What is the most accurate model in predicting the stroke risk of a person given 10 attributes.

Orfeas Gkourlias - 420172

2022-11-06

Preface

The project and paper were written for an assignment given by Hanzehogeschool Groningen. With the educational goal of providing a learning experience regarding machine learning and some Java. We were given the opportunity to choose any data set to our liking for the project, adhering to the requirements set, of course.

This made it so that the entire experience was more interesting, because of the agency in choosing something that interested me. Sadly, the first data set I had chosen did not meet the requirements, resulting in me lagging behind for a little while. Luckily, I was able to catch up with another interesting subject, concerning stroke patients. This subject intrigues me tremendously, because of people in my environment who have had the misfortune of suffering a stroke. Considering the ever growing importance and emphasizes on preventative care, and the significance of strokes on people, I thought this would be a perfect candidate for the machine learning project. I would like to thank my teachers for the guidance they offered in execution of the project.

Abstract

Every 40 seconds someone in the united states suffers from a stroke. With around 1 out of 5 people passing away as a result. The outlook for stroke patients are also varying in severity for the survivors. With extensive aftercare almost being a guarantee in all cases, strokes are also the leading cause of serious, long-term disabilities and impairments. With only 1 out of 10 patients making a full recovery.

The severity and prevalence of strokes cannot be understated. While modern aftercare usually provides favorable prognosis, the complete prevention of not only strokes, but many cardiovascular conditions has long been advocated for. The importance of preventative healthcare keeps being re-emphasized as people grow older. There are many preventative measures available and still being developed.

Considering the above, the ability to predict whether someone may suffer from a stroke would be an invaluable tool. This project has aimed to do just that. By taking patient data, examining that and then training a machine learning model to make predictions based off it. Using the powerful machine-learning and data analysis program called Weka, a prediction model was developed and trained to be 92 percent accurate in its classification. With emphasis on avoiding false negatives, the model will be identify high risk patients more often than not.

If this model could be made even more accurate, then it might serve as a primary detector of high risk stroke patients.

Abbreviations

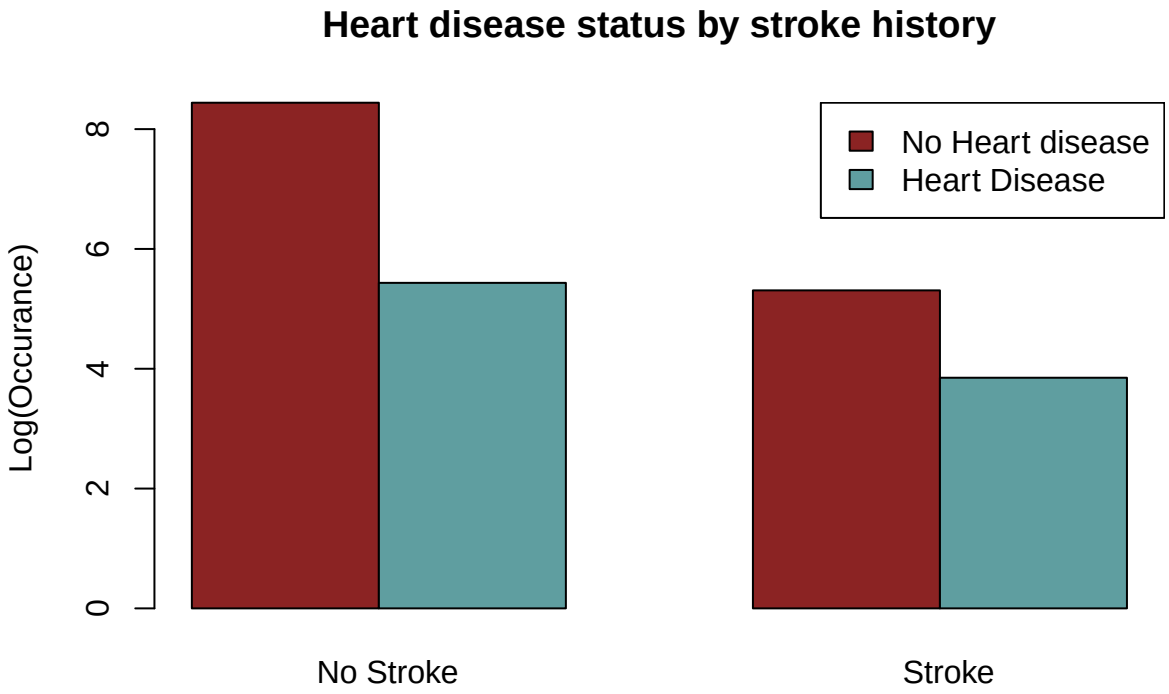
Table of Contents

Introduction

Materials and methods

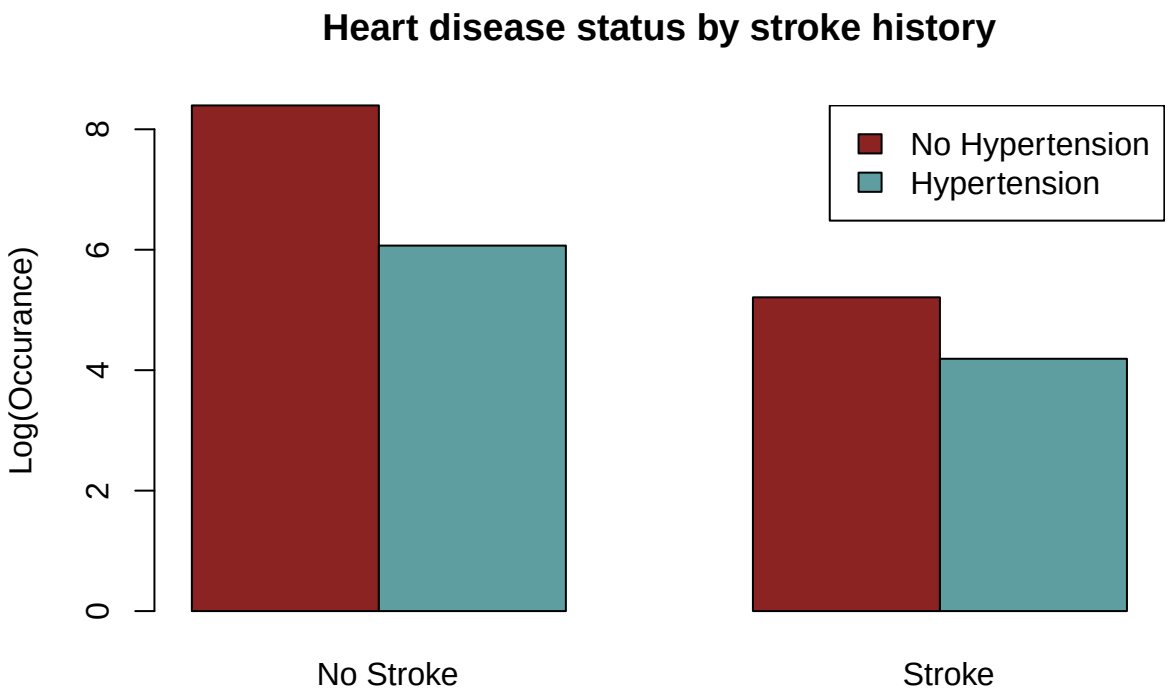
Results

Following the analysis of the data set and the attributes within, some important correlations can be observed. Correlations relevant to the research question mostly consist of heart and health attributes. The first plot regarding heart d isease differences in non-stroke and stroke patients will be shown.



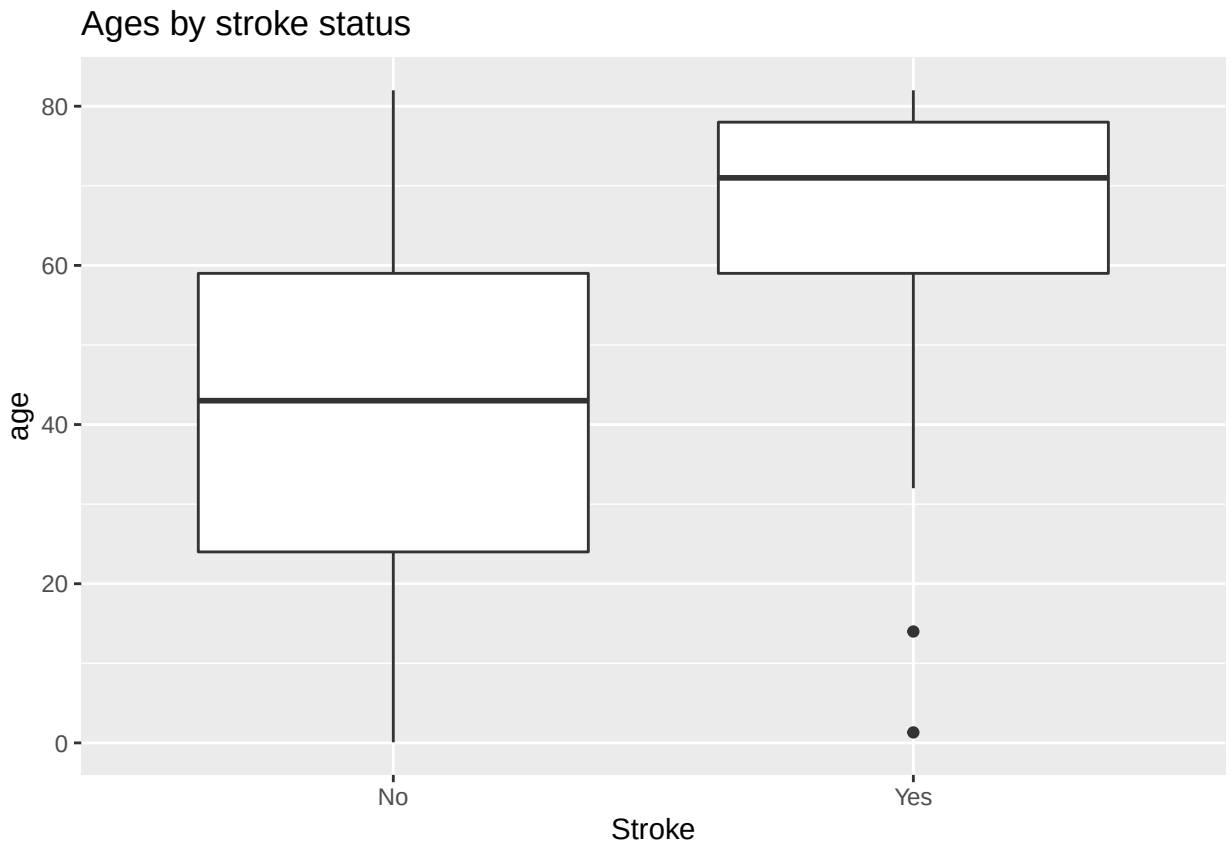
This plot of normalized data shows that there are proportionally more heart disease patients in the stroke group, than there are in the no stroke group. This ratio can be translated to a percentage. The percentage of patients who have not had a stroke is 4.71%. This percen tage is higher for the group of patients who have experienced a stroke. For that group the percentage is 18.88%. This means that it is 4 times as likely for someone who has had a stroke, to also have a heart condition. The literature on this subject seems to have come to a common consensus: There is a correlation between having a heart disease and a higher risk of a stroke, This paper published by the American Heart Association being one of the bigger published works on the topic: Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association (Keep in mind, the risk of developing heart disease AFTER a stroke is also quite big.)

The same can be seen in the case of hypertension, because that is also a heart condition.



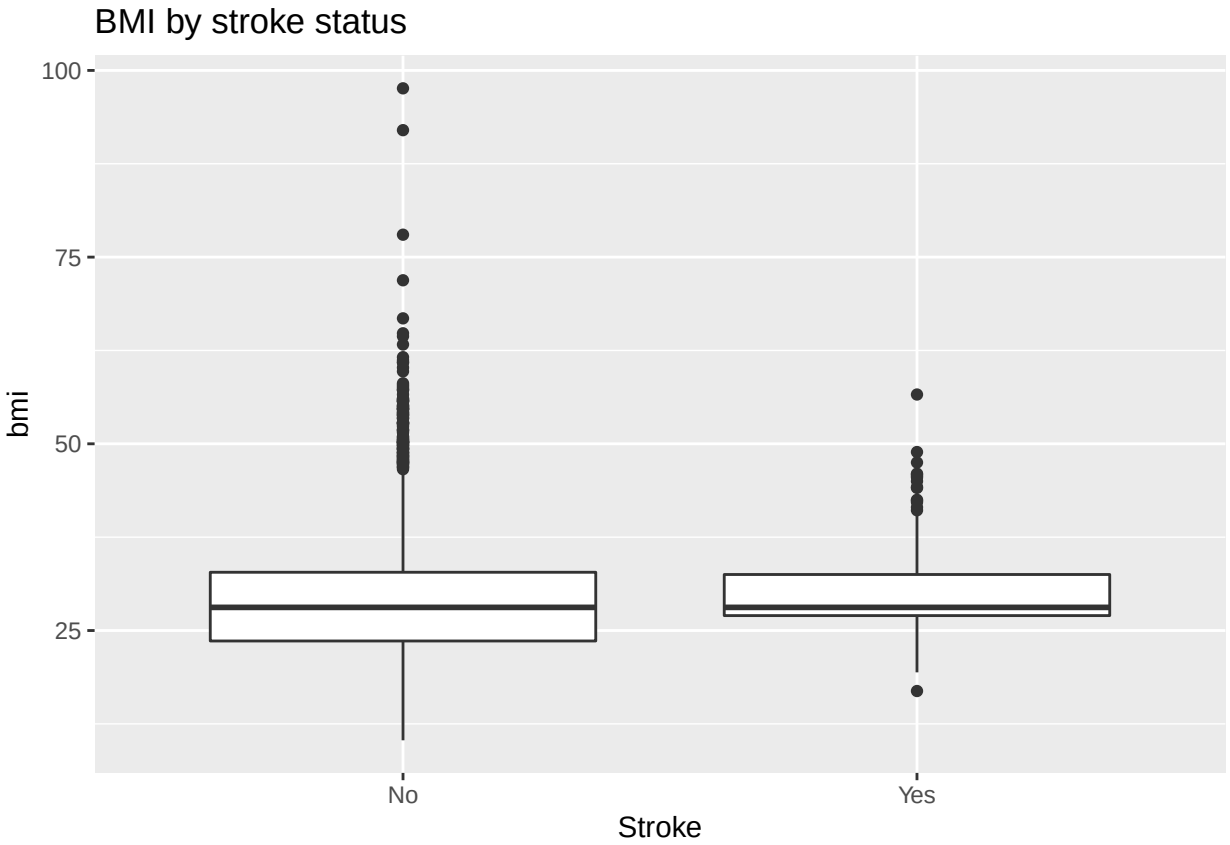
This plot is very similar to the prior one. That makes sense, because quite some heart conditions are accompanied by hypertension. Hypertension on its own is already classified as a heart condition too.

Age has been observed to also be an important attribute. The older a person beco mes, the higher the likelihood of a stroke. Especially when combined with other attributes, such as hypertension or heart conditions. According to this paper, the odds of having a stroke double for every ten years a person lives. Aging and ischemic stroke



This shows that the older a person is, the more likely they are to have experienced a stroke. This would of course make sense, considering that the older someone gets, the higher the chance one might encounter health issues.

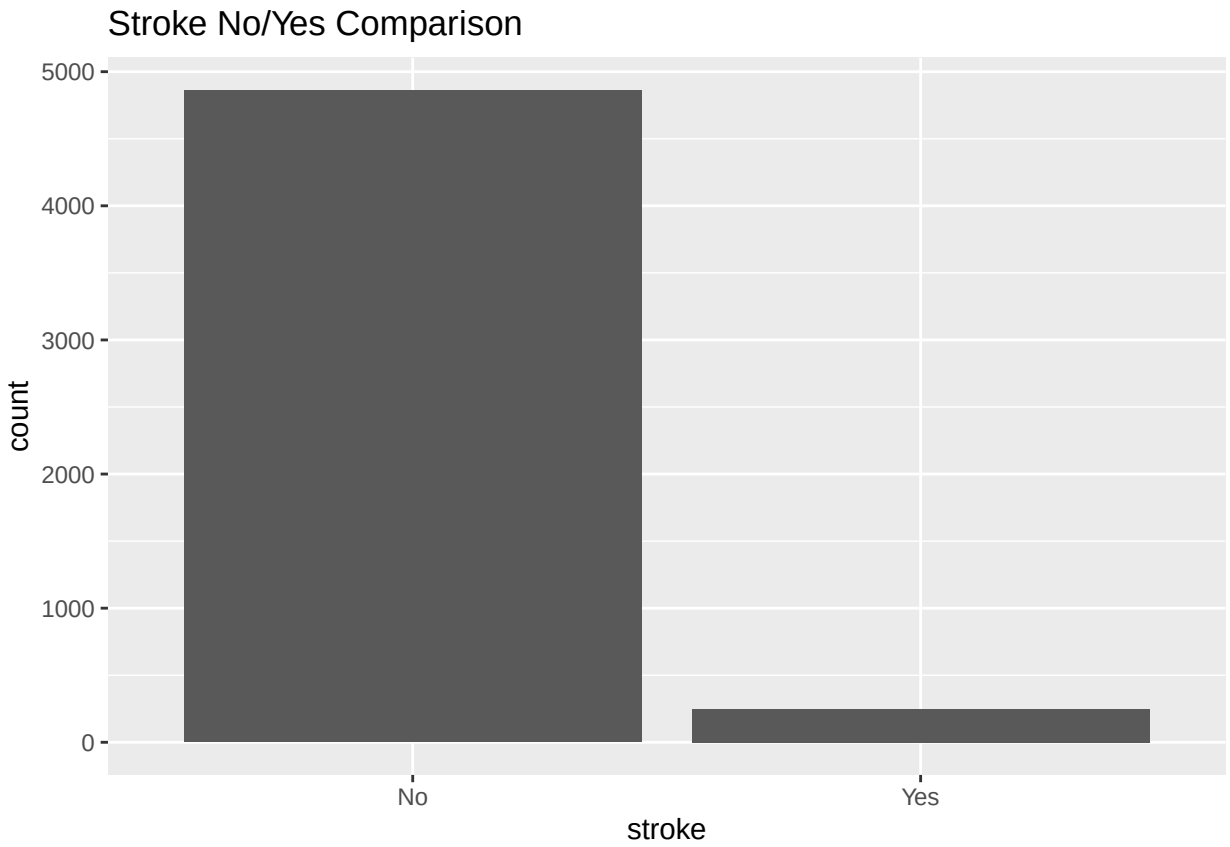
Another health related observation that may be relevant is that of the BMI means for stroke and no stroke patients. This can be visualized through a similar box plot as the prior one.



The patients who have had a stroke only have a slightly higher BMI. The no stroke group also has more extreme values on both ends.

Imbalance

Something important which can be observed in the data is the imbalance between stroke instances. A barplot will be used to demonstrate this imbalance



As can be seen, there is a great imbalance in the amount of patients who have had a stroke those who have not. To be exact: 0% of the data, consists of rows where stroke is equal to 0. The remaining 5% has experienced a stroke. This will be very important when selecting for different machine learn ing algorithms, because the weighting of occurrences will most likely have to be changed in this case.

Discussion and Conclusion

Now that the most important results have been shown, some points of contention could be considered. The aim of this project is to answer the original research question using machine learning. This may be done with varying algorithms and considering different attributes. Looking at the results, some of the attribute s may seem irrelevant at first, and some would argue these could be removed before applying any machine learning algorithms. But in this case, none of the attributes will be discarded. The program being used, Weka, has multiple functions which allow for the automation of this process. By allowing Weka to select the most influential attributes systematically, biases may be avoided.

In addition to the data selection, the imbalance is also important to consider. One way to tackle this problem is by generating samples in the minority class. That class being the 1 instances under the stroke attribute. Another option was to use SMOTE (Synthetic Minority Oversampling Technique). This method was chosen because it is effectively a more accurate version of oversampling. SMOTE will generate synthetic data points, so that the absence of balance will be remedied. Whether this should be remedied might also be debatable. Some may argue that the raw data ratio would be an accurate representation. But considering that the difference is this extreme, SMOTE was chosen regardless.

References

Appendices