Prediction of strokes using Machine Learning Algorithms

Abu Abdullah Dhrubo
Data Science
Hogskölan Dalarna
Borlänge, Sweden
h21abudh@du.se

Guran Yarikamrani
Business Intelligence
Hogskölan Dalarna
Borlänge, Sweden
v21borya@du.se

***Abstract*—Stroke is one of the leading causes of death worldwide, as reported by World Health Organization (WHO). Physicians provide possibilities for enhanced patient management by methodically collecting and documenting patients' health records as technology and medical diagnosis integrate. This study analyzes electronic health records to predict stroke outcomes using Machine Learning rooted in an unbalanced dataset. Using SMOTE for balancing datasets to model using the Logistic Regression approach and two more classification algorithms, K-Nearest Neighbor and Support Vector Machine, were utilized, and then a confusion matrix was used to demonstrate the performance and accuracy. The K-Nearest Neighbors classification outperforms the other models tested, with a classification accuracy of 89.2 percent. The outcomes demonstrated the critical relevance of over-sampling for unbalanced datasets like this.**

***Keywords*—Stroke, Disease, Prediction, Data Analytics, Data Science, Machine Learning**

I. INTRODUCTION

STROKE is a condition that affects the arteries that lead to and from the brain. It is the fifth-largest cause of death in the United States and a primary cause of disability. [1] A stroke happens when a blood artery carrying oxygen and nutrients to the brain becomes blocked or breaks (or ruptures). When this happens, a portion of the brain is deprived of blood (and oxygen), and it and brain cells die. [1]

A stroke is a type of cerebrovascular illness. This indicates that it influences the blood arteries that supply oxygen to the brain. Damage to the brain may begin if it does not receive enough oxygen. While many strokes are curable, some might result in disability or death. [2]

The goal of this study is to use Data Science and Machine Learning (ML) to develop an accurate model that can predict stroke outcomes based on individual patient's age, gender, marital status, work type, residence type, hypertension, heart disease, average glucose level, body mass index, smoking status and either they have experienced stroke or not. Providing relevant information for medical personnel to deploy necessary therapy and reduce risks and repercussions.

The most significant contribution of our study is that we applied various machine learning models to an openly accessible dataset. Previous studies have worked on Decision Tree, Naive Bayes, and Neural Network, which show acceptable accuracy in identifying stroke-prone patients. This paper [3] used, principal component analysis algorithm for reducing the dimensions and it determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient is suffering from stroke disease or not.

Research Questions:

Firstly, we tried to find out which model performs better? What is the performance of the model that was used?

Predicting the patient's need for medical help, this result is useful for preventing stroke. This project examines the application of the prediction, machine learning technique to predict stroke. Machine learning can be used to improve the accuracy of stroke predictions and help physicians take preventative health measures for those patients. [4] Some machine learning (ML) techniques have been shown to provide accurate predictions and are increasingly used in the diagnosis and prognosis of various diseases and health conditions. ML methods are data-based analytical methods that specialize in integrating several risk factors into one prediction algorithm. A variety of ML algorithms, including K-Nearest Neighbors (KNN), logistic regression (LR), and support vector machines (SVM), have been widely used to identify key features of patient conditions and to model disease progression.

Paper Structure: In Section 2, we focus on the basics of different machine learning techniques to make our journey to the next stage of the project easier. We start with the introduction of machine learning and three different models Logistics Regression, Support Vector Machine, and finally K-Nearest Neighbors. Section 3, this section describes the statistical methods and techniques used. In Section 4, we present the data analysis and the results, and finally, in Section 5, we conclude with the analysis.

II. LITERATURE REVIEW

Stroke is one of the leading causes of global mortality in developed countries. A stroke means that a heart attack or

bleeding occurs in the brain's blood vessels, and the disease usually occurs suddenly, leading to a lack of oxygen in the arteries of the brain. [1]

Symptoms of a stroke include dizziness, numbness, weakness, sudden severe headache, vision problems, difficulty walking or talking, dizziness, and speech. The most common cause of a stroke is a blood clot that blocks blood flow to an area of the brain. [5]

The advent of machine learning (ML) technology has greatly enriched healthcare services and created a new field of "smart healthcare". The aim of this study is to investigate the application of ML technology in the medical industry in predicting stroke, which can be useful for physicians in predicting a disease (stroke in this study).

III. Theory and Methodology

3.1 Proposed System. The application of the prediction technique, machine learning, is to predict stroke disease. Machine learning prevention of stroke helps physicians take preventive health measures for patients. Machine Learning (ML) is an integrated field of statistics, computer science, and engineering that facilitates the extraction of data based on pattern recognition and provides systems the ability to automatically learn and improve from experience without being explicitly programmed. [6]

The basic principle of machine learning is to build algorithms that can obtain input data and then predict results or outputs using statistical analysis over a satisfactory period. [6] The focus of machine learning in this study is to predict stroke and improve the accuracy and sensitivity of predictive models such as LR, SVM, and KNN in classified algorithms.

3.2 Dataset. The analysis was carried out using the stroke prediction dataset. [7] This dataset has 5110 observations and 12 features.

TABLE I Description of Dataset

Features	***Values***
Demographic data	
Number of individuals	5110
Age, years	1 – 82
Gender	Male/Female
Ever married	1/0
Work type	Private/Self-Employed/Govt_job/children/never worked
Residence type	Urban/Rural
Clinical data	
Hypertension	1/0
Heart disease	1/0
Avg Glucose Level	55.12 – 271.74
BMI [a]	10.3 – 97.6
Stroke	**1/0**
Other	
Smoking status	Formerly smoked/never smoked/smokes/Unknown

[a] BMI stands for Body mass index.

The output column stroke has a value of 1 or 0. The value 0 indicates that no stroke risk was detected, but the value 1 indicates that stroke risk was detected. In this dataset, the chance of 0 in the output column (stroke) exceeds the likelihood of 1 in the same column. In the stroke column alone, 249 rows have the value 1, while 4861 rows have the value 0. Data encoding was used to preprocess the data.

Figure I: observations with stroke and no stroke.

3.3 Imputation of data and reclassification Originally, the Dataset included 201 missing values for the Body Mass Index (BMI) feature. The mean BMI for the entire dataset was used to fill in these numbers. Furthermore, it was discovered that more than 30% of the observations have an Unknown smoking status, which can potentially be interpreted as missing data or a lack of knowledge regarding these characteristic values. To avoid this data being removed due to its volume, it was decided to re-categorize those persons using some estimates. Numeric features were also categorized due to the ease of model implementation.

3.4 Exploratory Data Analysis Using simple visualization in Figure. II showed that as the age of the person increases, the number of strokes also increases. As a result, the Dataset suggests that older persons are more prone to suffer from stroke.

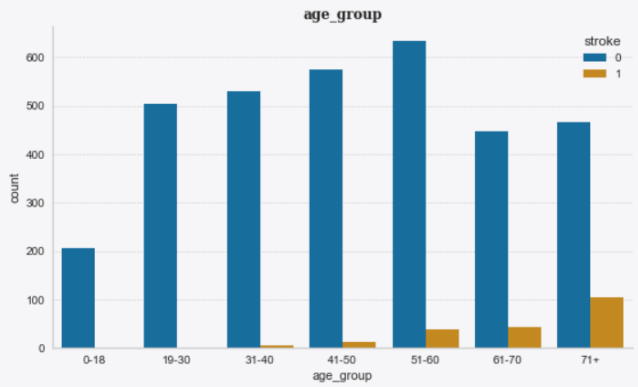

Figure II: observations with stroke and no stroke.

In figure III, overweight and obese people have higher BMIs hence they also have a higher record of stroke.

With these visualizations, we can observe that the dataset is not balanced because we have more observations of the patients who did not encounter stroke than the patients who have. And therefore, the predictions will have an impact on classification.

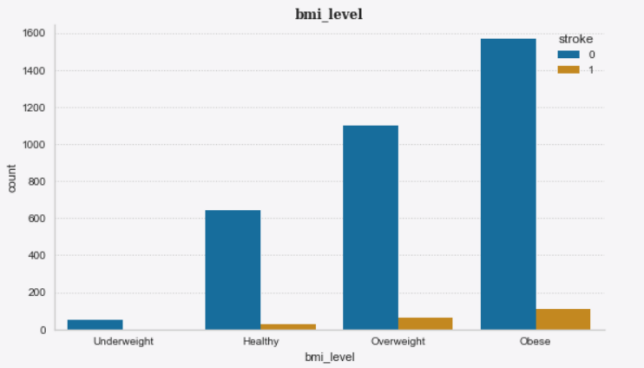

Figure III: observations with stroke and no stroke.

In figure IV, having a higher glucose level didn't have a strong relationship with the person experiencing a stroke.

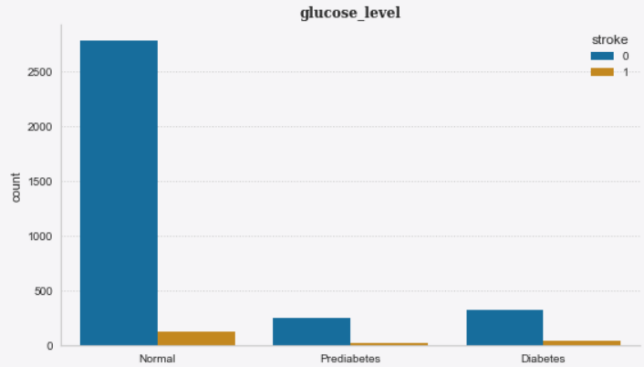

Figure IV: observations with stroke and no stroke.

3.5 Preprocessing Data preprocessing is required before developing a model to remove undesired noise and outliers from the dataset which might cause the model to deviate from its intended training. This stage focuses on everything that is impeding the model's efficiency. After gathering the appropriate dataset, these must be cleansed and processed for building the model. As previously indicated, the dataset used comprises twelve features.

To begin, the column id is excluded because it has no influence on model construction. The dataset is then checked for missing values and replaced if any are found. In this scenario, the null values in the variable BMI are filled with the mean of the data column. Label encoding turns the string literals in the dataset into integer values that the machine can understand. Strings must be translated to integers because the computer is typically educated on numbers. The obtained dataset has five columns of string data. During label encoding, all strings are encoded, and the entire dataset is converted to a series of numbers.

However, for our prediction, we did not use feature selection being that we are not the domain experts to determine whether we can say these features are important or not, to remove.

After completing data preparation, the next stage is to build the models. To improve the accuracy and efficiency of this work, the data is divided into training and testing data, with a 75/25 split. Upon splitting, the models are trained using 17 classification algorithms to be compared. The classification techniques used in the model evaluation are Ada Boost Classifier, Logistic Regression CV, Linear SVC, SVC, SGD Classifier, Passive Aggressive Classifier, Ridge Classifier CV, Gaussian Process Classifier, Gradient Boosting Classifier, K Nearest Neighbors Classifier, Random Forest Classifier, Extra Trees Classifier, Bagging Classifier, Bernoulli Naïve Bayes, Decision Tree Classifier, Perceptron, and Gaussian Naïve Bayes.

AUC Rate. This performance can be appropriate for the Accuracy and ROC rate of the predictive model. The area under the curve (AUC) gives the rate of successful classification by the logistic model. When AUC is approximately 0.5, the model has no discrimination capacity to distinguish between positive class and negative class. The AUROC is between 0 and 1, and AUROC = 1 means the prediction model is perfect. The further away the AUROC is from 0.5, the better; AUROC > 0.5, then we just need to invert the decision our model is making. As the Logistic Regression result, AUROC = 0.5, that is not good news because we just need to invert our model output to obtain a perfect mode.

Comparing the models, we conclude by selecting the best model for our prediction.

IV. RESULT

After running all the models, we come to have the results presented below table. It seems more than half of the models that we used did not perform as per the expectation. This raises the question if the data that has been fed has to do with this result or if the models that were used were inappropriate.

It shows the performance of 17 models based on different criteria, i.e., Accuracy, Precision, Recall, and AUC rate. For example, GaussianNB has the lowest accuracy of 0.22 on the test set.

TABLE II *Model Performance comparison*

No	ML Name	ML Train Accuracy	ML Test Accuracy	ML Precission	ML Recall	ML AUC
0	AdaBoostClassifier	0.943	0.945	0.000	0.000	0.500
6	LogisticRegressionCV	0.943	0.945	0.000	0.000	0.500
15	LinearSVC	0.943	0.945	0.000	0.000	0.500
14	SVC	0.943	0.945	0.000	0.000	0.500
9	SGDClassifier	0.943	0.945	0.000	0.000	0.500
8	RidgeClassifierCV	0.943	0.945	0.000	0.000	0.500
5	GaussianProcessClassifier	0.943	0.945	0.000	0.000	0.500
3	GradientBoostingClassifier	0.945	0.944	0.000	0.000	0.499
13	KNeighborsClassifier	0.944	0.941	0.167	0.020	0.507
7	PassiveAggressiveClassifier	0.934	0.937	0.231	0.061	0.525
4	RandomForestClassifier	0.970	0.935	0.000	0.000	0.495
2	ExtraTreesClassifier	0.970	0.933	0.000	0.000	0.493
1	BaggingClassifier	0.966	0.925	0.000	0.000	0.489
11	BernoulliNB	0.909	0.918	0.214	0.184	0.572
16	DecisionTreeClassifier	0.970	0.911	0.000	0.000	0.482
10	Perceptron	0.865	0.863	0.171	0.388	0.639
12	GaussianNB	0.229	0.224	0.066	1.000	0.590

And what we notice is that the results of these models are performing differently than our expectations. From which we can understand it is the result of using unbalanced data, therefore the models predicted randomly (with low accuracy).

The dataset contains 5110 observations, with 249 suggesting the probability of a stroke and 4861 proving the absence of a

stroke. While applying such data for training the machine learning models can result in inaccuracy, other metrics of performance, such as precision and recall can become insufficient. If such uneven data is not appropriately handled in the comparison of models, the results will be erroneous, and the prediction will result in misclassification. As a result, to acquire an efficient model, this unbalanced data must first be addressed.

For this objective, the SMOTE oversampling method is used. The Synthetic Minority Oversampling Technique, or SMOTE for short, is perhaps the most extensively used method for generating new samples. SMOTE begins by randomly selecting a minority class instance A and locating its k nearest minority class neighbors. The synthetic instance is then constructed by selecting one of its k nearest neighbors B at random and combining A and B in the feature space to form a line segment. The synthetic examples are created by convexly combining the two selected examples A and B. [8]

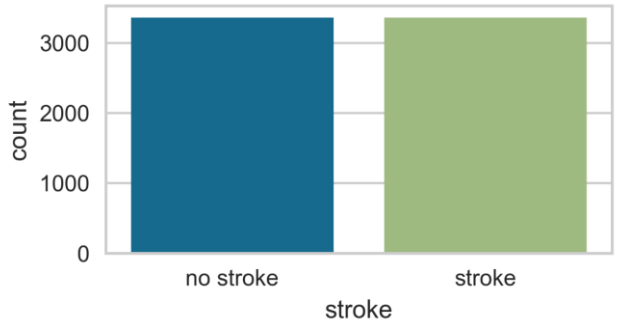

Figure V. observations after oversampling with SMOTE.

Figure V portrays the dataset has been balanced after oversampling using SMOTE.

Logistics Regression. We utilized this model, in which the y-intercept is defined by the best fit line and a cut-point of 0.5 (threshold). We're curious how effectively the model predicts whether a patient will have a stroke or not.

Support Vector Machine. We have used a linear kernel for our SVM model.

K-Nearest Neighbors. We have used selected 7 for k (k = 7).

TABLE III *After oversampling model performance*

Test Set	Precesion	Accuracy	Sensitivity	Specificity	AUROC
Lg Reg	0.824	0.794	0.749	0.824	0.801
SVM	0.917	0.867	0.810	0.917	0.802
KNN	0.937	0.892	0.842	0.937	0.892

Table II shows different matrices upon performing the LG, SVM, and KNN after oversampling using SMOTE. The AUROC value for KNN has the highest rate, in other words, it has the best performance among these three models in the case of the training and test data sets. We can also see in all the matrices the KNN model performed better.

A confusion matrix is a technique used to assess the performance of the machine learning classification models. The confusion matrix shows how frequently our models forecast accurately and how frequently they estimate erroneously. False positives and false negatives were assigned to incorrectly predicted values, while true positives and true negatives were allocated to correctly predicted values. After grouping all predicted values in the matrix, the accuracy, precision-recall trade-off, and AUC of the model were used to evaluate its performance.

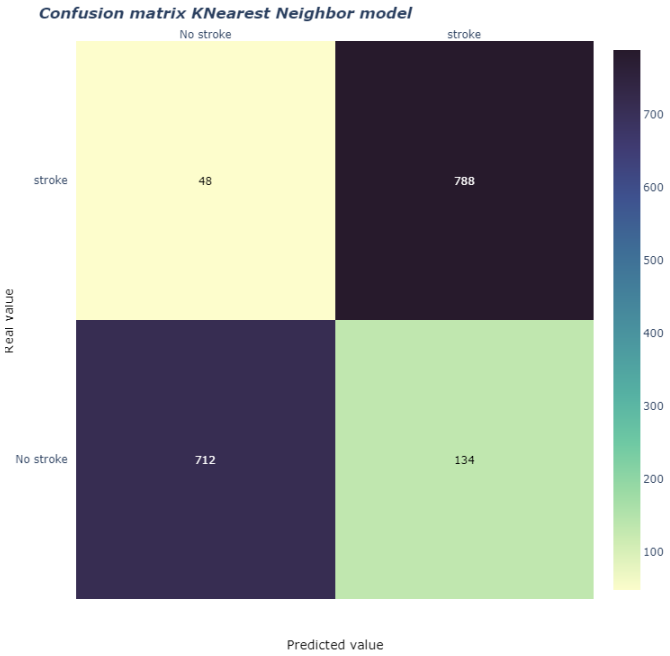

Figure VI. KNN Evaluation matrix.

Figure VI represents the confusion or evaluation matrix for the K Nearest Neighbors model.

With an accuracy of 0.892, the model estimates a total of 134 false negatives and 48 false positives. The area under the ROC curve (AUC) was calculated as well as the Receiver Operating Characteristic (ROC) curve. The greater the AUC, the more accurate the model is at classifying cases. The AUC was 0.8921, as indicated in the ROC curve in Figure VII.

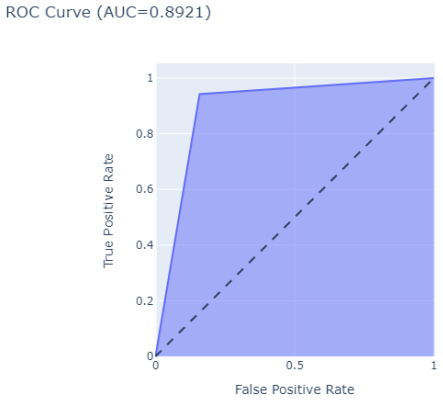

Figure VII. KNN ROC curve AUC score.

V. Conclusion

A stroke is a potentially fatal medical condition that must be treated as quickly as possible to prevent future consequences. The creation of a machine learning model might aid in the prevention of stroke and the associated decrease of its severe repercussions. This study looks at the performance of multiple machine learning algorithms in accurately predicting stroke based on specified physiological features. With a classification accuracy of 89.2%, the K-Nearest Neighbors classification exceeds the other models examined. The framework models may be improved in the future by utilizing a larger dataset and machine learning models such as AdaBoost, SVM, and Bagging. This will improve the framework's dependability as well as its presentation. The machine learning model may assist the public in identifying the possibility of a stroke developing in an adult patient in exchange for only supplying some basic information. In such a perfect scenario, it would assist patients in receiving early stroke treatment and rebuilding their lives following the incident.

References

[1] "About Stroke," American Heart Association, [Online]. Available: https://www.stroke.org/en/about-stroke#:~:text=Stroke%20is%20a%20disease%20that,or%20bursts%20(or%20ruptures).. [Accessed 2022].

[2] J. McIntosh, "Everything you need to know about stroke," Medical News Today, 11 March 2020. [Online]. Available: https://www.medicalnewstoday.com/articles/7624#definition. [Accessed May 2022].

[3] P. G. N. J. A. Sudha, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods".

[4] J. Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them from Scratch, Machine Learning Mastery, 2016.

[5] "Voice of MPN," [Online]. Available: https://www.voicesofmpn.com/glossary. [Accessed 2022].

[6] D. W. T. H. R. T. Gareth James, An Introduction to Statistical, Springer, 2021.

[7] "Stroke Prediction Dataset," 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

[8] H. H. Yunqian Ma, Imbalanced Learning: Foundations, Algorithms, and Applications, 2013.