



# NEST

Nurturing Excellence,  
Strengthening **Talent.**

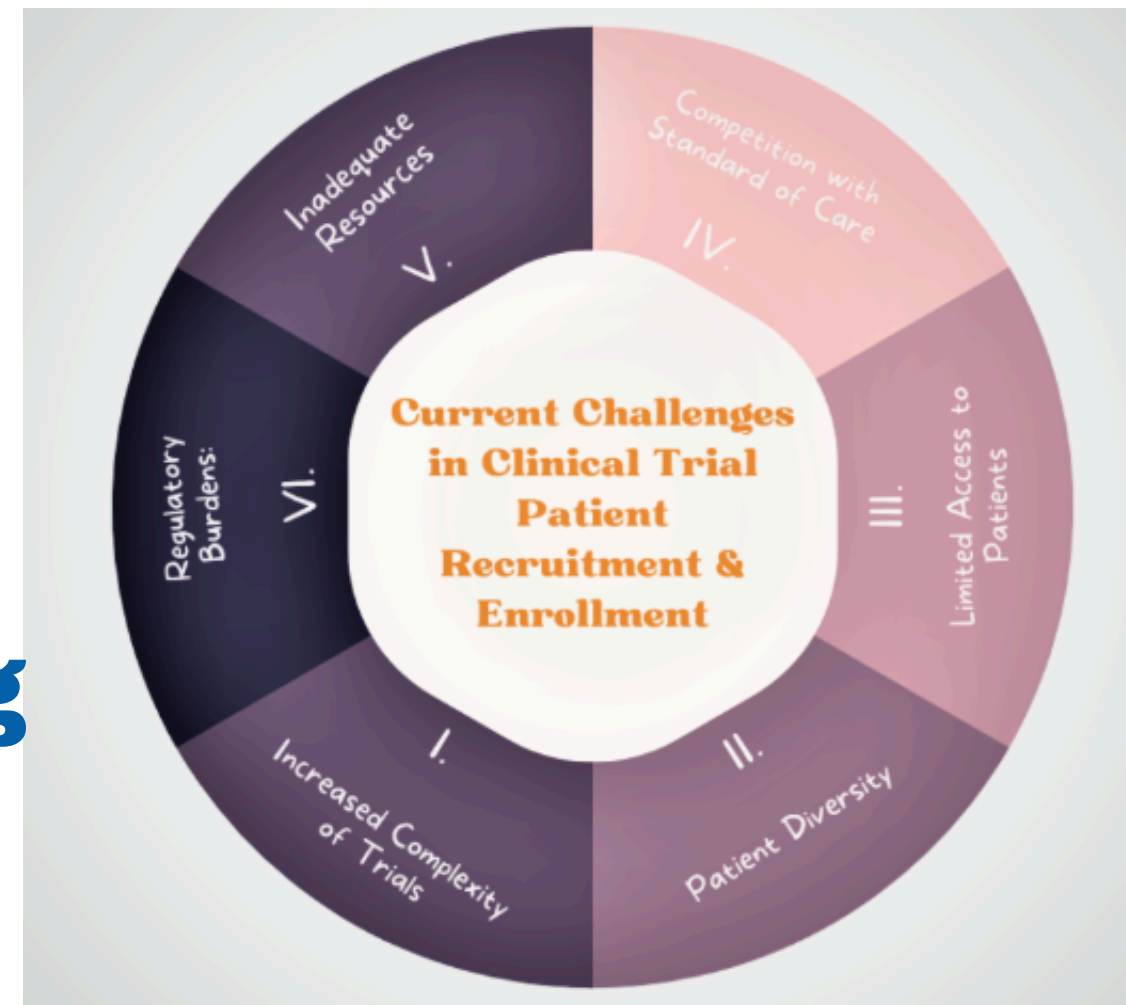
# NO DIRECTION





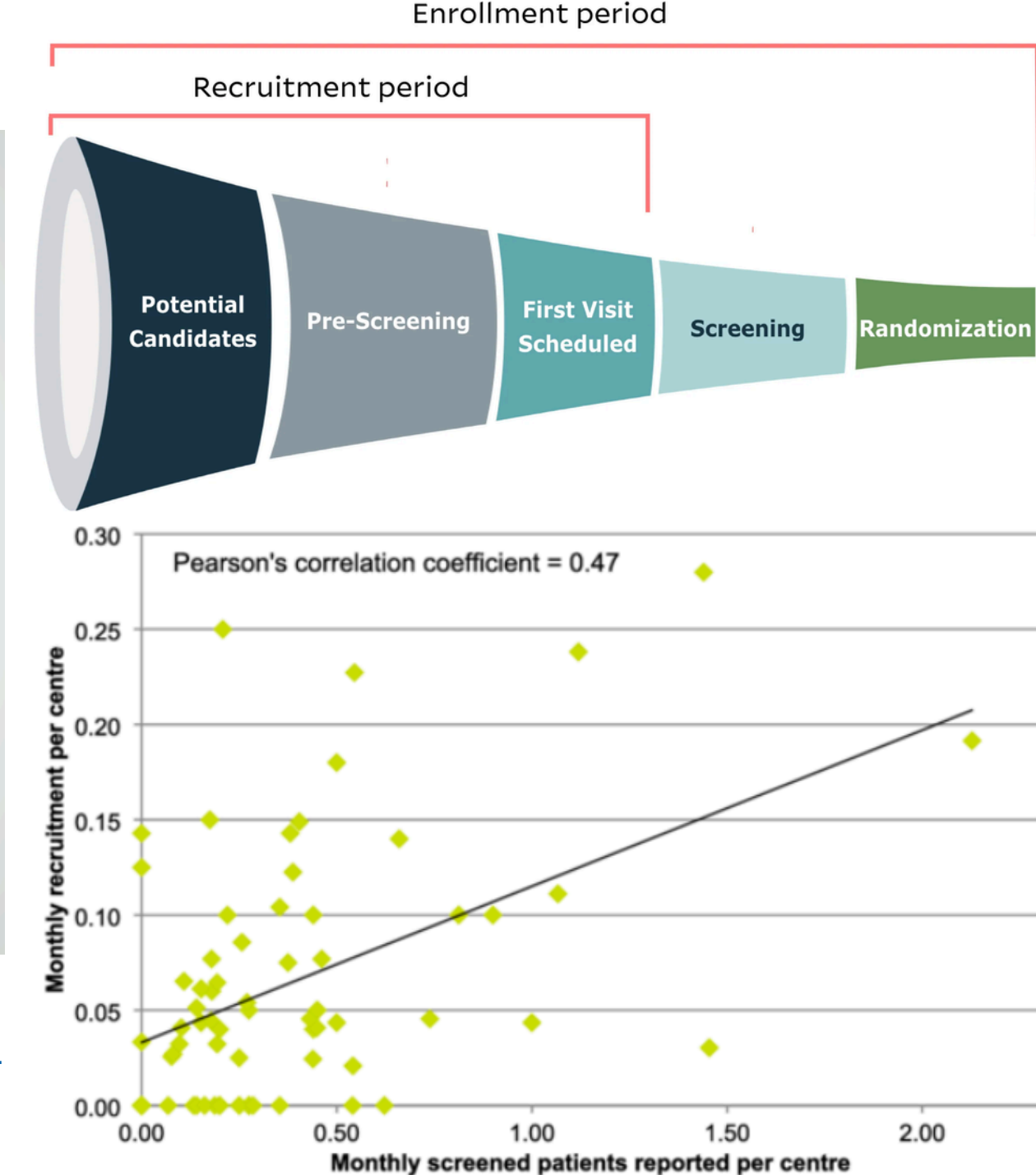
# Problem Statement – 4

## Utilizing data to predict recruitment rate (RR) in clinical trial for benchmarking



### Clinical Trial Delays and Mortality Statistics:

- 85% of clinical trials fail to recruit enough participants on time, causing delays in completion. [Source: CenterWatch, 2021.](#)
- 37% of trial sites fail to meet enrollment targets, and 11% fail to recruit a single patient. [Source: Biopharma Dive, 2020.](#)
- Recruitment rates vary by phase, with Phase I having lower rates due to smaller, specific populations. [Source: Applied Clinical Trials Online, 2019.](#)
- Recruitment rates differ geographically, with North America averaging 0.92 p/s/m and Europe 0.77 p/s/m. [Source: Tufts Center for the Study of Drug Development, 2020.](#)
- Rare disease trials have significantly lower RR, averaging 0.1–0.3 p/s/m due to limited eligible populations. [Source: Orphanet Journal of Rare Diseases, 2020.](#)
- Recruitment delays can cost \$600,000–\$8 million per day for large pharmaceutical trials. [Source: Pharmaceutical Technology, 2020.](#)



# Approach & methodology.

Overview	Methodology	Framework / tools used
<ul style="list-style-type: none"><li>• The project aims to develop an <b>AI model</b> that <b>predicts the recruitment rate of clinical trial participant</b> based on historical data. The model should <b>provide insights</b> into which factors (e.g., <b>conditions, age, sponsor etc.</b>) affect the recruitment process.</li><li>• <b>Clean and preprocess</b> structured (e.g., study design, age, sex) and unstructured (e.g., conditions, sponsor) data.</li><li>• <b>Use XGBoost</b> for structured data and <b>BERT</b> for unstructured text data.</li><li>• <b>Compare the outputs</b> of different models for improved prediction <b>accuracy</b>.</li><li>• Evaluate model performance using <b>RMSE, R<sup>2</sup>, and SMAPE</b>.</li><li>• Implement <b>SHAP and LIME</b> for <b>explainability</b> of model predictions.</li></ul>	<p><b>Data Collection &amp; Preprocessing</b></p> <ul style="list-style-type: none"><li>• To handle textual data effectively, we will vectorize the text fields using <b>BERT, TF-IDF, and TensorFlow's Text Vectorization Layer</b>.</li><li>• Identify and handle outliers using methods like <b>IQR (Interquartile Range)</b> or <b>Z-score</b>.</li><li>• Impute <b>missing data</b> using statistical methods or remove rows/columns with excessive missing values.</li></ul> <p><b>Model Developement:</b></p> <ul style="list-style-type: none"><li>• Implement <b>feature selection</b> through <b>PCA, LASSO regression</b>, and domain expert validation.</li><li>• Deploy <b>baseline models</b> (Linear Regression, Decision Trees) before advancing to complex models (Random Forest, XGBoost, Neural Networks)</li><li>• Apply <b>SHAP and LIME</b> for model interpretation and feature importance visualization</li></ul> <p><b>Accuracy Metrics</b></p> <ul style="list-style-type: none"><li>• <b>RMSE</b>: To measure model prediction error.</li><li>• <b>R<sup>2</sup></b>: To evaluate the model's fit and explanatory power.</li><li>• <b>SMAPE</b>: For percentage error to ensure robustness across varying data.</li></ul>	<p><b>scikit-learn:</b></p> <ul style="list-style-type: none"><li>• Purpose: <b>For data preprocessing</b> (scaling, encoding), feature selection, and model training (classification, regression).</li></ul> <p><b>TensorFlow/Keras:</b></p> <ul style="list-style-type: none"><li>• Purpose: To build and <b>train deep learning models</b>, including neural networks, for more complex patterns.</li></ul> <p><b>Pandas &amp; NumPy:</b></p> <ul style="list-style-type: none"><li>• Purpose: To <b>handle and manipulate data</b>, perform data cleaning, and numerical operations efficiently.</li></ul> <p><b>NLTK/SpaCy:</b></p> <ul style="list-style-type: none"><li>• Purpose: For text data preprocessing like <b>tokenization, lemmatization, and stopword removal</b>.</li></ul> <p><b>Matplotlib/Seaborn:</b></p> <ul style="list-style-type: none"><li>• Purpose: For <b>visualizing data distributions</b>, correlations, and model performance (like accuracy, ROC curves).</li></ul> <p><b>HuggingFace:</b></p> <ul style="list-style-type: none"><li>• Purpose: For using <b>BERT and other transformer models</b> to create <b>embeddings</b> from clinical trial text data, leveraging pre-trained medical domain models like <b>BioBERT and ClinicalBERT</b> for improved <b>text representation</b></li></ul>

# Model choice & setup

## Model Selection

### Baseline vs. Advanced Models:

- **Baseline Models:** Use simpler models like Linear Regression, Decision Trees, and Logistic Regression with **TF-IDF** to establish a performance baseline.
- **Advanced Models:** Implement more complex models like **Random Forest**, Gradient Boosting (XGBoost), and Neural Networks when using **BERT embeddings**.

### Performance Evaluation:

- **Compare** models using metrics like **RMSE**, **MAE**, and **R-squared** while ensuring proper cross-validation.

### Hyperparameter Tuning:

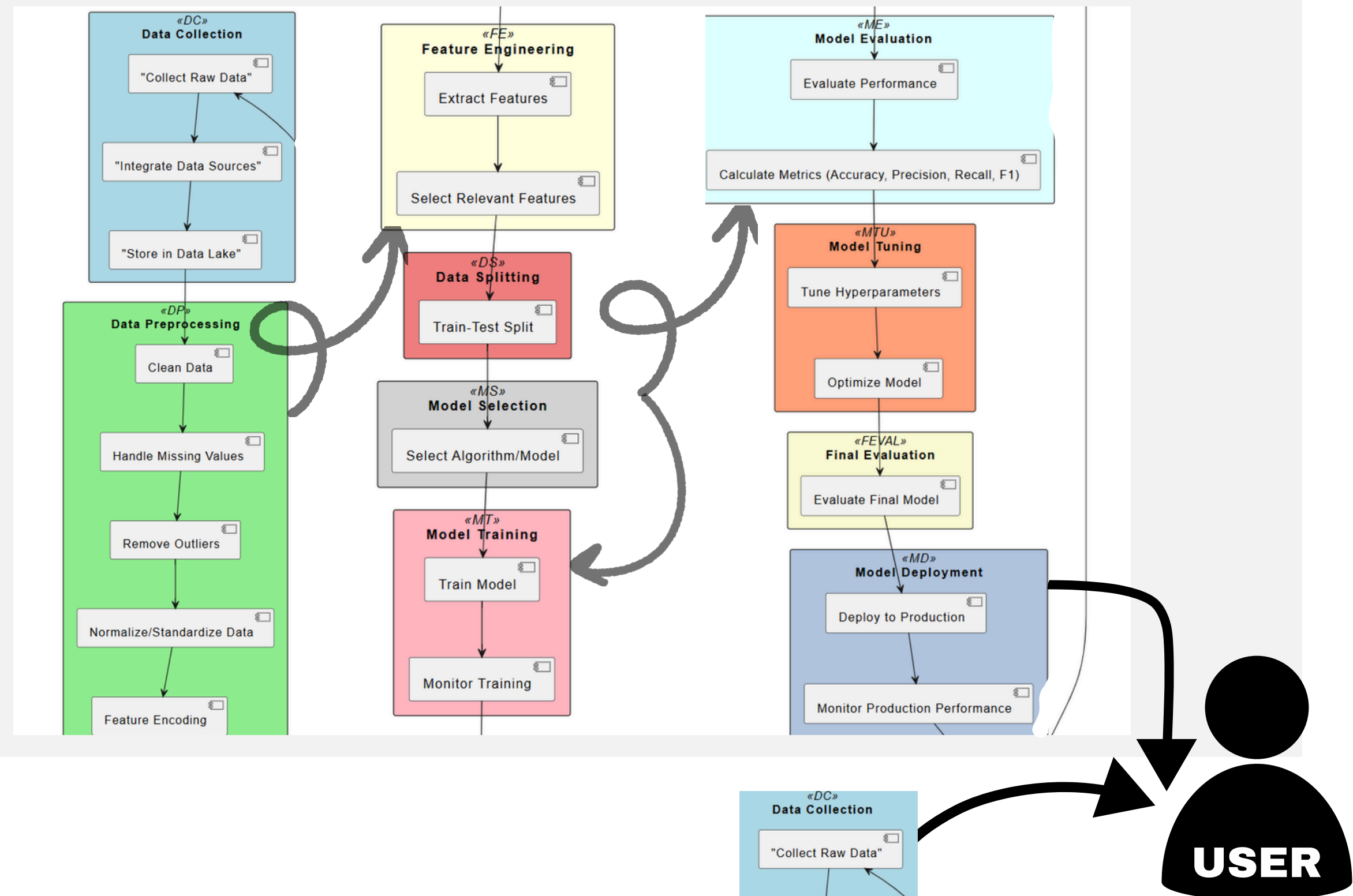
- Apply techniques like **Grid Search** and **Random Search** for optimal model configurations.

### Feature explainability:

- We will use techniques like **SHAP** (SHapley Additive Explanations) and Permutation Importance to **identify key predictors**.
- We will apply tools like **LIME** (Local Interpretable Model-agnostic Explanations) for visualizing **feature impact** on predictions.

## Model Architecture

### Technical flow of the end-to-end ML pipeline



# Model Training & Evaluation

## Evaluation Metrics

### Model Training Process:

The model training process will involve splitting the dataset into **training, validation, and test sets** to ensure **unbiased performance evaluation**. Both baseline models like **Logistic Regression with TF-IDF** and advanced models such as **Neural Networks with BERT** embeddings will be trained and compared. Hyperparameter tuning will be performed to optimize the model's performance.

### Evaluation Criteria and Metrics:

The evaluation will focus on measuring the **model's accuracy and reliability** using standard performance metrics. These metrics will help assess both **error magnitude and model fit** to ensure effective recruitment rate prediction.

### Discuss key performance metrics:

- **Root Mean Square Error (RMSE):** RMSE measures the standard deviation of the residuals (prediction errors) and gives more weight to larger errors due to squaring the differences.
- **Mean Absolute Error (MAE):** MAE measures the average magnitude of errors between predicted and actual values, making it easier to interpret.
- **R-squared (R<sup>2</sup>) Score:** R<sup>2</sup> measures how well the model explains the variance in the target variable, with values ranging from 0 to 1 (where 1 indicates a perfect fit)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (x_i - \hat{x}_i)^2}{N}}$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$R^2 = 1 - \frac{RSS}{TSS}$$



# Results and visualization

## Model Outcomes

### Prediction Accuracy:

- A **low RMSE** will indicate precise prediction of recruitment rates, enabling better trial planning and cost management.
- A **high  $R^2$  score** suggests the model effectively captures the variability in the recruitment rate due to the features considered.

### Key Findings:

- Influential Factors: **SHAP analysis** will highlight influential factors like **trial location and eligibility criteria**, with findings suggesting that stricter criteria may slow recruitment, while specific locations may consistently excel in recruitment speed.

### Performance Metrics:

- **RMSE,  $R^2$ , and SMAPE** will guide the model's reliability and utility.
- **Cross-validation** consistency will indicate generalizability across different trials.

### Implications for Stakeholders:

- **Pharmaceutical Companies:** Optimize trial timelines and **reduce delays** to save millions in revenue.
- **Clinical Sites:** Identify **recruitment bottlenecks** and address them proactively.
- **Regulatory Authorities:** Gain insights into **recruitment trends** for better oversight.

## Explainability

### Techniques for Interpretation:

#### SHAP (SHapley Additive Explanations):

- Provides a detailed breakdown of each feature's contribution to individual predictions.
- Highlights the most **influential factors** (e.g., trial location, eligibility criteria) for recruitment rate prediction.
- Summary plots and dependence plots will illustrate the **impact and interaction of features**.

#### LIME (Local Interpretable Model-agnostic Explanations):

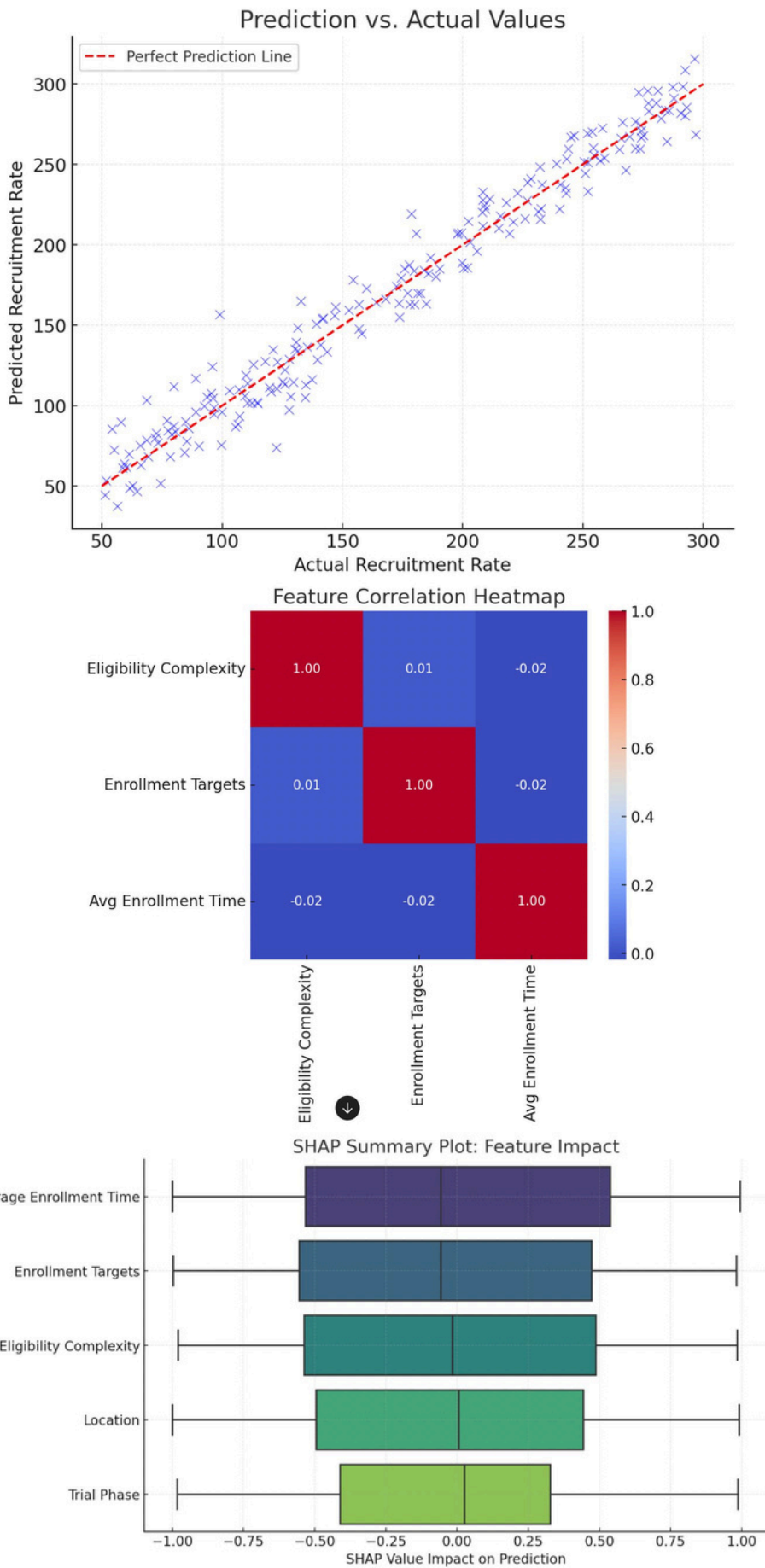
- Offers **localized explanations** for specific predictions.
- Useful for understanding how changes in input features affect individual outcomes.

### Feature Importance Analysis:

- Explains which **structured** (e.g., trial phase, location) and **unstructured** (e.g., eligibility criteria) features most impact the recruitment rate.
- Enables stakeholders to **identify actionable areas**, like optimizing trial locations or revising eligibility requirements.

### Transparency and Trust:

- Ensures stakeholders can understand why the model makes **specific predictions**.
- Facilitates **actionable decision-making** by connecting predictions to real-world factors.



# Challenges & Next Steps

## Limitations

### Data Quality and Availability:

- Missing or incomplete records, especially for key variables like eligibility criteria and outcomes, can impact model accuracy.
- **Limited data for rare conditions** or less common trial phases may reduce the model's robustness in these areas.

### Sampling Bias:

- Recruitment patterns vary across **geographies and trial types**, leading to potential **over-representation of well-funded or populous regions in the data**.
- Historical data may not fully represent diverse populations or **under-resourced areas**.

### Unbalanced Data:

- Datasets with a majority of trials **achieving high recruitment rates** may cause the model to underperform on trials with low or no enrollment, which are critical for prediction.

### Evolving Trends:

- External factors like **public health crises** (e.g., COVID-19), regulatory changes, or medical advancements are not fully captured in historical data, limiting the model's ability to **adapt to current dynamics**.

### Lack of Standardized Reporting:

- **Variability** in how clinical trial data is reported (e.g., different formats, terminologies, or languages) can **introduce inconsistencies** and reduce the reliability of predictions.

## Next Steps

- Include more **diverse datasets**, including rare conditions, global trial data, and failed trials to **reduce bias**.
- **Incorporate data** on public health events, policy changes, and demographic trends to capture evolving **recruitment dynamics**.
- Use **advanced imputation methods** and **NLP techniques** to clean and process **missing or unstructured data**.
- Use **transfer learning and advanced hyperparameter tuning** to increase **model generalization** and **prevent overfitting**.
- Develop **dynamic models** that adapt to new data and implement online learning for **real-time updates**.
- **Combine SHAP** with other **interpretability methods** and build user-friendly dashboards for **non-technical stakeholders**.
- Implement **bias detection and correction**, especially for **eligibility criteria and geographic representation**.
- **Test the model on external datasets** to ensure generalizability and conduct expert reviews for **real-world alignment**.
- Build tools to simulate recruitment scenarios and **integrate the model** into **trial management systems**.
- **Investigate patient behavior** and **causal relationships** to improve recruitment **prediction accuracy**.

# THANKYOU

## NO DIRECTION

**Bennett University**

SWAYAM PRASAD  
SAH

+91 99574 83591 || swayamps4567@gmail.com

SHREYA

+91 94519 54329 ||  
shreyyaaa369@gmail.com

ANANYA VERMA

+91 91180 37376 ||  
ananya.verma.may22@gmailcom

AMAN GUPTA

+91 91674 19990 ||  
10amangupta04@gmail.com

PRITESH PUNJ

+91 85951 68494 ||  
punjpritesh@gmail.com