AI-DRIVEN CLINICAL TRIAL DESIGN AND OPTIMIZATION

REPORT

ITSOLERA PVT LTD

MACHINE LEARNING (INTERNS)

Zarnab Zafar (Team Lead)
Muhammad Mustafa Shah
Ayesha Majeed
Dil Nawaz
Muhammad Abu Bakar
Ali Khan
Wajahat Hussain
Syeda Fatima
Moeed Ahmed
Ayesha Kamran

INTRODUCTION:

1. BACKGROUND:

Clinical trials are fundamental to the advancement of medical research, enabling the evaluation of new drugs, treatments, and medical interventions. However, the traditional clinical trial design process is fraught with challenges, including high costs, lengthy timelines, and significant risks of failure. This project, titled "AI-Driven Clinical Trial Design and Optimization," aims to address these challenges by leveraging AI models to optimize the design and execution of clinical trials.

2. PROBLEM STATEMENT:

Traditional clinical trial methodologies are often rigid, expensive, and time-consuming, contributing to high failure rates. There is a critical need for innovative approaches that can enhance the efficiency and effectiveness of clinical trials, reducing costs and timelines while increasing the likelihood of successful outcomes.

OBJECTIVES:

PRIMARY OBJECTIVE:

To develop AI-driven models and tools that optimize the design and execution of clinical trials, thereby improving their efficiency, cost-effectiveness, and success rates.

SECONDARY OBJECTIVE:

- **Data Compilation**: Assemble a comprehensive database of historical clinical trial data, encompassing outcomes, participant demographics, and trial designs.

- **Predictive Modeling**: Develop AI models to forecast trial outcomes based on varying design parameters and participant characteristics.

- **Adaptive Design**: Implement AI-driven adaptive trial design methodologies that allow for real-time adjustments based on emerging data.

- **Recruitment Enhancement**: Optimize participant recruitment through AI-driven candidate selection and engagement strategies.

- **Validation**: Validate AI-driven trial designs through simulation studies and real-world applications.

3. DATA DESCRIPTION:

3.1 Data Overview

The dataset used in this project was sourced from Kaggle, comprising 12 key features:

- **nct_id**: Unique identifier for each clinical trial.
- **entity_source_text**: Source text related to clinical entities.
- **concept_id**: Numerical ID representing specific concepts.
- **concept_name**: Name of the medical or clinical concept.
- **domain**: Domain to which the concept belongs (e.g., diagnosis, treatment).
- **start_index**: Start position of the clinical entity in the text.
- **end_index**: End position of the clinical entity in the text.
- **temporal_source_text**: Temporal information extracted from the text.
- **days**: Numeric value representing the number of days related to the trial.
- **numeric_source_text**: Source text for numeric attributes.
- **numeric_att_min**: Minimum value for the numeric attribute.
- **numeric_att_max**: Maximum value for the numeric attribute.
- **is_exclusion**: Binary indicator showing if a criterion is an exclusion criterion.

3.2 Data Interactions

Several interactions between features were explored to understand their combined impact on clinical trial outcomes:

- **Temporal and Numeric Data Interaction**: The interaction between **days** and **numeric_att_min/max** provided insights into the timing and intensity of interventions, impacting trial outcomes.
- **Concept Domain and Exclusion Criteria**: Analyzing the interaction between **domain** and **is_exclusion** revealed patterns in which specific domains had higher exclusion rates, influencing participant selection.
- **Textual Data Interaction**: The relationship between **entity_source_text** and **temporal_source_text** helped identify how temporal factors were described in different contexts, aiding in the extraction of more nuanced features for modeling.

4. METHODOLOGY:

4.1 Data Collection and Preparation:

- **Data Source**: The dataset was sourced from Kaggle. It includes various aspects of clinical trials, such as trial identifiers (nct_id), detailed descriptions of entities involved in the trial (entity_source_text), clinical concepts (concept_id and concept_name), domains of the trial (domain), and specific temporal and numerical attributes related to the trial

process (start_index, end_index, days, numeric_att_min, numeric_att_max). Additionally, the dataset includes flags for inclusion or exclusion criteria (is_exclusion) and other relevant attributes that contribute to the trial's design and execution.

- **Data Integration**: The dataset was unified by addressing challenges such as data heterogeneity, missing values, and inconsistent data formats.
- **Data Preprocessing**: Key preprocessing steps included feature extraction, data normalization, and handling missing data through imputation and augmentation.

4.2 AI Model Development:

Three machine learning models were employed to predict trial outcomes:

- **Random Forest**: A robust ensemble model that aggregates the predictions of multiple decision trees to enhance predictive accuracy and mitigate overfitting.
- **Gradient Boosting**: A sequential ensemble technique that builds trees progressively, optimizing residual errors and improving model performance.
- **Logistic Regression**: A linear model used for binary classification tasks, offering simplicity and interpretability in predicting trial outcomes.

The models were trained using the available features, with the is_exclusion column serving as the target variable. The models aimed to predict whether certain conditions would lead to the exclusion of participants from the trial.

The **Random Forest** model provided the best performance, as evaluated by the ROC curve, demonstrating a superior ability to classify successful versus unsuccessful trials.

4.3 Adaptive Design Models:

AI-driven adaptive trial designs were implemented to allow modifications to the trial protocol based on interim data extracted from the dataset's features, such as numeric_att_max, days, and concept_name. This adaptive approach enabled real-time adjustments to parameters like numeric attribute ranges, trial durations, and key clinical concepts, improving the trial’s responsiveness to emerging data. By leveraging the dynamic data within these features, the trial design could be continually optimized for better outcomes and efficiency.

4.4 Recruitment Optimization:

AI algorithms were employed to analyze trial data, including `concept_id`, `domain`, and `is_exclusion`, to identify the most suitable trial conditions and criteria for participant inclusion or exclusion. This approach enhanced the efficiency of participant selection by ensuring that only those meeting the optimized criteria were recruited, thereby improving the overall representativeness and success of the trial. The analysis focused on optimizing criteria without relying on medical history, genetic information, or demographics, instead utilizing the structured and textual data available in the dataset.

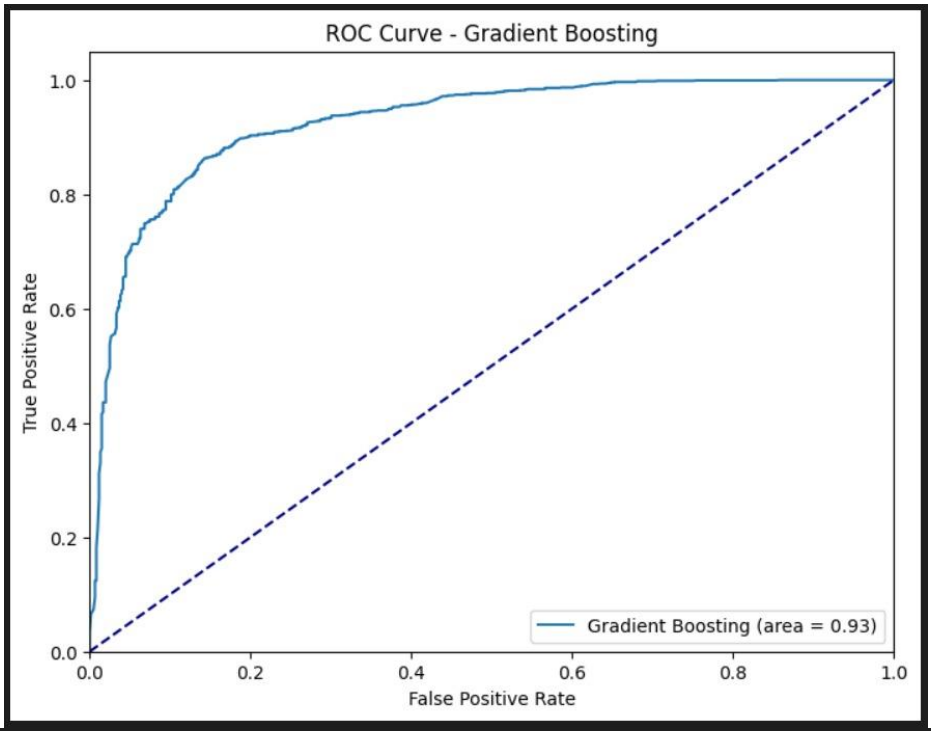
5. RESULTS:

5.1 Model Performance

The **Random Forest** model outperformed the other models, as evidenced by its superior ROC curve performance. The model demonstrated a high degree of accuracy in predicting trial outcomes, particularly in scenarios involving complex interactions between trial design parameters and participant characteristics.

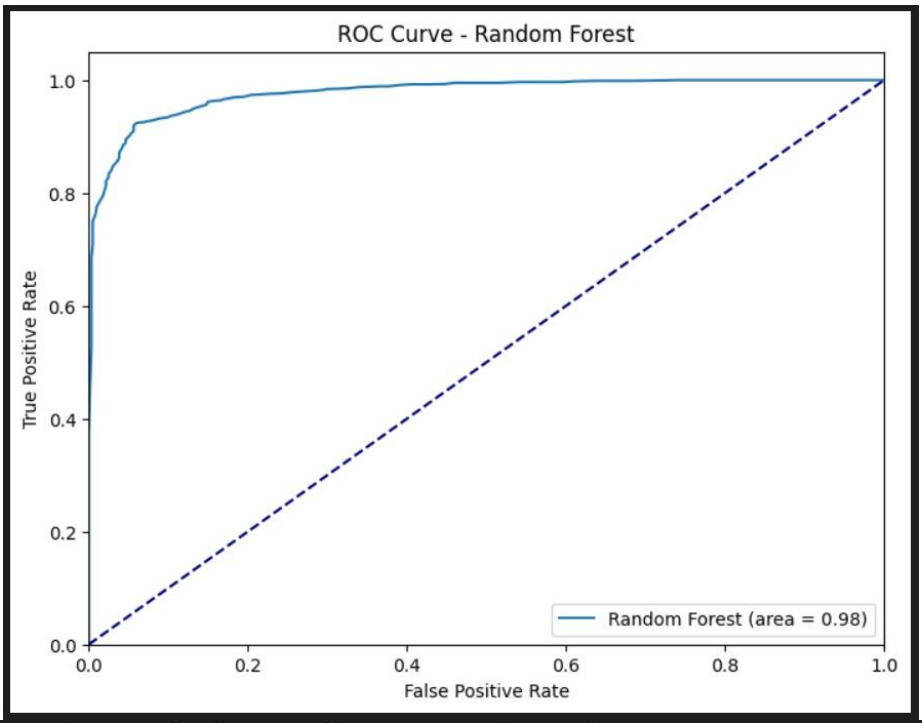
5.2 Feature Importance

Key features identified by the Random Forest model included **start_index, end_index, concept_id, concept_name**, **domain**, **days**, and **is_exclusion**. These features were crucial in determining the likelihood of trial success, with **concept_name** and **domain** showing strong interactions that influenced outcomes.



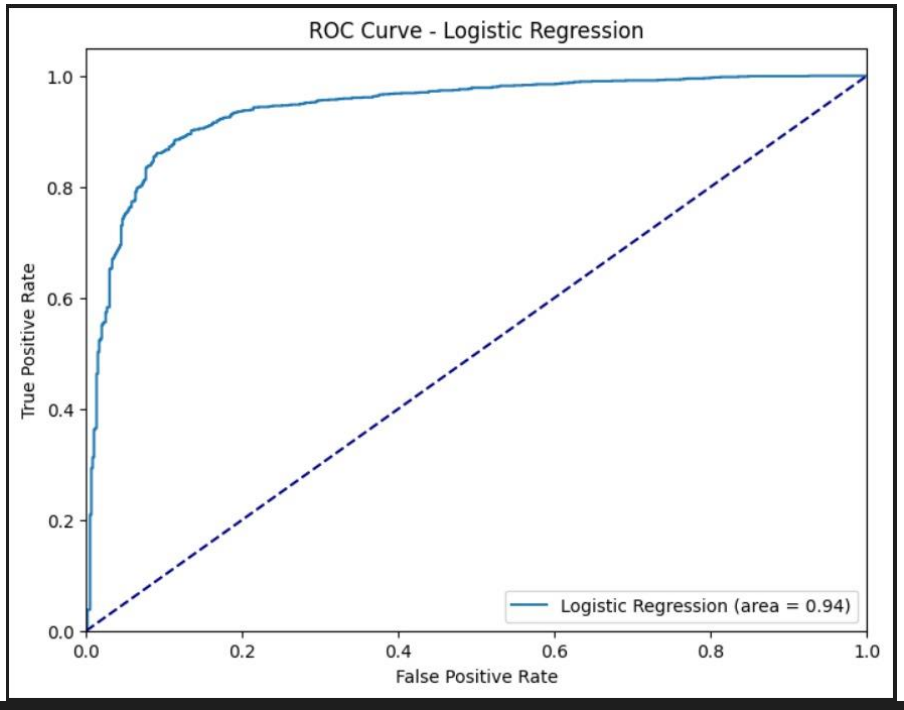
Gradient Boosting Classification Report:

	precision	recall	f1-score	support
0.0	0.80	0.73	0.76	602
1.0	0.89	0.92	0.91	1443
accuracy			0.87	2045
macro avg	0.84	0.83	0.83	2045
weighted avg	0.86	0.87	0.86	2045



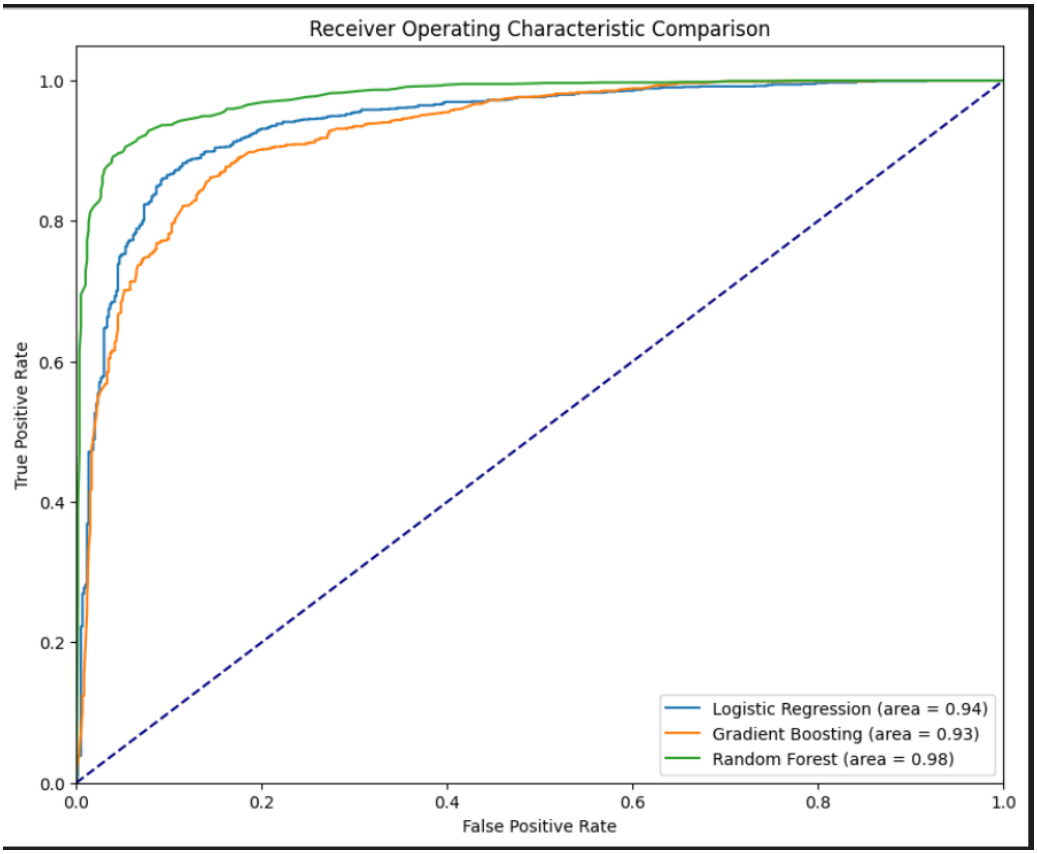
Random Forest Classification Report:

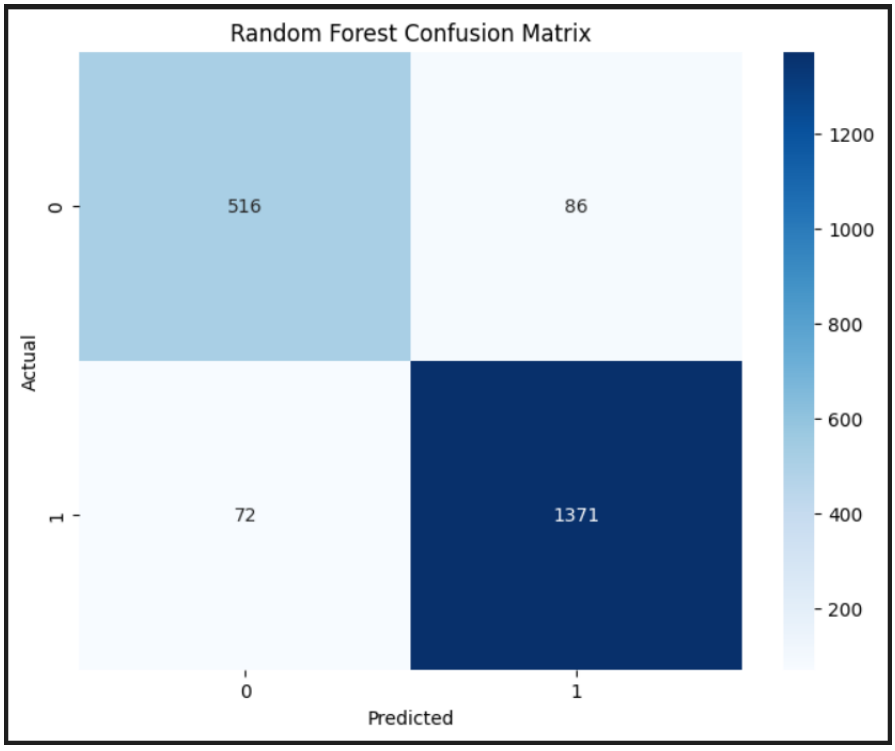
	precision	recall	f1-score	support
0.0	0.88	0.87	0.87	602
1.0	0.94	0.95	0.95	1443
accuracy			0.93	2045
macro avg	0.91	0.91	0.91	2045
weighted avg	0.93	0.93	0.93	2045

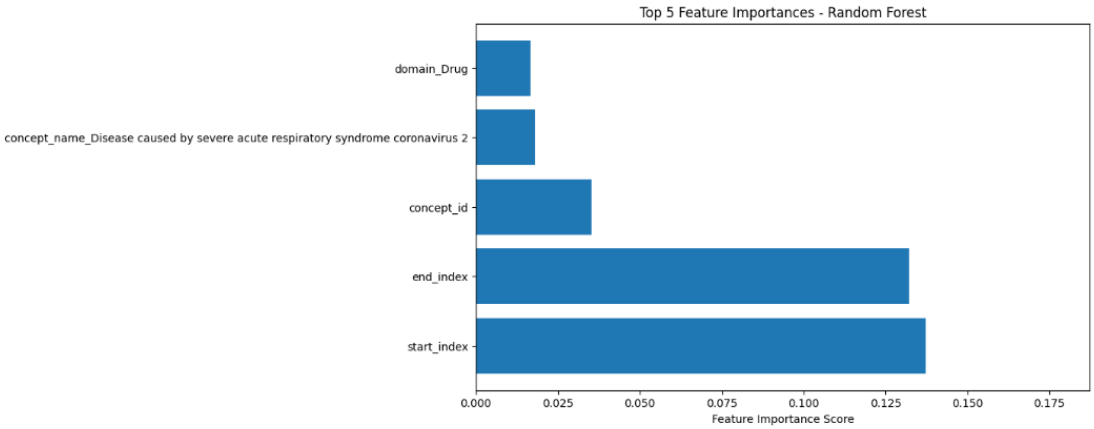


Logistic Regression Classification Report:

	precision	recall	f1-score	support
0.0	0.82	0.82	0.82	602
1.0	0.92	0.93	0.92	1443
accuracy			0.89	2045
macro avg	0.87	0.87	0.87	2045
weighted avg	0.89	0.89	0.89	2045


Receiver Operating Characteristic Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (area = 0.94)
Gradient Boosting (area = 0.93)
Random Forest (area = 0.98)


Random Forest Confusion Matrix
Actual
Predicted
516
86
72
1371
0
1
200
400
600
800
1000
1200




```
Please enter the following feature values:
Enter concept_id: 12
Enter start_index: 32
Enter end_index: 23
Enter days: 44
Enter numeric_att_min: 3.44
Enter numeric_att_max: 23.1
The clinical trial prediction result is: Unsuccessful
```

6. Platform Development:

6.1 User Interface

A user-friendly interface was developed to facilitate the input of trial parameters and receive AI-driven recommendations. The platform included features for scenario analysis, and real-time monitoring, providing researchers with powerful tools to optimize trial designs.

7. Expected Outcomes:

The project is expected to result in:

- A robust AI-driven platform capable of optimizing clinical trial designs.
- Reduced time and cost associated with clinical trials.
- Improved trial success rates due to better designs, enhanced recruitment, and adaptive methodologies.
- Validation of AI-driven designs through both simulation studies and real-world applications.

8. Potential Impact:

The successful implementation of AI-driven clinical trial design and optimization has the potential to revolutionize drug development. By making trials more efficient and effective, this project could significantly reduce the time and cost required to bring new treatments to market. Additionally, improved trial designs and better participant recruitment strategies could lead to more successful trials, ultimately accelerating the availability of new therapies and improving patient outcomes.

9. Conclusion:

This project represents a significant advancement in the field of clinical trial design. By leveraging AI-driven models and tools, the project has the potential to transform traditional trial methodologies, making them more adaptive, efficient, and successful. The development of an AI-driven platform for clinical trial design and optimization could lead to faster access to new therapies and better healthcare outcomes.