

WiseProphet

사용자 매뉴얼

WiseProphet v1.0

WISEiTECH

Copyrights 2019 WISEiTECH co., Ltd. All rights reserved.

Copyright Notice

Copyrights 2019 WISEiTECH co., Ltd. All rights reserved.

대한민국 경기도 과천시 과천대로 12 길 117 과천펜타원 G 동 11~13 층

Restricted Rights Legend

All WiseProphet Software and documents are protected by copyright laws and the Protection Act of Computer Programs, and international convention. WiseProphet software and documents are made available under the terms of the WISEiTECH License Agreement and may only be used or copied in accordance with the terms of this agreement. No part of this document may be transmitted, copied, deployed, or reproduced in any form or by any means, electronic, mechanical, or optical, without the prior written consent of WISEiTECH Co., Ltd.

이 소프트웨어(WiseProphet) 사용설명서의 내용과 프로그램은 저작권법, 컴퓨터프로그램보호법 및 국제 조약에 의해 보호받고 있습니다. 사용설명서의 내용과 여기에 설명된 프로그램은 WISEiTECH Co., Ltd.와의 사용권 계약 하에 서만 사용이 가능하며, 사용권 계약을 준수하는 경우에만 사용 또는 복제할 수 있습니다. 이 사용설명서의 전부 또는 일부분을 WISEiTECH 의 사전 서면 동의 없이 전자, 기계, 녹음 등의 수단을 사용하여 전송, 복제, 배포, 2 차적 저작물 작성 등의 행위를 하여서는 안 됩니다.

Trademarks

WiseProphet is a registered trademark of WISEiTECH Co., Ltd. Other products, titles or services may be registered trademarks of their respective companies.

WiseProphet 은 WISEiTECH Co., Ltd.의 등록 상표입니다. 기타 모든 제품들과 회사 이름은 각각 해당 소유주의 상표로서 참조용으로만 사용됩니다.

사용자 매뉴얼 정보

안내서 제목: Wise Prophet 사용자 매뉴얼

최초 발행일: 2019-04-26

최종 수정일: 2025-05-27

소프트웨어 버전: Wise Prophet v1.0

안내서 버전: 1.0.1

내용 목차

매뉴얼에 대하여	1
제품 정보 및 권장 사양	4
머신러닝의 기본 개념과 용어 정의	5
제 1 장. WiseProphet 소개	
1.1 개요	10
1.2 특징	10
1.3 주요 기능	12
제 2 장. 로그인	13
2.1 화면 구성	13
2.2 회원가입	14
제 3 장. 모델 학습	15
3.1 화면 구성	15
3.2 데이터 선택	16
3.2.1 파일 생성	16
3.2.2 로컬 파일	17
3.2.3 데이터베이스	18
3.3 데이터 탐색	19
3.3.1 탐색적 데이터 분석	19
3.3.2 데이터 스케일	20
3.3.3 데이터 분포	21
3.3.4 변수 유형 변경	22
3.3.5 변수 사용 여부 설정	23
3.3.6 목표 변수 설정	23
3.4 특징 선택	24

3.4.1	특징별 영향	24
3.4.2	상관관계	25
3.5	알고리즘 선택.....	26
3.5.1	클러스터링 모델.....	26
3.5.2	분류 모델.....	27
3.5.3	회귀 모델.....	28
3.5.4	파라미터 최적화.....	29
3.6	검증데이터 비율 설정.....	30
3.6.1	훈련-평가 데이터 분할	30
3.6.2	교차 검증.....	30
3.7	모델 실행.....	31
3.7.1	클러스터링	31
3.7.2	분류	31
3.7.3	회귀	33
3.7.4	모델 로그	34
제 4 장.	모델 관리	35
4.1	화면 구성.....	35
4.2	모델 관리	35
4.3	모델 수정	37
제 5 장.	모델 운영	38
5.1	화면 구성.....	38
5.2	설정 및 기능.....	39
5.2.1	스케줄 추가.....	39
5.2.2	스케줄 확인	39
제 6 장.	모니터링	40
5.1	화면 구성.....	40
5.2	사용자 및 모델 정보	40

제 7 장. 설정	42
5.1 사용자 설정	42
5.2 연결 설정	42
제 8 장. FAQ	44

매뉴얼에 대하여

매뉴얼의 대상

본 안내서는 **WiseProphet** 을 사용하여 데이터를 전처리하고 모델을 생성하여 데이터를 분류/예측하려는 사용자들을 대상으로 기술한다. WiseProphet 에서 제공하는 알고리즘은 오픈 소스인 Scikit-learn, Keras, Tensorflow 에서 제공하는 알고리즘을 기반으로 한다.

매뉴얼의 전제 조건

본 안내서는 WiseProphet 에서 모델을 생성하고 데이터를 예측하기 위해 필요한 과정을 설명하는 매뉴얼이다. 따라서 본 안내서를 원활히 이해하기 위해서는 다음과 같은 사항을 미리 알고 있어야 한다.

- 머신러닝의 이해 - “머신러닝의 기본 개념과 용어 정의” 참고

매뉴얼의 제한 조건

본 안내서는 WiseProphet 를 실무에 적용하거나 운용하는 데 필요한 모든 사항을 포함하고 있지 않다. 따라서 설치, 환경설정 등 운용 및 관리에 대해서는 각 제품 안내서를 참고하기 바란다.

참고

WiseProphet 의 설치 및 환경 설정에 관한 내용은 "운영 매뉴얼"을 참고한다.

안내서 구성

WiseProphet 사용자 안내서는 총 5 개의 장과 Appendix 로 구성되어 있다.

각 장의 주요 내용은 다음과 같다.

- 시작하기 전에:

- 제품 정보 및 권장 사양

- WiseProphet 제품 정보 및 하드웨어 권장 사양에 대해 간략히 설명한다.

- 머신러닝의 기본 개념 및 용어

- WiseProphet 을 사용하기 전에 머신러닝의 기본 개념에 대해 간략히 설명한다.

- 제 1 장: WiseProphet 소개

- WiseProphet 의 주요 특징에 대해 기술한다.

- 제 2 장: 로그인

- WiseProphet 의 로그인 화면에 대해 간략히 기술한다.

- 제 3 장: 모델 학습

- WiseProphet 에서 모델을 학습하기 위한 과정을 기술한다.

- 제 4 장: 모델 관리

- WiseProphet 에서 모델을 이용하여 신규 데이터를 예측하고 관리하는 과정을 기술한다.

- 제 5 장: 모델 운영

- WiseProphet 의 모델을 주기적으로 실행하기 위한 배치 스케줄 설정 방법을 기술한다.

- 제 6 장: 모니터링

- WiseProphet 의 모니터링 화면에 대해 간략히 기술한다.

- 제 7 장: 설정

- WiseProphet 의 설정에 대해 간략히 기술한다.

- 제 8 장: FAQ

- WiseProphet 관련 FAQ 에 대해 간략히 기술한다.

연락처

WISEiTECH Co., Ltd.
117, Gwacheon-daero 12-gil,
Gwacheon-si, Gyeonggi-do,
Republic of Korea
Tel: + 82-02-6246-1400
Fax: + 82-02-6246-1415
Email: contact@wise.co.kr
Web (Korean): <http://www.wise.co.kr/>
기술지원: <http://www.wise.co.kr/>

제품 정보 및 권장 사양

1. 구성 및 설치

본 제품은 사용자 매뉴얼, CD, 실행 파일로 구성되어 있다. 제품 설치의 사용자가 아닌 설치 담당자가 설치하며 고객에게 매뉴얼, CD, 실행파일을 전달한다.

2. 보증 기간

제품 구매일로부터 1 년간 무상으로 유지보수 서비스를 제공하며, 그 이후에는 유상으로 지원된다.

3. 하드웨어 사양

하드웨어 사양은 분석하려는 데이터양에 따라 다르기 때문에 본 매뉴얼에서는 100GB 기준으로 하드웨어 사양을 작성했다. GPU 장비는 필수 조건은 아니며 분석 성능을 최대로 올리려면 필요하다.

CPU	8Core 이상
Memory	128GB 이상
HDD	1TB 이상 (제품이용자 * 1TB)
GPU	RTX 2080 Ti 이상

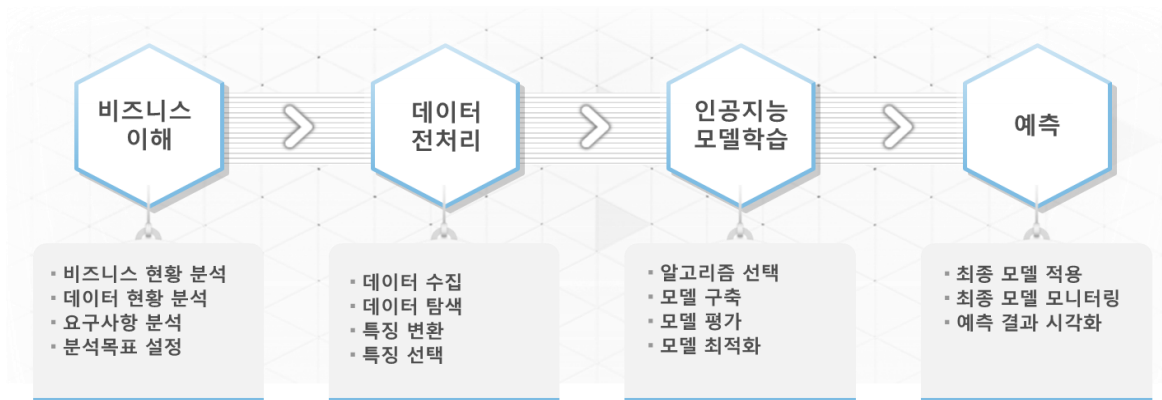
4. 지원 OS

- CentOS Linux
- OS X
- Redhat Enterprise Linux
- Solaris
- SUSE Linux
- Ubuntu Linux
- Microsoft Windows

머신러닝의 기본 개념

1. 머신러닝

머신러닝은 데이터 분석을 위한 모델 생성을 자동화하여 프로그램이 데이터를 바탕으로 학습하고 패턴을 찾아내는 것을 말한다. 학습을 통해 사람의 개입을 최소화하고 빠르게 의사결정을 할 수 있도록 지원한다.



[그림] 머신러닝 프로세스

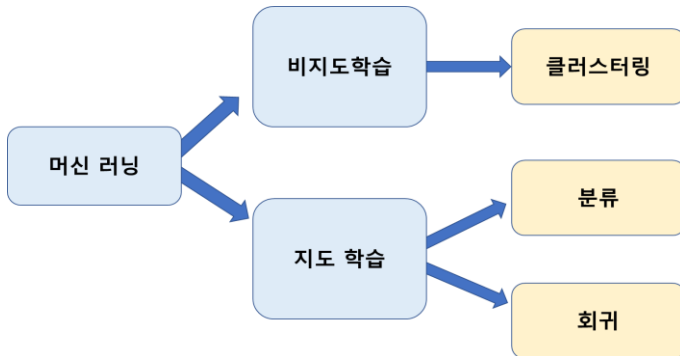
2. 데이터 전처리

데이터 전처리는 학습 모델을 생성하기 전에 데이터를 탐색하고 데이터 정제/변환, 데이터의 특징을 선택하는 작업을 말한다. 우수한 예측 모델을 구축하기 위해 수행해야 하는 중요한 작업이다. 데이터 전처리 과정은 다음과 같다.

구분	내용
1) 데이터 탐색	<ul style="list-style-type: none"> - 데이터 분석을 하기 전에 데이터를 탐색하는 단계 - 데이터의 분포 및 기초 통계량 (평균, 분산)을 확인' - 데이터의 상태 - 결측값 또는 이상값 여부 확인 - 데이터의 유형 (수치형, 범주형, 텍스트형 등)
2) 데이터 정제	<ul style="list-style-type: none"> - 결측값을 채우거나 노이즈가 많은 데이터와 이상값을 제거
3) 데이터 변환	<ul style="list-style-type: none"> - 데이터 스케일: 서로 다른 범위의 데이터의 범위를 지정된 범위로 변환하는 작업 예) 키/몸무게를 0~1 사이의 데이터 범위로 변환 - 범주형 데이터 변환: 범주형 데이터를 수치형 데이터(0,1)로 변환(One Hot Encoding)
4) 특징 선택	<ul style="list-style-type: none"> - 데이터를 분류/예측을 위해 최적의 특징을 선택하는 작업 - 필요한 특징들을 추가하거나 불필요한 특징을 제거하는 작업
5) 특징 추출	<ul style="list-style-type: none"> - 특징들의 조합으로 새로운 특징을 생성하는 것

3. 학습

학습 방법은 크게 지도학습과 비지도학습으로 나눌 수 있다.



[그림] 머신러닝 모델 구분

1) 지도 학습(Supervised Learning)

- 지도학습은 데이터에 대한 정답(레이블 또는 값)이 주어진 상태에서 학습시키는 방법이다.
- 각 데이터에 대한 정답을 레이블(label)이라고 표현하며, 레이블이 있는 데이터들의 집합은 학습 데이터(Training Set)라고 한다.
- 학습 데이터들을 기반으로 한 모델이 생성되고, 생성된 모델을 통해 어떠한 특징을 갖는 데이터가 어떤 정답(레이블 또는 값)에 속할 지 예측할 수 있다.

• 지도학습의 유형

지도학습은 유형에 따라 분류(Classification)와 회귀(Regression)으로 나뉜다.

- 분류(Classification)

: 어떤 카테고리에 해당하는 분류하는 기법 예) 스팸메일 분류 (스팸/정상메일 분류)

대표적인 알고리즘: 서포트 벡터 머신, 의사결정나무 등

- 회귀(Regression)

: 연속값(연속되는 수치)를 예측할 때 사용 예) 집값 예측, 주가 예측, 수요 예측

대표적인 알고리즘: 선형 회귀분석, 로지스틱 회귀분석 등

2) 비지도 학습(Unsupervised Learning)

비지도 학습은 데이터에 대한 정답(레이블)이 주어지지 않고, 데이터의 특징 만을 가지고 데이터의 숨겨진 패턴이나 그룹을 찾는 데 사용된다.

대표적인 알고리즘: k-means 클러스터링, PCA

예) 구글 뉴스 서비스: 비슷한 주제의 뉴스끼리 묶어줌.

단어 클러스터링: 유사한 단어끼리 묶어줌.

4. 알고리즘 선택

사용자의 분석 목적에 따라 학습 방법(분류/회귀)을 결정하고, 학습 방법에 따라 알고리즘을 선택한다. 알고리즘을 선택할 때에는 예측 정확도, 학습 시간, 사용 편의성을 고려한다.

학습유형에 따른 대표적인 알고리즘은 다음과 같다.

학습 유형	알고리즘
분류	랜덤 포레스트
	의사결정나무
	서포트벡터머신
회귀	선형 회귀분석
	엘라스틱 넷
클러스터링	k-means

5. 모델 평가

모델 평가는 학습 모델의 성능을 평가하고 모델을 활용하여 새로운 데이터에 대한 예측을 하는 단계이다. 모델을 평가할 때는 학습/평가 데이터를 나누어 수행한다.

특히, 모델 평가에서 고려해야 할 사항은 평가 데이터는 알고리즘 선택과 모델 학습 과정에서 쓰이지 않아야 한다.

모델 평가 방법은 학습/검증 데이터 분할 방법과 교차 검증 방법이 있다.

- 1) 학습/검증 데이터 분할: 학습데이터와 검증데이터를 나누어 평가하는 방법이다. 가장 많이 쓰이는 방법이다. 보통 80(학습데이터):20(평가데이터)로 나누어 평가한다.
- 2) 교차 검증: 학습/평가 데이터를 한 번만 나누는 것이 아니라 여러 번 하는 방식이다. 여러 세트를 나누어서 하나를 학습 데이터로, 나머지를 평가 데이터로 사용하여 여러 번 반복하여 정확도가 높은 모델을 선택한다.

모델 평가 지표는 학습 유형에 따라 분류, 회귀, 클러스터링으로 구분된다.

학습 유형	평가 지표	설명
분류	정확도(Accuracy)	예측값과 실젯값이 일치하는 비율
	정밀도(Precision)	예측값 중 실젯값이 발생하는 비율
	재현율(Recall)	모델에서 분류된 값이 정확하게 탐지한 정답비율
회귀	평균제곱근오차(RMSE)	예측값 - 실젯값의 제곱근
	평균절대오차(MSE)	예측값 - 실젯값의 절대값
	평균오차비율(MAPE)	예측값 - 실젯값의 비율

용어 정의

본 매뉴얼의 머신러닝과 관련된 용어는

구글 머신러닝 용어집(<https://developers.google.com/machine-learning/glossary/>)을 참고하여 한글로 표기한다.

명칭 (한글명)	명칭(영문명)	설명
학습 데이터	Training Set	모델 학습에 사용되는 데이터 집합
평가 데이터	Test Set	모델 초기 학습 시 모델을 테스트하는데 사용하는 데이터 집합
특징	Feature	예측을 수행하는 데 사용되는 입력 변수
특징 선택	Feature Selection	모델에 필요한 입력변수를 선택하는 작업
특징 추출	Feature Engineering (=Feature Extraction)	모델을 학습시키는 데 유용할 특성이 무엇인지 판단하고 로그 파일 및 기타 소스의 원시 데이터를 해당 특성으로 변환하는 과정
특징별 영향	Feature Importance	목표 변수에 대한 각 변수들의 상대적 중요도를 의미한다
데이터 스케일	Data Scale	특성 값 범위를 데이터 세트의 다른 특성 범위와 일치하도록 맞추는 작업. 예를 들어 데이터 세트의 모든 부동 소수점 특성을 0~1 범위로 맞추 수 있습니다. 어떤 특성의 범위가 0~500 이라면 각 값을 500 으로 나누어 특성을 조정할 수 있습니다.
1 분위	Q1	데이터의 하위 25% 값 데이터의 25%가 이 값보다 작거나 같음을 의미
3 분위	Q3	데이터의 상위 75% 값 데이터의 75%가 이 값보다 작거나 같음을 의미
유일값	Unique Value	데이터 중복 제거 후의 유일한 값을 의미
결측값	Missing Value	데이터가 빈 값일 경우
목표 변수	Target Value	예측하고자 하는 대상을 의미
정규화	Normalization	실제 값 범위를 표준 값 범위(일반적으로 -1~+1 또는 0~1)로 변환하는 과정입니다. 예를 들어 어떤 특성의 원래 범위가 800~6,000 인 경우, 뿔샘과 나뭇샘을 거쳐 값 범위를 -1~+1 로 정규화 할 수 있다.
매개변수	Parameter	머신러닝에서 스스로 학습하는 모델의 변수를 의미 학습이 반복됨에 따라 가중치 매개변수의 값이 학습한다.
범주형 데이터	Categorical Data	가능한 값의 불연속 집합을 갖는 특징 예) 성별(남자, 여자)
원-핫 인코딩	One-hot Encoding	범주형 데이터를 수치형 (0,1)로 변환해주는 작업
모델 학습	Model Training	최상의 모델을 결정하는 과정

L1 정규화	L1 regularization	가중치 절대값 합에 비례하여 가중치에 페널티를 주는 정규화 유형
초평면	Hyperplane	한 공간을 두 부분 공간으로 나누는 경계
분류 행렬	Confusion Matrix	실제값과 예측값을 비교하여 나타낸 매트릭스 분류 평가 지표로 사용
배치	Batch	모델 학습을 반복하는 작업

제 1 장

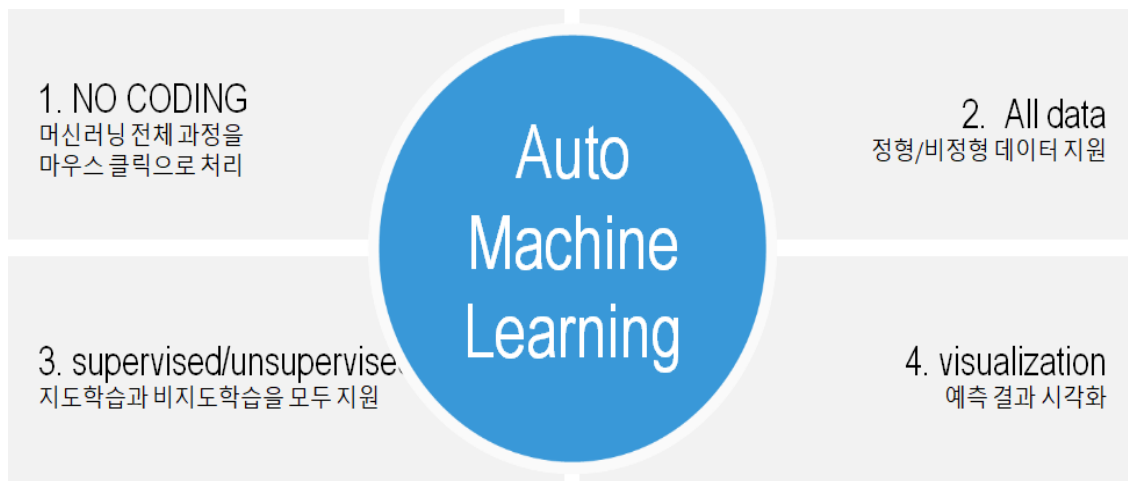
WiseProphet 소개

본 장에서는 WiseProphet 을 사용하기 전 주요 기능과 특징을 간략히 소개한다.

1.1. 개요

WiseProphet 은 데이터를 활용하여 손쉽게 예측 결과를 도출할 수 있는 머신 러닝 프로세스 자동화 플랫폼이다. 데이터를 수집하여 다양한 모델을 자동으로 실행하고 최적화된 알고리즘을 찾아 예측 결과를 도출한다.

1.2. 특징



1.2.1. 쉬운 사용

- 머신러닝 비전문가도 마우스 클릭만으로 머신러닝 적용 가능
- 손쉬운 유저 인터페이스

1.2.2. 모든 데이터 지원

- 정형 데이터뿐만 아니라, 텍스트, 이미지 등 비정형 데이터 지원
- 비정형 데이터로부터 특징을 추출하여 분석 가능

1.2.3. 다양한 알고리즘 지원

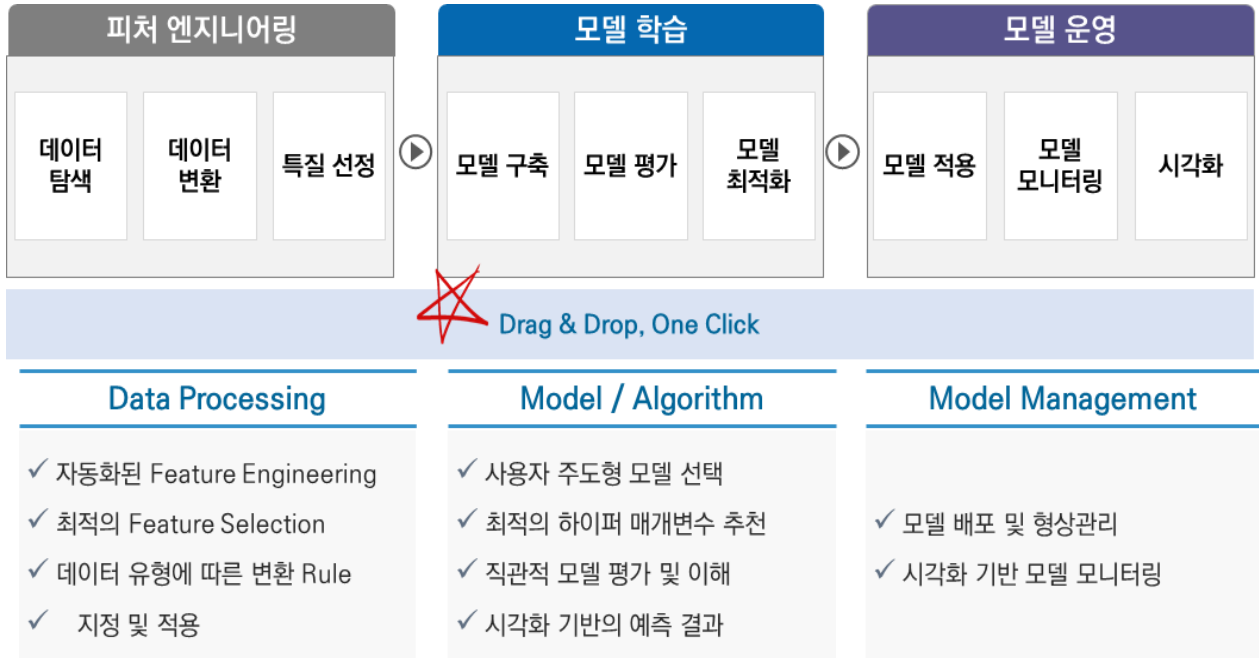
- 회귀, 분류, 클러스터링 알고리즘을 지원
- 다양한 알고리즘 및 비지도학습을 지원하며 지도학습과 연계 가능
- 신규 패턴 발굴을 위한 이상치 알고리즘과 예측 분석을 위한 지도 학습 알고리즘 지원

1.2.4. 특징 추출

- 피쳐 엔지니어링으로 데이터 전처리를 체계적으로 지원
- 특징 선택을 통한 특징 최적화 및 특징 중요도 제시

1.3. 주요 기능

WiseProphet의 주요 기능은 다음과 같다.

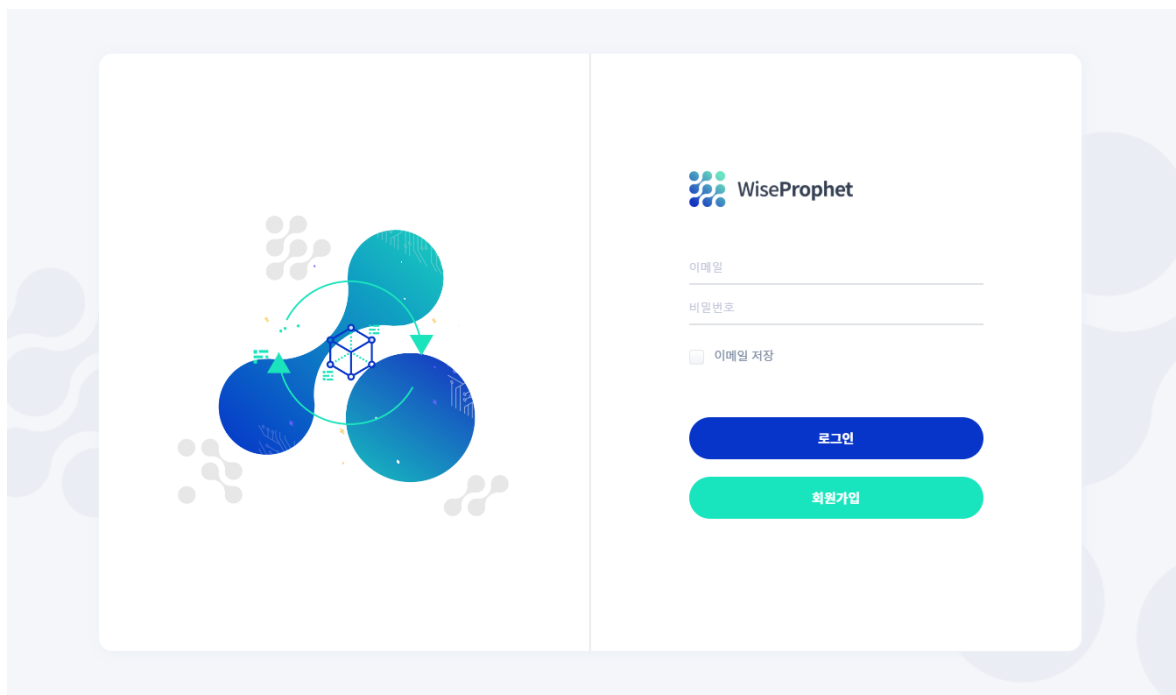


기능	설명
데이터 준비	다양한 정형, 비정형 형식의 데이터 원본 지원
공학적인 데이터 전처리	데이터 변환, 데이터 정제, 변수 스케일링
다양한 알고리즘 제공	예측 유형별 다양한 알고리즘 제공 다양한 오픈소스 기반 검증된 최적의 알고리즘 제공
학습 및 매개변수 튜닝 지원	반복적 모델 학습 하이퍼 매개변수 최적화/튜닝
모델 관리	다양한 환경으로의 손쉬운 모델 배포 주기적인 모델 모니터링 및 관리
모델 설명의 시각화	예측 결과의 이해를 위한 직관적인 모델 평가 지표 표시 시각화 기반의 모델 예측 결과 화면 제공

제 2 장 로그인

본 장에서는 사용자 등록 및 로그인을 위한 WiseProphet 사용법을 설명한다.

2.1. 화면 구성

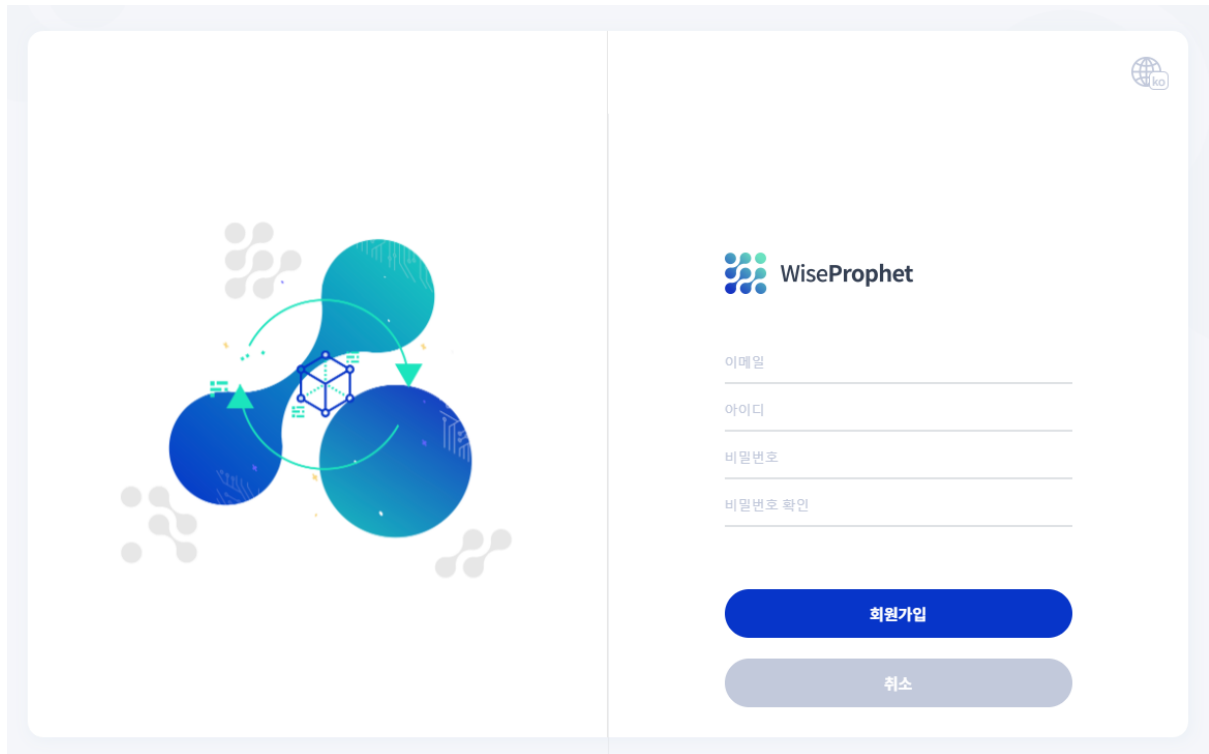


회원가입을 통해 사용자 등록을 하고 로그인하는 화면으로 구성된다.

- 1) 위의 화면은 로그인 메인 화면으로 이메일 및 비밀번호를 입력한 뒤 로그인 버튼을 클릭하면 모델 학습 메인 화면으로 이동한다.
- 2) 로그인 유지시간은 1 시간이며 1 시간이후에는 다시 로그인 해야 한다.
- 3) 회원가입 버튼을 클릭하여 사용자 등록을 한다.
- 4) 비밀번호는 5 회이상 틀리면 계정이 잠기며 다시 사용하기 위해서는 관리자 계정으로 요청 메일을 보내야 한다 (wiseprophet@wise.co.kr)

2.2. 회원가입

- 회원가입 버튼을 누르면 개인정보 처리방침서를 확인할 수 있으며 동의를 클릭하면 다음과 같이 이메일과 비밀번호를 설정하여 사용자 등록을 할 수 있다.



이메일

아이디

비밀번호

비밀번호 확인

회원가입

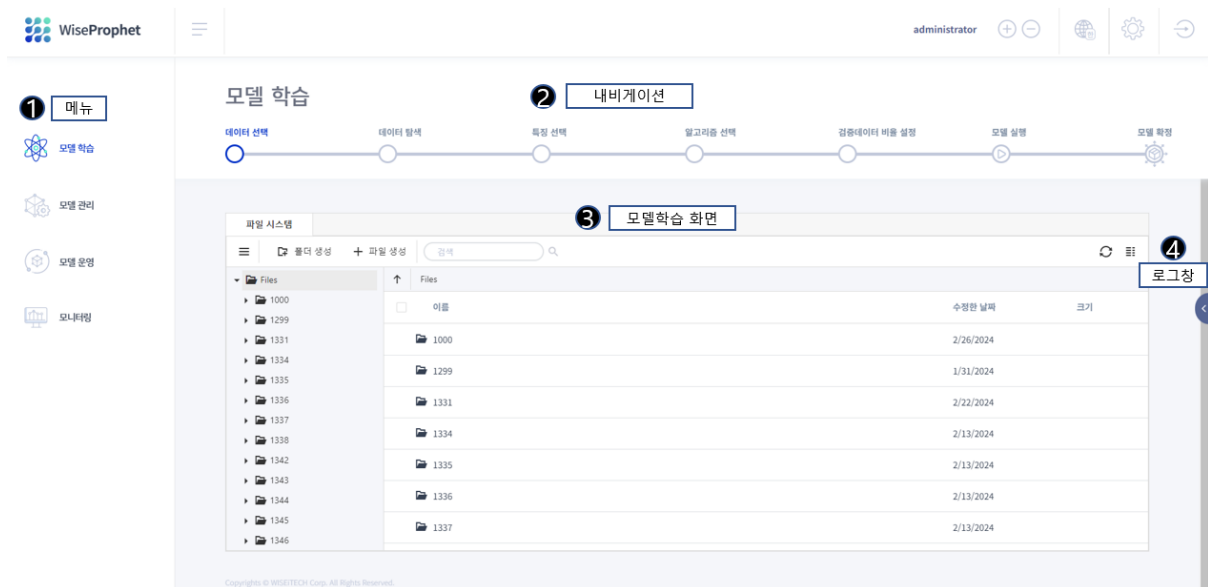
취소

제 3 장

모델 학습

본 장에서는 모델 학습을 위한 WiseProphet 사용법을 설명한다.

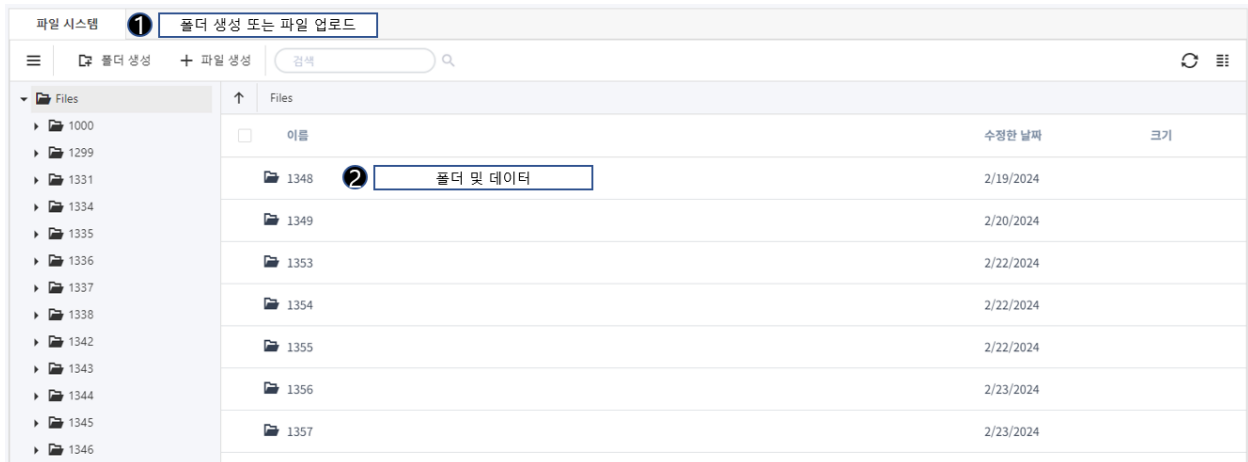
3.1. 화면 구성



데이터를 선택하고 모델을 학습하기 위한 화면은 1) 메뉴 화면 2) 내비게이션 3) 모델 학습 화면 4) 로그창으로 구성된다.

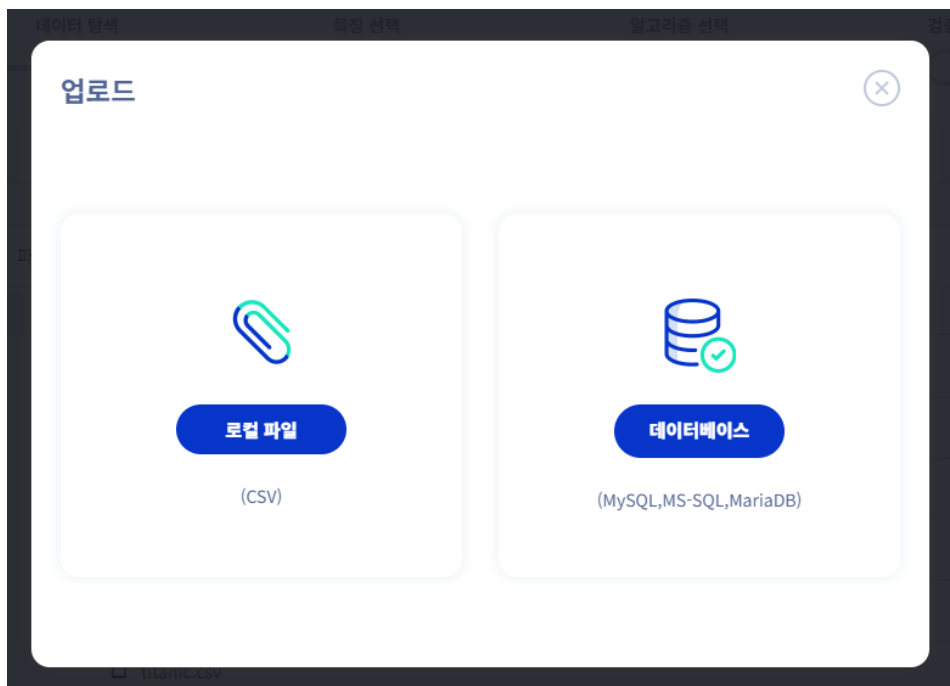
- 1) 메뉴 화면으로 분석 모델 생성, 분석 모델 관리, 모니터링 메뉴로 구성된다.
- 2) 내비게이션으로 현재의 모델 학습 단계를 보여주며, 데이터 선택, 데이터 탐색, 특징 선택, 모델 실행, 모델 확정까지 총 7 단계로 이루어져 있다.
- 3) 모델학습 메인 화면으로 학습 단계에 맞춰 학습 과정을 보여준다. 데이터 선택 - 데이터 탐색 - 특징 선택 - 알고리즘 선택 - 검증데이터 비율 설정 - 모델 실행 - 모델 확정 단계로 모델 학습을 수행할 수 있다.
- 4) 로그창: 모델 학습에 실행했던 데이터 분석 로그 기록들을 모두 확인할 수 있다.

3.2. 데이터 선택



데이터 선택 기본화면은 1) 폴더 생성 및 파일 업로드 2) 폴더 및 데이터 목록으로 구성된다.

3.2.1. 파일 생성



로컬 파일

- 1) 로컬 파일 버튼을 클릭하여 해당 데이터를 업로드한다.
- 2) 파일 타입은 CSV 형식을 지원한다.

데이터베이스

- 1) 데이터베이스는 선택한 테이블의 데이터를 불러온다.
- 2) 데이터베이스는 mysql, mssql, oracle 을 지원한다.

3.2.2. 로컬 파일



업로드할 데이터를 선택한 후 업로드 버튼을 누르면 파일이 업로드되며 데이터 선택 기본화면에서 파일을 선택할 수 있다.

이름	수정된 날짜	크기
deployResult	2/23/2024	
exeResult	2/26/2024	
featureResult	2/23/2024	
<input checked="" type="checkbox"/> iris_ori (1).csv	2/27/2024	5.1 KB
iris_test_outlier.csv	2/20/2024	6.8 KB
iris_test_outlier_smote.csv	2/26/2024	7.1 KB
lstm.csv	2/22/2024	206.8 KB

- 1) 분석할 데이터를 선택한다.
- 2) 파일 선택을 누르면 해당 데이터로 모델학습을 시작한다.
- 3) 선택한 데이터의 이름 변경/이동/삭제/다운로드를 할 수 있다.

3.2.3. 데이터베이스

데이터베이스를 이용하여 분석할 데이터를 불러올 수 있다.

업로드 ✕

* 데이터베이스 연결은 우측 상단의 [설정]-[연결 설정]에서 추가하실 수 있습니다.

① wise_demo : mysql

	데이터베이스	테이블명	테이블 설명
<input type="checkbox"/> ②	wise_demo	age_change	
<input type="checkbox"/>	wise_demo	AI_Gas	
<input type="checkbox"/>	wise_demo	AI_Gas_Paper	
<input type="checkbox"/>	wise_demo	awsbatch_test	
<input type="checkbox"/>	wise_demo	awsbatch_test2	
<input type="checkbox"/>	wise_demo	batchgg	

③ 뒤로 **확인**

- 1) 연결된 데이터베이스명을 선택한다.
- 2) 테이블명을 선택한다.
- 3) 선택 버튼을 클릭한다.

3.3. 데이터 탐색

데이터 탐색은 데이터의 구조와 특징을 파악하고 기초 통계량을 확인하는 과정이다.

Standard Scale | 검색어를 입력해주세요

데이터셋 정보 (결측값 : 177 중복값 : 0 컬럼수 : 9 데이터 갯수 : 891)

변수명	변수유형	최소값	최대값	유일값	결측값	평균	표준편차	사용	목표변수
PassengerId	Num...	1	891	891	0	446	257.35	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Survived	Num...	0	1	2	0	0.38	0.49	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Pclass	Num...	1	3	3	0	2.31	0.84	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Sex	Categ...			2	0	0	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Age	Num...	0.42	80	88	177	29.7	14.53	<input checked="" type="checkbox"/>	<input type="checkbox"/>
SibSp	Num...	0	8	7	0	0.52	1.1	<input checked="" type="checkbox"/>	<input type="checkbox"/>

데이터 초기화 | 뒤로 | 선택

- 데이터 탐색 초기화 화면이다.
- 데이터 탐색 화면은 1) 탐색적 데이터 분석 2) 데이터 스케일 3) 데이터 분포 4) 데이터 타입 변경 기능 5) 변수 사용 여부 설정 6) 목표 변수 설정으로 구성된다.

3.3.1. 탐색적 데이터 분석

데이터 탐색 단계에서는 데이터 선택 단계에서 업로드 한 데이터의 변수 목록과 기초 통계량을 확인할 수 있으며, 표 오른쪽 상단부분에 데이터셋에 대한 기본정보를 제공한다.

Standard Scale | 검색어를 입력해주세요

데이터셋 정보 (결측값 : 177 중복값 : 0 컬럼수 : 9 데이터 갯수 : 891)

변수명	변수유형	최소값	최대값	유일값	결측값	평균	표준편차	사용	목표변수
PassengerId	Num...	1	891	891	0	446	257.35	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Survived	Num...	0	1	2	0	0.38	0.49	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Pclass	Num...	1	3	3	0	2.31	0.84	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Sex	Categ...			2	0	0	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Age	Num...	0.42	80	88	177	29.7	14.53	<input checked="" type="checkbox"/>	<input type="checkbox"/>
SibSp	Num...	0	8	7	0	0.52	1.1	<input checked="" type="checkbox"/>	<input type="checkbox"/>

데이터 초기화 | 뒤로 | 선택

- 데이터 탐색 화면에서 확인할 수 있는 기초 통계 정보는 다음과 같다.

이름	내용
변수명	변수 이름
변수유형	변수의 유형 (수치형, 카테고리형)
최소값	가장 작은 값
최대값	가장 큰 값
1분위	전체 값의 하위 25%에 해당하는 값
3분위	전체 값의 하위 75%에 해당하는 값
유일값	데이터 중복 제거 후의 유일한 값
결측값	데이터가 공란으로 비어 있거나 누락된 값 **
사용	변수의 사용 여부
목표 변수	예측하고자 하는 대상 변수

** 각 변수에 결측값 변환 설정을 하지 않을 경우 자동으로 0으로 대체된다.

3.3.2. 데이터 스케일

각 데이터의 범위는 서로 다르기 때문에 범위를 동등하게 조정하는 작업이 필요하다. 이를 스케일링(Scaling) 작업이라 한다. 데이터의 범위를 늘리거나 줄이는 방식으로 여러 변수들이 같은 범위에 오도록 한다.

예를 들어, 키와 몸무게를 통해 100m 달리기에 걸리는 시간을 예측한다고 하면, 키와 몸무게는 범위가 다르기 때문에 더 큰 값을 가진 키 값이 결과값에 큰 영향을 끼칠 수 있다. 키와 몸무게를 0~1 범위로 변환하는 것이다. 예) 키 175 센티미터 → 0.175

데이터의 범위를 조정하는 방법은 표준화 스케일, 최소-최대 스케일, 로버스트 스케일, 최대-절대값 스케일, 정규화 스케일이 있다.

The screenshot shows the '데이터 탐색' (Data Exploration) step in a workflow. A dropdown menu is open over the 'Standard Scale' option, showing other scaling methods: MinMax Scale, Robust Scale, MaxAbs Scale, and Normalize. The table below shows variables with their types, min, max, unique, missing, and target values, along with scaling options.

변수명	변수유형	최소값	최대값	유일값	결측값	평균	표준편차	사용	목표변수
Passengerid	Nume...	1	891	891	0	446		<input type="checkbox"/>	<input type="checkbox"/>
Survived	Nume...	0	1	2	0	0.38		<input checked="" type="checkbox"/>	<input type="checkbox"/>
Pclass	Nume...	1	3	3	0	2.31	0.84	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Sex	Categ...			2	0	0	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Age	Nume...	0.42	80	88	177	29.7	14.53	<input checked="" type="checkbox"/>	<input type="checkbox"/>
SibSp	Nume...	0	8	7	0	0.52	1.1	<input checked="" type="checkbox"/>	<input type="checkbox"/>

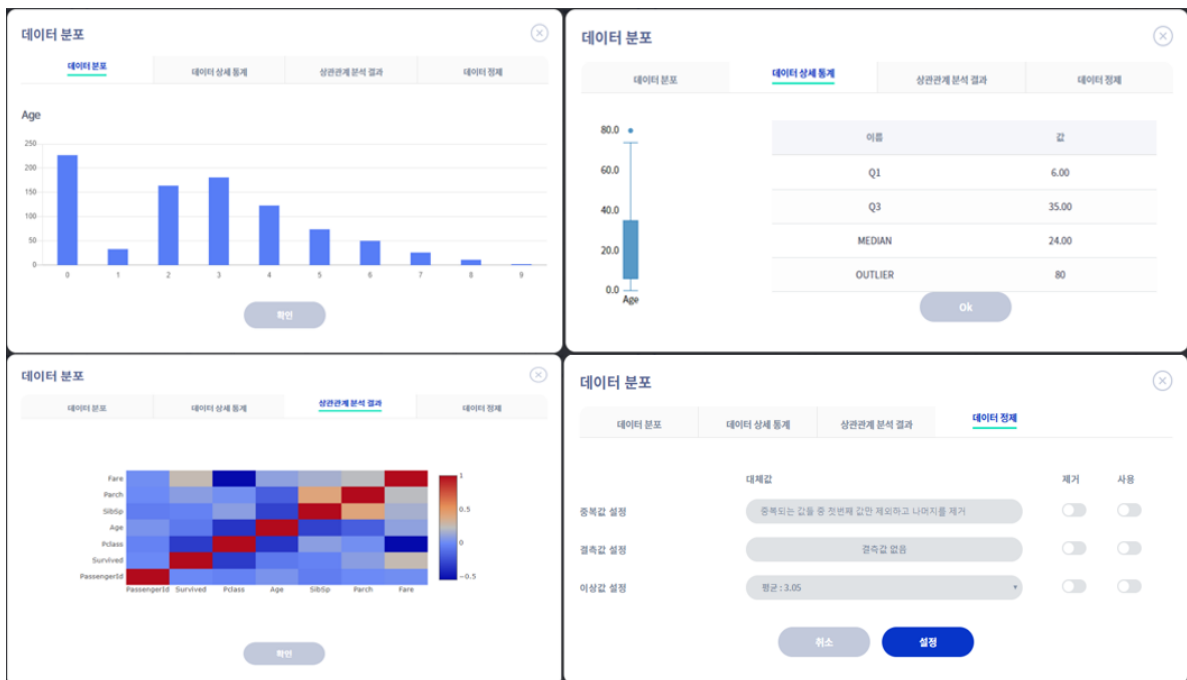
- 데이터 스케일 변경 화면이다.

- 데이터 스케일 방법은 변경할 수 있으며, 데이터 스케일은 표준화 스케일, 최대-최소 스케일, 로버스트 스케일, 최대-절대값 스케일을, 정규화 스케일을 지원한다.
- WiseProphet 에서 지원하는 데이터 스케일 방법은 다음과 같다.

이름	내용
표준화 (Standard Scale)	데이터의 범위를 정규분포로 변환한다. 정규 분포는 데이터의 중심값을 기준으로 좌우 대칭의 분포 형태를 나타낸다.
최대-최소 스케일 (Min-Max Scale)	데이터의 범위를 0~1 사이의 값으로 변환한다.
로버스트 스케일 (Robust Scale)	중앙값과 사분위 범위를 기준으로, 데이터의 범위를 변환한다. 이상값이 많을 경우 이상값의 영향을 최소화하기 위해 사용된다.
최대절대값 스케일 (Maxabs Scale)	최대 절대값과 0 이 각각 1,0 이 되도록 데이터 범위를 변환한다. 양수 데이터로만 구성된 데이터에는 MinMax Scale 과 유사하게 동작하며, 단점으로는 큰 이상치에 민감할 수 있다.
정규화 스케일	실제 값 범위를 표준 값 범위(일반적으로 -1~+1 또는 0~1)로 변환하는 과정

3.3.3. 데이터 분포

데이터의 수가 적으면 데이터의 값을 살펴볼 수 있지만, 데이터의 수가 많다면 숫자가 어떤 값 근처에 어떤 모양으로 모여 있는지 전반적인 형태를 살펴보는 작업이 필요하다.



데이터 값의 형태를 데이터 분포(Data Distribution)라 한다.

- 각 특징의 데이터 분포를 확인할 수 있는 화면이다.
- 돋보기 버튼을 클릭하면 해당 변수의 분포, 상세 통계, 변수들 간의 상관관계 확인할 수 있으며 변수의 중복값, 결측값, 이상값에 대한 데이터 정제를 할 수 있다.
- 데이터 분포 탭에서는 변수의 데이터 분포를 보여주는 히스토그램, 데이터 상세 통계에서는 1 분위, 3 분위, 평균, 이상치를 보여주는 박스플롯을 확인할 수 있으며 상관관계 분석 결과에서는 변수들 간의 상관관계를 확인할 수 있다.
- 히스토그램은 데이터 값이 가질 수 있는 범위를 몇 개의 구간으로 나누고, 각 구간에 해당하는 값의 빈도를 계산하는 방법이다.
- 1 분위, 3 분위수는 3.3.2의 표에 설명되어 있으며 이상치는 1 분위수보다 작거나 3 분위수보다 큰 관측값을 의미한다.
- 상관관계는 변수 A와 B 사이의 상관관계 정도를 나타내는 수치이다. -1과 1 사이의 값을 가지며 절대값이 1에 가까울수록 두 변수 간의 상관관계가 높다.
- 데이터 정제 탭에서는 변수의 중복값, 결측값, 이상값에 대하여 사용 유무 및 대체값 설정 그리고 제거를 할 수 있다. 사용 버튼을 클릭하면 제거 버튼 및 대체값 설정이 활성화되고, 제거 버튼을 클릭하여 제거를 하거나 대체값을 설정하여 데이터 정제를 한다.
- 데이터 정제를 설정하면 데이터 초기화 버튼이 활성화되며, 데이터를 원상태로 복원하고 싶을 경우 해당 버튼을 클릭하면 된다.

3.3.4. 변수 유형 변경

변수 유형을 변경할 수 있으며, 변수 유형에는 범주형, 수치형이 있다.

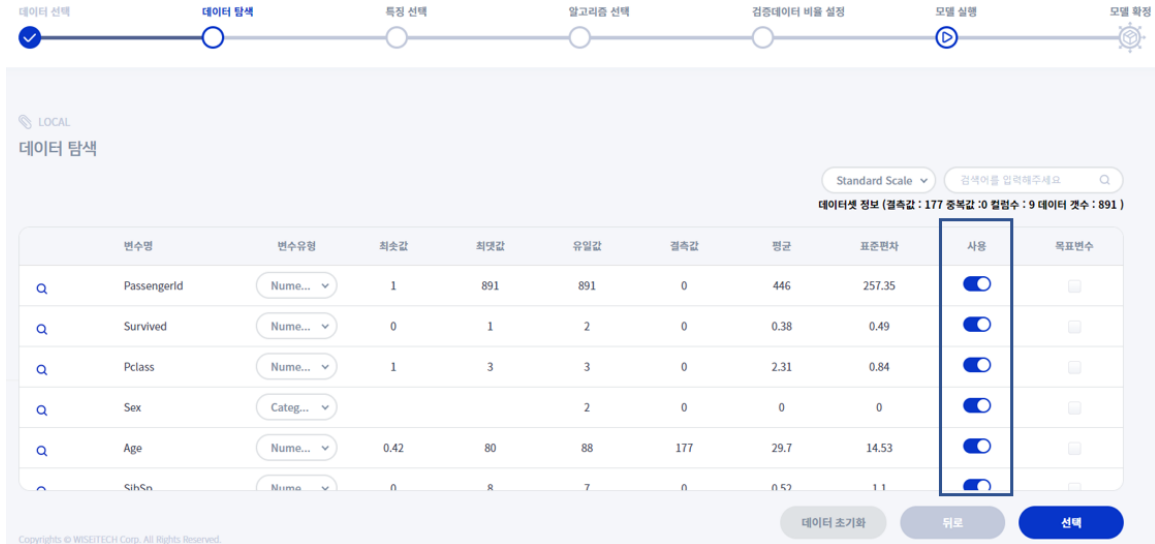
The screenshot shows a data analysis tool interface with a progress bar at the top containing steps: 데이터 선택, 데이터 탐색, 특징 선택, 알고리즘 선택, 검증데이터 비율 설정, 모델 실행, and 모델 최적. The '데이터 탐색' step is active. Below the progress bar, there's a search bar and a table of variables. The 'PassengerId' variable is selected, and a dropdown menu is open showing options: Numerical (selected), Categorical, Numerical, and Nume... (truncated). The table below shows various variables with their statistics and type indicators.

변수명	변수유형	최소값	최댓값	유일값	결측값	평균	표준편차	사용	목표변수
PassengerId	Nume...	1	891	891	0	446	257.35	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Survived	Categorical	0	1	2	0	0.38	0.49	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Pclass	Numerical	1	3	3	0	2.31	0.84	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Sex	Categ...			2	0	0	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Age	Nume...	0.42	80	88	177	29.7	14.53	<input checked="" type="checkbox"/>	<input type="checkbox"/>
SibSp	Nume...	0	8	7	0	0.57	1.1	<input checked="" type="checkbox"/>	<input type="checkbox"/>

At the bottom right, there are buttons for '데이터 초기화', '뒤로', and '선택'.

3.3.5. 변수 사용 여부 설정

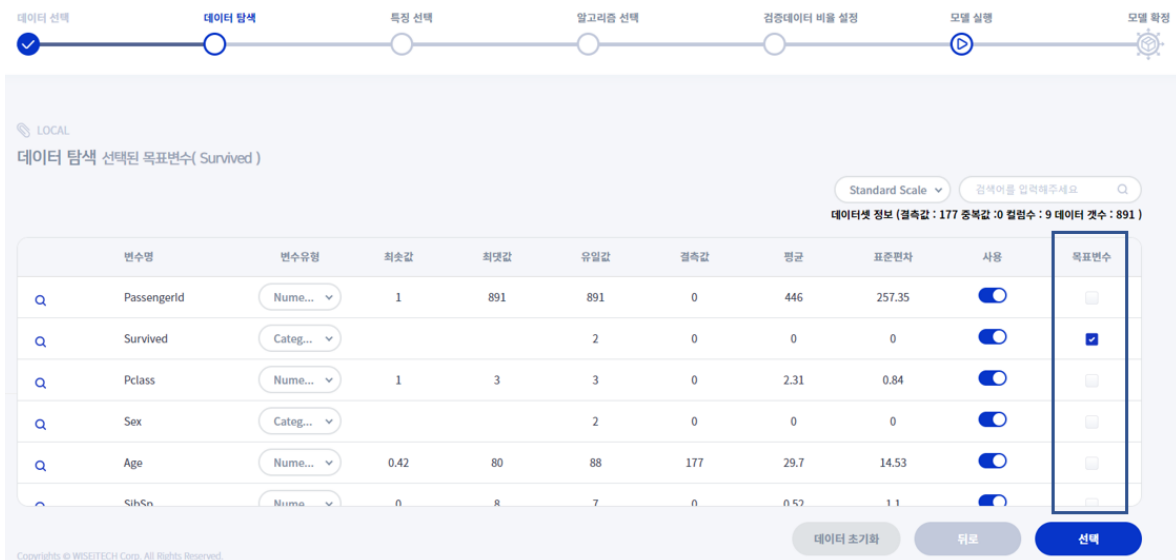
사용자는 예측에 필요한 변수의 사용 여부를 설정할 수 있다. 사용자는 분석 모델 생성시 불필요한 변수는 제외시킬 수 있다.



- 변수 사용 여부를 설정하는 화면이다.
- 불필요한 변수를 확인하여 사용 버튼을 클릭하여 사용 여부를 해제한다.

3.3.6. 목표 변수 설정

- 사용자는 예측하고자 하는 대상 변수를 목표변수로 설정한다.
예) 주가 예측의 목표 변수: 주가, 타이타닉 생존자 예측의 목표변수: 생존여부



- 목표 변수를 설정하는 화면이다.
- 사용자는 분석 목적에 맞는 변수를 확인하여 목표 변수로 설정하며, 클러스터링의 경우 목표 변수를 설정하지 않는다.
- 목표 변수를 설정할 경우 특징 선택 단계로 넘어가며, 설정하지 않을 경우 알고리즘 선택 단계로 넘어간다.

3.4. 특징 선택

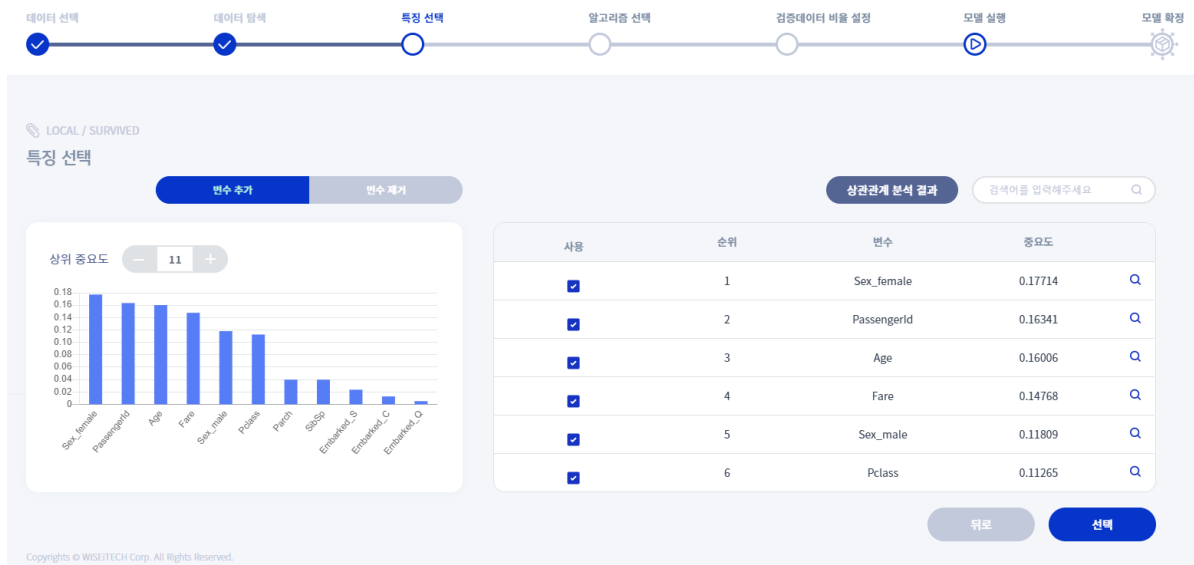
특징 선택은 목표 변수를 분류/예측하고자 할 때 최적의 특징들을 선택하는 것이다. 예를 들어 100m 달리기 선수들의 기록(목표 변수)을 예측할 때, 선수 평균 기록, 나이, 신장, 몸무게, 우승 경험 등 다양한 요인들이 영향을 미칠 것이다. 다양한 요인들 중에서 선수의 기록을 예측하는데 필요한 특징을 추가하거나 불필요한 특징은 제외한다.

특징 선택 단계에서는 데이터 탐색 단계에서 ‘사용함’으로 설정한 특징만을 이용하여 목표 변수에 미치는 중요도에 따라 특징을 추출합니다. 목표 변수를 선택하지 않았을 경우 해당 단계는 건너뛰게 된다.

최적의 특징을 선택하는 방법에는 (1) 특징별 영향 (2) 상관관계 분석방법이 있다.

3.4.1. 특징별 영향

특징별 영향은 목표 변수에 대한 각 변수들의 상대적 중요도를 의미한다. 예를 들어, 선수 기록을 예측하는 데 평균 기록 0.4, 나이 0.3, 신장 0.1, 몸무게 0.1의 중요도 값을 나타내면 선수 기록을 예측하는데 “평균 기록”이 40% 정도 영향을 미친다는 것을 의미한다.



- 특징 선택에서 특징별 영향 화면에서 변수 추가를 클릭한 화면이다.
- 1) 변수 추가를 클릭하면 중요도 순서가 높은 순서대로 특징을 선택할 수 있다
- 2) 플러스 버튼을 클릭하면 변수가 중요도가 높은 순으로 추가되고, 마이너스 버튼을 클릭하면 중요도가 낮은 순으로 제외된다.
- 3) 표 각 변수 오른쪽의 돋보기 버튼을 클릭하면 이전 데이터 탐색에서 설정한 스케일 방법에 의해 스케일된 각 특징별 데이터의 분포를 보여준다.

데이터 선택 데이터 탐색 **특징 선택** 알고리즘 선택 검증데이터 비율 설정 모델 실행 모델 확장

LOCAL / SURVIVED

특징 선택

변수 추가 변수 제거 상관관계 분석 결과 검색어를 입력해주세요

하위 중요도 - 0.00 + 이하 제거

사용	순위	변수	중요도
<input checked="" type="checkbox"/>	1	Sex_female	0.17714
<input checked="" type="checkbox"/>	2	PassengerId	0.16341
<input checked="" type="checkbox"/>	3	Age	0.16006
<input checked="" type="checkbox"/>	4	Fare	0.14768
<input checked="" type="checkbox"/>	5	Sex_male	0.11809
<input checked="" type="checkbox"/>	6	Pclass	0.11265

뒤로 선택

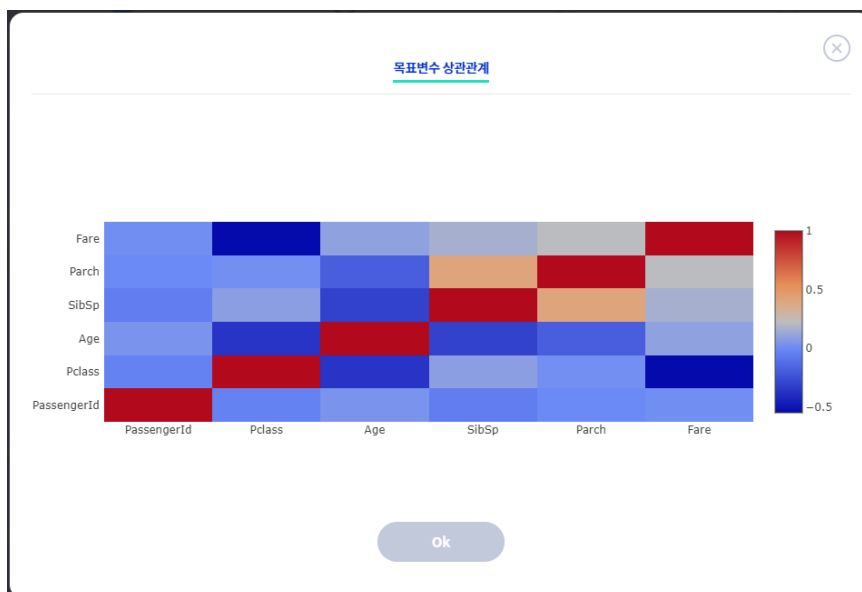
Copyrights © WISETECH Corp. All Rights Reserved.

- 특징 선택에서 특징별 영향 화면에서 변수 제거를 클릭한 화면이다.
- 1) 플러스, 마이너스 버튼을 누르면 0.01 단위로 증감하며, 지정 중요도 이하의 특징이 변수에서 제외된다.
- 2) 사용자는 중요도를 직접 입력하여 사용자가 지정한 중요도 이하의 특징을 변수에서 제외시킬 수 있다.

3.4.2. 상관 관계

상관 관계는 두 변수 간의 상관관계의 정도를 나타내는 수치이다. -1~1 사이의 값을 가지며 절대값이 1에 가까울수록 두 변수 간의 상관관계의 정도가 높다고 할 수 있다. 상관 관계를 이용하여 변수를 선택할 때는 두 변수 간의 상관관계가 너무 높은 경우

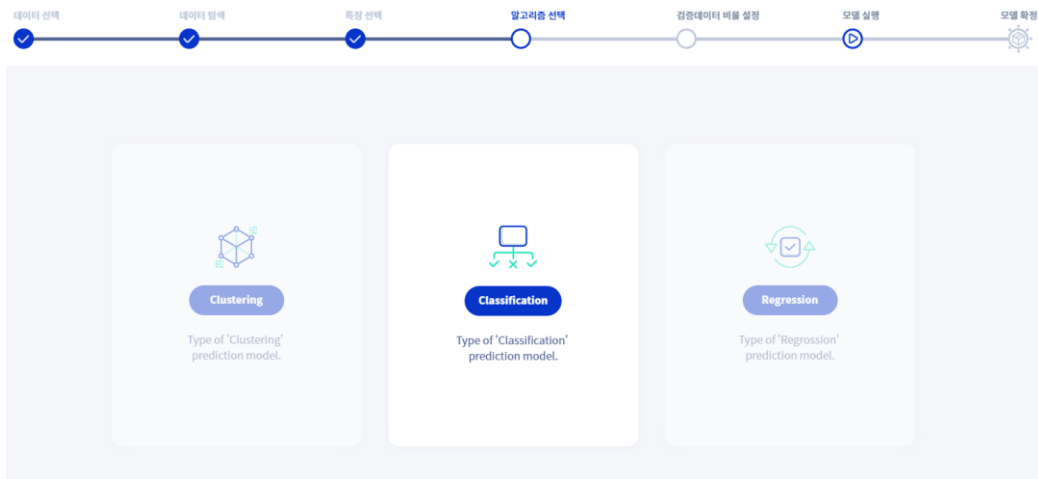
(80% 이상)는 두 변수 중 하나의 변수는 제외한다.



- 상관관계 분석 결과를 클릭하면 목표 변수로 설정된 분석 대상 변수와 다른 변수들 간의 상관관계를 확인할 수 있다.

3.5. 알고리즘 선택

알고리즘 선택 단계에서는 예측에 사용할 알고리즘을 선택한다. 예측 모델 유형에는 클러스터링, 분류, 회귀가 있다.

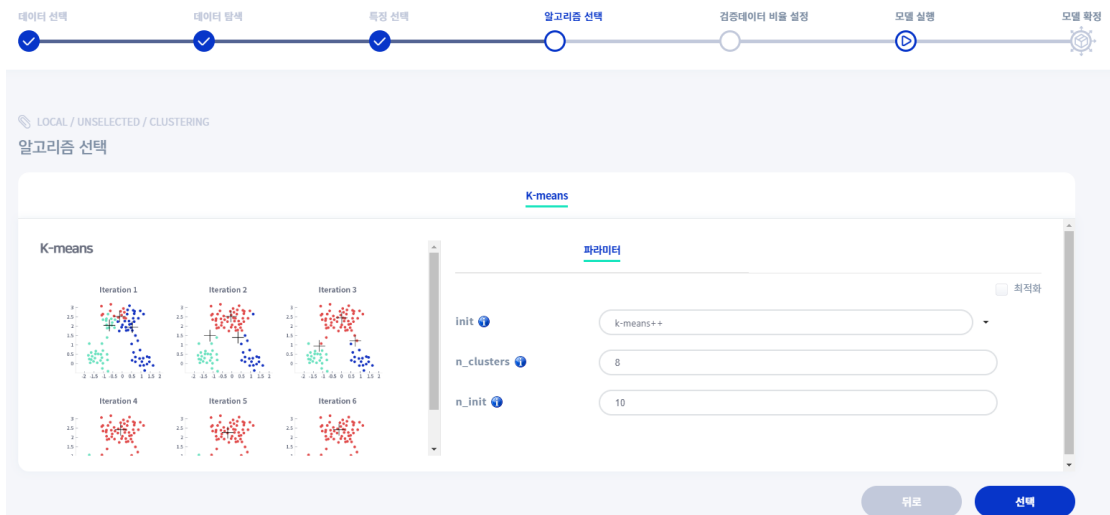


- 알고리즘 선택 화면이다.
- 1) 예측 모델 유형에는 클러스터링, 분류, 회귀가 있으며 데이터 탐색 단계에서 설정한 목표 변수의 유형에 따라 수치형은 회귀, 범주형은 분류, 설정하지 않은 경우 클러스터링 버튼이 활성화된다.
- 2) 예측 모델 유형을 선택하면 해당 유형의 알고리즘 선택 화면이 보인다.

3.5.1. 클러스터링 모델

클러스터링은 유사한 특징을 가진 데이터끼리 묶는 것을 의미한다. 대표적인 비지도 학습 기법 중의 하나로 목표 변수의 설정이 필요 없다.

- ① K-means: K 개의 Centroid 를 기반으로 K 개의 클러스터를 만들어주는 클러스터링 모델 중 하나이다.



- 1) 모델 유형에서 “클러스터링”을 선택한다.
- 2) K-means 를 선택한다.
- 3) 매개변수를 입력한다. 특히 n_clusters (클러스터의 수)는 반드시 입력해야 한다.

매개변수 명	설 명
init	클러스터 중심 초기화 방법 - k-means ++: 수렴 속도를 높이기 위해 효율적으로 초기 군집 중심을 선택. - random: 무작위로 k 개 선택
n_clusters	- 클러스터 개수
n_init	- 클러스터 중심의 초기화 횟수

3.5.2. 분류 모델

분류는 어떤 카테고리에 해당하는지 분류하는 기법으로 카테고리형 변수를 예측하기 위해 사용된다.

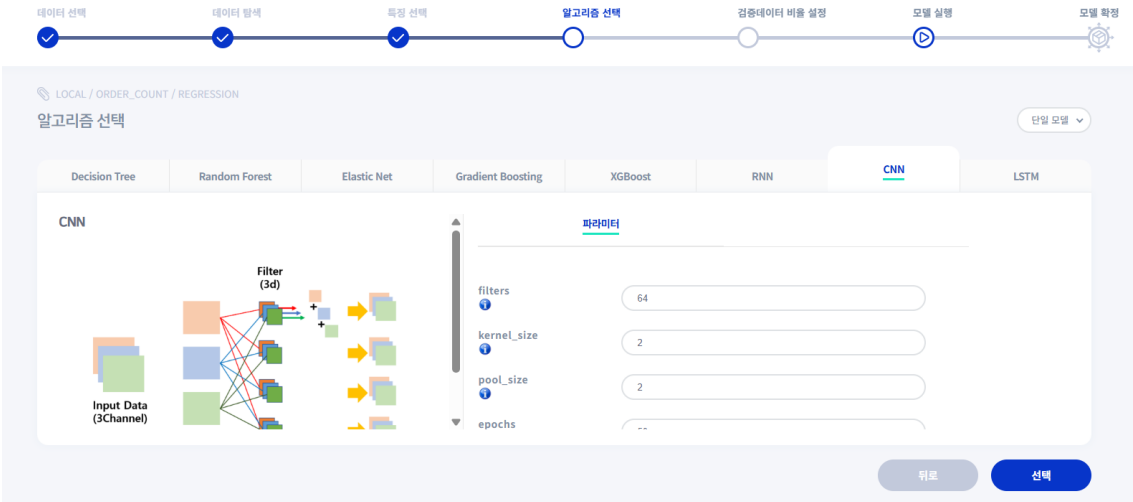
- ① 의사 결정 나무: 데이터를 분석하여 이들 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 나타내며 그 모양이 나무와 같다고 해서 의사결정나무라 불린다. 질문을 던져서 대상을 좁혀가는 ‘스무고개’놀이와 비슷한 개념이다.

- 1) 모델 유형에서 “분류”를 선택한다.
- 2) Decision Tree 를 선택한다.
- 3) 매개변수를 입력한다. 매개변수는 아래의 설명을 참고하여 설정한다.

매개변수 명	설 명
criterion	트리 분리 기준 - gini : 불순도, 집합에 이질적인 것이 얼마나 섞여 있는가를 나타내는 척도 - entropy : 불확실성의 정도

max_depth	-얼마나 깊게 트리를 만들 것인가의 기준
min_samples_leaf	-노드가 되려면 가지고 있어야 할 최소의 샘플 수

② CNN: 합성곱 계층을 사용하여 데이터를 학습하는 딥러닝 알고리즘이다. 데이터의 패턴을 분석하여 이산적인 클래스(범주형 값)를 예측하는 데 사용할 수 있다.
 ※ CNN 알고리즘은 GPU 가속이 필요한 연산을 포함하고 있으며, GPU 사용이 결재된 계정에서만 제공된다.



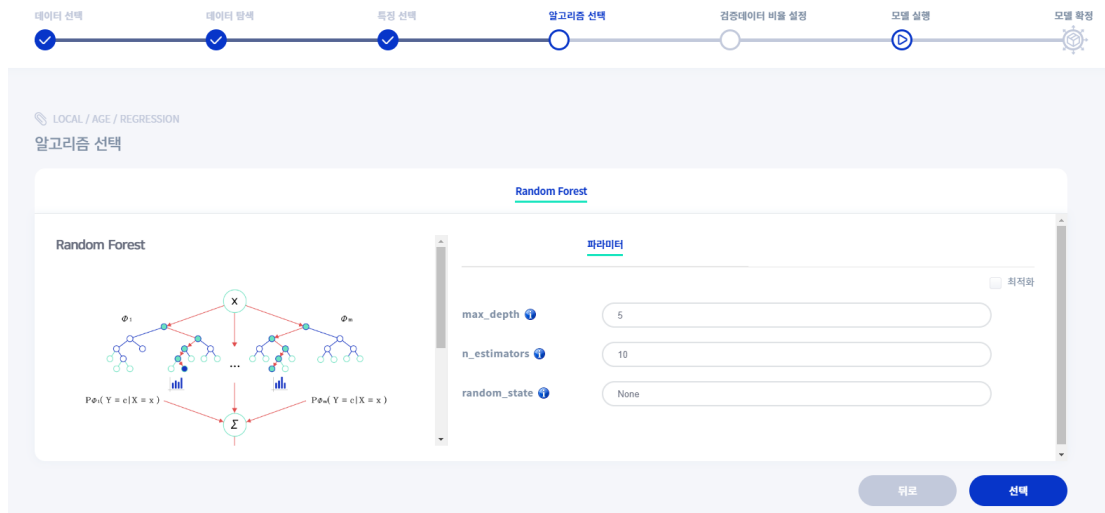
- 1) 모델 유형에서 “회귀”를 선택한다.
- 2) CNN 을 선택한다.
- 3) 매개변수를 입력한다. 매개변수는 아래의 설명을 참고하여 설정한다.

매개변수 명	설 명
filter	-특징맵의 개수. 각 필터는 입력 데이터에서 다른 특징을 추출하므로 필터 수가 많을수록 학습 능력이 향상된다.
kernel_size	-필터의 가로 x 세로 크기. 너무 크면 세부 특징을 놓치고, 너무 작으면 전체 구조를 파악하기 어렵다.
pool_size	-풀링 계층에서 사용되는 축소 영역의 크기입니다. 풀링은 연산량을 감소시켜 일반화 성능을 높인다.
epochs	-학습 횟수. 전체 학습 데이터를 몇 번 반복해서 학습할지 설정한다.
batch_size	-학습 시 한 번에 처리하는 데이터 샘플의 수. 클수록 학습 속도는 빨라지지만 더 많은 메모리를 사용한다.
loss	-손실함수. 각 배치마다 사용되며, 해당 배치에서 예측한 값과 실제값이 얼마나 다른지 계산한다. 이를 바탕으로 모델이 더욱 정확하게 예측할 수 있도록 학습 방향을 조정한다.

3.5.3. 회귀 모델

회귀는 연속값(연속되는 수치)를 예측할 때 사용된다.

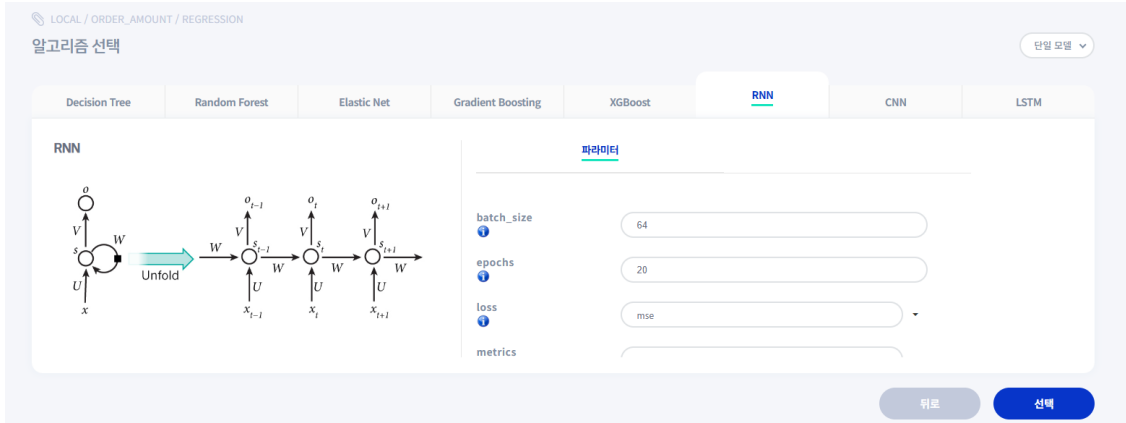
- ① 랜덤포레스트: 랜덤포레스트는 분류, 회귀 분석에 사용되는 앙상블 학습 방법의 일종으로 훈련 과정에서 구성된 다수의 결정 트리로부터 분류 또는 평균 예측치를 출력함으로써 동작한다. 여러 개의 의사 결정 트리를 만들고 투표를 시켜 다수결로 결과를 결정하는 방법이다.



- 1) 모델 유형에서 “회귀”를 선택한다.
- 2) Random Forest 를 선택한다.
- 3) 매개변수를 입력한다. 매개변수는 아래의 설명을 참고하여 설정한다.

매개변수 명	설 명
max_depth	-얼마나 깊게 트리를 만들 것인가의 기준
n_estimators	-트리의 수
random_state	-난수 초기값

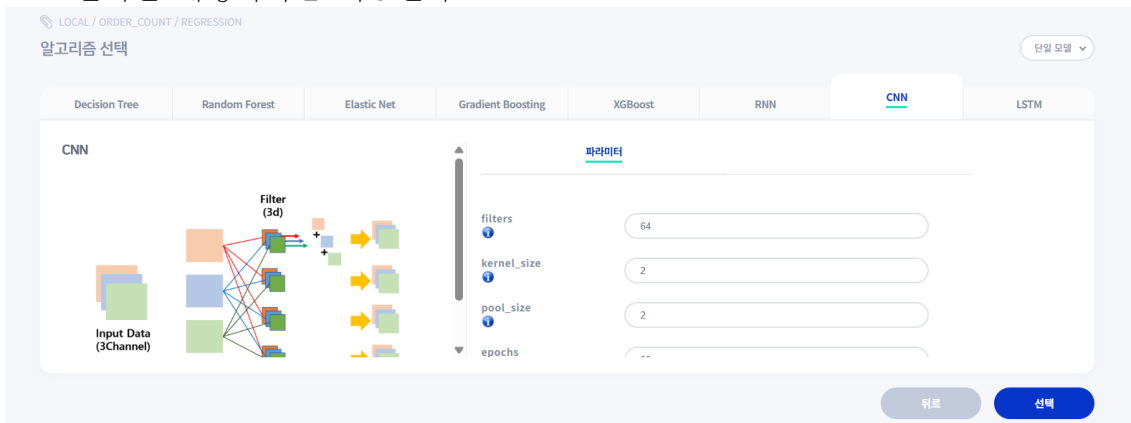
- ② RNN: 순서가 있는 데이터를 처리하는 데 적합한 딥러닝 알고리즘이다. 이전의 입력 정보를 기억하여 다음 예측에 반영하는 구조로, 시계열 데이터(날짜 칼럼이 있는 데이터)만 학습 가능하다.
 ※ RNN 알고리즘은 GPU 가속이 필요한 연산을 포함하고 있으며, GPU 사용이 결재된 계정에서만 제공된다.



- 1) 모델 유형에서 “회귀”를 선택한다.
- 2) RNN 을 선택한다.
- 3) 매개변수를 입력한다. 매개변수는 아래의 설명을 참고하여 설정한다.

매개변수 명	설 명
batch_size	- 학습 시 한 번에 처리하는 데이터 샘플의 수. 클수록 학습 속도는 빨라지지만 더 많은 메모리를 사용한다.
epochs	- 학습 횟수. 전체 학습 데이터를 몇 번 반복해서 학습할지 설정한다.
loss	- 손실함수. 각 배치마다 사용되며, 해당 배치에서 예측한 값과 실제값이 얼마나 다른지 계산한다. 이를 바탕으로 모델이 더욱 정확하게 예측할 수 있도록 학습 방향을 조정한다.
metrics	-모델의 성능을 평가하는 기준. 예측이 얼마나 정확한지 확인하기 위해 사용된다.
optimizer	-손실함수를 최소화하기 위해 가중치를 조정하는 알고리즘. 모델 학습의 속도와 안전성에 영향을 준다.
window_size	시계열 데이터를 학습할 때 한 번에 참고하는 데이터의 길이. 과거 데이터 중 몇 개 시점을 기반으로 다음 값을 예측할지 결정한다.

- ③ CNN: 합성곱 계층을 사용하여 데이터를 학습하는 딥러닝 알고리즘이다. 주로 데이터의 패턴을 분석하여 연속적인 값을 예측하는 데 사용된다.
 ※ CNN 알고리즘은 GPU 가속이 필요한 연산을 포함하고 있으며, GPU 사용이 결제된 계정에서만 제공된다.



- 1) 모델 유형에서 “회귀”를 선택한다.
- 2) CNN 을 선택한다.
- 3) 매개변수를 입력한다. 매개변수는 아래의 설명을 참고하여 설정한다.

매개변수 명	설 명
filter	-특징맵의 개수. 각 필터는 입력 데이터에서 다른 특징을 추출하므로 필터 수가 많을수록 학습 능력이 향상된다.
kernel_size	-필터의 가로 x 세로 크기. 너무 크면 세부 특징을 놓치고, 너무 작으면 전체 구조를 파악하기 어렵다.
pool_size	-풀링 계층에서 사용되는 축소 영역의 크기입니다. 풀링은 연산량을 감소시켜 일반화 성능을 높인다.
epochs	-학습 횟수. 전체 학습 데이터를 몇 번 반복해서 학습할지 설정한다.
batch_size	-학습 시 한 번에 처리하는 데이터 샘플의 수. 클수록 학습 속도는 빨라지지만 더 많은 메모리를 사용한다.
loss	-손실함수. 각 배치마다 사용되며, 해당 배치에서 예측한 값과 실젯값이 얼마나 다른지 계산한다. 이를 바탕으로 모델이 더욱 정확하게 예측할 수 있도록 학습 방향을 조정한다.

④ LSTM: RNN 의 한계를 보완한 구조로, 긴 시계열 데이터에서도 중요한 정보를 오래 기억할 수 있도록 설계된 딥러닝 알고리즘이다. 시계열 데이터(날짜 칼럼이 있는 데이터)만 학습 가능하다.

※ LSTM 알고리즘은 GPU 가속이 필요한 연산을 포함하고 있으며, GPU 사용이 결제된 계정에서만 제공된다.

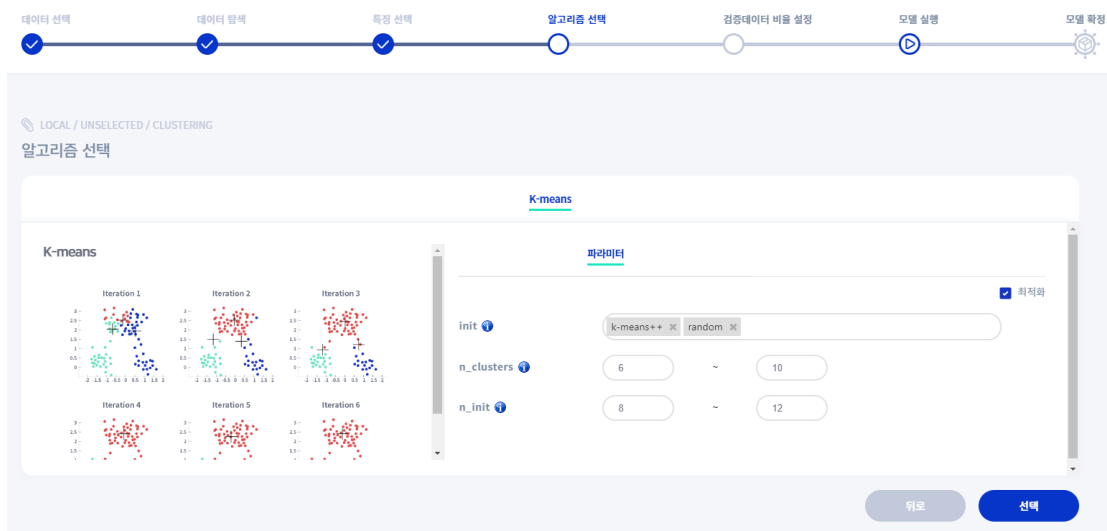
- 1) 모델 유형에서 “회귀”를 선택한다.
- 2) CNN 을 선택한다.
- 3) 매개변수를 입력한다. 매개변수는 아래의 설명을 참고하여 설정한다.

매개변수 명	설 명
batch_size	- 학습 시 한 번에 처리하는 데이터 샘플의 수. 클수록 학습 속도는 빨라지지만 더 많은 메모리를 사용한다.

epochs	- 학습 횟수. 전체 학습 데이터를 몇 번 반복해서 학습할지 설정한다.
loss	- 손실함수. 각 배치마다 사용되며, 해당 배치에서 예측한 값과 실제값이 얼마나 다른지 계산한다. 이를 바탕으로 모델이 더욱 정확하게 예측할 수 있도록 학습 방향을 조정한다.
metrics	-모델의 성능을 평가하는 기준. 예측이 얼마나 정확한지 확인하기 위해 사용된다.
optimizer	-손실함수를 최소화하기 위해 가중치를 조정하는 알고리즘. 모델 학습의 속도와 안전성에 영향을 준다.
window_size	시계열 데이터를 학습할 때 한 번에 참고하는 데이터의 길이. 과거 데이터 중 몇 개 시점을 기반으로 다음 값을 예측할지 결정한다.

3.5.4. 파라미터 최적화

최적화는 모델 생성시 알고리즘의 파라미터 최적화를 위해 사용된다. Wise Prophet 은 파라미터 최적화를 통해 최적의 모델 생성을 지원한다.



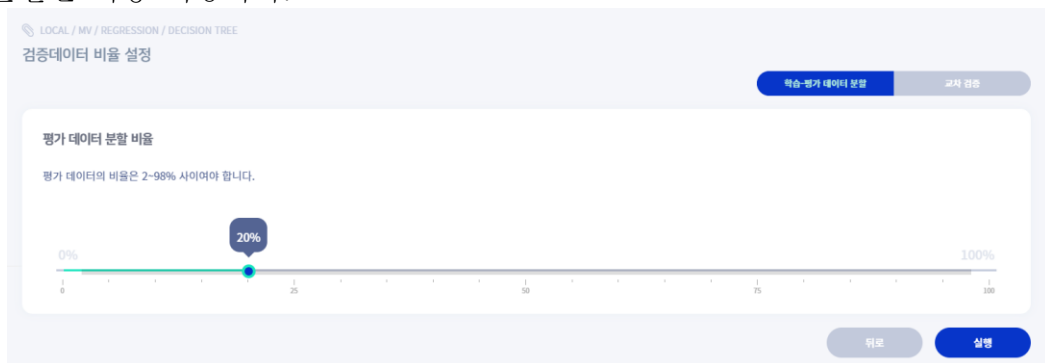
- 파라미터 최적화 화면이다.
- 1) 알고리즘 선택 화면에서 최적화 버튼을 클릭한다.
- 2) 알고리즘의 각 파라미터 별로 값의 범위를 지정한다.
- 3) 선택 버튼을 누르면 지정한 범위 내에서 최적의 파라미터 값을 찾아준다.
- 4) 최적화된 파라미터 값은 모델 실행 후에 모델 로그 창에서 확인할 수 있다.

3.6. 검증데이터 비율 설정

알고리즘을 선택하면 데이터를 학습데이터와 평가 데이터를 나누어 검증 단계를 거친다. 모델 검증 방법은 훈련-평가 데이터 분할과 교차검증을 지원한다.

3.6.1. 훈련-평가 데이터 분할

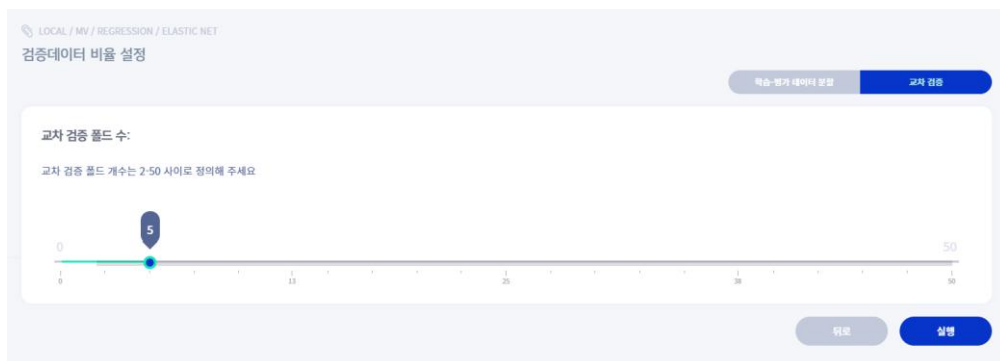
주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 검증을 실시하는 방법이다. 하나는 모형의 학습 및 구축을 위한 훈련용 데이터로 하나는 평가 데이터로 사용한다. 알고리즘 선택 단계에서 최적화를 선택한 경우 훈련-평가 데이터 분할만 사용 가능하다.



- 검증데이터 비율 설정화면에서 훈련-평가 데이터 분할 화면이다.
- 1) 훈련-평가 데이터 분할을 클릭한다.
- 2) 평가 데이터 분할 비율을 설정한다. 비율을 20%으로 설정하면 학습 데이터 비율을 80%, 평가 데이터 비율은 20%가 된다.
- 3) 실행 버튼을 클릭하면 예측 모델이 실행된다.

3.6.2. 교차 검증

교차 검증은 데이터를 폴드라고 하는 비슷한 크기의 부분 집합으로 나누고, 각각의 1 개 폴드를 평가 데이터로 나머지를 학습데이터로 폴드 수만큼 교차 검증을 실행하는 방식이다.



- 검증데이터 비율 설정화면에서 교차 검증 화면이다.
- 1) 교차 검증을 클릭한다.
- 2) 교차 검증 폴드 수를 설정한다.
- 3) 실행 버튼을 클릭하면 예측 모델이 실행된다.

3.7. 모델 실행

모델 실행은 클러스터링, 분류, 회귀 예측 유형에 따라 모델 평가 지표를 보여준다.

3.7.1. 클러스터링

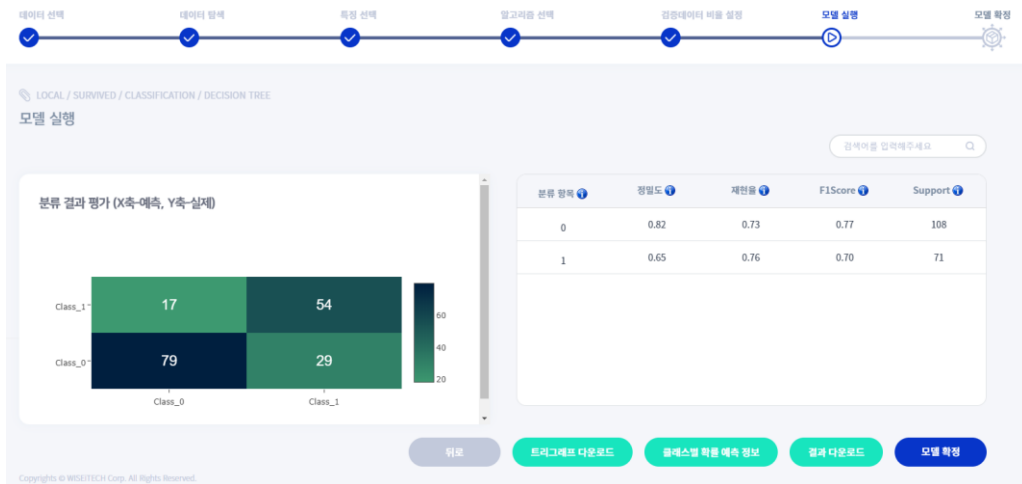
클러스터링 모델을 실행하면 사용자가 선택한 클러스터의 수 K 개의 클러스터 집합으로 나누어지며, 모델 실행 그래프를 통해 확인할 수 있다. 클러스터링의 경우 비지도 학습에 해당하므로 모델 실행 결과에 대한 세부 검증 지표는 나타내지 않는다.



- 클러스터링 모델 실행 화면이다.
- 1) 클러스터링 실행 결과를 시각화를 통해 확인할 수 있다. (색상=클러스터 구분)
- 2) 예측 실행 결과를 다운로드 받고 싶을 경우 결과 다운로드 버튼을 클릭하면 csv 형태로 파일을 다운받을 수 있다.
- 3) 모델 확장 버튼을 클릭하면 모델 저장 화면이 나타나고, 모델명을 입력하고 저장 버튼을 누르면 모델이 저장된다.

3.7.2. 분류

분류 예측 모델은 학습 및 검증 결과에 대한 실젯값과 예측값의 비교 결과를 분류 행렬 (Confusion Matrix)를 통해 시각화 하여 보여주며, 세부 지표인 정밀도, 재현율, F1-스코어, 지지도를 도표화하여 보여준다.



- 분류 모델 실행 화면이다.
- 1) 모델 실행 결과를 분류 행렬을 통해 시각화 하여 보여준다.
- 2) 분류 모델 평가 검증 지표를 도표화하여 표시해준다.
평가 검증 지표 세부 내용은 다음과 같다.

** 예시: 유방암 예측 - 유방암의 양성과 음성 분류 모델 가정

예측 결과

		양성	음성
실제 결과	양성	80	5
	음성	10	5

평가 지표	설명
정확도	전체 데이터 수에 대비하여 실제 결과와 일치한 수의 비율 (유방암 예측 정확도 = $(80 + 5) / 100 = 85\%$)
정밀도	양성으로 예측한 결과에서 실제 결과와 일치한 수의 비율 (유방암 예측 정밀도 = $80 / (80 + 10) = \text{약 } 88\%$)
재현율	실제 결과 양성인 것 중에서 예측 결과와 일치한 수의 비율 (유방암 예측 재현율 = $80 / (80 + 5) = \text{약 } 94\%$)
F1Score	정밀도와 재현율의 조화평균
Support	예측 합계

- 3) decision tree 알고리즘의 경우 트리그래프를 다운로드 받을 수 있으며, 분류형 모델일 경우 클래스별 확률 예측 정보 또한 다운로드 받을 수 있다.

3.7.3. 회귀

회귀 예측 모델은 학습 및 검증 결과에 대한 실젯값과 예측값의 추세선을 시각화하여 보여주며, 예측 결과에 대한 세부 검증 지표인 MSE, RMSE, MAPE 를 도표화하여 보여준다.



- 회귀 모델 실행 화면이다.
- 1) 모델 실행 결과를 실젯값과 예측값의 추세선으로 확인할 수 있다.
- 2) 예측 결과에 대한 세부 검증 지표를 확인한다. 세부 검증 지표에 대한 설명은 다음과 같으며, 각 평가항목의 오른쪽 아이콘에 마우스를 올리면 설명을 확인할 수 있다.

평가 지표	설 명
RMSE	예측값과 실젯값의 오차의 제곱근
MAPE	예측값과 실젯값의 오차를 백분율로 표현

3.7.4. 모델 로그

모델 로그 화면에서는 모델 학습에 실행했던 데이터 분석 로그 기록들을 확인할 수 있다.



- 모델 로그 화면이다.
- 1) 파이 차트는 모델의 예측값에 영향을 준 변수의 목록과 비율이다.
- 2) 클러스터링의 경우 방사선 형태의 차트로 표시되며, 마찬가지로 각각의 클러스터링 집합에 영향을 준 변수의 목록과 비율이다.
- 3) 모델 로그 화면에서는 각각의 데이터 분석에 사용된 알고리즘 유형, 정확도/스코어, 데이터 셋, 목표 변수를 확인할 수 있어 데이터 분석 결과 비교분석에 용이하다. 클러스터링의 경우 정확도/스코어 값이 실루엣 계수를 의미한다.
- 4) 알고리즘 선택 단계에서 최적화를 선택한 경우 최적 파라미터 목록을 보여 준다.

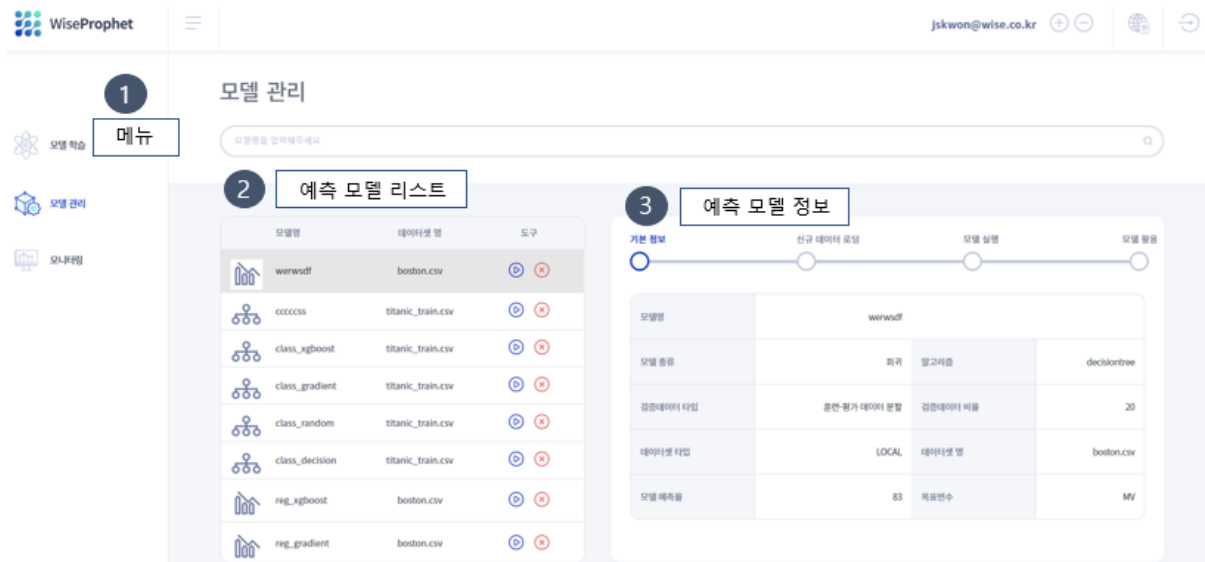
제 4 장 모델 관리

본 장에서는 모델 관리를 위한 WiseProphet 사용법을 기술한다.

4.1. 화면 구성

모델 관리 화면은 1) 메뉴 화면 2) 예측 모델 리스트 3) 모델 관리 메뉴로 구성된다. 모델 관리 화면에서는 모델 학습에서 생성한 예측 모델 리스트를 보여주며 모델 실행, 삭제, 신규데이터 입력 및 예측, 예측 결과 시각화 등의 모델 관리를 할 수 있다.

4.2. 모델 관리



- 모델 관리 화면이다.
- 1) 1 번은 모델 학습 화면과 동일한 메뉴 목록이다.
- 2) 모델 학습 화면에서 생성한 예측 모델의 리스트와 각 예측 모델의 모델명, 모델 유형, 데이터셋 명을 보여준다.
- 3) 예측 모델 리스트의 실행 버튼을 클릭하면 해당모델 생성시 설정한 값들을 불러올 수 있고, 기존 설정 값 변경을 통해 모델을 수정할 수 있다.

- 4) 3 번 예측 모델 정보는 모델 유형, 검증데이터 유형 및 비율, 데이터셋 유형, 목표변수 등의 해당 예측 모델에 대한 정보를 보여준다.
- 5) 3 번에서 신규 데이터 로딩 메뉴를 클릭하여 예측하고 싶은 데이터를 업로드 하여 모델 실행을 통해 예측할 수 있으며, 불필요한 모델은 모델 리스트의 삭제 버튼을 클릭하여 삭제할 수 있다. (단, 업로드 데이터의 변수 목록은 모델 학습 시 사용된 변수목록과 일치해야 한다)
- 6) 데이터가 성공적으로 업로드 되면 예측 모델이 실행된다.
- 7) 위의 과정을 통해 아래와 같이 예측 결과를 확인할 수 있다.



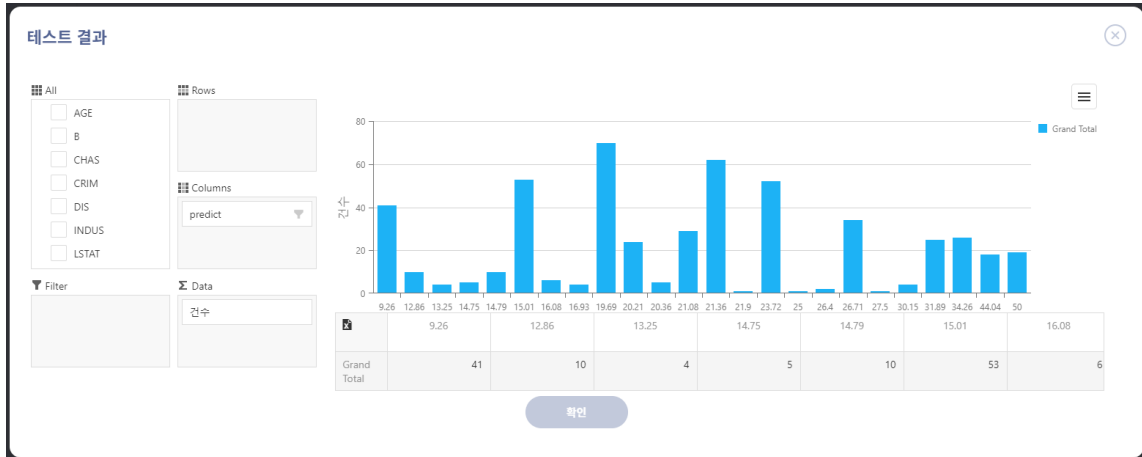
<분류 예측 모델 실행 시>

- 8) 모델 실행이 완료된 후에 모델 활용을 클릭하면 다음과 같은 화면이 나타나며, 업로드 데이터에 대한 예측 결과를 다운로드 받거나 모델을 다운로드 받아 활용할 수 있으며, 예측 결과를 시각화 할 수 있다.






- 9) 시각화를 클릭하면 다음과 같은 화면이 나오며, 시각화 하고자 하는 변수를 선택하면 컬럼 부분에 해당 변수가 추가되고, 로우나 필터에 추가하고 싶은 경우 해당 변수를 드래그 앤 드롭 하여 이동시킬 수 있다. 우측 상단의 버튼을 클릭하

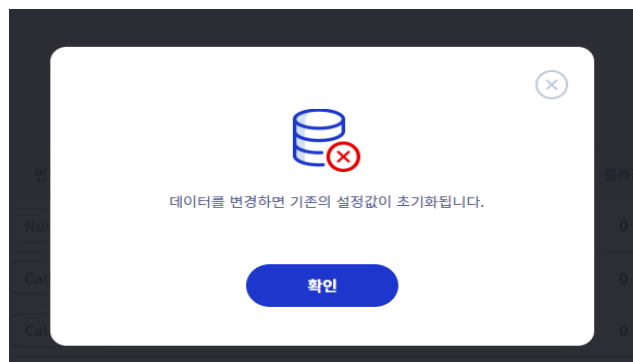
면 그래프를 다운받을 수 있으며, 그래프 아래 표 또한 엑셀 버튼 클릭을 통해 다운 가능하다.



4.3. 모델 수정

모델명	데이터셋 명	도구
 심장병 유무 예측	Heart.csv	 
 desdF_	heart-emp.txt	 

- 모델 관리 리스트에서  버튼을 클릭하면 모델을 수정할 수 있다.
- 1) 모델 수정 시 바로 데이터 탐색 단계로 넘어가며 기존 학습 화면과 동일하게 단계별로 모델을 수정할 수 있다.
- 2) 데이터 정제는 저장된 값으로 우선 변환되며 기존과 동일하게 설정 안된 결측값은 0으로 자동 변환된다.
- 3) 모델수정에서 변수의 컬럼 타입을 변경할 경우 저장된 설정이 초기화 된다.



- 4) 모델수정에서 데이터 초기화는 모델저장 시 설정된 값이 기본값이다.

4.4. 모델 API

기본 정보	신규 데이터 로딩		모델 실행	모델 활용
모델명	model A	알고리즘	decisiontree	
모델 종류	회귀	최적화	N	
검증데이터 타입	훈련-평가 데이터 분할	검증데이터 비율	20	
데이터셋 타입	LOCAL	데이터셋 명	A Shopping Mall Dataset.csv	
모델 예측을	0	목표변수	order_count	
API url	http://localhost:80/public/predict?id=6444&idx=1			

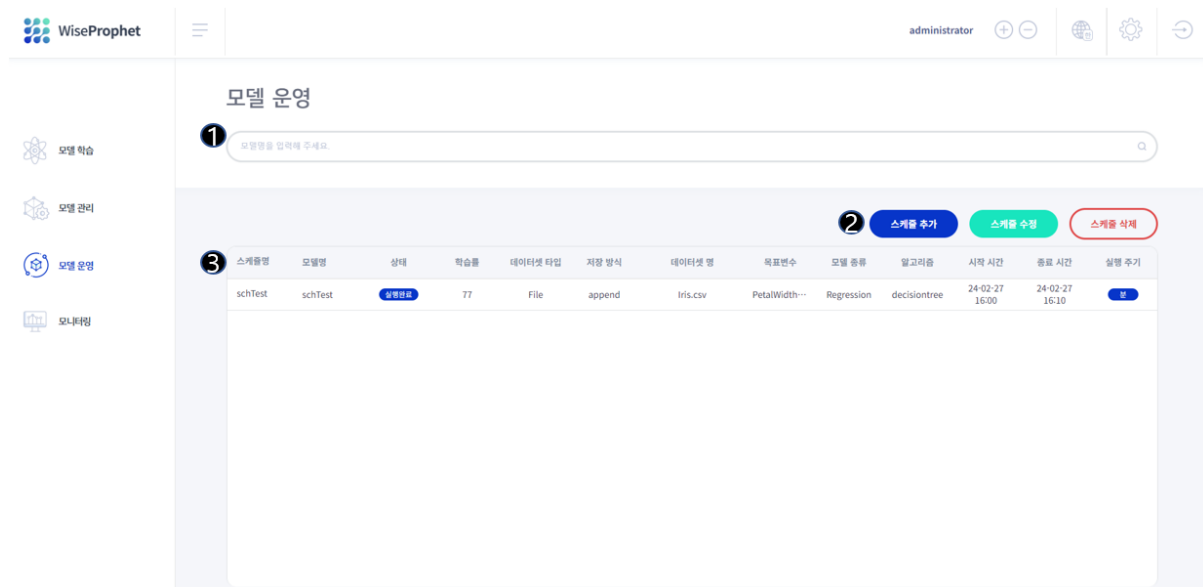
- WiseProphet 에서는 학습된 모델을 활용해 예측 결과를 확인할 수 있도록 모델 추론(예측) API 를 제공한다. 해당 API 는 우측의 예측 모델 정보 최하단에서 확인할 수 있다.
- 1) url 을 클릭하면 클립보드에 자동으로 복사된다.
※ 클립보드 복사가 지원되지 않는 환경에서는 알림 창에서 사용자가 직접 드래그해야 함.
- 2) 해당 API 를 통해 가장 최근에 사용된 데이터를 기반으로 모델 예측 결과를 얻을 수 있다.

제 5 장

모델 운영

본 장에서는 모델 운영을 위한 WiseProphet 사용법을 기술한다.

5.1. 화면 구성



생성한 모델로 스케줄을 설정하여 운영하는 화면으로 1) 스케줄 검색 2) 스케줄 관리 3) 스케줄 리스트 로 구성되어 있다.


- 1) 스케줄명을 입력하여 검색할 수 있다.
- 2) 스케줄을 추가 및 수정 그리고 삭제를 할 수 있다.
- 3) 현재 등록된 스케줄에 대한 모델, 상태, 실행 주기, 알고리즘 등을 확인할 수 있다.


5.2. 설정 및 기능

5.2.1 스케줄 추가

배치 스케줄 ×

스케줄명: *

시작 시간: * 

종료 시간: * 

모델명: * ▼

데이터셋 타입:

데이터셋 명:

재학습: * ▼

실행 단위: ▼ ▼ ▼ ▼ ▼

실행 주기: *

저장 위치 설정

저장 타입: * ▼

파일 저장명: *

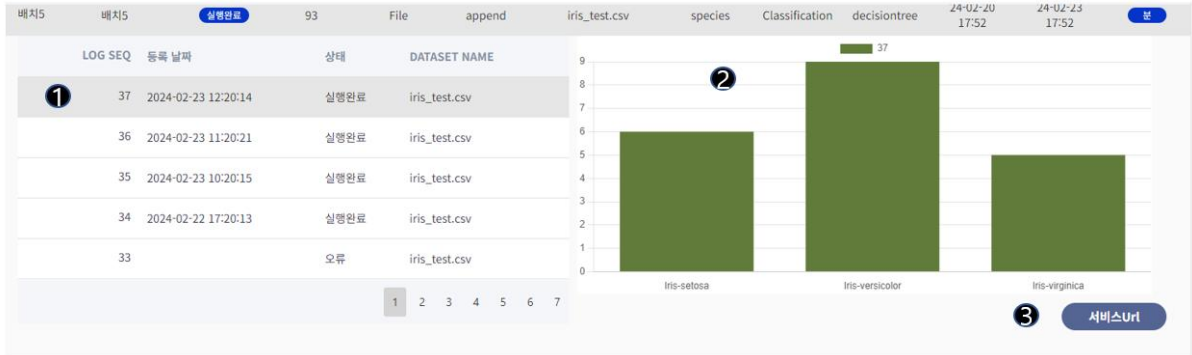
저장 방식: * ▼

모델 학습에서 생성한 모델을 활용하여 스케줄을 만들 수 있다. 스케줄 시작 시간과 종료시간을 설정할 수 있고 실행단위도 설정 가능하다.

‘재학습’은 모델을 새롭게 학습할지 여부를 설정하는 기능이다. ‘예’로 설정할 경우 기존 모델의 파라미터는 유지하되, 새 데이터를 기반으로 모델을 다시 학습한다. ‘아니오’로 설정할 경우 추가적인 학습 없이 기존에 저장된 모델을 불러와 예측만 수행한다.

모든 값들을 작성한 후 ‘Save’ 버튼을 누르면 스케줄에 바로 등록된다.

5.2.2 스케줄 실행 및 확인



스케줄이 추가하면 시작 시간에 맞춰서 스케줄이 등록되고 설정한 실행주기마다 모델이 실행된다. 해당 스케줄을 클릭 시 1) 로그가 나오며 해당 로그를 클릭하면 2) 결과에 대한 시각화를 확인할 수 있다. 3) 또한 결과를 JSON 형식으로 받아볼 수 있는 URL 을 제공한다.

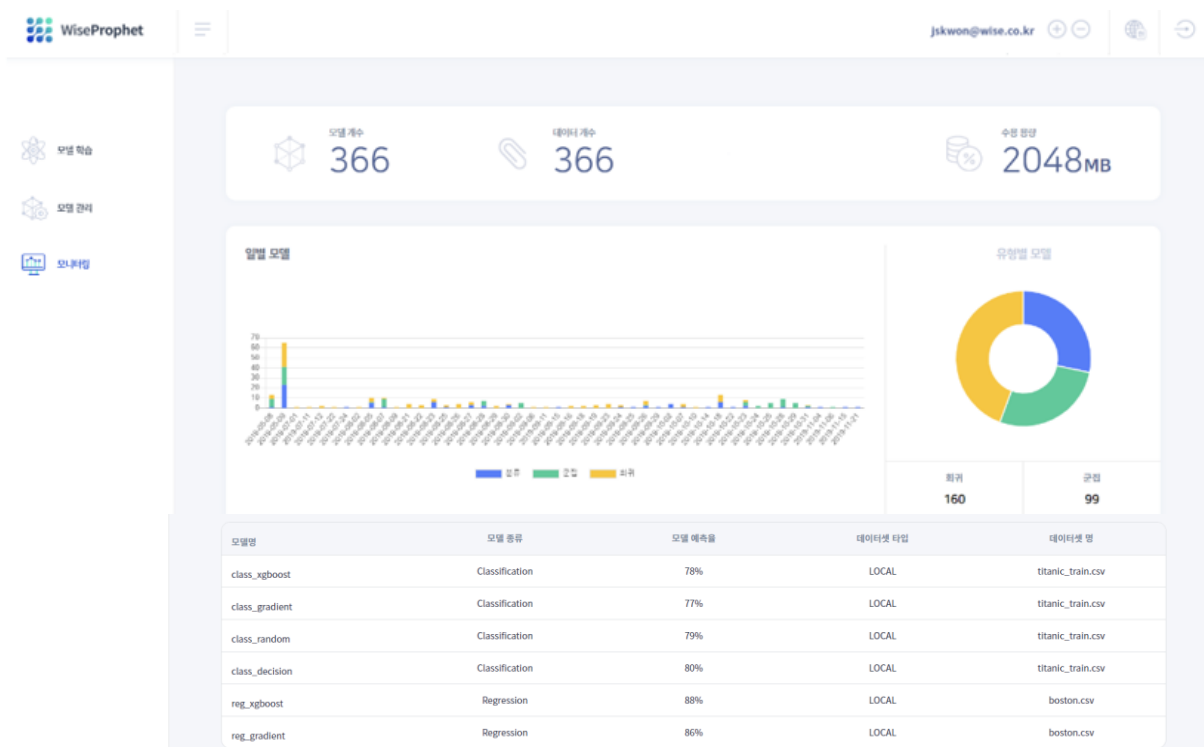
제 6 장 모니터링

본 장에서는 모델 모니터링을 위한 WiseProphet 사용법을 기술한다.

6.1. 화면 구성

모니터링 화면은 모델 정보, 데이터, 수용용량 등의 사용자 정보와 사용자가 생성한 모델에 대한 모니터링 화면으로 구성된다. 모니터링 화면에서는 일별 및 유형별 모델을 확인할 수 있다.

6.2. 사용자 및 모델 정보



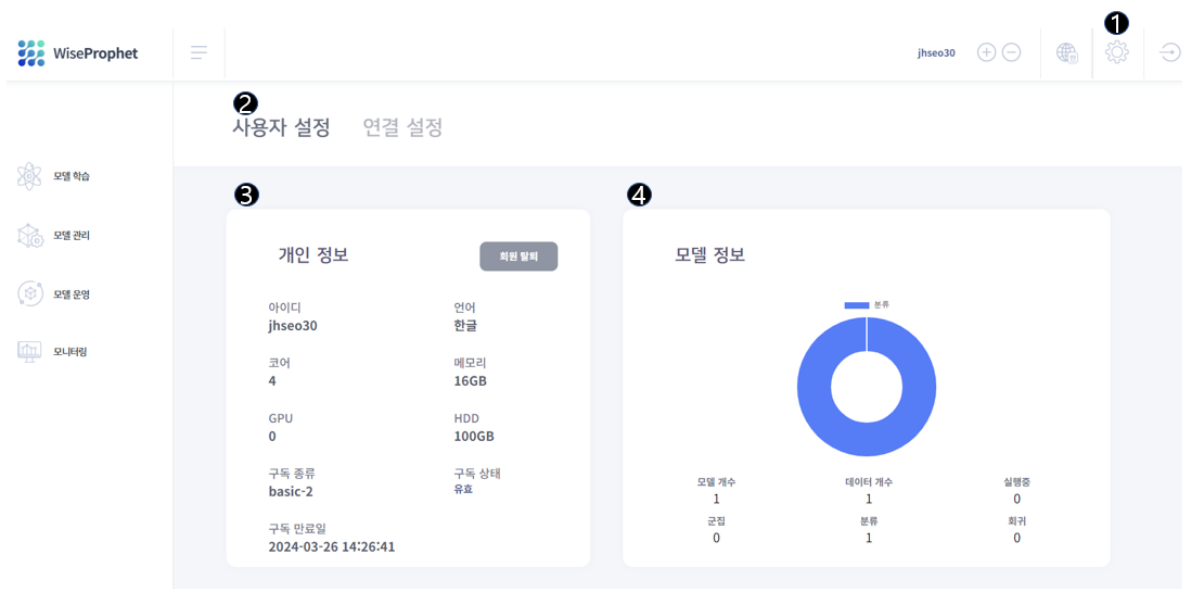
- 모니터링 화면이다.
- 1) 상단의 모델 개수는 사용자가 현재까지 생성하여 저장한 모델의 총 개수이며, 데이터 개수는 모델 생성시 사용한 데이터의 총 개수이다. 수용용량은 데이터 업로드 허용용량을 의미한다.

- 2) 유형별 모델 그래프는 회귀, 군집 및 분류모델의 비율을 시각적으로 보여주며, 일별 모델 그래프는 모델을 저장한 일별, 유형별 모델 개수를 나타낸다.
- 3) 하단에는 사용자가 저장한 모델 리스트를 보여주며, 모델의 종류, 예측률, 데이터셋 타입 그리고 데이터셋 명 정보를 확인할 수 있다.

제 7 장 설정

본 장에서는 사용자 설정을 위한 WiseProphet 사용법을 기술한다.

7.1. 사용자 설정



- 1) 해당 버튼을 클릭하면 설정화면으로 이동한다.
- 2) 사용자 설정에 대한 화면을 보여준다.
- 3) 사용자의 CPU, 메모리 등 서버 정보를 보여주며 회원 탈퇴가 가능하다.
- 4) 사용자가 생성한 모델 현황을 보여준다.

7.2. 연결 설정

WiseProphet | jhseo20

사용자 설정 | 연결 설정

+ 연결 추가

연결명	DB명	연결 종류	아이피	사용자	포트	DB 종류	소유자	설명	등록일	
wise_demo	wise_demo	db	15.164.216.49	root	3306	mysql		데모	2024-02-23 17:01:54	편집

- 1) 연결 설정에 대한 화면을 보여준다.
- 2) DB 접속 정보를 추가하여 모델 학습 등에서 사용할 수 있다.
- 3) 추가한 DB 접속 리스트를 보여준다.
- 4) DB 접속정보를 수정하거나 삭제할 수 있다.

제 8 장 FAQ

Q. 파일 업로드시 에러가 발생하는 경우

A. 파일 업로드시 에러가 발생하는 경우는 주로 지원하지 않는 파일 형식이거나 인코딩 문제입니다. 업로드 파일 형식은 CSV 형식만 가능하며 인코딩의 경우 UTF-8 으로 설정해줘야 합니다.

Q. 데이터 탐색 단계에서 ‘데이터는 최소 2 개 이상의 변수를 사용해야 합니다.’ 라는 에러가 발생하는 경우

A. 최소 2 개 이상의 변수가 사용되어야 각 변수 별 영향도를 계산하여 분석 모델을 만들 수 있습니다.

Q. 모델 관리에서 신규 데이터 입력 후 모델 실행 시 에러가 발생하는 경우

A. 모델 생성 시 사용한 데이터의 컬럼과 일치하지 않는 경우 에러가 발생합니다.

Q. ‘호출 URL 이 잘못 되었거나 서버가 종료되었습니다.’ 라는 에러 메시지가 뜨는 경우

A. 페이지를 새로 고침하고, 그 이후에도 에러가 발생하는 경우에는 서버를 다시 실행해야 합니다.

Q. ‘유일 값이 100 개 이상인 범주형 변수는 사용할 수 없습니다.’ 라는 메시지가 뜨는 경우

A. 범주형 변수의 경우, 원-핫 인코딩으로 더미 변수를 생성하여 유일 값의 개수만큼 컬럼이 증가하는데 유일 값이 너무 많을 경우 컬럼수가 과도하게 많아지고, 모델 생성에 사용할 특징 역시 증가함으로써 모델 성능이 저하되어 제외시킵니다.