|  |  |
| --- | --- |
| Name Of The Student | Sethulekshmi S |
| Internship Project Title | **RIO 125 - 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 |
| 06/09/2023 | 16/10/2023 | 125 | Remote internship | Jupyter notebook,  Microsoft word |

**TABLE OF CONTENT**

* List of figures
* Objective
* Description of Internship
* Internship Activities
* Approach / Methodology
* Abstract
* Problem definition
* Introduction
* Exploratory data analysis
* Data preprocessing
* Data modelling
* Outcome/Conclusion
* Link to code
* Risk
* Enhancement scope
* Reflection on the internship

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1 | Activities |
| Figure 2 | Data set info |
| Figure 3 | Statistical summary of the data set |
| Figure 4 | Histogram for numerical attribute |
| Figure 5 | Top 20 drugs based on no. of users |
| Figure 6 | Top 10 drugs with a 5-star rating for satisfaction |
| Figure 7 | Top 20 conditions |
| Figure 8 | Top 10 drugs with a 5-star rating for effectiveness |
| Figure 9 | Drug use based on race |
| Figure 10 | Drug use based on gender |
| Figure 11 | Drug use based on race & gender |
| Figure 12 | Drug use by age group |
| Figure 13 | Drug use age group & gender |
| Figure 14 | Count of yearly reviews |
| Figure 15 | Effectiveness of Lisinopril drug |
| Figure 16 | Effect of drug among both the gender |
| Figure 17 | Correlation between every pair of attributes |
| Figure 18 | Visualization of IQR method |
| Figure 19 | Percentage of effectiveness |

**OBJECTIVE**

The objective of this project is to build a model that predicts whether the given drug has a side effect or not. This product could be helpful for medical companies to provide detailed rating of the side effects of the product over their site. It could also be helpful for the patients who are buying drugs online to check the side effects of the drugs before buying it.

**DESCRIPTION OF INTERNSHIP**

TCS iON Remote internships are the career starters one can depend on. With relevant industry experience, graduates, postgraduates, and students across various domains—even freshmen—can add value to their resumes through a wide variety of internship programmes. End-to-end up-skilling solutions prepare, test, and analyse knowledge and kick start a career from the comfort of your location.

TCS iON is one of the largest online platforms for students, with the TCS iON Digital Learning Hub offering a wide range of courses, assessments, and events. TCS iON Digital Learning Hub also hosts digital discussion rooms designed for students to collaborate, share knowledge, participate in activities like quizzes, and provide feedback. TCS iON offers end-to-end support for up-skilling self-learners, right from K–12 to working professionals.

The TCS iON Remote Internship comprises various activities, including self-learning sessions, webinars, project work, and a few collaborative activities.

**INTERNSHIP ACTIVITIES**

TCS iON Remote internships RIO-125 include 125 hours of Internship on Industry Projects with Guidance and Webinars from Industry Mentors. It comprises of various activities including self-learning sessions, webinars, project work and few collaborative activities.

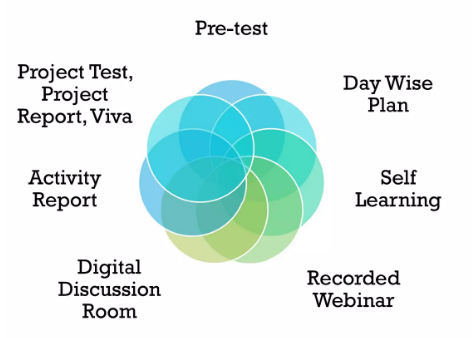


Figure 1 - Activities

**DAY WISE PLAN**

A day-wise plan or learning schedule gives an idea of how to spend the day and how much time to devote to each module for 30 days. A day-wise plan in a module under the table of contents in the course batch is a representative plan of activities that you would carry out during the tenure. This is to help students get a fair idea of how they may need to pace their efforts in doing the internship project. The industry mentor will post a planned set of activities for students to take in the digital discussion room over the course of the internship project.

**PRE – TEST**

A simple test for a quick knowledge check 1 hour duration was marked for 40 marks and 20 as passing marks. A pre-test is a general aptitude test that needs to be taken before commencing work on the project. The duration of the test is one hour, and it consists of questions to test analytical abilities, reasoning abilities, and language skills. Upon a successful attempt, it unlocks the details of the industry project and other nodes in the construct.

**SELF-LEARNING**

References to different learning resources enable students to successfully do projects and enrich learning. Self-learning reference materials are provided to help the students learn some of the concepts required to start working on the industry project.

**RECORDED WEBINAR**

Detailed description of the industry project, including hands-on environment, expected outcome, and introducing the industry guide Project-related learning sessions from industry experts to provide contextual, project-, or topic-level information.

**DIGITAL DISCUSSION ROOM**

Candidates can interact with each other and industry mentor regarding the project via digital discussion room.

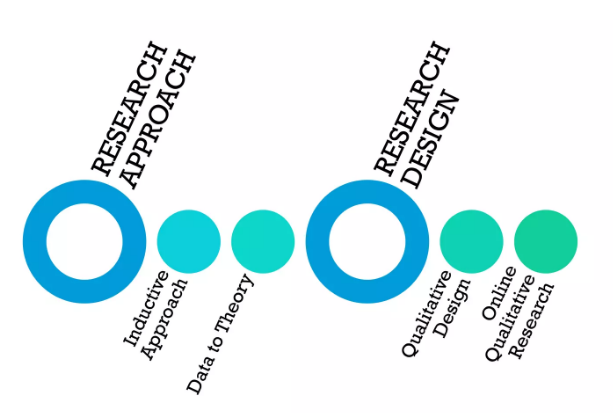
**ACTIVITY REPORT**

Daily activity report for documenting progress, including learning, project activities, mentor and peer interactions. (To be uploaded daily as per the activity schedule)

* 30 Activity Report
* 2 Interim Project Report
* Project Report
* Project Test – Project test is an online self-assessment on the internship project topic to be taken by students at the end of the internship.
* Viva – Viva to be conducted by Academic mentor.
* Viva Report

**APPROACH/METHODOLOGY**

The approach for this study is purely based on secondary research like the internet and articles on the same topic by various authors.



**ABSTRACT**

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. Drug toxicity (or adverse side effects) is a pressing health problem which is also an impediment to the development of therapeutically effective drugs. Despite many on-going efforts to determine the toxicity beforehand, computational prediction of drug side-effects remains a challenging task.

This thesis presents an approach to predict side-effects by utilizing side-information sources for the drugs, while simultaneously comparing state-of-the-art machine learning methods to improve accuracy. Specifically, the thesis implements a data analysis pipeline for obtaining side-information that is useful for the prediction task. This thesis then formulates the drug side-effect prediction as a machine learning problem: Given disease indications and structural features (as side-information sources) of drugs, for which some measurements of side-effect exist, predict side effect for a new drug. The present work investigates **Drug Dataset** taken from Kaggle.

The main focus of this project is learning about EXPLORATORY DATA ANALYSIS, DATA PRE-PROCESSING, FEATURE ENGINEERING, MODELING AND PREDICTION of data by supervised algorithms i.e. (Logistic Regression, Support Vector Machine, Random Forest Classifier, Decision Tree Classifier (classification)).

**PROBLEM DEFINITION**

**Overview**

Machine learning is used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to “produce reliable, repeatable decisions and results” and uncover “hidden insights” through learning from historical relationships and trends in the dataset (input).

Predicting drug side effects is an important topic in the drug discovery. Although several machine learning methods have been proposed to predict side effects, there is still space for improvements. Drug-related features are associated with side effects, and feature dimensions have specific biological meanings. Recognizing critical dimensions and reducing irrelevant dimensions may help to reveal the causes of side effects.

**Problem Statement**

The main agenda of this project is to build a model that classifies the side effects of a particular drug by age, gender and race. The prediction of side effects is a complex task that requires the consideration of multiple factors. To accurately predict the side effects of a drug, it is necessary to develop a model that takes into account relatable factors and their interactions. Machine learning techniques have shown promise in this regard, allowing for the development of predictive models that can accurately forecast air quality.

**INTRODUCTION**

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. However, wet experiments are costly and time-consuming. Since researchers collected drug data and compile them in the public databases, computational methods were developed for the side effect prediction.

Various models have been exercised in the literature to predict AQI, like statistical, deterministic, physical, and Machine Learning (ML) models. The traditional techniques based on probability, and statistics are very complex and less efficient. The ML-based AQI prediction models have been proved to be more reliable and consistent. Advanced technologies and sensors made data collection easy and precise. The accurate and reliable predictions through such huge environmental data require rigorous analysis which only ML algorithms can deal with efficiently.

The dataset is preprocessed and cleaned first, then methods of data visualization are applied to develop better insights and to investigate hidden patterns and trends. The data imbalance is identified and addressed with a resampling technique. Popular ML models are exercised in context with this resampling technique. Their performances are then compared through standard metrics.

Following are the steps followed during any data science projects:(Project Flow):

1. Define Goal (Business Objective)
2. Data collection and getting Data set details
3. EDA (Exploratory Data Analysis)
4. Data Preprocessing
5. Model Building
6. Evaluate and Compare Performance
7. Model Deployment

**EXPLORATORY DATA ANALYSIS**

This section of the present study deals with data exploration and analysis for finding various hidden patterns present in the dataset. Here we try to understand relationship of features with the outcome. EDA step is very important part in any type of project. EDA is nothing but Exploratory data analysis. Here we are getting insights from the data .Exploratory Data Analysis (EDA) helps in understanding the data sets by summarizing their main characteristics and plotting them visually. This step is very important before we apply and Machine learning algorithm. Check whether the data is normally distributed or not. If it is not normally distributed apply scaling technique. Represent the data set with the help of Graphs such as Line plot, Bar plot, Histogram plot, Pie chart, Box plot and try to get inferences/insights from it. Finally we should find the correlation in the data set .All the features should be independent to each other.

**DATA SET**

* The data set for the project is taken from kaggle, and the link to the dataset is provided below.

<https://www.kaggle.com/datasets/rohanharode07/webmd-drug-reviews-dataset>

* The dataset provides user reviews on specific drugs along with related conditions, side effects, age, sex, and ratings reflecting overall patient satisfaction. Data was acquired by scraping [WebMD](https://www.webmd.com/drugs/2/index) site. There are around 0.36 million rows of unique reviews and is updated till Mar 2020.
* The dataset contains 12 features.
* Data set contains 3 categorical columns and 13 numerical columns.
* **Features description**

1. Age: Range of age group within which the patient belongs.
2. Condition: Condition/symptom from which the patient is suffering.
3. Date: The date on which a patient taken that drug.
4. Drug : Name of the drug
5. Drug Id : Identity/code of drug
6. Ease of Use : Patient's 10-Star rating on the ease of use of drug
7. Effectiveness : Patient's 10-Star rating on the effectiveness of drug
8. Reviews : Patient's review
9. Satisfaction : Patient's 10-Star rating on satisfaction
10. Sex : Gender of the patient
11. Sides : Side effects of the drug
12. Useful Count : Number of users who found the review useful

**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 4.2).

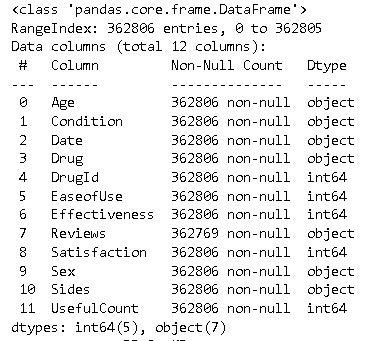


Figure 2 - data set info

**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.



Figure 3 - Statistical summary of the data set

**Distribution of each attribute**

A histogram was created for each numerical attribute. It allows us to identify the central tendency (mean, median) and the presence of outliers. It’s a skewed histogram suggesting a non-normal distribution. This information is valuable for understanding the data characteristics and potential insights it may provide. By calling the `hist()` method on the entire dataset, a histogram will be plotted for each numerical attribute, as show in Figure 4

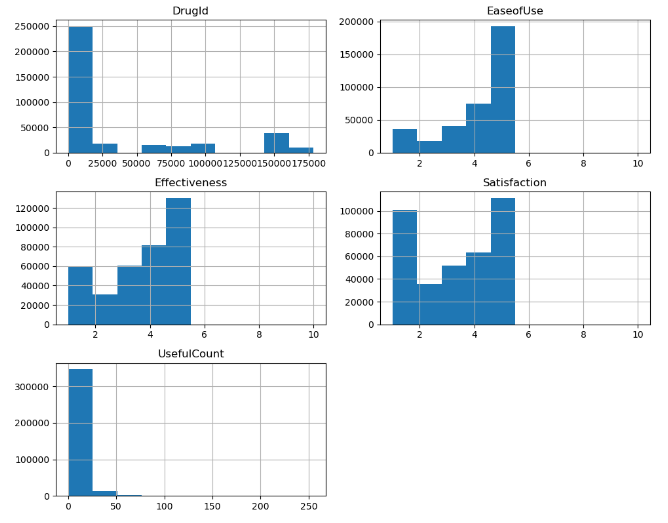


Figure 4 - Histogram for numerical attribute

* Name and race column added in the dataset.
* Missing values were detected and handled.
* Graphical representations were done to visualize various trends.
* Representation of top 20 drugs shows that (figure-5) Cymbalta tops the list followed by Lisinopril. It is to be noted that all the 20 drugs has been used by more than atleast 2000 people.
* The drug Lexapro was chosen first based on a 5-star rating for satisfaction given by the users. (Figure 6)
* As much as 50,000 users have reported other conditions followed by Pain topped the list with more than 25000 people reporting it. (Figure 7)
* The drug Lexapro was chosen first based on a 5-star rating for its effectiveness. (Figure 8)

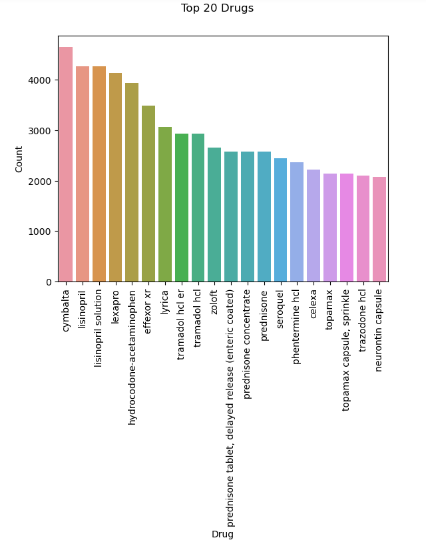


Figure 5 - Top 20 drugs based on no. of users

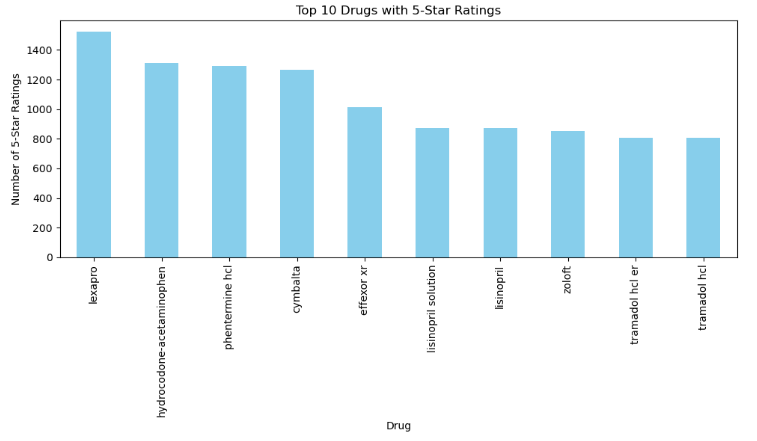


Figure 6 - Top 10 drugs with a 5-star rating for satisfaction

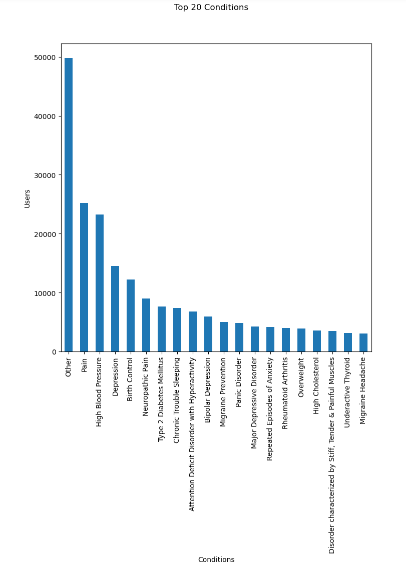


Figure 7 - Top 20 conditions

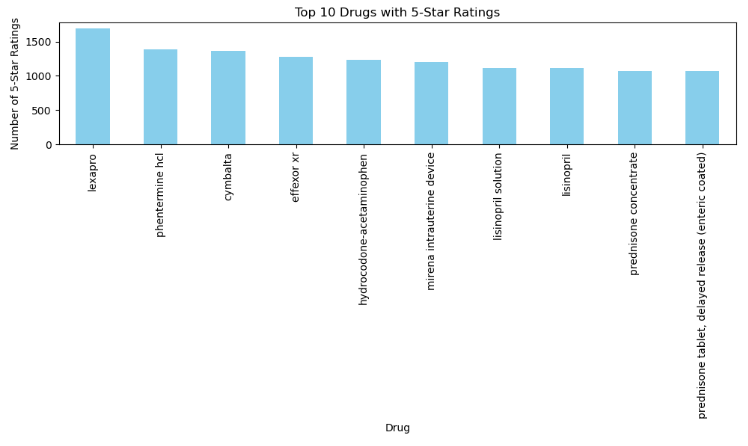


Figure 8 - Top 10 drugs with a 5-star rating for effectiveness

**EDA on Drug LISINOPRIL**

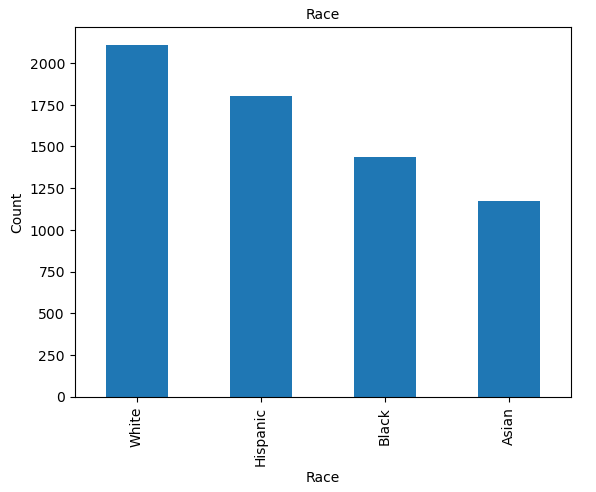
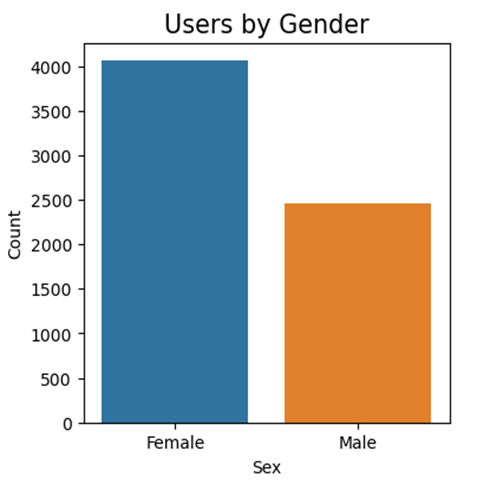
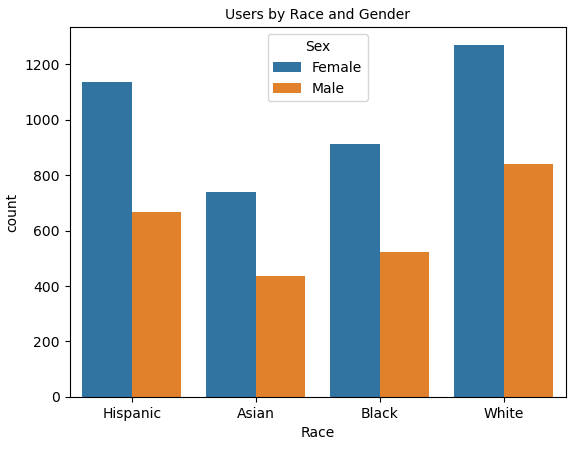
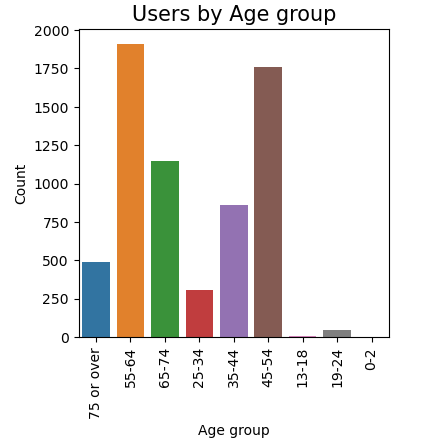
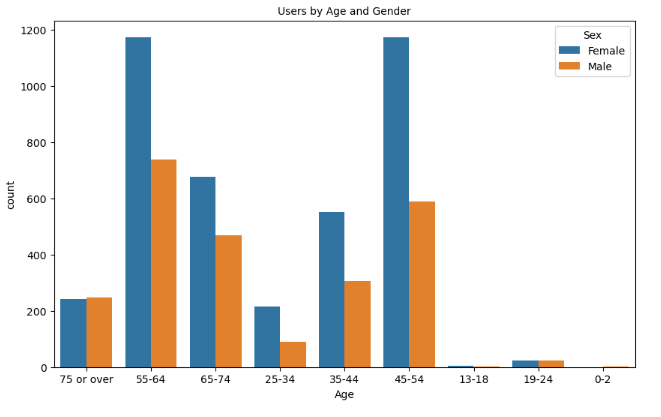
 

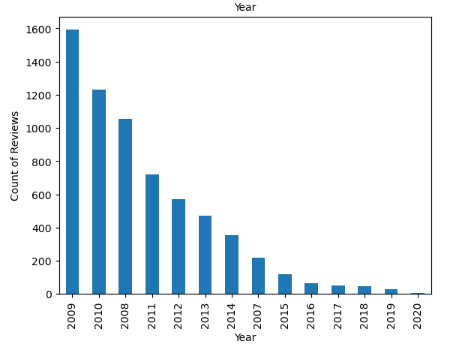
Figure 9 - Drug use based on race Figure 10 – Drug use based on gender

**Figure 11 – Drug use based on race & gender Figure 12 – Drug use by age group**

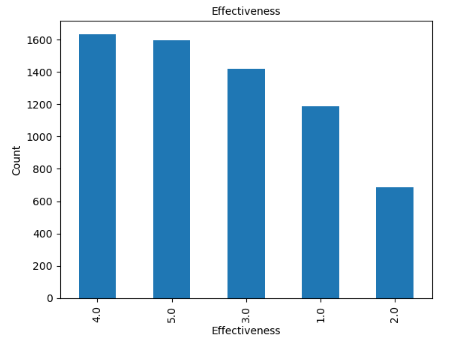
* Female users are predominant.
* White race is the predominant user.
* Age groups 55-64 are the predominant users.

**Figure 13 – Drug use age group & gender**

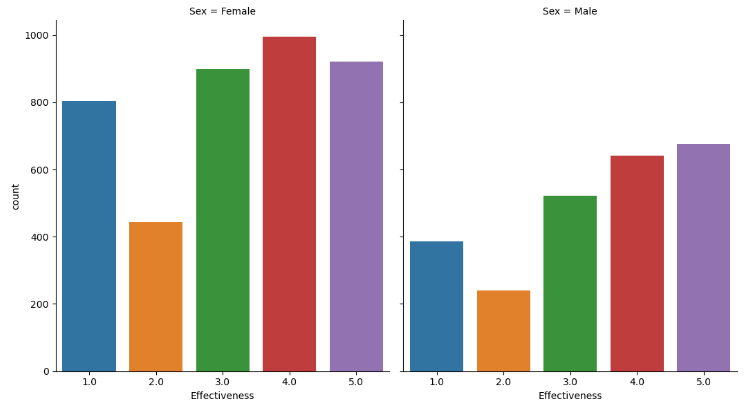


**Figure 14 – Count of yearly reviews**

* In each year, there is a gradual decrease in user reviews



**Figure 15 – Effectiveness of Lisinopril drug**



**Figure 16 – Effect of drug among both the gender**

* The drug is fairly effective with around 1600 users voting a 4 rating to it.
* Both the users have given a mixed response.
* Similarly EDA on drug Cymbalta is also done.

**DATA PREPROCESSING**

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 dataset are detailed in this chapter.

Model creation can be done by selecting a single drug. For that, the most frequently occurring drug in the dataset is selected for processing. After the detailed analysis, it was found that the drugs ‘**Lisinopril**’ and ‘**Lisinopril Solution**’ have the same drug ID. These two are the same; one is in tablet form, and the other is in solution form. Hence, it was selected for building the model. The dataset is extracted by taking the drug ID. The dataset with drug ID 6873 was only extracted from the parent dataset and checked in each column separately.

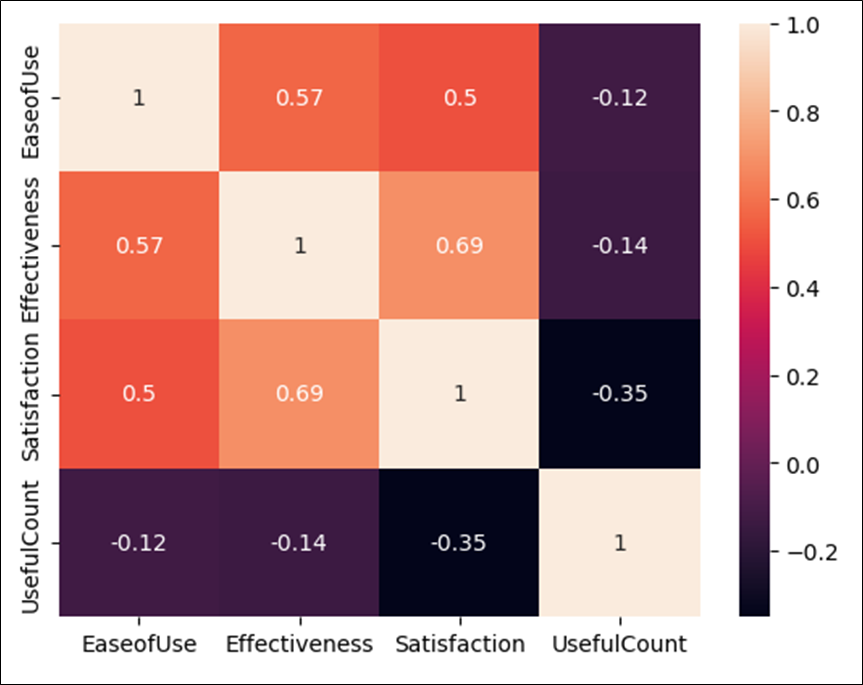
**Handling missing values**

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.

Here we treated the missing values by the method of dropping because the dataset contains each person’s reviews and satisfaction based on their experience with the drug being used.

**Correlation matrix**

A correlation-based feature selection method has been exploited in the present work to determine the optimal number of input variables (pollutants) when developing a predictive model. Statistical correlation-based feature selection algorithms compute correlations between every pair of the input variable and the target variable. The variables possessing the strongest correlation with the target variable are then filtered for further study. To compute and visualize a correlation matrix, the corr() function from the pandas library is used along with a heatmap plot from the seaborn library.



**Fig 17 - Correlation between every pair of attributes**

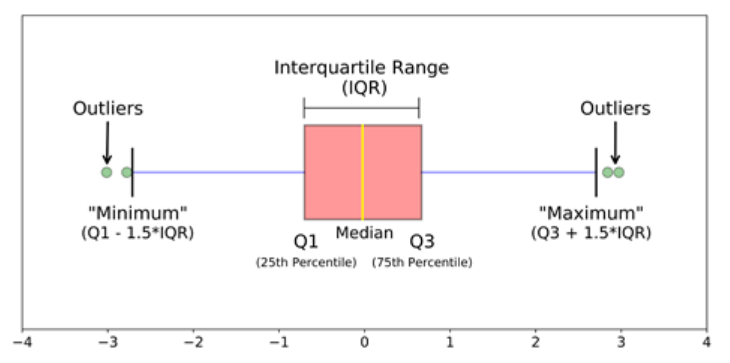
From the correlation matrix, we can infer that there is a strong relationship between effectiveness and satisfaction with the drug.

**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.

### Detection & Handling Outliers

The simplest way to detect an outlier is by graphing the features or the data points. Visualisation is one of the best and easiest ways to have an inference about the overall data and the outliers. Scatter plots and box plots are the most preferred visualisation tools to detect outliers. Interquartile range (IQR) technique method can be used to find the maximum and minimum values of data points that are outliers by calculating the boundaries. It is crucial to carefully consider the presence and impact of outliers on the analysis of the dataset. Outliers should be detected, evaluated, and handled appropriately for a cleaned dataset.



**Figure 18 – Visualization of IQR method**

In this project, the column ‘useful count’ has outliers that were calculated by the IQR method, and then it is dropped. The other numerical columns, such as ‘Satisfaction’ and 'Effectiveness', do not require any outlier treatment since they have discrete values.

**Encoding & Feature Scaling**

Both encoding and scaling are crucial parts of data transformation, aiming to enhance the quality, compatibility, and fairness of the data. Categorical variables need to be encoded into numerical values to effectively process them. This ensures that categorical data can be used as input for algorithms that only accept numerical inputs. The categorical features in the dataset are ‘Race’, ‘Age’, ‘Condition’, ‘Sex’. We are using the technique of Label Encoding to convert it into numerical values.

Feature scaling is the final step of data preprocessing. It involves transforming the independent variables of a dataset to a consistent range. By scaling the features, we ensure that no single variable dominates others, promoting fair comparisons. In this case, the drug dataset does not require any scaling.

**Feature selection/Feature deduction**

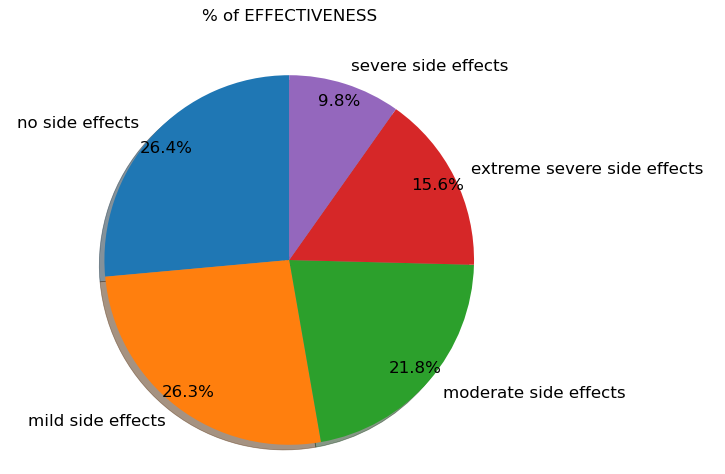
Feature selection involves selecting a subset of relevant features from the dataset, and this is a technique used for data reduction. It is often performed to remove irrelevant or redundant features from the dataset.

From the analysis, I dropped the unwanted columns from the extracted dataset, such as date, drug, review, and drug ID. The column ‘Name’ is also dropped since it does not have any effect on the side effects of the drug.

Then the ‘Sides’ column is taken for analysis. The dataset containing only the lisinopril drug is again divided into 4 based on effectiveness or satisfaction rating equal to 5, 4, 3, 2, and 1. This is done for analysing the side effects corresponding to various effectiveness or satisfactions. From the observation, it was found that the side effects shown are the same for all; there is no change with respect to the effectiveness. Hence, the ‘Sides’ column is also dropped.

**Target column**

Recoded the values of effectiveness to categorical values for making it the target column by labelling ‘extremely severe side effects', 'severe side effects', 'moderate side effects', 'mild side effects’ and ‘no side effects’ corresponding to effectiveness values equal to 1, 2, 3, 4, and 5, respectively. Then the percentages of effectiveness are drawn as a pie chart, which is shown below.



**Figure 19 – Percentage of effectiveness**

26.4% of people say that there are no side effects.

After preprocessing, the required dataset now consists of 7776 rows and 8 attributes, which includes the target feature.

**DATA MODELLING**

Predictive modeling in data science is a process of using data and statistical algorithms to make predictions about future outcomes or events. It's a subset of machine learning and is commonly used to forecast trends, behaviors, or events based on patterns and relationships discovered in data. Modeling techniques are based around the use of algorithms. We have to choose an appropriate algorithm or model to use for the prediction.

Data Modelling thus helps to increase consistency in naming, rules, semantics, and security. This, in turn, improves data analytics. The process of data modelling also enforces business rules, regulatory compliances, and government policies on the data. The emphasis is on the need for availability and organisation of data, independent of the manner of its application.

The choice of the model depends on the nature of the data and the problem you're trying to solve. Classification and Regression are two major prediction problems that are usually dealt with in data science.  The key distinction between Classification vs Regression algorithms is Regression algorithms are used to determine continuous values such as price, income, age, etc. and Classification algorithms are used to forecast or classify the distinct values such as True or False, Male or Female, Spam or Not Spam, etc. Common algorithms include linear regression, logistic regression, K-NN, decision trees, random forests and support vector machines.

In our dataset, the column ‘Effectiveness, is taken as the target column. As the initial step, the data has been split into training data and testing data, with a test size of 0.25. The subsequent task is to determine the most suitable algorithm, which can be challenging. In order to identify the most appropriate algorithm, the accuracy scores for different algorithms have been evaluated.

Different classification algorithms are:

* 1. **Logistic regression**: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.
  2. **KNN:** The k-nearest neighbours algorithm, also known as KNN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data. We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.
  3. **SVM (Support vector machine):** The main idea behind SVM is to find the best boundary (or hyperplane) that separates the data into different classes. There are different types of kernel methods in creating SVM models. They are linear, radial basis function and polynomial.
  4. **Gradient boosting classifier:** It is a popular boosting algorithm in machine learning used for classification and regression tasks. Boosting is one kind of ensemble Learning method which trains the model sequentially and each new model tries to correct the previous model. It combines several weak learners into strong learners.
  5. **Bernoulli Naive Bayes:** Naive Bayes is among the algorithms which are relatively faster than other classification algorithms. It works on the Bayes theorem of probability to predict the class of unknown data sets.
  6. **Decision tree classifier:** It uses a set of rules to make decisions, similarly to how humans make decisions. One way to think of a Machine Learning classification algorithm is that it is built to make decisions. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.
  7. **Random forest classifier:** It combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.
  8. **Ridge classifier:** It is a linear classification algorithm that is based on the Ridge regression algorithm. It is used to classify data into two or more classes based on features. The penalty term in Ridge Classifier is controlled by a parameter called alpha, which prevents overfitting.
  9. **Bagging classifier:** Bagging (or Bootstrap aggregating) is a type of ensemble learning in which multiple base models are trained independently in parallel on different subsets of the training data. Each subset is generated using bootstrap sampling, in which data points are picked at random with replacement.

Classification models were created and checked for accuracy.

|  |  |
| --- | --- |
| **Classification Algorithm** | **Accuracy score** |
| Logistic Regression | 0.501 |
| KNN | 0.552 |
| Linear SVM | 0.510 |
| RBF SVM | 0.556 |
| Polynomial SVM | 0.548 |
| Gradient boosting classifier | 0.606 |
| Bernoulli Naive Bayes | 0.271 |
| Decision tree | 0.621 |
| Random forest | 0.632 |
| Ridge classifier | 0.400 |
| Bagging classifier | 0.620 |

**Figure 20 - Comparison of accuracy score for various models**

The table illustrates a comparison of accuracy scores for different models. The selection of the model is based on the evaluation metrics and cross-validation results, aiming to identify the model with the highest accuracy. In this case, Figure 20 indicates that the Random Forest Classifier achieves the highest accuracy of 0.632, making it the best model for the given task. Hence, Random Forest Classifier is selected for deploying the model.

**Hyperparameter Tuning**

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters.

Hyperparameter tuning is basically referred to as tweaking the parameters of the model, which is basically a prolonged process. Some set of parameters that are used to control the behaviour of the model/algorithm and adjustable in order to obtain an improvised model with optimal performance is so-called Hyperparameters.

In the case of random forest there are parameters which either to increase the predictive power of the model or to make it easier to train the model. Following are the parameters we will be talking about in more detail.

We will try adjusting the following set of hyperparameters:

* n\_estimators = number of trees in the forest
* max\_features = max number of features considered for splitting a node
* max\_depth = max number of levels in each decision tree
* min\_samples\_split = min number of data points placed in a node before the node is split
* min\_samples\_leaf = min number of data points allowed in a leaf node

The two best strategies for Hyperparameter tuning are:

* Grid search CV: It involves specifying a list of values for each hyperparameter that you want to optimize, and then training a model for each combination of these values.
* Randomized search CV: In this the model selects the combinations randomly.

After the hyperparameter tuning of the random forest classifier, it was observed that the accuracy of the model was slightly reduced. Hence, I decided to go with the model without tuning, which already has an accuracy of 0.632.

**OUTCOME/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 ‘Lisinopril’ is done. The selection of the model is based on the evaluation metrics and cross-validation results, aiming to identify the model with the highest accuracy. In this case **Random Forest Classifier** achieves the highest accuracy of 0.632, making it the best model for the given task. People in the age group 55-64 mostly used this drug and within that the female ratio was higher. People with the condition ‘High blood pressure’ consumed this drug more. Over all by comparing with race also, the female ratio is higher.

**Link to code (Github link)**

<https://github.com/Sethu1994/TCSiON-Internship---Predicting-side-effects-of-drugs>

**Dataset link**

<https://www.kaggle.com/datasets/rohanharode07/webmd-drug-reviews-dataset>

**Loom video link (Video walkthrough)**

<https://www.loom.com/share/591bbd93904a44d99107aac2ab8f83bc>

**RISK**

* We can create a model that can only predict the side effects of a particular drug.
* We couldn’t analyse what would happen when people had multiple diseases.
* We are making the target based on the assumption that the drug whose effectiveness rating is high has no side effects. It will not always be correct.

**ENHANCEMENT SCOPE**

The project scope can be expanded by considering patients with multiple diseases. Prediction of side effects can be more precise and explanatory by considering such factors.

**REFLECTIONS ON THE INTERNSHIP**

**Skills learned**

* Medical related research
* Time management
* Documentation
* Stress management
* Analytical skills

**Analysis of Internship**

It helps to conduct research, analyse, understand, and bring out the best learning experience. Self-learning and webinars enable us to successfully complete the project and enrich learning.

**Conclusion**

It was an interesting journey and relevant topic in the current scenario.