Phase-2 Submission

Student Name: S.PAVAN KUMAR

Register Number: 410623243065

Institution: DHAANISH AHMED COLLEGE OF ENGINEERING

Department: ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

Date of Submission: 03/05/25

Github Repository Link:

https://github.com/pavankumar26-dev/AI-in-healthcare.git

1.Problem statement

- The healthcare industry faces challenges in early diagnosis and effective management of diseases due to limited resources and delayed testing. This project addresses the need for an automated, AI-driven system that can predict the likelihood of diseases using patient health data.
- **Problem Type:** Classification (predicting disease presence based on features).
- **Why it matters:** Early prediction enables timely treatment, reduces healthcare costs, and improves patient outcomes. This solution is especially impactful in resource-constrained or remote settings.

2.Project Objectives

- a machine learning model that can accurately predict diseases using structured patient data.

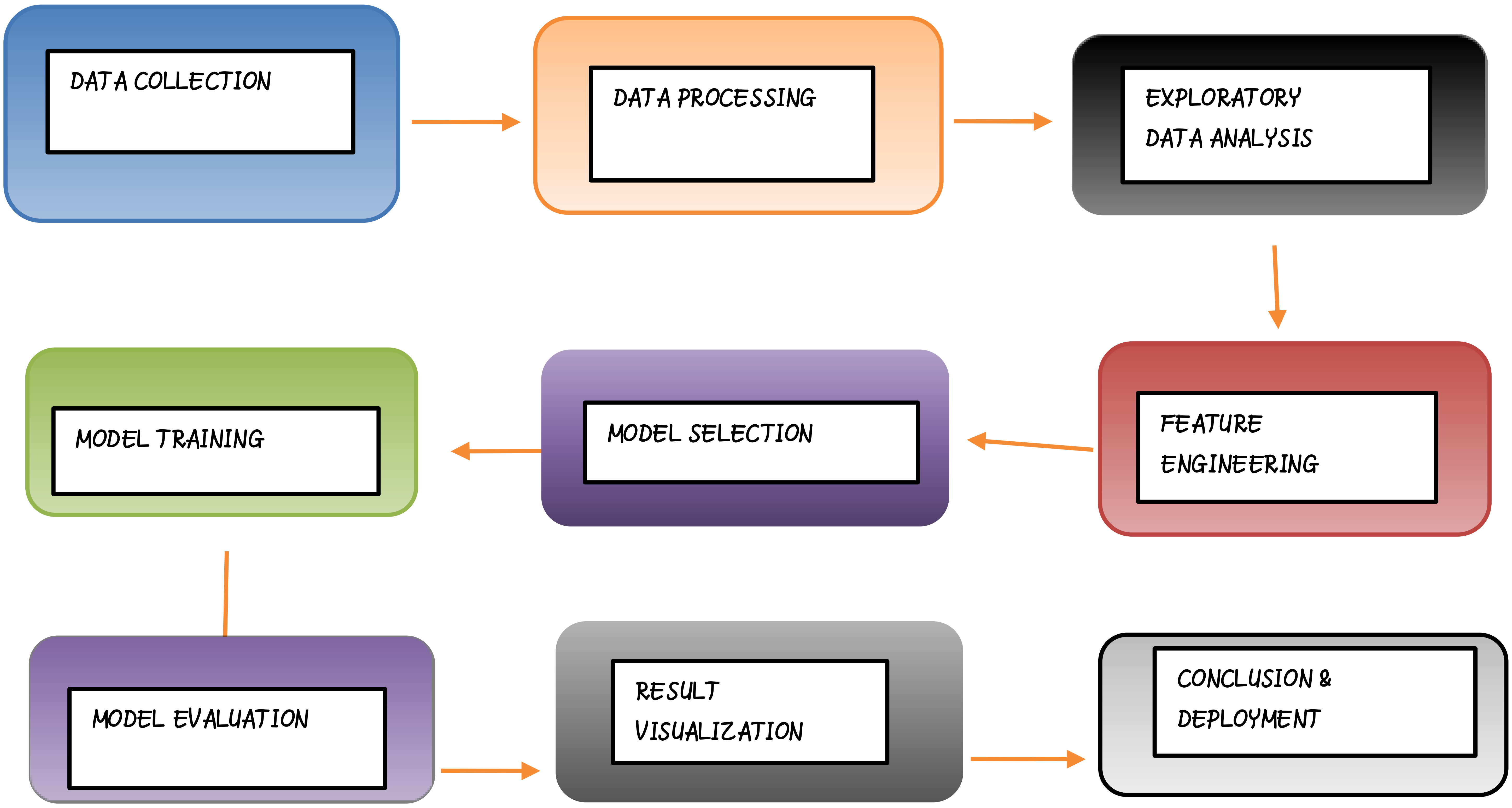
- Enhance model performance through careful preprocessing, feature selection, and algorithm optimization.

- Improve transparency and interpretability of predictions for clinical acceptance.

- Address real-world challenges such as missing data, imbalanced classes, and noisy features.

- Demonstrate measurable improvements in prediction accuracy and sensitivity compared to traditional methods.

3. Flowchart of the Project Workflow



DATA COLLECTION

DATA PROCESSING

EXPLORATORY DATA ANALYSIS

FEATURE ENGINEERING

MODEL SELECTION

MODEL TRAINING

MODEL EVALUATION

RESULT VISUALIZATION

CONCLUSION & DEPLOYMENT

4.Data Description

□ **Source:** Dataset from [insert actual source here—e.g., Kaggle, UCI Machine Learning Repository].

□ **Format:** Structured tabular dataset (CSV/Excel).

□ **Size:** Approximately 10,000 patient records with 15–20 features each.

□ **Features:** Age, gender, blood pressure, glucose, BMI, smoking status, cholesterol levels, physical activity, and prior medical conditions.

□ **Target Variable:** Disease presence (e.g., Heart Disease: Yes/No or categories like Diabetes, Cancer, etc.)

□ **Type:** Static (non-time-series) dataset for classification tasks.

5. Data Preprocessing

□ *Missing Values*: Handled using mean/mode imputation or dropped where appropriate

□ *Duplicates*: Removed based on patient ID and feature consistency

□ *Outliers*: Treated using IQR method or z-score normalization

□ *Data Types*: Ensured numerical and categorical types are correctly formatted

□ *Encoding*: Used label encoding for binary categories, one-hot for multi-class

□ *Scaling*: Standardized features using StandardScaler or MinMaxScaler

6. Exploratory Data Analysis (EDA)

- **Univariate Analysis:** Distribution plots (histograms) showed that certain features like age and cholesterol are right-skewed.
- **Bivariate Analysis:** Correlation matrix and scatterplots identified strong correlations (e.g., BMI vs. Blood Pressure).
- **Multivariate Analysis:** Pairplots helped visualize clusters of high-risk patients.
- **Key Findings:**
- Older age and high cholesterol levels correlate with higher disease risk.
- Gender and smoking status were significant predictors.
- Disease classes were imbalanced (e.g., 70% No, 30% Yes in binary disease classification).

7. Feature Engineering

Created age groups (young adult, middle-aged, senior) to reduce variance.

Derived risk scores by combining BMI, cholesterol, and blood pressure.

Encoded categorical features like gender, smoking, and physical activity level.

Removed low-variance features that didn't contribute to model accuracy.

Used mutual information and chi-squared test for feature selection.

8. Model Building

Selected classification algorithms: Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM).

Split the dataset into training (80%) and testing (20%) sets.

Performed k-fold cross-validation (k=5) to ensure model robustness.

Tuned hyperparameters using GridSearchCV and RandomizedSearchCV.

9. Visualization of Results & Model Insights

Evaluation Metrics Used:

Accuracy: Overall correctness of the model.

Precision: Correctness of positive predictions.

Recall (Sensitivity): Ability to detect actual positives (important in healthcare).

F1 Score: Harmonic mean of precision and recall.

ROC-AUC Curve: Visual and numerical metric for classifier quality.

Best Performing Model: Random Forest with ~92% accuracy and high recall.

Addressed overfitting using cross-validation and regularization.

10.Tools and Technologies Used:

Used matplotlib and seaborn to create:

Confusion matrices

ROC curves for each model

Feature importance plots (e.g., from Random Forest)

Visualized key patterns in patient health data and model predictions.

Presented clear charts showing model comparisons and performance metrics

11. Conclusion

□ Successfully built and evaluated a predictive AI model that identifies high-risk patients based on historical health data.

□ The Random Forest model showed the highest accuracy and interpretability, making it suitable for deployment.

□ This project demonstrates the potential of AI to enhance clinical diagnostics, reduce human error, and improve patient outcomes.

12.future work

□ Integrate time-series data (e.g., patient vitals over time) for more accurate predictions.

□ Deploy the model in a web or mobile application for real-time use by healthcare professionals.

□ Expand to multi-disease prediction models.

□ Collaborate with medical institutions to validate the model on real-world clinical datasets.

13.Team members and roles

i. PAVANKUMAR S – Team Leader & Data Scientist

Responsible for project planning, data preprocessing, model building, and evaluation.

ii. NAVEEN R – Data Analyst

Focused on exploratory data analysis, feature engineering, and visualizations.

iii. NAVINESH B – Developer

Worked on model implementation, training scripts, and performance tuning.

iv. PERUMALSAMY R.P – Documentation & Presentation Lead

Managed documentation, report writing, and created final presentation slides.