STUDENT NAME	:	HARINI.S
REGISTER NUMBER	:	422223243050
INSTITUTION	:	SURYA GROUP OF INSTITUTIONS
DEPARTMENT	:	B.TECH (AI&DS)
DATE OF SUBMISSION	:	9/05/2025
GITHUB REPOSITORY LINK	:	HTTPS://GITHUB.COM/HARINIHARINI123/HARINIII.GIT

1. Problem Statement:

Many diseases are detected at advanced stages, reducing treatment effectiveness. Manual diagnosis can lead to errors, affecting patient outcomes. Traditional methods struggle to predict disease onset and progression. Large amounts of patient data can be difficult to analyze and interpret. Developing tailored treatment plans can be complex and time-consuming.

2. Abstract :

This project focuses on developing an AI-powered disease prediction system to transform healthcare by enabling early detection and personalized medicine. The problem addressed is the delayed diagnosis of diseases, leading to poor patient outcomes. The objective is to create a predictive model that accurately forecasts disease likelihood based on patient data. Our approach involves collecting and preprocessing patient datasets, developing and training machine learning models, and evaluating their performance. The outcome is a robust predictive model that achieves high accuracy in disease prediction, enabling healthcare professionals to make informed decisions and improve patient care. By leveraging AI and machine

learning, this project aims to revolutionize healthcare by providing timely and targeted interventions. The system has the potential to enhance clinical decision-making and patient outcomes.

3. System Requirements:

Hardware Requirements:

Multi-core CPU (e.g., Intel Core i5 or higher) . Memory 8 GB RAM or more. Storage 256 GB or more free disk space

Software Requirements:

Programming Language Python 3.8 or higher. Machine Learning Frameworks scikit-learn, TensorFlow, or PyTorch. Data Analysis Libraries Pandas, NumPy, and Matplotlib. Operating System Windows, macOS, or Linux

Additional Requirements:

Dataset Relevant patient dataset for training and testing. Computational Resources GPU acceleration for faster model training (optional). These requirements ensure smooth development and deployment of the AI-powered disease prediction system.

4. Objectives :

Primary Objectives:

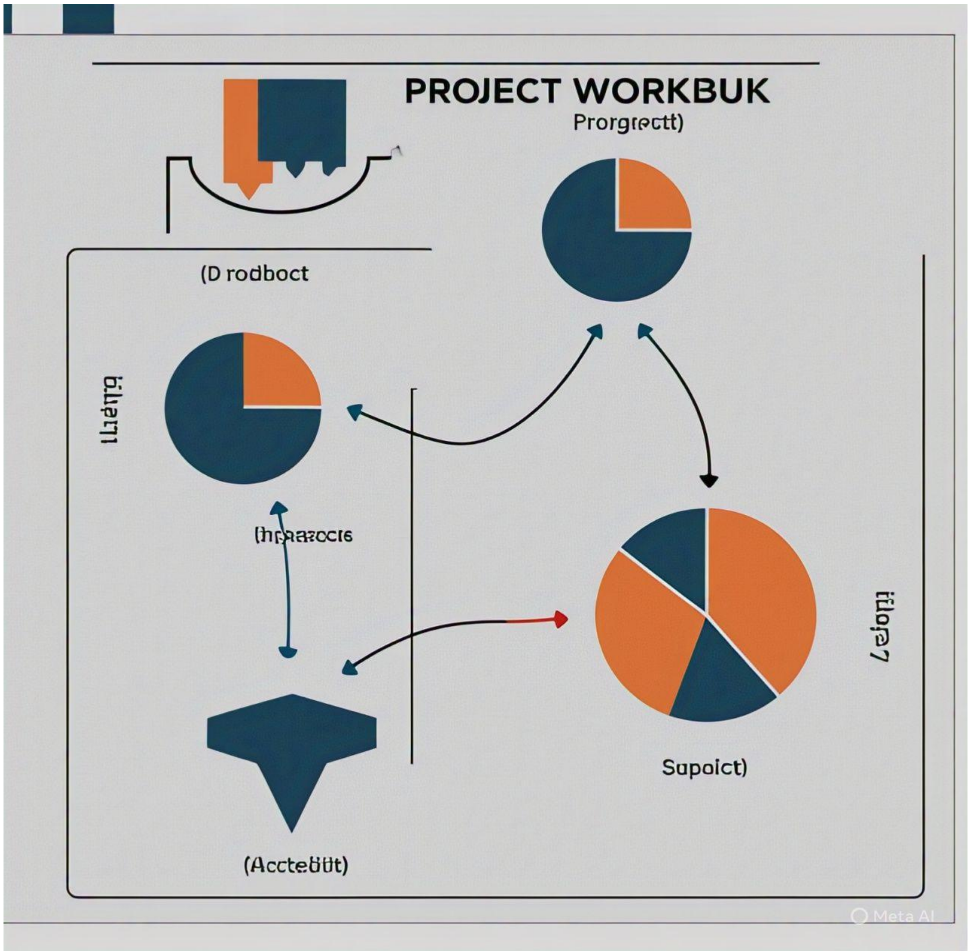
Develop Predictive Model Create an accurate AI-powered model to predict disease likelihood based on patient data. Improve Patient Outcomes Enable early detection and personalized medicine to enhance patient care and outcomes. Enhance Clinical Decision-Making Provide healthcare professionals with informed insights to make timely and targeted interventions.

Secondary Objectives:

Evaluate Model Performance Assess the accuracy, precision, and recall of the predictive model. Identify High-Risk Patients Detect patients at high risk of disease onset and enable proactive care. Improve Healthcare Efficiency Reduce healthcare costs and resource utilization through early detection and prevention. These objectives aim to harness the power of AI and machine learning to transform healthcare and improve patient lives.

5. Project Workflow Flowchart :

Data Collection- Gather patient data from various sources . Data Preprocessing - Clean, transform, and prepare data for analysis. Feature Engineering - Extract relevant features from patient data .



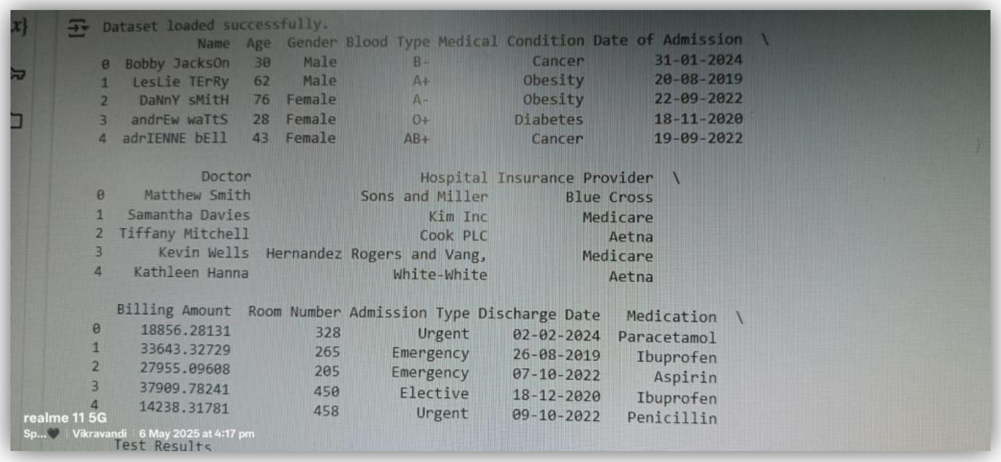
6. Dataset Description :

SOURCE: gts.ai

TYPE: Public

SIZE AND STRUCTURE: 30 rows and 12 columns

Healthbook 1()



```
Dataset loaded successfully.
            Name  Age  Gender Blood Type Medical Condition Date of Admission  \
0  Bobby JacksOn   30    Male         B-            Cancer        31-01-2024
1   LesLie TErRy   62    Male         A+           Obesity        20-08-2019
2    DaNnY sMitH   76  Female         A-           Obesity        22-09-2022
3   andrEw waTtS   28  Female         O+          Diabetes        18-11-2020
4  adrIENNE bEll   43  Female        AB+            Cancer        19-09-2022

             Doctor                    Hospital Insurance Provider  \
0     Matthew Smith             Sons and Miller        Blue Cross
1   Samantha Davies                     Kim Inc          Medicare
2  Tiffany Mitchell                    Cook PLC             Aetna
3       Kevin Wells  Hernandez Rogers and Vang,          Medicare
4    Kathleen Hanna                 White-White             Aetna

   Billing Amount  Room Number Admission Type Discharge Date   Medication  \
0     18856.28131          328         Urgent     02-02-2024  Paracetamol
1     33643.32729          265      Emergency     26-08-2019    Ibuprofen
2     27955.09608          205      Emergency     07-10-2022      Aspirin
3     37909.78241          450       Elective     18-12-2020    Ibuprofen
4     14238.31781          458         Urgent     09-10-2022   Penicillin

  Test Results
```

7.Data Preprocessing :

Handling Missing Values

```
df = df.dropna()
print("\nMissing values handled (rows with NaNs dropped).")
```

This line removes rows with missing values (NaNs) from the dataFrame df using dropna().

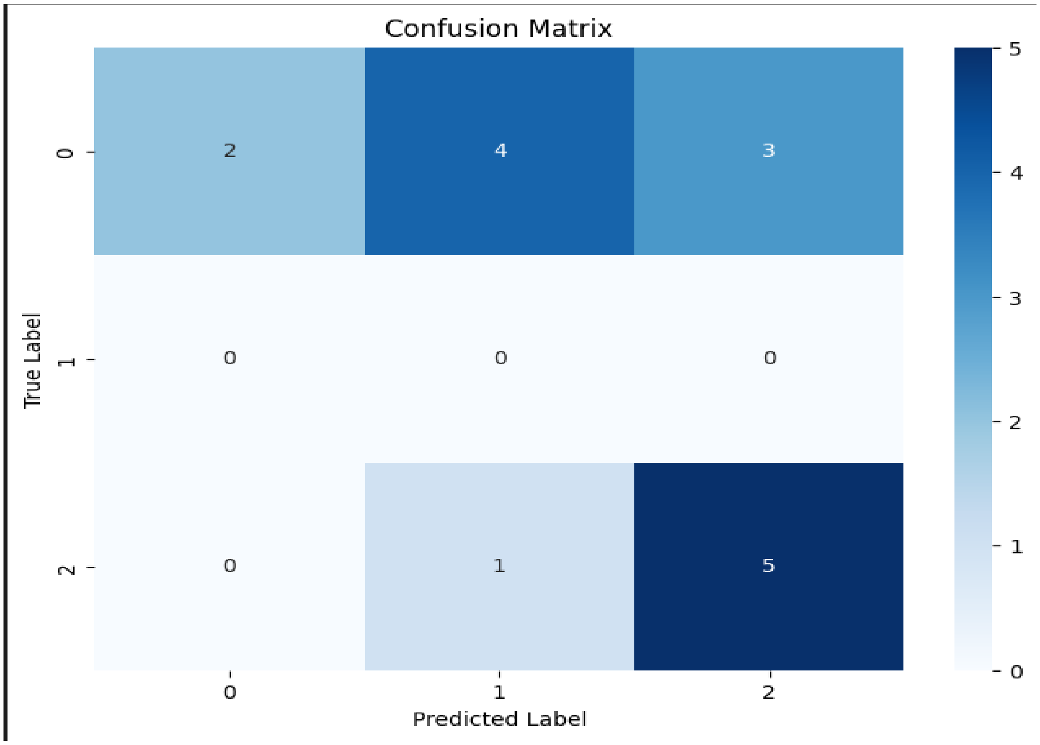
Encoding Categorical Features

```
categorical_cols = [col for col in df.columns if df[col].dtype == 'object']
print("\nCategorical columns:", categorical_cols)

label_encoders = {}
for col in categorical_cols:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le
print("\nCategorical columns encoded.")
print(df.head())
```

This block identifies categorical columns (columns with data type 'object') and converts them into numerical representations using LabelEncoder. It iterates through each categorical column, creates a LabelEncoder object, fits it to the data, transforms the column, and stores the encoder in a dictionary.

8. Exploratory Data Analysis (EDA) :

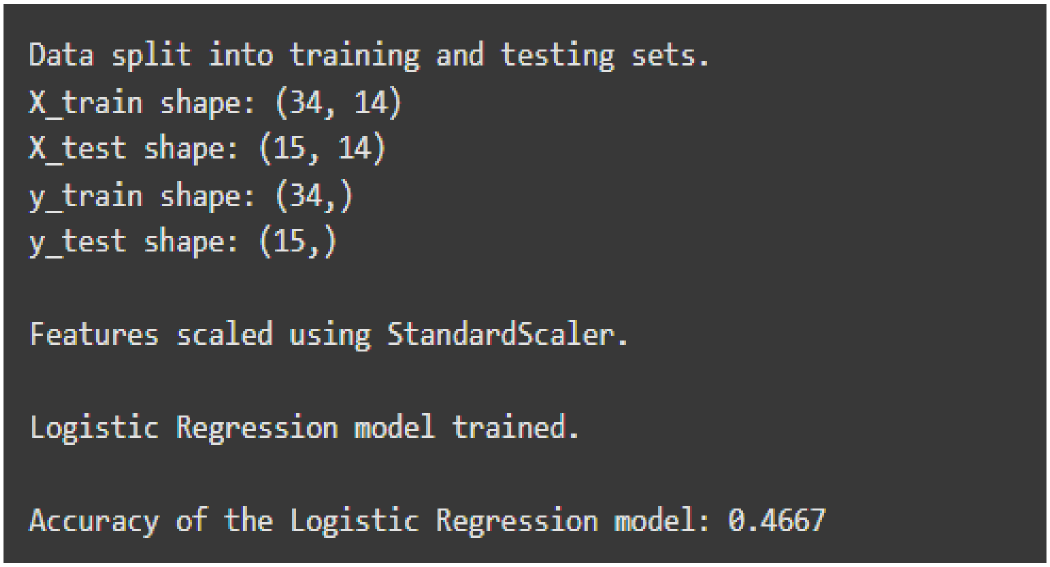
Data Visualization Use plots (e.g., histograms, scatter plots) to identify trends, patterns, and correlations in the data `df.hist()`, `df.corr()`. Distribution Analysis Examine data distributions to identify skewness, outliers, and anomalies. Correlation Analysis Identify relationships between variables using correlation coefficients and heatmaps. Pattern Identification Discover insights and trends in the data, such as relationships between variables or anomalies. EDA helps understand the data, identify potential issues, and inform machine learning model development.



9. Feature Engineering :

Feature Selection Identify and select relevant features that contribute to model performance. Feature Transformation Apply transformations to improve feature distribution and model performance. Feature Creation Generate new features from existing ones to capture complex relationships.   Encoding Categorical Variables Convert categorical variables into numerical variables for model compatibility. Dimensionality Reduction Reduce feature dimensionality using techniques like PCA or t-SNE to prevent overfitting . Effective feature engineering enhances model performance, interpretability, and reliability.

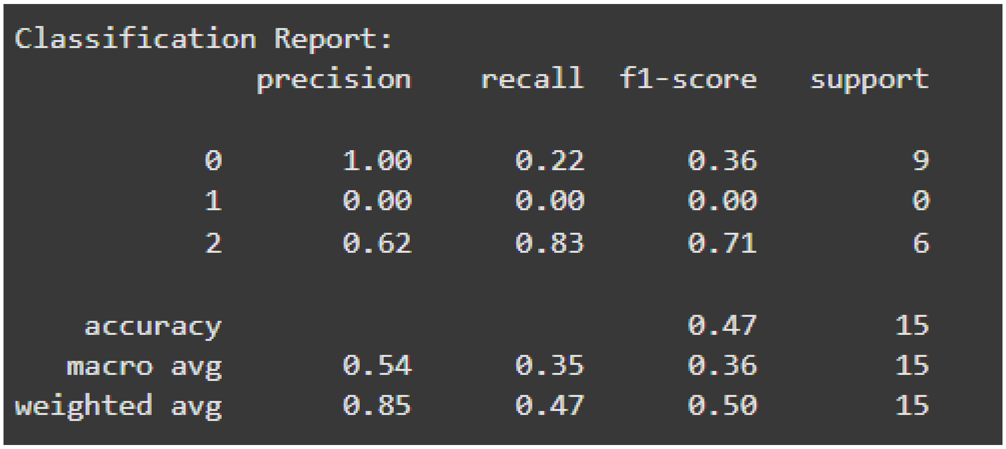
10. Model Building :



```
Data split into training and testing sets.
X_train shape: (34, 14)
X_test shape: (15, 14)
y_train shape: (34,)
y_test shape: (15,)

Features scaled using StandardScaler.

Logistic Regression model trained.

Accuracy of the Logistic Regression model: 0.4667
```

Algorithm Selection Choose suitable machine learning algorithms based on problem type (classification, regression, etc.) and data characteristics. Model Training Train models using training data and selected algorithms . Hyperparameter Tuning Optimize model hyperparameters to improve performance and prevent overfitting. These steps enable the development of robust and accurate machine learning models.

11. Model Evaluation :



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.22      0.36         9
           1       0.00      0.00      0.00         0
           2       0.62      0.83      0.71         6

    accuracy                           0.47        15
   macro avg       0.54      0.35      0.36        15
weighted avg       0.85      0.47      0.50        15
```

Metrics Selection Choose relevant evaluation metrics (e.g., accuracy, precision, recall, F1-score) based on problem type and requirementsThese steps help assess model performance, identify areas for improvement, and ensure the model meets problem requirements.

12. Deployment :

Deployment method: Gradio

Public link : https://04975cf4071fa7d7f3.gradio.live/

13. Source code :

https://colab.research.google.com/drive/1lD8YbdeoxSn0Lw1hy8xLNVzGywnCxyu4

14. Future scope :

Potential Enhancements:

Data Expansion Collect more data to improve model performance and generalizability. Model Optimization Experiment with advanced architectures and techniques (e.g., transfer learning, ensemble methods) to enhance accuracy. Real-time Deployment Integrate with real-time data sources for dynamic predictions.User Interface Improvements Enhance the UI for better user experience and accessibility. Explainability and Interpretability Incorporate techniques to provide insights into model decisions.

Applications:

Industry-specific Solutions Tailor the model for specific industries or use cases. Integration with Other Tools Combine with other tools or platforms for streamlined workflows.

15. Team Members and Roles :

Data cleaning	:	Shanmugapriya . M
EDA	:	Harini
Feature engineering	:	Mounika
Model development	:	Sumithira
Documentation and reporting	:	Shanmugapriya . M