스마트팜코리아 데이터마트 활용사례 수기 공모전

효율적인 이유자돈 관리를 위한 분석 및 솔루션 →

양돈 # 이유자돈 생존율 # 양돈기침 유형분류 # 음향데이터

스마트팜코리아
SMARTFARM KOREA

목차

스마트팜코리아
SMARTFARM KOREA

목차

스마트팜코리아
SMARTFARM KOREA

주제 선정 배경

스마트팜코리아 SMARTFARM KOREA

5년 사이 돼지농가 열에 아홉만 살아남았다

2023.01.25 23:14:12

크게보기

통계청, 5년간 가축동향조사 분석 결과 5년간 돼지농가수 618호, 9.8% 감소...대부분 1000마리 미만 농가

지난해 4분기 기준 양돈장 숫자가 역대 최저를 기록한 가운데 5년 사이 10% 가까운 농장이 감소한 것으로 분석되었습니다. 감소한 농장은 대부분 규모가 작은 농장입니다.

source : pigpeople.net

돼지고기 팔수록 손해..양돈농가 '울상'

우크라이나 러시아 전쟁 여파로 국제곡물가가 급등하면서 돼지 사룟값도 2년 새 2배 가까이 뛰었습니다.

문제는 돼지고기값은 계속 하락세에 있다는 점입니다.

한때 1킬로그램 7천8백 원까지 치솟았던 도매가격이 최근 4천3백 원대까지 떨어졌습니다.

돼지고기 1킬로그램 생산비는 5천5백 원,

농가는 한 마리를 출하하면 많게는 10만 원가량 손해를 보고 있습니다.

source : imbc.com

1 양돈 농가의 지속적인 하향세를 막아 국내 양돈 농가를 살리기 위함

-> 한국 양돈 경쟁력 강화

2 최근 우크라이나-러시아 전쟁 영향으로 인한 양돈 원가 급등의 대안책 필요

-> 인공지능 기술 기반 솔루션 도입

3 ICT 기반 농업 혁신 필요

-> 데이터를 활용한 의사결정 필요

목차

스마트팜코리아
SMARTFARM KOREA

활용사례 및 데이터 선정

스마트팜코리아
SMARTFARM KOREA

활용사례1. 데이터 분석을 통한 인사이트 도출

Objective : 이유두수 및 총산 대비 이유두수에 영향을 미치는 요인 분석

데이터셋 목록 (스마트팜 정형 데이터 셋)

- 번식로우 데이터 (스마트팜)
- 포유모돈 급이 번식 데이터 (스마트팜)
- 온도 및 습도 데이터 (기상청)

활용사례2. 머신러닝 기술을 통한 예측 솔루션 제공

Objective : 머신을 통한 양돈기침 유형분류로 질병에 대한 신속한 조치 시스템 구축

데이터셋 목록 (스마트팜 비정형 데이터 셋 – Audio)

- 양돈기침 데이터 (스마트팜)
- 온도 및 습도 데이터 (기상청)

목차

스마트팜코리아
SMARTFARM KOREA

활용사례1-1. '번식로우' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Create Data Pipeline

데이터 로딩

Preprocessing

1. 이유두수가 null이 아닌 sample 선택
2. 이유일이 2023년 03월 15일 미만,
분만일이 2015년 01월 01일 이상인 샘플 선택
3. 이유두수가 0 이상 30 미만인 샘플 선택
4. '지역' 컬럼 null값을 'unknown'으로 대치

외부데이터 결합

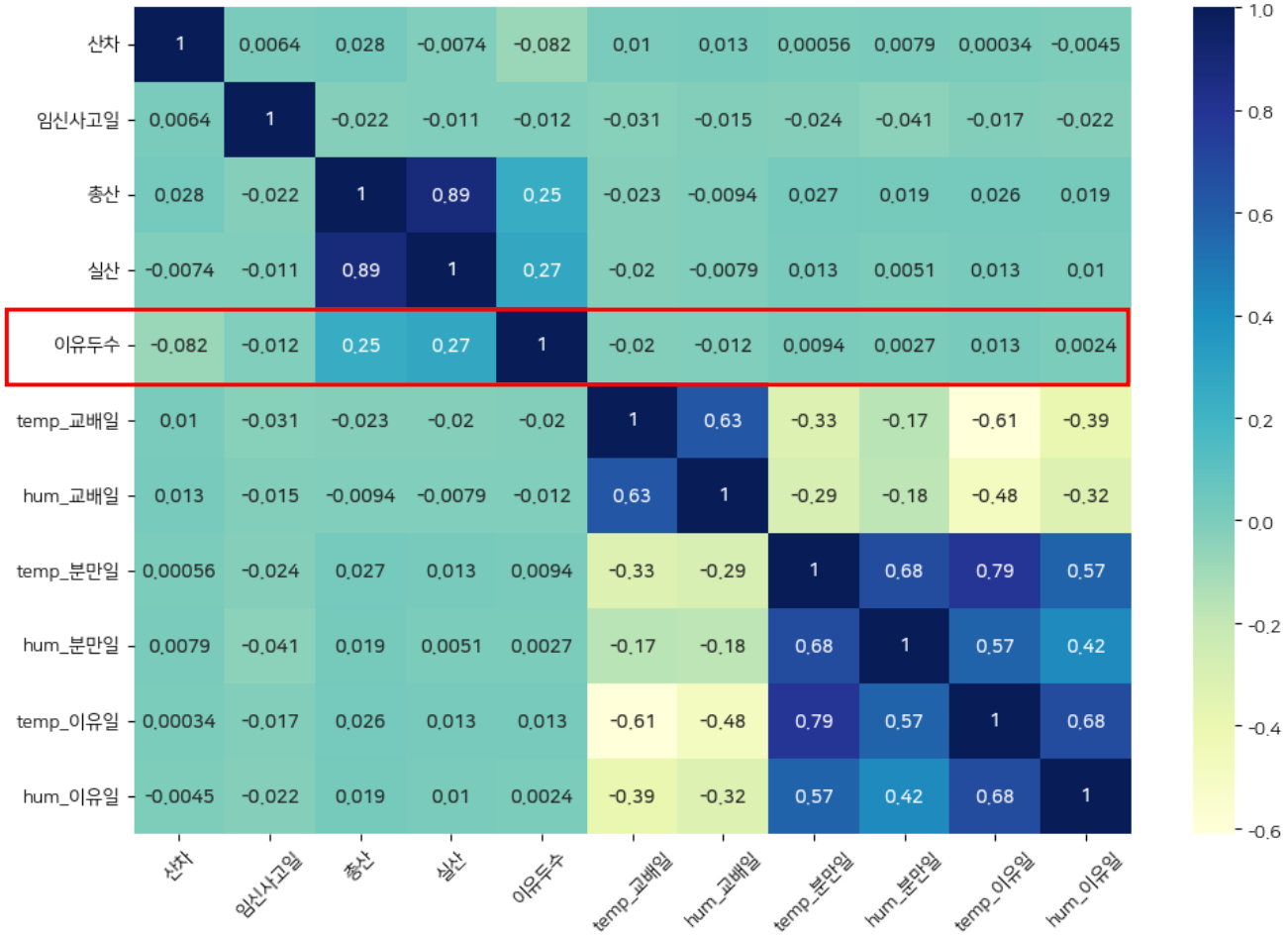
(*기상청 온도, 습도)

* 온도는 서울최고기온, 습도는 부산 평균습도 사용

활용사례1-1. '번식로우' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Data Analysis (Target : 이유두수)



수치형데이터에 대한 상관관계 분석

✓ 이유두수와 총산 & 실산의 상관관계가 낮다
총산과 별개로 성장환경이 중요하다는 것을 암시한다

✓ 온도나 습도와 같이 기후적 요인과는
큰 상관관계가 없는 것으로 보인다

활용사례1-1. '번식로우' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Data Analysis (Target : 이유두수)

범주형데이터 대한 분산분석

✓ 지역별 평균 이유두수 *ANOVA P-Value : 0.00000

지역	unknown	경기도	경상남도	경상북도	전라남도	전라북도	제주특별자치도	충청남도
이유두수	10.926203	10.218155	10.775912	10.508728	10.948655	9.855	10.250487	10.543819

➡ 경남과 전남의 평균 이유두수가 비교적 높다

✓ 규모별 평균 이유두수 *ANOVA P-Value : 0.00000

규모	1000두 이상	100~200	100두 미만	200~300	300~400	400~500	500~1000
이유두수	10.437462	9.882353	9.442804	10.699946	10.195407	11.148985	10.91099

➡ 400-500규모의 평균 이유두수가 비교적 높다

* 유효숫자는 소수점 아래 5자리

활용사례1-1. '번식로우' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Data Analysis (Target : 이유두수)

범주형데이터 대한 분산분석

✓ 교배 계절별 평균 이유두수 *ANOVA P-Value : 0.00000

season_교배일	봄	여름	가을	겨울
이유두수	10.562883	10.619143	10.600838	10.735276

➡ **겨울 및 여름에 교배한 그룹**의 이유두수가 비교적 높다

✓ 분만 계절별 평균 이유두수 *ANOVA P-Value : 0.00000

season_분만일	봄	여름	가을	겨울
이유두수	10.734233	10.559948	10.667946	10.543514

➡ **봄 및 가을에 분만한 그룹**의 이유두수가 비교적 높다

* 유효숫자는 소수점 아래 5자리

활용사례1-1. '번식로우' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Summary (Target : 이유두수)

1. 400-500규모의 평균 이유두수가 비교적 높다는 사실로 보아,
   적절한 크기의 농가에서 더 높은 이유두수가 나타난다는 것을 알 수 있다.
   이에 대한 이유는 관리의 용이함 등이 있을 수 있다.
   농가 확장 시 그에 맞는 관리 시스템 또한 잘 갖춰져야지 자돈들이 잘 성장할 수 있음을 의미한다.

2. 봄과 가을에 분만한 자돈들의 이유두수가 비교적 더 높다는 사실로 보아,
   봄과 가을에만 발생하는 어떠한 요인이 자돈 성장에 영향을 미친다는 사실을 알 수 있다.
   온도 혹은 습도는 상관관계가 낮았기 때문에 기후적 영향이 아닌 다른 요인에 의한 것으로 생각해 볼 수 있다.

목차

스마트팜코리아
SMARTFARM KOREA

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Preprocessing

데이터 로딩

전처리1

1. 급이설정량, 섭취량이 0일 경우 null 값 대치
(평균화 부분에서 계산에 제외하기 위함)

그룹핑 & 평균화

(keys : 분만일, 이유일, 총산, 분만시BCS)

외부데이터 결합

(*기상청 온도, 습도)

전처리2

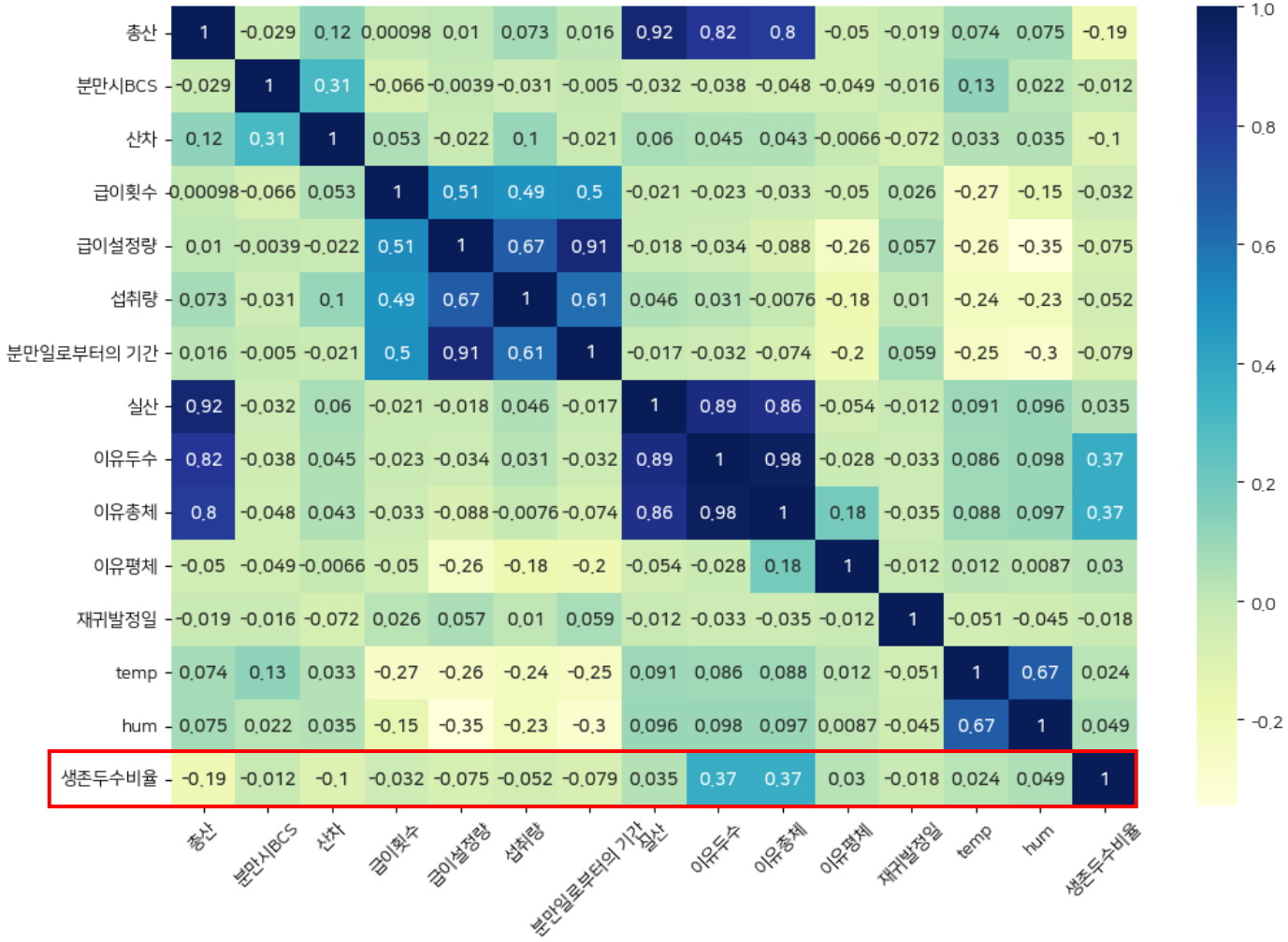
1. 총산이 0초과이고 이유두수가 0이상인 샘플 선택
2. 이유두수가 총산보다 작거나 같은 샘플 선택
3. Target으로 설정할 '생존두수비율' 변수 생성
(formula : 이유두수 / 총산)

* 온도는 서울최고기온, 습도는 부산 평균습도 사용

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Data Analysis (Target : 생존두수비율)



수치형데이터에 대한 상관관계 분석

- ✓ 총산과 생존두수비율의 상관관계가 음의 방향이다
  즉, 자돈들이 많이 태어날수록 생존비율이 낮다
- ✓ 먹이 관련 변수와 생존두수비율은
  음의 상관관계를 보인다
- ➡ 위 두 직관적이지 않은 사실에 대해 연구가 필요하다

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Data Analysis (Target : 생존두수비율)

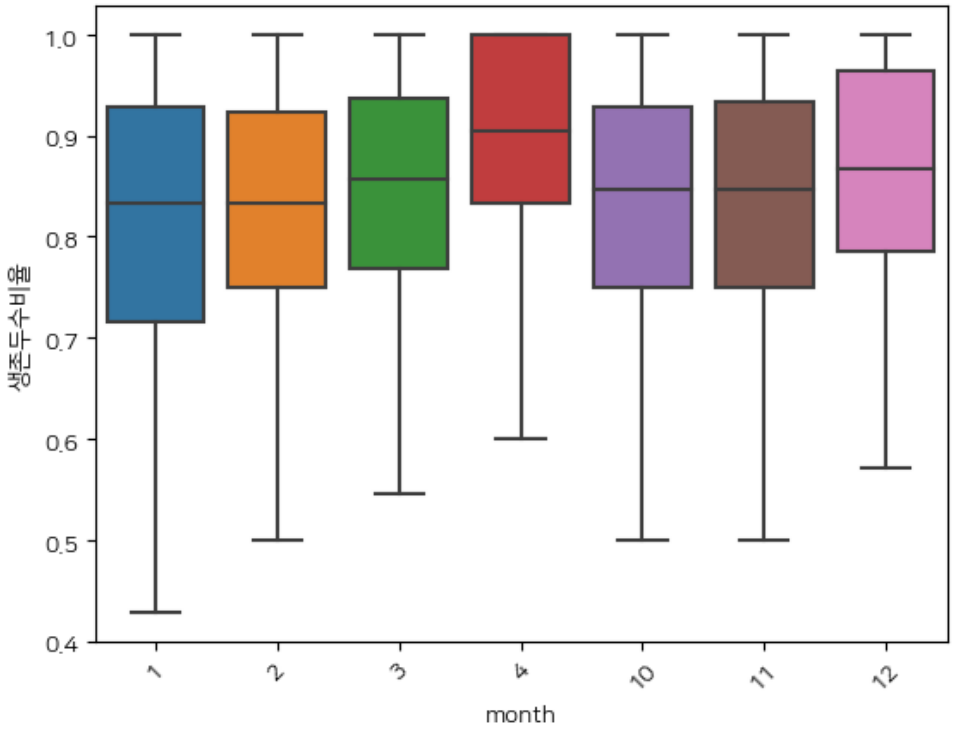
범주형데이터 대한 분산분석

✓ 분만 월별 평균 생존두수비율 *ANOVA P-Value : 0.00000

month	1	2	3	4	10	11	12
생존두수비율	0.814162	0.825059	0.84082	0.883968	0.834291	0.833883	0.857759

➡ **4월에 분만한 그룹**은 생존두수비율평균이 약 88.4%로 가장 높다

➡ 이는 앞서 연구된 '번식로우' 데이터 분석의 결과 중
'봄 및 가을에 분만한 그룹의 이유두수가 비교적 높다' 라는 사실을
뒷바침하는 또 다른 사실이라고 할 수 있다



* 유효숫자는 소수점 아래 5자리

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

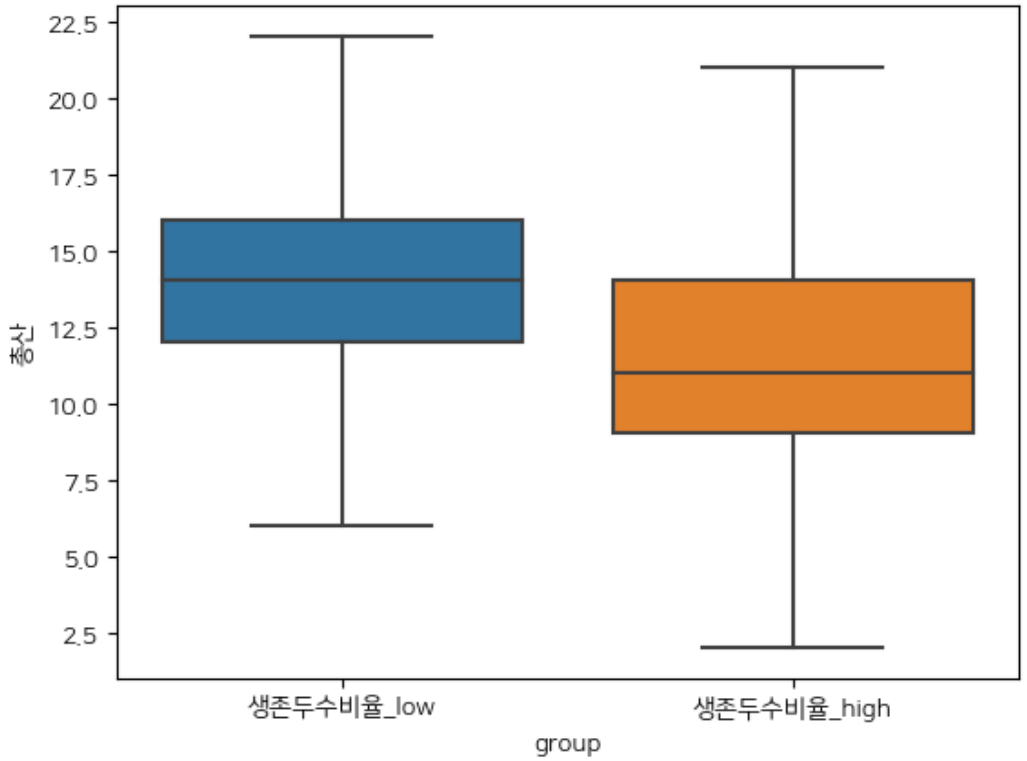
Data Analysis (Target : 생존두수비율)

그룹화 분석 - 생존두수비율의 Q3(상위25%) 이상 / 생존두수비율의 Q1(하위25%) 이하

✓ 그룹별 평균 총산 *ANOVA P-Value : 0.00000

group	생존두수비율_low	생존두수비율_high
총산	13.786241	11.469807

➡ 생존두수비율이 높은 그룹의 총산 평균은 낮은 그룹보다 통계적으로 유의미하게 낮다



* 유효숫자는 소수점 아래 5자리

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Data Analysis (Target : 생존두수비율)

그룹화 분석 - 생존두수비율의 Q3(상위25%) 이상 / 생존두수비율의 Q1(하위25%) 이하

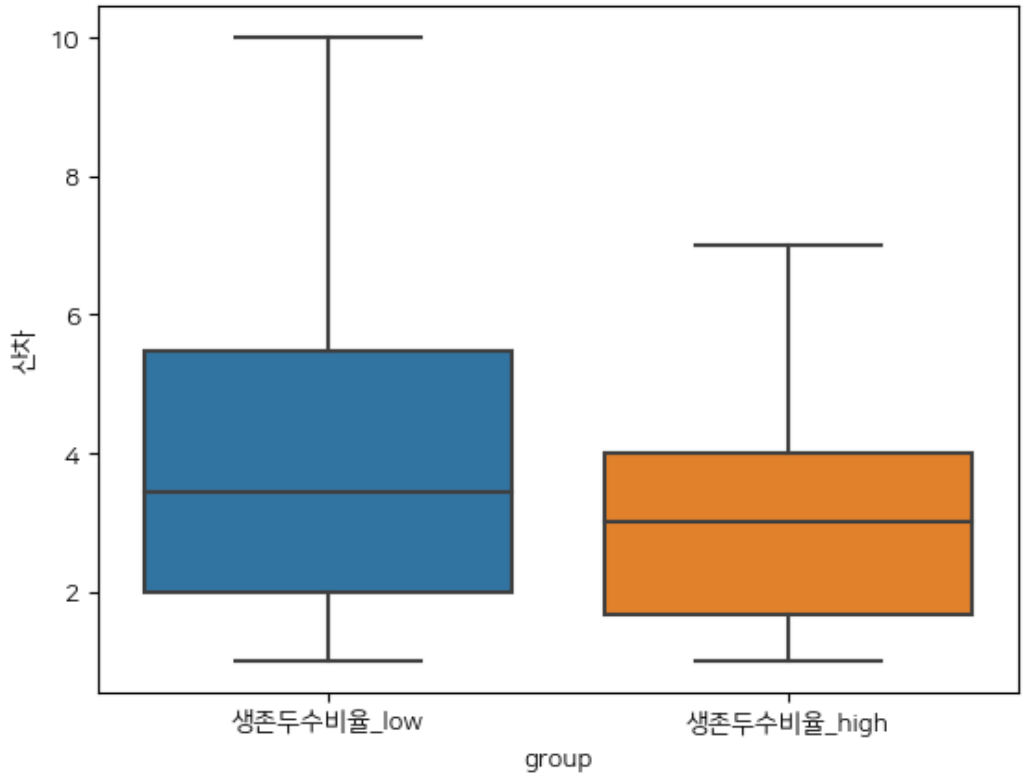
✓ 그룹별 평균 산차 *ANOVA P-Value : 0.00000

group	생존두수비율_low	생존두수비율_high
산차	3.893672	3.196812

생존두수비율이 높은 그룹의 산차 평균이 낮다

즉, 분만한 횟수가 많은 모돈의 자돈들은

이유기까지 생존할 확률이 더 적다는 것을 의미한다



* 유효숫자는 소수점 아래 5자리

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Data Analysis (Target : 생존두수비율)

그룹화 분석 - 생존두수비율의 Q3(상위25%) 이상 / 생존두수비율의 Q1(하위25%) 이하

1 급이횟수

*ANOVA P-Value : 0.04197

group	생존두수비율_low	생존두수비율_high
급이횟수	2.600784	2.568295

2 급이설정량

*ANOVA P-Value : 0.00013

group	생존두수비율_low	생존두수비율_high
급이설정량	6.225063	6.076922

3 섭취량

*ANOVA P-Value : 0.00148

group	생존두수비율_low	생존두수비율_high
섭취량	5.674796	5.539352

먹이 관련 변수들 모두 P-Value가 낮지는 않아 평균이 뚜렷이 유의미하게 차이가 난다고 보기엔 어렵다

그러나 세 변수를 통해 알 수 있는 점은 생존비율이 높은 그룹은 모돈이 평균적으로 미세한 차이로 더 적은 먹이를 먹는다

* 유효숫자는 소수점 아래 5자리

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

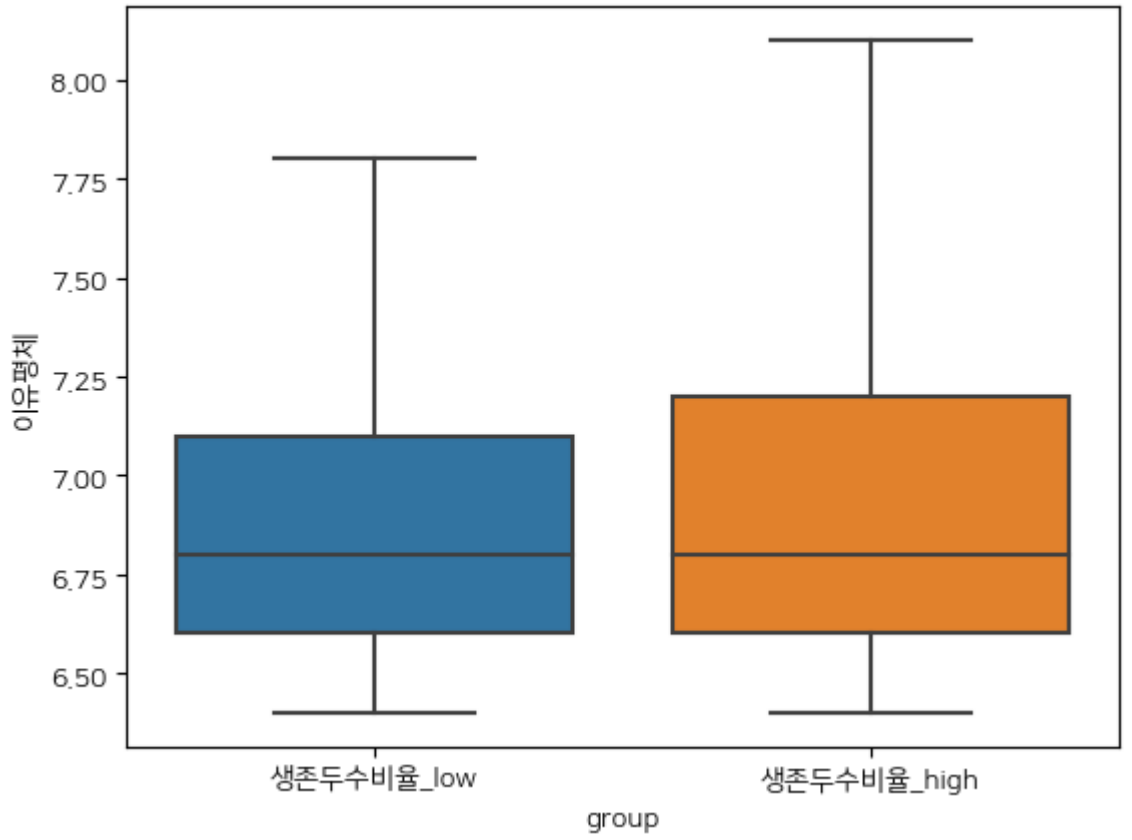
Data Analysis (Target : 생존두수비율)

그룹화 분석 - 생존두수비율의 Q3(상위25%) 이상 / 생존두수비율의 Q1(하위25%) 이하

✓ 그룹별 평균 이유평체 *ANOVA P-Value : 0.03355

group	생존두수비율_low	생존두수비율_high
이유평체	6.912776	6.955435

➡ 생존두수비율이 높은 그룹의 이유자돈들은
평균적으로 몸무게가 더 많이 나간다



* 유효숫자는 소수점 아래 5자리

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

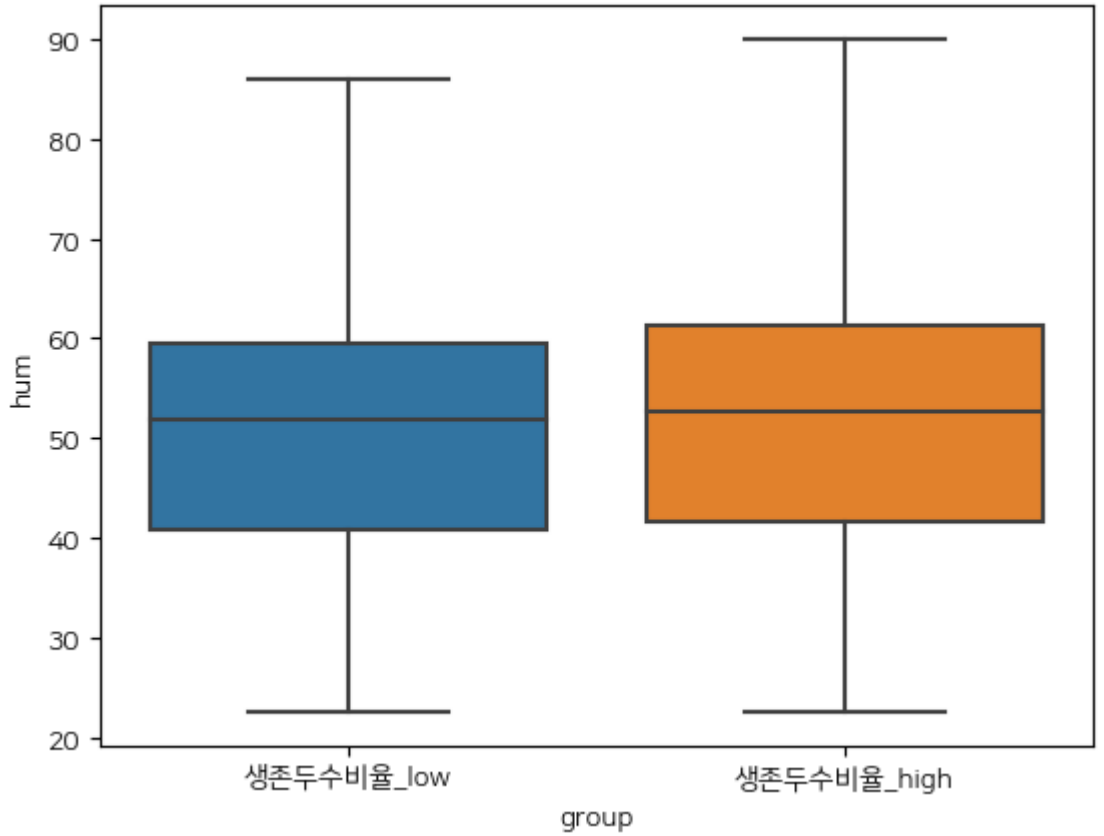
Data Analysis (Target : 생존두수비율)

그룹화 분석 - 생존두수비율의 Q3(상위25%) 이상 / 생존두수비율의 Q1(하위25%) 이하

✓ 그룹별 평균 부산광역시 평균습도 *ANOVA P-Value : 0.04044

group	생존두수비율_low	생존두수비율_high
hum	51.765356	53.273913

➡ 생존두수비율이 높은 그룹의
분만일에서 부산광역시 평균습도가 더 높았다



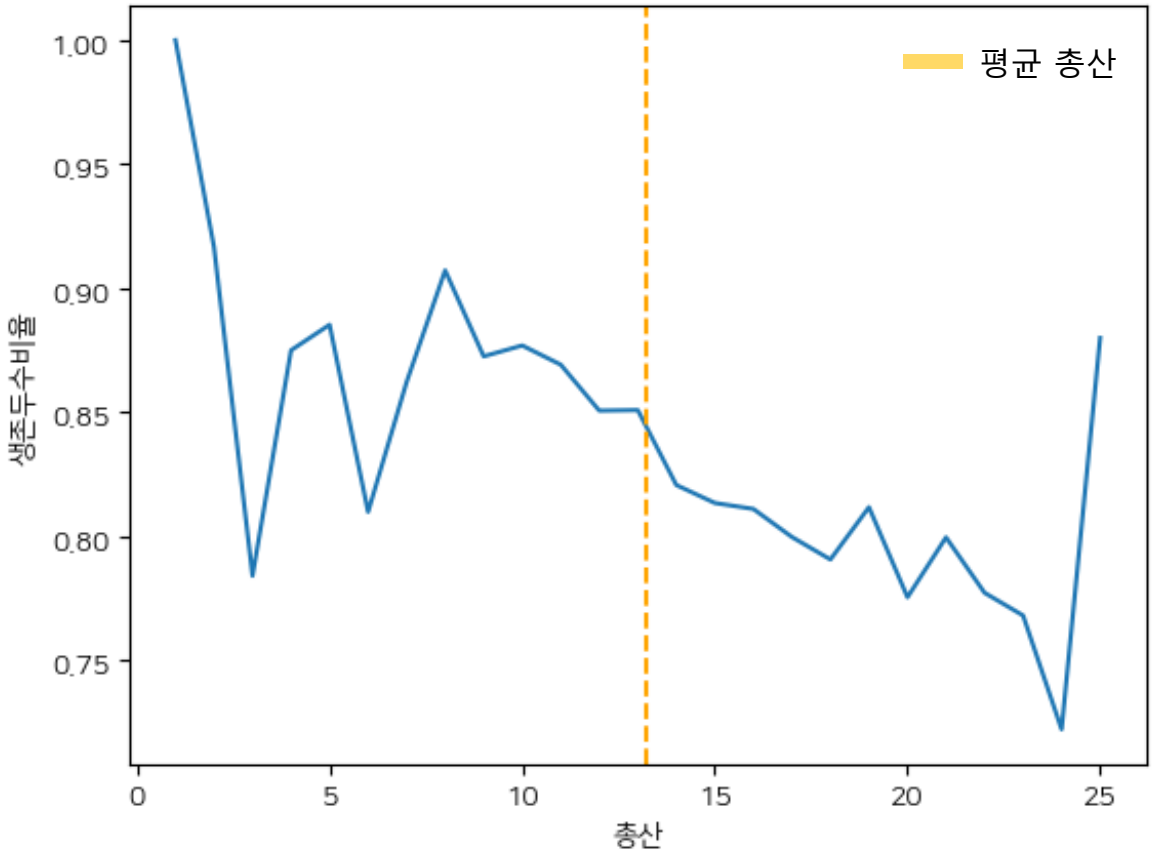
* 유효숫자는 소수점 아래 5자리

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Data Analysis (Target : 생존두수비율)

추가 분석 – 총산에 따른 생존두수비율 시각화



✓ high single digit 수준까지는 자돈들의 숫자와 생존확률이 양의 관계를 갖는 것으로 보인다

✓ 그러나, mid double digit 수준부터는 자돈들의 숫자와 생존확률이 음의 관계를 갖는 것으로 뚜렷이 변화한다

➡ 자돈들이 많이 태어날수록 그에 따른 케어시스템이 작동해주어야 하는데 그에 맞는 환경을 제공하기에는 충분치 않아 생존확률이 줄어드는 것으로 해석해 볼 수 있다

활용사례1-2. '포유모돈 급이 번식' 데이터 분석

스마트팜코리아 SMARTFARM KOREA

Summary (Target : 생존두수비율)

1 훌륭한 양돈을 위해서는 무조건 먹이를 많이 주고 많은 자돈들을 낳도록 하는 것이 중요한 것이 아니라,
적은 자돈이라도 질 좋은 환경에서 키워내는 것이 중요하다는 것을 알 수 있다.
이유자돈 생존비율이 더 높은 그룹은 약 2마리 정도로 자돈들의 수가 더 적고, 모돈 먹이량이 조금 더 적다.
또한 자돈의 절대적인 수는 낮지만 자돈 한 마리 당 평균 몸무게가 더 많이 나간다.
즉, 새끼를 많이 낳지는 않지만 더 건강한 새끼를 낳는다는 것이다.
자돈 수 만큼 그들을 케어할 수 있는 환경도 제공되어야 이유기까지 생존하는 자돈들이 많다는 것을 알 수 있다.

2 봄에 분만한 이유자돈들의 생존률이 상대적으로 더 높다는 것을 알 수 있다.
앞서 활용사례1 데이터에서 연구된 부분 중 '**봄에 이유자돈 수가 더 높다**'라는 사실을 뒷바침하는 사실이다.
하지만, 해당 데이터에서도 온도 혹은 습도와 생존두수비율의 상관관계는 크지 않았다.
왜 봄에 특별히 이유자돈 수가 더 많고 총산 대비 생존률이 높은지 추가적인 연구가 필요할 것으로 보인다.

목차

스마트팜코리아
SMARTFARM KOREA

활용사례2. 양돈기침 유형분류 솔루션

스마트팜코리아 SMARTFARM KOREA

Create Data Pipeline

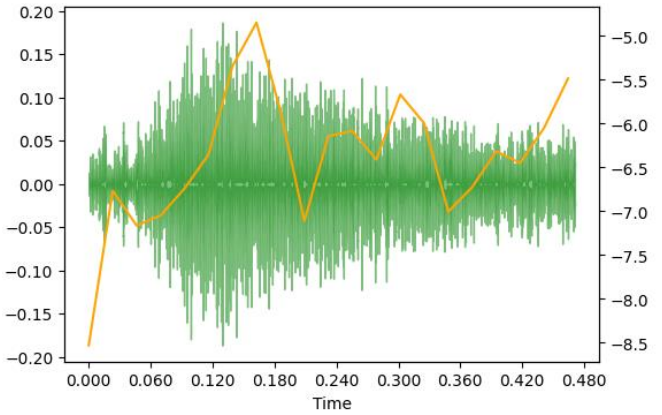
데이터 로딩

(양돈기침 음성 데이터)

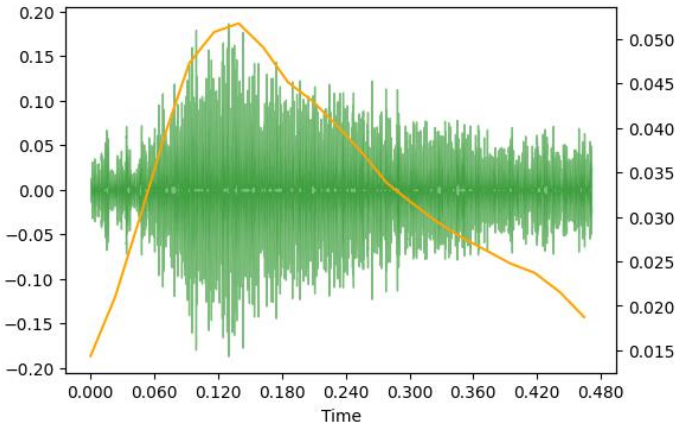
Audio Feature Extraction

1. ZCR (Zero Crossing Rate)
2. MFCC (dim : 32)
3. Chroma Frequencies (dim : 16)
4. RMS (Root Mean Square)

✓ Audio Wave & MFCC Feature



✓ Audio Wave & RSM Feature



활용사례2. 양돈기침 유형분류 솔루션

스마트팜코리아
SMARTFARM KOREA

Create Data Pipeline

Preprocessing & Feature Engineering

1. 웨이브 길이가 0인 샘플 제거
2. Transform ZCR vector to the scalar mean of the number of 'True' in ZCR vector
3. Add the mean & standard deviation on MFCC
4. Add the mean & standard deviation Chroma Frequencies
5. Add the mean, standard deviation, max, min, min-max range, min-max pct range on RMS

외부데이터 결합

(*기상청 온도, 습도)

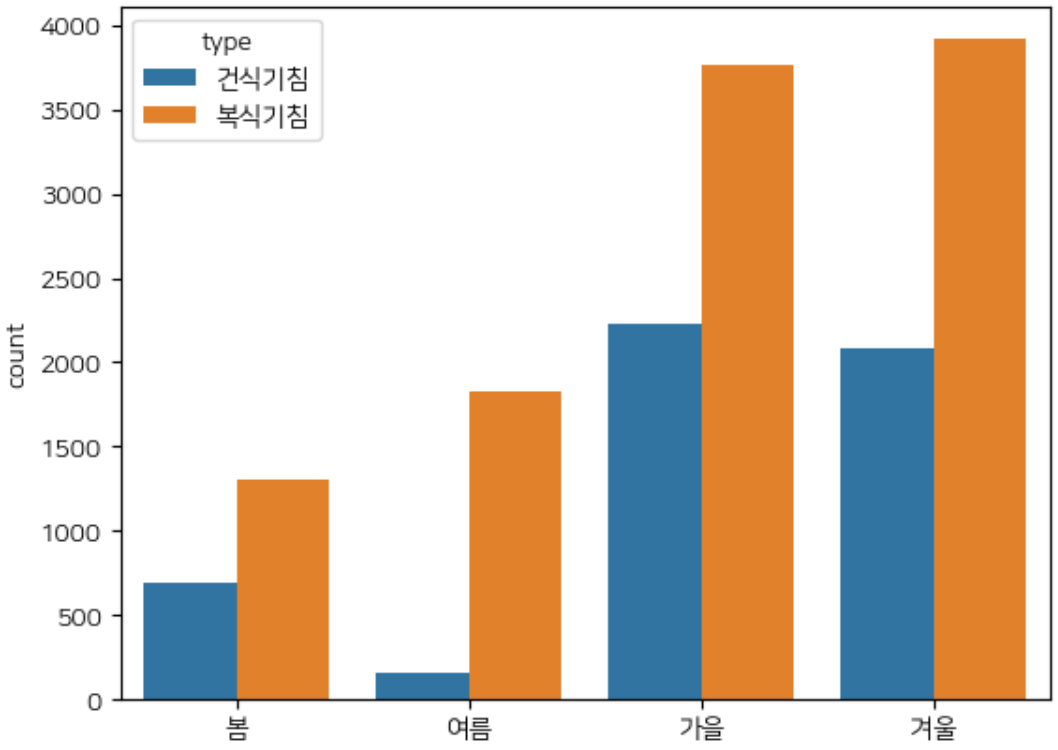
* 온도는 서울최고기온, 습도는 부산 평균습도 사용

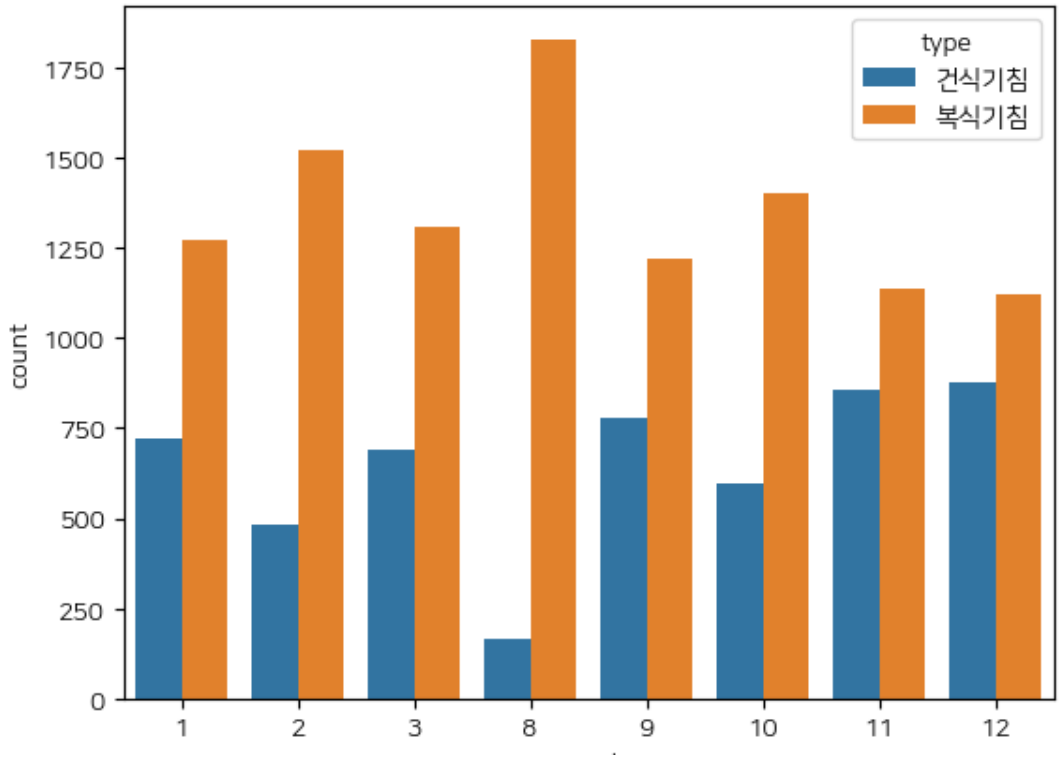
활용사례2. 양돈기침 유형분류 솔루션

스마트팜코리아 SMARTFARM KOREA

EDA

✓ 계절별 및 월별 기침 유형 분포도





➡ 본 데이터셋에 따르면 여름철 특히 8월 복식기침의 비율이 급격히 높아진다

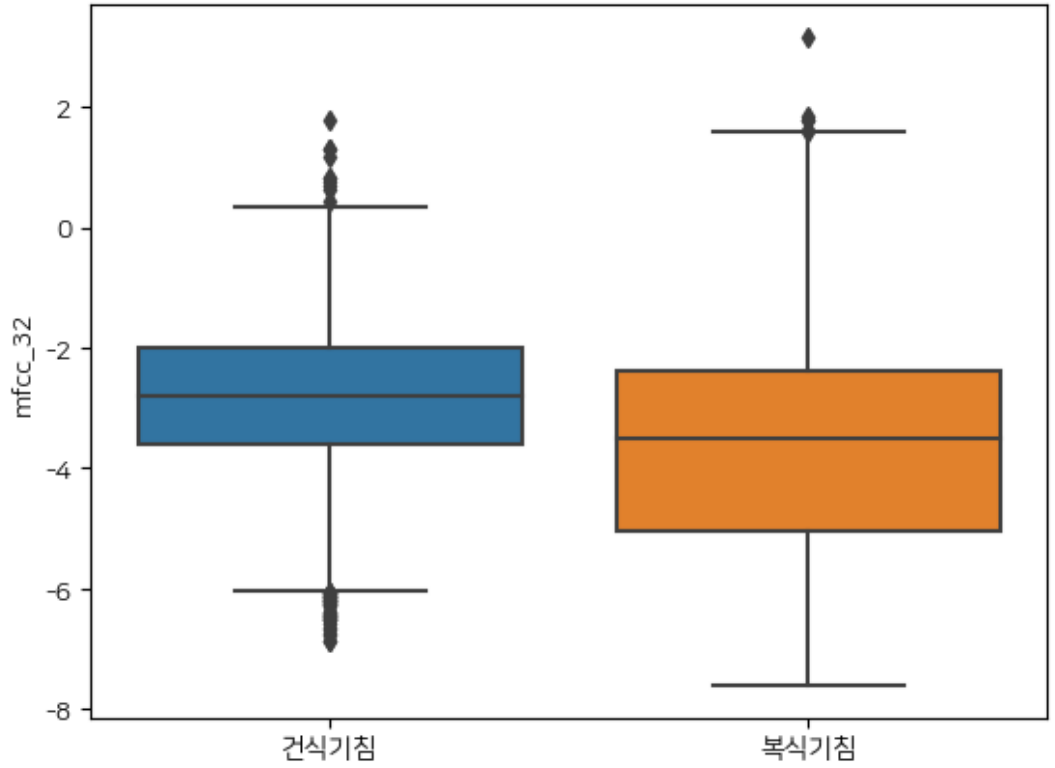
활용사례2. 양돈기침 유형분류 솔루션

스마트팜코리아 SMARTFARM KOREA

EDA

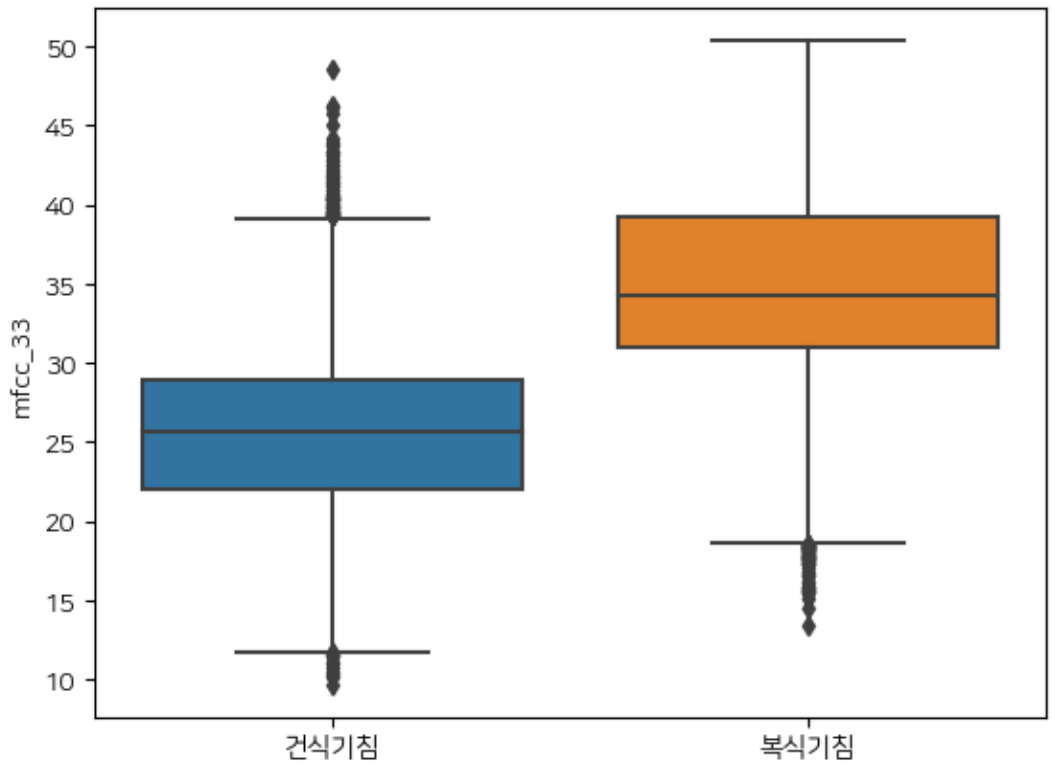
✓ MFCC vector 평균

(*P-Value : 0.00000)



✓ MFCC vector 표준편차

(*P-Value : 0.00000)



➡ MFCC 관련한 두 feature는 기침유형 그룹에 따른 평균의 차이가 통계적으로 있음을 보여준다

* 유효숫자는 소수점 아래 5자리

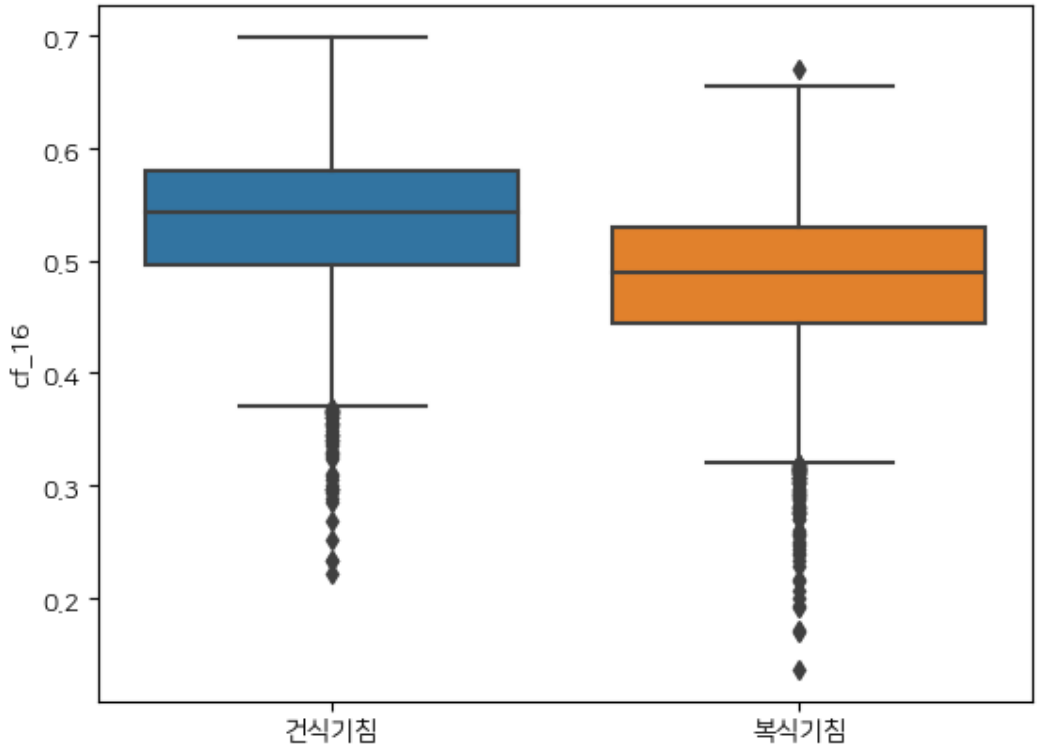
활용사례2. 양돈기침 유형분류 솔루션

스마트팜코리아 SMARTFARM KOREA

EDA

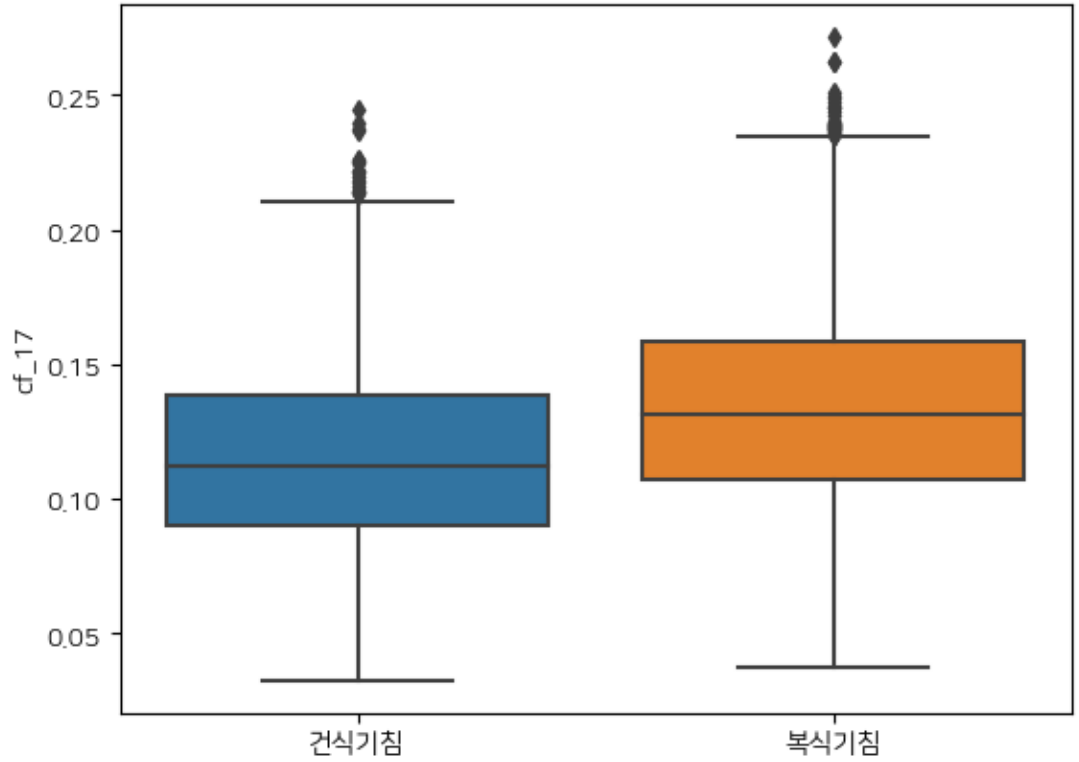
✓ Chroma Frequencies vector 평균

(*P-Value : 0.00000)



✓ Chroma Frequencies vector 표준편차

(*P-Value : 0.00000)



➡ Chroma Frequencies 관련한 두 feature는 기침유형 그룹에 따른 평균의 차이가 통계적으로 있음을 보여준다

* 유효숫자는 소수점 아래 5자리

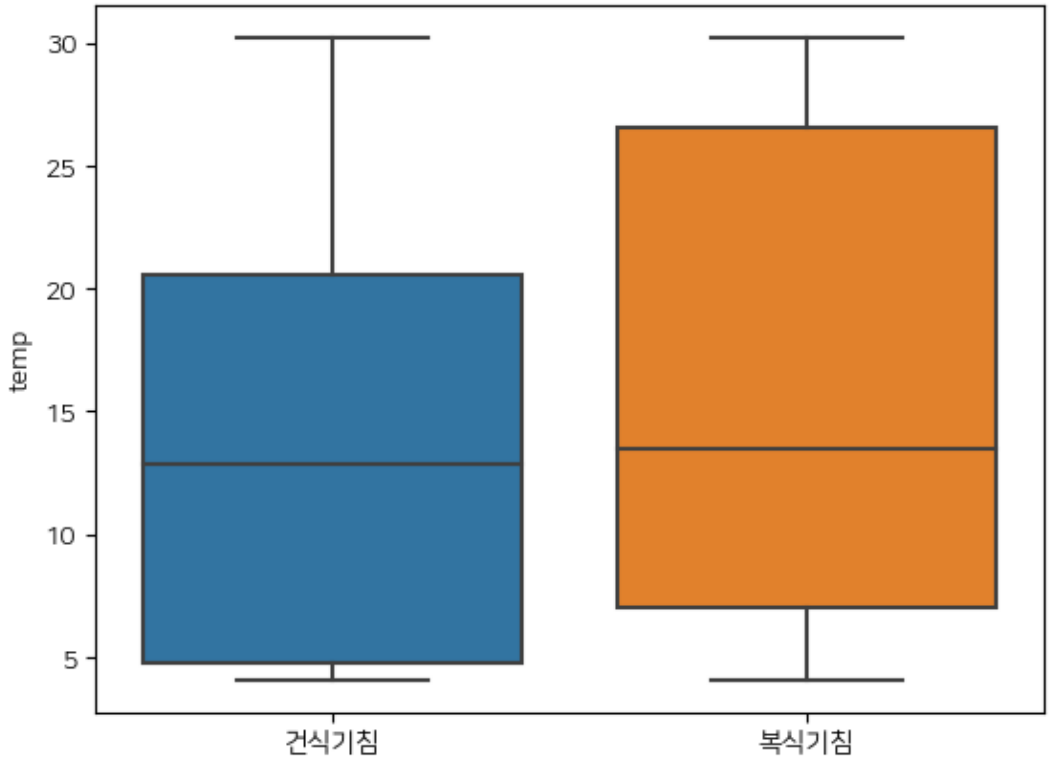
활용사례2. 양돈기침 유형분류 솔루션

스마트팜코리아 SMARTFARM KOREA

EDA

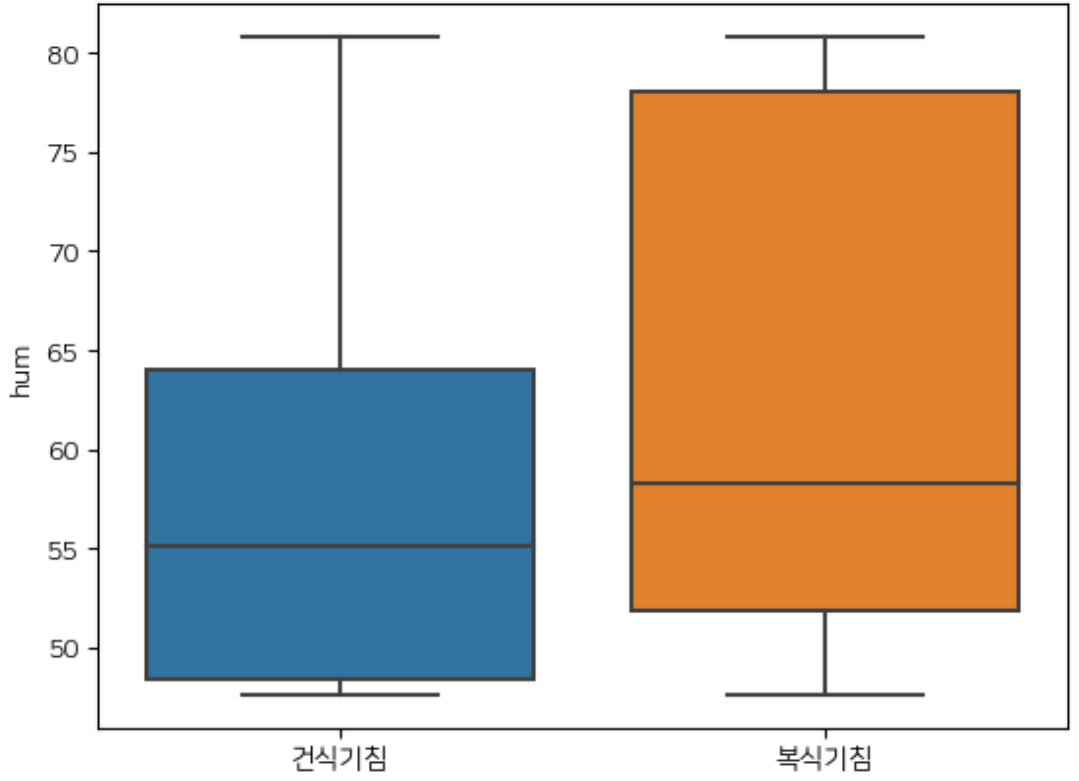
✓ 서울 최고기온

(*P-Value : 0.00000)



✓ 부산 평균습도

(*P-Value : 0.00000)



➡ 기후 관련한 두 feature는 기침유형 그룹에 따른 평균의 차이가 통계적으로 있음을 보여준다

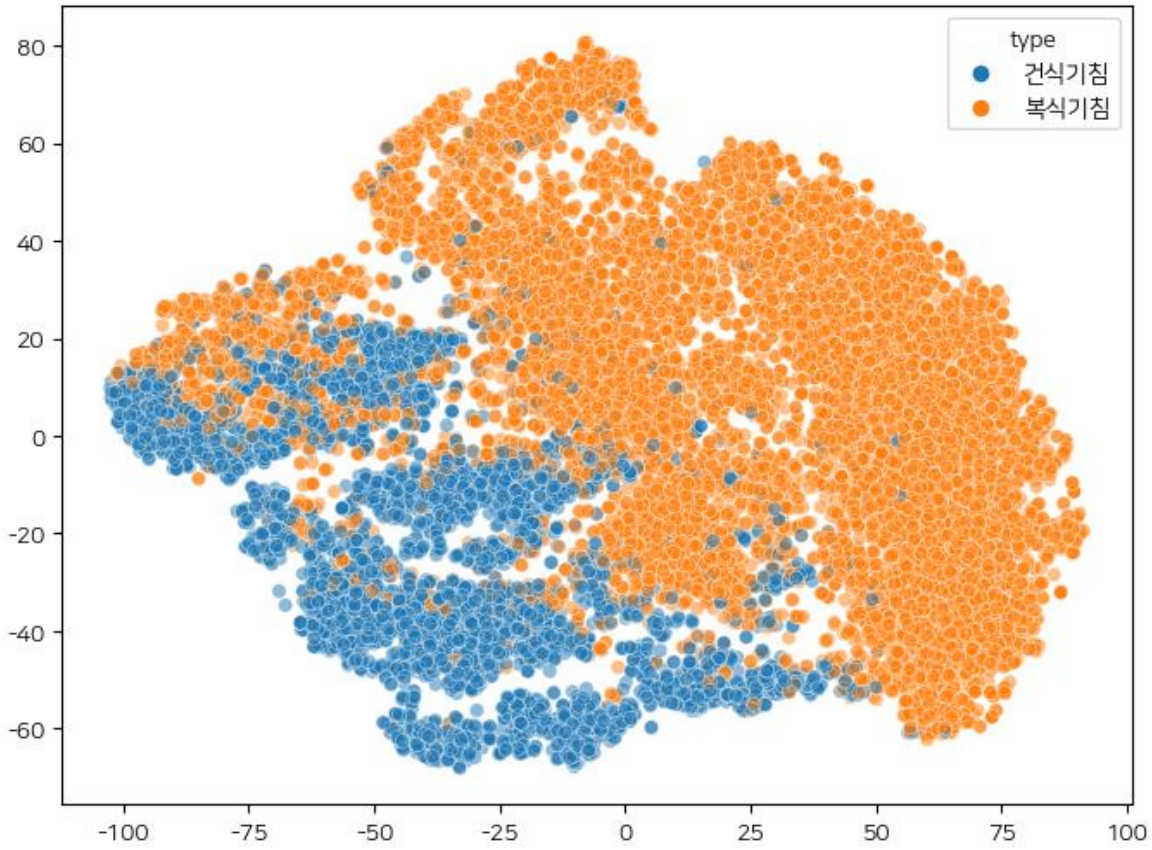
* 유효숫자는 소수점 아래 5자리

활용사례2. 양돈기침 유형분류 솔루션

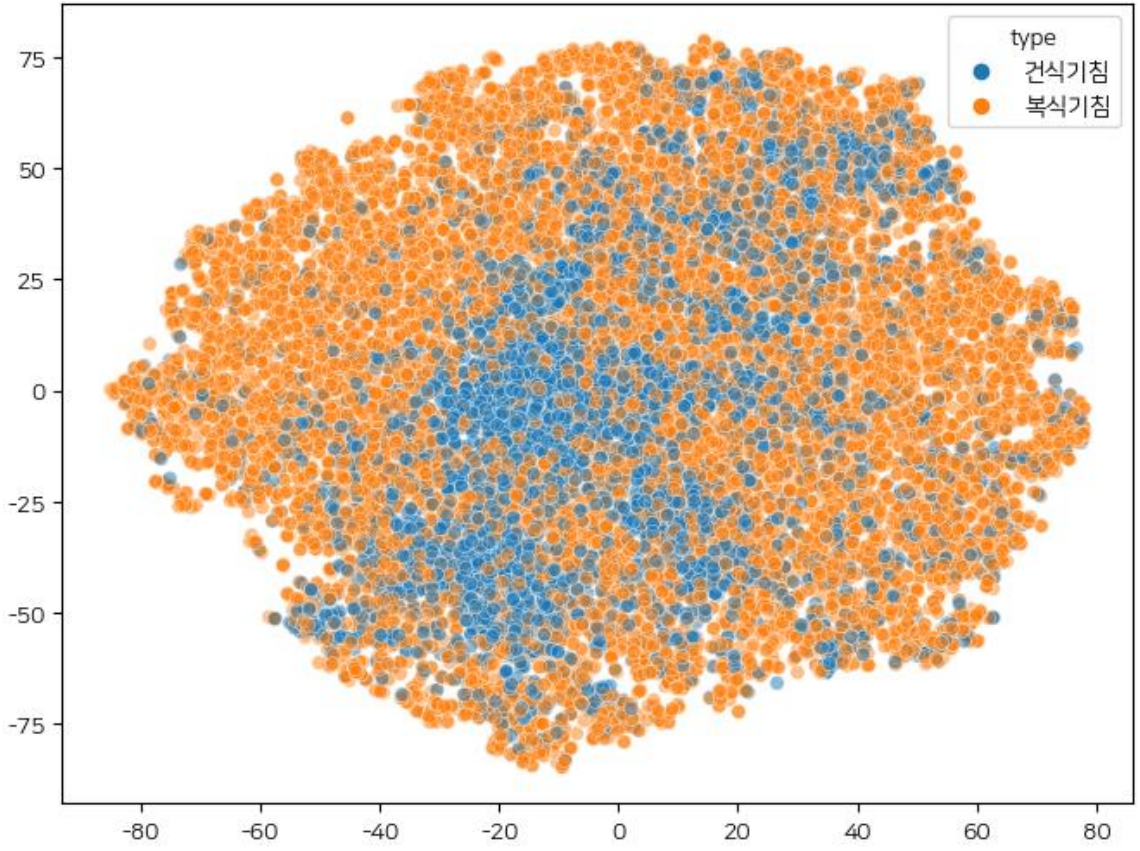
스마트팜코리아
SMARTFARM KOREA

EDA

✓ T-SNE Visualization on MFCC



✓ T-SNE Visualization on Chroma Frequencies



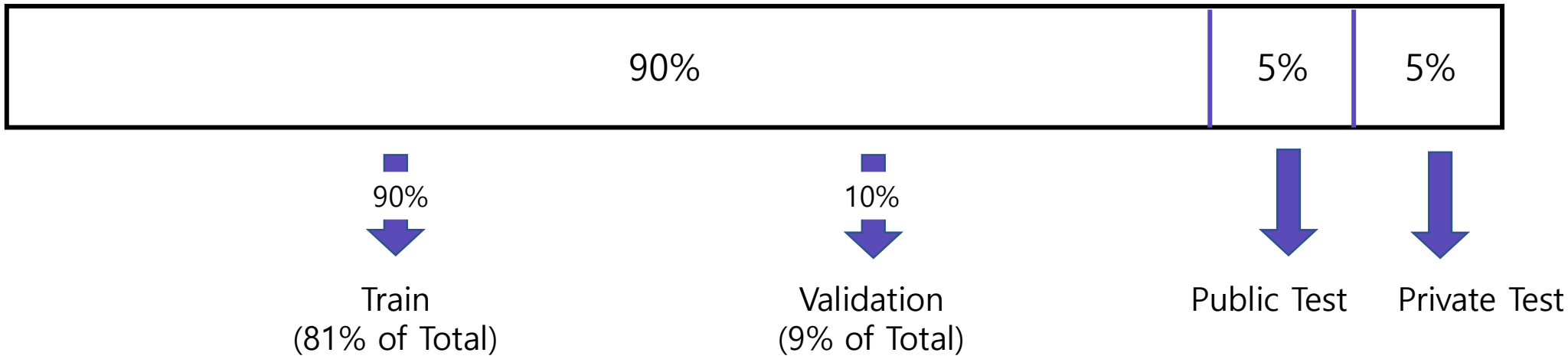
➡ 기침유형에 대해서는 Chroma Frequencies 보다 MFCC가 더 representation이 잘 되는 feature임을 보여준다

활용사례2. 양돈기침 유형분류 솔루션

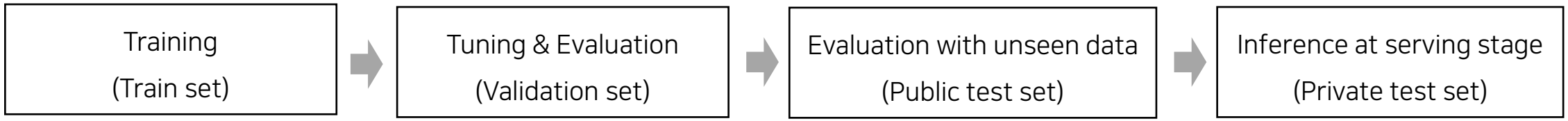
스마트팜코리아 SMARTFARM KOREA

Training & Evaluation System

✓ Dataset Split Strategy (Stratified sampling by Type)



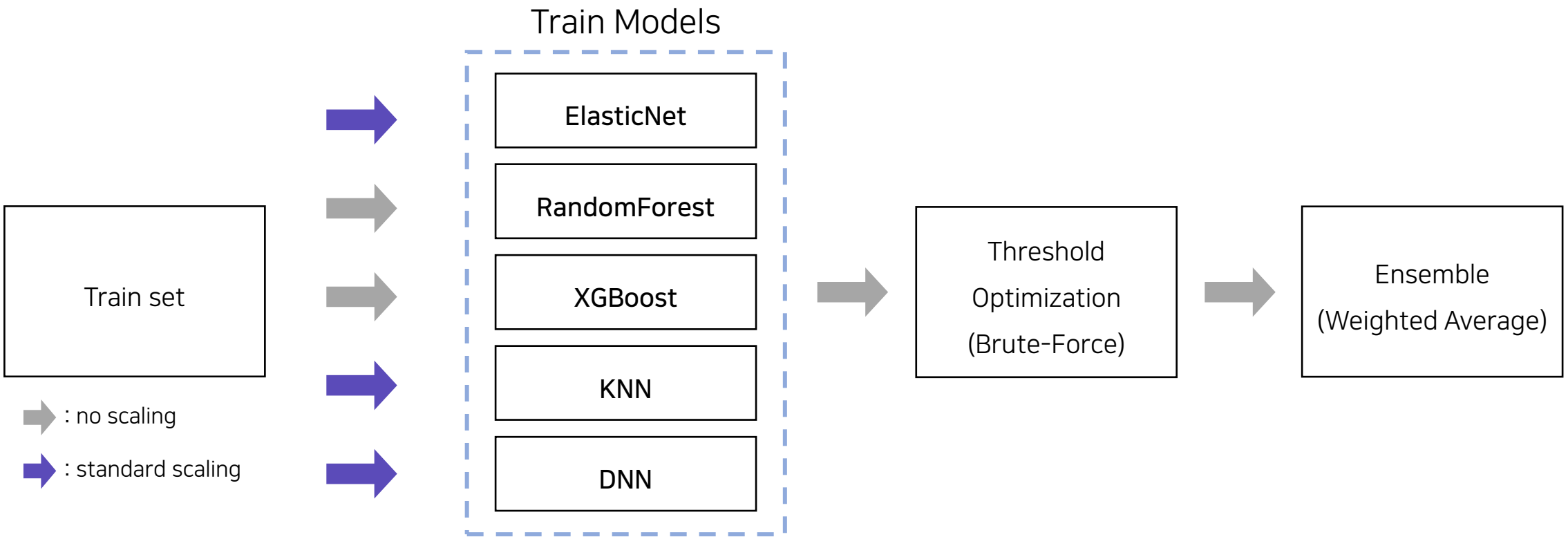
✓ Model Establishment Process



활용사례2. 양돈기침 유형분류 솔루션

스마트팜코리아
SMARTFARM KOREA

Solution Architecture



활용사례2. 양돈기침 유형분류 솔루션

스마트팜코리아 SMARTFARM KOREA

Result Summary

✓ Private data metric score

Models	Logloss	ROC AUC Score	Accuracy	F1
Logistic (Baseline)	0.1760	0.9785	0.9336	0.9239
ElasticNet	0.1784	0.9781	0.9386	0.9298
RandomForest	0.2522	0.9780	0.9286	0.9181
KNN	0.3011	0.9765	0.9248	0.9126
XGBoost	0.1090	0.9903	0.9612	0.9557
DNN	0.1547	0.9864	0.9536	0.9464
Ensemble	0.1674	0.9878	0.9574	0.9511

➡ Ensemble Weight

ElasticNet : 0.1

RandomFrorest : 0.2

KNN : 0.2

XGBoost : 0.25

DNN : 0.25

➡ Ensemble Threshold

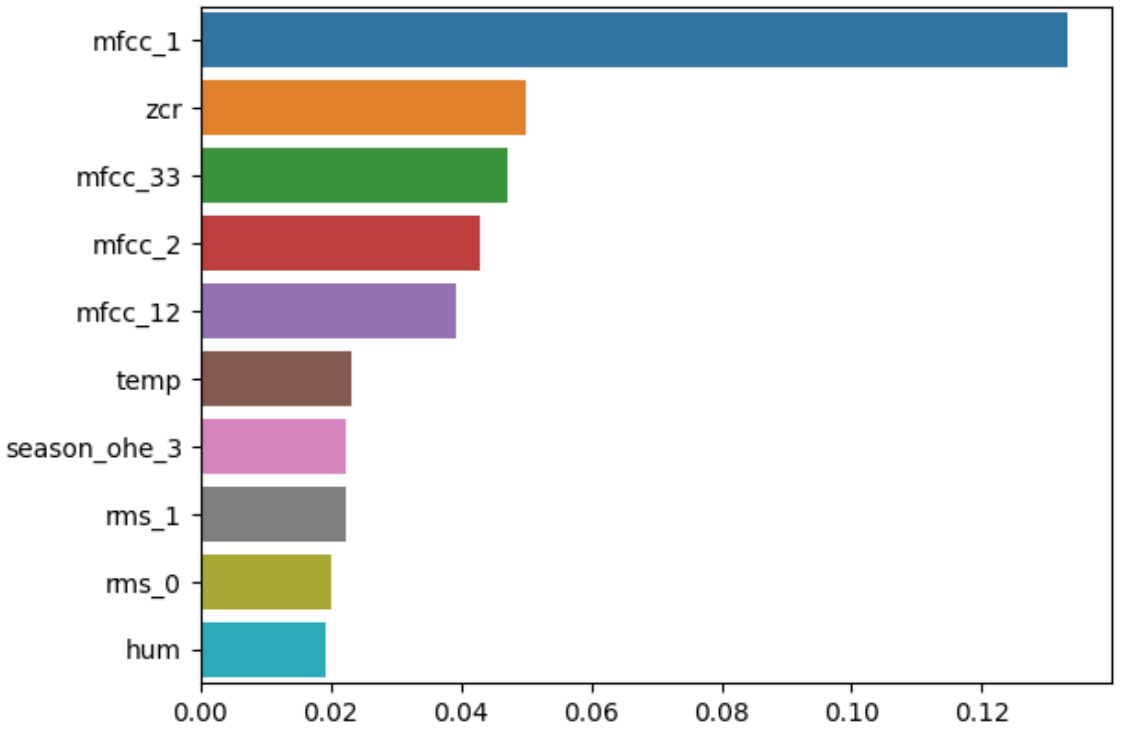
0.425 (on 복식기침)

활용사례2. 양돈기침 유형분류 솔루션

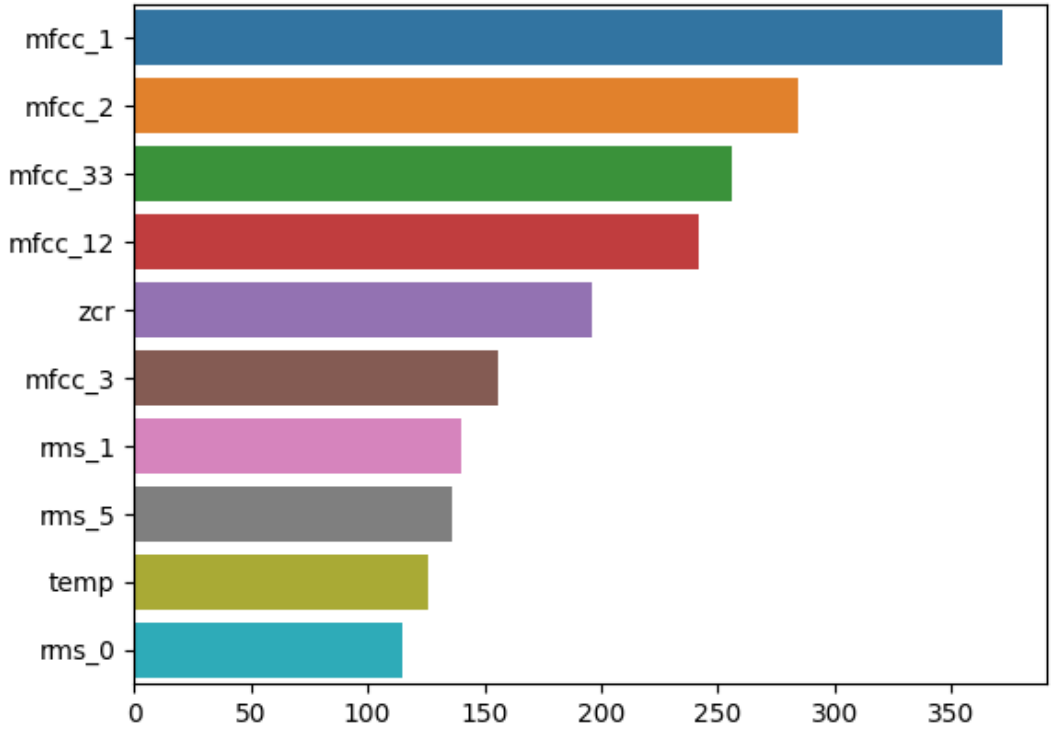
스마트팜코리아 SMARTFARM KOREA

Result Summary

✓ XGBoost Feature Importances



✓ RandomForest Feature Importances



두 모델의 Top5 변수 그룹이 모두 같음을 알 수 있다

Zero Crossing Rate의 평균과 MFCC vector 및 MFCC vector의 표준편차 feature가 유의미하게 작용했음을 알 수 있다

* mfcc_33 feature는 MFCC vector의 표준편차

목차

스마트팜코리아
SMARTFARM KOREA

정리 및 한계점

스마트팜코리아
SMARTFARM KOREA

✓ 정리

1. 관리가능한 사이즈의 농가 구성을 통해 모돈 및 자돈을 위한 최적의 환경을 만들어 주는 것이 무조건적으로 큰 사이즈의 구성을 가져가는 것보다 더 높은 효율을 보일 것이다.

2. 개발한 분류모델의 F1이 95%를 넘어서므로 머신을 통해 기침 유형을 분류하는 것이 더 효과적일 것으로 보인다. 이는 양돈농가 질병 대응 큰 도움을 줄 것이다.

✓ 한계점

1. 봄철 분만하는 자돈들의 경우 이유기까지 생존하는 수와 총산 대비 생존비율이 둘 다 높은 현상이 있는데, 본 데이터들을 통해서는 왜 이러한지 이유를 밝히기 어려웠다. 이에 추가적인 연구가 필요할 것으로 보인다.

2. 축산 퀄리티, 사료 종류, 모돈 신체 정보, 정상 양돈기침 등의 추가적인 데이터들이 제공된다면 더 자세한 분석과 더 좋은 모델 설계가 가능할 것으로 보인다.

Thank You

The End

양돈 # 이유자돈생존율 # 기침유형분류 # 음향데이터

스마트팜코리아
SMARTFARM KOREA

Appendix

스마트팜코리아 SMARTFARM KOREA

✓ References

양돈에 있어서 MSY, PSY, 분만율, 모돈회전율 등 계산방법 | 농사로 (nongsaro.go.kr)

http://www.liveinfo.kr/news/article.html?no=24072

가을철 돼지에 많이 발생하는 10가지 병 | 농사로 (nongsaro.go.kr)

[2022 신년특집-대담] "내가 생산한 돼지 품질에 관심을" "빨리빨리 사육 관행서 벗어나야" - 양돈타임스 (pigtimes.co.kr)

[누알사자] 작게 태어난 자돈들....도태해야 하는 이유와 기준은 이것! (pigpeople.net)

✓ Implementation

MyPortfolio/Competition/ETC/스마트팜코리아 at main · Cafelatte1/MyPortfolio (github.com)

✓ Contact

E-Mail : flash659@gmail.com

✓ 서울 최고기온과 부산 평균습도를 온도와 습도 feature의 대푯값으로 사용한 이유

타 competition 에서 연구한 부분으로 서울 최고기온과 부산 평균습도를 대푯값으로 사용했을 때 모델 성능이 조금 향상 되었습니다

그러나 본 task에도 적절한 방법인지는 추후에 다시 연구되어야 할 것입니다

또한 모델 input경우 최근 3년 (19,20,21) 평균의 온도 및 습도 값을 사용하였습니다. 이는 원활한 추론 시스템 구축을 할 수 있도록 합니다