|  |  |
| --- | --- |
| Name of the Student | Bryan Menezes |
| Internship Project Topic | Classification Model for Drug trial |
| Industry Mentor | Dr. Pradheep |
| Institute Name | CSMU-Chhatrapati Shivaji Maharaj University |

TCS iON Remote internship

RIO-210

|  |  |  |
| --- | --- | --- |
| Date | Day | Hour Spent |
|  | DAY 5 | 6.5 |
| Activities done today—Industry project work:  Data preprocessing:    This code is used to remove leading and trailing white space from all the string columns in a Data frame. This is done to ensure there are no unexpected spaces that could affect data analysis.    Using for col [‘Age’,’Condition’,’Sex’,’Reviews’]:  df=df[(df[col].astype(bool)&df[col].notnull())]  we can filter out the other section in the Sex column.  The pie graph shows us now how much the gender distribution between only male and females.    We define a function “sentiment polarity that used Textblob module to calculate the sentiment polarity and then we apply this function to the “Reviews” column of the data frame to create “sentiment score”.    We are mapping data into categories based on sentiment score. And data1 is used to store data into a new column.    Using print(df[‘side effects’].value\_counts()) we find out how how many patients have what type of side effects.  We use df,drop to drop DrugID, reviews and Side as DrugID was not necessary, Reviews are replaced by sentiment scores and Sides replaced by side effects.    This pie graph is created to show the percentage of side effects across the entire data set. This shows that Moderate is 69.02%, Mild is 19,46%, Severe is 10.69%, Extreme is 042% and no side effects is 0.41%.    Using df=df.drop we can remove “Others” from “Sex” column as it will help in beter for modeling.  Unising astype(int) and map we can change the Datatype of “Sex” column into integer namely making Male:0 and Female:1 simlarly changing Age data type form object to integer namely:  “0-2”:1  ”3-6”:2  ”7-12”:3  “13-18”:4  ”19-24”:5  ”25-34”:6  ”35-44”:7  “45-54”:8  “55-64”:9  “65-74”:10  “75 or over”:11  Changing the date format by using df['Date'] **=** pd**.**to\_datetime(df['Date']) from mm-dd-year to year-mm-dd.  Using from collections import Counter library to remove the drugs whose count is less than 5.    This value\_count piece of code is used to create a table that shows how many times each rating (0 to 6) has been given by three different aspect ‘Satisfaction’,Effectiveness’,and Ease of use.  Using value\_count we see that only 2 drugs have avalue point of 6. This means the 2 frugs are very good but since they are only 2 we will have to drop them as they may disturb the classification algorithm.    We can convert the Condition column from object to integerusing lable encoding. This allows for more classification models. | | |