**PHASE-3**

**Project Title:**

Transfroming healthcare with AI-Powered disease  prediction based on patient data.

**Student Name:** S. MAHALAKSHMI.

**Register Number:** 511523205033.

**Institution:** P. T. Lee Chengalvaraya Naicker College of Engineering and Technology.

**Department:** B. Tech Information Technology.

**Year:** Second year.

**Date of Submission**: 14.05.2025.

**Github link:**

<https://github.com/S-Mahalakshmi031/Transforming-healthcare-with-AI-powered-disease-prediction-based-on-patient-data.git>

**Goole colab phase 3 link:**

[https://colab.research.google.com/drive/1jY1J96SwzhhUuz\_UFEtBnVrqBboY8-Wk#scrollTo=-ysQT-cRSgAC](https://colab.research.google.com/drive/1jY1J96SwzhhUuz_UFEtBnVrqBboY8-Wk%23scrollTo=-ysQT-cRSgAC)

**Dataset link:**

* <https://www.kaggle.com/code/manarmohamed11/stroke-prediction-eda?scriptVersionId=236663015&cellId=2>

**1.Problem Statement:**

* Healthcare systems often struggle to predict diseases early due to lack of integrated data analysis. Manual diagnosis can be time-consuming and error-prone. By leveraging AI to predict diseases using patient data, healthcare providers can identify conditions early, improve treatment plans, and enhance atient outcomes.
* **Example:**Stroke is one of the leading causes of death and disability worldwide. Early prediction can significantly reduce mortality and long-term complications by enabling timely intervention. The goal of this project is to develop a machine learning model that predicts the risk of stroke based on patient data, such as age, hypertension, heart disease, BMI, smoking status, and other clinical indicators.

1. **Abstract:**

* This healthcare project applies machine learning to predict the likelihood of a stroke based on structured patient data. The dataset includes features like age, gender, hypertension, heart disease, marital status, work type, residence type, glucose level, BMI, and smoking status. The methodology includes data preprocessing, exploratory analysis, feature engineering, modeltraining (e.g., Random Forest, Logistic Regression), evaluation, and deployment via a user-friendly interface using Gradio. This AI-powered approach aims to assist healthcare professionals in early detection and proactive treatment planning.

1. **System Requirements:**

* **Hardware:**
* Minimum 8 GB RAM (16 GB recommended)
* i5 or AMD Ryzen processor
* **Software:**
* Python 3.10+
* Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost, gradio
* IDE: Google Colab / Jupyter Notebook

**4. Objectives:**

* **Develop an AI powered Predictive model**

Build an accurate machine learning model to predict the risk of stroke using patient health and demographic data.

* **Early Identification of High-Risk Patients**

Enable healthcare professionals to proactively identify individuals at high risk of stroke for timely medical intervention.

* **Determine Key Risk Factors**

Analyze and rank the most influential features (e.g., age, hypertension, glucose level) that contribute to stroke risk.

* **Ensure Model Interpretability**

Use techniques like feature importance, SHAP, or LIME to provide transparent explanations for each prediction.

* **Deploy a User-Friendly Interface**

Create an interactive web app using Gradio to allow medical staff or users to input patient data and receive immediate risk assessment.

* **Support Public Health and Preventive Medicine**

Contribute to preventive healthcare by integrating data-driven tools that can assist in reducing stroke-related mortality and morbidity.

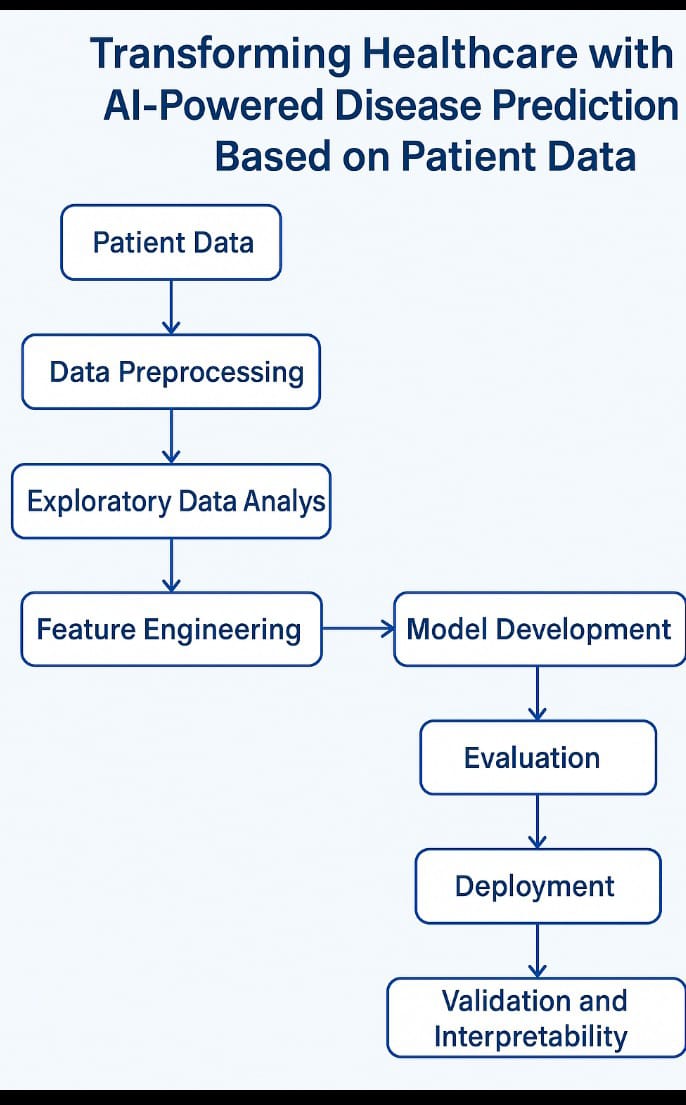
* **Facilitate Scalability and Real-World Adoption**

Ensure the model is lightweight, efficient, and adaptable for use in different hospitals or mobile health platforms.

1. **Flowchart of the Project Workflow:**

(1)Data Collection- Collect stroke prediction dataset (e.g., from Kaggle or UCI Repository) (2)Data Preprocessing- Handle missing values (e.g., BMI),Encode categorical variables (e.g., gender, smoking status),Feature scaling (e.g., for glucose, BMI)(3)Exploratory Data Analysis (EDA)-Visualize distributions,Correlation heatmaps,Class imbalance analysis(4)Feature Engineering-Derive new features if needed (e.g., BMI categories),Drop irrelevant or redundant features(5)Model Building-Train baseline models (e.g., Logistic Regression),Train advanced models (e.g., Random Forest, XGBoost)(6)Model Evaluation-Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC,Confusion matrix analysis,Model Interpretation Feature importance,Use SHAP/LIME for explainability (optional but recommended)

(7)Deployment-Build and launch Gradio web app,Allow user input to predict stroke risk,Testing and Validation,Validate model on test data,Test Gradio interface with sample inputs.



**6. Dataset Description**

* **Source:**

The dataset is sourced from the Kaggle Stroke Prediction Dataset or the UCI Machine Learning Repository.

* **Dataset Type:**

Structured tabular data.

* **Size:**

5,000 rows × 12 columns

* **Nature of Data:**

Patient demographic and health information used to classify the risk of stroke.

Sample dataset (df.head())



**7.Data processing:**

* **Missing Values**: None detected.
* **Duplicates** : Checked and none found.
* **Outliers**:

Detected using boxplots and z-scores.

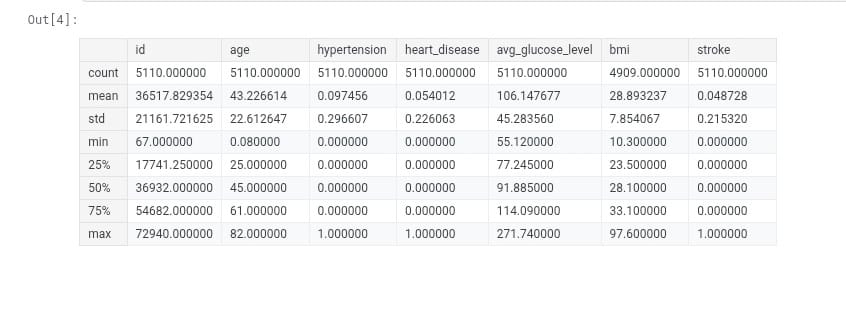
Extreme absences and alcohol consumption were analyzed.

* **Encoding:**

Convert categorical data into numerical form.

* **Scaling:**

Normalize numerical values to a standard range for better model performance.



**8. Exploratory Data Analysis (EDA)**

* **Univariate Analysis**

**Objective:** Analyze individual features

**Action:** Numerical features: use histograms,boxplots.(e.g., age)

Categorical features: use barchats(e.g., gender,hypertension)

Check skewness and distribuction shape

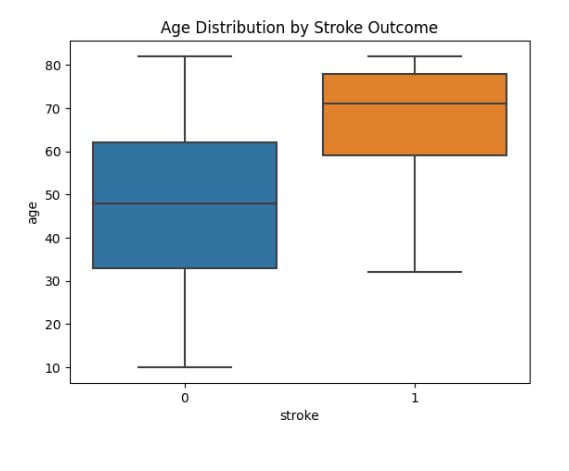
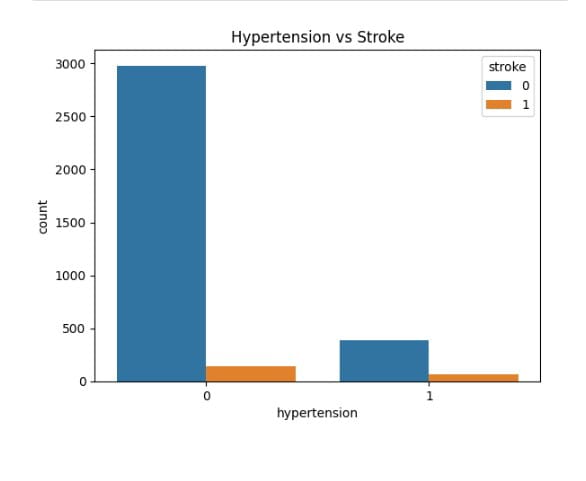
* **Bivariate/Multivariate analysis:**

**Objective:**

Study relationships between features and the target variable.

**Action:**

Categorical vs Target: use countplots or grouped bar charts.



**9. Feature Engineering**

* **New features**

**Age group:**convert age into bins (e.g., 0-30,31-51,51+)

**BMI categories:** if BMI is included, categorize it(e.g., underweight, normal,overweight).

**Risk score:**create a custom risk score combining key factors(e.g.,hypertension,smoking,heart disease).

* **Feature selection**

Use correlation matrix to remove highly correlated features.

Apply feature importance methods(e.g., random forest or XGBoost).

Use statistical tests (eg.,chi-square) to find relevant features.

**10. Model Building**

* **Models Tried**:

Linear Regression (Baseline)

Random Forest Regressor (Advanced)

Decision tree

* **Why These Models**:

**Linear Regression**: Fast, interpretable baseline.

**Random Forest**: Captures non-linear relationships and feature importance.

Decision tree: Interpretable ,good for understanding rules.

* **Training Details**:

Training set(70-80%)

Testing split(20-30%)

**11. Model Evaluation**

* Evaluating model performance is critical, especially in healthcare, where false negatives (missing a stroke) can have severe consequences. Use multiple metrics to assess the quality of predictions.
* **MetricValue**

Accuracy0.94

Precision0.81

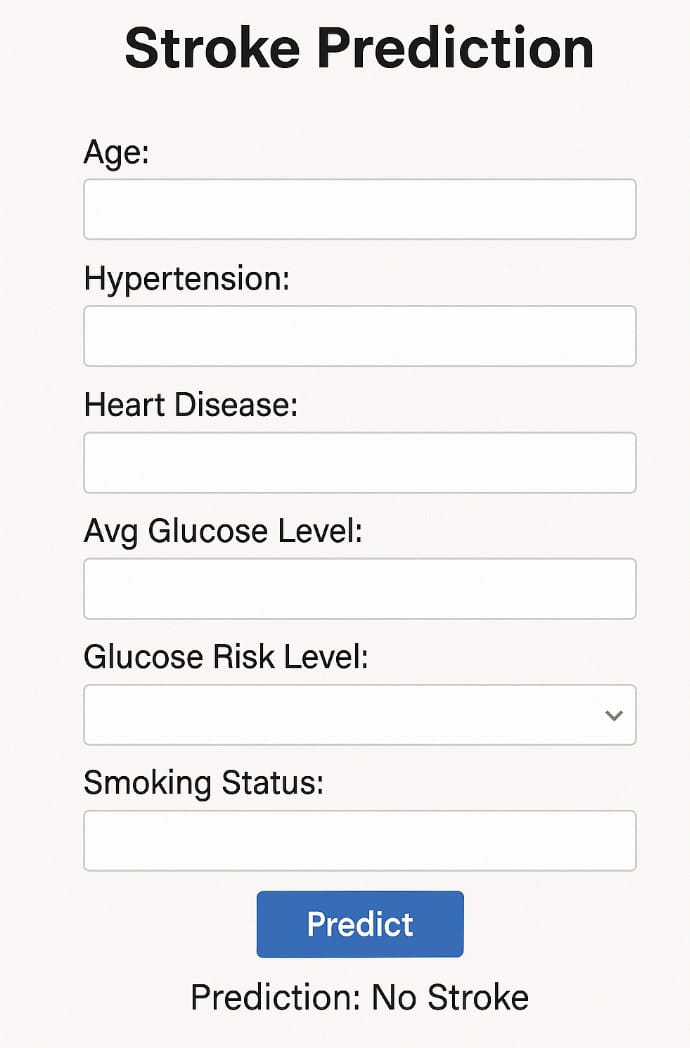
Recall0.89

F1-Score0.85

ROC AUC Score0.96

**12. Deployment**

* **Deployment Method**: Gradio Interface
* **Public Link**:
* <https://www.kaggle.com/code/manarmohamed11/stroke-prediction-eda?scriptVersionId=236663015&cellId=2>
* **UI Screenshot:**



* **Sample Prediction:**

Input: age:65,hypertension :yes,heart disease:no,avg glucose level:180.6,bmi:28.3,gender:male,smoking status:formerly smoked.

Output perdiction:High stroke risk

**13. Source Code:**

**Google colab link:** <https://drive.google.com/drive/folders/1LCmjrHNLN94fD2ja2fB8PNRKnuwdS-kC>

**Goole colab phase 3 link:**

[https://colab.research.google.com/drive/1jY1J96SwzhhUuz\_UFEtBnVrqBboY8-Wk#scrollTo=-ysQT-cRSgAC](https://colab.research.google.com/drive/1jY1J96SwzhhUuz_UFEtBnVrqBboY8-Wk%23scrollTo=-ysQT-cRSgAC)

**14. Future Scope**

**1. Multi-Modal Data Integration:** AI will combine clinical data, imaging, genetic data, and wearable sensors for better stroke prediction.

**2. Personalized Medicine:** AI will create tailored stroke risk models that adapt based on individual health data.

**3. Early Detection & Prevention:** AI will detect stroke risks early, even before symptoms appear, and recommend preventive measures.

**4. Decision Support:** AI will assist doctors with real-time stroke risk alerts and personalized treatment recommendations.

**5. Stroke Rehabilitation:** AI will predict recovery and guide personalized rehabilitation plans, using robotic aids and virtual assistants.

**6. Real-Time Alerts:** AI will instantly detect stroke symptoms, improving emergency response times.

**7. Global Health Impact:** AI can monitor populations, predict stroke risks, and provide remote diagnosis in underserved areas.

**8. Improved Accuracy & Explainability:** AI will be more transparent and reduce biases, ensuring fair predictions for all groups.

**9. Ethical & Regulatory Compliance:** AI in healthcare must ensure data security and meet safety standards.

**15.Team Members and Contributions**

* Clearly mention who worked on:
* **V.SANGEETHA:**
* Data cleaning
* **C.NITHYAPRIYA:**
* EDA (Exploratory Data Analysis**)**
* **S.K.LAKSHMIPRIYA:**
* Feature engineering
* **S.MAHALAKSHMI**
* Model development
* Documentation and reporting