Stroke risk prediction

Mihail Gonta [1]

[1] Technical University of Moldova, Faculty of Computing, Informatics and Microelectronics, Department of Informatics and Systems Engineering

ABSTRACT

This research investigates the escalating incidence of strokes, the second leading global cause of death, focusing on the surge in young adults. Leveraging a Kaggle-sourced dataset with 11 clinical features, the study employs comprehensive data preprocessing, exploratory data analysis, and logistic regression models to identify key factors influencing stroke occurrence and develop a predictive model. Findings reveal unexpected insights into glucose levels and BMI, emphasizing the need for nuanced risk assessments. Variable Importance in Projection analysis underscores the significance of age, average glucose level, smoking status, heart disease, and hypertension. Comparative analyses affirm the logistic regression model's balanced accuracy of 81.13%, while an XGBoost model outperforms with an accuracy of 96.15%. This research not only enhances our understanding of stroke risk factors but also provides a foundation for future predictive modeling efforts to guide early intervention and prevention strategies.

INTRODUCTION

As of the current date, a lot of medical research is done using the data analysis techniques. According to World Health Organization (WHO) stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths. Annually, 15 million people worldwide suffer a stroke. Of these, 5 million die and another 5 million are left permanently disabled, placing a burden on family and community. Stroke is uncommon in people under 40 years; when it does occur, the main cause is high blood pressure[1].

Stroke poses a significant health concern with severe repercussions, occurring when the blood flow to the brain is disrupted, leading to the demise of brain cells and dysfunction in specific brain regions. This interruption in blood supply can stem from an obstructed artery, a ruptured blood vessel causing intracranial bleeding, or a temporary reduction in blood flow to the brain.

The Medical Research Foundation found a new research that has shown a sharp increase in the incidence of stroke in young adults, in a study of more than 94,000 people of Oxfordshire. They found that between 2002-2010 and 2010-2018, there was a 67% increase in stroke incidents among the young adults (under 55 years), and 15% decrease among older adults (55 years or older)[2].

Our research focuses on conducting an educational study with the objective of analyzing a dataset comprising 11 medical features. The primary goal is to identify and elucidate the key factors that contribute to the description of the stroke variable within the dataset. Subsequently, our aim is to develop and train a novel predictive model that can effectively assess the risk of stroke. This study not only involves a comprehensive examination of the collected data but also encompasses the creation and optimization of a predictive model to enhance stroke risk prediction.

Initial assumptions

- Is there a correlation between age and strokes, and if so, how is this correlation distributed?
- Can body mass index and glucose levels contribute to the likelihood of experiencing a stroke?
- Is the assumption valid that smoking can trigger strokes?
- Is there a link between heart disease and an increased susceptibility to strokes?
- Does high blood pressure resulting from workload contribute to the risk of strokes?
- Is it accurate to claim that males, particularly those facing high work-related stress, are more prone to strokes?
- How are continuous and categorical data related, and what is the significance of this relationship?
- What is the importance of features and how can they be selected to enhance the accuracy of stroke predictions?

Materials & Methods

For this analysis, an open-source dataset was obtained from Kaggle. It is important to highlight that the source of the dataset is confidential, and its usage is restricted solely to educational purposes. Source: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/[3].

The initial dataset comprises 5110 observations and 12 variables. However, focusing solely on 11 clinical features: gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, smoking status, and stroke.

Variable	**Data Type**	**Description**
gender	Factor	Categorical variable indicating the gender (Male/Female)
age	Numeric	Age of the individuals in the dataset

hypertension	Factor	Categorical variable indicating the presence of hypertension (0: No, 1: Yes)
heart_disease	Factor	Categorical variable indicating the presence of heart disease (0: No, 1: Yes)
ever_married	Factor	Categorical variable indicating marital status (Yes/No)
work_type	Factor	Categorical variable indicating the type of work (children, govy job, never worked, private, self-employed)
Residence_type	Factor	Categorical variable indicating the residence type (Urban/Rural)
avg_glucose_level	Numeric	Average glucose level in the blood
bmi	Numeric	Body Mass Index (BMI)
smoking_status	Factor	Categorical variable indicating smoking status (smokes, formerly smoked, never smoked)
stroke	Factor	Categorical variable indicating stroke occurrence (0: No, 1: Yes)

Data preprocessing

At the very beginning were observed a significant number of entries with unknown or missing values that needed clarification or removal. Were encountered instances where certain observations exhibited missing data, such as "N/A" for BMI, "Unknown" for smoking status, and an ambiguous gender denoted as "Other," so those were removed from the dataset, including the unnecessary level of gender variable. Additionally, it was noted that the 'id' variable provided limited assistance and did not contribute significantly to the research objectives; hence, it was dropped from the dataset. This way our 5110 observations reduced to 3425.

Another step was converting certain columns to factors (categorical variables) and turning the "bmi" column into a double (numeric variable). Variables that were transformed as factors include "gender", "hypertension", "heart_disease", "ever_married", "work_type", "residence_type", "smoking_status" and "stroke".

Exploratory data analysis (EDA)

The EDA process involved visualization tools such as R and ggplot2 that were used to analyze data sets related to stroke events. The analysis included generating bar charts and density plots to examine the distribution of stroke across demographic and lifestyle factors in the images as well as blood pressure, cardiovascular disease, gender , type of occupation, marital status, smoking status, residential area, age, glucose level, and BMI. Bar charts based on percentages revealed the effect of these factors on tumor generation. Boxplots were used to analyze the distribution of statistical parameters such as age, mean glucose levels, BMI, etc., and revealed possible associations with stroke.

VIP (Variable Importance in Projection)

In this analysis, VIP (Variable Importance in Projection) was employed to assess the importance of features in predicting stroke using logistic regression models. The script utilized the "vip" package to generate variable importance plots for two logistic regression models. VIP provides insights into the contribution of each feature to the model's predictive performance. In the context of stroke prediction, VIP helped identify which variables, such as age, hypertension, heart disease, residence type, average glucose level, and smoking status, played crucial roles in the logistic regression models. The use of VIP assists in understanding the relative importance of different features and enhances interpretability, aiding researchers and practitioners in identifying key factors associated with stroke occurrences.

Logistic Regression

A logistic regression model was developed to predict strokes using the train function in R. The model considered various predictor variables, such as gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status. The training utilized the glm method with a binomial family, given the binary nature of the outcome variable (stroke or no stroke). The process involved 100-fold cross-validation (trainControl), ensuring robust evaluation by iteratively training on 99 subsets and validating on one. The resulting model is poised to offer insights into the relationships between the specified predictors and the likelihood of stroke occurrence.

Performance assessment of models

Various techniques were used to assess the performance of logistic regression models in predicting strokes. Key aspects of performance evaluation include confusion matrices, area under the receiver operating characteristic curve (AUC-ROC), and variable importance plots (VIP). The confusion matrices were used to analyze the model's accuracy, precision, recall, and F1 score on the training set. Additionally, the "pROC" package was used to create ROC curves for each model, visually depicting their discriminatory power. The AUC-ROC values were then computed to quantify the models' overall performance. Comparing multiple models, such as cv_stroke_model1 and cv_stroke_model2, facilitated the selection of the model with superior predictive capabilities.

EDA results

The obtained dataset is imbalanced

As we can observe from Figure 1 down bellow, the distribution of data in our sample shows that 5 out of 100 people suffer a stroke. Notably, these data exhibit significant bias with a zero accuracy score of 95%, suggesting that a naive model that randomly predicts strokes could achieve 95% accuracy.

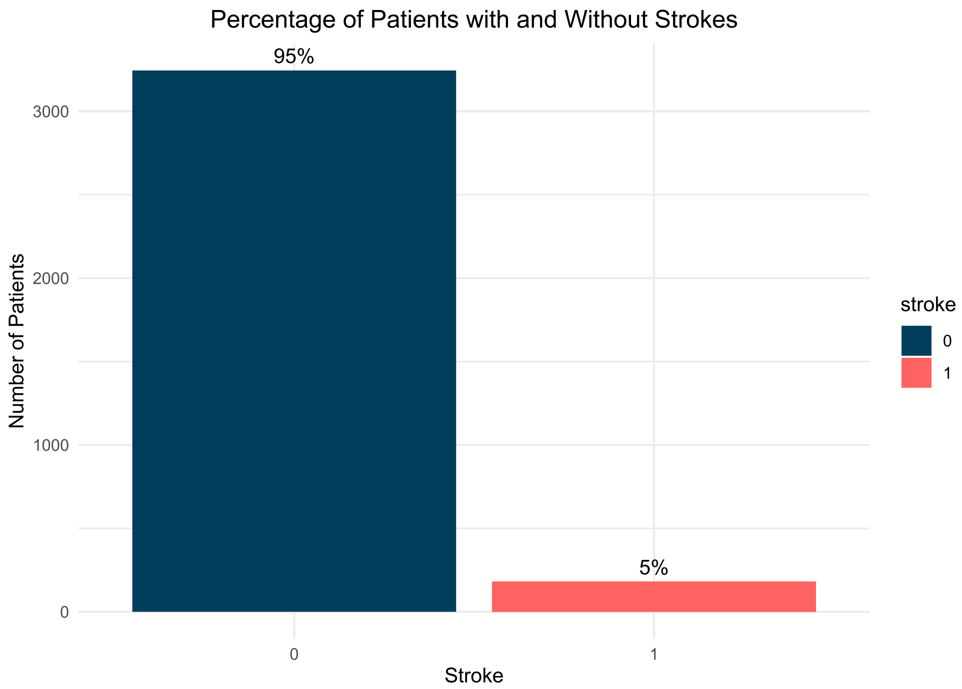

Figure 1. Distribution of Stroke in Patients

Therefore, as our data is imbalanced, during the modeling and training process, it is mandatory to address this bias by either oversampling or undersampling to achieve optimal results.

Older individuals exhibit a higher prevalence of strokes compared to their younger counterparts

The significance of age as a feature in predicting strokes is underscored by the non-normal distribution of age feature values, necessitating transformation for later analysis. Categorical features reveal a notable trend: older individuals exhibit a higher prevalence of strokes compared to their younger counterparts.

This aligns with our intuitive understanding that advancing age correlates with an increased stroke risk, this can be observed in Figure 2.

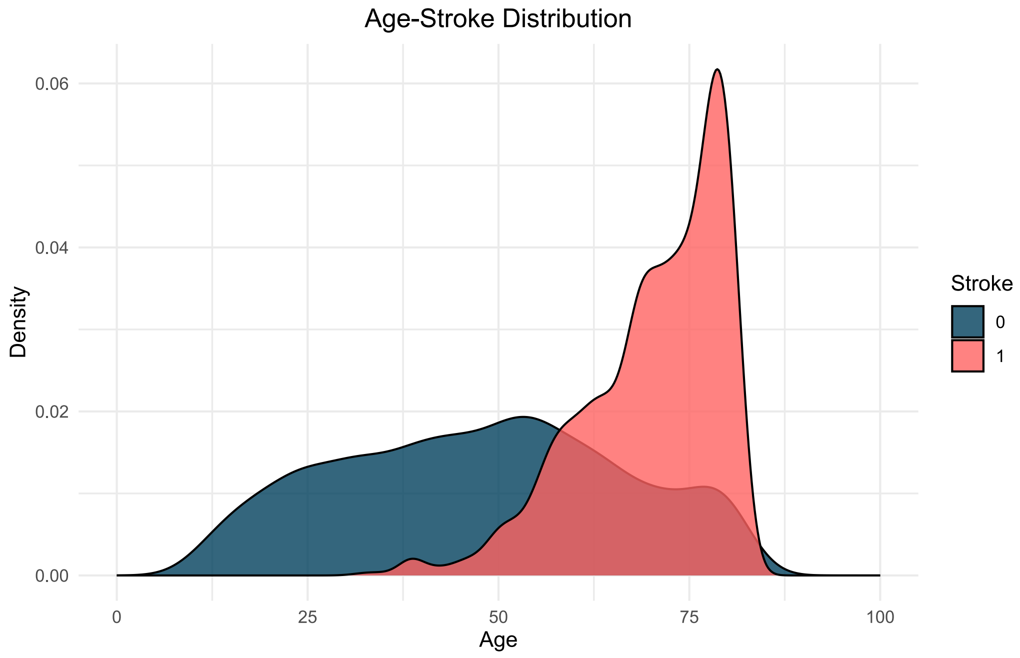

Figure 2. Distribution of Stroke by Age

Unexpectedly, strokes are more common in individuals with normal glucose levels

High blood glucose levels are often referred to as one of the causes of stroke risk[4]. But as we can see in Figure 3, the distribution of glucose levels exhibits a notable leftward skew, suggesting a predominant prevalence of lower glucose concentrations within the studied population.

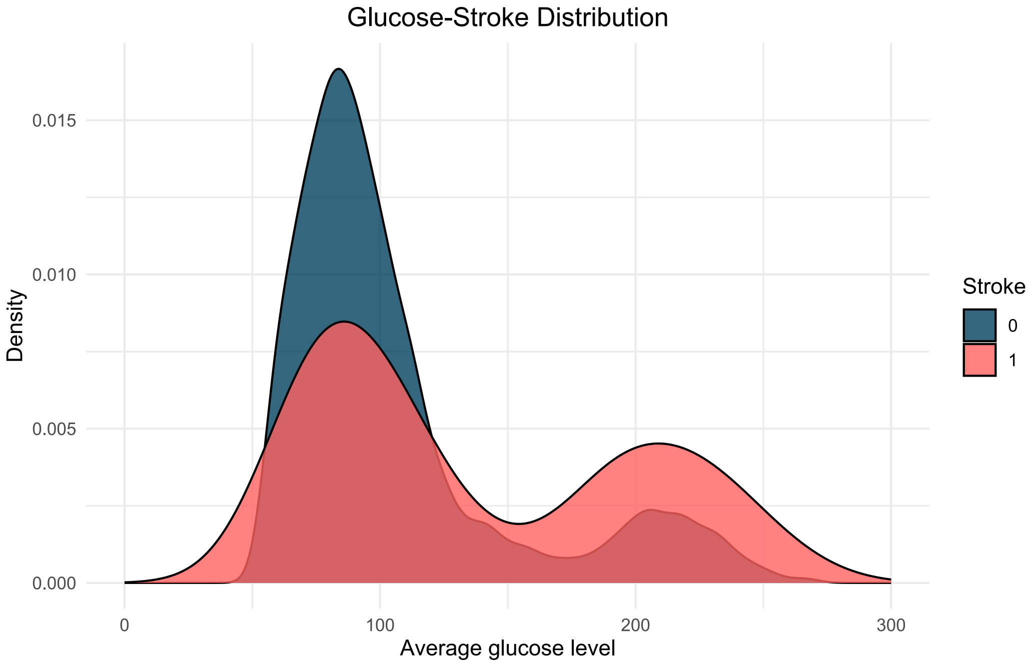

Figure 3. Distribution of Stroke by average glucose levels

The distribution of hypertension, illustrated in Figure 4, within populations underscores a notable health concern, as individuals afflicted by this condition confront a substantially heightened risk of experiencing a stroke. The data elucidates a significant discrepancy, revealing that those with hypertension face an approximate **10%** higher incidence of strokes compared to their counterparts without hypertension.

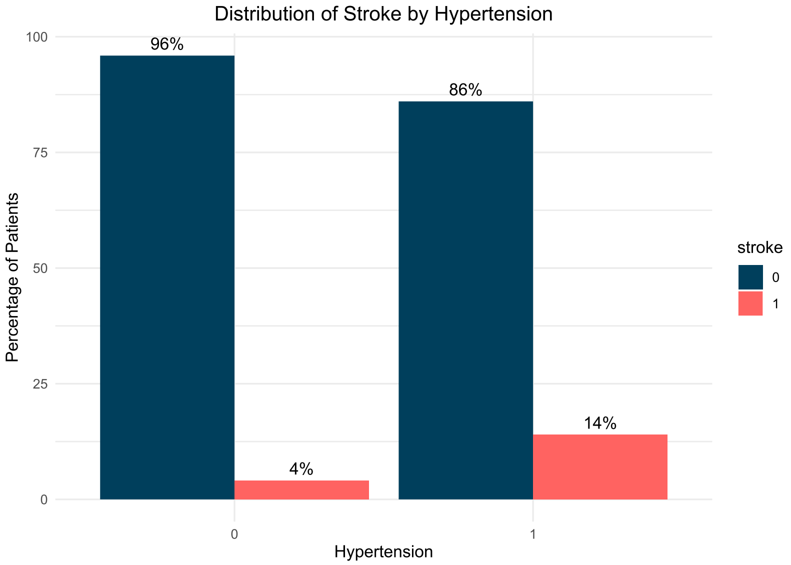

Figure 4. Distribution of Stroke by Hypertension

The distribution of heart disease, see Figure 5, within populations serves as a critical indicator of the potential risks individuals face, particularly concerning stroke incidence. Within this context, it becomes evident that individuals with pre-existing heart conditions encounter a significantly elevated risk, with almost **13%** experiencing strokes.

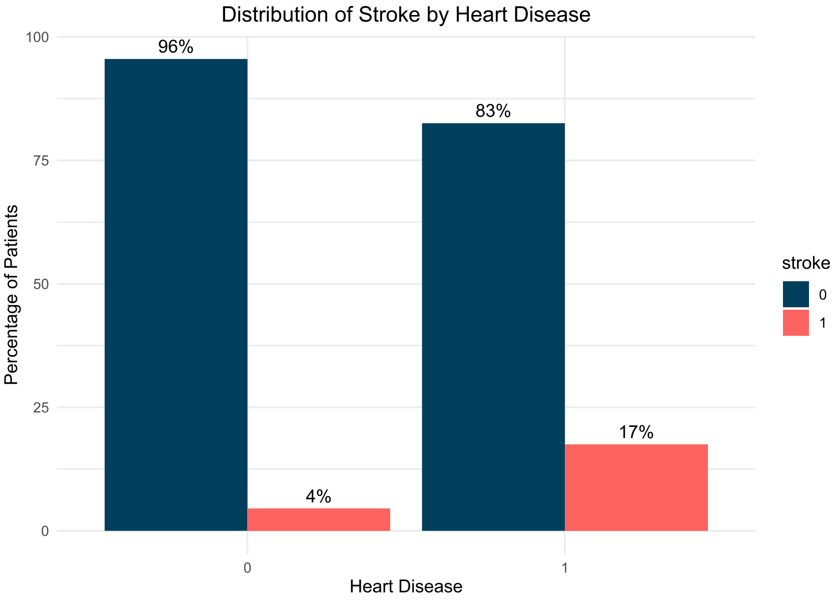

Figure 5. Distribution of Stroke by Heart Disease

Correlation between various features

The correlation coefficients, illustrated in Figure 6, reveal a moderate positive relationship between age and the likelihood of a stroke (r = 0.24). Additionally, hypertension, heart disease, and average glucose level demonstrate weaker positive correlations with stroke, with coefficients of 0.14 for each variable.

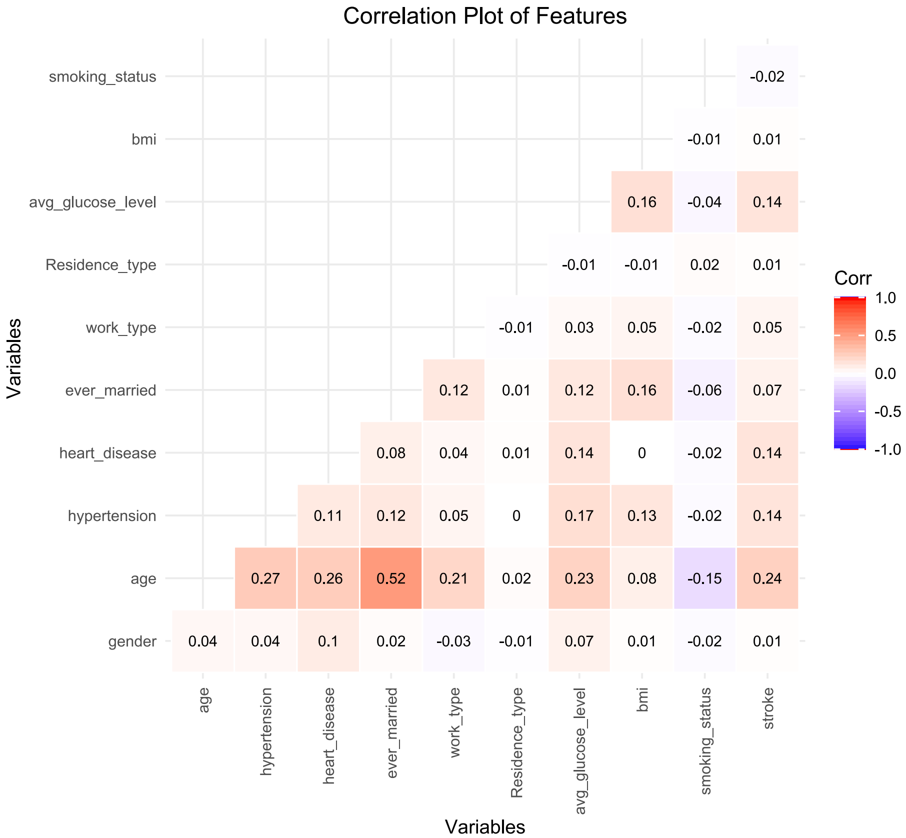

Figure 6. Correleation plot of dataset features

These findings suggest that as age increases, the probability of experiencing a stroke tends to rise, while the influence of hypertension, heart disease, and average glucose level is less pronounced but still present.

The exploratory data analysis (EDA) conducted on the dataset focusing on patients with and without strokes revealed several noteworthy findings. The data displayed a substantial imbalance, with only 5% of individuals experiencing strokes. Univariate analysis of continuous variables highlighted the significance of age as a predictor, revealing a non-normal distribution where older individuals exhibited a higher prevalence of strokes. Surprisingly, the distribution of glucose levels showed a leftward skew, indicating lower concentrations, yet strokes were observed across normative glucose levels. BMI distribution, skewed to the right, suggested a concentration within the healthy range, implying limited contribution to stroke likelihood. Categorical variables demonstrated varying impacts, with hypertension and pre-existing heart conditions correlating with a higher stroke risk, while gender, marital status, and residence type showed modest associations. Smoking and work type appeared to influence stroke likelihood, with former smokers and self-employed individuals exhibiting higher probabilities. Correlation analysis highlighted a moderate positive relationship between age and stroke, with weaker correlations for hypertension, heart disease, and average glucose level.

Imbalanced data was solved by oversampling

The initial imbalance ratio of the dataset, as calculated by imbalanceRatio, was 0.0555, indicating a substantial class imbalance with the minority class (stroke = 1) being underrepresented. To solve this problem was used the MWMOTE, or Majority Weighted Minority Oversampling Technique, that addresses class imbalance by selectively oversampling the minority class in a dataset. It identifies minority instances with fewer neighboring examples, emphasizing those that are potentially more challenging for the model to learn. By assigning weights to majority class instances based on their proximity to the minority class, MWMOTE guides the generation of synthetic samples that balance the class distribution.

The oversampled dataset contains a total of 6490 observations, with an equal distribution of 3245 instances for each class (stroke = 0 and stroke = 1), resulting in a balanced dataset with a class proportion of 0.5 for each class.

VIP findings

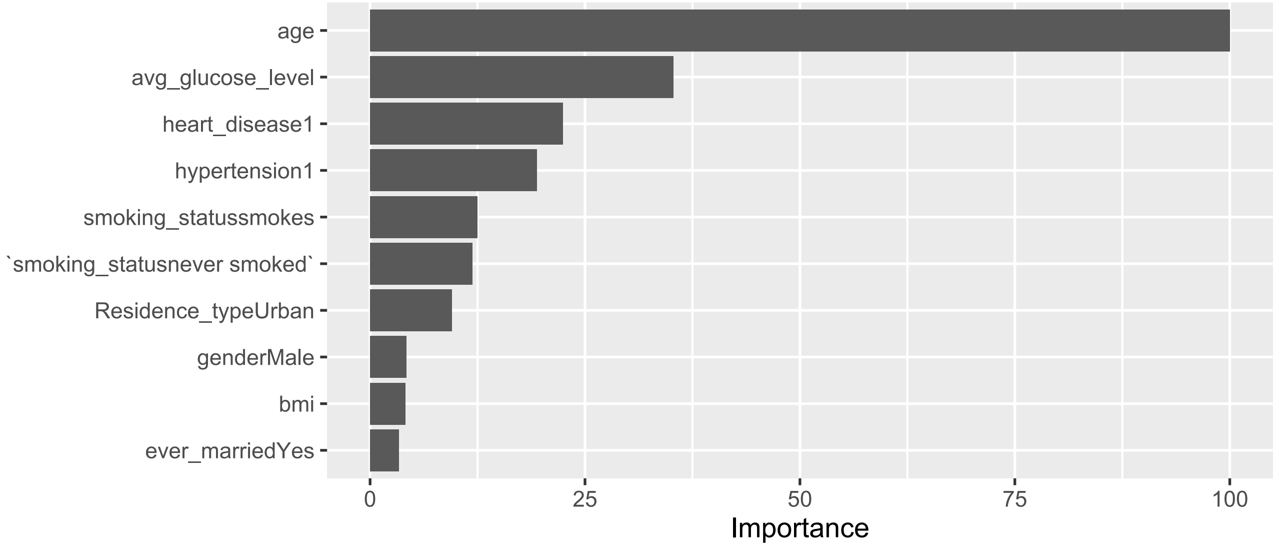

Figure 7. VIP of glm_model_1

Following the development of our initial logistic regression model, encompassing all features within our dataset, a crucial revelation emerged through the identification of Variable Importance in Prediction (VIP). This analysis brought to light the heightened significance of certain variables, notably age, average glucose level, smoking status and the presence of heart disease and hypertension. These findings underscore the pivotal role played by these specific factors in influencing and shaping the predictive power of our logistic regression model. As we delve deeper into the intricacies of VIP, a more nuanced understanding of feature importance is anticipated, paving the way for refined model enhancements and insightful interpretations.

All features can be preserved for the final model

Having identified features with pronounced impact on the model through our VIP findings, we proceeded to elevate our model refinement strategy. Subsequently, a second model was meticulously crafted, incorporating exclusively the most influential variables distilled from the preceding VIP analysis. To assess the efficacy of these adjustments, a rigorous comparative analysis ensued, employing the Akaike Information Criterion (AIC) method. This methodical evaluation not only gauges the performance disparities between the two models but also serves as a robust benchmarking tool, aiding in the discernment of the model iteration that strikes the optimal balance between simplicity and explanatory power. Here are the results:

Model	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
glm_model_2	6	3913.31	0.00	0.7	0.7	-1950.64
glm_model_1	11	3915.00	1.7	0.3	1.0	-1946.47

The AICc analysis reveals that both models, model1 and model2, exhibit comparable performance, as indicated by their close AICc values. The difference in AICc, measured by Delta_AICc, is relatively small (1.7), suggesting no substantial superiority of one model over the other. The AICc weights (AICcWt) further underscore the balanced consideration for both models, with a dominance of 0.7 for model2 and 0.3 for model1.

In light of these results, it is reasonable to conclude that the inclusion of all features in both models is justifiable, as the evidence provided by the AICc analysis does not strongly favor one model over the other. The cumulative weight (Cum.Wt) of 1.0 affirms the adequacy of retaining the entire feature set encompassed by both models.

Ultimately, this AICc analysis suggests that the models are equally plausible, and the inclusion of all features is warranted for a comprehensive understanding of the underlying data.

The ROC plot, see Figure 8, unequivocally illustrates comparable performance between the two models, implying that their discriminative abilities are nearly indistinguishable. Consequently, a fair conclusion can be drawn, suggesting that retaining all features is justifiable given the lack of substantial differentiation in performance between the models.

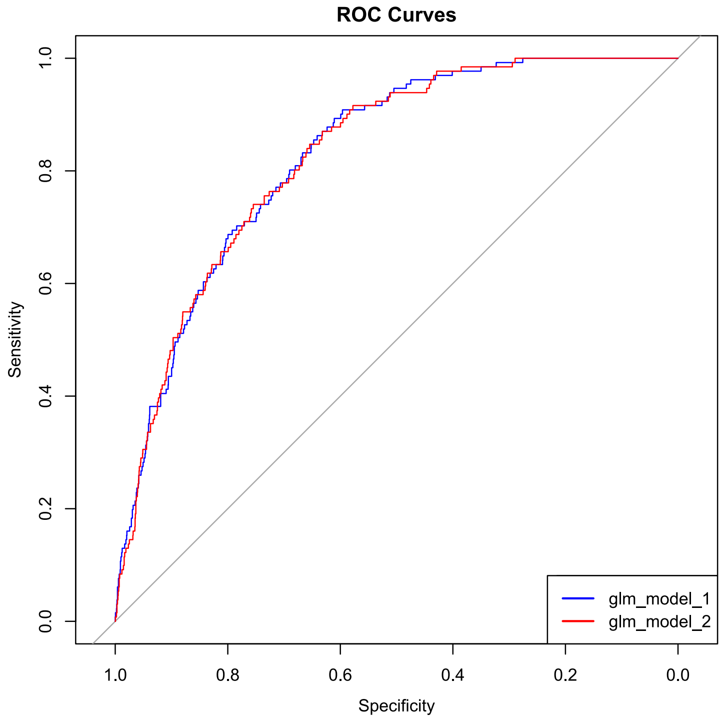

Figure 8. ROC of glm method models

Confusion matrix results

The logistic regression model (glm_model1) trained on the stroke dataset achieved an accuracy of 81.13%, indicating its ability to correctly classify instances. The model demonstrated good sensitivity (true positive rate) of 85.12%, suggesting its effectiveness in identifying positive cases of stroke. Additionally, the specificity (true negative rate) was 77.15%, indicating a reasonable ability to correctly identify non-stroke instances. The Kappa statistic, measuring the agreement beyond chance, was 0.6226, further supporting the model's overall performance. The positive predictive value (PPV) was 78.83%, signifying the proportion of predicted positive cases that were correctly classified. The negative predictive value (NPV) was 83.83%, indicating the proportion of predicted negative cases that were correctly identified. Overall, the model displayed a balanced accuracy of 81.13%, suggesting a fair equilibrium between sensitivity and specificity in stroke prediction. The results suggest that the model performs well in distinguishing between stroke and non-stroke cases, as evidenced by the comprehensive evaluation metrics.

XGBoost model showed better performance

XGBoost (Extreme Gradient Boosting) is an optimized distributed gradient boosting library. Yes, it uses gradient boosting (GBM) framework at core. Yet, does better than GBM framework alone. XGBoost was created by Tianqi Chen, PhD Student, University of Washington. It is used for supervised ML problems[5].

The xgBoost model (xgb_model) trained on the stroke dataset exhibited strong predictive performance, yielding a notable accuracy of 96.15%. The model showcased a remarkable ability to identify individuals at risk of stroke, with a sensitivity of 94.97%, while maintaining high specificity at 97.33% for correctly classifying non-stroke instances. The Kappa statistic, measuring agreement beyond chance, stood at 0.923, underscoring the model's robustness. Moreover, the positive predictive value (PPV) of 97.27% highlighted the model's precision in predicting positive cases, complemented by a solid negative predictive value (NPV) of 95.09%. The balanced accuracy of 96.15% reflected a well-calibrated equilibrium between sensitivity and specificity in stroke prediction.

Results and Discussions

The findings of this study shed light on critical aspects of stroke risk assessment, particularly in the context of the rising incidence among young adults. The comprehensive exploration of clinical features revealed unexpected insights, challenging preconceived notions about factors such as glucose levels and BMI. The Variable Importance in Projection (VIP) analysis emphasized the pivotal roles played by age, average glucose level, smoking status, heart disease, and hypertension in predicting stroke occurrences. The logistic regression model demonstrated a balanced accuracy of 81.13%, providing a robust foundation for risk assessment. Intriguingly, the introduction of the XGBoost model showcased superior predictive performance with an accuracy of 96.15%, suggesting the potential efficacy of advanced machine learning techniques in this domain.

This study contributes to the ongoing scientific discourse by not only enhancing our understanding of stroke risk factors but also by showcasing the utility of advanced modeling approaches. The unexpected correlations discovered, especially in glucose levels and BMI, underscore the need for nuanced risk assessments and challenge conventional assumptions in the field. While the logistic regression model proved effective, the XGBoost model's superior accuracy indicates avenues for further exploration into the application of machine learning in stroke prediction.

Limitations of this study include the inherent biases present in the dataset and the potential impact of unexplored variables. The exploration of additional clinical features and the inclusion of more diverse datasets could further refine predictive models. Despite these limitations, the study provides valuable insights into stroke risk assessment and sets the stage for future research endeavors, guiding early intervention and prevention strategies in the ongoing battle against the global health burden posed by strokes.

Link to data and code: https://github.com/Kaito999/stroke-risk-prediction

REFERENCES

1. WHO EMRO | Stroke, Cerebrovascular accident | Health topics. *World Health Organization - Regional Office for the Eastern Mediterranean* http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html.
2. Medical Research Foundation | World Stroke Day 2022: study reveals…. *Medical Research Foundation* https://www.medicalresearchfoundation.org.uk/news/world-stroke-day-2022-study-reveals-sharp-rise-in-stroke-cases-among-young-adults-in-oxford.
3. Find Open Datasets and Machine Learning Projects | Kaggle. https://www.kaggle.com/datasets.
4. Let&rsquo;s Talk About the Connection Between Diabetes&nbsp;and Stroke. *www.stroke.org* https://www.stroke.org/en/help-and-support/resource-library/lets-talk-about-stroke/diabetes.
5. Beginners Tutorial on XGBoost and Parameter Tuning in R Tutorials & Notes | Machine Learning. *HackerEarth* https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/.