Internship Project Title	ALIYA P SUBEER
Name of the Company	Classification Model - Build a Model that Classifies the Side Effects of a Drug
Name of the Industry Mentor	DEBASHIUS ROY
Name of the Institute	ICT ACADEMY OF KERALA

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
6/09/2023	11/09/2023	22	COLAB	PYTHON SKLEARN
Milestone #	1	Milestone:	EXPLORATORY DATA ANALYSIS	

TABLE OF CONTENT

1. ACKNOWLEDGEMENT

First I would like to thank TCSION for giving me the opportunity to do an internship within the organization.

I also would like all the people that worked along with me to support with their patience and openness they created an enjoyable working environment.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

I am highly indebted TCSion authorities for the facilities provided to accomplish this internship.

I would like to thank my mentor Mr. Debashius Roy for his constructive criticism throughout my internship.

I would like to thank ICT internship for their support and advices to get and complete internship in above said organization.

I am extremely great full to my colleagues and friends who helped me in successful completion of this internship.

2. OBJECTIVE

To create a Classification Model.
Build a Model that Classifies the Side Effects of a Drug.

3. INTRODUCTION

Machine learning (ML) is a state-of-the-art approach that has extensive applications in categorization, prediction, and forecasting. Machine learning techniques are being used in a variety of fields such as medicine, engineering, education, manufacturing and production, weather forecasting, traffic management, robotics, and more. It is one of the most advanced concepts of artificial intelligence (AI), and provides a strategic approach to developing automated, intricate and unbiased algorithmic techniques for multimodal and dimensional biomedical or mathematical data analysis. Machine learning has already shown potential in pharmaceuticals and medicine for finding ways to effectively collect and use various types of data for better analysis, prevention, and treatment of individuals.

Healthcare is an important industry that offers value-based care to millions of people. Healthcare specialists and stakeholders around the world are looking for innovative ways to deliver on quality, value and outcome. Machine learning (ML) based applications embedded with real-time patient data available from different healthcare systems in multiple countries can increase the efficacy of new treatment options which were previously unavailable. It has found wide applications in precision medicine and personalized treatments. Using ML techniques, side effects of drugs both beneficial and adverse can be classified into categories. This can help make more intelligent decisions for precision medicine, personalized treatments, and drug repurposing. Drug classifiers based on side effects can also be an informational resource designed to assist licensed healthcare practitioners in caring for their patients and/or to serve consumers viewing this service as a supplement to, and not a substitute for, the expertise, skill, knowledge and judgment of healthcare practitioners.

A side effect is an unwanted secondary effect that occurs in addition to the desired therapeutic effect of a drug or medication. Side effects can vary for each individual depending on their disease state, age, weight, gender, ethnicity, and general health. They can occur when commencing, decreasing/increasing dosages, or ending a drug or medication regimen and may lead to non-compliance with prescribed treatment. Severe side effects may require adjusting the dosage or prescribing a second medication. Lifestyle or dietary changes may also help minimize side effects. Classifying the side effects for each drug is a challenging task. Machine learning techniques can make such tasks easier without compromising accuracy. Pharmacogenetic research has uncovered significant differences among racial and ethnic groups in the metabolism, clinical effectiveness, and side-effect profiles of many clinically important drugs. These differences must be taken into account in the design of cost management policies such as formulary implementation, therapeutic substitution, and step-care protocols. These programs should be broad and flexible enough to enable rational choices and individualized treatment for all patients, regardless of race or ethnic origin.

4. DESCRIPTION OF INTERNSHIP

Machine learning models have been developed to classify side effects of drugs based on age and gender using a dataset with user-generated text acquired by scraping the WebMD site. The dataset includes both demographic and clinical data and provides user reviews on specific drugs along with related conditions, side effects, age, sex, and ratings reflecting overall patient satisfaction. WebMD is an organization that provides information, support, and reference material about health subjects through a team of doctors and health experts across a broad range of specialty areas. The dataset contains 362806 instances and 12 features including categorical, numerical, and text data. Dataset provides user reviews on specific drugs along with related conditions, side effects, age, sex, and effectiveness reflecting overall patient satisfaction. The structure of the data is that a patient with a unique ID purchases a drug that meets his condition and writes a review and rating for the drug he/she purchased on the date. Afterwards, if the others read that review and find it helpful, they will click UsefulCount.

5. INTERNSHIP ACTIVITIES

Activity 1: Familiarize the topic and search for a dataset.
Familiarize the TCSION platform and understands what I have to do and what kind of dataset I want. For that I search on different sites like Kaggle. And finalize a WebMD dataset of 368206 entries.
Activity 2: Finalize a dataset from Kaggle and do basic analysis.
Do the basic analysis like info, describe, columns, shape, null values, datatypes, unique count etc. Then I familiarize what each column in the dataset about.
Activity 3: Data preprocessing
Missing value handling, Data Wrangling, Cleaning irrelevant data, data with inconsistent datatypes etc.
Activity4: Exploratory data analysis non graphical
Count of gender, Drugs, DrugId, most mentioned drug and DrugId, most mentioned condition etc.
Activity 5: Exploratory data analysis graphical
Plotting of graph based on each parameter. Count plot of top 20 drugs used, ease of use satisfaction rate etc.

6. APPROACH / METHODOLOGY

The dataset includes both demographic and clinical data. It contains 362806 instances and 12 features including categorical, numerical and text data. Dataset provides user reviews on specific drugs along with related conditions, side effects, age, sex, and effectiveness reflecting overall patient satisfaction. The structure of the data is that a patient with a unique ID purchases a drug that meets his condition and writes a review and rating for the drug he/she purchased on the date. Afterwards, if the others read that review and find it helpful, they will click UsefulCount.

7. ASSUMPTIONS

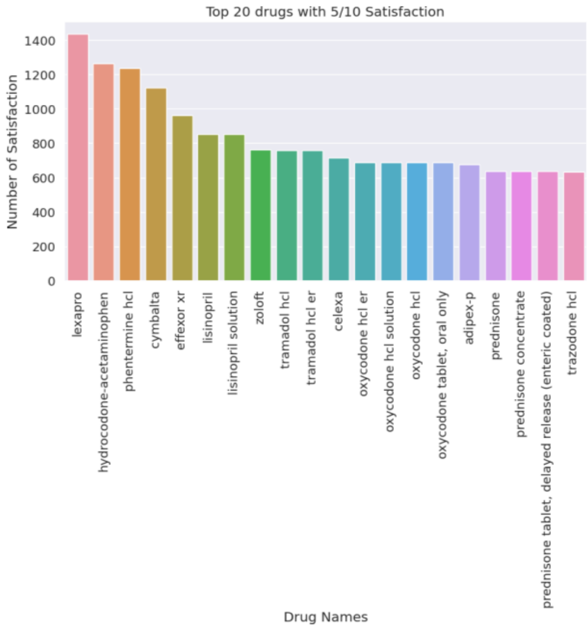
The data collected from WebMD is representative of the population as a whole.
The data collected is accurate and free from errors.
The machine learning techniques used to classify side effects are appropriate for the dataset.
The classification model developed is generalizable to other datasets.
The age and gender information provided in the dataset is accurate and complete.

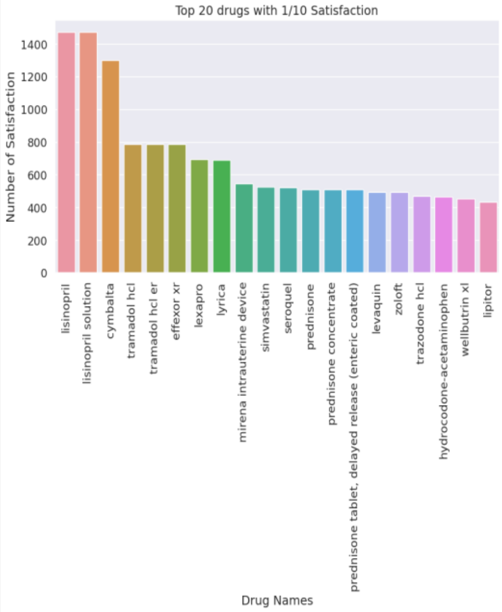
8. EXCEPTIONS / EXCLUSIONS

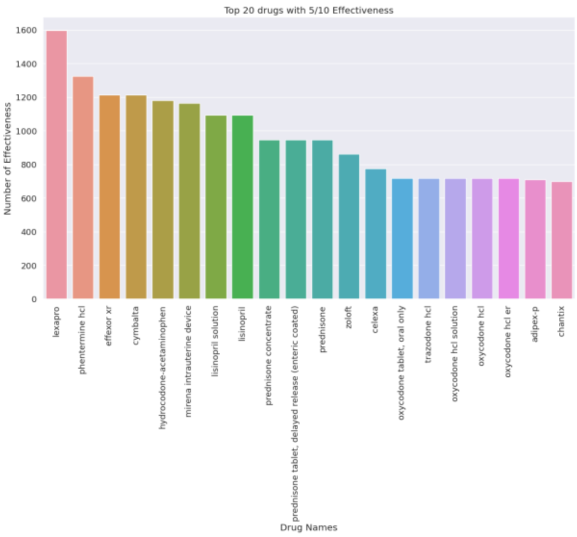
The dataset may not include all possible side effects for each drug.
The dataset may not include information on all drugs or medications.
The dataset may not be up-to-date or accurate.
The classification model developed may not be able to accurately predict side effects for all individuals.
The classification model developed may not be able to account for all factors that can impact side effects.

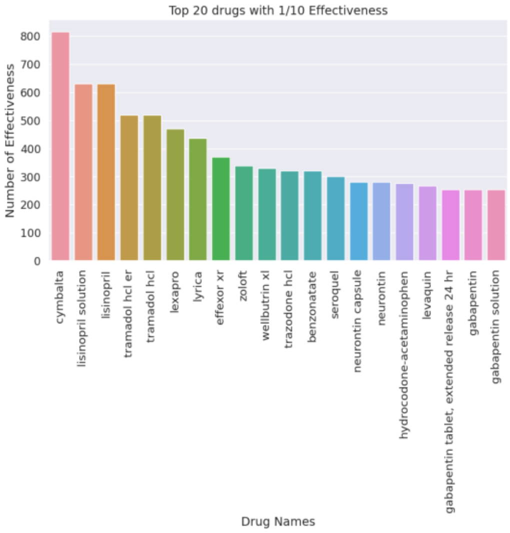
9. CHARTS, TABLES, DIAGRAMS

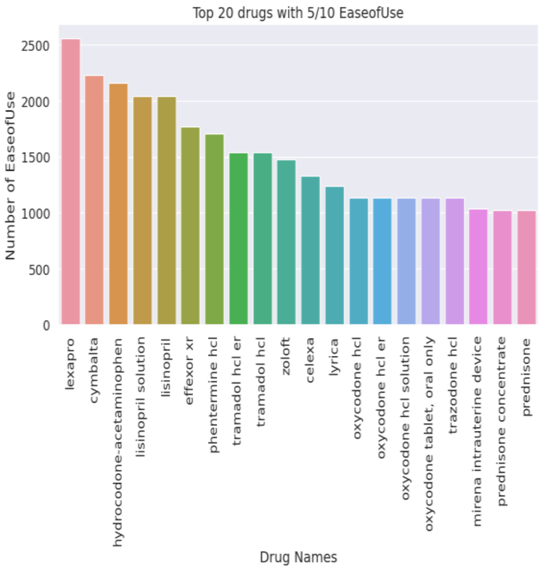
Exploratory data analysis

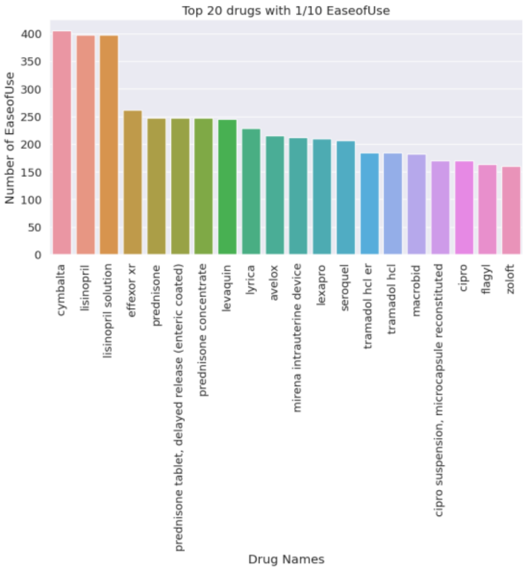


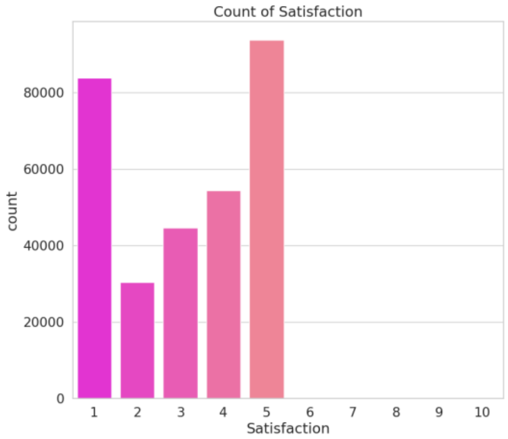


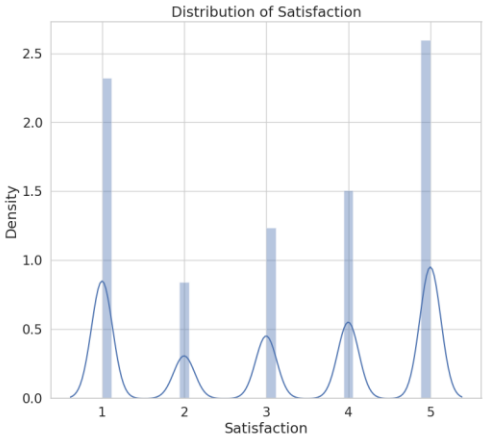


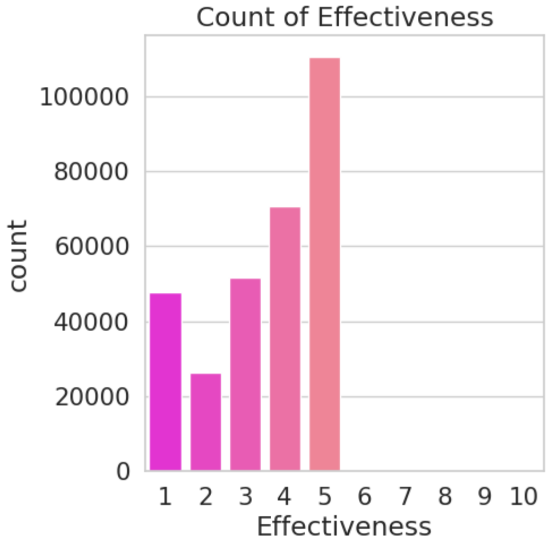


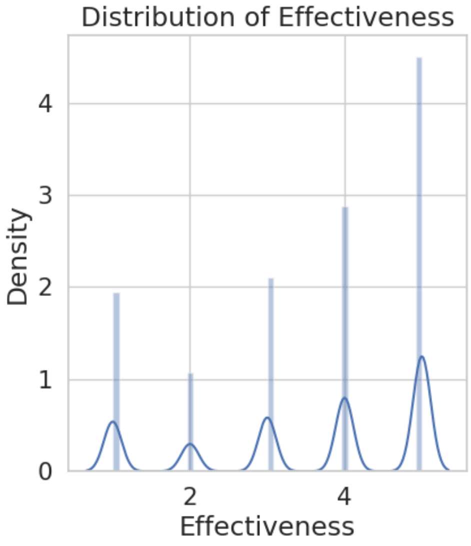

Top 20 drugs with 5/10 EaseofUse
Number of EaseofUse
0
500
1000
1500
2000
2500
lexapro
cymbalta
hydrocodone-acetaminophen
lisinopril solution
lisinopril
effexor xr
phentermine hcl
tramadol hcl er
tramadol hcl
zoloft
celexa
lyrica
oxycodone hcl
oxycodone hcl er
oxycodone hcl solution
oxycodone tablet, oral only
trazodone hcl
mirena intrauterine device
prednisone concentrate
prednisone
Drug Names

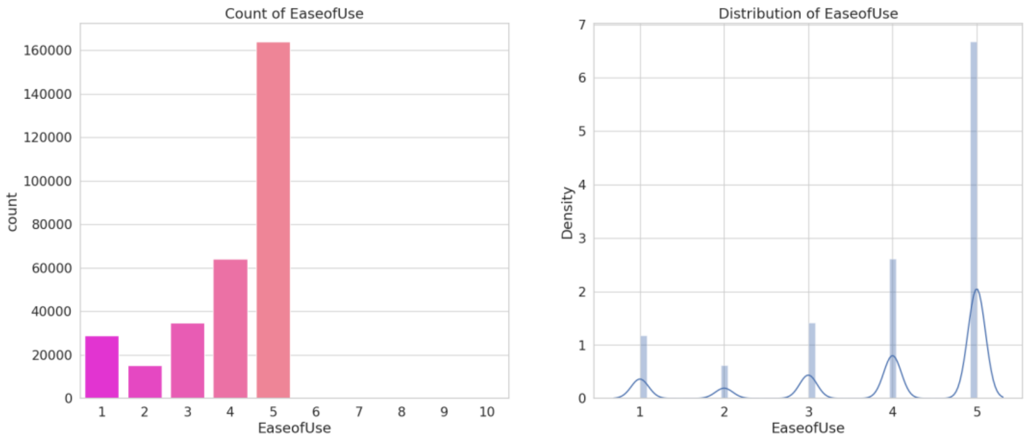

Top 20 drugs with 1/10 EaseofUse
Number of EaseofUse
0
50
100
150
200
250
300
350
400
cymbalta
lisinopril
lisinopril solution
effexor xr
prednisone
prednisone tablet, delayed release (enteric coated)
prednisone concentrate
levaquin
lyrica
avelox
mirena intrauterine device
lexapro
seroquel
tramadol hcl er
tramadol hcl
macrobid
cipro suspension, microcapsule reconstituted
cipro
flagyl
zoloft
Drug Names

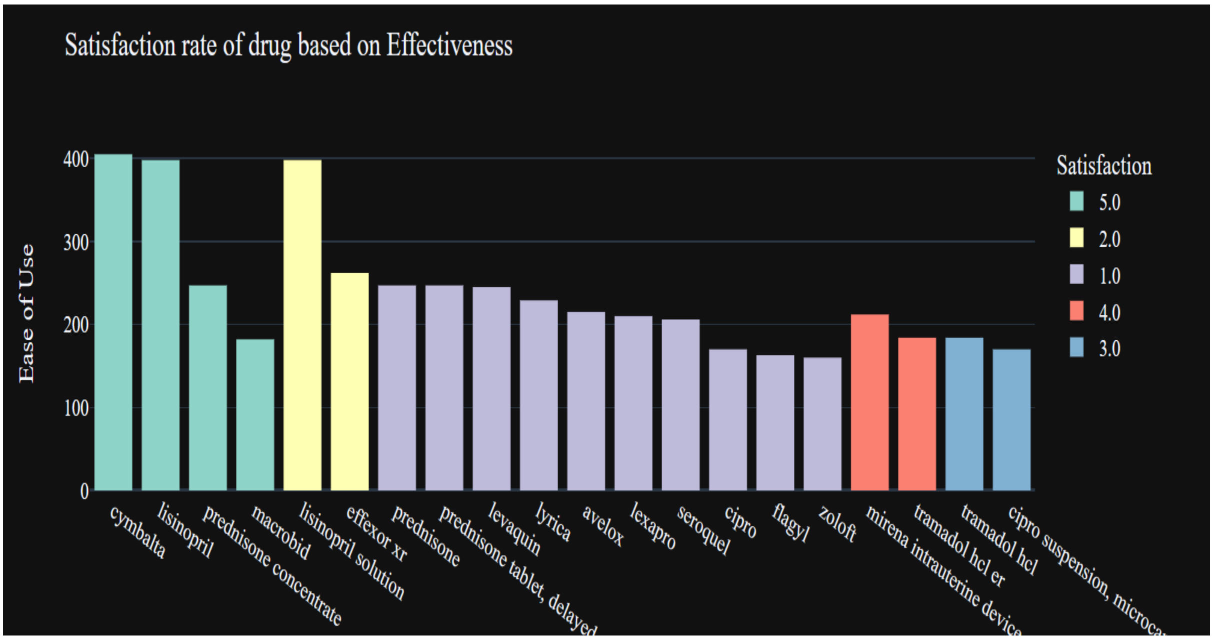

Count of Satisfaction
count
0
20000
40000
60000
80000
1
2
3
4
5
6
7
8
9
10
Satisfaction

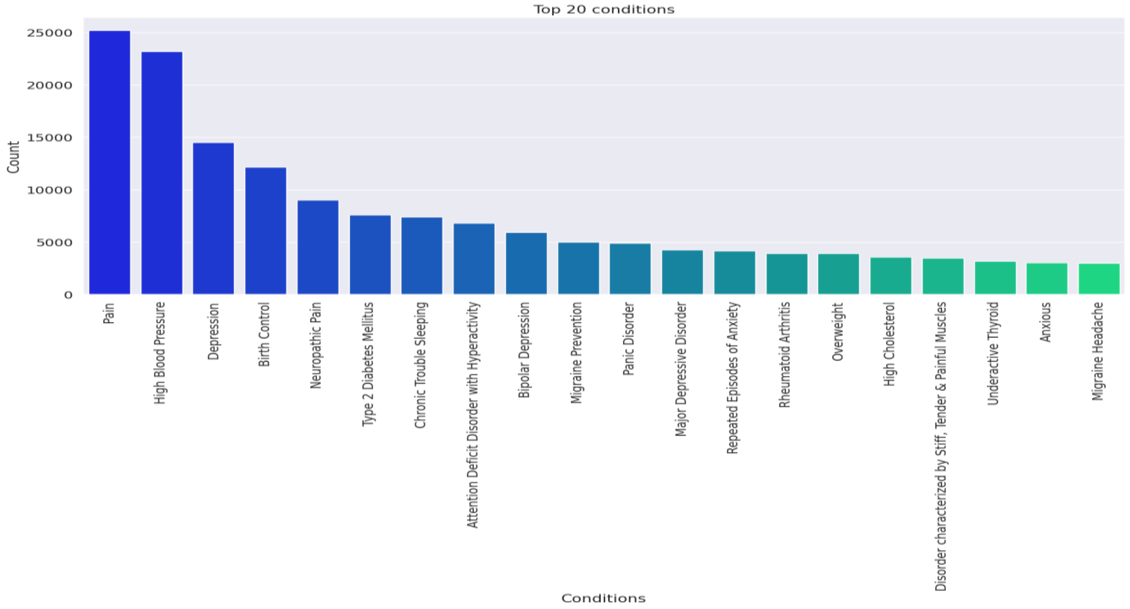

Distribution of Satisfaction
Density
0.0
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
Satisfaction


Count of Effectiveness
count
0
20000
40000
60000
80000
100000
1
2
3
4
5
6
7
8
9
10
Effectiveness

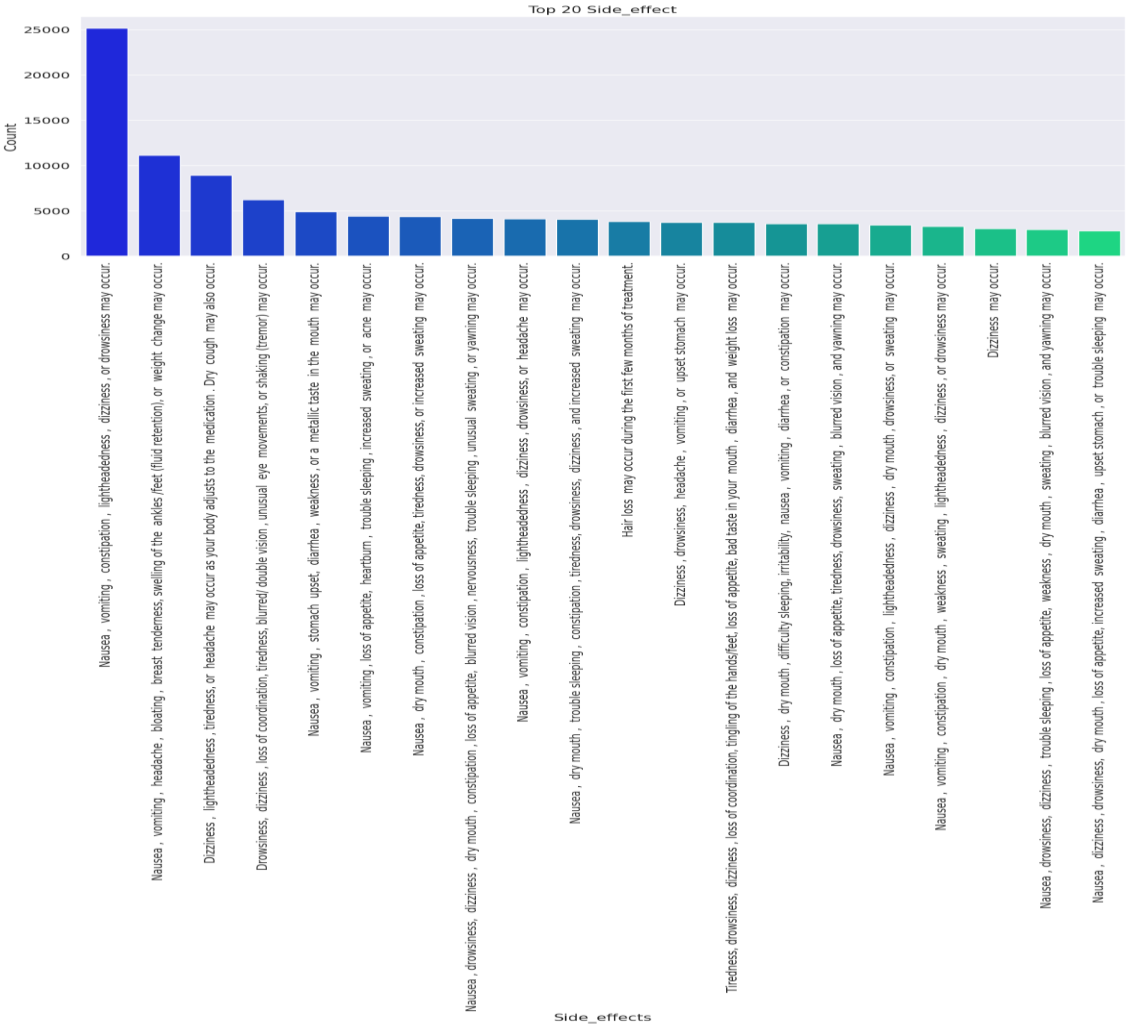

Distribution of Effectiveness
Density
0
1
2
3
4
2
4
Effectiveness


Count of EaseofUse
count
160000
140000
120000
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
7
8
9
10
EaseofUse
Distribution of EaseofUse
Density
7
6
5
4
3
2
1
0
1
2
3
4
5
EaseofUse


Satisfaction rate of drug based on Effectiveness
Ease of Use
400
300
200
100
0
cymbalta
lisinopril
prednisone concentrate
macrobid
lisinopril solution
effexor xr
prednisone
prednisone tablet, delayed
levaquin
lyrica
avelox
lexapro
seroquel
cipro
flagyl
zoloft
mirena intrauterine device
tramadol hcl er
tramadol hcl
cipro suspension, microca
Satisfaction
5.0
2.0
1.0
4.0
3.0


Top 20 conditions
Count
25000
20000
15000
10000
5000
0
Pain
High Blood Pressure
Depression
Birth Control
Neuropathic Pain
Type 2 Diabetes Mellitus
Chronic Trouble Sleeping
Attention Deficit Disorder with Hyperactivity
Bipolar Depression
Migraine Prevention
Panic Disorder
Major Depressive Disorder
Repeated Episodes of Anxiety
Rheumatoid Arthritis
Overweight
High Cholesterol
Disorder characterized by Stiff, Tender & Painful Muscles
Underactive Thyroid
Anxious
Migraine Headache
Conditions

\



For the Lisinopril drug as example, I concluded as this drug have side effects but its effectiveness and satisfaction rate is good. Females are more affected by the side effects and females are the main users. age group of 55-64 having adverse effects. 45-64 is almost the same side effects. reviews count decreased yearly. That means users body are adjusted to side effects or however users have the mentality to cure the sickness and not too much worry about side effects.

10.CHALLENGES & OPPORTUNITIES

- The dataset may not be representative of the entire population.
- The data collected may not be accurate or complete.
- The classification model developed may not be generalizable to other datasets.
- The age and gender information provided in the dataset may not be accurate or complete.
- Developing a classification model that can accurately predict side effects for all individuals is a challenging task.
- The dataset may not include all possible side effects for each drug.
- The dataset may not include information on all drugs or medications.
- The dataset may not be up-to-date or accurate.
- Developing a classification model that can accurately predict side effects for all individuals is a challenging task.
- There is a need to account for all factors that can impact side effects.
- However, developing such a model can help healthcare professionals make more informed decisions about prescribing medications and help patients make more informed decisions about their health.
- Another opportunity is that the model can be used to identify previously unknown side effects of drugs, which can lead to new discoveries and better treatment options. Additionally, the model can be used to identify patterns in side effects across different drugs and patient populations, which can help researchers better understand the underlying mechanisms of drug side effects. Finally, the model can be used to develop personalized treatment plans based on an individual's age, gender, and other factors that may impact their risk of experiencing side effects.

11.RISK VS REWARD

12. Risk	Reward
As a reward, Healthcare professionals make more informed decisions about prescribing medications and help patients make more informed decisions about their health. It can also be used to identify previously unknown side effects of drugs, which can lead to new discoveries and better treatment options. Additionally, the model can be used to identify patterns in side effects across different drugs and patient populations, which can help researchers better understand the underlying mechanisms of drug side effects.	As a risk, developing a classification model that can accurately predict side effects for all individuals is a challenging task. There is a need to account for all factors that can impact side effects, and the dataset may not include all possible side effects for each drug. The dataset may also not include information on all drugs or medications, and it may not be up-to-date or accurate. It is important to carefully consider the risks and rewards of any treatment or intervention before making a decision. Healthcare professionals should work closely with their patients to ensure that they understand the potential benefits and harms of different treatments and can make informed decisions about their care.

12.REFLECTIONS ON THE INTERNSHIP

Developing a classification model that can accurately predict side effects of drugs based on age and gender. The project also highlighted the importance of data preprocessing and cleaning to ensure that the data is accurate and complete. Additionally, the project demonstrated the potential of machine learning techniques to identify previously unknown side effects of drugs and patterns in side effects across different drugs and patient populations. Future research could focus on developing more accurate classification models that can account for all factors that can impact side effects and can be generalized to other datasets.

13.RECOMMENDATIONS

- Data Collection: Collect data from multiple sources to increase the diversity of the dataset and improve the generalizability of the model.
- Feature Engineering: Explore different feature engineering techniques to extract meaningful information from the dataset. Consider incorporating additional features such as drug dosage, treatment duration, and patient medical history to improve the model's performance.
- Model Selection: Experiment with different machine learning algorithms and evaluate their performance using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Consider ensemble methods or deep learning models to further enhance the model's predictive capabilities.
- Hyper parameter Tuning: Optimize the hyper parameters of the selected model using techniques such as grid search or random search. This can help improve the model's generalization ability and prevent overfitting.
- Validation and Testing: Use appropriate validation techniques such as cross-validation to estimate the model's performance on unseen data. Perform rigorous testing on an independent test set to assess the model's real-world performance.
- Interpretability: Use interpretable machine learning models such as decision trees or logistic regression to provide insights into the factors influencing side effects.
- Continual Improvement: Regularly update the model with new data to ensure its relevance and accuracy over time. Monitor its performance in real-world scenarios and refine it based on feedback from healthcare professionals and end-users.

14.OUTCOME / CONCLUSION

In conclusion, from the EDA analysis, I concluded as this drug have side effects but its effectiveness and satisfaction rate is good. Females are more affected by the side effects and females are the main users. age group of 55-64 having adverse effects. 45-64 is almost the same side effects. Reviews counts decreased yearly. That means users body are adjusted to side effects or however users have the mentality to cure the sickness and not too much worry about side effects.

Developing a classification model for side effects of drugs using the WebMD dataset is a challenging task that requires careful consideration of the data, machine learning techniques, and evaluation metrics. While there are limitations and challenges associated with this approach, such as the need to account for all factors that can impact side effects and the accuracy of the dataset, there are also significant opportunities for improving patient care and advancing medical research. By developing more accurate classification models, researchers can help healthcare professionals make more informed decisions about prescribing medications and help patients make more informed decisions about their health. Additionally, these models can be used to identify previously unknown side effects of drugs and patterns in side effects across different drugs and patient populations, which can lead to new discoveries and better treatment options. Overall, the development of classification models for side effects of drugs using the WebMD dataset has the potential to significantly improve patient outcomes and advance medical research.

15.ENHANCEMENT SCOPE

Regularly update the model with new data to ensure its relevance and accuracy over time. Monitor its performance in real-world scenarios and refine it based on feedback from healthcare professionals and end-users.

16.LINK TO CODE AND EXECUTABLE FILE

CLASSIFICATION_MODEL_SIDE_EFFECTS_OF_DRUGS/internship_EDA_report.ipynb at main · aliyapsubeer/CLASSIFICATION_MODEL_SIDE_EFFECTS_OF_DRUGS (github.com)