COMPUTER DESIGN SYSTEMS. THEORY AND PRACTICE

Vol. 6, No. 1, 2024

Yurii Patereha[1], Mykhaylo Melnyk[2]

[1]Computer Aided Design Department, Lviv Polytechnic National University, S. Bandery street 12, Lviv, Ukraine, E-mail: yurii.i.patereha@lpnu.ua, ORCID 0009-0002-5110-008X
[2]Computer Aided Design Department, Lviv Polytechnic National University, S. Bandery street 12, Lviv, Ukraine, E-mail: mykhaylo.r.melnyk@lpnu.ua, ORCID 0000-0002-8593-8799

PREDICTION OF THE OCCURRENCE OF STROKE BASED ON MACHINE LEARNING MODELS

Received: March 11, 2024 / Revised: April 01, 2024 / Accepted: April 05, 2024

https://doi.org/

Abstract. The research conducted in the medical domain addressed a topic of significant importance, steadily growing in relevance each year. The study focused on predicting the onset of strokes, a condition posing a grave risk to individuals' health and lives. Utilizing a highly imbalanced dataset posed a challenge in developing machine learning models capable of effectively predicting stroke occurrences. Among the models examined, the Random Forest model demonstrated the most promising performance, achieving precision, recall, and F1-score metrics of 90%. These findings hold potential utility for healthcare professionals involved in stroke diagnosis and treatment.

Keywords: cerebrovascular accident, decision tree, randomized forest, stacking, synthetic Minority Over-sampling Technique, Grid Search, machine-learning

Introduction

Stroke stands as one of the most prevalent causes of mortality and disability globally. Often, individuals at heightened stroke risk remain unaware, foregoing necessary medical attention. Even upon consultation, diagnosis can prove challenging, leading to treatment delays and severe outcomes. Thus, there arises a pressing need for an effective stroke prediction system to mitigate risk and enhance quality of life. Within this framework, machine learning methodologies emerge as potent tools for stroke risk analysis and prediction. Consequently, it becomes imperative to explore the efficacy of diverse machine learning techniques in accurately predicting stroke risk and identifying optimal solutions for this predictive challenge.

The application of machine learning methods in forecasting stroke risk holds promise for early risk detection and timely intervention. This dissertation is dedicated to exploring various machine-learning approaches for analyzing patient data and forecasting stroke risk. The findings obtained could prove invaluable for healthcare professionals engaged in stroke diagnosis and treatment.

Employing machine learning methods enables the timely identification of high-risk stroke patients, facilitating the recommendation of preventive measures and enhancing treatment success rates, thereby improving overall health and quality of life. Moreover, these results can inform the development of novel stroke prevention and treatment strategies. Consequently, given these considerations, the need for a program to improve these statistics and save lives becomes apparent.

Problem Statement

There is a need to improve data pre-processing for more plausible results, to develop rapid, efficient, and optimized machine learning models, to achieve robust performance evaluation metrics for machine

learning models, including accuracy, and to search for and identify optimal hyperparameters for the machine learning models employed.

In their paper [1], the authors aimed to propose a stroke prediction model using a stacking ensemble classifier. The proposed model demonstrated an accuracy rate of 97%. However, the study presents some limitations that may restrict its applicability. For instance, the data preparation process is not adequately described. In another study [2], the authors conducted research proposing a machine-learning approach for stroke diagnosis utilizing unbalanced data. The results indicated that the Support VectorMachine model achieved the highest accuracy of 99.99%. Nonetheless, the article has several drawbacks: it utilizes a rather limited dataset, potentially impacting result accuracy. Overall, the papers [1-2] attained remarkably high metric scores, including addressing data imbalance issues. However, the notably high accuracy rates exceeding 98%-99% suggest potential overfitting of the models. In the paper [3], the authors achieved the highest accuracy with Random Forest, reaching 96.01%. However, the study has some limitations: the authors did not address data imbalance, raising questions about the accuracy rates [4].

Hence, while the reviewed studies exhibit numerous strengths, they also possess several limitations that need to be addressed. For instance, addressing the issue of unbalanced data to ensure accurate results and tackling underfitting models to obtain genuine and qualitative outcomes.

Review of Modern Information Sources on the Subject of the Paper

An analytical review of scientific sources was conducted following the standardized PRISMA methodology [5].

Initially, duplicates were identified and removed among 59 documents from the Scopus database and 29 documents from the Google Scholar database, totalling 88. Subsequently, 13 duplicates were identified and removed, resulting in 75 relevant documents.

Following this, an analysis of the obtained works was performed based on their titles, which best matched the thesis topic, and through abstract review. A further 44 documents were excluded, leaving 31 relevant documents.

Finally, a comprehensive analysis of the obtained full-text publications was conducted, considering the number of citations for each publication. Four documents were removed due to only having abstracts available and an additional 12 documents were excluded after a full review. These operations resulted in obtaining 15 of the most relevant full-text publications for qualitative analysis.

In the article [6], the authors presented a study proposing a machine-learning approach for stroke diagnosis using imbalanced data. The Random Over-Sampling (ROS) technique was utilized in this work to balance the data. Eleven classifiers were evaluated in this study, including Support VectorMachine Random Forest, K-nearest Neighbour, Decision Tree, Naive Bayes, Voting Classifier, AdaBoost, Gradient Boosting, Multi-Layer Perception, and Nearest Centroid. The results indicated that Support Vector Machine achieved the highest accuracy of 99.99%. Random Forest showed the second-highest accuracy at 99.87%. However, the article has several limitations: it employs a relatively limited dataset, which may impact result accuracy; it does not consider factors such as genetics and lifestyle, which could influence the risk of stroke development. Therefore, while the study represents a significant advancement in medical analytics, it also has certain limitations that warrant attention.

The authors of article [7] sought to propose a stroke prediction model utilizing machine learning classifiers and a stacking ensemble classifier. The study also suggests methodologies to enhance classification accuracy, such as feature selection and model parameter optimization. The proposed stacking prediction model amalgamated Random Forest, K-nearest Neighbour, Logistic Regression, Support VectorMachine, and Naive Bayes as base classifiers, with Random Forest selected for meta-learning. The achieved accuracy was 97%. However, the study possesses limitations that may impede its applicability. For instance, the process of data selection and preparation for analysis lacks sufficient description, potentially affecting result accuracy. Additionally, the authors did not compare their method with other machine-learning techniques.

In article [8], various machine learning algorithms including Logistic Regression, Decision Tree, Random Forest, K-nearest Neighbour, Support VectorMachine, and Naive Bayes were utilized to train five distinct models for accurate prediction. The Naive Bayes algorithm emerged as the best performer with an accuracy of approximately 82%. The study also provides essential recommendations for parameter optimization and feature selection. However, the work suffers from drawbacks such as insufficient information regarding the features utilized in the study.

In articles [9-11], the authors employed machine learning algorithms like Support VectorMachine, DNN, Random Forest, Decision Tree, Naive Bayes, and Logistic Regression. Random Forest achieved the highest accuracy at 96.01%. An advantage of this work is the utilization of an ensemble to combine the classifiers mentioned above and reduce potential errors in loss functions. Furthermore, the authors employed a relatively large dataset for training and testing their method. Nevertheless, the study lacks sufficient information about the classifiers utilized and their hyperparameters, and it does not compare the method with other alternatives, limiting the comparison of results and the determination of superior methods.

In articles [12-13], three different machine learning algorithms (Logistic Regression, Random Forest, XGBoost) were employed. The training dataset comprised 3178 patients. The accuracy values for Logistic Regression, Random Forest, and XGBoost were 0.89%, 0.89%, and 0.88%, respectively. A limitation of this study is its reliance on a limited amount of medical data collected in Japan, potentially restricting its applicability in other countries with different patient characteristics. Additionally, the article lacks sufficient information about the machine learning methods utilized and their parameters, hindering the reuse of methods for further research.

In articles [14-15], a revised version of the genetic convolution algorithm was introduced for stroke prediction based on various symptoms. The proposed model underwent comparison with machine learning models including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Support VectorMachine. Achieving an accuracy of 83.2%, the proposed minimal genetic convolution approach surpassed Logistic Regression by 4.2%, Naive Bayes by 1.2%, Decision Tree by 17.2%, and Support Vector Machine by 3.2%. However, the study bears some limitations worth noting. Particularly, the limited amount of data (only 1,000 patients) may diminish the model's accuracy and efficacy.

Furthermore, from the distribution diagram of works by type (Fig. 1), it is apparent that articles predominantly constitute the literature on this topic. This further underscores the significance of this subject, reaffirming people's efforts to propose pragmatic solutions.

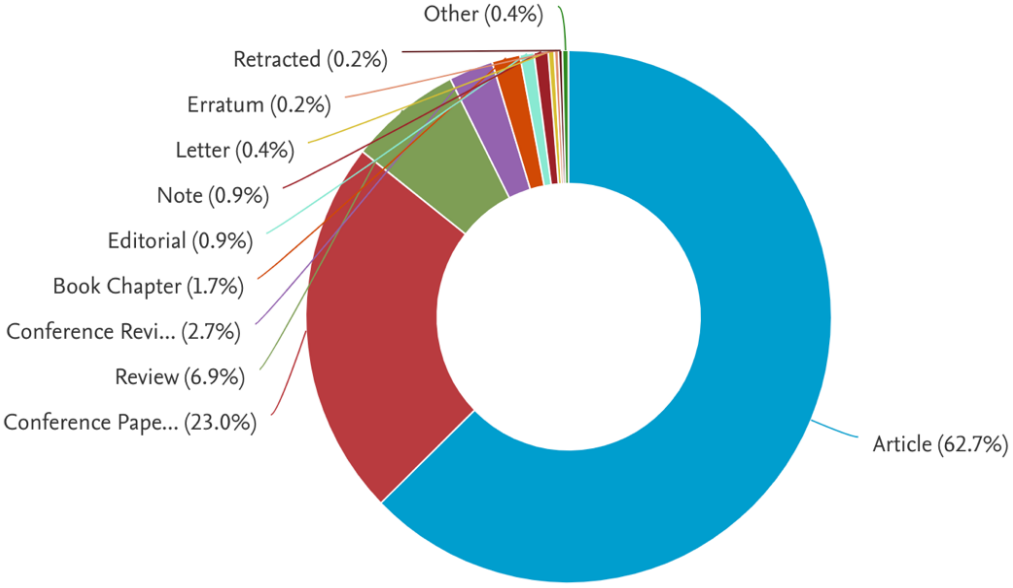

Fig. 1. Pie chart of distribution of works by type

Objectives and Problems of Research

The review and analysis of the selected articles have shown that there is currently significant interest in this topic. However, it can be said that many of them have some issues:

- Significant complexity of models using a combination of multiple machine learning methods, which did not actually improve accuracy for the authors but only increased the models' processing time and memory usage compared to using a combination of fewer methods.
- Relatively high levels of accuracy achieved due to the lack of processing very imbalanced data.
- Absence of optimal parameter tuning for different machine learning methods.

It is worth noting that each of the reviewed studies also has many positive and useful aspects that should be considered in one's research.

Therefore, the work aims to apply machine-learning methods to predict the risk of stroke.

Main Material Presentation

The dataset used to solve the given task - predicting the risk of stroke, was obtained from the DataHack Analytics Vidhya company website [16].

Therefore, the dataset consists of:

- 11 columns (different indicators and parameters of patients).
- 4981 rows (values of indicators and parameters) for each patient.

The first 5 and last 5 rows of the dataset are shown in Figure 2.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
2	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
3	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
4	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
...	...	...	...	...	...	...	...	...	...	...	...
4976	Male	41.0	0	0	No	Private	Rural	70.15	29.8	formerly smoked	0
4977	Male	40.0	0	0	Yes	Private	Urban	191.15	31.1	smokes	0
4978	Female	45.0	1	0	Yes	Govt_job	Rural	95.02	31.8	smokes	0
4979	Male	40.0	0	0	Yes	Private	Rural	83.94	30.0	smokes	0
4980	Female	80.0	1	0	Yes	Private	Urban	83.75	29.1	never smoked	0

4981 rows × 11 columns

Fig. 2. View of the first 5 and last 5 rows of data in the data set

Table 1 provides an overview of the data columns within the dataset:

Table 1

Stroke Data Set

Column Name	Type(Values) of the column	Description of the column
gender	String(Male, Female, Other)	Patient's gender
age	Integer	Patient's age
hypertension	Integer(1, 0)	Whether the patient has hypertension or not
heart_disease	Integer(1, 0)	Whether the patient has heart disease or not
ever_married	String(Yes, No)	Whether the patient is married or not
work_type	String(Govt_job, Never_worked, Private, Self-employed, children)	Categories for the work of the patient
Residence_type	String(Urban, Rural)	Categories for the residence type of the patient
avg_glucose_level	Float	Average patient's glucose level
bmi	Float	Patient's Body Mass Index (BMI)
smoking_status	String(formerly smoked, never smoked, smokes, unknown)	Categories for the smoking status of the patient
stroke	Integer(1, 0)	Whether the patient has a stroke or not

The dataset exhibits a significant imbalance, with the "no stroke" category (value equal to 0) occurring 4733 times, while the "stroke" category (value equal to 1) appears only 248 times.

Data preprocessing is crucial for preparing data for machine learning model training. Given the data structure and task at hand, the following preprocessing steps will be performed:

- Outlier removal using the interquartile range method.
- Categorical data encoding using the one-hot-encoding method.
- Addressing unbalanced data using the SMOTE method.
- Splitting the dataset into training (80%) and testing (20%) subsets.
- Attribute scaling using the min-max-scaler method.

For our task, we employed Decision Tree, Random Forest, and Stacking classification models.

Given the dataset size of 4981 rows and 11 columns, which is not particularly large, Grid Search was utilized for hyperparameter tuning.

Performance metrics are essential for assessing model effectiveness. The following metrics were employed for evaluation: accuracy, precision, recall, f1-score, and accuracy.

Results and Discussion

To solve the problem, we will use 5 classification models:

- Decision Tree classifier,
- Random Forest classifier.
- K-Neighbors classifier,
- AdaBoost classifier,
- Stacking classifier.

Overall, our dataset contains various types of data in machine learning.

Categorical features:

• Nominal - gender, ever_married, work_type, Residence_type, smoking_status.

• Binary - hypertension, heart_disease, stroke.

Numeric features:

• Continuous - avg_glucose_level, bmi.

• Discrete - age.

Now let's move on to visualizing our data. First, let's examine the distribution of our features Figures 3, 4, 5:

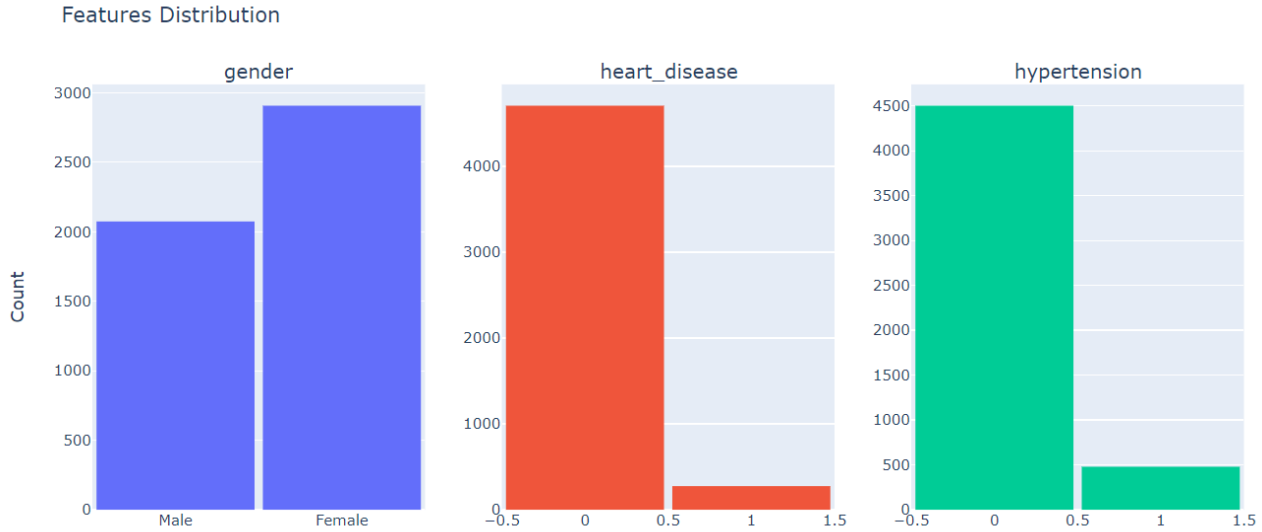

Fig. 3. Distribution of features - gender, heart_disease, and hypertension.

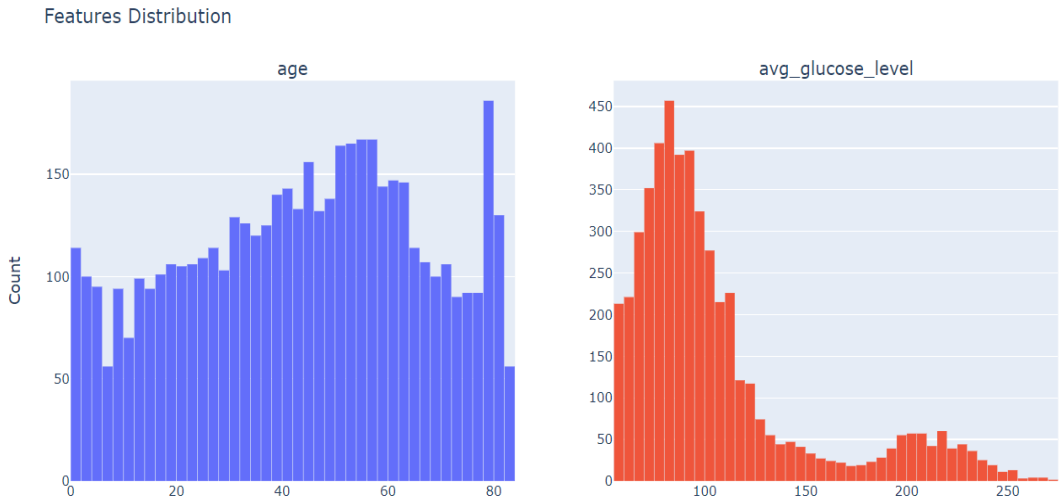

Fig. 4. Distribution of features – age andavg_glucose_level.

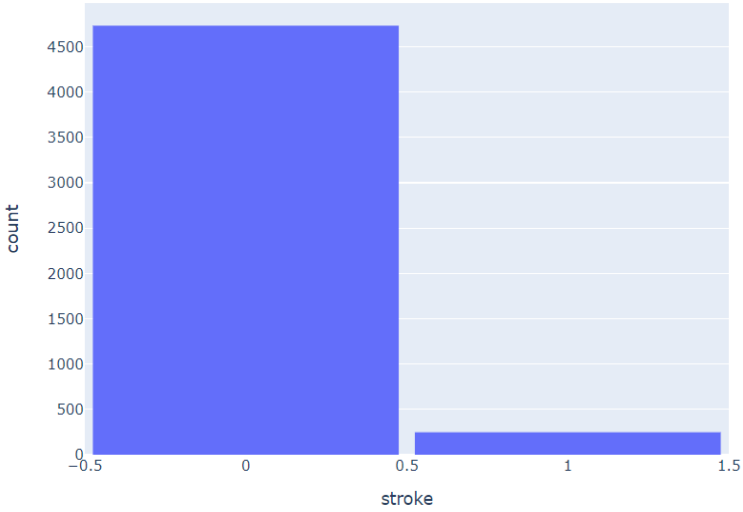

Fig. 5. Distribution of features – stroke.

So, we can propose the consequences:

• The number of married individuals outweighs the number of never-married individuals by a factor of 2.

• The most common type of occupation is private work, with self-employed individuals being five times less common, and the number of people working in government jobs almost equaling the number of unemployed individuals.

• The distribution of residential areas shows that approximately equal numbers of people live in urban and rural areas.

• The age distribution of individuals in the dataset resembles a uniform distribution.

• The average glucose level indicator most frequently falls within the range of 85-90, while it is least common in the ranges of 175-180 and 250+.

• The frequency of average glucose levels significantly drops after the 100 mark.

• Our dataset is highly unbalanced.

• The number of individuals without a stroke significantly exceeds the number of those who had one.

• The ratio between individuals without a stroke and those with one is 9:1.

• It is crucial to balance the dataset to ensure high-quality results from the models.

In this dataset, there are no missing values, indicating that operations such as deletion or averaging

of missing data are not necessary.

Now let's move on to removing outliers from our dataset. First, let's see which features have outliers:

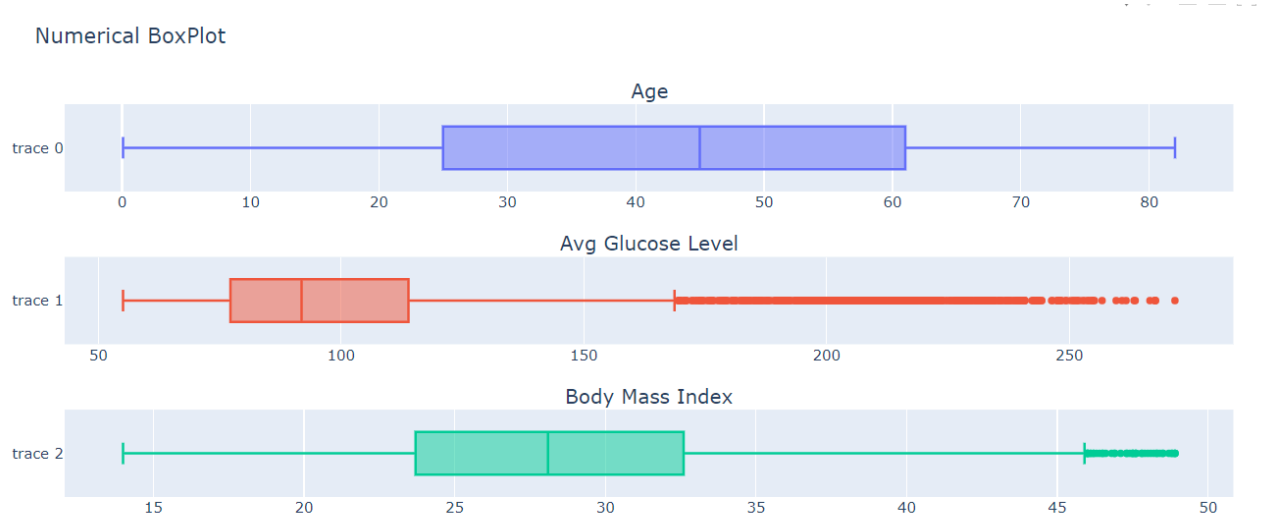

Fig. 6. Box plot of numerical features

We can conclude that there are outliers present in the following characteristics in our dataset: average glucose level and body mass index.

Now, we will proceed with attribute scaling using the min-max scaler method. The outcomes of attribute scaling are depicted in Figure 7:

	age	hypertension	heart_disease	avg_glucose_level	bmi
0	0.816895	0	1	0.447548	0.708464
1	0.975586	0	1	0.447341	0.579937
2	0.597168	0	0	0.447548	0.639498
3	0.963379	1	0	0.447548	0.313480
4	0.987793	0	0	0.447548	0.470219

Fig. 7. Scaling attribute outcomes - age, avg_glucose_level, and bmi

After performing all preprocessing operations on the data (except data splitting), let's examine the correlation matrix of the data to see possible dependencies before starting the construction of our machine-learning models. The correlation matrix is depicted in Figure. 8.

The results of the dependency analysis are as follows:

- The probability of stroke occurrence correlates most strongly with the age characteristic.
- The probability of stroke occurrence also correlates with the marital status characteristic, although not very strongly, but stronger than other features.
- The probability of stroke occurrence weakly correlates with hypertension and heart problems.
- Other characteristics seem to have little influence on the probability of stroke occurrence.

Finally, let's perform the data-splitting operation into training and testing datasets. We will split the data in a ratio of 80% to 20%, where 80% represents the training data and 20% represents the testing data for the models.

Results of Model Evaluation: Next, we will proceed to examine the outcomes of training and testing our models, starting with an assessment of the performance of the base models using default hyperparameters on the testing dataset.

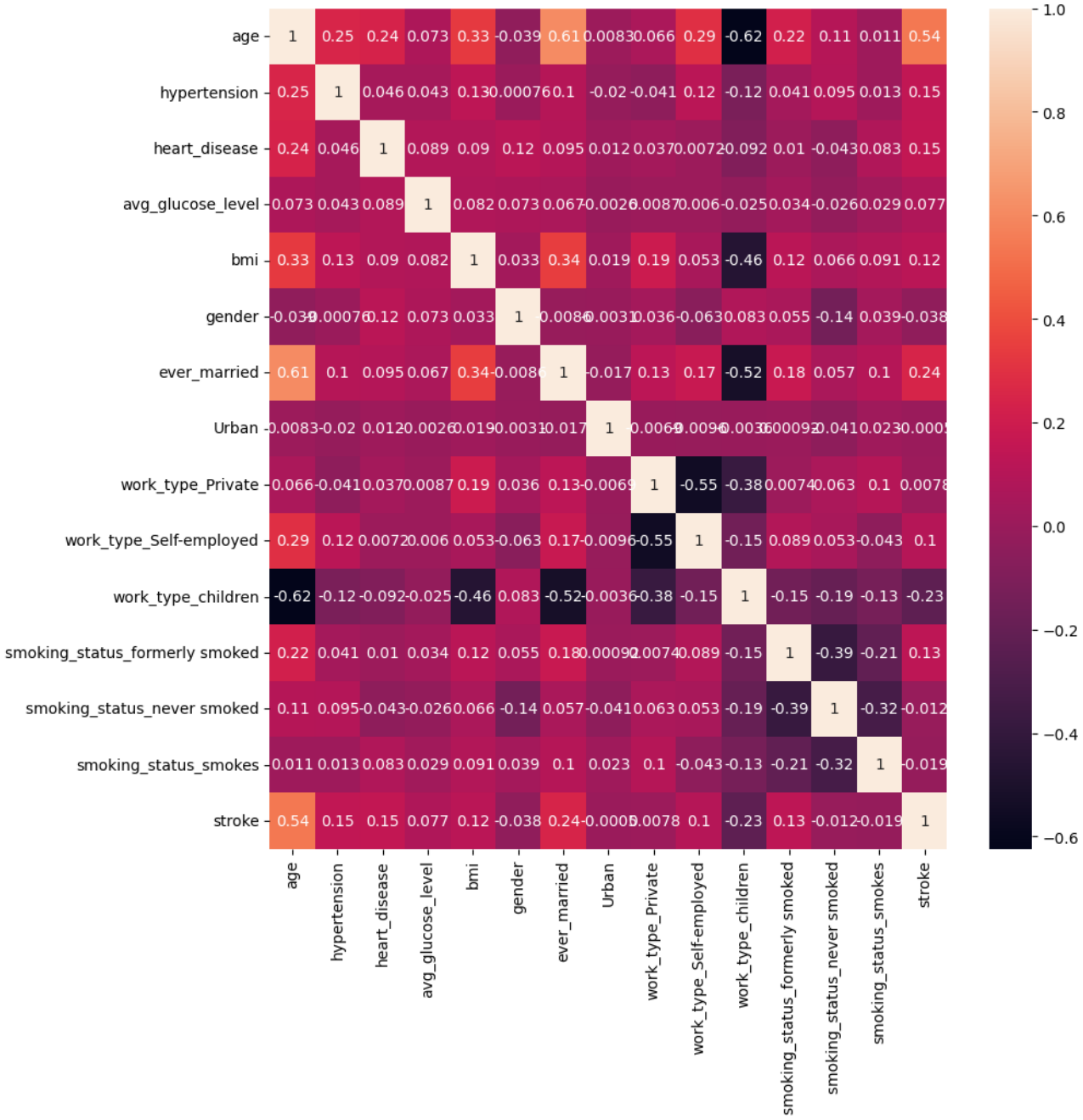

Fig. 8. Correlation matrix of data in a data set

The testing results of the basic Decision Tree Classifier model are presented in Table 2:

Table 2

Results of the basic Decision Tree Classifier model

class	precision	recall	f1-score	accuracy
0(no stroke)	0.89	0.91	0.81	0.97
1(stroke)	0.82	0.80		

The observations from this table are as follows: The model demonstrated a relatively consistent and equitable classification of the data, as evidenced by the close precision and recall values for both classes, suggesting effective differentiation between them. However, the f1-score is 76%, indicating a moderate level of performance.

Next, let's examine the outcomes of testing the basic Random Forest Classifier, as depicted in Table 3:

Table 3

Results of the basic Random Forest Classifier model

class	precision	recall	f1-score	accuracy
0(no stroke)	0.93	0.94	0.86	0.9
1(stroke)	0.87	0.86		

From the data presented in this table, the following conclusions can be drawn: The model demonstrated stable and balanced classification of the data, similar to the baseline Decision Tree Classifier model. Both precision and recall values are closely aligned for the two classes. Additionally, the f1-score is notably high at 86%, indicating promising performance and suitability for further refinement.

Next, let's examine the outcomes of testing the basic Stacking Classifier, as depicted in Table 4:

Table 4

Results of the basic Stacking Classifier model

class	precision	recall	f1-score	accuracy
0(no stroke)	0.94	0.93	0.87	0.91
1(stroke)	0.87	0.88		

The following observations can be made from the provided table: The model demonstrated stable classification, similar to the base Random Forest Classifier model. Both precision and recall scores exhibit close alignment for both classes. Notably, the f1-score improved to 87%, marking it as the highest result among the baseline models.

In summary, the Random Forest Classifier model emerged as the top performer among the baseline models. It showcased stability, avoiding confusion between classes, and achieved an impressive f1-score of 86%, indicating robust performance for this dataset.

Moving forward, let's examine the outcomes of optimizing our models' hyperparameters using the Grid Search method. Notably, the Random Forest Classifier model showed the most significant improvement. It produced a model with the following hyperparameters: n_estimators – 500; criterion – entropy; max_depth – 30.

The outcomes of this optimized model are illustrated in Table 5:

Table 5

Results of the improved Random Forest Classifier model

class	precision	recall	f1-score	accuracy
0(no stroke)	0.93	0.94	0.9	0.91
1(stroke)	0.89	0.86		

The insights gleaned from these findings are as follows: There has been a notable 4% increase in the f1-score, reaching 90%, signifying a commendable enhancement. Additionally, the model exhibited balanced classification.

Comparison with the model trained on unbalanced data:To assess the influence of imbalanced data, we will construct a random forest classifier model without employing the SMOTE method and examine its performance.

Table 6

Results of the Random Forest Classifier model without SMOTE method

class	precision	recall	f1-score	accuracy
0(no stroke)	1	0.96	0.98	0.96
1(stroke)	0	0	0	

The insights drawn from this table are as follows: The model exhibits poor performance and lacks

stability; it fails to classify class 1 (stroke) entirely; Although the accuracy metric suggests 96% accuracy, it does not reflect the model's actual performance; With precision, recall, and f1-score for class 1 (stroke) all at 0, the model fails to differentiate this class. Thus, it is evident that addressing the issue of highly unbalanced data is crucial, as the model's training results will be inaccurate. Using the accuracy metric alone for assessing classification models is inadequate.

Conclusions

The study conducted research in the medical field, which holds significant importance for individuals and continues to gain relevance over time. Its primary objective was to predict the occurrence of stroke, a critical menace to human health and well-being.

The Random Forest Classifier, identified as the most effective model, exhibited the following performance metrics: Precision - 90%; Recall - 90%; F1-score - 90%. Additionally, an accuracy score of 91% was obtained. However, to illustrate the limitations of relying solely on accuracy, the Random Forest Classifier was trained on unbalanced data, resulting in an accuracy rate of 96%.

These findings hold significant implications as they can aid healthcare professionals in implementing preventive measures more efficiently, thereby enhancing the prospects of preserving patients' lives and well-being by mitigating the risks associated with stroke.

References

[1] Abedi V., Avula V., Chaudhary D., Shahjouei S., Khan A., Griessenauer C. J., Li J., Zand R. Prediction of Long-Term Stroke Recurrence Using Machine Learning Models. Journal of Clinical Medicine. 2021. Vol. 10, № 6. C. 1286. https://doi.org/10.3390/jcm10061286

[2] Melnykova N., Chereshchuk L. Application of machine learning methods for predicting the risk of stroke occurrence. Proceedings of the VI International Scientific and Practical Conference. Sofia, Bulgaria. 2023. pp. 210-216. International Science Group, 2023. ISBN 9798891451926.

[3] Ashrafuzzaman Md., Saha S., Nur K. Prediction of Stroke Disease Using Deep CNN Based Approach. Journal of Advances in Information Technology. 2022. Vol. 13, № 6. https://doi.org/10.12720/jait.13.6.604-613

[4] Sun X. Predictive model analysis of stroke disease based on machine learning. SPIE, 2023. https://doi.org/10.1117/12.2669554

[5] Preferred Reporting Items for Systematic Reviews and Meta-Analyses. 2023.

[6] Biswas N., Uddin K. M. M., Rikta S. T., Dey S. K. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics. 2022. Vol. 2. C. 100116. https://doi.org/10.1016/j.health.2022.100116

[7] Mostafa S. A., Elzanfaly D. S., Yakoub A. E. A Machine Learning Ensemble Classifier for Prediction of Brain Strokes. International Journal of Advanced Computer Science and Applications (IJACSA). 2022. Vol. 13,№ 12. https://doi.org/10.14569/IJACSA.2022.0131232

[8] Sailasya G., Kumari G. L. A. Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. International Journal of Advanced Computer Science and Applications (IJACSA). 2021. Vol. 12, № 6. https://doi.org/10.14569/IJACSA.2021.0120662

[9] Khan M. K. Computer Science and Engineering.

[10] Uchida K., Kouno J., Yoshimura S., Kinjo N., Sakakibara F., Araki H., Morimoto T. Development of Machine Learning Models to Predict Probabilities and Types of Stroke at Prehospital Stage: the Japan Urgent Stroke Triage Score Using Machine Learning (JUST-ML). Translational Stroke Research. 2022. Vol. 13, № 3. C. 370–381. https://doi.org/10.1007/s12975-021-00937-x

[11] Mezher M. A. Genetic Folding (GF) Algorithm with Minimal Kernel Operators to Predict Stroke Patients. Applied Artificial Intelligence. 2022. Vol. 36, № 1. C. 2151179. https://doi.org/10.1080/08839514.2022.2151179

[12] Tegistu B. S. Brain stroke prediction model using deep neural network (dnn). 2021.

[13] Pitchai R., Dappuri B., Pramila P. V., Vidhyalakshmi M., Shanthi S., Alonazi W. B., Almutairi K. M. A., Sundaram R. S., Beyene I. An Artificial Intelligence-Based Bio-Medical Stroke Prediction and Analytical System Using a Machine Learning Approach. Computational Intelligence and Neuroscience. 2022. P. e5489084.

https://doi.org/10.1155/2022/5489084

[14] Rohit A. P. V., Chowdary M. U., Ashish G. B. S., Anitha V., Sana S. Ml approach for brain stroke prediction using ist database. 2022. Vol. 7, № 10. https://doi.org/10.33564/IJEAST.2023.v07i10.008

[15] Telu V., Padimi V., Ningombam D. D. Optimizing Predictions of Brain Stroke Using Machine Learning. Journal of Neutrosophic and Fuzzy Systems. 2022. Vol. 2. С. 31–43. https://doi.org/10.54216/JNFS.020203

[16] DataHack : Biggest Data hackathon platform for Data Scientists.

Юрій Патерега[1], Михайло Мельник[2]

[1]Кафедра систем автоматизованого проектування, Національний університет Львівська політехніка, вул. Степана Бандери 12, Львів, Україна, E-mail: yurii.i.patereha@lpnu.ua, ORCID 0009-0002-5110-008X

[2] Кафедра систем автоматизованого проектування, Національний університет Львівська політехніка, вул. Степана Бандери 12, Львів, Україна, E-mail: mykhaylo.r.melnyk@lpnu.ua, ORCID 0000-0002-8593-8799

ПРОГНОЗУВАННЯ ВИНИКНЕННЯ ІНСУЛЬТУ НА ОСНОВІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Отримано: березень 11, 2024 / Переглянуто: квітень 01, 2024 / Прийнято: квітень 05, 2024

Анотація. Дослідження, проведені в галузі медицини, стосуються важливої теми, інтерес до якої з кожним роком зростає. Дослідження було зосереджено на прогнозуванні початку інсульту, стану, що становить серйозний ризик для здоров'я та життя людей. Використання надзвичайно незбалансованого набору даних стало проблемою для розробки моделей машинного навчання, здатних ефективно передбачати випадки інсульту. Серед розглянутих моделей модель Random Forest продемонструвала найбільш багатообіцяючу продуктивність, досягнувши 90% показників точності, запам’ятовування та оцінки F1. Ці висновки можуть бути корисними для медичних працівників, які займаються діагностикою та лікуванням інсульту.

Ключові слова: порушення мозкового кровообігу, дерево рішень, рандомізований ліс, накопичення, техніка надмірної вибірки синтетичних меншин, пошук у сітці, машинне навчання