Brain Stroke Prediction model

INTRODUCTION TO DATA SCIENCE FOR ENGINEERS

CS22B1082
Thallapally Nimisha

PROJECT OVERVIEW

- **Objective:** Develop a predictive model to assess stroke risk based on health and lifestyle data.
- **Purpose:** Aid healthcare providers in early identification of individuals at risk for stroke.
- **Goal:** Leverage patient data for accurate, data-driven stroke prediction.

Key Components:

- Data pre-processing
- Model training and evaluation
- Performance metrics

DATASET OVERVIEW

Features Included:

- **Demographics:** Gender, Age
- **Medical History:** Hypertension, Heart Disease
- **Lifestyle & Environment:** Ever Married, Work Type, Residence Type, Smoking Status
- **Health Metrics:** Average Glucose Level, BMI
- **Target Variable:** Stroke (1 = Yes, 0 = No)

STEPS INVOLVED

- **Data Collection and Exploration:**
  - Loaded and examined the dataset.
  - Assessed data distribution and initial patterns.
- **Data Preprocessing:**
  - **Handling Missing Values and Duplicate Values:** Imputed missing values for continuous features with mean and categorical features with mode.
  - **Encoding Categorical Variables:** Applied one-hot encoding for categorical variables, such as smoking status and residence type.
  - **Normalization:** Standardized numerical features using StandardScaler.
- **Feature Engineering:**
  - **Correlation Analysis:** Identified and removed highly correlated features to prevent multicollinearity.
  - **Principal Component Analysis (PCA):** Reduced dimensionality to retain 95% of variance, simplifying the feature space.
  - **Addressing Class Imbalance:**
    - Applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the distribution of the target variable (Stroke) classes.
- **Model Selection and Training:**
  - Chose Random Forest Classifier for its robustness and ability to handle imbalanced datasets.
  - Trained the model, adjusting class weights to further improve balance.
- **Model Evaluation:**
  - Assessed performance using confusion matrix, accuracy, precision, recall, and F1-score.
  - Verified model’s ability to identify stroke and non-stroke cases effectively.
- **Insights and Final Model Deployment:**
  - Analyzed feature importance to understand key predictors of stroke.

CORRELATION ANALYSIS

- **Purpose of Correlation Analysis:**
    - Identify relationships between features and the target variable (Stroke).
- **Process:**
    - Calculated correlation matrix to visualize feature relationships.
- **Findings:**
    - Some features like Age and Hypertension showed stronger correlations with the Stroke variable, hinting at their predictive importance.

PRINCIPAL COMPONENT ANALYSIS (PCA)

- **Purpose of PCA:**
  - Reduce dimensionality of data while retaining as much variance as possible.
- **Implementation:**
  - Applied PCA on standardized numerical features to capture essential variance.
- **Results:**
  - Reduced feature space to a few principal components, which significantly represented the original data.
  - Reduced overfitting risk by removing less informative components, leading to a more generalized model.

METHODOLOGY

Data Preprocessing:

- **Handling Missing Values:** Mean and mode imputation for continuous and categorical data, respectively.
- **Encoding Categorical Variables:** Applied one-hot encoding to categorical features.
- **Normalization:** Standardized features using StandardScaler.
- **Addressing Class Imbalance:** Used SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset.

Model Selection:

- **Classifier Used:** Random Forest Classifier
- **Training:** Adjusted class weights for imbalanced target variable.

WHY RANDOM FOREST CLASSIFIER

- **Reasons for Choosing Random Forest:**
  - **Robustness to Noise:** Random Forest is less sensitive to noisy data and outliers due to its ensemble structure.
  - **Handling Imbalanced Data:** The algorithm can handle imbalanced classes well, especially when class weights are adjusted.
  - **Interpretability:** Provides feature importance scores, enabling insights into which factors are most influential in predicting stroke.

- **Benefits in This Model:**
  - Achieved high accuracy with strong precision and recall metrics, indicating well-balanced performance.
  - Offers high flexibility and works well with a mix of categorical and numerical data, as found in our dataset.

MODEL EVALUATION

Metrics:

- **Accuracy:** 92.09%
- **Precision, Recall, F1-Score:**
  - **Class 0 (No Stroke):** Precision 0.95, Recall 0.89, F1-score 0.92
  - **Class 1 (Stroke):** Precision 0.90, Recall 0.95, F1-score 0.92
- **Interpretation:** High accuracy with strong precision and recall across both classes, indicating balanced performance.

Thank you!