

A predictive analytics approach for stroke prediction using machine learning and neural networks

Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, Nishtha Jain, Bharadwaj Veeravalli, Deepu John

Abstract:

The medical condition known as the stroke is when the blood supply to the brain in a human's body is interrupted causing brain cells to die leaving many negative effects like speech difficulties, paralysis, memory issues and brain bleeding. A stroke can affect anybody but tends to be higher in people with high blood pressure, diabetes, family history of strokes, obesity, and age. This paper analyzes different factors in EHR for stroke prediction in an attempt to be able to help identify individuals at higher risk of having a stroke. In the model, the Convolutional Neural Network model was the best but there were also experiments with SVM, Lasso, and Elastic Net. The NN had a precision of 0.80, SVM with 0.67, Lasso with 0.78, and ElasticNet with 0.79.

Introduction:

The problem described in this paper is that strokes are one of the leading causes of death in the United States. Stroke prediction could lead to early detection in patients with high risk factors which could lead to an overall decrease in the amount of strokes in the United States. The logic of the paper is to aggregate several risk factors from a given dataset of patient electronic health records and use these features to predict the probability of a specific patient of actually having a stroke. The paper highlights several different machine learning algorithms which were used on the data and showed that many of them were relatively successful. With more and cleaner data, even more accurate predictions would most likely be able to be made.

Why is this paper important to the discipline:

This paper is important to the discipline of artificial intelligence within the medical field because by studying patient records you could increase overall patient health and wellness. In this specific case the medical condition being focused on is strokes, but since this was proved to be effective, similar techniques can be used for other conditions like cancer and heart attack prediction, two other leading causes of death in the United States. New technologies like machine learning and artificial intelligence could revolutionize how we perceive, prevent, and treat patient health.

Research question:

The key question that this paper is addressing is whether we can accurately predict whether a stroke will occur in a patient based on several risk factors.

Methodology:

Our first method is further understanding which deep learning model we want to work with in accurately predicting a patient will experience a stroke based on several risk factors. Our options are using the Convolutional Neural Network model (CNN), SVM, Lasso, and Elastic Net. These were the models that were used in the parent paper as well. By further understanding the parameters that can be tuned within each model we can get a more accurate result.

Findings:

In the model, the Convolutional Neural Network model was the best but there were also experiments with SVM, Lasso, and Elastic Net. The NN had a precision of 0.80, SVM with 0.67, Lasso with 0.78, and ElasticNet with 0.79.

Conclusions and Recommendations:

It can be drawn that it is fully possible to predict whether or not a stroke will occur in a patient with relatively high accuracy. This could be related to business by a company building a highly accurate ML model and selling it to insurance companies or hospitals. A government agency could use this model to push new initiatives as to which risk factors are more prevalent for strokes. A Non-government organization could use this model to help people in need who do not have access to information like this to live a better lifestyle.

Literature Review:

A theoretical concept used in the development of the paper is the dependency on risk factors in predicting the possibility of a stroke to see which risk factors were most important. Another concept used were statistical techniques to evaluate the models. One of the authors of this paper is Soumyabrata Dev, an assistant professor at University College Dublin. One of his previous works was how numerous weather parameters affect the occurrence and amount of rainfall. He sheds light on this issue because rainfall prediction is not a common topic.

Pros and Cons:

If I could have done this research differently, I would have focused more on cleaning the data. I think the PCA component of the research was interesting but feature engineering could also be used. Also, I would focus more on tuning hyperparameters because that was not covered too much. Specifically using Grid Search CV. The pros of this research was that several machine learning models were used because it showed that some models performed better than others.