Journal of Information Technology and Architecture
Vol. 15. No. 4, December 2018, Pages 521-536

머신러닝 알고리즘을 이용한 기상 조건에 따른 노면 상태 예측 모델 연구

이민우[1] · 김영곤[1] · 김강화[1] · 전용주[1] · 용환성[2] · 이석준[3,*]

[1]디토닉 주식회사 기술연구소
[2]한양대학교 생산서비스경영학과
[3]건국대학교 경영대학
minwoo.lee@dtonic.io, kyg@dtonic.io, reinforcement@dtonic.io,
richard@dtonic.io, hsyong71@naver.com, seogjun@konkuk.ac.kr

(2018년 11월 21일 접수; 2018년 12월 16일 수정; 2018년 12월 19일 채택)

요약: 기상 변화는 도로상의 차량 안전에 큰 영향을 미친다. 기상상태가 변화함에 따라, 노면은 미끄러워질 수 있고, 미끄러운 노면은 차량의 제동거리를 증가시켜 운전자가 운전에 심혈을 기울여야 되기 때문이다. 이러한 기상 변수와 노면 상태 및 차량 안전성 간 관계의 복잡성에 따라 기상 변화에 따른 차량 안전성을 예측하기 위해서는 머신 러닝 모델 도입 필요성이 제기되고 있다. 본 연구에서는 기존 기상 관측 장비가 없는 지역(미계측 지역)에서의 상세한 기상 데이터 및 노면 상태를 관측해 차량 안전성을 예측하는 최적의 기계학습 모델을 개발하였다. 이를 위해, 도로의 기하 정보와 ASOS의 기상정보를 활용하여 설명변수를 가공하였다. 또한, 모델 평가 기준에 합당한 검증 방식을 적용하여, 가장 합리적인 머신러닝 모델을 선정하였다. 그 결과, 특정 지역에 대해서 1400개의 데이터로 80% 이상의 정확도로 노면 상태를 예측할 수 있음을 확인했다. 본 연구를 통해, 미계측 지역에 대한 노면 상태를 예측하고, 그에 대한 마찰력을 유추한다면, 해당 도로의 위험성을 운전자에게 알리고, 사고 위험도를 낮춰 사회적 비용을 감소시킬 수 있다.

주제어: 도로기상, 빅데이터, 도로 노면상태, 머신러닝, 모델 평가

Prediction of Road Surface State Caused by Weather Condition Using Machine Learning Model

Min-Woo Lee[1], Young-Gon Kim[1], Kang-Hwa Kim[1], Yong-Joo Jun[1],
Hwan-Seong Yong[2], and SeogJun Lee[3,*]

[1]D-Tonic, Korea
[2]Department of Business Administration, Hanyang University
[3]Business Dept, Konkuk University, Seoul, Korea

(Received November 21, 2018; Revised December 16, 2018; Accepted December 19, 2018)

Abstract: The meteorological change affects vehicle safety. Slippery road caused by the weather increases the braking distance of vehicle. In result, drivers should drive more carefully. Because of the complexity of the relationship among vehicle safety, road surface state, and meteorological factors, machine learning model is considered to predict weather related vehicle safety. In this paper, we develop a machine learning model predicting vehicle safety by collecting detail weather data and road surface state of an area through ASOS where no weather stations are installed(so-called unmeasured area). To select the most reasonable machine learning model, we define model eval-

*Corresponding author

uation criteria and apply them to various machine learning models. As a result, the selected model predicts road surface state of the target area with more than 80% of accuracy by using only 1,400 samples. If road surface state and friction of the road of the unmeasured area is predicted with high accuracy, social costs can be saved by decreasing accident risk through driver alert.

Keywords: Road Weather, Big Data, Road Surface State, Machine Learning, Model Evaluation

1. 서 론

도로의 기상상태는 당시의 도로가 얼마나 위험한지를 나타내는 중요한 척도다. 그 중, 도로의 노면 상태는 노면의 마찰력을 가늠할 수 있는 중요한 요소이다. 도로의 마찰력은 자동차의 제동 거리를 증가시켜, 운전자에게 안전 운전의 필요성을 요구하게 된다. 실제 지난 2017년 한 해 동안 발생한 216,335건 중 비나 눈이 오는 날에 발생한 날 총 12,362건의 교통사고가 발생했다[1]. 2017년 전국 평균 강수일수가 17.1일인 점을 고려했을 때, 하루 723건의 교통사고가 비나 눈이 올 때 발생한다. 이는 강수가 발생하지 않은 날 일간 교통사고 횟수가 593건인 점을 고려하면 높은 확률로 교통사고가 발생한다고 볼 수 있다. 따라서 도로 노면 상태를 예측하는 것은 운전자가 사전에 안전 운전이 필요한 지역을 인지하게 하고, 사고를 예방하는 데 중요한 역할을 할 수 있다.

일반적으로 도로 노면 상태를 예측할 때 도로 노면 상태를 설명할 수 있는 대표적인 변수는 기상 관측 데이터지만, 도로 노면의 상태는 같은 기후 조건에서 도로의 주변 환경에 따라 달라질 수 있다. 교량의 경우 평지구간과 달리 지열이 없고 태양열에 의해서만 열을 얻기 때문에 노면 결빙에 취약하다[2]. 도로 주변의 건축물, 나무 등은 그늘을 발생시켜 도로의 노면 상태 변화 속도를 감소시킨다[3]. 한국의 경우 도시화로 인해 고층 빌딩이 많이 건축되고, 도로 주변에는 가로수 역시 전국 도로의 33.3%를 차지하고 있다[4]. 한국의 국토의 경우, 70% 이상이 산지로 구성된 지형적 특성으로 인해 고속도로가 늘어나면서, 비탈면, 터널, 교량 등도 함께 증가하고 있다[5]. 따라서, 넓은 영역에서 도로의 노면 상태를 예측하기 위해서는 기상정보에 대한 설명변수뿐만 아니라, 도로의 주변 환경을 설명할 수 있는 변수가 필요하다.

하지만 도로의 주변 환경 중 주변 구조물에 대한 정보를 수집하는 것은 매우 어렵다. 같은 도로지만 가로수나, 건축물 등이 서로 다를 수 있으므로, 매우 촘촘한 간격의 정보 수집을 요구하기 때문이다. 또한, 아직 자동화하여 정보를 수집할 수 있는 기술의 부재 역시 해당 정보 수집을 어렵게 하는 요인이다. 다만, 센서 기술의 발전으로 인해 도로의 경사도나 도로의 선형에 대한 정보 같은 도로 기하구조 정보는 스마트폰이나 이동형 장비 등을 통해 수집할 수 있다. 그중에 특히 도로의 종단 경사도는 비가 내릴 경우, 물이 고일 수 있는 지점을 설명할 수 있으므로, 노면의 상태 변화 속도에 영향을 미칠 수 있다.

본 연구에서는 실제 도로의 기하구조가 노면 예측에 미치는 영향을 살펴보기 위해, 도로 기하 정보 중에 종단 경사도 정보를 가공 작업을 통해 설명변수로 활용하였다. 또한, 노면 상태에 직접적인 영향을 미치는 기상변수들을 시간과 공간에 관해 설명할 수 있도록 가공하여 설명변수로 사용하였다. 또한, 가공한 데이터 세트를 사용하여 선정된 모델을 학습하여 모델의 정확도를 검증하고, 모델 평가 기준에 근거하여, 본 연구에서 제안하는 설명변수와 가장 적합한 예측 모델을 선정한다.

2. 문헌 연구

2.1 기상 조건과 노면 상태 및 마찰력

기상 조건와 노면 상태, 노면 마찰력 등은 서로가 상관관계를 갖는다[6]. 특히 노면 상태와 마찰력은 강한 양의 상관관계를 보인다. 하지만 노면 상태와 마찰력을 사용하여 서로를 정확히 예측하는 것은 어렵다. 노면 상태를 통해 마찰력을 예측하는 경우, 노면 상태로 유추할 수 있는 것은 정성적 수치인 마찰계수이기 때문에 정량적 수치인 마찰력을 정확히 계산할 수 없다. 반면에, 마찰력을 활용하여 노면 상태를 예측하게 된다면, 도로포장 종류나 차종 등에 따라 달라지는 마찰력에 의해 잘못 예측하는 경우가 발생한다. 따라서 노면 변화 예측 관련 연구는 연구의 성격에 따라 마찰력을 예측할 것인지 노면 상태를 예측하는 것인지 선택하는 것이 필요하다. 본 절에서는 기상정보에 따른 노면의 마찰력 또는 노면 상태 등의 노면 변화를 예측한 관련 연구를 정리하였다.

Table 1. Features of Weather Observation Equipment

Type		Installation Number (EA)	Density (km^2)	Observation Element										
				DT	WD	WS	T	RH	P	PE	SC	LE	SE	RW
AWS	KMA	494	13	○	○	○	○	○	○	○	×	×	×	×
	SK tehx	1089	3	○	○	○	○	○	○	○	○	○	×	×
	Seoul City	26	13	○	○	○	○	○	○	○	×	×	×	×
ASOS		96	67	○	○	○	○	○	○	○	○	○	○	×
RWIS		35	197	○	○	○	○	○	○	○	X	X	○	○

Abbreviations Explain

DT : Date/Time / WD : Wind Direction / WS : Wind Speed T : Temperature / RH : Relative Humidity / P : Precipitation

PE : Precipitation Existence / SC : Sky Cord / LE : Lightning Existence / SE : Snow Existence / RW : Road Weather

2.1.1 기상 조건과 노면 상태 관련 연구

Kim et al.[7]은 기상인자별 사고분석 자료를 기반으로 실시간 기상정보와 교통정보를 연계하여 기상상황별 도로위험을 예보할 수 있는 노면상태 위험지도를 개발하였다. 기상정보는 기상청 지상기상관측망의 관측장비인 AWS (Automatic Weather System), ASOS (Automated Surface Observation System)에서 관측된 기상 데이터 (강수량, 기온, 평균풍속)를 사용하였고, 교통정보는 표준노드 링크 ID와 교통사고 사고정보 데이터를 결합하였다. 도로에 위험을 미치는 노면상태를 강우와 강설로 구분하였고, 강수지속시간, 기온, 해당 지점의 교통사고 유 무를 분석하여 노면 상태 위험도를 상·중·하로 구분하였다. Lee et al.[8]은 머신러닝 분류 모델을 이용하여 건조 (dry), 습윤 (moist), 젖음 (wet)노면으로 구분되는 노면 상태를 예측하였다. 머신러닝 분류 모델에 사용된 입력 데이터는 sk thex에서 운용 중인 서울지역 AWS에서 관측된 강수량, 기온, 평균 풍속, 습도 데이터를 사용하였다. Sin et al.[2]은 교량구간의 노면이 결빙되는 조건을 분석하였다. 노면센서를 교량에 부착하여 노면 온도, 대기 온도, 습도 등의 기상정보를 수집하였고, 수집된 기상정보를 통해 교량구간의 이슬점 온도를 계산하였다. 분석결과 교량지점의 결빙 발생조건은 이슬점온도가 0°C 아래에 형성되었을 때 노면 온도가 이슬점 온도 아래로 떨어지게 되면 노면이 결빙된다는 점을 알아내었다. Kim et al.[9]은 복합 센서에서 수집한 대기온도, 노면온도, 습도를 이용하여 이슬점 온도 및 습윤 정도를 판정하여 노면상태를 마름 (Dry), 젖음 (Wet), 결빙 위험 (Frozen danger), 적설 (Snow piled)로 분류하였고, 복합 센서 외에 적외선 카메라를 통합 운영하여 노면이 결빙되는 조건뿐만 아니라 강설 상태 및 강설 강도까지 구분하여 노면 상태를 분류할 수 있는 연구를 진행하였다. Berrocal et al.[10]은 노면상태가 결빙 (Ice)으로 변화하는 조건을 알아보기 위해 AWS에서 관측된 강수량과 기온데이터를 수집하였다. 수집된 AWS 데이터를 MCMC (Markov chain Monte Carlo) 알고리즘에 적용하였고, 대기온도가 0°C 이하 일 때, 강수가 존재 할 때 노면이 결빙된다는 조건을 발견하였다. Crevier et al.[11]은 RWIS에서 관측된 노면온도, 대기온도, 강수량 데이터를 도로상태 예측 수치 모델인 METRo(Model of the Environment and Temperature)의 입력 데이터로 사용하였다. METRo는 입력 데이터의 관측시간 기준 24시간 후의 노면상태를 습윤(Moist), 결빙 (Ice), 적설 (Snow)의 형태로 예측하였다.

2.1.2 기상 조건과 노면 마찰력

Kim et al.[6]은 기상 조건과 노면 마찰력 간의 상관관계를 입증하기 위해 AWS에서 측정된 강수량 데이터와 차량에 부착한 노면 센서에서 측정된 노면 마찰력 간의 상관관계를 분석하였고 강우량이 증가할수록 노면의 마찰력은 감소한다는 결과를 도출하였다. Lee et al.[12]은 기상청 AWS에서 관측된 강우량과 도로기하구조 조건을 조합하여 도로의 노면의 마찰력을 산정하는데 주요 인자로 사용되는 수막현상(Hydroplaning)을 예측하였다.

2.2 기상 조건 관측 방법

기상 관측 데이터란 특정 시점의 대기 상태를 나타내는 관측 목적에 맞는 기상요소(강수량, 기온, 기압 등)를 관측하여 기록한 데이터이다. 기상 관측 데이터는

주로 위험 지역(도심지, 산, 바다)을 평가하거나 위험 인자(폭우, 폭설, 낙뢰, 가뭄, 해일 등)를 예측하는 데 사용된다. 신뢰도 높은 기상 분석 및 예측을 수행하기 위해서는 관측 목적에 맞는 기상 관측 자료를 수집해야 한다. 기상청에서는 여러 관측 목적에 맞는 기상 관측 자료를 수집하기 위해 다양한 기상 관측망을 운용 중이며, 기상청 외에도 정부 기관, 지방자치단체, 민간기업에서도 지상기상관측망을 구축하고 있다[13].

2.2.1 기상 관측 목적에 따른 분류

관측망 대부분이 지상기상관측에 한정되어 있는 데 반해 기상청은 가장 다양한 기상 관측 목적에 따른 기상 관측망을 보유하고 있다. 기상청에서 정의한 관측 목적에 따른 관측망 분류는 Table 1과 같다.

방재기상관측 (AWS)

AWS는 국지적인 위험기상 현상을 탐지하는 장비이다. AWS는 높은 공간 해상도를 지닌 기상 데이터를 생산할 수 있으나, 자동 관측장비이기 때문에 적설량과 같은 실측 데이터가 부재하다는 단점을 가진다. 우리나라의 대표적인 AWS 관측망으로는 한국 기상청(KMA), SK techx, 서울시에서 운용 중인 AWS 관측망이 있다.

종관기상관측(ASOS)

종관기상관측(ASOS) 데이터는 자동 관측이 가능한 기상요소와 더불어 자동 관측이 어려운 현상(운량, 운형, 적설량 등)에 대해서 수동 관측한다. AWS와의 가장 큰 차이점으로는 지방청과 기상대에 설치되어 있고, 관리자가 상주하여 시정, 운량, 운형, 증발량, 지중온도, 적설량을 수동 관측한다. ASOS는 관리자가 주기적으로 장비를 점검하기 때문에, 기상관측장비의 이상 발생 시 자료수신 현황 및 긴급복구를 할 수 있어, 오류 데이터가 생성되는 것을 방지할 수 있다.

도로기상정보시스템 (RWIS)

RWIS는 어느 특정지역이 아닌 여러 지역을 통과하는 고속도로 노선 특성상 기존의 기상청 자료(광역기상정보) 만을 사용하기에 어려움이 발생하여 고속도로에 특화된 기상정보를 실시간 수집 · 가공 · 활용하기 위한 시스템이다[5]. RWIS는 AWS, ASOS가 관측하는 기상 관측과 더불어 노면 온도, 노면 수막 두께, 도로시정 등 노면 상태 예측에 필요한 데이터를 추가로 관측하여, 노면 상태 예측에 활용하기 위한 정보를 제공하는 장점이 있다. 그러나 특정 지점의 정보만을 수집하고 제공하기 때문에 전체 도로구간에 대한 노면 예측이 어렵고, 구축비용이 비싸다는 단점이 있다[14].

2.3 예측 모델 및 평가 기준

2.3.1 예측 모델(Machine Leaning)

머신러닝은 경험을 통해 학습하는 것을 컴퓨터가 수행할 수 있도록 가르치는 과정을 말한다. 경험을 통해 학습되기 때문에 미리 결정된 방정식을 모델로 의존하지 않고 계산 방법을 사용하여 데이터에서 직접 정보를 학습한다. 머신러닝 알고리즘의 성능은 학습한 데이터 수에 비례한다[15]. 머신러닝모델은 예측하고자 하는 종속 변수의 형태에 따라 회귀 모델과 분류모델로 나눠진다.

회귀 모델은 강수량의 변화, 기온의 변화 등 실수로 이루어진 연속적인 값을 예측하는 모델로 그 종류로는 선형 회귀, 비선형 회귀, 가우시안 프로세스 회귀 등이 있다.

분류 모델은 둘 이상의 불연속적인 데이터를 특정 그룹이나 클래스로 구분하여 분류하는 모델이다. 예를 들어 노면 상태가 건조 (dry), 습윤 (moist), 젖음 (wet) 이상 3 가지의 상태로 구분되면, 분류 모델은 노면 상태별 데이터가 가지는 특징을 분석하고 입력 데이터 (설명변수)가 어떤 노면 상태의 특징과 보다 더 상관도가 높은지를 분류한다.

본 연구의 목적은 기상 데이터를 이용하여 노면 상태라는 정성적 데이터를 예측하는 것으로 회귀모델보다 분류 모델을 사용하는 것이 합리적인 예측이 가능하다. 본 장에서는 대표적 분류 모델인 로지스틱 회귀 (Logistic Regression), 의사결정나무 (Decision Tree), SVM (Support Vector Machine), k-NN (k-Nearest Neighbor)과 모델 선정과 평가 기준에 관해 설명하고, 6장에서 모델 선정기준을 모델들을 평가하였다.

가. 로지스틱 회귀 (Logistic Regression)

로지스틱 회귀는 종속변수와 설명변수 간의 인과관계를 로지스틱 함수를 이용하여 분류하는 모델이다[16]. 종속변수는 정성적 데이터로 이루어져 있어야 하며 새로운 설명변수가 주어졌을 때 종속변수의 각 집단에 속할 확률이 얼마인지를 추정해 준다. 추정 확률의 기준치에 따라 종속변수를 분류하는 목적으로 사용할

수 있다. Kim et al.[17]은 강우와 도로와 같이 넓은 범위에 존재하는 데이터간의 형식적 통계 비교 및 분포 함수를 지정하는 관계식을 구하는데 로지스틱 회귀가 용이하다고 평가하였다.

나. 의사결정나무 (Decision Tree)

의사결정나무는 변수 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 특징에 따라 분류하고 예측한다. 규칙들을 조합하는 과정이 도표화되어 나타나는데 그 모양이'나무'와 같다고 해서 의사결정나무라 불린다. 분류 및 예측의 과정이 도표화되어 표현되기 때문에 연구자가 데이터 분류 및 예측 과정을 이해하고 설명하기 쉽다는 장점이 있다[18].

다. SVM (Support Vector Machine)

SVM은 입력공간과 관련된 비선형문제를 고차원적인 특정 공간으로 변화시켜 선형문제로 적용한다[19]. 두 변수 사이에 존재하는 여백(margin)을 최대화하여 선형 결정 경계선을 찾아 데이터를 분류하여 분류 성능을 극대화하는 방법이다. 여기서 여백은 분류를 위한 선형 결정 경계선과 이 경계선에 가장 가까운 데이터 사이의 거리를 말한다[20]. SVM은 선형과 비선형 분류를 모두 지원한다. 뿐만 아니라 SVM은 적은 양의 데이터로 구축이 가능하다[21]. 하지만 비선형 모델을 사용할 경우, 설명변수를 해석하기 어려워지고, 연산량이 많다는 단점이 있다.

라. k-NN (k-Nearest Neighbor)

k-NN은 비모수기반 패턴기반의 알고리즘으로 예측하고자 하는 관측치가 있을 때 이와 가장 가까운 거리에 있는 k개의 관측치를 결정한 후 이들의 특성을 이용해 관심 관측치를 예측하는 과정을 거쳐 데이터를 분류한다[15]. 알고리즘이 간단하여 구현이 쉽고, 수치기반 데이터를 분류할 때 우수한 성능을 보이는 장점을 지닌다.

마. 딥 러닝 (Deep Learning)

딥 러닝은 머신러닝과 달리 연구자가 데이터의 특징을 모델에 설정할 필요 없이 모델 스스로 분석한 후 답을 내는 방식이다. 분류 과정을 처음부터 끝까지 모델이 학습하기 때문에 감독자가 개입함으로써 생길 수 있는 오류를 최소화하여 높은 정확도를 얻을 수 있다는 장점이 있다[22]. 딥러닝은 높은 공간해상도와 시간 정확도가 요구되는 장기 강수량 예측과 같은 시간적 상관관계 (temporal correlation)가 요구되는 데이터를 예측하는데 강력한 성능을 보인다[23]. 다만, 복잡한 예측 과정으로 인해, 모델에 대한 해석이 어렵고 가중치 수가 많아져 학습 연산량 및 메모리가 증가하는 단점이 있다[24].

2.3.2 모델 선정기준 및 평가

데이터에 적합한 모델을 선정하기 위해서는 모델을 선정하기 위한 기준과 그 선정기준에 따라 모델을 평가할 수 있는 평가 기준이 필요하다. 본 연구에서는 모델을 선정하기 위해 예측력, 해석력, 효율성, 안정성 등 4가지의 선정기준에 따라 각 모델의 적합성을 평가하였다. 이를 위해, F1-Score나 특성 중요도, 교차 검증과 같은 모델 평가 방식들을 사용하여 각 예측 모델이 노면 상태 예측에 적합한지 평가하였다.

가. 선정기준

신뢰할 수 있는 데이터 예측 결과를 얻기 위해서는 모델 평가 이전에 데이터의 특성과 목적에 맞는 모델을 선정해야 한다. Niazi et al.[25]은 모델의 선정기준을 간단성(simplicity), 유효성(valid), 강건성(robustness) 등으로 정의하였다. Kim et al.[26]은 위 3가지 선정기준을 확장한 4가지 모델 선정기준을 Table 2과 같이 정의하였다.

모델의 예측력(Predictability)은 모델이 갖는 예측 정확도를 의미한다. 예측력은 훈련 데이터를 통해 학습된 모델을 테스트 데이터로 평가한다. 예측력에 대한 평가는 예측할 종속변수에 따라 달라진다. 평가 방식에 대한 자세한 설명은 다음 목에서 설명하도록 한다.

해석력(Interpretability)은 설명변수와 종속변수간의 상관관계가 유효한 정도를 평가하는 요소이다. 보통 변수간의 유효성을 평가할 때 예측력의 정도를 사용하

Table 2. Model Selection Criteria

Model evaluation criteria	Description
Predictability	Does the model predicts well?
Interpretability	Do explanatory variables describe dependent variable well?
Efficiency	Is the model constructed only with essential explanatory variables?
Robustness	Does the result of the model is stable whenever test sets are different?

는데 이는 설명변수를 모델에 입력했을 때 출력되는 종속변수가 실제 데이터와의 얼마나 일치하는지 나타내는 지표이다.

효율성(Efficiency)은 데이터 세트에 불필요한 설명변수가 존재하는지에 대한 여부를 평가하는 요소이다. 효율성에 대한 평가는 예측모델을 훈련할 때 각 설명변수에 대한 활용도를 사용한다. 모델에 따라 각 설명변수의 활용도에 대한 표현이 다르므로, 모델에 따라 다른 방식으로 효율성을 평가한다.

안정성(Robustness)은 모델 구동에 사용한 입력 데이터나 변수를 변경했을 때 모델의 예측력이 변화하지 않고 얼마나 둔감한지를 나타내는 요소이다. 만일 입력 데이터나 변수를 변경할 때마다 모델의 예측력이 크게 변화한다면 연구자는 해당 모델을 신뢰할 수 없지만 반대로 모델의 예측력이 큰 변화 없이 둔감하게 반응한다면 해당 모델을 신뢰할 수 있다.

본 연구에서는 다양한 기준으로 모델을 평가하기 위해, Kim et al.[26]에서 정의한 4가지 선정기준에 따른 예측 모델 평가를 진행하였다.

나. 모델 평가

혼동행렬(Confusion matrix)은 Table 3와 같이 분류 모델이 예측한 결과가 참 (True)인 경우와 예측한 결과가 틀린 (False) 경우를 나누어 분류한 행렬로써 분류 모델의 성능을 시각화 할 수 있다.

혼동행렬에서 오차율과 정확도는 Table 4와 같이 계산한다. 정확도는 전체 관측치 중 실제값과 예측값이 일치한 정도를 나타내고 범주의 분포가 균형을 이룰 때 효과적인 평가지표이다. 오차율은 모델이 전체 관측치와 실제값의 예측치가 다른 정도를 나타낸다.

Table 3. Confusion Matrix of Classification Model

		Predicted	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

Table 4. Evaluation Index of Accuracy and Error ratio

Criteria	Definition	Formula
Accuracy	Ratio at which the model correctly classifies sample	$\frac{TP+TN}{TP+FN+FP+TN}$
Error ratio	Ratio at which the model did not correctly classifies sample	$1-\frac{TP+TN}{TP+FN+FP+TN}$

정확도와 오차율이 모델 전체 데이터에 대한 신뢰도를 평가한다면, Table 5의 관측치들은 모델의 성능을 상세하게 평가 할 수 있는 관측치들이다. 예를 들어 하나의 종속변수 클래스의 비율이 90%인 모집단을 사용하여 구축한 모델이 90%의 정확도를 갖더라도, 나머지 10%의 클래스에 대한 예측 정확도는 90%라고 확신할 수 없다. 위 예와 같이 정확도와 오차율만으로 모델이 적합하다고 설명하기 어렵다면, Table 5의 평가 기준들을 목적에 맞게 사용하여 모델의 상세한 예측력을 평가하는 것이 중요하다.

본 연구의 목적인 노면 상태 예측 역시 정확도만으로 모델을 평가하기 어렵다. 만약 건조, 습윤, 젖음 등 3가지 노면 상태를 예측한다고 가정하면, 예측 모델은 각각의 노면 상태에 대해 높은 정확도로 예측하는 것이 필요할 것이다. 특히 젖음과 습윤 상태의 노면 예측력을 평가하기 위해서는 참 부정(True Negative)에 대한 비율이 얼마나 낮은지도 중요한 척도가 될 수 있다. 운전자가 젖은 도로를 운행하고 있을 때, 예측 모델이 노면 상태가 젖어있음을 예측하지 못한다면, 큰 사고로 이어질 수 있기 때문이다. 따라서 본 연구에서는 예측력을 평가하기 위해, 각 노면 상태에 따른 예측 정확도를 평가하였으며, 예측 정확도는 정확도, 재현율, F1-Score 값을 사용하였다.

3. 예측 모델 평가 프로세스

Figure 1은 본 연구에서 활용한 예측 모델 평가 과정을 설명한다. 예측 모델에 대한 평가를 위해서 크게 데이터 세트 생성 단계와 모델 평가 단계로 분류하여 진행하였다. 데이터의 생성 단계는 실제값을 수집하고, 데이터를 마이닝하며, 설명 변수 및 데이터 세트를 생성하는 과정을 포함한다. 모델 평가 단계에서는 훈련 데이터를 활용하여 대상 모델을 학습하고, 정확도 및 각 설명변수의 중요도 등을 비교하여 노면 상태 예측에 가장 적합한 모델을 평가한다.

3.1 데이터의 생성 및 수집

데이터 세트 생성 단계에서는 예측 대상과 관련한 실제값들을 수집하고, 그 실제값들이 설명변수로 가공하

Table 5. Various Measured Value of Error ratio

Criteria	Definition	Formula	
Precision	ratio of samples that the model positively classifies as positive	$\frac{TP}{TP+FP}$	
Recall	ratio of positive affirmative samples classified as positive	$\frac{TP}{TP+FN}$	
Fall-out	ratio of false negative samples that the model incorrectly classified as positive	$\frac{FP}{TN+FP}$	
F1 score	Harmonic mean of Precision and Recall	$2\times\frac{Precision\times Recall}{(Precision\times Recall)}$	
ROC* Curve	Graphical plot which illustratios the performance of a binary classifier system as its discrimination threshold is varied..	x axis	$\frac{TN}{TN+FP}=Fall-out$
		y axis	$\frac{TP}{TP+FN}=Recall$

ROC* : Receiver Operating Characteristic

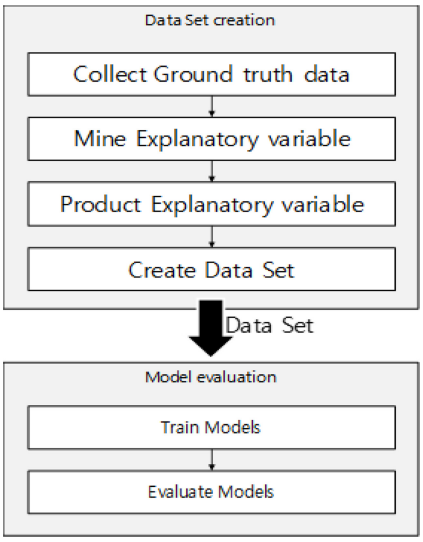

Figure 1. Model evaluation process

였다. 데이터 세트 생성 단계에서는 가장 우선적으로 실제값 수집이 이루어져야 한다. 실제값을 수집하는 것은 분석의 시작이자, 전체 과정 중 가장 중요한 단계이다. 아무리 좋은 설명변수와 파라미터 설정이 된 모델일지라도, 실제값에 대한 신뢰도가 낮다면, 예측 모델에 대한 신뢰도 역시 낮아진다. 따라서 신뢰도가 높은 데이터를 많이 수집하기 위해서는 처리 과정을 통해 데이터가 필터링(Filtering)되는 것을 고려하여 최대한 많은 데이터를 수집하는 것이 좋다. 본 연구에서 진행한 노면 예측을 위한 노면 정보 수집의 경우, Kim et al.[6]이 제안한 빅데이터 시스템과 RCM411[27] 센서를 사용하여 노면 정보를 수집하였다.

데이터가 수집되면, 여러 실제값들에 대한 데이터 마이닝을 통해 데이터 세트를 구성하는 설명변수를 찾는다. 일반적으로 데이터 세트가 실제 분석에서는 미리 존재하는 경우가 드물어서, 데이터 마이닝 과정을 통해, 예측 대상을 설명할 수 있는 설명변수를 찾는 과정이 필요하다. 데이터 마이닝을 통해 찾은 이 설명변수들은 반드시 예측 대상을 설명할 수 있어야 한다. 설명변수들이 만약 예측 대상에 대한 설명력이 부족하다면, 모델의 예측 정확도를 감소시키거나, 과대 적합을 야기할 수 있기 때문이다. 만약 가공한 데이터 세트에 위와 같은 문제가 있다면, 마이닝 과정을 다시 수행하는 것이 좋다. 데이터 마이닝이 끝나면, 각 실제값들로부터 설명변수를 가공하여, 종속변수와 정합한 뒤, 데이터 세트를 생성한다.

3.2 데이터 세트 가공

수집한 데이터를 모델의 학습 및 평가에 사용하기 위해서는, 종속변수와 종속변수를 설명할 수 있는 설명변수를 포함하는 데이터 세트가 필요하다. 종속변수의 경우 실제 현상을 예측하는 경우가 많아 이상치에 대한 처리 이외의 가공을 하지 않는 경우가 많다. 이는 종속변수가 실제값 그 자체이거나 정확한 실제값을 수집하는 과정에서 이미 처리 과정을 수행하는 경우가 많기

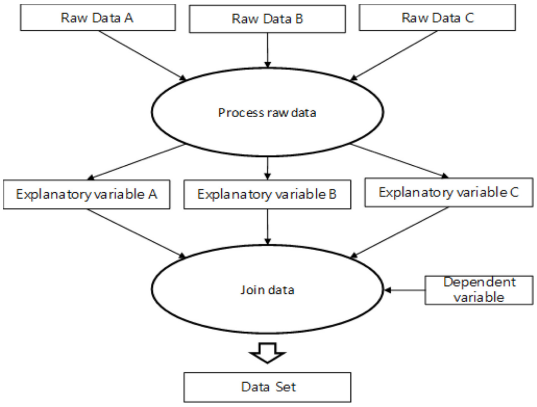

Figure 2. Processing Data Set from multiple raw data

때문이다. 설명변수는 Figure 2와 같이, 하나 또는 여러 개의 실제값을 가공하여 생산한다. 설명변수들이 가공되면, 각 설명변수들을 하나의 데이터 세트로 결합한다.

3.3 설명변수의 처리

3.3.1 노면 상태

도로에는 주변 건물, 가로수, 도로의 포장 종류, 도로의 경사도 등 노면 상태에 영향을 주는 다양한 주변 환경들이 존재한다. 이 도로 주변 환경들은 주로 노면 상태의 변화에 직접적인 영향을 미치는 것이 아니라, 노면 상태의 변화 속도에 영향을 준다. 본 연구에서는 도로 주변 환경 중 종단 경사도를 활용한다.

종단 경사도는 도로 기하구조 정보에 속한다. 기하구조 데이터는 이동형 관측 기기를 통해 수집한 종단 경사도와 좌표, 고도 등의 기본 정보와 추가 가공과정을 통해 생성된 곡선 구간 여부, 곡선 반경, 곡선 부 길이, 포장 종류 등의 정보를 포함한다. 기하구조 데이터는 시작지점을 기준으로 정렬되어 있으며, 연속된 두 행 사이의 공간거리는 10 m이다. 각 행의 좌표 정보는 점의 형태로 이상치에 대한 보정작업이 완료된 상태이다. 대상 도로는 서울시 도시화 고속도로이며, 본 연구에서는 그 중 내부순환로에 대한 기하구조 정보를 활용한다.

종단 경사도는 현재 도로가 내리막길인지 오르막길인지를 나타내는 변수이다. 종단 경사도가 음수일 때 해당 도로는 내리막의 형태를 띠고 있으며, 양수일 때 오르막길의 형태를 띤다. 도로의 내리막길과 오르막길은 관측 지점에서의 수막 두께에 영향을 미친다. 예를 들어 내리막길에서 오르막길로 변하는 극소점 부분은

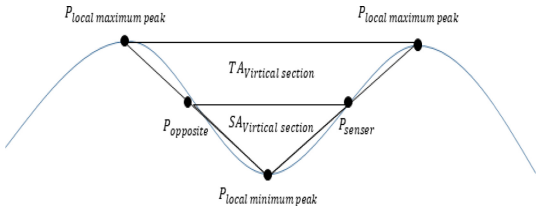

Figure 3. The vertical surface of road

물이 고여 수막 두께가 두꺼울 수 있다. 반대로 극대점 부분은 상대적으로 수막 두께가 얇다. 이 수막 두께는 노면에서의 건조 속도비에 영향을 미친다. 수막 두께가 증가할수록 노면이 마르는 데 필요한 시간이 증가하며, 얇을수록 감소하기 때문이다. 본 연구에서 건조 속도비는 아래 수식을 통해 계산하였다.

$$Dry\ speed\ ratio = \frac{SA_{Virtical\ section}}{TA_{Virtical\ section}}$$

건조 속도비를 구하기 위해서는, Figure 3와 같이 종단 경사도의 극점과 관측 지점을 활용하여 총 종단면 넓이($TA_{Virtical\ section}$)와 부분 종단면 넓이($SA_{Virtical\ section}$)를 계산한다. 총 종단면 넓이는 두 극대점과 하나의 극소점이 이루는 삼각형의 넓이를 의미한다. 부분 종단면 넓이는 관측 지점과 극소점을 꼭짓점으로 포함하는 삼각형의 넓이를 의미한다.

이 건조 속도비는 노면 상태 변화에 대한 설명력을 부여하는 것이 목적이므로, 정확한 건조 속도비를 계산하지 않는다. 다만, 길이가 아닌 넓이의 비를 사용함으로써, 해당 경사 구간의 물 수용력에 대한 설명력을 부여하였다. 종단면 넓이 대신 점과 점 사이의 고도 차를 사용할 수 있지만, 같은 높이라도 종단 경사도가 달라지면, 수막 두께 역시 다를 수 있기 때문이다.

또한, 해당 구간의 건조 속도비에 대한 정확한 설명력을 부여하고자 한다면, 횡단 경사도에 대해 고려가 필요할 것이다. 강수가 발생하여 도로에 물이 축적되는 것은 종단 경사도에 의해 영향을 받는 것이 사실이지만, 횡단 경사에 의해 물이 갓길로 배출되기 때문에 실제 축적되는 양은 예상보다 적을 것이기 때문이다. 하지만 횡단 경사에 의해 갓길로 물이 배출되더라도, 노면 위에 수분은 남아 있을 수 있다. 또한, 노면 위에 남은 수분이 구간의 종단면 넓이에 의해 달라질 수 있으므로, 횡단 경사도를 고려하지 않아도 건조 속도비가 충분한 설명력을 가질 수 있다. 따라서 본 연구에서 건조 속도

Table 6. Weather explanatory variable

Explanatory Variable	Description
acc_precipitation	Accumulated precipitation from precipitation end time to current sensing time
time_dist_min_pet_st	Time distance represented as minutes between the end time of precipitation and sensing time
avg_temperature_pet_st	Average temperature from precipitation end time to current sensing time
sd_temperature_pet_st	Standard deviation of temperature from precipitation end time to sensing time
avg_temperature_pst_pet	Average temperature from precipitation start time to precipitation end time
sd_temperature_pst_pet	Standard deviation of temperature from precipitation start time to precipitation end time
avg_temperature_pst_st	Average temperature from precipitation start time to sensing time
sd_temperature_pst_st	Standard deviation of temperature from precipitation start time to sensing time

Table 7. Classification report of machine learning models

Balanced Data Set(data size = 60000)															
	k-NN			Logistic Regression			Decision Tree			SVM			Deep learning		
	dry	moist	wet	dry	moist	wet	dry	moist	wet	dry	moist	wet	dry	moist	wet
precision	94%	90%	97%	58%	63%	66%	94%	91%	95%	59%	67%	80%	68%	75%	83%
recall	95%	92%	94%	77%	62%	62%	95%	91%	95%	84%	64%	49%	80%	69%	76%
f1-score	94%	91%	96%	66%	44%	53%	94%	91%	95%	69%	65%	61%	74%	72%	80%
Imbalanced Data Set(data size = 60000, dry=24%, moist=18%, wet=58%)															
precision	93%	87%	97%	52%	60%	72%	93%	85%	97%	62%	73%	71%	63%	72%	86%
recall	95%	85%	97%	64%	25%	79%	94%	86%	96%	54%	28%	89%	79%	49%	86%
f1-score	94%	86%	97%	57%	35%	75%	94%	85%	97%	58%	40%	79%	70%	59%	86%

비를 가공할 때, 종단 경사도만을 사용하여, 물 수용력에 대한 설명력을 부여하였다.

3.3.2 기상 조건

기상정보는 노면 상태 변화에 가장 직접적인 영향을 미치는 변수이다. 눈과 비는 도로 노면의 변화를 유발하며, 온도는 노면 상태가 변하는데, 직접적인 영향을 미친다. 하지만 노면 상태를 예측하기 위해 기상정보를 가공하지 않고 설명변수로 사용한다면, 정확한 예측이 어려울 수 있다. 기상 변수가 노면 상태에 미치는 영향을 설명할 때, 시간에 대해 고려가 필요하기 때문이다. 예를 들어, 도로의 노면은 현재 발생하고 있는 강수의 영향도 받지만, 1시간 전, 눈과 같은 기상 변수는 2~3일 전에 내린 경우에도 현재 노면 상태에 영향을 미친다. 따라서 기상정보를 활용하여 설명변수로 가공할 때에는 반드시 시간에 대해 고려가 필요하다. 본 연구에서는 ASOS의 분당 강수량과 기온, 시간당 적설량 정보를 활용하여 Table 6과 같이 8개의 설명변수를 가공하였다.

누적 강수량(acc_precipitation)은 강수 발생 시점부터 강수 종료 시점까지 누적된 강수량을 측정한 값이다. 누적 강수량 값은 현재 노면의 상태에 직접적 또는 간접적으로 영향을 미치는 설명변수다. 강수량은 비와 눈에 의해 발생한 강수의 양을 의미하기 때문에, 누적 강수량을 가공하기 위해서 ASOS의 분당 강수량과 시간당 적설량을 사용한다. ASOS의 분당 강수량은 00시 00분부터 현재까지 발생한 누적 강수량 정보를 제공한다. 하지만 강우 발생 시점이 00시 00분 이전일 경우 직전 발생한 강수의 총량으로 보기 어려우므로, 이에 대한 추가 작업이 필요하다. 적설량은 분당 강수량에서 정확하게 측정하지 못한 눈에 대한 정보를 보충 설명하기 위해 사용한다.

일반적으로 적설량과 강수량은 10:1의 비율로 표현한다. 즉, 1 cm의 적설량이 관측되었다면, 강수량으로는 0.1 mm라고 기록된다. 하지만 노면 상태 변화 측면에서 바라볼 경우, 눈은 녹으면서 증발하기 때문에 비보다 노면 상태 변화 속도가 상대적으로 느리다. 또한, 눈이 겨울철에 발생한다는 점을 고려하면, 추운 날씨에 의해 눈이 얼고 녹음을 반복하기 때문에 노면이 건조 상태가 되는 데 필요한 시간이 증가한다. 이러한 이유로 적설량의 강수량 표현은 단순히 단위 변환하는 것으로 대체하였다.

강수 종료 시점부터 관측 시점 사이의 시간적 거리는 이전에 발생한 강수가 현재 노면 상태에 미치는 영향을

설명하는 변수이다. 강수는 관측 시점 당시에 발생하고 있을 수 있고, 1시간 또는 하루 전에 종료될 수 있다. 여름철 약한 비가 내리는 중에 노면 상태를 예측한다면 노면은 젖어있다. 하지만 아무리 강한 비가 내리더라도 3시간 전에 내린 비는 도로 관리가 잘 되어 있다면 노면이 말라 있을 수 있다. 노면은 강수가 발생하는 도중에는 항상 젖었거나 얼어있다. 따라서 도로 노면이 실제로 변하는 시간은 강수 종료 시점이므로, 두 시점 사이의 시간적 거리는 강수 시작 시점이 아닌 강수 종료 시점과 관측 시점 사이의 시간으로 정의한다. 또한, 강우 종료 시점이 관측 시점보다 늦을 수 있으므로, 음수의 값을 포함한다.

온도는 강수 발생 시점부터 관측 시점까지 시간을 3개의 구간으로 정의하고, 각 구간에서의 온도 정보를 표현하기 위해 평균과 표준 편차를 각각 계산한다. 평균 온도는 노면 위의 수분이 증발하는 속도에 직접적인 영향을 미친다. 평균 온도가 높다면 노면이 마르는 속도는 증가하고, 낮으면 감소한다. 또한, 노면이 평균 온도가 낮다면 노면이 어는 경우도 발생한다. 하지만 노면이 어는 경우는 항상 평균 온도가 낮을 때만이 아니다. 겨울철 노면은 2~3일 전 발생한 강수의 영향을 받을 수 있으므로, 3일 전에 내린 비가 2일 전 발생한 강추위로 인해 노면이 어는 경우가 발생할 수 있기 때문이다. 겨울철 노면의 상태는 시각의 영향을 받는다. 새벽과 저녁은 온도가 영하를 기록하지만, 아침과 낮에는 상온을 유지하는 예도 있다. 온도의 표준 편차는 시간에 따라 달라지는 온도를 설명하는 변수로 활용한다. 표준 편차는 유사한 평균 온도일 때의 노면 상태 예측하는 부분에서 효율적이다.

3.4 모델 평가

본 장에서는 4가지 모델 평가 기준에 따라 각 머신러닝 모델들을 검증한다. 총 5가지 노면 상태를 검출할 수 있지만, 결빙(Ice) 및 살얼음(Slush)에 대한 실제값(Ground truth)은 그 수가 충분하지 않아 본 연구에서는 건조(Dry), 습윤(Moist), 젖음(Wet)의 노면 상태를 갖는 실제값만을 사용하였다. 모델 예측 및 평가는 scikit-learn 라이브러리를 활용하였으며, 모든 파라미터 설정은 기본값을 사용하였다

3.4.1 예측력(Predictability)

예측력은 모델 선정에 있어서, 가장 중요한 선정기준이 된다. SVM과 같이 데이터 크기 및 데이터의 스케일, 불균형 데이터 세트에 민감한 모델을 검증하기 위해서 노면 상태별로 20,000개의 데이터를 사용하였으며, 데이터는 시간과 관계없이 무작위로 선택하였다. 또한, 일부 모델에서는 데이터 세트를 표준화(Normalization)하여 사용하였다.

각 모델에 대한 예측 정확도는 Table 7과 같다. 정확한 모델의 예측력을 평가하기 위해서 정확도, 재현율, f1-score를 노면 상태별로 적용하였다. 정확도 계산 결과 전반적으로 k-NN과 의사결정 트리의 정확도가 높았다. 반면 로지스틱 회귀와 SVM, 딥 러닝의 정확도가 낮았다. 이는 해당 설면 변수에 가장 적합한 모델은 수식기반 모델보다 로직이나 통계 기반 모델이 더 적합하다는 것을 의미한다.

3.4.2 해석력(Interpretability) & 효율성(Efficiency)

해석력과 효율성은 모두 설명변수와 종속변수 사이의 관계를 평가하기 위한 요소이다. 또한, 두 선정기준 중 하나의 기준을 충족한다면, 다른 기준도 충족할 확률이 높다. 이러한 두 선정기준의 유사성에 따라, 본 연구에서는 두 선정기준에 대한 평가를 각 모델에서 제공하는 설명변수의 활용도를 사용하여 진행한다.

의사결정 트리는 특성 중요도(Feature Importance)를 사용하여 평가하였다. 특성 중요도는 설명변수들이 의사결정 트리의 각 노드에서 분기 조건으로 활용된 비율을 의미한다. 따라서 중요도가 가장 높은 설명변수는 의사결정 트리를 구축할 때 가장 많은 영향을 미친다고 할 수 있다. 반대로 특성 중요도가 0에 가깝다면, 해당 설명변수는 모델을 구축을 위해 영향을 미치지 않으니,

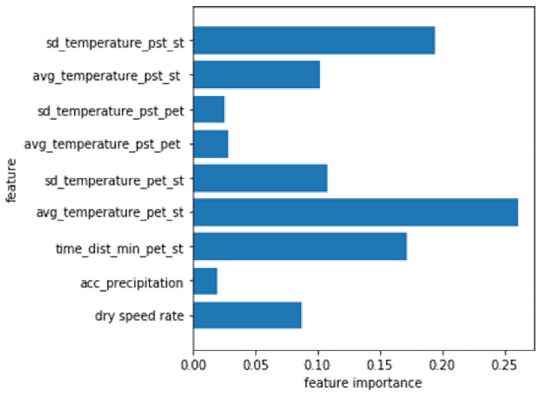

Figure 4. Feature importance of decision trees

데이터 세트에서 제거해도 무방하다. 노면 상태 예측을 위해 구축한 의사결정 트리의 특성 중요도는 Figure 4과 같다.

의사결정 트리의 경우, 모든 설명변수를 고르게 사용하였다. 그 중, 의사결정 트리를 구축하기 위해 가장 많이 활용한 설명변수에는 avg_temperature_pet_st (Average temperature from precipitation start time to sensing time)와 시간적 거리가 있다. 이 설명변수들이 높은 이유는 수집한 실제값의 특징 때문이다. [6]에서 설명한 바와 같이 노면 상태 데이터를 수집하기 위해서는 차량에 노면 센서를 부착하고 해당 도로를 이동하면서 노면 상태를 검출해야 한다. 따라서 비나 눈이 발생하는 날 노면의 상태를 수집하기 위해서는 대상 도로까지 이동해야 하는데 대부분 눈과 비가 이동 시간 중에 그치는 경우가 많다. 그 결과, 노면의 상태는 비나 눈이 그친 뒤 수집되는 경우가 많아졌고, 모델의 예측에서도 강수 발생 시점과 관측 시점 사이의 시간적 거리와 평균 온도가 가장 많은 영향을 끼친다. 건조 속도비의 경우 높은 비율은 아니지만, 중요도가 10%에 가깝다는 것은 의사결정 트리에서 노면의 상태를 예측하는데 건조 속도비가 영향을 미치고 있다는 것을 의미한다.

로지스틱 회귀는 모델 학습을 하면서, 각 특성에 대한 가중치를 부여하기 위해 계수(Coefficient)를 사용한다. 계수는 음수 또는 양수 값을 갖고, 양수일수록 해당 설명변수가 모델 예측에 대한 양의 상관도가 크다는 것을 의미한다. 노면 상태 예측을 위한 각 설명변수의 계수는 Figure 5과 같다.

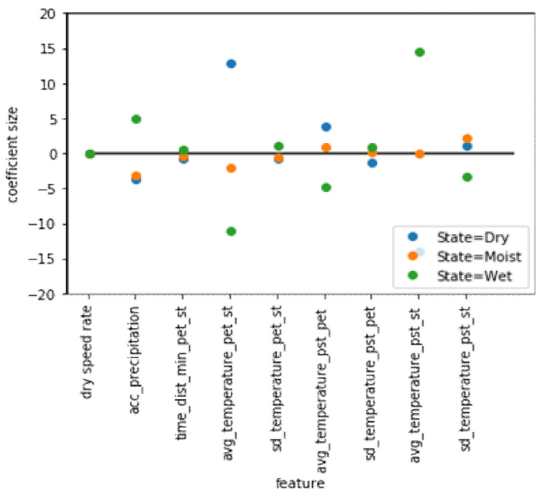

Figure 5. Coefficient of Logistic Regression

로지스틱 회귀는 이진 분류를 기본으로 한다. 만약 로지스틱 회귀를 사용하여 여러 클래스를 분류한다면, 각 클래스에 대한 모델이 학습되어야 가능하다. 따라서 이 같은 경우에는 클래스별 설명변수에 대한 계수가 존재한다. 로지스틱 회귀에서는 의사결정 트리와 달리 건조 속도비와 시간적 거리에 대한 계수가 0에 가까우므로, 노면 상태 예측에 영향을 거의 미치지 않았다고 할 수 있다.

반면에, avg_temperature_pet_st는 의사결정 트리와 마찬가지로 노면 예측에 큰 영향을 미칠 뿐만 아니라, 해당 설명변수가 노면의 상태에 어떤 영향을 미치는지 명확하게 설명하고 있다. 먼저 건조 상태는 평균 온도와 양의 상관관계를 보이는데, 이는 평균 온도가 증가함에 따라 노면 상태가 건조할 확률이 높다는 것을 의미한다. 반대로 젖음 상태는 평균 온도와 음의 상관관계를 갖는다.

k-NN은 클래스가 존재하는 데이터와 존재하지 않는 데이터 사이의 통계적 거리를 계산하므로, 모델에서 설명변수에 대한 영향도를 확인할 방법이 없다. 따라서 K-NN의 설명변수에 대한 영향도 검증 방법으로는 데이터 세트를 사용하여 설명변수와 종속변수가 각각 하나인 데이터 세트를 만들고, 모델을 학습하여 정확도를 평가하는 것으로 대체하였다. 이 방법은 다른 설명변수의 개입이 없는 모델을 통해 예측 정확도를 평가하기 때문에 순수하게 설명변수가 노면 상태에 미치는 영향을 확인할 수 있다. Table 8은 각 설명변수에 따른 예측 정확도를 계산한 결과이다.

SVM은 로지스틱 회귀와 같이 계수를 제공한다. 다만, 종속변수에 대한 설명변수의 영향도는 선형 SVC 모델에서만 제공한다. 선형 SVC 모델은 노면 상태에 따른 설명변수의 계수를 제공한다. Figure 6는 Linear SVC로 데이터를 학습하였을 때 가장 높은 양, 음의 상

Table 8. Precision of k-NN per explanatory variable

Explanatory variable	Precision(%)
dry speed rate	34%
acc_precipitation	56%
time_dist_min_pet_st	75%
avg_temperature_pet_st	91%
sd_temperature_pet_st	91%
avg_temperature_pst_pet	69%
sd_temperature_pst_pet	68%
avg_temperature_pst_st	92%
sd_temperature_pst_st	92%

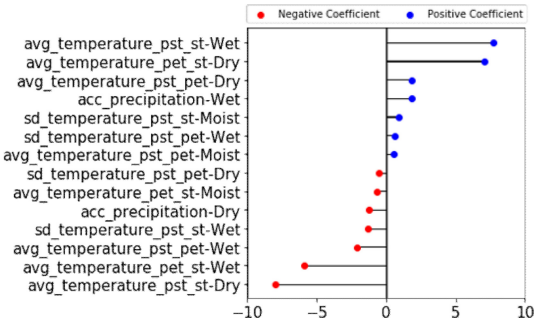

Figure 6. Negative and positive importance of each explanatory variables for linear SVC.

관관계를 갖는 7개 설명변수의 계수를 그래프 형태로 나타내고 있다.

딥러닝에서의 계수는 각 설명변수에 대한 은닉 계층(Hidden Layer)의 각 퍼셉트론(Perceptron)에서의 계수를 나타내는 값이기 때문에 해석이 어렵다. 하지만 딥러닝이나 SVM이 설명변수들의 특징을 더 잘 설명하는 경우도 있다. 예를 들어, 만약 설명변수가 이미지의 각 픽셀의 색상 정보라면, 학습된 딥러닝 모델의 계수 정보를 사용하여, 딥 러닝이 이미지를 분류할 때 중요하게 생각한 패턴이나, 특징을 이미지 형태로 확인할 수 있기 때문이다.

3.4.3 안정성(Robustness)

안정성이 높은 모델은 여러 데이터 세트로 모델을 평가하였을 때, 예측 정확도가 일정하다. 본 연구에서는 교차 검증(Cross-validation)을 통해 서로 다른 데이터로 모델을 학습했을 때의 예측 정확도를 검증하였다(Table 9). 검증 결과, 전반적으로 각 모델들의 정확도에 대한 표준 편차는 큰 차이를 보이지 않았다. 교차 검증에 대한 평균 정확도는 k-NN이 94%로 가장 높았다. 각 모델의 정확도에 대한 표준 편차는 크게 차이는 없었지만, 그중에서도 K-NN이 0.25%로 학습 데이터에 따른 정확도 변동 폭이 가장 좁았다.

Table 9. Cross-Validation by Model

Model Name	Average accuracy	Standard deviation
Decision tree	93.53%	0.31%
k-NN	94.20%	0.25%
Deep Learning	75.92%	0.85%
Logistic Regression	61.26%	0.56%
SVC	66.52%	0.73%

3.5 예측 모델 선택에 대한 시사점

3, 4장에서는 4가지 선정기준에 따른 최적의 예측 모델 선정을 위한 평가를 진행하였다. 모델 평가에서 사용한 이 선정기준들은 준비된 데이터 세트에 예측 모델이 얼마나 적합한지를 평가하기 위한 척도로 사용하였다. 하지만 적합한 모델을 평가하기 위해서는 앞서 정의한 선정기준들 이외에도 데이터의 수집 환경이나, 데이터 수집 난이도 등을 반드시 고려해야 한다. 이동형 노면 센서를 사용하여 실제값을 수집하기 위해서는 큰 노력이 필요하다. 비나 눈이 발생하면, 종료 전에 수집 대상 도로에 도착하여 노면의 상태를 수집해야 한다. 여름에는 노면의 상태가 빨리 변하므로 강우 종료 후에 노면 상태를 검출하면, 건조 상태의 데이터가 많다. 겨울의 경우, 대상 도로의 관리가 잘되어, 눈이 내려서 노면 상태를 수집한다고 하더라도 결빙된 지역을 찾아보기 어렵다.

노면 상태 예측에 필요한 데이터의 양도 중요한 모델 선정기준이 될 수 있다. 실제로 약 300만 개의 실제 수집한 데이터를 가공 및 처리하면, 약 33만 개의 분석 가능한 데이터가 생산된다. 약 300만 개의 데이터를 수집하기 위해 월 10회 이상 1년 동안 도로의 노면 상태를 수집하였다. 그중에서 습윤과 결빙, 살얼음 등의 노면 상태 정보는 건조와 젖음 상태의 노면 상태 정보 수집보다 더 큰 노력이 필요하다. 이렇게 노면 상태의 비율이 불균형하다는 점을 고려하면, 각 노면 상태별로 충분한 양의 자료를 수집하는 것은 많은 시간적, 비용적 노력이 필요하다.

위와 같이 실제값을 수집하기 어려운 데이터의 경우, 수집한 데이터를 효율적으로 사용하는 모델을 선택하는 것이 바람직하다. 이러한 데이터 활용적인 측면을 검증하기 위해, 본 연구에서는 불균형 데이터에 대한 예측 정확도와 데이터 세트 크기에 따른 모델 성능 평가를 진행하였다.

실제 노면 상태를 수집하게 되면, 건조 또는, 젖음 상태의 데이터의 비율이 다른 상태들의 데이터 수보다 월등히 많을 때도 있다. 이러한 불균형 데이터는 SVM과 같은 예민한 모델에서 정상적으로 소수의 상태를 예측하기 어렵다. 반면에 k-NN과 의사결정 트리처럼 데이터 불균형에 둔감한 모델은 데이터가 축적됨에 따라 충분히 소수의 상태도 검출할 수 있다. 불균형 데이터의

검증은 모든 노면 상태가 포함된 임의의 60000개의 데이터에 대한 예측 정확도를 평가하였고, 결과는 Table 7과 같다. 정확도에서는 모든 데이터가 각 노면 상태에 대해서 유사한 정확도로 노면 상태를 예측하였다. 하지만 SVM과 딥 러닝은 습윤 상태에 대한 재현율이 다른 노면 상태보다 낮았다. 상대적으로 의사결정 트리와 k-NN은 모든 노면 상태에 대해서 유사한 정확도와 재현율을 보이는 것을 확인할 수 있다.

데이터 세트 크기에 따른 성능을 평가하는 것은 실제값을 수집하기 위한 비용을 절감하기 위한 중요한 척도가 된다. 실제값을 수집하는 것은 시간적, 금전적으로 비용이 발생하기 때문에, 적은 데이터 세트를 사용하여, 높은 예측 정확도를 보이는 모델을 선정하는 것이 타당하다. 데이터 세트 크기에 따른 성능 평가는 가장 예측 정확도가 높은 k-NN과 의사결정 트리를 대상으로 평가하였다. 평가에 사용한 데이터는 총 10000개로 10개부터 10개씩 증가시켜 예측 정확도를 계산하였으며, 그 결과는 Figure 7과 같다.

최소 데이터 수 비교를 위한 두 모델의 목표 예측 정확도는 연구의 정량적 목표인 80%로 결정하였다. 그 결과, 의사결정 트리는 1400개의 데이터 세트만으로도 80%가 넘는 예측 정확도를 보였다. 반면 k-NN은 2500개가 넘어가면서 예측 정확도가 80%를 넘는 것을 확인할 수 있었다. 두 모델을 비교하였을 때에는 의사결정 트리가 k-NN보다 80%의 정확도를 달성하는 데 필요한 데이터의 수가 더 적은 것을 확인할 수 있다.

의사결정 트리가 K-NN보다 낮은 데이터 수에서 효율이 높은 이유는 분기를 사용하기 때문이다. 분기는 하나의 값이 아닌 조건으로 정의한다. 노면 상태 예측을 예를 들면, 대부분의 기상 상황에서 노면은 건조 상태이다. 하지만 비나 눈이 내려, 노면이 젖거나 습윤 상태일 때에는 하나의 값으로 정의하는 것보단 특정 조건을 만족하는 경우가 많다.

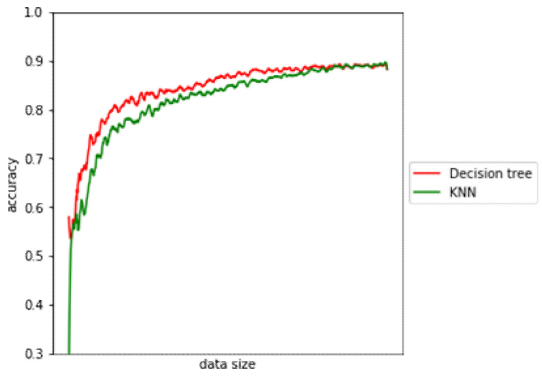

Figure 7. Precision based on data size

또한, 의사결정 트리에서 트리를 구성할 때 상위노드에서 분기에 사용한 설명변수를 하위노드의 분기에 사용할 수 있다. 이는 노면 상태를 예측할 때, 노면이 젖거나 습윤 상태일 때를 정의하는 조건을 범위 형태로 표현하는 것을 가능하게 한다. 이는 기온, 강수량과 같은 기상정보를 분류할 때 특정 범례를 지정하는 것과 유사한 역할을 하게 된다.

따라서, 의사결정 트리 기반 모델이 K-NN 기반 모델보다 노면 상태 예측에 더욱 적합하고, 더불어 적은 양의 데이터로 높은 정확도를 보이게 된다.

4. 결 론

본 연구에서는 노면 마찰력의 유추가 가능한 노면의 상태를 예측하는 연구를 진행하였다. 노면 상태 예측을 위한 머신러닝 모델은 평가 기준에 근거하여 검증하였다. 본 연구의 결과에 의하면, 노면 상태 예측은 의사결정 트리와 같이 논리 기반 모델이 가장 높은 정확도를 보였다. 그뿐만 아니라, 의사결정 트리는 모든 설명변수를 활용하여 노면 상태를 예측하였으며, 80%의 예측 정확도로 노면 상태를 예측하는 데 필요한 데이터 수도 가장 적었다. 하지만 데이터의 불균형과 수집된 데이터의 가공과정에서의 이상치 제거로 인해, 데이터 세트 수보다 더 많은 데이터가 필요하므로, 최소 데이터 세트 수에 약 10배 많은 데이터가 필요하다. 그뿐만 아니라, 기상 조건은 계절별, 월별로 달라서, 월 단위로 충분한 양의 데이터 수집이 필요하다.

본 연구가 갖는 한계점은 실제값을 수집한 대상 지역이 한정되어 있다는 것이다. 비록 60000개의 균형 데이터 세트를 활용하여 90%가 넘는 정확도를 보였지만, 한정된 지역에서 수집한 결과이기 때문에 더 넓은 지역을 포함한다면 높은 정확도를 장담할 수 없다. 간단한 예로 강원도 산간 지역의 도로의 노면 상태 예측은 도시 지역의 노면 상태 예측과 다른 기상 및 주변 환경 조건을 갖는다. 특히

대상 도로로 선정한 내부순환로와 같은 도시화 고속도로는 가까운 주변 도로와 비교해도 그 환경이 다르다. 또한, 대상 도로가 도시화 고속도로로 한정되어 있으므로, 유사한 기하구조를 갖는다. 종단 경사도만으로 표현한 건조 속도비가 예측 모델에서 충분한 영향력을 미

치고 있다는 것 역시, 대상 도로가 매우 제한적이라는 것을 의미한다.

수집 지역의 확장은 결빙 및 살얼음 상태 정보 수집과도 관계가 있다. 앞서 언급했듯이, 도시화 고속도로는 사고의 위험도를 낮추기 위해, 강수 발생 시 노면에 대한 관리가 철저하다. 그뿐만 아니라, 차량의 통행량이 여타 도로보다 월등히 많으므로, 노면 온도가 다른 도로보다 높다. 따라서 눈이 내리더라도, 실제 노면은 눈이 쌓이지 않고 금세 녹아 버리는 경우가 대부분이므로, 결빙과 살얼음 상태의 검출이 어렵다. 따라서 상대적으로 도로가 잘 결빙되는 도로에서 정보를 수집하는 것이 필요하다.

ACKNOWLEDGMENT

이 연구는 기상청 「기상 · 지진See-At기술개발연구」(KMIPA-2017-5020)의 지원으로 수행되었습니다.

REFERENCES

[1] The Road Traffic Authority, http://taas.koroad.or.kr/sta/acs/exs/typical.do?menuId=WEB_KMP_OVT_UAS_ASA

[2] Geon Hun Sin, “Bridge Road Surface Frost Prediction and Monitoring System.”, International Journal of Contents, Vol. 11. No. 11, 2011

[3] Richard Steed, Clifford Mass, “Roadwat Icing and Weather”, https://atmos.washington.edu/cliff/Roadway3.html

[4] Korea Forest Service, “National urban forest statistics”, 2016

[5] Korea Meteorological Administration, http://www.kma.go.kr/lifenindustry/download/disasters/load.pdf

[6] Seung hyun Kim, Kang Hwa Kim, Nim Woo Lee, Yeon Ju Jo, “Improving Vehicle Safety Through Road Surface Condition and Weather Data Analysis.”, Journal of Information Technology and Architecture, Vol. 14. No. 4. pp. 375-386, 2017

[7] Joon Hyung Kim, “Development of a Road Hazard Map Considering Meteorological Factors.”, Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, Vol. 35. No. 3. pp. 133-144, 2017

[8] Moo Hun Lee, Min Gyu Kim, “Development of Real-time Road Surface Condition Determination Algorithm sing an Automatic Weather System”, IT Convergence and Security, 2016

[9] Jong Woo Kim, Young Woo Jung, Jin Won Nam, “Study on the Development of Road Icing Forecast and Snow Detection System Using State Evaluation Algorithm of Multi Sensoring Method”, Journal of the Korea Institute for Structural Maintenance and Inspection, Vol. 17. No. 5. pp. 113-121, 2013

[10] Veronica J. Berrocal, Adrian E. Raftery, Tilmann Gneiting, and Richard C. Steed, “Probabilistic Weather Forecasting for Winter Road Maintenance”, Technical Report Department of Statistics university of Washington, 2007

[11] Louis P. Crevier, Yves Delage, “METRo: A New Model for Road-Condition Forecasting in Canada”, Journal of Applied Meteorology, Vol. 40, 2001

[12] Jong Hak Lee, Jeonghoon Roh, Seok Ju Park, “Development of Hydroplaning Estimation on an uninterrupted Road”, International Journal of Highway Engineering, Vol. 19. No. 6. pp. 147-153, 2017

[13] Won Seok Park, “An Improvement of the Heavy Rain Warning Area using AWS Precipitation Data : the case of capital region cities”, 2017

[14] Korea Institute of Civil Engineering and Building Technology, “Development of Technologies for Advancement of HazardousRoadway Information Service under Inclement Weather”, 2011

[15] Yong Cheon Sin, “A Research on Contract Price Prediction Model for Real Estate Auction Based on Machine Learning Technology”, 2017

[16] Cheol Woo Jung, “A Study on the Factors Influencing Drowsiness by Binary Logistic Regression Analysis”, 2007

[17] Jiwon Kim, Hani S. Mahmassani, Jing Dong, “Likelihood and Duration of Flow Breakdown Modeling the Effect of Weather”, Journal of the Transportation Research Board, 2010

[18] Ji Hyun Lee, “Optimal Condition of Pruning for VariousTypes of Dataset in Decision Tree Classification”, 2018

[19] Jae Yeon Park, Ryu Jae Pil, Shin Hyun Joon, “Predicting KOSPI Stock Index using Machine Learning Algorithms with Technical Indicators”, Journal of Information Technology and Architecture, Vol. 13. No. 2. pp. 331-340, 2016

[20] Hyo mi Kim, “Multiclass Classification of Microarray Gene Expression Data using SVM”, 2002

[21] Scikit-Learn, http://scikit-learn.org/stable/modules/svm.html#svc
[22] Won Ki Park, “Introduction to Deep Learning and Financial Application”, 2017
[23] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong Wang-chun Woo, “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting”, Neural Information Processing Systems, 2015
[24] Woo Jeong Joo, Sang Gil Kang, “Forward-Propagating Weight Quantization Method based on Deep Learning”, Journal of Information Technology and Architecture, Vol. 15. No. 2. pp. 245-252, 2018
[25] Muaz A. Niazi, “Introduction to the Modeling and Analysis of Complex Systems: a Review”, Complex Adaptive Systems Modeling, 2016
[26] Eun Ha Kim et al., “A Study on Finding Welfare Blind Area Using Social Security Information System”, Ministry of Health and Welfare Policy Report, 2015
[27] Teconer Inc., http://www.teconer.fi/en/winter.html, 2015

이민우 (Min Woo, Lee)
전남대학교 컴퓨터공학과 (공학사)
연세대학교 컴퓨터과학과 박사 수료
디토닉주식회사 기술연구소 선임연구원
<관심분야> : 빅데이터, 자료 구조, 공간 기하학
E-mail : minwoo.lee@dtonic.io

김영곤 (Young Gon Kim)
강원대학교 도시·환경방재공학 석사
디토닉주식회사 기술연구소 연구원
<관심분야> : 빅데이터 통계분석, 도로기상
E-mail : kyg@dtonic.io

김강화 (Kang Hwa Kim)
한동대학교 전산전자공학부 학사
디토닉주식회사 기술연구소 수석연구원
<관심분야> : 빅데이터 인프라 구축/처리(엔지니어링), 빅데이터 통계분석
E-mail : reinforcement@dtonic.io

전용주 (Yong Joo Jun)
고려대학교 전기전자전파공학 학사
디토닉 주식회사 대표이사
<관심분야> : 빅데이터 비즈니스 모델, C-ITS, V2X
E-mail : richard@dtonic.io

용환성 (Hwan seong, Yong)
고려대학교 컴퓨터정보통신공학 석사
한양대학교 생산서비스경영학과 박사 과정
<관심분야> : 프로젝트 거버넌스 · 편익관리, 인공지능, 증강현실, 블록체인
E-mail : hsyong71@naver.com

이석준 (Seog Jun, Lee)
고려대학교 산업공학 학사
고려대학교 산업공학 석사
University of Wisconsin 산업공학 박사
건국대학교 경영대학 경영정보학과 교수
<관심분야> : EA, 정보화 성과평가, IT-ROI, e-Health, 의사결정 등
E-mail : seogjun@konkuk.ac.kr