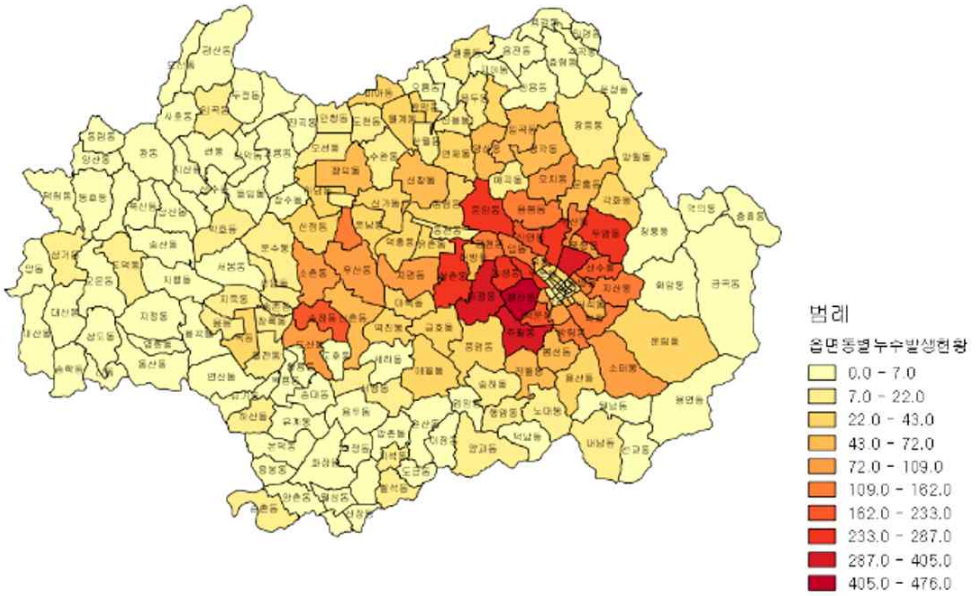
[그림 3-36] 도로종류별 상수관로 분포



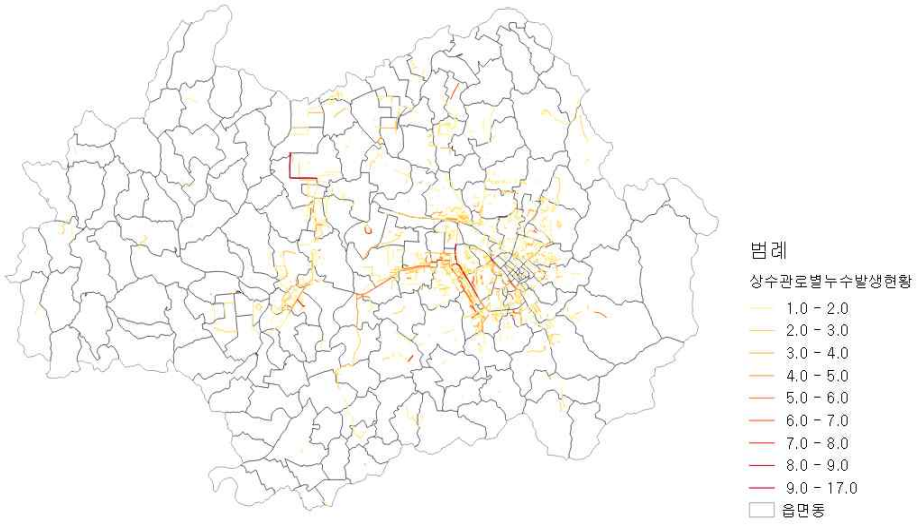
[그림 3-37] 관재질별 상수관로 분포



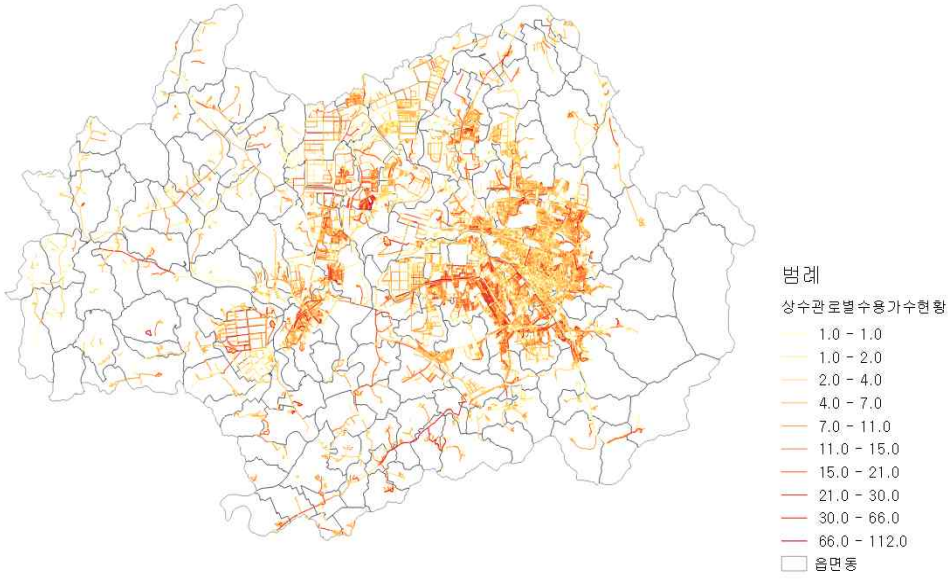
[그림 3-38] 토양배수등급별 상수관로 현황



[그림 3-39] 읍면동별 누수다발지역 현황



[그림 3-40] 상수관로별 누수발생 현황

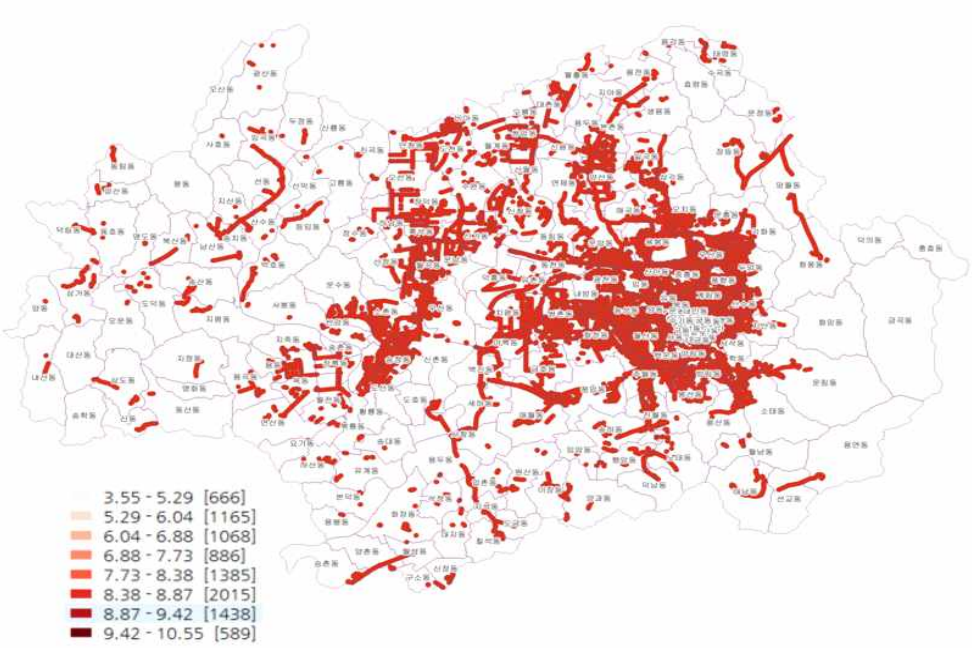


[그림 3-41] 상수관로별 수용가수 현황

3.5.2 분석결과 시각화

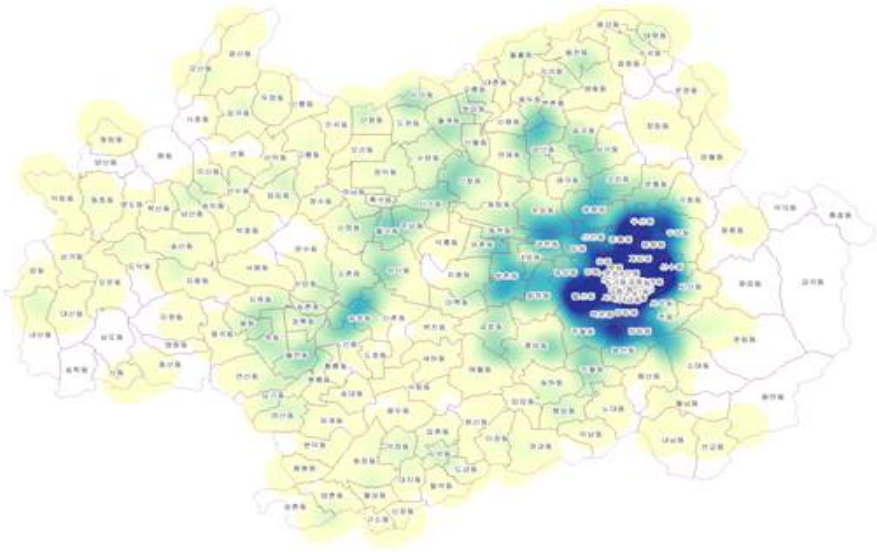
K시 누수 위험도 상위관로 분포 현황은 아래 그림과 같다.

분석결과 시각화는 예측모델로부터 도출한 누수 위험도 수치를 GIS 상수관로에 속성정보 컬럼 추가하고 그것을 시각화 단계 등급화로 나눠 사용하였다.



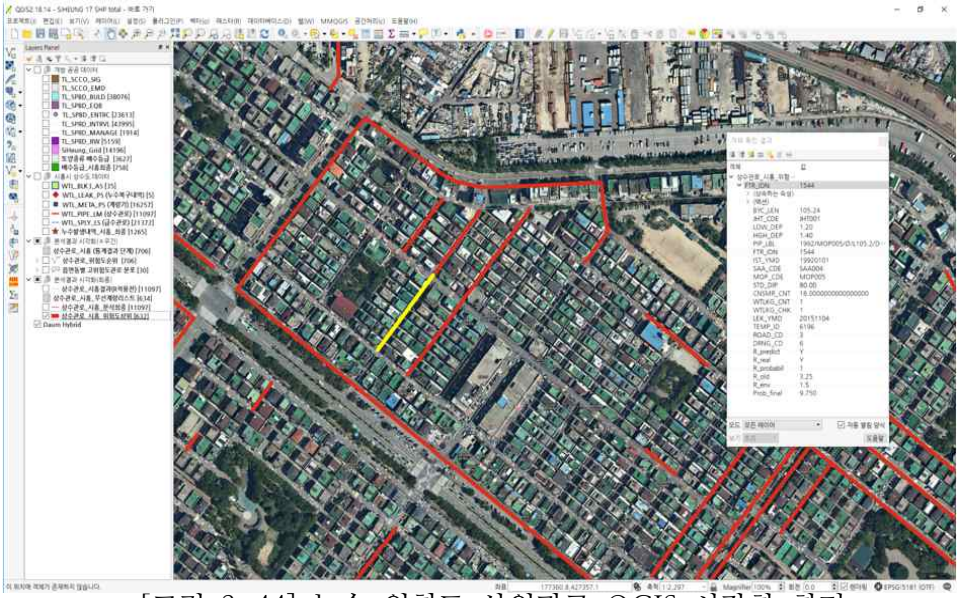
[그림 3-42] K시 누수 위험도 상위관로 전체 현황

분석결과를 히트맵(Heat Map)으로 살펴보면, K시의 구도심인 우측 집중되어 있는 지역을 중심으로 한 반경 1.5 ~ 2.0km 내에 누수 위험도 상위관로가 최다 분포되어 있음을 시각적으로 확인 할 수 있다.



[그림 3-43] 누수 위험도 상위관로 분포 히트맵

분석결과로 산출한 K시 누수 위험도 상위관로들은 QGIS 프로그램상에서 영상지도 배경으로 상수관로 속성정보와 함께 확인 할 수 있다.



[그림 3-44] 누수 위험도 상위관로 QGIS 시각화 화면

아래와 같이 확대해서 누수 위험도 상위 상수관로의 위치를 확인할 수 있다.

[그림 3-45] 누수 위험도 상위관로 상세정보 보기

누수 위험도 상위관로들의 최종 분석 결과는 위치를 확인할 수 있는 그림과 함께 아래 리스트가 함께 제공되며 세부 속성정보 내역은 다음과 같다.

FTR_IDN	IST_YMD	MOP_CDE	PIP_DIP	PIP_LEN	JHT_CDE	PIP_DEP	COM_NAM	BLK_NAM
15183	19840101	MOP010	150	114.91	JHT002	1		지원14
38429	19840101	MOP010	150	175.62	JHT002	1		지원14
659001553	19810804	MOP003	100	68.57	JHT004	1	제일수도	지원14
1630305	19830101	MOP010	350	71.71	JHT002	1.2		지원14
39884	19840101	MOP010	100	90.68	JHT002	1		지원14
15169	19910727	MOP013	30	30.5	JHT000	0.6	해성수도	지원14
38207	19890101	MOP013	100	83.11	JHT004	0.7		지원14
1630515	19830101	MOP010	350	0.45	JHT002	1.2		지원14
1630362	19880101	MOP001	100	15.01	JHT002	0.9		지원14
15175	19860706	MOP013	40	61.97	JHT000	0.6	서봉수도	지원14

[표 3-9] 누수 위험도 상위관로 컬럼 속성정보

Ⅳ. 연구 결과

4.1 누수 위험도 예측결과의 현실적 검증

4.1.1 현실적 검증 과정

실제 누수 위험도가 높게 나온 상수관로에 대해 담당자, 누수점검업체와 함께 현장을 방문하여 지하 매설된 상수관로의 누수 여부 상태를 확인하였다.

아래 그림은 누수 위험도 상위지역을 실제 방문한 현장 사진이다.

[그림 3-46] 상수도 누수 위험도 상위지역 현장 방문

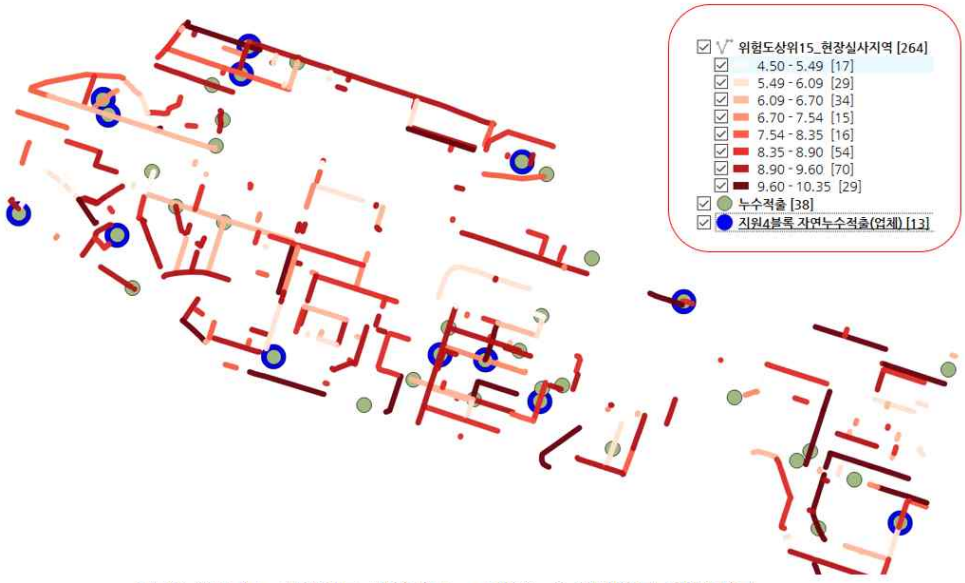
아래 그림은 누수점검업체 담당자가 누수 위험도가 높다고 판단된 상수관로를 실제 점검하는 현장과 점검 장비를 확인한 사진이다.

[그림 3-47] 누수점검업체와 상수관로 현장 확인

[그림 3-48] 누수점검업체 현장 점검 장비

4.1.2 현실적 검증 결과

데이터 수급이후 예측 분석결과가 산출되는 기간 동안 자연적으로 발생한 누수내역 데이터를 가지고 지자체 공간 내 특정 영역(블록 단위)에 분포하는 누수 위험도가 높은 264개 상수관로를 대상으로 사업초기 데이터 수급시점 이후부터 발생한 13개 자연누수 가운데 11개 관로가 서로 일치하여 수치상으로는 84% 예측 정확도를 나타냈다고 볼 수 있다.



[그림 3-49] 누수 위험도 상위관로 VS 자연누수발생 지점 비교

여기서 누수 위험도 상위관로 264개는 예측분석모델을 통해 산출한 K시 전체 누수 위험도 상위 판별관로 9,212개(K시 전체 상수관로의 15%)에 포함된다.

아래는 예측분석모델을 통해 산출된 누수 위험도 상위관로들에 대해 실제 누수가 발생한 상수관로들과 비교를 통해 일치여부를 확인한 결과이다.

[표 4-1] 누수 위험도 상위관로 VS 자연누수발생 일치 비교

관로번호	블록명	누수원인	일치 여부
39022	-	노후	○
659002460	-	노후	○
1531058	-	노후	○
13524	-	노후	○
39981	-	노후	○
51612	-	노후	X
40066	-	노후	○
37990	-	노후	X
39142	-	노후	○
38095	-	노후	○
39581	-	노후	○
41978	-	노후	○
659002416	-	노후	○

위의 표에서 보듯이 해당 블록의 13개 자연누수 발생중 11개 관로가 서로 예측분석 결과와 일치하여 84%의 판별 정확도를 나타냈다고 볼 수 있다.

최종 모형의 평가지표 중인 하나인 정확도가 86.74%인 점과 비교해 볼 때, 80% 대의 비슷한 수준의 결과 정확도를 보인다.

이것은 앞에서 선행연구들 가운데 예측 분석 결과에 대해 현실적 검증 단계를 적용한 체납 예측(63.85%), 유동인구수 예측(75%) 정확도와 비교해 볼 때, 현실적으로 적용해도 상수관로 누수 위험도 정확도(K시 샘플링 84% 정확도)가 상당히 높은 우수한 예측모델이라고 평가할 수 있다.

4.2 연구 모형 요약 및 비교

4.2.1 연구 모형 요약

본 연구의 연구모형 프로세스를 요약하면 다음과 같다.

[표 4-2] 연구모형 프로세스 요약

항 목	세부 내용	비 고
적용분야	상수관로 누수 위험도 예측	고장 예측 분야
데이터 수집	내부 데이터, 외부 데이터 수집	지자체 및 공공 데이터
데이터 전처리	상수관로 속성 데이터 추출, 데이터 지오코딩 및 공간 매칭 파생변수 생성, 변수 범주화, 결측치 처리	표준 데이터셋 생성 (QGIS)
데이터 분할	훈련세트 60%, 검증세트 20%, 테스트세트 20% 데이터 분할 생성	전처리(R) 오버샘플링 적용
모형 성능 비교 평가	로지스틱 회귀, 그래디언트부스팅, 랜덤포레스트	모델링(R)
최종 모형 채택	성능평가지표 : 정확도/F스코어/이항편차/AUC	누수 위험도 지수 (0.0~1.0)
분석 결과 시각화	누수 위험도를 상수관로와 결합 및 시각화	공간 시각화(QGIS)
현실적 검증	예측분석 결과와 실제 발생 데이터를 비교 평가	누수 위험도 상위 10% VS 누수발생내역 비교

4.2.2 선행 연구 요약

기존 선행 연구들을 요약하면 다음과 같다.

[표 4-3] 기존 선행연구 요약 및 비교

선행 연구	적용 분야	예측 기법	통계적 성능 지표	현실적 검증 항목
사례1	범죄율 예측	다중회귀분석(6개의 주요 변수 도출) 기반으로 AMOS 베이지안 추론 수행	ROC 곡선 AUC값이 0.861	집계구당 범죄발생 예측건수 VS 블록당 범죄 발생 건수(관할 경찰서 범죄관리대장) 비교
사례2	체납 예측	의사결정트리 및 로지스틱회귀	의사결정트리(76.294%, ±0.075, 로지스틱회귀(75.654%, ±0.035)	체납 예측 결과와 실제 납부 형태 비교
사례3	영화 흥행 예측	회귀분석(유의미한 변수) 및 나이브 베이즈 분류	나이브 베이즈 분류의 정확도(Accuracy)는 91%	총 관객 수가 2,650,000명 이상 흥행 판단, 56편 중 51편의 영화가 실제 흥행 여부와 동일
사례4	온라인 판매 예측	시계열 예측방법 ARIMA 모형	반팔 티셔츠와 아우터웨어의 예측값과 실제 판매량의 오차율은 각각 ±1.5%, ±8%	2018년도 판매량 예측 VS 실제 2018년 판매량 비교
사례5	유동인구수 예측	다중 회귀 분석, 의사결정나무, 시계열 분석	회귀분석 60%, 의사결정나무 75%, 시계열 예측 67%	인구수 예측값 VS 유원지 시간대별 인구 수

4.2 기존 연구와의 차별성

본 연구모델의 차별성은 기존 선행연구들과는 달리 예측한 결과를 실제 발생한 데이터와 비교한 점과 선행연구들이 예측결과 값의 해상도(분석 단위)가 너무 낮아 현실에 적용하기가 무리가 있는 반면, 본 연구모형은 실제 표준화 모델로 개발되어 지자체에 실제 적용해서 검증된 모형이라는 차이점이 있다.

[표 4-4] 기존 연구와의 차별성

항 목	적용 분야	예측 기법	통계적 성능 지표	현실적 검증 항목	실제검증
연구모형	상수도 누수 위험도 예측	로지스틱 회귀, 그래디언트부스팅, 랜덤포레스트	정확도/F스코어/이항편차/AUC (93.99%/0.9361/2424.67/0.8568)	누수 위험도 상위 10% VS 누수발생 내역 비교(6개월 발생데이터 80% 일치)	O 자연누수 검증
선행연구	범죄율 예측	다중회귀분석(6개의 주요 변수 도출) 기반으로 AMOS 베이지안 추론 수행	ROC 곡선 AUC값이 0.861	집계구당 범죄발생 예측건수 VS 블록당 범죄 발생 건수(관할 경찰서 범죄관리대장) 비교	X (과거 자료)
	체납 예측	의사결정트리 및 로지스틱회귀	의사결정트리(76.294%, ±0.075, 로지스틱회귀(75.654%, ±0.035)	체납 예측 결과와 실제 납부 현황 데이터 비교	△ (정확도 낮음)
	영화흥행 예측	회귀분석(유의미한 변수) 및 나이브 베이즈 분류	나이브 베이즈 분류의 정확도(Accuracy)는 91%	총 관객 수가 2,650,000명 이상 흥행 판단, 56편 중 51편의 영화가 실제 흥행 여부와 동일	△ (해상도 낮음)
	온라인 판매 예측	시계열 예측방법 ARIMA 모형	반팔 티셔츠와 아우터웨어의 예측값과 실제 판매량의 오차율은 각각 ±1.5%, ±8%	2018년도 판매량 예측 VS 2018년 판매량 비교	X (과거 자료)
	유동인구수 예측	다중 회귀 분석, 의사결정나무, 시계열 분석	회귀분석 60%, 의사결정나무 75%, 시계열 예측 67%	인구수 예측값 VS 유원지 시간대별 인구 수	△ (정확도 낮음)

Ⅴ. 연구의 한계 및 극복 방안

5.1 연구의 한계점

머신러닝(Machine Learning) 예측 모델의 특성상 예측 모델의 결과에 영향을 미치는 요인(변수)들은 매우 다양하고 복잡한데 현실적으로 이 요인들을 모두 고려하지 못하는 한계가 있으며, 모델을 만드는데 있어서 정제된 데이터의 양이 충분히 많지 않고 제한적인 데이터만 사용한다는 현실적인 문제가 선행연구 모델 대부분에 존재했다고 추측되므로 추가적인 핵심 변수의 확보와 높은 성능을 위한 모델 정교화 작업이 필요하다고 하겠다.

그리고 빅데이터 분석 모형의 성능 평가 검증과 실제 발생 데이터 기반으로 현실적 검증 간의 비교를 업종별 분야별로 다양하게 확인해 보려 하였으나 그 사례들이 충분히 많지 않아서 만족할만한 시사점을 도출 할 수는 없었다.

5.2 향후 발전 방향

빅데이터 예측 모델의 성능 평가와 별개로 예측 시점 이후에 실제로 발생한 데이터에 의한 현실적 검증을 효과적으로 할 수 있는 방안을 업종별 분야별로 연구를 진행하여 하나의 〈예측 모델의 현실적 검증 평가방법론〉분야로 발전 가능하도록 정립이 필요하다고 하겠다.

참고문헌

1.국내문헌

김주영. (2018). 공간 빅데이터를 활용한 범죄 발생 위험지역 예측 모형 구축. 경상대학교 석사학위논문.

정재안. (2019). 빅데이터 기반 체납 수용가 예측 모델 개발. 배재대학교 석사학위논문.

김세윤. (2018). 데이터 분석을 통한 한국영화 흥행 예측. 숭실대학교 석사학위논문.

오지연. (2019). 빅데이터 분석을 활용한 온라인 판매 수요 예측. 한신대학교 석사학위논문.

허찬. (2018). 기지국 데이터를 활용한 기계학습 기반의 유동인구 분석 및 예측 모델 연구 고려대학교 석사학위논문.

행정안전부. (2017). 공공빅데이터 상수도 누수 위험도 표준분석모델 매뉴얼.

LX국토정보공사. (2016). 빅데이터 활용 상수도 누수 절감 의사결정 지원

동경수도국. (2015). 동경 수도국 유수율 향상 사례 연구

ABSTRACT

A Study on the Establishment and Verification of the Prediction Model for Water Leakage Risk through the Analysis of Big Data

Park, Jin-Woo

Major in Smart Convergence Technical Consulting

Dept. of Smart Convergence Consulting

Graduate School of Knowledge Service Consulting

Hansung University

Looking at the various cases in the public and private sectors where the big data predictive analysis model is applied, the performance verification by the model evaluation of the machine learning predictive analysis model is basically included, but there are fewer realistic verification cases than expected to confirm that the results of the analysis predicted by the machine learning technique actually fit the data.

Therefore, we would like to look at the cases in each field that predict crime rates, overdue charges, movie shows, online sales, and floating population, and look for implications and development directions so that we can establish them as a "Realistic Verification Evaluation Methodology for Predictive Models" for effective practical verification.

【Keywords】 Big data analysis, water leak risk, machine learning, predictive model, public big data