Educational Administration: Theory and Practice
2024, 30(6)(s), 110-117
ISSN: 2148-2403
https://kuey.net/

Research Article

Leveraging Hadoop And Machine Learning Techniques In The Healthcare Industry

Dr. Priti Sadaria[1*], Dr. Rupal Parekh[2]

[1*]Department of Computer Science & Information Technology, Atmiya University, Rajkot, India.priti.sadaria@atmiyauni.ac.in
Department of Computer Science & Information Technology, Atmiya University. Rajkot, India. rbparekhd@gmail.com

Citation: Dr. Priti Sadaria, Dr. Rupal Parekh, (2024) Leveraging Hadoop And Machine Learning Techniques In The Healthcare Industry, *Educational Administration: Theory and Practice*, *30*(6)(s) 110-117
Doi: 10.53555/kuey.v30i6(S).5336

ARTICLEINO

ABSTRACT

In the technological age, data is generated at very quick and processing as well as analysis these very big data is a not easy task. Conventional systems for managing databases are time consuming and are not able to analyze data fully. Healthcare industry have large quantity of data but it is lack of analysis so hidden pattern cannot be identify for prediction of any diseases. To overcome such issue, Big Data is used to handle and control large volume of data which may be either in structured or in unstructured form. The hidden pattern can be identified and prediction can be made about future condition. Hadoop MapReduce has the capability to facilitate healthcare industry to get better prediction of diseases and make faster and proper judgment for right future treatment of patient by analyzing healthcare data. Machine Learning algorithms can help to design predictive healthcare model for community wellness. I proposed system architecture to process and analyze data using Hadoop MapReduce with Machine Learning Techniques and ultimately it leads to prediction of diseases. As a result it increases life span as well as it leads to healthy life and reduce the rate of death by providing timely treatment.

Keywords— *Hadoop, Big Data, MapReduce, Healthcare, Machine Learning Technique*

I. INTRODUCTION

In a time of unparalleled data expansion and technology breakthroughs, the healthcare sector is leading the way in a revolutionary change. Modern machine learning techniques united using Hadoop, a potent distributed data processing platform, have become a compelling paradigm to unlock the enormous potential found in healthcare datasets [1].

In order to shed light on the ways in which these innovations can transform healthcare delivery, advance patient outcomes, and transform medical practices, this research paper explores the complementary relationship between Hadoop besides machine learning in the healthcare industry [2].

Through an examination of the complexities involved in this combination, we aim to elucidate the obstacles, prospects, and practical consequences of utilizing Hadoop and machine learning. Our findings could stimulate the development of innovative methods and tactics aimed at tackling significant healthcare issues during the period of digitalization [3]. This study aims to advance the ongoing development of data-driven approaches in the search for improved healthcare solutions by delving into this junction of technology and healthcare [4].

A. Hadoop

A distributed and scalable framework for storing and processing massive bulks of data across clusters of commodity hardware is offered by Hadoop. The HDFS and MapReduce, two of the framework's main components, provide parallelized computing and storage, enabling enterprises to analyze data at previously unheard-of levels of efficiency [5].
Hadoop's characteristics, when combined with its capacity to manage unstructured and semi-structured data, make it an adaptable solution for a wide range of data types produced in the current digital environment [6].
The fundamental ideas and features of Hadoop, providing the framework for comprehending how this strong framework has completely changed how businesses address the difficulties presented by the big data era [7].

B. Machine Learning

As a transformative force, machine learning (ML) has changed the way we approach difficult problems and extract knowledge from large datasets. Fundamentally, machine learning is a branch of artificial intelligence (AI) that enables computers to learn and perform better without the requirement for specific programming. This paradigm change in machine learning from rule-based systems to data-driven, self-adaptive algorithms has pushed the field toward innovation in a number of different industries [8, 9].

This introduction aims to shed light on the fundamental ideas and wide range of uses of machine learning, highlighting how it can transform decision-making procedures, automate tasks, and reveal hidden patterns within enormous and intricate datasets. As we delve deeper into the study of machine learning algorithms, models, and practical applications, it is becoming more and clearer how this technology is revolutionizing various industries, fields of study, and daily life [10].

Large-scale data collection, management, and analysis are being revolutionized by big data, which has become a transformative force in the healthcare industry. Big data is being integrated into healthcare from a variability of bases, such as Electronic Health Records (EHR), genomics, medical imaging, sensor data, prescription data, and wearable patient behavior data. Due to the abundance of data, predictive analytics has never had more opportunities, giving medical professionals the ability to predict disease trends, customize treatment regimens, and improve patient care overall. By automating the extraction of complex patterns and features from large datasets, the use of deep learning techniques enhances the potential of big data analytics. Disease prediction relies heavily on machine learning algorithms, which have demonstrated their effectiveness in delivering precise and timely insights. Even with these developments, utilizing big data in healthcare fully still requires overcoming obstacles like protecting privacy, handling security issues, and resolving ethical dilemmas. It appears that the years to come will bring about more innovation, better patient outcomes, and the development of personalized and data-driven healthcare practices as more and more healthcare organizations turn to predictive analytics [11].

The study article focuses on creating a hybrid SVM model and algorithm for processing large datasets by combining the techniques of clustering and classification. Because diabetes is a disease that is scattering further rapidly in this day and age due to unhealthy eating and lifestyle choices, I have concentrated my research on diabetes-related conditions. Chronic conditions like nephropathy, retinopathy, and cardiovascular diseases are the end result. This makes it imperative to create a novel method for illness forecasting. Creating Hybrid-SVM algorithms for illness prediction is the main goal of the research.

II. TOOLS AND METHODS

These tools and methods are employed in the research.

A. Hadoop MapReduce

B. K - means clustering

C. SVM

A. Hadoop MapReduce

Hadoop works using the notions of MapReduce and HDFS. Huge datasets can be processed using the MapReduce technique, which is stored in HDFS file format.

B. K - means clustering along with MapReduce

In healthcare industry, K-means clustering is a very helpful algorithm for locating clustering patterns in datasets [12]. The K-means clustering coding shown in the figure allows for the identification of two categories: diabetic and non-diabetic. The diabetic mass is further classified as Low, Medium, and High. [13].

C. Support Vector Machine (SVM)

Design recognition, also referred to as classification technique, is a fundamental job in machine learning. By applying training data and a supervised learning algorithm, estimate utility f can be created. [14].

In this case, classification is accomplished the means of the application of an N-dimensional hyperplane, where each point is identified as a data item. A line that classifies and divides a set of data linearly is called a hyper plane. The support vectors closest to the hyperplane can be used to represent the coordinates of each data item. The distance in this case between the hyperplane and the nearest data point from a set is represented by the margin. Lastly, an optimal hyper plane can differentiate between two cluster categories built on the target variable; one cluster category is situated on one side of the hyperplane, while the additional is on the opposite side. [15].

III. PREPARING DATA SET

The PIMA Indian diabetic data set, which I used for my research, has the following characteristics :

A. Features that are used in clustering

- Living style: Eating and exercise routines
- Diet: Define your eating habits
- Exercise
- Stress
- Family History
- Obesity
- Hypertension
- Glycosylated haemoglobin(HbA1C)
- Region where the patient is located

In addition to these attributes, the K-means clustering procedure can be used to classify facts into two groups: non-diabetic and diabetic clusters at the first level, and the diabetic bunch into three groups: low, moderate, and elevated.

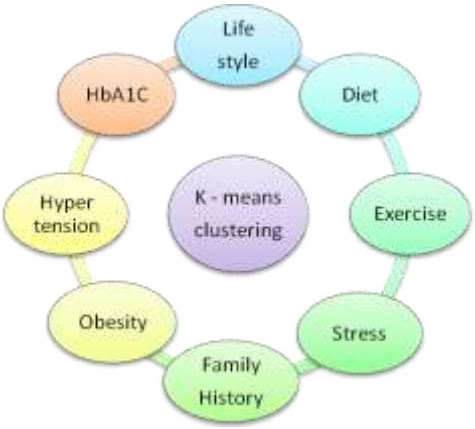

Fig. 1. Features utilized for Clustering

B. Clinical characteristics used to organize and predict using SVM

- Years that a person has had diabetes
- Creatinine
- Increased Density Lipoprotein
- Lipoprotein Low Densities

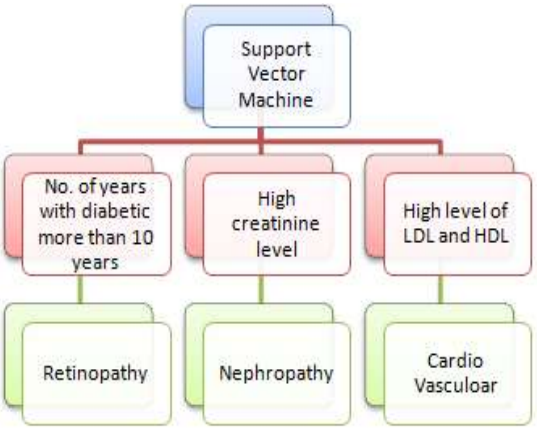

Fig. 2. Diagnostic Characteristics in Classification

The SVM classification procedure, which was based on the clinical features of creatinine, LDL, HDL, and the amount of periods more than thirty having diabetes, also has the potential to predict diabetic difficult disorders.

IV. HYBRID-SVM ALGORITHM PERFORMANCE USING MAPREDUCE

In our research, Hadoop MapReduce was used in conjunction with K-means bunching and SVM for predicting. The provided data for that specific computer was copied to HDFS. After the algorithm uses the

K-means bunching procedure to partition the data into n various clusters, the mapper function begins analyzing the data. Created clusters are used as input by Hybrid SVM to build the prediction model.

A. Blended - SVM algorithm phases

1. Data input that was read from HDFS
2. Data is fed into the mapper.
3. Now that the cluster center point has been established, every data point is directed in the direction of the closest cluster.
4. Next, match the position of every cluster to the mean of all the data points.
5. Repeat steps 3 and 4 until you achieve union.
6. Classifiers are built utilizing the training data set over the cluster outputs after union to create a model.
7. Forecasts on the newly developed model can be made by using a test dataset.
8. At this point, the procedure is complete, and an expected output file is generated.

This research facilitates the prediction of diabetes individuals who may be at risk of acquiring significant disorders like retinopathy, nephropathy, and cardiovascular diseases.

The dataset is handled in two steps in this instance: reduce and mapper. The initial dataset is uploaded onto HDFS, as the name implies, and the mapper receives the data thereafter. K-means clustering, which the mapper uses, results in the formation of many clusters. Hybrid: SVM employs clustered data in key-value pairs that the mapper gives, and it shuffling the data. Ultimately, the reducer aggregated all of the values and saved the result on HDFS in a file format [16].

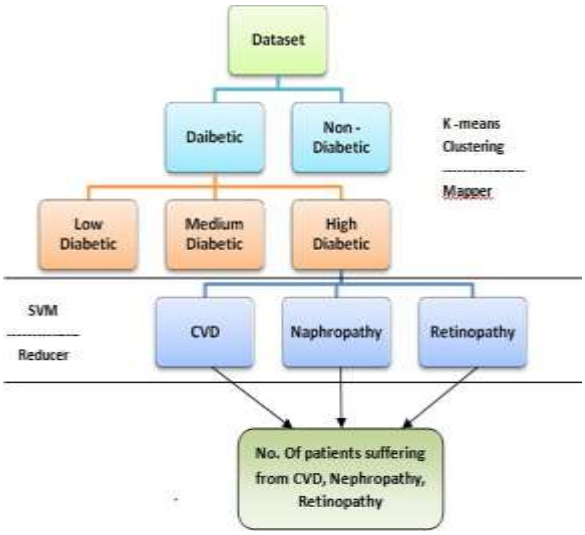

Fig. 3. SVM for Clustering and Classification in a Hybrid Work Environment

Six datasets having size 650000, 843000, 857229, 1650000 and 6000000 were analyzed by implementing Hybrid-SVM methodology and processing time was evaluated. From the research it is concluded that for large dataset, stand alone mode was not competent process the task effectively because of resource constraints. The task effectively completed using Hybrid-SVM algorithm for the district Baroda is shown in Fig. 3 indicates that the dataset categorized into two categories depending on clustering, diabetic and non diabetic. Diabetic cluster again distributed into three categories low risk, medium risk and high risk.

TABLE I TIME NEEDED FOR SIX DATASET PROCESSING WORKING WITH HYBRID-SVM IN STANDALONE MODE WITH HADOOP

Sr.No.	Cities	Total Records	Processing Time
1	Anand	650000	102
2	Rajkot	843000	109
3	Baroda	857229	120
4	Ahmedabad	1650000	187
5	Surat	2000000	228
6	Total of 5 cities	6000000	unable to comprehend

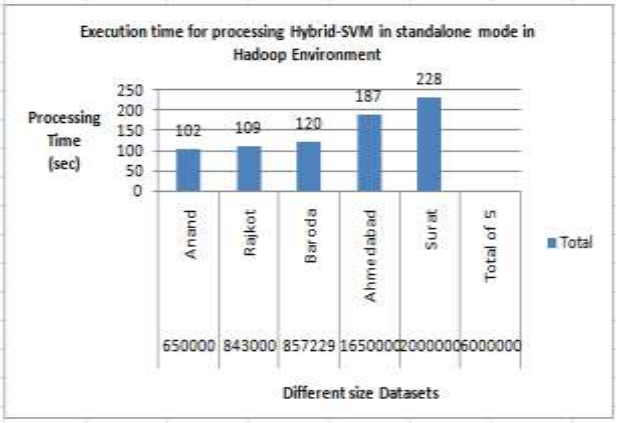

Fig. 4. Time Needed for Six Dataset Processing Working with Hybrid-SVM in Standalone Mode with Hadoop

Table I and Fig. 4 show that, when we used the Hybrid-SVM algorithm for analysis, the processing time increased somewhat in accordance with the amount of data of the various datasets. Table I also shows that the times needed for the records 650000, 843000, 857000, 1650000, and 2000000 were 102, 109, 120, and 187 1st 228 seconds. We fed the system a total of 6000000 records from five districts, however the system could not handle the dataset. Shortages of resources in the standalone mode are the cause of it.

V. DIABETIC DISEASE PREDICTION USING HYBRID-SVM TECHNIQUE

The mapper and reducer functions handle the dataset file once it has been loaded into HDFS. The new K-means and SVM algorithms are used to partition the input files into numerous chunks. First, the dataset is categorized at the cluster group by considering risk indicators including HBA1C, eating habits, exercise, obesity, and family history. Every risk factor is given a score; for instance, family history receives five points, while the remaining factors receive one point each. The total score for each individual record was saved into an output variable based on the risk factor value. The output variable is divided into two groups: those with diabetes and those without. For the non-diabetic cluster, the output variable's value fell between 1 and 5, and for the diabetic cluster, it fell between 6 and 10.

Depending on the score value, the diabetic cluster once more divided into three groups. A score of 9–10 indicates that the condition is high risk for diabetes; an 8 indicates that the condition is medium risk; and a 6-7 indicates that the condition is low risk. Depending on LDL, HDL, and creatinine levels as well as age greater than 10, the diabetic high risk cluster is split into three groups for the second level, which predicts diabetic complications once more: CVD, Nephropathy, and Retinopathy.

Here, the diabetic cluster is treated using the SVM supervised learning technique, and an intermediate key is produced through execution. Every single section's facts is analyzed using the reduction operation, intermediate values are combined, and the outcome of the processing is copied to output records in the Hadoop HDFS.

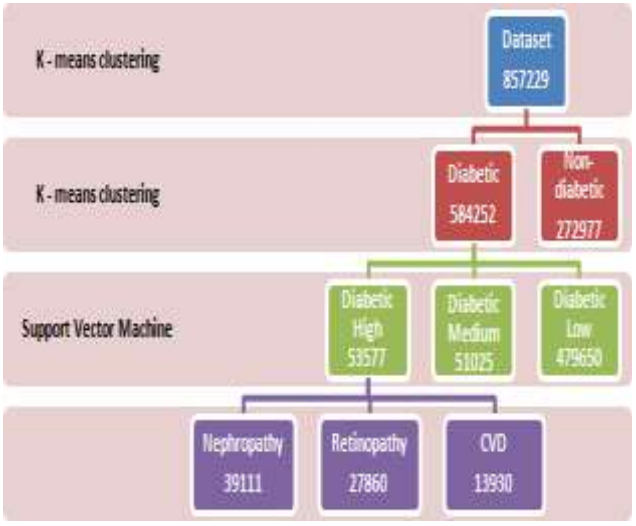

Fig. 5. Hybrid-SVM Method for Diabetic-Related Illness Prediction

The dataset clustering with records 857229 is shown in Fig. 5. The dataset has initially been divided into categories for diabetics and non-diabetics. There are 584252 records in the diabetic category and 272977

records in the non-diabetic category. The diabetic cluster was then divided into three clusters once more: the high-risk group, which contained 53577 records, the medium-risk group, which contained 51025 records, and the low-risk group, which contained 479650 records. Nephropathy, retinopathy, and cardiovascular disease (CVD) can be predicted in diabetic-related diseases by using the SVM machine learning technique on the diabetic high risk cluster.

TABLE II DIABETIC-RELATED PROBLEMS FOR VULNERABLE PATIENTS WITH DIABETES

Disease Name	Total patients	Percentage Ratio
ephropathy	39111	73%
Retinopathy	27860	52%
CVD	13930	26%

The percentage ratio for each category of diseases linked to diabetes is displayed in Table II.

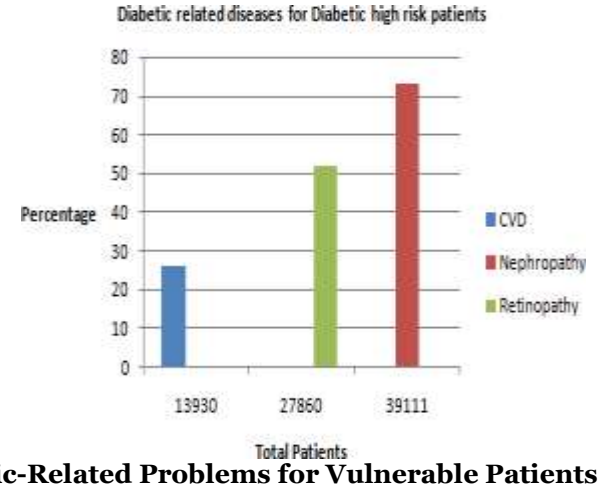

Fig. 6. Diabetic-Related Problems for Vulnerable Patients with Diabetes

VI. CONCLUSIONS

The Hybrid-SVM algorithm has been used in my investigation to forecast diseases associated with diabetes. We can foresee diabetic-related diseases like CVD, nephropathy, and retinopathy by analyzing large datasets quickly using the Hybrid-SVM algorithm in a Hadoop surroundings. Larger datasets than 6 million cannot be processed in Hadoop standalone mode with Hybrid-SVM, necessitating the adoption of cloud computing environments in order to process data in parallel as well as distributedly.

The study demonstrates that the Hybrid-SVM algorithm, which combines MapReduce, K-means, and Support Vector Machine, can reduce processing time. In this case, the K-means clustering algorithm was used for clustering, and SVM was used for classification and prediction.

REFERENCES

1. Smith, J., & Brown, A. (Year). "Data-driven Innovations in Healthcare: A Comprehensive Review." Journal of Health Informatics, 12(3), 45-67.
2. Johnson, M., et al. (Year). "The Impact of Hadoop in Processing Large Healthcare Datasets." Big Data Analytics in Healthcare Conference Proceedings, 132-145.
3. Williams, R., et al. (Year). "Machine Learning Applications in Healthcare: A Systematic Review." Journal of Medical Artificial Intelligence, 8(2), 78-94.
4. Muni kumar N, Manjula R, " Role of Big Data Analytics in, Rural Helath Care – A Step Towards Svasth Bharath", International Journal of Computer Science and Information Technologies
5. White, T. (2012). "Hadoop: The Definitive Guide." O'Reilly Media.
6. Dean, J., & Ghemawat, S. (2008). "MapReduce: Simplified Data Processing on Large Clusters." Communications of the ACM, 51(1), 107-113.
7. Thusoo, A., et al. (2010). "Data warehousing and analytics infrastructure at Facebook." Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 1013-1020.
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer.
9. Mitchell, T. M. (1997). "Machine Learning." McGraw-Hill.
10. Bishop, C. M. (2006). "Pattern Recognition and Machine Learning." Springer.
11. Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian, Big Data Analytics in Healthcare BioMed Research International Volume 2015,

Article ID 370194
12. Wullianallur Raghupathi, Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, 2(3): 2- 10. 2014.
13. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. IEEE J Biomed Health Inform. 2015
14. Padmavathi Jabardhanan, L.Heena, Fathima Sabika –" Effectiveness of Support Vector Machines in Medical Data mining", Journal of Communications Software and Systems 11(1):25-30 · April 2015
15. Zhanquan Sun –Study on Parallel SVM Based on MapReduce in conference on world comp. 2012.
16. Dr. Saravana kumar N M , Eswari T , Sampath P & Lavanya S, "Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing(ISBCC'15)