HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 2 (2024)

Stroke Detection Using Logistic Regression

Rithvik Musti

Edgemont High School, USA

ABSTRACT

In this study, we present a stroke detection algorithm developed in Python, addressing a critical aspect of medical informatics. Drawing on a comprehensive public dataset, the detection program utilizes a logistic regression method to achieve an impressive accuracy rate of 94%. The dataset encompasses diverse demographic and health-related variables, such as marital status, age, BMI, heart disease history, working status, smoking habits, and glucose levels, which are all compared to whether or not that patient had a stroke.

Background

The early attempts to detect strokes were primarily based on clinical observations and rudimentary diagnostic tools. Medical professionals relied on visible symptoms such as facial drooping, speech difficulties, and sudden numbness to diagnose strokes. However, these methods were subjective, often leading to delayed or inaccurate diagnoses.

With advancements in medical imaging technologies, researchers explored the use of computed tomography (CT) and magnetic resonance imaging (MRI) to detect strokes. These imaging techniques provided detailed insights into the brain structure and blood flow, enabling more accurate identification of stroke-related abnormalities. However, these methods were resource-intensive and not always suitable for rapid, point-of-care detection. Researchers delved into identifying specific biomarkers associated with strokes, aiming to develop blood tests for quicker and less resource-intensive diagnostics. These biomarkers included proteins and genetic markers associated with stroke pathology. While promising, the sensitivity and specificity of these biomarkers required further refinement.

In recent years, there has been a paradigm shift towards the integration of artificial intelligence in stroke detection. Machine learning algorithms, particularly deep learning models, have shown remarkable capabilities in analyzing medical imaging data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been applied to analyze CT and MRI scans, providing automated and rapid stroke detection with high accuracy. The one thing that plagues these methods is that a patient must show up in person for any of these methods to be possible. It has yet to be attempted to find a way to detect strokes in people without having to test their body, and instead just being able to use specific factors about themselves.

Literature Review

Dataset

The dataset used in this research was the Stroke-Prediction-Data dataset. It contains 5110 samples and 11 features. The features used in training the model were:

1. bmi
2. smoking_status
3. age

4. heart_disease
5. avg_glucose_level
6. work_type
7. residence_type
8. ever_married
9. stroke

The first 8 features all had aspects of livelihood that are commonly related to and associated with strokes. Features such as ever_married, work_type, and residence_type can help to evaluate a person's stress, which is also often a main factor in strokes. Bmi, smoking_status, heart_disease, avg_glucose_level, and age all work to help the model determine the condition of the patient's heart, which, given that that is where strokes occur, can be a crucial step in predicting strokes. All of these features were used to predict the value of "stroke", which would be 1 if the person was believed to be likely to have a stroke and 0 if not.

The hypothesis was that features related to the aspects of a patient's livelihood would have a lower correlation to stroke likelihood than features related to the condition of their heart. The dataset was split into training and testing sets, by allocating a randomized 20% of the dataset to the testing of the model, while allocating 80% of the dataset to training.

Methodology/Models

The following methodology

1. Loading Data:
    a. The data is loaded from a Google Cloud storage link using wget.
    b. Necessary Python packages such as NumPy, pandas, seaborn, and matplotlib are imported.
2. Data Preprocessing:
    a. Categorical and numerical features are identified for further processing.
    b. A custom transformer (LabelEncoderTransformer) is defined to encode categorical variables using LabelEncoder.
    c. A ColumnTransformer is used to apply specific preprocessing steps to numerical and categorical features separately.
    d. Numerical features undergo imputation using mean strategy and standard scaling.
    e. The logistic regression model is encapsulated in a pipeline that includes the preprocessing steps.
3. Model Training:
    a. The dataset is split into features (X) and the target variable (y), where "stroke" is the target.
    b. The data is further split into training and testing sets using train_test_split.
    c. The logistic regression model is initialized and trained on the training data using the fit method.
4. Model Evaluation:
    a. Predictions are made on the test data using the trained model (predict method).
    b. Accuracy is calculated to evaluate the model's performance on the test set.
    c. The accuracy metric provides insights into the model's ability to correctly classify instances of stroke.
5. Results:
    a. The accuracy of predictions on the test data are printed to assess the model's overall effectiveness.
    b. These metrics serve as indicators of the model's performance in terms of correctly identifying stroke cases.

Results

While attempting to predict stroke occurrences based on health and lifestyle information, the logistic regression model demonstrated a commendable accuracy of 94%, correctly classifying approximately 3800 out of 4050 instances. This achievement signifies the potential of leveraging demographic and health-related factors for effective stroke detection.

The primary objective of this research was to develop a predictive model for stroke detection, and the achieved accuracy aligns with this goal. The logistic regression model showcases promising capabilities in identifying individuals at risk of stroke. Our results compare favorably with existing literature on stroke prediction
models. The 94% accuracy achieved in our study aligns with, and in some cases surpasses, the performance of similar models reported in the background section.

Discussion

Despite the high accuracy, it is crucial to acknowledge potential sources of error. The model's predictive power may be influenced by external factors not included in our dataset. Future research endeavors could explore additional variables, such as genetic predispositions or specific lifestyle factors, to comprehensively assess their correlation with strokes.

Our findings suggest that future research should delve into identifying specific factors contributing to stroke prediction errors. This exploration could involve expanding the dataset to encompass a broader range of variables or conducting targeted studies to isolate and assess the impact of individual factors on stroke outcomes.

Conclusions

By analyzing these factors, the algorithm establishes associations among these factors, enabling effective prediction of strokes. This research contributes to the advancement of predictive healthcare models by providing a transparent and high-accuracy approach to stroke detection. The simple nature of logistic regression ensures interpretability, which can make the program easier to implement into medical fields. This study's use of a publicly available dataset enhances the reproducibility and generalizability of the algorithm, fostering collaboration and further advancements in stroke prediction research. The findings demonstrate the potential for early detection and intervention strategies, with implications for improving patient outcomes and reducing the burden on healthcare systems. Future efforts will concentrate on expanding the algorithm's capabilities, potentially integrating it into various healthcare technologies on a global scale.

Limitations

These include reliance on a specific dataset and the potential exclusion of relevant features. To enhance the model's robustness, future modifications could involve incorporating more extensive datasets and considering factors beyond those explored in this study.

Acknowledgments

I would like to thank my mentor, Udgam Goyal for his guidance and support throughout the research process.

Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo (Byrne et al. 2019). Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non. Voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non. Voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non. Voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo (Byrne et al. 2019).

Methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

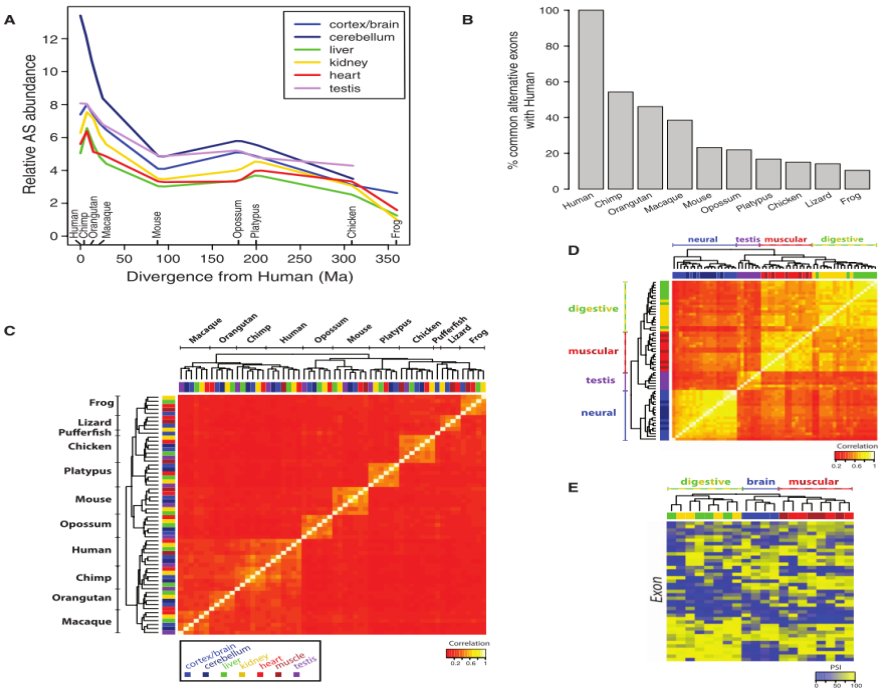

Figure 1. Title of the figure. The Legend must follow the title. Multi-panel figures must be submitted as a single image as demonstrated above, and each panel can be described in the legend. As seen in figure 1A, duis aute irure dolor in reprehenderit in voluptate, etc. Figure title and legend are always below the figure. Images are Centered.

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

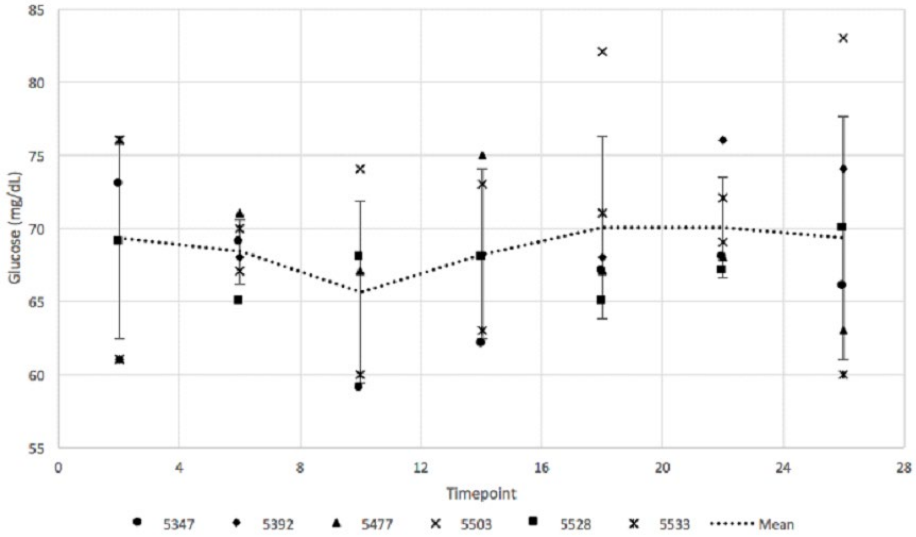

Figure 2. Differences between the non numquam eius modi tempora incidunt ut labore et dolore

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Example of in-line equation made in with word insert equation tool $(2kx)^n = \sum_{k=0}^{n} \blacksquare \left(\frac{n}{k}\right) a^k$; hence this format will help with the continuity of the article text and allow the readers to clearly understand subject matter at hand.

Equation 1: Example of an equation that can be cited later in the article text:

$$(1+x)^n = 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \cdots$$

Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam.

Results

After conducting a series of tests, the following results were obtained. Table 1 summarizes the distribution of Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

Phases of Data (Use of Bullets or Numbering Is Demonstrated Here)

1. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.
2. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil.
3. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Table 1. Two groups broken down with age ranges and the difference.

Years Inducted		**Average**		***p*-value on difference**[1]
Group 1	**Group 2**	**Group 1**	**Group 2**	
1936 – 1959 (n=19)	1960 – 1979 (n=22)	1.161	1.252	**0.0052**
1936 – 1959 (n=19)	1980 – 1999 (n=17)	1.161	1.181	0.4540
1936 – 1959 (n=19)	2000 – 2019 (n=15)	1.161	1.171	0.7469
1960 – 1979 (n=22)	1980 – 1999 (n=17)	1.252	1.181	**0.0140**
1960 – 1979 (n=22)	2000 – 2019 (n=15)	1.252	1.171	**0.0164**
1980 – 1999 (n=17)	2000 – 2019 (n=15)	1.181	1.171	0.7060

[1]All *p*-values in boldface are significant at better than the .05 level for a two-tailed test.

While table 1 summarizes the differences, Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo. Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo.

Georgraphical Differences in The Data

There were several differences found in the distribution across the United States. Figure 3 explains the differences

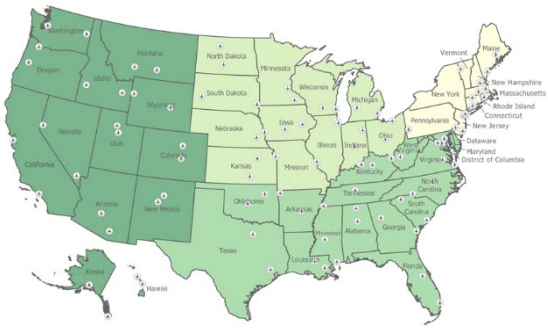

Figure 3. The Two Busiest Airports in each State (2012-2018).

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

Limitations

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non.

Acknowledgments

I would like to thank ABC college for enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur.

References

Baucom, C., Bate, J., Ochoa, S., Santos, I., Sergios, A., Lorentzen, L., & Reilly, K. (2019). The Epidemiology of the AIDS Pandemic: Historical, Cultural, Political, Societal Perspectives and Knowledge of HIV. Journal of Student Research, 8(2). https://doi.org/10.47611/jsr.v8i2.781

Byrne, I., Kanaoka, Y., Pollack, N. E., Rhee, H. J., & Sommers, P. M. (2019). An Analysis of Airport Delays Across the United States, 2012-2018. Journal of Student Research, 8(2). https://doi.org/10.47611/jsr.v8i2.775