Stroke Analysis and Prediction

Harshitha Chollangi

December 2023

1 Abstract

Stroke is a major global health concern as well as a leading cause of mortality and morbidity. Strokes, according to the World Health Organization (WHO), cause a significant number of deaths and disabilities worldwide. As a result, there is an increasing demand for effective predictive analytics to aid in the early detection and prevention of strokes. This paper presents a detailed analysis and prediction of strokes using machine learning models and a dataset of 5110 observations with 12 key attributes. These characteristics include vital patient information such as gender, age, diseases, and smoking status. We hope to uncover patterns, risk factors, and potential predictors of strokes using advanced machine learning techniques, providing valuable insights for healthcare professionals and policymakers.

2 Introduction

Stroke is one of the leading causes of death and disability in the world. The purpose of this paper is to develop a reliable stroke prediction model that can be used to aid in clinical decision-making. The dataset contains 5110 observations with 12 attributes capturing critical patient information such as gender, age, diseases, and smoking status. The goal is to predict stroke probability based on these input parameters. Linear Models, Random Forest, Logistic Regression, Naive Bayes Classification, K-Nearest Neighbor Classification, Linear SVM, Simple Decision Tree Modelling, and XGBoost have all been used to predict strokes. Early detection of warning signs is critical in stroke prevention, as are lifestyle factors such as smoking, drinking, BMI, average glucose level, and maintaining heart and kidney health.

3 System Methodology

The datasets were taken from Kaggle. Following data collection, the next step is data preprocessing, which includes dealing with missing values, managing imbalanced data, and performing label encoding specific to this dataset. The preprocessed dataset, as well as machine learning algorithms, are required for model construction. Accuracy metrics such as Accuracy Score, Precision Score, Recall Score, F1 Score, and Receiver Operating Characteristic (ROC) curve are used to compare them. The model comparison identifies the best model in terms of accuracy metrics to move forward with the deployment phase. During the deployment phase, a Shiny app is used to collect data from users, and the deployed model predicts the likelihood of a stroke based on the data collected.

The flow chart of the proposed system's methodology is shown in Fig. 1.

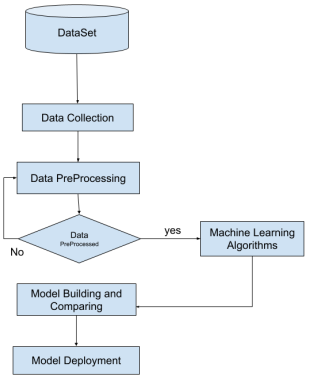

Figure 1: System Methodology

3.1 Dataset

The stroke prediction dataset, obtained from Kaggle [2], consists of 5110 rows and 11 columns, including 'gender', 'age', 'hypertension', 'heart_disease', 'ever_married', 'work_type', 'Residence_type', 'avg_glucose_level', 'bmi', 'smoking_status', and 'stroke'. The 'stroke' column, serving as the target variable, holds binary values '1' or '0', signifying potential stroke risk or no stroke risk, respectively. Data preprocessing is conducted to address imbalances and enhance accuracy.

The summarized stroke dataset is presented in Table 1.

Table 1: Stroke Dataset

Attribute Name	**Type (Values)**	**Description**
gender	String literal (Male, Female, Other)	Gender of the patient
age	Integer	Age of the patient
hypertension	Integer (1, 0)	Presence of hypertension
heart_disease	Integer (1, 0)	Presence of heart disease
ever_married	String literal (Yes, No)	Marital status of the patient
work_type	String literal	Category of work
Residence_type	String literal (Urban, Rural)	Patient's residence type
avg_glucose_level	Floating-point number	Average glucose level in blood
bmi	Floating-point number	Body Mass Index (BMI) of the patient
smoking_status	String literal	Smoking status of the patient
stroke	Integer (1, 0)	Stroke status (1: Risk, 0: No Risk)

3.2 Data Preprocessing

Data preprocessing is a crucial step before model building to eliminate unwanted noise and outliers from the dataset, ensuring proper training and efficient model performance. The dataset, as mentioned in Table I, contains 12 attributes. To prepare the data for model building, the following preprocessing steps are applied:

- The 'id' column is dropped as its presence does not significantly impact model building.

- Null values in the dataset are checked and filled if any are found. In this case, the 'bmi' column's null values are filled with the mean of the column data.
- After removing null values, the next task is label encoding to handle categorical variables.

These steps ensure that the dataset is cleaned and ready for further model building and analysis. The cleaned dataset is then used for training the machine learning models.

3.2.1 Correlation Analysis

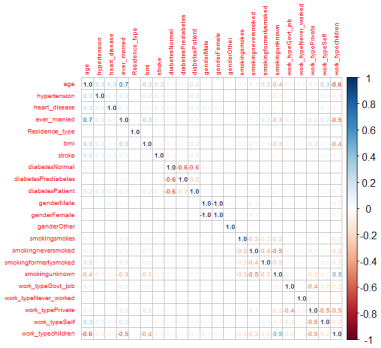

Figure 2: Correlation Matrix

The correlation matrix above illustrates the relationships between different attributes, highlighting the first five with the highest correlation to the 'Stroke' column: age, hypertension, heart_disease, ever_married, and residential_type.

3.3 Model Building

3.3.1 Splitting the Data

After preprocessing and addressing imbalanced data, the dataset is split into 80

3.3.2 Linear Models

Multilinear, polynomial, and improved linear models are explored for stroke prediction:

- **Multilinear Model:** Extends simple linear regression to multiple variables.
- **Polynomial Model:** Introduces polynomial terms for non-linear relationships.
- **Improved Linear Models:** Includes regularization techniques like L1, L2, Elastic Net, Ridge, and Lasso Regression.

These models provide interpretability and capture complex relationships, compared during evaluation with other algorithms.

3.3.3 Classification Algorithms

Various algorithms are employed for model training and stroke risk prediction:

- **Logistic Regression:** Linear model for binary classification.
- **K-Nearest Neighbors (KNN):** Non-parametric, lazy learning algorithm.
- **Support Vector Machine (SVM):** Powerful algorithm for linear and non-linear tasks.
- **Naïve Bayes:** Probabilistic algorithm effective for text classification.
- **Decision Tree:** Tree-like model with interpretability.
- **Random Forest:** Ensemble learning method with multiple decision trees.
- **XGBoost:** Optimized gradient boosting algorithm.

Model evaluation metrics in the table provide a comprehensive view of each model's performance.

4 Detailed Result and Model Comparison

4.1 Age and Stroke Risk

Stroke percentage increases with age, peaking near the end of the 70s. No male strokes below age 40; several females experience strokes before 40, including cases around 30, 15, and below 1 year old.

4.2 Gender Disparities

Hypertension slightly affects females more than males. Married females have lower stroke risk than unmarried individuals. Heart disease impact is similar for both genders.

4.3 Occupation and Stroke Risk

Self-employed individuals have a higher stroke percentage than private and governmental sectors, suggesting lower job security impact.

4.4 Useful Classification

Gender-based classification is particularly useful for understanding stroke risk factors. Observations on age, hypertension, marital status, and occupation provide insights into potential predictors of stroke.

4.5 Model Comparison

Based on the table, Linear Models (multilinear, polynomial, improved LM), Logistic Regression, SVM, and XGBoost outperform others in accuracy, precision, recall, and F1 score.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9559902	0.375	0.05769231	0.1
KNN	0.9323553	0.12195122	0.09615385	0.10752688
SVM	0.9576202	0.5	0.01923077	0.03703704
Naive Bayes	0.5607172	0.07652174	0.84615385	0.14035088
Decision Tree	0.9576202	0.95762021	1	0.97835137
Random Forest	0.9437653	0.16	0.07692308	0.1038961
XGBoost	0.9568052	0.95758564	0.99914894	0.97792586

Table 2: Model Evaluation Metrics

5 Discussion

5.1 Future Work

5.1.1 Dataset Expansion

Expanding the dataset to include more demographic, health, and lifestyle data for enhanced accuracy and generalizability.

5.1.2 Age Analysis Refinement

Investigating specific age ranges, creating age brackets, or exploring nonlinear relationships for improved precision.

5.1.3 Lifestyle Factors Exploration

Incorporating additional lifestyle factors like diet, physical activity, and stress levels to enhance predictive accuracy.

5.2 Limitations

5.2.1 Dataset Constraints

Imbalanced dataset with fewer stroke cases may impact prediction accuracy. Addressing this imbalance and acquiring a more representative dataset could improve model performance.

5.2.2 Age Analysis Complexity

Further complexity in age analysis, exploring nonlinear relationships and interactions with other risk factors for a nuanced understanding of age-related stroke risk.

5.2.3 Lack of Lifestyle Data

Enhancing the model with more comprehensive lifestyle data beyond basic factors could improve predictive capabilities.

6 Conclusion

While the current work provides valuable insights into stroke prediction, future research opportunities exist to enhance accuracy and broaden the model's scope. Addressing dataset lim-

itations, refining age analysis, and exploring additional lifestyle factors could contribute to the development of a more robust and effective stroke prediction model.

6.1 Shiny App Link

To access the Shiny app, visit: Shiny App Link

References

[1] Qasim Alhammad, *Build and Deploy a Stroke Prediction Model Using R*, Kaggle, https://www.kaggle.com/code/qasimalhammad/build-and-deploy-a-stroke-prediction-model-using-r/report

[2] Fernando Sorian, *Stroke Prediction Dataset*, Kaggle, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[3] Author(s), *Title of the Article*, International Journal of Advanced Computer Science and Applications (IJACSA), DOI: http://dx.doi.org/10.14569/IJACSA.2021.0120662