Stroke Risk Factors

Julia Decker and Rai García-Figueras

[jcd163@miami.edu](mailto:jcd163@miami.edu), [rxg1292@miami.edu](mailto:rxg1292@miami.edu)

University of Miami

Prof. Vanessa Aguiar

CSC642

May 7, 2023

## Background / Introduction

A stroke is a serious medical condition that occurs when the blood supply to a part of the brain is interrupted or reduced, leading to those parts of the brain becoming damaged or dying. A stroke can result in permanent brain damage, long-term disability, or even death.1  Strokes are a leading cause of death around the world and, according to the World Health Organisation (WHO), account for approximately 11% of total global deaths. 2 The prevalence of stroke is expected to rise worldwide due to the increasing incidence of risk factors such as high blood pressure, diabetes, obesity, and the aging population.3

There are two main ways in which blood supply can be obstructed or prevented from reaching the brain, and thus there are two main types of strokes. The first and most common type of strokes are Ischemic strokes. These occur when a blood clot blocks an artery that supplies blood to the brain. On the other hand, hemorrhagic strokes happen when a blood vessel within the brain ruptures and bleeds into the surrounding tissue. Both types of strokes can cause sudden and permanent neurological damage, such as paralysis, difficulty speaking or understanding speech, and memory loss.3 The extent of damage depends on the location and size of the affected area in the brain, as well as how quickly medical treatment is received. There are many risk factors that can increase the probability of having a stroke. Researchers in the field have suggested that some contributing factors may include high blood pressure, smoking, diabetes or high blood sugar, high salt-intake, high cholesterol, heart disease, and a family history of stroke. 1,2,3,4

This study aims to answer the question: What biological and behavioral features are most prevalent in people who have experienced strokes?

Despite advancements in medical research and treatment options, predicting stroke occurrence in individuals is still a challenging task. Stroke prediction can be a critical step in identifying individuals at high risk of stroke, allowing for early intervention and prevention strategies to be implemented. By utilizing predictive models and risk assessment tools, clinicians can identify patients who may benefit from more attentive management of cardiovascular risk factors, such as hypertension, diabetes, and high cholesterol.1,3,4 Early identification of at-risk patients can also aid in the development of targeted public health campaigns aimed at reducing the burden of stroke on populations. Overall, stroke prediction is a critical component of stroke prevention and a valuable tool for healthcare providers in the fight against this debilitating disease.

State-of-the-Art

Machine learning models have become a key focus of recent stroke prediction research, with a goal of accurately identifying individuals at high risk based on demographic, clinical, and lifestyle factors. Most recent studies focus on classification techniques using common machine learning algorithms such as decision trees, random forests, support vector machines and neural networks among others. These models perform analysis on datasets with a combination of traditional risk factors like age, sex, hypertension, and diabetes, alongside novel factors such as genetics, biomarkers, stress biomarkers, food intake and imaging data. The potential for machine learning to harness vast, regularly compiled datasets and provide precise individualized preemptive factors is continuing to gain increasing interest in this field. 2,5

One such study, analyzed the same Kaggle dataset used in this project and aimed to design a robust framework for the long-term risk prediction of stroke occurrence5. The study evaluated various models, including naive Bayes, random forest, and decision tree algorithms. It also proposed a stacking method that combined the results of the four base classifiers and fed them into a logistic regression meta-classifier. The findings demonstrated that the stacking method outperformed other techniques with an AUC of 98.9%, F-measure, precision, and recall of 97.4%, and an accuracy of 98%. The study concluded that machine learning can be used to predict stroke early and prevent severe consequences. The stacking method was the best-performing approach and the primary recommendation of the study.5 It is also possible that over-fitting occurred to that particular data set as some variables were removed and the age bracket was limited.

Another similar study, which trained six different machine learning models with this same dataset, was performed by Gangavarapu Sailasya et al. (2021). The models used were Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors, Support Vector Machine, and Naïve Bayes Classification. The study did not combine or expand on these models like the previous one and therefore did not manage to achieve as high of an accuracy. Its best performing model was the Naïve Bayes Classification, which achieved an accuracy of 82%.

## 

## Methodology

The dataset used in the study is the Stroke Prediction Dataset found in Kaggle7. Each row within the dataset contains the information pertaining to a specific patient. The unaltered dataset has 12 attributes and 5110 instances for each attribute. They include a mixture of numerical and categorical data.

The attributes and ensuing values are as follows:

| **Attribute Name** | **Type (Values)** | **Breakdown** | **Description** |
| --- | --- | --- | --- |
| **1.** **id** | Numerical |  | Identification number for each patient in the dataset. |
| **2.** **gender** | Categorical ("Male", "Female" or "Other") | Female – 2994 (58.59%)  Male – 2115 (41.38%)  Other – 1 (0.02%) | Participant’s gender. |
| **3.** **age** | Numerical | Min: 0.08  Mean: 43.23  Max: 82 | Age of each patient. |
| **4.** **hypertension** | Binary (0,1) | 0 – 4612 (90.25%)  1 – 498 (9.75%) | If the patient suffers from hypertension (1) or not (0). |
| **5.** **heart\_disease** | Binary (0,1) | 0 – 4834 (94.60%)  1 – 276 (5.40%) | If the patient suffers from heart disease (1) or not (0). |
| **6.** **ever\_married** | Binary (‘Yes’,’No’) | Yes – 3353 (65.62%)  No – 1757 (34.38%) | If the patient has ever married. |
| **7.** **work\_type** | Categorical ("children", "Govt\_job", "Never\_worked", "Private" or "Self-employed”) | Children – 687 (13.44%)  Govt\_job – 657 (12.86%)  Never\_worked – 22 (0.43%)  Private – 2925 (57.24%)  Self-emploeyed – 819 (17.42%) | Type of employment the patient is involved in. |
| **8.** **residence\_type** | Binary ("Rural" or "Urban”) | Rural – 2514 (49.20%)  Urban – 2596 (50.80%) | Type of residence the patient lives in. |
| **9.** **avg\_glucose\_level** | Numerical | Min: 55.12  Mean: 106.15  Max: 271.74 | Average glucose level in blood measured in mg/dL. |
| **10.** **bmi** | Numerical | Min: 10.3  Mean: 28.89  Max: 97.60 | Body Mass Index (BMI) measurement for each patient |
| **11.** **smoking\_status** | Categorical ("formerly smoked", "never smoked", "smokes" or "Unknown") | Formerly smoked – 885 (17.32%)  Never smoked – 1892 (37.03%)  Smokes – 789 (15.44%)  Unknown – 1544 (30.22%) | Smoking habits of each patient. |
| **12.** **stroke** | Binary (1,0) | 1 – 249 (4.87%)  0 – 4861 (95.13%) | Target variable informing on whether a patient has suffered a stroke (1) or not (0) |

*Table 1: Variables information adapted from Healthcare dataset on Kaggle*

Pre-Processing and Data Cleaning

The dataset had to be pre-processed before any analysis could be performed. The first step involved removing unnecessary attributes from the dataset. The variables id and work\_type were removed and not included in our analysis. ‘id’ was removed as it did not contribute any meaningful information regarding the subjects of the study, while we decided to remove work\_type as, throughout testing, it would lower the accuracy of the models when included. The next step involved cleaning the data with the removal of missing values. The dataset contained 201 missing values within the ‘bmi’ attribute as well as 1544 entries in smoking\_status labeled as ‘unknown’; these were all removed. After removing the data, including patient information and additional details, the dataset was reduced to 3426 instances.

Next, we had to deal with the unbalanced data of our class variable stroke to avoid a biased model performance. After removal of the missing values, the stroke variable contained 180 positive instances of stroke and 3246 negative instances. To fix this we used the ROSE (Random Over-Sampling Examples) R package to generate synthetic data. ROSE is based on bootstrapping and is designed to assist with binary classification when dealing with rare classes. After running the dataset through ROSE, the resulting class variable is comprised of 1724 and 1702 instances of negative and positive cases respectively, a much more manageable difference. After this, the dataset is finally ready to be split into the training and testing sets. The dataset was divided into an 80/20 split for training and testing respectively. The training set was made up of 2740 of the samples while the testing had the remaining 686 samples.

Models

Our analysis involved training four supervised learning machine learning models on the training dataset for the correct classification of our class variable, stroke. The models used were logistic regression, decision tree, random forests, and support vector machines (SVM). The models were then evaluated using the testing set to calculate the accuracy of their classifications. This section explains each model in a bit more detail as well as talk about their implementation in R.

Logistic Regression

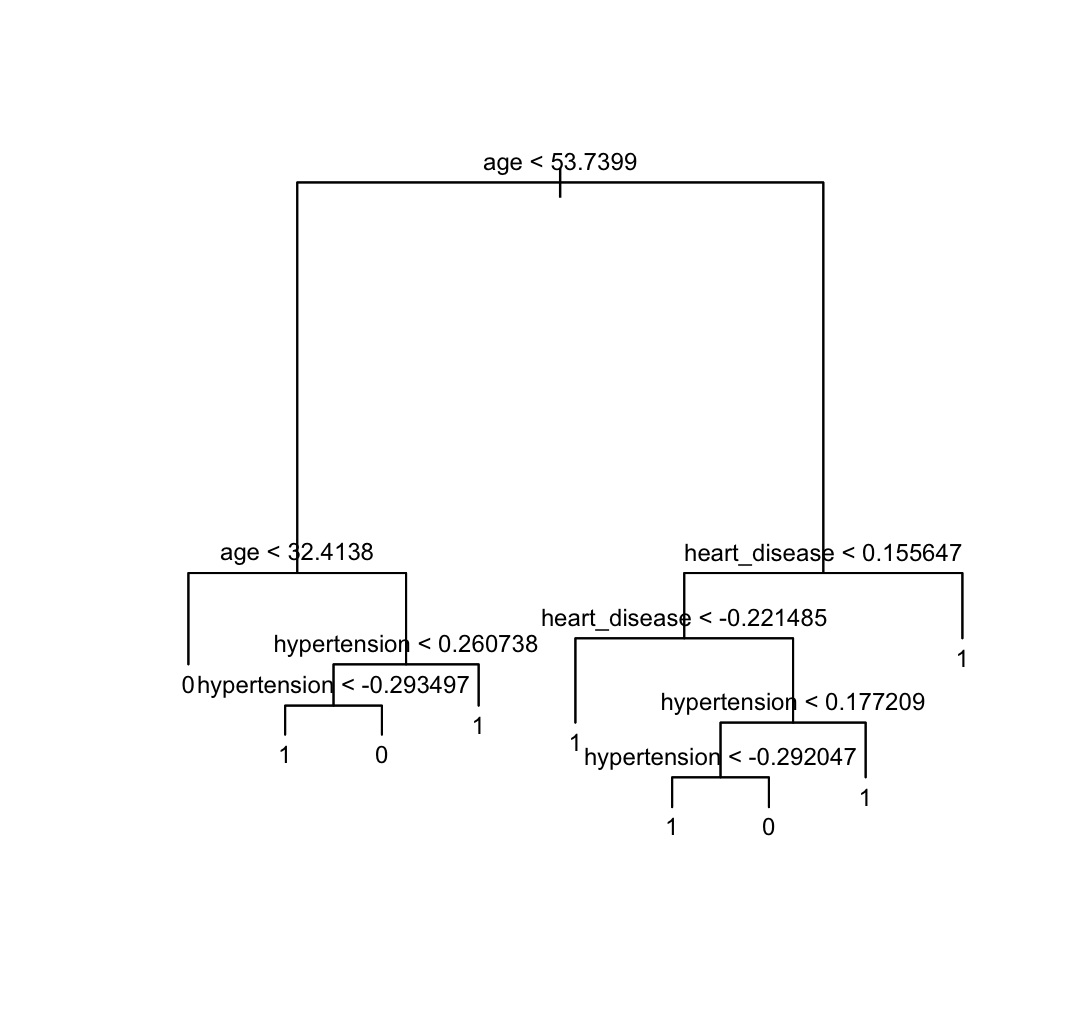
Logistic regression is a statistical technique that investigates the relationship between one or more predictor variables and a binary outcome variable. The outcome variable usually takes binary values, such as yes or no, true or false, or 0 and 1. The model of logistic regression leverages a logistic function to quantify the link between the predictor variables and the likelihood of the binary outcome. This logistic function generates an S-shaped curve that ranges from 0 to 1, and it represents the probability of the outcome variable being 1, given a certain set of predictor variables. By estimating the coefficients of the predictor variables, logistic regression can forecast the probability of the outcome variable.

In R this model is fitted through the general linear model function (GLM), with the binomial value for the family parameter.

Decision Tree

Decision trees are a popular machine learning algorithm for classification and regression tasks. They make decisions by repeatedly dividing the dataset based on the values of the predictor variables until every observation has a decision or prediction attached to it. This iterative process selects the most suitable variable to split the data based on criteria like information gain or Gini impurity. The final product is a decision tree that can then be applied to new data by following the split decisions. Not only can decision trees handle both numerical and categorical data, but they're also straightforward to interpret. However, these trees may not generalize well to new data and are more susceptible to overfitting than other models.

This algorithm is implemented in R through the ‘tree’ package. We retained the default parameters in the building of our model. Hyperparameter optimisation for the "size" parameter of the decision tree model was done using cross-validation to see if pruning the tree for better results was necessary. However, the optimisation proved that the best tree ‘size’ was already being employed in the original model. The tree diagram of the model can be seen in figure 1 below.



*Figure 1: Tree Diagram*

Random Forest

Random Forest is a powerful ensemble learning algorithm that addresses the overfitting problem of decision trees by generating multiple trees, each built using a different subset of the available features and a random sampling of the training data. These trees then work together to generate a final prediction, with their individual predictions being aggregated using the mode (for classification problems) or the mean (for regression problems). By leveraging this technique, Random Forests can handle high-dimensional data with a mix of continuous and categorical variables, while remaining versatile, scalable, and robust. These properties have made Random Forests a popular choice for tackling classification, regression, and feature selection challenges across a variety of domains.

​​ We used the randomForest() function to implement the random forest model. The number of variables used at each split was set to the square root of the number of columns, and the model was trained with importance set to TRUE to obtain a measure of variable importance.

Support Vector Machines (SVM)

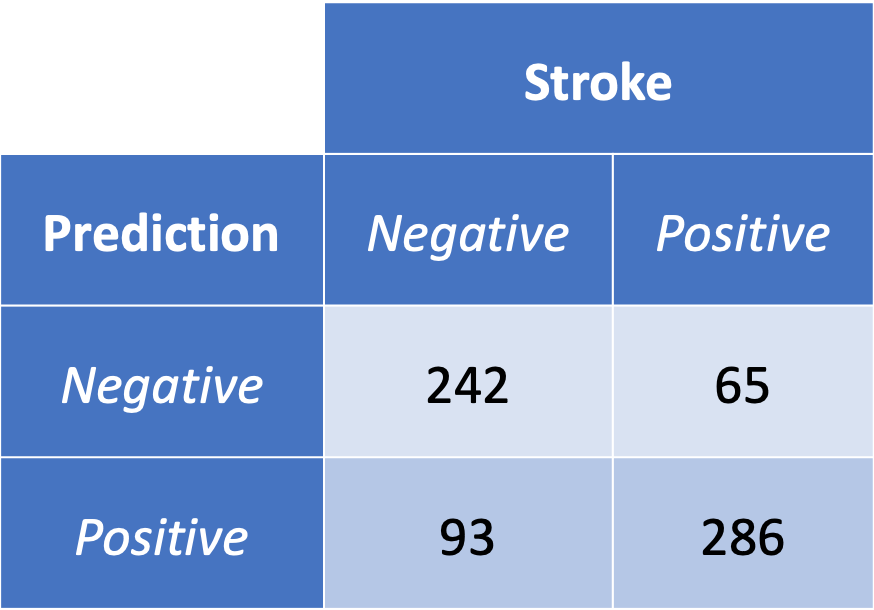
SVMs are another common machine learning algorithm used for both classification and regression problems. It works by mapping the data into a feature space with a high number of dimensions, and then identifying a hyperplane that partitions the data points into distinct classes. Notably, SVM selects the hyperplane with the largest margin, which is the distance between the hyperplane and the nearest data points. This makes SVM a good choice for problems with intricate decision boundaries, as well as for processing large datasets efficiently. Additionally, SVM is particularly useful in problems with high-dimensional feature spaces, where linear classifiers might struggle to perform.

In this study, two SVM models were implemented using the e1071 library. The first SVM model was a linear kernel with cost=10, and the second SVM model was a radial kernel with gamma=1 and cost=1. Cross-validation was performed to tune the parameters of both SVM models, and the best-performing model was selected to predict the outcome of the testing dataset.

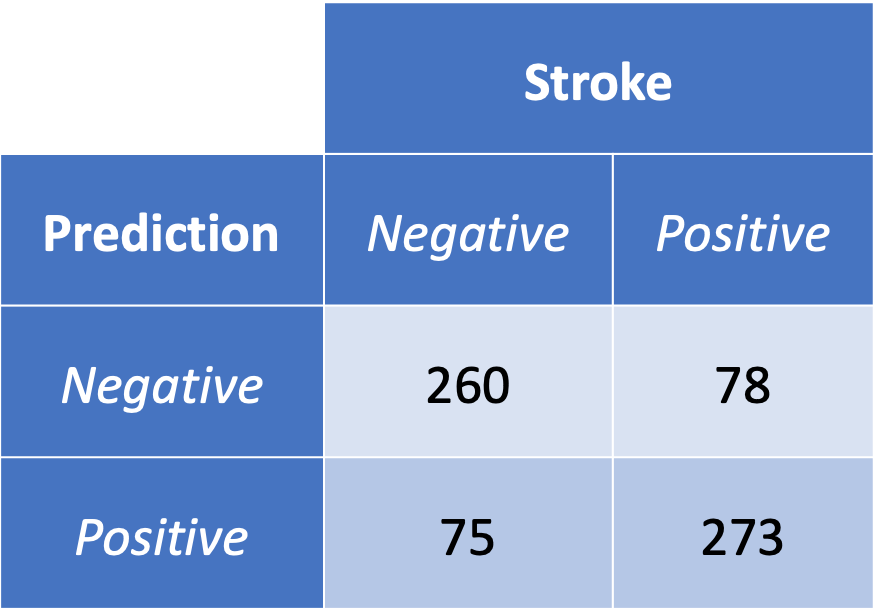
Results

The results showed that Random Forest Classification had the best accuracy, precision, recall and F1 score. This model used for this evaluation had 500 trees and a variable tested per split size of 3.

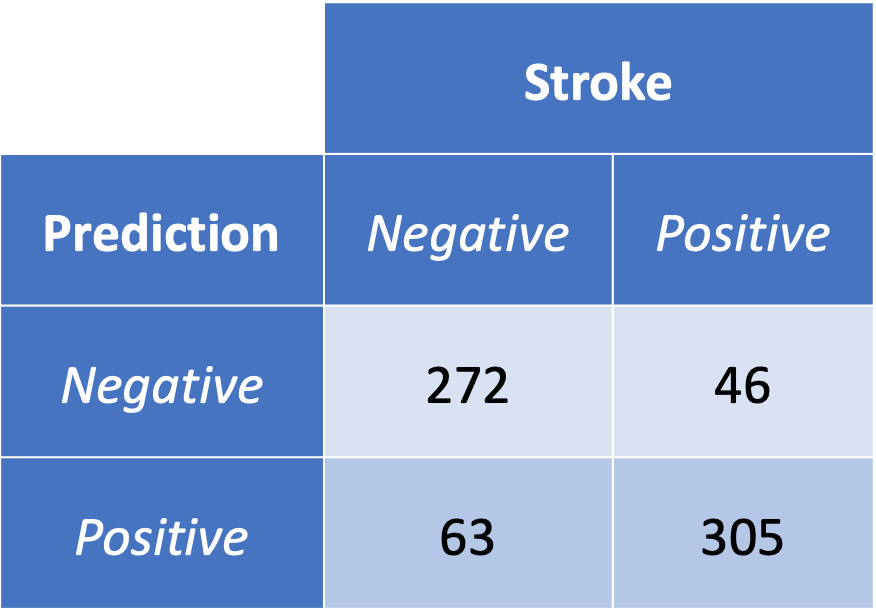
Logistic Regression



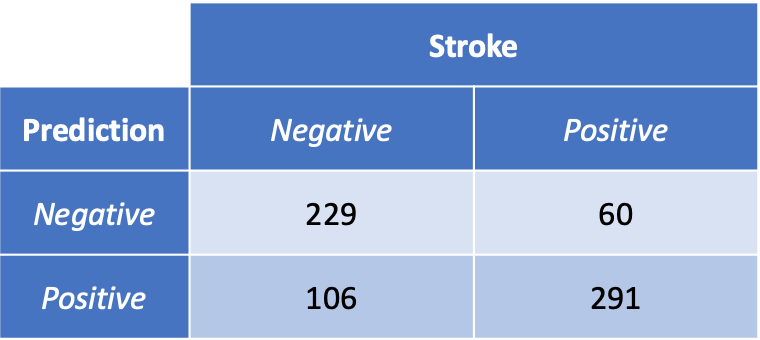
DT



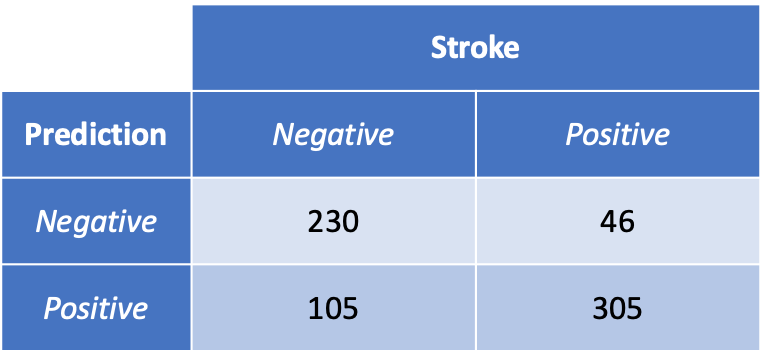
Random Forest

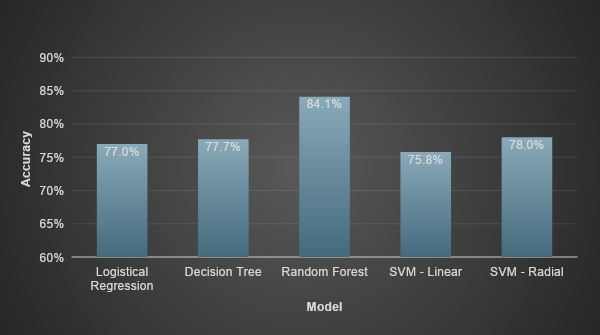


SVM - Linear Kernel



SVM - Radial Kernel

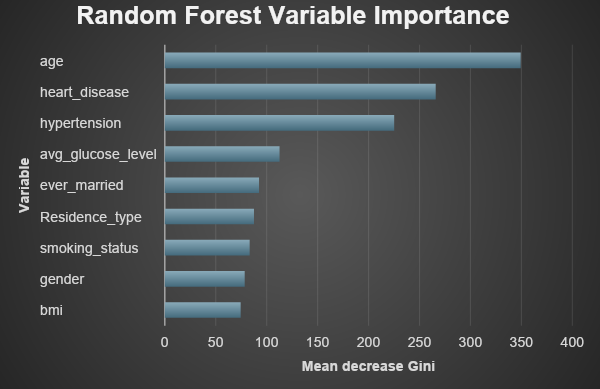




*Figure 2: Accuracy percent for each model*

| model | Accuracy | Precision | Recall | F1 Score |
| --- | --- | --- | --- | --- |
| Logistic Regression | 77.0% | 81.5% | 75.5% | 78.4% |
| Decision Tree | 77.7% | 77.8% | 78.4% | 78.1% |
| **Random Forest** | **84.1%** | **82.9%** | **86.9%** | **84.8%** |
| SVM - Linear | 75.8% | 73.3% | 82.9% | 77.8% |
| SVM - Radial | 78.0% | 74.4% | *86.9%* | 80.2% |

*Table 2: Comparison metrics of each model*



*Figure 3: Gini ( top variable importance) score for the best RF model*

Discussion/Conclusion

This analysis tested the data using logistic regression, a single decision tree, a random forest and two types of support vector machines to create classifier prediction models. Among the models tested the most accurate was the random forest classifier. It also had the best precision, recall and F1 scores. The Gini score for this model suggests that the most relevant variables used in prediction were age, heart disease, hypertension, average blood glucose, and marriage. Future statistical analysis might include other tests such as stacking models as well as more expansive degrees of hyper parameter tuning.

By running this analysis, we gained some possible insight into the biological and behavioral features that might be most prevalent in people who have experienced strokes. This analysis predicted that among the factors tested, those most correlated with the onset of a stroke event include increased hypertension, heart disease, average glucose level, age and smoking. While this study provided some information on the topic, it is not conclusive whether any of these are causal. In some ways, there might be a variety of intermediary factors or conditions that would lead to both the onset of stroke as well as the predictor. For example, whether the root cause of high glucose level manifests from metabolic disorder, activity levels or nutritional choices. And other factors like age are not modifiable which does not provide any insight to possible interventions.So it could be interesting to compare samples in set ranges of age, with others similar in age to see if there is any correlation to potential risk factors that might affect certain age populations. Overall, studies like this may provide a foundation for insights towards lifestyle recommendations with the hopes of mitigating risk for individuals through awareness strategies. To expand upon this model future studies might include other variables as predictors. In particular, some interesting information that might also be interesting to consider could be in biomarkers related to sleep quality, blood tests, alcohol intake, food choices like nutrient consumption and salt intake levels, chronic stress levels and emotional health.3,8,9 Another nuance to consider are the different types of strokes such as ischemic vs hemorrhagic, and strokes that may be correlated with environmental factors such as heat stroke.

Sources:

1. Feigin VL, Norrving B, George MG, Foltz JL, Roth GA, Mensah GA. Prevention of stroke: a strategic global imperative. *Nature Reviews Neurology*. 2016;12(9):501-512. doi:https://doi.org/10.1038/nrneurol.2016.107
2. Sailasya G, Kumari GLA. Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. *International Journal of Advanced Computer Science and Applications*. 2021;12(6). doi:<https://doi.org/10.14569/ijacsa.2021.0120662>
3. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. *Circulation*. 2022;145(8). doi:https://doi.org/10.1161/cir.0000000000001052
4. Sharma S. Prevent brain stroke. www.who.int. Published October 29, 2016. https://www.who.int/southeastasia/news/detail/29-10-2016-prevent-brain-stroke
5. Dritsas E, Trigka M. Stroke Risk Prediction with Machine Learning Techniques. *Sensors*. 2022;22(13):4670. doi:<https://doi.org/10.3390/s22134670>
6. Centers for Disease Control and Prevention. Stroke Facts. Centers for Disease Control and Prevention. Published May 25, 2021. <https://www.cdc.gov/stroke/facts.htm>
7. fedesoriano. Stroke Prediction Dataset. www.kaggle.com. Published 2021. Accessed April 4AD. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
8. Iacoviello L, Bonaccio M, Cairella G, et al. Diet and primary prevention of stroke: Systematic review and dietary recommendations by the ad hoc Working Group of the Italian Society of Human Nutrition. *Nutrition, Metabolism and Cardiovascular Diseases*. 2018;28(4):309-334. doi:<https://doi.org/10.1016/j.numecd.2017.12.010>
9. O’Donnell MJ. The Roles of Nutrition and Physical Activity in Chronic Disease in North Carolina. the Lancet. 2010;376(9735). doi:<https://doi.org/10.1016/S0140-6736(10)60834-3>

Github:

<https://github.com/DeckerDreams/Statistics>