Brain Stroke Analysis – Project Report

Index

1. Introduction

Stroke is a medical emergency that occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Early detection of risk factors can help in prevention and timely treatment. With the advancement of data science and machine learning, it is possible to analyze health data and predict stroke risks more effectively. This project focuses on analyzing a brain stroke dataset, performing exploratory data analysis (EDA), applying preprocessing techniques, and building predictive models to identify factors contributing to stroke.

2. Problem Statement

The dataset under study is highly imbalanced, with very few cases of stroke compared to non-stroke. The challenge lies in handling this imbalance effectively, while also ensuring proper preprocessing of the dataset to prepare it for predictive modeling. The ultimate goal of this project is to develop a reliable machine learning model that can accurately predict the likelihood of stroke based on patient information.

3. Dataset Description

The dataset contains several patient attributes, both categorical and numerical. These include gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), and smoking status. The target variable is "stroke," which indicates whether the patient has suffered a stroke (1) or not (0). For example, age is recorded as a continuous numerical feature, while marital status is categorical with values such as "Yes" or "No." Features like hypertension and heart disease are binary, coded as 0 for "No" and 1 for "Yes." Other features such as work type, residence type, and smoking status include multiple categories and required encoding before analysis.

4. Data Preprocessing

To ensure the dataset was ready for modeling, several preprocessing steps were undertaken. Missing values were identified and appropriately handled. Duplicate records were checked and removed to maintain data integrity, and unique rows were verified. Outliers, particularly in average glucose level and BMI, were detected and addressed, as they could otherwise mislead the model's training process. Categorical variables were encoded into numerical form using label or one-hot encoding methods. Given that the dataset was imbalanced, with very few stroke cases, techniques such as SMOTE (Synthetic Minority Oversampling Technique) were considered to balance the classes.

5. Exploratory Data Analysis (EDA)

The target variable distribution showed a significant imbalance, with the majority of patients not experiencing a stroke. The correlation matrix revealed no strong linear correlations among the numerical features. Analysis of numerical features provided key insights. Patients who experienced a stroke were found to be older on average than those who did not. The average glucose level was right-skewed and contained many outliers, with stroke patients generally having higher median values. BMI did not show a strong correlation with stroke occurrence, as median values were similar across both groups, though the presence of many outliers was noted.

Categorical feature analysis also revealed important patterns. Females were slightly more likely than males to suffer a stroke. Interestingly, a large proportion of stroke patients did not have hypertension or heart disease, suggesting that these conditions, while important, are not the only drivers of stroke. Most stroke patients were married, and those working in the private sector appeared more likely to suffer from stroke than individuals in government jobs or those who had never worked. Furthermore, more stroke cases were reported among urban residents compared to rural ones. An unexpected observation was that many stroke

patients had never smoked, which indicated that smoking status might not be a strong predictor of stroke in this dataset.

6. Data Preparation for Modeling

In preparing the dataset for model training, outliers were either removed or transformed to reduce their influence. Features were normalized and scaled to ensure consistency across variables, particularly those measured on different scales. The dataset was then split into training and testing sets to evaluate model performance fairly. Handling the imbalance in the target variable was a critical step, as it directly impacted the model's ability to correctly identify stroke cases.

7. Model Building

Several machine learning models were implemented, including a Decision Tree classifier. The dataset was divided into training and test sets, and balancing techniques such as oversampling were applied to improve the model's ability to detect minority class cases. Hyperparameter tuning was also performed to enhance model accuracy and efficiency. The Decision Tree was chosen for its interpretability, allowing healthcare professionals to understand the rules and thresholds used to predict stroke risk. Other models such as Logistic Regression, Random Forest, or Support Vector Machines may also be considered for comparison.

8. Results and Evaluation

The models were evaluated using several metrics, including accuracy, precision, recall, F1-score, and the ROC-AUC curve. Accuracy alone was not considered sufficient due to the imbalance in the dataset. Precision and recall provided a better understanding of how well the model was performing on the minority class. The Decision Tree classifier demonstrated interpretable results, although precision-recall trade-offs highlighted the challenges of imbalanced classification. The ROC-AUC curve was used as the most reliable performance measure, as it accounted for the sensitivity and specificity of the model.

9. Conclusion

The analysis indicated that age and average glucose level were strong predictors of stroke risk. Categorical features such as work type, marital status, and residence type also showed noticeable influence. On the other hand, factors like hypertension, heart disease, and smoking status, while traditionally associated with stroke, were not the strongest predictors in this dataset. This highlights the complexity of stroke prediction and the importance of analyzing multiple features together rather than relying on a single health indicator.

10. Future Work

Future improvements to this study could include the use of ensemble methods such as Random Forests and Gradient Boosting to improve prediction accuracy. Deep learning models could also be explored to capture more complex patterns within the dataset. Additionally, access to larger and more balanced datasets would improve the generalizability of the models. Integrating these findings into a predictive healthcare system could provide real-time assistance to doctors in assessing stroke risk, ultimately contributing to better healthcare outcomes.