PREDICTING STROKE RISK: A MULTI-ALGORITHM APPROACH USING MACHINE LEARNING

Satya Kiran Kota **Akshit Sharma**
Sai Srinivas Munagala **Nitish Kumar Pineeti**

Abstract

Stroke, a condition characterized by burst blood vessels in the brain, causes brain damage owing to an interruption in blood supply. The World Health Organization (WHO) identifies it as the leading cause of mortality and disability worldwide (Organization, 2020) Understanding how strokes happen is critical for prompt intervention. This project trains models using biomedical data from the Stroke Prediction dataset using machine learning methods. The project emphasizes addressing class imbalance using Sampling techniques. Rigorous model comparisons will validate the suggested approach and provide insights for clinical use and future studies.

1. Introduction

Stroke remains a pressing concern globally, with significant mortality and disability rates, as highlighted by the Centers for Disease Control and Prevention in 2023 (for Disease Control & Prevention, 2023). In India alone, strokes are the fourth leading cause of death, further emphasizing the urgency of early detection and intervention strategies. Disruptions in blood flow to the brain can result in severe consequences, making it imperative to identify warning signs promptly. However, early detection and treatment are crucial for minimizing damage and improving outcomes. This is where Machine Learning comes in.

This project focuses on developing a precise and accurate stroke prediction system using machine learning algorithms. It integrates physiological data from various sources to conduct a comprehensive analysis. Utilizing algorithms such as Logistic Regression, Decision Tree, and Voting Classifier, the model addresses class imbalance within stroke prediction datasets. By exploiting comprehensive feature selection techniques and rigorous evaluation metrics, the aim is to accurately identify stroke cases while maintaining clinical accountability.

In stroke prediction using machine learning, imbalanced datasets present significant challenges. The scarcity of stroke cases compared to non-stroke instances can skew model training, leading to biased predictions favoring the majority class. Addressing this class imbalance necessitates cautious data sampling techniques like oversampling (Mohammed et al., 2020) to avoid overfitting or information loss. Furthermore, traditional evaluation metrics like accuracy may not accurately reflect model performance. Metrics such as precision, recall, and F1-score become crucial for assessing model effectiveness in capturing stroke cases amidst class imbalance.

2. Procedure and Methodology

This section includes a dataset description, a flow diagram, and details regarding the study's process and methodology being employed.

2.1. Proposed System

Data is processed to enable model construction, utilizing preprocessed datasets and various machine learning techniques such as Logistic Regression, Decision Tree classification, and Voting Classifier. After all the alternative models are created and evaluated, a comparison between them will be made using accuracy, precision, recall, and F1 scores.

2.2. Dataset

The stroke prediction dataset (Federico, 2021) going to be used, comprises 5110 rows and 12 columns, with the output column *stroke* containing binary values (1 or 0) indicating stroke risk presence. Notably, instances of 0 outnumber instances of 1 in this dataset, with 4861 rows indicating no stroke risk and 249 rows indicating stroke risk.

From Figure 1, it is clear that this dataset is imbalanced. To tackle this class imbalance and improve accuracy, we are considering employing data preprocessing techniques, with Random Oversampling and SMOTE techniques being potential options.

2.3. Preprocessing

Before model construction, data preprocessing is essential to eliminate noise and outliers that could impede the

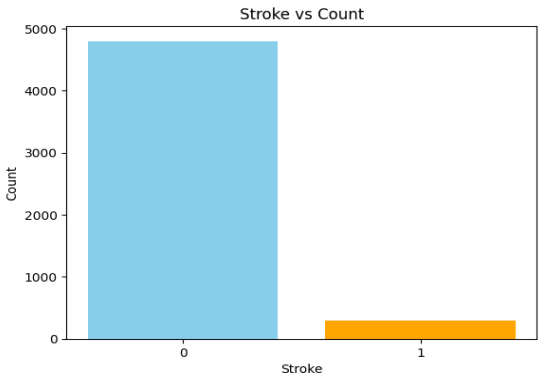

Figure 1. Data Distribution in the original Dataset

model's intended training process. After gathering the relevant dataset, it undergoes cleaning and preparation for model development. These preprocessing steps ensure that the dataset is cleaned, standardized, and balanced, thus preparing it for the subsequent stages of model development and evaluation.

In this dataset, with 12 characteristics initially, the column *id* is to be excluded as it does not contribute to model construction. Subsequently, the dataset is examined for null values, which are then filled if detected. In the case of the *BMI* column, null values are to be imputed using the mean of the column's data. Failure to address these imbalances can lead to inaccurate findings and ineffective forecasts. Therefore, dealing with this imbalance is paramount for obtaining an efficient model, achieved through the utilization of some Sampling techniques.

Following the data preparation and handling of the imbalanced dataset, the subsequent step involves model construction. To enhance accuracy and efficiency, the data is divided into training and testing sets at an *80:20* ratio. Subsequently, the model undergoes training utilizing various classification methods, including decision tree classification, voting classifier, and logistic regression.

3. Intended Experiments

We planned to perform experiments using the following algorithms:

3.1. Decision Trees

Decision Trees (DT) (Quinlan, 1986) are hierarchical structures that mimic a tree with decision nodes and leaf nodes, and they divide data according to particular attributes. They imitate the decision-making process that humans go through and are intuitive and simple to understand. DTs are helpful for some problem domains because they may be used for both regression and classification tasks, and they require less data preprocessing than other methods.

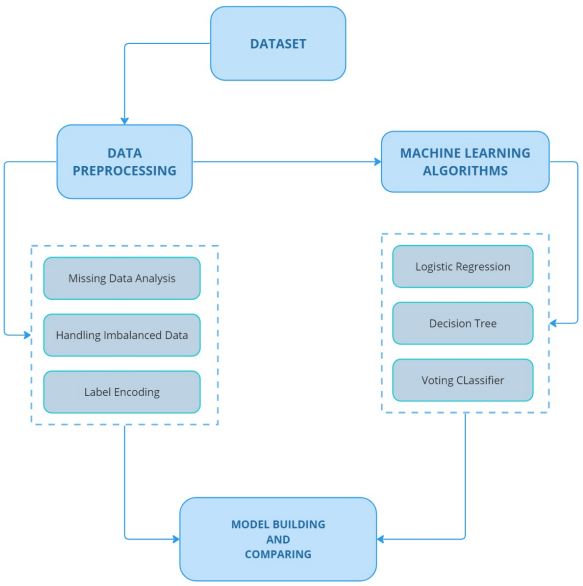

Figure 2. Graphical Demonstration of our Proposal

3.2. Logistic Regression

Logistic Regression (LR) is a widely used algorithm for binary classification tasks. It models the probability of a binary outcome based on one or more predictor variables. LR predicts the probability of the dependent variable belonging to a particular category, typically between 0 and 1, using a logistic function. It is suitable for problems where the relationship between independent variables and the categorical outcome needs to be understood and modeled.

3.3. Voting Classifier

A Voting Classifier combines the predictions of multiple individual models to generate a final prediction. It can operate in two modes:

3.3.1. Soft Voting

In soft voting (Wang et al., 2013), the predicted probabilities from each model are averaged, and the class with the highest average probability is selected. This approach considers the confidence levels of each model's predictions, providing a weighted average for the final decision.

3.3.2. Hard Voting

In hard voting (Habib & Tasnim, 2020), the mode value of the predicted classes is chosen as the final output. This method disregards probability values and focuses solely on the most commonly predicted class by the individual models. Hard voting is straightforward and effective when models produce diverse predictions.

References

Dritsas, E. and Trigka, M. Stroke risk prediction with machine learning techniques. *Sensors (Basel)*, 22(13):4670, 2022. doi: 10.3390/s22134670.

Federico, S. Stroke prediction dataset. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, 2021.

for Disease Control, C. and Prevention. Stroke facts. https://www.cdc.gov/stroke/facts.htm, 2023.

Habib, A.-Z. S. B. and Tasnim, T. An ensemble hard voting model for cardiovascular disease prediction. In *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–6. IEEE, 2020.

Mohammed, R., Rawashdeh, J., and Abdullah, M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. Unknown, 2020. doi: 10.1109/ICICS49469.2020.239556. URL https://ieeexplore.ieee.org/document/9078901. Accessed: 2024-02-23.

Organization, W. H. The top 10 causes of death, 2020. URL https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed: 2024-02-23.

Quinlan, J. R. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

Wang, H., Yang, Y., Wang, H., and Chen, D. Soft-voting clustering ensemble. In Zhou, Z.-H., Roli, F., and Kittler, J. (eds.), *Multiple Classifier Systems*, pp. 307–318, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-38067-9.