도시계획학 박사학위논문

앙상블 학습을 이용한 경유자동차의 미세먼지 배출농도 예측 연구

Ensemble Learning to Predict Particulate Matter Concentrations Emitted by Diesel Vehicles

2023년 2월

서울대학교 대학원

환경계획학과

이 상 준

앙상블 학습을 이용한 경유자동차의 미세먼지 배출농도 예측 연구

지도교수 장 수 은

이 논문을 도시계획학박사 학위논문으로 제출함
2022 년 1 월

서울대학교 대학원
환경계획학과 교통학전공
이 상 준

이상준의 박사 학위논문을 인준함
2023 년 2 월

위 원 장 한 상 진 (인)

부위원장 허 성 호 (인)

위 원 정 규 수 (인)

위 원 임 동 욱 (인)

위 원 장 수 은 (인)

국문 초록

경유자동차는 디젤엔진 특성으로 다른 차량에 비해 미세먼지(PM)를 압도적으로 많이 배출한다. 2019년 6월 기준으로 한국의 경유자동차는 총 997만여 대로 전체 차량에서 차지하는 비중이 42.5%에 이른다. 반면 미국과 중국, 일본은 디젤차 비중이 1~3% 수준에 그쳐 한국은 이들 국가에 비해 경유자동차 비중이 높은 편이다. 그러므로 우리나라는 경유자동차 대기오염저감에 관한 연구와 정책이 더욱 중요하다. 이러한 연구를 진행하려면 무엇보다도 경유자동차 PM 배출에 미치는 영향요인을 명확하게 규명하는 기초연구가 필요하다. 이에 본 연구는 경유자동차 PM 배출 예측모형을 제안하고, 이 모형에서 도출된 PM 배출의 주요인을 확인하고자 한다.

그러나 기존의 경유자동차 PM 배출 예측모형은 다음과 같은 한계점을 갖는다. 기존 연구에서는 대부분 전통적인 통계기법을 활용하여 예측성능이 비교적 낮은 편이다. 경유자동차 PM과 배출요인과의 인과관계는 매우 복잡하며, 외생변수 통제가 어려운 PM 측정방식을 채택하고 있다. 이러한 한계를 극복하기 위해 본 연구는 빅데이터이자 변인 통제된 자동차 배출가스 정밀검사 자료를 이용하여 머신러닝기법이 적용된 경유자동차 PM 배출 예측모형을 제시하였다.

본 연구에서는 세 가지 단계를 거치면서 세 가지 연구 목표를 달성하였다. 각 단계별 내용은 다음과 같다. 첫째, '예측모형의 정확도를 높였다'. 이를 위해 머신러닝기법인 앙상블 학습기반 PM 배출 예측모형을 구축하였다. 먼저 1차 앙상블 학습 예측모형은 KD-147모드와 Lug-Down3모드로 구분하고, 배출가스검사 합격과 불합격 데이터를 분류하여 앙상블 학습 기반 20개 모형을 분석하였다. 여기서 통계기법을 대표하는 회귀분석과 의사결정나무, Bagging을 대표하는 랜덤포레스트, 나머

지 3개 모형은 Boosting을 대표하는 CatBoost, LightGBM, XGBoost를 선정하였다. 2차 앙상블 학습에서는 차종별 PM 배출 예측모형을 구축하였다. 예측모형의 성능은 최적의 하이퍼파라미터 튜닝을 통해 예측성능을 향상시켰다. KD-147모드 6개 모형 중 CatBoost R^2가 0.815로 분석되었나 선형회귀모형의 R^2는 0.649로 두 모형 간의 예측성과지표 편차는 높았다. 이 정도 편차는 모든 부스팅모형과 통계모형에서 나타났다.

둘째, '경유자동차 PM 배출의 주요인을 규명하였다' 앙상블 학습 경유자동차 PM 배출 예측모형은 입력변수 간의 영향력을 수치화하여 순열특성 중요도(Permutation Feature Importance: PFI)를 분석하였다. 모형별로 PFI를 비교해보면 모형별로 다소 차이를 보이고 있으나 공통적인 PM 배출요인은 배출가스등급, 연식, 배기량, 총중량으로 도출되었다. 차종별 PM 배출요인의 차이점은 특수차는 적재중량, 승합차는 승차인원이 선정되었다. 이는 차량별 제작 목적과 PM 배출 주요인이 일치하기 때문이다.

셋째 '경유자동차 PM 배출 주요인을 관련 정책에 활용하였다.' 사례분석의 주요 목적은 앙상블 학습 PM 예측모형에서 도출된 PM 배출의 주요인을 미세먼지 절감 및 환경 관련 정책에 적용하기 위함이다. 현재 환경개선부담금 산정방식은 다방면으로 문제점을 안고 있다. 이러한 문제점을 개선하기 위해 본 연구에서는 PM 배출요인과 주요인별 PFI를 환경개선부담금 산정계수의 가중치로 반영하였다. 다음으로 PM 배출 고·중·저농도 차량을 분류하거나 차종 및 지역별에 따라 기존 산정방식과 개선방안의 자동차 1대당 환경개선부담금 변화를 비교해보았다. 지역계수 대신 중량계수와 배출가스등급계수를 산정식에 적용해본 결과에서는 환경개선부담금 부과대상자의 형평성을 한층 고려된 것으로 나타났다. 배출가스등급과 연식의 PFI는 환경개선부담금 산정계수 가중치에 적용시키면 고농도 PM 배출 운전자에게 부담금이 더 전가되는 구조를 확인

하였다. 이는 오염자 부담원칙 강화에 부합됨으로 해석할 수 있다.

본 연구에서는 앙상블 학습 경유자동차 PM 배출 예측모형의 성능이 우수함을 검토하였고, PM 배출 주요인을 규명하였다. 이 예측모형의 활용방안으로는 PM 배출 저감정책효과를 평가하거나 향후 친환경 정책 및 전략 수립에 기초자료로 이용될 것으로 사료된다.

주요어 : 경유자동차, 미세먼지, PM, 배출요인, 배출가스정밀검사 자료, 머신러닝, 앙상블 학습, 순열 특성 중요도

학 번 : 2010-31246

목 차

표 목차

그림 목차

제 I 장 서 론

제1절 연구의 배경 및 목적

미세먼지란 대기 중에 떠다니거나 흩날려 내려오는 아주 미세한 크기의 물질을 말한다. 이수지·김호(2019)에서 미세먼지(PM10)는 10/1000 mm보다 작은 먼지이며, 초미세먼지(PM2.5)는 2.5/1000 mm보다 작은 먼지이다. 일반적으로 초미세먼지는 입자의 크기가 작을수록 독성이 강하다고 알려져 있는데 이는 동일한 질량 농도일 때 초미세먼지의 입자 수가 훨씬 많고 표면적도 넓어서 수용체가 유해물질과 더 많이 흡착하기 때문이다. 다시 말해 미세먼지 배출은 우리의 건강과도 직결되는 문제이므로 미세먼지 저감정책은 매우 중요하다.

교통부문에서 발생하는 미세먼지는 연소성 미세먼지와 비연소성 미세먼지로 양분된다. 내연기관 운행차에서 배출되는 연소성 미세먼지는 도로이동오염원이며, 내연기관 내 연료연소와 무관하게 차량이 도로를 이동할 때 발생 미세먼지와 노면 미세먼지의 재비산에 의한 미세먼지 등이 이에 해당한다. 장수은(2021) 연구에 따르면 유럽 주요국과 우리나라는 이미 비연소성 미세먼지의 배출량이 연소성 미세먼지를 추월했으며, 격차는 더 커지고 있는 추세이다. 그러나 비연소성 미세먼지는 다양한 요인의 복합체이기 때문에 연구의 한계점이 많으며, 관련 자료 또한 미미한 실정이다. 반면 연소성 미세먼지로 인해 문제는 여전히 발생하고 있어 현실적으로 매우 필요한 연구일 뿐만 아니라 관련 자료가 빅데이터 형태로 구축되어 연구결과의 신뢰성을 확보할 수 있다.

국가미세먼지정보센터(2022)에서 매년 집계하는 우리나라 대기오염물질 배출량 통계를 살펴보면 2019년 자동차 운행에 따른 도로이동오염원의 PM2.5, PM10 배출량이 전국기준으로 6,183톤, 6,719톤이며, 서울시 배출량은 321톤, 349톤이다. 서울의 경우 도로이동오염원이 모든 PM 배출원 대비 12%로 적지 않은 비중을 차지고 하고 있다. 전국 PM 배출량

을 차종별로 살펴보면 경유자동차 비중이 높은 레저용 차량(RV), 화물차, 특수차의 PM 배출량이 전체 차종 대비 90%이며, 특히 초미세먼지 PM2.5의 경우 전체 유종 대비 경유자동차 비중이 99%를 차지한다.

그리고 경유자동차는 디젤엔진 특성상 다른 차량에 비해 PM과 NO_X를 압도적으로 많이 배출한다(국립환경과학원, 2015). 미세먼지의 1차 원인은 매연, 검댕 등의 입자이며, NO_X는 공기 중 화학반응을 통해 2차 발생원이 된다(환경부, 2016). 여기서 경유자동차가 배출하는 미세먼지는 기계적 특성과 배출량을 견주어 봤을 때 도로이동오염원의 주요 원인이라 할 수 있다.

이런 현상을 종합해보면 교통부문 도로이동오염원의 PM 발생 주범은 경유자동차에서 배출한 연소성 미세먼지임을 알 수 있다. 앞서 언급한 미세먼지는 인체에 유해한 것으로 알려져 있어 연소성 미세먼지의 다양한 원인을 규명할 필요성이 있다. 이러한 연구는 현실적인 연소성 미세먼지 저감정책 제안이 가능하고, 국민 건강 증진으로 이어져 실용적이면서 매우 중요한 연구이기도 하다.

한편 2019년 6월 기준으로 한국의 경유자동차는 총 997만여 대로 전체 차량에서 차지하는 비중이 42.5%에 이른다. 특히 경유차는 화물차의 93.5%, 승합차의 84.9%를 차지하고 있다(한국의 사회동향, 2019). 반면 미국과 중국, 일본은 디젤차 비중이 1~3% 수준에 그쳐 한국은 이들 국가에 비해 경유자동차 비중이 높은 편이다. 따라서 타 국가보다 우리나라의 경유자동차 대기오염에 관한 연구가 활발히 진행될 필요가 있다.

EU의 경우에는 2020년 기준 경유자동차 비중이 24.8%에 달하지만, 무부하검사(Idling) 방식을 채택하고 있어 실제 도로의 주행패턴을 반영하지 못하는 실정이다(한정헌 2022). 그러나 우리나라는 한국교통안전공단에서 조사하는 배출가스 정밀검사(Inspection/Maintenance: I/M)자료가 있다. 이 자료는 운행 경유자동차의 배출수준을 분석할 수 있는 유일한 빅데이터이다. 이 데이터는 2019년 73만여 대 검사차량의 PM 측정결과치와 다양한 차량제원에 관련된 정보가 있어 충분한 표본수 확보로 통계

적 신뢰도가 높은 것으로 알려져 있다. 특히 I/M자료의 장점은 PM 단위가 농도(%)이므로 질량으로 변환하는 회귀분석모형이 필요 없고, 원자료 고유의 특성 분석이 가능하다는 점이다. 이러한 자료를 활용한다면 예측력이 우수한 모형 구축이 가능하고, 이를 토대로 경유자동차 PM 배출요인을 규명할 수 있다.

이에 본 연구는 전통적인 통계기법을 활용하여 경유자동차 PM 배출요인을 분석한 기존연구의 한계를 극복하기 위해 머신러닝기법 중 앙상블 학습이 적용된 예측모형을 제시하고자 한다. 앙상블 학습은 여러 가지 우수한 학습모형을 조합해 예측력을 제고시키는 기법이다. 무엇보다도 단일모형보다 여러 개의 모형을 구축하는 데 효율적이며, 단일모형에서 놓칠 수 있는 작은 패턴 반영이 가능한 장점이 있다(Rokach, 2010). 또한 앙상블 학습은 머신러닝 분석 대회에서 가장 많은 우승을 차지한 바 있으며, 회귀, 분류, 순위 문제를 해결할 수 있다(Breiman and Cutler, 2014; Siroky, 2009). 더불어 앙상블 학습은 종속변수와 입력변수간의 정량적 영향력을 모형별로 비교 분석이 가능하다. 다시 말해 앙상블 학습 예측모형에서 도출된 다양한 PM 배출요인을 종합적으로 검토할 수 있어 분석결과의 객관성 확보에 도움이 된다. 따라서 본 연구는 앙상블 학습 통해 예측모형 구축과정을 순차적으로 소개하고, 모형의 우수성을 검토하고자 한다.

우수성이 확인된 모형은 경유자동차의 PM 배출요인 분석한다. 여기서 규명된 PM 배출요인은 선행연구의 PM 배출요인과 비교·검토한 후 시사점을 도출하고, 다양한 활용방안을 논의한다. 본 연구에서 제안한 활용방안은 PM 배출 저감 정책효과를 평가하거나 향후 친환경 정책 및 전략 수립에 기초자료로 이용될 수 있다. 특히, 경유자동차 평가관리 방안과 관련 정책인 배출허용기준 강화, 배출가스등급제, 미세먼지 저감장치 부착, 환경개선부담금 등 다양한 분야의 개선방안 도출에 기여할 것으로 판단된다.

제2절 연구의 범위

본 연구는 경유자동차의 1차 생성 미세먼지의 배출요인을 분석대상으로 설정한다. 경유자동차가 휘발유자동차 보다 2차 생성 미세먼지인 NOx(질소산화물)를 특정조건에서 28배나 많이 배출하고 있으나(국립환경과학원, 2015), 경유차동차에 대한 NOx 검사는 2021년 1월부터 시행되고 있다. 본 연구는 2019년 기준 배출가스 정밀검사를 받은 경유자동차를 대상으로 연구를 진행한다. 분석자료의 시간적 범위는 2019년이므로 경유자동차에 의한 2차 생성 미세먼지는 자료의 한계로 분석대상에서 제외하기로 한다.

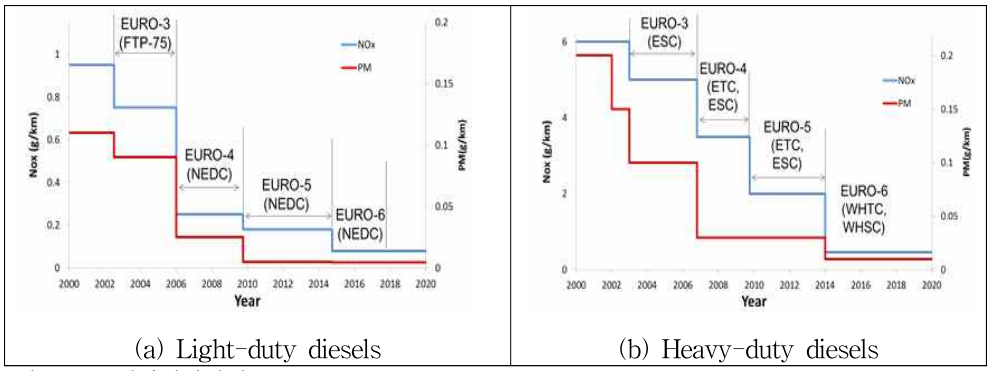
경유자동차의 배출가스 정밀검사방식은 KD-147모드과 Lug-Down3모드로 양분된다. KD-147모드는 차량이 도로를 주행하는 부하검사방식이며, Lug-Down3모드는 엔진회전수를 제어하는 무부하검사방식이다. 두 가지 자동차 검사방식에 따른 본 연구의 분석대상 차종은 <표 1-1>과 같다.

<표 1-1> 경유자동차 미세먼지 검사방식에 따른 적용 차종

검사방식	적용 차종
한국형 경유 147 (KD-147모드)	승용자동차 승합, 화물, 특수자동차(중형 이하)
엔진회전수 제어방식 (Lug-Down3모드)	승합, 화물, 특수자동차(대형) 중형 화물, 특수자동차 중 일반형에서 특수용도로 구조를 변경한 자동차

자료 : 대기환경보전법 시행규칙, 제97조

2019년 기준 배출가스 정밀검사대상인 경유자동차는 배출가스허용기준과 배출저감장치 기술이 적용된다. 배출가스허용기준은 <그림 1-1>과 같이 유럽기준을 준용하고 있다. 2002년 7월에 유로3 기준이 적용되기 시작하면서 2014년부터 유로6 기준이 적용되고 있다.

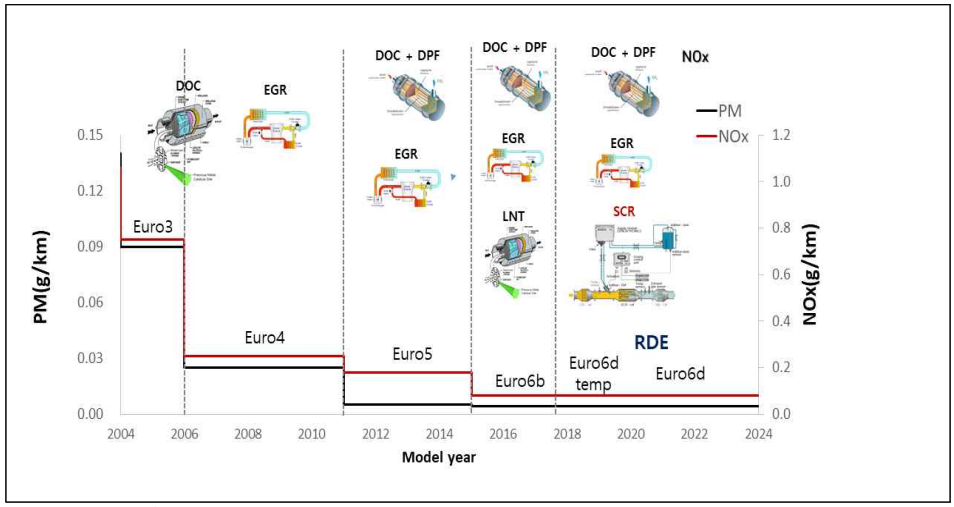


(a) Light-duty diesels (b) Heavy-duty diesels

자료 : 국립환경과학원(2018)

<그림 1-1> 경유자동차 PM, NOx 배출허용기준

자동차 제작사는 배출허용기준을 준수하기 위해 <그림 1-2>와 같이 디젤산화촉매기(DOC, Diesel Oxidation Catalyst), 디젤입자상물질 여과장치(DPF, Diesel Particulate Filter), 배기가스 재순환장치(EGR, Exhaust Gas Recirculation), 선택적 촉매환원장치(SCR, Selective Catalytic Reduction) 등 다양한 저감장치가 경유자동차에 적용되고 있다(한정헌, 2022).

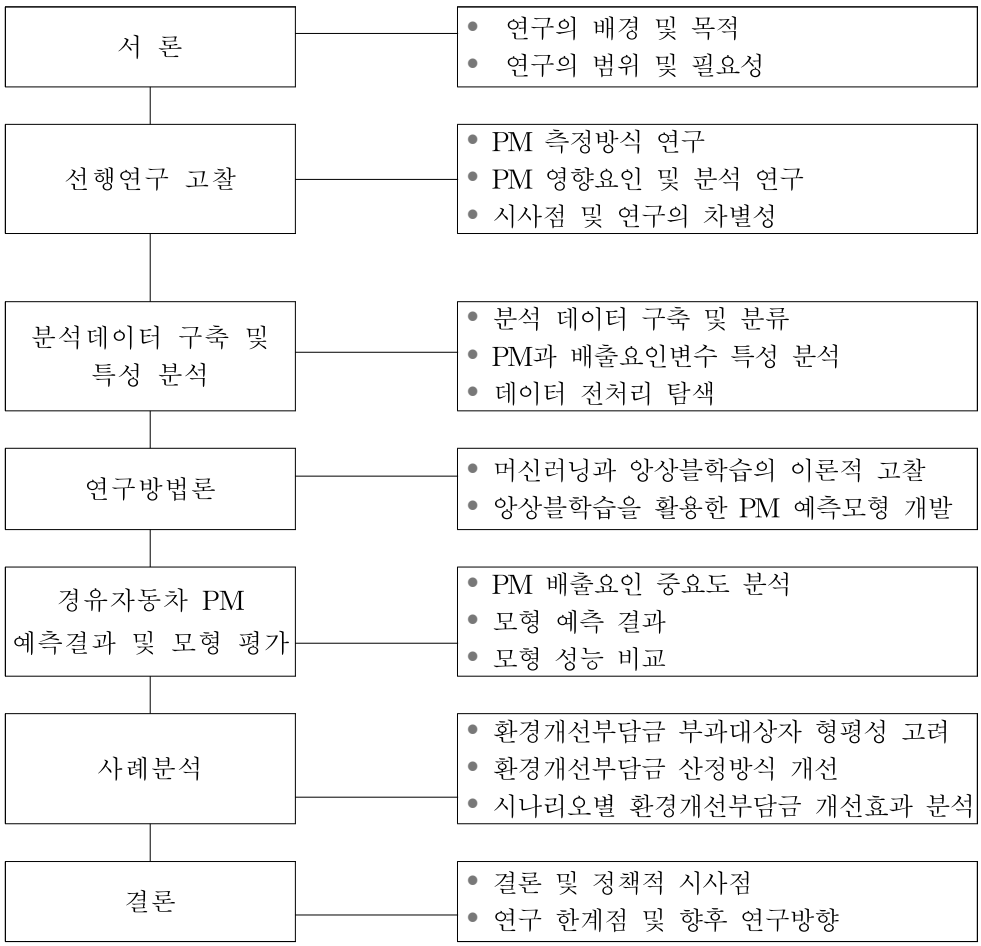


자료 : 국립환경과학원(2018)

<그림 1-2> 경유자동차 PM, NOx 배출저감기술

제3절 연구의 수행체계

본 연구는 총 7장으로 구성한다. 2장에서는 PM 측정방식과 선행연구를 검토하여 본 연구의 차별성으로 제시한다. 3장은 분석데이터 구축 및 특성을 분석한다. 4장에서 앙상블학습을 활용한 PM 예측모형을 구축한다. 5장은 경유자동차 PM 배출요인과 모형의 성능 비교평가와 예측결과를 제시한다. 6장에서는 예측모형의 사례분석을 적용하여 환경개선부담금 개선방안을 제안한다. 본 연구의 수행과정은 <그림 1-3>과 같다.



<그림 1-3> 연구 수행 흐름도

제Ⅱ장 선행연구 고찰

제1절 개요

이 장에서는 자동차 PM의 측정방식과 영향요인 그리고 요인의 분석기법의 분류체계를 구성하고, 이 기준으로 선행연구들에 대해 논의한다. 우선 일반적인 자동차 PM 측정방식은 실험실(Laboratory) 측정(Franco et al., 2013), 고정 실도로(Stationary roadside) 측정(Bukowiecki et al., 2010), 차량 탑재 이동(mobile on-board) 측정(Ježek et al., 2015)으로 분류할 수 있다. 이와 같은 측정방법은 연구 목적과 환경에 따라 선택 및 결합하여 사용하고 있다. 다음으로 PM의 영향요인은 차량 내부요인(Karjalainen et al., 2014)과 차량 외부요인(Suleiman et al., 2019)으로 양분된다. 이 기준으로 선행연구를 검토하고 그 영향요인을 구체적으로 파악한다. 마지막으로 PM의 영향요인을 분석하는 방법론에 따라 기존 문헌을 고찰한다. 전통적인 통계기법과 데이터 과학(Data Science)기법이 광범위하게 활용되고 있다. 통계기법은 기술통계(descriptive statistics) 분석부터 상관관계 분석, 분산분석, 회귀분석, 로지스틱 회귀분석, 로짓모형에 이르기까지 다양한 방법론이 사용되고 있다(Geller et al., 2006; Karjalainen et al., 2014; Gulia et al., 2017; Krecl et al., 2017; 한진석, 2020; 한정헌, 2022). 데이터 과학기법은 데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야다. 데이터 과학은 데이터를 통해 실제 현상을 이해하고 분석하는데 데이터 마이닝, 머신러닝, 딥러닝과 연관된 방법론을 통합하는 개념이다. 데이터 과학기법은 ANN, 유전자알고리즘, SVM, XGBoost 등 다양한 알고리즘과 머신러닝기법 등이 활용되고 있다(Elangasinghe et al.,2014; Lešnika et al., 2019; Suleiman et al.,2019; Xu et al., 2020). 따라서 이 두 가지 분석방법론의 장단점을 확인한 후 본 연구에 적용할 분석기법에 대해 논의하기로 한다.

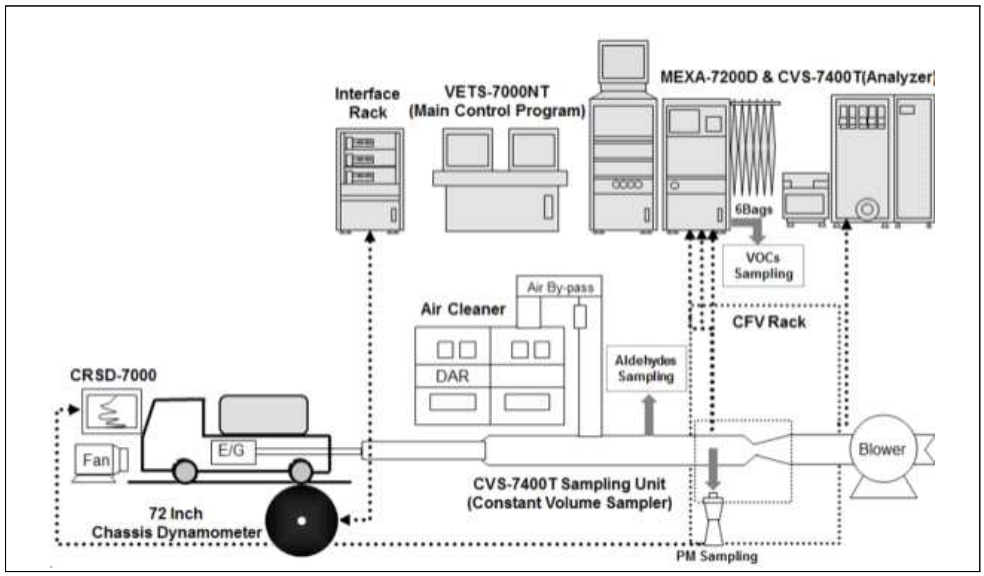
제2절 자동차 배출 미세먼지 농도 측정방식

1. 실험실 측정법

국내의 실험실 측정방식은 국내 자동차 배출계수 산출, 제작차 배출가스 인증시험 등이다. 실험실 측정은 운행 자동차의 실제 도로주행 패턴을 모사한 주행모드를 기반으로 시험자동차를 차대동력계로 주행시킨 후 배출가스 분석기 등으로 배출물질을 측정한다(한국교통안전공단, 2020). 시험주행모드는 부하, 가속, 감속, 정속, 급가감속, 평균속도 등 다양한 차량 주행패턴을 평지, 언덕길, 내리막길과 같은 실제도로에 적용시켜 차량의 오염물질 배출을 측정하는 방식이다. 이 측정방식은 차량중량, 차종 등 변수 선정이 비교적 용이하기 때문에 PM 배출에 영향을 미치는 요인을 분석하는데 매우 효과적이다(Rönkko et al., 2006; Rönkö et al., 2007). 또한 개별 자동차의 오염물질 배출량을 배기관에서 직접 측정함으로써, 데이터의 정확도가 높은 장점이 있다(Franco et al., 2013). 그러나 질량분석이 가능한 PM 측정장치(MDT, Mini-dilution tunnel particulate measurement system), 정용량 시료채취장비(CVS, Constant Volume Sampler) 등은 고가일 뿐만 아니라, 대당 배출가스 측정과 분석에 대략 1~2일의 시간이 소요되는 단점이 있다. 따라서 차대동력계-CVS 장비를 활용한 PM 측정은 제작차 배출가스 인증 등 제한된 실험차량에만 적용할 수 있는 실정이다(한정헌, 2022).

차대동력계는 자동차가 평지를 주행할 때 제동장치를 활용하여 주행저항을 측정한다. 시험자동차의 차대동력계 주행 중에 배출되는 배기가스는 블로워에 의해 빠른 속도로 외부공기와 희석되어 오리피스를 통과한다. 이때 임계속도까지 가속되었을 때 희석 배기가스가 희석 터널 내의 유량이 증가하지 않고 일정한 값을 유지하게 된다. 희석터널에서 샘플링된 입자상물질의 농도를 측정하고, 희석터널의 유량을 이용하여 중량단위의 배출율을 산정할 수 있다. 여기서 자동차 배기가스와 외부공기 간

의 희석비율은 탄소균형(Carbon balance) 방법으로 계산할 수 있다. CVS(ConstantVolume Sampler)라는 측정장치를 이용하면 자동차 주행 중 배기가스 유량이 급격하게 변하지 않도록 일정한 CVS 유량값을 측정할 수 있어 결과의 신뢰성을 높일 수 있는 장치이다(박준홍 외, 2013).

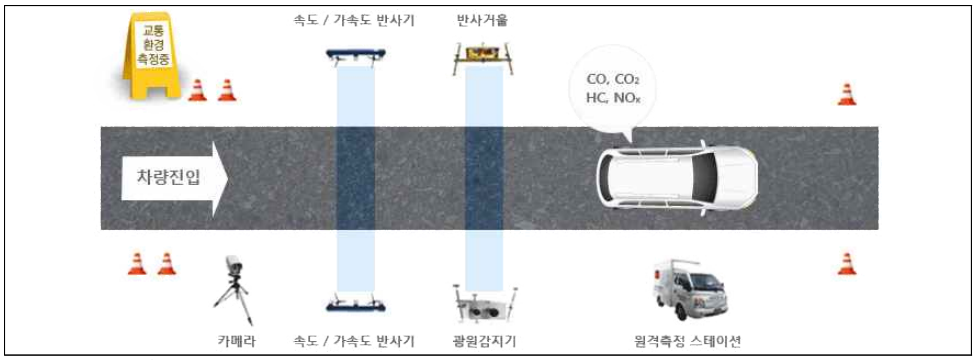


자료 : Jung et al.(2019)

<그림 2-1> 경유차 차대동력계-CVS 활용 PM 측정

2. 고정 실도로 측정법

도로변 고정된 장소에서 임의 차량에 대해 측정하는 <그림 2-2>의 도로측정방법은 도로교통, 기상조건 등을 반영하므로, 엔진연소 배출값 뿐만 아니라 비연소 배출 측정값을 얻을 수 있다. 그러나 PM 배출에 영향을 미칠 수 있는 주변의 다양한 환경에 도출될 수 있는 PM 배출시설에 대한 고려가 필요하다(Bukowiecki et al., 2010). 국내의 경우, 매연측정장비의 신뢰성 문제 등으로 경유자동차에 대해서는 측정하지 않고 있다(한정헌, 2002).



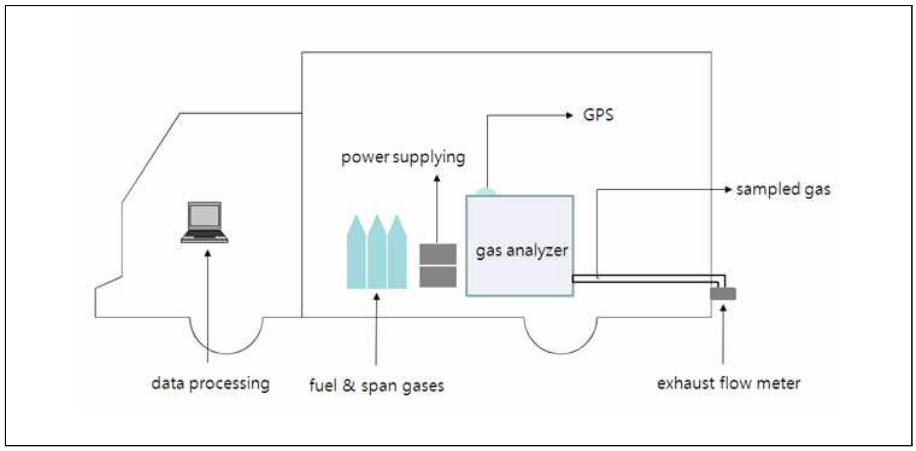
자료 : 한국환경공단(https://www.keco.or.kr/kr/business/climate/contentsid/1534/index.do)

<그림 2-2> 고정 실도로 배출가스 측정방식

3. 이동식 차량 측정법

이동식 배출가스 측정은 배출측정을 위한 장비(PEMS, Portable Emission Measurement Systems)를 <그림 2-3>과 같이 시험 차량에 탑재하여, 실제 도로를 주행하면서 배출가스를 측정한다. 차량 내부에 센서를 탑재하고, 센서를 배기관 등에 연결하여 분석하는데(Ježek et al., 2015; Pirjola et al., 2004), 국내 제작차 실외도로주행 시험방법 등에 활용되고 있다. 다만, 이 방법은 실제 도로에서 측정하지만 현재 제한된 시험차량을 대상으로 적용 가능한 실정이다.

도로상의 차량추적 측정방법(Wang et al., 2012; Park et al., 2011, Kittelson et al., 2004)은 특정차량의 배출계수(EFs, Emission Factors) 측정을 목표로 관찰 차량을 지정된 거리만큼 따라가며, PM 배출을 측정한다(Pirjola et al., 2004; Karjalainen et al., 2014). 이 방식은 관찰 차량에 대한 실도로 측정이 장점이나, 관찰 차량에 대한 제한된 측정이기 때문에 전체 교통량과의 인과관계 분석에 어려움이 있다. 또한 주변 차량의 PM이 관찰 차량의 PM 측정에 영향을 미칠 수 있어 데이터 신뢰성이 떨어지는 단점이 있다(Ježek et al., 2015).



자료 : 국립환경과학원(2012)

<그림 2-3> PEMS 기반 제작차 실도로 배출가스 측정방식

제3절 자동차 미세먼지 배출 요인분석 연구

자동차 PM 배출 요인분석 연구는 다양하게 이루어지고 있다. 관련 연구는 PM 배출요인과 분석방법론으로 분류할 수 있다. 요인은 차량의 내부요인과 외부요인으로 분석방법론은 통계기법과 데이터 과학기법으로 양분할 수 있다.

첫째, PM 배출요인은 차량 내부요인과 차량 외부요인으로 양분된다. 차량 내부요인은 차량업종, 등록지, 연식, 총주행거리, 배출가스 저감장치 유무(Geller et al., 2006; Bikas and Zervas., 2007; Karjalainen et al., 2014; 한정헌, 2022). 특히 미세먼지를 저감하는 DPF와 질소산화물을 저감하는 SCR를 부착한 차량이 오염물질별 배출허용기준을 얼마나 만족하는가를 분석한 사례가 대부분이다(Geller et al., 2006; Bergmann et al., 2009; Dallmann et al., 2011; Karjalainen et al., 2014). Guan et al.(2014)는 도로교통부문 NO_x 저감을 위한 SCR 장치의 기술 동향을 전반적으로 검토하였으며, Quiros et al.(2016)는 대형화물자동차의 연료에(경유(SCR 부착), 천연가스, 하이브리드) 따라 실도로의 NO_x 배출량을 비교 분석하였다. 차량 외부요인은 교통량, 교통혼잡, 풍속, 기온, 운전자 운전 습관 등이 있다(Elangasinghe et al., 2014; Krecl et al., 2017). 이 같은 요인들은 변인 통제가 불가능하고 요인들이 복잡하기 때문에 인과관계 규명에 어려움이 따른다. 따라서 변인 통제가 가능한 차량 내부요인 의한 PM 배출 요인분석이 상대적으로 과학적인 연구라 할 수 있다.

둘째, PM 배출 요인분석에 적용되는 분석방법론에 따라 구분된다. 전통적인 통계기법인 기술통계, 회귀모형, 로짓모형 등을 이용한 다수의 선행연구가 수행되었다(Bikas and Zervas, 2007; 서울연구원, 2015; Krecl et al., 2017; 한정헌, 2022). 반면 데이터 과학기법을 활용한 미세먼지 영향요인 분석 관련 연구는 빅데이터 분석에 적합하며, 모형 강건성(Robustness)이 우수한 것으로 알려져 있어 최근 들어 활발한 연구가 진행 중이다(Lesnik et al., 2019; Suleiman et al., 2019; Xu et al., 2020).

<표 2-1> 통계기법 활용한 미세먼지 영향요인 분석 관련 연구

연구	배출물질	측정 방식	방법론	영향요인	연구대상
Geller et al.(2006)	PM, PN	실험실	통계기법 (회귀모형)	저감장치	EURO3 디젤, 가솔린, EURO4 디젤
Bikas and Zervas(2007)	PM, CO, HC, NO_x	실험실	기술통계	주행거리, 규제기준	경유차(승용) EURO3, 1.9L 디젤엔진
Bergmann et al.(2009)	PM, CO_2	실험실	기술통계	저감장치	EURO4 디젤
Karjalainen et al.(2014)	PM	실도로 (고정식 이동식)	기술통계	저감장치	가솔린 실험차량
서울연구원 (2015)	PM_{10}, $PM_{2.5}$, NO_x,	실험실	기술통계	연식, 차종, 배출, 저감장치	경유차
Kumar and Goel(2016)	PM_1, $PM_{2.5}$, PM_{10}	실도로 (고정식 이동식)	통계기법 (미세먼지 산정식)	지체상황	실험차량 (고정식, 이동식)
Gulia et al.(2017)	NO_X, $PM_{2.5}$	실도로 (고정식)	통계기법 (통계모형 + 결정적모형)	계절, 교통량	도로 주행차량
Krecl et al,(2017)	NO_x, $PM_{2.5}$, PN, 매연	실도로 (고정식)	통계기법 (상관분석)	교통량, 속도	도로 주행차량
Krecl et al,(2018)	NO_x, $PM_{2.5}$, PN, 매연	실도로 (고정식)	통계기법 (회귀모형)	기후, 차량상태, 운전자 운전습관 및 주행패턴, 교통량, 차중량	LDV, HDV
한진석 (2020)	$PM_{2.5}$, NO_x,	실험실	통계기법 (이항 로짓모형)	차량업종, 등록지, 연식, 총주행거리, 배출가스 저감장치 유무	경유차
한정헌 (2022)	PM	실험실	통계기법 (회귀모형)	차량연식, 차량중량, 영업용, 외산, 등록지역	경유차

<표 2-2> 데이터 과학기법을 활용한 미세먼지 영향요인 분석 관련 연구

연구	배출물질	측정 방식	방법론	영향요인	연구대상
Elangasinghe et al.(2014)	$PM_{2.5}$, PM_{10}	실도로 (고정식)	ANN[1)]	풍속, 풍향, 교통량	도로 주행차량
Lešnika et al. (2019)	PM_{10}	실도로 (이동식)	유전자알고리즘, 다중회귀모형	기온, 교통량, 풍속, 날씨	도로 주행차량
Suleiman et al. (2019)	$PM_{2.5}$, PM_{10}	실도로 (고정식)	ANN, BRT[2)], SVM[3)]	교통량,	LGV, HGV
Xu et al.(2020)	CO_2, $PM_{2.5}$	실도로 (이동식) GPS, 구글맵	XGBoost[4)]	교통혼잡, 운전자연령, 차량연식, 배출가스 규제기준 교통량 차량조건, 도로조건	69개 표본차량, GPS자료

주: 1) Artifical Neural Network(ANN)
2) Boosted Regression Trees(BRT)
3) Support Vector Machines(SVM)
4) Etreme Gradient Boosting(XGBoost)

제4절 시사점 및 본 연구의 차별성

1. 선행연구의 시사점

선행연구를 검토한 바 자동차 PM 측정방식 및 자료 샘플링 방법과 그 요인을 분석하고, 분석모형을 제시하는 연구로 세 가지 연구유형으로 분류할 수 있다. 이 세 가지 연구유형 관점에서 본 연구의 필요성은 다음과 같다.

첫째, 일반적인 PM 측정방법으로는 실험실 측정(Franco et al., 2013), 고정 실도로 측정(Bukowiecki et al., 2010), 이동식 차량 측정(Ježek et al., 2015) 등이 있으며, 연구 목적에 따라 측정방법을 결합해서 사용하고 있다. 실험실 측정의 대표적인 예는 국내 자동차 배출계수 산출, 제작사 배출가스 인증시험 등이다. 실험실 측정은 운행 자동차의 실제 도로주행 패턴을 모사한 주행모드를 기반으로, 시험자동차를 차대동력계로 주행시킨 후 배출가스 분석기 등으로 배출물질을 측정한다.

PM 배출자료는 연구 목적에 따라 고정 실도로 측정, 이동식 차량 측정 등 다양한 방법을 통해 표본을 추출하고 있으며, 대규모의 데이터를 활용한 연구는 배출가스 정밀검사 자료를 이용하고 있다. 대다수의 사례 연구는 1~2만 건 내외의 PM 측정 자료를 사용하고 있으며, 10만 건 이상 대규모 표본을 사용한 연구는 Beydoun and Guldmann (2006), 국립환경과학원(2015), 한진석(2020), 한정헌(2022) 등 소수에 불과한 실정이다. 또한, 운행중인 휘발유자동차를 대상으로 한 연구는 다수 수행되었으나, 경유자동차를 대상으로 한 연구는 소수이다. 무엇보다도 I/M (Inspection /Maintenance)의 빅데이터를 활용한 경유차 PM을 예측한 연구는 극소수만 존재한다.

둘째, PM 배출요인은 다양하지만 크게 차량 내부요인, 차량 외부요인이 있다. 차량 내부요인은 연구자가 관심 있는 요인들의 통제가 가능하고 인과관계 규명이 명확한 장점이 있는 반면 차량 외부요인은 요인 통

제가 어려워 인과관계 불확실성이 높은 편이다. 그러므로 경유자동차 PM 배출요인은 차량 내부요인으로 한정하여 요인별 영향력을 분석한다면 선행연구보다 명확한 요인규명이 연구가 가능할 것으로 판단된다.

셋째, 데이터 과학기법 중 머신러닝을 활용한 분석방법론이다. 근래 들어 머신러닝을 이용한 연구가 활발히 이루어지고 있다(Suleiman et al., 2019, Xu et al., 2020). 그러나 PM 배출요인을 규명하거나 예측에 머신러닝을 활용한 연구는 저조한 편이며, 연구 초기 단계에 해당된다. 따라서 이 같은 연구 동향을 지속적으로 발전시키려면 머신러닝 또는 딥러닝 기반의 예측력이 높은 모형을 활용하거나 개발하는 연구가 필요하다.

2. 본 연구의 차별성

지금까지 선행연구를 고찰한 결과 본 연구의 차별성은 다음과 같다. 본 연구의 목적은 경유자동차 PM 배출 농도의 예측력을 높일 수 있는 모형을 구축하는 것이다. 즉, 모형의 강건성(Robustness)을 확보하기 위해 필수조건은 다음과 같다. 첫째, 입력변수와 예측치의 정확한 인과관계 규명이 필요하다. 둘째, 변수의 영향력을 가늠하기 위해 변인통제가 가능해야 한다. 셋째, 통계적 신뢰성을 확보하기 위해 표본수 확보가 필요하다.

따라서 이 같은 조건이 충족 가능한 PM 측정방식과 자료, PM에 영향을 미치는 입력변수 그리고 적합한 분석모형을 선택해야 한다. 본 연구에서 이용할 자료는 대규모 운행중인 경유자동차 제원정보와 연계된 PM 측정결과치가 약 68만 건에 달하기 때문에 배출요인 분석결과의 통계적 신뢰도가 높다. 이 자료는 실험실 자료이므로 변인 통제도 가능하고 차량과 관련된 입력변수를 이용하여 인과관계 규명에 적합하다. 또한 분석자료가 빅데이터이므로 기존의 전통적인 통계기법보다 머신러닝기법인 앙상블 모형 활용이 적절할 것으로 판단된다. <표 2-3>과 같이 본 연구는 앞서 전술한바 이 같은 조건을 모두 충족하는 경유자동차 PM

배출 농도 예측모형을 개발하고자 한다.

선행연구의 차별성을 정리한 결과 본 연구는 실험실 방식으로 측정한 PM자료를 활용한 앙상블 학습을 활용한 경유자동차 미세먼지 예측모형 연구로서 유일한 연구로 확인되었다. 학문적 기여도 측면에서는 경유자동차 PM과 입력변수간의 인과관계를 밝히는 데 일조할 것으로 판단된다.

<표 2-3> 연소성 미세먼지 관련 연구 비교 분석

구분	측정방식			영향요인		분석방법론	
	실험실	실도로 고정식	실도로 이동식	차량 내	차량 외	통계 기법	머신러닝 기법
표본수 확보	○	×	×	○	△	×	○
변인 통제	○	×	×	○	×	△	△
인과관계 규명	○	△	△	○	△	△	○
본 연구	√			√			√

주: ○=적합, △=일부 적합, ×=미흡

제Ⅲ장 데이터 구축 및 특성 분석

제1절 데이터 구축

1. 데이터 범위

본 연구는 운행 경유차의 PM 배출요인 분석을 위해 대규모 정밀검사 자료를 수집하고, 분석하는데 중점을 둔다. 분석 데이터는 2019년 자동차 검사 중 종합검사를 받은 차량 1,179만 대 중 한국교통안전공단에서 수행한 정밀검사자료 약 73만여 건을 수집하였다. 검사방식은 부하검사방식인 KD-147모드, ASM-Idling모드, Lug-Down3모드가 있으며, 무부하 방식(급가속)이 있다. 여기서 본 연구 목적에 부합되는 경유자동차 검사 데이터만 사용한다. 경유자동차 분석데이터는 일종의 실험실 측정 부하검사인 KD-147모드와 실도로 주행패턴은 아니지만 현재 중·대형 화물, 승합, 특수차를 대상으로 시행중인 Lug-Down3모드 데이터도 분석범위에 포함한다. 따라서 최종 분석데이터는 KD-147모드가 181,344대이며, Lug-Down3모드는 386,124대로 총 567,468대이다.

<표 3-1> 경유자동차 정밀검사모드별 차종 분류기준

구분	KD-147모드	Lug-Down3모드
검사대상	승용차(소형, 중형, 대형) 승합차(소형) 화물차(소형) 특수차(소형)	승합차(중형, 대형) 화물차(중형, 대형) 특수차(중형, 대형)
엔진특성	승용형 엔진, 화물형 소형 엔진	화물형 중대형 엔진 승합은 통상 화물형 엔진 사용

<표 3-2> 경유자동차 정밀검사모드별 차종별 데이터 수

구분		승용	승합	화물	특수	합계
KD-147	대수	73,790	17,203	90,351	-	181,344
	비중(%)	40.7	9.5	49.8	-	100
Lug-Down3	대수	-	41,996	293,196	50,932	386,124
	비중(%)	-	10.9	75.9	13.2	100

2. 데이터 분류

데이터의 분류체계는 KD-147와 Lug-Down3모드로 양분하고 검사방식에 부합되는 분류체계로 데이터를 구분한다. 자동차관리법에서는 자동차를 4개 차종(승용, 승합, 화물, 특수)으로 구분하는 반면, 대기환경보전법은 3개 차종(경차, 승용, 화물차)으로 구분하고 있어, 두 개 법령을 모두 만족하는 차종 구분은 현실적으로 불가능하다. 따라서 경유자동차 정밀검사모드인 KD-147모드와 Lug-Down3모드에 따라 <표 3-1>과 같이 차종과 크기를 구분한다. 그리고 KD-147모드와 Lug-Down3모드의 검사 데이터를 구분하는 또 하나의 이유는 검사방식이 다르기 때문이다. KD-147모드는 실험실에서 실도로를 차량이 직접 주행하여 PM배출농도(%)를 측정하는 방식이고, Lug-Down3모드는 부하방식으로 엔진출력과 PM배출농도(%)를 동시에 검사하는 방식이다. 이 두 가지 검사모드 방식과 그 특성을 정리하면 다음과 같다.

KD-147모드는 예열모드와 주행모드로 구분된다. 예열모드는 자동차를 예열시키는 과정으로 측정대상 자동차의 상태가 정상으로 확인되면 자동차를 차대동력계 위에 엔진정격 출력의 40% 부하로 50±6.4km/h의 차량속도로 40초 동안 주행한다. 주행모드는 <그림 3-1>과 같이 2.16km의 거리를 평균속도 53.0km/h로 40초 가량 주행한 후 147초 동안 정지상태에서부터 83.5km/h까지 최고속도로 도달하는 과정에서 급가속, 가속, 정속, 감속, 급감속을 반복하며, 미세먼지 농도(%)를 측정한다. 미세먼지 농도는 부분유량 채취방식의 광투과식 분석방법을 채택한 측정기를 사용

한다. 운전허용오차는 상한 및 하한 속도는 규정된 시간 1초 이내의 속도곡선 상에서 가장 높거나 낮은 속도보다 3.2km/h 이내의 속도로 한다. 기어 변속할 때와 감속구간(128~147초)과 같이 허용오차보다 더 큰 속도변화는 2초 이내에 일어나면 허용한다. 허용횟수는 0초에서부터 147초까지 전체 구간에서 3회 이하로 한다(한국교통안전공단, 2022).

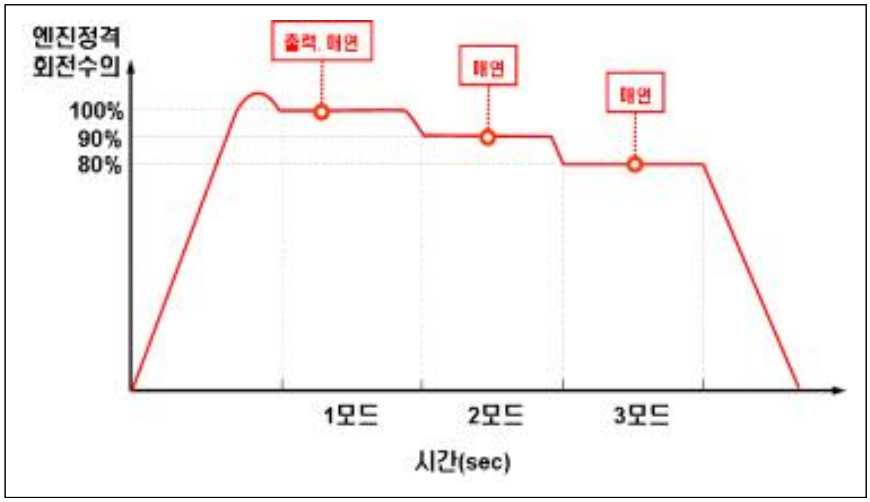
자료 : 환경부(2012)

<그림 3-1> KD-147모드 차량 주행 그래프

Lug-Down3모드는 승합·화물·특수 중 대형차와 화물·특수차 중 특수용으로 구조변경한 중형차, KD-147모드 검사가 불가능한 차량의 경우를 대상으로 엔진정격출력, 엔진정격회전수, 매연농도 등을 측정한다. 엔진출력과 PM배출농도(%)를 동시에 검사하여 부적합 원인을 파악할 수 있는 실용적인 검사방법이라 할 수 있다.

<그림 3-2>와 같이 차대동력계상 주행상태로 엔진정격회전수의 100%에서 1모드, 엔진정격회전수의 90%에서 2모드, 엔진정격회전수의 80%에서 3모드로 구성된다. 1모드에서는 엔진정격출력, 엔진정격회전수 및 매연농도를 측정하고, 2모드와 3모드에서 각각 엔진회전수와 매연농도를

측정한다. 매연농도 측정은 부분유량채취방식의 광투과식 분석측정기를 사용하며, 각각의 모드에서 측정한 값 전부가 배출허용기준 이내일 경우에만 적합으로, 어느 한 모드에서라도 기준이 초과되면 부적합으로 판정된다(한국교통안전공단, 2022).



자료 : 한국교통안전공단(2009)

<그림 3-2> Lug-Down3모드 부하검사 그래프

3. 데이터 항목

본 연구에서는 KD-147모드, Lug-Down3모드에 따라 수집된 배출가스 정밀검사자료의 세부 항목과 표본의 분포 특성 등을 살펴본다. 차량연식, 주행거리, 배기량, 차량중량, 용도, 등록지역 등 차량특성에 따른 영향요인을 분석한다.

먼저 데이터 항목은 종속변수인 PM과 설명변수인 차량특성 데이터로 구분된다. 다양한 차량특성 데이터로 구성된 설명변수는 수치형 변수로 차량연식, 차량중량, 연간 주행거리 등을 사용하고, 범주형 변수로는 유로기준, 크기, 차종, 용도, 등록지역 등을 적용한다.

<표 3-3> 정밀검사 데이터 주요 수집항목

종류	세부항목
차량특성	제작일자, 차량연식, 최초등록일, 배기량, 연료(휘발유, 경유, LPG), 차량중량, 총중량, 적재량, 주행거리, 차종(승용·,화물,승합,특수), 크기(대형,중형,소형,경형), 용도(자가,영업,관용,개인택시), 차량길이, 차량너비, 차량높이, 승차정원, 동력전달장치, 연비, 저감장치 부착
검사항목 및 검사결과	PM, CO, HC 기준치(1~6) / PM, CO, HC 측정치(1~6) / 매연, CO, HC 측정결과(1~6), 판정내용(10종), 배출가스검사 합격·불합격 결과치
기타	미세먼지(PM), 검사모드, 검사연도, 검사구분, 배출가스 등급(1~5), 구조변경내용, 구조변경일자, 정밀검사 시작일·종료일, 등록지역(시·도, 시·군·구)

<표 3-4> 분석데이터 주요변수

종류	주요변수
수치형	미세먼지(PM), 차량연식, 배기량, 차량중량, 총중량, 적재량, 주행거리, 차량길이, 차량너비, 차량높이, 승차정원, 연비
범주형	저감장치 부착, 배출가스 등급(1~5), 연료(휘발유, 경유), 등록지역(시·도, 시·군·구), 차종(승용·,화물,승합,특수), 크기(대형,중형,소형,경형), 용도(자가, 영업용, 관용, 개인택시), 유로(유로3, 유로4, 유로5, 유로6)

추가적으로 설명변수 식별, 병합, 분리 등을 통해 기존 설명변수를 추가변수로 생성한다. 첫째, PM 배출에 미치는 영향요인을 분석하기 위해 '배출규제 적용시점', '자동차 종류', '정밀검사모드'에 따라 모형을 분류할 필요가 있다. '자동차 종류 기준'과 '정밀검사 모드'에 따른 분류는 비교적 단순하다. 그러나 유로기준은 조금 복잡하다. 우리나라는 유럽기준을 준용해 운영하고 있다. 유로기준에 따른 '배출규제 적용시점' 분류는 차종, 크기에 따라 약간의 시차가 존재하여 분류체계가 조금 복잡하지만

<표 3-5>와 같이 정리할 수 있다. 차종별로 규제 적용시점과 생산시점이 다소 상이한 것을 볼 수 있다. '인증일'은 최초 배출규제 적용일자이며, '생산일'은 전체 차량에 대한 적용일자이다. 또한 '인증일'과 '생산일'간의 차이는 일종의 새로운 유로기준 규제 적용에 대한 유예기간으로, 새로 도입되는 유로기준 이전차량과 병행되어 생산된 기간으로 설명할 수 있다. 음영은 배출가스 저감장치(DPF)가 전면적으로 도입·적용된 시점으로, 차종에 따라 적은 수준이긴 하나 적용시점에 차이가 있다.

<표 3-5> 차종(크기)별 배출규제 현황

차종	'2002년 7월1일 이후' (Euro3)		'2006년 1월1일 이후' (Euro4)	
	인증기준	생산기준	인증기준	생산기준
· 승용 소형·중형·대형	2002.7.1	2004.1.1	2006.1.1	2006.1.1
· 화물 소형 · 승합 소형 · 특수 소형	2002.7.1	2004.1.1	2007.11.	2008.1.1
· 화물 중형·대형 · 승합 중형·대형 · 특수 중형·대형	2003.1.1	2004.9.1	2006.1.1	2008.1.1
차종	'2009년 9월1일 이후' (Euro5)		'2014년 1월1일 이후' (Euro6)	
	인증기준	생산기준	인증기준	생산기준
· 승용 소형·중형·대형	2009.9.1	2011.1.1	2014.9.1	2015.9.1
· 화물 소형 · 승합 소형 · 특수 소형	2010.9.1	2012.1.1	2015.9.1	2016.9.1
· 화물 중형·대형 · 승합 중형·대형 · 특수 중형·대형	2009.9.1	2010.10.1	2014.1.1	2015.1.1

주) : 상기 표 ' '는 '대기환경보전법 시행규칙' 〈별표17〉의 규제이며, ( )는 대응되는 유로기준 규제이다. 국내는 2006.1.1. 이후부터 EU와 동일한 유로기준을 적용하고 있다.

제2절 데이터 특성 분석

1. 기초통계 분석

이 절에서는 KD-147모드, Lug-Down3모드에 따라 수집된 배출가스 정밀검사 자료의 특성 분석 및 기술통계 자료를 제시한다. 차량검사 시 함께 수집되는 차량 제원과 미세먼지 배출 관련 변수들의 자료 특성과 분포를 먼저 살펴본 후, 입력변수인 차량제원과 종속변수인 PM의 특성을 파악하고 서로의 영향관계를 확인한다.

1.1 KD-147모드

KD-147모드로 검사된 주요 데이터들의 기초통계 분석결과는 <표 3-6>과 같다. 종속변수인 평균 PM농도(%)는 8.2에 비해 분산은 13.5로 큰 편이기 때문에 표본 간의 이질성이 큰 것으로 판단된다. 평균 차량 연식은 약 10년, 평균 주행거리는 15만km 이상인 것으로 나타났다.

다음으로는 차량의 규모(대형, 중형, 소형) 및 차종(승용, 승합, 화물, 특수)별 데이터 기초통계 결과를 살펴본다. 차량의 크기와 차종은 배기량, 적재량 등과 영향이 있는 변수이며, 이는 PM 배출과도 연관이 있을 개연성이 높다. 분석 결과를 살펴보면 대형 차종일수록 평균 연식이 오래되고, 주행거리가 많으며, PM 배출농도가 높은 것으로 나타났다. 연비는 중형이 가장 우수한 것으로 분석되었다.

차종별 분석 결과에서는 화물 및 특수차량에 비해 승용 및 승합차량의 연식이 평균적으로 더 긴 것을 알 수 있다. 평균 PM 농도는 승합차에서 가장 높으며, 평균 주행거리 또한 가장 긴 것으로 분석되었다. 연비는 승용차가 가장 우수한 것으로 나타났다. 이 같은 데이터 특성은 PM 배출 예측모형의 변수 선정에 직관적으로 도움을 줄 뿐만 아니라 모형의 정확도 향상에 많은 영향을 미칠 것으로 사료된다.

<표 3-6> 데이터 기초통계 분석(KD-147모드)

구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	9.7	2,742	492	2,389	157,854	3.8
std	4.1	903	637	495	94,972	0.8
min	2	1,325	0	1,396	1	3
median	9	2,855	0	2,497	141,402	4
max	35	39,105	25,000	16,991	2,244,437	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	5.5	4.9	1.8	1.9	12.1	8.2
std	2.8	0.5	0.1	0.3	3.1	13.5
min	1	0.5	1.5	0.6	2.7	0
median	5	5.1	1.8	1.9	11.4	3
max	49	13	2.5	23.3	185	100

<표 3-7> 차량 규모별 기초통계 분석(KD-147모드)

대형						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배기 가스등급
mean	11.7	2,479	22	2,485	171,150	4.2
std	4.3	947	615	460	89,673	0.8
min	2	1,600	0	2,143	87	3
median	12	2,435	0	2,476	160,378	4
max	27	39,105	25,000	16,991	2,136,740	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	7.2	4.8	1.9	1.7	11.7	11.2
std	1.9	0.3	0.1	0.1	1.8	16.7
min	2	3.5	1.5	1.4	2.7	0
median	7	4.7	1.9	1.8	11.2	5
max	39	12.9	2.5	4	37	100
중형						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배기 가스등급
mean	10.1	2,592	246	2,270	156,442	3.7
std	4	1,186	709	631	97,805	0.8
min	2	1,325	0	1,396	1	3
median	10	2,255	0	1,995	138,057	3
max	35	12,425	4,600	7,684	2,244,437	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	6.6	4.8	1.9	1.8	13.4	7.3
std	3.2	0.6	0.1	0.3	2.6	13.2
min	1	2.7	1.5	1.3	4	0
median	5	4.7	1.9	1.7	13.3	1
max	49	13	2.5	4	112	100

<표 계속>

소형						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배기 가스등급
mean	8.8	2,959	857	2,499	156,654	3.7
std	4.1	262	246	236	92,665	0.7
min	2	1,390	0	1,422	17	3
median	8	2,955	1,000	2,497	141,262	4
max	30	3,495	1,000	2,957	1,447,333	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	3.9	5.1	1.8	2	10.9	8.5
std	1.4	0.2	0.1	0.3	3.3	13
min	1	0.5	1.5	0.6	5.7	0
median	3	5.1	1.7	2	10.9	4
max	35	9	2.5	23.3	185	100

<표 3-8> 차종별 기초통계 분석(KD-147모드)

승용						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	10	2,113	0	2,070	145,430	3.6
std	3.7	278	0	283	75,086	0.8
min	2	1,325	0	1,396	1	3
median	10	2,125	0	1,995	133,582	3
max	30	2,995	0	2,993	2,244,437	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	6	4.6	1.9	1.7	13.8	6.6
std	1.3	0.2	0.1	0.1	2.3	13.3
min	2	3.5	1.5	1.3	6.4	0
median	5	4.6	1.9	1.7	13.8	1
max	34	8.7	2.4	3.2	24.2	100
화물						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	9.4	3,167	989	2,597	166,422	3.8
std	4.4	903	554	474	106,881	0.8
min	2	1,970	200	1,998	17	3
median	8	2,975	1,000	2,497	147,126	4
max	35	39,105	25,000	16,991	1,447,333	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	3.9	5.1	1.8	2	11	9
std	1.5	0.4	0.1	0.3	3.2	13.3
min	2	0.5	1.5	0.6	3.2	0
median	3	5.1	1.7	2	10.9	4
max	35	12.9	2.5	23.3	185	100

<표 계속>

승합						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	10.5	3,129	0	2,636	167,008	4
std	4.4	934	0	531	94,650	0.7
min	2	2,080	0	1,991	1,284	3
median	10	2,950	0	2,497	154,849	4
max	26	13,295	0	11,149	1,201,201	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	11.2	5.2	1.9	1.9	11.2	10.7
std	4.1	0.6	0.1	0.3	2.1	15.1
min	1	3.5	1.5	1.3	4	0
median	11	5.1	1.9	1.9	10.5	6
max	49	13	2.5	4	98	100
특수						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	7.4	3,814	295	2,781	142,809	3
std	3.3	2,120	1,272	893	144,789	1
min	2	1,390	0	1,969	319	3
median	7	3,450	0	2,497	108,944	3
max	28	26,050	17,500	12,780	2,136,740	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	3.2	3.1	5.4	1.8	2.5	11.2
std	0.6	1.3	0.6	0.1	0.4	1.9
min	3	1	3.6	1.6	1.4	2.7
median	3	3	5.2	1.8	2.6	11.2
max	5	12	9.7	2.5	3.9	22

유로 배출규제 적용기준에 따른 기초통계 분석 결과는 <표 3-9>와 같다. EURO5 이상 차량의 평균 PM 농도가 크게 감소하는 것으로 나타났다. 특히 EURO3과 EURO6의 평균 PM농도는 15% 이상 차이를 보일 만큼, 배출규제 적용 시점에 따른 PM농도의 차이가 크다고 할 수 있다. 연비 또한 배출가스 규제기준이 EURO6에 가까울수록 더욱 우수해지는 것으로 분석되었다.

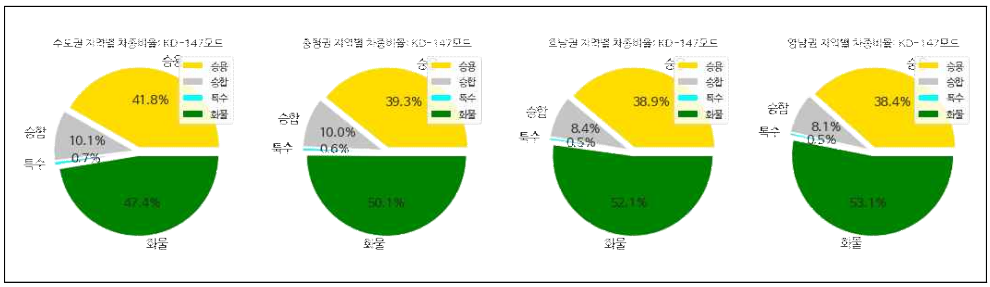
<표 3-9> 유로(EURO) 기준별 기초통계 분석(KD-147모드)

EURO3						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	16.5	3,020	651	2,710	219,508	5
std	2.6	1,300	884	713	114,827	0
min	13	1,475	0	1,493	1	4
median	16	2,775	500	2,607	203,404	5
max	35	38,420	25,000	16,991	2,244,437	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	5.9	5.0	1.8	1.9	11.5	17.0
std	3.4	0.7	0.1	0.2	5.1	16.8
min	1	3.5	1.5	1	2.7	0
median	6	5.0	1.8	1.8	10	12
max	39	12.6	2.5	3.9	185	100
EURO4						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	10.9	2,660	372	2,341	176,400	4
std	1.6	708	524	410	84,600	1
min	9	1,455	0	1,493	1	3
median	10	2,825	0	2,497	164,320	4
max	14	25,390	17,000	11,149	2,033,947	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	5.9	4.9	1.8	1.9	12.1	10.7
std	2.9	0.4	0.1	0.3	1.8	14.9
min	1	2.7	1.5	1.4	3.2	0
median	5	5	1.9	1.8	11.5	6
max	39	12.7	2.5	23.3	23.5	100
EURO5						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	6.8	2699	531	2317	126,409	3
std	1.5	761	553	379	78,305	0
min	4	1325	0	1,396	163	3
median	7	2895	600	2,497	111,584	3
max	9	12485	5,000	7,640	1,076,237	4
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	5.0	4.9	1.8	1.9	12.4	3.5
std	2.3	0.4	0.1	0.3	2.7	8.3
min	1	0.5	1.6	0.6	4	0
median	5	5.1	1.8	1.9	11.8	0
max	49	13	2.5	4	24.2	100

<표 계속>

EURO6						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	3.5	2,703	413	2,226	87,013	3
std	0.6	1,333	834	559	59,092	0
min	2.0	1,410	0	1,396	1,012	3
median	3.0	2,715	0	2,199	73,335	3
max	5.0	39,105	25,000	12,742	702,800	3
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	5.4	5.0	1.9	1.9	12.5	1.4
std	2.9	0.6	0.1	0.4	2.9	5.3
min	1.0	3.6	1.6	1.4	3.7	0.0
median	5.0	5.0	1.9	1.8	12.9	0.0
max	39.0	12.9	2.5	4.0	21.1	99

지역별 차종 비율의 편차는 크지 않은 것으로 나타났으며, 대도시가 많은 수도권 및 충청권 지역의 경우 승용, 승합차의 비율이 더욱 높아지는 것으로 나타났다. 반면 영남권의 물동량 거점지역과 주요 고속도로를 이용한 비율이 높은 만큼 타 지역보다 화물차량의 비율이 비교적 높은 것을 확인할 수 있다.



<그림 3-3> 지역별 차종 비율(KD-147모드)

데이터를 그룹별로 분류한 기준은 그룹 1이 모든 승용차 포함되었으며, 그룹 2에는 소형 및 중형의 승합, 화물, 특수차량, 그룹 3은 나머지 중대형 승합, 화물, 특수차량으로 분류하였다. 중대형 차량으로 구성된 그룹 3의 연식이 다른 그룹에 비해 특히 높은 것으로 나타났고 총중량, 평균 PM 배출량, 주행거리도 높은 것으로 분석되었다.

1.2 Lug-Down3모드

Lug-Down3모드의 주요 데이터를 기초통계 분석결과는 <표 3-11>과 같다. KD-147모드와 유사하게 PM 농도(%)의 평균에 비해 분산이 큰 편이기 때문에 표본 간의 이질성이 큰 것으로 판단된다. 평균 차량 연식은 약 10년 이상으로 KD-147모드보다 조금 차령이 긴 것으로 나타났다. 평균주행거리는 약 38만km를 상회하는 것으로 나타나 KD-147모드보다 주행거리에서 큰 차이를 보였다. 이는 Lug-Down3모드 검사차량이 중대형 화물차 및 특수차로서 영업용으로 사용하는 차량이 대부분이기 때문이다.

<표 3-10> 데이터 기초통계(Lug-Down3모드)

구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	10.6	15,838	6,760	7,878	380,713	3.8
std	5.8	9,813	7,270	3,519	300,424	0.9
min	1	0	0	1	1	2
median	9	12,270	3,500	6,606	309,822	3
max	36	248,200	125,002	19,543	9,413,420	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	6.5	9	2.4	3.2	6.9	7.1
std	11.8	2.4	0.2	0.5	2.5	11.1
min	1	0.7	0.2	0.2	2	0
median	3	8.7	2.5	3.3	7.5	3
max	74	129.1	21.7	39.5	30	100

다음으로는 차량의 규모(대형, 중형) 및 차종(승합, 화물, 특수)별 데이터 기초통계 분석 결과는 앞선 KD-147모드와 확연한 차이점이 발견되었다. 대형차량이 평균 연식과 PM 배출농도에서 모두 낮은 값을 보인 점은 흥미롭다. 반면 주행거리는 대형차량에서 평균적으로 더욱 높았다. 차량 연비는 중형차량이 우수한 것으로 나타났으며, 총중량 및 적재량의 경우 중형차량에 비해 대형차량이 높았다.

차종별 분석 결과 화물차량의 연식이 승합 및 특수차량의 연식보다 평균적으로 더 긴 것으로 나타났다. KD-147모드에서는 승합차량의 연식이 가장 길었던 점과 확연히 대비된다. 화물차량은 다른 차종에 비해 연비도 높고 평균 PM농도 또한 높았다. 특수차량의 경우 평균주행거리가 가장 높고 연비는 가장 낮은 것으로 분석되었다. 이 같은 특성은 PM 배출농도의 원인을 파악하는 데에 직관적인 도움을 받을 수 있을 것으로 예상된다.

<표 3-11> 차량 규모별 기초통계 분석(Lug-Down3모드)

대형						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	10.2	19,402	8,672	9,327	415,254	3.7
std	5.9	9,397	7,769	3,056	316,855	0.9
min	1	0	0	1	1	2
median	9	15,205	5,000	9,960	350,084	3
max	36	248,200	125,002	19,543	9,413,420	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	7.8	9.9	2.5	3.4	5.8	6.2
std	13.6	2.3	0.1	0.5	1.9	10.4
min	1	0.8	2	0.2	2	0
median	3	9.8	2.5	3.4	5.9	2
max	74	129.1	17	39.5	30	100
중형						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	11.8	6,799	1,912	4,204	293,100	4.1
std	5.4	1,626	1,024	1,131	232,008	0.9
min	1	0	0	2	1	3
median	12	6,680	2,000	3,933	226,837	4
max	33	22,760	8,000	14,866	6,721,568	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	3.3	6.8	2.1	3	9.8	9.2
std	2.9	1	0.2	0.4	1.4	12.5
min	1	0.7	0.2	0.3	4.2	0
median	3	6.6	2.1	3	10	5
max	67	61.8	21.7	31	17.6	100

<표 3-12> 차종별 기초통계 분석(Lug-Down3모드)

승합						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	7.4	13,828	3	10,725	364,476	3.3
std	3.7	2,798	101	2,855	301,641	0.7
min	1	2,400	0	1	1	2
median	6	15,015	0	12,344	296,916	3
max	30	21,060	14,300	17,787	3,548,033	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	38.8	11.2	2.4	3.3	6.4	4.5
std	11.2	1.5	0.1	0.2	1.4	8.2
min	1	4.9	1.8	1.8	3.3	0
median	45	12	2.5	3.4	6	1
max	74	13	2.9	4	17.6	99
화물						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	11.1	15,090	6,689	7,012	365,976	3.9
std	5.9	10,511	6,920	3,141	267,563	0.9
min	1	0	0	2	1	2
median	10	10,685	4,000	6,299	304,567	4
max	34	248,200	125,002	19,543	9,413,420	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	2.7	9	2.3	3.2	7.4	7.8
std	0.5	2.5	0.2	0.5	2.5	11.8
min	1	0.7	0.2	0.2	2	0
median	3	8.8	2.4	3.2	7.8	4
max	35	129.1	21.7	39.5	16.2	100
특수						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	10.6	21,177	12,138	10,219	468,511	3.8
std	5.8	7,391	7,403	3,621	420,283	0.9
min	1	3,010	0	2	1	3
median	9	25,270	16,500	11,946	382,482	3
max	36	36,380	40,000	17,990	7,332,207	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	2.2	7.1	2.4	3.4	5.1	5.1
std	0.4	0.9	0.2	0.3	2.7	8.7
min	1	4.5	1.7	1.9	2.1	0
median	2	6.9	2.5	3.3	3.7	2
max	9	13	2.8	4.2	30	100

유로 배출규제 적용기준에 따른 기초통계 분석 결과는 <표 3-19>와 같다. 앞선 KD-147모드와는 다르게 EURO4 및 EURO6에서 평균 PM농도가 크게 감소하는 것으로 나타났다. EURO6과 3은 약 10가량의 평균 PM농도의 차이를 보인다. 연비는 EURO6으로 올수록 더욱 우수해지고 주행거리는 더욱 짧아지는 것으로 나타났다.

<표 3-13> 유로(Euro) 기준별 기초통계 분석(Lug-Down3모드)

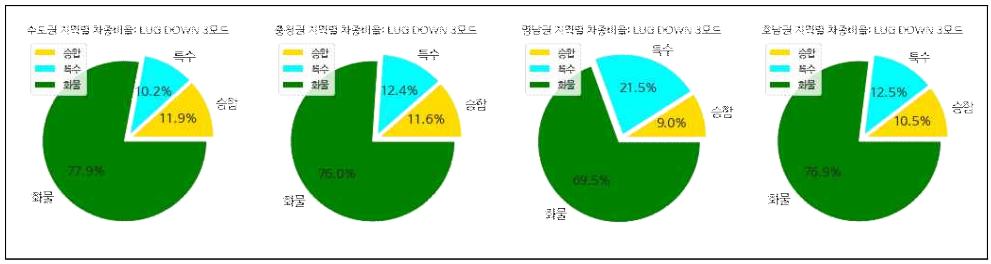
EURO3						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	17.4	14,588	6,717	7,828	386,663	5
std	3.8	9,542	6,645	3,648	316,681	0.1
min	13	0	0	1	1	2
median	17	10,155	3,900	6,606	306,500	5
max	36	248,200	125,002	19,543	6,721,568	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	3.8	8.2	2.3	3.1	6.9	12.3
std	6.8	2.3	0.2	0.5	2.6	13.9
min	1	0.7	0.2	0.2	2.1	0
median	3	7.7	2.4	3.1	7.6	8
max	67	126.9	21.7	39.5	30	100
EURO4						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	11.1	15,503	6,842	7,300	442,268	3.8
std	0.8	9,988	7,327	3,251	323,603	0.5
min	9	400	0	2	1	2
median	11	11,585	3,300	5,899	386,110	4
max	12	40,000	27,500	15,928	9,413,420	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	4.9	8.9	2.4	3.2	7.2	6.1
std	9.3	2.4	0.2	0.5	2.6	9.2
min	1	0.8	1.7	0.9	2	0
median	3	8.7	2.5	3.2	7.8	3
max	67	121.1	21.4	36.4	16.6	99

<표 계속>

EURO5						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	7.5	16,487	6,750	7,930	414,444	3.1
std	1.2	10,075	7,635	3,575	308,253	0.2
min	4	2,500	0	2	1	2
median	8	13,925	3,000	5,899	361,828	3
max	9	40,000	27,500	16,162	8,369,702	5
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	8.4	9.3	2.4	3.3	7	5.1
std	14.1	2.4	0.2	0.4	2.6	8.6
min	1	4.5	1.7	0.3	2.5	0
median	3	9.4	2.5	3.3	7.2	2
max	67	127.1	3.5	4.2	17.6	99
EURO6						
구분	연식 (년)	총중량 (kg)	적재량 (kg)	배기량 (cc)	주행거리 (km)	배출 가스등급
mean	4	17,167	6,787	8,277	286,223	3
std	1	9,515	7,665	3,353	213,821	0.1
min	1	2,470	0	2	1	3
median	4	14,375	4,000	6,728	248,908	3
max	5	39,990	27,500	16,353	7,332,207	4
구분	승차정원 (인)	길이 (m)	너비 (m)	높이 (m)	연비 (km/l)	PM (%)
mean	9.3	9.8	2.4	3.4	6.7	2.1
std	14.8	2.4	0.2	0.4	2.4	5.8
min	1	3	1.7	0.4	2.5	0
median	3	10	2.5	3.4	6.4	0
max	74	129.1	2.9	4	16.2	100

지역별로 등록된 차량의 차종 비율을 <그림 3-4>와 같다.지역별 차종 비율의 편차는 크지 않고 수도권 지역에서 화물차량과 승합차량의 비율이 높은 것으로 나타났다. 앞선 KD-147모드 처럼 영남권의 경우 주요 도로를 이용한 물동량이 집중되는 지역인 만큼 특수차량의 비율이 타 지역과 크게 차이를 보이는 것을 확인할 수 있다.

Lug-Down3모드에는 그룹 1이 포함되지 않으므로 그룹 2와 3의 기초통계만 분석했다. 두 그룹 간에 평균 PM 배출농도는 큰 차이는 없는 것으로 보이며, 배출가스 등급도 평균 4등급으로 동일하다. 주행거리와 총중량이 그룹 3이 2에 비해 두배 가량 많은 것으로 나타났다.

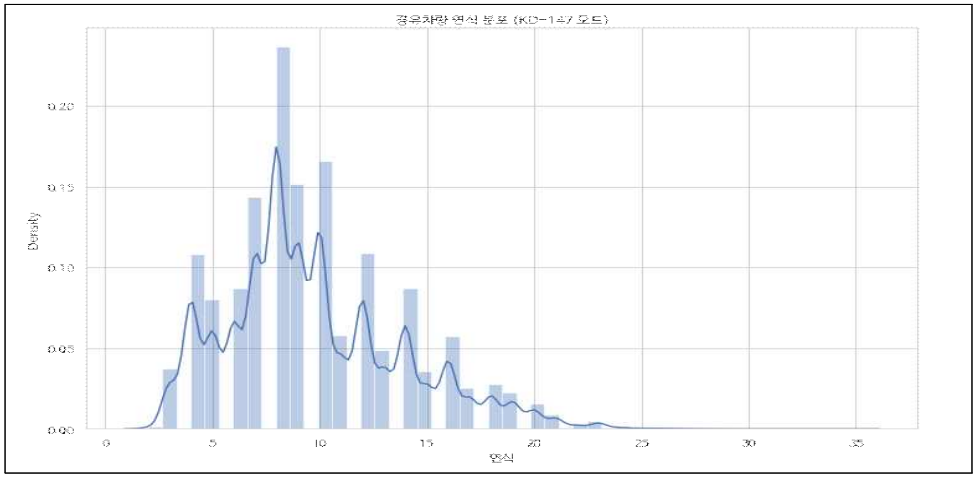


<그림 3-4> 지역별 차종 비율(Lug-Down3모드)

2. 차량 정밀검사 데이터 분석

2.1 차량연식

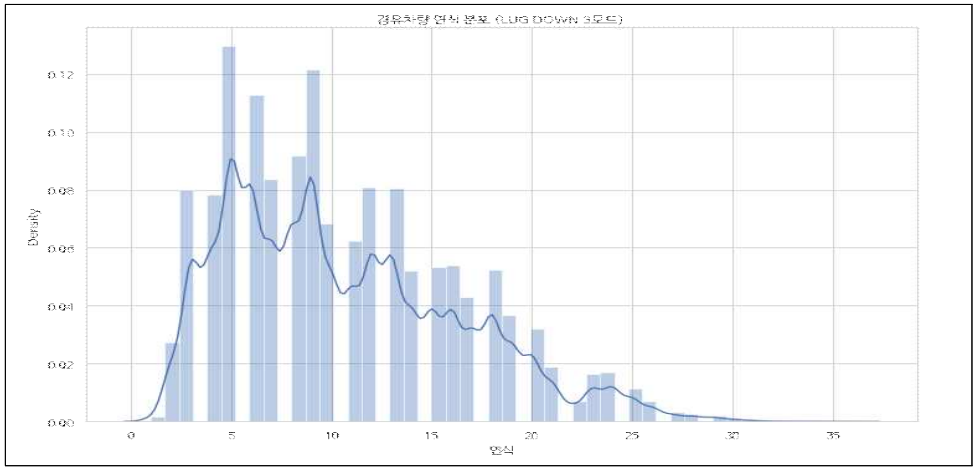
차량연식은 차량제작일자, 연식, 최초등록일로 구성되어 있다. 차량제작일자는 차량이 생산된 '년월일'이고, 연식은 '몇년형'으로 칭하는 차량 판매 시 제시되는 차량제원의 형식이며, 최초등록일은 차량을 구매하고 차량등록일을 의미한다. 본 연구에서는 차량연식은 2019년을 기준으로 차량제작일자를 뺀 연도를 차량연식으로 산정하였다.



<그림 3-5> 차량 연식 분포(KD-147모드)

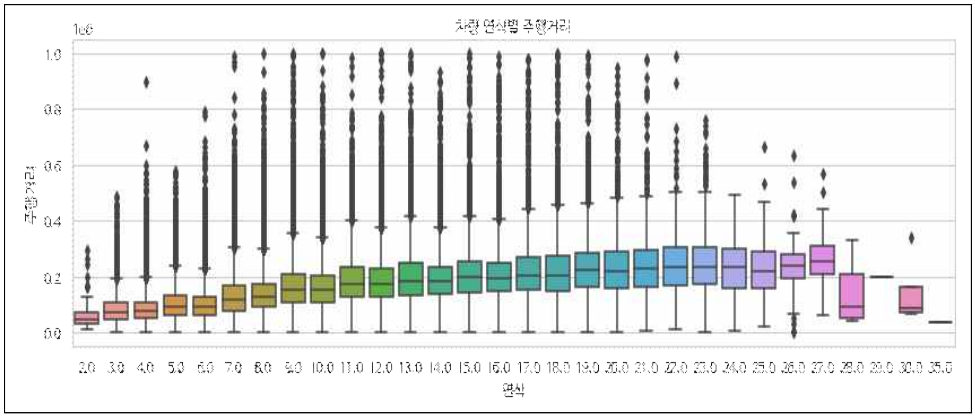
<그림 3-5>의 차량연식 분포를 살펴보면, KD-147모드 데이터의 최빈

값은 약 8년 부근에 형성되어 있으며, Lug-Down3모드는 약 5년으로 나타났다. 그러나 차량연식의 전반적인 분포를 보면 Lug-Down3모드 검사 차량 연식이 높은 차량 분포가 많은 것으로 확인된다.

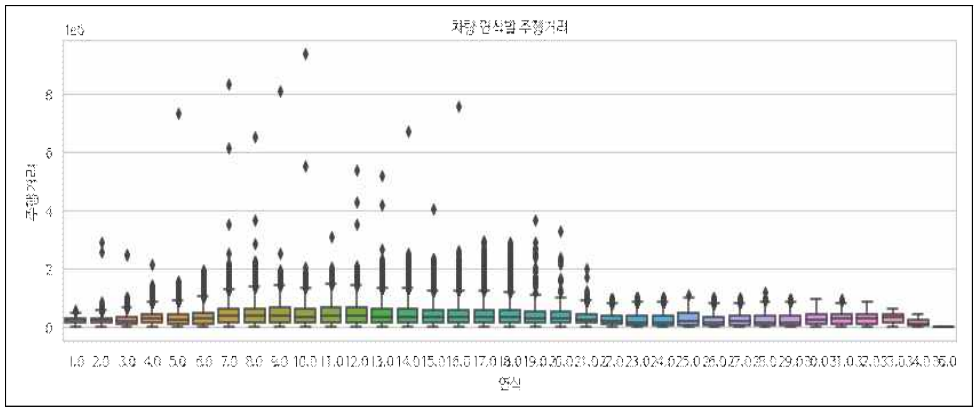


<그림 3-6> 차량 연식 분포(Lug-Down3모드)

차량연식별 주행거리는 KD-147모드 차량은 25년까지는 증가추세를 보이며, Lug-Down3모드는 20년까지 증가추세를 보인다. Lug-Down3모드 차량은 대부분의 사업용도로 사용하기 때문에 주행거리는 KD-147모드 차량 많으나 연식이 오래된 차량 비중은 낮은 편이다.



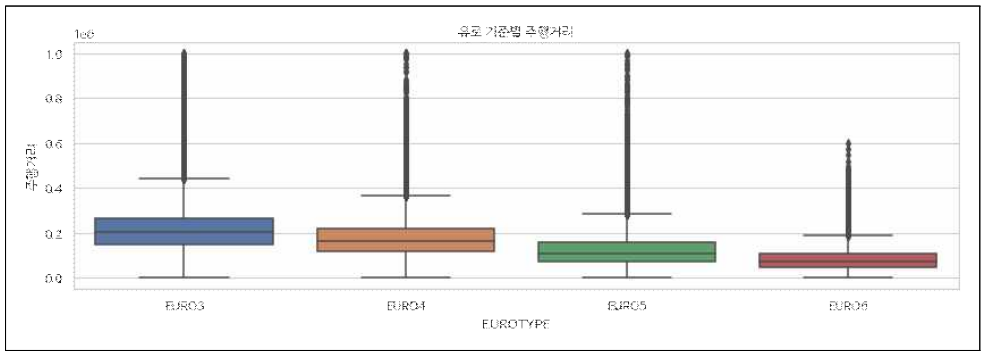
<그림 3-7> 차량 연식별 주행거리(KD-147모드)



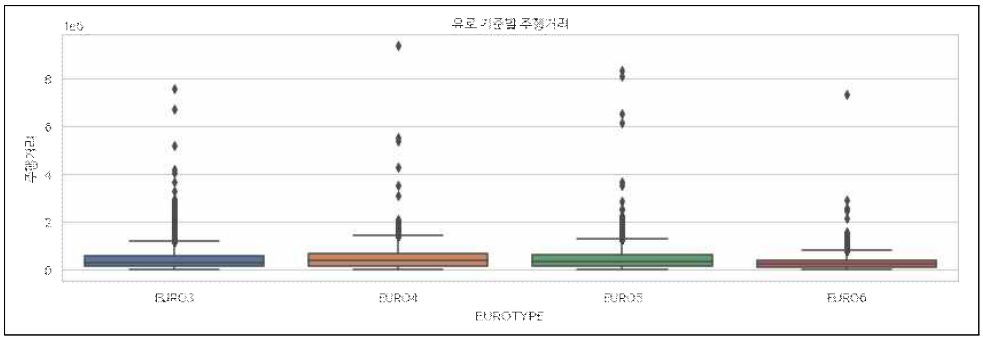
<그림 3-8> 차량 연식별 주행거리(Lug-Down3모드)

2.2 배출가스 규제기준(EURO)

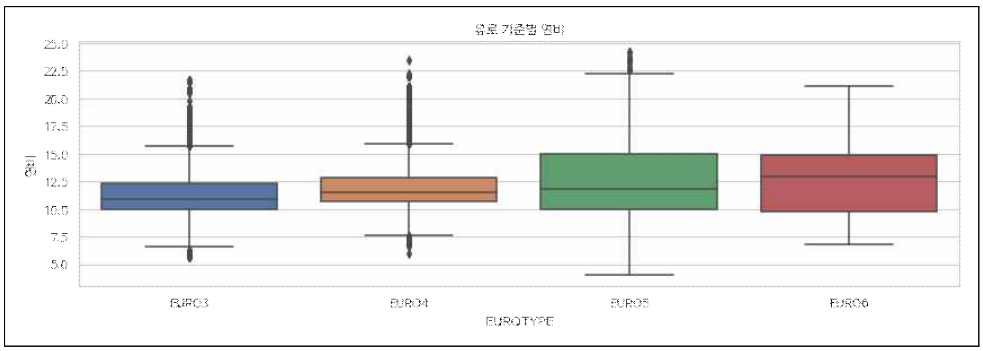
우리나라 배출가스 규제기준은 유로기준에 따른다. 현재는 EURO6 규제가 적용되고 있으며, 차량이 배출하는 일산화탄소, 질소산화물, 미세먼지 및 탄화수소를 특정 수치 이하로 제한해야 한다. <그림 3-9>과 <그림 3-10>은 유로기준별 KD-147모드와 Lug-Down3모드별 주행거리 분포를 나타낸 것이다. KD-147모드는 EURO 기준이 상위로 갈수록 평균 주행거리가 감소하는 것으로 나타났다. 반면 Lug-Down3모드는 EURO6 기준부터 감소하는 경향을 보인다. 연비를 살펴보면 KD-147모드는 유로기준에 따른 차이는 거의 미미하며, Lug-Down3모드는 EURO6에서 연비가 감소하는 것으로 나타났다.



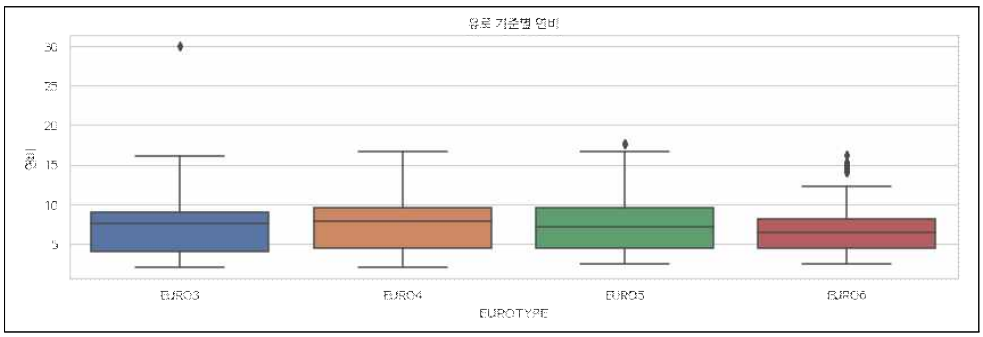
<그림 3-9> 유로 기준별 주행거리(KD-147모드)



<그림 3-10> 유로 기준별 주행거리(Lug-Down3모드)



<그림 3-11> 유로 기준별 연비(KD-147모드)



<그림 3-12> 유로 기준별 연비(Lug-Down3모드)

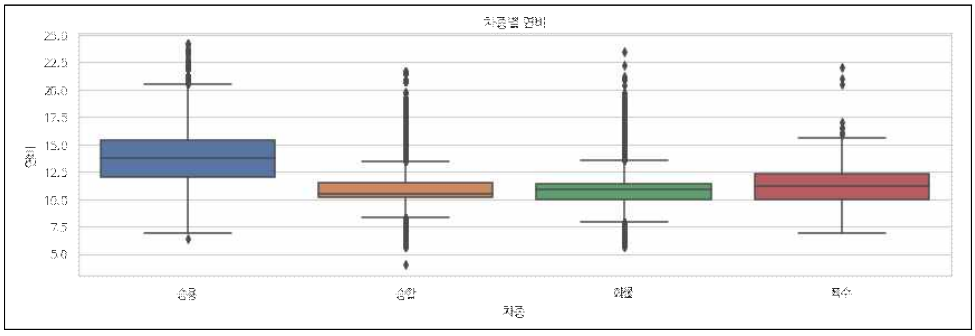
2.3 차종

본 연구에서 차종은 승용, 승합, 화물, 그리고 특수차량으로 구분한다. 승용자동차는 10인 이하를 운송하기에 적합하게 제작된 자동차이며, 승

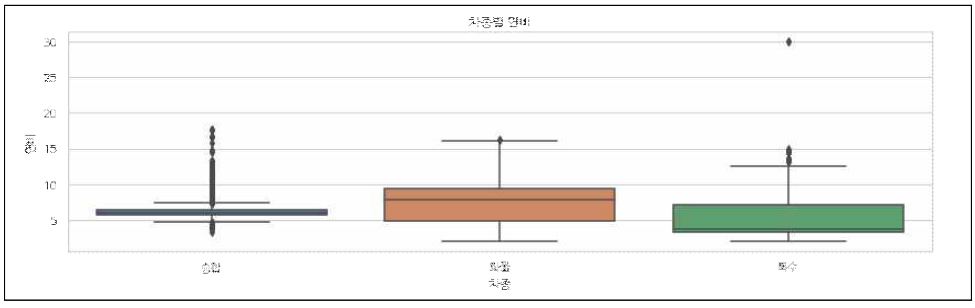
합차는 11인 이상 운송 가능한 차종이 이에 해당된다. 화물차는 적재공간을 갖추며 총 적재화물의 무게가 운전자를 제외한 총 승객의 무게보다 높은 차를 의미한다. 특수차량은 견인, 구난작업 등의 특수한 작업을 수행하기에 적합한 자동차를 의미한다.

KD-147모드의 차종별로 연비가 크게 다르지 않은 것으로 나타났으며 화물 차종에 높은 연비를 가진 이상치가 다량 확인되었다. Lug-Down3 모드의 경우 화물차의 연비 중앙값이 가장 높았다.

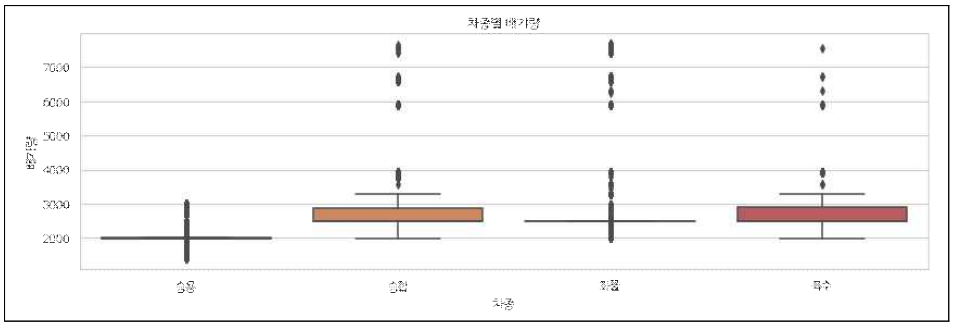
차종별 배기량의 상자도표는 <그림 3-13>과 <그림 3-14>에 제시하였다. KD-147모드에 승용차인 경우 다른 차종에 비해 배기량이 유독 낮다는 것을 확인할 수 있으며, 화물 및 특수차량의 경우에 이상치가 다수 분포해있다. Lug-Down3모드는 화물차량의 연비 중앙값이 가장 낮으며, 승합차량과 특수차량은 유사한 수준을 유지하고 있는 것을 확인할 수 있었다.



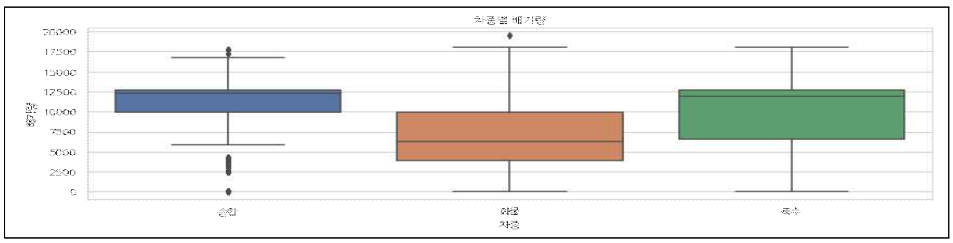
<그림 3-13> 차종별 연비(KD-147모드)



<그림 3-14> 차종별 연비(Lug-Down3모드)



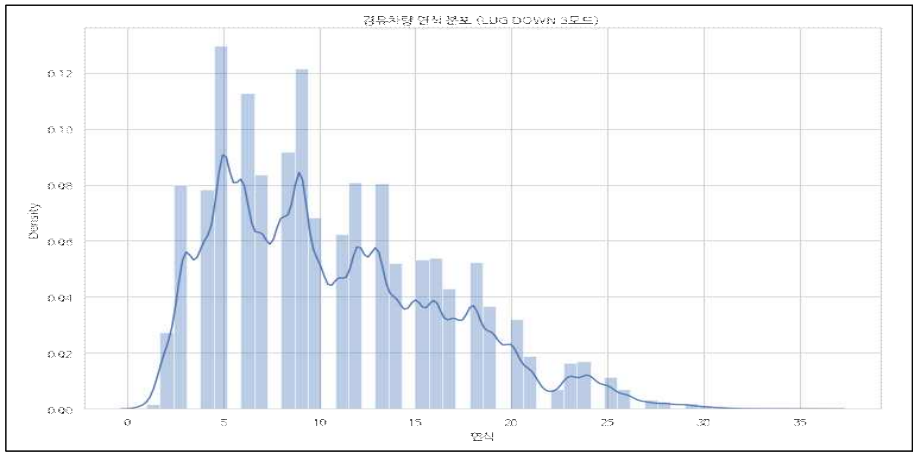
<그림 3-15> 차종별 배기량(KD-147모드)



<그림 3-16> 차종별 배기량(Lug-Down3모드)

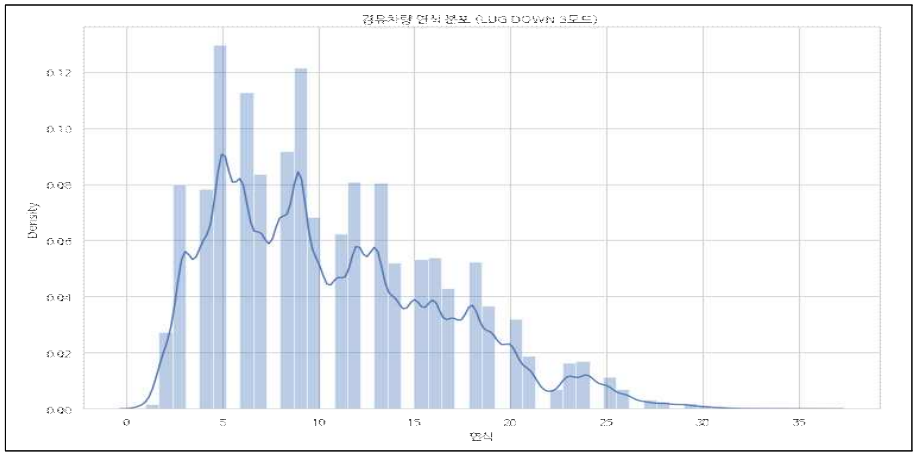
2.4 배기량

KD-147모드와 Lug-Down3모드의 배기량 분포를 살펴보면 KD-147모드의 경우 표본에 따라 배기량에 큰 차이가 없이 특정값에 밀집되었다. 반면 Lug-Down3모드는 표본별로 배기량 차이가 크게 나타났다.



<그림 3-17> 차량 배기량 분포(KD-147모드)

앞선 분석에서도 보았듯이 KD-147모드는 승용차 비중이 높고 Lug-Down3모드는 화물, 특수차량 비율이 높기 때문에 두 정밀검사 종류별 수집되는 차량 데이터 분포에 차이가 있음을 다시 확인할 수 있다.



<그림 3-18> 차량 배기량 분포(Lug-Down3모드)

<표 3-14> 지역별 평균 배기량

단위 : cc

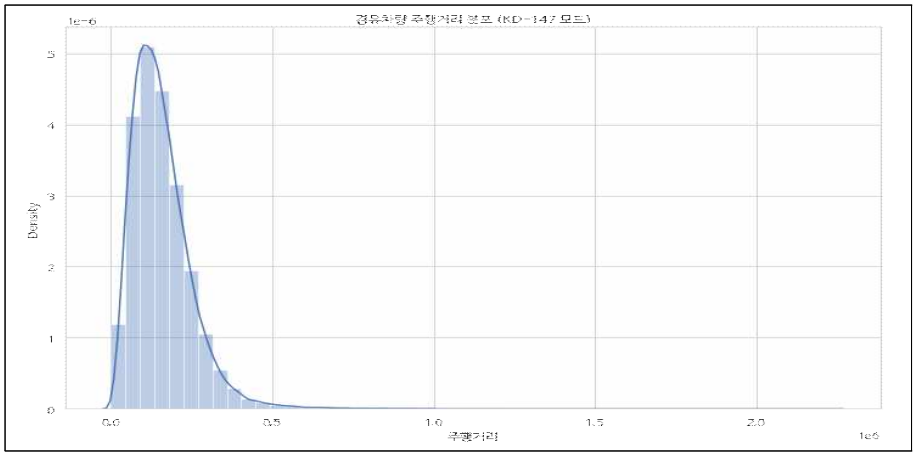
KD-147모드					
서울	부산	대구	인천	광주	대전
2,370.2	2,415.1	2,408.1	2,319.7	2,399.1	2,403.7
울산	경기	충북	충남	경북	경남
2,371.7	2,403.2	2,381.5	2,399.3	2,432.8	2,409.6
Lug-Down3모드					
서울	부산	대구	인천	광주	대전
7,092.4	9,145.9	7,445.0	8,719.1	7,759.5	8,274.6
울산	경기	충북	충남	경북	경남
9,106.4	7,061.3	8,178.2	7,805.5	9,079.6	7,901.7

2.5 주행거리

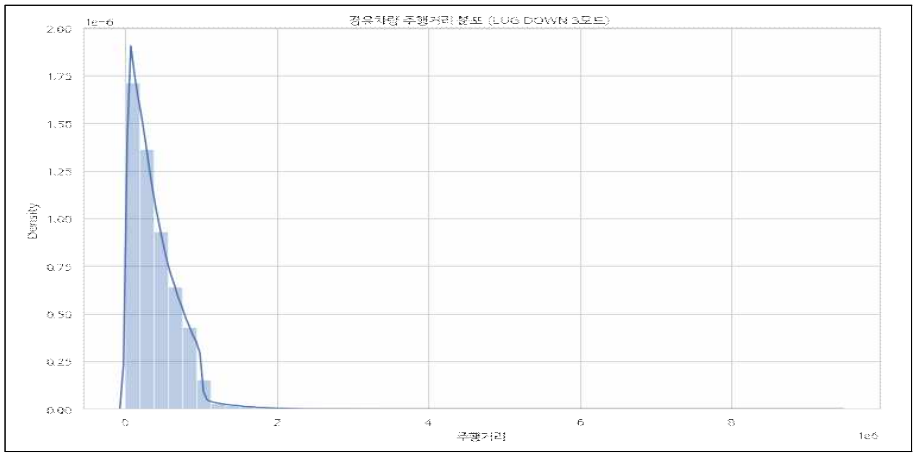
앞선 기초통계 분석에서 KD-147모드와 Lug-Down3모드의 조사 표본의 평균 주행거리에 큰 차이가 있음을 확인하였다. 주행거리 분포는 KD-147모드의 경우 정규분포에 가까우며, Lug-Down3모드는 왜도

(Skewness)가 우측으로 치우쳐져 있다. Lug-Down3모드는 영업용 차량이 많아 주행거리가 훨씬 긴 것을 알 수 있다.

지역별 평균 주행거리를 살펴보면 두 모드에서 모두 서울 지역이 가장 짧은 주행거리를 가지는 것으로 나타났다. KD-147모드에서 평균 주행거리가 가장 긴 지역은 대구, Lug-Down3모드에서는 부산으로 드러났다. 두 도시 모두 영남권에 속해있는 화물 물동량이 많은 지역이라는 공통점이 있다.



<그림 3-19> 차량 주행거리 분포(KD-147모드)



<그림 3-20> 차량 주행거리 분포(Lug-Down3모드)

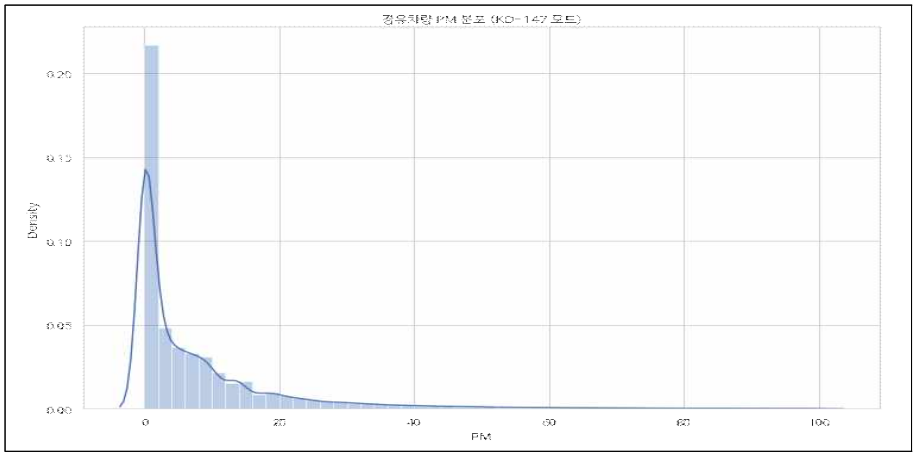
<표 3-15> 지역별 평균 주행거리

단위 : km

KD-147모드					
서울	부산	대구	인천	광주	대전
145,548	159,067	174,489	139,595	163,152	168,046
울산	경기	충북	충남	경북	경남
148,865	163,101	162,491	160,426	159,694	165,601
Lug-Down3모드					
서울	부산	대구	인천	광주	대전
341,426.8	427,620.0	379,634.1	396,262.7	388,362.2	386,920.0
울산	경기	충북	충남	경북	경남
372,100.8	365,967.1	388,115.9	356,703.0	425,214.2	377,400.8

3. PM 농도 데이터 분석

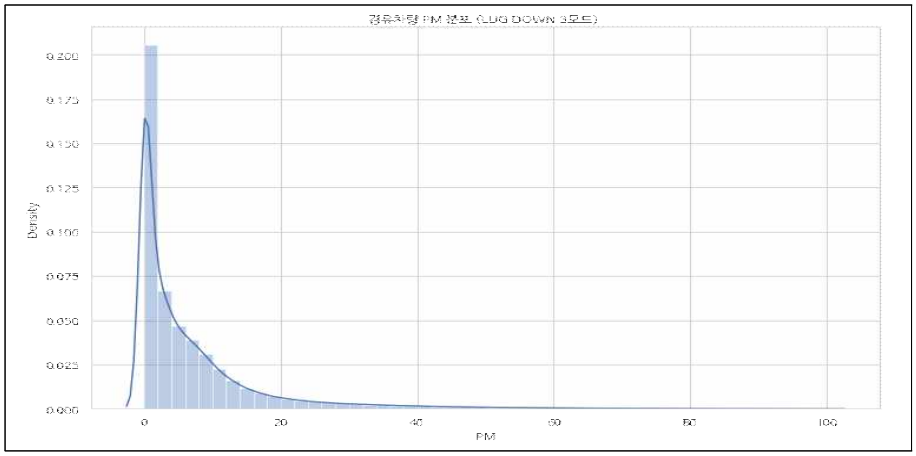
PM 농도의 데이터 단위는 %이고, 0~100%로 다양하게 분포되어 있다. PM 농도 데이터와 차량정밀검사 데이터의 관계를 파악하고자 분포 그래프 및 상자도표를 통해 기초적인 상관관계를 분석하였다.



<그림 3-21> 차량 PM 배출농도 분포(KD-147모드)

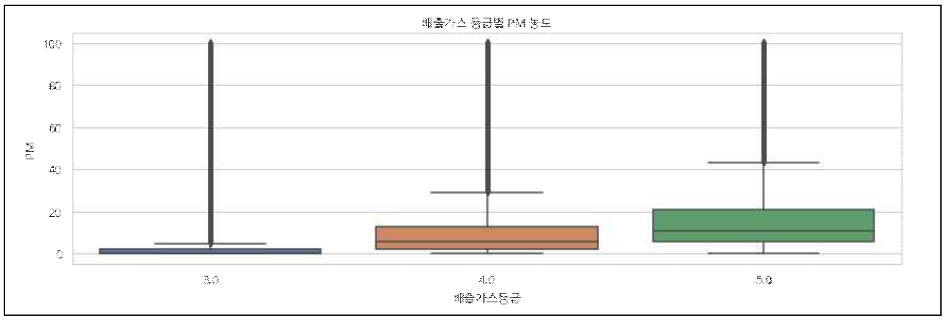
<그림 3-21>에 정밀검사모드별 PM 농도 그래프를 보면 두 검사 모두에서 PM 농도의 대부분은 0-20% 사이에 분포하고 있으며, 왜도(Skewness)가 오른쪽으로 치우친 분포 형태를 나타내고 있다. 일부 차량의 경우 20%를 넘는 큰 값을 보이는데 해당 값들은 분포에서 크게 떨

어져 있다. 따라서 이러한 높은 수치들을 논리적 이상치로 제거하거나, 해당 차량에서 유독 높게 측정된 원인을 규명한다면 더욱 정확한 분석 결과를 얻을 수 있을 것으로 예상된다.

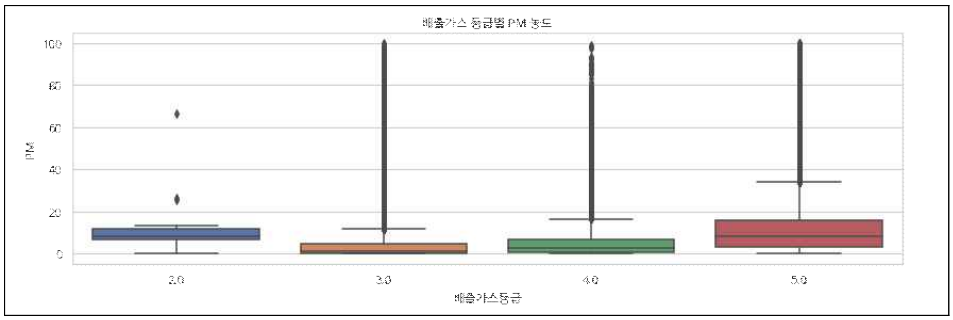


<그림 3-22> 차량 PM 배출농도 분포(Lug-Down3모드)

배출가스등급과 PM 농도와의 관계를 분석한 상자도표는 <그림 3-23>과 <그림 3-24>이다. KD-147모드에서 배출가스 등급이 올라갈수록 PM 농도의 중앙값이 올라가며 최댓값과 최솟값의 차이도 커진다. 또한 이상치는 세 등급에서 모두 다수 분포하는 것을 관측할 수 있다. Lug-Down3 모드는 배출가스 2등급 차량도 포함하여 조사한 결과, 모든 등급의 이상치를 제외한다면 2등급 차량의 평균 PM이 다른 등급에 비해 높게 형성되는 것을 알 수 있다.

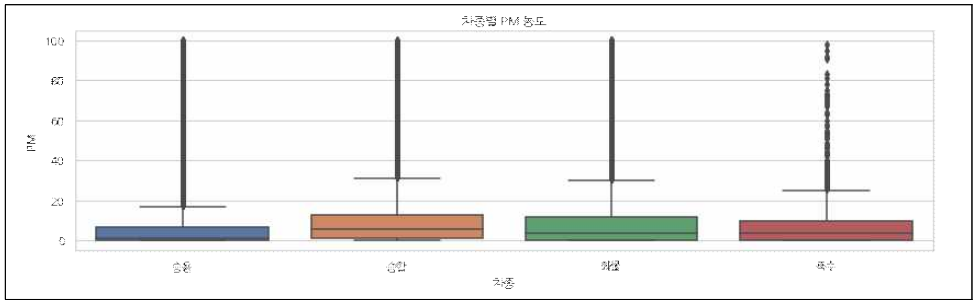


<그림 3-23> 배출가스 등급별 PM 농도(KD-147모드)

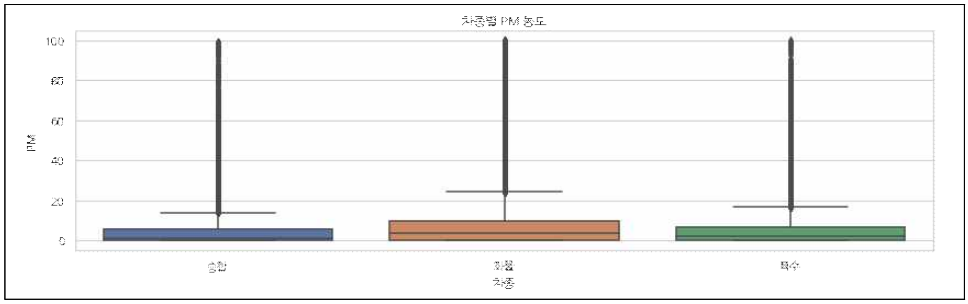


<그림 3-24> 배출가스 등급별 PM 농도(Lug-Down3모드)

차종별 PM 농도는 KD-147모드에서는 승합차량과 화물차량의 PM 중앙값이 다른 차종에 비해 약간 높은 것으로 분석되었으며, 승용차의 PM 값은 확연히 낮다는 것을 확인할 수 있다. 반면 Lug-Down3모드에서는 화물차량의 평균 PM 농도가 가장 높게 나타났다. 이는 배출가스 정밀검사 및 차종별 PM 분포에 차이가 존재함을 확인하였다.

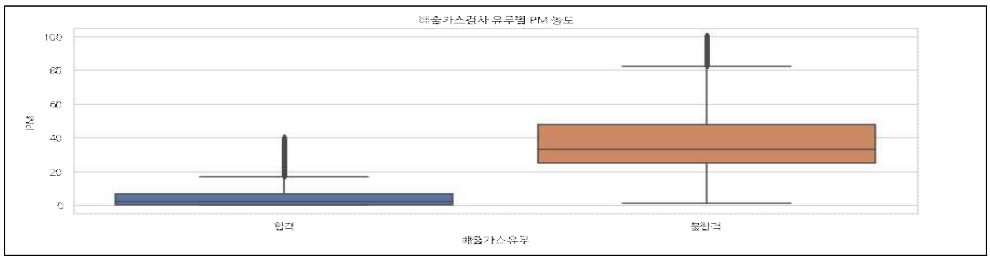


<그림 3-25> 차종별 PM 농도(KD-147모드)

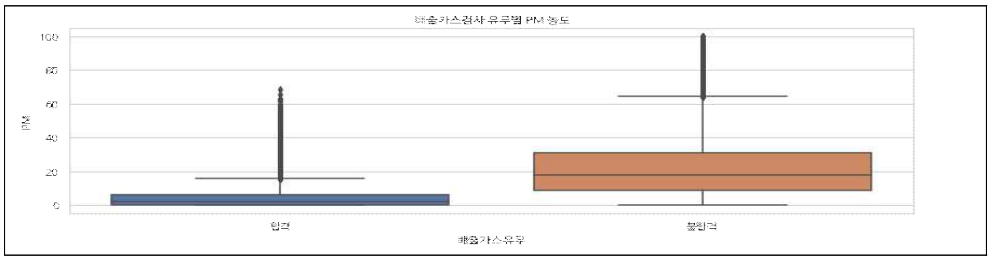


<그림 3-26> 차종별 PM 농도(Lug-Down3모드)

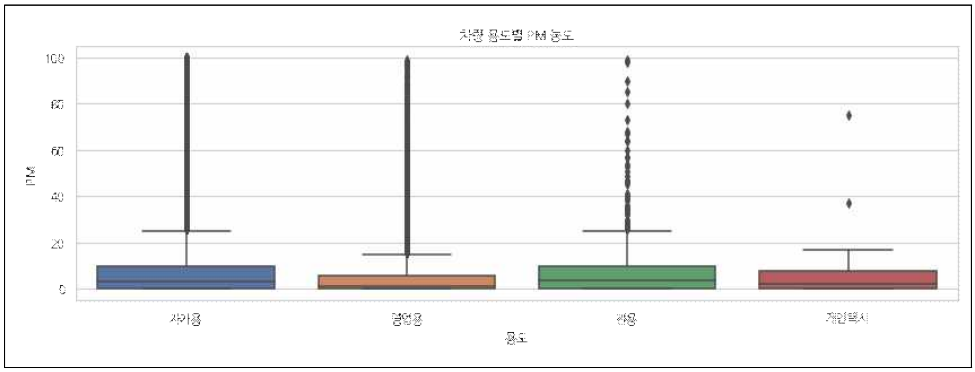
배출가스검사 합격유무별 PM 농도의 분포를 살펴보면 두 정밀검사모드 모두 PM농도가 합격차량은 낮고 불합격 차량은 높은 것을 확인할 수 있었다. KD-147모드에서는 합격차량과 불합격차량의 PM농도의 중앙값은 20% 이상의 차이를 보이며, 불합격 차량에서도 이상치가 많이 존재하는 것을 알 수 있다. Lug-Down3모드에서도 마찬가지로 PM 중앙값의 차이가 15% 이상을 상회한다.



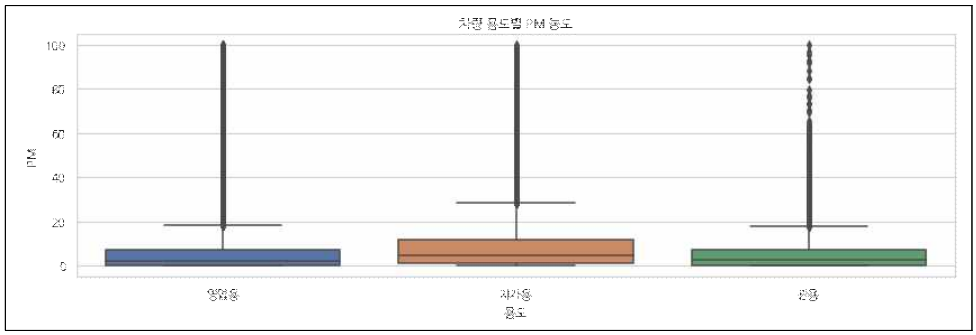
<그림 3-27> 배출가스검사 합격유무별 PM 농도(KD-147모드)



<그림 3-28> 배출가스검사 합격유무별 PM 농도(Lug-Down3모드)



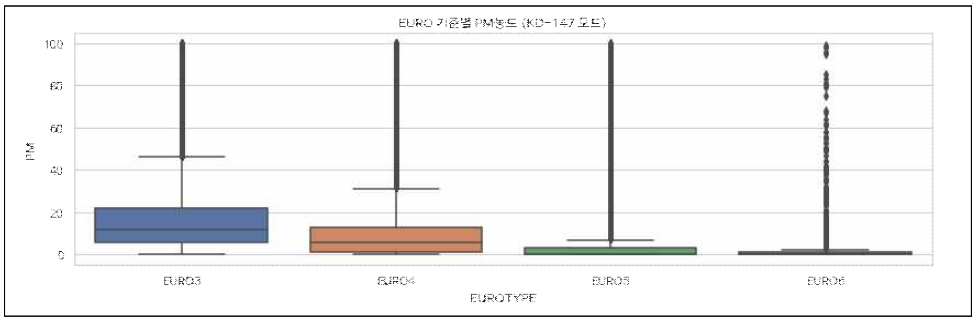
<그림 3-29> 차량 용도별 PM 농도(KD-147모드)



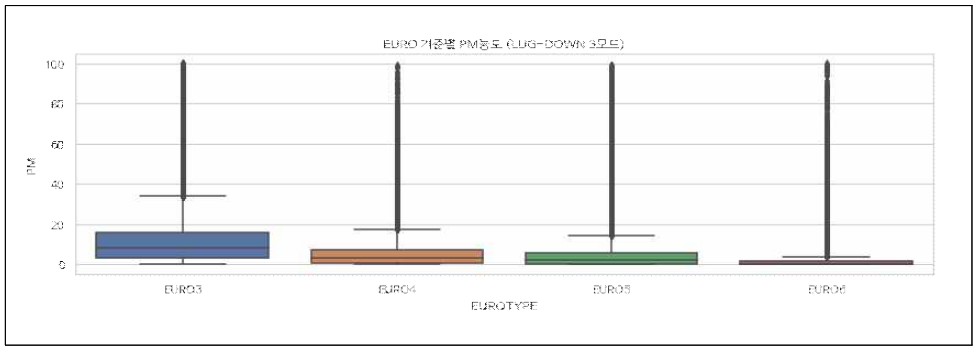
<그림 3-30> 차량 용도별 PM 농도(Lug-Down3모드)

<그림 3-29>와 <그림 3-30>은 차량의 용도별로도 PM 분포에 차이를 확인하기 위해 상자도표를 작성하였다. KD-147모드에서 개인택시의 이상치가 가장 적은 것으로 도출되었고, 영업용 차량의 최빈값이 가장 낮으나 이상치가 다수 분포한다는 특징을 보였다. Lug-Down3모드는 세 차량 용도 모두 이상치를 포함하고 있으며, 자가용의 중앙값이 다른 두 차량 용도에 비해 가장 높다는 것을 알 수 있다.

EURO기준별 PM 농도는 EURO3~EURO6 순차적으로 PM이 감소하는 것을 알 수 있다. 이는 배출기준이 엄격할수록 PM 배출농도 평균값은 작아지는 것을 의미한다. 다만 Lug-Down3모드의 분산이 KD-147모드보다 더 크므로 KD-147모드 데이터가 모형의 적합도 측면에서 더 높다고 할 수 있다.

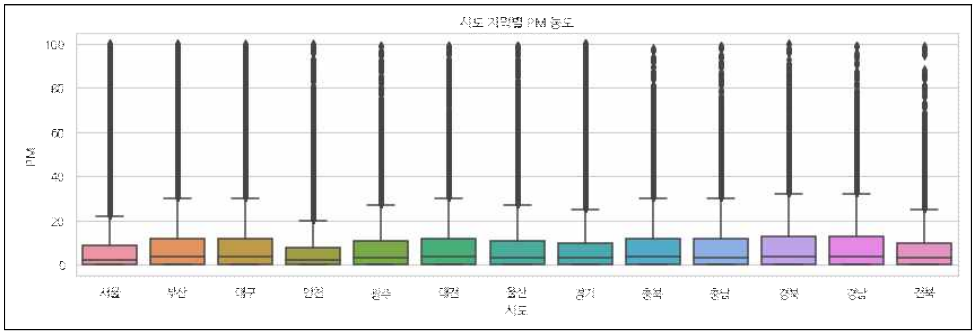


<그림 3-31> EURO기준별 PM 농도(KD-147모드)

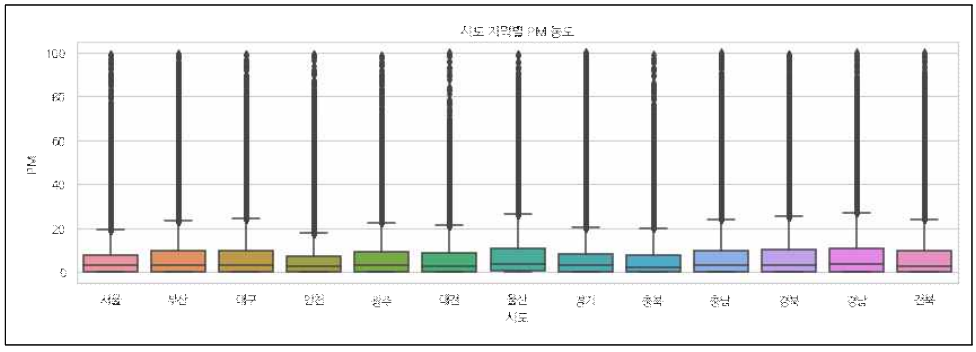


<그림 3-32> EURO기준별 PM 농도(Lug-Down3모드)

<그림 3-33>과 <그림 3-34>는 시도별 지역에 따라 PM 농도의 차이를 나타냈다. 두 가지 검사모드 모두 이상치를 제외하면, 지역별 PM 농도에 눈에 띄는 큰 차이는 보이지 않는다. PM의 최빈값과 최댓값이 가장 낮은 지역은 공통적으로 서울과 인천으로 확인되었다.



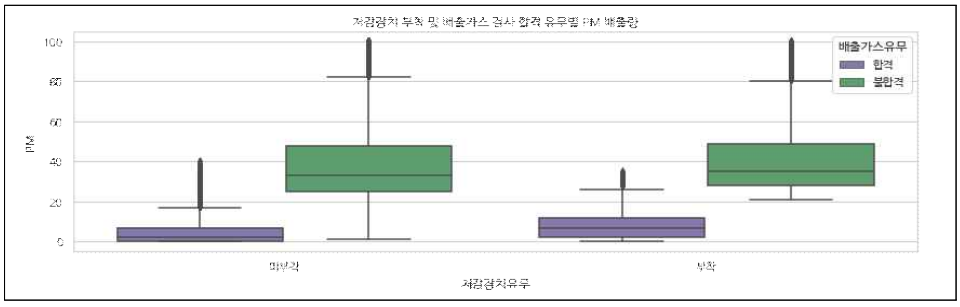
<그림 3-33> 시도 지역별 PM 농도(KD-147모드)



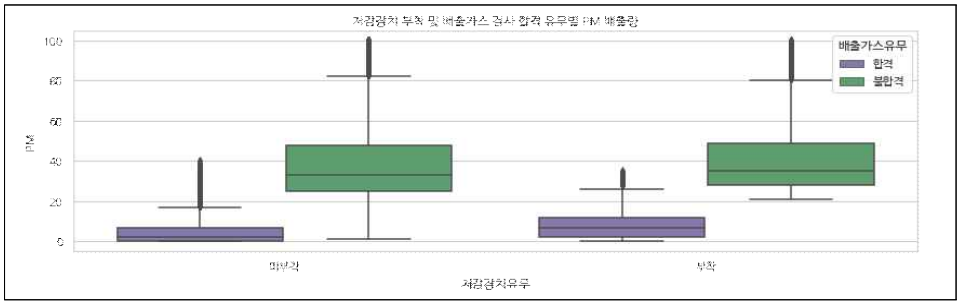
<그림 3-34> 시도 지역별 PM 농도(Lug-Down3모드)

<그림 3-35~38>은 저감장치 부착과 배출가스검사 합격유무, EURO 기준별 배출가스검사 합격유무를 상자도표로 작성하였다. 앞선 기술통계 분석에서 언급한바, KD-147모드 및 Lug-Down3모드 모두에서 배출가스검사 합격 유무에 따른 PM 분포에 큰 차이가 존재한다. 배출가스검사에 합격한 차량의 경우 그렇지 않은 차량에 비해 이상치도 적고 값의 변동 역시 작다는 것을 알 수 있다. 또한 배출가스 저감장치를 부착한 차량의 중앙값 및 최댓값이 미부착 차량에 비해 모두 높은 것으로 분석되었다.

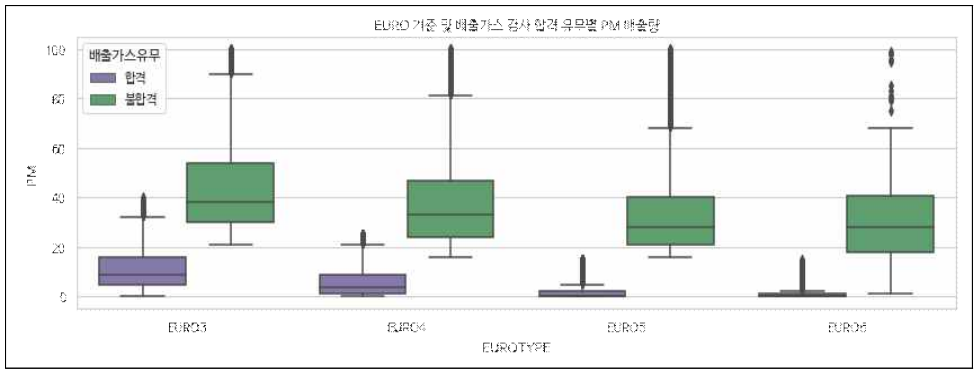
EURO기준별 배출가스검사 합격유무 분석결과는 상위 기준으로 갈수록 합격 대비 불합격 비율이 낮아지는 추세를 확인할 수 있었다. 특히 EURO3과 EURO6의 합격과 불합격 비율을 비교해보면 EURO6의 불합격 비율이 현저히 낮아지는 것으로 분석되었다. 다만 배출가스검사 합격 차량에도 일부 PM 고농도 배출차량이 존재하는 것으로 확인되었다. 이는 저감장치가 PM 고농도 배출을 저감 기능이 완벽히 발휘되지 않는 것으로 해석된다.



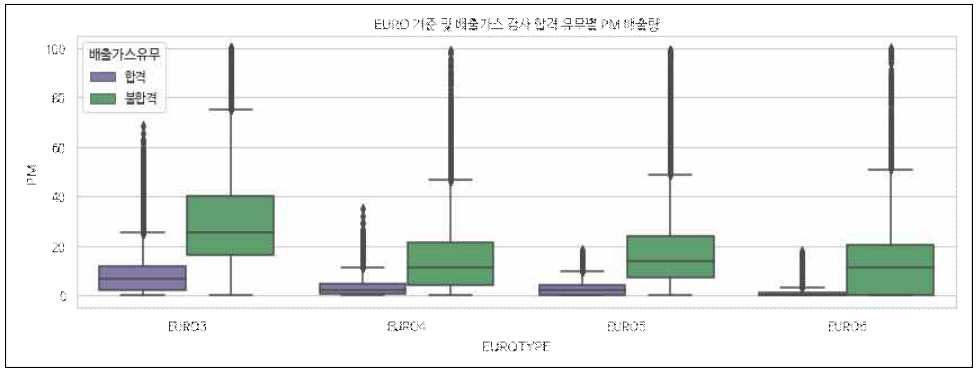
<그림 3-35> 저감장치 및 배출검사 합격 유무별 PM 농도(KD-147모드)



<그림 3-36> 저감장치 및 배출검사 합격 유무별 PM 농도(Lug-Down3모드)



<그림 3-37> EURO기준의 배출검사 합격 유무별 PM 농도(KD-147모드)



<그림 3-38> EURO기준의 배출검사 합격 유무별 PM 농도(Lug-Down3모드)

제Ⅳ장 연구 방법론

제1절 개요

이 장에서는 경유자동차 PM 배출요인 분석과 예측을 위해 머신러닝기법 중에 하나인 앙상블 학습을 활용하기 위해 관련 이론을 검토한다. 이와 같은 예측방법론은 채택한 배경은 다음과 같다.

첫째, 경유차 PM 배출요인이 다양하고 복잡하기 때문에 이 같은 문제를 풀 때 규칙기반 인공지능보다는 머신러닝이 더 효율적으로 해결할 수 있다. 둘째, 경유차 PM 배출에 영향을 미치는 요인을 파악하기 위해 요인별 영향력의 크기를 분석해야 한다. 요인별 영향력을 분석하기에는 딥러닝 기법보다 머신러닝 기법을 활용하는 것이 설명력이 더 뛰어나다. 셋째, 경유차 PM 배출요인의 특징을 분석하고 새로운 알고리즘을 적용하는 것보다 컴퓨터가 직접 특징을 추출하는 것이 더 빠르고 정확하기 때문에 앙상블 학습을 활용하고자 한다.

2절은 머신러닝기법에 대해 살펴본다. 본 연구는 경유차 PM 배출요인 분석이 연구 목적이므로 회귀모형이 적합하다. 따라서 분류와 회귀 문제를 모두 적용할 수 있는 앙상블 학습을 고찰한다.

3절에서는 앙상블 모형의 개별모형에 대해 제안한다. 개별모형의 개념, 발전과정, 분석방법론 등을 설명한다. 또한 모형의 장단점, 모형별 차이점 등을 상세히 정리한다.

4절은 본 연구에서 제안할 앙상블 학습 예측모형 구축방법론을 제시한다. 모형 설계와 분류체계를 구축하고, 데이터 전처리 과정에서 최종 예측모형 구축과정까지 단계별로 상세히 설명한다.

제2절 머신러닝

머신러닝은 인공지능에 속한 분야로, 분석자료를 통해 기계가 스스로 학습한 후 어떠한 판단이나 예측을 하는 것을 의미한다. 인공지능을 구현하는데 대표적인 방법 중 하나이며 인간처럼 학습시키고 더욱 똑똑하게 만드는 방법이다. 인공지능에게 데이터를 많이 학습시킬수록 정확한 결과를 얻어낼 수 있다. 이러한 규칙을 이용한 머신러닝 기술은 다양한 데이터를 통한 예측치 제시하거나 개인이 선호하는 상품, 동영상을 추천해주는 등 다양한 분야에 활용되고 있다.

머신러닝은 지도학습과 비지도학습으로 분류된다. 지도학습은 입력값에 정답을 학습시키면 새로운 입력값이 주어졌을 때 미리 학습된 데이터를 통해 예측하는 방법이다. 지도학습의 대표적인 알고리즘들은 선형 서포트 벡터 머신(Linear Support Vector Machine: Linear SVM), 나이브 베이즈(NaiveBayes), 로지스틱 회귀(Logistic Regression), 커널 서포트 벡터 머신(Kernel Support Vector Machine) 등이 있다. 이 알고리즘은 분류를 목적으로 예측기법이며, 분류와 회귀 문제를 해결할 수 있는 알고리즘은 의사결정 나무(Decision Tree),랜덤 포레스트(Random Forest), 그래디언트 부스팅(Gradient Boosting), 극단부스트(XGBoost), Light GBM, CatBoost, 선형회귀(Linear Regression) 등이 있다.

비지도학습은 정답이 없는 데이터들을 자동으로 군집하여 규칙을 스스로 발견하게 하는 학습 방식이다. 비지도 학습은 미분류 데이터가 활용되며, 밀도기반 군집(Density-Based Spatial Clustering), 계층적 군집(Hierarchical Clusting), 특이값 분해(Singular Value Decomposition) 등 있다. 이 같은 머신러닝 기법을 정리하면 아래 <표 4-1>과 같다.

<표 4-1> 머신러닝 기법 분류체계

구분	유형	분석모형	유형
지도학습 (Supervised learning)	예측 (Prediction)	선형 서포트 벡터머신 (Linear Support Vector Machine)	분류
		나이브 베이즈 (Naive Bayes)	분류
		로지스틱 회귀분석 (Logistic Regression)	분류
		커넬 서포트 벡터머신 (Kenel Support Vector Machine)	분류
		의사결정나무 (Decision Tree)	**분류/회귀**
		랜덤 포레스트 (Random Forest)	**분류/회귀**
		그래디언트 부스팅 (Gradient Boosting)	**분류/회귀**
		극단부스트 (XGBoost)	**분류/회귀**
		LightGBM	**분류/회귀**
		CatBoost	**분류/회귀**
		선형 회귀분석 (Linear Regression)	**회귀**
비지도학습 (Unsupervised learning)	반응형 (Have a Response)	가우시안 혼합모형 (Gaussian Mixture Model)	군집화
		K-평균 (K-Mean)	군집화
		밀도기반 군집 (Density-Based Spatial Clustering)	군집화
		계층적 군집 (Hierarchical Clusting)	군집화
	차원축소 (Demension Reduction)	주성분 분석 (Principal Component Analysis)	차원축소
		특이값 분해 (Singular Value Decomposition)	차원축소
		잠재 다리클레 할당 (Latent Dirichlet Allocation)	차원축소

제3절 앙상블 학습

앙상블 학습(Ensemble Learning)은 여러 가지 우수한 학습모형을 조합해 예측력을 향상시키는 기법이다. 단일모형에 비해서 분류 성능 우수하고 최적의 단일 모형을 생성하는 시간이 적게 소요된다. 특히 여러 학습이 적용된 서로 다른 예측모형을 생성 후, 예측모형의 분석결과를 종합하여 최종 예측결과를 도출해 내는 방법을 말한다. 이런 기법을 모형에 병합하면 단일 편향에 영향을 받지 않아 과적합의 문제를 해결할 수 있다. 무엇보다도 단일모형보다 여러 개의 모형을 구축하는 데 효율적이며, 단일모형에서 놓칠 수 있는 작은 패턴 반영이 가능한 장점이 있다(Rokach, 2010). 또한 앙상블 학습은 머신러닝 분석 대회에서 가장 많은 우승을 차지한 바 있으며, 회귀, 분류, 순위 문제를 해결할 수 있다(Breiman and Cutler, 2014; Siroky, 2009).

최근 제안된 앙상블 학습은 약한 학습(Weak Leaner)을 활용하여 강건한 학습(Strong Leaner)을 만드는 것으로 적당한 예측을 가진 단일 모형을 이용하기보다 다양한 모형을 조합하는 것이 더 좋은 결과를 도출할 수 있다는 것을 가정한다. 앙상블 학습은 분석과정에 따라 의사결정나무, 배깅(Bagging), 랜덤 포레스트(Random Forest), 부스팅(Boosting), 경사부스팅(Gradient Boosting), 스태킹(Stacking) 기법 등으로 구분할 수 있다.

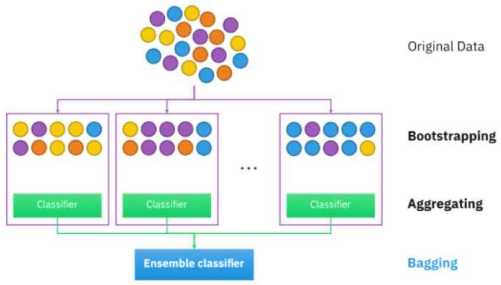
첫째, 배깅(Bagging) 방법은 Breiman(1996)에 의해 개발된 분류 앙상블 방법이다. 배깅은 전체데이터에서 일부데이터를 복원 추출한 후 부스트랩(Bootstrap)을 통해 N개의 부분 데이터 집합(Subset data)을 만들어 내고, 이러한 부분 데이터 집합에 모형을 N번 적합하여 예측값을 각각 계산한다. 마지막으로 이러한 예측 평균값을 최종 예측치를 정하게 된다.

둘째, 부스팅(Boosting)은 이전 모형 학습이 다음 모형 학습으로 이루어지면서 순차적으로 학습한다. 이 같은 방식으로 학습데이터의 샘플 가중치를 조정해 학습을 진행하는 것이 특징이라 할 수 있다. 즉. 최초의 모형을 구성한 후, 종속변수가 아닌 잔차를 업데이트하는 방식으로 모형

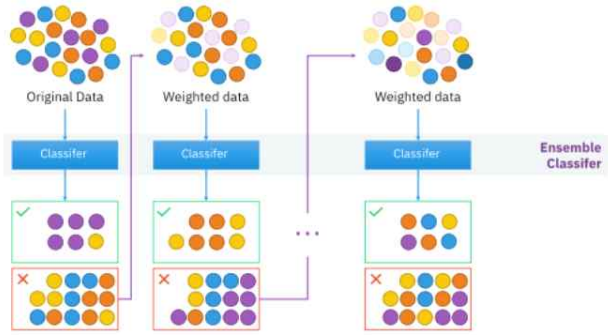
을 수정하게 된다.

셋째, 스태킹(Stacking)은 배깅과 부스팅 기법과 다른 접근법을 가지고 있다. 기본적으로 스태킹은 개별모형(Base Learner)이 예측한 데이터를 다시 최종메타 데이터세트로 사용해 학습을 수행한다. 즉, 개별모형(Base Learner)에서 예측한 결과를 다시 한번 최종 메타데이터 세트로 나누어 최종메타모형(Meta Learner)을 통해 예측결과를 제시한다.

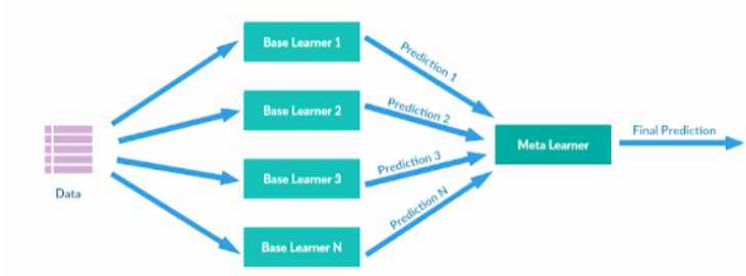
<그림 4-1> 앙상블 학습 발전과정



<그림 4-2> 배깅(Bagging) 개념[1)]



<그림 4-3> 부스팅(Boosting) 개념 [2)]



<그림 4-4> 스태킹(Stacking) 개념[3)]

1) https://en.wikipedia.org/wiki/Bootstrap_aggregating)
2) https://en.wikipedia.org/wiki/Boosting_(machine_learning)
3) https://subinium.github.io/introduction-to-ensemble-1

1. 통계기법

1.1 선형회귀분석

선형회귀 분석(Linear Regression Analysis)은 변수들 사이의 상관관계를 분석하는 데 사용하는 통계학적 방법이다. 선형회귀 분석은 독립변수 x, 종속변수 y, 상수항 c 사이의 관계를 모델링 하는 방법이다. 두 변수 사이의 관계일 경우 단순 선형회귀(Simple Linear Regression)라고 하며, 여러 개의 변수를 다루는 다중 선형회귀(Multiple Linear Regression)도 있다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \qquad \text{(식 4-1)}$$

여기서,

y : 종속변수

$\beta_1, \beta_2, \beta_k$: x_1, x_2, x_k 독립변수의 계수

u : 오차항

이 회귀분석을 사용할 때 종속변수와 독립변수 관계를 설정하는 문제가 가장 중요할 것이다. 다수의 연구에서는 독립변수의 선택 문제를 단계적 회귀분석법(stepwise regression)을 활용하여 해결하려 노력해 왔다 그러나 단계적 회귀분석법은 변수 선택에 있어서 제한이 있는데, 구체적으로는 변수 설정 및 기각 유의수준을 통제하기 어렵고 독립변수간의 상관계수가 높을수록 변수 설정이 어렵다는 점 등 여러 제한사항이 발생한다(Hocking, 1976; Berk. K, 1978). 이 때, 단계적 회귀분석법의 대안으로 제시된 부분 회귀분석법(all subsets regression)은 선택 가능한 모든 회귀모형을 사전에 선정한 척도로 가장 좋은 모형을 선택하게 하는 방법으로 단계적 회귀분석법보다 높은 신뢰성이 있는 것으로 알려져 있다.

1.2 의사결정 나무

의사결정나무(Decision Tree)는 의사결정규칙(decision rule)을 도식화시켜 대상 집단을 각각의 소집단으로 분류(classification) 또는 예측(prediction)하는 분석 방식이다. 분석과정이 복잡하지 않고 단순하여 쉽게 설명이 가능하다는 장점이 있다(최종후 외, 1998).

모형 구축 시의 빠른 수행능력과 결과에 대한 해석의 용이함으로 분류문제에 많이 활용되는 데이터 마이닝 기법이다. 의사결정 나무 알고리즘은 데이터 집합을 점점 작은 부분집합으로 분리해 나가는데, 더이상 분리가 되지 않는 단계에 최종 결과치들이 포함된다. 그리고 분리기준과 정지규칙, 가지치기에서 서로 다른 형성기준으로 조합하여 여러 알고리즘으로 분류된다. 알고리즘으로는 CHAID, QUEST, CART 등 대표적으로 세 가지 알고리즘이 있다. CHAID 알고리즘은 목표변수가 이산형인 경우와 연속형일 경우가 있으며, 이산형일 경우 카이제곱 검정통계량을 사용하고, 연속형이면 F-검정을 이용하여 다지 분리(multiway split)를 수행하게 된다. QUEST 알고리즘은 명목형 변수에 국한되어, 카이제곱 검정통계량을 사용하며, 예측변수의 측도에 따라 서로 다른 분리 기준을 사용한다(박용우, 2016). 이와 같은 대표적인 알고리즘을 <표 4-2>와 같이 정리하였다.

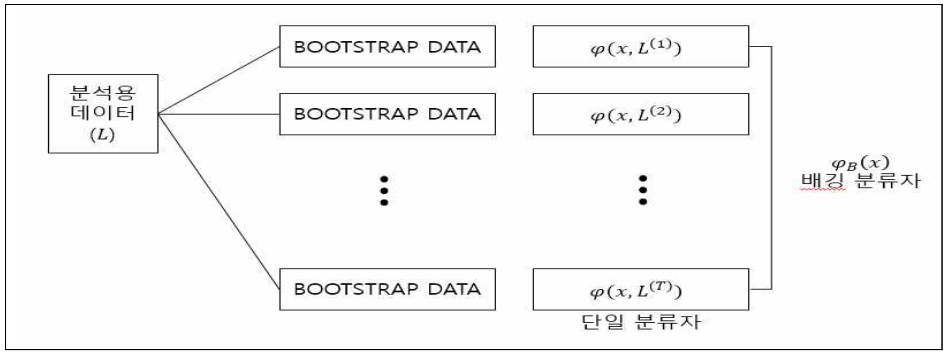
<표 4-2> 의사결정나무 알고리즘

구분	CHAID	QUEST	CART
목표변수	명목형, 순서형, 연속형	명목형, 순서형, 연속형	명목형
예측변수	명목형, 순서형, 연속형	명목형, 순서형, 연속형	명목형, 순서형, 연속형
분리기준	카이제곱-검정, F-검정	지니지수, 분산 감소	카이제곱-검정, F-검정
가지치기	알고리즘 미포함	알고리즘 포함	알고리즘 포함
결손값의 대체규칙	알고리즘 미포함	알고리즘 포함	알고리즘 포함

2. 배깅

배깅은 붓스트랩 aggregating의 약어로 다양한 모형을 구축하기 위해 데이터를 재구성한다. 이때, 동일 데이터를 반복복원추출을 통해 다양한 데이터세트를 만들고, 이 다양한 데이터세트를 각각 학습시켜서 평균(voting)을 통해 결과를 도출하는 것이다. 배깅 알고리즘은 불안정한 분류방법의 예측력을 상당히 향상시키는 것으로 알려져 있으며, 목표변수의 형태에 따라 분류분석과 회귀분석에도 사용할 수 있다.

일반적으로 나무 구조모형(Tree Structure Model)은 편향이 작으나 분산이 크기 때문에 안정성이 떨어지는 경향이 있어서 깊이가 깊게 성장한 모형은 훈련용 데이터 집합에 대해 과적합 되는 경향이 있다. 배깅은 데이터에 따라 노드에서 선택되는 변수와 분리 값이 쉽게 변경되는 이와 같은 불안정성을 해결하고자 제안된 알고리즘으로 나무 구조모형을 독립적으로 생성해 낮은 예측모형을 결합해서 강한 예측모형을 생성하고, 분산을 줄이고자 하는 것이 목적이다. 배깅이 불안정한 모형의 예측 수행력을 향상하긴 하지만, 붓스트랩 표본의 수가 증가할수록 메모리 사용량이 증가하게 되고 배깅을 하지 않은 모형에 비교해 해석력이 떨어지는 단점이 있다(Breiman, 1996).



자료 : 이영섭 외, (2005)

<그림 4-5> 배깅(Bagging) 알고리즘 과정

배깅 알고리즘 과정은 <그림 4-5>와 같다. 모집단으로부터 추출된 학습 데이터세트(training data set) L에서 단순 복원 임의추출에 의해 부스트랩(Booststrap) 분석용 데이터를 추출한다. 이와 같은 방법을 T번 반복하여 T개의 부스트랩 분석용 데이터를 생성한다. 각각의 부스트랩 분석할 데이터에 적절한 분류 알고리즘를 적용한 단일분류자 $\phi(x, L^t)$를 형성하여 T개의 단일 분류자 집합 $\phi(x, L^t)$, $(t = 1,2,\cdots,T)$을 얻는다. 이런 단일 분류자를 결합하는 방식에는 목표 변수가 연속형, 범주형이냐에 따라 구분된다. 연속형인 경우 평균(average), 범주형이면 다중 투표(majority vote)를 활용하게 된다. 이와 같이 형성되어진 분류자를 배깅 분류자($\phi_B(x)$)라 한다(이영섭 외, 2005).

Breiman(1996)에 의하면 분석용 데이터가 불안정(unstable)하면, 배깅 분류자의 결합을 통해서 분류 성능이 향상된다. 반면 분석용 데이터가 안정적(stable)인 경우에는 배깅 과정을 통해서 얻어진 배깅 분류자는 분석용 데이터에서 얻어진 단일 분류자와 비슷하다.

배깅의 알고리즘을 구체적으로 설명하면 <표 4-3>과 같다.

<표 4-3> 배깅(Bagging) 알고리즘 단계별 과정

각 과정별 단계	
1단계	훈련용 데이터 집합의 표본을 $L = (x_i, y_i)^{n_{i=1}}$라고 하고, K개의 부스트랩 자료 $L^{*(k)}$, $k = 1,2,\cdots,K$를 생성한다.
2단계	각 부스트랩 자료 $L^{*(k)}$에 대한 모형 $f^{(k)}(x)$를 생성한다.
3단계	K개의 예측모형을 결합하여 최종 모형 $\hat{f}$를 만든다. 최종 모형을 만드는 방법은 아래와 같다. 3-1) 회귀모형은 $\widehat{f(x)} = \sum_{k=1}^{K} \frac{f^{(k)}(x)}{K}$ 와 같이 평균을 취한다. 3-2) 분류모형은 $\widehat{f(x)} = \arg max_y \sum_{k=1}^{K} I(f^{(k)}(x) = y)$와 같이 다수결로 예측한다.

의사결정나무는 분류(Classification)와 회귀(Regression) 모형으로 구분된다. 이 모형은 연구대상을 분류와 예측, 그리고 순차적 결정할 문제들에서 의사결정을 용이하게 해주는 강력한 방법이다. 하나의 의사결정 트리가 분류 목적을 위해 사용될 때 분류트리, 회귀를 목적으로 사용될 때, 회귀나무로 불린다(Timofeev, 2004).

의사결정 나무 알고리즘 중 회귀나무(Regression Tree)는 먼저 N개의 관측치별로 대해 자료가 p개의 입력변수 x와 1개 종속변수를 y로 구성하면, (x_i, y_i), $i = 1,2,\cdots,N$, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$와 같이 된다. 그 다음 변수들과 점들을 분리하고, 나무의 구조를 결정하는 알고리즘이 작동된다. 이러한 경우 한 개의 분할을 M 지역 $R_1, R_2, \cdots, R_m$으로 나누게 된다면, 각 지역에서 상수 c_m 을 산정하고 아래와 같은 반응(response)에 대한 식을 가진다(이재득, 2021).

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \qquad \text{(식 4-2)}$$

전체적인 과정 중 나무의 크기가 너무 크면 오버피팅(Overfitting)을 가지고 나무의 크기가 너무 작으면 중요한 성질을 놓치는 언더피팅(Underfitting)을 가진다. 따라서 모형의 복잡도를 조절하는 나무의 최적 크기를 결정하는 주요 요인은 최소 절점 크기를 가진 큰 나무(T_0)의 부분집합인 나무 $_T(T \subset T_0)$를 구하는 것이다. 여기서 터미널 노드는 m, 영역을 m에서 R_m로 나타내고, $|T|$를 T에서의 터미널 노드 개수를 표현하고, 아래와 같은 산출식으로 나타낼 수 있다.

$$N_m = Number\left[x_i \in R_m\right] \qquad \text{(식 4-3)}$$

$$\hat{c_m} = \frac{1}{N_m} \sum_{x_i \in R_m} y_i \qquad \text{(식 4-4)}$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \widehat{c_m})^2 \quad \text{(식 4-5)}$$

여기서 $Q_m(T)$는 노드 불순도(squared-error node impurity)를 측정한다. 그리고 비용 복잡성(cost complexity)의 기준을 다음과 같이 설정한다.

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T| \quad \text{(식 4-6)}$$

여기서 α는 모형의 복잡성을 조절하기 모수를 튜닝하면서 나무의 크기와 자료의 적합성 간의 상충관계(trade-off)를 조정한다. 그리하여 각 α에 대하여 $C_\alpha(T)$를 최소화시키는 하위 나무(subtree) $T_\alpha(T_\alpha \subseteq T_0)$를 구한다.

하지만 회귀나무 지수는 나무의 크기에 따라 민감도가 낮은 단점을 보완하고자 분류나무(Classification Tree) 모형을 사용한다. 이런 민감도가 높은 모형의 불순도 측정 지수로써 대표되는 두 가지 지수가 존재한다. 지니 지수(Gini Index)와 엔트로피 지수(Entrophy Index)이다. 먼저 노드 m, 영역 R_m, 그리고 N_m개의 관측치가 있을 때, 노트 m에서 k클래스의 관측치들의 비율을 $\widehat{p_{mk}}(0 \le \widehat{p_{mk}} \le 1)$라고 설정하면, 지니 지수(G)와 엔트로피 지수(Entrophy Index)는 다음과 같이 정의된다(이재득, 2021).

$$\widehat{p_{mk}} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) \quad \text{(식 4-7)}$$

$$\text{Gini Index} : G = \sum_{k=1}^{K} \widehat{p_{mk}}(1 - \widehat{p_{mk}}) \quad \text{(식 4-8)}$$

$$\text{Entrophy Index} : D = -\sum_{k=1}^{K} \widehat{p_{mk}} \log(\widehat{p_{mk}}) \quad \text{(식 4-9)}$$

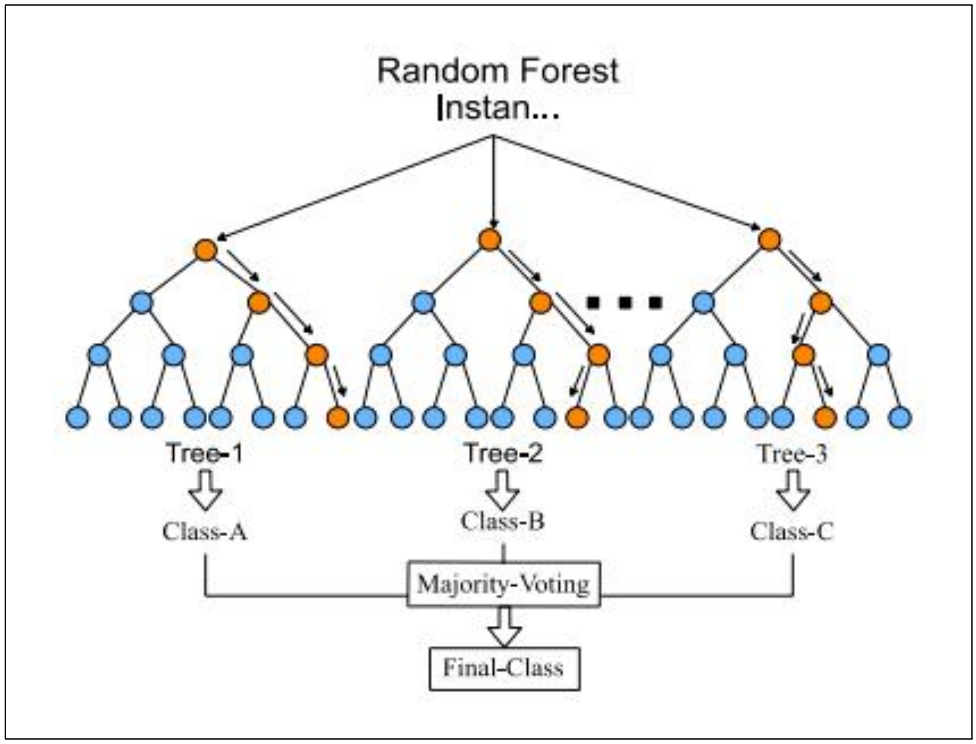
여기서 $\widehat{p_{mk}}(0 \leq \widehat{p_{mk}} \leq 1)$이므로, 엔트로피 지수(D)≥0이다. 그리고 $\widehat{p_{mk}}$가 모두 0 혹은 1에 가까이 있으면, 엔트로피 지수는 0에 가까이 가게 될 것이다. 따라서 지니 지수와 엔트로피 지수는 모두 m차 노드에서 순수(pure)하다면 각각 작은 값을 가지게 된다.

2.1 랜덤포레스트

랜덤 포레스트(Random forest)는 배깅과 의사결정나무 기법을 차용한 앙상블 학습 중 하나이다. 배깅은 분류 모형의 학습 오류인 편향(bias)와 분산(variance)의 균형을 적절히 조절하기 위해 사용되는 기법이다. 편향은 분류 모형이 예측치와 실제치의 차이 정도를 나타내는 지표이고, 분산은 특정 학습데이터에만 정확한 예측 성능을 보이며, 그 나머지 데이터는 잘 분류하지 못하는 정도를 나타내는 지표이다. 편향과 분산은 상호 상충관계(trade off)에 있으며, 예측 정확도를 높이기 위해 모형을 과도하게 학습시킨다면 데이터의 분산이 높아지고, 분산을 줄이기 위해 다양한 데이터에 안정적으로 학습을 시행한다면 편향이 높아지게 된다. 이때, 의사결정나무 기법이 분산이 크다는 한계점이 있기에 이 점을 보완하고자 만들어진 것이 랜덤포레스트(Random Forest)이다. 이와 같은 문제점을 보완할 때 사용하는 기법이 바로 배깅인 것이다. 배깅 기법은 각 의사결정트리의 예측치 평균으로 편향을 낮게 유지하며, 분산은 중심극한정리(central limit theorem)에 의해 낮아지게 한다. 이때 중심극한정리는 모집단의 분포와 상관없이 표본의 크기가 큰 경우 표본 평균이 정규분포를 따른다는 특징이 있다(안경민, 2021).

Breiman(2001)이 고안한 랜덤 포레스트(Random Forest)는 결국 편향이 적은 다수의 의사결정나무 예측 결과 중 다수의 결과를 선택하여 분산과 편향이 작아지는 학습 결과를 도출하는 방법이라고 할 수 있다. 또한, 의사결정나무 기법과 달리 랜덤 포레스트에 대한 이론설명 및 결과에 대한 해석은 어렵다는 단점이 있지만 예측에 있어서는 매우 높은 정확도를 보이는 방법으로 알려져 있다. 특히나 변수의 개수가 많을 경우

배깅 혹은 부스팅과 유사하거나 더 높은 성능을 보이며, 조율모수의 부재로 실제 자료분석에 대한 접근성이 높은 편이다(Mohammad et al., 2021).



자료: Chen et. al., (2016)

<그림 4-6> 랜덤포레스트 개념

랜덤 포레스트는 크게 두 가지 장점이 있다. 과적합(Overfitting) 문제와 높은 예측력이다. Breiman and Cutler (2014)에 의하면 트리의 수가 많으면 예측 오차가 줄어 과적합 현상이 최소화 될 수 있으며, 높은 예측력과 안정적인 모형 형태가 가능하다고 한다. 그리고 독립변수의 수와 형태에 상관 없이 활용 가능하므로 방대한 자료로부터 핵심 변수를 찾는데 용이함을 보인다(김성진·안현철, 2016). 알고리즘은 아래 <표 4-4>와 같다(Siroky, 2009).

<표 4-4> 랜덤포레스트(Random Forest) 알고리즘 단계별 과정

각 과정별 단계	
1단계	For B_i, $i = 1,\ldots,B$ a. 학습용 데이터 사이즈가 N인 붓스트랩 데이터 샘플 S를 추출 b. 하위 알고리즘에 따라 최소 터미널 절점 크기 $n_{\min}$을 구할 때까지 붓스트랩된 데이터를 사용한 T_b를 구한다. 전체 P변수 집합에서 무작위로 p 변수를 선택하고, 이 변수 중 최적의 변수/분할점 선택 절점를 두 개의 daughter nodes로 분할시킨다.
2단계	앙상블 나무의 최종 산출물은 다음과 같으며, $(T_b)_1^B$
3단계	새로운 관측치 혹은 out-of-bag 관측치를 예측한다. 회귀문제인 경우, $\widehat{f_{rf}^B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$ 분류문제인 경우, $\widehat{C}_b(x)$를 b번째 트리의 클래스 예측이라고 하면, $\widehat{C_{rf}^B}(x)$는 다수결의 선택에 의해 $(\widehat{C}_b(x))_1^B$ 와 같이 나온다.

3. 부스팅

Boosting은 Bagging과 다르게 일반적인 모형에 집중되어 있지 않고, 맞추기 어려운 문제를 맞추는데 초점이 맞춰 있다. 배깅은 독립적인 데이터를 가지고 복원추출을 통해 독립적으로 예측하지만, 부스팅은 이전 모형이 다음 모형에 영향을 준다. 즉, 이전 모형에서 가장 큰 오차를 보인 관측치들이 나타날 확률이 높도록 가중치를 부여한다. 이런 과정을 거치면 관측치의 오차를 더 잘 맞출 수 있도록 하는 알고리즘이라 할 수 있다. Boosting기법은 Adaboost, Gradient Boosting Model(GBM), XGBoost, LightGBM, CatBoost으로 발전되고 있다.

3.1 AdaBoost

Freund and Schapire(1999)에 의해 개발된 앙상블 방법이다. 약한 분류기(weak classifier)들은 한 번의 과정을 하나씩 순서대로 학습을 진행한다. 먼저 학습된 분류기는 정분류 데이터와 오분류 데이터로 구분하여 다음 분류기에 전달한다. 이전 분류기로부터 받은 정보를 활용하는 다음 분류기는 오분류 데이터들의 가중치(weight)를 높이는데, 이는 이전 분류기가 상호보완적으로 오분류된 샘플의 가중치를 변경하면서 잘못 분류되는 데이터에 집중학습이 될 수 있도록 한다. 최종 분류기(strong classifier)는 약한 분류기들에 다른 가중치를 적용, 조합하여 학습을 진행한다. 여기서 부스팅에 사용되는 분류기는 조금이라 예측력이 높은 모형 구축에 효과가 있는 것으로 알려져 있다. 이는 예측력이 약한 모형을 결합하여 예측력이 강한 모형을 구축하는 과정으로, 다양한 분야에 적용되는 알고리즘이다.

배깅이 데이터 집합에서 단순히 표본을 뽑아 각 모형에 적용하는 방법이라면, 에이다부스트(AdaBoost) 방법은 이전 모형들이 예측하지 못한 오류 데이터에 가중치를 부여하여 다음 모형이 더 잘 예측되도록 하는 방법이다. 부스팅은 임의 분류보다 조금 향상된 약한 예측모형을 합쳐서 오분류로 인한 오차율을 줄이는 앙상블 방법의 아이디어를 사용하였다(Valiant, 1984; Kearns and Valiant, 1989).

부스팅의 트리는 이전 트리에 영향을 받고, 최소 깊이로 만들어질 뿐만 아니라, 최종 모형에 반영되는 크기 또한 다르게 된다. 이때 부스팅의 변수 중요도는 오차 제곱 감소 정도를 의미하는데, 특히 각 결과변수로부터 만들어지는 오차 제곱값이 감소하는 것은 앙상블에서 각 트리로부터 향상된 값이 나오게 된다. 각 결과변수의 향상도는 전체 앙상블에 대해 평균을 낸 후 전체 중요도 값으로 사용한다. 부스팅은 다른 앙상블 방법들과 마찬가지로 예측모형 결과에 대한 해석이 불가능하다는 단점을 가지고 있지만, 목적이 예측모형의 수행력을 높이고자 한다면 적절한 방

법이 될 수 있다(Friedman, 2001).

<표 4-5>에서 주의 깊게 살펴볼 부분은 2단계에서의 2-3과 2-4 과정이다. 부스팅에 사용되는 예측모형 G_m은 랜덤한 추출보다 조금 더 좋은 예측력을 갖는다고 가정하면, G_m의 (가중) 오분류율은 0.5보다 작게 되므로 2-3) 과정의 $\alpha_m > 0$이 된다(Schapire, 1990).

<표 4-5> 에이다 부스트(AdaBoost) 알고리즘 단계별 과정

각 과정별 단계	
1단계	초기 개별 관측치들의 가중치로 $w_i = \frac{1}{N}, i = 1,2,\cdots,N$를 부여한다.
2단계	아래 과정을 M $(m = 1,2,\cdots,M)$회 반복한다. 2-1) 가중치 w_i를 사용한 표본으로 약한 예측모형 $G_m(x)$를 생성한다. 2-2) 생성된 약한 예측모형의 오분류율을 다음과 같이 정의한다. $$err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$$ 2-3) 이때 개별 예측모형의 가중치를 다음과 같이 정의한다. $$\alpha_m = \log((1-err_m)/err_m)$$ 2-4) 오분류된 표본에 가중치를 추가하고, 제대로 예측된 모형에는 가중치를 줄이는 방식으로 가중치를 갱신한다. 갱신한 가중치는 다음과 같다. $$w_i e^{[\alpha_m I(y_i \neq G_m(x_i))]}, \ i = 1,2,\cdots,N$$
3단계	*i번째 각 표본의 기본값을 I번째 모형의 예측값과 곱한 후, 이 값을 k개 전반에 걸쳐 통합하여 부스팅 된 모형의 예측값 $G(x)$를 구한다.* $$G(x) = sign[\sum_{m=1}^{M} \alpha_m G_m(x)]$$

그러면 2-4) 과정에서 각 관측치에 할당되는 가중치가 G_m에 의해서 오분류된 관측치를 증가시켜 정분류된 관측치를 기존값과 같게 만든다. 가중치를 정규화시켜 총합이 1이 되면, AdaBoost 알고리즘이 매번 반복

하면서 오분류된 관측치의 가중치는 증가시키고, 정분류된 관측치는 감소시키면서 예측모형을 구축해 간다.

3.2 Gradient Boosting

Friedman(2001)은 분류와 회귀를 포괄하는 방법으로 기울기 강하(gradient descent) 알고리즘을 부스팅 알고리즘으로 사용하는 방법을 제안하였다. 즉, 주어진 손실함수와 약한 예측모형에 대해서 손실함수를 최소화할 수 있는 가법모형을 찾고, 결과변수를 가장 잘 예측할 수 있는 식으로 초기화한다. 이후 경사 또는 잔차를 구하고 손실함수를 최소화할 수 있는 잔차에 맞는 모형을 구축한다. 이렇게 구해진 모형을 이전 모형에 더하고, 이 과정을 반복 시행한다. 이처럼 경사 부스팅(gradient boosting)은 부스팅 알고리즘이 일반화된 것으로 오류에 초점을 맞춰 모형을 학습하는 앙상블 방법이다(Friedman, 2001).

경사 부스팅은 사건 확률 $\hat{p_i} = \frac{1}{1+\exp[-f(x)]}$를 이용하여 모형화를 함에 있어서, $f(x)$는 $[-\infty, \infty]$ 범위의 모형 예측식을 의미한다. 경사 부스팅의 알고리즘을 요약하면 <표 4-6>와 같다(Friedman, 2001).

<표 4-6> 경사부스팅(Gradient Boosting) 알고리즘 단계별 과정

각 과정별 단계	
1단계	깊이 D와 반복횟수 K선택한다.
2단계	응답 값의 평균 y를 구한 후, 이를 각 표본에 대한 기본 예측값으로 사용한다.
3단계	아래 과정들을 $K(k=1,2,\cdots,K)$회 반복한다.
	3-1) 각 표본에 대해 관측값과 현재 예측값의 차이인 잔차를 구한다.
	3-2) 잔차값을 응답값으로 사용한 후, 회귀 트리를 깊이 D에 맞춘다.
	3-3) 앞에 맞춘 회귀 트리를 사용해 각 표본에 대해 예측값 구한다.
	3-4) 앞에 만든 예측값과 앞 반복과정에서 만들어진 예측값을 더해 각 표본의 예측값을 수정한다.

경사부스팅과 배깅의 차이점은 배깅의 경우 분류기들이 상호영향을 주지 않지만 부스팅은 이전 분류기의 학습 결과를 기반으로 샘플링된 다음 분류기 데이터의 가중치를 조정한다는 것이다. 부스팅 알고리즘으로써 처음으로 개발된 AdaBoost 알고리즘은 가중선형결합에 의해 최종 분류기를 설정하는 알고리즘이다. AdaBoost 알고리즘은 초기에는 모두 동일한 확률로 복원추출을 하지만 매번 반복마다 잘못 분류 관측치의 가중치는 증가시키고 제대로 분류된 가중치는 감소시키면서 예측모형을 만들어 간다.

경사 부스팅은 널리 사용되는 방법이며 이분형 자료 또는 연속형 자료에서 예측력이 우수한 강력한 방법이다. 하지만, 매개변수를 잘 조정해야 한다는 것과 훈련시간이 오래 걸린다는 것이 단점이다. 또한, 트리 기반 모형의 특성상 희소한 고차원 데이터에서는 예측력이 떨어지는 경향이 있다. 경사 부스팅의 중요한 매개변수는 나무모형의 개수와 이전 나무모형의 오차를 바로 잡는 학습률을 조절하는 것이다. 이 두 매개변수는 매우 밀접하게 연관되어 있고 트리의 개수를 크게 하면 모형이 복잡해지고 과적합될 가능성이 증가한다. 따라서 보통 경사 부스팅에서는 최대 깊이를 5보다 작게 해서 각 나무 모형의 복잡도를 낮추는 경향이 있다. 또한 부스팅 알고리즘 특성상 오류분류율을 계속 보완하려고 하므로 이상치에 민감할 수도 있다(Muller et al., 2016).

3.3 XGBoost

Chen and Guestrin(2016)이 개발한 XGBoost(eXtreme-Graditent-Boosting)의 기본형은 그래디언트 부스팅(gradient boosting)으로(Friedman, 2001; Friedman, 2002), 기존의 그래디언트 부스팅의 느린 속도와 오버피팅(Overfitting) 규제 부재와 같은 단점을 보완하여 빠른 학습 속도와 탁월한 성능을 선보이는 분류 모형이다. XGBoost는 그래디언트 부스팅과 마찬가지로 하이퍼파라미터들을 설정할 수 있다. 지도학습으로 독립변수를 학습하여 결과변수의 값을 예측하는 방법이며, 대부분 작은 학습률로 많은 수의 나무모형을 결합하는 것이기 때문에 초기에 추

가된 나무모형이 중요하다. 사실 나무모형 학습에서 가장 많은 시간이 필요한 부분은 정렬된 순서로 데이터를 가져오는 것이다. 전체 데이터 집합을 단일 구역에 저장하고 미리 정렬된 항목을 선형 모형으로 만들어 분할 검색 알고리즘을 실행하게 되는데, XGBoost과적합을 방지하기 위해 정규화된 모형을 통합하고 결측치를 자동 처리한다. 경사 부스팅에서 나무모형 가지치기는 음의 손실이 발생하면 멈추지만, XGBoost는 최대 깊이까지 진행한 후에 손실함수가 일정 값에 도달하지 못하면 역방향으로 가지치기를 진행한다. 또한, 다른 방법들과 달리 XGBoost는 범주형 변수를 자체적으로 처리할 수 없으므로 단일핫인코딩(one-hot encoding), 레이블인코딩(label encoding), 평균인코딩(mean encoding)과 같은 다양한 인코딩을 수행해야 한다.

XGBoost 알고리즘은 훈련용 손실을 최소화하면서 과적합을 축소하기 위해 나무모형의 복잡도를 조절하는 방식으로 최적화된 모형을 구축하는 과정을 사용한다. XGBoost의 목적함수는 다음과 같이 설정한다.

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \quad \text{(식 4-10)}$$

여기서 K는 나무모형의 수, 그리고 Ω는 나무모형의 복잡도에 영향을 줄 수 있는 모든 상황을 포함하는 함수이다. 깊이가 0인 나무모형에서 시작해서, 가지치기를 시행하였을 때 새로 습득하는 정보가 많아질수록 나무모형의 깊이는 더 깊어지게 된다.

XGBoost의 목표는 수식에서 목적 함수를 최소화할 t번째 최소화할 f_t을 찾는 것이다. 형태는 트리별로 트리의 구조와 잎(leaf) 점수를 가진 f의 형태로써 각 단계에서 트리의 손실을 계산하며, 단계별 손실 확인이 가능하다. 다음과 같은 식으로 표현 가능하다(김종성 외, 2019).

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)}) + \sum_{k=1}^{K} \Omega(f_k) \quad \text{(식 4-11)}$$

새로 습득된 정보는 왼쪽 자식 노드점수와 오른쪽 자식 노드점수를 합

한 값에서 미분 시 점수를 뺀 값의 절반에서 복잡도 비용을 제외한 값을 의미하며, 다음과 같이 정의된다.

$$Gain = \frac{1}{2}[\frac{G_L^2}{H_L-\lambda}+\frac{G_R^2}{H_R-\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}]-\gamma \quad \text{(식 4-12)}$$

새로 습득된 정보가 최대가 되도록 가지를 치는 나무모형을 생성한 후, 마지막으로 점수가 높은 나무모형을 조합하여 부스팅을 함으로써 K 개의 나무모형이 조합된 최적의 예측모형을 구축한다(Chen et al., 2016).

3.4 LightGBM

LightGBM(Light Gradient Boosting Machine)은 Microsoft에서 개발한 모형으로 잎 분할(Leaf-Wise) 방식을 활용하여 정확도를 높여 XGBoost의 한계점을 보완하고자 만든 모형이다(Wang and Wang, 2020). LightGBM은 말 그대로 “Light” 가벼움 뜻하며, 빅데이터를 다룰 수 있고 실행시킬 때 메모리를 적게 차지하므로 처리 속도가 빠른 장점이 있다. 무엇보다도 LightGBM이 인기가 있는 이유는 결과의 정확도에 초점을 맞추기 때문이다.

그 비결은 <그림 4-7>과 같이 XGBoost 모형 알고리즘과 비교해보면 쉽게 이해할 수 있다. XGBoost 모형은 레벨 분할방식(Level Wise, Depth-first)을 사용하여 의사결정트리의 노드들이 뿌리 노드와 가까운 노드를 우선 순회하고 수평으로 확장하는 형태이다. 반면 LightGBM은 최대 델타 손실(Max Delta Loss)이 큰 노드에서 분할하여 수직 성장하는 잎 분할방식을 이용한다. 이때 수직으로 발전되어 동일한 잎을 성장시킬 때 더 큰 손실을 줄일 수 있다는 것을 가정한다. 이로써 정확도와 속도가 뛰어난 성능을 보여주는 알고리즘이라 할 수 있다(안경민, 2021).

(a) XGBoosting (b) LightGBM

<그림 4-7> Boosting과 LightGBM 알고리즘 비교(Wang and Wang, 2020)

LightGBM의 도출 수식은 아래과 같다. 지도학습세트가 $X=(x_i, y_j)_{i=1}^{n}$ 이라고 할 때, 다음과 같은 특성 손실함수 $L(y, f(x))$의 기대값을 최소화하는 특정 함수 $f^*(x)$에 대한 근사치 $\hat{f}(x)$을 찾는 것을 목표로 한다(Ke et al., 2017).

$$\hat{f}=\arg\min E_y,\ {}_X L(Y,f(x)) \qquad \text{(식 4-13)}$$

위와 같은 기본식을 기본으로 알고리즘 별 각 단계를 거쳐 마지막으로, 분할을 추가한 후의 식을 최종 목적으로 한다.

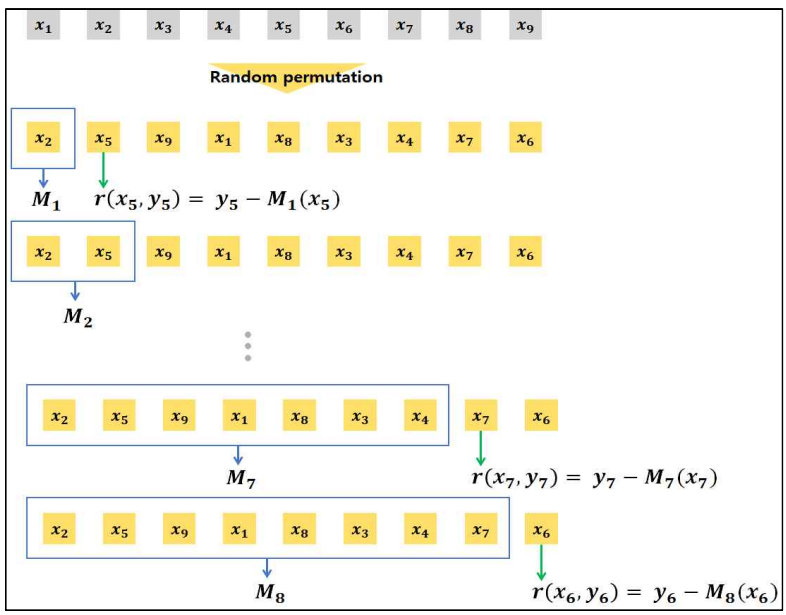
다음 <표 4-7>은 LightGBM 알고리즘의 과정이다(안경민, 2021). 마지막 3단계 과정에서 I_L과 I_R은 각각 왼쪽과 오른쪽 가지의 표본 집합이다. 즉 LightGBM은 트리를 수직으로 확장하여 성장시켜 많은 데이터와 기능을 처리하는데 탁월한 방법이다.

<표 4-7> LightGBM 알고리즘 단계별 과정

각 과정별 단계	
1단계	$\hat{f}=\arg\min E_y,\ {}_X L(Y,f(x))$을 기본식으로 다수의 T 회귀트리를 통합하여 최종 모형에 근사치를 부여한다. $f_T(X)=\sum_{t=1}^{T} f_t(X)$
2단계	2-1) t단계를 거치며 학습시키며, $\Gamma=\sum_{t=1}^{n} L(y_i, F_{t-1}(x_i)+f_t(x_i))$ 2-2) Newton's method 활용해 근사치를 도출시킨다. $\Gamma_t \cong \sum_{t=1}^{n}(g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i))$ $\Gamma_t \cong \sum_{j=1}^{j}(\sum_{i\in I_j} g_i)w_j)+\frac{1}{2}(\sum_{i\in I_j} h_i+\lambda)w_j^2)$
3단계	분할을 추간 후의 목표기능 함수는 다음과 같다. $G=\frac{1}{2}(\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i+\lambda}+\frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i+\lambda}+\frac{(\sum_{i\in I_I} g_i)^2}{\sum_{i\in I_I} h_i+\lambda})$

3.5 CatBoost

CatBoost(Category Boosting)는 범주형 변수를 처리하는 데 높은 성능을 보유하고 있는 알고리즘이다(Prokhorenkova et al., 2018). 이전에 살펴본 Gradient Boosting은 부스팅의 진행마다 학습데이터가 전체 테스트데이터의 잔차로 나타나기 때문에 학습데이터가 테스트데이터로 수렴하는 분포로 변화하게 된다. 이런 분포 변화는 오버피팅(Overfitting) 또는 예측 변화(Prediction Shift)가 나타나는 문제가 발생한다. 또한 이전 범주형 변수는 추가되는 이진변수가 있는데, 이때 수많은 이진변수가 추가되면 통계량 증가로 계산 시간(Computation Time)과 메모리 소모(Memory Consumption)가 증가하게 된다.



자료 : Prokhorenkova et. al,(2018) 그림 수정

<그림 4-8> 순차적 부스팅(Ordered boosting)의 원리

이러한 단점을 보완하고자 CatBoost는 Ordered Boosting 기법을 사용하여 범주형 변수 전처리와 오버피팅(Overfitting) 문제를 보완해준다. Ordered Boosting은 기존의 부스팅이 모든 잔여 오차를 차례로 학습하

는 것과 달리 일부 데이터의 잔여 오차를 계산하여 모형을 구축하며, 그 뒤에 데이터의 잔차는 이 모형이 예측한 값을 다시 사용한다(안경민, 2021).

즉, Ordered Boosting에 Random Permutation을 통해 데이터 순서를 변경하여 오버피팅(Overfitting)을 방지한다(Prokhorenkova et. al, 2018). 다시 말해 범주형 변수 전처리를 위해서 CatBoost 알고리즘은 Random Permutation을 거친 데이터세트에서 같은 범주를 가진 변수들의 평균 표본 값을 계산하며, 이는 식 (4-14)와 같다(안경민, 2021).

$$\widehat{x_k^i} = \frac{\sum_{j=1}^{n} [x_j^i = x_k^j] \circ y_j + \alpha P}{\sum_{j=1}^{n} [x_j^i = x_k^j] \circ y_j + \alpha} \qquad \text{(식 4-14)}$$

순열이 $x_k = (x_k^1, \cdots, x_m^k)$ 인 경우,

α = 가중치(;이전 값의 가중치)$\alpha > 0$)

P = 사전확률(;이전 값)

이 방법은 오차를 최소화하는 것에 도움을 주며, 특히 범위가 작은 범주에서 성능이 월등하다. 또한 변수 조합(Feature Combinations)을 사용하여 동일한 정보형태를 가진 변수들을 하나로 묶어 트레이닝 속도를 향상시킬 수 있다. 그리고 초기 하이퍼 파라미터값이 최적화된 상태이기에 파라미터 튜닝 절차가 필요 없다는 이점이 있다(Prokhorenkova et al., 2018).

제4절 앙상블 학습 경유자동차 PM 예측모형 구축

1. 경유자동차 PM 예측 과정

본 연구는 경유차 PM 배출 예측과 요인분석이 주요 목적이다. 요인들의 영향력을 분석하기에는 머신러닝기법이 유리한 것으로 알려져 있다. 그 이유는 요인들의 설명력이 딥러닝기법보다 머신러닝기법이 더 뛰어나기 때문이다. 그리고 경유차 PM 예측은 회귀에 대한 문제이므로 머신러닝기법이 더 적합하다. 여기에 앙상블 학습을 적용하여 여러 가지 우수한 학습모형을 조합하면 예측력이 향상되고 단일모형에서 놓칠 수 있는 작은 패턴반영이 가능하다. 또한 분석데이터가 빅데이터이므로 과적합 문제를 해결하고 오차와 예측성을 높이기 위해 앙상블 학습을 적용한 경유자동차 PM 예측모형을 구축하고자 한다.

본 연구의 경유차 PM 예측모형 구축과정은 <그림 4-9>과 같이 4단계로 이루어진다. 첫째, 데이터 전처리과정이다. 데이터의 Null 값은 제외하고 기존 데이터의 변수들의 특성을 반영한 범주형 변수를 추가하였다. 데이터의 이상치 제거는 수학적 모형인 레버리지 분석(Leverage analysis) 방법론을 차용한다. 레버리지는 실제 실측치가 예측치에 미치는 영향력을 나타낸 값이다. 레버리지 분석을 통해 레버리지(H^*) 경계와 잔차(R) 경계 밖에 있는 이상치는 제거한다.

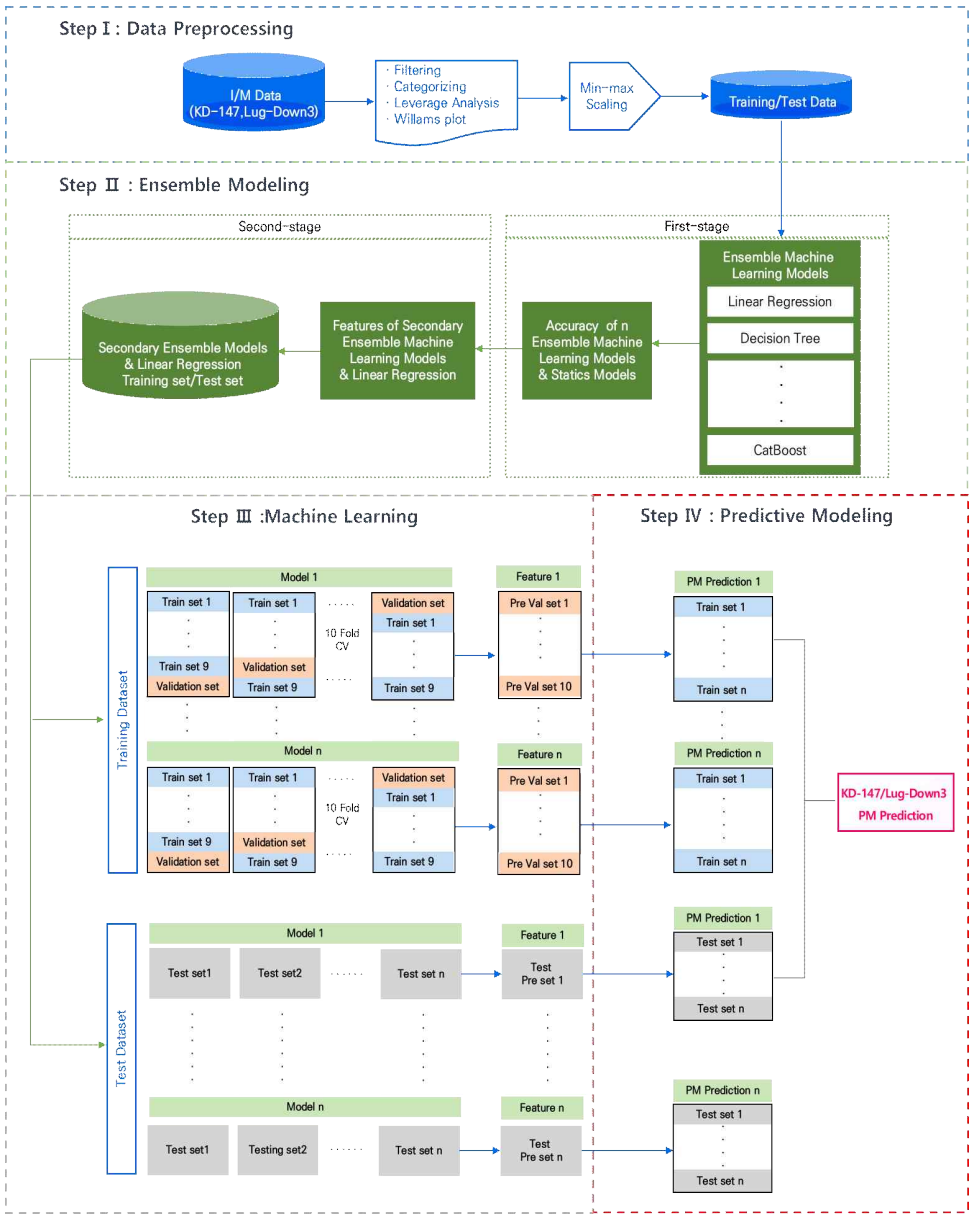
둘째, 다양한 앙상블 학습기반 예측모형을 구축한다. 1차 모형 구축은 Python Pycaret는 AutoML(Automated Machine Learning)을 지원하는 파이썬 라이브러리로서 여러개의 모형을 탐색하는 패키지을 활용하고자 한다. Pycaret에서는 분류(classification), 회귀(regression), 군집(clustering), 이상감지(anomaly detection), 자연어처리(NLP), 연관규칙(association rules)에 활용된다. 1차 앙상블 학습 기반 모형들은 신뢰성을 검증하기 위해 회귀, 배깅, 부스팅 알고리즘을 기반한 20개 예측모형을 구축한다. 여기서 예측력이 높은 상위 n개 모형과 통계기법, 배깅, 부

스팅을 대표하는 모형을 선정하여 2차 앙상블 학습 예측모형을 구축한다. 2차 앙상블 학습기반 모형들은 경유자동차 PM 배출 주요원인을 확인하고 경유자동차 PM 배출을 예측 및 모형의 성능을 비교 검증한다.

셋째, 학습 세트(Training Set)과 테스트 세트(Test Set)으로 구분하여 예측모형을 구축한다. 이러한 이유는 모형의 성능을 정확하게 평가하기 위함이다. 학습 세트는 모형의 학습을 위해 사용되며, 테스트 세트는 학습된 모형의 성능을 평가하기 위해 사용된다. 테스트 세트는 학습에 사용되지 않았으므로 모형이 알지 못하는 데이터를 포함하고 있어, 모형의 정확성을 검증하는 데 활용된다. 본 연구의 모형의 학습, 검증, 테스트의 과정을 명확하게 설명하고, 예측결과와 성능도 학습 세트와 테스트 세트를 분리해서 제시한다. 여기서 학습과 테스트의 비율은 전체 데이터 세트에서 학습 세트 70%, 테스트 세트는 30%로 지정한다.

넷째, 검증 세트(Validation)로 구분하여 모형을 교차검증을 수행한다. 예측모형은 학습하면서 학습과정에서 교차검증을 통해 파라미터와 확률경계를 최적화한다. 일반적으로 학습에 사용하지 않은 표본에 대해 종속변수를 알아내고자 하는 것을 예측이라고 하며, 예측성능이 표본 내에서는 좋지만 표본 외에서 좋지 않은 경우를 과적합(Overfitting)이라고 한다. 과적합문제가 발생할 경우 새로운 설명변수 데이터에 대해 전혀 예측하지 못하기 때문에 목적에 부합하지 못하게 된다. 이를 해결하기 위하여 교차검증을 고려한다. 교차검증은 수차례 모형 학습과정 및 검증과정을 통해 예측모형의 통계적 신뢰도를 높일 수 있다. 또한 데이터 수가 부족하면 그만큼 적게 성능 평가가 이루어져 검증 성능의 신뢰도가 낮아질 수 있고, 검증 데이터를 늘리면 학습 데이터가 적어지기 때문에 정상적인 학습이 이루어지지 않는다. 이 같은 문제를 해결하기 위한 검증방법론이 K폴더 교차검증(K-fold-Cross Validation)이다(김종성 외, 2019). 이때 CV의 K의 수는 자료의 양에 따라 달라지고, 주로 3~10 사이의 값이 사용된다. 또한 검증자료 생성 시 불균형적인 클래스에서 샘플링 편향 현상이 일어나는 것을 방지하기 위해 계층적 샘플링(Stratified

sampling)기법을 적용하면 평가의 신뢰도를 높일 수 있다(최용욱, 2020). 일반적으로 K의 수는 3~10 사이의 값이 사용하여 평가의 신뢰성을 검증하므로 본 연구에서는 K=10을 사용한다.



<그림 4-9> 앙상블 학습 예측모형 구축과정

2. 경유자동차 PM 예측모형 설계

경유자동차 PM 예측모형 설계는 분석데이터의 특성 분석과 앙상블 학습 예측모형의 이론적 고찰을 통해 두 가지 가설을 제기하는 바이다.

○ 가설1 : 경유자동차 PM 배출요인은 검사방식에 따라 다를 것이다.
  - 검사방식: KD-147모드, Lug-Down3모드
○ 가설2 : 경유자동차 PM 배출요인은 차종에 따라 다를 것이다.
  - 차종: 승용, 화물, 승합(KD-147모드), 화물, 특수, 승합(Lug-Down3모드)

위 두 가지 가설을 설정한 이유는 다음과 같다. 첫째, KD-147모드와 Lug-Down3모드의 검사방식이 다르기 때문이다. KD-147모드는 실험실에서 실도로를 차량이 직접 주행하여 PM배출농도(%)를 측정하는 방식이고, Lug-Down3모드는 차량 부하방식으로 엔진출력과 PM배출농도(%)를 동시에 검사하는 방식이다. 두 검사방식은 검사방식에도 차이가 있고, PM 측정치도 KD-147모드는 단일 측정치이나 Lug-Down3모드는 3개 측정치이므로 검사방식별로 모형을 분류한다.

둘째, 차종에 따라 경유자동차 PM 배출요인 다를 개연성이 높으므로 검사방식별 차종별 모형을 분류하여 구축한다. 여기서 KD-147 검사모드 대상에 특수차가 포함되지만 전체 조사차량 대비 특수차 비중이 현저히 작으므로 특수차의 PM 배출요인을 밝혀줄 예측모형의 정확도가 매우 낮을 것으로 예상된다. 따라서 KD-147 모드 차종별 PM 예측모형에는 특수차를 제외한다.

이와 같은 근거를 바탕으로 본 연구는 검사방식과 차종에 따라 개별모형을 구축하고자 한다. 이는 가설에 따라 경유자동차 PM 배출요인이 차이를 증명하고자 한다.

3. 경유자동차 PM 예측모형 구축

3.1 예측모형 입력변수 선정

입력변수는 수치형변수와 범주형 변수로 양분하였다. 수치형 변수는 차량연식, 차량총중량, 주행거리, 연비, 배기량, 길이, 너비, 높이 등이며, 범주형 변수는 저감장치유무, 배출가스등급, 유로기준, 사업용도 등이 있다. 배출가스등급은 1~5등급으로 적용하였으며, 배출가스규제기준은 EURO4 이전와 EURO5 이후로 구분하고, 저감장치, 배출가스검사 사업용도는 더미변수로 처리하였다.

<표 4-8> 앙상블 학습 예측모형 입력변수

모형		단위	변수 형태
종속변수		PM 배출농도	수치형 변수
입력 변수	차량연식	년	
	차량총중량	kg	
	차량적재량	kg	
	주행거리	km	
	배기량	cc	
	길이	m	
	너비	m	
	높이	m	
	연비	l/km	
	승차정원	인	
	배출가스등급	2~5등급	범주형 변수
	저감장치 유무	유(1), 무(0)	
	사업용도	사업용(1), 비사업용(0)	
	유로	유로4 이전(1), 유로5 이후(0)	

3.2 예측모형 분류

예측모형 설계에서 설정한 연구 가설에 따른 예측모형을 구축을 위해 다음과 같이 모형을 분류하고자 한다. 먼저 1차 앙상블 학습 예측모형에서 자동차 검사방식에 따라 KD-147모드와 Lug-Down3모드로 구분하여 예측모형 분석한다. 그리고 하위모형으로는 배출가스검사에서 합격 판정을 받은 차량에 대해 예측모형을 구축한다. 그 이유는 합격 판정 받은 차량 데이터는 분석데이터의 일관성이 담보되어 예측력이 높을 것으로 사료된다. 반면 불합격 판정을 받은 차량에 대해 예측모형을 구축하지 않은 이유는 <표 4-9>와 같이 불합격데이터 비율은 전체 데이터에 15% 내외이므로 샘플수가 부족하고 불합격 판정 차량의 PM 측정치의 범위가 광범위한 관계로 예측모형의 성능이 낮을 것으로 예상된다. 따라서 불합격 모형을 따로 구축하지 않는다.

2차 앙상블 학습 예측모형은 연구 가설을 토대로로 배출가스 검사방식과 차종별 분류하여 경유자동차 PM 예측을 수행하도록 한다. 이는 모형의 예측성능과 PM 배출요인도 다를 것으로 예상되기 때문이다.

<표 4-9> 경유자동차 배출가스검사 합격 및 불합격 비율

합격		불합격		합계	
대	비중(%)	대	비중(%)	대	비중(%)
484,290	**85.3**	83,178	**14.7**	567,468	100

3.3 데이터 전처리 탐색

자료 전처리 탐색은 예측모형 구축을 위한 데이터 준비과정이다. 일반적으로 데이터 정제, 데이터 처리, 이상치 제거, 특성 추가, 데이터 확장(data augmentation), 입력변수 스케일링(feature scaling) 등의 과정을 거쳐야 한다.

1) 이상치 제거

이상치란 데이터 내에 비현실적인 값이나 비정상적으로 극단값을 말한다. 통상 데이터를 분석할 경우 이상치가 존재하면 분석결과에 지대한 영향을 미칠 수 있으므로 제거하는 것이 중요하다. 일반적으로 이상치를 제거할 때 데이터의 평균값이 사용된다.

본 연구의 분석방법인 앙상블 학습기법의 예측력은 학습과정에서 이용된 데이터 포인트(Data points)의 정확도에 따라 결정된다. 정확도를 높이기 위해서는 수학적 모형인 레버리지 분석(Leverage analysis) 방법론을 이용하기로 한다. 레버리지 분석은 대용량자료의 이상치 탐색에 매우 효율적인 방법론으로 알려져 있다(Mohammadi et al., 2012; Khamehchi and Bemani; 2021, Bemani et al., 2022). 그 이유는 레버리지 분석은 종속변수와 설명변수의 이상치를 동시에 탐색하므로 대용량자료의 이상치 탐색에 적절하다.

레버리지란 실제 종속변수가 예측치에 미치는 영향을 나타낸 값을 의미한다. 레버리지 분석은 H(Hat matrix)기반 분석을 기본전제로 한다. 모형의 잔차(R)와 H값의 기준에서 벗어난 이상치를 탐색한다. H는 영향도 행렬 또는 hat 행렬이라고 한다.

식(4-15)은 2차원 행렬 X, 역행렬 T로 구성되었다. 여기서 X는 $m \times n$로 나타내며, m과 n은 영향도 행렬의 대각성분으로 정의된다. 레버리지는 관측치가 예측치에 미치는 영향, 즉, 예측지점을 자기 자신의 위치로 끌어 당기는 정도를 나타내는 것이다. (Mohammadi et al., 2012; Khamehchi and Bemani; 2021, Bemani et al., 2022).

식(4-16)의 H^*값은 n개 학습데이터와 모형의 p개 파라미터를 의미한다. 즉, H^*값은, 파라미터 p개를 n개 변수로 나눈값이다. 여기서, H^*의 대각성분이 1이고 나머지 성분이 모두 0이면 예측치와 종속변수값이 일치함을 의미한다. 일반적으로 잔차의 범위는 $-3 \le R \le 3$이며, H지표 범위는 $0 \le H \le H^*$이다. 여기서, H^*는 레버리지값의 임계치이며, 이 범위

밖에 탐색된 관측치는 이상치(High Leverage)이라 칭하며, 이를 제거하기로 한다(Mohammadi et al., 2012).

$$H = X(X^T X)^{-1} X^T \quad \text{(식 4-15)}$$

$$H^* = 3\frac{p+1}{n} \quad \text{(식 4-16)}$$

초기 분석데이터 수는 KD-147모드가 182,792대이며, 특수차가 1,458대로 데이터 샘플수가 적은 관계로 특수차 데이터를 제외한 181,344대로 분석한다. Lug Down3모드의 초기 분석데이터 수는 386,124대이다. 분석데이터에 Null 값을 제외한 데이터 수는 KD-147모드가 150,669대, Lug-Down3모드가 340,424대이다. 이상치를 제거한 데이터 수는 KD-147모드가 137,723대, Lug-Down3모드가 315,792대로 실제 모형에 적용될 데이터수 이다. Null 값을 제외 데이터에서 이상치를 제거한 데이터 비율은 KD- 147모드, Lug-Down3모드 각각 8.6%, 7.2%이다.

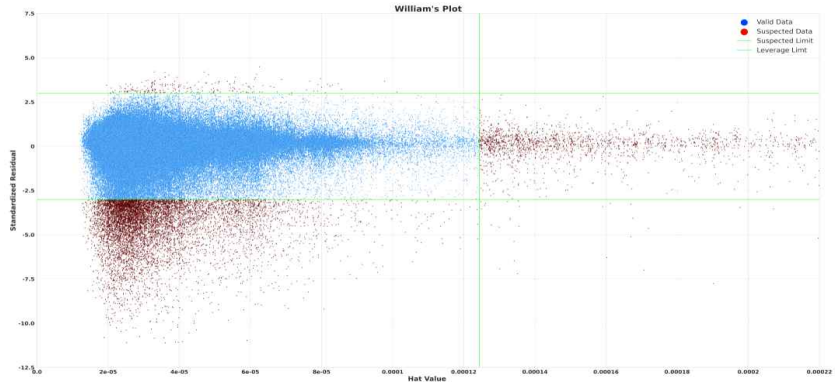
이상치 제거 결과는 Willam plot으로 제시한다. 여기서 Hat value(H^*) 경계와 잔차(R) 경계 안에 존재하는 데이터는 신뢰성 영역으로 포함시키고, 그 밖에 존재하는 관측치는 이상치로 간주하여 분석데이터에서 제외한다. 본 연구에서는 모형 분류기준에 따라 검사방식과 차종별로 이상치를 제거하였다. 그 결과는 <그림 4-10>~<그림 4-19>에 제시한다.

<표 4-10> 분석데이터 현황

단위 : 대

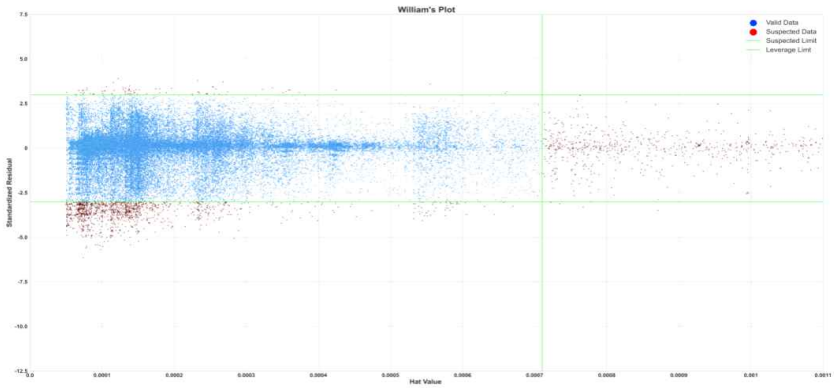
구분	KD-147모드	Lug-Down3모드
분석데이터	181,344	386,124
Null 값 제거 데이터	150,669	340,424
이상치 제거 데이터	137,723	315,792
합격데이터	122,693	263,950
불합격데이터	14,770	48,187

<그림 4-10> Willam plot(KD-147모드 통합데이터 전차종)

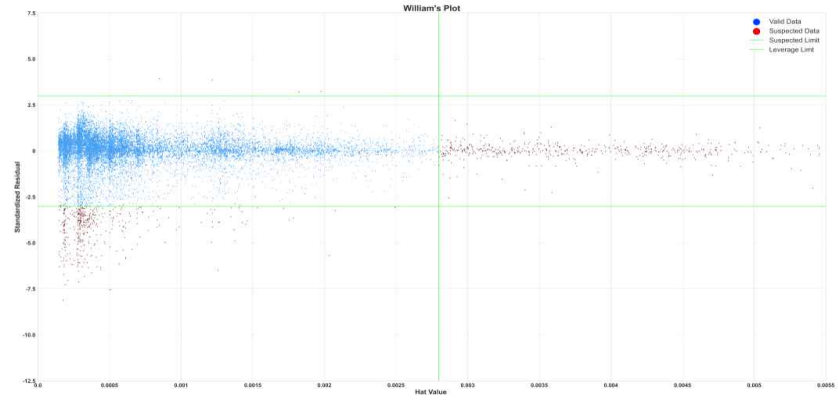


<그림 4-11> Willam plot(Lug-Down3모드 통합데이터 전차종)

<그림 4-12> Willam plot(KD-147모드 합격데이터 전차종)

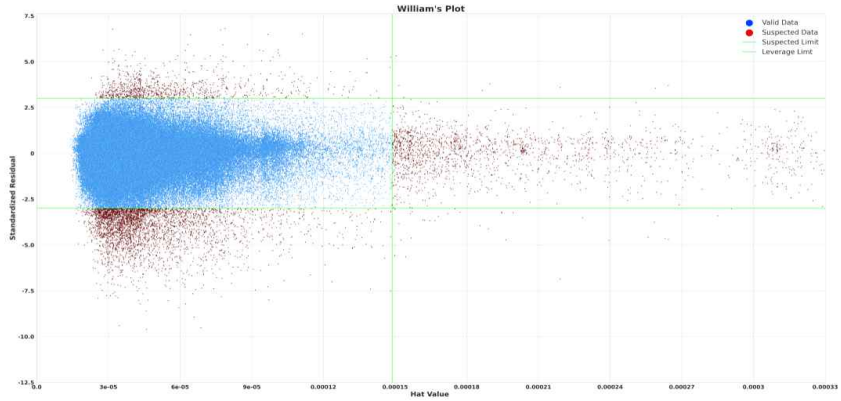


<그림 4-13> Willam plot(KD-147모드 합격데이터 승용)

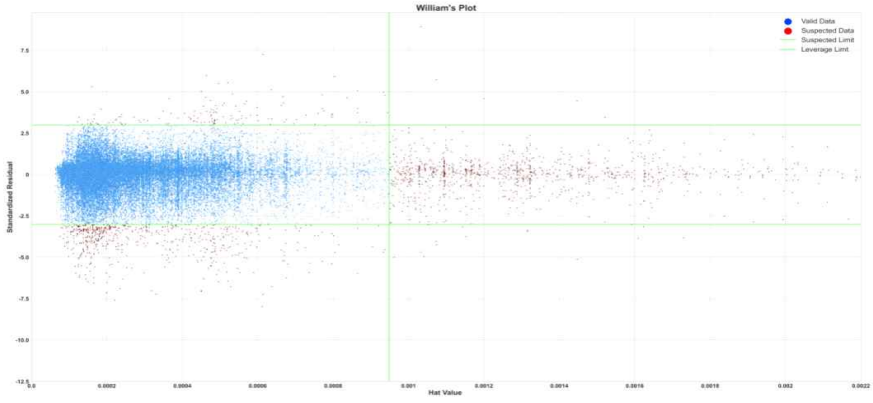


<그림 4-14> Willam plot(KD-147모드 합격데이터 승합)

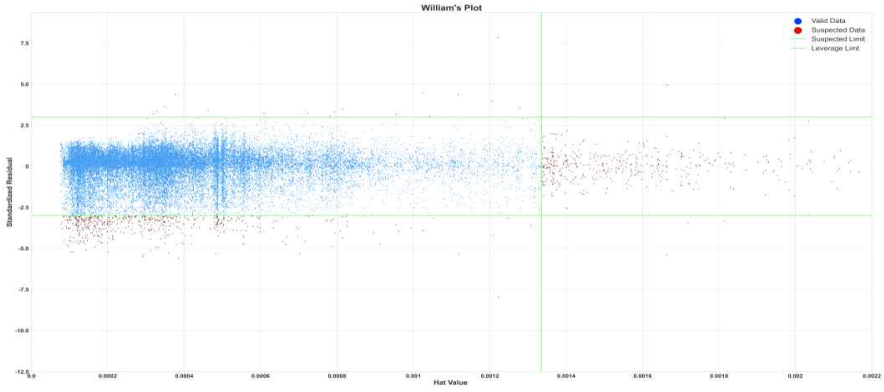
<그림 4-15> Willam plot(KD-147모드 합격데이터 화물)



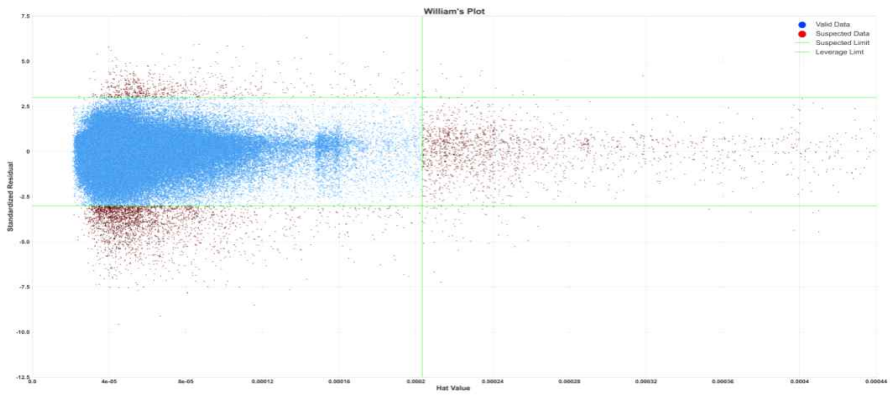
<그림 4-16> Willam plot(Lug-Down3모드 합격데이터 전차종)



<그림 4-17> Willam plot(Lug-Down3모드 합격데이터 특수)



<그림 4-18> Willam plot(Lug-Down3모드 합격데이터 승합)



<그림 4-19> Willam plot(Lug-Down3모드 합격데이터 화물)

2) 입력변수 스케일링(Feature Scaling)

변수는 각 특성에 따라 변수의 범위가 다르다. 만약 변수 값을 그대로 모형에 적용하면 모형 학습시에 스케일이 큰 변수의 영향을 더 많이 받아 최적해를 찾기 어렵고 학습시간 또한 더 오래 소요된다. 그러므로 입력변수를 그대로 모형에 적용하지 않고 입력변수 스케일링(Feature scaling) 과정을 거쳐야 한다. 입력변수 스케일링은 Min-Max Scaling, Stardard Scaling 등 다양한 방법론이 있으나 본 연구의 주요변수는 양의 값만 존재하므로 Min-Max Scaling을 선택한다. Min-Max Scaling은 (식 4-17)과 같이 입력변수에 최소값을 뺀 뒤에 (최대값-최소값)을 나눈 비율을 의미하며, 범위는 [1, 0] 이내로 유지된다.

$$x^{*} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad \text{(식 4-17)}$$

제Ⅴ장. 경유자동차 PM 예측결과 및 평가

제1절 1차 앙상블 학습 기반 예측모형 평가

1. 모형 평가지표 선정

본 연구는 앙상블 학습을 통해 통계모형, 배깅모형, 부스팅모형들을 구축하였으며, 통계기법보다 머신러닝기법의 예측력이 향상된 것을 증명하기 위해 모형평가를 수행한다. 일반적으로 예측모형을 평가할 때 실측치와 예측치의 잔차를 사용한다. 즉, 모형의 실측치와 예측치의 차이 값을 기반으로 한 지표가 주로 평가지표로 활용된다. 따라서 본 장에서 모형평가는 다음과 같은 평가지표를 사용한다.

평균절대오차(Mean Absolute Error, MAE)는 실제치와 측정치의 차이의 절대 오차평균이다. 식(5-1)을 통해 MAE 값이 작으면 예측치와 실측치 오차가 작음을 의미한다. 평균제곱오차(Mean Squared Error, MSE)는 모형 오차를 제곱하여 평균을 취한 값이다. 식(5-2)와 같이 산정된 MSE는 예측의 정확도의 척도로 많이 사용하며, 수치가 작을수록 정확성이 높다. 평균 제곱근 편차(Root Mean Square Error, RMSE)는 식(5-3)를 통해 PM 예측값과 평균값을 비교한다. RMSE는 MSE에 제곱근을 취한 값이며, MAE에 비해 직관성은 떨어지지만 극단값에 덜 민감한 강점을 보인다. RMSE값이 클수록 모형의 예측치와 실제치의 오차가 큰 것을 의미한다. 결정계수 R^2는 모형의 예측치와 실제치의 상관관계 정도를 나타내는 상관계수(R)을 제곱한 것과 같다. 식(5-4)에 산정된 R^2가 1에 가까우면 변수 사이의 관계를 완전히 설명해 주고 있음을 의미한다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{predicted} - y_i^{actual}\right| \quad \text{(식 5-1)}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i^{predicted} - y_i^{actual})^2 \quad \text{(식 5-2)}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{predicted} - y_i^{actual})^2} \quad \text{(식 5-3)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^{predicted} - y_i^{actual})^2}{\sum_{i=1}^{n}(y_i^{actual} - \overline{y_i^{actual}})^2} \quad \text{(식 5-4)}$$

2. 모형 성능 평가 및 비교

1차 앙상블 학습 예측모형은 KD-147모드와 Lug-Down3모드로 구분해서 통계모형, 배깅과 부스팅 알고리즘을 기반한 20개 모형을 분석하였다. 20개에 대한 모형 설명은 <표 5-1>과 같이 제시하는 바이다.이 모형들 중에서 예측성능이 현저히 떨어지는 모형을 제외하였다. <표 5-2>~<표 5-7>에서 14개 모형의 평가지표를 제시하였으며, 이중 정확도가 높으면서 통계, 배깅, 부스팅 기법을 대표할 수 있는 모형 6개를 선정하였다. 회귀를 대표하는 선형회귀모형과 의사결정나무를 선택하였으며, Bagging을 대표하는 랜덤포레스트모형을 선정하였다. 나머지 3개 모형은 Boosting을 대표하는 모형이며, 예측력이 상위 1~3위인 CatBoost, LightGBM, XGBoost모형을 선정하였다.

KD-147모드의 CatBoost, LightGBM, XGBoost 모형의 R^2는 0.626~0.622이며, 랜덤 포레스트 0.562, 선형회귀모형 0.464, 의사결정나무 0.283로 산정되어 예측모형의 신뢰성이 비교적 낮은 것으로 확인되었다. Lug-Down3모드는 CatBoost, LightGBM, XGBoost, 랜덤 포레스트 모형의 R^2는 0.587~0.558이며, 선형회귀모형 0.439, 의사결정나무 0.229로 산정되어 KD-147모드 모형들 보다는 예측력이 떨어지는 것으로 나타났다.

합격데이터 예측모형의 경우 KD-147모드에서 CatBoost모형의 예측력이 가장 높았으며, 모형의 R^2는 KD-147모드가 0.807, Lug-Down3모드에서 0.716로 나타났다. 분석데이터로 모든 예측모형을 구축하는 것 보다

합격데이터만 이용한 예측모형의 성능이 더 높은 것으로 알 수 있다. 반면 불합격데이터로 구축한 예측모형의 성능 현저히 낮은 것을 확인하였다. 이는 합격데이터의 특성이 경유자동차 PM 배출 예측의 설명력을 더 높일 수 있음을 방증하는 것이다. 따라서 2차 앙상블 학습 예측모형은 합격데이터로만 이용하기로 한다.

<표 5-1> 1차 앙상블 학습 예측모형

대분류	모형	모형 약어
Regression	Extra Trees Regressor	et
	AdaBoost Regressor	ada
	Bayesian Ridge	br
	Decision Tree Regressor	dt
	Dummy Regressor	dummy
	Elastic Net	en
	Huber Regressor	huber
	K Neighbors Regressor	knn
	Least Angle Regression	lar
	Lasso Regression	lasso
	Lasso Least Angle Regression	llar
	Linear Regression	lr
	Orthogonal Matching Pursuit	omp
	Passive Aggressive Regressor	par
	Ridge Regression	ridge
Bagging	Random Forest Regressor	rf
Boosting	AdaBoost Regressor	ada
	CatBoost Regressor	catboost
	Extreme Gradient Boosting	xgboost
	Gradient Boosting Regressor	gbr
	Light Gradient Boosting Machine	lightgbm

<표 5-2> 1차 앙상블 학습 예측모형 성능 비교(KD-147모드 통합데이터)

Model	*MAE*	*MSE*	*RMSE*	R^2
CatBoost	**2.808**	**15.921**	**3.990**	**0.626**
XGBoost	**2.815**	**16.075**	**4.009**	**0.623**
LightGBM	**2.826**	**16.128**	**4.016**	**0.622**
Gradient Boosting	2.955	17.636	4.199	0.586
Random Forest	**2.995**	**18.677**	**4.322**	**0.562**
Ridge Regression	3.390	22.752	4.770	0.466
Bayesian Ridge	3.390	22.752	4.770	0.466
Linear Regression	**3.395**	**22.835**	**4.778**	**0.464**
Extra Trees	3.234	23.048	4.801	0.459
OMP	3.692	25.012	5.001	0.413
Elastic Net	3.700	25.056	5.005	0.412
Lasso Regression	3.702	25.074	5.007	0.412
Decision Tree	**3.618**	**30.547**	**5.527**	**0.283**
KNN	4.307	37.147	6.095	0.129

<표 5-3> 1차 앙상블 학습 예측모형 성능 비교(Lug-Down3모드 통합데이터)

Model	*MAE*	*MSE*	*RMSE*	R^2
CatBoost	**2.567**	**11.899**	**3.449**	**0.587**
XGBoost	**2.571**	**11.982**	**3.461**	**0.585**
LightGBM	**2.600**	**12.199**	**3.493**	**0.577**
Random Forest	**2.629**	**12.744**	**3.570**	**0.558**
Gradient Boosting	2.685	13.100	3.619	0.546
Extra Trees	2.786	15.037	3.878	0.479
Ridge Regression	2.963	16.167	4.021	0.439
Bayesian Ridge	2.963	16.167	4.021	0.439
Linear Regression	**2.963**	**16.167**	**4.021**	**0.439**
Elastic Net	3.006	16.791	4.098	0.418
Lasso Regression	3.008	16.807	4.100	0.417
OMP	3.070	17.923	4.234	0.379
AdaBoost	3.838	19.900	4.459	0.310
Decision Tree	**3.299**	**22.219**	**4.714**	**0.229**

<표 5-4> 1차 앙상블 학습 예측모형 성능 비교(KD-147모드 합격데이터)

Model	*MAE*	*MSE*	*RMSE*	R^2
CatBoost	**1.200**	**2.705**	**1.645**	**0.807**
LightGBM	**1.203**	**2.739**	**1.655**	**0.805**
XGBoost	**1.210**	**2.747**	**1.657**	**0.804**
Gradient Boosting	1.276	3.133	1.770	0.777
Random Forest	**1.259**	**3.145**	**1.773**	**0.776**
Extra Trees	1.336	3.837	1.959	0.726
Linear Regression	1.585	4.923	2.219	0.649
Ridge Regression	1.585	4.923	2.219	0.649
Bayesian Ridge	1.585	4.923	2.219	0.649
Decision Tree	**1.482**	**5.149**	**2.269**	**0.633**
OMP	1.889	6.163	2.482	0.560
Elastic Net	1.896	6.216	2.493	0.557
Lasso Regression	1.900	6.253	2.501	0.554
AdaBoost	2.240	6.532	2.551	0.534

<표 5-5> 1차 앙상블 학습 예측모형 성능 비교(Lug-Down3모드 합격데이터)

Model	*MAE*	*MSE*	*RMSE*	R^2
CatBoost	**1.314**	**3.069**	**1.752**	**0.716**
XGBoost	**1.314**	**3.094**	**1.759**	**0.714**
LightGBM	**1.336**	**3.172**	**1.781**	**0.707**
Random Forest	**1.327**	**3.222**	**1.795**	**0.702**
Gradient Boosting	1.395	3.479	1.865	0.678
Extra Trees	1.400	3.761	1.939	0.652
Least Angle Regression	1.548	4.283	2.070	0.604
Bayesian Ridge	1.548	4.283	2.070	0.604
Linear Regression	**1.548**	**4.283**	**2.070**	**0.604**
Ridge Regression	1.548	4.283	2.070	0.604
Elastic Net	1.621	4.692	2.166	0.566
OMP	1.624	4.714	2.171	0.564
AdaBoost	1.667	5.098	2.258	0.529
Decision Tree	2.011	5.494	2.341	0.492

<표 5-6> 1차 앙상블 학습 예측모형 성능 비교(KD-147모드 불합격데이터)

Model	*MAE*	*MSE*	*RMSE*	R^2
LightGBM	**7.449**	**91.438**	**9.555**	**0.369**
CatBoost	**7.469**	**91.835**	**9.577**	**0.366**
Random Forest	**7.455**	**92.574**	**9.614**	**0.361**
XGBoost	**7.523**	**94.063**	**9.694**	**0.350**
Gradient Boosting	7.714	96.788	9.830	0.332
Bayesian Ridge	8.159	107.926	10.381	0.255
Ridge Regression	8.147	107.941	10.382	0.255
Linear Regression	**8.175**	**108.766**	**10.421**	**0.250**
Extra Trees	8.056	109.371	10.453	0.244
OMP	8.229	109.998	10.481	0.241
Elastic Net	8.447	114.156	10.677	0.213
Lasso Regression	8.483	115.028	10.718	0.207
AdaBoost	9.779	135.931	11.648	0.062
Decision Tree	**8.977**	**138.758**	**11.767**	**0.044**

<표 5-7> 1차 앙상블 학습 예측모형 성능 비교(Lug-Down3모드 불합격데이터)

Model	*MAE*	*MSE*	*RMSE*	R^2
Random Forest	**6.084**	**58.832**	**7.669**	**0.614**
CatBoost	**6.170**	**59.552**	**7.716**	**0.609**
XGBoost	**6.180**	**60.194**	**7.758**	**0.605**
LightGBM	**6.226**	**60.710**	**7.791**	**0.602**
Extra Trees	6.364	65.173	8.072	0.572
Gradient Boosting	6.463	65.736	8.107	0.569
Linear Regression	**6.961**	**77.219**	**8.787**	**0.493**
Bayesian Ridge	6.961	77.219	8.787	0.493
Ridge Regression	6.961	77.222	8.787	0.493
Least Angle Regression	6.961	77.227	8.787	0.493
AdaBoost	7.248	78.548	8.862	0.485
Elastic Net	7.304	85.230	9.232	0.441
Lasso Regression	7.342	86.141	9.281	0.435
Decision Tree	**8.019**	**100.425**	**10.021**	**0.341**

제2절 2차 앙상블 학습 기반 예측모형 결과 및 평가

1. 하이퍼파라미터 최적화

하이퍼파라미터(Hyper-parameter)란 모형이 훈련하면서 업데이트하는 가중치 파라미터나 입력 파라미터를 제외하고 모델을 설계할 때 직접 설정해줘야 하는 파라미터를 말한다. 본 분석모형은 베이지안 최적화(Bayesian Optimize) 알고리즘을 통해 산출된 최적의 하이퍼파라미터를 각 모형에 적용하여 모형평가를 수행하였다. 일반적으로 분석가는 직관적으로 하이퍼파라미터를 찾는다. 이러한 방법을 Manual Search라고 한다. 이보다 더 체계적인 방법으로는 Grid Search와 Random Search가 있다(Shahriari et al., 2015).

GridSearch는 탐색하고자 하는 하이퍼파라미터의 값들을 일정한 간격을 두고 각각에 테스트를 거쳐 가장 높은 성능이 나타낼 때 하이퍼파라미터 값을 최종적으로 선택한다. 가능한 모든 경우의 수에 대해 테스트를 하기 때문에 탐색하려는 하이퍼파라미터의 종류가 많을수록 시간이 기하급수적으로 증가한다. 반면, Random Search는 Grid Search와 유사하지만 탐색 구간 내의 후보 하이퍼파라미터 값들을 무작위 추출 후 선정한다는 차이가 있다. Grid Search에 비해 불필요한 반복을 대폭 줄이면서 정해진 간격 사이의 값에 대해서도 확률적인 탐색이 가능하므로 최적의 하이퍼파라미터 값을 더 빨리 찾을 수 있으나 정확도는 다소 떨어진다. 그러나 Grid Search와 Random Search 모두 이전 조사과정에서 얻어진 하이퍼파라미터 값들의 성능에 대한 사전지식을 다음 테스트에 '사전 지식'을 전혀 반영하지 못하는 한계가 존재한다. 새로운 하이퍼파라미터에 대한 조사를 수행할 때 사전지식을 반영하면서 전체적인 과정을 체계적으로 수행할 수 있는 방법으로 베이지안 최적화(Bayesian Optimization)가 있다(Snoek et al., 2012).

베이지안 최적화 방법론은 목적함수 f에 대해 함수값 f(x)를 최대로

만드는 최적해 x를 탐색하는 방법이다. 이때 목적함수 f는 표현식을 명시적으로 알지 못하는 black-box 함수의 형태를 갖는다. 즉, 입력에 따른 출력의 결과가 직접적인 수의 형태로 나타나지 않으며, 출력 값을 구하는 과정 또한 오랜 시간이 소요된다. 따라서 가능한 적은 수의 x 후보에 대해서만 함수값을 연산하며, f를 최대로 만드는 최적해 x를 빠르고 효과적으로 탐색하는 것이 목적이다. 본 연구에서는 베이지안 최적화 알고리즘을 이용하여 <표 5-8>, <표 5-9>와 같이 전차종과 차종별 예측 모형의 하이퍼파라미터를 최적화하였다.

<표 5-8> 하이퍼파라미터 튜닝(KD1-47모드)

Model	Hyper parameter	Min	Max	전차종	화물	승합	승용
LightGBM	n_estimators	200	800	726	287	288	341
	learning_rate	0.01	0.2	0.0697	0.1118	0.1469	0.0590
	max_depth	10	30	22	15	10	23
	min_child_samples	0	70	20	23	21	3
	num_leaves	32	255	38	44	52	76
	colsample_bytree	0.5	1	0.6552	0.9167	0.7085	0.8113
	subsample	0.7	1	0.9052	0.9988	0.8190	0.7201
	reg_alpha	0	1	0.7354	0.5726	0.1863	0.6235
	reg_lambda	0	1	0.1551	0.9245	0.3456	0.8595
CatBoost	n_estimators	200	800	261	259	566	327
	learning_rate	0.01	0.2	0.1912	0.1745	0.0817	0.0743
	max_depth	1	16	11	12	9	15
XGBoost	n_estimators	200	800	259	254	217	282
	learning_rate	0.01	0.2	0.0552	0.0211	0.0198	0.0474
	max_depth	10	30	10	24	36	10
	min_child_weight	1	10	9	7	3	9
	max_leaves	32	255	116	104	83	108
	colsample_bytree	0.5	1	0.7590	0.6696	0.6010	0.8213
	subsample	0.7	1	0.7632	0.7370	0.9075	0.9159
	reg_alpha	0	1	0.9976	0.0201	0.1409	0.7990
	reg_lambda	0	1	0.8384	0.1560	0.7468	0.7336
Decision tree	max_depth	1	16	10	8	5	9
	min_samples_split	2	5	4	4	4	4
	min_samples_leaf	1	4	2	1	1	2
Random forest	n_estimators	200	800	264	489	269	287
	max_depth	1	16	15	13	15	12
	min_samples_split	2	5	2	2	2	2
	min_samples_leaf	1	4	3	1	1	3

<표 5-9> 하이퍼파라미터 튜닝(Lug-Down3모드)

Model	Hyper parameter	Min	Max	전차종	화물	승합	특수
LightGBM	n_estimators	200	800	712	690	728	306
	learning_rate	0.01	0.2	0.1640	0.1100	0.0586	0.0552
	max_depth	10	30	28	25	17	13
	min_child_samples	0	70	14	6	7	6
	num_leaves	32	255	93	60	37	32
	colsample_bytree	0.5	1	0.6204	0.6152	0.5758	0.8678
	subsample	0.7	1	0.9700	0.8557	0.9435	0.9692
	reg_alpha	0	1	0.1162	0.5308	0.9139	0.3381
	reg_lambda	0	1	0.9035	0.8836	0.2999	0.6615
CatBoost	n_estimators	200	800	558	560	552	240
	learning_rate	0.01	0.2	0.0558	0.1205	0.1174	0.2000
	max_depth	1	16	15	14	9	8
XGBoost	n_estimators	200	800	248	255	255	255
	learning_rate	0.01	0.2	0.0586	0.0993	0.0353	0.0994
	max_depth	10	30	13	12	11	12
	min_child_weight	1	10	8	6	1	6
	max_leaves	32	255	117	100	117	100
	colsample_bytree	0.5	1	0.6210	0.6569	0.5782	0.6569
	subsample	0.7	1	0.7878	0.7824	0.9107	0.7824
	reg_alpha	0	1	0.3528	0.9227	0.8116	0.9227
	reg_lambda	0	1	0.6833	0.6061	0.0168	0.6062
Decision tree	max_depth	1	16	11	11	7	9
	min_samples_split	2	5	4	2	2	2
	min_samples_leaf	1	4	1	1	2	1
Random forest	n_estimators	200	800	233	275	273	290
	max_depth	1	16	16	16	15	12
	min_samples_split	2	5	3	2	3	2
	min_samples_leaf	1	4	1	1	1	1

2. 모형 예측 결과 및 성능 평가 및 비교

2차 앙상블 학습에서는 합격데이터 예측모형을 구축한다. 또한 배출검사방식과 차종에 따라 데이터를 분류하여 예측모형을 분석하였다. 모형 성능 평가결과는 학습 세트와 테스트 세트로 구분하여 성능지표를 제시하였다. 모형의 예측력을 판단하기 위해서는 테스트 세트의 성능지표를 참고하면 된다.

경유자동차 PM 배출 예측모형의 예측결과는 <그림 5-1>, <그림 5-2>과 같다. 모형의 적합도는 우수한 것으로 판단된다. <그림 5-3>, <그림 5-4>은 예측모형의 정확도를 비교한 산점도이다. 예측모형 성능은 KD-147모드 CatBoost모형의 R^2가 0.815, *RMSE*는 1.619로 모형의 정확도가 가장 높은 것으로 나타났다. 반면 선형회귀분석에는 R^2가 0.649, *RMSE*는 2.231로 모형의 정확도가 가장 낮은 것으로 분석되었다. 예측성능을 비교해 볼 때 모형의 신뢰성이 높은 순서대로 나열하면 부스팅모형, 배깅모형, 통계모형 순으로 나타났다. 부스팅 모형 중 CatBoost, LightGBM, XGBoost의 예측력은 시나리오에 따라 미미한 차이를 보이며, 모형의 분류기준과 시나리오에 따라 모형 예측력 순위가 달라지기는 하지만 예측성능은 유사하다. 그 다음으로 랜덤 포레스트, 의사결정나무, 선형회귀 순으로 예측력이 낮은 것으로 분석되었다. 이 같은 경향은 KD-147모드와 Lug-Down3모드 둘 다 동일하게 나타났다. 그리고 예측결과에 대한 검증곡선은 <그림 5-5>, <그림 5-6>에 제시하였다. 검증곡선은 모형에 따라 모형의 성능이 어떻게 변하는지를 파악하기 위한 성능지표이다. 본 연구의 제시한 검증 곡선은 모형의 과적합 또는 과소적합이 일어나지 않은 것으로 확인되었다.

차종별로 분류한 예측모형은 KD-147모드의 경우 승용차 예측성능이 가장 뛰어났고 Lug-Down3모드는 화물차의 예측력이 가장 우수하였다. 반면 승합차의 예측모형은 두 검사모드 모두 낮은 것으로 나타났다. 차종별 예측모형의 예측결과와 성능 비교 그리고 검증곡선 그림은 부록에 제시하는 바이다.

<표 5-10> 2차 앙상블 학습 예측모형 성능 비교(KD-147모드 전차종)

Model	*MAE*	*MSE*	*RMSE*	R^2	*MAE*	*MSE*	*RMSE*	R^2
	Test set				Training set			
CatBoost	1.184	2.621	1.619	0.815	1.070	2.087	1.445	0.850
LightGBM	1.185	2.634	1.623	0.814	1.102	2.211	1.487	0.841
XGBoost	1.186	2.648	1.627	0.813	1.082	2.141	1.463	0.846
Random Forest	1.193	2.701	1.644	0.810	1.067	2.109	1.452	0.848
Decision Tree	1.226	2.960	1.720	0.791	1.180	2.672	1.635	0.808
Linear Regression	1.574	4.976	2.231	0.649	1.581	4.952	2.225	0.644

<표 5-11> 2차 앙상블 학습 예측모형 성능 비교(KD-147모드 승용)

Model	*MAE*	*MSE*	*RMSE*	R^2	*MAE*	*MSE*	*RMSE*	R^2
	Test set				Training set			
CatBoost	0.891	1.665	1.290	0.814	0.810	1.272	1.128	0.863
LightGBM	0.893	1.666	1.291	0.814	0.783	1.176	1.084	0.874
XGBoost	0.896	1.694	1.302	0.811	0.780	1.179	1.086	0.873
Random Forest	0.896	1.703	1.305	0.810	0.826	1.349	1.161	0.855
Decision Tree	0.928	1.943	1.394	0.783	0.879	1.611	1.269	0.827
Linear Regression	1.361	3.974	1.993	0.557	1.376	4.052	2.013	0.564

<표 5-12> 2차 앙상블 학습 예측모형 성능 비교(KD-147모드 승합)

Model	*MAE*	*MSE*	*RMSE*	R^2	*MAE*	*MSE*	*RMSE*	R^2
	Test set				Training set			
CatBoost	1.773	6.152	2.480	0.724	1.391	3.470	1.863	0.854
LightGBM	1.831	6.693	2.587	0.699	1.348	3.462	1.861	0.854
XGBoost	1.825	6.505	2.550	0.708	0.825	1.520	1.233	0.936
Random Forest	1.831	6.639	2.577	0.702	1.323	3.462	1.861	0.854
Decision Tree	1.921	7.759	2.786	0.652	1.789	6.590	2.567	0.722
Linear Regression	2.029	8.452	2.907	0.620	1.953	7.872	2.806	0.668

<표 5-13> 2차 앙상블 학습 예측모형 성능 비교(KD-147모드 화물)

Model	*MAE*	*MSE*	*RMSE*	R^2	*MAE*	*MSE*	*RMSE*	R^2
	Test set				Training set			
CatBoost	1.453	3.789	1.947	0.776	1.219	2.525	1.589	0.853
LightGBM	1.450	3.722	1.929	0.780	1.287	2.790	1.670	0.837
XGBoost	1.462	3.833	1.958	0.774	1.011	1.818	1.348	0.894
Random Forest	1.455	3.763	1.940	0.778	1.275	2.731	1.653	0.841
Decision Tree	1.508	4.228	2.056	0.750	1.446	3.745	1.935	0.782
Linear Regression	1.746	5.544	2.355	0.673	1.722	5.524	2.350	0.678

<표 5-14> 2차 앙상블 학습 예측모형 성능 비교(Lug-Down3모드 전차종)

Model	*MAE*	*MSE*	*RMSE*	R^2	*MAE*	*MSE*	*RMSE*	R^2
	Test set				Training set			
CatBoost	1.280	2.913	1.707	0.736	1.089	2.011	1.418	0.813
LightGBM	1.283	2.946	1.717	0.733	1.053	1.885	1.373	0.825
XGBoost	1.288	2.944	1.716	0.733	1.150	2.229	1.493	0.793
Random Forest	1.288	2.971	1.724	0.730	1.106	2.103	1.450	0.804
Decision Tree	1.336	3.347	1.829	0.696	1.291	2.961	1.721	0.725
Linear Regression	1.585	4.488	2.119	0.593	1.581	4.415	2.101	0.590

<표 5-15> 2차 앙상블 학습 예측모형 성능 비교(Lug-Down3모드 화물)

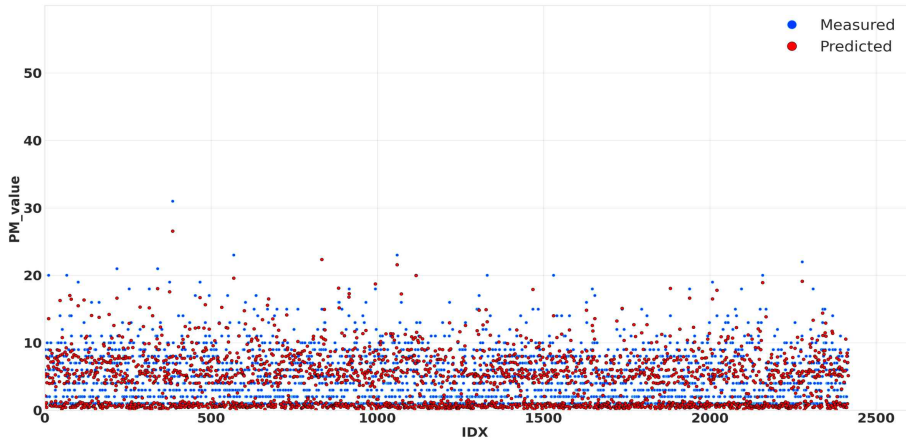
Model	*MAE*	*MSE*	*RMSE*	R^2	*MAE*	*MSE*	*RMSE*	R^2
	Test set				Training set			
CatBoost	1.386	3.411	1.847	0.721	1.104	2.064	1.437	0.832
LightGBM	1.395	3.436	1.854	0.719	1.204	2.413	1.553	0.804
XGBoost	1.388	3.432	1.853	0.720	1.052	1.893	1.376	0.846
Random Forest	1.394	3.480	1.865	0.716	1.149	2.260	1.503	0.816
Decision Tree	1.462	4.062	2.015	0.668	1.382	3.443	1.856	0.720
Linear Regression	1.696	5.152	2.270	0.579	1.694	5.128	2.265	0.583

<표 5-16> 2차 앙상블 학습 예측모형 성능 비교(Lug-Down3모드 승합)

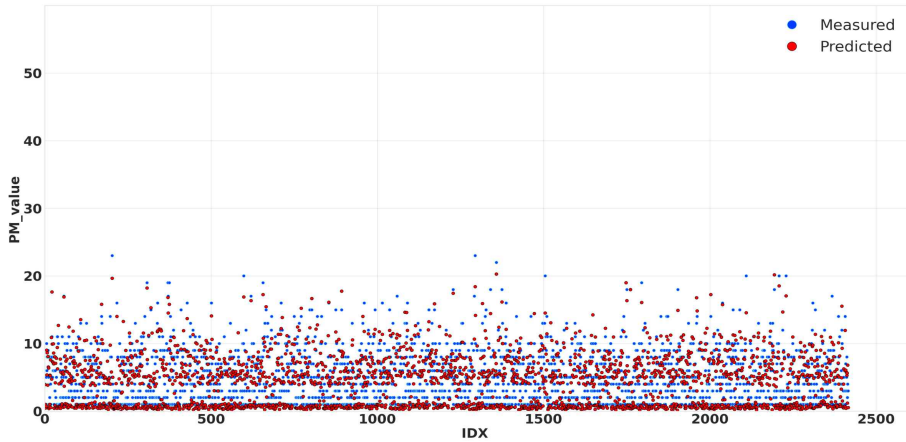
Model	*MAE*	*MSE*	*RMSE*	R^2	*MAE*	*MSE*	*RMSE*	R^2
	Test set				Training set			
CatBoost	0.966	1.917	1.385	0.658	0.639	0.759	0.871	0.846
LightGBM	0.978	1.956	1.398	0.651	0.761	1.051	1.025	0.787
XGBoost	0.975	1.980	1.407	0.647	0.770	1.068	1.033	0.784
Random Forest	0.984	2.018	1.421	0.640	0.716	0.976	0.988	0.802
Decision Tree	1.093	2.673	1.635	0.523	1.037	2.221	1.490	0.550
Linear Regression	1.303	3.610	1.900	0.356	1.264	3.199	1.788	0.352

<표 5-17> 2차 앙상블 학습 예측모형 성능 비교(Lug-Down3모드 특수)

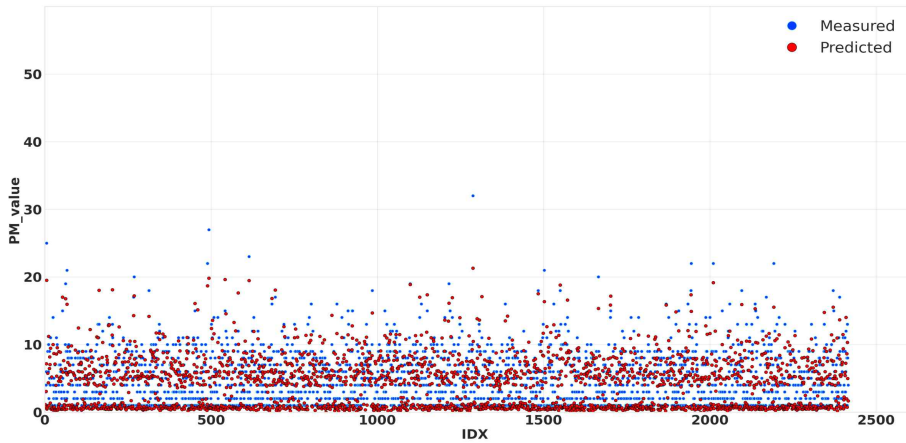
Model	*MAE*	*MSE*	*RMSE*	R^2	*MAE*	*MSE*	*RMSE*	R^2
	Test set				Training set			
CatBoost	1.017	1.893	1.376	0.790	0.928	1.488	1.220	0.843
LightGBM	1.018	1.895	1.377	0.790	0.953	1.575	1.255	0.834
XGBoost	1.018	1.945	1.395	0.785	0.704	0.906	0.952	0.904
Random Forest	1.012	1.916	1.384	0.788	0.904	1.463	1.210	0.846
Decision Tree	1.049	2.144	1.464	0.763	1.012	1.936	1.391	0.796
Linear Regression	1.353	3.284	1.812	0.636	1.372	3.487	1.867	0.632



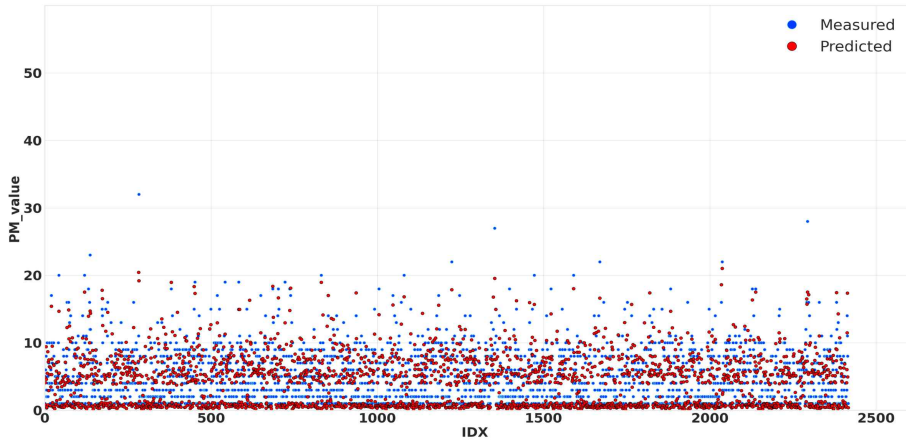
(a) CatBoost



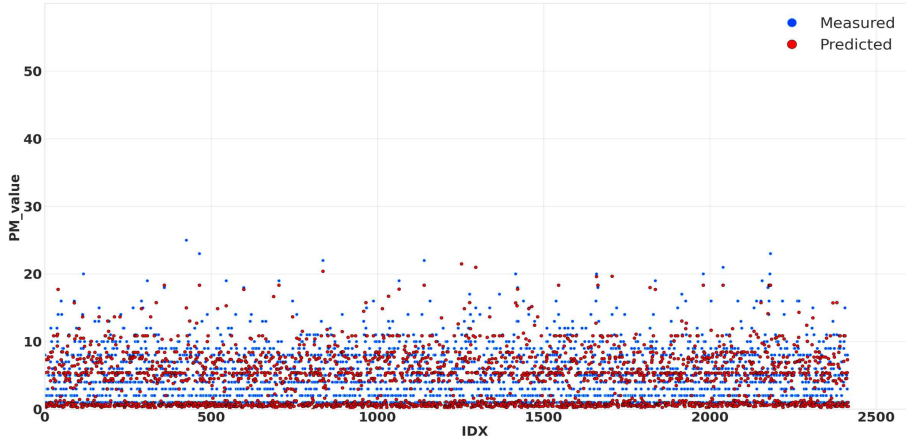
(b) LightGBM



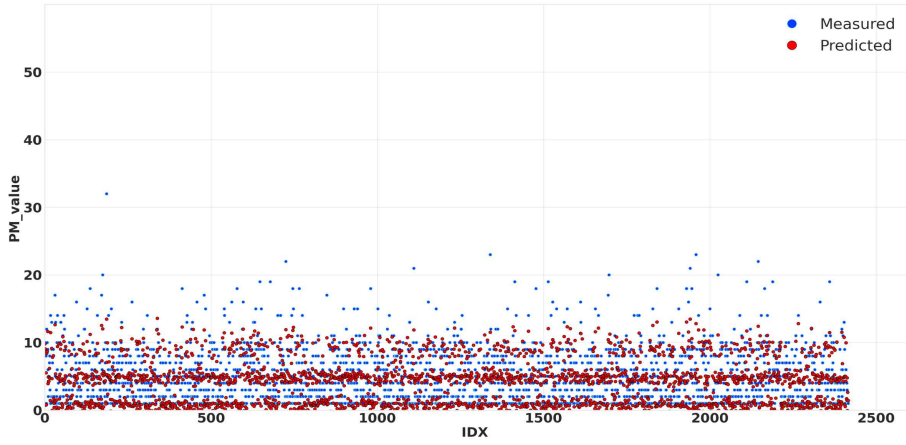
(c) XGBoost



(d) Random Forest

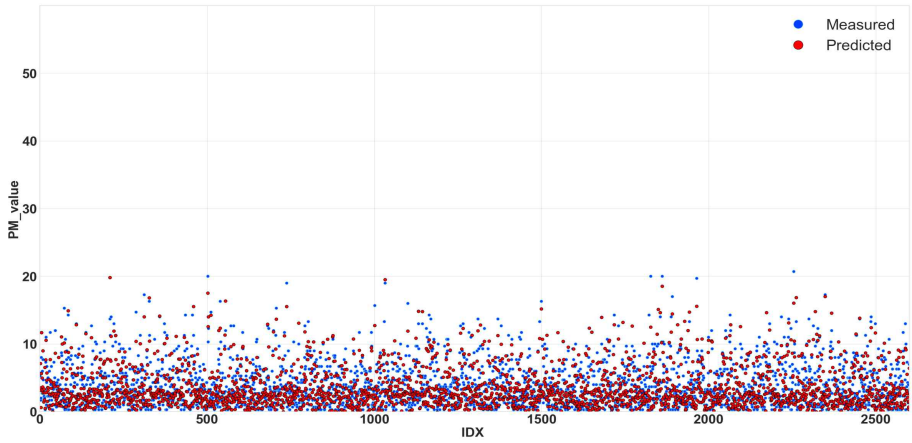


(e) Decision Tree

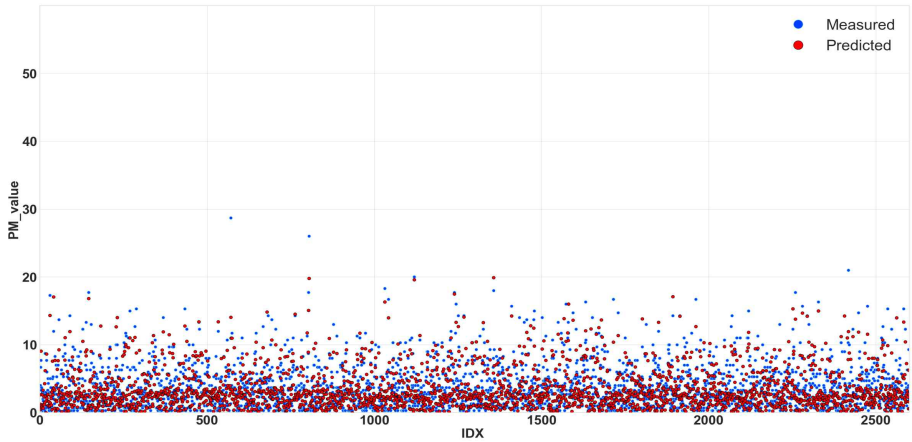


(f) Linear Regression

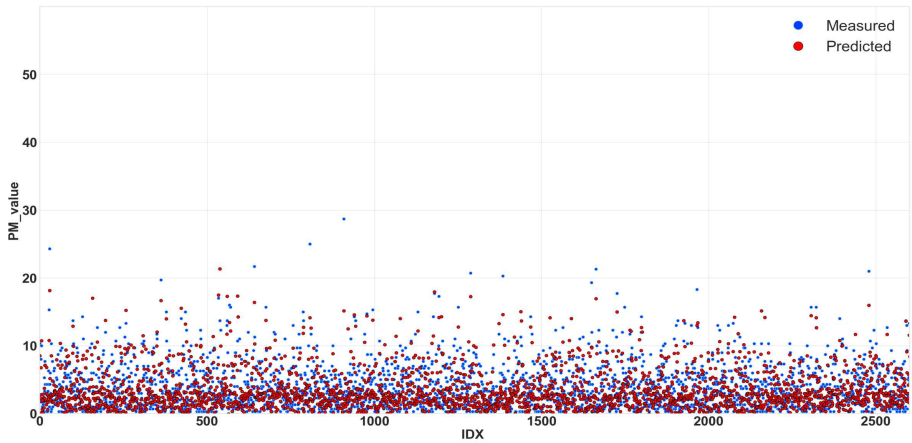
<그림 5-1> 2차 앙상블 학습 예측모형 예측결과(KD-147모드)



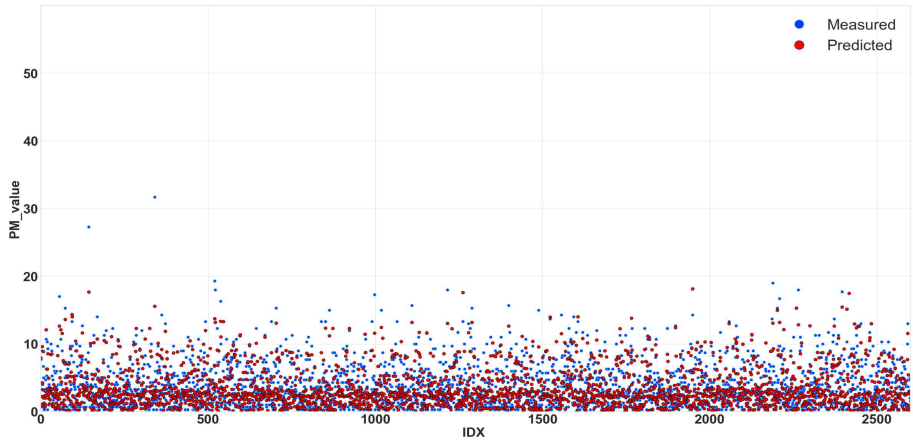
(a) CatBoost



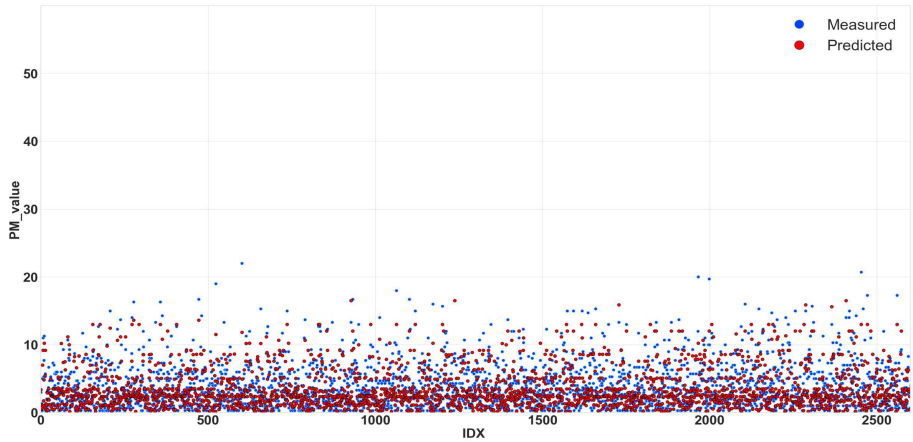
(b) LightGBM



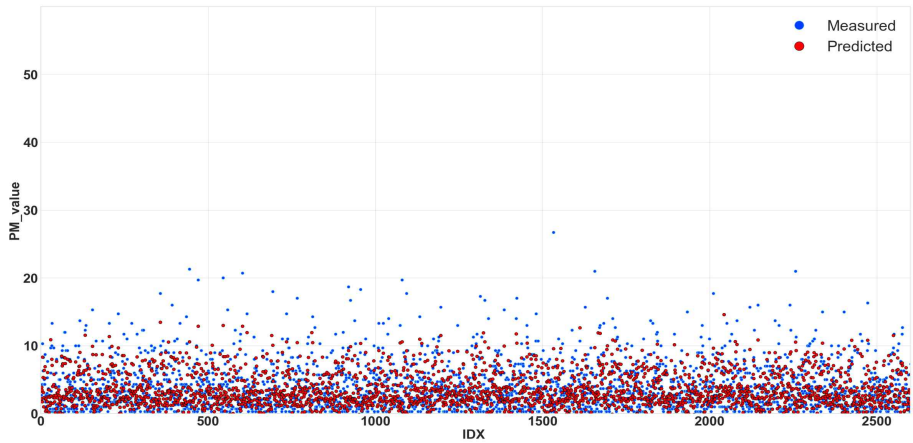
(c) XGBoost



(d) Random Forest

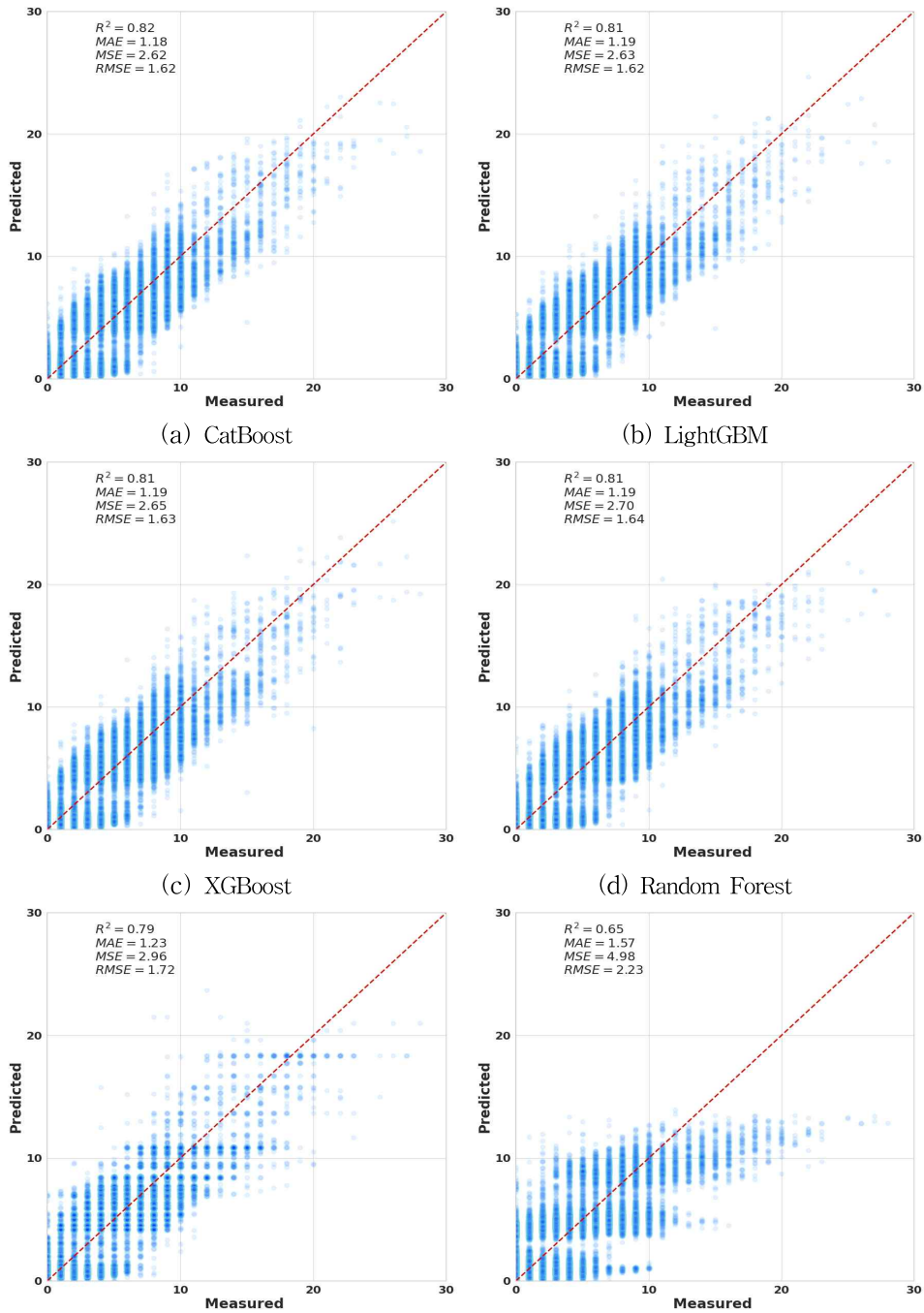


(e) Decision Tree

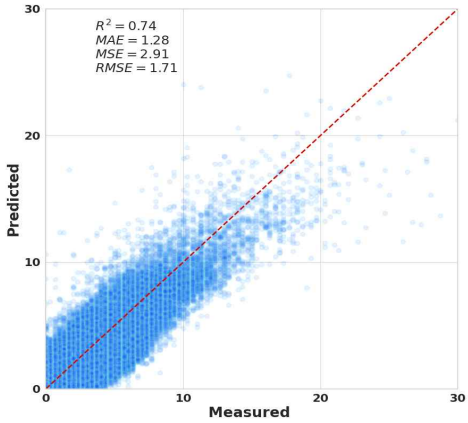


(f) Linear Regression

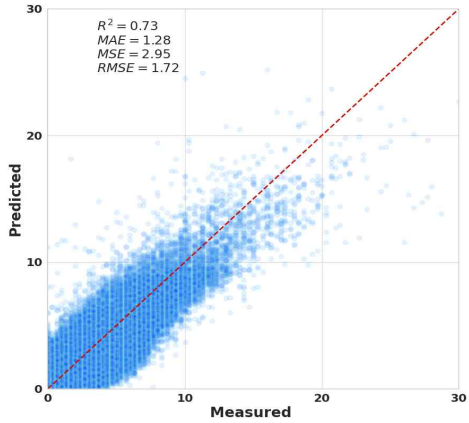
<그림 5-2> 2차 앙상블 학습 예측모형 예측결과(Lug-Down3모드)



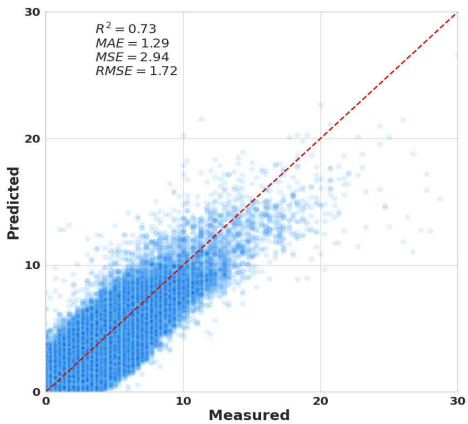
<그림 5-3> 2차 앙상블 학습 예측모형 정확도(KD-147모드)



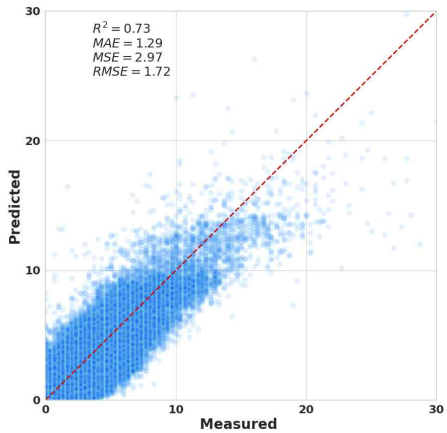
(a) CatBoost



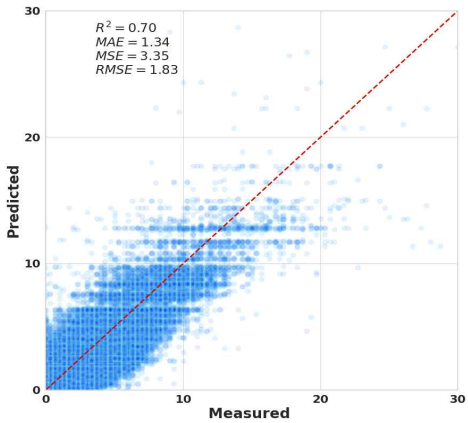
(b) LightGBM



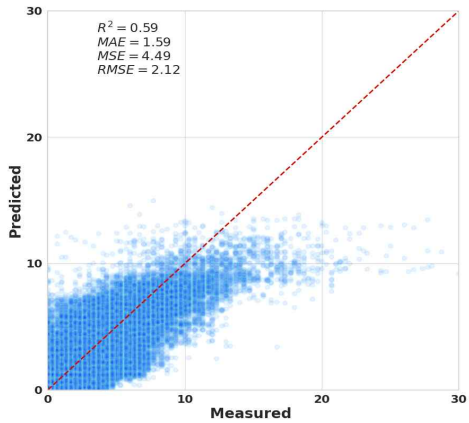
(c) XGBoost



(d) Random Forest

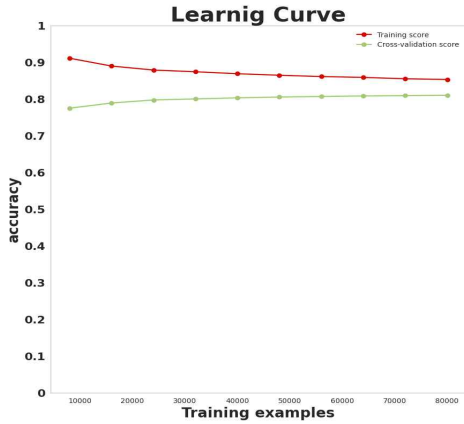


(e) Decision Tree

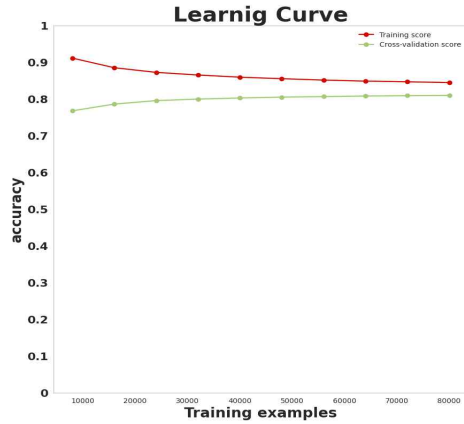


(f) Linear Regression

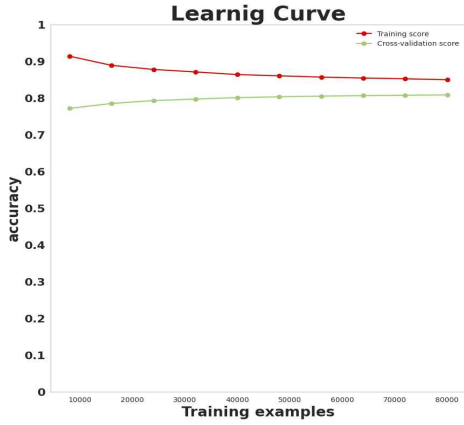
<그림 5-4> 2차 앙상블 학습 예측모형 정확도(Lug-Down3모드)



(a) CatBoost

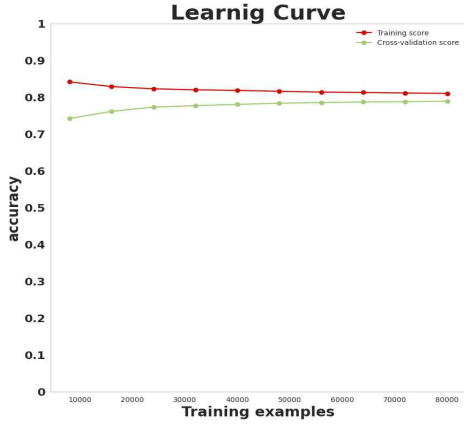


(b) LightGBM

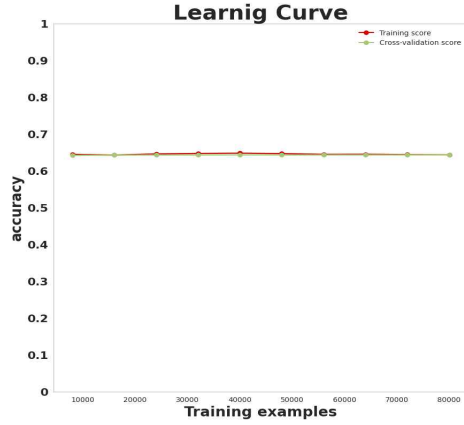


(c) XGBoost

(d) Random Forest

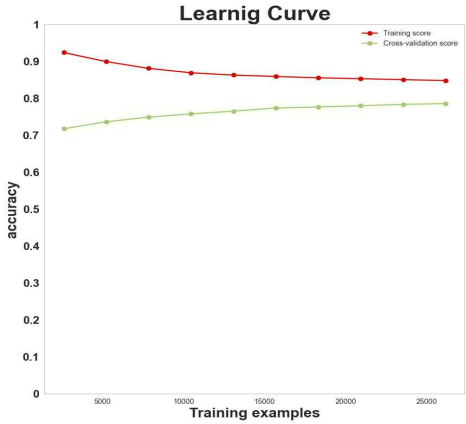


(e) Decision Tree

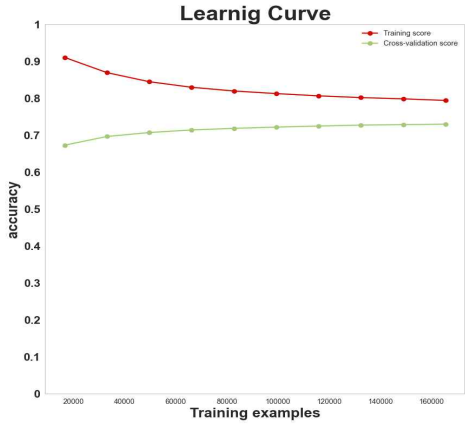


(f) Linear Regression

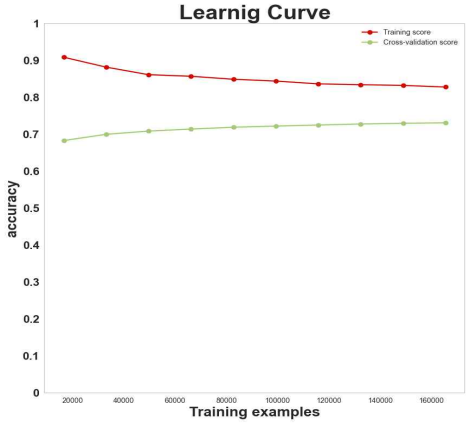
<그림 5-5> 2차 앙상블 학습 예측모형 학습곡선(KD-147모드)



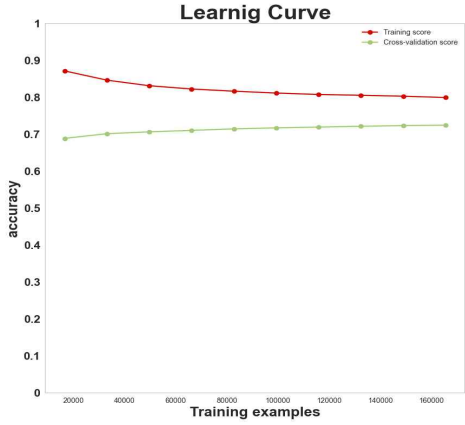
(a) CatBoost



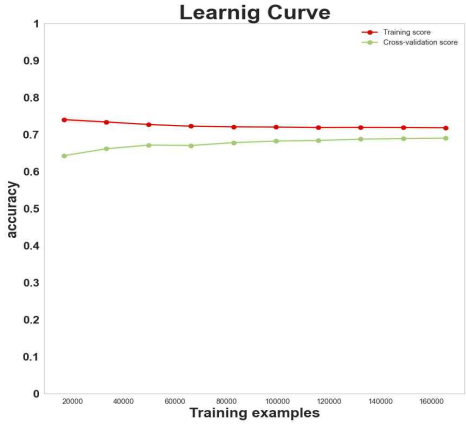
(b) LightGBM



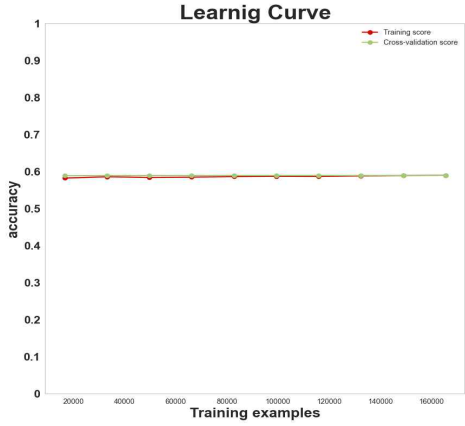
(c) XGBoost



(d) Random Forest



(e) Decision Tree



(f) Linear Regression

<그림 5-6> 2차 앙상블 학습 예측모형 학습곡선(Lug-Down3모드)

제3절 경유자동차 PM 배출요인 분석 결과

1. 변수 중요도

변수 중요도(Variable Importance)는 학습된 모형을 통해 예측시 각 변수들의 영향력을 수치화 한 것이다. 전통적인 통계기법과 달리 랜덤 포레스트, XGBoost, LightGBM, CatBoost와 같은 머신러닝기법은 많은 강점이 있음에도 불구하고, 블랙박스모형이기 때문에 종속변수와 설명변수의 설명력을 확보하기 어렵다는 단점도 있다. 반면 통계기법인 회귀분석의 경우 종속변수와 설명변수의 인과관계 해석이 용이하다. 따라서 선형회귀모형을 포함 6개 예측모형의 변수 중요도를 분석하였다. 변수 중요도라는 척도를 통해 어느 변수가 예측성능에 어느 정도 역할을 하는지를 확인이 필요하다.

먼저 변수 중요도 지표는 순열 특성 중요도(Permutation Feature Importance: PFI)를 사용한다. PFI는 변수의 특성과 실측치 간의 관계를 끊어내도록 특성값들을 교란 후 모형 예측치의 오류 증가량을 측정한다. 예컨대, 하나의 특성이 빠지고 모형의 오류가 증가한다면 모형이 예측할 때 해당 특성에 대한 의존도가 높아진다는 의미이기 때문에 중요한 특성으로 분류하고 반면 오류가 감소한다면 중요하지 않은 특성으로 분류할 수 있다. 즉, 무작위로 섞을 때 예측값이 실제값보다 얼마나 차이가 더 생겼는지를 통해 해당 변수의 영향력을 파악할 수 있다.

PFI를 변수 중요도 지표로 사용하는 이유는 첫째, 모든 모형에 적용할 수 있다. 학습모형과 데이터만 있으면 변수 중요도를 산정할 수 있기 때문에 모형의 학습과정, 내부 구조에 대한 정보가 필요 없어서 어느 모형이든 적용이 가능한 장점이 있다. 둘째, 만약 입력변수 하나를 제거한다면, 입력변수의 차원이 하나 줄어들기 때문에 모형에 입력변수를 조정하기 위해 모형을 재학습시켜야 한다. 그러나 입력변수를 노이즈화하면 입력변수의 차원이 동일하기 때문에 학습된 모형에 데이터를 입력하여

PFI를 분석할 수 있다. Permutation Feature Importance 산정 알고리즘은 <표 5-18>과 같다(Fisher et. al, 2019).

<표 5-18> Permutation Feature Importance 알고리즘 단계별 과정

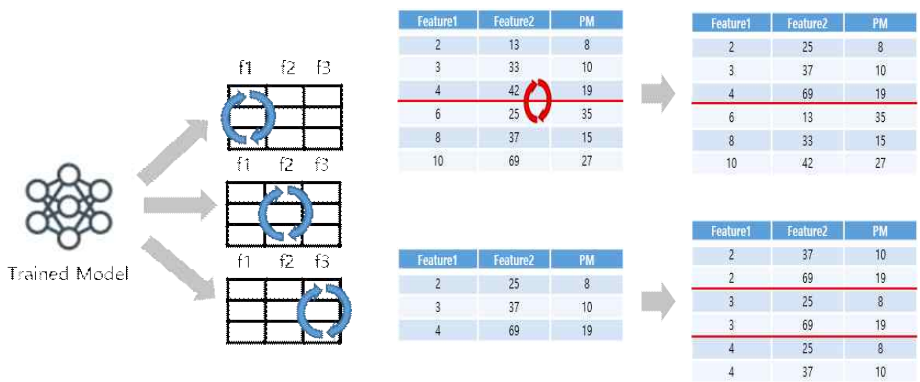
각 과정별 단계	
1단계	기본모형의 오류를 측정한다. $e^{origin} = L(y, f(X))$
2단계	각 특성 $j = 1, \cdots\cdots, p$에 대해 아래 과정을 수행한다. 2-1) 데이터 X에서 특성 j를 섞어가며 특성행렬 X^{perm}를 생성한다. 이는 특성 j와 실제 결과값 y간의 관계를 끊는다. 2-2) 섞여진 데이터 예측치를 기반으로 오류를 측정한다. $e^{perm} = L(Y, f(X^{perm}))$ 2-3) Permutation Feature Importance 계산한다. 순열 특성 중요도 FI는 아래 2가지 계산법 모두 사용할 수 있다. $FI^j = e^{perm} / e^{orig}$ (5-1) $FI^j = e^{perm} - e^{orig}$ (5-2)
3단계	FI 값에 따라 특성을 내림차순으로 정렬하고 평균값을 제시한다.

PFI의 특징은 데이터가 n개 존재할 때 무작위 경우의 수는 n(n-1)이 되므로 빅데이터의 경우 전체 데이터세트를 절반으로 나누어 나눠진 절반끼리 바꿔서(swap) 무작위(permutation)로 섞는 것을 대체한다. 또한, 오류비율이나 차이를 통해 중요도를 측정하는데 차이보다는 비율을 사용하는 것이 다른 모형과 비교 가능하기 때문에 일반적으로 식(5-2)와 같이 오차의 평균을 중요도 지표로 사용한다.

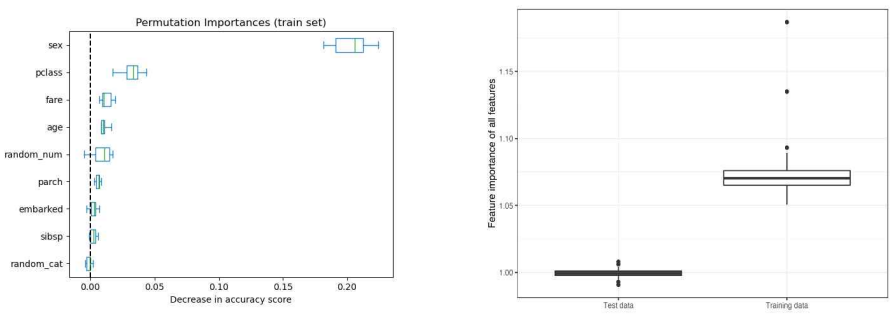
PFI의 장점은 특정 변수를 제거하고 재학습을 시키는 것이 아니라 특정 변수 하나를 섞는 것이므로 상대적으로 계산량 감소하여 계산속도가 빠르다. 또한 한 변수에 대해서만 섞기 때문에 상대적으로 일관된 PFI를

측정할 수 있다. 이는 변수의 특성과의 모든 상호작용을 고려할 수 있음을 의미한다. 다시 말해 특성을 섞으면 다른 특성과 상호작용 효과도 상쇄된다. 이는 동시에 단점이 되기도 하는데, 두 특성의 중요도 측정에 두 특성의 상호작용 중요성이 포함되기 때문이다. 그러나 특성 간 상호작용이 최소하도록 변수를 잘 선정한다면 중요도가 전체 성능에 근사할 수 있다.

a) 1단계 과정



b) 2단계 과정



c) 3단계 과정(특성별 PFI 예시)

자료 : https://christophm.github.io/interpretable-ml-book/feature-importance.html

<그림 5-7> Permutation Feature Importance 산정과정

마지막으로 PFI 산정의 딜레마는 학습 세트 또는 테스트 세트 중 무엇을 사용해야 하는지 불명확하다. 학습 세트는 모형 예측에 얼마나 각 특성에 의존했는지를 파악하는 것이고, 테스트 세트는 학습하지 않은 데이터 세트에 대한 모형성능에 어느 정도 기여하는 특성을 알고 싶은지 판단해야 하는 것이다. 어떤 세트를 선택하는 것이 더 나은 선택인지 아직 정확한 정답은 없다. 결국 예측(학습 세트)을 위해 각 특성에 얼마나 의존하는 모형인지, 보이지 않는 데이터(테스트 세트)에 대한 모형의 예측성능에 어느 정도 기여하는 특성을 알고 싶은지 판단해야 하는 것이다. 본 연구는 경유자동차의 PM 배출 예측모형에 어떤 영향요인에 의존하는지를 파악하기 위해 학습 세트의 중요도를 PM 배출요인의 영향력 척도로 사용한다.

2. PM 배출요인 중요도 분석결과

2.1 모형별 입력변수 중요도 분석결과

입력변수 중요도 분석결과는 <그림 5-18>과 <그림 5-25>와 같이 입력변수 중요도인 PFI 수치로 제시하였다. <표 5-19>~<표 5-26>에서 모형별 입력변수 중요도의 상대적 비중을 제시하였다. 그 이유는 입력변수중요도의 우선순위를 쉽게 이해하기 위함이다. 입력변수 중요도는 통계모형, 배깅모형, 부스팅모형에 따라 다소 차이를 보였다. 모형별 입력변수 중요도의 상대적 비중은 일부 모형에서 차이를 보인다. KD-147모드에서 부스팅모형과 선형회귀모형을 비교해보면 연식과 배출가스등급 변수의 PFI 비중의 차이가 많은 것으로 분석되었다. 연식의 PFI 비중은 CatBoost 11%, Random Forest 5.5% 선형회귀모형 18.1%로 산출되었다. 배출가스등급은 CatBoost 45.2%, Random Forest 57.6% 선형회귀모형 71.81%로 나타났다. 선형회귀모형의 경우 입력변수 중에서 연식과 배출가스등급 PFI 비중이 절대적으로 큰 것을 확인할 수 있었다. Lug-Down3모드 경우 연식, 배출가스등급, 배기량 순으로 비중이 큰 것으로 나타났다. 연식의 경우 PFI 비중이 LightGBM 35.2%, Desion Tree 31.8% 선형회귀모형 74.11%로 분석되었다. 배기량의 경우 CatBoost 11.7%, Random Forest 14.7% 선형회귀모형 1.7%로 산정되었다. 선형회귀모형의 입력변수 PFI 비중은 연식, 배출가스등급, 유로5 전후에만 치중되었다.

이렇듯 모형별 PFI가 부스팅모형과 배깅모형은 유사하지만 통계기법이 적용된 의사결정나무, 선형회귀모형과 일부 입력변수에서는 차이가 있음을 발견하였다. 본 연구에서는 입력변수간 상대적 PFI 비중을 모형별 PM 배출 주요인의 선별기준으로 활용한다. 선별기준은 모형별 변수중요도를 상대적 PFI 비중의 평균값을 기준으로 우선순위를 정한다.

<표 5-19> 예측모형별 입력변수 상대적 중요도 분석결과(KD-147 전차종)

구분	CatBoost	LightGBM	XGBoost	Random Forest	Decision Tree	Linear Regression	Mean
연식	11.0	11.9	13.7	5.5	9.8	18.1	11.7
배기량	5.9	5.5	5.3	4.6	5.9	0.8	4.7
주행거리	2.9	2.4	3.8	2.6	1.1	0.4	2.2
배출가스등급	45.2	48.9	46.7	57.6	49.5	71.8	53.3
총중량	13.5	13.0	12.2	10.9	10.9	0.1	10.1
적재중량	1.5	1.0	1.5	2.8	2.4	1.3	1.8
승차정원	0.9	0.5	0.4	0.2	0.1	0.8	0.5
길이	2.3	1.5	1.7	0.7	0.8	1.7	1.4
넓이	3.8	2.5	1.9	1.4	2.2	0.1	2.0
높이	3.1	1.8	2.7	2.1	2.7	0.5	2.1
연비	3.7	2.9	2.8	1.5	1.9	1.4	2.3
저감장치 유무	1.3	1.4	1.4	0.9	0.8	1.8	1.3
유로5 전후	4.7	6.5	5.9	9.3	11.9	1.0	6.5
사업/비사업	0.1	0.1	0.1	0.0	0.0	0.2	0.1
합계	100.0	100.0	100.0	100.0	100.0	100.0	100.0

<표 5-20> 예측모형별 입력변수 상대적 중요도 분석결과(KD-147 승용)

구분	CatBoost	LightGBM	XGBoost	Random Forest	Decision Tree	Linear Regression	Mean
연식	14.9	18.2	11.3	4.8	6.7	5.3	10.2
배기량	6.9	5.9	4.9	5.4	8.1	0.6	5.3
주행거리	4.0	4.3	4.5	2.8	1.8	0.0	2.9
배출가스등급	30.7	46.8	55.4	64.6	53.1	77.5	54.7
총중량	13.5	14.3	10.3	12.9	15.8	0.2	11.2
적재중량	0.0	0.0	0.0	0.0	0.0	0.0	0.0
승차정원	1.3	0.3	0.2	0.1	0.1	6.6	1.4
길이	5.1	2.8	2.2	1.9	3.2	1.0	2.7
넓이	4.0	2.0	4.1	1.3	2.3	0.9	2.4
높이	4.4	1.3	1.2	1.5	6.0	1.3	2.6
연비	5.3	3.3	4.2	1.5	2.3	5.3	3.6
저감장치 유무	0.9	0.7	0.8	0.8	0.6	0.8	0.8
유로5 전후	8.8	0.1	0.9	2.5	0.1	0.2	2.1
사업/비사업	0.0	0.0	0.0	0.0	0.0	0.4	0.1
합계	100.0	100.0	100.0	100.0	100.0	100.0	100.0

<표 5-21> 예측모형별 입력변수 상대적 중요도 분석결과(KD-147 승합)

구분	CatBoost	LightGBM	XGBoost	Random Forest	Decision Tree	Linear Regression	Mean
연식	25.1	55.3	37.2	50.6	58.7	76.8	50.6
배기량	5.4	3.8	4.8	3.4	1.6	2.6	3.6
주행거리	9.9	9.9	15.9	8.0	1.4	0.4	7.6
배출가스등급	22.3	8.7	13.1	17.9	21.8	5.1	14.8
총중량	4.6	3.7	6.3	2.8	0.9	5.8	4.0
적재중량	0.0	0.0	0.0	0.0	0.0	0.0	0.0
승차정원	7.2	3.0	3.1	4.7	6.4	0.9	4.2
길이	2.9	3.6	2.2	1.6	2.1	0.5	2.2
넓이	4.5	1.7	4.0	1.5	0.8	1.0	2.3
높이	2.6	1.9	2.5	1.7	0.6	0.3	1.6
연비	7.3	5.6	6.0	4.7	3.0	1.7	4.7
저감장치 유무	3.6	2.7	2.7	3.0	2.5	4.3	3.1
유로5 전후	4.5	0.0	2.1	0.0	0.0	0.1	1.1
사업/비사업	0.1	0.1	0.1	0.0	0.0	0.5	0.1
합계	100.0	100.0	100.0	100.0	100.0	100.0	100.0

<표 5-22> 예측모형별 입력변수 상대적 중요도 분석결과(KD-147 화물)

구분	CatBoost	LightGBM	XGBoost	Random Forest	Decision Tree	Linear Regression	Mean
연식	20.3	27.7	36.0	20.2	14.3	35.6	25.7
배기량	7.2	8.5	7.6	8.8	7.3	3.5	7.1
주행거리	2.6	6.0	9.3	2.9	1.0	0.5	3.7
배출가스등급	54.6	36.5	26.6	56.2	65.9	51.6	48.6
총중량	2.5	4.3	3.4	3.2	3.7	0.0	2.9
적재중량	1.2	1.2	1.4	1.3	1.6	3.7	1.7
승차정원	0.6	0.4	0.5	0.1	0.1	0.1	0.3
길이	1.2	2.5	2.2	0.8	0.7	0.1	1.2
넓이	2.7	2.1	2.5	2.0	2.1	0.3	1.9
높이	1.1	2.3	2.5	0.8	1.0	0.1	1.3
연비	2.5	3.6	3.8	1.6	1.0	0.2	2.1
저감장치 유무	2.3	2.5	2.5	1.9	1.5	3.3	2.3
유로5 전후	0.8	2.2	1.4	0.0	0.0	0.7	0.8
사업/비사업	0.2	0.3	0.4	0.1	0.0	0.3	0.2
합계	100.0	100.0	100.0	100.0	100.0	100.0	100.0

<표 5-23> 예측모형별 입력변수 상대적 중요도 분석결과(Lug-Down3 전차종)

구분	CatBoost	LightGBM	XGBoost	Random Forest	Decision Tree	Linear Regression	Mean
연식	28.2	35.2	36.7	29.5	31.8	74.1	39.2
배기량	11.7	13.3	9.8	14.7	13.6	1.7	10.8
주행거리	4.0	5.4	4.7	3.3	1.0	0.0	3.1
배출가스등급	20.5	9.3	18.8	28.9	28.2	9.8	19.3
총중량	3.6	5.3	6.3	4.5	5.7	1.7	4.5
적재중량	4.4	4.8	4.0	3.3	3.0	0.0	3.3
승차정원	1.0	1.2	0.8	1.0	1.3	0.0	0.9
길이	5.7	5.2	3.1	3.5	4.3	0.1	3.6
넓이	5.5	5.5	3.6	3.4	3.4	0.5	3.7
높이	3.9	4.9	3.9	2.2	1.3	0.0	2.7
연비	4.8	5.6	4.8	3.3	4.5	0.2	3.9
저감장치 유무	3.1	1.6	1.9	1.4	0.8	0.0	1.5
유로5 전후	0.7	1.6	0.5	0.0	0.3	11.9	2.5
사업/비사업	2.8	1.1	1.2	1.0	0.7	0.0	1.1
합계	100.0	100.0	100.0	100.0	100.0	100.0	100.0

<표 5-24> 예측모형별 입력변수 상대적 중요도 분석결과(Lug-Down3 특수)

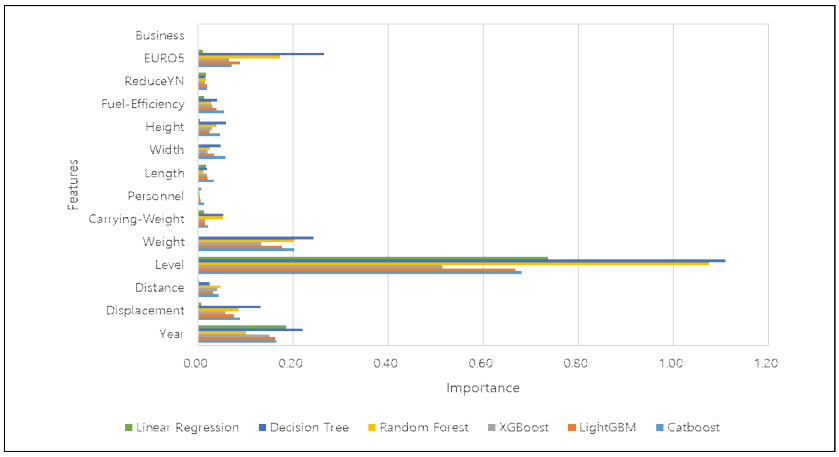
구분	CatBoost	LightGBM	XGBoost	Random Forest	Decision Tree	Linear Regression	Mean
연식	32.2	29.6	29.3	21.5	20.9	49.8	30.5
배기량	3.6	10.8	7.2	11.1	10.9	1.0	7.4
주행거리	6.7	3.7	10.4	2.9	1.5	0.2	4.2
배출가스등급	28.7	35.3	28.9	44.9	43.1	16.6	32.9
총중량	2.7	2.1	3.1	2.9	4.1	0.2	2.5
적재중량	6.0	4.5	4.1	5.9	6.5	0.0	4.5
승차정원	1.5	0.1	0.1	0.0	0.0	0.4	0.4
길이	3.1	1.1	0.9	1.3	1.2	14.4	3.6
넓이	3.3	2.6	2.9	2.9	3.7	1.4	2.8
높이	3.7	2.9	3.6	1.6	1.6	0.1	2.2
연비	5.1	4.7	6.2	3.6	5.0	6.4	5.2
저감장치 유무	0.8	0.3	0.7	0.3	0.3	0.0	0.4
유로5 전후	0.8	0.4	0.9	0.1	0.5	9.5	2.0
사업/비사업	1.9	2.0	1.7	1.1	1.0	0.0	1.3
합계	100.0	100.0	100.0	100.0	100.0	100.0	100.0

<표 5-25> 예측모형별 입력변수 상대적 중요도 분석결과(Lug-Down3 승합)

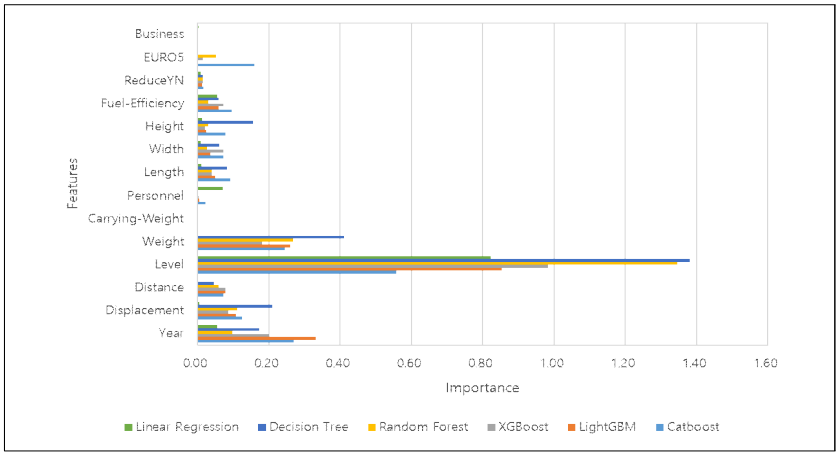
구분	CatBoost	LightGBM	XGBoost	Random Forest	Decision Tree	Linear Regression	Mean
연식	37.6	39.0	36.5	47.2	57.8	60.1	46.4
배기량	9.0	11.3	10.4	16.4	17.8	8.0	12.2
주행거리	11.7	10.9	13.2	11.1	6.3	0.0	8.9
배출가스등급	3.3	3.4	4.9	0.9	0.0	6.5	3.2
총중량	8.3	6.6	8.3	4.8	0.6	0.1	4.8
적재중량	0.1	0.2	0.2	0.1	0.0	0.0	0.1
승차정원	6.3	6.3	5.2	4.6	2.5	0.4	4.2
길이	5.6	3.1	2.7	1.5	1.2	1.9	2.7
넓이	0.4	0.5	0.7	0.3	0.0	1.0	0.5
높이	3.9	3.9	4.0	2.0	1.6	2.6	3.0
연비	9.2	11.0	9.6	8.3	10.7	1.9	8.5
저감장치 유무	3.4	2.8	2.6	2.3	1.2	0.0	2.0
유로5 전후	0.2	0.4	0.7	0.0	0.0	17.5	3.1
사업/비사업	0.8	0.7	1.1	0.4	0.4	0.0	0.6
합계	100.0	100.0	100.0	100.0	100.0	100.0	100.0

<표 5-26> 예측모형별 입력변수 상대적 중요도 분석결과(Lug-Down3 화물)

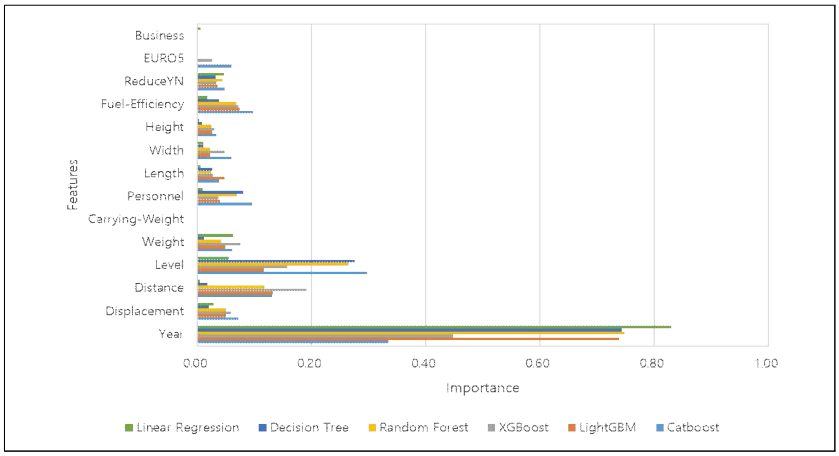
구분	CatBoost	LightGBM	XGBoost	Random Forest	Decision Tree	Linear Regression	Mean
연식	25.0	38.9	43.1	24.8	22.8	76.6	38.5
배기량	8.6	7.8	5.7	14.0	13.5	0.4	8.4
주행거리	5.3	5.0	5.8	4.0	4.1	0.2	4.1
배출가스등급	17.5	9.0	9.8	27.3	25.5	6.7	16.0
총중량	6.5	5.4	6.0	6.1	8.5	4.7	6.2
적재중량	5.9	5.5	4.4	3.5	4.2	0.5	4.0
승차정원	0.8	0.4	0.2	0.3	0.6	0.0	0.4
길이	5.7	6.5	5.2	3.4	5.0	0.2	4.3
넓이	5.0	5.6	5.1	5.1	5.8	0.6	4.5
높이	5.8	5.6	4.6	3.3	2.6	0.0	3.6
연비	6.2	6.6	6.8	5.1	5.6	0.0	5.1
저감장치 유무	3.9	1.9	2.1	1.9	1.2	0.0	1.8
유로5 전후	0.5	0.9	0.1	0.1	0.0	10.1	2.0
사업/비사업	3.1	1.0	1.1	1.1	0.7	0.0	1.2
합계	100.0	100.0	100.0	100.0	100.0	100.0	100.0



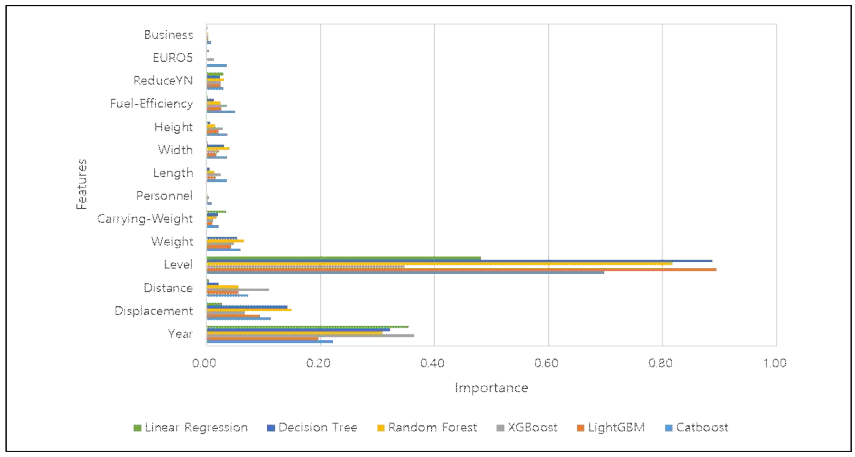
<그림 5-8> KD-147모드 전차종 PM 배출요인 중요도(PFI)



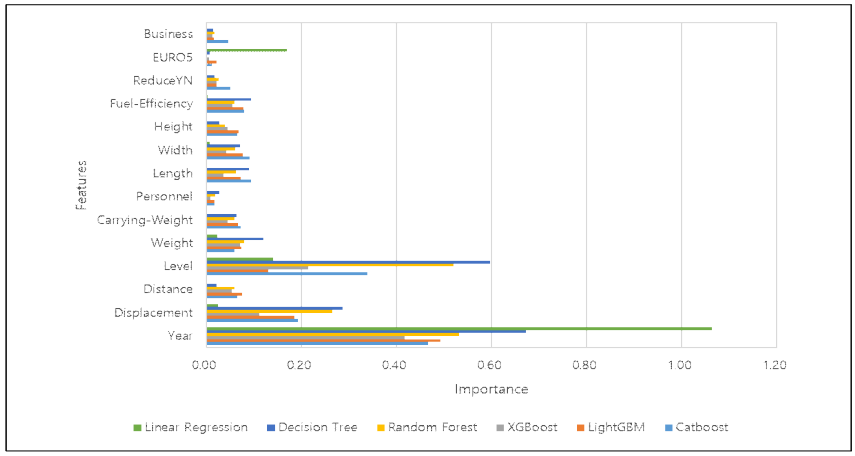
<그림 5-9> KD-147모드 승용차 PM 배출요인 중요도(PFI)



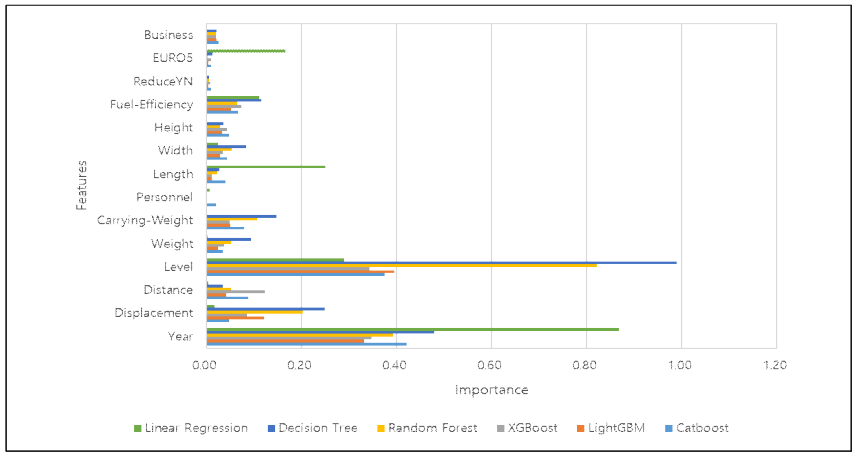
<그림 5-0> KD-147모드 승합차 PM 배출요인 중요도(PFI)



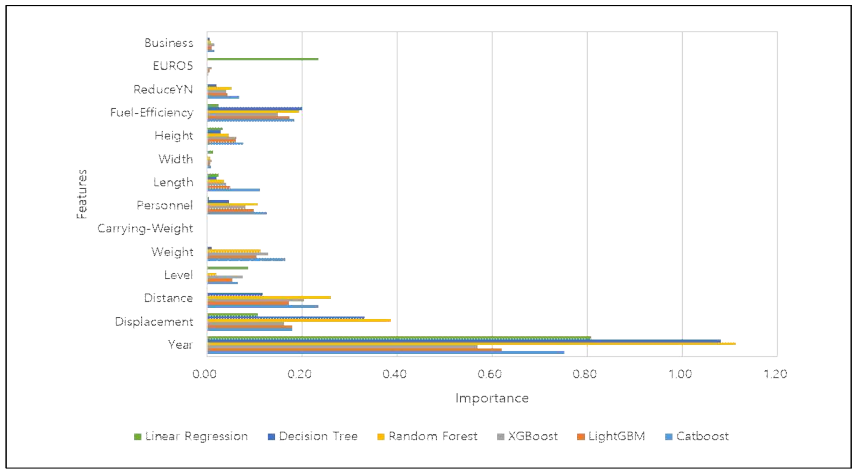
<그림 5-11> KD-147모드 화물차 PM 배출요인 중요도(PFI)



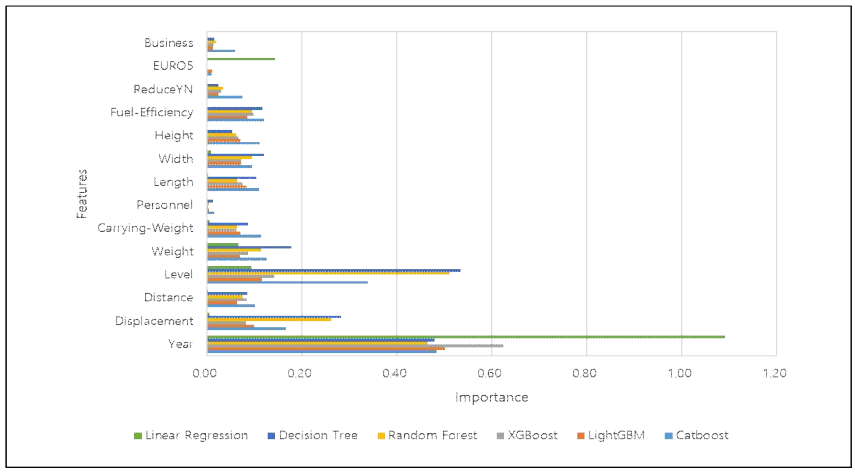
<그림 5-12> Lug-Down3모드 전차종 PM 배출요인 중요도(PFI)



<그림 5-13> Lug-Down3모드 특수차 PM 배출요인 중요도(PFI)



<그림 5-14> Lug-Down3모드 승합차 PM 배출요인 중요도(PFI)



<그림 5-15> Lug-Down3모드 화물차 PM 배출요인 중요도(PFI)

2.2 모형별 PM 배출요인 선정

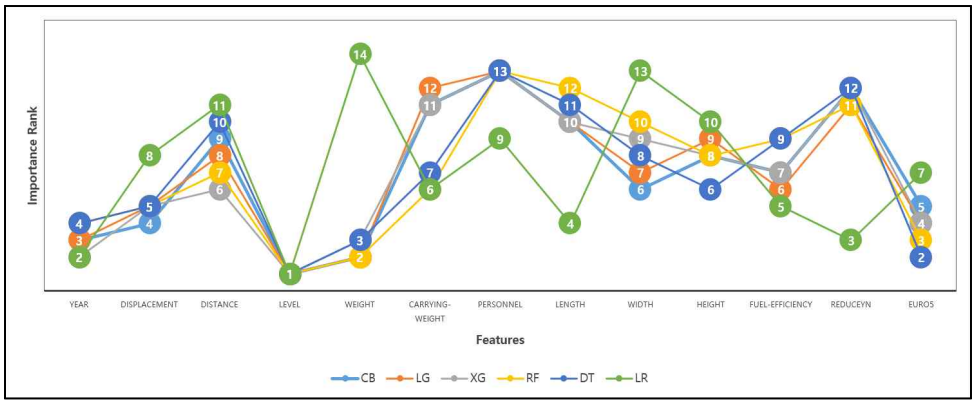
PM 배출 주요인은 앙상블 학습 기반으로 배출가스검사 합격차량 데이터만을 활용한 예측모형에서 도출되었다. 입력변수의 중요도는 배출가스 검사방식과 차종에 따라 일부 다르게 분석되었다. 먼저 검사방식에 상관없이 공통적인 PM 배출 주요인은 배출가스등급, 연식, 배기량, 총중량으로 분석되었다. PM 배출요인의 차이점은 검사방식에 따라 다르게 나타

났다. KD-147모드는 배출가스등급이 가장 중요도가 높았으나 Lug-Down3모드는 연식의 중요도가 1위를 차지하였다. 차종별 차이점은 특수차는 적재중량, 승합차는 승차인원과 연비가 선정되었다. 이는 특수차는 적재를 목적으로 운행하는 차량이 대부분이고, 승합차는 적재보다 다인 승객 운행이 주된 목적에서 기인한 것으로 사료된다.

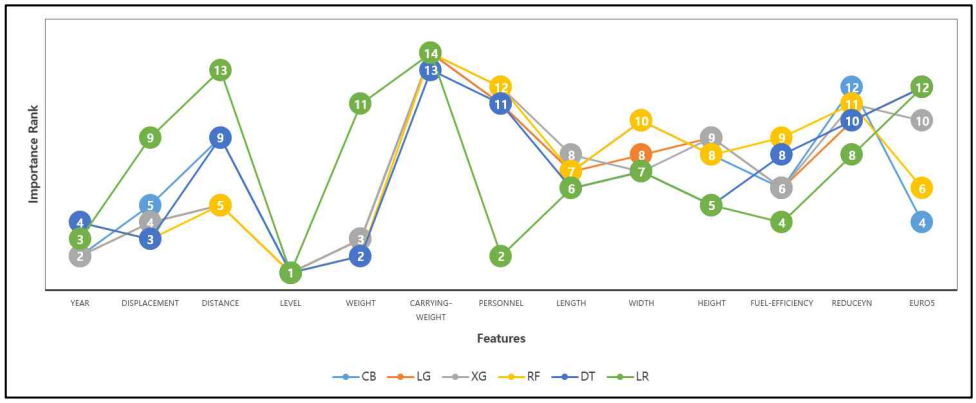
<표 5-27> 차종별 경유자동차 PM 배출 주요인 분석결과

검사방식	순위	전차종	승용	승합	화물
KD-147	1	배출가스등급	배출가스등급	연식	배출가스등급
	2	연식	총중량	배출가스등급	연식
	3	총중량	연식	주행거리	배기량
	4	EURO5	배기량	연비	주행거리
	5	배기량	연비	승차인원	총중량
검사방식	순위	전차종	특수	승합	화물
Lug-Down3	1	연식	배출가스등급	연식	연식
	2	배출가스등급	연식	배기량	배출가스등급
	3	배기량	배기량	주행거리	배기량
	4	총중량	연비	연비	총중량
	5	연비	적재중량	총중량	연비

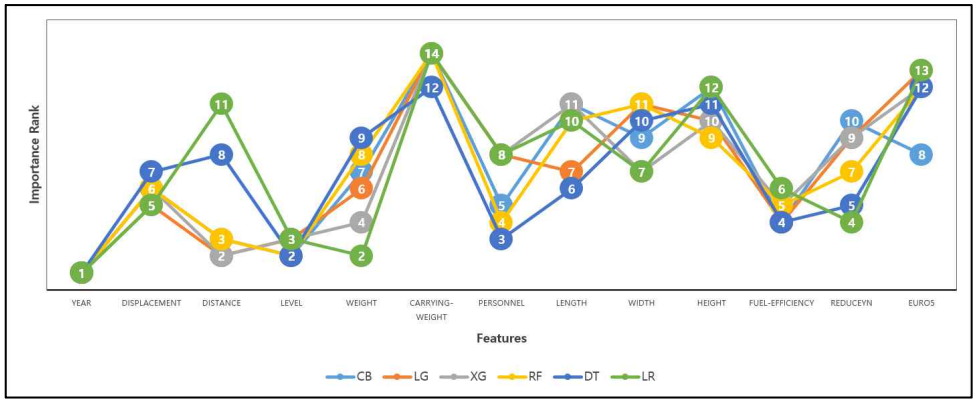
앞서 경유자동차 PM 배출 예측모형에서 도출된 입력변수의 평균 PFI 비중을 기준으로 PM 배출요인의 우선순위를 선정한다. <그림 5-26>~<그림 5-33>을 살펴보면 부스팅모형의 입력변수 중요도 순위는 유사한 패턴을 보이고 있으나 의사결정나무, 선형회귀모형은 다른 패턴이 발견되었다. 선형회귀모형의 차종별 PM 배출요인 보면 KD-147모드 승용차의 경우 승차인원, 연비, EURO5가 주요 순위에 포함되어 있고, Lug-Dow3모드 화물차는 EURO5, 적재중량이 주요 변수로 선정하였다.



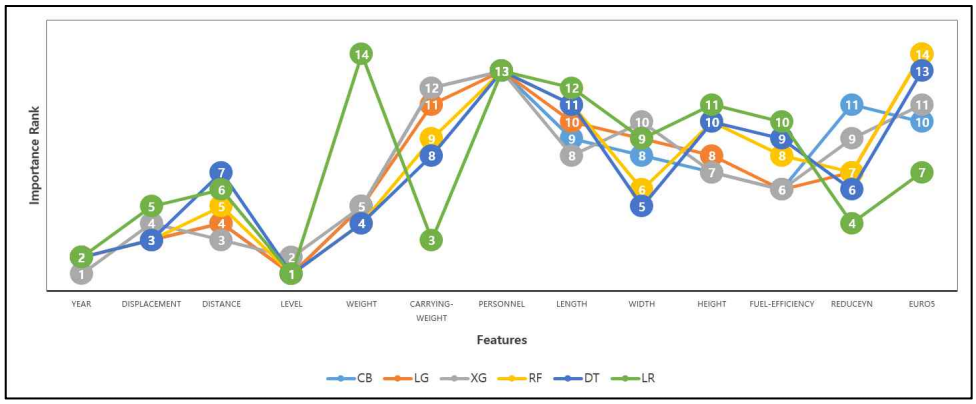
<그림 5126> KD-147모드 전차종 PM 배출요인 순위



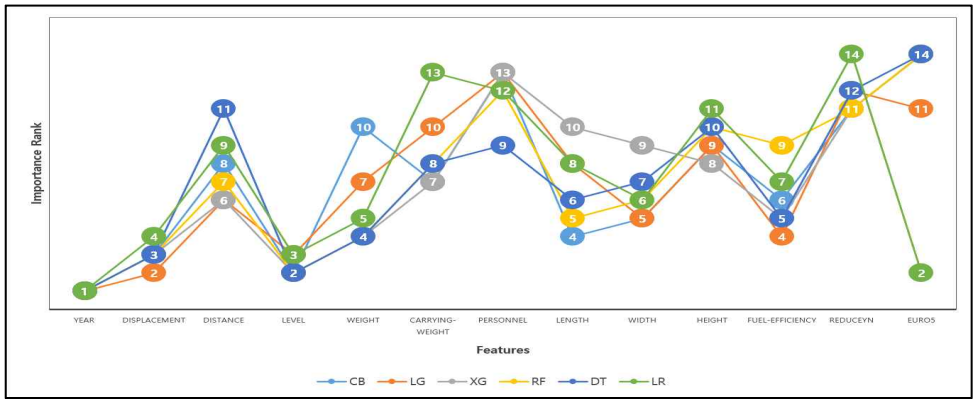
<그림 5-17> KD-147모드 승용차 PM 배출요인 순위



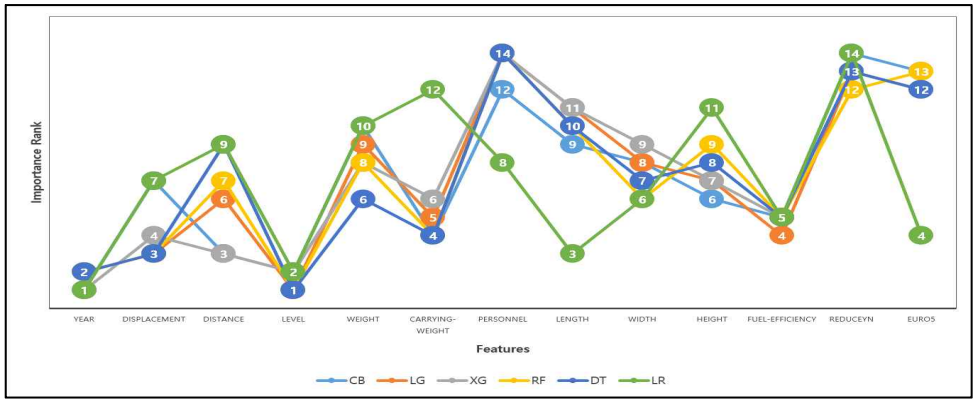
<그림 5-18> KD-147모드 승합차 PM 배출요인 순위



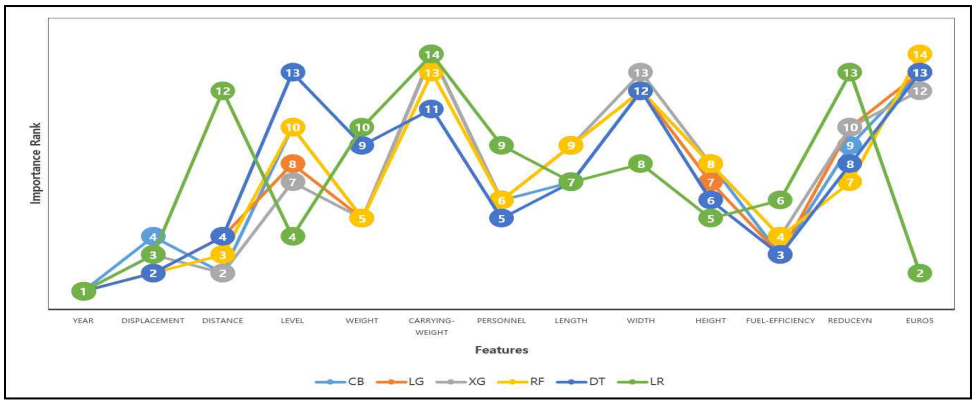
<그림 5-19> KD-147모드 화물차 PM 배출요인 순위



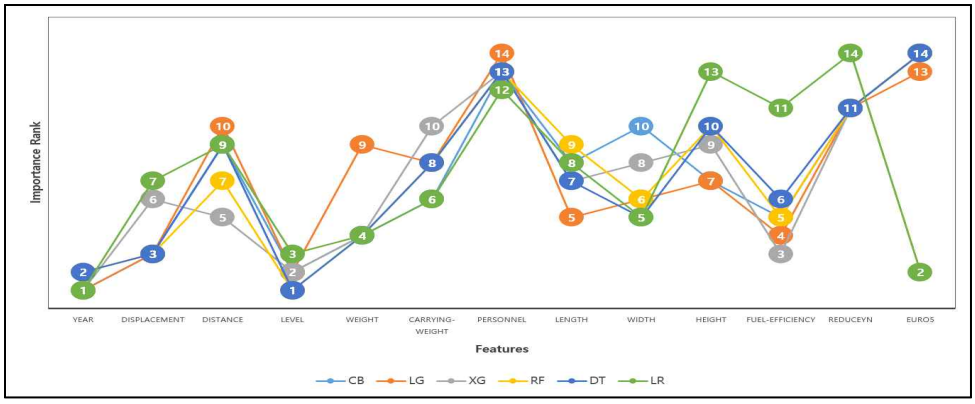
<그림 5-20> Lug-Down3모드 전차종 PM 배출요인 순위



<그림 5-21> Lug-Down3모드 특수차 PM 배출요인 순위



<그림 5-22> Lug-Down3모드 승합차 PM 배출요인 순위



<그림 5-23> Lug-Down3모드 화물차 PM 배출요인 순위

3. 선행연구와 경유자동차 PM 배출요인 비교 분석

경유자동차 PM 배출 예측모형에서 확인된 PM 배출 주요인은 차량 내부요인이다. 따라서 제2장 선행연구 고찰에서 제시한 경유자동차 PM 배출요인 중 차량 내부요인만을 한정해서 본 연구와 비교하였다. 본 연구는 배출가스검사,배출가스등급, 배기량, 연식, 총중량, 총주행거리 등 PM 배출 주요원인으로 도출되었다. 통계기법과 앙상블 학습에서 도출된 경유자동차 PM 배출요인은 다소 차이점을 발견하였다. 한정헌(2022)와 국산/외산 변수를 제외하고 동일한 변수를 적용하였으나 배기량, 주행거리, 배출가스등급, 연비 요인은 다르게 도출되었다. 배출가스저감장치가 주요인은 선정된 연구는(Geller et al., 2006; Bergmann et al., 2009; Karjalainen et al., 2014; 한진석, 2020) 많은 편이나 본 연구는 배출가스저감장치 변수가 주요인에서 제외되었다. 이는 제3장에서 확인한바 배출가스저감장치 장착 차량 중 고농도 배출 차량의 비중이 일부 차지하고 있기 때문이다.

<표 5-28> 경유자동차 PM 배출의 차량 내부요인 선행연구 비교

연구	경유자동차 PM 배출요인
Geller et al.(2006)	배출가스저감장치
Bikas and Zervas(2007)	주행거리, 규제기준
Bergmann et al.(2009)	배출가스저감장치
Bukowiecki et al (2010)	차량중량
Karjalainen et al.(2014)	배출가스저감장치
Krecl et al. (2018)	차량상태, 운전자 운전습관 및 주행패턴
Suleiman et al.(2019)	차량중량
한진석(2020)	차량업종, 등록지, 연식, 총주행거리, 배출가스저감장치
Xu et al.(2020)	운전자연령, 차량연식, 배출가스, 규제기준, 차량조건
한정헌(2022)	차량연식, 차량중량, 영업용, 외산, 등록지역
본 연구	**배출가스등급, 연식, 총중량, 배기량, EURO5 전후, 연비, 총주행거리**

Ⅵ. 사례분석

사례분석의 주요 목적은 본 연구의 앙상블 학습 예측모형에서 도출된 경유자동차 PM 배출의 주요인을 미세먼지 절감 및 환경 관련 정책에 적용하기 위함이다. 현재 친환경정책의 일환으로 환경개선부담금 제도 시행하고 있다. 그러나 현재 환경개선부담금 산정원칙과 방식 등 다방면으로 문제가 발생하고 있는 실정이다. 이러한 문제점을 개선하기 위해 본 연구에서는 PM 배출요인과 요인별 중요도 분석 결과를 사례분석에 활용하기로 한다. 이 연구 결과가 환경개선부담금 제도 개선방안에 적용된다면 정책적 연구 기여도 측면에서 일조할 것으로 예상된다.

사례분석 과정은 먼저 환경개선부담금 제도 현황을 검토하고, 문제점을 분석한다. 그리고 개선방안을 제안하고, 분석 대안 및 시나리오를 설정한다. 대안별 시나리오에 따라 다양한 정책 사례분석을 통해 실용성과 활용성이 높은 분석 결과를 제시하도록 한다. 마지막으로 사례분석에 대한 종합평가를 수행한다.

제1절 환경개선부담금 정책 검토

1. 환경개선부담금 제도의 도입 배경

환경개선부담금 제도 도입 배경은 환경오염의 원인자로 하여금 환경개선에 필요한 비용을 부담하게 하여 환경개선을 위한 투자재원을 합리적으로 조달함으로써 국가의 지속적인 발전의 기반이 되는 쾌적한 환경으로 조성하는 데 이바지하는 것으로 목적으로 1991년 ‘환경개선 중기 종합계획’에 의거 도입되었다(환경부, 2015).

일반적으로 정부의 환경문제 해결을 위한 정책은 직접규제과 간접규제로 구분할 수 있다. 직접규제는 오염물질의 배출허용기준 위반에 대한 지도·단속 등과 같이 획일적으로 강제하는 방식을 말하며, 간접규제는

쓰레기 종량제와 같이 경제적 동기부여를 통해 오염원인자로 하여금 자발적으로 오염저감을 하도록 유도하는 방식을 의미한다. 환경개선부담금 제도는 간접규제의 일환으로 오염인자에게 오염물질 처리비용을 부담토록 하여 오염저감을 유도하고, 환경투자재원을 안정적으로 확보하기 위한 제도이다. 부과대상은 “건물 각층 바닥면적 합계가 160㎡이상인 시설물과 「자동차관리법」에 의해 등록된 경유자동차”로서 3월과 9월 연 2회 부과되고 있다(이소영·조현구, 2017).

2. 환경개선부담금 제도의 주요 내용

2.1 적용대상

현재 환경개선부담금 부과대상은 경유를 연료로 사용하는「자동차관리법」에 따라 등록된 자동차에 부과하고 있다. 반면 부과감면대상은 저공해자동차, 유로5 및 유로6 경유자동차, 국가유공자나 장애인 자동차 1대, 저감장치 부착자동차(3년), 저공해엔진으로 개조 또는 교체한 자동차 등에 대해 부담금을 면제해주고 있다.

환경개선비용 부담법(면제대상 자동차 등에 관한 규정)

제4조(저공해자동차 등 대한 면제)
제2조제1호의 "저공해자동차" 및 "유로5 경유차"와"유로6 경유차"에 대하여는 환경개선부담금을 면제함

제5조(배출가스저감장치 부착자동차 등에 대한 면제)
저감장치를 부착하거나 저공해엔진으로 개조 또는 교체한 자동차에 대하여는 다음의 기간 동안 환경개선부담금을 면제한다.
1. 저감장치 부착 자동차 : 저감장치의 보증기간
2. 저공해엔진으로 개조 또는 교체한 자동차 : 개조 또는 교체의 지속기간

미세먼지 계절관리제 기간 동안
환경부는 배출가스 5등급 경유차량이 운행하지 못하는 기간 동안 차량 등록지를 기준으로 환경개선부담금을 감면

부과경감 대상은 배기량 3,000cc 이하 일반형 소형 화물자동차 중 최대적재량이 800kg 이상인 화물자동차의 기준부과금액(20,250원→15,190원)을 사실상 100분의 25로 경감해주고 있다. 단, 「화물자동차 운송사업법」에 사용되는 자동차 등은 제외한다(환경부, 2015).

2.2 적용기준

유로(EURO)는 자동차의 배기가스 배출량을 줄이기 위해 유럽 연합에서 시행하고 있는 규제기준으로 현재는 유로6가 적용되고 있다. <표 6-1>과 같이 적용기준을 살펴보면 다음과 같은 특징이 있다.

유로1은 1992년에 시행한 최초의 유럽 배출가스 규제했으며, 우리나라는 1994년부터 적용하였다. 상용차는 연료를 전자제어로 분사함으로써 불필요한 연료소모를 막아 배기가스를 줄여주고, 공통적으로는 질소산화물 생성을 줄여주는 EGR 밸브를 개발해 장착하였다 유로2는 유로1배출규제만 강화됐을 뿐, 처리방식은 별반 차이가 없었다. 그러나 정부차원에서 DPF 개조를 권장해서 DPF가 장착된 차량들이 유로 2부터 많아지기 시작했다. 우리나라에서는 유로3 이하 2006년 이전에 생산된 경유자동차는 배출가스 5등급으로 분류되며, 운행제한(비상저감조치 발령일, 12월~익년 3월 수도권 계절관리제 기간, 녹색교통지역) 위반시 과태료를 부과하였다. 그나마 일부 차종은 정부 차원에서 DPF 장착을 지원해주었으나 상용차 중에 사제로 장착한 DPF는 관리를 안해서 제 기능을 못 하는 경우가 많았다. 유로5까지는 SCR 방식인 배기가스 라인 쪽에서 분사하는 방식이 적용됐지만, 유로6부터는 차량은 UWS 방식인 요소수를 엔진 실린더에 직접 분사하는 방식이 적용되었다.

유로1~유로6까지 시간이 경과하면서 배출가스 규제기준은 더 엄격해졌고, 환경개선부담금의 감면을 받는 유로5이상 차량은 SCR과 요소수 방식을 사용하는 반면 유로4이하에서 EGR 방식을 사용하는 것이 차별성을 지닌다. 국내에 시판되는 경유자동차 중 약 2012년 이후 생산된 경유자동차부터는 환경개선부담금이 면제되고 있다.

<표 6-1> 경유자동차(승용) EURO1~6 배출가스 규제기준

단위 : mg/km

구분		EURO1	EURO2	EURO3	EURO4	EURO5	EURO6
저감장치		DPF + EGR		DPF/CRDi + EGR	DPF/DOC + EGR	SCR + UWS	SCR + UWS (직접분사)
적용시점	유럽	1992	1996	2000	2005	2009	2014
	한국	1994	2000	2005	2008	2011	2015
CO		2720	1000	660	500	500	500
HC + NO_X		970	700	560	300	230	170
PM		140	80	50	25	5	4.5

주: DPF(Diesel Particulate Filter, 디젤입자상물질 여과장치)
EGR(Exhaust Gas Recirculation, 배기가스 재순환장치)
CRDi(Common Rail Direct Injection, 커먼레일 연료분사장치)
EGR(Exhaust Gas Recirculation, 배기가스 재순환장치)
DOC(Diesel Oxidation Catalyst, 디젤 산화촉매기)
SCR(Selective Catalytic Reduction, 선택적 촉매환원장치)
UWS(Urea water solution, 요소수 방식)

2.3 환경개선부담금 산정방식

환경개선부담금은 부과대상 차종에 대해 환경개선비용 부담법(제10조)의 적용계수를 살펴보면 대당 기본부과금액, 오염유발계수, 차령계수, 지역계수를 적용하고 있다. 여기서 대당 기본부과금액은 환경개선부담금의 기준 부과금액인 20,250원에서 2.102(2022년도 부과금 산정지수)를 곱하여 산정하고 있으며, 부과금 산정지수는 매년 환경부에서 고시하고 있다. 부과기준일(6.30, 12.31) 기준으로 현재 자동차 등록원부상 소유자에게 연간 2회 부과하고 있다.

환경개선비용 부담법 제10조(개선부담금의 산정기준)

제9조제2항에 따른 자동차에 대한 개선부담금은 다음의 계산식에 따라 산정한다.

대당(臺當) 기본 부과금액 × 오염유발계수 × 차령계수(車齡係數) × 지역계수 × 연간 부과횟수(2회)

제2항에 따른 대당 기본 부과금액, 오염유발계수, 차령계수 및 지역계수는 대통령령으로 정한다.

제2절 환경개선부담금 제도 문제점

1. 환경개선부담금 수입 현황

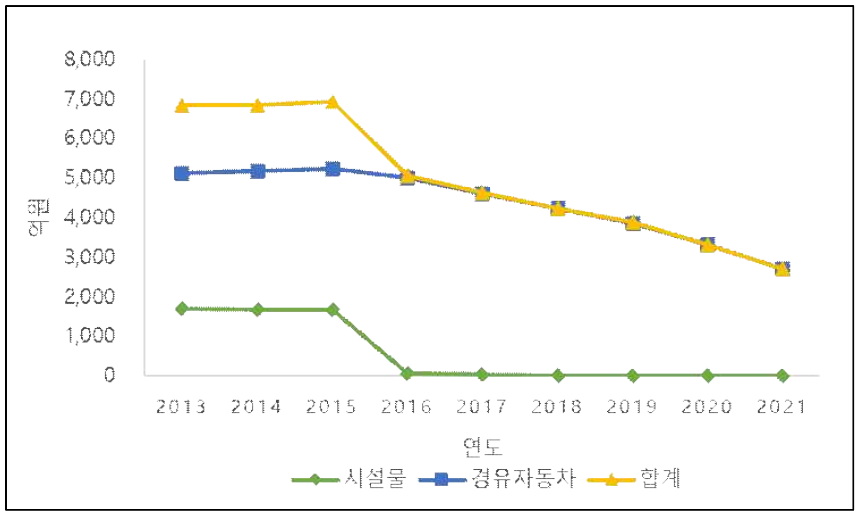
환경개선부담금은 시설물과 경유자동차을 대상으로 징수되어 환경개선특별회계에 귀속되었다. <표 6-2>와 같이 환경개선부담금은 2013년 6,840억원에서 2021년 2,713억원으로 4,127억원이 감소되었다. 2016년 5,062억원이 징수되어 환경부 부담금 총수입(2016년 2조 7,080억원)의 20% 내외를 차지한다(기획재정부 2017). 지자체에 대한 징수교부액(10%)을 제외하고 환경개선특별회계에 귀속되어 대기 및 수질환경개선사업비와 개발연구비 등으로 사용되고 있다. 그러나 시설물부담금은 2015년 하반기에 폐지되어 과년도 체납액만을 징수하고 있으며, 경유차 부담금 역시 2016년에는 감소 추세로 전환되었다. 이는 2012년 이후 생산된 경유자동차부터 환경개선부담금이 면제되고 있고, 전기차 및 수소차와 같은 친환경 자동차의 보급량은 증가추세이므로 환경개선부담금의 수입 지속적으로 감소될 것으로 예상된다.

<표 6-2> 환경개선부담금 수입 추이

단위 : 백만원

구분	2013년	2014년	2015년	2016년
시설물	171,583	167,782	168,360	6,706
경유자동차	512,393	517,113	523,662	499,542
합계	683,976	684,895	692,022	506,248
구분	2018년	2019년	2020년	2021년
시설물	1,427	884	698	552
경유자동차	422,207	386,877	331,039	270,759
합계	423,634	387,761	331,737	271,311

자료 : 기획재정부(2014~2022)



<그림 6-1> 환경개선부담금 수입 추이

2 환경개선부담금 산정기준 및 방식 검토

기존의 환경개선부담금 부과계수를 살펴보면 배기량을 토대로 부과되는 오염유발계수, 도시의 인구 규모로 부과되는 지역계수, 자동차의 차량 연도로 부과되는 차령계수로 구성되고 있다. 이 같은 산정기준 및 방식이 합리적이고, 현실을 반영하면서 적합한지를 비교 검토하고자 한다.

2.1 오염유발계수

오염유발계수를 살펴보면 배기량이 높아질수록 유발계수는 급격히 증가한다. 배기량 2000cc를 1.0을 기준으로 비교해보면 2,500cc는 1.25배, 3,500cc는 1.75배 등의 배출량이 발생되는 것을 의미한다. <그림 6-2>에서 보는 바와 같이 오염유발계수는 배기량이 증가할수록 적용계수가 높아지고 있다.

국립환경과학원(2013)에 의하면 차량의 배출가스 부피 유량을 종속변수로 두고 독립변수로서 배기량, 차량 출력, 차량 중량을 각각 선정하여 각 조합의 상관성을 각각 산출하였다 분석 결과 세 변수가 모두 유의한

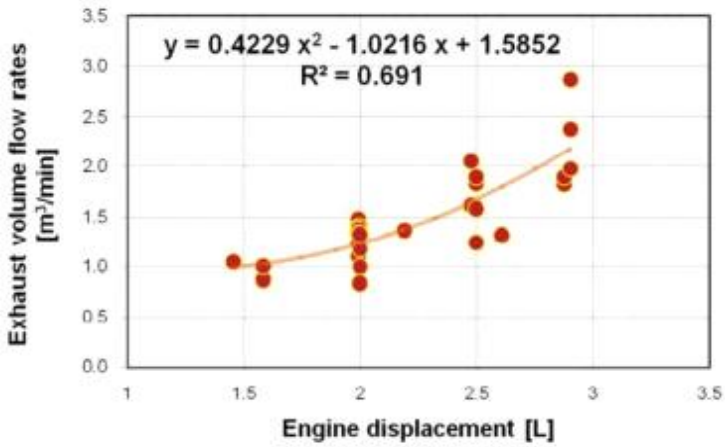
결과를 보였으며 그 중 배기량의 설명 계수가 가장 높게 나타났다.

즉, 배기량이 높아질수록 배출가스 양은 많아지는 것으로 나타나 상관관계가 높은 것을 알 수 있다. 그러나 계수가 3,500cc 초과부터 급격히 증가하는 특성을 보이고 있어 이에 대한 검토가 요구된다.

<표 6-3> 환경개선부담금 오염유발계수

엔진총배기량(cc)	2,000이하	2,000초과~2,500이하	2,500초과~3,500이하	3,500초과~6,500이하	6,500초과~10,000이하	10,000초과
오염유발계수	1	1.25	1.75	2.64	4.5	5

자료 : 법제처(2022), 환경개선비용 부담법



<그림 6-2> 경유자동차 배기량별 배출가스 부피유량

자료 : 국립환경과학원(2013)

2.2 차령계수

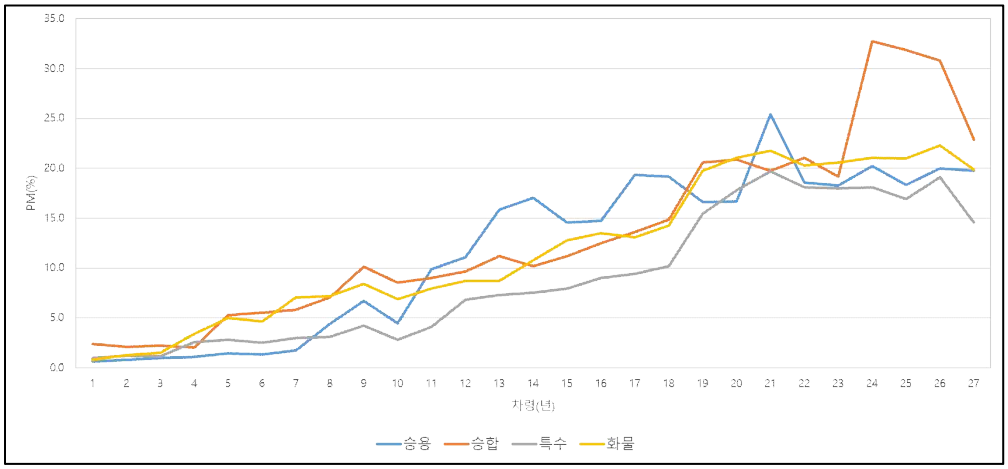
차령계수는 자동차의 최초 등록일을 기준으로 날 수에 비례하여 계산하되「대기환경보전법」제46조에 따른다. 차령계수는 차량 연식별로 유발계수를 부여하고 있으며, 차령연도가 높아질수록 오염유발량이 높아짐을 의미한다.

<그림 6-3>을 살펴보면 10년 이상 차량도 지속적으로 PM(%)이 증가하는 추이를 보이고 있다. 그러나 현재 차령계수는 10년 이상 차량은 동일한 차령계수 1.16을 적용하고 있어 합리적이지 못하다. 또한 차령이 10년 이상 차량은 유로4 이전 차량으로 다양한 차종이 포함되어 있어 일괄적으로 부과하는 방식은 현실에 맞지 않는다.

<표 6-4> 환경개선부담금 차령계수

차령	3미만	3년이상~4년미만	4년이상~6년미만	6년이상~8년미만	8년이상~10년미만	10년이상
차령계수	0.5	1	1.04	1.08	1.12	1.16

자료 : 법제처(2022), 환경개선비용 부담법



<그림 6-3> 경유자동차 차령별 평균 PM 배출 추이

2.3 지역계수

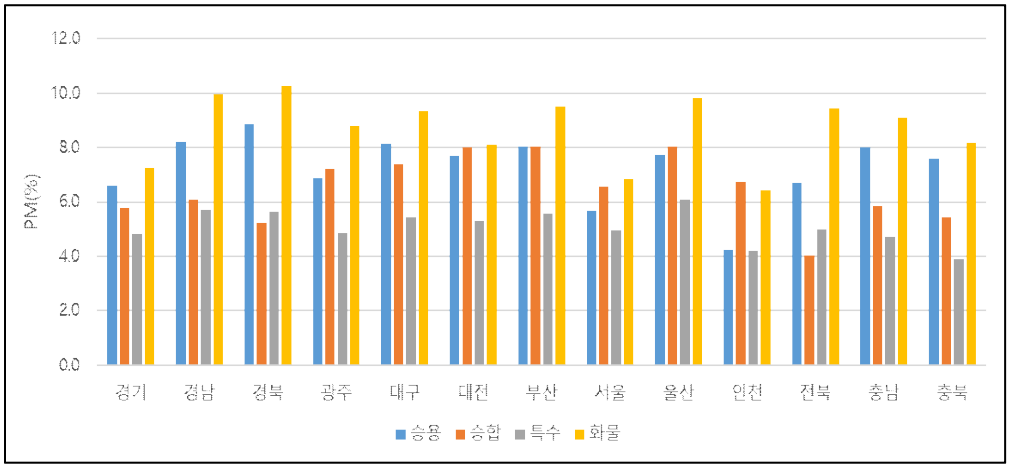
지역계수는 인구의 적용은 특별시, 광역시, 특별자치시, 시·군 단위로 하고, 「주민등록법」에 따라 해당지역에 등록된 인구를 기준으로 한다. 인구를 기준으로 500만명이상 1.53, 100만명이상 500만명이하 1.00, 50만명이상 100만명이하 0.87, 10만명이상 50만명이하 0.85, 10만명미만 0.40

이다. 지역계수의 의미는 지역별로 인구가 많으면 오염이 미치는 정도가 많아짐을 의미한다. 이는 인구가 많아질수록 해당지역의 피해인구가 높아지므로 인구대비 부과계수를 적용한 것으로 해석된다. 그러나 <그림 6-4>를 보면 서울, 인천이 대당 평균 PM(%) 가장 낮은 것으로 분석되었다. 서울, 인천의 경우 타지역보다 대당 PM 배출이 적음에도 불구하고 더 높은 환경개선부담금을 징수하고 있어 오염자 원인부담 원칙에 어긋난다.

<표 6-5> 환경개선부담금 지역계수

지역인구수 (인)	500만이상	100만이상~ 500만미만	50만이상~ 100만미만	10만 이상~ 50만미만	10만미만
지역계수	1.53	1.0	0.87	0.85	0.4

자료 : 법제처(2022), 환경개선비용 부담법



<그림 6-4> 2019년 지역별 경유자동차 평균 PM

3. 환경개선부담금 제도의 문제점

현재 환경개선부담금의 문제점은 오염부담원칙의 실효성, 부과대상자의 형평성, 산정방식의 현실성 부족 등으로 제도의 신뢰성이 저하되고

있어 수년 내에 세수 부족으로 제도 폐지 수순에 접어들 것이 자명하다. 현재 환경개선부담금의 부과기준과 산정방식의 문제점은 다음과 같다.

첫째, 오염부담원칙을 강화하는 것이다. 환경개선부담금은 경유차가 휘발유차나 LPG차 등 경쟁차에 비해 대기 오염물질을 상대적으로 많이 배출하기 때문에 오염물질 배출원인자인 경유차 운행자에게 오염물질의 피해비용을 환경개선부담금을 통하여 전가하고자 하는 목적에서 제도가 도입되었다. 그러나 현재 환경개선부담금 부과대상이 제도 취지에 부합하고 있는지 검토 및 연구해 볼 필요성이 있다. 그 이유는 자동차의 기술발전과 교통환경이 급격히 변화했음에도 불구하고, 2007년부터 시행된 환경개선부담금 부과산정기준이 오염부담원칙을 여전히 반영하고 있는지 검토할 필요성이 있다.

둘째, 산정방식이 현실성을 반영하지 못한다. 기존의 오염계수는 운행거리와 상관없이 배기량과 연식에 따라 일률적으로 부과되는 구조로서 운행 축소에 따른 오염저감 효과를 반영할 수 없다. 지역계수 또한 오염원인자 원칙에 입각해보면 인구가 많아질수록 계수값이 높아지는 것이 합당한 것인지 생각해 볼 문제이다. 인구가 많은 대도시일수록 세금을 부과하는 시민이 많아지므로 이중과세의 특성이 강하다. 단순히 인구수만을 기준으로 지역계수를 선정하는 것보다 지역면적, 인구밀도 등 다양한 변수를 반영한 계수 산정법이 개선이 요구된다. 또한 자동차 등록지를 기준으로 지역계수를 적용하는 것도 문제이다. 자동차 특성상 자동차 등록지에 머무르는 것이 아니고 타지역으로 이동하는 수단이므로 이동특성이 반영된 계수가 필요할 것으로 판단된다(이소영·조현구, 2017; 장재민 외, 2021).

셋째, 세수 부족으로 인한 재정 감소이다. 앞서 전술한 바와 같은 현상은 지속될 것이며, 환경개선부담금 제도 존폐가 걸려 있는 문제이다. 따라서 환경개선부담금 대상을 변경하거나 제도 보완이 시급한 실정이다.

제3절 사례분석 결과

1. 환경개선부담금 제도의 개선대안 설정

본 연구의 경유자동차 PM 예측모형은 PM 주요 배출요인을 제시하였다. 여기서 선정된 PM 배출요인은 환경개선부담금제 개선방안 도출에 활용하는 데 주안점을 둔다. 환경개선부담금제의 개선방안은 과태료와 부담금 부과방식 두 가지 측면으로 접근할 수 있다. 과태료는 행정법상 의무불이행에 대한 금전적 제재를 의미하며, 부담금은 관련법률에서 정하는 바에 따라 부과하는 조세 외의 금전지급의무 말한다. 이에 환경개선부담금제도의 해결방안은 오염부담원칙에 따른 부과대상자의 형평성 확보, 산정항목 및 방식 개선에 대해 가지 측면에서 제안하고자 한다.

첫째, 과태료 부과방식은 법률적 규제가 필요하다. 그러나 경유자동차의 배출허용기준 초과한 운전자는 일정수준 이상 오염물질을 배출한 원인자로서 이에 과태료를 부과하면 여러 가지 문제점 해결이 가능하다. 이 같은 정책이 지속적으로 이행된다면 경유자동차 운전자는 벌금 납부 부담으로 이어져 미세먼지 배출 감소에 적극적으로 동참할 것이며, 향후에는 미세먼지 저감효과로 이어질 것으로 예상된다.

둘째, 부담금 산정방식 개선방안은 오염자부담원칙 준수가 담보되어야 한다. 즉, 경유자동차의 PM 배출 기여도를 정확히 반영한 환경개선부담금 부과방식 수정이 필요하다. 기존의 환경개선부담금 산정항목은 PM 기여도에 따른 차등 부과하기 위해 본 연구에서 선정된 경유자동차 PM 배출 주요인을 적용한다. 제5장에서 제시한 요인들은 배출가스등급, 중량, 연식, 주행거리, 배기량이다. 이 요인들은 기존의 환경개선부담금 산정항목과 계수에 반영한다면 개선효과를 확인할 수 있을 것으로 예상된다.

그리고 현재 환경개선부담금 산정계수에 대한 적합성은 면밀히 검토해 봐야 한다. 기존의 오염유발계수는 산정항목으로는 합리적이나 계수의 차령계수 또한 차령이 10년 이상 된 자동차도 연식이 오래될수록 평균

PM이 증가하는 추세를 보이고 있다. 기존의 10년 이상 자동차의 차령계수가 일정한 것에 대해 일부 수정이 필요하다. 지역계수는 오염자 원인부담 원칙에 어긋나므로 기존 산정항목에서 제외하고 대체항목을 추가하기로 한다.

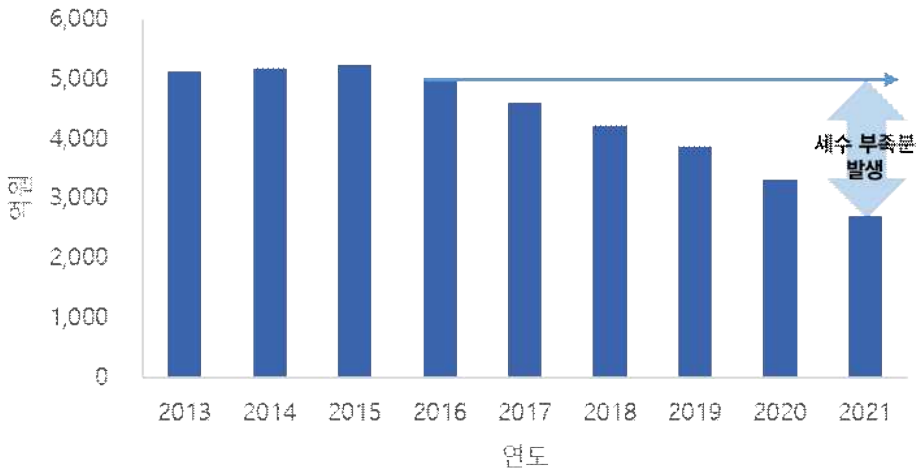
경유자동차 환경개선부담 도입 취지에 부합하기 위해서는 사전적인 오염배출저감 목적을 달성할 수 있어야 한다. 그래서 환경개선부담금은 기존 유사제도와 이중부담을지지 않게 설계되어야 하고, 부담대상 설정은 징수비용과 부담금 수입을 비교하여 효율적인 규모를 설정해야 한다. 그러기 위해서는 과태료 부과와 부담금 산정방식 개선 두 가지 대안을 제안한다.

대안1은 배출가스검사에서 불합격 판정을 받은 운전자에게 과태료를 부과하는 방안이다. 대안2는 환경개선부담금의 산정항목을 수정하고 산정방식을 개선하여 경유자동차 운행자의 형평성을 반영하기 위한 대안이다. 대안1과 대안2는 세부적으로 시나리오를 설정한 후 사례분석을 통해 관련정책의 개선 및 파급효과를 비교 분석한다.

2. 대안1 사례분석

2.1 시나리오 설정

<그림 6-5>와 같이 환경개선부담금은 2016년부터 감소 추세로 돌아서 급격하게 세수 부족분이 발생하고 있다. 이 같은 추세를 유지하면 환경개선부담금제도는 폐지될 것으로 예상된다. 환경개선부담금 제도를 유지하기 위해 세수 부족분 충당해야 한다. 2016년부터 분석 기준연도인 2019년 기준으로 산정한 세수 부족금액은 약 1,126억원이다. 이 세수 부족분을 해결하기 위해 다양한 예상 시나리오를 <표 6-6>과 같이 설정하였다. 시나리오 내용은 세수 부족분을 전부 환수하는 시나리오 또는 부분적으로 환수가 가능한 시나리오로 구성하였다.



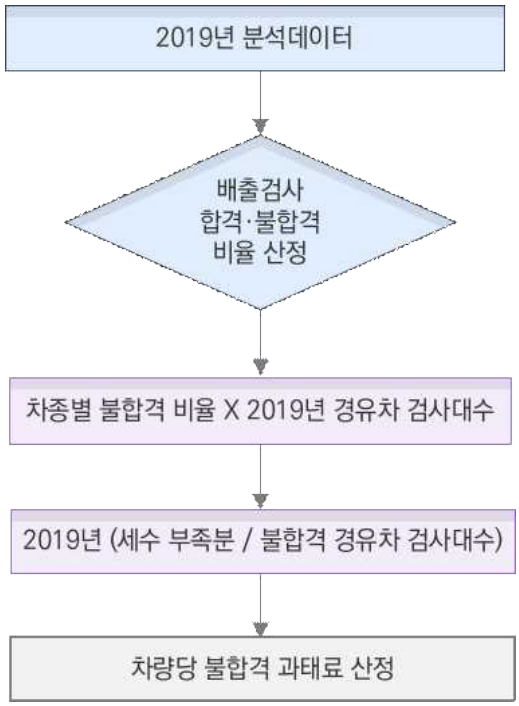
<그림 6-5> 환경개선부담금 세수 부족분

<표 6-6> 대안1 시나리오 내용

시나리오	내용
S1	· 불합격 과태료 부과로 세수 부족분 100% 환수
S2	· 불합격 과태료 부과로 세수 부족분 80% 환수
S3	· 불합격 과태료 부과로 세수 부족분 50% 환수

2.2 배출가스검사 불합격 운전자 과태료 부과방식

본 연구의 분석데이터에는 배출가스검사 합격/불합격 이력이 존재한다. 이 분석데이터를 활용하면 <그림 6-6>과 같이 배출가스검사 차량 불합률 산정하고, 이를 2019년 경유자동차 검사대수로 환산하면 연간 과태료 징수금액 추정이 가능하다. PM 배출허용기준은 <표 6-7>과 <표 6-8>을 준용하여 배출가스검사 합격과 불합격 차량을 분류하였다. 전 차종 불합격 비율은 14.7%이며, 전체 차종 중 화물차 불합격 비율은 16.9%로 가장 높은 것으로 나타났다. <표 6-9>를 기준으로 배출가스검사 불합격 차량에 대해 과태료 부과 시 적정 부과금을 산정하였다.



<그림 6-6> 배출가스검사 불합격 운전자 과태료 부과 산정 과정

<표 6-7> KD-147모드 배출가스검사 PM 배출기준

제작일자	PM 배출농도 기준
1992년 12월 31일 이전	45% 이하
1993년 1월 1일부터 1995년 12월 31일까지	40% 이하
1996년 1월 1일부터 2000년 12월 31일까지	35% 이하
2001년 1월 1일부터 2007년 12월 31일까지	25% 이하
2008년 1월 1일부터 2016년 8월 31일까지	15% 이하
2016년 9월 1일부터 2017년 12월 31일까지	8% 이하
2018년 1월 1일 이후	8% 이하

자료 : 한국교통안전공단(2022)

<표 6-8> Lug-Down3모드 배출가스검사 PM 배출기준

차종	제작일자	PM 배출농도 기준		
		1모드	2모드	3모드
차량총중량 5.5톤 이하 자동차	1995년 12월 31일 이전	70% 이하		
	1996년 1월 1일부터 2000년 12월 31일까지	60% 이하		
	2001년 1월 1일부터 2007년 12월 31일까지	50% 이하		
	2008년 1월 1일 이후	20% 이하		
차량총중량 5.5톤 초과 자동차	1995년 12월 31일 이전	50% 이하		
	1996년 1월 1일부터 2000년 12월 31일까지	45% 이하		
	2001년 1월 1일부터 2007년 12월 31일까지	30% 이하		
	2008년 1월 1일 이후	15% 이하		

자료 : 한국교통안전공단(2022)

<표 6-9> 2019년 경유차 배출가스검사 차종별 합격 및 불합격 비율

분석 데이터	합격		불합격		합계	
	대	비중(%)	대	비중(%)	대	비중(%)
승용	66,845	91.8	6,002	8.2	72,847	100
승합	50,778	86.7	7,776	13.3	58,554	100
화물	315,224	83.1	64,094	16.9	379,318	100
특수	51,443	90.7	5,306	9.3	56,749	100
계	484,290	**85.3**	83,178	**14.7**	567,468	100
2019년 배출검사차량	합격		불합격		합계	
	대		대		대	
승용	2,033,901		181,677		2,215,578	
승합	743,514		114,057		857,571	
화물	2,710,673		551,268		3,261,941	
특수	113,049		11,592		124,641	
계	5,601,137		858,594		6,459,731	

2.3 시나리오별 개선효과 분석결과

시나리오 1과 같이 불합격 차량을 대상에게 과태료를 100% 부과한다면 차량당 131,145원을 부과하면 환경개선부담금 세수 감소분을 충당이 가능한 것으로 분석되었다. 시나리오2의 경우 104,916원, 시나리오3은 65,572원이 적정 부과금액으로 산출되었다.

현재 등록지역이 서울인 운전자가 부담하는 환경개선부담금은 차종과 배기량에 따라 차이를 보이지만 연간 10~20만원 정도 납부하고 있다. 이 기준으로 비교해보면 과태료의 수준은 적정한 것으로 판단된다. 또한 불합격 운전자의 과태료 부과는 이용자의 형평성, 오염원인자 부담원칙, 재정 확보 모두가 가능한 적절한 미세먼지 절감정책이라 할 수 있다.

3. 대안2 사례분석

3.1 시나리오 설정

대안2의 기본 산정원칙은 산정항목 및 산정계수를 수정하거나 새롭게 도입하는 대안으로서 지역계수를 제외하고, 배출가스등급과 중량계수를 추가한다. 시나리오 설정은 기존 산정계수를 준용하거나 기존 산정계수와 새롭게 추가된 계수를 혼용하거나 본 연구에서 제안한 계수를 적용하는 시나리오로 설정하였다.

시나리오는 <표 6-10>과 같이 환경개선부담금 부과방식에 오염자 부담원칙, 부과대상자의 형평성 고려, 경유자동차 PM 주요 배출요인이 반영되도록 내용을 구성하였다. 시나리오1의 경우 지역계수는 제외하고 배출가스등급계수로 대체하였으며, 계수값은 차령계수를 준용하였다. 차령계수는 10년 이상 차량에게도 기존의 계수값 급간을 적용하여 20년 미만 차량까지 반영하였다. 시나리오2는 본 연구에서 선정된 PM 배출 주요인 항목과 산정계수에 요인별 중요도 가중치를 반영하였다. 시나리오3은 기

존의 오염부과계수과 같이 배기량이 높을수록 계수값이 큰 점을 착안하여 배출가스 4, 5등급의 계수값을 크게 설정하였다. 3개 시나리오는 차종별, 지역별, 오염자 배출 기여도별 환경개선부담금의 파급효과 분석한다.

<표 6-10> 대안2 시나리오 내용

시나리오	산정계수
S1	오염유발계수(기존) × 배출가스등급계수(차령계수 준용)× 차령계수(수정) × 연간 2회
S2	오염유발계수(가중치) × 배출가스등급계수(가중치) × 차령계수(가중치) × 중량계수(가중치) × 연간 2회
S3	오염유발계수(가중치) × 배출가스등급계수(오염유발계수 준용) × 차령계수(가중치) × 중량계수(가중치) × 연간 2회

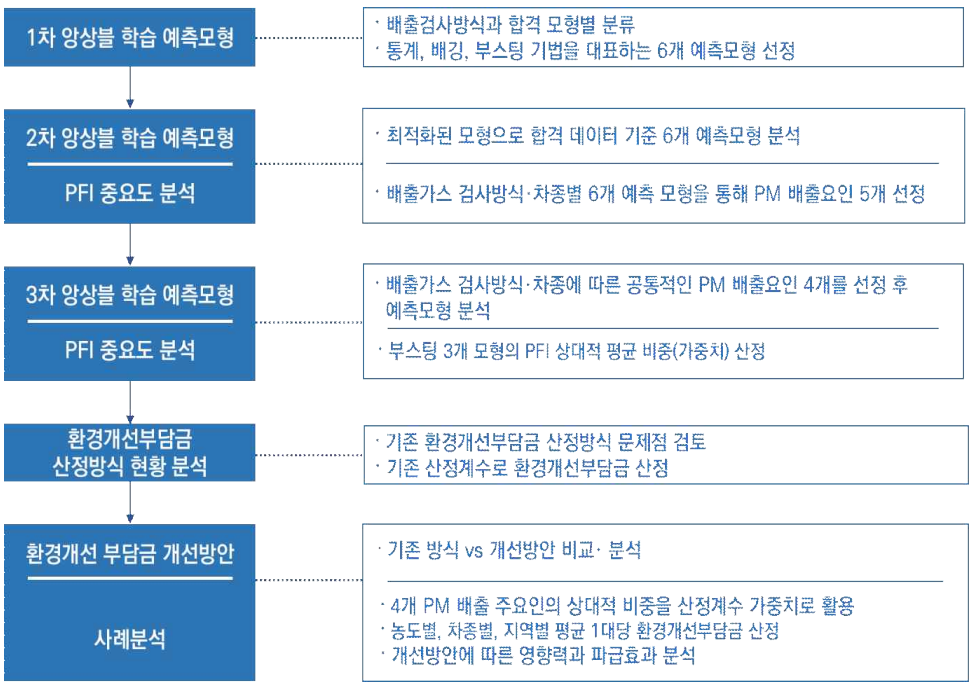
3.2 환경개선부담금 산정방식 개선과정

본 연구에서 경유자동차 PM과 입력변수간의 인과관계의 설명이 뛰어난 PM의 배출요인은 배출가스등급, 연식, 배기량, 중량으로 밝혀졌다. 이 요인들은 환경개선부담금 산정계수에 반영시켜 보다 합리적인 산정방식을 제안하고자 한다.

분석기준연도는 2019년이며, 기존 산정방식에 적용되고 있는 산정항목인 오염계수 즉, 배기량 항목은 유지한다. 기준 부과금액은 20,250원이며, 2019년도 부과금 산정지수 2.037를 적용한다. 환경개선부담금 산정방식 개선방안은 첫째, 기존의 차령계수를 현실 수준에 맞게 수정한다. 둘째, 배출가스등급계수, 중량계수를 추가로 반영하는 산정방식을 제시한다.

환경개선부담금 산정방식 개선 과정은 <그림 6-7>과 같이 아래와 같다. 1차 앙상블 학습에서는 배출검사방식과 합격모형으로 분류한 후 통계, 배깅, 부스팅 기법을 대표하는 6개 모형을 선정하였다. 2차 앙상블 학습에서는 하이퍼파라미터 튜닝으로 모형 최적화를 통해 합격데이터만

분류하여 검사방식과 차종별 6개 예측모형을 구축하였다. 2차 앙상블 학습 예측모형에 도출된 PM 배출요인은 공통적 특성을 분석하여 배출가스등급, 연식, 배기량, 총중량을 PM 배출의 주요인으로 선정하였다. 다음 과정에서는 4개의 배출요인을 3차 앙상블 학습에 다시 적용시킨 후 PM 배출요인의 PFI를 상대적 비중으로 산정하여 변수의 중요도를 정량화하였다. 상대적 비중은 예측력이 높은 부스팅모형인 CatBoost, LightGBM, XGBoost 3개 모형의 평균값을 적용하여 산정계수의 가중치로 사용한다. 기존의 환경개선부담금 산정방식에 대해 문제점을 검토하고, 개선방안을 제시한다. 3차 앙상블 학습에서 도출된 4개 PM 배출 주요인은 환경개선부담금 산정계수에 적용하기로 한다. 마지막으로 현재 환경개선부담금 산정금액과 3개 시나리오의 산정결과를 비교 분석한 후 그에 따른 영향력과 파급효과를 분석한다.



<그림 6-7> 환경개선 부담금 산정방식 개선과정

3차 앙상블 학습 예측모형에서 도출된 PM 배출 주요인의 상대적 중요도는 <표 6-11>과 같이 산출되었다. 연식, 배기량, 배출가스등급, 총중량의 중요도는 검사방식에 따라 차이를 보이고 있다. 현실적으로 차량검사방식으로 환경개선부담금의 대상을 분류하기가 어려우므로 요인별 중요도 가중치의 통일이 필요하다. 따라서 KD-147모드와 Lug-Down3모드의 배출요인 중요도의 평균값을 산정계수로 활용하고자 한다. 최종적으로 산정식에 반영될 PM 배출요인별 중요도 비중이 유사한 연식과 배출가스등급 그리고 배기량과 총중량을 각각 동일하게 적용하기로 한다. 연식과 배출가스등급 중요도 비중은 34%이며, 배기량과 총종량은 16.1%이다. 여기서 산정계수 가중치는 배기량과 총중량 평균비중을 배출가스등급과 연식 평균비중으로 나누어서 평균 산정계수 가중치는 2.12로 결정되었다.

<표 6-11> PM 배출 주요인 산정계수 가중치

구분	KD-147	Lug-Down3	평균	산정계수 비중	산정계수 가중치
연식	0.181	0.468	0.325	0.340	2.12
배출가스등급	0.501	0.208	0.355		
배기량	0.127	0.225	0.176	0.161	1
총중량	0.191	0.099	0.145		

환경개선부담금 산정계수는 <표 6-11>~<표6-15>와 같이 시나리오에 따라 적용하였다. 개선된 환경부담금 산정방식과 개선효과를 분석하기 위해 아래와 같은 순서대로 진행된다.

① 기존 오염계수(배기량계수) 가중치 사용
② 중량계수는 ①과 같이 준용하고 K-평균 군집화기법을 활용해서 기존 6개 등급으로 결정
③ 차령계수(연식계수) 기존 등급을 준용하고 PM 주요변수 상대적 가중치를 활용하여 2.12를 보정

④ 배출가스등급계수도 ③과 같이 준용하고 배출가스등급은 2~5 등급으로 4개 등급을 적용

⑤ 적용대상은 KD147, LugDown3 모두 포함하며, 면제 차량은 EURO5, EURO6, 저감장치 부착 차량 -> 산정대상에서 제외

⑥ 시나리오별 환경개선부담금 징수금액 산정

⑦ K-평균 군집화기법을 통해 경유자동차 PM 배출농도별 6등급 설정

⑧ PM 배출농도, 차종, 지역에 따라 시나리오별 환경개선부담금 개선효과 분석

<표 6-12> 시나리오별 오염유발계수

단위 : cc

시나리오	2,000이하	2,000초과~2,500이하	2,500초과~3,500이하	3,500초과~6,500이하	6,500초과~10,000이하	10,000초과
S1	1	1.25	1.75	2.64	4.5	5
S2	0.50	1.00	1.04	1.08	1.12	1.16
S3	0.50	1.00	1.04	1.08	1.12	1.16

<표 6-13> 시나리오별 차령계수

단위 : 년

시나리오	3미만	3이상~4미만	4이상~6미만	6이상~8미만
S1	0.5	1	1.04	1.08
S2	8 이상~10 미만	10 이상~15 미만	15 이상~20 미만	20 이상
S3	1.12	1.16	1.2	1.24

<표 6-14> 시나리오별 중량계수

단위 : kg

시나리오	2,115미만	2,115이상~2,685미만	2,685이상~4,595미만	4,595이상~7,885미만	7,885이상~19,170미만	19,170이상
S2 S3	0.50	1.00	1.04	1.08	1.12	1.16

<표 6-15> 시나리오별 배출가스등급계수

시나리오	2등급	3등급	4등급	5등급
S1	0.5	1.04	1.12	1.24
S2	1.06	2.20	2.37	2.63
S3	1.00	1.75	2.64	5.00

<표 6-16> 경유자동차 PM 배출농도별 6등급

등급	1등급	2등급	3등급	4등급	5등급	6등급
	저농도		중농도		고농도	
PM농도(%)	0~4	4~12	12~24	24~41	41~65	65~100

3.3 시나리오별 개선효과 분석결과

자동차 1대당 환경개선부담금 부과금액은 KD-147모드와 Lug - Down3 모드로 구분해서 산정하였으며, 개선효과는 두 가지 검사모드 차량을 합친금액을 비교한다. 자동차 1대당 환경개선부담금 부과금액은 기존 부과금액 보다 모든 시나리오에서 증가하였다. 현재 환경개선부담금 징수금액이 감소 추세인 관계로 산정계수를 높이거나 산정항목을 추가할 경우 세수 확보의 가능성을 검토하였다. 이는 산정계수 조정을 통해 부과금액의 감소로도 유도할 수 있어 정책의 유연성을 확인하기 위함이다.

무엇보다 사례분석에서 중요한 연구결과는 미시행시 대비 시나리오별 시행에 따라 PM 배출농도, 차종, 지역별로 파급효과를 검토하는 것이다. 이에 본 연구는 오염부담원칙에 따른 부과대상자의 형평성 확보, 산정항목 및 방식의 현실성 제고에 대한 개선효과를 등급별 자동차 1대당 환경개선부담금 부과금액 대비 전체등급 1대당 환경개선부담금 부과금액의 총합계 비중 변화로 비교한다. 이는 모집단 내의 분류 등급에 따라 상대적 부담금액의 변화를 알 수 있기 때문이다. 예컨대, 미시행과 시나리오1을 비교해보았을 PM 배출이 작은 동일그룹 내에 때 시나리오1 부과금액 비중이 감소하게 되면 환경개선부담금 산정방식의 개선에 따라 해당 차량운전자에게 상대적으로 부담금을 적게 부과됨을 의미한다.

미시행 대비 모든 시나리오에서 PM 고농도 배출차량이 더 많은 금액이 부과되었다. 시나리오 중 시나리오3이 환경개선부담금을 가장 많이 부담하는 것으로 나타났다. PM 배출농도, 지역, 차종에 따른 분류기준으로 미시행과 시나리오를 비교해보면 다음과 같은 특이점이 발견되었다.

경유자동차 PM 배출농도는 6개 등급으로 분류하였고 고·중·저 농도로 재분류하여 부과금액 비중을 비교하였다. 미시행시 고농도와 저농도 그룹의 부과금액이 167,782원, 78,137원이며, 시나리오3은 651,781원, 192,627원으로 산정되었다. 고농도와 저농도 그룹의 부과금액을 나누어 보면 미시행시 2.1배이지만 시나리오3은 3.4배로 분석되었다. 이는 시나리오3이 미시행시보다 고농도와 저농도 그룹의 부담금액 차이가 더 많다는 것을 알 수 있다. PM 배출 고농도 그룹은 미시행시 부과금액 비중이 36.1%이며 시나리오3의 경우 37.3%로 1.2%가 증가하였다. 그러나 저농도 그룹의 경우 미시행시는 24.5%, 시나리오3은 20.6%로 약 4% 감소하였다. 다시 말해 환경개선부담금 산정방식을 개선됨에 따라 PM 고농도 그룹의 부담금 비중은 증가한 반면 저농도 그룹의 부담금 비중은 감소함을 의미한다.

지역별로 개선효과를 비교해보면 부담금액 비중이 큰 폭으로 감소한 지역은 서울이며, 나머지 지역은 소폭 감소 또는 증가하는 것으로 나타났다. 이러한 이유는 기존의 지역계수값에서 서울 지역이 가장 높았으나 지역계수를 산정방식에 제외하고, 지역별로 차이가 없는 중량계수와 배출가스등급계수를 추가한 결과이다.

서울지역의 경우 계수값이 높은 기존의 지역계수를 산정방식에서 제외하고, 지역별로 차이가 없는 중량계수와 배출가스등급계수를 추가한 결과이다.

차종별 개선효과는 미시행와 모든 시나리오를 비교해보면 시나리오1의 부과금액 비중 변화는 미미하나 시나리오 2, 3은 일부 변화가 발견되었다. 부과금액 비중은 승용차와 특수차는 감소하나 화물차와 승합차는 증가하였다. 부과금액은 특수차가 높으며, 승합차, 승용차, 화물차 순으로 나타났다. 부과금액이 가장 높고 낮은 특수차와 화물차를 차이를 살펴보면 특수차가 화물차보다 미시행시 3.4배, 시나리오3은 2.5배 차이를 보였다. 이는 개선방안이 적용되면 기존 환경개선부담금의 차종별 차이가 작아지는 것을 알 수 있다.

<표 6-17> PM 배출농도별 자동차 1대당 환경개선부담금(KD-147모드)

단위 : 원, %

PM농도		미시행		S1	
		금액	비중	금액	비중
100~65	고농도	103,510	19.4	129,460	19.7
65~41		101,935	19.1	127,029	19.3
41~24	중농도	103,380	19.3	129,902	19.7
24~12		101,675	19.0	126,115	19.1
12~4	저농도	87,810	16.4	105,298	16.0
4~0		36,014	6.7	40,935	6.2
합계		534,325	100	658,739	100
PM농도		S2		S3	
		금액	비중	금액	비중
100~65	고농도	367,763	19.4	593,168	20.1
65~41		369,451	19.5	585,072	19.8
41~24	중농도	373,837	19.8	599,253	20.3
24~12		363,350	19.2	576,886	19.5
12~4	저농도	308,598	16.3	455,776	15.4
4~0		109,243	5.8	141,548	4.8
합계		1,892,243	100	2,951,704	100

<표 6-18> PM 배출농도별 자동차 1대당 환경개선부담금(Lug-Down3모드)

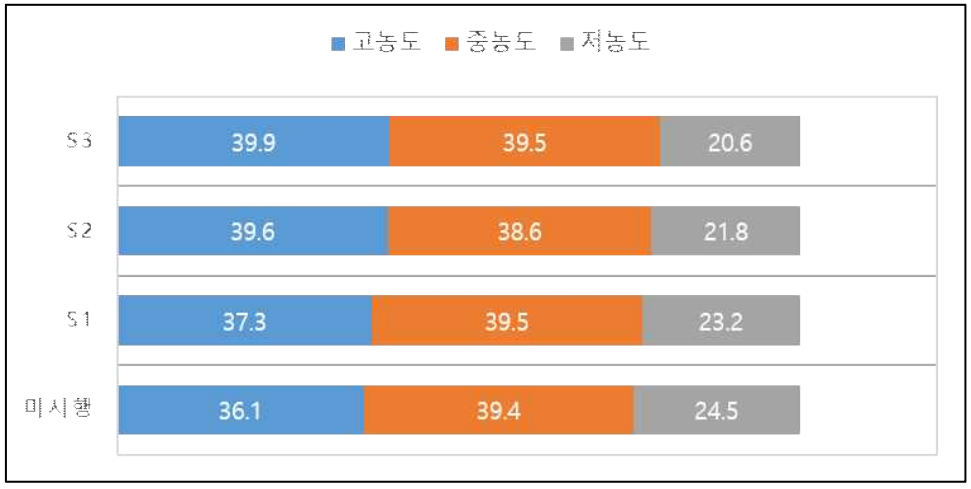
단위 : 원, %

PM농도		미시행		S1	
		금액	비중	금액	비중
100~65	고농도	230,364	18.7	344,170	19.9
65~41		236,148	19.2	343,132	19.8
41~24	중농도	236,922	19.3	334,203	19.3
24~12		239,039	19.4	331,606	19.2
12~4	저농도	187,890	15.3	251,903	14.6
4~0		98,917	8.0	126,012	7.3
합계		1,229,279	100	1,731,026	100
PM농도		S2		S3	
		금액	비중	금액	비중
100~65	고농도	454,648	20.3	708,853	20.4
65~41		455,000	20.3	701,482	20.2
41~24	중농도	432,834	19.3	681,606	19.6
24~12		419,536	18.7	677,124	19.5
12~4	저농도	319,521	14.3	491,700	14.1
4~0		158,039	7.1	217,824	6.3
합계		2,239,578	100	3,478,588	100

<표 6-19> PM 배출농도별 자동차 1대당 환경개선부담금(총 차량)

단위 : 원, %

PM농도		미시행		S1	
		금액	비중	금액	비중
100~65	고농도	167,782	17.2	238,245	17.8
65~41		184,529	18.9	260,018	19.5
41~24	중농도	190,860	19.6	263,735	19.8
24~12		193,265	19.8	263,129	19.7
12~4	저농도	160,772	16.5	212,179	15.9
4~0		78,137	8.0	97,907	7.3
합계		975,346	100	1,335,213	100
PM농도		S2		S3	
		금액	비중	금액	비중
100~65	상위	411,785	19.6	651,781	19.9
65~41		422,097	20.0	656,711	20.0
41~24	중위	412,485	19.6	653,200	19.9
24~12		400,813	19.0	643,721	19.6
12~4	하위	316,561	15.0	481,966	14.7
4~0		141,920	6.7	192,627	5.9
합계		2,105,660	100	3,280,006	100

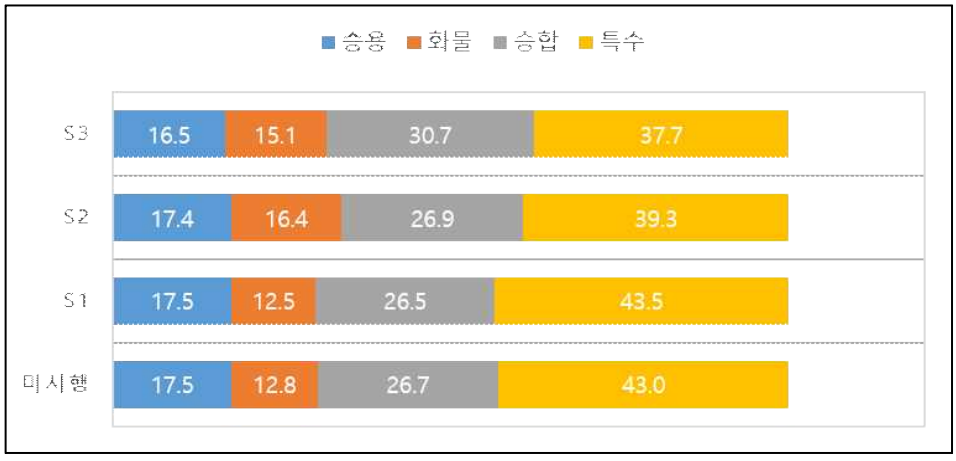


<그림 6-8> 농도별 1대당 환경개선부담금 비중

<표 6-20> 차종별 자동차 1대당 환경개선부담금 산정결과

단위 : 원, %

KD147	미시행		S1		S2		S3	
차종	금액	비중	금액	비중	금액	비중	금액	비중
승용	52,244	19.1	62,741	19.2	171,359	18.4	200,106	14.6
화물	60,659	22.1	72,293	22.2	211,482	22.8	336,340	24.5
승합	104,707	38.2	124,607	38.2	341,396	36.7	582,884	42.5
특수	56,415	20.6	66,292	20.3	204,770	22.0	251,027	18.3
합계	274,025	100	325,933	100	929,006	100	1,370,357	100
LugDown3	미시행		S1		S2		S3	
차종	금액	비중	금액	비중	금액	비중	금액	비중
화물	148,049	34.0	197,867	34.1	250,743	33.6	382,117	35.4
승합	86,337	19.8	113,856	19.6	185,007	24.8	197,618	18.3
특수	201,674	46.2	269,208	46.3	311,167	41.7	498,341	46.2
합계	436,059	100	580,931	100	746,916	100	1,078,076	100
총합계	미시행		S1		S2		S3	
차종	금액	비중	금액	비중	금액	비중	금액	비중
승용	200,293	17.5	260,608	17.5	422,101	17.4	582,223	16.5
화물	146,995	12.8	186,150	12.5	396,489	16.4	533,958	15.1
승합	306,381	26.7	393,814	26.5	652,563	26.9	1,081,225	30.7
특수	492,475	43.0	647,223	43.5	951,686	39.3	1,329,103	37.7
합계	1,146,144	100	1,487,795	100	2,422,839	100	3,526,509	100



<그림 6-9> 차종별 1대당 환경개선부담금 비중

<표 6-21> 지역별 자동차 1대당 환경개선부담금(KD-147모드)

단위 : 원, %

지역	미시행		S1		S2		S3	
	금액	비중	금액	비중	금액	비중	금액	비중
경기	43,981	20.0	45,363	3.1	190,603	6.0	281,622	6.0
경남	41,912	19.1	864,541	58.6	598,252	18.9	747,601	15.9
경북	4,792	2.2	64,696	4.4	277,587	8.8	453,164	9.6
광주	10,088	4.6	53,063	3.6	221,898	7.0	345,791	7.3
대구	16,688	7.6	55,183	3.7	234,894	7.4	378,512	8.0
대전	10,591	4.8	55,518	3.8	232,497	7.3	363,467	7.7
부산	19,155	8.7	52,672	3.6	223,049	7.0	350,274	7.4
서울	38,261	17.4	34,858	2.4	146,348	4.6	185,624	3.9
울산	9,303	4.2	54,714	3.7	233,340	7.4	371,528	7.9
인천	12,358	5.6	32,093	2.2	134,424	4.2	176,084	3.7
전북	4,329	2.0	55,140	3.7	228,279	7.2	354,546	7.5
충남	4,278	1.9	53,132	3.6	221,674	7.0	341,427	7.3
충북	4,114	1.9	54,110	3.7	225,147	7.1	348,751	7.4
합계	219,852	100	1,475,083	100	3,167,991	100	4,707,459	100

<표 6-22> 지역별 자동차 1대당 환경개선부담금(Lug-Down3모드)

단위 : 원, %

지역	미시행		S1		S2		S3	
	금액	비중	금액	비중	금액	비중	금액	비중
경기	106,178	4.4	152,068	4.5	205,329	4.9	280,282	4.3
경남	165,490	6.9	254,648	7.6	326,057	7.8	505,881	7.8
경북	180,924	7.6	278,709	8.3	330,450	7.9	548,824	8.5
광주	180,776	7.6	232,517	7.0	298,646	7.2	452,945	7.0
대구	161,628	6.8	206,898	6.2	272,561	6.5	399,457	6.2
대전	174,834	7.3	223,715	6.7	278,227	6.7	429,630	6.6
부산	216,617	9.1	279,085	8.3	330,409	7.9	544,569	8.4
서울	153,335	6.4	124,014	3.7	168,407	4.0	214,842	3.3
세종	142,474	6.0	217,754	6.5	273,763	6.6	430,220	6.6
울산	190,166	8.0	257,260	7.7	306,573	7.3	505,680	7.8
인천	137,132	5.7	171,879	5.1	205,861	4.9	308,604	4.8
전남	166,808	7.0	286,046	8.6	339,443	8.1	564,855	8.7
전북	152,063	6.4	228,401	6.8	288,333	6.9	452,012	7.0
충남	132,570	5.6	226,473	6.8	287,540	6.9	442,178	6.8
충북	126,781	5.3	204,491	6.1	260,764	6.2	393,563	6.1
합계	2,387,776	100	3,343,956	100	4,172,364	100	6,473,541	100

<표 6-23> 지역별 자동차 1대당 환경개선부담금(총 차량)

단위 : 원, %

지역	미시행		S1		S2		S3	
	금액	비중	금액	비중	금액	비중	금액	비중
경기	87,202	4.3	87,202	4.3	200,836	5.0	280,691	4.5
경남	208,938	10.3	208,938	10.3	389,601	9.7	562,311	9.0
경북	157,265	7.8	157,265	7.8	320,181	7.9	530,242	8.5
광주	138,492	6.8	138,492	6.8	272,516	6.8	416,463	6.7
대구	121,135	6.0	121,135	6.0	257,808	6.4	391,253	6.3
대전	128,356	6.3	128,356	6.3	259,852	6.4	403,044	6.5
부산	169,692	8.4	169,692	8.4	299,036	7.4	487,792	7.8
서울	113,620	5.6	113,620	5.6	159,021	3.9	202,410	3.3
세종	142,474	7.0	142,474	7.0	273,763	6.8	430,220	6.9
울산	145,309	7.2	145,309	7.2	281,701	7.0	460,118	7.4
인천	99,900	4.9	99,900	4.9	179,610	4.5	259,907	4.2
전남	166,808	8.2	166,808	8.2	339,443	8.4	564,855	9.1
전북	123,529	6.1	123,529	6.1	271,340	6.7	424,433	6.8
충남	116,403	5.7	116,403	5.7	274,727	6.8	422,578	6.8
충북	107,183	5.3	107,183	5.3	251,636	6.2	382,079	6.1
합계	2,026,306	100	2,026,306	100	4,031,072	100	6,218,396	100

제Ⅷ장 결론 및 향후 연구

제1절 결론

경유자동차는 디젤엔진 특성으로 다른 차량에 비해 미세먼지(PM)를 압도적으로 많이 배출한다. 2019년 6월 기준으로 한국의 경유자동차는 총 997만여 대로 전체 차량에서 차지하는 비중이 42.5%에 이른다. 반면 미국과 중국, 일본은 디젤차 비중이 1~3% 수준에 그쳐 한국은 이들 국가에 비해 경유자동차 비중이 높은 편이다. 이 같은 여건에서 여전히 경유자동차는 미세먼지를 배출하면서 우리 인체를 위협하고 있다. 특히 우리나라는 경유자동차 비중이 타국가에 비해 높기 때문에 경유자동차 평가관리 방안 또는 배출허용기준 강화, 배출가스등급제, 미세먼지 저감장치 부착, 환경개선부담금제 등 관련정책의 개선방안을 지속적으로 제기해야 한다. 이 같은 정책을 수립 및 시행하기 위해서는 경유자동차 PM의 영향요인 파악이 매우 중요하다. 그러므로 경유자동차 PM 배출 예측의 정확도를 제고시키고, 영향 원인 중요도를 명확하게 규명하는 연구가 필요한 실정이다.

본 연구에서는 분석데이터의 특성과 연구방법론을 다음과 같이 요약한다. 자동차 배출가스 정밀검사(Inspection/Maintenance: I/M) 자료는 PM의 최초 측정 단위인 농도(%)이므로 질량으로 변환하는 회귀분석이 필요 없고, 원자료의 특성을 분석할 수 있는 강점이 있다. 예측모형 방법론은 머신러닝 기법인 앙상블 학습을 통해 경유자동차의 PM 배출요인을 식별하고, 다양한 예측모형 구축을 이용하여 PM 배출요인별 영향력을 분석하였다. 또한, 레버리지 분석 통해 분석데이터를 전처리하고, 예측모형의 신뢰성을 향상시켰다. 본 연구는 실험실 방식으로 측정한 경유자동차 PM과 차량제원 데이터를 활용하고, 앙상블 학습을 이용한 PM 배출 예측모형을 구축하였다.

경유자동차 PM 배출 예측모형 구축과정은 다양한 앙상블 학습을 적용하였다. 1차 앙상블 학습 예측모형은 KD-147모드와 Lug-Down3모드로 구분하고, 배출가스검사 합격과 불합격 데이터를 분류하여 회귀모형, Bagging과 Boosting 알고리즘을 기반한 20개 모형을 분석하였다. 여기서 예측성능이 현저히 떨어지는 모형은 제외시켰다. 이 모형 중 정확도가 높으면서 통계기법, Bagging, Boosting을 대표할 수 있는 모형 6개를 선정하였다. 통계기법을 대표하는 회귀분석과 의사결정나무를 선택하였으며, Bagging을 대표하는 랜덤포레스트 모형을 선정하였다. 나머지 3개 모형은 Boosting을 대표하는 모형이며, 예측성능 상위 1~3위인 CatBoost, LightGBM, XGBoost모형을 선정하였다.

2차 앙상블 학습에서는 합격데이터만을 활용하고, 차종별 PM 예측모형을 구축하였다. 2차 앙상블 학습 PM 예측모형은 하이퍼파라미터 튜닝으로 모형 최적화를 통해 예측성능을 향상시켰다. 모형의 예측성능 평가 결과는 학습 세트와 테스트 세트로 구분하여 성능지표를 제시하였다. KD-147모드의 경우 CatBoost모형 R^2가 0.815, $RMSE$는 1.619로 모형의 정확도가 가장 높은 것으로 나타났다. 반면 선형회귀분석에는 R^2가 0.649, $RMSE$는 2.231로 모형의 정확도가 가장 낮은 것으로 분석되었다. 예측성능을 비교해보면 모형의 신뢰성이 높은 순서대로 나열하면 부스팅 모형, 배깅모형, 통계모형 순으로 나타났다. 부스팅 모형 중 CatBoost, LightGBM, XGBoost의 예측력 차이는 미미하며, 모형 분류기준과 시나리오에 따라 모형 예측력 순위가 달라지기는 하지만 예측성능은 유사하다. 차종별로 분류한 예측모형은 KD-147모드의 경우 승용차의 예측성능이 가장 뛰어났고 Lug-Down3모드는 화물차의 예측력이 가장 우수하였다. 반면 승합차의 예측모형은 두 검사모드 모두 낮은 것으로 나타났다.

입력변수 중요도는 통계모형, 배깅모형, 부스팅모형에 따라 다소 차이를 보였다. 그러나 입력변수 중요도의 상대적 비중을 모형별로 비교해보면 일부 모형에서 차이를 보인다. 부스팅모형과 통계모형을 비교해보면 연식과 배출가스등급 변수의 PFI 상대적 비중의 차이가 많은 것으로 분

석되었다. 선형회귀모형의 경우 입력변수 중에서 연식과 배출가스등급 PFI 비중이 매우 큰 것으로 분석되었다. 또한 배출가스검사방식과 차종에 따라 PM 배출 주요인의 PFI가 일부 다르게 산정되었다. 먼저 검사방식에 상관없이 공통적인 PM 배출 주요인은 배출가스등급, 연식, 배기량, 총중량으로 분석되었다. PM 배출 주요인의 차이점은 검사방식에 따라 다르게 나타났다. KD-147모드에서 배출가스등급이 PFI가 가장 높았으나 Lug-Down3모드의 경우 연식의 중요도가 1위를 차지하였다. 차종별 차이점은 특수차는 적재중량, 승합차는 승차인원이 선정되었다. 이는 특수차는 적재를 목적으로 운행하는 차량이 대부분이고, 승합차는 적재보다 다인 승객 운행이 주된 목적에서 기인한 것으로 사료된다. 본 연구와 선행연구의 PM 배출 주요인은 대부분 일치하였다. 그러나 본 연구의 경우 배출가스등급이 주요인에 포함된 것과 선행연구에 많은 요인으로 제시되었던 배출가스저감장치 포함되지 않는 점이 특징이라 할 수 있다. 이는 분석데이터 특성에서 확인되었듯이 배출가스저감장치 장착 차량 중 고농도 배출 차량의 비중이 일정부분 차지하고 있는 요인이 반영되었을 것으로 판단된다.

사례분석의 주요 목적은 본 연구의 앙상블 학습 예측모형에서 도출된 경유자동차 PM 배출의 주요인을 미세먼지 절감 및 환경 관련 정책에 적용하기 위함이다. 현재 환경개선부담금 산정원칙과 방식은 다방면으로 문제점을 안고 있다. 이러한 문제점을 개선하기 위해 본 연구에서는 PM 배출요인과 요인별 중요도 분석결과를 사례분석에 활용하였다. 환경개선부담금제의 개선방안은 두 가지 측면으로 접근한다. 첫 번째 대안은 경유자동차의 배출허용기준 초과한 운전자는 일정수준 이상 오염물질을 배출한 원인자이기 때문에 과태료를 부과하는 방안이다. 두 번째, 대안은 경유자동차의 PM 배출 기여도가 반영한 환경개선부담금 부과방식으로 개선하는 것이다. 기존의 환경개선부담금 산정항목은 PM 배출 기여도에 따라 차등 부과하기 위해 본 연구에서 선정된 경유자동차 PM 배출 주요인을 적용한다.

대안별 시나리오에 따라 사례분석 결과는 다음과 같다. 대안1의 경우 배출가스검사에서 불합격 판정을 받은 운전자에게 과태료를 부과한다면 자동차 1대당 131,145원을 부과하면 환경개선부담금 세수 감소분을 충당이 가능한 것으로 분석되었다. 대안2의 경우 지역계수 대신 중량계수와 배출가스등급계수를 산정식에 적용하는 것이 환경개선부담금 취지에 더 적합한 것으로 판명되었다. 또한 평균 차량연식이 약 10년임을 감안하면 최소 20년까지 차령계수는 차등 적용하는 것이 합리적이다. 본 연구의 도출된 배출가스등급과 차령계수의 가중치를 적용한 개선방안과 기존의 산정방식으로 책정한 자동차 1대당 환경개선부담금을 비교한 결과, 고농도와 저농도 그룹의 부담금액 차이는 기존방식은 2.1배이지만 개선방안은 3.4배로 분석되었다. 이는 더 많은 PM 고농도 배출 운전자에게 부담금이 더 전가되는 구조를 확인할 수 있었다. 또한 오염계수 중 배기량이 높은 계수값을 배출가스 4, 5등급 차량에게 해당계수값에 적용하면 최대 708,853원이 부과되었다. 이는 오염자부담원칙 강화 기조에 적합한 산정계수임을 방증하였다. 차종별로 부담금이 가장 높고 낮은 특수차와 화물차의 차이를 살펴보면 특수차가 화물차보다 미시행시 3.4배, 시나리오3은 2.5배 차이를 보였다. 이는 개선방안이 적용되면 기존 환경개선부담금의 차종별 차이는 작아지는 것을 의미한다. 지역별로 검토해보면 서울 지역의 부담금만 감소 폭이 큰 편이고, 나머지 지역은 소폭 증감하는 것으로 분석되었다.

이와 같이 본 연구의 전 과정에서 앙상블 학습 경유자동차 PM 배출 예측모형의 예측성능의 우수함과 PM 배출요인을 명확히 규명하였다. 앙상블 학습 경유자동차 PM 배출 예측모형 활용방안으로는 환경개선부담금 산정방식 개선뿐만 아니라 다양한 미세먼지 저감정책 활용에도 적합할 것으로 판단된다. 이러한 연구는 향후 친환경 정책 및 전략 수립에 밑거름이 될 것으로 기대한다.

제2절 향후 연구

본 연구는 전통적인 통계적 기법을 활용한 기존연구의 한계점을 극복하기 위하여 머신러닝 기법인 앙상블 학습을 통해 경유자동차 PM 배출농도(%)기반 예측모형을 구축하고, PM 배출 주요인을 제시하였다. 그러나 연구과정에서 나타난 연구의 한계와 향후 과제는 다음과 같다.

첫째, 머신러닝기법은 종속변수와 입력변수들 간의 비선형관계를 한층 더 유연하게 모델링할 수 있는 강점이 있음에도 불구하고, 종속변수와 입력변수간의 인과관계를 직관적으로 해석하기 어려운 단점이 있다. 이런 한계를 극복하기 위해 본 연구는 앙상블 학습을 통해 다양한 모형의 PM 배출요인을 제시하였다. 앙상블 학습은 종속변수와 입력변수간의 정량적인 영향력을 분석에는 출중하기 때문이다. 그러나 앙상블 학습 또한 정확한 인과관계의 설명력을 명확히 서술하기에는 다소 부족한 점이 있다.

둘째, 횡단면 자료의 한계이다. 분석데이터의 시간적 범위는 2019년 9월 자동차 정밀검사자료이다. 횡단면자료는 시계열자료보다 데이터 학습에 한계가 존재한다. 순차적인 데이터 학습이 가능하다면 과거의 학습과 연결이 가능하므로 예측성능을 더 높일 수 있을 것으로 판단된다.

셋째, 본 연구의 경유자동차 PM 배출 예측모형에는 복합변수를 고려하지 못한 한계가 있다. 자동차 배출에 영향을 미치는 요인은 차량출력, 엔진실린더수, 타이어 휠 크기, 수동/자동, 연료인젝션 등 다양한 요인들을 반영하지 못하였다. 자동차 정밀검사 항목을 입력변수로 사용하였으나 모형의 우수성을 확보하려면 검사항목에 대한 복합변수를 활용한 연구가 필요할 것으로 사료된다.

마지막으로 전기차 및 수소차와 같은 친환경자동차의 대중화 길목에서 비연소성 PM 배출요인 분석이 병행되어야 한다. 특히 전기차의 경우 경유자동차에 비해 차량 중량이 높으므로 비연소성 PM을 더 많이 배출하고 있다. 본 연구에서 PM 배출의 주요인인 중 하나인 차량중량을 새로

운 기준으로 배기량이 없는 친환경차의 비연소성 PM 배출까지 고려한 관련 정책에 대한 논의와 연구가 필요한 시점이다. 향후 비연소성 PM 관련 데이터가 충분히 확보된다면 연소성과 비연소성 PM 배출요인을 복합적으로 분석할 수 있는 연구가 필요할 것으로 판단된다.

참 고 문 헌

[국내연구]

강광규(2009), “경유차 환경개선부담금제도 개선방안”, 환경포럼, 13(4), pp. 1-8.

국립환경과학원(2012),「대형차 배출가스 결함확인검사 도입 연구(II) - 이동식 배출가스 측정장비(PEMS) 활용」, NIER-RP2012-211.

국립환경과학원(2013),「운행경유차 매연 농도 검사결과의 배출질량 환산 기법 연구」, NIER-RP2013-284.

국립환경과학원(2015),「운행차 검사결과 데이터베이스를 활용한 운행차 배출기준 개선 방안 마련」, NIER-RP2015-354.

국립환경과학원(2015),「자동차 배출오염물질 관리와 환경」, NIER-GP 2015-169.

국립환경과학원(2018),「자동차 실도로 주행시 배출가스 측정제도 도입에 따른 오염물질 개선효과 평가(Ⅱ)」, NIER-RP2018-120.

국립환경과학원(2019),「실제 도로주행 기반 자동차 대기오염물질 배출」.

기획재정부(2014～2022), 「부담금운용종합보고서」.

김성진, 안현철(2016), “기업신용등급 예측을 위한 랜덤 포레스트의 응용. 산업혁신연구”, 32(1), pp. 187-211.

김종성, 이준형, 김동현, 최창현, 이명진, 김형수(2019), “머신러닝 기반의 호우피해 발생확률 예측 모형 개발”, 한국방재학회 논문집, 19(6), pp. 115-127.

박용우, 백세흠, 박신형, 권오훈(2016), “의사결정나무를 이용한 고속도로 공사구간 사고 심각도에 관한 연구”, 대한교통학회지, 34(6), pp. 535-547.

박준홍, 이종태, 김정수, 김선문, 안근환(2013), “WLTP 주행모드에서의 경유차 입자상물질 개수 배출 특성”, 한국분무공학회지, 18(3), pp. 155-160.

법제처(2022)「환경개선비용 부담법」.

서울연구원(2015),「서울시 운행경유차의 매연배출평가와 맞춤관리방안」, 2015-PR-07, pp. 1-130.

안경민(2021), “통계적 매칭과 머신러닝 앙상블 기법을 활용한 기업혁신 및 경영성과 예측 모형 개발”, 동국대 박사논문.

이소영, 조현구(2017), “제도분석의 관점에서 본 경유차 환경개선부담금 변화과정 연구: 정책모형간 비교를 중심으로”. 한국정부학회, 29(3), pp. 537-561.

이수지, 김호(2019), "미세먼지가 건강에 미치는 영향", 국토 제452호, 특집_미세먼지 해결을 위한 국토분야의 대응방안 6, pp. 42-48

이영섭, 오현정, 김미경(2005), “데이터 마이닝에서 배깅, 부스팅, SVM 분류 알고리즘 비교 분석”, 응용통계연구, 18(2), pp. 343-354.

이재득(2021), “투자와 수출 및 환율의 고용에 대한 의사결정 나무, 랜덤 포레스트와 그래디언트 부스팅 머신러닝 모형 예측”, 무역학회지, 46(2), pp. 281-299.

장수은(2021), “미세먼지 주범은 경유차? 문제는 비연소성 미세먼지: 미세먼지 정책 전환, 빠를수록 좋다”, 한국경제지리학회지, 24(2), pp. 210-212.

장재민, 이유봉, 한정헌(2021), “탄소중립시대를 대비한 환경개선부담금 개선방안 연구”, 대한교통학회지, 40(5), pp. 631-642.

최용욱, 윤대웅, 최준환, 변중무(2020), “베이지안 최적화를 이용한 암상 분류 모델의 하이퍼 파라미터 탐색”. 지구물리와 물리탐사, 23(3), pp. 157-167.

통계청(2019), 「한국의 사회동향 2019」.

한국교통안전공단(2008),「운행경유차 정밀검사방법 개선에 관한 연구」.

한국교통안전공단(2009),「대형경유차 및 건설기계 배출가스 검사방법 개선에 관한 연구」.

한국교통안전공단(2022),「자동차종합검사 교육교재(업무 매뉴얼)」.

환경부(2015),「환경개선부담금 업무편람」.

환경부(2016),「정부합동 미세먼지 관리 특별대책」.

환경부(2022),「국가미세먼지정보센터, 미세먼지자료」.

한정헌(2022), “배출가스 정밀검사 대용량 자료를 활용한 경유자동차의 1차 생성 미세먼지 실배출 요인 연구”, 서울대학교 환경대학원 박사논문.

한진석(2020), “경유 화물자동차 매연 배출 요인 분석”, 환경정책, 28(3), pp. 1-18.

[국외연구]

Aman, M., Solangi, K., Hossain, M., Badarudin, A., Jasmon, G., Mokhlis, H., Bakar, A., Kazi, S.N.(2015), “A review of safety, health and environmental (SHE) issues of solar energy system”, Renewable and Sustainable Energy Reviews, 41, pp. 1190-1204.

Bai, Z., Liu, Q., & Liu, Y.(2022). “Groundwater potential mapping in Hubei Region of China using machine learning, ensemble learning, deep learning and AutoML methods.” Natural Resources Research (New York, N.Y.), 31(5), pp. 2549-2569.

Bemani, A., Baghban, A., Mohammadi, A. H., Andersen, P. Ø.(2020), “Estimation of adsorption capacity of CO2, CH4, and their binary mixtures in Quidam shale using LSSVM: application in CO2 enhanced shale gas recovery and CO2 storage”, Journal of Natural Gas Science and Engineering, 76, pp. 103204.

Bemani, A., Kazemi, A., Ahmadi, M., Yousefzadeh, R., & Moraveji, M. K.(2022), “Rigorous modeling of frictional pressure loss in inclined annuli using artificial intelligence methods”, Journal of Petroleum Science and Engineering, 211, pp. 110203.

Bergmann, M., Kirchner, U., Vogt, R. and Bente, T.(2009), “On-road and laboratory investigation of low-level PM emissions of a modern diesel particulate filter equipped diesel passenger car”, Atmospheric Environment, Vol. 43(11), pp. 1908-1916.

Berk, K.(1978), “Comparing subset regression procedures”, technometrics, 20(1), pp. 1-6.

Beydoun, M. and Guldmann, J.(2006), “Vehicle characteristics and emissions: logit and regression analyses of I/M data from Massachusetts, Maryland, and Illinois”. Transportation Research Part D, Vol. 11 pp. 59 - 76.

Bikas, G., Zervas, E.(2007), “Non regulated pollutants emitted from EURO3 diesel vehicles as a function of their mileage”. Energy Fuel, Vol. 21(5), pp. 2731 - 2736.

Breiman, L.(1996), “Bagging predictors”, Machine Learning, 24(2), pp. 123-140.

Breiman, L.(2001), “Random forests”, Machine Learning, 45(1), pp. 5-32.

Bukowiecki, N., Lienemann, P., Hill, M., Furger, M., Richard, A., Amato, F., Prévôt, A. S. H., Baltensperger, U., Buchmann, B. and Gehrig, R.(2010), “PM10 emission factors for non-exhaust particles generated by road traffic in an urban street canyon and along a freeway in Switzerland”, Atmospheric Environment, Vol. 44, pp. 2330-2340.

Chen W, Yun Y, Wen M, Lu H, Zhang Z, Liang Y.(2016), “Representative subset selection and outlier detection via isolation forest”, Analytical Methods, 8(39): pp. 7225-7231.

Chen, T and Guestin, C.(2016), “XGBoost: a scalable tree boosting system”, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785 - 794

Dallmann, T. R., Harley, R. A. and Kirchstetter, T. W.(2011), “Effects of diesel particle filter retrofits and accelerated gleet turnover on drayage truck emissions at the port of Oakland”. Environ. Sci. Technol. 45, pp. 10773 - 1077.

DL4J(2016), Introduction to Deep Neural Networks, DEEPLEARNING4J, A Web Page, (Available in : https://deeplearning4j.org/neuralnet-overview).

Elangasinghe, M. A., Singhal, N., Dirks, K. N., Salmond, J. A. and Samarasinghe, S.(2014), “Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering”, Atmospheric Environment, Vol. 94, pp. 106-116.

Fisher, A., Rudin, C., Dominici, F.(2019), “All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously”. J. Mach. Learn. Res., 20(177), pp. 1-81.

Franco, V., Kousoulidou, M., Muntean, M., Ntziachristos, L., Hausberger, S., and Dilara, P.(2013), “Road vehicle emission factors development: a review”. Atmospheric Environment, Vol. 70, pp. 84-97.

Freund, Y., Schapire, R.(1999), “Large margin classification using the perceptron algorithm”, Machine Learning, 37(3), pp. 277-296.

Friedman, J. H.(2001), “Greedy function approximation: a gradient boosting machine”, The Annals of Statistics, 29(5), pp. 1189-1232.

Friedman, J. H.(2002), “Stochastic gradient boosting, Computational Statistics and Data Analysis”, 38(4), pp. 367-378

Geller, M. D., Ntziachristos, L., Mamakos, A., Samaras, Z., Schmitz, D. A., Froines, J. R. and Sioutas, C.(2006), “Physicochemical and redox characteristics of particulate matter (PM) emitted from gasoline and diesel passenger cars”, Atmospheric Environment, Vol. 40, pp. 6988-7004.

Guan, B., Zhan, R., Lin, H. and Huang, Z.(2014), “Review of state of the art technologies of selective catalytic reduction of NOX from diesel engine exhaust,” Applied Thermal Engineering, Vol. 66(1-2), pp. 395-414.

Gulia, S., Nagendra, S. M. S. and Khare, M.(2017), “A system based approach to develop hybrid model predicting extreme urban NOx and PM2.5 concentrations”. Transportation Research Part D, Vol. 56, pp. 141-154.

Hinton, G. E., Simon O., and The. Y.W.(2006), "A fast learning algorithm for deep belief nets", Neural computation 18(7), pp. 1527-1554

Hochreiter, H and Schmidhuber, J.(1997), “Long short-term memory”, Neural Computation, 9(8), pp. 1735-1780.

Hocking, R.(1976), “The analysis and selection of variables in linear regression”, Biometrics, 32(1), pp. 1-49.

Ježek, I., Drinovec, L., Ferrero, L., Carriero, M. and Močnik, G.(2015), “Determination of car on-road black carbon and particle number emission factors and comparison between mobile and stationary measurements”, Atmospheric Measurement Techniques, Vol. 8, pp. 43 - 55.

Jung, S., Mun, S., Chung, T., Kim, S., Seo, S., Kim, I., ... & Hong, Y. (2019), “Emission characteristics of regulated and unregulated air pollutants from heavy duty diesel trucks and buses”, Aerosol and Air Quality Research, 19(2), pp. 431-442.

Karjalainen, P., Pirjola, L., Heikkilä, J., Lähde, T., Tzamkiozisc, T., Ntziachristosc, L., Keskinena, J. and Rönkkö, T.(2014), “Exhaust particles of modern gasoline vehicles: a laboratory and an on-road study”, Atmospheric Environment, Vol. 97, pp. 262-270.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. Y.(2017), “Lightgbm: A highly efficient gradient boosting decision tree”, In Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3149-3157

Kearns, M., Valiant, L.(1989), “Crytographic limitations on learning Boolean formulae and finite automata”, Annual ACM Symposium on Theory of Computing: Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing; 14-17 May 1989, pp. 433-444.

Khamehchi, E., Bemani, A.(2021), Prediction of pressure in different two-phase flow conditions: machine learning applications. Measurement, 173, pp. 108665.

Kim, W-G., Kim, C-K., Lee, J-T., Kim, J-S., Yun, C-W. and Yook, S-J.(2017), “Fine particle emission characteristics of a light-duty diesel vehicle according to vehicle acceleration and road grade”, Transportation Research Part D, Vol. 53, pp. 428-439.

Krecl, P., A. C. Targino, T. P. Landi, and M. Ketzel.(2018), “Determination of black carbon, PM2.5, particle number and NOx emission factors from roadside measurements and their implications for emission inventory development”, Atmospheric Environment, Vol. 186, pp. 229-240.

Krecl, P., Johansson, C., Targino, A. C., Ström, J. and Burman, L.(2017), “Trends in black carbon and size-resolved particle number concentrations and vehicle emission factors under real-world conditions”, Atmospheric Environment, Vol. 165, pp. 155-168.

Kumar, P., and Goel, A.(2016), “Concentration dynamics of coarse and fine particulate matter at and around signalised traffic intersections”, Environmental Science, Vol. 18, pp. 1220-1235.

Lee, J., Kim, J., & Ko, W. (2019), “Day-ahead electric load forecasting for the residential building with a small-size dataset based on a self-organizing map and a stacking ensemble learning method”. Applied Sciences, 9(6), pp. 1231.

Lee, S. H., Kwak, J. H., Lee, S. Y. and Lee J. H.(2015), “On-road chasing and laboratory measurements of exhaust particle emissions of diesel vehicles equipped with after treatment technologies (DPF, urea-SCR)”, International Journal of Automotive Technology, Vol. 16(4), pp. 551-559.

Lešnika, U., Mongus, D. and Jesenkob, D.(2019), "Predictive analytics of PM10 concentration levels using detailed traffic data", Transportation Research Part D, Vol. 37, pp. 141-154.

Liu, H., Rodgers, M. O. and Guensler, R.(2019), “The impact of road grade on vehicle accelerations behavior, PM2.5 emissions, and dispersion modeling”, Transportation Research Part D, Vol. 75, 2019, pp. 297-319.

Martinet, S., Liu, Y., Louis, C., Tassel, P., Perret, A., Chaumond, A. and André M.(2017), “EURO6 unregulated pollutant characterization and statistical analysis of after-treatment device and driving-condition impact on recent passenger-car emissions”, Environmental Science & Technology, Vol. 51, pp. 5847-5855.

Mohammad, Z. K., Batelaan, O., Fadaee, M., Hinkelmann, R.(2021), “Ensemble machine learning paradigms in hydrology: A review”, Journal of Hydrology, 598, 126266.

Mohammadi, A. H., Eslamimanesh, A., Gharagheizi, F., Richon, D.(2012), “A novel method for evaluation of asphaltene precipitation titration data”, Chemical Engineering Science, 78, pp. 181-185.

Mohammadi, A. H., Gharagheizi, F., Eslamimanesh, A., & Richon, D.(2012), “Evaluation of experimental data for wax and diamondoids solubility in gaseous systems”, Chemical engineering science, 81, pp. 1-7.

Molnar, C.,(2022), “Interpretable machine learning a guide for making black box models explainable”, https://christophm.github.io/interpretable-ml-book/feature-importance.html

Montella, A.(2011), “Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types”, Accident Analysis and Prevention, 43(4), pp. 1451-1463.

Müller, G., Guido, S.(2016), “Introduction to machine learning with python: a guide for data scientists”.

Natekin, A., Knoll, A.(2013), “Gradient boosting machines, a tutorial”, Frontiers in neurorobotics, 7, pp, 21.

Osorio, C., Nandur, K.(2015), “Urban transportation emissions mitigation: Coupling high-resolution vehicular emissions and traffic models for traffic signal optimization”, Transportation Research Part B, Vol. 81, pp. 520-538.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A.(2018), “CatBoost: unbiased boosting with categorical features”, In Advances in Neural Information Processing Systems, pp. 6638-6648.

Quiros, D. C., A. Thiruvengadam, S. Pradhan, M. Besch, P. Thiruvengadam, and B. Demirgok et al.(2016), “Real-world emissions from modern heavy-duty diesel, natural gas, and hybrid deisel trucks operating along major California freight corridors,” Emission Control Science and Technology, Vol. 2(3), pp. 156-172,

Rasmussen, C. E.(2003), “Gaussian processes in machine learning. In summer school on machine learning”, Springer, Berlin, Heidelberg.

Rokach, L.(2010), “Ensemble-based classifiers”, Artificial Intelligence Review, 33, pp. 1-39.

Rönkkö, T., Virtanen, A., Vaaraslahti, K., Keskinen, J., Pirjola, L. and Lappi, M.(2006), "Effect of dilution conditions and driving parameters on nucleation mode particles in diesel exhaust: laboratory and on-road study", Atmospheric Environment, Vol. 40, pp. 2893 - 2901.

Schapire, R.(1990), "The Strength of waek learnability", Machine Learning, 5(2), pp. 197-227.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., De Freitas, N. (2015), "Taking the human out of the loop: a review of Bayesian optimization", Proceedings of the IEEE, 104(1), pp. 148-175.

Siroky, D. S.(2009), "Navigating random forests and related advances in algorithmic modeling. Statistics Surveys", 3, pp. 147-163.

Snoek, J., Larochelle, H., Adams, R. P.(2012), "Practical bayesian optimization of machine learning algorithms", Advances in neural information processing systems, 25, pp. 1-9.

Suleiman, A., Tight, M. R. and Quinnm, A. D.(2019), "Applying machine learning methods in managing urban concentrations of traffic-related particulate matter(PM10 and PM2.5)", Atmospheric Pollution Research, Vol. 10, pp. 134-144.

Timofeev, R.(2004), "Classification and regression trees (CART) theory and applications", Humboldt University, Berlin

Valiant, L.(1984), "Deductive Learning [and Discussion]", Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences, 312(1522), pp. 441-446.

Wang, Y., Wang, T.(2020), "Application of improved LightGBM model in blood glucose prediction", Applied Sciences, 10(9), pp. 2076-3417.

Wolpert, D. (1992), "Stacked generalization". Neural Networks, 5(2), pp. 241-259.

Xu, J., Saleh, M. and Hatzopoulou, M.(2020), “A machine learning approach capturing the effects of driving behaviour and driver characteristics on trip-level emissions”, Atmospheric Environment, Vol. 2Interpretable Machine Learning

부 록

<표 부록-1> 인구수별 기준에 따른 지자체별 적용대상 지역계수('15년 7월 기준)

시도	시군구	기준 인구수	지역계수
서울특별시	전체	500만이상	1.53
부산광역시	전체	100만이상 500만미만	1
부산광역시	기장군	10만미만	0.4
대구광역시	전체	100만이상 500만미만	1
대구광역시	달성군	10만미만	0.4
인천광역시	전체	100만이상 500만미만	1
인천광역시	강화군	10만미만	0.4
인천광역시	옹진군	10만미만	0.4
광주광역시	전체	100만이상 500만미만	1
대전광역시	전체	100만이상 500만미만	1
울산광역시	전체	100만이상 500만미만	1
세종특별자치시	전체	10만이상 50만미만	0.85
경기도	수원시	100만이상 500만미만(특별권역)	1
경기도	성남시	50만이상 100만미만	0.87
경기도	의정부시	10만이상 50만미만(특별권역)	0.87
경기도	안양시	50만이상 100만미만(특별권역)	0.87
경기도	부천시	50만이상 100만미만(특별권역)	0.87
경기도	광명시	10만이상 50만미만(특별권역)	0.87
경기도	평택시	10만이상 50만미만(특별권역)	0.87
경기도	동두천시	10만미만(특별권역)	0.87
경기도	안산시	50만이상 100만미만(특별권역)	0.87
경기도	고양시	100만이상 500만미만	1
경기도	과천시	10만미만(특별권역)	0.87
경기도	구리시	10만이상 50만미만(특별권역)	0.87
경기도	남양주시	10만이상 50만미만(특별권역)	0.87
경기도	오산시	10만이상 50만미만(특별권역)	0.87
경기도	시흥시	10만이상 50만미만(특별권역)	0.87
경기도	군포시	10만이상 50만미만(특별권역)	0.87
경기도	의왕시	10만이상 50만미만(특별권역)	0.87
경기도	하남시	10만이상 50만미만(특별권역)	0.87
경기도	용인시	50만이상 100만미만(특별권역)	0.87
경기도	파주시	10만이상 50만미만(특별권역)	0.87
경기도	이천시	10만이상 50만미만(특별권역)	0.87
경기도	안성시	10만이상 50만미만	0.85
경기도	김포시	10만이상 50만미만(특별권역)	0.87
경기도	화성시	10만이상 50만미만(특별권역)	0.87
경기도	광주시	10만이상 50만미만	0.85
경기도	양주시	10만이상 50만미만(특별권역)	0.87
경기도	포천시	10만이상 50만미만	0.85

<표 계속>

시도	시군구	기준 인구수	지역계수
강원도	속초시	10만미만	0.4
강원도	삼척시	10만미만	0.4
강원도	홍천군	10만미만	0.4
강원도	횡성군	10만미만	0.4
강원도	영월군	10만미만	0.4
강원도	평창군	10만미만	0.4
강원도	정선군	10만미만	0.4
강원도	철원군	10만미만	0.4
강원도	화천군	10만미만	0.4
강원도	양구군	10만미만	0.4
강원도	인제군	10만미만	0.4
강원도	고성군	10만미만	0.4
강원도	양양군	10만미만	0.4
충청북도	청주시	50만이상 100만미만	0.87
충청북도	충주시	10만이상 50만미만	0.85
충청북도	제천시	10만이상 50만미만	0.85
충청북도	청원군	10만이상 50만미만	0.40
충청북도	보은군	10만미만	0.4
충청북도	옥천군	10만미만	0.4
충청북도	영동군	10만미만	0.4
충청북도	증평군	10만미만	0.4
충청북도	진천군	10만미만	0.4
충청북도	괴산군	10만미만	0.4
충청북도	음성군	10만미만	0.4
충청북도	단양군	10만미만	0.4
충청남도	천안시	50만이상 100만미만	0.87
충청남도	공주시	10만이상 50만미만	0.85
충청남도	보령시	10만이상 50만미만	0.85
충청남도	아산시	10만이상 50만미만	0.85
충청남도	서산시	10만이상 50만미만	0.85
충청남도	논산시	10만이상 50만미만	0.85
충청남도	계룡시	10만미만	0.4
충청남도	금산군	10만미만	0.4
충청남도	연기군	10만미만	0.4
충청남도	부여군	10만미만	0.4
충청남도	서천군	10만미만	0.4
충청남도	청양군	10만미만	0.4
충청남도	홍성군	10만미만	0.4
충청남도	예산군	10만미만	0.4
충청남도	태안군	10만미만	0.4
충청남도	당진시	10만이상 50만미만	0.85
전라북도	전주시	50만이상 100만미만	0.87
전라북도	군산시	10만이상 50만미만	0.85

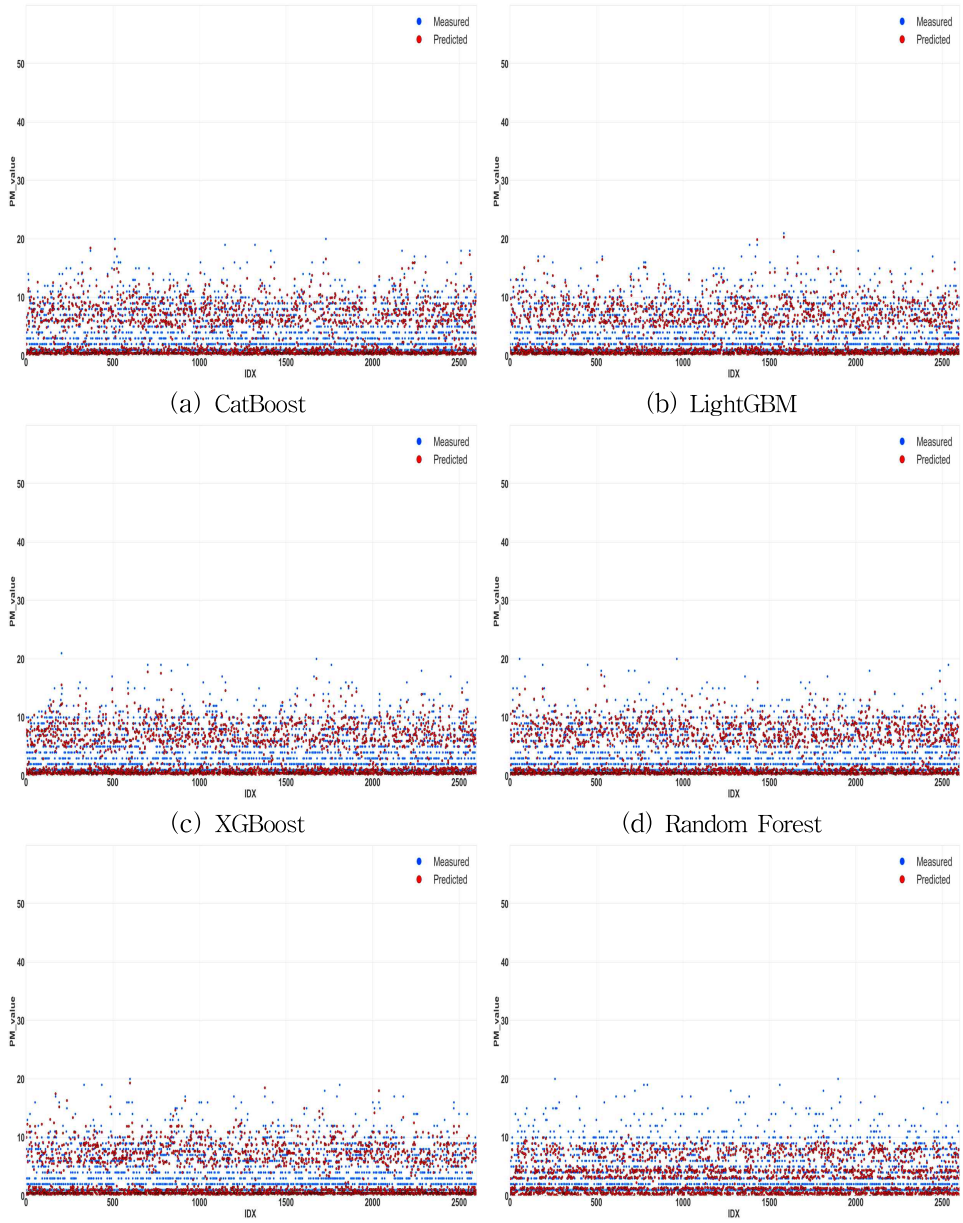
<표 계속>

시도	시군구	기준 인구수	지역계수
전라북도	익산시	10만이상 50만미만	0.85
전라북도	정읍시	10만이상 50만미만	0.85
전라북도	남원시	10만미만	0.4
전라북도	김제시	10만미만	0.4
전라북도	완주군	10만미만	0.4
전라북도	진안군	10만미만	0.4
전라북도	무주군	10만미만	0.4
전라북도	장수군	10만미만	0.4
전라북도	임실군	10만미만	0.4
전라북도	순창군	10만미만	0.4
전라북도	고창군	10만미만	0.4
전라북도	부안군	10만미만	0.4
전라남도	목포시	10만이상 50만미만	0.85
전라남도	여수시	10만이상 50만미만	0.85
전라남도	순천시	10만이상 50만미만	0.85
전라남도	나주시	10만미만	0.4
전라남도	광양시	10만이상 50만미만	0.85
전라남도	담양군	10만미만	0.4
전라남도	곡성군	10만미만	0.4
전라남도	구례군	10만미만	0.4
전라남도	고흥군	10만미만	0.4
전라남도	보성군	10만미만	0.4
전라남도	화순군	10만미만	0.4
전라남도	장흥군	10만미만	0.4
전라남도	강진군	10만미만	0.4
전라남도	해남군	10만미만	0.4
전라남도	영암군	10만미만	0.4
전라남도	무안군	10만미만	0.4
전라남도	함평군	10만미만	0.4
전라남도	영광군	10만미만	0.4
전라남도	장성군	10만미만	0.4
전라남도	완도군	10만미만	0.4
전라남도	진도군	10만미만	0.4
전라남도	신안군	10만미만	0.4
경상북도	포항시	50만이상 100만미만	0.87
경상북도	경주시	10만이상 50만미만	0.85
경상북도	김천시	10만이상 50만미만	0.85
경상북도	안동시	10만이상 50만미만	0.85
경상북도	구미시	10만이상 50만미만	0.85
경상북도	영주시	10만이상 50만미만	0.85
경상북도	영천시	10만이상 50만미만	0.85
경상북도	상주시	10만이상 50만미만	0.85
경상북도	문경시	10만미만	0.4

<표 계속>

시도	시군구	기준 인구수	지역계수
경상북도	경산시	10만이상 50만미만	0.85
경상북도	군위군	10만미만	0.4
경상북도	의성군	10만미만	0.4
경상북도	청송군	10만미만	0.4
경상북도	영양군	10만미만	0.4
경상북도	영덕군	10만미만	0.4
경상북도	청도군	10만미만	0.4
경상북도	고령군	10만미만	0.4
경상북도	성주군	10만미만	0.4
경상북도	칠곡군	10만이상 50만미만	0.4
경상북도	예천군	10만미만	0.4
경상북도	봉화군	10만미만	0.4
경상북도	울진군	10만미만	0.4
경상북도	울릉군	10만미만	0.4
경상남도	창원시	50만이상 100만미만	0.87
경상남도	창원시(마산)	10만이상 50만미만	0.85
경상남도	창원시(진해)	10만이상 50만미만	0.85
경상남도	진주시	10만이상 50만미만	0.85
경상남도	통영시	10만이상 50만미만	0.85
경상남도	사천시	10만이상 50만미만	0.4
경상남도	김해시	10만이상 50만미만	0.85
경상남도	밀양시	10만이상 50만미만	0.85
경상남도	거제시	10만이상 50만미만	0.85
경상남도	양산시	10만이상 50만미만	0.85
경상남도	의령군	10만미만	0.4
경상남도	함안군	10만미만	0.4
경상남도	창녕군	10만미만	0.4
경상남도	고성군	10만미만	0.4
경상남도	남해군	10만미만	0.4
경상남도	하동군	10만미만	0.4
경상남도	산청군	10만미만	0.4
경상남도	함양군	10만미만	0.4
경상남도	거창군	10만미만	0.4
경상남도	합천군	10만미만	0.4
제주특별자치도	제주시	10만이상 50만미만	0.85
제주특별자치도	서귀포시	10만이상 50만미만	0.85

2. 2차 앙상블 학습 예측모형 예측결과

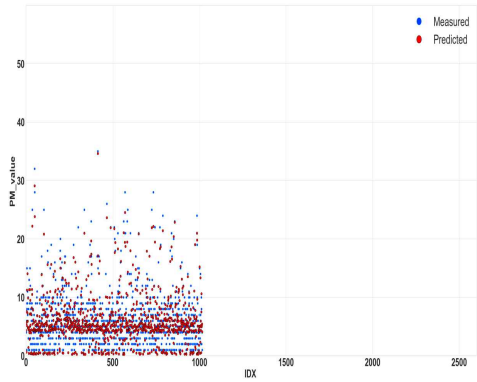


(a) CatBoost (b) LightGBM

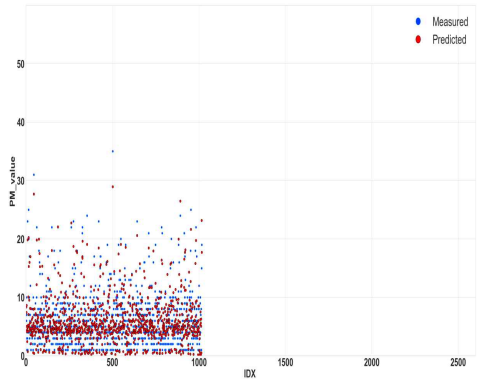
(c) XGBoost (d) Random Forest

(e) Decision Tree (f) Linear Regression

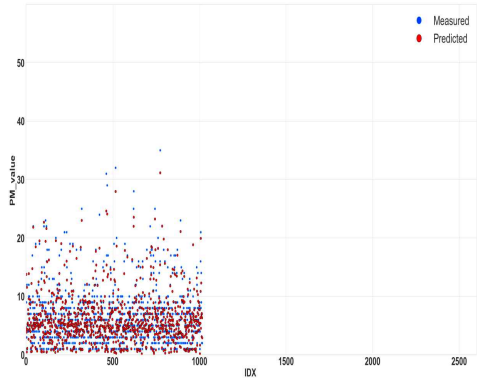
<그림 부록-1> 2차 앙상블 학습 예측모형 예측결과(KD-147모드 승용)



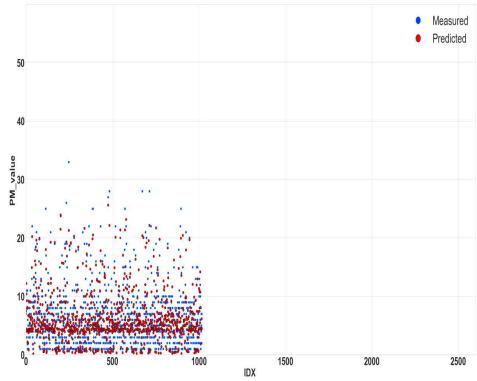
(a) CatBoost



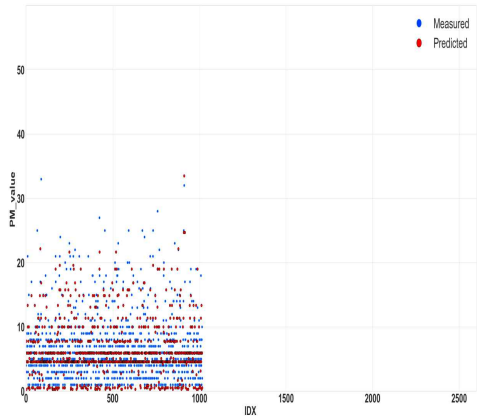
(b) LightGBM



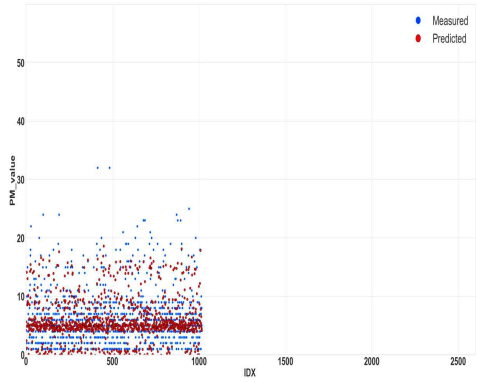
(c) XGBoost



(d) Random Forest

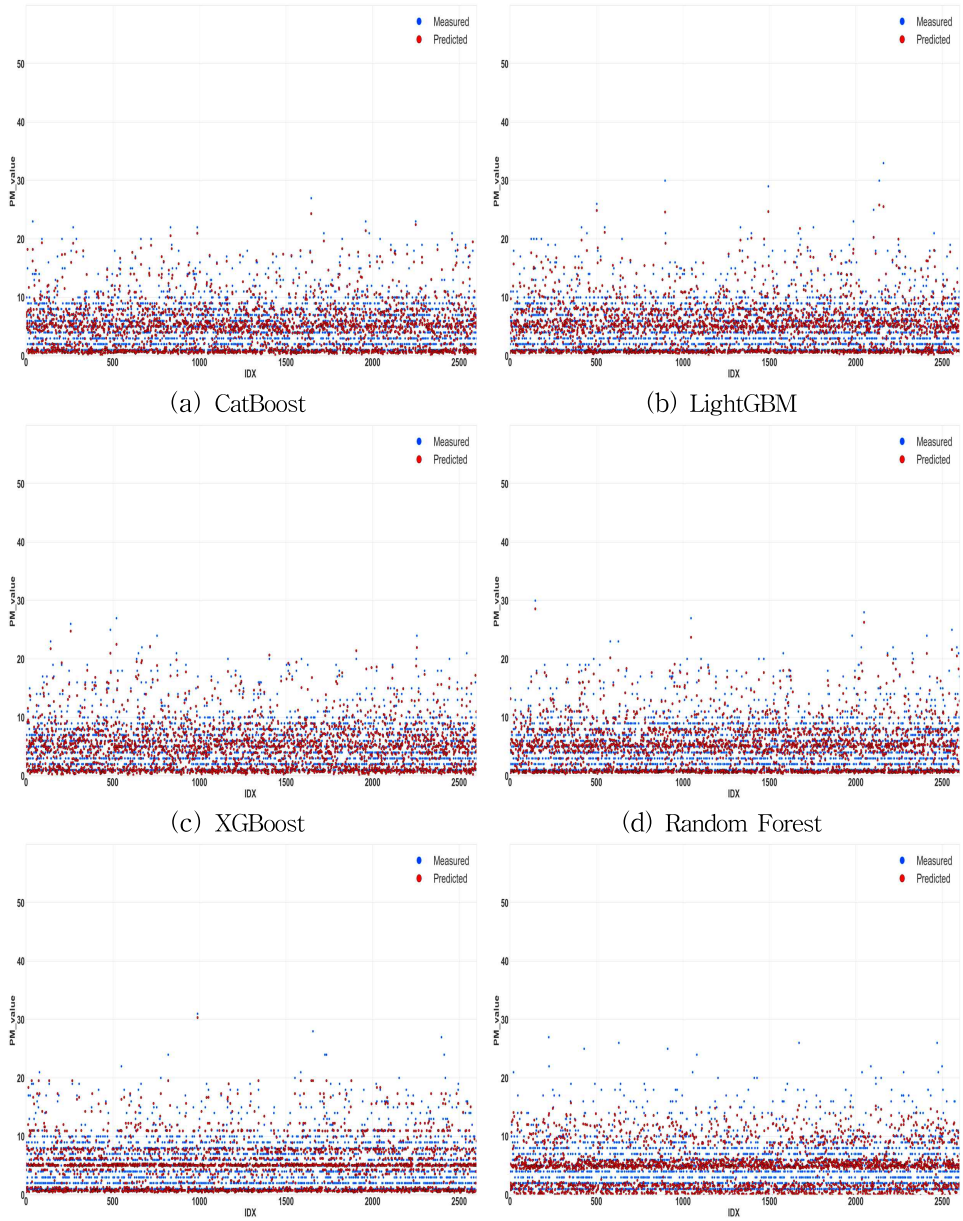


(e) Decision Tree



(f) Linear Regression

<그림 부록-2> 2차 앙상블 학습 예측모형 예측결과(KD-147모드 승합)

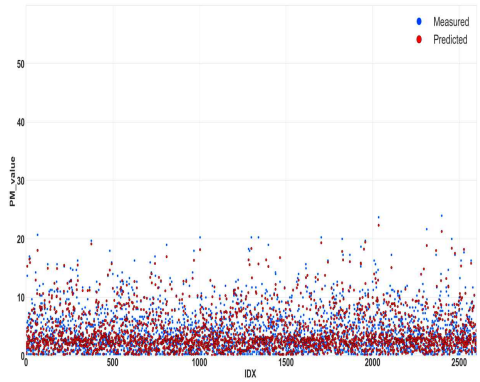


(a) CatBoost (b) LightGBM

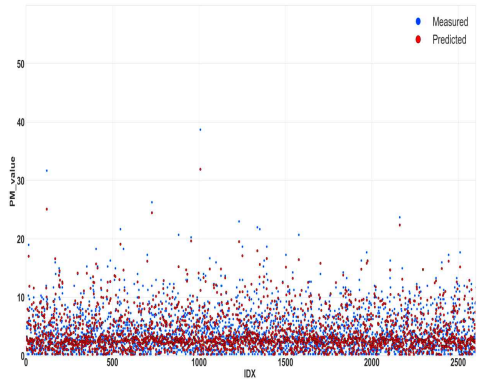
(c) XGBoost (d) Random Forest

(e) Decision Tree (f) Linear Regression

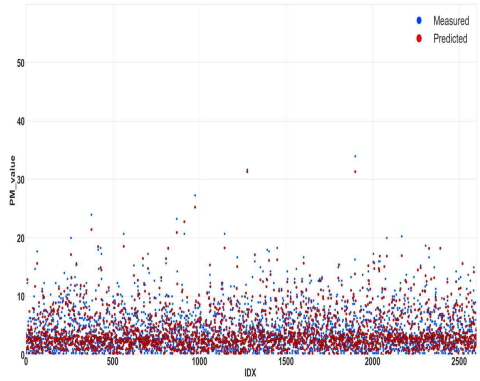
<그림 부록-3> 2차 앙상블 학습 예측모형 예측결과(KD-147모드 화물)



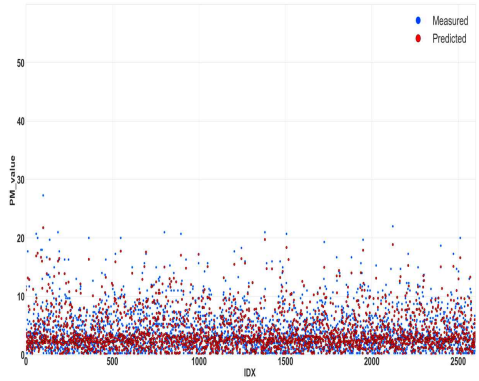
(a) CatBoost



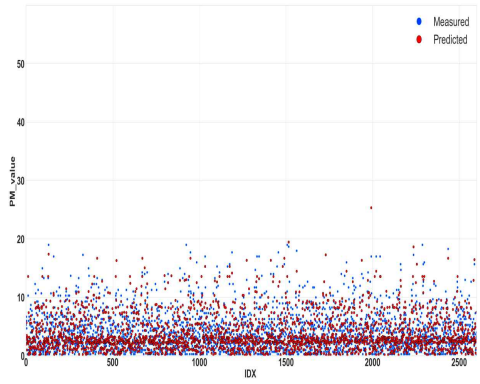
(b) LightGBM



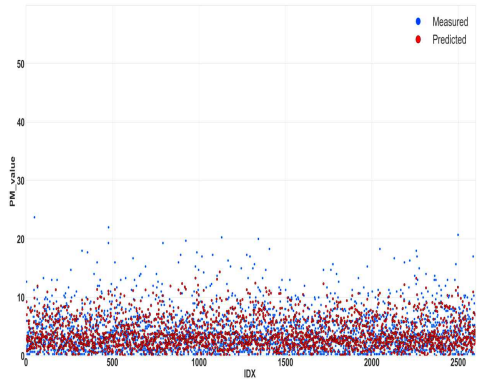
(c) XGBoost



(d) Random Forest

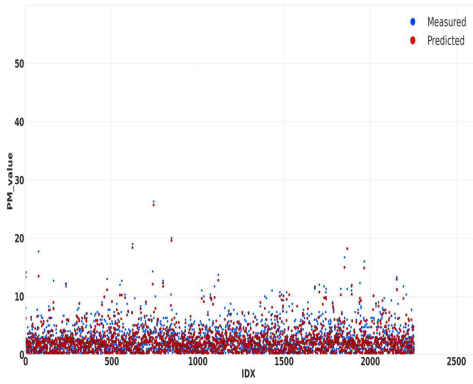


(e) Decision Tree

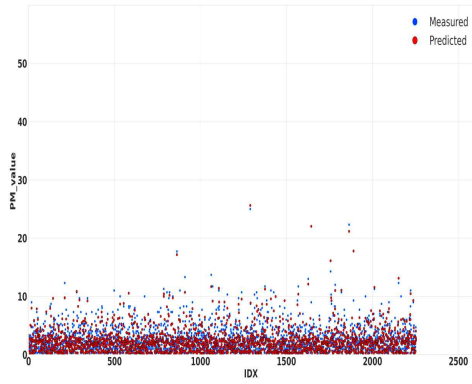


(f) Linear Regression

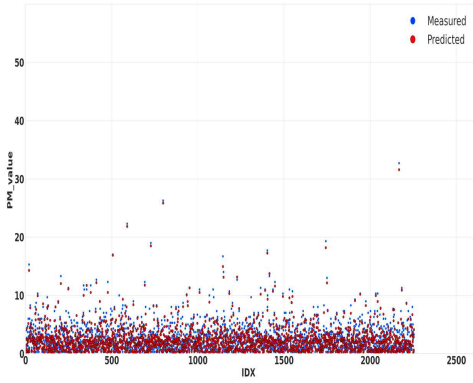
<그림 부록-4> 2차 앙상블 학습 예측모형 예측결과(Lug-Down3모드 화물)



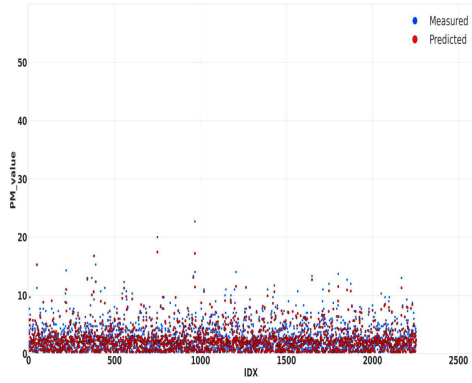
(a) CatBoost



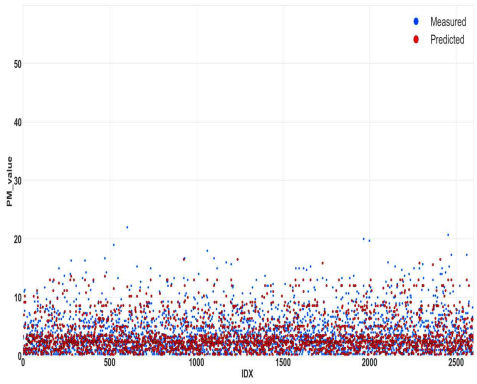
(b) LightGBM



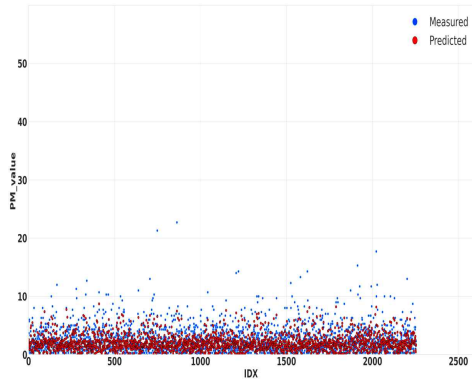
(c) XGBoost



(d) Random Forest

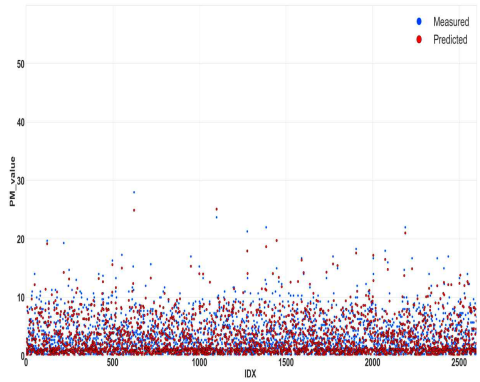


(e) Decision Tree

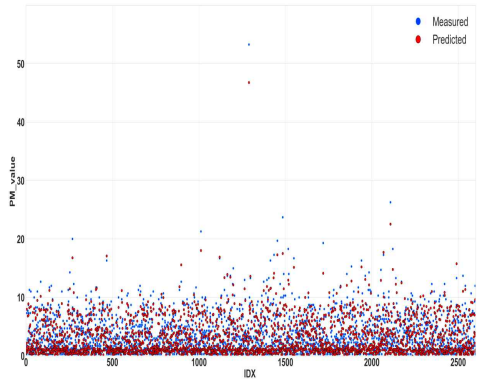


(f) Linear Regression

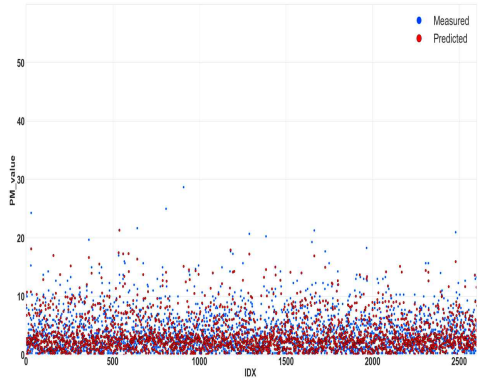
<그림 부록-5> 2차 앙상블 학습 예측모형 예측결과(Lug-Down3모드 승합)



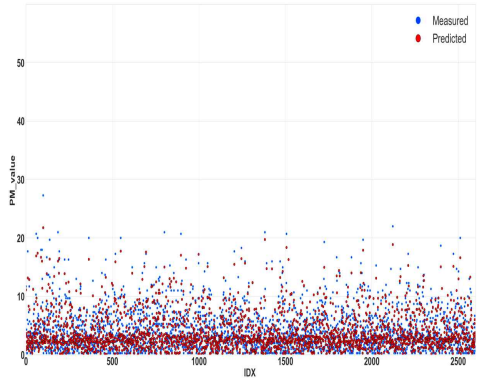
(a) CatBoost



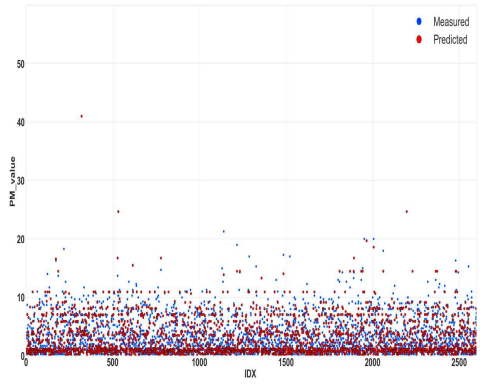
(b) LightGBM



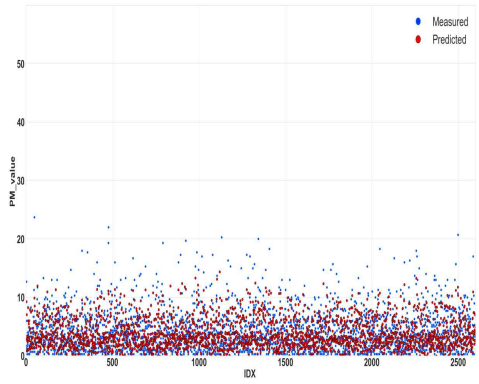
(c) XGBoost



(d) Random Forest



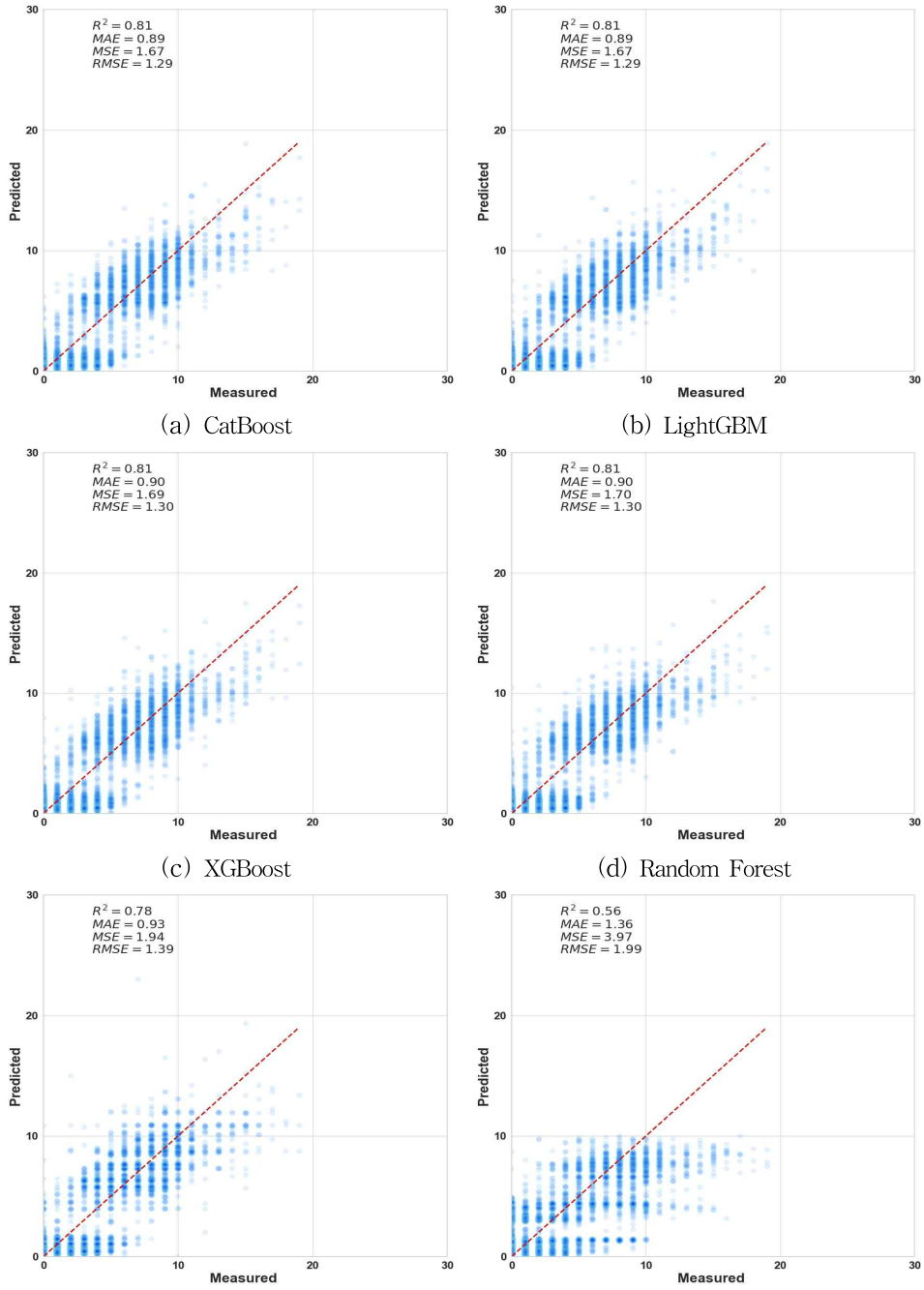
(e) Decision Tree



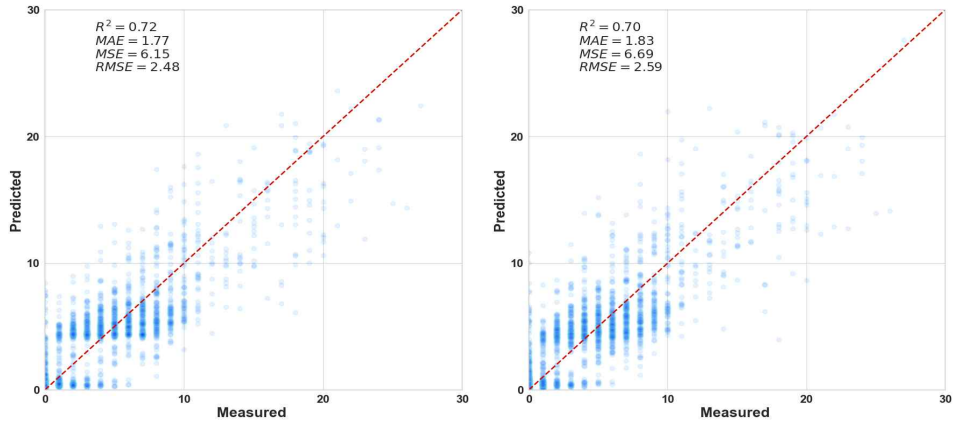
(f) Linear Regression

<그림 부록-6> 2차 앙상블 학습 예측모형 예측결과(Lug-Down3모드 특수)

3. 2차 앙상블 학습 차종별 예측모형 정확도

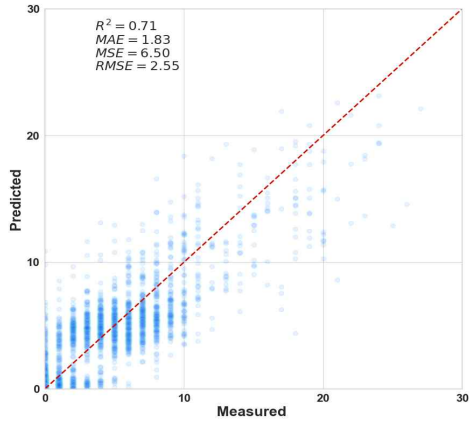


<그림 부록-7> 2차 앙상블 학습 예측모형 정확도(KD-147모드 승용)

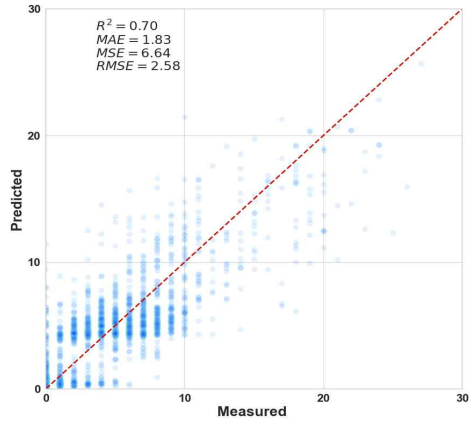


(a) CatBoost

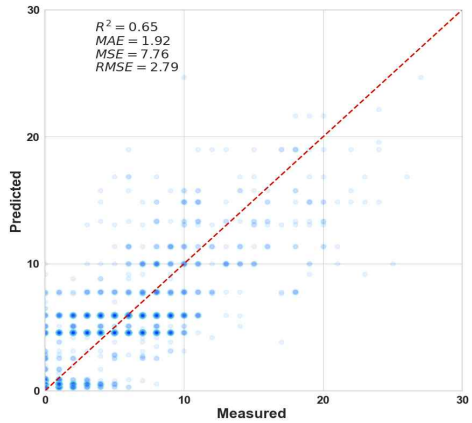
(b) LightGBM



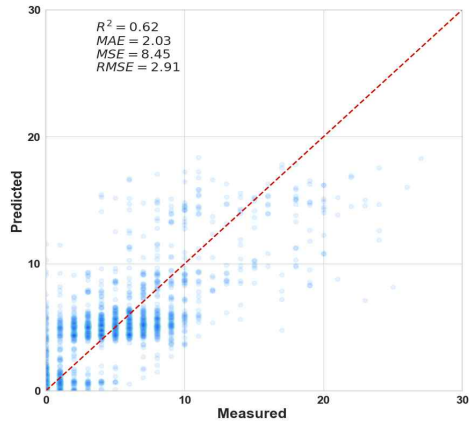
(c) XGBoost



(d) Random Forest

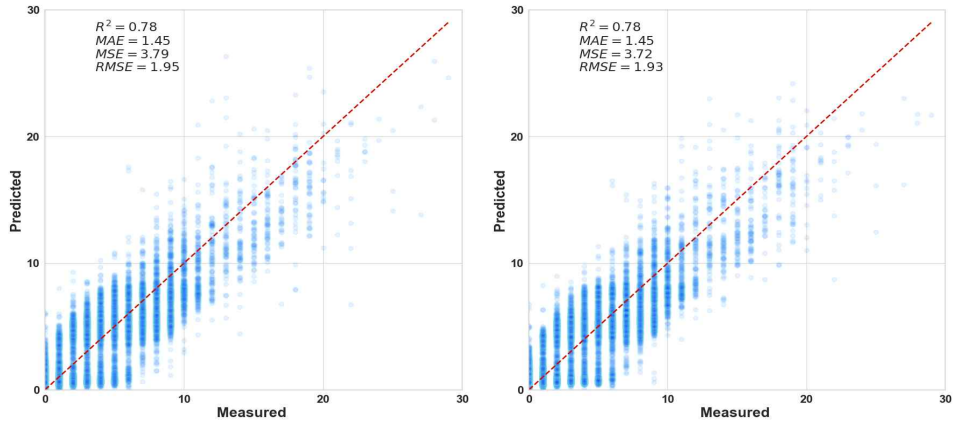


(e) Decision Tree



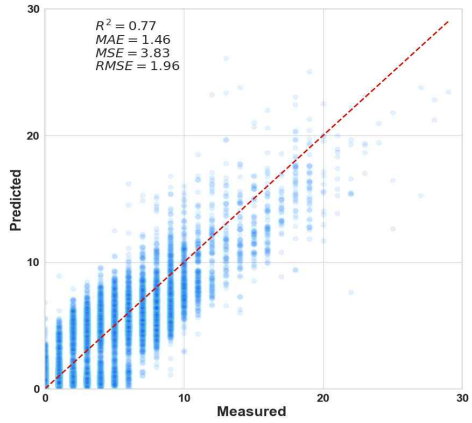
(f) Linear Regression

<그림 부록-8> 2차 앙상블 학습 예측모형 정확도(KD-147모드 승합)

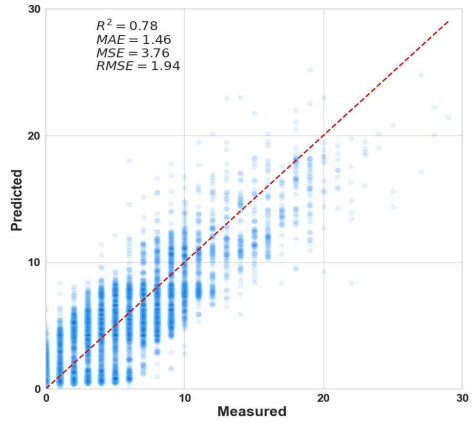


(a) CatBoost

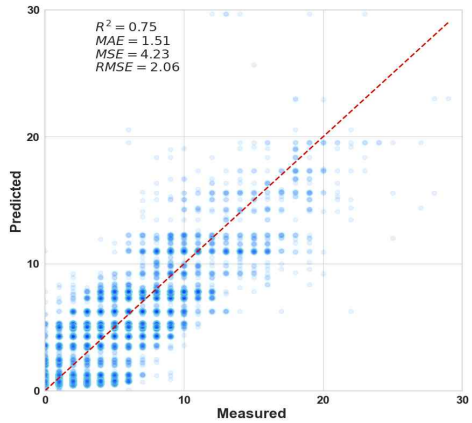
(b) LightGBM



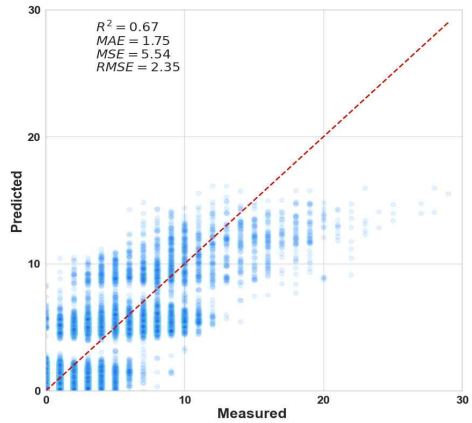
(c) XGBoost



(d) Random Forest

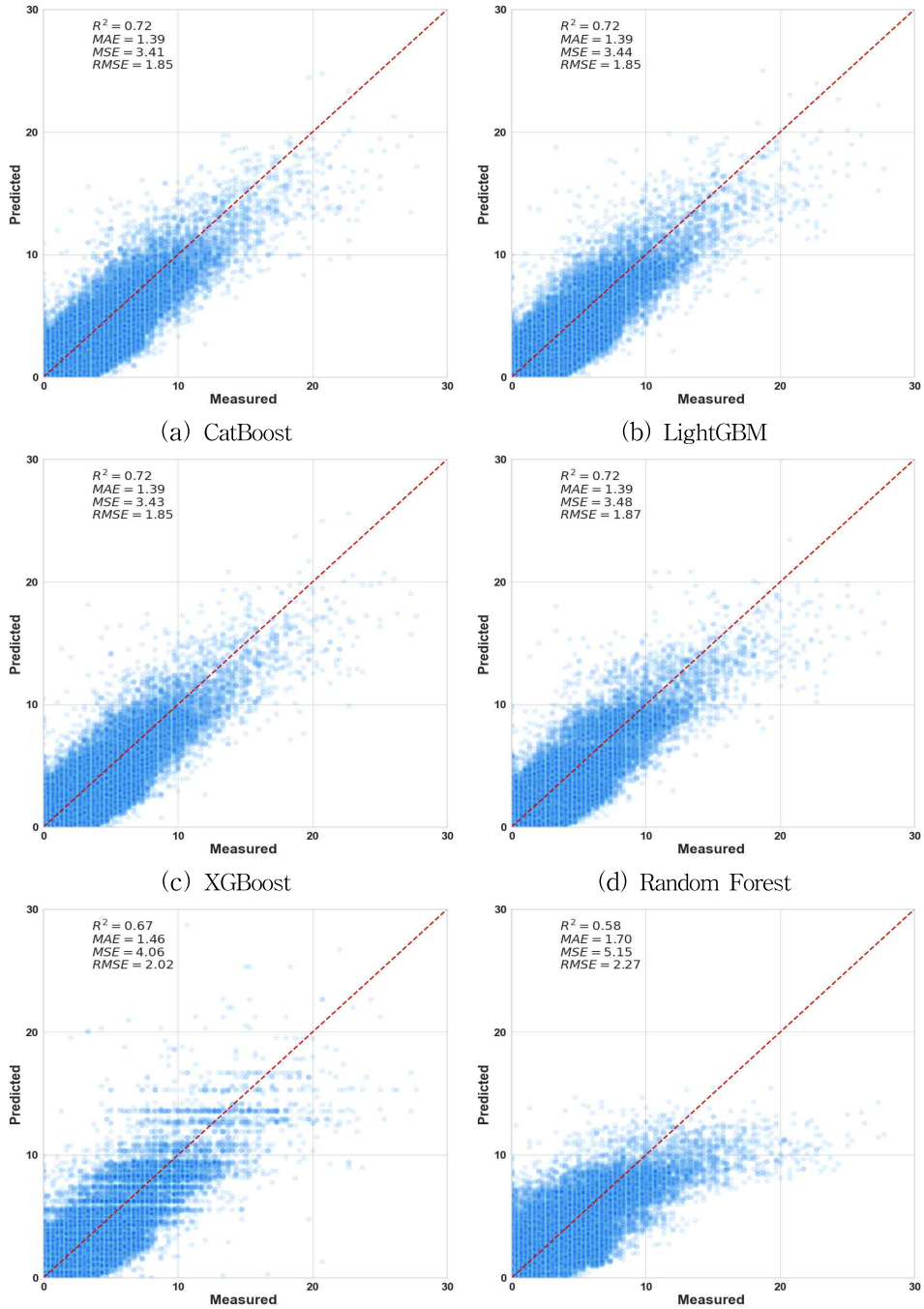


(e) Decision Tree



(f) Linear Regression

<그림 부록-9> 2차 앙상블 학습 예측모형 정확도(KD-147모드 화물)



(a) CatBoost

(b) LightGBM

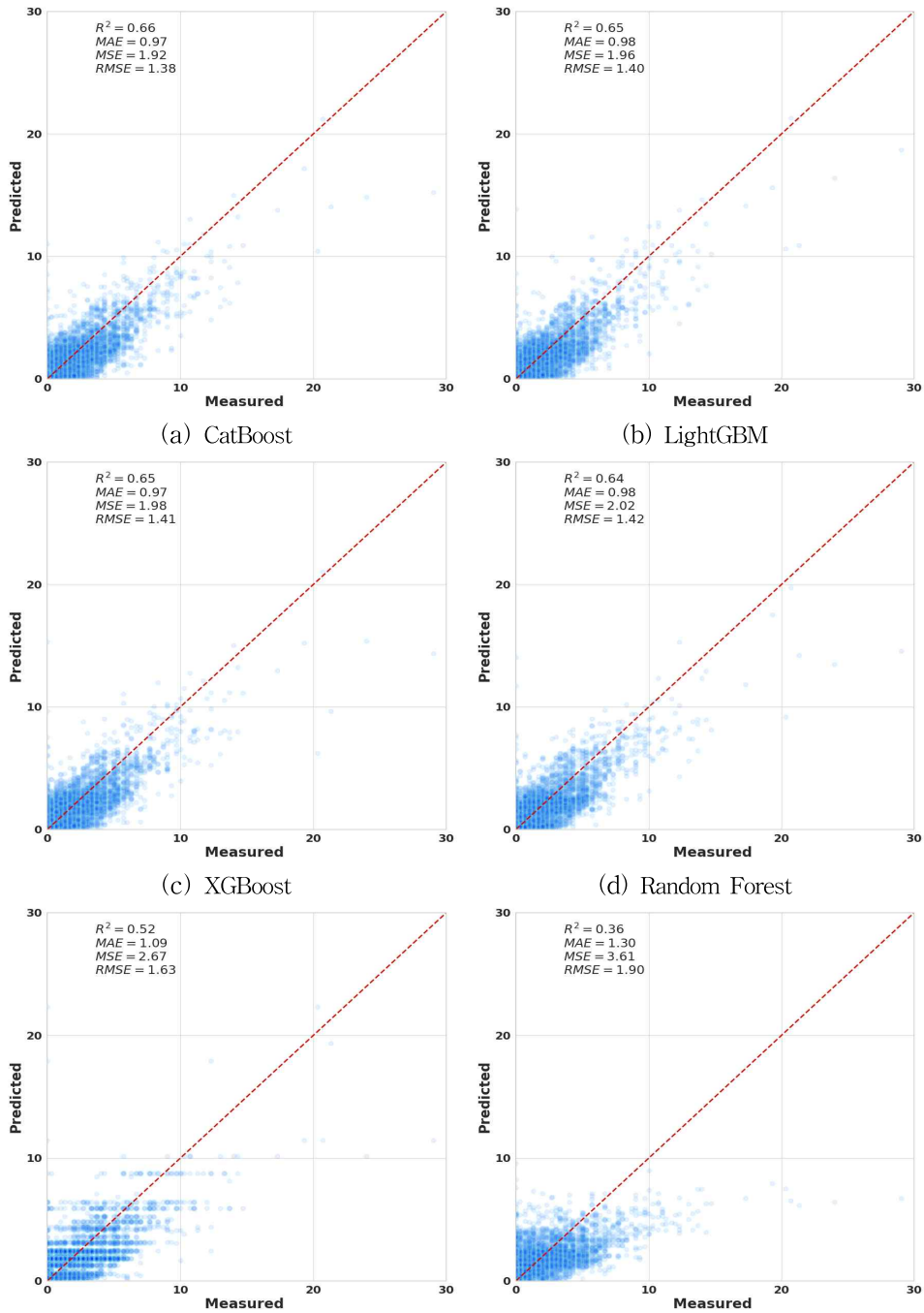
(c) XGBoost

(d) Random Forest

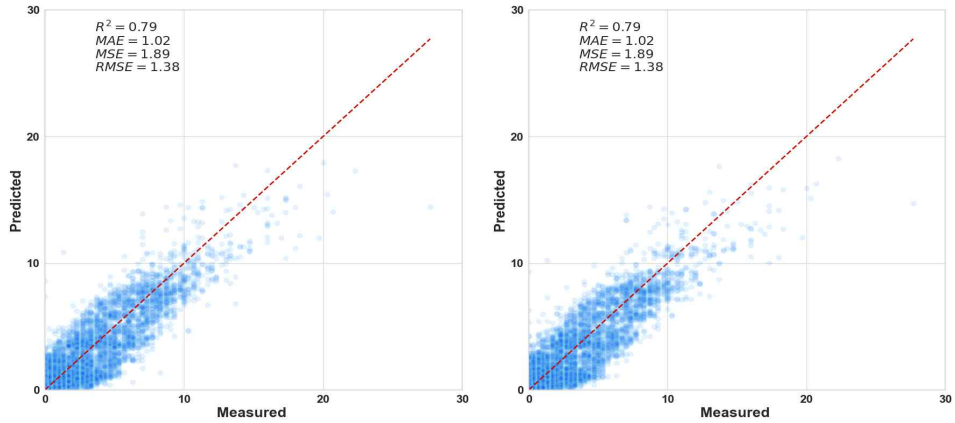
(e) Decision Tree

(f) Linear Regression

<그림 부록-10> 2차 앙상블 학습 예측모형 정확도(Lug-Down3모드 화물)

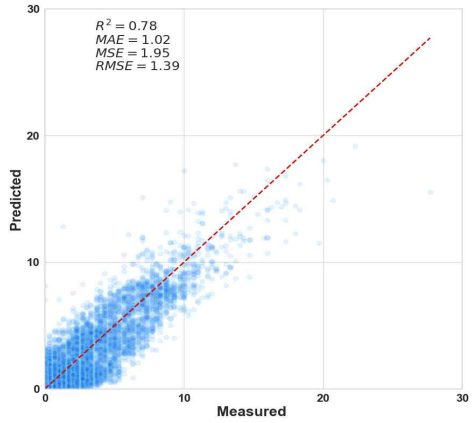


<그림 부록-11> 2차 앙상블 학습 예측모형 정확도(Lug-Down3모드 승합)

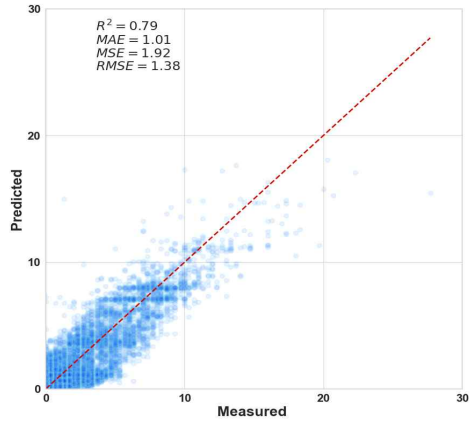


(a) CatBoost

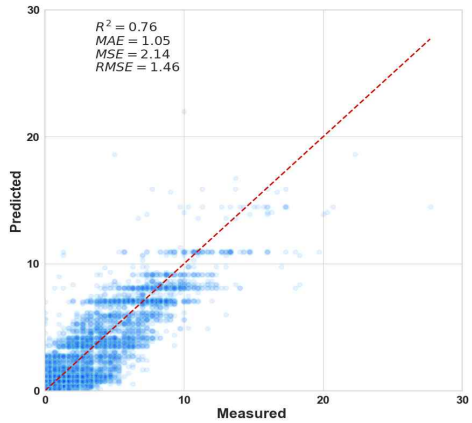
(b) LightGBM



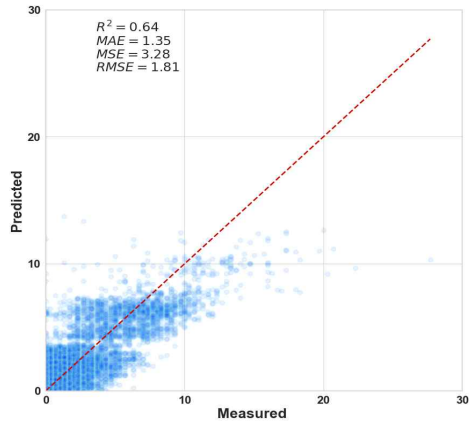
(c) XGBoost



(d) Random Forest



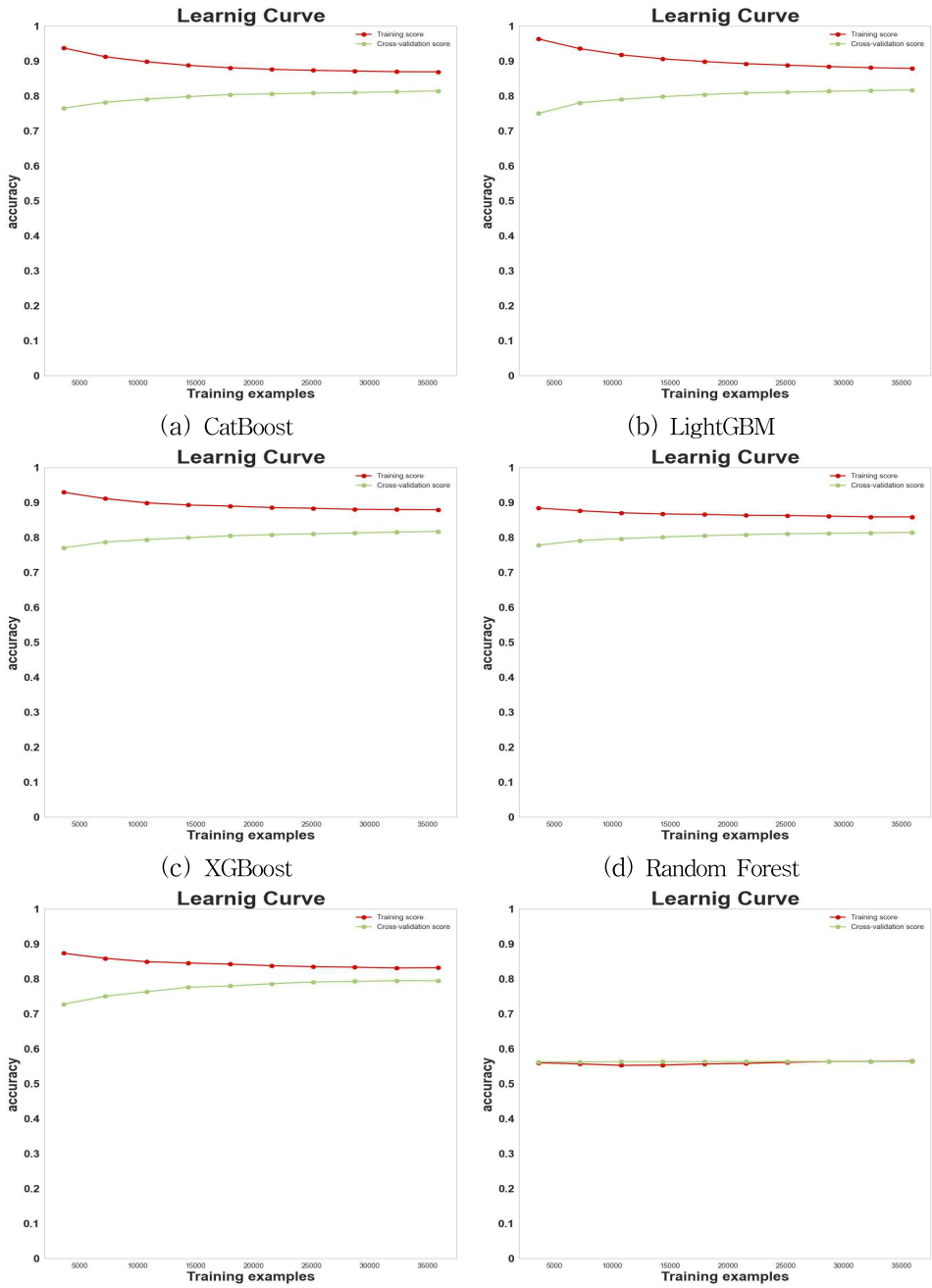
(e) Decision Tree



(f) Linear Regression

<그림 부록-12> 2차 앙상블 학습 예측모형 정확도(Lug-Down3모드 특수)

4. 2차 앙상블 학습 차종별 예측모형 학습곡선 검증결과

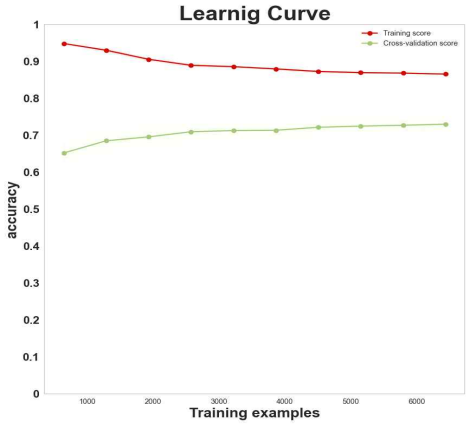


(a) CatBoost (b) LightGBM

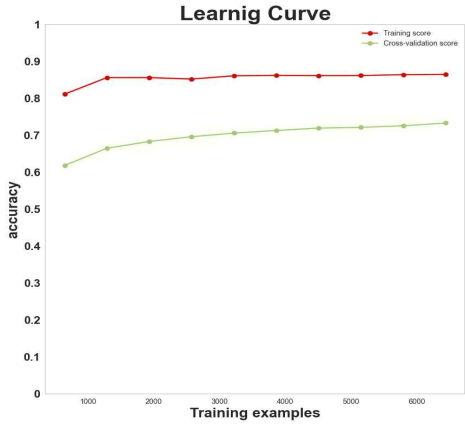
(c) XGBoost (d) Random Forest

(e) Decision Tree (f) Linear Regression

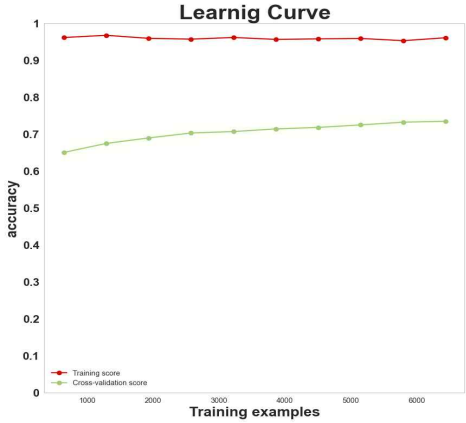
<그림 부록-13> 2차 앙상블 학습 예측모형 학습곡선(KD-147모드 승용)



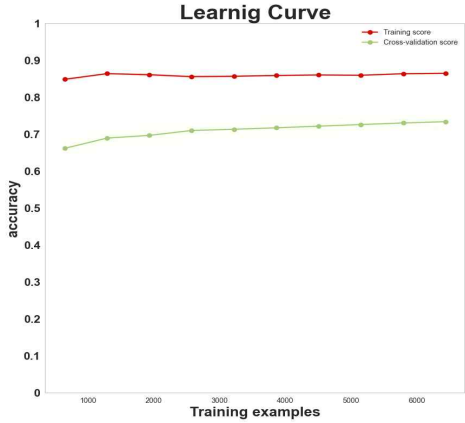
(a) CatBoost



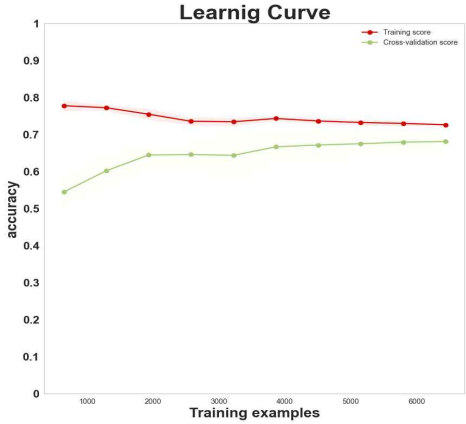
(b) LightGBM



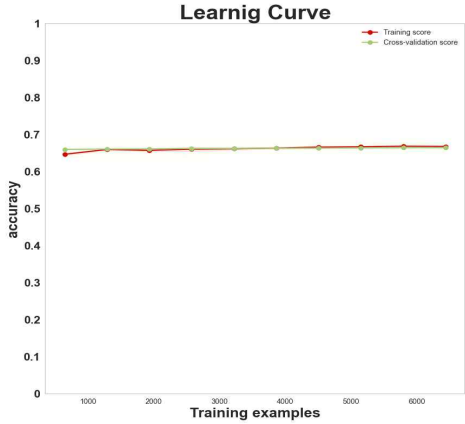
(c) XGBoost



(d) Random Forest

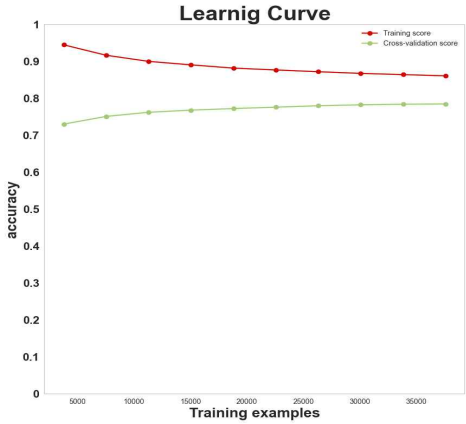


(e) Decision Tree

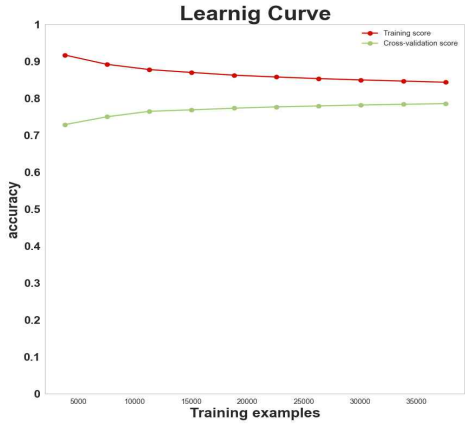


(f) Linear Regression

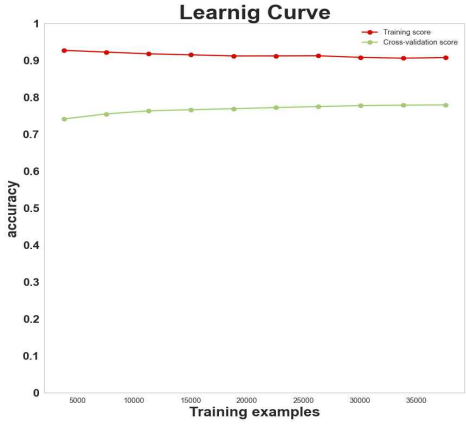
<그림 부록-14> 2차 앙상블 학습 예측모형 학습곡선(KD-147모드 승합)



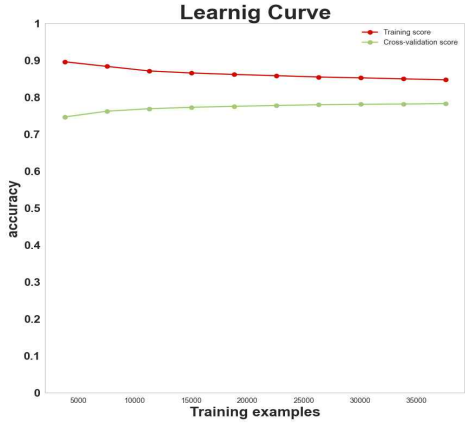
(a) CatBoost



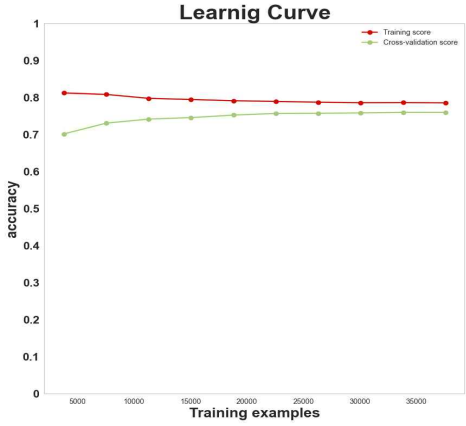
(b) LightGBM



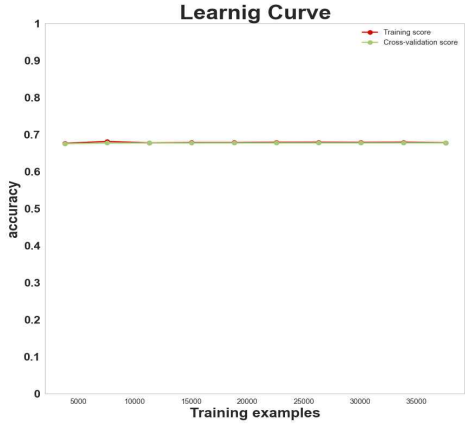
(c) XGBoost



(d) Random Forest

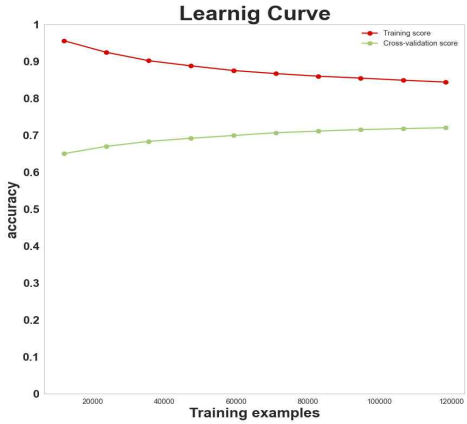


(e) Decision Tree

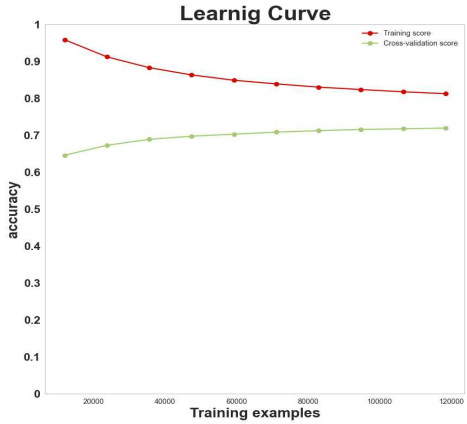


(f) Linear Regression

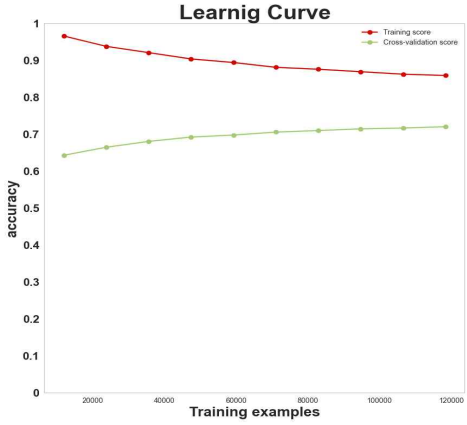
<그림 부록-15> 2차 앙상블 학습 예측모형 학습곡선(KD-147모드 화물)



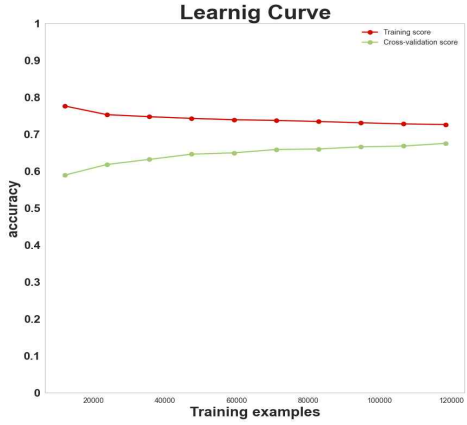
(a) CatBoost



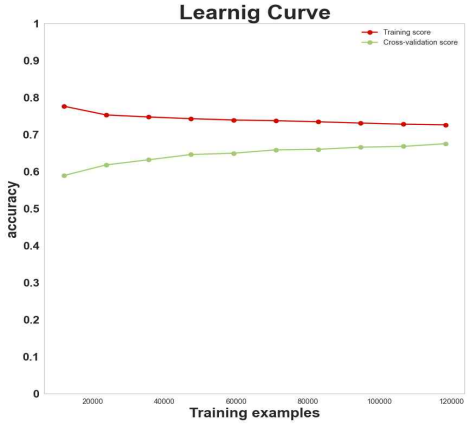
(b) LightGBM



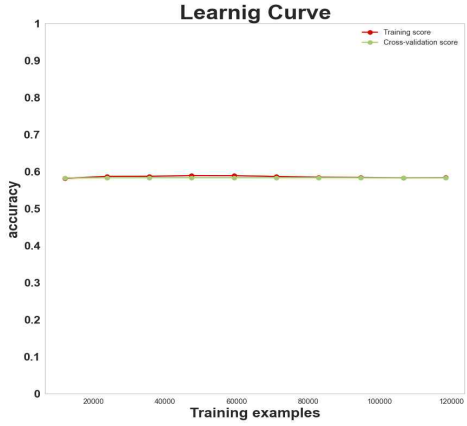
(c) XGBoost



(d) Random Forest

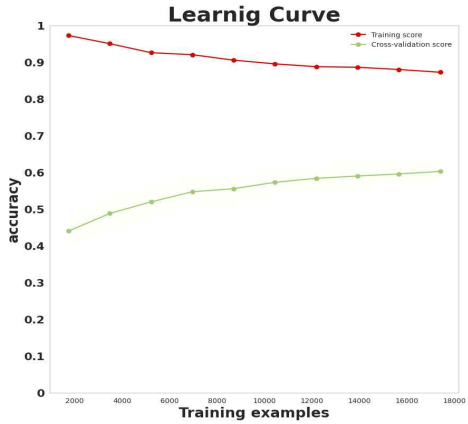


(e) Decision Tree

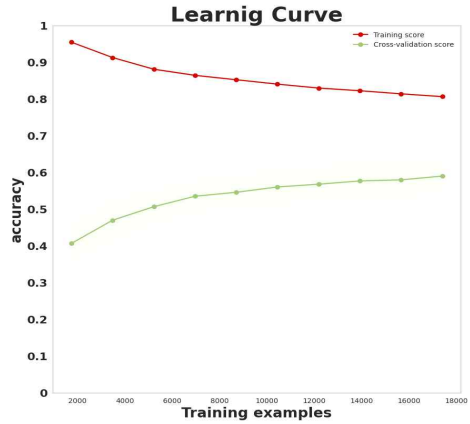


(f) Linear Regression

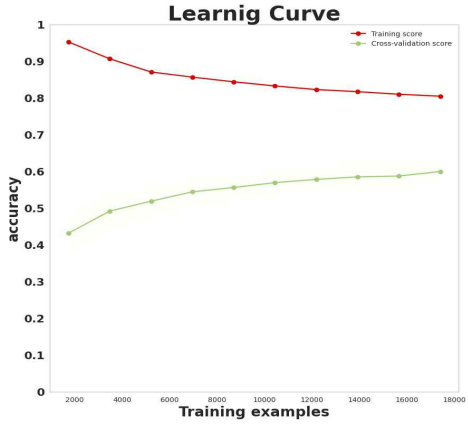
<그림 부록-16> 2차 앙상블 학습 예측모형 학습곡선(Lug-Down3모드 화물)



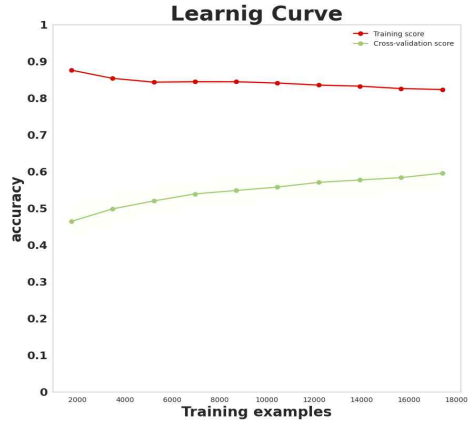
(a) CatBoost



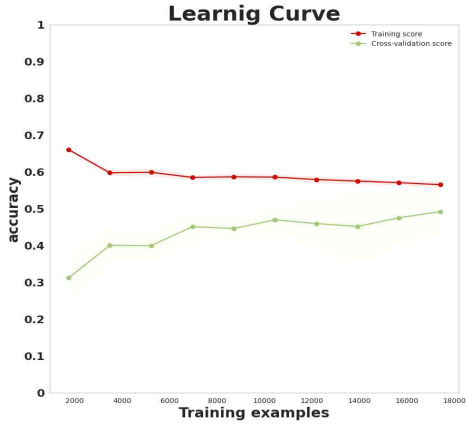
(b) LightGBM



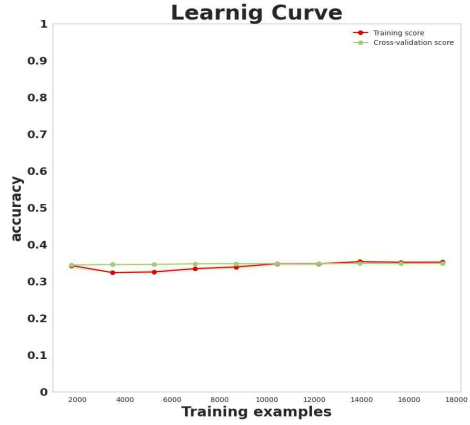
(c) XGBoost



(d) Random Forest



(e) Decision Tree



(f) Linear Regression

<그림 부록-17> 2차 앙상블 학습 예측모형 학습곡선(Lug-Down3모드 승합)

Abstract

Ensemble Learning to Predict Particulate Matter Concentrations Emitted by Diesel Vehicles

Lee, Sang-Jun
Transportation Studies Major
Department of Environmental Planning
Graduate School of Environmental Studies
Seoul National University

Diesel vehicles emit an amount of Particulate Matter(PM) compared to other vehicles due to their diesel engine characteristics. As of June 2019, there were 9.97 million diesel vehicles in Korea, accounting for 42.5% of the total in the nation. On the other hand, diesel vehicles account for only 1-3% of all vehicles in the U.S., China, and Japan. Therefore this study is focused on ways to reduce air pollution from diesel vehicles in Korea and is crucial policies for reducing PM. To achieve this goal, a basic study is needed to identify the key factors affecting PM emissions from diesel vehicles. The proposed prediction model aims to improve the accuracy of PM prediction, allowing for a better understanding of the contributing factors and the development of targeted policies.

This study also addresses the limitations of existing PM emission prediction models for diesel vehicles, which include their low accuracy with traditional statistical methods and complexity of the relationship between PM emissions and contributing factors. The authors propose a solution that involves applying machine learning techniques and

utilizing big data and controlled I/M(Inspection and Maintenance) data to enhance accuracy.

This study has three research goals, which were achieved in three stages. The first stage aimed to improve predictive performance with a prediction model using ensemble learning. The first stage of the study divided the ensemble learning prediction model into two modes: KD-147 and Lug-Down3. Analysis of 20 models involved classifying emission test pass/fail data using ensemble learning. These models included regression analysis, decision tree, random forest and three models representing CatBoost, LightGBM, and XGBoost. The statement implies that the performance of a predictive model was optimized by tuning its hyperparameters. Of the six models, the CatBoost model achieved the highest R^2 value at 0.815, which indicates a strong correlation between predicted and measured values. On the other hand, the linear regression model showed a lower R^2 value of 0.649, indicating weaker correlation between predicted and measured values. Hence, the statement highlights a significant difference in prediction performance between the two models.

In the second stage, permutation feature importance(PFI) was calculated for the PM emission prediction model for diesel vehicles using ensemble learning. This helped to identify the common PM emission factors, including Korean emission standards, fuel efficiency, displacement, and weight. The differences in the main factors for each vehicle type were found to be loading weight for special truck and the number of passengers for van. These findings show that the main factors affecting PM emissions align with the intended use of each vehicle type.

The third stage of the study aimed to reflect the main factors of diesel vehicle PM emissions in related policies. The purpose of this

case analysis was to use these main factors derived from an ensemble learning prediction model to inform PM reduction and environmental policies. The environmental improvement charge per vehicle was calculated based on the importance of each PM emission factor, and vehicles were classified into high, medium, and low concentrations in terms of their PM emissions. The study also evaluated how the environmental improvement charges per vehicle change by type of vehicle and region. This information can help with designing targeted policies to effectively reduce PM emissions from diesel vehicles and improve air quality.

In order to consider the equity of those subject to environmental improvement charges, weight coefficient and Korean emission standards coefficient were additionally applied to the calculation formula instead of the regional coefficient. Applying the derived Korean emission standards of this study and the PFI of the model year as weights made it possible to confirm the structure in which the levy was further transferred to the drivers of high-concentration PM emitting vehicles.

This study reviewed the predictive performance of PM emission prediction models for vehicles through ensemble learning and identified the main factors of PM emissions. The model can be used as basic data for evaluating the effectiveness of PM emission reduction policies or establishing other eco-friendly policies and strategies in the future.

Keywords: Diesel vehicle, PM emission, Emission factors, Inspection and Maintenance Data, Machine Learning, Ensemble Learning, Permutation Feature Importance

Student Number: 2010-31246