Capstone Project 2: Predicting Stroke Risk Based on Patient Health Indicators

1. Project Overview

Stroke remains one of the leading causes of death and disability worldwide. Early identification of individuals at high risk is crucial for timely intervention and reducing healthcare burdens.
This project aims to build machine learning models to predict the likelihood of stroke occurrence based on clinical and demographic health indicators.

2. Problem Identification

Problem Statement

Stroke is a major global health issue that often results in severe disability or death. While many strokes are preventable, early identification of high-risk individuals is a persistent challenge due to the interplay of multiple health factors.
The objective of this project is to develop a predictive classification model that estimates stroke risk using health metrics such as hypertension status, heart disease presence, glucose levels, and BMI.

Context

Preventive care organizations, healthcare providers, insurance companies, and public health institutions are seeking ways to leverage data science to improve early detection of stroke risk.
This project focuses on using accessible individual health indicators to create interpretable and actionable predictive models to aid proactive healthcare planning.

Success Criteria

- Achieve a model AUC (Area Under the ROC Curve) greater than 75%.
- Identify and interpret key predictors of stroke.
- Create clear visualizations for non-technical healthcare professionals.

- Maintain model interpretability to build clinical trust.

3. Data Sources

- **Dataset**: Stroke Prediction Dataset
- **Source**: Kaggle (uploaded by fedesoriano)
- **Access**: link here
- **Description**: 5,110 individual records containing indicators like age, gender, hypertension, heart disease, average glucose level, BMI, and smoking status.

4. Data Wrangling

- **Missing Values**: BMI missing values imputed using median strategy.
- **Categorical Encoding**: One-hot encoding for categorical features such as gender, work type, and smoking status.
- **Feature Scaling**: StandardScaler applied to numeric features to standardize magnitudes.
- **Data Splitting**: 70% training, 30% testing split with stratification to preserve class distribution.

5. Exploratory Data Analysis (EDA)

- **Stroke Prevalence**: Only ~5% of individuals in the dataset had experienced a stroke, leading to class imbalance.
- **Key Feature Insights**:

    - Higher age and glucose levels were associated with increased stroke risk.
    - Hypertension and heart disease strongly correlated with stroke occurrence.
- **Visualizations**:
    - Distribution plots for age, glucose level, and BMI.
    - Correlation heatmaps to reveal relationships among features.
    - Boxplots of health indicators stratified by stroke status.

6. Modeling

Models Built

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

Hyperparameter Tuning

- **Random Forest**: Tuned with GridSearchCV for number of estimators, max depth, minimum samples split/leaf.
- **SVM**: Tuned with GridSearchCV for C parameter, kernel type, and gamma setting.

7. Model Performance

Model	Accuracy	Recall (Stroke)	Precision (Stroke)	F1 Score (Stroke)	ROC AUC
Logistic Regression	0.74	0.76	0.13	0.22	0.84

Random Forest (Tuned)	0.87	0.31	0.14	0.19	0.92
SVM (Tuned)	0.71	0.73	0.11	0.20	0.81

Final Model Selected

- **Random Forest Classifier (Tuned)**
  - Best combination of overall accuracy (87%) and ROC AUC (0.92).
  - Better balance between precision and recall compared to other models.

8. Feature Importance and Explainability

- **Feature Importance (Random Forest)**:
  - Age, average glucose level, and heart disease were the most influential features.
- **SHAP (SHapley Additive exPlanations)**:
  - SHAP summary plot provided transparent feature contributions for individual predictions, supporting clinical interpretability.

9. Learning Curve Analysis

- A learning curve plotted for the Random Forest model showed steady convergence between training and cross-validation scores.
- No major overfitting or underfitting detected.

10. Insights and Limitations

Insights

- Age and glucose levels are critical indicators of stroke risk.
- Heart disease and hypertension significantly increase stroke likelihood.
- Preventive interventions can be better targeted to elderly individuals with chronic conditions.

Limitations

- Class imbalance (few positive stroke cases) limited recall performance.
- Only a limited set of health indicators were available—missing deeper clinical factors like family history, medication use, etc.
- Ethical caution needed to ensure model support, not replacement, of medical judgment.

11. Conclusion

The stroke prediction model developed in this project demonstrates strong potential for aiding preventive healthcare planning by identifying individuals at elevated risk.
The Random Forest model achieved an excellent ROC AUC of 92%, validating the ability of machine learning to leverage simple health indicators for meaningful prediction.

Future improvements include exploring more complex models (e.g., XGBoost), using SMOTE/ADASYN techniques to address imbalance, and integrating richer clinical datasets.