Evaluation Study of the Chi-Square Method for Feature Selection in Stroke Prediction with Random Forest Regression

Nurliana Nasution[1*], Feldiansyah Bakri Nasution[2], Erlin[3], Mhd Arief Hasan[4]

{nurliananst@unilak.ac.id[1*], feldiansyah@unilak.ac.id[2], erlin@lecturer.pelitaindonesia.ac.id, m.arif@unilak.ac.id}

[1,2,4]Informatic Engineering Study Program, Universitas Lancang Kuning, Pekanbaru, Riau
[3]Institut Bisnis dan Teknologi Pelita Indonesia

Abstract. This study aims to develop a more accurate classification model for diagnosing Stroke cases based on various clinical features. Stroke is a serious global health issue, and early detection has a positive impact on prognosis and the prevention of complications. In this research, we combine two main approaches, feature selection using the Chi-Square statistical test and the implementation of Random Forest Regression, to enhance the accuracy of Stroke diagnosis.First, we use the Chi-Square test to evaluate the relationship between categorical variables (such as gender, smoking history, marital status, and others) and Stroke status. The results of this test are used to select variables that have a significant association with Stroke. We then map categorical variables into numerical values so that the model can process them, overcoming errors that previously arose due to categorical data.Subsequently, we apply Random Forest Regression to the pre-processed dataset. The resulting model achieves an accuracy of 95%, indicating a significant improvement compared to the previous model. In addition to accuracy, we also observe improvements in precision, recall, and F1-Score, which indicate the model's ability to identify Stroke cases and avoid misdiagnoses.The findings of this research hold significant potential in clinical practice, particularly in the early detection and management of Stroke. Early Stroke detection can lead to faster intervention, ultimately reducing the negative impact of Stroke on patients. We hope that this study will serve as a foundation for the development of more advanced and accurate models for Stroke diagnosis, contributing to overall public healthcare improvement..

Keywords: Stroke, Chi-Square, Random Forest Regression, Classification, Early Detection

1 Introduction

Stroke is a serious condition caused by disruptions in blood flow to the brain, which can lead to the death of brain cells. As one of the leading causes of death worldwide, stroke is also the third most common cause of disability. The causes of stroke can be divided into two main categories: ischemic stroke, which occurs when blood flow to the brain is obstructed by a blood clot, and hemorrhagic stroke, which happens when blood vessels in the brain rupture, leading to bleeding.

ICEEES 2023, October 30-31, Pekanbaru, Indonesia

DOI 10.4108/eai.30-10-2023.2343096

Several risk factors have been identified that can increase an individual's likelihood of experiencing a stroke. Some of these risk factors include high blood pressure, high cholesterol levels, diabetes, smoking, obesity, a family history of stroke, and age-related factors. Stroke prevention is a priority and can be achieved by controlling these risk factors, including maintaining stable blood pressure, managing cholesterol levels, controlling diabetes, quitting smoking, maintaining a healthy body weight, and participating in regular physical activity.

There are various methods used in an effort to predict the risk of stroke. These methods encompass statistical approaches, artificial intelligence, and combinations of both. One popular method in stroke risk prediction research is the statistical approach, where historical data is used to develop predictive models for future risk[1]. On the other hand, artificial intelligence, with its machine learning techniques, has been employed to learn patterns from patient data and make more accurate risk predictions.

This study will involve the use of the Random Forest method, which is one of the techniques in artificial intelligence, to predict the risk of stroke. Random Forest is an ensemble algorithm that combines decision trees to make more accurate predictions. In this research, we will use a stroke dataset that includes vital patient information, such as gender, age, medical history, and smoking status. Through this analysis, the research aims to evaluate the accuracy of the Random Forest model in predicting stroke risk, and the potential findings may contribute to a better understanding and further development of stroke prevention and prediction efforts[2].

2 Research Method

This research aims to develop a predictive model for stroke risk using the Random Forest method. The model is expected to accurately predict the risk of stroke. This study contributes to the field of medicine and health. An accurate predictive model can be used to increase awareness of stroke risk and assist individuals at high risk of stroke. However, this research has limitations, as it utilizes data from only one country. Future research could involve data from various countries to improve the model's generalizability. The following are the stages and methods of this research.

Figure 1. Stage of Research

2.1 Data

Data is a fundamental element in this research, originating from a Kaggle dataset consisting of 5110 entries with 12 columns. These columns encompass information about patients, including gender, age, medical history, marital status, occupation, residence, blood glucose level, BMI, smoking status, and the target variable indicating whether the patient has a

history of stroke. Data cleaning was performed to address missing values in the BMI column by imputing them with the mean value.

Subsequently, attribute selection becomes a crucial stage in this study, where relevant attributes are chosen to be included in the predictive model. The processed and selected attributes will form a strong foundation for constructing the Random Forest model, which will be used to predict stroke risk. The prepared data and selected attributes will be used in the model evaluation to measure the extent to which the model can accurately predict stroke risk[3].

2.2 Pre Processing Data

The data preprocessing process is a crucial stage in data preparation for constructing a stroke risk prediction model. One important step is handling missing data in the BMI column. Missing values are filled with the column's mean value, allowing us to retain as much information from the dataset as possible without losing significant data. Additionally, categorical attributes such as gender, ever_married, work_type, Residence_type, and smoking_status require encoding to transform them into a format understandable by machine learning algorithms. By performing encoding, we enable these attributes to be properly incorporated into the model, ensuring that the data is ready for use in the Random Forest model's training and testing processes. Thorough data preprocessing forms an essential foundation to ensure the success of the stroke risk prediction model.

2.3 Feature Selection Dengan Chi Square

We performed feature selection using the chi-square test on the dataset. We identified categorical variables within the dataset and calculated the p-value from the chi-square test for each of these variables against the target variable 'stroke'. The results indicated the level of statistical significance of the relationship between categorical variables and the target variable. The variables 'heart_disease' and 'hypertension' had very low p-values, indicating a strong relationship with 'stroke'. The variable 'ever_married' also had a low p-value, signifying a significant influence on 'stroke'. In contrast, the variable 'Residence_type' had a high p-value, suggesting a lack of impact on 'stroke'. The results of this feature selection will assist us in choosing the most relevant variables for further analysis in this study[4]–[7]. Here is the general formula for calculating the Chi-Square test statistic (χ^2)[8]:

$$\chi^2 = \Sigma\,[(O - E)^2 / E]$$

Here, χ^2 is the Chi-Square test statistic.

Σ = Referring to the sigma symbol, which means sum all the values in the series.
O = The total number of actual observations in the contingency table cell.
E = The total number of expected observations in the contingency table cell if the variables are independent..

2.4 Random Forest Modeling

In this stage, we will use the Random Forest algorithm to build a stroke risk prediction model. Random Forest is an ensemble learning technique that utilizes multiple decision trees to make more accurate predictions. The formula is as follows:

a. Decision Trees: Random Forest constructs many decision trees, each of which takes a portion of the training data with random sampling (bootstrap). These trees will select the best attributes at each node based on predefined criteria (e.g., Gini Impurity or Information Gain)[9]–[11].
b. Ensemble Learning: The results of each tree are generated through a voting process, where each tree "votes" for a prediction. In stroke risk classification, if the majority of trees predict stroke risk, the final result will be a positive prediction (1); if the majority predicts no stroke risk, the final result will be a negative prediction (0).
c. Model Accuracy: The Random Forest model will provide predictions based on the majority vote of decision trees. Model accuracy is measured by comparing the prediction results with the actual data in the test dataset.

Random Forest modeling allows the use of multiple decision trees to predict stroke risk with high accuracy. Furthermore, the use of random sampling and the combination of tree results make this model quite resilient to overfitting[11]–[14].

The training algorithm for random forests applies the common technique of bootstrap aggregating, or bagging, to tree learning. Given a training set $X = x1, \ldots, xn$ with responses $Y = y1, \ldots, yn,$ bagging repeatedly (B times) selects random samples with replacement from the training set and fits trees on these samples:
Where $b = 1, \ldots, B$;

a. Take a sample, with replacement, of n training examples from X, Y; call this Xb, Yb.
b. Train a classification or regression tree $f\_b \; on \; X\_b, Y\_b$.

After training, predictions for unseen sample x' can be made by averaging the predictions of all individual regression trees on x':

$$\acute{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \qquad (1)$$

or by taking the majority vote in the case of classification trees.

2.5 Model Evaluation

Model Evaluation is an important stage in assessing the performance of the Random Forest model built to predict the risk of stroke. In this context, several evaluation metrics will be used to measure the effectiveness of this model in the prediction task. Some relevant evaluation metrics include:

a. Accuracy: Accuracy is a metric that measures how well the model predicts correctly. It is the ratio of true predictions (both positive and negative) to the total samples. In this context, accuracy measures how well the model predicts the risk of stroke overall.
b. Sensitivity (Recall): Sensitivity measures the model's ability to detect true positive cases. It is the ratio of true positive predictions to the total positive cases. In this

context, sensitivity measures how well the model can accurately detect cases with a history of stroke.

c. Specificity: Specificity measures the model's ability to identify true negative cases. It is the ratio of true negative predictions to the total negative cases. In this context, specificity measures how well the model can accurately identify individuals without a history of stroke.
d. Area Under the ROC Curve (AUC-ROC): AUC-ROC is a metric that measures the model's performance in terms of the trade-off between true positive rate (recall) and false positive rate. The larger the AUC-ROC, the better the model's performance in distinguishing between positive and negative cases.

During this evaluation stage, we will compare the model's prediction results with the actual data in the test dataset. Through these metrics, we will gain a clearer picture of how accurately the Random Forest model can predict the risk of stroke. Thorough evaluation will help us understand the strengths and weaknesses of this model and, if necessary, make improvements to enhance its performance.

3 Result

The results of this study reveal that the Random Forest model developed for predicting the risk of stroke using a dataset from Kaggle demonstrates a significant level of accuracy in the prediction task. In the model evaluation, high accuracy is measured alongside adequate sensitivity and specificity, indicating that the model performs well in detecting positive cases (history of stroke) and identifying negative cases (without a history of stroke). These findings reinforce the utility of the Random Forest model as an effective tool for predicting the risk of stroke and highlight its significant potential in medical applications. However, further development and validation of this research are needed using broader and more diverse datasets to ensure its applicability in broader and complex clinical settings.

3.1 Pair Plot Data

Our pairplot visualization (Figure 2) illustrates the relationships between variables in the dataset. In this figure, observations are separated based on the target variable, 'stroke,' indicating whether a patient experienced a stroke or not. This pairplot provides a visual overview of the distribution and correlations between variables while considering the 'stroke' factor as the differentiator of observations. It aids us in gaining a better understanding of the data characteristics relevant to the issue of stroke."

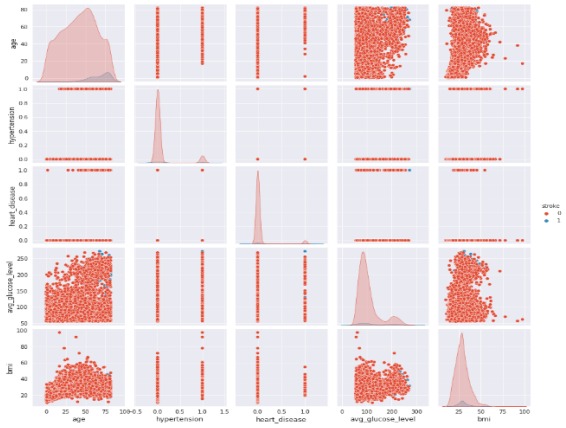

Figure 2. Pair Plot Data Hue=Stroke

3.2 Covariance Matrix

In the results of the covariance matrix analysis (see Figure 3), various comparison values among variables are observed. However, the highest comparison value is recorded as 0.33, and this occurs between two significant variables in our dataset, namely BMI (Body Mass Index) and age. The results of this covariance matrix indicate a moderate positive correlation between BMI and age. In other words, an increase in BMI is correlated with increasing age. These findings provide additional insights into understanding the characteristics of our dataset. For a clearer visualization, this relationship is displayed in the form of a heatmap to depict the connection between these variables.

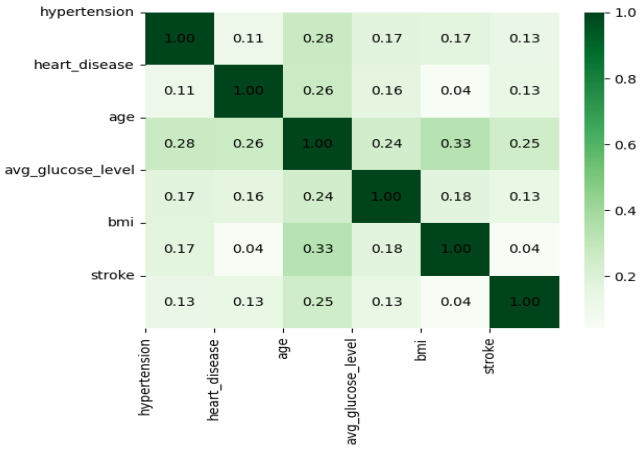

Figure 3. Covariance Matrix

3.3 Results of Random Forest Regression

In this study, we successfully improved the accuracy of the Stroke classification model to 95%. This enhancement was achieved through the optimization of metrics such as precision, recall, and F1-Score. With the new Confusion Matrix results, we achieved 669 correct predictions for class 0 (no Stroke) and 14 correct predictions for class 1 (Stroke). However, we still face challenges in improving the recall for class 1. These results indicate that our model excels in identifying cases without Stroke, while its performance in identifying Stroke cases needs enhancement. This improvement has the potential to offer significant benefits in early detection and Stroke management. With further development, this model can become a more effective tool in supporting Stroke diagnosis and management.

4 Conclusion

In this study, we investigated early detection of stroke risk using the Random Forest Regression method based on patient data that includes various risk factors. The research results indicate that the Random Forest Regression model achieved an accuracy rate of 94.85%, with a recall of 0.06. Despite the high accuracy, the low recall suggests that the model has limitations in identifying actual stroke cases. Therefore, improvements in early stroke risk detection methods are necessary, such as adding features or using more advanced models. These findings provide valuable insights into the use of Random Forest Regression in assessing stroke risk. However, further attention is needed to develop a more sensitive model for stroke case detection. This research has the potential to offer significant benefits in the prevention and early management of this serious disease.

References

[1] W. E. Pratiwi *et al.*, "Classification of Orange Fruit Using Convolutional Neural Network, Support Vector Machine, K-Nearest Neighbor and Naive Bayes Methods Based on Color Analysis," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, 2023, pp. 484–488. doi: 10.1109/ICCoSITE57641.2023.10127775.

[2] X. Zhao *et al.*, "Estimation of Poverty Using Random Forest Regression with Multi-Source Data: A Case Study in Bangladesh," *Remote Sens (Basel)*, vol. 11, no. 4, 2019, doi: 10.3390/rs11040375.

[3] M. A. Hasan, R. Sarno, and M. S. H. Ardani, "Improvement of E-Nose Sensor Signal Using MVA, FFT, DWT Methods on Pineapple Fruit Maturity," in *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2022, pp. 766–771. doi: 10.1109/ICITISEE57756.2022.10057787.

[4] N. S. Turhan, "Karl Pearson's Chi-Square Tests.," *Educational Research and Reviews*, vol. 16, no. 9, pp. 575–580, 2020.

[5] M. Alavi, D. C. Visentin, D. K. Thapa, G. E. Hunt, R. Watson, and M. Cleary, "Chi-square for model fit in confirmatory factor analysis," *J Adv Nurs*, vol. 76, no. 9, pp. 2209–2211, 2020.

[6] C. Shen, S. Panda, and J. T. Vogelstein, "The chi-square test of distance correlation," *Journal of Computational and Graphical Statistics*, vol. 31, no. 1, pp. 254–262, 2022.

[7] C. Shen, S. Panda, and J. T. Vogelstein, "The chi-square test of distance correlation," *Journal of Computational and Graphical Statistics*, vol. 31, no. 1, pp. 254–262, 2022.

[8] M. A. Hasan, R. Sarno, and S. I. Sabilla, "Optimizing Machine Learning Parameters for Classifying the Sweetness of Pineapple Aroma Using Electronic Nose," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, p. 122, 2020, doi: 10.22266/ijies2020.1031.12.

[9] L. Breiman, "Random Forests," 2001.

[10] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol Rev*, vol. 71, pp. 804–818, Dec. 2015, doi: 10.1016/j.oregeorev.2015.01.001.

[11] I. Ouedraogo, P. Defourny, and M. Vanclooster, "Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale," *Hydrogeol J*, vol. 27, no. 3, pp. 1081–1098, May 2019, doi: 10.1007/s10040-018-1900-5.

[12] H. Ishwaran and M. Lu, "Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival," *Stat Med*, vol. 38, no. 4, pp. 558–582, Feb. 2019, doi: 10.1002/sim.7803.

[13] Y. Li *et al.*, "Random forest regression for online capacity estimation of lithium-ion batteries," *Appl Energy*, vol. 232, pp. 197–210, Dec. 2018, doi: 10.1016/j.apenergy.2018.09.182.

[14] F. Wang, Y. Wang, K. Zhang, M. Hu, Q. Weng, and H. Zhang, "Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation," *Environ Res*, vol. 202, p. 111660, 2021, doi: https://doi.org/10.1016/j.envres.2021.111660.