Journal of Process Management and New Technologies
Vol. 12, Issue 1-2, 2024, pp. 110-120

PREDICTIVE MODELING OF STROKE OCCURRENCE USING PYTHON FOR IMPROVED RISK ASSESSMENT

Đorđe PUCAR[1*], Vladimir ŠIMOVIĆ[2]

[1]Faculty of Applied Management, Economics and Finance, Belgrade, Belgrade, University Business Academy in Novi Sad, Serbia, djordje@mef.edu.rs
[2]Faculty of Economics and Engineering Management, Novi Sad, University Business Academy in Novi Sad, Serbia, vladimir.simovic@outlook.com

Abstract: *This paper examines the use of Machine Learning (ML) techniques, particularly Logistic Regression and Random Forests, to predict the occurrence of strokes. It integrates demographic, clinical, and lifestyle factors. The study uses Python as the primary tool for model development and analysis, focusing on binary classification to categorize individuals as either having had a stroke or not. The dataset includes attributes such as age, gender, hypertension, smoking status, and more, which are used to train and evaluate the models. Through extensive experimentation and evaluation, the paper demonstrates the effectiveness of Logistic Regression and Random Forests in stroke prediction. Logistic Regression provides a straightforward baseline, while Random Forests offer higher predictive accuracy. The findings highlight the importance of ML-based approaches in healthcare risk assessment and showcase Python's versatility in facilitating such analyses.*

Keywords: *Python, ML, binary classification, Logistic Regression, Random Forests*

Original scientific paper
Received: 09.05.2024
Accepted: 18.06.2024
Available online: 29.06.2024
DOI: 10.5937/jpmnt12-50921

1. Introduction

Stroke is a serious health issue that can lead to disability and even death. It's crucial to have effective ways to assess the risk and prevent it. Stroke is the second most frequent cause of death worldwide and one of the major contributors to disability in the elderly (Maier et al., 2016). Traditional risk assessment methods often struggle to accurately predict stroke occurrences because they can't fully capture the complex relationships between different risk factors. Machine Learning (ML) offers a better solution. ML uses advanced algorithms to analyze a wide range of variables, such as demographics, health indicators, and lifestyle choices, to identify individuals at high risk of having a stroke. This allows for targeted interventions and personalized preventive measures. A stroke is a sudden disruption in the blood supply to the brain, affecting one or more blood vessels that nourish the brain. This results in a disturbance or deficiency in the brain's oxygen supply, causing damage or impairment to brain cells (Merdas, 2024). The study mentioned aims to develop and evaluate

* Corresponding author

ML models specifically designed for predicting stroke occurrences. This research project uses a detailed set of data that covers various risk factors linked to strokes. By using Machine Learning algorithms, we hope to make stroke predictions more accurate and reliable. By using Machine Learning algorithms, the researcher hopes to make stroke predictions more accurate and reliable (Wang et al., 2020). This will help with early treatment and better outcomes for people at risk. Machine Learning algorithms, particularly Random Forest, can be effectively used in long-term outcome pre diction of mortality and morbidity of stroke patients (Fernandez-Lozano et al., 2021). This shows how Machine Learning can be useful in healthcare and how important it is to use technology to tackle complex health problems effectively. According to 2021 statistics, 1 in 6 deaths from cardiovascular disease was due to stroke (Poorani et al., 2023).

2. Methods

The research employed a comprehensive dataset packed with essential variables for forecasting stroke events. This data encompassed a wide array of details, including demographic information like age and gender, critical health factors such as high blood pressure, heart conditions, average glucose levels, and body mass index (BMI), as well as lifestyle elements like smoking habits. To prepare the dataset for analysis, a meticulous preprocessing phase was undertaken to ensure data integrity and suitability for Machine Learning algorithms. This involved addressing any missing data points through appropriate imputation methods, ensuring the dataset was complete and ready for analysis. Categorical features were transformed using techniques like binary classification or label encoding, converting them into a format that Machine Learning models could understand and process effectively. Machine learning is a type of artificial intelligence that focuses on constructing computerized algorithms to automatically improve performance through experience (Zu et al., 2023). Furthermore, numerical data underwent normalization, a process that standardized the scale across different features, promoting consistency and aiding model convergence during the training phase, ensuring optimal performance and accuracy.

To bolster the predictive potency of the models employed, sophisticated feature engineering methodologies were meticulously implemented. Rigorous techniques like dimensionality reduction and feature selection were judiciously applied, enabling the identification and preservation of the most pertinent and informative attributes for accurately forecasting stroke occurrences. This strategic approach facilitated the streamlining of the models' learning processes by honing their focus on the most significant features extracted from the comprehensive array of variables encompassed within the dataset. Consequently, the model's proficiency in precisely anticipating stroke events was markedly augmented.

After completing the preprocessing steps, the dataset underwent a crucial step of partitioning into training and testing sets using a method called stratified sampling. This technique was employed to maintain a balanced distribution of classes across both sets, ensuring that the models are trained and evaluated on representative data. Subsequently, a range of Machine Learning algorithms were selected and implemented to predict stroke occurrences based on the processed dataset. Moreover, as for the method for developing models, the most frequently used method in the included studies was logistic regression, which is consistent with a recent review, indicating a preference for logistic regression models in this specific field (Bonkhoff et al., 2021). These algorithms encompassed a variety of approaches, including Logistic Regression, Random Forests, support vector machines, gradient boosting. Each algorithm was assessed for its effectiveness in capturing patterns within the data and making accurate predictions regarding stroke events. The process of thoroughly evaluating the models involved a meticulous assessment phase. This phase utilized widely accepted

classification metrics to accurately gauge and measure the performance of the models. Metrics such as accuracy, precision, recall, and the F1 score were calculated. These metrics provided crucial insights into the model's predictive abilities, specifically their effectiveness in correctly classifying instances related to the occurrence of strokes.

The predictive models developed through this comprehensive study hold immense potential for healthcare professionals. These models can serve as invaluable tools in facilitating the early identification of individuals who are at an elevated risk of experiencing a stroke. By enabling timely intervention and personalized preventive measures tailored to the specific needs of each patient, these models possess the remarkable capacity to significantly improve patient outcome-s and elevate the overall quality of healthcare delivery. Early detection and proactive measures empower healthcare providers to take necessary steps that can potentially save lives and mitigate the devastating consequences of strokes.

2.1. Data Preprocessing and Visualization

Before analyzing the dataset with Machine Learning techniques, it underwent a comprehensive data preparation process to ensure its suitability. This process involved several crucial steps to handle missing values, encode categorical features, and normalize numerical data. To ensure the data preprocessing step goes correctly, libraries such as pandas, NumPy, matplotlib and seaborn were used for either data preparation or data visualization.

First step was to load the comma separated value (CSV) file which had the stroke data, the dataset came from kaggle.com. To load the CSV data file, we used the pandas library, as seen in Figure 1. The variable used to store this massive amount of data is stroke.

```
[1]: import pandas as pd
     import numpy as np

[2]: import matplotlib.pyplot as plt
     import seaborn as sns
     %matplotlib inline

[3]: stroke = pd.read_csv('stroke.csv')
```

Figure 1. Reading the CSV file
Source: Kaggle.com

Visualization of the raw data, seen in Figure 2., can be done with the head method on the stroke dataset. The visualization of the unprepared data will help to determine the next steps for data manipulation.

```
[4]: stroke.head(4)
```

[4]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1

Figure 2. Raw Data visualization
Source: Kaggle.com

For machine learning, the best practice is for data to be numerical in nature, if we are not dealing with the NLP type of models, so we have quite a few table columns that are in need to

be processed like the gender, ever married, work type, residence type, smoking status columns. Some columns like the duplicated id column which we get from the dataset are not needed and could be deprecated completely to avoid confusion. Next step is to figure out which data is missing or null (Nans), to achieve this, we could either use describe method to get a table which will show number of usable data, or a visualization technique using a heatmap.

To easily understand what is missing, the seaborn library heatmap method is used to identify columns that have null or missing data seen in Figure 3.

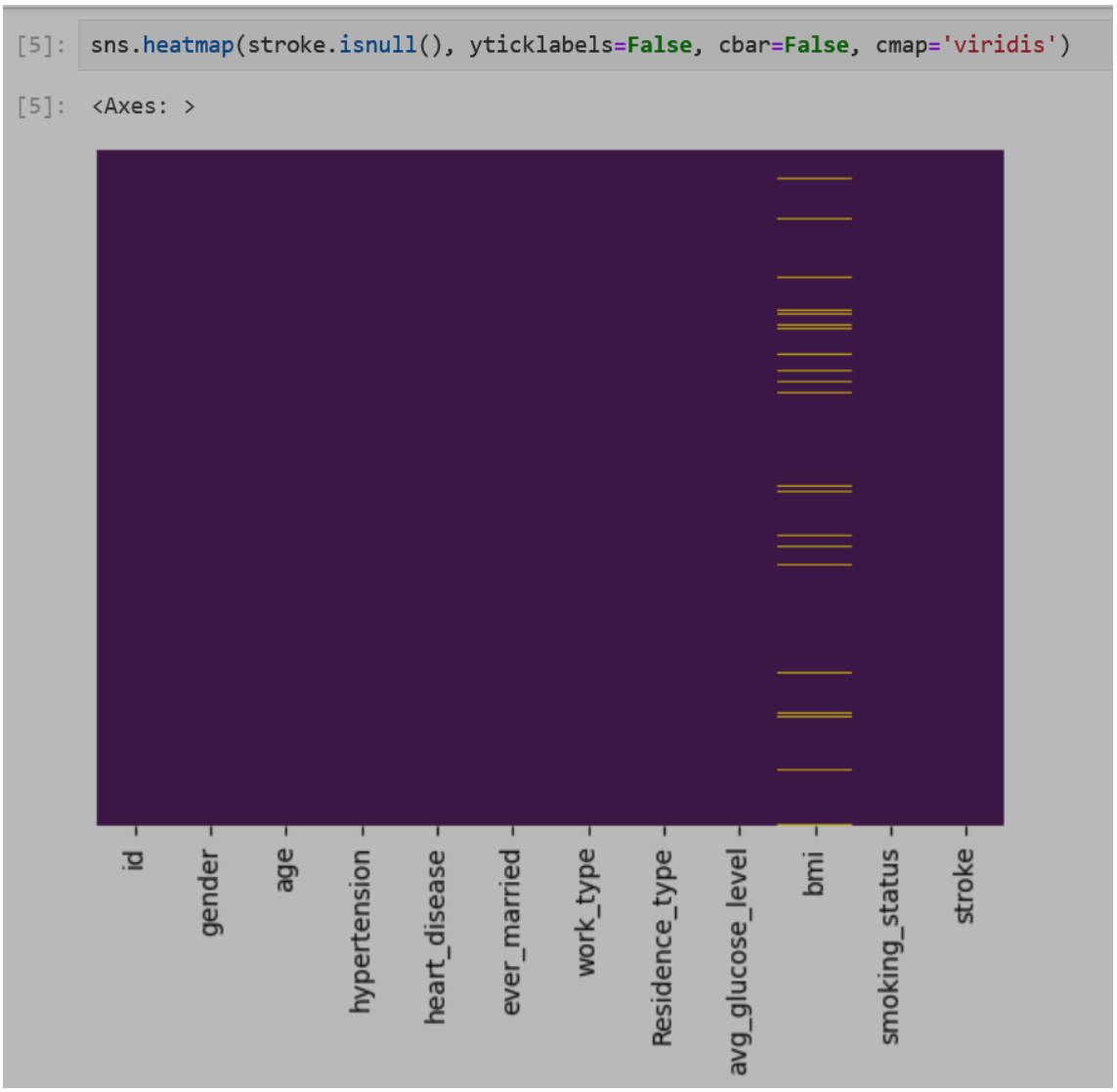


Figure 4. Sns heatmap representation of data via columns
Source: Kaggle.com

Via data plotting using the heatmap, we can conclude that most of the data is present, the only data that is missing in small amounts are colored in yellow and are mostly based in the BMI column. To avoid Nans or null data to mess up our model training we need to either fill it up with some value or drop the Nan values completely from our dataset. Because we are dealing with a rather large dataset (5110 entries) the best option is to use mean value replacement in the Nan fields which is done by calling mean method and to fill the missing data the use of fillna method is required. Doing this we are left with no nulls in our dataset.

To better understand the correlation of data with the stroke column, and for research of correct methods that will be used to train our models via Logistic Regression and Random Forest, for that the correlation will be done in x axis for the stroke and y axis in smoking_status seen in Figure 5.

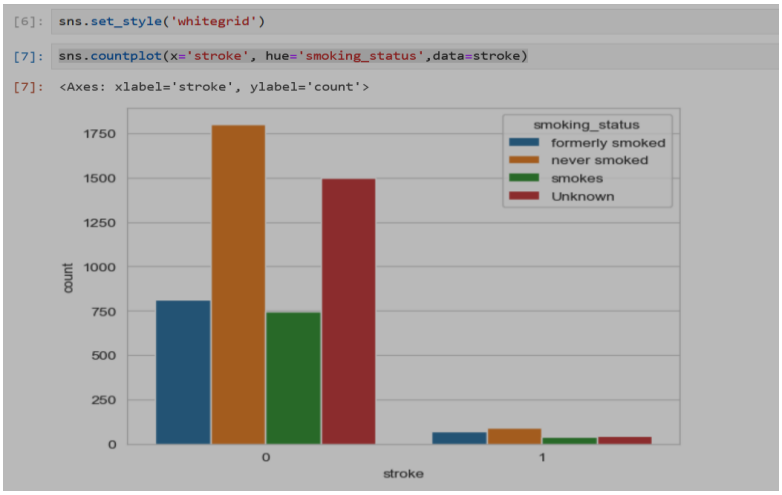


Figure 5. Sns countplot representation of stroke: smoking_status
Source: Author's calculation

The correlation will be also done in x axis for the smoking_status and y axis for age to better understand the demographics of people in the dataset seen in Figure 6.

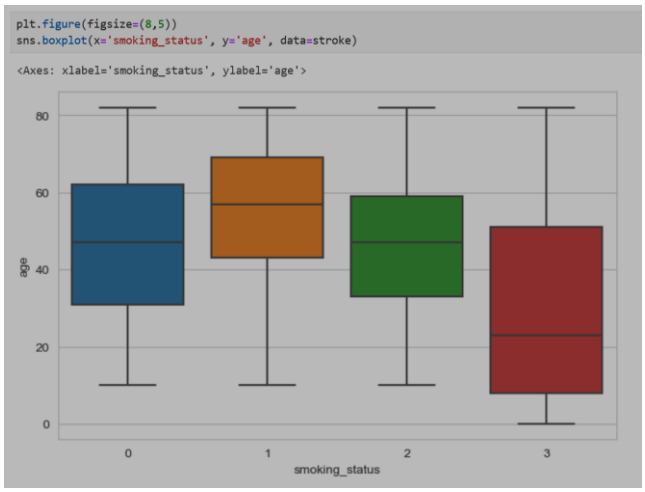


Figure 6. Sns boxplot representation of smoking_status : age
Source: Author's calculation

In the plot, on the X-axis we can see four numbers which represent the smoking status of the demographics. The smoking mapping is done so that the 0 represents never smoked, 1 represents formerly smoked, 2 represent active smoking and 3 is unknow.

After these visualization techniques, the data needs to be converted to numerical counterparts so that the models could learn of the numerical representation of those values. Instead of having a gender column we can opt to have a column named Male which will have as it values either 0 or 1 as know as a True False statement. In the same method we will also be

replacing the other values with numerical numbers, which can be seen in table of context below, in tables below.

Table 1. Gender

Gender	Gender (numerical)
Male	0
Femle	1

Source: Author's calculation

Table 2. Smoking status

Smoking status	Smoking_status (numerical)
Never smoked	0
Formerly smoked	1
Active smoking	2
Unknown	3

Source: Author's calculation

Table 3. Married status

Married	Married (numerical)
Yes	0
No	1

Source: Author's calculation

Table 4. Residence status

Residence Type	Residence Type (numerical)
Urban	0
Rural	1

Source: Author's calculation

After converting all labels to numerical data, we are left with this table as seen as in Figure 7. which concluded the part of preparing the data for training.

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	smoking_status	stroke	Male	Married	Urban
0	9046	67.0	0	1	228.69	36.600000	1	1	1	1	1
1	51676	61.0	0	0	202.21	28.893237	0	1	0	1	0
2	31112	80.0	0	1	105.92	32.500000	0	1	1	1	0
3	60182	49.0	0	0	171.23	34.400000	2	1	0	1	1
4	1665	79.0	1	0	174.12	24.000000	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...

Figure 7. Numerical representation of data
Source: Author's calculation

2.2. Model Training

The training of Machine Learning models in this study focused on two prominent algorithms: Logistic Regression and Random Forests. The Random Forest (RF) algorithm for regression and classification has considerably gained popularity since its introduction in 2001. Meanwhile, it has grown to a standard classification approach competing with Logistic Regression in many innovation-friendly scientific fields (Couronne et al., 2018).

These models were chosen for their ability to handle binary classification tasks effectively and their interpretability in the case of Logistic Regression, and their capacity to capture

complex nonlinear relationships in the case of Random Forests. A major challenge in big and high-dimensional data analysis is related to the classification and prediction of the variables of interest by characterizing the relationships between the characteristic factors and predictors (Hajipour et al., 2020).

Fitting a Logistic Regression model to the preprocessed dataset was the training step for Logistic Regression. With the goal of determining the ideal coefficients that define the decision boundary dividing stroke occurrence cases from non-occurrence, the model was trained using the training set. In order to minimize the logistic loss function, the model's parameters were iteratively modified during training using gradient descent optimization techniques.

The regularization parameter, which regulates the amount of regularization used to the model to minimize overfitting, was optimized by hyperparameter tuning. To evaluate the model's performance over a number of folds and guarantee its generalizability and robustness, cross-validation techniques were utilized.

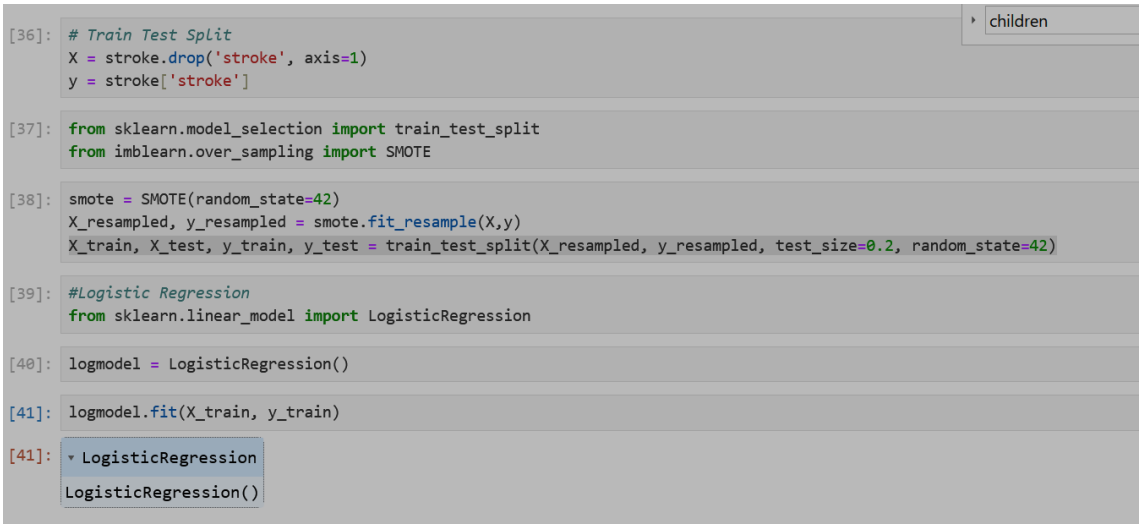
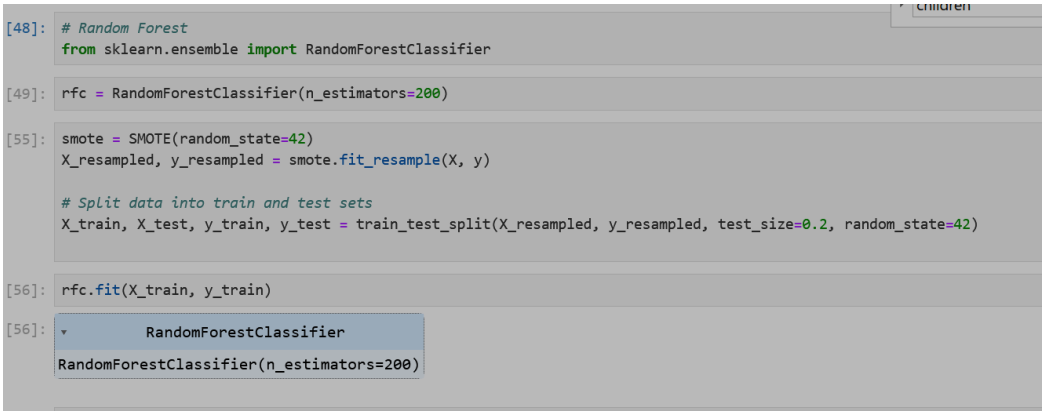


Figure 8. Logistic Regression with Smote
Source: Author's calculation

During model training, the Synthetic Minority Over-sampling Technique (SMOTE) was used to resolve the class imbalance in the dataset. When working with imbalanced data, SMOTE resampling often achieves better performance in predicting stroke occurrence (Wu & Fang, 2020). SMOTE oversamples the minority class by taking each minority class sample and injecting synthetic examples along the line segments joining any/all the k minority class nearest neighbors, resulting in the minority class being more generic (Jing, 2022). By interpolating between current samples, SMOTE creates synthetic samples for the minority class, balancing the class distribution and lessening the effect of class imbalance on model performance. SMOTE increases the model's forecast accuracy for the incidence of strokes and strengthens its capacity to learn from minority class events by oversampling the minority class. The Logistic Regression and Random Forest models were subsequently trained using the oversampled dataset, guaranteeing a more balanced representation of stroke and non-stroke cases in the training data.

Training of the Random Forest model involved the construction of an ensemble of decision trees, where each tree was trained on a bootstrapped subset of the training data. During the training process, feature subsets were randomly sampled at each split to introduce diversity among the trees and reduce overfitting. Hyperparameter tuning was performed to optimize the

number of trees in the forest, the maximum depth of each tree, and the minimum number of samples required to split a node.



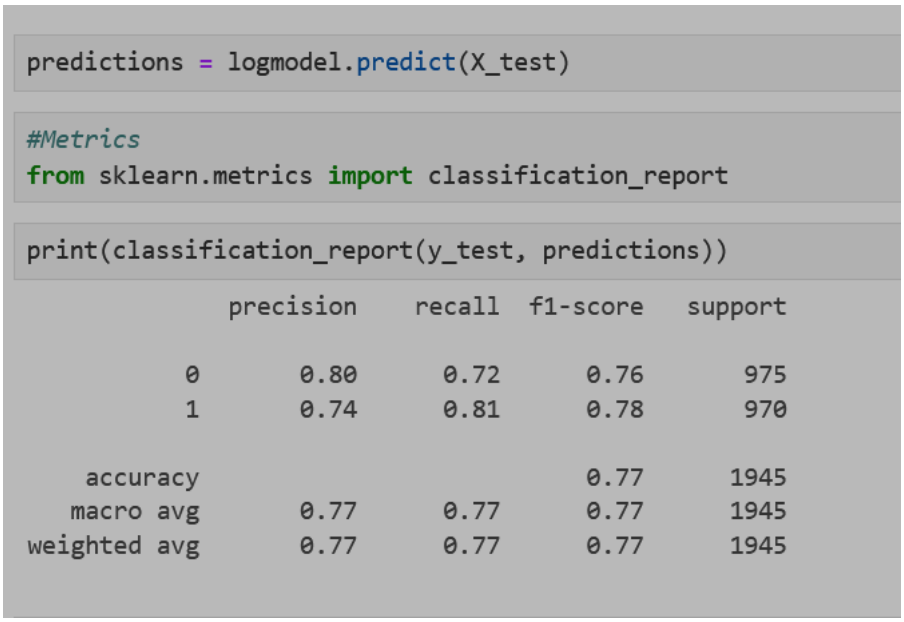
```
[48]: # Random Forest
      from sklearn.ensemble import RandomForestClassifier

[49]: rfc = RandomForestClassifier(n_estimators=200)

[55]: smote = SMOTE(random_state=42)
      X_resampled, y_resampled = smote.fit_resample(X, y)

      # Split data into train and test sets
      X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.2, random_state=42)

[56]: rfc.fit(X_train, y_train)

[56]: RandomForestClassifier
      RandomForestClassifier(n_estimators=200)
```

Figure 8. Random Forest Classifier with SMOTE
Source: Author's calculation

3. Results

Predicting the risk of stroke is a crucial task that can potentially save lives. Two widely used Machine Learning models, Logistic Regression and Random Forests, offer distinct approaches to tackle this challenge. Logistic Regression, a straightforward statistical model, excels in accurately identifying individuals at high risk of stroke. Its precision and recall scores are impressive, meaning it minimizes both false positives and false negatives. However, Logistic Regression 's linear decision boundary may struggle to capture intricate relationships between various risk factors contributing to stroke occurrence.

Printing out the classification report and confusion matrix we can see the results of the Logistic Regression algorithm.



```
predictions = logmodel.predict(X_test)

#Metrics
from sklearn.metrics import classification_report

print(classification_report(y_test, predictions))
```

	precision	recall	f1-score	support
0	0.80	0.72	0.76	975
1	0.74	0.81	0.78	970
accuracy			0.77	1945
macro avg	0.77	0.77	0.77	1945
weighted avg	0.77	0.77	0.77	1945

Figure 9. Classification report Logistic Regression
Source: Author's calculation

On the other hand, Random Forests employ a sophisticated ensemble learning technique that combines multiple decision trees. This approach allows Random Forests to effectively model complex, nonlinear interactions among features, potentially uncovering hidden patterns missed by simpler models. Random Forests consistently demonstrate robust performance across multiple evaluation metrics, making them a powerful tool for stroke risk assessment. While the interpretability of individual decision trees within the ensemble may be somewhat compromised, the overall predictive accuracy of Random Forests often outweighs this concern, particularly in high stakes medical applications where accurate risk estimation is paramount. The classification for the Random Forest is better than the Logistic Regression as seen in Figure 10.

```
print(classification_report(y_test, rfc_pred))
```

	precision	recall	f1-score	support
0	0.96	0.91	0.93	975
1	0.91	0.96	0.94	970
accuracy			0.94	1945
macro avg	0.94	0.94	0.94	1945
weighted avg	0.94	0.94	0.94	1945

```
confusion_matrix(y_test, predictions)
```

```
array([[770, 205],
       [165, 805]], dtype=int64)
```

Figure 10. Classification report Random Forest
Source: Author's calculation

It is noticeable that the precision, recall score as others are better in Random Forest algorithm then the Logistical Regression which is to be expected.

The model works effectively in recognizing cases of no stroke, as seen by the relatively high precision, recall, and F1-score for class 0 (no stroke). More specifically, the precision of 0.96 indicates that 96% of the time the model is right when it predicts an instance to not have a stroke. With a recall of 0.91, the model accurately predicts 91% of all cases in which there is no stroke. For class 0, a balanced indicator of the model's performance is provided by the F1-score, which takes precision and recall into account.

In the same manner, class 1 (stroke) has high precision, recall, and F1-score values, suggesting good performance in detecting stroke cases. With a precision of 0.91, the model is 91% accurate 91% of the time when it predicts that an incident would involve a stroke. With a recall of 0.96, the model is able to accurately identify 96% of all stroke cases. Class 1 has a strong F1-score as well, suggesting a balance between recall and precision.

This systematic review is limited by small sample sizes and suboptimal data sources for the included models, and thus the reported model performance may be overly optimistic (Zu et al, 2023).

Overall, the Random Forest model's high accuracy (0.94) and evenly distributed performance in both groups show how well it is at forecasting the occurrence of strokes. One of the greatest strengths of ML is its ability to endlessly process data and tirelessly perform an iterative task (Mainali et al., 2021).

3. Conclusion

The goal in this work was to create predictive models that would identify stroke incidence based on personal and medical characteristics. We trained and assessed classification models using a range of Machine Learning approaches, leveraging an extensive dataset from a survey of stroke patients from Kaggle.com.

We prepared the dataset for model training by carefully preparing the data, which included handling missing values, encoding categorical variables, and utilizing SMOTE to resolve class imbalance. The preprocessed data was then used to train Random Forest and Logistic Regression classifiers.

By calculating precision, recall, and F1-score values, the Logistic Regression model was able to attain an accuracy of 77%, according to the results. Even though Logistic Regression performed rather well, a Random Forest classifier was trained in order to undertake additional research.

In comparison to Logistic Regression, the Random Forest model fared better, attaining a 94% accuracy rate with precision, recall, and F1-score values of 0.91 and 0.96 for both classes. The Random Forest model performs better than other models because it can capture intricate correlations between input features, which makes it a good choice for jobs involving nonlinear classification. Furthermore, by reducing the impact of class imbalance during model training, SMOTE was applied, which improved the models' resilience and generalization capacities.

In summary, this study shows how Machine Learning techniques can be used to predict the risk of stroke by taking into account several demographic and health-related variables. Machine learning models are feasible in predicting early stroke outcomes. An enriched feature bank could improve model performance (Su et al., 2022). By helping healthcare professionals identify and intervene early, the created model can improve patient outcomes and lessen the burden of stroke-related morbidity and mortality. Ideally, the treating physician could predict the outcome under different circumstances beforehand and thus make an informed treatment decision (Maier et al., 2016). In order to improve the model's predictive capability and ease its application in clinical settings, more investigation should be conducted into the incorporation of other features and optimization strategies. AI tools could play key role in the management of the post-stroke patients. Specifically, state-of-the-art ML models have already been used for the prediction of the functional outcomes in the rehabilitation of post-stroke patients (Kokkotis et al, 2022).

References

Bonkhoff, A. K., & Grefkes, C. (2022). Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence. Brain, 145(2), 457-475.

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC bioinformatics, 19, 1-14.

Fernandez-Lozano, C., Hervella, P., Mato-Abad, V., Rodríguez-Yáñez, M., Suárez-Garaboa, S., López-Dequidt, I., Estany-Gestal, A., Sobrino, T., Campos, F., Castillo, J., Rodríguez-Yáñez, S., & Iglesias-Rey, R. (2021). Random forest-based prediction of stroke outcome. Scientific reports, 11(1), 10071. https://doi.org/10.1038/s41598-021-89434-7

Hajipour, F., Jozani, M. J., & Moussavi, Z. (2020). A comparison of regularized Logistic Regression and Random Forest Machine Learning models for daytime diagnosis of obstructive sleep apnea. Medical & Biological Engineering & Computing, 58(10), 2517–2529. doi:10.1007/s11517-020-02206-9

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (15.03.2024.)

Jing, Y. (2022). Machine Learning Performance Analysis to Predict Stroke Based on Imbalanced Medical Dataset. In CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms (pp. 1-7). Nanjing, China.

Kokkotis, C., Moustakidis, S., Giarmatzis, G., Giannakou, E., Makri, E., Sakellari, P., ... & Aggelousis, N. (2022). Machine Learning Techniques for the Prediction of Functional Outcomes in the Rehabilitation of Post-Stroke Patients: A Scoping Review. BioMed, 3(1), 1-20. https://doi.org/10.3390/biomed3010001

Maier, O., & Handels, H. (2016). Predicting Stroke Lesion and Clinical Outcome with Random Forests. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (pp. 219–230). Springer International Publishing. https://doi.org/10.1007/978-3-319-55524-9_21

Mainali, S., Darsie, M. E., & Smetana, K. S. (2021). Machine learning in action: stroke diagnosis and outcome prediction. Frontiers in neurology, 12, 734345. https://doi.org/10.3389/fneur.2021.734345

Merdas, H. M. (2024). Elastic Net–MLP–SMOTE (EMS)-Based Model for Enhancing Stroke Prediction. Medinformatics, 1(2), 73-78.

Poorani, K., Karuppasamy, M., Jansi Rani, M., & Prabha, M. (2023). Classifier Comparison for Stroke Prediction Ensembling SMOTE+ENN using Machine Learning Approach. Research Square. https://doi.org/10.21203/rs.3.rs-1675863/v1

Su, P. Y., Wei, Y. C., Luo, H., Liu, C. H., Huang, W. Y., Chen, K. F., ... & Lee, T. H. (2022). Machine learning models for predicting influential factors of early outcomes in acute ischemic stroke: registry-based study. JMIR Medical Informatics, 10(3), e32508. https://doi.org/10.2196/32508

Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., ... & Bray, B. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. PloS one, 15(6), e0234722. https://doi.org/10.1371/journal.pone.0234722

Wu, Y., & Fang, Y. (2020). Stroke prediction with machine learning methods among older Chinese. International journal of environmental research and public health, 17(6), 1828. https://doi.org/10.3390/ijerph17061828

Zu, W., Huang, X., Xu, T., Du, L., Wang, Y., Wang, L., & Nie, W. (2023). Machine learning in predicting outcomes for stroke patients following rehabilitation treatment: A systematic review. Plos one, 18(6), e0287308. https://doi.org/10.1371/journal.pone.0287308

© 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).