Stroke Risk Prediction Based on Patient Health Indicators

By Manuel Ramirez

Problem Statement

Problem:

Stroke is a leading cause of death and disability worldwide.

Early identification of high-risk individuals can save lives.

Predicting stroke is challenging due to complex risk factors.

Project Objectives

Goals:

Develop a machine learning model to predict stroke risk.

Use health indicators like age, glucose level, hypertension.

Support preventive healthcare and decision-making.

Dataset Overview

Dataset:

Source: Kaggle — Stroke Prediction Dataset

5,110 patient records

Features: Age, Gender, BMI, Hypertension, Heart Disease, Smoking Status

Target: Stroke (0 = No, 1 = Yes)

Data Preparation

Steps Taken:

- Imputed missing values (BMI) using median.
- Encoded categorical features with one-hot encoding.
- Standardized numerical features.
- Train-test split: 70% training / 30% testing (stratified).

Models Built

Algorithms:

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (SVM)

Hyperparameter Tuning:

- GridSearchCV optimization for Random Forest and SVM.

Model Performance

- **Model Metrics Summary:**

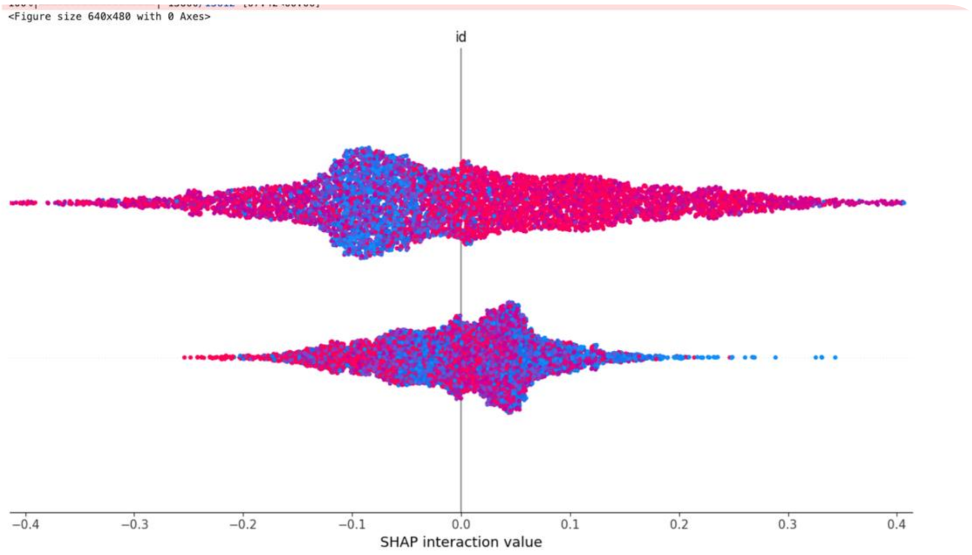
Model	Accuracy	Recall (Stroke)	F1-Score	ROC AUC
Logistic Regression	0.74	0.76	0.22	0.84
Random Forest (Tuned)	0.87	0.31	0.19	0.92
SVM (Tuned)	0.71	0.73	0.20	0.81

Key Insights

Findings:

- Age, Average Glucose Level, Heart Disease were top predictors.
- Class imbalance affected recall despite high overall accuracy.
- SHAP analysis improved model interpretability for clinical use.

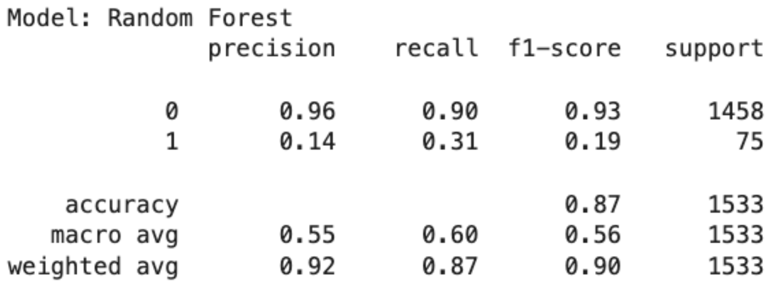
SHAP Feature Importance Plot



```
<Figure size 640x480 with 0 Axes>
```

"Feature importance derived from SHAP analysis. Age, average glucose level, and heart disease status were the strongest predictors of stroke risk."

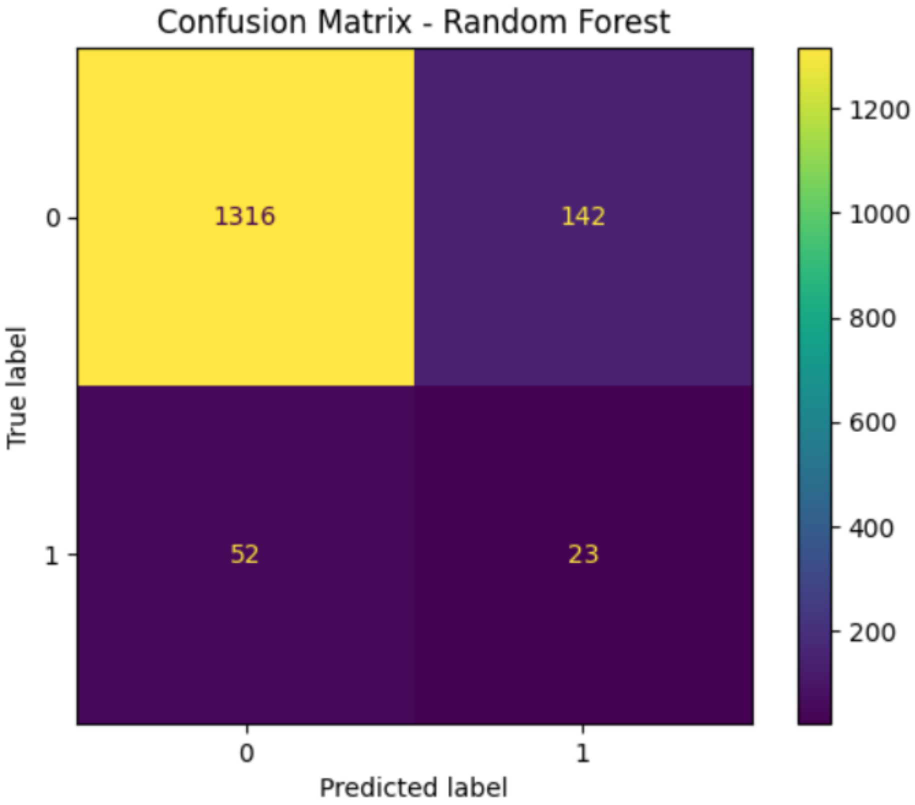
Confusion Matrix for Final Model (Random Forest)



```
Model: Random Forest
              precision    recall  f1-score   support

           0       0.96      0.90      0.93      1458
           1       0.14      0.31      0.19        75

    accuracy                           0.87      1533
   macro avg       0.55      0.60      0.56      1533
weighted avg       0.92      0.87      0.90      1533
```



"Confusion Matrix showing Random Forest model performance. While 'No Stroke' cases are classified accurately, 'Stroke' prediction remains challenging due to class imbalance.”

Challenges & Limitations

Challenges:

- Severe class imbalance (only ~5% positive stroke cases).
- Limited feature set (missing deeper clinical history).

Future Work:

- Apply SMOTE for better balance.
- Test advanced models (e.g., XGBoost).
- Expand dataset with more clinical features.