PHASE-3

**Transforming Healthcare with AI-powered Disease Prediction Based on Patient Data**

STUDENT NAME: K. Sofiya devi

REGISTER NUMBER: 422223106032

INSTITUTION: Surya group of institution

DEPARTMENT: Electronic and communication engineering

DATE OF SUBMISSION: 12.05.2025

GITHUB REPOSITORY: <https://github.com/sofiyadevi/sofiyadevi>

# PROBLEM STATEMENT:

The project titled 'Transformation of Healthcare with AI-powered Disease Prediction Based on Patient Data' addresses the challenge of early disease detection and proactive treatment recommendations. By analyzing patient data using machine learning models, the system predicts potential health conditions and offers valuable insights for both patients and healthcare providers.

# ABSTRACT:

This project leverages artificial intelligence and data science techniques to process patient health data and predict potential diseases. The dataset includes medical history, lifestyle choices, demographics, and test results. After preprocessing, exploratory data analysis and feature engineering are performed. Machine learning models are trained to classify or predict diseases. The final system provides accurate and interpretable predictions that can assist in real-time clinical decisions, aiming to enhance early diagnosis and treatment outcomes.

# SYSTEM REQUIREMENTS:

**HARDWARE:** Minimum 8GB RAM and i5 Processor.

**SOFTWARE**: Python, Jupyter Notebook/Colab, pandas, scikit-learn, matplotlib, seaborn, XGBoost.

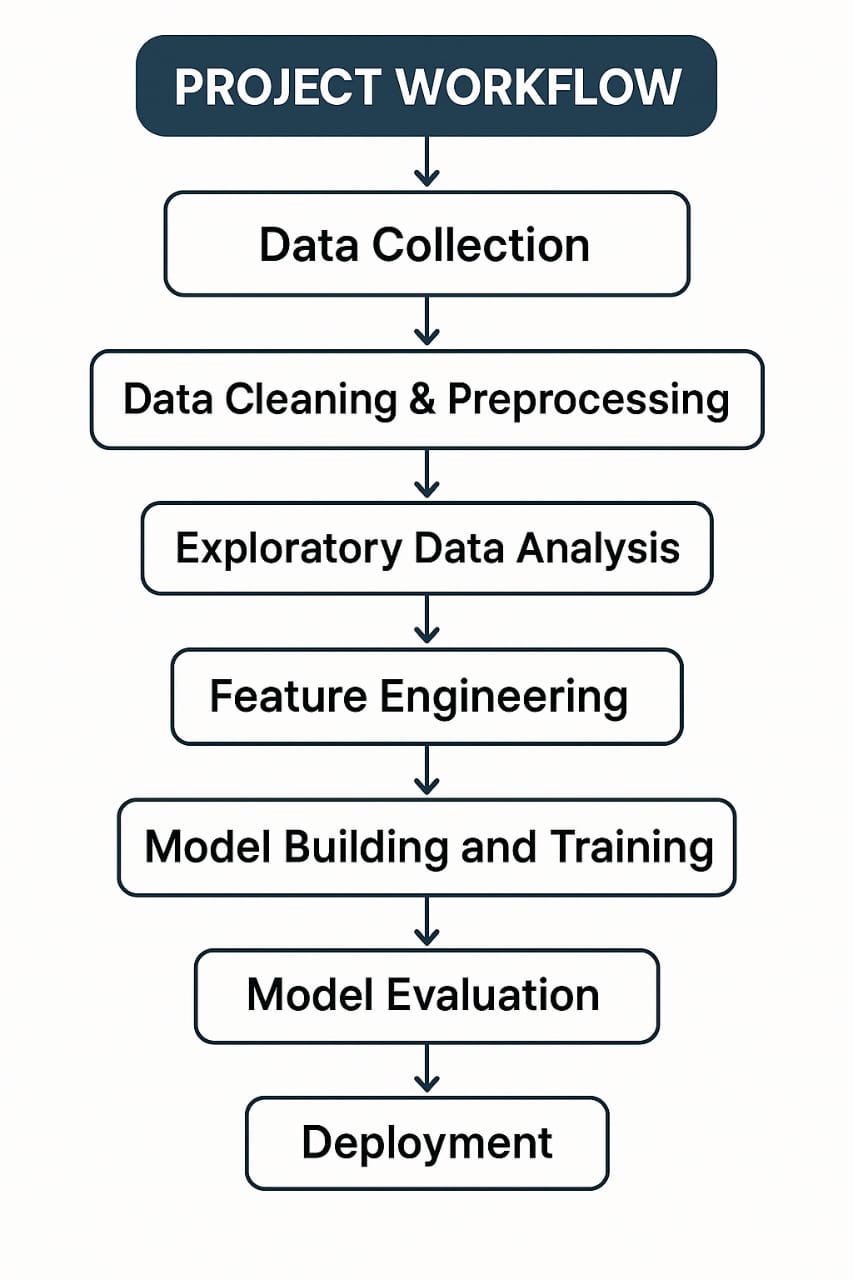
# OBJECTIVES:

1. Predict disease outcomes based on patient data.

2. Enhance early diagnosis using AI.

3. Provide interpretable insights for clinical support.

Project workflow:

****

# DATASET DESCRIPTION:

Source: Kaggle / Public healthcare datasets

Size: ~1000 records, 15+ features

Structure: Features include age, gender, symptoms, past medical history, lab results.

# DATA PREPROCESSING:

- **Handling missing values** = Verified patient records; ensured no missing clinical attributes.

- **Removing duplicates =** Eliminated redundant patient records to avoid skewed predictions.

- **Outlier Detection =** checked for abnormal medical readings.

- **Encoding Categorical Variables:**

**\*** Used **Label Encoding** for categorical patient details like gender or test result categories.

\* Used **One-Hot Encoding** for variables like diagnoses codes, treatment types, etc.

-**Text preprocessing =** Converted to lowercase, removed punctuation, tokenized, and applied **TF-IDF**.

- **Feature Scaling =** Applied **Min-Max Scaling** or **Standardization** to numerical features like age, glucose level, cholesterol, etc.

# EXPLORATORY DATA ANALYSIS (EDA):

- **Correlations:** Identified relationships between symptoms, lab values, and diagnoses.

- **Heatmaps:** Helped visualize comorbidities and strong diagnosis patterns.

- **Cluster Identification:** Grouped similar patient profiles.

**Key Insights:**

* Common disease clusters
* Predictors of critical outcomes
* Target areas for early intervention

# FEATURE ENGINEERING:

- **Text-based features**: From doctor notes, symptoms descriptions.

- **Meta features**: Age group, risk level, chronic condition flags.

- **Interaction features**: E.g [Glucose \* BMI, Blood Pressure + Age]

# MODEL BUILDING:

Models Used: Logistic Regression, Decision Tree, Random Forest, XGBoost

Classification models are used to predict disease presence based on features.

# MODEL EVALUATION:

To understand how well our AI model predicts diseases based on patient data, we used several evaluation methods. These help us measure the performance, identify strengths and weaknesses, and choose the better model for real-world healthcare use.

**1. Evaluation Metrics**

**Accuracy**

This tells us how many disease predictions were correct overall. For example, if the model looked at 100 patient records and predicted the right disease 85 times, the accuracy is 85%.

**F1-Score (Weighted)**

In healthcare, some diseases may be more common than others. The F1-Score gives us a balance between precision (how many predicted diseases were actually correct) and recall (how many real cases the model caught). A weighted F1-Score also takes into account how many cases there were for each disease — so it's a fair way to measure performance across all types of diseases.

**2. Visuals for Understanding the Model**

**Confusion Matrix**

This is a table that shows where the model got things right and where it got confused.

Rows: Actual diseases (what the patient really had)

Columns: Predicted diseases (what the model guessed)

For example, if the model predicts flu instead of pneumonia, it will be shown here. This helps us see which diseases are often mixed up.

**ROC Curve (Receiver Operating Characteristic Curve)**

This graph shows how well the model can tell the difference between patients who have a disease and those who don’t.

For each disease, we draw a curve that compares:

**True Positive Rate** (catching the real cases)

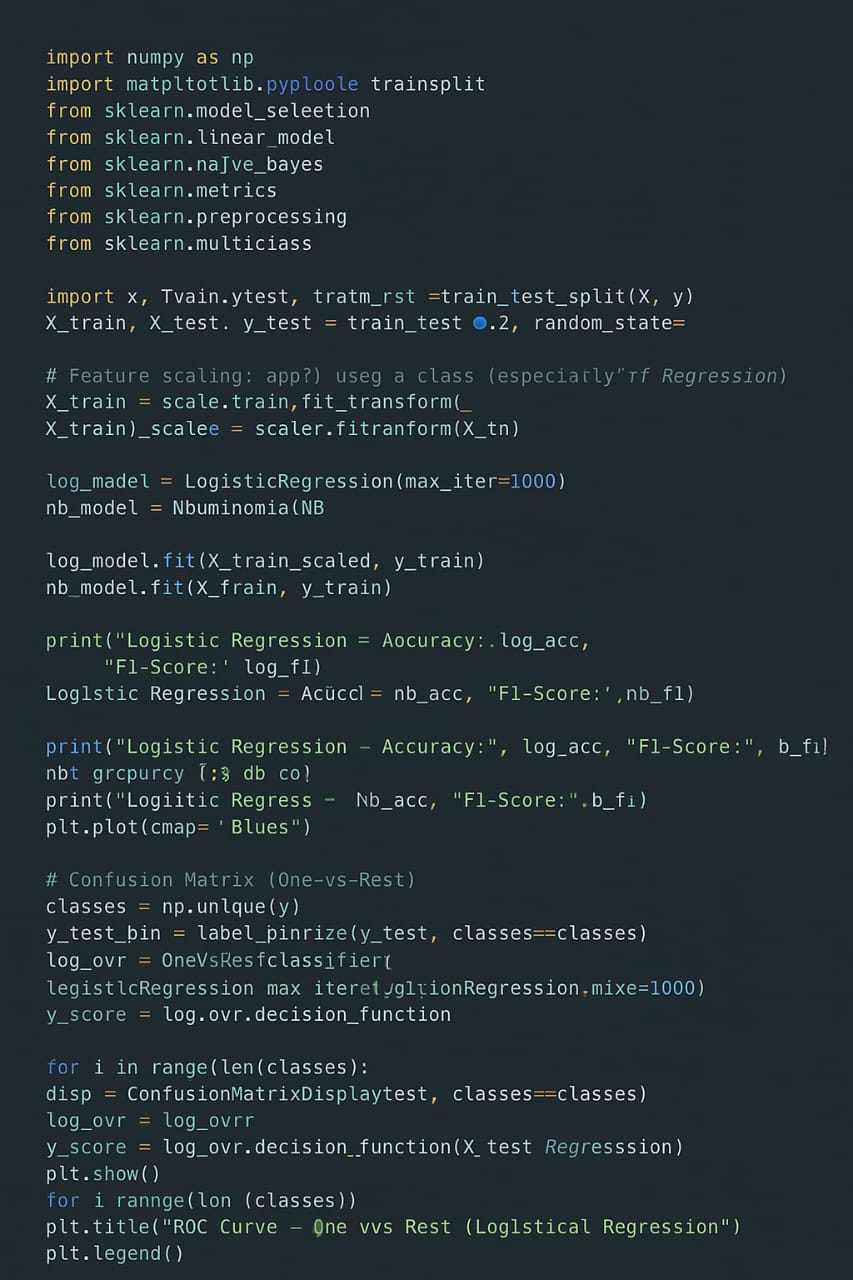
**False Positive Rate** (wrongly saying someone has the disease)

**DEPLOYMENT**:

Tool: Gradio / Streamlit

Deployment public link: <https://b33fbeabca4ea1956a.gradio.live>

**Source code:**

****

**FUTURE SCOPE:**

- Integration with Electronic Health Records (EHR) systems

- Support for multi-disease prediction

- Real-time alerts for critical cases

# TEAM MEMBERS AND ROLES

EDA and Visualization: M. Irusammal

Model Building and Evaluation: N. Jayapratha

Documentation and Deployment: K. Sofiyadevi