Computational Intelligence Project

3. Model Performance Report: Stroke Prediction System

1. Introduction

The objective of this project is to build a machine-learning system that predicts the likelihood of a stroke using patient demographic and medical attributes. Two models were developed and evaluated: Random Forest Classifier and Logistic Regression. The dataset contains severe class imbalance, requiring SMOTE for oversampling.

2. Dataset Overview and Preprocessing

- Total samples: 10,000

- Stroke cases: 485

- Non-stroke: 9,515

- Missing values handled using median imputation.

- Label encoding applied to categorical features.

- StandardScaler used for feature scaling.

- Train-test split: 80/20 with stratification.

- SMOTE applied on training data.

3. *Model raining*

Random Forest: n_estimators=200,

class_weight='balanced'- High accuracy ut poor minority detection.

Logistic Regression:

- *max_iter=1000, class_weight='balanced'- Slightly better recall for stroke class.*

4.Model Performance Results

Random Forest:

```
=== 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
```

Logistic Regression:

```
=== LOGISTIC REGRESSION RESULTS ===
              precision    recall  f1-score   support

           0       0.95      0.68      0.79      9515
           1       0.05      0.33      0.09       485

    accuracy                           0.66     10000
   macro avg       0.50      0.50      0.44     10000
weighted avg       0.91      0.66      0.76     10000

ROC-AUC Score: 0.5011089381389124
```

5. Comparative Discussion

Random Forest yields higher accuracy but fails to detect actual stroke cases. Logistic Regression performs slightly better in identifying stroke patients, but both models still struggle heavily with imbalance.

6. Conclusion

Both models show limited clinical usefulness due to low recall for positive cases. Future improvements include adjusting class weights, using XGBoost or LightGBM, feature engineering, and improving dataset quality.

Group no-3

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