|  |  |
| --- | --- |
| Internship Project Title | Classification Model-Build a Model that classifies the side effects of a Drug. |
| Name of the Company | TCS-ion |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | Tatyasaheb Kore Institute of Engineering and Technology  (TKIET), Warananagar |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 20-01-2025 | 19-02-2025 | | 125hrs | | Virtual Internship Project | Google Colab |
| Milestone # | 3 | Milestone: | | Completed TCS iON RIO-125: Classification Model - Build a Model that Classifies the Side Effects of a Drug - Internship | | |

**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

**ACKNOWLEDGEMENT**

First, I would like to thank TCSION for giving me the opportunity to do an internship within the organization.

I also would like all the people that worked along with me to support with their patience and openness they created an enjoyable working environment.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

I am highly indebted TCS-ion authorities for the facilities provided to accomplish this internship.

I would like to thank my mentor Mr. Debashis Roy for his constructive criticism throughout my internship.

I would like to thank ICT internship for their support and advices to get and complete internship in above said organization.

I am extremely great full to my colleagues and friends who helped me in successful completion of this internship.

**Table of Contents**

|  |  |  |
| --- | --- | --- |
|  | **Abstract** | **4** |
| **1** | **Problem Definition** | **5** |
| **1.1** | **Project Overview** | **5** |
| **1.2** | **Problem Statement** | **5** |
| **1.3** | **Objective** | **5** |
| **1.4** | **Domain Knowledge** | **5** |
| **2** | **Introduction** | **6** |
| **3** | **Data Analysis** | **7** |
| **4** | **Cleaning & Pre processing** | **9** |
| **5** | **Visualizations** | **12** |
| **6** | **Splitting the dataset** | **16** |
| **7** | **Model Selection** | **16** |
| **8** | **Result** | **19** |
| **9** | **Conclusions** | **19** |

**ABSTRACT**

In pharmacology, a drug is a chemical substance, typically of known structure, which, when administered to a living organism, produces a biological effect. A pharmaceutical drug, also called a medication or medicine, is a chemical substance used to treat, cure, prevent, or diagnose a disease or to promote well-being.

A side effect is an unwanted symptom caused by medical treatment. All medicines can cause side effects, including prescription, over-the-counter and complementary medicines. Complementary medicines include herbal preparations, vitamins, and some products dispensed by naturopaths and other practitioners of complementary medicine.

Here in this internship, I need to create a dataset of 4 lakh patients with details such as names, age, gender, race, side effects. First, we will perform preprocessing on the data then performing classification algorithms on this data. Feature engineering, feature selection and classification model will be selected.

**1.PROBLEM DEFINITION**

**1.1. PROJECT OVERVIEW**

In this internship I need to create a dataset of 4 lakh patients with details such as names, age, gender, race, side effects. First, we will perform preprocessing on the data then performing classification algorithms on this data. Feature engineering, feature selection will be done and classification model will be selected.

Actually, we have taken a dataset with 362806 rows and 14 columns. We need to do only classification of side effects of a particular drug from the dataset.

**1.2. PROBLEM STATEMENT**

This is a multilevel classification problem which classifies the effectiveness of a particular drug by analysing the dataset. The goal of the internship is to predict the effectiveness of a drug with respect to age, gender and race.

**1.3. OBJECTIVE**

Side effects and Effectiveness of a particular drug need to be addressed. The objective of this project is to build a classification model that classifies the side effects of a particular drug by age, gender and race.

**1.4. DOMAIN KNOWLEDGE**

The different tools used in this project:

* Python
* Machine Learning

**2.INTRODUCTION**

A side effect is usually regarded as an undesirable secondary effect which occurs in addition to the desired therapeutic effect of a drug or medication. Side effects may vary for each individual depending on the person's disease state, age, weight, gender, ethnicity and general health.

Side effects can occur when commencing, decreasing/increasing dosages, or ending a drug or medication regimen. Side effects may also lead to non-compliance with prescribed treatment. When side effects of a drug or medication are severe, the dosage may be adjusted or a second medication may be prescribed. Lifestyle or dietary changes may also help to minimize side effects.

In this Internship, we are creating a Classification Model for classifying the side effects of a particular drug chosen from the dataset.

**3.DATA ANALYSIS**

**3.1. Importing Python libraries**

For analyzing and creating a model we have imported a number of python libraries. That includes:

* pandas
* numpy
* matplotlib.pyplot
* seaborn
* sklearn.preprocessing
* sklearn.model\_selection.train\_test\_split

**3.2. Importing the data**

The data imported using the function ‘read\_csv()’. It is a function from the library ‘pandas’. ‘pandas’ is imported as ‘pd’. So, we have used ‘pd.read\_csv(filename)’ for reading the dataset.

**3.3. Data source**

Google Drive Link: https://drive.google.com/file/d/1qc2WKsZ\_7sDCijMrlRyzWRDmyYYy3ZkI/view?usp=sharing

**3.4. Dataset Description**

**Name:** It contains the names of different patients whose data are collected.

**Age:** It is a range column. It contains the range of age group within which the patient belongs.

**Condition:** It is a categorical column which contains the various conditions of patients or symptoms.

**Date:** It is a date column which contains the date on which a patient taken that drug.

**Drug:** It is a categorical column which contains the name of various drugs patients are taking.

**Drug Id:** It is a unique value for each Drug.

**Ease of Use:** It is a kind of rating given to each drug based on its easiness in using.

**Effectiveness:** It is also a kind of rating for side effects of a particular drug.

**Reviews:** These are categorical columns. It contains the reviews of different patients on having particular drug.

**Satisfaction:** It is also a rating provided by patients when they are satisfied with the drug they used.

**Sides:** These are categorical column which contain all the side effects of drugs.

**Useful count:** These shows the usefulness of drug as a ranking.

**Sex:** These are gender of various patients. It contains male and female.

**4.DATA CLEANING AND PRE-PROCESSING**

**4.1. DRUG SELECTION:**

From finding the mode of Drugs column we could easily find the most commonly occurring drug in that column.

We have taken the drug named ‘cymbalta’. We dropped all other rows which is having other drugs.

**4.2. DROPPING COLUMNS:**

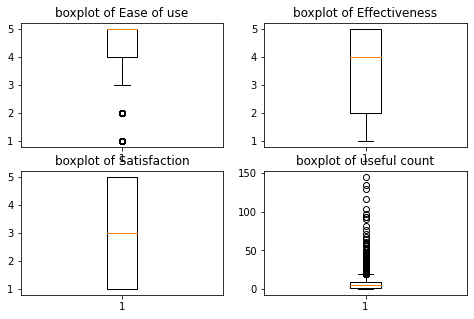
We have dropped the columns 'Date', 'Drug', 'Drug Id', 'Sides', 'Name', ’Reviews’.

The columns Reviews was having majority of its values as unique.So I dropped it.

The column name doesn’t have to deal with the Side effects of a drug.

The column Drug, Drug Id, and Sides are having similar values for the same drug .So I dropped it.

**4.3. FINDING OUTLIERS AND REMOVING:**.

****

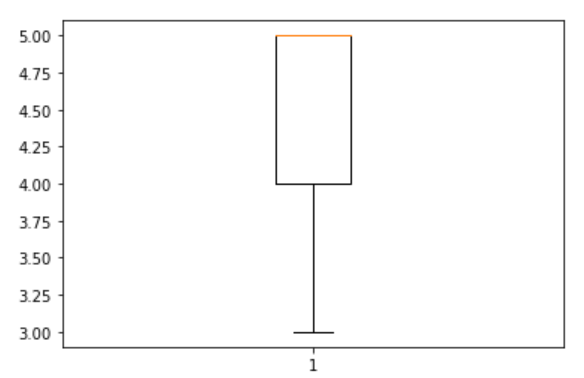
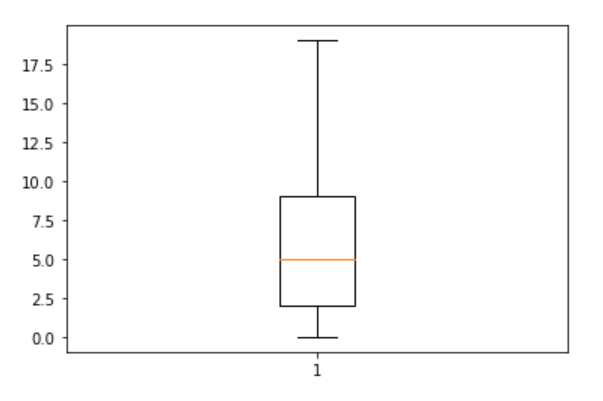
Boxplot of 4 different columns are given above.

The outliers present in 2 columns Useful count and Ease of Use is dropped.

The dataset without removing outliers is copied to another variable ‘data\_cpy’.

After dropping outliers, the boxplot appears as such.

Useful count Easeofuse



**4.4. RECODING VALUES:**

* 2 values in column Age is recoded.It contained a blank space as well as a value as ‘7-Dec’.So I recoded them to the range values.
* The values of Effectiveness column is recoded:
  + 1 = extreme severe side effects
  + 2 = severe side effects
  + 3 = moderate side effects
  + 4 = mild side effects
  + 5 = no side effects.

Now we have a clean data set after completing all the pre – processing steps.

**4.5. ENCODING:**

For creating a model, we have to encode all the categorical variables using label encoding or one hot encoding. Here we used label encoding for this purpose.

LABEL ENCODING: In label encoding in Python, we replace the categorical value with a numeric value between**0 and the number of classes minus 1.**If the categorical variable value contains 5 distinct classes, we use (0, 1, 2, 3, and 4).

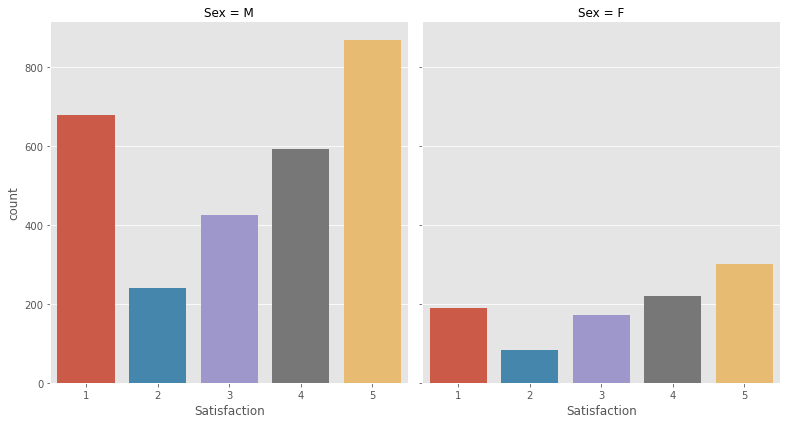
Here we have done label encoding to columns "Age", "Condition", "Sex". Now we have our prediction column as the only categorical value.

**5.VISUALIZATIONS**

Data visualization is a visual (or graphic) representation of data to**find useful insights** (i.e. trends and patterns) in the data and making the process of data analysis easier and simpler.

I have imported libraries of matplotlib, Seaborn, pandas for doing the visualization.

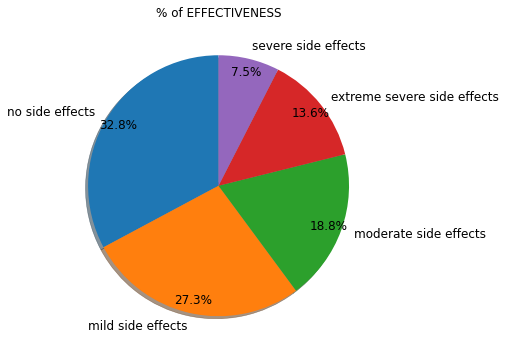
**5.1. CATPLOT OF SATISFACTION WITH RESPECT TO SEX COLUMN**



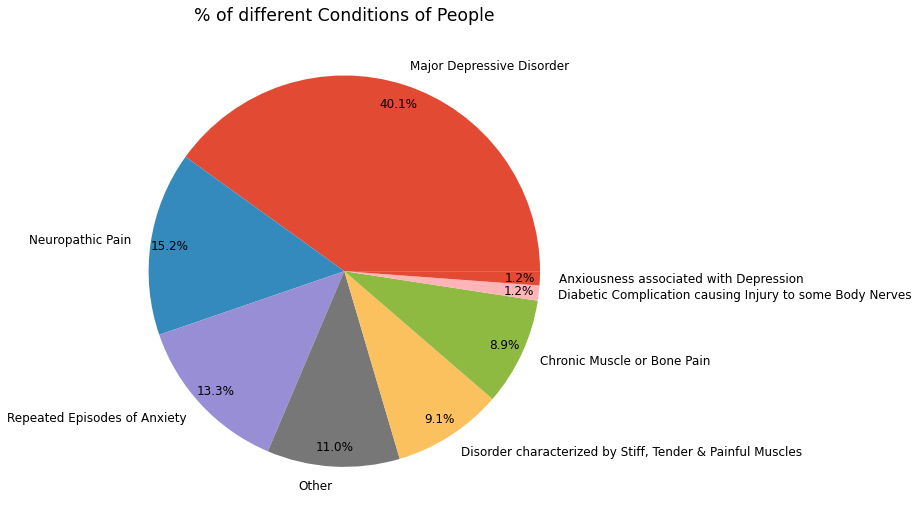
In this graph, in the x-axis we have the different Satisfaction and on y-axis we have the count of each.This graph is splitted with respect to gender, that is male and female. Overall the male ratio is higher. Different satisfaction ratings were given in different colours.

**5.2.PIE CHART OF PERCENTAGE OF EFFECTIVENESS OF CYMBALTA DRUG.**

In the pie chart given below each of the effectiveness is given in different colours.Almost 32% of people has no side effects on using the drug ‘cymbalta’.And 13% of patients are having extremely severe side effects.



**5.3.PIE CHART OF PERCENTAGE OF DIFFERENT CONDITIONS OF PEOPLE:**



The pie chart given above shows the percentage of different conditions of people. From this graph it is clear that, Persons with Major Depressive Disorder mostly take this drug.

# 5.4. COUNT PLOT OF VARIOUS AGE GROUPS TAKEN FOR THIS AND THEIR SEX IS DISPLAYED BELOW.

# The graph given below is a count plot which shows different age groups on x-axis from 0 to above 75. The count is given on y-axis. The age group 45-54 is consuming this Drug more and a higher ratio of males consume this. Male ratio is shown in orange colour and female ratio in blue colour.

# 

# 5.5. EFFECT OF DRUG ON AGE GROUP 45-54(MOST COMMONLY USED BY)

# 

# From graph 5 it is clear that patients in the age group 45-54 consume this drug more.So here we depicted the effectiveness of that drug in the age group 45-54.34% of people doesn’t have any side effects. Only 13% resulted in having extremely severe side effects.

# 5.6.SCATTER PLOT OF RELATING VARIABLES SATISFACTION AND SEX

# 

Different conditions are given on different color lines and sex is given on x-axis and Satisfaction on y-axis.

**6.SPLITTING THE DATASET**

To create a model, we need to split the dataset as train and test data. For that we are dividing the dataset into two variables ‘X’ and ‘Y’. ‘Y’ is the target value that we are going to predict so this contains the column ‘Effectiveness. All the other columns fall into the category ‘X’.Then we are performing the splitting operations using the train\_test\_split() function.

**7.MODEL SELECTION**

Ihave applied almost 6 classification algorithms to this dataset and checked. The accuracy remains somewhat constant even after doing Standardization, Min max scaling and Normalization.

We also applied these algorithms to the dataset without removing outliers.

**KNN**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

Accuracy: 58%

Accuracy without removing outliers:59%

**DECISION TREE**

Decision Tree algorithm belongs to the family of supervised learning algorithms. In Decision Trees, for predicting a class label for a record we start from the **root** of the tree.

We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Accuracy:50%

Accuracy without removing outliers: 47%

**GRADIENT BOOSTING**

Gradient boosting is a [machine learning](https://data-flair.training/blogs/machine-learning-tutorial/)technique for regression and classification problems. That produces a prediction model in the form of an ensemble of weak prediction models.

Accuracy:62%

Accuracy without removing outliers: 68%

**LINEAR SVM**

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Accuracy:53%

Accuracy without removing outliers: 30%

**RANDOM FOREST CLASSIFIER**

The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction

Accuracy:55%

Accuracy without removing outliers: 62%

**BERNOULLI NAIVE BAYES**

**Naive Bayes** is a classification algorithm of Machine Learning based on Bayes theorem which gives the likelihood of occurrence of the event. Naive Bayes classifier is a probabilistic classifier which means that given an input, it predicts the probability of the input being classified for all the classes. It is also called conditional probability.

Bernoulli Naive bayes used for discrete data and it works on Bernoulli distribution. The main feature of Bernoulli Naive Bayes is that it accepts features only as binary values like true or false, yes or no, success or failure, 0 or 1 and so on. So when the feature values are binary we know that we have to use Bernoulli Naive Bayes classifier.

Accuracy:48%

Accuracy without removing outliers: 32%

**8.RESULT**

From the Accuracy calculated from different models it is noticed that Gradient Boosting algorithm is having higher accuracy. So modelling is done with that algorithm. When we are removing outliers and checking the accuracy it is less compared to the dataset without removing outliers. The accuracy can be increased when more data set is available to train the model.

**9.CONCLUSION**

In the dataset I have done all the necessary pre-processing steps. Then I have made some useful visualizations out of it. After that Classification of side effects of a drug called ‘Cymbalta’ is done.

The Side effect of the drug is best modelled by Gradient Boosting without doing any scaling and standardization. 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.