Dear Intern

Interim project report is an inherent component of your internship. We are enclosing a reference table of content for the interim project report.

The key objective of this report is for you to capture how far you have got in completing the internship work against milestones expected to be achieved within a specific duration and seek the mentor’s feedback. Depending on the internship project and your progress (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections or leave some sections blank from the table of content mentioned below. You can also add additional sections. You can refer the project presentation to view the milestones related to your internship project. Please populate milestone# (1 / 2 / 3) and the milestone description in the interim project report based on the milestone for which you are submitting the interim project report.

You can refer the project presentation to view the milestones related to your internship project.

|  |  |
| --- | --- |
| 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 |
| 15.06.2023 | 19.06.2023 | | 23.5 hours | | VS code, Google colab | Python |
| Milestone # | 1 | Milestone: | | Create, clean and sanitize a dataset and carry out the preprocessing of data. | | |

**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**

* Classify 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**

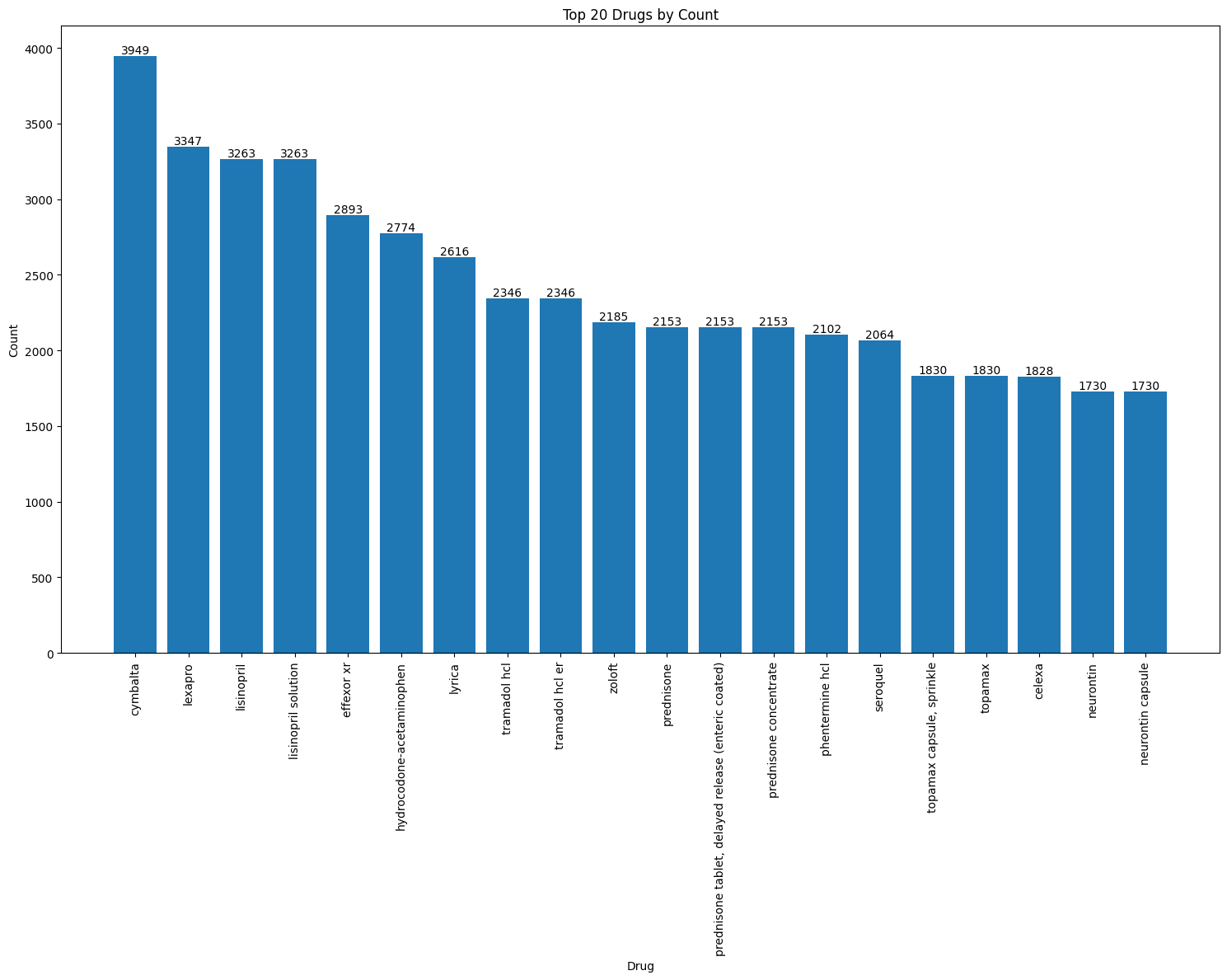
* Data Collection
* Preparation of the dataset (WebMD dataset on drugs review is the base here).
* Datasets usually contain raw and disoriented data. For the better performance of the model, data should be cleaned.
* Pre-process the data.
* Carry out exploratory data analysis.
* Feature selection.
* 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**

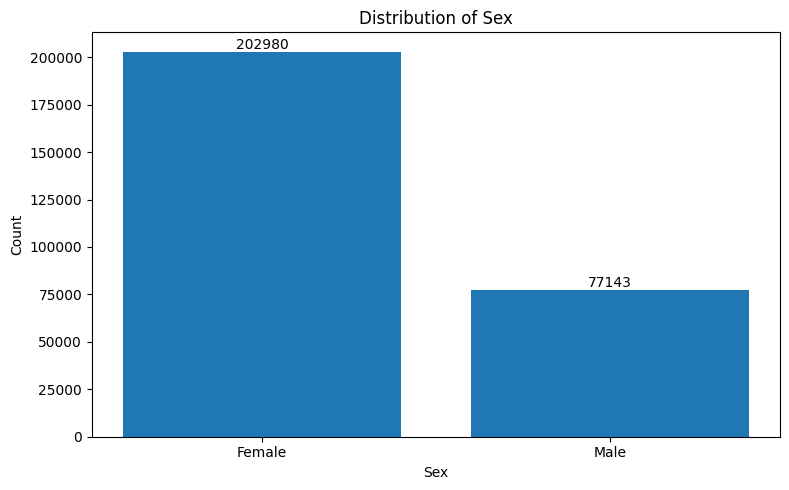
**Exceptions / Exclusions**

**Charts, Table, Diagrams**

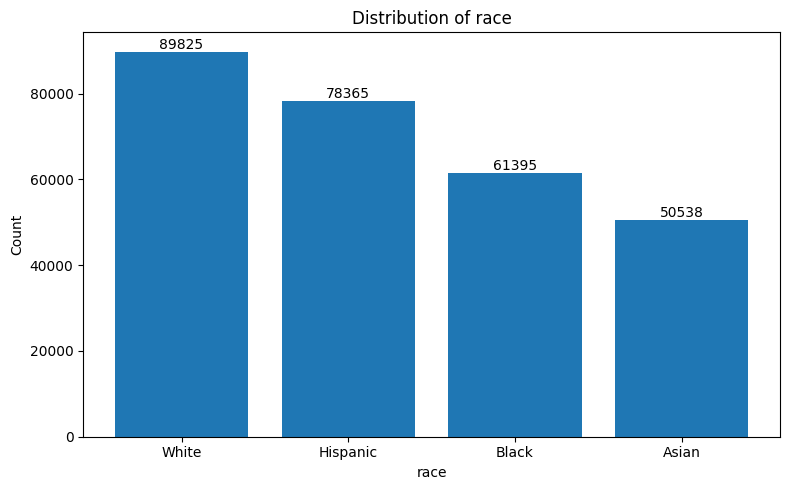
1. Top 20 Drugs by count

****

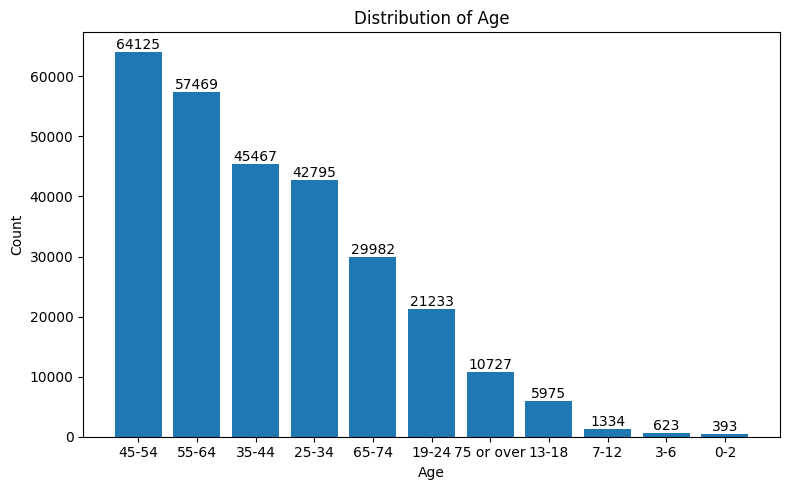
* Unique values : 5014
* Number of unique identities with count 1: 1046,
* Top counts
* cymbalta 3949
* lexapro 3347
* lisinopril 3263
* lisinopril solution 3263
* effexor xr 2893



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



* White – 32%
* Hispanic – 27%
* Black – 21%
* Asian – 18%



Highest drug usage was among 45-54 age group and the least among 0-2.

**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**