인공지능을 활용한 가상 캐릭터의 특성 기반 마약 의심 SNS 포스팅 탐지 솔루션

발표자 : 안 상 선

sangsun.ahn@m-robo.com

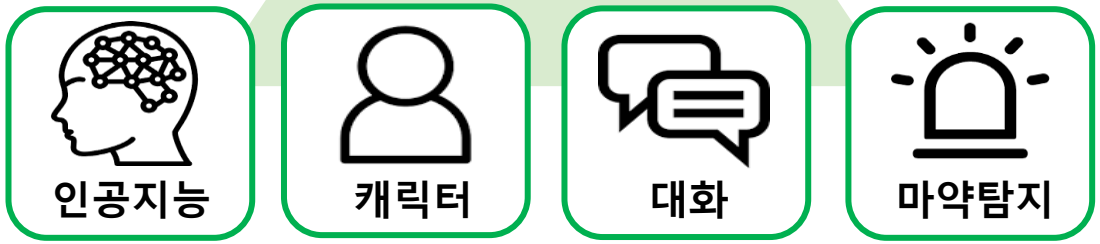
www.ahnboat.xyz

1. 아이디어 컨셉

- **대한민국은 마약류 사범은 2021년 현재 1만6163명**으로, UN기준으로 인구 10만명당 마약류 사범이 20명 미만(약 1만여명)인 경우 부여하는 **마약청정국의 지위를 사실상 잃은 상태**임
- 특히 마약을 **거래하는 플랫폼인 텔레그램**, **결제에 쓰이는 암호화폐**는 **익명성을** 특징으로 하고 있어 단속 및 적발이 어려운데, 이 같은 취약점을 노리고 최근 인스타그램 및 X(트위터)로 마약 판매 포스팅이 폭증하고 있음
- 이에 **인공지능을 이용해서 마약판매 의심 포스팅을 쉽고 빠르게 탐지하는 솔루션**을 제안함

A.I를 활용한 캐릭터-특성 기반 마약 판매 의심 포스팅 탐지 솔루션

Drug Catcher



기획배경

마약 오남용의 급증으로 인한 사회문제

- **대한민국은 마약류 사범은 2021년 현재 1만6163명**으로, UN기준으로 인구 10만명당 마약류사범이 20명 미만(약 1만여명)인 경우 부여하는 **마약청정국의 지위를 사실상 잃은 상태**임

- 현재 마약류 뿐 만 아니라, **펜타닐 등 신종 중독성 약물의 오남용**으로 인해 실제 마약 중독자 수치는 1만 6천여명을 훨씬 초과하는 것으로 알려져 있음

연합뉴스TV
자전거 안장 뜯자 '마약' 와르르…250억원대 적발
자전거 안장 뜯자 '마약' 와르르…250억원대 적발 [앵커] 해외 항공특송화물을 이용해 자전거 안장과 야구 배트 등에 마약을 숨겨 국내로 들여와 판매…
3일 전

한국농어촌방송
올해 4월까지 마약 적발 '역대 최대'…여행자 반입 1320% 증가
[한국농어촌방송=홍채린 기자] 윤태식 관세청장이 급증하는 마약범죄를 근절하기 위해 전국 세관 마약조사관 회의를 개최하고 "마약과의 전쟁에 승리…
1개월 전

동아일보
청주지검, 국제우편 이용 국내 마약 반입 내·외국인 22명 적발
국제우편을 이용해 마약을 국내로 들여온 내·외국인 22명이 검찰에 적발됐다. 청주지검 형사3부(부장검사 안창주)는 태국에서 필로폰 997.01 g (2…
2023. 3. 29.

연합뉴스TV
美, '좀비마약' 펜타닐 등 불법의약품 제조장비판매 中업체 제재
美, '좀비마약' 펜타닐 등 불법의약품 제조장비판매 中업체 제재 미국 재무부는 '좀비마약'으로 불리는 펜타닐 등 불법 의약품 생산과 관련된 중국…
3주 전

라디오코리아
샌프란시스코서 210만 여명 죽음에 이르게 할 수 있는 펜타닐 압수
[앵커멘트]지난달(5일) 1일 CA주가 샌프란시스코에서 마약과의 전쟁을 선포한 이후 불과 6주 동안 무려 210만 여명 이상을 죽음에 이르게 할 수 있는…
4일 전

안전신문
정부, 청년 사망원인 1위 펜타닐 관리방안 논의… "강력한 마약대책 실행할 것"
최근 국내에서 마약 범죄가 자주 발생하는 가운데 미국에서 많은 사망자를 발생케 한 펜타닐을 국내에서 관리키 위한 방안이 논의됐다.

문제의 원인 분석(1)

마약류 반입 및 중독 급증의 원인 : 구매 환경의 편의성 증대

- 해외 여행 및 외국인의 입국 증대로 인해 **마약의 밀반입 규모가 자연적으로 늘어나는 경향**을 보이고 있음

- 특히, 마약을 **거래하는 플랫폼인 텔레그램**과 **결제에 활용되는 암호 화폐**는 **익명성**을 특징으로 하고 있어 단속 및 적발이 어려움 단점이 있음

SBS 뉴스

마약 거래에 사용되는 텔레그램... 그 실체는? [뽀얀거탑]

[골룸] 뽀얀거탑 382 : 마약 거래에 사용되는 텔레그램..그 실체는? 한국 내 온라인 마약 거래의 72.8%가 한 SNS에서 진행되는 것, 알고 계셨나요?

9시간 전

경북도민일보

텔레그램·가상자산 이용 마약 판매책 3명 붙잡혀

텔레그램으로 정보를 주고받은 뒤 가상자산으로 결제하는 방식으로 수백 명이 투약할 수 있는 양의 마약을 판매한 일당이 경찰에 붙잡혔다.

6일 전

프레시안

대구경찰, 텔레그램 통해 마약 판매 20대 3명 검거

대구 성서경찰서는 13일 텔레그램을 이용해 마약을 판매한 혐의(마약류관리법 위반)로 A(23)씨 등 20대 3명을 검거해 조사 중이다.

6일 전

암호화폐로 필로폰 구매...제주경찰, 마약사범 무더기 검거

이창준 기자 headlinejeju@headlinejeju.co.kr | 승인 2022.09.15 10:46 | 댓글 0

1월부터 집중단속 결과 75명 검거...작년 대비 2배 가까이 증가
필로폰 판매.투약 76% 차지...연령대는 40대가 가장 많아

3. 문제 원인 분석(2)

- 최근의 마약 유통 실태를 보면 기존에 현장 단속 위주로는 한계가 있으며, 현장 단속의 경우, 마약의 중증 중독자, 유통책 위주로 검거되는 경향이 있음

- 따라서 **급증하고 있는 마약류 사범 초범이나 1회성으로 투약하는 경우에는 단속에 한계**가 있음

- 이에 **마약 관련 은어가 섞인 "대화내용" 분석을 통해서 미리 마약 유통을 조기에 모니터링 할 수 있는 "이상탐지" 솔루션**을 제안함

Service Flow Chart

마약 및 중독성 약물에 대한 “은어” 및 대화 규칙 학습

↓

대화 캐릭터 설정
(마약 중독 단계에 따른 캐릭터 특성 생성)

↓

특성 별 대화 생성

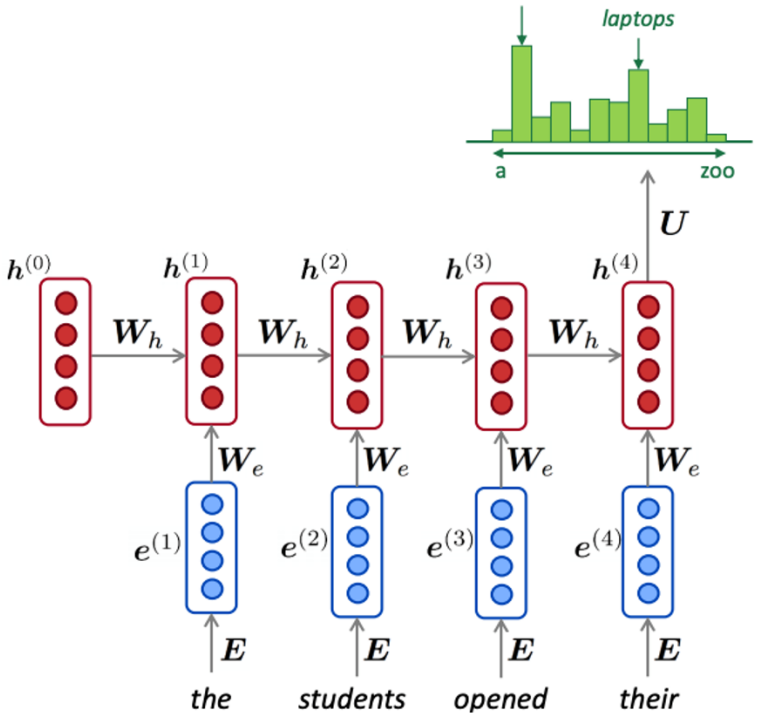
↓

실제 마약 거래 대화와 유사도 분석

↓

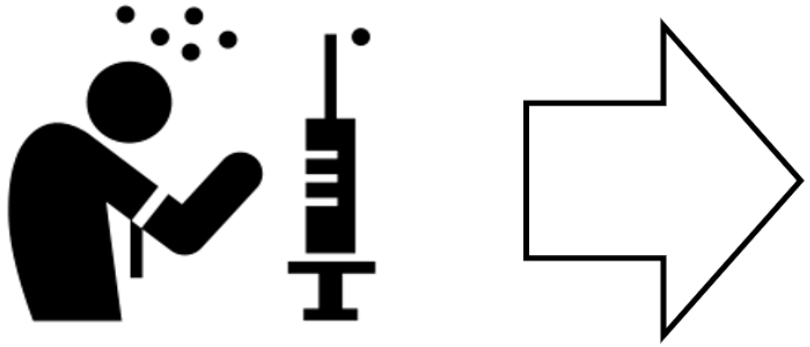
유사도를 반영한 특성별 대화 재생성

종료



캐릭터 프로파일 설정

대화 캐릭터 설정
(마약 중독 단계에 따른 캐릭터 특성 생성)



실험적 사용단계

호기심 : 80%
적발에 대한 걱정 : 80%
가격 민감도 : 80%
구매 강박감 : 0%

습관적 사용단계

호기심 : 65%
적발에 대한 걱정 : 65%
가격 민감도 : 60%
구매 강박감 : 15%

심화적 사용단계

호기심 : 50%
적발에 대한 걱정 : 55%
가격 민감도 : 30%
구매 강박감 : 60%

강박적 사용단계

호기심 : 40%
적발에 대한 걱정 : 55%
가격 민감도 : 40%
구매 강박감 : 80%

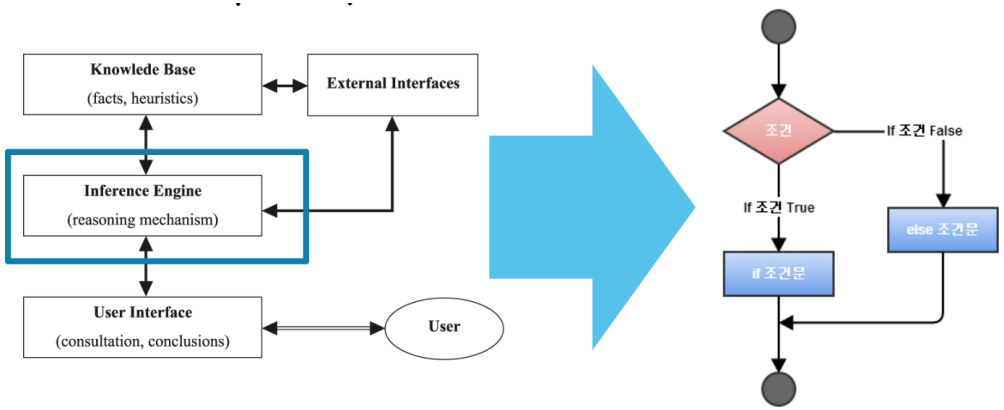
방법론 : 전통적인 프로그래밍 방법

Expert Model(전문가모델)

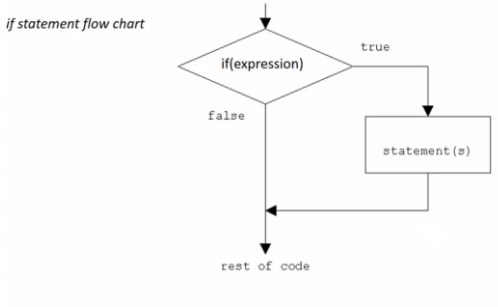
■ 기존의 전통적인 프로그래밍(Conventional Programming)

- 인공지능 기술의 응용 분야 중 하나로 인간이 특정분야에 대하여 가지고 있는 전문적인 지식을 정리하고 표현한 뒤 컴퓨터에 기억시켜서 이를 활용함
- 일반인도 이 전문지식을 이용할 수 있도록 하는 모델로, 1971년 DENDRAL의 화학분자구조식 추정 모형이 대표적임
- 마약거래 탐지를 위해서는 전문가의 탐지 “Flow Chart”의 “판단“에 대한 기준, 기준의 수, 선후관계 등을 설계해야 함

[그림] 전문가 모델과 Flow Chart



[그림] 마약거래 탐지 모델 개발을 위한 Flow Chart



방법론 : 기계학습(Machine Learning Model)

기계학습 모델

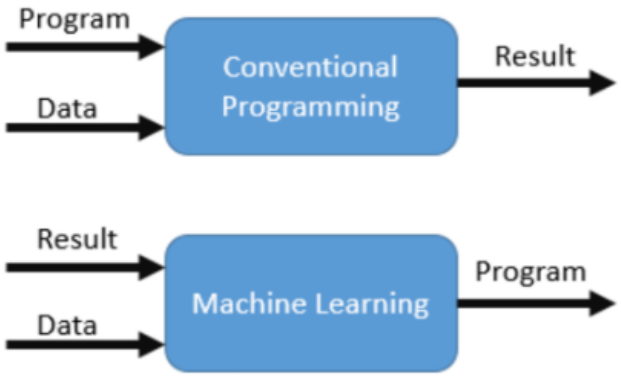
■ 기존의 전통적인 프로그래밍(Conventional Programming)

- 작성된 프로그램에 데이터를 입력하면 결과 값(분류 값)을 출력

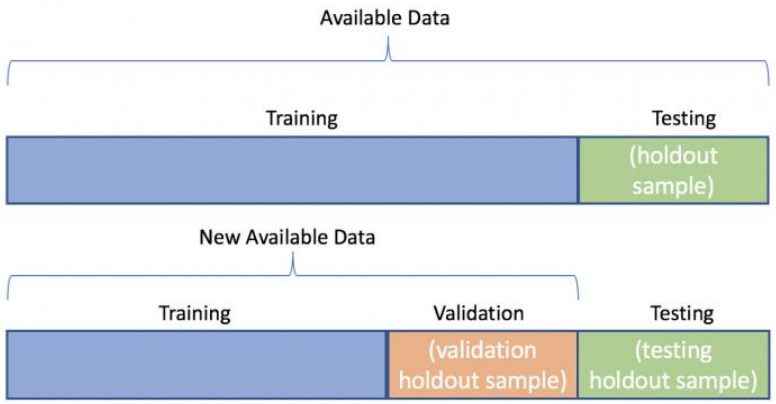
■ 머신러닝(기계학습) 모형

- 라벨링된 데이터, 즉 입력 데이터(X)와 출력 값(Y, 분류 값)을 사용해서 새로운 데이터에 대해서 학습
- 주어진 데이터를 모형의 학습(Training)과 테스트(Test), 검증(Validation) 부분으로 분할하여 모형 도출 및 평가

[그림] 전통적인 프로그래밍 방법과 머신러닝 비교



[그림] 머신러닝에서의 데이터 분할(학습, 테스트 데이터)



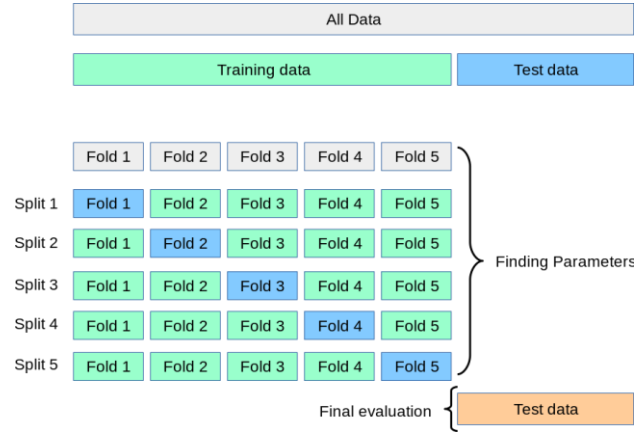
방법론 : 기계학습(Machine Learning Model)

기계학습 모델

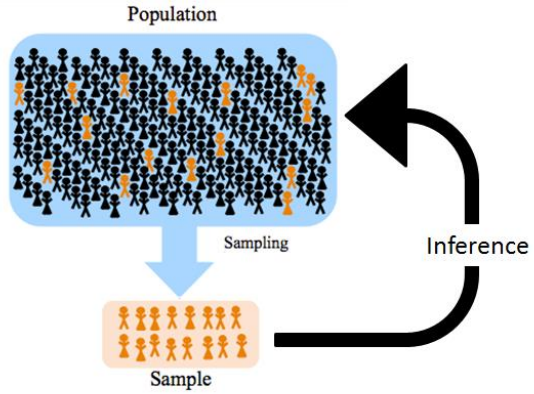
■ 머신러닝(기계학습) 모형의 통계 모형 비교

- 기계학습 모델에 사용되는 데이터는 전체 모집단을 대표하는 표본이며, 모델의 성능은 별도의 테스트 데이터 세트에 대해서 평가하며, 교차검증(Cross Validation test)를 통해서 Bias로 인한 문제를 조정함
- 통계 모형(통계적 추정 모형)은 모집단에 대한 가정을 통해서 주어진 샘플을 근거로 모집단의 값(평균, 분산 등)을 추정하며 샘플링 Bias의 문제를 감안해 “유의수준” “신뢰구간” “1, 2종 오류“ 등의 개념으로 오류 가능성을 제시

[그림] 전통적인 프로그래밍 방법과 머신러닝 비교



[그림] 머신러닝에서의 데이터 분할(학습, 테스트 데이터)



방법론 : 기계학습(Machine Learning Model)

[참고] 범죄 암수(Hidden Crime) 추정

■ 범죄 암수 추정

- 검찰청의 범죄백서나 언론 보도를 통해 접하는 범죄 단속 건수는 검거 또는 적발된 경우로 실제 발생한 범죄보다 그 수치가 적을 수 밖에 없으며, 실제 범죄 발생 건수는 검거율 등의 지표를 이용해서 범죄 등을 추산해서 사용하고 있음

〈표〉 마약류 검거 인원 현황

<표 2> 2016년 형사사법 기관별 범죄사건처리 현황

단위: 건, 명

구분	전체범죄			계	마약류 범죄	유해화학물질 관련 범죄
	발생건수	검거건수	검거인원, 수용인원	검거인원, 수용인원	검거인원, 수용인원	검거인원, 수용인원
경찰	1,818,933	1,522,342	1,867,707	8,922	8,920	2
검찰	2,008,290	1,691,370	2,020,196	14,232	14,214	18
법원	-	-	449,973	9,244	8,944	300
교도소	-	-	57,675	3,073	3,000	73

[그림] 마약 암수 추정

기계학습 방법

분류모형(Classification Model)

- 기계학습의 한 유형으로 입력된 데이터를 사전에 정의된 Class 또는 카테고리로 분류하는 작업을 수행하는 모델로,
  다양한 분류 알고리즘을 사용해서 데이터와 데이터의 레이블 값을 학습 시키고 모델을 생성

〈참고〉 분류모형의 프로세스

분류 모형의 프로세스

1) 데이터 수집 : 입력 데이터와 Class 레이블로 구성

2) 데이터 전 처리 : 모델에 적합한 데이터 형태로 변환

3) 데이터 분할 : 전체 데이터를 학습 데이터와 테스트 데이터로 분할

4) 모델 선택 : 로지스틱 회귀, 의사결정 나무, 근접(이웃)분류 등의 모델 사용

5) 모델 학습 : 선택한 모델에 학습 데이터를 제공해 모델을 학습

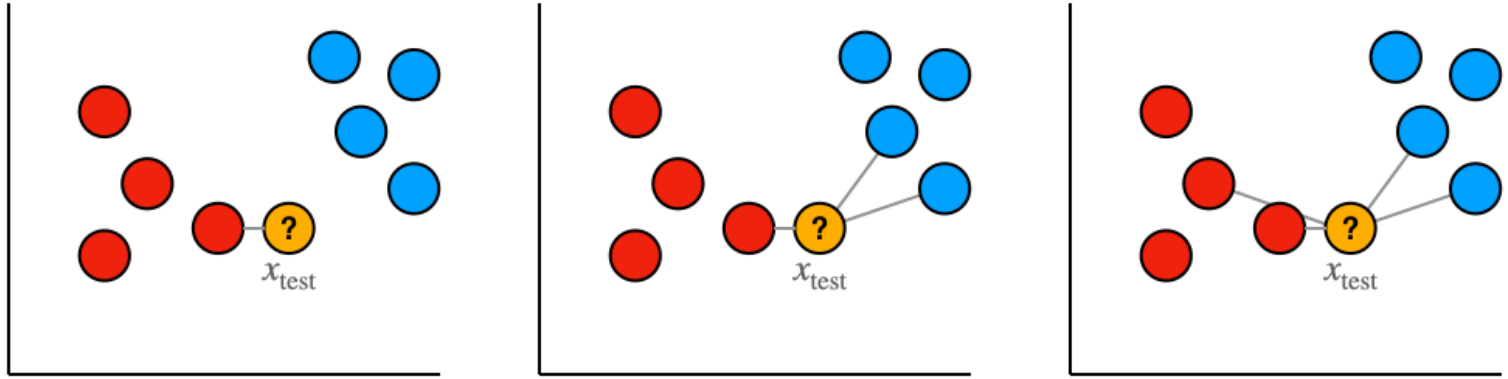
6) 모델의 평가 및 조정

분류 모형 : 세부 모형

K-이웃분류모형(K-Nearest Neighbor, KNN)

- 지도학습(Supervised Learning)의 한 종류로, 새로운 데이터 포인트를 분류할 때 기존의 학습 데이터 중에서 가장 유사한 데이터 포인트의 클래스로 분류
- 새로운 데이터 포인트를 분류할 때 주변의 k개 이웃 데이터 포인트들의 클래스를 고려하여 다수결로 클래스를 결정함
- 아래 그림에서 K값이 1이면, 제일 가까운 빨간색이, K값이 3이면 파란색 2개, 빨간색1개로 빨간색으로 판정하고, K=4 일 때는 빨간색, 파란색이 각각 2:2이므로 판정이 어려움, 이 때문에 K 값은 주로 홀수로 설정

[그림] 이웃 분류모형 : 그룹 판정 방법 예시

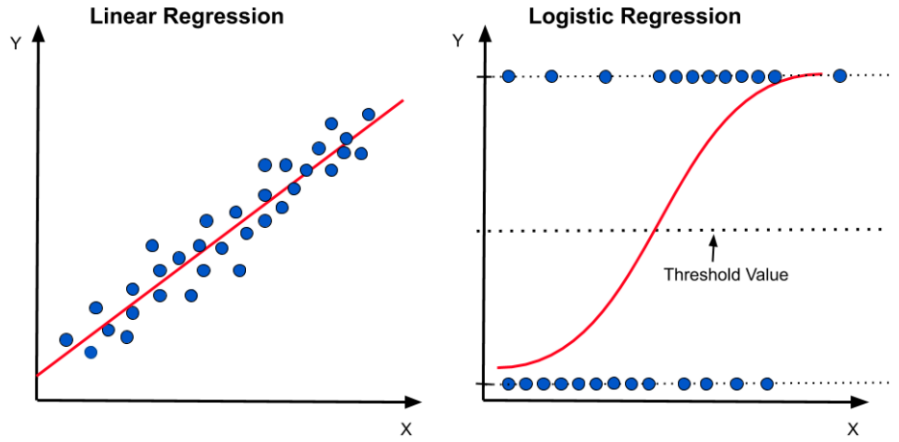


분류 모형 : 세부 모형

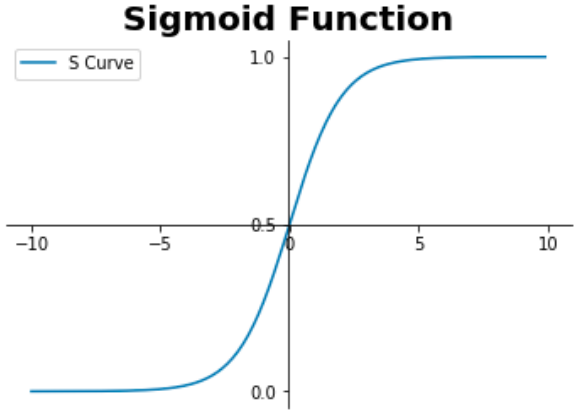
로지스틱 (회귀)모형

- 범주형 변수로 구성된 종속 변수를 대상으로 한 통계 기반의 모형으로, 독립 변수의 선형 조합을 사용하여 종속 변수가 특정 범주에 속할 확률을 예측하며, 분류 문제에 주로 사용됨
- 모형에 사용되는 Sigmoid 함수에 따르면 지수x 값이 커지면 함수 값이 1에 가까워 지고, 반대로 작아지면 0에 가까워 지는 특성이 있음
- 이에 따라 회귀모형과 동일하게 종속변수가 수치형 값으로 표시가 되지만, 특정 집단(0또는 1)에 속할 확률을 알려주기 때문에 분류 모형의 특성을 가지고 있음

[그림] 로지스틱 모형 (회귀모형과 비교)



[그림] 로지스틱 모형(Sigmoid 함수)

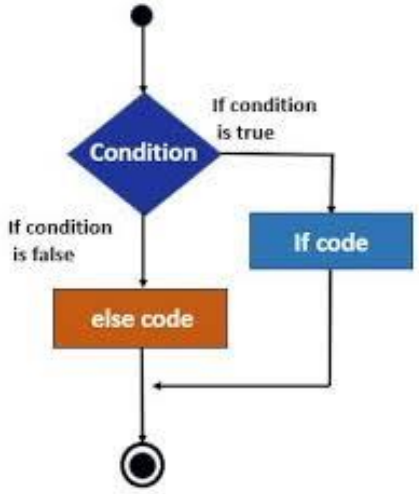


분류 모형 : 세부 모형

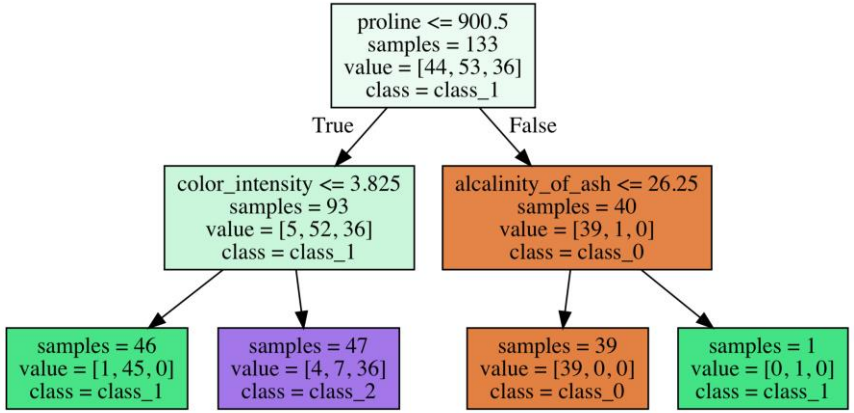
의사결정 나무 모형

- 데이터를 분할하여 트리 구조를 생성하는 알고리즘으로. 각 노드에서 특정 기준에 따라 데이터를 분할하고, 노드에서 최종 예측을 수행
- 전체 집단을 몇 개의 소그룹으로 분류하거나 예측하는 분석 방법으로 복수의 조건(if)문을 사용해서 대상을 분류함
- 의사결정나무는 해석이 용이하고 범주형 및 연속형 변수를 모두 처리할 수 있어 다양한 분야에서 활용되고 있음

[그림] 순서도의 조건문



[그림] 의사결정 나무 모형



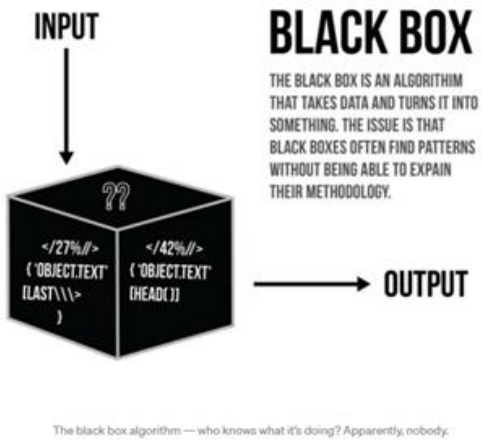
블랙박스와 설명가능한 인공지능

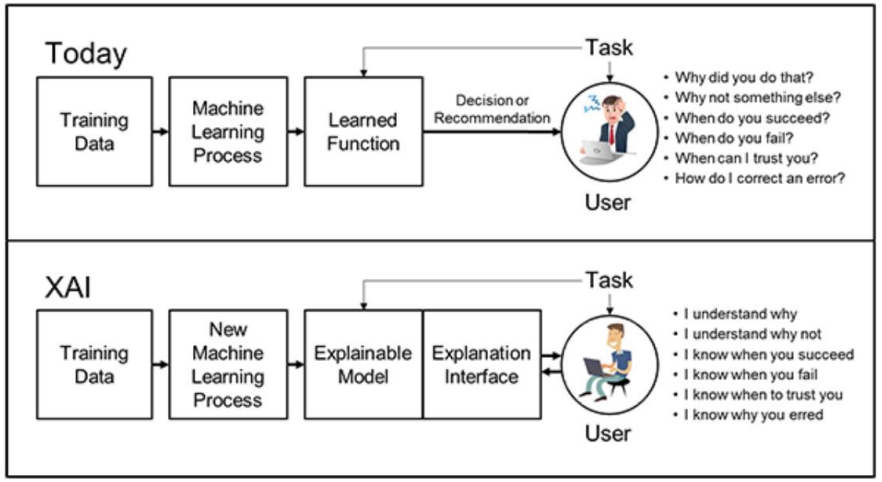
설명가능한 인공지능(Explainable AI)

■ **머신러닝 모델과 블랙박스**

- 전통적으로 머신러닝 모델은 높은 정확성에도 불구하고, '블랙박스'라는 오명을 받아왔는데, 이는 예측의 논리를 명확하게 설명하거나 식별할 수 있지만 왜 그런 것인지 알 수 없음
- 이를 해결하기 위해 설명가능한 인공지능(Explainable AI)라는 연구 분야가 최근 주목 받고 있으며, 이는 1. 머신러닝 모델의 신뢰, 2. 머신러닝 모델의 비교 가능성, 3. 규제 및 거버넌스에 대한 요구사항 충족 등 제고에 기여할 것으로 평가받고 있음

[그림] 머신러닝 모델의 블랙박스와 설명가능한 인공 지능 개념도





https://www.linkedin.com/pulse/explaining-explainable-ai-part-1-vatsal-nayak-/?trackingId=T6xjnbbRYVaGJoNzaGvzeg%3D%3D

사용모델

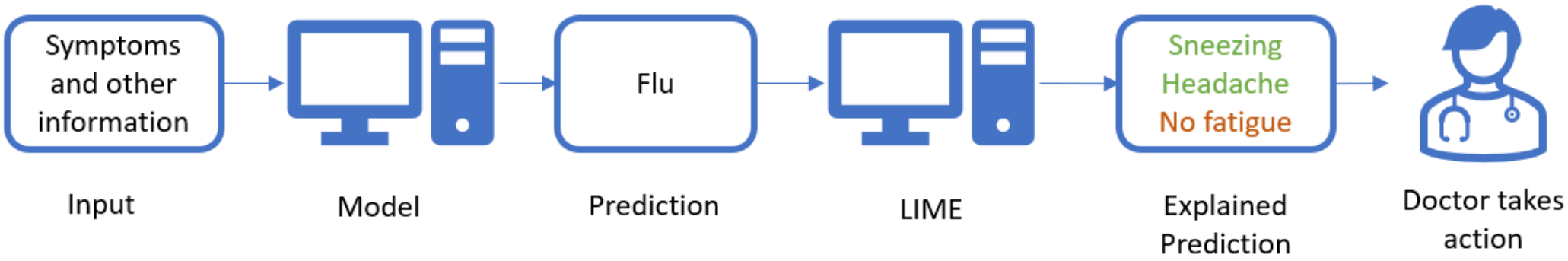
LIME(Local Interpretable Model-agnostic Explanations)

- **LIME 모델은** blackbox 모델을 지역적으로 근사함으로써 설명 가능성을 제공하는 알고리즘이라 할 수 있음
- 지역적 설명 가능성(지역적 근사)

  : 머신러닝 모델에 따라 부동산 담보 대출은 거절됐지만, 자세히 분석해 본 결과 수원 영통구 시세 1억~2억원 상당의 빌라에서 대출 거절 가능성이 커진 것으로 보였다 => 확인해 보니 "빌라왕" 사기 지역이었음

  : 병 증세를 입력하면 병명을 출력하는 머신 러닝 모델을 분석해보니, 녹색 부분(재채기, 두통)은 독감예측에 기여했으나, 빨간색 항목(피로 없음, No fatigue)는 독감 예측에 기여하지 못하는 것으로 판명

[그림] 지역적 예측 예시 : 독감 예측 및 사후 조치



사용모델

LIME(Local Interpretable Model-agnostic Explanations)

- **LIME 모델은 머신러닝 모델의 예측 결과를 설명하기 위해 사용되는 분석 모형으로, 복잡하고 해석하기 어려운 블랙박스 모델의 결과에 대해 쉽게 설명할 수 있다는 장점이 있음**

- **주요 특징은 다음과 같음**
  1. 모델 종류에 구애받지 않음 (Model-agnostic) 다양한 머신러닝 모델에 적용 가능
     (예: 딥러닝, 랜덤 포레스트, SVM 등)
  2. 개별 예측 결과에 대한 국소적 설명 (Local Interpretation) 전체 모델이 아닌, 특정 입력에 대한 모델의 예측 결과를 설명
  3. 사람이 이해하기 쉬운 형태의 설명 (Interpretable Explanations) 모델의 예측에 영향을 미친 주요 특성(features)을 식별하고, 이를 사람이 이해할 수 있는 형태로 제시

〈참고〉 파이썬에서 사용방법

https://towardsdatascience.com/understand-the-workings-of-black-box-models-with-lime-92203f906431

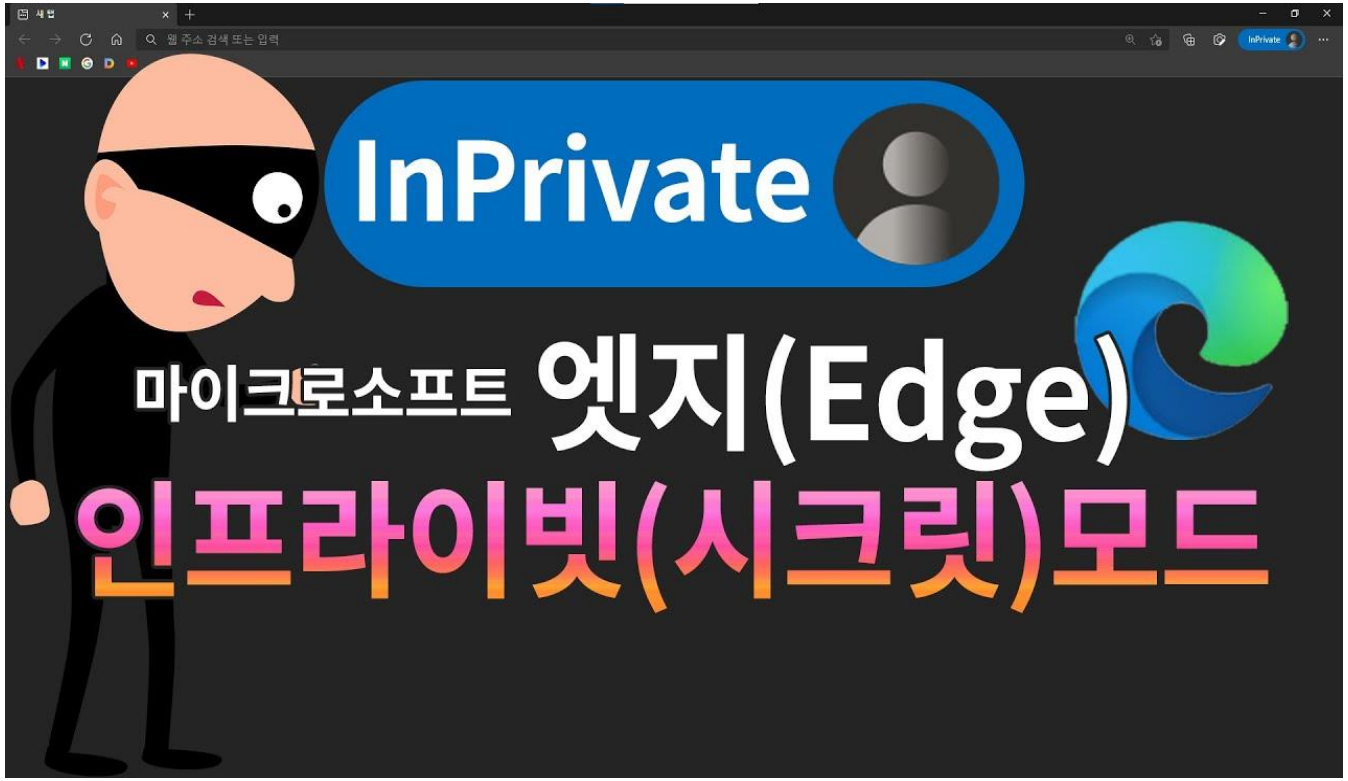
사용모델

LIME(Local Interpretable Model-agnostic Explanations)

- 기계 학습 모델의 예측 결과를 해석하고 설명하는 데 사용되는 모델
- 특히 "블랙 박스" 모델, 즉 내부 동작이 복잡하고 해석이 어려운 모델에 대한 설명을 생성하는 데 유용

Local
Interpretable
Model-agnostic
Explanations

데이터 수집



데이터 수집 (VPN 권장)

파이낸셜뉴스
온라인 무법지대](하)SNS에 '아...

국민일보
아이스·작대기·얼음 팔아요" 너무 쉬워진 S...

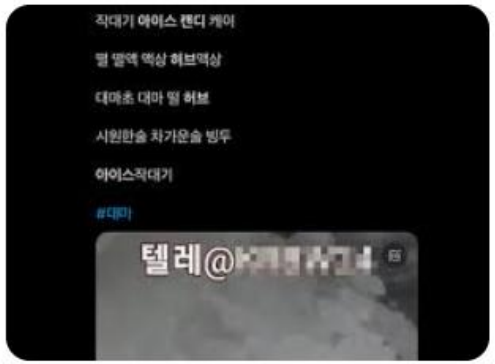
대구노컷
대구노컷

국민일보
아이스·작대기·얼음 팔아요" ㄴ

...

Medium
아이스작대기1위 업체...

매일경제
한겨울에 왜 아이스, ...



법률신문
이슈 인사이드] 캔디, 아이스, 허브...

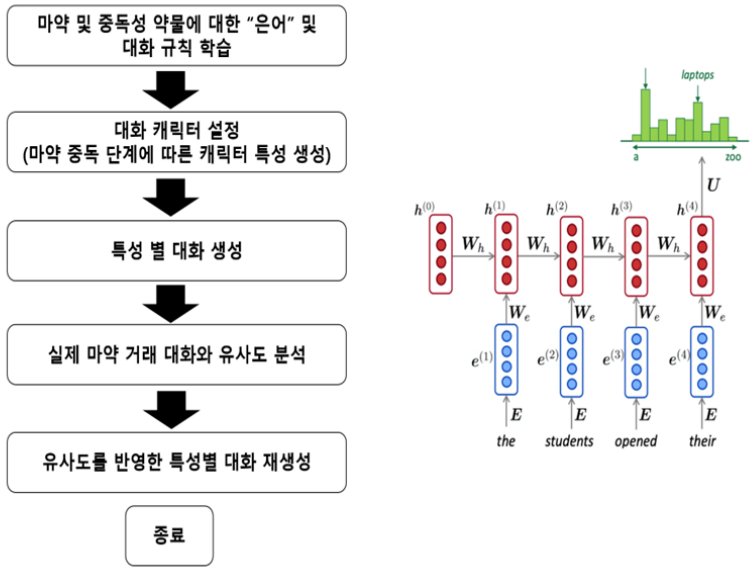
GrabCAD
아이스작대기 [텔레@DROP1004] 아...

모델링 및 결과 분석

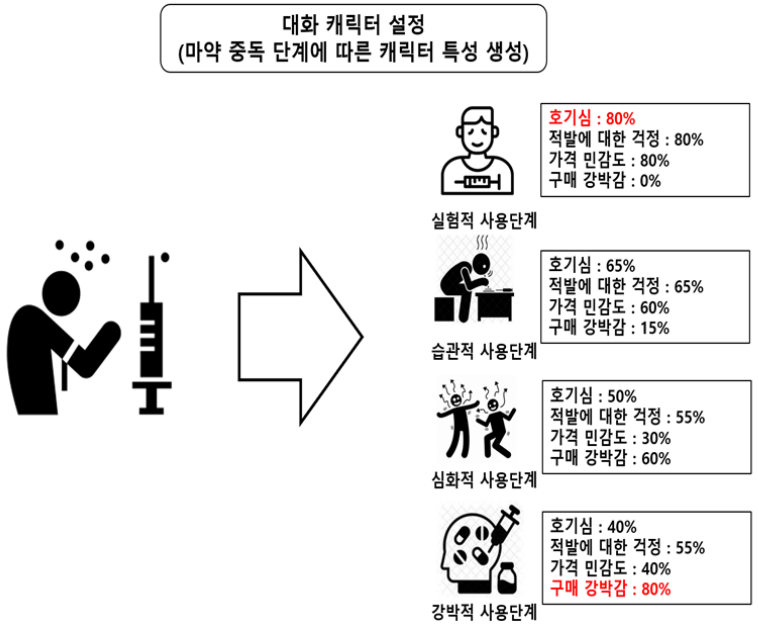
1단계 : 캐릭터- 특성 기반 포스팅 생성

- 데이터 수집 및 학습 : 마약 판매 및 의심 SNS 포스팅을 수집한 뒤에 “은어” 및 특정 연관 단어 도출을 위한 자연어 분석을 진행
- 이후 마약 판매자의 특성에 기반한 캐릭터 함수를 정의하고, GPT2 및 BERT 등 텍스트 생성 알고리즘을 이용해서 포스팅 글 생성

[그림] SNS포스팅(글) 생성 프로세스



[그림] 포스팅 생성을 위한 캐릭터 특성 설정

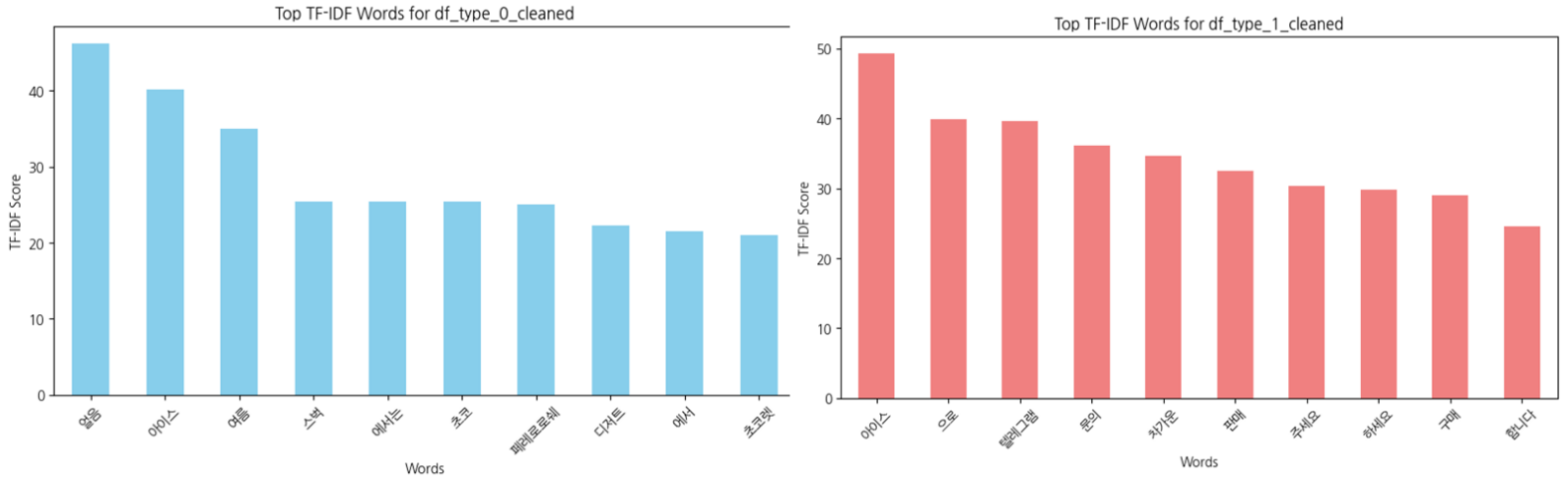


2단계 : 생성 및 실제 SNS 포스팅 데이터 분석

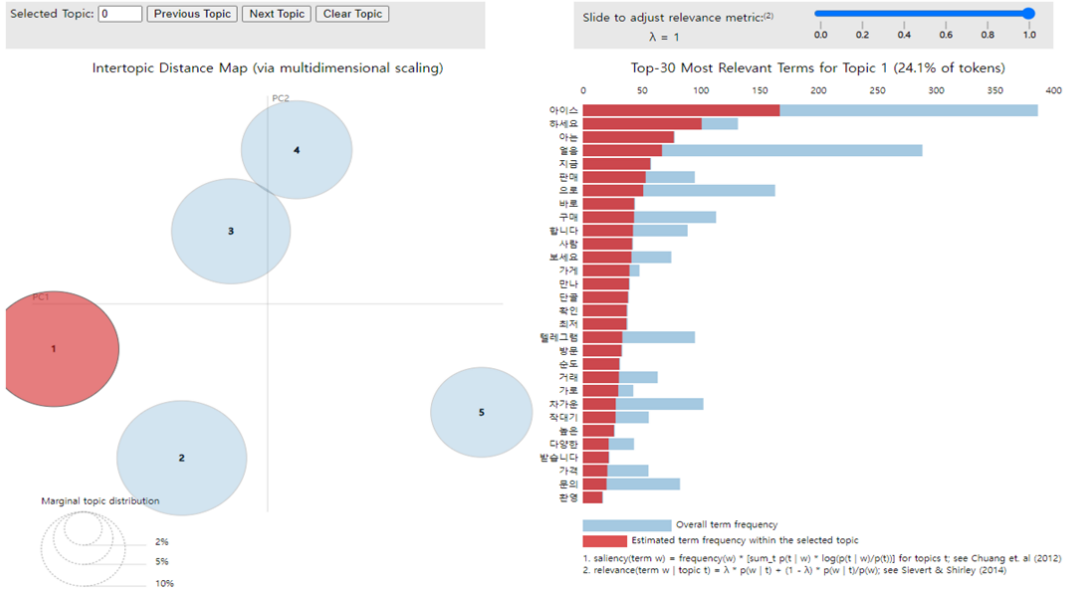
- 1단계에서 생성된 데이터 및 실제 마약 판매 의심 포스팅 데이터를 사용해서 약 1200 여개의 SNS 포스팅(텍스트)를 생성했음
- 마약거래 의심 포스팅의 주요 키워드인 “아이스”는 실제 아이스아메리카노, 얼음 판매, 빙상 공연 등 일상적인 포스팅에도 사용되는 문제가 있음
- 텍스트 분석 결과 마약 판매 의심 포스팅은 특정 단어 간의 연관 정도가 강하게 나타났으며, 토픽 모델링 분석을 통해 마약 판매 의심 토픽을 도출했음

[그림] WordCloud : 정상 포스팅(좌측) 및 마약 판매 의심 포스팅 (우측)

[그림] TF-IDF 분석 : 정상 포스팅(좌측) 및 마약 판매 의심 포스팅 (우측)



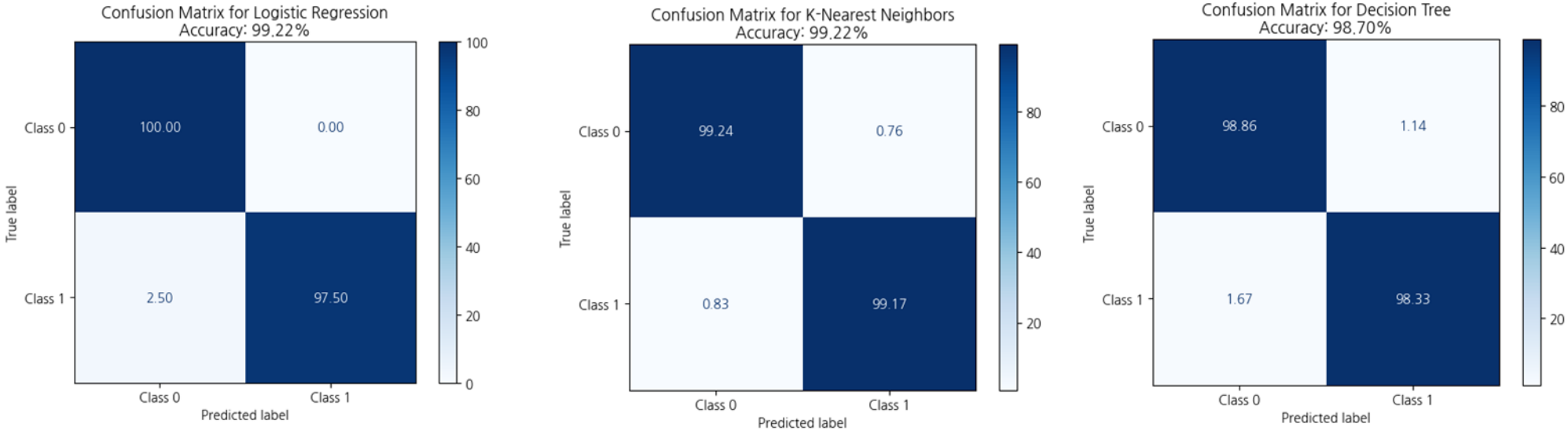
[그림] 토픽 모델링 : 마약 판매 의심 토픽 도출



3단계 : 포스팅 탐지 알고리즘

- 도출된 마약거래 의심 포스팅에 대한 “특징”을 근거로 실제 정상 포스팅과 마약 판매 의심 포스팅을 구분하는 알고리즘을 인공지능 모형으로 개발
- 기계학습 방법 중 하나인 분류모형(Classification Model)을 사용해 포스팅(텍스트 데이터)를 사전에 정의된 Class(정상, 마약 판매 의심)으로 정확히 구분하는 것을 평가
  * 기본 분류모형인 로지스틱 회귀모형, 이웃분류모형, 의사결정나무 모형을 사용

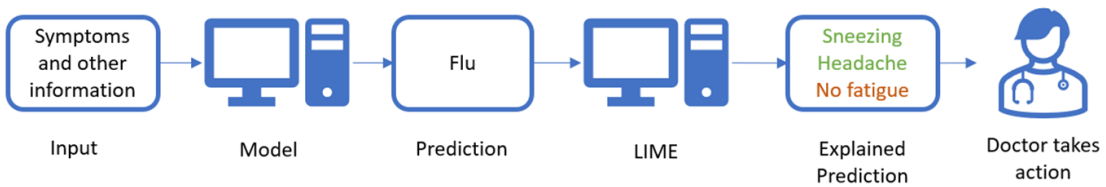
[그림] 분류모형의 정확도 평가 (평균 : 약 99%)



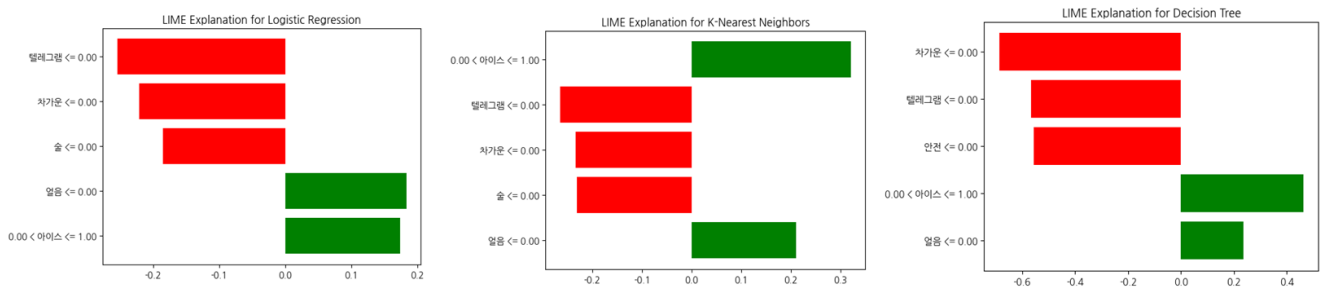
4단계 : 탐지 알고리즘의 매커니즘 설명 : LIME 모델

- LIME 모델은 blackbox 모델을 지역적으로 근사함으로써 설명 가능성을 제공하는 알고리즘으로 아래의 독감예측 모델에 따르면 지역적인 설명이 가능
- 병 증세를 보니, 녹색 부분(재채기, 두통)은 독감예측에 기여했으나, 빨간색 항목(피로 없음, No fatigue)는 독감 예측에 기여하지 못하는 것으로 판명
- 이 같은 방법으로 위의 분류모델에 대해서 “마약거래 의심 단어”를 분석한 결과 각 모델별로 유의미한 단어와 기여도가 낮은 단어를 추출했음

[그림]LIME 모델 예시 : 독감의 증세에 대한 예측 기여도



[그림] LIME모델을 적용한 각 분류모델의 “의심 키워드” 기여도



감사합니다