|  |  |
| --- | --- |
| Internship Project Title | TCS iON 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 |
| 08-09-2023 | 20-10-2023 | | | 75 | | Virtual Internship Project | Google Colab, Excel, Chat gpt, |
| Milestone # | 2 | Milestone: | | | Day 15: Students should be able to split the dataset into training and testing datasets. They should also be able to build a classifier and fit the data to model | | |
|  |  |  |  | |  |  |  |

**CONTENTS**

[ACKNOWLEDGEMENT 2](#_Toc142154617)

[OBJECTIVE 3](#_Toc142154618)

[INTRODUCTION 3](#_Toc142154619)

[INTERNSHIP ACTIVITIES 4](#_Toc142154620)

[METHODOLOGY 4](#_Toc142154621)

[DATASET 4](#_Toc142154622)

[EXPLORATORY DATA ANALYSIS 5](#_Toc142154623)

[NULL VALUE ANALYSIS AND OUTLIER TREATING 5](#_Toc142154624)

## 

## ACKNOWLEDGEMENT

The Internship on the topic “Classification Model-Build a Model that Classifies the Side Effects of a Drug” was taken as a part of the course completion of Data Science and Analytics at ICT Academy, Kerala.

I sincerely express my gratitude to all teachers and the director of the ICT Academy, as well as TCS-ion, for encouraging and supporting me throughout the project.

I am also thankful to our project guide, mentor, Debashis Roy, who gave us constant support throughout.

## OBJECTIVE

The objective of this project is to develop a classification model that can accurately predict and classify the side effects of a specific drug based on the age, gender, and race of the patients.

## INTRODUCTION

The field of healthcare and pharmaceuticals has always been at the forefront of technological advancements, seeking innovative ways to ensure patient well-being and safety. In this context, the internship topic of "Classification Model: Build a Model that Classifies the Side Effects of a Drug" assumes paramount significance. This internship delves into the realm of data science and machine learning, where the objective is to create a predictive model that can effectively categorize the potential side effects of various drugs.

Unleashing the power of machine learning in health care redefined treatment precision, personalized care, and unlocked new possibilities. With real-time patient data, drug side effects are classified with accuracy, empowering smarter decisions for better outcomes. Discover the future of healthcare, where innovation meets value-based care.

Side effects are unwanted effects that can occur alongside the desired effects of a drug or medication. They vary based on factors like age, disease, gender, race, and health. Starting, changing, or stopping medication can trigger side effects that lead to non-compliance. Severe cases may require dosage adjustments or additional medication. Lifestyle changes can help minimize side effects. Classifying side effects for each drug is challenging, but machine learning eases the process while maintaining accuracy. Research reveals differences in drug response among racial and ethnic groups, emphasizing the need for inclusive treatment policies that accommodate individual needs.

For this project, a dataset of drug effects lacked ’Name’, ‘Race’, ‘Gender’ and ‘Age’ features. The dataset used for classification includes the drug name, side effects, etc. of different users of drugs. Datasets were combined. Supervised machine learning classifiers are used in building the model and fitting the data into the model.

## 

## INTERNSHIP ACTIVITIES

For the first fifteen days, the activities were:

●    Watched the welcome kit videos.

●    Attended the RIO – pre-assessment test.

●    Went through the day-wise plan.

●    Read the project reference material.

●    I went through the different classifications of YouTube videos.

●    Gone through tutorials explaining data analysis

●    Created a dataset with the given requirements.

●    Worked with the data set by visualizing the data.

●    Done Exploratory Data Analysis (EDA)

* Gone through tutorials on LinkedIn
* Watched videos about data mining
* Reviewed my code and resolved errors.
* Analyzed data and created more graphs
* Watched videos related to model building
* Self-Learning about Data Models and Machine Learning.

## METHODOLOGY

### DATASET

 The provided dataset comprises essential information regarding patients and their interactions with different drugs. It is structured with distinct features that capture various aspects of these interactions, offering insights into their potential effects and outcomes.

|  |  |  |
| --- | --- | --- |
| **NAME OF FEATURE** |  | **DESCRIPTION** |
| NAME |  | Name of patients |
| GENDER |  | Gender of the patients |
| AGE |  | Age of the patients |
| DRUG NAME |  | Name of various drugs |
| RACE |  | Race of patients |
| SIDE EFFECTS |  | column which contain all the side effects of drugs. |

Table 4.1: Dataset Description

The dataset includes the six features shown in the table above and 400000 records. This dataset is valuable for conducting analyses related to drug interactions, gender-based responses, racial disparities, and the prevalence of side effects associated with different drugs. Researchers and analysts can utilize this dataset to draw meaningful conclusions about the relationships between these factors and make informed decisions in the fields of healthcare and pharmaceuticals.

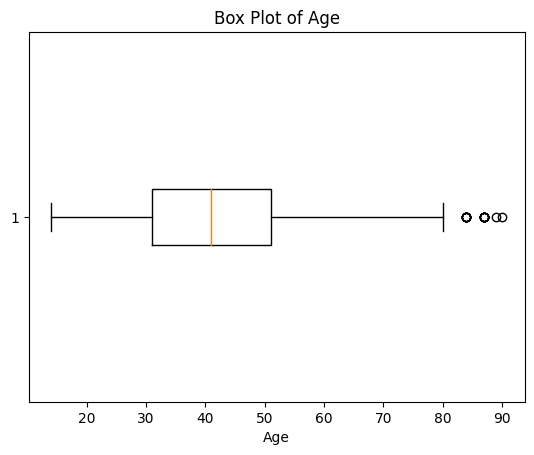
### EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis helps to gather insights, make better sense of the data, and remove irregularities and unnecessary values from the data. It also helps you prepare your dataset for analysis and allows a machine learning model to predict our dataset better by giving you more accurate results.

The types of graphs used for visualization of the dataset are Bar charts, Histogram, Pie charts, Box plot, Matplotlib and Seaborn are the two Python libraries that have been used here to generate the graphs.

### NULL VALUE ANALYSIS AND OUTLIER TREATING

## At the beginning of this EDA, null values were investigated. The age column was missing 20 records. Outliers were sought using box plots. It discovered 19 outliers. The outliers were dealt with, and on fields where age had nan values, the mean of age was replaced. There were no additional outliers or null values. The fig displays the box plot following the treatment of outliers and null values.

****

### VISUALIZATION

The figure below displays the patients' ages.  Most people are between the ages of 20 and 60.  The graph displays an ordinary bell curve.

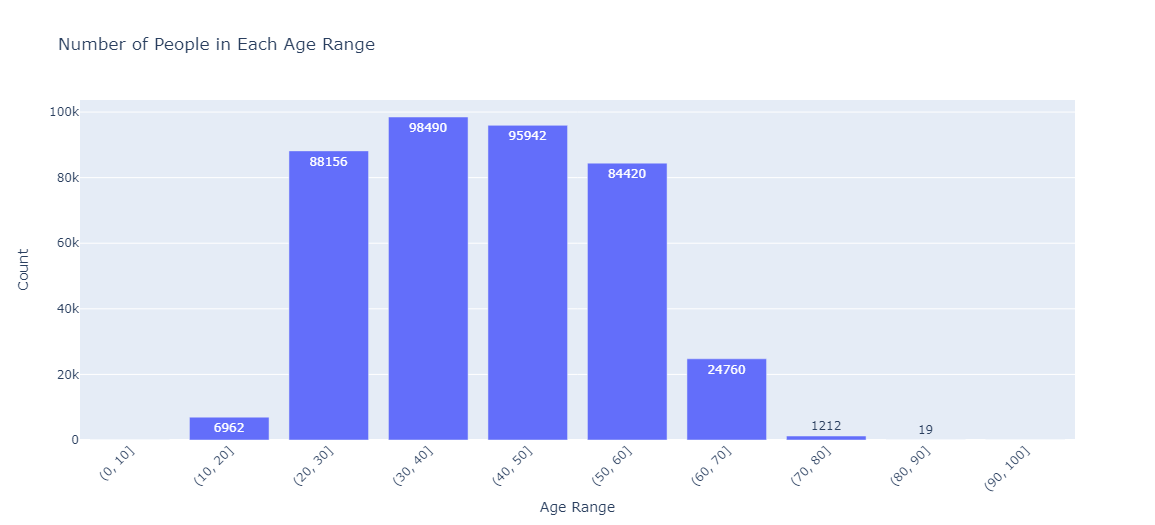


Fig:-2 Age Range

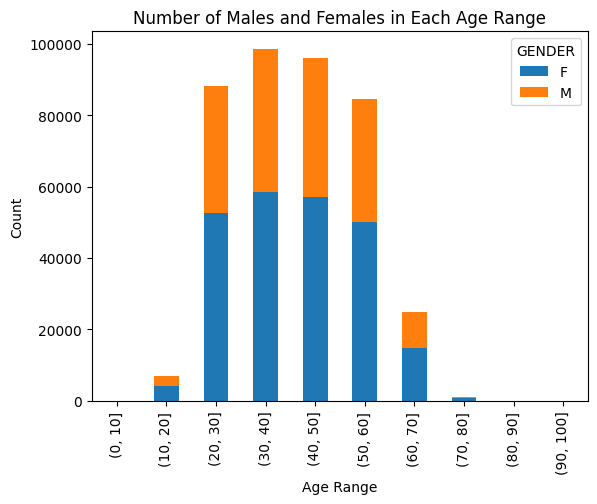
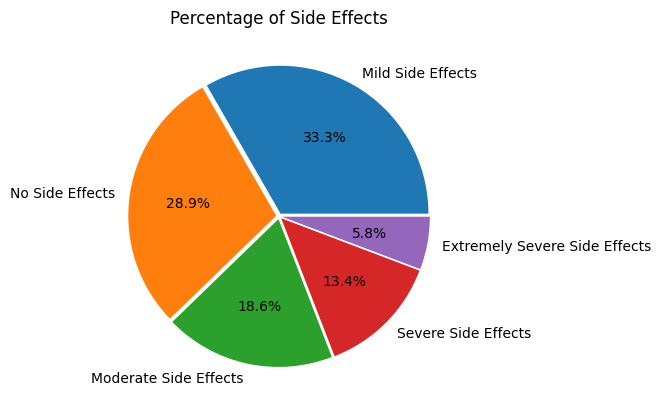
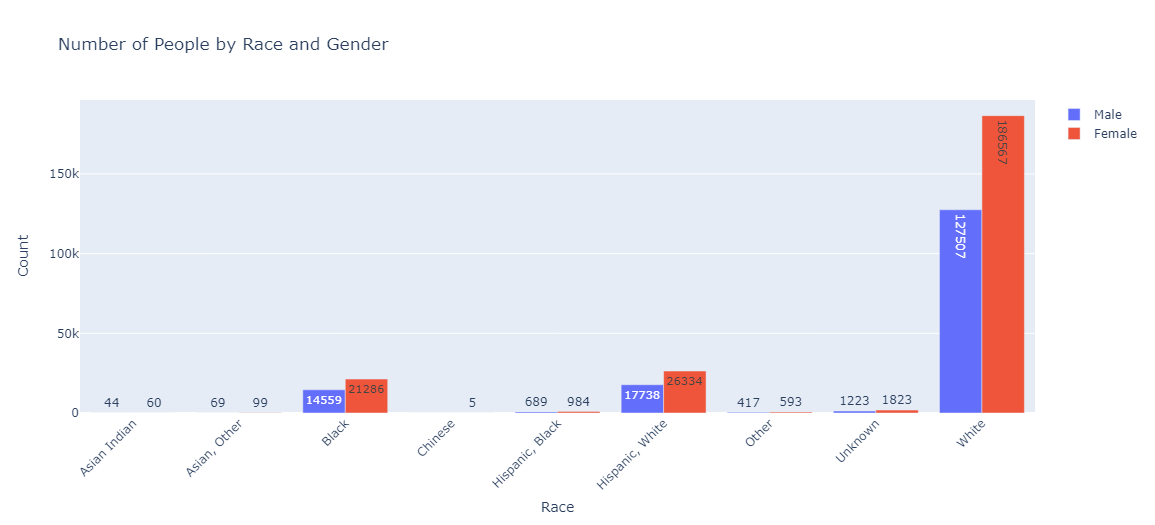


Fig:-3 Number of males and Females in each Age range



The pie chart displays the proportion of medications with varying degrees of side effects compared to those with none. The graphical representation demonstrates that 62.2% of patients who took the medications had no or just minor side effects that were not seriously damaging to their health. The race of the patients is depicted in the graph below, which is divided into males and females. patients make of 40.6% men and 59.4% women. Compared to other races, the majority of individuals are of the white race.



The most popular medications are Paxil (sometimes referred to as paroxetine and used to treat depression, panic attacks, and OCD.), Nexium, and lamictal. The top ten medications have mild to moderate negative effects.

