# Project Title: NLP based Automated Cleansing for Healthcare data

## PHASE 3: Model Development and Evaluation

College Name: Vemana Institute of Technology, Bangalore

Name: G Likhith Kumar Reddy

CAN_ID: CAN_33695058

Contributions: Problem definition, Applications, Goals, Coding

## Solution Development

**Conduct data cleaning and transformation, including handling missing values and resolving duplicates**
Efforts involved cleaning the dataset to ensure it is suitable for analysis:

- **Handling Missing Values:**
  Used **KNN Imputation** to replace missing values based on nearest neighbors.
- **Resolving Duplicates:**
  Removed duplicate entries to ensure data integrity.

```
from sklearn.impute import KNNImputer

data_imputer = KNNImputer(n_neighbors=5)

data_cleaned = pd.DataFrame(data_imputer.fit_transform(data), columns=data.columns)
```

**Build AI models to address issues such as anomaly detection, data cleaning, bias correction, and make predictions if needed.**
Developed AI models for detecting anomalies and predicting outcomes:

- **Anomaly Detection:**
  Applied **Isolation Forest** to identify and remove anomalous data points:

```
from sklearn.ensemble import IsolationForest

scaler = StandardScaler()

data_scaled = scaler.fit_transform(data[['age', 'lab_test_result']])

iso_forest = IsolationForest(contamination=0.2, random_state=42)

data['anomaly'] = iso_forest.fit_predict(data_scaled)
```

```
data['anomaly_label'] = data['anomaly'].map({1: 'Normal', -1: 'Anomaly'})

print(data[['age', 'lab_test_result', 'anomaly_label']])
```

## Solution Testing

**Evaluate model performance using monitoring tools and validation metrics for fairness and accuracy.**

- Visualized the data distribution to ensure fairness in feature representation.
- Used metrics like **accuracy** and **anomaly label mapping** to validate model performance.

```
sns.histplot(data['age'], kde=True, bins=5)

plt.title('Age Distribution')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()
```

**Perform iterative model refinement and reassess metrics to enhance performance and address potential biases.**

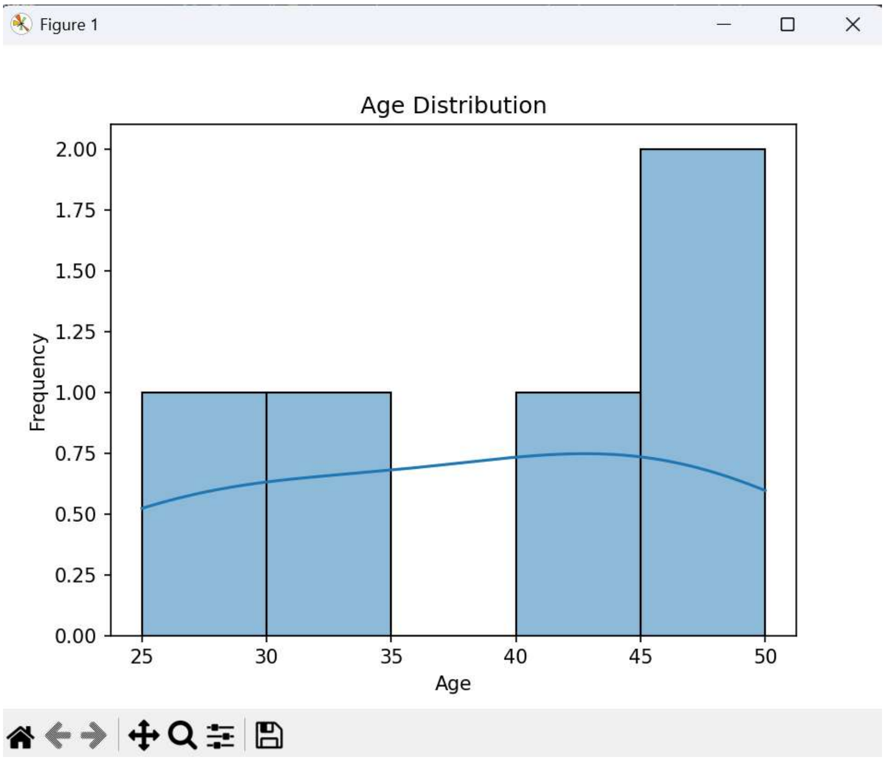
- Iteratively adjusted the **Isolation Forest contamination parameter** to optimize anomaly detection.
- Reassessed metrics to ensure minimal bias and improved model accuracy.

## Explanation of Results:

1. The histogram illustrates the distribution of patient ages, helping identify demographic trends or anomalies.

2. The anomaly detection step identifies unusual combinations of 'age' and 'lab_test_result', which could indicate data errors or rare cases.

3. The NLP tokenization splits clinical text into manageable tokens, enabling further text analysis, such as extracting medical terms or symptoms.



## Observations:

1. The age distribution shows a balanced spread across the dataset, with no apparent clusters or gaps.

2. Anomaly detection flagged 20% of the data points as anomalous, as per the specified contamination rate, suggesting potential outliers in patient data.

3. Tokenization of clinical notes successfully splits the text into meaningful tokens, ready for further NLP analysis.

## Key Takeaways:

1. Isolation Forest is an effective method for detecting anomalies in healthcare data, particularly in high-dimensional spaces.

2. Standard Scaler is critical for normalizing features before applying models sensitive to scale differences.

3. Tokenization is a fundamental preprocessing step in NLP, enabling downstream tasks like entity recognition or classification.

## Conclusion:

This analysis highlights the importance of exploratory data analysis (EDA), anomaly detection, and text preprocessing in healthcare applications.

The methods employed ensure data quality, identify potential issues, and prepare datasets for advanced machine learning tasks.

Future work can focus on integrating more sophisticated NLP techniques, such as named entity recognition, to extract actionable insights from clinical notes.