

데이터 예측 분석

관능 평가(맛 점수)와 성분 분석(일반 분석 · 대사체)이 함께 담긴 연구실 엑셀 하나를 올리면,
변수 분포 점검 → 전처리 → 그룹 차이 검정 → 맛별 마커 성분 → 맛 예측식과 교차검증까지 한
흐름으로 계산하고, AI직원이 그 결과를 설명합니다.

앱 열기 예시 데이터 김치 32제품 × 속성 3단계 · 60변수 2026년 10월

무엇을 푸나

"이 맛은 어떤 성분으로 설명되고, 성분 값만으로 얼마나 정확히 예측되는가?"

문제 ①

엑셀이 분석 준비가 안 되어 있다

머리글이 여러 줄로 병합되어 있고, 같은 성분이 다른 단위로 두 번 적혀 있거나, 0이 '미검출'인지 실제 0인지 섞여 있습니다. 분석 전에 이것부터 사람이 일일이 찾아야 합니다.

문제 ②

분석 단계가 여러 도구에 흩어져 있다

분포 확인, log · z-score 변환, ANOVA와 사후검정, LDA, 상관 + FDR, LASSO, 회귀, 교차검증을 각각 다른 프로그램과 스크립트로 돌리고 결과를 다시 엑셀로 모읍니다.

문제 ③

성능이 부풀려지기 쉽다

같은 제품을 여러 속성 단계에서 켜 반복측정 데이터는, 같은 제품의 다른 시점이 학습과 검증에 함께 들어가면 예측 성능이 실제보다 좋게 나옵니다.

이 앱은 **표를 그대로 올리면 구조를 스스로 읽고**, 분포 점검에서 찾은 신호로 전처리 규칙을 제안하고, **제품 단위로 묶은 교차검증**으로 맛별 예측 성능을 정직하게 보여 줍니다. 모든 계산은 화면 안에서 바로 돌고, 사람은 규칙을 고치며 결과가 어떻게 바뀌는지 봅니다.

여섯 단계

2열의 단계 버튼을 차례로 누르면 됩니다. 각 단계의 결과는 다음 단계의 입력입니다.

1 불러오기 머리글 · 블록 · 그룹 · 시료명 열 인식	2 분포 히스토그램 · 상자그림 · Q-Q Q · 품질 신호	3 전처리 제외 · LOD 대체 · $\log_{10}$ · z-score	4 차이검정 ANOVA + Tukey · FDR · LDA	5 마커 Pearson + FDR · LASSO	6 예측 맛별 회귀식 · 제품 단위 교차검증
---	---	---	--	---	--

단계	방법	무엇을 알 수 있나
불러오기	다단 머리글 해석(블록 줄 · 단위 줄 · 변수 이름 줄), 그룹 열 · 시료명 열 자동 판별	시료 수, 블록별 변수, 반복측정 구조 (제품 × 단계), 중복 의심 열
분포	기술통계, 히스토그램 + 밀도, 그룹별 상자그림, 정규 Q-Q, 누적분포, z-score 열지도	치우침 · 0 값 · 극단값 · 결측 · 중복 열, 그룹을 잘 가르는 변수(η^2)
전처리	상수 · 중복 열 제외, 0을 검출한계(LOD)의 절반으로 대체, 치우친 변수 $\log_{10}$, z-score 표준화	변수별 규칙과 근거, 변환 전 · 후 분포 비교 — 규칙은 변수마다 직접 바꿀 수 있음
차이검정	일원 ANOVA + Tukey HSD + 문자 표기, BH-FDR 보정, 축소 LDA(제품 하나씩 빼고 판별)	그룹(숙성 단계)에 따라 달라지는 맛 · 성분, 성분만으로 단계가 얼마나 갈리는지
마커	맛 점수와 성분의 Pearson 상관 + FDR, LASSO 경로와 제품 단위 교차검증으로 λ 선택(1-SE 규칙)	맛별로 함께 움직이는 성분 후보와, 서로 겹치지 않는 핵심 마커
예측	접기마다 표준화 → LASSO → 상위 변수로 최소제곱, 제품 단위 5-fold × 50회 반복	맛별 R^2 · RMSE · MAE · Bias · CCC, 관측 vs 예측 그림, 마커 선택 빈도, 최종 회귀식

화면으로 보기

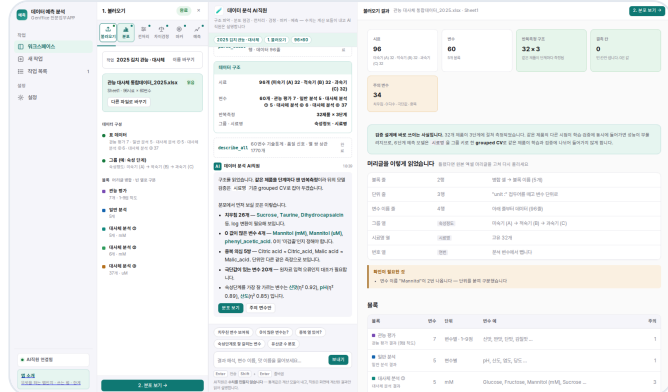
예시 데이터(김치 32제품 × 미숙기 · 적숙기 · 과숙기, 관능 7 · 일반 분석 5 · 대사체 48 변수)로 진행한 화면입니다. 그림을 누르면 크게 봅니다.

1 · 불러오기

머리글을 어떻게 읽었는지 먼저 보여 줍니다

병합된 블록 줄, "unit : mM" 단위 줄, 변수 이름 줄을 찾아 다섯 블록으로 나눕니다. 반복되는 세 범주 열은 속성 단계 그룹으로, 32개 이름이 단계마다 반복되는 열은 제품 (교차검증 묶음 키)으로 씁니다.

같은 성분이 mM과 μM 으로 두 번 적힌 **중복 의심 5쌍**을 상관계수 0.999 초과와 비율 ($\times 1000$)로 찾아냅니다.



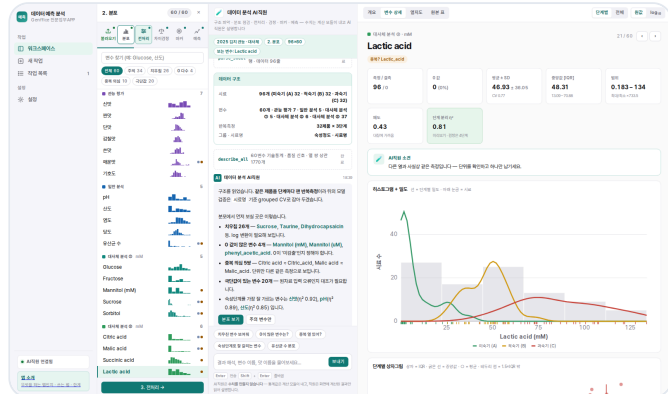
불러오기 결과 — 시료 · 변수 · 반복측정 구조 · 머리글 해석 · 블록

2 · 분포

변수 하나하나의 분포와 품질 신호

개요에서는 60개 변수의 작은 분포 그림과 신호 (치우침 · 0 값 · 극단값 · 중복)를 한 번에 보고, 상세에서는 히스토그램 · 그룹별 상자그림 · Q-Q · 누적분포를 나란히 봅니다.

- 원값 / $\log_{10}$ 전환, 단계별 / 전체 전환
- 치우침 큰 순 · 단계 차이(η^2) 큰 순 · 0 값 많은 순 정렬
- 울타리 밖 시료를 누르면 모든 그림에 표시

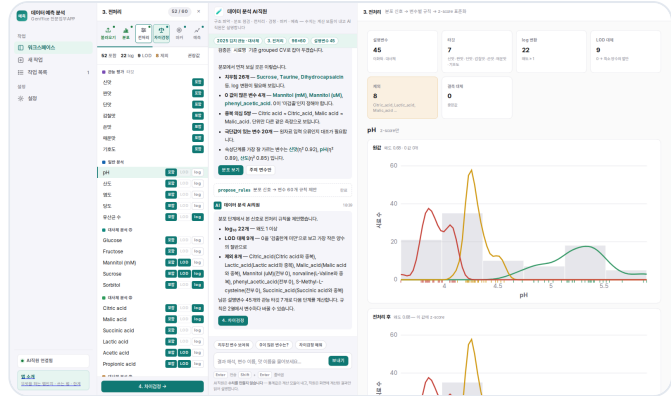


변수 상세 — Lactic acid: 단계가 진행될수록 커지는 전형적인 속성 지표

3 · 전처리

분포 신호에서 규칙을 제안하고, 사람이 고칩니다

0이 있는 변수는 LOD 대체, 왜도가 큰 변수는 $\log_{10}$, 값이 일정하거나 다른 열과 같은 변수는 제외를 제안합니다. 2열에서 변수마다 포함 · LOD · log를 바꾸면 뒤 단계 결과를 지우고 새 규칙으로 다시 계산합니다.

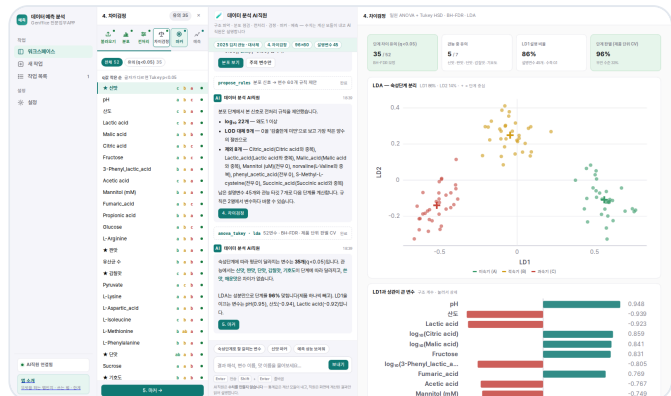


전처리 — 변수별 규칙과 근거, 변환 전 · 후 분포 비교

4 · 차이검정

숙성 단계에 따라 무엇이 달라지나

변수마다 일원 ANOVA와 Tukey 사후검정을 하고, 여러 번 검정한 효과를 FDR로 보정합니다. LDA는 성분만으로 단계를 얼마나 가를 수 있는지를 제품 하나씩 빼고 켜봅니다.

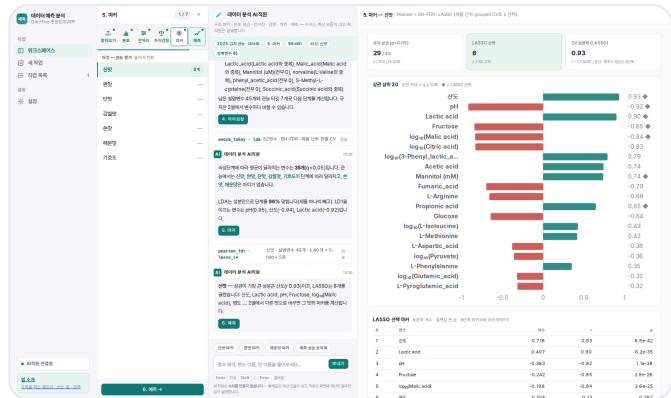


차이검정 — LDA 분리 그림, 가르는 성분, 혼동표, 변수별 ANOVA 표

5 · 마커

맛별로 함께 움직이는 성분

상관 상위 20개와 FDR q값, LASSO 경로와 λ 별 교차검증 오차를 보여 줍니다. 상관이 커도 LASSO가 고르지 않은 성분은 이미 고른 성분과 정보가 겹치는 경우입니다(예: pH와 산도).



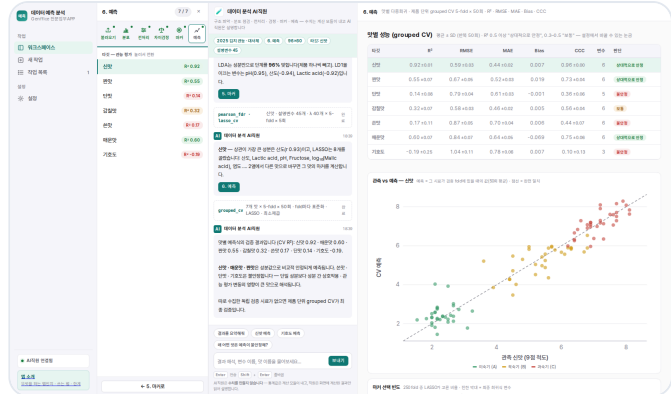
마커 — 신맛: 상관 막대, 선택 마커 표, LASSO 경로, 교차검증 오차 곡선

6 · 예측

맛별 예측식과 정직한 성능

같은 제품의 세 시점은 항상 같은 접기에 묶고, 표준화 · 마커 선택 · 회귀를 접기마다 학습 데이터로만 다시 합니다. 예시 데이터에서는 신맛이 가장 잘 예측되고, 단맛 · 쓴맛 · 기호도는 성분만으로는 불안정합니다.

예시 결과(교차검증 R^2) — 신맛 **0.92** · 매운맛 **0.60** · 짭맛 **0.55** · 감칠맛 0.32 · 쓴맛 0.17 · 단맛 0.14 · 기호도 -0.19



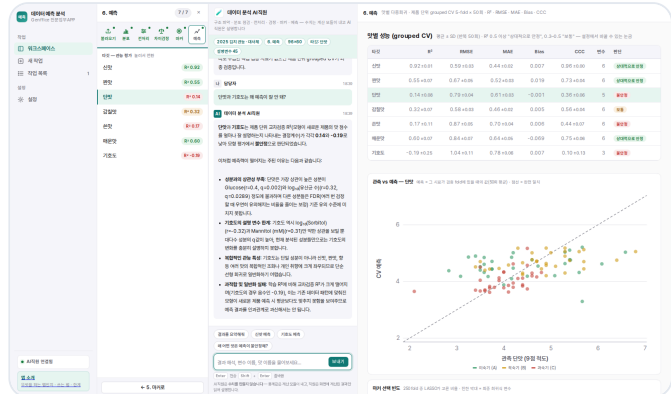
예측 — 맛별 성능 표, 관측 vs 예측, 마커 선택 빈도, 최종 회귀식

AI직원

계산된 결과를 읽고 설명합니다

3열에 질문하면 AI직원이 지금 화면에 계산된 값만 인용해 답합니다. 변수 이름이나 맛 이름을 말하면 해당 화면으로 옮겨 주고, 단계가 끝날 때마다 결과를 요약해 보고합니다.

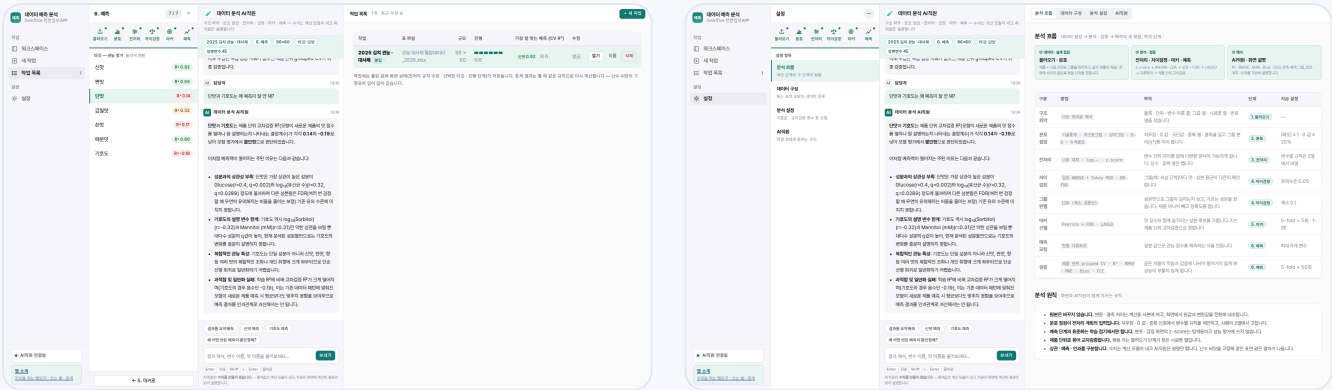
- 새 수치를 계산하거나 지어내지 않습니다
- 상관 · 예측 · 인과를 구분하고, 낮은 R^2 를 과신하지 않게 말합니다
- 연결되지 않은 환경에서는 정해진 질문에 규칙 기반으로 답합니다



"단맛과 기호도는 왜 예측이 잘 안 돼?" — 교차검증 R^2 와 상관 · FDR 값을 근거로 설명

작업 · 설정

왼쪽 메뉴는 워크스페이스 · 새 작업 · 작업 목록 · 설정 네 가지입니다.



작업 목록 — 표 파일 하나가 작업 하나. 진행 단계와 가장 잘 맞는
예측을 한눈에

설정 — 분석 흐름 · 데이터 구성 · 분석 설정 · 시직원

워크스페이스	지금 열린 작업의 단계 화면. 작업이 없으면 시작 화면과 최근 작업을 보여 줍니다.
새 작업	작업 이름을 정하고 표 파일(.xlsx · .csv)을 올리거나 예시 데이터로 시작합니다. 화면 어디에 파일을 끌어다 놓아도 됩니다.
작업 목록	저장된 작업을 열고, 이름을 바꾸고, 지웁니다. 작업에는 올린 표와 전처리 규칙 수정 · 선택한 맛 · 진행 단계가 저장되어, 다시 열면 보던 단계로 돌아갑니다.
설정	앱의 분석 흐름과 데이터 구성을 설명하고, 치우침 기준 · 유의수준 · 교차검증 접기 수와 반복 횟수 · 회귀식 최대 변수 수 · 판정 R ² 기준 등을 바꿉니다. 저장하면 지금 작업을 새 설정으로 다시 계산합니다.

믿을 수 있게

결과를 논문 · 보고서에 옮길 수 있도록 지키는 원칙입니다.

1

원본은 바꾸지 않습니다

변환과 결측 처리는 계산용 사본에만 하고, 화면에서 원값과 변환값을 전환해 대조할 수 있습니다.

2

해석이 틀릴 수 있는 곳은 드러냅니다

머리글 추정, 중복 이름, 0 값, 척도 밖 값을 숨기지 않고 표시합니다 — 사람이 원본과 대조할 수 있어야 합니다.

3

누수를 막습니다

같은 제품은 같은 접기에 묶고, 예측 단계의 표준화 · 마커 선택 · 회귀는 접기마다 학습 데이터로만 합니다.

4

같은 표면 같은 결과

교차검증 난수의 씨앗을 고정해, 작업을 다시 열거나 다른 사람이 계산해도 같은 값이 나옵니다.

5

AI직원은 수치를 만들지 않습니다

통계값은 계산 모듈이 내고, AI직원에게는 화면의 계산 요약만 보냅니다. 원본 표 전체는 보내지 않습니다.

6

상관 · 예측 · 인과를 구분합니다

마커는 "함께 움직이는 성분 후보"이고, 예측식은 성분으로 맛 점수를 추정하는 식입니다. 원인을 말하지 않습니다.

데이터 형식

연구실에서 쓰던 엑셀을 고치지 않고 그대로 올리는 것을 목표로 했습니다.

항목	읽는 방법
파일	.xlsx · .csv (최대 20MB). 값이 있는 첫 시트를 읽습니다.
행	한 줄 = 시료 한 개.
머리글	위에서부터 블록 줄(병합 가능) · 단위 줄(unit : mM) · 변수 이름 줄. 블록 줄과 단위 줄은 없어도 됩니다.
블록 안 묶음	빈 열은 같은 블록 안의 하위 묶음(예: 당류 유기산 LC-MS)으로 읽습니다.
그룹 열	반복되는 소수 범주(예: 미숙기 (A) · 적숙기 (B) · 과숙기 (C)).
시료명 열	글자가 가장 다양한 열. 같은 이름이 그룹마다 있으면 반복측정으로 보고 교차검증 묶음 키로 씁니다.
예측 대상	머리글에 "9점 척도"처럼 척도가 적힌 블록(관능 평가).
빈 칸 · 0	빈 칸은 결측(중앙값으로 채움), 0은 값입니다. 0이 '검출한계 미만'이면 전처리에서 LOD 대체를 겁니다.

지원

관능 평가

맛 점수 — 예측
대상

지원

기본 지표 ·

대사체

pH · 산도 · 염도 ·
당도 · 군수, 당류 ·
유기산 · 아미노산

확장 예정

사진

색 · 조직 · 양념
분포를 이미지
임베딩으로

확장 예정

미생물

16S / ASV 군집을
CLR 변환해 같은
흐름에

한계와 다음 단계

지금 결과를 어디까지 믿어도 되는지와, 다음에 붙일 기능입니다.

항목	지금	상태
독립 검증	모델을 만들 때 쓰지 않은 별도 시료가 없으면, 제품 단위 교차검증이 가능한 가장 엄격한 검증입니다. 새 시료를 측정하면 같은 형식으로 올려 회귀식을 그대로 적용하는 화면을 붙일 수 있습니다.	다음 단계
반복측정 검증	일원 ANOVA는 시료를 독립으로 가정해 p값이 다소 작게 나올 수 있습니다. 보고용으로는 제품을 임의효과로 둔 혼합모형을 함께 보는 것이 안전합니다.	다음 단계
다중모달	사진 · 미생물 데이터는 자리만 두었습니다. 표 데이터 모형으로 기준 성능을 확보한 뒤 하나씩 더해 성능이 실제로 오르는지 확인하는 순서를 권합니다.	확장 예정
보고서 내보내기	표와 그림은 화면에서 확인합니다. 논문용 표 · 그림 묶음 내보내기는 아직 없습니다.	다음 단계
계산 위치	통계 계산은 브라우저에서 돌아 수백 시료 × 수백 변수 규모에 맞습니다. 예측 단계(맛 7개 × 250회 적합)는 수십 초 걸릴 수 있습니다.	동작

결과는 탐색과 의사결정을 돕는 참고 자료입니다. 마커와 예측식은 원인을 증명하지 않으며, 보고 · 출판 전에는 원자료 대조와 전문가 검토가 필요합니다.

Genffice 전문업무APP · 데이터 예측 분석 · [앱 열기](#) · [소개 슬라이드](#) · [소개 영상](#) · [PDF](#)