|  |  |
| --- | --- |
| 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 | ICT Academy of Kerala |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 09-09-2023 | 30-10-2023 | 125hrs | Virtual Internship Project | Jupyter Notebook |

**ACKNOWLEDGEMENT**

The internship focused on developing a classification model aimed at categorizing the side effects of various drugs, which was a crucial component of the Data Science and Analytics course at ICT Academy, Kerala. I would like to express my heartfelt gratitude to all the dedicated teachers and the director of ICT Academy, as well as the team at TCS-iON, for their unwavering encouragement and support throughout the duration of this project.

I also want to extend my sincerest thanks to our project guide and mentor, Debashis Roy, whose constant guidance and assistance were instrumental in our success. Last but certainly not least, I offer my gratitude to the divine for guiding us towards the successful completion of this internship project with their blessings

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter** | **Tittle** | **Page no.** |
|  | ABSTRACT | **4** |
| **1** | PROBLEM DEFINITION  1.1 Over view  1.2 Problem Statement  1.3 Objective | **5** |
| **2** | INTRODUCTION | **6** |
| **3** | EXPLORATORY DATA ANALYSIS  3.1 Importing Python Libraries  3.2.Importing the data  3.3.Data source  3.4 Description of the dataset  3.5 overview of the dataset  3.6 Descriptive Statistical Summary 3.7 Selection of drug | **7** |
| **4** | 4.DATA CLEANING AND PRE-PROCESSING  4.1 Handling missing values  4.2 Text Processing  4.3 Outliers  4.4 Feature Engineering | **10** |
| **5** | VISUALIZATIONS | **14** |
| **6** | ENCODING | **18** |
| **7** | SPLITTING THE DATASET | **19** |
| **8** | MODEL SELECTION   * LOGISTIC REGRESSION * KNN * DECISION TREE * SUPPORT VECTOR MACHINE  1. Linear SVM 2. Polynomial SVM 3. Radial SVM  * RANDOM FOREST CLASSIFIER * GRADIENT BOOSTING * EXTREME GRADIENT BOOSTING | **19** |
| **9** | RECOMMENDATIONS | **22** |
| **10** | CONCLUSION | **23** |

**ABSTRACT**

Within the realm of pharmacology, a drug is defined as a chemical compound with a well-established structure that, when administered to a living organism, elicits a specific biological response. In the context of healthcare, these chemical substances, also known as pharmaceutical drugs or medications, serve various purposes such as treating, curing, preventing, diagnosing diseases, or enhancing overall well-being.

Side effects, on the other hand, represent unintended and often undesirable symptoms resulting from the use of medical treatments, and this can occur with any form of medication, including prescribed drugs, over-the-counter options, and complementary medicines like herbal remedies, vitamins, and products recommended by practitioners of complementary medicine.

In the course of this internship, the primary objective is to create a comprehensive dataset containing the details of approximately 400,000 patients. These details will encompass crucial information such as their names, ages, genders, races, and the side effects experienced. The internship process will be divided into several stages.

**1.PROBLEM DEFINITION**

**1.1. Overview**

In this internship project, our main goal is to work with a dataset containing 362,806 patient records. We will focus on classifying the side effects of a particular drug within this dataset. Our process involves data preprocessing, the application of classification algorithms, feature engineering and selection, and, ultimately, selecting the most effective classification model. The project aims to enhance our understanding of drug-related side effects and contribute to improved patient care and medication management.

**1.2. Problem Statement**

This internship project addresses a complex multilevel classification challenge, where the objective is to determine the efficacy of a specific drug by leveraging the dataset. The primary aim of this internship is to forecast the drug's effectiveness while considering variables such as age, gender, and race. Through this analysis, we aim to gain insights into the factors influencing a drug's performance, ultimately contributing to more informed decisions in healthcare and pharmaceutical management.

**1.3. Objective**

The central focus of this project is the assessment of both the side effects and effectiveness of a specific drug. Our primary objective is to construct a classification model capable of categorizing the side effects of this drug, taking into account the variables of age, gender, and race. By accomplishing this, we aim to provide a valuable tool for better understanding how these factors impact the drug's performance, with potential implications for more informed healthcare decisions.

**2.INTRODUCTION**

A side effect, in the context of drug or medication usage, is typically an unintended and often unwanted outcome that occurs alongside the intended therapeutic effects of the drug. These side effects can vary widely from person to person, depending on factors such as their medical condition, age, weight, gender, ethnic background, and overall health.

Side effects can manifest when starting, adjusting dosages, or discontinuing a drug regimen. In some cases, side effects can be so significant that they lead to non-compliance with prescribed treatment. When this happens, healthcare providers may opt to modify the dosage, or they might prescribe a different medication altogether. Additionally, lifestyle or dietary adjustments can sometimes help mitigate the impact of side effects.

Within the framework of this internship, our primary focus is on constructing a Classification Model. This model's purpose is to categorize and classify the side effects associated with a specific drug, using data from the provided dataset. The goal is to enhance our understanding of how different factors influence the occurrence and severity of side effects, ultimately contributing to more effective and personalized healthcare management.

**3. EXPLORATORY DATA ANALYSIS**

**3.1 Importing Python Libraries**

To facilitate our data analysis and model creation, we've harnessed the power of various Python libraries. These indispensable tools include:

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

These Python libraries are fundamental to our data analysis and model development process, enabling us to efficiently work with data, visualize insights, and construct a robust classification model.

**3.2.Importing the data**

The data has been imported using the 'read\_csv()' function, which is part of the 'pandas' library. In the code, 'pandas' is typically imported with the alias 'pd' to make it easier to reference the library. Therefore, the data is read from a CSV file using the following syntax: 'pd.read\_csv(filename)'. This function reads the data from the specified CSV file and loads it into a 'pandas' DataFrame, which is a powerful data structure for working with structured data in Python.

**3.3.Data source**

Google Drive Link: https://drive.google.com/file/d/1kyYcDaCVF9KVXwNJ4nF9jP9BvX8Y0vLX/view?usp=sharing

**3.4 Description of the dataset**

Name : Name of the patient

Race : Race of the patient

Age : Age of the patient

Condition : Condition/symptom from which the patient is suffering

Date : Date of usage

Drug : Name of the drug

DrugId : Identity/code of drug

EaseofUse : Patient's 10-Star rating on the ease of use of drug

Effectiveness : Patient's 10-Star rating on the effectiveness of drug

Reviews : Patient's review

Satisfaction : Patient's 10-Star rating on satisfaction

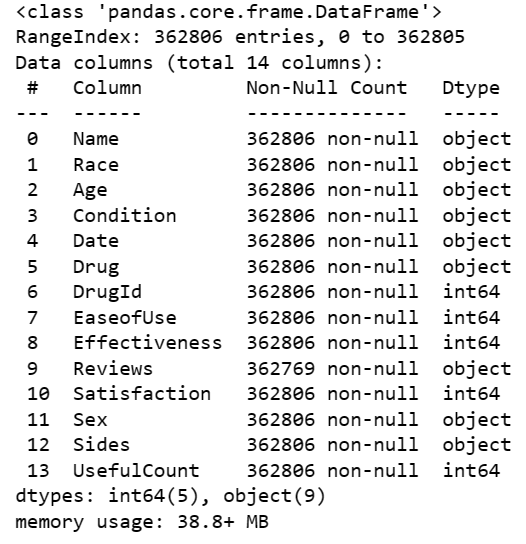
Sex : Gender of the patient

Sides : Side effects of the drug

UsefulCount : Number of users who found the review useful

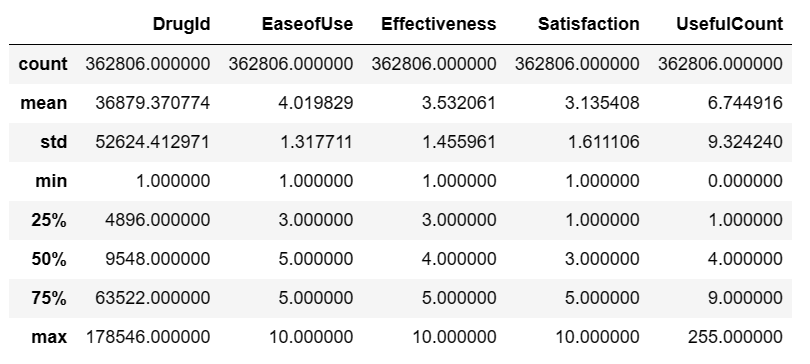
**3.5 overview of the dataset**

The ‘info()’ function of the Data Frame provides a summary of the dataset, including the total number of rows, the types of attributes present, and the count of non-null values for each attribute. By calling ‘df.info()’ a concise description of the dataset is obtained as below (Fig 1). This dataset contains 9 categorical variables and 5 integer variables.



**Fig-1**

**3.6 Descriptive Statistical Summary**

Descriptive statistical summary provides a concise overview of the key statistical measures for each attribute in the dataset. This summary includes measures such as mean, standard deviation, minimum, maximum, quartiles, and count. It helps to understand the central tendency, dispersion, and distribution of the data. To obtain the descriptive statistical summary, the describe() function in pandas is used.

**Fig-2**

**3.7 Selection of drug**

In this project, we have chosen a specific drug named "Lisinopril" with a unique DrugId of 6873 as our focal point for building a classification model. The primary objective of this model is to analyze and classify the side effects associated with the administration of Lisinopril. By using this drug as the subject of our study, we aim to gain insights into the factors and variables that influence the occurrence and severity of side effects in individuals who use Lisinopril. This classification model will ultimately assist in better understanding and managing the side effects of this particular medication.

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

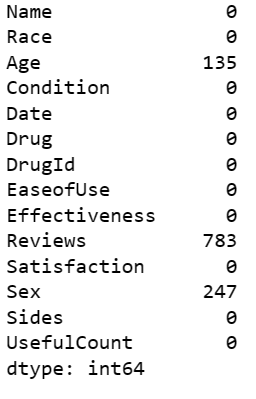
Data pre-processing refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for building and training Machine Learning models. It is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn. Therefore, it is extremely important to pre-process our data before feeding it into our model.

Data pre-processing mainly includes the steps of Data Cleaning, Data Transformation and Data Reduction. Data Cleaning involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Data Transformation is the step of converting the data into a suitable format for analysis. Data Reduction means reducing the size of the dataset while preserving the important information. The steps of pre-processing performed with the drug side effects dataset are detailed in this chapter.

**4.1 Handling missing values**

The concept of missing values is important to comprehend in order to efficiently manage data. If the researcher, programmer, or academician does not properly handle the missing figures, he or she may get to the wrong conclusion about the data, which will have a significant impact on the modeling phase. It is a significant problem in data analysis since it has an impact on the outcomes. It is difficult to have total faith in the insights when you know that several items are missing data. It may reduce the statistical power of research and lead to erroneous results owing to skewed estimates. Problems due to missing values are, Statistical power, or the chance that the test would reject the null hypothesis when it is erroneous, is lowered in the absence of evidence, The loss of data might cause parameter estimations to be skewed, It has the ability to reduce the representativeness of the sample, It might also make the analysis of the study more challenging.

Figure 3 shown below presents a view of the missing values in each feature of the dataset. Observe that among all other features, Reviews has the most missing values and Age has the least missing values.



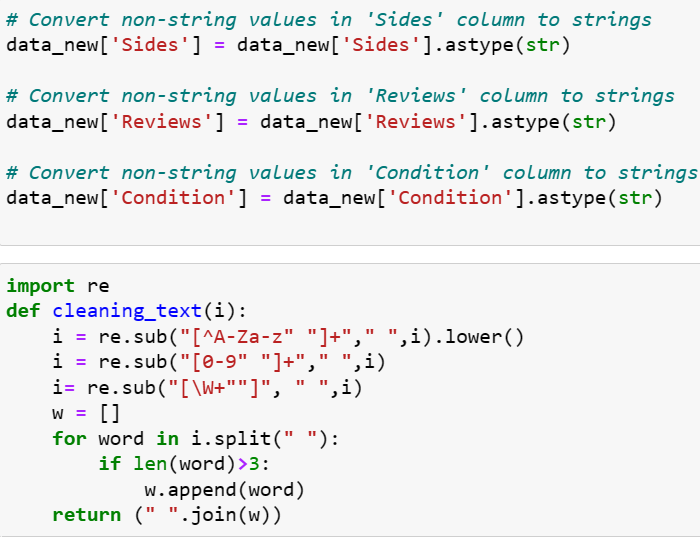
**Fig-3**

**4.2 Text Processing**

Text processing is a broad field that involves the manipulation and analysis of text data, often using computational methods and tools. It encompasses a wide range of tasks and techniques.

Text processing is a vital component of natural language processing (NLP) and plays a significant role in applications like information retrieval, chatbots, recommendation systems, and many more. It enables us to extract valuable insights from unstructured text data, making it a crucial field in the era of big data and digital communication.

Fig-4 below shows the codes used for text processing.

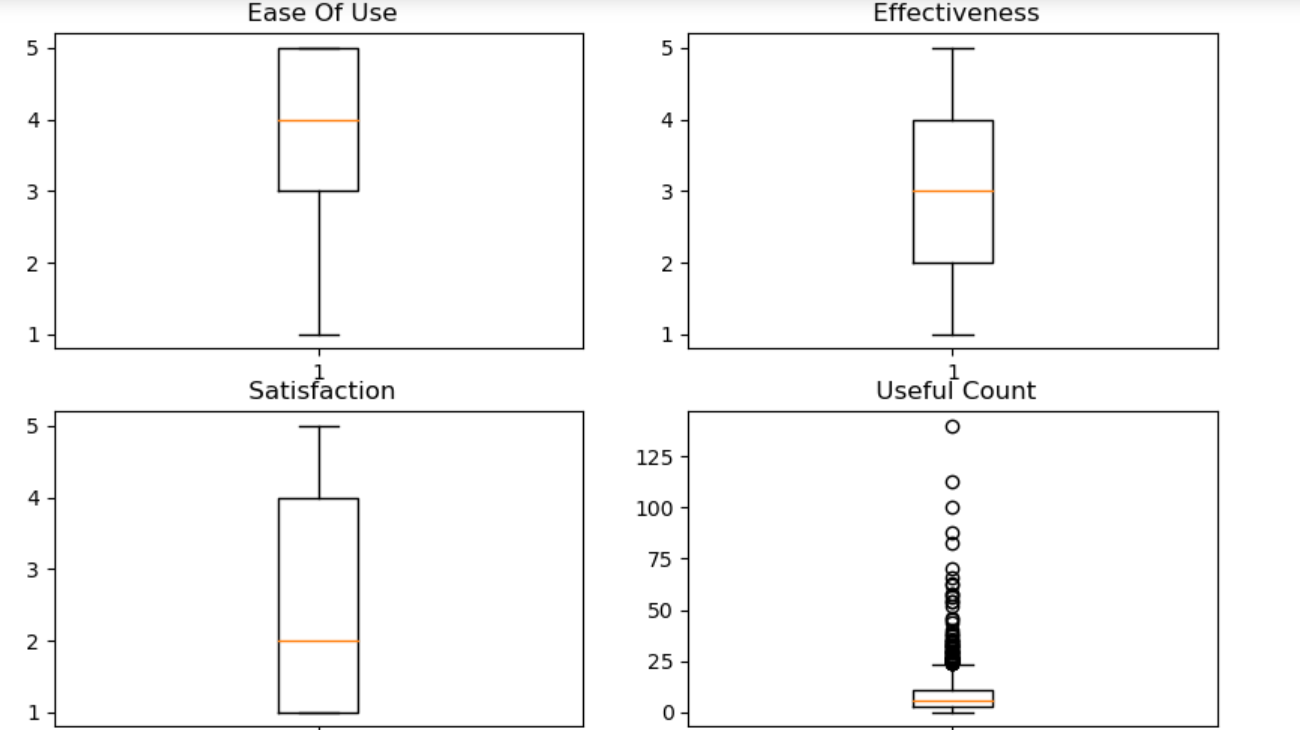


**Fig-4**

**4.3 Outliers**

An outlier is a piece of data that is an abnormal distance from other points. In other words, it’s data that lies outside the other values in the set. They are the extremely high or extremely low values in the data set. A simple way to find an outlier is to examine the numbers in the data set. We will see that most numbers are clustered around a range and some numbers are way too low or too high compared to the rest of the numbers. Such numbers are known as outliers.

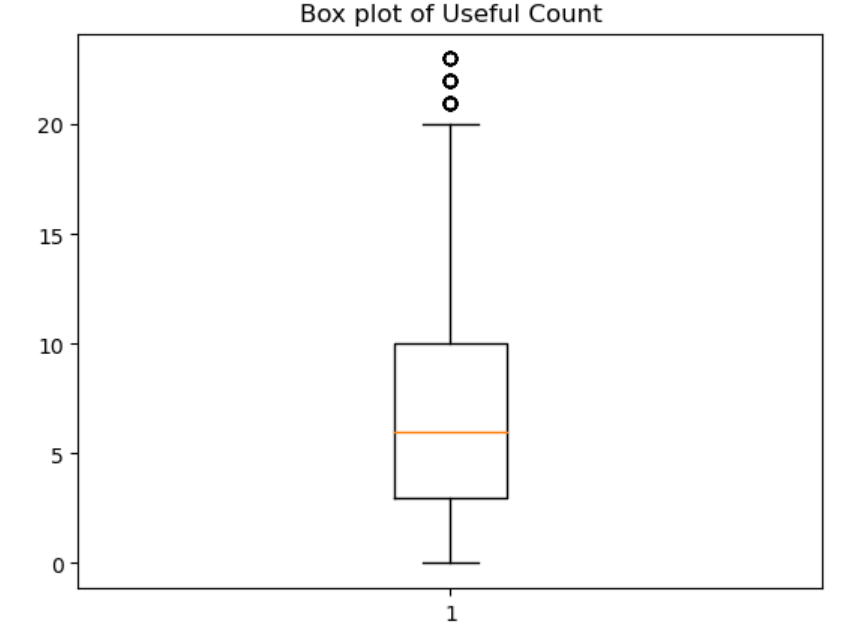
In addition, a box plot was generated for each attribute, categorized by class. A boxplot is a standardized visualization that presents the distribution of data based on a five-number summary, including the minimum, first quartile (Q1), median, third quartile (Q3), and maximum values. This plot helps identify outliers and provides insights into their values.

.

**Fig-5**

Boxplot of 4 different columns are given above.

The outliers present in Useful count.

After dropping outliers the boxplot appears as such.

**4.4 Feature Engineering**

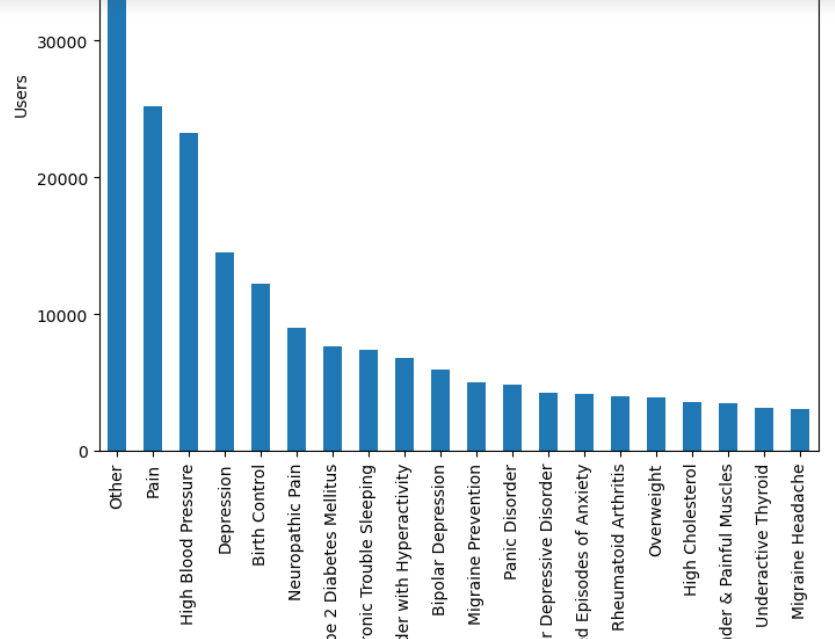
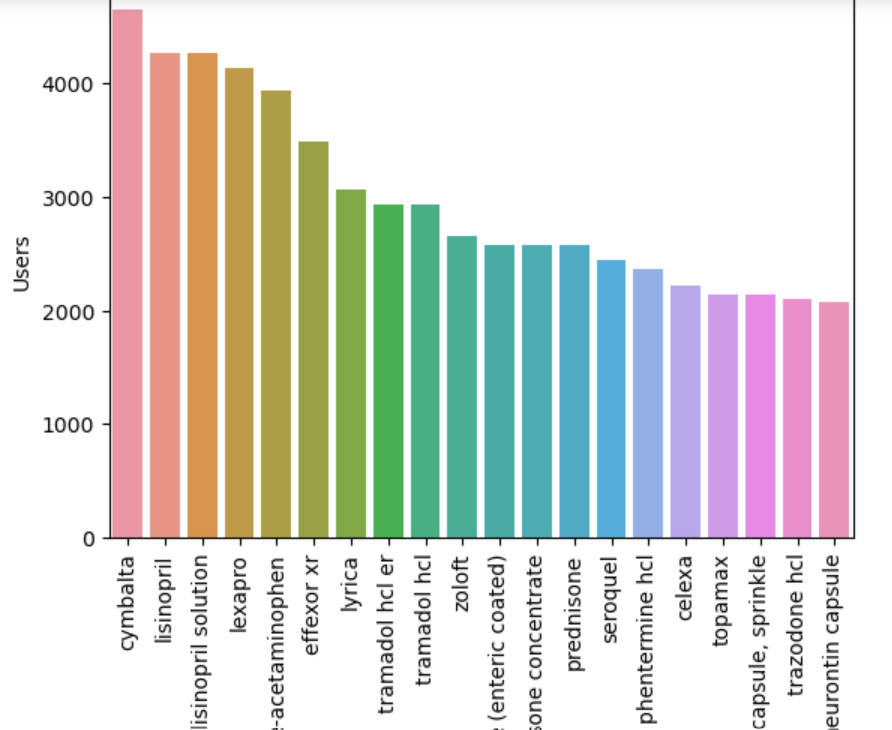
Feature engineering is the process of creating new features or transforming existing ones in a dataset to improve the performance of machine learning models. It's a crucial step in the data preprocessing pipeline and can significantly impact the accuracy and effectiveness of models.

Here we created a new column named "SideEffects" in a DataFrame based on the values in an existing column called "Satisfaction." Used NumPy's `np.where` function to perform conditional assignments. The resulting "SideEffects" column indicate the severity of side effects based on the "Satisfaction" ratings.

Then created a new column named "SideEffects" in the DataFrame `data\_new` by assigning the values from the `data1` array. The "SideEffects" column now contains the values "Nil," "Mild," "Moderate," "Severe," or "Extreme," depending on the corresponding values in the "Satisfaction" column. This additional column can be useful for further analysis and visualization of side effects in dataset. Then we removed two columns, "Satisfaction" and "DrugId," from the DataFrame using the `drop` method.

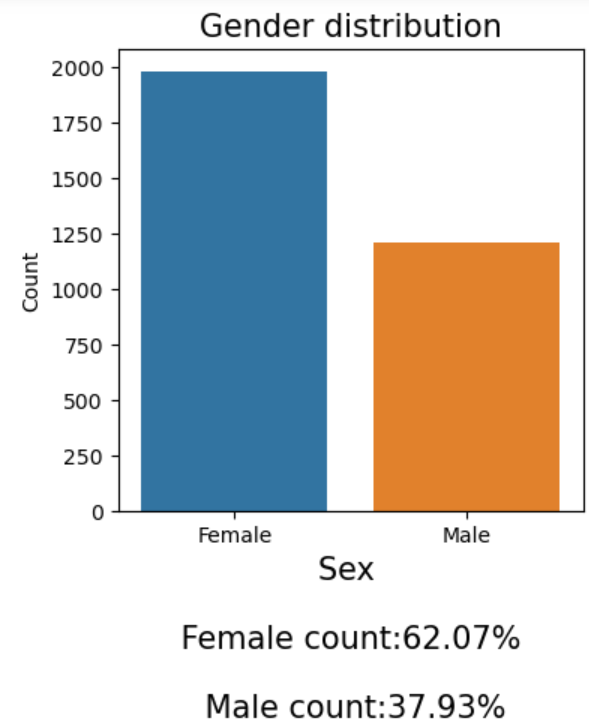
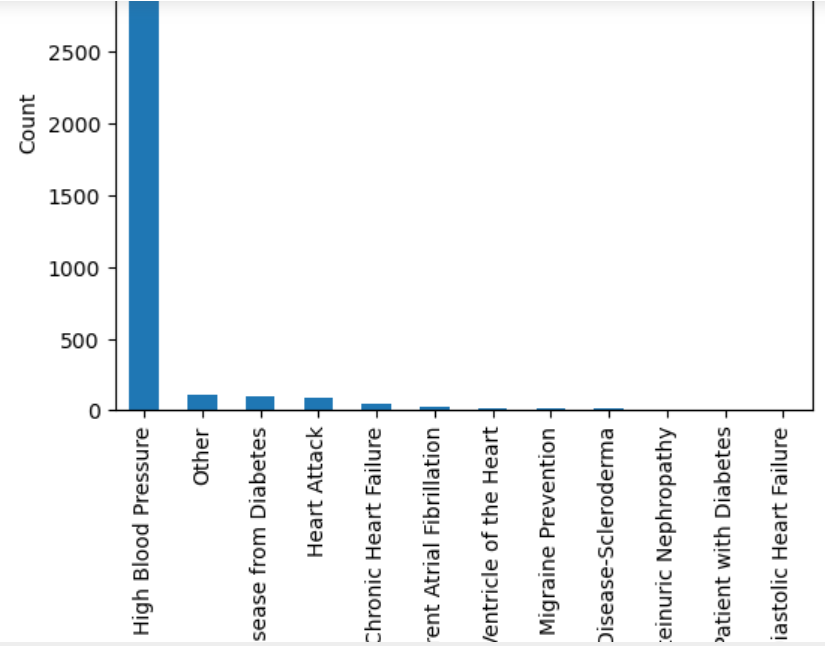
**5.VISUALIZATIONS**

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. ****

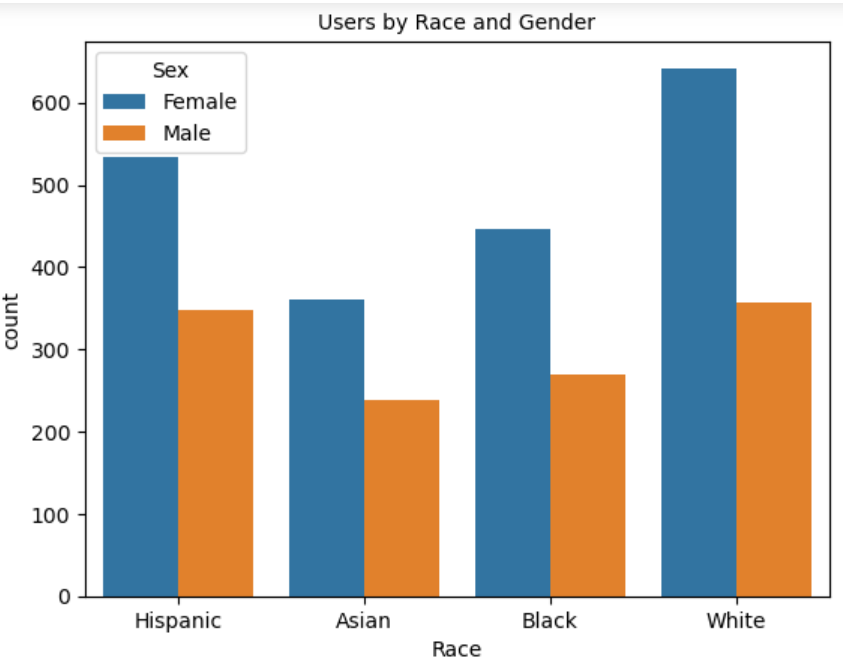
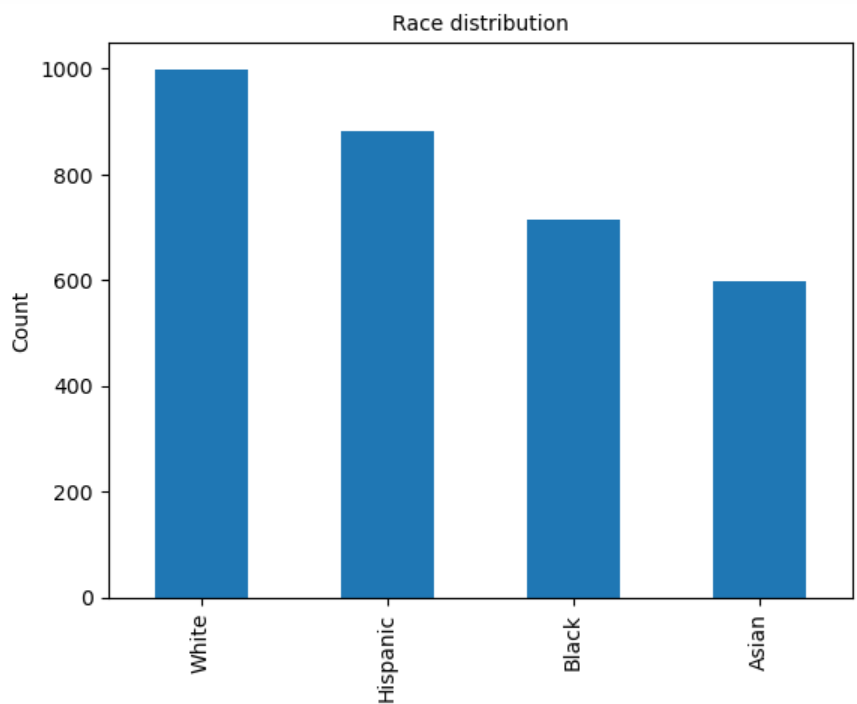
**Fig-7**

**Fig-6**



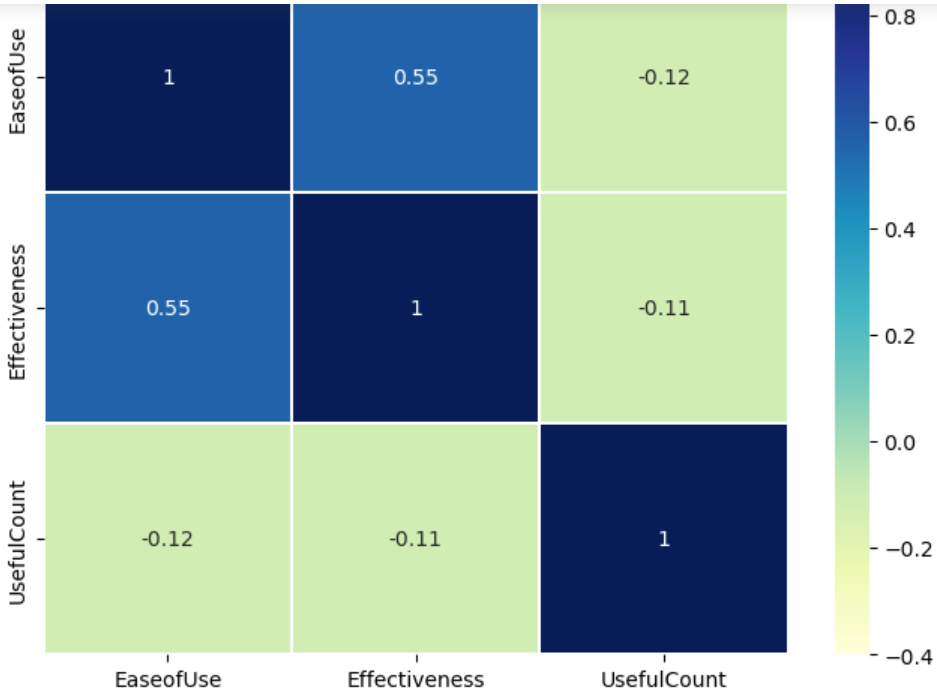
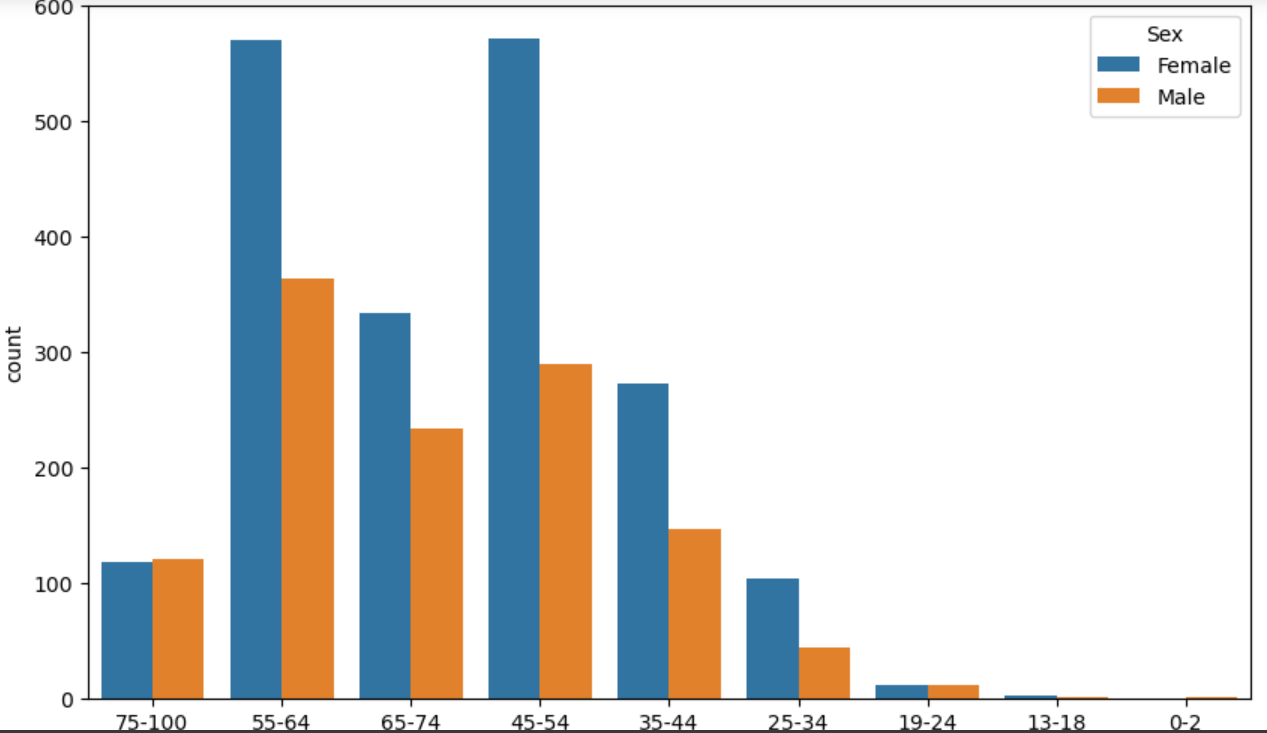
**Fig-9**

**Fig-8**



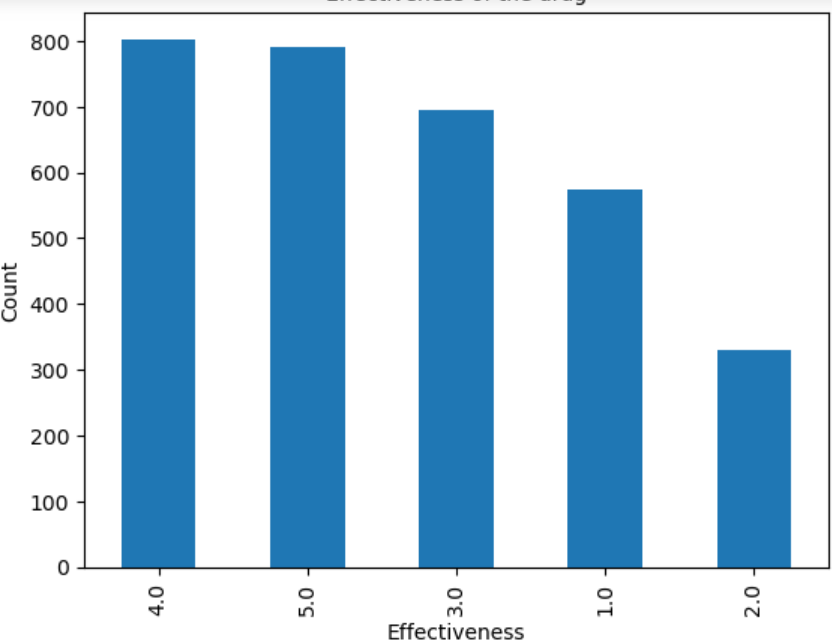
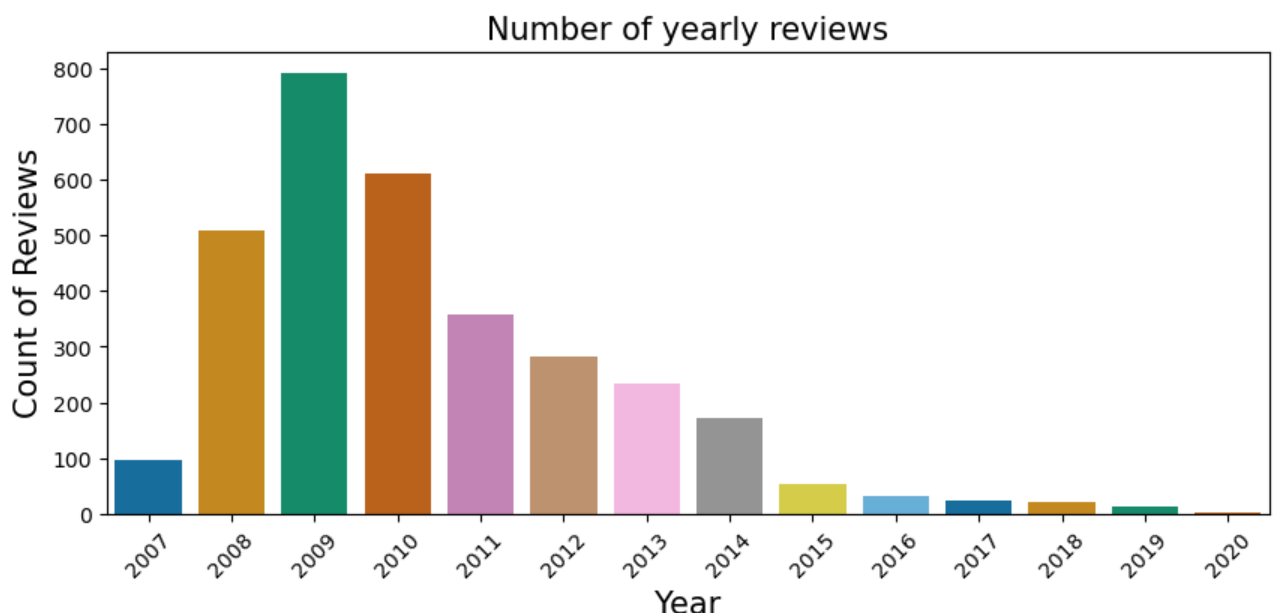
**Fig-11**

**Fig-10**



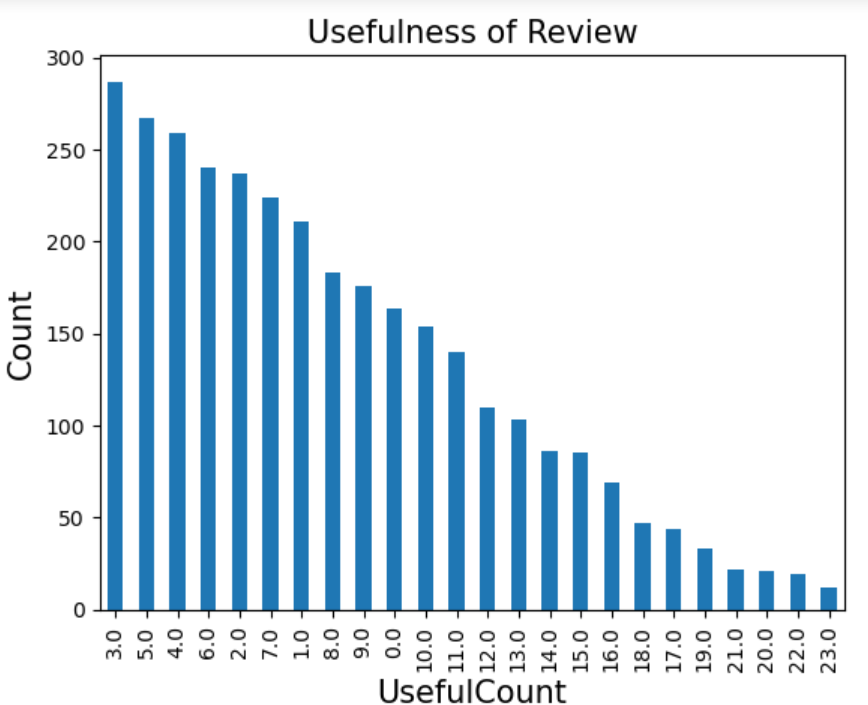
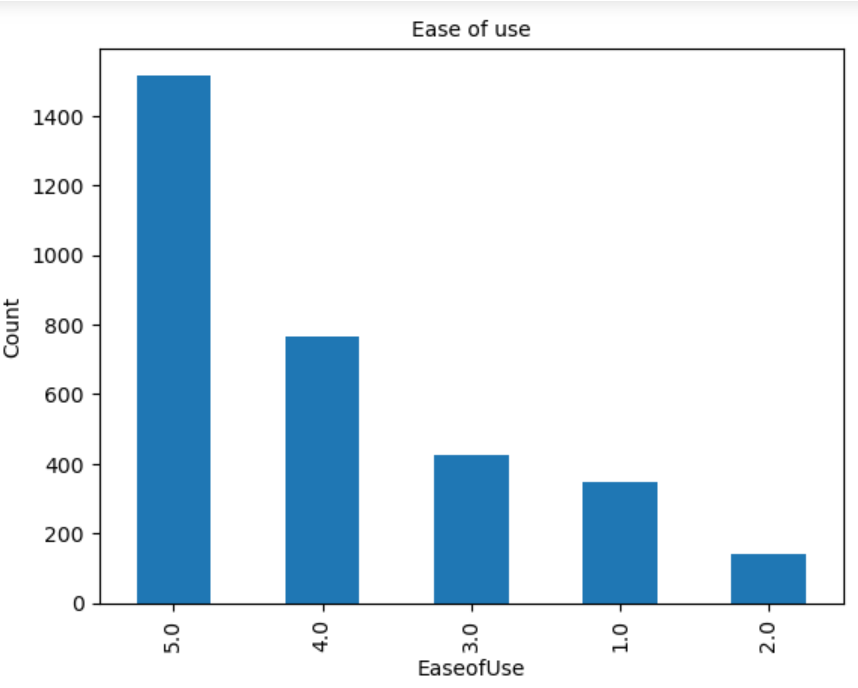
**Fig-13**

**Fig-12**



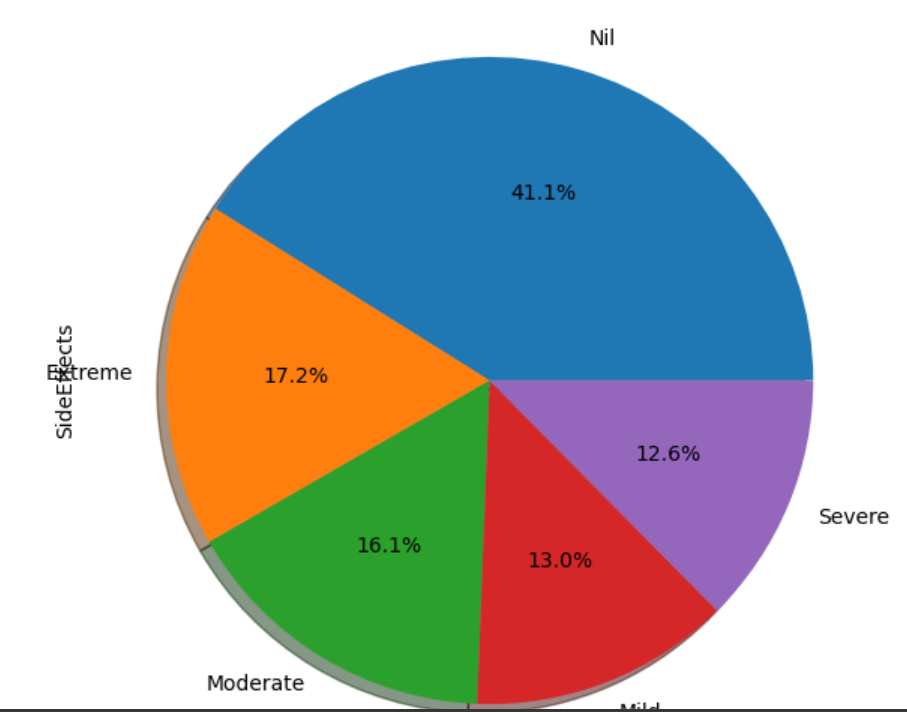
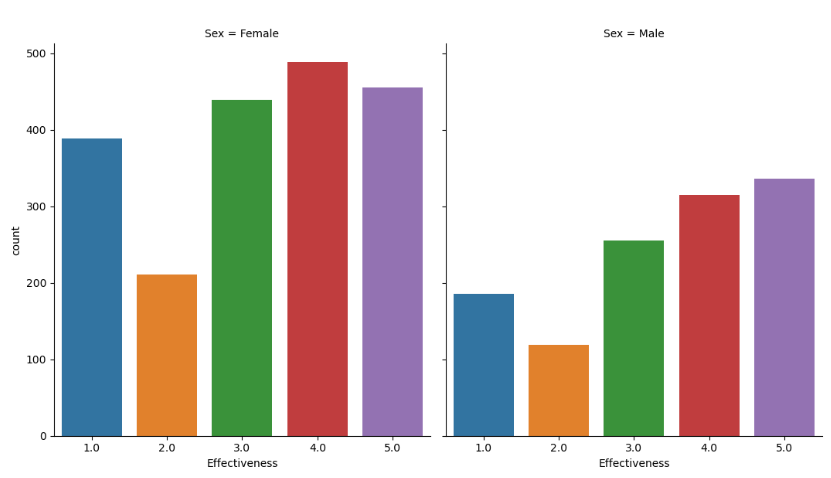
**Fig-15**

**Fig-14**



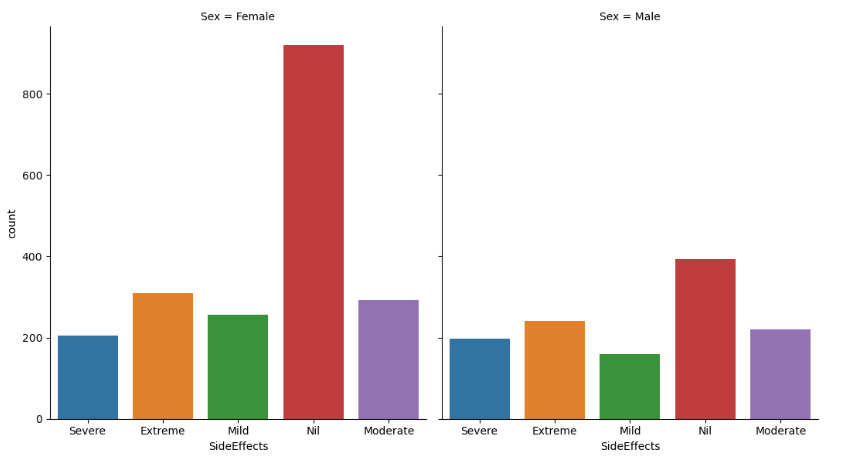
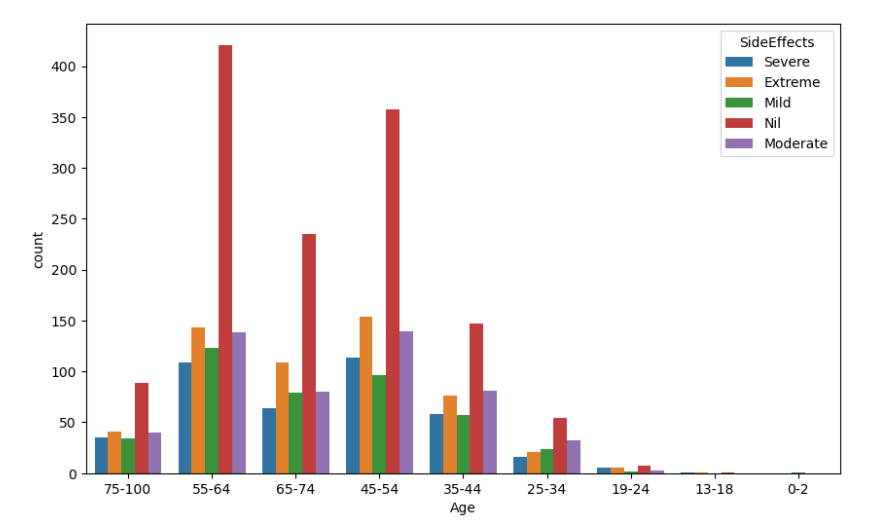
**Fig-17**

**Fig-16**



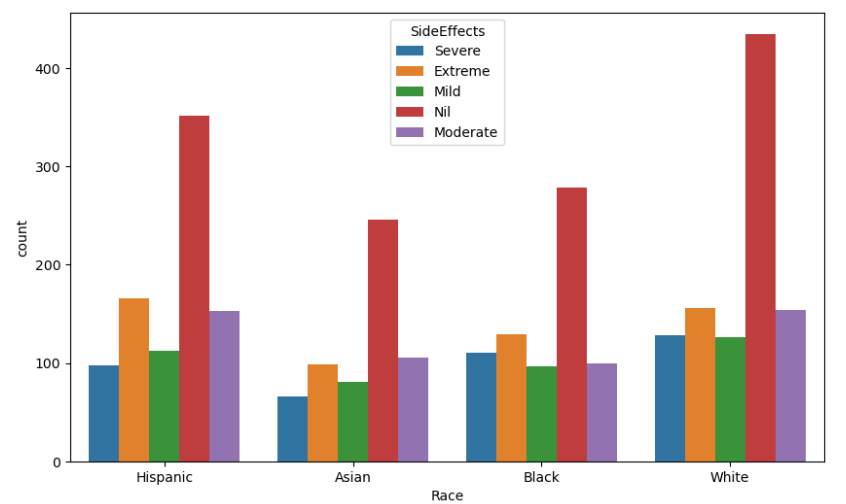
**Fig-19**

**Fig-18**



**Fig-21**

**Fig-20**



**Fig-22**

* Fig-6 shows the bar plot of top 20 drugs, Cymbalta has the maximum number of users
* Fig-7 shows the bar plot of top 20 conditions. Out of 1806 Conditions, Other Conditions tops the chart followed respectively by Pain and High Blood Pressure. According to this plot **we select a particular drug Lisinopril, one featuring in the top 5.**
* Fig-8 shows the bar plot of top conditions for using Lisinopril. Maximum users consume the drug for High blood pressure. The drug is mainly used for 12 conditions listed above.
* Fig-9 shows the count plot of gender distribution. Females uses the drug more compared to males.
* Fig-10 shows the bar plot of race distribution. Lisinopril drug is mostly consumed by Whites followed by Hispanic followed by Black and least used by Asian.
* Fig-11 shows the count plot of users by race and gender. Females dominate the consumption of Lisinopril drug within each race.
* Fig-12 shows the count plot of users by age and gender. Consumption of the drug by female users is visbly dominant across age groups above 24 and below 75 years. Plus, people in the age groups 45 to 54 and 55 to 64 accounted for maximum usage of Lisinopril.
* Fig-13 shows the correlation map. Correlation is used to find relationship between variables. Here, variables are weakly or moderately correlated to each other.
* Fig-14 shows the count plot of noumber of yearly reviews. Lisinopril’s yearly reviews peaked in 2009 and reduced gradually thereafter.
* Fig-15 shows the bar plot of effectiveness of drug. Most users found Lisinopril drug effective.But it cannot be considered the best drug as seen from the above graph.
* Fig-16 shows the bar plot of ease of use. Most users rated Lisinopril highly for its ease of use.
* Fig-17 shows the bar plot of usefulness of reviews. As with usefulness of reviews, Lisinopril users throw a mixed bag with most not finding the reviews beneficial.
* Fig-18 shows the cat plot of effectiveness and gender. Males and females have similar kind of effectiveness review for the drug. Graphs look different as the male consumption of the is drug is low compared to females.Most males were totally satisfied with the effectiveness of the drug. But thats not the case with females.
* Fig-19 shows the pie chart of side effects. Majority of consumers did not have any side effects. Many had Extreme side effects, some had Moderate side effects, a few had Mild side effect and a very few had Severe side effect.
* Fig-20 shows the count plot of side effects with respect to age. In almost all age groups majority of consumers did not have any side effects. No side effects topped in the age groups between 55-64.
* Fig-21 shows the cat plot of side effects with respect to gender. For females, majority did not have any side effects. In the case of males, Extreme and Moderate side effect cases are not very low compared to no side effects.
* Fig-22 shows the count plot of side effects with respect to race. Side effects proved race neutral with proportionate results.

**6.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", “SideEffects”. Now we have our prediction column as the only categorical value.

**7.SPLITTING THE DATASET**

To create a machine learning model, it's essential to divide the dataset into training and testing sets. To do this, we separate the data into two distinct variables: 'X' and 'Y'. 'Y' represents the target variable we aim to predict, in this case, it contains the 'Effectiveness' column. All other columns in the dataset are grouped into the 'X' variable, which serves as the input features for our model.

Next, we utilize the `train\_test\_split()` function to perform the actual data splitting. This function partitions the dataset into training data (used for model training) and test data (used for model evaluation). This step is crucial to assess the model's performance and ensure it generalizes well to unseen data.

**8.MODEL SELECTION**

Model selection is a critical step in machine learning and data science, where you choose the best algorithm or model for your specific problem. It involves evaluating and comparing different models to determine which one performs best on your dataset.

Model selection is an iterative process, and it's common to experiment with multiple algorithms and hyperparameter settings to find the best-performing model for your particular problem. The selection process may involve comparing different models, refining the chosen model, and continuously evaluating and improving its performance.

Ihave applied almost 7 classification algorithms to this dataset and checked. The accuracy remains somewhat constant even after doing Standardization, Min max scaling and Normalization.

**LOGISTIC REGRESSION**

Linear regression is a fundamental statistical and machine learning technique used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It's commonly employed for tasks such as predicting numerical values or understanding the association between variables.

Accuracy: 53.83%

**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 : 55.08%

**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:55%

**SUPPORT VECTOR MACHINE**

A Support Vector Machine (SVM) is a powerful and versatile machine learning algorithm used for both classification and regression tasks. It is particularly well-suited for problems with complex decision boundaries and high-dimensional data.

SVM is widely used in various domains, including image classification, text categorization, bioinformatics, and finance. Its ability to handle non-linear data and the versatility of its kernel functions make it a valuable tool for many machine learning tasks.

1. **Linear SVM**

Linear Support Vector Machine (Linear SVM) is a type of Support Vector Machine that is specifically designed for linear classification problems. It aims to find the optimal hyperplane that best separates data points belonging to different classes.

Accuracy: 53.52%

1. **Polynomial SVM**

Polynomial Support Vector Machine (Polynomial SVM) is an extension of the Support Vector Machine (SVM) algorithm that can be used for non-linear classification problems. While the standard Linear SVM is suitable for linearly separable data, Polynomial SVM is designed to handle data that cannot be linearly separated by mapping it to a higher-dimensional space using polynomial kernel functions.

Accuracy: 54.56%

1. **Radial SVM**

Radial Support Vector Machine (Radial SVM), often referred to as the Radial Basis Function (RBF) SVM, is a popular variant of the Support Vector Machine (SVM) algorithm. It is primarily used for non-linear classification and regression tasks. Radial SVM uses the radial basis function as a kernel to map data into a higher-dimensional space, where it can find non-linear decision boundaries.

Accuracy: 54.61%

**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:55.08%

Accuracy after fine-tuning : 55.71%

**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: 55.86%

**EXTREME GRADIENT BOOSTING**

Extreme Gradient Boosting, often referred to as XGBoost, is a powerful and versatile machine learning algorithm that has gained popularity for its effectiveness in a wide range of tasks, including classification, regression, and ranking. XGBoost is known for its high predictive accuracy, speed, and scalability.

Accuracy: 55.55%

**9.RECOMMENDATIONS**

Based on the derived EDAs, the following are recommended with regard to the use of Lisinopril:

1.It cannot be considered the best drug as its effectiveness is not that great.

2.This drug is normally used by people with high blood pressure.

3.Majority of consumers did not have any side effects from the consumption of this drug.

4.The drug is safe for use by those between the age group 45-74 years (maximum nil side effects)

5.Females in general are better to use the drug as majority of them did not have any side effects from the consumption of this drug.

6.Effectiveness of the drug is almost same regardless of race and gender.

**10.CONCLUSION**

Performed various pre-processing steps and obtained a clean dataset for exploratory data analysis and modelling. Various visualisations were done on the dataset and selected a particular drug Lisinopril. Maximum users consume this drug for high blood pressure.

The best classification model was obtained by Gradient Bossting model. On performance evaluation, maximum accuracy of 55.86% has been achieved.

By race, Whites topped in the usage of this drug with a higher female ratio. Domination of female users were consistent across age groups above 24 and below 75 years. Also, people in the age groups 45 to 64 accounted for maximum usage of of the drug Lisinopril. Most rated it highly for its ease of use. Side effects of the drug proved race neutral. Genderwise, side effects proved least in females.In almost all age groups majority of consumers did not have any side effects. In terms of effectiveness of the drug, it is not the best.