STAT 451 Group 10

Introduction

This project will develop and compare predictive models using the Stroke Prediction Dataset. The dataset contains various health factors–including age, gender, and medical history–for a diverse patient population, along with stroke occurrence indicators. The goal of this project is to build classification models that can predict stroke probability based on these health factors. We will evaluate multiple feature combinations and classification algorithms to identify the optimal approach and draw insights through comprehensive evaluation and visualization.

Code for Data

```
# pip install kagglehub[hf-datasets]
import kagglehub
from kagglehub import KaggleDatasetAdapter

# Load Stroke Prediction Dataset
file_path = "healthcare-dataset-stroke-data.csv"
dataset = kagglehub.load_dataset(KaggleDatasetAdapter.HUGGING_FACE, "fedesoriano/stroke-prediction-dataset",
                                 file_path)
stroke_df = dataset.to_pandas()
print(stroke_df.head(1))
```

```
     id gender   age  hypertension  heart_disease ever_married work_type  \
0  9046   Male  67.0             0              1          Yes   Private

  Residence_type  avg_glucose_level   bmi   smoking_status  stroke
0          Urban             228.69  36.6  formerly smoked       1
```

Questions

1. Which features (e.g., age, BMI, heart disease, smoking status) are most predictive of stroke occurrence?
2. Does the inclusion of interaction terms (like age × smoking status or hypertension × heart disease) improve model performance?
3. Which machine learning model performs best for predicting stroke: Logistic Regression, Random Forest, or Neural Network?

Variables

Our stroke prediction model will utilize multiple variables, excluding patient ID to prevent sequence-based patterns. Key predictors include age, hypertension and heart disease (binary), average glucose level (continuous), BMI (with imputation for missing values), and smoking status (one-hot encoded). Demographic and socioeconomic factors will be incorporated as contextual variables: gender and work type (one-hot encoded), marital status, and residence type (binary encoded). The target variable is stroke occurrence (binary). Our modeling approach prioritizes health metrics as primary predictors while using demographic and socioeconomic factors as contextual variables. We'll focus on handling class imbalance in the target variable and investigating interaction effects between key risk factors such as age, hypertension, and glucose levels.

Methods

1. Data preprocessing: (i) Encode categorical variables into numeric variables via one-hot encoding or label encoding. (ii) Handle the missing values (e.g., BMI) via imputation. (iii) Normalize continuous variables to ensure consistent scaling across variables. (iv) Address imbalance in stroke cases via class weighting. (v) Continuous variables can be binned to extract features.
2. Models: (i) Logistic regression. (ii) Random Forest. (iii) SVM. (iv) Neural network.
3. Training and evaluation: (i) Split the data into train/validation/test sets. (ii) Implement k-fold cross-validation to ensure robust performance. (iii) Train multiple models and tune hyperparameters using grid search. (iv) Evaluate models using multiple metrics: Accuracy, Precision, Recall, F1, AUROC.
4. Unsupervised learning: Apply clustering algorithms (k-Means or DBSCAN) to identify patient subgroups and examine how these features correlate with stroke occurrence.
5. Visualization: (i) PCA or t-SNE for dimensionality reduction; (ii) Generate feature importance plots. (iii) Develop ROC and precision-recall curves for model comparison. (iv) Create partial dependence plots to visualize relationships between features and stroke probability.
6. Fairness analysis: Compare feature importance for different subgroups, e.g., male vs. female.