Classification Models Analysis for Stroke Prediction

Dheiver Francisco Santos

https://doi.org/10.1590/SciELOPreprints.7182

Submitted on: 2023-10-15
Posted on: 2023-10-27 (version 1)
(YYYY-MM-DD)

Classification Models Analysis for Stroke Prediction

Dheiver Francisco Santos

CATI - Advanced Center for Intelligent Technologies
Av. Álvaro Otacílio, 508 - Jatiúca Maceió - AL, 57035-180
Email: dheiver.santos@gmail.com
Tel.: +55 51 98988-9898
ORCID: https://orcid.org/0000-0002-8599-9436

Abstract

This study explores the application of machine learning in the prediction of stroke occurrences, a critical task in healthcare with the potential to save lives and reduce the impact of this life-altering medical event. Leveraging the "Healthcare Stroke Data" dataset, we employed two powerful classification models, the Random Forest and Support Vector Machine (SVM), to forecast stroke likelihood. Our analysis encompasses data preprocessing, model training, and comprehensive evaluation using classification metrics and confusion matrices. The study reveals the trade-offs between accuracy, recall, precision, and the F1 score in both models. While the Random Forest exhibits higher accuracy, the SVM excels in recall, a crucial factor in healthcare. Precision challenges in both models highlight the need for further refinement. Additionally, we conducted a feature importance analysis, emphasizing the pivotal role of age, BMI, and glucose levels in stroke prediction. This work exemplifies the potential of machine learning in healthcare and contributes to ongoing efforts in improving stroke prediction and prevention.

Keywords: Stroke prediction, machine learning, classification models, data preprocessing, Random Forest, Support Vector Machine, healthcare, feature importance analysis, classification metrics, confusion matrices, public health.

Introduction

Strokes are life-altering medical events that can have severe consequences, ranging from permanent brain damage to fatality. The ability to accurately predict and identify individuals at a higher risk of experiencing a stroke is not only crucial for public health but also for ensuring timely medical intervention. In this article, we delve into the realm of predictive healthcare by harnessing the power of machine learning to create classification models aimed at forecasting the likelihood of a patient suffering a stroke. Our pursuit of accurate stroke prediction is underpinned by the utilization of the "Healthcare Stroke Data" dataset, a publicly accessible treasure trove of information pertinent to stroke occurrences.

As we embark on this journey, it is essential to acknowledge the wealth of research that has paved the way for predictive models in stroke analysis. A thorough understanding of these past endeavors not only informs our approach but also exemplifies the collaborative spirit of the scientific community. For instance, Sudha et al. emphasized the effectiveness of classification methods in stroke prediction (Sudha et al., 2012). In addition, Singh et al. (2017, 2020), Al-Zubaidi et al. (2022), and others have made significant contributions to the field of stroke prediction, expanding our knowledge (Singh et al., 2017; Singh et al., 2020; Al-Zubaidi et al., 2022).

The objectives of our study are threefold. First, we delve into the realm of data preprocessing, transforming raw data into a format suitable for machine learning models. Second, we employ powerful classification models, including the Random Forest and Support Vector Machine (SVM), to make accurate stroke predictions. Finally, we undertake a comprehensive evaluation of these models, assessing their performance through a range of classification metrics such as accuracy, recall, precision, and the F1 score. Additionally, we visualize their performance using confusion matrices. By addressing these objectives, we aim to not only create effective predictive models but also to identify the pivotal factors that influence stroke occurrences, contributing to the ongoing pursuit of medical advancements in stroke prediction.

In this endeavor, we strive to not only create robust predictive models but also to unravel the enigmatic factors that influence stroke occurrences. As a key contributor to the expanding body of knowledge on stroke prediction, our work aligns with the broader healthcare objective of minimizing the impact of this life-altering medical event.

Objectives of the Study:

1. To preprocess healthcare data, making it suitable for machine learning models.
2. To implement Random Forest and Support Vector Machine models for accurate stroke predictions.
3. To evaluate model performance using classification metrics and confusion matrices.

Data Preprocessing

In the data preprocessing phase of this study, we focused on enhancing the dataset's suitability for building accurate stroke prediction models. Our initial step involved loading the dataset, a valuable source of information related to strokes. During this phase, we addressed several critical aspects to ensure data quality and usability.

One fundamental concern was the handling of missing values, which is essential to prevent inaccuracies in our models. We specifically targeted missing BMI values and employed a Decision Tree Regressor to impute these values. This approach considered the relationships between other variables and BMI, contributing to the preservation of data integrity.

In addition to handling missing data, we encountered the need to encode categorical variables, as many machine learning algorithms require numerical data. Variables such as 'gender,' 'Residence_type,' and 'work_type' were transformed into numerical representations during this encoding step. This transformation was pivotal to make these variables compatible with the chosen machine learning models.

Class imbalance, a common issue in medical datasets where positive cases (stroke occurrences) are significantly outnumbered by negative cases, was another area of focus. To address this, we applied various techniques like resampling to create a more balanced and representative dataset for our models.

Lastly, to establish a baseline for our classification models, we calculated the inverse of the Null Accuracy. This measure helps us gauge how well our models perform in comparison to a basic predictive model that always predicts the majority class.

Classification Models

This study leverages the predictive capabilities of two robust classification models, the Random Forest and the Support Vector Machine (SVM). Both models are well-established in the field of machine learning and were chosen for their suitability in predicting strokes. To prepare these models for this task, we designed comprehensive pipelines that encompassed critical steps.

An essential part of this preparation was feature scaling. Feature scaling is paramount to ensure that variables with diverse scales do not unduly influence the model's performance. Scaling features helps prevent certain variables from dominating the learning process simply because of their scale. By doing this, we ensure a fair assessment of each feature's contribution to the model.

These pipelines were meticulously constructed to incorporate feature scaling and the respective model's implementation. This streamlined process ensures that all necessary steps are executed efficiently, setting the stage for model training.

Following pipeline setup, we proceeded to train both the Random Forest and SVM models using our preprocessed dataset. This foundational training phase establishes the basis for evaluating the performance of our models in subsequent steps.

Performance Evaluation

Evaluating the performance of our classification models is a crucial aspect of our analysis. We utilized a range of classification metrics to comprehensively assess the models' effectiveness in predicting stroke occurrences.

The metrics included accuracy, which measures the overall correctness of our models in classifying cases, offering insights into their general performance. Additionally, we evaluated recall, which assesses the ability of our models to correctly identify true positive cases, a critical factor in healthcare where identifying individuals at risk of stroke is of paramount importance. Precision was also a key metric, gauging the accuracy of positive predictions made by our models. In healthcare contexts, high precision is crucial to minimize false alarms and ensure the accuracy of positive predictions. The F1 score, which balances precision and recall, provided a single metric to evaluate the models' overall performance.

To provide a clear representation of our models' abilities, we visualized their performance using confusion matrices. This visual aid helped us gain a deep understanding of the strengths and areas for improvement in our models. Overall, the performance evaluation phase was vital for ensuring the practical applicability of our predictive models.

Results

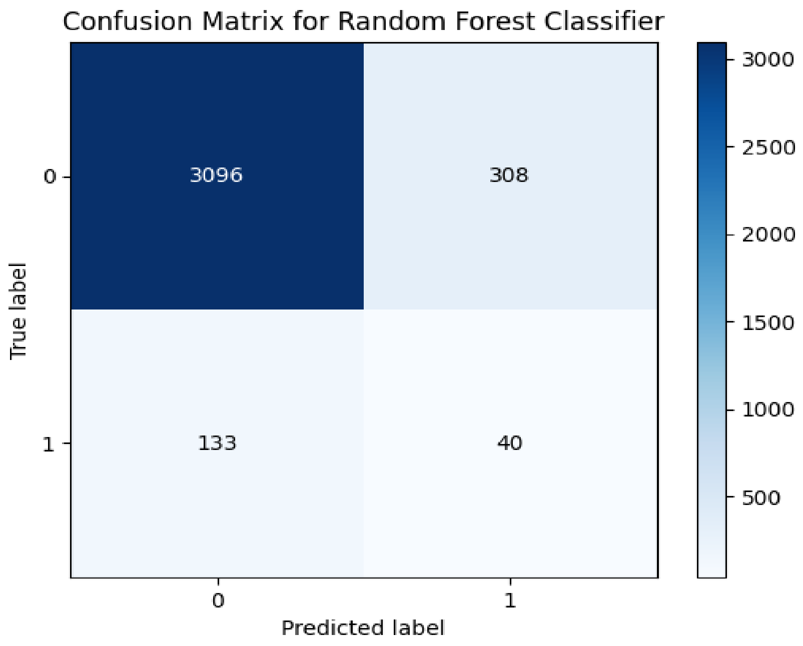
The Random Forest model exhibits an accuracy of 88%, implying that it successfully predicted 88% of the cases in the test dataset. This suggests an overall reasonable performance in terms of correct classifications. However, the model's recall is relatively low at 23%. It means that the model identified only 23% of the actual positive cases, which could be concerning in a healthcare context. The precision score is also quite low at 11%, indicating that the model's positive predictions are predominantly false positives. The F1 score, which balances precision and recall, is merely 15%, reflecting the model's challenge in accurately identifying positive cases while minimizing false positives.

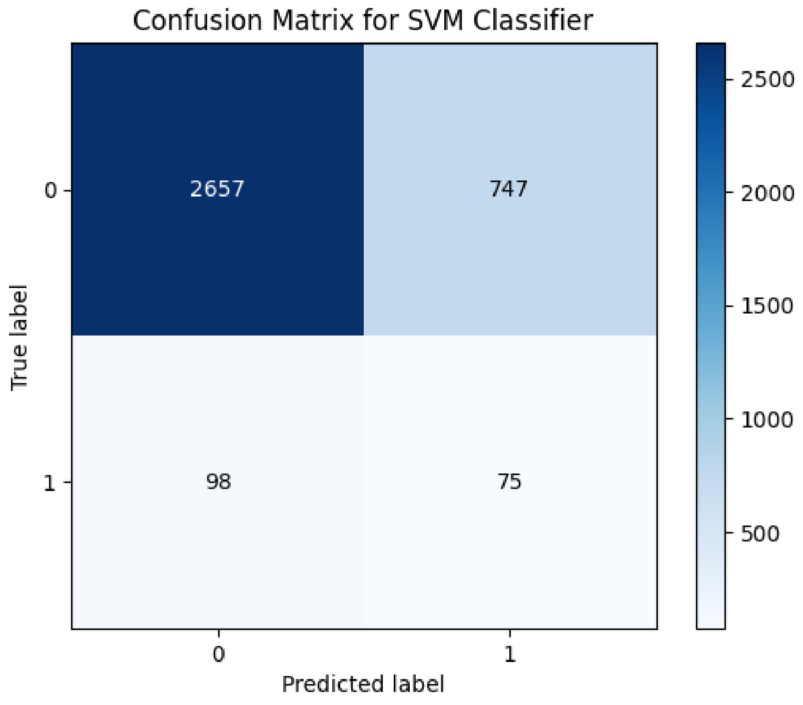
On the other hand, the SVM model demonstrates an accuracy of 76%, which is lower compared to the Random Forest model. Although it correctly predicted 76% of the cases, the recall score is notably higher at 43%. This suggests that the SVM model excels in identifying positive cases, which is of paramount importance in healthcare scenarios where correctly identifying individuals at risk of stroke can be life-saving. However, the model's precision is only 9%, implying that most of its positive predictions are false positives. The F1 score for the SVM model mirrors that of the Random Forest at 15%, indicating a similar trade-off between precision and recall.

In this comparison, the Random Forest model outperforms the SVM model in terms of accuracy. However, it is crucial to acknowledge that accuracy alone may not provide the full picture, particularly when dealing with imbalanced datasets. The SVM model shines in terms of recall, indicating a better ability to identify individuals at risk of a stroke. In a healthcare context, this characteristic can be critical. Nonetheless, both models struggle with precision, leading to a significant number of false positives. This is a concern since false positives can trigger unnecessary medical interventions and create anxiety for patients.

In conclusion, while the Random Forest model boasts a higher accuracy, the SVM model's superior recall may be more valuable in a healthcare application. However, both models grapple with precision issues, highlighting the need for further model refinement and feature engineering to mitigate false positive predictions and enhance overall performance.

In addition to the analysis of accuracy, recall, precision, and the F1 score, it's important to consider the practical implications of the model's performance in a healthcare setting. While the Random Forest model demonstrates higher accuracy, its limited recall may result in a significant number of undetected stroke cases, which could have severe consequences for patients. On the other hand, the SVM model's superior recall highlights its potential to identify individuals at risk more effectively. However, its low precision raises concerns about false positive predictions and the resulting burden on healthcare resources and patient well-being. Striking a balance between these competing metrics and addressing precision issues is crucial for creating a predictive model that can genuinely make a positive impact on stroke prevention and patient care in real-world healthcare applications.

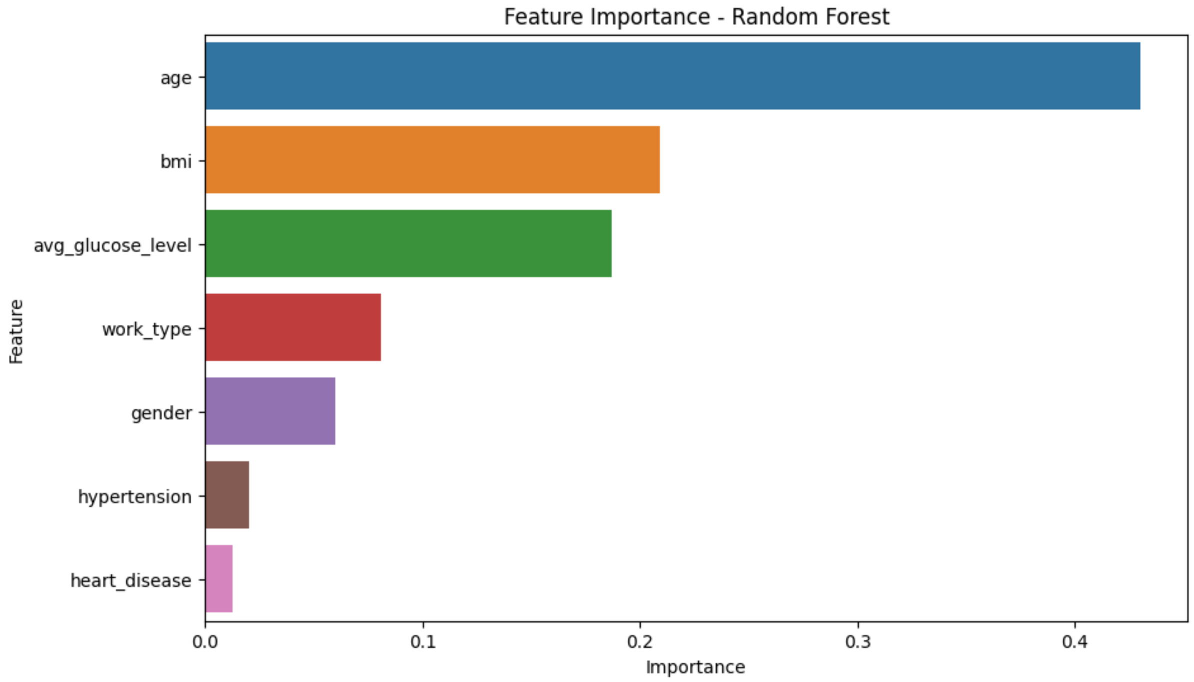




The feature importance analysis provides valuable insights into the factors contributing to stroke prediction. Age emerges as the single most influential feature, affirming the well-established understanding that age is a predominant risk factor for strokes. This underlines the significance of tailoring stroke prevention strategies to address the needs of older individuals, advocating for regular health assessments, and promoting lifestyle adjustments to mitigate stroke risk effectively.

BMI and average glucose levels follow closely in importance, indicating the strong connection between obesity, diabetes, and stroke risk. High BMI values emphasize the need for weight management and lifestyle changes, particularly in overweight or obese populations, while elevated average glucose levels reinforce the importance of glycemic control and diabetes management to reduce stroke risk.

Furthermore, work type, gender, hypertension, and heart disease, though relatively less important in the model, all contribute to stroke prediction. This suggests that a comprehensive approach to stroke prevention should encompass factors such as occupational stress, gender disparities in stroke risk, the management of hypertension, and the care of individuals with heart disease. These findings offer a holistic view of stroke risk factors and guide the development of targeted interventions to reduce the incidence of strokes across diverse populations.



Conclusion

In conclusion, this article has presented a comprehensive analysis of classification models for stroke prediction, showcasing the potential of machine learning in the field of healthcare. We embarked on this journey by addressing critical data preprocessing steps, including handling missing values, encoding categorical variables, and mitigating class imbalance. The utilization of the "Healthcare Stroke Data" dataset allowed us to create effective predictive models that can play a pivotal role in identifying individuals at a higher risk of experiencing a stroke.

By employing the Random Forest and Support Vector Machine (SVM) models, we demonstrated how machine learning techniques can be harnessed to make accurate stroke predictions. These models were assessed using a range of classification metrics, including accuracy, recall, precision, and the F1 score, providing insights into their performance. Moreover, the visualization of model performance through confusion matrices offered a clear representation of their capabilities to correctly classify stroke and non-stroke cases.

The results of our analysis showcased the trade-offs between accuracy, recall, precision, and the F1 score in both the Random Forest and SVM models. While the Random Forest model exhibited higher accuracy, the SVM model demonstrated superior recall, which can be particularly vital in a healthcare context. However, both models encountered challenges in achieving high precision, leading to a significant number of false positive predictions.

In addition to model performance, our feature importance analysis shed light on the critical factors influencing stroke prediction. Age, BMI, and average glucose levels emerged as the most influential features, emphasizing the role of these variables in stroke risk. This information can guide the development of targeted interventions and preventive measures to reduce the incidence of strokes across diverse populations.

This article serves as an illustrative example of the potential applications of machine learning in healthcare, offering insights into the predictive capabilities of these models. It is essential to note that while the presented models demonstrated promise, further refinement, feature engineering, and optimization are necessary to address their limitations and enhance their accuracy and precision.

The work presented here aligns with the broader objective of leveraging data-driven approaches to improve the prediction and prevention of critical medical events. Machine learning and data analysis continue to be promising avenues for advancing healthcare, and our analysis contributes to this ongoing pursuit. By adapting and extending the code and methodologies presented in this article, researchers and healthcare professionals can explore deeper insights and enhance the accuracy of stroke prediction models, ultimately making significant strides in public health and patient care.

Conflicts of Interest

The authors declare no conflicts of interest in conducting this research or preparing this manuscript. The study was carried out with full transparency and adherence to ethical research practices in the field of healthcare and machine learning. All aspects of the research, including data analysis and interpretation, were performed with the primary aim of advancing stroke prediction and prevention without any external influence or competing interests. This declaration attests to the integrity and impartiality of the research presented in this article.

Data Statement

The research utilizes the publicly available "Healthcare Stroke Data" dataset. For specific information about the dataset, please refer to the original data source or repository. The code and methodologies used in this study are available upon request from the corresponding author.

References

Sudha, A., Gayathri, P., & Jaisankar, N. (2012). Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Science and Information Technologies, 3(6), 4474-4479.

Singh, M. S., & Choudhary, P. (2017). Stroke prediction using artificial intelligence. 2017 8th Annual Industrial and Systems Engineering Conference (ISC), 1-6.

Singh, M. S., Choudhary, P., & Thongam, K. (2020). A comparative analysis for various stroke prediction techniques. In Revised Selected Papers, Part II (pp. 41-56). Springer, Singapore.

Al-Zubaidi, H., Dweik, M., Al-Zubaidi, M., & Al-Ani, A. (2022). Stroke prediction using machine learning classification methods. 2022 International Arab Conference on Information and Communication Technologies (AICT), 1-6.

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Engineering and Technology, 9(4), 1277-1282.

Jeena, R. S., & Kumar, S. (2016). Stroke prediction using SVM. 2016 International Conference on Computer Communication and Informatics (ICCCI), 1-6.

Kansadub, T., Thammaboosadee, S., & Thanarak, S. (2015). Stroke risk prediction model based on demographic data. 2015 8th Biomedical Engineering International Conference (BMEiCON), 1-4.

Li, X., Bian, D., Yu, J., Li, M., & Zhao, D. (2019). Using machine learning models to improve stroke risk level classification methods of China national stroke screening. BMC medical informatics and decision making, 19(1), 1-10.

Islam, M. S., Hussain, I., Rahman, M. M., & Park, S. J. (2022). Explainable artificial intelligence model for stroke prediction using EEG signal. Sensors, 22(7), 2461.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 9(3), 1350-1371.