|  |  |
| --- | --- |
| 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 |
| 12/11/2023 | 8/12/2023 | | 125 hrs | | Virtual Internship Project | Google Colab,Jupyter Notebook |
| Milestone # | 3 | Milestone: | | Completed TCS iON RIO- 125 Classification Model - Build a Model that Classifies the Side Effects of a Drug | | |

**TABLE OF CONTENTS**

* Acknowledgement
* Abstract
* Overview
* Objective
* Introduction
* Methodology
* Internship Activities
* Charts, Table, Diagrams
* Algorithms
* Predictive Modelling and Fine Tuning
* Conclusion

**ACKNOWLEDGEMENT**

I would like to take a moment to express my heartfelt appreciation to all those who have played a pivotal role in the development of this report. This document signifies a significant milestone in my ongoing project, and it wouldn't have been possible without the invaluable contributions and support of numerous individuals and organizations. The internship, focused on the topic of "Classification Model - Building a Model to Classify Drug Side Effects," was undertaken as a part of the course requirements for Data Science and Analytics at ICT Academy, Kerala.

First and foremost, I extend my sincere gratitude to all the teachers and the director of ICT Academy, as well as TCS-ion, for their unwavering encouragement and support throughout the project. Your guidance and belief in me have been instrumental in this journey. I want to express my deep appreciation to my dedicated mentor, Debashis Roy, whose guidance, expertise, and unwavering commitment to the project have been indispensable in shaping this report. Your mentorship has been truly invaluable, and I am immensely grateful for the profound insights you have shared.

I acknowledge and deeply appreciate the contributions of the dataset participants, whose valuable data forms the foundation of my research. Your willingness to share your information has been vital to the success of my project. Lastly, I wish to express my heartfelt gratitude to my friends and family for their unwavering support and encouragement throughout this journey. Your belief in me has been an unwavering source of motivation and strength.

**ABSTRACT**

The project focuses on the development of a robust and accurate classification model aimed at predicting drug side effects. With the increasing complexity of pharmaceutical data and the critical need for early detection of adverse reactions, this endeavor holds paramount importance in the healthcare and pharmaceutical sectors. The project utilizes advanced data science and machine learning techniques to create a model capable of classifying potential side effects associated with various drugs.

Ultimately, the project aims to provide a valuable tool for healthcare professionals, pharmaceutical researchers, and regulatory agencies to proactively identify potential drug side effects. By contributing to early intervention and patient safety, this classification model has the potential to revolutionize drug development and healthcare decision-making processes.

For my current internship project, I've obtained a dataset containing information on 400,000 patients. This dataset includes details such as the names of the drugs administered, patient names, ages, genders, drug names, races, and recorded side effects. The overarching goal of this project is to conduct a comprehensive analysis, encompassing exploratory data analysis (EDA), data preprocessing, correlation analysis, feature reduction, feature engineering, and classification on the data then performing classification algorithms on this data and select the best model out of it.

Through rigorous experimentation and evaluation, this project strives to demonstrate the model's effectiveness in accurately predicting drug side effects, thereby improving patient care, reducing healthcare costs, and advancing drug safety standards.

**OVERVIEW**

The dataset at hand is dedicated to the classification of side effects associated with a variety of drugs, with a specific emphasis on delivering actionable insights for the pharmaceutical and healthcare sectors. This dataset comprises 400,000 records and encompasses six key columns, including patient age, gender, drug name, and reported side effects. My primary objective revolves around the classification of these side effects for individual drugs within the dataset, a task that holds great significance in improving drug safety and patient care within the healthcare industry.

**OBJECTIVE**

The objective of this project is to construct a robust classification model specifically designed to categorize the side effects of a particular drug based on age, gender, and race. In addition to analyzing and predicting side effects, this project also aims to assess the drug's overall effectiveness. By focusing on these critical aspects, the project seeks to provide valuable insights into how drug reactions vary across different demographic factors, ultimately contributing to more personalized and informed healthcare decision-making.

**DOMAIN KNOWLEDGE**

The different tools used in this project:

• Python

• Machine Learning

**INTRODUCTION**

In an era marked by remarkable advancements in healthcare and pharmaceuticals, the development of novel drugs has achieved unprecedented heights, holding immense potential to alleviate suffering and enhance lives. However, alongside this expanding array of pharmaceutical options, there arises a critical imperative to comprehend and effectively manage the associated side effects.

Side effects, typically viewed as undesirable secondary effects occurring alongside a drug's intended therapeutic benefits, exhibit variability among individuals based on factors such as their health conditions, age, weight, gender, ethnicity, and overall well-being. These side effects may manifest when initiating, adjusting dosages, or discontinuing a drug regimen, and they sometimes lead to non-compliance with prescribed treatments. In cases of severe side effects, dosage adjustments or alternative medications may be necessary, while lifestyle or dietary changes can also play a role in mitigating these effects.

Against this backdrop, the Data Science and Analytics Internship embarked on a transformative mission to construct a robust predictive model, with the aim of selecting the best-performing model among the options available. The primary objective of this internship report is to comprehensively document the journey undertaken, encompassing data collection, preprocessing, and the construction of machine learning models. The ultimate goal is to empower healthcare professionals and researchers with a systematic tool for classifying and predicting the side effects of drugs.

In an era where precision medicine and patient well-being take center stage, this project serves as a guiding light, illuminating the intricate landscape of drug-related side effects. By simplifying the process of classification, it contributes to safer and more informed healthcare decisions. Specifically, this internship focuses on the creation of a classification model dedicated to categorizing the side effects of a chosen drug within the dataset, thereby facilitating a more profound understanding of pharmaceutical outcomes.

**METHODOLOGY**

**Data Collection and Preprocessing**

Data Sources: Describe the sources from which the dataset was obtained, including any external databases, data providers, or APIs.

Data Cleaning: Explain the steps taken to clean and prepare the dataset, addressing issues such as missing values, duplicates, and inconsistent formatting.

Feature Engineering: Detail any feature engineering techniques applied to create new variables or transform existing ones to enhance the dataset's quality and relevance.

**Exploratory Data Analysis (EDA)**

Summary Statistics: Provide an overview of the dataset's key statistics, including measures of central tendency, variability, and distribution.

Data Visualization: Describe the visualizations used to explore data patterns, such as histograms, bar plots, box plots, and count plots.

Correlation Analysis: Discuss the analysis of feature correlations to identify potential relationships between variables.

**Handling Missing Data**

Missing Data Assessment: Find the missing values in the dataset.

Imputation Techniques: Describe the strategies used to handle missing data, such as mean imputation, median imputation, or advanced imputation methods like regression imputation and finally handle the missing data.

**Outlier Detection and Treatment**

Outlier Identification: Explain the techniques employed to identify outliers in the dataset, such as the Z-score method or the Interquartile Range (IQR) method.

Outlier Handling: Describe the approaches used to handle outliers, including removal, transformation, or the use of robust algorithms.

**Feature Selection and Engineering**

Feature Importance: Discuss how feature importance was assessed, whether through statistical tests, machine learning models, or domain knowledge.

Feature Scaling: Check the scope of feature scaling and feature engineering in the dataset.

**Model Development and evaluation**

Training and Testing: Detail the process of splitting the dataset into training and testing sets.

Model Selection: After finding the metrics of all models, decide on the best one.

Cross validation: Check cross-validation techniques on the selected model.

Hyperparameter Tuning: Describe efforts to optimize model hyperparameters for improved performance.

**Results Interpretation**

Inference: Provide insights and interpretations of the model results, including the significance of features and their impact on predictions.

Conclusion: Mention in brief the conclusion that we get from this project.

Limitations: Discuss any limitations encountered during the methodology, such as data constraints, model assumptions, or computational resources.

**INTERNSHIP ACTIVITIES**

### Dataset Acquisition and Preparation:

**Imported necessary Python libraries**

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

· pandas

· numpy

· matplotlib.pyplot

· seaborn

· sklearn.preprocessing

· sklearn.model\_selection.train\_test\_split

**Importing the data**

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

**Data source**

Github Link:[TCSiON-Internship-RIO-125-Classification-Model-/INTERNSHIP TCS (1).ipynb at main · Themaqtron/TCSiON-Internship-RIO-125-Classification-Model- (github.com)](https://github.com/Themaqtron/TCSiON-Internship-RIO-125-Classification-Model-/blob/main/INTERNSHIP%20TCS%20(1).ipynb)

**Dataset Description:**

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

Age: It is a column which contains the age of the patients, but has some data entry error

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

Side effects: This categorical column contains all the side effects of drugs.

Gender: This is the gender of patients namely male or female.

Race: It refers to the classification of humans into distinct groups based on shared physical characteristics, such as skin color, hair texture, and facial features.

### Exploratory Data Analysis (EDA):

**1. Univariate Analysis:**

1. **Name**:

Name of the patient doesn't give any contribution for the modeling hence it is dropped.

1. **Age**:

Age is a skewed distribution, and majority patients are in the age group of 24-55.

1. **Race:**

Among the patients majority are whites (78.5%) and Chinese are minority (negligible %)

1. **Gender:**

59.4% are females, 40.5% are males, and females are more.

1. **Drug name:**

Top 10 drugs used are paxil, lamictal, nexium, levoxyl, lexapro, ambien, retin-a, citalopram, prozac, effexor. Out of this paxil is the most preferred drug. Majority drugs are used to deal with depression and anxiety disorders, skin problems, esophagus issues. 101 drugs are least preferred ones out of a total of 502 drugs in the dataset.

1. **Side effects:**

It is the target column in the dataset. Patients with mild effects are the 33.3%, people with extreme side effects are 5.8 % of total patients.

**2. Bivariate Analysis:**

1. **Gender Vs Age:**

Majority patients are in the age group of 20-59, and out of that too females are higher than males in all groups.

Irrespective of gender people of same age are suffering from illness-got this inference from boxplot.

1. **Gender Vs Side Effects:**

Irrespective of the gender, patients with mild side effects are max, and those with extreme.

1. **Gender Vs Race:**

Majority patients are whites, Chinese are minority.

1. **Gender Vs Drug Name:**

Top 10 drugs used by males and females are same, this indicates both have similar health issues, mostly they are affected by depression or mental health issues.

1. **Drug Name Vs Race:**

These top 10 drugs are generally preferred by whites, Hispanic whites and blacks, these people suffer with common illness that is depression.

1. **Drug Name Vs Side Effects:**

Ambien, citalopram, nexium, retin-a are the ones having no case of extreme side effects, but all have severe side effects only nexium out of this is having least patients with severe side effects too. Nexium, paxil and lamictal are the drugs which are either having very mild side effects or no side effects at all.

1. **Age Vs Race:**

China and Asian Indian have only young patients, white, Hispanic white and blacks have patients from young age to senior citizens.

1. **Age Vs Side Effects:**

Severity trends: across all age groups, "mild side effects" are consistently the most reported, followed by "moderate side effects." however, the prevalence of "severe side effects" and "extremely severe side effects" tends to increase with age, with the 60+ age group having a relatively higher proportion of these severe side effects.

1. **Drug Name Vs Age:**

Lamictal (lamotrigine) is a widely used anticonvulsant prescribed primarily to patients in the age

group of 40 to 55. Its primary role is in managing epilepsy and bipolar disorder, providing much-

needed stability to those with these conditions.

In contrast, Retin-A finds popularity among younger individuals and is primarily employed to

address various skin concerns such as acne, wrinkles, hyper pigmentation, and rough texture. Its

cosmetic benefits make it a go-to choice for individuals seeking youthful and clear skin.

Citalopram, on the other hand, is a medication that tends to be favored by older patients. It

serves as an antidepressant, offering relief to those battling depression and specific anxiety

disorders. Its effectiveness in alleviating these conditions has made it a trusted option among

healthcare professionals.

Similarly, Effexor is another antidepressant that tends to be preferred by older patients. It is

prescribed to manage major depressive disorder, generalized anxiety disorder, and panic

disorder, providing much-needed support to those grappling with these mental health

challenges.

Lastly, Levoxyl (levothyroxine) is a thyroid hormone replacement medication primarily

prescribed to treat hypothyroidism. It plays a vital role in restoring thyroid hormone levels to

normal, thereby helping individuals maintain proper metabolic function. Each of these

medications addresses distinct health needs, catering to a wide range of patients across

different age groups.

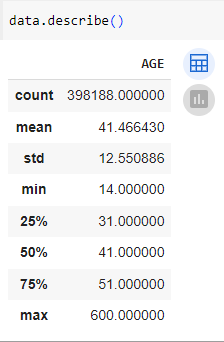
**3. Multivariate Analysis:**

**1. Drug Name Vs Gender Vs Age**

Younger patients are having skin concerns (males are higher than females), older patients are suffering from depression irrespective of gender.

**4. Summary Statistics:**

Used the describe function to get the statistical summary of numerical feature in the dataset.



Other than describe function has tried all basic EDA: head, info, shape, unique, data types.

**5. Preprocessing:**

1. Dealt with duplicate values:

Found the duplicate values in the dataset and dropped it.

1. Finding and handling missing values:

Found the missing values in the dataset, it was found only in 1 column ‘age’. After studying about the various methods of filling missing values, finally decided to fill it with median, also because the age column is showing skewed distribution.

1. Outliers Detection and Handling:

Outliers are data points that significantly deviate from the typical values in a dataset, potentially indicating errors or unusual observations.

Found outliers only in age column, used IQR method and found the outliers, also made use of boxplot for visualization. The outliers in this column show a clear indication that it was due to data entry errors.

After studying about outliers and the various methods to handle them like trimming, capping, flooring, winsorization, imputation decided to drop values which are above 100.

1. Encoding:

Encoding refers to the process of converting categorical data (non-numeric data) into a numerical format that machine learning algorithms can work with. There are several types of encoding methods in data science.

Label Encoding: This method assigns a unique integer or label to each category in a categorical variable. It is suitable for ordinal data, generally used for target column.

One-Hot Encoding: One-hot encoding creates binary columns (0 or 1) for each category in a categorical variable. Each category gets a separate column, and the presence of the category is indicated by a 1 in its respective column. This method is suitable for nominal data, where there is no inherent order among categories. It can be done with sk learn pr get dummies code.

Frequency Encoding, Ordinal Encoding are examples of other types of encoding.

As my dataset is quite large, carrying out One Hot Encoding was very time consuming. Hence I encoded Gender, Race, Drug Name and Side Effects with Label Encoding.

1. Feature Scaling:

Feature Scaling refers to the process of preparing and transforming data to ensure that it is suitable for analysis or modeling. Data scaling is important because it can have a significant impact on the performance of various machine learning algorithms. Scaling ensures that all features (variables) contribute equally to the analysis, preventing features with larger magnitudes from dominating the results. Here are some common types of data scaling in data science:

1. Min Max Scaling

2. Standardisation

3. Normalization

4. Log Transformation and others

My dataset contains only 1 numerical feature and it is discrete. So I decided not to do scaling, instead preferred algorithms which are not affected by magnitude of features.

1. Feature Engineering:

Age is a discrete variable in our dataset and i have grouped it for visualization purposes as it offered me better inference.

1. Correlation:

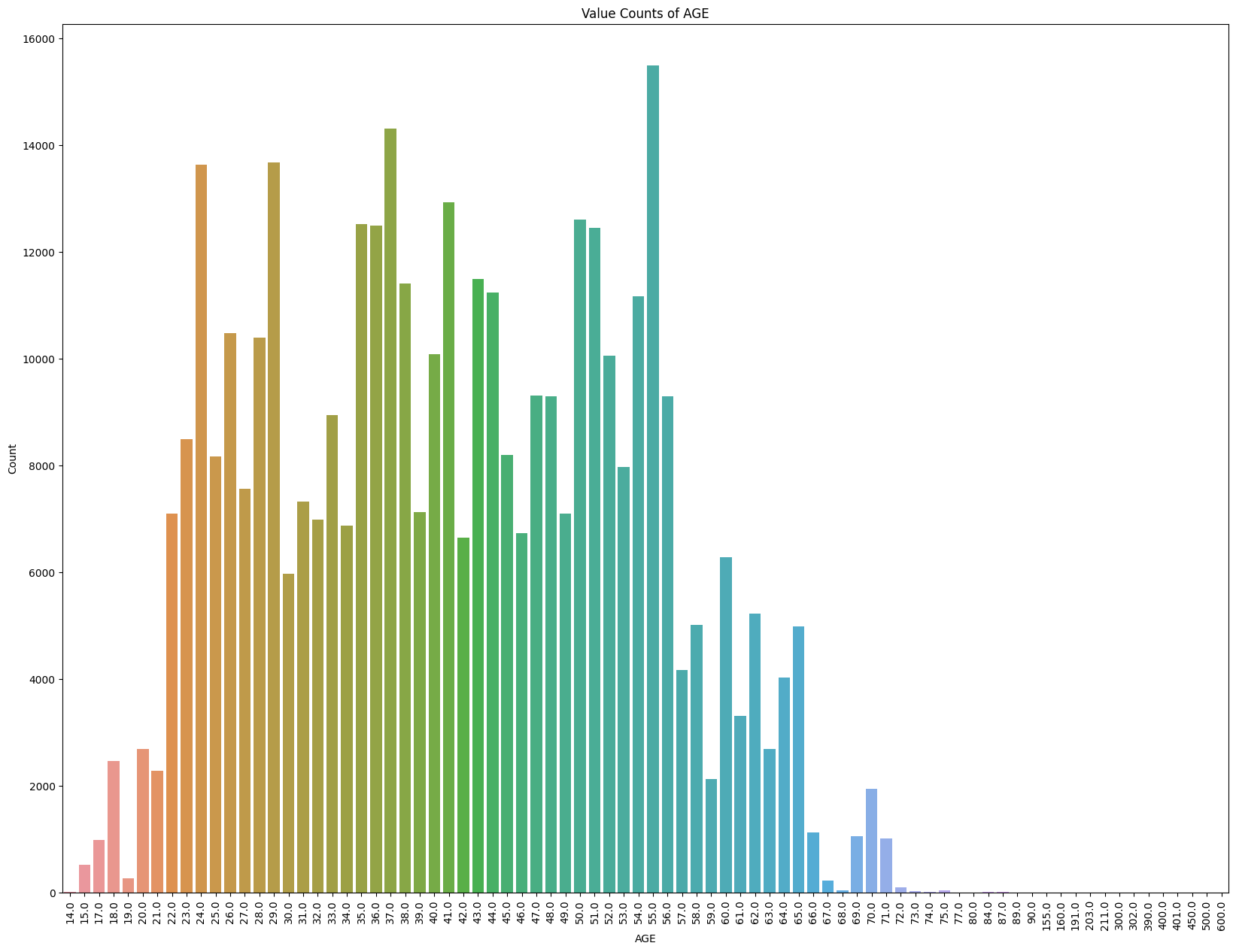
Correlation is a statistical measure that quantifies the degree to which two variables are related or associated with each other. It assesses the strength and direction of the linear relationship between two variables. Correlation is commonly used in data analysis and statistics to understand how changes in one variable may affect another.

I have checked the correlation between all variables and found no significant relation between the features. So I haven’t dropped anything.

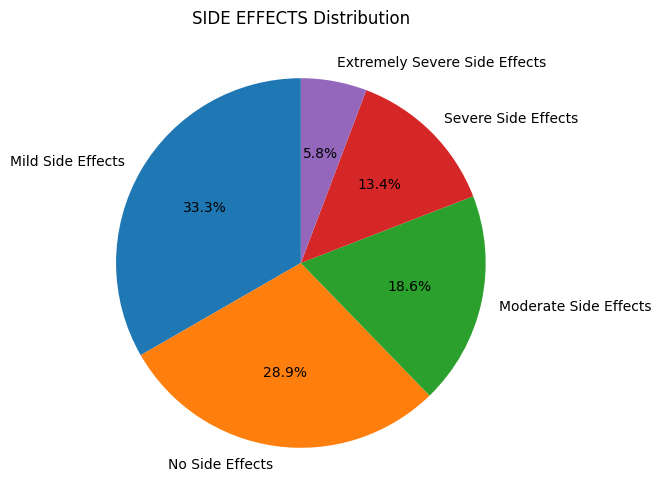
**CHARTS, TABLE, DIAGRAMS**

* **UNIVARIATE PLOTS:-**

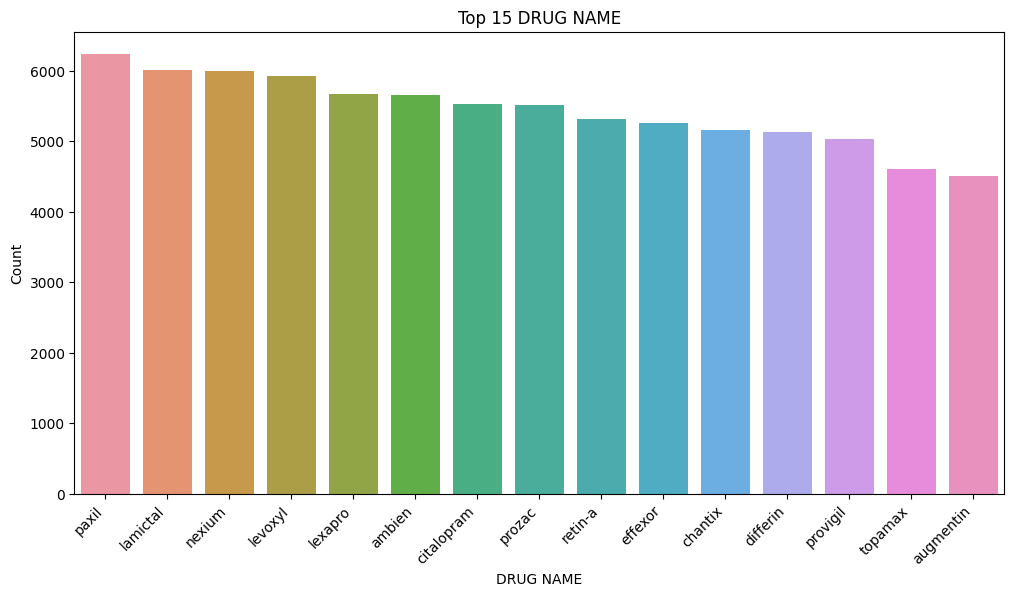
**1. AGE:**

****

**2. SIDE EFFECTS**:

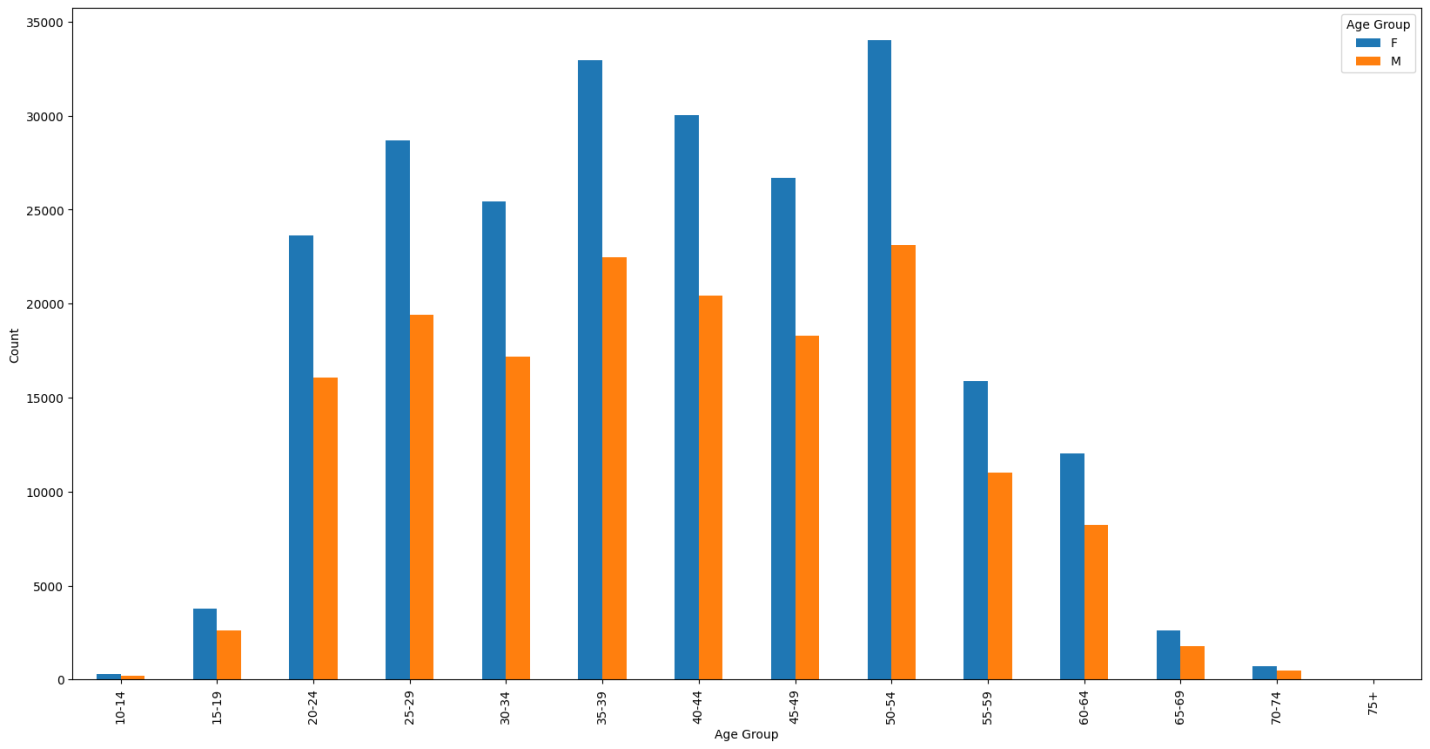


**3.DRUG NAME:**

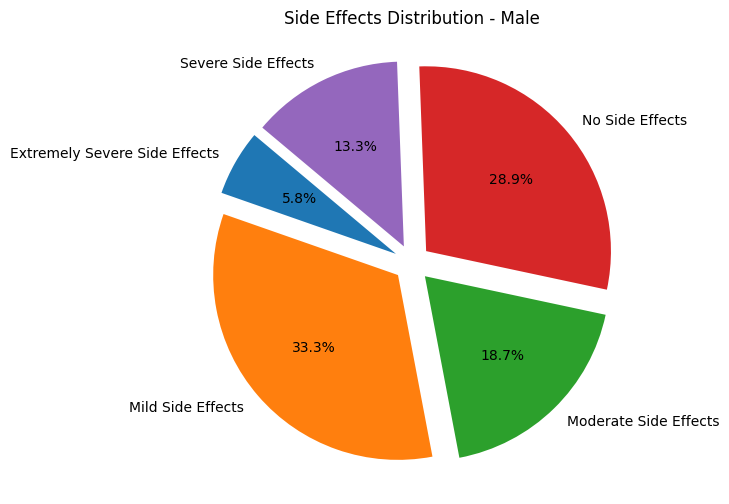


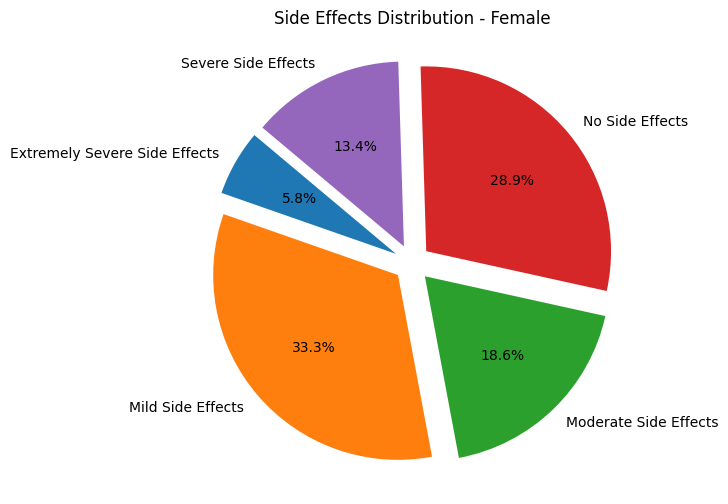
* **BIVARIATE PLOTS:-**

**1. Age Vs Gender:**

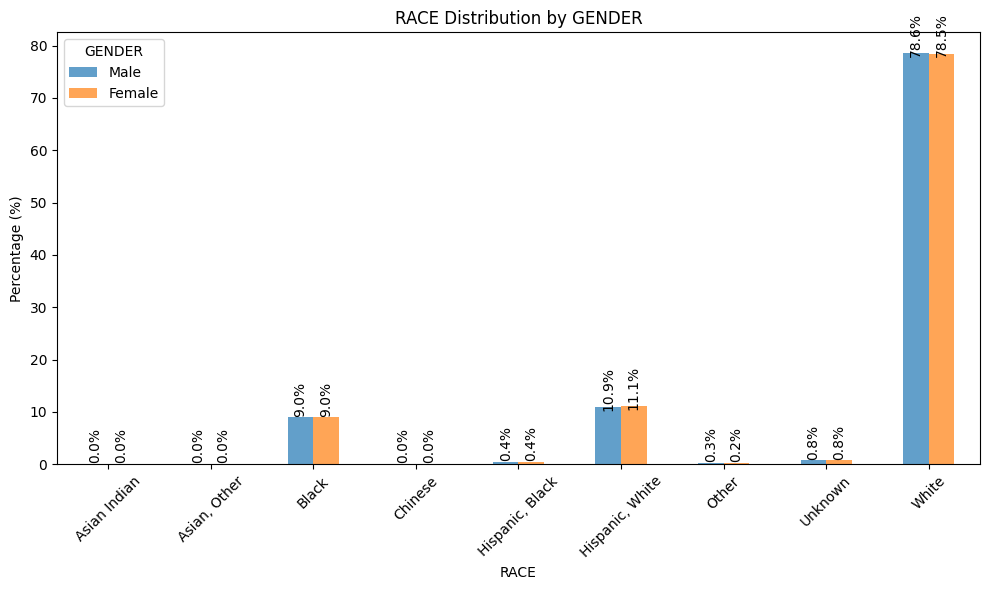


**2. Gender Vs Side Effects:**

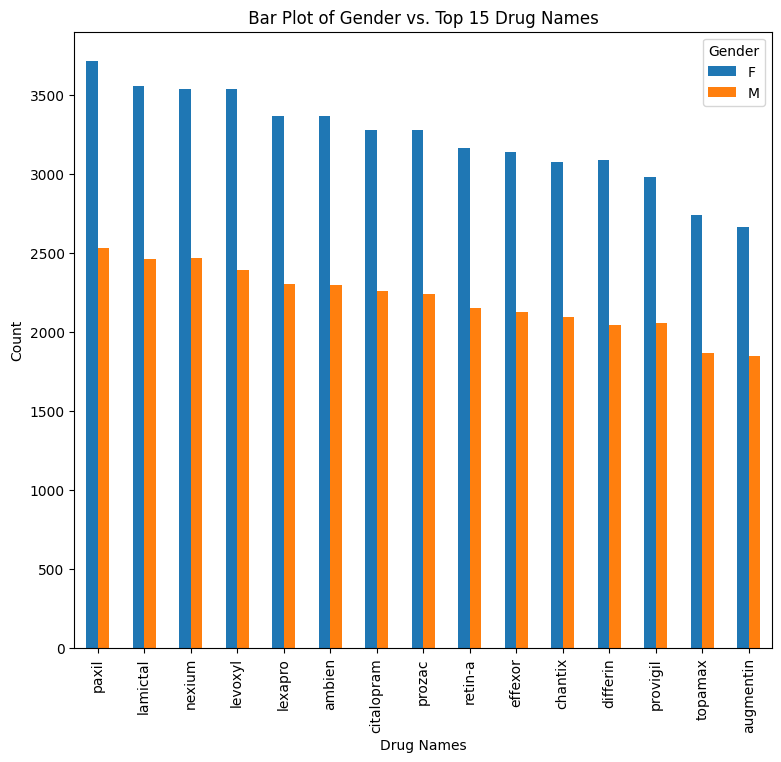




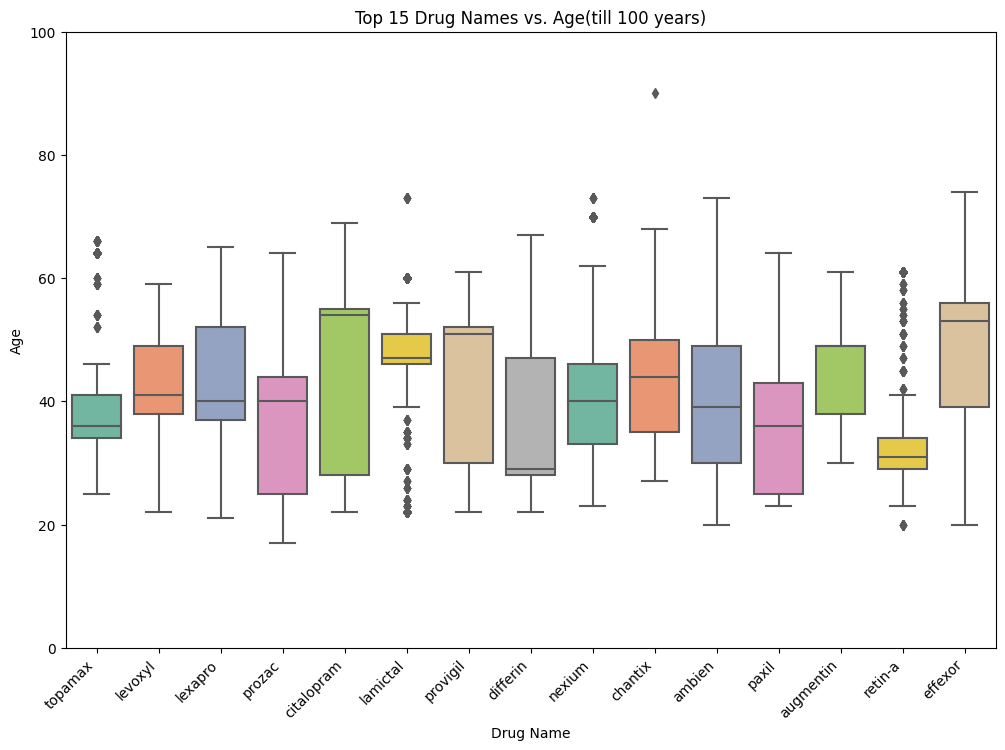
**3. Gender Vs Race:**



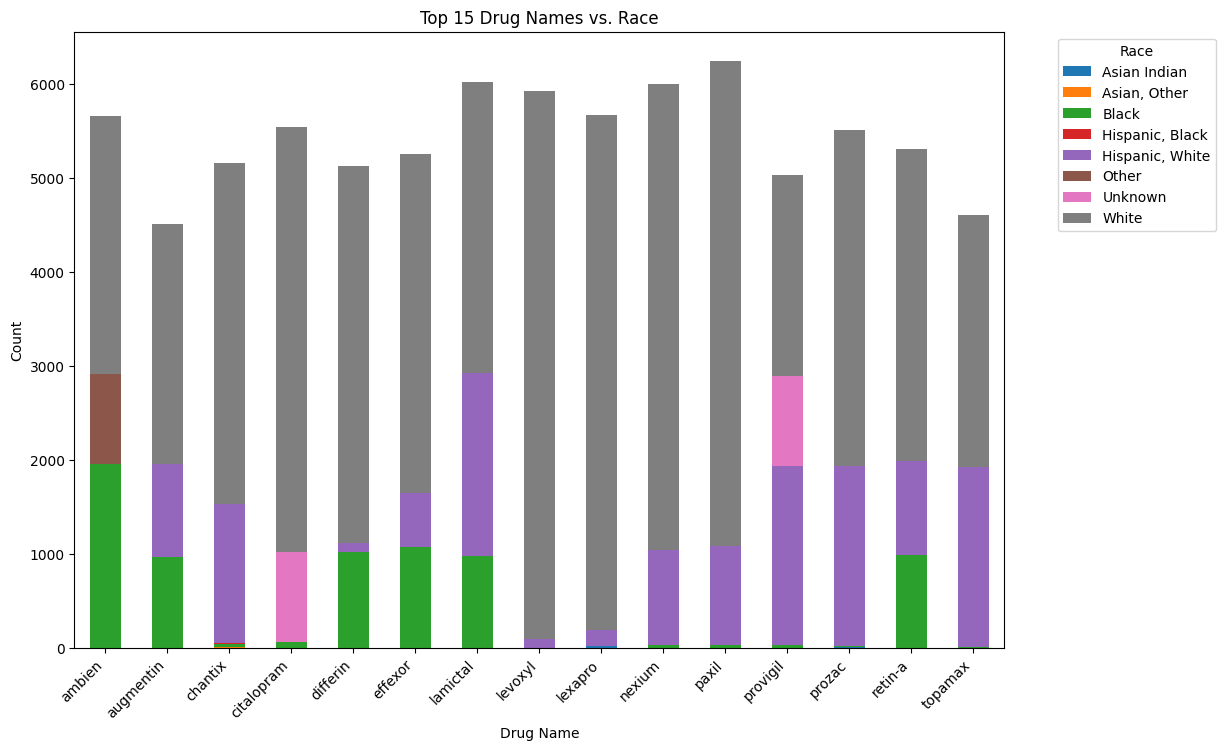
1. **Gender Vs Drug Name:**

****

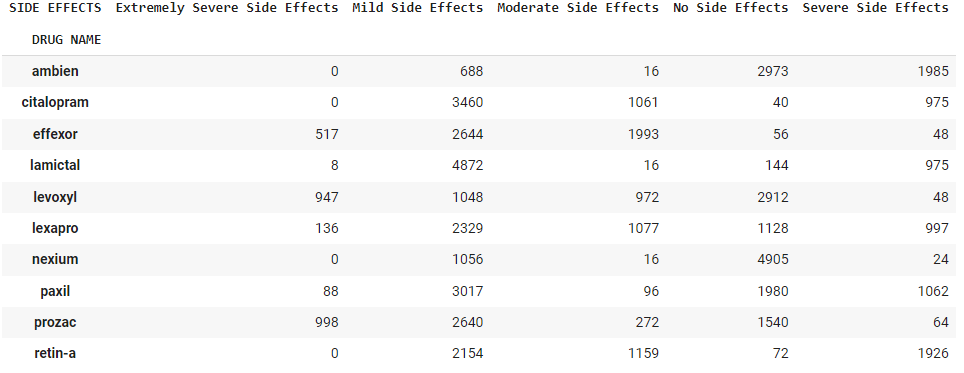
1. **Drug name Vs Age:**

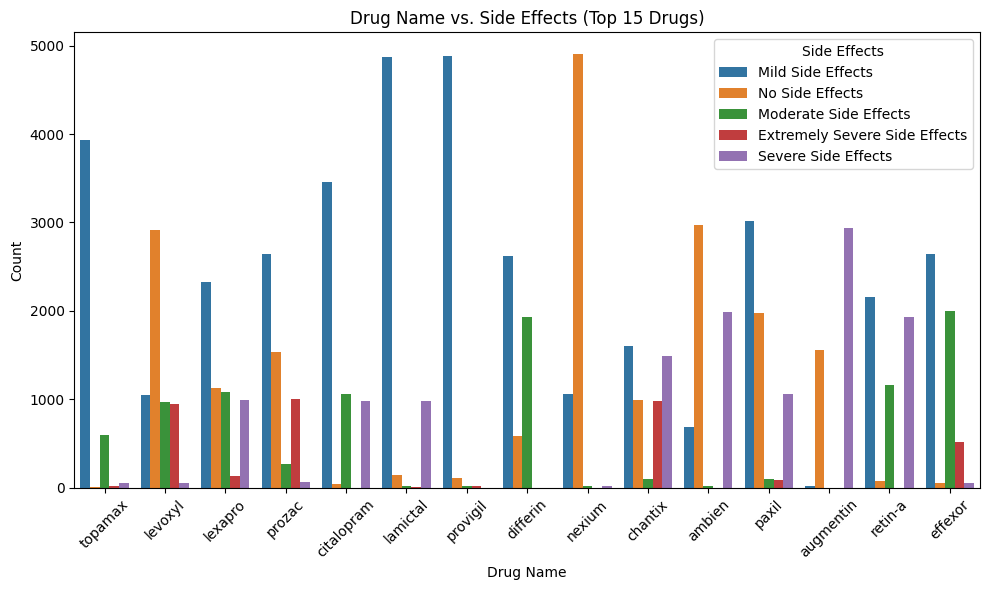
****

1. **Drug Name Vs Race:**

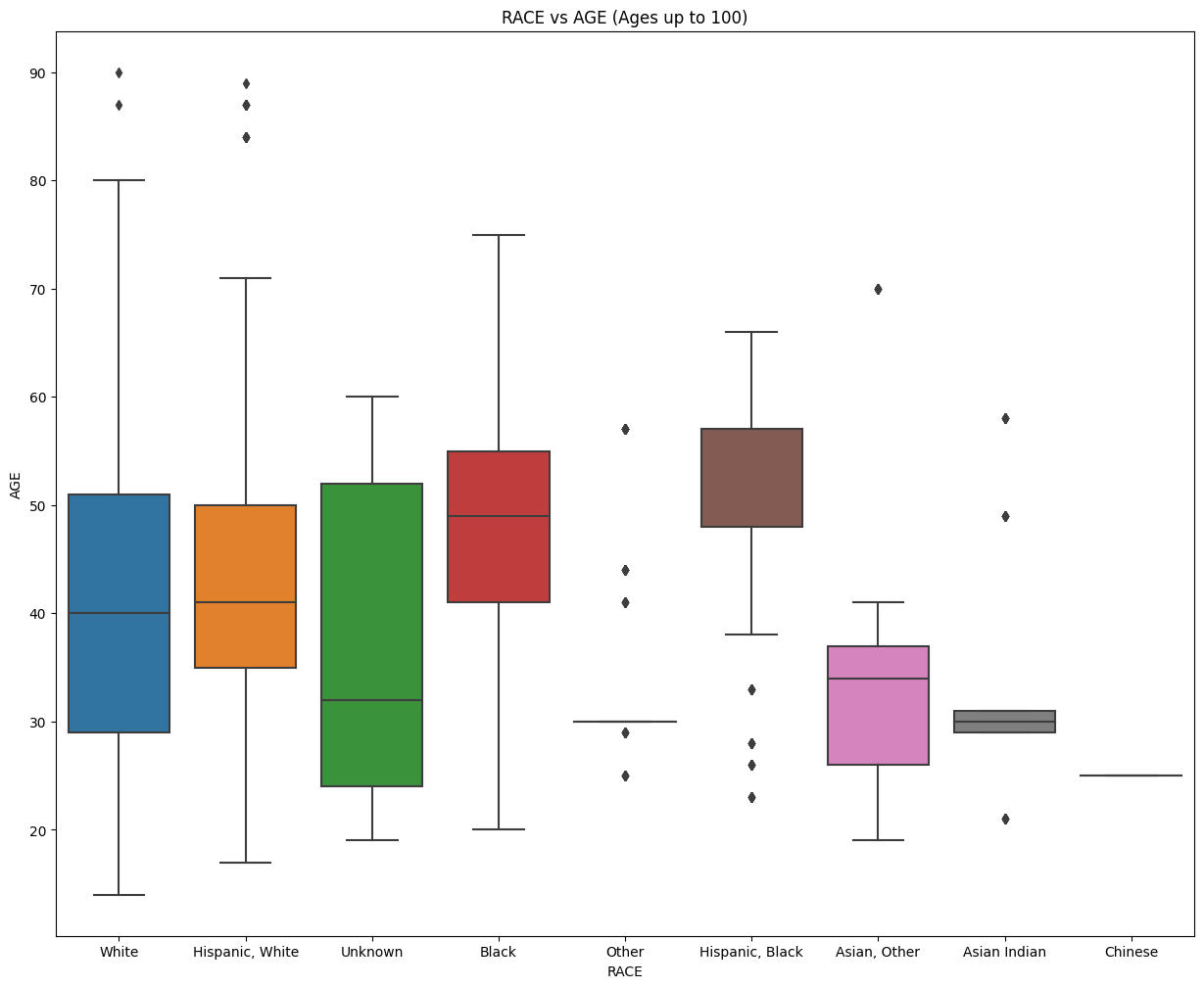
****

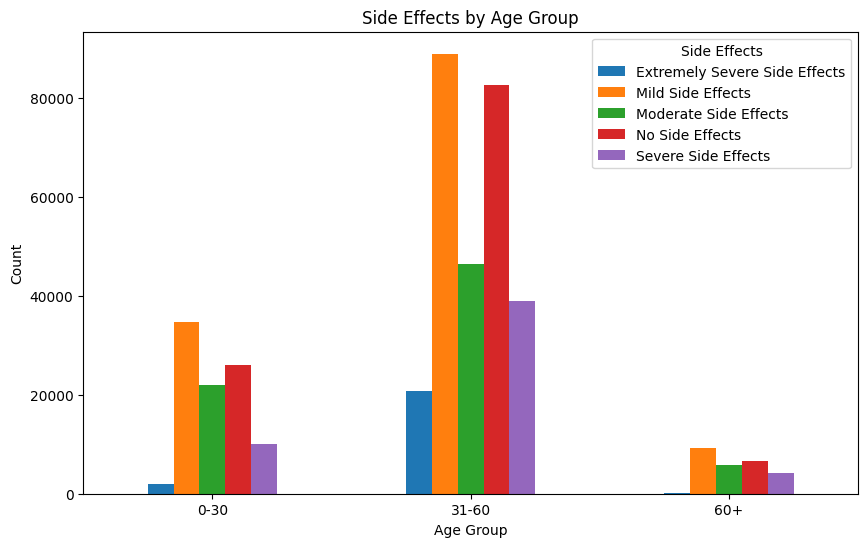
1. **Drug Name Vs Side effects:**



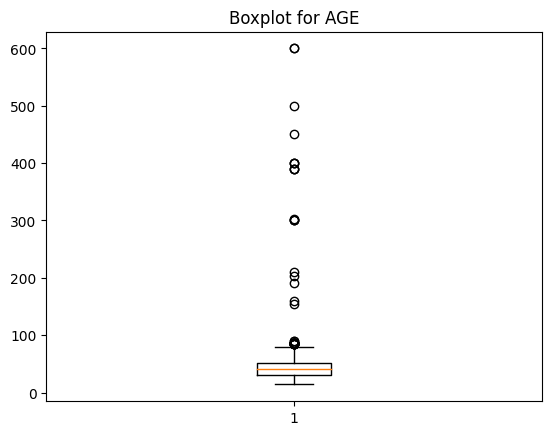


**7. Race Vs Age:**





**OUTLIER DETECTION**



**ALGORITHMS**

When selecting the most suitable algorithm for your dataset it is crucial to take two key factors into account: the scale of features and the presence of outliers. Certain algorithms exhibit robustness when dealing with datasets characterized by varying feature scales. Moreover, the existence of outliers, which are data points significantly different from the majority, can introduce challenges to statistical analyses and model predictions.

Several algorithms are known for their ability to perform well under these conditions, as they are generally less affected by feature magnitudes and the presence of outliers. Nonetheless, it is considered best practice to address outliers during data preprocessing to enhance model performance and generalization. Strategies for handling outliers may involve their removal, data transformation, or the utilization of robust algorithms designed to handle outliers effectively.

Before deciding on the most appropriate algorithm for your dataset, it is essential to assess the nature of your data thoroughly. This assessment should encompass an evaluation of feature ranges, scales, and the presence of outliers. By considering these factors, you can make an informed choice, ensuring that your chosen algorithm aligns with your data’s characteristics and thereby enhancing the accuracy and reliability of your analysis or modeling endeavors. I am considering the following algorithms after

Considering all the above explanation:

1. Decision trees: Decision trees split the data based on feature thresholds, and the order or magnitude of features does not affect the outcome. Only the relative ordering of feature values matters.

2. Random forests: Random forests are an ensemble of decision trees. Since decision trees in a random forest work independently, the magnitude of features does not significantly impact the overall model.

3. Gradient Boosting Machines (GBM): GBM builds an ensemble of weak learners (typically Decision trees) in a sequential manner. Similar to random forests, the magnitude of features does not have a significant effect on the final model.

4. Naive Bayes: Naive Bayes is a probabilistic algorithm based on Bayes theorem. It assumes that features are conditionally independent given the class label, making it less sensitive to feature scaling.

5. Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel: SVM with an RBF kernel only considers the pairwise distance between samples in a high-dimensional space. Thus, the scale of individual features does not matter. (DROPPED)

## PREDICTIVE MODELLING AND FINE TUNING

## 1. Train Test Split

In the process of preparing a machine learning model, one crucial step is the "train-test split." This step involves dividing the available dataset into two distinct subsets: the training set and the test set. In this particular scenario, the data has been split using a test size of 0.25, meaning that approximately 25% of the data has been set aside for testing the model's performance. To ensure reproducibility, a random state of 42 has been employed. This separation into training and test sets is fundamental for evaluating the model's generalization performance. By training the model on the larger training set and then assessing its predictive capabilities on the separate test set, we can gain insights into how well the model is likely to perform on new, unseen data.

Number transactions X\_train dataset:  (298641, 4)

Number transactions y\_train dataset:  (298641, 1)

Number transactions X\_test dataset:  (99548, 4)

Number transactions y\_test dataset:  (99548, 1)

## 2 Algorithms Decided

Algorithms that are generally less affected by the magnitude of features was considered to be tried out:

Decision Tree: A Decision Tree is a machine learning algorithm that makes decisions by recursively partitioning the data into subsets based on feature values. It selects the best features and thresholds to create splits that minimize impurity or maximize information gain. Decision Trees are interpretable and can capture complex relationships, but they can also over fit if not controlled properly.

Random Forest: Random Forest is an ensemble learning method that combines multiple Decision Trees to improve predictive accuracy and control over fitting. Each tree is trained on a subset of the data with replacement (bootstrap sampling) and a subset of features. The final prediction is made by aggregating the predictions of individual trees. Random Forests are robust and effective for a wide range of tasks.

Gradient Boosting Machines (GBM): GBM is another ensemble technique that builds an ensemble of weak learners (typically Decision Trees) sequentially. It starts with an initial model and then fits subsequent models to the errors of the previous ones. GBM focuses on improving the model's performance with each iteration. It's a powerful algorithm but can be more sensitive to hyperparameters and may require careful tuning.

Naive Bayes: Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that features are conditionally independent given the class label. Despite its "naive" assumption, Naive Bayes often performs well in text classification and other high-dimensional problems. It's computationally efficient and requires relatively small amounts of training data.

## 3. Comparison of Metrics

While accuracy is a straightforward measure of a classifier's performance, it might not provide a comprehensive evaluation, particularly in cases of imbalanced class distribution. In scenarios where classes are unevenly represented, accuracy can be misleading, as it doesn't account for the specific challenges posed by imbalanced data. Therefore, alternative metrics such as precision, recall, and F1 score are essential for a more nuanced assessment. These metrics focus on aspects like true positive rate (recall) and the balance between precision and recall (F1 score), which are particularly relevant when dealing with imbalanced classes. The Decision Tree algorithm performs better compared to other algorithms based on the metrics.

Metrics of Decision Tree is as follows:

.Accuracy = 0.9863985213163499, Precision = 0.9864405884238486

Recall = 0.9863985213163499, F1 Score = 0.9863622857744598

However, recognizing the importance of addressing data imbalances, a strategic decision was made, that is to consider other metrics instead of accuracy.

**4. Best Model and Cross Validation**

Decision Tree performed well on the basis of metrics, hence it is considered as the best model.

Cross-validation is a crucial technique used to evaluate the performance of machine learning models and mitigate the risk of over fitting. It involves dividing the dataset into multiple subsets (folds), training the model on a subset, and evaluating its performance on the remaining data. This process is repeated several times, rotating the subsets used for training and testing. Stratified cross-validation, a variant of cross-validation, ensures that each fold maintains the same class distribution as the original dataset, which is particularly useful for imbalanced classes.

Cross-Validation Precision: **0.9863593616546146**

Cross-Validation Recall: **: 0.986323077582729**

Cross-Validation F1 Score: **0.9863195530972095**

## 5. Hyperparameter Tuning

Randomized Search CV is a technique used for hyperparameter tuning in machine learning. It efficiently explores a predefined hyperparameter space by randomly selecting combinations of hyperparameters to evaluate the model's performance.

In this specific case, Randomized Search CV was preferred over Grid Search CV due to the size of the dataset. Grid Search CV explores all possible hyperparameter combinations, which can be impractical for large datasets as it may lead to excessive computational time. After applying Randomized Search CV and identifying the best set of hyperparameters, the model's performance further improved.

Consequently, the model was fine-tuned using the best hyperparameters obtained from Randomized Search CV, resulting in even better performance. This improved performance, combined with the efficiency of Randomized Search CV on a large dataset, led to the decision to finalize this model as the best option.

Metrics of the model after Hyperparameter Tuning:

Mean Precision (Decision Tree): **0.9852784128948934**

Mean Recall Score (Decision Tree): **0.985245551690797**

Mean F1 Score (Decision Tree): **0.9852484451807402**

**CONCLUSION**

In summary, my project has effectively tackled the essential task of developing a model for classifying drug side effects, a matter of paramount significance within the healthcare and pharmacological sectors. Leveraging comprehensive data analysis, meticulous feature engineering, and the application of cutting-edge machine learning techniques, I have attained notable milestones in this pursuit.

My model showcases its potential in elevating drug safety and patient care by providing precise predictions and categorizations of side effects. This not only equips healthcare professionals with invaluable insights but also empowers patients to make informed decisions regarding their treatment choices.

Throughout the course of this project, I encountered several challenges, encompassing issues related to data quality, the intricacies of model interpretability, and ethical considerations concerning sensitive medical information. Nonetheless, my unwavering commitment to overcoming these challenges has yielded a robust and ethically sound solution.

Looking ahead, my work unlocks promising avenues for personalized medicine, cost-effective healthcare, and streamlined drug development processes. Collaborative ventures with healthcare providers and pharmaceutical companies hold the potential for the practical implementation of my model in clinical settings, ultimately enhancing patient outcomes.

In a broader context, this project contributes to the ongoing endeavors to enhance drug safety and healthcare delivery. It underscores the transformative potential of data-driven approaches in addressing complex medical issues. As I bring this phase of my project to a close, I remain enthusiastic about exploring further research opportunities and leaving a lasting imprint on the field of healthcare.

I extend my heartfelt gratitude to all those who provided support and contributed to this project, and I eagerly welcome any inquiries or collaborative ventures aimed at advancing this pivotal area of study.

Github Repository Link: <https://github.com/Shreejith25/ML_Projects/blob/c85d547ad67cc122a19ea41f568e42b41c794272/Drug_Classinfication_TCS_ION%20InternShip/Drugbatch_Side_effect.ipynb>