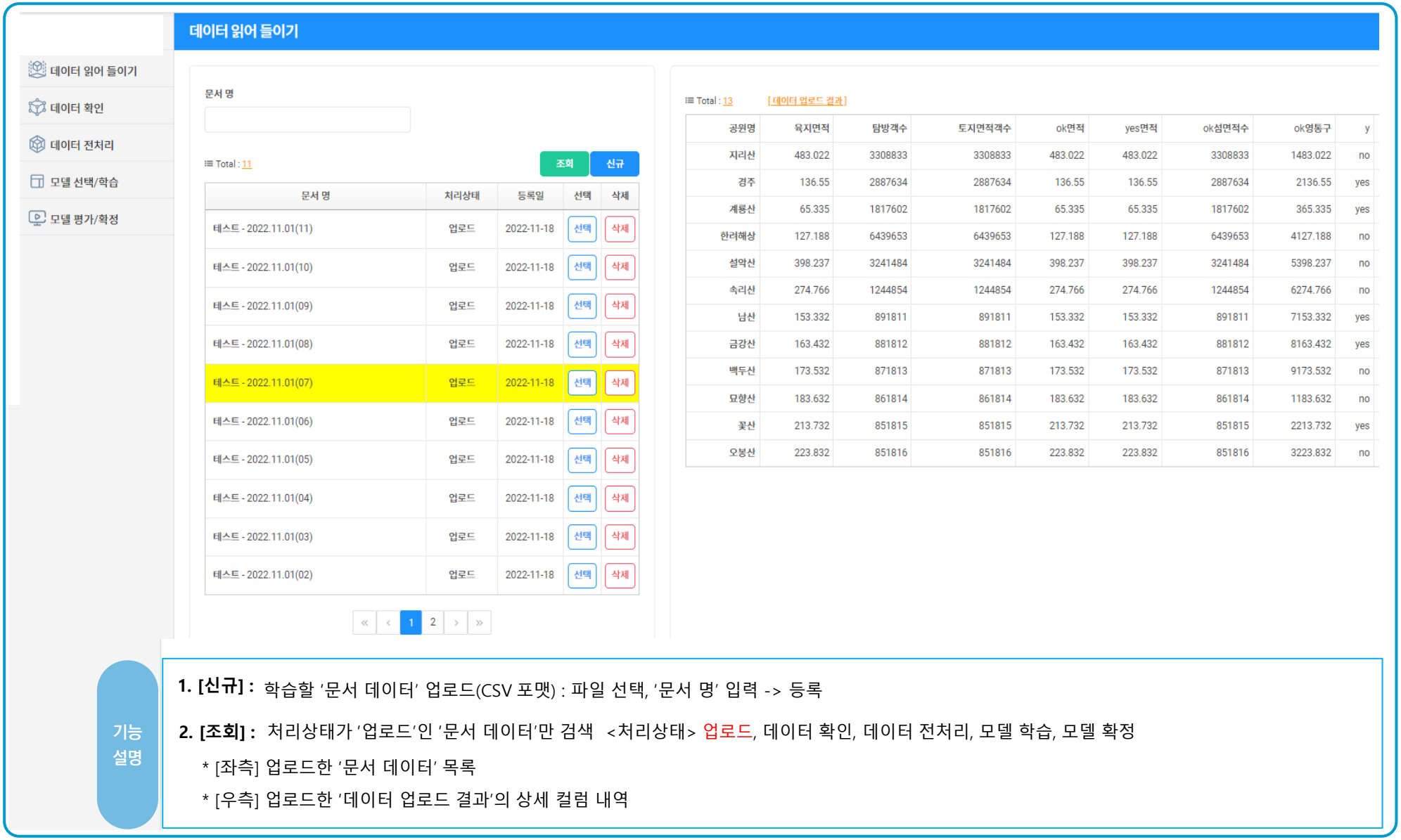
올바로시스템 노후 장비 교체 및 클라우드 인프라 증설

화면 설계서

[머신러닝 소프트웨어]

문서번호	B2022_01
작성일자	2022.09.

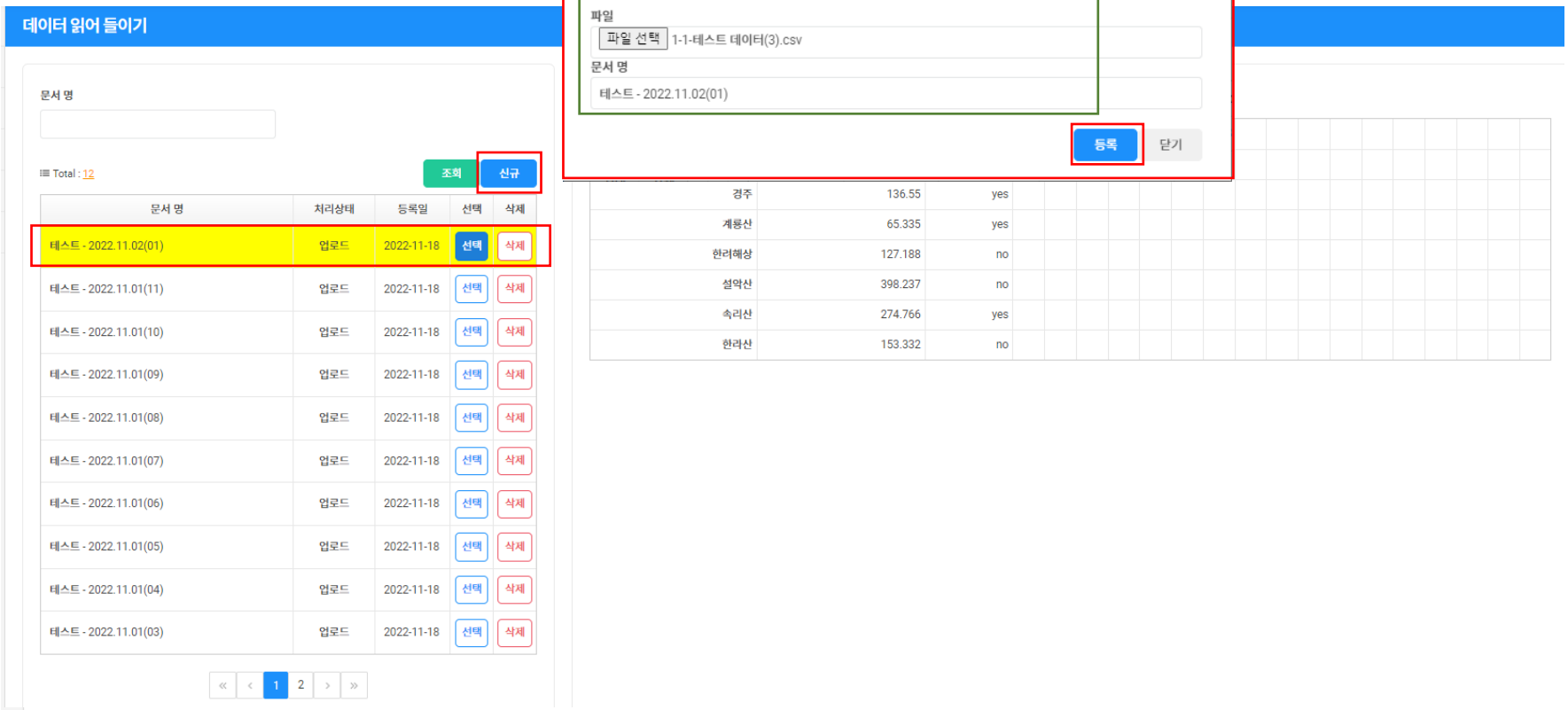
1. 데이터 읽어 들이기 (1/2)



기능 설명

1. **[신규] :** 학습할 '문서 데이터' 업로드(CSV 포맷) : 파일 선택, '문서 명' 입력 -> 등록
2. **[조회] :** 처리상태가 '업로드'인 '문서 데이터'만 검색 <처리상태> 업로드, 데이터 확인, 데이터 전처리, 모델 학습, 모델 확정
   * [좌측] 업로드한 '문서 데이터' 목록
   * [우측] 업로드한 '데이터 업로드 결과'의 상세 컬럼 내역

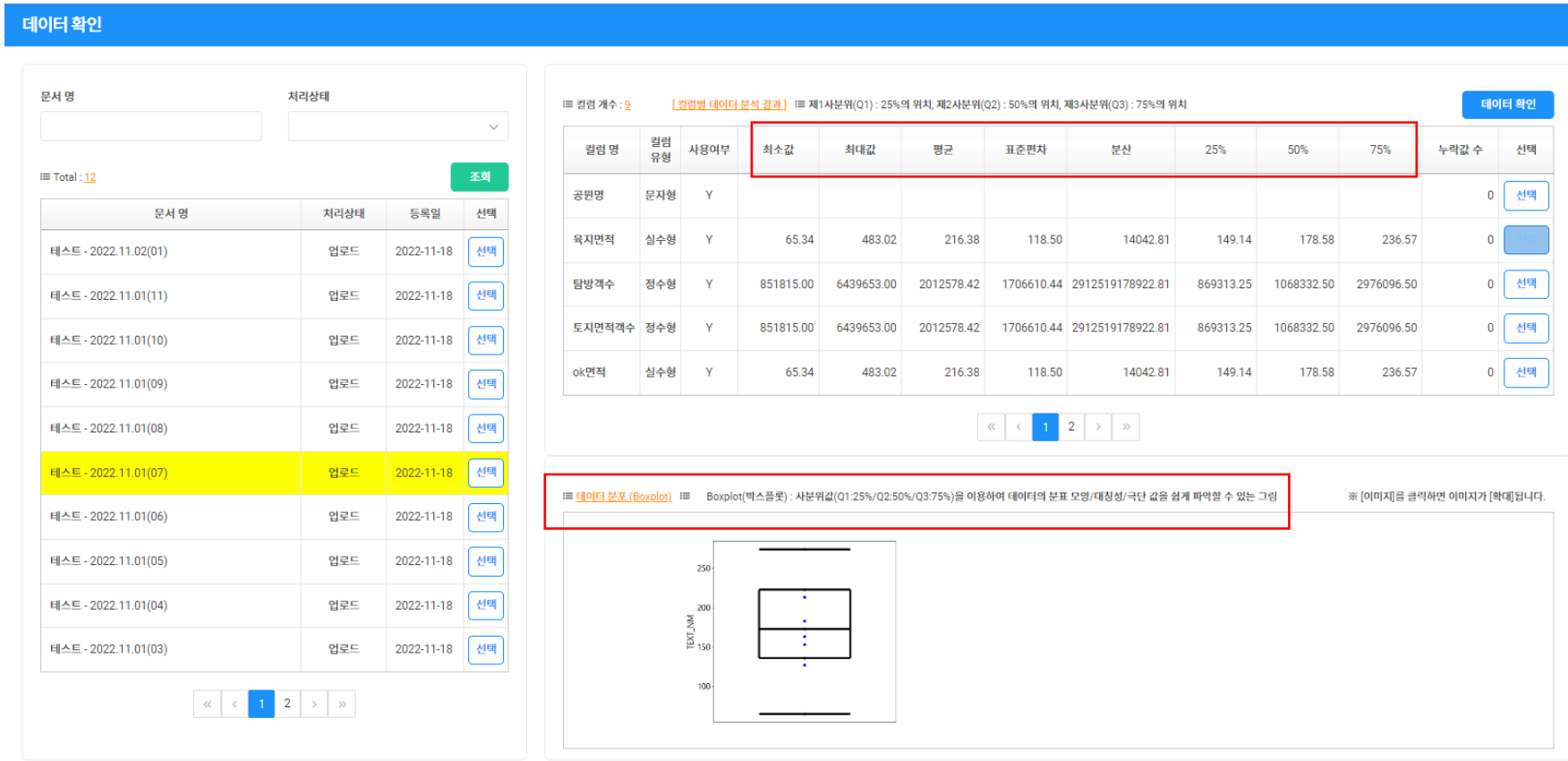
1. 데이터 읽어 들이기 (2/2)



기능 설명

1. **[신규] :** 학습할 '문서 데이터' 업로드(CSV 포맷) : 파일 선택, '문서 명' 입력 -> 등록
2. **[조회] :** 처리상태가 '업로드'인 '문서 데이터'만 검색 <처리상태> 업로드, 데이터 확인, 데이터 전처리, 모델 학습, 모델 확정
   * [좌측] 업로드한 '문서 데이터' 목록
   * [우측] 업로드한 '데이터 업로드 결과'의 상세 컬럼 내역

2. 데이터 확인



기능 설명

1. [조회] : <처리상태> 업로드, 데이터 확인

2. 컬럼별 데이터 분석 결과

(1) 컬럼명, 컬럼유형, 사용여부, 표준편차, 분산

(2) 사분위 : 데이터를 4등분한 지점 (참고) IQR(4분위수 범위)=Q3–Q1-> 데이터를 쌓아올렸을 때 25% 지점(1분위수)에 있는 데이터와 75% 지점(3분위수)에 있는 데이터의 차이

* 최소값 : 제 1사분위에서 1.5 IQR1을 뺀 위치
* 최대값 : 제 3사분위에서 1.5 IQR을 더한 위치
* 제 1사분위(Q1) : 25%의 위치, 즉 전체 데이터 중 하위 25%에 해당하는 값
* 제 2사분위(Q2) : 50%의 위치로 중앙값(median)
* 제 3사분위(Q3) : 75%의 위치, 전체 데이터 중 상위 25%에 해당하는 값

3. 데이터 분포(Boxplot) : 사분위값(Q1:25%/Q2:50%/Q3:75%)을 이용하여 데이터의 분포 모양/대칭성/극단 값을 쉽게 파악할 수 있는 그림

3. 데이터 전처리 (1/5)

데이터 전처리

문서 명 / 처리상태

Total : 34 [조회]

문서 명	처리상태	등록일	선택
테스트 - 2022.11.02(01)	업로드	2022-11-18	선택
테스트 - 2022.11.01(11)	업로드	2022-11-18	선택
테스트 - 2022.11.01(10)	데이터 확인	2022-11-18	선택
테스트 - 2022.11.01(09)	업로드	2022-11-18	선택
테스트 - 2022.11.01(08)	업로드	2022-11-18	선택
테스트 - 2022.11.01(07)	데이터 확인	2022-11-18	선택
테스트 - 2022.11.01(06)	업로드	2022-11-18	선택
테스트 - 2022.11.01(05)	업로드	2022-11-18	선택
테스트 - 2022.11.01(04)	업로드	2022-11-18	선택
테스트 - 2022.11.01(03)	업로드	2022-11-18	선택

« ‹ 1 2 3 4 › »

칼럼 개수 : 14 [목표변수 적용] ※[목표변수는 반드시 1 칼럼만 선택하셔야 모델학습이 가능합니다] ※목표변수(결과변수) : 추정하거나 예측하고 싶은 목적 데이터 (예)등급/가격/성별/학력

칼럼 명	칼럼 유형	최소값	최대값	평균	표준편차	분산	25%	50%	75%	누락값 수	목표변수 선택
공원명	문자형									0	☐
육지면적	실수형	65.34	483.02	216.38	118.50	14042.81	149.14	178.58	236.57	0	☐
탐방객수	정수형	851815.00	6439653.00	2012578.42	1706610.44	2912519178922.81	869313.25	1068332.50	2976096.50	0	☐
토지면적객수	정수형	851815.00	6439653.00	2012578.42	1706610.44	2912519178922.81	869313.25	1068332.50	2976096.50	0	☐
ok면적	실수형	65.34	483.02	216.38	118.50	14042.81	149.14	178.58	236.57	0	☐
yes면적	실수형	65.34	483.02	216.38	118.50	14042.81	149.14	178.58	236.57	0	☐
ok섬면적수	정수형	851815.00	6439653.00	2012578.42	1706610.44	2912519178922.81	869313.25	1068332.50	2976096.50	0	☐
ok영통구	실수형	365.34	9173.53	4241.38	2942.37	8657568.40	1973.17	3675.51	6494.41	0	☐
토지면적객수.1	정수형	51815.00	887634.00	279245.08	294964.43	87004015286.45	69313.25	166647.50	341538.00	0	☐
ok장안구	실수형	765.34	9483.02	4191.38	2693.42	7254532.68	2178.58	3718.78	5580.47	0	☐

« ‹ 1 2 › »

데이터 정제 (Data Cleansing) [데이터 전처리 적용]

[데이터 전처리]구분	☐제거 여부 선택	☐사용 여부 선택
중복값 처리	☐	☑
결측값 처리	☐	☑
이상값 처리	☐	☑

기능 설명

1. **[조회] :** <처리상태> 업로드, 데이터 확인, 데이터 전처리, 모델 학습
2. **[목표변수 적용]**
   (1) 목표변수로 사용하고자 하는 경우 : check box 선택 처리
   (2) [목표변수 적용] 버튼 클릭 -> 1)적용한 '목표변수' 데이터를 Update 처리 2) 처리상태 : '데이터 전처리'로 변경 처리

3. 데이터 전처리 (2/5)

데이터 전처리

문서 명 / 처리상태

Total : 33 [조회]

문서 명	처리상태	등록일	선택
테스트 - 2022.11.21(01)	업로드	2022-11-21	선택
테스트 - 2022.11.02(01)	업로드	2022-11-18	선택
테스트 - 2022.11.01(11)	업로드	2022-11-18	선택
테스트 - 2022.11.01(10)	데이터 확인	2022-11-18	선택
테스트 - 2022.11.01(09)	업로드	2022-11-18	선택
테스트 - 2022.11.01(08)	업로드	2022-11-18	선택
테스트 - 2022.11.01(07)	데이터 확인	2022-11-18	선택
테스트 - 2022.11.01(06)	업로드	2022-11-18	선택
테스트 - 2022.11.01(05)	업로드	2022-11-18	선택
테스트 - 2022.11.01(04)	업로드	2022-11-18	선택

« ‹ 1 2 3 4 › »

칼럼 개수 : 14 [목표변수 적용] ※ [목표변수는 반드시 1 칼럼만 선택하셔야 모델학습이 가능합니다] ※목표변수(결과변수) : 추정하거나 예측하고 싶은 목적 데이터 (예)등급/가격/성별/학력

칼럼 명	칼럼 유형	최소값	최대값	평균	표준편차	분산	25%	50%	75%	누락값 수	목표변수 선택
공원명	문자형									0	
육지면적	실수형	65.34	483.02	216.38	118.50	14042.81	149.14	178.58	236.57	0	
탐방객수	정수형	851815.00	6439653.00	2012578.42	1706610.44	2912519178922.81	869313.25	1068332.50	2976096.50	0	
토지면적객수	정수형	851815.00	6439653.00	2012578.42	1706610.44	2912519178922.81	869313.25	1068332.50	2976096.50	0	
ok면적	실수형	65.34	483.02	216.38	118.50	14042.81	149.14	178.58	236.57	0	
yes면적	실수형	65.34	483.02	216.38	118.50	14042.81	149.14	178.58	236.57	0	
ok섬면적수	정수형	851815.00	6439653.00	2012578.42	1706610.44	2912519178922.81	869313.25	1068332.50	2976096.50	0	
ok영통구	실수형	365.34	9173.53	4241.38	2942.37	8657568.40	1973.17	3675.51	6494.41	0	
토지면적객수.1	정수형	51815.00	887634.00	279245.08	294964.43	87004015286.45	69313.25	166647.50	341538.00	0	
ok장안구	실수형	765.34	9483.02	4191.38	2693.42	7254532.68	2178.58	3718.78	5580.47	0	

« ‹ 1 2 › »

데이터 정제 (Data Cleansing) [데이터 전처리 적용]

※ [결측값, Missing Value] 알려지지 않고, 수집되지 않거나 잘못 입력된 데이터 세트의 값

※ [이상값, 극단값, Outiler] 특정 데이터 변수의 분포에서 비정상적으로 벗어난 값

[데이터 전처리]구분	제거 여부 선택	사용 여부 선택
중복값 처리	☐	☑
결측값 처리	☐	☑
이상값 처리	☐	☑

기능 설명

3. [데이터 전처리 적용]

(1) 중복값 처리 : [처리방법] 중복되는 값 중 첫번째 값만 제외하고 나머지를 제거

(2) 결측값 처리 : Missing Value - 알려지지 않고, 수집되지 않거나 잘못 입력된 데이터 세트의 값 [처리방법] 평균(mean)값

(3) 이상값 처리 : Outiler - 이상값 = 극단값 = 이상점 : 특정 데이터 변수의 분포에서 비정상적으로 벗어난 값 [처리방법] 평균(mean)값

3. 데이터 전처리 (3/5)

기능 설명

[Data Cleaning (데이터 정제)]

(1) 중복값(Duplicated Value)

1) 중복 데이터 확인 – duplicated()

* 전에 나온 행들과 비교하여 중복되는 행이면 True를 반환하고, 처음 나오는 행이면 False를 반환함

* [형식] DataFrame.duplicated(keep = 'first'|'last'|False, subset = ...)

2) 중복 데이터 제거 – drop_duplicates()

* 중복되는 행을 제거하고, 고유한 관측값(observed value)을 가진 행들만 남김

* [방법] 중복된 것들 중 하나만 남기고 제거하기 (예) df.drop_duplicates(['title'],keep='last') – 마지막에 있는 데이터만 남김.

(2) 결측값(Missing Value)

1) 설명

* Missing Feature = NA(Not Available), 값이 표기되지 않은 값

* 결측값은 '데이터가 없다'는 의미이며, 결측값 처리는 데이터 종류와 관계없이 필수적으로 해야 하는 데이터 전처리 과정임.

2) 데이터 결측값을 채우는 방법 (Data Imputation)

① 평균값으로 대체 (Mean Imputation)

② 새로운 값으로 대체(Substitution)

③ Hot deck imputation : 다른 변수에서 비슷한 값을 갖는 데이터 중에서 하나를 random sampling하여 그 값을 복사해 오는 방법

④ Cold deck imputation : 다른 변수에서 비슷한 값을 갖는 데이터 중에서 하나를 골라 그 값으로 결측값을 대체하는 방식
- 하나를 random sampling하는 것이 아니라 어떠한 규칙(예, k번째 sample의 값을 취해온다는 등)에서 하나를 선정하는 방법

⑤ Regression imputation

* 결측값이 존재하지 않는 변수를 feature로 삼고, 결측값을 채우고자 하는 변수를 target으로 삼아 regression task를 진행하는 방법

* (문제점) 데이터 내의 다른 변수를 기반으로 결측값을 예측하는 것이기 때문에 변수 간 관계를 그대로 보존할 수 있지만 동시에 예측값 간 variability는 보존하지 못함.

⑥ Stochastic regression imputation

* Regression imputation 방법에 random residual value를 더해서 결측치의 최종 예측값으로 대체하는 방법

* regression 방법의 이점을 모두 갖는데다 random component를 갖는 데에서 따르는 이점이 있음.

3. 데이터 전처리 (4/5)

기능 설명

3) 결측값 처리 방법

① 결측값 확인

* 결측값 여부 확인 (형식) df["col"].isnull()
* 결측값 개수 확인 (형식) df["col"].isnull().value_counts()

② 결측값 제거(Deletion)

* 목록 삭제(Listwise)
  - 결측값이 있는 행은 전부 삭제 (형식) df = df.dropna(axis = 0) - default 값은 0
  - 결측값이 있는 열은 전부 삭제 (형식) df = df.dropna(axis = 1)
* 단일값 삭제(Pairwise)
  - 행 전체가 결측값이 행만 삭제 (형식) df = df.dropna(how = 'all')
  - 행의 결측값이 2개 초과인 행만 삭제 (형식) df = df.dropna(thresh = 2)
  - 특정 열들중에 결측값이 있을 경우에 해당 행을 삭제 (형식) df = df.dropna(subset=['col1', 'col2', 'col3'])

③ 결측값 대치(Imputation) – 결측값을 특정값으로 대치하는 방법

* 대치 값 종류
  - 최빈값(mode) : 범주형에서 결측값이 발생시, 범주별 빈도가 가장 높은 값으로 대치하는 방법
  - 중앙값(median) : 숫자형(연속형)에서 결측값을 제외한 중앙값으로 대치하는 방법
  - 평균(mean) : 숫자형(연속형)에서 결측값을 제외한 평균으로 대치하는 방법
  - Similar case Imputation : 조건부 대치
  - Generalized Imputation : 회귀분석을 이용한 대치

[참고] 결측값(Missing Value) 처리 기준 : 결측값이 전체 데이터에 얼만큼 차지하는 지 비율에 따라 처리 방법을 달리함.

* 10% 미만 : 해당 row 값을 삭제하거나 다른 값으로 치환
  - 이 때 치환은 전체 데이터의 평균값이나 최빈값을 활용할 수도 있고, 외부 다른 데이터나 다른 컬럼의 값을 기반으로 유추해서 사용할 수도 있음.
* 10 ~ 50% : 값을 채워줄 수 있는 모델(결측값을 채우기 위한 수식 작성)을 만들어 처리하는 것이 좋음.
  - (예) 키 / 성별 / 체중 / 나이 -> 체중의 몇몇 값이 비어있는 경우
    <방법> 키/성별/나이 값을 적절히 조합하여 대략적인 체중의 값을 유추 -> 수식을 머신러닝을 통해 만들어서 사용
* 50% 이상 : 해당 컬럼을 삭제하는 것이 좋음.

3. 데이터 전처리 (5/5)

기능 설명

(3) 이상값(Outlier)

1) 설명

* 이상값 = 극단값 = 이상점 : 특정 데이터 변수의 분포에서 비정상적으로 벗어난 값
* 데이터 집합에서 대부분의 다른 sample들과 현저한 차이를 보이는 sample 또는 변수값

2) 데이터에서 이상값을 탐지하는 방법

① Standard Deviation : 데이터의 분포가 정규분포를 이룰 때, 데이터의 표준편차를 이용해 이상치를 탐지하는 방법

② IQR(Interquartile Range) with Box plots

* 데이터의 분포가 정규 분포를 이루지 않거나 한 쪽으로 편향(skewed)된 경우, 데이터의 IQR 값을 이용해 이상치를 탐지하는 방법

③ Isolation Forest

* 결정 트리 계열의 비지도 학습 알고리즘으로 High dimensional 데이터 셋에서 이상치를 탐지할 때 효과적인 방법
* Data Set을 결정 트리 형태로 표현해 정상 데이터를 분리하기 위해서는 Tree의 깊이가 깊어지고 반대로 이상치는 Tree의 상단에서 분리할 수 있다는 개념을 이용함 -> 즉, 데이터에서 이상치를 분리하는 것이 더 쉽다는 것임.

④ DBScan (Density Based Spatial Clustering of Applications with Noise)

* DBSCAN(밀도 기반의 클러스터링 알고리즘)으로 어떠한 클러스터에도 포함되지 않는 데이터를 이상값으로 탐지하는 방법

3) 이상값 처리 방법

① 이상값 제거 (Delete) : 극단적으로 크거나 작은 값을 제거함으로써 분석 값을 조금 더 보정하는 방법

* 상/하위 이상값를 확인 한 결과, 상위 극단치만 있는 것으로 확인된 경우 -> 이상치를 갖고 있는 행만 제거

② 이상값 대체(Replacement) : 이상값을 대체함으로써 Data Set을 보정하는 방법

* 하한값/상한값 결정 후, 하한값보다 작으면 하한값으로, 상한값보다 크면 상한값으로 대체
* 중위수로부터 n편차 큰 값으로 대체
* 평균의 표준편차 * n범위를 초과하는 값일 경우, 평균 +- (표준편차 * n) 값을 하한/상한값으로 지정

③ 데이터 셋 축소/과장(Scaling)

④ 데이터 셋 최소최대 척도(MinMax Scaling) : 최대값을 1, 최소값을 0으로 변환한 뒤, 각 구간값을 0~1사이 스케일로 적용하는 방법

⑤ 데이터 셋 정규화(Normalize)

4. 모델 선택/학습 (1/6)

모델 선택/학습

문서 명 / 처리상태

Total : 22 [조회]

문서 명	처리상태	등록일	선택
테스트 - 2022.11.11(03)	모델 학습	2022-11-11	선택
테스트 - 2022.11.11(02)	모델 학습	2022-11-11	선택
테스트 - 칼럼 6개(2)	모델 학습	2022-11-11	선택
테스트 - 칼럼 12개-	모델 학습	2022-11-11	선택
테스트 - 칼럼 11개	모델 학습	2022-11-11	선택
테스트 - 칼럼 10개	모델 학습	2022-11-11	선택
테스트 - 칼럼 9개	모델 학습	2022-11-11	선택
테스트 - 칼럼 8개	모델 학습	2022-11-11	선택
테스트 - 칼럼 7개	모델 학습	2022-11-11	선택
테스트 - 칼럼 6개	모델 학습	2022-11-11	선택

« ‹ 1 2 3 › »

분류(Classification) 모델 [그래디언트 부스팅(Gradient Boost)] [의사 결정 트리] [Random Forest] [SVM(Support Vector Machine)] [신경망(Neural Network)]

군집(Clustering) 모델 [K-평균 군집화(K-means Clustering)]

※ 아래 화면의 [이미지]를 클릭하면 이미지가 [확대]됩니다.

Total : 0 모델 학습 목록

※정밀도(Precision) : 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율

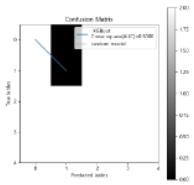
※ 재현율(Recall) : 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율

※ F1-score : 정밀도와 재현율의 조화평균 -> 2*(정밀도*민감도)/(정밀도+민감도)

학습 구분	정밀도	재현율	F1-score	학습일자	학습횟수	선택
Gradient Boost	0.5	0.5	0.5	2022-11-17	1	선택
의사결정트리	0.4	0.4	0.4	2022-11-17	1	선택
Random Forest	0.4	0.4	0.4	2022-11-17	1	선택
SVM	0.4	0.4	0.4	2022-11-17	1	선택
신경망	0.25	0.25	0.25	2022-11-17	1	선택

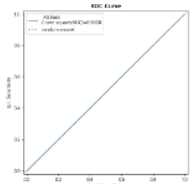
Confusion Matrix (혼동행렬) / (신경망) 훈련 데이터 대 검증 데이터 손실 그래프

※ 학습을 통한 예측성능을 측정하기 위해 예측결과 값과 실제 값을 비교하기 위한 표



ROC Curve ※ [이미지]를 클릭하면 이미지가 [확대]됩니다.

※ (수신자 판단 곡선) 모델의 효율성을 민감도와 특이도를 이용하여 그래프로 나타낸 것.



기능 설명

1. 모델 (알고리즘) 선택

(1) 분류(Classification) : 그래디언트 부스팅(Gradient Boost), 의사 결정 트리(Decision Tree), Random Forest, SVM(Support Vector Machine), 신경망(Neural Network)

(2) 군집(Clustering) : K-평균 군집화(K-means Clustering)

4. 모델 선택/학습 (2/6)

모델 선택/학습

문서 명 / 처리상태

Total : 22 [조회]

문서 명	처리상태	등록일	선택
테스트 - 2022.11.11(03)	모델 학습	2022-11-11	선택
테스트 - 2022.11.11(02)	모델 학습	2022-11-11	선택
테스트 - 칼럼 6개(2)	모델 학습	2022-11-11	선택
테스트 - 칼럼 12개-	모델 학습	2022-11-11	선택
테스트 - 칼럼 11개	모델 학습	2022-11-11	선택
테스트 - 칼럼 10개	모델 학습	2022-11-11	선택
테스트 - 칼럼 9개	모델 학습	2022-11-11	선택
테스트 - 칼럼 8개	모델 학습	2022-11-11	선택
테스트 - 칼럼 7개	모델 학습	2022-11-11	선택
테스트 - 칼럼 6개	모델 학습	2022-11-11	선택

« ‹ 1 2 3 › »

분류(Classification) 모델: 그레디언트 부스팅(Gradient Boost) | 의사 결정 트리 | Random Forest | SVM(Support Vector Machine) | 신경망(Neural Network)

군집(Clustering) 모델: K-평균 군집화(K-means Clustering)

※ 아래 화면의 [이미지]를 클릭하면 이미지가 [확대]됩니다.

Total : 0 모델 학습 목록

※정밀도(Precision) : 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율

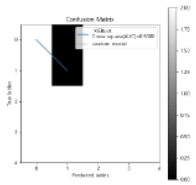
※ 재현율(Recall) : 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율

※ F1-score : 정밀도와 재현율의 조화평균 -> 2*(정밀도*민감도)/(정밀도+민감도)

학습 구분	정밀도	재현율	F1-score	학습일자	학습횟수	선택
Gradient Boost	0.5	0.5	0.5	2022-11-17	1	선택
의사결정트리	0.4	0.4	0.4	2022-11-17	1	선택
Random Forest	0.4	0.4	0.4	2022-11-17	1	선택
SVM	0.4	0.4	0.4	2022-11-17	1	선택
신경망	0.25	0.25	0.25	2022-11-17	1	선택

Confusion Matrix (혼동행렬) / (신경망) 훈련 데이터 대 검증 데이터 손실 그래프

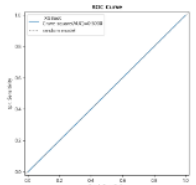
※ 학습을 통한 예측성능을 측정하기 위해 예측결과 값과 실제 값을 비교하기 위한 표



ROC Curve

※ [이미지]를 클릭하면 이미지가 [확대]됩니다.

※ (수신자 판단 곡선) 모델의 효율성을 민감도와 특이도를 이용하여 그래프로 나타낸 것.



기능 설명

2. 모델 학습 목록

(1) 학습하고자 하는 '문서'를 선택하고, 우측 상단의 분류 또는 군집 모델을 클릭한다.

(2) 알림창 – '[분류 모델] / [군집 모델]을 학습하시겠습니까?' -> [예] 선택 -> 알림창 – '[분류 모델] 학습이 정상적으로 처리 되었습니다.' 메시지가 나타남.

(3) [모델 학습 목록] – 학습된 모델의 학습결과가 나타남 -> [선택] 버튼 클릭 : 1) Confusion Matrix(혼동행렬) / (신경망) 훈련 데이터 대 검증 데이터 손실 그래프

2) ROC Curve - Receiver Operating Characteristic 곡선

4. 모델 선택/학습 (3/3)

기능 설명

[학습 모델]

(1) 그래디언트 부스팅(Gradient Boost)

* 분류의 여러 개의 결정 트리를 사용하는 기법인 '앙상블'과 '배깅', '부스팅'을 결합한 알고리즘
* 의사결정 트리의 단점인 오버피팅(overfitting, 과적합) 문제를 개선할 수 있음.

[참고] * Bagging : 서로 무관한 약분류기를 병렬로 만들고, 그 분류 결과를 다수결로 최종 결과를 결정하는 기법 (예) Random Forest
* Boosting : 여러 개의 분류기를 사용한다는 점은 배깅과 같지만 배깅이 여러 개의 분류기를 서로 무관하게 학습시키는 데 반해, Boosting은 바로 전 분류기의 결과를 기초로 다음 분류기를 학습한다는 점이 다름.

(2) 의사결정 트리 (Decision Tree)

* 어떤 항목에 대한 관측 값과 목표 값을 연결시켜주는 예측 모델로써 결정 트리를 사용하는 학습 방법

(3) Random Forest

* 다수의 의사 결정 Tree를 만들고 구 Tree들의 분류를 집계하여 최종 분류하는 학습 방법
* 의사결정 트리의 단점인 오버피팅(overfitting, 과적합) 문제를 개선할 수 있음.

(4) SVM (Support Vector Machine)

* 분류의 여래 개의 결정 트리를 사용하는 기법인 '앙상블'과 '배깅', '부스팅'을 결합한 알고리즘
* 의사결정 트리의 단점인 오버피팅(overfitting, 과적합) 문제를 개선할 수 있음.

(5) k-평균 군집화(K-means Clustering)

* 주어진 데이터를 k개의 클러스터로 묶는 비지도 학습 방법
* 데이터의 특징만으로 비슷한 데이터끼리 모아 군집화된 클래스로 분류

4. 모델 선택/학습 (3/3)

기능 설명

[학습 모델]

(1) 그래디언트 부스팅(Gradient Boost)

* 분류의 여래 개의 결정 트리를 사용하는 기법인 '앙상블'과 '배깅', '부스팅'을 결합한 알고리즘
* 의사결정 트리의 단점인 오버피팅(overfitting, 과적합) 문제를 개선할 수 있음.

[참고] * Bagging : 서로 무관한 약분류기를 병렬로 만들고, 그 분류 결과를 다수결로 최종 결과를 결정하는 기법 (예) Random Forest
* Boosting : 여러 개의 분류기를 사용한다는 점은 배깅과 같지만 배깅이 여러 개의 분류기를 서로 무관하게 학습시키는 데 반해, Boosting은 바로 전 분류기의 결과를 기초로 다음 분류기를 학습한다는 점이 다름.

(2) 의사결정 트리 (Decision Tree)

* 어떤 항목에 대한 관측 값과 목표 값을 연결시켜주는 예측 모델로써 결정 트리를 사용하는 학습 방법

(3) Random Forest

* 다수의 의사 결정 Tree를 만들고 구 Tree들의 분류를 집계하여 최종 분류하는 학습 방법
* 의사결정 트리의 단점인 오버피팅(overfitting, 과적합) 문제를 개선할 수 있음.

(4) SVM (Support Vector Machine)

* 분류의 여래 개의 결정 트리를 사용하는 기법인 '앙상블'과 '배깅', '부스팅'을 결합한 알고리즘
* 의사결정 트리의 단점인 오버피팅(overfitting, 과적합) 문제를 개선할 수 있음.

(5) k-평균 군집화(K-means Clustering)

* 주어진 데이터를 k개의 클러스터로 묶는 비지도 학습 방법
* 데이터의 특징만으로 비슷한 데이터끼리 모아 군집화된 클래스로 분류

4. 모델 선택/학습 (3/3)

기능 설명

[Confusion Matrix (혼동행렬)]

(1) 설명

* 분류 모델이 얼마나 헷갈리고(confused) 있는지도 함께 보여주는 지표
  - 이진 분류에서 예측 오류가 얼마인지와 더불어 어떠한 유형의 예측오류가 발생하고 있는지를 함께 나타내는 지표
* 학습(Training)을 통한 Prediction 성능을 측정하기 위해 예측결과 값(Value)과 실제 값을 비교하기 위한 표

[참고] TP : 맞는 것을 올바르게 예측한 것 FN : 맞는 것을 틀렸다고 잘못 예측한 것
FP : 틀린 것을 맞다고 잘못 예측한 것 TN : 틀린 것을 올바르게 예측한 것

(2) 정밀도(Precision rate)

* 양성으로 예측된 결과의 정확한 예측의 비율을 의미하는 모델의 성능 지표
* 정밀도 = TP / (TP + FP) = 양성으로 예측한 경우 중에 실제 양성인 비율

(3) 재현율 (Recall rate)

* 실제 양성(positive)이었던 데이터 중 모델이 양성으로 판정된 건수 비율
* 재현율 = TP / (TP + FN)

(4) F1-score

1) 설명

* 정밀도(Precision Rate)와 민감도(Sensitivity Rate)의 조화평균으로 주로 분류 클래스 간의 데이터가 불균형이 심각할 때 사용함.
* 정확도의 경우, 데이터 분류 클래스가 균일하지 못하면 머신러닝 성능을 제대로 나타낼 수 없기 때문에 F1 Score를 사용함.
* 둘을 따로 볼 경우 Trade-off 관계가 발생하여 판단 어려움, 이 경우 둘을 조화 평균한 F1-Score 사용
* F1-Score는 정밀도와 재현율을 동등하게 계산한 경우이며
  정밀도에 더 가중치를 주고 싶은경우 β를 1 이상으로 입력
  재현율에 더 가중치를 주고 싶은경우 β를 0~1로 입력

2) F1-score = 2 * (정밀도 * 민감도) / (정밀도 + 민감도)

= 2 * (((TP / (TP + FP)) * (TP / (TP + FN)) / ((TP / (TP + FP)) + (TP / (TP + FN)))

4. 모델 선택/학습 (3/3)

기능 설명

[ROC (Receiver Operating Characteristic) Curve]

1) 설명

* 혼동행렬만으로는 모델의 평가 척도가 부족해서 모델의 효율성을 평가하는 척도로 사용됨.
* 모델의 효율성을 민감도와 특이도를 이용하여 그래프로 나타낸 것. 즉, 잘못된 Negative를 줄이는데 초점

2) 민감도 (Sensitivity Rate) = 재현율(Recall Rate) = TPR(True Positive Rate)

* 실제 Positive인 것 중 올바르게 Positive를 맞춘 것의 비율 - 실제 정답을 얼마나 맞췄느냐?
* 실제 양성 중 모델이 정확하게 양성으로 예측한 비율을 나타내는 모델의 성능 지표
* TP / (TP + FN) = 실제 진단결과가 양성 중에 양성을 양성이라고 맞춘 비율

3) 특이도 (Specificity Rate)

* TN / (TN + FP) = 실제 진단결과가 음성 중에 음성을 음성이라고 맞춘 비율

5. 모델 평가/확정

모델 평가/확정

문서 명 / 처리상태

Total : 22 [조회]

문서 명	처리상태	등록일	선택
테스트 - 2022.11.17 (15개)	모델 학습	2022-11-17	선택
테스트 - 2022.11.17 (14개)	모델 학습	2022-11-17	선택
테스트 - 2022.11.17 (13개)	모델 학습	2022-11-17	선택
테스트 - 2022.11.17 (12개)	모델 학습	2022-11-17	선택
테스트 - 2022.11.17 (11개)	모델 학습	2022-11-17	선택
테스트 - 2022.11.16(04)	모델 학습	2022-11-16	선택
테스트 - 2022.11.16(05)	모델 학습	2022-11-16	선택
테스트 - 2022.11.15(01)	모델 학습	2022-11-15	선택
테스트 - 2022.11.11(03)	모델 학습	2022-11-11	선택
테스트 - 2022.11.11(02)	모델 학습	2022-11-11	선택

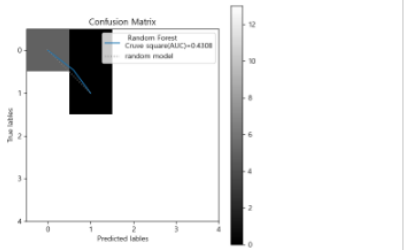
« ‹ 1 2 3 › »

Total : 5 모델 학습 목록

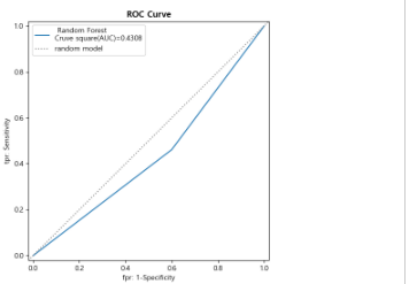
※정밀도(Precision) : 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율
※ 재현율(Recall) : 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율
※ F1-score : 정밀도와 재현율의 조화평균 -> 2*(정밀도*민감도)/(정밀도+민감도)

학습 구분	정밀도	재현율	F1-score	학습일자	학습횟수	확정일자	선택	확정
Gradient Boost	0.47	0.47	0.47	2022-11-17	3		선택	확정
의사결정트리	0.47	0.47	0.47	2022-11-17	2		선택	확정
Random Forest	0.28	0.28	0.28	2022-11-16	1		선택	확정
SVM	0.47	0.47	0.47	2022-11-17	1		선택	확정
신경망	0.55	0.55	0.55	2022-11-17	4		선택	확정

Confusion Matrix (혼동행렬) / (신경망) 훈련 데이터 대 검증 데이터 손실 그래프
※오차행렬 : 학습을 통한 예측성능을 측정하기 위해 예측결과 값과 실제 값을 비교하기 위한 표



ROC Curve ※ [이미지]를 클릭하면 이미지가 [확대]됩니다.
※ (수신자 판단 곡선) 모델의 효율성을 민감도와 특이도를 이용하여 그래프로 나타낸 것.



기능 설명

1. 모델 확정

(1) 좌측 [문서목록]에서 확정하고자 하는 '문서' 선택

(2) '모델 학습 목록'에서 확정하고자 학습 모델의 [확정] 버튼 클릭

(3) 알림창 – '[모델 확정] 처리 하시겠습니까?' -> [예] 선택 -> 알림창 – [모델 확정]이 정상적으로 처리 되었습니다.' 메시지가 나타남.