Data Mining Project

의료약물 정보 간의 연관성 분석

생명공학과 202002134 오재은

바이오메디컬공학부 202104415 윤수진

LIST

Chapter. 1

Data set

Chapter. 2

Hypothesis

Chapter. 3

Analysis

Chapter. 4

Conclusion

Chapter. 5

Expected Effect

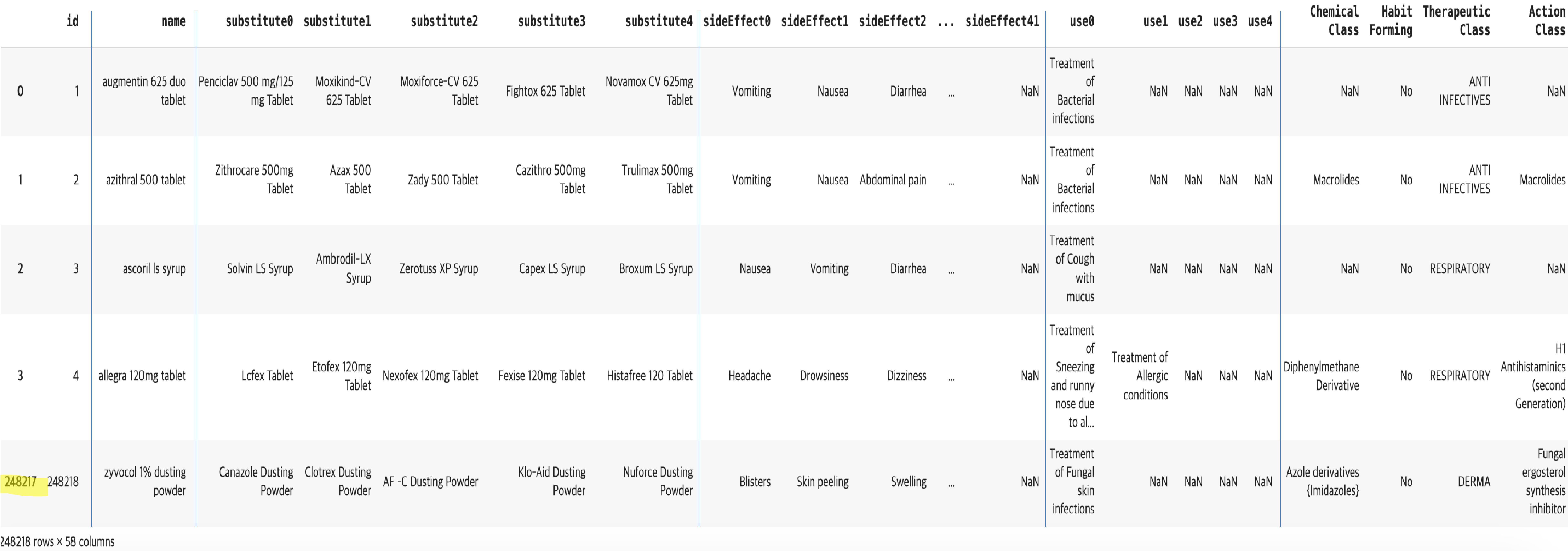
Chapter. 1

Data Set

큰 데이터 # 문자열 변수

“250k Medicines Usage, Side Effects and Substitutes”

ID Name Substitute 0~4 Side Effect 0~41 Use0~4 Chemical class / Habit Forming /Therapeutic class / Action Class



	id	name	substitute0	substitute1	substitute2	substitute3	substitute4	sideEffect0	sideEffect1	sideEffect2	...	sideEffect41	use0	use1	use2	use3	use4	Chemical Class	Habit Forming	Therapeutic Class	Action Class
0	1	augmentin 625 duo tablet	Penciclav 500 mg/125 mg Tablet	Moxikind-CV 625 Tablet	Moxiforce-CV 625 Tablet	Fightox 625 Tablet	Novamox CV 625mg Tablet	Vomiting	Nausea	Diarrhea	...	NaN	Treatment of Bacterial infections	NaN	NaN	NaN	NaN	NaN	No	ANTI INFECTIVES	NaN
1	2	azithral 500 tablet	Zithrocare 500mg Tablet	Azax 500 Tablet	Zady 500 Tablet	Cazithro 500mg Tablet	Trulimax 500mg Tablet	Vomiting	Nausea	Abdominal pain	...	NaN	Treatment of Bacterial infections	NaN	NaN	NaN	NaN	Macrolides	No	ANTI INFECTIVES	Macrolides
2	3	ascoril ls syrup	Solvin LS Syrup	Ambrodil-LX Syrup	Zerotuss XP Syrup	Capex LS Syrup	Broxum LS Syrup	Nausea	Vomiting	Diarrhea	...	NaN	Treatment of Cough with mucus	NaN	NaN	NaN	NaN	NaN	No	RESPIRATORY	NaN
3	4	allegra 120mg tablet	Lcfex Tablet	Etofex 120mg Tablet	Nexofex 120mg Tablet	Fexise 120mg Tablet	Histafree 120 Tablet	Headache	Drowsiness	Dizziness	...	NaN	Treatment of Sneezing and runny nose due to al...	Treatment of Allergic conditions	NaN	NaN	NaN	Diphenylmethane Derivative	No	RESPIRATORY	H1 Antihistaminics (second Generation)
248217	248218	zyvocol 1% dusting powder	Canazole Dusting Powder	Clotrex Dusting Powder	AF -C Dusting Powder	Klo-Aid Dusting Powder	Nuforce Dusting Powder	Blisters	Skin peeling	Swelling	...	NaN	Treatment of Fungal skin infections	NaN	NaN	NaN	NaN	Azole derivatives {Imidazoles}	No	DERMA	Fungal ergosterol synthesis inhibitor

248218 rows × 58 columns

248,218 rows × 58 columns

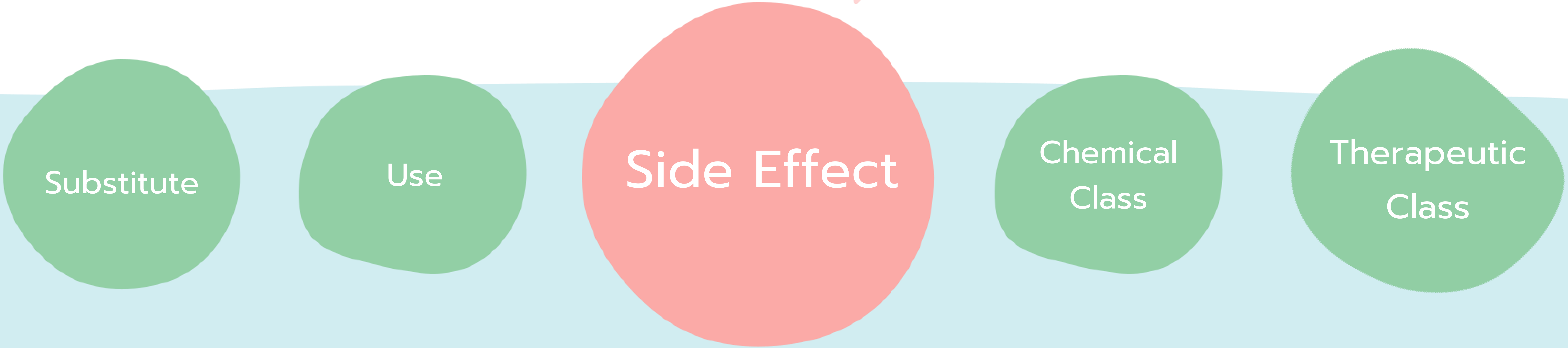
Dataset link: https://www.kaggle.com/datasets/shudhanshusingh/250k-medicines-usage-side-effects-and-substitutes

Chapter. 1

Motivation

“250k Medicines Usage, Side Effects and Substitutes”

의약품의 **부작용**에 밀접하게 영향을 미치는 **Attribute**는 무엇일까 ??


Substitute

Use

Side Effect

Chemical Class

Therapeutic Class

Chapter. 2

Hypothesis

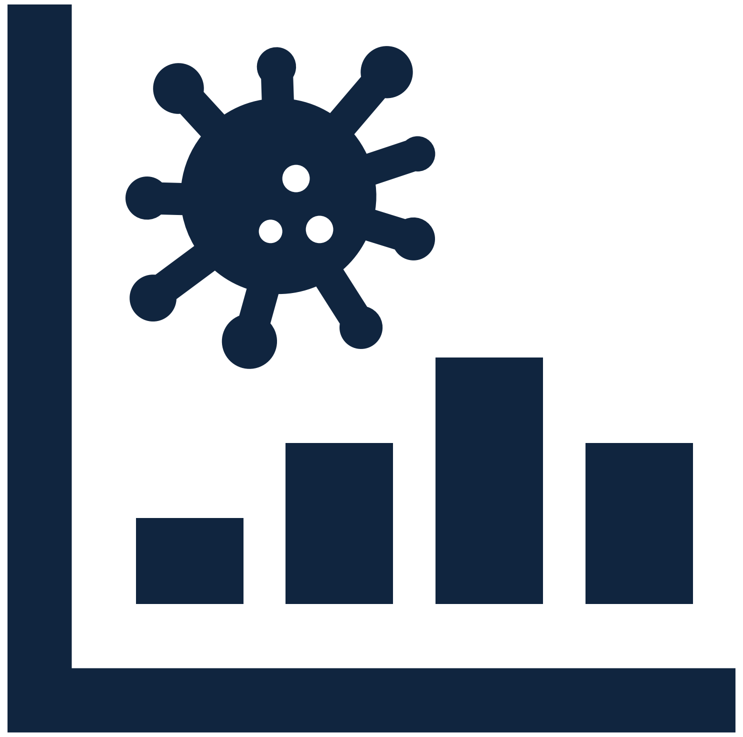
Side Effect <-> The Other Attribute

Side Effect를 중심으로 side effect와 특정 Attribute 사이에 연관성이 있다고 가정

=> **Regression**

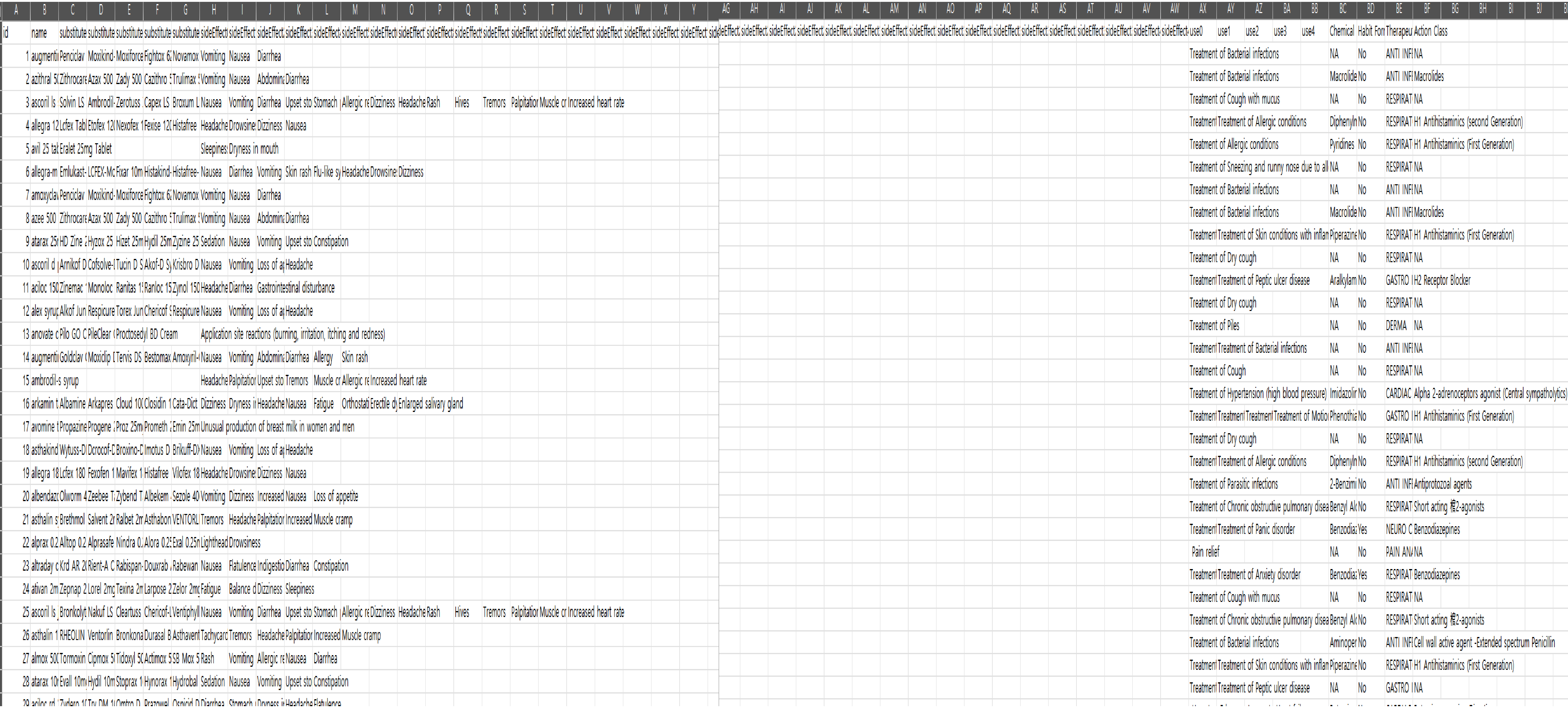
Chapter. 3

기법 선택 이유



=> Regression

전체 데이터에서 side effect에 영향을 주는 요인들을
분석하기 위해 regression model을 생성하여
side effect에 영향을 미치는 요인들에 대한
importance를 확인하고자 함

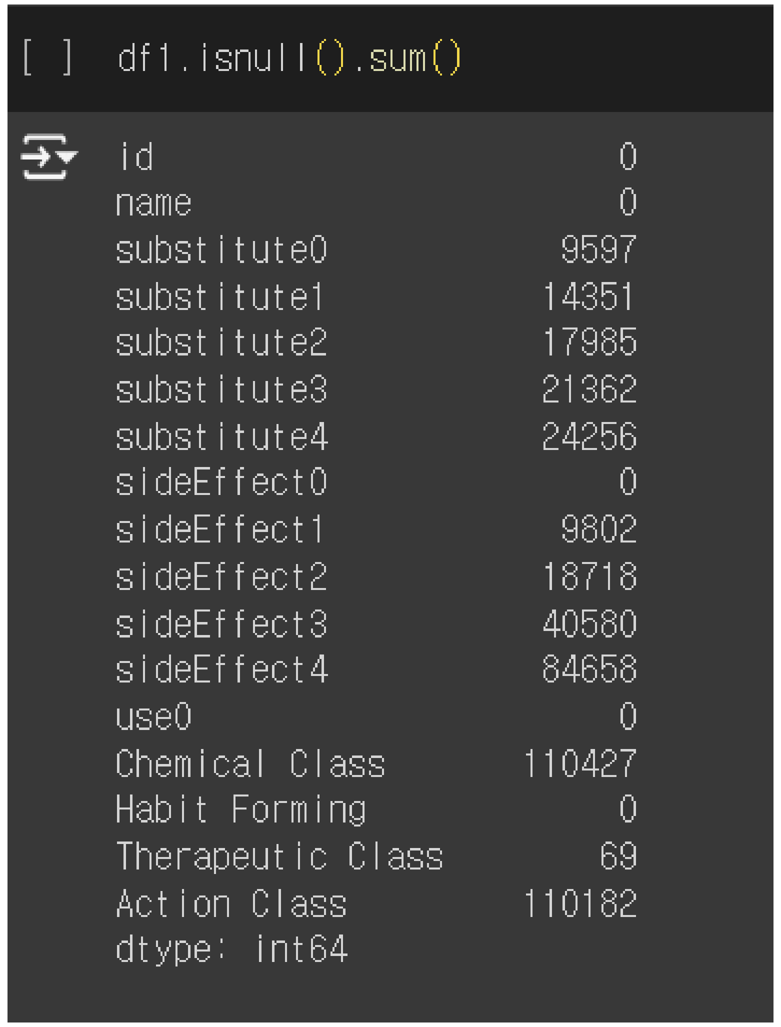


A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
id	name	substitute	substitute	substitute	substitute	substitute	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sid
1	augmenti	Penciclav	Moxikind-	Moxiforce	Fightox 6:	Novamox	Vomiting	Nausea	Diarrhea																
2	azithral 5(	Zithrocare	Azax 500	Zady 500	Cazithro 5	Trulimax 5	Vomiting	Nausea	Abdomina	Diarrhea															
3	ascoril ls	Solvin LS	Ambrodil-	Zerotuss	Capex LS	Broxum L	Nausea	Vomiting	Diarrhea	Upset sto	Stomach (	Allergic re	Dizziness	Headache	Rash	Hives	Tremors	Palpitatior	Muscle cr	Increased heart rate					
4	allegra 12	Lcfex Tabl	Etofex 12(	Nexofex 1	Fexise 12(	Histafree	Headache	Drowsine:	Dizziness	Nausea															
5	avil 25 tal	Eralet 25mg Tablet					Sleepines:	Dryness in mouth																	
6	allegra-m	Emlukast-	LCFEX-Mc	Fixar 10m	Histakind-	Histafree-	Nausea	Diarrhea	Vomiting	Skin rash	Flu-like sy	Headache	Drowsine:	Dizziness											
7	amoxyclav	Penciclav	Moxikind-	Moxiforce	Fightox 6:	Novamox	Vomiting	Nausea	Diarrhea																
8	azee 500	Zithrocare	Azax 500	Zady 500	Cazithro 5	Trulimax 5	Vomiting	Nausea	Abdomina	Diarrhea															
9	atarax 25r	HD Zine 2	Hyzox 25	Hizet 25m	Hydil 25m	Zyzine 25	Sedation	Nausea	Vomiting	Upset sto	Constipation														
10	ascoril d	Arnikof D	Cofsolve-(	Tucin D S	Akof-D Sy	Krisbro D	Nausea	Vomiting	Loss of ap	Headache															
11	aciloc 150	Zinemac 1	Monoloc	Ranitas 1!	Ranloc 15	Zynol 150	Headache	Diarrhea	Gastrointestinal disturbance																
12	alex syrup	Alkof Jun	Respicure	Torex Jun	Chericof S	Respicure	Nausea	Vomiting	Loss of ap	Headache															
13	anovate c	Pilo GO C	PileClear (	Proctosedyl	BD Cream		Application site reactions (burning, irritation, itching and redness)																		
14	augmenti	Goldclav (	Moxiclip [	Tervis DS	Bestomax	Amoxyril-(	Nausea	Vomiting	Abdomina	Diarrhea	Allergy	Skin rash													
15	ambrodil-s syrup						Headache	Palpitatior	Upset sto	Tremors	Muscle cr	Allergic re	Increased heart rate												
16	arkamin t	Albamine	Arkapres	Cloud 10(	Closidin 1	Cata-Dict	Dizziness	Dryness ii	Headache	Nausea	Fatigue	Orthostati	Erectile dy	Enlarged salivary gland											
17	avomine t	Propazine	Progene 2	Proz 25m	Prometh 2	Emin 25m	Unusual production of breast milk in women and men																		
18	asthakind	Wytuss-DI	Dcrocof-D	Broxino-D	Imotus D	Brikuff-D>	Nausea	Vomiting	Loss of ap	Headache															
19	allegra 18	Lcfex 180	Fexofen 1	Mavifex 1	Histafree	Vilofex 18	Headache	Drowsine:	Dizziness	Nausea															
20	albendazo	Olworm 4	Zeebee T:	Zybend T	Albekem -	Sezole 40	Vomiting	Dizziness	Increased	Nausea	Loss of appetite														
21	asthalin s	Brethmol	Salvent 2r	Ralbet 2m	Asthabon	VENTORLI	Tremors	Headache	Palpitatior	Increased	Muscle cramp														
22	alprax 0.2	Alltop 0.2	Alprasafe	Nindra 0.:	Alora 0.25	Exal 0.25n	Lighthead	Drowsiness																	
23	altraday c	Krd AR 2(	Rient-A C	Rabispan-	Douxrab /	Rabewan	Nausea	Flatulence	Indigestio	Diarrhea	Constipation														
24	ativan 2m	Zepnap 2	Lorel 2mg	Texina 2m	Larpose 2	Zelor 2mg	Fatigue	Balance d	Dizziness	Sleepiness															
25	ascoril ls_	Bronkolyt	Nakuf LS	Cleartuss	Chericof-L	Ventiphyll	Nausea	Vomiting	Diarrhea	Upset sto	Stomach (	Allergic re	Dizziness	Headache	Rash	Hives	Tremors	Palpitatior	Muscle cr	Increased heart rate					
26	asthalin 1	RHEOLIN	Ventorlin	Bronkona	Durasal B	Asthavent	Tachycard	Tremors	Headache	Palpitatior	Increased	Muscle cramp													
27	almox 50(	Tormoxin	Cipmox 5(	Tidoxyl 5(	Actimox 5	SB Mox 5	Rash	Vomiting	Allergic re	Nausea	Diarrhea														
28	atarax 10(	Evall 10m	Hydil 10m	Stoprax 1	Hynorax 1	Hydrobal	Sedation	Nausea	Vomiting	Upset sto	Constipation														
29	aciloc rd	Zydero 1(	Try DM 1(	Omtro D	Prazowel	Ospicid D	Diarrhea	Stomach (	Dryness ir	Headache	Flatulence														

AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ
deEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	sideEffect	use0	use1	use2	use3	use4	Chemical	Habit For	Therapeu	Action Class				
																	Treatment of Bacterial infections					NA	No	ANTI INFE	NA				
																	Treatment of Bacterial infections					Macrolide	No	ANTI INFE	Macrolides				
																	Treatment of Cough with mucus					NA	No	RESPIRAT	NA				
																	Treatment	Treatment of Allergic conditions				Diphenyln	No	RESPIRAT	H1 Antihistaminics (second Generation)				
																	Treatment of Allergic conditions					Pyridines	No	RESPIRAT	H1 Antihistaminics (First Generation)				
																	Treatment of Sneezing and runny nose due to all					NA	No	RESPIRAT	NA				
																	Treatment of Bacterial infections					NA	No	ANTI INFE	NA				
																	Treatment of Bacterial infections					Macrolide	No	ANTI INFE	Macrolides				
																	Treatment	Treatment of Skin conditions with inflan				Piperazine	No	RESPIRAT	H1 Antihistaminics (First Generation)				
																	Treatment of Dry cough					NA	No	RESPIRAT	NA				
																	Treatment	Treatment of Peptic ulcer disease				Aralkylam	No	GASTRO I	H2 Receptor Blocker				
																	Treatment of Dry cough					NA	No	RESPIRAT	NA				
																	Treatment of Piles					NA	No	DERMA	NA				
																	Treatment	Treatment of Bacterial infections				NA	No	ANTI INFE	NA				
																	Treatment of Cough					NA	No	RESPIRAT	NA				
																	Treatment of Hypertension (high blood pressure)					Imidazolir	No	CARDIAC	Alpha 2-adrenoceptors agonist (Central sympatholytics)				
																	Treatment	Treatment	Treatment	Treatment of Motio		Phenothia	No	GASTRO I	H1 Antihistaminics (First Generation)				
																	Treatment of Dry cough					NA	No	RESPIRAT	NA				
																	Treatment	Treatment of Allergic conditions				Diphenyln	No	RESPIRAT	H1 Antihistaminics (second Generation)				
																	Treatment of Parasitic infections					2-Benzimi	No	ANTI INFE	Antiprotozoal agents				
																	Treatment of Chronic obstructive pulmonary disea					Benzyl Alc	No	RESPIRAT	Short acting 棺2-agonists				
																	Treatment	Treatment of Panic disorder				Benzodia:	Yes	NEURO C	Benzodiazepines				
																	Pain relief					NA	No	PAIN ANA	NA				
																	Treatment	Treatment of Anxiety disorder				Benzodia:	Yes	RESPIRAT	Benzodiazepines				
																	Treatment of Cough with mucus					NA	No	RESPIRAT	NA				
																	Treatment of Chronic obstructive pulmonary disea					Benzyl Alc	No	RESPIRAT	Short acting 棺2-agonists				
																	Treatment of Bacterial infections					Aminoper	No	ANTI INFE	Cell wall active agent -Extended spectrum Penicillin				
																	Treatment	Treatment of Skin conditions with inflan				Piperazine	No	RESPIRAT	H1 Antihistaminics (First Generation)				
																	Treatment	Treatment of Peptic ulcer disease				NA	No	GASTRO I	NA				
																	[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]				

Data Preprocessing

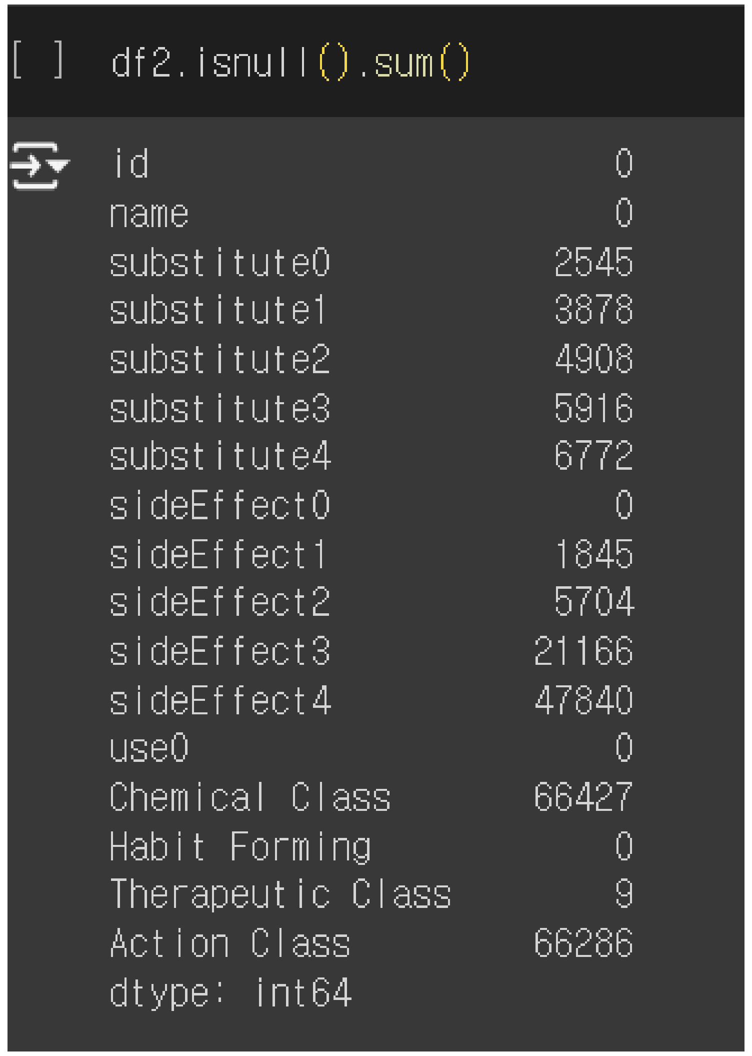
1. 각 attribute 마다 missing value의 비율이 45% 이상이면 attribute 제거
→ dimension 축소



```
[ ]  df1.isnull().sum()

     id                      0
     name                    0
     substitute0          9597
     substitute1         14351
     substitute2         17985
     substitute3         21362
     substitute4         24256
     sideEffect0             0
     sideEffect1          9802
     sideEffect2         18718
     sideEffect3         40580
     sideEffect4         84658
     use0                    0
     Chemical Class     110427
     Habit Forming           0
     Therapeutic Class      69
     Action Class       110182
     dtype: int64
```

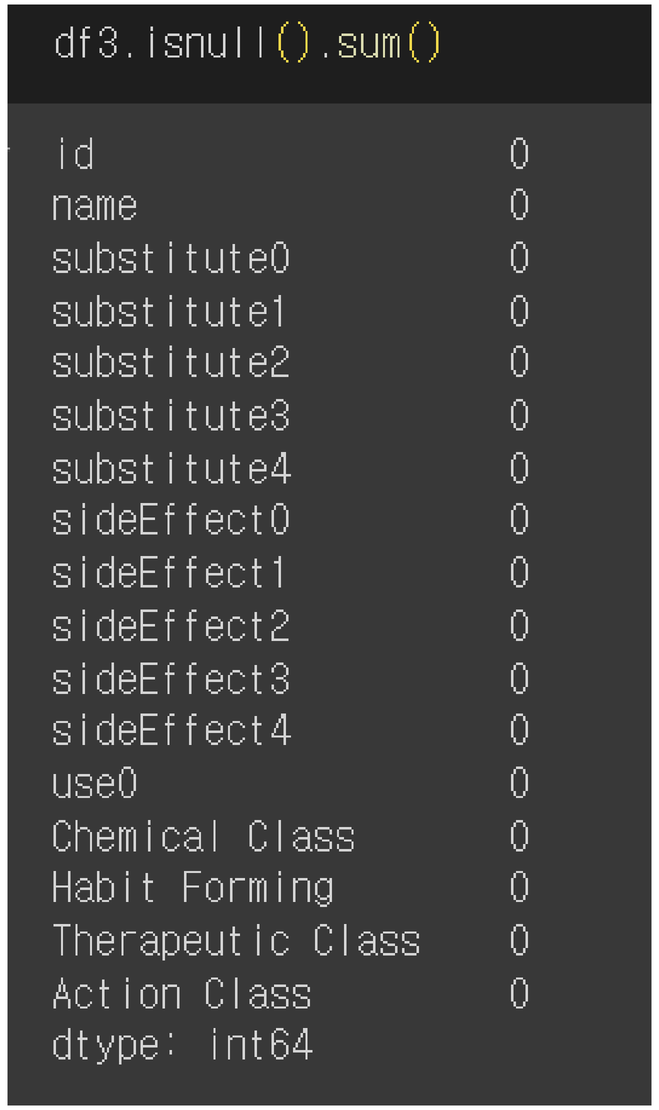
2. Use0(약물 사용 목적) 에서 빈도가 높은 상위 15개의 사용 목적만 남기고 나머지 값은 제거



```
[ ]  df2.isnull().sum()

     id                      0
     name                    0
     substitute0          2545
     substitute1          3878
     substitute2          4908
     substitute3          5916
     substitute4          6772
     sideEffect0             0
     sideEffect1          1845
     sideEffect2          5704
     sideEffect3         21166
     sideEffect4         47840
     use0                    0
     Chemical Class      66427
     Habit Forming           0
     Therapeutic Class       9
     Action Class        66286
     dtype: int64
```

3. 나머지 null 값 확인 후 전부 제거



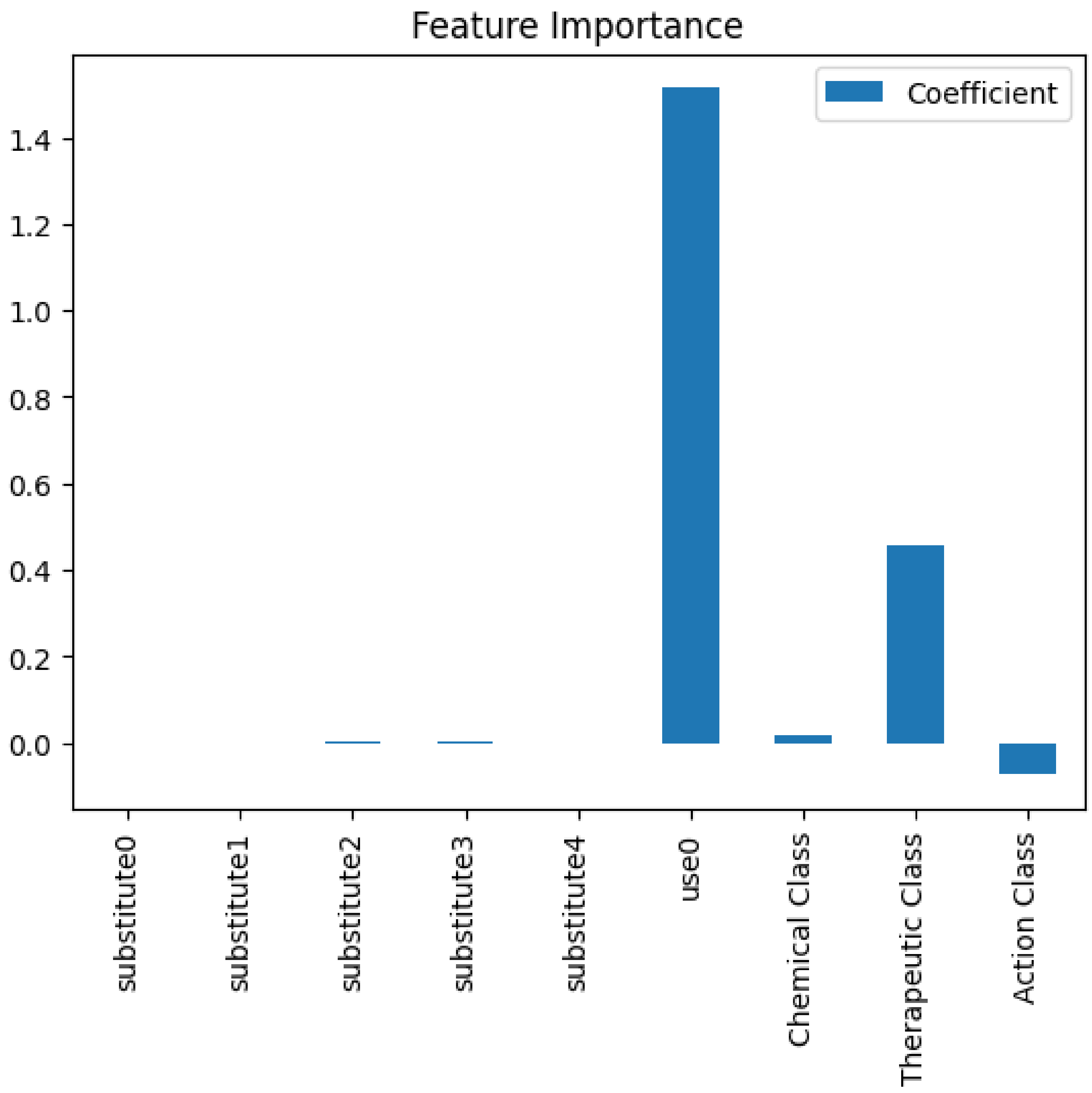
```
df3.isnull().sum()

id                   0
name                 0
substitute0          0
substitute1          0
substitute2          0
substitute3          0
substitute4          0
sideEffect0          0
sideEffect1          0
sideEffect2          0
sideEffect3          0
sideEffect4          0
use0                 0
Chemical Class       0
Habit Forming        0
Therapeutic Class    0
Action Class         0
dtype: int64
```

4. Label encoding
- 모든 문자열 → 숫자로 변환

5. 전처리 후 Data Shape : (34859,16)

Linear Regression



Mean Squared Error: 128.46215053720954

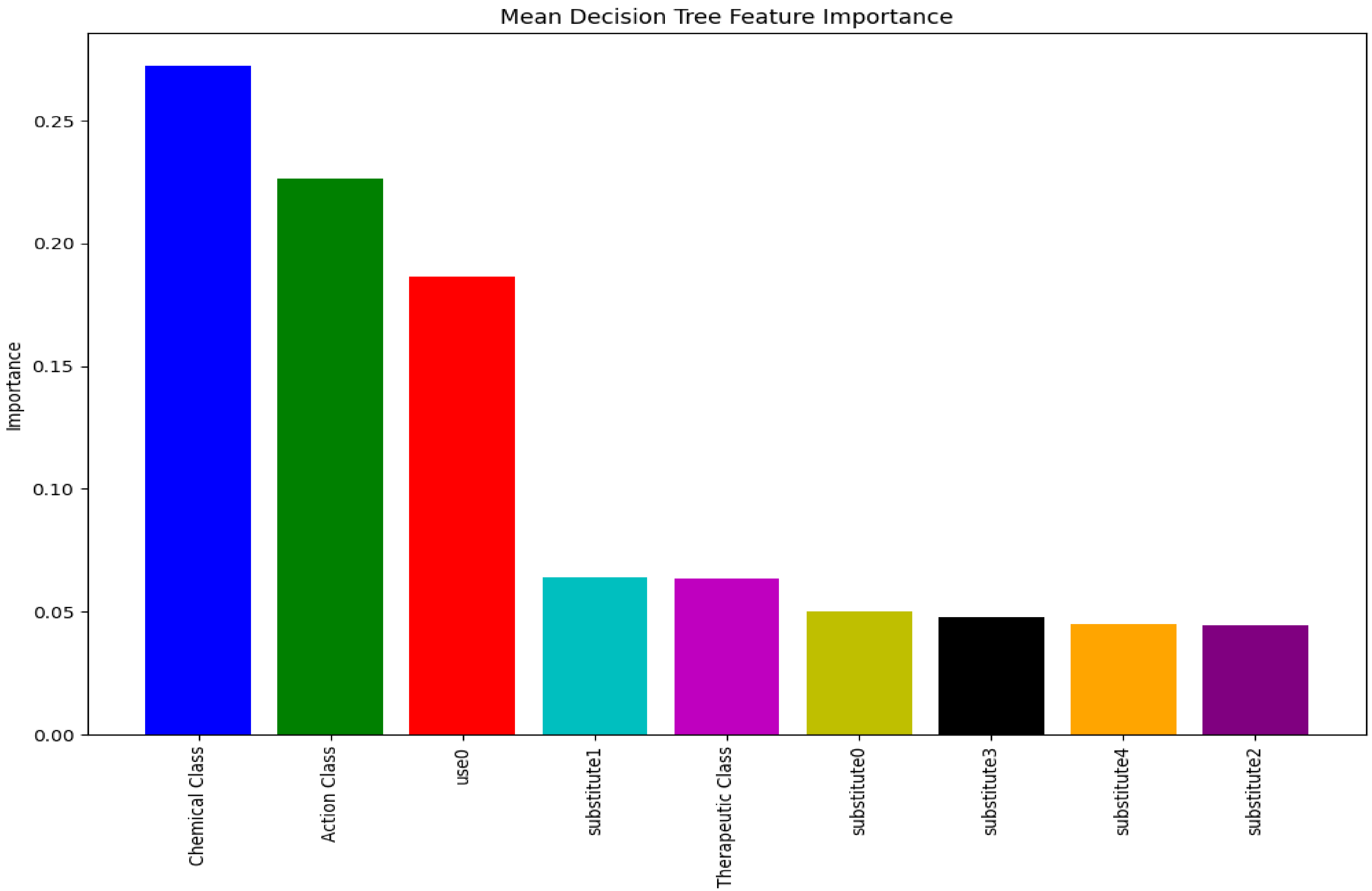
R^2 Score: 0.20824906022893785

	Coefficient
substitute0	-0.002113
substitute1	-0.001490
substitute2	0.001770
substitute3	0.000999
substitute4	-0.000981
use0	1.515184
Chemical Class	0.016376
Therapeutic Class	0.454804
Action Class	-0.074955

한계: R^2 Score 값이 0.2로 낮음

Decision Tree

Side effect 0~4까지 총 5개의 모델을 생성 → 수치의 평균을 내어 Importance 시각화



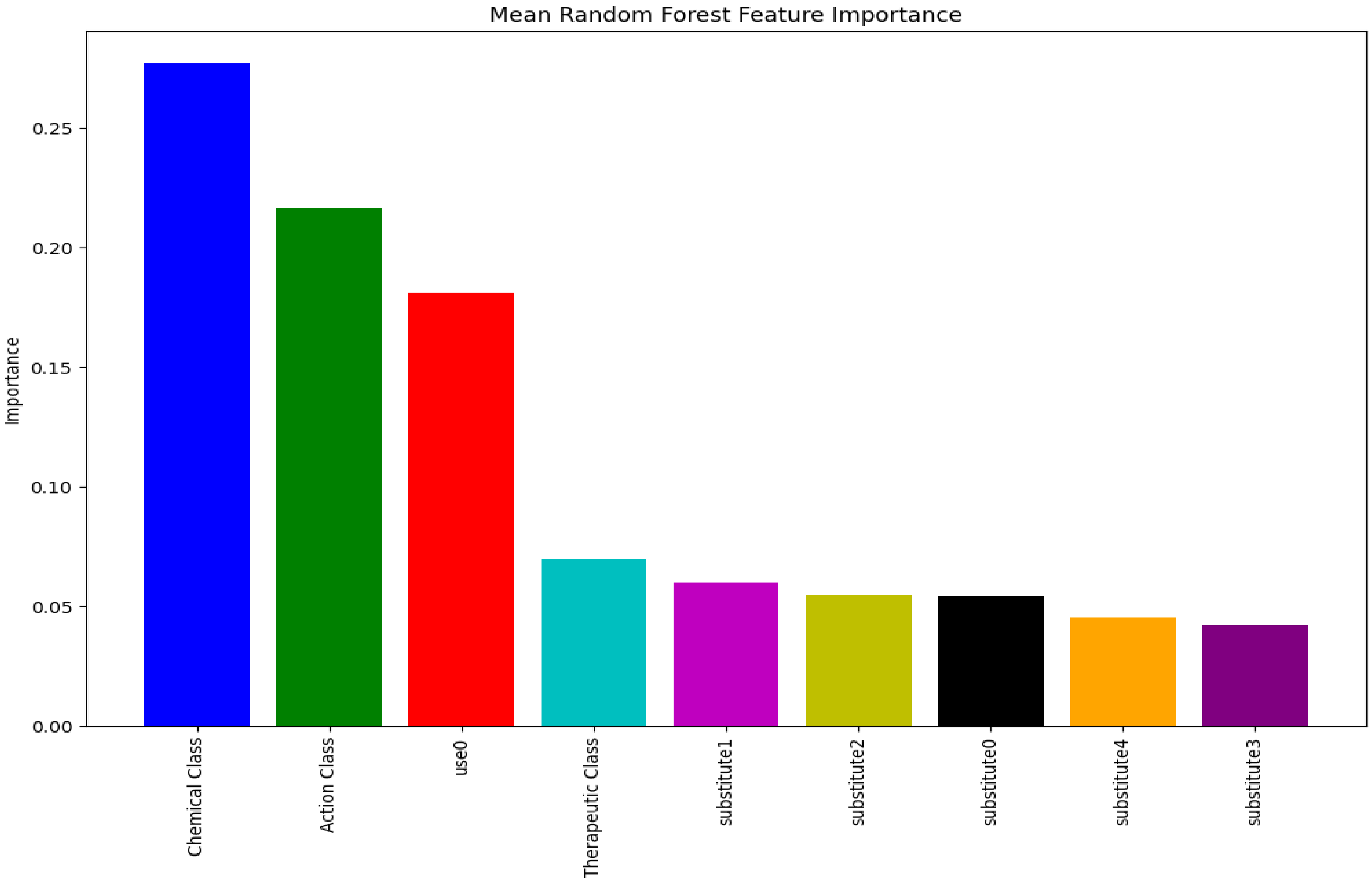
Decision Tree Feature Importance:

	Importance
Chemical Class	0.272018
Action Class	0.226378
use0	0.186317
substitute1	0.064070
Therapeutic Class	0.063603
substitute0	0.049972
substitute3	0.047883
substitute4	0.045149
substitute2	0.044610

두 경우 모두 Chemical Class, Action Class 가 높은 Importance로 나타남

Random Forest

Side effect 0~4까지 총 5개의 모델을 생성 → 수치의 평균을 내어 Importance 시각화

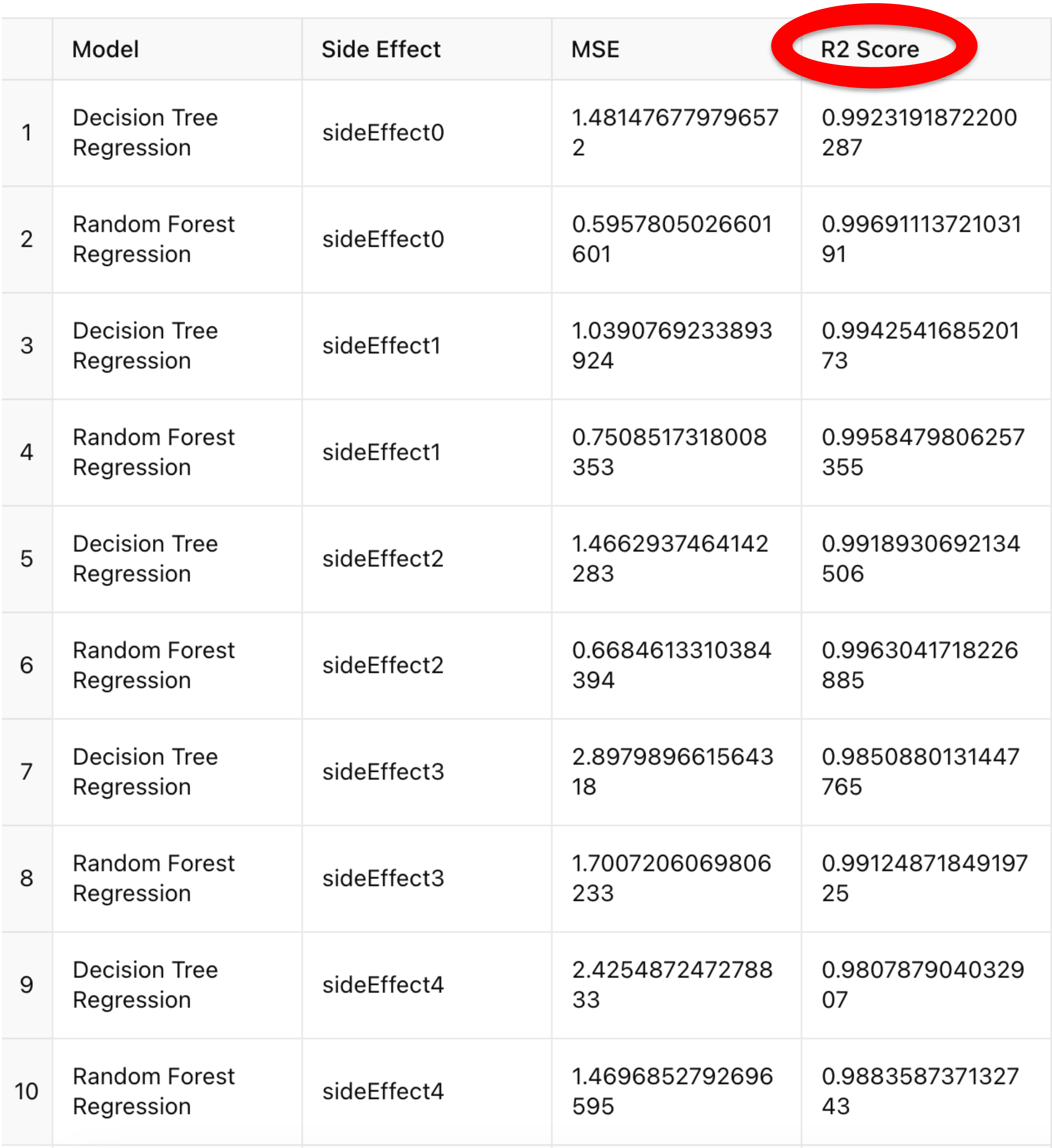


Random Forest Feature Importance:

	Importance
Chemical Class	0.276567
Action Class	0.216396
use0	0.181050
Therapeutic Class	0.069960
substitute1	0.059942
substitute2	0.054579
substitute0	0.054125
substitute4	0.045281
substitute3	0.042101

두 경우 모두 Chemical Class, Action Class 가 높은 Importance로 나타남

Decision Tree & Random Forest



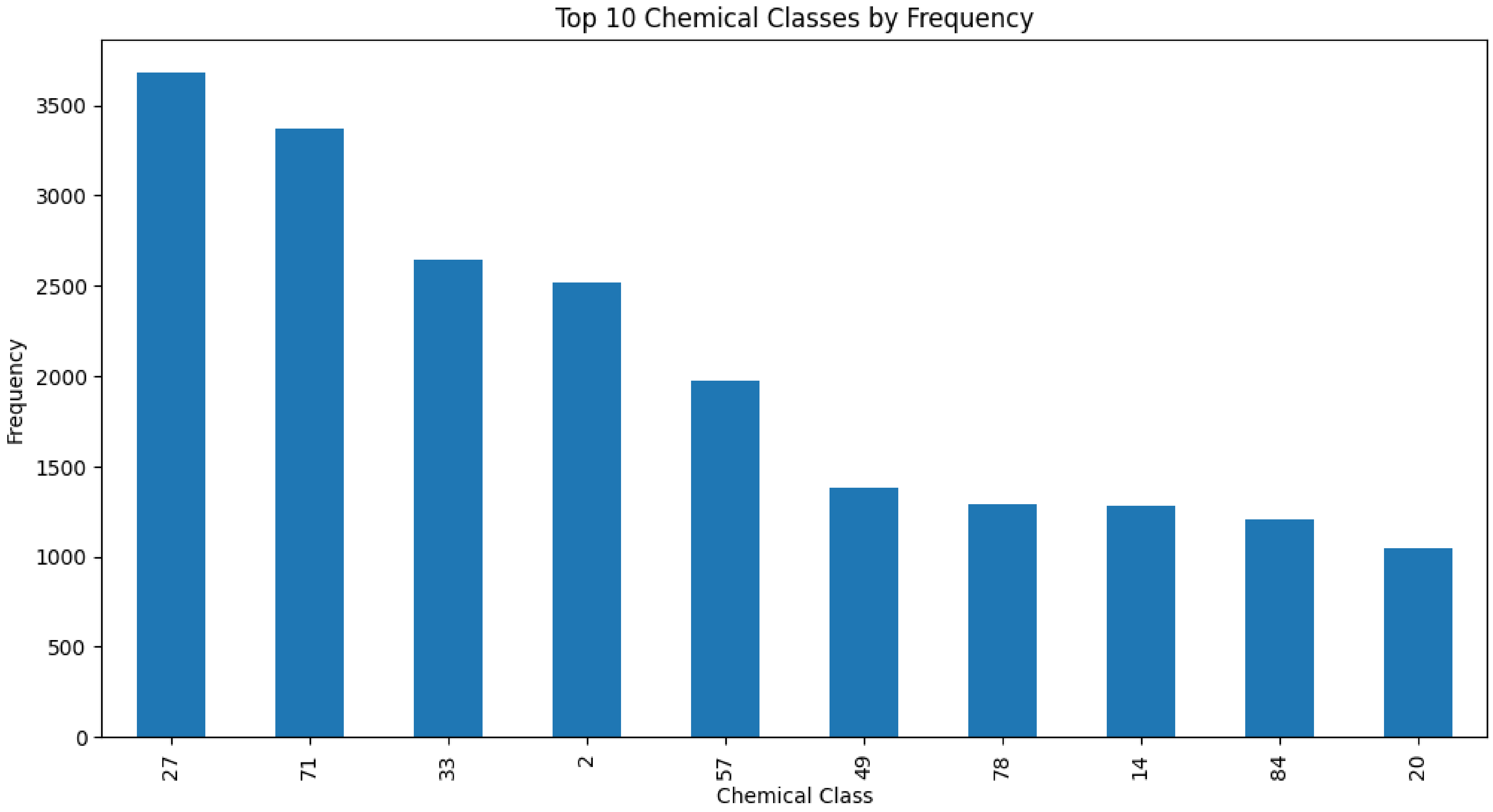
	Model	Side Effect	MSE	R2 Score
1	Decision Tree Regression	sideEffect0	1.481476779796572	0.9923191872200287
2	Random Forest Regression	sideEffect0	0.5957805026601601	0.9969111372103191
3	Decision Tree Regression	sideEffect1	1.0390769233893924	0.994254168520173
4	Random Forest Regression	sideEffect1	0.7508517318008353	0.9958479806257355
5	Decision Tree Regression	sideEffect2	1.4662937464142283	0.9918930692134506
6	Random Forest Regression	sideEffect2	0.6684613310384394	0.9963041718226885
7	Decision Tree Regression	sideEffect3	2.897989661564318	0.9850880131447765
8	Random Forest Regression	sideEffect3	1.7007206069806233	0.9912487184919725
9	Decision Tree Regression	sideEffect4	2.425487247278833	0.980787904032907
10	Random Forest Regression	sideEffect4	1.4696852792696595	0.988358737132743

Side effect 0,1,2,3,4의 R^2 score 값이 모두 높음

* R square : 독립변수가 종속변수를 얼마나 잘 설명하는지를 나타냄 (0에서 1사이의 값)

Chapter. 3

Chemical Class & Side effect



1. Chemical Class 내에서 빈도 수 체크

→ 27, 71, 33이라는 chemical class순으로 가장 많이 존재

2. 특정 chemical class 내에서 side effect 빈도 수 체크

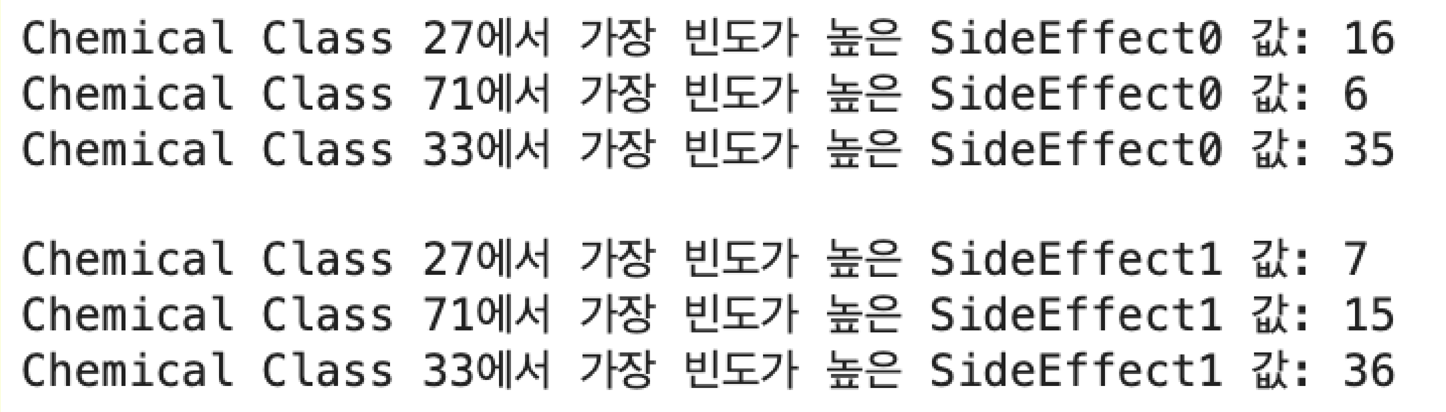
→ Chemical Class 27에서 가장 빈도가 높은 SideEffect 값: 16

Chemical Class 71에서 가장 빈도가 높은 SideEffect 값: 6

Chemical Class 33에서 가장 빈도가 높은 SideEffect 값: 35

Label encoding 값을 다시 decoding하면 어떤 chemical class가 어떤 side effect 발생에 영향을 미치는지에 대한 이해를 도울 수 있음

Data Decoding



```
Chemical Class 27에서 가장 빈도가 높은 SideEffect0 값: 16
Chemical Class 71에서 가장 빈도가 높은 SideEffect0 값: 6
Chemical Class 33에서 가장 빈도가 높은 SideEffect0 값: 35

Chemical Class 27에서 가장 빈도가 높은 SideEffect1 값: 7
Chemical Class 71에서 가장 빈도가 높은 SideEffect1 값: 15
Chemical Class 33에서 가장 빈도가 높은 SideEffect1 값: 36
```

```
Chemical Class 27에서 가장 빈도가 높은 SideEffect0 값: 16 [Akathisia (inability to stay still)]
Chemical Class 71에서 가장 빈도가 높은 SideEffect0 값: 6 [Abnormal liver function tests]
Chemical Class 33에서 가장 빈도가 높은 SideEffect0 값: 35 [Argyria (skin turns blue or bluish-grey)]

Chemical Class 27에서 가장 빈도가 높은 SideEffect1 값: 7 [Accomodation disorder]
Chemical Class 71에서 가장 빈도가 높은 SideEffect1 값: 15 [Allergic skin rash]
Chemical Class 33에서 가장 빈도가 높은 SideEffect1 값: 36 [Blisters on skin]

Chemical Class 27: [Acetic acid Derivatives]
Chemical Class 71: [Amino Acids, Peptides, and Analogues]
Chemical Class 33: [Acylsalicylamides (Benzoic acids and derivatives)]
```

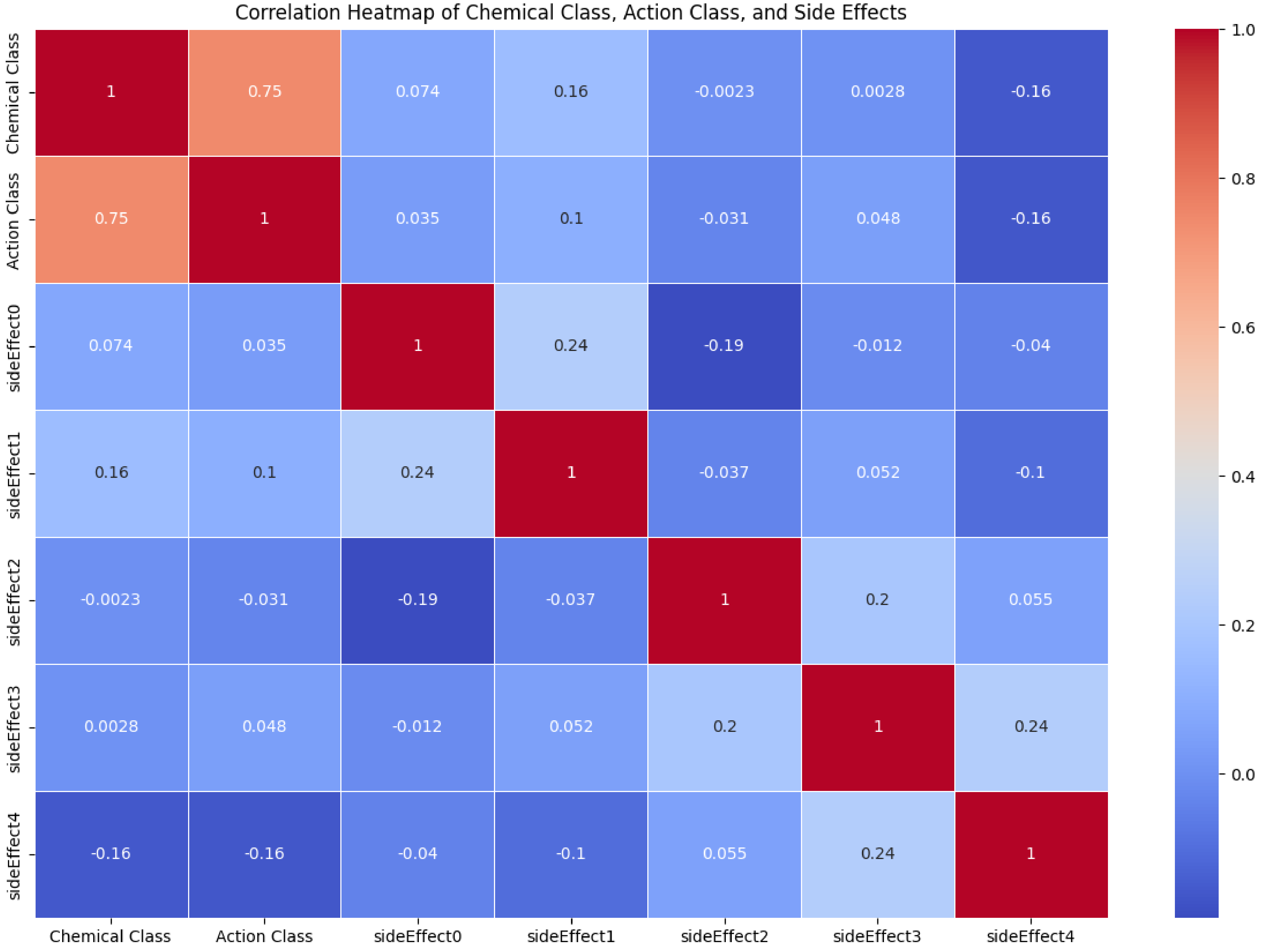
Chapter. 3

Chemical<->Action<->Side Effect

Correlation matrix Heat Map

- **Action class**와 **Chemical class** 사이의 상관관계가 높은 것으로 보아 서로 **종속적인 관계**가 있는 것으로 추측

- **다른** 독립 변수들도 연관성의 요인으로 고려해보는 것이 필요



Correlation Heatmap of Chemical Class, Action Class, and Side Effects

	Chemical Class	Action Class	sideEffect0	sideEffect1	sideEffect2	sideEffect3	sideEffect4
Chemical Class	1	0.75	0.074	0.16	-0.0023	0.0028	-0.16
Action Class	0.75	1	0.035	0.1	-0.031	0.048	-0.16
sideEffect0	0.074	0.035	1	0.24	-0.19	-0.012	-0.04
sideEffect1	0.16	0.1	0.24	1	-0.037	0.052	-0.1
sideEffect2	-0.0023	-0.031	-0.19	-0.037	1	0.2	0.055
sideEffect3	0.0028	0.048	-0.012	0.052	0.2	1	0.24
sideEffect4	-0.16	-0.16	-0.04	-0.1	0.055	0.24	1

Chapter. 4

Conclusion & 한계점

결론

- 'Chemical class' 와 'Action class'가 약물의 부작용에 높은 중요도를 가진다.

한계점

- 문자열 데이터 & 많은 Null 값으로 전처리 과정에 어려움, 다양한 기법을 사용해도 유의미한 결과를 찾는 데에 많은 시간 소요 발생
- Label encoding한 수치들을 다시 Decoding 하는 과정이 필요

Chapter. 5

Expected Effect

기대효과 1

* 부작용이 약물의 어떤 특성과 연관이 있는지 확인

=> 의약품 연구에 활용 가능.

기대효과 2

* 연관성 있는 attribute의 종류에 따라 어떤 부작용이 발생할 수 있는지 예측

=> 대체 의약품 개발에 도움.

THANK YOU