2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech) | 978-1-7281-9782-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/iCareTech49914.2020.00030

Generalization of Deep Neural Network of Hospital Readmission Prediction Models for Diabetes Patients Using Apache Spark Clustering

Fatma Al-Rubaei
Computer Engineering Department, Smart Pace
Gaza, Palestine
fatma1211994@hotmail.com

Mohammed Alhanjouri
Computer Engineering Department
The Islamic University of Gaza
Gaza, Palestine
mhanjouri@iugaza.edu.ps

Abstract—**The high readmission rate, which is the percentage of patients who admitted to the hospital within a specific period after they have been discharged, is a major concern for many healthcare organizations. Reducing it would help significantly improving healthcare services by lowering the pressure on reception, focusing the resources on cases that need special care which could help to save their lives, being cost-efficient, and finally, it would provide a better life quality for the patients. Reducing the readmission rate could be done by studying the relation between the readmission rate and the factors of the patients such as age, race, number of diagnoses, and others. Many studies have studied this relation using different machine learning techniques such as unsupervised learning, which includes clustering algorithms such as k-means and supervised learning which includes regression and classification algorithms such as KNN, decision trees, and neural networks. This research uses a dataset of diabetic patients, which was gathered from 130 different US hospitals for years 1999-2008 with more than 100,000 instances and 55 different attributes to study the relationship between the readmission rate and other factors of the patients to determine the most influential factors that lead to a higher readmission rate and determine the most important factors that help to reduce readmission rate. Our novel solution consists of two stages, the first stage is developing predictive models for predicting readmission rates accurately, then use those models to identify the critical risk factors. As we experiment with different machine learning models, we report an accuracy of 53%, 35.7%, 35%, 11.6%, 50% on a held-out test set for KNN, Decision Trees, Random Forest, Support Vector, and deep neural networks respectively, and using ablation study we identified the top ten influential risk factors. As for the second stage, we reduced the computational time for training the machine learning models using the Spark cluster-computing framework to distribute the dataset across 4 workers to speed up the training process, and we report a 32% time reduction compared with running the models without Spark.**

Keywords— Data; Diabetic Patient; Spark Cluster; Readmission Rate

I. Introduction

The field of Artificial intelligence for medical diagnosis is growing at an exponential rate, covering many technical areas such as image processing, classification, telecommunication and human-computer interaction. Readmission rate became recently a significant area of interest for healthcare officials and stakeholders. The readmission rate can be defined as the percentage of patients who are admitted to the hospital within a specific period after they have been discharged. As the readmission rate increases exponentially, health care reform in the US has made the readmission rate a leading factor for improving healthcare services and for potential savings [1]. Identifying the conditions that are mostly associated with readmission cases could significantly decrease their rate of occurrence. In addition, it could potentially help stakeholders to focus their resources on those conditions which could help reducing the readmission rate in healthcare facilities, that would become more cost-efficient and provide more savings.

As the field of big data grows, stakeholders became more and more interested in seeking data-driven solutions. A statistical study, conducted by the Healthcare Research and Quality (AHRQ), on the readmission rate of patients within 30 days of the first discharge [2] found that in 2011, about 3.3 million readmission cases were recorded for adults in the US hospitals, with estimated associated costs of 41.3 billion US dollars. Furthermore, the study found that the three conditions with the highest readmission rate among Medicare patients (is an insurance program) were congestive heart failure, septicemia, and pneumonia. These conditions resulted in about $4.3 billion in hospital costs. In addition, the study found that the three conditions with the highest readmission rate among Medicaid patients were mood disorders, schizophrenia, and diabetes, with estimated associated costs of 839 million US dollars in hospital costs.

Readmission is fairly common and costs hospitals in the US billions of dollars each year. However, it is preventable and many studies have been conducted to provide many different solutions starting from random guessing to more sophisticated solutions. For instance, a study in [3] introduced an objective method called the LACE index, where the letter (L) stands for the length of stay; the letter (A) stands for the acuity of the admission; the letter (C) stands for comorbidity of the patient (measured with the Charlson comorbidity index score); and the letter (E) stands for emergency department use (measured as the number of visits in the six months before admission). According to the study authors, the LACE index can be used to quantify the risk of death or unplanned readmission within 30 days after discharge from the hospital. Other studies used machine learning models to predict risk of readmission [4, 5]. Such data-driven models are more complex but they are significantly more accurate in predicting the risk.

978-1-7281-9782-1/20/$31.00 ©2020 IEEE
DOI 10.1109/iCareTech49914.2020.00030

In this research, we provide a novel data-driven solution using deep neural networks to predict the risk of readmission and determining the top 10 factors that are associated the most with increasing the readmission rate. In addition, we used a computer-clustering technique to improve time efficiency.

II. Related Works

Several previous studies analyzed risk factors that predict diabetic patients' readmission rates. Bhuvan et el. [1] identified the high-risk factors of readmission using machine learning methods to build a system for identifying diabetic patients facing high risk of future readmission. Strong predictors for readmission were identified as, number of inpatient visits, admission type and discharge disposition. This study was useful for healthcare providers to improve inpatient diabetic care. The study also included an analysis for cost suggesting that $252.76 million can be saved across 98,053 diabetic patients.

Another statistical study [2] focused on the impact of HbA1c measurement on hospital readmission rates by analyzing about 70,000 clinical databases of patient records. This study presented a statistical model that suggests that the relationship between readmission rate probability and HbA1c measurement depends primary on diagnosis. In addition, logistic regression was used to fit the relations between them. Results displayed that the measurement of HbA1c is a useful predictor of the readmission rate, which will contribute to reducing the rate of hospital readmissions and consequent costs of patient healthcare.

Another study [3] presented how diabetes develops and how to manage it for children aged less than 15 years and consequently evaluate the outcomes of those children. The authors focused on the proportion of children who are managed and kept out of hospitals, and comparing readmission rates and concentrations of hemoglobin in children who have been admitted to hospitals and who have not.

K. Saravananathan and T. Velmurugan [4] carried out a study to analyze diabetic data using Classification Algorithms in Data Mining. They used different classification techniques; J48, CART, SVMs, and kNN, on a medical dataset to find the optimal solution for diabetes readmission. Accuracy, specificity, sensitivity, precision, and error rate were calculated for the given dataset as performance indicators.

Mingle conducted a study [5] to predicting diabetic patients' readmission rates, the study suggests that applying a machine learning approach to a larger feature set can improve on readmission rate models potentially, improve patient outcomes and lower inpatient healthcare costs on hospitals. This study targets diabetic patients only to improve the accuracy of readmission risk for healthcare conditions. Results showed that while healthcare providers may make the decision not to obtain HbA1c for diabetics during hospital stay. There are other useful factors for predicting readmission rates, that may be valuable in developing strategies to reduce readmission rates and the costs associated with caring for these individuals. The proposed machine resulted in a 26% improved learning approach using more than 100 thousand patients from 130 hospitals in the United States over a year period compared with LACE, which was derived from 4800 patients over a 4 years period.

Drincic et al. [6] developed a diabetes resource nurse program as a novel model of inpatient diabetes care, which aimed at increasing the knowledge of staff nurses, and evaluate the impact of this program on readmission rates. They analyzed discharge patients' records before 18 months and after 18 months of the intervention period. The overall 30-day readmission rate for patients with diabetes decreased significantly from 20.1% (pre) to 17.6% (post) intervention.

III. Materials and Methods

A. Dataset

The dataset used in this study was extracted from the Health Facts database which is a national data warehouse that collects clinical records across US-hospitals by Beata Strack in his study [7] and was donated to UCI machine learning repository. The dataset was collected from 10 years (1999-2008) of clinical care at 130 US hospitals [8, 9]. It has about 100,000 instances of diabetes inpatient encounters with 55 different integer attributes [10]. Each encounter in the data set belongs to an inpatient that has any kind of diabetes registered as a diagnosis and stated at least 1 day and at most 14 days in the hospital. Table 1 provides a list of the dataset features and their descriptions.

Table 1. List of Features and Their Descriptions [11].

	Feature name	Description
1	Encounter ID	A numerical attribute, where each encounter has its own identifier
2	Patient number	A numerical attribute, where each patient has a unique identifier
3	Race	A nominal attribute, patient race, e.g. Asian, Caucasian, African American
4	Gender	A nominal attribute, patient gender, e.g. female, male, or unknown
5	Age	A numerical attribute, which interval of 10-years the patient age belongs to (0-10, 10-20..., 90-100)?
6	Weight	A numerical attribute, weight of the patient measured in pounds.
7	Admission type	A nominal attribute, where an encounter can have 9 different admission types.
8	Discharge disposition	A nominal attribute, where an encounter can have 29 different types of discharge disposition.
9	Admission source	A nominal attribute, where an encounter can have 21 different types of admission sources, e.g. emergency room, physician referral, and transfer from a hospital.
10	Time in hospital	A numerical attribute, period in days between admitting the patient and discharging him/her.
11	Payer code	A nominal attribute, where an encounter can have 23 different types of payer code.
12	Medical specialty	A nominal attribute, where an admitting physician of the encounter can have 73 different types of medical specialty.
13	Number of lab procedures	A numerical attribute, number of procedures or test that was performed in the lab.
14	Number of procedures	A numerical attribute, how many procedures (excluding lab test) was taken place in the lab.
15	Number of medications	A numerical attribute, how many distinct medications given to the patient?
16	Number of outpatients	A numerical attribute, how many times has the patient had outpatient visits in the last year?

17	Number of emergency visits	A numerical attribute, how many times has the patient had emergency visits in the last year?
18	Number of inpatient visits	A numerical attribute, how many times has the patient had inpatient visits in the last year?
19	Diagnosis 1	A nominal attribute, 848 different values indicates the first diagnosis.
20	Diagnosis 2	A nominal attribute, 923 different values indicates the second diagnosis.
21	Diagnosis 3	A nominal attribute, 954 different values indicates the third diagnosis.
22	Number of diagnoses	A numerical attribute, how many diagnoses have the patient that was entered into the system?
23	Glucose serum test results	A nominal attribute, it's an indication of the range of the glucose serum test results and whether the test took place or not. E.g. normal, >300, none.
24	A1C test result	A nominal attribute, it's an indication of the range of A1c results and whether the test took place or not.
25	Change of medications	A nominal attribute, it's an indication of whether there was a change in diabetic medications.
26	Diabetes medications	A nominal attribute, it's an indication of whether there was a medication prescribed. E.g. yes, no.
27	24 features for diabetes medications	24 different features indicate whether the drug was prescribed or a modification in its dosage was taken place.
28	Readmitted	A nominal attribute, the number of days for the patient to be readmitted. The values are: "<30" if the patient readmitted within 30 days of being discharged, ">30" if the patient readmitted after 30 days of his discharge, and "No" if the patient was not readmitted.

B. Data Pre-Processing

In this section, we provide details about the pre-processing techniques that were performed on the dataset before using it to train the machine learning algorithms,

1) Dealing with missing data

A missing value is an empty (?) cell in the table of the dataset. We dealt with such missing values by implementing the following strategies:

- There were some missing data in race attribute (2%) which contains (Caucasian, African American, Asian, Hispanic, other), so we will replace all missing values with the category (other).
- In weight, Medical specialty, and payer code attributes most of the values are missing (>50%), so we ignored this attribute in processing.

2) Data transformation

Some of dataset attributes are nominal, and this will cause difficulty with data processing later, so we need to transform all nominal data to numerical as follow [12-14]:

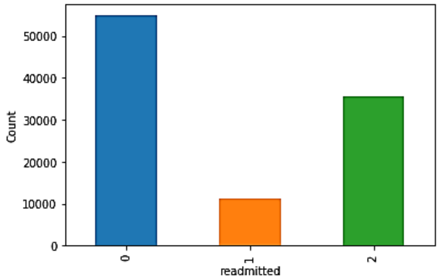
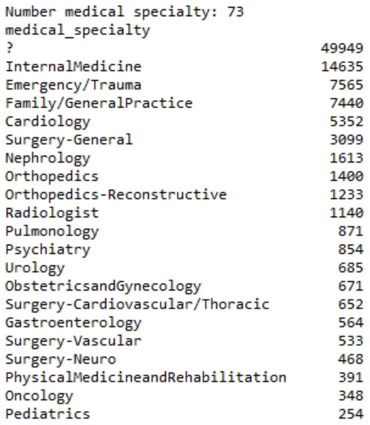


Fig. 1. Count Samples of Readmission Rate.



```
Number medical specialty: 73
medical_specialty
?                                    49949
InternalMedicine                     14635
Emergency/Trauma                      7565
Family/GeneralPractice                7440
Cardiology                            5352
Surgery-General                       3099
Nephrology                            1613
Orthopedics                           1400
Orthopedics-Reconstructive            1233
Radiologist                           1140
Pulmonology                            871
Psychiatry                             854
Urology                                685
ObstetricsandGynecology                671
Surgery-Cardiovascular/Thoracic        652
Gastroenterology                       564
Surgery-Vascular                       533
Surgery-Neuro                          468
PhysicalMedicineandRehabilitation      391
Oncology                               348
Pediatrics                             254
```

Fig. 2. High Range Variables in Medical Specialty Factor.

- The race attribute has four categories (Caucasian, African American, Asian, Hispanic, other), we will transform them into (1, 2, 3, 4, 5).
- The gender attribute contains (Male and Female, Unknown/Invalid), we will transform these into (1, 2, 3).
- The age attribute contains groups of ages: ([0-10], [10-20], [20-30], [30-40], [40-50], [50-60], [60-70], [70-80], [80-90], [90-100]). We will use data reduction to label these groups from bin 1 to 10: (1,2,3,4,5,6,7,8,9,10).
- Glucose serum test with results: (None, >200, >300, Normal), will be transformed to numeric: (0,2,3,1).
- A1C Test Result attribute which have result > 7 for greater than 8% and normal (between 7–8%) and none, will be transformed into groups (none=0, normal=1, ">7"=2, ">8"=3).
- Insulin attribute consists of: (NO, UP, DOWN, Steady), we will transform it to (0, 1, -1, 0.5).
- Change of medications attribute with (change or No-change) will be replaced with (1, 0).
- Readmission attribute with values (No, <30, >30) will be converted to numeric (0, 1, 2).

3) Data exploration

First, we analyzed the main factor in our study (Readmitted Attribute), to get the number of samples in each state of readmission to the hospital (no, <30, >30) transformed to (0, 1, 2). Fig. 1 shows the counts of readmission rates for our dataset, which shows a percentage of 11% for patients who returned to hospital during 30 days, 35% of patients returned to hospital

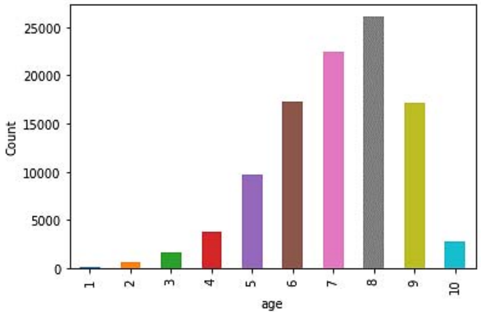


Fig. 3. Statistical Graph for Age Range Factor.

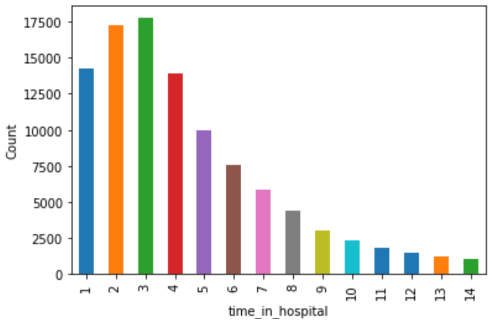


Fig. 4. Time in Hospital Rate.

after 30 days. This means that there is a total of 46% of all patients who returned to hospital within short or long periods of time. The (medical specialty) factor has 73 variables, if we analyze the high range of this variable, we get the highest 20 disease specialties that were readmitted to hospital, as presented in Fig. 2. The (Age) factor, was analyzed to obtain the most important age range that affect the readmission to hospital. Fig. 3 shows that patients aged from 60-90 years old are majority of the readmitted to hospital patients.

The (Time in hospital) factor represents the time that a patient spends in hospital for treatment. Fig. 4 shows the time in hospital rate for the dataset from 1 day to 14 days. Fig. 4 indicates that longer stays at hospitals are less frequent than short stays.

Taking into account the quality of healthcare and operational costs of patients staying in hospitals for a long time puts us in a desire to make a tradeoff between the readmission rate and healthcare budgeting.

The (Number diagnosis) factor means the number of diagnoses entered to the system for each patient. Fig. 5 views readmission encounter frequency for each number of diagnoses. A spike is noticed at 9 diagnoses count, which means that about 50,000 encounters which are about 50% of the total encounters has the same number of diagnoses. Furthermore, no number of diagnoses higher than 9 is observed as shown in Fig. 5. This indicated that a strong relationship exists between diabetic patients and patients who have 9 diagnoses, and further investigation should be carried out to provide a better understanding of this anomaly.

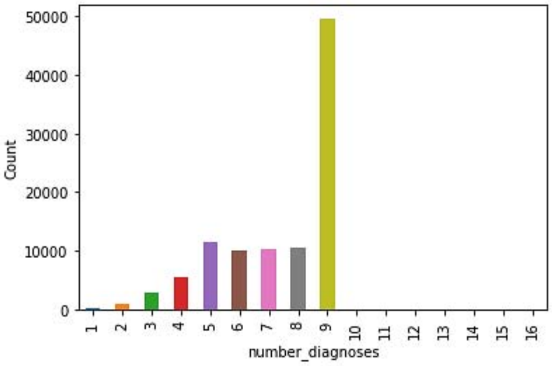


Fig. 5. Number of Diagnoses Entered to the System.

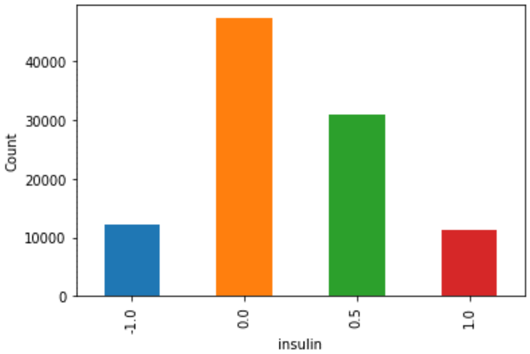


Fig. 6. Insulin Prescription Rates Analysis.

Insulin is one of the (24 features for diabetes medications) factor in dataset. This feature indicates whether the drug is prescribed or there has been a dose change. Values: “up” if the dose was increased during the encounter, “down” if the dose was reduced, “steady” if the dose did not change, and “no” if the medication was not prescribed. Fig. 6 presents the insulin prescription rates for patients with variable 1 for up, -1 for down, 0.5 for steady and 0 for no. As can be seen, the highest rate is for 0, i.e. the drug was not prescribed.

The (race) factor represented in Fig. 7 shows that the vast majority of encounters belonged to patients of the Caucasian race. This could have many indications. For instance, most of diabetic patients are indeed Caucasians, in this case, it means that the (race) factor is irrelevant to the task of predicting the readmission rate regardless of other races.

The (diabetes medications) factor is an indication of whether there was a diabetic medication prescribed (e.g. Yes, No) as illustrated in Fig. 8. It is clear that the distribution is skewed toward patients who have a diabetic medication prescribed. This could have multiple interpretations, for example, it could mean that some medications develop complications that lead to the patient being readmitted. Another explanation is to assume that medications are prescribed to patients with a bad health condition where diabetes is more developed and they need constant and special care.

So, we present some feature that have differences in values and rates, that will help us in determine the relation between this risk factors and readmission to hospital by using a group of method for statistical and classification models to compare the accuracy and the dependency of these factors.

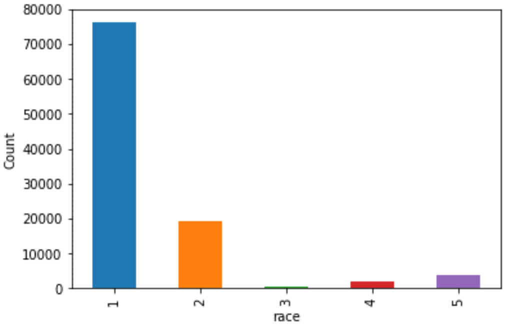


Fig. 7. Race Factor to Present Highest Race of Patients.

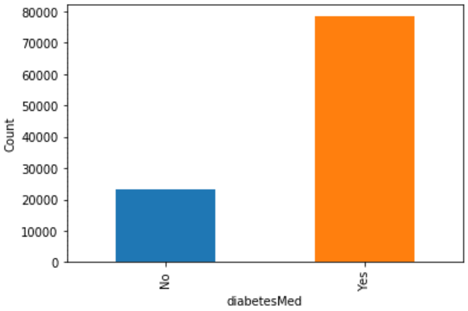


Fig. 8. Diabetes Medications Prescribed or Not.

IV. Building Models and Training

Here we report our results on the classification task. Given an encounter, we are trying to predict which of the following classes it belongs to:

- Class "0": The patient would not be readmitted.
- Class "1": The patient will be readmitted within the next 30 days of discharge.
- Class "2": The patient will be readmitted after 30 days of discharge.

The most important task is identifying the influence of risk factors that help healthcare providers to reduce the readmission rate by focusing more resources on handling those factors.

As an example, we can summarize the important features that can be used to feed classifiers as: (num_lab_ procedures), (num_medications), (time_in_hospital), (number_inpatient), (age), (insulin), (number_diagnoses), (num_procedures), (discharge_disposition_id), (admission_ type_id) [15].

Fig. 9 shows the training and testing accuracy when different number of neighbors were used to train the KNN algorithm as one of the five classifiers which were used to test selected factors as features.

Table 2 compares the performance of the used machine learning algorithms. It is summarizing accuracy in training, testing, and determine the execution time of classifiers which were used to model extracted features in previous section. We used many classification tools built in Python Language such as: K-Nearest Neighbors (KNN), Decision Trees, Random Forest, Support Vector Machines (SVMs), and Deep Neural Network (DNN).

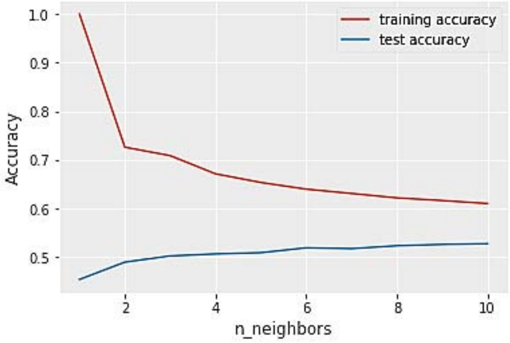


Fig. 9. Accuracy of Training and Testing in KNN.

TABLE. 2. COMPARISON OF ACCURACIES AND EXECUTION TIME FOR DIFFERENT MODELS.

Model	Accuracy Training	Accuracy Test	Execution Time
KNN	60%	53%	161.025s
Decision Trees	52.8%	35.7%	10.646s
Random Forest	98.2%	35%	125.213s
SVMs	100%	11.6%	352.5s
DNN	66%	50%	111.025s

The used features are not necessary to be risk factors, therefore, to identify the influence of individual risk factors that leads to an increase in readmission rate, we used an ablation study of the factors. An ablation study is simply deleting one factor at a time from the training process and comparing the performance with using it. If the performance decreased by a significant margin, then that is an indication of the importance of the factor that was removed. But this approach consumed a lot of time, so we need to select one classifier and improve it to extract the risk factors which indicate readmission rate.

We can notice that DNN does relatively great on both the test set and execution time, and for further reducing its execution time we will explain the use of Apache Spark in the following section.

V. Predictive Analytics Using Apache Spark

Spark is the most active open source cluster computing in the big data world. It is an effective platform for analyzing large datasets. Apache Spark is an alternative to Hadoop MapReduce framework. The most important feature of Apache Spark is the cluster computing within memory which increases application processing speed. Fig. 10 shows the architecture of Apache Spark, which is based on two main abstractions: *Resilient Distributed Dataset (RDD)*, and *Directed Acyclic Graph (DAG)*.

To work deeper in Spark architecture [16], there is a key node containing the driver that drives the application. The code written as a driver or if you use the interactive shell, the shell functions as a driver.

Within the driver, the Spark context is created, which works with the group manager to manage various functions. The task is divided into multiple tasks that are distributed to the worker node. The RDD is created at any time in the context of Spark.

The contract of the worker is to perform the essential tasks. These tasks are performed on the RDDs that are divided into the working node and the result is returned to the Spark context.

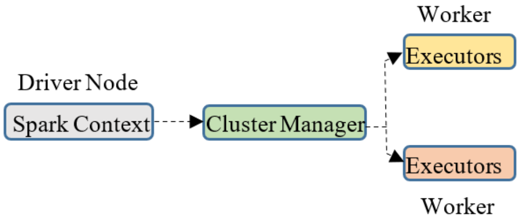


Fig. 10. Spark Architecture [17].

TABLE. 3. SUMMARY OF RESULTS IN SELECTED MODELS USING SPARK CLUSTER.

Model	Accuracy training	Accuracy Test	Execution Time
DNN	0.62	0.53	75.326

As the number of workers increases, jobs can be divided into more departments and implemented in parallel with multiple systems. It will be much faster.

After initialized Spark cluster with master and three worker nodes, and use Spark-submit to execute python application to the master, Table 3 presents the results of the experiment after using Spark cluster.

By comparing results, the execution time that was consumed in Spark cluster was decreased, and to summarize, the performance of using Spark cluster to execute tasks takes lower time.

VI. CONCLUSION

Our experimental results show a very promising solution to the costly and high readmission rate. We experimented with many different predictive models, and it successfully and accurately predicted unplanned readmissions for diabetic patients. Our best and most promising model (DNN) reported an accuracy of 66% on a training set and 50% on the testing set. This could help healthcare providers to reduce the readmission rate by providing special treatment for diabetic patients with a high risk of being readmitted in short period of time. We concluded that machine learning algorithms could be reliable in predicting readmission rates.

In addition to our work in developing predictive models, we used data exploration and analysis techniques to give a deeper insight into the critical risk factors that influence the readmission rate. Using an ablation study on the random forest model, we successfully identified the top ten critical risk factors. This could help future studies and healthcare providers to increase the quality of healthcare services and reduce the operational costs. We concluded that the type of discharge disposition, patient age, and the number of medications that the patient use are the most influential risk factors in readmission rates to hospitals.

Finally, after showing our promising work on predictive models and identifying the critical risk factors, we contributed a faster pipeline to train and execute our work. In Big Data and especially when we work with medical data, it is critical to reduce the execution time as much as possible. We report a 32% execution time reduction on training machine learning algorithms using Apache Spark which shows promising results for future work on big medical data analysis.

REFERENCES

[1] Bhuvan M S, Ankit Kumary, Adil Zafarz, Vinith Kishore, "Identifying Diabetic Patients with High Risk of Readmission ", arXiv:1602.04257v1 [cs.AI] 12 Feb 2016.

[2] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, " Impact of HbA1c Measurement on Hospital Readmission Rates:Analysis of 70,000 Clinical Database Patient Records ", Hindawi Publishing Corporation, BioMed Research International, Volume 2014, Article ID 781670, 11 pages, http://dx.doi.org/10.1155/2014/781670.

[3] P G F Swift, J R Hearnshaw, J L Botha, G Wright, N T Raymond, K F Jamieson, " A decade of diabetes: keeping children out of hospital", BMJ: British Medical Journal, Vol. 307, No. 6896 (Jul. 10, 1993), pp. 96-98, https://www.jstor.org/stable/29720335

[4] K. Saravananathan and T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining", Indian Journal of Science and Technology, Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/93874, November 2016.

[5] Damian Mingle, " Predicting Diabetic Readmission Rates: Moving Beyond Hba1c ", Current Trends Biomedical Engineering & Biosciences, Volume 7 Issue 3 - August 2017, DOI: 10.19080/CTBEB.2017.07.555715

[6] Andjela Drincic, Elisabeth Pfeffer, Jiangtao Luo, Whitney S. Goldner, " The effect of diabetes case management and Diabetes Resource Nurse program on readmissions of patients with diabetes mellitus", Journal of Clinical & Translational Endocrinology 8 (2017) 29–34

[7] John Billings, Jennifer Dixon, Tod Mijanovich and David Wennberg, "Case finding for patients at risk of readmission to hospital: development of algorithm toidentify high risk patients", BMJ: British Medical Journal, Vol. 333, No. 7563 (12 August 2006), pp. 327-330

[8] Charlie Xu, Stephone Christian, Christina Pan, "Beating Diabetes: Predicting Early Diabetes Patient Hospital Readmittance to Help Optimize Patient Care",12/15/2017

[9] Michael Kahn, Diabetes 130-US Hospitals for Years 1999-2008 Data Set, UCI Machine Learning Repository: Diabetes Data Set , UCI Center: archive.ics.uci.edu/ml/datasets/diabetes

[10] Hephzibah Munnangi, MS, Dr. Goutam Chakraborty, "Predicting Readmission of Diabetic Patients using the high performance Support Vector Machine algorithm of SAS® Enterprise Miner™", Oklahoma State University, Stillwater,OK

[11] Darwin E. Asper, "Predicting hospital readmissions in patients with diabetes:importance of diabetes education and other factors", A Dissertation Submitted to the Faculty of The College of Educationin Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy, Florida Atlantic University, Boca Raton, FL, August 2009

[12] Kathleen M. Dungan, M.D., "The Effect of Diabetes on Hospital Readmissions", Journal of Diabetes Science and Technology , Volume 6, Issue 5, September 2012

[13] William D. Spector, Ryan Mutter, Pamela Owens and Rhona Limcangco, "Thirty-Day, All-cause Readmissions for Elderly Patients Who Have an Injury-relatedInpatient Stay", Medical Care, Vol. 50, No. 10 (October 2012), pp. 863-869

[14] J.Archenaa and Dr E.A.Mary Anita, "Health Recommender System using Big data analytics", Journal of Management Science and Business Intelligence, 2017, 2–2, Aug. 2017, pages 17-24

[15] Daniel J. Rubin, "Correction to: Hospital Readmission of Patients with Diabetes", Published online: 13 March 2018, LLC, part of Springer Nature 2018

[16] Cilcia Pinto, A Spark Based Workflow For Probabilistic Linkage Of Healthcare Data, Brazilian Research Council White Paper, 2013

[17] Apache Spark Architecture, https://www.edureka.co/blog/spark-architecture/