### **STROKE PREDICTION : A MACHINE LEARNING REGRESSION APPROACH**

- By Kishore K(220701134)

### **ABSTRACT**

Stroke is a major medical condition and a significant contributor to global morbidity and mortality rates. Precise and prompt prediction of the risk of stroke is crucial for preventive medicine and early clinical intervention. The present study suggests a Logistic Regression-based predictive model for the detection of stroke, a commonly applied statistical approach ideal for binary classification problems. The research employs a publicly accessible healthcare dataset with features like age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), and smoking status. Thorough data preprocessing methods, such as missing value handling, encoding categorical variables, and feature scaling, were used to make the model reliable. The Logistic Regression model was tuned and tested on classification metrics including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The experimental results show that the model performs well in predicting stroke risk, promising its applicability as a tool to be integrated into clinical decision support systems. This study emphasizes the potential of machine learning methods, especially interpretable models such as Logistic Regression, in accelerating early stroke risk assessment and encouraging proactive healthcare provision.

**INTRODUCTION**

Stroke is a potentially fatal medical emergency that results from either the interruption or decrease of blood flow to part of the brain, resulting in a lack of brain tissue oxygen and nutrients. It is among the most common causes of death and long-term disability globally. Based on the World Health Organization (WHO), millions of individuals endure strokes every year, with a large percentage having enduring impairments. Early detection and preventive measures are therefore crucial in minimizing the risk and severity of stroke complications.

Over the last few years, machine learning has become a highly effective tool in the medical industry, providing novel means to forecast illnesses through the learning process using patient data. Predictive algorithms can aid physicians by classifying high-risk subjects prior to symptoms occurring, thus allowing early intervention by physicians. Among several machine learning models, Logistic Regression is notable due to its simplicity, interpretability, and accuracy for binary classification problems—thus especially being an apt choice for medical diagnosis procedures like stroke prediction.

### This study is geared towards building a stroke prediction model based on Logistic Regression. The model is learned on a healthcare dataset with demographic and clinical attributes such as age, hypertension, heart disease, average glucose level, and body mass index (BMI). Via strenuous data preprocessing and performance assessment, the study seeks to determine whether the model can effectively classify subjects as at-risk or not at-risk for stroke.

### The long-term objective of this research is to illustrate the capability of interpretable machine learning techniques in improving clinical decision-making and enabling proactive medicine. The system in question may be a worthwhile asset for hospitals and telemedicine platforms, assisting with early diagnosis and stroke event prevention.

### **LITERATURE REVIEW**

### The use of machine learning in medical diagnosis has made significant progress over the past ten years, especially in the early detection and prevention of life-threatening diseases such as stroke. There have been several studies investigating different algorithms and methods for enhancing the accuracy and reliability of stroke prediction systems.

### Kumar et al. (2020) also constructed a prediction model for stroke based on Decision Tree and Random Forest algorithms. It was observed that their results indicated that ensemble methods were more accurate in prediction compared to individual classifiers. Nonetheless, they reported a compromise between model simplicity and interpretability, which is important in clinical usage.

### Chaurasia and Pal (2014) applied Naïve Bayes and Decision Tree classifiers to a heart disease dataset and highlighted the viability of utilizing minimal input features in successful disease prediction. Although their model was successful, it was not generalizable because of low feature diversity.

### Ali et al. (2019) compared Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Logistic Regression for stroke classification. Their findings indicated that Logistic Regression, being less complex, provided competitive accuracy and had good interpretability, which made it appropriate for use by healthcare professionals.

### Saini et al. (2021) utilized deep learning methods for stroke prediction with real-time data from wearable sensors. While neural networks exhibited encouraging performance, they needed big datasets and significant computational power, which restricts deployment in low-resource environments.

### Tiwari and Gupta (2022) suggested an ensemble of Gradient Boosting and Logistic Regression for stroke prediction on the Kaggle stroke dataset. Their hybrid model enhanced prediction accuracy at the cost of higher system complexity, which became a concern regarding real-time use.

### We can see from the literature examined that although intricate models have high accuracy, more straightforward models such as Logistic Regression provide efficiency, performance, and interpretability—important considerations in clinical decision-making scenarios. This research thus endeavors to capitalize on these findings by applying and assessing a logistic regression model, which has been specifically designed for stroke prediction based on structured healthcare data.

### **METHODOLOGY**

**Dataset Description**

The dataset used for this project was sourced from the public Kaggle Stroke Prediction Dataset, which contains anonymized health-related records of individuals. It includes both categorical and numerical features relevant to stroke risk such as gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status. The target variable is stroke, a binary indicator (0 = No Stroke, 1 = Stroke), which makes this a binary classification problem.

**Data Preprocessing**

To ensure high-quality input for the model, the dataset underwent several preprocessing steps:

* Handling Missing Values: Instances with missing BMI values were imputed using the mean of the column to maintain dataset integrity.
* Encoding Categorical Variables: Categorical features such as gender, ever\_married, work\_type, Residence\_type, and smoking\_status were transformed using one-hot encoding to make them compatible with machine learning algorithms.
* Feature Scaling: Numerical features like age, average glucose level, and BMI were scaled using StandardScaler to normalize the feature values to a standard distribution (mean = 0, std = 1). This was particularly important for ensuring optimal performance of algorithms like Logistic Regression.
* Balancing the Dataset: Given the dataset’s imbalance (fewer stroke cases than non-stroke), SMOTE (Synthetic Minority Oversampling Technique) was applied to oversample the minority class and reduce prediction bias.

**Model Selection and Training**

The Logistic Regression algorithm was chosen as the core model due to its effectiveness in binary classification tasks and its interpretability in medical domains. The dataset was split using an 80-20 train-test ratio to provide sufficient data for model learning while preserving a subset for unbiased evaluation.

Model hyperparameters, such as the regularization strength (C) and solver (liblinear), were tuned using Grid Search with Cross-Validation to optimize the trade-off between bias and variance.

**Evaluation Metrics**

To evaluate the effectiveness of the model in predicting stroke risk, several key classification metrics were employed:

* **Accuracy** – Measures the percentage of correct predictions over total predictions.
* **Precision** – Indicates how many predicted stroke cases were actually correct.
* **Recall (Sensitivity)** – Reflects the model's ability to detect actual stroke cases.
* **F1-Score** – Combines precision and recall into a single metric to assess overall performance.
* **ROC-AUC Score** – Evaluates the model's ability to distinguish between positive and negative classes at various thresholds.

Screens screenshot of a computer

AI-generated content may be incorrect.

### **EXPERIMENTAL ANALYSES**

To evaluate the accuracy and reliability of the stroke prediction system, the dataset was split into training and test sets using an 80:20 ratio. Prior to training, feature normalization was applied using StandardScaler to ensure that all clinical features contributed equally. The primary model used for classification was Logistic Regression, chosen for its interpretability and effectiveness on binary outcomes.

The model was trained on the normalized training data, and predictions were made on the test set. Evaluation was conducted using standard classification metrics such as Accuracy, Precision, Recall, and F1-Score. Logistic Regression showed promising results, particularly in balancing precision and recall, which is crucial in healthcare-related applications to avoid false negatives.

A table with numbers and text

AI-generated content may be incorrect.

Results for Model Evaluation:

The results show that Logistic regression performs the best with the highest accuracy, making it the model of choice for predicting stroke.

**VISUALIZATIONS**

 **Correlation Heatmap**: A heatmap was created to analyze the relationship between features such as age, hypertension, heart disease, BMI, and average glucose level with the stroke variable. It showed a strong positive correlation between stroke occurrence and factors like age and hypertension.

A chart with red and blue squares

AI-generated content may be incorrect.

 **ROC Curve**: The Receiver Operating Characteristic (ROC) curve for the logistic regression model exhibited a good area under the curve (AUC), indicating strong discriminative power.

A graph with a line

AI-generated content may be incorrect.

 **Confusion Matrix**: A confusion matrix highlighted the model’s performance in correctly classifying stroke and non-stroke cases, revealing high specificity and acceptable sensitivity.

A graph with numbers and a blue square

AI-generated content may be incorrect.

 **Feature Importance Plot**: Coefficients from the logistic regression model were plotted to show the most influential features, with age, hypertension, and average glucose level among the top predictors.

A graph with a bar graph

AI-generated content may be incorrect.

### **CONCLUSION**

In conclusion, the gold price prediction models explored in this study demonstrated the viability of machine learning techniques in forecasting commodity prices. Among the evaluated models, Support Vector Regression (SVR) and Random Forest emerged as strong performers, with SVR offering better consistency and Random Forest showcasing robustness in handling non-linear patterns. The use of data augmentation through Gaussian noise proved beneficial in enhancing model generalization and reducing overfitting, particularly in high-variance models.

Despite promising results, the models exhibited some limitations in capturing extreme fluctuations and rare market movements. This highlights the need for incorporating more diverse and dynamic features that influence gold prices, such as macroeconomic indicators, global financial news sentiment, interest rates, and geopolitical developments.

### 

### **REFERENCES**

[1] World Health Organization, “Stroke Fact Sheet,” 2023.  
[2] B. Johnson et al., “Machine Learning Approaches for Stroke Prediction: A Review,” Journal of Biomedical Informatics, vol. 122, 2021.  
[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," KDD, 2016.  
[4] M. G. Tsai and C. H. Lin, “Logistic Regression in Healthcare Predictive Modeling,” Health Informatics Journal, vol. 25, no. 3, 2019.  
[5] A. R. Smith et al., “Data-Driven Clinical Prediction Models for Stroke,” Journal of Stroke and Cerebrovascular Diseases, vol. 30, no. 2, 2021.  
[6] S. Sharma and D. Jain, “Prediction of Stroke Using Machine Learning Techniques,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 4, 2020.