KSCE Journal of Civil and Environmental Engineering Research
Vol. 43, No. 6: 841-849/ December, 2023
DOI: https://doi.org/10.12652/Ksce.2023.43.6.0841

ISSN 1015-6348 (Print)
ISSN 2799-9629 (Online)
www.kscejournal.or.kr

Information Technology **정보기술**

건설 리스크 도출을 위한 SVM 기반의 건설프로젝트 문서 분류 모델 개발

강동욱* · 조민건** · 차기춘*** · 박승희****
Kang, Donguk*, Cho, Mingeon**, Cha, Gichun***, Park, Seunghee****

Development of SVM-based Construction Project Document Classification Model to Derive Construction Risk

ABSTRACT

Construction projects have risks due to various factors such as construction delays and construction accidents. Based on these construction risks, the method of calculating the construction period of the construction project is mainly made by subjective judgment that relies on supervisor experience. In addition, unreasonable shortening construction to meet construction project schedules delayed by construction delays and construction disasters causes negative consequences such as poor construction, and economic losses are caused by the absence of infrastructure due to delayed schedules. Data-based scientific approaches and statistical analysis are needed to solve the risks of such construction projects. Data collected in actual construction projects is stored in unstructured text, so to apply data-based risks, data pre-processing involves a lot of manpower and cost, so basic data through a data classification model using text mining is required. Therefore, in this study, a document-based data generation classification model for risk management was developed through a data classification model based on SVM (Support Vector Machine) by collecting construction project documents and utilizing text mining. Through quantitative analysis through future research results, it is expected that risk management will be possible by being used as efficient and objective basic data for construction project process management.

Keywords : Construction delays, Construction accident, Data classification, Text mining, Machine learning

초 록

건설프로젝트는 공기 지연, 건설 재해 등 다양한 요인으로 인한 리스크가 존재한다. 이러한 건설 리스크를 기반으로 건설프로젝트의 공사 기간의 산정 방법은 주로 감독자 경험에 의존한 주관적 판단으로 이루어지고 있다. 또한, 공기 지연과 건설 재해로 지연된 건설프로젝트 일정을 맞추기 위한 무리한 단축 시공은 부실시공 등의 부정적인 결과를 초래하며, 지연된 일정으로 인한 사회 기반 시설물 부재로 경제적 손실이 발생한다. 이러한 건설프로젝트의 리스크 해결을 위한 데이터 기반의 과학적 접근과 통계적 분석이 필요한 실정이다. 실제 건설프로젝트에서 수집되는 데이터는 비정형 텍스트 형태로 저장되어 있어 데이터를 기반으로 한 리스크를 적용하기 위해서는 데이터 전처리에 많은 인력과 비용을 수반하기 때문에 텍스트 마이닝을 활용한 데이터 분류 모델을 통한 기초자료를 요구한다. 따라서, 본 연구에서는 건설프로젝트 문서를 수집하여 텍스트 마이닝을 활용하여 SVM(Support Vector Machine) 기반의 데이터 분류 모델을 통해 리스크 관리를 위한 문서 기초자료 생성 분류 모델을 개발하였다. 향후 연구 결과를 통해 정량적인 분석을 통해서 건설프로젝트 공정관리 등에 있어 효율적이고 객관적인 기초자료로 활용되어 리스크 관리가 가능해질 것으로 기대된다.

검색어 : 공기 지연, 건설 재해, 데이터 분류, 텍스트 마이닝, SVM

* 정회원 · 성균관대학교 글로벌스마트시티융합공학과 석사과정 (Sungkyunkwan University · kdu7964@skku.edu)
** 정회원 · 성균관대학교 글로벌스마트시티융합공학과 박사과정 (Sungkyunkwan University · raonik6713@skku.edu)
*** 정회원 · 성균관대학교 글로벌스마트시티융합공학과 연구교수 (Sungkyunkwan University · ckckicun@skku.edu)
**** 종신회원 · 교신저자 · 성균관대학교 건설환경공학부 교수 (**Corresponding Author** · Sungkyunkwan University · shparkpc@skku.edu)

Received October 13, 2023/ **revised** October 28, 2023/ **accepted** October 30, 2023

Copyright ⓒ 2023 by the Korean Society of Civil Engineers
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. 서론

건설프로젝트는 정해진 공정계획에 따라 완수하는 것이 프로젝트 성공의 중요한 요소이다(Islam et al., 2019). 그러나 건설프로젝트의 특성상 인력, 자재, 장비 등의 자원들이 투입되는 복잡한 과정과 외부 건설환경의 변화들로 인하여 리스크 요인들에 노출되어 있다(Chun et al., 2001). 다양한 건설 리스크 요인 중 예기치 않은 공기 지연은 공사 금액 증가, 발주처 클레임 등으로 인하여 경제적 손실을 초래하고 있다(Park et al., 2023). 또한, 공사 기간을 맞추기 위한 무리한 단축 시공은 부실시공, 설계변경 등의 부정적인 결과를 초래할 수 있으며, 건설업의 이익률을 감소하는 부정적인 영향이 발생한다(Park, 2012). 전 세계적으로도 건설프로젝트가 예정된 공정계획보다 지연되는 사례는 빈번하게 발생하고 있으며, 이는 지금까지 연구 결과에도 불구하고 개선이 되지 않음을 나타낸다(Glenigan, 2015). 또한, 이러한 공기 지연으로 인한 경제적 손실을 전 세계적으로 추산하면 매년 6,200억 달러에 달하며, 약 1조 1,000억 달러 규모의 잠재적인 손실이 발생한다(Oxford Economics, 2017).

건설프로젝트의 공기 지연 요인은 설계변경부터 민원 발생까지 다양하게 발생하였다(Mulholland and Christian, 1999). 가장 많이 보고된 공기 지연 리스크 요인은 '불확실한 기상'(Al-Refaie et al., 2021). 또한, 자재 및 노동 생산성의 부정확한 산정(Kim et al., 2020)과 자원 부족(El-Sayegh et al., 2021) 등이 있다. 설계단계와 시공단계에서 발생한 공기 지연의 경우는 발주자의 짧은 공기 산정(Clegg et al., 2018)과 시공사의 부족한 시공 능력(Jitpaiboon et al., 2019), 건설프로젝트의 복잡한 특성을 반영하지 못한 단편적인 시공 기간 산출(Koppenjan et al., 2011) 등이 있다. 이는 경험적인 기준으로 산정된 무리한 공사 기간 산출이 원인으로 건설프로젝트 공기 지연이 발생하였다. 또한, 건설프로젝트의 완수 가능성보다 경험과 정치적인 문제에 의해 공기를 짧게 설정하는 관행에 의해 발생한다(KICT, 2020).

국내 건설업의 경우, 2021년 산업재해로 인한 사고사망자가 828명 중 417명으로 산업군 중 가장 높은 50.36%로 나타났다(Kosha, 2022). OECD 국가의 건설업 건설 재해 사망사고 실태 분석 보고서에서 국내 건설 재해로 인한 사고 사망자 수는 전체 35개국 회원국 중 2번째로 근로자 수 대비 많은 건설 재해가 발생하였다(Choi, 2020). 하지만, 건설 재해를 예방하기 위한 연구는 건설프로젝트 현황 파악 부족으로 제대로 이루어지지 않고 있으며, 국내 건설 재해 감소를 위한 연구의 방향을 참고할 수 있는 정형화된 자료가 부족한 상황이다(Yang and Lim, 2021).

최근 경험 의존적 공사 기간 산출 방법에서 실제 데이터 기반의 공기 산정을 위한 건설프로젝트 리스크가 다양한 방법을 통해 연구되고 있다. 국내 건설사의 해외 건설프로젝트 공정관리를 위하여 군집화 작업을 수행하여 24개의 공기 지연 리스크 요인을 도출하여 대응 방향성을 제안하였으며(Kim, 2022), 건설프로젝트의 공기 지연으로 인한 발주자의 클레임 비용을 감소시키기 위해 공기 지연 일수와 귀책 여부를 판별하였다(Lee, 2023). Durdyev and Hosseini(2020)은 1985년부터 2018년까지의 발표된 논문을 분석하여 건설프로젝트의 리스크 원인을 예방하고 해결방안을 제시하였다. 또한, Wang et al.(2020)은 상하이에서 수행된 건설 프로젝트 감독일지를 분석하여 공기 지연 리스크를 정의하였다. Yu et al.(2021)는 대규모 건설사업에서 일정 및 비용 관리 문제를 해결하기 위하여 지연관리지수(Delay Management Index, DMI)를 개발하고 사례 분석을 통해 지연 예방체계를 제안하였다.

Park(2012)은 정부와 민간 간의 공기 산정 방법을 조사하여 공기 산정 방향성을 제시하였다. Kang et al.(2017)은 기존 건설 데이터 및 경험 자료를 통해 공기 산정 정형화 산정 방법론 및 산정체계를 실적으로 구조화하여 공기를 자동으로 생성하였다. 몬테카를로 시뮬레이션을 통한 공기 산정 모델 연구는 건설프로젝트 공정표를 수집하여 도출된 공기 변화 확률분포로 공기 산정 모델을 개발하였으나, 상세한 공기 지연 리스크 분류의 한계를 제시하였다(Park et al., 2010).

또한, 건설 재해를 예방하기 위한 다양한 연구가 진행되었다. Jo(2012)는 건설 재해를 분석하여 건설 재해 요인을 도출하고 예방하는 방향성을 제시하였다. Ha et al.(2018)은 건설 재해 보상 지급액 데이터를 수집하여 몬테카를로 시뮬레이션을 통해 건설 현장 인명사고 발생확률과 손해배상액 확률분포를 분석하였다. 그리고 Zhang et al.(2019), Choi et al.(2021), Yang and Lim (2021)은 건설 재해 사례를 수집하여 텍스트 마이닝을 통해 비정형 텍스트 데이터를 분석하여 건설 재해 요인을 도출하고 예방을 위한 비정형 데이터 활용을 제안하였다.

건설프로젝트의 공기 지연과 건설 재해에 대하여 리스크 요인을 도출하는 연구는 일부 사례에서만 실적 데이터를 사용하고 있으며, 확보하기 쉬운 사고 사례 데이터를 중심으로 진행되었다. 그 결과, 실제 건설프로젝트에서 발생하는 다양한 요인에 의한 리스크에 대응하는 능력이 제한적이며, 계획 공정표와 준공공정표와의 비교의 부재로 실제 지연분석에 대한 현장 실용성과 적용에 한계가 있다.

본 논문은 건설프로젝트 수행 과정에서 리스크 요인을 가지고 있을 것으로 판단되는 건설프로젝트 문서 공사일지와 송수신 공문에서 건설프로젝트 중 가장 복잡한 공정 과정 및 리스크 요인을 가지고 있는 토공사로 한정하여 프로젝트 문서분류를 통해 비정형 데이터를 정형화하여 기초 데이터를 도출하고 데이터베이스에 구축하고자 한다. 건설 현장에서의 공사일지와 송수신 공문은 시공단

계에서 발생하는 리스크 발생 데이터와 구간별 작업 데이터를 통해 잠재적인 공기 지연과 건설 재해 데이터를 도출할 수 있을 것으로 판단된다. 또한, 송수신 공문을 통해 발주처와 시공사 및 설계사의 시공단계에서의 공정 정보 확인을 통해 공사일지의 신뢰성을 높일 수 있다. 따라서, 본 연구에서는 비정형 데이터로 저장된 공사일지와 송수신 공문을 수집하여 머신러닝을 통한 건설프로젝트 문서 분류 모델을 개발하여 건설프로젝트 리스크 요인을 도출하고 기초 데이터를 생성하고자 한다.

2. 건설프로젝트 문서 분류 모델 개발

본 연구에서 제안하는 건설프로젝트 문서 분류 모델 개발은 비정형 데이터인 공사일지와 송수신 공문을 고려할 수 있도록 데이터 수집(Data collection), 데이터 전처리(Pre-processing), 데이터 수치화(Data normalization), 분류 모델 개발(Model development), 모델 분석(Analysis) 5단계로 진행하였다. 건설프로젝트 문서분류 프로세스를 Fig. 1에 나타내었다.

데이터 수집 단계에서는 2009년부터 2022년까지 수행된 국내 도로공사 현장 12개 구간 75개 공구의 공사일지와 송수신 공문을 수집하였고, 데이터 전처리 단계에서는 수집한 건설프로젝트 문서를 수치화할 수 있도록 문장 단위의 작업 내용을 분류할 수 있도록 데이터 전처리를 수행하였다. 데이터 수치화 단계에서 문장으로 전처리된 텍스트를 TF-IDF(Term Frequency-Inverse Document Frequency) 기법을 통해 수치화하였다. 건설프로젝트 문서분류 모델의 경우 다양한 머신러닝(Naive Bayes, Support Vector Machine, Random Forest, Gradient Boost, Extreme Gradient Boosting) 모델을 적용하였다. 모델 분석단계에서 성능을 평가하여 최적의 모델을 선별하였다.

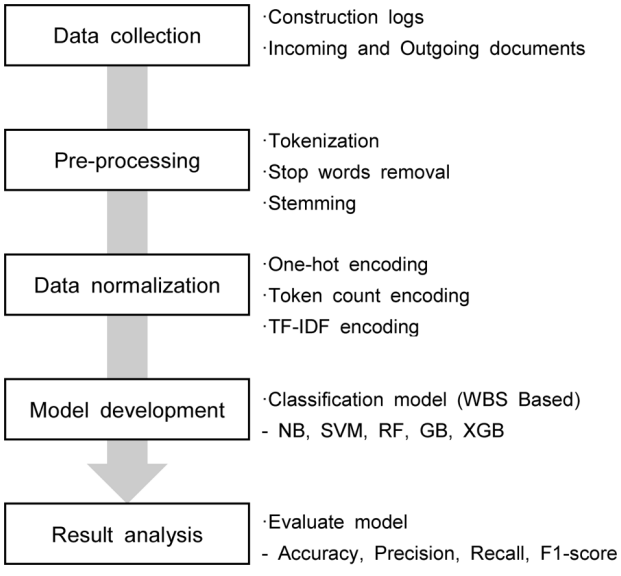

Fig. 1. Research Flow Chart to Document Classification Model

2.1 데이터 수집 및 전처리

수집된 공사일지와 송수신 공문은 작업 일자를 기반으로 당일 작업 내용과 감독 일지, 기상요건에 대해 기술되어 있으나, 공정별·구간별 내용이 정형화 되어 있지 않은 단순한 텍스트 뭉치의 비정형 텍스트 데이터이다. 공사일지의 경우 당일 공사 작업 내용이 상세히 기록되어 시공단계에서 발생하는 공기 지연과 건설 재해 발생 데이터를 확보할 수 있으며, 연속적인 작업 내용을 통해 흐름을 파악하고 계획 공정과 실제 공정의 비교를 통해 공기 지연일을 도출할 수 있다. 또한, 송수신 공문을 통해 공기 지연에 대한 리스크 요인을 비교 분석할 수 있다. 본 연구에서는 수집된 건설프로젝트 문서에서 가장 많은 리스크 요인이 발생하는 토공사를 중심으로 도로공사 현장 12개 구간, 75개의 공사일지와 송수신 공문을 수집하고 한국도로공사 7단계의 작업 분류체계(Work Breakdown Structure, WBS) 유형으로 분류하였다.

본 연구에서는 수집된 비정형 텍스트 데이터를 머신러닝 기법을 통한 분류 모델에 적용하기 위하여 전처리 작업을 수행하였다. 먼저, 작업 내용을 뭉쳐놓은 데이터를 문장 단위의 형식으로 변환하기 위하여 정규 표현식(Regular expression)을 활용한 문장 토큰화(Tokenization) 작업을 수행하였다. 또한, 불용어 처리(Stop words removal)(Lee, 2022)을 통해 문장에 불필요한 특수문자와 같은 요인들을 제거하였다. 그 후 Stemming 작업을 통해 문장 텍스트 데이터의 차원을 축소하고 일관성을 증가시키고 텍스트 분석의 성능을 향상하였다.

2.2 데이터 수치화

텍스트 데이터를 분류 모델에 입력하기 전에 문장을 수치형 데이터로 변환하는 embedding 작업을 수행하였다. 데이터 수치화 작업은 텍스트 데이터를 머신러닝이나 통계 분석에 적합한 형태로 변환하는 중요한 전처리 과정이다(Singh and Singh, 2020). 본 연구에서는 TF-IDF 기법(Lee et al., 2019)을 적용하여 토큰화된 문장을 수치 데이터로 변환하였다.

TF(Term frequency)는 텍스트가 주어질 때 단어가 몇 번 출현했는지 보여주는 수치이다. TF 기법은 “텍스트가 있을 때 단어가 여러 번 출현한다면 그 여러 번 출현한 만큼 연관성이 높을 것이다.” 라는 가설로 사용하는 값이다(Christian et al., 2016).

$$TF = tf(t,d) \tag{1}$$

Eq. (1)의 $tf(t, d)$는 문서(문장) d에서 단어 t가 출현한 횟수를

구한 것으로 단어의 출현 빈도를 나타낸다. 즉, 출현 빈도가 가장 높은 단어는 문장에서 가장 중요한 단어이다. 하지만 TF 값은 단순히 단어 빈도가 높다고 판단하기 때문에 특정 단어가 문서나 문장 전체와 관련 없는 'a'와 같은 연관성이 없는 단어가 중요하다고 도출될 수 있다. 이러한 TF 값의 통계적 오류를 해결하기 위해 IDF(Inverse document frequency) 값을 활용한다. 어떤 단어는 문서(문장)에서 특정 단어가 문서 내에 중요한 의미를 갖는지 통계적인 값으로 나타낸다.

$$IDF = \log \frac{D}{1+df(t)} \tag{2}$$

Eq. (2)는 역문서의 출현 빈도를 나타낸다. 전체 문서 D에서 특정 단어 t가 출현한 문서의 역수를 의미한다. 즉, 한 단어가 문서 전체에 얼마나 공통적으로 출현하는지 나타낸다.

$$TF-IDF = tf(t,d)*\log \frac{D}{1+df(t)} \tag{3}$$

따라서, TF-IDF 값은 Eq. (3)와 같이 계산된다. TF-IDF의 수학적 정의는 단어 빈도와 역문서 빈도의 곱(Lee et al., 2019)이라고 볼 수 있다. 여러 문서가 있을 때, 특정 단어의 중요한 정도값을 통해 의미를 갖는지 보여준다. 이와 같이 TF-IDF는 어떤 문서가 주어질 때, 각 단어별 출현 빈도를 통해 문장의 연관성을 수치화 한다.

2.3 머신러닝 기반 건설프로젝트 문서 분류 모델

머신러닝 기반의 건설프로젝트 문서 분류 모델의 경우, TF-IDF로 수치화한 데이터 세트를 한국도로공사 작업분류 체계 유형으로 다중 분류하였다. 머신러닝 알고리즘은 집중적인 정보 처리를 통해 최적의 의사결정을 위한 신뢰할 수 있는 결과를 예측한다(Raschka, 2015). 우수한 5가지의 머신러닝 알고리즘(NB, SVM, RF, GB, XGB)에 적용하여 분류 모델을 개발하고 성능을 평가하였다.

NB(Naive Bayes)은 베이즈 정리를 기반으로 하는 확률적 분류 모델로, 간단하고 빠르게 학습할 수 있고 작은 데이터 세트에도 사용할 수 있는 알고리즘이다(Saritas and Yasar, 2019). SVM(Support Vector Machine)은 데이터를 구분하는 결정 경계(decision boundary)와 각 클래스의 가장 가까운 데이터 포인트 사이의 마진을 최대화하여 새로운 데이터에 더 높은 성능을 보여주는 알고리즘이다(Alirameza et al., 2022). RF(Random Forest)는 의사결정나무(Decision Tree)를 여러 개 결합한 앙상블(Ensemble) 방법의 하나로, 데이터의 부트스트랩 샘플링(Bootstrap Sampling)과 무작위 특성 선택(Random Feature Selection)을 통해 예측 결과를 도출하는 알고리즘이다(Bressan et al., 2020). GB(Gradient Boost)는 약한 성능의 예측 모델들, 특히 의사결정 트리(Decision tree)를 순차적으로 학습시켜 강력한 예측 모델을 만드는 앙상블 기법이며, 회귀, 분류 등을 포함한 다양한 학습 작업에서 높은 성능을 보이고 고차원 대용량 데이터 세트에서도 효과적인 알고리즘이다(Natekin and Knoll, 2013). 마지막으로 XGB(Extreme Gradient Boosting)는 그래디언트 부스팅 알고리즘 개념을 기반으로 하는 다양한 최적화와 정규화 기법을 적용한 알고리즘이다(Li et al., 2019).

머신러닝 기반 문서 분류 모델의 성능은 하이퍼파라미터(Hyper-parameter)에 크게 의존하므로, 최적화하는 것이 중요한 요소로 작용한다(Raschka, 2015). 본 연구에서는 머신러닝 기반 문서 분류 모델의 일반화를 위해 5겹 교차 검증(5-fold cross-validation) (Normawati and Ismi, 2019) 방식을 사용하여 하이퍼파라미터를 설정하였다. 5겹의 교차 검증을 사용함으로써 모델의 일반화 성능에 대한 더 신뢰성 있는 결과를 얻을 수 있다(Wong and Yeh, 2019).

건설프로젝트 문서 분류 모델 검증 및 평가는 한국도로공사 WBS 분류체계 7단계로 구성된 데이터로 학습된 5가지 머신러닝 알고리즘을 적용하여 검증용 데이터를 사용하여 정확도(Accuracy), 정밀도(Preicision), 재현율(Recall), F1 점수(F-1 score)를 계산하여 정량적 평가를 진행하였고, 최적의 분류 성능 모델을 분석하였다.

분류성능을 평가하는 데 사용되는 혼동 행렬(Confusion Matrix)은 실제 클래스와 예측 클래스 간의 관계를 나타내는 표현으로, 머신러닝 알고리즘의 성능을 시각화할 수 있다(Bressan et al., 2020). 혼동 행렬은 이진 분류를 위해 생성된 표는 Fig. 2에 나타내었다. 주대각선은 평가된 기법의 정확도를 나타내며, 실제 결과를 조직화 된 구조로 결합한다. 혼동 행렬은 주로 이진 분류에서 사용되지만, 다중 클래스 분류에도 확장될 수 있다.

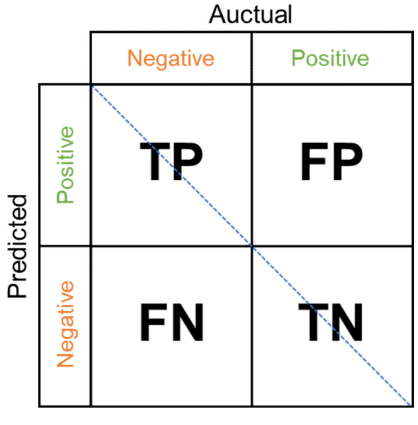

Fig. 2. Definition of the Confusion Matrix

정확도는 Eq. (4)와 같이 전체 예측 중 올바르게 예측된 비율을 의미하며, Eq. (5)는 정밀도로 Positive로 예측되었을 때 실제로 Positive인 비율을 의미하고 Eq. (6)은 재현율로 실제 Positive일 때 Positive로 예측된 비율을 의미한다. 또한, Eq. (7)에서 F1 점수는 정밀도와 재현율의 조화 평균을 의미한다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 = \frac{2*\mathrm{Precision}*recall}{\mathrm{Precision}+recall} \quad (7)$$

3. 건설프로젝트 문서 분류 모델 평가

본 연구에서는 비정형 텍스트 데이터인 건설프로젝트 문서를 텍스트 마이닝(Text Minig)을 활용하여 머신러닝 기반 분류 모델을 통해 최적의 건설프로젝트 문서 분류 모델을 제안하고자 한다. 건설프로젝트 문서를 수집하여 텍스트 마이닝을 활용한 머신러닝 기반 분류 모델은 총 7단계 유형의 WBS 기준에 따라 문장 단위의 분류 모델을 개발하고 정량적 평가를 진행하였다. 평가는 Accuracy와 Precision, Recall, F1-score를 고려하여 판단하였다.

3.1 분류 모델의 성능 분석 및 평가

본 연구에서는 수집된 건설프로젝트 문서에서 토공사로 한정된 공사일지를 수집하고 전처리 작업이 수행된 데이터는 Table 1과 같이 한국도로공사 7단계의 작업 분류체계(WBS)(Park et al., 2023) 분류 기준으로 클래스를 지정하였다.

7단계의 작업 분류체계 중 3단계(시설물, 작업관리 활동1, 작업관리 활동2) 그룹으로 학습되어 활용하였다. 단계별 특징을 살펴보면 시설물은 '본선(Main line), IC/JC(Interchange/Juction), 지선/부체도로(Side/Access road)'으로 3구간으로 1차 분류하였다. 작업관리 활동1의 경우 '흙깎기(Soil cut), 흙쌓기(Soil fill), 토공 기타(Earth work, etc)'로 3공정으로 2차 분류하였으며, 작업관리 활동2는 '노상(Subsoil), 노체(Subgrade), 리핑(Reaping), 발파(Blasting), 비탈면 보호공(Slope protection), 연약지반처리(Soft ground improvement), 옹벽 기타(Concrete retaining wall/ Retaining wall), 흙깎기 기타(Soil cut, etc), 흙쌓기 기타(Soil fill, etc), 토공 기타(Earth work, etc)'로 10가지 공정을 3차 분류하였다. 분류된 데이터를 변환하여 3개 구간, 3가지 유형의 10개 작업 공정으로 문장 단위의 분류 작업을 수행하였다.

Table 2에 나타낸 것과 같이, TF-IDF 기법을 통해 수치화된 데이터를 머신러닝 알고리즘(NB, SVM, RF, GB, XGB)에 적용한 건설프로젝트 문서 분류 모델은 Accuracy가 83.56~91.78%, F1-score가 0.79~0.92로 높은 분류성능을 나타내었다. 그 중, SVM 알고리즘의 건설프로젝트 문서 분류 모델이 가장 높은 Accuracy (91.78%)와 F1-score (0.92)로 분석되었다.

Table 2. Predictive Performance of Machine Learning Algorithms Using Data Normalization

Evaluation / Methods	Accuracy (%)	Precision	Recall	F1-score
NB	83.56	0.90	0.76	0.79
SVM*	91.78*	0.93*	0.92*	0.92*
RF	90.18	0.92	0.90	0.90
GB	90.18	0.93	0.90	0.91
XGB	86.76	0.90	0.84	0.86

*The best algorithm with the highest predictive performance

Table 1. Construction Work Breakdown Structure (WBS) Classification Criteria

	WBS Classification criteria		
Lv3 (Construction Kind)	Lv4(Facility)	Lv7-1(Activity1)	Lv7-2(Activity2)
Earthwork	Interchange/Junction(IC/JC)	Soil cut	Soil, Reaping, Blast, Soil cut, etc
	Main line	Soil fill	Subsoil, Subgrade, Soil fill etc
	Side/ Access road	Earth work, etc	Slope protection Soft ground improvement Earth retaining wall Concrete retaining wall Retaining wall Earth work etc

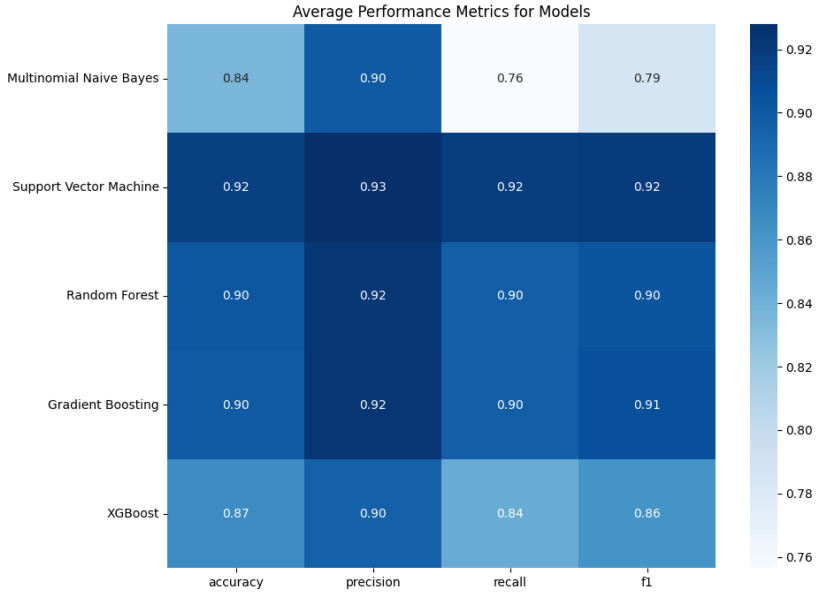

Fig. 3. Average Performance Metrics for Models

Fig. 3과 같이, TF-IDF에 의해 수치화된 데이터 세트를 WBS 분류 기준으로 분류한 5가지 모델 중 데이터를 선형으로 분리하는데 효과적인 SVM 기법의 분류 모델이 가장 높은 분류성능을 나타내는 것으로 평가되었다. 또한, TF-IDF 벡터화 기법을 적용하여 고차원의 수치 벡터로 변환하였을 때 SVM 기법이 효과적이라는 것으로 나타났다. 따라서, 본 연구에서는 선형 텍스트 데이터와 TF-IDF의 특성을 고려한 비정형 텍스트 데이터 특성에 가장 높은 성능의 SVM 기반의 분류 모델을 건설프로젝트 문서 최적의 모델로 제안하였다.

3.2 분류 모델의 분석 결과

본 연구에서는 제안한 분류 모델의 한국도로공사 WBS 기준으로 토공사를 한정하여 단계별 예측정확도를 평가하였으며, Fig. 4와 같이, 작업 분류체계 단계별로 혼동 행렬(Confusion Matrix)을 도출하였다.

시설물의 분포는 3개 구간 본선(37.33%), IC/JC(49.32%), 지선/부체도로(13.36%)로 구분되어 데이터가 비교적 적은 지선/부체도로는 Accuracy가 84.21%로 평가되었다. 시설물에 대한 데이터가 가장 많은 IC/JC의 경우는 Accuracy 90.36%의 높은 정확도를 보여주었다. 또한, 본선의 경우 Accuracy가 93.18%로 Overall accuracy의 91.78%보다 높은 분류성능을 보이며, 건설프로젝트의 특성상 형식이 유사한 본선은 텍스트 형태로 인하여 높은 분류성능 특징을 가지고 있다. 작업관리 활동1 유형의 경우 흙깎기(25.17%), 흙쌓기(28.08%), 토공 기타(46.75%)의 분포를 보여주고 있으며, 토공사의 경우 흙깎기와 흙쌓기를 위한 부가적인 토공 기타에 많은 소요일수가 발생함을 알 수 있었다. 또한, 작업관리 활동1

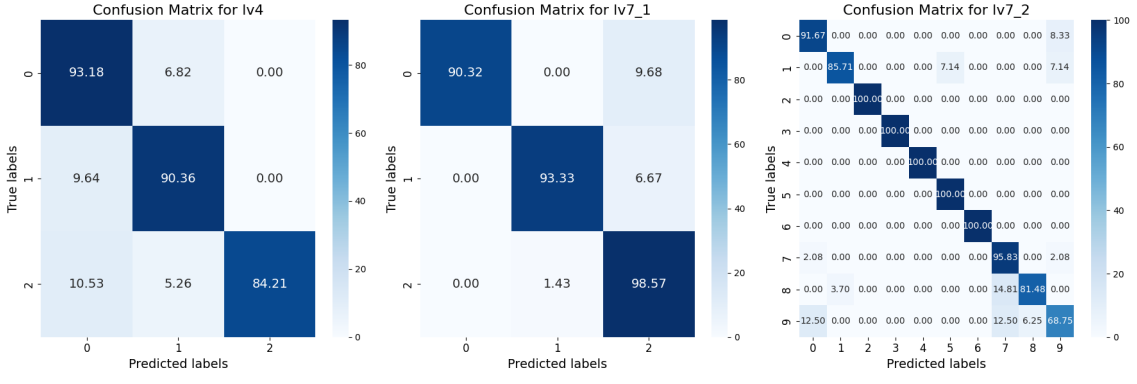

Fig. 4. Confusion Matrix for Proposed Model

유형은 적절한 데이터 비율을 통해 Accuracy 90% 이상의 높은 분류정확도를 보여주었다. 작업관리 활동2 공정 분포는 노상(6.50%), 노체(9.25%), 리핑(1.71%), 발파(3.25%) 비탈면 보호공(6.16%), 연약지반처리(2.23%), 옹벽 기타(7.02%), 흙깎기 기타(21.75%), 흙쌓기 기타(10.62%), 토공 기타(31.51%) 순으로 나타났다. 작업관리 활동2 분류는 10개의 공정 유형으로 분류를 세분화하였으며, Average accuracy는 92.34%로 높은 분류정확도를 보이는 것으로 평가되었다. 하지만, 토공 기타의 경우 토공사의 공정에 속해 있지 않는 건설 재해와 공기 지연 리스크가 포함되어 있다. 따라서, 데이터의 특성이 다양하고 복잡하여 낮은 분류 Accuracy 68.75%를 보여주었다.

4. 결 론

건설프로젝트에서 공기 지연 요인과 건설 재해 등 다양한 요인에 의하여 리스크가 발생 되고 있다. 건설 리스크로 인한 공기 지연은 공사 비용을 증가시키고 공사 기간을 맞추기 위한 무리한 시공은 건설 재해로 이어질 수 있다. 또한, 기반 시설물의 부재는 교통의 흐름을 방해하는 등 막대한 경제적 손실을 초래한다. 건설 재해는 사회문제를 발생시키기 때문에 예방과 해결책 수립이 중요하다. 이러한 건설 리스크를 해결하기 위해 실제 건설프로젝트 데이터 기반의 공기 산정이 요구된다. 건설프로젝트 문서는 감독자의 수기를 통해 기록되어 비정형 텍스트 데이터의 특징을 가지고 있으며, 정량적 분석을 위한 문서의 디지털화가 필요하다. 따라서, 본 연구에서는 건설 리스크 도출을 위한 SVM 기반의 건설프로젝트 문서 분류 모델을 개발하였다. 이를 위해 도로공사 현장 12개 구간, 75개 공구의 건설프로젝트 문서인 공사일지와 송수신 공문을 수집하였다. 건설프로젝트 8개 공사 중 자연재해나 기술적 어려움을 겪는 토공사를 한정하여 한국도로공사 WBS 분류 기준을 모델에 적용하였다. WBS 분류는 7단계 중 3단계 과정으로 진행되었으며, '시설물, 작업관리 활동1, 작업관리 활동2'그룹으로 분류하여 WBS 기준을 산정하였다. 건설프로젝트 문서의 비정형 텍스트 데이터를 분류하기 위하여 다양한 분야에서 사용되는 분류 알고리즘을 비교하였다. 수집된 데이터는 전처리 과정을 거쳐 TF-IDF를 통해 수치화하여 분류 알고리즘 모델들에 적용한 결과, SVM 분류 모델이 Accurcy 94.44%, Precision 0.87, Recall 0.83, F1-score 0.85로 가장 높은 분류성능을 보여주었다.

본 연구에서 개발된 건설 리스크 도출을 위한 건설프로젝트 문서 분류 모델은 건설공정관리의 중요한 측면에서 2가지 의미가 있다.

첫째, 건설프로젝트는 공기 지연과 건설 재해 등 다양한 요인에 리스크를 내포하고 있으나, 계획 공정표는 주로 감독자의 경험에 의한 주관적 판단으로 공사 기간이 산정되는 경우가 많다. 본 연구를 통해 개발된 건설프로젝트 문서 분류 모델은 데이터 기반으로 건설 리스크를 관리하는 가능성을 제시한다. 둘째, 실제 시공단계에서 얻어진 건설프로젝트 문서 데이터는 대부분 비정형 텍스트 데이터로 구성되어 데이터 전처리에 많은 인력과 비용이 소요된다. 본 연구에서는 텍스트 마이닝을 활용한 SVM 기반의 데이터 분류를 통해 구간별 및 공정별로 효율적으로 분석하고 건설프로젝트의 공기 지연과 건설 재해에 대한 정보를 도출할 수 있다. 도출된 건설 리스크 기초자료는 향후 건설 현장에서의 리스크관리 및 효율적인 프로젝트 관리에 이바지할 것으로 기대한다.

하지만 본 연구에서는 국내 도로공사 현장에 한정하여 다양한 건설프로젝트의 복잡한 상황과 환경을 모두 반영되지 못했을 가능성이 있다. 또한, 정량적인 분류를 통해서 공기 지연과 건설 재해 요인과 해결책을 정성적으로 분석하기에 한계가 있다. 따라서, 향후 보다 다양한 건설프로젝트 문서를 확보하여 분류 모델을 적용하고 확보된 실제 데이터를 기반으로 공기 산정을 위한 시뮬레이션 모델을 적용하여 공정관리하고 건설 재해 요인을 통한 안전관리 대응 연구가 필요하다.

Acknowledgements

This research was conducted with the support of the "National R&D Project for Smart Construction Technology (22SMIP-A158708-03)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation. and This research was supported by RS-2023-00248092 of Technology Development Program on Disaster Restoration Capacity Building and Strenthening funded by Ministry of Interior and Safety(MOIS, Korea).

ORCID

Donguk Kang https://orcid.org/0009-0004-4410-1660
Mingeon Cho https://orcid.org/0000-0002-6375-717X
Gichun Cha https://orcid.org/0000-0002-6327-8742
Seunghee Park https://orcid.org/0000-0002-7518-4890

References

Aliramezani, M., Koch, C. R. and Shahbakhti, M. (2022). "Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques:

A review and future directions." *Progress in Energy and Combustion Science*, Elsevier, Vol. 88, 100967, https://doi.org/10.1016/j.pecs.2021.100967.

Al-Refaie, A. M., Alashwal, A. M., Abdul-Samad, Z. and Salleh, H. (2021). "Weather and labor productivity in construction: a literature review and taxonomy of studies." *International Journal of Productivity and Performance Management*, Emerald Publishing Limited, Vol. 70, No. 4, pp. 941-957, https://doi.org/10.1108/IJPPM-12-2019-0577 .

Bressan, T. S., de Souza, M. K., Girelli, T. J. and Junior, F. C. (2020). "Evaluation of machine learning methods for lithology classification using geophysical data." *Computers and Geosciences*, Elsevier, Vol. 139, 104475, https://doi.org/10.1016/j.cageo.2020.104475.

Choi, S. Y. (2020). "Comparison analysis of deaths in construction industry in OECD countries." *Construction and Economy Research Institute of Korea* (in Korean).

Choi, S. J., Kim, J. H. and Jung, K. (2021). "Development of prediction models for fatal accidents using proactive information in construction sites." *Journal of the Korean Society of Safety*, KOSOS, Vol. 36, No. 3, pp. 31-39, https://doi.org/10.14346/JKOSOS.2021.36.3.31 (in Korean).

Christian, H., Agus, M. P. and Suhartono, D. (2016). "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)." *ComTech: Computer, Mathematics and Engineering Applications*, Binus University, Vol. 7, No. 4, pp. 285-294, https://doi.org/10.21512/comtech.v7i4.3746.

Chun, Y. G., Lee, G. U., Kim, Y. T. and Hyun, C. T. (2001). "A study on the estimation standard of optimal construction duration for reinforced concrete apartment house." *Proceedings of 2nd Korea Institute of Construction Engineering and Management*, KICEM, Seoul, Korea, pp. 531-534 (in Korean).

Clegg, S., Killen, C. P., Biesenthal, C. and Sankaran, S. (2018). "Practices, projects and portfolios: Current research trends and new directions." *International Journal of Project Management*, Elsevier, Vol. 36, No. 5, pp. 762-772, https://doi.org/10.1016/j.ijproman.2018.03.008.

Durdyev, S. and Hosseini, M. R. (2020). "Causes of delays on construction projects: a comprehensive list." *International Journal of Managing Projects in Business*, Emerald Publishing Limited, Vol. 13, No. 1, pp. 20-46, https://doi.org/10.1108/IJMPB-09-2018-0178.

El-Sayegh, S. M., Manjikian, S., Ibrahim, A., Abouelyousr, A. and Jabbour, R. (2021). "Risk identification and assessment in sustainable construction projects in the UAE." *International Journal of Construction Management*, Taylor & Francis, Vol. 21, No. 4, pp. 327-336, https://doi.org/10.1080/15623599.2018.1536963.

Glenigan (2015). *UK Industry Performance Report.*

Ha, S. G., Kim, T. H., Son, K. Y., Kim, J. M. and Son, S. H. (2018). "Quantification model development of human accidents on external construction site by applying probabilistic method." *Journal of the Korea Institute of Building Construction*, KIC, Vol. 18, No. 6, pp. 611-619, https://doi.org/10.5345/JKIBC.2018.18.6.611 (in Korean).

Islam, M. S., Nepal, M. P., Skitmore, M. and Kabir, G. (2019). "A knowledge-based expert system to assess power plant project cost overrun risks." *Expert Systems with Applications*, Elsevier, Vol. 136, No. 1, pp. 12-32, https://doi.org/10.1016/j.eswa.2019.06.030.

Jitpaiboon, T., Smith, S. M. and Gu, Q. (2019). "Critical success factors affecting project performance: An analysis of tools, practices, and managerial support." *Project Management Journal*, PMI, Vol. 50, No. 3, pp. 271-287, https://doi.org/10.1177/8756972819833545.

Jo, J. H. (2012). "A study on the causes analysis and preventive measures by disaster types in construction fields." *Journal of the Korea Safety Management & Science,* Vol. 14, No. 1, pp. 7-13, https://doi.org/10.12812/KSMS.2012.14.1.007 (in Korean).

Kang, S. H., Jung, Y. S., Kim, S. R., Lee, I. H., Lee, C. W. and Jung, J. H. (2017). "Preliminary scheduling based on historical and experience data for airport project." *Korea Institute of Construction Engineering and Management*, KICEM, Vol. 18, No. 6, pp. 26-37, https://doi.org/10.6106/KJCEM.2017.18.6.026 (in Korean).

Kim, J. S. (2022). *Analysis of project delay using big data*. Msc. thesis, Hanyang University, Seoul, Korea (in Korean).

Kim, S., Chang, S. and Castro-Lacouture, D. (2020). "Dynamic modeling for analyzing impacts of skilled labor shortage on construction project management." *Journal of Management in Engineering*, ASCE, Vol. 36, No. 1, 04019035, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000720.

Koppenjan, J., Veeneman, W., van der Voort, H., ten Heuvelhof, E. and Leijten, M. (2011). "Competing management approaches in large engineering projects: The Dutch RandstadRail project." *International Journal of Project Management*, Vol. 29, No. 6, pp. 740-750, https://doi.org/10.1016/j.ijproman.2010.07.003.

Korea Institute of Civil engineering and building Technology (KICT) (2020). *Guidelines for ensuring adequate construction time* (in Korean).

Korea Occupational Safety and Health Agency (KOSHA) (2019). *2019 Large accident report book,* pp. 9 (in Korean).

Lee, W. J. (2022). "A study on the use of stopword corpus for cleansing unstructured text data." *The Journal of the Convergence on Culture Technology*, IPACT, Vol. 8, No. 6, pp. 891-897, https://doi.org/10.17703/JCCT.2022.8.6.891 (in Korean).

Lee, G. S. (2023). *Methods to analyze the delays for extension of time claims.* Msc. thesis, Hanyang University, Seoul, Korea (in Korean).

Lee, J. H., Lee, M. B. and Kim, J. W. (2019). "A study on Korean language processing using TF-IDF." *The Journal of Information Systems*, KAIS, Vol. 28, No. 3, pp. 105-121, https://doi.org/10.5859/KAIS.2019.28.3.105 (in Korean).

Li, W., Yin, Y., Quan, X. and Zhang, H. (2019). "Gene expression value prediction based on XGBoost algorithm." *Frontiers in genetics*, Frontiers, Vol. 10, 1077, https://doi.org/10.3389/fgene.

2019.01077.

Mulholland, B. and Christian, J. (1999). "Risk assessment in construction schedules." *Journal of Construction Engineering and Management,* Vol. 125, No. 1, pp. 8-15, https://doi.org/10.1061/(ASCE)0733-9364(1999)125:1(8).

Natekin, A. and Knoll, A. (2013). "Gradient boosting machines, a tutorial." *Frontiers in Neurorobotics*, Frontiers, Vol. 7, https://doi.org/10.3389/fnbot.2013.00021.

Normawati, D. and Ismi, D. P. (2019). "K-fold cross validation for selection of cardiovascular disease diagnosis features by applying rule-based datamining." *Signal and Image Processing Letters*, ASCEE, Vol. 1, No. 2, pp. 62-72, https://doi.org/10.31763/simple.v1i2.3.

Oxford Economics (2017). *Global Infrastructure Outlook*.

Park, G. S. (2012). *A study on the estimation of an approproate construction duration of apartment*. Msc. thesis, Seoul National University of Science and Technology (in Korean).

Park, J. H., Cho, M. G., Eom, S. H. and Park, S. K. (2023). "Quantification of schedule delay risk of rain via text mining of a construction log." *KSCE Journal of Civil and Environmental Engineering Research,* KSCE, Vol. 43, No. 1, pp. 109-117, https://doi.org/10.12652/Ksce.2023.43.1.0109 (in Korean).

Park, J. H., Choi, H. S., Cho, Y., Bang, K. S., Yun, S. H. and Paek, J. H. (2010). "A study on the development of probabilistic duration estimation module using monte carlo simulation." *Journal of the Architectural Institute of Korea Structure and Construction*, AIK, Vol. 26, No. 5, pp. 101-108 (in Korean).

Raschka, S. (2015). *Python machine learning,* 1st ed., Packt publishing Ltd.

Saritas, M. M. and Yasar, A. (2019). "Performance analysis of ANN and Naive Bayes classification algorithm for data classification." *International Journal of Intelligent Systems and Applications in Engineering*, Elsevier, Vol. 7, No. 2, pp. 88-91, https://doi.org/10.18201//ijisae.2019252786.

Singh, D. and Singh, B. (2020). "Investigating the impact of data normalization on classification performance." *Applied Soft Computing*, Elsevier, Vol. 97, Part B, 105524, https://doi.org/10.1016/j.asoc.2019.105524.

Wang, G., Liu, M., Cao, D. and Tan, D. (2020). "Identifying high-frequency-low-severity construction safety risks: An empirical study based on official supervision reports in Shanghai." *Engineering, Construction and Architectural Management*, Emerald Publishing Limited, Vol. 29, No. 2, pp. 940-960, https://doi.org/10.1108/ECAM-07-2020-0581.

Wong, T. T. and Yeh, P. Y. (2019). "Reliable accuracy estimates from k-fold cross validation." *IEEE Transactions on Knowledge and Data Engineering*, IEEE, Vol. 32, No. 8, pp. 1586-1594, https://doi.org/10.1109/TKDE.2019.2912815.

Yang, S. W. and Lim, H. C. (2021). "Semantic network analysis on the research trends of construction accident." *Journal of the Architectural Institute of Korea*, AIK, Vol. 37, No. 6, pp. 231-236 (in Korean).

Yu, J. H. and Kim, O. K. (2021). "A case study on the prevention of construction delays using the delay management index in program level construction projects." *Journal of the Korea Institute of Building Construction,* Vol. 21, No. 4, pp. 347-359, https://doi.org/10.5345/JKIBC.2021.21.4.347 (in Korean).

Zhang, F., Fleyeh, H., Wang, X. and Lu, M. (2019). "Construction site accident analysis using text mining and natural language processing techniques." *Automation in Construction*, Elsevier, Vol. 99, pp. 238-248, https://doi.org/10.1016/j.autcon.2018.12.016.