medical sciences forum

Proceeding Paper

Use of Machine Learning to Predict the Glycemic Status of Patients with Diabetes †

Alessandro Massaro [1,2], Nicola Magaletti [1], Gabriele Cosoli [1], Angelo Leogrande [1,2,]* and Francesco Cannone [3]

[1] LUM University-Giuseppe Degennaro, Casamassima (BA), Italy; massaro@lum.it (A.M.); magaletti@lumenterprise.it (N.M.); cosoli@lumenterprise.it (G.C.)
[2] LUM Enterprise s.r.l
[3] Emtesys s.r..l, Piazza Giuseppe Massari, 6 – 70122 Bari (BA), Italy; info@emtesys.com
* Correspondence: leogrande.cultore@lum.it or angelo.economics@gmail.com

† Presented at the 2nd International Electronic Conference on Healthcare, 17 February–3 March 2022. Available online: https://iech2022.sciforum.net/.

Abstract: In this work, a machine learning methodology is used to predict the progress of the glycemic values of six patients with diabetes. Eight different algorithms are compared i.e. ANN, PNN, Polynomial Regression, Gradient Boosted Trees Regression, Random Forest Regression, Simple Regression Tree, Tree Ensemble Regression, Linear Regression. The algorithms are classified based on the ability to minimize four statistical errors, namely: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Mean Signed Difference. Following the analysis, an ordering of the algorithms by predictive efficiency is proposed. Data are collected within the "*Smart District 4.0 Project*" with the contribution of the Italian Ministry of Economic Development.

Keywords: Machine Learning; Predictions; Telemedicine, ANN-Artificial Neural Network.

1. Introduction

In the following analysis the case of the use of machine learning algorithms for the prediction of the glycemic state of two patients is proposed. Through the analysis of the historical series detected with a frequency of 3 minutes it was possible to predict the future trend of the patients' glycemic status. However, to better choose the algorithms to be used for the prediction, a comparative analysis of eight different algorithms was carried out, i.e. ANN-Artificial Neural Network with Perceptron Multilayer, PNN-Probabilistic Neural Network, Polynomial Regression, Gradient Boosted Trees Regression, Random Forest Regression, Simple Regression Tree, Tree Ensemble Regression, Linear Regression. The choice of the best performing algorithm was made considering both the value of the R-square and the ability to minimize also various statistical errors detected.

The data that was processed was produced within a research project carried out by LUM Enterprise s.r.l. in collaboration with the company Noovle s.r.l. and financed by the Ministry of Economic Development of the Government of the Italian Republic. Specifically, the objective of the research project was to monitor the health status of the bus drivers from a glycemic point of view to verify whether they had any problems that could cause damage to passengers during travel. The monitoring of the glycemic health status of the drivers was carried out using special detection devices capable of detecting the condition of the individuals observed every 3 minutes. Specifically, the analysis was conducted for two different patients synthetically indicated as patient A and patient B. The historical series of the data collected differs significantly for patient A and for patient B, in fact, while for patient A there are 243 observations in the case of patient B the observations are about 13,204.

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Med. Sci. Forum* **2021**, *1*, x. https://doi.org/10.3390/xxxxx

Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Thanks to the use of telemedicine systems, it is possible to assist the patient population by means of a set of remotely operating medical tools. The detected data can be analyzed using DSS systems integrated with models of artificial neural networks also applied to the prediction of the values detected by patients [1] This prediction is useful for identifying critical elements such as to also foresee the intervention. In particular, the use of the artificial neural network-ANN with Multilayer Perceptron allows to realize the de-hospitalization process thanks to the use of intelligent sensors oriented to the measurement of patient data [2]. The relevant data using telemedicine tools can also be integrated in the analysis of big data for the analysis of the state of health of patient population [3]. The solution is therefore efficient both for the individual analysis of the patient's health condition and for the implementation of real health policies.

Telemedicine platforms can be used both to monitor the physical condition of individual patients and to carry out overall analyzes of the reference population. Telemedicine platforms can therefore be used both for the specific objectives of medicine aimed at individual patients and for more general purposes aimed at solving public health issues [4].

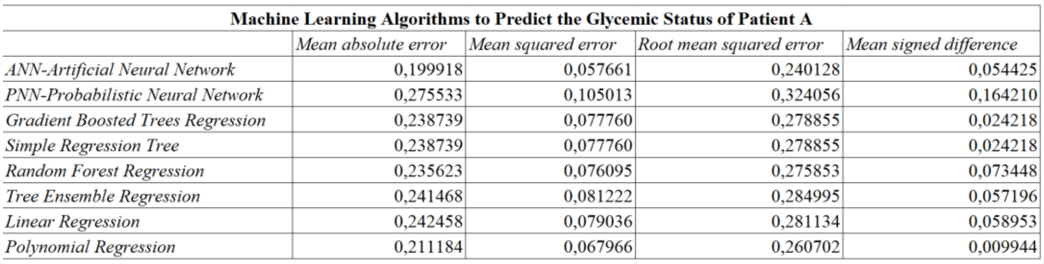
The article continues as follows: the second paragraph contains “Machine Learning and Predictions” while the third paragraph concludes.

2. Machine Learning and Predictions

Eight different machine learning algorithms for predicting the glycemic status of six different diabetes patients are analyzed below. Predictions are made for each patient by identifying the most efficient algorithm based on the historical series of surveys. Specifically, the 70% of the dataset has been used as learning rate while the remaining 30% is used for the prediction. The choice of the algorithm in terms of predictive efficiency is made based on the analysis of four different statistical errors. The Four statistical errors analyzed are: “*Mean Absolute Error*”, “*Mean Squared Error*”, “*Root Mean Squared Error*”, “*Mean Signed Difference*”.

Patient A. In the case of Patient A the best predictor of the glycemic state is the Artificial Neural Network-ANN algorithm with Perceptron Multilayer. Specifically, in the case of patient A the following order of algorithms in terms of prediction is proposed:

- Artificial Neural Network-ANN with a payoff equal to 7;
- *Polynomial Regression* with a payoff equal to 7;
- Gradient Boosted Trees Regression with a payoff equal to 14;
- *Random Forest Regression* with a payoff equal to 16;
- *Simple Regression Tree* with a payoff equal to 18;
- *Tree Ensemble Regression* with a payoff equal to 23;
- *Linear Regression* with a payoff equal to 27;
- Probabilistic Neural Network-PNN with a payoff equal to 32.



Machine Learning Algorithms to Predict the Glycemic Status of Patient A

	Mean absolute error	*Mean squared error*	*Root mean squared error*	*Mean signed difference*
ANN-Artificial Neural Network	0,199918	0,057661	0,240128	0,054425
PNN-Probabilistic Neural Network	0,275533	0,105013	0,324056	0,164210
Gradient Boosted Trees Regression	0,238739	0,077760	0,278855	0,024218
Simple Regression Tree	0,238739	0,077760	0,278855	0,024218
Random Forest Regression	0,235623	0,076095	0,275853	0,073448
Tree Ensemble Regression	0,241468	0,081222	0,284995	0,057196
Linear Regression	0,242458	0,079036	0,281134	0,058953
Polynomial Regression	0,211184	0,067966	0,260702	0,009944

Figure 1. Machine Learning Algorithms to Predict the Glycemic Status of Patient A**.**

In this case the Artificial Neural Network-ANN with Multilayer Perceptron has the following parameters: the maximum number of iterations is equal to 100, the number of hidden layers is equal to 1, the number of hidden neurons per layer is equal to 10. Since the value of threshold is equal to 150, as it can be viewed in the Figure 2, the glycemic status of the patient A is essentially under the critical level of 150. The algorithm predicts a level of the glycemic status for the patient A lower than the threshold of 150.

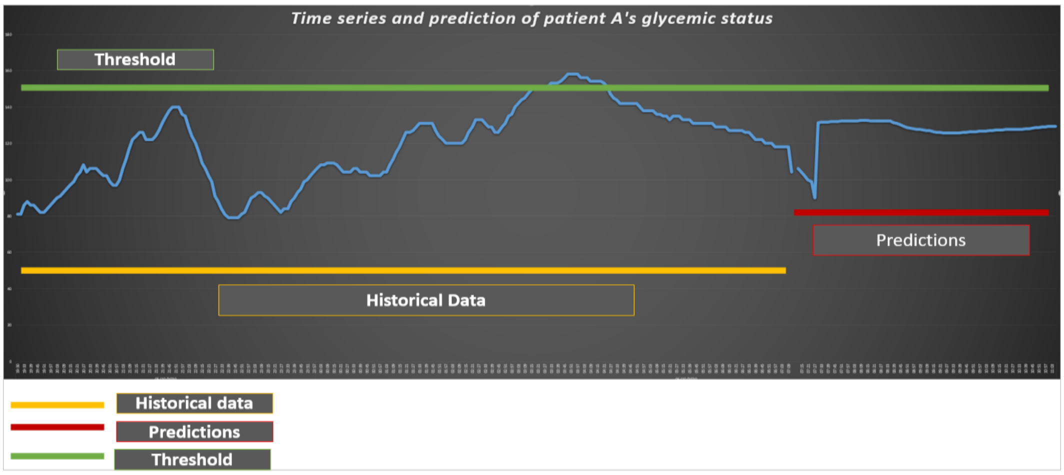
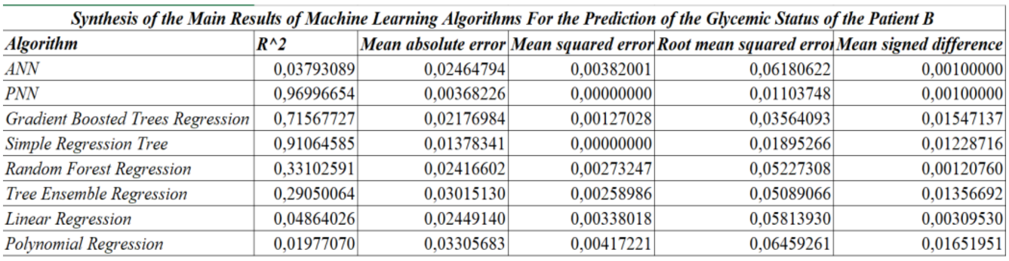


Figure 2. Machine Learning Algorithms to Predict the Glycemic Status of Patient A**.**

Patient B. In the case of Patient B the best predictor of the glycemic state is the Probabilistic Neural Network-PNN. In the case of the patient B the choice of the best predictor is realized through either the maximization of the R-Squared, either the minimization of the four statistical errors indicated. Specifically, in the case of patient B the following order of algorithms in terms of prediction is proposed:

- Probabilistic Neural Networks-PNN with a payoff equal to 6;
- *Simple Regression Tree* with a payoff equal to 13;
- Gradient Boosted Trees Regression and Random Forest Regression with a payoff equal to 19;
- *Linear Regression* with a payoff equal to 27;
- Tree Ensemble Regression and Artificial Neural Network-ANN with a payoff equal to 28;
- *Polynomial Regression* with a payoff equal to 40.



Synthesis of the Main Results of Machine Learning Algorithms For the Prediction of the Glycemic Status of the Patient B

Algorithm	*R^2*	*Mean absolute error*	*Mean squared error*	*Root mean squared erro*	*Mean signed difference*
ANN	0,03793089	0,02464794	0,00382001	0,06180622	0,00100000
PNN	0,96996654	0,00368226	0,00000000	0,01103748	0,00100000
Gradient Boosted Trees Regression	0,71567727	0,02176984	0,00127028	0,03564093	0,01547137
Simple Regression Tree	0,91064585	0,01378341	0,00000000	0,01895266	0,01228716
Random Forest Regression	0,33102591	0,02416602	0,00273247	0,05227308	0,00120760
Tree Ensemble Regression	0,29050064	0,03015130	0,00258986	0,05089066	0,01356692
Linear Regression	0,04864026	0,02449140	0,00338018	0,05813930	0,00309530
Polynomial Regression	0,01977070	0,03305683	0,00417221	0,06459261	0,01651951

Figure 3. Synthesis of the Main Results of Machine Learning Algorithms for the Prediction of the Glycemic Status of the Patient B**.**

As we can see from the Figure 4 the glycemic status of patient B is generally over the level of 150 that is considered the threshold. Effectively especially in the second part of the historical data the glycemic status of the patient B has overcome the threshold level. The prediction shows the presence of value that are very closed to the threshold level with some maximum in which the patient B exceeds the threshold values.

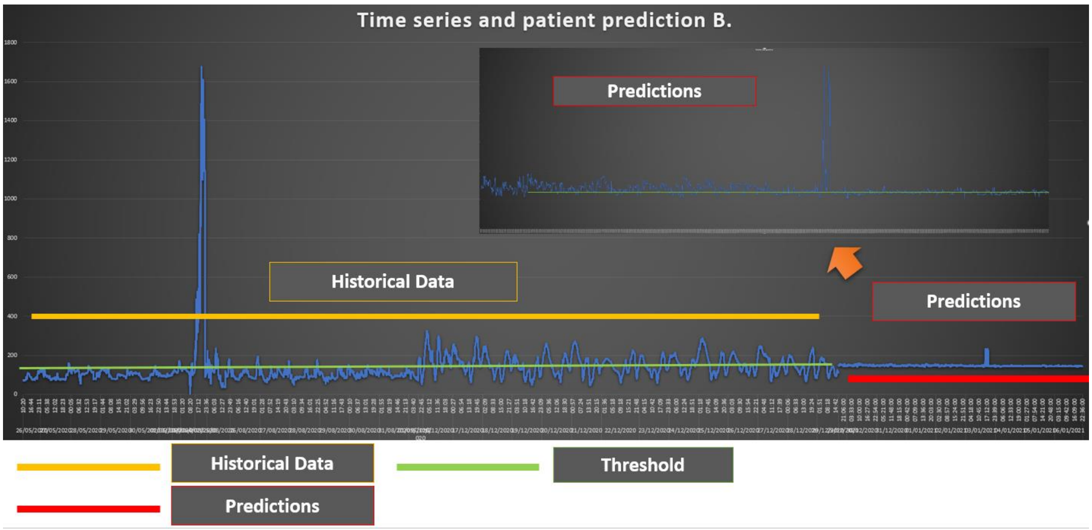
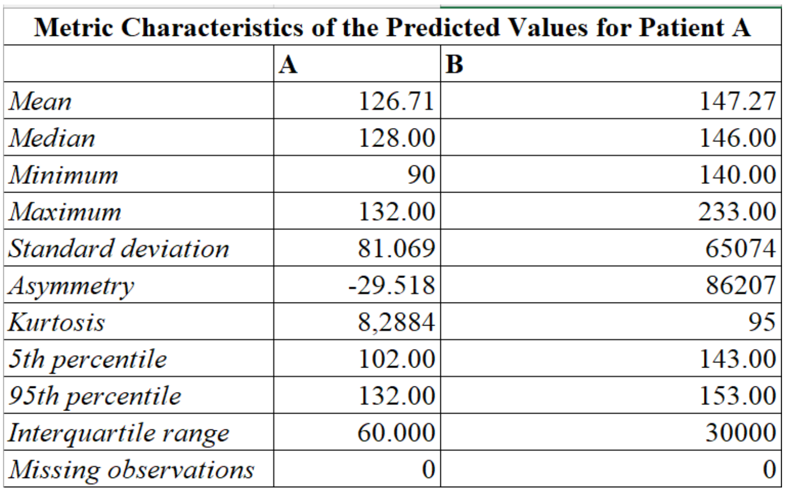


Figure 4. Time Series and Patient Prediction B.

There are many differences between the dataset of the patient A and the dataset of the patient B. Specifically, the dataset of patient A is composed of 243 observations that are timely consecutive while the dataset of the patient B is composed of 13204 discontinued observations. The analysis shows that the Artificial Neural Network-ANN with Multilayer Perceptron is more efficient in respect to PNN-Probabilistic Neural Network in the prediction applied to smaller and more timely coherent datasets. On the other hand, the PNN-Probabilistic Neural Network has a better performance in prediction in the case of larger and timely discontinued datasets.

The confrontation between the value of the prediction of patient A and the value of the prediction of patient B is presented in the Figure 5 and shows that for both the patients the level of the glycemic status is under the threshold level. The following equation has been used to de-normalize data $\boldsymbol{y} = \boldsymbol{y}' * \boldsymbol{stdev}(\boldsymbol{y}) + \boldsymbol{mean}(\boldsymbol{y})$ [5] where y′ is the predicted value and y is the value in the input dataset.

But for the case of Patient B there is a risk greater than that of the Patient A as showed in the mean, median and maximum level.



Metric Characteristics of the Predicted Values for Patient A		
	A	**B**
Mean	126.71	147.27
Median	128.00	146.00
Minimum	90	140.00
Maximum	132.00	233.00
Standard deviation	81.069	65074
Asymmetry	-29.518	86207
Kurtosis	8,2884	95
5th percentile	102.00	143.00
95th percentile	132.00	153.00
Interquartile range	60.000	30000
Missing observations	0	0

Figure 5. Metrics Characteristics of the Predicted Values for Patient A and Patient B.

As we can see from data in Figure 5, patient B has worse level of glycemic status in respect to patient B. And, in the prediction of the glycemic state of health of patient B there are data very close, and in some cases higher than the threshold values.

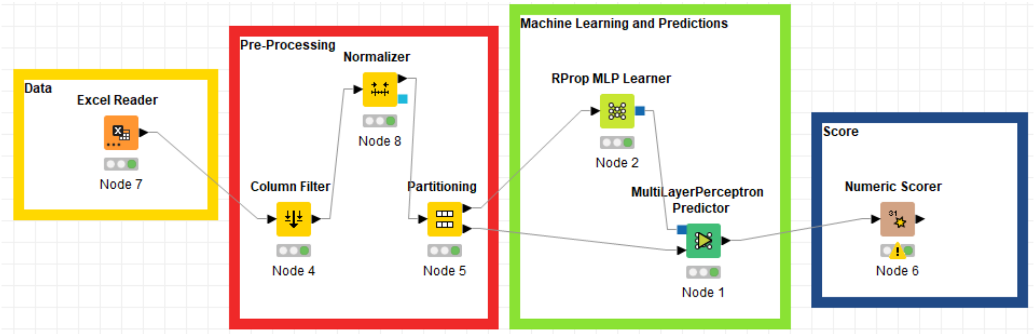


Figure 6. KNIME workflows of the Artificial Neural Network-ANN with Multilayer Perceptron.

Figure 6 shows the 4 phases of the creation of a machine learning algorithm with predictive capability based on the Artificial Neural Network-ANN with Multilayer Perceptron. As can be seen from Figure 6 there are four phases:

- *Data*: consists of a single KNIME node which has the function of reading the data entered in the Excel format;
- *Preprocessing*: consists of a group of 3 different KNIME nodes. The first node consists of "*Column Filter*" and is a node in which it is possible to select the columns of interest through which to carry out the prediction activity. The second node consists of "*Normalizer*" and it is a node that compresses data in the range from 0 to 1. The third node consists of "*Partitioning*" and is a node in which the data is divided into two different groups: 70% is used for the training of the neural network while the remaining 30% is used for the actual prediction;
- *Machine Learning and predictions:* is the central part of the data analysis process aimed at predictions and consists of two nodes. The first known is "*RProp MLP Learner"* is used for neural network training. It has hyperparameters that can be modified according to the analytical needs and which, however, in the analyzed case were used in the basic version. The second node of KNIME is "*MultilayerPerceptron Predictor*" and is the node containing the real data prediction.
- *Score*: the last phase consists of the "*Numeric Scorer*" node which allows to evaluate the predictive efficiency of the neural network through the analysis of both the R-square and the statistical errors.

3. Conclusions

In this article, eight different machine learning algorithms were used to predict the health status of two patients with blood glucose. The choice of the best predictor among the eight algorithms was made considering the performance of the algorithms in terms of maximization of the R-square and minimization of statistical errors. Subsequently, the metric characteristics of the series were analyzed to verify the trend of the performance status of both patient A and patient B. Finally, the structure of the artificial neural network used was analyzed in detail with an indication of the various KNIME nodes used. for prediction.

The use of machine learning algorithms for prediction makes it possible to identify the critical elements in the health management of patients with diabetes can also have a life-saving impact on the monitored patients.

Acknowledgments: All the applications have been deployed by a unique IT collaborative framework developed within the Smart District 4.0 Project: the italian Fondo per la Crescita Sostenibile, Bando "Agenda Digitale", D.M. Oct. 15th, 2014, funded by "Ministero dello Sviluppo Economico". This is an initiative funded with the contribution of the Italian Ministry of Economic Development aiming to sustain the digitization process of the Italian SMEs. Authors thank to the partner Emtesys s.r.l. for the collaboration provided during the work development

References

1. A. Massaro, V. Maritati, N. Savino, A. Galiano, D. Convertini, E. De Fonte and M. Di Muro, "A Study of a health re-sources management platform integrating neural networks and DSS telemedicine for homecare assistance," Infor-mation, vol. 7, no. 176, p. 9, 2018.
2. A. Massaro, V. Maritati, N. Savino and A. Galiano, "Neural networks for automated smart health platforms orient-ed on heart predictive diagnostic big data systems," 2018 AEIT International Annual Conference , no. IEEE, pp. 1-5, 2018.
3. A. Massaro, V. Maritati, D. Giannone, D. Convertini, & A. Galiano (2019) "LSTM DSS automatism and dataset op-timization for diabetes prediction", Applied Sciences, 9(17), 3532.
4. A. Massaro, Electronics in Advanced Research Industries: Industry 4.0 to Industry 5.0 Advances, John Wiley & Sons ed., 2021.
5. S. overflow, "Stack overflow," [Online]. Available: https://stackoverflow.com/questions/32888108/denormalization-of-predicted-data-in-neural-networks. [Accessed 07 01 2022].