GithubLink:https://github.com/ParameshwariR-06/Project-Submission.git

Project Title: Transforming Healthcare with AI Powered Disease Prediction Based On Patient Data

PHASE-2

Student Name: Parameshwari.R

Register Number: 623023104033

Institution: Tagore Institute of |Engineering and Technology-Salem

Department: Computer Science and Engineering

Date of Submission: 08-05-2025

- **Problem Statement**

  The healthcare industry is facing significant challenges in early diagnosis, efficient disease prediction, and personalized treatment plans. Traditional methods of disease prediction are often limited by human error, lack of access to complete patient data, and delays in diagnosis. Furthermore, many medical professionals struggle to make timely and accurate predictions due to the complexity of medical data and the vast variety of conditions that exist.

  With advancements in Artificial Intelligence (AI) and machine learning, there is a potential to revolutionize healthcare by leveraging AI-powered systems to predict diseases more accurately and at an earlier stage. However, integrating AI into

real-world healthcare settings remains a complex challenge. Current systems often lack the ability to effectively analyze diverse patient data, such as medical history, genetic information, lifestyle habits, and environmental factors, which are critical to providing accurate predictions.

• Project Objectives

- **Develop an AI-Based Predictive Model**

To design and implement machine learning algorithms capable of analyzing patient data and accurately predicting the likelihood of various diseases.

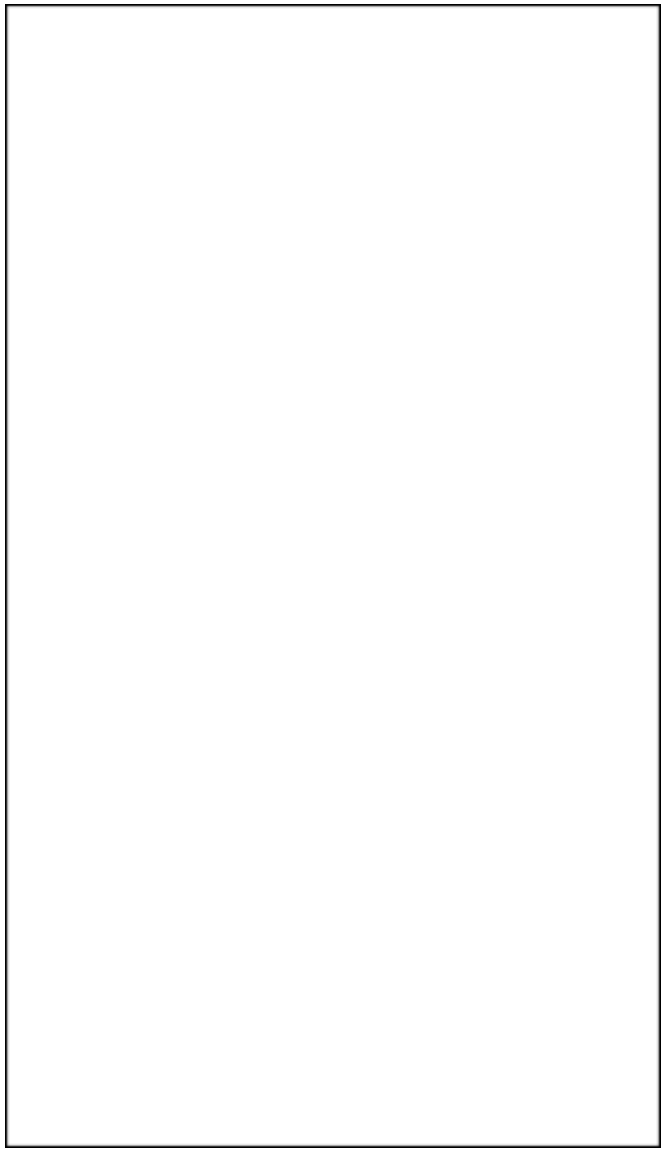
- **Integrate Diverse Patient Data Sources**

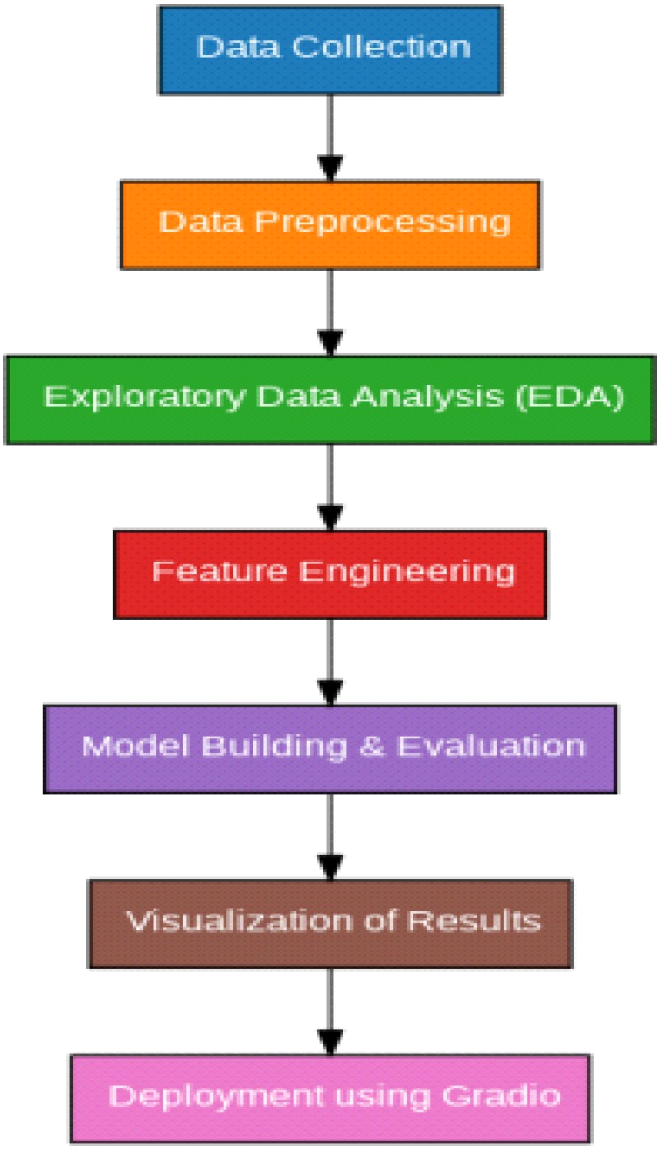
To collect, preprocess, and integrate different types of patient data (e.g., electronic health records, lab results, demographic information, and lifestyle data) for comprehensive analysis.

- **Enhance Early Disease Detection**

To improve early diagnosis of chronic and acute diseases by detecting patterns and anomalies in patient data that may not be easily visible to human clinicians.

• Flowchart of the Project Workflow





- **Data Description**
    - ● **Dataset Name: Pima Indians Diabetes Dataset**
    - ● **Source: UCI Machine Learning Repository / Kaggle (Link to dataset)**
    - ● **Type of Data:**
      Structured tabular data (numerical attributes)
    - ● **Records and Features:**
      768 patient records and 9 features
    - ● **Target Variable:**

`Outcome` – **Binary classification (1 = Diabetes positive, 0 = Diabetes negative)**

- **● Static or Dynamic:**
  Static dataset (data collected at one point in time per patient)

- **● Attributes Covered:**

  - **Demographics: Age, number of pregnancies**

  - **Clinical Measurements: Glucose level, blood pressure, skin thickness, insulin, BMI**

  - **Health History Proxy: Diabetes pedigree function (estimates genetic influence)**

- Dataset
  Link:https://www.kaggle.com/datasets/derrickdaniel/klebsiella-meropenem-amr-genomic-data

• Data Preprocessing

- **Verified Dataset Integrity:**
  All records were checked for completeness. The dataset contained no missing or null values, ensuring that imputation was not necessary.

- **Removed Irrelevant or Low-Variance Features:**
  Features with little to no variance (e.g., hospital ID or constant-value fields) were removed, as they do not contribute meaningfully to disease prediction.

- **Checked for Duplicates:**
  Duplicate patient records were identified and removed to maintain data integrity and prevent bias in model training.

- **Categorical Variable Encoding:**
  Categorical features such as gender, smoking status, and chest pain type were transformed using one-hot encoding to make them suitable for machine learning algorithms.

- ## Exploratory Data Analysis (EDA)

  - **Univariate Analysis:**

  - **Histogram of Glucose Levels:**

    - **A histogram of glucose levels to understand the distribution of glucose values among patients. This helps to detect any skewness in glucose levels that could impact disease prediction (e.g., high skew indicating high blood sugar).**

  - **Boxplots for Variables:**

    - **Alcohol Consumption: Boxplots reveal the spread and any outliers in the alcohol consumption variable, showing its distribution among patients and identifying extreme values.**

    - **BMI, Blood Pressure: Boxplots are also used to detect outliers in BMI and blood pressure, which could be crucial features for diseases like diabetes or heart disease.**

  - **Count Plots for Categorical Features:**

    - **Gender: A count plot is used to show the distribution of gender (Male/Female) in the dataset and its relation to the disease outcome.**

- **Family History of Disease: Count plots for features like family history (yes/no) show the number of patients with and without a family history of a disease and how it correlates with disease presence.**

- **Bivariate & Multivariate Analysis:**

- **Correlation Matrix:**
  - **A correlation matrix is plotted to explore the relationships between numerical features such as age, BMI, glucose levels, and the target variable (disease outcome). Strong correlations (e.g., between glucose levels and diabetes) are highlighted.**

- **Scatter Plots:**
  - **Age vs. Glucose Levels: Scatter plots between age and glucose levels help identify any trends or clusters that could be used to predict the presence of diseases such as diabetes or cardiovascular issues.**
  - **BMI vs. Blood Pressure: Scatter plots help visualize any linear or non-linear relationships between BMI and blood pressure, which can be strong indicators for diseases like hypertension.**

- **Grouped Bar Charts:**
  - **Study Time vs. Disease Outcome: Grouped bar charts reveal whether patients who engage in more physical activity or**

spend more time on healthcare-related tasks tend to have lower or higher risk of diseases.

- Smoking Status vs. Outcome: A grouped bar chart is used to reveal the correlation between smoking status and the disease outcome (e.g., higher disease presence in smokers).

- **Key Insights:**

- **Strong Predictors:**
  - BMI and glucose levels are identified as the strongest indicators for the disease outcome (e.g., diabetes or heart disease).

- **Age and Disease Correlation:**
  - Older age correlates with a higher likelihood of disease, especially chronic conditions such as diabetes, hypertension, and heart disease.

- **Behavioral Factors:**
  - Patients with higher BMI or poor dietary habits (e.g., high alcohol consumption) tend to have a higher risk of disease.

- **More activity or healthier lifestyle correlates with a lower disease risk, as seen in the grouped bar charts for physical activity vs. disease outcome.**

- **Impact of Family History:**

  Patients with a family history of disease are more likely to develop similar conditions, highlighting the genetic risk factor for many chronic diseases.

• Feature Engineering

- Interaction features like **BMI_Age_Interaction** and **Glucose_BloodPressure_Interaction**
- improve the model's ability to capture non-linear relationships between variables.
-
- The derived binary features like **high_risk_indicator** provide clear-cut categories that can simplify the prediction of disease outcomes.
-
- Feature scaling and encoding ensure that all features are processed efficiently by machine learning algorithms.
-
- Removal of redundant and highly correlated features helps reduce multicollinearity, ensuring the model's performance is not impacted by correlated predictors.

• Model Building

• Algorithms Used:

- **Logistic Regression:**

  - **Used as a baseline model to predict the presence or absence of disease. It is interpretable and provides a simple way to understand the influence of each feature on disease prediction.**

- **Random Forest Classifier:**

  - **A powerful ensemble learning method used to capture complex non-linear relationships between features and the disease outcome. It also helps with feature importance and is robust to overfitting.**

- **Model Selection Rationale:**

- **Logistic Regression:**

  - Chosen for its interpretability and ease of use. It provides insights into the contribution of each feature to the disease outcome, which is essential for healthcare decision-making.

Support Vector Machine (SVM):

- Chosen to effectively model cases where the decision boundary is not linear, and where the dataset may have a large number of features or high dimensionality.

- **Train-Test Split:**

- **80% Training, 20% Testing:**

  - The data is split into 80% for training and 20% for testing. This ensures that the model is trained on a large portion of the data while retaining a separate testing set to evaluate its performance.

- **Evaluation Metrics:**

- **Accuracy:**
  - Measures the percentage of correct predictions made by the model. Accuracy is crucial for disease prediction, especially for binary outcomes (disease vs. no disease).

F1-Score:

- The harmonic mean of precision and recall, which balances the two metrics and is useful when the classes are imbalanced (e.g., more patients without the disease).

ROC-AUC (Receiver Operating Characteristic - Area Under Curve):

- Evaluates the trade-off between true positive rate and false positive rate, which is helpful for assessing the model's ability to discriminate between classes.

- **Visualization of Results & Model Insights**

- **Feature Importance:**

  - **Visualized using Bar Plots from Random Forest:**

    - **The Random Forest Classifier was used to determine the importance of each feature in predicting the disease outcome. Feature importance was visualized using bar plots to highlight which features are most impactful in the prediction process.**

    - **Model Comparison**:

  - **Evaluation of MAE, RMSE, and Accuracy for All Models:**

    - **Bar plots were created to compare the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Accuracy across the different models: Logistic Regression, Random Forest Classifier, SVM, and Gradient Boosting.**

    - **Residual Plots:**

  - **Checked Prediction Errors Against Actual Disease Outcomes:**

    - **Residual plots were created to visualize the differences between predicted values and actual outcomes. These plots help check for any bias in the model (e.g., systematic over- or under-predictions) and ensure that the model generalizes well to unseen data.**

    - **User Testing:**

  - **Integrated Model into a User Interface:**

    - **A Gradio interface was developed to test the disease prediction model. Users can input patient data, such as BMI, age, glucose levels, blood pressure, etc., and get a real-time prediction of the likelihood of a disease (e.g., diabetes or heart disease)**

Tools and Technologies Used

- **Programming Language**: Python 3
- **Notebook Environment**: Google Colab

- **Key Libraries**:

- pandas, numpy for data handling
- matplotlib, seaborn, plotly for visualizations
- scikit-learn for preprocessing and building machine learning model
- Gradio for interactive user interface

- ## Team Members and Contributions

Prabhavathi: *Focused on* ***data cleaning****,* ***feature engineering****, and* ***model development****.*

Nivetha: *Led the* ***EDA****,* ***model evaluation****, and* ***documentation/reporting****.*

Parameshwari: *Handled* ***data preprocessing****,* ***feature engineering****, and the* ***Gradio interface development*** *for user interaction.*