Evaluating ML Binary Classification for Predicting Stroke-Related Mortality

Dalia Kriksciuniene, Virgilijus Sakalauskas,
Vilnius University

Abstract

Stroke is a leading cause of mortality and long-term disability worldwide, forcing effective predictive models to identify at-risk individuals and optimize treatment plans. In this study, we evaluate the performance of various machine learning (ML) algorithms in predicting stroke-related mortality. Five binary classification models—Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machines (XGBoost), Support Vector Machine (SVM), and Neural Networks (MLPClassifier) - were applied to a dataset containing clinical and demographic features of stroke patients. Each model was trained and evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. Also the feature importance was analyzed to find the key predictors of stroke mortality across different models. After performing a comparative analysis of the models, we disclosed the Random Forest and XGBoost performance advantages over simpler models, proposing superior accuracy and interpretability. This research highlights the potential of ML models in predicting stroke-related mortality and provides insights into the most critical features driving these predictions.

Introduction

The rise of applications of machine learning (ML) algorithms also influenced the healthcare facilities in predicting patient's outcome of the disease. Machine learning models have the potential for detecting patterns and correlations that escape traditional statistical methods. In stroke care, such models could unlock powerful new ways to identify patients at risk of mortality early, leading to faster interventions and tailored treatments. Yet, despite the promise of ML, not all algorithms are created equal. The task of choosing the right model, one that balances accuracy with interpretability, is a critical challenge that has yet to be fully addressed.

In this study, we aim to evaluate the performance of five prominent ML classifiers—Logistic Regression, Random Forest, Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Neural Networks (MLPClassifier) for predicting stroke-related mortality.

The experimental research was done using the database of registered stroke cases of the neurology department of Clinical Centre in Montenegro. The initial database consists of the structured 944 patients records with 58 variables. For our research we have selected 10 variables and apply specific coding to its value. The full description of used database is provided.

Using this clinical and demographic data from stroke patients, we have analysed not only the accuracy of these models but also the key features that drive their predictions. By providing a comprehensive comparison of these approaches, our goal is to shed light on which methods offer the greatest promise for improving stroke care, both in terms of predictive power and interpretability.

A brief overview of ML techniques

In this research, several machine learning (ML) techniques are employed to predict stroke-related mortality.

Logistic Regression - Logistic Regression is a linear model used for binary classification. It estimates the probability that a given input belongs to a certain class, based on a linear combination of input features.

Random Forest Classifier - Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance. It averages the predictions of several decision trees to reduce overfitting and variance. For each tree, predictions are made based on majority voting. The general structure for decision trees is based on recursively splitting the dataset according to feature values that minimize the Gini impurity.

Gradient Boosting Machines (XGBoost) - XGBoost (Extreme Gradient Boosting) is a powerful machine learning method that builds multiple small models (decision trees) in sequence, each one correcting the errors of the previous model. Instead of training all models at once, XGBoost builds one tree at a time, and the next tree focuses on fixing the mistakes made by the earlier trees.

Support Vector Machine (SVM) - Support Vector Machine (SVM) is a machine learning method used for classification tasks, which helps us separate data into two different groups (like predicting "survival" or "non-survival"). The goal of SVM is to find the best boundary (called a hyperplane) that separates the two classes as clearly as possible.

Neural Networks (MLPClassifier) - A neural network consists of layers of interconnected neurons. Each neuron applies a weighted sum of inputs followed by a non-linear activation function. The neural network is trained by minimizing a loss function.

In models like Random Forest and XGBoost, feature importance is derived from the contribution of each feature to splits in the decision trees. For models like SVM and Neural Networks, permutation importance can be used, which measures the change in model performance when a feature's values are randomly shuffled.

Stroke Clinical Data description

The database applied for the experimental research consist of stroke patient clinical data records registered by the neurology department of Clinical Centre of Montenegro, operating in Podgorica, Montenegro. The original database consists of the structured 944 records of stroke patients, 58 variables, where 50 of them are coded by scale values of {1,2,3} corresponding to "Yes, No, Unspecified", and 8 variables consisting of the demographic data, admission date and discharge date from hospital. The data was collected between 02/25/2017 and 12/18/2019. The demographic data of stroke patients varies by age (from 13 to 96 years), and gender (485-male, 427-female). For our research we have cleansed the initial stroke database, recoded some variables and finally got the 11 variables database for research (Sakalauskas et al. 2022). The example of database structure and data records is presented in Table.

	1 Days till Hospital	2 Vital Status	3 Stroke Type	4 Treatment methods	5 Health Status	6 Age	7 Gender	8 Past Stroke	9 Stroke Symptoms	10 Health complications	11 Smoke Status
1	1	1	3	0	1	17	2	0	0	0	2
2	0	1	3	4	2	16	1	0	12	0	2
3	0	0	2	4	0	13	2	1	12	0	3
4	0	0	1	24	0	96	1	0	23	4	2
5	0	0	1	24	9	94	2	0	23	4	3
6	0	0	4	4	0	93	2	0	123	2	3
7	0	0	1	24	0	94	1	0	12	4	2
8	1	0	1	24	1	93	2	1	123	23	2
9	0	1	1	24	0	91	1	0	2	0	2
10	1	0	2	4	0	91	1	0	23	4	1
11	0	1	1	24	0	91	1	0	13	0	2
12	0	0	2	4	0	90	2	0	2	0	2
13	0	1	3	4	0	92	1	0	1	0	3
14	0	1	1	24	4	91	2	1	23	0	2
15	0	0	1	24	0	90	2	0	123	2	2
16	1	0	1	12	3	91	2	0	12	0	2
17	0	0	1	24	1	89	2	1	123	2	2

Variable name	**Meaning and coding of data**
Days till Hospital	– the number of days after stroke till hospital admission
Vital Status	– 1:Event (death), 0: Alive/censored
Stroke Type	- 1: Ischemic, 2: Hemorag, 3: SAH, 4: Unspecified
Treatment methods	- 0:No treatment,1:Anticoagulation, 2:Dual Antiplatelet Therap 3:Thrombolysis, 4:Others, Two digit codes: mean combined treatme methods, e.g. 24:means 2 and 4 are applied
Health Status	- Health score before stroke from 0:best to 9:worst: 0: Witho symptoms; 1: Without significant disability despite symptoms; Minor disability; 3: Moderate disability, but able to wa independency; 4: Moderate disability, not able to walk independenc 5: Major disability; 9: Unknown
Age	– Patient age, years
Gender	– 1:Male, 2:Female, 9:Unspecified
Past Stroke	– Stroke in past. 1:Yes, registered in patient health record, 0:No
Stroke Symptoms	0:No symptoms, 1:Impaired consciousness, 2:Weakness/paresi 3:Speech disorder (aphasia), Several digit codes: 123-means all symptoms
Health complications	0:unspecified, 1: other CV (cardiovascular) complications 4:oth complications, Several digit codes: 23:means 2 and 3
Smoke Status	1-Smokes, 2-No, 3-Smoked before

Performance comparison of binary classification

Before applying the machine learning models, the dataset underwent several preprocessing steps: handling missing values, encoding categorical variables, splitting initial database to train-test sets and scaling where it is necessary.

The dataset was split into training and testing sets using train_test_split() from the sklearn library. The initial split ratio was set at 70% for training and 30% for testing. Each of the binary classification models described in section 3 was implemented in Python using corresponding python libraries. The function used in our program trains and evaluates the following models:

Logistic Regression:	LogisticRegression() from scikit-learn.
Random Forest Classifier:	RandomForestClassifier() from scikit-learn.
XGBoost:	XGBClassifier() from xgboost.
Support Vector Machine (SVM):	SVC() from scikit-learn.
Neural Network (MLPClassifier):	MLPClassifier() from scikit-learn.

To assess and compare the models, we used four evaluation metrics:

Accuracy:	The percentage of correct predictions.
Precision:	The percentage of correct predictions along the predicted positive cases.
Recall:	The percentage of correct predictions along the actual positive cases
F1-Score:	The harmonic mean of precision and recall.

The performance of each model was recorded in terms of these metrics and compared across all models to identify the best-performing one.

Performance of ML models

The Table present the comparison of performance data for all binary classification models in case of training set size is 70% - 188 records representing alive status and 96 not. The random state parameter here is set equal to 42.

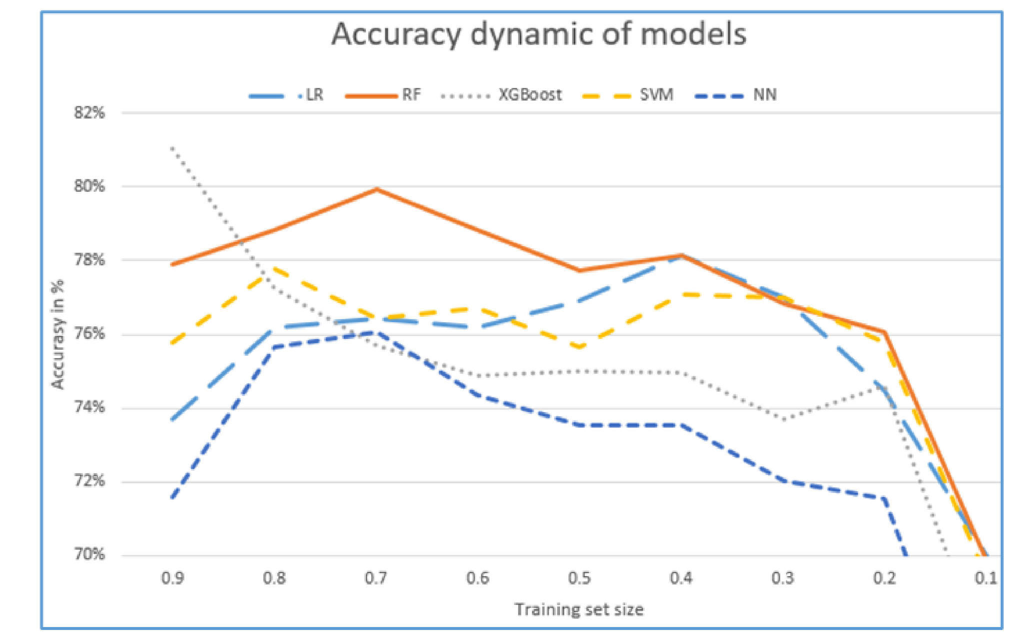
The performance metrics used are accuracy, precision, recall, and F1 score, which offer insights into the models' predictive power and reliability.

		Accuracy	*Precision*	*Recall*	*F1 score*
Logistic regression	*0-alive*	76.41%	0.77	0.91	0.84
	1-dead		0.73	0.48	0.58
Random Forest	*0-alive*	79.93%	0.80	0.92	0.86
	1-dead		0.78	0.56	0.65
XGBoost	*0-alive*	75.70%	0.80	0.85	0.82
	1-dead		0.66	0.58	0.62
SVM	*0-alive*	76.41%	0.76	0.94	0.84
	1-dead		0.78	0.42	0.54
MLPClassifier	*0-alive*	76.06%	0.80	0.85	0.82
	1-dead		0.66	0.59	0.63

In summary, Random Forest is the best-performing model in this study for stroke-related mortality prediction, especially in terms of overall accuracy and its ability to identify patients likely to survive. However, improving predictions for the mortality class remains a key challenge for all models, as evidenced by lower precision and recall for class "1-dead.„

Impact of the training set size on model accuracy

Our next experiment will measure the impact of the training set size on model accuracy. Figure presents the dynamic of accuracy of all models by changing training set size.



This graph illustrates the dynamic changes in the accuracy of five machine learning models based on the proportion of the training set used, ranging from 0.9 to 0.1. The y-axis represents accuracy percentages, while the x-axis shows the training set size as a proportion of the total dataset.

Feature importance score

Based on the feature importance table (Table below) across multiple models, we can select features most strongly influencing predictions for stroke-related mortality. In the table, the most influencing the classification accuracy feature is marked by number 1 and the weakest feature influence – by number 9.

	LR	*RF*	*XG Boost*	*SVM*	*MLP Classifier*	*Average score*
Health_Status	2	3	2	1	1	1.8
Age	1	1	9	4	2	3.4
Stroke_Symptoms	5	2	4	2	6	3.8
Health_complications	10	4	1	5	3	4.6
Smoke	6	6	3	3	5	4.6
Stroke_Type	4	7	5	6	4	5.2
Treatment_methods	3	5	6	9	8	6.2
Days_till_Hospital	8	9	7	8	9	8.2
Past_Stroke	7	10	8	7	10	8.4
Gender	9	8	10	10	7	8.8

As we can derive from the column Average score, the Health Status, Age and Stroke Symptoms are the most critical features, dominating across all models and suggesting that patient-specific health factors provide the most valuable information for mortality prediction.

Conclusion

In this study, we evaluated the effectiveness of five binary classification machine learning models—Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Neural Networks (MLPClassifier)—for predicting stroke-related mortality using a dataset of clinical and demographic features from the neurology department of Clinical Centre of Montenegro. By comparing these models across key performance metrics such as accuracy, precision, recall, and F1-score, we gained insights into their predictive power and reliability. Additionally, we examined feature importance to identify the most critical factors driving mortality predictions.

Our results indicate that ensemble-based models like Random Forest and XGBoost outperformed other methods, delivering higher accuracy and interpretability. These models consistently identified Health Status, and Age and Stroke Symptoms as the most influential predictors, underscoring the importance of these variables in stroke outcome prediction. While Neural Networks showed competitive performance, particularly in terms of precision and recall, the model's lack of interpretability remains a limitation in clinical applications where understanding the driving factors is crucial.

Interestingly, simpler models like Logistic Regression, while offering less accuracy, provided clearer insights into feature importance, making them potentially valuable in settings where transparency and ease of interpretation are critical. Conversely, SVM, while delivering good results for certain metrics, struggled with generalizability across different test set sizes.

Ultimately, our findings highlight the potential of machine learning models in predicting stroke-related mortality, with Random Forest and XGBoost standing out as the most robust options. These models offer both strong performance and the ability to interpret feature importance, making them suitable for real-world clinical applications. This research underscores the promise of data-driven approaches in improving stroke care by facilitating early identification of high-risk patients and guiding more personalized treatment strategies.

Contacts dalia.kriksciuniene@knf.vu.lt
virgilijus.sakalauskas@knf.vu.lt

Vilnius University
Institute of Social Sciences and Applied Informatics