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 | 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 |
| 7-9-2023 | 27-10-2023 | 125 | Remote Internship | Jupyter Notebook |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Outcome / Conclusion
* Link to code and executable file
* Research questions and responses

**Acknowledgements**

I would like to express my heartfelt gratitude to all those who made this internship experience a meaningful and insightful journey. The successful completion of the project, "- Build a Model that Classifies the Side Effects of a Drug." was made possible by the support, guidance, and collaboration of TCSiON organization

I am immensely thankful to my Industry mentor, Debashis Roy , whose expertise and guidance played a pivotal role in shaping this project. Their patience, wisdom, and willingness to share knowledge have been instrumental in my learning and development throughout this internship

This project would not have been possible without the unwavering support of my family and friends. Your encouragement and belief in my abilities were a constant source of motivation.

Lastly, I acknowledge the ICT Academy of Kerala, and online resources that provided the foundation of knowledge necessary to undertake this project.

Thank you all for your contributions and support in making this internship project a reality.

Sincerely,

Blessy John

.

**Objective**

The objective of the project is to build a model that classifies the side effects of a drug.

**Introduction / Description of Internship**

TCS iON Remote Internships are the job starter you can depend on. With relevant industry experience – graduates, post-graduates, students across various domains, and even freshers can add value to their resumes from our various internship programs. Our end-to-end upskilling solutions help you prepare, test and analyze your knowledge and kickstart a career from the comfort of your home.

TCS iON is on of the largest online pltform 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.

**Internship Activities**

TCS ION Remote Internship has two major learning components spread across 125 hours of engagement for successfully completion of program.

* TCS iON RIO-125: Classification Model - Build a Model that Classifies the Side Effects of a Drug
* Digital Discussion Room RIO-125: Classification Model - Build a Model that Classifies the Side Effects of a Drug.

**Classification Model - Build a Model that Classifies the Side Effects of a Drug**

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

**DAY WISE PLAN**

Day wise plan or learning schedule gives an idea on how to spend day and how much time to devote to each module for 30 days.

**PRE TEST**

A simple test for a quick knowledge check.1 hour duration marked for 40 marks and 20 as passing marks.Pre test is a general aptitude test needed to be taken before commencing work on the project.

The duration of the test is one hour and it comprises of questions to test analytical abilities,reasoning abilities and language skills.Upon successful attempt ,it unlocks the details of the industry project and other nodes in the construct.

**SELF LEARNING**

References to different learning resources to enable students to successfully do project,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.As soon as students have seen the industry project and have understood the objective,guidelines and project outcome,students may start going through the self learning content.

**DIGITAL DISCUSSION ROOM**

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

**ACTIVITY REPORT**

Daily activity report for documenting progress.

* Day 1- 30 Activity Report
* Industry project
* 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.The test will have 10 objective type questions.

* Viva
* Viva Report

**Approach / Methodology**

The main approches used in this study are researche on internet and reading and refering articles based on the project.And also refer the materials given by the company and understand the various algorithms used and how to interpret it.

**Assumptions**

* There is sufficient and accessible data available for training and evaluating the

Classification model. This includes data on drugs, their usage, and associated side

effects.

* The data we have is of reasonable quality, with few mistakes or inconsistent data points.
* The dataset's attributes are pertinent for classifying drug side effects. Information on the medications, patient characteristics, and descriptions of side effects are all included.
* Assume that it is possible and can produce useful results for the problem of classifying drug side effects using machine learning techniques.

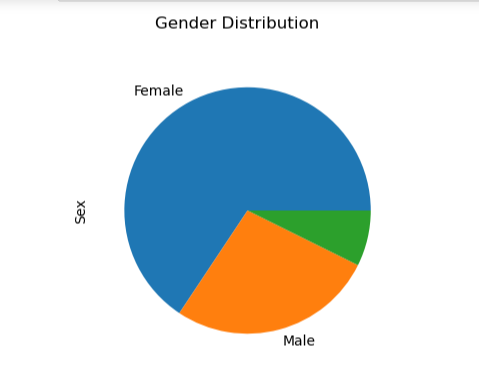
**Exceptions / Exclusions**

Clearly specifying exceptions and exclusions helps manage expectations and avoid misunderstandings.

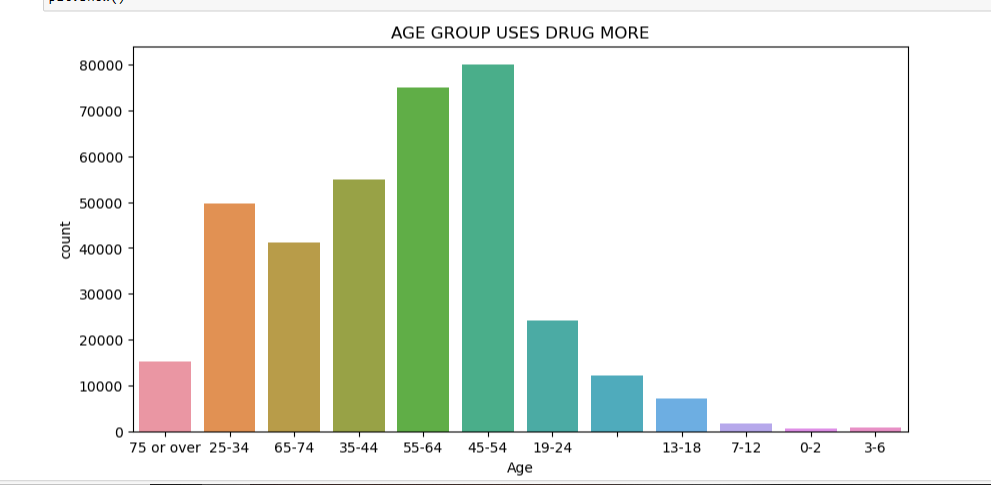
Some Exceptions/Exclusions encountered in our project are

* Non-Medical Advice: It makes no recommendations for medical diagnosis or treatment. Only research and informational purposes are served by it.
* Limited Drug Coverage: Due to data shortages or project scope restrictions, the project may concentrate on a small number of drugs, drug classes, or side effects.
* Assessment of Severity: Only the presence or absence of side effects may be evaluated , not their severity or intensity.

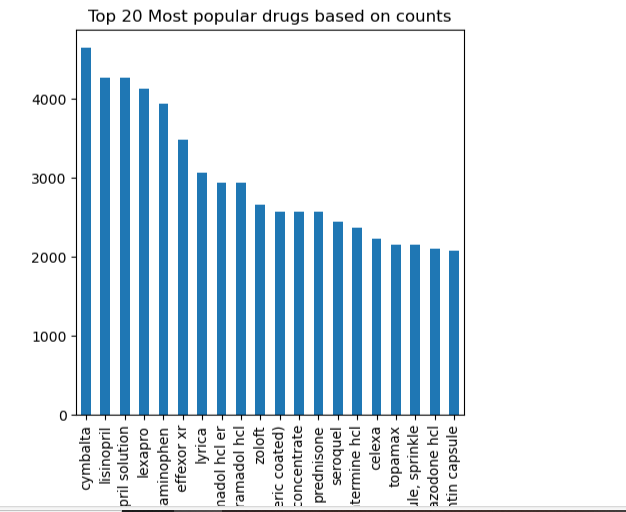
**Charts, Table, Diagrams**

****

Here the gender distribution is visible,we can see there is another data entry which is neither male or female in the gender column.



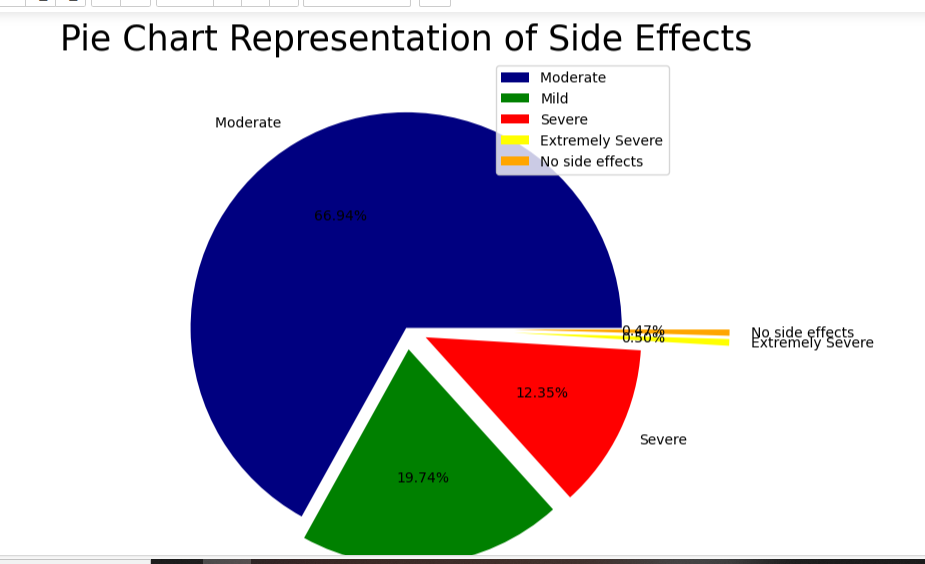
Here we can see that age 45-54 uses more drugs than others



 The most commonly used drug is Cymbalta.

• The top four drugs has count of reviews around 4000 and above.

• If we look at the top 20, around 2000 people has given reviews on each drug

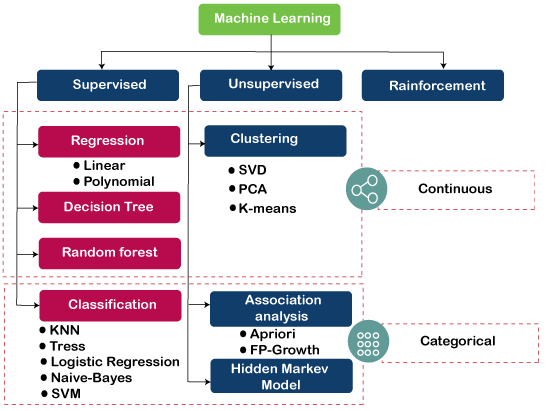


**Algorithms**

Machine Learning algorithms are the programs that can learn the hidden patterns from the data, predict the output, and improve the performance from experiences on their own. Different algorithms can be used in machine learning for different tasks, such as simple linear regression that can be used **for prediction problems** like**stock market prediction,**and**the KNN algorithm can be used for classification problems.**

Machine Learning Algorithm can be broadly classified into three types:

1. **Supervised Learning Algorithms**
2. **Unsupervised Learning Algorithms**
3. **Reinforcement Learning algorithm**



**Logistic Regression**

Logistic regression is the supervised learning algorithm, which is used to predict the categorical variables or discrete values. It can be used for the classification problems in machine learning, and the output of the logistic regression algorithm can be either Yes or NO, 0 or 1, Red or Blue, etc. Logistic regression is similar to the linear regression except how they are used, such as Linear regression is used to solve the regression problem and predict continuous values, whereas Logistic regression is used to solve the Classification problem and used to predict the discrete values.

**Decision Tree Algorithm**

A decision tree is a supervised learning algorithm that is mainly used to solve the classification problems but can also be used for solving the regression problems. It can work with both categorical variables and continuous variables. It shows a tree-like structure that includes nodes and branches, and starts with the root node that expands on further branches till the leaf node. The internal node is used to represent the features of the dataset, branches show the decision rules, and leaf nodes represent the outcome of the problem.

### Random Forest Algorithm

Random forest is the supervised learning algorithm that can be used for both classification and regression problems in machine learning. It is an ensemble learning technique that provides the predictions by combining the multiple classifiers and improve the performance of the model.

It contains multiple decision trees for subsets of the given dataset, and find the average to improve the predictive accuracy of the model. A random-forest should contain 64-128 trees. The greater number of trees leads to higher accuracy of the algorithm.

To classify a new dataset or object, each tree gives the classification result and based on the majority votes, the algorithm predicts the final output.

### NAIVE BAYES

Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

### BERNOLI NAIVE BAYES

Bernoulli Naive Bayes is a part of the family of Naive Bayes. It only takes binary values. The most general example is where we check if each value will be whether or not a word that appears in a document.In case of small amount of data or small documents(for example in text classification), Bernoulli Naive Bayes gives more accurate and precise results as compared to other models

**Challenges and opportunities**

**Challenges**

Data Quality and Availability: Access to high-quality labeled data for drug side effects can be limited, making it challenging to train accurate models.

Ambiguity in Textual Data: Drug side effect descriptions in text can be ambiguous, making it difficult to extract meaningful features.

Clinical Heterogeneity: Challenge: Patients may experience side effects differently due to factors like age, gender, and pre-existing conditions.

Ethical and Privacy Concerns:Challenge: Handling sensitive patient data requires strict privacy and ethical considerations.

Regulatory Compliance:Challenge: Developing a model for healthcare applications requires adherence to strict regulatory standards.

**Opportunities**

Improved Patient Safety: Accurate side effect classification models can help healthcare professionals and patients make informed decisions about drug usage, ultimately enhancing patient safety.

Drug Development:Identifying side effects early in the drug development process can save time and resources, leading to safer and more effective medications.

Personalized Medicine:Opportunity: Machine learning models can enable personalized treatment plans, considering individual patient characteristics and genetic factors.

Early Detection:Models can help detect and predict side effects even before they become severe, allowing for timely intervention.

Reduced Healthcare Costs: By preventing adverse side effects, these models can contribute to reduced healthcare costs associated with treating complications.

Research and Collaboration:Collaboration between healthcare professionals, researchers, and data scientists can drive innovation in this field, leading to better models and more accurate predictions.

**Outcome / Conclusion**

In conclusion, developing a model for classifying the side effects of drugs is a complex and important endeavor. While it comes with its set of challenges, including data quality, class imbalances, and interpretability issues, there are numerous opportunities for significant positive impacts on healthcare and patient well-being.

Accurate side effect classification models offer the potential to enhance patient safety by providing healthcare professionals and patients with informed insights into the potential risks associated with specific medications. Additionally, these models can streamline drug development by identifying side effects early, thus saving time and resources and contributing to the creation of safer and more effective medications.

Despite the challenges, the field of drug side effect classification holds great potential for making a significant and positive impact on healthcare and public health. As data quality, model interpretability, and ethical considerations are addressed, we can expect continued progress in this critical area of healthcare and artificial intelligence.

**Link to code and executable file**

https://github.com/Blessyshino/Internship-on-classification-model-side-effects-of-drugs/blob/main/TcsiON%20internship%20(1).ipynb

**Research questions and responses**

1. How can we improve the accuracy of drug side effect classification models, especially for rare or underrepresented side effects?

Response: This question underscores the challenge of class imbalance in side effect data. Possible research directions could involve exploring advanced techniques for handling class imbalances, such as cost-sensitive learning, resampling methods, or generative adversarial networks to generate synthetic examples of rare side effects. Additionally, improving the quality and quantity of labeled data, perhaps through active learning or crowd-sourcing, could contribute to enhanced model accuracy.

2. How can we maintain privacy and ethical standards while handling sensitive patient data for drug side effect classification?

Response: Research in this area can explore privacy-preserving machine learning techniques, including federated learning, secure multi-party computation, and homomorphic encryption. These methods allow data to be analyzed without exposing individual patient records. Ensuring compliance with regulations like GDPR and HIPAA and conducting ethical impact assessments are also critical research directions.

3. What impact does patient heterogeneity, including factors like age, gender, and comorbidities, have on drug side effect prediction?

Response: This question highlights the need for personalized medicine. Research can delve into methods for stratified analysis, which considers patient characteristics in side effect prediction. Understanding how these factors influence side effect risk and developing models that account for them can be pivotal in improving patient safety and outcomes.