Journal of KONBiN 2023
Volume 53, Issue 1
DOI 10.5604/01.3001.0016.3235

Dariusz AMPUŁA ORCID 0000-0002-9036-9498
Military Institute of Armament Technology (Wojskowy Instytut Techniczny Uzbrojenia)

PREDICTIVE DATA MINING MODELS IN THE TESTS OF PROPELLING CHARGES

Predykcyjne modele data mining w badaniach ładunków miotających

Abstract: *In the article, in the introduction, the concept of predictive data mining models and was defined and the purpose of the article was specified. Then, the method of building predictive models was characterized and the elements of ammunition were indicated, the test results of which were prepared for the building of models, and the types of ammunition in which the propellant charge is present were indicated. The results of building four data mining models are presented. Predictive models for C&RT, CHAID and exhaustive CHAID decision trees were designed and built. The fourth model analyzed was the SANN model, i.e. the model of neural networks. For each of the tree models, a schema of the designed tree, the rate of false predictions and the parameters of goodness of fit of the built models are shown. For the SANN model, the parameters of the selected neural network were additionally characterized. An analysis of the built models was made and, based on the obtained results, the best designed predictive data mining model was indicated. At the end, the graphical form of the workspace predefined by the GC Advanced Comprehensive Classifiers project is shown.*

Keywords: artificial intelligence, predictor, charge, predictive model, tests

Streszczenie: *W artykule we wstępie zdefiniowano pojęcie predykcyjnych modeli data mining oraz określono cel artykułu. Następnie, scharakteryzowano metodę budowy modeli predykcyjnych oraz wskazano elementy amunicji, których wyniki badań zostały przygotowane do budowy modeli a także wskazano rodzaje amunicji w których występuje przedmiotowy ładunek miotający. Przedstawiono wyniki budowy czterech modeli predykcyjnych data mining. Zaprojektowano oraz zbudowano predykcyjne modele dla drzew decyzyjnych typu C&RT, CHAID oraz wyczerpujący CHAID. Czwartym analizowanym modelem był model SANN czyli model sieci neuronowych. Dla każdego z modeli drzew przedstawiono schemat zaprojektowanego drzewa, stopę błędnych przewidywań oraz pokazano parametry dobroci dopasowania zbudowanych modeli. Dla modelu SANN scharakteryzowano dodatkowo parametry wybranej sieci neuronowej.*

Dokonano analizy zbudowanych modeli oraz na podstawie otrzymanych wyników, wskazano najlepszy zaprojektowany predykcyjny model data mining. Na końcu pokazano graficzną postać przestrzeni roboczej predefiniowaną projektem GC Advanced Comprehensive Classifiers.

Słowa kluczowe: sztuczna inteligencja, predyktor, ładunek, model predykcyjny, badania

1. Introduction

As predictive data mining models, we will consider any type of designed statistical model that allows you to make predictions for new test results obtained, i.e. in practice it has a module that allows you to make this prediction for new observations. Without having such a module, the built model cannot be called predictive.

The aim of this article was to show the possibility of designing and building predictive data mining models for qualitative dependent variables for the tested 100 mm propellant charges. These charges [4, 7] have the largest number of test results, thanks to which the designed predictive models will be the most real. Four predictive data mining models based on Statistica software were designed in the article. A detailed analysis of the built models was carried out and the best obtained model was indicated, which can be used during the analysis of the new test results obtained. The remaining three models built can also be used in practice, but it should be remembered that the use of these models will introduce larger evaluation errors when making predictions of new observations, which may lead to determining incorrect values of the dependent variable sought.

2. Build of predictive models

The software [8] allows us to design and build several predictive data mining models. In our case, we will use the modules: standard C&RT classification trees with implementation, standard CHAID classification with implementation, exhaustive CHAID for classification CHAID with implementation and SANN (Statistica Automatic Neural Networks) for classification with implementation, i.e. from the neural network module. The acronym C&RT (Breiman, Friedman, Olshen and Stone 1984) is a classic decision tree algorithm, the acronym CHAID stands for Chi-squre Automatic Interaction Detection, (1980) and the exhaustive model CHAID (Biggs, de Ville, 1991) it is often denoted with the acronym XAID (Exhaustive CHAID). The process of designing and building the above-mentioned models requires extensive practical knowledge from the designer, because it is quite advanced technically and substantively.

When designing our predictive models, we will use the so-called a workspace, which will graphically show us how to proceed. This space is predefined by the GC Advanced Comprehensive Classifiers project, which represents the starting node to which we connect the input data. These data are the results of diagnostic tests of tested 100 mm propellant charges, defined as input predictors for individual predictive models.

The propelling charge [9] is the part of the cartridge that is designed to impart a certain energy to the projectile when fired from a barreled weapon. The source of the propellant energy are the gases formed from the burning of powders. The propellant charge consists of a certain amount of powder placed in a predetermined order along with other components of the charge in a case or bag. 100 mm caliber propellant charges are used in 100 mm caliber ammunition, in cartridges with a fragmentation, high-explosive, hydro-blast, armor-piercing-tracer and flash-smoke projectiles.

In reserve in the warehouses of the military units of the Polish Army still contain 100 mm caliber cartridges, although a few years ago the administrator carried out the disposal of a part of this caliber by selling it to a specific entrepreneur, but later the sale process was under the microscope of control administration. Each attempt to use this ammunition or even its dispose of it requires the administrator of this ammunition to carry out laboratory diagnostic tests, the purpose of which is to determine its current technical condition. One of the tested elements of this ammunition is its propellant charge, which is tested in accordance with the procedures included in the test methodology [10].

The dependent variable in all models is the DEC variable, which in this case can take six values (B5, B3, BP, BS, PS and R). This dependent variable is the postdiagnostic decision made on the basis of the obtained results of tests of particular characteristics of propellant charges. A detailed description of the possible postdiagnostic decisions can be found in the test methodology [10].

All the tested characteristics of the propellant charges were divided, in accordance with the test methodology, into three classes of importance: A, B and C. Depending on the inconsistencies detected during laboratory tests, the tested lot of propellant charges receives a specific postdiagnostic decision.

In the designed predictive data mining models for 100 mm propellant charges for the first laboratory diagnostic tests, three predictors were adopted in accordance with the test methodology, which were: the number of inconsistencies in the importance class A (LA), the number of inconsistencies in the importance class B (LB) and the number of inconsistencies in the C (LC) importance class.

A very important aspect when conducting analyzes is the fact that in the practical applications of predictive data mining, it is necessary to carefully and carefully check the input data before performing this analysis to be sure that the input data does not incorrect numerical values or incorrectly coded values.

Observations (test results) that were not used to build the model itself will be used to evaluate the already built model. Thanks to such a procedure, on the basis of specific goodness of fit statistics, it will be possible to reliably assess the prognostic validity (accuracy) of each built model and thus compare the models and thus determine the best designed model.

3. The result of building predictive models

The article presents the designed and then built predictive data mining models for the tested 100 mm propellant charges. Our models were classification type models because the dependent variable was the DEC variable denoting the adopted postdiagnostic decision. Quantitative predictors were the obtained test results (observations) of the individual tested features of 100 mm propellant charges.

When designing our predictive data mining models for 100 mm propellant charges, we started with a standard C&RT classification tree [2]. The analyzed test sample of the test results was randomly divided into the learning sample and the test sample. This division was obligatory for each data mining model created. These models will be fitted to the created learning sample and will be assessed using the test sample. In order to specify this split, the predictors and the dependent variable have already been defined, which will be the same for all predictor models.

The working space for all four predictive data mining models is shown in Fig. 1. We see the input data 100 mm powder charges, the starting node, the two resulting sets of observations: learning data and test data, and four data mining predictive models.

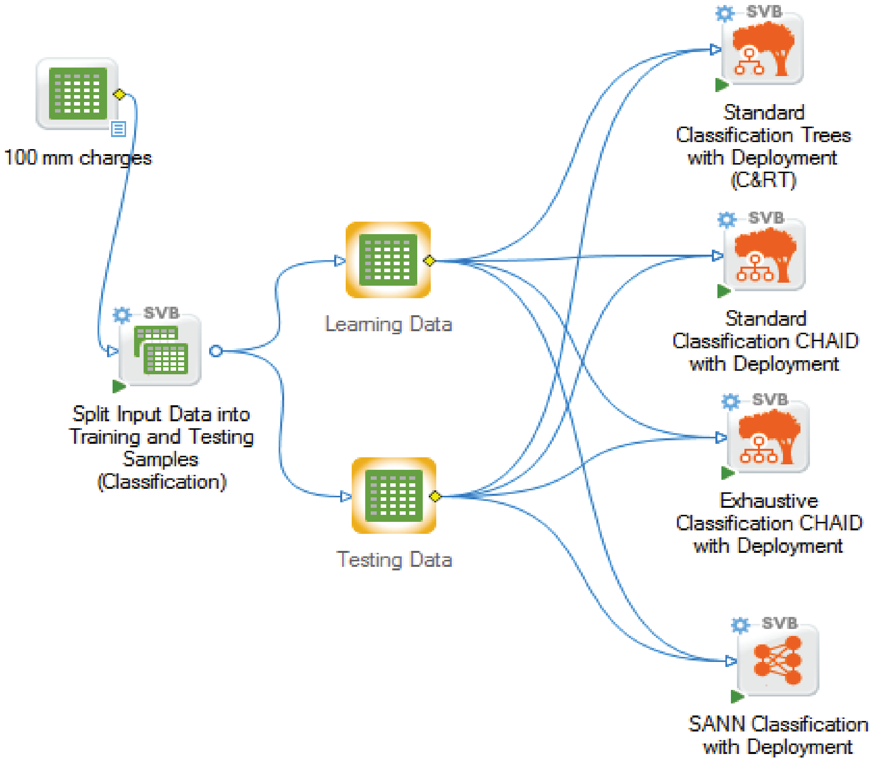

SVB
100 mm charges
SVB
Split Input Data into Training and Testing Samples (Classification)
Learning Data
Testing Data
SVB
Standard Classification Trees with Deployment (C&RT)
SVB
Standard Classification CHAID with Deployment
SVB
Exhaustive Classification CHAID with Deployment
SVB
SANN Classification with Deployment

Fig. 1. Workspace for designed predictive models

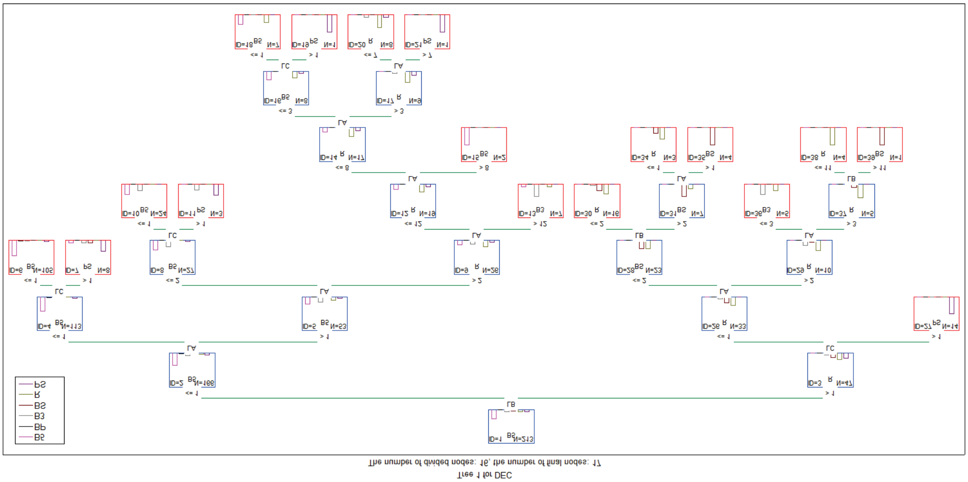

Tree 1 for DEC
The number of divided nodes: 16, the number of final nodes: 17

Fig. 2. Tree schema for the C&RT model

The software [8], after building a predictive model of standard C&RT classification trees, will automatically generate the information needed to implement this model, which allows you to classify new observations using this model. After building our C&RT model, we received a classification tree [6], the schema of which is shown in Fig. 2. This tree has sixteen divided nodes and seventeen final nodes - leaves. So it is quite an expanded C&RT tree.

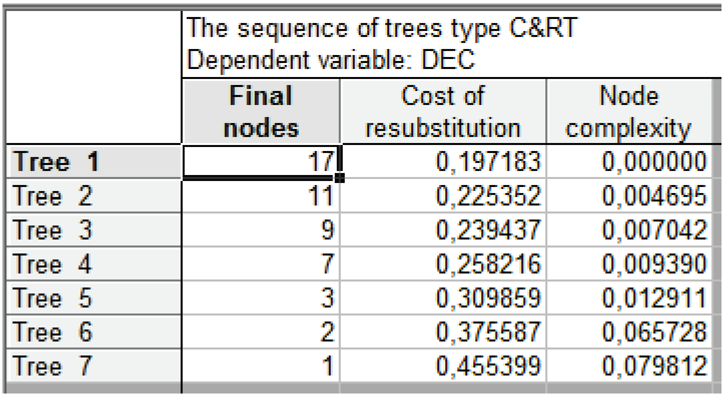
The sequence of tree shown in Table 1 indicates that the built tree 1 was chosen as the best due to the lowest value of the cost of resubstitution, which for this tree is 0.197183. The cost of resubstitution in classification problems evaluates the accuracy of the built model. It is the proportion of cases incorrectly classified by the classification model built on the basis of all cases.

After building this model, a large number of documents appeared in the workspace, including containing the predicted and observed values for the cases from the test sample, which were placed in a sheet called Testing PMML CTrees (Fig. 5). Additionally, the final results and analyzes for this model have been placed in the report documents workbook. The calculated false prediction rate for this model used to classify the cases from the test sample is 0.262443438914027 (taken from the Documents of the report).

One of the tools for evaluating the designed model is the goodness of fit, which will be calculated for many data sources, i.e. for each built model separately. The summary of the goodness of fit for the C&RT type model is shown in Table 2, which shows that percentages discordance of this model i.e. [5] the percentage of inconsistent classifications is 26.5766. The chi-square statistic [5] is used to assess the independence of the tested variables, while the G-square statistic [5] is another measure of fit, it is the equivalent of the chi-square statistic but based on the likelihood ratio.

Table 1

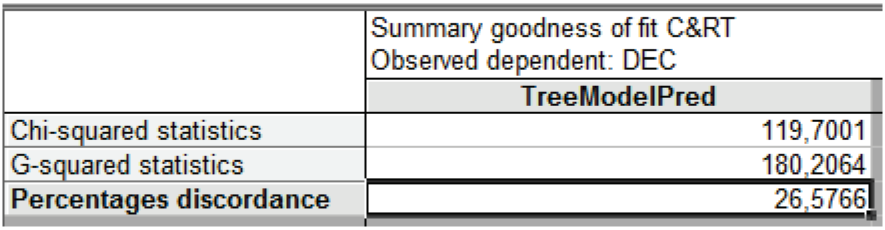
Sequence of trees for the C&RT model



	The sequence of trees type C&RT Dependent variable: DEC		
	Final nodes	Cost of resubstitution	Node complexity
Tree 1	17	0,197183	0,000000
Tree 2	11	0,225352	0,004695
Tree 3	9	0,239437	0,007042
Tree 4	7	0,258216	0,009390
Tree 5	3	0,309859	0,012911
Tree 6	2	0,375587	0,065728
Tree 7	1	0,455399	0,079812

Table 2

Summary goodness of fit for the C&RT model



	Summary goodness of fit C&RT Observed dependent: DEC
	TreeModelPred
Chi-squared statistics	119,7001
G-squared statistics	180,2064
Percentages discordance	26,5766

The goodness of fit statistics reliably assesses the prognostic validity (accuracy) of each built model and thus gives the opportunity to compare the created models.

The next model designed and built is the standard CHAID classification model [3]. The model was developed on the same data results as before. The procedure of building this model is the same as in the case of the C&RT model.

As a result of building our predictive model of the CHAID type, we obtained a tree [6], the schema of which is shown in Fig. 3. The tree consists of only two divided nodes and three final nodes. So it is a very simple decision tree.

The false prediction rate for this model used to classify the test sample cases is 0.280542986425339 (taken from the Documents of the report) and is slightly higher than for the C&RT model.

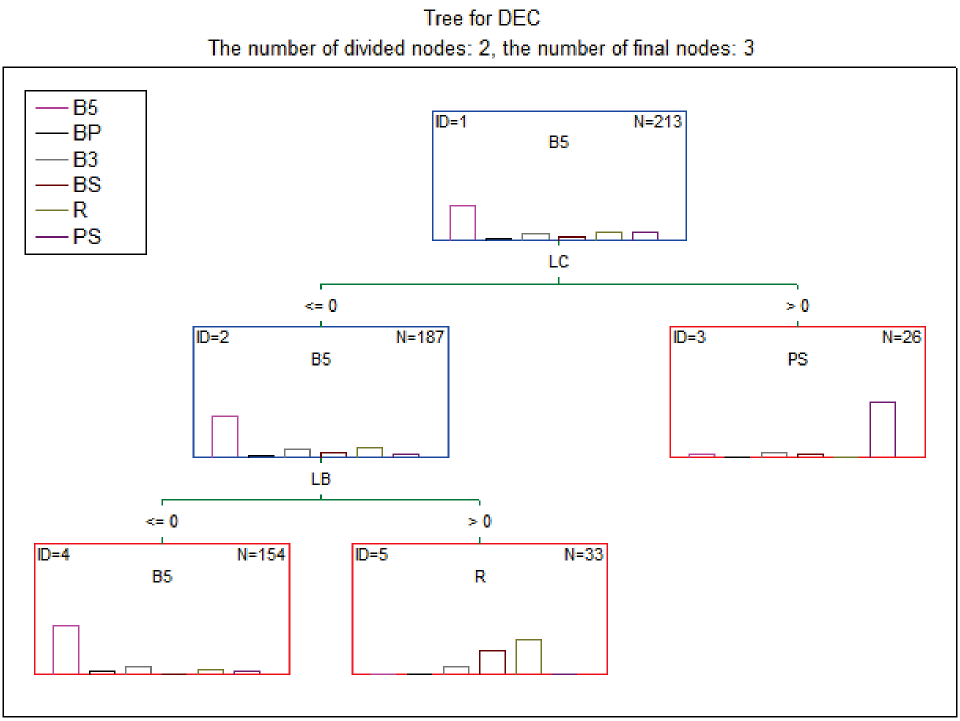
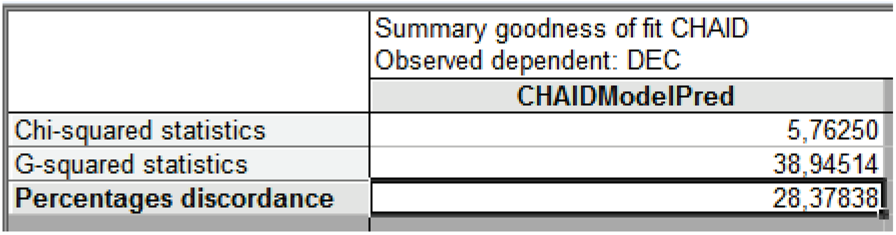


Fig. 3. Tree schema for the CHAID model

The summary of goodness of fit is presented in Table 3, which shows that the percentages discordance is 28.3738 and it is also higher than in the case of the C&RT model, which suggests that this model is slightly worse than the previous model.

Table 3

Summary of goodness of fit for the CHAID model



	Summary goodness of fit CHAID Observed dependent: DEC
	CHAIDModelPred
Chi-squared statistics	5,76250
G-squared statistics	38,94514
Percentages discordance	28,37838

Another designed and built model is the exhaustive CHAID (XAID) model. The same programming procedures were also used during the design and build of this type of model.

After carrying out the process of building an exhaustive CHAID model, we obtained a tree [6], the structure of which is shown in Fig. 4. It consists of two divided nodes and four final nodes. This tree is slightly larger than the previous CHAID tree.

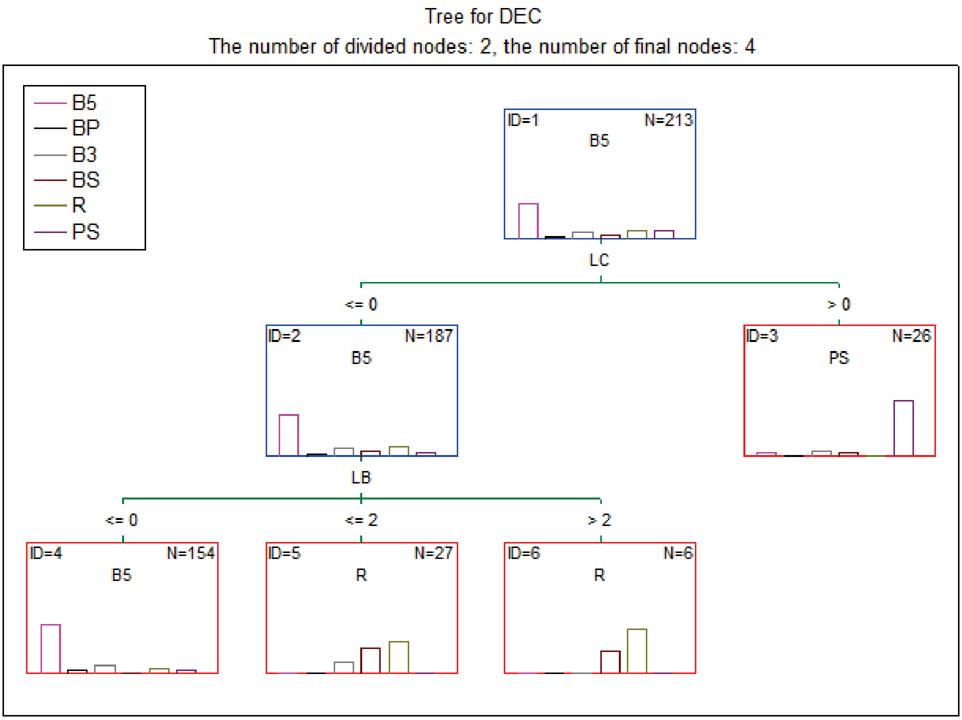


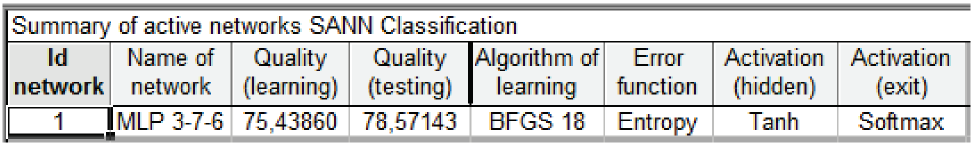
Fig. 4. Tree schema for the exhaustive CHAID model

The false prediction rate for this model used to classify the observations from the test sample and the parameters of goodness of fit were the same as for the case of the CHAID tree. Despite the fact that the resulting tree is slightly larger, the rest of the model parameters are identical to those for the CHAID tree.

The last designed and built predictive data mining model is the SANN model, i.e. the model of neural networks [1]. A summary of active networks is shown in Table 4. The table shows that this network has seven hidden neurons in the hidden layer, the quality of testing, which indicates the quality of the developed model, is at the level 78.57143%. Another very important factor in choosing the built network is its topology. The applied learning algorithm indicates that the epoch value of 18 was obtained at which the quality values presented in the summary were reached. When designing the network, the mutual entropy error function of the hyperbolic tangent of the hidden layer activation were used. A softmax function was used as the output layer activation function. The quality of the validation set was not determined because this set was not distinguished in this analysis.

Table 4

Summary of an active neutral network

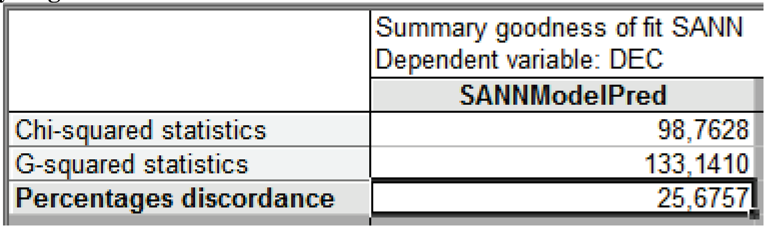


Summary of active networks SANN Classification							
Id network	Name of network	Quality (learning)	Quality (testing)	Algorithm of learning	Error function	Activation (hidden)	Activation (exit)
1	MLP 3-7-6	75,43860	78,57143	BFGS 18	Entropy	Tanh	Softmax

The false prediction rate for this model of neural network used to classify the observations from the test sample is 0.253393665158371 and is slightly lower than for the C&RT model.

Table 5

Summary of goodness of fit for the SANN model

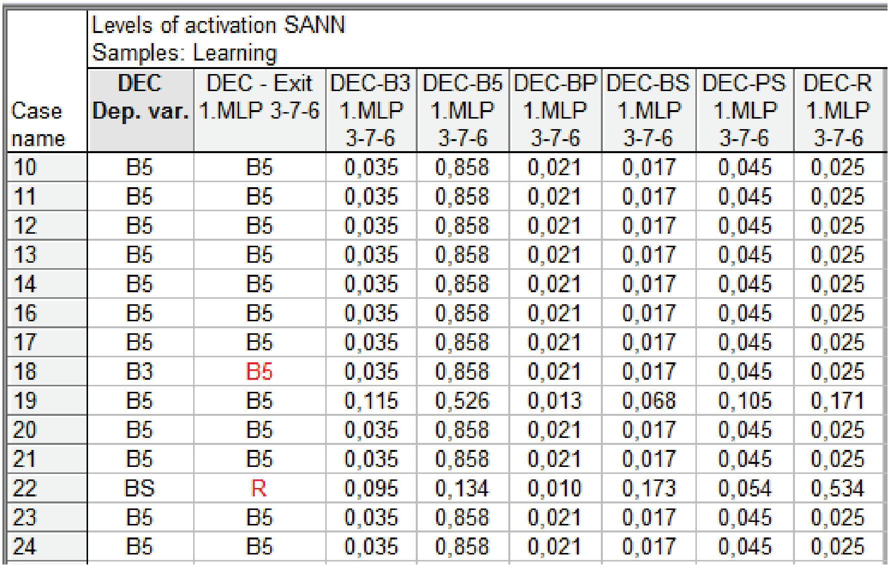


	Summary goodness of fit SANN Dependent variable: DEC
	SANNModelPred
Chi-squared statistics	98,7628
G-squared statistics	133,1410
Percentages discordance	25,6757

A summary of the goodness of fit is presented in Table 5, which shows that the percentages discordance is 25.6757 and it is also lower than in the C&RT model, which suggests that this model is a slightly better model.

Table 6

Excerpt from a prediction sheet including activation levels



	Levels of activation SANN Samples: Learning							
Case name	**DEC Dep. var.**	DEC - Exit 1.MLP 3-7-6	DEC-B3 1.MLP 3-7-6	DEC-B5 1.MLP 3-7-6	DEC-BP 1.MLP 3-7-6	DEC-BS 1.MLP 3-7-6	DEC-PS 1.MLP 3-7-6	DEC-R 1.MLP 3-7-6
10	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
11	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
12	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
13	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
14	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
16	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
17	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
18	B3	B5	0,035	0,858	0,021	0,017	0,045	0,025
19	B5	B5	0,115	0,526	0,013	0,068	0,105	0,171
20	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
21	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
22	BS	R	0,095	0,134	0,010	0,173	0,054	0,534
23	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025
24	B5	B5	0,035	0,858	0,021	0,017	0,045	0,025

Table 6 presents a fragment of the prediction sheet together with activation levels of individual obtained values of the dependent variable. This fragment shows that some of the postdiagnostic decisions determined by our built model of neural networks are different (in red). This suggests the need to verify these tested lots of propellant charges in order to check whether the person making the assessment did not make a mistake.

The final form of the working space for the built data mining predictive models is shown in Fig. 5. Here we can see new nodes of learning and test sets created for each built model, as well as a goodness of fit node for many data sources and report documents in which we have a number of final parameters of each built model.

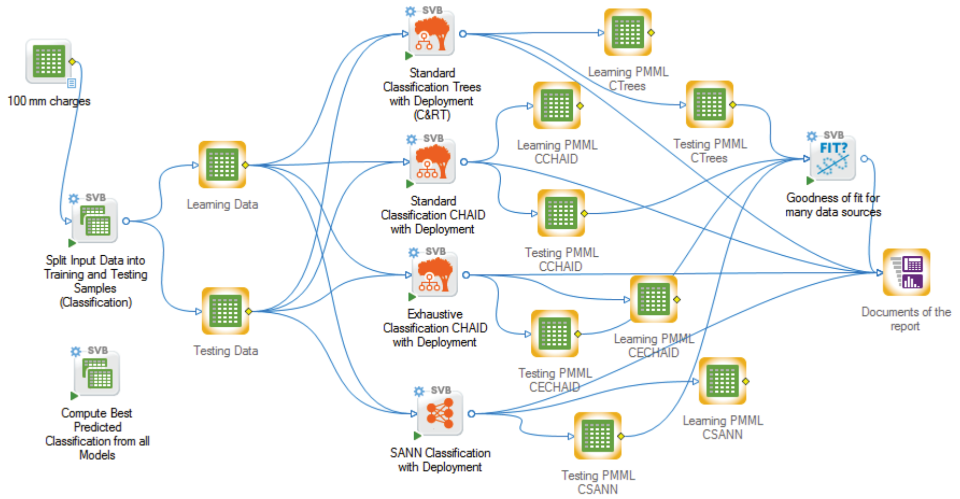

100 mm charges
SVB
Split Input Data into Training and Testing Samples (Classification)
SVB
Compute Best Predicted Classification from all Models
Learning Data
Testing Data
SVB
Standard Classification Trees with Deployment (C&RT)
SVB
Standard Classification CHAID with Deployment
SVB
Exhaustive Classification CHAID with Deployment
SVB
SANN Classification with Deployment
Learning PMML CTrees
Learning PMML CCHAID
Testing PMML CTrees
Testing PMML CCHAID
Learning PMML CECHAID
Testing PMML CECHAID
Learning PMML CSANN
Testing PMML CSANN
SVB
FIT?
Goodness of fit for many data sources
Documents of the report

Fig. 5. The final form of the workspace for all built models

In addition, we see Compute Best Predicted Classification from all Models node to which we can connect a new data source in which the software will calculate the classification prediction of the dependent variable for new observations. It should be remembered that the structure of the input file with the new data should be the same as the structure of the source file used to build our model. In particular, check that the same predictors and the dependent variable are selected, and that the names of all the variables match (this is especially important in the case of neural network models). In this way, we can make predictions for new data based on the designed and built and then selected as our best predictive data mining model.

In order to better illustrate the obtained results, Table 7 lists the values of the obtained percentage of inconsistent classifications for the built predictive models. The table shows that the best value of this parameter was obtained when building the SANN model.

Table 7

Summary of the percentage of inconsistent classifications for built models

No.	Prediction model	Percentages discordance
1.	C&RT	26,5766
2.	CHAID	28,3784
3.	XAID	28,3784
4.	SANN	25,6757

4. Summary

The article presents a method of designing and building predictive data mining models for the results of 100 mm propellant charges. Four predictive models based on the same laboratory diagnostic results were developed. The final result obtained for the individual built models was obviously different in terms of the obtained parameters indicating the quality of the built models.

A thorough analysis of the obtained parameters of the built predictive models showed a number of data that clearly indicate the answer. The best quality parameters were achieved for the SANN model, i.e. for neural networks. There is also the possibility of predictive assessment for new test results using the other three developed models, but it should be remembered that they have higher values of assessment errors and obtaining a correct postdiagnostic decision may be subject to greater errors, which may lead to the determination of an incorrect value of the dependent variable.

In the opinion of the author, the purpose set at the beginning of the article has been fully achieved. It is therefore possible to design and build data mining predictive models based on the results of diagnostic tests of artillery cartridges elements, which are 100 mm propellant charges.

Implementation of this statistical predictive model for practical use is conditioned by its acceptance by the management of the test facility dealing with these diagnostic tests. The formal requirement is the need to create terminals at the test stations and to connect them with the software [8] owned by the Institute, so that the designed evaluation model can be used.

The combination of various techniques in Statistica software is currently the strongest known method of creating classification predictions, suitable for use even in very difficult conditions (situations) where the predictor variables are strongly and nonlinearly related to each other.

From the obtained values of qualitative parameters, such as the false of predictive rate of the built predictive data mining models, it can be seen that the predictions of all models have similar accuracy, which may mean that the other built models can also be implemented. A similar suggestion can be made by analyzing the obtained vales of the goodness of fit for the individual built models.

In conclusion, the presented example was intended to show how many existing, very advanced methods of predictive data mining can be combined and used in an implementation that will perform prediction for new test results. When performing this type of analysis, one should remember about a number of formal requirements that must be met in order to properly design a predictive data mining model.

As you can see, artificial intelligence is creeping ever wider even into the diagnostic tests of special elements of technical objects, which are, after all, propelling charges completed in artillery cartridges. It is only a matter of time before it is put into practice and the human element is removed from the evaluation process. This fact will cause that the obtained prediction value of the dependent variable for new observations will be at a higher level of credibility, which will indirectly increase the level of security of long-term stored munitions.

5. References

1. Ampuła D.: Applying of neutral networks for testing of tracers with using empirical data. Scientific Journal of Polish Naval Academy, No. 3, 2019.
2. Ampuła D.: Decision trees in the tests of artillery igniters. Journal of KONBiN, Vol. 50, Iss. 1, 2020, DOI 10.2478/jok-2020-0007.
3. Ampuła D.: Prediction of post-diagnostic decisions for tested hand grenade' fuses using decisions trees. Problems of Mechatronics 12, 2(44), 2021.
4. Cards from laboratory tests of 100 mm propellant charges. Archive Military Institute of Armament Technology (MIAT).
5. Electronic manual Statistica, Statsoft Poland 2022.
6. Łapczyński M., Demski T.: Data mining – predictive methods. Material from course, Statsoft Poland 2019, pp. 7-31.
7. Reports from tests of ammunition – archive MIAT.
8. Statistica 13.3 PL – computer software. Statsoft Poland 2018.
9. The handbook – Ammunition of land forces. Publishing House Ministry of National Defence, Warsaw 1985, pp. 184-193.
10. The methodology of tests of artillery charges – handbook. Index N-5005°, archive MIAT, 1987.