Dear Intern

Project report is an inherent component of your internship. We are enclosing a reference table of content for the project report. Depending on the internship project (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections from the table of content mentioned below. You can also add additional sections. The key objective of this report is for you to systemically document the project work done.

|  |  |
| --- | --- |
| Internship Project Title | Classification Model - Build a Model that Classifies the Side Effects of a Drug |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | ICT Academy of Kerala |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
|  |  |  | VS code, Google colab | Python |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses

**Acknowledgements**

I would like to express my sincere gratitude for their invaluable support, guidance, and contributions throughout my TCS iON internship and research project on data science. I appreciate TCS iON granting me an opportunity to pursue this internship and learn much concerning the interesting area of data science. My research objectives could not have been accomplished without the assistance of TCS iON's resources, amenities, and friendly working atmosphere.

Finally, I want to extend my heartfelt thanks to my close companions and loved ones for their perpetual encouragement, inspiration, and compassion throughout this project. They continually inspired me and believed in my potential, which helped me stay inspired and laser-focused.

**Objective**

* Develop a classification model that classifies the side effects if a drug by age, gender, and race.

**Introduction / Description of Internship**

The discovery and evolution of new pharmaceuticals are integral to enriching people's quality of life on a global scale. Drugs might have healing advantages, but they can also have unforeseen consequences such as side effects. For patient safety and maximizing therapeutic results, it is crucial to recognize and classify these side effects, which can range from minor distress to catastrophic effects.

For myriad stakeholders in the healthcare ecosystem, the classification of side effects of drugs has substantial implications. A methodical strategy for recognizing and comprehending pharmaceutical side effects can aid healthcare professionals in dealing with patients and designing treatments by empowering them to make well-informed decisions. Regulatory bodies can use this research to better drug safety assessments and make sure that patients and healthcare providers are properly informed about the hazards associated with medications. Additionally, pharmaceutical firms might make use of the knowledge gleaned from this study to enhance medication development procedures and optimize drug safety profiles.

The principal objective of this study is to employ data science methods to construct a classification model for pharmacological side effects. Founded on data from clinical trials, the program will try to categorize pharmacological side effects. By fulfilling this goal, I hope to enhance knowledge about drug side effects and give regulatory organizations and medical professionals a practical tool for weighing the pros and cons of various prescription drugs.

Machine learning and data science approaches have recently revolutionized many industries, including healthcare. Classification models have been established as beneficial tools for studying and classifying complex datasets, such as drug side effects. The classification models often used in Python for researching pharmacological side effects are logistic regression, decision trees, random forest, support vector machine, neural networks, Naive-Bayes, etc.

**Internship Activities**

The internship activities are based on an organized and object-oriented day-wise plan provided. It started with the pre-assessment test. Also, following the day-wise plan to engage in various learning techniques and project activities.

**Approach / Methodology**

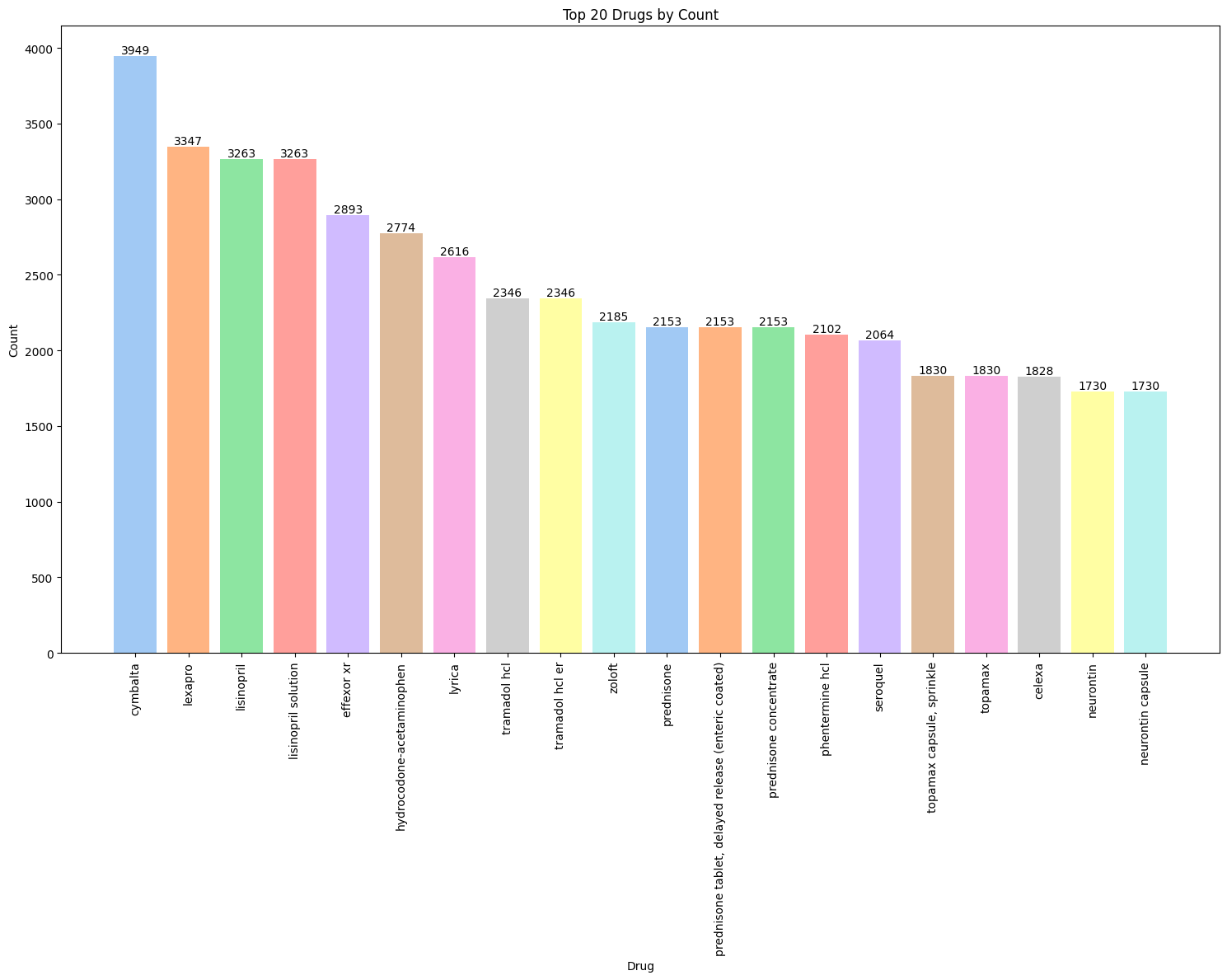
* Data Collection
* Preparation of the dataset (WebMD dataset on drugs review is the base here).
* Datasets usually contain raw and disoriented data. For the better performance of the model, data should be cleaned.
* Pre-process the data.
* Carry out exploratory data analysis.
* Feature selection.
* Split the data for training and testing.
* Feature Engineering of the data.
* Determine an appropriate classification model (e.g., Logistic Regression, Decision Trees, Random Forests, SVM, Neural Networks) based on the research objectives and the characteristics of the dataset.
* Evaluate the trained model's performance on the test set to assess its accuracy and generalization ability. Use appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
* If the model's performance is not satisfactory, consider fine-tuning the model's hyperparameters to improve its performance.
* Documentation and Reporting.

**Assumptions**

* The accessibility of correct and trustworthy data
* Independence of observations
* Relevance and comprehensiveness of features
* The linearity assumption in logistic regression
* Handling of class imbalance and the generalizability of constructed models.

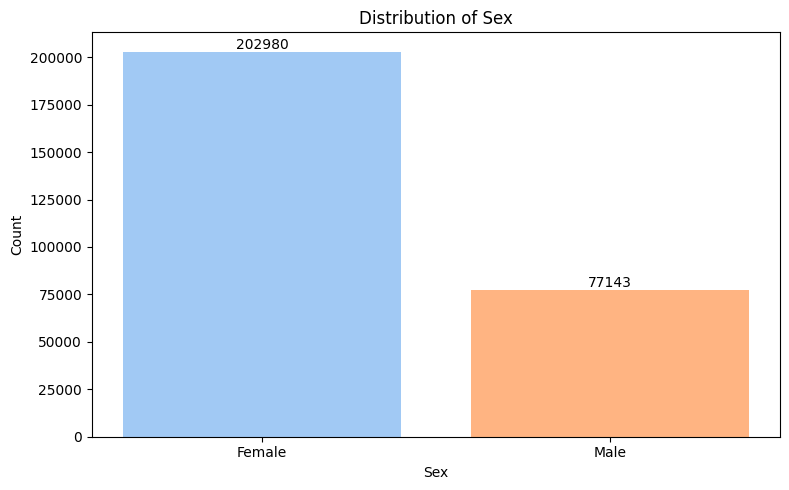
**Charts, Table, Diagrams**

1. Top 20 Drugs by count

****

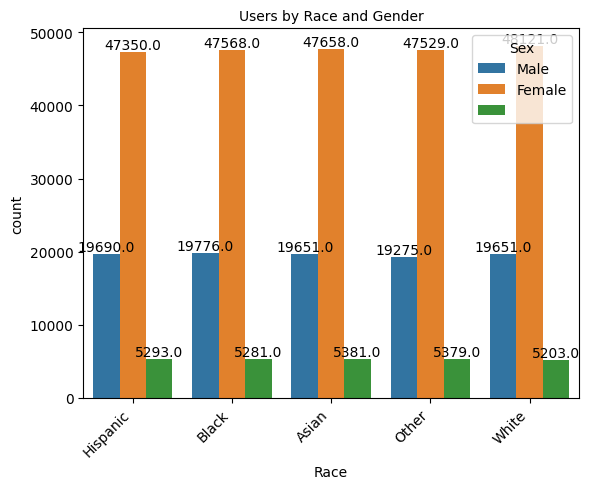
* Unique values : 5014
* Number of unique identities with count 1: 1046,
* Top counts
* cymbalta 3949
* lexapro 3347
* lisinopril 3263
* lisinopril solution 3263
* effexor xr 2893

1. Distribution of sex



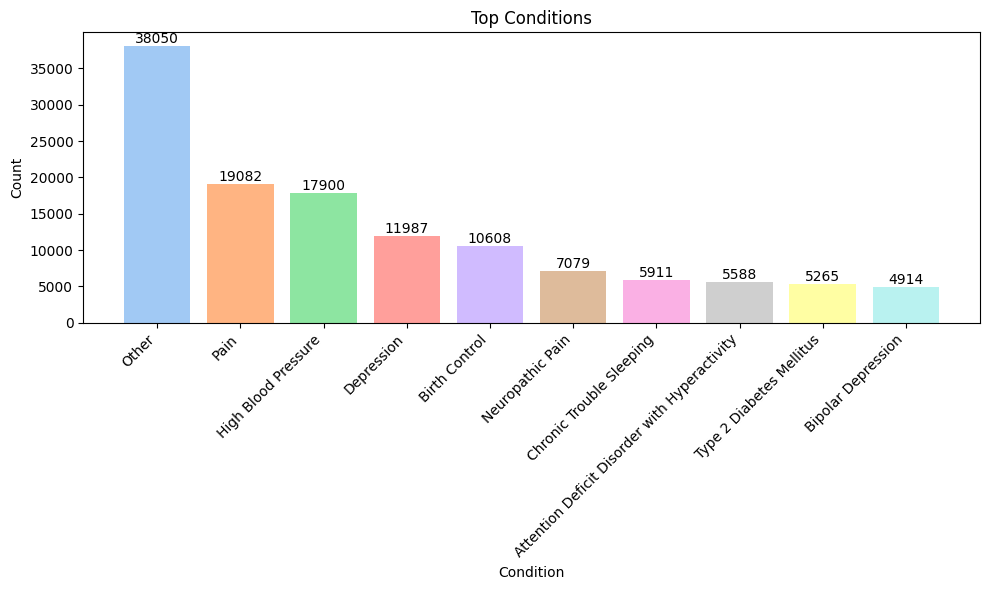
* Female 202980
* Male 77143
* 72% of the population is Female

1. Distribution of Race and Gender



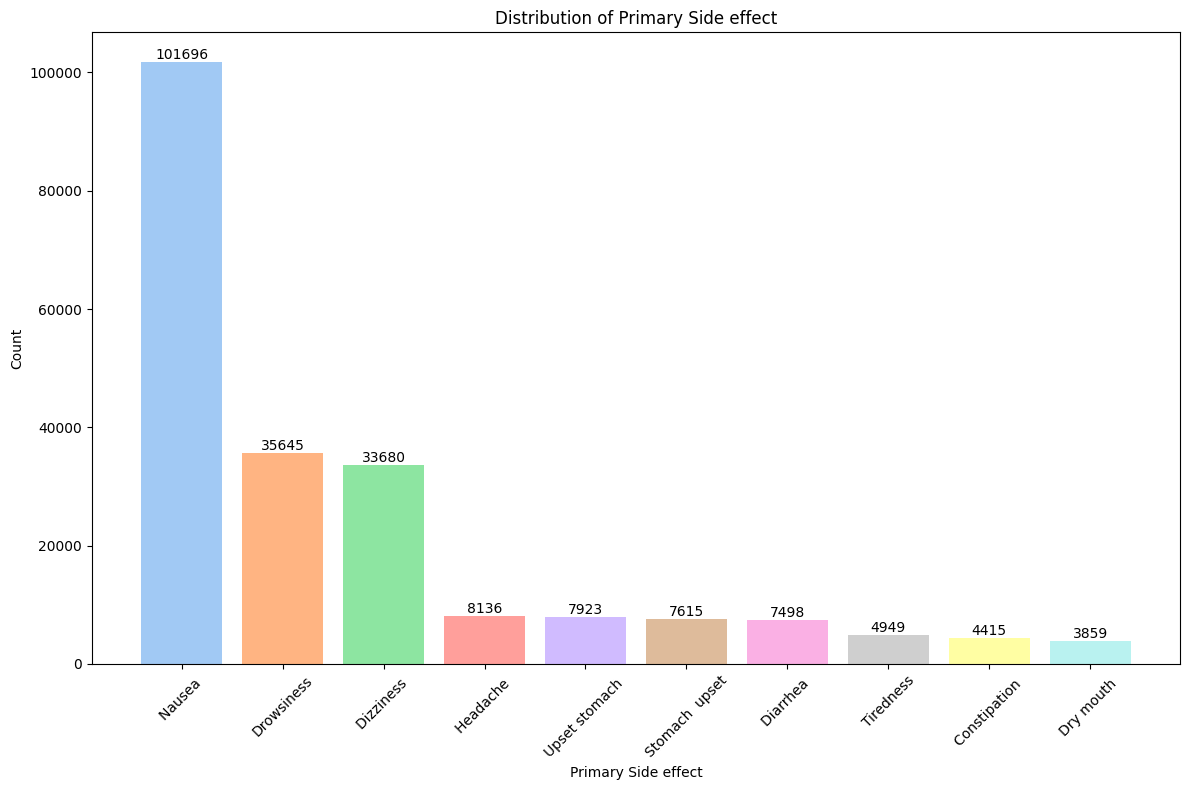
* No matter the race, Female population is the highest drug user.

1. Top conditions drugs were used for



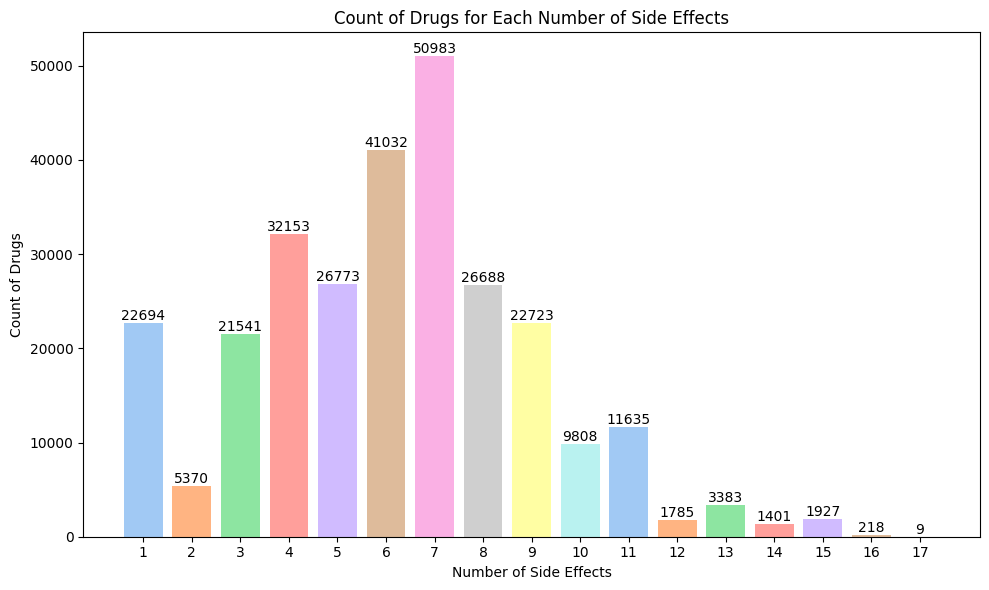
Pain was the condition that needed higher drug dependency other than for un named conditions, followed by high blood pressure, and depression.

1. Top Primary side effects for drug usage



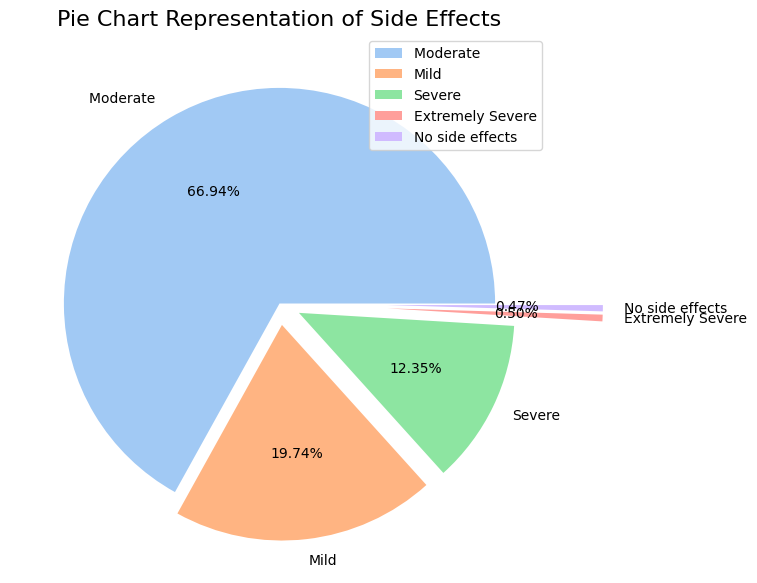
From the Side effects separated by comma, each were categorized into different columns. Among them, in the primary side effect, nausea topped the list followed by drowsiness and dizziness.

1. Drug Vs No.of Side effects



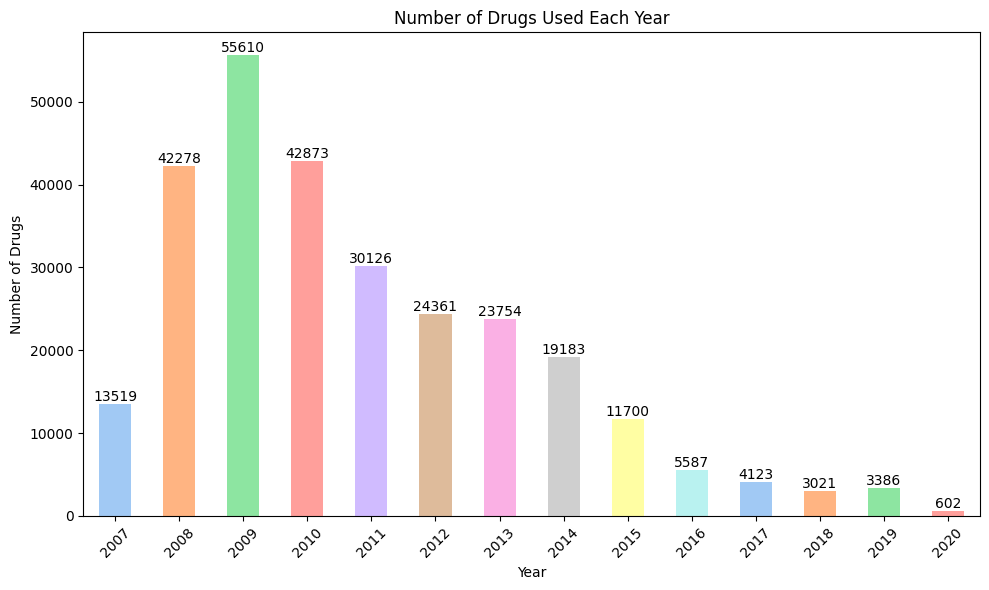
50983 drugs recorded 7 side effects with which it topped the list among total 17 side effects. And 9 drugs recorded 17 side effects which was the highest.

1. Pie chart Representation of side effects



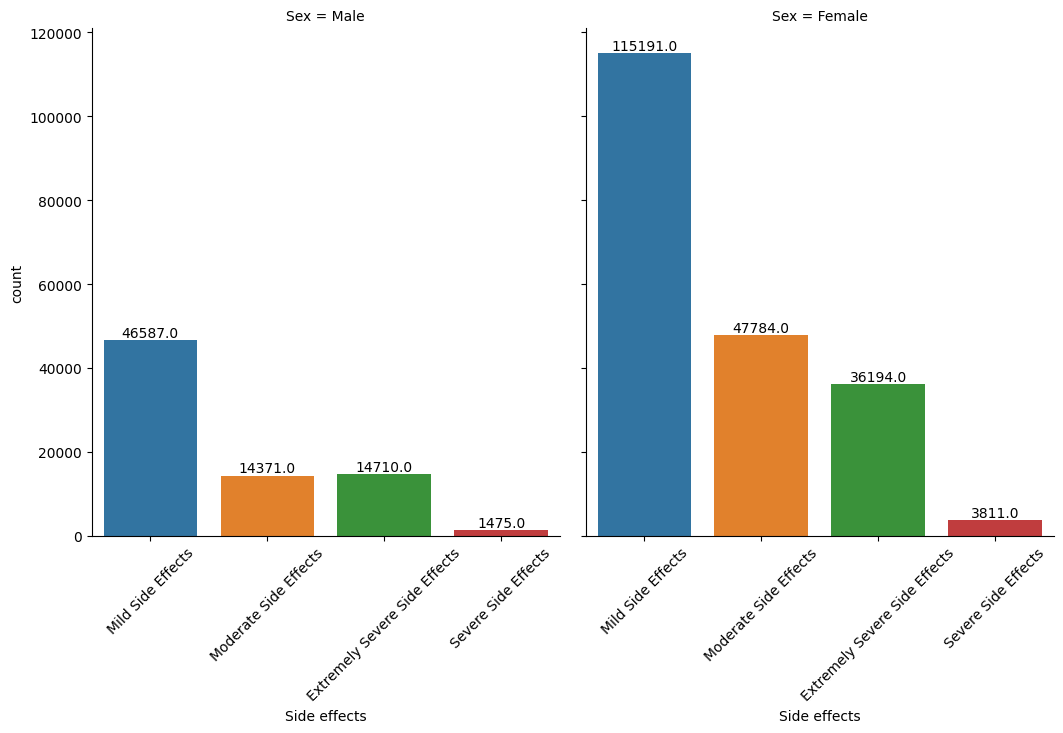
99 percent of the drugs showed side effects with varying severity. Moderate side effects were the highest with 66.94% and Extremely severe the lowest with 0.50 %.

1. Number of drugs used each year



Therapeutic drug usage over the years has treasonously reduced. Highest drug usage was seen in the year 2009 and lowest in the recent years.

1. Severity of Side effects Vs Sex



Female gender had more side effects to the drugs than males.

**Algorithms**

1. **Logistic Regression**

The ability to precisely classify the presence or absence of particular side effects is crucial in the study of drug side effects. An effective tool for this is logistic regression, a widely accepted classification technique. In binary classification problems, the dependent variable might have one of two possible outcomes. Logistic regression is a statistical model employed for these types of conditions. In contrast to linear regression, which makes predictions about continuous values, logistic regression models the likelihood that the target variable will fall into one of several classes. For this, a logistic curve is fitted to the data, enabling probabilistic predictions and categorization.

1. **Decision Tree Classifier**

A well-liked machine learning approach for classification and regression is the decision tree classifier. Each internal node reflects a judgment based on a feature, and each leaf node corresponds to a class label. It divides the dataset into subsets depending on the values of independent variables. Because they can manage both category and numerical variables, decision trees are particularly helpful for handling the many types of data used in drug side effects research.

1. **Support Vector Machines**

Support Vector Machines are models for supervised learning that examine the data and create a boundary to divide it into various classes. SVM seeks to identify an ideal hyperplane that maximizes the separation between the data points of various classes while preserving a margin of separation. Through the use of several kernel functions, including linear, polynomial, and radial basis function (RBF) kernels, SVM is capable of handling both linear and non-linear classification problems.

1. **Random Forest Classifier**

A decision tree ensemble learning technique called Random Forest Classifier uses several different decision trees to aggregate predictions. It works by building a collection of decision trees, where each tree is trained on a random part of the dataset and uses a random subset of characteristics for splitting. Through voting or average, the predictions of different trees are combined to get the final prediction. Both classification and regression tasks can be handled by the Random Forest Classifier, which excels at handling high-dimensional and noisy data.

1. **KNN Classifier**

The KNN classifier is a non-parametric algorithm that assigns labels to data points based on the majority vote of their nearest neighbors in the feature space. The algorithm measures the distance between a test data point and the training data points to determine its K nearest neighbors. The class label of the test data point is then determined by the majority class among its neighbors. KNN is a simple yet effective algorithm that can handle both binary and multi-class classification problems. The KNN classifier offers several advantages in the context of drug side effects classification. Firstly, it is simple to understand and implement, making it accessible even to those without extensive machine learning knowledge. Secondly, KNN is a non-parametric algorithm, meaning it does not make assumptions about the underlying data distribution. This flexibility allows it to handle complex relationships and adapt well to different types of data. Thirdly, KNN can handle multi-class classification without requiring additional modifications, making it suitable for datasets with multiple side effects.

**6. Naive Bayes classifier**

The Naive Bayes classifier is based on Bayes' theorem and assumes that features are conditionally independent given the class label. This assumption allows for efficient and fast training and classification, making Naive Bayes a popular choice for text categorization, spam filtering, and other classification problems. Naive Bayes classifiers use probability distributions to estimate the likelihood of a given class label based on the observed features. The Naive Bayes classifier offers several advantages in the context of drug side effects classification. Firstly, it is computationally efficient and requires a relatively small amount of training data. This makes it suitable for large datasets and real-time applications. Secondly, Naive Bayes can handle high-dimensional feature spaces and effectively deal with irrelevant features by assuming independence among them. Thirdly, the algorithm is robust to missing data and can handle categorical and numerical features without requiring extensive preprocessing. Additionally, Naive Bayes classifiers can provide interpretable results by estimating class probabilities based on observed feature values.

**Challenges & Opportunities**

Challenges encompassed data quality, class imbalance, feature selection, and interpreting complex models. However, the opportunity for improved safety assessment, personalized medicine, pharmacovigilance, and data-driven insights arises from tackling these issues. Researchers can use classification models to meaningfully contribute to the classification of pharmacological side effects and ultimately enhance patient care and drug development processes by recognizing and overcoming the hurdles.

**Risk Vs Reward**

Risks include issues with interpretability, model performance, ethical and legal issues, and model generalization. However, the benefits of a well-run study hold better medication safety evaluation, righter patient care, and personalized treatment, advances in pharmacological understanding, and contributions to public health. Researchers can maximize the potential advantages and significantly advance the science of classifying drug side effects by properly regulating and limiting the hazards.

**Reflections on the Internship**

The opportunity for professional and personal growth given by the internship experience enabled me to implement my acquired knowledge and abilities to use in a real-world context. Throughout the internship, many important lessons came to light that have advanced knowledge of the classification of drug side effects and the larger field of data science. Applying several classification methods, such as logistic regression, decision trees, random forests, support vector machines (SVM), K-nearest neighbours (KNN), and Naive Bayes, in real-world settings was made possible by the internship. This experience helped me clarify the advantages, constraints, and suitable applications of each approach. Proper evaluation of classification models using appropriate metrics such as accuracy, precision, recall, F1-score, and AUC-ROC was emphasized during the internship. The importance of model interpretation, including feature importance analysis and visualizations, was also recognized to gain insights into the relationships between features and drug side effects.

Exploring more advanced techniques, such as deep learning algorithms or ensemble methods, to further improve the performance and interpretability of the classification models. This would involve delving into complex neural network architectures, transfer learning, and model ensembles to leverage the full potential of data science in drug side effects classification. Future improvements can involve incorporating advanced techniques, incorporating domain expertise, exploring longitudinal data analysis, and ensuring ethical considerations are at the forefront. Overall, the internship served as a valuable learning experience and laid the foundation for further exploration and growth in the field of drug side effects classification.

**Recommendations**

During the internship, transparent communication between the mentor and intern is crucial for a successful and fulfilling experience. Transparent communication fosters a better understanding of expectations, goals, and feedback, leading to improved learning outcomes and a stronger mentor-intern relationship.

**Outcome / Conclusion**

|  |  |
| --- | --- |
| Model | Accuracy (%) |
| Logistic Regression | 57.57 |
| KNN Classifier | 76.55 |
| Naïve Bayes Classifier | 57.36 |
| Random Forest Classifier | 94.92 |
| Decision Tree classifier | 60.74 |

**Enhancement Scope**

During the internship, there were several opportunities to enhance the overall experience and maximize the learning and growth potential.

**Link to code and executable file**

<https://github.com/Ardraky/Internship/blob/main/Project/Ardra_K_Y_CSED.ipynb>