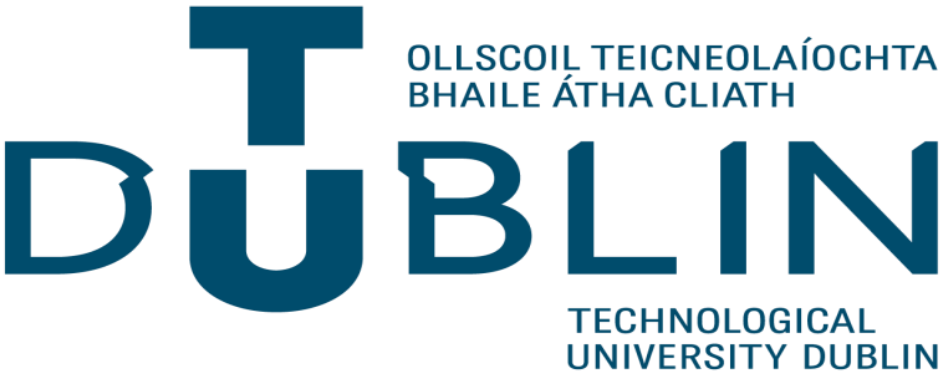


DATA MINING – ASSIGNMENT 2

ANALYSIS AND MODELLING OF PHARMACEUTICAL DATA FOR OPINION MINING AND DRUG CHARACTERISTICS

Student Name	Vyoma Mohan
Student Id	D22124454
Subject Code	DATA 9900
Subject	Data Mining
Course Id	TU059
Name of the assignment	Analysing pharmaceutical data for opinion mining and drug characteristics
Deadline	8th Jan 2023
Stream	Data Science (DS)
Year	First Year

Text-mining Assignment Submission Cover Sheet

This Assessment Cover Sheet **must** be included on all Assessment submissions.

Assignment Title	Analysis and Modelling of Pharmaceutical Data for Opinion Mining and Drug Characteristics
Module	DATA 9900
Student Name (same as Student Card)	Vyoma Mohan
Student Number	D22124454
Programme	TU059
Part-Time/Full-Time	Full-time
Year of Study (First Year, Second Year, etc)	First year

Late Submissions: Assessment submitted after the deadline will have a late penalty applied.

Academic Integrity for assessment in TU Dublin Programmes

Each student is responsible for knowing and abiding by TU Dublin Academic Regulations and Policies. Any student in breach of these regulation/policies will be subject to action in accordance with the University's procedures for breaches of assessment regulations. Please refer to the General Assessment Regulations at
https://tudublin.libguides.com/c.php?g=674049&p=4794713
https://www.tudublinsu.ie/advice/exams/breachesofregulations/

All students are expected to complete their courses/programmes in compliance with University regulations. No student shall engage in any activity that involves attempting to receive a grade by means other than honest effort, for example:

1. No student shall complete, in part or in total, any examination or assessment for another person.
2. No student shall knowingly allow any examination or assessment to be completed, in part or in total, for themselves by another person.
3. No student shall plagiarise or copy the work of another and submit it as their own work.
4. No student shall falsify any data. Falsification is the invention of data, its alteration, its copying from any other source, or otherwise obtaining it by unfair means, or inventing quotations and/or references.
5. No student shall use aids or devices excluded by the lecturer in undertaking course work or assessments/ examinations.
6. No student shall knowingly procure, provide, or accept any materials that contain questions or answers to any examination or assessment to be given at a subsequent time.
7. No student shall provide their assignments, in part or in total, to any other student in current or future classes of this module/ programme unless authorised to do so by the lecturer.
8. No student shall submit substantially the same material in more than one module/programme without prior authorization.
9. No student shall alter graded assignments or examinations and then resubmit them for regrading, unless specifically authorised to do so by the lecturer.
10. All programming code and documentation, unless correctly referenced, submitted for assessment or existing in the student's computer accounts must be the students' original work or material specifically authorized by the lecturer.
11. Collaborating with other students to develop, complete or correct course work is limited to activities explicitly authorized by the lecturer.
12. For all group assignments, each member of the group is responsible for the academic integrity of the entire submission. Consequently, all group members must satisfy themselves that all elements of their submission adhere to the academic integrity statement points above.

By submitting coursework, either physically or electronically, you are confirming that it is your own work (or, in the case of a group submission, that it is the result of joint work undertaken by members of the group that you represent) and that you have read and understand the University's Regulations and Policies covering Academic Integrity (see General Assessment Regulations).

Coursework may be submitted to an electronic detection system in order to help ascertain if any plagiarised material is present. If you have queries about what constitutes plagiarism, please speak to your lecturer.

Student Signature	Vyoma Mohan
Date	07-01-2023

SECTION 1: INTRODUCTION AND PROBLEM DEFINITION

Purpose of Analysis:

In this scenario, we are tasked by a pharmaceutical company to derive insights from the given set of data which contains information about customer reviews about drugs produced by the company. The dataset includes information about the overall rating, the benefits and the side effects of the drugs as described by the consumer along with other data.

Our goal is to derive insights about what sort of drugs are preferred by the customers. We also need to find out which are the indicators that influence customer experience about the given drugs and if there are any rules that will help us determine if a drug will be received well. In an ideal scenario, we would be able to settle for a drug for a problem where the drug is effective and has a good experience while using it.

Source of dataset:

Link to dataset: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29

The sources for the dataset are stated as Surya Kallumadi and Felix Gräßer. They have used it in their paper "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning".

The researchers obtained the data by scraping drug.com and drugslib.com. These two websites had their data structured different so, they tried to match up whichever characteristics they could and ended up with the current dataset. The scraping was done using beautiful soup *(Ref - [1] (Gräßer, Kallumadi, Malberg, & Zaunseder, 2018))* . For analysis, we retrieve the dataset from the UCI database.

Type of mining tasks performed:

- **Classification tasks:** Classification tasks are done in order to classify a drug into a "Positive" or "Negative" category based on the characteristics of the drug.
- **Association rules:** Used in order to understand whether there is some relation between the variables and the outcome for Sentiment based on how frequently they occur together.
- Analysing sentiment by condition.

Assumptions before starting:

- All the drugs in the dataset are produced by the same company (this is for the given scenario as stated in the assignment requirements. In actuality, it is scraped from a site with user reviews for different drugs).
- The categories for effectiveness are: Highly effective, Considerable effective, Moderately effective, Marginally effective and Ineffective. This is stated from the best case (most effective) to the worst case (least effective).
- The side effects are categorized into: Extremely severe, Severe, Moderate, Mild, No side effects. This is stated from the least desirable to the most desirable.
- The words "medicine" and "drug" are used interchangeably in this report.
- At times the words customer, consumer and patient may be used interchangeably in the report.

Note for analysis:

Most of the analysis is done on the training data.

About the variables:

The training dataset and the test dataset have the same set of variables. These variables are described below:

Variable Name	Data Type	Variable Type	Discrete/ Continuous	Range/Possible values
DrugName	Text (String)	Nominal	Discrete	N/A
Rating	Numeric	Interval	Discrete	1-10
Effectiveness	Text (String)	Ordinal	Discrete	Highly Effective, Considerably Effective, Moderately Effective, Marginally Effective, Ineffective
Side effects	Text (String)	Ordinal	Discrete	No side effects, Mild side effects, Moderate side effects, Severe side effects, Extremely severe side effects
Condition	Text (String)	Nominal	Discrete	N/A
BenefitsReview	Text (String)	Nominal	Discrete	N/A
SideEffectsReview	Text (String)	Nominal	Discrete	N/A
CommentsReview	Text (String)	Nominal	Discrete	N/A

The descriptions of the variables are given below:

DrugName – The name of the drug for which the review is about.

Rating – A rating for the drug provided by the consumer on a scale of 1-10.

Effectiveness – States how well the drug works for the given problem. The definition of effectiveness is given as: *"In medicine, the ability of an intervention (for example, a drug or surgery) to produce the desired beneficial effect." (Ref – [16])*

Side Effects – States how severe the side effects for the condition are.

Condition – The disease or illness that the patient is taking the consumer in order to cure.

BenefitsReview – Addresses the pros of taking the medicine.

SideEffectsReview – Addresses what side effects are experienced when taking the medicine.

CommentsReview – The overall review from the consumer.

Section summary:

Thus, the problem statement is defined and the dataset is described in detail.

SECTION 2: DATA CLEANING

Assessing NA counts:

For the first step of cleaning, we will check the NA counts for each column. This is just to get an initial estimate. We can remove these rows later during the machine learning phase.

```
Unnamed: 0           0
urlDrugName          0
rating               0
effectiveness        0
sideEffects          0
condition            1
benefitsReview       0
sideEffectsReview    2
commentsReview       8
dtype: int64
```

Fig: NA counts of all the columns.

Non-Alphanumeric comments:

Some of the comments are just punctuations. They are shown below:

	Unnamed: 0	urlDrugName	rating	effectiveness	sideEffects	condition	benefitsReview	sideEffectsReview	commentsReview
249	1843	yasmin	10	Highly Effective	No Side Effects	birth control	I've been on yasmin four years now, it works s...	None	.
910	3239	nuvaring	9	Highly Effective	Mild Side Effects	prevent pregnancy	had a more regular period, knew that i would b...	i feel as if i get headaches more, and pissed ...	...
1385	2995	sulfasalazine	10	Highly Effective	No Side Effects	psoriatic arthritis	Significantly reduced pain and swelling in low...	Slight yellowing of fluids and in whites of eyes.	.
1408	3024	cipro	10	Highly Effective	Mild Side Effects	rare kidney infection	My daughter is playing now finally and she see...	Blistering rash	.
2271	781	naproxen	1	Ineffective	Extremely Severe Side Effects	carpal tunnel	releved pain for 1 hour but then has no more e...	gave me extreme cramps, swelling, dizzyness, n...	????
2500	1119	effexor	6	Moderately Effective	Severe Side Effects	depression	I noticed significant changes in the beginning...	Horrific nausea and feelings of body shock whe...	-
3104	1664	climara	2	Marginally Effective	Moderate Side Effects	total hysterctomy	---	Constant issues with the patch not staying on....	---

Fig: Comments with no words and only punctuations. See "CommentsReview" column.

There is a tiny portion of such rows in the data. These rows are removed. We check the rows: BenefitsReview, SideEffectsReview, CommentsReview for such entries and remove them.

Renaming columns:

The columns are renamed from pascal case and the starting letter is capitalized.

Drop the column that is not described:

There is a column called "Unnamed". It contains numerical data. We are not sure what this data is as it also not described in the source. So, we remove this column from the data so that it does not interfere with the analysis.

Attribute Information:

1. urlDrugName (categorical): name of drug
2. condition (categorical): name of condition
3. benefitsReview (text): patient on benefits
4. sideEffectsReview (text): patient on side effects
5. commentsReview (text): overall patient comment
6. rating (numerical): 10 star patient rating
7. sideEffects (categorical): 5 step side effect rating
8. effectiveness (categorical): 5 step effectiveness rating

Fig: Column descriptions from the UCI website.

Additional Notes:

- There are some conditions described that sound very similar, but it is not confirmed whether it is a typo or intentional. For example, there are some rows like “acid reflex” and “acid reflux”. The words reflex and reflux mean different things, so it is not confirmed whether they are different conditions or not. Thus, it is left unchanged in the data.

TRANSFORMATION

Adding a column to categorize the sentiment based on rating:

A column can be added to the dataset to distinguish the positive-negative reviews instead of using the 10-point rating scale. This has less categories, making classification easier. For this analysis we will consider 1-5 as a negative review and 6-10 as a positive review.

	DrugName	Rating	Effectiveness	SideEffects	Condition	BenefitsReview	SideEffectsReview	CommentsReview	Sentiment
0	enalapril	4	Highly Effective	Mild Side Effects	management of congestive heart failure	slowed the progression of left ventricular dys...	cough, hypotension , proteinuria, impotence , ...	monitor blood pressure , weight and asses for ...	Negative
1	ortho-tri-cyclen	1	Highly Effective	Severe Side Effects	birth prevention	Although this type of birth control has more c...	Heavy Cycle, Cramps, Hot Flashes, Fatigue, Lon...	I Hate This Birth Control, I Would Not Suggest...	Negative
2	ponstel	10	Highly Effective	No Side Effects	menstrual cramps	I was used to having cramps so badly that they...	Heavier bleeding and clotting than normal.	I took 2 pills at the onset of my menstrual cr...	Positive
3	prilosec	3	Marginally Effective	Mild Side Effects	acid reflux	The acid reflux went away for a few months aft...	Constipation, dry mouth and some mild dizzines...	I was given Prilosec prescription at a dose of...	Negative
4	lyrica	2	Marginally Effective	Severe Side Effects	fibromyalgia	I think that the Lyrica was starting to help w...	I felt extremely drugged and dopey. Could not...	See above	Negative

Fig: Dataset with Sentiment column added

Section summary:

Thus, the dataset is cleaned and the “Sentiment” column is included in order to do machine learning. This cleaned dataset is put into a csv file in order to perform ML tasks on it. The same procedure for cleaning and storing is followed for both the training data, as well as the test data.

SECTION 3 : DATA EXPLORATION

For the initial data exploration state, we will try to visualize some basic facts about the dataset at hand.

Number of drugs produced by the company:

Process to get visual: Here, we try to find how many drugs are produced by the company that are mentioned in the reviews. In order to get this number, we use 'nunique' in python which returns the number of unique values in a column (Not including null). This is run on the 'DrugName' column.

Result: It is found that 502 unique drugs are produced by the company.

Number of conditions covered by the company's drugs:

Process to get visual: The goal is to find the number of conditions that drugs that company produces are meant to treat. In order to get this number, we use 'nunique' in python which return the number of unique values in a column (Not including null). This is run on the 'Condition' column.

Result: It is found that 1422 conditions are addressed by the drugs produced by the company.

Distribution of Effectiveness of drugs in the reviews:

Process to get visual: Our goal is to find what is the proportion for the Effectiveness of the drugs produced by the company (according to the reviews. The value for some drugs may be repeated due to multiple reviews). In order to get this, we need the counts of each category of Effectiveness. This is done by using 'value_counts' in python. This data is then used to produce a Pie-chart.

```
Highly Effective          1326
Considerably Effective     927
Moderately Effective       414
Ineffective                246
Marginally Effective       186
Name: Effectiveness, dtype: int64
```

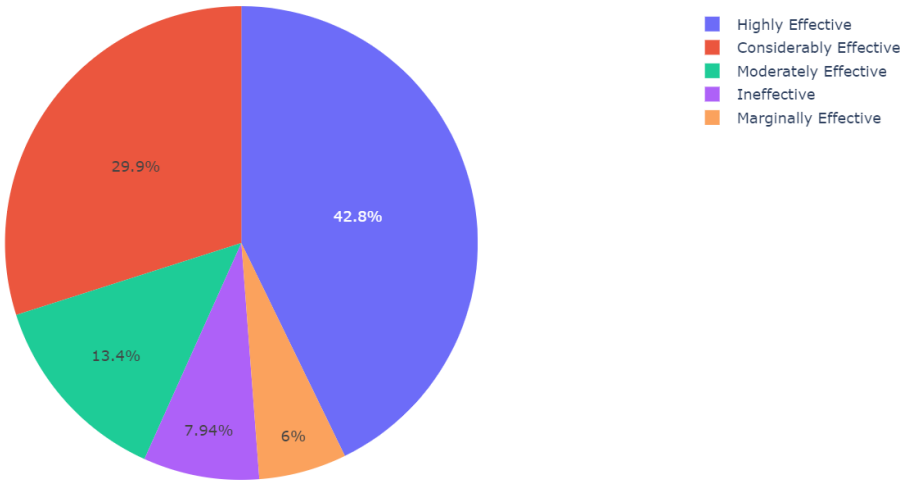


Fig: Pie-chart showing distribution of effectiveness of drugs in reviews

Interpretation: It is good to note that a majority of the drugs fall under the “Highly Effective” category (42.8%). The least number of drugs fall under the “Marginally Effective” category (6%).

Distribution of Side Effects according to the reviews:

Process to get visual: The goal is to get the distribution of side effect severity in the reviews (Value may be repeated when there is more than 1 review for a drug). In order to get the visual, we need the count in each category of severity. This is done by using ‘value_counts’ in python. The data then used to produce a Pie-chart.

```
Mild Side Effects              1017
No Side Effects                 928
Moderate Side Effects           612
Severe Side Effects             368
Extremely Severe Side Effects   174
Name: SideEffects, dtype: int64
```

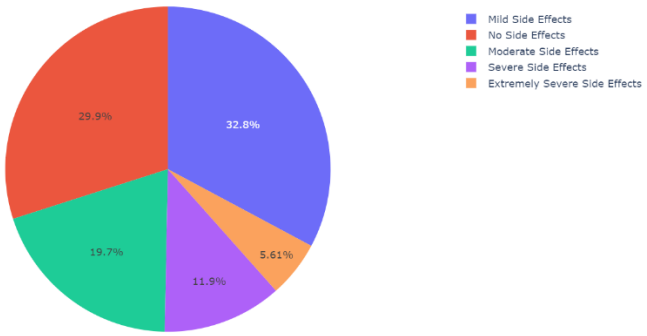


Fig: Pie-chart showing distribution of side-effects in reviews

Interpretation: A majority of the values fall under “Mild Side Effects” (32.8%). This is closely followed by “No Side Effects” (29.9%). The least number of values lie in “Extremely Severe Side Effects”.

Top 10 conditions based on number of reviews:

Process to get visual:

In this part, we attempt to study which conditions are commonly present among the reviews. This could be useful to study the skew of data. If a condition is repeated often, there is a good chance that reviews will be related to the same set of drugs. Performing a count using the drug names might address the point better, but there are too many drugs (502) to get proper categories. Thus, is better to visualize with conditions.

First, we take the counts of all the conditions for study. This is obtained by grouping by ‘Condition’ and then taking a count.

```
display(condi_count.sort_values('Count',ascending = False).head(10))
```

```
1422
```

Condition	Count
depression	235
acne	165
anxiety	63
insomnia	54
birth control	48
high blood pressure	42
allergies	37
asthma	33
acid reflux	33
migraines	31

Fig: Counts of conditions in the dataset

This list is order from conditions with the largest number of reviews to the lowest number of reviews. The first 10 rows are shown above. We can see that in total there are 1422 conditions. This becomes very tough to visualize. So, in order to get a neater visual, we will visualize only the top 10 conditions. This is shown in a bar chart as:

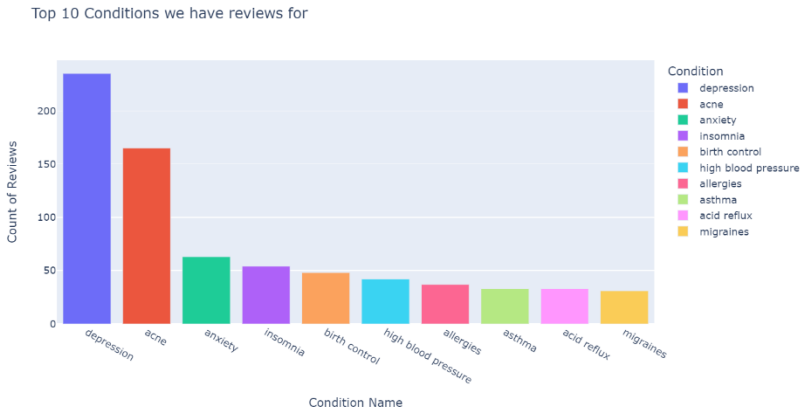


Fig: Bar-plot showing the counts of top-10 conditions

Interpretation: Thus, we can see that depression has the highest number of reviews (235). We could expect depression related meds to have more impact on the results when compared to other conditions. This is simply because there are a significantly higher number of reviews in the set for them.

Section summary:

Thus, exploratory data analysis is done to get some basic facts about the data and study the distribution of the columns.

SECTION 4: DATA ANALYSIS

Q1: Which drugs are the best performing?

Why will this be useful? : If we know which the best performing drugs are, we can study their characteristics and then find ways to improve similar drugs.

How the visual is obtained: The first thing we do is to find the average rating of all the drugs. To do this, we group the data by the drug names and take the mean. This will give us the average ratings for all the drugs in the dataframe. We see that some of the drugs have an average rating of 10 (maximum rating). Thus to get the best performing drug, we will filter those drugs which have an average rating of 10.

Then we take a count of how many drugs have an average rating of 10. It is found that 50 drugs have the maximum average rating of 10. These drugs are displayed in a table with the help of plotly as shown below.

Visual:

Names of medicines with an average rating of 10

Medicine Name	Medicine Name	Medicine Name	Medicine Name	Medicine Name
androgel	bystolic	exelon	methimazole	prograf
antivert	cataflam	flovent	naltrexone	qvar
asmanex	delestrogen	follistim	nasacort-aq	ribavirin
atripla	dextrostat	fosamax-plus-d	nasarel	sanctura-xr
avapro	elocon	geodon	neoprofen	suboxone
axert	erythra-derm	haldol	phendimetrazine	tekturna
azopt	estrasorb	hydrocortisone	polymyxin-b	trimethoprim
baraclude	estratest	hytrin	ponstel	triphasil
bisoprolol	estrostep-fe	lidocaine	pravachol	vigamox
buprenorphine	eulexin	metformin-extended-releas	progesterone	zestoretic

Fig: Names of the drugs with a 10-rating

Q2: Find the best rating for a drug for the conditions and find the distribution of the reviews.

Why will this be useful? : In this part, we will try to get the highest rating for a condition. This will help us understand which conditions we need to focus on while developing new drugs. (For example, if the best rating for a medicine is 1, more work is needed to develop medicines for that particular condition as there is no suitable alternative). Note that, for this process, the best rating can be influenced by a single review. If there are multiple reviews, all having the same maximum value, they are all considered.

How the visual is obtained: First, we need to find the best rated drugs for all the conditions. In order to do this, we first group by the condition and then do a maximum transform on it. We will get the indexes for the best rated medicines, and we will get them from the all reviews dataframe.

In order to get the distributions for the best ratings of a drug, we will take a count for the rating category each best review belongs to. This is expressed in a graph as:

Visual:

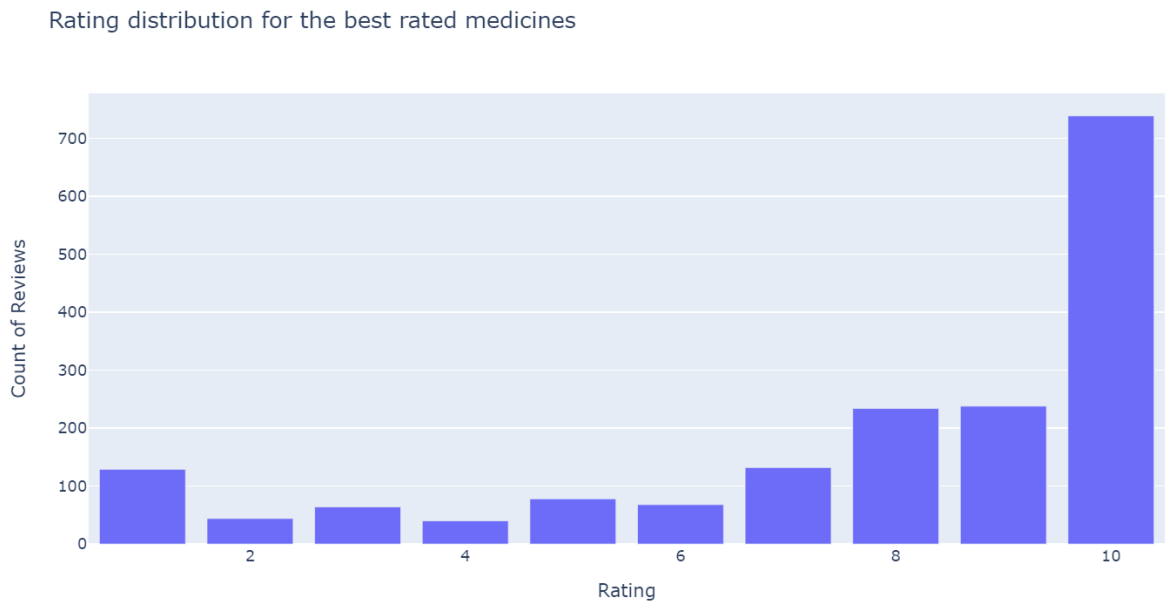


Fig: Rating distribution for the best rated medicines for a condition

From the graph, we can see that there is at least one 10-rating review for most of the conditions. The 4-rating has the lowest number of reviews.

We will expand on this analysis in the sections that follow.

Q3. Study the characteristics of the best-rated medicines for the top 10 conditions.

Why will this be useful? : In this section, we will be studying the side effects and effectiveness of the best rated medicines for the top 10 conditions. This will help gain an initial insight into whether the ratings are influenced by side effects or effectiveness. We will be able to study which sort of medicine appeals to the people.

This study can be done for all conditions and not just the top 10 (overall analysis follows this section). However, in this section, we attempt to put the results into a visual since this part is only going to help us form an initial impression. If there is a lot of data (such as data for all the conditions) the visuals end up looking crowded with no clarity. Thus, we go for visualizing only for the top conditions here.

How the visual is obtained :

Studying the Side effects of the best medicines for the top 10 conditions:

First, we filter and obtain the best medicine reviews only for the top 10 conditions from the previous step. Then we make a bubble plot using the data.

Here, we plot the name of the medicine against the side effect severity. The rating determines the size of the bubble in the bubble plot. The bubbles are coloured according to the condition the medicine is meant to address. Doing this produces the following visual:

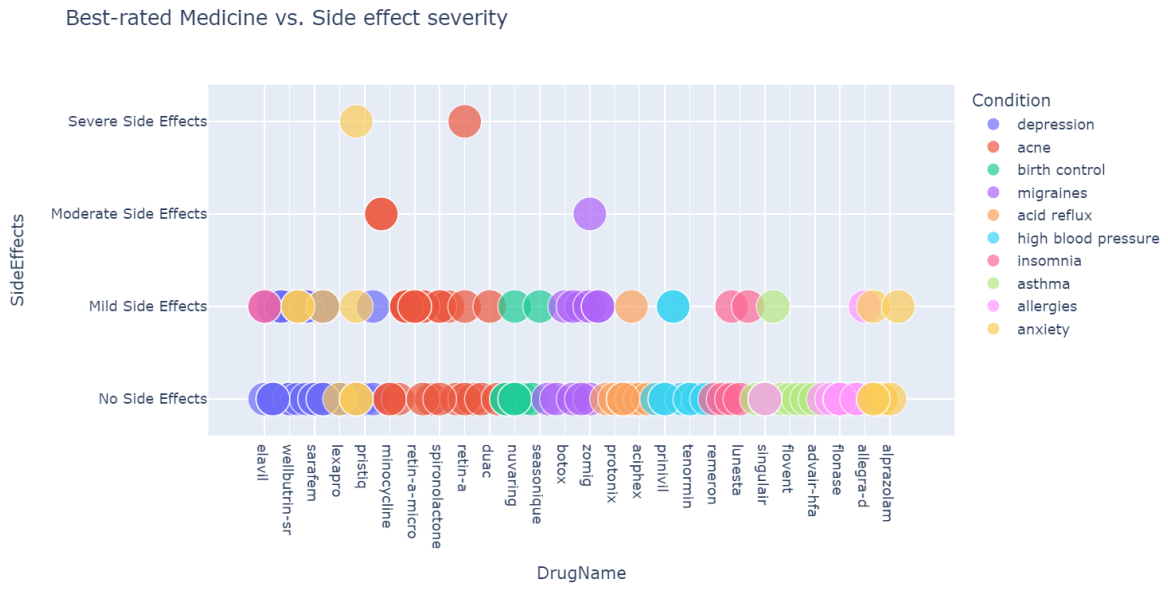


Fig: Checking the side effects for the best rated medicines

Interpretation:

Here, we can see that most of the medicines in this category have no side effects or mild side effects. There are a couple medicines for acne and anxiety however, that have a high rating even though they have severe side effects. There are no reviews that pertain to medicines with extremely severe side effects in this category.

The size of the bubbles is determined by the rating. Here, all the bubbles are of the same size (10). So, all the best medicines for the top 10 conditions have a 10 rating.

Studying the effectiveness of the best medicines for the top 10 conditions:

Here, again we use the filtered list which contains the best medicine reviews for the top 10 conditions. We plot the medicine names against the effectiveness. The condition is denoted by the colour of the bubbles. The size is represented by the rating score. Doing the above, we get the visual:

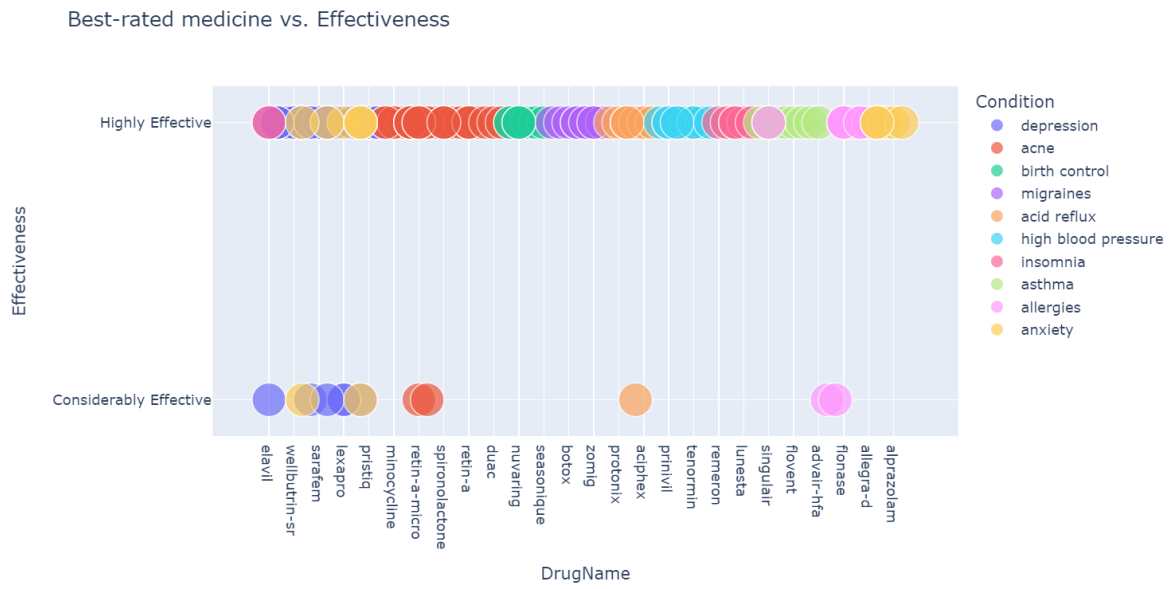


Fig: Studying the effectiveness of the best rated medicines

Interpretation:

Here, we can see that most of the best rated medicines fall into the highly effective category. There are also some entries for Considerably effective. Again, the size of the bubble represents rating. Here, the bubbles are of the same size (10-rating). The entries for the other 3 categories are not present.

Q4. Study people's opinion about medicine which is highly effective but has extremely severe side effects

Why will this be useful? : It will be interesting to study people's opinion on medicines which have extremely severe side effects but is highly effective i.e, when the drug has the intended effect but causes other problems. The goal here is to see if people think that such medicines are worth the trouble.

How the visual is obtained: First, we filter out all the reviews with 'highly effective' in the effectiveness column and also 'extremely severe side effects' in the side effects column. Then we will take a count of each rating category and put it in a bar chart. The image is as follows:

Visual:

The visual which depicts the rating distribution of medicines which are highly effective but have extremely severe side effects is expressed as a bar chart with review count as the y-axis and the rating as the x-axis as given below:

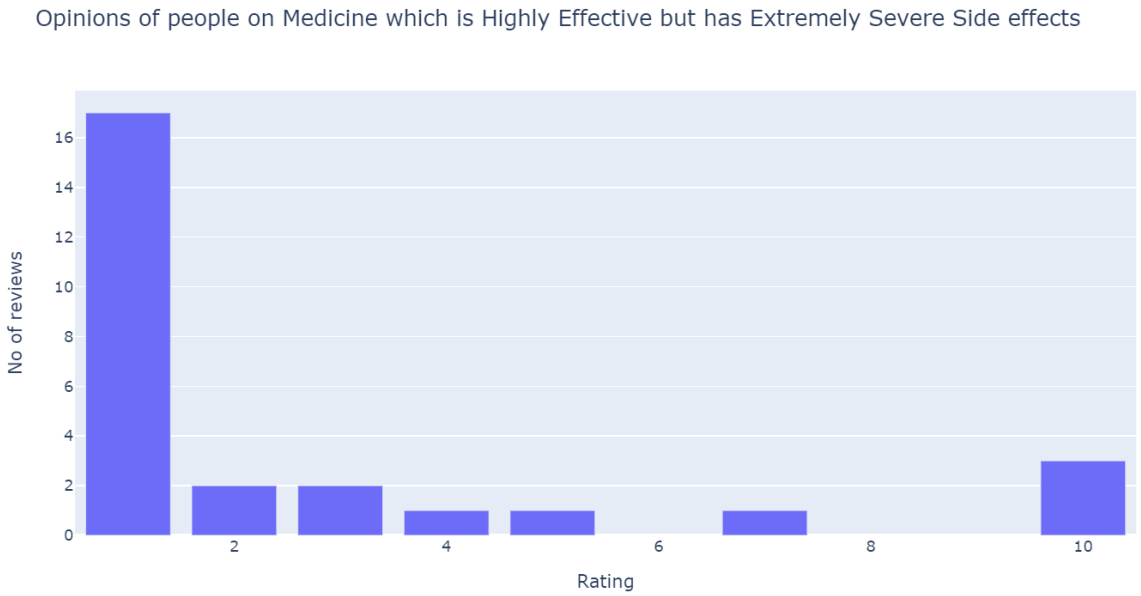


Fig: Rating distribution for Highly Effective medicines which have Extremely Severe side effects

Interpretation:

Here we can see that even though the medicine is highly effective, there is a clear bias to rating-1 when considering medicines which are highly effective but have severe side effects. Thus, the overall opinion of people towards such medicines seems to be negative. The minimum rating given is 1, but some medicines do have a 10 rating in this set. So, we can say that the rating range is from 1-10.

Q5. Study people's opinion about medicine which is ineffective and also has extremely high severe side effects

Why will this be useful? : There may be some drugs which are ineffective and also have extremely severe side effects. It is important to find alternatives for such drugs if there aren't any. We will study people's opinion on such drugs to verify if they are expressing discontent. We will also analyse the alternates for the same condition and compare the opinions of people.

How the visual is obtained: First, we filter out the reviews from the set which say that the medicine is ineffective and also has extremely severe side effects. Then we will make a count of the reviews under each rating. The result is displayed as a bar graph as follows:

Visual:

The visual which depicts the rating distribution of medicines which are ineffective but have extremely severe side effects is expressed as a bar chart with review count as the y-axis and the rating as the x-axis as given below:

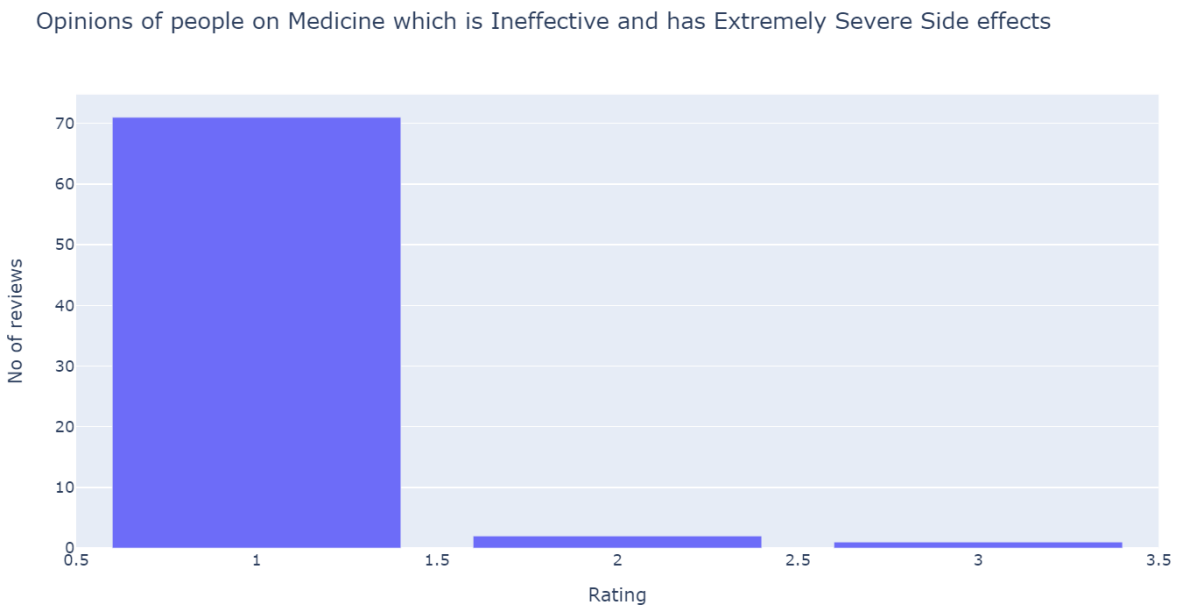


Fig: Rating distribution for Ineffective medicines which have Extremely Severe side effects

Interpretation:

As expected, the ratings for such medicines are low. The range for ratings is 1-3, with rating 1 having the majority of the count.

Studying the alternatives of such medicines:

While assessing such medicines, it is also prudent to take note of other medicines that treat the same condition. Do they have better ratings than the above medicines?

How the visual for alternates is obtained: We take the set of conditions that the above medicines (ineffective-ext.severe side effects) treat. Then we get all the reviews for those conditions. After getting them, we filter out the reviews that deal with medicines that are ineffective with extremely severe side effects. This is the data for the alternatives. It is also plotted in a bar graph as:

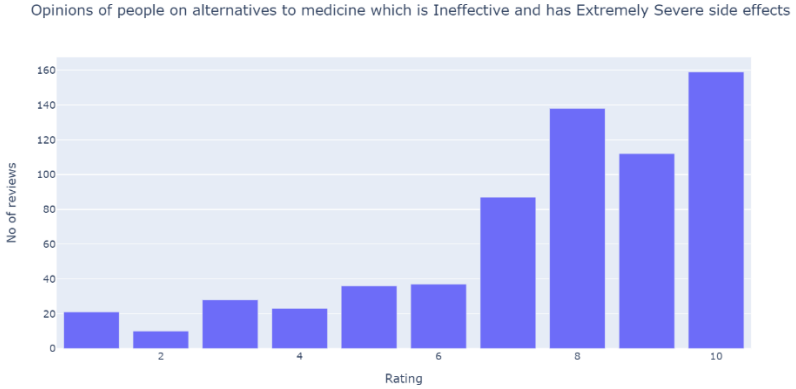


Fig: Rating distribution for medicines which serve as an alternate to the conditions for medicines which are ineffective but have extremely severe side effects.

Interpretation of above:

Here, we can see that a rating-10 has the majority of values and there is a much wider range with reviews ranging from rating 1 to rating 10. Thus, the people seem to have a more positive opinion about the alternates.

(However, it is more important to note the proportion of numbers in each category rather than the actual number since the total number of reviews for the alternates is more than the total number of reviews for the medicines in the ineffective-extremely severe side effect category)

Q6: Constructing word clouds on the reviews in order to find the most common words in the reviews.

Why is this important? : Word clouds give us a bird's eye view into the corpus related to a topic. In this case, it is the reviews of the drugs. The size of the words in the cloud are related to how often the words occur in the corpus. Using this, we can get the most commonly used words in the reviews.

How the visual is obtained: We essential pick the columns we need to make word clouds out of and feed them to the word cloud widget. The preprocess text removes most stopwords and numbers. The visual is obtained from orange by using the following widget configuration:

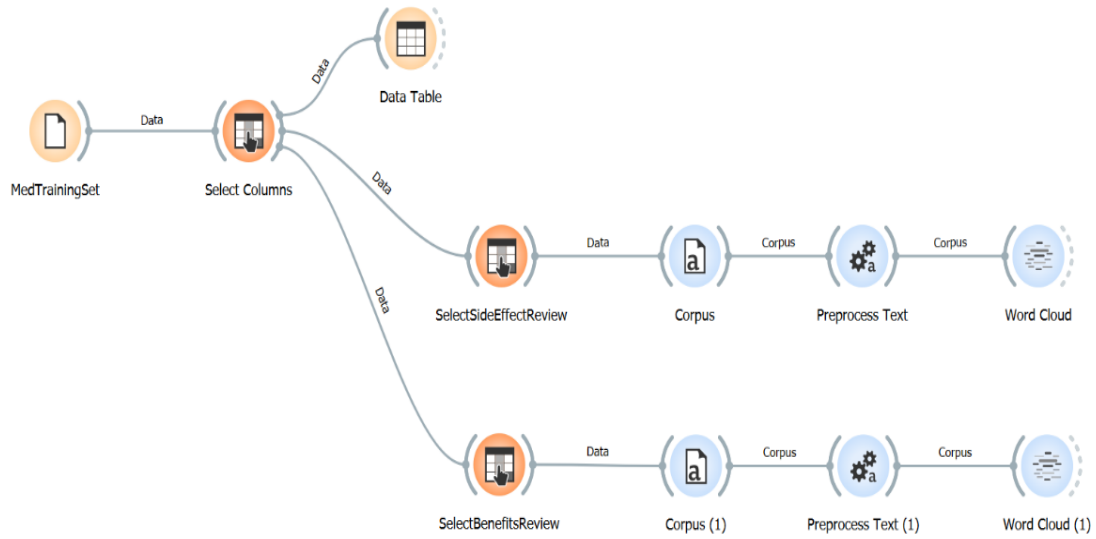


Fig: Orange widgets to generate word clouds

We generate word clouds for the review aspects for side effects and benefits. These word clouds are given below:

Word cloud for Side effects reviews:

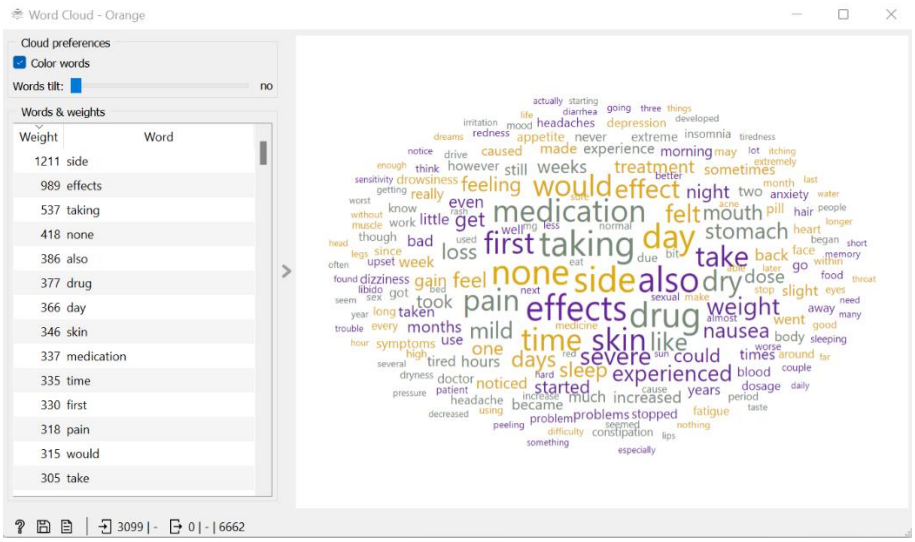


Fig: Word cloud for SideEffectsReview column

Word cloud for Benefits reviews:

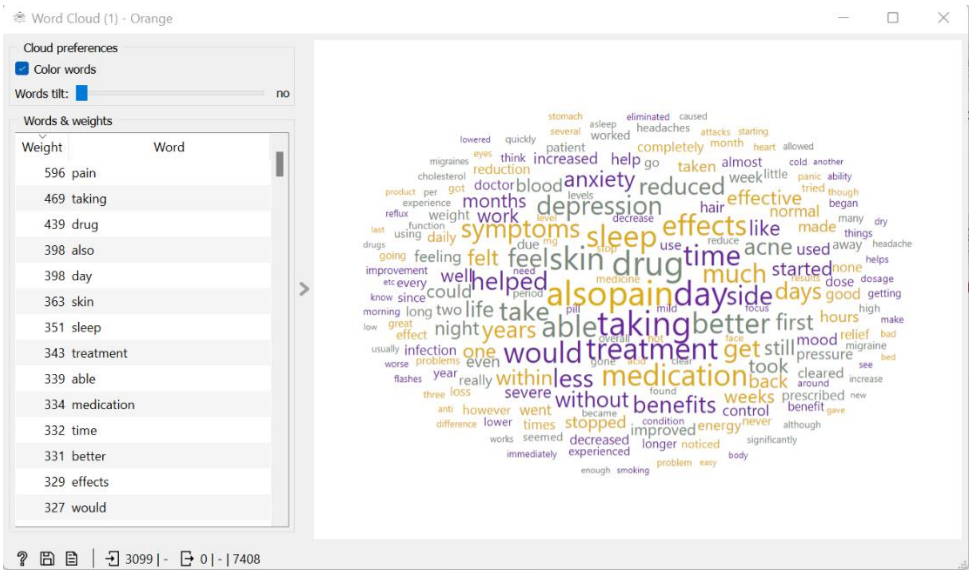


Fig: Word cloud for Benefits column

Q7: An initial sentiment analysis on the benefits review rows

Why is this useful? : We can do an initial sentiment analysis on a review column of the data. This is just to understand the general tone of the reviews of the column. It does not have any overarching value for the whole column. This can be useful to perform an initial check to see how VADER is performing on the reviews in that column before going for the ML aspect.

How the visual is obtained:

The VADER opinion mining tool is run on the benefits review column to get the sentiment of the column. A heat-map can be generated based on the corpus returned by the sentiment analysis widget. The orange configuration is given as:

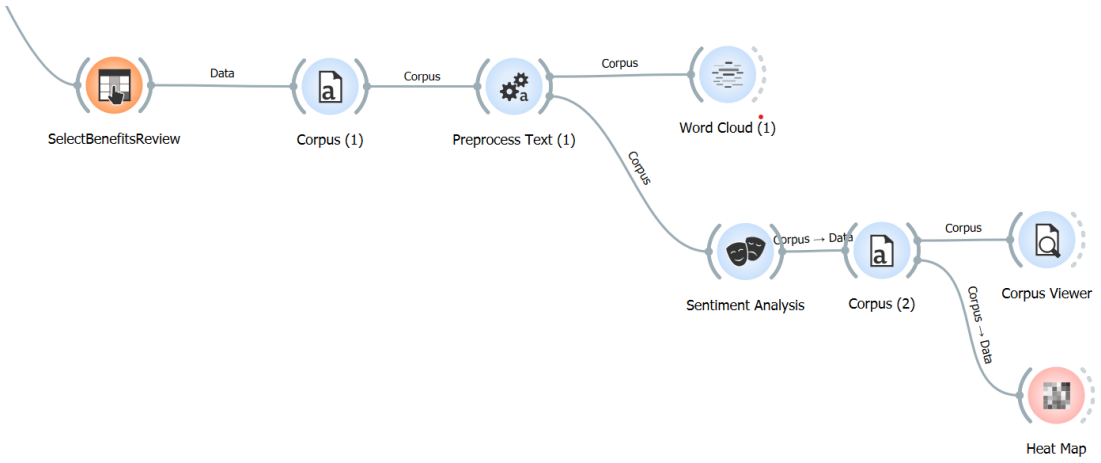


Fig: Orange widget configuration to check VADER

From this, we can study the specifics in the corpus output and get a general picture in the heatmap widget.

Screenshots and interpretation:

The output from corpus viewer is shown as:

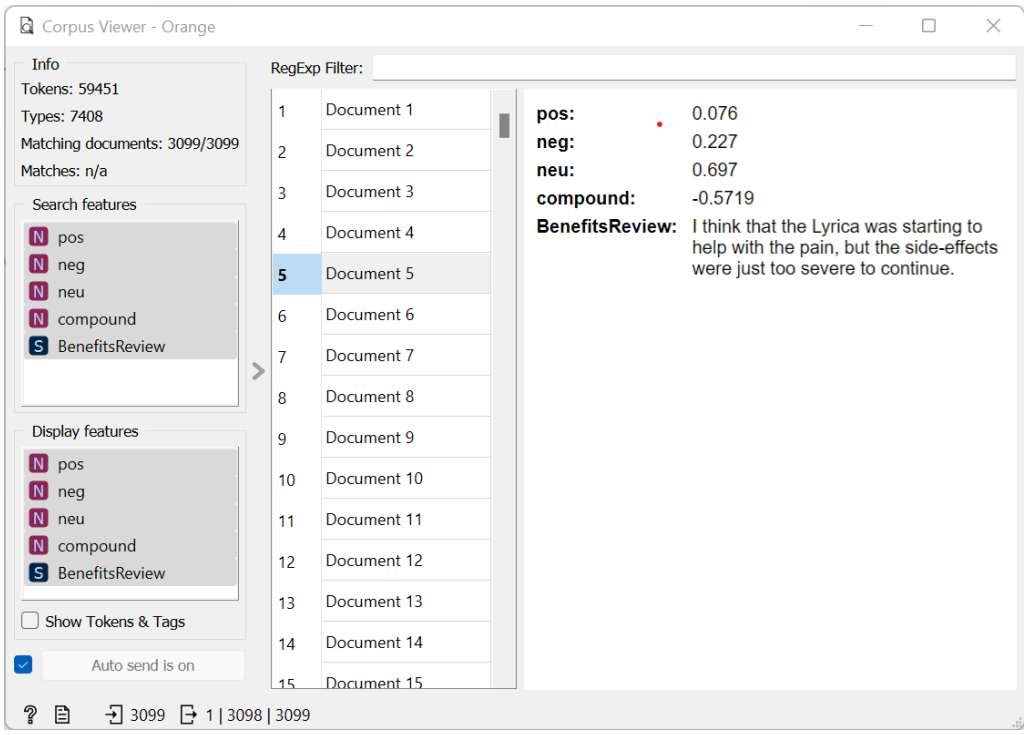


Fig: VADER output with pos, neg, neu and compound values for the BenefitsReview

Here, we can see the specific review and the sentiment values associated with it. “Pos” field shows the chance of the review being positive. “Neg” shows the chance of the review being negative. “Neu” shows the chance of the review being neutral. “Compound” is the combined sentiment that is being expressed in the statement.

So, for instance, in the above, the overall sentiment in the given review according to VADER is negative (-0.57). On reading the review, it looks accurate.

The heatmap version is shown as such:

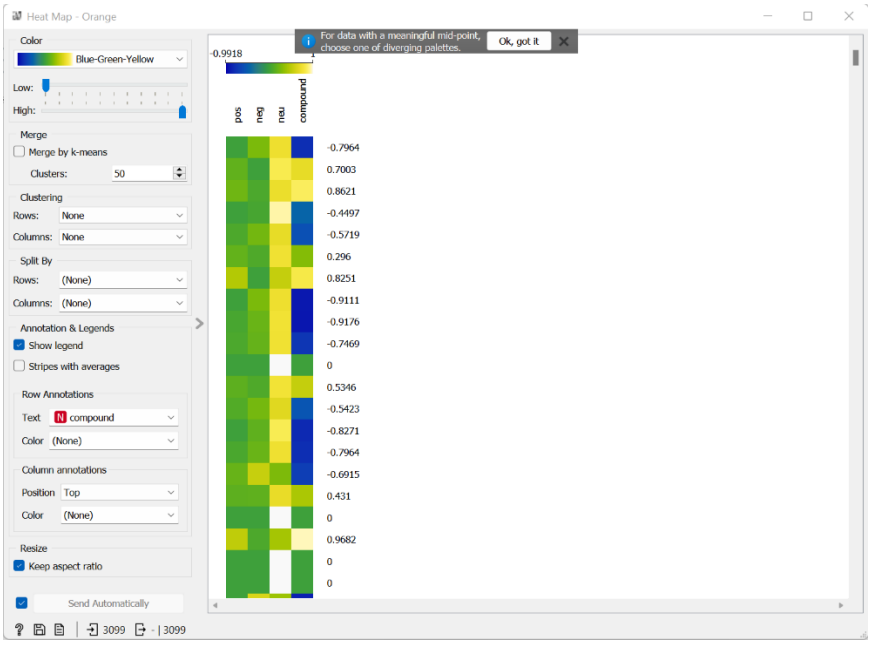


Fig: Heatmap version of VADER values. Yellow tends towards positive, Blue is negative, Green is the middle ground between the two

Here, a blue colour means negative, whereas a yellow means a positive sentiment. This heatmap can be used to quickly looks at all the results produced from the sentiment analysis.

Thus, a few reviews are checked, and VADER seems to be preforming in a satisfactory manner.

Section Summary:

Thus, the descriptive analysis is carried out for the data. We have created visuals for some subsets of the data in order to understand them better. We have checked the performance of VADER on the reviews.

SECTION 5: CORRELATION ANALYSIS

Usually before ML is carried out on a set of data, we do some analysis to verify if some of the fields are correlated in order to justify using them for ML. Therefore, we will check some combinations of fields to check whether there is a correlation between the fields.

A major portion of the ML section will be related to classifying the data according to whether the sentiment in the data is positive or negative. So, we will do correlation analysis on the Rating column with other columns in the data.

For correlation analysis, we will still retain the 10 point rating scale. And perform correlation analysis on it.

Correlation between the rating and the effectiveness:

Variables compared:

Variable	Independent/Dependent	Type	Continuous/Discrete
Effectiveness	Independent	Ordinal	Discrete
Rating	Dependent	Ordinal	Discrete

Between two non-numeric variables, it is best to use the chi-square test and then Cramer V to analyse the effect.

The contingency table is generated as follows:

Rating / Effectiveness	1	2	3	4	5	6	7	8	9	10
Considerably Effective	33	15	29	20	42	40	171	334	186	57
Highly Effective	32	11	12	12	28	21	66	182	285	677
Ineffective	168	39	27	5	5	2	0	0	0	0
Marginally Effective	34	20	53	33	23	15	3	4	0	1
Moderately Effective	37	17	25	37	61	78	109	38	8	4

Fig: Contingency table for Effectiveness vs. Rating

This is fed to the chi-square function to get the following results:

```
stat, p, dof, expected = chi2_contingency(rat_effect_cont)
print(stat)
print(p)
print(dof)
# Since more than 2 categories for the variable, cannot do fisher exact
```

```
3326.0024692529396
0.0
36
```

Fig: Chi-square test results showing chi-sq value, p-value and degrees of freedom

Since there are more than 2 categories for one of the variables, we cannot use the Fisher exact test defined by python. So, we will skip that test.

From the above p value, the finding is seen to be statistically significant. So, we can proceed with finding the effect. Finding the Cramer's V for these variables is given as:

```
#There are 36 degrees of freedom
#For them, 0.5 shows a strong effect for cramer's v
```

```
3099
4
0.517988955502901
```

Fig: Cramer V test results

From this, we can see there is a strong effect (0.5 Cramer V for 36 degrees of freedom).

Correlation between rating and side effects:

Variables compared:

Variable	Independent/Dependent	Type of Variable	Continuous/Discrete
Side effects	Independent	Ordinal	Discrete
Rating	Dependent	Ordinal	Discrete

Between two non-numeric variables, it is best to use the chi-square test and then Cramer V to analyse the effect.

The contingency table is generated as follows:

Rating	1	2	3	4	5	6	7	8	9	10
SideEffects										
Extremely Severe Side Effects	135	17	10	1	3	1	3	1	0	3
Mild Side Effects	7	13	15	13	34	46	128	272	243	246
Moderate Side Effects	28	20	48	47	65	66	113	143	60	22
No Side Effects	23	6	15	10	22	30	70	123	171	458
Severe Side Effects	111	46	58	36	35	13	35	19	5	10

Fig: Contingency table for Side Effects vs. Rating

This is fed to the chi-square function to get the following results:

```
# Chi-sq test for rating vs side-effects
stat2, p2, dof2, expected2 = chi2_contingency(rat_se_cont)
print(stat2)
print(p2)
print(dof2)
# Since more than 2 categories for the variable, cannot do fisher exact
```

```
2446.2134438458093
0.0
36
```

Fig: Results of chisq test showing chi-sq, p-value and degrees of freedom

Since there are more than 2 categories for one of the variables, we cannot use the Fisher exact test defined by python. So, we will skip that test.

From the above p value, the finding is seen to be statistically significant. So, we can proceed with finding the effect. Finding the Cramer's V for these variables is given as:

```
#There are 36 degrees of freedom
#For them, 0.4 shows a strong effect for cramer's v
```

```
3099
4
0.4442284721296169
```

Fig: Cramer V results

From this, we can see there is a strong effect (0.4 Cramer V for 36 degrees of freedom).

Correlation between Effectiveness and Side Effects:

This is interesting to analyse from the point of view that "If a medicine is more effective, does it also come with more side effects?".

Variables compared:

Variable	Independent/Dependent	Type of variable	Discrete/Continuous
Effectiveness	Independent	Ordinal	Discrete
Side Effects	Dependent	Ordinal	Discrete

Between two non-numeric variables, it is best to use the chi-square test and then Cramer V to analyse the effect.

The contingency table is generated as follows:

SideEffects / Effectiveness	Extremely Severe Side Effects	Mild Side Effects	Moderate Side Effects	No Side Effects	Severe Side Effects
Considerably Effective	30	361	196	255	85
Highly Effective	27	480	206	527	86
Ineffective	74	20	46	30	76
Marginally Effective	17	37	42	35	55
Moderately Effective	26	119	122	81	66

Fig: Contingency table for Effectiveness vs. Side effects

This is fed to the chi-square function to get the following results:

```
# Chi-sq test for effectiveness vs sideeffects
stat3, p3, dof3, expected3 = chi2_contingency(eff_se_cont)
print(stat3)
print(p3)
print(dof3)
# Since more than 2 categories for the variable, cannot do fisher exact
```

```
673.0147599196276
7.117623315175725e-133
16
```

Fig: Results of chisq test showing chisq value, p-value and degrees of freedom

Since there are more than 2 categories for one of the variables, we cannot use the Fisher exact test defined by python. So, we will skip that test.

From the above p value, the finding is seen to be statistically significant (7.11 e^-133). So, we can proceed with finding the effect. Finding the Cramer's V for these variables is given as:

```
#There are 16 degrees of freedom
#For them, 0.2 shows a strong effect for cramer's v
```

```
3099
4
0.2330083640182357
```

Fig: Results of Cramer's V

From this, we can see there is a strong effect (0.2 Cramer V for 16 degrees of freedom).

However, it is important to note that even though it says there is a strong effect, this does not indicate the direction. That would have to be verified further.

Section summary:

Thus, correlational analysis is carried out between some fields in the data. We will keep these conclusions in mind while proceeding with the ML part of the analysis.

SECTION 6: MACHINE LEARNING

About the section:

In this section, we will train ML models on the dataset and try to predict the sentiment (based on rating) of the review. We also try to identify the characteristics of a medicine leading to it getting a positive or a negative review. We will be carrying out classification and association tasks in this section. We will also study the models which perform the best and the most explainable models in detail. We use two new columns for machine learning. They are described in the table below:

Attributes created for ML:

Variable Name	Data Type	Variable Type	Discrete/ Continuous	Range/ Possible values
Sentiment	Text (String)	Ordinal	Discrete	Positive, Negative
Compound	Numeric	Ratio	Continuous (Made discrete for analysis)	Normalized to lie between [-1,+1]

CLASSIFICATION ALGORITHMS

Why Classification?:

In the above data, we have all the attributes from the previous section along with an additional column which states whether the review is positive or negative.

Using the classification algorithm, we can predict if the opinion on the medicine will be positive or negative. This will help the company predict how a launch of a new medicine will go based on previous data. It can also be helpful to find what the characteristics of a desirable drug are i.e, one that gets a positive review.

Modification to the data:

In addition to the columns already present in the data, we have one more column called sentiment. This has the value “Positive” or “Negative” based on the rating for that medicine.

Orange widget alignment for classification:

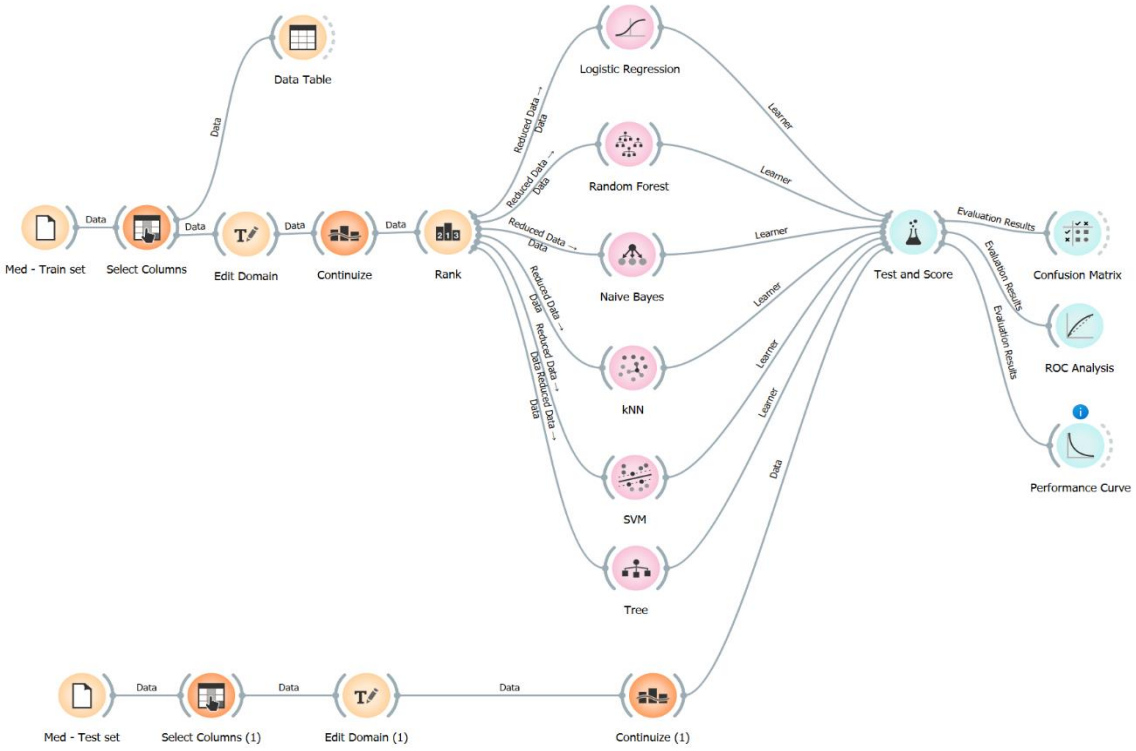


Fig: Orange widget configuration for running all the classification models

Loading the attributes:

Features: Effectiveness, SideEffects

Metas: DrugName, Condition, BenefitsReview, SideEffectsReview, CommentsReview

Targets: Sentiment

Ignored: Rating

Reason for ignoring rating: Rating has a direct relation with Sentiment (which we are trying to predict). Sentiment is directly derieved from rating, so if we keep rating we would have good ML results, but it wouldn't make sense in the real world because we wouldn't have the ratings before launch (when studying sentiment ternds about a drug).

The "Select Columns" widget is used to drop the Rating attributes.

Imposing an order on the ordinal attributes:

Here, SideEffects and Effectiveness are ordinal in nature, i.e, there is some ordering among the attributes. This ordering needs to be preserved when doing the ML modelling so that the relations established are closer to the actual scenario in the real world.

The "Edit Domain" widget is used to impose ordering on the ordinal attributes. This essentially converts the categories into numerical categories, thus providing an order to the data. The two columns are given order as follows:

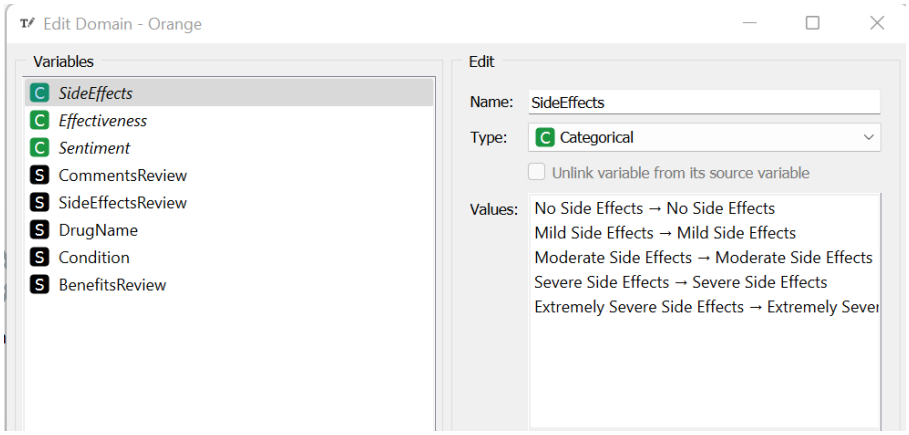


Fig: Setting order for side effects

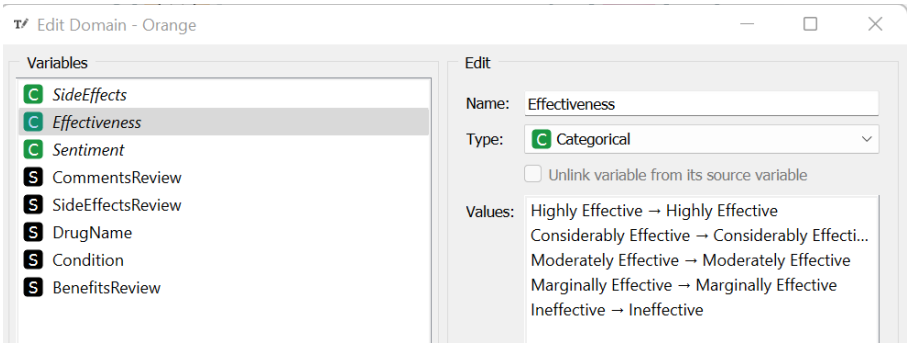


Fig: Setting order for effectiveness

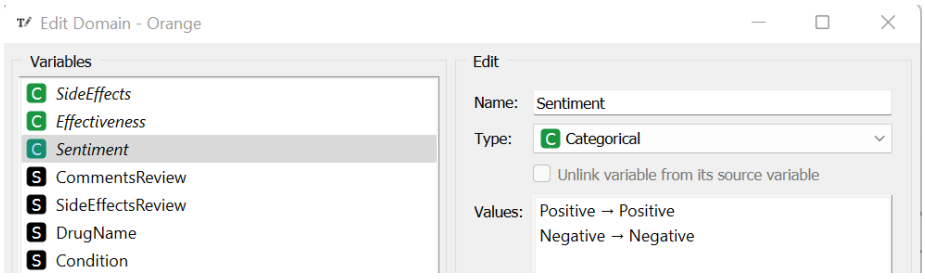


Fig: Setting order for sentiment

As we can see, the categorical attributes are arranged from the most positive scenario to the most negative scenario.

Describing how to handle each type of features/targets:

The “continuize” widget is used to tell orange how to handle each type of attribute. We select the option to treat categorical variables as ordinal and treat other types of variables as they are i.e, no preprocessing.

Ranking the features:

We perform ranking on the features to see which correlates more closely to the result. Since there are only 2 attributes, and the others are just metas, we will not reject any features based on the results of the ranking. We use gain ratio and gini index to check. The ranking results are given below:

Fig: Feature ranking for classification

Training the models:

The following classification algorithms are trained to find the results of the “Sentiment” column in the data:

1. Logistic Regression
2. Random Forest
3. Naïve Bayes
4. kNN
5. SVM
6. Tree

The results of all the algorithms are checked in “Test and Score”. The data for testing comes from the testing file. It has gone through the same preprocessing as the Training Data. The results of the algorithms are described below:

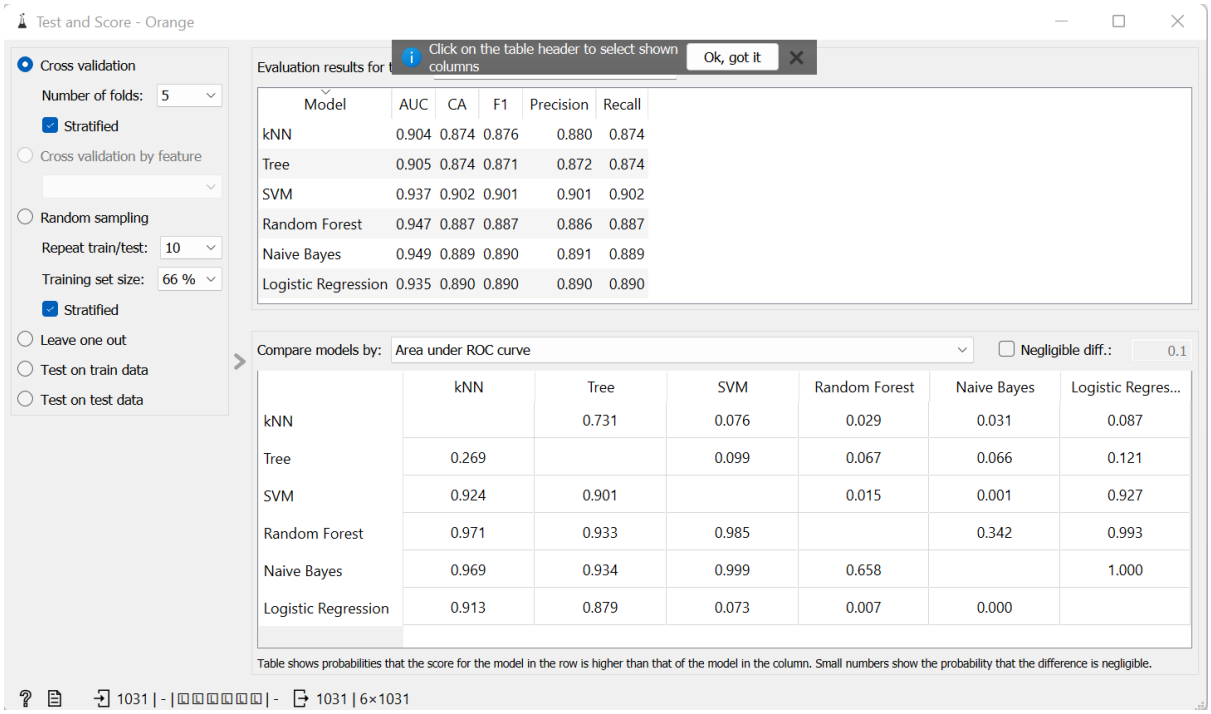


Fig: Results of the classification algorithms

We can also check the results in more detail using the confusion matrix widget. The confusion matrix for the best performing algorithm (Naïve Bayes) is included below:

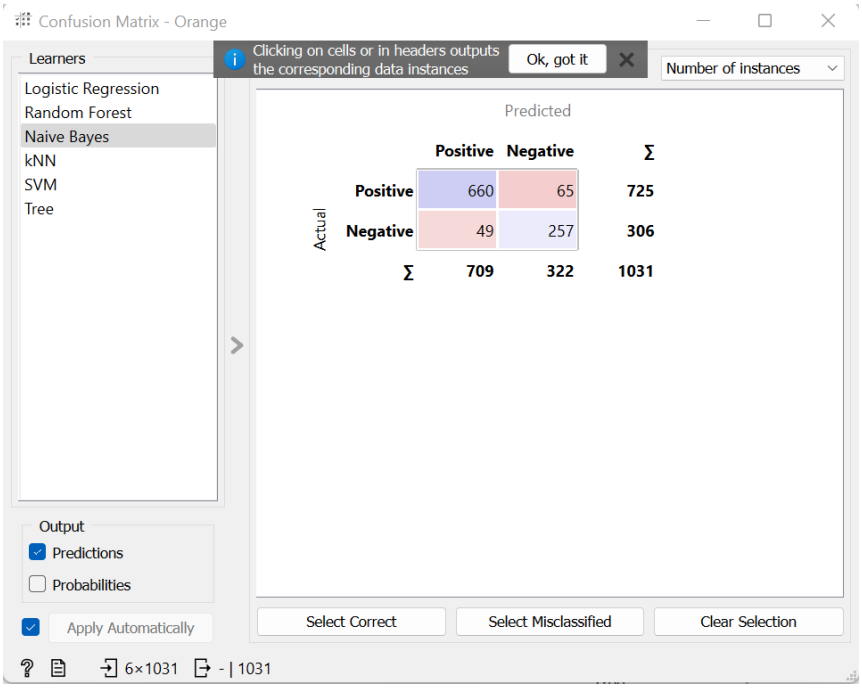


Fig: Confusion matrix for Naïve Bayes

The purple cells are the correct predictions, whereas the pink cells are incorrect predictions.

We can also study other characteristics of the models like the ROC curve and the performance curve.

ROC curve:

The ROC curve addresses sensitivity and specificity. The more area under the ROC curve, the more useful the predictions are and the less likely the results are obtained due to random chance. The ROC curves for the models are displayed below:

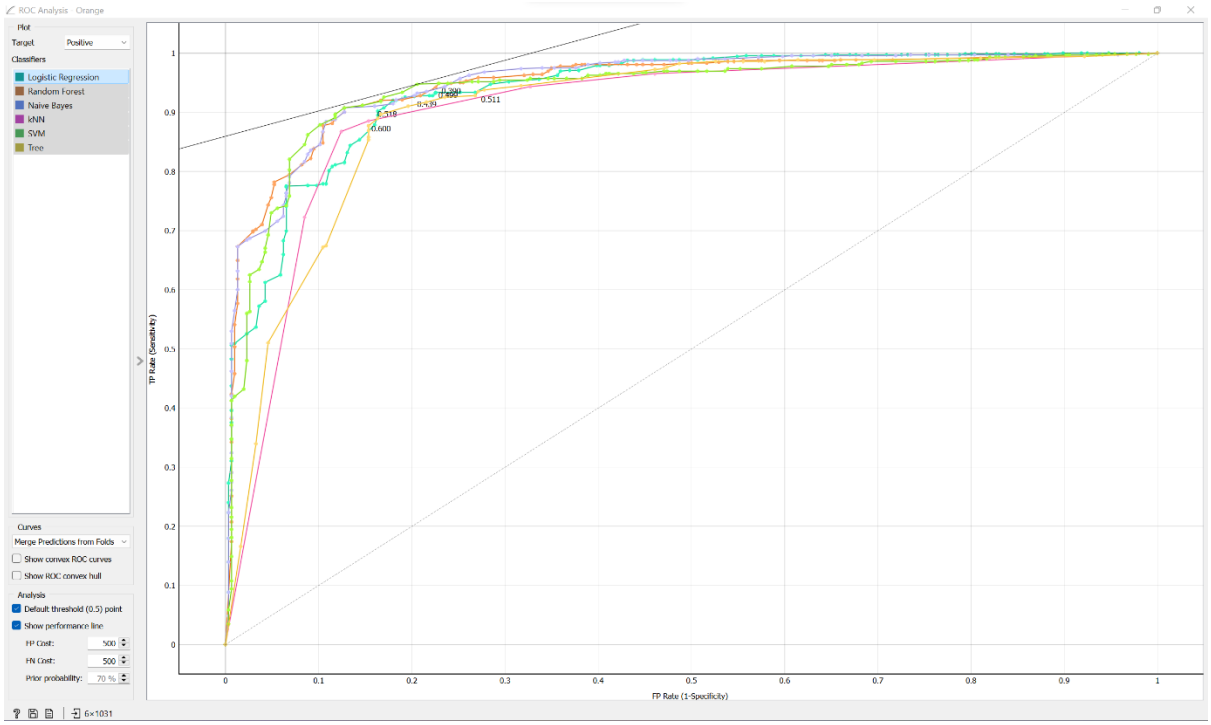


Fig: ROC curves of the 6 classification algorithms

There is a lot of area under the curves, so we can say that there is a very small chance that the results are obtained due to random chance.

Performance curve:

Performance curves are used the analyse the fit of a model. If we are making changes to the model over time, we can use the performance curves to check whether we are making good changes. The performance curve of the models are given below:

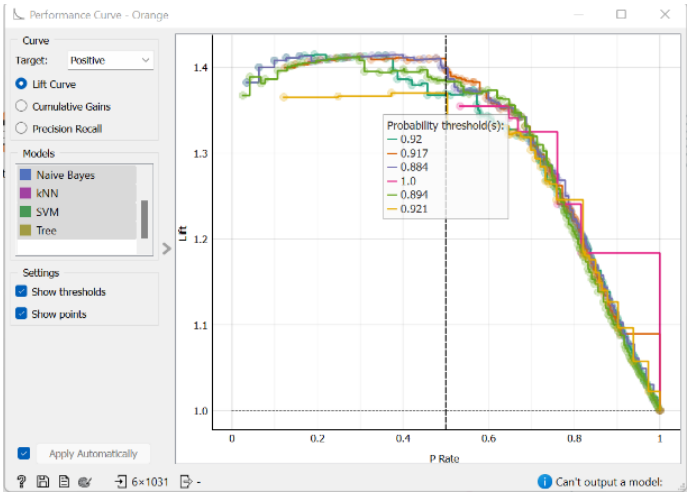


Fig: Performance curves of the 6 algorithms

Studying the Naïve Bayes Predictions:

Setup to analyse:

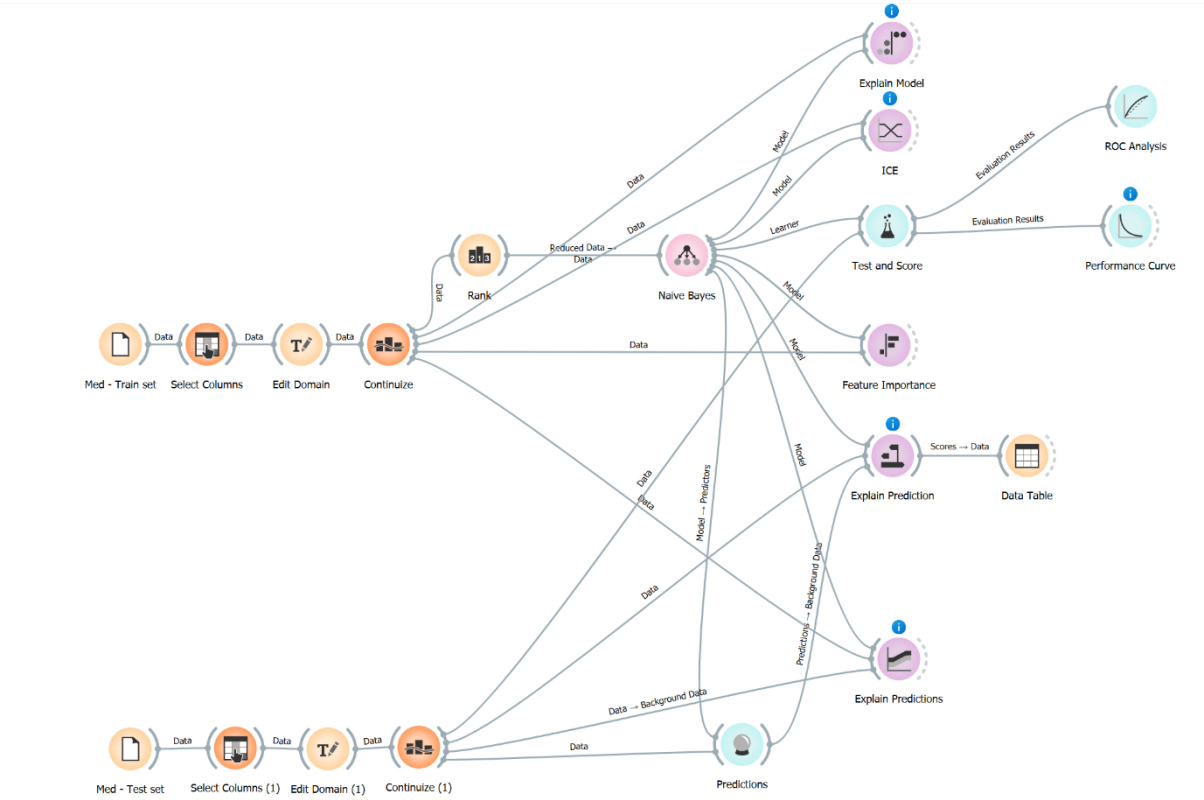


Fig: Orange widget configuration to explore Naïve Bayes

The above is the orange widget setup that will help us analyse the Naïve Bayes predictions.

Feature Importance:

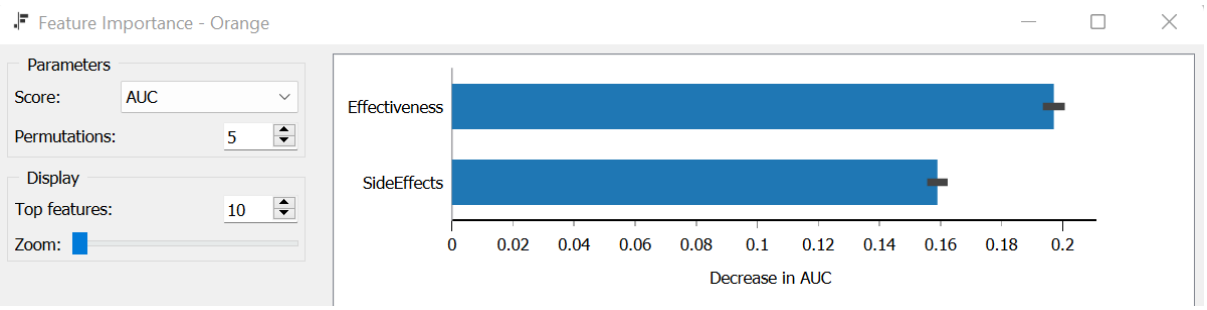


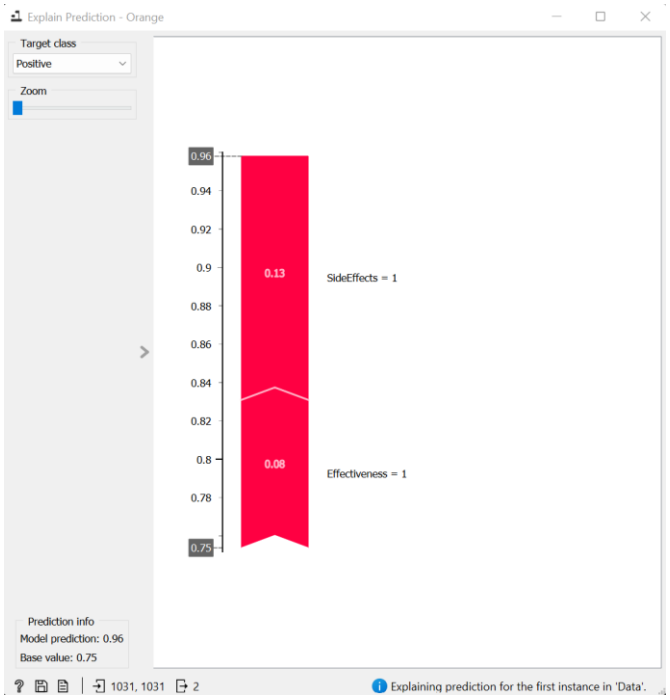
Fig: Feature importance for Naïve Bayes

According to the feature importance widget, effectiveness has more effect on the overall prediction. We will also explore this for each category.

Explain prediction:

Explain prediction is used to understand why a model classifies something into a particular category. The size of the arrow gives how much weightage that feature has for that particular prediction. It shows the

weightage of the features for a single instance. The values in red increase the chance of a prediction of the category. Values in blue decrease the chance that it will be in a category. We will check the sentiment for the first instance in the data. *[Ref - 17]*



Here, we can see that side effects has a bigger impact to putting in in the "Positive" sentiment from the size of the arrow. Since they both are red, these values influence the instance being put into positive category.

The model has predicted it is Positive with "96%" probability.

This analysis can be done for any instance in the data.

Fig: Explain predictions for Naïve Bayes

Explain predictions:

Here, we can explore what combinations of the features lead to which results. This is an interactive graph, so we can hold our cursor at different points to explore different combinations. We can also change according to the category we are trying to study.

For example:

Here, we include a screen-shot for the Positive category. If it is coloured in red, it increases the chance of something being in a category.

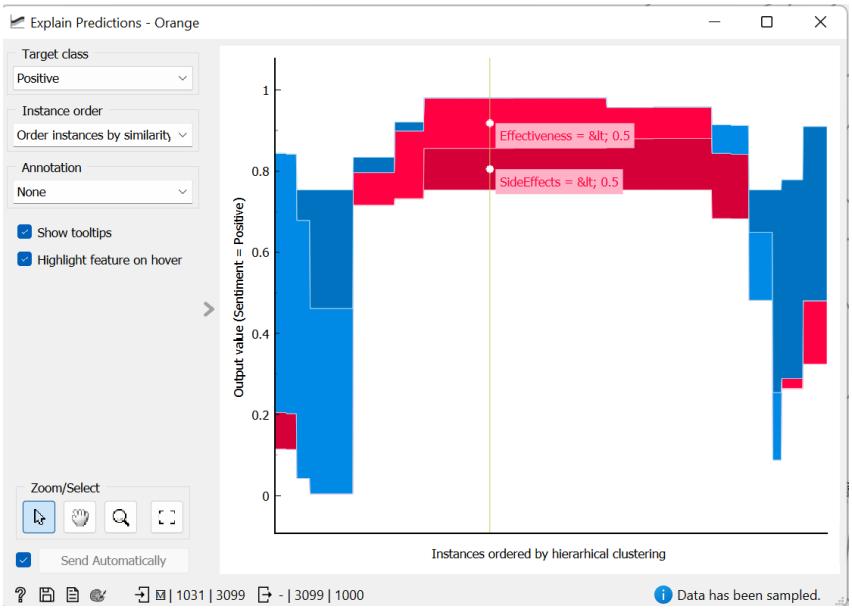


Fig: Explain predictions for Naïve Bayes

So here, an effectiveness with less that equal to 1.5 (We'll consider Highly Effective and Considerably effective) and Side Effects less than or equal to 1.5 (No side effects and mild side effects) increases the chance of a Positive Sentiment.

The values are derived from the enum values we have chosen in the Edit domain stage.

ICE:

Individual component Expectation widget shows how the chance of getting a class varies with each feature.

When checking for the chance of getting a "Positive" result:

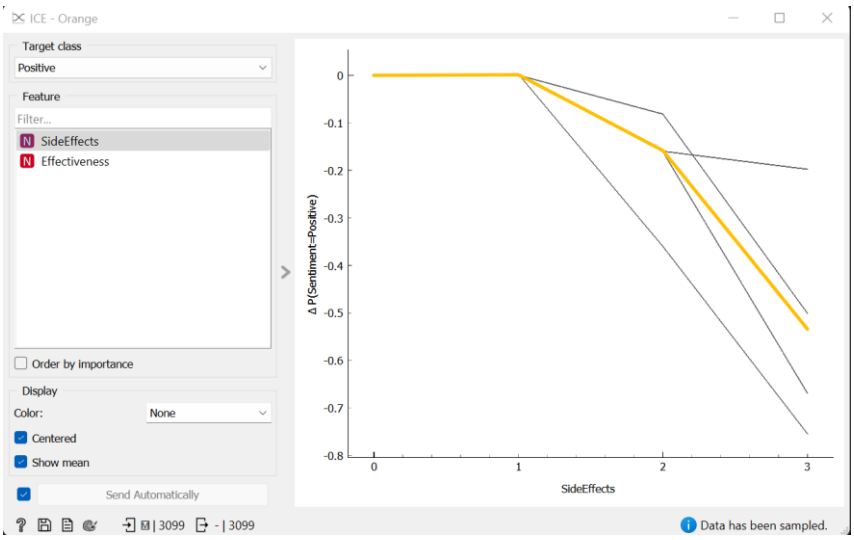


Fig: ICE widget results for Naïve Bayes for positive class side effects

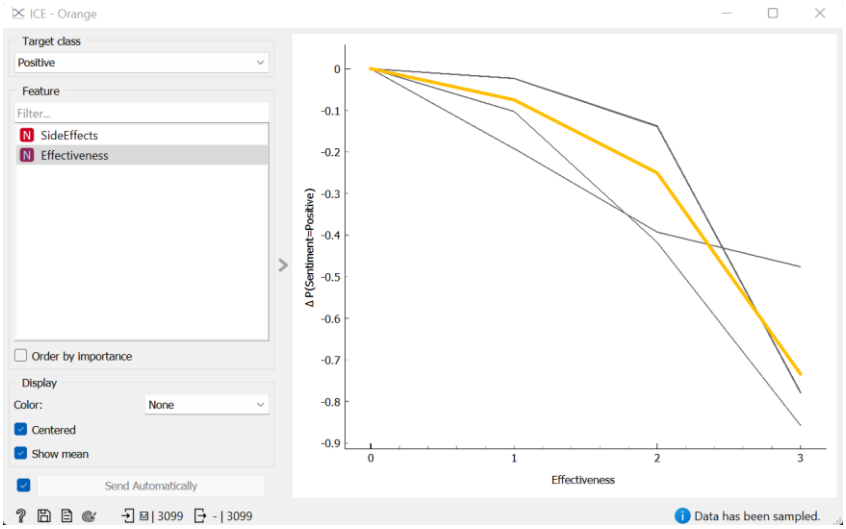
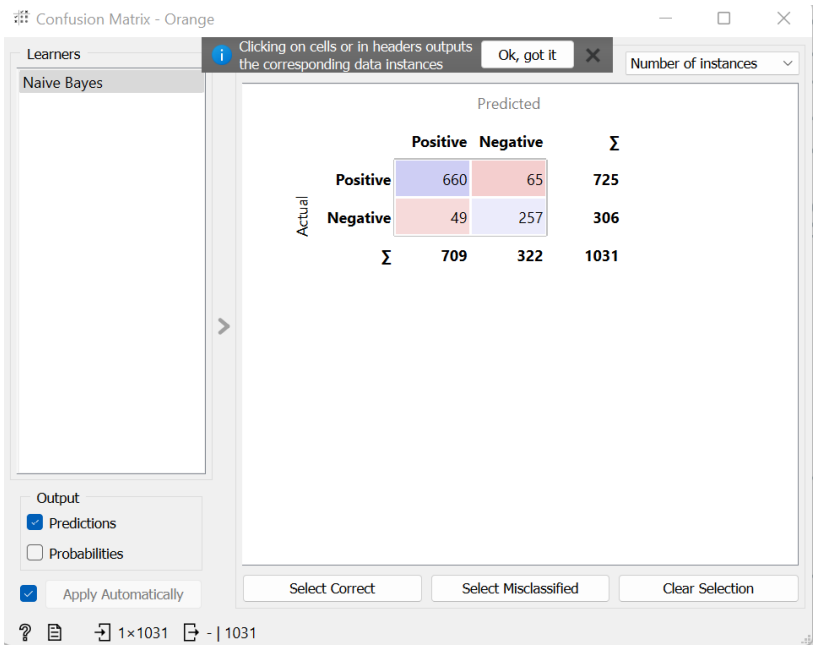


Fig: ICE widget results for Naïve Bayes for positive class effectiveness

We can see there is a marked downward trend for effectiveness and side effects as the enum value increases. This is expected because we arranged the values from most desirable to least desirable for side effects and effectiveness.

Confusion Matrix for Naïve Bayes:



Here,

Sensitivity = 0.91

Specificity = 0.83

For an algorithm related to medicine, a high specificity is recommended.

Thus, the model explanation widgets for Naïve Bayes are analysed.

Studying the random forest prediction:

Setup to analyse:

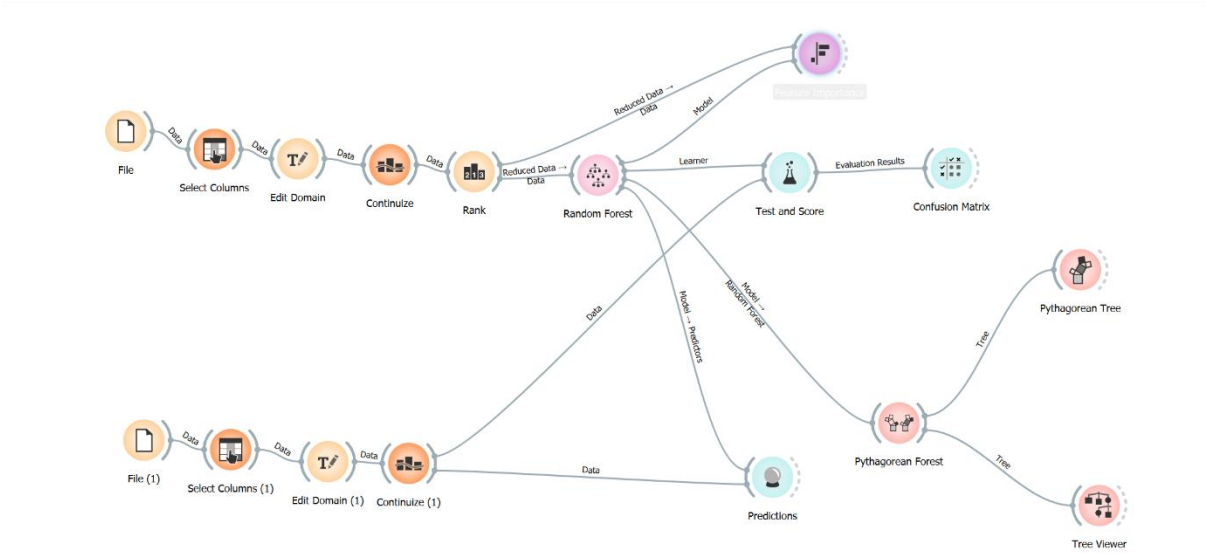


Fig: Widget configuration to study random forest in detail

Random Forests:

For random forests, the algorithm generates a set of trees called Pythagorean trees. The combined results of all these trees are returned as the final solution. Using orange, we can generate the Pythagorean trees and then study them.

The set of Pythagorean trees generated is:

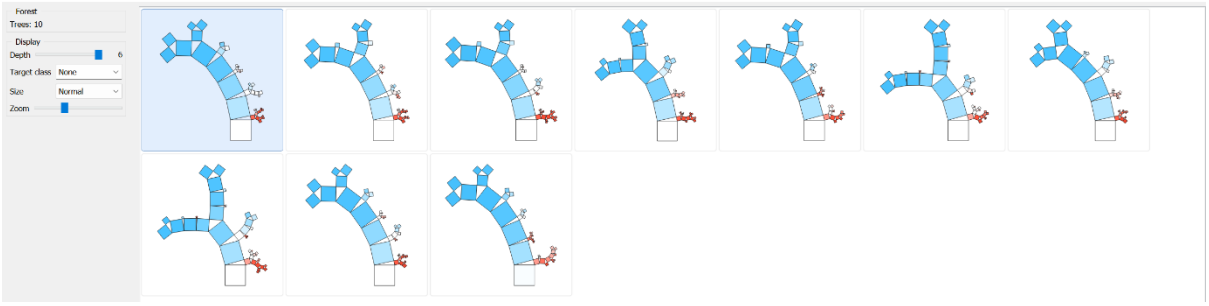


Fig: All the trees in the random forest classifier

The best tree is the one with the shortest height and most densely coloured branches. Therefore, the first tree in the second row is chosen for study.

The content inside the tree is given as:

Positive side of the tree:

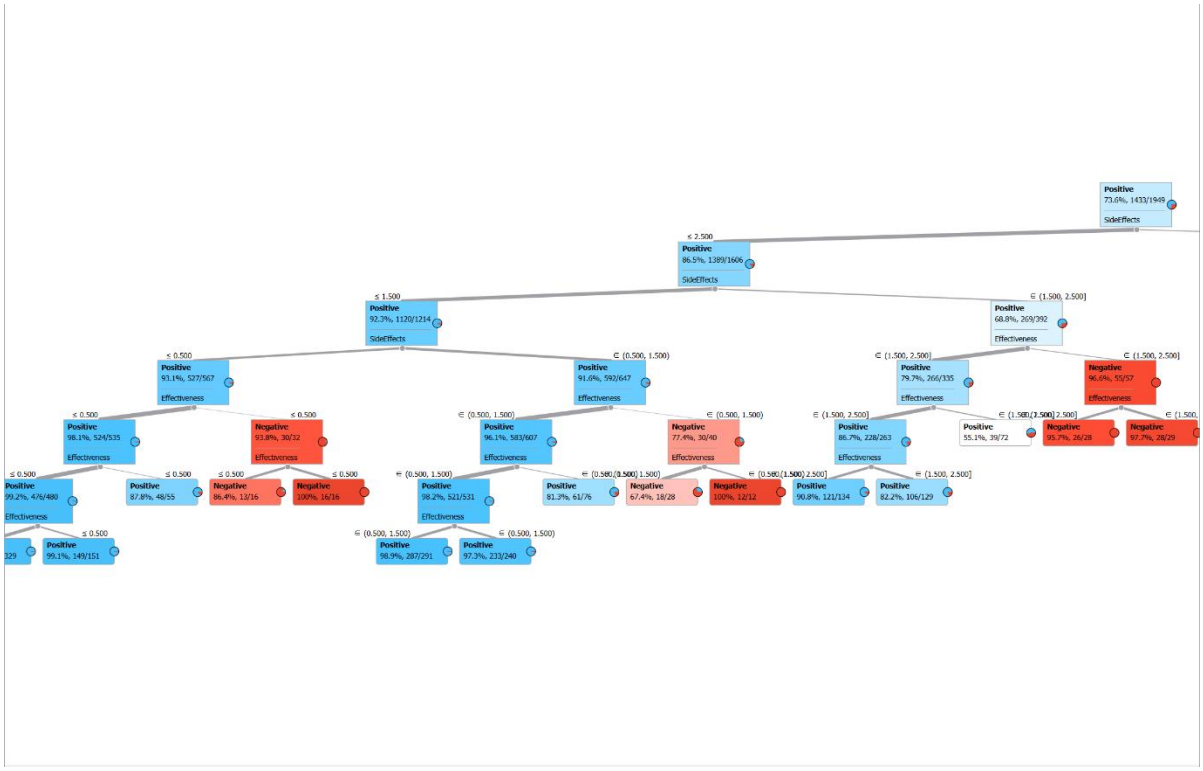


Fig: Half of the tree generated by random forest

(The tree is too large to have in a single picture, so the positive label side and negative label side are shown in two screenshots)

Negative label side of the tree:

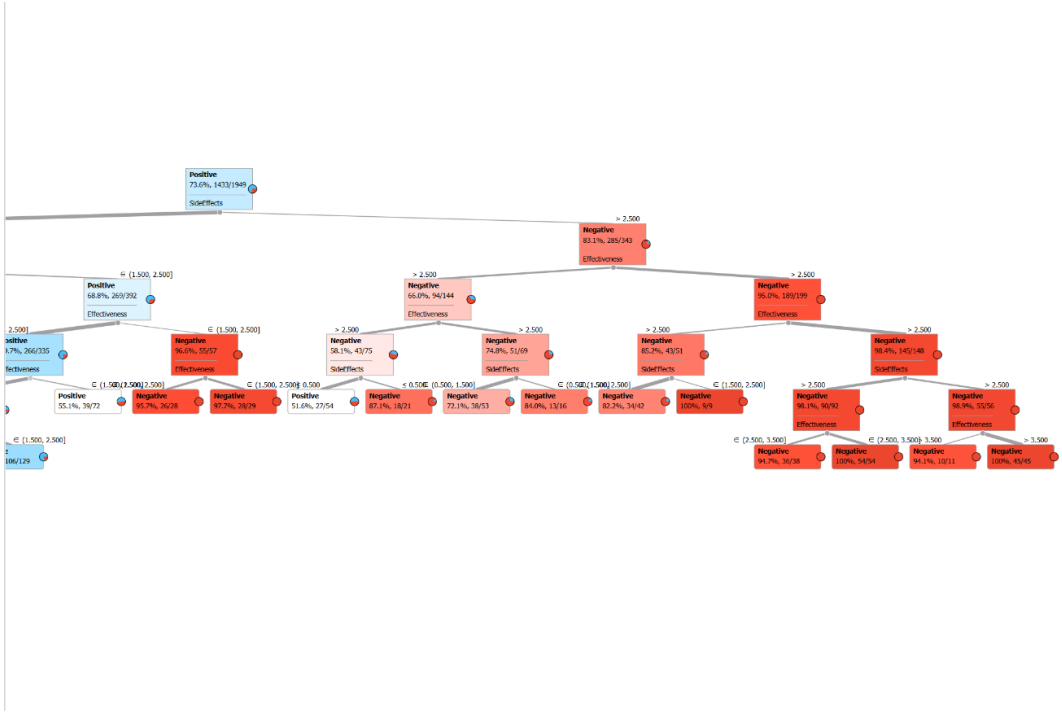


Fig: Other half of the tree generated by random forest

As we can see, it is hard to derive rules from the tree directly, so we will go for mining association rules later.

Confusion matrix:

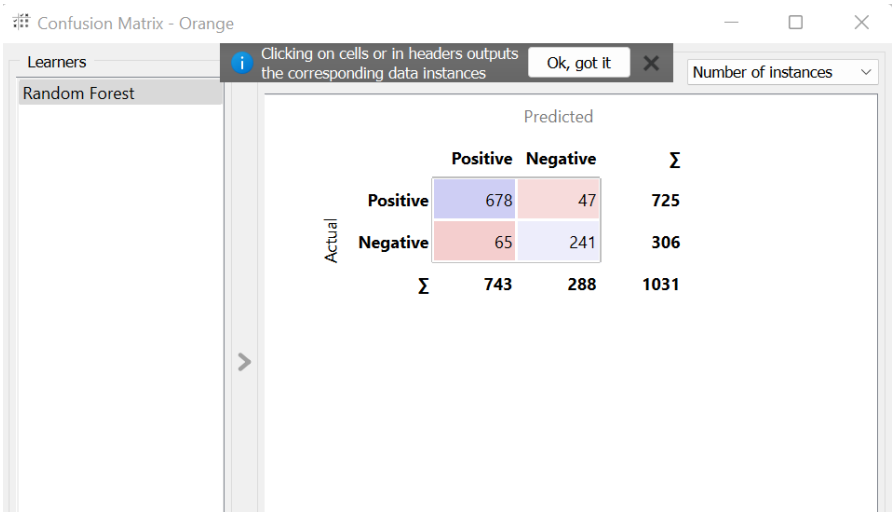


Fig: Confusion matrix for random forest

For this matrix,

Sensitivity = 0.93

Specificity = 0.78

Thus, the random forest classifier and the results it produces is studied.

Studying the decision tree prediction

Setup to analyse:

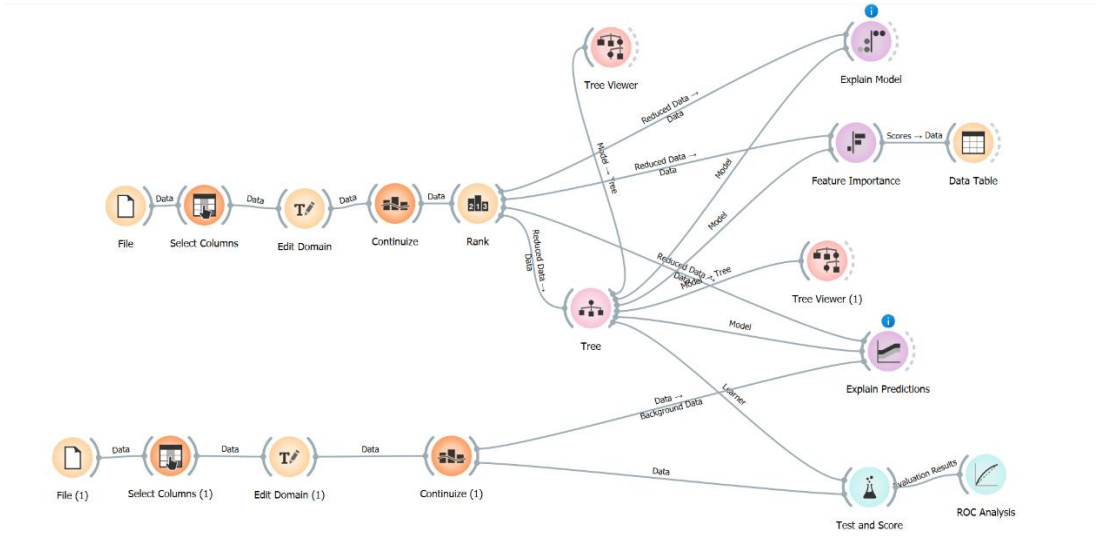


Fig: Orange widget configuration to analyse decision trees

The decision tree is not as accurate compared to the other models, but the final tree is very clear and easy to understand.

The tree generated by this model is given by:

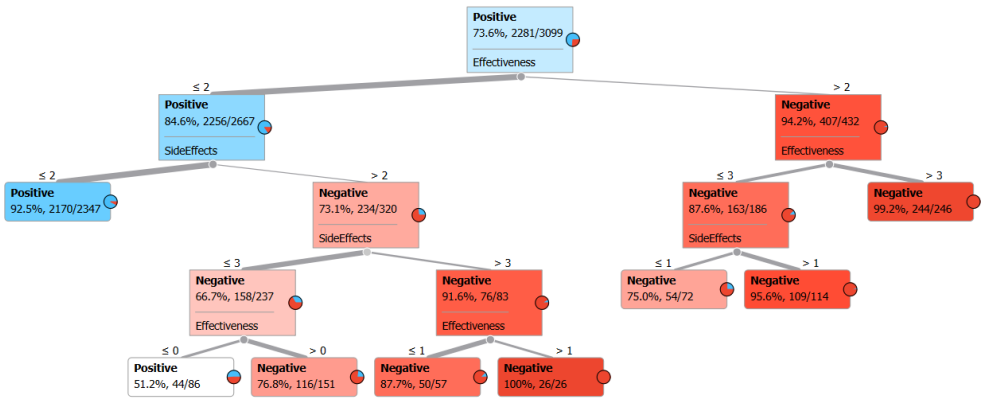


Fig: Decision tree model

<u>Inferences from above tree:</u>

Here, we can see that the effectiveness is the most important decider. If the effectiveness is in the first 2 categories (Highly effective, Considerably effective), and also the side effects are in the first 2 categories (No side effects, Mild side effects) then the review is positive. Otherwise, the review is mostly negative.

<u>Confusion matrix for decision tree:</u>

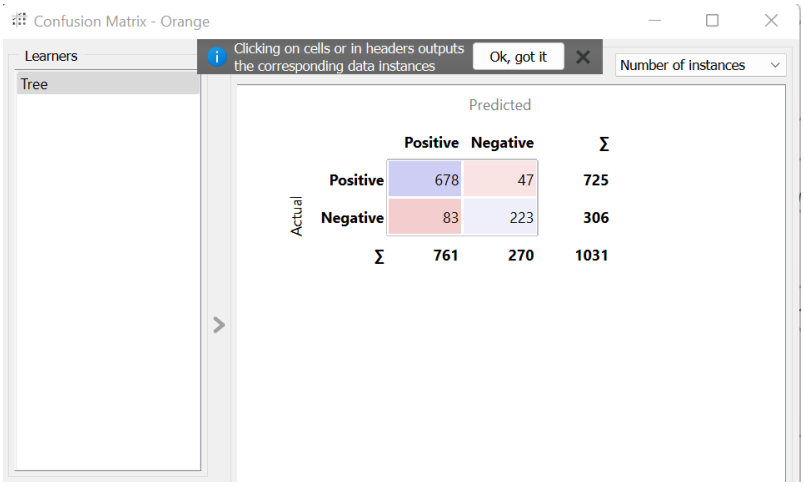


Fig: Confusion matrix for decision tree

Sensitivity = 0.93

Specificity = 0.72

Thus, the decision tree and the results it produces are analysed.

ASSOCIATION RULES

Association algorithms are used to generate rules that are applicable to the data in the dataset. These rules can also be used to predict the characteristics of future data for that set and help us understand the nature of the data.

There are two important metrics for association:

Support – How often a rule appears in a dataset

Confidence – How often the rule is true in the dataset

We examine association rules in two ways. In one, we include only the fields in the dataset. In another, we use VADER to derive the opinions in the reviews and put it as a separate column and then mine for rules.

Orange setup:

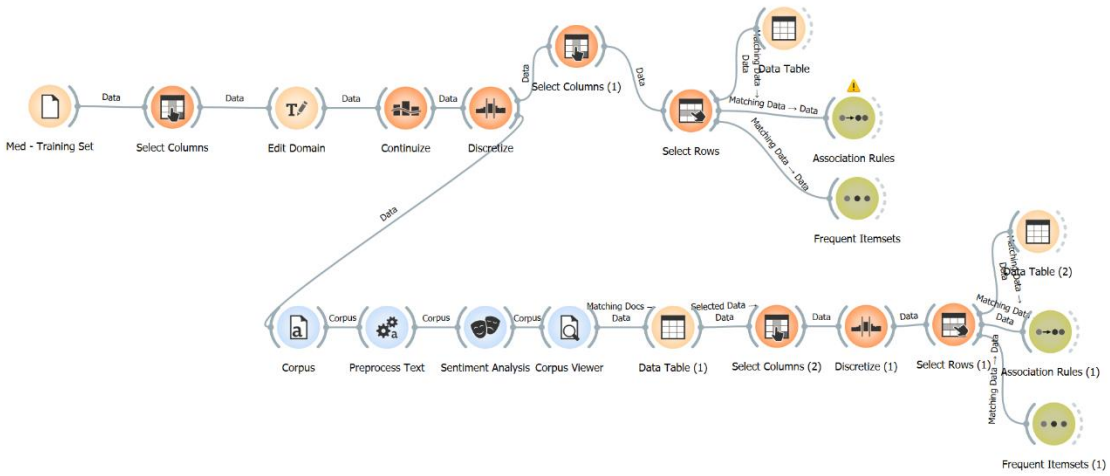


Fig: Orange widgets for association rule mining

As we can see, the top half of the widget configuration does not use VADER for opinion mining. The bottom half generates a column "Compound" which contains the compound score generated by VADER for the document. "Select rows" makes sure only the rows with all variables filled are used. "Discretize" is used to convert continuous attributes into discrete ones that can be used in association.

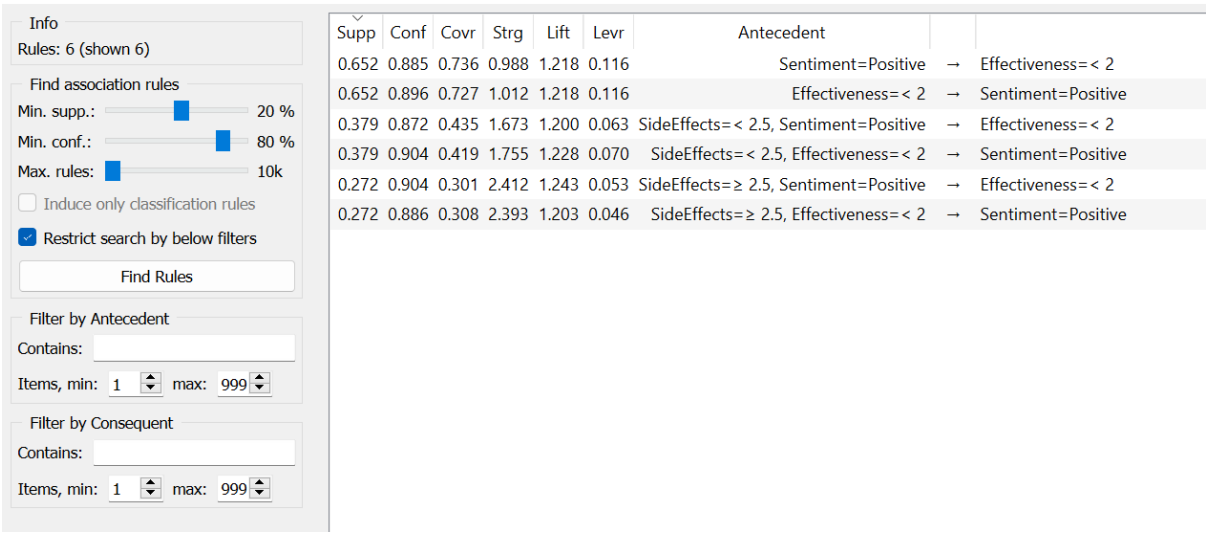
Note: In both cases, we have used different thresholds for support and confidence. This is based on the approximate number of rules/ frequent itemsets that we want.

Without VADER:

Confidence – 80%

Support – 20%

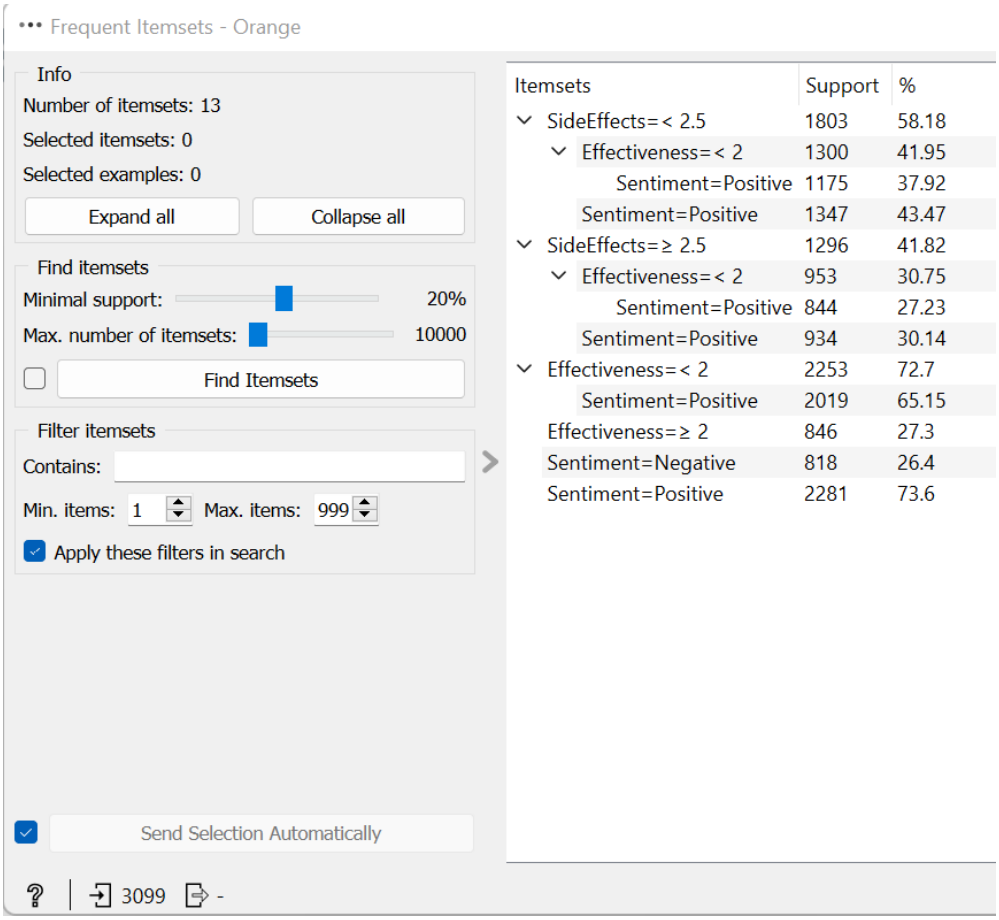
Association Rules:



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.652	0.885	0.736	0.988	1.218	0.116	Sentiment=Positive	→	Effectiveness=< 2
0.652	0.896	0.727	1.012	1.218	0.116	Effectiveness=< 2	→	Sentiment=Positive
0.379	0.872	0.435	1.673	1.200	0.063	SideEffects=< 2.5, Sentiment=Positive	→	Effectiveness=< 2
0.379	0.904	0.419	1.755	1.228	0.070	SideEffects=< 2.5, Effectiveness=< 2	→	Sentiment=Positive
0.272	0.904	0.301	2.412	1.243	0.053	SideEffects=≥ 2.5, Sentiment=Positive	→	Effectiveness=< 2
0.272	0.886	0.308	2.393	1.203	0.046	SideEffects=≥ 2.5, Effectiveness=< 2	→	Sentiment=Positive

Fig: Association rules mined without using VADER column

Frequent Itemsets:



Itemsets	Support	%
SideEffects=< 2.5	1803	58.18
Effectiveness=< 2	1300	41.95
Sentiment=Positive	1175	37.92
Sentiment=Positive	1347	43.47
SideEffects=≥ 2.5	1296	41.82
Effectiveness=< 2	953	30.75
Sentiment=Positive	844	27.23
Sentiment=Positive	934	30.14
Effectiveness=< 2	2253	72.7
Sentiment=Positive	2019	65.15
Effectiveness=≥ 2	846	27.3
Sentiment=Negative	818	26.4
Sentiment=Positive	2281	73.6

Fig: Frequent itemsets generated without VADER column

From this, we are able to see some rules like:

- If the effectiveness is in the first two categories (Highly effective, Considerably effective) then the sentiment is positive.
- If the side effects are in the first two categories (No side effects, Mild side effects) and the effectiveness is in the first two categories (Highly effective, Considerably effective) then the sentiment is positive.
- If the effectiveness is in the first two categories (Highly effective, Considerably effective) and the side effects are in the last 3 categories (Moderate side effects, Severe side effects, Extremely severe side effects), the sentiment is positive.

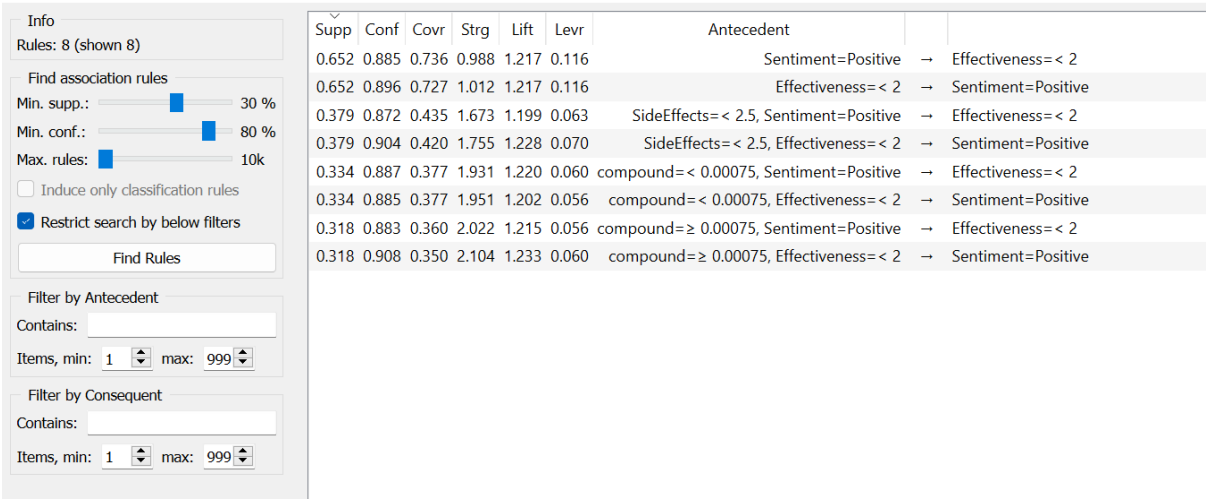
With VADER:

Support – 30%

Confidence – 80%

Before doing opinion mining using VADER, we need to preprocess the text. Stopwords in English are removed. Numbers are also removed. POS tagging of nouns and verbs is done. We do not include any addition vocabulary while running VADER. VADER produces a positive, negative, neutral and compound column. We will use the compound column as it is derived from the other 3.

Association Rules:



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.652	0.885	0.736	0.988	1.217	0.116	Sentiment=Positive	→	Effectiveness=< 2
0.652	0.896	0.727	1.012	1.217	0.116	Effectiveness=< 2	→	Sentiment=Positive
0.379	0.872	0.435	1.673	1.199	0.063	SideEffects=< 2.5, Sentiment=Positive	→	Effectiveness=< 2
0.379	0.904	0.420	1.755	1.228	0.070	SideEffects=< 2.5, Effectiveness=< 2	→	Sentiment=Positive
0.334	0.887	0.377	1.931	1.220	0.060	compound=< 0.00075, Sentiment=Positive	→	Effectiveness=< 2
0.334	0.885	0.377	1.951	1.202	0.056	compound=< 0.00075, Effectiveness=< 2	→	Sentiment=Positive
0.318	0.883	0.360	2.022	1.215	0.056	compound=≥ 0.00075, Sentiment=Positive	→	Effectiveness=< 2
0.318	0.908	0.350	2.104	1.233	0.060	compound=≥ 0.00075, Effectiveness=< 2	→	Sentiment=Positive

Fig: Association rules mined using the compound column from VADER

Frequent Itemsets:

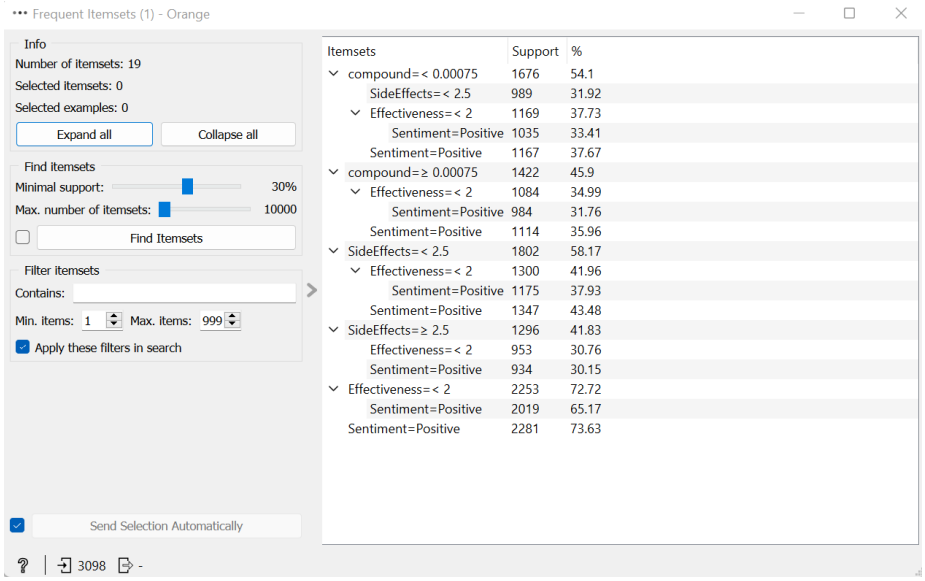


Fig: Frequent itemsets mined using the compound column from VADER

The rules generated after adding VADER are similar to the rules generated without it. However, there are two rules in this set which are interesting. For one rule “compound<= 0.00075, effectiveness <= 2 -> sentiment = ‘Positive’” and the other is “compound>= 0.00075, effectiveness <= 2 -> sentiment = ‘Positive’”. From these two, we can see that adding the column did not generate much changes in the rules at least at this threshold (There are opposite values for compound, but the rest of the rule is the same).

ANALYZING SIDE EFFECT SENTIMENTS

Goal of analysis:

There is an interesting point that reviews may sometimes be subjective. The goal of this part is to verify whether some conditions i.e, side effect reviews that included words like depression, anxiety may lead to a higher frequency of negative reviews instead of positive ones. We compare the frequency distribution of sentiment in reviews containing such keywords against reviews containing other side effects as keywords in order to verify the above.

Orange setup:

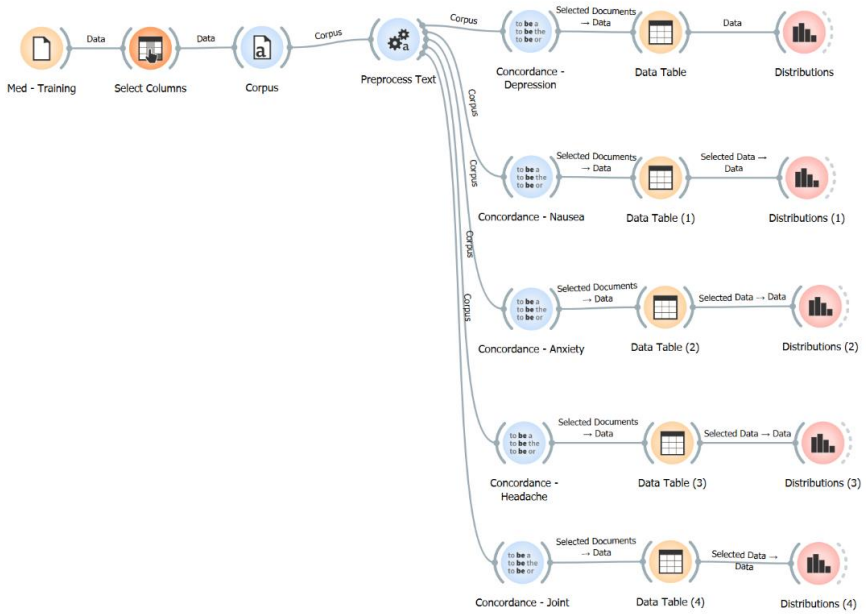


Fig: Orange widgets for analysing each side effect separately

Here, after preprocessing (removing numbers, stopwords, etc), concordance is used in order to find the Side Effect Reviews which have the specified keyword. After filtering those out, we can study the distribution.

Concordance query example:

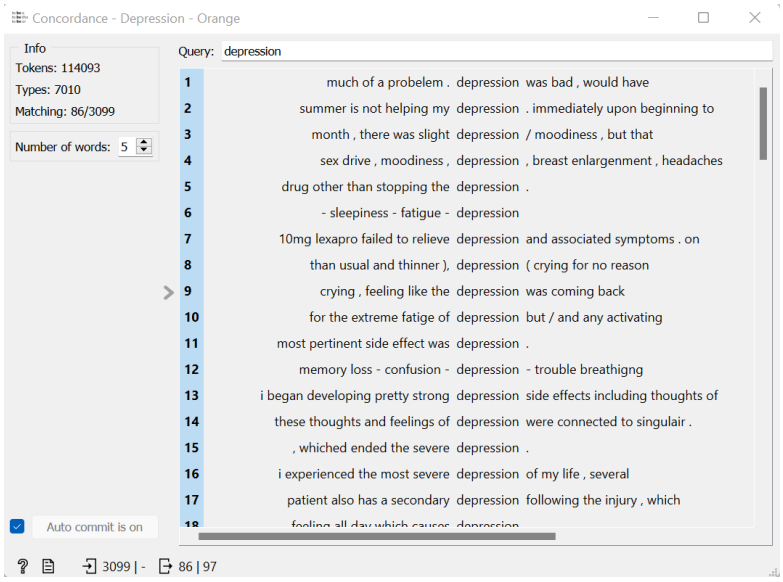


Fig: Concordance query example. Here, query word is 'depression'

Distribution example:

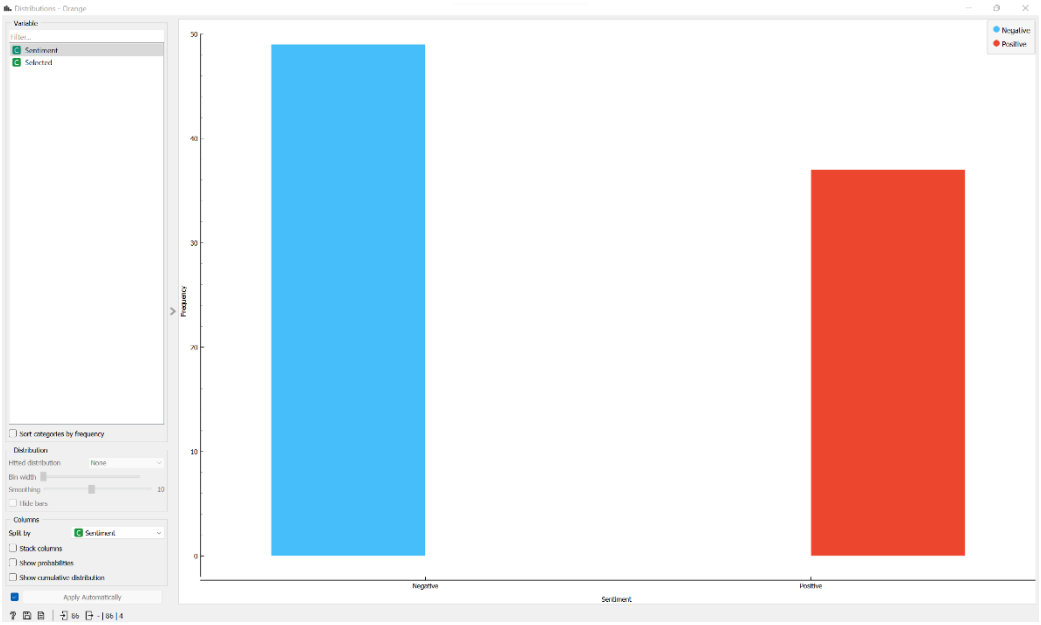


Fig: The distribution of reviews for depression

Results of analysis:

The results of the analysis after checking the distributions is tabulated below:

Side Effect	Query Keyword	No of matches	% of Negative sentiment	% of Positive sentiment
Depression	depression	86	56.98%	43.02%
Nausea	nausea	200	33.50%	66.50%
Anxiety	anxiety	83	55.42%	44.58%
Headache	headache	92	26.09%	73.91%
Joint pain	joint	35	57.14%	42.86%

Thus, from the above, we can see that Depression, Anxiety, Joint pain have a greater number of negative reviews than positive ones. For Nausea and Headache, it is vice versa.

Since joint pain also has a higher number of negative reviews than positive ones, we cannot conclude that our initial premise is true. Thus, the mental stress of the consumer may not have such a huge impact on the skew of reviews.

Thus, research can be extended into:

- Comparing more painful side effects vs. less painful ones.
- If a side effect is more common (such as headache or nausea), are consumers willing to overlook it and leave a positive review?

Thus, the reviews are segregated based on the side effects mentioned and analysed.

ML section summary:

Classification summary:

Algorithm Name	Accuracy	Precision	Recall	Settings info if any
Naïve Bayes	0.949 (94%)	0.891	0.889	N/A
Random Forest	0.948 (94%)	0.891	0.892	10 trees. Don't split subsets smaller than 5.
Logistic regression	0.935 (93%)	0.890	0.890	Regularization type: Ridge L2
SVM	0.935 (93%)	0.901	0.902	Cost: 1, Epsilon: 0.1. Kernel – RBF. Iteration limit: 100.
Tree	0.905 (90%)	0.872	0.874	Induced binary. Min 2 instances in node. Don't split subsets smaller then 5. Max tree depth 100.
kNN	0.904 (90%)	0.880	0.874	5 neighbours, Euclidean distance

Association summary:

Method	Min. Confidence	Min. Support	Rules
Without VADER compound column	80%	20%	- If the effectiveness is in the first two categories (Highly effective, Considerably effective) then the sentiment is positive. - If the side effects are in the first two categories (No side effects, Mild side effects) and the effectiveness is in the first two categories (Highly effective, Considerably effective) then the sentiment is positive. - If the effectiveness is in the first two categories (Highly effective, Considerably effective) and the side effects are in the last 3 categories (Moderate side effects, Severe side effects, Extremely severe side effects), the sentiment is positive.
With VADER compound column	80%	30%	Same rules as above. We also find that the 'compound' column doesn't have much impact on rules

Most important variable:

As shown in the Naïve Bayes section, the most important variable is Effectiveness followed by SideEffects.

Section summary:

Thus, various machine learning algorithms are run on the dataset. We perform classification tasks and mine association rules. The results of the ML models are tabulated above.

SECTION 7: OTHER RESEARCH

Survey of papers that use the same dataset or are in the same domain

The dataset in UCI was generated as part of a paper by *(Gräßer, Kallumadi, Malberg, & Zaunseder, 2018)*. They scraped the websites drugs.com and druglib.com in order to obtain the dataset. Their research was done from the perspective of 'post market surveillance' i.e, watching how a drug performs after it is released and monitor for ADRs (Adverse Drug Reactions). Their focus was on sentiment analysis on unstructured drug reviews and classification. The data is scraped using beautiful soup. They wanted to predict overall patient satisfaction, transferability over conditions and transferability over sources. One interesting thing is that they ran different models for the data from drugs.com and druglib.com. For preprocessing, they converted the alphabets to lower case, removed numbers and terms with high frequency. However, stopwords were not removed. They were able to achieve an overall accuracy of 75.29% for the dataset from drugs.com and 70.06% for the data from drugslib.com.

(Shiju & He, 2022) have also used the data from drugs.com in their research to classify drug reviews using transformer-based language models. They used beautiful soup to scrape the data from drugs.com. They applied n-grams in order to study the data in context. They applied the algorithms like BERT, XLNet and Electra in order to classify the reviews into above average and below average. Of all the algorithms they used, BioClinicalBest and Electra performed the best.

A medical recommender system was proposed by *(Ramya, Sumitha, Ranjani, & Ahamed, 2022)* in their paper. It would take an age-wise, gender-wise and side-effect-wise grouping when recommending the medicines. It would separate side effects from other aspects in the reviews while learning. They used a semi-regulated approach for learning. CNN, CSTM, WKSM were used when checking the drugs the consumers were given. They used SVM, decision tree and random forest models.

An aspect-based approach was adopted by *(Imani & Noferesti, 2022)* in their research where they would split the review into various different aspects like effectiveness, side effects, etc and then they would run a pre-trained deep directional transformer along with BERT on it. BERT can help understand the words in context and a 9-word window was used by them to study the data. Their prepocessing included sentence detection, tokenization, part of speech tagging, specifying semantic classes, replacing verbs and lemmatization. Their syntax trees were generated using Stanford core NLP tool. This would help in parts of speech analysis. They also obtained their data from drugs.com and drugslib.com. With a fine-tuned BERT, they were able to reach an F measure of upto 78.05%.

(Joshi & Abdelfattah, 2021) also used the UCI drug dataset on Multi-class text classification of online drug reviews. They used Term Frequency-Inverse Document Frequency (TF-IDF) Vectorization. TD-IDF represents the importance of a term considering all other documents too. If a word appears many times in a document, but less times in other documents it is considered to be important. They ran several ML models like Multinominal Naïve Bayes, Multinomial Logistic Regression, Linear Support Vector Classifier (SVC), Decision Trees, Extra Trees, and Random Forests. Of these, the best scores were achieved from Linear SVC with a precision of 0.8832, a recall of 0.8817 and an F-score of 0.8825.

(J, Cambria, & Trueman, 2021) also used the same websites as a source for their data, but their goal was to create a strongly labelled dataset which can serve as a corpus for other research. Their dataset was called DUSE and it was generated with the help of SenticNet which is a neurosymbolic artificial intelligence framework for sentiment analysis. This dataset is evaluated using baseline models

like Logistic regression and Naïve Bayes. The Logistic regression and Naïve Bayes are used to represent the Bag of words features, whereas Gated Recurrent Units represent the context independent features. BERT is used for sentiment analysis like *(Imani & Noferesti, 2022).* K-train python library is also used in order to create a baseline. When such baselines are created, the performance of BERT is also improved.

(Yadav & Vishwakarma, 2020) also trained models like SVM, Decision Tree, Random Forests, Naïve Bayes, and K-Nearest Neighbour on Medicinal drug reviews in order to perform sentiment analysis. What is interesting about their work is that while most of the other research usually removes all numbers from the input text data, they convert their numbers into words i.e, 1 to "one", etc in their research. They wanted to find the relationship between the polarity and the popularity of drugs (i.e, is there any strong sentiment about a drug vs. how often it is used).

The drug reviews were used to get the top drugs for a given condition in the hopes of curbing self-medication without any background info by *(Garg, 2021)*. He used the patient reviews to predict the sentiment using various vectorization processes like Bow, TD-IDF, Word2Vec and Manual Feature analysis. Smote techniques were used in order to provide more data for the drugs for rare conditions. Logistic regression and Multinominal Naïve Bayes are used to predict the sentiment. Multinomial Naïve Bayes had the highest precision of 0.93%.

(Min, 2019) proposed using a weakly supervised mechanism at first to pre-train the data and then used labelled data to fine tune the model. This is done in order to reduce the effect of noise in the model. He used Convolutional Neural Network (CNN) and Bi-directional Long Short-term Memory (Bi-LSTM) named as WSM-CNN-LSTM to complete the task for ADR reviews sentiment classification.

Aspect level drug reviews were done using BiGRU and Knowledge transfer by *(Han, Liu, & Jing, 2020)*. They proposed pretraining and multi-task learning model. At first the pretrained weight learned from short textlevel sentiment analysis is used to initialize the weight of the model. Then two BiGRU networks are applied to generate the bidirectional semantic representations. Attention mechanism is used to obtain the target-specific representation for aspect-level drug review. The main goal was to create "SentiDrugs" which is a dataset for aspect-level review sentiment classification.

A prediction system based on patient evaluation (i.e, reviews from drugs.com and druglib.com) was proposed by *(Jayale & Desai, 2021).* They also take data from webmd in order to provide a more comprehensive outlook on the conditions of the patients. Drug prediction is done for COVID-19 based on protein-to-protein reactions and availability.

There are some other research papers that are about the same topic, but do not use data from drugs.com or druglib.com. The content in those papers is discussed below:

(Cavalcanti & Prudêncio, 2017) worked on aspect extraction and aspect classification on drug reviews. These concepts are touched several times in the above papers too. The dataset used by them contained reviews related to ADHD, Anxiety and AIDS. Sometimes ADRs might not be detected until the drug goes to the market. This sort of opinion mining will help companies keep a watch on reviews and get news about ADRs. Their goal was to establish a method to get the aspects from reviews and produce a corpus that can be used to train future models. They use a supervised method based on domain knowledge and linguistic features. They worked on the aspects like overall, effectiveness, side effects, dosage and cost. Conditional random fields and Hidden Markov models are used to detect adverse drug

reactions. MedTagger is used to provide additional context on the data. The steps include preprocessing, aspect extraction, aspect classification and creating syntactic trees.

(Sampathkumar, Chen, & Luo, 2014) also used Hidden Markov Models to mine ADRs. Their data is sourced from medications.com and stayhealthy.com. The steps followed in their research was Information Retrieval, Text processing and Information extraction. The text processing is done using a chain of NLP tools. HTML tags are removed from the data, punctions and stopwords are removed and stemming is done. Named Entity Recognition is also used. SIDER is used to get the side effects. Viterbi decoding is used along with the HMM. If a side effect occurs more than once in a review, the duplicate is removed. The evaluation is done using 10-fold validation. They experiment with and without stopword removal. It is found that that part does not have much impact on the result. One interesting thing that they point out is that reviews are very subjective and sometimes users may not know what to describe.

(Gopalakrishnan & Ramaswamy, 2017) perform patient opinion mining using supervised learning. They compare their approach (Neural Networks) to SVM. They chose SVM for comparison because they felt that it is a commonly used technique for this problem. They obtained their data from askapatient.com. Scraping is done using Java and parts of speech tagger is used during preprocessing. The used the KNIME tool to implement probabilistic neural network along with radial bias functions. The results are measured using F-scores, precision and recall. The F-score they got for SVM is 79.7%, for probabilistic neural network it is 85.5 and for radial bias function neural network is is 92.5%. Thus, a marked improvement is seen in RBFN.

(Korkontzelos et al., 2016) preferred to work with tweets and forum posts while conducting their analysis. Their data was sourced from Dailystrength and twitter. Their goal was to create a spontaneous reporting system. They used two medical experts to annotate their dataset. Only the data where the two experts agreed was used. They followed the usual guidelines like pos tagging, stemming, tokenization, negation, normalization, polarity lexicon, etc. They also used n-grams. They used a 10 x 10 fold validation to evaluate the results.

Now, we will find similar techniques used in the paper that were followed in our own analysis. In preprocessing, we follow techniques similar to *(Imani & Noferesti, 2022)* and *(Sampathkumar, Chen, & Luo, 2014)* with removing the stopwords, numbers and doing the POS tagging. We have run ML models to classify the reviews based on sentiment which are similar to *(Gopalakrishnan & Ramaswamy, 2017)* and *(Joshi & Abdelfattah, 2021)*. We have used classification models like Naïve Bayes, Decision trees, Random forests, etc. During the analysis phase, we have tried to identify the popular drugs and study their characteristics similar to *(Yadav & Vishwakarma, 2020).*

Section summary:

Thus, various papers that use the dataset, datasource or are in the same domain are studied and comparison is performed between them and our own analysis.

SECTION 8: ACTION PLANS BASED ON ANALYSIS

This section aims to highlight the ethical and legal issues and some suggestions observed during the analysis and machine learning phase.

1: High rated drugs not present for some conditions

In question [2] of analysis, we have found that for some conditions, the highest rating for the drug lies between 1-4. When developing new medicines, it is good to focus on these conditions first in order to provide good alternatives for the customers since they don't have any such option at the moment.

2: People's opinion on medicine with extremely severe side effects

From question [4] of analysis, we can see that people still have a bad opinion of medicines with extremely severe side effects (peak at rating 1) but highly effective. Thus we can see that even though the medicine in that category works well for the particular condition, people still don't want to take the risk of creating new problems.

When we create a graph for medicines with severe side effects but are highly effective, we get the following graph:

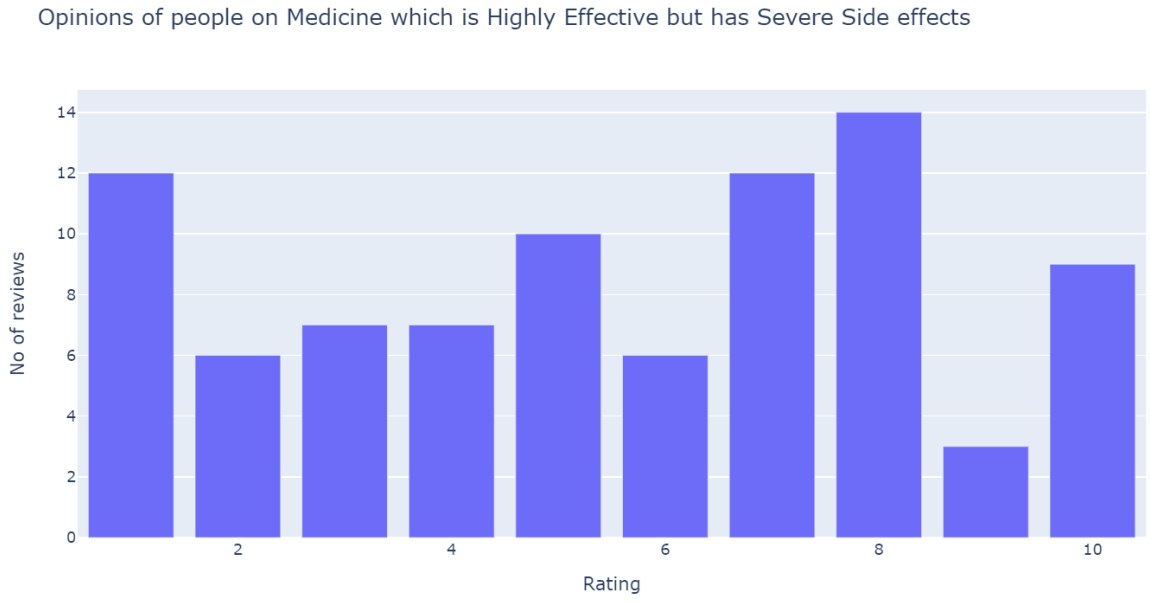


Fig: Opinion distribution of people on medicines that are Highly effective but have Severe side effects

We can see that this has much more varied opinions and a slight peak at rating 8 instead of a pronounced peak at rating 1. Thus improving the side effects by even one category might have a good impact on people's opinions on the medicine.

3: Medicines which are ineffective and have extremely severe side effects

We have studied the medicines which are ineffective but have extremely severe side effects in question [5] of analysis. People have a very negative outlook on such medicines with rating-1 having most review and some reviews in rating 2 and 3. We also derived the alternative medicines for such conditions and their reviews had a slight peak at 10.

This being the case, the question arises of why such medicines which are ineffective but have extremely severe side effects are still in circulation. There should be additional analysis done taking into consideration the cost, method of production and availability. If there is still no reason found as to why these medicines are being used, it is suggested to pull these medicines from circulation and use the better alternate options instead. This will ensure that patients do not have to suffer side effects when there is a much better option available.

If the cost analysis shows that these medicines are cheaper and therefore more widely used, steps should be taken in order to reduce production costs or resolve issues related to bringing down the costs of the more expensive medicines as everyone deserves access to good health care.

4: Some conditions treated by drugs that are highly effective but have severe side effects

When we take a count of the conditions treated by the medicines that are highly effective but have extremely severe side effects, we get the following:

```
birth control                                    2
sore throat                                      1
contraception                                    1
crohns disease, psoriatic arthritis              1
seizures                                         1
acne                                             1
grand mal seizures                               1
skin wound/infection                             1
add                                              1
adhd                                             1
weightloss                                       1
cornea transplant rejection                      1
i kept getting pregnant. no more stairs.... please    1
hysterectomy                                     1
u/i                                              1
acid reflux, gerd                                1
cystic acne                                      1
controceptive; help with pmsd                    1
headaches                                        1
prostatitis                                      1
high blood pressure                              1
tooth infection                                  1
puncture wound                                   1
acute sinusitis/bronchitis                       1
tooth abscess                                    1
parkinson's                                      1
Name: Condition, dtype: int64
```

Fig: Conditions that are treated by the medicines that are highly effective but have extremely severe side effects.

In this diagram, we can see stuff like acid reflux, weight loss and acne that are not serious and that sometimes go away on their own. Here, we can argue whether it is necessary to use medicine that can produce more serious side effects in order to treat something that usually isn't much of an issue.

Analysis must be done to weed out the medicines where the side effects are more severe than the condition the medicine is treating. Such drugs must not be used on people.

5: Results of the correlational analysis

From the correlational analysis we have discovered correlations between effectiveness and rating, side effects and rating and effectiveness and side effects.

Pharmaceutical companies must keep the first two correlations in mind when designing the medicines.

For the correlation between effectiveness and side effects, it must be analysed more in depth. Causal relationship and the direction of correlation should be explored more. Since both these factors influence the ratings, if we find the relation between the two, we can establish a trade-off with the optimal values for both in order to produce a drug that will be liked by the consumers.

6: The importance of effectiveness vs. side-effects

While checking the feature importance while doing classification models, it is found that effectiveness has a higher impact. Also, the gain ratio of effectiveness is much more when compared to side effects. Thus, from this, we can conclude that people first look for effectiveness in their medicines and are a bit lenient about the side effects.

From the ICE performed during classification, we also note that as the side-effects become worse or as the effectiveness decreases, the chance of getting a positive review also decreases.

7: Association rules

The rules obtained from association are:

- If the effectiveness is in the first two categories (Highly effective, Considerably effective) then the sentiment is positive.
- If the side effects are in the first two categories (No side effects, Mild side effects) and the effectiveness is in the first two categories (Highly effective, Considerably effective) then the sentiment is positive.
- If the effectiveness is in the first two categories (Highly effective, Considerably effective) and the side effects are in the last 3 categories (Moderate side effects, Severe side effects, Extremely severe side effects), the sentiment is positive.

These further cements the importance of effectiveness and side effects.

Right now, we have taken only two factors into consideration, but it will be nice to have more data related to drug costs, patient age, patient gender, other history, severity of condition, etc. This will help to derive more complex models that are closer to the real world.

The above rules may be taken as a very simplified version of the scenario.

8: VADER scores

Including the sentiment scores from VADER did not produce much changes to the rules. Thus, it would be good to find a way to establish causality between the effectiveness/side effects and the sentiment. Then it could explain why there is no significant difference to the rules, because then the sentiment becomes a field that is derived leading to us including a collinear column when including sentiment scores.

9: Side effects and sentiments

Understanding the relationship between side effects and sentiments is an interesting area that can be explored further with more data. This should also include details about the consumers like age, gender, etc as it will influence the conditions they can have and provide some data about their general tolerance to pain, etc.

Some areas that can be explored are:

- What side effects are considered to be unacceptable
- Do some medicines have permanent side effects
- Does tolerance increase/decrease with the age of the consumer
- Do some subcategory of patients have a higher tolerance to side effects
- Which medicines have side effects worse than the condition itself
- Does the condition that the patient have influence their general attitude in the review
- Does lower cost ensure a higher tolerance

Answering these questions will help us find which medicines are in dire need of alternatives and would help us find a baseline for customer tone in reviews.

10: Cost

Having the cost field would be very useful because it would help us analyse whether customers are more lenient towards medicines with a lower cost. This raises the ethical issue of whether bad medicines with low costs should still be available. This can be answered by gauging people's reactions about such medicines.

Section summary:

Thus, some issues and questions that arises after the analysis are discussed and some areas for further research and solutions are pointed out. We have also included suggestions for some fields to be included in order to get a representation that is more closer to the real world scenario.

CONCLUSION:

Thus, the pharmaceutical data is cleaned, analysed and various ML models are run on it. Based on the insights from the analysis and the explanation from ML some action plans have been suggested to keep in mind when creating new drugs and when deciding which drugs are too harmful to be used.

REFERENCES:

[1] Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. *Proceedings of the 2018 International Conference on Digital Health*. https://doi.org/10.1145/3194658.3194677

[2] Shiju, A., & He, Z. (2022, June 1). Classifying Drug Ratings Using User Reviews with Transformer-Based Language Models. https://doi.org/10.1109/ICHI54592.2022.00035

[3] Ramya, S. P., Sumitha, B., Ranjani, R., & Ahamed, M. A. (2022, August 1). A Comparative Study on Aspects Level Drug Reviews using Back Propagation Neural Networks. https://doi.org/10.1109/ICESC54411.2022.9885360

[4] Imani, M., & Noferesti, S. (2022). Aspect extraction and classification for sentiment analysis in drug reviews. *Journal of Intelligent Information Systems*, *59*(3), 613–633. https://doi.org/10.1007/s10844-022-00712-w

[5] Joshi, S., & Abdelfattah, E. (2021, May 1). Multi-Class Text Classification Using Machine Learning Models for Online Drug Reviews. https://doi.org/10.1109/AIIoT52608.2021.9454250

[6] J, A. K., Cambria, E., & Trueman, T. E. (2021). DUSE: A New Benchmark Dataset for Drug User Sentiment Extraction. *2021 International Conference on Data Mining Workshops (ICDMW)*. https://doi.org/10.1109/icdmw53433.2021.00028

[7] Yadav, A., & Vishwakarma, D. K. (2020, September 1). A Weighted Text Representation framework for Sentiment Analysis of Medical Drug Reviews. https://doi.org/10.1109/BigMM50055.2020.00057

[8] Garg, S. (2021). Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. https://doi.org/10.1109/confluence51648.2021.9377188

[9] Min, Z. (2019). Drugs Reviews Sentiment Analysis using Weakly Supervised Model. *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. https://doi.org/10.1109/icaica.2019.8873466

[10] Han, Y., Liu, M., & Jing, W. (2020). Aspect-Level Drug Reviews Sentiment Analysis Based on Double BiGRU and Knowledge Transfer. *IEEE Access*, *8*, 21314–21325. https://doi.org/10.1109/ACCESS.2020.2969473

[11] Jayale, R. S., & Desai, S. (2021, September 1). Aspect-Level Drug Reviews Sentiment Analysis and COVID-19 Drug prediction using PPI & Deep Learning. https://doi.org/10.1109/CCGE50943.2021.9776369

[12] Cavalcanti, D., & Prudêncio, R. (2017). Aspect-Based Opinion Mining in Drug Reviews. *Progress in Artificial Intelligence*, 815–827. https://doi.org/10.1007/978-3-319-65340-2_66

[13] Sampathkumar, H., Chen, X., & Luo, B. (2014). Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model. *BMC Medical Informatics and Decision Making*, *14*(1). https://doi.org/10.1186/1472-6947-14-91

[14] Gopalakrishnan, V., & Ramaswamy, C. (2017). Patient opinion mining to analyze drugs satisfaction using supervised learning. *Journal of Applied Research and Technology*, *15*(4), 311–319. https://doi.org/10.1016/j.jart.2017.02.005

[15] Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., & Gonzalez, G. H. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, *62*, 148–158. https://doi.org/10.1016/j.jbi.2016.06.007

[16] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/efficacy. (2011, February 2). Retrieved January 7, 2023, from www.cancer.gov website: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/efficacy

[17] Ljubljana, B. L., University of. (n.d.). Explaining Predictive Models. Retrieved January 7, 2023, from orangedatamining.com website: https://orangedatamining.com/blog/2021/2021-02-10-explaining-models/

[18] What are Association Rules in Data Mining (Association Rule Mining)? (n.d.). Retrieved January 7, 2023, from Business Analytics website: https://www.techtarget.com/searchbusinessanalytics/definition/association-rules-in-data-mining#:~:text=The%20strength%20of%20a%20given

APPENDIX

[1] Python code screenshots included in a separate PDF.