Dear Intern

Project report is an inherent component of your internship. We are enclosing a reference table of content for the project report. Depending on the internship project (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections from the table of content mentioned below. You can also add additional sections. The key objective of this report is for you to systemically document the project work done.

|  |  |
| --- | --- |
| 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 | 25/03/2021 | 125 | Jupyter Notebook | MS Excel, Jupyter |

**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**

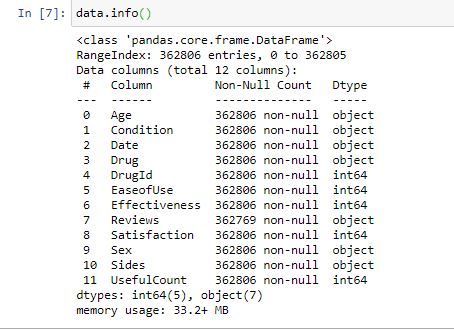
By various Exploratory data analysis, we can come to an assumption that the drug is rated good for the body by chemist, it has a slight side effect of the dataset mainly for depression. The condition attribute mainly concentrates on Major depressive disorder, Chronic muscle or bone pain, Neuropathic 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 issue 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**



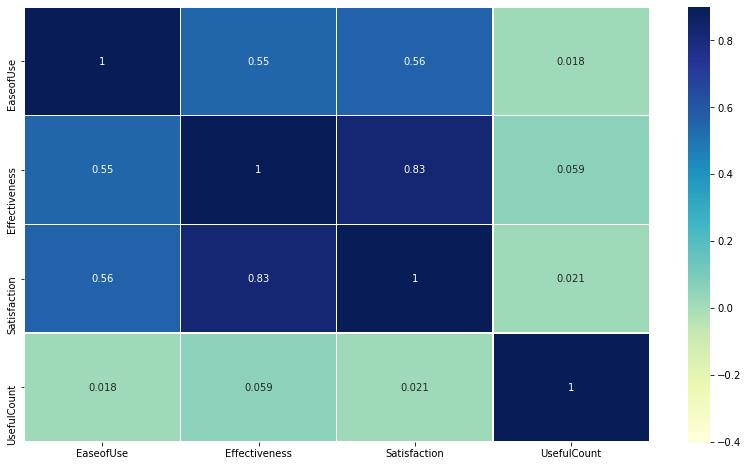
1. Age (Categorical) : Age Range
2. Condition (categorical): Condition of patient
3. Date (Categorical) : Date of patient report
4. Drug (categorical): Name of drug
5. Drug Id (categorical): Unique identification of a drug
6. Ease of Use (Numerical) : 5 step ease of use rating
7. Effectiveness (Numerical) : 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**

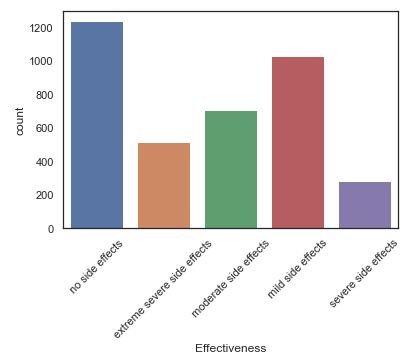
Afterwards, I have filtered the dataset with one drug to predict the side effect of that particular drug which occurs most using **mode function**. The drug thus obtained is **‘cymbalta’** with drug id **‘91491’**. For the sake of simplicity dataset is converted to one specific drug for better prediction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

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

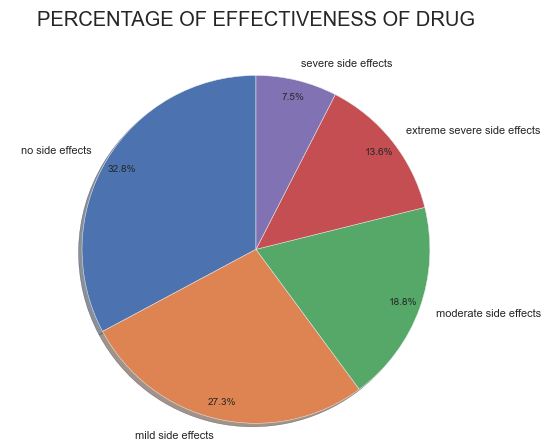


Correlation matrix shows the parameters of the dataset of different ranges than [-1, 1]. Where we will remove the low and high correlated feature from the dataset.

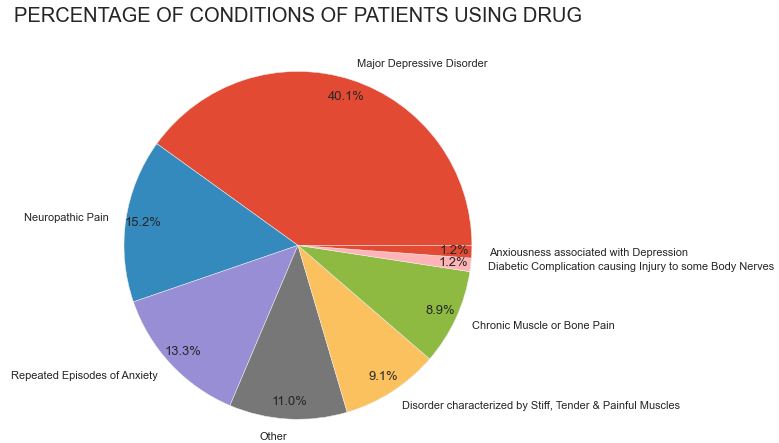
* Plotted a count plot on Effectiveness to analyze the count and found that ‘No side effects’ show maximum count.



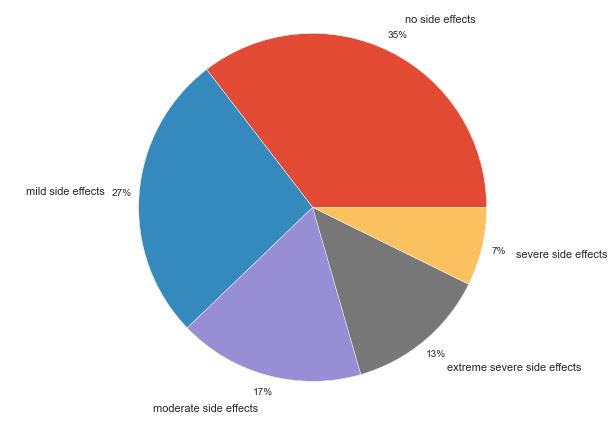
* Plotted a pie chart to analyze the percentage of effectiveness of drug.



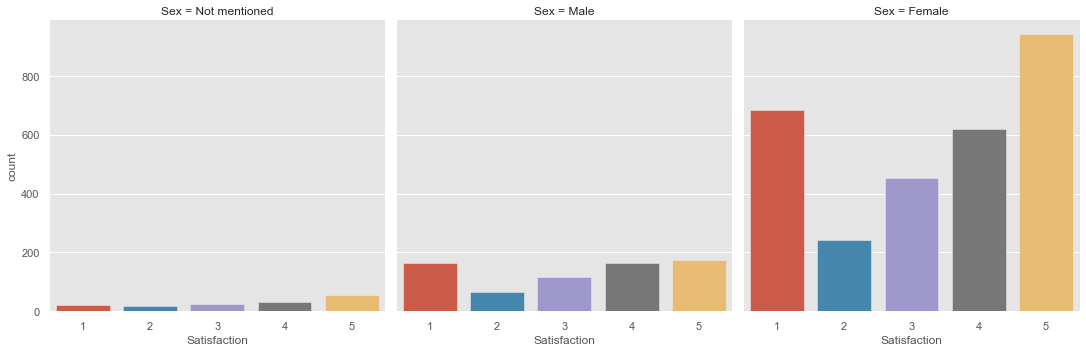
* Plotted a pie chart to analyze the percentage of conditions of patients using the drug.



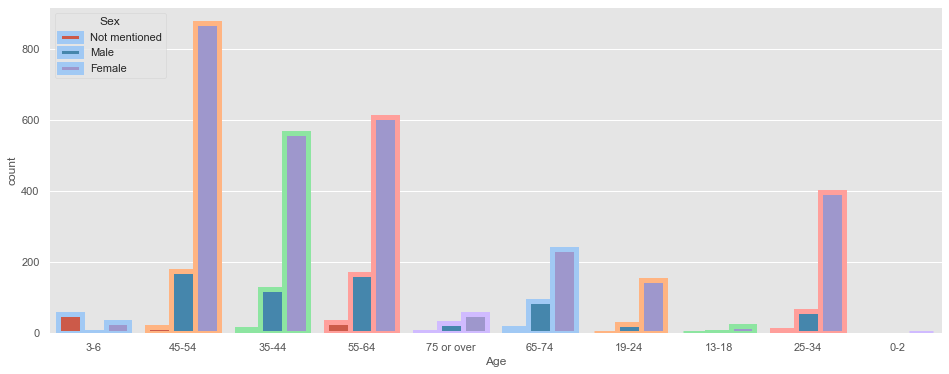
* From this plot, we can understand that patients with condition ‘Major depressive Disorder’ has highest percentage. So, we can plot a pie chart with patients having side effects with the condition ‘Major depressive Disorder’



* 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.

* **ALGORITHMS**

***1 – LOGISTIC REGRESSION***

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

***2 - THE K-NEAREST NEIGHBORS (KNN)***

KNN algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It’s easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

***3 - LINEAR SVM***

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

***4 – DECISION TREE***

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

***5 – RANDOM FOREST***

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

***6 – 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.

***7 – BERNOULLI NAIVE BAYES***

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions

**CONCLUSION**

The Side effect of the drug is best modelled by Gradient Boosting without doing any scaling and standardization (accuracy of **62.3%**). People in the age group 45-54 mostly used this drug and within that the male ratio was higher. People with the condition 'Major Depressive Disorder' consumed this drug more. Over all by comparing with race also, the male ratio is higher.

**REFLECTIONS ON THE INTERNSHIP:**

It was a great experience. The start of the internship was bitter, but gradually made into the track. Learned about the project documentation and understood how to make the best out the given data. Gained more knowledge on machine learning algorithms and its uses. The webinar helped a lot to grasp the idea of the reference dataset. Overwhelmed with the mentor support from TCS – iON.

**LINK TO CODE AND EXECUTABLE FILE:**

[soorajs3006/TCS\_INTERNSHIP: Drug Classification (github.com)](https://github.com/soorajs3006/TCS_INTERNSHIP)