Brain Stroke Prediction

Group 4 Team Members:

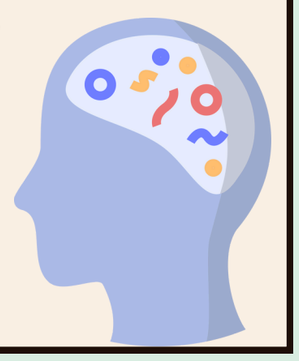
- 2602149243 Steve Aditya Gandha
- 2602180222 Sherly Patricia
- 2602117033 Dionisius Avelino
- 2602148562 Manisha Arie Paramartha
- 2602103980 Yonathan Immanuel Nurhan

ABOUT THE DATASET

The dataset used in this research is about brain stroke prediction and it is taken from Kaggle.

Our dataset contains 4981 different observations with 11 different variables for each observation such as Gender, Age, Hypertension, Heart Disease, Marital Status, Work Type, Residence Type, Average Glucose Level, BMI, Smoking Status, and Stroke which is the target variable.

This dataset will be split into training data and testing data.



INTRODUCTION

- An early brain stroke prediction is a form of prevention against strokes that is required at this time as it is the second leading cause of death in the world, with a death rate of around 5.5 million people per year.
- This research aims to create a random forest model that can predict a person's risk of having a brain stroke. By giving information related to the required variables, people can be more aware of its warning signs and symptoms, which are very helpful in preventing brain stroke.

END-TO-END MACHINE LEARNING STEPS

I. Data Selection

The dataset used is about brain stroke that contains 4981 observations and 11 variables. The dataset will then be split into 80% for training and 20% for testing, with 0 being chosen as the random state.

II. Feature Engineering

- Binning

  Data pre-processing technique to convert a continuous value into an interval or bins. Binning will be done on the age variable where they will be grouped into a new variable called "age band"
- Normalization

  Data pre-processing technique to transform variables into specific scale or range. Normalization are used for two variables which are "BMI" and "average_glucose_level"
- Encoding

  Data pre-processing technique to transform categorical variables into numerical variables. Two variables that are transformed are 'gender' with one-hot encoding and 'smoking_status' with label encoding
- Random Sampling

  A form of sampling where data points are selected randomly from a dataset without any predetermined pattern. Non-stroke observations will be randomly selected because of the unbalanced amount of stroke and non-stroke observations

III. Model Selection

By creating a good model, a good prediction related to brain stroke can be made. The model that used to provide a predictive analysis in this research is Random Forest. It can handle both classification and regression problems. The random forest model will be used to handle classification (handling categorical variables).

RESULTS & CONCLUSION

The brain plays an integral part in the human body, and it is essential for the body control and proper function. This research uses Random Forest machine learning to predict brain strokes. Through data preprocessing, training, splitting and testing, it achieves an **accuracy of 80%** which is considered good enough for a single model to perform as brain stroke predictor. This means that the model can be used to predict upcoming brain stroke, so that they can be more aware and do some early prevention.

IV. Evaluation Metrics

Evaluation Metrics are measures that are used to assess the performance of a model. It can help to know how well and effectively a model is used to make predictions. For this dataset which involves classification, the evaluation metrics used are accuracy, precision, recall, and F1 score.

REFERENCES

[1] E. S. Donkor, “Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life,” Stroke Research and Treatment, vol. 2018, pp. 1–10, Nov. 2018.

[2] Centers for Disease Control and Prevention, “About Stroke,” https://www.cdc.gov/stroke/about/?CDC_AAref_Val=https://www.cdc.gov/stroke/about.htm. [Accessed: May 19, 2024].

[3] National Institute of Neurological Disorders and Stroke, “Brain Basics: Preventing Stroke,” https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-preventing-stroke#:~:text=It%20is%20the%20most%20common,that%20burden%20through%20biomedical%20research. [Accessed: May 19, 2024].

[4] Javatpoint, "Machine Learning - Random Forest Algorithm," https://www.javatpoint.com/machine-learning-random-forest-algorithm. [Accessed: May 19, 2024].

[5] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of Brain Stroke Severity Using Machine Learning," Revue d’Intelligence Artificielle, vol. 34, no. 6, pp. 753-761, Dec. 2020

[6] M. S. Sirsat, E. Fermˊe, and J. Cˆamara, “Machine Learning for Brain Stroke: A Review,” Journal of Stroke and Cerebrovascular Diseases, vol. 29, no. 10, p. 105162, Oct. 2020.

[7] V. Krishna, J. S. Kiran, P. P. Rao, G. C. Babu, and G. J. Babu, “Early Detection of Brain Stroke Using Machine Learning Techniques.” Trichy, India: IEEE, 2021, pp. 1489–1495.

[8] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, “Performance analysis of machine learning approaches in stroke prediction,” in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, Nov. 2020.

[9] K. Akash, H. Shashank, S. .S, and T. A.M, “Prediction of Stroke Using Machine Learning,” 06 2020.

[10] E. Dritsas and M. Trigka, “Stroke Risk Prediction with Machine Learning Techniques,” Sensors, vol. 22, no. 13, p. 4670, Jun. 2022.

[11] [V. Mariano, J. A. T. Vasquez, M. R. Casu, and F. Vipiana, “Efficient Data Generation for Stroke Classification Via Multilayer Perceptron.” Denver, CO, USA: IEEE, 2022, pp. 890–891.

[12] S. Rahman, M. Hasan, and A. K. Sarkar, “Prediction of Brain Stroke Using Machine Learning Algorithms and Deep Neural Network Techniques,” European Journal of Electrical Engineering and Computer Science, vol. 7, no. 1, pp. 23–30, Jan. 2023.

[13] GeeksForGeeks, "What is Feature Engineering" https://www.geeksforgeeks.org/what-is-feature-engineering. [Accessed: June 10, 2024].

[14] M, Kabungo, "Data Bining Explained" https://medium.com/@mose.kabungo/binning-explained-557aa3cce591. [Accessed: June 10, 2024].

[15] Deepchecks, "Normalization in Machine Learning" https://deepchecks.com/glossary/normalization-in-machine-learning/#:~:text=Normalization%20in%20machine%20learning%20is%20the%20process%20of%20translating%20data,when%20Euclidean%20distance%20is%20used [Accessed: June 10, 2024].

[16] GeeksForGeeks, "Feature Encoding Techniques Machine Learning" https://www.geeksforgeeks.org/feature-encoding-techniques-machine-learning/ [Accessed: June 10, 2024].

[17] R. Gangwal, "Types of Sampling and Sampling Technique" https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/. [Accessed: June 10, 2024].

[18] IBM, "What is Random Forest" https://www.ibm.com/topics/random-forest [Accessed: June 10, 2024].

[19] S. Kumar Agwal, "Metrics to Evaluate Your Classification Model to Take the Right Decisions" https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/[Accessed: June 10, 2024].