ORIGINALAN NAUČNI RAD – ORIGINAL SCIENTIFIC PAPER 10.7251/ZRPIM2201633B

PREDICTING THE IDEAL WEIGHT IN THE PROCESS OF DIALYSIS IN CHILDREN USING REGRESSION ALGORITHMS

Marija Blagojević[1], Danijela Milošević[1], Katarina Mitrović[1], Mirjana Kostić[2], Dušan Paripović[2], Nataša Gojgić[1]

[1]University of Kragujevac, Faculty of Technical Sciences Čačak, Sv. Save 65, 32000 Čačak, Serbia, marija.blagojevic@ftn.kg.ac.rs
[2]University children's clinic Tiršova, Belgrade, Tirsova 10, 11000 Belgrade, Serbia

ABSTRACT

This study shows the use of regression algorithms in predicting the ideal weight of children during the process of haemodialysis. The data regarding height and weight for calculating the BMI (Body Mass Index) are collected at the University Children's Hospital Tirsova in Belgrade. In addition, bioimpedance, haematocrit and blood pressure are measured. The collected data are pre-processed and transformed before the application of regression algorithms. The application of regression algorithms predicts the ideal weight of the patient, which, in practice, is determined by the attending physician. Two types of errors are used for the evaluation: mean absolute error (MAE) and root mean squared error (RMSE). The results indicate satisfactory accuracy, while future work refers to the implementation of algorithms through a smart watch which would enable the timely receipt of the information regarding the ideal weight.

Keywords: haemodialysis, regression, ideal weight.

INTRODUCTION

Determining the quantity of free fluid that should be removed in patients with end-stage chronic kidney disease during haemodialysis could reduce the morbidity and mortality of this vulnerable group of patients. Namely, the life expectancy of children under the age of 14 who have the end-stage renal disease is 30 years from the moment of initiating the therapy to replace renal function with dialysis (US Renal Data System [USRDS], 2008). These children have a 50 times higher mortality rate compared to their healthy peers (Chesnaye et al., 2014). Cardiovascular diseases are the leading cause of premature mortality in this population. Increased awareness that death caused by cardiovascular diseases is a more probable outcome since the beginning of renal function replacement has directed the attention of nephrologists to the prevention of cardiovascular diseases (Keith, Nichols, Gullion, Brown, & Smith 2004). More precise control of volume status could provide a more rational application of renal function replacement with haemodialysis. Unlike bioelectric impedance, which is a non-invasive method of determining the excess of free fluid, standard methods are invasive and involve the use of ionizing radiation (the dilution method which includes the use of deuterium or tritium). Children represent a specific population in which growth significantly affects the change in fluid distribution, so the option of a quick, simple, non-invasive method without side effects would be of great medical importance.

Determining the exact amount of fluid to be removed during the process of haemodialysis is extremely difficult and requires an extensive clinical experience. Furthermore, the assessment is repeated at each haemodialysis session. Determining the amount of fluid would enable the optimization of haemodialysis not only in children and adolescents but also in the adult population. It should be taken into account that this is the therapy that requires significant funds, so in addition to medical justification, financial justification should be added, which could provide a more rational use of the funds intended for the treatment of patients with chronic kidney diseases.

Blagojević et al. (Blagojević, Milošević, Mitrović, Kostić, & Paripović, 2022) applied an artificial neural network to predict the amount of free fluid. In relation to that research, the analysis has now been extended to the machine learning algorithms related to regression.

This paper aims at investigating the possibility of applying regression algorithms (linear regression, k-means, decision tree, support vector regression and multilayer perceptron) to determine the amount of fluid.

THEORY OF ALGORITHMS AND METHODOLOGY

Linear regression

Linear regression algorithm is used to predict or visualize a relationship between two different variables. The linear regression deals with two types of variables: the dependent variable and the independent variable. As per rule, the independent variable stands by itself, without being impacted by the other one(s). The regression model is the technique used to predict/solve the dependent variable.

K-means

K-Means clustering is an unsupervised learning algorithm which divides the data into different clusters. While being different from the datasets belonging to another cluster, the datasets in one cluster are similar to each other. Unlike in supervised learning, the data is unlabelled.

Decision tree

Decision Tree Analysis is a predictive modelling tool that is applied within several different fields. In general, decision trees are created via an algorithmic approach which identifies ways to split a dataset based on various conditions. It is the most widely used and practical method for supervised learning for both classification and regression tasks.

Support vector regression

Support Vector Regression is a regression algorithm used for both linear and non-linear regressions (Pedamkar). It is based on the principle of the Support Vector Machine. Unlike SVM which is a classifier used for predicting discrete categorical labels, SVR is a regressor utilised for predicting continuous ordered variables.

When it comes to simple regression, the key goal is to minimize the error rate, whereas in SVR the main idea is to fit the error inside a certain threshold.

Multilayer perception

A multilayer perceptron (MLP) is a deep-learning method and it refers to a feed-forward artificial neural network generating a set of outputs from a set of inputs. What characterizes the MLP are several layers of input nodes connected as a directed graph between the input and output layers. For training the network, MLP deploys backpropagation.

The methodology is defined according to the common data mining process practice. The steps are defined as follows:

Data selection

The data were selected from the University Children's Hospital Tiršova after the measurement of relevant parameters during the process of haemodialysis in children.

Data pre-processing and transformation

In this phase, the data are pre-processed to obtain clean and appropriate data for the subsequent analysis. This means that the data were transposed, columns became rows and vice versa.

Model creation, testing and evaluation

Weka software [5] was used for the creation, testing and evaluation of the model,. The input parameters are the following: Date, Measurement, OH (L), TBW (L), V (L), ECW (L), ICW (L), ECW/ICW (L), BMI (kg/m2), LTI (kg/m2), FTI (kg/m2), LTM/rel LTM (kg), FAT/TT (kg), ATM (Kg), BCM (kg), Age [months], Weight before HD [kg], Height (cm), TP [m²], BMI [kg/m²], TA Sy (mmHg), while Ideal weight [kg] is the output parameter . The following approaches were used for the testing: cross-validation and splitting data in the ratio 70:30 (70% as training and 30% as a testing set). In order to evaluate the accuracy, the following errors were used: mean absolute error (Interpretation of evaluation metrics for regression analysis [IEMRA], 2022) and root mean squared error (How to Interpret Root Mean Square Error [RMSE], 2022).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{prediction}_i - \text{actual}_i)^2} \qquad (1)$$

where:

- Σ means “sum”
- Pi is the predicted value for the i^{th} observation in the dataset
- Oi is the observed value for the i^{th} observation in the dataset
- n is the sample size

$$\text{MAE} = (1/n) * \Sigma|yi - xi| \qquad (2)$$

where:

- Σ: means “sum”
- yi: The observed value for the i^{th} observation
- xi: The predicted value for the i^{th} observation
- n: The total number of observations

RESULTS AND DISCUSSION

After the selection of algorithms, the errors (MAE and RMSE) and the correlation coefficient were calculated. The results are presented in Table 1

Table 1. The errors in different regression algorithms.

Algorithm name	Correlation coefficient	Mean absolute error	Root mean squared error
Linear regression	0.9491	0.1718	0.5688
Kmeans	0.866	0.565	0.8016
Decision tree	0.5957	1.4396	1.987
SVR	0.9098	0.6361	0.7794
MLP	0.9815	0.7246	0.8709

The Greedy Stepwise was used for the optimization of the proposed model. The most important parameters, as the results showcased, are: Date, OH (L), LTM/rel LTM (kg) and TA Sy (mmHg).

Table 2 presents the results for the same algorithms, with the most important parameters.

Table 2. The errors in different regression algorithms with the optimized parameters.

Algorithm name	Correlation coefficient	Mean absolute error	Root mean squared error
Linear regression	0	2,07	2,24
Kmeans	0,77	1	1,49
Decision tree	0,59	1,44	1.98
SVR	0,62	1,41	1,92
MLP	0,67	1,35	1,83

Figure 1 presents the visualization of the errors given in Table 1.

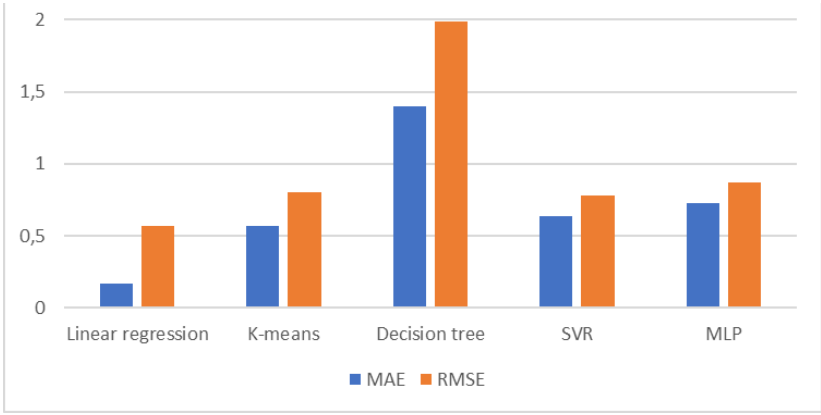

Figure 1. Visualization of MAE and RMSE without the optimization of the parameters.

CONCLUSIONS

The presented results lead towards several conclusions:

- regression algorithms for predicting the ideal children's weight could be implemented with satisfactory accuracy
- the optimized parameters are Date, OH (L), LTM/rel LTM (kg) and TA Sy (mmHg)
- the optimization of the parameters did not provide better accuracy.

The main limitation of the study refers to a necessity for gathering more data to obtain more precise results. Future research is related to the use of smart watches for obtaining the information about patients' future treatment.

Acknowledgement

This study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and these results are part of the Grant No. 451-03-68/2022-14/200132 with the University of Kragujevac – Faculty of Technical Sciences Čačak. Also, the study was supported by the Eureca project "Advanced development of hemodialysiss system with predictive fluid balance in body for kids".

LITERATURE

Blagojevic, M., Milošević, D., Mitrović, K., Kostić, M., & Paripović, D. (2022). Predicting the ideal weight in the process of hemodialysis in children using artificial neural networks. *In Proceedings Coast conference* (pp 26-29). Herceg Novi, Montenegro.

Chesnaye, N., Bonthuis, M., Schaefer, F., Groothoff, J.W., Verrina, E., Heaf, J.G., Jankauskiene, A., Lukosiene, V., Molchanova, E.A., Mota, C., Peco-Antić, A., Ratsch, I.M., Bjerre, A., Roussinov, D.L., Sukalo, A., Topaloglu, R., Van Hoeck, K., Zagozdzon, I., Jager, K.J., & Van Stralen, K.J. (2014). ESPN/ERA–EDTA registry. Demographics of paediatric renal replacement therapy in Europe: a report of the ESPN/ERA–EDTA registry. *Pediatr Nephrol, 29*, 2403–10.

How to Interpret Root Mean Square Error (RMSE). (2022). Report. Retrieved June 12, 2022, from https://www.statology.org/how-to-interpret-rmse/

Interpretation of evaluation metrics for regression analysis (mae, mse, rmse, mape, r-squared, and adjusted r-squared). (2022). Report. Retrieved June 12, 2022, from https://traindatahub.com/?p=291

Keith, D.S., Nichols, G.A., Gullion, C.M., Brown, J.B., & Smith, D.H. (2004). Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization. *Arch Intern Med*, *164,* 659−63.

US Renal Data System. (2008). USRDS 2008 annual data report: atlas of chronic kidney disease and end-stage renal disease in the United States. *National Institutes of Health, National Institute of Diabetes and Digestive and Digestive and Kidney Diseases, Vol. 2014*.