EVALUATING PATIENT SATISFACTION THROUGH DRUG REVIEWS

Vidhi Patel

Introduction

Background:

The pharmaceutical industry continuously seeks to understand patient experiences with medications to improve drug efficacy, safety, and patient satisfaction. Traditionally, this understanding has been gleaned from clinical trials, post-marketing surveillance, and direct feedback from healthcare professionals. [1] However, these methods can be limited in scope, costly, and may not fully capture the patient's perspective. With the advent of online platforms, forums, and social media, patients increasingly share their experiences with medications through reviews. These text-based reviews are rich, unstructured data sources that offer direct insights into patient sentiment, side effects, effectiveness, and overall satisfaction with treatment. [2] Machine learning, particularly NLP (Natural Language Processing) techniques, can automate the analysis of this vast amount of text data, identifying patterns, sentiments, and correlations that would be impractical to analyse manually.

Motivation:

Understanding patient satisfaction with specific medications is vital for healthcare professionals, pharmaceutical companies, and regulatory agencies. The motivation behind this project is that by analysing drugs we can help identify trends in patient sentiment, highlight areas for improvement, and contribute to the enhancement of patient-centred care. [3] Moreover, it can also help in making informed decisions about drug prescription and patient education.

Goal:

The primary goal of this project is to build a predictive model that can evaluate patient satisfaction based on drug reviews. [4] By utilizing natural language processing (NLP) techniques and machine learning algorithms, we can compare the performance of different models based on their F1 score, recall, and precision, accuracy and select the best model. [5] We aim to predict sentiment labels, identifying factors that influences patient’s satisfaction towards healthcare and medicinal drugs. It will also provide actionable insights for the healthcare professionals.

Methodology

Data Extraction:

In the data extraction phase of this project, the existing dataset is imported into the analysis environment using Python’s pandas library. The appropriate function is selected based on the file format—read_csv() for CSV. This phase involves a preliminary data examination to understand the dataset's structure with functions like head(), describe(), and info(). This step is crucial to identifying any missing values or potential inconsistencies. We explored the dataset from the UCI Machine Learning Repository consisting of about 215,063 sample. Each sample includes fields like Drug Name, condition, user review, rating, review date, and useful count. We focused on text analysis to predict review sentiment, categorizing ratings from 1 to 10 into positive, negative, and neutral classes.

Data preprocessing & feature engineering:

In this stage, we utilized various NLP techniques to extract key features from the textual data. The process began with tokenization and lemmatization, essential for standardizing the text for analysis. We also implemented sentiment analysis to identify and extract subjective expressions from user reviews that convey user experiences. As part of the data preprocessing for NLP analysis, the text was normalized by converting it to lowercase, and noise reduction was achieved by removing stopwords and punctuation. Finally, the clean, structured data was saved, setting the stage for detailed sentiment analysis and the subsequent modeling phases.

Exploratory Data Analysis (EDA):

In this stage, we conducted an initial analysis using descriptive statistics to gain insights into the central tendencies, dispersion, and overall shape of the dataset's distributions. We visualized the distribution of ratings across different conditions using various plots, such as histograms, box plots, and scatter plots, to identify any inherent biases or trends in patient feedback and explore relationships between numerical features and sentiment ratings. [7] Additionally, we used correlation matrices to investigate potential linear relationships or multicollinearity among the engineered features. Sentiment analysis was performed on sample reviews, and the frequency of positive, negative, and neutral sentiments was plotted to better understand common patient perceptions. We also investigated the presence of outliers or anomalies in the data, particularly focusing on features like the useful count, which might affect the model's performance.

Sentiment Analysis:

Sentiment Analysis is the technique used to determine and understand the emotions expressed in a piece of text, such as reviews, comments, or social media posts. [6] It involved analysing the language to identify whether the sentiment is positive, negative, or neutral. This is a useful technique for the project as it will give a way to measure the patient satisfaction. We can find patterns in patients’ feelings and highlight the areas that need attention. It’s a practical tool to guide decisions and make healthcare better by focusing on what matters most to patients.

Model Selection:

Several machine learning models were evaluated, including Logistic Regression, Random Forest, Decision Tree, SVM, and Naive Bayes. Each model was trained on the training dataset and validated using a separate testing set to ensure robustness and prevent overfitting.

Evaluation and Optimization:

The models were assessed based on accuracy, precision, recall, and F1-score. Hyperparameter tuning was performed to optimize each model's performance. The best-performing model was then selected for deployment based on its ability to accurately classify sentiments in new drug reviews.

Machine Learning:

We have used the following supervised classification algorithms:

1. **Naive Bayes Classifier:** The Naive Bayes Classifier is a statistical algorithm that applies Bayes' theorem for classification tasks, with an assumption of independence among predictors. This algorithm will be applied to predict patient satisfaction from drug reviews, leveraging its probabilistic approach that assumes feature independence. It is particularly useful for large datasets and works well with text classification, making it ideal for analysing sentiment in patient feedback.

2. **Support Vector Machines (SVM):** Support Vector Machines are well-suited for text classification tasks due to their ability to handle high-dimensional data. They work effectively in both linear and non-linear scenarios, making it suitable for analysing complex relationships in drug reviews. It will help in identifying key features and sentiments that are influencing patient satisfaction.

3. **Decision Tree:** A Decision Tree is a flowchart-like tree structure where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. It will be employed for its interpretability, mapping out the decisions that lead to the prediction of patient satisfaction. This model is beneficial for understanding the factors that contribute most significantly to patient sentiments.

4. **Logistic Regression:** Logistic Regression is a predictive analysis algorithm used for binary classification problems, modelling the probability of a default class based on one or more independent variables. [8] Logistic Regression will be utilized to model the probability of patient satisfaction, offering a straightforward and efficient way to assess the influence of review features on the satisfaction outcome. It's particularly adept at binary classification problems, such as determining if a review is positive or negative.

5. **Random Forest Classifier:** Random Forests excel in capturing complex relationships within data and mitigating overfitting by aggregating the predictions of multiple decision trees. This ensemble learning method combines the outputs of numerous individual decision trees, each trained on a subset of the dataset and with a random subset of features. The collective decision-making process of these trees enhances the model's predictive performance and generalization capabilities.

Dataset Description

This dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating reflecting the overall patient satisfaction. The data was obtained by crawling online pharmaceutical review sites.

Data Size	110.63 MB
Data Types	int64 / string / date
Target	review column is the target variable for sentiments
Features	**drugName**
	condition
	review
	rating
	date
	usefulcount
	uniqueId

DrugName (categorical): The name of the drug that the patient is reviewing. This feature will be used to group reviews by drug and analyze the effectiveness of each drug for specific conditions.

Condition (categorical): The name of the condition that the patient is reviewing the drug for. This feature will be used to identify reviews related to Depression, Anxiety, High Blood Pressure, and Type 2 Diabetes.

Review (text): The patient's review of the drug. This feature will be used to extract insights on the effectiveness and potential side effects of drugs for specific conditions.

Rating (numerical): A 10-star patient rating reflecting overall patient satisfaction with the drug. This feature will be used to understand the level of patient satisfaction with different drugs for specific conditions.

Date (date): The date on which the review was entered. This feature will be used to analyze trends over time in patient reviews and ratings.

UsefulCount (numerical): The number of users who found the review useful. This feature will be used to identify reviews that are likely to clarify the effectiveness and potential side effects of drugs for specific conditions**.**

Dataset Link : **https://www.kaggle.com/code/harshjain123/drugs-review-sentiment/input**

Data Cleaning and Processing:

Importing the required libraries and reading the dataset

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly as px
%matplotlib inline
from wordcloud import WordCloud
from wordcloud import STOPWORDS
import nltk
import string
#nltk.download('punkt')
#nltk.download('stopwords')
from nltk.corpus import stopwords
from textblob import TextBlob
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report,confusion_matrix, mean_squared_error
```

In [2]:
```
data = pd.read_csv("data.csv")
```

```
/var/folders/c2/gthqp2qj1kl6ktjxxxfjgdqr0000gn/T/ipykernel_45135/1572660008.py:1: DtypeWarning: Columns (0,6) have
mixed types. Specify dtype option on import or set low_memory=False.
  data = pd.read_csv("data.csv")
```

In [3]:
```
data.head()
```

Out[3]:

	uniqueID	drugName	condition	review	rating	date	usefulCount	effectiveness	sideEffectsReview	commentsReview	sideEffects
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27	Considerably Effective	SOMETIMES TROUBLE BREATHING AND TROUBLE URINAT...	I used five rings daily, before every intimate...	Extremely Severe Side Effects
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192	Marginally Effective	muscle pain, loss of mobility, depression, head...	I take the drug once a day at night with a sma...	Extremely Severe Side Effects
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17	Moderately Effective	I have only had one side effect due to mixing ...	Just take the pills every 8 hours to tame the ...	Moderate Side Effects
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	03-Nov-15	10	Considerably Effective	reddness, flaking, sensitive skin. Was not abl...	take once per day in the morning	Moderate Side Effects
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37	Considerably Effective	My only major complaint is that since Suboxone...	This was part of a treatment for Adult ADD. I...	Extremely Severe Side Effects

In [5]:
```
data.describe()
```

Out[5]:

	rating
count	164403.000000
mean	6.994635
std	3.266290
min	1.000000
25%	5.000000
50%	8.000000
75%	10.000000
max	10.000000

- The dataset contains 1,644,043 ratings, which suggests a substantial volume of data for analysis.
- The average rating is approximately 6.99, indicating a generally positive trend in the dataset.
- The ratings range from a minimum of 1 to a maximum of 10, with the median rating being 8, which shows a high central tendency towards the upper end of the rating scale.
- The initial dataset consisted of 1,644,043 entries with 11 variables, but it contained 900 missing values in the 'condition' variable.

- After cleaning, the dataset was reduced to 1,635,03 entries by removing missing values, ensuring a more accurate and reliable dataset for analysis.
- The dataset also had 31 duplicate entries which were identified and removed, further refining the data quality.

```
In [6]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 164403 entries, 0 to 164402
Data columns (total 11 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   uniqueID           164403 non-null  object
 1   drugName           164403 non-null  object
 2   condition          163503 non-null  object
 3   review             164403 non-null  object
 4   rating             164403 non-null  int64
 5   date               164403 non-null  object
 6   usefulCount        164403 non-null  object
 7   effectiveness      164403 non-null  object
 8   sideEffectsReview  164403 non-null  object
 9   commentsReview     164403 non-null  object
 10  sideEffects        164403 non-null  object
dtypes: int64(1), object(10)
memory usage: 13.8+ MB
```

```
In [9]: data.isnull().sum()

Out[9]: uniqueID               0
        drugName               0
        condition            900
        review                 0
        rating                 0
        date                   0
        usefulCount            0
        effectiveness          0
        sideEffectsReview      0
        commentsReview         0
        sideEffects            0
        dtype: int64
```

```
In [10]: data.shape

Out[10]: (164403, 11)
```

```
In [11]: data.dropna(inplace=True)
```

```
In [12]: data.isnull().any()

Out[12]: uniqueID             False
         drugName             False
         condition            False
         review               False
         rating               False
         date                 False
         usefulCount          False
         effectiveness        False
         sideEffectsReview    False
         commentsReview       False
         sideEffects          False
         dtype: bool
```

```
In [13]: data.shape

Out[13]: (163503, 11)
```

```
In [14]: data.duplicated().sum()

Out[14]: 31
```

```
In [15]:
# Drop duplicate rows
data.drop_duplicates(inplace=True)
```

```
In [16]: keep_conditions = ["ADHD", "Anxiety", "Insomnia","Weight Loss","Migraine", "Obesity","Depression", "High Blood Press

         # Filter the DataFrame to only keep the records with the specified conditions
         df = data[data['condition'].isin(keep_conditions)]

         df.drop(['uniqueID'],axis =1,inplace=True)

         df.head()
```

Out[16]:

	drugName	condition	review	rating	date	usefulCount	effectiveness	sideEffectsReview	commentsReview	sideEffects
1	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192	Marginally Effective	muscle pain, loss of mobility, depresion, head...	I take the drug once a day at night with a sma...	Extremely Severe Side Effects
11	L-methylfolate	Depression	"I have taken anti-depressants for years, with...	10	09-Mar-17	54	Highly Effective	Increased risks for breast cancer and conditio...	Prescribed to take whenever flare-up for perio...	Extremely Severe Side Effects
15	Liraglutide	Obesity	"I have been taking Saxenda since July 2016. ...	9	19-Jan-17	20	Highly Effective	Mild drowsiness accompanied by a sense of well...	took one 5mg tablet daily in the am	No Side Effects
21	Trazodone	Insomnia	[illegible]ave insomnia, it's horrible. My story...	10	03-Apr-16	43	Moderately Effective	I have found that if I take it too early prior...	I pill a day and all has been well. =D	Mild Side Effects
27	Daytrana	ADHD	"Hi all, My son who is 12 was diagnosed when h...	10	12-Jan-17	11	Ineffective	I always experiencing oily bowel movement. I s...	The initial skin infection is clearing up.	Mild Side Effects

click to scroll output; double click to hide

```
In [17]: df.shape

Out[17]: (35774, 10)
```

In the preprocessing stage, the data was meticulously filtered to retain only those records pertaining to conditions deemed most pertinent to the study's focus: ADHD, Anxiety, Insomnia, Weight Loss, Migraine, Obesity, Depression, and High Blood Pressure. This targeted approach ensured a tailored dataset, conducive to a more nuanced analysis of drug reviews.

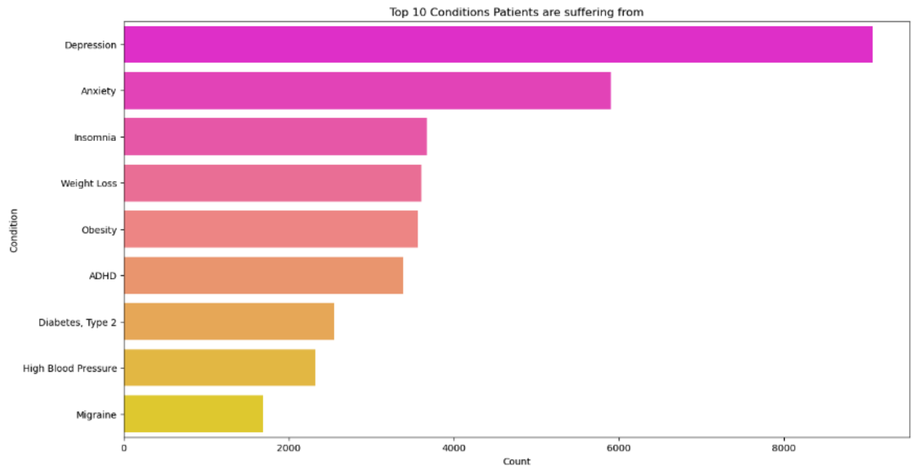
```
In [19]: # Top 20 most popular drugs
         df['drugName'].value_counts().nlargest(20)

Out[19]: Phentermine               1515
         Bupropion / naltrexone     945
         Contrave                   912
         Escitalopram               864
         Liraglutide                748
         Lexapro                    678
         Bupropion                  667
         Venlafaxine                586
         Lorcaserin                 572
         Belviq                     562
         Desvenlafaxine             515
         Alprazolam                 489
         Pristiq                    486
         Trazodone                  481
         Zolpidem                   477
         Mirtazapine                475
         Clonazepam                 462
         Sertraline                 459
         Duloxetine                 453
         Cymbalta                   427
         Name: drugName, dtype: int64
```

Results and Analysis

EDA (Exploratory Data Analysis):

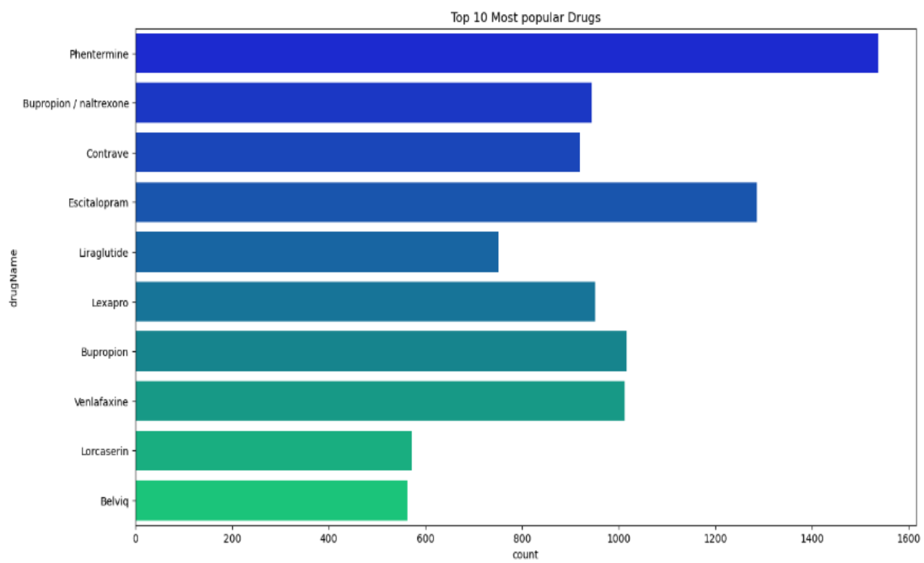
First, we started with understanding what are the most frequent conditions that patients are suffering from. We visualized the conditions with the number of counts.

```
In [20]: plt.figure(figsize=(15,8))
         sns.countplot(data=data, y='condition', order=df['condition'].value_counts().index[:10], palette="spring")
         plt.title('Top 10 Conditions Patients are suffering from')
         plt.xlabel('Count')
         plt.ylabel('Condition')
         plt.show()
```



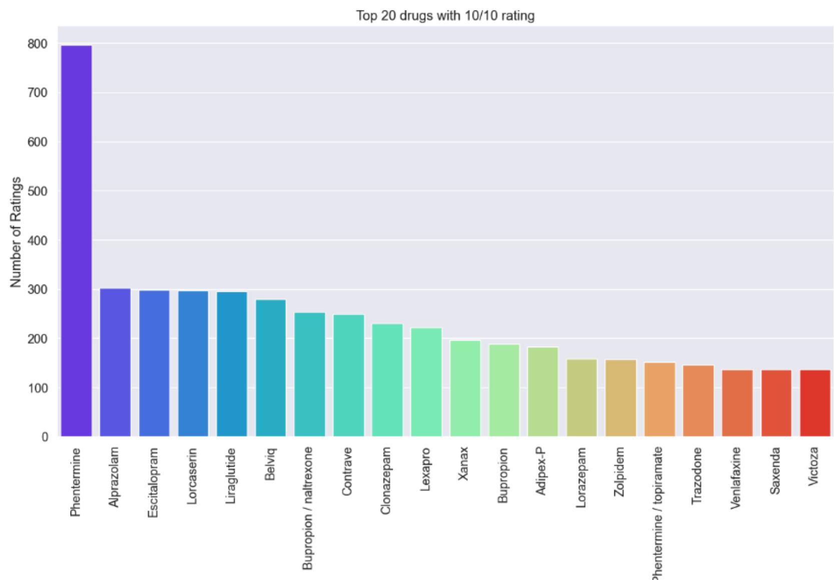
We also plotted the frequent drugs used by the patients by the drug names.

```
In [21]: plt.figure(figsize=(15,8))
         sns.countplot(data=data, y='drugName', order=df['drugName'].value_counts().index[:10], palette="winter")
         plt.title('Top 10 Most popular Drugs')
         plt.show()
```

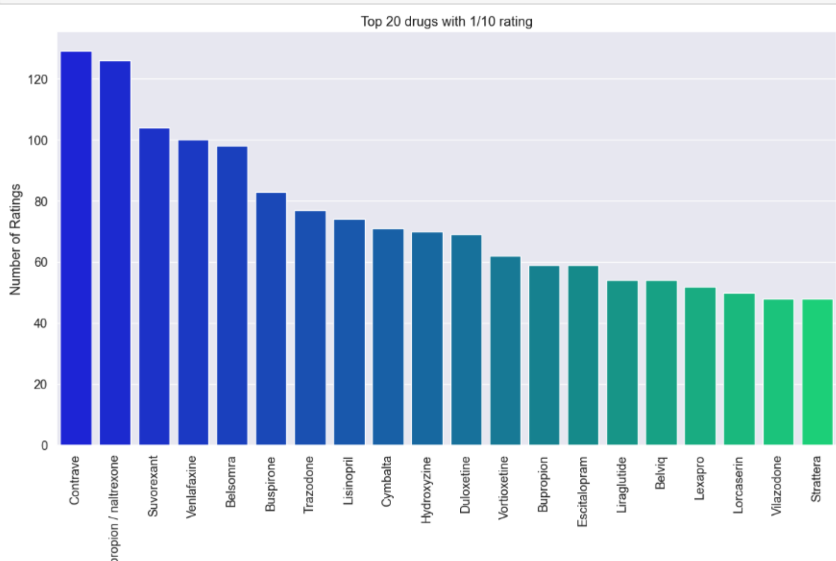


Then we decided to check that which are the drugs that have 10 stars rating and which are the drugs with 1-star ratings.

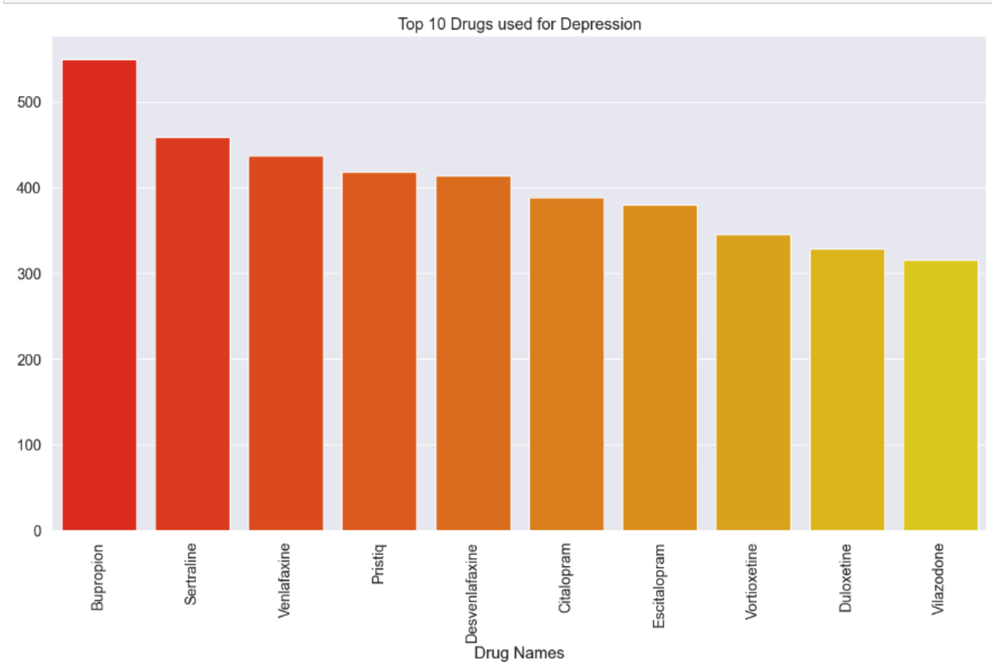
In [24]:
```
sns.set(font_scale = 1.2, style = 'darkgrid')
plt.rcParams['figure.figsize'] = [15, 8]

rating = dict(df.loc[df.rating == 10, "drugName"].value_counts())
drugname = list(rating.keys())
drug_rating = list(rating.values())

sns_rating = sns.barplot(x = drugname[0:20], y = drug_rating[0:20], palette = 'rainbow')

sns_rating.set_title('Top 20 drugs with 10/10 rating')
sns_rating.set_ylabel("Number of Ratings")
sns_rating.set_xlabel("Drug Names")
plt.setp(sns_rating.get_xticklabels(), rotation=90);
```



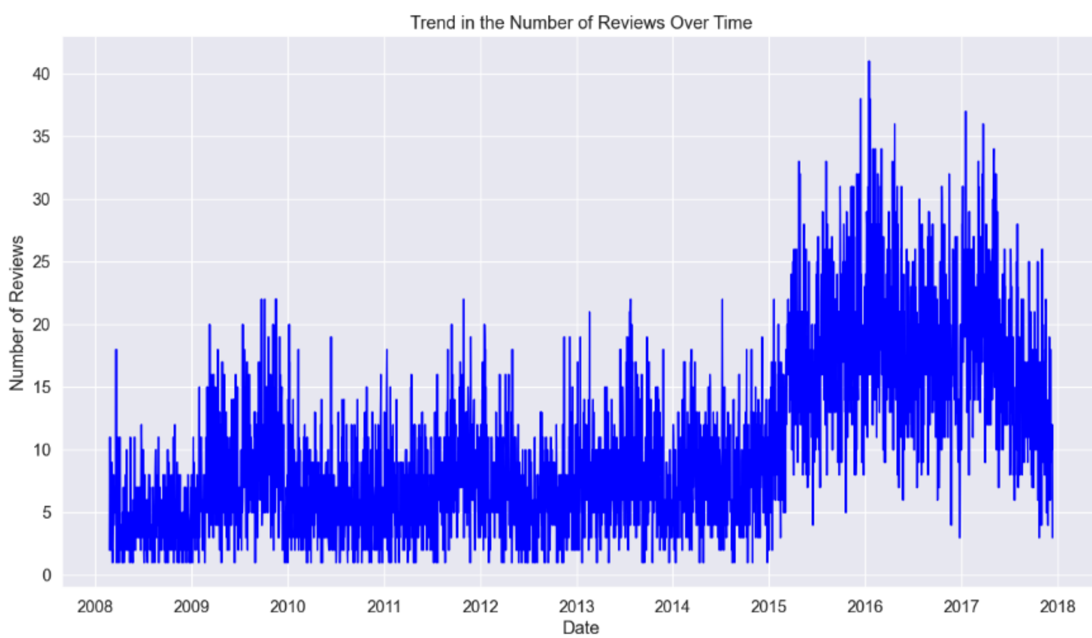
In [25]:
```
sns.set(font_scale = 1.2, style = 'darkgrid')
plt.rcParams['figure.figsize'] = [15, 8]

rating = dict(df.loc[df.rating == 1, "drugName"].value_counts())
drugname = list(rating.keys())
drug_rating = list(rating.values())

sns_rating = sns.barplot(x = drugname[0:20], y = drug_rating[0:20], palette = 'winter')

sns_rating.set_title('Top 20 drugs with 1/10 rating')
sns_rating.set_ylabel("Number of Ratings")
sns_rating.set_xlabel("Drug Names")
plt.setp(sns_rating.get_xticklabels(), rotation=90);
```



Further, we created a bar chart to showcase the top 10 drugs that are used for depression.



For the below data, we wanted to show the number of patient reviews through time. We noticed that there are increased number of patients using drug and writing reviews as time passed. This could be due to either decreased cost of drugs or more availability of drugs. It could also be due to increase in the use of online platforms for reviewing drugs. Lastly, it could be possible due to the healthcare industry changing their marketing strategies and increasing the drug accessibility.



```
In [29]: # let's make a new column named "review sentiment"

         df.loc[(df['rating'] >= 5), 'Review_Sentiment'] = 1
         df.loc[(df['rating'] < 5), 'Review_Sentiment'] = 0

         df['Review_Sentiment'].value_counts()

Out[29]: 1.0    28298
         0.0     7476
         Name: Review_Sentiment, dtype: int64
```

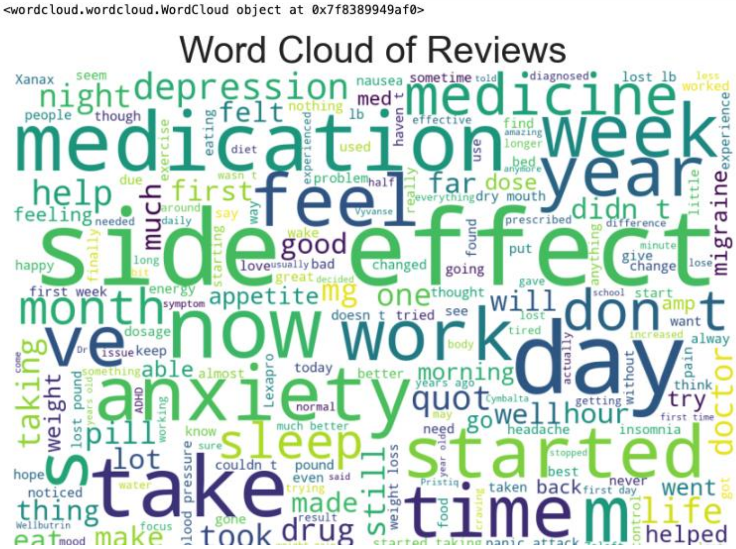
A critical part of the analysis focused on sentiment classification based on user ratings. To facilitate this, a new feature named Review_Sentiment was engineered:

Sentiment Categorization Logic: Ratings were categorized into two sentiments. Ratings equal to or above 5 were labeled as positive (1), and ratings below 5 as negative (0). This binary classification allows for a straightforward understanding of user sentiment in the dataset.

Sentiment Distribution Insights: Upon applying this logic to the dataset, it was found that most reviews were positive, with 28,298 instances. In contrast, 7,476 reviews were categorized as negative.

Implications for Model Training: This distribution is crucial for training classification models as it indicates a significant class imbalance that models will need to account for. Strategies such as class weight adjustment or resampling might be necessary to ensure model robustness.

```
In [30]: sentiments = " ".join([text for text in df['review'][df['Review_Sentiment'] == 1]])

         from wordcloud import WordCloud
         from wordcloud import STOPWORDS

         stopwords = set(STOPWORDS)
         wordcloud = WordCloud(background_color = 'white', stopwords = stopwords, width = 1200, height = 800).generate(sentim

         plt.rcParams['figure.figsize'] = (20, 7)
         plt.title('Word Cloud of Reviews', fontsize = 30)
         print(wordcloud)
         plt.axis('off')
         plt.imshow(wordcloud)
         plt.show()

<wordcloud.wordcloud.WordCloud object at 0x7f8389949af0>
```



Words related to specific conditions like "depression", "anxiety", and "migraine" provide insight into the types of health issues being addressed. This could indicate the effectiveness of medications in treating these conditions as perceived by users.

Utility in Model Improvement: The Word Cloud also reveals terms that could be influential in sentiment analysis and could be used to refine NLP models. The prevalence of these terms can aid in feature selection for machine learning algorithms.

In [31]:
```
def get_sentiment(text):
    blob = TextBlob(text)
    return blob.polarity

def get_sentiment_label(text):
    blob = TextBlob(text)
    if blob.polarity > 0:
        result = 'positive'
    elif blob.polarity < 0:
        result = 'negative'
    else:
        result = 'neutral'
    return result
```

In [32]:
```
get_sentiment_label("I love this medicine")
```

click to scroll output; double click to hide

Out[32]:
```
'positive'
```

In [33]:
```
get_sentiment_label("I hate this medicine")
```

Out[33]:
```
'negative'
```

We used function **get_sentiment_label** to get an idea of whether the review is positive or negative. Below, we summarized as to how many sentiments each for positive, negative, and neutral are present in the dataset.

In [36]:
```
df[['review','sentiment','sentiment_label']]
```

Out[36]:

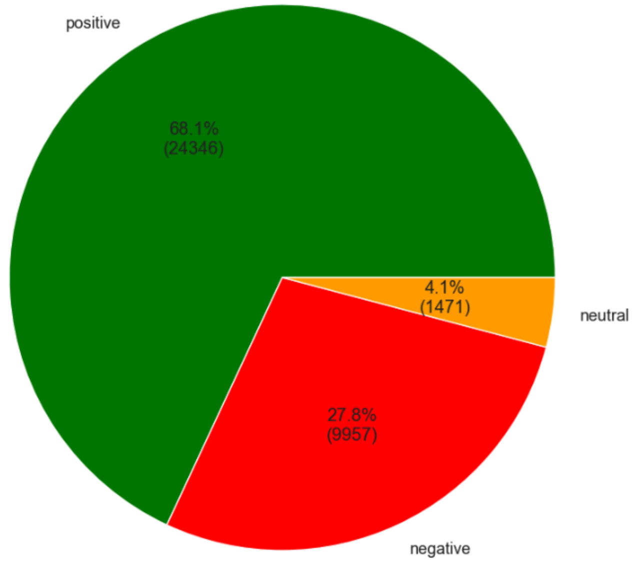
	review	sentiment	sentiment_label
1	"My son is halfway through his fourth week of ...	0.168333	positive
11	"I have taken anti-depressants for years, with...	0.275000	positive
15	"I have been taking Saxenda since July 2016. ...	0.209259	positive
21	"I have insomnia, it's horrible. My story...	0.061503	positive
27	"Hi all, My son who is 12 was diagnosed when h...	0.070798	positive
...	...	...	...
161276	"I started taking this medication 10 years ago...	-0.166667	negative
161277	"I just got diagnosed with type 2. My doctor p...	0.048611	positive
161285	"This is the third med I've tried for anx...	-0.100694	negative
161286	"I was super against taking medication. I&#039...	-0.046667	negative
161289	"I have only been on Tekturna for 9 days. The ...	-0.100000	negative

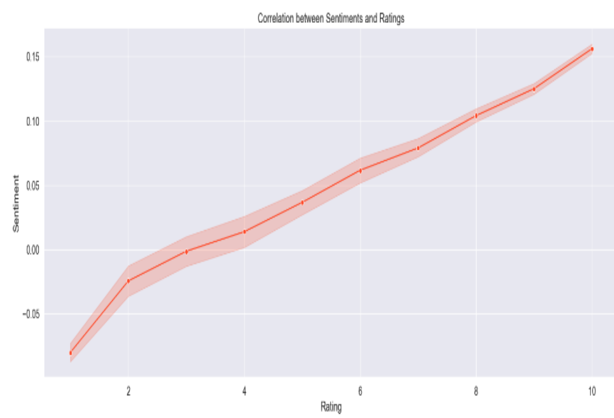
35774 rows × 3 columns

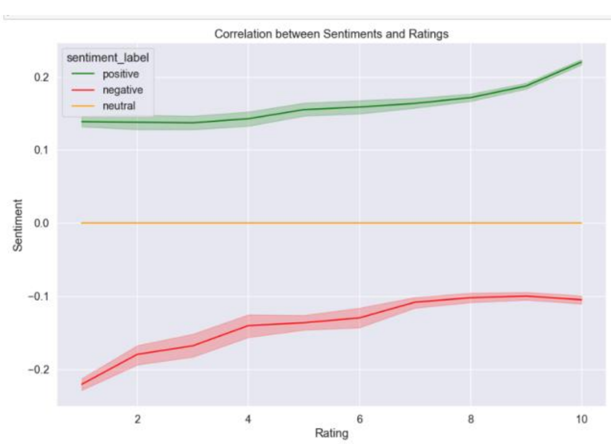
In [44]:
```
df['sentiment_label'].value_counts()
```

Out[44]:
```
positive    24346
negative     9957
neutral      1471
Name: sentiment_label, dtype: int64
```

We found out that most of the patients have reviewed the drugs as Positive (68.1%). There are a few patients who gave Negative (27.8%) reviews. And the rest are Neutral (4.1%).



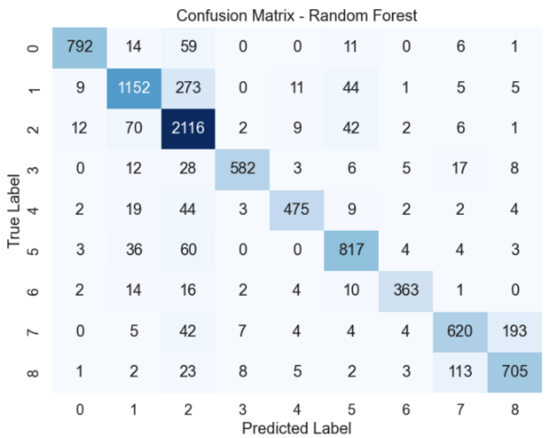




Below we have the Confusion matrix and accuracies of each of the models we used. According to the data, SVM performed the best.

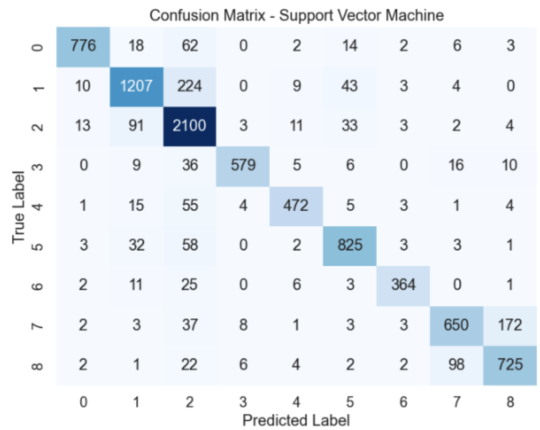
Accuracy of Random Forest model: 0.8521914132379249

	precision	recall	f1-score	support
ADHD	0.96	0.90	0.93	883
Anxiety	0.87	0.77	0.82	1500
Depression	0.80	0.94	0.86	2260
Diabetes, Type 2	0.96	0.88	0.92	661
High Blood Pressure	0.93	0.85	0.89	560
Insomnia	0.86	0.88	0.87	927
Migraine	0.95	0.88	0.91	412
Obesity	0.80	0.71	0.75	879
Weight Loss	0.77	0.82	0.79	862
accuracy			0.85	8944
macro avg	0.88	0.85	0.86	8944
weighted avg	0.86	0.85	0.85	8944



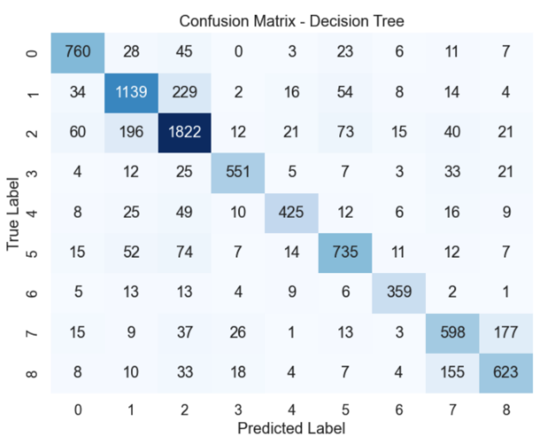
Accuracy of SVM model: 0.8606887298747764

	precision	recall	f1-score	support
ADHD	0.96	0.88	0.92	883
Anxiety	0.87	0.80	0.84	1500
Depression	0.80	0.93	0.86	2260
Diabetes, Type 2	0.96	0.88	0.92	661
High Blood Pressure	0.92	0.84	0.88	560
Insomnia	0.88	0.89	0.89	927
Migraine	0.95	0.88	0.92	412
Obesity	0.83	0.74	0.78	879
Weight Loss	0.79	0.84	0.81	862
accuracy			0.86	8944
macro avg	0.89	0.85	0.87	8944
weighted avg	0.87	0.86	0.86	8944



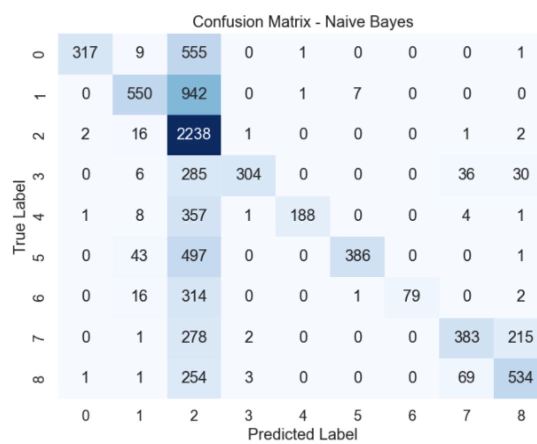
Accuracy of Decision Tree model: 0.7839892665474061

	precision	recall	f1-score	support
ADHD	0.84	0.86	0.85	883
Anxiety	0.77	0.76	0.76	1500
Depression	0.78	0.81	0.79	2260
Diabetes, Type 2	0.87	0.83	0.85	661
High Blood Pressure	0.85	0.76	0.80	560
Insomnia	0.79	0.79	0.79	927
Migraine	0.87	0.87	0.87	412
Obesity	0.68	0.68	0.68	879
Weight Loss	0.72	0.72	0.72	862
accuracy			0.78	8944
macro avg	0.80	0.79	0.79	8944
weighted avg	0.78	0.78	0.78	8944



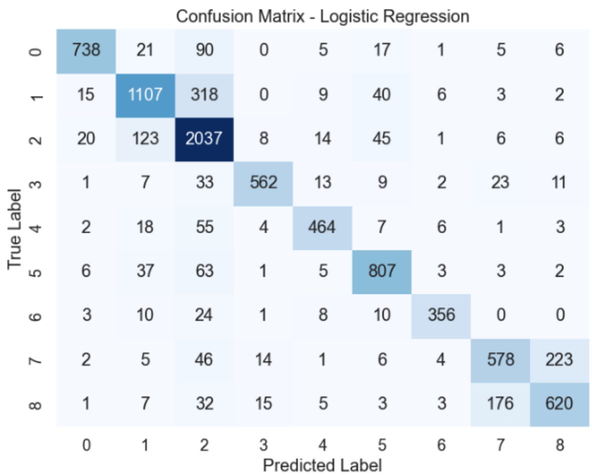
Accuracy of Naive Bayes model: 0.5566860465116279

	precision	recall	f1-score	support
ADHD	0.99	0.36	0.53	883
Anxiety	0.85	0.37	0.51	1500
Depression	0.39	0.99	0.56	2260
Diabetes, Type 2	0.98	0.46	0.63	661
High Blood Pressure	0.99	0.34	0.50	560
Insomnia	0.98	0.42	0.58	927
Migraine	1.00	0.19	0.32	412
Obesity	0.78	0.44	0.56	879
Weight Loss	0.68	0.62	0.65	862
accuracy			0.56	8944
macro avg	0.85	0.46	0.54	8944
weighted avg	0.76	0.56	0.55	8944



Accuracy of Logistic Regression model: 0.8127236135957067

	precision	recall	f1-score	support
ADHD	0.94	0.84	0.88	883
Anxiety	0.83	0.74	0.78	1500
Depression	0.76	0.90	0.82	2260
Diabetes, Type 2	0.93	0.85	0.89	661
High Blood Pressure	0.89	0.83	0.86	560
Insomnia	0.85	0.87	0.86	927
Migraine	0.93	0.86	0.90	412
Obesity	0.73	0.66	0.69	879
Weight Loss	0.71	0.72	0.71	862
accuracy			0.81	8944
macro avg	0.84	0.81	0.82	8944
weighted avg	0.82	0.81	0.81	8944



This is the Accuracy of all the models:

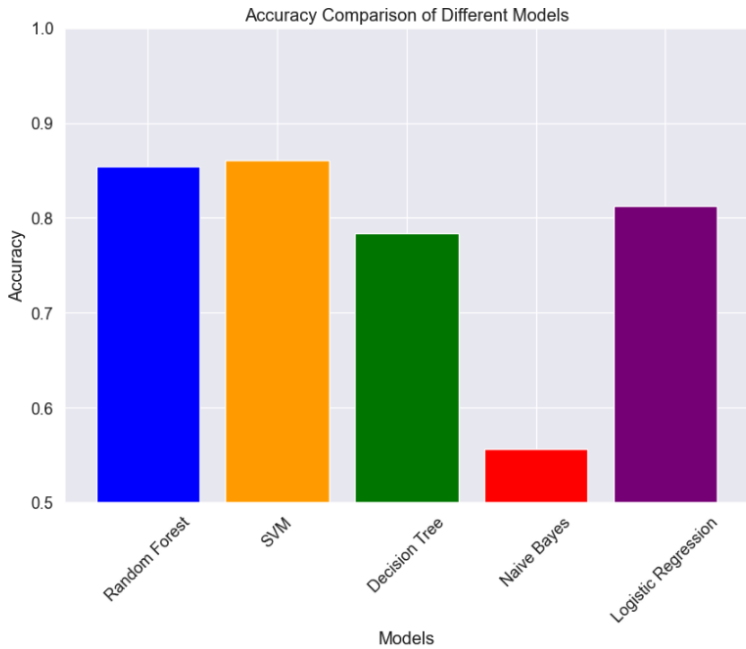
SVM - 86% (This model performed well compared to all other models)

Random Forest – 85%

Naive Bayes –61%

Logistic Regression –81%

Decision Tree- 78%



Why did naive bayes failed: Naive Bayes models are based on the assumption of independence between features, meaning that each feature contributes independently to the probability of a certain outcome. In the context of text classification, Naive Bayes assumes that each word's presence in a document is independent of other words. However, this assumption may not hold true in many real-world scenarios, including sentiment analysis of drug reviews.

Drug reviews often contain complex and nuanced language, with dependencies and correlations between words that may affect the overall sentiment expressed. Naive Bayes may struggle to capture these intricate relationships between words and sentiments, leading to suboptimal performance compared to more sophisticated models like Random Forest and Support Vector Machine (SVM), which are better equipped to handle such complexities.

Additionally, Naive Bayes tends to perform well when the independence assumption holds true or when the feature space is relatively simple. In the case of sentiment analysis of drug reviews, where the language is often rich and varied, this assumption may be too simplistic to accurately capture the underlying sentiment patterns.

Overall, the limitations of Naive Bayes in capturing complex dependencies and nuances in text data likely contributed to its lower performance compared to other models in this project.

Conclusion

This project successfully leveraged machine learning techniques to predict patient conditions from drug reviews, providing a practical tool in the realm of personalized medicine. By employing a TF-IDF vectorizer and various predictive models, the project identified the SVM model as the most effective, due to its superior accuracy in classifying patient conditions. Furthermore, the project extends its utility by offering recommendations for five popular drugs, enhancing the decision-making process for healthcare providers. While these recommendations are based on general trends and individual responses to drugs may vary, they still offer valuable insights.

This approach not only aids in understanding patient experiences but also supports healthcare professionals in tailoring treatments to individual needs, thereby improving therapeutic outcomes. The methodologies and findings of this project could significantly influence future strategies in patient care and drug recommendation, marking a substantial advancement in healthcare services.

References

1] Sentiment Analysis in Drug Reviews using Machine Learning and Deep Learning Techniques**.** Link

2] Exploring Drug Sentiment Analysis with Machine Learning Techniques. Link

3] Vimala Balakrishnan and Ethel Lloyd-Yemoh. Stemming and lemmatization: A comparison of retrieval performances. IACSIT, 2014. Link

4] A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. Link

5] Drug review sentimental analysis based on modular lexicon generation and a fusion of bidirectional threshold weighted mapping CNN-RNN. Link

6] Sentiment Analysis of User-Generated Content on Drug Review Websites. Link

7] Sentiment Classification of Drug Reviews Using Machine Learning Techniques.Link

8] Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. Link

9] Use of Sentiment Analysis for Capturing Patient Experience from Free-Text Comments Posted Online Link