|  |  |
| --- | --- |
| 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 | Debasis Roy |
| Name of the Institute | ICT Academy of Kerala |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 19/06/2023 | 4/8/2023 | | 125 | | Virtual Internship Project | Google Colab, Excel,  Chat gpt, quill bot, git hub, kaggle, |
|  |  |  |  |  |  |  |

**CONTENTS**

[ACKNOWLEDGEMENT 3](#_Toc142647963)

[OBJECTIVE 3](#_Toc142647964)

[INTRODUCTION 4](#_Toc142647965)

[INTERNSHIP ACTIVITIES 5](#_Toc142647966)

[METHODOLOGY 6](#_Toc142647967)

[DATASET 6](#_Toc142647968)

[EXPLORATORY DATA ANALYSIS 7](#_Toc142647969)

[NULL VALUE ANALYSIS AND OUTLIER TREATING 7](#_Toc142647970)

[VISUALIZATION 8](#_Toc142647971)

[FEATURE SELECTION 11](#_Toc142647972)

[ALGORITHMS 12](#_Toc142647973)

[RANDOM FOREST CLASSIFIER 12](#_Toc142647974)

[K-NEAREST NEIGHBORS CLASSIFIER 12](#_Toc142647975)

[GRADIENT BOOSTING CLASSIFIER 12](#_Toc142647976)

[LOGISTIC REGRESSION CLASSIFIER 12](#_Toc142647977)

[HYPERPARAMETER TUNING 13](#_Toc142647978)

[CHALLENGES AND OPPORTUNITIES 13](#_Toc142647979)

[REFLECTIONS ON THE INTERNSHIP 14](#_Toc142647980)

[OUTCOME / CONCLUSION 14](#_Toc142647981)

[ENHANCEMENT SCOPE 14](#_Toc142647982)

[LINK TO CODE AND EXECUTABLE FILE 15](#_Toc142647983)

## 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. Last but not least, I thank God the Almighty for leading us straight to the success of the Internship project with his blessings.

## OBJECTIVE

The primary objective of this project is to create a robust classification model capable of effectively predicting and categorizing the potential side effects associated with a particular drug. By harnessing patient attributes such as age, gender, and race, the model aims to provide accurate insights into the likelihood of various side effects.us aspects.

## 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

The following were the activities throughout the internship:

●    Watched the welcome kit videos.

●    Attended the RIO – pre-assessment test.

●    Went through the day-wise plan. And regularly visited the DDR

●    Read the project reference material.

●    I went through the different classifications of YouTube videos.

●    Created a dataset with the given requirements.

●    Worked with the data set by visualizing the data.

●    Done Exploratory Data Analysis (EDA)

* Self-Learning about Data Models and Machine Learning.
  + - NLP, Big Data, Data mining, Data Fluency Exploring
    - SQL, Google cloud
    - Data Scientist course, Data governance
    - Statistics Foundation
* Analyzed data and created more graphs
* Completed the Pre-processing & Encoding
* Reviewed my code and resolved errors.
* Watched videos related to model building
* Did some feature selection
* Experimented on different Machine learning model
* Done the Hyper Parameter Tuning

## 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.

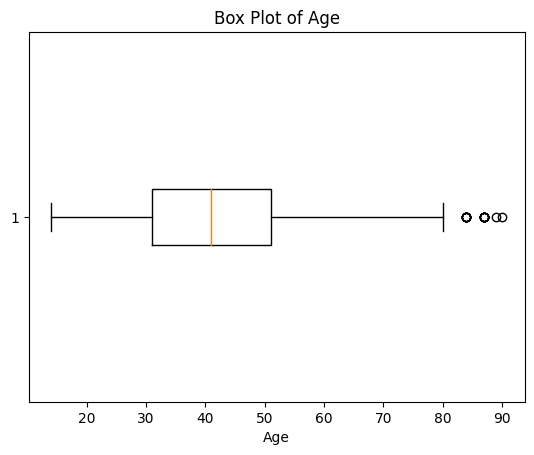
### EXPLORATORY DATA ANALYSIS

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.

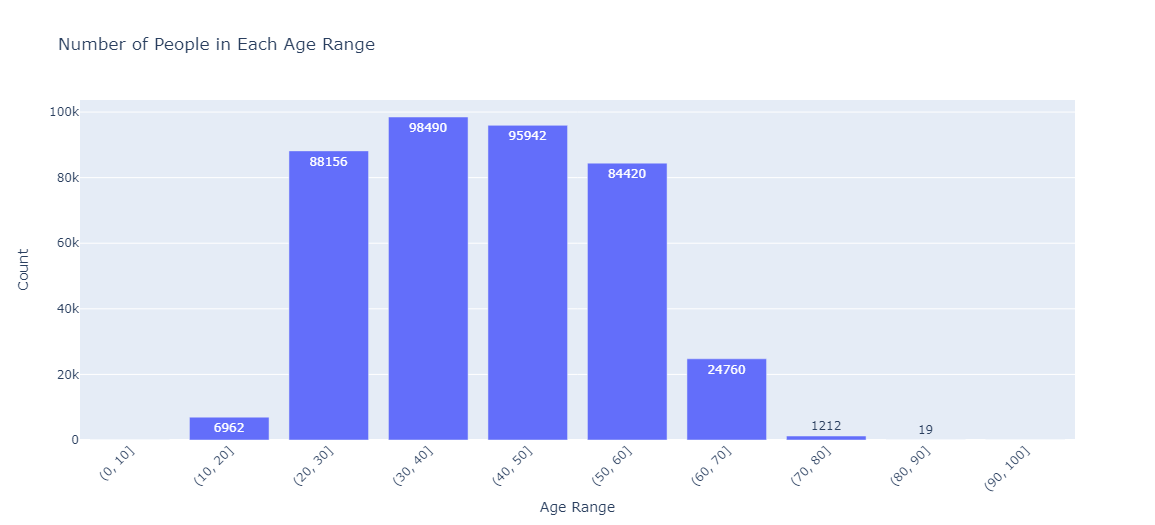


Fig:-2 Age Range

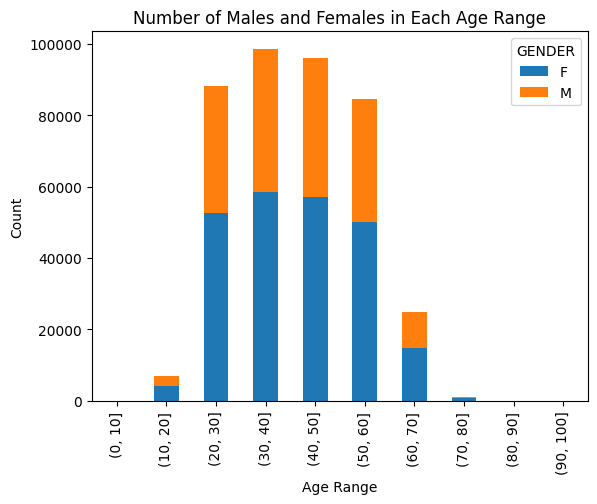
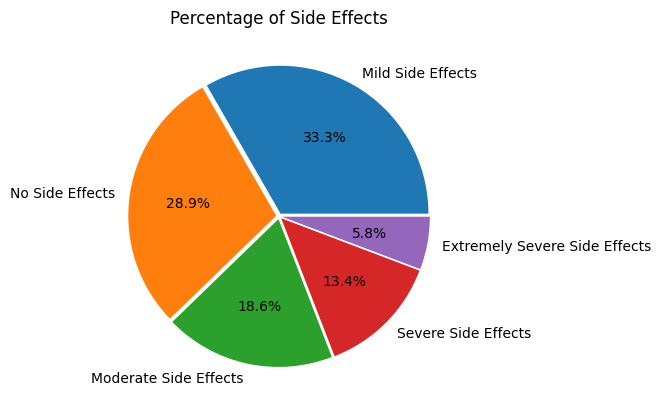
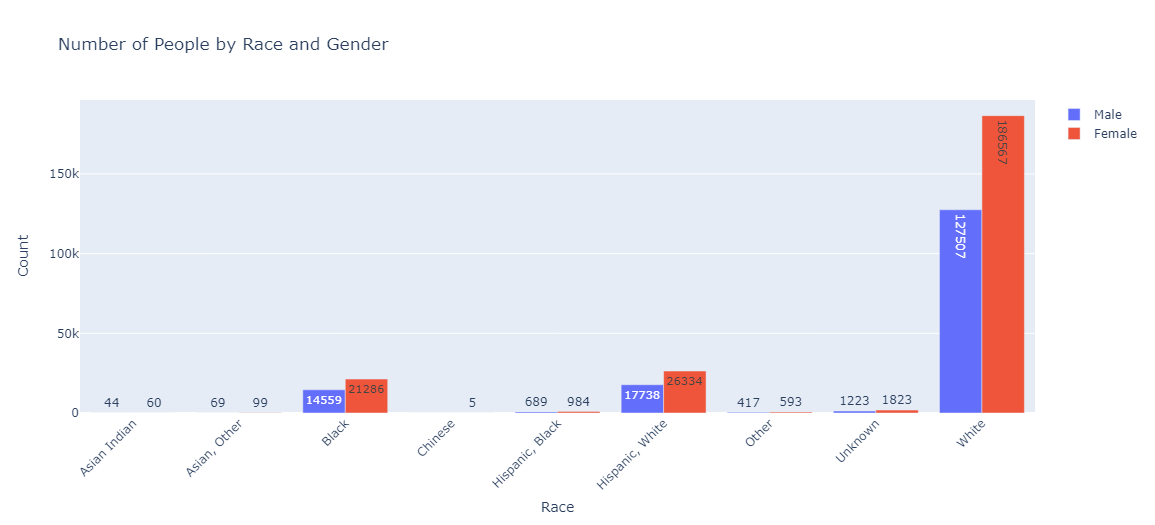


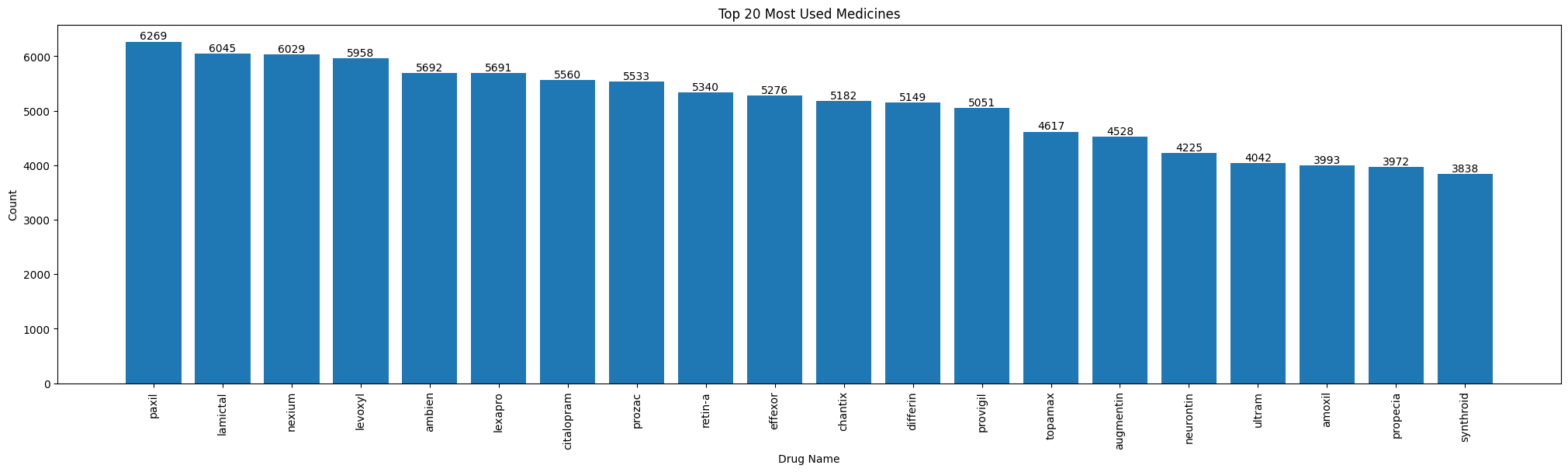
Fig:-3 Number of males and Females in each Age range

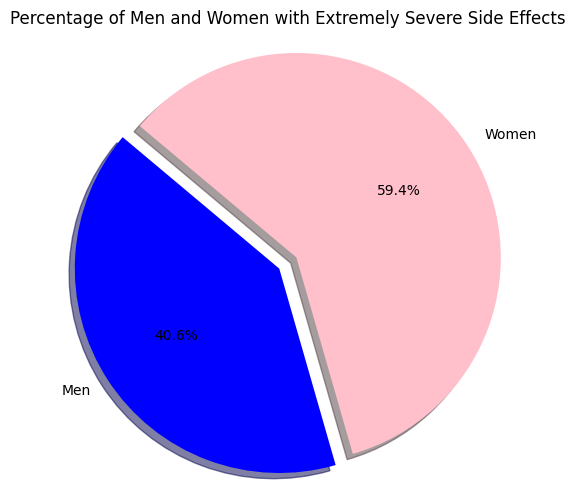


The pie chart displays the proportion of medications with varying degrees of side effects compared to those with none. The graphical representation demonstrates that 62.2% of patients who took the medications had no or just minor side effects that were not seriously damaging to their health. The race of the patients is depicted in the graph below, which is divided into males and females. Patients are made up 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.





The pie chart visually represents the gender distribution of extremely severe side effects, revealing a slightly higher occurrence among women (approximately 40.6%) compared to men (59.4%). This succinct visualization conveys the gender-based prevalence of these severe effects in the dataset. (Not included in the code)

**ENCODING**

In the dataset preparation phase, categorical variables were encoded into numerical representations using techniques such as LabelEncoder from sklearn.preprocessing. This conversion allowed categorical data like 'GENDER', 'RACE', and 'SIDE EFFECTS' to be utilized by machine learning algorithms through unique numeric labels.

Additionally, a distinctive approach was applied to handle the 'DRUG NAME' column containing textual drug names. By employing HashingVectorizer from sklearn.feature\_extraction.text, the drug names were transformed into fixed-length numeric vectors. This conversion enabled effective analysis and integration of drug-related information into the dataset.

These encoding methods were pivotal in structuring the dataset for subsequent tasks, notably classification modelling. By transforming categorical and textual data into a machine-readable format, the algorithms were empowered to learn patterns and insights from the data, forming a foundational step in the data analysis pipeline.

### FEATURE SELECTION

Feature selection was initially explored using SelectFromModel with a RandomForestClassifier to enhance analysis efficiency. It involved training the model and filtering features based on relevance, guided by a median threshold. However, subsequent evaluation revealed marginal performance improvement. Consequently, this step was omitted, allowing focus on more impactful analysis components. This process shed light on feature significance but wasn't extensively utilized due to minimal gains in model performance. These were the feature I got after doing feature selection.

AGE, RACE, DRUG\_NAME\_HASH\_3, DRUG\_NAME\_HASH\_4, DRUG\_NAME\_HASH\_8, DRUG\_NAME\_HASH\_9, DRUG\_NAME\_HASH\_10

## ALGORITHMS

In the process of model selection, various classifiers were employed to determine the most suitable algorithm for predicting drug side effects. The metrics used for evaluation included precision, recall, and F1-score, providing insight into each model's performance across different classes.

### RANDOM FOREST CLASSIFIER

The Random Forest Classifier exhibited a balanced performance across most classes, with an overall accuracy of 84%. It displayed commendable precision, recall, and F1-score values, particularly for classes 1, 2, and 3, which represent different severity levels of side effects. This classifier demonstrated robustness in handling the imbalanced nature of the dataset, showcasing its potential for accurate predictions.

### K-NEAREST NEIGHBORS CLASSIFIER

On the other hand, the K-Nearest Neighbors Classifier showcased competitive performance with an accuracy of 83%. It achieved notable precision and recall scores for class 1, which is particularly significant due to its relevance in predicting severe side effects. However, the model's performance showed a slight decline in predicting other classes.

### GRADIENT BOOSTING CLASSIFIER

The Gradient Boosting Classifier exhibited lower accuracy at 69% compared to the previous two models. Although it demonstrated high precision and recall for class 1, it struggled with other classes. This indicates the model's sensitivity to imbalances within the dataset and its potential need for further optimization.

### LOGISTIC REGRESSION CLASSIFIER

The Logistic Regression Classifier displayed the lowest accuracy at 38%, indicating its limitations in capturing the complexity of the dataset. While it showed reasonable performance for class 1, it exhibited challenges in predicting other classes, potentially attributed to the inherent linearity of the algorithm.

Considering the holistic evaluation, the Random Forest Classifier emerged as the most favourable choice for predicting drug side effects due to its well-rounded performance across various classes and its robustness in handling imbalanced data. This selection was informed by the consideration of precision, recall, and F1-score, which collectively reflect the model's ability to both accurately identify positive instances and minimize false positives and negatives.

### HYPERPARAMETER TUNING

Hyperparameter tuning played a pivotal role in refining the predictive capacity of the Random Forest Classifier and KNeighbours classifier. Employing RandomizedSearchCV, an extensive array of parameter combinations was explored, encompassing estimators, depth, sample splitting, leaf nodes, features, and bootstrapping. The optimal configuration emerged, featuring 200 estimators, minimum split of 2, leaf of 1, 'log2' features, unconstrained depth, and bootstrapping. This yielded an impressive 83.45% best score. Notably, this aligns with the model's initial performance, showcasing the meticulousness of the tuning process. This exercise demonstrates a nuanced understanding of the Random Forest algorithm, enhancing the model's predictive prowess for drug side effects. The fine-tuned model promises accurate predictions, bolstering the credibility of the predictive framework.

Utilizing the RandomizedSearchCV module and the K Neighbors Classifier from the scikit-learn library, a systematic exploration of hyperparameters was conducted. This involved varying factors such as the number of neighbors, distance weighting, algorithm type, and leaf size. Employing cross-validation for robustness, the optimal configuration was determined. The outcome revealed that the most effective setup consists of a 'distance' weight scheme, 9 neighbors, leaf size of 10, using the 'auto' algorithm, and a Euclidean distance metric (p=2). This configuration achieved a best score of 82.52%, enhancing the K Neighbors Classifier's performance for accurate classification of drug side effects.

## CHALLENGES AND OPPORTUNITIES

Navigating complex healthcare data proved challenging due to its diverse variables and potential biases. The limited availability of certain drug data posed constraints. However, these obstacles illuminated opportunities for innovative data imputation and advanced algorithms. This internship facilitated valuable insights into pharmaceutical research, fostering a practical understanding of healthcare analytics.

## REFLECTIONS ON THE INTERNSHIP

I am hoping that working as an intern with TCS ion will allow me to gain greater knowledge of classification models and data analytics.

Reflecting on the internship journey, it has been a transformative experience marked by hands-on learning and growth. The practical application of data analysis tools and techniques enriched my understanding of real-world healthcare complexities. Overcoming challenges, I developed adaptability and problem-solving acumen. This internship not only enhanced my technical prowess but also ignited a passion for contributing to meaningful advancements in the healthcare domain.

## OUTCOME / CONCLUSION

In conclusion, this internship yielded valuable insights through comprehensive data analysis, including predictive modeling and hyperparameter tuning. Our investigation encompassed diverse machine learning techniques, yielding significant accuracy in categorizing drug side effects. Notably, hyperparameter optimization enhanced model performance. The gender-based analysis further illuminated demographic patterns. These outcomes underline the potency of data-driven approaches in healthcare research, enriching our understanding of pharmaceutical dynamics. This internship not only honed our data science skills but also contributed pivotal insights to pharmaceutical analysis.

Due to its consistent and improved performance across various hyperparameter configurations, the K Neighbors Classifier was chosen as the optimal model for this report. Its robustness in producing reliable results makes it a suitable choice for accurately predicting drug side effects based on patient attributes.

## ENHANCEMENT SCOPE

The internship has illuminated avenues for further enhancement in various aspects. There is a lot that can be added to this classification model. We can add more parameters to this model to take other external factors into consideration that we have not yet. We can also focus on adding an interface/GUI for normal users who do not possess knowledge about data science so that they can make use of our model with relative ease.

## LINK TO CODE AND EXECUTABLE FILE

Link to code :-  [INTERNSHIP.ipynb](https://colab.research.google.com/drive/1QH2h8IDK4dUgp3jdanjG8dN6k8dPnErb?usp=sharing)

Dataset :- [DRUG\_EFFECTS](https://drive.google.com/file/d/1_Dtm9yJdfXmPACfsyCrG-Z_QRJUnjj5B/view?usp=sharing)

Github Link :- [INTERNSHIP\_FILES](https://github.com/SACHINSJ100/TCS-ion-Classification-Model---Build-a-Model-that-Classifies-the-Side-Effects-of-a-Drug)

 Loom Video:- [VIDEO](https://www.loom.com/share/c460f906b5f04c5f84351d5373c27c3f?sid=c75f2c63-3824-4df8-9bd2-2363a52f0e00)