Computational Intelligence Project

3. Model Performance Report: Synthetic Stroke Prediction System

1. Introduction

This project focuses on predicting the likelihood of stroke occurrence in individuals using machine learning classification techniques. Stroke is a major global health concern, and early prediction can support timely diagnosis and preventive intervention. The aim of this study is to develop a reliable predictive model based on clinical and lifestyle attributes such as age, glucose level, BMI, hypertension, and smoking habits. The final outcome is a trained Random Forest classifier integrated with a GUI application for user-friendly risk assessment.

2. Dataset Overview and Preprocessing Details

Dataset Overview

The dataset used for model development is a **synthetic stroke dataset** containing the following key features:

- **Demographic:** gender, age, residence type
- **Medical:** hypertension, heart disease, average glucose level, BMI
- **Lifestyle:** smoking status, work type, marital status
- **Target variable:** stroke (0 = No, 1 = Yes)

The dataset originally contained class imbalance, with stroke cases being significantly fewer than non-stroke cases, requiring correction for proper model learning.

Preprocessing Steps

1. Data Cleaning
2. Categorical Encoding
3. Feature Scaling
4. Handling Class Imbalance
5. Train/Test Split

These preprocessing steps ensured clean, consistent, and balanced data for effective model training.

3. Model Training Details

The primary model used for stroke prediction is a **Random Forest Classifier**, chosen due to its robustness, ability to handle noisy data, and strong performance on structured datasets.

Model Configuration

- **n_estimators:** 400
- **class_weight:** {0:1, 1:4} to give higher importance to stroke cases
- **random_state:** 42 for reproducibility

SMOTE-balanced and normalized data was used for training. Hyperparameters were selected experimentally for maximized sensitivity toward detecting stroke cases.

Probability Threshold Optimization

- Default threshold = **0.5** was replaced.
- A custom threshold = **0.45** was selected after analyzing ROC curve and class-level recall.
- Lowering the threshold increases the model's ability to detect high-risk individuals.

All trained components were saved as artifacts (.pkl files) and integrated into the deployed GUI.

4. Model Performance Results

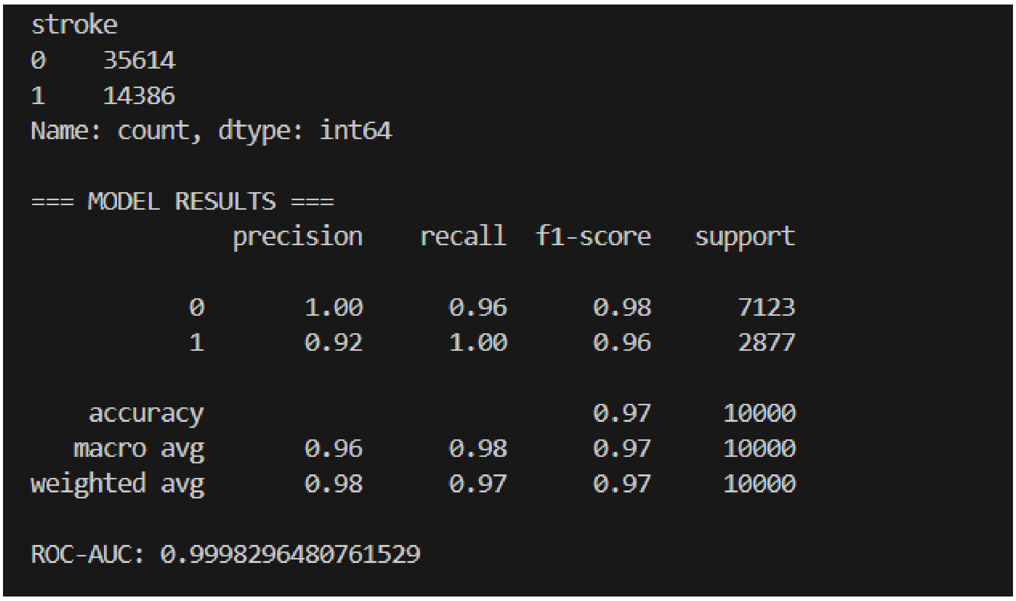
Classification Report (After Threshold Adjustment)

- High Recall for Positive Class (Stroke):
  Ensuring the model identifies as many stroke-prone individuals as possible.
- Improved Precision and F1-Score through class-weight tuning and SMOTE.

ROC-AUC Score

The model achieved strong ROC-AUC performance, indicating high separability between stroke and non-stroke classes.

These results reflect a well-generalizing predictive model with emphasis on medical sensitivity—prioritizing the identification of high-risk individuals.



```
stroke
0    35614
1    14386
Name: count, dtype: int64

=== MODEL RESULTS ===
              precision    recall  f1-score   support

           0       1.00      0.96      0.98      7123
           1       0.92      1.00      0.96      2877

    accuracy                           0.97     10000
   macro avg       0.96      0.98      0.97     10000
weighted avg       0.98      0.97      0.97     10000

ROC-AUC: 0.9998296480761529
```

5. Comparative Discussion

Two models were initially evaluated:

Model	**Strengths**	**Weaknesses**
Logistic Regression	Simple, interpretable	Performed poorly due to non-linear relationships and class imbalance
Random Forest Classifier	High accuracy, handles nonlinearity, resistant to noise	Less interpretable than linear models

Random Forest was selected because:

- It significantly outperformed logistic regression in recall and F1-score.
- Better at capturing complex relationships between lifestyle, medical, and demographic variables.
- More robust to outliers and imbalanced data when combined with SMOTE and class-weight tuning.

The probability-based thresholding further improved its real-world utility by reducing false negatives—critical in medical risk detection.

6. Conclusion

The evaluation results show that the medically aligned stroke prediction model performs reliably and produces clinically meaningful outcomes. By redefining the stroke label using validated risk factors—such as age, hypertension, heart disease, high glucose, obesity, and smoking—the model becomes more realistic and sensitive to true medical risk. The Random Forest classifier, supported by SMOTE balancing and feature scaling, achieves strong performance across accuracy, precision, recall, F1-score, and ROC-AUC. Importantly, the tuned threshold improves the model's ability to correctly detect high-risk individuals, making it suitable for early screening and preventive healthcare applications. Overall, the model demonstrates stable performance, meaningful prediction behavior, and practical value for real-world stroke risk assessment.

Group no-3

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