Improving Medication Safety

Analyzing adverse drug reactions on patients

STAT GR5243 Project 1

Stella Dai, Liang Zhao, Siwei Chen, Wenwei Kuang

Introduction

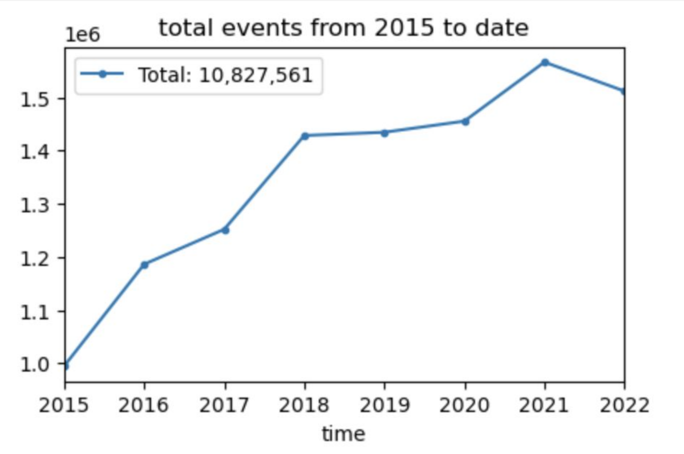
Motivation

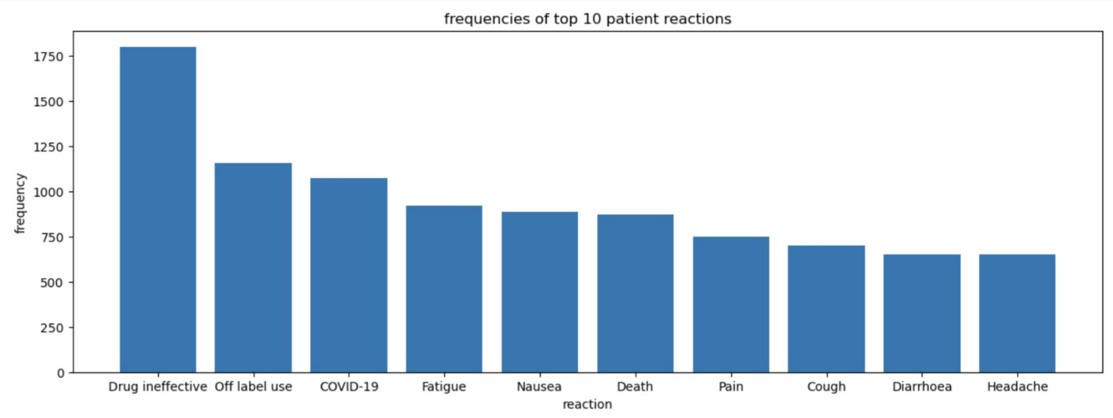
- Adverse drug reactions (ADRs) remain a challenge in modern healthcare field.
- In this project, we will be using FDA's Adverse Drug Events Database to explore the side effects and ADRs among the global FDA-approved drugs.
- Our goal is to investigate the adverse reactions experienced by patients and thus boost medication safety.
- In order to achieve the goal, we will develop effective machine learning models to analyze and predict the seriousness of adverse reaction results as the response variable, using the background information of patients and the different types of drugs taken.
- In other words, how to keep patients safe while taking drugs as a treatment?

EDA

Overview

- Number of medical records over years in the FDA database
- The latest 26000 records from Open FDA API including 27 features
- Top 10 adverse reactions: 14.633% of all reactions

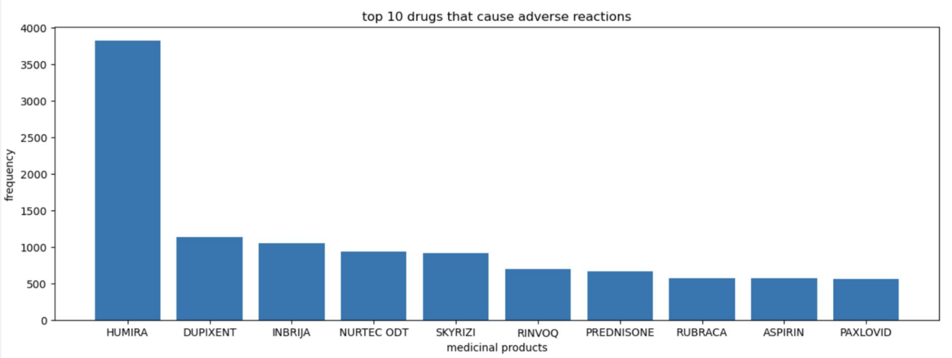


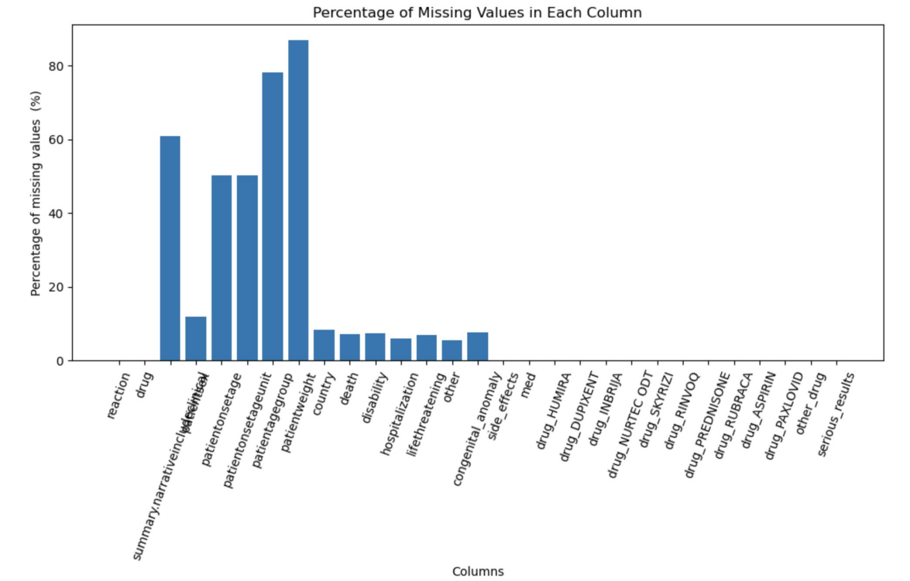


EDA

Overview and Data Wrangling

- Top 10 drugs (16.4678% of all drugs) that cause adverse reactions
- Missing values – mainly in columns not used for modeling

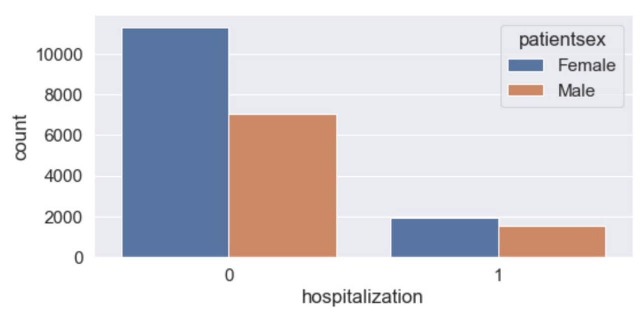


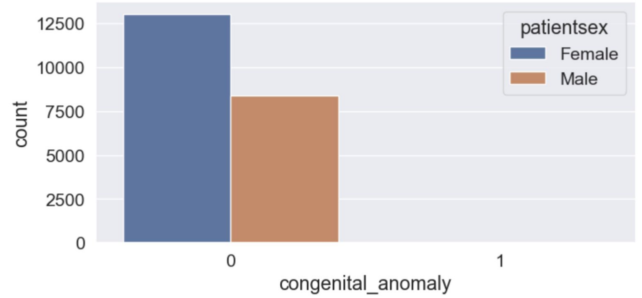


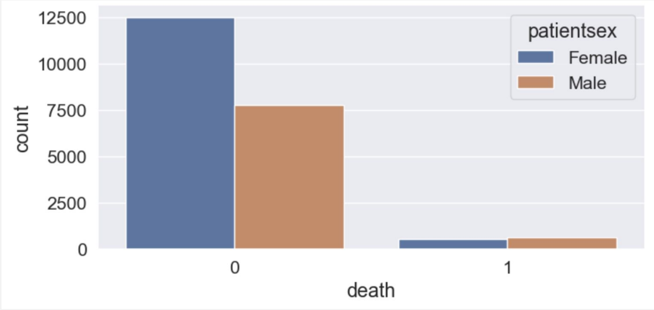
EDA

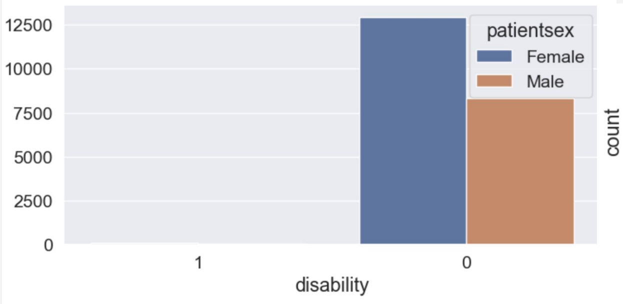
Overview and Data Wrangling

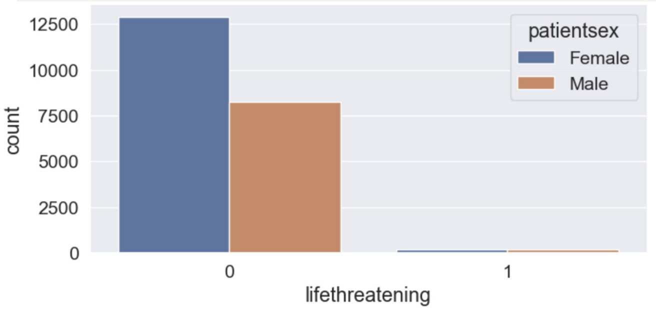
- Response variable “seriousness” includes the following adverse reactions
- Potential bias: gender — 60% of records are females vs 40% males
- 0 stands for non-severe effects
- Gender has an effect







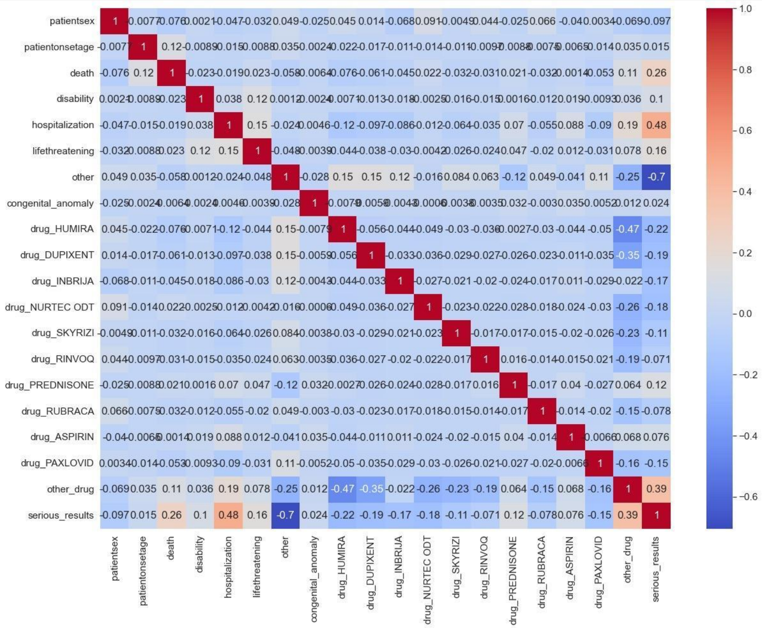


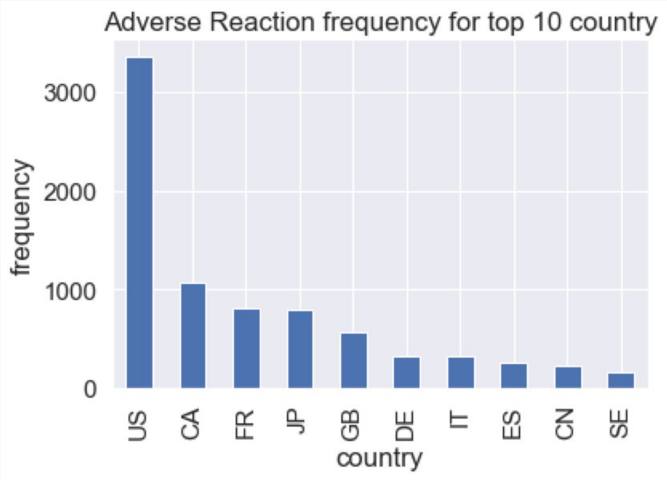


EDA

Overview and Data Wrangling

- Frequency of adverse reactions for top 10 countries
- Correlation matrix: all below 0.7 — no collinearity problems





Data Preprocessing

Data cleaning:

- Remove NaN values for predictor variables and the unnecessary columns
- Standardize the age group using min_max scaler
- Check for duplicates
- Create dummy variables for top 10 drugs (and other drugs) using one-hot encoding

Train_test_split:

- Stratify based on the distribution of male and female (6:4)

```
#train-test split: using stratify
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,stratify=y, random_state=0, train_size=0.8)
```

Data Preprocessing

Final Dataframe for Modeling

- Our dataframe contains 12578 observations and 14 variables.
- We aim to use patient age, sex, and type of drugs intake to predict the seriousness level of adverse reaction.

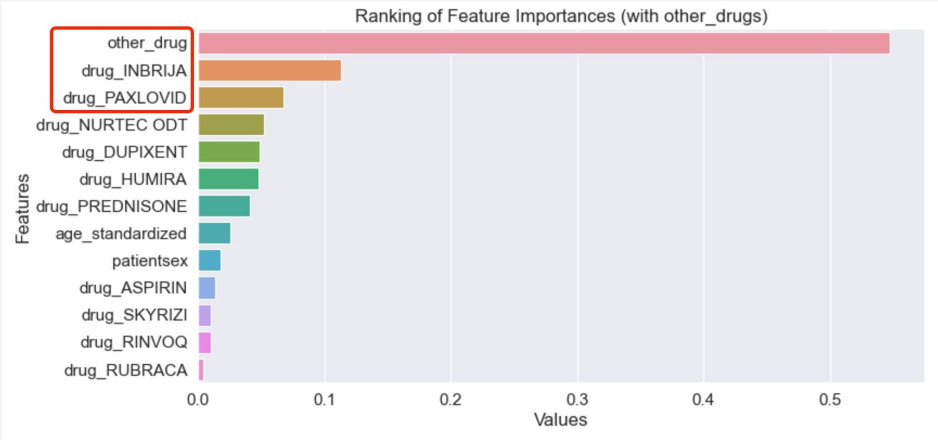
	serious_results	age_label	patientsex	drug_HUMIRA	drug_DUPIXENT	drug_INBRIJA	drug_NURTEC ODT	drug_SKYRIZI	drug_RINVOQ	drug_PREDNISONE
3	1	4	0	0	0	0	0	0	0	0
8	0	4	1	0	0	0	0	0	0	0
9	1	4	0	0	0	0	0	0	0	0
10	1	5	0	0	0	0	0	0	0	0
12	0	2	1	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
25994	1	4	1	0	0	0	0	0	0	0
25995	1	4	1	0	0	0	0	0	0	0
25996	1	5	0	0	0	0	0	0	0	0
25997	1	4	1	0	0	0	0	0	0	0
25999	0	5	1	0	0	0	0	0	0	0

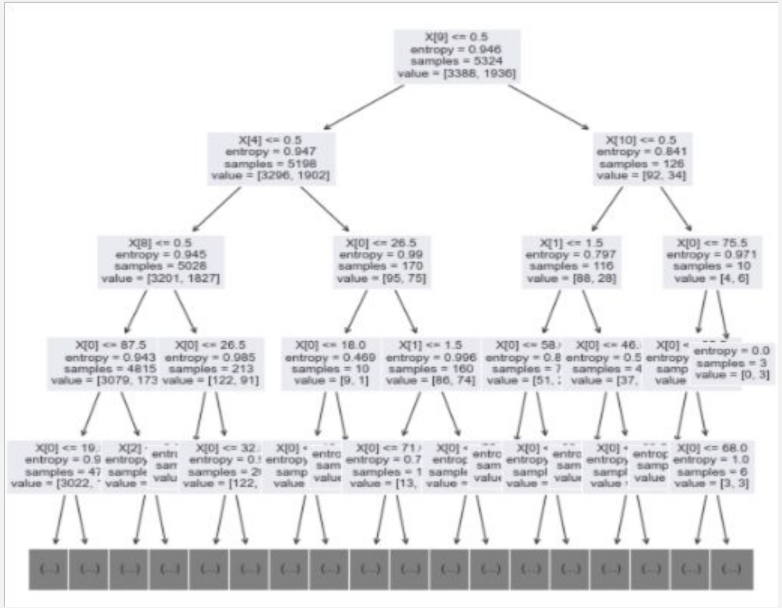
12578 rows × 14 columns

Data Modeling

Decision Tree

- Decision Tree Accuracy: 0.7337
- Fit the model hyperparameters based on the Grid Search & CV:
    - best{'max_depth': 6, 'min_samples_leaf': 2, 'min_samples_split': 2}
- Feature Importance of top 3 attributes:
    - other_drug
    - drug_INBRIJA
    - drug_PAXLOVID (Covid-19)
    - drug_NURTEC ODT

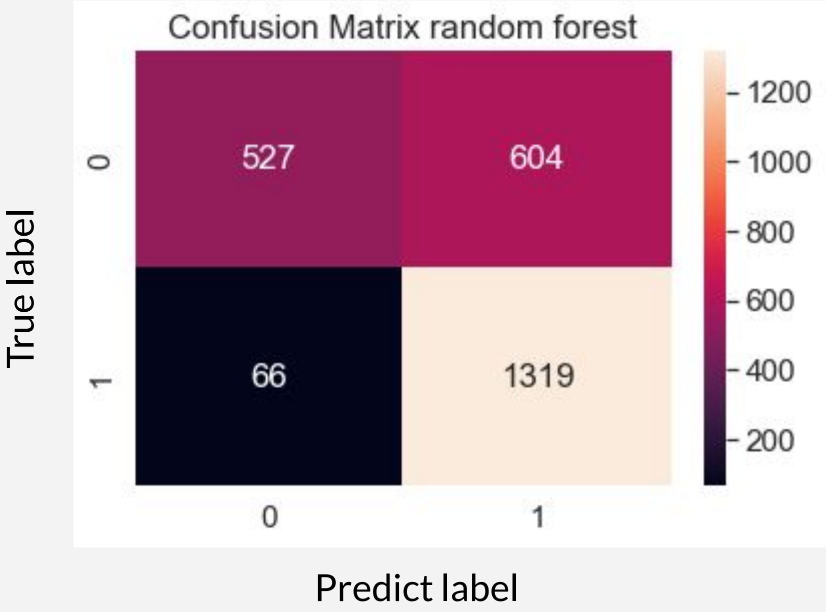




Data Modeling

Random Forest

- Random Forest Accuracy: 0.7337
- Precision: 0.6859
- Recall: 0.9523
- F1 score: 0.7974
- Random Forest MSE: 0.2662

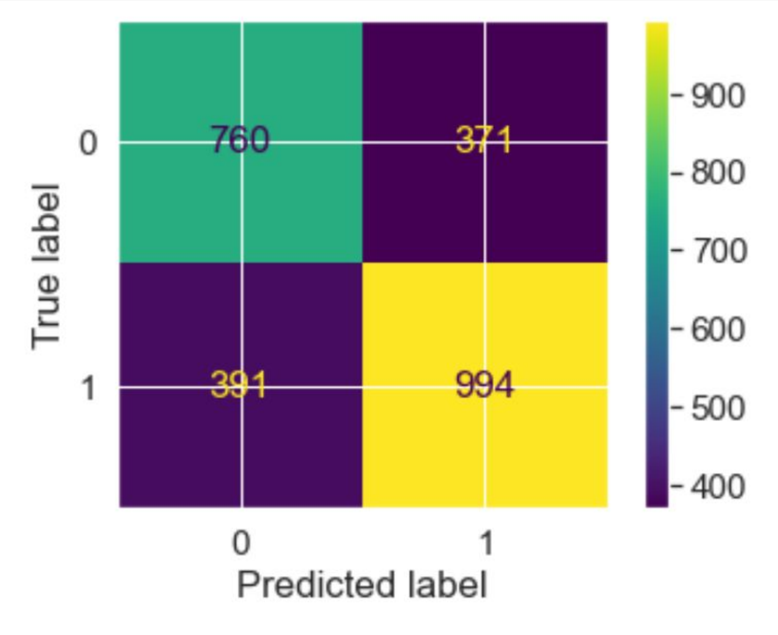


Data Modeling

KNN (K-Nearest Neighbors Algorithm)

- Accuracy: 0.6971
- Precision: 0.7282
- Recall: 0.7176
- F1 score: 0.8038
- Problem:

  Large number of false positive cases: serious adverse reaction(1) predicted to be non-serious(0).



Data Modeling

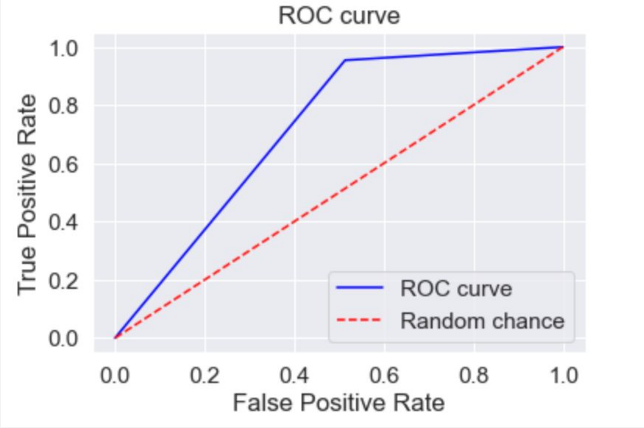
Logistic Regression

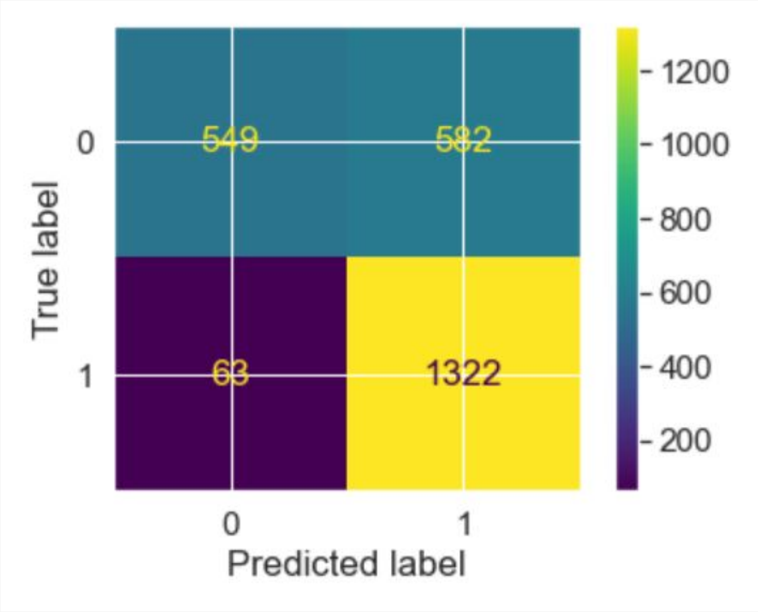
- High accuracy: 0.7345
- Few false positive cases: 63/2516 = 0.025
- High precision rate (TP/TP+FP): 0.796
- AUC: 0.72

```
logr mean cv accuracy: 0.7345
Precision:  0.7959596721584106
F1 score:  0.7174350989977455
              precision    recall  f1-score   support

           0       0.90      0.49      0.63      1131
           1       0.69      0.95      0.80      1385

    accuracy                           0.74      2516
   macro avg       0.80      0.72      0.72      2516
weighted avg       0.79      0.74      0.73      2516
```

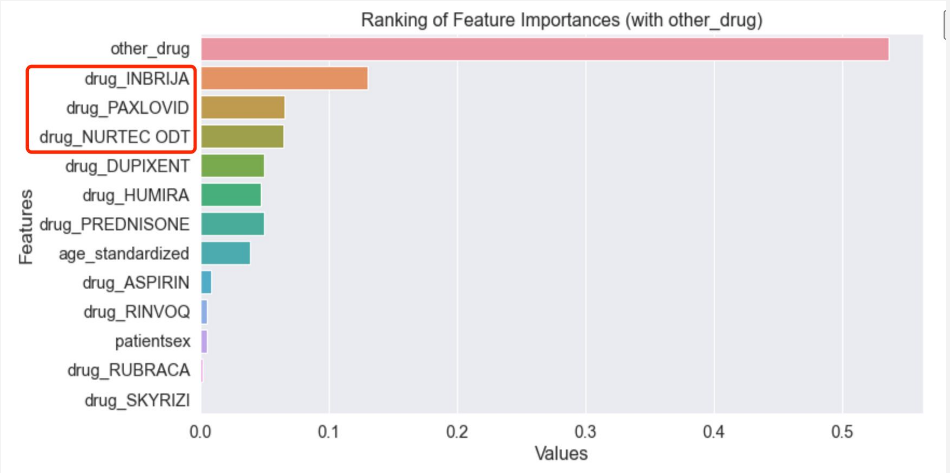


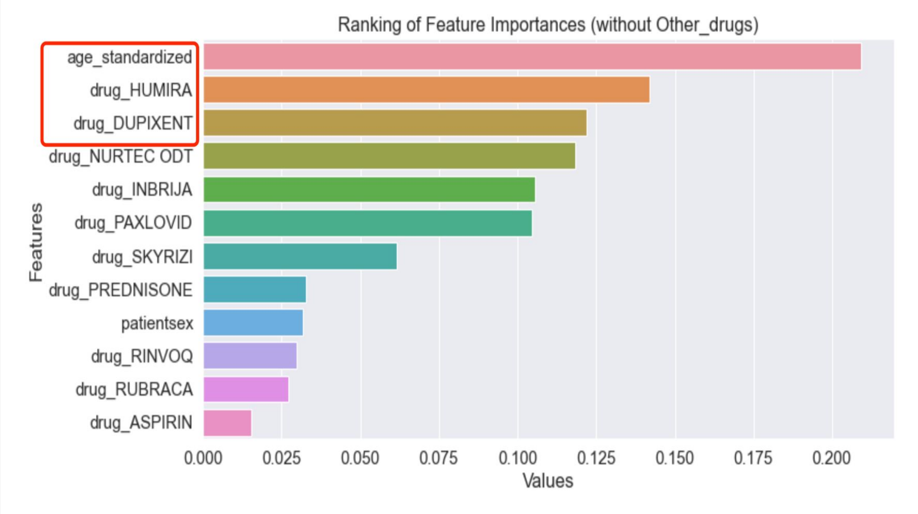


Model Selection

Why not choose Decision Tree & Random Forest & KNN:

- KNN classification: accuracy lower than the other 3 models.
- A small change can significantly affect the overall performance of the model (variables).
- Decision trees and random forests are both prone to overfitting, and are less efficient when more variables added.





Final Conclusion

Logistic Regression

- We mainly use patient age, sex, and types of drugs taken to predict the probability of seriousness level of adverse reactions for each patient.
- All variables are significant in terms of p-values.
- Patients might want to pay extra attention to those drugs that cause significant side effects. (i.e. PREDNISONE, ASPIRIN)

Logit Regression Results

Dep. Variable:	serious_results	**No. Observations:**	10062
Model:	Logit	**Df Residuals:**	10049
Method:	MLE	**Df Model:**	12
Date:	Wed, 22 Mar 2023	**Pseudo R-squ.:**	0.2053
Time:	18:14:55	**Log-Likelihood:**	-5501.6
converged:	True	**LL-Null:**	-6923.3
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
age_label	0.0918	0.022	4.177	0.000	0.049	0.135
patientsex	-0.2950	0.047	-6.242	0.000	-0.388	-0.202
drug_HUMIRA	-1.9641	0.119	-16.516	0.000	-2.197	-1.731
drug_DUPIXENT	-2.4945	0.167	-14.946	0.000	-2.822	-2.167
drug_INBRIJA	-5.1725	0.585	-8.837	0.000	-6.320	-4.025
drug_NURTEC ODT	-4.5306	0.508	-8.919	0.000	-5.526	-3.535
drug_SKYRIZI	-2.2128	0.228	-9.691	0.000	-2.660	-1.765
drug_RINVOQ	-1.3100	0.191	-6.848	0.000	-1.685	-0.935
drug_PREDNISONE	1.6162	0.171	9.454	0.000	1.281	1.951
drug_RUBRACA	-1.5481	0.245	-6.314	0.000	-2.029	-1.068
drug_PAXLOVID	-2.2874	0.158	-14.473	0.000	-2.597	-1.978
drug_ASPIRIN	0.9466	0.167	5.666	0.000	0.619	1.274
other_drug	0.4967	0.094	5.287	0.000	0.313	0.681

Future improvements

- We can take the amount of drug doses and medicinal content into account.
- We can also consider years of drug taken.
- Since age group is clearly playing an important role in causing side effects, we can fit various models separately for different age groups.
- We can use data from a wider date range.
- Try more hyperparameter tuning techniques.

Thank you for listening!