RIO-210: Build a classification model for Drug Trials Dataset

Internship report

Submitted to

Department of Mathematics & Statistics

Faculty of Science & Technology

Under TCS ION INDUSTRY HONOUR PROGRAM

Vishwakarma University, Pune (Maharashtra)

By

Gauri Bharat Wable

Under the supervision of

INDUSTRY MENTOR

MR. HIMDWEEP WALIA

TATA CONSULTANCY SERVICES

FACULTY MENTOR

DR. NAZIA WAHID

MS. RICHA PANCHGAUR

VISHWAKARMA UNIVERSITY, PUNE

CERTIFICATE

This is to certify that the project of titled "**RIO-210: Build a classification model for Drug Trials Dataset** "submitted by **Gauri Bharat Wable** is an original work and has not been previously submitted in part or full for the award of any degree or diploma to this or any other university. The project is submitted to **Vishwakarma University Pune and TCS-ION Industry Honor Program,** in partial fulfillment of the requirement for the award of the degree of **Master of Science** in the subject of **Statistics-Big Data Analytics**

Date:

Dr. Nazia Wahid

Faculty Mentor

Ms.Richa Panchgaur
Faculty Mentor

Head of Department

(Mathematics and Statistics)

Acknowledgement

I have satisfaction upon completion of this project work entitled “RIO-210: Build a classification model for Drug Trials Dataset” at the Department of statistics of ‘Vishwakarma University Pune’. The project is submitted to Vishwakarma University Pune and TCS-ION Industry Honor Program, during academic year 2023-2024.

I take this opportunity to express our gratitude to Industry mentor MR. HIMDWEEP WALIA. & Faculty mentor DR. NAZIA WAHID, Ms. RICHA PANCHGAUR Assistant Professor, Faculty of Science & Technology, Vishwakarma University, Pune for their valuable guidance and help provided us during the course of completion of the project.

I am thankful to our HOD for providing all necessary facilities, timely co-operation and her valuable guidance and help during the completion of my project.

DECLARATION

I,

Gauri Bharat Wable (202200175)

Here by declare that the work embodied in this project entitled “RIO-210: Build a classification model for Drug Trials Dataset” carried out by under the supervision of Industry mentor MR. HIMDWEEP WALIA, & Faculty mentor DR. NAZIA WAHID, Assistant Professor, Faculty of Science & Technology, Vishwakarma University, Pune is an original work and does not contain any work submitted for the award of any degree in this university or any other university.

Gauri Bharat Wable

M.Sc. Statistics-Big Data Analytics,
Department Of Mathematics & Statistics,
Vishwakarma University, Pune.

Internship Project Title	RIO-210: Build a classification model for Drug Trials Dataset
Name of the Company	TCSiON
Name of the Industry Mentor	Himdweep Walia
Name of the Institute	Vishwakarma University

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
14 April 2024	12 June 2024	227	Google colab	Python: Pandas, Numpy, Matplotlib, Seaborn, sklearn

TABLE OF CONTENT

- **Acknowledgements**

I would like to express my gratitude to Himdweep Walia, TCSiON, and all those who supported and guided me throughout this internship. Their insights and encouragement were invaluable.

- **Objective**

The objective of this project is to build a classification model to predict the side effects of a particular drug based on patients' Age, Gender, and Race.

- **Introduction / Description of Internship**

During this internship, I focused on creating a dataset, preprocessing data, and developing a machine learning model to classify drug side effects. The internship provided hands-on experience in data science and machine learning.

- **Internship Activities**

Data generation and preprocessing

Model training and evaluation

Data and analysis

Reporting and documentation

- **Approach / Methodology**

Data visualization Generation: Created a synthetic dataset with 400,000 patient records, including columns for Name, Age, Gender, Race, and Side Effects.

Data Preprocessing: Handled missing values, encoded categorical variables, and split data into training and testing sets.

Model: Used Logistic Regression and random forest classifier to build a classification model.

Model Evaluation: Evaluated the model using accuracy, precision, recall, F1-score, and confusion matrix.

- **Assumptions**

The dataset is synthetically generated and may not accurately represent the real population.

The selected features (Age, Gender, Race) are considered adequate for predicting side effects.

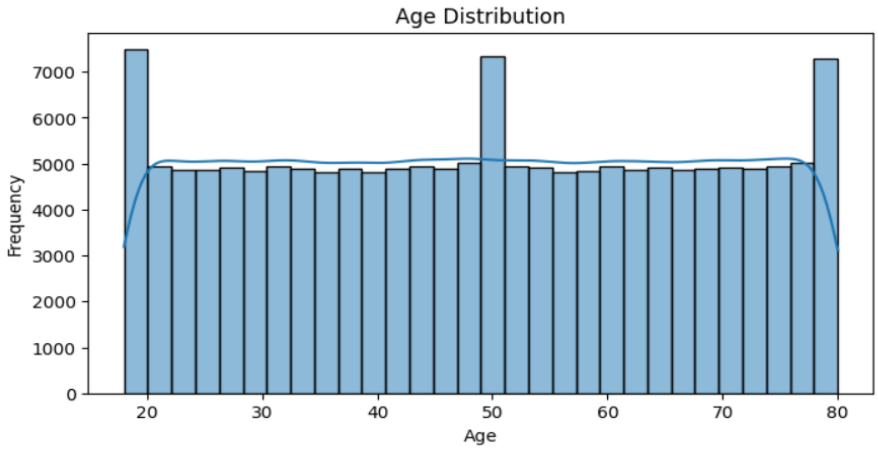
- **Exceptions / Exclusions**

External factors that could influence side effects but are not included in the dataset (e.g., medical history, lifestyle) were excluded from consideration.

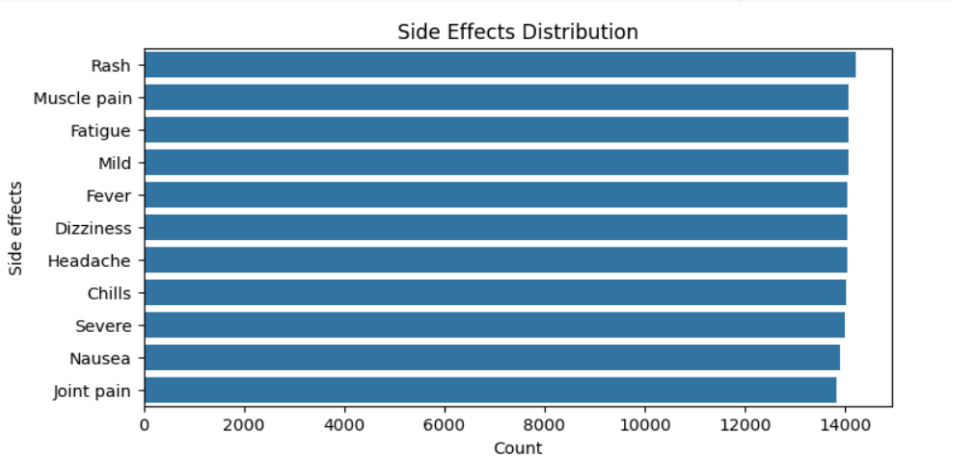
The dataset is synthetic and may not capture the complexities of the real world.

- **Charts, Table, Diagrams**

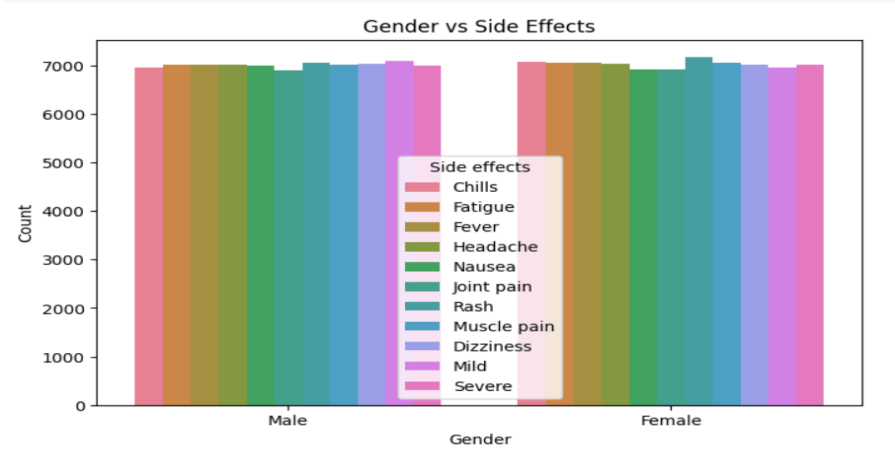
1. **Age Distribution:** The histogram shows the distribution of patient ages, indicating the age range of participants in the dataset



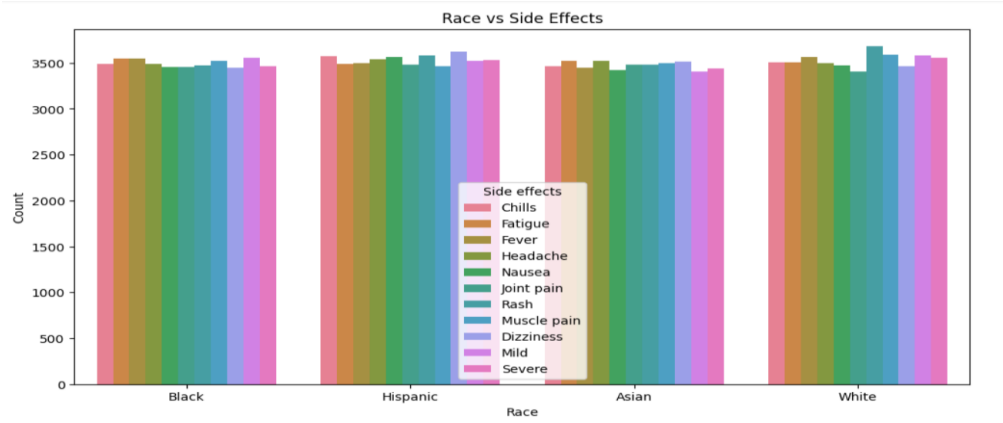
2. **Side Effects Distribution**: A count plot displaying the frequency of each side effect in the dataset.



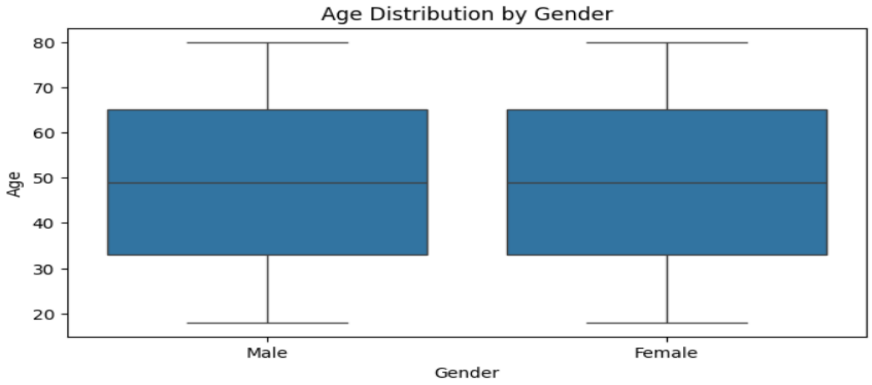
3. **Gender vs Side Effects:** Shows the distribution of side effects among different genders.



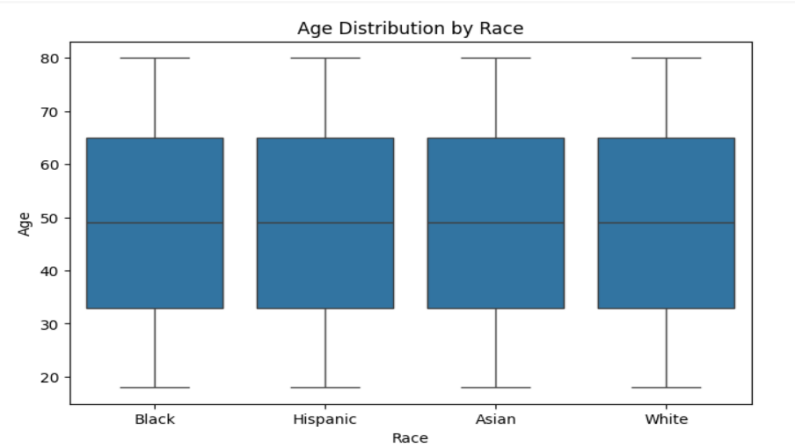
4. **Race vs Side Effects**: Illustrates how side effects vary across different racial groups.



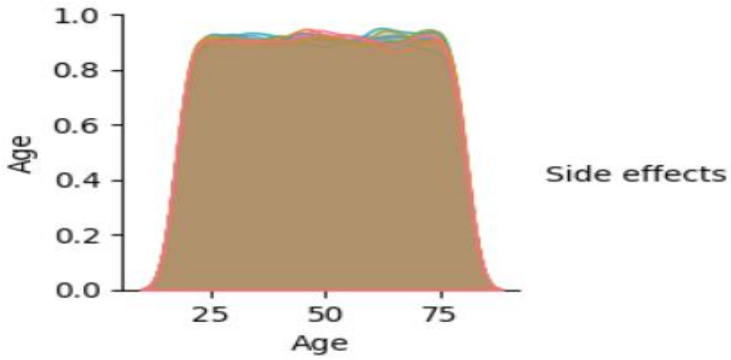
5. **Age vs Gender Box Plot**: Presents the distribution of ages by gender, highlighting any differences



6. **Age vs Race Box Plot**: Demonstrates the distribution of ages among different races.



7. **Pair plot: for side effects**



- **Algorithms**

1 **Categorical Encoding**: We use OneHotEncoder to change categorical features like names, gender, and race into numbers that the computer can understand.

2 **Pipeline and Classifiers**: We create a plan (pipeline) for each model (Logistic Regression, Random Forest, SVM) that includes the preprocessing step.

3 **Model Training and Evaluation**: We train each model and then check how well it works using accuracy, classification reports, and confusion matrices.

- **Challenges & Opportunities**

Challenges:

Ensuring the synthetic dataset accurately reflects real-world data.
Handling class imbalances in the dataset.

Opportunities:

Gaining practical experience with machine learning workflows.
Improving the model with additional features and data.

- **Risk Vs Reward**

Risk:

Potential inaccuracies in the synthetic dataset affecting model performance.

Reward:

Developing a robust model that can be used to predict drug side effects, contributing to healthcare

- **Reflections on the Internship**

This internship helped me learn a lot about building machine learning models and using data science methods. It was a great experience that improved my technical and analytical abilities.

- **Recommendations**

Integrate additional patient data (e.g., medical history) for a more accurate model.
Validate the model with real-world data for better reliability.

- **Outcome / Conclusion**

The project successfully created a classification model that predicts drug side effects with a good level of accuracy. The knowledge and skills acquired during this internship are incredibly valuable for future data science projects.

- **Enhancement Scope**

1.Use a variety of datasets: Include different types of data to make the model more versatile.

2.Try advanced algorithms: Experiment with more complex machine learning methods like Random Forest or Gradient Boosting.

3.Understand feature importance: Use techniques to better understand which factors are most important in the model's predictions.

- **Link to code and executable file**
  https://colab.research.google.com/drive/1TA_kVDDXuw_PNbV8tl68HwzxizUOyqCr?usp=sharing