|  |  |
| --- | --- |
| 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 | 29/4/2021 | 125 | Jupyter notebook | Excel, Jupyter |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Charts, Table, Diagrams
* Algorithms
* Reflections on the Internship
* Conclusion
* Link to code and executable file

**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 Because of unintended side effects, a great number of approved drugs were even withdrawn from the market.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 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 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.

**ABOUT THE DATASET**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate, Text | **Number of Instances:** | 4141 |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 5 |
| **Associated Tasks:** | Classification, Regression | **Missing Values:** | N/A |

### Data Description:

* Drug Name (categorical): name of drug
* rating (numerical): 10 star patient rating
* effectiveness (categorical): effectiveness rating
* condition (categorical): name of condition
* side effects(text): patient review

The dataset provides patient reviews on specific drugs along with related conditions. Furthermore, reviews are grouped into reports on the three aspects benefits, side effects and overall comment. Additionally, ratings are available concerning overall satisfaction as well as a 5 step side effect rating and a 5 step effectiveness rating. The data is split into a train (80%) a test (20%) partition and stored in two .tsv (tab-separated-values) files The data was obtained by crawling online pharmaceutical review sites. The intention was to study  
  
(1) sentiment analysis of drug experience over multiple facets, i.e. sentiments learned on specific aspects such as effectiveness and side effects,  
(2) the transferability of models among domains, i.e. conditions, and  
(3) the transferability of models among different data sources .

**CHARTS:**

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

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

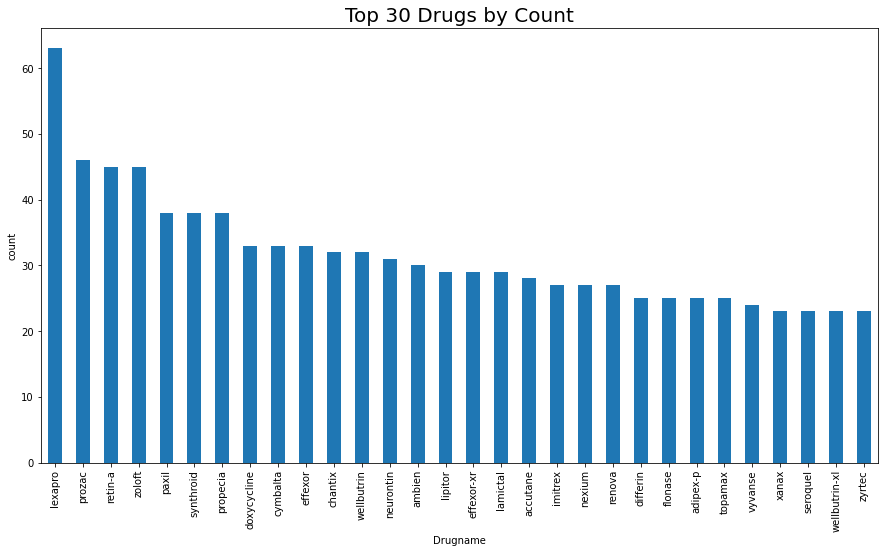
plt.figure(figsize = (15,8))

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

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

plt.xlabel("Drugname")

plt.ylabel("Count")



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

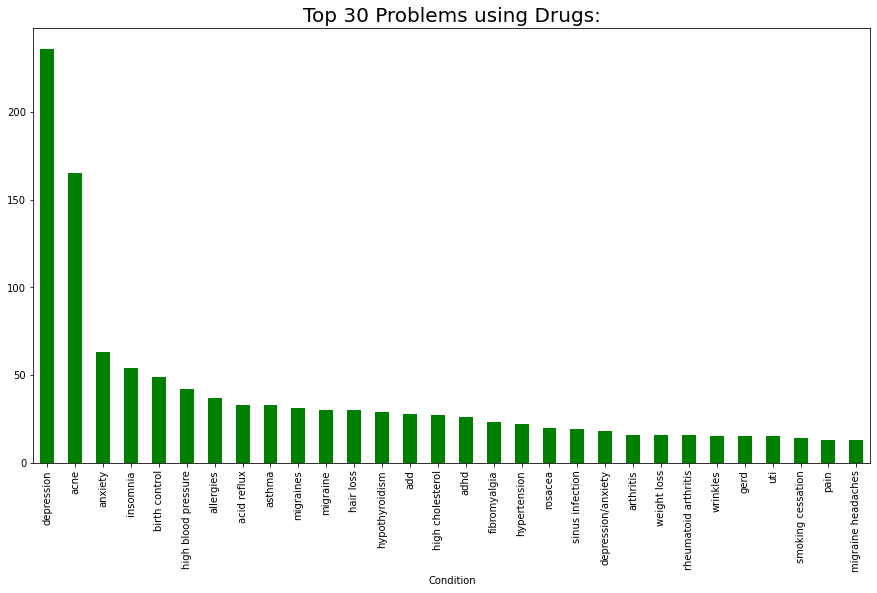
1. **Top-30 patients problems using Drugs by Condition:**

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.

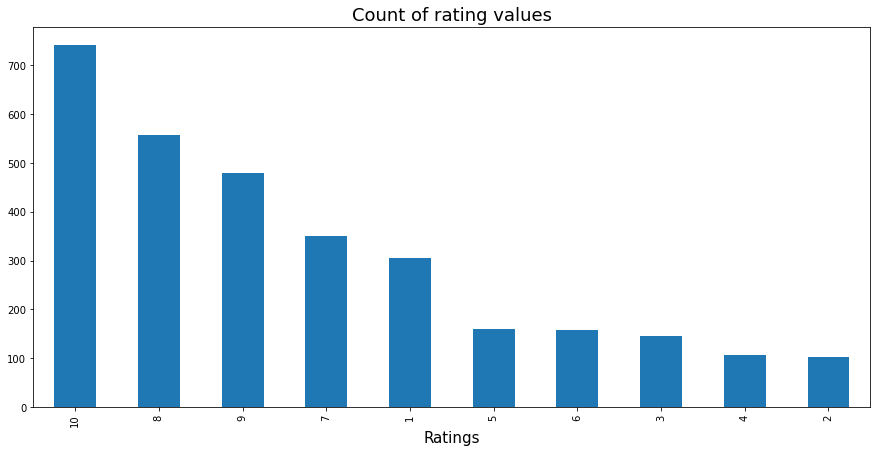
1. **Count the number of ratings**

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

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

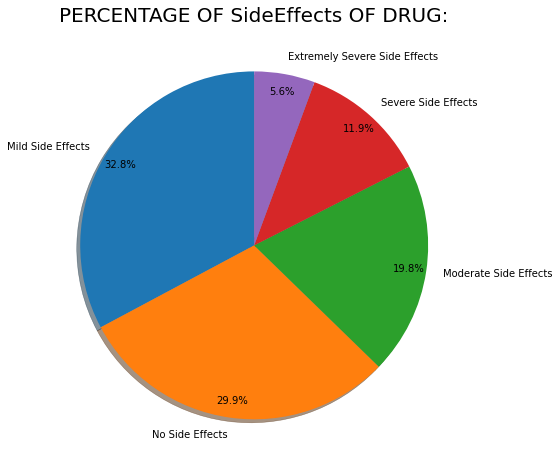
plt.xlabel('Ratings',fontsize = 15)

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.

# Check the Percentages of Side Effects Of DRUG



In this figure we can identified that severe side effects has 11.9% and extremely severe side effects has 5.6% i.e severe side effects of drugs is < 12%, remaining drugs side effect not effected much.

**Prepare data for training**

Two tasks will be performed which is resultant data is then divided into training and test sets. Training and Testing Data The data is split into training (80%) and testing (20%) 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 RMSE 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. It is used to examine the association of (categorical or continuous) independent variable(s) with one dichotomous dependent variable. This is in contrast to linear regression analysis in which the dependent variable is a continuous variable.The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

1. ***Random Forest Classifier****:* It is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. The RandomForestClassifier class of the sklearn.ensemble library is used to Classifier problems via random forest. The most important parameter of the RandomForestClassifier class is the n estimators parameter. This parameter defines the number of trees in the random forest.
2. ***Bernoulli Naive Baye’s***

Naive Baye’s methods are a set of supervised learning algorithms based on applying Bayes’ theorem [BernoulliNB](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#sklearn.naive_bayes.BernoulliNB) implements the Naive Baye’s training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued variable.

1. ***Decision Tree Classifier***

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

**5. *K-Nearest Neighbors (KNN):***

It 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

***6 .* 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.

***7 . 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.

**CONCLUSION**

The Side effect of the drug is not severe, it is less effected the people by depression, acne and anxiety, insomnia, high blood pressure. The best model by K-Nearest Neighbours(KNN) . Considering our dataset attributes we derived 7 models where K-Nearest Neighbours(KNN) has the highest accuracy which is 59% accuracy score.

|  |  |  |
| --- | --- | --- |
| **Models** | **Accuracy** | **After scaling Accuracy** |
| Logistics Regression | 34.726688 | 45.016077 |
| RandomForestClassifier | 43.56913 | 38.102893 |
| Bernoullis Naïve Bayes | 40.675241 | 40.996784 |
| DecisionTreeClassifier | 39.389067 | 38.10289 |
| SupportVectorMachine(SVM) | 39.389067 | 38.102893 |
| K-Nearest Neighbours(KNN) | 50.1610305 | 59.90338 |
| Gradient BoostingClassifier | 45.65916 | 47.42765 |

**REFLECTIONS ON THE INTERNSHIP:**

This internship was a great experience. The start of the internship was littles bit difficult, 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.

**LINK TO CODE AND EXECUTABLE FILE:**

<https://github.com/Anjana247/TCSiON-Internship>

**Reference:**

[**https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29**](https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29)

**Kagggle**

**Google**