e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

COMPUTATIONAL HEALTH CARE ANALYSIS USING HADOOP – STROKE PREDICTION

Bobby Prathikshana M.[*1], Nivetha V. [*2], Rinubha P. [*3],
Dr. M. Marimuthu[*4], Dr. P. Velvadivu[*5]
[*1,2,3]M.Sc Decision and Computing Science Dept. of Computing
Coimbatore Institute of Technology, India.
[*4,5]Guide, Assistant Professor, Dept. of Computing Coimbatore Institute of Technology, India.

ABSTRACT

An estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes. According to the World Health Organization, ischemic heart disease and stroke are the leading causes of death worldwide. When a patient, who has suffered stroke is treated, the healthcare institution can obtain huge volumes of valuable data like the time of stroke, influential events preceding the stroke, such as psychological stress, heavy workload or the atmospheric influence. With such information, the doctors would be able to understand the disease better and predict occurrence of strokes to a degree of precision and can provide preventive measures. This study focuses on various techniques to analyse and retrieve the required information from big data in the stroke prediction dataset. PySpark is used to build a predictive model to analyse the possibility of occurrence of stroke in Hadoop environment. And Hive Queries are used in order to extract the information needed.

Keywords: Stroke, Hive, Pyspark, Predictive Model, Mapreduce, Hadoop.

I. INTRODUCTION

According to a study of more than 56 million deaths in 2001, 7.1 million died of heart disease and 5.4 million died of stroke. This indicates that stroke is the second leading cause of death in the world after heart disease, accounting for nearly 10% of the total reported deaths. Stroke is the third leading cause of death in the United States, and approximately 137,000 Americans have lost their lives. Get sick every year. In 2006, 6 out of every 10 stroke deaths were women. In the United States, people have a stroke every 40 seconds, and people die of a stroke every 3-4 minutes. In the United States alone, the cost of this disease is estimated to be approximately $73.7 million. Stroke is one of the main causes of disability in the world. According to a report in 2005, about 1.1 million people have had a stroke, but they have difficulty living in daily activities. According to a study of 450,229 residents of Mashhad, in Iran, it was found that stroke occurred nearly ten years earlier than in Western countries, and the incidence in Iran was also higher than in most countries. Most research on automatic diagnosis of stroke and its subtypes has focused on imaging techniques, computed tomography and magnetic resonance imaging. For example, computed tomography images have been used to diagnose stroke and its subtypes. In order to improve image quality and reduce noise, the symmetry line of the skull was determined, and then a histogram was made for the hemisphere of the brain. Haemorrhagic and chronic stroke are distinguished by a histogram. We use waveforms for acute stroke diagnosis and normal imaging. The accuracy and recovery rate obtained are 90% and 100%, respectively. There are several factors that affect the occurrence of stroke, including genetics, age, gender, and race, as well as certain medical conditions, such as high blood pressure, hypercholesterolemia, heart disease, and diabetes. Quitting smoking, alcohol and daily activities can also reduce the risk of stroke. Using the above-mentioned risk factors and data extraction techniques, a decision support system can be developed. In addition to the doctor's knowledge and experience, the system can also be used to predict stroke. Due to human demand for knowledge and the ever-increasing amount of data, it is inevitable to develop methods for automatically extracting knowledge from this data. Data extraction is about extracting knowledge and attractive models from large amounts of data. The knowledge-based data mining techniques that can be extracted can be divided into three categories: pattern classification, clustering the data and association rule mining.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

II. LITERATURE REVIEW

Many researchers have already used a machine learning-based approach to predict stroke. Govindarajan et al. conducted a study to categorize stroke disorders using a combination of text mining and a machine learning classifier, and collected data from 507 patients. For their analysis, they used different machine learning approaches for training purposes using ANN, and the SGD algorithm gave the best value, 95%. Amini et al. conducted research to predict stroke incidence, collected 807 healthy and unhealthy subjects in their study, and categorized 50 risk factors for stroke, diabetes, cardiovascular disease, smoking, hyperlipidaemia, and alcohol consumption. They used two techniques that had the best accuracy of the c4.5 decision tree algorithm and it was 95% and for the nearest neighbor K the accuracy was 94%. Cheng et al. published a report on estimating the prognosis of an ischemic stroke. In their analysis, 82 data from patients with ischemic stroke were used, two ANN models were used to determine the precision, and 79% and 95% were used. Cheon et al. conducted a study to predict mortality in stroke patients. used 15099 patients to identify stroke occurrence. They used a deep neural network approach to detect strokes. The authors used PCA to extract the medical history and predict a stroke. They have an area under the curve (AUC) of 83%. Singh et al. conducted a stroke prediction study for artificial intelligence. In their research, they used a different method to predict stroke in the Cardiovascular Health Study (CHS) dataset. And they used the decision tree algorithm to extract features from principal component analysis. They used a neural network classification algorithm to precisely build the model they received. Chin et al. conducted a study to detect automated early ischemic stroke. In their study, the main goal was to develop a system that uses CNN to automate primary ischemic stroke. They collected 256 images to train and test the CNN model. Area that cannot be caused by stroke, they used the data extension method to increase the image collected. His CNN method showed an accuracy of 90 °. Jung et al. conducted a study to develop a stroke severity index. They collected 3,577 data from acute ischemic stroke patients. They used various data mining and linear regression techniques for their predictive models. Its prediction function gave the best result of the k-model for the nearest neighbor (95% CI). Monteiro et al. conducted a study to use machine learning to predict the functional outcome of an ischemic stroke. In their research, they applied this technique to a patient who died three months after admission. They achieved an AUC value of over 90%. Kansadub et al. conducted a study to predict stroke risk. For the study, the authors used Naive Bayes, Decision Tree, and the Neural Network to analyze the data and predict a stroke. In their study, they used precision and AUC as a rating of their pointer. In all of these algorithms, they classified the decision tree and the naive Bayes gave the most accurate. Adam et al. conducted a study on the classification of ischemic stroke. They used two models: a logistic regression and a decision tree algorithm to classify ischemic stroke.

III. DATASET AND SOURCE

The dataset that we used was the healthcare dataset on stroke prediction, which was imported from the kaggle website. It included various columns that help in the prediction of stroke like the age, gender, ever_married, presence of hypertension, heart disease, work_type, residence_type,average glucose levels, bmi, smoking_status, stroke. With the help of these influential factors, prediction of stroke is carried forward.

id	gender	age	hypertensi	heart_dise	ever_marr	work_type	Residence	avg_gluco	bmi	smoking_s	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly s	1
51676	Female	61	0	0	Yes	Self-emplc	Rural	202.21	N/A	never smo	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smo	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-emplc	Rural	174.12	24	never smo	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly s	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smo	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smo	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smo	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smo	1
58202	Female	50	1	0	Yes	Self-emplc	Rural	167.41	30.9	never smo	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smo	1
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-emplc	Urban	233.29	48.9	never smo	1
68794	Female	79	0	0	Yes	Self-emplc	Urban	228.7	26.6	never smo	1
64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
4219	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly s	1
70822	Male	80	0	0	Yes	Self-emplc	Rural	104.12	23.5	never smo	1

This is a sample of our dataset. It had a total of 12 attributes and 5110 datavalues.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

IV. METHODOLOGIES

4.1 PySpark

PySpark is a Python API written in Python to support Apache Spark. Apache Spark is a distributed framework that can perform big data analytics. Apache Spark is written in Scala and can be integrated into Python, Scala, Java, R and SQL languages. Basically, it is a computer engine that can process large amounts of data and process it in batch and parallel systems. PySpark was used to implement machine learning on the dataset, which included algorithms like the random forest, logistic regression and decision tree. We have vectorized and converted into a computable format. We take those vectors and store them in an assembler. We have created a pipeline to encode the data and we have fitted the model. After fitting the model we could see that vectors and features are included.

4.1.1 Logistic Regression

Logistic regression, a statistical model, uses Logistic function to model the binary dependent variable, despite the existence of many complex extensions. In classification analysis, logistic regression (or logit regression) was developed to estimate the parameters of the logistic model. (Binary regression table).

We have split the data set into train and test 0.7,0.3 respectively, and applied logistic regression to classify whether the person has been affected by stroke or not.

stroke	rawPrediction	prediction	probability
0	[2.83083888243126...]	0.0	[0.94431972693716...]
0	[3.31370082882613...]	0.0	[0.96489585060542...]
0	[3.10621295962882...]	0.0	[0.95714829730486...]
0	[3.49493380341456...]	0.0	[0.97054327638806...]
0	[2.15488079009535...]	0.0	[0.89612398826406...]
0	[2.86188252053739...]	0.0	[0.94592966516320...]
0	[2.91786886175706...]	0.0	[0.94872272273244...]
0	[2.67502484827357...]	0.0	[0.93553673272220...]
0	[3.12532961434219...]	0.0	[0.95792555892635...]
0	[2.92463265502394...]	0.0	[0.94905077029613...]
1	[2.68316264653270...]	0.0	[0.93602576906613...]
0	[3.13280980646937...]	0.0	[0.95822601139269...]
0	[3.09530741981314...]	0.0	[0.95669876518920...]
0	[3.61932683358818...]	0.0	[0.97389881848191...]
1	[2.54220035484266...]	0.0	[0.92704777690949...]
0	[3.11406013256536...]	0.0	[0.95746900039978...]
0	[3.28600917008622...]	0.0	[0.96394571182553...]

Accuracy : 0.922334949618163

We have the raw prediction and the result for stroke in 0 and 1 and it will display the probability to which the person is affected by stroke. We get around 92.33% accuracy for logistic regression.

4.1.2 Decision Tree

A decision tree is a block diagram similar to a tree structure, where each internal node specifies an attribute test, each branch represents the result of the test, and each leaf node has a class name.

stroke	rawPrediction	prediction	probability
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[741.0,31.0]	0.0	[0.95984455958549...]
0	[537.0,91.0]	0.0	[0.85509554140127...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
1	[537.0,91.0]	0.0	[0.85509554140127...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
1	[537.0,91.0]	0.0	[0.85509554140127...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]
0	[2097.0,11.0]	0.0	[0.99478178368121...]

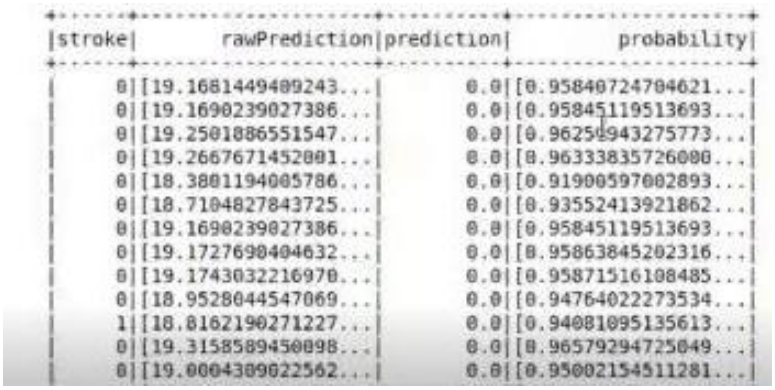
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

Accuracy : 0.9260140905005085

The next algorithm that has been implemented is the decision tree classifier. When we fit this classifier we get 92.6% accuracy. The accuracy is comparatively lower than that of the logistic regression.

4.1.3 Random Forest

Random forests or random decision forests is a method used for classification, regression and other tasks that are computed by constructing a multitude of decision trees at training time and returns the class that has most frequently occurred, that is, the mode of the classes (classification) or the norm prediction (regression) of each of the trees.



stroke	rawPrediction	prediction	probability
0	[19.1681449409243...	0.0	[0.95840724704621...
0	[19.1690239027386...	0.0	[0.95845119513693...
0	[19.2501886551547...	0.0	[0.96259943275773...
0	[19.2667671452001...	0.0	[0.96333835726080...
0	[18.3801194005786...	0.0	[0.91900597002893...
0	[18.7104827843725...	0.0	[0.93552413921862...
0	[19.1690239027386...	0.0	[0.95845119513693...
0	[19.1727690404632...	0.0	[0.95863845202316...
0	[19.1743032216970...	0.0	[0.95871516108485...
0	[18.9528044547069...	0.0	[0.94764022273534...
1	[18.8162190271227...	0.0	[0.94081095135613...
0	[19.3158589450098...	0.0	[0.96579294725049...
0	[19.0004309022562...	0.0	[0.95002154511281...

Accuracy : 0.7922334949618163

We've next fit the random forest classifier. Its accuracy is about 92.33% which is almost similar to logistic regression. From the accuracies of the three algorithms, the decision tree has produced almost more accuracy when compared with random forest and logistic regression. These predictions would help the doctors to prevent the occurrence of stroke in patients and come up with measures.

4.2 Hive

Hive is a data warehousing infrastructure tool for processing structured data in Hadoop that resides on Hadoop to aggregate big data and facilitate querying and analysis. Hive provides various functionalities like reading, writing and managing huge volumes of data in distributed storage. It runs SQL-like queries called HQL (Hive Query Language) that are converted internally to MapReduce jobs. Hive supports DDL (Data Definition Language), DML (Data Manipulation Language), and UDF (User Defined Functions). A number of queries were used to retrieve required information from the processed dataset. The queries and the outputs are as follows:

4.2.1 Query 1:

SELECT COUNT(work_type) FROM stroke_data WHERE smoking_status = 'smokes' AND stroke= 1 GROUP BY residence_type;

Rural	18
Urban	24

4.2.2 Query 2:

SELECT residence_type,work_type, COUNT (work_type) FROM stroke_data WHERE stroke=1 AND hypertension=1 GROUP BY residence_type,work_type;

Rural	Govt_job	3
Urban	Govt_job	5
Rural	Private	15
Urban	Private	18
Rural	Self_employed	14
Urban	Self_employed	11

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

4.2.3 Query 3:

SELECT Id,gender,age, residence_type FROM stroke_data WHERE avg_glucose_level>200 and bmi > 30 GROUP BY residence_type;

9046	Male	67	Urban
13861	Female	52	Urban
64778	Male	82	Rural
43717	Male	57	Urban
54401	Male	80	Urban
47269	Male	74	Rural
19824	Male	76	Rural
17004	Female	70	Urban
2458	Female	78	Rural
56841	Male	58	Rural
45277	Female	74	Rural
41069	Female	45	Rural
53401	Male	71	Rural
13491	Male	80	Rural
44033	Male	56	Rural
71279	Female	71	Urban
11762	Female	76	Urban
17308	Female	72	Urban
46703	Male	68	Urban
24669	Female	77	Rural
59437	Female	57	Urban
20426	Female	78	Urban

Govt_Job	657
Never_Worked	22
Private	2925
Self_employed	819
Children	687
Work_type	1

4.3 Mapreduce

MapReduce is a data processing tool with which data is processed in parallel in a distributed manner. MapReduce is a paradigm with two phases, the mapping phase and the reduction phase. In the mapper, the input is provided as a key-value pair. The output of the mapper is fed as input to the reducer. The reducer will not run until the mapper has ended. The reducer also takes the input in key-value format, and the reducer output is the final output. In this dataset, we've used mapreduce to return the count of the number of people, categorised based on their work types. The work types include government job, private, self employed, never worked and children.

From the result, we conclude that there are 657 government employees, 2925 private employees, 819 self employed people, 22 people who have never worked and 687 children.

V. RESULTS AND DISCUSSION

As we have shown in this study, the pyspark machine learning algorithms like the random forest, logistic regression and decision tree performs very well for the Healthcare Stroke prediction dataset. However, we realize that out of the 3 algorithms decision tree classifier gives better accuracy and out performed other 2 algorithms. To get a better refined approach for improving accuracy, we could use pruning techniques on the features before applying the classifier model to the dataset. In this paper, we have presented the machine learning approaches combining the elements of data imputation, feature selection and prediction using pyspark in Hadoop environment. We provide extensive data analysis using the hive queries and mapreduce results to arrive at better decisions in healthcare regarding the personalized treatment for Stroke. This could possibly provide a well structured analysis on various triggering factors of the disease. Further, our method can be used for identifying potential risk factors for diseases without performing clinical trials.

VI. CONCLUSION

From the Analysis, we've come to know that people are being exposed to various influential factors that contribute to the occurrence of stroke. When people tend to follow a proper diet, exercise regularly and stress-free lifestyle, many factors like hypertension, imbalance in BMI, fluctuations in sugar levels and heart diseases wouldn't affect them. Thereby, promoting the reduction in the number of people being affected by stroke. Also, on the professional front, our analysis could provide better insights in getting a step further towards personalised treatment for Stroke and related heart diseases.

VII. REFERENCES

[1] https://www.javatpoint.com/machine-learning-random-forest-algorithm
[2] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
[3] https://www.geeksforgeeks.org/decision-tree/
[4] https://www.geeksforgeeks.org/understanding-logistic-regression/#:~:text=Logistic%20regression%20is%20basically%20a,regression%20IS%20a%20regression%20model.
[5] https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/
[6] American Heart Association. Heart Disease and Stroke Statistics 2009 Update. American Heart Association, Dallas, Texas, 2009.
[7] Khosla, A.; Cao, Y.; Lin, C.C.Y.; Chiu, H.K.; Hu, J.; Lee, H. An integrated machine learning approach to stroke prediction. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–28 July 2010
[8] Nahla F.Omran, Abdelmgeid A.Ali, H. A. S. F. A.- el ghany, E. M. (2019). Stroke Prediction using Distributed Machine Learning Based on Apache Spark. International Journal of Advanced Science and Technology, 28(15), 89 - 97.