 **디지탈쉽**

**Ai 기반 자율 데이터 전처리 기술**

**IND-DSHIP-CM-02-006**

|  |
| --- |
| 문서 정보 |

문서 개요

이 문서는 사용자가 인공지능(ai)를 기반으로 하여 데이터를 자동적으로 전처리할 수 있는 기술에 대해 설명하는 문서입니다

담당

오무송 소장 (msoh@digitalship.co.kr)

권진수 연구원(jskwon@digitalship.co.kr)

이정준 연구원 ([jjlee@digitalship.co.kr](mailto:jjlee@digitalship.co.kr))

문서 버전 및 이력

|  |  |  |  |
| --- | --- | --- | --- |
| 버전 | 일자 | 이력사항 | 작성자 |
| 1.0 | 2023-10-24 | 초안 작성 | 이정준 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

|  |
| --- |
| 목차 |

1. [ai 기반 자동데이터 전처리 소개](#_Toc124783216) 3

[1.1 자동 데이터 전처리기술의 배경](#_Toc124783217) 3

1.2 자동 데이터 전처리기술의 필요성 4

[1.3 국내외 동향](#_Toc124783218) 5

[2. 전처리 방법](#_Toc124783219) 7

[2.1 데이터 상태 측정 알고리즘 모듈화](#_Toc124783220) 7

2.2 전처리 기법 8

3. 데이터 상태 측정 및 imputation 모듈 기반 전처리 자동화기술 14

3.1 전문가 provenance 학습모델 14

3.2 전처리 자동화 기술 및 소스 code 예시 16

4. 기술 검증 및 평가 22

# Ai 기반 자동데이터 전처리 소개

Ai 기반 자동 데이터 전처리는 인공지능 기술을 활용해 데이터 전처리 과정을 자동화하는 방법이다. 데이터 전처리는 일반적으로 머신러닝, 데이터 분석 작업을 수행하는 초기 단계에서 주로 이루어지며, 원시 데이터를 분석, 모델링 목적에 맞게 적합한 형태로 변환하는 과정을 포함한다. 데이터 전처리는 어떤 식으로 작업이 이루어지느냐에 따라서, 원시 데이터의 품질 향상, 모델 성능 개선, 비용 절감 등 분석 및 모델링 작업에서 핵심적인 요소이다.

## **자동 데이터 전처리 기술의 배경**

* 빅데이터 분석 및 인공지능 모델링 업무를 수행하는데 있어서 데이터 전처리는 선행되어져야한다. 다음은 최근 우리나라의 전처리 과정을 진행하는데 있어서 발생하는 문제점들을 서술한 내용이다.

1. **많은 양의 데이터 처리**   
   21세기 현대 사회에서는 빅데이터 시대라고 할 수 있을만큼. 막대한 양의 데이터가 생성되고 있다. 이러한 빅데이터를 효과적으로 처리하기 위해서는 많은 시간과 비용을 필요로 한다.
2. **전처리 과정의 복잡성**   
   후술하겠지만, 데이터 전처리는 결측치 처리, 이상치 탐지 및 제거, 변수 변환, 범주형 변수 인코딩 등 다양한 작업을 포함하는 복잡한 과정이다. 각각의 작업은 특별한 기술과 지식을 요구하며, 잘못된 처리는 분석 결과에 큰 영향을 미칠 수 있었다.

Ai 기반 자동 데이터 전처리는 서술한 문제점들을 해결하기 위해서 개발되었다. 기존에는 사용자가 직접 데이터를 확인하고, 그에 맞는 전처리 작업을 수행하였지만, 이제는 인공지능 알고리즘이 데이터를 파악하여 적절한 전처리 작업을 선택하거나, 파라미터를 학습하고 적용해 원시데이터를 효율적으로 처리할 수 있게 되었다.

* 또, 데이터 자동화 기술을 도입할 시, 나타날 수 있는 긍정적인 효과는 다음과 같다.

1. **분석결과의 일관성 유지**   
   수작업으로 전처리를 진행하면 사람마다 방법이나 기준이 달라질 수 있어 결과에 일관성이 부족할 수 있다. 반면, 자동화 도구를 사용하면 동일한 기준과 절차로 일관된 결과를 얻을 수 있다.
2. **AI와 머신러닝 발전**   
   AI와 머신러닝 분야의 발전으로 인해 알고리즘이 스스로 최적의 전처리 방법을 학습하고 적용할 수 있는 가능성이 열렸다.
3. **분석 모델 성능 향상**   
   원시 데이터 그대로 사용하는 것보다 적절히 전처리된 데이터를 사용하면 분석 모델의 성능을 크게 향상시킬 수 있다.

## **적절한 전처리 자동화 기술의 필요성**

* 데이터는 인공지능(AI)의 핵심 원천이며, 양질의 데이터는 AI의 성능과 결과에 직결된다. 그러나 '양질의 데이터'라는 것은 단순히 오류가 없고 완전한 정보만을 의미하는 것이 아니다. 올바른 결과를 얻기 위해서는 양질의 데이터 뿐만 아니라 적절한 전처리 과정도 필요하다. 이런 의미에서 "Trash in, trash out(쓰레기를 넣으면 쓰레기가 나온다)"라는 말이 주목할 만한 의미를 갖게 되었다.   
    
  전처리 과정은 원시 데이터를 분석 가능한 형태로 가공하는 작업이다. 결측치 처리, 이상치 탐지 및 제거, 변수 변환 등 다양한 작업을 통해 데이터 품질을 향상시키고 분석에 적합하게 만든다. 왜냐하면 AI 모델은 학습 데이터에 기반하여 패턴을 학습하고 예측을 수행하기 때문이다.하지만 여기서 중요한 점은 전처리 과정에서 소요되는 시간과 비용이다. 대규모의 데이터셋을 수작업으로 처리하는 것은 상당한 시간과 비용이 들게 된다. 이러한 비용 부담과 함께 인력 한계로 인해 전체적인 전처리 작업이 소홀히 될 수 있었다.그래서 자동화된 전처리 기술의 필요성을 강조해야 한다. 자동화된 도구와 기법은 일관성 있고 효율적인 전처리 프로세스를 구축할 수 있으며, 시간과 비용을 절감할 수 있다. 더욱 중요한 것은 자동화된 방식으로 진행되면서 발생할 수 있는 인간 실수나 주관적인 판단으로 인한 편향성도 최소화될 수 있다는 점이다.

<그림 > GIGO를 설명하는 자료

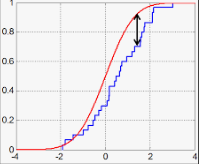
## **국내외 전처리 관련 기술의 동향**

* 국내 기술 동향 및 수준
  + 와이즈프로핏(wiseprophet) : 와이즈 프로핏은 인공지능과 머신러닝 기술을 활용한 데이터 분석 및 모델 설계 전문 기업으로 알려져 있다. 그들의 주력 제품인 AutoML 기반 예측모델 자동화 플랫폼은 다양한 산업 분야에 큰 영향을 미치고 있다.이 플랫폼은 예측 정비, 이상 거래 탐지 등의 핵심적인 역량을 갖추고 있어, 다양한 상황에서 사용자에게 깊이 있는 인사이트를 제공한다. 이는 복잡한 데이터 패턴을 파악하고, 예측 가능한 결과를 도출해내는 데 중요한 역할을 한다.또한, 공공 부문부터 금융, 제조업, 유통업, 미디어 등 다양한 산업에 적용 가능하도록 설계되어 있다.
  + 빅스테이션 : 비정형 데이터에 대한 깊이 있는 이해를 가능하게 하는 특별한 기술을 보유하고 있다. 그들의 솔루션은 다양한 비정형 데이터에서 발생하는 트렌드, 패턴, 시간 빈도, 연관 관계 등의 복잡한 요소들을 체계적으로 분석한다. 이를 통해 비즈니스 환경의 현재 상황을 정확하게 파악하고, 그에 따른 합리적인 의사결정 지원이 가능하다.
  + bigin : bigin은 e-commerce 웹사이트의 데이터를 활용하여 가치 있는 인사이트를 제공하는 서비스로 알려져 있다. 웹사이트 방문자들의 행동 패턴을 꼼꼼히 분석하여 잠재 고객 추천 및 마케팅 전략 개발에 도움을 준다. 사용자 경험이 중요한 e-commerce 산업에서 bigin의 솔루션은 고객 관계 관리와 맞춤형 서비스 제공에 필수적인 역할을 한다
* 국외 기술 동향 및 수준
  + 미국 : 주요 대학과 연구기관들이 AI를 활용한 데이터 전처리 자동화에 관심을 갖고 있고, 동시에 산업계에서도 많은 스타트업들이 해당 분야에서 혁신적인 솔루션을 제공하려는 시도를 하고 있다. 예를 들어보자면, 미국의 경우 Trifacta라는 회사가 대표적이다. Trifacta는 사용자 중심의 인터페이스와 AI를 활용하여 데이터 전처리 작업을 간소화하고 자동화한다. 그들의 솔루션은 비즈니스 사용자와 데이터 과학자 모두가 쉽게 사용할 수 있도록 설계되었다.
  + 일본 : ai와 관련된 최신 기술 개발에 투자하는 국가이다. 일본의 주요 IT기업들은 데이터 전문기관(DMA) 설립 등으로 실무자 양성 및 컴퓨터 비전, 음성 처리 자율 주행 등 다양한 분야에서 AI와 함께하는 실제 사례를 보여주고 있다. 예를 들면, RPA(Robotic Process Automation) 회사인 UiPath가 주목받고 있다. UiPath는 업무 프로세스를 자동화하기 위해 로봇 소프트웨어를 활용하지만, 이것은 궁극적으로 많은 양의 비구조화된 데이터 처리와 관련있다. 따라서 UiPath의 RPA 솔루션은 일부 상황에서 복잡한 데이터 전처리 작업을 대체할 수 있는 잠재력이 있다.
  + 중국 : 인공지능 분야에 큰 힘이 되기 위해 지속적인 노력을 기울이고 있다. 많은 스타트업과 대형 IT 기업들이 AI를 활용하여 데이터 전처리 영역에서 혁신적인 솔루션과 서비스를 제공하려는 경쟁력있는 시장이다. 4Paradigm이 유명하다. 4Paradigm은 AI 및 머신러닝 솔루션을 제공하며, 그 중 하나인 'AutoML' 플랫폼은 자동 피처 엔지니어링 기능을 포함하여 데이터 전처리를 자동화한다. 이 회사의 목표는 AI 기술을 보다 접근 가능하게 만드는 것으로, 데이터 전처리 자동화 기술은 그 목표 달성에 핵심적인 부분이다.

# 2. 전처리 방법

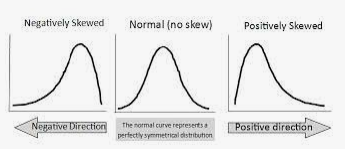
## **데이터 상태 측정 알고리즘 모듈화**

* 데이터 전처리 필요성 측정을 위한 상태 측정 알고리즘들

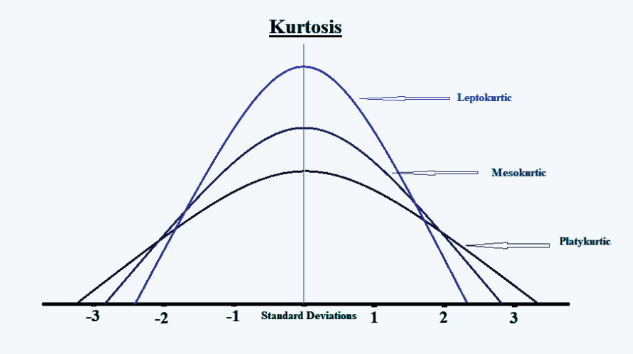


<그림 > Kolmogorov-Smirnov Test

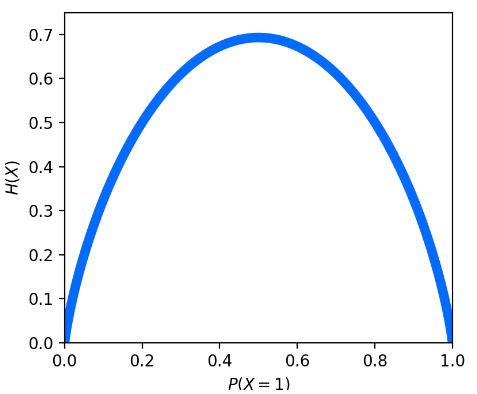
Kolmogorov-Smirnov Test: 두 확률 분포가 같은지를 비모수적으로 검증하는 방법이다. 이 방법은 각 분포의 누적 분포 함수(CDF)를 비교하여 그 차이가 유의한지를 판단한다. 특히 데이터 전처리 과정에서는 이 검정을 사용하여 변환된 데이터가 정규분포와 같은 특정 분포를 따르는지 확인한다. 예를 들어, 로그 변환 등의 방법을 사용해 데이터를 정규화하려는 경우, 해당 검정을 통해 변환 후의 데이터가 실제로 정규분포에 가까워졌는지 확인할 수 있다.

 Skewness: 왜도(skewness)는 확률 분포의 비대칭성을 나타내는 척도다. 왜도 값이 0인 경우, 해당 분포는 완전한 대칭성을 가진 것으로 간주된다. 양수 값은 오른쪽 꼬리(즉, 보다 큰 값)가 더 길고, 음수 값은 왼쪽 꼬리(즉, 보다 작은 값)가 더 길다는 것을 의미한다. 데이터 전처리 과정에서 왜도 계산은 중요한데, 그 이유는 대부분의 기계 학습 알고리즘이 입력 데이터가 대칭적인 정규분포에 가깝게 가정되기 때문이다.

<그림 > 그래프의 왜도

 Kurtosis: 첨도란 확률 분포의 '뾰족함' 혹은 꼬리 부분의 굵기를 측정하는 지수다. 첨도값이 3보다 큰 경우, 해당 분포는 '뚱뚱하며' 긴 꼬리를 가진 것으로 간주되며 이러한 형태의 분산과 함께 이상치 발생 가능성이 커진다. 반면 첨도값이 3보다 작으면 해당 분포는 '마른' 상태로 평탄하며 짧은 꼬리를 가진 것으로 간주된다. 첨도 계산은 데이터의 이상치 존재 여부와 분포 형태를 이해하는 데 중요한 도구가 될 수 있다.

<그림 > kurtosis 그래프

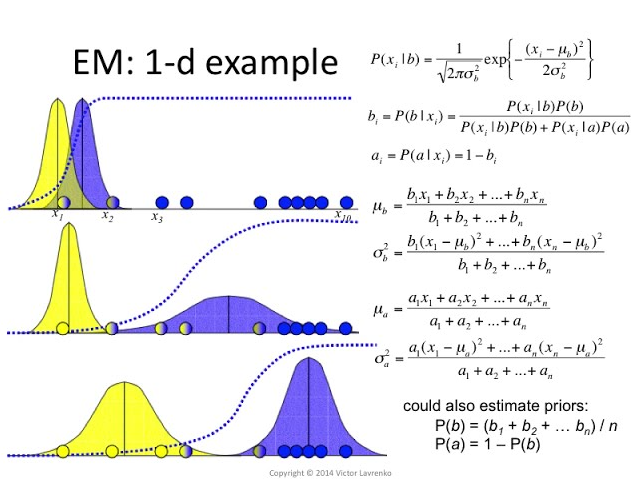
 Information Entropy: 정보 엔트로피는 주어진 데이터 세트의 복잡성 혹은 불확실성을 측정하는 방법이다. 이는 정보 이론에서 유래된 개념으로, 각각의 가능한 사건이 발생할 확률에 기반한다. 엔트로피 값이 클수록 해당 데이터 세트의 복잡성이 커지며 예측하기 어려워진다는 것을 의미한다. 따라서, 정보 엔트로피는 특히 분류 문제에서 클래스 불균형을 판단하거나, 결정 트리 알고리즘에서 최적의 분기점을 찾는 등 다양한 상황에서 사용될 수 있다.

<그림 > information entropy 그래프

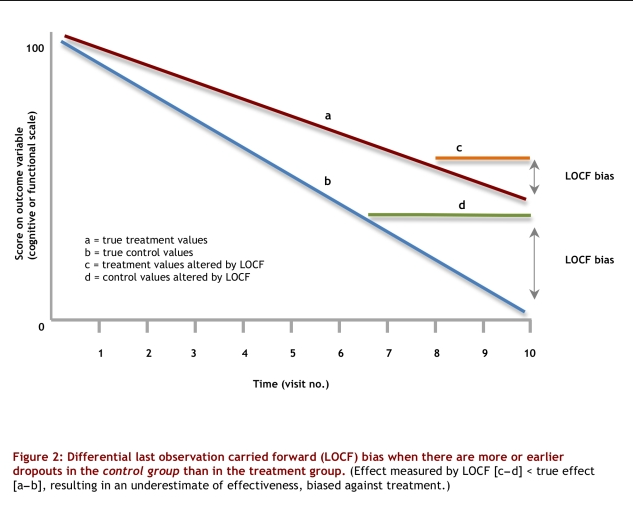
## **전처리 기법**

* 원시 데이터에서 전처리가 필요한 데이터들은 결측치(missing value), 이상치(outlier), 불일치 값(inconsistent value)이 존재한다. 결측치는 데이터 세트에서 값이 누락된 상태를 의미한다. 이는 데이터 수집 과정에서 오류, 정보의 부재 등 다양한 원인으로 발생할 수 있다. 결측치가 많은 경우, 데이터의 통계적 속성을 왜곡시키거나 분석 결과를 불완전하게 만들 수 있다. 이상치는 기대되는 범위에서 크게 벗어난 값이다. 이상치는 측정 오류 또는 실제 데이터 분포의 극단적인 변동으로 발생할 수 있다. 이상치가 포함된 데이터로 모델을 학습시키면 모델이 일반적인 패턴보다 이러한 극단적인 값을 과도하게 반영할 위험이 있어 결과에 부정적인 영향을 줄 수 있다. 예를 들어, 20대 남성의 신장을 조사한 데이터가 있는데, 통상 170~185사이의 분포를 보이는데, 간혹 가다 신장이 200이 넘는 데이터가 존재한다면, 이것이 이상치라고 할 수 있겠다. 다음으로는 불일치 값이다. 데이터 세트 내에서 일관성이 없는 값들을 의미한다. 이는 동일한 변수에 대해 다른 형식, 단위, 혹은 표현 방식이 사용되었을 때 발생한다.예를 들어, 날짜 데이터에서 일부는 '2023-10-25'의 형식으로 저장되어 있고, 일부는 '10/25/2023'의 형식으로 저장되어 있을 경우 이를 'inconsistent value'라고 할 수 있다. 또 다른 예로는 성별 데이터에서 '남성', '남', 'M' 등 다양한 표현이 사용되었을 때 이를 일관성 없는 값이라고 할 수 있다. 이런 일관성 없는 값은 데이터 분석을 어렵게 만들며, 이를 해결하기 위해 데이터 전처리 과정에서 일관성 있는 값으로 통일시키는 작업이 필요하다.

|  |  |  |
| --- | --- | --- |
| **값** | **정의** | **영향** |
| 결측치 | 알려지지 않고, 수집되지 않거나 잘못 입력된 데이터 세트의 값 | 데이터 통계적 속성과 분석 결과 왜곡 |
| 이상치 | 기대값의 범위에서 크게 벗어난 값. | 일반적으로 갖고 있는 패턴에 비해 극단적으로 값을 학습(과적합) |
| 불일치 값 | 데이터 표준 값의 형식과 상이한 형식을 가지는 값 | 데이터 분석 작업 어려움 |

* 이러한 3가지 값은 그에 맞는 적절한 전처리 프로세스를 거쳐야 한다. 데이터 분석 및 모델 설계 과정에서 전처리 방법은 정해져 있지 않다. 프로젝트 및 연구과제 목적이 다양하듯이 그에 따라 유연하게 수행되어져야 한다. 이 3가지 값 발생 상황외에도 데이터와 관련하여 여러 문제가 있을 수 있다. 이것을 데이터 품질 이슈가 발생했다고 볼 수 있다. 각 데이터별 데이터 품질 이슈를 해결할 수 있는 전처리 기법들을 알아보자. 데이터 무시, 데이터 삭제, 데이터 대체 등 많은 기법이 존재하는데, 해당 기술문서에서는 이 3가지를 위주로 설명한다.
  + 데이터 무시 : 데이터를 그대로 두는 형태. 처리하기가 간편하며, 데이터의 기존 분포를 유지한다는 장점이 있음.
  + 데이터 삭제 : 데이터 상에 결측치가 있는 로우나 컬럼을 전체적으로 제거하는 방법이다. 데이터가 없는 값들을 삭제하는 방법의 주요 강점은 간단하고 직관적인 방법인데, 분석에 사용되는 데이터의 손실을 최소화할 수 있다는 점이다. 그러나 결측치가 너무나도 많은 경우에는 데이터 손실이 크게 발생하고, 습득한 데이터의 양이 보다 많이 줄어들어 왜곡된 결과를 야기하게 된다.
  + 데이터 대체
    - 평균값 대체 : 결측치를 처리하는 가장 기본적인 방법 중 하나로, 해당 변수의 평균값으로 결측치를 대체하는 방식이다.
      * 특징 : 이 방법의 가장 큰 특징은 간단하고 직관적이라는 점이다. 또한, 이 방법을 사용하면 원래 데이터의 평균과 분산을 유지할 수 있어 데이터의 전반적인 분포를 왜곡하지 않는다.
      * 강점 : 평균값 대체 방법이 강점을 보이는 상황은 결측치가 무작위로 발생하고, 결측치의 비율이 상대적으로 낮을 때이다. 이런 경우, 평균값 대체 방법은 전체 데이터의 패턴을 잘 유지해준다.
      * 단점 : 첫째, 해당 변수의 분산을 과소평가할 수 있다. 이는 평균값이 결측치를 대체하게 되면, 그 변수의 분산이 실제보다 작게 계산될 수 있기 때문이다. 둘째, 데이터가 극단값을 가지고 있거나, 왜곡된 분포를 가질 때 평균값 대체 방법은 심각한 왜곡을 초래할 수 있다. 즉, 데이터의 분포가 대칭적이거나, 중앙에 몰려 있지 않은 경우 평균값 대체는 적절하지 않다. 마지막으로, 결측치가 특정 패턴을 가지고 발생하는 경우(예를 들어, 특정 그룹에서만 결측치가 발생하는 경우 등)에도 이 방법은 적절하지 않다. 이런 상황에서 평균값 대체를 사용하면, 데이터의 편향을 증가시킬 수 있기 때문이다.
    - 중간값 대체 : 결측치를 해당 변수의 중간값으로 대체하는 방식이다.
      * 특징 : 이 방법은 특히 이상치의 영향을 크게 받지 않으며, 데이터의 분포가 왜곡되어 있거나 극단값이 있을 때 유용하다. 중간값 대체 방법의 주요 특징은 데이터의 중심 경향성을 잘 나타내주는 특성 때문에 이상치의 영향을 상대적으로 덜 받는다는 점이다. 이는 평균값 대체 방법이 이상치에 취약한 반면, 중간값은 데이터의 중앙에 위치하기 때문에 이상치의 영향력을 줄여준다.
      * 강점 : 중간값 대체 방법이 강점을 보이는 상황은 변수의 분포가 왜곡되어 있거나, 극단값이 존재하는 경우이다. 이런 상황에서 평균값 대체 방법을 사용하면, 대체된 값이 데이터의 왜곡을 더욱 심화시킬 수 있다. 그러나 중간값은 데이터의 중앙에 위치하므로 이런 극단값 문제에 영향을 받지 않는다.
      * 단점 : 첫째, 이 방법은 결측치의 발생 원인이나 패턴을 고려하지 않는다. 예를 들어, 결측치가 특정 그룹에서 주로 발생한다면, 그 그룹의 특성을 반영하지 못하고 중간값으로 단순히 대체해버리는 문제가 있다. 둘째, 결측치가 많을 경우 데이터의 분포를 왜곡시킬 수 있다. 중간값은 원래의 데이터 분포를 유지하는 데 도움이 되지만, 결측치가 너무 많을 경우에는 오히려 데이터의 분포를 왜곡시키는 결과를 초래할 수 있다.
    - 최빈값 대체 : 결측치를 해당 변수의 최빈값으로 대체하는 방식이다. 이 방법은 주로 범주형 데이터에서 사용되며, 데이터의 대표값을 결측치에 대입해 데이터의 일관성을 유지하고, 결측치로 인한 정보 손실을 최소화하려는 목적으로 사용된다.
      * 특징 : 이 방법은 특히 범주형 데이터에서 유용하며, 가장 자주 등장하는 값을 결측치에 대입함으로써 데이터의 전체적인 패턴을 유지할 수 있도록 돕는다. 평균이나 중간값은 연속형 데이터에 적용하기 쉽지만, 범주형 데이터에서는 적용하기 어렵다. 이런 경우, 가장 많이 나타나는 범주를 결측치에 대입하는 최빈값 대체 방법이 효과적일 수 있다.
      * 강점 : 범주형 데이터에서 결측치가 발생했을 때 유용하게 사용가능. 특히, 하나의 범주가 압도적으로 많이 나타나는 경우, 그 범주를 결측치에 대입하는 것은 데이터의 일관성을 유지하는 데 도움이 될 수 있다.
      * 단점 : 첫째, 결측치가 특정 패턴을 가지고 발생하는 경우, 그 패턴을 무시하고 단순히 최빈값으로 대체하는 것은 데이터의 편향을 증가시킬 수 있다. 둘째, 결측치가 많을 경우, 최빈값으로 대체하는 것은 데이터의 다양성을 줄이고, 데이터의 분포를 왜곡시킬 수 있다. 셋째, 최빈값이 여러 개인 경우, 어떤 값을 선택해야 할지 결정하기 어렵다는 문제가 있다. 이런 경우, 추가적인 기준이 필요하며, 그 기준에 따라 결과가 달라질 수 있다.
    - 0 또는 상수값 대체 : 0 또는 상수값 대체 방법은 결측치를 0이나 특정 상수값으로 대체하는 방식이다.
      * 강점 : 이 방법은 간단하게 적용할 수 있다는 장점이 있다. 0 또는 상수값 대체 방법의 특징은 모든 결측치를 동일한 값으로 대체한다는 것이다. 이는 결측치 처리를 간단하게 만들어주지만, 동시에 데이터의 원래 분포와 패턴을 왜곡시킬 가능성이 있다. 이 방법이 강점을 보이는 상황은 결측치가 발생한 이유가 '해당 항목이 없음'을 의미할 때다. 예를 들어, 설문조사에서 응답하지 않은 항목이 '해당사항 없음'을 의미하는 경우, 0으로 대체하는 것이 적절하다.
      * 단점 : 그러나, 이 방법은 여러 가지 단점을 가진다. 첫째, 모든 결측치를 동일한 값으로 대체하면, 데이터의 분포와 분산을 왜곡시킬 수 있다. 둘째, 결측치가 무작위로 발생한 것이 아니라, 일정한 패턴이나 원인이 있는 경우, 이 방법은 그 원인이나 패턴을 무시하게 된다. 이는 데이터 분석의 정확성을 저하시킬 수 있다. 셋째, 상수값으로 대체하는 것은 결측치가 실제로 그 상수값을 가지는 경우와 구분하기 어렵다. 이는 데이터를 해석하는 데 혼란을 줄 수 있다.따라서, 0 또는 상수값 대체 방법은 결측치 처리에 있어 간단한 선택지일 수 있지만, 그 결과가 데이터의 원래 패턴을 왜곡시키지 않는지, 데이터 분석의 정확성을 저하시키지 않는지 반드시 확인해야 한다.

<그림 6> em 알고리즘

* + - EM(expectation-maximazation) : 기대-최소 알고리즘으로 불리우는 이 방법은 결측 데이터 또는 숨겨진 데이터가 있는 경우 최대우도추정법을 적용하는 통계적 방법이다. 이 알고리즘은 두가지 단계가 존재한다. 기대 단계(expectation)에서는 현재의 추정 값을 이용하여 결측치에 대체될 기대값을 계산하고, 최대화 단계(maximization)에서는 이 기대값을 사용하여 모수를 다시 추정한다. 이 두가지 단계를 반복적으로 수행하여 결측치를 채우게 되는 것이다.
      * 특징 : 기대값 단계에서는 현재의 모수 추정치를 기반으로 결측치의 기대값을 계산하고, 최대화 단계에서는 계산된 기대값을 이용하여 새로운 모수를 추정한다. 이 두 단계를 반복하면서 모수의 추정치가 수렴할 때까지 계속한다.
      * 강점 : 이 방법은 결측치가 무작위로 발생했을 때 효과적이며, 결측치가 발생한 데이터가 다른 관측치와 함께 특정 분포를 이루고 있다는 가정 하에 잘 작동한다.
      * 단점 : 하지만, 이 방법은 알고리즘의 복잡성으로 인해 계산 시간이 오래 걸리거나, 결측치가 무작위로 발생하지 않았을 때, 즉 결측치의 발생에 일정한 패턴이나 원인이 있을 경우에는 잘못된 추정을 할 수 있다는 단점이 있다.

< 그림 7 > locf 그래프

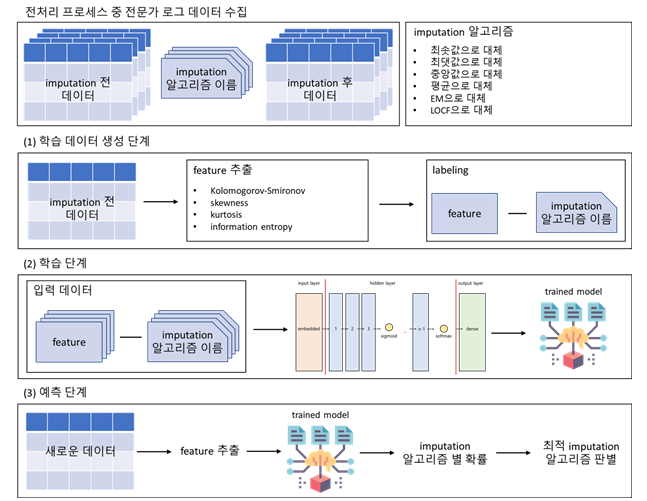
* + - LOCF(last observation carried forward : LOCF(Last Observation Carried Forward) 방법은 결측치가 발생한 경우, 이전의 마지막 관측치로 해당 결측치를 대체하는 방법이다. 이 방법은 주로 시계열 데이터나 연속적인 관측치가 있는 데이터에서 사용된다.
    - 특징 : LOCF의 주요 특징은 다음과 같다. 첫째, 이 방법은 간단하고 직관적이다. 둘째, 데이터의 시간적 연속성을 유지한다는 점에서 장점이 있다. 셋째, 빠른 계산 속도를 가지고 있다.
    - 강점 : LOCF는 상황에 따라 강점을 보이는 경우가 있다. 예를 들어, 결측치가 발생하기 직전의 데이터가 결측치가 발생한 시점의 데이터를 잘 대표하는 경우에는 LOCF가 좋은 선택이 될 수 있다. 또한, 데이터가 시간에 따라서 크게 변하지 않는 경우에도 LOCF 방법이 유용하다.
    - 단점 : 첫째, 결측치가 발생한 시점의 실제 상황을 정확하게 반영하지 못하는 경우가 있다. 이전의 관측치가 결측치가 발생한 시점의 상황을 잘 반영하지 못하면, 왜곡된 결과를 초래할 수 있다. 둘째, 데이터가 시간에 따라 변하는 경향이 있는 경우, LOCF 방법은 이러한 경향을 잘 반영하지 못한다. 즉, 시간의 흐름에 따른 변화를 무시하게 되므로, 시계열 데이터의 패턴을 왜곡할 수 있다.

3. 상태 측정 및 imputation 모듈 기반 전처리 자동화 기술 개발

데이터 분석과 모델 설계는 큰 비즈니스 가치를 창출할 수 있는 중요한 활동이다. 이러한 활동의 기본이 되는 것은 바로 '데이터 전처리'다. 데이터 전처리는 원시 데이터를 분석 가능한 형태로 변환하는 과정으로, 이 과정에서 데이터의 품질을 향상시키고, 데이터의 노이즈를 제거하며, 결측치를 대체하고, 이상치를 탐지하고 수정하는 등의 작업을 포함한다. 이러한 작업들은 데이터 분석의 결과를 크게 좌우하는 중요한 요소이다.데이터 전처리는 데이터 분석 및 모델 설계의 목적에 따라서 전처리 방식이 다양하게 달라진다. 예를 들어, 회귀 분석을 위한 모델 설계를 목적으로 한다면, 데이터의 정규성을 확인하고 이를 향상시키는 전처리 작업이 필요하다. 또한, 분류 문제를 해결하기 위한 모델 설계를 목적으로 한다면, 불균형한 클래스 분포를 고려한 전처리 작업이 필요하다. 데이터 전처리는 심지어 실무자가 매번 새로운 데이터를 직접 확인하고 전처리 기법을 모색해야하는 과제이기도 하다. 즉, 각 데이터마다 특성이 다르기 때문에, 그에 따른 적절한 전처리 기법을 적용해야 한다. 이는 결코 간단한 작업이 아니다. 이러한 이유로 데이터 전처리는 시간적으로 많은 비용이 소요된다. 실제로 데이터 분석 작업의 상당 부분을 데이터 전처리가 차지한다고 해도 과언이 아니다.

## **3.1 전문가 provenance 학습모델**

전문가 provenance 학습모델이란 데이터 처리 과정에서 전문가의 지식과 경험을 기반으로 한 알고리즘이나 처리 방법론을 학습하고, 이를 바탕으로 새로운 데이터에 대한 최적의 처리 방법을 예측하는 AI 모델이다. "Provenance"는 원산지나 출처를 의미하는 단어로, 여기서는 데이터의 변형이나 처리 과정에서 전문가의 결정 및 알고리즘이 어떻게 적용되었는지 그 출처와 흐름을 추적하려는 의미로 사용된다. 따라서 전문가 provenance 학습 모델은 주어진 문제 상황에 따라 전문가의 결정을 모방하거나 그에 기반한 최적의 알고리즘 선택 능력을 가지게 된다. 이러한 방식은 특히 복잡하거나 도메인 지식이 필요한 문제에서 유용하다. 예를 들어, 의료 분야에서는 환자 정보를 다루는데 있어 다양한 변수와 조건들, 그리고 전문적인 지식이 요구된다. 이런 경우 전문가 provenance 학습 모델은 의사나 간호사 등 의료 전문가들이 어떤 판단을 내릴지 예측하는 데 활용될 수 있다. 자사에서는 이 provenance 모델을 참조하여 자동 데이터 전처리 기술을 개발했다.



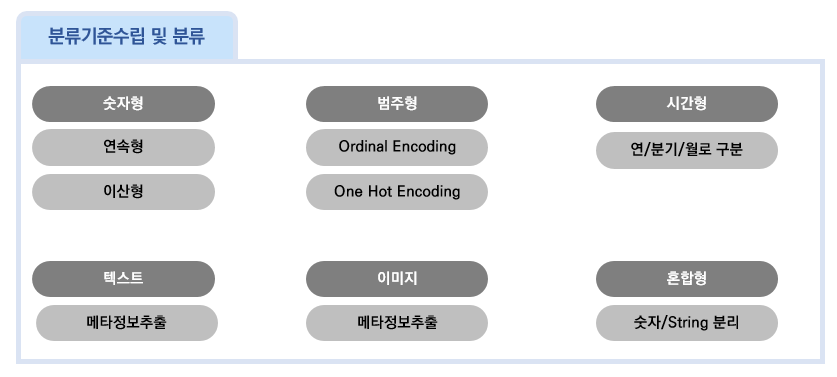
< 그림 > 전문가 provenance 학습 모델을 통한 최적 imputation 알고리즘 판별 프로세스

1. 데이터 전처리: 이 단계에서 raw 데이터를 받아서 필요한 형태로 변환한다. 이 과정에서 결측치나 오류 등을 처리하기 위해 다양한 imputation 알고리즘이 사용될 수 있다. 중요한 것은 어떤 알고리즘이 사용되었느냐 하는 것으로, 이 정보를 추후에 학습 데이터 생성 단계에서 활용한다.
2. 학습 데이터 생성: 이 단계에서 전처리된 데이터를 바탕으로 feature(특징인자)를 추출하고, 해당 feature에 대해 어떤 imputation 알고리즘이 사용되었는지 labeling 한다. 즉, 각 feature와 연결된 imputation 방법론이 함께 표시된 학습 데이터 세트를 생성한다.
3. 모델 학습: 생성된 학습 데이터 세트를 AI 모델에 입력하여 training 시킨다. AI 모델은 각각의 feature와 연관된 최적의 imputation 알고리즘이 무엇인지 배우게 된다.
4. 예측: 실제 운용 상황에서 새로운 입력이 주어질 때 AI 모델은 각각의 feature에 대해 가장 확률이 높은 imputation 알고리즘을 선택한다.

* 본사에서 개발한 데이터 전처리 자동화 기술은 전처리 과정 상 발생할 수 있는 여러 비용, 시간적 문제를 해결하는데 큰 기여를 할 것으로 예상된다. 이 기술은 AI가 데이터의 상태를 직접적으로 파악하고, 데이터에서 필요한 전처리 값들을 자동적으로 식별한 후, 결측치 대체(imputation) 알고리즘을 모듈화하여 자동으로 전처리를 수행한다. 이를 통해 데이터 전처리의 시간적 비용을 크게 줄일 수 있으며, 전처리 과정에서의 인간이 발생시킬 수 있는 실수를 최소화할 수 있다. 이는 결국 데이터 분석의 효율성과 정확성을 향상시킬 수 있을 것이다.

## **3.2 전처리 자동화 기술 및 소스 code 예시**

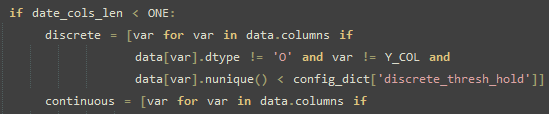
* 자동 전처리를 위한 데이터 분류 기준 수립



<그림 > 전처리 분류 기준

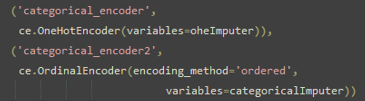
* + 텍스트 데이터와 이미지 데이터는 목적에 따라 전처리 방향이 개별적으로 수립이 필요하기 때문에, 전처리 대신 파일의 메타정보를 추출하여 전처리에 참고할 수 있도록 한다
  + 수치형 데이터는 주로 이산형, 연속형 데이터로 존재하는데, unique value 가 10 미만이면 이산형 데이터, 반대로 10 이상인 경우, 연속형 데이터로 구분할 수 있도록 기준을 마련했다. 추가로, 결측치의 경우 중간값 대체 또는 최대값 대체를 사용하여 결측치를 처리하도록 했다

<그림 9> 수치형 데이터 분류기준 코드2



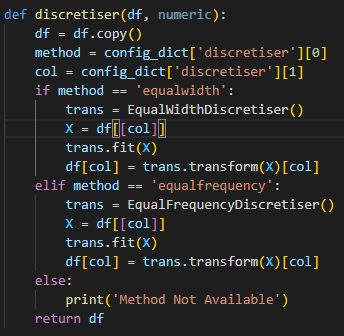
<그림 8>수치형 데이터 분류기준 코드1

* + 범주형 데이터의 경우, y컬럼은 라벨 인코딩을 사용하고, ordered ordinal encoding이나 one-hot encoding을 사용하여 전처리를 자동화하였다.



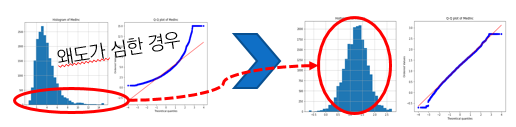
<그림 10>범주형 데이터 전처리

* 전처리 전 특징 분석
  + 결측치, 범주형 변수 등 특징 외에도 선형관계 탐색, 분포 탐색, 이상치 탐색, 스케일링 분석 등 데이터별 특징들이 존재한다.
    - 선형관계탐색과 데이터 분포 탐색의 경우, 사람의 판단능력을 상당히 필요로 한다. 따라서, 자동화하는 데에는 한계가 있으나, 모듈로 개발할 수 있도록 하였다.
    - 분포 탐색을 용이하게 하기 위해서 discretiser를 따로 개발하였다. 이 함수를 통해 데이터의 분포와 상관없이 동일한 크기의 구간을 생성할 수 있다. 즉, 데이터의 분포가 일정하지 않거나, 분포 형태에 대한 가정이 어려울 때 유용하게 사용될 수 있다.



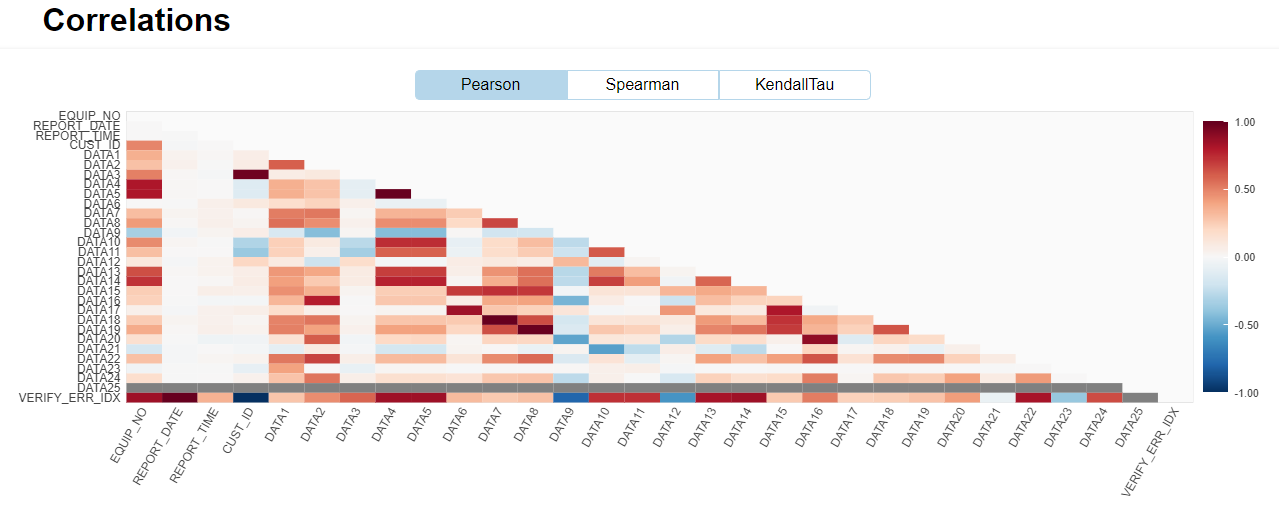
<그림 11> discretiser를 이용한 분포 탐색

* + 특징에 맞게 전처리 후, 생성된 데이터셋의 가치를 판단하는 것은, 해당 데이터셋을 실제로 적용하는 단계에서 적극 활용할 수 있다.
    - 왜도, 인코딩, 상관관계, 이상치 식별, 결측치 식별 등 통계적 수치분석에 따라서 전처리 방향을 확립하고, 이와 같은 eda 결과에 따라 전처리 과정에 피드백할 수 있도록 하였다
      * 왜도 분석: 왜도는 가우시안 형태가 아니라, 한쪽 끝으로 치우쳐져 있는 경우, 그 절대값이 증가한다. 또한, 왜도의 값이 클수록 유의미한 통계분석이 어렵게 된다. 그래서 왜도가 높은경우, 전체 변수 중에서 상위 3개의 왜도 변수를 추출하고 로그로 변환시켜 이를 완화할 수 있도록 했다



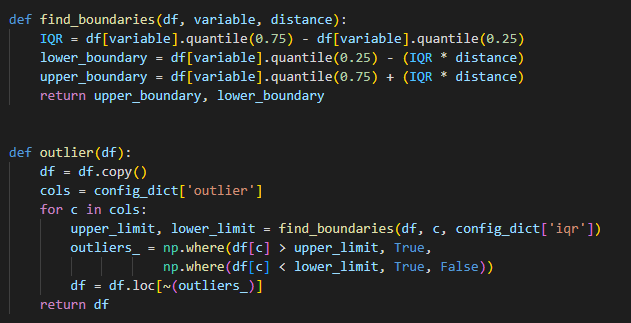
<그림 12> 왜도 처리

* + - * 상관관계 분석 : 하나의 독립변수와 다수의 종속변수 간의 상관관계를 분석하는 방법으로, 상관관계 분석을 통해 예측값에 큰 영향력을 갖는 변수 정보를 제공하고자 했다.



<그림 13> 상관분석

* + - * 이상치 식별 : 이상치는 향후 데이터 셋을 분석하는데 통계적 분석의 왜곡, 모델 성능 저하, 과적합의 위험을 증가시키는 등 여러 문제점을 발생시킨다. 이를 방지하기 위해, 절대값 상관계수 상위 5개를 대상으로 사분위법(IQR)[[1]](#footnote-1)을 사용하여 가중치가 1.5이상 초과한 값을 이상치로 간주하였다.



<그림 14> IQR 사용을 통한 이상치 식별

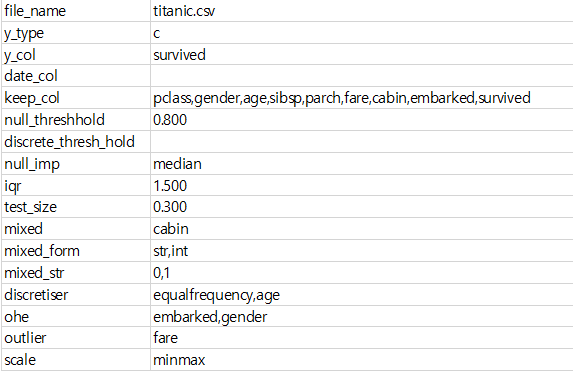
* 전처리 자동화 프로세스
  + 전처리 자동화를 위한 환경변수 파일을 지정.

1. 지도학습 머신러닝의 경우에는 레이블 칼럼을 지정한다.
2. 훈련 데이터와 평가 데이터 비율의 기준을 지정한다.
3. 문자형으로 오분류 방지를 위해 날짜 칼럼을 개별적으로 지정한다.
4. 전처리 효율성 향상을 위해서 필요한 칼럼과 불필요한 칼럼을 지정한다.
5. Null 한계값을 지정하여, 해당 값이 특정 비율 이상일 경우 삭제한다.
6. 이상치 처리를 위해 IQR thresh hold를 지정한다
7. 데이터 변환 연구개발 절차를 진행하면서 새로히 추가할 환경변수 항목을 식별할 수 있다

* 자동화 적용

1. 자동화 프로그램인 파이썬을 실행한다.
2. Argument 변수로 환경변수가 선언된 파일을 지정한다.

* 실제 개발한 환경 변수 파일 예시(사용한 데이터는 타이타닉 데이터를 사용하였다)

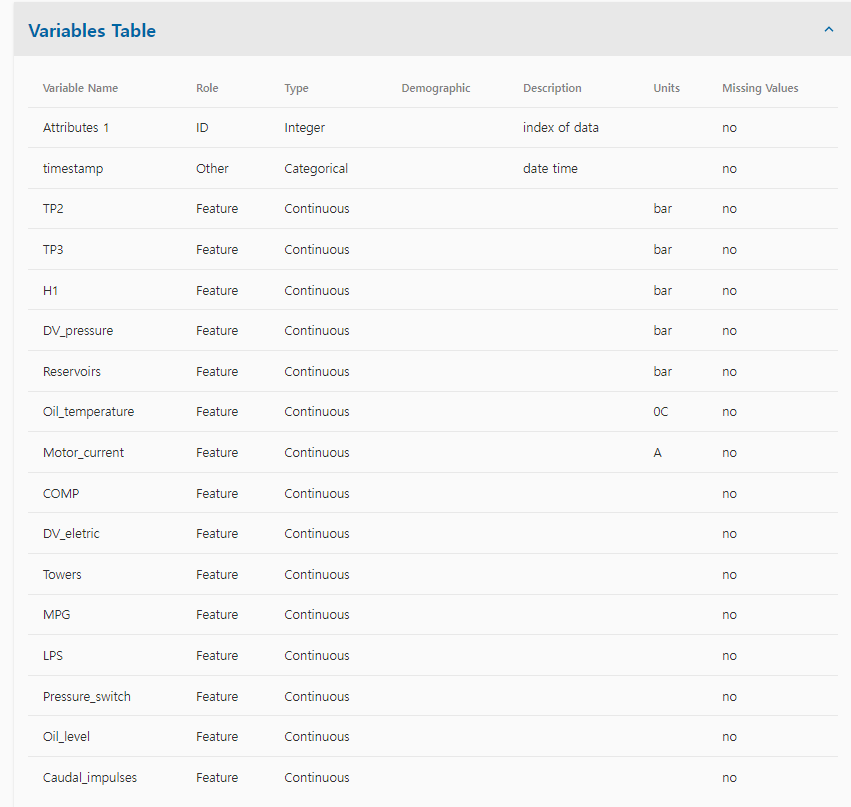


<그림 14>환경변수 파일 예시

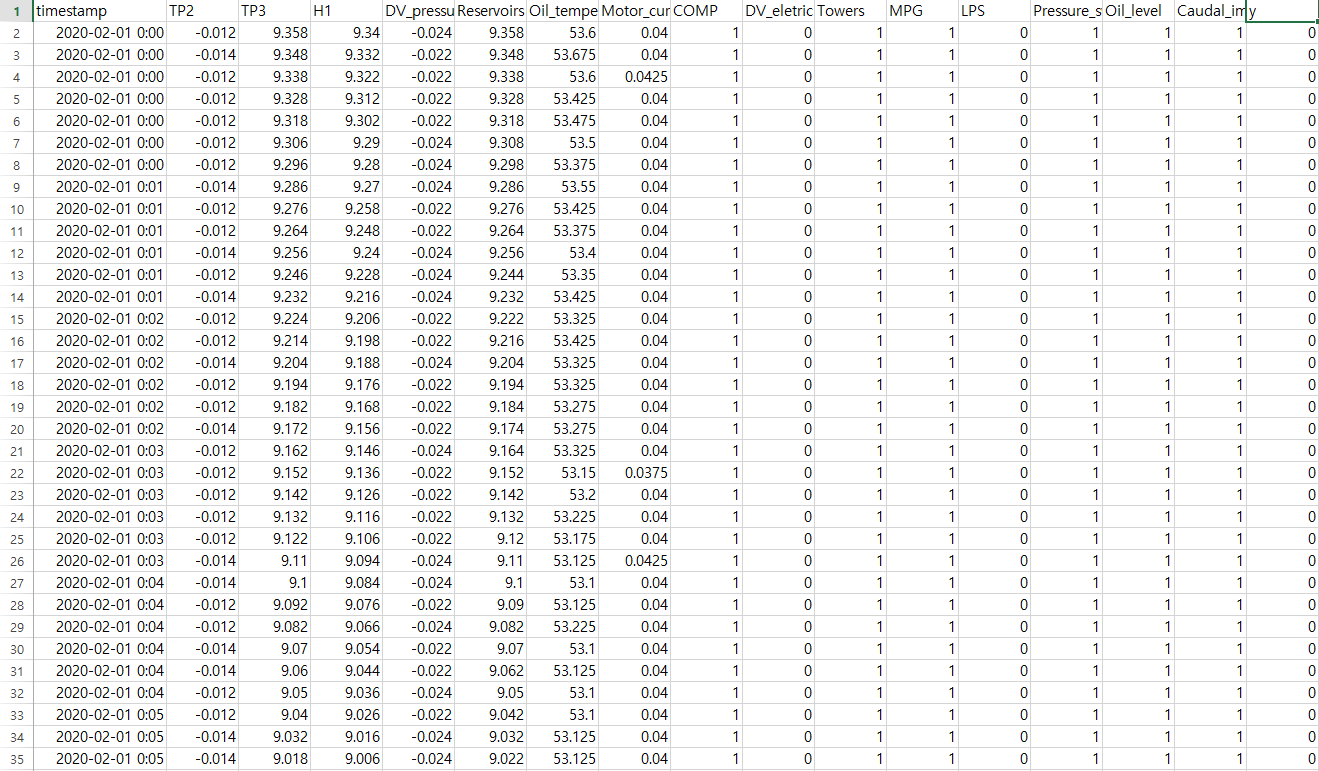
* + 좌측 열은 환경 변수이고, 우측 열은 환경 변수에 해당되는 값이다.
  + Null\_threshhold를 이용하여 null값의 비율이 80%이상인 경우 해당 칼럼은 제외하도록 하였다.
  + Null\_imp는 결측치 처리방식을 의미한다. 예시파일에서는 median, 중간값 대체 방식으로 설정하였다.
  + Discretiser는 equal frequency로 AGE 컬럼에 적용하였음을 알 수 있다.
  + Outlier를 통해 fare칼럼을 이상치 범위(IQR)1.5에 따라 이상치를 처리하였음을 알 수 있다.
  + Scale을 통해 데이터의 범위도 조정할 수 있다. Min-max 방식을 사용하여 스케일링 했음을 보여준다.

# 4. 기술 검증 및 평가

* 데이터 분석 및 모델링 과정에서 모델의 성능을 평가하는 것은 필수적인 절차다. 이 과정은 모델의 예측력을 정량적으로 측정하며, 모델이 실제 상황에 얼마나 잘 일반화될 수 있는지 알아보는 중요한 역할을 한다.모델의 성능 평가는 학습 과정에서 발생할 수 있는 다양한 문제를 사전에 파악하고, 이를 개선하기 위한 방향성을 제시한다. 예를 들어, 모델이 학습 데이터에 과적합되어 있다면, 새로운 데이터에 대한 예측력이 떨어질 수 있다. 이런 문제를 미리 인지하고, 모델의 복잡도를 조정하거나 정규화 기법을 적용해 과적합을 방지할 수 있다.또한, 모델의 성능을 평가하는 과정에서는 다양한 평가 지표를 활용한다. 분류 모델의 경우 정확도, 정밀도, 재현율, F1 점수 등을 사용해 모델의 성능을 측정하고, 회귀 모델의 경우 MSE, RMSE, MAE 등을 활용한다. 이러한 지표들은 모델의 성능을 다양한 관점에서 평가하고, 모델의 강점과 약점을 파악하는 데 도움을 준다.결국, 모델의 성능을 평가하는 것은 '얼마나 잘 동작하는가'를 알아보는 것 이상의 의미가 있다. 이는 모델의 개선 가능성을 찾아내고, 이를 통해 최종적으로 더 높은 성능의 모델을 만들어내는 필수적인 과정이라고 할 수 있다.
* 자사에서는 모델 성능을 검증하는데 유용한 데이터로써, [미국 지하철 전력 유지보수 데이터](http://www.archive.ics.uci.edu/dataset/791/metropt+3+dataset)를 사용하였다.
* 컬럼 및 데이터

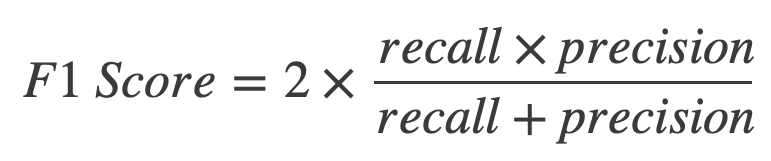


<그림 15> 데이터 컬럼



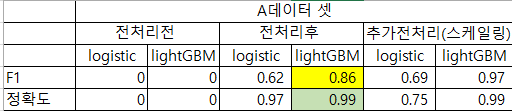
<그림 16> 미국 지하철 전력 유지보수 데이터

* 전처리는 자사에서 개발한 전처리 자동화 기술을 사용하였고, 모델은 lightgbm[[2]](#footnote-2)을 사용하였다.
* 주요 평가 지표는 f1 score이다. 자사는 f1 score가 85%이상을 기록하는 것을 주요 목표로 하고 있다.   
  주로 분류모델에서 사용되는 머신러닝 평가지표라고 할 수 있다. 기존의 accuracry(정확도)는 데이터가 불균형할 때 그 평가 기능이 매우 저하된다. 그러나 f1 스코어는 precision과 recall의 조화 평균을 계산한 값으로, 모델이 얼마나 많은 true값을 찾아내었는지(recall),그리고 찾아낸 참값 중에서도 실제로 참값이 얼마나 많은지(precision)을 동시에 고려하므로, 불균형한 데이터에서 유리하다.



<그림 17> f1 score 공식

* 검증 및 평가 결과  
  위와 같이, f1 스코어가 85% 이상을 기록했다. 이를 통해, 개발한 ai기반 자율 데이터 전처리 기술은 수집한 데이터셋에서 실제 사용이 가능할 정도로 준수한 성능을 보이고 있다. 이후, 해당 전처리 자동화 기술을 한국저작권 위원회에 프로그램 등록을 신청하여, 정식으로 기술이 인정되었다.



<그림 18> 미국 지하철 데이터를 이용한 전처리 자동화 기술 평가 결과

[참고자료]

스마트 해상 물류관리사 – ict 요소기술 개론  
<http://kocw-n.xcache.kinxcdn.com/data/document/2023/upa/leeseoyoungchohyuksoo0525/03.pdf>

Qlik – 데이터 관리

<https://help.qlik.com/ko-KR/sense/May2023/pdf/%EB%8D%B0%EC%9D%B4%ED%84%B0%20%EA%B4%80%EB%A6%AC.pdf>

kistep - 텍스트 마이닝에 의한 기술용어 기반의 예비타당성조사 편익기간 추정 연구

<https://www.kistep.re.kr/board.es?mid=a10305080000&bid=0002&act=view&list_no=25440&tag=&nPage=38>

데이터 사이언스 스쿨 – 가우시안 혼합모형과 em 방법

<https://datascienceschool.net/03%20machine%20learning/18.01%20%EA%B0%80%EC%9A%B0%EC%8B%9C%EC%95%88%20%ED%98%BC%ED%95%A9%EB%AA%A8%ED%98%95%EA%B3%BC%20EM%20%EB%B0%A9%EB%B2%95.html>

1. IQR : Q1(1사분위)와 Q3(3사분위) 간 범위 차이. 이상치를 식별하는 범위이다. [↑](#footnote-ref-1)
2. Gbm(약한 예측모델인 결정트리를 연속적으로 학습시켜 강한 예측 모델을 만드는 알고리즘)에서 더욱 강화된 모델. 트리 분할(leaf-wise)을 사용하여, 기존보다 손실을 크게 줄일 수 있는 방향으로 트리를 성장시키는 방법 [↑](#footnote-ref-2)