Journal of Medical Systems (2021) 45: 11
https://doi.org/10.1007/s10916-020-01686-4

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Different Scenarios for the Prediction of Hospital Readmission of Diabetic Patients

Cristiana Neto[1] · Fábio Senra[2] · Jaime Leite[2] · Nuno Rei[2] · Rui Rodrigues[2] · Diana Ferreira[1] · José Machado[1]

Received: 1 October 2020 / Accepted: 1 December 2020 / Published online: 7 January 2021

Abstract

Hospitals generate large amounts of data on a daily basis, but most of the time that data is just an overwhelming amount of information which never transitions to knowledge. Through the application of Data Mining techniques it is possible to find hidden relations or patterns among the data and convert those into knowledge that can further be used to aid in the decision-making of hospital professionals. This study aims to use information about patients with diabetes, which is a chronic (long-term) condition that occurs when the body does not produce enough or any insulin. The main purpose is to help hospitals improve their care with diabetic patients and consequently reduce readmission costs. An hospital readmission is an episode in which a patient discharged from a hospital is admitted again within a specified period of time (usually a 30 day period). This period allows hospitals to verify that their services are being performed correctly and also to verify the costs of these re-admissions. The goal of the study is to predict if a patient who suffers from diabetes will be readmitted, after being discharged, using Machine Leaning algorithms. The final results revealed that the most efficient algorithm was Random Forest with 0.898 of accuracy.

Keywords Data mining · Diabetes · Weka · RapidMiner studio · Prediction · Readmission

Introduction

One way to characterize health systems is by using readmission metrics, i.e., to check if the patient returns to the hospital after their initial discharge [8]. There are three types of readmissions: planned, unplanned and unavoidable. The unavoidable readmissions are highly predictable due to the nature of the pathology or patient's condition [6]. Since planned readmissions are also easy to anticipate, the focus of this study is unplanned readmissions.

In 2012 the Hospital Readmissions Reduction Program (HRRP) was established which in reducing the payments to hospitals with excessive readmissions, where the critical aspect are the 30 days after a patient discharge. If a readmission does not happen in that time, the hospital receives a monetary benefit [7]. In a study made about the HRRP it was concluded that this had a great impact in reducing the readmissions, however its priority remains low and hospital leaders give more attention to factors such as patient safety, patient experience, and adherence to guidelines [10].

Despite all the continuous scientific and technological advances, diabetes remains a disease haunted by frequent hospital readmissions. Patients with diabetes account for approximately 480,958 hospital in-patient stays per year, with a 30-day readmission rate of 97,784, accounting for a 20.3% hospital readmission rate [6].

According to American Diabetes Association, in 2018, 34.2 million Americans suffered from diabetes, but this number is increasing every year. Diabetes are the seventh cause of death in the U.S.A (2017) and has a cost of 327 billion of dollars in direct and indirect estimated costs [3]. Also, it has been estimated that about 366 million people worldwide suffer from diabetes and that number can escalate to 552 million in 2030 [13].

Since these patients are among the most costly and their readmission has several repercussions to the hospitals, an accurate prediction can lead to better medical care for the

This article is part of the Topical Collection on *Health Information Systems & Technologies*
Guest Editors: Álvaro Rocha and Joaquim Gonçalves

✉ Cristiana Neto
cristiana.neto@algoritmi.uminho.pt

Extended author information available on the last page of the article.

patient while admitted, as well as hospital cost reductions by avoiding his/her readmission.

The discharge of a patient is often a decision made by health professionals, and thus it's inevitably subjective and more prone to errors. Data Mining (DM) enables the limit of this human subjectivity in decision-making processes, handling the large amounts of data collected on a daily basis, at an increasing speed with the help of the growing power of computers [5].

On a business perspective, our goal is to predict, using DM techniques, if a patient with diabetes will be readmitted within a period of 30 days after being discharged. DM classification algorithms help to discover patterns and connections that would be hard to find otherwise [2].

This paper presents a great contribution to the scientific community since it achieved promising results compared with other works related to the topic that also uses DM to predict hospital readmission for patients with diabetes. As this study is based in a medical context, the constant search for highly accurate predictions is essential for better healthcare providing.

The rest of the paper is organized as follows: in Section "Related work" it is presented some works related to the topic, in Section "Methodology" it is presented the methodology used in this study, the Section "Results and discussion" shows the results and their discussion and, finally, the conclusions are drawn in Section "Conclusions".

Related work

Research on the topic revealed several studies that attempted to achieve the goal outlined in this paper. Most of these studies used a dataset very similar to the one used in this study. The state-of-the-art results opened opportunities for improvement, establishing the context for the development of this study. Table 1 compares different studies related to the topic.

Methodology

To carry out the DM process we followed the CRISP-DM methodology, presented in Fig. 1. This methodology is more complete comparatively to others such as SEMMA (Sample, Explore, Modify, Model, Assess) [16]. The CRISP-DM steps will be described next. To conduct this study, two tools were used: RapidMiner for the Data Preparation phase and Weka for the Modeling and Evaluation phases. These tools were chosen due to their usability and the vast number of classifiers available.

Business understanding

Hospital readmissions are not only a quality indicator of healthcare systems, but also a financial problem for several

Table 1 Analysis of the related work

Authors	Title	Description
Shankar and Manikandan	Predicting the risk of readmission of diabetic patients using deep neural networks	This paper predicts whether a patient discharged from the hospital will return within 30 days or not using Linear Support Vector Machine, Random Forest (RF), Multilayer Perceptron (MLP) and a Deep Neural Network. The best accuracy value presented is 0.840 [17].
Duggal, Shukla, Chandra, Shukla and Khatri	Predictive risk modelling for early hospital readmission of patients with diabetes in India	This study classified the patients into two different risk groups of readmission (Yes or No) within 30 days of discharge based on patients' characteristics using 2-year clinical and administrative data. Five different DM classifiers were used, namely, Naïve Bayes (NB), Logistic Regression, RF, Adaboost, and Neural Networks. RF was found to be the optimal classifier for this task, with an accuracy of 0.876 [4].
Tamin and Iswari	Implementation of C4.5 algorithm to determine hospital readmission rate of diabetes patient	This study aims to use a decision tree, more precisely C4.5, to determine the rate of hospital readmission of patients with diabetes. It distinguishes itself from other articles for using three values of the readmitted attribute, that is, *<30*, *>30*, and *NO* in some scenarios. The best accuracy result achieved was 0.745 [20].

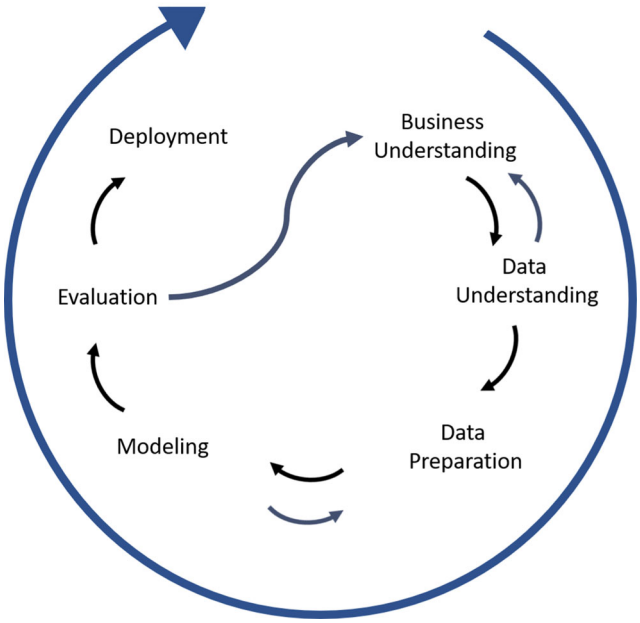

Fig. 1 CRISP-DM methodology steps

nations [6]. According to Agency for Healthcare Research and Quality (AHRQ) there were about 3.3 million 30-day readmissions in the United States. Those readmissions contributed to about 41 billion dollars in hospital costs. AHRQ also states that the third cause of readmissions for Medicaid[1] patients is Diabetes Mellitus (23,700 readmissions) [9]. The Centers for Medicare & Medicaid Services (CMS) have taken some measures to reduce hospital readmission rates and improve the quality of healthcare in the United States [13]. As it was said earlier, readmissions are a serious problem with several consequences, with rates between 8.5 and 13.5%, but when the focus goes to readmissions of patients who suffer from diabetes the rate goes up to 14.4–21.0%. With the number of diabetics increasing annually these rates tend to grow [15]. In this sense, to help reduce these readmissions, the business aim of this project is to classify cases of patients with diabetes susceptible to the occurrence of hospital readmission.

Data understanding

The dataset used in this study contains data related to the risk of hospital readmission of patients with diabetes collected over the span of 10 years (1999-2008) in 130 hospitals in the USA. It contains information related to 101766 patients and has 50 attributes that relate to the patient's personal data, hospital episode, laboratory tests and other tests, drugs, therapy and also data on the patient's clinical history [1, 19].

[1]Medicaid in the United States is a federal and state program that helps with medical costs for people with limited income and resources.

The instances of the dataset are represented by 50 variables present in Table 2.

Most of the attributes are related to the patients' personal features, their clinical history and drugs that can be administered or changed to the patient during the episode. Furthermore, it is possible to verify that 27 of 50 attributes are associated with drugs, and the patients' personal characteristics are the remaining attributes (these can be divided in 5 variables for personal features, 10 for the patient's encounter and, finally, 7 for his/her historical data).

Data preparation

In this section, the data pre-processing that was performed on the dataset before applying the Data Mining Models (DMM) is described.

This process started with the removal of the attributes:

- *acetohexamide*, *cytoglipton* and *examite*, since these had only one value;
- *medical_specialty*, *weight*, for having too many missing values;
- *encounter_id*, *patient_nbr* and *payer_code*, because these were considered unnecessary attributes for the classification.

Next, the missing values and outliers were treated since they could negatively influence the analysis process due to the lack of information and the creation of instability in the attributes' values, respectively. The missing values of the nominal attributes were replaced by the mode and the numerical ones were replaced by the mean. Outliers and instances that had the attributes related to death or hospice were removed, since these patients could not be readmitted.

In order to facilitate the classification process, it was decided to group the values of some attributes:

- *Abnorm* was used to replace the values *>7* and *>8* in the *A1Cresult* attribute, and the values *>200* and *>300* in the *max_glu_serum* attribute.
- The age values of [0-10] and [10-20] were replaced by *child/young*, the values between [20-30] and [50-60] were replaced by *adult* and the values between [60-70] and [90-100] were replaced by *elderly*.
- All three *diag* attributes, these being the *diag_1*, *diag_2* and *diag_3*, were replaced by an id that represents the value range where it belongs.
- Finally, the values *<30* and *>30* of the attribute *readmitted* were replaced by *YES* and *NO* respectively.

The next action was to map all the values of every polynomial attribute, except for the target attribute, to a unique integer so that the normalization process could be applied to obtain better performance when training the models.

Table 2 Attributes of the dataset

Attribute	Description
encounter_id	unique identifier of an encounter
patient_nbr	unique identifier of a patient
race	race of the patient (Caucasian, African American,...)
gender	patient's gender
age	patient's age grouped in 10-year intervals ([0, 10], ..., [90, 100])
weight	patient's weight in pounds
admission_type_id	integer identifier corresponding to 8 distinct values (1-emergency, 2-urgent, 3-elective...)
discharge_disposition_id	integer identifier corresponding to 29 distinct values (1-discharged to home, 2-Discharged/transferred to another short term hospital...)
admission_source_id	integer identifier corresponding to 25 distinct values (1-Physician Referral, 2-Clinic Referral...)
time_in_hospital	number of days between admission and discharge
payer_code	payment method corresponding to 23 distinct values (MC – Medicare, SP – self-pay...)
medical_specialty	specialty of the admitting physician, corresponding to 84 distinct values such as cardiology and internal medicine
num_lab_procedures	number of lab tests performed during the encounter
num_procedures	number of procedures (other than lab tests) performed during the encounter
num_medications	number of distinct generic names administered during the encounter
number_outpatient	number of outpatient visits of the patient in the year preceding the encounter
number_emergency	number of emergency visits of the patient in the year preceding the encounter
number_inpatient	number of inpatient visits of the patient in the year preceding the encounter
diag_1	the primary diagnosis (coded as first three digits of ICD9)
diag_2	secondary diagnosis (coded as first three digits of ICD9)
diag_3	additional secondary diagnosis (coded as first three digits of ICD9)
number_diagnoses	number of diagnoses entered to the system
max_glu_serum	value that indicates the range of the result or if the test was not taken. (>200, >300, normal, none - if not measured)
A1Cresult	value that Indicates the range of the result or if the test was not taken. (>8, >7 (>7 and <= 8), normal (if the result was less than 7%) and none (if not measured))
metformin, *repaglinide*, *nateglinide*, *chlorpropamide*, *glimepiride*, *acetohexamide*, *glipizide*, *glyburide*, *tolbutamide*, *pioglitazone*, *rosiglitazone*, *acarbose*, *miglitol*, *troglitazone*, *tolazamide*, *examide*, *citoglipton*, *insulin*, *glyburide - metformin*, *glipizide - metformin*, *glimepiride - pioglitazone*, *metformin - rosiglitazone*, *metformin - pioglitazone*	features that indicates whether the drug was prescribed or there was a change in the dosage(up-dosage was increased down-dosage was decreased)
change	indicates if there was a change in diabetic medications (either dosage or generic name)
diabetesMed	indicates if there was any diabetic medication prescribed (yes, no)
readmitted	value to predict and represent the number of days to inpatient readmission (<30, >30, NO)

The final step of the data preparation process was to balance the dataset by applying Oversampling to the minority class and Undersampling to the majority class applying then a randomization process to ensure that the data is disperse. This balancing was performed without equalizing the classes totally in order to avoid too many synthetic data, making the models more reliable.

Modeling

The chosen tool to perform this step was WEKA. This software has numerous DM methods available: trees-based algorithms like RF and J48, bayesian learning algorithms such as NB, rules-based algorithms like zeroR and lazy learning like IBk. In this sense, in order to include these different types of algorithms, this study used the IBk, J48, RF, NB and MLP algorithms.

After having the dataset prepared and the algorithms chosen, the next step was to select the sampling methods. Cross-validation with 10 folds and holdout sampling (70% to the train set and 30% test set) were the sampling methods chosen for this study.

Finally, some scenarios were elaborated to see how the algorithms reacted to the withdrawal or introduction of new attributes, in order to study the influence of these attributes on the readmission of diabetic patients, as can be seen in Fig. 2.

Scenario	Description	Attributes
S1	All the remaining attributes after the data preparation step	All except acetohexamide, cytoglipton, examite, medical_specialty, weight, encounter_id, patient_nbr, payer_code
S2	The attributes more related with the patient and with his/her episodes and tests performed	Age, Race, Gender, Time in hospital, Number of lab tests performed during the encounter, Number of procedures (other than lab tests) performed during the encounter, Number of distinct generic names administered during the encounter, Number of outpatient, emergency, inpatient visits of the patient in the year preceding the encounter, Number of diagnoses entered to the system
S3	The attributes related to medications and their changes	All medications, Change (Indicates if there was a change in diabetic medications) and diabetesMed (Indicates if there was any diabetic medication prescribed)
S4	The attributes that show more diverse values	All diagnoses, Age, A1CResult, Race, Gender, Metformin, Glipizide, Glyburide, Insulin, Change in diabetic medication, Diabetes medication, Admission type, Discharge disposition, Admissions source, Time in hospital, Number of lab tests performed, Number of procedures (other than test), Number of medications administered during the encounter, Number of outpatient visits, Number of inpatient visits, Number of diagnoses entered in the system
S5	The attributes more related with the patient, his/her hospital admission and the way he/she leaves the hospital	All diagnoses, Age, Gender, Race, Type of Admission, Discharge disposition and admission source
S6	The twelve first attributes that have more weight by correlation with the attribute "readmitted", obtained using the Weight by Correlation RapidMiner operator.	Number of inpatient visits, Discharge disposition, Number of diagnoses, time in hospital, Age, Number of medications during the encounter, A1CResult, Metformin, Diabetes Medication Race, Number of lab tests performed, Insulin

Fig. 2 Scenarios created with different attributes

Table 3 Performance results of the DMM obtained for Scenario I

Classifiers	Percentage split			Cross Validation		
	Accuracy	Precision	Kappa statistic	Accuracy	Precision	Kappa statistic
NB	0.535	0.614	0.136	0.537	0.615	0.138
IBk	0.800	0.845	0.609	0.812	0.856	0.631
J48	0.845	0.846	0.667	0.849	0.850	0.675
RF	0.896	0.910	0.771	0.898	0.911	0.776
ML	0.654	0.660	0.284	0.663	0.660	0.283

Evaluation

For each scenario it was performed an evaluation and comparison between the algorithms used. To measure the performance of the algorithms three principal metrics were taken into account [18]:

- *Accuracy*, which gives the number o correctly classified instances;
- *Precision*, which measures the classifier exactness;
- *Kappa Statistic*, which measures if the result can be trustfull or if it has occured by chance. A k value between 0.61 and 0.80 represents substantial agreement and between 0.81 and 1 represents almost perfect agreement [12].

Accuracy (1) and Precision (2) can be calculated using the values obtained from the confusion matrix for True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) [14]. Kappa statistic can be represented by Eq. 3, where where p_o is the observed agreement, and p_e is the expected agreement.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$k = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

Results and discussion

In this section it is presented the results obtained from the application of the DMM on the previously described scenarios. For the first scenario, presented in Table 3, all attributes coming from the Data Preparation phase were considered. This was the scenario where the best results were obtained for all three metrics, and the algorithm with the best performance was RF achieving 0.898 of accuracy.

In S2, Table 4, attributes related with the patients' personal information and their medical diagnostics were considered. The results were slightly worse than in the previous scenario but once again the RF algorithm had the best performance with an accuracy of 0.873. This reveals that not considering medication related attributes worsens the solution.

Table 5, presents the third scenario and also the one considered to have the worst performance. In S3 only attributes related to medication were considered without any actual context about the patient. These attributes proved to not have any strong relation with our target attribute, *readmitted*, and so the lowest values for each of the three metrics were obtained.

Moving on to S4, Table 6, RapidMiner was used to obtain the attributes with more variate values. This was the second-best performing scenario, only worse than S1, and the algorithm RF once again achieved an accuracy of 0.898.

Another scenario, S5, presented in Table 7, considered attributes that were related to the patient and his discharge from the hospital information. The results obtained were worse than the previous scenarios (except for S3), meaning

Table 4 Performance results of the DMM obtained for Scenario II

Classifiers	Percentage split			Cross validation		
	Accuracy	Precision	Kappa statistic	Accuracy	Precision	Kappa statistic
NB	0.616	0.593	0.122	0.616	0.593	0.121
IBk	0.758	0.769	0.507	0.768	0.778	0.525
J48	0.800	0.806	0.557	0.809	0.811	0.580
RF	0.868	0.874	0.711	0.873	0.878	0.721
MLP	0.627	0.611	0.103	0.612	0.590	0.118

Table 5 Performance results of the DMM obtained for Scenario III

Classifiers	Percentage split			Cross Validation		
	Accuracy	Precision	Kappa statistic	Accuracy	Precision	Kappa statistic
NB	0.443	0.534	0.010	0.451	0.540	0.018
IBk	0.619	0.662	0.034	0.620	0.638	0.033
J48	0.613	0.719	0.011	0.617	0.655	0.019
RF	0.621	0.673	0.041	0.622	0.648	0.041
MLP	0.610	0.558	0.000	0.603	0.541	0.011

Table 6 Performance results of the DMM obtained for Scenario IV

Classifiers	Percentage split			Cross Validation		
	Accuracy	Precision	Kappa statistic	Accuracy	Precision	Kappa statistic
NB	0.621	0.616	0.193	0.622	0.617	0.192
IBk	0.802	0.845	0.612	0.814	0.855	0.634
J48	0.844	0.844	0.663	0.850	0.850	0.676
RF	0.896	0.908	0.771	0.898	0.910	0.775
MLP	0.658	0.652	0.265	0.663	0.653	0.261

Table 7 Performance results of the DMM obtained for Scenario V

Classifiers	Percentage split			Cross Validation		
	Accuracy	Precision	Kappa statistic	Accuracy	Precision	Kappa statistic
NB	0.612	0.579	0.051	0.613	0.578	0.049
IBk	0.738	0.735	0.440	0.756	0.753	0.475
J48	0.793	0.805	0.537	0.805	0.812	0.564
RF	0.836	0.838	0.642	0.841	0.844	0.652
MLP	0.657	0.680	0.169	0.637	0.624	0.120

Table 8 Performance results of the DMM obtained for Scenario VI

Classifiers	Percentage split			Cross Validation		
	Accuracy	Precision	Kappa statistic	Accuracy	Precision	Kappa statistic
NB	0.626	0.609	0.161	0.624	0.605	0.152
IBk	0.768	0.781	0.528	0.780	0.792	0.552
J48	0.828	0.831	0.625	0.832	0.835	0.633
RF	0.876	0.883	0.728	0.881	0.888	0.741
MLP	0.658	0.647	0.239	0.654	0.647	0.253

that only considering these attributes is not enough to obtain the best prediction model.

Lastly, in the sixth scenario, presented in Table 8, the RapidMiner was once again used to obtain the attributes with more correlation weight to the *readmitted* attribute. Alongside S4, this scenario showed similar results while still not reaching the level of performance achieved in S1. In this scenario there are important features for the prediction such as age and race, considered good predictors for hospital readmissions by several studies [11].

Across all scenarios the best results came from S1, where all attributes after Data Preparation were considered. The worst scenario was S3, using all the medications, showing that to obtain a good classification, the attributes need to be related with information about the patient and not only with the medications and their changes, which was predictable. Also, when using the scenario with the attributes more correlated to the target (S6), the results proved to be good because as it takes into account the most relevant features for the prediction.

Trough the analysis of all tables, the RF algorithm stands out by getting above 0.8 of accuracy in 4 out of 6 scenarios independently of the sampling method. The worst performance was with NB algorithm showing the lowest results in all scenarios.

In what regards the sampling methods used, the results show that there is no significant differences in the final performance of the models for both of the methods. However, Cross Validation performer slightly better, which was expected since it uses all data to train the model, dividing it into k folds and allowing all data to be used for testing and training.

Comparing these results with the related work analysed in Section "Related work", this work presents better results overall. For example, the analysed study with better results presented an accuracy of 0.876, in turn, our study was able to build a model that achieved an accuracy of 0.898. However, a direct comparison can not be established since the datasets are different. The study carried out by Shankar and Manikandan obtained also a lower accuracy value (0.840). Finally, the study carried out by Tamin and Iswari, although it obtained its best results with RF, like in the present work, our better accuracy value obtained can explained by the different approaches taken in the data preparation and/or scenarios.

Conclusions

This project was mainly focused on the application of DM techniques to predict the early (less than 30 days) readmission of patients who suffer from diabetes taking into account several characteristics such as age, number of emergency entries or time spent in the hospital. The readmissions in hospitals are still a big problem for several countries and are the cause of spending lots of money that could be used to improve the health facilities.

The best results were obtained with RF, which presented 0.898 (first and fourth scenario) and 0.873 (second scenario) of accuracy. The third best result was reached using J48 algorithm that obtained 0.849 of accuracy in the first scenario. The best results for Precision (0.910) and Kappa Statistic (0.771) were also observed with RF in S1 and S4. These results show that, with this dataset, using all attributes (S1) or using at least the most correlated ones (S2) are the best approaches to predict the hospital readmission of diabetic patients.

In this case study, the presence of false positives or negatives could mean that a person was classified as readmitted when it was not or vice-versa. Thus, although the classification errors in this case are not as dangerous as predicting the presence of diseases in patients, it remains very important to obtain a good classification in order to show if the care given to the patients was good or if it resulted in their hospital readmission.

For future work, the dataset could be enriched by filling in the missing values and adding new features encompassing more context of the patients, such as attributes that show eating habits, economic conditions and insurance type, so the prediction can be more accurate and complete. The introduction of these type of attributes could bring more information to support the prediction. It would be interesting to collect a bigger dataset in order to enrich the study. Also, more metrics could be evaluated and compared.

Funding This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors. The used dataset is public and submited on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. John Clore (jclore@vcu.edu), Krzysztof J. Cios (kcios@vcu.edu), Jon DeShazo (jpdeshazo@vcu.edu), and Beata Strack (strackb@vcu.edu). This data is a de-identified abstract of the Health Facts database (Cerner Corporation, Kansas City, MO).

References

1. Archive.ics.uci.edu, and 2020, UCI Machine Learning Repository: Diabetes 130-US Hospitals For Years 1999-2008 Data Set.

https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008. Accessed 25 Sep 2020.
2. Bharati, M., and Ramageri, M., Data mining techniques and applications. *Indian Journal of Computer Science and Engineering* 1, 2010.
3. Centers for Disease Control and Prevention, National Diabetes Statistics Report. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf. Accessed 25 Sep 2020, 2020.
4. Duggal, R., Shukla, S., Chandra, S., Shukla, B., and Khatri, K., Predictive risk modelling for early hospital readmission of patients with diabetes in India. *International Journal of Diabetes in Developing Countries* 36(4):519–528, 2016.
5. Ferreira, D., et al., Predictive data mining in nutrition therapy. In: *2018 13th APCA International Conference on Automatic Control and Soft Computing (CONTROLO)*, pp. 137–142: IEEE, 2018.
6. Goudjerkan, T., and Jayabalan, M., Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron. *Int. J. Adv. Comput. Sci. Appl.* 10(2), 2019.
7. Graham, E., Saxena, A., and Kirby, H., Identifying high risk patients for hospital readmission. *SMU Data Science Review* 2(1):22, 2019.
8. Hempstalk, K., and Mordaunt, D., Improving 30-day readmission risk predictions using machine learning. Health Informatics New Zealand (HiNZ) Conference, 2016.
9. Hines, A. L., Barrett, M. L., Jiang, H. J. et al., Conditions With the Largest Number of Adult Hospital Readmissions by Payer 2011: Statistical Brief #172. In: *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs* https://www.ncbi.nlm.nih.gov/books/NBK206781/. Accessed 25 Sep 2020, 2014.
10. Joynt, K., Figueroa, J., Orav, E., and Jha, A., Opinions on the Hospital Readmission Reduction Program: results of a national survey of hospital leaders. *The American journal of managed care*22(8), 2016.
11. Kroch, E., Duan, M., Martin, J., and Bankowitz, R., Patient factors predictive of hospital readmissions within 30 days. *J. Healthc. Qual.* 38(2):106–115, 2016.
12. McHugh, M. L., Interrater reliability: the kappa statistic. *Biochem Med, (Zagreb)* 22(3):276–282, 2012.
13. Munnangi, H., and Chakraborty, G., Predicting readmission of diabetic patients using the high performance support vector machine algorithm of sas® enterprise miner™. In: *Proc. of SAS Global Forum*, 2015.
14. Neto, C. et al., Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients. *Entropy* 21(12):1163, 2019.
15. Rubin, D., Donnell-Jackson, K., Jhingan, R., Golden, S., and Paranjape, A., Early readmission among patients with diabetes: a qualitative assessment of contributing factors. *Journal of Diabetes and its Complications* 28.6:869–873, 2014.
16. Shafique, U., and Qaiser, H., A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research* 12(1):217–222, 2014.
17. Shankar, G. S., and Manikandan, K., Predicting the risk of readmission of diabetic patients using deep neural networks. In: *Innovations in Computer Science and Engineering*, pp. 385–392. Singapore: Springer, 2019.
18. Silva, C., Oliveira, D., Peixoto, H., Machado, J., and Abelha, A., Data mining for prediction of length of stay of cardiovascular accident inpatients. In: *International Conference on Digital Transformation and Global Society*, pp. 516–527. Cham, 2018.
19. Strack, B., et al., Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014.
20. Tamin, F., and Iswari, N., Implementation of C4. 5 algorithm to determine hospital readmission rate of diabetes patient. In: *2017 4th International Conference on New Media Studies (CONMEDIA)*, pp. 15–18: IEEE, 2017.

Affiliations

Cristiana Neto[1] · Fábio Senra[2] · Jaime Leite[2] · Nuno Rei[2] · Rui Rodrigues[2] · Diana Ferreira[1] · José Machado[1]

Fábio Senra
a82108@alunos.uminho.pt

Jaime Leite
a80757@alunos.uminho.pt

Nuno Rei
a81918@alunos.uminho.pt

Rui Rodrigues
a74572@alunos.uminho.pt

Diana Ferreira
diana.ferreira@algoritmi.uminho.pt

José Machado
jmac@di.uminho.pt

1 Algoritmi Research Center, Braga, Portugal

2 University of Minho, Braga, Portugal