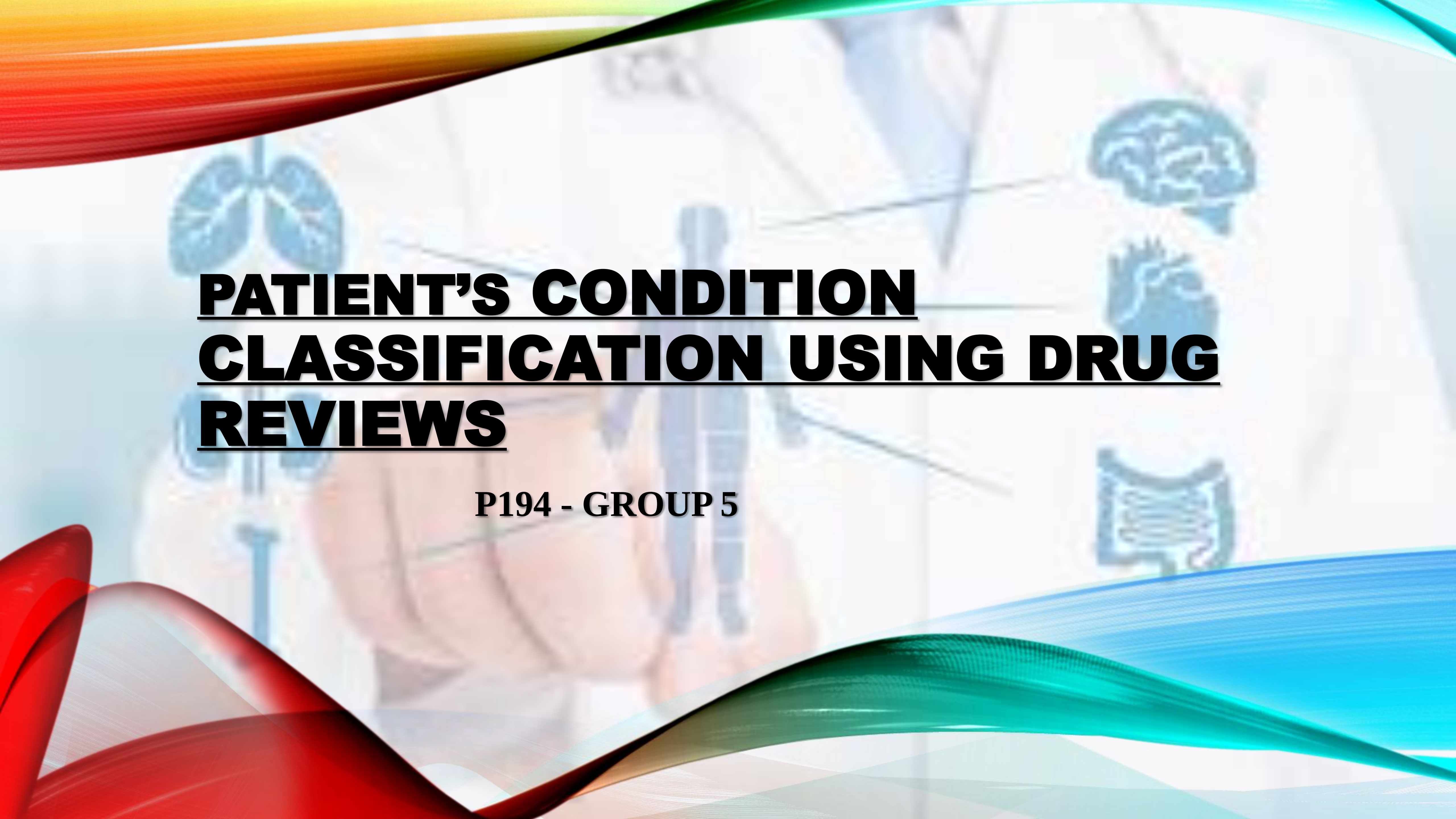


PATIENT’S CONDITION CLASSIFICATION USING DRUG REVIEWS

P194 - GROUP 5

GROUP MEMBERS

Aji Thomas

Shubham Ajitkumar Sutar

Manohar Kumar Sinha

Jagadeesh Korukonda

Manali Vijay Mane

Sachin Biradar

Margam Navya

Mentor : Mr. R P Adhvaith

CONTENT

- Abstract
- Introduction
  - Objective
  - Project Architecture
  - Challenges Faced
- Data Collection
- EDA
- Data Visualization
- Feature Engineering/Data Pre-processing
- Model Building
- Model Deployment
- Result

ABSTRACT

Natural Language Processing (NLP) is a subfield of machine learning that makes it possible for computers to understand, analyse, manipulate and generate human language.

NLTK is a popular open-source suite of Python libraries. Rather than building all of your NLP tools from scratch, NLTK provides all common NLP tasks so you can jump right in.

The two key techniques in Text Summarization are extraction and abstraction.

Extraction is a process that assesses large amounts of textual data to 'extract' short and definitive summaries.

Abstraction programs create summaries by creating new text based on the assessment of the original source text.

INTRODUCTION

Objective

- This is a sample dataset which consists of 161297 drug name, condition reviews and ratings from different patients.
- Our goal is to examine how patients are feeling using the drugs their positive and negative experiences so that we can recommend him a suitable drug.
- By analysing the reviews, we can understand the drug effectiveness and its side effects.

Project Architecture

- Data Collection
- EDA and Data Cleaning
- Data Visualization
- Model building
- Model Deployment

Challenges Faced

- Limited Time
- We solved the issues raised during this period by sharing work load equally among us.

DATA COLLECTION

THE DATASET PROVIDES PATIENT REVIEWS ON SPECIFIC DRUGS ALONG WITH RELATED CONDITIONS

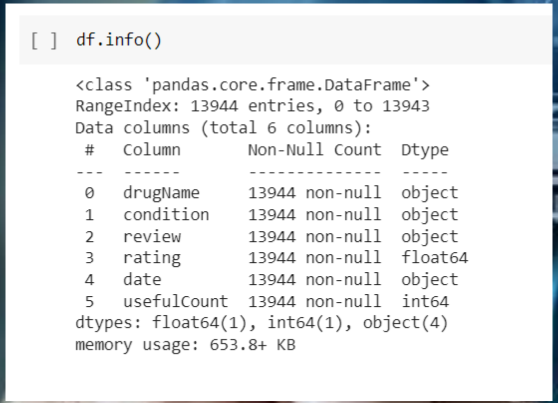
In this dataset, we can see many patients conditions but we have focused only on the below conditions from the patients reviews

a. Depression

c. High Blood Pressure

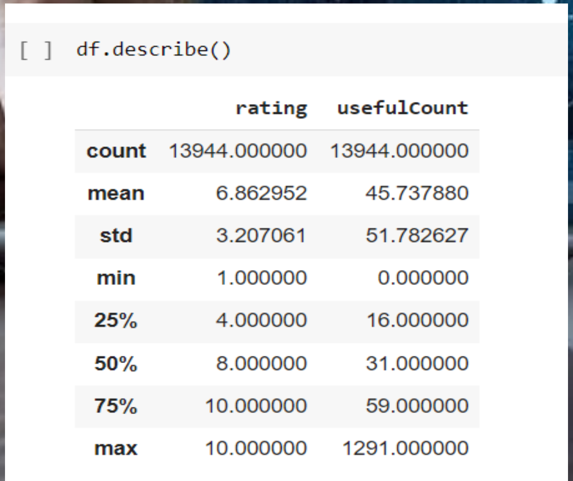
d. Diabetes, Type 2

EDA



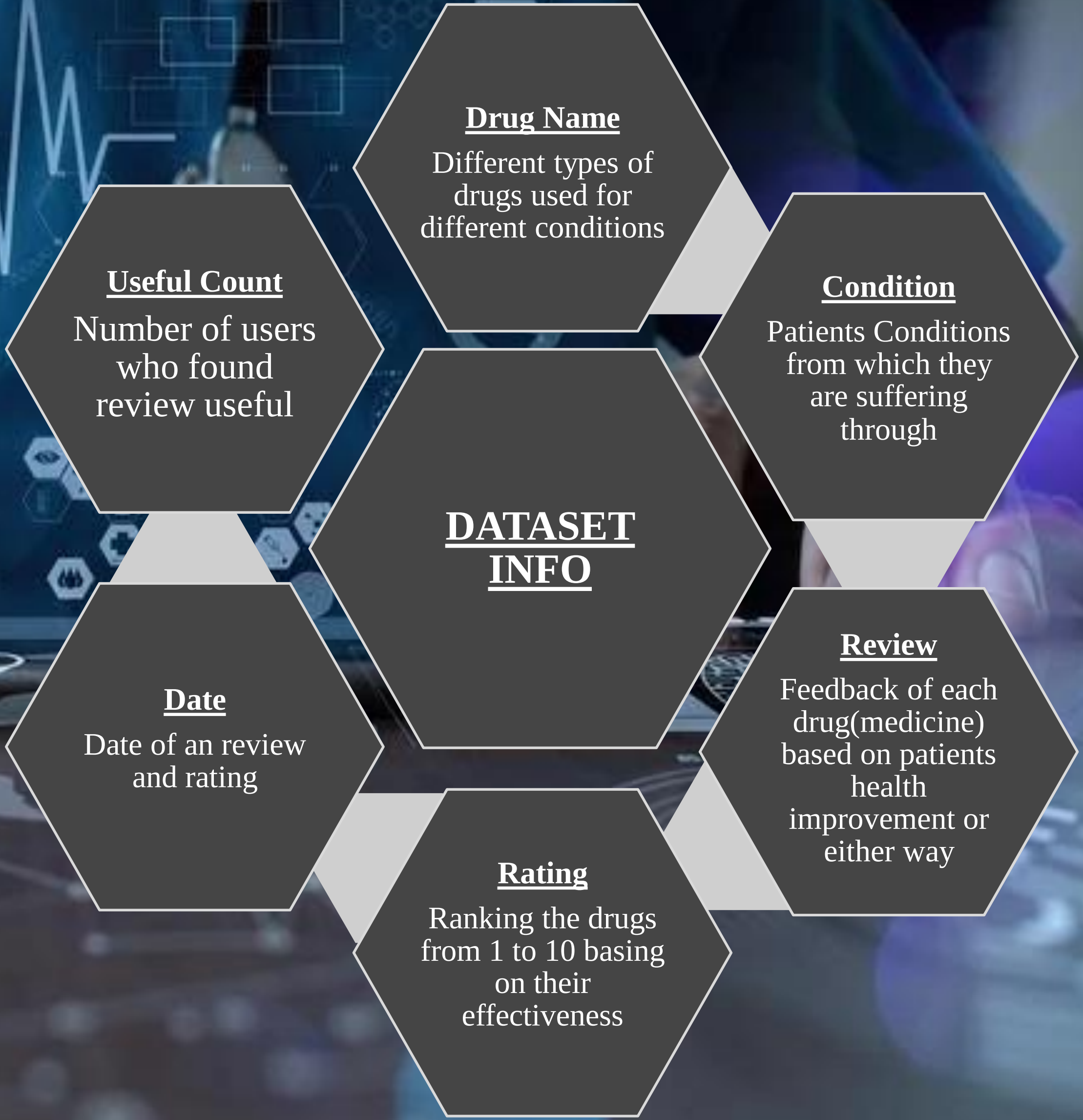
```
[ ]  df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 13944 entries, 0 to 13943
     Data columns (total 6 columns):
      #   Column       Non-Null Count  Dtype
     ---  ------       --------------  -----
      0   drugName     13944 non-null  object
      1   condition    13944 non-null  object
      2   review       13944 non-null  object
      3   rating       13944 non-null  float64
      4   date         13944 non-null  object
      5   usefulCount  13944 non-null  int64
     dtypes: float64(1), int64(1), object(4)
     memory usage: 653.8+ KB
```



```
[ ]  df.describe()
```

	rating	usefulCount
count	13944.000000	13944.000000
mean	6.862952	45.737880
std	3.207061	51.782627
min	1.000000	0.000000
25%	4.000000	16.000000
50%	8.000000	31.000000
75%	10.000000	59.000000
max	10.000000	1291.000000



DATASET INFO

Drug Name
Different types of drugs used for different conditions

Condition
Patients Conditions from which they are suffering through

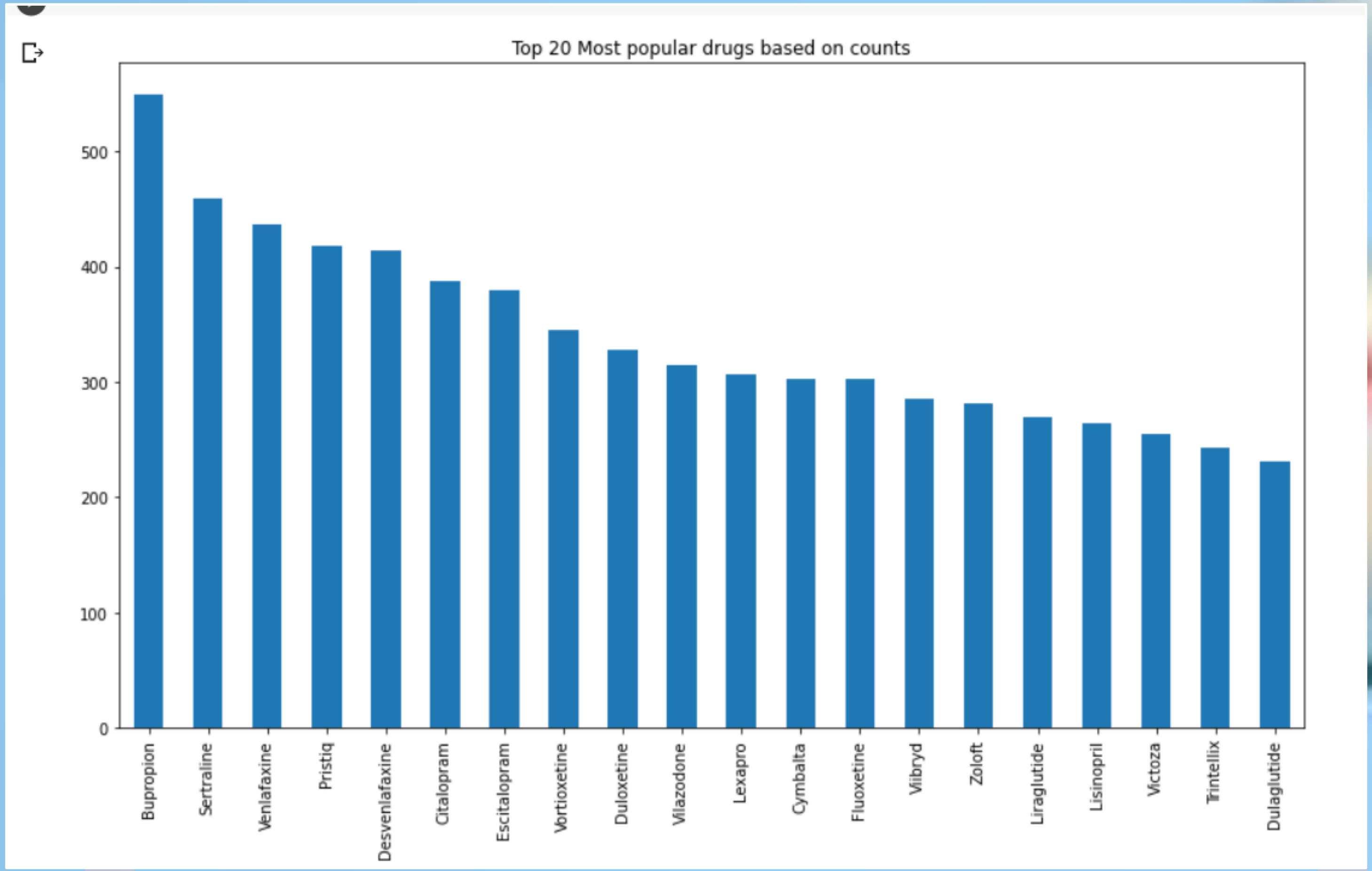
Review
Feedback of each drug(medicine) based on patients health improvement or either way

Rating
Ranking the drugs from 1 to 10 basing on their effectiveness

Date
Date of an review and rating

Useful Count
Number of users who found review useful

DATA VISUALIZATION

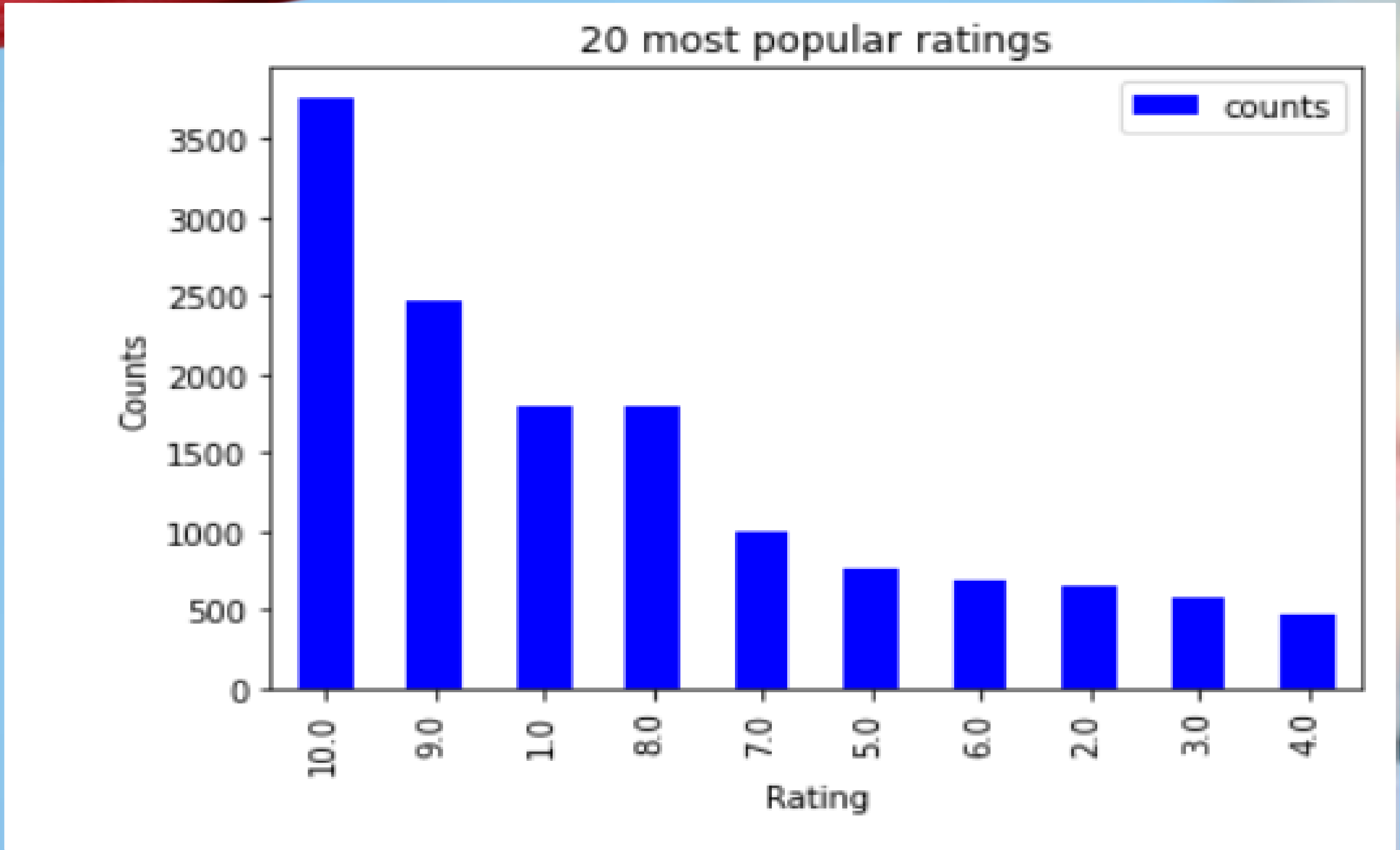


Bar Plot : Top 20 most popular drug based on counts

A bar plot shows the relationship between a numeric and a categoric variable.

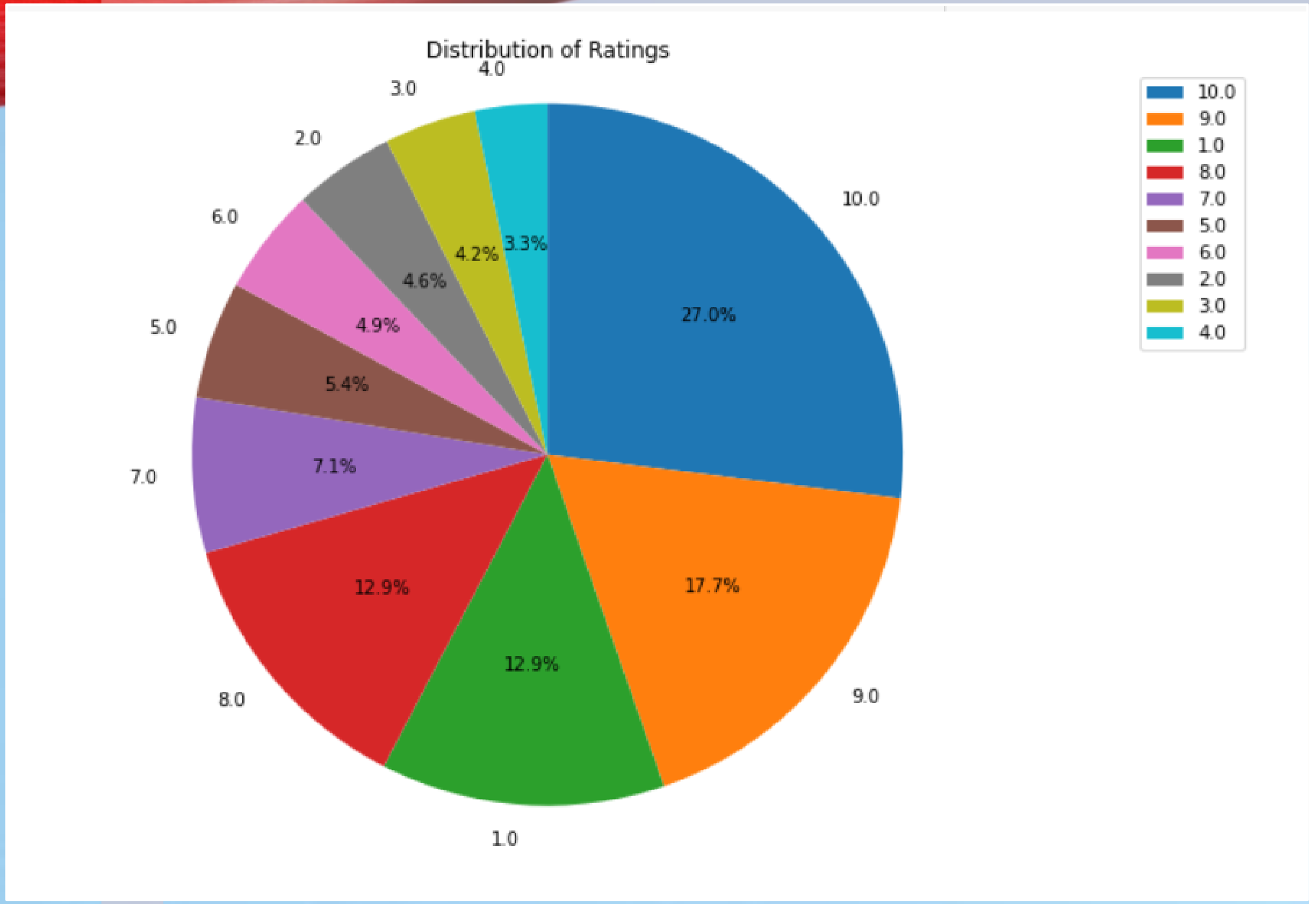
The bar plot illustrated the top 20 most popular drugs based on the counts.

And the result follows Bupropion, sertraline and venlafaxine are mostly used drugs based on useful count.



Bar Plot : Top 20 most popular ratings

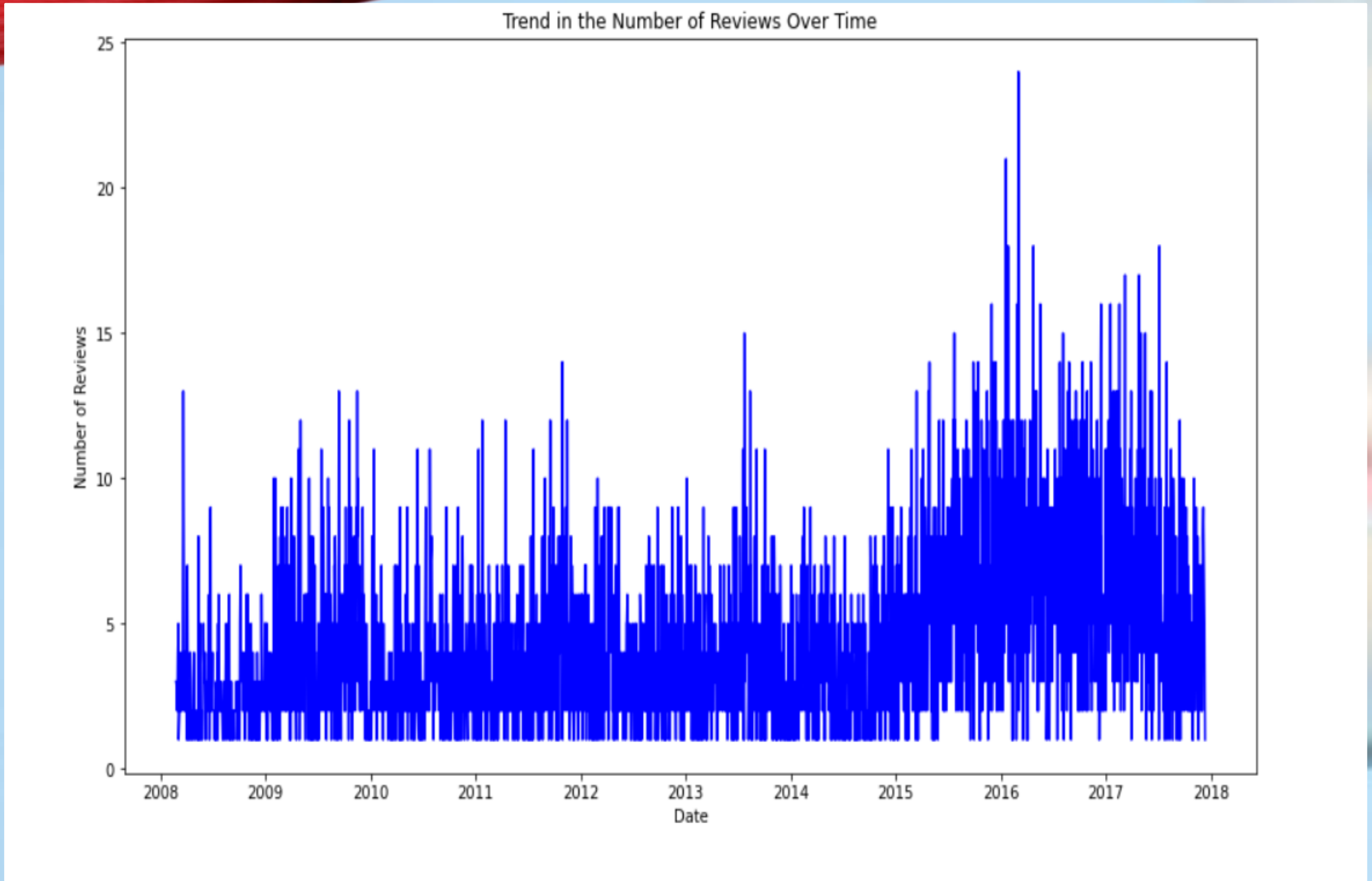
The following illustrated plot shows the top 20 most popular ratings basing on counts and the output states most rated one are 10, 9,1 and 8.



Pie Chart : Distribution of Ratings

Pie Chart is a circular statistical graphic, which divides into slices to illustrate numerical proportion.

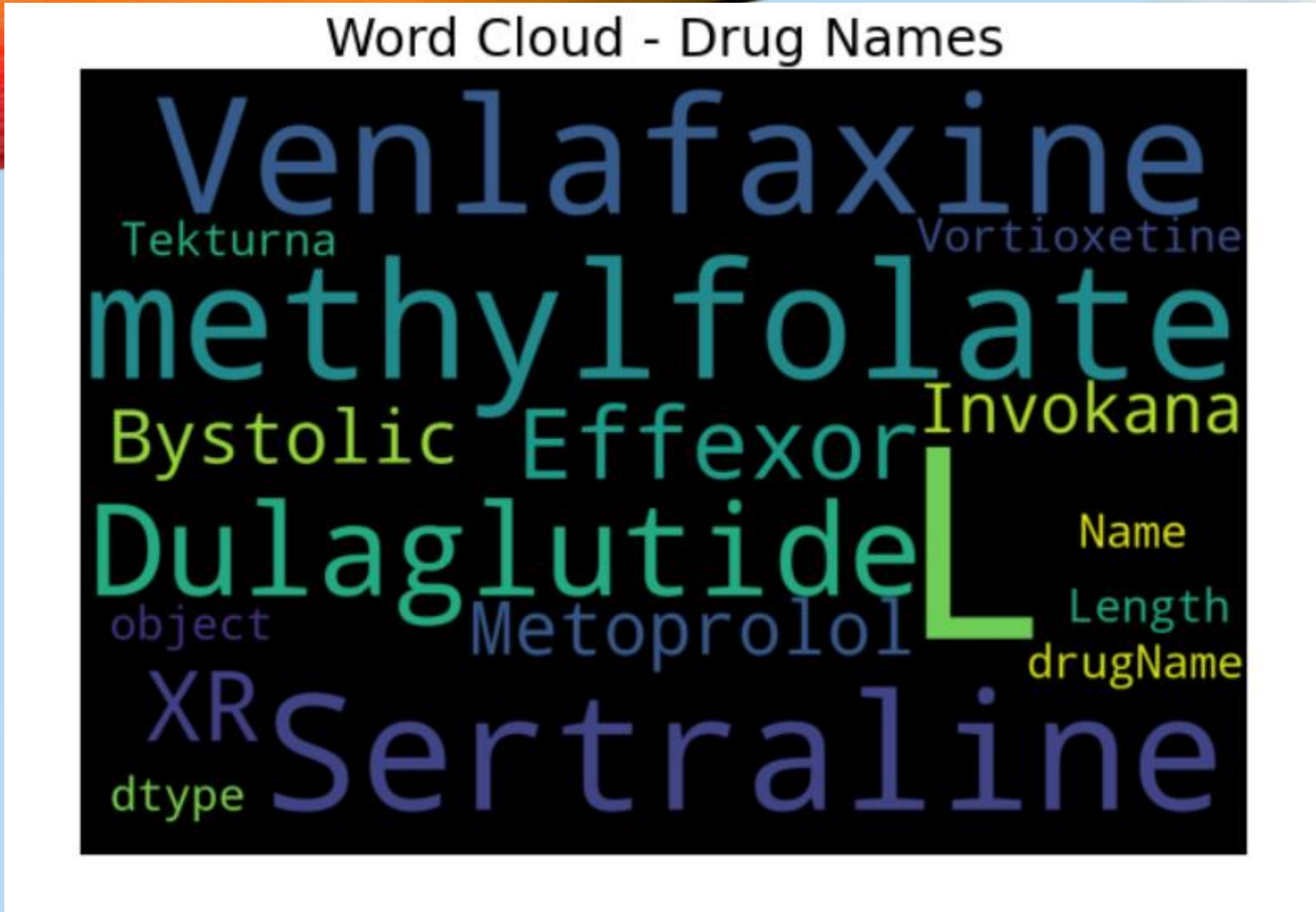
The chat define that distribution of ratings over dividing into percentage.



LINE CHART : Trend in the number of reviews over time

Line chart is a simple, two-dimensional chart with an X and Y axis, each point representing a single value.

The chart represents about trend in reviews over time by the patients.



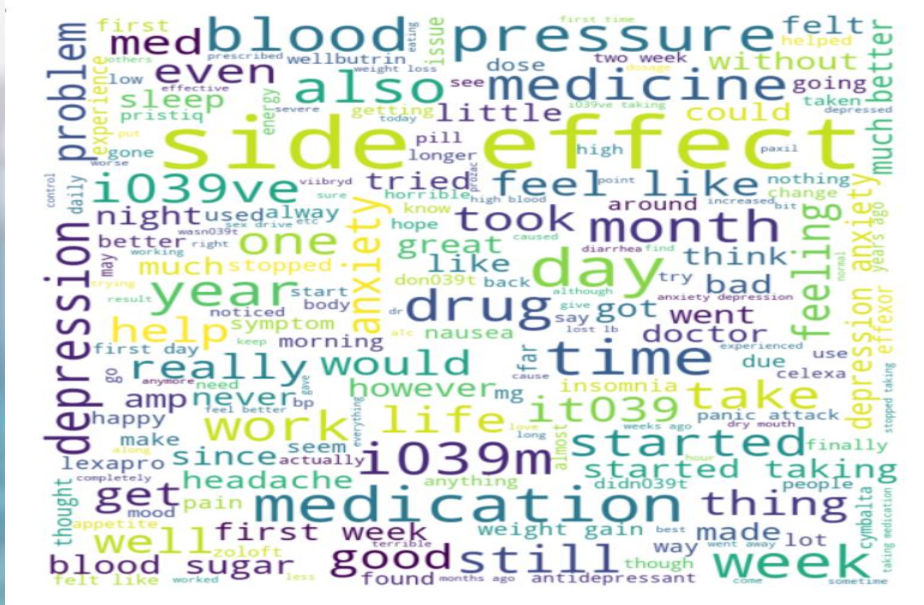
Word Cloud : Drug Names

Word cloud is basically a visualization technique to represent the frequency of words in a text where the size of the word represents its frequency.

And the picture represent about Drug name that are mostly commonly used.

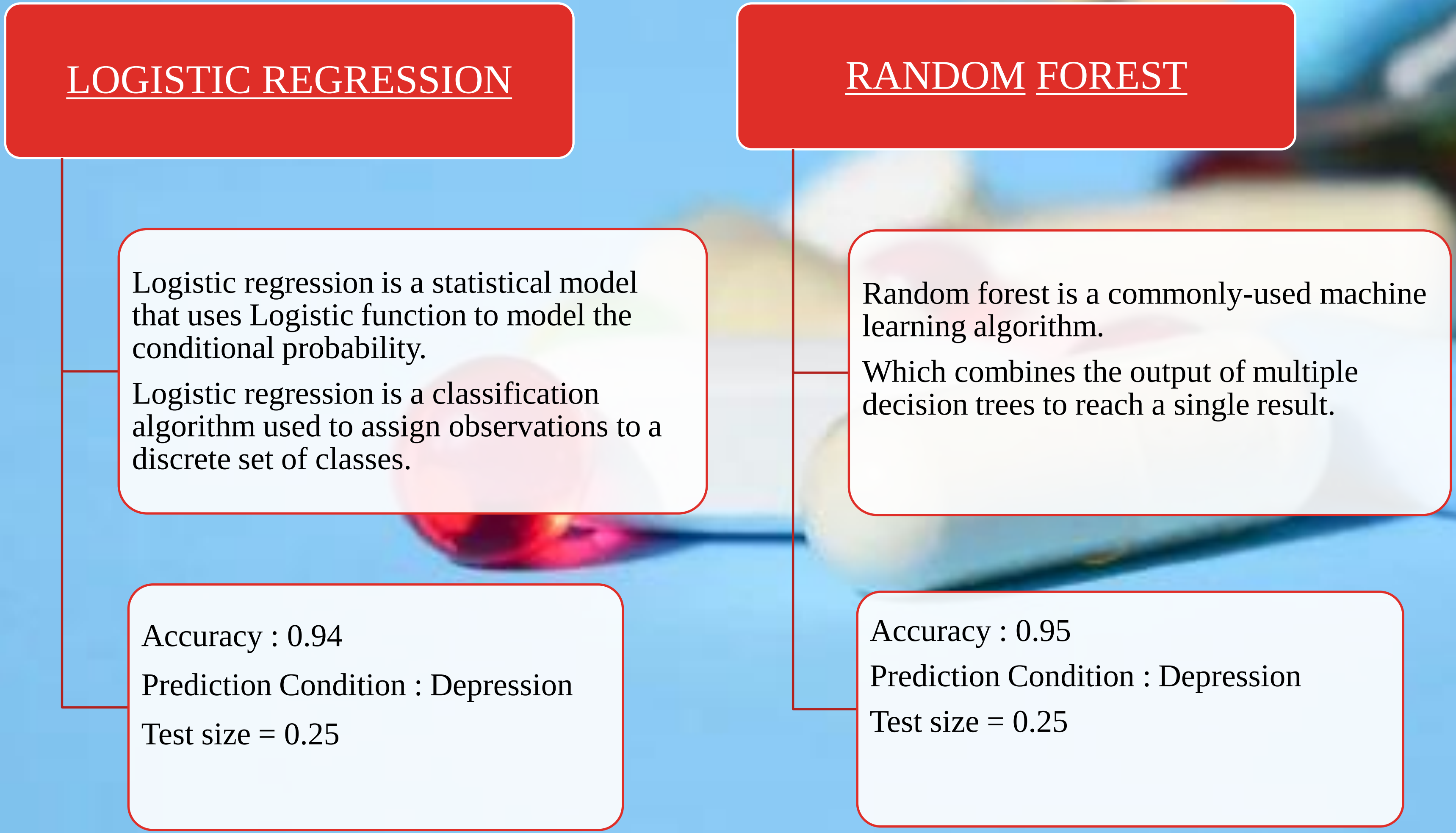
FEATURE ENGINEERING/DATA PRE-PROCESSING

- Converting into lowercase
- Remove punctuation
- Tokenize the review
- Remove stop words
- Join the cleaned tokens back together



Word cloud of reviews

MODEL BUILDING



LOGISTIC REGRESSION

Logistic regression is a statistical model that uses Logistic function to model the conditional probability.

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes.

Accuracy : 0.94

Prediction Condition : Depression

Test size = 0.25

RANDOM FOREST

Random forest is a commonly-used machine learning algorithm.

Which combines the output of multiple decision trees to reach a single result.

Accuracy : 0.95

Prediction Condition : Depression

Test size = 0.25

SUPPORT VECTOR MACHINE

SVM is a supervised machine learning algorithm used for both classification and regression.

The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

It is more effective in high dimensional spaces and is relatively memory efficient.

Accuracy : 0.96

Prediction Condition : Depression

Test size = 0.25

GRADIENT BOOSTING

Gradient boosting is a type of machine learning boosting.

It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error.

The key idea is to set the target outcomes for this next model in order to minimize the error.

Accuracy : 0.92

Prediction Condition :Depression

Test size = 0.25

DECISION TREE

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks.

It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes

Accuracy : 0.94

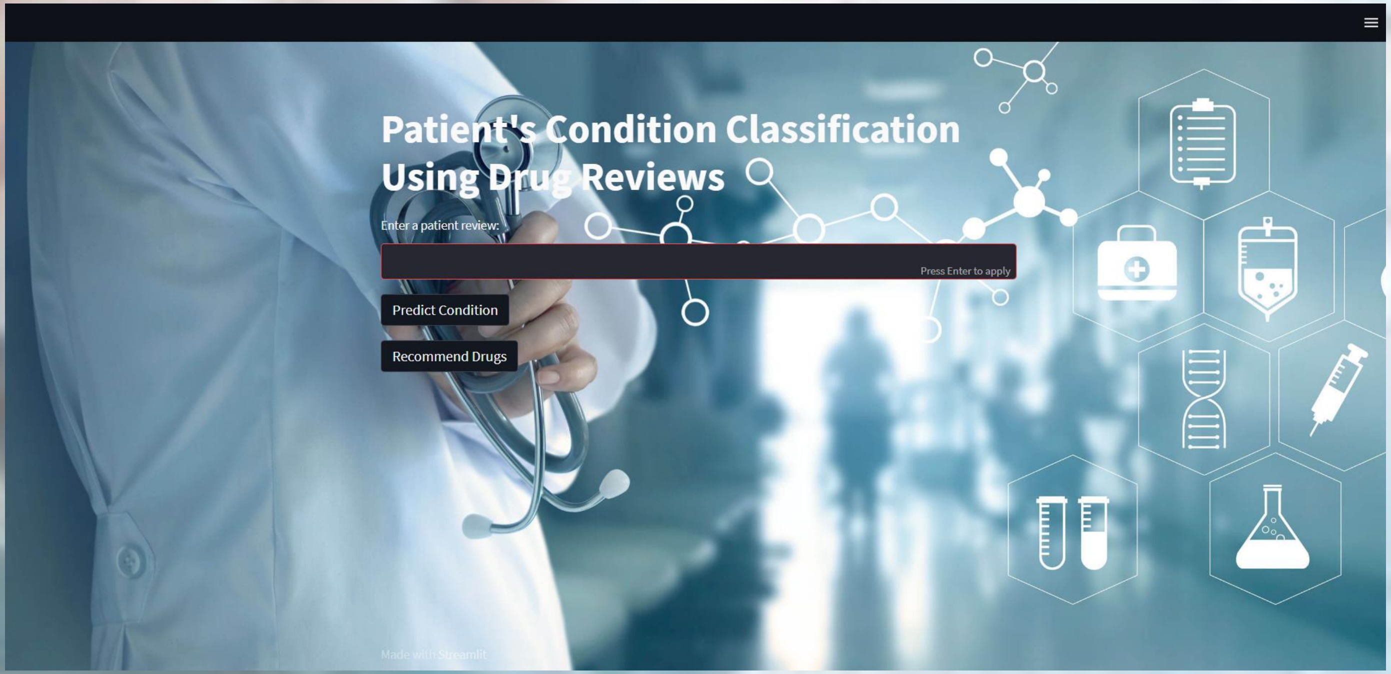
Prediction Condition : Depression

Test size = 0.25

MODEL DEPLOYMENT

Streamlit is an open-source app framework in python language. It helps us create beautiful web apps for data science and machine learning in a little time.

It is compatible with major python libraries such as scikit-learn, keras, PyTorch, latex, numpy, pandas, matplotlib, etc.

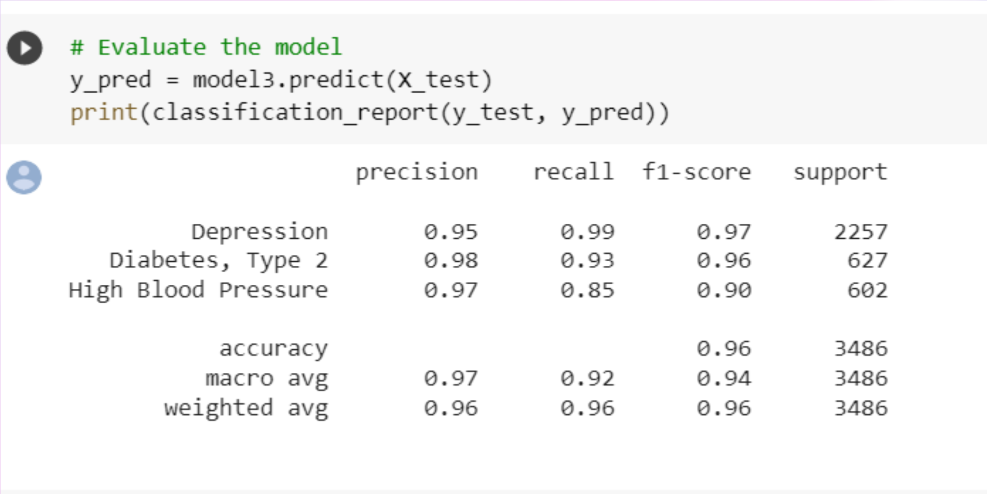


Patient's Condition Classification Using Drug Reviews

Enter a patient review:

RESULT

- From above all the algorithms we can say that the best result showing algorithm is SVM (support vector machine) with the highest accuracy of 0.96 among rest of the algorithms.

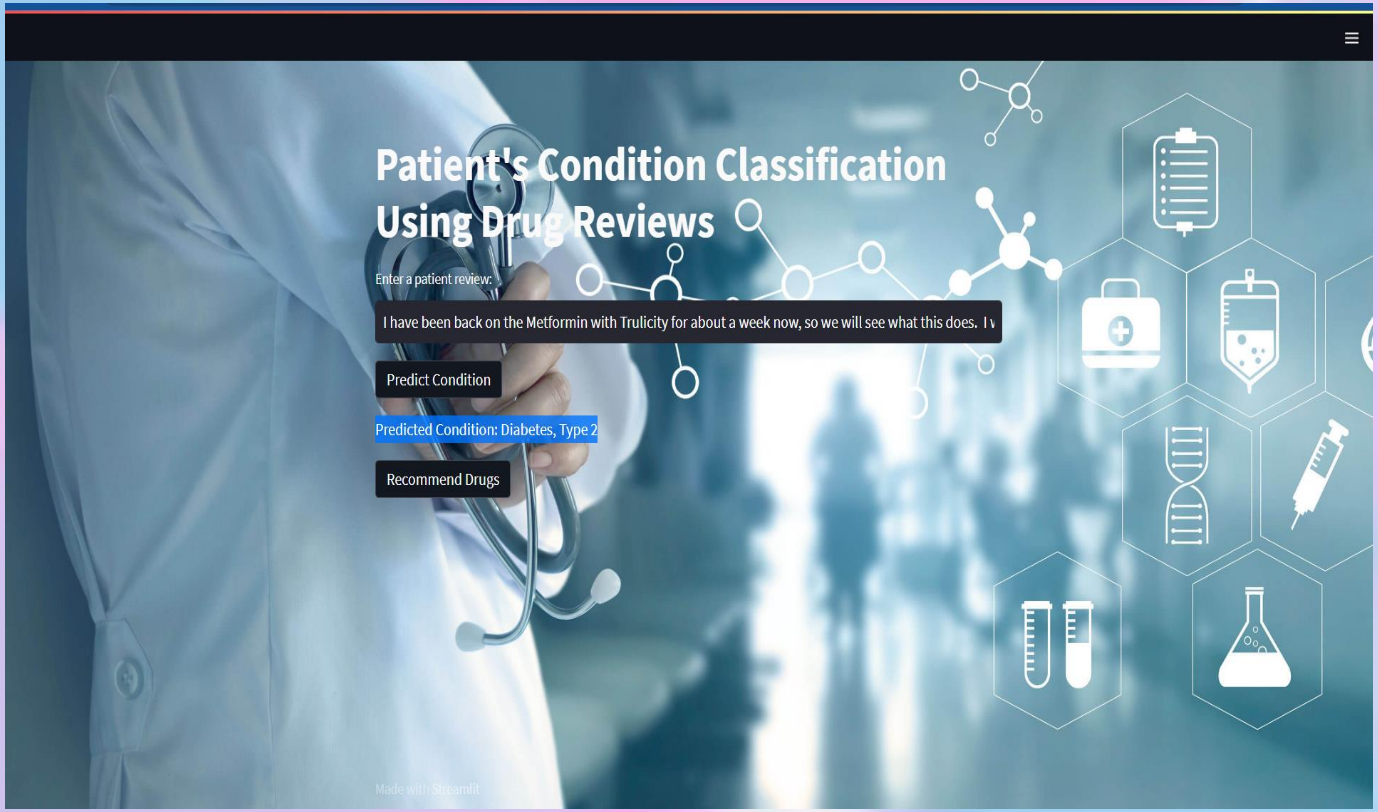


```
# Evaluate the model
y_pred = model3.predict(X_test)
print(classification_report(y_test, y_pred))
```

```
                     precision    recall  f1-score   support

         Depression       0.95      0.99      0.97      2257
   Diabetes, Type 2       0.98      0.93      0.96       627
High Blood Pressure       0.97      0.85      0.90       602

           accuracy                           0.96      3486
          macro avg       0.97      0.92      0.94      3486
       weighted avg       0.96      0.96      0.96      3486
```

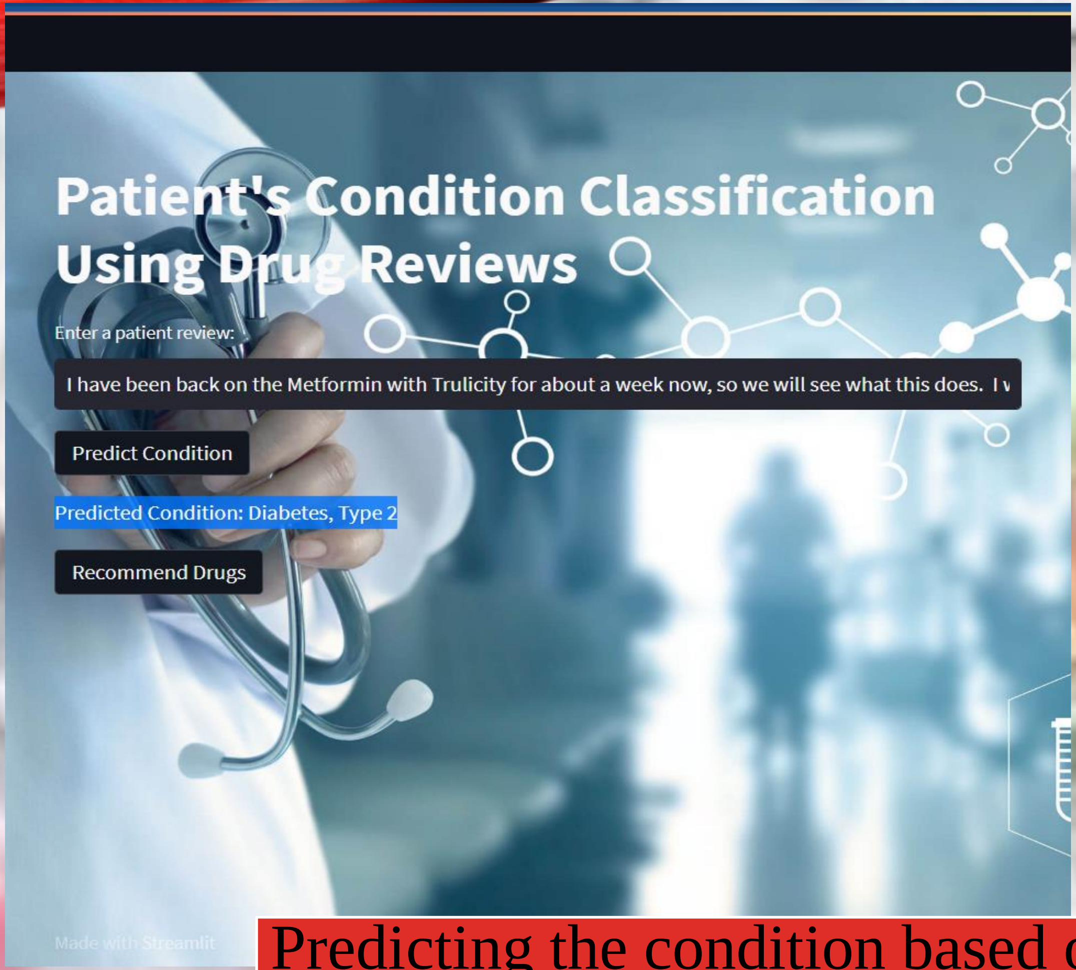


Patient's Condition Classification Using Drug Reviews

Enter a patient review:

I have been back on the Metformin with Trulicity for about a week now, so we will see what this does. I v

Predicted Condition: Diabetes, Type 2



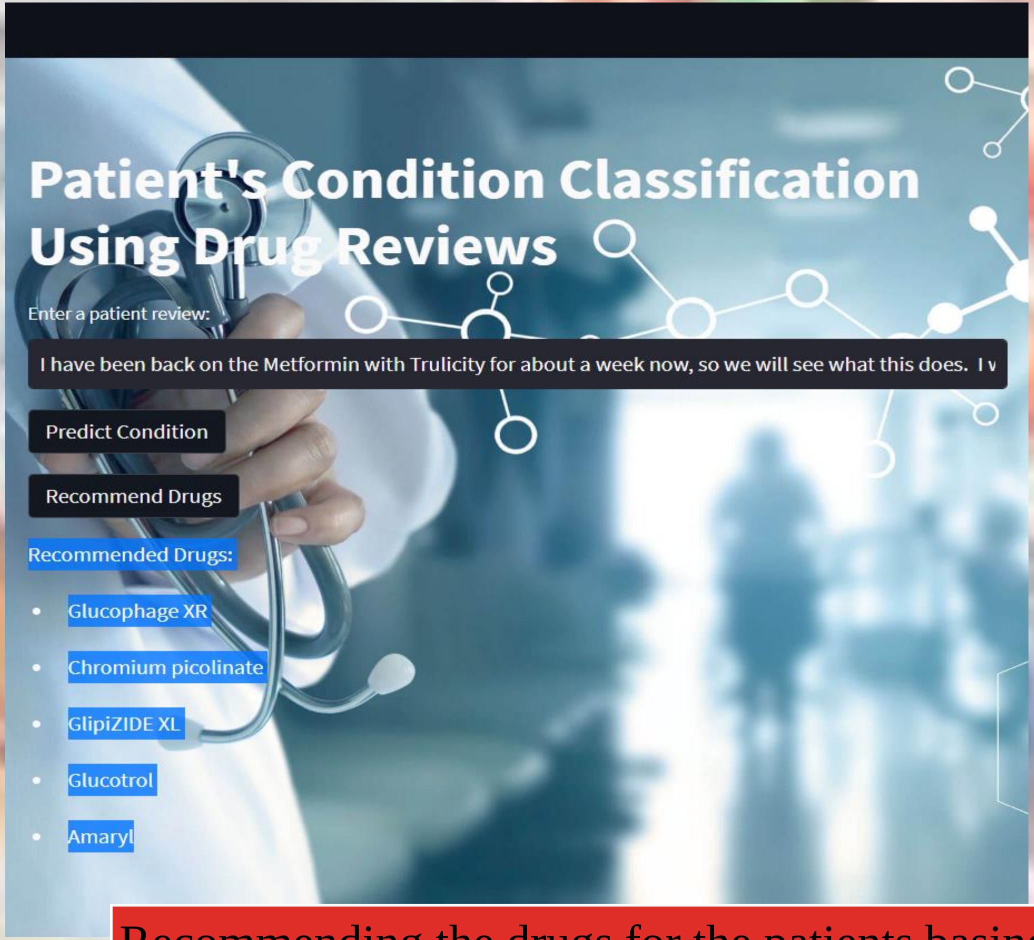
Patient's Condition Classification Using Drug Reviews

Enter a patient review:

I have been back on the Metformin with Trulicity for about a week now, so we will see what this does. I v

Predicted Condition: Diabetes, Type 2

Predicting the condition based on patients review



Patient's Condition Classification Using Drug Reviews

Enter a patient review:

I have been back on the Metformin with Trulicity for about a week now, so we will see what this does. I v

Recommended Drugs:

- Glucophage XR
- Chromium picolinate
- GlipiZIDE XL
- Glucotrol
- Amaryl

Recommending the drugs for the patients basing on patient review to predicted condition