A Project Report Entitled

NLP Classification of Patient Conditions through Drug Reviews

Project Report submitted to

RAYAT SHIKSHAN SANSTHA's,

SADGURU GADAGE MAHARAJ COLLEGE, KARAD

(An Autonomous College)

DEPARTMENT OF STATISTICS

FOR THE PARTIAL FULFILLMENT OF THE DEGREE
MASTER OF SCIENCE
IN
STATISTICS

SUBMITTED

BY,

Mr. Rudray Sanjay Magdum

M.Sc. II (Statistics)

Under the guidance of

Miss. K. H. Powar

2023-2024

CERTIFICATE

This is to certify that the project report entitled '**NLP Classification of Patient Conditions through Drug Reviews'** being Submitted by **Mr. Rudray Sanjay Magdum** as partial fulfillment for the award of degree of M.Sc. (Statistics) is a record of work carried out by him under my supervision and guidance. To the best of my knowledge the matter presented in the survey has not been submitted earlier.

Place: Karad
Date:

Miss. K. H. Powar
Project guide

Dr. Mrs. Patil S. P.
PG Co-ordinator
Department of Statistics,
Sadguru Gadage Maharaj College, Karad.

Index

Chapter No.	Name of Chapter	Page No.
1.	Acknowledgement	4
2.	Introduction	5
3.	Background	6
4.	Significance Of Dataset	7
5.	Data Description	8
6.	Scope Of The Project	9
7.	Methodology	10
8.	Objective	11
9.	Data Preprocessing	12
10.	Exploratory Data Analysis (EDA) 1. Distribution of Rating 2. Wordcloud 3. Boxplot	13
11.	Classification/Model building 1. MultinomialNB 2. Navie Baye's 3. Random Forest 4. SVM	17
12.	Overall summary of classification model	22
13.	Common Observation	23
14.	Refferences	24
15.	Appendix	25

ACKNOWLEDGEMENT

I have put considerable effort into this survey, but it would not have been possible without the generous support and assistance of many individuals. I would like to extend my sincere thanks to all of them. It is my great pleasure to express my gratitude to ***Miss. K. H. Powar*** for her invaluable guidance and constant encouragement throughout the project, as well as for providing necessary facilities and information regarding the survey. Additionally, I would like to thank all the non-teaching staff of our department for their help and cooperation.

I am grateful to all the teachers for dedicating their valuable time to this endeavor, and I extend my thanks to all my friends and research students for their cooperation and assistance.

Introduction

In contemporary healthcare discussions, there is a growing acknowledgment of the importance of patient experiences and their narratives in understanding medical conditions and treatment efficacy. The conventional barriers to open dialogue surrounding healthcare are gradually dissipating, fostering an environment where individuals feel more comfortable sharing their insights, challenges, and responses to various medical interventions. One notable source of this valuable information is embedded within the vast repository of patient reviews on pharmaceutical interventions.

Our project, titled "NLP Classification of Patient Conditions through Drug Reviews," delves into the expansive realm of natural language expressions present in patient feedback. The "Condition" column encapsulates the specific medical contexts for which patients are providing feedback, offering a glimpse into the multifaceted ways individuals express their experiences with diverse health conditions. On the other hand, the "Review" column encapsulates the nuanced narratives, opinions, and outcomes shared by patients in response to specific drug treatments.

Our primary objective is to leverage Natural Language Processing (NLP) techniques to gain profound insights from this repository of patient reviews. Employing advanced language analysis and machine learning methodologies, our goal is to develop a classification system that can identify patterns, sentiments, and potential indicators within these drug reviews. In simpler terms, we aim to discern common themes, emotional nuances, and the effectiveness of various drug interventions based on patient testimonials. This endeavor holds the promise of providing a deeper understanding of how individuals articulate their experiences with different medications and, consequently, aids in the classification of patient conditions.

Ultimately, our aspiration is for this project to contribute to a more informed understanding of patient perspectives on drug interventions, facilitating healthcare professionals and researchers in comprehending the efficacy and challenges associated with specific treatments. Through this, we anticipate a valuable enhancement in the discourse around healthcare, fostering an environment where patient experiences play a pivotal role in shaping our understanding of medical conditions and the effectiveness of pharmaceutical interventions.

Background

This initiative centers around employing Natural Language Processing (NLP) to unravel insights within patient reviews, particularly focusing on drug-related experiences. In contemporary healthcare, there is a notable trend of individuals openly sharing their perspectives and feedback on pharmaceutical interventions. This project harnesses a comprehensive dataset featuring essential elements such as "Drug Name," "Condition," "Review," "Rating," "Date," and "Useful Count." The "Condition" column encapsulates the diverse medical contexts for which patients provide feedback, offering a nuanced portrayal of their experiences with various health conditions. Simultaneously, the "Review" column captures the intricate narratives, opinions, and outcomes shared by patients in response to specific drug treatments.

The project operates on the premise that this classification system can contribute significantly to understanding patient perspectives on drug interventions. By scrutinizing the language used in reviews and discerning commonalities, emotional nuances, and the perceived effectiveness of drug treatments, the project aims to provide valuable insights for healthcare professionals and researchers. This endeavor is expected to enhance our comprehension of the efficacy and challenges associated with specific pharmaceutical interventions, fostering an environment where patient experiences play a pivotal role in shaping our understanding of medical conditions and the effectiveness of drug treatme

Significance of the Dataset

Comprehending patient perspectives on drug interventions is a nuanced and multifaceted challenge within contemporary healthcare. Conventional research and analytical methods may fall short in fully encapsulating the intricacies of individual experiences and the diverse support mechanisms evident in patient reviews. The dataset employed in this project presents a distinctive opportunity to unravel the intricate weave of natural language expressions embedded within authentic patient narratives related to pharmaceutical interventions.

The significance of this dataset lies in its capacity to transcend the limitations of standardized research methods, offering a more holistic understanding of patient experiences. By delving into the authentic language used by patients and discerning the nuances within their testimonials, the dataset becomes a rich source of information for constructing a robust classification system. This system, in turn, holds the potential to uncover patterns, sentiments, and valuable indicators that traditional research approaches might overlook.

In essence, the dataset serves as a valuable reservoir of real-world patient experiences, providing a nuanced and authentic lens through which healthcare professionals and researchers can better comprehend the effectiveness, challenges, and emotional dimensions associated with specific drug interventions. This authenticity and depth contribute significantly to advancing our understanding of patient perspectives, enriching the discourse on pharmaceutical interventions and their impact on various medical conditions.

Data Description

The dataset contains 161,297 rows and 7 columns.
The data spans from 2018 to 2021, capturing patient drug reviews over this time period.

"Unnamed" : The "Unnamed" column appears to be an index or identifier and should be checked for its relevance. If it doesn't provide meaningful information, it might be advisable to drop this column.

"drugName" : The "drugName" column specifies the names of drugs under review.

"condition" : The "condition" column details the medical conditions or reasons for which the drugs are prescribed.

"review" : Patient reviews are provided in the "review" column, offering qualitative insights into their experiences.

"rating" : The "rating" column quantifies patient satisfaction with the drugs, presumably on a numerical scale.

"date" : The "date" column records the date of each review, allowing for temporal analysis.

"usefulCount" : "usefulCount" indicates the number of users who found a particular review helpful, suggesting its relevance or impact within the community.

```
data
```

Out[2]:

	Unnamed: 0	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	April 27, 2010	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	December 14, 2009	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	November 3, 2015	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	November 27, 2016	37
...	...	...	...	...	...	...	...
161292	191035	Campral	Alcohol Dependence	"I wrote my first report in Mid-October of 201...	10.0	May 31, 2015	125
161293	127085	Metoclopramide	Nausea/Vomiting	"I was given this in IV before surgey. I immed...	1.0	November 1, 2011	34
161294	187382	Orencia	Rheumatoid Arthritis	"Limited improvement after 4 months, developed...	2.0	March 15, 2014	35
161295	47128	Thyroid desiccated	Underactive Thyroid	"I've been on thyroid medication 49 years...	10.0	September 19, 2015	79
161296	215220	Lubiprostone	Constipation, Chronic	"I've had chronic constipation all my adu...	9.0	December 13, 2014	116

161297 rows × 7 columns

Scope Of The Project

This project's focus is on employing Natural Language Processing (NLP) methodologies to scrutinize conversations surrounding patient experiences with drug interventions. The dataset under consideration is characterized by two pivotal columns: "Condition" and "Review." The project aims to delve into these patient narratives to construct a classification system, offering a more nuanced understanding of the diverse medical contexts and authentic responses related to specific pharmaceutical treatments.

Unlike conventional research methods, this approach acknowledges the complexity of individual experiences, aiming to capture the intricacies present in patients' unfiltered expressions. By leveraging NLP techniques, the project endeavors to distill valuable patterns and sentiments embedded in patient reviews, enriching our comprehension of the efficacy, challenges, and emotional dimensions associated with various drug interventions. In essence, the scope of this initiative extends beyond traditional analyses, providing a comprehensive exploration of patient perspectives on pharmaceutical interventions through the lens of natural language expressions.

Methodology

❖ Data Collection: Assemble a dataset featuring essential columns, including "Condition," "Review," and "Rating," capturing diverse patient experiences with drug interventions.

❖ Text Preprocessing: Cleanse and preprocess the text data, eliminating irrelevant details, addressing typos, and standardizing formats to enhance data quality.

❖ Tokenization: Break down patient reviews into tokens, facilitating a granular analysis to uncover patterns and sentiments related to drug interventions.

❖ Data Visualization: Present the results through visualizations, facilitating a clearer interpretation of the classification outcomes and contributing to comprehensive insights.

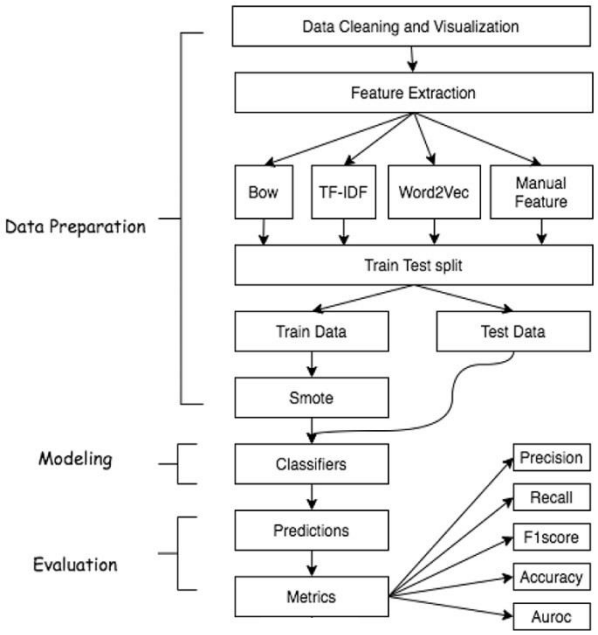
❖ Classification Model Development: Implement machine learning algorithms to construct a robust classification system. Train the model to categorize patient conditions based on their articulated experiences.

❖ Performance Evaluation: Assess the accuracy and effectiveness of the classification model through metrics such as precision, recall, and F1 score, ensuring its reliability in identifying diverse medical contexts.

❖ Feature Importance Analysis: Explore the significance of different features within patient reviews to understand which aspects contribute most to the classification outcomes.

❖ Iterative Refinement: Continuously refine the classification model based on feedback and emerging insights, optimizing its performance over time.

❖ Iterative Analysis: Iteratively review and enhance the classification methodology based on emerging findings, ensuring adaptability and continuous improvement throughout the project lifecycle.



Objectives

- To Develop a robust classification model to categorize patient conditions based on drug reviews, providing a detailed understanding of diverse medical contexts.

- To Investigate and analyze significant features within patient reviews to comprehend the factors contributing to the classification outcomes.

- To Present the classification outcomes through visualizations, facilitating a clear interpretation of patient condition categorizations for improved comprehension.

- To Assess the accuracy, precision, recall, and F1 score of the classification model to ensure its effectiveness in identifying and categorizing patient conditions.

Data Preprocessing

- Text Cleaning:
  - Remove any irrelevant characters, HTML tags, or special characters that may not contribute to the analysis.
  - Convert text to lowercase to ensure uniformity.

- Tokenization:
  - Tokenize the text into individual words or subwords. You can use libraries such as NLTK for tokenization.

- Remove Stopwords:
  - Remove common stopwords (e.g., 'and', 'the', 'is') that don't contribute much to the meaning of the text.

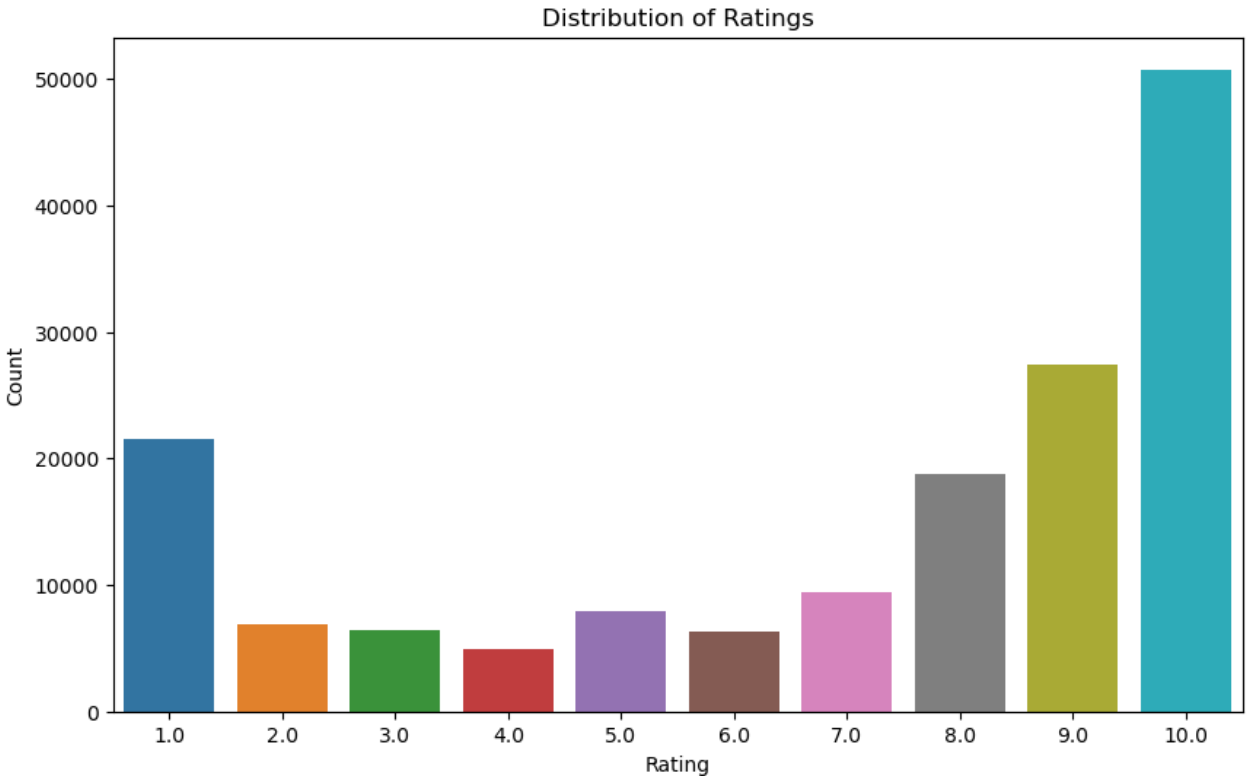
- Lemmatization :
  - Reduce words to their base or root form to consolidate similar words.

- Join Tokens Back to Text (Optional):
  - If necessary, join the processed tokens back into a text format.

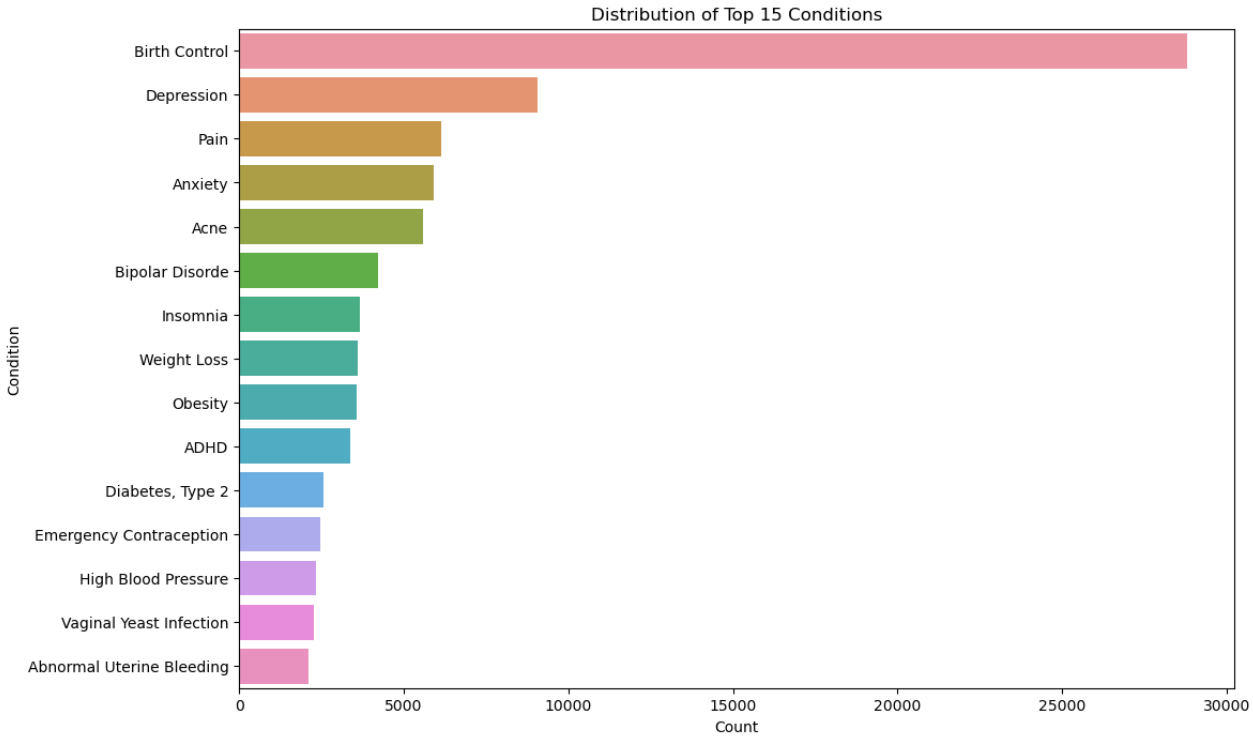
Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics and patterns within your Drug review dataset.

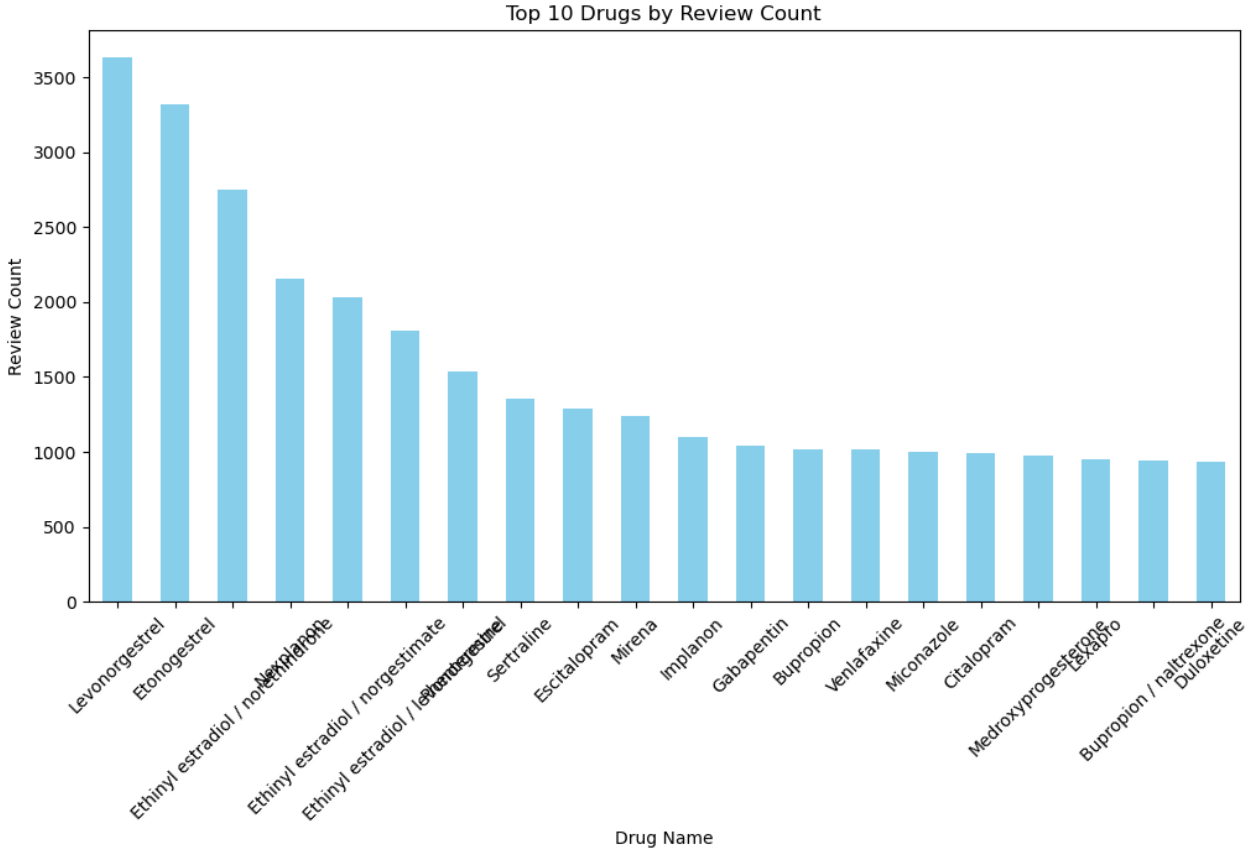
1) Distribution of Ratings:



From the above graph we will see that there is rating of 10.0 is of count 50000.Which are given to the drug have highest satisfaction to the patient.



From the above graph we have seen that Birth Control have highest review count followed by Depression and Pain, Anxiety and Acne.

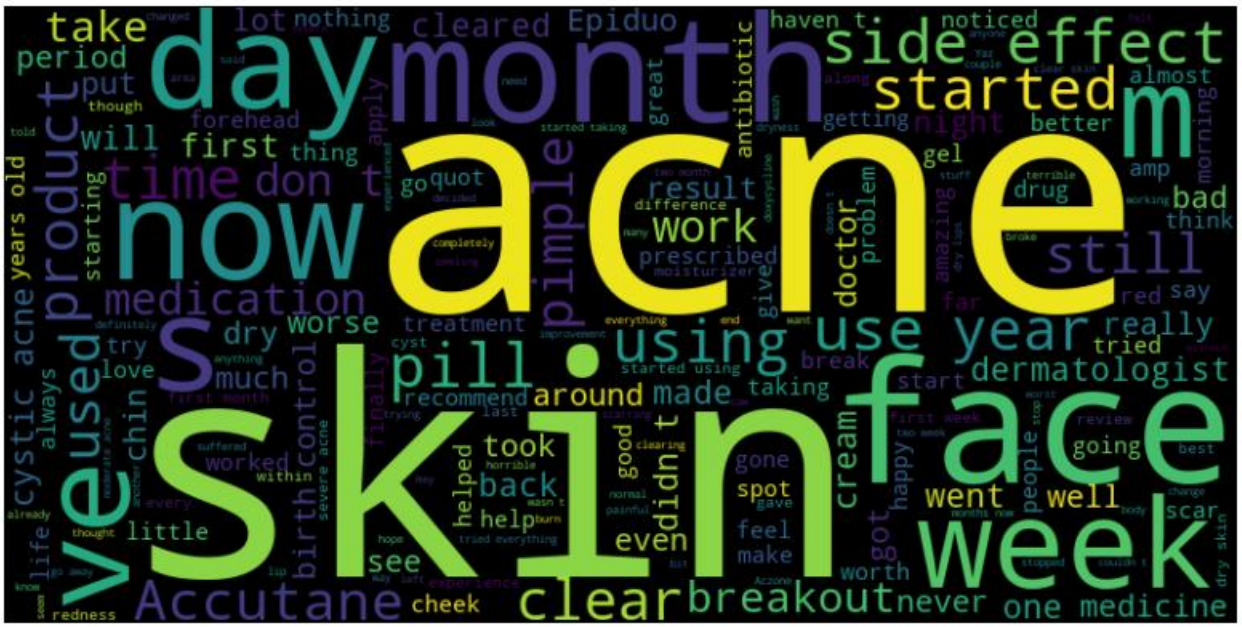


From the above graph we have seen The Drugs have positively skewed distribution.

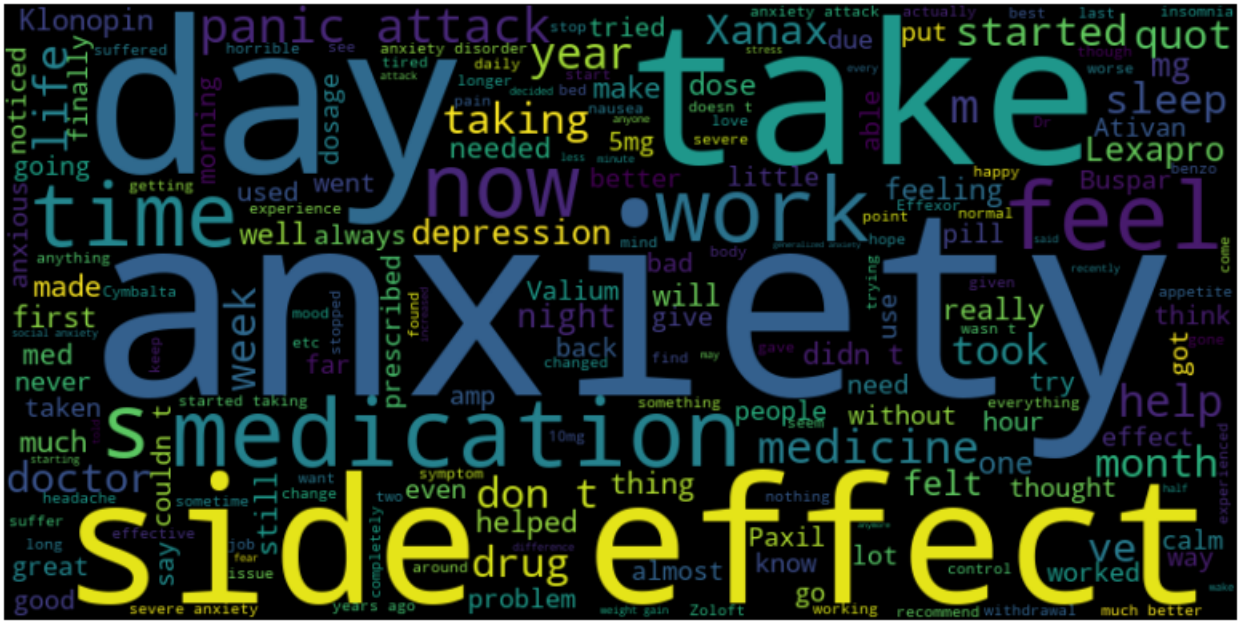
2) Word Cloud :

Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the document.

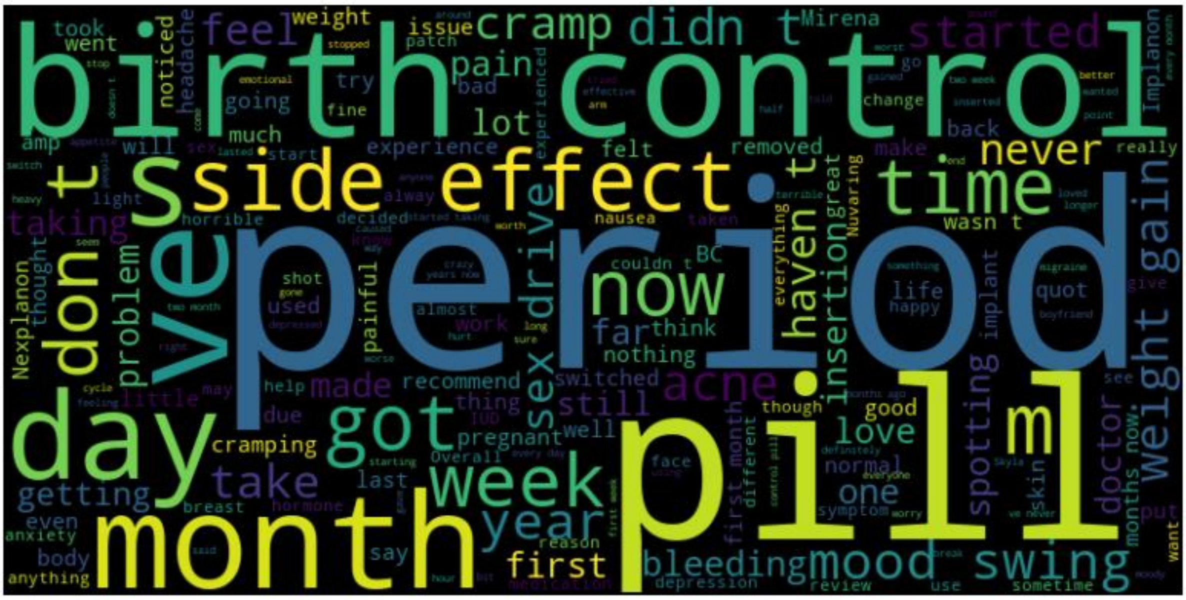
Word Cloud for Acne



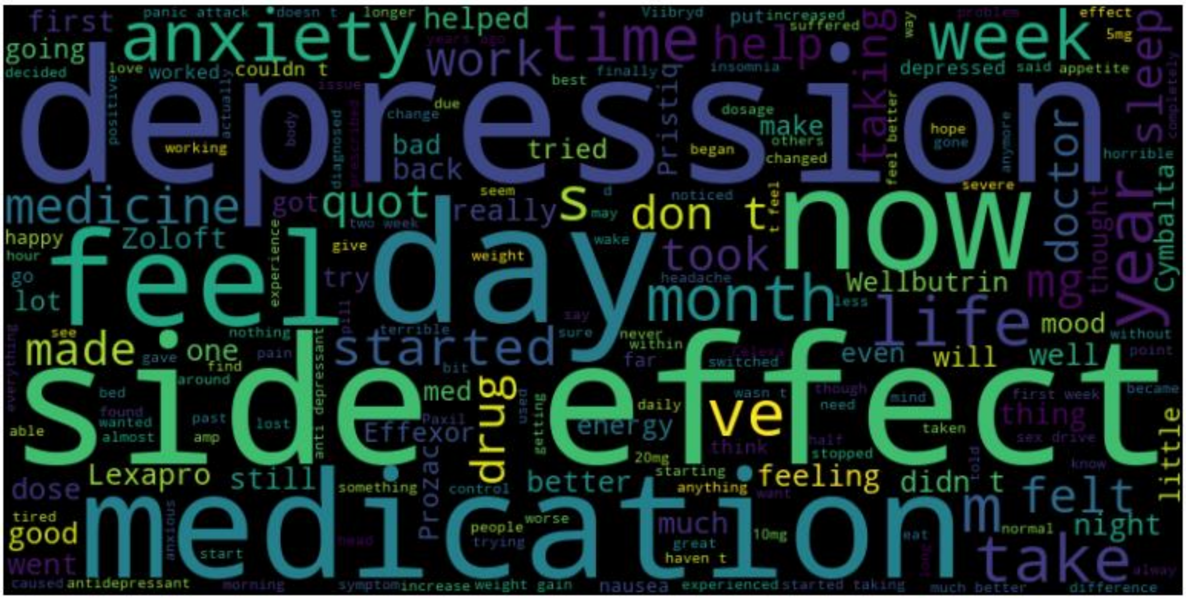
Word Cloud for Anxiety



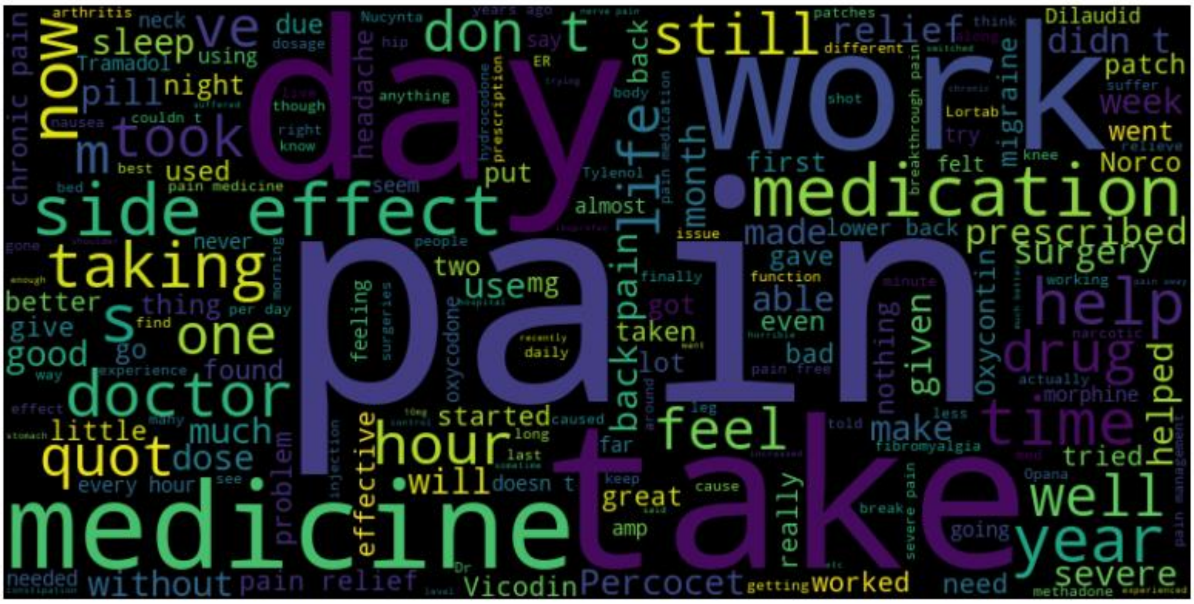
Word Cloud for Birth Control



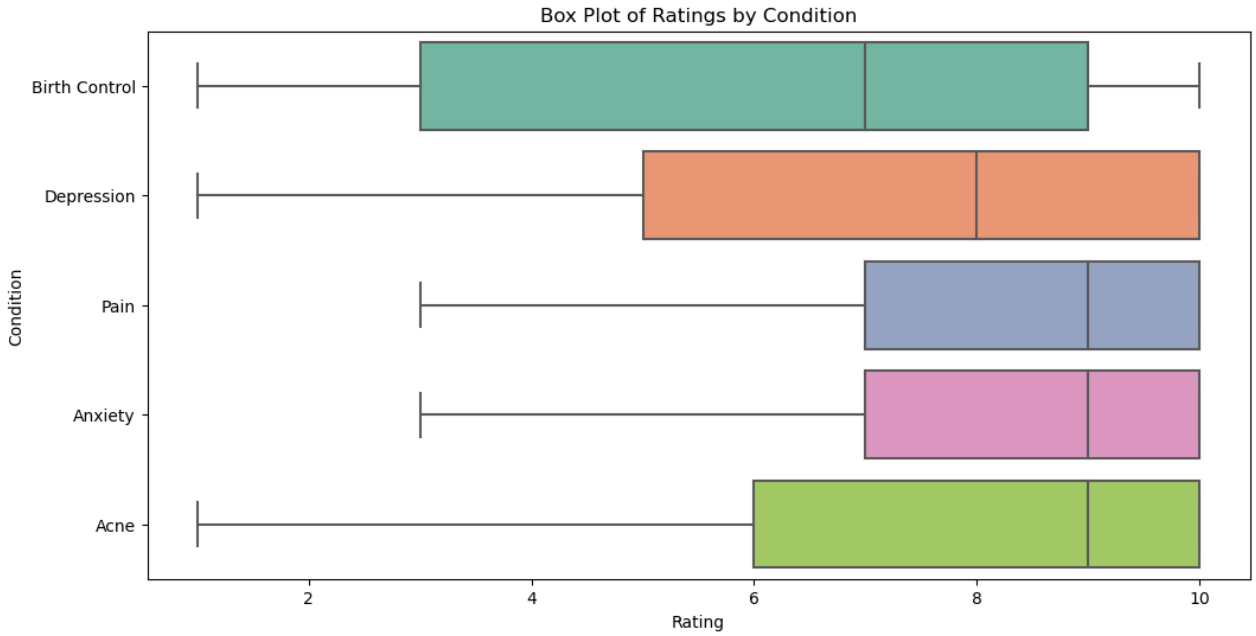
Word Cloud for Depression



Word Cloud for Pain



3) Box Plot:



In descriptive statistics, a box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.
We draw boxplot to check outliers.

Classification

The "Classification" section elucidates the methodologies employed in categorizing patient conditions through drug reviews. Multinomial Naive Bayes, Support Vector Machine, Random Forest, and Naive Bayes algorithms were utilized, each undergoing meticulous configuration, predictor selection, and iterative refinement. Results showcase key performance metrics, facilitating a comparative analysis of algorithmic efficacy. Significant insights, including feature importance and interpretability, augment the understanding of patient conditions. Visualizations aid in result interpretation. The section concludes with a succinct summary of findings, paving the way for potential future enhancements in algorithmic exploration and feature engineering.

1. MultinomialNB:

Multinomial Naive Bayes (MultinomialNB) is a variant of the Naive Bayes algorithm designed specifically for multinomially distributed data. It's commonly used in text classification tasks, where the features (words or terms) are discrete and represent the frequency of terms in a document.

I initiated the project by employing the 'Multinomial Naive Bayes' algorithm for classification. This choice was driven by its suitability for text data and its efficient handling of multiple classes, aligning well with the diverse patient conditions in drug reviews. The initial phase involved training the model with various predictors, considering a comprehensive approach to capture nuanced patterns in patient narratives.

Accuracy on Top 15 Conditions: 0.81

	precision	recall	f1-score	support
ADHD	0.96	0.75	0.84	657
Abnormal Uterine Bleeding	0.71	0.04	0.08	376
Acne	0.95	0.80	0.87	1081
Anxiety	0.79	0.57	0.66	1190
Bipolar Disorde	0.92	0.54	0.68	882
Birth Control	0.82	0.99	0.90	5807
Depression	0.59	0.85	0.70	1804
Diabetes, Type 2	0.95	0.78	0.85	497
Emergency Contraception	1.00	0.82	0.90	532
High Blood Pressure	0.91	0.70	0.79	451
Insomnia	0.87	0.72	0.79	707
Obesity	0.65	0.57	0.61	687
Pain	0.90	0.86	0.88	1289
Vaginal Yeast Infection	0.98	0.91	0.94	459
Weight Loss	0.68	0.63	0.65	713
accuracy			0.81	17132
macro avg	0.85	0.70	0.74	17132
weighted avg	0.82	0.81	0.80	17132

The Multinomial Naive Bayes model achieves an accuracy of 81% in classifying patient conditions related to the top 15 health concerns.

In summary, the Multinomial Naive Bayes model demonstrates a solid performance in classifying patient conditions related to the top 15 health concerns. Notable precision and recall values, along with balanced F1-scores, affirm its effectiveness across a range of medical categories.

2.Naive Bayes:

The 'Naive Bayes' algorithm, known for its simplicity and efficiency, played a crucial role in the project. Its application was aimed at discerning patterns and relationships within the drug reviews, particularly focusing on the independence assumption. The model underwent iterations to optimize the selection of relevant predictors, with the goal of achieving accurate classification while adhering to the naive Bayes framework.

Accuracy (Naive Bayes): 0.81

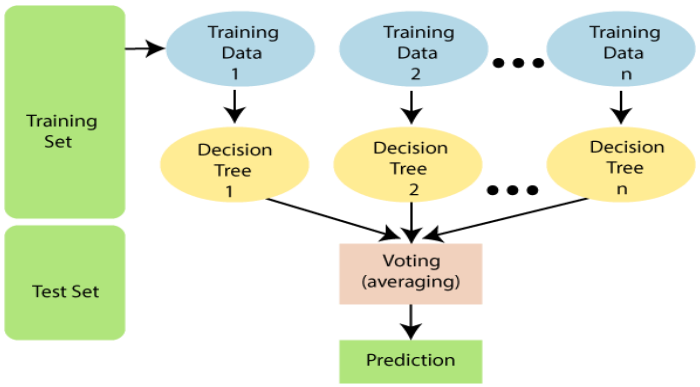
	precision	recall	f1-score	support
ADHD	0.96	0.75	0.84	657
Abnormal Uterine Bleeding	0.71	0.04	0.08	376
Acne	0.95	0.80	0.87	1081
Anxiety	0.79	0.57	0.66	1190
Bipolar Disorde	0.92	0.54	0.68	882
Birth Control	0.82	0.99	0.90	5807
Depression	0.59	0.85	0.70	1804
Diabetes, Type 2	0.95	0.78	0.85	497
Emergency Contraception	1.00	0.82	0.90	532
High Blood Pressure	0.91	0.70	0.79	451
Insomnia	0.87	0.72	0.79	707
Obesity	0.65	0.57	0.61	687
Pain	0.90	0.86	0.88	1289
Vaginal Yeast Infection	0.98	0.91	0.94	459
Weight Loss	0.68	0.63	0.65	713
accuracy			0.81	17132
macro avg	0.85	0.70	0.74	17132
weighted avg	0.82	0.81	0.80	17132

The Naive Bayes model achieves an accuracy of 81% in classifying patient conditions based on drug reviews.

In summary, the Naive Bayes model demonstrates a solid performance in classifying patient conditions. Notable precision and recall values, along with balanced F1-scores, affirm its effectiveness across a spectrum of medical categories.

3.Random Forest:

The project further incorporated the 'Random Forest' algorithm, leveraging its ensemble learning capabilities to enhance classification accuracy. I initially utilized a set of predictors and iteratively refined the model to achieve a balance between feature richness and efficiency. The ensemble nature of Random Forest proved beneficial in capturing complex relationships within patient reviews, contributing to a more robust classification process.



Accuracy (Random Forest): 0.89

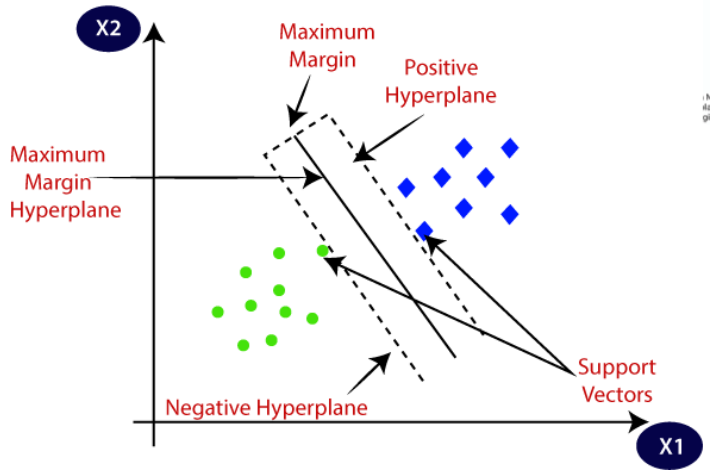
	precision	recall	f1-score	support
ADHD	0.93	0.89	0.91	657
Abnormal Uterine Bleeding	0.99	0.57	0.73	376
Acne	0.96	0.91	0.93	1081
Anxiety	0.83	0.79	0.81	1190
Bipolar Disorde	0.91	0.77	0.83	882
Birth Control	0.92	0.99	0.96	5807
Depression	0.78	0.86	0.82	1804
Diabetes, Type 2	0.94	0.86	0.90	497
Emergency Contraception	0.99	0.94	0.96	532
High Blood Pressure	0.89	0.78	0.83	451
Insomnia	0.84	0.83	0.83	707
Obesity	0.81	0.72	0.77	687
Pain	0.89	0.92	0.90	1289
Vaginal Yeast Infection	0.94	0.93	0.94	459
Weight Loss	0.80	0.80	0.80	713
accuracy			0.89	17132
macro avg	0.89	0.84	0.86	17132
weighted avg	0.89	0.89	0.89	17132

The overall accuracy of the model is 89%, indicating the proportion of correctly classified instances across all classes.

In summary, the Random Forest model demonstrates high accuracy and balanced performance in classifying patient conditions. Classes like "Birth Control" exhibit particularly strong predictive capabilities, while the overall macro and weighted averages emphasize the model's effectiveness across diverse medical conditions.

4.Support vector Machine (SVM) :

In the subsequent stage, I introduced the 'Support Vector Machine' (SVM) algorithm for classification tasks. SVM's ability to handle high-dimensional data and its effectiveness in distinguishing between different classes made it a fitting choice for the project. The model was configured to optimize both precision and recall, aiming to strike a balance between accurate classification and capturing diverse patient conditions expressed in drug reviews.



Accuracy (SVM): 0.86

	precision	recall	f1-score	support
ADHD	0.93	0.85	0.89	657
Abnormal Uterine Bleeding	0.71	0.33	0.45	376
Acne	0.95	0.87	0.91	1081
Anxiety	0.76	0.75	0.75	1190
Bipolar Disorde	0.84	0.73	0.78	882
Birth Control	0.93	0.97	0.95	5807
Depression	0.72	0.81	0.77	1804
Diabetes, Type 2	0.89	0.87	0.88	497
Emergency Contraception	1.00	0.97	0.98	532
High Blood Pressure	0.85	0.83	0.84	451
Insomnia	0.84	0.83	0.83	707
Obesity	0.68	0.64	0.66	687
Pain	0.89	0.93	0.91	1289
Vaginal Yeast Infection	0.98	0.93	0.96	459
Weight Loss	0.70	0.68	0.69	713
accuracy			0.86	17132
macro avg	0.84	0.80	0.82	17132
weighted avg	0.86	0.86	0.86	17132

The Support Vector Machine (SVM) model achieves an accuracy of 86% in classifying patient conditions based on drug reviews.

In summary, the SVM model demonstrates a solid performance in classifying patient conditions, with notable precision and recall values. The balanced F1-scores and overall accuracy affirm its effectiveness across a spectrum of medical categories

Overall Summary Of Classification Models:

1. Random Forest:
    - Accuracy : 89%
    - Key Observations: Achieves high accuracy and balanced performance in categorizing patient conditions. Notable precision and recall values across various health concerns, with particularly strong performance in "Birth Control."

2. Support Vector Machine (SVM):
    - Accuracy : 86%
    - Key Observation: Demonstrates a solid performance, with balanced precision and recall. Notable recall values in classes like "Birth Control" and "Emergency Contraception."

3. Naïve Bayes:
    - Accuracy : 81%
    - Key Observation: Shows robust performance in classifying patient conditions. Notable precision and recall values in various health concerns, including "Emergency Contraception" and "ADHD."

4. Multinomial Naïve bayes(NB):
    - Accuracy: 81%
    - Key Observation: Effective in categorizing patient conditions related to top 15 health concerns. Strong precision and recall values in classes like "Emergency Contraception" and "Weight Loss."

Common Observation:

- All models exhibit solid accuracy, with Random Forest leading at 89%.
- Precision and recall values vary across health concerns, emphasizing the need for tailored model evaluation based on specific medical categories.
- "Birth Control" consistently shows high performance in all models, reflecting distinctive patterns in patient reviews for this condition.
- The models demonstrate a balanced trade-off between precision and recall, crucial for reliable classification of diverse patient conditions.

In conclusion, the classification models collectively showcase effectiveness in categorizing patient conditions using drug reviews. The choice of model may be tailored based on specific considerations such as precision, recall, or overall accuracy, depending on the priorities of the classification task.

Reference

[1] UCI. *Drug Review Dataset (Drugs.com) Data Set*. Retrieved from https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com

[2] Kim, S., & Liu, H. (2017). Toward a Fully Automated Deep Learning System for Predicting Diagnoses and Classifying Eeg Records in Real-Time. *Neuroinformatics*, *15*(4), 349–361. [DOI: 10.1007/s12021-017-9331-6]

[3] Thu Dinh and Goutam Chakraborty. 2020. Detecting Side Effects and Evaluating the Effectiveness of Drugs from Customers' Online Reviews using Text Analytics, Sentiment Analysis, and Machine Learning Models. sas-global-forum-proceedings (2020), 1–23

[4] Sairamvinay Vijayaraghavan and Debraj Basu. *Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms*. Retrieved from. https://arxiv.org/abs/2003.11643

[5] Manek, A. S., Pallavi, R. P., Bhat, V. H., Shenoy, D. P., Mohan, M. C., Venugopal, K. R. and Patnaik, L. M. (2013). Sentrep: Sentiment classification of movie reviews using efficient repetitive pre-processing, 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013), pp. 1–5.

[6] Natural Language Processing from Scratch – http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35671.pd f

[7] Cambria, Erik; Schuller, Björn; Xia, Yunqing; Havasi, Catherine (2013). "New Avenues in Opinion Mining and Sentiment Analysis". IEEE Intelligent Systems

[8] Snyder, Benjamin; Barzilay, Regina (2007). "Multiple Aspect Ranking using the Good Grief Algorithm". Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference

[9] Chirag Sangani Stanford University, USA "Sentiment Analysis of App Store Reviewx, Susannah, and Maeve Duggan. "Health online 2013. 2013." URL:http://pewinternet.org/Reports/2013/Health-online.aspx

Appendix:

Data Importing

```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings("ignore")
```
In [1]:

```
data = pd.read_csv('D:\Drug\Drugdata.tsv', sep='\t')
data
```
In [2]:

```
data.info()
null_counts = data.isnull().sum()
```
In [3]:

```
print(null_counts)
```
In [4]:

```
# Remove unnecessary columns

df = data[['drugName', 'condition', 'review', 'rating']]
```
In [5]:

```
# Handle missing data if needed

df.dropna(inplace=True)
df
```
In [6]:

```
# Check the distribution of conditions
condition_counts = df['condition'].value_counts()
print("Distribution of Conditions:")
print(condition_counts)
```
In [7]:

Data Preprocessing / Text Mining

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
import re
```
In [8]:

```
# Convert text to lowercase
df['review'] = df['review'].str.lower()
```
In [9]:

```
# Remove special characters and numbers
df['review'] = df['review'].apply(lambda x: re.sub(r'[^a-zA-Z\s]', '', x))
```
In [10]:

```
# Tokenization
df['review'] = df['review'].apply(lambda x: word_tokenize(x))
```
In [11]:

```
# Remove stop words
stop_words = set(stopwords.words('english'))
df['review'] = df['review'].apply(lambda x: [word for word in x if word not in
stop_words])
```

In [12]:

```
stopwords.words('english')
```

```
#import nltk
#nltk.download('wordnet')
#nltk.download('omw-1.4')
# Lemmatization
lemmatizer = WordNetLemmatizer()
df['review'] = df['review'].apply(lambda x: [lemmatizer.lemmatize(word) for
word in x])
```

In [14]:

```
df['review'][5]
```

```
# Join tokens back into sentences
df['review'] = df['review'].apply(lambda x: ' '.join(x))
```

In [16]:

```
df
```

Data Visualization

```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [18]:

```
# Distribution of Ratings
plt.figure(figsize=(10, 6))
sns.countplot(x='rating', data=df)
plt.title('Distribution of Ratings')
plt.xlabel('Rating')
plt.ylabel('Count')
plt.show()
```

```
# Top 10 Drugs by Review Count
top_drugs = df['drugName'].value_counts().nlargest(20)
plt.figure(figsize=(12, 6))
top_drugs.plot(kind='bar', color='skyblue')
plt.title('Top 10 Drugs by Review Count')
plt.xlabel('Drug Name')
plt.ylabel('Review Count')
plt.xticks(rotation=45)
plt.show()

# Distribution of Conditions
plt.figure(figsize=(12, 8))
sns.countplot(y='condition', data=df,
order=df['condition'].value_counts().index[:15])
plt.title('Distribution of Top 15 Conditions')
plt.xlabel('Count')
plt.ylabel('Condition')
plt.show()
```

```
# Plot the distribution of ratings
plt.figure(figsize=(8, 5))
sns.histplot(df['rating'], bins=5, kde=True, color='orange')
plt.title('Distribution of Ratings')
plt.xlabel('Rating')
plt.ylabel('Count')
plt.show()

# Plot box plots of ratings for top conditions
top_conditions = condition_counts.head(5).index
plt.figure(figsize=(12, 6))
sns.boxplot(x='rating', y='condition', data=df, order=top_conditions,
palette='Set2')
plt.title('Box Plot of Ratings by Condition')
plt.xlabel('Rating')
plt.ylabel('Condition')
plt.show()


from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Assuming you have a DataFrame df_top_conditions with columns 'condition' and
'review'
df_top_conditions = pd.read_csv('D:\Drug\Drugdata.tsv', sep='\t')
top_conditions = condition_counts.head(5).index
df_top_conditions =
df_top_conditions[df_top_conditions['condition'].isin(top_conditions)]

# Combine all reviews for each condition
condition_reviews = {}
for condition in top_conditions:
    reviews = " ".join(df_top_conditions[df_top_conditions['condition'] ==
condition]['review'])
    condition_reviews[condition] = reviews

# Generate and display word clouds
for condition, reviews in condition_reviews.items():
    wordcloud = WordCloud(width=800, height=400,
background_color='black').generate(reviews)

    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.title(f'Word Cloud for {condition}')
    plt.axis('off')
    plt.show()
```

Classification:

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report
```

In [26]:

```
tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_features=5000)
X = tfidf_vectorizer.fit_transform(df['review'])
y = df['condition']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Additional: Check class distribution in test set
test_condition_counts = pd.Series(y_test).value_counts()
print("Distribution of Conditions in Test Set:")
print(test_condition_counts)


# Filter the data to include only reviews related to the first 15 most common
conditions
top_conditions = condition_counts.head(15).index
df_top_conditions = df[df['condition'].isin(top_conditions)]
top_conditions

# Split the data into training and testing sets
X_top_conditions = tfidf_vectorizer.transform(df_top_conditions['review'])
y_top_conditions = df_top_conditions['condition']
X_train_top, X_test_top, y_train_top, y_test_top =
train_test_split(X_top_conditions, y_top_conditions, test_size=0.2,
random_state=42)

model_top_conditions = MultinomialNB()
model_top_conditions.fit(X_train_top, y_train_top)


# Naive Bayes
model_nb = MultinomialNB()
model_nb.fit(X_train_top, y_train_top)


# Random Forest
from sklearn.ensemble import RandomForestClassifier
model_rf = RandomForestClassifier(n_estimators=100, random_state=42)
model_rf.fit(X_train_top, y_train_top)

# Support Vector Machine (SVM)
from sklearn.svm import SVC
model_svm = SVC(kernel='linear', C=1)
model_svm.fit(X_train_top, y_train_top)
```

```
# Step 6: Make predictions and evaluate the model
y_pred_top_conditions = model_top_conditions.predict(X_test_top)

# Evaluate accuracy
accuracy_top_conditions = accuracy_score(y_test_top, y_pred_top_conditions)
print(f'Accuracy on Top 10 Conditions: {accuracy_top_conditions:.2f}')

# Display classification report
print(classification_report(y_test_top, y_pred_top_conditions))

# Naive Bayes
y_pred_nb = model_nb.predict(X_test_top)
accuracy_nb = accuracy_score(y_test_top, y_pred_nb)
print(f'Accuracy (Naive Bayes): {accuracy_nb:.2f}')
print(classification_report(y_test_top, y_pred_nb))


# Random Forest
y_pred_rf = model_rf.predict(X_test_top)
accuracy_rf = accuracy_score(y_test_top, y_pred_rf)
print(f'Accuracy (Random Forest): {accuracy_rf:.2f}')
print(classification_report(y_test_top, y_pred_rf))


# SVM
y_pred_svm = model_svm.predict(X_test_top)
accuracy_svm = accuracy_score(y_test_top, y_pred_svm)
print(f'Accuracy (SVM): {accuracy_svm:.2f}')
print(classification_report(y_test_top, y_pred_svm))
```

Thank you.....