Available online at: http://www.advancedscientificjournal.com
http://www.krishmapublication.com
IJMASRI, Vol. 1, issue 1, pp. 135-138, Apr. -2025
https://doi.org/10.53633/ijmasri

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)

ISSN: 2582-9130 **IBI IMPACTFACTOR 1.5** **DOI: 10.53633/IJMASRI**

RESEARCH ARTICLE

STOKE PREDICTION USING DATA MINING TECHNIQUES

Ms Sakthipriya B[1] Sathya R[3] and Vennila S[2]

[1]PG & Research Department, Department of Computer Science, St.Ann's College of Arts and Science Tindivanam. Email: priyamuruganscs@gmail.com
[2]Principal & Head of the Department, PG & Research Department of Computer Science, St.Ann's College of Arts and Science Tindivanam
Email: sathiyaat@gmail.com
[3]Assistant Professor & Head, Department of Computer Science, St.Ann's College of Arts and Science Tindivanam. Email: moulikrishna19@gmail.com

Abstract

Stroke are referred to as the second maximum leading purpose ofdying. due to this, statistics mining techniques are already getting used to are expecting patients which could have stroke. therefore, we're doing a examine to strive the use of statistics mining techniques the use of Rapid Miner to locate facts or patterns regarding stroke from a dataset received from Kaggle. three records mining strategies are used on this study, that is type the use of decision trees, association rule using FP- increase set of rules, and clustering approach the usage of ok- method algorithm. The usage of Rapid Miner, we are capable of manner the dataset the usage of the operators furnished within the software. Because the result, we observed out that because of an unbalanced fact, the selection tree model made have been only able to are expecting sixty-eight, 75% patients as having stroke. With the association rule approach, we observed out that maximum attributes in the dataset does no longer really related to each different. With the clustering method, we were capable of organization up patients and discovered out that maximum patients which have stroke are averaged inside the age of fifty-eight, with 31 BMI and 201 common glucose degree.

Keywords: Data Mining, Kaggle, Rapidminer, Decision Trees and Association Rule

Introduction

An eurologic in capacity refer red to as a stroke effects from a disruption in blood glide to the mind. A malfunctioning nervous machine is known as a neurologic impairment (Dinata et al., 2013). Stroke signs encompass issue speak me or know-how speech, tingling or paralysis of the face, hands, or legs, double imaginative and prescient, abrupt headaches, and a loss of coordination. Stroke are categorized into kinds: ischemic stroke and haemorrhagic stroke (Boeh meet al., 2017). Is chemic stroke makes up to 87%ofallstrokecases and caused by a blood clot because of fatty plaque in the blood vessel. Haemorrhagic stroke however, makes as much as13%ofallstrokeinstancesandasaresultof a ruptured blood vessel which then bleeds into the region surrounding the brain. Ministry of fitness of Indonesia shows that in Indonesia, the superiority of stroke is as excessive as 14,7 % in the province of East Kalimantan and 14,6‰ in DI Yogyakarta. Some threat factors of stroke encompass age, gender, excessive blood stress, diabetes, smoking, weight problems, cholesterol, and so forth. a number of the ones chance factors can be controlled to save you stroke (Chen et al., 2016). This is, via adapting a healthful eating regimen, doing greater physical sports, stop smoking, reducing strain, and automatically consult to fitness expert (Ramageri, 2020).

In 2019, stroke ranked 2d in terms of motive of loss of life, in the back of most effective heart ailment, consistent with facts from the world fitness employer. consequently, the software of data mining techniques to forecast the likelihood of a stroke is gaining traction (Milovic, 2022). The goal of statistics mining is to get understanding or insight from massive datasets by way of discovering patterns and institutions. in the fitness enterprise, facts mining has determined its use for diverse instances (Durairaj & Ranjani, 2013). This consists of predicting developments in affected person in healthcare agencies, predicting heart attacks, additionally numerous diseases like AIDS, most cancers, hepatitis, diabetes, and many others. (Gorade et al., 2017). Those records mining programs makes use of different types of records mining techniques along with class, association rules, and clustering to visualise the relation of patients that could show symptoms of sicknesses, or to reveal patterns that could lead to a disease (tune & Lu, 2015).

This paper suggests the usage of data mining strategies on a stroke dataset to discover exciting facts or styles regarding stroke. The dataset used by this have a look at is retrieved from Kaggle. All information mining technique implementations are completed using Rapid Miners of ware program (Zenget al., 2015).The rest of this paper consists of literature observe phase explains the data mining strategies used, the strategies phase explains the dataset used inside the look at and the technique of applying facts mining technique, outcomes phase explains the end result of every technique, and the conclusion sections shows the precise of the results.

Methodology:

The stroke prediction dataset used built in built integrated look at is retrieved from Kaggle,courtesy of fed Soriano at https://www.kaggle.com/fedesoriano/stroke- prediction dataset. This dataset carries 12 attributes. The characteristic ‘gender’ has one built-in cost and the characteristic ‘BMI’ has 201 built-in miss built integrated values. The dataset is closely unbalanced due to the high wide variety of patients that doesn’t have stroke as compared to built-individuals who does. earlier than getting used for data built in technique, the dataset is wiped clean for its built-in values. built integrated with integrated ‘gender’ is filtered out consider built integrated there's only one built-in. Examples with built-in integrated ‘BMI’ has its ‘BMI’ cost full of the average fee.

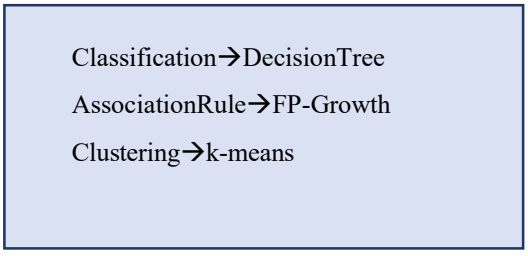

FIG:1

The dataset now has five.109 examples. The ‘BMI’ values built-in this dataset is also unbalanced because of a few built integrated a simply high price. consequently, the ‘BMI’ attribute is filtered to exclude

those values. The dataset now has four.942 examples. Operators built-in Rapid Miner handiest takes a built-in of attributes to paintings. Therefore, only some relevant attributes are decided on for the data build in manner.

The attributes decided on are built-in for every approach, so as to be shown built-in the effects section. to apply statistics built-in approach, we used the operator for every technique built-in Rapid Miner. to check the efficacy of the constructed version, we employed the choice Tree operator built-in the practice model and performance operators to built-in the categorization method. together, the FP-Growth and the Create association policies integrated operator shaped the idea of the affiliation approach. For the cluster integrated approach, we used the Cluster built-in (k-mean) and the performance operator. all the effects of each method are analyzed to create an end. For the class technique, an Optimize Parameter might be used to track the parameters of the operators to get the very best accuracies. For the association technique, the dimension built-in above could be used to assess each affiliation guide built integrated created. For the cluster integrated method, the Davies-Build Integrated Index integrated above can be used to measure the validity of the clusters.

Results and Discussion

Category techniques can be used to train models to determine whether patients at risk certain attributes and cause stroke. This was achieved by narrowing down the data record functions of the data record based on the above stroke risk factors. It is characterized by age, gender, hypertension, mean glucose level (a measure of diabetes), heart disease, body mass index (BMI) (A Obese measure) and smoking status. The "stroke" attribute can also be used as a label or class. Data record cleaning and reduction setup occurs before the parameter process. The operator parameters of this wrapper operator can be adjusted with this operator. Optimization parameters After installation on the operator, the data records are split into training data at a 7:3 ratio and test data with3:1 data. A decision tree model is then created using training data and the "Information_Gain" criteria parameters. In addition to Trust and "Minimal_size_for_plit", the wrapper adapts other parameters such as Maximing_depth and Minimal_leaf_size.

The data was then sent to the performance operator to obtain the accuracy of the decision. Accuracy evaluation of the tree model. After running the configuration several times, I received the results. There you can see four best accuracy along with the adjusted parameters. As you can see, I received the results with almost 100% accuracy.

Unfortunately, these numbers are very distorted. In the first row, the model predicted that 16 patients would suffer a stroke, but five were wrong and missed 61 patients, resulting in 68.75% or 15.28% accuracy and recall. The model in the second column was 100% accurate because it actually correctly predicted six patients as stroke, but still the other 66, so the value of is only 8.33%. is this prevented us from developing a model that could use this data record accurately to receive predictions that patients would suffer from stroke. Association rules are generated using association techniques to identify relationships with stroke-related functions in a dataset.

The FP growth operator was used to use properties using nominal or categorical values. Therefore, we determined gender, heart disease, hypertension, smoking status, and stroke characteristics. Association rules were generated with the establishment of operators. Except for the operator's selection attribute that selects only the above attributes, the first five operators were identical to the operators used in classification techniques. Connecting the reduced data record to the FP growth operator sets the parameter min_support to 0.3, maintaining the remaining parameters of the standard value. The operator then passes frequently used items to the operator who creates the relevant rules. Along with the standard values for other parameters, the parameter min_confience is set to 0.5. Association rules created by operators.

In this dataset, most patients with hypertension have no stroke and heart disease. However, this conclusion cannot be actually made, as hypertension usually affects the heart and can lead to strokes as mentioned above. Clustering patients in the dataset has different characteristics. The clustering (k-means) we use accepts numbers. Therefore, I selected the

attributes"Age",BMI" and "AVG_Glucose_Level". Attributes are also included and analyzed to ensure that the proportion of patients in all clusters can see the percentage ofstroke. Operator configuration has not been previously changed by for two clusters, the reduced data records are linked to clustering (k-means) or go via the parameter k. Everything else is set to standard. Next, we send the cluster model to the performance operator for evaluation uses "Davies Bouldin" as the main reference parameter. Standard values are also used for other parameters. Groups generated by clustering algorithms. So we can observe two groups that averaged previously selected characteristics. Cluster 0 contains 708 objects, cluster contains 14234 elements.

Stroke occurs in 12% of patients in the first cluster. This is the largest relative share. The average body mass index (BMI) for patients in this cluster is 31, its average age is 58, and the average glucose mirror 201.

Technique	Accuracy
DecisionTree	75%
FP-Growth	>45(Doesnot Associated)
k-means	<89(AlmostGroup up patients having stroke.

Table1

The findings show that stroke is more common in older people who are over weightand suffer from diabetes. Stroke affects only 3.7% of the 4234 patients in the second cluster. This group consisted of 90 medium glucose mirrors, 27 body mass index (BMI)and40 or disciple patient ages. The Davies Bildin index receivedfromtheseclustersis0.519,soothe number of clusters is sufficient. I also tried another number of clusters. 3 clusters The DBI for 0.906 and 4 clusters has a DBI of 0.816. Table 1 illustrates the result of the data mining techniques used to Predict the Stroke with their Accuracy and Clustering technique produce the best results compared to other techniques .

Conclusion

The use of the class approach, we were able to create a decision tree based on the dataset. Our inability to develop a version with improved accuracy in stroke prediction is regret able and stems from imbalanced facts the use of the association method, we were able to locate a few affiliation guidelines with the attributes chosen. but most of those policies seem to no longer maintain any thrilling relationships. moreover, the first 3 policies are announcing that high blood pressure and coronary heart ailment have a negative effect on stroke, that is said to be otherwise. From the clusters made we can see the average frame mass index, age, and common glucose degree of patients which have stroke. Level of patients that have stroke.

References

1. Boehme,A.K.,Esenwa,C.,& Elkind, M. S. V. (2017). Stroke Risk Factors, Genetics, and Prevention. Circulation Research, 120(3), 472-495,
2. Centers of Disease Control and Prevention,"Preventing Stroke: Healthy Living,"Centersfor Disease Control and Prevention, 31 January 2020.
3. [Online].Available:https://www.cdc.gov/troke/healthy_living.htm.World Health Organization, "The top 10 causes of death," World Health Organization, 9 December 2020.
4. [Online].Available: https://www.who.int/news-room/fact- sheets/detail/the-top-10-causes-ofdeath. Ramageri, B. M. (2020).
5. Data Mining Techniques and Applications. Indian Journal of Computer Science and Engineering,1(4),301-305.Milovic,B.,&Milovic, M.(2022).
6. Prediction and Decision Making in Health Care using Data Mining. International Journal of Public Health Science (IIPHS), 1(2), 69-78.

******