Journal of Research in Engineering and Computer Sciences
February 2024, Vol. 2, No. 1, pp. 61-72

# Stroke Prediction Using Machine Learning

Niharika Patil and Alex Sumarsono

1. Department of Engineering, California State University, East Bay, Hayward, United States

**Abstract:**
Stroke, a cerebrovascular event, represents a significant global health concern due to its substantial impact on morbidity and mortality. It occurs when there is a sudden interruption or reduction of blood supply to the brain, leading to the impairment of brain function. As the second leading cause of death globally, stroke demands urgent attention, and early detection is pivotal for effective intervention. This study addresses the global health concern of strokes by leveraging machine learning models for early detection and risk assessment. The study employs logistic regression, random forest, naive Bayes, and support vector machine algorithms to create a robust predictive model. Key objectives include data cleaning, addressing class imbalance, model evaluation, and deployment. The research contributes to the growing literature on machine learning applications in healthcare by presenting a holistic approach to stroke prediction. Results indicate that while random forest achieves high accuracy, logistic regression provides a balanced sensitivity-specificity trade-off. The models are deployed through an interactive Shiny app, enhancing accessibility and usability for healthcare professionals. Future work involves refining models, incorporating additional features.

*Keywords: Stroke, machine learning models, predictive model, risk assessment, Shiny app deployment.*

## INTRODUCTION

In recent years, the intersection of healthcare and machine learning has presented unprecedented opportunities for enhancing predictive analytics and improving patient outcomes [1]. Among the myriad of medical conditions, stroke stands out as a leading cause of morbidity and mortality worldwide [2]. Early detection and timely intervention are critical factors in mitigating the devastating effects of strokes. This paper explores a machine learning approach to stroke prediction.

Stroke, characterized by a sudden interruption of blood flow to the brain, poses a significant public health challenge [3]. Machine learning models have shown promise in analyzing complex patterns within large datasets, facilitating the identification of subtle risk factors, and improving the accuracy of predictive models [4].

Key objectives of this study include:

- To perform data cleaning and preparation including splitting data into training and testing.
- To deal with class imbalance.
- To evaluate and select the best suited predictive model.
- To deploy the model for use.

This paper contributes to the growing literature on machine learning applications in healthcare by presenting a holistic approach to stroke prediction. The remaining sections of this paper are structured as follows: Section II delves into the existing literature on the subject. Section III outlines the methodology employed in this study. Section IV provides a detailed analysis of the results; Section V shows web application deployment of the model and Section VI concludes the paper.

## RELATED WORK

The logistic regression model, recognized as the logit model, is extensively utilized in classification and predictive analytics. The estimation of the probability of a specific event occurrence, such as the likelihood of an individual having a stroke, is conducted based on a dataset of independent variables. The outcome is a probability confined within the 0 to 1 range, with predictions for binary classification made by interpreting probabilities - 0.5 or less predicts 0, while over 0 predicts 1. The random forest algorithm, a frequently employed machine learning approach, amalgamates outputs from multiple decision trees to derive a single result. An ensemble of decision trees, each constructed from a bootstrap sample drawn with replacement from a training set, constitutes the random forest. Feature bagging introduces randomness to enhance dataset diversity and mitigate correlation among decision trees.

Naive Bayes, grounded in Bayes' Theorem, assumes independence among predictors, rendering it valuable for text classification. Despite simplifying assumptions, Naive Bayes classifiers excel due to their efficiency, particularly when coupled with kernel density estimation for scenarios with undefined data distributions. A powerful supervised learning algorithm, Support Vector Machine (SVM), is employed for classification and regression tasks. The hyperplane that best separates classes in a high-dimensional space is sought by SVM. Its suitability for the intricate nature of stroke prediction is attributed to its ability to handle non-linear relationships and complex decision boundaries. Upon the foundation of machine learning models developed [5],[6],[7],[8] and extensive research in the domain of predictive modeling for stroke risk assessment [9],[10],[11], this study is built. Various machine learning techniques to forecast the likelihood of stroke based on diverse sets of input features have been explored in previous investigations. Studies employing logistic regression, decision trees, support vector machines, and ensemble methods such as random forests [12],[13] have been employed, demonstrating the potential for accurate stroke prediction and often emphasizing the significance of feature selection and model interpretability.

Additionally, the integration of advanced techniques such as deep learning and ensemble learning has been explored to enhance predictive accuracy [1]. Some studies have focused on the integration of electronic health records (EHRs) and comprehensive patient histories to improve model robustness and generalizability [14].

In this context, the extension of the application of predictive modeling to a Shiny app interface is contributed by our research, offering a user-friendly tool for healthcare professionals. The ongoing discourse in stroke prediction research is enriched by the synthesis of insights from previous studies and the novel deployment on an interactive platform.

## METHODOLOGY

The methodology employed in this study comprised a structured seven-step process, as depicted in Figure 1: data understanding, data preprocessing, exploratory data analysis, data preparation, model building, model evaluation, and deployment.

### Data Understanding

The initial step involved understanding of the stroke prediction dataset sourced from Kaggle [11]. This encompassed an examination of data distribution, identification of data types, and investigation of relationships between variables. The dataset comprises information on 5110 individuals, with the response variable exhibiting binary nature (∈ {0,1}), indicating whether a person is prone to a stroke or not. A thorough understanding of the problem, encompassing the target variable, features, and their interrelationships, is essential for guiding subsequent steps in the analysis.

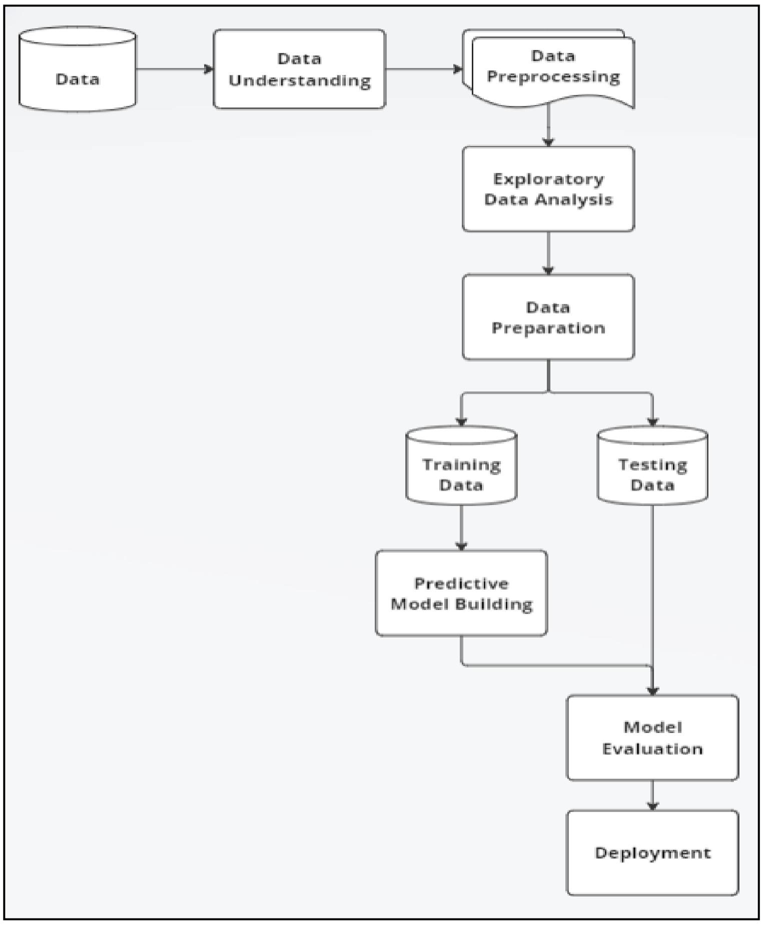


**Fig. 1: Methodology**

### Data Preprocessing

To ensure data integrity, identification, and subsequent resolution of missing values and outliers were conducted. Data preparation encompassed the cleaning of data through the removal of duplicate rows, correction of errors, and transformation of data into a suitable format. The NA values observed in the bmi column were addressed through mean matching imputation using the MICE package. Columns containing categorical values underwent an update to factors.

### Exploratory Data Analysis

Exploratory data analysis played a pivotal role in understanding the data's characteristics and patterns. The bivariate analysis showed preliminary insights on stroke prediction for each column, the potential predictor variables as follows.

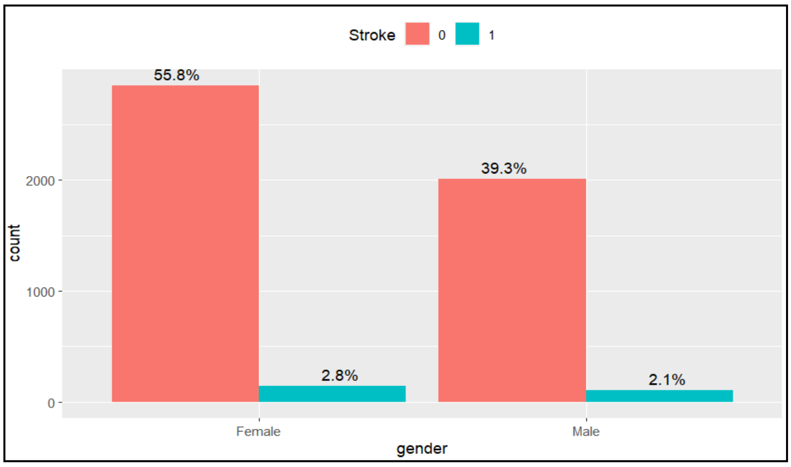


**Fig. 2: Stroke occurrence in gender**

In this study, it was found out that 58.6% (2994) were female, out of which 2.8% had stroke and 41.4% (2115) were male, out of which 2.1% had stroke.

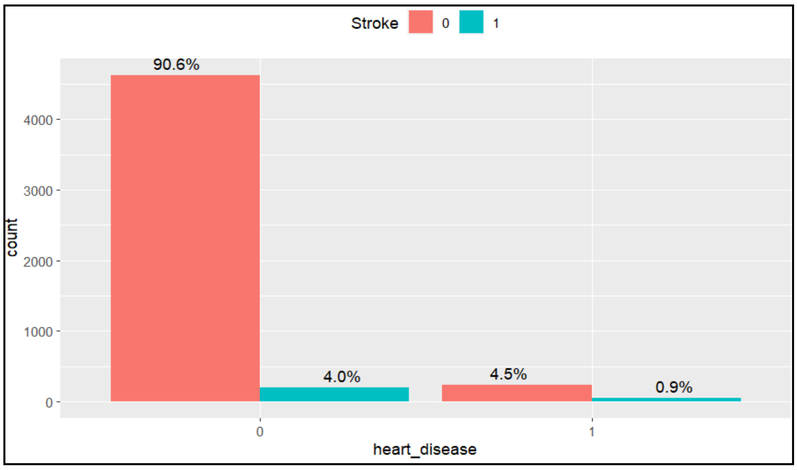


**Fig. 3: Stroke Occurrence in cardiac patients**

The individuals having heart disease were only 5.4% (276). The stroke occurrence was higher in individuals with no heart disease.

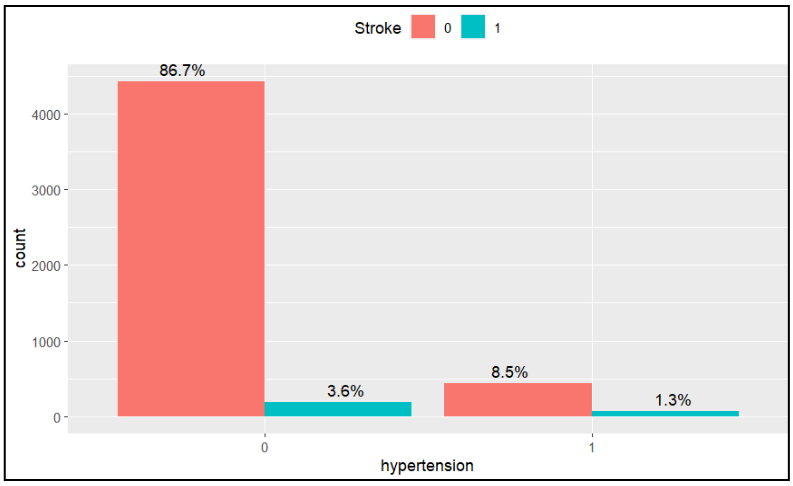


**Fig. 4: Stroke occurrence in hypertension patients**

The individuals having hypertension were only 9.8% (498). The stroke occurrence was higher in individuals with no hypertension.

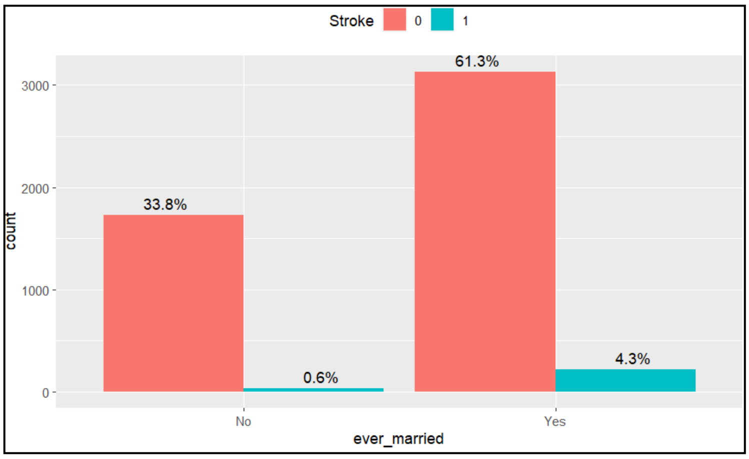


Fig. 5: Stroke occurrence in married people

The married individuals were 65.6% (3353), and stroke occurrence was higher in this group.

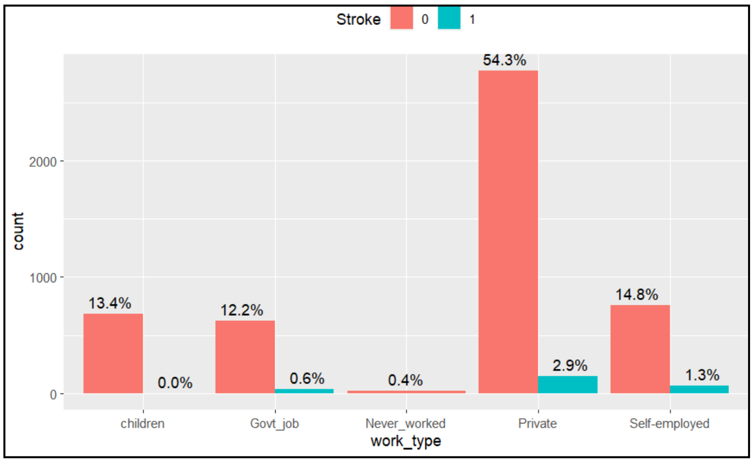


Fig. 6: Stoke occurrence in work types

The individuals working in private sector were 57.4% (29244), and the stroke occurrence was higher in this group.

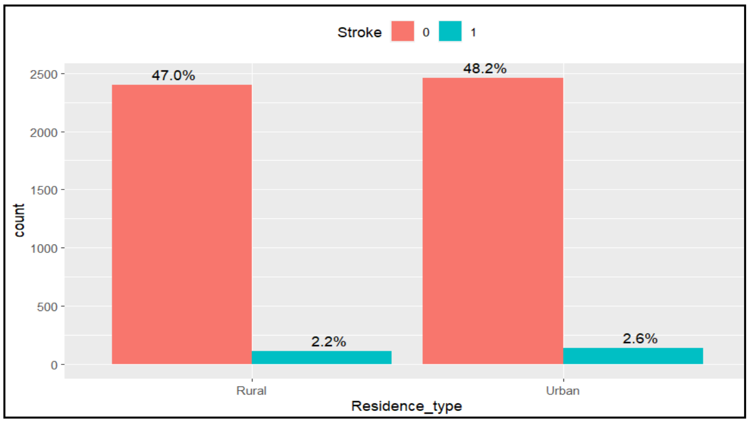


Fig. 7: Stroke occurrence in residence types

The number of individuals staying in rural and urban areas were similar. The occurrence of stroke was slightly higher in urban group.

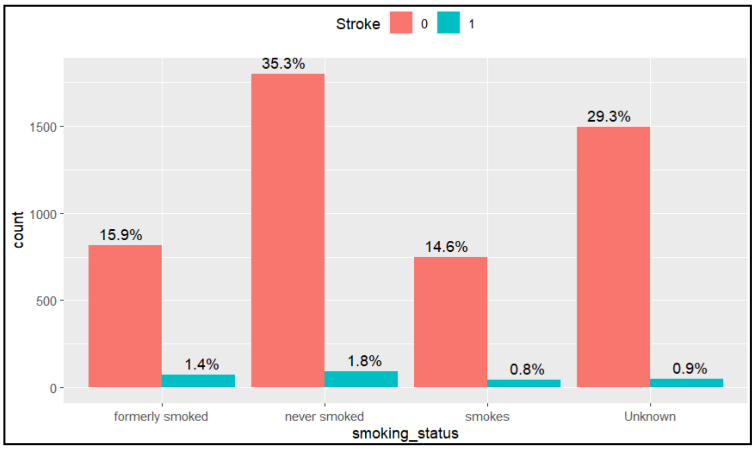


**Fig. 8: Stroke occurrence as per smoking status**

The individuals who never smoked were 37.1% (1892). The occurrence of stroke was higher in this group followed by individuals who formerly smoked.

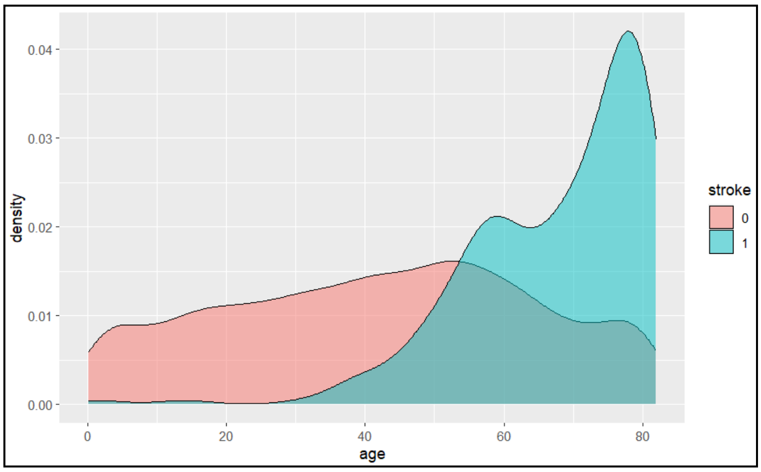


**Fig. 9: stroke occurrence as per age**

The figure above shows that stroke occurrence is higher in individuals whose age is between 45 to 60 years and highest in age between 70 to 90 years.

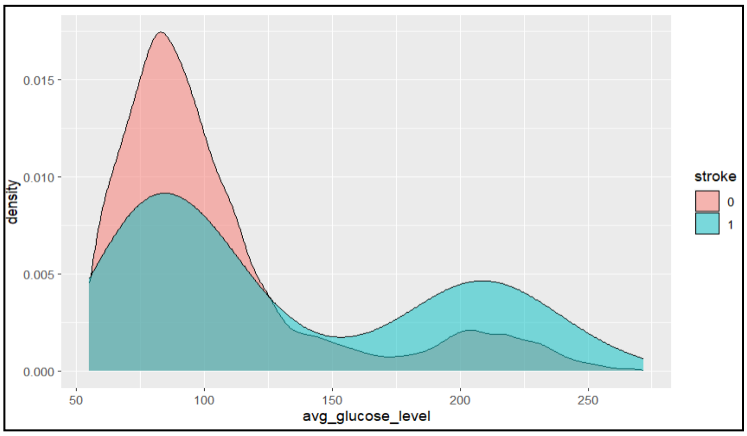


**Fig. 10: Stroke occurrence as per average blood glucose**

The figure above shows stroke occurrence for individuals having average blood glucose between 50 to 110 and 180 to 230. The stroke occurrence is lesser comparatively in individuals having average blood glucose between 50 to 120.

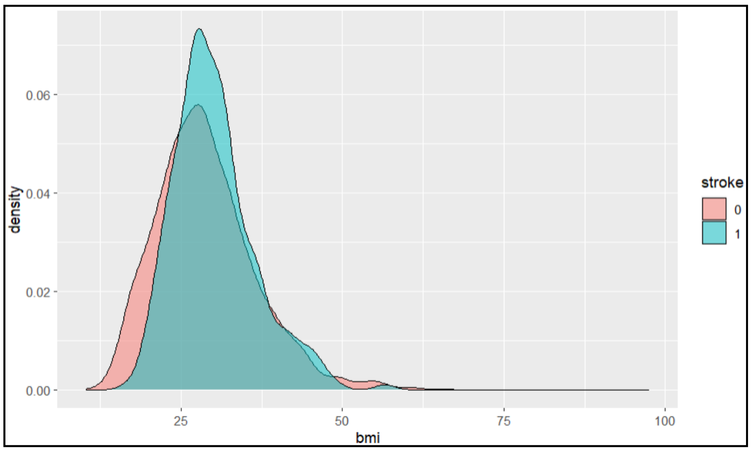


**Fig. 11: Stroke occurrence as per BMI**

The figure above shows that stroke occurrence is higher in the individuals having higher BMI.

The bivariate analysis revealed that individuals aged over 45 have a higher likelihood of experiencing a stroke. Additionally, those with a BMI between 25 and 50, average glucose levels ranging from 50 to 110 and 180 to 230, non-smokers, employed in the private sector, without heart disease and hypertension, and of female gender, are associated with an elevated risk of stroke.

The valuable insights into individual associations provided by the outcomes of the bivariate analysis are not only utilized but also contribute to the groundwork for the construction of robust machine learning models in subsequent multivariate analyses. Identifying variables such as age, BMI, average glucose levels, smoking history, occupational sector, and specific health conditions that show significant relationships with the likelihood of stroke is essential for creating effective predictive models.

### Data Preparing and Model Building

The dataset exhibited an imbalance in class distribution of response variable stroke. 95.12 % was no stroke class and 4.87% was stroke class. To rectify this, oversampling techniques were applied using the ROSE package to achieve a balanced distribution of classes. There were equal cases (3402) of stroke and no stroke classes. The data was then split into training and testing sets (70:30). The training set was used to train the model, while the testing set was used to evaluate its performance. This study employs a diverse set of classifying machine learning algorithms to develop a robust stroke prediction model. The models considered include Multiple Logistic Regression, Random Forest, Naïve Bayes, and Support Vector Machine (SVM).

## RESULTS

During exploratory data analysis (EDA), preliminary insights into stroke prediction for each column were unveiled, encompassing the investigation of relationships between variables and understanding the characteristics of the dataset. Notable associations with the likelihood of stroke emerged, highlighting variables such as age, BMI, average glucose levels, smoking history, occupation, and specific health conditions. The data's intricacies were explored, providing a foundational understanding for subsequent analyses and the construction of predictive models. Model evaluation was performed by assessing the performance of the model on the testing data. Evaluation metrics appropriate for the problem type were selected. For classification, metrics such as accuracy, area under the curve (AUC), precision, recall, and F1-score were utilized. Insightful findings regarding the performance and suitability of the given dataset were revealed through the evaluation of the stroke prediction models. Logistic Regression, Random Forest, Naïve Bayes, and Support Vector Machine (SVM) were scrutinized based on identified metrics for classification models.

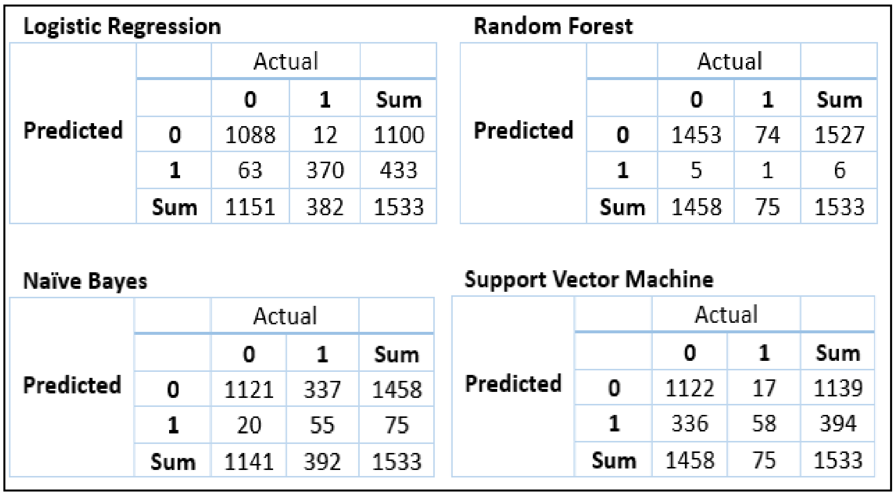
An accuracy of 75.08%, a specificity of 84%, and an AUC of 86.59% were demonstrated by the logistic regression model. Despite its comparatively lower accuracy, a balanced approach with high specificity is achieved by logistic regression, rendering it a compelling choice when the minimization of false positives is imperative. This is particularly crucial in healthcare applications where the prediction of the absence of a stroke for a patient holds equal importance.

**Table I: Model Metrics**

| Model | Accuracy | Specificity | AUC |
|---|---|---|---|
| **Logistic Regression** | 75.08 | 84 | 86.59 |
| **Random Forests** | 94.85 | 1.33 | 81.14 |
| **Naïve Bayes** | 76.71 | 73.33 | 75.11 |
| **Support Vector Machine** | 76.97 | 77.33 | 77.14 |

**Table II: Model Metrics**

| Model | Precision | Recall | F1 |
|---|---|---|---|
| **Logistic Regression** | 98.91 | 84 | 90.84 |
| **Random Forests** | 95.15 | 1.33 | 2.62 |
| **Naïve Bayes** | 98.25 | 73.33 | 83.98 |
| **Support Vector Machine** | 98.51 | 77.33 | 86.64 |



**Logistic Regression**

| | | Actual | | |
|---|---|---|---|---|
| | | **0** | **1** | **Sum** |
| **Predicted** | **0** | 1088 | 12 | 1100 |
| | **1** | 63 | 370 | 433 |
| | **Sum** | 1151 | 382 | 1533 |

**Random Forest**

| | | Actual | | |
|---|---|---|---|---|
| | | **0** | **1** | **Sum** |
| **Predicted** | **0** | 1453 | 74 | 1527 |
| | **1** | 5 | 1 | 6 |
| | **Sum** | 1458 | 75 | 1533 |

**Naïve Bayes**

| | | Actual | | |
|---|---|---|---|---|
| | | **0** | **1** | **Sum** |
| **Predicted** | **0** | 1121 | 337 | 1458 |
| | **1** | 20 | 55 | 75 |
| | **Sum** | 1141 | 392 | 1533 |

**Support Vector Machine**

| | | Actual | | |
|---|---|---|---|---|
| | | **0** | **1** | **Sum** |
| **Predicted** | **0** | 1122 | 17 | 1139 |
| | **1** | 336 | 58 | 394 |
| | **Sum** | 1458 | 75 | 1533 |

**Figure 12: Confusion Matrix**

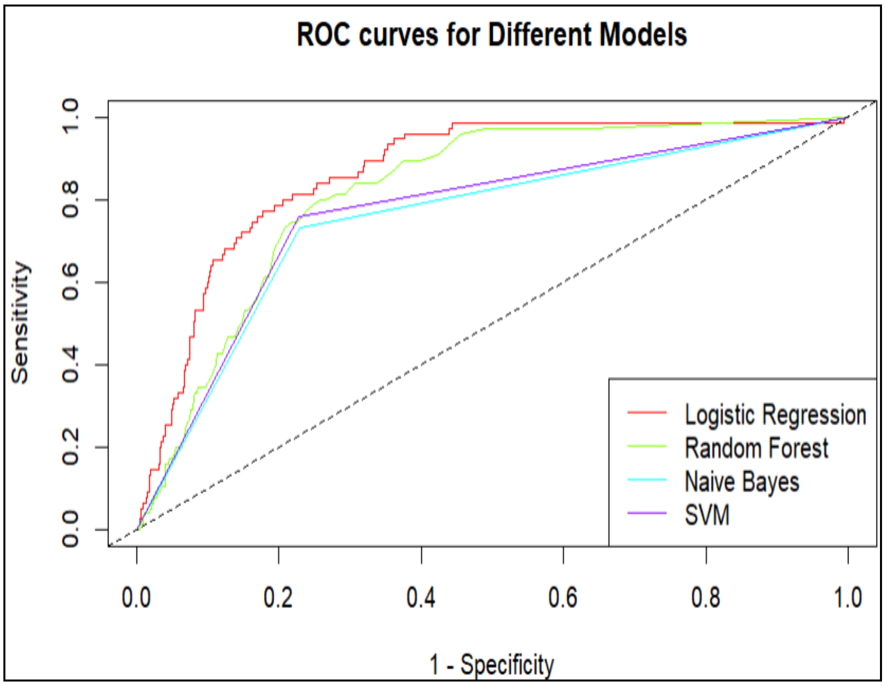


**Figure 13: ROC curves**

An accuracy of 76.71%, a specificity of 73.33%, and an AUC of 75.11% were exhibited by Naïve Bayes. An accuracy of 76.97%, a specificity of 77.33%, and an AUC of 77.14% were achieved by SVM. The competitive accuracy and balanced performance of SVM position it as a strong candidate for healthcare applications. The algorithm's ability to handle non-linear relationships and complex decision boundaries aligns well with the intricate nature of stroke prediction.

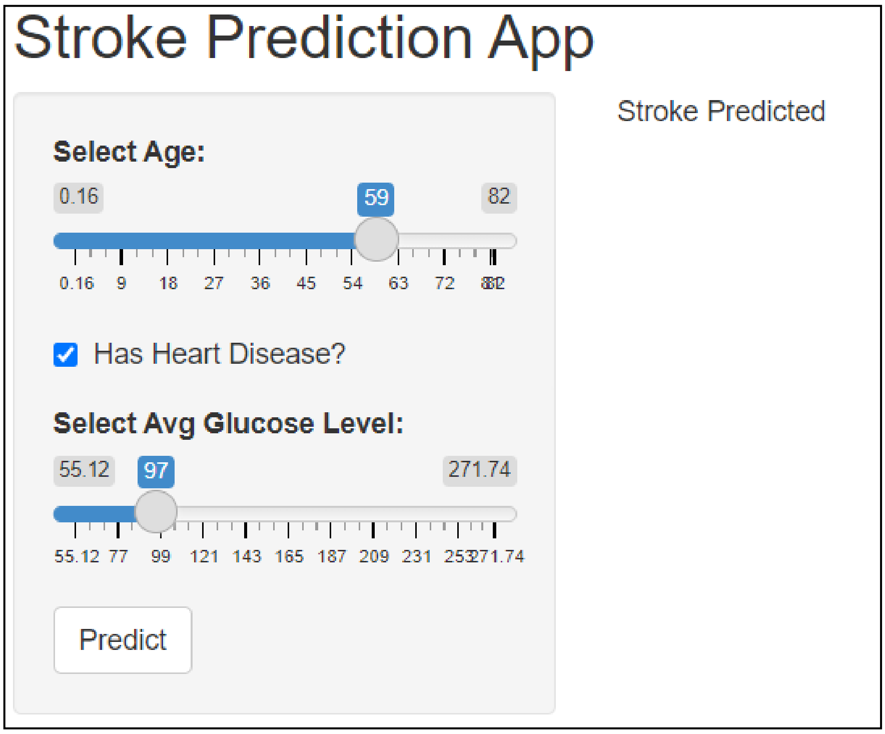
The higher accuracy achieved by some algorithms for stroke prediction can be attributed to several factors, such as their ability to handle complex data relationships, utilize ensemble methods, incorporate randomness, and perform effective feature selection. In this study, Random Forest's superior accuracy is likely due to the combined strengths of these factors, while Logistic Regression's high specificity makes it valuable for scenarios where minimizing false positives is crucial, even when accuracy alone is not sufficient. The understanding of these factors aids in the selection of the most suitable algorithm for specific needs and the effective interpretation of model performance.

These results underscore the importance of considering multiple metrics in model evaluation. While Random Forest demonstrated superior accuracy, the balance between sensitivity and specificity i.e., F1 metric provided by Logistic Regression makes it a compelling choice.

## DEPLOYMENT

The stroke prediction model was deployed on a Shiny app to enhance accessibility and usability. Utilizing Shiny, a R package, an interactive web application was created directly from R scripts. Through this platform, demographic and health-related features for a given patient can be seamlessly input, and real-time predictions regarding stroke risk can be obtained. The user-friendly interface enables the application of predictive models in real-world healthcare scenarios, promoting wider accessibility and fostering the integration of data-driven decision-making in clinical practices.

The Shiny app deployment aligns with the broader objective of translating machine learning models into practical tools for healthcare professionals. This interactive platform is a tangible outcome of the research, bringing the benefits of predictive analytics directly to end-users and reinforcing the potential impact of data science in the healthcare domain.



Stroke Prediction App

Stroke Predicted

Select Age: 59

☑ Has Heart Disease?

Select Avg Glucose Level: 97

**Figure 14: Web Application of the deployed model**

## CONCLUSION

This study investigated the application of machine learning models for stroke prediction. Four different models, namely Logistic Regression, Random Forest, Naive Bayes, and Support Vector Machine, were employed and evaluated based on various metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve. Suitable for analysis and prediction, the logistic regression model was chosen to deploy for its ability to estimate the probability of event occurrences, such as stroke likelihood, based on a set of independent variables. In handling binary classification tasks, providing probabilities bounded between 0 and 1, and offering a balanced approach with high specificity, its effectiveness is valuable for scenarios were minimizing false positives, such as predicting the absence of a stroke, is crucial. The following key takeaways have been observed:

### Metrics Matter

While Random Forest achieved the highest accuracy, Logistic Regression offered a better balance between sensitivity and specificity (F1-score). This highlights the importance of considering various metrics beyond just accuracy when choosing the best model for stroke prediction.

### Balancing Act

Each algorithm exhibited its strengths and trade-offs. The choice of the most suitable model depends on the specific needs of the application. For example, if minimizing false positives is crucial, Logistic Regression might be preferred despite its slightly lower accuracy.

## Problem Relevance

The choice of the most suitable model in this paper depends on clinical requirements, emphasizing the importance of understanding the trade-off between minimizing false positives and false negatives in healthcare applications.

Also, the foundation for several avenues of future research and development has been established with the successful deployment of the stroke prediction model on the Shiny app. Firstly, the predictive performance of the models can be refined by incorporating additional relevant features or exploring deep learning algorithms. Continuous data collection and integration of new patient records can contribute to the model's adaptability and improved accuracy over time.

Furthermore, additional functionalities can be incorporated into the Shiny app, such as personalized risk factor analysis, explanatory model outputs, and integration with electronic health records. These advancements can facilitate a more comprehensive and user-centric approach, enabling informed decisions to be made by healthcare professionals based on a deeper understanding of the underlying model predictions.

Moreover, the integration of real-time data feeds and continuous model retraining can ensure that the predictive capabilities of the deployed system remain robust and up-to-date. In summary, the future scope encompasses the continuous refinement of predictive models, the expansion of the Shiny app's capabilities, and collaboration with the healthcare community to foster a more seamless and impactful integration of predictive analytics into routine clinical practices.

## REFERENCES

[1] Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthcare Analytics, 2, 100032. https://doi.org/10.1016/j.health.2022.100032

[2] Katan, M., & Luft, A. (2018). Global Burden of Stroke. Seminars in Neurology, 38(02), 208–211. https://doi.org/10.1055/s-0038-1649503

[3] National Heart, Lung and Blood Institute. (2022, March 24). Stroke - What Is a Stroke? Www.nhlbi.nih.gov. https://www.nhlbi.nih.gov/health/stroke

[4] Mainali, S., Darsie, M. E., & Smetana, K. S. (2021). Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. Frontiers in Neurology, 12. https://doi.org/10.3389/fneur.2021.734345

[5] IBM. (2022). What Is Logistic Regression? IBM. https://www.ibm.com/topics/logistic-regression

[6] IBM. (2023). What is Random Forest? | IBM. Www.ibm.com. https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly

[7] What are Naive Bayes classifiers? | IBM. (n.d.). Www.ibm.com. Retrieved November 29, 2023, from https://www.ibm.com/topics/naive-bayes#:~:text=the%20next%20step-

[8] Kanade, V. (2022, September 29). What Is a Support Vector Machine? Working, Types, and Examples. Spiceworks. https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/I. Chourib, G. Guillard, I. R. Farah, and B. Solaiman, "Stroke Treatment Prediction Using Features Selection Methods and Machine Learning Classifiers," IRBM, Mar. 2022, doi: https://doi.org/10.1016/j.irbm.2022.02.002

[9] Rahman, M., Islam, S., Sadia Binta Sarowar, & Meem Tasfia Zaman. (2023). Multiple Disease Prediction using Machine Learning and Deep Learning with the Implementation of Web Technology. https://doi.org/10.1109/aibthings58340.2023.10292488.

[10] Sharma, S. (2023). Stroke Prediction Using XGB Classifier, Logistic Regression, GaussianNB and BernaulliNB Classifier. https://doi.org/10.1109/iccpct58313.2023.10245269.

[11] Al-Zubaidi, H., Dweik, M., & Al-Mousa, A. (2022, November 1). Stroke Prediction Using Machine Learning Classification Methods. IEEE Xplore. https://doi.org/10.1109/ACIT57182.2022.10022050

[12] Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. International Journal of Advanced Computer Science and Applications, 12(6). https://doi.org/10.14569/ijacsa.2021.0120662

[13] Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019). Predicting Stroke from Electronic Health Records. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2019, 5704–5707. https://doi.org/10.1109/EMBC.2019.8857234

[14] FEDESORIANO. (2021). Stroke Prediction Dataset. Www.kaggle.com. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[15] Alanazi, E., Abdou, A., & Luo, J. (2020). Predicting risk of stroke from lab tests using machine learning algorithms (Preprint). JMIR Formative Research. https://doi.org/10.2196/23440

[16] Mridha, K., Ghimire, S., Shin, J., Aran, A., Uddin, Md. M., & Mridha, M. F. (2023). Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study with a Web Application for Early Intervention. IEEE Access, 11, 52288–52308. https://doi.org/10.1109/ACCESS.2023.3278273

[17] Amann, J. (2021). Machine Learning in Stroke Medicine: Opportunities and Challenges for Risk Prediction and Prevention. Advances in Neuroethics, 57–71. https://doi.org/10.1007/978-3-030-74188-4_5

[18] Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. International Journal of Advanced Computer Science and Applications, 12(6). https://doi.org/10.14569/ijacsa.2021.0120662

[19] Dritsas, E., & Trigka, M. (2022). Stroke Risk Prediction with Machine Learning Techniques. Sensors, 22(13), 4670. https://doi.org/10.3390/s22134670