KJCEM 21. 6. 046~055
November 30, 2020

www.kicem.or.kr (pISSN 2005-6095 / eISSN 2465-9703)
http://dx.doi.org/10.6106/KJCEM.2020.21.6.046

XGBoost를 활용한 고속도로 콘크리트 포장 파손 예측

이용준[1*] · 선종완[2]

[1]한국건설기술연구원 인프라안전연구본부 도로관리통합센터 박사후연구원 · [2]한국건설기술연구원 인프라안전연구본부 도로관리통합센터 수석연구원

Predicting Highway Concrete Pavement Damage using XGBoost

Lee, Yongjun[1*], Sun, Jongwan[2]

[1]Post-Doctoral Researcher, Department of Infrastructure Safety Research, Korea Institute of Civil Engineering and Building Technology
[2]Senior Researcher, Department of Infrastructure Safety Research, Korea Institute of Civil Engineering and Building Technology

Abstract : The maintenance cost for highway pavement is gradually increasing due to the continuous increase in road extension as well as increase in the number of old routes that have passed the public period. As a result, there is a need for a method of minimizing costs through preventative grievance Preventive maintenance requires the establishment of a strategic plan through accurate prediction old Highway pavement. herefore, in this study, the XGBoost among machine learning classification-based models was used to develop a highway pavement damage prediction model. First, we solved the imbalanced data issue through data sampling, then developed a predictive model using the XGBoost. This predictive model was evaluated through performance indicators such as accuracy and F1 score. As a result, the over-sampling method showed the best performance result. On the other hand, the main variables affecting road damage were calculated in the order of the number of years of service, ESAL, and the number of days below the minimum temperature -2 degrees Celsius. If the performance of the prediction model is improved through more data accumulation and detailed data pre-processing in the future, it is expected that more accurate prediction of maintenance-required sections will be possible. In addition, it is expected to be used as important basic information for estimating the highway pavement maintenance budget in the future.

Keywords : Highway Pavement, Damage Prediction, Maintenance, Machine Learning, XGBoost

1. 서론

1.1 연구의 배경 및 목적

현재 급속한 도시화로 인하여 고속도로는 연장의 지속적인 증가와 함께 국민 생활환경과 산업 활동의 기반시설로서 그 역할이 매우 중요하게 인식되고 있다. 그러나 건설 후 공용년수가 오래된 노후 노선이 증가함에 따라 유지관리의 중요성이 점차 강조되고 있다. 도로 포장은 건설과 동시에 환경적, 기능적, 구조적 요인 등 다양한 요인에 의해 다양한 형태로 파손이 시작되고 발전되며 보수의 시기를 놓치거나 적절한 유지보수가 이루어지지 않는다면 막대한 보수비용을 발생시킨다(Korea Expressway Corporation, 2018). 이에 관리자의 입장에서는 한정적인 보수예산을 효율적으로 활용하기 위한 전략적인 유지관리 계획 수립이 필요하다.

전략적인 유지관리 계획 수립을 위해서는 포장의 유지보수 시기 및 공법 결정을 위한 의사결정체계가 필요하며, 이 과정에서 해당 도로구간의 파손 예측이 선행되어야 한다. 즉, 과거 및 현재의 포장상태를 기반으로 파손 예측을 통한 잔존수명(기대수명)의 추정과정은 유지보수가 필요한 구간의 예측과 예산의 추정에 중요한 기초정보로 활용되기 때문이다(Do et al., 2011). 도로포장 분야에서 일반적으로 말하는 열화예측 혹은 파손모형(deterioration model)이란 포장의 열화과정에 대한 해석을 다루는 내용으로 주로 유지보수가 요구되기까지의 기대수명과 그 기대수명 내에서 열화특성에 따른 상태 변화과정을 표현하는 것으로 정의할 수 있다(Kwon et al., 2012).

일반적으로 도로포장의 파손 예측 모델은 회귀분석법, 생존분석법, Markov 모델, 확률론적 기반 모델 등 다양하며최근 컴퓨터 사양의 급격한 발전과 함께 기계학습 라이브러리 개발로 인하여 머신러닝 기법을 활용한 예측 모델 개발이

* **Corresponding author:** Lee, Yongjun, Department of Architectural Engineering, Korea Institute of Civil Engineering and Building Technology, Goyang-Si 10223, Korea
E-mail: yongjunlee@kict.re.kr
Received August 25, 2020: **revised** -
accepted September 17, 2020

여러 분야에서 연구가 활발하게 진행되고 있다(Lee et al., 2019).

따라서 본 연구에서는 고속도로 콘크리트 포장 열화 예측을 위해 머신러닝 기법인 분류학습을 기반으로 하는 파손 예측 방법을 제안하려 한다.

1.2 연구의 범위 및 방법

도로포장은 한순간에 파손되는 것이 아니라 공용년수에 따른 여러 변수들에 의한 스트레스 누적으로 인하여 점차 열화가 진행되다가 파손된다. 머신러닝 분석을 통한 파손 예측을 위해서는 영향인자와 그 결과(포장상태)로 구성된 데이터 셋을 구축하여야 한다.

이에 본 연구에서는 도로파손에 영향을 주는 변수들을 기존문헌 고찰을 통해 축하중 교통량(Equivalent Single Axle Load; ESAL) 및 환경 변수 등 총 7개의 독립변수 데이터와 고속도로 포장상태지수(Highway Pavement Condition Index; HPCI)를 기반으로 한 포장 상태를 종속변수로 하는 데이터 셋을 구축하였다.

구축된 데이터 셋을 기반으로 의사결정나무 기법 중 하나인 XGBoost를 활용한 머신러닝 분석을 수행하여 고속도로 콘크리트 포장 파손 예측모델을 개발하였으며 그 성능을 평가하였다.

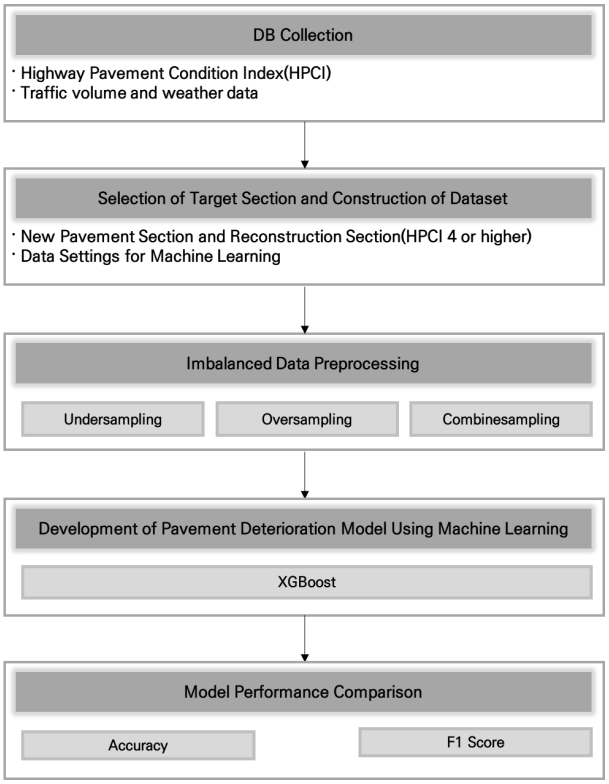

Fig. 1. Research Methodology

2. 기존 연구 고찰

도로포장 파손예측 모델과 관련된 연구는 크게 통계적인 방법론과 마르코프 체인 등을 활용한 확률론적 기법으로 구분되어 왔다.

먼저 통계적인 방법으로는 Moon et al. (2017)은 노후 콘크리트 포장 전체 구간의 파손 종류와 파손에 영향을 미치는 환경 인자를 분석하고 회귀분석을 통해 평탄성과 표면손상에 대한 예측 모델을 개발하고 활용방안을 제시하였다.

마르코프 체인 모델을 활용한 방법으로는 Han et al. (2017)은 일반국도 포장상태평가지수인 NHPCI (National Highway Pavement Condition Index) 데이터를 기반으로 ESAL과 포장강도지수인 SNP (Structural Number of Pavement)를 설명변수로 활용하여 모형을 개발하였으며 일반국도 포장의 서비스 등급별 기대수명과 파손속도의 변화, 그리고 이 파손과정의 불확실성을 제시하였다.

최근 들어 많은 양의 데이터를 빠르고 다양하게 분석이 가능한 딥러닝 기법을 활용하여 도로포장의 파손을 예측하는 연구가 활발히 이루어지고 있다.

Huyan, J. (2019)는 포장의 전반적인 상태를 예측할 수 있는 상태 데이터 셋을 기반으로 XGBoost 모델과 Gradient Boosting Decision Tree (GBDT) 모델을 활용하여 예측모델을 개발하고 성능을 비교하였으며, 연구 결과, 대부분의 조건에서 XGBoost 방법이 GBDT보다 성능이 우수한 것으로 분석되었다.

Gong et al. (2019)은 XGBoost 모델과 랜덤포레스트 모델을 활용하여 거북등 균열과 세로 균열 두 가지 유형의 피로균열을 예측하기 위해 트리 기반 예측모델을 제안하였으며, 연구결과 Gradient Boosting Machine (GBM)기반인 XGBoost 모델이 랜덤포레스트보다 예측성능이 더 우수한 것으로 분석되었다.

도로포장 파손 예측은 아니지만 도로시설물인 교량에 대한 파손 예측 모델에도 머신러닝 기법이 활용되었다. Lim et al. (2019)은 프리스트레스트 콘크리트 I형 교량의 바닥판을 대상으로 파손 예측 모델을 개발하였다. 모델은 인공지능 모델인 심층신경망(Deep Neural Networks; DNN) 기법과 XGBoost를 활용하여 개발하였으며 연구결과 더 높은 정확도를 보이는 XGBoost 모델을 최종 모델로 선정하였다.

국내외 연구의 동향을 살펴보면 기존 통계분석이나 마르코프 체인과 같은 확률론적인 방법도 많이 활용하고 있지만 방대한 양의 데이터를 빠르게 분석하고 데이터가 축적될수록 예측의 정확도가 향상되는 머신러닝의 활용도 높아지고 있다.

따라서 본 연구에서는 기존문헌 고찰을 통해 예측 성능이

우수한 의사결정나무기법 중 부스팅 방식인 XGBoost를 활용하여 국내 고속도로 콘크리트 포장 파손을 예측하는 기법을 제안하고 성능을 평가하였다.

3. 머신러닝 방법론

XGBoost는 선형 모델이나 트리 기반 모델에서의 과적합 문제를 해결하고, 규모가 큰 데이터 셋의 안정성과 훈련 속도를 향상시키기 위한 목적으로 Tianqi Chen과 Carlos Guestrin이 소개한 방법이다. eXtreme Gradient Boosting의 약자로 boosting algorithm 기반 모델이며, 회귀와 분류, 순위 및 사용자 정의 objective을 지원하는 유연한 모델이다(T. Chen et al., 2017).

분류학습 중 부스팅 방식은 배깅과 유사하게 초기 샘플 데이터를 조작하여 다수의 분류기를 생성하는 기법 중 하나지만 가장 큰 차이는 순차적 방법이라는 것이다. 부스팅 기법은 이전 분류기의 학습 결과를 토대로 다음 분류기의 학습 데이터의 샘플 가중치를 조정해 학습을 진행하는 방법이다.

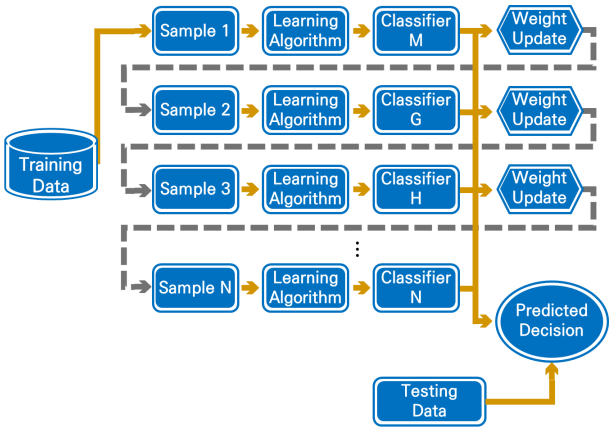

Fig. 2. Boosting Flow Chart

〈Fig. 2〉와 같이 부스팅 방식은 먼저 학습 데이터와 테스트 데이터를 적당한 비율로 랜덤하게 추출하여 분할한다. 그 다음 테스트 셋에서 부스트트랩 샘플링 기법을 이용해 샘플을 추출하고 특정한 학습 알고리즘에 적용하여 분류기를 생성한다. 이렇게 생성된 분류기의 분류결과를 통해 잘못 분류한 데이터와 추출되지 않은(학습에 이용되지 않은) 데이터에는 가중치를 부여하여 다음 학습에 이용한다. 이러한 일련의 과정을 부스팅 라운드라고 한다. 이렇게 총 n번의 부스팅 라운드를 거쳐서 완성된 모형들을 이용해 최종적인 분류모형이 만들어지게 된다.

예를 들어, 학습기 M에 대해 Y를 예측할 확률은 다음과 같다〈Fig. 2〉.

$$Y = M(X) + error_1 \quad (1)$$

$error_1$에 대해 조금 더 상세히 분류할 수 있는 모델 G가 있다면(단, $error_1 \rangle error_2$) 식 (2)로 표현할 수 있을 것이다.

$$error_1 = G(X) + error_2 \quad (2)$$

여기에 $error_2$를 더 세밀하게 분리할 수 있는 모델 H가 존재한다면(단, $error_2 \rangle error_3$) 식 (3)으로 표현할 수 있을 것이다.

$$error_2 = H(X) + error_3 \quad (3)$$

식 (1)에서 식 (2)와 식 (3)을 적용하면 식 (4)로 표현할 수 있다.

$$Y = M(X) + G(X) + H(X) + error_3 \quad (4)$$

이렇게 하면 학습기 M을 단독으로 사용했을 때보다 정확도가 높아진다. 그러나 분류기 M, G, H의 성능이 각각 다른데, 모두 같은 비중으로 분류를 진행하게 되면 임의의 데이터에 대해 간섭하며 오류를 높일 수 있으므로 각 모델 앞에 비중(weights)을 두고 머신러닝으로 최적의 비중을 찾아 식 (4)의 모델보다 훨씬 더 좋은 성능($error_3 \rangle error_4$)을 내는 분류기를 만들 수 있다.

근래의 그래디언트 부스팅(Gradient Boosting)의 경우 뛰어난 예측 성능을 가지고 있지만, 수행 시간이 오래 걸리는 단점으로 인해 최적화 모델 튜닝이 어려웠다. 하지만 XGBoost와 LihgtGBM 등 기존 그래디어트 부스팅의 예측 성능을 한 단계 발전시키면서도 수행 시간을 단축시킨 알고리즘이 계속 등장하면서 정형 데이터의 분류 영역에서 가장 활용도가 높은 알고리즘으로 부스팅 계열 알고리즘에서 가장 각광을 받고 있다. 압도적인 수치의 차이는 아니지만, 분류에 있어서 일반적으로 다른 머신러닝보다 뛰어난 예측 성능을 나타낸다.

4. 머신러닝을 활용한 파손 예측 모델 개발

4.1 대상구간 및 변수 선정

본 연구에서는 고속도로 포장 파손 예측 모델 개발을 위해 2008년~2017년까지의 HPCI 데이터를 고속도로 포장관리시스템에서 수집하였다.

고속도로 데이터의 포장형식은 총 6개의 형식으로 AP교면, AP복합단면, AP토공, CP교면, CRCP, JCP로 나뉘어져있

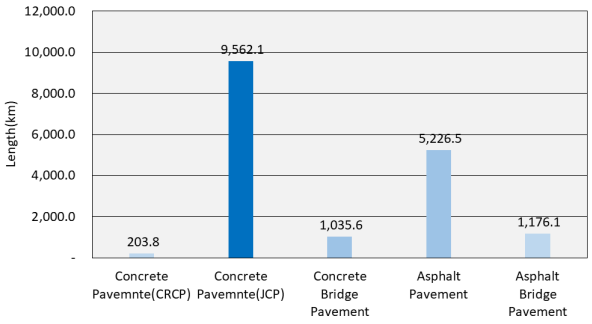

Fig. 3. Ratio by Pavement Type (2018)

으며 그 중 많은 비중을 차지하고 있는 콘크리트 JCP를 대상으로 분석하였다〈Fig. 3〉.

콘크리트 JCP포장의 HPCI는 포장상태조사를 통하여 확보된 표면손상 및 평탄성의 정량화된 데이터를 변수로 식(5)와 같이 산정되며 유지보수 대상구간 선정 및 각종 현황분석에 활용된다.

$$HPCI_{10m} = 5 - 0.8 \times IRI^{0.7} - 0.85 \times \log(1 + 10 \times 2.5 \times SD) \quad (5)$$

여기서, $HPCI_{100m} = \sum_{n}^{n+10} (HPCI_{10m})$

평탄성(International Roughness Index) = 평탄성지수(m/km)

표면손상(Surface Distress) = 노면손상 환산면적(㎡)

〈Table 1〉은 고속도로 도로포장상태조사 평가결과에 의한 등급기준이다.

Table 1. Leveling Standard According to The Highway Pavement Condition Index (HPCI)

Level	HPCI	Condition	Measure
Level 1	More than 4.00	Very good	Do nothing
Level 2	More than 3.50 and less than 4.00	Good	Preventive repair
Level 3	More than 3.25 and less than 3.50	Somewhat good	Repair and maintenance if necessary
Level 4	More than 3.00 and less than 3.25	Normal	Repair and maintenance
Level 5	More than 2.50 and less than 3.00	Somewhat poor	Improvement if necessary
Level 6	More than 2.00 and less than 2.50	Poor	Improvement
Level 7	Less than 2.00	Very poor	Improvement on a preferential basis

source: Korea Expressway Corporation (2018)

포장은 시간이 경과됨에 따라 여러 파손원인 변수들에 의한 스트레스 누적으로 인하여 점차 파손되는 것이기 때문에 도로포장의 정확한 공용연수를 파악하여, 누적된 하중량을 고려할 수 있어야 한다.

그러나 수집된 고속도로 포장 상태 데이터는 유지보수 이력데이터가 누락되어 있어 대상구간의 공용년수가 불확실하였다.

따라서 초기 상태에서 공용연수의 증가에 따른 포장 파손을 예측하기 위해, 수집한 데이터 범위의 기준년도가 되는 해당년도에 데이터 중 HPCI가 4 이상이며, SD 값이 $0m^2$, IRI 값이 1.18m/km 이하인 대상만을 초기 상태로 정의하였다〈Table. 1. 참조〉.

IRI의 값의 경우, 국내 도로포장구조설계요령(2011)에서 초기 IRI의 값을 1.18m/km~1.41m/km로 제시하고 있어, 본 연구에서는 데이터의 신뢰성 확보를 위하여 1.18m/km 값을 기준으로 하였다.

이러한 데이터 가공 과정을 거쳐 〈Table 2〉와 〈Fig. 4〉와 같이 19개 노선, 42개 지역에서 총 34,449개 데이터를 수집하였다.

Table 2. Data Collection Status

	N	Route No.	Area
Data Collection Status	34,449	1, 10, 12, 15, 16, 20, 25, 30, 35, 37, 40, 45, 50, 55, 60, 65, 110, 253, 300	Seoul et 41 area

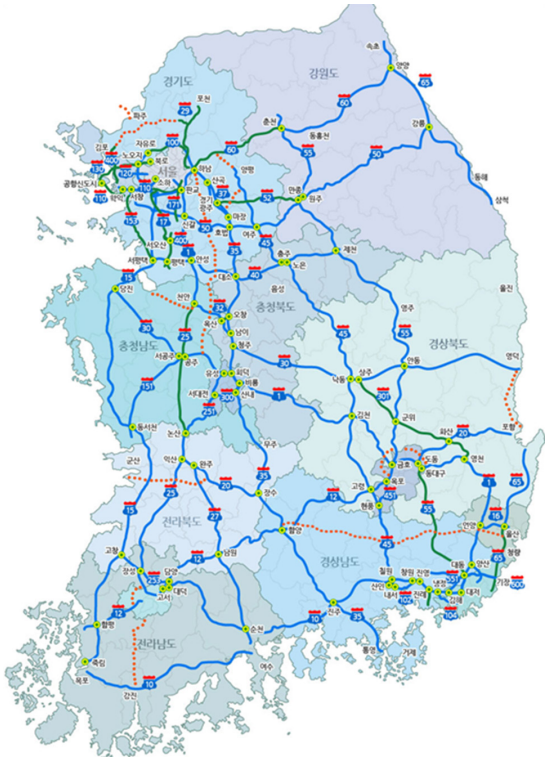

Fig. 4. Route Information for The Target Section

최근 들어 폭염, 한파, 집중호우 등 국내외 이상기후 현상으로 인해 1950년대 이후 기온 상승률은 20세기 전체 기간

에 비하여 약 1.5배 이상 증가하였다. 특히, 최근 100년 동안 국내 평균기온이 약 0.74℃ 상승되는 등 급격한 기후변화로 인하여 집중적인 강우현상, 폭설 등 기상이변이 증가하고 있는 추세이다(Ko, 2020). 이와 같이 환경의 변화로 도로포장의 상황은 더욱 열악해짐에 따라 표면손상, 소성변형, 포트홀 등과 같은 조기 파손으로 평균수명이 점점 감소하고 있다.

이에 한국도로공사(2010)에서는 동결융해와 제설재의 영향을 받는 콘크리트 도로시설물의 내구성능 저하 방지 및 경감을 위해 강설일, 강설량, 온도특성, 해발을 〈Table 3〉과 같이 고려하고 있다. 특수환경에 해당하는 지역의 정의는 ① ~ ④ 항목 중 3개 항목 이상이 포함되어야 하고, ①은 반드시 해당되어야 하며, ⑤번 항목은 하나에만 해당되어도 특수환경 노출지역으로 구분하였다(Koh et al., 2015).

Table 3. Exposure Environment Grading Guideline

		Instruction
① Number of freezing and thawing days per year	Daily average temperature (Note 1) less than 0℃	45 days or more
	Lowest daily average temperature (Note 2) less than -2℃	90 days or more
② Annual deicing agent usage		13 tons / two-lane·km or more
③ Annual accumulated amount of snowfall		60cm or more
④ Number of snowing days per year		14 days or more
⑤ Average altitude of the route		450m or more

source: Koh et al. (2015)

또한, 포장의 수명(공용성)은 교통량, 대형차량 비율, 평탄성 등 다양한 인자들의 영향을 받으며, 특히 교통하중 특성과 기초의 두께 등이 포장의 수명에 큰 영향을 미친다(You et al., 2002; Loizos et al., 2005; Yang et al., 2005).

Table 4. Data Set

		Average	Standard deviation	Minimum value	Maximum value
Environment Factors	1. Service Year	5.6	2.1	2.0	8.0
	2. Number of days with intense heat	104.0	69.8	1.0	283.0
	3. Precipitation	7,857.5	2,739.6	2,306.0	14,095.0
	4. Number of days with the average temperature of less than 0℃	45.0	46.1	0.0	166.0
	5. Number of days with the lowest temperature of less than -2℃	279.4	131.3	18.0	610.0
	6. Number of snow days	500.5	209.8	61.0	1,038.0
Transportation Factor	7. ESAL	39,329.9	38,133.1	2.0	215,440.0

교통변수의 경우 총 교통량(Annual Average Daily Traffic; AADT)과 같은 절대 값 보다는 하중의 강도 예를 들어, ESAL 크기가 포장의 수명에 더 큰 영향을 미친다는 연구결과가 많이 있다(Gharaibeh et al., 2003; Lee, 2013).

따라서 본 연구에서는 〈Table 4〉와 같이 고속도로 콘크리트 포장에 파손을 주는 인자인 공용년수, 폭염일수, 강수량, 평균기온 0도 이하일수, 최저기온 −2도 이하 일수, 적설일수, ESAL을 입력데이터로 하는 데이터 셋을 구축하였다. 그러나 대상 구간의 위치 좌표 정보가 누락되어 있어 기상자료를 정확하게 맵핑하기에는 다소 어려워 대상구간과 가장 가까운 기상관측소 데이터를 활용하였다.

4.2 데이터 전처리

데이터 분류는 〈Table 5〉와 같이 이진분류 기법을 활용하기 위하여 고속도로 콘크리트 포장 유지보수 기준인 HPCI가 3이하일 경우 유지보수가 필요한 구간〈Table 1〉, 즉 파손된 구간으로 분류하여, 파손으로 인한 유지보수가 필요한 구간은 '1', 유지보수가 필요하지 않은 구간은 '0'으로 데이터를 분류하였다.

Table 5. Data Classification

	HPCI	Service Year	Classification
Section A	3.2	6	0
Section B	2.8	4	1
Section C	1.5	8	1
⋮			
Section N	4.2	2	0

〈Table 6〉은 데이터 분류 결과이며 유지보수가 필요하지 않은 구간(0) 데이터가 총 데이터의 약 92% 비율을 차지하고 있어 비대칭 데이터 형태를 보이고 있다.

Table 6. Result of Data Classification

	No Damage (0)	Damage (1)
N	31,716 (92.1%)	2,733 (7.9%)

클래스 레이블이 불균형한 분포를 가진 데이터 세트를 학습시킬 때 예측 성능의 문제가 발생할 수 있는데, 이는 이상 레이블을 가지는 데이터 건수는 매우 적기 때문에 제대로 다양한 유형을 학습하지 못하는 반면에 정상 레이블을 가지는 데이터 건수는 매우 많기 때문에 일방적으로 정상 레이블로 치우친 학습을 수행해 제대로 된 이상 데이터 검출이 어려워지기 쉽다. 즉, 현재 수집된 데이터를 통해 도로포장 파손 예측 모델을 개발하게 되면 대부분의 결과들이 유지보

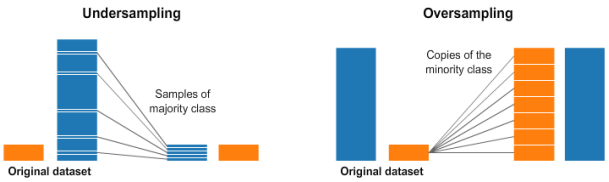

Fig. 5. Data Sampling Methodology

수가 필요 없는 구간으로 예측할 것이다.

지도학습에서 극도로 불균형한 레이블 값 분포로 인한 문제점을 해결하기 위해서는 적절한 학습데이터를 확보하는 방안이 필요한데 대표적으로 언더 샘플링과 오버 샘플링 방법이 있다〈Fig. 5〉.

언더 샘플링은 많은 데이터 세트를 적은 데이터 세트 수준으로 감소시키는 방식이며, 정상 레이블 데이터를 이상 레이블 데이터 수준으로 줄여 버린 상태에서 학습을 수행하여 과도하게 정상 레이블로 학습/예측하는 부작용을 개선할 수 있지만, 너무 많은 정상 레이블 데이터를 감소시키기 때문에 정상 레이블의 경우 오히려 제대로 된 학습을 수행할 수 없다는 단점이 있어 잘 적용하지 않는 방법이다. 본 연구에서는 언더 샘플링 기법중 Editetd Nearest Neighbours (ENN) 방법을 활용하였다. ENN은 K-Nearest Neighbor (KNN)을 사용해 다수 클래스 데이터를 축소, 이웃한 데이터 중 자신과 같은 클래스보다 다른 클래스의 데이터가 많을 경우 해당 데이터는 제외하는 방법이다(Dennis, L. et al., 1972).

오버 샘플링은 이상 데이터와 같이 적은 데이터 세트를 증식하여 학습을 위한 충분한 데이터를 확보하는 방법이다. 동일한 데이터를 단순히 증식하는 방법은 과적합이 되기 때문에 의미가 없으므로 원본 데이터의 피처 값들을 아주 약간만 변경하여 증가시킨다. 본 연구에서는 오버 샘플링 방법 중 Synthetic Minority Over-sampling Technique (SMOTE) 활용하였다. SMOTE는 적은 데이터 세트에 있는 개별 데이터들의 K 최근접 이웃을 찾아서 이 데이터와 K개 이웃들의 차이를 일정 값으로 만들어서 기존 데이터와 약간 차이가 나는 새로운 데이터들을 생성하는 방식이다.

또한, 언더 샘플링과 오버 샘플링을 함께 결합한 혼합 샘플링 방법도 존재한다. 혼합 샘플링은 다수 클래스 데이터를 중요하지 않은 개체 수를 제거한 후, 이상 데이터에서 개체 수를 증가시키는 기법으로, 오버샘플링 기법에 비해 자료의 노이즈나 과적합 문제를 줄일 수 있고 언더 샘플링 기법에 비해 데이터 손실을 줄일 수 있다(Batista et al., 2004).

본 연구에서는 SMOTE와 ENN을 함께 사용하는 혼합 샘플링 방법을 활용하였다

따라서 본 연구에서는 여러 데이터 샘플링 기법을 통해 〈Fig. 6〉과 같이 여러 개의 데이터 셋을 구축하여 XGBoost 기법을 활용한 파손 예측 모델을 개발하였다

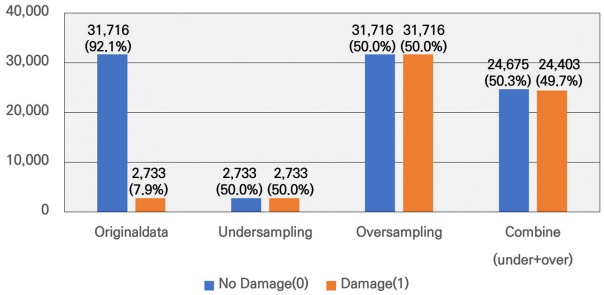

Fig. 6. Result of Data Sampling

4.3 최적 파라미터 선정

하이퍼 파라미터는 머신러닝 알고리즘을 구성하는 주요 구성 요소이며, 파라미터 값을 조정해 알고리즘의 예측 성능을 개선할 수 있다.

GridSearchCV API를 이용해 Classifier나 Regressor와 같은 알고리즘에 사용되는 하이퍼 파라미터를 순차적으로 입력하면서 편리하게 최적의 파라미터를 도출할 수 있는 방안을 제공한다. 즉, 데이터 세트를 Cross-validation을 위한 Train/Test 세트로 자동으로 분할한 뒤에 하이퍼 파리미터 그리드에 기술된 모든 파라미터를 순차적으로 적용해 최적의 파라미터를 찾을 수 있게 해준다.

본 연구에서는 〈Fig. 7〉과 같은 모델링 구조를 가진 XGBoost 모델의 파라미터 튜닝을 위해서 먼저 Learning Rate를 선택하고 이 학습률에 맞는 트리 개수를 선정하였다. 그리고 max_depth, min_child_weight, gamma 등의 파라

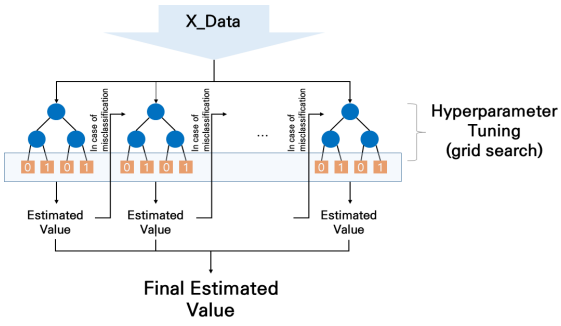

Fig. 7. XGBoost Modelling Structure

Table 7. Set of Values for Best Hyperparameters

	Original Data	Under Sample	Over Sample	Combine Sample
learning_rate	0.01	0.01	0.01	0.01
n_estimators	10,000	10,000	10,000	10,000
max_depth	6	9	3	9
min_child_weight	3	5	3	1
gamma	0.1	0.3	0.0	0.0
reg_alpha	1e-05	1e-05	1e-05	1e-05
subsample	0.5	0.7	0.8	0.7
colsample_bytree	0.8	0.9	0.8	0.8

미터를 선정하고 다시 학습률을 낮춰 위의 과정을 반복하여 최적 파라미터를 선정하였다.

〈Table 7〉은 각각의 샘플링된 데이터들을 적용한 XGBoost 모델의 최적 파라미터 선정 결과를 요약한 표이다.

4.4 모델 성능 평가

현재 수집할 수 없는 미래의 데이터를 투입하는 경우를 모사하기 위해, 데이터를 8:2 비율로 학습(training)과 검증(testing) 데이터로 분할하였으며, 학습 데이터는 다시 5등분하여, 그 중 하나를 학습된 모델의 성능을 확인(validation)하는 데이터로 확인하고 나머지 4/5를 학습에 활용하는 과정을 확인 데이터 셋으로 바꾸는 과정을 총 5회 반복하여 수행하는 5-fold cross validation을 사용하여 검증하였다.

또한, 과적합을 방지하기 위하여 오류함수의 성능 개선이 없을 경우 조기 중단할 수 있는 최소 반복 횟수를 100회로 선정하였다. 마지막으로 데이터 전처리를 통해 임의적으로 뽑은 샘플이 편향 됐을 가능성이 있으므로 검증 데이터는 원본 데이터(Original Data)를 활용하여 예측모델의 성능을 평가하였다.

〈Table 8〉과 〈Fig. 8〉은 예측모델의 학습 반복 횟수에 따른 손실함수에 대한 결과이다. 샘플링 된 데이터를 적용한 예측 모델들 모두 과적합은 발생하지 않았으나 언더 샘플링 데이터를 활용한 예측 모델의 경우 다른 샘플링 데이터를 활용한 모델들에 비해 학습 데이터와 검증 데이터의 손실함수 차이가 커 성능이 다소 떨어질 것으로 보인다. 이는 과도하게 너무 많은 정상 레이블 데이터를 감소시켰기 때문인 것으로 판단된다.

Table 8. Loss Function Result

	Original Data	Under Sample	Over Sample	Combine Sample
n_tors(Early Stopping)	1,914	748	4,979	986
Train Log Loss	0.13	0.28	0.18	0.12
Test Log Loss	0.14	0.36	0.20	0.15

분류 예측 모델에 사용되는 성능 평가 지표는 〈Table 9〉와 같이 학습된 분류 모델이 예측을 수행하면서 얼마나 오차가 발생하는지 함께 보여주는 지표인 오차행렬(Confusion Matrix)를 기반으로 산정이 가능하다.

Table .9 Confusion Matrix

Predict \ Actual	True	False
True	True Positive(TP)	False Positive(FP)
False	False Negative(TN)	True Negative(TN)

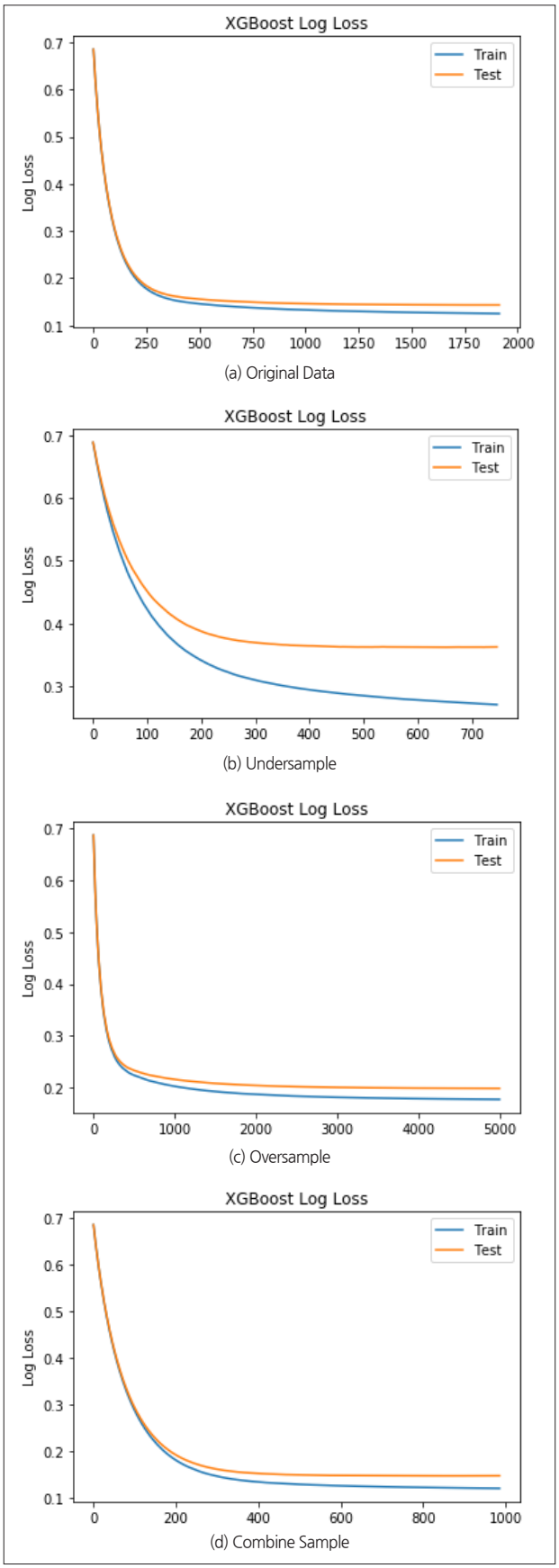

Fig. 8. Loss Function according to the Number of Training Iterations

〈Table 10〉은 오차행렬 값을 조합하여 예측모델의 성능을 측정할 수 있는 지표들을 요약한 것이다.

Table 10. Performance Metrics

	Performance Metrics
Accuracy	(TP + TN) / (TP+FN+FP+TN) ; Accuracy is how close a measured value is to the actual (true) value.
Precision	TP / (TP + FP) ; Precision quantifies the number of positive class predictions that actually belong to the positive class.
Recall	TP / (TP + FN) ; Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.
F1 score	2 / (1 / Precision + 1 / Recall) ; F-Measure provides a single score that balances both the concerns of precision and recall in one number.

〈Table 11〉은 데이터 샘플링에 따른 예측모델의 정확도를 비교한 결과로써 예측모델들의 정확도는 0.90 이상으로 매우 높았으나 언더 샘플링 된 데이터를 적용한 예측모델이 다른 샘플링 된 데이터를 적용한 예측 모델들보다 정확도가 다소 떨어지는 것으로 분석되었다.

Table 11. Accuracy Result

	Original Data	Under Sample	Over Sample	Combine Sample
Accuracy	0.95	0.91	0.95	0.95

분류의 평가방법도 일반적으로는 실제 결과 데이터와 예측 결과 데이터가 얼마나 정확하고 오류가 적게 발생하는가에 기반하지만, 단순히 이러한 정확도만 가지고 판단하게 되면 잘못된 평가 결과에 빠질 수 있다.

따라서 학습보다는 검증 과정의 결과가 현실을 더 반영하며, 정확도 보다는 정밀도와 재현율을 혼합한 F1 스코어가 더욱 보수적인 기준이므로 검증과정의 F1 스코어를 중심으로 모델의 성능을 최종적으로 평가하였다.

〈Fig. 9〉는 샘플링된 데이터들을 적용한 XGBoost 모델들의 예측 성능을 최종적으로 비교한 결과이며, 성능 지표는 F1 스코어 0.7 (Kohavi et al., 1997)을 기준으로 하였다. 분석 결과를 요약하면 유지보수가 필요 없는 구간(0)의 예측 성능은 F1 스코어가 예측모델 모두 0.95가 넘어 예측 성능이 우수하였으나, 파손으로 인하여 유지보수가 필요한 구간(1)의 예측 성능은 F1 스코어가 모델 모두 0.7미만으로 기준을 만족하지 못하였다.

오버 샘플링 기법을 적용하게 될 경우 원본 데이터를 활용하는 것보다 성능은 향상되었지만, 언더 샘플링 기법과 혼합 샘플링 기법의 경우 오히려 F1 스코어가 감소하였다. 이는 데이터 불균형을 해결하였지만 원본 데이터의 이상치 데이터 제거 등 세밀한 데이터 전처리 작업과 상관분석을 통한 영향력이 높은 변수들만을 고려한 추가 분석이 필요한 것으로 판단된다.

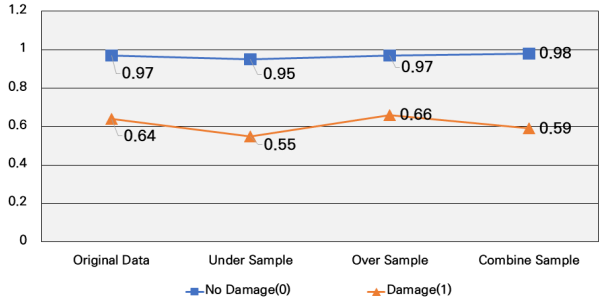

Fig. 9. F1 Score Result

마지막으로 〈Fig. 10〉은 SHAP 방법을 활용하여 변수의 중요도를 산정하였다. SHAP는 SHapley Additive exPlanations의 약자로 입력 데이터에 대해 섀플리 값과 피처 간 독립성을 활용하여 입력 데이터의 중요도를 식 (6)과 같이 산정한다. 분석 결과, 공용년수, 축하중 교통량, 평균 최저기온 -2도 이하 일수 순으로 도로 파손에 영향을 주는 주요 변수로 나타났다.

$$\Phi_i(v) = \sum_{S \subseteq N \setminus i} \frac{|S|!(n-|S|-1)!}{n!}(v(S \cup i) - v(S)) \quad (6)$$

여기서, Φ_i: i 데이터에 대한 섀플리 값
n: 참여자
S: 총 그룹에서 i번째 인무를 제외한 모든 집합
$v(S)$: i번째 피처를 제외하고 나머지 부분 집합이 결과에 공헌한 기여도
$v(S \cup i)$: i번째 피처를 포함한 전체 기여도

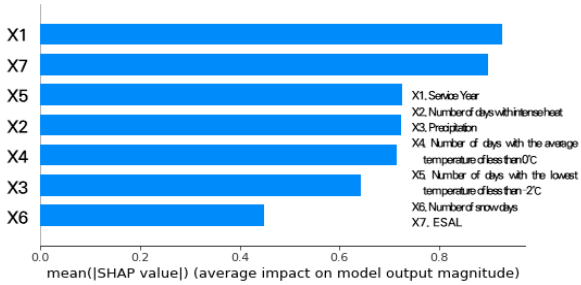

Fig. 10. SHAP Feature Importance

5. 결론

본 연구에서는 고속도로 포장 파손 예측을 위해 머신러닝 기법 중 머신러닝을 활용한 예측 기법을 제안하였다.

먼저 도로파손에 영향을 주는 변수들을 기존문헌 고찰을 통해 선정된 ESAL 및 환경 변수 총 6개의 독립변수 데이터와 고속도로 포장상태지수(HPCI)에 따른 상태 클래스(0, 1)를 종속변수로 하는 데이터 셋을 구축하였다. 구축된 데이터

셋을 기반으로 머신러닝 분석을 통한 고속도로 콘크리트 포장 파손 예측을 위해 의사결정나무 기법인 XGBoost를 활용하여 예측 기법을 제안하고 성능을 평가하였다.

구축된 데이터 셋의 불균형 문제를 해결하기 위해 다수 범주의 데이터 개수를 소수 범주의 데이터 개수로 맞추어 불균형을 해소하는 언더 샘플링과 소수 범주의 데이터 개수를 다수 범주의 데이터 개수로 맞추어 불균형을 해소하는 언더 샘플링, 그리고 두 가지 방법을 혼합한 샘플링 기법을 통해 데이터 불균형 문제를 해결하였다. 데이터 불균형 문제가 해결된 샘플링 데이터들에 XGBoost 기법을 활용하여 예측 모델을 개발하고 F1 스코어를 통해 모델의 성능을 평가하였다.

분석 결과 오버 샘플링 기법이 포장 파손 예측 성능이 가장 우수한 결과를 보였으며 언더샘플링과 혼합샘플링의 경우 원본 데이터를 활용했을 때보다 예측모델의 성능수준이 감소하였다.

마지막으로 SHAP 방법을 활용하여 도로파손에 영향을 주는 주요 변수를 산정하였으며 공용년수, ESAL, 최저 평균 기온 –2도 이하 일수 순으로 산정되었다.

그러나 본 연구에서 개발한 콘크리트 파손 예측 모델은 유지보수가 필요 없는 구간(0)의 예측 성능은 우수하였으나, 파손으로 인하여 유지보수가 필요한 구간(1)의 예측 성능은 성능 수준 기준을 만족하지 못하여 현장에 바로 적용하기에는 다소 어려움이 있어 보완이 필요하다.

따라서 향후 연구를 통해 세밀한 데이터 전처리와 추가 분석을 통해 예측 모델의 성능을 향상시킨다면 보다 정확한 유지보수 필요 구간의 예측이 가능해져 장래 고속도로 포장 유지보수 예산의 추정에 중요한 기초정보로 활용될 수 있을 것이라 기대된다.

감사의 글

이 논문은 2020년 교량관리시스템 운영 위탁 과제의 연구비 지원에 의해 수행되었으며, 논문을 작성할 수 있도록 데이터를 제공해준 한국도로공사에 감사드립니다.

Reference

Batista, G.E.A.P.A., Ronaldo C. Prati and Maria Carolina Monard (2004). “A study of the behavior of several methods for balancing machine learning training data.” SIGKDD Explorations, 6(1), pp. 20-29.

Dennis L., Wilson (1972). “Asymptotic properties of nearest neighbor rules using edited data.” IEEE Transactions on Systems, Man, and Cybernetice, 3, pp. 408-421.

Do, M., Lee, Y., Lim, K., and Kwon, S. (2011). “Estimation of Performance and Pavement Life using National Highway Pavement Condition Index.” *KSCE Journal of Civil Engineering*, 15(2), pp. 261-270.

Gharaibeh, N., and Darter, M. (2003). “Probabilistic analysis of highway pavement life for Illinois.” *Transportation Research Record 1823*, No. 03-4294, pp. 111-120.

Gong, H., Sun, Y., and Huang, B. (2019). “Gradient Boosted Models for Enhancing Fatigue Cracking Prediction in Mechanistic-Empirical Pavement Design Guide.” *Journal of Transportation Engineering, Part B: Pavementsm*, 145(2), 04019014-1-04019014-10.

Han, D., Do, M., and Kim, B. (2017). “Internal Property and Stochastic Deterioration Modeling of Total Pavement Condition Index for Transportation Asset Management.” *International Journal of Highway Engineering*, 19(5), pp. 1-11.

Ju, Huyan (2019). “Development of Machine Learning Based Analytical Tools for Pavement Performance Assessment and Crack Detection.” Doctoral Thesis.

Ko, M. (2020). “Climate Changes and the Future of the Environment.” *Archives of 21st KIPA Public Leadership Seminar*, pp. 1-58.

Korea Expressway Corporation (2010). “Guideline for Exposure to Environment.”

Korea Expressway Corporation (2018). “2017 Investigation and analysis of highway pavement condition.”

Kwon, S., Jeong, K., and Sun, Y. (2012). ”A Study on Decision Criteria of traffic volumes for Choosing of Modified Asphalt Pavement in Korea National Highway.” *International Journal of Highway Engineering*, 4(3), pp. 25-33.

Lee, Y. (2013). “A Study on the Method of Establishing Road Maintenance Strategy Considering the Forecasting Traffic Demand.” Master Thesis.

Lee, Y., and Lee, M. (2016). “A Study on Estimating of Probability Distribution and Mean Life of Bridge Member for Effective Maintenance of the Bridge.” *Korea Journal of Construction Engineering and Management*, KICEM, 17(4), pp. 57-65.

Lee, Y., Sun, J., and Lee, M. (2019). “Development of Deep Learning Based Deterioration Prediction Model for the Maintenance Planning of Highway Pavement.” *Korea Journal of Construction Engineering and Management*, KICEM, 20(6), pp. 34-43.

Lim, S., and Chi, S. (2019). “Xgboost application on bridge management systems for proactive damage estimation.” Advanced Engineering Informatics, 41, pp. 1-14.

Loizos, A., and Karlaftis, M.G. (2005). “Prediction of pavement crack initiation from in-service pavements: A duration model approach.” *Journal of the*

Transportation Research Board, 1940, TRB, pp. 38-42.
Moon, K., You, T., Kim, J., and Park, J. (2017). "An Application Study on Selecting Proper and Otimized Sections for Remodeling of Aged Pavement." Expressway & Transportation Research Institute, Report No. KECRI-2017-32-534.9607.
Kohavi, R., and John, G.H. (1997). "Wrappers for feature subset selection." Artificial Intelligence, 97, Issues1-2, pp. 273-324.
Chen, T., and Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." KDD'16, pp. 785-794.
Yang, J., Gunaratne, M., Lu, J.J., and Dietrich, B. (2005). "Use of recurrent Markov chains for modeling the crack performance of flexible pavements." *Journal of Transportation Engineering*, 131(11), pp. 861-872.
You, P., and Lee, D. (2002). "Methodology of a Probabilistic Pavement Performance Prediction Model Based on the Markov Process." *International Journal of Highway Engineering*, 4(4), pp. 1-12.

요약 : 도로연장의 지속적인 증가와 공용기간이 상당히 경과한 노후 노선이 늘어남에 따라 도로포장에 대한 유지관리비용은 점차 증가하고 있어, 예방적 유지관리를 통해 비용을 최소화 하는 방안에 대한 필요성이 제기되고 있다. 예방적 유지관리를 위해서는 도로포장의 정확한 파손 예측을 통한 전략적 유지관리 계획 수립이 필요하다. 이에 본 연구에서는 고속도로 콘크리트 포장 파손 예측 모델 개발을 위해 머신러닝 분류기반 모델 중 성능이 우수한 XGBoost 기법을 사용하였다. 먼저 데이터 샘플링을 통해 데이터 불균형 문제를 해결하고 샘플링된 데이터들에 XGBoost 기법을 활용하여 예측모델을 개발하고. F1 소코어를 통해 성능을 평가하였다. 분석 결과 오버 샘플링 기법이 가장 좋은 성능 결과를 보였으며, 도로파손에 영향을 주는 주요 변수로 공용년수, ESAL, 최저 평균 최저기온 -2도 이하 일수 순으로 산정되었다. 향후 더 많은 데이터 축적 및 세밀한 데이터 전처리 작업을 통해 예측모델의 성능이 향상된다면 보다 정확한 유지보수 필요 구간의 예측이 가능해질 것으로 판단되므로 장래 고속도로 포장 유지보수 예산의 추정에 중요한 기초정보로 활용될 수 있을 것이라 기대된다.

키워드 : 고속도로 포장, 파손예측, 유지관리, 머신러닝, XGBoost