

K-최근접이웃기법

K-최근접 이웃기법 (K-nearest neighbor method)은 데이터가 주어졌을 때 기존 데이터 가운데 가장 가까운 K개 이웃의 정보를 이용하여 유사성(가장 가까운 이웃)을 기준으로 데이터를 분류하는 방법입니다. 서로 가깝게 있는 데이터를 “이웃” 이라고 합니다. 새로운 케이스가 들어 오면, 각 기존 데이터와의 거리가 계산되고 가장 많은 수의 가장 가까운 이웃이 포함된 범주에 새 케이스가 할당됩니다. 검사할 최근접 이웃(nearest neighbor)의 수 K를 지정할 수 있습니다. 또한 연속적인 대상 값을 계산하는 경우 가장 가까운 이웃의 평균 또는 중앙값 대상 값이 사용되어 새 케이스의 예측값을 가져옵니다.

메뉴 호출하기

- 고급분석 > 분류분석 > 지도 학습 > K최근접이웃기법



• 변수설정 탭

K최근접이웃기법

변수설정

분석옵션

자료분할

출력옵션

데이터

전체변수

id

bweight

lowbw

gestwks

preterm

matage

hyp

sex

① 종속변수(필수)

>

<

* 분류분석 진행 - 질적변수 선택

* 회귀분석 진행 - 양적변수 선택

설명변수

② 질적변수(선택-1개이상가능)

>

<

③ 양적변수(선택-1개이상가능)

>

<

도움말

재설정

확인

취소

메뉴 요소	설명
① 종속변수	종속변수에 해당하는 변수를 전체변수로부터 선택할 수 있습니다. 한 개의 변수가 필수적으로 선택되어야 하며 양적변수와 질적변수 모두 사용이 가능합니다. 종속변수에 결측치가 존재하는 관측치는 분석에서 제외됩니다.
② 질적변수	설명변수 중 질적변수에 해당하는 변수를 전체변수로부터 선택할 수 있습니다. 종속변수와 중복하여 선택할 수 없습니다. 질적변수와 양적변수 중 적어도 하나 이상의 변수를 선택해야 분석이 가능합니다.
③ 양적변수	설명변수 중 양적변수에 해당하는 변수를 전체변수로부터 선택할 수 있습니다. 종속변수와 중복하여 선택할 수 없습니다. 질적변수와 양적변수 중 적어도 하나 이상의 변수를 선택해야 분석이 가능합니다. 설명변수가 양적일 때는 낮은 예측능력을 보일 수 있습니다.

- 분석옵션 탭

K최근접이웃기법

변수설정 분석옵션 자료분할 출력옵션

① 분석방법

☒ 분류분석(classification) ☐ 회귀분석(regression)

K값 (K이웃)

도움말 재설정 확인 취소

메뉴 요소	설명
① 분석방법	<p>[변수설정] 탭에서 종속변수로 선택한 자료형에 따라 아래 기법 중 하나를 선택합니다.</p> <ul style="list-style-type: none"> 분류분석 (classification) (Default) : 종속변수로 질적변수를 택한 경우 선택합니다. 회귀분석 (regression) : 종속변수로 양적변수를 택한 경우 선택합니다. - K값 (K이웃) : 몇 번째로 가까운 데이터까지 살펴볼 것인지를 설정합니다. 1 이상의 정수만 입력 가능하며, Default는 5입니다.

• 자료분할 탭

K최근접이웃기법

변수설정 분석옵션 **자료분할** 출력옵션

변수목록

id
bweight
lowbw
gestwks
preterm
matage
hyp
sex

① 훈련 및 검증(필수)

● 분할검증

② ● 모든 데이터를 훈련에 이용

○ 비율에 따라 임의로 분할

훈련(train) 자료 %

시험(test) 자료 %

○ 변수로 분할

분할변수(1-훈련, 2-시험)

>

<

③ ○ 교차검증

○ Leave-one-out 교차검증

● K-fold 교차검증 K 10

④ 예측(선택)

분할변수(1-예측, 2-훈련 및 검증)

>

<

도움말 재설정 **확인** 취소

메뉴 요소	설명
① 훈련 및 검증	<p>K-최근접이웃모형 적합에 사용될 데이터를 훈련자료(training data)와 시험자료(test data)로 분할하는 방식으로 다음 2가지 옵션 중 1개를 선택할 수 있습니다.</p> <ul style="list-style-type: none"> 분할검증 (Default) : 훈련자료와 시험자료로 분할된 자료로 모형을 1회 검증하는 방법입니다. 교차검증 : 훈련자료와 시험자료를 변경해가며 여러 차례 반복 검증하는 방법입니다.
② 분할검증	<p>[분할검증]을 선택하는 경우 다음의 3가지 옵션이 활성화되어 이 중 1개를 선택할 수 있습니다.</p> <ul style="list-style-type: none"> 모든 데이터를 훈련에 이용 (Default) : 시험자료 없이 모든 개체를 모형 적합에 사용합니다. 비율에 따라 임의로 분할 : 훈련자료와 시험자료의 비율을 설정하여 임의로 분할하는 방식입니다. Default 값은 훈련자료 70%, 시험자료가 30% 입니다. 사용자는 훈련자료에 0~100을 입력할 수 있으며, 시험자료에는 100에서 입력한 값을 뺀 수치가 자동으로 입력됩니다. 임의로 분할된 개체들 중 훈련자료와 시험자료의 인덱스를 저장하려면 [출력옵션]-[저장]-[자료분할지표]를 선택합니다. 변수로 분할 : 훈련자료와 시험자료로 사용될 개체가 결정되어 있는 경우 이 옵션을 선택합니다. 이때, 훈련자료에 해당하는 개체는 1, 시험자료에 해당하는 개체는 2의 값을 갖는 인덱스 변수를 분할변수로 지정해주어야 합니다.

• 자료분할 탭

K최근접이웃기법

변수설정 분석옵션 **자료분할** 출력옵션

변수목록

id
bweight
lowbw
gestwks
preterm
matage
hyp
sex

① 훈련 및 검증(필수)

● 분할검증

② ● 모든 데이터를 훈련에 이용
○ 비율에 따라 임의로 분할
훈련(train) 자료 %
시험(test) 자료 %
○ 변수로 분할
분할변수(1-훈련, 2-시험)
> <

③ ○ 교차검증
○ Leave-one-out 교차검증
● K-fold 교차검증 K 10

④ 예측(선택)
분할변수(1-예측, 2-훈련 및 검증)
> <

도움말 재설정 확인 취소

메뉴 요소	설명
③ 교차검증	<p>[교차검증]을 선택하는 경우 다음의 2가지 옵션이 활성화되어 이 중 1개를 선택할 수 있습니다.</p> <ul style="list-style-type: none"> • Leave-one-out 교차검증 : 한 개체를 시험자료로 사용하고 나머지 개체를 모두 훈련자료로 하여 모형을 적합하는 방식으로 모든 개체에 대해 이 과정을 반복한 뒤, 전체 개체 수만큼의 모형으로부터 얻은 정확도의 평균을 모형의 최종 정확도로 계산합니다. • K-fold 교차검증 : 전체 개체를 K개의 그룹으로 임의로 분할하여, 하나의 그룹을 시험자료로 사용하고 나머지 그룹을 모두 훈련자료로 하여 모형을 적합하는 방식으로 K개의 그룹에 대해 이 과정을 반복한 뒤, 그룹 수만큼의 모형으로부터 얻은 정확도의 평균을 모형의 최종 정확도로 계산합니다. <p>- K : [교차검증]-[K-fold 교차검증]을 선택할 경우 활성화됩니다. K-fold 교차검증에 사용할 K의 값을 입력합니다. 2 이상의 정수만 입력 가능하며, 전체 개체 수보다 더 큰 정수가 입력되는 경우 자동으로 Leave-one-out 교차검증을 실시합니다. Default는 10입니다.</p>

- 자료분할 탭

K최근접이웃기법

변수설정 분석옵션 **자료분할** 출력옵션

변수목록

id
bweight
lowbw
gestwks
preterm
matage
hyp
sex

① 훈련 및 검증(필수)

② 분할검증

☒ 모든 데이터를 훈련에 이용

☐ 비율에 따라 임의로 분할

훈련(train) 자료 %

시험(test) 자료 %

☐ 변수로 분할

분할변수(1-훈련, 2-시험)

>

<

③ 교차검증

☐ Leave-one-out 교차검증

☒ K-fold 교차검증 K

④ 예측(선택)

분할변수(1-예측, 2-훈련 및 검증)

>

<

도움말 재설정 **확인** 취소

메뉴 요소	설명
④ 예측 > 분할변수	K-최근접이웃모형 적합에 사용될 훈련 및 검증 데이터와 해당 모형으로부터 예측값을 얻을 예측 데이터가 분할되어 있는 경우 사용됩니다. 훈련 및 검증에 사용되는 개체는 2, 예측에 사용되는 개체는 1의 값을 갖는 인덱스 변수를 분할변수로 지정해주어야 합니다. 예측분할변수를 지정하지 않아도 분석이 가능합니다. 예측분할변수가 지정된 경우, 예측에 해당하는 개체에 해당하는 예측값이 엑셀 시트에 "Predicted_pred_KNN"라는 변수명으로 저장됩니다.

- 출력옵션 탭

K최근접이웃기법

변수설정 분석옵션 자료분할 **출력옵션**

저장

훈련자료

① ☐ 적합값

시험자료

② ☐ 예측값

③ ☐ 자료분할지표

도움말 재설정 **확인** 취소

메뉴 요소	설명
① 적합값	적합값을 괄호 안의 변수명으로 저장합니다. (Fitted_KNN_Train)
② 예측값	[자료분할] 탭에서 '비율에 따라 임의로 분할' 또는 '변수로 분할' 을 택할 경우 예측값이 활성화됩니다. 예측값을 괄호 안의 변수명으로 저장합니다. (Predicted_test_KNN)
③ 자료분할지표	각 관측값이 훈련 혹은 시험자료 중 어떤 자료로 사용되었는지 여부를 괄호 안의 변수명으로 저장합니다. (Partition_idx_KNN)