International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 7 Issue 5 ǁ 2019 ǁ PP. 42-46

A Data-Driven Approach to Predicting Stroke Risk from Lifestyle Factors

Praveen Kumar Misra,
Dr.Shakuntala Misra National Rehabilitation University, Lucknow

Abstract
Stroke is one of the leading causes of death and long-term disability worldwide, affecting millions of individuals every year. According to the World Health Organization, nearly 15 million people suffer a stroke annually, resulting in approximately 5.5 million deaths and 5.5 million cases of permanent disability. These alarming statistics highlight the urgent need for early detection and effective prevention strategies. Lifestyle factors such as physical activity, diet, smoking habits, alcohol consumption, and medical conditions play a crucial role in determining an individual's risk of stroke. This study presents a data-driven approach to predicting stroke risk using lifestyle-related data and machine learning techniques. Various lifestyle attributes are analyzed to identify patterns and relationships associated with stroke occurrence. Predictive models are trained and evaluated to classify individuals based on their potential stroke risk. The results demonstrate that machine learning can effectively analyze lifestyle data to assist in early risk assessment. Such predictive systems can support healthcare professionals in identifying high-risk individuals and promoting timely lifestyle interventions, ultimately contributing to improved public health outcomes and reduced stroke incidence.
Keywords: *Stroke Prediction, Machine Learning, Lifestyle Factors, Healthcare Analytics, Risk Assessment, Data-Driven Healthcare.*

I. Introduction

Stroke is one of the leading causes of death and disability worldwide. The World Health Organisation (WHO) estimates that stroke causes about 15 million strokes annually worldwide. Of them, 5.5 million result in death and an additional 5.5 million cause permanent disabilities. According to the World Health Organisation, one in four individuals over 25 will experience a stroke at some point in their lives.

These figures demonstrate the enormous toll that stroke takes on people, the medical system, and society at large. In order to avoid strokes and lessen their effects, early detection and risk assessment might be extremely important. The risk of stroke is significantly influenced by lifestyle factors. This study explores how human lifestyle data can be used to train machine learning algorithms to predict the risk of stroke.

II. Objectives

The main goal of this research is to create reliable prediction models that can evaluate the risk of strokes by analyzing a wide range of lifestyle factors related to human life and behavior like age, gender, average glucose level, smoking habits, etc. With a wide range of parameters, including lifestyle decisions, medical history, and demographic data, the research aims to discover important stroke risk factors and develop predictive models.

III. Methods

A. Data Collection and Preprocessing:

- A dataset is gathered from the Data Science website Kaggle that includes labels for stroke occurrences and information on human lifestyle.Data cleaning procedures are utilised to fix inconsistencies and missing values.Numerical representations that are appropriate for machine learning algorithms are created by encoding categorical information.Numerical features are normalised via feature scaling, which guarantees that each feature contributes equally to the model.

B. Exploratory Data Analysis (EDA):

- Histograms and scatter plots are examples of visualisations used to investigate the relationship between the goal variable (the occurrence of strokes) and the distribution of features.
- Potential correlations between lifestyle factors and the risk of stroke are identified with the use of this analysis.

C. Model Building and Training:

- Feature Selection: To determine the most relevant lifestyle factors for predicting the risk of stroke, feature selection approaches such as correlation analysis or chi-square testing can be utilised. This can shorten the training period and enhance model performance.
- For predicting the risk of stroke, machine learning algorithms like **K-Nearest Neighbours (KNN)** and **Logistic Regression** are used.
- A useful method for examining the connection between several characteristics and a binary result (stroke or no stroke) is logistic regression.
- KNN is a non-parametric technique that uses the labels of data points' closest neighbours in the feature space to classify the data points.
- GridSearchCV is used for hyperparameter tuning, which finds the best setup for each model to maximise accuracy on the training set.
- In this step, several combinations of hyperparameter values are tried, and their performance on the training set of data is assessed.
- The preprocessed data is used to train the selected models.

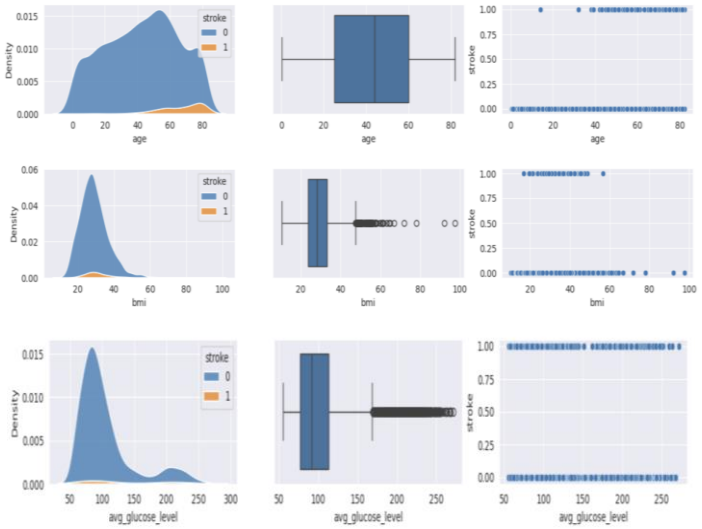

Fig. 1 Factors like age, bmi and glucose level are considered

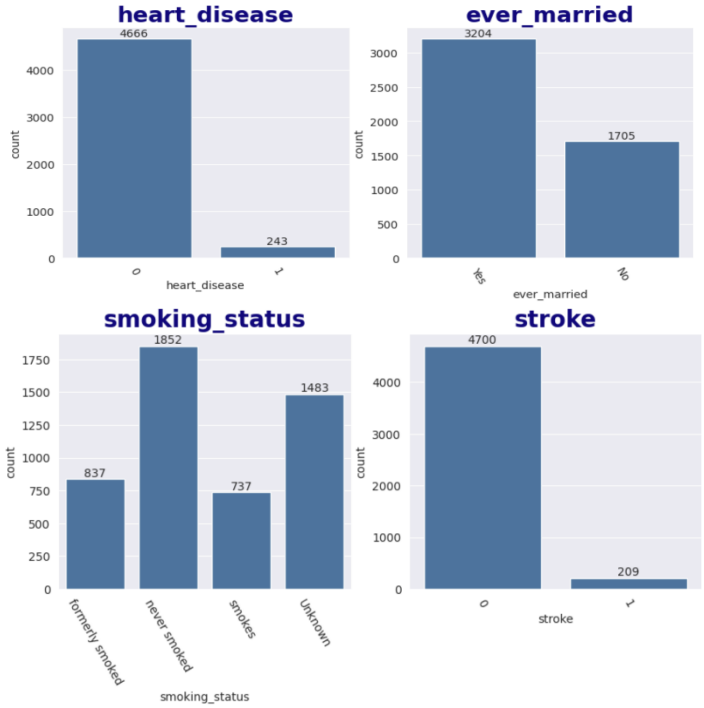

Fig 2. Some other factors selected for the prediction of strokes using feature selection

D. Model Evaluation:

- Metrics like accuracy, precision, recall, and F1-score are used to assess the performance of the trained models on an independent testing dataset.
- The total percentage of precise predictions is measured by accuracy.
- The percentage of positive forecasts that are actually true positives is measured by precision.
- The percentage of real positive cases that the model accurately detects is measured by recall.
- To see the distribution of accurate and inaccurate predictions for every model, confusion matrices are created.
- The number of correctly and wrongly predicted instances of each class (stroke/no stroke) is displayed in confusion matrices.

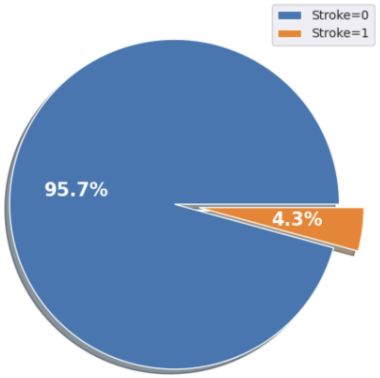

Fig. 3 A Pie-Chart showing the stroke percentage

IV. Results

- Significant correlations between a few lifestyle factors and the risk of stroke may be found by the investigation. For instance, it may be discovered that smoking, high blood pressure, and inactivity are all significant risk factors for stroke.
- Based on lifestyle data, the trained machine learning models may be able to predict stroke risk with a reasonable degree of accuracy.

- The number of correctly and wrongly predicted instances of each class (stroke/no stroke) is displayed in confusion matrices.
- One model may outperform the other in this study's stroke risk prediction, according to the comparison of the models. While KNN may be easier to construct but possibly less interpretable, logistic regression may be a better fit for analysing the association between characteristics and stroke risk.
- One model may outperform the other in this study's stroke risk prediction, according to the comparison of the models. While KNN may be easier to construct but possibly less interpretable, logistic regression may be a better fit for analysing the association between characteristics and stroke risk.

Future Work:

- Other lifestyle factors may be investigated and added to the model in subsequent studies.
- For maybe better results, more intricate machine learning models like Gradient Boosting Machines or Random Forests could be researched.

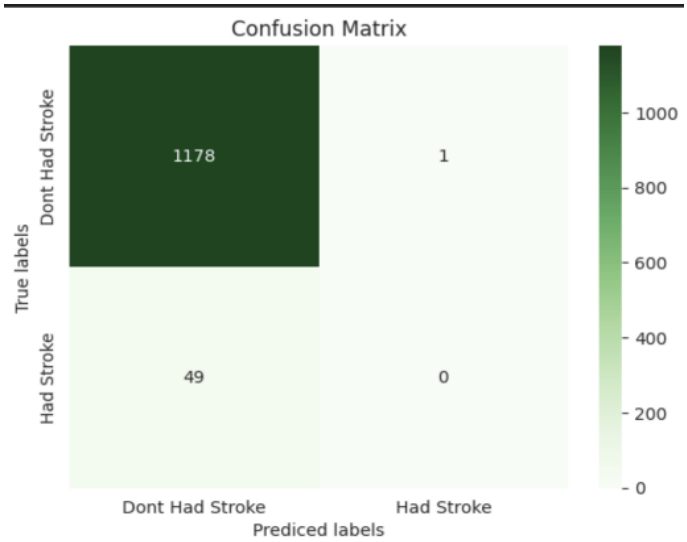

Fig. 4 A Confusion Matrix representing true and predicted values using KNN Classifier

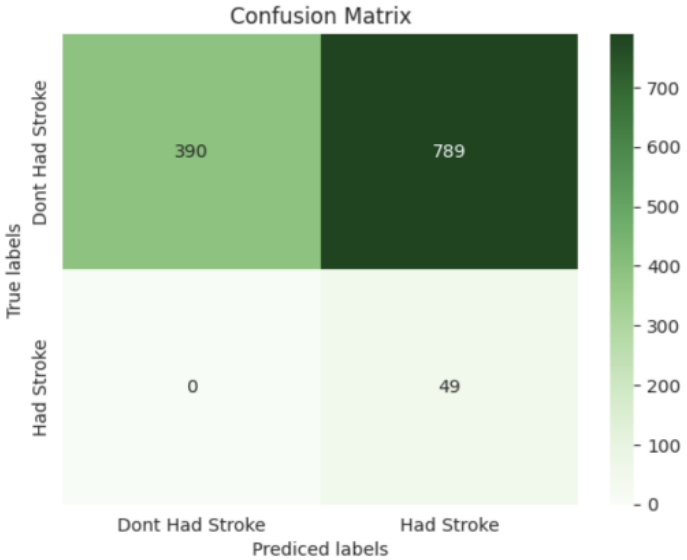

Fig. 5 A Confusion Matrix representing true and predicted values using Logistic Regression

V. CONCLUSION

This study not only shows how lifestyle factors might predict the risk of stroke, but it also shows how machine learning algorithms can be used to improve stroke prevention and early detection methods. Through the examination of several lifestyle factors like, age, gender, smoking status, and medical background, machine learning models are able to accurately determine those who are more likely to suffer a stroke. Aiming to

encourage healthy living and lower the occurrence of strokes, this research's conclusions can guide the development and execution of treatment programmes and customised risk assessment tools. To increase their effectiveness, these programmes can be tailored to target particular risk profiles or demographic groups.

Furthermore, the findings of this research can operate as a basis for the creation of new and innovative digital health solutions and mobile applications that enable people to track and control their stroke risk in real time. These tools can reduce the incidence of stroke and enhance overall health outcomes by using machine learning algorithms to deliver tailored advice, lifestyle changes, and preventive measures.

By making use of the power of data-driven insights, policymakers, healthcare professionals, and individuals alike can collaborate to create a healthier society and reduce the burden of stroke-related disabilities and deaths.

REFERENCES

[1]. Feigin, V. L., Norrving, B., & Mensah, G. A. (2017). Global burden of stroke. Circulation Research, 120(3), 439–448.

[2]. Wolf, P. A., D'Agostino, R. B., Belanger, A. J., & Kannel, W. B. (1991). Probability of stroke: A risk profile from the Framingham Study. Stroke, 22(3), 312–318.

[3]. O'Donnell, M. J., Chin, S. L., Rangarajan, S., et al. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke. The Lancet, 388(10046), 761–775.

[4]. Kannel, W. B., & Wolf, P. A. (1998). Epidemiology of stroke. In Handbook of Clinical Neurology. Elsevier.

[5]. Benjamin, E. J., Virani, S. S., Callaway, C. W., et al. (2018). Heart disease and stroke statistics—2018 update. Circulation, 137(12), e67–e492.

[6]. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[7]. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.

[8]. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.

[9]. Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[10]. Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern Classification (2nd ed.). Wiley.