|  |  |
| --- | --- |
| Name Of The Student | BASIL CHACKO MATHEW |
| Internship Project Topic | RIO-125: Classification Model - Build a Model that Classifies the Side Effects of a Drug |
| Name of the Organization | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | ICT academy kerala |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 10/04/2022 | 22/05/2022 | 120 | Anaconda | Excel, jupyter |

TABLE OF CONTENT

* Acknowledgements
* Objective
* Introduction
* Internship Activities
* Methodology
* Assumptions
* Exceptions / Exclusions
* About the dataset
* Attribute information
* Charts, Diagrams
* Prepare data for training
* Algorithms
* Reflections on the Internship
* Conclusion
* Link to code and executable file

**Acknowledgement**

The internship opportunity I had with TCSion was a great chance for learning and professional development. I express my deepest thanks to Debashis Roy, Industry Mentor for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make life easier.

It is my radiant sentiment to place on record my best regards, deepest sense of gratitude to all faculty member of ICT Academy of Kerala for their careful and precious guidance which were extremely valuable for my study both theoretically and practically.

I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

**Objective**

In the Pharma industry, the most common and high-priority question to be answered is “whether a particular drug has side effects over various types of people based on the reviews in the dataset !”

Now a days massive data generated from the search engines has widened the perspective of the market research and analysis in the drug data. With the help of other parameters we will predict whether a drug is safe or not. Its success is based on the available data on <https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29> . The main given objective of this project to build a classification model that classifies the side effects of a particular drug by reviews. The model need to have good amount of accuracy and have to meet the industry standards.

**Introduction**

The project guidelines clearly mentioned that we are expected to create a model that classifies the trial data of a drug based on their age, gender and race. We also entrusted to create a dataset of 4143 patients containing the following details for each patient based on various attributes according to the data. At the end of the project we should be able to create a dataset, clean the dataset, sanitize it and preprocess the data to perform data partitioning and handle missing values. Create training and testing sets. Build a classifier and fit the data to the model

**Internship Activities**

The activity mainly concentrates on how we make up to the objective of the internship. The given resources were very useful to kick start our internship and the day wise plan helps us to calculate the overall time and amount of work to be done each day and what extra we can do about it. We can explore different aspects of this data which vary from EDA to the final prediction model for the 30days

**Methodology**

The Methodology used here will be the Linear Strategy which consist in sequential phases with no feedback loops. The project solution is not released until the final phase is reached. This strategy is characterized by clearly defined goal solution and requirements, zero or few change request of the scope, routine and repetitive process inside the project, use of pre-established formulas and templates. Main objective of the project will be combining the three reviews as the independent variable making as x with y as side effects. The pre-defined steps includes data cleaning, EDA, PCA, data preprocessing , feature processing, splitting to test and train set, applying machine learning algorithms, comparison of machine learning algorithms. Opting the best prediction model

**Assumptions**

By various Exploratory data analysis we can come an assumption that the drug are rated good for the body by chemist, it have a slight side effect of the dataset mainly for depression .The condition attribute mainly concentrate on insomnia, depression and other mental problem related to brain issues. So taking has condition which have a less correlated value with respected to other features so dropped it. The main issues was accuracy with respect to the output of the model based on various trial and errors came to an assumption that the reviews which was three feature or columns were concluded to one as ‘combined Review’.

**Exceptions**

Exceptions were made as many when considering the transform of textual data to a numerical data. Compared a lot of nltk libraries which is natural language processing tool kit which is used in preprocessing of the dataset, had to except the list to string function where an immersive effects on the accuracy of the machine learning model had. The .values method of the pandas libraries had latter issues converting the string to n dimension array explicitly reconsidered when normalization and scaling of the feature variable.

**About the dataset**

The dataset provides patient reviews on specific drugs along with related conditions. Reviews and ratings are grouped into reports on the three aspects benefits, side effects and overall comment.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate, Text | **Number of Instances:** | 4143 | **Area:** | N/A |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 8 | **Date Donated** | 2018-10-02 |
| **Associated Tasks:** | Classification, Regression, Clustering | **Missing Values?** | N/A | **Number of Web Hits:** | 62471 |

The dataset provides patient reviews on specific drugs along with related conditions. Furthermore, reviews are grouped into reports on the three aspects benefits, side effects and overall comment. Additionally, ratings are available concerning overall satisfaction as well as a 5 step side effect rating and a 5 step effectiveness rating. The data is split into a train (75%) a test (25%) partition (see publication) and stored in two .tsv (tab-separated-values) files, respectively.

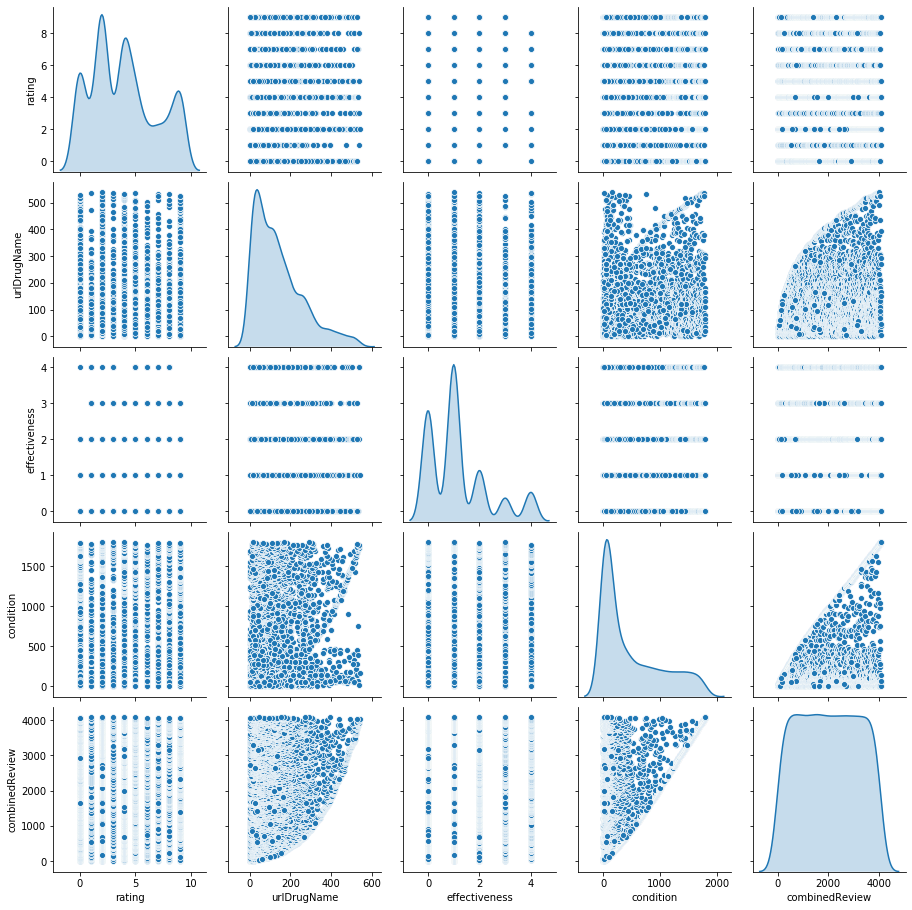
**Attribute Information:**

1. urlDrugName (categorical): name of drug  
2. condition (categorical): name of condition  
3. benefitsReview (text): patient on benefits  
4. sideEffectsReview (text): patient on side effects  
5. commentsReview (text): overall patient comment  
6. rating (numerical): 10 star patient rating  
7. sideEffects (categorical): 5 step side effect rating  
8. effectiveness (categorical): 5 step effectiveness rating

**Charts, diagram**

****

Correlation matrix shows the parameters of the dataset of different ranges than [-1, 1]. Where we will remove the low and high correlated feature from the dataset so looking at the graph we can consider that the review feature have an good correlation between them so approached by making the three reviews as one has combinedReview thus not considering others



To plot multiple pairwise bivariate distributions in a dataset we can use the **pairplot**() function. This shows the relationship for (n,2) combination of variable in a Data Frame as a matrix of plots and the diagonal plots are the univariate plots. Our reluctant pair plot is adversely dispersed so a linear model will not work well it need classifier like random forest classifier which can make linear separation in our data-set.

**Prepare data for training**

Two tasks will be performed which is resultant data is then divided into training and test sets. Training and Testing Data The data is split into training (75%) and testing (25%) data sets through random sampling. The following machine learning techniques were considered in the experiment. As the dataset is labeled properly, it is considered to be used for supervised learning. In order to find out the best machine learning technique, different machine learning techniques were tested and based on the RMSE further decision was taken.

**Algorithms**

**1. Random Forest Classifier**: an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. The Random Forest Classifier class of the sklearn.ensemble library is used to Classifier problems via random forest. The most important parameter of the Random Forest Classifier class is the n estimator’s parameter. This parameter defines the number of trees in the random forest. We will start with n estimator=20 to see how the algorithm performs

**2. Naive bayes**: the algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It’s easy to implement and understand, but has a major drawback of becoming significantly moderate as the size of that data in use grows which shown -0.48 as accuracy

**3. SVM(Support Vector Machine):** The main reason to use an SVM instead is because the problem might not be linearly separable. In that case, we will have to use an SVM with a non linear kernel (e.g. RBF).Another related reason to use SVMs is if you are in a highly dimensional space. SVM have been reported to work better for text classification. Where in our case it gives a 0.48 accuracy

**Conclusion**

Systematic reviews and meta-analyses are at the top of the hierarchy of evidence as they should be based on rigorous and reproducible methods for synthesizing a comprehensive dataset, and there exists some previous advice regarding methods for the systematic assessment of adverse effects but are nevertheless interested in conducting unbiased evaluations of adverse effects.  So considering our dataset attributes we derived 3 model where random forest classifier has the highest accuracy which is 52% accuracy score.

|  |  |
| --- | --- |
| accuracy of  random forest regressor | 0.52 |
| accuracy of  SVM | 0.48 |
| accuracy of  naive bayes | 0.48 |

**Reflections on the Internship:**

Had a slow start but catches up in the middle learned a lot about textual data transform combining of words to a numerical data frame. Studied about nltk tools and libraries. The webinar helped a lot to grasp the idea of the reference dataset. Great mentor support from tcsion.

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

https://github.com/BasilcM/TCSPROJECT