Data Science Canvas

Project: STROKE PREDICTION USING PATIENT HEALTH RECORDS

Team: STREAM SAVANTS

Problem Statement

Business Case & Value Added

Early stroke risk detection using patient health records enables preventive care and lowers long-term treatment costs. The model supports clinicians in identifying and managing high-risk patients before emergencies, thus saving lives and resources.

Model Selection

Stacking ensemble model with Random Forest and Gradient Boosting achieves best recall and overall accuracy. These methods were chosen for their ability to deal with class imbalance and provide clinical interpretability. Alternative tested approaches include individual RF and GB models.

Model Requirements

Required: High accuracy, recall for true positive stroke cases (target ≥85%), transparent feature importances, scalability for large patient datasets, and handling of imbalanced classes

Skills

Project development relies on expertise in Python, scikit-learn, pandas, healthcare domain knowledge and EDA/ML validation.

Data Landscape

We are using pre available data from Mendeley data set.

Key attributes: age, gender, marital status, work type, residence, BMI, glucose level, hypertension, heart disease, smoking status, and stroke outcome.

Software & Libraries

Utilized: pandas
numpy
matplotlib
seaborn
scikit-learn
jupyter
jupyterlab
ipython

Execution & Evaluation

Model Evaluation

Metrics: Accuracy (~85%), Recall (~80-85%), Precision, F1-score, ROC-AUC. Five-fold cross-validation.

Data Storytelling

Results communicated using visualizations (bar plots, confusion matrices, feature rankings), clear summary tables, and interpretable key drivers. Feature importances and recall/accuracy are translated for healthcare audiences and explained in clinical terms.

Data Collection & Preparation

Data Selection & Cleansing

Clean missing BMI data (median imputation), verify and filter demographic/vital indicators. Validate and flag outliers with EDA (none significant). Standardize all features for model input.

Data Collection

Data captured from hospital EMRs or health databases for missing vital signs or risk indicators. Properties: Must be recent, structured, and span all demographics and risk categories

Data Integration

Input data unified ingest from csv and predictions and other statistics can be integrated to end systems for interpretability. in current scope we are just generating the csv files for out without any integration to end systems.

Explorative Data Analysis

Perform univariate, Bivariate and multivariate analysis and feature correlation to identify patterns, distributions and data quality issues.