Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 591–596

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Healthcare Interoperability and Pervasive Intelligent Systems (HiPIS 2017)

Understanding Stroke in Dialysis and Chronic Kidney Disease

Mariana Rodrigues[a], Hugo Peixoto[b], Marisa Esteves[b], José Machado[b]* and Abelha[b]

[a]*University of Minho, Campus Gualtar, Braga 4710, Portugal*
[b]*Algoritmi Center, University of Minho, Campus Gualtar, Braga 4710, Portugal*

Abstract

Patients with severe kidney failure need to be carefully monitored. One of the many treatments is called Continuous Ambulatory Peritoneal Dialysis (CAPD). This kind of treatment intends to maintain the blood tests as normal as possible. Data Mining and Machine Learning can take a simple and meaningless blood's test data set and build it into a Decision Support System. Through this article, Machine Learning algorithms will be explored with different Data Mining Models in order to extract knowledge and classify a patient with a stroke risk or not, according to their blood analysis.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Data Mining, Classification, Dialysis, Stroke Risk, Chronic Kidney Disease.

1. Introduction

Patients with Acute Renal Failure (ARF) can't live without the dialysis treatment. Dialysis is the procedure of removing all the waste substances and water from the blood using artificial machine that works similar as a kidney[1]. The most popular types of dialysis are Hemodialysis and Peritoneal. This article is focused on a blood test's data set of patients with peritoneal dialysis. Peritoneal Dialysis is a treatment where the patients must place about two quarts of cleansing fluid into their belly and later drain it. Nowadays, 11 % of the Europe's dialysis patients, are undergoing Continuous Ambulatory Peritoneal Dialysis (CADP) treatment[2]. These patients are dependent of this treatment and can't live without it. So, monitoring this patient's and know if everything is under control is extremely important.

* Corresponding author. Tel.: +351 253 604 430; fax: +351 253 604 471.
E-mail address: jmac@di.uminho.pt

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.296

Machine Learning and Data Mining are two revolutionary features of Decision Support Systems in Healthcare. But, how will a dataset predict a patient from a stroke risk? When we are faced with a situation where the kidney is injured, all values related to glomerular filtration are modified. Some blood analysis may influence each other. For example, when a patient has high creatinine and high calcium, it's likely to have high urea's values as well.

The main objective of this current paper is to determine through patient's information and blood analysis, what is the risk of a patient to suffer stroke risk, knowing that dehydration that is directly related to stroke risk, which can happen when the BUN/CR ratio is superior to 15 on older patients. In this article, it will be explained all the phases to predict stroke risk: The Business Understanding and Data preparation will be referenced, the different models used will be exposed, and finally the data mining techniques will be compared and carefully analyzed, concluding which one is the best and the reason why. In this article, the algorithms with the best results were IBK and Random Tree. Data Mining is one of the processes of machine learning. It takes a meaningless from a big data set, pre-processes it, understands it and finally analyzes patterns, turns them into useful information and even predicts and classifies. Data mining tools predict future behaviors, allowing medical teams to make proactive decisions[3]. Data mining is a huge advantage in healthcare. Some experts believe the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending[4]. A Clinical Decision Support System is a system design to provide a decision support for the medical team. A Decision support system is helpful and reminds the staff members, patients or others to knowledge a certain condition with intelligently, automatically, filtered information and presented at appropriate times, to enhance health and health care[5]. This decision support system also helps to decrease the number of Medical errors, because the staff team is more informed.

In this case, it is extremely important to take blood tests as input and output all the important diagnosis that should be considered by the doctor. It saves time for the medical team, and it can prevent a worse situation. For example, in this article, all the patient's that have a stroke risk, are diagnosed by simple machine learning classification's algorithm. This helps the doctor to understand that these patients should be more monitored.

2. Background and Related Work

Urea is connected directly with digestive and urinary systems. Urea is a waste molecule of the processes of protein digestion. At the end, it is carried by bloodstream to the kidney, where is excreted in urine. Creatinine on the other hand is a waste of muscle activity and it also is expelled by the human body through kidney's filtration. In this study, all the patients have kidney failure, which means that the urea and creatinine will not be properly expelled through urine. In these cases, dialyses is the only method that can helps to decrease Urea's and Creatinine's blood values. However, most of the times it isn't enough, so these patients need monitoring to make sure they are most of the times in non-dangerous values. Any BUN and Creatinine ratio value above 15 is considered dehydration[6]. Dehydration is one of the main causes of stroke risk in patients. It is associated with acute cerebral infarction[6]. A study done in an hospital discovered that patients with the ratio value's above 15, with age's above the 64 years old, and with acute ischemic stroke (AIS) have a big percentage of stroke risk than other patients who don't have this factors[6]. The data set doesn't give any information about the AIS, but it gives information about the other remaining factors. All the patients that are above 64 years old and have a ratio of Bun and creatinine above 15, should be monitored and be seen by a specialist to make sure they don't have AIS as well.

2.1. Related Work

Nowadays is more and more important to improve healthcare services. One of the ways is to predict bad situations to happen. Some works about data mining in healthcare field have been done during the last few years. It's the case of the article of 2014 *"Preventing patient cardiac arrhythmias by using data mining techniques"* [7]. This article explores data mining to prevent a patient cardiac Arrhythmias through a database with old cases. This study is important in healthcare services because it predicts Cardiac Arrhythmias with a 95% of Sensitivity, showing that Data Mining models are very useful for supporting the decision making process and for preventing future critical events such as Cardiac Arrhythmia. This study used the CRISP-DM approach to test each algorithm with different scenarios.

3. Methodology, Materials and Methods

3.1. Methodology

The Data mining process will explore the different phases of CRISP-DM (Cross Industry Standard Process for Data Mining): Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation and Deployment[8].

3.2. Materials

The dataset used had blood analysis and urine analysis of patients that were going through a CADP treatment in a major health care unity in the north of Portugal. This data set had an episode's collection between 2011 and 2017. It is a dataset for doctors to understand if the values are normal or not and to compare with past episodes. This data set had information like the age, gender, episode's date, blood analysis and urine analysis. Every line has a different episode. Each patient can have different episodes. The data set is composed by 850 cases.

3.3. Methods

The algorithms of classification in Weka that this article will approach are NaiveBayes (*NB*), Logistics Regression (*LR*), Multilayer Perceptron (*MP*), Random Tree (*RT*) and K-neighbors (*IBK*). The *NB* is a classification algorithm based in Bayes' Theorem. It defends that there is independence among predictors. It assumes that the presence of a feature in a class is unrelated to the presence of any other feature. It is use full for big data sets. It is also a fast algorith[9]. The *LR* uses an equation as the representation, similar to linear regression[10]. The main purpose in this paper in this function is to explain the relationship between one dependent binary variable and the stroke risk's column nominal[11]. Another one is the *MP* algorithm. It is an artificial neural network that have one or more layers. The data flows in one direction from input until output. It can solve problems that are not linearly separable[12]. The *RT* can deal with classification and regression problems. It is a tree algorithm, which means that is trained with the same parameters and uses different training sets. It is "random" because not all the variables are used at each node to find the best split, but a random subset of them[13]. And finally, the *IBK* which is an instance-based learning (IBK) is a parametric learning algorithm. K-nearest neighbor (k-NN) learning is the most popular realization of IBK. It can do distance weighting. The IBK algorithm is one of the simplest and fastest machine learning algorithm[16].

4. Data Mining Process

4.1. Business Understanding

All the patients that go through a CADP procedure should be always under vigilance. But there are always some cases that are more worrisome than others. Those cases should be monitored and recommended more frequent dialysis in order to prevent a more serious situation: the death of the patient because of a stroke risk. The main objective of this article is to predict if a patient has stroke risk or not, using the patient's information and blood analysis. The final results should be very accurate regarding the main nature of the system and its appliance to healthcare domain

4.1. Data Understanding

The columns chosen of the data set to use were all values that are directly influenced by kidney failure and that should be monitored in a patient that goes under a CADP treatment. All these substances are supposed to be expelled by the kidneys in a normal body. The columns of the final dataset after data preparation are Age, Calcium,

Chlorides, Creatinine, Ferritin, Iron and Urea. Figure 1 presents the normal distribution of the target attribute: risk of stroke.

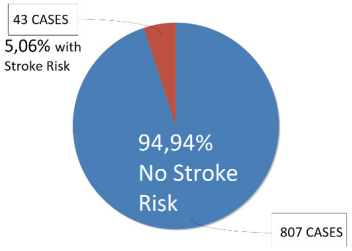

Figure. 1. Distribution of cases by risk of stroke in percentage.

4.2. Data Preparation

Most of the studies found[6] instead of using Urea's level, they use BUN's level. BUN value can be related with Urea's values through the following equation:

$$Urea(mg/dL) = BUN(mg/dL) \times 2{,}14 \quad (1)$$

Since the main objective of this work is to analyze the ratio of BUN/CREATININE level, age and use a classification algorithm to determine if the patient has risky values or not, some values had to be calculated and a new column had to be created. All the cases under 64 years old were eliminated because they were not in stroke risk. After that, the new column named "stroke risk Classification" is classified with the following rules: If the patient has BUN/Creatinine > 15, then the classification is "Dehydration (Stroke risk)". The other remaining cases have a classification as "No Dehydration (No Stroke risk)".

4.3. Data Modelling

This section, will explore the different Data Mining Models (*DMM*) used to get all the results. The DMM has the following formula:

$$DMM = \{Approach, Scenarios, Data\ Mining\ Techniques, Sampling\ Models, Data\ Approaches, Target\} \quad (2)$$

For this article, it was decided to give to each DMM the following methods: Approach: {Classification}; Scenarios: {S1; S2}; DMT: {NB, LR, MP, RT, IKB}; SM: {without oversampling, with oversampling}; DA: {cross-validation, percentage split}; Target: {Stroke Risk}.

The final objective is to predict if a patient has a Stroke risk, which means that the appropriated approach is classification and the target is to understand if the patient has Stroke Risk or not. The Scenarios (S) can be described as S1: {Gender, Age, Blood Analysis} and S2: {remove Gender and Age}. The blood analysis has Calcium, Chlorides, Creatinine, Ferratin and Iron. This article will also explore if the age and gender have or not a big influence in the results. At the end, there were 40 different models to compare and analyze: DMM = {1Approach, 2Scenarios, 5Data Mining Techniques, 2 Sampling Models, 2Data Approaches, 1Target}.

4.4. Evaluation

The first approach was use the dataset as achieved during the initial Data Preparation phase. Although, some algorithms tested presented low accuracy and the results were not as near as desired. Therefore, and given the disparity in the target attribute oversampling came as the natural solution to improve the achieved results. The

oversampling method consisted in doubling all the dataset entries and the final data set that was used had 1700 lines (850 x 2), with 86 cases of patients with stroke risk. Table 1 presents the results of sensitivity, specificity and accuracy for the 4 DMT using oversampling and for the first scenario (S1 - all attributes) using cross validation and percentage split with 66% / 33%.

Table 1. Results of NB, LR, MP, RT and IBK using Cross Validation and Percentage Split on Scenario 1.

	Cross Validation (10 Folds)			**Percentage Split (66% / 33%)**		
DMT	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**	**Accuracy**
NB	23,26 %	99,19 %	95,35 %	39,29 %	98,91 %	96,02 %
LR	43,02 %	99,38 %	96,53 %	46,43 %	98,91 %	96,37 %
MP	61,63 %	99,13 %	97,24 %	64,29 %	98,91 %	97,23 %
RT	95,35 %	99,38 %	99,18 %	71,43 %	99,64 %	98,27 %
IBK	95,35 %	99,88 %	99,65 %	71,43 %	99,27 %	97,92 %

In Table 2 are presented the results achieved for Scenario number 2. In this scenario the influence of Gender and Age are explored for the main goal of classifying the patient stroke risk.

Table 2. Results of NB, LR, MP, RT and IBK using Cross Validation and Percentage Split on Scenario 2.

	Cross Validation (10 Folds)			**Percentage Split (66% / 33%)**		
DMT	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**	**Accuracy**
NB	13,95 %	99,63 %	95,29 %	17,86 %	99,45 %	95,50 %
LR	31,40 %	98,64 %	95,24 %	35,71 %	98,18 %	95,16 %
MP	56,98 %	98,51 %	96,41 %	60,71 %	97,82 %	96,02 %
RT	95,35 %	99,88 %	99,65 %	98,92 %	91,67 %	98,62 %
IBK	13,95 %	99,63 %	95,29 %	64,29 %	99,45 %	97,75 %

After analyzing each algorithm and their respective results, it's necessary to choose the best DMM for each algorithm. For that, it was decided to create a threshold both for sensitivity as well as specificity and accuracy. The threshold create was 95% of Sensitivity, 95% Specificity and 95% Accuracy. Table 3 presents the achieved DMM.

Table 3. Best results achieved for threshold (95% of Sensitivity, 95% Specificity and 95% Accuracy).

DMT	**SM**	**DA**	**Scenario**	**Sensitivity**	**Specificity**	**Accuracy**
RT	Over sampling	Cross Validation	Scenario 1	95,35 %	99,38 %	99,18 %
IBK	Over sampling	Cross Validation	Scenario 1	95,35 %	99,88 %	99,65 %
RT	Over sampling	Cross Validation	Scenario 2	95,35 %	99,88 %	99,65 %
IBK	Over sampling	Cross Validation	Scenario 2	95,35 %	99,88 %	99,65 %

5. Discussion

In this article, it was possible to classify patients with stroke risk using Machine Learning classification's algorithms having good results. First, there's a big evidence that shows that using Cross Validation achieves better results than using percentage split. This is can be due the fact that cross validation uses all cases for learning and for tests in different iterations. Both scenarios achieve good results for the chosen DMM. However, when checking the sensitive values, it can notice a big difference between the first and the second scenario, meaning that Age and Gender could be as important as initially suspected. For example, comparing Scenario 1 the algorithms NB, LR and MP with Cross-Validation, have better results than Scenario 2. Regarding the used algorithms, *NB* is an algorithm that assumes that features are independent. Maybe that's why this algorithm fails, because to predict the value of

stroke risk, the data set needs to understand all the features and the relationship between them. *LR* didn't work because the data set didn't represent a linear regression. *MP* had better results than LR because it can solve problems that are not linearly separable (probably this model is non-linear), as it is in LR. Indeed, the two best results were: The IBK and the Random Tree achieving 95.35% of Sensitivity, 99.88% of Specificity and 99.65% of Accuracy. Anyway, these two algorithms have presented good results and they were better than the others, making them a still the best choice.

4. Conclusion and Future Work

Data mining applied to healthcare as proven to be a breakthrough for the years to come. It offers the possibility to discover hidden patterns from the big data that health organizations present. These patterns can be used by decision support systems to determine diagnoses, prognoses and treatments for patients in healthcare organizations and present them as aids to physicians. The focus of this work was to predict some condition through blood test analysis of CADP's patients to prevent poor outcome. Urea and creatinine are directly associated with dehydration, and dehydration is also associated with stroke risk. In this paper RT and IBK had the best results compared to all the other algorithms mentioned specially because they're the only algorithms that have a sensitivity, Specificity and Accuracy higher than 95%. Using this algorithm would give the medical staff a highly quality with a low error informed diagnosis. For a future work, one of the things that could be done is try to find out more diagnosis that can be evaluated through blood analysis and with urine analysis. Also, it is known that high Urea in the blood affects the nervous system, and cognitive intelligence. Perhaps there is a certain value in the blood that can be a biomarker for the destruction of the brain. In data mining, it would be also interesting to try to add more columns and features to find more relationships between the different substances of the blood.

Acknowledgments

This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

References

1. Dialysis. Available at: MedicinNet ⟨http://www.medicinenet.com/dialysis/article.htm#1whatis⟩ Last accessed May 2017
2. Arsh K. Jain, Peter Blake, Peter Cordy, Amit X. Global Trends in Rates of Peritoneal Dialysis.
3. An Introduction to Data Mining. Available at: ⟨http://www.thearling.com/text/dmwhite/dmwhite.htm⟩ Last accessed May 2017
4. What Is Data Mining in Healthcare. Available at: ⟨ https://www.healthcatalyst.com/data-mining-in-healthcare⟩ Last accessed May 2017
5. Eta S, Ed D. Clinical Decision Support Systems Agency for Healthcare Research and Quality; 2009.
6. Jon W. Schrocka,G. Michael,D. Kristin. Elevated blood urea nitrogen/creatinine ratio is associated with poor outcome in patients with ischemic stroke. MetroHealth Medical Center Department of Emergency Medicine, Case Western Reserve University School of Medicine, United States, 2012.
7. Portela F, Santos M, Silva A, Abelha A, Machado J. Preventing Patient Cardiac Arrythmias by Using Data Mining Techniques; 2014.
8. Braga A., Portela F, Santos M, Abelha A, Machado J, Silva A, Rua F. Data Mining to Predict the use of VasoPressors in Intensive Medicine Patients
9. 6 Easy Steps to Learn Naive Bayes Algorithm. Available at: ⟨https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/14.⟩ Last accessed May 2017
10. Logistic Regression for Machine Learning. Available at: ⟨http://machinelearningmastery.com/logistic-regression-for-machine-learning/ ⟩ Last accessed May 2017
11. What is Logistic Regression. Available at: ⟨ http://www.statisticssolutions.com/what-is-logistic-regression/ ⟩ Last accessed May 2017
12. Multi-Layer Perceptron. Available at: ⟨ http://neuroph.sourceforge.net/tutorials/MultiLayerPerceptron.html ⟩ Last accessed May 2017
13. RandomTrees. Available at: Opencv⟨http://docs.opencv.org/2.4/modules/ml/doc/random_trees.html⟩ Last accessed May 2017