INTERNSHIP: PROJECT REPORT

Internship Project Title	Shakchi Prasad
Name of the Company	RIO-125: Classification Model - Build a Model that Classifies the Side Effects of a Drug
Name of the Industry Mentor	TCS iON
Name of the Institute	Debashis Roy
	PrepInsta Pvt Ltd.

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
1-05-2023	30-07-2023	125	Remote	• Google Colaboratory • Pycharm IDLE • Kaggle • YouTube • Research papers, etc.

TABLE OF CONTENT

Acknowledgements

I would like to thank TCS for the opportunity to be a part of the RIO – 125 Program. I am grateful for the staff at TCS for diligently curating the RIO experience and looking over my work on the same. Lastly, I am greatly thankful to my Institute – PrepInsta Pvt. Ltd. for creating this opportunity and allowing me to be go on this journey.

Objective

To build a Classification model that classifies the Side effects of a Drug.

Description of Internship

The TCS RIO – 125 is an initiative that allows people to complete an internship of 30 days over the course of 3 months, comprehensive of 125 hours. At its base it is designed to teach the intern to build an industry grade project in the above timeframe.

Internship Activities

Self-learning:

This activity is the backbone of the program. Self-learning via resources provided by TCS themselves as well as any other available source.

A considerable amount of the time devoted to this internship was spent learning the essentials of Machine Learning, Supervised Learning, Classification and the algorithms associated with it and so much more to build the knowledge base required to work on the assigned project.

Industry project building:

An Industry Grade classification model that classifies the side effects of a drug is the final outcome required of this program.

Once, the theoretical knowledge of the concepts of classification was gained, the next step was to focus on learning the steps of building a machine learning and project and properly executing them.

Theory

Machine Learning:

Machine Learning is a subset of artificial intelligence that involves training algorithms to learn from data and make predictions or decisions without explicit programming. It includes both supervised (classification, regression) and unsupervised (clustering, dimensionality reduction) learning techniques. Supervised learning deals with labeled data for classification or regression tasks, while unsupervised learning finds patterns in unlabeled data without explicit guidance.

Classification and its Algorithms:

Classification is a supervised learning task where the goal is to assign predefined labels (classes) to data instances based on their features.

Ensemble Learning:

Ensemble learning combines multiple models to improve prediction performance. Techniques like Bagging, Boosting, and Stacking aggregate individual model predictions, reducing overfitting and enhancing overall accuracy.

Support Vector Machines (SVM):

SVM is a powerful supervised learning algorithm used for classification and regression tasks. It creates a hyperplane that maximizes the margin between classes, making it effective for complex decision boundaries.

k-Nearest Neighbors (KNN):

KNN is a simple classification algorithm that assigns labels to new data points based on the majority class of their nearest neighbors in the feature space.

Random Forests:

Random Forests use an ensemble of decision trees to improve prediction accuracy. Each tree is built on random subsets of data and features, reducing variance and providing robust predictions.

Decision Trees:

Decision trees recursively split data based on features to create a tree-like structure, making it easy to interpret and understand how predictions are made.

Evaluation Metrics:

Evaluation metrics measure model performance. Accuracy gauges overall correctness, precision measures true positive rate, recall assesses sensitivity, F1-score balances precision and recall, and the confusion matrix presents the results.

Project Building:

For the drug side effect classification project, the process begins with data preprocessing, involving cleaning, encoding, and scaling the dataset. Next, feature engineering selects relevant attributes and optimizes the feature space. The ensemble training model combines powerful classification algorithms like SVM, KNN, Decision Trees, and Random Forests to achieve accurate predictions and robust generalization. Hyperparameter tuning ensures optimal model configurations, followed by rigorous evaluation using appropriate metrics.

Data Preprocessing:

Data preprocessing involves cleaning, scaling, and transforming data to prepare it for modeling. Tasks include handling missing values, encoding categorical variables, and normalization.

Model Selection and Hyperparameter Tuning:

Model selection involves choosing the best algorithm for the task, while hyperparameter tuning optimizes model parameters using techniques like cross-validation and grid search.

Overfitting and Underfitting:

Overfitting occurs when a model fits the training data too closely, leading to poor generalization. Underfitting happens when the model is too simple to capture underlying patterns, resulting in low performance. Techniques like regularization and cross-validation help mitigate these issues.

Industry Project

Abstract:

This drug side effect classification project presents an ensemble training model that combines SVM, KNN, Random Forests, and Decision Tree algorithms to predict drug side effects accurately. The model's social benefits lie in its ability to enhance drug safety evaluation, thereby reducing adverse reactions and promoting public health. By leveraging the combined strengths of diverse algorithms, the ensemble model offers improved accuracy (>94%) and robustness, making it invaluable for pharmaceutical research and healthcare decision-making. The project's findings signify its potential application in optimizing drug safety assessment, benefiting patients, healthcare providers, and society at large.

Literature survey:

https://core.ac.uk/download/pdf/145239566.pdf
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0774-y

https://www.ibg.uu.se/digitalAssets/147/c_147658-l_3-k_report-xiaodong-liu.pdf
https://www.researchgate.net/publication/315845346

The above research papers and articles were used to get an understanding on the solutions already existing in the field on the same topic.

Dataset:

The data required to build the model culminated in its final form is known as the dataset.

It was gathered over the course of a few days in the project and thoroughly pre-processed to convert into the dataset shown below, i.e., a form that would be most affective to train the model on its details of effectiveness and make the correct prediction of its side effects (Mild, Severe, Negligible, etc.)

The dataset used in this project has 3106 records and the following columns:

1. Drug Name
2. Rating
3. Effectiveness
4. Condition
5. Side-effects

	urlDrugName	rating	effectiveness	condition	sideEffects
0	enalapril	4	Highly Effective	management of congestive heart failure	Mild Side Effects
1	ortho-tri-cyclen	1	Highly Effective	birth prevention	Severe Side Effects
2	ponstel	10	Highly Effective	menstrual cramps	No Side Effects
3	prilosec	3	Marginally Effective	acid reflux	Mild Side Effects
4	lyrica	2	Marginally Effective	fibromyalgia	Severe Side Effects

The above snippet showcases the first 5 columns of the dataset. The entire dataset can be found at the link provided underneath.

Coding:

After the data gathering, cleaning and pre-processing is completed the actual coding of the project starts. This project is built on Python using the cloud based IDLE – Google Collaboratory. The link to the notebook has been attached below.

IMPORTING LIBRARIES

Importing the right data processing, visualization libraries (Numpy, Pandas, Matplotlib) as well as Sci-kit libraries to run the classification algorithms in Python.

IMPORTING DATASET

Once the apt libraries are imported, the dataset is imported into the notebook to start manipulation and model building.

MANAGING NULL VALUES

The characteristics (shape, description, etc.) of the data are cleared up, the dataset is checked for any NULL values present. The null value holding records are removed from the dataset as those records cannot be fed into the model.

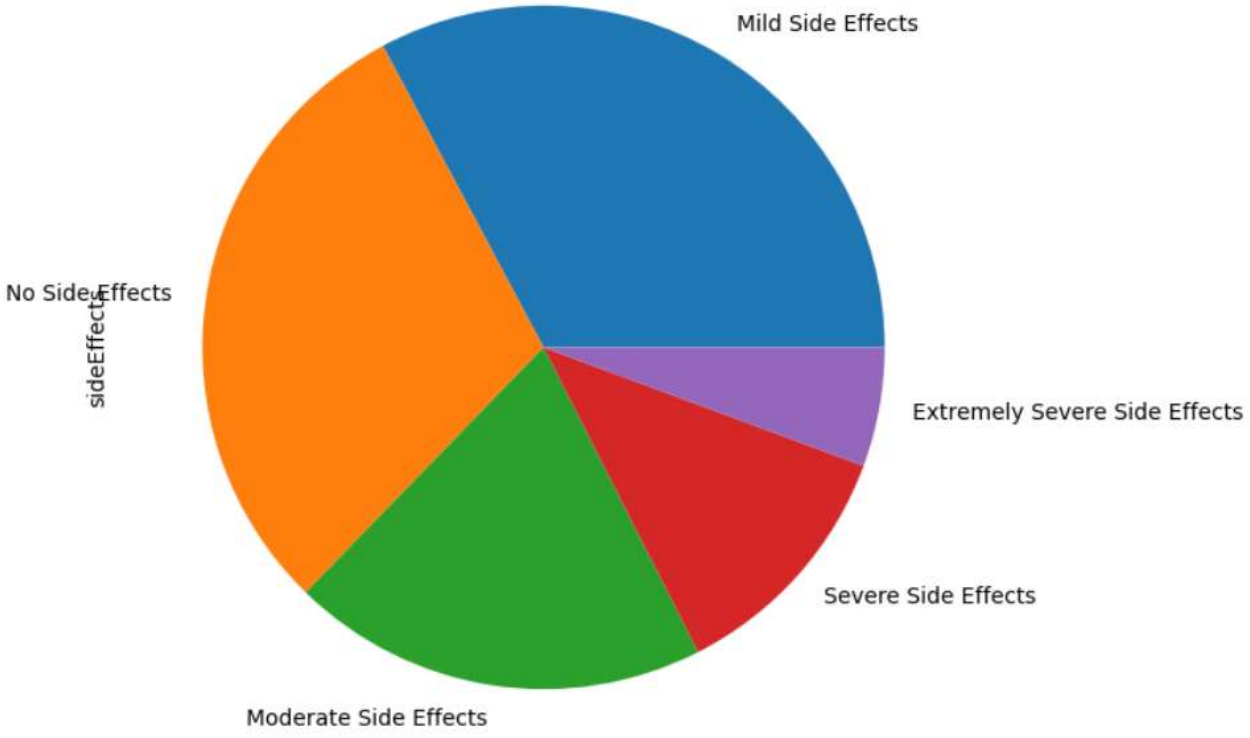
```
#showcases dimensions of data

df.shape

#dataset holds 3106 records, i.e data of 3106 drugs

(3106, 5)
```

DATA VISUALIZATION



The above pie-chart depicts the distribution of the Side effects present in the dataset.

Data Visualization is the step where the different features (columns) available are visualized using graphing techniques (Bar graphs, histograms, etc.) to get a better picture of the data and its statistical distribution, frequency, etc.

This is a very important step in feature selection.

ENCODING CATEGORICAL DATA

Conversion of text-based data into numerical counterparts for the model's benefit by assigning all unique values an integer.

	urlDrugName	rating	effectiveness	condition
0	0	4	0	0
1	1	1	0	1
2	2	10	0	2
3	3	3	1	3
4	4	2	1	4

The dataset snippet shown above, encoded into numerical form.

SPLITTING DATA

The data is then split into:

Training data – To teach the model; 70%

Testing data – To check the accuracy of the predictions of the model; 30%

BUILDING THE MODEL

The most important step in all of the project, building of the actual model. As this project uses a format of ensemble learning, a number of classification algorithms are employed.

The data is run through, SVM first, then Random Forest and Decision tree and then finally the K-Nearest Neighbor algorithm.

PREDICTION

The final step in the journey of the model is to get its prediction after all the training is done

ACCURACY

Finally, the accuracy of the model is calculated. The above model provides an accuracy of 94.17%.

The key tasks required to build the over-all project have been

Conclusion:

In conclusion, the drug side effect classification project successfully utilizes an ensemble training model to predict drug side effects effectively. The ensemble approach significantly enhances drug safety evaluation, leading to minimized adverse reactions and improved patient well-being. The model's accuracy and reliability foster trust in medical treatments, ultimately advancing public health. Healthcare providers and pharmaceutical researchers can utilize this approach to make informed decisions about drug usage, ensuring safer medications for the global population. This project's success highlights the significance of ensemble models in promoting healthcare standards and protecting patients from potential harm.

Future Scope:

The future scope of this project involves expanding the ensemble model to incorporate emerging machine learning techniques, such as gradient boosting and deep learning. Integrating these advanced algorithms could further elevate the model's predictive capabilities, leading to more precise drug side effect classification. Additionally, incorporating natural language processing and sentiment analysis can enhance the model's understanding of text-based drug information, enabling a comprehensive assessment of potential side effects. Furthermore, real-time monitoring of adverse drug reactions and post-marketing surveillance can be integrated into the model, providing continuous updates on drug safety profiles, which will be crucial for promoting safer medications and continually improving healthcare outcomes.

Reflections on the Internship

This internship turned out to be a wonderful opportunity as it allowed me to build a fully-functioning project wholly on my own merit. The research and self-learning as well as problem solving experience will surely be a skill highly beneficial for my future goals in the field of Machine Learning and Data Science.

REQUIRED LINKS

Loom Video:

https://www.loom.com/share/f178abd2a58d4adc9fc52a2045fc81fd?sid=4a5a9af4-ec0a-4ea2-ab6a-1bb693108f62

Git-Hub Repository:

https://github.com/Shakchi-Prasad/TCS-RIO-125---Drug-Side-effects-Classification

Google Collaboratory Notebook of The Project:

https://colab.research.google.com/drive/1e1XCoWvqewCRcPU8RNv4oSZajGQNIEkb?usp=sharing

Dataset:

https://docs.google.com/spreadsheets/d/110dJ2mMBMKkT8OWt1h1QCdBXwWTvysx93R-lcwOSQ9A/edit?usp=sharing