Stroke Prediction

David Obele & Zachery Davis

Introduction:

According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. In this project, we aim to build a machine-learning model capable of predicting the likelihood of stroke occurrences based on key health and demographic features.

The Question:

This project seeks to determine which factors most significantly contribute to stroke risk and which types of people are at risk for stroke. Using various machine learning techniques, such as logistic regression and random forest, we want to find out which model performs the best on unseen data. Overall, we want to create an accurate and well-evaluated model through comprehensive analysis and iterative improvements.

The Data:

The dataset used in this project was obtained from a publicly available healthcare dataset on Kaggle. It consists of 5,110 samples with 11 relevant features and a target variable indicating stroke occurrence. The features include gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), and smoking status.

Pre-processing:

The project followed several pre-processing steps, including data cleaning by handling missing BMI values through mean imputation, feature engineering by encoding categorical variables using one-hot encoding, and data normalization to ensure numerical features like age, average glucose level, and BMI are on the same scale. We experimented with various machine learning models, including Logistic Regression and Random Forest. Early iterations faced issues such as class imbalance, which were mitigated by applying SMOTE for oversampling the minority class.

Data Understanding/Visualization:

The target variable stroke indicates whether the patient experienced a stroke or not. Initial exploratory data analysis (EDA) revealed an average age of 43.2 years and a stroke incidence of approximately 4.9%. Visualizations such as histograms, box plots, and correlation heatmaps were used to explore feature distributions and relationships.

Modeling and Evaluation:

Model performance was evaluated using appropriate classification metrics, including accuracy, precision, recall, F1-score, and ROC AUC. The best-performing model achieved an accuracy of around 96%, with balanced precision and recall. Cross-validation confirmed that the model generalizes well to unseen data. Key questions from the introduction were answered, including identifying the most influential features such as age, average glucose level, and BMI.

Conclusion:

Through this project, we gained valuable insights into stroke risk factors. The initial problem of accurately predicting stroke occurrences was addressed successfully, with the model achieving high performance across key evaluation metrics. We learned that feature selection, handling class imbalance, and hyperparameter tuning play critical roles in enhancing model performance. Future work could involve expanding the dataset, incorporating additional health metrics, or experimenting with deep learning models. Regular evaluations with updated data can further improve the model's reliability.

Impact:

The project has significant social and ethical implications. Positively, it can lead to better healthcare decisions and proactive stroke prevention through risk prediction. However, there are ethical concerns related to data privacy, potential biases in the dataset, and over-reliance on automated healthcare systems. Addressing these issues through transparency, fairness evaluations, and regular model updates can mitigate negative impacts while maximizing societal benefits.

References:

Code: https://github.com/Zachery-Davis/Data-Mining-Project-5
Dataset: Stroke Prediction Dataset | Kaggle
Presentation: Slides