|  |  |
| --- | --- |
| Internship Project Title | RIO-125: Classification Model-Build a Model that Classifies the Side Effects of Drugs |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Himalaya Aashish |
| Name of the Institute | ICT ACDEMY OF KERALA |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 23/02/2021 | 30/03/2021 | | 18 | | Jupyter notebook | Excel, jupyter |
| Milestone # | 1 | Milestone: | | Day5: Preprocessing steps   * Visualization:   Graph is using to check the drugname and condition of the patients.  Using this we have to analyse the condition of patients and usage of drugs | | |

**TABLE OF CONTENT :**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Charts, Table, Diagrams

**Acknowledgement**

The internship opportunity provided by TCSiON was a great chance for learning and professional development. I express my deepest thanks to the industry mentor for taking part in useful decision & giving necessary advices and guidance, and also express my deepest thanks to all faculty member of ICT Academy of Kerala for their careful and precious guidance which were extremely valuable for my learning.

I consider this as the biggest opportunity that has improve my career development.

**Objective**

Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. Drugs can help to treat diseases, but usually come with side effects or adverse reactions. Because of unintended side effects, a great number of approved drugs were even withdrawn from the market. Therefore, recognizing potential side effects helps to reduce costs and avoid risks in the drug discovery.

It could also be helpful for the patients who are buying drugs online to check the side effects of the drugs before buying it.

The main objective of this project to build a classification model that classifies the side effects of a particular drug by age, gender and race. The model need to have good amount of accuracy and have to meet the industry standards.

**Introduction / Description of Internship**

The project guidelines clearly mentioned that we are expected to create a model that classifies the trial data of a drug based on the patients review . At the end of the project we should be able to create a dataset, clean the dataset, sanitize it and preprocess the data to perform data partitioning and handle missing values. Create training and testing sets. Build a classifier and fit the data to the model.

**Internship Activities**

The internship activity is mainly concentrates on how we make up to the objective of the internship. The given resources were very useful to our internship and the day wise plan helps us to calculate the overall time and amount of work to be done each day.

**Approach / Methodology**

The Approach / Methodology used here will be the Linear Strategy which consist in sequential phases with no feedback loops. The project solution is not released until the final phase is reached. This strategy is characterized by clearly defined goal solution and requirements. The pre-defined steps includes data cleaning, data preprocessing , feature processing, splitting to test and train set, applying machine learning algorithms, comparison of machine learning algorithms.

**Assumptions**

By various Exploratory data analysis we can come an assumption that the drug are rated good for the body by chemist, it have a slight side effect of the dataset mainly for depression from the condition attribute mainly concentrate on acne, anxiety, insomnia, birth control, high blood pressure allergies and other mental problem related to brain issues.

**Charts:**

1. **Check the top-30 Drugs by count:**

top30\_drugs = data\_df.DrugName.value\_counts ()[:30]

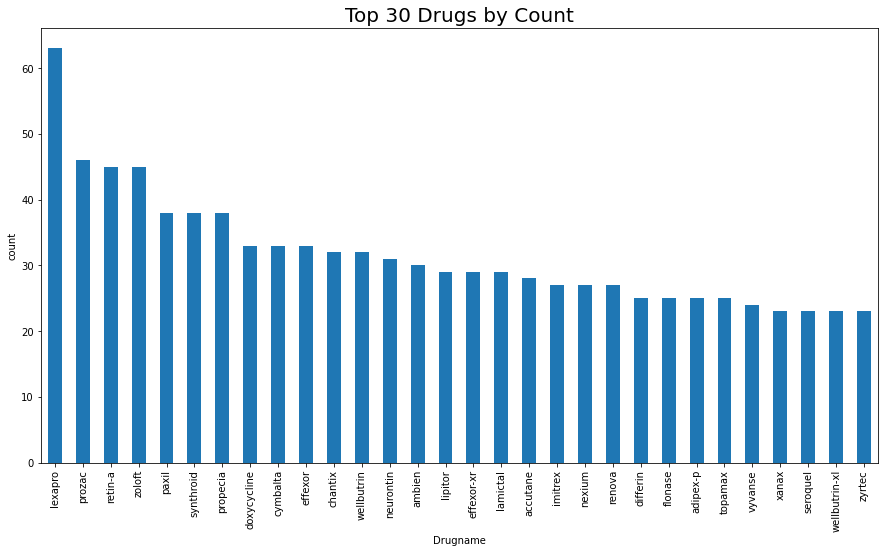
plt.figure(figsize = (15,8))

top30\_drugs.plot(kind = 'bar');

plt.title('Top 30 Drugs by Count',fontsize = 20);

plt.xlabel("Drugname")

plt.ylabel("Count")



* Here morethan 60% sed drug is Lexapro,its mainly used for depression and anxiety.

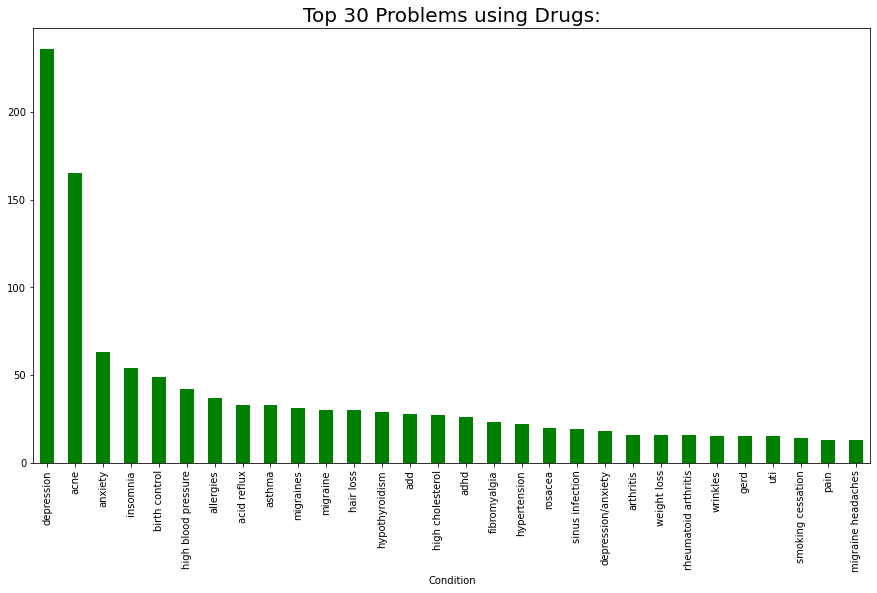
1. **Top-30 patients problems using Drugs by Condition:**

top30\_problems = data\_df.condition.value\_counts()[:30]

top30\_problems.plot(kind="bar", figsize = (15,8), fontsize = 10,color="green")

plt.title('Top 30 Problems of Patientsusing Drugs:',fontsize = 20)

plt.xlabel("Condition")



* From above graph we can identified that usage of drugs mainly effected by depression, depression is twice as big as anyone of the condition, and acne then also effected by anxiety, insomnia ,birth control, high blood pressure allergies etc.

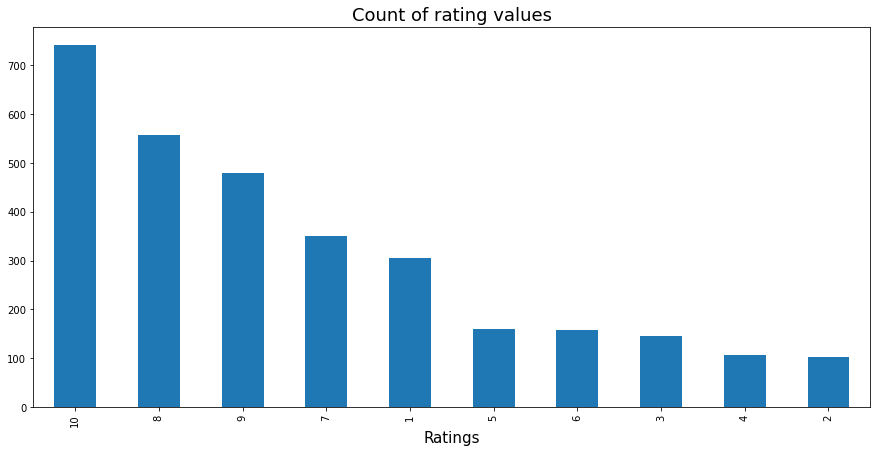
1. **Count the number of ratings**

plt.figure(figsize = (15,7))

data\_df.rating.value\_counts().plot(kind = 'bar')

plt.xlabel('Ratings',fontsize = 15)

plt.title('Count of rating values',fontsize = 18)



* Most people choose the values; 10, 8, 9,7,1 and the number of 10 is more than as many as the others.
* With this, we can see that the percentage of positives is higher than negative, and people's reactions are extreme.
* Here we'll consider ratings more than 5 as positive comments and less than or equal to 5 as negative comments.