

Prediction of new prescription requirements for diabetes patients using big data technologies

334

Received 13 May 2020
Revised 16 July 2020
17 August 2020
Accepted 25 August 2020

Batuhan Bakırarar

Biostatistics, Ankara University School of Medicine, Ankara, Turkey

Cemil Yüksel

Surgical Oncology,

University of Health Sciences, Ankara Oncology Training and Research Hospital, Ankara, Turkey, and

Yasemin Yavuz

Biostatistics, Ankara University School of Medicine, Ankara, Turkey

Abstract

Purpose – The study aimed to evaluate the effectiveness of using large data sets for new diabetes patient prescriptions.

Design/methodology/approach – This study consisted of 101,766 individuals, who had applied to the hospital with a diabetes diagnosis and were hospitalized for 1–14 days and subjected to laboratory tests and medication.

Findings – With the help of Mahout and Scala, data mining methods of random forest and multilayer perceptron were used. Accuracy rates of these methods were found to be 0.879 and 0.849 for Mahout and 0.870 for Scala.

Originality/value – The mahout random forest method provided a better prediction of new prescription requirements than the other methods according to accuracy criteria.

Keywords Big data, Classification, Data mining, Diabetes mellitus

Paper type Research paper

Introduction

Diabetes mellitus (DM) is a complex and metabolic chronic disease associated with a state of high blood glucose level or hyperglycemia, occurring from deficiencies in insulin secretion, action or both. Autoimmune destruction of the pancreatic gland, insulin secretion anomalies and insulin resistance play a role in the development of diabetes. Diabetes is divided into two categories: type 1 diabetes is caused by insulin secretion disorder, and type 2 diabetes is the development of insulin resistance. The chronic metabolic imbalance associated with this disease puts patients at high risk for long-term macro and microvascular complications and dysfunction of some organs which, if not provided with high-quality care, lead to frequent hospitalization and complications, including elevated risk for cardiovascular and renal diseases (CVDs) [1]. In addition, ketotic and non-ketotic coma may develop.



The International Diabetes Federation estimates that there are approximately 387 million people diagnosed with diabetes across the globe with two-thirds of them being adults aged 20-65 years and the proportion of deaths before 60 years ranging from 36% to 73% [2]. In general, 1.4 million newly diagnosed cases in the US are reported every year. If this trend continues, it is projected that in 2050, one in three Americans will have diabetes. Diabetes, with its associated side effects, remains the seventh leading cause of mortality in the United States [3]. In addition, epidemiological studies report that diabetes causes more deaths in Americans every year compared to breast cancer and acquired immunodeficiency syndrome (AIDS) combined [3]. The most important criterion in the treatment of diabetes is glycemic control. Drugs are being evaluated to manage DM including oral GLP-1 analogs (Glucagon-like peptide), glucokinase activators, glucagon receptor antibodies, metformin, sodium-glucose co-transporter-2 (SGLT-2) inhibitors. The purpose is to keep the HbA1c below 7% [4]. However, advances in diabetes treatment continue. New treatment regimens are especially emerging to reduce cardiovascular mortality and renal transplant need. Therefore, in order to ensure good metabolic control in diabetic patients and to ensure improved health and longevity, a combination of changes in lifestyle, pharmacological treatment and prescription change should be used (in cases where the treatment cannot be benefited) [5]. Countries with the highest prevalence of diabetes account for 60% of the world's population, so researching new effective treatment regimens is essential [6].

At least 95% of clinical data recorded in the healthcare sector being in video format indicate the importance of multimedia data within big data. Big data technologies have started to be used for it is not possible to store and analyze all these data with standard database solutions and classical statistical methods. Big data is defined as “how businesses, states, hospitals, and organizations integrate datasets which are mostly not structured and continue to accumulate endlessly, are away from structurality to the extent that they cannot be analyzed with traditional association-based database techniques and are very big, raw and growing exponentially, and explore information that has remained hidden and surprise correlations through methods of statistics and data mining” [7–9]. Today, data are still kept in different formats in hospital systems, and it is almost impossible to collect this data in the same format for multi-center studies. Therefore, there are no studies conducted with big data in the literature apart from a few articles. Mahout and Scala used in the study are software packages like SPSS (Statistical Package for the Social Sciences) that enables analysis of big data. The two software packages were analyzed for their performance on the same data that was compared, and the results they provided for data mining methods were presented. This study aimed to find the factors affecting the variable “New Prescription” and to determine whether patients need medication change using the big data technologies of Mahout and Scala.

Methodology

Datasets

Health Facts is a database that records details of hospital data in the USA including electronic medical records, demographics, hospital procedures, laboratory findings, pharmacy data and hospital death rates. Diabetes is a disease affected by genetic and environmental factors. There are more than 100 genetic differences in diabetes [10]. This gives us a chance to customize the treatment. However, it may be difficult to apply a fixed treatment given that there are so many factors. Different responses to antidiabetics have increased the importance of pharmacogenetics. The dataset used in our study was retrieved from the data in the Health Fact database (Cerner Corporation, Kansas City, MO). This dataset involved hospitals in the Central (18 hospitals), Northeast (58 hospitals), Southern (28 hospitals) and Western (16 hospitals) regions of the USA between 1999 and 2008. There were 50 variables regarding

patient and hospital results in the dataset. Approval was obtained to use the data set [11, 12]. Out of this dataset, 101,766 individuals who had applied to the hospital with a diagnosis of diabetes, were hospitalized for 1 to 14 days and were subjected to laboratory tests and medication were included in our study. Hence, variables that had high missing data percentages had some categories with very few data and were deemed to have no effect on the dependent variable were excluded from the study, and 10 out of 49 independent variables were included in the study [11]. It must be noted that even when data sets with the same features and data are taken from two different centers, their characteristics differed. These differences were also revealed in the statistical tests to be performed (For example, the average age was different, the gender ratio was different). For this reason, it was a more correct approach to use different data mining methods and/or different data mining software (packages) for the data set used. We aimed to do this in our study and presented results in a table format. The scope of the research relied on the available data.

Big data technologies

Through the latest technologies, big data provides the opportunity to analyze the data types which are impossible to be analyzed with standard methods such as text, audio and video analyses [13, 14].

Hadoop. Hadoop is an open-coded library that was developed in Java and runs the applications necessary for processing and analyzing big data on the set formed by multiple servers. Hadoop is composed of Hadoop Distributed File System (HDFS) and MapReduce.

HDFS combines the disks of multiple servers to use them as a single virtual disk for storing a huge amount of data that cannot be stored in one server [15–17]. MapReduce is used for processing the large-scale data stored on HDFS. It is composed of the Map function developed to filter data and the Reduce function used for having outputs from data [18].

Machine learning. Machine Learning is also called automatic modeling and tests the data with several models to achieve the best fit possible. The velocity and volume of big data technologies make use of machine learning importance [19, 20].

Machine learning libraries. The most commonly used machine learning algorithms are Mahout and Scala. The Mahout algorithm has features such as data preparation, modeling and accessing information via a model. It is frequently used for classification and clustering. The most used classification algorithms in Mahout are Logistic Regression, Naïve Bayes and random forest and the most used clustering algorithms are k-means, Canopy and MinHash [19, 20].

Scala is also regarded as a programming language as it involves object-oriented and functional programming languages. It has its own compiler, so it can compile and run Java codes easily. Since it can use all libraries and features offered by Java, it is possible to produce all projects in Java in Scala, too [19, 20].

Data pre-processing procedure

Variables were evaluated by using the gain ratio, information gain and chi-squared. Attributed evaluation variable importance methods and the variables which were considered to be insignificant by the three methods and were thought as less important by clinical evaluation were excluded from the data set. Data were randomly divided into two datasets: training data (80%) and test data (20%). Following these procedures, the data were transferred to the big data technologies of Mahout and Scala, prediction of new prescriptions was predicted by using 10 independent variables with the help of random forest and multilayer perceptron algorithms. Mahout and Scala used in the study are software

applications like SPSS that enable analysis of big data. Although these software algorithms are already in the literature, their use is not common. In the study, these methods were also evaluated to compare the results. Gain ratio, information gain and chi-squared attributed evaluation methods are the methods used routinely in data mining, giving the degree of importance of the independent variables based on the result variable.

We used multilayer perception as a neural network algorithm. In fact, the study looked at methods such as support vector machine, J48 and logistic regression. Since the two methods that give the best results are multilayer perceptron and random forest, the results of these methods were included in the study.

Ethical issue

The dataset was approved and obtained for use from the data in the Health Fact database (Cerner Corporation, Kansas City, MO), [11, 12].

Results

There were 101,766 patients in the study and 78,363 (77.0%) of these patients required a new prescription, while 23,403 (23.0%) did not require a new prescription. The majority of patients applying to the emergency department required new prescriptions (76.4%). The average length of hospital stay for patients requiring new prescriptions was 4.5 days, while patients who did not need a new prescription stayed for an average of 4.1 days. The descriptive statistics regarding the explanatory variables which formed the dataset of the study on the level of the dependent variable are given in [Tables 1 and 2](#).

Variables	New prescription		
	No (n = 23403)	Yes (n = 78363)	
Race, n (%)	Caucasian	18052 (23.0)	60322 (77.0)
	Afro-American	4413 (23.0)	14799 (77.0)
	Asian	289 (19.2)	1219 (80.8)
	Hispanic	167 (26.0)	476 (74.0)
	Others	487 (23.9)	1552 (76.1)
Gender, n (%)	Female	12922 (23.6)	41788 (76.4)
	Male	10482 (22.3)	36575 (77.7)
Age, n (%)	0-9	29 (17.8)	134 (82.2)
	10-19	92 (13.3)	601 (86.7)
	20-29	343 (20.7)	1316 (79.3)
	30-39	927 (24.5)	2850 (75.5)
	40-49	2281 (23.5)	7406 (76.5)
	50-59	3856 (22.3)	13402 (77.7)
	60-69	4873 (21.7)	17612 (78.3)
	70-79	5878 (22.5)	20192 (77.5)
	80-89	4284 (24.9)	12915 (75.1)
Admission type, n (%)	90-100	850 (30.4)	1945 (69.6)
	Emergency	12725 (23.6)	41267 (76.4)
	Highly emergency	3958 (21.4)	14524 (78.6)
	Upon Patient's request	4263 (22.6)	14608 (77.4)
	Newborn	3 (25.0)	9 (75.0)
	Trauma	7 (30.4)	16 (69.6)
Readmitted, n (%)	Others	49 (15.2)	273 (84.8)
	<30	2247 (19.8)	9112 (80.2)
	≥30	7228 (20.3)	28319 (79.7)
	No	13931 (25.4)	40935 (74.6)

Table 1. Descriptive statistics of qualitative variables in the dataset

Mahout random forest

In the first stage of the Mahout random forest method, the training data was run in the Mahout environment and was divided into several numbers to experiment with until the fittest model was created. The Map procedure was performed to filter the data and create key/value pairs and Reduce was performed to reduce the data.

The time taken for the procedures of map and reduce to be completed was approximately nine minutes (541,348 ms) in the training model. As for the comparison between the numbers of read and written bytes, the reading from the file was much lower than the numbers during processing on HDFS (FILE Number of bytes read = 3,369, HDFS Number of bytes read = 5,624,479) (Figure 1).

While the model creation required 5,624,095 bytes of data during reading, its processing and writing stage required 4,440,974 bytes of data. Even in this stage, an advantage of about 25% was achieved with the reduction process. The training model was created in 3 minutes 42 seconds. This process would have taken about 30 minutes if it could have been distributed to and processed on five computers with standard hardware used today. The Mahout technology selects the numbers of nod and depth for the decision tree in the random forest method in such a way that they provide the most ideal outcome with the least processing

Table 2.
Descriptive statistics of quantitative variables in the dataset

Variables	New prescription			
	No (n = 23403)		Yes (n = 78363)	
	Mean ± SD	Median (Min-Max)	Mean ± SD	Median (Min-Max)
Time in hospital (day)	4.1 ± 2.9	3.0 (1.0–14.0)	4.5 ± 3.0	4.0 (1.0–14.0)
Number of laboratory procedures	41.9 ± 19.1	43.0 (1.0–129.0)	43.5 ± 19.8	45.0 (1.0–132.0)
Number of procedures	1.4 ± 1.7	1.0 (0.0–6.0)	1.3 ± 1.7	1.0 (0.0–6.0)
Number of medication	13.2 ± 7.0	12.0 (1.0–69.0)	16.8 ± 8.2	15.0 (1.0–81.0)
Number of diagnosis	7.3 ± 1.9	8.0 (1.0–16.0)	7.4 ± 1.9	9.0 (1.0–16.0)

Note(s): SD: standard deviation, Min: minimum, Max: maximum

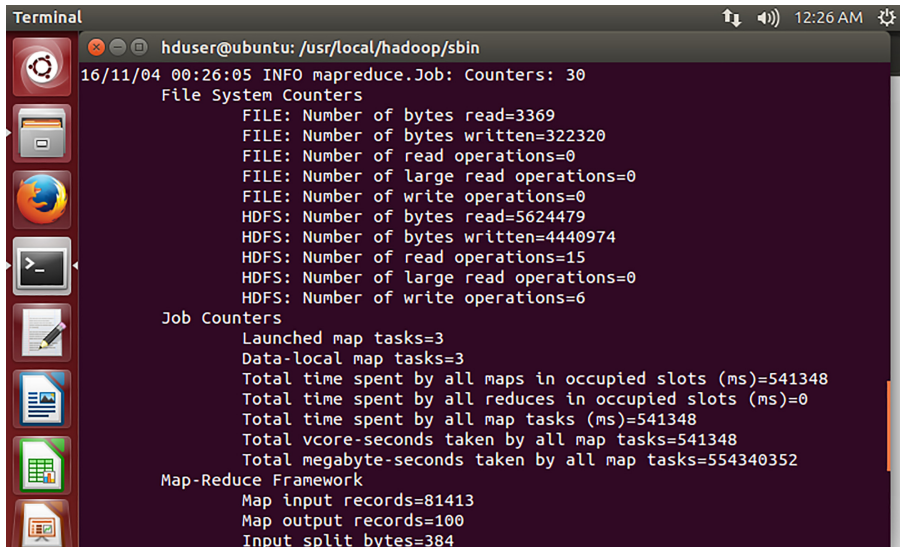


Figure 1.
Read, write, byte amounts and processing times when creating Mahout random forest training model

power (hardware use) in the least time and creates the model according to that number of nodes and depths.

In our model, the optimum number of nodes (Forest num Nodes) was found 269,245, mean number of nodes (Forest mean num Nodes) 2,692 and depth (Forest mean max Depth) 26 (Figure 2).

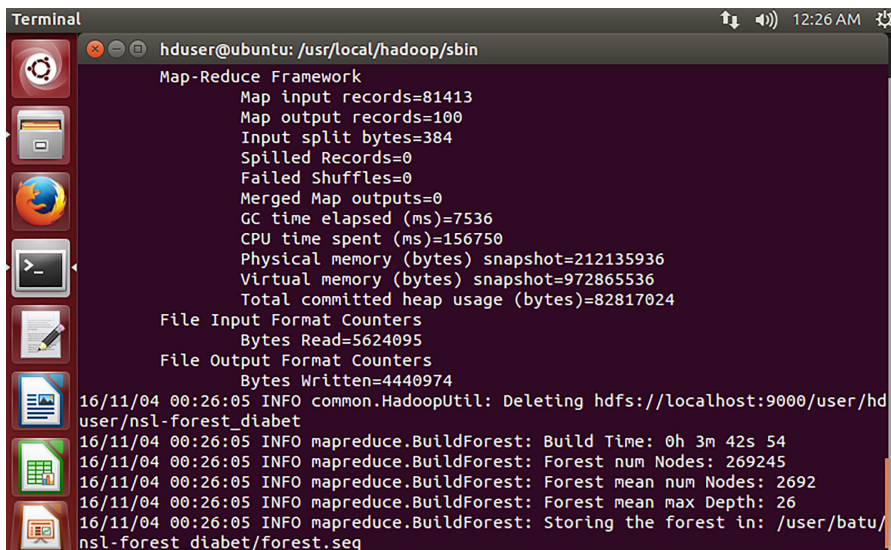
The time passed for the procedures of map and reduce performed with the entire hardware was approximately nine seconds (8724 ms) when creating the Mahout random forest test model. As for the comparison between the numbers of read and written bytes, the numbers of read and written bytes during reading from the file were higher than the numbers during processing on HDFS (FILE Number of bytes read = 4,472,619, HDFS Number of bytes read = 1,396,768). The reason is that the number of processes on HDFS decreased because the model was created during the training stage.

For the test model, data read required 1,396,642 bytes, and write required 4,112,43 bytes in the Mahout random forest method. The fact that the number of bytes required for data write decreased by one-third proves the success of Mahout technology in the reduction process. This significantly shortens the time spent on test data outcomes.

In the last stage, accuracy and F -measure values were obtained through the test data with the model created using the training data with the help of the Mahout random forest. The accuracy and F -measure value were found to be 0.879 and 0.662.

One of the tree diagrams of the random forest is shown in Figure 3. The accuracy value was calculated as 0.872 and the F -measure value was calculated as 0.659. These values, calculated from the single tree structure were close to the values calculated by random forest.

By using this tree structure it can be concluded that when the number of medications is greater than eight, the prescription change is probably needed. When the number of medications is less than or equal to eight and the Number of diagnoses is greater than five and Readmitted is greater than or equal to thirty, the prescription change is probably needed and so on (Figure 3).



```

Terminal
hduser@ubuntu: /usr/local/hadoop/sbin
Map-Reduce Framework
  Map input records=81413
  Map output records=100
  Input split bytes=384
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=7536
  CPU time spent (ms)=156750
  Physical memory (bytes) snapshot=212135936
  Virtual memory (bytes) snapshot=972865536
  Total committed heap usage (bytes)=82817024
File Input Format Counters
  Bytes Read=5624095
File Output Format Counters
  Bytes Written=4440974
16/11/04 00:26:05 INFO common.HadoopUtil: Deleting hdfs://localhost:9000/user/hd
user/nsl-forest_diabet
16/11/04 00:26:05 INFO mapreduce.BuildForest: Build Time: 0h 3m 42s 54
16/11/04 00:26:05 INFO mapreduce.BuildForest: Forest num Nodes: 269245
16/11/04 00:26:05 INFO mapreduce.BuildForest: Forest mean num Nodes: 2692
16/11/04 00:26:05 INFO mapreduce.BuildForest: Forest mean max Depth: 26
16/11/04 00:26:05 INFO mapreduce.BuildForest: Storing the forest in: /user/batu/
nsl-forest_diabet/forest.seq

```

Figure 2. Time spent, byte quantity, node and depth count of the created tree for the Mahout random forest training model

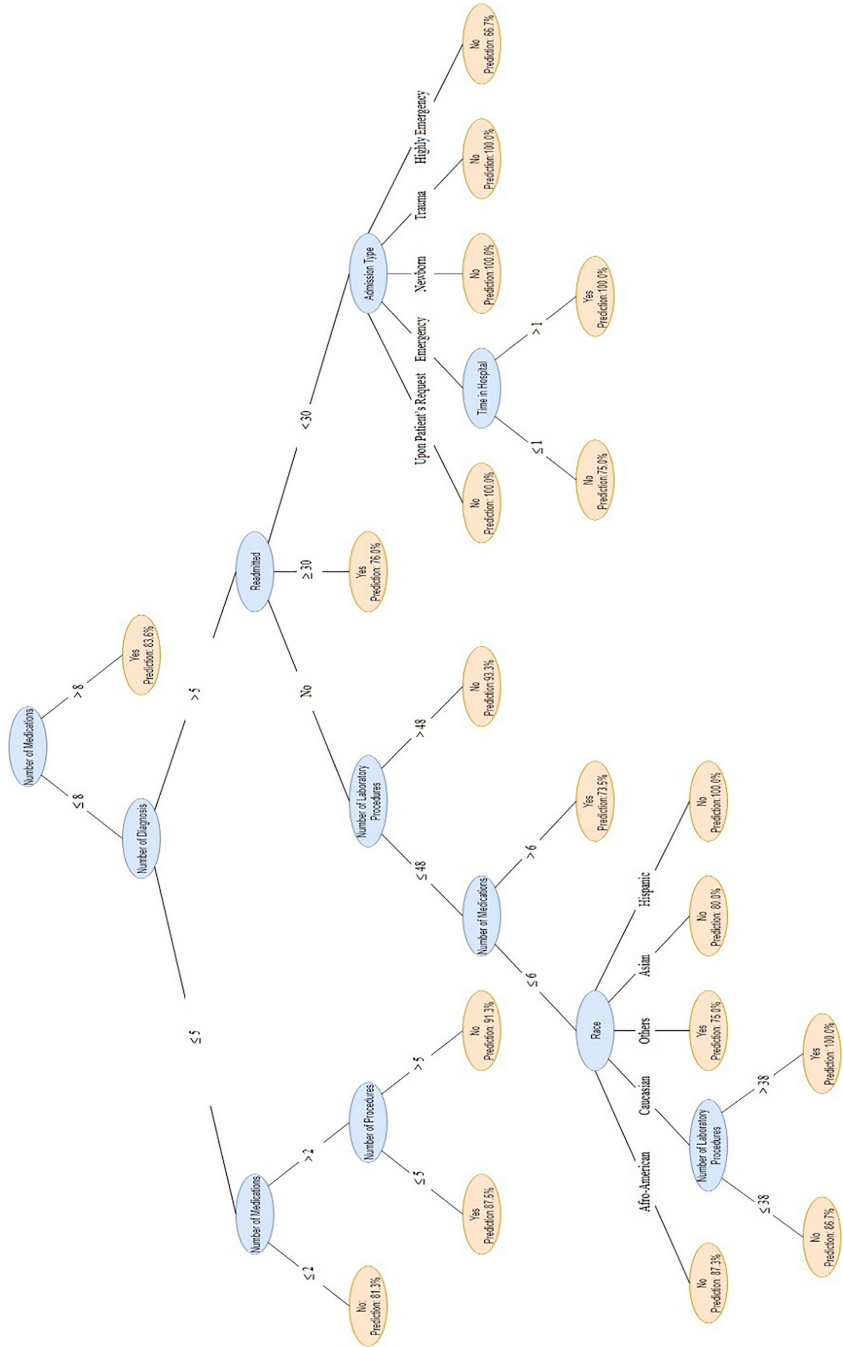


Figure 3.
One of the tree
diagrams of the
random forest model

Scala random forest

Parameters are provided manually within the code in the Scala method. Hence, outcomes of the random forest model were calculated using the parameters created for Mahout. In the test model, the Accuracy and *F*-measure value were found to be 0.849 and 0.604.

Mahout multilayer perceptron

In the multilayer perceptron analysis conducted with the Mahout technology, the scenarios were similar to those in the random forest and the fittest parameters were selected automatically. In the test model, the Accuracy and *F*-measure value were found to be 0.849 and 0.580, respectively.

Scala multilayer perceptron

Parameters are provided manually within the code in the Scala method. Hence, outcomes of the multilayer perceptron model were calculated using the parameters created for Mahout. The Accuracy and *F*-Measure value were found to be 0.870 and 0.709 in the test model.

Performance comparison of classification methods by big data technologies

Accuracy and *F*-measure values for the random forest and multilayer perceptron methods using the Scala and Mahout technologies are given in Table 3. Accuracy of the Mahout and Scala technologies by the random forest and multilayer perceptron methods were found to be 0.879 and 0.849 and 0.849 and 0.870, respectively. According to these rates, the accuracy of random forest using the Mahout technology was higher than others.

Discussion

There are many studies on data mining methods for the prediction of diabetes disease. Malik *et al.* [21] used linear data logistic regression, ANN, linear support vector machine and radial basis function support vector machine methods in their study to diagnose and predict diabetes disease with 175 (87 healthy and 88 type 2 diabetes) people. The accuracy values of these methods were found as 75.86, 80.70, 77.93 and 84.09 respectively. Farran *et al.* [22] used logistic regression, support vector machine, k nearest neighbors and multifactor dimension reduction methods for predicting diabetes in the study they planned with 10,632 patients. The accuracy values of these methods were 80.7, 81.3, 78.6 and 78.3 respectively. Tapak *et al.*, [23] in their study with 6,500 patients, used logistic regression, linear discriminant analysis, fuzzy c-mean, support vector machine, neural network and random forest methods and found accuracy values for these methods as 0.935, 0.925, 0.859, 0.986, 0.931 and 0.930.

In Rajesh *et al.*'s study [24] of 768 patients, they used many data mining algorithms and provided the best results from RMD Tree. However, due to the over-fitting of data problem in this method, they preferred C 4.5 which was the second method with the accuracy value of 0.910 and giving the best result Meng *et al.* [25], in their study of 1,487 (735 diabetes or prediabetes patients and 752 control) people used logistic regression, artificial neural network

		Random forest	Multilayer perceptron
Mahout	Accuracy	0.879	0.849
	<i>F</i> -measure	0.662	0.580
Scala	Accuracy	0.849	0.870
	<i>F</i> -measure	0.604	0.709

Table 3.
Performance comparison of mahout and Scala technologies by classification methods

and decision tree methods for the prediction of diabetes or prediabetic patients. They found the accuracy values of these methods as 0.761, 0.822 and 0.807.

El-Sappagh *et al.* [26] implemented a framework and tested the accuracy of this system in their study with 60 patients diagnosed with diabetes. The accuracy value of the system was found as 0.977. They compared their framework with existing CBR systems and a set of five machine-learning classifiers and their system outperformed all of these methods. Deep learning is a method that has become recently popular. It is preferred for analysis that includes time data such as image processing and survival. It gives better results in such data, but the data for modeling purposes like our study gives similar results. In our study, data mining methods were preferred because our main purpose was modeling, classification-based analysis.

The data in our study included 101,766 individuals who applied to the hospital with suspicion of diabetes. Using the data mining technologies of random forest and multilayer perceptron with the help of big data technologies, prediction of new prescription was made on these data. Accuracy of the random forest and multilayer perceptron methods using the Mahout technology were found to be 0.879 and 0.849 respectively whereas Accuracy of the random forest and multilayer perceptron methods using the Scala technology were found to be 0.849 and 0.870. It is a big data-based software program that analyzes for data mining algorithms such as Mahout Scala. For example, the logic is similar to the Student-*t*-test in both SPSS software and R programming language. For analysis, 5 SSD Cloud Servers are run simultaneously and the results are processed by means of the map and reduce functions. The use of servers with higher vertical configurations only reduces processing time and does not cause any change in performance criteria.

The results are better because the large data enables the relationships between variables, and even with the low probability of realization, it is possible to add to the training data and learn. This is one of the purposes of recommending the use of big data in studies. These results suggest that big data technologies provide good results for diabetes patients and that it is necessary to use them for new knowledge discovery about this disease.

Conclusion

Big data research will continue to increase since diabetes is a major health problem. Diabetes with hyperglycemia and many complications can be reversed with appropriate approaches. New treatment modalities should develop as diabetes and complications create high costs [27]. Large data sets should be used in the development of these treatment modalities. With the increasing size of data, big data will continue to expand in years to come, and each data scientist will have to manage more data each year. These data will be more diverse, bigger and faster and can be a potentially exciting opportunity for the future. Thus, big data will be the new frontier for scientific data research and business applications.

The use of big data provides advantages in the healthcare sector by allowing for more testing or more qualities for research; this brings about faster validity of studies and acquisition of sufficient examples for education when small amounts of examples are positively available. On the path to medical informatics, big data obtained from all levels of medical data is utilized, and the best possible ways of analyzing, deriving and answering as many medical questions as possible are being found to improve patients' health [28].

Future research may focus on using all data of patients and diseases to offer clinicians more diagnoses, treatments and ways of helping the patients. Each of these technologies can be developed and whether the results are the same in different populations can be tested using large-volume and diversified datasets. These technologies provide a short overview of opportunities that can be accessed through big data analysis in the data mining and healthcare field.

In this study, the Mahout random forest method provided a better prediction of new prescription requirements than the other methods according to accuracy criteria. Consequently, when the clinical parameters (risk factors) used in a new prescription prediction are known, a model can be created with the Mahout random forest method, and this model can be used as an alternative in the diagnosis of patients and be an assistant tool for clinicians. It is the models created as a result of the analysis using big data that can be used as an assistant tool in the diagnosis and diagnosis stages by transforming them into software or algorithms. Most hospitals do not have such ancillary software for physicians, and physicians continue diagnosis and treatment based on their experience and knowledge level.

Conflict of Interest: None

References

1. DeHart A, Richter G. Hemangioma: recent advances. *F1000Res*. 2019; 8. doi: [10.12688/f1000research.20152.1](https://doi.org/10.12688/f1000research.20152.1).
2. Type 2 diabetes statistics and facts. [cited 2020 April]. Available from: <https://www.healthline.com/health/type-2-diabetes/statistics>.
3. American Diabetes Association. Statistics about diabetes. [cited 2020 April]. Available from: www.diabetes.org/diabetes-basics/statistics/.
4. American Diabetes Association. Standards of medical care in diabetes-2019 abridged for primary care providers. *Clin. Diabetes*. 2019; 37(1): 11-34. doi: [10.2337/cd18-0105](https://doi.org/10.2337/cd18-0105).
5. Marín-Peñalver JJ, Martín-Timón I, Sevillano-Collantes C, Del Cañizo-Gómez FJ. Update on the treatment of type 2 diabetes mellitus. *World J. Diabetes*. 2016; 7(17): 354-95. doi: [10.4239/wjd.v7.i17.354](https://doi.org/10.4239/wjd.v7.i17.354).
6. Fernandez A, Quan J, Moffet H, Parker MM, Schillinger D, Karter AJ. Adherence to newly prescribed diabetes medications among insured Latino and white patients with diabetes. *JAMA Intern Med*. 2017; 177(3): 371-9. doi: [10.1001/jamainternmed.2016.8653](https://doi.org/10.1001/jamainternmed.2016.8653).
7. Vinod B. Leveraging BIG DATA for competitive advantage in travel. *J Revenue Pricing Manage*. 2013; 12(1): 96-100. doi: [10.1057/rpm.2012.46](https://doi.org/10.1057/rpm.2012.46).
8. Rubinstein IS. Big data: the end of privacy or a new beginning?. *Int Data Priv Law*. 2013; 3(2): 74-87.
9. Altunışık R. Büyük Veri: Fırsatlar Kaynağı mı yoksa yeni sorunlar yumağı mı?. *Yildiz Social Science Review*. 2015; 1(1): 45-76. [in Turkish].
10. Prasad RB, Groop L. Genetics of type 2 diabetes-pitfalls and possibilities. *Genes*. 2015; 6(1): 87-123. doi: [10.3390/genes6010087](https://doi.org/10.3390/genes6010087).
11. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, *et al*. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *Biomed Res Int*. 2014; 2014: 781670. doi: [10.1155/2014/781670](https://doi.org/10.1155/2014/781670).
12. Machine Learning Repository. Diabetes 130-US hospitals for years 1999-2008 data set. [cited 2020 April]. Available from: <https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008>.
13. Pentreath N. *Machine learning with spark*. Birmingham: Packt; 2015.
14. Karau H, Konwinski A, Wendell P, Zaharia M. *Learning spark : lightning-fast big data analysis*. Beijing: O'Reilly Media; 2015.
15. Lohr S. The age of big data. *The New York Times* [Internet]. 2012 Feb 11. [cited 2020 April]. Available from: <https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all>.
16. Gobble MM. Big data: the next big thing in innovation. *Res-TechManag*. 2013; 56(1): 64-6. doi: [10.5437/08956308x5601005](https://doi.org/10.5437/08956308x5601005).

17. Court D. Getting big impact from big data. *McKinsey Q.* 2015; 1(1): 52-60.
18. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. In: *Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation*, 2004 Dec 6-8, San Francisco, CA. San Francisco, CA. USENIX Association. 2004; 6: 10.
19. Tiwary C. *Learning Apache mahout*. Birmingham: Packt; 2015.
20. Giacomelli P. *Apache mahout cookbook : a fast, fresh, developer-oriented dive into the world of Apache mahout*. Birmingham: Packt; 2013.
21. Malik S, Khadgawat R, Anand S, Gupta S. Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. *Springerplus.* 2016; 5(1): 701. doi: [10.1186/s40064-016-2339-6](https://doi.org/10.1186/s40064-016-2339-6).
22. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ Open.* 2013; 3(5): e002457. doi: [10.1136/bmjopen-2012-002457](https://doi.org/10.1136/bmjopen-2012-002457).
23. Tapak L, Mahjub H, Hamidi O, Poorolajal J. Real-data comparison of data mining methods in prediction of diabetes in Iran. *Healthc Inform Res.* 2013; 19(3): 177-85. doi: [10.4258/hir.2013.19.3.177](https://doi.org/10.4258/hir.2013.19.3.177).
24. Rajesh K, Sangeetha V. Application of data mining methods and techniques for diabetes diagnosis. *Int J Eng Innov Technol.* 2012; 2(3): 224-9.
25. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung J Med Sci.* 2013; 29(2): 93-9. doi: [10.1016/j.kjms.2012.08.016](https://doi.org/10.1016/j.kjms.2012.08.016).
26. El-Sappagh S, Elmogy M, Riad AM. A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis. *Artif Intell Med.* 2015; 65(3): 179-208. doi: [10.1016/j.artmed.2015.08.003](https://doi.org/10.1016/j.artmed.2015.08.003).
27. Polonsky WH, Henry RR. Poor medication adherence in type 2 diabetes: recognizing the scope of the problem and its key contributors. *Patient Pre. Adherence.* 2016; 10: 1299-307. doi: [10.2147/PPA.S106821](https://doi.org/10.2147/PPA.S106821).
28. Mahabub A. A robust voting approach for diabetes prediction using traditional machine learning techniques. *SN Appl Sci.* 2019; 1(12): 1667. doi: [10.1007/s42452-019-1759-7](https://doi.org/10.1007/s42452-019-1759-7).

Corresponding author

Cemil Yüksel can be contacted at: cemil8537@hotmail.com