공학석사 학위논문

조선 생산 리드타임 예측을 위한 기계학습 방법론에 관한 연구

A study on Machine Learning for Prediction of the Shipbuilding Lead Time

지도교수 우종훈

2018년 2월

한국해양대학교 대학원

조선해양시스템공학과

김지혜

본 논문을 김지혜의 공학석사 학위논문으로 인준함.

위원장 남 종 호 (인)
위 원 박 주 용 (인)
위 원 우 종 훈 (인)

2018년 2월

한국해양대학교 대학원

목 차

List of Tables

List of Tables

List of Figures

List of Figures

List of Figures

<영문초록>

A study on Machine Learning for the Prediction of the Shipbuilding Lead Time

Kim, Ji Hye

Department of Naval Architecture and Ocean Systems Engineering
Graduate School of Korea Maritime and Ocean University

Abstract

In recent years, big data technology, which is one of the biggest issues in IT field, has been applied in various fields as data has increased exponentially compared to the past, however, in the shipbuilding and offshore industries, the use of big data related technology is relatively rare compared to other manufacturing industries such as automobile and electronics industries. But, shipbuilding and offshore industry is one-piece manufacturing industry, and statistics-based analysis such as the Big Data methodology can be very effective because vast amounts of data are generated throughout the entire life cycle and are highly variable in the manufacturing environment. As a result, the big data-based machine learning research is progressing slowly in the shipbuilding industry.

However, this is limited to the design field that manages the fixed variables and it is difficult to apply it in terms of production management such as lead time which is the basis of construction activity. In particular, the standard data such as production lead time is highly variable due to various process variables so, it is necessary to study changing from causation viewpoint to correlation to solve it.

Therefore, in this paper, I has constructed a prediction model applying machine learning and deep learning algorithm to improve the standard data for the time factor of production lead time. In order to predict the variable lead time considering the various properties of the product in comparison with the standard lead time, I collect data from several shipyards and apply various machine learning and deep learning algorithms to predict the production lead time according to the process. Respectively. To analyze the data, open source such as R and Python language was used and a lead time prediction model based on the algorithm was created. Various evaluation indices were used to evaluate the prediction model generated by the analysis algorithm. In addition, I compared the results of machine learning and deep learning algorithms with those of previous studies, and the decision support for the establishment of standard information according to various process variables is made possible.

KEY WORDS: Production Management; Standard Data; Big Data; Statistical Analysis; Machine Learning; Deep Learning.

〈국문초록〉

조선 생산 리드타임 예측을 위한 기계학습 방법론에 관한 연구

김지혜

한국해양대학교 대학원
조선해양시스템공학과

Abstract

최근 IT 분야에서 가장 큰 화두인 빅데이터 기술은 과거에 비해 데이터가 기하급수적으로 증가함에 따라 다양한 분야에서 적용되고 있지만 자동차, 전자 업종 등의 다른 제조업에 비해 조선 및 해양산업에서는 빅데이터 관련 기술의 활용사례가 상대적으로 드문 실정이다. 하지만 조선 및 해양산업은 일품 제조 산업으로 영업, 설계, 건조, 유지 보수 등 전체 수명 주기에서 방대한 데이터가 생성되고, 제조 환경에 따라 변동성이 크기 때문에 빅데이터 방법론과 같은 통계 기반 분석이 큰 효과를 발휘할 수 있다. 이로 인해 빅데이터 기반의 기계학습 연구는 조선업에서도 서서히 진행되고 있으나 이는 고정변수를 관리하는 설계 분야에 한정되어 있으며 건조 활동의 근간이 되는 기준 정보, 즉 원단위나 시수, 리드타임 등의 생산관리 관점에서는 적용에 어려움을 겪고 있다. 특히 생산 리드타임이라는 기준정보는 다양한 공정변수로 인한 변동성이 크기 때문에 이를 해결하기 위해서 현실적으로 한계가 있는 영역에 대해 Causation에서 Correlation 관점에 따른 연구가 필요하다고 본다.

따라서 본 논문에서는 생산 리드타임이라는 시간요소에 대한 기준정보 체계 개선을 위해 기계학습 및 딥러닝 알고리즘을 적용한 예측모델을 구축하였다. 기존에 관리되는 조선소의 표준 리드타임에 대비하여 제품의 다양한 속성을 고려한 변동 리드타임을 예측하기 위해 여러 조선소의 데이터를 수집하였고 공정에 따른 생산 리드타임을 예측하기 위한 다양한 기계학습 및 딥러닝 알고리즘을 적용하였다. 데이터를 분석하기 위해서 R과 Python 언어 등의 오픈소스를 활용하였으며 알고리즘에 따른 리드타임 예측모델을 생성하였다. 분석 알고리즘에 따라 생성된 예측모델의 평가를 위해 여러 가지 평가지표를 활용하였다. 또한, 기존연구 결과에 비해 기계학습과 딥러닝 알고리즘에 따른 유의미한 결과를 비교하여 조선소에서의 활용성을 검토해보고 다양한 공정변수에 따른 기준정보 수립에 대한 의사결정 지원을 가능하도록 하였다.

KEY WORDS: 생산관리 Production Management; 기준정보 Standard Data; 빅데이터 Big Data; 통계분석 Statistical Analysis; 기계학습 Machine Learning; 딥러닝 Deep Learning.

제 1 장 서 론

1.1 연구 배경

최근 IT 분야에서 가장 큰 화두가 되고 있는 것은 단연 빅데이터라고 할 수 있다. 과거에 비해 데이터의 양이 기하급수적으로 증가함에 따라 이를 일반적인 데이터베이스로는 관리하기 어렵기 때문에 대용량데이터의 수집, 저장, 분석 등을 체계적으로 수행하기 위한 다양한 기술들이 여러 분야에서 활용되고 있다. 빅데이터란 일반적인 데이터베이스, 소프트웨어로 관리하기 어려운 정도의 큰 규모로 기존의 방법이나 도구로 수집, 저장, 분석 등이 어려운 정형 또는 비정형 데이터를 모두 포함하는 대용량 데이터를 의미한다. 따라서 빅데이터 및 이에 기반을 둔 고도분석은 각종 산업분야에서 확대되고 있으며 이를 도입한 다양한 활용사례가 증가하고 있는 만큼 기업에서는 보다 빠르고 규모가 큰 데이터의 분석으로 기존의 IT 시스템을 혁신하고자 데이터의 가치를 극대화하는 방법을 적극적으로 찾고 있다.

하지만 다른 제조 산업에 비해 조선 및 해양플랜트 산업에서는 빅데이터의 활용 사례가 상대적으로 부족하다. 특히 조선/해양 산업에서도 생산 관리 관점에서의 여러 문제점을 해결하기 위해 다양한 시도 가운데 시간요소에 대한 기준정보 체계를 수립하기 위한 연구가 다방면에서 진행 중이다. 조선소에서 기존에 적용한 엔지니어링 분석 방법론은 제품 정보에 기반을 두어 물량 및 시수뿐만 아니라 생산 및 조달 리드타임을 분석하는 것으로 인과관계에 기반을 두었다고 볼 수 있다. 하지만 이는 오랜 연구에도 불구하고 현재까지도 미성숙하며 다양한 공정변수로 인해 현업에 적용한 사례가 드문 실정이다. 따라서 현실적으로 한계가 있는 영역에 대해서 causation에서 correlation 관점에 따른 연구 병행이 필요하다고 본다. 이를 위해 조선업에서도

빅데이터에 기반을 둔 통계분석을 활용하여 과거 데이터의 분석 및 학습을 통해 유의미한 예측 모델을 개발할 수 있다면 조선 생산 관점에서의 문제점을 해결할 수 있을 것이라 기대한다.

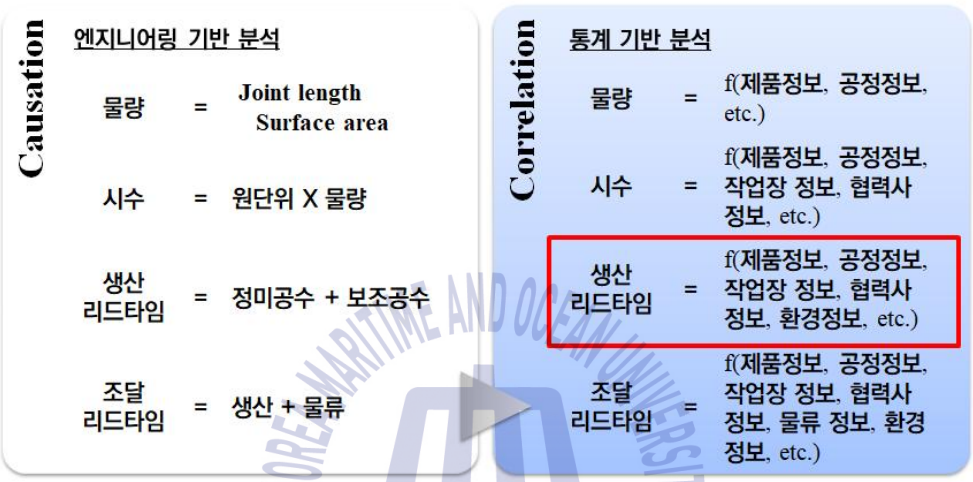


Fig. 1 Method of shipbuilding lead time prediction

따라서 본 논문에서는 생산 리드타임이라는 시간요소에 대한 기준정보 체계 개선을 위해 기계학습 및 딥러닝 알고리즘을 적용한 예측모델을 구축하였다. 기존에 관리되는 조선소의 표준 리드타임에 대비하여 제품의 다양한 속성을 고려한 변동 리드타임을 예측하기 위해 여러 조선소의 데이터를 수집하였고 공정에 따른 생산 리드타임을 예측하기 위한 다양한 기계학습 및 딥러닝 알고리즘을 적용하였다. 데이터를 분석하기 위해서 R과 Python 언어 등의 오픈소스를 활용하였으며 알고리즘에 따른 리드타임 예측모델을 생성하였다. 분석 알고리즘에 따라 생성된 예측모델의 평가를 위해 여러 가지 평가지표를 활용하였다. 또한, 기존연구 결과에 비해 기계학습과 딥러닝 알고리즘에 따른 유의미한 결과를 비교하여 조선소에서의 활용성을 검토해보고 다양한 공정변수에 따른 기준정보 수립에 대한 의사결정 지원을 가능하도록 하였다.

1.1.1 조선소의 생산관리

조선업은 일품 제조 산업으로 영업, 설계, 건조, 유지 보수 등 전체 수명 주기에서 방대한 데이터가 생성되고 특히 건조 활동의 근간이 되는 시수, 공정 리드타임과 같은 시간에 대한 기준정보가 체계적으로 수립되지 않아 빅데이터 기술 적용에 어려움을 겪고 있다.

대형 조선소의 설계 및 생산기술력은 세계 선두자리로 지키고 있는 반면 중소형 조선소는 이에 대한 역량이 미흡할 뿐만 아니라 생산관리 관점에서의 기술개발을 위한 인프라 구축에도 한계가 있다. 또한 계획과 관리능력 향상을 위한 기술개발은 인력 부족 및 자본 등의 문제로 실현되지 못하고 있으며 분석에 적용하기 위한 데이터의 수가 대형 조선소에 적어 계획/실적 동기화에 대한 대처가 어려운 실정이다(우종훈 등, 2015).

하지만 대형 조선소에서도 생산관리 관점에서의 문제점을 피할 수는 없었다. 세계적으로 해양플랜트의 발주가 증가함에 따라 대형 조선소를 위주로 해양플랜트 산업이 활성화되면서 이는 대표적인 고부가가치 선박으로 자리를 잡게 되었다. 해양플랜트는 매우 복잡한 구조와 기능으로 인해 일반 상선에 비해 높은 단가와 납기가 긴 특징으로 가지는데, 이로 인해 건조과정에서의 다양한 문제점이 발생하게 된다. 대표적으로 의장품을 설치하는 과정에서 입고지연에 따른 추가 시수 급증이 경쟁력 상실의 큰 요인으로 자리매김하고 있어 이를 해결하기 위한 다양한 시도가 진행 중이다. 또한 ICT 융합기술을 적용한 기술 경쟁력 향상을 위해 대형 조선소를 시작으로 최적의 생산계획을 수립하여 품질향상, 납기준수, 원가절감을 할 수 있는 다양한 빅데이터 기술을 개발 중이다.

1.1.2 빅데이터 방법론

빅데이터란 일반적인 데이터베이스, 소프트웨어로 관리하기 어려운 정도의 큰 규모로 기존의 방법이나 도구로 수집, 저장, 분석 등이 어려운 정형 또는 비정형 데이터를 모두 포함하는 대용량 데이터를 의미한다. 빅데이터의 핵심적인 특징은 3V로 요약하는 것이 일반적이다. 데이터의 양(Volume), 데이터 생성 속도(Velocity), 형태의 다양성(Variety)을 의미하며 최근에는 가치(Value)나 복잡성(Complexity)이 추가되기도 한다. 따라서 빅데이터는 단순히 대용량 데이터만을 의미하는 것이 아니라 데이터의 수집, 저장, 분석, 체계화를 위한 도구, 플랫폼, 분석기법 등을 포괄하는 용어로 변화하고 있으며, 대용량 데이터를 활용/분석하여 가치 있는 정보를 추출하고 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술을 말한다(강만모 등, 2012).

Table 1 Characteristics of Big Data environment

구분	기존	빅데이터 환경
데이터	- 정형화된 수치자료 중심	- 비정형의 다양한 데이터 - 문자 데이터(SMS, 검색어) - 영상 데이터(CCTV, 동영상) - 위치 데이터
하드웨어	- 고가의 저장장치 - 데이터베이스 - 데이터웨어하우스	- 클라우드 컴퓨팅 등 비용효율적인 장비 활용 가능
소프트웨어/분석 방법	- 관계형 데이터베이스 - 통계패키지(SAS, SPSS) - data mining - machine learning - knowledge discovery	- 오픈소스 형태의 무료 소프트웨어 - Hadoop, NoSQL - 오픈 소스 통계솔루션(R) - 텍스트 마이닝(text mining) - 온라인 버즈 분석(opinion mining) - 감성 분석(sentiment analysis)

빅데이터 환경에서는 기존의 정형화된 수치자료 중심의 데이터에서 벗어나 문자, 영상 이외에 비정형의 다양한 데이터를 분석할 수 있도록 여러 소프트웨어 환경이 만들어졌다(Table 1). 이를 위해 빅데이터 플랫폼에서는 데이터를 수집, 저장, 처리 및 관리할 수 있으며 빅데이터를 분석하거나 활용하는 데 필요한 필수 인프라(Infrastructure)라고 할 수 있다. 가장 대표적인 하둡(Hadoop)은 오픈소스 분산처리기술로 많은 양의 데이터를 저장 및 처리할 수 있는 솔루션이고, 오픈소스 R은 통계 기법부터 다양한 기계학습 모델링까지 적용할 수 있는 데이터 분석 솔루션이다. 일반적으로 하둡은 대용량의 데이터를 여러 서버에 분산적으로 처리하기 위해 기업에서 적용한 사례가 많으므로 본 논문에서는 R이나 Python 언어 등의 오픈소스를 활용한 조선소 데이터 분석에 초점을 두어 연구를 수행하였다(Fig. 2). 빅데이터 분석을 위해서는 데이터 저장 및 통계뿐만 아니라 다양한 알고리즘을 활용한 데이터 분석을 기반으로 예측이 갖는 의미를 찾아내는 것이 중요하다.

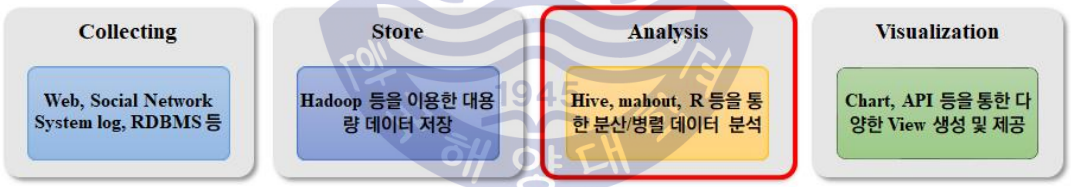


Fig. 2 Technology of Big Data modeling

많은 양의 데이터를 수집하여 빅데이터 인프라에서 적용하기 위해서는 분석 절차가 체계적으로 정리되어야 한다. 빅데이터 분석 단계는 크게 데이터 이해, 분석, 고급 분석에 따라 진행되는데 이는 일반적으로 가장 기본적인 데이터마이닝 기법을 따른다. 데이터마이닝(Data Mining)은 많은 데이터 가운데 과거에 알지 못했던 유용한 상관관계를 발견하여 미래에 실행 가능한 정보를 추출해 내고 이를 의사 결정에 이용하는 과정을 말한다. 본 논문에서의 데이터 분석 과정은 데이터 마이닝의 가장 핵심 프로세스인 KDD(Knowledge Discovery in Databases)에 따라 진행되었다. 이는 복잡하고 많은 양의 데이터로부터

사용자가 원하는 유용하고 가치 있는 지식을 찾아내는 과정으로 넓은 의미로 데이터 분석의 전반적인 과정이라고 할 수 있다(Maimon & Rokach, 2011). KDD는 Fig. 3과 같이 대규모의 데이터로부터 분석을 수행하기 위해 데이터를 추출(Extraction)하여 전처리(Pre-processing)와 변환과정(Transformation)을 거쳐 분석(Data Mining)하고 결과를 해석하는 일련의 프로세스로 구성되어 있다.

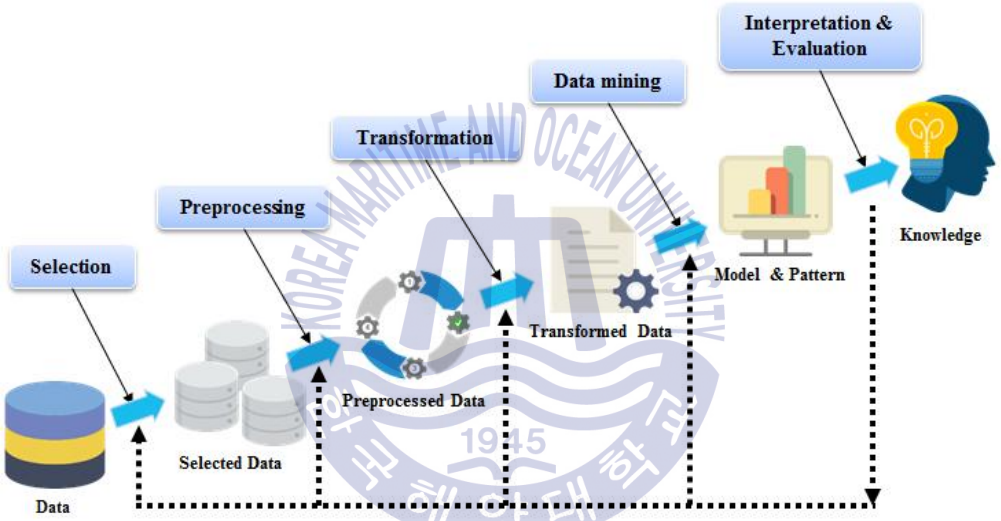


Fig. 3 The process of KDD(Knowledge Discovery in Databases)

데이터마이닝은 데이터 분석 과정의 핵심요소이며 분석을 위한 데이터를 전처리, 변환하는 과정이나 결과를 해석하고 평가하는 것은 넓은 의미로는 데이터 분석에 해당하기 때문에 이런 관점에서 데이터마이닝은 KDD의 구성요소라기보다는 KDD의 전 과정을 포괄하는 개념이라고 할 수 있다.

1.2 관련 연구 동향

1.2.1 제조업 빅데이터 연구 사례

최근 제조 산업에서의 빅데이터는 기존의 분석 대상이었던 조업 데이터뿐만 아니라 ERP, MES, FDC 등 다양한 공정 데이터, 유동에서 확보할 수 있는 POS 데이터, AS 서비스에서 확보할 수 있는 필드 클레임 데이터 등 다양한 형태로 나타나고 있다.

장영재(2012)에서는 제조업에서 활용이 가능한 빅데이터의 범주를 크게 제조 장비 데이터, 운영 통합 데이터, 고객 경험 데이터로 구별하고 있다. 빅데이터의 제조 활용은 가장 많이 언급되고 있는 분야로 반도체나 LCD와 같이 자동화 장비로 구성된 제조 산업에서 많이 활용되고 있다. 운영 통합 데이터는 POS 데이터 형태의 세일즈 및 마케팅과 관련된 정형 데이터로써 기존 제조데이터와 기업 내 다양한 데이터의 통합을 통해 새로운 가치를 창출하고자 한다. 특히 시시각각 변화하는 고객의 요구에 대응하기 위한 전사적인 데이터 통합을 통해 더 나은 의사결정을 내릴 수 있도록 빅데이터 기술을 통한 실시간 의사결정을 지원하고 있다. 마지막으로 고객 경험 데이터는 제품의 사용 후기 혹은 실시간 유입되는 제품 사용정보의 비정형과 정형 데이터를 모두 말한다. 텍스트 마이닝 기술의 발달로 애프터서비스와 관련된 고객의 의견을 분석하여 재품개발에 다양한 아이디어를 제공함으로써 품질 개선과 제품 생산 비용의 절감을 할 수 있다.

조성준과 강석호(2016)에서는 제조업에서 적용될 수 있는 빅데이터를 제품 개발, 제조 공정, 영업 및 마케팅, AS 서비스로 구분하여 설명하고 실제 사례를 제시하였다. 대표적으로 제품 개발 영역에서 자동차 부품 설계 프로세스의 효율성을 향상시키기 위한 텍스트 마이닝 기법을 적용한 사례가 있다. 이는 설계 검증 데이터에서 텍스트 속에 내재된 유의미한 정보를 추출하여 문제를 분석함으로써 기존 조회 시스템에 대비하여 주요 문제에 대한 파악에 소요되는 인력 및 시간을 감소하여 설계 과정의 효율성을 개선하고자 하였다.

정세훈과 심춘보(2014)에서는 용접의 빅데이터 분석 및 추출을 통하여 용접사의 숙련된 패턴을 분석하여 용접 작업에 적용되는 비용을 절감하고자 하였다. 이를 위해 다량의 패턴 변수에 R의 알고리즘과 회귀분석을 적용하여 전력소비량과 와이어 소모 길이에 대한 패턴 구조를 확인하였다.

1.2.2 조선업 빅데이터 연구 사례

본 논문의 선행연구인 함동균(2016)에서는 의장품 중 후행작업에서 많은 지연이 야기되는 배관재의 제작 공정부터 설치 공정까지의 리드타임을 예측하여 조달관리의 수준을 높이기 위한 연구를 수행하였다. 해당 연구에서는 배관공정의 공급망을 6개의 공정으로 나누어 리드타임을 정의하였으며 이를 예측하기 위하여 SPSS를 활용한 다중선형회귀분석과 PLS 회귀분석을 수행하였다. 하지만 해당 연구에서는 공정별 리드타임의 오차율이 크게 나타났으며 단순한 전처리를 통해 데이터의 노이즈를 줄이는 데에 한계점이 있었다.

Hur, at al.(2015)에서는 조선소에서의 공수를 예측하기 위하여 선박 설계 및 생산 과정에서 공수와 관련된 데이터만을 분석하였다. 선박블록 및 공수와 관련된 변수를 정의하고 다중선형회귀분석과 의사결정나무를 활용하여 예측모델을 생성하였다. 해당 연구에서는 공수예측의 정확성을 위하여 분기별, 월별, 일별로 구분하여 예측모델을 생성하였고 그 결과 모델 측면에서는 의사결정나무, 기간 측면에서는 일별에서 가장 설명력 있는 예측모델을 제시하였다. 하지만 조선소에서의 선박 건조기간이 평균적으로 1~2년이라는 점을 감안할 때 장기간의 데이터를 수집하는 데에 한계를 보였고, 작업장 이외에 외적 요인들의 고려가 부족하다고 판단된다.

Lee, at al.(2014)에서는 해양 구조물 제조 과정에서 발생하는 다양한 종류의 불량을 사전에 점검하기 위하여 텍스트 마이닝 기법을 적용하였다. 텍스트 로그데이터를 통해 불량 추이 분석, 연관 불량 분석 등을 파악하여 시각화함으로써 제조 과정에 도움이 되는 유의미한 지식들을 추출하여 활용하고자 하였다.

National Information Society Agency(NIA) (2016)에서 발표한 빅데이터 시범사업 중 조선소의 제조 프로세스 분석을 위하여 빅데이터 클라우드 서비스를 개발한 사례가 있다. 조선업의 특성상 대규모 프로젝트들이 동시다발적으로 진행되기 때문에 공정 현황 및 지연 원인을 파악하는 데에 어려움이 따른다. 이를 해결하기 위해 프로세스 마이닝을 기반에 둔 제조 공정 빅데이터 분석을 통해 공정 지연 및 부하를 분석하여 업무 효율을 향상시키고자 하였다.

김성훈 등(2016)에서는 대표적인 빅데이터 기술인 하둡(Hadoop)을 적용한 빅데이터 플랫폼을 제시하고 이를 이용하여 해양 플랜트 상부의 중량 추정 과정에 적용해봄으로써 조선 해양 분야에서의 빅데이터 적용 가능성에 대해 연구하였다. 하둡(Haddop) 기반의 빅데이터 플랫폼을 제시한 것에는 의의가 있지만 실제 서버에 적용하는 데에는 어려움을 보였으며 데이터의 수도 부족하여 단순한 선형회귀분석만 적용한 한계점을 보였다.

현재까지 빅데이터 적용 사례는 다양한 제조업 및 산업분야에서는 늘어나고 있는 실정이지만 조선업에서는 실질적으로 적용된 사례가 드문 것을 확인하였다. 다품종 소량생산인 조선업의 특성으로 인해 데이터의 수는 계속해서 쌓여감에도 불구하고 체계적인 빅데이터 플랫폼 구축이 어려운 점이 큰 이유이며 건조 활동의 근간이 되는 기준정보, 특히 원단위의 시수, 리드타임 등 시간 정보 체계가 미흡하기 때문에 최신 ICT 기술 및 빅데이터 기술 적용에 어려움을 겪고 있다고 판단된다.

제 2 장 분석 알고리즘 및 적용방안

2.1 기계학습 알고리즘

기계학습 또는 머신러닝(Machine Learning)은 분석하고자 하는 데이터로부터 지속적인 학습을 통해 만들어진 새로운 데이터의 작업으로, 해결하고자 하는 문제에 대한 답을 얻어내는 방법이다. 기계학습은 통계(Statistics), 컴퓨터 과학(Computer Science), 데이터 마이닝(Data Mining) 등 여러 분야와 밀접한 관련이 있으며 문자 인식, 쇼핑몰 추천 시스템, 스팸 메일 필터링 등과 같은 다양한 사례로 이미 일반 소비자들에게도 낯선 기술이 아니다. 즉, 다양한 확률, 조합 이론, 수학적 최적화 기법, 통계, 알고리즘, 컴퓨터 구조를 활용해 이상적인 학습모델을 구축하는 기술로 연구자의 경험적 지식 습득과 그 응용 방법까지 포함하는 포괄적인 융합 기술이라고 할 수 있다(이재구 등, 2014). 기계학습은 일반적인 프로그래밍과는 다르게 입력값과 출력값을 동시에 고려하여 하나의 모델을 구축하는 것으로 컴퓨터에 명시적으로 프로그래밍하지 않고 학습할 수 있는 능력을 부여하는 컴퓨터 과학이라고 할 수 있다.

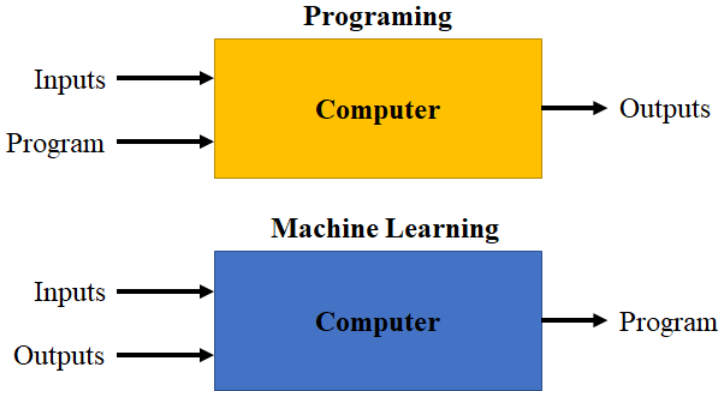


Fig. 4 Machine Learning

기계학습은 데이터에 대한 모델을 만드는 것이 주요 목적이기 때문에 입력 데이터에 대한 고려뿐만 아니라 분석 대상에 맞는 알고리즘을 선택하는 것도 중요하다. 기계학습 알고리즘은 학습 데이터 혹은 훈련 데이터(training data)를 어떻게 이용하느냐에 따라서 즉, 훈련 데이터에 레이블이 있는 경우와 그렇지 않은 경우에 따라 각각 지도학습(supervised learning)과 비지도 학습(unsupervised learning)으로 분류된다(Fig. 5). 레이블이라는 것은 학습 데이터의 속성을 우리가 분석하고자 하는 관점에서 정의하는 것이다. 지도학습은 미리 준비된 훈련 데이터에 알고리즘을 적용하여 도출된 예측 및 추론을 통해 컴퓨터가 스스로 답을 찾도록 한다. 예를 들어 사진을 주고 '이 사진은 고양이'라고 알려주는 방식으로 컴퓨터는 미리 학습된 결과를 바탕으로 고양이 사진을 구분할 수 있다. 따라서 지도학습은 명확한 입력값과 출력값이 존재하며 이러한 지도학습에는 크게 분류(classification)와 회귀(regression)가 있다. 반면 비지도 학습은 지도학습과는 다르게 데이터에 대한 정답을 알지 못한 채 컴퓨터가 알아서 분류를 하고 의미 있는 값을 찾는 것으로 배움의 과정이 없다. '이 사진이 고양이'라는 배움의 과정 없이 컴퓨터가 스스로 학습해야 하기 때문에 컴퓨터의 높은 연산능력이 요구된다. 이러한 비지도학습의 대표적인 모델로 군집(clustering)모델이 있다. 이 외에도 학습 수행 결과에 대해 적절한 보상을 주면서 피드백을 통해 학습하는 강화 학습(reinforcement learning)이 있다.

본 논문의 목적은 조선소에서 리드타임이라는 기준정보를 개선하기 위한 예측모델을 만드는 것이다. 즉, 예측대상인 '시간'이 출력값으로 정의되며 예측모델을 생성하기 위해 조선소 공정정보가 입력값으로 정의된다. 따라서 본 논문에서의 기계학습 알고리즘은 비지도 학습이 아닌 지도학습 알고리즘을 적용하는 것이 타당하며 그 중 수치예측에 대표적인 회귀분석, 인공신경망, 의사결정나무를 활용한 예측모델을 만들고자 한다.

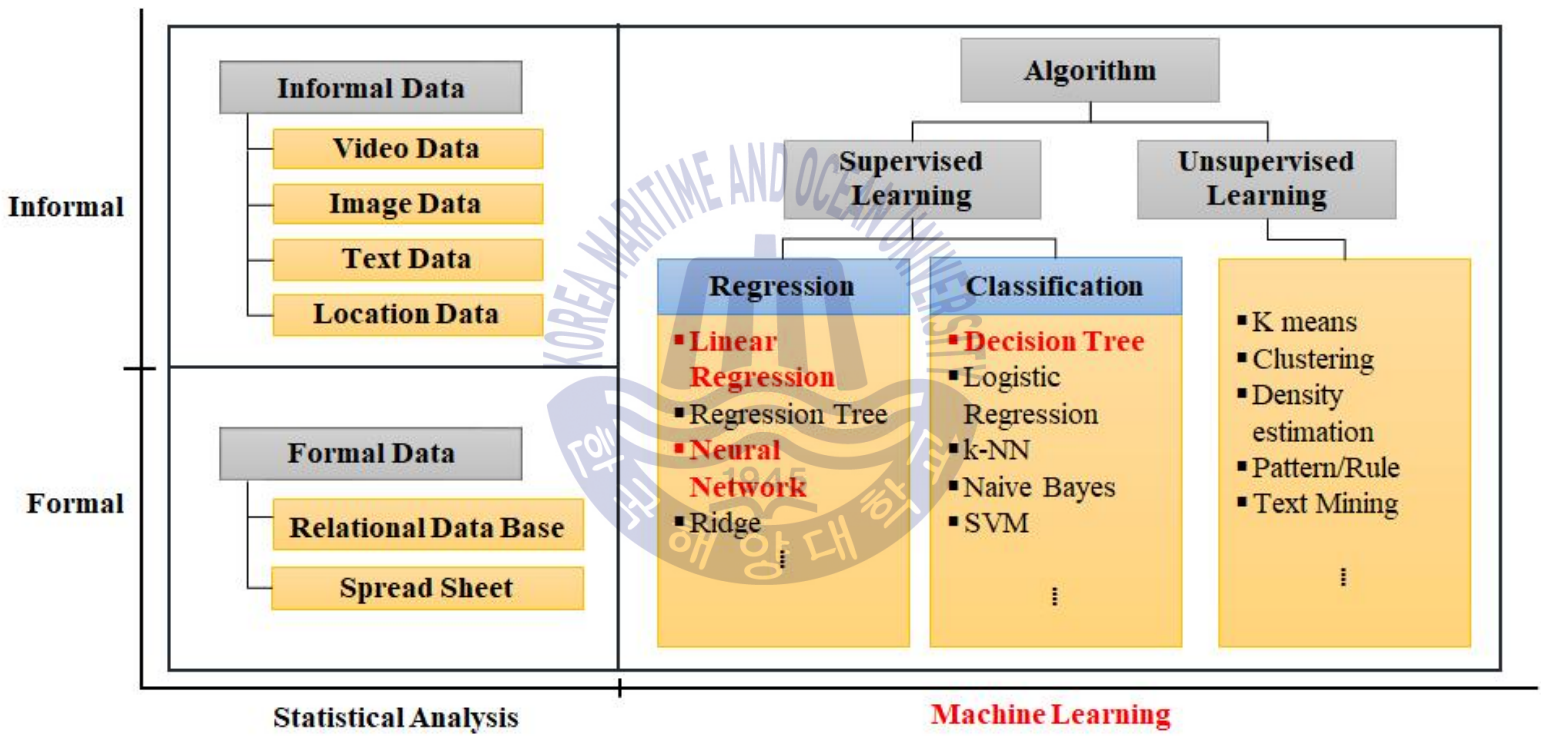


Fig. 5 Algorithm type of machine learning

2.1.1 회귀분석

수치예측 알고리즘에서 가장 기본적으로 사용되는 것이 회귀분석이다. 최근 커뮤니케이션 과학에서 가장 많이 사용되는 분석기법으로 예측변인이 변화함에 따라 결과값이 얼마나 변화하는지를 예측할 수 있다.

회귀분석은 변수와 변수 사이의 관계를 알아보기 위한 통계적 분석방법으로 독립변수의 값에 따라 종속변수의 값을 예측하기 위해 활용되는 기법이다. 여기서 독립변수(Independent Variable)는 종속변수에 영향을 미치는 변수를 의미하고 종속변수(Dependent Variable)는 분석의 대상이 되는 변수로 독립변수에 따라 값이 변하는 변수를 의미한다. 회귀분석은 독립변수의 수에 따라 단순회귀분석과 다중회귀분석으로 분류되는데 만약 하나의 종속변수와 하나의 독립변수 사이의 관계를 분석하고자 한다면 단순회귀분석을 적용하고, 하나의 종속변수와 두 개 이상의 독립변수 사이의 관계를 분석하고자 한다면 다중회귀분석을 적용하게 된다. 특히 다중 공선성의 문제를 해결하기 위해서는 사전에 데이터 전처리 과정에서 상관계수가 높은 변수의 유무를 확인하여 차원 축소를 하는 것이 바람직하다.

2.1.2 인공신경망

기계학습 알고리즘 중 인공신경망(artificial neural network, ANN) 생물학적으로 뇌가 감각 입력의 자극에 어떻게 반응하는지에 대한 이해로부터 얻어진 모델을 사용해서 입력 신호와 출력 신호 간의 관계를 모델화하는 알고리즘이다(Fig. 6). 마치 뇌가 막대한 병렬 프로세서를 생성하기 위해 뉴런이라는 세포로 연결된 망을 사용하듯이 인공신경망은 학습 문제를 풀기 위해 인공 뉴런이나 노드의 망을 사용한다.

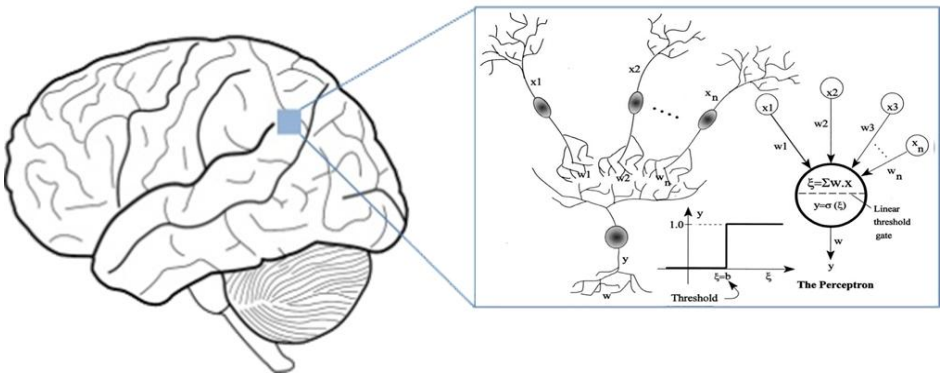


Fig. 6 Artificial neural network

일반적으로 피드포워드 신경망을 이용한 다층 퍼셉트론으로 입력층, 은닉층, 출력층의 3가지 층으로 구성되어 있는 다층 퍼셉트론을 의미하기도 한다. 입력층은 원본데이터 필드에서 -1과 1사이의 값으로 맵핑된 값을 읽어온 뒤 출력변수로 출력하는 기능을 한다. 은닉층은 입력층에서 전달되는 출력값과 가중치를 입력값으로 취하고 마지막 층인 출력층으로 출력값과 가중치를 전달하는 층으로 입력데이터와 최종 예측값 모두와 직접적으로 연결되어 있지 않기 때문에 은닉층으로 불린다. 일반적으로 하나의 출력값을 가지나 추정문제의 경우 복수개의 출력값 가지는 인공신경망을 사용하게 된다. 각각의 층(Layer)들은 앞서 설명한 결합함수를 가진 노드로 이루어져 있고 은닉층의 경우 일반적으로 2개에서 8개의 노드수를 가지며 은닉층 수가 2개 이상일 경우를 MLP(Multi Layer Perceptron)라 하며 많은 노드 수와 복수개의 은닉층이 더 정확한 결과를 낼 수 도 있지만 과적합(Overfitting)의 위험을 가지고 있으므로 적절하게 은닉층의 수를 정의하는 것이 중요하다(함동균 2016).

2.1.3 의사결정나무

의사결정나무(decision tree)는 귀납적 추론을 기반으로 하는 알고리즘으로 실무적으로 가장 많이 사용되고 있는 지도학습 모델 중 하나이다. 의사결정나무의 경우 일반적으로 분류문제를 적용하기 위해 사용되지만 연속형

변수를 추정하기 위해서도 적용이 가능한 알고리즘이며 특히 선형식이 아닌 불연속 데이터도 활용할 수 있다. 분석과정이 나무구조에 의해서 표현되기 때문에 판별 분석(Discriminant Analysis), 회귀분석(Regression Analysis), 신경망(Neural Networks) 등과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다.

2.2 딥러닝 알고리즘

머신러닝의 연구가 다양한 산업분야에서 활용되고 있다. 90년대에 들어서면서 컴퓨터의 학습 방법론에 중점을 뒀던 기존의 접근법보다는 실생활에서 필요한 문제를 해결할 수 있는 실용적인 머신러닝 연구가 주류를 이뤘다. 이는 2000년대 초까지 주류를 이뤄왔으며 앞서 언급했듯이 통계학 기반의 분류 및 회귀모델 등으로 분류된다. 기존의 머신러닝은 컴퓨터가 주어진 훈련 데이터로부터 특징을 추출하는 과정에서 사람이 사전에 정의하거나 개입하는 부분이 많으므로 이러한 과정에서 발생하는 오류가 존재하기 마련이었다. 또한, 다수의 신경층을 이용하는 접근은 비선형 문제, 신경층의 수에 따른 가중치 수의 한계, 과적합 등의 다양한 문제점으로 활용되지 않았다. 그러나 이러한 문제점들을 컴퓨터의 계산 성능, 데이터의 증가와 알고리즘의 발달로 인해 접근의 효용성이 밝혀지게 되면서 딥러닝 기술은 현재 인공지능 분야에서 필수적인 요소로 자리매김하였다(김의중, 2016).

Fig. 7 Conceptual diagram of AI, Machine Learning, Deep Learning

딥러닝(deep learning)은 패턴 인식 문제 또는 특징점 학습을 위해 많은 수의 신경층을 가지도록 모델을 구성하는 기계학습 기술들을 의미한다(문성은 등, 2016). 특히 딥러닝의 가장 큰 장점은 컴퓨터가 스스로 학습이 가능하다는 것이다. 기존의 머신러닝은 훈련 데이터의 특징 및 패턴 추출을 위해서 사람이 개입하였지만 딥러닝은 이러한 것들을 컴퓨터가 스스로 학습할 수 있도록 하기 때문에 보다 진보된 인공지능 기술이라 할 수 있다. 이는 신경망의 깊이가 매우 깊은(deep) 구조를 가지고 있고 각 층마다 고려되는 변수가 많아졌기 때문에 가능한 일이다. 딥러닝은 기존의 머신러닝에 비해 개선된 신경망을 가지고 있으며 과적합 문제를 해결하기 위한 알고리즘이라고 할 수 있다.

하지만 딥러닝 신경망의 깊이가 깊어질수록 연산에 필요한 복잡도 및 계산량도 비례적으로 많아짐에도 불구하고 신경망의 깊이를 늘릴 수밖에 없는 이유는 앞에서 설명한 머신러닝의 한계점이 가장 크다고 할 수 있다. 따라서 딥러닝에서는 신경망의 깊이를 결정하는 것이 중요하며 이를 위해서는 많은 양의 데이터 및 우수한 컴퓨터 성능이 필요하다. 인공지능 분야의 핵심기술인 딥러닝이 급속도로 발전하게 되면서 이를 조선업에서도 적용한다면 보다 예측도 높고 체계적인 생산 기준정보를 수립할 수 있을 뿐만 아니라 다양한 패턴 분석을 통해 생산 시뮬레이션을 위한 데이터로도 활용할 수 있을 것이라 기대하는 바이다. 따라서 본 논문에서는 수치예측에 많이 활용되는 딥러닝 알고리즘 중 가장 기본적인 다층 퍼셉트론과 순환 신경망을 적용하여 예측모델을 생성하고자 한다.

2.2.1 Multi-Layer Perceptron (MLP)

다층 퍼셉트론(Multi-Layer Perceptron)을 이해하기 위해서는 먼저 퍼셉트론이 무엇인지를 이해해야 한다. 퍼셉트론은 인간의 신경세포인 뉴런을 매우 단순히 모사하여 계산 가능한 형태로 만든 알고리즘으로 인공신경망에서 활용된 구조를 말한다. 기본적인 퍼셉트론은 데이터의 입력층과 출력층만 있는 구조로 단층 퍼셉트론이라고도 하는데 이는 활성 함수가 1개 밖에 없는 구조이기 때문에 비선형적으로 분리되는 데이터에 대해서는 제대로 된 학습이 불가능하다는 문제점이 있다. 따라서 이를 극복하기 위한 방안으로 입력층과 출력층 사이에 하나 이상의 중간층을 둔 다층 퍼셉트론이 고안되었다.

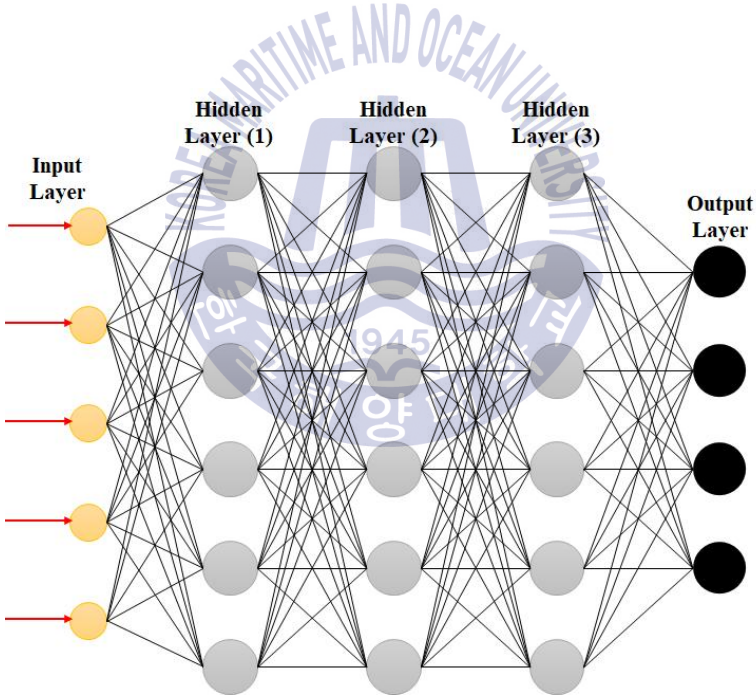


Fig. 8 Conceptual diagram of Multi-Layer Perceptron

Fig. 8과 같이 다층 퍼셉트론은 입력층과 출력층 사이에 은닉층(Hiddne Layer)라는 하나 이상의 중간층을 가지는 계층구조로 각 층 내의 연결과

출력층에서 입력층으로의 직접적인 연결은 존재하지 않은 전방향(Feedforward) 네트워크이다. 이를 바탕으로 입력층과 출력층 사이에 여러 개의 은닉층이 있는 인공 신경망을 심층 신경망(Deep Neural Network)이라 부르며, 심층 신경망을 학습하기 위해 고안된 특별한 알고리즘들을 딥러닝(Deep Learning)이라 부른다.

2.2.2 Recurrent Neural Network (RNN)

순환 신경망은 시계열 데이터에서 탁월한 성능을 보여주는 알고리즘으로 매 순간마다 인공신경망 구조를 쌓아올린 것이다. 쉽게 말해 현재 들어온 입력 데이터와 과거에 입력 받았던 데이터를 동시에 고려한다는 뜻이다. 일반적인 인공신경망은 입력 데이터가 들어오면 입력층부터 출력층까지 차례대로 연산을 진행하는데, 이 때 입력 데이터는 모든 노드를 한 번만 지나가게 된다. 즉, 시간에 대한 흐름을 무시하고 주어진 데이터만을 의존하여 학습을 수행하게 된다. 하지만 순환 신경망은 과거의 출력값이 은닉층에서 새로운 입력 데이터로 활성화되면서 기억 능력을 가지게 된다. 이는 일반적인 인공신경망과 가장 큰 차이점으로 마치 사람이 과거의 기억에 의존하여 판단하는 과정과 유사하기 때문에 비슷한 인공신경망에 비해 더 적은 수의 노드만으로도 복잡한 데이터를 모델링할 수 있게 한다. 따라서 순환 신경망 알고리즘을 활용하여 배관재 공급망 프로세스의 시계열 데이터를 분석한다면 각 공정의 세부적인 작업시간에 영향을 미치는 요소들을 파악하여 전제적인 공급망 리드타임을 예측할 수 있을 것이다.

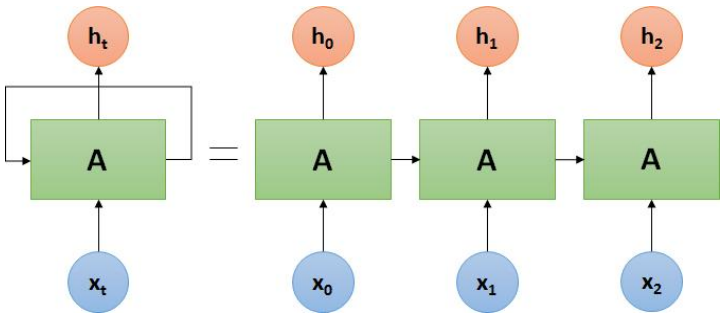


Fig. 9 Recurrent neural network

2.2.3 딥러닝 라이브러리

본 논문에서는 딥러닝 알고리즘을 구현하기 위한 라이브러리로 Python 언어 기반의 'Keras'를 활용하였다. Keras는 딥러닝 모델을 위해 직관적인 API를 제공하고 있으며 내부적으로는 텐서플로우(Tensorflow), 티아노(Theano), CNTK 등의 딥러닝 전용 엔진이 구동되지만 사용자가 쉽게 접근할 수 있는 구조로 이루어져 있다. 케라스는 모듈화, 최소주의, 쉬운 확장성, Python 기반이라는 4가지 특징을 가지고 있다. 케라스에서 제공하는 모듈은 독립적으로 설정이 가능하며 최소한의 제약사항으로 서로 연결되어 있고, 각 모듈은 짧고 간결하여 사용하기 쉽다는 장점이 있다. 또한 새로운 클래스나 함수로 쉬운 모듈을 추가할 수 있으며 고급 연구에 필요한 다양한 표현이 가능하다.

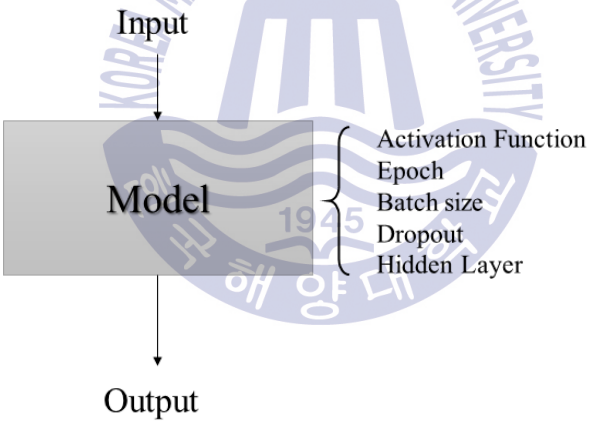


Fig. 10 Structure of Keras library

케라스에서는 Fig. 10에서처럼 다양한 파라미터를 설정하여 딥러닝 모델을 구축할 수 있다. 모델을 학습하기 위해 가장 기본적으로 정의하는 파라미터는 Batch size와 Epoch가 있다. Batch size는 모델을 학습할 때 몇 개의 샘플로 가중치를 갱신할 것인지를 지정하는 값이고, Epoch는 모델의 학습 반복 횟수를 의미한다. 이 외에 모델을 학습하기 위한 활성화 함수를 지정하는 Activation Function과 과적합을 해결하기 위한 Dropout 설정, 그리고 Hidden layer의 수에

따라 다양한 딥러닝 모델을 구축할 수 있다.

딥러닝 모델을 구축할 때는 가장 중요한 것이 학습과정에서의 기울기 값과 과적합의 문제를 해결하는 것이다. 기본적으로 신경망의 구조에서 레이어의 값이 늘어날수록 역전파 과정에서의 기울기 값이 사라지는 문제가 발생하기 때문에 네트워크의 파라미터를 효과적으로 학습시킬 수 없게 되며 error rate가 낮아지지 못한 채 수렴해버리는 문제가 발생하게 된다. 이를 해결하기 위해 activation function을 ReLU(Rectified Linear Unit)을 선택하였다. 기존에 많이 사용된 Sigmoid 함수는 값을 변형하면서 vanishing gradient 문제가 발생하게 되었는데 Fig. 11에서처럼 ReLU함수를 사용하게 되면 0보다 작을 때는 0을 사용하고 0보다 큰 값에 대해서는 해당 값을 그대로 사용하기 때문에 기존의 문제를 해결할 수 있으며 계단이 단순해져 학습이 빠르다는 장점이 있다.

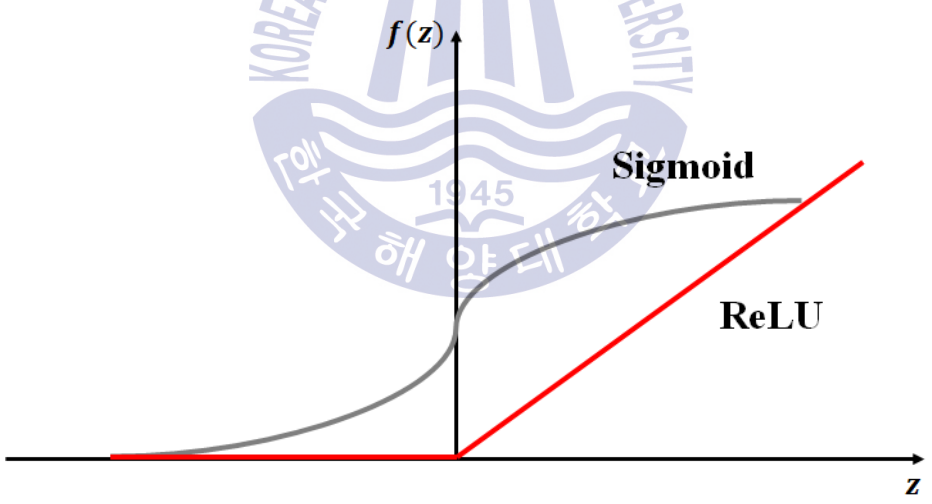


Fig. 11 Neural network activation function (ReLU)

또한, 데이터의 수가 너무 적거나 많을 경우 학습한 데이터에만 최적화되어 학습되지 않은 데이터에 대한 예측 성능이 저하되는 현상이 나타나게 된다. 보통 이러한 과적합 문제는 training data를 많이 모으는 방법도 있지만

regularization을 통해 가중치를 설정하는 방법이 있다. 가장 대표적인 방법이 dropout(Fig. 12)이 있다. dropout은 전체 weight의 값을 학습에 참여시키지 않고 무작위로 hidden layer에 있는 unit을 없애고 학습하는 방법이다. 너무 많은 weight가 오히려 학습을 방해할 수 있으며 일부만 사용해도 균형 잡힌 결과를 도출할 수 있는 것으로 판단되어 딥러닝에서 많이 활용되는 방법이다.

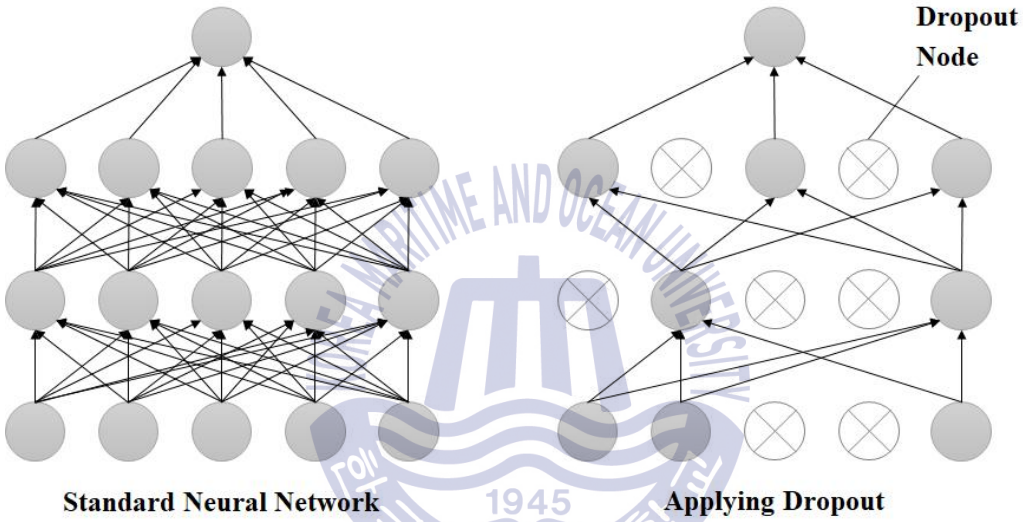


Fig. 12 Dropout (Regularization Method)

이러한 다양한 방법들을 활용하여 딥러닝 라이브러리를 이용한 모델을 구축하고자 하며 다양한 파라미터를 정의한 모델의 평가를 통해 가장 유의미한 예측모델을 갖는 설정값을 찾고자 한다.

2.3 데이터 분석과정

데이터 분석과정은 KDD 이론을 바탕으로 Fig. 13과 같이 크게 데이터 수집, 데이터 처리, 모델 구축, 데이터 평가의 4단계로 정의하였다. 전반적인 과정은 먼저 생산 리드타임을 예측하고자 하는 조선소 데이터를 공정에 따라 수집한 후 모델을 구축하기 위한 다양한 데이터 처리기법을 적용하게 된다. 데이터 처리를 위해서 공정변수 정의 및 데이터 탐색을 통해 수집된 조선소 데이터의 현황을 파악할 수 있다. 다음 단계에서는 예측모델을 생성하기 위한 데이터 분류를 수행하고 다양한 기계학습 및 딥러닝 알고리즘을 적용하여 알고리즘에 따른 예측모델을 생성할 수 있다. 최종적으로는 데이터의 평가를 위해 다양한 평가지표를 활용하여 예측모델의 성능평가를 수행할 수 있고 이를 통해 조선소 데이터에 따른 평가뿐만 아니라 알고리즘에 따른 평가 결과를 통해 조선소 데이터의 분석 과정에서의 의사결정을 지원할 수 있다. 분석과정의 단계별 상세내용은 다음과 같다.

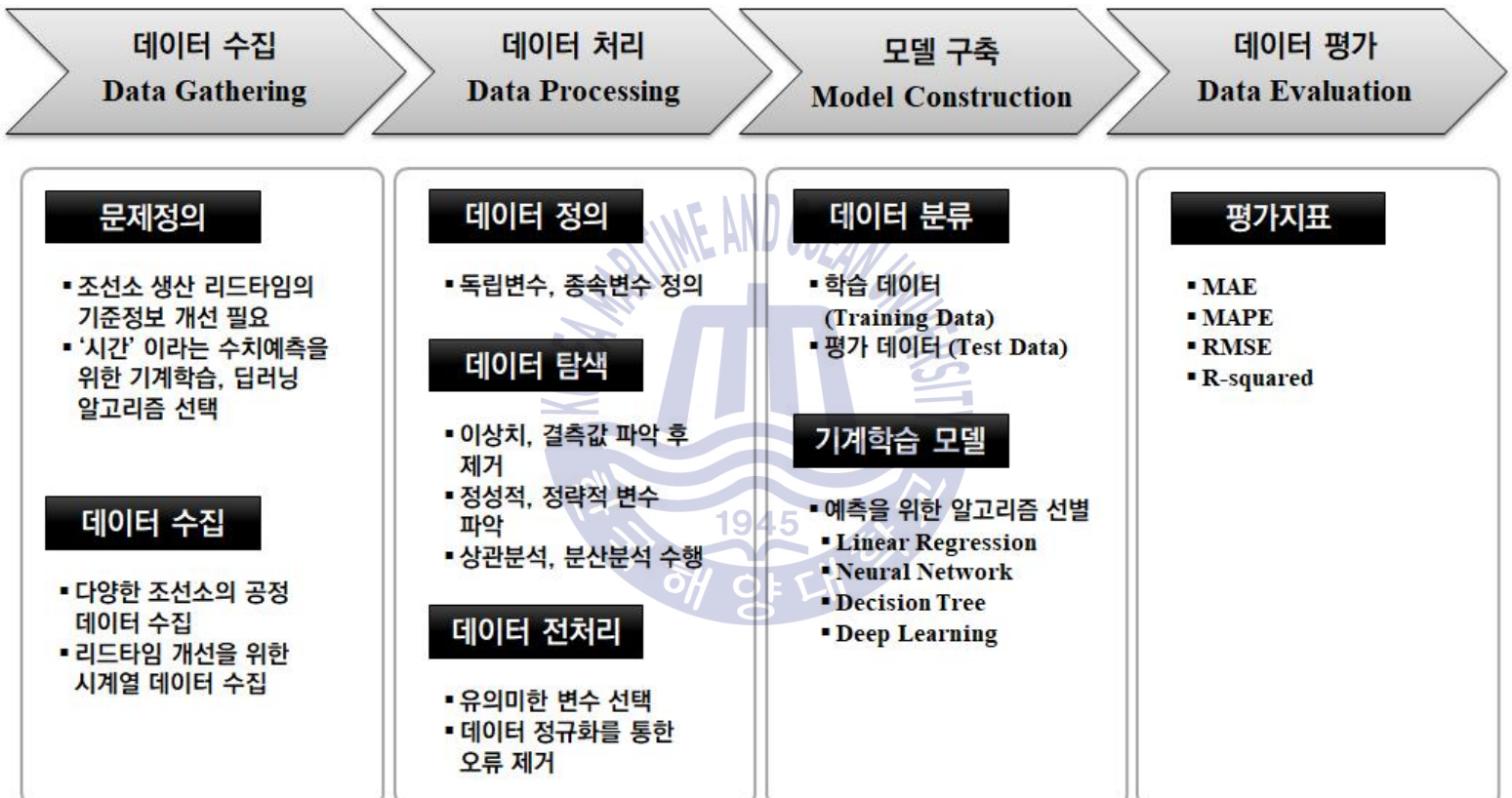


Fig. 13 Data analysis process

2.3.1 데이터 수집

데이터 수집 단계에서는 문제정의 후 이에 맞는 분석목적에 따라 데이터를 수집하는 단계이다. 본 논문에서의 연구목적은 조선소의 생산 리드타임이라는 기준정보를 개선하기 위함이다. 이를 위해서는 '시간'이라는 연속형 변수를 예측하기 위한 알고리즘을 적용하는 것으로 문제정의를 하고 이에 맞는 데이터를 수집하였다. 수집된 조선소의 데이터는 3가지 유형으로 Table 2와 같이 선박 블록 절단, 선박 블록 탑재, 해양플랜트 배관재 공급망 데이터로 구성되며 모든 데이터로부터 예측하고자 하는 대상은 생산 리드타임으로 정의할 수 있다.

Table 2 Collection data

분석 데이터	예측 대상
선박 블록 절단 실적 데이터	블록 절단공정 계획 리드타임
선박 블록 탑재 실적 데이터	블록 탑재 리드타임
해양플랜트 배관재 공급망 데이터	배관재 공급망 리드타임

2.3.2 데이터 처리

수집된 조선소 데이터를 정의, 탐색, 수정, 전처리 등을 수행하는 포괄적인 작업을 데이터 처리 단계로 정의하였다. 데이터를 분석하기 위해서는 알고리즘 적용을 위한 변수 선택부터 오류가 있는 데이터를 처리하는 작업이 필수적으로 수행되어야 한다. 특히 가장 중요한 데이터 전처리(Pre-processing) 작업은 유의미한 변수를 추출하기 위한 다양한 Feature Engineering 방법론을 적용하여 데이터를 정교하게 만드는 과정을 의미한다(부록 A 참고).

먼저 알고리즘 적용을 위한 독립변수와 종속변수를 정의하고 각각의 변수들을 탐색하여 이상치나 결측값이 있는지를 파악하여 제거함으로써 분석결과의 오류를 사전에 방지하고자 한다. 결측값이란 수집된 데이터 중

몇몇 변수들의 값이 측정되지 못한 값을 의미하는 것으로 결측값이 있는 상태로 모델을 만들 경우 변수 간의 관계가 왜곡될 가능성이 있기 때문에 정확성이 떨어지게 된다. 이상치란 수집된 데이터와 동떨어진 관측치로 모델을 왜곡할 가능성이 있는 관측치를 말한다. 일반적으로 결측값이나 이상치가 무작위로 발생한 경우는 단순히 제거하는 것이 가장 간단한 방법이지만 만약 이러한 값들이 임의로 발생한 것이 아닌데 제거할 경우 오히려 왜곡된 모델이 생성될 수 있으므로 데이터의 정확한 분석을 통해 파악하는 것이 중요하다.

기본적으로 이상치나 결측값을 찾기 위한 쉽고 간단한 방법은 변수의 분포를 시각화하는 것이다. 본 논문에서는 Histogram, Boxplot 등의 그래프를 통해 데이터 분포나 이상치를 확인하여 처리하였다. Histogram을 통해서는 데이터의 분포를 확인하여 왜도 등의 데이터 비대칭성을 파악하기 쉬우며 Boxplot을 통해서는 데이터의 이상치를 시각적으로 쉽게 판단할 수 있다.

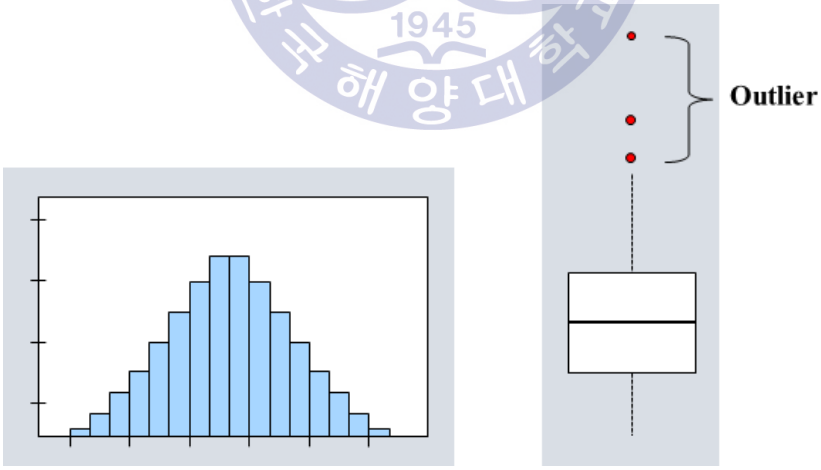


Fig. 14 Histogram and Boxplot

또한 통계분석은 변수의 형태 및 개수에 따라 분석기법을 다르게 적용하는데 일반적으로 독립변수의 수에 따라 단변량(one variable)과 다변량 (Multi variable)으로 나눌 수 있다(Fig. 15). 단변량 데이터는 변수 자체의 분포나 형태에 초점을 맞추어 분석을 한다면 다변량 데이터는 변수와 변수 사이의 관계에 초점을 맞추기 때문에 이를 위한 상관분석을 수행하는 것은 데이터 분석 관점에서 중요한 과정 중 하나라고 할 수 있다. 이를 위해서는 자료의 형태에 따라 알맞은 분석방법을 적용해야하기 때문에 본 논문에서는 연속형 자료에는 상관분석, 범주형 자료에는 분산분석을 활용하고자 한다.

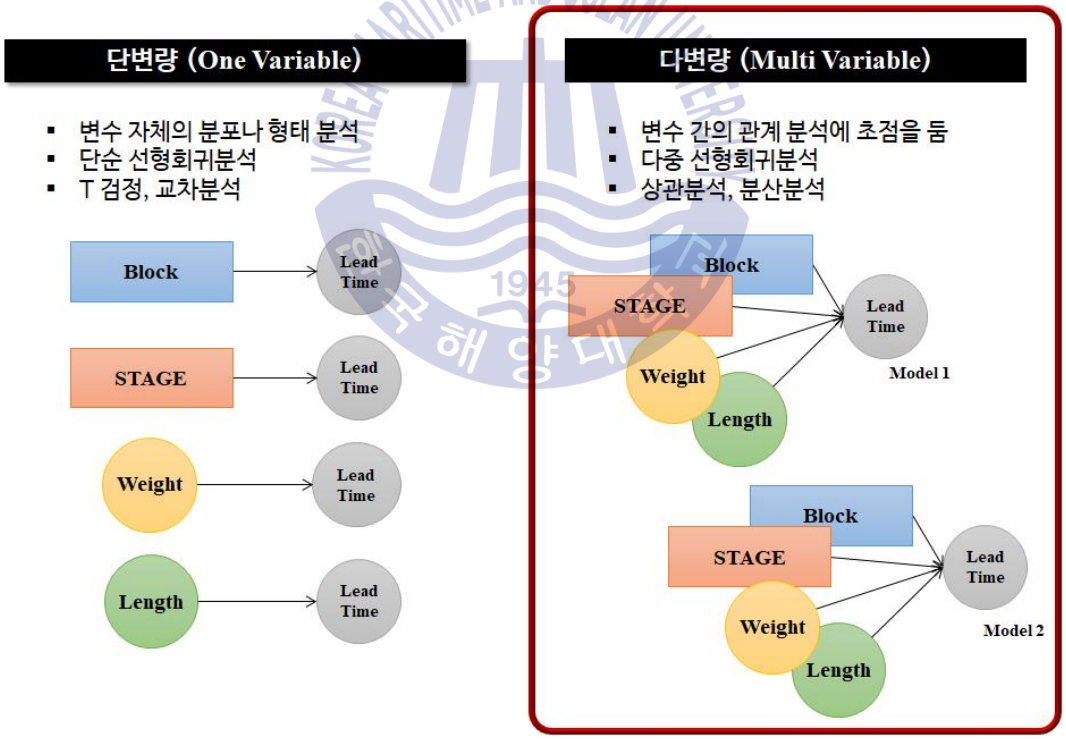


Fig. 15 Analysis method by variable type

상관분석(Correlation Analysis)은 연속형인 두 변수 간에 어떠한 선형적 관계를 가지고 있는지를 분석하는 통계방법으로 두 변수 간의 상관관계 정도를 상관계수(Correlation coefficient)로 파악할 수 있다(Fig. 16). 상관계수는 두 변수 간의 상관성을 판단하는 지표로 만약 상관계수가 1에 가까우면 강한 양의 상관관계, -1에 가까우면 강한 음의 상관관계, 0에 가까우면 상관관계가 매우 낮다고 판단할 수 있다. 상관계수는 변수 간의 상관관계 정도만을 나타낼 뿐 서로 간의 인과관계를 설명하는 것이 아니기 때문에 추후 알고리즘 적용을 위한 척도로써 활용하고자 한다. 본 논문에서는 두 변수간의 상관관계를 구하기 위해 보편적으로 이용되는 피어슨 상관계수(Pearson correlation coefficient)를 통해 연속형 변수 간의 상관관계를 분석하여 독립변수 정의를 위한 의사결정에 활용하고자 한다.

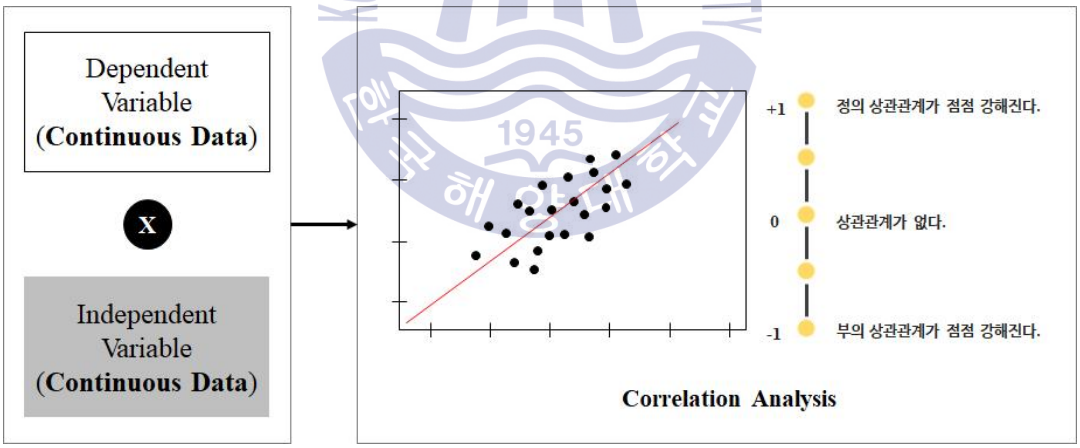


Fig. 16 Correlation analysis

분산분석(ANOVA: Analysis Of Variable)은 3개 이상의 집단 간 평균을 비교하기 위한 분석방법으로 독립변수는 범주형 자료, 종속변수는 연속형 자료임을 가정한다. 기본적으로 분산의 개념을 이용하는 방법으로 분산을 계산할 때처럼 편차의 각각의 제곱합을 해당 자유도로 나누어서 얻게 되는 값을 이용하여 수준평균들 간의 차이가 존재하는 지를 판단하게 되는데 이는 P-value와 유의수준을 비교하여 집단에 따라 관측값의 차이가 존재하는지를 판단할 수 있다. 분산분석은 측정하고자 하는 값에 영향을 미치는 요인(Factor)의 수에 따라 구분하게 되는데 1개인 데이터는 일원분산분석, 2개인 데이터는 이원분산분석, 그 이상인 데이터는 다중분산분석을 활용할 수 있다. 본 논문에서 활용된 범주형 변수와 종속변수인 생산 리드타임 간의 차이가 존재하는지를 R 프로그래밍 state 패키지의 aov() 함수를 활용하여 분산분석을 수행하였다. 분산분석의 결과는 기본적으로 P-value가 유의수준보다 낮을 경우 관측값의 차이가 없다고 판단된다.

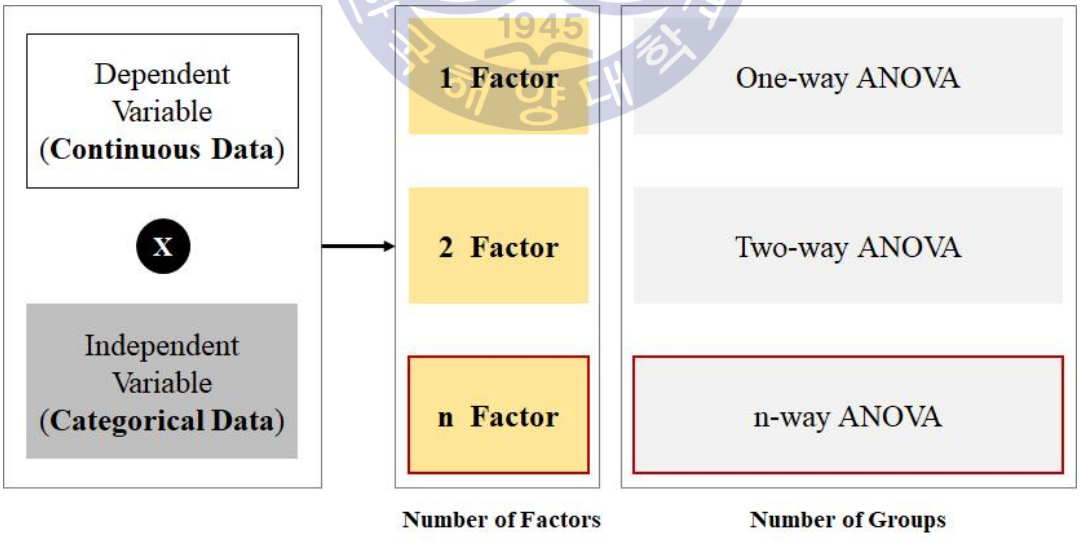


Fig. 17 Analysis of variable

2.3.3. 모델 구축

예측 모델 구축을 위한 학습용 데이터와 평가용 데이터를 분류하고 조선소별 예측모델 생성을 위한 기계학습 및 딥러닝 알고리즘을 적용하는 단계이다. 학습 데이터는 예측모델을 구축하는데 사용되고 평가 데이터는 새로운 데이터에 적용하여 모델의 실제 예측도를 평가하는데 사용되는 데이터로 보통 7:3 정도의 비율에 따라 임의적으로 분류한다(Fig. 18). 수집된 데이터의 전처리는 기본적으로 동일한 방법을 적용하게 되며 예측대상인 공정별 리드타임과의 상관관계를 분석하기 위한 상관분석과 분산분석을 수행하게 된다. 최종적으로 정의된 공정변수를 바탕으로 데이터를 분류하여 알고리즘을 적용하게 되는데 이 때 기계학습 알고리즘은 R 언어를 사용하고 딥러닝 알고리즘은 Python 언어를 사용하여 모델을 구축할 수 있으며 최종적으로 생성된 예측모델을 평가용 데이터에 적용하게 된다.

2.3.4 데이터 평가

마지막으로는 여러 평가지표를 적용한 예측모델의 성능평가를 수행하는 단계이다. 예측모델의 성능 평가지표는 크게 MAE, MAPE, RMSE를 활용하며 회귀분석의 설명력을 분석하는 R^2를 추가적으로 분석하였다. 이러한 평가지표는 예측모델에서 산출된 예측값과 실제 데이터에서 얻은 실적값의 오차율 및 정밀도를 정량적 지표로 산출할 수 있으므로 이를 통해 알고리즘에 따른 조선소 데이터 별로 유효한 알고리즘의 비교가 가능하며 분석 케이스에 따른 예측모델의 성능평가를 가능하게 한다(부록 A 참고).

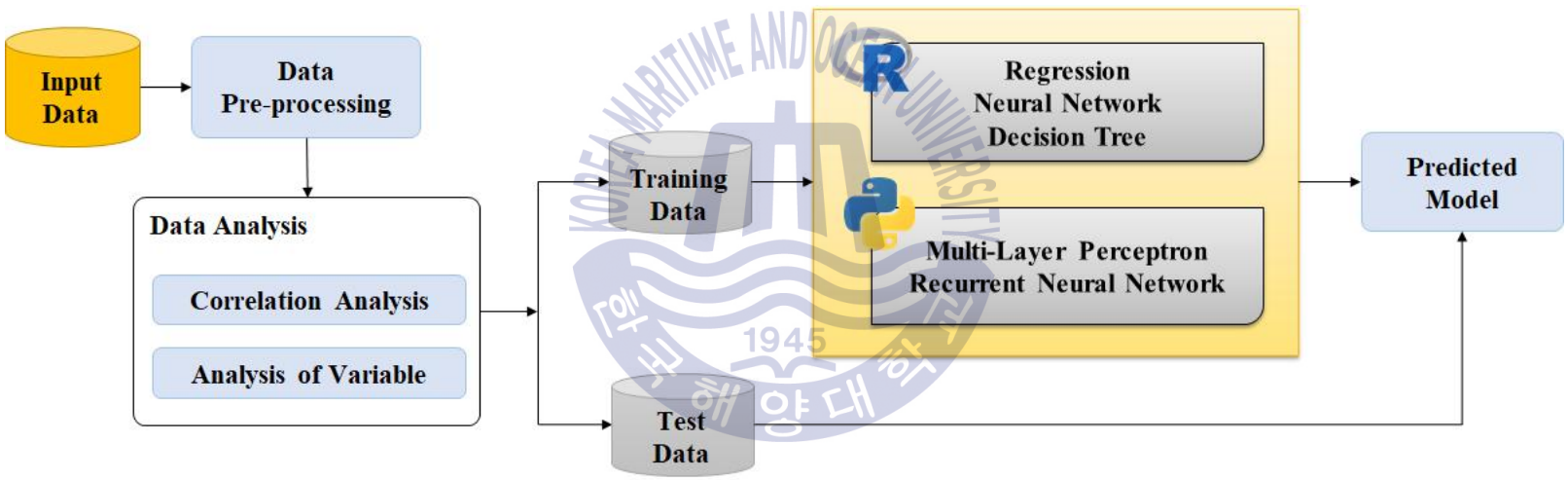


Fig. 18 Process of Data Analysis

제 3 장 예측모델 구축을 위한 조선소 데이터 분석

조선소 생산 리드타임을 예측하기 위해서 본 연구에서는 3개의 조선소 데이터를 공정별로 수집하여 분석을 수행하였다. 본 논문의 선행연구에서는 기본적인 통계분석을 위한 SPSS 프로그램을 이용하여 선형회귀와 PLS 회귀분석을 수행하여 리드타임 예측모델을 구축하였다. IBM에서 제공하는 SPSS는 기본적인 통계분석부터 회귀분석 이외에 다양한 데이터 분석을 위한 패키지를 제공하고 있지만 단순 통계분석 이외에 복잡한 알고리즘 적용의 한계가 나타났고 데이터의 수가 많아지고 복잡해짐에 따라 분석 속도가 느려지는 단점이 나타났다. 이를 해결하기 위해 다양한 알고리즘과 패키지를 제공하는 여러 가지 오픈소스(R, Python 등)를 활용하여 많은 양의 데이터를 처리한다면 보다 빠르고 심도 있는 데이터 분석이 가능할 것이다.

따라서 본 논문에서는 기본적인 통계분석을 기반으로 기계학습과 딥러닝까지 분석이 가능한 플랫폼을 활용하여 조선소의 데이터를 분석하고자 한다(Fig. 19). 이러한 분석 플랫폼은 최신 데이터 마이닝 기법을 적용할 수 있는 라이브러리와 데이터의 시각화까지 가능한 언어 및 개발환경을 제공하기 때문에 다양한 알고리즘의 적용을 통한 예측모델의 구현이 가능하다. 이를 위해 수집된 조선소 데이터의 분석을 수행하여 조선 생산 리드타임을 예측하기 위한 모델을 생성하여 기준정보의 체계를 수립하고자 한다.

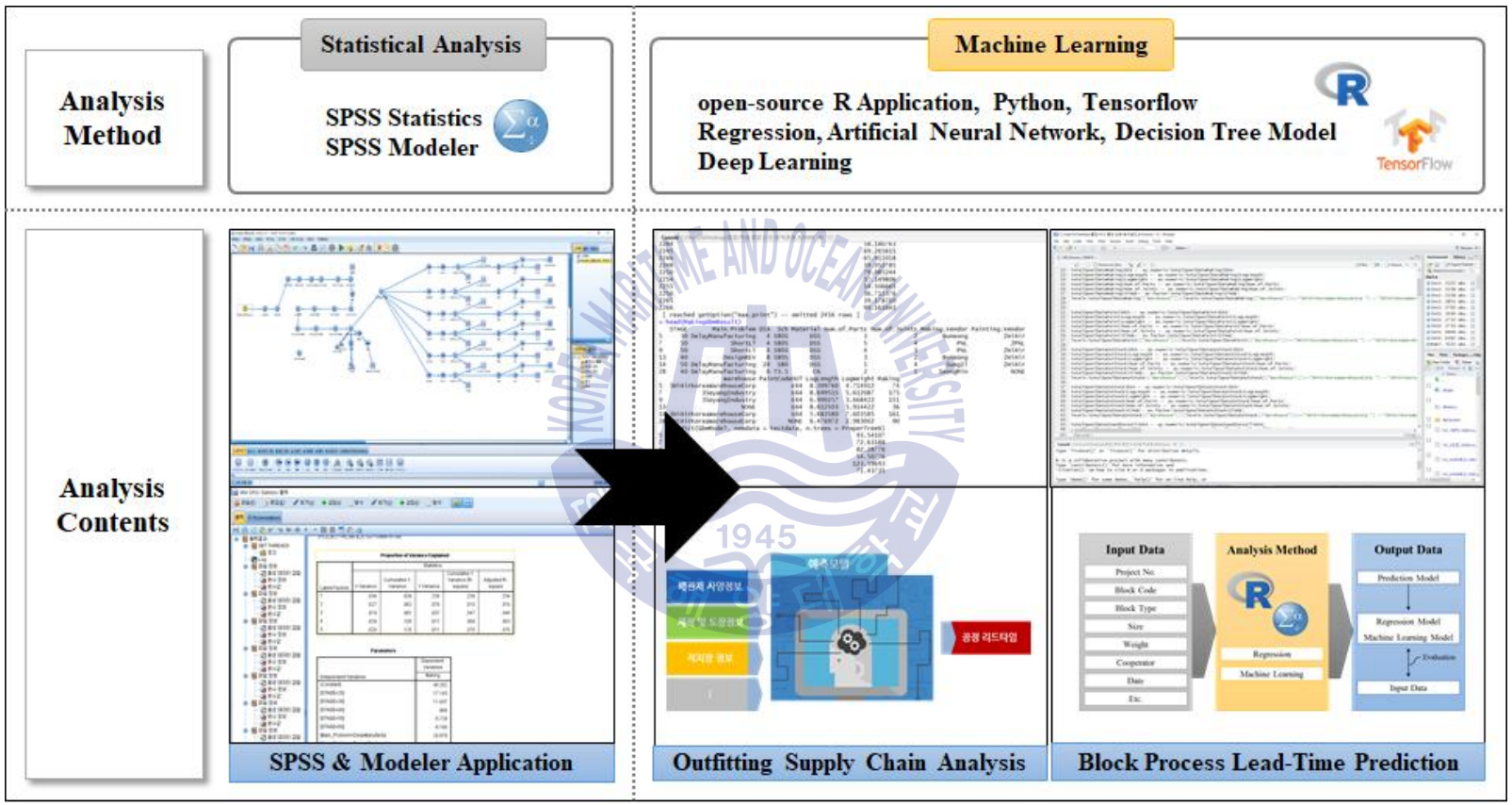


Fig. 19 Application cases of data analysis methodology

3.1 블록 절단공정 중일정 계획 데이터 분석

중소형 조선소 절단공정의 중일정 실적정보를 학습 데이터로 활용하여 계획정보를 예측하였다. 해당 조선소에서는 블록의 절단공정에 대한 중일정 계획을 위해 과거 실적 데이터를 분석하여 계획 데이터의 절단 리드타임을 예측하고자 한다. 절단 공정의 계획 기준 실적 데이터는 5년간 저장된 데이터로 총 64,270개로 구성되어 있으며 다양한 공정정보에서 리드타임 예측을 위한 독립변수를 1차적으로 선별하였다(Table 3). 독립변수는 3개의 연속형 변수인 무게(kg), 강수량(mm), 계획 리드타임(day)과 4개의 범주형 변수인 선종, 블록타입, 블록방향, 계획 협력업체로 구성되어 있다. 정의된 독립변수는 실무자의 의견을 반영한 것으로 실제 절단 리드타임과의 연관성을 분석하기 위하여 상관분석과 분산분석을 추가적으로 수행하였다. 기본적으로 결측값이 존재하는 데이터는 무작위로 발생하였기 때문에 단순 제거를 통해 63,950 rows 데이터만 추출하였다.

Table 3 Block cutting process data

Data	Contents	
Collection Data	- 절단공정 계획기준 실적 데이터 (64,270 rows) - 중일정 계획정보 (2,832 rows)	
Input Data	Continuos Data	Weight (kg)
		Precipitation (mm)
		Planning L/T (day)
	Categorical Data	Ship Type
		Block Group
		Block Direction
		Planning Cooperation
Output Data	- Lead Time (day)	

범주형 변수와 종속변수인 연속형 변수 사이의 관계는 분산분석으로 확인할 수 있다. 분산분석은 F-value와 P-value로 판단이 가능하며 블록의 절단공정 데이터에서 선별된 4개의 범주형 변수와 종속변수인 리드타임 간의 분산분석 결과는 Table 4와 같다. Table 4에서 모든 P-value 값이 유의수준인 0.05에 미치지 못하기 때문에 대립가설이 채택되어 모든 범주형 변수의 집단에 따라 종속변수인 리드타임의 차이가 존재한다는 결과를 얻을 수 있었다. 따라서 1차적으로 선별한 독립변수 중 4개의 범주형 변수는 리드타임에 영향을 미치는 것으로 판단하여 예측모델 생성을 위한 데이터에 활용하고자 한다.

Table 4 Analysis of Variable (a)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cooperation	61	3366504	55189	114.28	<2e-16
Block Group	14	1001382	71527	148.11	<2e-16
Block Direction	2	77692	38846	80.44	<2e-16
Ship Type	12	4662408	388534	804.54	<2e-16

연속형 변수 간의 상관분석은 독립변수인 3개의 연속형 변수와 종속변수인 리드타임 사이의 상관관계를 분석하기 위하여 수행하였다. 본 논문에서는 상관계수의 값을 0.65 기준으로 판단하여 분석한 결과 Fig. 20과 Table 5와 같이 나타났다. 결과적으로 강수량과 계획 리드타임 간의 상관관계가 상대적으로 높은 것으로 나타났기 때문에 차원축소를 수행해야 한다. 하지만 블록 절단 시 계획 기준일로부터 강수량 데이터를 수집하는 것은 현실적으로 어렵기 때문에 독립변수에서 제외하고자 하므로 차원축소 대신 단순히 독립변수에서 삭제하는 것을 선택하였다. 따라서 1차적으로 선별한 3개의 연속형 변수 중 강수량을 제외하여 총 2개의 연속형 변수인 무게와 계획 리드타임만 예측모델 생성을 위한 데이터에 활용하고자 한다.

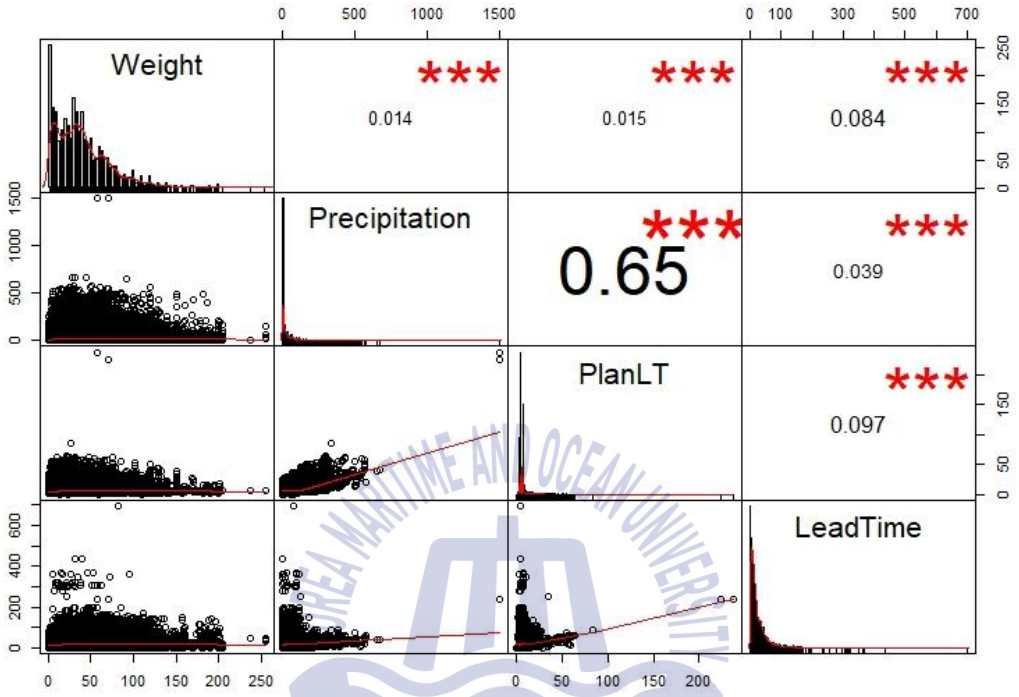


Fig. 20 Graph of Correlation Analysis (a)

Table 5 Correlation Analysis Result (a)

	Weight	Precipitation	Plan L/T	Lead Time
Weight	1	-	-	-
Precipitation	0.01	1	-	-
Plan L/T	0.02	0.65	1	-
Lead Time	0.08	0.04	0.1	1

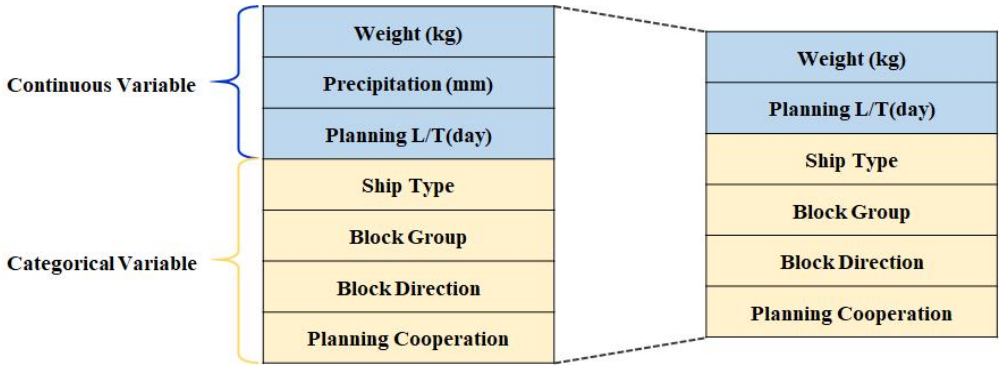


Fig. 21 Modified independent variable (a)

상관분석과 분산분석의 결과를 적용하여 최종적으로 정의된 독립변수는 Fig. 21과 같다. 1개의 연속형 변수를 제외하여 총 6개의 독립변수를 정의하여 분석을 하고자 한다.

데이터 분석과정에서는 기본적으로 분석을 위한 데이터를 정의한 후 데이터의 탐색과 전처리를 수행하게 된다. 데이터의 탐색은 흔히 분포를 확인하는 과정을 의미하는데 데이터의 분포는 Histogram으로, 이상치는 Boxplot으로 확인할 수 있다.

중일정 계획의 실적 데이터에서 최종적으로 정의된 독립변수 중 연속형 변수인 무게와 계획 리드타임의 이상치를 Fig. 22와 같이 확인하였다. 일반적으로 이상치가 단순 발생이거나 분포에 영향을 줄 경우는 삭제하기 때문에 이를 제거하는 작업을 수행하였다. 또한, 종속변수의 왜도가 심하거나 편차가 심할 경우 분석에 오류를 범할 가능성이 있으므로 표준화나 정규화를 수행하는 것이 좋다. Fig. 23과 같이 종속변수인 리드타임의 분포를 분석한 결과 왜도가 지나친 것을 확인하였으므로 시간이라는 특성을 반영해 정규화가 아닌 표준화를 위한 로그변환을 수행하였다.

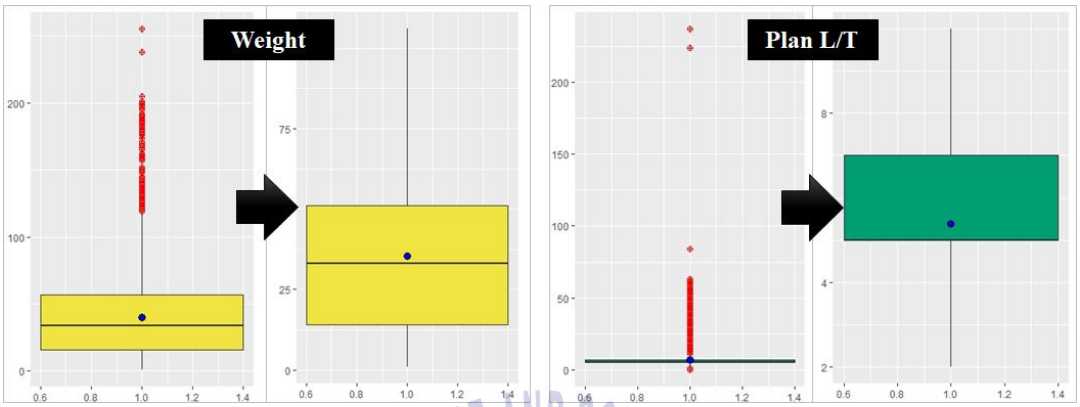


Fig. 22 Outlier treatment of continuos variable (a)

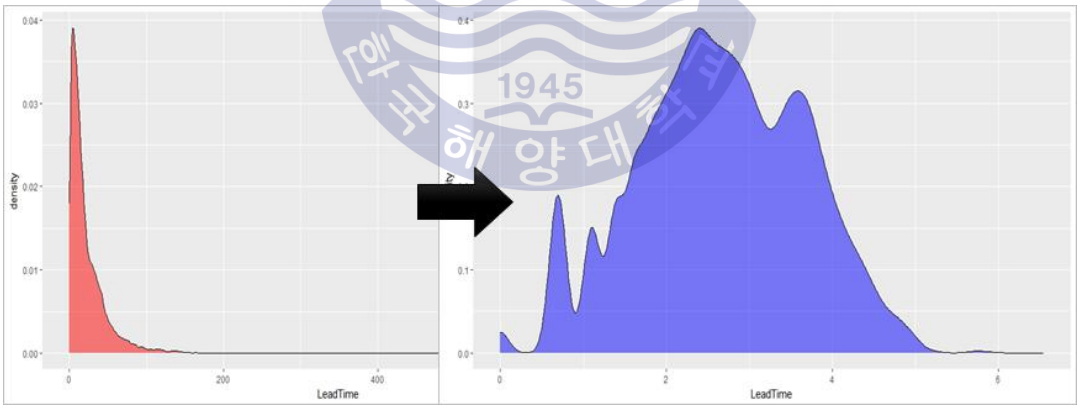


Fig. 23 Log transformation of dependent variable (a)

3.2 블록 탑재공정 실적 데이터 분석

두 번째 사례는 조선소의 블록 공종별 탑재 리드타임을 예측하기 위한 탑재공정의 실적 데이터를 분석한 연구이다. 조선소에서는 선박에 탑재되는 블록의 다양한 정보를 실적 데이터로써 저장 및 관리하고 있다. 블록의 호선별로 작업 대상에 따라 블록 정보를 관리하고 있어 블록의 다양한 공정정보를 활용하여 예측 모델을 생성할 수 있다. 선박의 블록 탑재공정은 블록코드, 블록타입, 방향 등의 정성적 데이터뿐만 아니라 길이, 폭, 면적 등의 정량적 데이터의 다양한 공정정보를 관리한다. 이러한 블록과 관련된 정보들을 입력변수로 실적 리드타임을 목표변수로 정의하여 분석 알고리즘에 적용할 수 있다(Fig. 24).

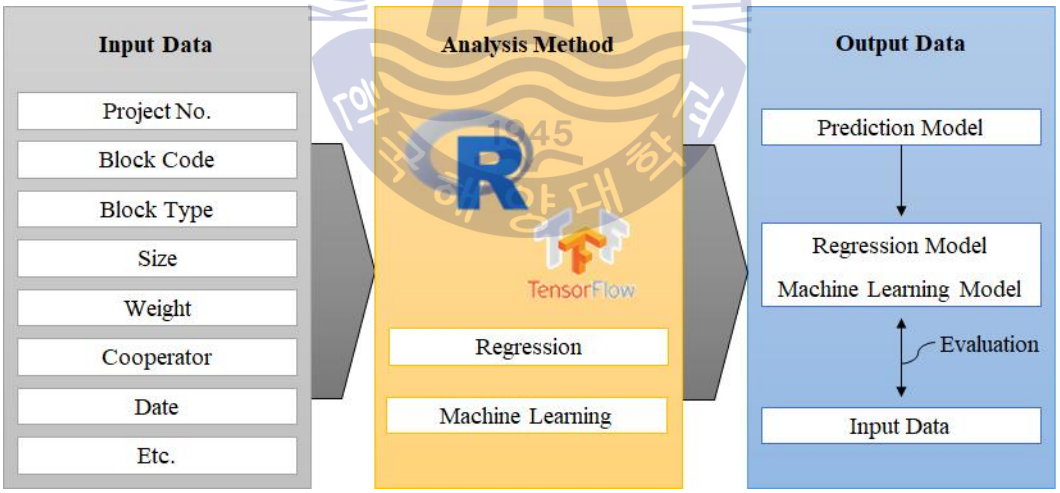


Fig. 24 Data analysis process of block erection data

해당 조선소에서 수집된 데이터는 블록 공종별 실적 데이터와 탑재블록의 정보를 포함하는 데이터로 구분된다. 호선에 따른 블록 탑재 실적 데이터는 최근 3년 간 수집된 약 2만개의 데이터로 구성되어 있고 이 외에 탑재블록 데이터에는 블록에 따른 정량적 정보를 포함하고 있다. 독립변수는 7개의 연속형 변수로 길이(m), 폭(m), 높이((m), 면적(m^2), 부재중량(ton), 순중량(ton), 총중량(ton), 계획 리드타임(day)으로 구성되어 있고, 7개의 범주형 변수는 프로젝트 번호, 대구분, 대공종, 블록그룹, 방향, 모양, 시리얼 번호로 구성되어 있다(Table 6). 정의된 독립변수는 실무자의 의견을 반영한 것으로 실제 절단 리드타임과의 연관성을 분석하기 위하여 상관분석과 분산분석을 추가적으로 수행하였다.

Table 6 Block erection process data

Data	Contents	
Collection Data	- 호선 블록 공종별 실적공기 자료 (22,758 rows) - 블록 데이터 (384 rows)	
Input Data	Continuos Data	Length (m)
		Width (m)
		Height (m)
		Area (m^2)
		Sub Weight (ton)
		Net Weight (ton)
		Weight (ton)
		Planning L/T (day)
	Categorical Data	Project No.
		Division
		Construction
		Block Group
		Direction
		STAGE
		Block Serial No.
Output Data	- Lead Time (day)	

범주형 변수와 종속변수인 연속형 변수 사이의 관계는 분산분석으로 확인할 수 있다. 분산분석은 F-value와 P-value로 판단이 가능하며 블록의 탑재공정 데이터에서 선별된 7개의 범주형 변수와 종속변수인 리드타임 간의 분산분석 결과는 Table 7과 같다. Table 7에서 모든 P-value 값이 유의수준인 0.05에 미치지 못하기 때문에 대립가설이 채택되어 모든 범주형 변수의 집단에 따라 종속변수인 탑재 리드타임의 차이가 존재한다는 결과를 얻을 수 있었다. 따라서 1차적으로 선별한 독립변수 중 7개의 범주형 변수는 리드타임에 영향을 미치는 것으로 판단하여 예측모델 생성을 위한 데이터에 활용하고자 한다.

Table 7 Analysis of Variable (b)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Block Group	10	943840	94384	240.352	<2e-16
Direction	2	51605	25803	65.707	<2e-16
Block Serial	21	110968	5284	13.456	<2e-16
Shape	8	86624	10828	27.574	<2e-16
Project No.	17	62928	3702	9.426	<2e-16
Division	6	216448	36075	91.866	<2e-16
Construction	4	613738	153435	390.726	<2e-16

연속형 변수 간의 상관분석은 독립변수인 8개의 연속형 변수와 종속변수인 리드타임 사이의 상관관계를 분석하기 위하여 수행하였다. 본 논문에서는 상관계수의 값을 0.65 기준으로 판단하여 분석한 결과 Fig. 25와 Table 8과 같이 나타났다. 해당 데이터의 연속형 변수는 대부분 블록의 물성치를 나타내는 데이터로 이 중 총중량은 부재중량과 순중량의 총합과 같으며 블록의 면적은 길이와 너비의 곱과 같다. 따라서 총중량, 부재중량, 순중량 사이의 상관관계가 높은 것으로 나타났으며 면적은 모든 연속형 변수와의 상관관계가 지나치게 높은 것으로 나타났기 때문에 상관분석의 결과를 반영하여

차원축소가 아닌 변수의 단순 제거를 통해 독립변수를 수정하였다. 따라서 1차적으로 선별한 8개의 연속형 변수 중 순중량, 부재중량, 면적, 폭, 너비를 제외하여 최종적으로 3개의 연속형 변수인 길이, 총중량, 계획 리드타임만 예측모델 생성을 위한 데이터에 활용하고자 한다.

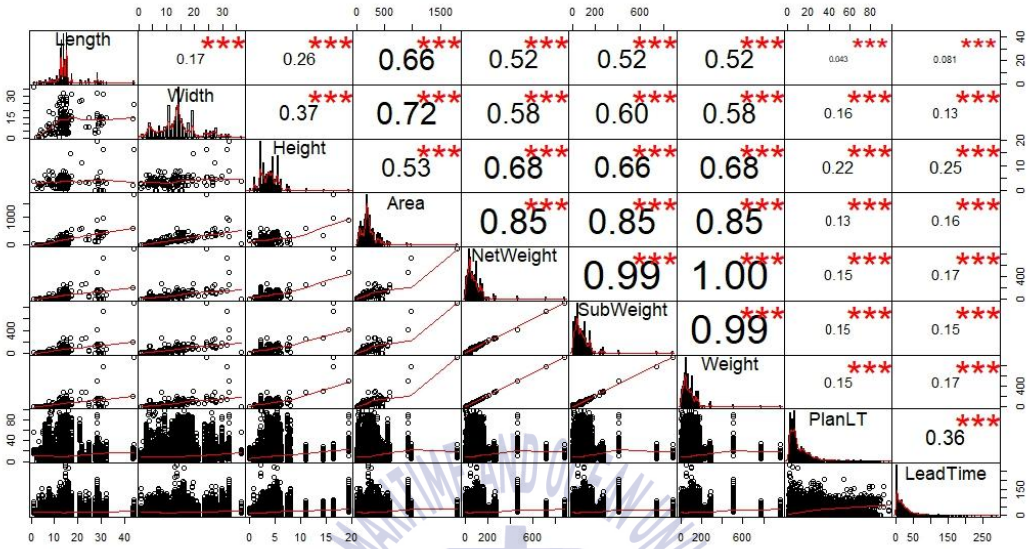


Fig. 25 Graph of Correlation Analysis (b)

Table 8 Correlation Analysis Result (b)

	Length	Width	Height	Area	Net Weight	Sub Weight	Weight	Plan L/T	L/T
Length	1	-	-	-	-	-	-	-	-
Width	0.17	1	-	-	-	-	-	-	-
Height	0.26	0.37	1	-	-	-	-	-	-
Area	0.66	0.72	0.53	1	-	-	-	-	-
Net Weight	0.52	0.58	0.68	0.85	1	-	-	-	-
Sub Weight	0.52	0.6	0.66	0.85	0.99	1	-	-	-
Weight	0.52	0.58	0.68	0.85	1	0.99	1	-	-
Plan L/T	0.04	0.16	0.22	0.13	0.15	0.15	0.15	1	-
L/T	0.08	0.13	0.25	0.16	0.17	0.15	0.17	0.36	1

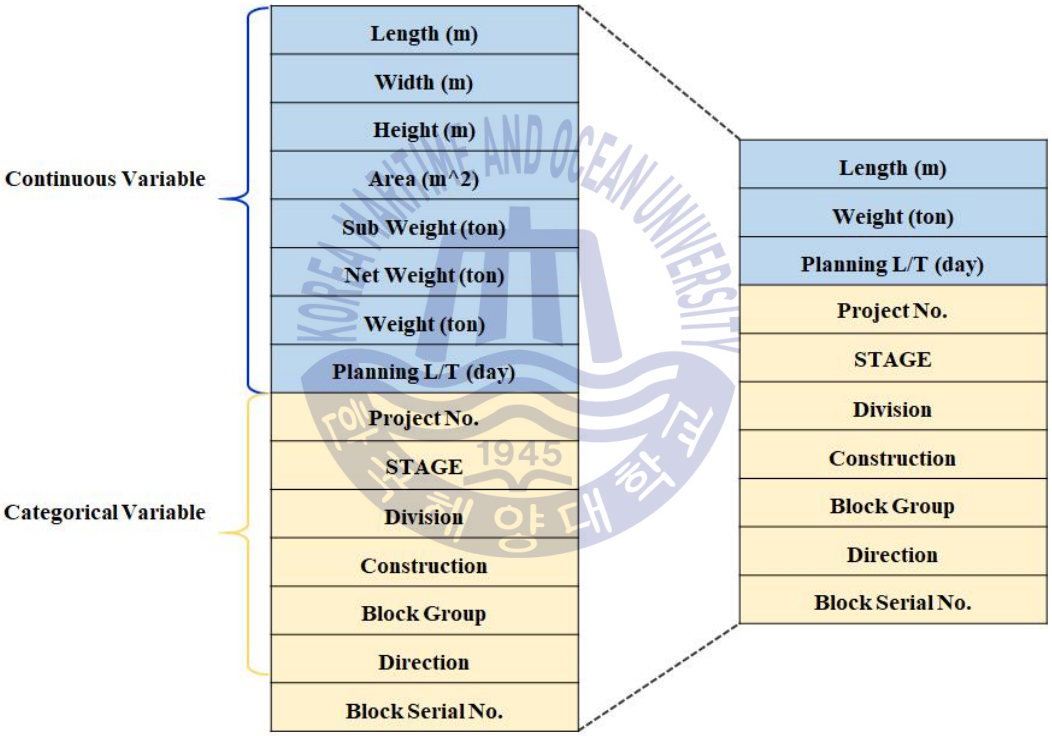


Fig. 26 Modified independent variable (b)

상관분석과 분산분석의 결과를 적용하여 최종적으로 정의된 독립변수는 Fig. 26과 같다. 3개의 연속형 변수를 제외한 총 12개의 독립변수를 정의하여 분석을 하고자 한다.

블록 탑재 실적 데이터에서 최종적으로 정의된 독립변수 중 연속형 변수인 총중량, 길이, 계획 리드타임의 이상치를 Fig. 27과 같이 확인하였다. 일반적으로 이상치가 단순 발생이거나 분포에 영향을 줄 경우는 삭제하기 때문에 이를 제거하는 작업을 수행하였다. 또한, 종속변수의 왜도가 심하거나 편차가 심할 경우 분석에 오류를 범할 가능성이 있으므로 표준화나 정규화를 수행하는 것이 좋다. Fig. 28과 같이 종속변수인 리드타임의 분포를 분석한 결과 왜도가 지나친 것을 확인하였으므로 시간이라는 특성을 반영해 정규화가 아닌 표준화를 위한 로그변환을 수행하였다.

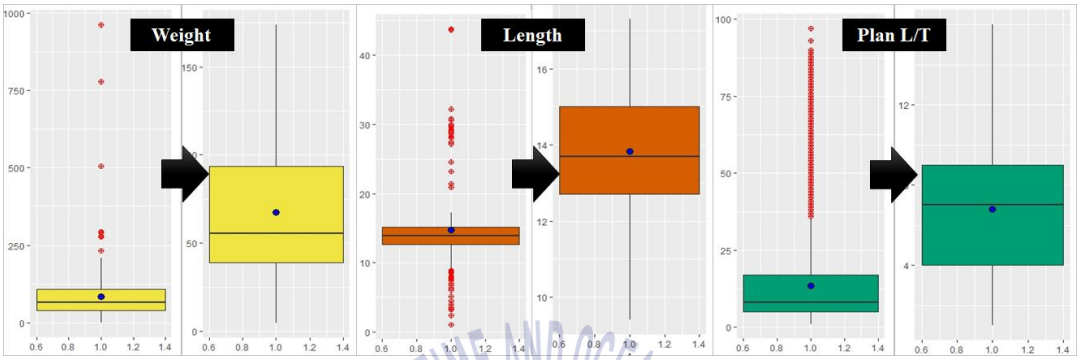


Fig. 27 Outlier treatment of continuos variable (b)

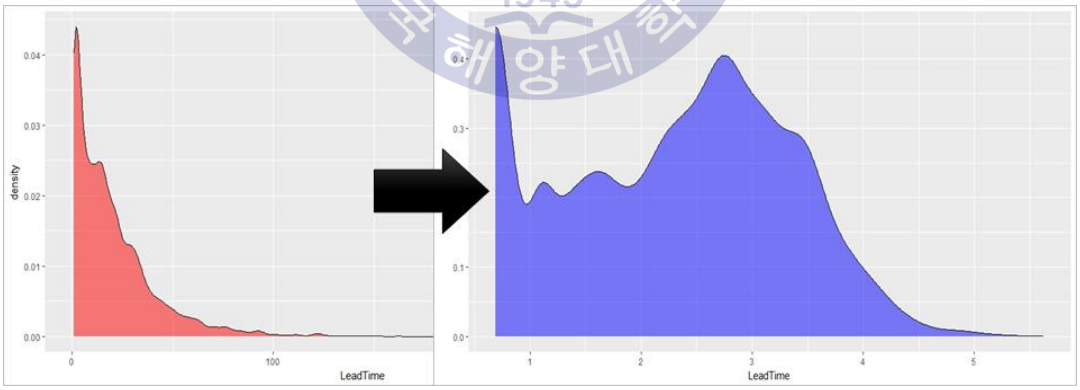


Fig. 28 Log transformation of dependent variable (b)

3.3 해양플랜트 배관재 공급망 데이터 분석

세 번째 사례는 해양플랜트 배관재의 공급망 관리를 위해 배관재 데이터를 분석한 연구이다. 조선소에서의 선박 및 해양플랜트 건조는 설계부터 생산까지 복잡한 공정으로 이루어져 있을 뿐만 아니라 선체에 조립되는 다양한 종류의 의장품들이 복잡한 공급망을 통해 관리되고 있다. 특히 해양플랜트 의장 공정의 대부분을 차지하고 있는 배관재는 적절한 조달관리가 어렵고 수작업에 따른 관리의 한계로 납기 지연에 따른 문제점이 발생하는 경우가 있다. 따라서 납기 관리의 필요성이 높아짐에 따라 다양한 데이터 분석을 시도할 필요성이 있다.

해양플랜트의 배관재도 선박 블록 데이터처럼 공정에 따른 배관재의 다양한 정보를 관리하고 있다. 공급망 데이터에는 배관재의 사양 정보뿐만 아니라 시계열 정보 등 제작 공정부터 설치 공정에 관련된 다양한 공정정보를 관리하기 때문에 이를 바탕으로 공급망 리드타임과 관련된 데이터를 추출할 수 있다. 배관재의 공급망은 Fig. 29와 같이 크게 6개의 공정으로 제작공정부터 설치공정까지 공정절점별로 리드타임이 관리되고 있다. W/O 발행일을 시작으로 제작, 도장, 사외적치, 사내적치, 설치대기, 설치 순으로 공정이 진행된다. 본 논문의 선행연구인 함동균(2016)에서는 전체 6개의 공정에 따른 리드타임을 예측하였지만 사외재고 이후의 공정에서는 다양한 외적요인 및 데이터의 결함으로 인해 예측도가 현저히 떨어진 것으로 판단하여 본 논문에서는 제작과 도장공정의 리드타임만을 예측하고자 한다.

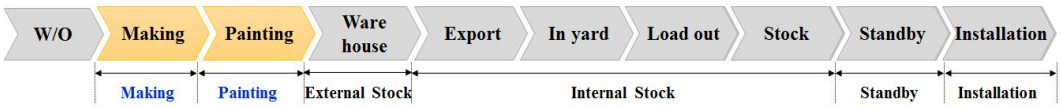


Fig. 29 Supply chain process of offshore outfitting

수집된 배관재 공급망 데이터로부터 예측모델을 생성하기 위한 독립변수와 종속변수를 정의하였다. 최근 작업된 하나의 호선에 설치된 배관재 데이터 16,814개를 수집할 수 있었고 이 중 선박 블록 데이터와 마찬가지로 1차적으로 리드타임과 관련된 데이터를 Table 9와 같이 추출하였다. 독립변수는 4개의 연속형 변수인 무게(ton), 길이(mm), 직경(mm), joint 수와 10개의 범주형 변수인 지연여부, 폐기여부, 긴급여부, 적용L/T, STAGE, 주요 문제점, 협력사, 관통여부, 수정횟수, 재료로 구성되어 있다. 종속변수는 2개의 리드타임을 구분하여 정의하였다. 본 연구에서 예측하고자 하는 대상은 제작공정와 도장공정의 리드타임이기 때문에 공급망에 따라 리드타임을 구분하는 작업을 추가적으로 수행하였다. 따라서 종속변수는 제작 리드타임, 도장 리드타임으로 정의하고 각 공정 리드타임을 예측하기 위한 분석 및 예측은 구분하여 수행하였다.

Table 9 Spool supply chain process data

Data	Contents	
Collection Data	- 배관재 공급망 실적 데이터 (16,814 rows)	
Input Data	Continuos Data	DIA
		Length (mm)
		Weight (ton)
		Joint
	Categorical Data	Put off
		Disuse
		Emergency
		Apply LT
		STAGE
		Main Problem
		Cooperation
		Penetration
		Rev No.
		Material
Output Data	- Lead Time (day)	

먼저 연속형 변수 간의 상관분석은 독립변수인 4개의 연속형 변수와 종속변수인 2개의 리드타임 사이의 상관관계를 분석하기 위하여 수행하였다. 본 논문에서는 상관계수의 값을 0.65 기준으로 판단하여 분석한 결과 Fig. 30과 Table 10과 같이 나타났다. 상관계수를 살펴보면 직경과 중량 사이의 상관관계가 상대적으로 높은 것으로 나타났지만 Fig. 30에서 'DIA'의 그래프를 살펴보면 데이터의 분포가 정량적이 아닌 정성적 특징을 나타낸 것으로 판단되었다. 따라서 'DIA' 변수를 연속형 변수가 아닌 범주형 변수로 데이터의 유형을 변환하는 것이 필요하다. 이 외에는 연속형 변수 사이의 상관관계가 낮은 것으로 나타났기 때문에 예측모델 생성에 있어서 다중공선성의 영향을 주지 않을 것으로 판단되어 모두 독립변수로 활용하고자 한다. 따라서 최종적으로 독립변수로 정의된 연속형 변수는 3가지로 길이, 중량, joint 수로 구성된다.

범주형 변수와 종속변수인 연속형 변수 사이의 관계는 분산분석으로 확인할 수 있다. 분산분석은 F-value와 P-value로 판단이 가능하며 배관재의 공급망 데이터에서 선별된 9개의 범주형 변수와 종속변수인 2개의 리드타임 간의 분산분석을 수행하였다. 여기서는 공정에 따라 리드타임을 구분하였기 때문에 분산분석 역시 공정 리드타임 별로 분석하였으며 결과는 Table 11~12와 같다. 분산분석 결과 F-value 및 P-value 값이 유의수준인 0.05에 미치지 못한 변수는 지연여부(Put off), 폐기여부(Disuse), 관통(Penetration)으로 나타났기 때문에 이 3개의 범주형 변수에 따라 제작과 도장 리드타임의 차이가 존재하지 않는다는 결과를 얻었다. 따라서 1차적으로 선별한 독립변수 중 10개의 범주형 변수 중 3가지의 변수를 제외한 7개의 범주형 변수만을 리드타임 예측모델 생성을 위한 데이터에 활용하고자 한다.

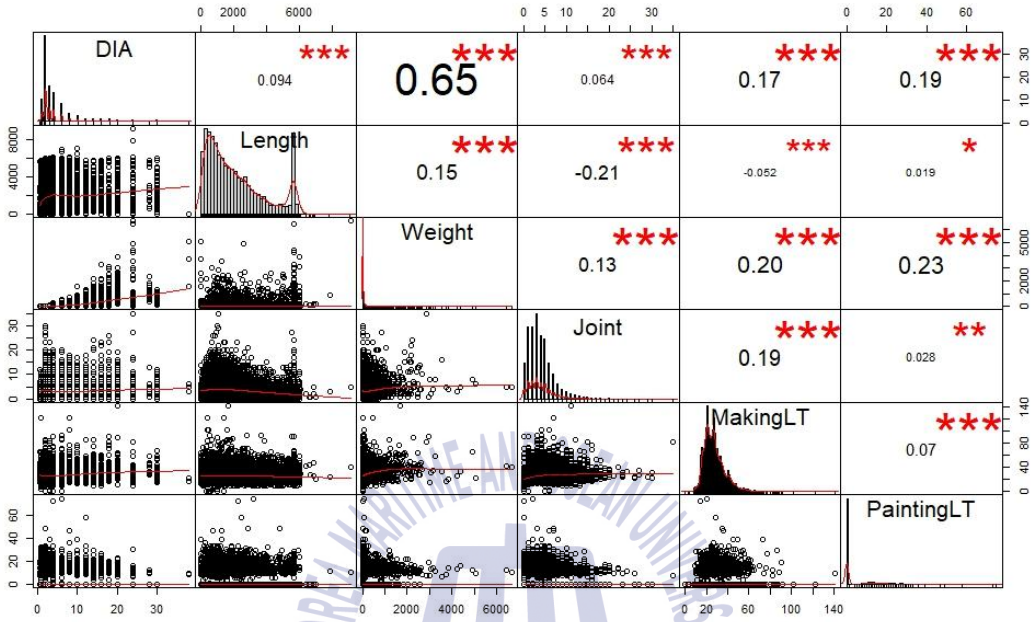


Fig. 30 Graph of Correlation Analysis (c)

Table 10 Correlation Analysis Result (c)

	DIA	Length	Weight	Joint	Making L/T	Painting L/T
DIA	1	-	-	-	-	-
Length	0.09	1	-	-	-	-
Weight	0.65	0.15	1	-	-	-
Joint	0.06	-0.21	0.13	1	-	-
Making L/T	0.17	-0.05	0.20	0.19	1	-
Painting L/T	0.19	0.02	0.23	0.03	0.07	1

Table 11 Analysis of Variable (c)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Emergency	1	9082	9082	112.194	< 2e-16
Put Off	1	16	16	0.193	0.6608
Disuse	1	266	266	3.284	0.07
Apply LT	2	491	246	3.035	0.0481
STAGE	4	11937	2984	36.864	< 2e-16
Main Problem	7	46767	6681	82.53	< 2e-16
Penetration	1	178	178	2.203	0.1378
REV No.	4	5926	1482	18.302	5.55E-15
Material	4	13014	3254	40.191	< 2e-16

Table 12 Analysis of Variable (d)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Emergency	1	901	901	30.183	4.05E-08
Put Off	1	46	45.5	1.525	0.217
Disuse	1	10	9.6	0.321	0.571
Apply LT	2	2080	1040.2	34.846	8.50E-16
STAGE	4	12442	3110.4	104.199	< 2e-16
Main Problem	7	6189	884.1	29.617	< 2e-16
Penetration	1	75	74.7	2.503	0.114
REV No.	4	2088	522.1	17.49	2.66E-14
Material	4	9780	2444.9	81.903	< 2e-16

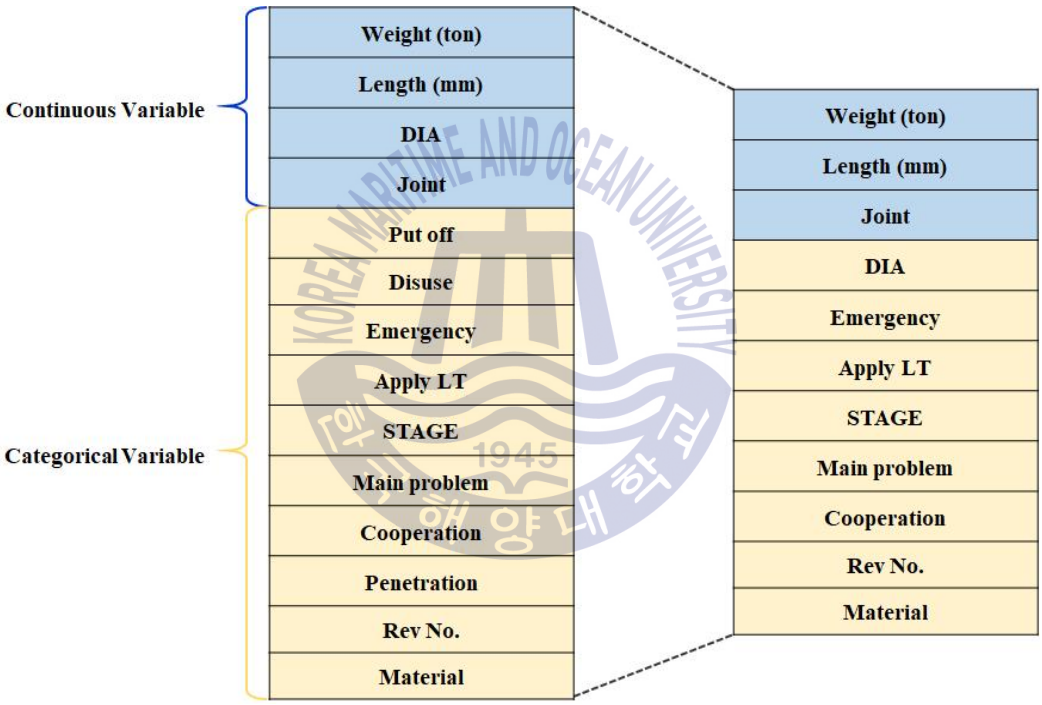


Fig. 31 Modified independent variable (c)

상관분석과 분산분석의 결과를 적용하여 최종적으로 정의된 독립변수는 Fig. 31과 같다. 1개의 연속형 변수를 범주형 변수로 변환하였으며 기존의 변수 중 3개의 범주형 변수를 제외한 총 11개의 독립변수를 정의하여 분석을 하고자 한다.

배관재 공급망 데이터에서 최종적으로 정의된 독립변수 중 연속형 변수인 총중량, 길이, joint 수의 이상치를 Fig. 32와 같이 확인하였다. 일반적으로 이상치가 단순 발생이거나 분포에 영향을 줄 경우는 삭제하기 때문에 이를 제거하는 작업을 수행하였다. 또한, 종속변수의 왜도가 심하거나 편차가 심할 경우 분석에 오류를 범할 가능성이 있으므로 표준화나 정규화를 수행하는 것이 좋다. Fig. 33과 같이 종속변수인 리드타임의 분포를 제작과 도장 공정에 따라 분석한 결과 제작 리드타임은 비교적 분포가 고르게 나타났기 때문에 추가적인 표준화를 수행하지 않았다. 반면 도장 리드타임은 outlier로 인한 왜도가 지나친 것을 확인하였기 때문에 데이터를 제거하여 축소하였다.

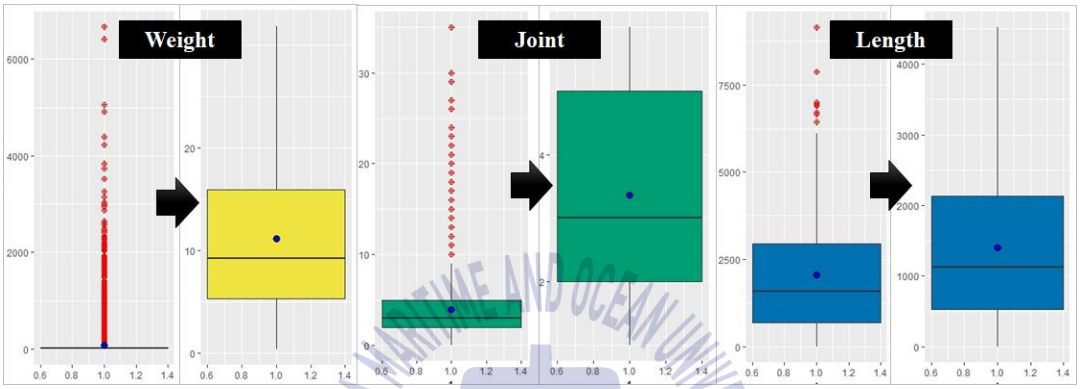


Fig. 32 Outlier treatment of continuos variable (c)

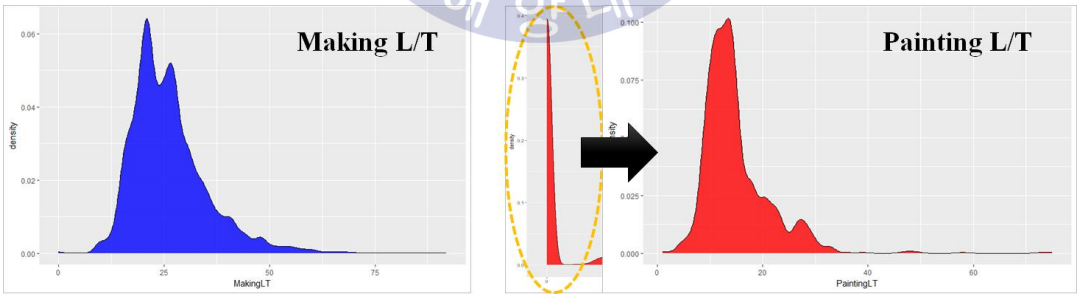


Fig. 33 Log transformation of dependent variable (c)

제 4 장 예측모델 결과분석

조선 생산 리드타임을 예측하기 위해 수집된 3가지의 데이터를 분석하여 데이터 전처리 및 변수 정의한 결과를 바탕으로 예측모델을 구축하고자 한다. 수치예측을 위해 활용하고자 하는 알고리즘은 크게 기계학습과 딥러닝이기 때문에 각각의 알고리즘에 다른 예측모델을 구축하고 이에 대한 성능평가를 위해 여러 가지 평가지표를 활용하고자 한다.

4.1 기계학습 예측모델 결과

수집된 데이터를 통해 예측하고 하는 생산 리드타임은 선박 탑재공정 리드타임, 선박 절단공정 리드타임, 배관재 공급망 리드타임으로 구성되어 있다. 데이터의 다양한 전처리 기법을 적용한 것 이외에도 리드타임과 관련된 독립변수를 정의하기 위한 상관분석과 분산분석을 수행하여 학습을 위한 데이터를 추출하는 작업을 수행하였다. 본 논문에서는 이러한 데이터 전처리에 따른 예측모델의 결과가 어떻게 달라지는지를 판단하기 위하여 공정에 따른 리드타임의 예측모델을 두 가지의 케이스로 분류하여 구축하고자 한다.

먼저 첫 번째 비교대상인 ‘Case 1’의 경우는 데이터 전처리를 하지 않는 Raw Data를 의미하는 것으로 초기에 수집된 데이터를 있는 그대로 활용한 것이다. 이 때 독립변수는 1차적으로 정의된 변수들을 활용한 것으로 상관분석과 분산분석의 결과가 적용되지 않는 데이터를 의미한다.

두 번째 비교대상인 ‘Case 2’는 이상치와 결측값 처리 및 데이터 분포 변환 등의 전처리와 상관분석과 분산분석의 결과를 적용한 데이터를 의미한다. 데이터의 정규화에 따른 결과 및 리드타임과의 상관성이 높은 독립변수 정의가 예측 결과에 어떠한 영향을 미치는지를 판단하기 위한 데이터라고 할 수 있다. 최종적으로 정의된 분석 케이스에 따른 데이터의 수는 Table 13과 같으며

각각의 케이스에 따라 예측모델 구축을 위한 학습용 데이터와 평가용 데이터가 다르다는 것을 알 수 있다.

Table 13 Number of data according to analysis case

	블록 절단공정 리드타임	블록 탑재공정 리드타임	배관재 제작 리드타임	배관재 도장 리드타임
Case 1	63,987 rows	21,868 rows	11,876 rows	11,876 rows
Case 2	49,290 rows	13,554 rows	7,994 rows	928 rows

첫 번째 분석 사례인 블록 절단공정 리드타임의 예측결과는 Table 14와 Fig. 34와 같다. 먼저 기계학습 알고리즘에 따른 예측결과는 모든 분석 케이스에서 의사결정나무 모델이 가장 좋은 성능을 나타내었다. 3가지 평가지표인 MAE, MAPE, RMSE 모두 의사결정나무 모델에서 좋은 결과를 나타냈으며 특히 오차율이 다른 두 개의 알고리즘보다 낮은 것을 확인하였다. 분석 케이스에 따른 예측결과를 살펴보면 모든 알고리즘에서 Case 2의 데이터가 가장 좋은 평과결과를 보였다. 특히 회귀분석 모델에서도 모델의 설명력이 높아지는 것을 확인한 것으로 보아 데이터의 전처리 유무에 따라 예측모델의 결과가 달라진다는 결론을 얻었다.

Table 14 Machine learning result of block cutting lead time

	Regression		ANN		Tree	
	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2
MAE	13.32	7.42	12.91	7.96	11	7.72
MAPE	169%	95%	162%	91%	126%	82%
RMSE	21.91	10.39	21.55	11	19.51	10.84
R^2	0.24	0.36				

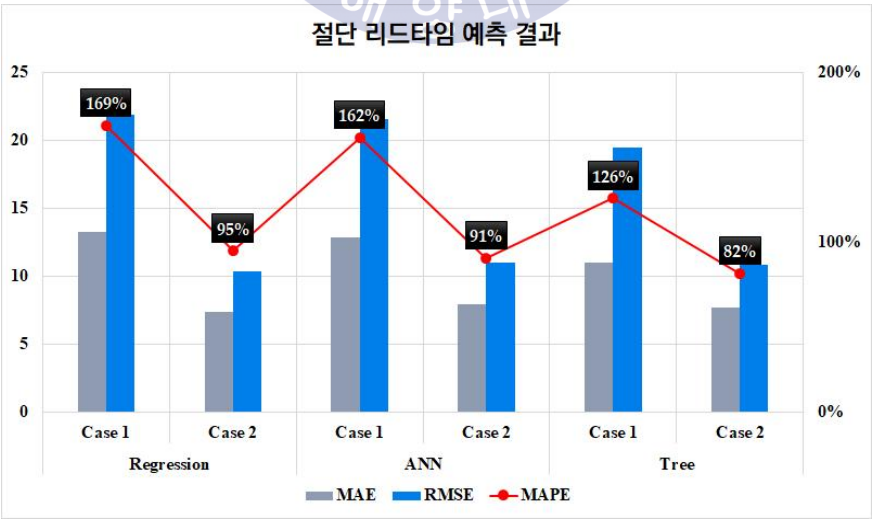


Fig. 34 Machine learning result of block cutting lead time

두 번째 분석 사례인 블록 탑재공정 리드타임의 예측결과는 Table 15와 Fig. 35와 같다. 마찬가지로 기계학습 알고리즘에 따른 예측결과는 모든 분석 케이스에서 의사결정나무 모델이 가장 좋은 성능을 나타내었다.

분석 케이스에 따른 예측결과를 살펴보면 모든 알고리즘에서 Case 2의 데이터가 가장 좋은 평과결과를 보였다. 특히 인공신경망 모델에서는 데이터 케이스에 따른 예측모델의 결과가 큰 편차를 보였다. 또한 회귀분석 모델에서도 모델의 설명력이 높아지는 것을 확인한 것으로 보아 마찬가지로 데이터의 전처리 유무에 따라 예측모델의 결과가 달라진다는 결론을 얻었다.

Table 15 Machine learning result of block erection lead time

	Regression		ANN		Tree	
	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2
MAE	12	9.13	13.58	8.65	9.36	8.21
MAPE	255%	133%	334%	126%	148%	117%
RMSE	18.71	15.54	20.18	15.29	15.73	14.21
R^2	0.34	0.38				

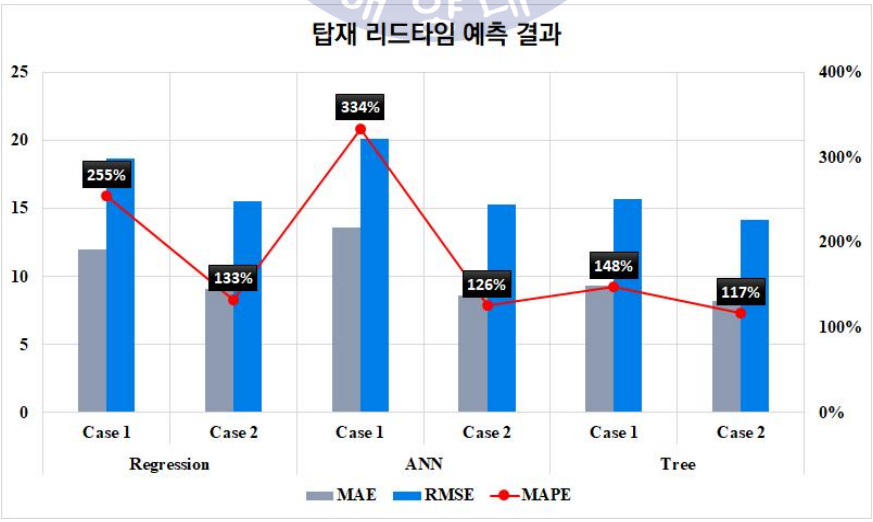


Fig. 35 Machine learning result of block erection lead time

세 번째 분석 사례인 배관재 공급망 리드타임의 예측결과는 Table 16~17과 Fig. 36~37과 같다. 먼저 제작 리드타임을 살펴보면 블록 절단과 탑재 리드타임 결과에 비해 낮은 예측 오차율과 잔차를 보였으며 전반적으로 알고리즘에 따른 결과는 비슷하게 나타났다.

반면 분석 케이스에 따른 예측결과는 회귀분석에서 다소 차이를 보였으나 인공신경망과 의사결정나무 모델에서는 대체적으로 좋은 결과를 보였다. 특히 회귀분석의 설명력이 높아진 것으로 보아 데이터의 전처리의 유무가 영향을 미친 것으로 판단된다.

Table 16 Machine learning result of spool making lead time

	Regression		ANN		Tree	
	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2
MAE	8.46	4.54	5.08	4.34	4.96	4.28
MAPE	35%	19%	20%	18%	19%	18%
RMSE	10.95	5.86	6.98	5.65	6.98	5.57
R^2	0.38	0.43				

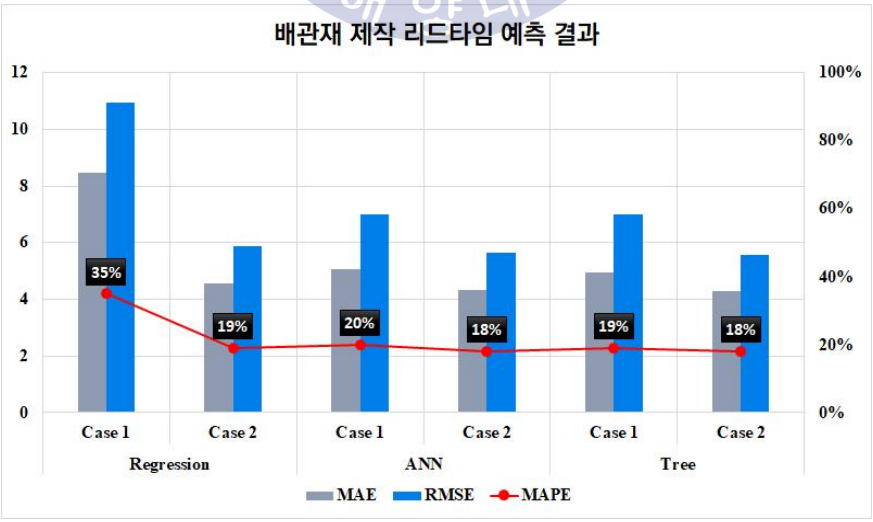


Fig. 36 Machine learning result of spool making lead time

반면 배관재 도장 리드타임의 결과를 살펴보면 제작 리드타임에 비해 오차율의 평균이 높은 것으로 나타났지만 알고리즘에 따른 차이가 크지 않다는 것을 확인하였다.

분석 케이스에 따른 예측결과는 마찬가지로 Case 2의 데이터가 가장 좋은 평가결과를 보였으며, 회귀분석의 설명력 또한 높아진 것을 확인하였기 때문에 데이터의 전처리의 유무가 중요하다는 결론을 얻었다.

Table 17 Machine learning result of spool painting lead time

	Regression		ANN		Tree	
	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2
MAE	10.57	3.78	4.48	2.98	4.35	3.66
MAPE	32%	25%	34%	25%	49%	24%
RMSE	11.5	5.14	6.37	4.36	3.59	5.15
R^2	0.33	0.44				

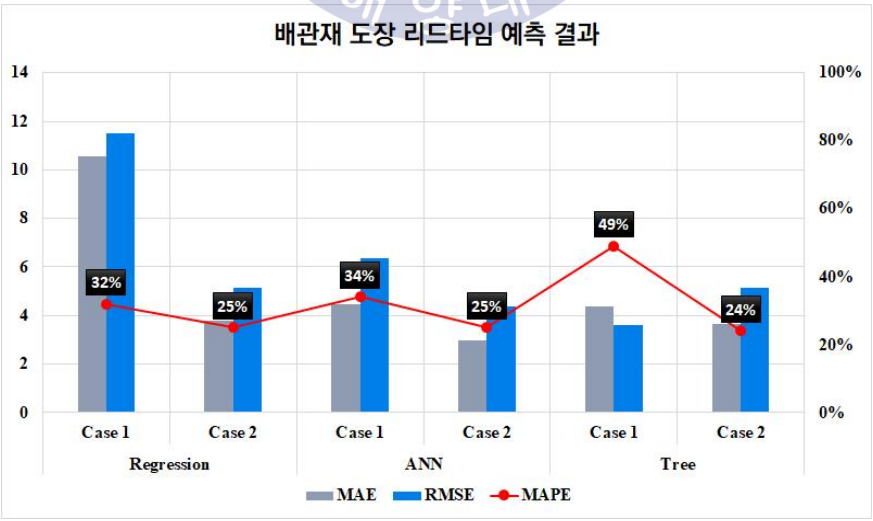


Fig. 37 Machine learning result of spool painting lead time

4.2 딥러닝 예측모델 결과

기계학습 알고리즘은 회귀분석, 인공신경망, 의사결정나무를 활용하여 총 3가지의 예측모델 구축하여 결과를 분석하였다. 본 논문에서는 딥러닝 알고리즘 중 수치예측을 위해 다층 퍼셉트론과 순환신경망 모델을 적용하고자 한다.

딥러닝 알고리즘을 적용하기 위한 라이브러리로 'Keras'를 활용하였다. 'Keras' 라이브러리는 다양한 파라미터를 설정하여 딥러닝 모델을 구축할 수 있다. 모델을 학습하기 위해 가장 기본적으로 정의하는 파라미터는 Batch size와 Epoch가 있다. Batch size는 모델을 학습할 때 몇 개의 샘플로 가중치를 갱신할 것인지를 지정하는 값이고, Epoch는 모델의 학습 반복 횟수를 의미한다. 이 외에 모델을 학습하기 위한 활성화 함수를 지정하는 Activation과 Dropout 설정, Hidden layer의 수에 따라 다양한 딥러닝 모델을 구축할 수 있다.

본 논문에서는 학습 반복 횟수인 Epoch는 200회로 고정하고 활성화 함수는 수치예측에서 가장 많이 활용되는 'Relu'를 적용하고자 한다. 따라서 기본적인 다중 퍼셉트론 모델을 적용하되 은닉층과 배치사이즈의 값에 따른 결과와 종속변수의 데이터 분포에 따른 결과 및 추가적으로 순환 신경망 모델을 비교하기 위한 분석 케이스를 Table 18과 같이 분류하였다. Case 1 데이터는 다층 퍼셉트론을 적용한 3개의 은닉층과 배치사이즈의 값이 100인 모델로 비교대상의 기준이 되도록 정의하였다. 이와 비교하기 위해 배치사이즈에 따라 Case 2,3을 정의하고 은닉층의 수에 따라 Case 4,5를 정의하였다. 또한 기본 모델인 Case 1과 비교하기 위해 종속변수인 리드타임의 표준화를 적용한 모델을 Case 6으로 정의하였고, 순환 신경망 모델을 적용한 데이터를 Case 7로 정의하였다.

Table 18 Analysis case for deep learning model

	Data	Hidden Layer	Batch Size
Case 1	• MLP • Output Log(x)	3	100
Case 2	• MLP • Output Log(x)	3	50
Case 3	• MLP • Output Log(x)	3	30
Case 4	• MLP • Output Log(x)	5	100
Case 5	• MLP • Output Log(x)	10	100
Case 6	• MLP • Output Log(o)	3	100
Case 7	• RNN • Output Log(o)	-	100

첫 번째 분석 사례인 블록 절단공정 리드타임의 예측결과는 Table 19와 Fig. 38과 같다. Case 1을 기준으로 batch size와 hidden layer에 따른 결과를 비교해보면 예측의 오차율이 큰 변화는 보이지 않았으나 전반적으로 기계학습 알고리즘보다는 좋은 성능을 보였다. 특히 종속변수의 표준화를 적용한 Case 6과 Case 7에서의 성능평가가 눈에 띄게 좋아진 것을 확인할 수 있었다.

Table 19 Deep learning result of block cutting lead time

	Case1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
MAE	6.18	6.04	6.25	6.25	6.60	7.59	7.24
MAPE	108%	106%	115%	110%	126%	69%	75%
RMSE	8.14	8.13	8.23	8.21	8.37	11.13	8.56

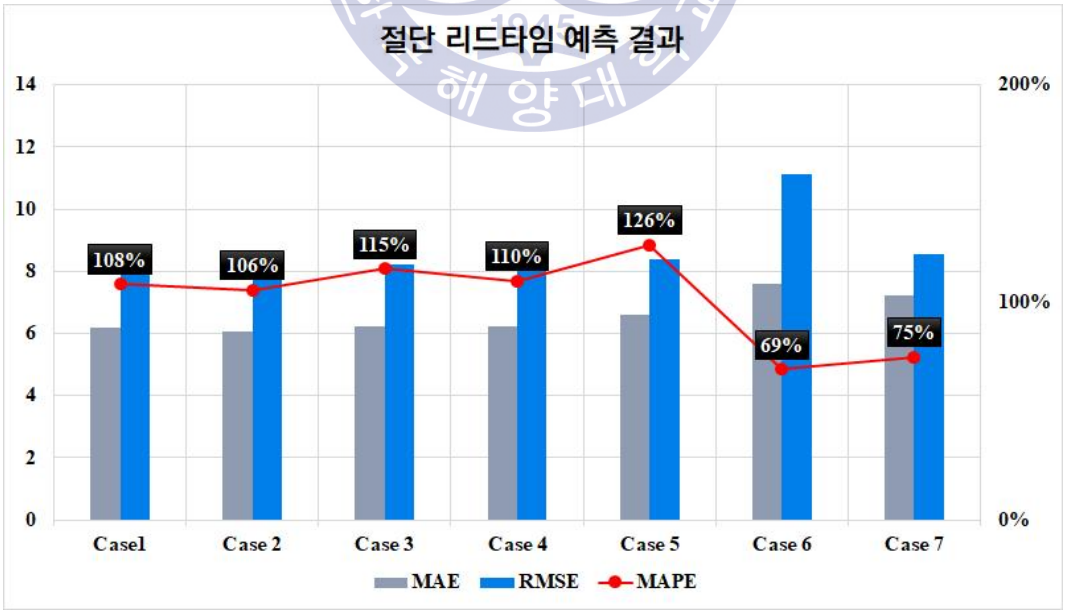


Fig. 38 Deep learning result of block cutting lead time

두 번째 분석 사례인 블록 탑재공정 리드타임의 예측결과는 Table 20과 Fig. 39와 같다. Case 1을 기준으로 batch size와 hidden layer에 따른 결과를 비교해보면 마찬가지로 예측의 오차율이 큰 변화는 보이지 않았으나 전반적으로 기계학습 알고리즘보다는 좋은 성능을 보였다. 절단 리드타임과 동일하게 종속변수의 표준화를 적용한 Case 6과 Case 7에서의 성능평가가 눈에 띄게 좋아진 것을 확인할 수 있었다.

Table 20 Deep learning result of block erection lead time

	Case1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
MAE	6.78	6.65	6.46	6.63	6.49	6.68	6.90
MAPE	128%	125%	115%	126%	116%	144%	70%
RMSE	12.03	12.15	12.07	11.78	12.29	11.74	13.21

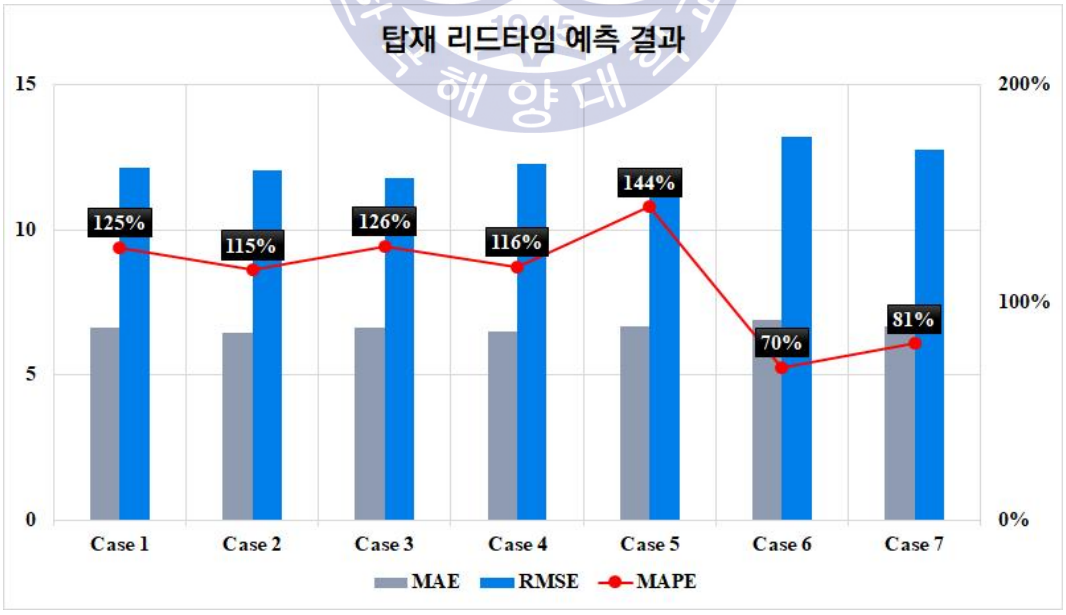


Fig. 39 Deep learning result of block erection lead time

세 번째 분석 사례인 배관재 공급망 리드타임의 예측결과는 Table 21~22와 Fig. 40~41과 같다. 제작 리드타임의 예측결과를 살펴보면 분석 케이스에 따라 큰 차이는 없었으나 batch size가 작은 데이터에서 기존의 예측모델보다 오차율이 낮아졌다. 또한, 블록 절단 및 탑재 리드타임에 비해 리드타임의 분포가 고르게 나타났기 때문에 데이터의 표준화 효과가 크게 작용하지 않은 것으로 판단되어 Case 6과 Case 7에서의 예측도가 비슷한 결과를 나타냈다.

Table 21 Deep learning result of spool making lead time

	Case1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
MAE	4.11	4.18	4.00	4.12	4.61	4.27	4.16
MAPE	16%	18%	16%	16%	17%	16%	16%
RMSE	5.49	5.49	5.38	5.56	6.18	5.83	5.67

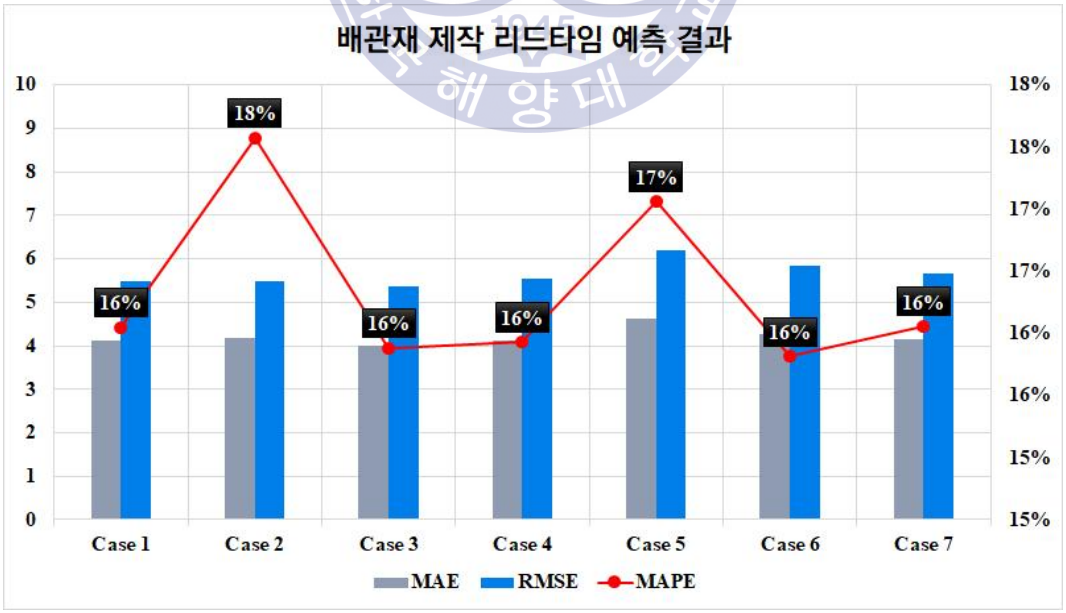


Fig. 40 Deep learning result of spool making lead time

도장 리드타임의 예측결과도 마찬가지로 분석 케이스에 따라 큰 차이는 없었으나 batch size가 작은 데이터에서 가장 높은 정확도를 나타냈다. 또한 데이터 전처리 과정에서 이상치가 있는 리드타임을 제거했기 때문에 데이터의 표준화 효과가 미비한 것으로 판단되어 제작 리드타임의 예측결과와 동일하게 Case 6과 Case 7에서의 예측도가 비슷한 결과를 나타냈다.

Table 22 Deep learning result of spool painting lead time

	Case1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
MAE	3.38	3.23	3.36	3.81	4.48	4.40	3.95
MAPE	24%	22%	22%	24%	27%	24%	23%
RMSE	5.20	5.13	5.29	5.99	6.91	6.84	6.30

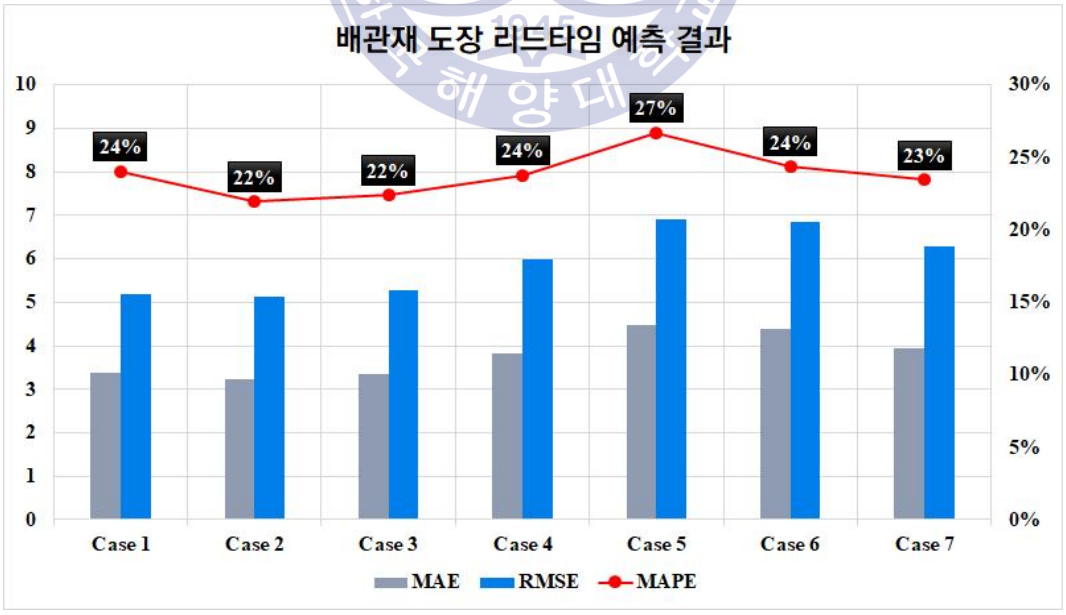


Fig. 41 Deep learning result of spool painting lead time

4.3 알고리즘에 따른 결과분석

본 논문에서는 수집된 3가지의 공정 데이터를 기반으로 조선 생산 리드타임을 예측하기 위한 모델을 구축하였다. 기계학습 알고리즘의 회귀분석, 인공신경망, 의사결정나무, 그리고 딥러닝 알고리즘을 적용한 예측모델의 성능평가를 최종적으로 분석한 결과는 Table 23~26, Fig. 42~45와 같다.

조선소의 공정 데이터에 따르면 해양플랜트 배관재 공급망 리드타임의 예측모델이 가장 좋은 성능을 보였고 그 다음 블록 절단 리드타임, 블록 탑재 리드타임 순으로 예측도가 좋았다. 반면, 알고리즘에 따른 예측모델의 결과는 모든 데이터에서 회귀분석, 인공신경망, 의사결정나무, 딥러닝 순으로 성능평가가 좋은 것을 확인하였다.

수집된 데이터의 리드타임의 평균 및 편차는 공정에 따라 차이가 존재하기 때문에 MAPE 값의 차이가 나타난 것으로 보인다. 또한, 모든 데이터가 비선형적 특징을 갖고 있기 때문에 일반적인 회귀분석 보다 딥러닝 알고리즘에서 예측의 정확도가 높아진 것으로 판단된다.

Table 23 Final prediction result of block cutting lead time

	Regression	Neural Network	Decision Tree	Deep Neural Network
MAE	7.42	7.96	7.72	6.65
MAPE	95%	91%	82%	71%
RMSE	10.39	11	10.84	9.59

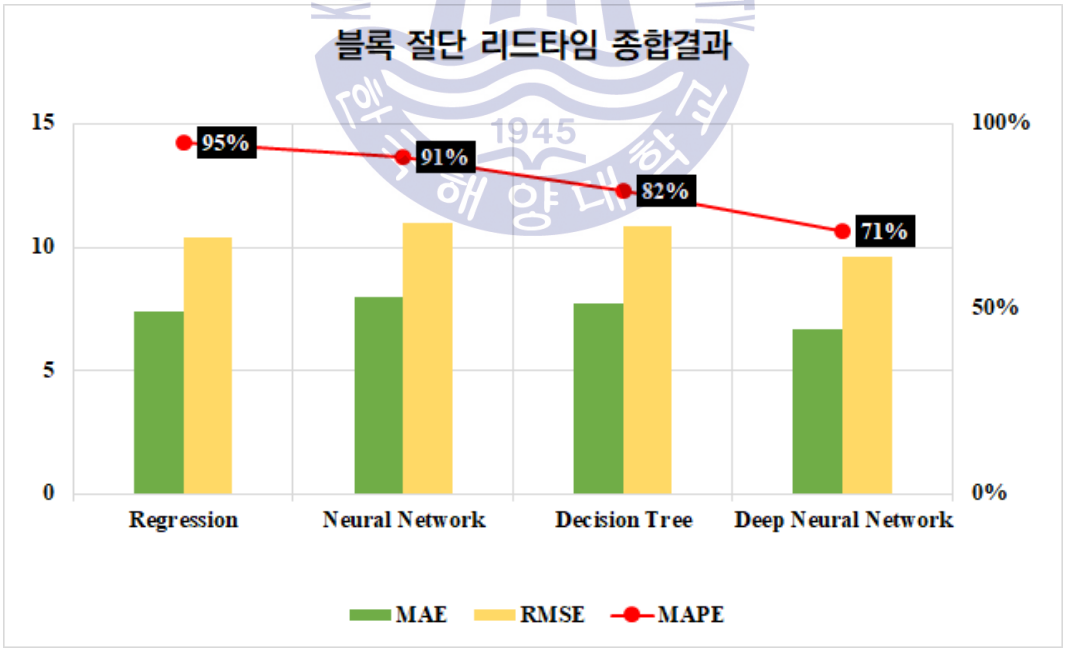


Fig. 42 Final prediction result of block cutting lead time

Table 24 Final prediction result of block erection lead time

	Regression	Neural Network	Decision Tree	Deep Neural Network
MAE	9.13	8.65	8.21	7.59
MAPE	133%	126%	117%	69%
RMSE	15.54	15.29	14.21	11.130

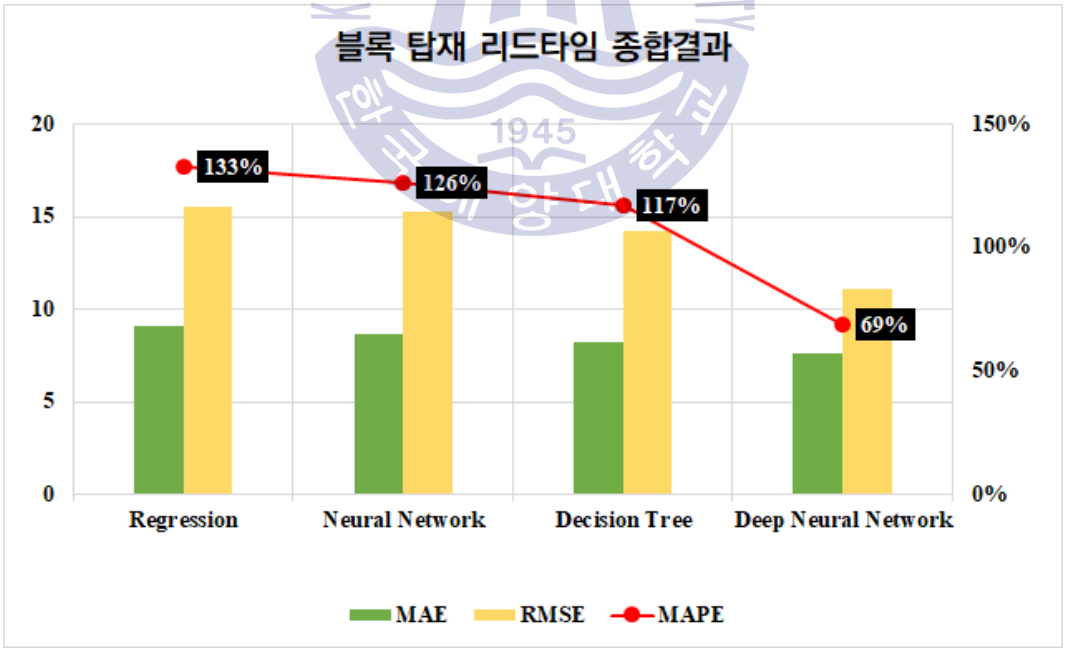


Fig. 43 Final prediction result of block erection lead time

Table 25 Final prediction result of spool making lead time

	Regression	Neural Network	Decision Tree	Deep Neural Network
MAE	4.54	4.34	4.28	4.00
MAPE	19%	18%	18%	16%
RMSE	5.86	5.65	5.57	5.38

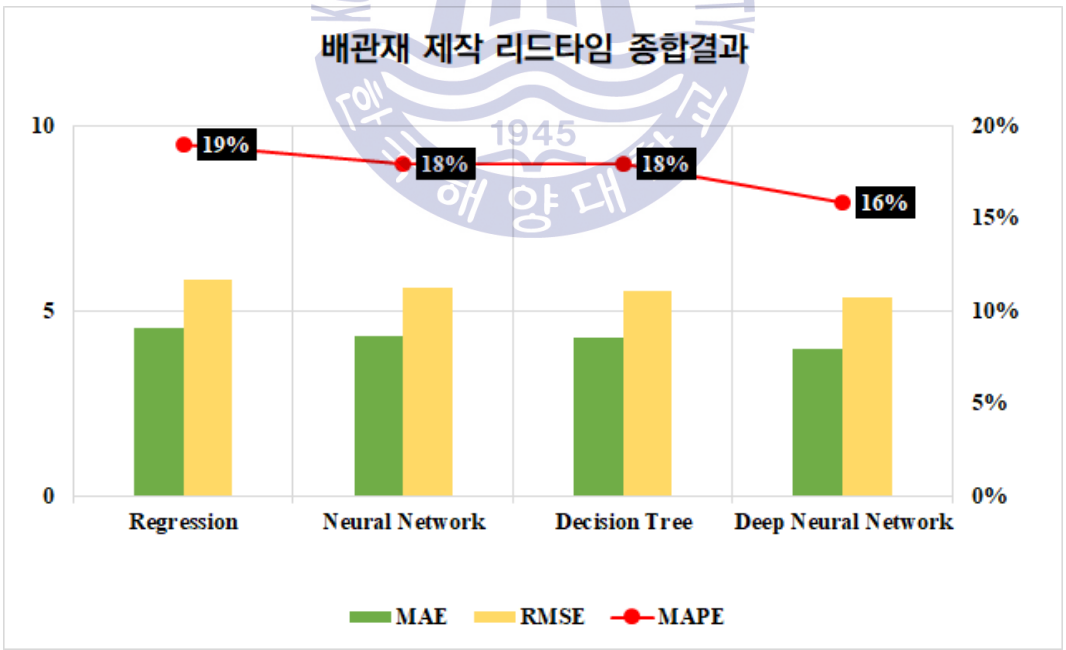


Fig. 44 Final prediction result of spool making lead time

Table 26 Final prediction result of spool painting lead time

	Regression	Neural Network	Decision Tree	Deep Neural Network
MAE	3.78	2.98	3.66	3.23
MAPE	25%	25%	24%	22%
RMSE	5.14	4.36	5.15	5.13

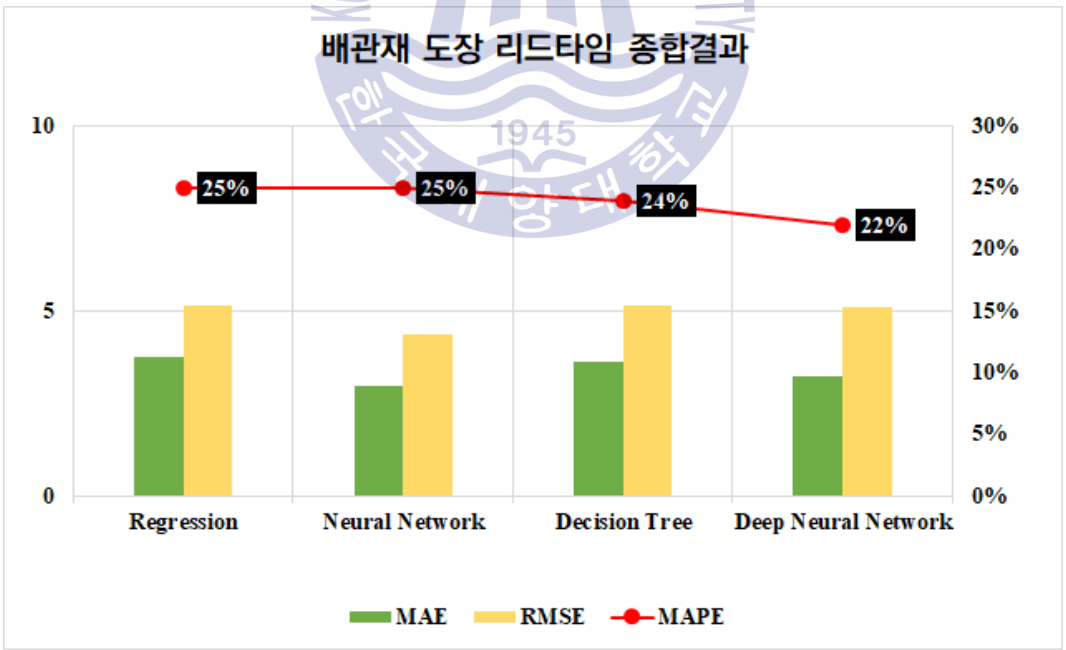


Fig. 45 Final prediction result of spool painting lead time

제 5 장 결 론

5.1 연구결론

본 논문은 조선소에서 관리되는 생산 리드타임이라는 기준정보를 체계적으로 수립하기 위해 빅데이터 분석 방법론을 적용하였다. 조선 생산 관점에서의 리드타임이라는 기준정보는 다른 변수와는 다르게 변동성이 크기 때문에 기존의 엔지니어링 방법론으로는 물량이나 시수 산출에 한계점이 있는 실정이다. 따라서 조선소의 블록 및 배관재의 다양한 제품 속성과 리소스를 고려한 생산 리드타임의 예측모델을 생성하여 생산 환경에 따른 변동성이 심한 기준정보를 개선하고자 조선소의 데이터를 분석하였다. 데이터의 분석 및 예측모델 생성은 R과 Python 등의 오픈소스를 활용하였으며 개발환경에서 제공하는 다양한 기계학습 및 딥러닝 알고리즘을 적용하여 수행할 수 있었다. 본 논문에서 수집한 조선소의 데이터는 총 3가지이며 분석 결과는 다음과 같다.

첫 번째 분석사례는 선박 블록의 절단공정 리드타임을 예측한 연구이다. 절단공정의 중일정 계획에 대한 실적 데이터를 수집하여 알고리즘에 따른 리드타임 예측모델을 생성하였다. MAPE의 평균값은 80% 정도로 나타났으며 딥러닝 알고리즘에서 가장 좋은 예측도를 나타냈다.

두 번째 분석사례는 선박 블록의 탑재공정 리드타임을 예측한 연구이다. 블록의 다양한 공정변수를 수집하여 독립변수로 활용하여 예측모델을 생성한 결과 MAPE의 값은 평균값은 100% 정도로 모든 분석사례에서 가장 높게 나타났으며 알고리즘 측면에서는 딥러닝 알고리즘에서 가장 좋은 예측도를 나타냈다.

세 번째 분석사례는 해양플랜트 배관재의 공급망 리드타임을 예측한 연구이다. 앞의 사례와 동일하게 분석 알고리즘에 따른 예측모델을 생성한 결과 제작 리드타임의 MAPE 평균값은 18%, 도장 리드타임의 MAPE 평균값은 25%로 나타났다.

알고리즘에 따른 분석결과 공정 데이터에 따라 차이는 존재하였지만 모든 리드타임 예측모델에서 딥러닝 알고리즘을 적용한 예측모델의 성능이 우수하다는 것을 검증하였다. 또한 기존연구에 비해 분석과정에서 데이터 전처리를 중점적으로 수행한 결과 예측모델의 오차율이 낮아진 것을 확인하였다. 이를 통해 기존에 관리되고 있는 표준 리드타임과 대비하여 예측 리드타임을 통한 기준정보의 체계적인 관리가 가능할 것이라고 본다. 또한, 작업 계획 시 예측 리드타임의 기준을 통해 빠른 의사결정을 지원할 수 있으며 공정 데이터에 따른 분석 기법 및 변수 설정에 대한 인사이트를 확보할 수 있을 것이다.

5.2 향후 과제

본 연구에서는 생산 리드타임이라는 수치예측을 위한 기계학습 및 딥러닝 알고리즘을 적용하였다. 기계학습 알고리즘은 기본적으로 예측분야에서 많이 활용되지만 딥러닝 알고리즘은 단순한 수치예측 뿐만 아니라 시계열 데이터를 분석하기 위한 다양한 방법론을 제시하고 있다. 따라서 향후에는 조선 생산 관점에서의 다양한 공정별 시계열 예측이 가능하다면 작업 계획 시 더 나은 의사결정이 가능할 것이라 본다.

참고문헌

강만모, 김상락, 박상무, 2012. 빅데이터의 분석과 활용. *정보과학회지*, 30(6), pp.25-32.

김성훈, 노명일, 김기수, 2016. 조선 해양 산업에서의 응용을 위한 하둡 기반의 빅데이터 플랫폼 연구. *한국CDE학회 논문집,* 21(3), pp.334-340.

김연진, 2013. 빅데이터 기반의 고도분석체계 도입을 통한 기업혁신: 사례와 방법론. *ie 매거진,* 20(1), pp.43-49.

김영주 등, 2013. 선박설계 자동화를 위한 빅데이터 기술 및 분석기법 연구. *한국통신학회 학술대회논문집*, pp.213-215.

김의중, 2016. *알고리즘으로 배우는 인공지능, 머신러닝, 딥러닝 입문*. 3nd Ed. 위키북스:파주.

김지원 등, 2015. 다양한 딥러닝 알고리즘과 활용. *정보과학회지*, 33(8), pp.25-31.

문성은, 장수범, 이정혁, 이종석, 2016. 기계학습 및 딥러닝 기술동향. *한국통신학회지(정보와통신)*, 33(10), pp.49-56.

미래창조과학부(NIA), 2016. *2016 글로벌 빅데이터 융합 사례집*, 대구:한국정보화진흥원

서지혜, 용환승, 2016. 텐서플로우를 이용한 LSTM과 GRU의 성능평가. *한국정보과학회 학술발표논문집*, pp.211-213.

우종훈, 남종호, 홍성인, 신종계, 2015. 중소조선소 경쟁력 강화를 위한 생산관리 고도화의 필요성 및 전략. *대한조선학회지*, 52(1), pp.4-13.

우종훈 등, 2015. 중소형조선소 시뮬레이션 기반 생산관리 시스템 개발. *대한조선학회지*, 52(1), pp.22-30.

이재구, 이태훈, 윤성로, 2014. Big Data 분석을 위한 Machine Learning. *한국통신학*

회지(정보와통신), 31(11), pp.14-26.

이훈혜, 2013. *제조업 경쟁력 강화를 위한 빅데이터 활용 방안*, 산업연구원.

장영재, 2012. 제조 분야에서의 빅데이터 기술 활용. *한국통신학회지(정보와통신)*, 29(11), pp.30-35.

정세훈, 심춘보, 2014. 용접 빅데이터 환경에서 상관분석 및 회귀분석을 이용한 작업 패턴 분석 모형에 관한 연구. *한국전자통신학회*, 9(10), pp.1071-1078

조성준, 강석호, 2016. 머신러닝(인공지능)의 산업 응용. *ie 매거진*, 23(2), pp.34-38.

한성호, 김회, 2016. 한국 제조업 빅데이터 도입 사례 연구, *(재)인천테크노파크*.

함동균, 2016. *조선소 의장품 조달관리를 위한 데이터마이닝 방법론에 관한 연구*. 석사학위논문. 부산:한국해양대학교.

함동균, 백명기, 박중구, 우종훈, 2016. 해양플랜트 의장품 조달관리를 위한 배관 공정 리드타임 예측 모델에 관한 연구. *대한조선학회 논문집*, 53(1), pp.29-36.

함동균, 이용길, 우종훈, 2016. 조선소 의장품 조달관리를 위한 빅데이터기반 시뮬레이션 연구. *대한산업공학회 춘계공동학술대회 논문집*, pp.3142-3149.

황규옥, 2001. 디지털 조선소를 위한 생산관리시스템의 발전방향. *대한조선학회지*, 38(1), pp.42-46.

Hur, M.H. et al., 2015. A study on the man-hour prediction system for shipbuilding. *Journal of Intelligent Manufacturing*, 26(6), pp.1267-1279.

Lee, S.K. et al., 2014. Knowledge discovery in inspection reports of marine structures. *Expert Systems with Applications*, 41(4), pp.1153-1167.

Maimon, O., Rokach, L., 2011. *Data Mining and Knowledge Discovery Handbook*, 2nd Ed. Springer US: New York.

부록 A

A.1 데이터 분석 플랫폼

R은 통계분석과 그래픽을 위한 언어이자 환경이라고 정의한다. 수치 분석, 기계학습 분야 개발에서 중요한 도구이며 오픈소스로서 개인부터 기업까지 누구나 쉽게 접근이 가능한 장점이 있다. 특히 많은 수의 분석 패키지를 제공할 뿐만 아니라 자동화된 사용자 정의 함수를 만들어 배포, 공유하기 때문에 어느 상용 툴보다도 빠르고 광범위하게 분석 기능이 확장되고 있다. 또한, 시각화 부문에서도 강력한 그래프 기능을 가지고 있기 때문에 빅데이터 분석을 위해 많은 사용자들이 사용하고 있다.

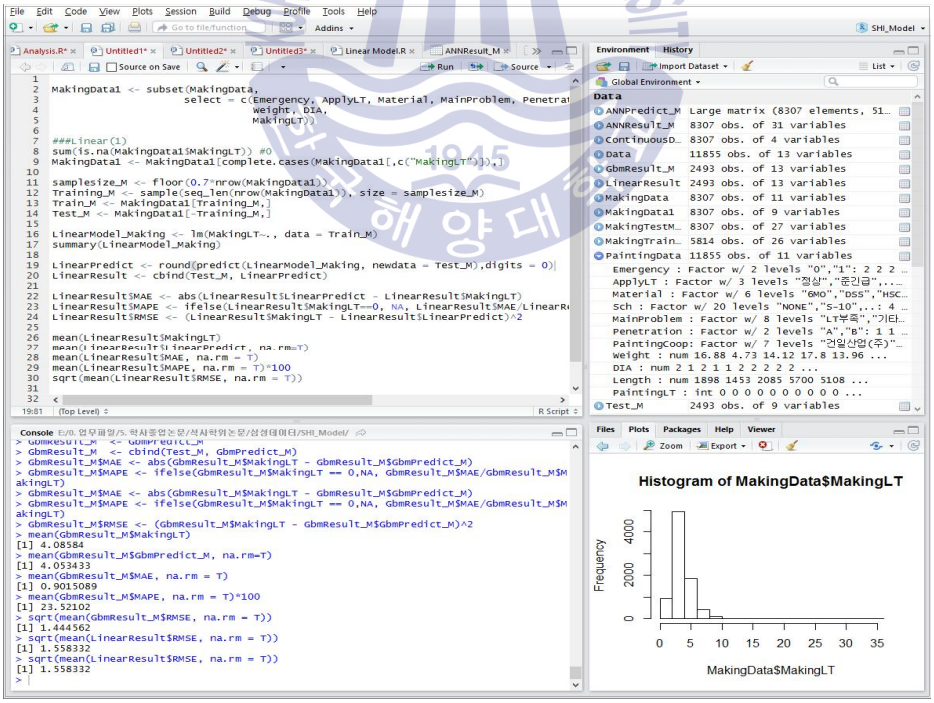
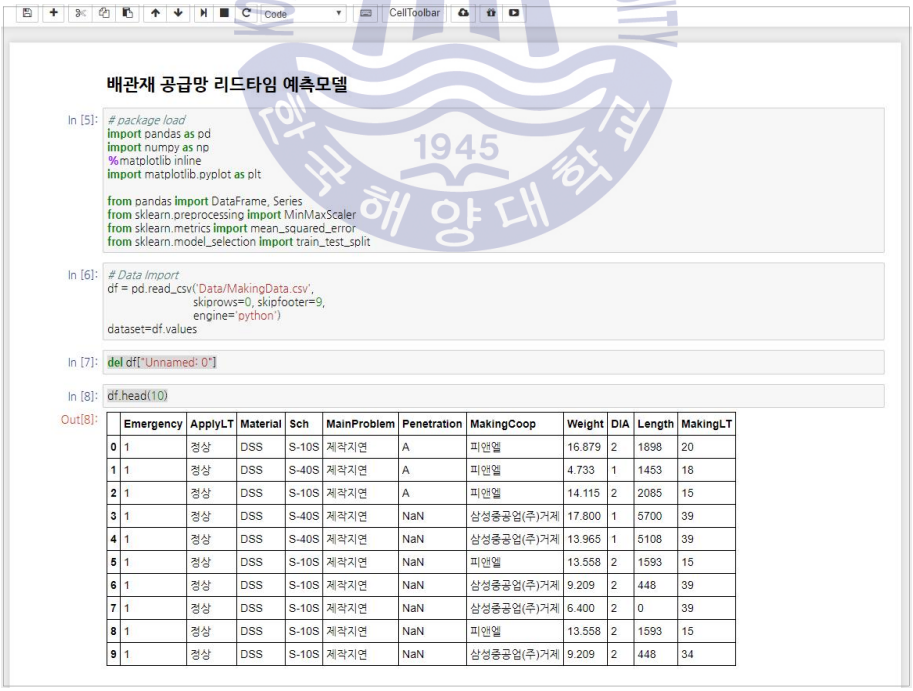


Fig. 46 R studio

딥러닝 알고리즘 구현을 위해서 가장 먼저 오픈소스로 공개되는 다양한 딥러닝 프레임워크를 분석할 필요가 있다. 특히 딥러닝을 구현하기 위해서는 가장 기본적으로 선택해야 하는 것이 어떤 프로그래밍 언어를 사용할 것인지, 컴퓨터 운영체계 등을 판단하는 것이다. 선행연구에서 기본적으로 사용한 R은 통계분석에 최적화된 언어로 머신러닝 알고리즘을 구현하기 위한 패키지 및 라이브러리가 한정적이다. 따라서 딥러닝 구현을 위해서 Python 언어를 활용하였다. Python은 대부분의 라이브러리들이 빠른 속도로 업데이트 되고 특히 구글에서 오픈소스로 공개한 텐서플로우(TensorFlow) 구현이 가능하다는 장점이 있다. 텐서플로우는 머신러닝 및 딥러닝 프레임워크로 CPU와 GPU 모드가 모두 가능하며 강화학습의 알고리즘도 동시에 지원하는 라이브러리이다. 이러한 딥러닝 프레임워크를 Jupyter Notebook 환경에서 구현할 수 있다.



배관재 공급망 리드타임 예측모델

In [5]:
```
# package load
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt

from pandas import DataFrame, Series
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
```

In [6]:
```
# Data Import
df = pd.read_csv('Data/MakingData.csv',
                 skiprows=0, skipfooter=9,
                 engine='python')
dataset=df.values
```

In [7]:
```
del df["Unnamed: 0"]
```

In [8]:
```
df.head(10)
```

Out[8]:

	Emergency	ApplyLT	Material	Sch	MainProblem	Penetration	MakingCoop	Weight	DIA	Length	MakingLT
0	1	정상	DSS	S-10S	체작지연	A	피앤엘	16.879	2	1898	20
1	1	정상	DSS	S-40S	체작지연	A	피앤엘	4.733	1	1453	18
2	1	정상	DSS	S-10S	체작지연	A	피앤엘	14.115	2	2085	15
3	1	정상	DSS	S-40S	체작지연	NaN	삼성중공업(주)거제	17.800	1	5700	39
4	1	정상	DSS	S-40S	체작지연	NaN	삼성중공업(주)거제	13.965	1	5108	39
5	1	정상	DSS	S-10S	체작지연	NaN	피앤엘	13.558	2	1593	15
6	1	정상	DSS	S-10S	체작지연	NaN	삼성중공업(주)거제	9.209	2	448	39
7	1	정상	DSS	S-10S	체작지연	NaN	삼성중공업(주)거제	6.400	2	0	39
8	1	정상	DSS	S-10S	체작지연	NaN	피앤엘	13.558	2	1593	15
9	1	정상	DSS	S-10S	체작지연	NaN	삼성중공업(주)거제	9.209	2	448	34

Fig. 47 Python in Jupyter

A.2 데이터 전처리

데이터 분석 중 가장 많은 시간이 소요되는 단계가 바로 데이터 전처리이다. 일반적으로 데이터 전처리는 데이터의 추가, 삭제, 변환 등의 작업을 의미하는데 데이터 분석가는 업무 시간 중 80% 정도를 데이터 수집 및 전처리 과정에 사용된다고 할 만큼 가장 중요한 단계라고 할 수 있다. 데이터 전처리는 Fig. 48과 같이 크게 데이터 셋 확인, 결측값 처리, 이상값 처리, Feature Engineering의 순서로 진행된다.

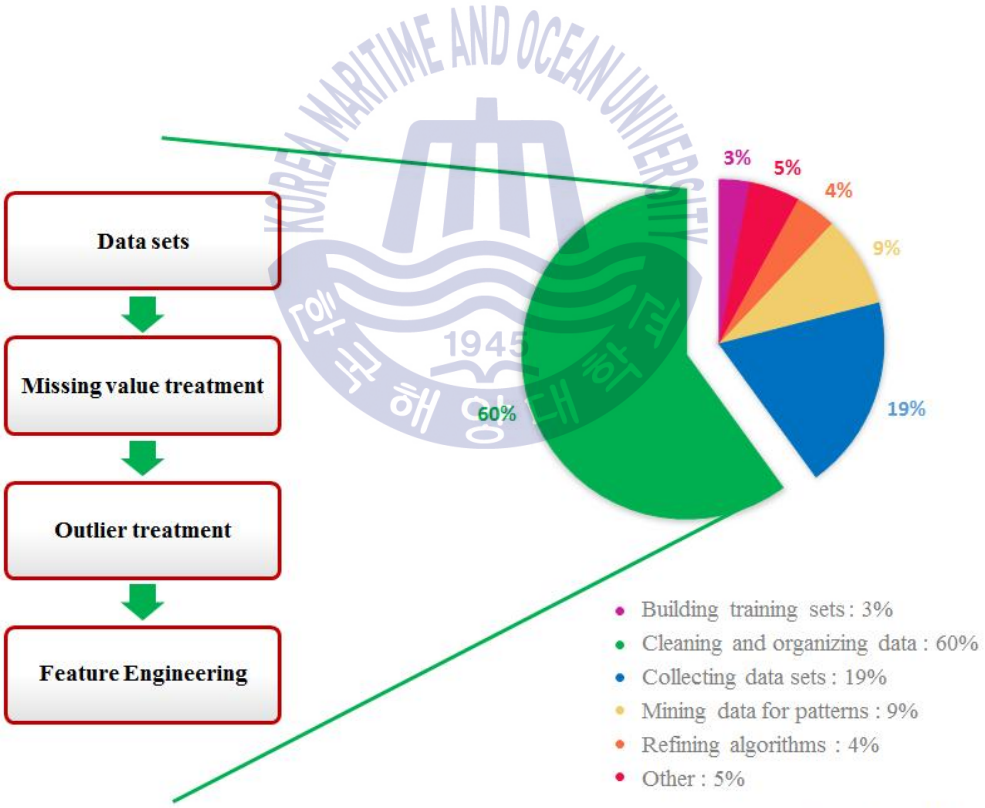


Fig. 48 Data Pre-processing process

A.2.1 데이터 셋 확인

분석하고자 하는 데이터 셋을 확인하는 단계이며 크게 변수와 Raw Data를 확인하는 작업으로 분류된다. 분석을 위한 독립변수와 종속변수의 정의, 각 변수의 유형이 범주형인지 연속형인지, 변수의 데이터 타입이 문자형인지 숫자형인지 등을 확인해야 한다. 분석 알고리즘에 따라 적용되는 데이터 유형이 다르며 예측에 있어서도 결과에 영향을 미치기 때문에 가장 우선적으로 수행해야하는 단계이다.

Raw Data를 확인하는 작업은 단변수와 다변수를 분석하는 것이다. 단변수란 변수 하나를 의미하는 것으로 일반적으로 Histogram이나 Boxplot을 사용하여 변수의 평균, 최빈값, 중간값 등의 다양한 분포를 확인하게 된다. 변수 2개 사이의 관계를 분석하기 위해서는 이변수 분석이 활용된다. 두 변수 간의 관계는 Table 27과 같이 데이터 유형에 맞는 그래프와 분석 방법을 적용할 수 있다.

Table 27 Method of analysis by data type

	그래프	분석방법
연속형 X 연속형	- Scatter plot	- Correlation 분석 (두 변수 간 상관관계 분석)
범주형 X 범주형	- 누적 막대그래프 - 100%기준 누적 막대그래프	- Chi-Square 분석 (두 변수가 독립적인지 파악)
범주형 X 연속형	- 누적 막대그래프 - 범주 별 Histogram	범주의 종류에 따라 - 2개: T-test/Z-test - 3개 이상: ANOVA (집단 별 평균 차가 유의한지 파악)

A.2.2 결측값 처리 (Missing value treatment)

결측값이란 측정한 데이터 중 몇몇 변수들의 값이 측정되지 못한 경우를 의미하는데 이러한 결측값이 있는 상태로 모델을 만들 경우 변수 간의 관계가 왜곡될 수 있기 때문에 모델의 정확성이 떨어지게 된다. 결측값이 발생하는 유형은 무작위로 발생하느냐, 결측값의 발생이 다른 변수와 관계와 있는지 여부에 따라 처리하는 방법이 다르다.

가장 단순한 방법은 결측값이 발생한 모든 관측치를 삭제하거나 (전체 삭제, Listwise Deletion), 데이터 중 모델에 포함시킬 변수들 중 관측값이 발생한 모든 관측치를 삭제하는 방법(부분 삭제)이 있다. 전체 삭제는 간편한 반면 관측치가 줄어들어 모델의 유효성이 낮아질 수 있고, 부분 삭제는 모델에 따라 변수가 제각각 다르기 때문에 관리 Cost가 늘어난다는 단점이 있다. 데이터를 삭제하게 되면 좀 더 간결하고 해석하기 쉬운 모델을 만들 수 있으며 성능 및 안정성이 향상될 수 있으나 결측값이 무작위로 발생한 것이 아닌데 관측치를 삭제한 데이터를 사용할 경우 왜곡된 모델이 생성될 수 있다. 따라서 일반적으로 삭제는 결측값이 무작위로 발생한 경우에 사용된다.

결측값이 발생한 경우 다른 관측치의 평균, 최빈값, 중간값 등으로 대체할 수 있는데, 모든 관측치의 평균값 등으로 대체하는 일괄 대체 방법과, 범주형 변수를 활용해 유사한 유형의 평균값 등으로 대체하는 유사 유형 대체 방법이 있다. 결측값의 발생이 다른 변수와 관계가 있는 경우 대체 방법이 유용한 측면은 있지만, 유사 유형 대체 방법의 경우 어떤 범주형 변수를 유사한 유형으로 선택할 것인지는 자의적으로 선택하므로 모델이 왜곡될 가능성이 존재한다.

결측값이 없는 관측치를 트레이닝 데이터로 사용해서 결측값을 예측하는 모델을 만들고, 이 모델을 통해 결측값이 있는 관측 데이터의 결측값을 예측하는 방법으로 Regression이나 Logistic regression을 주로 사용한다.

대체하는 방법보다 조금 덜 자의적이나, 결측값이 다양한 변수에서 발생하는 경우 사용 가능 변수 수가 적어 적합한 모델을 만들기 어렵고, 또 이렇게 만들어 진 모델의 예측력이 낮은 경우에는 사용하기 어려운 방법이다.

A.2.3 이상값 처리 (Outlier treatment)

이상값이란 데이터/샘플과 동떨어진 관측치로, 모델을 왜곡할 가능성이 있는 관측치를 말한다. 이상값을 찾아 기 위한 쉽고 간단한 방법은 변수의 분포를 시각화하는 것이다. 일반적으로 하나의 변수에 대해서는 Boxplot이나 Histogram을, 두개의 변수 간 이상값을 찾기 위해서는 Scatter plot을 사용한다. 시각적으로 확인하는 방법은 직관적이지만 자의적이기도 하고 하나 씩 확인해야 해서 번거로운 측면이 있다.

두 변수 간 이상값을 찾기 위한 또 다른 방법으로는 두 변수 간 회귀 모형에서 Residual, Studentized residual(혹은 standardized residual), leverage, Cook's D값을 확인하면 된다. 이상값이 Human error에 의해서 발생한 경우에는 해당 관측치를 삭제하면 됩니다. 단순 오타나, 주관식 설문 등의 비현실적인 응답, 데이터 처리 과정에서의 오류 등의 경우에 사용합니다.

절대적인 관측치의 숫자가 작은 경우, 삭제의 방법으로 이상치를 제거하면 관측치의 절대량이 작아지는 문제가 발생한다. 이런 경우 이상값이 Human error에 의해 발생했더라도 관측치를 삭제하는 대신 다른 값(평균 등)으로 대체하거나, 결측값과 유사하게 다른 변수들을 사용해서 예측 모델을 만들고, 이상값을 예측한 후 해당 값으로 대체하는 방법도 사용할 수 있다.

이상값이 자연 발생한 경우, 단순 삭제나 대체의 방법을 통해 수립된 모델은 설명/예측하고자 하는 현상을 잘 설명하지 못할 수도 있다. 자연발생적인 이상값의 경우, 바로 삭제하지 말고 좀 더 찬찬히 이상값에 대해 파악하는 것이 중요하다. 자연 발생한 이상값을 처리하는 또 다른 방법으로는 해당 이상값을 분리해서 모델을 만드는 방법이 있다.

A.2.4 Feature Engineering

Feature Engineering이란, 기존의 변수를 사용해서 데이터에 정보를 추가하는 일련의 과정으로 새로 관측치나 변수를 추가하지 않고도 기존의 데이터를 보다 유용하게 만드는 방법론이다(Fig. 49).

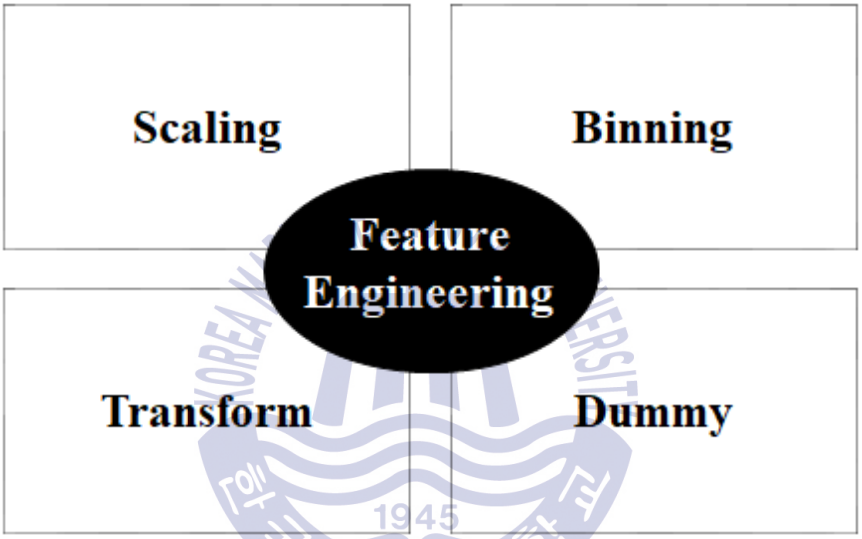


Fig. 49 Feature Engineering

Scaling은 변수의 단위를 변경하고 싶거나, 변수의 분포가 편향되어 있을 경우, 변수 간의 관계가 잘 드러나지 않는 경우에는 변수 변환의 방법을 사용한다. 가장 자주 사용하는 방법으로는 Log 함수가 있고, 유사하지만 좀 덜 자주 사용되는 Square root를 취하는 방법도 있다.

Binning은 연속형 변수를 범주형 변수로 만드는 방법이다. 예를 들어 나이라는 연속형 변수가 존재하는 경우 10대, 20대, 30대 등의 식으로 범주형 변수로 변환할 수 있다. Binning에는 특별한 원칙이 있는 것이 아니기 때문에, 분석가의 Business 이해도에 따라 창의적인 방법으로 수행할 수 있다.

Transform은 기존에 존재하는 변수의 특징을 이용해 다른 변수를 만드는 방법이다. 예를 들어 일자별 강수량 데이터를 기간별로 나눈 새로운 변수로 만들거나 시계열 데이터로부터 새로운 수치 데이터를 추가하는 것을 의미한다. Transform에도 Binning과 마찬가지로 특별한 원칙이 있는 것은 아니기 때문에 사용자의 이해도에 따라 수행할 수 있는 방법이다.

Dummy는 Binning과는 반대로 범주형 변수를 연속형 변수로 변환하기 위해 사용된다. 사용하고자 하는 분석 방법론에서 필요한 경우에 주로 사용되며 원-핫(One-hot) 인코딩이라고도 한다. 이는 범주형 변수를 해당하는 칸의 정보를 1로 나머지를 0으로 표시하는 방법으로 주로 사용되며 대부분의 기계학습 및 딥러닝 알고리즘에서는 모든 데이터를 숫자로 넣어주어야 하기 때문에 가장 많이 사용되는 방법이라고 할 수 있다.

A.3 예측모델 평가지표

예측모델의 성능평가 및 결과를 비교하기 위한 지표로 MAE, MAPE, RMSE, Adjusted R-squared를 활용하였다. 해당 지표들로 모델의 예측값과 실적값을 비교하여 어느 정도의 오차를 가지고 있는지를 나타내거나 회귀모형이 어느 정도의 설명력을 갖고 있는지 평가할 수 있다.

먼저 MAE(Mean Absolute Error)는 평균절대오차로 예측값과 실적값의 차이를 절댓값으로 나타낸 것을 의미하며 식 (1)에 나타난 듯이 예측값 y_i와 실적값 $\hat{y_i}$의 차이에 대한 절댓값으로 계산된다.

$$MAE = \left| y_i - \hat{y_i} \right| \tag{1}$$

MAPE(Mean Absolute Percentage Error)는 평균절대오차비율로 예측값과 실적값의 사이의 오차를 백분율로 환산한 값이다. 따라서 MAPE값이 작을수록 오차율이 작다는 것을 의미하며 계산식은 식 (2)와 같다.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| (y_i - \hat{y_i}) / \hat{y_i} \right| \tag{2}$$

RMSE(Root Mean Square Error)는 평균제곱근오차로 보통 정밀도라고 표현되는 값이며 잔차들을 하나의 척도로 종합할 때 사용된다. RMSE값이 작을수록 정밀도가 높다는 것을 의미하며 계산식은 식 (3)과 같다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{n}} \tag{3}$$

결정계수란 적합도를 평가하기 위한 척도로 회귀분석으로 만든 모형이 실제 데이터에 얼마나 적용되었는지를 판단할 수 있다. 일반적으로 독립변수의 수가 증가하게 되면 결정계수가 증가하는데 이때 결정계수의 값은 종속변수에 영향을 미치지 않는 즉, 유의하지 않는 독립변수가 적용되어도 더 높은 값을 나타낼 수도 있다. 따라서 다중회귀분석에서는 독립변수의 수가 다른 모형들을 비교하기 위해 수정된 결정계수(Adjusted R-squared)를 사용한다. n을 표본의 개수, p를 독립변수의 개수라고 했을 때 수정된 결정계수는 식 (4)와 같으며 결정계수보다 항상 작은 것이 특징이다.

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) \tag{4}$$