Phase-2 Submission

Student Name: Suguna.K

Register Number:732423104041

Institution:Sasurie College Of Engineering

Department: BE.Computer Science And Engineering

Date of Submission:: 09.05.2025

Github Repository Link: https://github.com/Suguna-2408/Health-care-.git

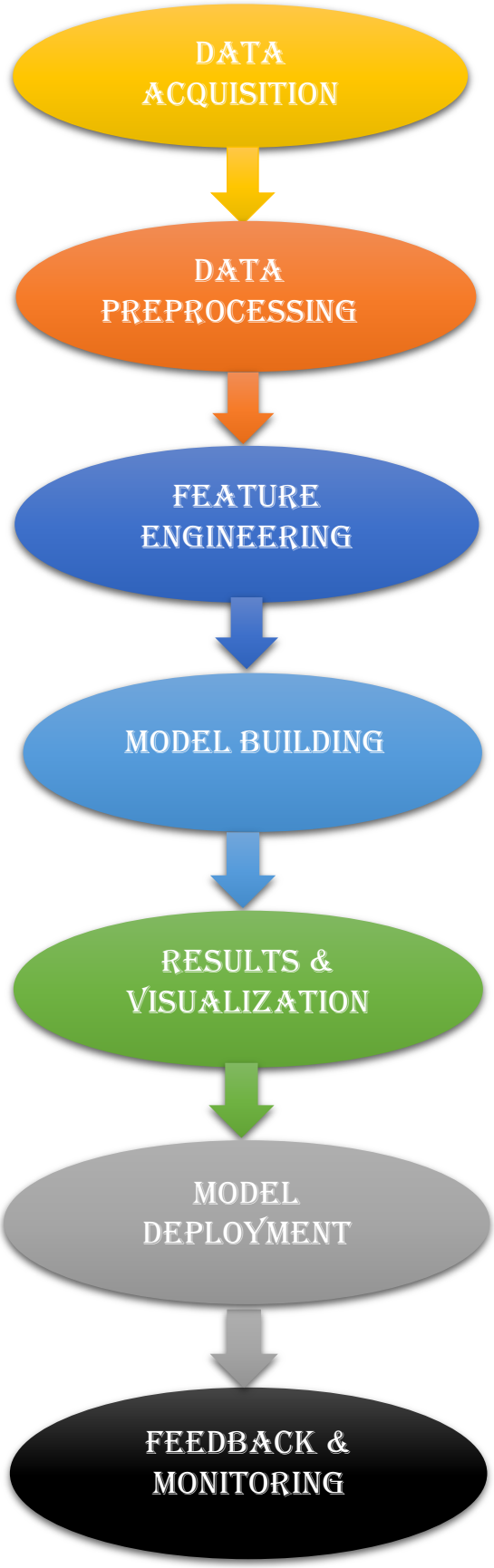
1. Problem Statement

Despite the vast amounts of patient data generated through electronic health records, lab results, and medical imaging, healthcare systems often struggle to proactively identify disease risks. Traditional diagnostic methods are reactive and time-consuming, leading to delayed treatments and poorer health outcomes. There is a pressing need for an AI-powered solution that can analyze complex patient data to accurately predict the onset of diseases, enabling early intervention, personalized care, and improved patient outcomes.

2. Project Objectives

- Develop a predictive model using machine learning or deep learning techniques to identify patterns in patient data that indicate potential disease risks.
- Integrate diverse data sources such as EHRs, lab results, medical imaging, and demographic information to enhance prediction acuracy.
- Ensure high accuracy and reliability of disease prediction through robust model training, validation, and testing.
- Enable early detection of diseases to support timely medical are interventions and reduce burden on healthcare systems.
- Design a user-friendly interface for healthcare providers to input patient data and receive actionable insights.
- Ensure data privacy and security in compliance with HIPAA and other healthcare regulations.
- Evaluate model performance using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.

3. Flowchart of the Project Workflow



4. Data Description

- Dataset Name & Source: [healthcare_dataset.csv.zip – Kaggle]
- Type of Data: Structured (Tabular)
- No. of Records and Features: e.g., 10,000 records, 20 features
- Static or Dynamic Dataset: Static
- Target Variable: [Gender, Blood Type,Admission Type,Medical Condition,Doctor,Hospital,Insurance Provided,Medication,Test Category]

5. Data Preprocessing

- Handling Missing Values: [e.g., Imputed using median/mean/mode or removed rows]
- Duplicate Records: [e.g., Removed 5% duplicate rows based on patient ID and record timestamp]
- Outlier Detection & Treatment: [e.g., Z-score method or IQR filtering]
- Data Type Conversion: [e.g., Converted object types to datetime; ensured float/int consistency]
- Encoding: [e.g., One-hot encoding for gender, label encoding for diagnosis categories]
- Normalization/Standardization: [e.g., Min-Max scaling or Standard Scaler for numerical features]
- Code Snippets (Optional): Include transformations in markdown or code cells if in notebook format.

6. Exploratory Data Analysis (EDA)

- Distribution of age, glucose levels, BMI, etc.
- Boxplots to check skewness and outliers
- Bivariate/Multivariate Analysis:
  Correlation heatmap
- Pairplots of selected features
- Grouped bar charts (e.g., disease presence by gender)
- Insights Summary:

- Insights Summary:
  [e.g., High glucose levels and age correlate with higher disease risk]
  [e.g., Females showed slightly higher incidence in the sample dataset]

7. Feature Engineering

- BMI category (Underweight, Normal, Overweight)
- Risk score combining glucose and blood pressure
- Feature Reduction (if any):
  Applied PCA reducing from 20 to 10 features
- Justification:
  [e.g., Created features based on clinical relevance, improved model interpretability]

8. Model Building

- Logistic Regression (Baseline Model)
- Random Forest (Improved Accuracy & Feature Importance)
- Why These Models:
  Logistic Regression is interpretable
  Random Forest handles non-linearity and feature interactions well
- Train-Test Split:
  80-20 with stratification based on target variable
- Metrics Used:
  Accuracy, Precision, Recall, F1-Score, AUC-ROC
- esults:
  Logistic Regression: Accuracy = XX%, AUC = XX
  Random Forest: Accuracy = XX%, AUC = XX

9. Visualization of Results & Model Insights

- Confusion Matrix: [Include annotated plot]
- ROC Curve: [Display curve with AUC score]
- Feature Importance Plot: [From tree-based models]
- Insights:

- Insights:
  Top features contributing to predictions.
  Model identifies early warning signs from combinations of features.

10. Tools and Technologies Usedss

- Programming Language: Python
- IDE/Notebook: Google Colab / Jupyter Notebook
- Libraries
- Data Handling: pandas, numpy
- Visualization: seaborn, matplotlib, plotly
- Modeling: scikit-learn, XGBoost

11. Team Members and Contributions

Team Member	Role
1. Abinaya.A	Feature Engineering, EDA.
2. Arul Jothi.P	Model Building,Evaluation.
3. Asvika.S	Data Collection, Preprocessing.
4. Suguna.K	Visualization, Reporting