2024년도 공공기관 용역과제

AI개발 수행내역서

과제명	AI기반 약물 예측모델 개발 및 시각화
담당자	문예빈

2025년 01 월 07 일

AI개발 수행내용

1. 사업과제 : AI기반 약물 예측모델 개발 및 시각화

2. 개요 및 현황

2.1 추진배경 및 목적

○ 의료 및 약물 데이터베이스의 장기적인 축적으로 인해, 데이터 기반 인공지능 기술을 활용한 약물 분류 모델에 대한 수요가 점진적으로 증가하고 있음.

○ 신약 개발, 약물의 안전성 확보, 그리고 맞춤형 치료 솔루션의 필요성이 증가함에 따라, 약물의 분류 및 특성에 대한 분석 기술이 점차 중요해지고 있는 상황.

○ 약물 데이터를 분석하고 복잡한 특성 및 상호작용 패턴을 학습함으로써, 특정 약물의 효능, 적응증, 및 부작용 가능성을 사전에 분류하여 의료 산업에 기여하고자 함

○ 초기 단계로 특정 약물 데이터를 기반으로 분류 모델을 시범적으로 구축하고, 이를 기반으로 다양한 약물군 및 질환 관련 데이터로 확장하여 의료 데이터 분석의 활용도를 높이고자 함.

2.2 과제 범위

과제구분		내용
AI	AI기반 약물 예측모델 구현	원시 데이터 수집 및 데이터셋 구축
		데이터 전처리, 표준화, 상관관계 분석 (EDA도구 활용)
		예측모델 선정 및 학습
		Accuracy 등 평가지표를 활용한 모델 성능 평가
		웹 API 및 프로토타입 구축.
		예측 모델 시각화
		예측모델 웹기반 시스템 구축
		테스트

2.3 과제 추진 방법

1) 구축 대상 선정 기준

○ 데이터 접근성 및 활용성

■ 데이터 수집 및 관리의 용이성

■ 종속변수에 영향을 미치는 다양한 독립변수에 대한 정보 포함여부를 통한 모델학습의 유용성

○ 예측모델 개발 효율성

■ 모델 학습 및 평가 과정 간소화를 위한 다른 환경 기초데이터에 비해 변수가 상대적으로 단순한 구조 여부

■ 개발된 모델을 통해 다른 환경기초 데이터에 적용 가능 여부

○ 의료문제 해결 기여도 및 경제성

■ 예측모델을 통해 약물 개발 및 안전성 평가에 기여할 수 있는 지 여부 (예: 부작용 감소, 치료 효율 향상 등)

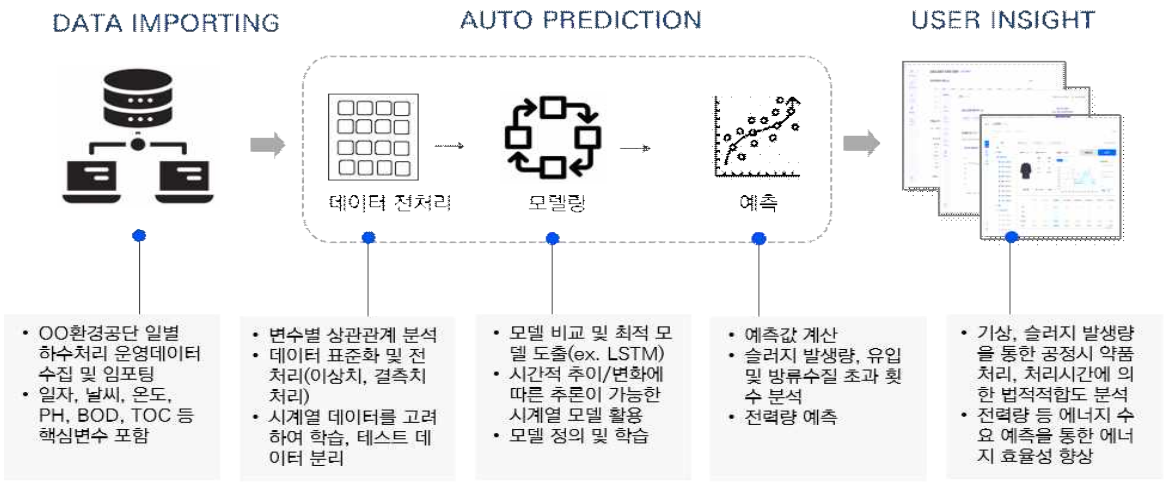
■ 약물 개발 과정의 효율성을 높여 연구 및 개발 비용 절감 효과 여부

■ 약물의 효과적 분류를 통해 맞춤형 치료와 같은 의료 혁신을 통한 사회적 비용 감소 효과 여부

2) AI 예측 분석모델 적용 대상

약물관리 기능	수집 데이터	예측모델인자(독립변수)	AI예측 분석 대상
약물	- 환자의 나이, 성별, 혈압, 콜레스테롤 수치, 나트륨-칼륨 비율, 약물종류	- Age: 환자의 나이 - Sex: 환자의 성별 - Bp: 혈압 (Blood Pressure) - Cholesterol: 콜레스테롤 수치 - Na_to_K: 나트륨과 칼륨 비율	- 환자의 특성(나이, 성별 등)에 따른 약물 분류 - 각 변수(나이, 성별, 혈압 등)와 약물 간의 관계 분석 - 약물에 대한 예측 모델 개발 및 새로운 환자에 대한 약물 추천

3) AI 분석모델 구축 프로세스

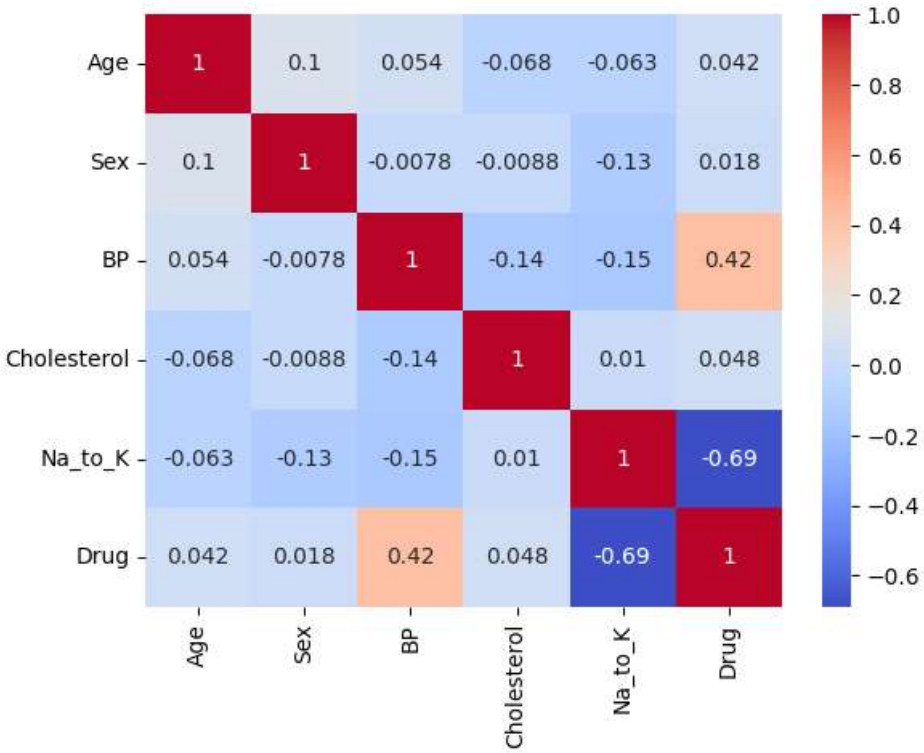


연구개발 주요 결과물

1. 데이터 수집
   - 약물 데이터(엑셀)
   - https://www.kaggle.com/datasets/prathamtripathi/drug-classification/data

2. 데이터 분석

2.1 약물 데이터 상관관계(Heatmap)



○ EDA 히스토그램, 히트맵 변수별 분포를 통해 정규분포 여부, 데이터 변환(ex. 로그변환) 필요성 및 변수 간 관계를 유추

○ 약물 예측 모델링을 위한 대상 설정 : 약물(Drug)

○ 약품에 영향을 미치는 요인 분석

■ 환자의 나이, 성별, 혈압, 콜레스테롤 수치, 나트륨-칼륨 비율

2.2 탐색적 데이터 분석

○ 결측치 및 중복값 통계

Overview | Alerts 2 | Reproduction

Dataset statistics

Number of variables	6
Number of observations	200
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	50.8 KiB
Average record size in memory	259.9 B

Variable types

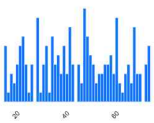
Numeric	2
Categorical	4

○ 주요 변수별 데이터 분포(Histogram)

Age

Real number (ℝ)

Distinct	57	**Minimum**	15
Distinct (%)	28.5%	**Maximum**	74
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	44.315	**Memory size**	1.7 KiB

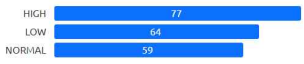


BP

Categorical

High correlation

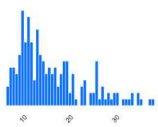
Distinct	3
Distinct (%)	1.5%
Missing	0
Missing (%)	0.0%
Memory size	12.1 KiB



Na_to_K

Real number (ℝ)

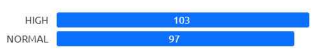
Distinct	198	**Minimum**	6.269
Distinct (%)	99.0%	**Maximum**	38.247
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	16.084485	**Memory size**	1.7 KiB



Cholesterol

Categorical

Distinct	2
Distinct (%)	1.0%
Missing	0
Missing (%)	0.0%
Memory size	12.2 KiB



○ 데이터 전처리

First rows | Last rows

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	F	HIGH	HIGH	25.355	DrugY
1	47	M	LOW	HIGH	13.093	drugC
2	47	M	LOW	HIGH	10.114	drugC
3	28	F	NORMAL	HIGH	7.798	drugX
4	61	F	LOW	HIGH	18.043	DrugY
5	22	F	NORMAL	HIGH	8.607	drugX
6	49	F	NORMAL	HIGH	16.275	DrugY
7	41	M	LOW	HIGH	11.037	drugC
8	60	M	NORMAL	HIGH	15.171	DrugY
9	43	M	LOW	NORMAL	19.368	DrugY

3. 데이터 학습 및 모델정의

3.1 모델정의 및 학습

○ 분류 모델 정의 : KNN

```
# model - KNN
N_NEIGHBORS = 6
knn = KNeighborsClassifier(n_neighbors=N_NEIGHBORS)
knn.fit(X_train, y_train)
```

3.2 모델 시각화

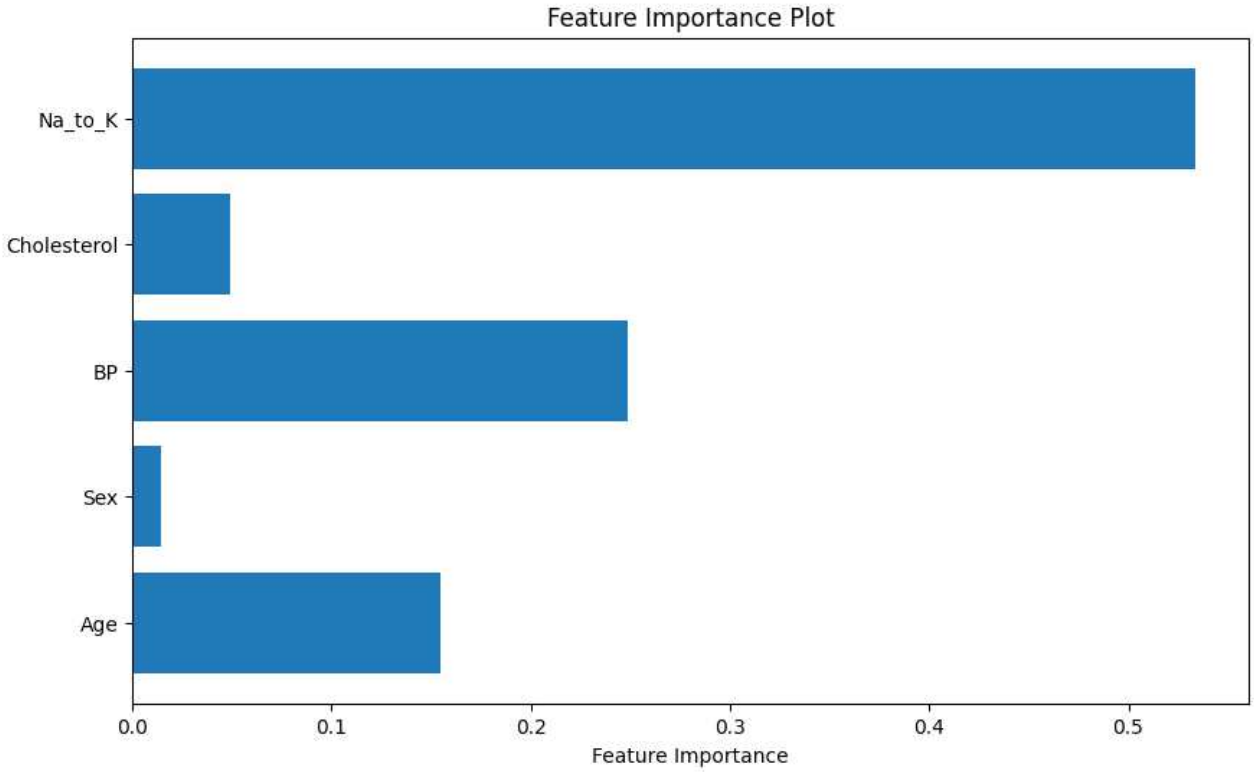
○ 특성 중요도 - RandomForest

```
# Feature Importance
feature_importances = pd.DataFrame({
    'Feature': X.columns,
    'Importance': model.feature_importances_
}).sort_values(by='Importance', ascending=False)

print("\nFeature Importances:")
print(feature_importances)
```

```
Feature Importances:
       Feature  Importance
4       Na_to_K    0.533735
2           BP    0.248586
0          Age    0.154324
3  Cholesterol    0.048865
1          Sex    0.014490
```

○ 특성 중요도 시각화

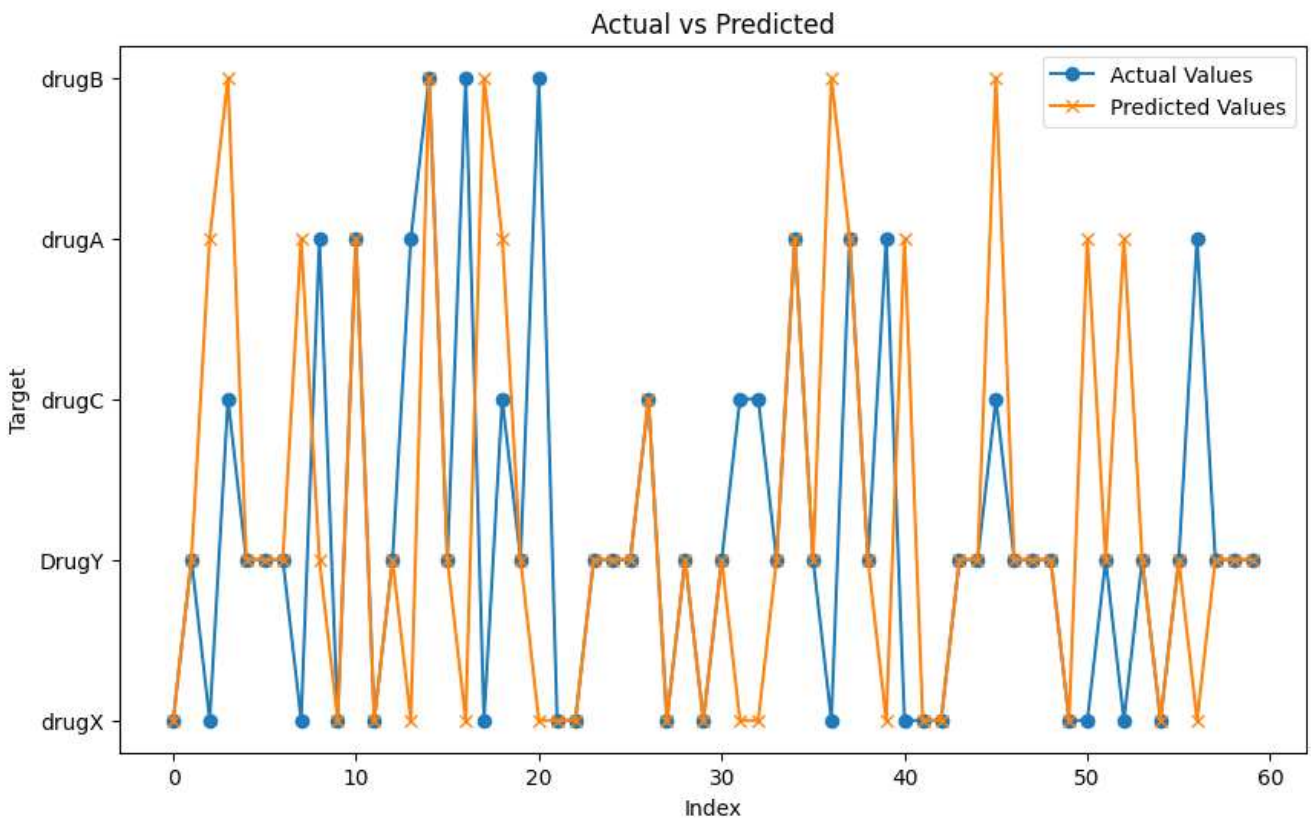
```
# visualize
importances = model.feature_importances_
features = X.columns

plt.figure(figsize=(10, 6))
plt.barh(features, importances)
plt.xlabel('Feature Importance')
plt.title('Feature Importance Plot')
plt.show()
```



3.3 모델 예측

○ 예측값 vs 실제값 비교

```
# 예측값과 실제값 비교 그래프
plt.figure(figsize=(10,6))
plt.plot(y_test.values, label='Actual Values', marker='o')
plt.plot(y_pred, label='Predicted Values', marker='x')
plt.legend()
plt.title('Actual vs Predicted')
plt.xlabel('Index')
plt.ylabel('Target')
plt.show()
```



4. 프로토타이핑(화면)

4.1 모델 예측

○ 사용자 입력값에 따른 약물 예측