Data Mining Statistical Methods

C&RT trees and Neural networks

Martin Sebera

Faculty of Sports Studies, Masaryk University

- Orcid: 0000-0003-3750-1549
- Scopus Author ID: 36165218300
- ResearcherID: M-9818-2018
- https://www.muni.cz/en/people/55084-martin-sebera

CONTENT

- Standards statistical methods
  - Disadvantages, limitations ☹
- Datamining methods – classification, regression
  - Benefits ☺
  - C&RT trees, MLP Neural networks
- Examples

Short list of frequently used methods

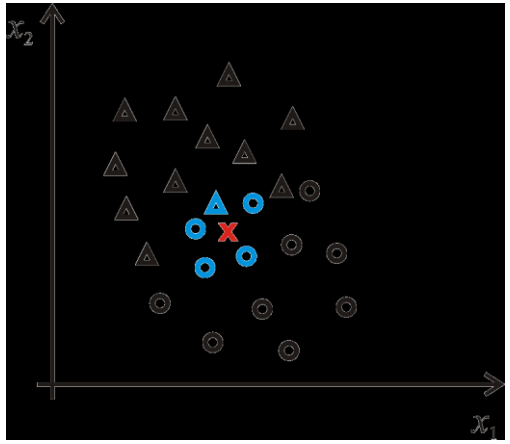
- Basic statistical characteristics (mean, median, standard deviation, …)
- t-test and ANOVA (analysis of variance)
- Correlation analysis and Factor analysis
- Linear regression and Test chí2
- assumptions: normality of data, homogenity of variances
  ($\rightarrow$ parametric vs. nonparametric methods)
  nominal, ordinal, categorical variables cannot be combined in model

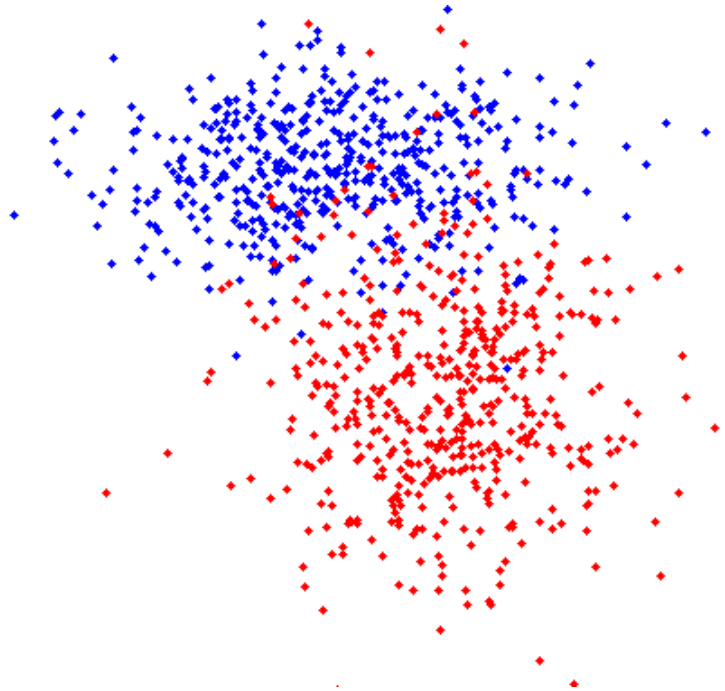
Classification and regression

- Classification - the process of classifying patterns into given classes based on the features of the classified object
- Regression - the process of interleaving function data

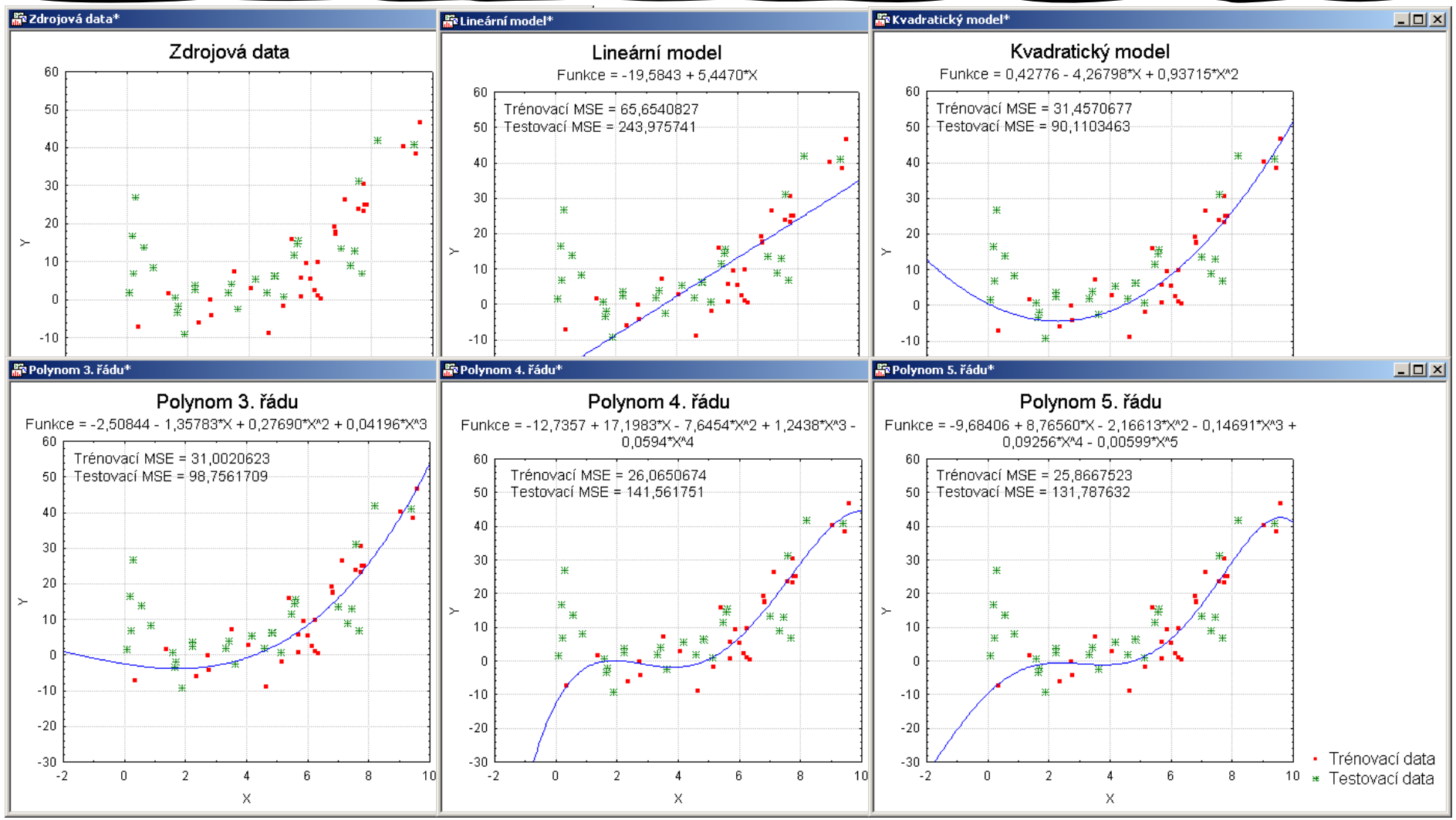
Classification

- E.g. linear classifier, k-NN (k-nearest neighbors) classifier, trees, neural networks, ...





Regression (approximation)



Classification and regression tools

- Tools that are used for regression can almost always be used for classification.
- Not all classification tools are usable for regression.
- Artificial neural networks (MLP, RBF,...)
- Trees and their variants
- Genetic and evolutionary techniques
- Classical discriminant methods - k-nearest neighbors, Bayesian classifier, Support-vector machine (SVM), ...

Data preparation

- Editing data to the standard dataset format

	x_1	x_2	…	x_n	d_1	d_2	…	d_m
s_1	0,84	0,96	…	0,14	0,99	0,53	…	1
s_2	0,51	0,04	…	0,12	0,78	0,23	…	0
s_3	0,62	0,21	…	0,87	0,25	0,57	…	1
…	…	…	…		…	…	…	…
s_N	0,37	0,83	…	0,17	0,64	0,09	…	1

Data preprocessing

- Filtering - most often noise (sliding averaging, median filter - impulse noise removal)
- Completion of missing data - only in an emergency, it is supplemented on the basis of the average resp. frequency
- Normalization - the goal is to unify the numerical ranges of values

Data preparation

- Dividing data into training, validation and testing set
- Training set - known output → classifier learning
- Validation set - known output, but we will not provide it to the classifier (comparison of the classification result with the real output) → validation
- Test set - known output, we measure the success of the classifier

Classification and regression trees

Principle: Gradual hierarchical dividing of data space into subgroups so that in the leaves of the created tree there are (homogeneous) groups of data belonging (in case of classification) to one class.

- Based on the gradual division of the symptom space (similar to searching in the botanical key).
- Classify an object into the corresponding class based on flags
- Simple, fast
- Easy visualization → good interpretation of results
- More resistant to outliers and missing values.

Classification accuracy

We can influence the accuracy of the classification and thus the structure of the tree:

- misclassification cost matrix
- a priori probabilities of representation of individual classes (priors)
- proportional representation of patterns of individual classes in the data (case weights, count variable)
- change of one parameter affects the others (different expressions of the same)

Classification accuracy

- ROC curve (Receiver Operating Characteristics) - area under the plotted curve - quality of the classification
- graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

Sensitivity = TP / (TP + FN), Specificity = TN / (FP + TN),

Positive predictive value = TP / (TP + FP),

Negative predictive value = TN / (FN + TN),

Efficiency = (TP + TN) / (TP + TN + FP + FN)

TP – true positive, TN – true negative, FP – false positive, FN – false negative

Classification trees (Nonparametric method)

- Samples are classified linearly and hierarchically into a finite (small) predetermined number of classes.
- Sequence of decisions.
- The root contains the entire data file.
- Two (binary tree) or more branches grow from each node. Each sheet represents one of the groups.
- The creation: choose a variable that divides the data into the most homogeneous subgroups possible.

Classification trees (Nonparametric method)

- Stopping the growth of the tree. There are ways to:
  - further subdivision is not statistically significant
  - the size of the error in the subnodes (growth stops when the percentage success of the incorrect classification exceeds the specified value)
  - number of samples in the end node
  - number of terminal nodes

C&RT

- C&RT is a binary tree build by splitting node into two child nodes repeatedly.
- Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).
- The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.

Example 1

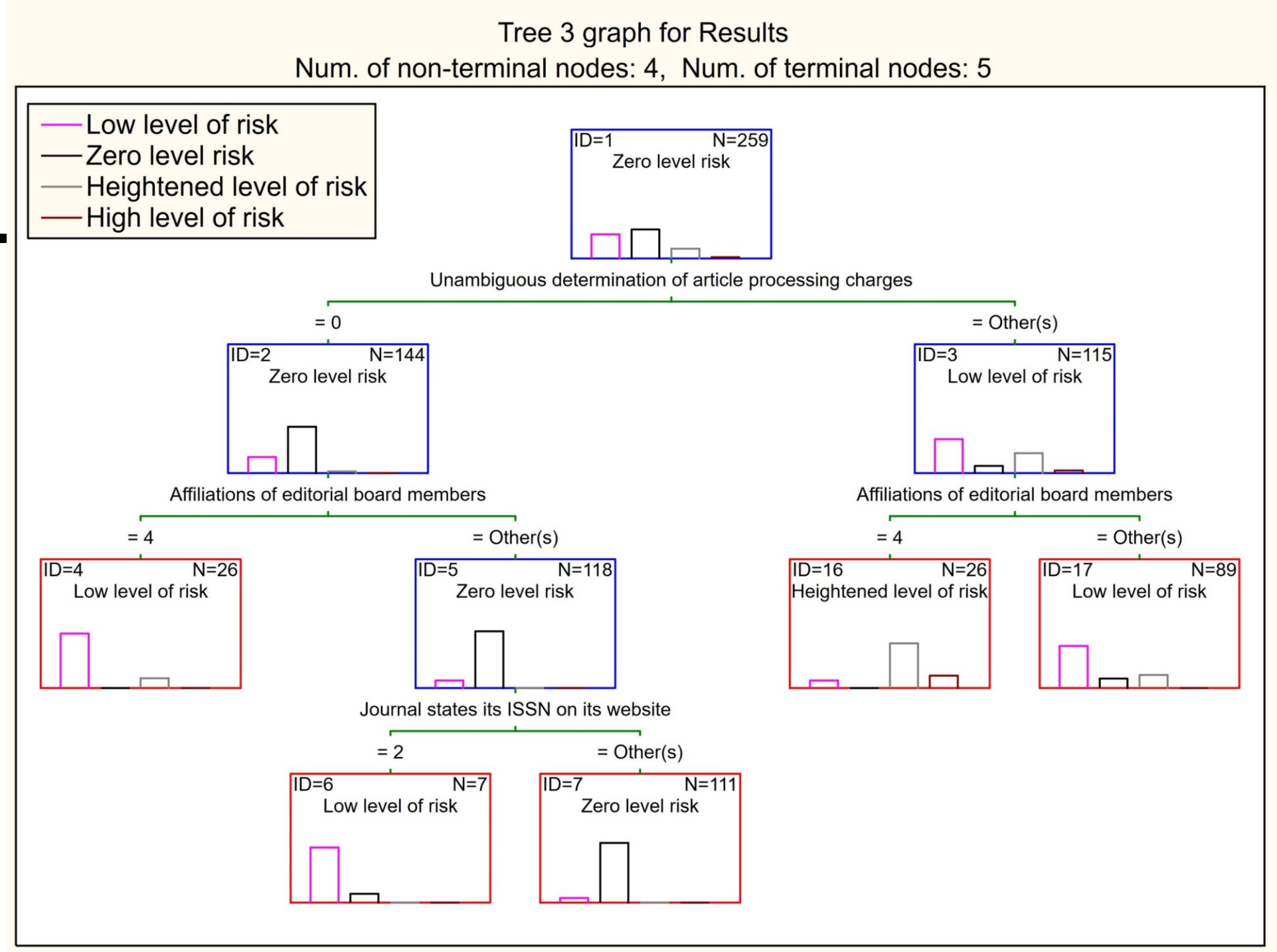
Kratochvíl, J., Plch, L., Sebera, M., & Koritáková, E. (2020). Evaluation of untrustworthy journals: Transition from formal criteria to a complex view. *Learned Publishing*, 33(3), 308–322. https://doi-org.ezproxy.muni.cz/10.1002/leap.1299

Affiliations of editorial board members	Unambiguous determination of article processing charges	Description of the review process	Accurate information about the journal's citation metrics in Journal Citation Reports and Scopus	Accurate information about the journal's indexing in Web of Science and Scopus	Journal states its ISSN on its website	Free and open access to full text	Proclamation of dubious metrics/databases	TOTAL	Results
1	2	0	0	0	0	0	0	3	Heightened level of risk
0	2	1	0	0	0	0	0	3	Heightened level of risk
0	0	0	0	0	0	0	0	0	Zero level risk
0	0	0	0	0	0	0	0	0	Zero level risk
0	0	0	0	0	0	0	0	0	Zero level risk
1	1	0	0	0	0	0	0	2	Low level of risk
0	0	0	0	0	0	0	0	0	Zero level risk
0	1	0	0	0	0	0	0	1	Low level of risk
0	0	0	0	0	0	0	0	0	Zero level risk
1	2	1	0	0	0	0	0	4	Heightened level of risk
1	0	0	0	0	0	0	0	1	Low level of risk
0	0	0	0	0	0	0	0	0	Zero level risk
2	0	0	0	0	0	0	0	2	Low level of risk
0	2	0	0	0	0	0	0	2	Low level of risk
0	0	0	0	0	0	0	0	0	Zero level risk
0	0	0	0	0	0	0	0	0	Zero level risk
0	0	0	0	0	0	0	0	0	Zero level risk

Predatory journals: criteria

- Unambiguous determination of article processing charges
- Affiliation of editorial board members
- Journal ISSN on its website
- Review Time
- Description of Peer-Review
- Proclamation of indexing WoS/Scopus/ERIH/Medline
- Accessibility of full texts
- Publisher

- An evaluation of 259 biomedical journals
- using the list of 8 criteria
- The most common reason for failure to comply was:
  - sufficient editorial information and
  - declaration of article processing charges.


Tree 3 graph for Results
Num. of non-terminal nodes: 4, Num. of terminal nodes: 5
Low level of risk
Zero level risk
Heightened level of risk
High level of risk
ID=1 N=259
Zero level risk
Unambiguous determination of article processing charges
= 0
= Other(s)
ID=2 N=144
Zero level risk
ID=3 N=115
Low level of risk
Affiliations of editorial board members
Affiliations of editorial board members
= 4
= Other(s)
= 4
= Other(s)
ID=4 N=26
Low level of risk
ID=5 N=118
Zero level risk
ID=16 N=26
Heightened level of risk
ID=17 N=89
Low level of risk
Journal states its ISSN on its website
= 2
= Other(s)
ID=6 N=7
Low level of risk
ID=7 N=111
Zero level risk

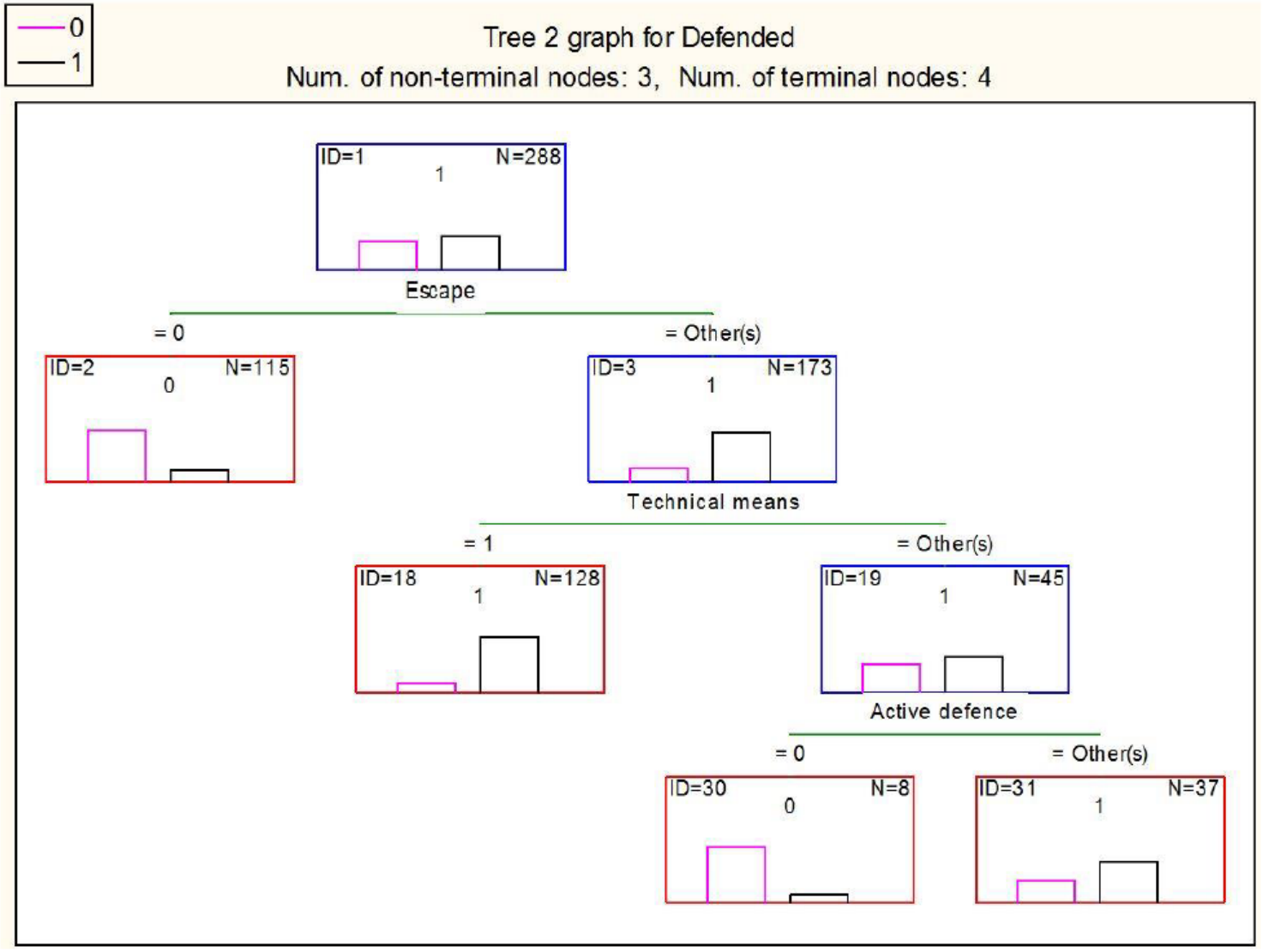
EXAMPLE 2

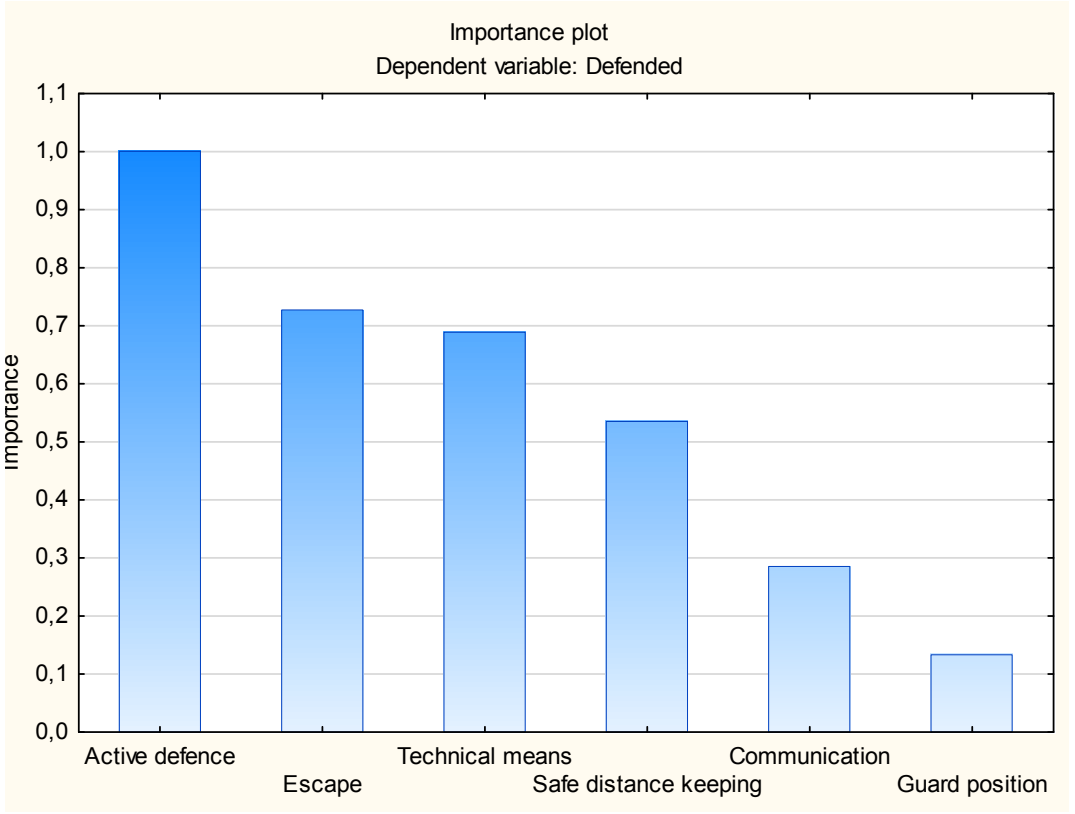
Vít M., Reguli Z., Sebera M., Cihounková J., & Bugala M. (2016). Predictors of children s successful defence against adult attacker. *Archives of Budo. (12), p. 141-150*.

- The paper is based on the presumption that the probability of successful defence of a child against an adult attacker is influenced by diversity of variables with different predictive values.

Object	Expert	Guard position	Communication	Safe distance keeping	Active defence	Technical means	Escape	defended
A_002	E1	0	0	0	1	0	0	0
A_003	E1	0	0	0	1	0	0	0
A_004	E1	0	0	0	0	0	0	0
A_005	E1	1	0	1	1	1	0	0
A_006	E1	0	0	0	1	0	0	0
A_007	E1	0	0	0	1	0	1	1
A_008	E1	1	0	1	1	0	0	1
A_009	E1	0	0	1	1	1	0	0
A_010	E1	0	0	1	1	1	0	1
A_011	E1	1	1	1	1	1	0	0
A_013	E1	0	0	1	1	0	1	1
A_014	E1	0	0	1	1	1	1	1
A_015	E1	0	1	1	1	1	1	1
A_017	E1	0	0	0	1	1	0	0
A_018	E1	0	0	0	0	0	0	0
A_019	E1	0	0	0	0	0	0	0

288 defense situations were evaluated by 6 self-defense experts in 6 criteria.





- **The best predictors:** Active defence, Escape and Technical means
- Communication and Safe distance keeping varied in the fifth position.
- Guard position was found the weakest predictor.

Neural network (NN)

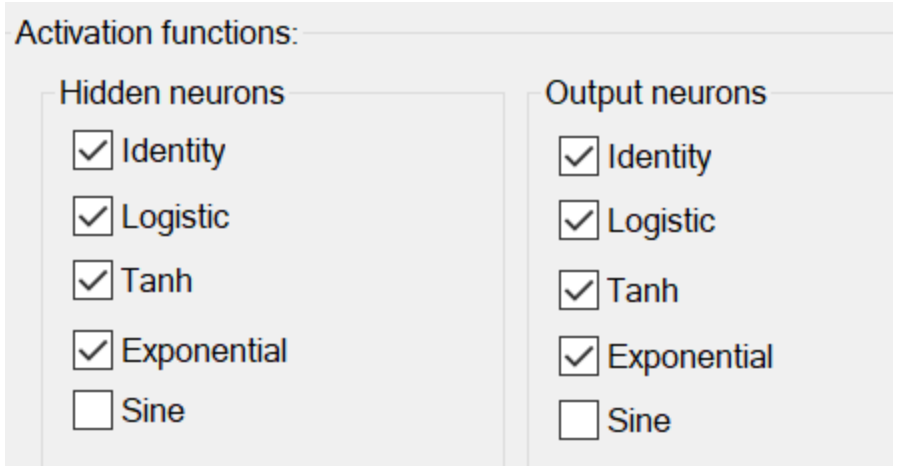
- A tool for nonlinear modeling.
- Many inputs generate an output that is a nonlinear function of the weighted sum of these inputs.
- The weights assigned to each of the inputs are obtained on the basis of a learning process, where the generated outputs are compared with the so-called target outputs.
- The obtained deviations between the known values and the obtained outputs serve as feedback for the adjustment of the weights.

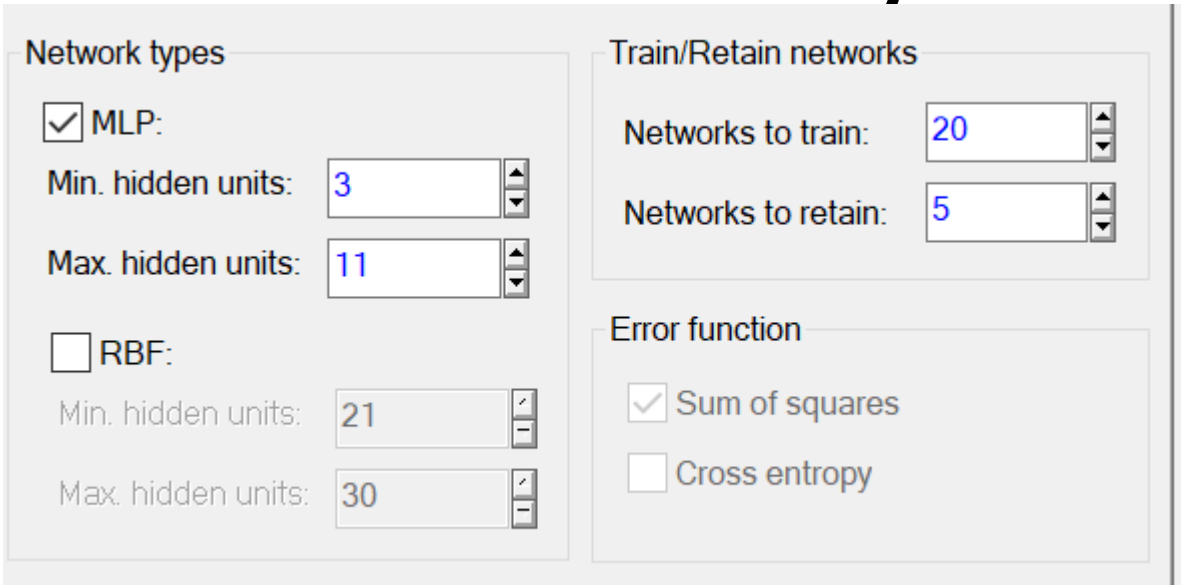
Neural network

- NN is a method in artificial intelligence
- NN teaches computers to process data in a way that is inspired by the human brain.
- It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.
- In other words, it is a very complex regression, where I have one dependent and many independent

Neural network - MLP

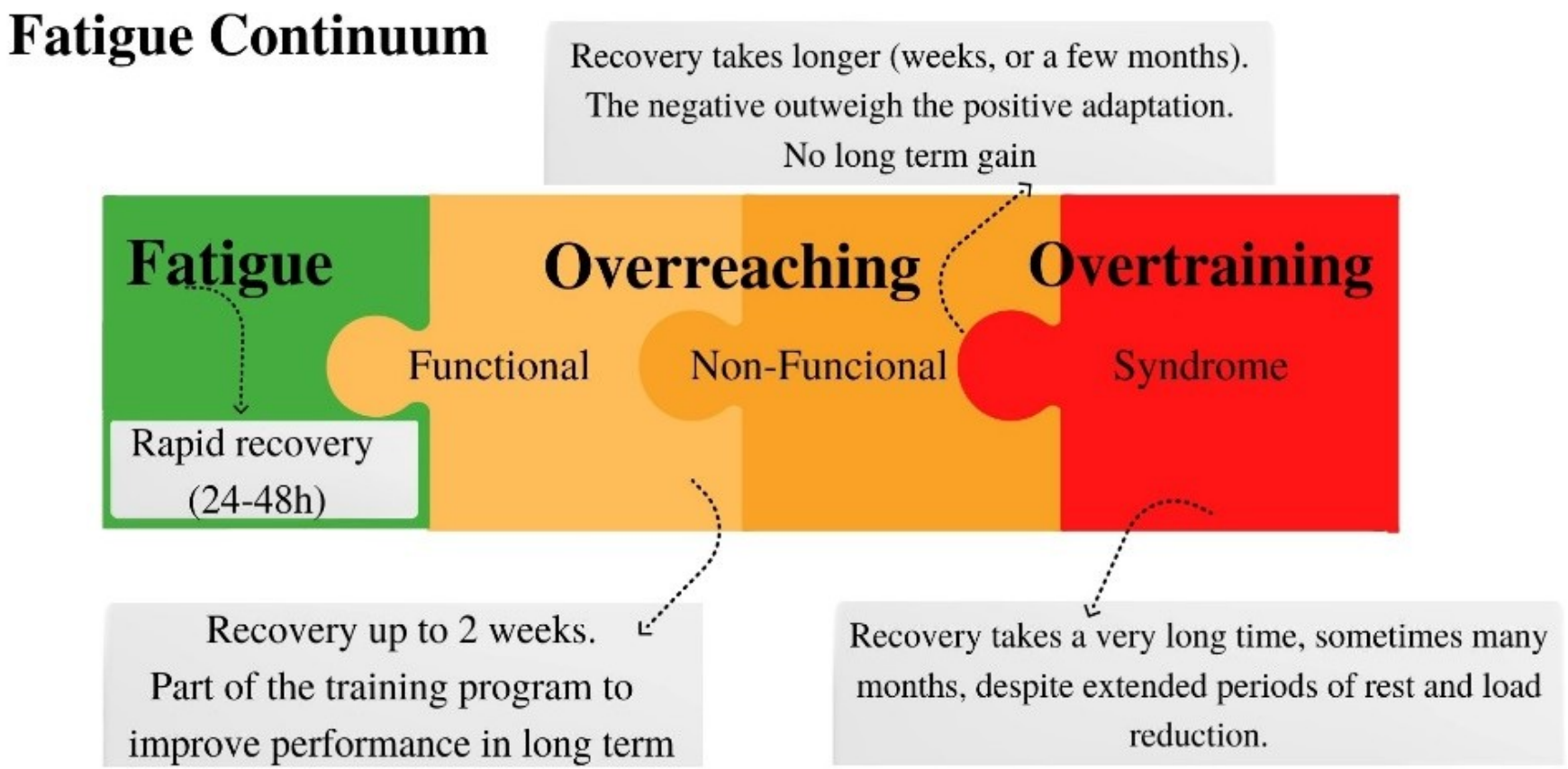
- Multilayer Perceptron (MLP): class of feed-forward neural networks
- 3 types of layers - the input layer, output layer and hidden layer
- Activation Functions: **defines how the weighted sum of the input is transformed into an output from a node or nodes in a layer of the network**

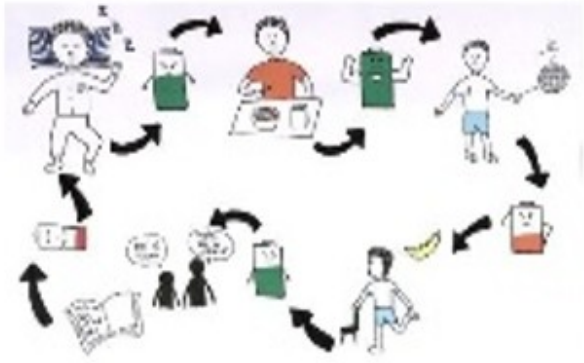




EXAMPLE 3 - Bernaciková M., Kumstát M., Buresová I., Kapounková K.,Struhár I., Sebera M. & Paludo, A. C. (2022). Diagnosing and preventing chronic fatigue in Czech youth athletes: mobile application

Frontiers in Physiology, section Exercise Physiology. In press

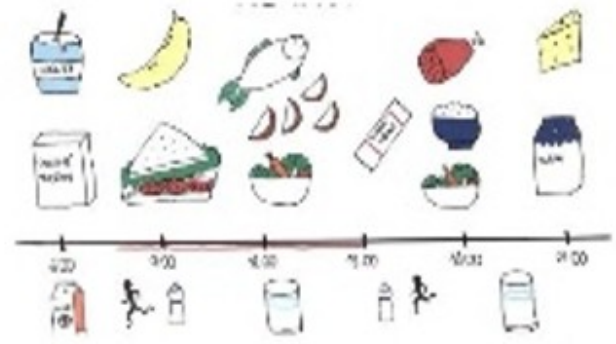




INTRODUCTION

SLEEP

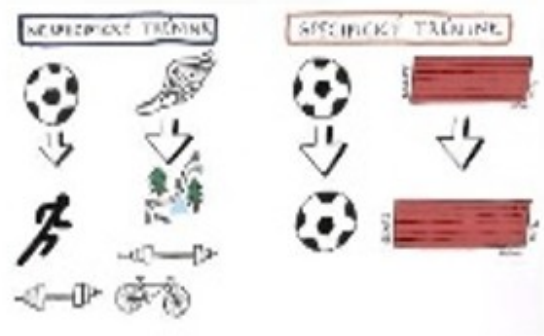
NUTRITION



NUTRITION AND TRAINING

REGENERATION

COMMUNICATION



TRAINING PARAMETERS

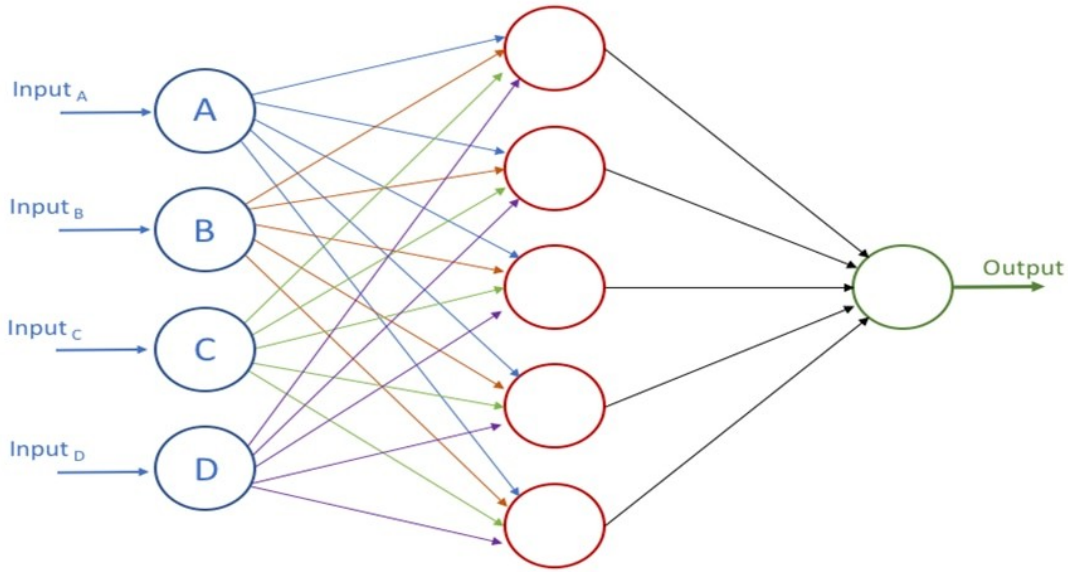
MONITORING PROCESS

Bernaciková M., Kumstát M., Buresová I., Kapounková K.,Struhár I., Sebera M. & Paludo, A. C. Diagnosing and preventing chronic fatigue in Czech youth athletes: mobile application

type of sport (cycling, football, ice hockey, gymnastics, swimming)
another sport (yes / no)
regeneration (yes / no)
1 day available (yes / no)
frequent illness (yes / no)
Injuries (yes / no)
disease / condition IMUNO (yes / no)
food. Intolerance (yes / no)
trainings of the week
training length (hours)
total number of hours / week
tournaments, races / year
sports total number of hours / week
sleep (1-5)
nutrition: appetite (1-5)
energy (1-5)
appetite for training (1-5)
training quality (1-5)
degree of risk of fatigue (1-10)

Dependent
Continuous
Categorical

MLP 30-11-1



Bernaciková M., Kumstát M., Buresová I., Kapounková K.,Struhár I., Sebera M. & Paludo, A. C. Diagnosing and preventing chronic fatigue in Czech youth athletes: mobile application

type of sport (cycling, football, ice hockey, gymnastics, swimming)
another sport (yes / no)
regeneration (yes / no)
1 day available (yes / no)
frequent illness (yes / no)
Injuries (yes / no)
disease / condition IMUNO (yes / no)
food. Intolerance (yes / no)
trainings of the week
training length (hours)
total number of hours / week
tournaments, races / year
sports total number of hours / week
sleep (1-5)
nutrition: appetite (1-5)
energy (1-5)
appetite for training (1-5)
training quality (1-5)
degree of risk of fatigue (1-10)

MLP 30-11-1

Dependent
Continuous
Categorical

Input A
Input B
Input C
Input D

Sensitivity analysis	Average
regeneration	23 024 168
type of sport	559 030
1 free day	364 464
disease / condition IMUNO	266 686
appetite for training	9 420
appetite	8 264
training quality	7 846
energy	7 835
sleeping	6 980
tournaments, races / year	4 459
total number of hours / week	672
trainings / week	115
training length (hours)	49
food intolerance	1
Injuries	1
frequent illness	1
another sport	1

```
double algoritmus_vyplneno_160220c_5_MLP_30_8_1_MeanInputs[9]={ 5.32000000000000e+000, 2.30000000000000e+000, 1.35051020408163e+001, 4.22340425531915e+001, 2.9591836

void algoritmus_vyplneno_160220c_5_MLP_30_8_1_ScaleInputs(double* input, double minimum, double maximum, int size)
{
 double delta;
 long i;
 for(i=0; i<size; i++)
 {
    delta = (maximum-minimum)/(algoritmus_vyplneno_160220c_5_MLP_30_8_1_max_input[i]-algoritmus_vyplneno_160220c_5_MLP_30_8_1_min_input[i]);
    input[i] = minimum - delta*algoritmus_vyplneno_160220c_5_MLP_30_8_1_min_input[i]+ delta*input[i];
 }
}

void algoritmus_vyplneno_160220c_5_MLP_30_8_1_UnscaleTargets(double* output, double minimum, double maximum, int size)
{
  double delta;
  long i;
  for(i=0; i<size; i++)
  {
    delta = (maximum-minimum)/(algoritmus_vyplneno_160220c_5_MLP_30_8_1_max_target[i]-algoritmus_vyplneno_160220c_5_MLP_30_8_1_min_target[i]);
    output[i] = (output[i] - minimum + delta*algoritmus_vyplneno_160220c_5_MLP_30_8_1_min_target[i])/delta;
   }
}

void algoritmus_vyplneno_160220c_5_MLP_30_8_1_ComputeFeedForwardSignals(double* MAT_INOUT,double* V_IN,double* V_OUT, double* V_BIAS,int size1,int size2,int layer)
{
  int row,col;
  for(row=0;row < size2; row++)
    {
      V_OUT[row]=0.0;
      for(col=0;col<size1;col++)V_OUT[row]+=(*(MAT_INOUT+(row*size1)+col)*V_IN[col]);
      V_OUT[row]+=V_BIAS[row];
      if(layer==0) V_OUT[row] = exp(V_OUT[row]);
   }
}

void algoritmus_vyplneno_160220c_5_MLP_30_8_1_RunNeuralNet_Regression ()
{
  algoritmus_vyplneno_160220c_5_MLP_30_8_1_ComputeFeedForwardSignals((double*)algoritmus_vyplneno_160220c_5_MLP_30_8_1_input_hidden_weights,algoritmus_vyplneno_16022
  algoritmus_vyplneno_160220c_5_MLP_30_8_1_ComputeFeedForwardSignals((double*)algoritmus_vyplneno_160220c_5_MLP_30_8_1_hidden_output_wts,algoritmus_vyplneno_160220c_
}

int main()
```

Software

- STATISTICA
  - ❑TIBCO Software Inc. (2020). Data Science Workbench, version 14. http://tibco.com.
- SPSS 28
  - ❑IBM SPSS Statistics, 28.0.0.0 (190)
- I have to learn „R“
  - ❑language and environment for statistical calculations, https://www.r-project.org/

Sources

- Kratochvíl, J., Plch, L., Sebera, M., & Koritáková, E. (2020). Evaluation of untrustworthy journals: Transition from formal criteria to a complex view. *Learned Publishing*, 33(3), 308–322. https://doi-org.ezproxy.muni.cz/10.1002/leap.1299
- Vít M., Reguli Z., Sebera M., Cihounková J., & Bugala M. (2016). Predictors of children s successful defence against adult attacker. *Archives of Budo. (12), p. 141-150.*
- Bernaciková M., Kumstát M., Buresová I., Kapounková K.,Struhár I., Sebera M. & Paludo, A. C. (2022). Diagnosing and preventing chronic fatigue in Czech youth athletes: mobile application. *Frontiers in Physiology, section Exercise Physiology*. In press.

Conclusion

Everything here are statistical games ☺

Remember the most important and difficult thing of all statistical calculations is:

factual interpretation of the results !

Hvala na pažnji

Thank you for your attention