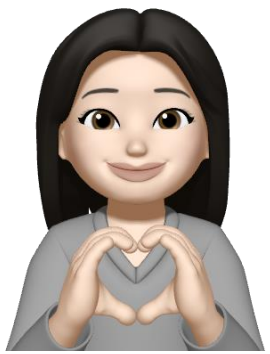
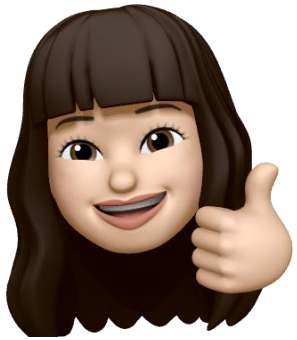
등록 상품 자동검수(Filtering) 모델

KITA DIMA 3기 1조
김도연B 김인영 김희진 유진우 윤주영



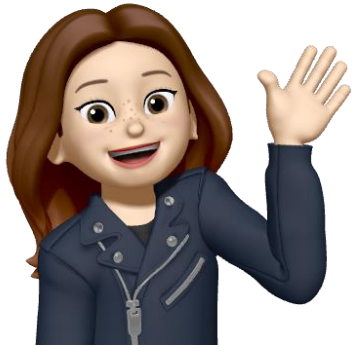
김도연B



김인영

김희진

유진우



윤주영

EC21 Global B2B Marketplace

국내 1위 글로벌 온라인 B2B 마켓 플레이스

온라인 B2B 마켓플레이스, 글로벌 시장 조사, 온/오프라인 마케팅,

무역 교육 프로그램 및 국제 협력

셀러의 상품을 사이트에 노출하여 바이어를 유입

등록된 바이어에게 등록된 카테고리나 관심정보와 매칭되는 셀러 정보를 제공

250만 이상 회원

700만 이상 제품

월별 약 350만 방문자

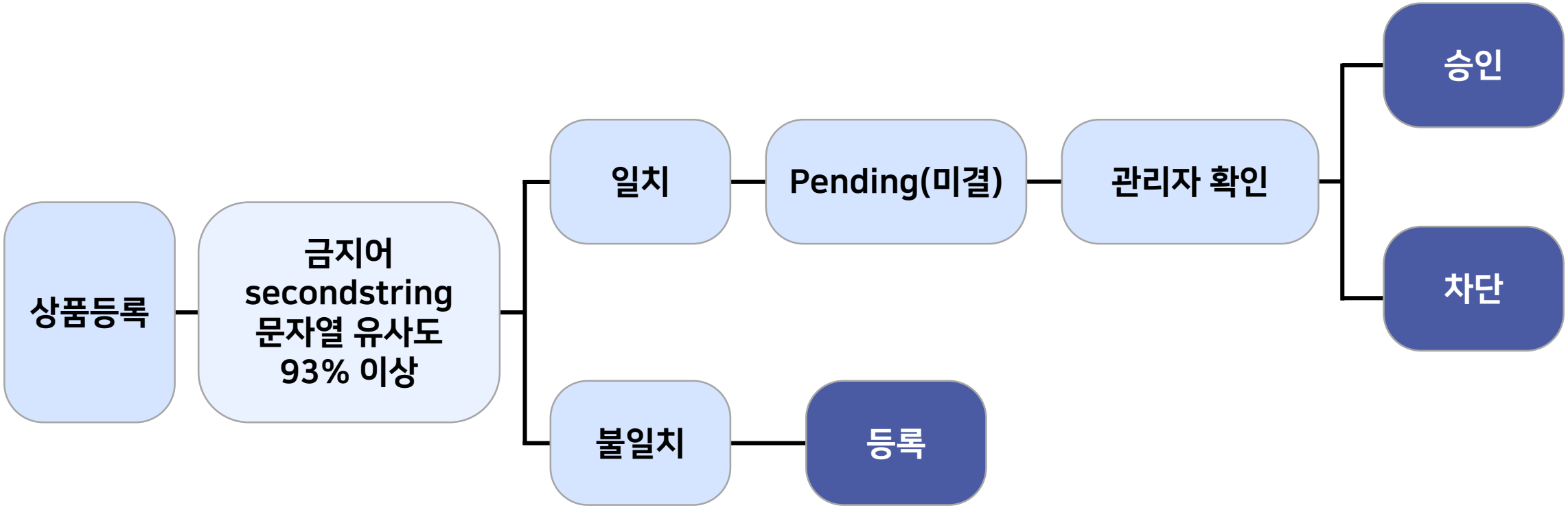
온라인 의약품 불법판매

2022.09.23.

최근 5년
온라인
불법 의약품 광고·판매 적발
13만4440건
매해 2만5000건 이상 적발

https://health.chosun.com/site/data/html_dir/2022/09/23/2022092300974.html

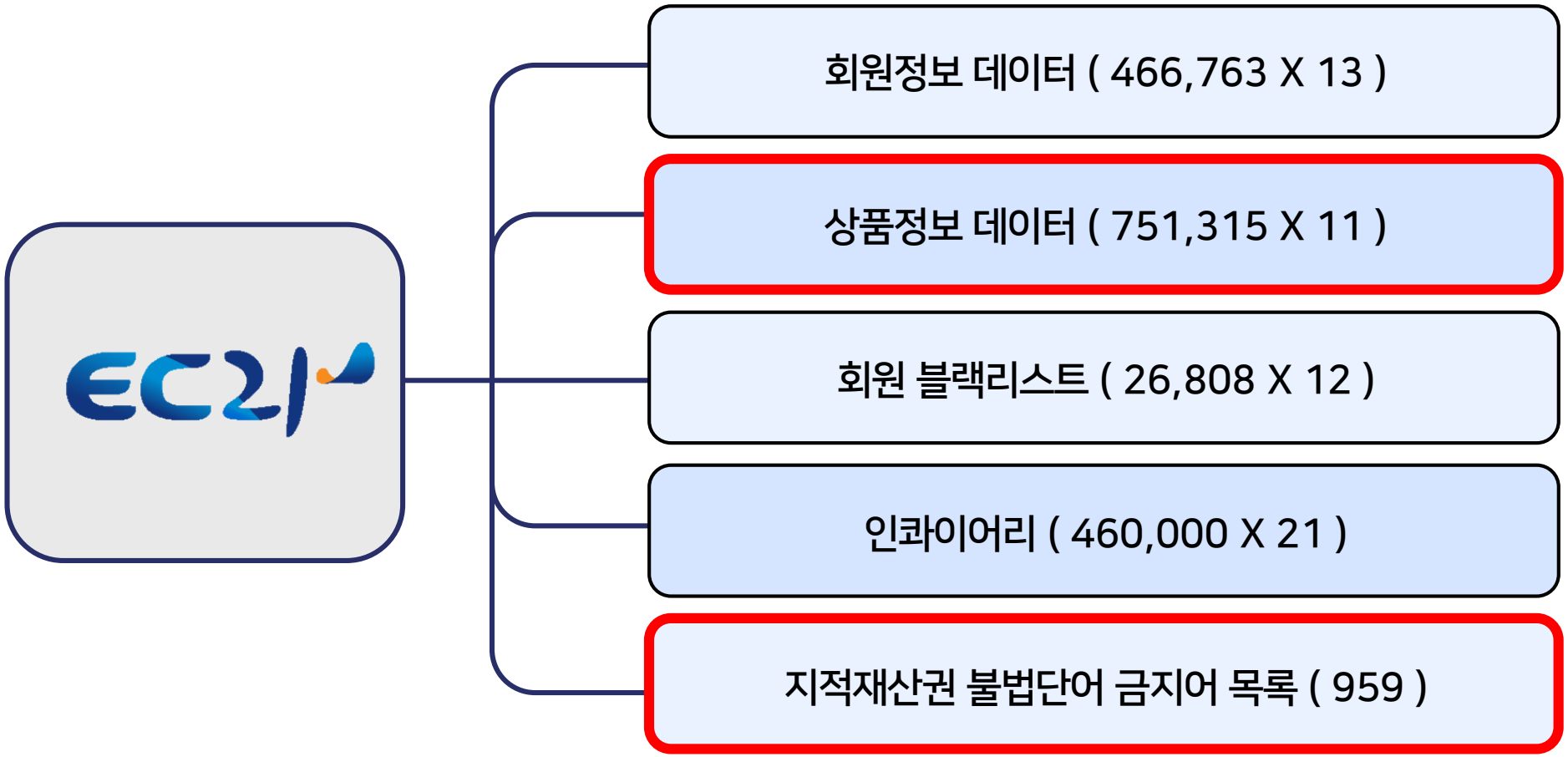
EC21 Global B2B Marketplace



EC21 Global B2B Marketplace

문제점

- 자동 검수를 피해간 단어들로 대량 상품등록
- 금지어 목록과의 단어 유사도만 확인하는 한계점
- 관리자의 경험만으로 상품 판단



등록 상품 자동 검수 (Filtering) 모델

지적재산권 불법단어 금지어 목록 (959)

KEYWORD	사유
adidas	IPR 침해 (IPR Infringement)
I phone	IPR 침해 (IPR Infringement)
Salvinorin	금지약품 (Prohibited Drugs)
Trenbolone Acetate	금지약품 (Prohibited Drugs)
Black Money	금지품목 (Prohibited Items)
gold bar	금지품목 (Prohibited Items)
sex pill	성인용품(Explicit Adult)

.
.
.

> 상품정보 DATA

> 2018 - 2023.10

> 751,315개

MEMBER_ID	GCATALOG_ID	CATALOG_ID	CATALOG_NM	CATEGORYM_ID	KEYWORD
회원ID	상품그룹ID	상품ID	상품명	상품카테고리ID	상품키워드

DISPLAY	CATALOG_DESC	INPUT_DT	UPDATE_DT	REMOTEIP	SHOW_CHK
상품노출여부	상품설명	등록일시	수정일시	등록시IP주소	상품상태 (승인/거부/대기)

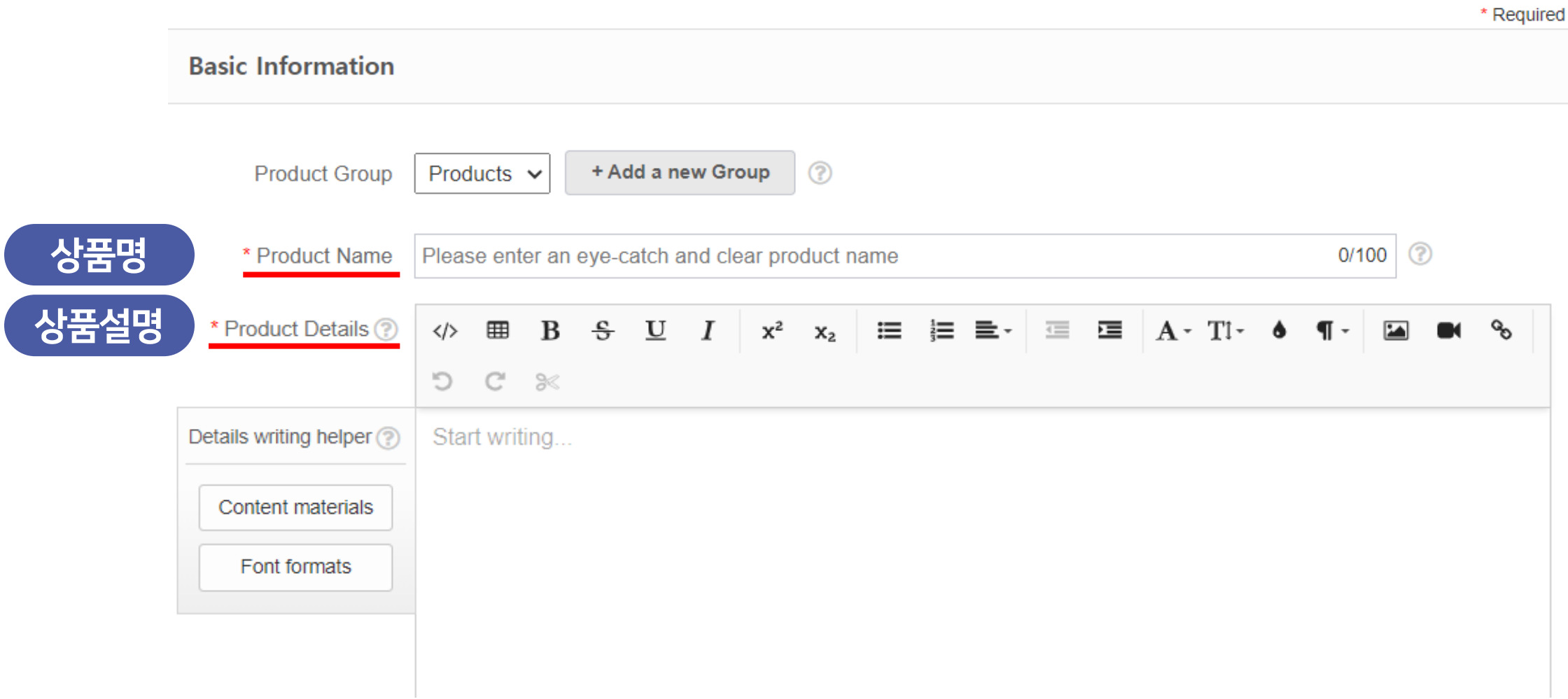
컬럼추가 JUDGE : 상품 유형 표시 (이상상품 : 0 / 정상상품 : 1)

독립변수 선정 과정

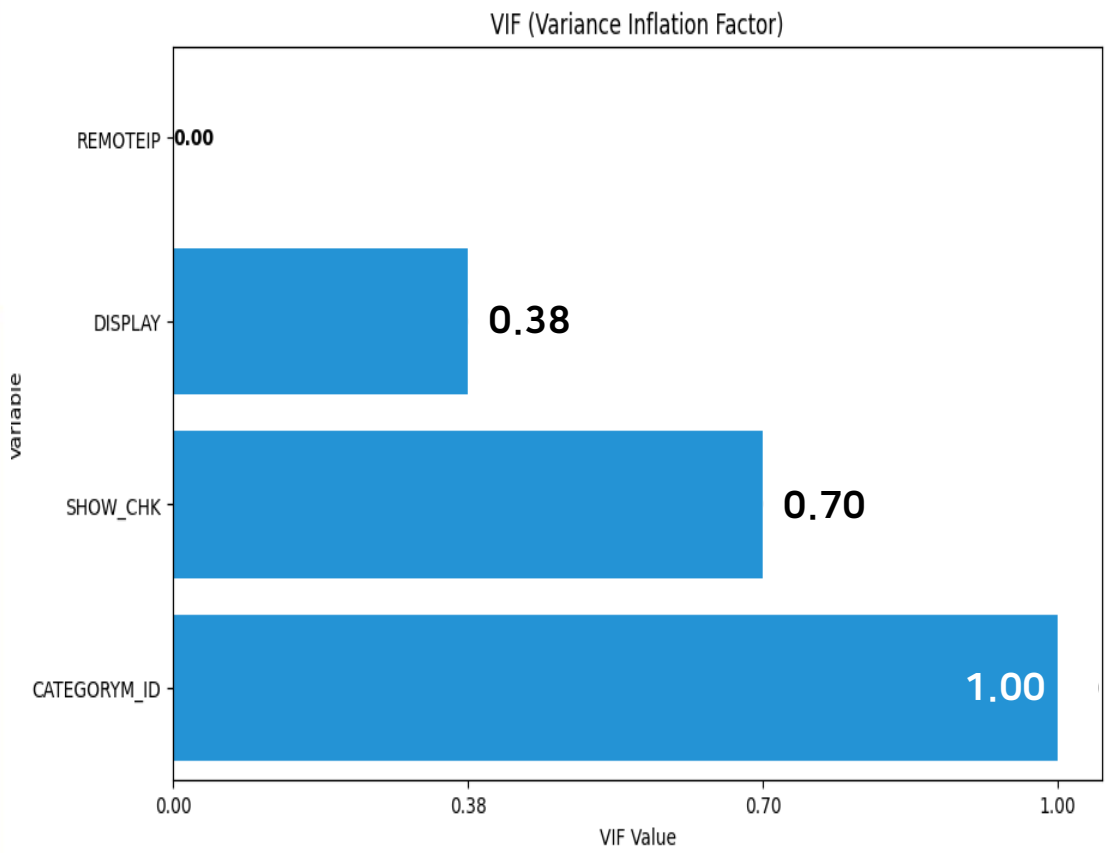
> VIF

> Heatmap

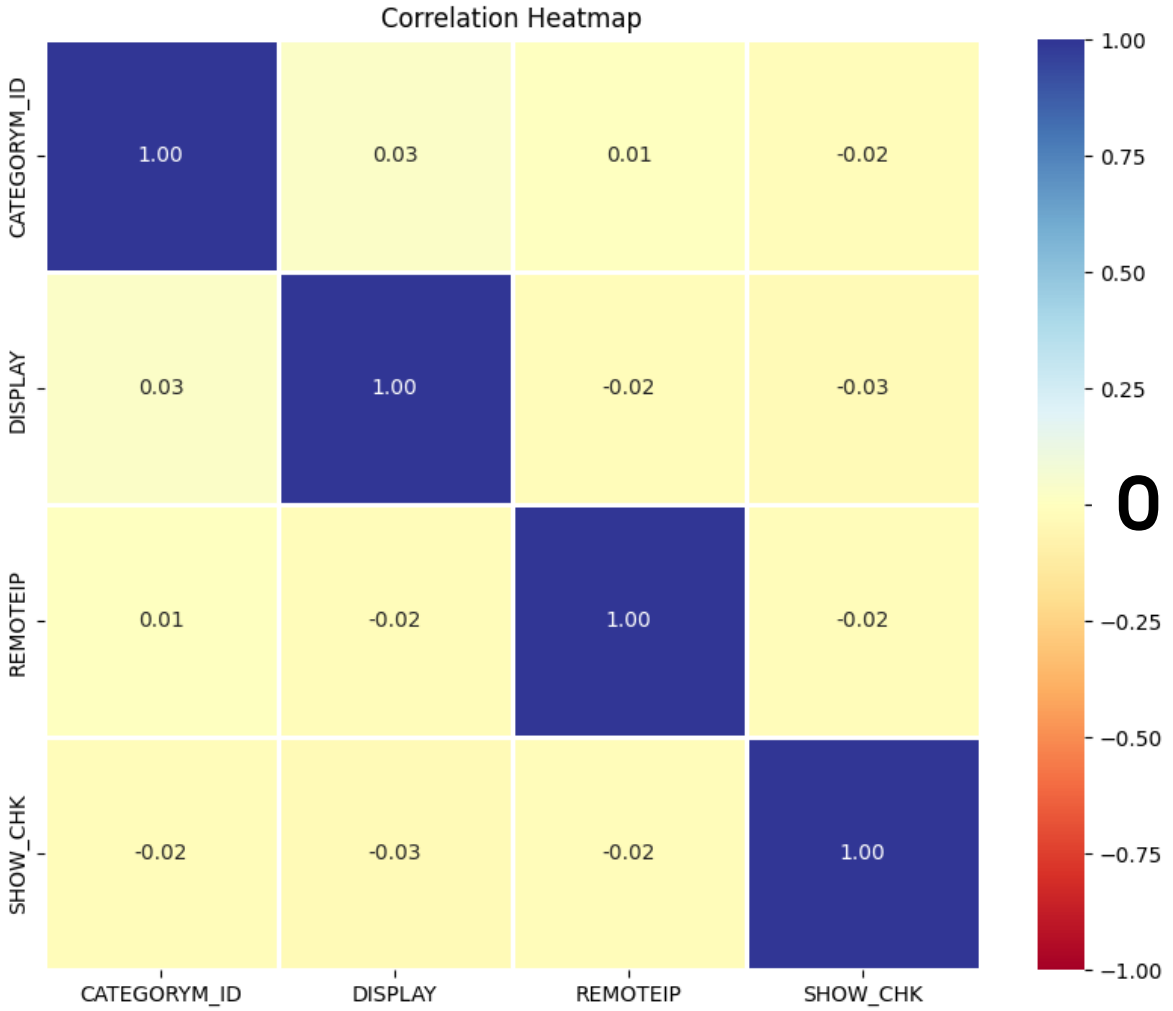
EC21 상품등록 페이지



> VIF 독립 변수 간의 상관 관계로 인해 회귀 분석 결과의 불안정성을 측정



> Heatmap 변수 간의 상관 관계를 시각적으로 표현 / 변수 간의 강도를 보여줌



독립변수 컬럼 선정

MEMBER_ID	GCATALOG_ID	CATALOG_ID	CATALOG_NM	CATEGORYM_ID	KEYWORD
회원ID	상품그룹ID	상품ID	상품명	상품카테고리ID	상품키워드

DISPLAY	CATALOG_DESC	INPUT_DT	UPDATE_DT	REMOTEIP	SHOW_CHK
상품노출여부	상품설명	등록일시	수정일시	등록시IP주소	상품상태 (승인/거부/대기)

컬럼추가 JUDGE : 상품 유형 표시 (이상상품 : 0 / 정상상품 : 1)

데이터 전처리

자연어 처리 모델을 적용하기 전 데이터를 정제하여
불필요한 정보를 배제하고 핵심적인 의미를 정확하게 분석하기 위함

데이터 전처리

① 중국어 제거

	CATALOG_NM	CATEGORYM_ID	DISPLAY	CATALOG_DESC
1	boc 4 4 bromo phenylamino piperidinecas 4439...	212815	1	1 名称和标识符1 1 名称1 叔丁氧羰基 4 （4 溴苯基氨基）哌啶1 2 同义词1 哌啶...
2	high purity n tert butoxycarbonyl 4 piperidone...	212815	1	1 名称和标识符1 1 名称n 叔丁氧羰基 4 哌啶酮1 2 同义词1 叔丁氧羰基 4 哌啶...

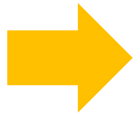
상품 설명에 중국어가 포함된 레코드의 개수 : 10,358

EC21 상품등록시 영어를 권장함

데이터 전처리

② HTML 태그 제거

CATALOG_DESC
Black tea ctc grade: pf packing...
<p class="MsoNormal" style="background: white;...



CATALOG_DESC
Black tea ctcgrade: pfpacking 65kg/bagqt 26 to...
\nPOLYDAIRY cow-mats make life of Animals mor...

※ 상품설명 (‘CATALOG_DESC’) 결측 치 존재 시, 상품명(‘CATALOG_NM’)으로 대체

데이터 전처리

③ 불완전 태그, 개행문자 제거

불완전 태그

<p>, <th>

CATALOG_DESC
www.cropgroupcn.comContact information : <p data-mce-style="font-family: Arial, Helv

CATALOG_DESC
www.cropgroupcn.comContact information :

개행 문자

₩r / ₩n / ₩₩

CATALOG_DESC
\nDescription:\nInnotox is a South Korean prod...

CATALOG_DESC
Description:Innotox is a South Korean prod...

※ 상품설명 ('CATALOG_DESC') 결측 치 존재 시, 상품명('CATALOG_NM')으로 대체

데이터 전처리

④ 소문자 변경, 구두점 제거

'CONTACT US TO PLACEORDERWhats-App:₩xa0+1 (661)-429-2164₩xa0₩xa0SHIPPING AND DELIVERY:* Delivery takes just 2₩xa0to 3₩xa0working days, nocustom troubles because the company takes careof all procedures.* We also offer wholesale and retail purchase, withgood discounts the more you buy, the better yourdiscount.* Tracking number is issued immediately aftershipment is done..'

처리 후

'contact us to placeorderwhats app xa0 1 661 429 2164 xa0 xa0shipping and delivery delivery takes just 2 xa0to 3 xa0working days nocustom troubles because the company takes careof all procedures we also offer wholesale and retail purchase withgood discounts the more you buy the better yourdiscount tracking number is issued immediately aftershipment is done '

데이터 전처리

④ 불용어 제거, 단어 토큰화, 어근추출

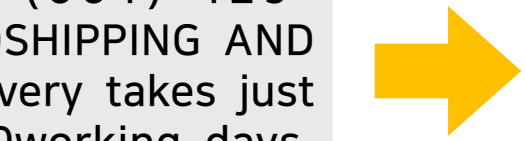
'contact us to placeorderwhats app xa0 1 661 429 2164 xa0 xa0shipping and delivery delivery takes just 2 xa0to 3 xa0working days nocustom troubles because the company takes careof all procedures we also offer wholesale and retail purchase withgood discounts the more you buy the better yourdiscount tracking number is issued immediately aftershipment is done '

처리 후

'contact', ‘us', 'placeorderwhats', 'app', '1', '661', '429', '2164', 'shipping', 'delivery', 'delivery', 'take', '2', '3', 'working', 'day', 'nocustom', 'trouble', 'company', 'take', 'careof', 'procedure', 'also', 'offer', 'wholesale', 'retail', 'purchase', 'withgood', 'discount', 'buy', 'better', 'yourdiscount', 'tracking', 'number', 'issued', 'immediately', 'aftershipment', 'done'

> 전처리 전 상품정보 DATA

CATALOG_NM 상품명	CeraVing Anti Aging Gel Serum for Face To Boost Hydration and Hyaluronic Acid
CATALOG_DESC 상품설명	CONTACT US PLACEORDERWhats-App:₩xa0+1 (661)-429-2164₩xa0₩xa0SHIPPING AND DELIVERY:* Delivery takes just 2₩xa0to 3₩xa0working days, nocustom troubles because the company takes careof all procedures.* We also offer wholesale and retail purchase, withgood discounts the more you buy, the better yourdiscount.* Tracking number is issued immediately aftershipment is done..



> 전처리 후 상품정보 DATA

CATALOG_NM 상품명	['ceraving', 'anti', 'aging', 'gel', 'serum', 'face', 'boost', 'hydration', 'hyaluronic', 'acid']
CATALOG_DESC 상품설명	['contact', 'us', 'placeorderwhats', 'app', '1', '661', '429', '2164', 'shipping', 'delivery', 'delivery', 'take', '2', '3', 'working', 'day', 'nocustom', 'trouble', 'company', 'take', 'careof', 'procedure', 'also', 'offer', 'wholesale', 'retail', 'purchase', 'withgood', 'discount', 'buy', 'better', 'yourdiscount', 'tracking', 'number', 'issued', 'immediately', 'aftershipment', 'done']

모델 선정 과정

> Machine Learning

> Deep Learning

Machine Learning

> Decision Tree Classifier

> Logistic Regression

> Random forest classifier

> Support Vector Machine

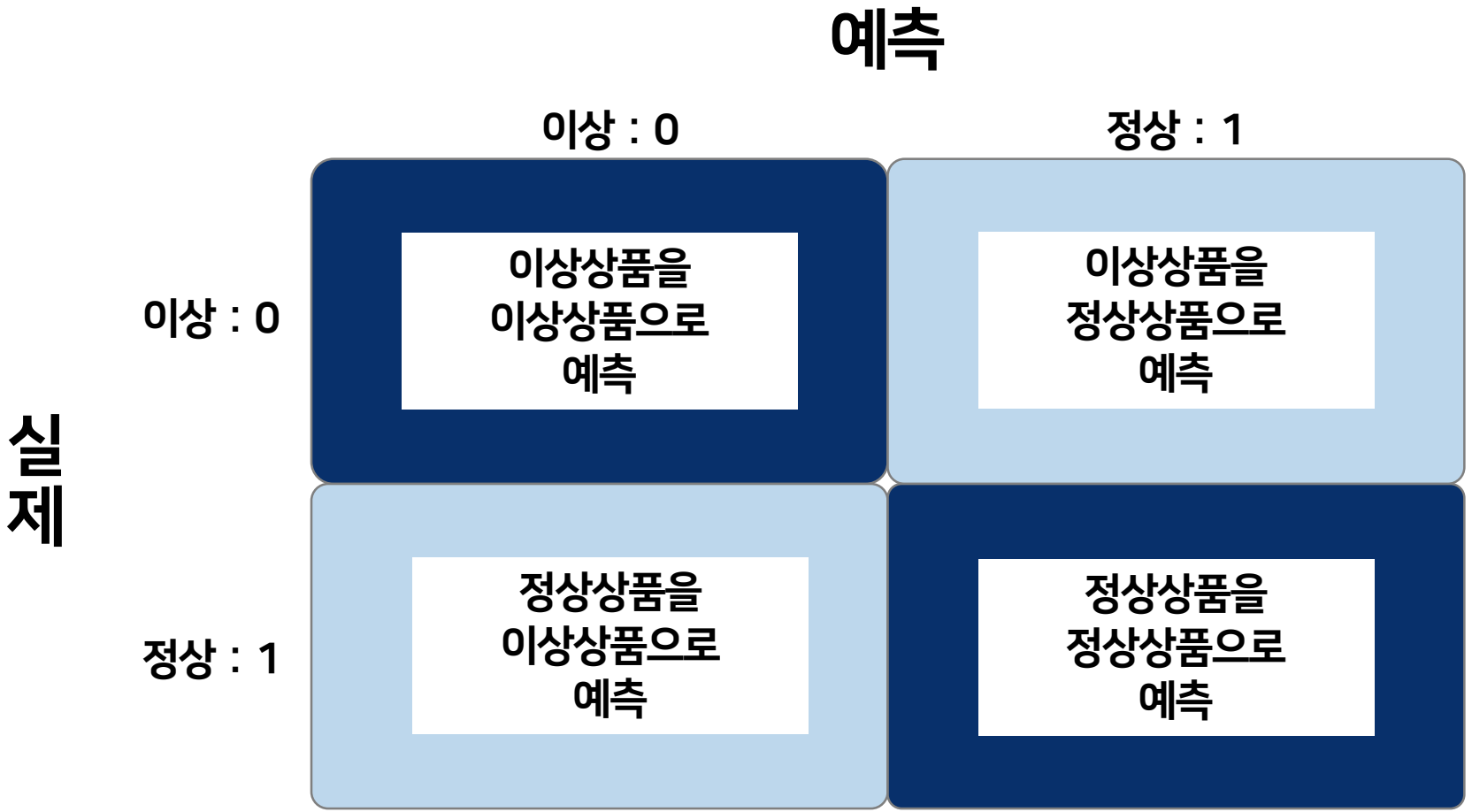
Deep Learning

> Long Short-Term Memory

테스트 데이터 (120개)

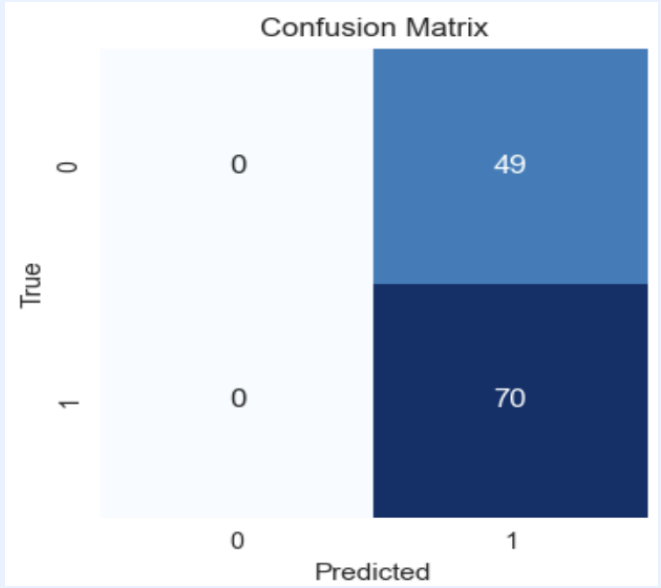
모델

Confusion Matrix



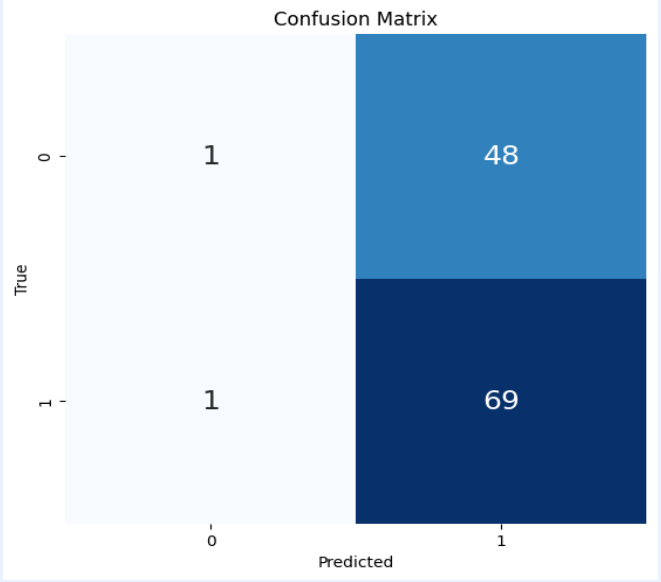
Machine Learning

> Decision Tree Classifier



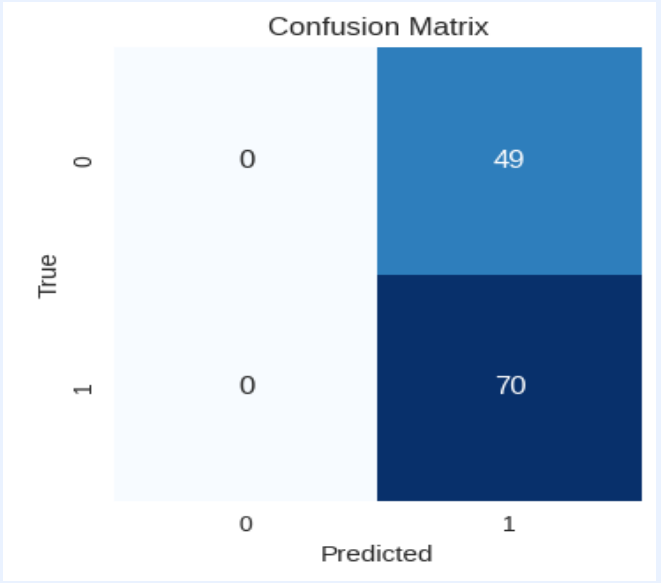
정확도 : 0.59

> Random forest classifier



정확도 : 0.58

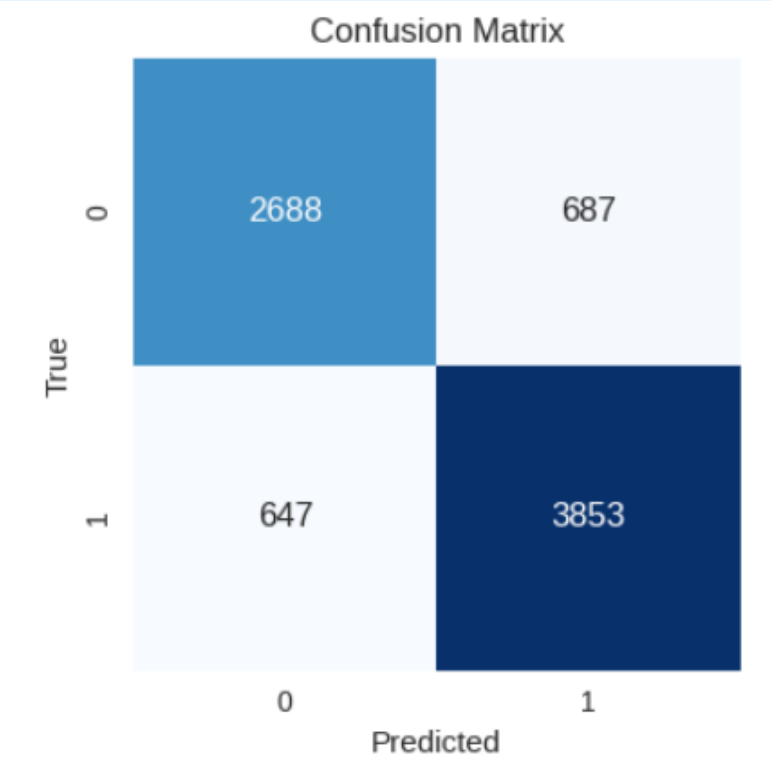
> Logistic Regression



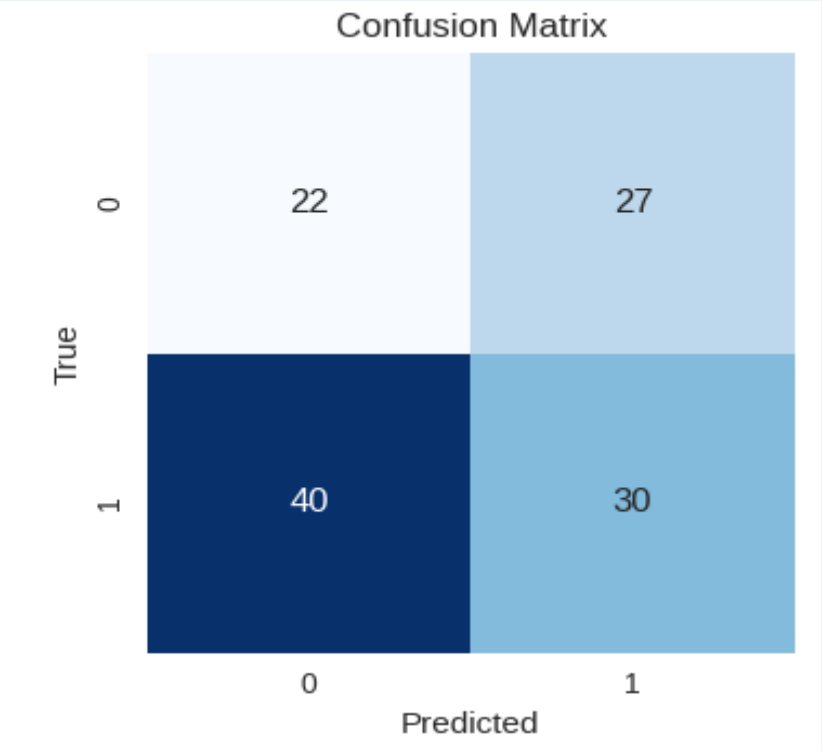
정확도 : 0.59

Machine Learning

> Support Vector Machine



학습 데이터 정확도 : 0.83



평가 데이터 정확도 : 0.44

> Machine Learning

어둠 → 어두워 / 어둡다 / 어두운

> Deep Learning

“스위치가 어디있지?”

“앞이 안보여”

“정전인가?"

채택모델

> LSTM

Long short-term memory

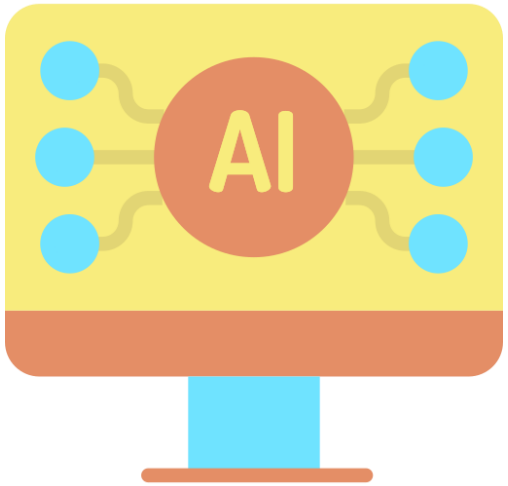
단기 기억이 아닌 장기기억으로 글을 읽고 해석

문장 속에 있는 단어와 단어 간의 관계를 판단

ex) 학술 논문, 친구간 문자 메시지, 상품의 설명, 어떤 형태와 종류의 텍스트

모델 학습 과정

등록된 상품



모델 학습

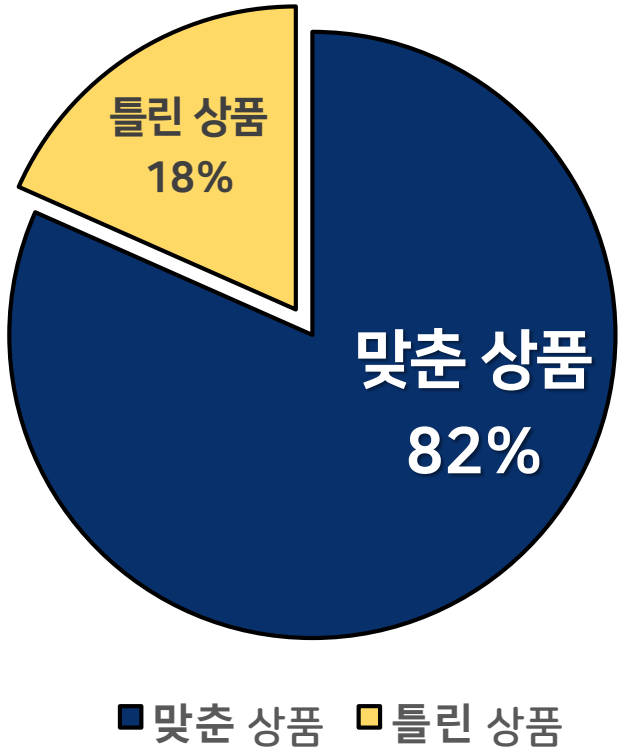
이상상품과 정상상품을
동일한 비율로 학습

예측

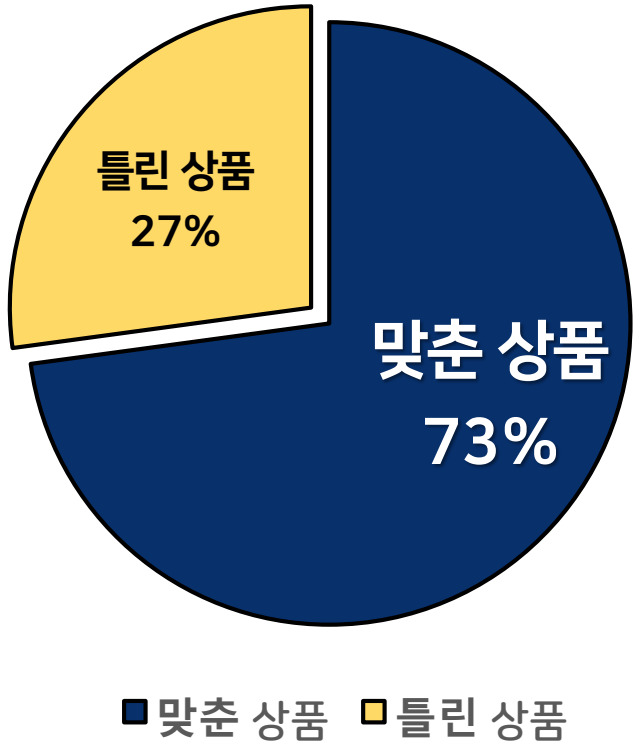
Long short-term memory

LSTM 결과

전체 데이터 개수 : 120
예측 성공한 데이터 개수 : 92
맞춘 이상 상품 : 40　　맞춘 정상 상품 : 52
틀린 이상 상품 : 9　　틀린 정상 상품 : 19

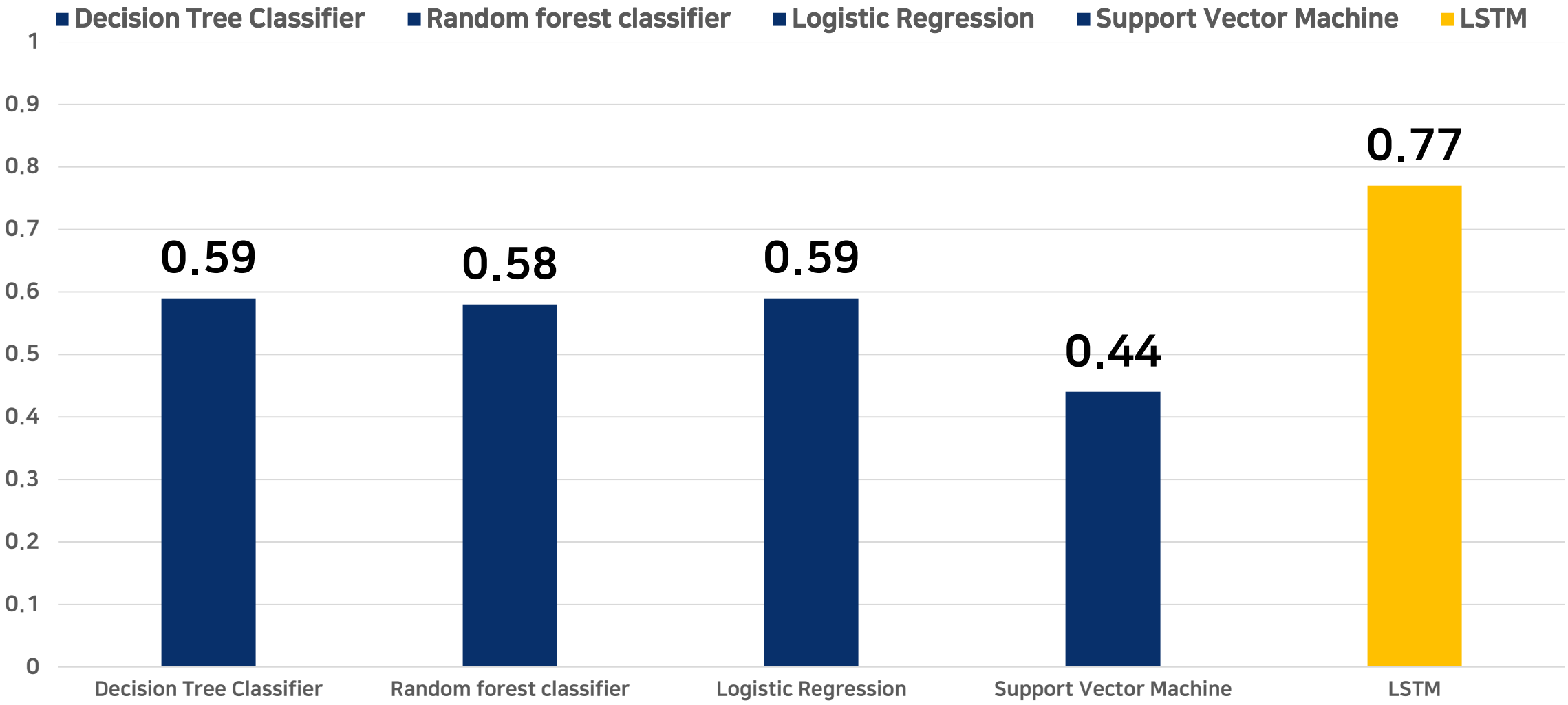


이상상품



정상상품

평가 데이터 정확도 (모델)


Decision Tree Classifier
Random forest classifier
Logistic Regression
Support Vector Machine
LSTM
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.59
0.58
0.59
0.44
0.77
Decision Tree Classifier
Random forest classifier
Logistic Regression
Support Vector Machine
LSTM

Long short-term memory

LSTM 결과

	CATALOG_NM	CATALOG_DESC	judge	y_pred	predict_proba
91	shoes slippers loafers slides men platform wom...	Product descriptions from the supplier₩nOvervi...	1	1	0.984684
84	Free sample price 1M x 1M fire resistant rescu...	Product descriptions from the supplier₩nOvervi...	1	1	0.984486
16	Nordic Ins Style Pangpang Fat Mug Creative Nov...	Brand₩nDongyang Porcelain, OEM, ODM₩nMaterial₩...	1	1	0.984400
58	Calendar 2023 - Vertical 11x17 2023 Wall Calen...	Size₩tLarge₩nYear₩t2023₩nBrand₩tKarto₩nMateria...	1	1	0.984163
80	Factory Wholesale 13 Gauge Nylon PU Coated Glo...	Product descriptions from the supplier₩nOvervi...	1	1	0.984149
...	...	...	...	...	...
102	Marvel Hasbro Legends Series Iron Man Mark 46,...	marvel irnoman avengers disney	0	0	0.127014
26	Apple iPhone 11, 64GB, Black - Unlocked (Renewed)	This phone is unlocked and compatible with any...	0	0	0.078926
13	Honey and rose Herbal Cigarettes - Tobacco and...	A Good Helper To Quit Smoking - Quitting smoki...	0	0	0.055820
66	Sony PS5 Playstation 5 Console Disc Version - ...	Sony PS5 Playstation 5 Console Disc Version - ...	0	0	0.049630
22	Humo's' Herbal Cigarettes - Tobacco & Nicotine...	🌿 **ALL-NATURAL**: Humo's tobacco-free cigarettes a...	0	0	0.048055

Long short-term memory

LSTM 결과

judge	y_pred	predict_proba
1	1	0.984684
...	...	...
0	0	0.048055

judge	y_pred	predict_proba
1	1	0.984684
1	1	0.984486
1	1	0.984400
1	1	0.984163
1	1	0.984149
...	...	...
0	0	0.127014
0	0	0.078926
0	0	0.055820
0	0	0.049630
0	0	0.048055

> Fuzzywuzzy

‘레벤슈타인거리’를 통해 금지어 단어유사도 확인

시퀀스(문자열, 리스트 등) 간의 유사성을 비교하는 데 사용

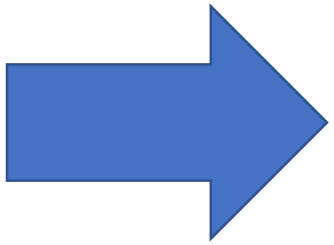
일치하는 부분을 찾고 유사성을 비교하여 유사도를 측정

Fuzzywuzzy

jordans

jor-dan

jordann



Jordan

Fuzzywuzzy

	상품명+상품설명	예측	금지어단어 (카테고리)
8	nike 2020 2021 inter milan fourth football soccer ...	0	Nike (IPR 침해)
39	nike running air shoe elevate running experience premium running shoe ...	0	Nike (IPR 침해)
43	jordan brooklyn fleece jordan style top bottom hoodie ...	0	jordan (IPR 침해)
45	adidas backpack travel style spacious ...	0	adidas (IPR 침해)
47	sony wireless mouse increase productivity precision wireless ...	0	유사단어 없음 확인 필요

결론

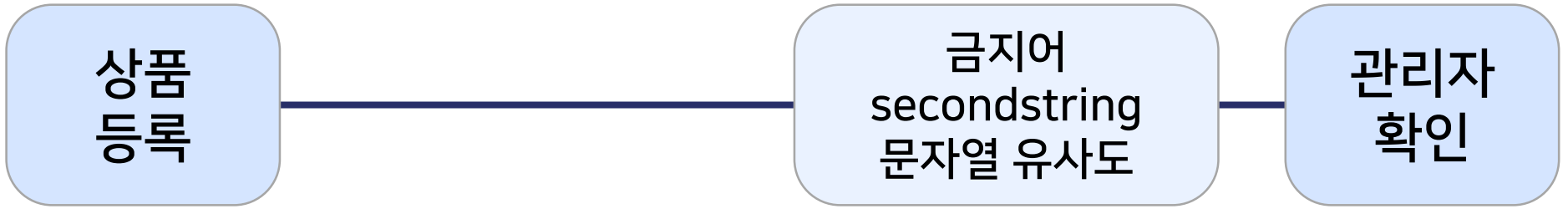
기존

Item Name	Item Desc	Actions
Iphons 14 pro max 256gb deep purple	brand new original complete accessory warranty	정상상품으로 분류됨

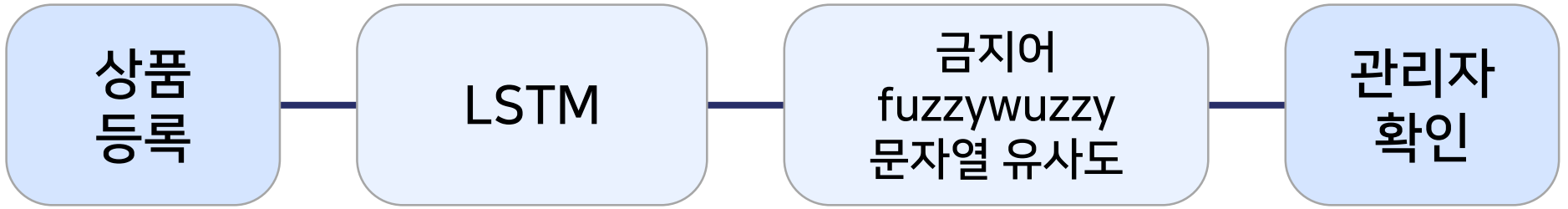
대안

Item Name	Item Desc	Actions
Iphons 14 pro max 256gb deep purple	brand new original complete accessory warranty	이상검출도 : 91% 이상단어검출: iphone (IPR infringement) 금지어유사도: 93% Approve Reject

> 기존


상품
등록
금지어
secondstring
문자열 유사도
관리자
확인

> 대안


상품
등록
LSTM
금지어
fuzzywuzzy
문자열 유사도
관리자
확인

EC21 Global B2B Marketplace

기존 문제점

- 자동 검수를 피해간 단어들로 대량 상품등록
- 금지어 목록과의 단어 유사도만 확인하는 한계점
- 관리자의 경험만으로 상품 판단

자동 검수 모델 활용 방안

이상 검출도 + 금지어 유사도

Threshold > n

1. 자동화

자동 블락

2. 반자동화

관리자 확인

감사합니다.