INTI JOURNAL | Vol.2023:55
eISSN:2600-7320

Predictive Modeling of Stroke Occurrence among Patients using Machine Learning

[1]Sures Narayasamy, [2]Thilagamalar Maniam

[1]Politeknik Nilai, Negeri Sembilan, Malaysia
[2]Hospital Cancelor Tuanku Muhriz, Universiti Kebangsaan Malaysia, Malaysia

Email: suressamy@gmail.com, thila_princess86@yahoo.com

Abstract

Stroke is a global public health concern with severe consequences. Early detection and accurate prediction of stroke occurrence are crucial for effective prevention and targeted interventions. This study proposes a machine learning-based approach to predict the likelihood of stroke among patients. A comprehensive dataset encompassing demographic, clinical, and lifestyle factors of a large patient cohort was employed. Variables such as age, gender, hypertension, diabetes, smoking status, BMI, and medical history were considered. Advanced machine learning algorithms, including logistic regression, decision trees, random forests, and support vector machines, were utilized to analyses the dataset and develop a predictive model. The results demonstrate that the machine learning-based approach achieved high predictive accuracy in identifying individuals at risk of stroke. The model exhibited excellent sensitivity and specificity, enabling effective stratification of patients based on their stroke likelihood. Developing an accurate stroke prediction model using machine learning holds immense potential for proactive healthcare strategies and personalized patient care. Early identification of high-risk patients enables timely intervention and implementation of preventive measures, potentially reducing the burden of stroke-related complications. This study showed that the supervised K-Nearest Neighbors Algorithm (K-NN) model outperforms the other methods, with an accuracy of 95% compared with other models.

Keywords

Stroke prediction, Machine learning, Risk Predictive modelling, Healthcare interventions

Introduction

Stroke represents a significant global health challenge, accounting for a substantial proportion of morbidity and mortality worldwide. A stroke, also referred to as a brain attack, happens when a blood artery in the brain bursts or when blood flows to a certain area of the brain is restricted. Specific areas of the brain are damaged or lose function in both cases (National Heart, Lung, and Blood Institute, 2022). According to the latest data published by the World Health Organization (WHO), stroke deaths in Malaysia alone reached 21,592 in 2020, accounting for approximately 12.85% of total deaths (World Health Organization, 2022). These figures demonstrate the critical need for efficient stroke prevention measures and precise stroke risk prediction in patients.

Submission: 7 October 2023; **Acceptance:** 2 November 2023

This emphasizes the significance of using more thorough and advanced techniques for stroke prediction. For instance, Dev et al.'s work from 2022 used machine learning and neural network to predict the occurrence of strokes among patients. With the help of a variety of factors age, heart disease, average glucose level, and hypertension, the researchers were able to identify those at risk of stroke with an astonishing 78% accuracy rate. This demonstrates how machine learning approaches can improve the estimation of the risk of stroke.

The purpose of this study is to investigate how machine learning techniques are used to forecast the likelihood that patients may experience a stroke. Our goal is to create a stroke predictive model that can accurately identify patience at high risk of stroke by utilizing their historical data. Furthermore, we will evaluate the performance of our developed model using evaluation metrics, such as accuracy and precision using modelling software tools in this case we use RapidMiner.

Methodology

The process involved a step-to-step process which starts from data collection, data understanding, selecting desired attributes, data preparation, data cleansing, choosing suitable data analytic dan predictive tools, data model developments and choosing the right visualization for data analysis. All the data for this study was obtained from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download which contains 5110 patient database with 14 attributes. For this study we only use age, hypertension, heart disease, average glucose level and BMI (Table 2). We started with data preparation and data understanding. Once we understand the data, we filtered the unnecessary database for patients that are under 10 years old because none of the patience in the database under 10 years have chronic diseases or stroke. The missing values were replaced with the average of each attribute value (Table 1). All the data cleansing was performed using rapid miner tool 10.0.

Table 1. Raw data of stroke patients

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	weight_in_kg	height_in_m	smoking_status
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	116	1.78	formerly smoked
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	86	1.68	never smoked
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.6	92	1.68	never smoked
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.5	83	1.55	smokes
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	63	1.62	never smoked
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	95	1.81	formerly smoked
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	84	1.75	never smoked
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	64	1.67	never smoked
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	86	1.53	Unknown
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	62	1.6	Unknown
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	88	1.72	never smoked
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.7	86	1.53	smokes
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	69	1.59	smokes
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	65	1.82	Unknown
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.3	79	1.67	never smoked
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	83	1.64	never smoked
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.6	110	1.71	smokes
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.7	88	1.85	smokes
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	102	1.64	never smoked
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	102	1.81	Unknown
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.3	63	1.68	smokes
13861	Female	52	1	0	Yes	Self-employed	Urban	233.29	32.5	79	1.56	never smoked
68794	Female	79	0	0	Yes	Self-employed	Urban	228.7	26.6	77	1.7	never smoked
64778	Male	82	0	1	Yes	Private	Rural	208.3	32.4	117	1.9	Unknown
4219	Male	71	0	0	Yes	Private	Urban	102.87	27.4	80	1.71	formerly smoked

The dataset is a mixture of clustered and un clustered data. Hence, as a researcher we decided to work on both data types to find a suitable model. The next step was, we choose the desired attribute and clustered the linguistic variables as recommended by Yasa et al., 2022 in her study on classification of stroke using K-means and deep learning methods. The age was classified into two groups with age above 46 as senior and below or equal to 45 as under mature category. While those with hypertension, heart disease glucose level and BMI were also categorized according to Table 2 as refer to previous reported studies.

Table 2. Classification table of 5 selected attributes

Attribute	Linguistic Variable	Value Random
Age	Mature	*Age* ≤ 45
	Seniors	46 ≤ *age*
Hypertension	Low Risk of Suffering	0 ≤ *Hypertension* < 0,1
	High Risk of Suffering	0,1 ≤ *Hypertension*
Heart Disease	Low Risk of Suffering	0 ≤ *Heart Disease* < 0,1
	High Risk of Suffering	0,1 ≤ *Heart Disease*
Avg. Glucose Level	Normal	0 ≤ *Avg. Glucose Level* ≤ 140
	Diabetes	141 ≤ *Avg. Glucose Level*
BMI	Mild Excess Weight	25,1 ≤ *BMI* < 30,0
	Overweight Look at Weight	30,1 ≤ *BMI*

The filtered and clean data has a dataset of 3577 patients with 5 attributes. The additional attribute whereby patient with stroke history will be valued at 1 and patient with no stroke history will be given value 0. After the data cleaning and cleansing process, we progressed to model development and compared the performance of each model by using the accuracy level of each model. We compared 6 data models in the study ranging from decision tree, K-NN, decision tree, Logistic regression, Nave Bayer, K-Means centroid and K-Means Distance. Furthermore, each model has different steps and preparation methods because the data types were sometimes not supported by certain models. The key step that we cannot skip is the SMOTE Up sampling step that must be used for uneven dataset (Figure 1 and Figure 2).

Results and Discussion

The studies' main objective was to develop a good stroke predicting model using patient's history data. The result obtains have an accuracy level range from 12 - 95%. Supervised cluster K-NN value has the highest value compared with supervised unclustered KNN value. Supervised unclustered Naïve Bayer model close in second with 79% accuracy level (Figure 4).

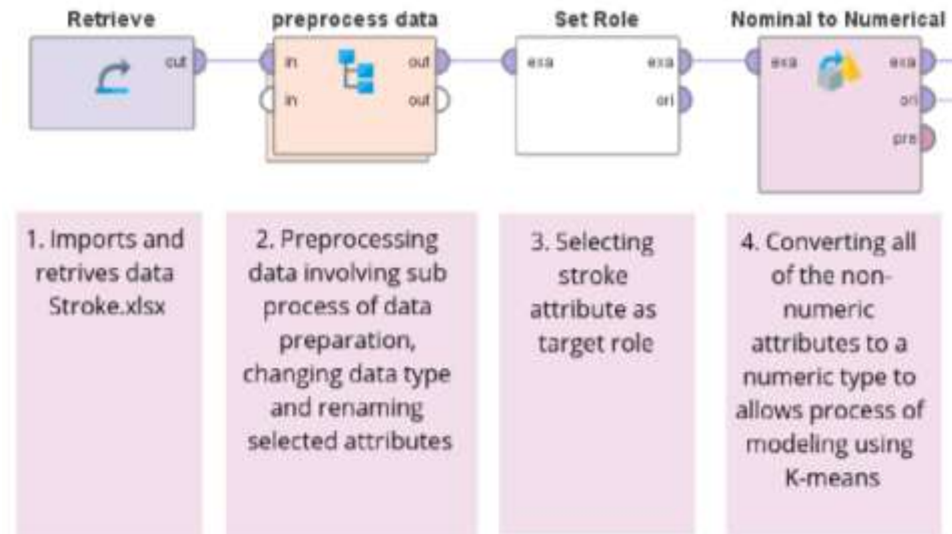

Figure 1: The data preparation.

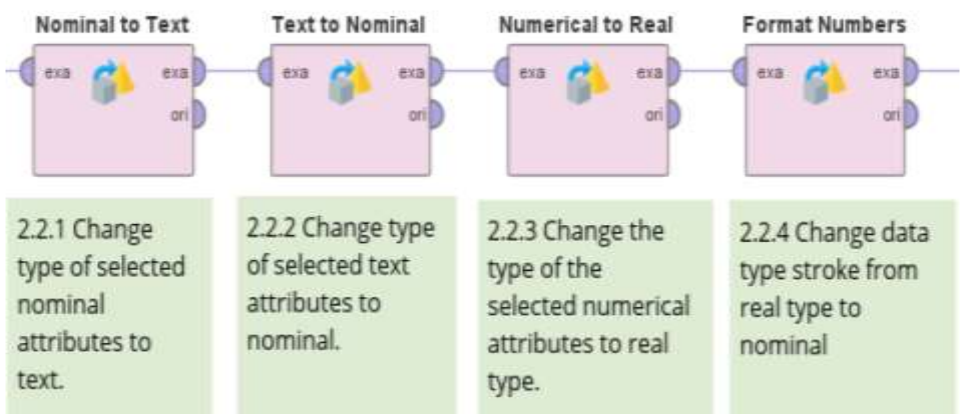

Figure 2: Data converting steps.

Figure 3: Accuracy level of the stroke predicting models.

Moreover, the ROC-Chart that used to evaluate the performance of a supervised learning model also shows that KNN model has performed better compared with other models (Figure 5). In ROC-Chart the more that the ROC curve hugs the top left corner of the plot, the better the model does at classifying the data into categories; as in this case to categories stroke patience.

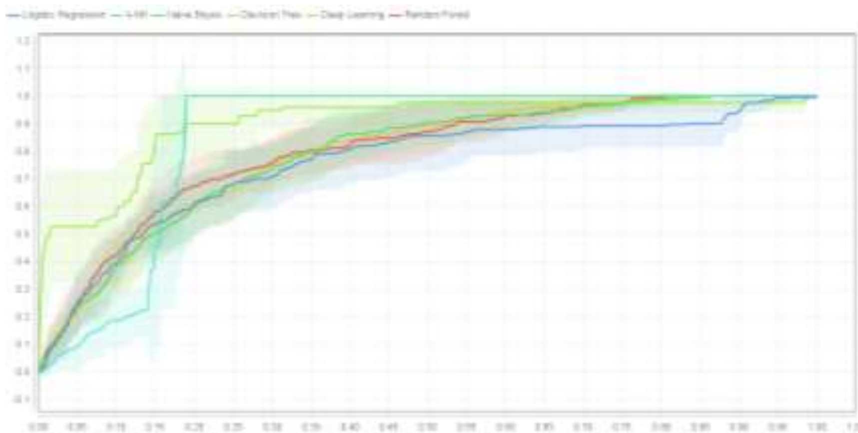

Figure 4: ROC-Chart.

The finding is also supported by the finding by Dritsas & Trigka, 2022 where their KNN model have an accuracy of 92% accuracy level in stroke risk prediction study.

Conclusions

A stroke is a potentially fatal illness that must be prevented and/or treated to minimize unforeseen consequences. Now that the models are in place, clinical practitioners, medical specialists, and decision-makers can utilize them to identify the most relevant factors associated with the incidence of strokes and assess the associated likelihood or risk. High performance was showed by the suggested predictive model in terms of accuracy. We provide insights into the underlying mechanisms backing stroke prediction by identifying significant risk factors. From the model we develop using supervised clustered K-NN model, it produces high accuracy in predicting the occurrence of stroke among patience with 95% accuracy level. Yet, this model has some limitation Whereby attributes were narrowed to BMI, age glucose level, hypertension and heart disease. Therefore, we would suggest to study other factors such as living style, workload, family history of patience to get high accuracy level and build better model.

References

Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthcare Analytics, 2, 100032.

Dritsas, E., & Trigka, M. (2022). *PMC PubMed Central.* Retrieved from Stroke Risk Prediction with Machine Learning Techniques: 22(13):4670.

Lee, S. (2014). Plasma Focus Radiative Model: Review of the Lee Model Code. Journal of Fusion Energy, 33, 319–335.

National Heart, Lung, and Blood Institute. (2022, March 24). *What Is a Stroke?* Retrieved from National Heart, Lung, and Blood Institute: https://www.nhlbi.nih.gov/health/stroke

World Health Organization. (2022, October 29). *World Stroke Day 2022*. Retrieved March 1, 2023, from https://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-day-2022#:~:text=The%20Global%20Stroke%20Factsheet%20released,a%20stroke%20in%20their%20lifetime.

Yasa, P., Rusjayanthi, N., & Mohd Luthfi, W. (2022). Classification of Stroke Using K-Means and Deep Learning Methods. Lontar Komputer, 13, 23- 34.