SYNTHETIC STROKE PREDICTION
Using Machine Learning

"Prevention is better than cure — with machine learning, prediction becomes prevention."

Introduction

Stroke is one of the most critical medical emergencies, often leading to severe disability or death.
Predicting the chances of a stroke before it occurs can save lives through early diagnosis and preventive care.
Using machine learning, we can analyze health and lifestyle patterns to identify people at higher risk and support timely medical intervention.

Objective

- Develop a synthetic stroke prediction model using ML.
- Handle missing data and class imbalance effectively.
- Compare multiple algorithms for the most accurate prediction.
- Prepare the model for deployment in real-world systems.
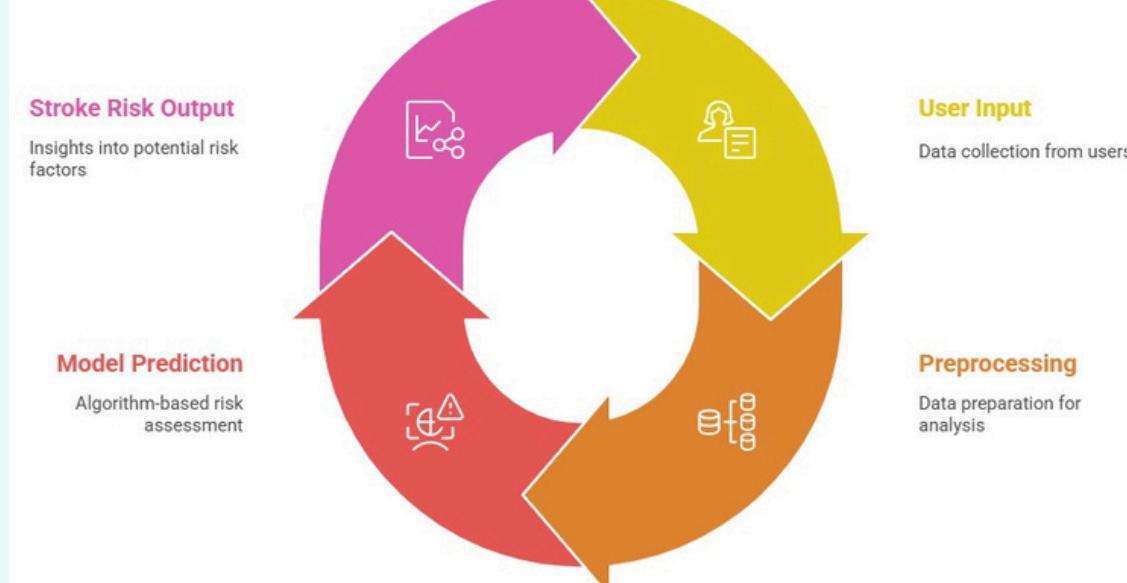
Machine Learning Model Used

A Random Forest Classifier model was used for stroke prediction. It works by building multiple decision trees and combining their outputs to improve accuracy and stability.
Why Random Forest?
- Handles both numerical and categorical data effectively
- Resistant to overfitting
- Provides feature importance insights
- High accuracy and interpretability
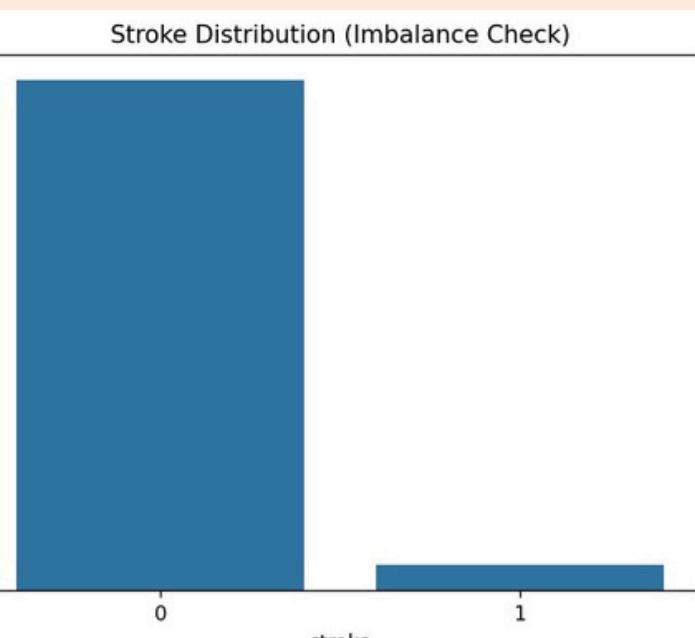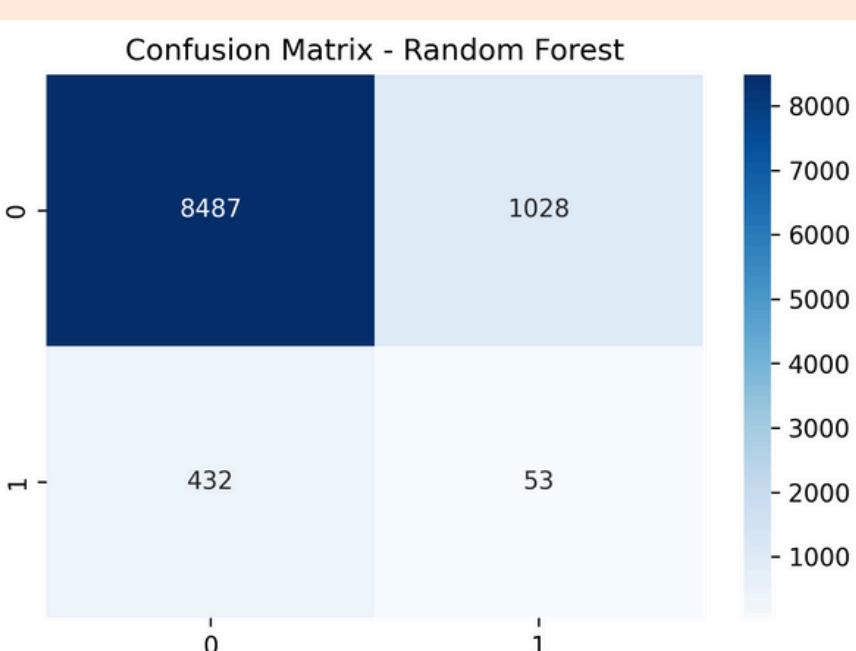
Workflow

Stroke Risk Output
Insights into potential risk factors

User Input
Data collection from users

Model Prediction
Algorithm-based risk assessment

Preprocessing
Data preparation for analysis

Technologies Used

- VS Code
- Python
- Pandas
- Scikit-learn
- Matplotlib
- Imbalanced-learn

Visualisations

Confusion Matrix - Random Forest

	0	1
0	8487	1028
1	432	53

Stroke Distribution (Imbalance Check)

Model Evaluation

```
=== RANDOM FOREST RESULTS ===
              precision    recall  f1-score   support

           0       0.95      0.89      0.92      9515
           1       0.05      0.11      0.07       485

    accuracy                           0.85     10000
   macro avg       0.50      0.50      0.49     10000
weighted avg       0.91      0.85      0.88     10000

ROC-AUC Score: 0.47558125802449747
```

Sample Output

Stroke Prediction System with AI Suggestions
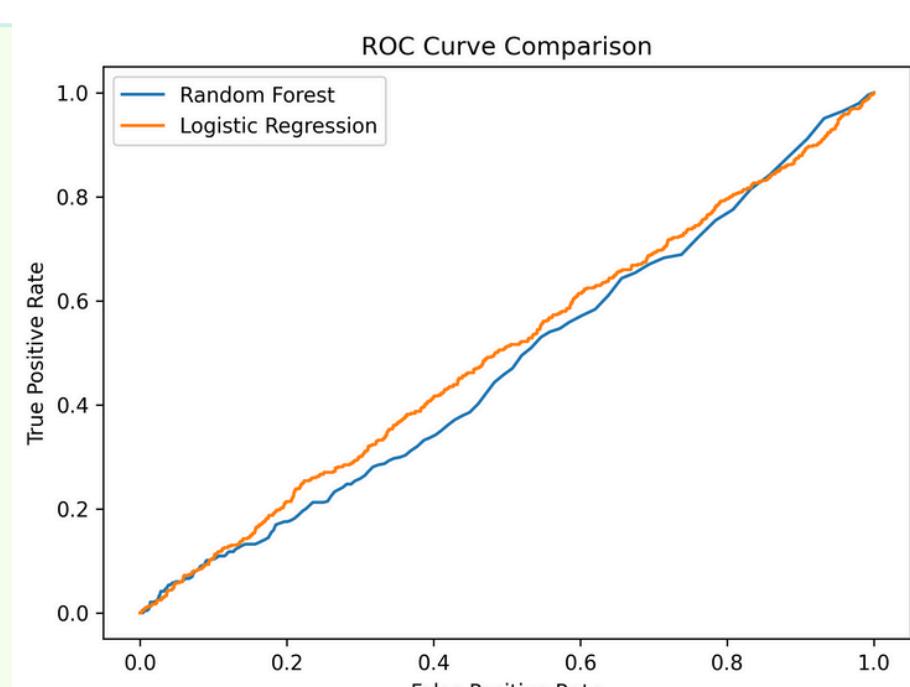
|	
Gender	Female
Age	59
Hypertension	No
Heart Disease	Yes
Ever Married	Yes
Work Type	Govt job
Residence Type	Rural
Avg Glucose Level	99
BMI	87
Smoking Status	never smoked

Predict & Get Suggestions

☑ No Stroke Detected.

AI-Powered Suggestions:

Performance Summary

- Random Forest performed significantly better than Logistic Regression on all major metrics
- Random Forest showed higher recall, important for detecting patients who may be at risk
- Logistic Regression provided good interpretability but lower predictive power

ROC Curve Comparison
— Random Forest
— Logistic Regression

Guided By::
Dr. Abhilasha Joshi

Aakanksha Sah 202301070148
Purva Nigade 202301070146
Amruthavarshini Repalle 202301070158
Pratiksha Shinde 202301070155