

INTERNSHIP: PROJECT REPORT

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 Student	Vardaan Vishnu
Name of the Industry Mentor	Debashis Roy
Name of the Institute	VIT BHOPAL UNIVERSITY

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
7-11-22	2-01-23	125	VS Code, Chrome, Windows 10, GitHub	Python, Colab, libraries

TABLE OF CONTENT

- Acknowledgements
- Objective
- Introduction / Description of Internship
- Internship Activities
- Approach / Methodology
- Charts, Table, Diagrams
- Algorithms
- Challenges & Opportunities
- Reflections on the Internship
- Outcome / Conclusion
- Enhancement Scope
- Link to code and executable file

Acknowledgements

I thank TCS iON for the guidance and support as well as for providing necessary information regarding the project. I also thank my industry guide, Mr. Debashis Roy, and the other mentors for being helpful and co-operative during this internship.

Objective

The objective of this project is to develop a classification model that classifies the side effects of a particular drug by age, gender, and race.

Introduction / Description of Internship

The Purpose of this internship is to build a Classification Model. A Classification models are a subset of supervised machine learning. A classification model reads some input and generates an output that classifies the input into some category. Here we intend to use these tools to create a model that can categorize different side effects that can occur due to age, gender and race.

We use various tools and setup environments for the development of this model.

Since the goal here is to develop a classification model that classifies the side effects of a particular drug by age, gender, and race, we start with gathering data that will help us model the algorithm. We then proceed to clean, sanitize, and order that data into a concise and clean dataset. Afterwards of this we perform certain exploratory analysis in order to gather an idea about the data we have and make sense of it. Then we use Machine Learning algorithms to create models that classify the side effects caused due to age, gender, and race.

Internship Activities

Some of the activities to be done during the internship include Pre-project test, activity report, interim project reports, final project report, project test.

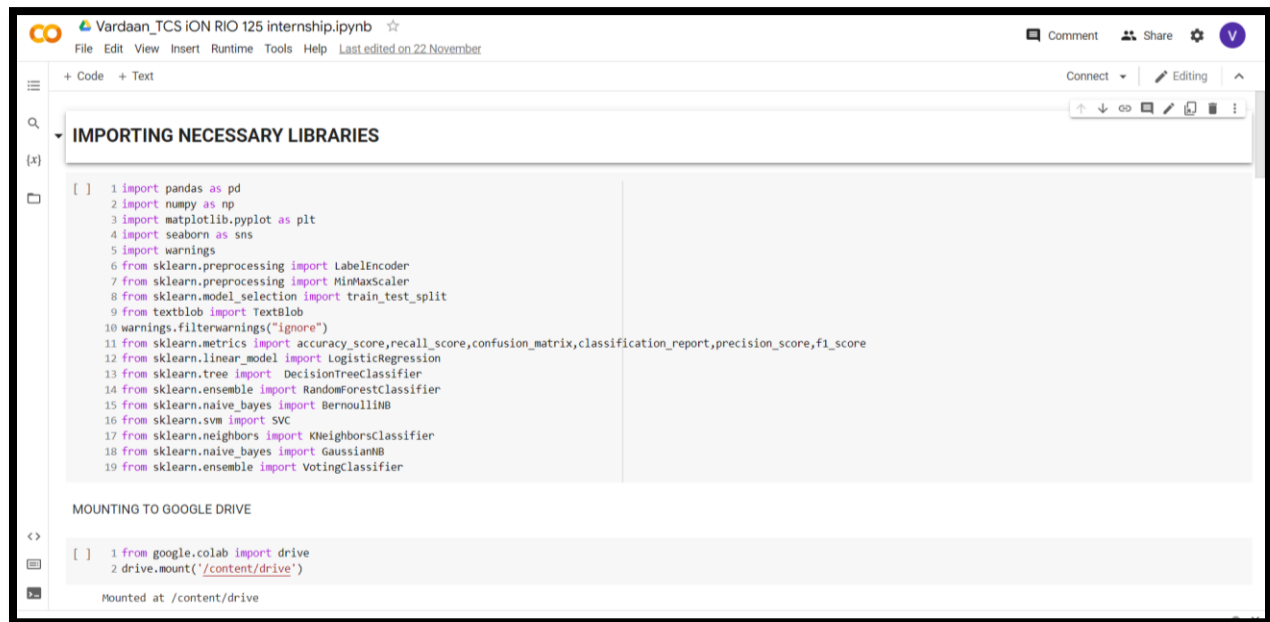
Approach / Methodology

1. Create a dataset. Here we use the WebMD dataset from Kaggle.
2. Clean the dataset and sanitize the data. Datasets usually contain raw and disoriented data and we need to clean it and make it proper for better performance with our classification model.
3. We pre-process the data according to our needs and utilize.
4. So, we remove the outliers, remove missing values, and then use the dataset.
5. We carry out exploratory data analysis.
6. We create various bar plots and pie charts to understand and have firm grasp about our dataset.
7. We then proceed to use only those parameters that help us achieve our goals,
8. Thus, irrelevant columns are not utilized further for data analysis and are dropped instead.
9. We Split the dataset into training and testing sets for the purpose of having different values when training our model and different values when testing it for real world.
10. We the create a Classifier that trains on our training dataset and make a model that fits the data.
11. Then we run our classification model that on our testing dataset and can make proper inferences from it too.
12. We use different algorithms to build our model so that we can gauge which one of the listed methods is best for creating our desired classification model.
13. We use logistic regression, Naïve-bayes, Random Forest Classifier, Decision Tree and K-Nearest neighbor algorithms to carry out the functionality.

INTERNSHIP: PROJECT REPORT

14. We have thus created our classification model and tested it too. Now we make the final analysis and the conclusion and our project is completed.

Charts, Table, Diagrams



The screenshot shows a Jupyter Notebook interface with the title "Vardaan_TCS ION RIO 125 internship.ipynb". The notebook is divided into two sections. The first section, titled "IMPORTING NECESSARY LIBRARIES", contains a code cell with the following Python code:

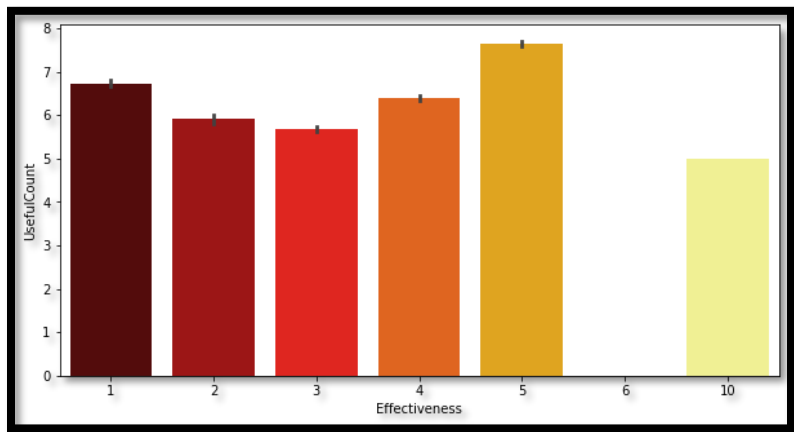
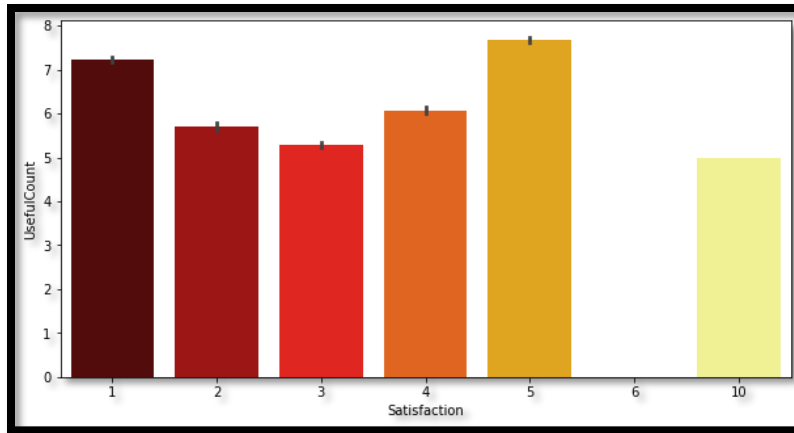
```
[ ] 1 import pandas as pd
    2 import numpy as np
    3 import matplotlib.pyplot as plt
    4 import seaborn as sns
    5 import warnings
    6 from sklearn.preprocessing import LabelEncoder
    7 from sklearn.preprocessing import MinMaxScaler
    8 from sklearn.model_selection import train_test_split
    9 from textblob import TextBlob
   10 warnings.filterwarnings("ignore")
   11 from sklearn.metrics import accuracy_score, recall_score, confusion_matrix, classification_report, precision_score, f1_score
   12 from sklearn.linear_model import LogisticRegression
   13 from sklearn.tree import DecisionTreeClassifier
   14 from sklearn.ensemble import RandomForestClassifier
   15 from sklearn.naive_bayes import BernoulliNB
   16 from sklearn.svm import SVC
   17 from sklearn.neighbors import KNeighborsClassifier
   18 from sklearn.naive_bayes import GaussianNB
   19 from sklearn.ensemble import VotingClassifier
```

The second section, titled "MOUNTING TO GOOGLE DRIVE", contains a code cell with the following Python code:

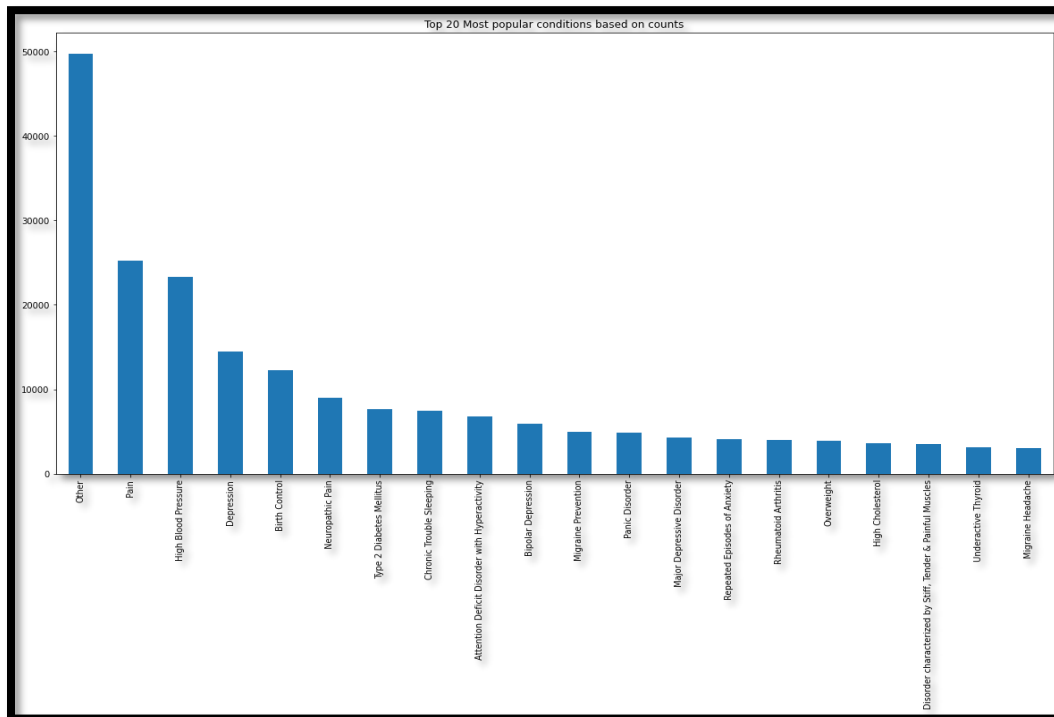
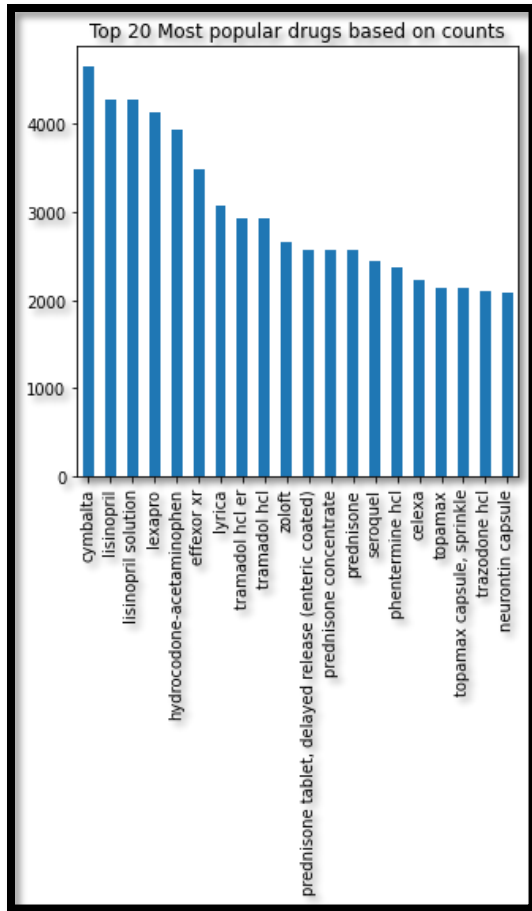
```
[ ] 1 from google.colab import drive
    2 drive.mount('/content/drive')
```

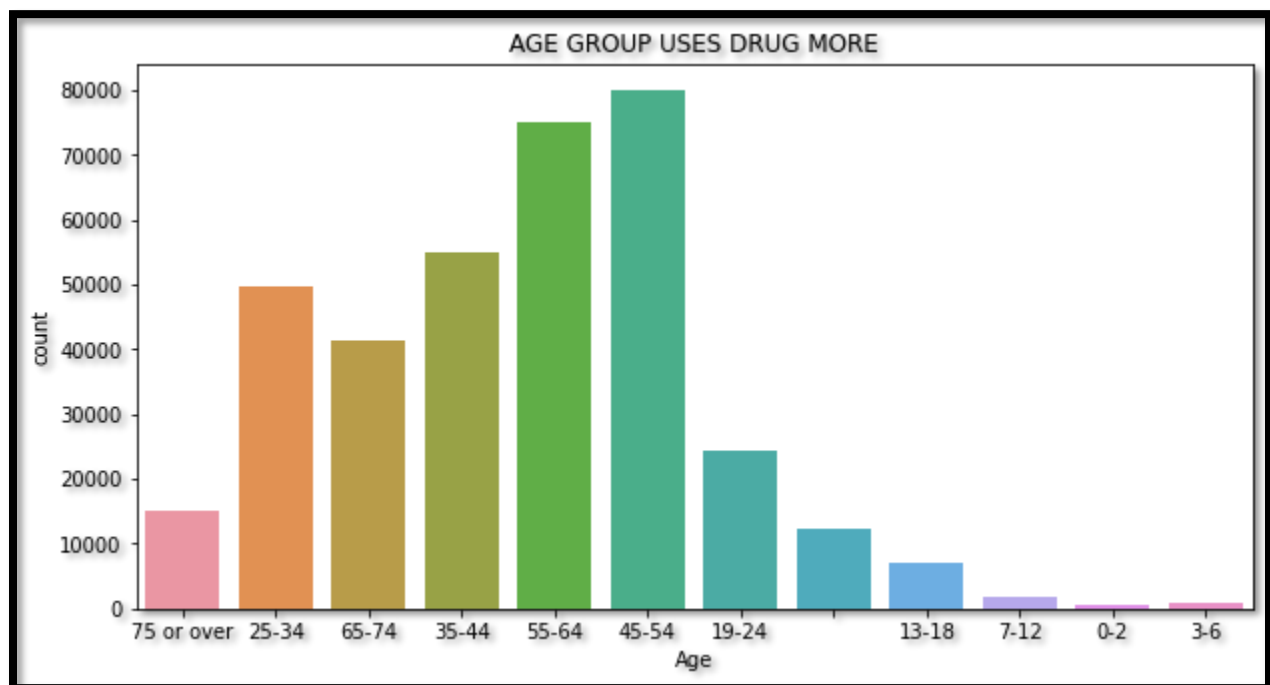
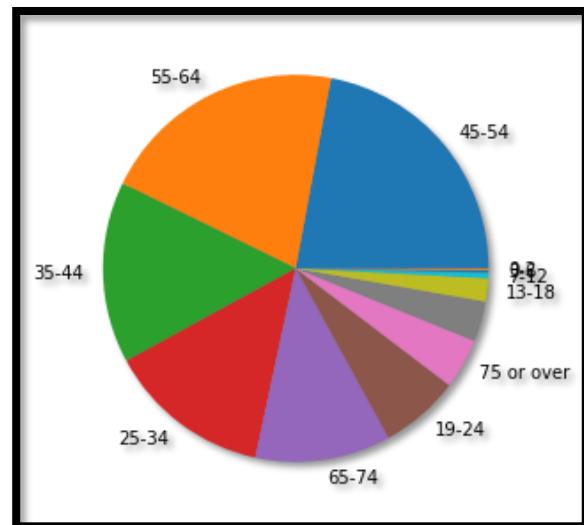
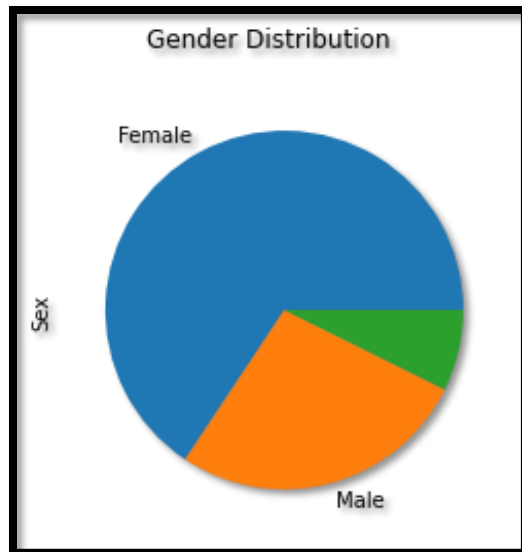
Below the code cells, the output shows "Mounted at /content/drive".

INTERNSHIP: PROJECT REPORT



INTERNSHIP: PROJECT REPORT





Algorithms

Some of the ML Algorithms used are Naïve Bayes Classification, K-nearest neighbor, Decision Tree, Random Forest, and Logistic Regression. They help us to build the classifier model.

Challenges and Opportunities

It is a big challenge to learn the basics of exploratory data analysis since the field is so wide and needs to be understood thoroughly to have favorable outcomes, it is also an opportunity to learn more and expand my capabilities through this learning experience.

Reflections on the Internship

I am hoping that working as an intern with TCS ion will allow me to gain greater knowledge of classification models and data analytics.

Outcome / Conclusion

As an outcome, we have completed the importing of various libraries, cleaning and sanitizing of dataset, exploratory analysis of the dataset and further splitting of dataset into the training and testing sets.

Then we created a simple classifier using logistic regression and trained it on our training set. Finally, we tested our model on the testing set and we got the accuracy of 69%.

Enhancement Scope

There is a lot that can be added to this classification model. We can add more parameters to this model to gauge in other external factors into consideration that we have not till now. We can also focus on adding an interface/GUI for normal users who do not possess the knowledge about data science, so that they can make use of our model with relative ease.

Link to code and executable file

<https://github.com/vardaan11/TCS-iON-RIO-125-Internship>

Link to Project Video

<https://www.loom.com/share/dee4386bbaf5401392aa6fed5e265967>