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 | RIO-125:Classification Model: Build a Model that Classifies the Side effects of Drugs. |
| Name of the Company | TCS ion |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | ICT KERALA OF ACADEMY |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 1/09/22 | 30/09/22 | 125 | Remote internship | Word,python,loom |

**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
* **Objective** :

Medicine Side-effect Analysis : To build a classification model that classifies side effects of a particular drug by Age, Race and Gender

* **Introduction:**

Machine learning (ML) is an innovative approach that has extensive applications in classification, prediction, forecasting etc.. Machine learning techniques are being used in a variety of areas such as medicine, engineering, education, manufacturing and production, forecast, traffic management, robotics etc. It is one of the most advanced concepts of artificial intelligence (AI), and provides a strategic approach to developing automated, complex and objective algorithmic techniques for multimodal and dimensional biomedical or mathematical data analysis. Machine learning has already shown potential in pharma and medicine for finding ways to effectively collect and use lots of different types of data for better analysis, prevention, and treatment of individuals.

Healthcare is an important industry which offers value-based care to millions of people. Healthcare specialists and stakeholders around the globe are looking for innovative ways to deliver on quality, value and outcome. Machine learning (ML) based applications embedded with real-time patient data available from different healthcare systems in multiple countries can increase the efficacy of new treatment options which were unavailable before. It has found wide applications in precision medicine and personalised treatments. Using ML techniques, side effects of drugs both beneficial and adverse can be classified into categories. That can help to make smarter decisions for precision medicine, personalised treatments and to repurpose the drug. Drug classifiers based on side effects can also be an informational resource designed to assist licensed healthcare practitioners in caring for their patients and/or to serve consumers viewing this service as a supplement to, and not a substitute for, the expertise, skill, knowledge and judgment of healthcare practitioners.

A side effect is usually regarded as an undesirable secondary effect which occurs in addition to the desired therapeutic effect of a drug or medication. Side effects may vary for each individual depending on the person's disease state, age, weight, gender, ethnicity and general health. Side effects can occur when commencing, decreasing/increasing dosages, or ending a drug or medication regimen. Side effects may also lead to non-compliance with prescribed treatment. When side effects of a drug or medication are severe, the dosage may be adjusted or a second medication may be prescribed. Lifestyle or dietary changes may also help to minimize side effects. It is a challenging task to classify the side effects for each drug. Machine learning techniques can make such tasks easier to do without compromising the accuracy. Pharmacogenetic research in the past few decades has uncovered significant differences among racial and ethnic groups in the metabolism, clinical effectiveness, and side-effect profiles of many clinically important drugs. These differences must be taken into account in the design of cost management policies such as formulary implementation, therapeutic substitution and step-care protocols. These programs should be broad and flexible enough to enable rational choices and individualized treatment for all patients, regardless of race or ethnic origin.

This work includes machine learning models for classifying side effects of drugs based on age and gender using a dataset with user generated text. Data was acquired by scraping the WebMD site. WebMD is an organization which provides information, support and reference material about health subjects through a team of doctors and health experts across a broad range of specialty areas. The dataset includes both demographic and clinical data. It contains 362806 instances and 12 features including categorical, numerical and text data. Dataset provides user reviews on specific drugs along with related conditions, side effects, age, sex, and effectiveness reflecting overall patient satisfaction. The structure of the data is that a patient with a unique ID purchases a drug that meets his condition and writes a review and rating for the drug he/she purchased on the date. Afterwards, if the others read that review and find it helpful, they will click usefulCount.

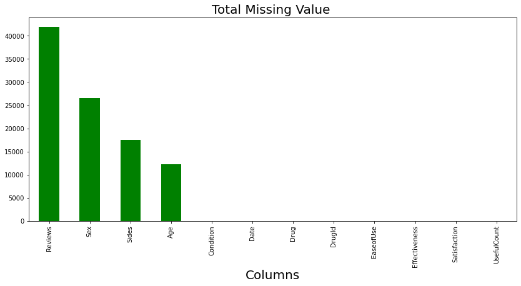
* **Approach/Methodology:**

1. Understanding the dataset
2. Data cleaning and preprocessing

* Finding the missing values and handling it
* Outlier detection and handling

1. Exploratory Data analysis
2. Encoding the categorical data into numerical values using label encoding/one hot encoding.
3. Feature engineering and selection
4. Splitting the dataset into train data and test data
5. Building the model
6. Fine tuning
7. **Data collection and getting Data set details**: Once we get any data set during the project ,we should know the meaning of each and every single feature in it.Following are the steps involved in getting the data set details: a)Find the number columns ,number of rows. b)What is training and testing data set ratio? c)Find the dimension of the data. d)Know what all different data types in the data set.
8. **Data Preprocessing**:

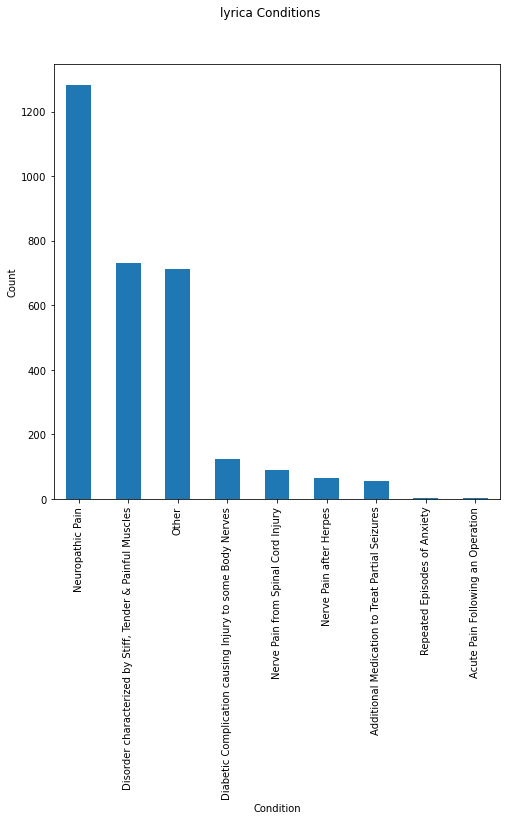
Data preprocessing is noting but cleaning the data and removing unwanted material out of your data like noise,duplicate records,inconsistancy,missing values.Following are the steps involved,Data Cleaning step plays a very important role.Once data cleaning is done we check which column is having and NA(Null) values,If there are any null values fill them with different imputation techniques such as Mean,Median,Mode Imputation techiniques.we check the if there is any outlier,and remove it using IQR method.also I do some feature reduction also using VIF method.

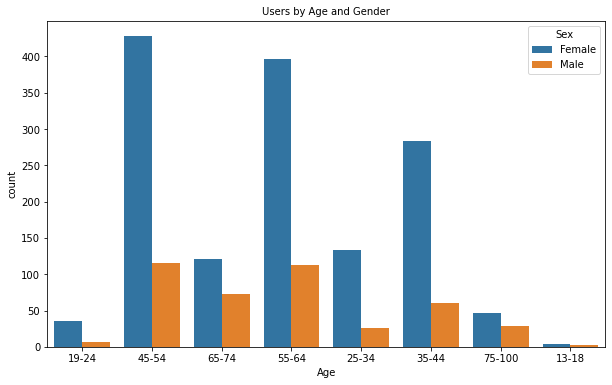


**3)EDA :(Exploratory Data Analysis)**

Here we are exploring and vizualizing data.Try to understand the pattern in your data.Here we try to understand relationship of features with the outcome.EDA step is very important part in any type of project.EDA is noting but Exploratory data analysis.Here we are getting insights from the data.Exploratory Data Analysis (EDA) helps in understanding the data sets by summarizing their main characteristics and plotting them visually.This step is very important before we apply and Machine learning algorithm.Following are the steps involved: a)We should know about Business moments such as Measure of Central tendency(Mean,Median.Mode),Measure of dispersion(Variance,Standatd deviation,Range),Skewness,Kurtosis. b)Check whether the data is normally disrtibuted or not.If it is not Normally distributed apply scaling technique. c)Graphical representation :Represent the data set with the help of Graphs such as Line plot,Bar plot,Histogram plot,Pie chart,Box plot and try to get inferences/insights from it. d)we should find the correlation in the data set.All the features should be independent to each other.

Some of the EDA’s





1. **Model Building step:**

First step in Model building is partitioning the data.Here we alredy have train and test data set seprately. In our project we are dealing with classification probelm i.e o find whether the given drug has a side effect or not with the help of reviews.Following are the models used for Model building in this project: 1)Logistic regression. 2)Decision tree classifier. 3) Random forest classifier. 4) Naive Bayes classifier . 5) Support Vector Machine(SVM) Classifier.(Linear SVM)(polynomial svm)(Radial svm). 6) Gradient boost .7)Extreme Gradient boost. 8)light bgm. 9)KNN model.From the model I found that Random forest is the best model.

**6)Evalute and Compare Performance :**

In this step we finalised the models based on accuracy ,precison,recall and f1 score values.We get all these values from Confusion Matrix.We are checking here the performance of the classification models.And also we check for the errors in the different models.We are comparing the actual value versus predicted values.After fine tuning I got the maximum accuracy of 67.46% for Random forest .

* **Internship activities:**

**Day wise plan**:learning schedule gives an idea on how to spend day and how much time to devote to each module.

**Pre-Test**:A simple test for a quick knowledge check.1 hour duration marked for 50 and 20 as passing mark.pre test is a general aptitude test,needed to be taken before commencing work on the project.

**Self-Learning**: References to different learning resources to enable students to successfully do project,and enrich learning.

**Recorded Webinar**: Detailed description of the industry project,including hands-on environment,expected outcome,and introducing industry guide.

**Digital Discussion Room**: Candidates can interact with each other and industry mentor regarding the project Via Digital Discussion room.

**Activity Report**: Daily activity report for documenting progress including learning,project activities,mentor and peer interactions.

* **Reccomendations:**

Based on the derived EDAs, the following are recommended with regard to the use of lyrica:

1) The drug is safe for use by those up to 24 years (maximum nil side effects)

2) Females in general are better to use the drug under a physician’s guidance given the high chances of extremely severe side effects.

* **Conclusion:**

Performed various pre-processing steps and obtained a clean dataset for exploratory data analysis and modelling. Various visualisations were done on the dataset and selected a particular drug lyrica. Maximum users consume this drug for neuropathic pain. The best classification model was obtained by Random Forest. On performance evaluation, maximum accuracy of 67.46% has been achieved. By race, Whites topped in the usage of this drug with a higher female ratio. Domination of female users were consistent across age groups above 12 and below 75 years. Also, people in the age groups 45 to 54 accounted for maximum usage of lyrica. Most users found lyrica as highly effective, especially females. Most rated it highly for its ease of use as well. Side effects of the drug proved race neutral. Genderwise, side effects proved extreme in females.

* **Link to code and executable file:**

<https://www.loom.com/share/5d5cd65261ad4016844de7051da25a6d>