Dear Intern

Interim project report is an inherent component of your internship. We are enclosing a reference table of content for the interim project report.

The key objective of this report is for you to capture how far you have got in completing the internship work against milestones expected to be achieved within a specific duration and seek the mentor’s feedback. Depending on the internship project and your progress (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections or leave some sections blank from the table of content mentioned below. You can also add additional sections. You can refer the project presentation to view the milestones related to your internship project. Please populate milestone# (1 / 2 / 3) and the milestone description in the interim project report based on the milestone for which you are submitting the interim project report.

You can refer the project presentation to view the milestones related to your internship project.

|  |  |
| --- | --- |
| Internship Project Title | RIO-125: Classification Model - Build a Model that Classifies the Side Effects of a Drug Batch 01 |
| Name of the Company | TCS- iON |
| Name of the Industry Mentor | Himalaya Aashish |
| Name of the Institute | ICT ACADEMY OF KERALA |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 24/02/2021 | 11/03/2021 | | 38 | | Jupyter Notebook | MS Excel, Jupyter |
| Milestone # | 2 | Milestone: | | Day15: Student should be able split the dataset into training and testing sets. They should also be able to build a classifier and fit the data to the model. | | |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses
* **ACKNOWLEDGEMENTS**

The internship opportunity I had with TCS - iON was a great chance for learning and professional development. I take this opportunity to express my profound gratitude and deep regards to Himalaya Ashish, Industry Mentor, for his exemplary guidance, monitoring and constant encouragement throughout the course of this project.

I am highly indebted to all the faculty members of ICT Academy of Kerala for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. Lastly, I thank almighty, my parents and friends for their constant encouragement without which this project would not be possible.

I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

* **OBJECTIVE**

Now a days massive data generated from the search engines has widened the perspective of the market research and analysis in the drug data. With the help of other parameters, we will predict whether a drug is safe or not. Side effects and Effectiveness of a particular drug need to be addressed.

The main given 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 needs 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 their age, gender and race. We also entrusted to create a dataset of 3,62,797 patients containing the following details for each patient based on various attributes according to the data. At the end of the project, we should be able to create a dataset, create useful visualizations, 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 activity mainly concentrates on how we make up to the objective of the internship. The given resources were very useful to kick start our internship and the day wise plan helps us to calculate the overall time and amount of work to be done each day and what extra we can do about it. We can explore different aspects of this data which vary from EDA to the final prediction model for the 30 days.

* **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, zero or few change requests of the scope, routine and repetitive process inside the project, use of pre-established formulas and templates. The pre-defined steps include data cleaning, EDA, PCA, data pre-processing, feature processing, splitting to test and train set, applying machine learning algorithms, comparison of machine learning algorithms and Opting the best prediction model.

* **ASSUMPTIONS**

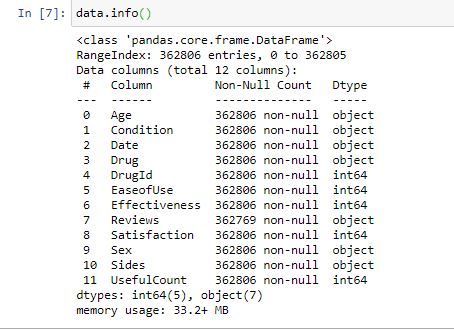
By various Exploratory data analysis we can come to an assumption that the drug are rated good for the body by chemist, it have a slight side effect of the dataset mainly for depression .The condition attribute mainly concentrate on Major depressive disorder , Chronic muscle or bone pain, Neuropathic pain, depression and other mental problem related to brain issues. So taking has condition which have a less correlated value with respected to other features so dropped it. The main issues was accuracy with respect to the output.

* **DATASET DESCRIPTION**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate, Text | **Number of Instances:** | 3,62,806 |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 12 |
| **Associated Tasks:** | Classification, Regression, Clustering | **Missing Values:** | N/A |

The dataset provides patient reviews on specific drugs along with related conditions. Additionally, ratings are available concerning overall satisfaction as well as a 5-step side effect rating and a 5-step effectiveness rating

* **ATTRIBUTE INFORMATION**



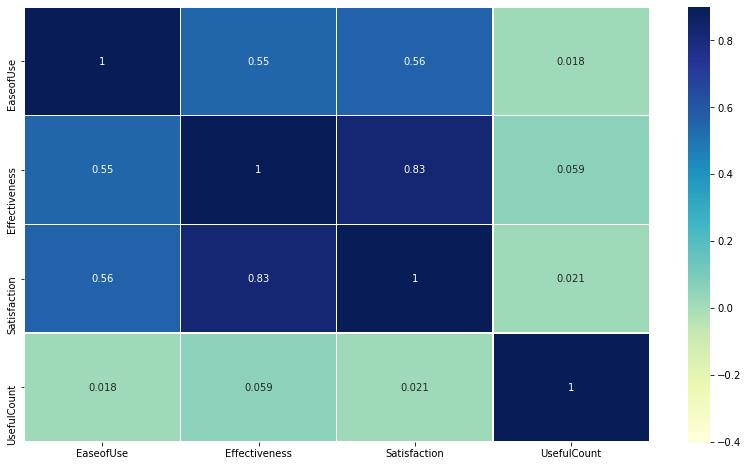
1. Age (Categorical) : Age Range
2. Condition (categorical): Condition of patient
3. Date (Categorical) : Date of patient report
4. Drug (categorical): Name of drug
5. Drug Id (categorical): Unique identification of a drug
6. Ease of Use (Numerical) : 5 step ease of use rating
7. Effectiveness (Numerical) : 5 step effectiveness rating
8. Reviews (Text) : Patient review on drug
9. Satisfaction(Numerical) : 5 step satisfaction rating
10. Sides (Text) : patient on side effects
11. Useful Count (Numerical) : Count performed
12. Sex (Categorical) : Sex of patient

After preprocessing, many of the attributes are dropped for better prediction.

* **EXCEPTIONS**

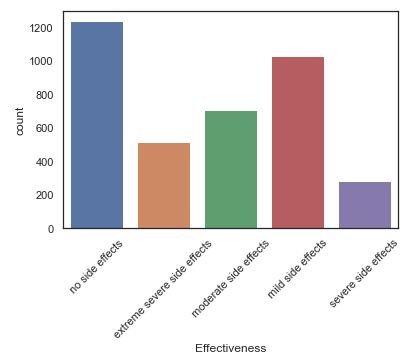
Afterwards, I have filtered the dataset with one drug to predict the side effect of that particular drug which occurs most using mode function. The drug thus obtained is **‘cymbalta’** with drug id **‘91491’**. For the sake of simplicity dataset is converted to one specific drug for better prediction which has an instance of **4648**.

* **CHARTS/TABLES/DIAGRAMS**
* Checked Correlation – Spotted a high correlation between ‘Effectiveness’ and ‘Satisfaction’.

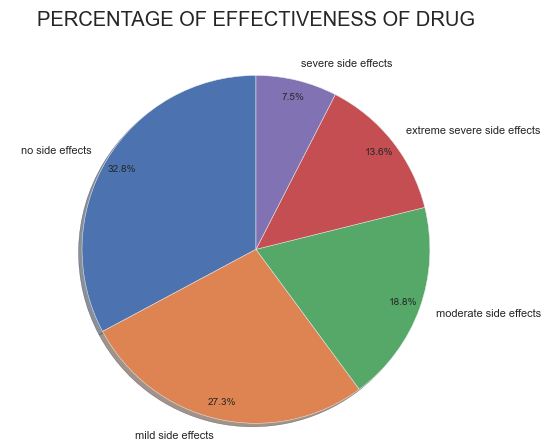


Correlation matrix shows the parameters of the dataset of different ranges than [-1, 1]. Where we will remove the low and high correlated feature from the dataset.

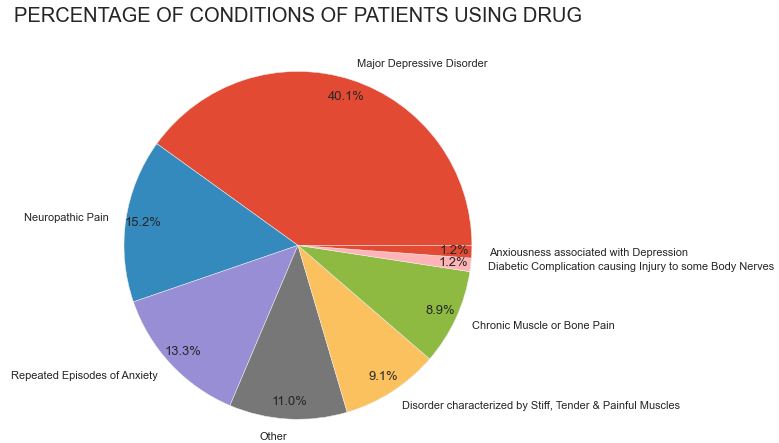
* Plotted a count plot on Effectiveness to analyze the count and found that ‘No side effects’ show maximum count.



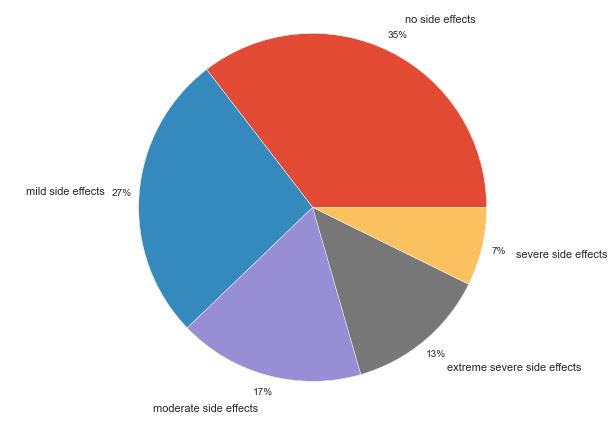
* Plotted a pie chart to analyze the percentage of effectiveness of drug.



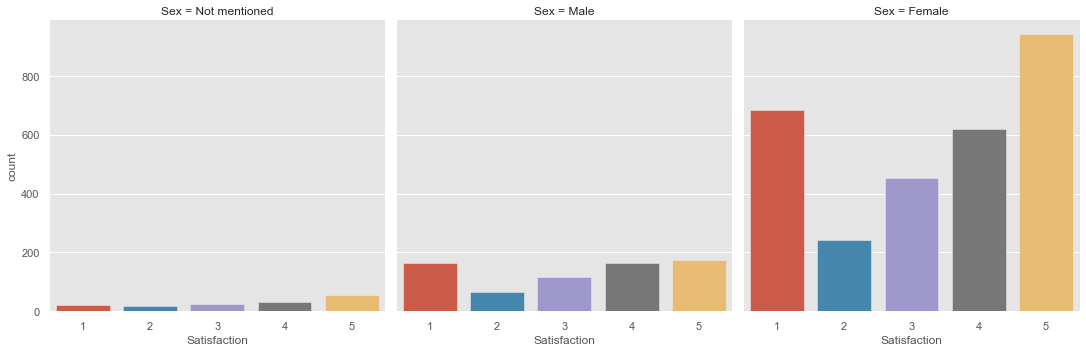
* Plotted a pie chart to analyze the percentage of conditions of patients using the drug.



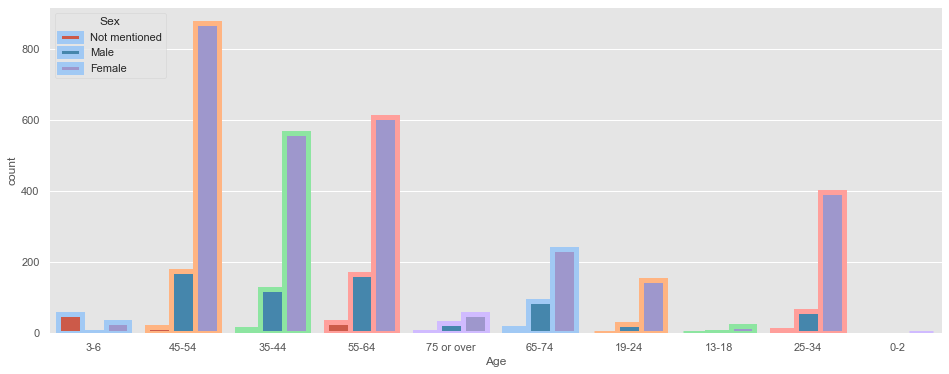
* From this plot, we can understand that patients with condition ‘Major depressive Disorder’ has highest percentage. So, we can plot a pie chart with patients having side effects with the condition ‘Major depressive Disorder’



* Plotted a Cat plot of satisfaction with respect to sex.



# Plotted Count Plot displaying the age groups and their sex .from this, we got an insight that the age group 45 -54 used this drug mostly and within that male ratio is higher.



* Plotted Pie plots on the effectiveness of drug on each age groups with their percentages.
* **PREPARE DATA FOR TRAINING**

Two tasks will be performed with resultant data and then further then divided into training and test sets. Say, Training data and testing data. The data is split into training (75%) and testing (25%) data sets through random sampling. The following machine learning techniques were considered in the experiment. As the dataset is labeled properly, it is considered to be used for supervised learning. In order to find out the best machine learning technique, different machine learning techniques were tested and based on the accuracy further decision was taken.