Journal of Information System, Technology and Engineering
Volume 2, No. 3, pp. 318-322
E-ISSN: 2987-6117
http://gemapublisher.com/index.php/jiste
Received: July 2024
Accepted: August 2024
Published: September 2024

Data Management Analysis for Predicting Stroke using RapidMiner

Johanes Fernandes Andry[1], Kosasi[2], Hendy Tannady[3]

Universitas Bunda Mulia[1,2], Universitas Esa Unggul[3]

Correspondence Email: hendy.tannady@esaunggul.ac.id[3]

Abstract

Stroke are known as the second most leading cause of death. Because of this, data mining techniques are already being used to predict patients that may have stroke. Therefore, we are doing a study to try using data mining techniques using RapidMiner to find information or patterns regarding stroke from a dataset obtained from Kaggle. Three data mining techniques are used in this study, that is classification using decision trees, association rule using FP-Growth algorithm, and clustering technique using k-Means algorithm. Using RapidMiner, we are able to process the dataset using the operators provided in the application. As the result, we found out that due to an unbalanced data, the decision tree model made were only able to predict 68,75% patients as having stroke. With the association rule technique, we found out that most attributes in the dataset does not really associated with each other. With the clustering technique, we were able to group up patients and found out that most patients that have stroke are averaged in the age of 58, with 31 bmi, and 201 average glucose level.

Keywords: stroke, data mining, decision tree.

INTRODUCTION

A neurologic disability known as a stroke results from a disruption in blood flow to the brain. A malfunctioning nervous system is known as a neurologic impairment (Dinata et al., 2013). Stroke symptoms include difficulty speaking or understanding speech, tingling or paralysis of the face, arms, or legs, double vision, abrupt headaches, and a lack of coordination. Stroke are classified into two types: ischemic stroke and hemorrhagic stroke (Boehme et al., 2017). Ischemic stroke makes up to 87% of all stroke cases and caused by a blood clot because of fatty plaque in the blood vessel. Hemorrhagic stroke on the other hand, makes up to 13% of all stroke cases and caused by a ruptured blood vessel which then bleeds into the area surrounding the brain. Ministry of Health of Indonesia shows that in Indonesia, the prevalence of stroke is as high as 14,7‰ in the province of East Kalimantan and 14,6‰ in DI Yogyakarta. Some risk factors of stroke include age, gender, high blood pressure, diabetes, smoking, obesity, cholesterol, etc. Some of those risk factors can be controlled to prevent stroke (Chen et al., 2016). That is, by adapting a healthy diet, doing more physical activities, stop smoking, reducing stress, and routinely consult to health expert (Ramageri, 2020).

DOI: https://doi.org/10.61487/jiste.v2i3.95

In 2019, stroke ranked second in terms of cause of death, behind only heart disease, according to data from the World Health Organization. Consequently, the application of data mining techniques to forecast the likelihood of a stroke is gaining traction (Milovic, 2022). The goal of data mining is to get knowledge or insight from massive datasets by discovering patterns and associations. In the health industry, data mining has found its use for various cases (Durairaj & Ranjani, 2013). This includes predicting trends in patient in healthcare organizations, predicting heart attacks, also various diseases like AIDS, cancer, hepatitis, diabetes, etc. (Gorade et al., 2017). These data mining applications uses different kinds of data mining techniques such as classification, association rules, and clustering to visualize the relation of patients that may show symptoms of diseases, or to show patterns that may lead to a disease (Song & Lu, 2015).

This paper shows the use of data mining techniques on a stroke dataset to find interesting information or patterns regarding stroke. The dataset used by this study is retrieved from Kaggle. All data mining technique implementations are done using RapidMiner software (Zeng et al., 2015). The rest of this paper consists of literature study section explains the data mining techniques used, the methods section explains the dataset used in the study and the process of applying data mining technique, results section explains the result of each technique, and the conclusion sections shows the summary of the results.

METHOD

The stroke prediction dataset used in this study is retrieved from Kaggle, courtesy of fedesoriano at https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. This dataset contains 12 attributes. The attribute 'gender' has one missing value and the attribute 'bmi' has 201 missing values. The dataset is heavily unbalanced due to the high number of patients that doesn't have stroke compared to those who does. Before being used for data mining process, the dataset is cleaned for its missing values. Example with missing 'gender' is filtered out since there is only one missing. Examples with missing 'bmi' has its 'bmi' value filled with the average value. The dataset now has 5.109 examples. The 'bmi' values in this dataset is also unbalanced due to a few records having a really high value. Therefore, the 'bmi' attribute is filtered to exclude those values. The dataset now has 4.942 examples. Operators in RapidMiner only takes a certain type of attributes to work. Therefore, only some relevant attributes are selected for the data mining process. The attributes selected are different for each technique, which will be shown in the results section. To apply data mining technique, we used the built-in operator for each technique in RapidMiner. To test the efficacy of the built model, we employed the Decision Tree operator in conjunction with the Apply Model and Performance operators to determine the categorization method. Together, the FP-Growth and the Create Association Rules operator formed the basis of the association method. For the clustering technique, we used the Clustering (k-Means) and the Performance operator. All the results of each technique are analyzed to create a conclusion. For the classification technique, an Optimize Parameter will be used to tune the parameters of the operators to get the highest accuracies. For the association technique, the measurement explained above will be used to evaluate each association rules created. For the clustering technique, the Davies-Bouldin Index explained above will be used to measure the validity of the clusters.

RESULT AND DISCUSSION

A model can be trained using the classification technique to determine if a patient with a certain attribute is at risk of having a stroke. We accomplished this by narrowing the dataset's features based on the aforementioned stroke risk factors. Age, gender, hypertension,

DOI: https://doi.org/10.61487/jiste.v2i3.95

average glucose level (a measure of diabetes), heart disease, body mass index (BMI) (a measure of obesity), and smoking status are the features. As a label or class, the'stroke' attribute is also present. The setup for cleaning and shrinking the dataset is before the Optimize Parameter operation. The parameters of the operator inside this wrapper operator can be tuned using this operator. Following its incorporation into the Optimize Parameter operator, the dataset is partitioned into training data with a ratio of 7:3 and testing data with a ratio of 3:1. The decision tree model is then built using the training data and the criteria parameter set to 'information_gain'. In addition to confidence and "minimal_size_for_split," the wrapper adjusts other parameters like maximal_depth and minimal_leaf_size.

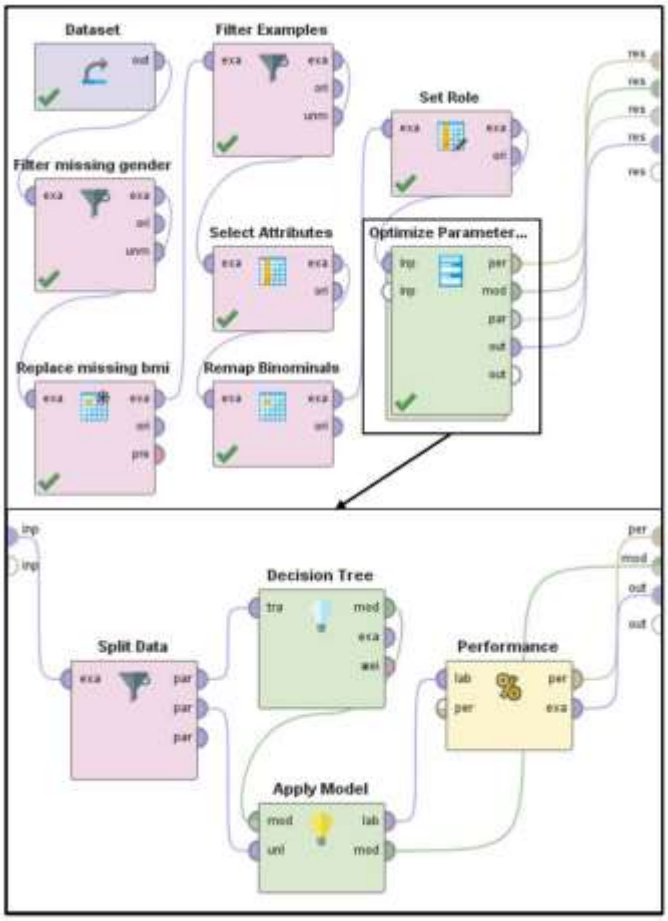

Figure 1. Operator Configuration for Classification Technique

Next, the testing data and model were supplied to the Apply Model operator. The data was subsequently sent to the Performance operator for decision tree model accuracy evaluation. After running the configuration several times, we got the result. There, we can see the four highest accuracies along with the tuned parameters. As we can see, we got results with almost 100% accuracy. However, these numbers are, unfortunately, heavily skewed towards the prediction of someone that does not have stroke because the number of examples is too high compared to those who does. On the first row, the model predicted 16 patients to have stroke, but 5 were incorrect and also misses 61 patients bringing the precision and recall value 68,75% and 15,28% respectively. The model on the second row actually got 100% precision by correctly predicted 6 patients to have stroke, but still misses the other 66, so the recall value is only 8,33%. Because of this, we were unable to develop a model that could

DOI: https://doi.org/10.61487/jiste.v2i3.95

accurately use this dataset to forecast which patients would suffer a stroke. Association rules are generated using the association technique in order to identify relationships between stroke-related features in the dataset. We were able to employ characteristics with either nominal or categorical values with the FP-Growth operator. Hence, we settled on the gender, heart disease, hypertension, smoking status, and stroke features. In the setup of the operators that were employed to generate the rules for the association. Except for the Select Attributes operator, which now only selects the aforementioned attributes, the first five operators were identical to the ones used on the classification technique. Connecting the reduced dataset to the FP-Growth operator, we have set the parameter min_support to 0.3 and kept the remaining parameters at their default values. The operator then passes the frequently used itemset to the operator that creates the association rules. Along with the default values for the other parameters, the min_confidence parameter is set to 0.5.

The association rules created from the operators. The result is filtered so it only shows rules that includes the attribute 'stroke' since we only interested in that. As we can see on the result, most rules have a lift value really close to 1, despite having a high confidence. Therefore, the items on these rules are most likely to be independent, which means that they just happened to appear together. The first three rules have a lift value a bit lower than 1. Which means the antecedent have a negative effect on the consequent. The rule on number 6, for example, means that most patients that have hypertension does not have stroke and heart disease in this dataset. We cannot really take that conclusion, however, since usually, high blood pressure can have effects on the heart and may also lead to stroke as mentioned above.

The clustering technique is used to create clusters that can group up the patients in the dataset with distinct characteristics. The Clustering (k-Means) operator we used accepts numerical values. Therefore, we chose the attributes 'age', 'bmi', and 'avg_glucose_level'. The attribute 'stroke' is also included and parsed to number so we can see the proportion of the patients in each cluster that have stroke. The configuration of the operators is unchanged from previously; however, the Select Attributes now selects the aforementioned attributes, and the 'stroke' attribute now has an extra Parse Number operator. With 2 clusters, the reduced dataset is linked to the Clustering (k-Means) operation through the parameter k. Everything else is set to default. Next, we send the cluster model to the Performance operator for assessment, using 'Davies Bouldin' as the primary criterion parameter. Default values are also used for other parameters. groupings produced by the Clustering algorithm. There, we can observe the two groups that have averaged the traits we chose earlier. Cluster 0 contains 708 objects, while cluster 1 contains 4234 items. Stroke occurs in 12% of patients in the first cluster, which is the largest relative proportion. The average body mass index (BMI) of the patients in this cluster is 31, their average age is 58, and their average glucose level is 201. Stroke is more common in elderly people who are overweight and have diabetes, according to this finding. Stroke affects just 3,7 percent of the 4234 patients in the second cluster. This group included individuals with a mean glucose level of 90, a body mass index (BMI) of 27, and a patient age of 40 or younger. The Davies-Bouldin Index we got from these clusters are 0,519 so the number of clusters are good enough. We also tried with a different number of clusters, with 3 clusters have a DBI of 0,906 and 4 clusters have a DBI of 0,816.

CONCLUSION

Using the classification technique, we were able to create a decision tree based on the dataset. Our inability to develop a model with improved accuracy in stroke prediction is regrettable and stems from imbalanced data. Using the association technique, we were able to find some association rules with the attributes chosen. But most of those rules seem to not

DOI: https://doi.org/10.61487/jiste.v2i3.95

hold any interesting relationships. Furthermore, the first three rules are saying that hypertension and heart disease have a negative effect on stroke, which is said to be otherwise. From the clusters made we can see the average body mass index, age, and average glucose level of patients that have stroke.

REFERENCES

Dinata, C. A., Syafrita, Y., & Sastri, S. (2013). Gambaran Faktor Risiko dan Tipe Stroke pada Pasien Rawat Inap di Bagian Penyakit Dalam RSUD Kabupaten Solok Selatan Periode 1 Januari 2010 - 31 Juni 2012. *Jurnal Kesehatan Andalas, 2*(2), 57-61.

Boehme, A. K., Esenwa, C., & Elkind, M. S. V. (2017). Stroke Risk Factors, Genetics, and Prevention. *Circulation Research, 120*(3), 472-495.

Centers of Disease Control and Prevention, "Preventing Stroke: Healthy Living," Centers for Disease Control and Prevention, 31 January 2020. [Online]. Available: https://www.cdc.gov/stroke/healthy_living.htm. [Accessed 9 May 2021].

World Health Organization, "The top 10 causes of death," World Health Organization, 9 December 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Accessed 9 May 2021].

Ramageri, B. M. (2020). Data Mining Techniques and Applications. *Indian Journal of Computer Science and Engineering, 1*(4), 301-305.

Milovic, B., & Milovic, M. (2022). Prediction and Decision Making in Health Care using Data Mining. *International Journal of Public Health Science (IJPHS), 1*(2), 69-78.

Durairaj, M., & Ranjani, V. (2013). Data MIning Applications in Healthcare Sector: A Study. International Journal of Scientific & Technology Research, 2(10), 29-35.

Gorade, S. M., Deo, A., & Purohit, P. (2017). A Study of Some Data Mining Classification Techniques. *International Research Journal of Engineering and Technology, 4*(4), 3112-3115.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry, 27*(2), 130-135.

Zeng, Y., Yin, S., Liu, J., & Zhang, M. (2015). *Research of Improved FP-Growth Algorithm in Association Rules Mining*. London: Hindawi.

Chen, R., Ovbiagele, B., & Feng, W. (2016). Diabetes and Stroke: Epidemiology, Pathophysiology, Pharmaceuticals and Outcomes. *The American Journal of the Medical Sciences, 351*(4), 380-386.

DOI: https://doi.org/10.61487/jiste.v2i3.95