Health Risk Factors Contribution in Predicting Stroke Risk

May 3, 2025

Hannah Abele, Dean Chenzaie, Diya Gupta, Rhea Hemrajani, Kelly Punsalan, Nora Wiktorowicz

Executive Summary

This study explores the predictive value of various health risk factors in determining stroke risk using machine learning classification methods. Drawing on the Kaggle "Stroke Risk Prediction Dataset," we analyzed 7,000 patient records that include binary indicators for key symptoms (e.g., chest pain, shortness of breath, dizziness, anxiety, nausea) and demographic data such as age. Our central research question asked: *Which health risk factors are most useful in predicting stroke risk?* We applied both logistic regression and gradient boosting models (GBM) to compare predictive performance and interpretability. The logistic regression model initially appeared to yield perfect classification results (F1-score = 1.00), but upon further examination, we discovered this likely resulted from overfitting or improper validation, such as training and testing on the same data. To improve robustness, we turned to a gradient boosting model, which builds sequential decision trees to minimize prediction error. This model achieved a 96% overall accuracy, with F1-scores of 0.94 for stroke-negative and 0.97 for stroke-positive predictions. Feature importance analysis revealed that age was by far the most significant predictor of stroke, with a feature importance score of 0.61—substantially higher than the next most influential factors: anxiety (0.03), nausea/vomiting (0.028), and cold extremities (0.027). Some features, such as neck/jaw/back pain and swelling, ranked lower in importance. Our findings aligned with existing medical literature emphasizing the strong link between aging and stroke risk, though the prominence of psychological symptoms like anxiety was unexpected and warrants further exploration. Challenges included managing variable correlation (e.g., fatigue and shortness of breath) and addressing concerns about the dataset's origin and representativeness. We noted possible biases in participant recruitment and the limitations of binary symptom encoding in capturing clinical nuance. Future work will involve improving data validation procedures, extending the model to more diverse populations, and exploring causal pathways behind age-related risk. This research demonstrates the potential of ensemble learning methods in healthcare prediction while emphasizing the importance of rigorous data integrity and thoughtful model evaluation.

Introduction

Stroke remains one of the leading causes of death and long-term disability worldwide. According to the World Health Organization, nearly 15 million people suffer from strokes each year, with roughly five million left permanently disabled. Despite being so prevalent, strokes are notoriously difficult to predict with high precision, particularly due to the interplay of both physiological and psychological symptoms that often go unnoticed or unreported until it's too late. In light of this, our study seeks to explore how health risk factors, ranging from chest pain and nausea to anxiety and fatigue, can be used to predict stroke risk using machine learning models.

The central question guiding our research is: Which health risk factors are most useful in predicting stroke risk? By answering this, we hope to offer not only insight into the key symptoms that warrant closer clinical attention, but also to demonstrate how machine learning can support preventative healthcare strategies. Our dataset, drawn from the publicly available Kaggle "Stroke Risk Prediction Dataset," consists of 7,000 anonymized patient records. Each record includes binary indicators (1 = symptom present, 0 = symptom absent) for various health risk factors and one continuous variable: age. These variables were selected based on established stroke symptoms outlined by medical institutions such as the American Stroke Association, Mayo Clinic, and the WHO.

We used two different machine learning approaches to model stroke risk: logistic regression and gradient boosting machines (GBM). The logistic regression model appeared to achieve perfect accuracy (F1-score = 1.00), but upon further investigation, we suspected overfitting or flawed testing procedures such as training and testing on the same data. To address this, we turned to a GBM model, which builds decision trees sequentially to correct errors and improve prediction. This model proved to be significantly more robust, achieving 96% accuracy and producing a ranked list of feature importances that allowed us to better interpret which symptoms matter most.

One of our most notable findings was that age emerged as the most important predictor of stroke risk by a wide margin, with a feature importance score of 0.61. This aligns with established medical literature linking age and stroke risk, particularly for individuals over the age of 65.

Surprisingly, psychological symptoms such as anxiety/feeling of doom ranked second, which was unexpected and indicates potential underexplored relationships between mental health and stroke. Other influential but less significant features included nausea/vomiting and cold extremities, while symptoms like neck/jaw/back pain and swelling ranked lower in importance.

In building our models, we encountered several challenges. These included issues with variable correlation, such as the overlap between fatigue and shortness of breath, and concerns about dataset representativeness and validity. Given that the data was self-reported and binary in format, it lacked the nuance found in clinical settings. This limitation raised questions about sampling bias, especially considering that participation may have skewed toward more tech-savvy individuals or those with easier access to digital platforms. Moreover, the binary encoding of symptoms failed to capture intensity or duration, which could have improved model sensitivity.

Despite these challenges, our results highlight the power of ensemble learning in identifying patterns that could be life-saving. Our study demonstrates how machine learning can be an effective complement to clinical judgment—flagging patients who may otherwise fall through the cracks and prompting early interventions. As we proceed through the rest of this paper, we will detail the methodology used to develop and validate our models, discuss the significance of each feature in the prediction process, and explore the ethical and practical considerations surrounding data reliability. Ultimately, our goal is to provide actionable insights that improve stroke prediction and inform future work in preventative healthcare modeling.

Data

Our dataset examines key risk factors that contribute to a person's likelihood of experiencing a stroke. The studied variables include chest pain, shortness of breath, irregular heartbeat, fatigue, dizziness, swelling (edema), pain in the neck/jaw/shoulder/back, excessive sweating, persistent cough, and nausea/vomiting. These factors were selected through analysis of leading medical authorities such as the American Stroke Association, Mayo Clinic, Cleveland Clinic, Harrison's Principles of Internal Medicine, Stroke Prevention, Treatment, and Rehabilitation, The Stroke Book, and the World Health Organization. Each of the 7000 records represents an individual's medical symptoms and risk assessment, with binary indicators (1 = symptom present, 0 =

symptom absent). Age is also included as a critical factor, given that an increase in age is strongly associated with a heightened risk of stroke.

According to UTMB health, stroke is the fifth leading cause of death for US citizens, and is a prominent cause of long-term disability ("Did You Know:"). Since stroke is such a prevalent cause of health issues in the US, the phenomenon we are interested in is calculating the risk of a stroke based on the presence or absence of the following symptoms: chest pain, shortness of breath, irregular heartbeat, fatigue and weakness, dizziness, swelling (edema), pain in neck/shoulder/back, excessive sweating, persistent cough, nausea/vomiting, high blood pressure, chest discomfort (activity), cold hands/feet, snoring/sleep apnea, anxiety/feeling of doom, and age. Stroke symptoms are often overlooked or attributed to other health issues, leading to delayed diagnosis, inadequate preventative measures, and increased risk of severe complications or fatal outcomes (Johnson). According to the CDC, stroke prevalence has increased by 15.7% for adults aged 45 to 64 and 14.6% for adults aged 18 to 44 from 2011 to 2023 (Torres). As the age of onset for strokes is becoming younger and younger over time, this study becomes increasingly relevant to American health. This dataset will be useful for examining the risk of stroke by age group, the risk factors that are majorly present in the American population, and the correlation between each risk factor and the overall risk of stroke, which enables us to determine which risk factor is most correlated with incidence of stroke. Through this analysis, we hope to identify a better way to quantify the risk of stroke, enhance our understanding of its causes, and perhaps provide insight into ways we can better combat it.

However, there are several challenges that we initially ran into with our data. Luckily, it could be resolved with some initial cleaning. All the variables initially presented as "0,1." Thus, we changed them to "Y,N" in order to more easily interpret our variables (as the "0,1" represented "Y,N"). This will make it easier to analyze our data in the future. Next, we needed to ensure that our code was able to create a "count" for the binary variables to be able to perform analysis, such as calculating percentages of the symptoms likelihood to impact having a stroke.

Although we fixed some initial issues, we anticipate several future challenges. First, we may need to alter the "age" variable when creating visualizations and descriptions, as it is currently an integer (float variable), while all other variables are binary variables. Second, we anticipate that several of the variables will correlate. For example, the variable "anxiety and feeling of doom"

may closely relate to the variable “increased heart rate” or “high blood pressure,” as anxiety causes tension, sweating, increased blood pressure, and increased heart rate. During our analysis, it may be difficult to discern which variables independently contribute to increasing the likelihood of having a stroke.

Overall, even though our data does present some challenges, creating a model that can discern stroke risk from simple health data can have a large impact on future stroke preventative measures.

Method

We selected to use the Kaggle dataset, “Stroke Risk Prediction Dataset.” To recap, our research question is as follows: Which health risk factors are most useful in predicting stroke risk? One observation in our study is one individual, specifically their age and risk factors. Examples of risk factors in the dataset include chest pain, shortness of breath, and irregular heart rate.

We are utilizing classification to model our research question. Our data is categorical, showing a “1” if the individual experiences a risk factor, and “0” if they do not. Classification is most appropriate when working with categorical data. As we are not predicting numerical values, regression analysis may not be appropriate. Furthermore, we will be utilizing supervised learning. In supervised learning, the model is able to predict outcomes for new data using predictions or classifications. As the labelled datasets have a known input and corresponding output (“0” or “1”), the model will ideally be able to predict unforeseen inputs. Therefore, this may be better suited for categorical data and a classification model.

Modeling

Next, we determined which modeling approach to utilize. We will start by using a train-test model to accurately predict stroke risk. We will incorporate our own data to ensure accuracy. A logistic regression model will then be used to predict stroke probability based on the coefficient values, where larger, positive values are indicative of a higher risk for stroke. We will contrast this model with a gradient boosting model (GBM). This model builds sequential decision trees, where each tree corrects the errors of the previous one. The GBM ranks risk factors based on

their importance of predicting stroke, providing probability scores for stroke risk, which will help us identify key contributors.

Measuring Success

To measure the success of our approach, we will assess outcomes of our prediction model with a confusion matrix, and look at the accuracy rate of the predictions. When assessing the confusion matrix, we will know that our predictions have been successful if our model can accurately predict true positives and true negatives for a high percentage of individuals. Furthermore, we can utilize a F1- score to evaluate the accuracy of the classification. This utilizes the number of false positives and false negatives to create a score. A score of "1" shows perfect precision, while a "0" shows poor precision. A higher score will show that our model correctly minimizes the number of false positives and false negatives.

Considerations

While we are very confident in our dataset and the ability to analyze it, there are a couple of weaknesses that have stood out to us. First, there are a large number of variables in our dataset. Even though at times a larger number of variables does appear to provide clarity, as there are more metrics to analyze, it can also be counterproductive when there are too many observations to comb through. Furthermore, this leads into the second weakness that goes along with too many variables: correlation. Specifically looking at the variables in this dataset that impact stroke risk, a lot of them can correlate with each other. For example, anxiety and irregular heartbeat and shortness of breath are typically correlated, and together could have an exponential impact on stroke risk. This correlation and covariance between variables will need to be addressed. If the approach fails, from this unfortunate outcome we may learn that our model or strategy is not properly suited for the dataset. We may move forward utilizing a different modeling approach.

Significant preparation isn't necessitated for the data as it is already aggregated and collated in the data repository we've appropriated for this project. However, if we were to extend the scope of our data usage, it would be necessary to converge the disparate data streams, and likely would require us to Extract, Transform, Load (ETL) to ensure legibility and searchability. Currently, our chief concern is one-hot encoding the categorical variable responses currently maintained in the

dataset into binary variables (0,1). PCA could be utilized when analyzing the dataset to understand the covariance and correlation between variables, considering the high volume of variables in this dataset that could have high covariance. Detecting the correlation and covariance between variables may then help us simplify our dataset, facilitating better data integrity.

Communication of Results

In terms of communicating our results, we will present: 1) a table of regression coefficients, 2) a gradient boosting model, and 3) a confusion matrix. Our table of regression coefficients will show which risk factors are more strongly correlated with stroke risk. If $p<0.05$, the variable is statistically significant. Next, our gradient boosting model will give us probability scores for each risk factor, where a higher probability score is more indicative of stroke risk. Lastly, we will use a confusion matrix to analyze the accuracy of the predictions, identifying the number of false positives and negatives.

Results

Our prediction question was as follows: Which health risk factors are most influential in predicting stroke risk? To address this, the team created a train test model on the data, a logistic regression model, and a gradient boosting model.

Logistic Regression Model

When we created the logistic regression model to predict the binary stroke risk target variable, the resulting confusion matrix showed 0 false positives and 0 false negatives. In other words, the model perfectly classified all instances. The F1 scores for both classes were 1.00, indicating 100% accuracy in predicting both true positives and true negatives. While these results might seem ideal, we were initially suspicious of this result because we thought it would be unlikely for a simple predictive model to have perfect accuracy on real-world data. We thought there could be something wrong with the model, like for example testing it on the same training dataset might not be meaningful for the actual performance of the model. With these findings in mind, we recognized that we should try another kind of predictive model to contrast with these results.

Gradient Boosting Model

Next, we ran the gradient boosting model. This provided a more accurate and well-rounded analysis, as it ranked the features by level of importance. This allowed us to more effectively answer our research question, of which health risk factors are most influential in predicting stroke risk. Results show that age is by far the most important feature in predicting stroke risk, scoring 0.610267 in importance. This is followed by anxiety/feeling of doom, nausea/vomiting, and cold hands/feet, scoring 0.029289, 0.027703, and 0.027550, respectively. The features that ranked the least important, in order of increasing importance, were pain in neck/jaw/shoulder/back, chest discomfort, and swelling, scoring 0.022954, 0.023266, and 0.024520, respectively.

The confusion matrix and classification report show that the gradient boosting model is highly accurate. The confusion matrix shows that the model predicted 4,483 true negatives (predicted 0 and it was 0), and 409 false negatives (predicted 0 and it was 1). The model predicted 183 false positives and 8925 true positives. When predicting false results (0's), the model was 94% accurate (as the f1-score is 0.94 for "0" results). Therefore, the model predicts more false negatives (predicting 0 when the score is 1). This may have harsh consequences, as in the healthcare world it is unhelpful if the model is predicting a patient is not at risk, when they are. When predicting true positive results (1's), the model was 97% accurate (as the f1-score is 0.97 for "1" results). Overall the model is 96% accurate in making predictions. This shows that the gradient boosting model accurately predicts stroke risk by feature, overall.

As discussed in the results analysis above, age is by far the most important feature in predicting stroke risk. This matches the team's initial hypothesis, as age is frequently associated with stroke risk and increasing risk of other factors. Our external research showed that after 65, your risk of a stroke greatly increases. It is important to note that age may be highly associated with other features, as several of these features increase with age (chest discomfort, pain in neck/jaw/shoulder/back, fatigue and weakness). We found anxiety/feeling of doom to be a surprising result, scoring second in importance in predicting stroke risk. We were not previously aware that anxiety/feeling of doom was a symptom or predictor of a stroke.

Conclusion

In Summary

In this project, we used the Kaggle "Stroke Risk Prediction Dataset" to explore which health risk factors most accurately predict stroke risk, using classification models suitable for our categorical data. We built a logistic regression model and a gradient boosting model (GBM) to assess feature importance and prediction accuracy. While the logistic regression model initially showed perfect accuracy, further investigation revealed limitations, likely due to testing flaws. The GBM, by contrast, offered more robust and interpretable results, with "age" emerging as the most influential factor, far surpassing others like anxiety and nausea. Challenges included data cleaning issues and variable correlation, which may have affected regression performance and feature importance rankings. As next steps, we plan to further investigate the relationship between aging and stroke risk, expand our dataset to include younger stroke cases, and explore discrepancies between statistical correlation and real-world causality. These efforts aim to improve model precision and help identify meaningful intervention points for stroke prevention.

Challenges

While initially analyzing our data, we encountered a few concerns that needed to be addressed. Most of our dataset was binary, meaning that the numbers '0' and '1' represented whether the stroke risk symptom was present or not, respectively. When we initially cleaned the data, we determined that it would be easier to visualize the entire file if the '0's and '1's were converted into 'Y' and 'N', which we performed. However, this presented challenges when performing regression and other analysis. We needed numbers, not Y or N. Therefore, when we were initially running our tests, we were getting errors and even precision reports with a value of 1, telling us our model was perfect, when we could easily ascertain it was not considering the complexity of the subject at hand. As a result, we pivoted towards the gradient boosting test to represent our data more accurately. This overall inability to perform a completely correct linear regression model is still one of our primary concerns. Furthermore, our variables could be highly correlated, such as age, pain in neck, shoulder, and back, and chest pain. This would skew our regression results through coefficients that are higher than they should be. Consequently, feature

importance ranking may be inaccurate and something we would definitely focus on a lot more had we had more time to explore the subject.

An overarching challenge with our dataset was the verifiability. We discovered that the source of our dataset may not be legitimate late in our progress, which is obviously a significant caveat that must be explicated. While the dataset was very convenient due to the fact it's highly structured, it's important to bear in mind the convolution of the topic at hand and how a dataset that relies simply on binary answers, without any incorporation of nuance, may not provide the most robust answers. While we weren't privy to this data, it's also important to mention that there may be concerns about how the sample data was collected. Doctor-patient confidentiality and HIPAA regulations ensure that medical records are sealed to the general public, and can only be accessed by physicians and under court-orders. Thus, when crafting a dataset such as this one, all information was probably given over electively, and it's an important facet of any research study to determine why any individual would share their information. In this case, there's no indication that individuals were compensated for their participation, so this may not have been a factor, but the mode of communication may be. For instance, should this study have publicized on social media that they were looking for participants, only those who are tech-savvy may have come across it. Thus, those who have limited or no social media presence may have never encountered it, such as the elderly.

Next Steps

Given the analysis we've done so far, it appears that it would be fruitful to continue to delve into factors we've identified that are of significance. For instance, age was the most significant predictor of stroke risk with a relatively high score (0.61), however the next predictor, anxiety, is a very distant second (0.02). Therefore, further research into what aspects of aging contribute to the increased risk – whether that be habits, lifestyle choices, or anything else – would likely even further improve our accuracy score. Additionally, even those below certain age thresholds are definitely prone to stroke risks, instances of which may have been neglected in our sample. Thus, going forward, we should try to cast a wider sample to ensure that those prone to strokes outside those traditional age categories can be detected and their risk mitigated. Additionally, we should seek clarity on the modes of communication being used to recruit participants, ensuring that the sample is representative across all demographics. Furthermore, it would be worthwhile to

investigate why the correlation for certain variables is high while their feature importance is rather low. Determining why this disparity between correlation and causation exists may be paramount, so as to determine if – notionally – there are certain traits that may appear in those of advanced ages that could be unrelated but tend to overlap. Having a deeper, more comprehensive understanding of the different variables involved could allow us to simultaneously improve predictive accuracy and ascertain what preventative measures could be taken before anything adverse occurs.

References

"Did You Know: The Truth About Stroke." *UTMBHealth*, www.utmbhealth.com/services/neurology/procedures-conditions/stroke/stroke-facts. Accessed 2 May 2025.

Johnson, Alison. "The Consequences of a Delayed or Missed Diagnosis of a Stroke." *UK and International Law Firm*, Penningtons Manches Cooper, 8 Feb. 2023, www.penningtonslaw.com/news-publications/latest-news/2023/the-consequences-of-a-delayed-or-missed-diagnosis-of-a-stroke#:~:text=If%20there%20is%20a%20delay,speech%2C%20mobility%20and%20personality%20changes.

Torres, Amanda. "What to Know about the Rising Stroke Rates in Younger People." *NewYork-Presbyterian*, 27 Jan. 2025, healthmatters.nyp.org/what-to-know-about-the-rising-stroke-rates-in-younger-people/#:~:text=The%20rates%20of%20stroke%20are,Joshua%20Willey.