Journal of KONBiN 2020
Volume 50, Issue 1
DOI 10.2478/jok-2020-0007

Dariusz AMPUŁA
Military Institute of Armament Technology (Wojskowy Instytut Techniczny Uzbrojenia)

DECISION TREES IN THE TESTS OF ARTILLERY IGNITERS

Drzewa decyzyjne w badaniach zapłonników artyleryjskich

Abstract: *The article addressed the method for building decision trees paying attention to the binary character of the tree structure. The methodology for building our decision tree for KW-4 igniters was presented. It involves determining features of tested igniters and applied predictors, which are necessary to create the correct model of the tree. The classification tree was built based on the possessed test results, determining the adopted post-diagnostic decision as the qualitative independent variable. The schema of the resultant classification tree and the full structure of this tree together with the results in end nodes were shown. The obtained graphic and tabular sequence of the designed tree was characterized, and the prediction accuracy was evaluated on the basis of the resultant matrix of incorrect classifications. The quality of the resultant predictive model was assessed on the basis of the chosen examples by means of the 'ROC' curve and the graph of the cumulative value of increase coefficient.*
Keywords: decision trees, igniter, node, branch, leaf, laboratory tests

Streszczenie: *W artykule opisano metodę budowy drzew decyzyjnych zwracając uwagę na binarny charakter struktury drzewa. Przedstawiono metodykę budowy drzewa dla zapłonników typu KW-4, określając cechy badanych zapłonników oraz zastosowane predyktory, które są niezbędne do tworzenia prawidłowego modelu drzewa. Na podstawie posiadanych wyników badań, zbudowano drzewo klasyfikacyjne, określając jako jakościową zmienną niezależną przyjętą decyzję podiagnostyczną. Pokazano schemat powstałego drzewa klasyfikacyjnego oraz pełną strukturę tego drzewa łącznie z wynikami w węzłach końcowych. Scharakteryzowano uzyskaną graficzną i tabelaryczną sekwencję zaprojektowanego drzewa oraz oceniono trafność predykcji na podstawie powstałej macierzy błędnych klasyfikacji. Oceniono na wybranych przykładach jakość powstałego modelu predykcyjnego za pomocą krzywej „ROC" oraz wykresu skumulowanej wartości współczynnika przyrostu.*
Słowa kluczowe: drzewa decyzyjne, zapłonnik, węzeł, gałąź, liść, badania laboratoryjne

1. Introduction

Decision or classification trees [4] appeared in the literature in the context of sociological research. Brian Ripley considers that an article by J.N. Morgan and J.A. Sonquist from 1963 was the first publication devoted to them. Although it was printed in the *Journal of the American Statistical Association*, it regarded sociological issues. Decision trees (and regression trees) settled in statistics in 1984, thanks to book [2]. In the field of machine learning, decision trees began to be popularized a little earlier, that is in the late 70s by Quinlan, who was unaware of the works of Morgan and others, but referred to the earlier works of psychologists and mentioned them as those that led him to the idea of constructing a decision tree.

Using the language of mathematics [4], trees are defined as not directed acyclic and coherent graphs. It is convenient to present them as directed trees with the only distinguishable vertex, called the root, which is the initial vertex of the tree.

The purpose of this article was to design and build a decision tree based on previous results of laboratory tests of artillery igniters. KW-4 igniters were used to achieve this goal. The test results base of these igniters is the most numerous, which gives a high probability of developing a decision tree with a high-quality level of its work. This article shows us the possibility of applying decision tree theory to support making post-diagnostic decisions for the tested artillery igniters. The decision tree designed in this article concerned the evaluation module of tested igniters in the scope of the first laboratory diagnostic tests. The purpose of using the designed decision tree algorithms is primarily to obtain a tree model with the best predictive accuracy for the new test results.

2. The method for building decision trees

It is a good idea to imagine decision trees [4] as objects in whose roots the entire learning sample is concentrated, and the next elements of this sample are shifted along the branches from top to bottom through the nodes. In each node, a decision is made to select a particular branch, along with which the sample element will continue. In this way, at each node that is not a leaf, the elements of the learning sample are divided into subgroups. The leaf is called the end node of the tree. Under each node that is not a leaf, most often next to branches, the criterion for dividing the subgroup that reaches that node into smaller subgroups that reach the next nodes (children's nodes) is recorded. The criterion of the division made in a given node is common (the same) for all elements of the learning sample that were in the node. The elements of the sample are moved to one end node, which is the leaf of the tree, which is usually assigned a label of the class of the analysed problem of discrimination and from which come most elements of the learning sample that reach this leaf.

When building a decision tree [4], the question ‘Why to classify elements of the learning sample whose class membership is known to us?’ will arise. The decision tree is

built to enable the classification of future observations of the classes we do not know. The decision tree is built on the basis of a learning sample. The learning sample determines the form of the conditions for the division of elements going to a given node into subgroups going further to the nodes. It is also up to the learning sample which node is accepted as leaf, where no division occurs anymore and all observations that reach it are classified in this, and not the other way.

It should be noted here an essential thing - if the decision tree is to be used to classify future observations, then the division conditions must be in the form of conditions imposed only on the values of the observation vector, and not on the belonging of these elements to classes.

By designing decision trees, only binary trees are considered: that is, the trees whose nodes (except leaves) have two children each. Such trees usually turn out to be the best, which means that the basic principle of dividing the elements of the learning sample, which were in a given node, will be dividing into two parts, i.e., we will pass to two nodes - children. This division should be based on the best separation of this sample.

When all these requirements are given by the designer of a given decision tree, the software [8] will automatically look for the proportion (fraction) of observation of a given class in a node, considering them as approximations of the probability of occurrence of a given class in this node. Such a solution to the distribution of data results will be the best local division for a given designed decision tree.

3. The methodology for building the tree for KW-4 igniters

The test results of KW-4 igniters [3, 7] were prepared for the designed decision tree of the first laboratory diagnostic tests. These igniters are used in artillery cartridges of 57 mm to 152 mm calibres both in high-explosive blast-fragmentation projectiles, antitank-tracer projectiles and also in blank projectiles. The results of the so-called scientific-research inquiries that are not relevant to other test results, were eliminated. Only those tests were taken for analysis, in which the type of test specified in the test methodology [6] was one for test samples stored in the stores of the Polish Army's economic departments, which means that only lots of igniters stored in the storage subset specified as ‘K’ and ‘L’ were considered. All these restrictions were intended to create a homogeneous data results set that can be analysed by the designed decision trees.

KW-4 igniter is an impact igniter. According to [1], it consists of a primer cap with pyrotechnic mass, anvil, sealant, body, incendiary charge and a special seal. It can withstand an internal pressure of about 300 MPa. During laboratory tests, the following were checked: correct operation of the primer cap and ignition charge, corrosion of individual parts of the igniter and humidity of powder and powder cube.

All properties (features) of KW-4 igniters were tested according to [6] and were divided into five classes of importance (inconsistencies): A, B, C, D and E. Depending on

the number of detected inconsistencies in individual importance classes during diagnostic laboratory tests, a post-diagnostic decision is obtained, according to the evaluation module.

In the case of the designed decision tree for the analysed igniters, for the first laboratory diagnostic tests, five different pieces of data (predictors) of the tested features were adopted, which were the information obtained after the diagnostic tests, namely. Those were the following predictors: the number of the inconsistencies in the importance class A (LA), the number of the inconsistencies in the importance class B (LB), the number of the inconsistencies in the importance class C (LC), the number of the inconsistencies in the importance class D (LD) and the number of the inconsistencies in the importance class E (LE).

Therefore, in our case, the results of all tested features of a given igniter lot were the values of the predictors. These parameters were written in numerical form, i.e.; if no inconsistencies of a given class were found during the diagnostic test, then, the value zero was provided. However, if inconsistencies were found in the tests, then a specific number of these inconsistencies was given. The searched value was obtained based on a specific post-diagnostic decision in accordance with the test methodology [6].

While designing a decision tree, we were dealing with a classification tree due to the fact that it was possible to obtain several different post-diagnostic decisions, depending on the number of inconsistencies received during the leading laboratory tests. According to the evaluation module in the test methodology [6], six different post-diagnostic decisions could be obtained as a result of the first laboratory tests .

While building our classification tree, additional auxiliary parameters were adopted, whose task was to design the best classification tree for the test results of the analysed igniters. The values of these parameters have been introduced in the software [8].

To sum up, the subject of the classification during the design and building process of our decision tree was a set of data obtained during the first laboratory tests of KW-4 igniters. The set of these data was regarded as a learning set for the new received diagnostic test results. Each tested lot of igniters was characterized by the obtained results of the tested features of these igniters, recorded in numerical form and entered into the database, which constituted the learning set. A classification tree was built using the designed C&RT algorithm.

4. Results of building a decision tree with the C&RT method

For building our decision tree, the Classification and Regression Tree (C&RT) algorithm was adopted, which works on recursive binary division of objects from the learning sample into two subsets. The problem of ending the division of the set, i.e., the initial ordering, was solved by applying the heuristic rule that the division should be abandoned if there is no significant decrease in the degree of object diversity in the tested set.

In the case of building a tree for the first laboratory tests of KW-4 igniters, a qualitative dependent variable, called 'DEC', was adopted, which is the post-diagnostic decision obtained after the first laboratory tests. The decision may take the form of 'B5', 'B3', 'B', 'Z' and 'W'. A detailed description of possible diagnostic decisions is presented in [6]. The database of test results for these igniters was prepared according to such a key that all data entered into this database constituted a certain homogeneity.

The so-called costs of incorrect classification at 'equal' levels were adopted as a criterion for the predictive validity of building a tree. In such a way, the reduction of costs will minimise the proportion of misclassified cases, while a priori probabilities will be proportional to the size of the classes, and the misclassification costs in each class will be the same.

The value of a priori probability at the 'estimated' level was also assumed in the case of our qualitative dependent variable. Then, the minimisation of costs means the smallest proportion of misclassification cases, provided that a priori probabilities are proportional to the size of the classes and the costs of incorrect classifications are equal in all classes as previously adopted.

The rules for the division of the designed tree in terms of goodness of fitting, as in the Gini index, which specifies the measure of heterogeneity (inconsistency) of the node, are also indicated.

The next step in building the tree was to choose the 'stop' criterion (rule). The level 'cut at the error of wrong classification' was adopted, in which the minimal number of end nodes was determined at a level of '50', which means that the process of pruning the tree will begin when the number of cases in the leaf of the tree reaches this value. The maximum number of nodes at the value level '1000' is also indicated.

In the last step, the option of estimating the model error using multiple cross-validation (multiple cross-test) was specified. The number of repetitions for the v-fold cross-test was determined at the level of '10', as this value is recommended by the authors of the software.

As a result of the completion of the process of building our tree model, among the designed trees, the programme automatically selected one tree with the best parameters. This is tree number 3, the scheme of which is shown in fig. 1.

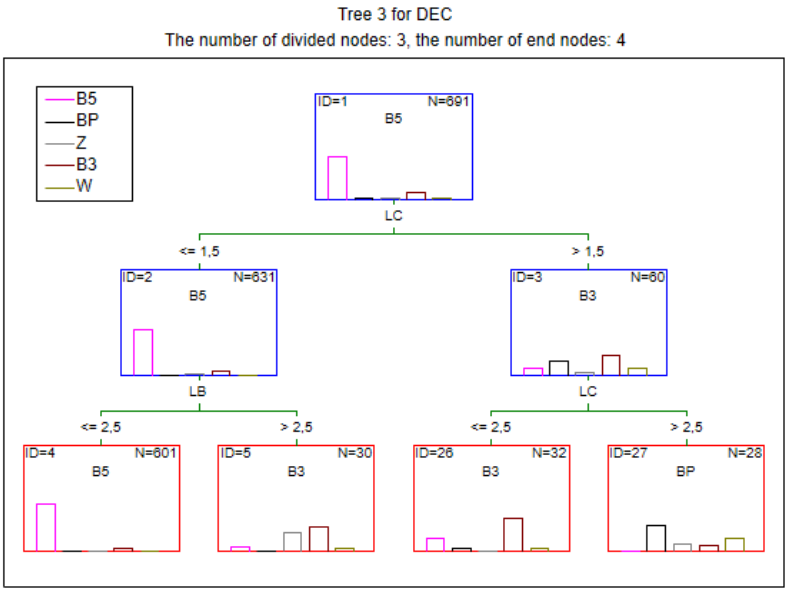


Fig. 1. The scheme of the tree number 3 for first laboratory tests

As can be seen in fig. 1, tree number 3 was selected, which has 3 divided nodes and 4 end nodes (leaves), thus, it is little developed. The presented scheme does not include some nodes that have been removed due to the pruning process (pruning-one of the nodes has the ID number 27). The exact number structure of tree number 3 is shown in fig. 2.

The structure of the tree number 3 (KW4 RB=1)
Dependent variable: DEC
Options: Quality dependent, Tree number 3

Node number	Left branch	Right branch	Node size	N class B5	N class BP	N class Z	N class B3	N class W	Chosen class	Divide variable	Divide constant
1	2	3	691	552	22	21	81	15	B5	LC	1,5
2	4	5	631	544	6	17	58	6	B5	LB	2,5
4			601	541	6	6	44	4	B5		
5			30	3	0	11	14	2	B3		
3	26	27	60	8	16	4	23	9	B3	LC	2,5
26			32	8	2	0	20	2	B3		
27			28	0	14	4	3	7	BP		

Fig. 2. The structure of the tree number 3 for first laboratory tests

In the presented table, you can see a description of the exact data in all nodes, including the total number of the given node, the number of the adopted individual classes and the selected class that is the name of each node. Branches in divided nodes and parameters for the division of these nodes were also determined.

An important advantage of the C&RT algorithm [5] is the simultaneous comparison of the cost of re-substitution with the error rate calculated on the test set. Figure 3 shows us

just such a sequence of costs, which is created from the data contained in the tree sequence table, shown in fig. 4. Although tree number 2 has the lowest CT cost (cross-test), however, tree number 3 was chosen by software as the one with the 'correct size'.

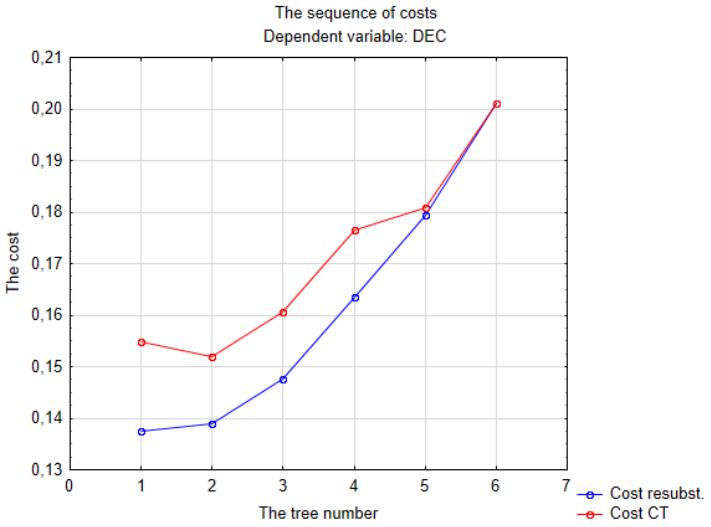


Fig. 3. The sequence of costs of tree number 3 for first laboratory tests

The sequence of trees (KW4 RB=1)
Dependent variable: DEC
The best tree marked *

	Final nodes	SC cost	SC std. error	Cost of resubstitution	Node complexity
Tree 1	6	0,154848	0,013762	0,137482	0,000000
Tree 2	5	0,151954	0,013656	0,138929	0,001447
*Tree 3	4	0,160637	0,013969	0,147612	0,008683
Tree 4	3	0,176556	0,014505	0,163531	0,015919
Tree 5	2	0,180897	0,014644	0,179450	0,015919
Tree 6	1	0,201158	0,015250	0,201158	0,021708

Fig. 4. The sequence of trees for first laboratory tests

The next step in the analysis of building the tree model was to assess the accuracy of the prediction. The simplest tool to assess the correctness of classification [4] is the created matrix of incorrect classifications (fig. 5). This matrix compares the observed classes and the predicted classes. In our case, we received automatically calculated prediction errors, which are located in the row '% from line' for the individual predicted post-diagnostic decisions.

	Matrix of classification 3 (KW4 RB=1) Dependent variable: DEC Options: Quality dependent, Attempt to analysis						
	Observed	Predicted B5	Predicted BP	Predicted Z	Predicted B3	Predicted W	Together in the line
Number	B5	541			11		552
% from column		90.02%	0.00%		17.74%		
% from line		98.01%	0.00%	0.00%	1.99%	0.00%	
% from total		78.29%	0.00%	0.00%	1.59%	0.00%	79.88%
Number	BP	6	14		2		22
% from column		1.00%	50.00%		3.23%		
% from line		27.27%	63.64%	0.00%	9.09%	0.00%	
% from total		0.87%	2.03%	0.00%	0.29%	0.00%	3.18%
Number	Z	6	4		11		21
% from column		1.00%	14.29%		17.74%		
% from line		28.57%	19.05%	0.00%	52.38%	0.00%	
% from total		0.87%	0.58%	0.00%	1.59%	0.00%	3.04%
Number	B3	44	3		34		81
% from column		7.32%	10.71%		54.84%		
% from line		54.32%	3.70%	0.00%	41.98%	0.00%	
% from total		6.37%	0.43%	0.00%	4.92%	0.00%	11.72%
Number	W	4	7		4		15
% from column		0.67%	25.00%		6.45%		
% from line		26.67%	46.67%	0.00%	26.67%	0.00%	
% from total		0.58%	1.01%	0.00%	0.58%	0.00%	2.17%
Number	Total groups	601	28		62		691
% together		86.98%	4.05%	0.00%	8.97%	0.00%	

Fig. 5. The matrix of incorrect classifications for first laboratory tests

	Importance of predictors 3 (KW4 RB=1) Dependent variable: DEC Options: Quality dependent, Tree number 3	
	Variable rank	Importance
LC	100	1,000000
LD	82	0,816631
LB	43	0,425306
LA	38	0,382962
LE	0	0,000000

Fig. 6. The ranking of importance of predictors for first laboratory tests

For example, for the observed BP class, the prediction error for the 'Predicted B3' class is 0.29%. It should be emphasized here that the usefulness of building a tree model is not only determined by the accuracy of the prediction of the entire solution, but also by the accuracy of the prediction of individual classes of the model. The matrix of incorrect classifications is also the starting point for calculating other measures, e.g., sensitivity, specificity, F measure or G mean.

The next step was to compile the rankings of predictors' importance and, if need be, to prepare their graph. The table in figure 6 shows us such a ranking of predictor importance. It shows that the LE predictor is not statistically significant, and the 'LC' predictor is the most significant that is most important (variable rank).

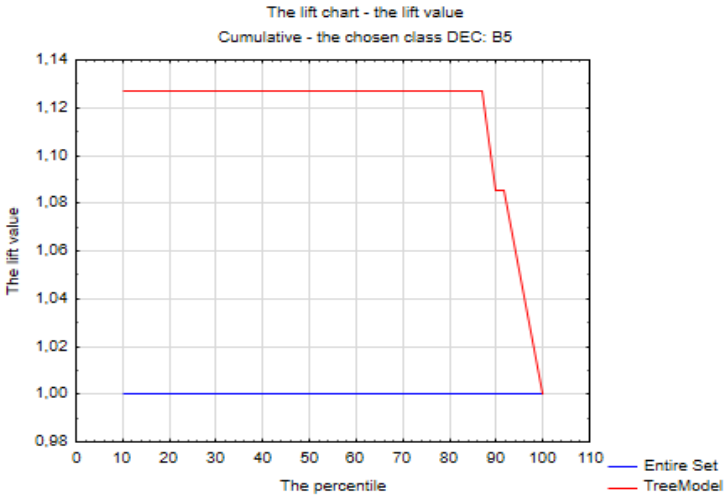


Fig. 7. The lift chart for B5 decision

There is one more tool in the software [8] to assess the quality of the resulting predictive model. It is a graph of the cumulative value of the lift chart, which is a graphical summary of the model's usability to predict the value of a dependent variable.

An example of the graph of the increasing value for the B5 class is shown in fig. 7. Graphs of this type were created for all post-diagnostic decisions. Axis (y) displays the increasing values, that is the multiples of the reference line, and axis (x) defines the 'percentile' values. Intuitively speaking, the percentile is that falls below the values of a given percentage of samples.

From this graph, it can be read that taking, for example, 10% of cases most likely classified to B5 class (with the highest probability of classification) we will receive a sample that contains about 1,128 times more cases than if the selection was random. The resulting graph is created for changing sets, which contain an increasing number of cases with the highest probability of getting to the class. The created model in this way consequently shows that the next one contains the previous one (in this sense, the graph is cumulative).

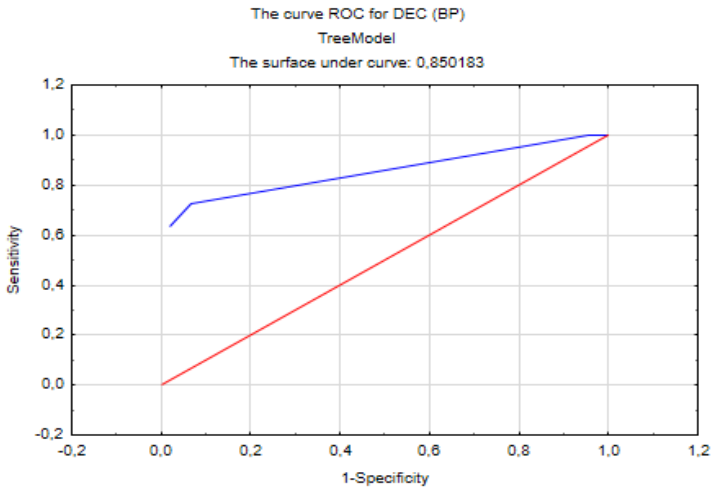


Fig. 8. The ROC curve for BP decision

Another tool for assessing the performance of the built model is the ROC curve (Receiver Operating Characteristic), which has many applications. In particular, it is used to assess the quality of the model, which predicts belonging to different classes. A very useful indicator of the quality of the resulting model is the area under this ROC curve: the larger it is, the better the built model will be. An example of such a curve for the BP class is shown in figure 8. As you can see, the area under the curve is 0.850183, which means that the built model in this class is good. ROC curves have been developed for all possible kinds of dependent variables in the model, which is for post-diagnostic decisions.

5. Summary

The article was an attempt to design and build a decision tree model according to the C&RT method for KW-4 igniters, which are used in artillery cartridges from 57 mm to 152 mm calibres. The purpose specified at the beginning of the article has been fully achieved. The necessary parameters for building the correct tree were determined, which ultimately led to the building and selection of the best classification tree model. Thanks to the specialized computer software [8], a tree model was built, whose best form (number 3) was automatically chosen by this software. The performed tree pruning procedure limited its size and presented the form of this tree, which is the most optimal for the tree predictors assumed at the beginning. The pruning process is carried out according to the principle that it is better to have a smaller model because it is more universal in practice.

Thanks to designing and building our classification tree, we can now make a quick assessment of newly tested lots of KW-4 igniters. The implementation of this decision tree to the assessment process requires connecting the computers with the new test results to the

software [8], where our designed and built evaluation model is installed. The new predictor values obtained from the tests enable the designed tree model to quickly evaluate these new lots of igniters by automatically determining the post-diagnostic decision. Obviously, the use of the built classification tree requires, in practice, the permission to implement this tool. From a practical point of view, the use of this tool seems to be a necessity, because the built classification tree automatically indicates the correct post-diagnostic decision for the newly tested lots of igniters. In this way, the weakest link in the assessment process, which is a human element, is avoided.

It appears that artificial intelligence, which storms most scientific fields, will be also used for evaluating of the results after testing special elements of technical objects, which are KW-4 igniters. Similar classification trees can be designed and built for other types of tested artillery igniters. A necessary condition is to have a database that will contain the results of the previously conducted diagnostic laboratory tests.

6. References

1. Amunicja wojsk lądowych. Ministry of National Defence Publishing House, Warszawa 1985.
2. Breiman L., Friedman J.H., Olsen R.A., Stone C.J.: Classification and Regression Trees. Chapman & Hall, 1984.
3. Cards from laboratory tests of igniters type KW-4. Archive Military Institute of Armament Technology (MIAT).
4. Koronacki J., Ćwik J.: Statystyczne systemy uczące się. Akademicka Oficyna Wydawnicza Exit, Warszawa 2008.
5. Łapczyński M., Demski T.: Data mining – metody predykcyjne. Statsoft Polska – materials from course, Kraków 2019.
6. Metodyka badań diagnostycznych amunicji – Indeks N-5001b – Archiwum WITU, 1985.
7. Reports from tests of ammunition. Archive MIAT.
8. Statystyka 13.3 PL, computer software. Statsoft Polska 2018.

DRZEWA DECYZYJNE W BADANIACH ZAPŁONNIKÓW ARTYLERYJSKICH

1. Wprowadzenie

Drzewa zwane decyzyjnymi lub klasyfikacyjnymi [4] pojawiły się w literaturze w kontekście badań socjologicznych. Brian Ripley uważa za pierwszą publikację im poświęconą artykuł J.N. Morgana i J.A. Sonquista z roku 1963, co prawda wydrukowany w czasopiśmie *Journal of the American Statistical Association*, ale dotyczący tematyki socjologicznej. Drzewa decyzyjne (i regresyjne) zadomowiły się w statystyce w roku 1984 dzięki książce [2]. W dziedzinie uczenia maszynowego drzewa decyzyjne zaczęły być popularyzowane nieco wcześniej, bo w końcu lat 70. przez Quinlana, nieznającego zresztą prac Morgana i innych, za to powołującego się na wcześniejsze prace psychologów i wymieniającego je jako te prace, które naprowadziły go na ideę konstrukcji drzewa decyzyjnego.

Posługując się językiem matematyki [4], drzewa definiuje się jako nieskierowane grafy acykliczne i spójne. Wygodnie jest je przedstawić jako drzewa skierowane, mające jedyny dający się wyróżnić wierzchołek, nazywany korzeniem, będący wierzchołkiem początkowym drzewa.

Celem tego artykułu było więc zaprojektowanie oraz zbudowanie drzewa decyzyjnego w oparciu o dotychczasowe wyniki badań laboratoryjnych zapłonników artyleryjskich. Do realizacji tak postawionego celu wzięto zapłonniki typu KW-4. Baza wyników badań tych zapłonników jest najliczniejsza, co daje duże prawdopodobieństwo opracowania drzewa decyzyjnego o wysokim poziomie jakościowym jego pracy. Artykuł ten pokazuje nam możliwość zastosowania teorii drzew decyzyjnych do wspomagania podejmowania decyzji podiagnostycznych dla badanych zapłonników artyleryjskich. Projektowane w tym artykule drzewo decyzyjne dotyczyło modułu ocenowego badanych zapłonników w zakresie pierwszych laboratoryjnych badań diagnostycznych. Celem stosowania projektowanych algorytmów drzew decyzyjnych jest przede wszystkim otrzymanie modelu drzewa o najlepszej trafności predykcyjnej dla nowych wyników badań.

2. Metoda budowy drzew decyzyjnych

Drzewa decyzyjne [4] dobrze jest wyobrazić sobie jako obiekty, w których korzeniach jest skupiona cała próba ucząca i następnie kolejne elementy tej próby są przesuwane wzdłuż gałęzi, z góry na dół przez węzły, w każdym jest podejmowana decyzja o wyborze gałęzi, wzdłuż której będzie przesuwany element próby. W ten sposób w każdym węźle, który nie jest liściem, jest dokonywany podział elementów próby uczącej, na podgrupy. Liściem nazywa się węzeł końcowy drzewa. Pod każdym węzłem, który nie jest liściem, najczęściej przy gałęziach, zapisywane jest kryterium podziału podgrupy docierającej do tego węzła na mniejsze podgrupy, które dotrą do węzłów następnych (węzłów dzieci). Kryterium podziału dokonywanego w danym węźle jest wspólne (takie samo) dla wszystkich elementów próby uczącej, które znalazły się w tym węźle. Elementy próby są przesuwane aż do któregoś węzła końcowego, czyli liścia drzewa, któremu zwykle przypisuje się etykietę tej klasy analizowanego problemu dyskryminacji, z której pochodzi najwięcej elementów próby uczącej, które dotarły do tego liścia.

Budując drzewo decyzyjne [4], nasuwa się pytanie, po co klasyfikować elementy próby uczącej, których przynależność do klas jest nam znana. Drzewo decyzyjne jest po to budowane, aby umożliwić klasyfikowanie przyszłych obserwacji, o których nie wiemy, do jakich klas należą. Drzewo decyzyjne budowane jest na podstawie próby uczącej. To od niej zależy postać warunków podziału elementów trafiających do danego węzła na podgrupy trafiające dalej do węzłów. To od niej też zależy, który węzeł zostaje uznany za liść, gdzie żaden podział już nie następuje i wszystkie obserwacje, które doń dotrą zostają tak, a nie inaczej zaklasyfikowane.

Zauważyć należy tu rzecz niezwykle istotną – jeśli drzewo decyzyjne ma służyć do klasyfikowania przyszłych obserwacji, to warunki podziału muszą mieć postać warunków narzuconych wyłącznie na wartości wektora obserwacji, a nie na przynależność tych elementów do klas.

Projektując drzewa decyzyjne, rozważa się w większości tylko drzewa binarne, czyli takie, których węzły (poza liśćmi) mają po dwoje dzieci. Takie drzewa najczęściej okazują się najlepsze, czyli podstawową zasadą podziału elementów próby uczącej, które znalazły się w danym węźle, będzie podział na dwie części, tzn. przechodzić będziemy do dwóch węzłów – dzieci. Podział ten powinien polegać na najlepszym rozdzieleniu tej próby.

Gdy te wszystkie wymagania zostaną podane przez projektanta danego drzewa decyzyjnego, oprogramowanie [8] samo automatycznie będzie szukało proporcji (frakcji) obserwacji danej klasy w węźle, uznając je za przybliżenia prawdopodobieństw pojawienia się w tym węźle obserwacji danej klasy. Takie rozwiązanie podziału wyników danych, będzie dla danego projektowanego drzewa decyzyjnego podziałem lokalnie najlepszym.

3. Metodyka budowy drzewa dla zapłonników KW-4

Do projektowanego drzewa decyzyjnego dla pierwszych laboratoryjnych badań diagnostycznych przygotowano wyniki badań zapłonników typu KW-4 [3, 7]. Zapłonniki te są stosowane w nabojach artyleryjskich kalibru od 57 mm do 152 mm, i to zarówno w pociskach odłamkowo-burzących, jak i w pociskach przeciwpancerno-smugowych, a także w pociskach ślepych. Wyeliminowano wyniki badań tzw. dociekań naukowo-badawczych, które nie są miarodajne do pozostałych wyników badań. Do analizy wzięto tylko badania, w których określony w metodyce badawczej [6] rodzaj badania wynosił wartość jeden dla próbek badawczych składowanych w magazynach oddziałów gospodarczych Wojska Polskiego, co oznacza, że rozpatrywano tylko badane partie zapłonników składowane w podzbiorze przechowywania określonym jako „K" i „L". Wszystkie te ograniczenia miały na celu stworzenie jednorodnego zbioru wyników danych, który będzie mógł być analizowany przez projektowane drzewa decyzyjne.

Zapłonnik typu KW-4 jest zapłonnikiem uderzeniowym. Składa się on według [1] ze spłonki zapalającej z masą pirotechniczną, kowadełka, uszczelniacza, kadłuba, ładunku zapalającego i specjalnego uszczelnienia. Wytrzymuje on ciśnienie wewnętrzne ok. 300 MPa. Podczas badań laboratoryjnych sprawdzane były m.in.: prawidłowe działanie spłonki zapalającej oraz ładunku zapalającego, kontrolowana była również korozja poszczególnych części zapłonnika, a także sprawdzana była wilgotność podsypki i kostki prochowej.

Badane wszystkie właściwości (cechy) zapłonników KW-4 według [6] podzielone zostały na pięć klas ważności (niezgodności): A, B, C, D i E. W zależności od liczby wykrytych niezgodności w poszczególnych klasach ważności podczas diagnostycznych badań laboratoryjnych, otrzymuje się określoną, zgodnie z modułem ocenowym decyzję podiagnostyczną.

W przypadku projektowanego drzewa decyzyjnego dla analizowanych zapłonników, dla pierwszych laboratoryjnych badań diagnostycznych, przyjęto pięć różnych danych (predyktorów) badanych cech, którymi były informacje uzyskane po przeprowadzonych badaniach diagnostycznych, a mianowicie były to predyktory: liczba niezgodności w klasie ważności A (LA), w klasie ważności B (LB), w klasie ważności C (LC), w klasie ważności D (LD) i w klasie ważności E (LE).

Wyniki wszystkich badanych cech danej partii zapłonnika stanowiły więc w naszym przypadku wartości predyktorów. Parametry te były zapisane w postaci liczbowej, tzn. jeżeli podczas badania diagnostycznego nie stwierdzono żadnych niezgodności danej klasy, wówczas dostarczona została wartość zero. Natomiast, jeżeli w badaniach wykryto niezgodności, wówczas podana była ich konkretna liczba. Wartością szukaną była uzyskana konkretna decyzja podiagnostyczna zgodnie z zapisami metodyki badawczej [6].

W naszym projektowanym drzewie decyzyjnym mieliśmy do czynienia z drzewem klasyfikacyjnym z uwagi na fakt, że istniała możliwość uzyskania kilku różnych decyzji podiagnostycznych, w zależności od liczby niezgodności otrzymanych podczas prowadzonych badań laboratoryjnych. Zgodnie z modułem ocenowym w metodyce badawczej

[6], w wyniku przeprowadzonych pierwszych badań laboratoryjnych można było otrzymać sześć różnych decyzji podiagnostycznych.

Podczas budowy naszego drzewa klasyfikacyjnego przyjęto dodatkowe parametry pomocnicze, których zadaniem było zaprojektowanie najlepszego drzewa klasyfikacyjnego dla posiadanych wyników badań analizowanych zapłonników. Wartości tych parametrów zostały wprowadzone do oprogramowania [8].

Reasumując, przedmiotem klasyfikacji podczas projektowania i budowy naszego drzewa decyzyjnego był zbiór danych uzyskanych podczas pierwszych badań laboratoryjnych zapłonników typu KW-4. Zbiór tych danych był traktowany jako zbiór uczący dla nowych otrzymywanych wyników badań diagnostycznych. Każda badana partia zapłonników była charakteryzowana przez otrzymane wyniki badanych cech tych zapłonników, zapisane w postaci liczbowej i wprowadzone do bazy danych, która stanowiły zbiór uczący. Zaprojektowane zostało drzewo klasyfikacyjne zbudowane za pomocą algorytmu C&RT.

4. Wyniki budowy drzewa decyzyjnego metodą C&RT

Do budowy naszego drzewa decyzyjnego przyjęto algorytm C&RT (Classification and Regression Trees), którego działanie polega na rekurencyjnym binarnym podziale obiektów z próby uczącej na dwa podzbiory. Problem zakończenia podziału zbioru, tj. porządkowania wstępnego, został rozwiązany przez zastosowanie reguły heurystycznej, mówiącej o tym, że należy zaniechać podziału, gdy nie następuje znaczący spadek stopnia zróżnicowania obiektów w badanym zbiorze.

W przypadku budowy drzewa dla pierwszych laboratoryjnych badań zapłonników typu KW-4, przyjęto jakościową zmienną zależną oznaczoną jako „DEC", która oznacza uzyskaną decyzję podiagnostyczną po pierwszych badaniach laboratoryjnych. Decyzja ta może przyjmować postać: „B5", „B3", „BP", „Z" i „W". Szczegółowy opis możliwych decyzji diagnostycznych został przedstawiony w [6]. Baza wyników badań tych zapłonników była przygotowana według takiego klucza, aby wszystkie dane wprowadzone do tej bazy stanowiły pewną jednorodność.

Przyjęto jako kryterium trafności predykcyjnej budowanego drzewa tzw. koszty błędnych klasyfikacji na poziomie „równe", wówczas minimalizacja kosztów spowoduje minimalizację proporcji przypadków błędnie zakwalifikowanych, przy czym prawdopodobieństwa a priori będą proporcjonalne do liczebności klas, a koszty błędnych klasyfikacji w każdej z klas będą takie same.

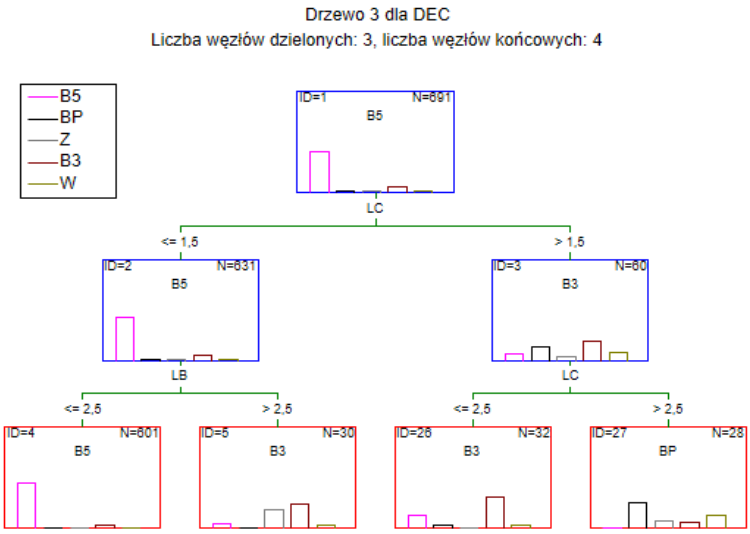
Założono także wartość prawdopodobieństwa a priori na poziomie „szacowane" w przypadku naszej jakościowej zmiennej zależnej, wówczas minimalizacja kosztów oznacza najmniejszą proporcję błędnych klasyfikacji przypadków, o ile prawdopodobieństwa a priori są proporcjonalne do wielkości klas, a koszty błędnych klasyfikacji są równe we wszystkich klasach, co zostało wcześniej przyjęte.

Zaznaczono również reguły podziału projektowanego drzewa w zakresie dobroci dopasowania jako indeks (miarę) Giniego, który określa miarę niejednorodności (niespójności) węzła.

Kolejnym krokiem w budowie drzewa był wybór kryterium (reguły) „stopu". Przyjęto poziom „przytnij przy błędzie złej klasyfikacji", w którym określono minimalną liczność węzła końcowego na poziomie wartości „50", co oznacza, że proces przycinania drzewa rozpoczynał się będzie w chwili, kiedy liczba przypadków w liściu drzewa osiągnie tę wartość. Zaznaczono także maksymalną liczbę węzłów na poziomie wartości „1000".

W ostatnim kroku określono opcję szacowania błędu modelu z wykorzystaniem wielokrotnej walidacji krzyżowej (wielokrotnego sprawdzianu krzyżowego). Określono liczbę powtórzeń dla v-krotnego sprawdzianu krzyżowego na poziomie wartości „10", gdyż taką wartość polecają autorzy oprogramowania.

W efekcie zakończenia procesu budowy modelu naszego drzewa, spośród zaprojektowanych drzew program wybrał automatycznie jedno drzewo o najlepszych parametrach. Jest to drzewo numer 3, którego schemat został pokazany na rys. 1.



Rys. 1. Schemat drzewa nr 3 dla pierwszych badań laboratoryjnych

Jak widać z rys. 1, drzewo nr 3 posiada trzy węzły dzielone oraz cztery węzły końcowe (liście), jest więc mało rozbudowane. Przedstawiony schemat nie uwzględnia niektórych węzłów, które zostały usunięte dzięki zastosowanemu procesowi przycinania (pruning – jeden z węzłów ma numer ID = 27). Dokładna struktura liczbowa drzewa nr 3 została przedstawiona na rys. 2.

W przedstawionej tabeli widać opis dokładnych danych we wszystkich węzłach z wyszczególnieniem liczności całkowitej danego węzła, liczności poszczególnych przyję-

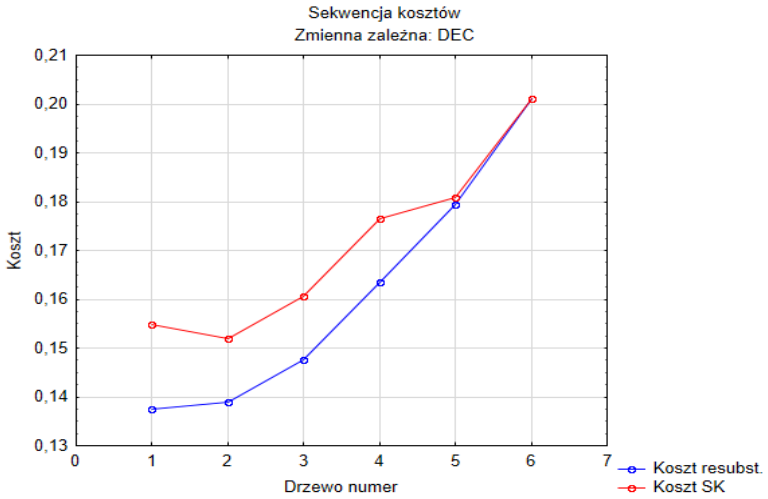
tych klas, a także wybraną klasę, czyli nazwę każdego węzła. Określono także gałęzie w węzłach dzielonych oraz parametry podziału tych węzłów.

Ważną zaletą algorytmu C&RT [5] jest jednoczesne zestawienie kosztu resubstytucji ze współczynnikiem błędu obliczonym na zbiorze testowym. Rys. 3 przedstawia nam właśnie taką sekwencję kosztów, która powstaje z danych zawartych w tabeli sekwencji drzew, pokazanej na rys. 4. Chociaż drzewo nr 2 ma najniższy koszt SK – sprawdzianu krzyżowego, to jednak drzewo nr 3 zostało wybrane przez oprogramowanie jako to o „właściwym rozmiarze".

Struktura drzewa 3 (KW4 RB=1)
Zmienna zależna: DEC
Opcje: Jakościowa zależna, Drzewo numer 3

Nr węzła	Lewa gałąź	Prawa gałąź	Liczność węzła	N klasy B5	N klasy BP	N klasy Z	N klasy B3	N klasy W	Wybrana klasa	Podział zmienna	Podział stała
1	2	3	691	552	22	21	81	15	B5	LC	1,5
2	4	5	631	544	6	17	58	6	B5	LB	2,5
4			601	541	6	6	44	4	B5		
5			30	3	0	11	14	2	B3		
3	26	27	60	8	16	4	23	9	B3	LC	2,5
26			32	8	2	0	20	2	B3		
27			28	0	14	4	3	7	BP		

Rys. 2. Struktura drzewa nr 3 dla pierwszych badań laboratoryjnych



Rys. 3. Sekwencja kosztów drzewa nr 3 dla pierwszych badań laboratoryjnych

Kolejnym krokiem w analizie zbudowanego modelu drzewa była ocena trafności predykcji. Najprostszym narzędziem do oceny poprawności klasyfikacji [5] jest powstała macierz błędnych klasyfikacji (rys. 5). Macierz ta zestawia klasy obserwowane i klasy przewidywane. W naszym przypadku otrzymaliśmy automatycznie wyliczone błędy pre-

dykcji, które znajdują się w wierszu „Procent z wiersza" dla poszczególnych przewidywanych decyzji podiagnostycznych.

Przykładowo, dla klasy obserwowane „BP" błąd predykcji dla klasy „Przewidywana B3" wynosi 0,29%. Należy tu podkreślić, że o przydatności zbudowanego modelu drzewa nie decyduje wyłącznie trafność predykcji całego rozwiązania, ale również trafność predykcji poszczególnych klas modelu. Macierz błędnych klasyfikacji jest także punktem wyjścia do obliczenia innych miar, np. czułości, specyficzności, miary F (F measure) czy średniej G (G-mean).

Sekwencja drzew (KW4 RB=1)
Zmienna zależna: DEC
Najlepsze drzewo oznaczono *

	Końcowe węzły	SK koszt	SK std. błąd	Resubstytucji koszt	Węzeł złożoność
Drzewo 1	6	0,154848	0,013762	0,137482	0,000000
Drzewo 2	5	0,151954	0,013656	0,138929	0,001447
*Drzewo 3	4	0,160637	0,013969	0,147612	0,008683
Drzewo 4	3	0,176556	0,014505	0,163531	0,015919
Drzewo 5	2	0,180897	0,014644	0,179450	0,015919
Drzewo 6	1	0,201158	0,015250	0,201158	0,021708

Rys. 4. Sekwencja drzew dla pierwszych badań laboratoryjnych

Macierz klasyfikacji 3 (KW4 RB=1)
Zmienna zależna: DEC
Opcje: Jakościowa zależna, Próba do analizy

	Obserw.	Przewidywana B5	Przewidywana BP	Przewidywana Z	Przewidywana B3	Przewidywana W	Łącznie w wierszu
Liczba	B5	541			11		552
Procent z kolumnny		90.02%	0.00%		17.74%		
Procent z wiersza		98.01%	0.00%	0.00%	1.99%	0.00%	
Procent z ogółu		78.29%	0.00%	0.00%	1.59%	0.00%	79.88%
Liczba	BP	6	14		2		22
Procent z kolumnny		1.00%	50.00%		3.23%		
Procent z wiersza		27.27%	63.64%	0.00%	9.09%	0.00%	
Procent z ogółu		0.87%	2.03%	0.00%	0.29%	0.00%	3.18%
Liczba	Z	6	4		11		21
Procent z kolumnny		1.00%	14.29%		17.74%		
Procent z wiersza		28.57%	19.05%	0.00%	52.38%	0.00%	
Procent z ogółu		0.87%	0.58%	0.00%	1.59%	0.00%	3.04%
Liczba	B3	44	3		34		81
Procent z kolumnny		7.32%	10.71%		54.84%		
Procent z wiersza		54.32%	3.70%	0.00%	41.98%	0.00%	
Procent z ogółu		6.37%	0.43%	0.00%	4.92%	0.00%	11.72%
Liczba	W	4	7		4		15
Procent z kolumnny		0.67%	25.00%		6.45%		
Procent z wiersza		26.67%	46.67%	0.00%	26.67%	0.00%	
Procent z ogółu		0.58%	1.01%	0.00%	0.58%	0.00%	2.17%
Liczba	Ogół grup	601	28		62		691
Procent łącznie		86.98%	4.05%	0.00%	8.97%	0.00%	

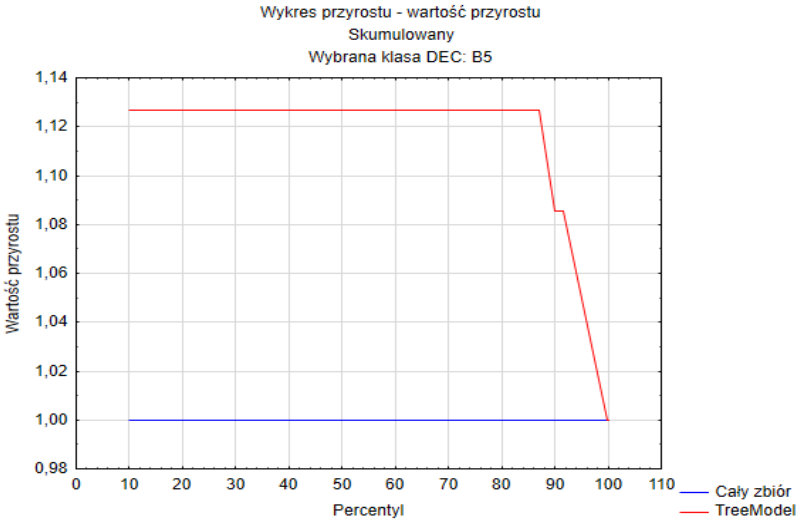
Rys. 5. Macierz błędnych klasyfikacji dla pierwszych badań laboratoryjnych

Kolejnym krokiem było zestawienie rankingu ważności predyktorów oraz ewentualne sporządzenie ich wykresu. Tabela przedstawiona na rys. 6 przedstawia nam taki ranking ważności predyktorów. Widać w niej, że predyktor „LE" jest nieistotny statystycznie, a predyktor „LC" jest najbardziej istotny, czyli najbardziej ważny (zmn. ranga).

Ważność predyktorów 3 (KW4 RB=1)
Zmienna zależna: DEC
Opcje: Jakościowa zależna, Drzewo numer 3

	zmn. ranga	Ważność
LC	100	1,000000
LD	82	0,816631
LB	43	0,425306
LA	38	0,382962
LE	0	0,000000

Rys. 6. Ranking ważności predyktorów dla pierwszych badań laboratoryjnych



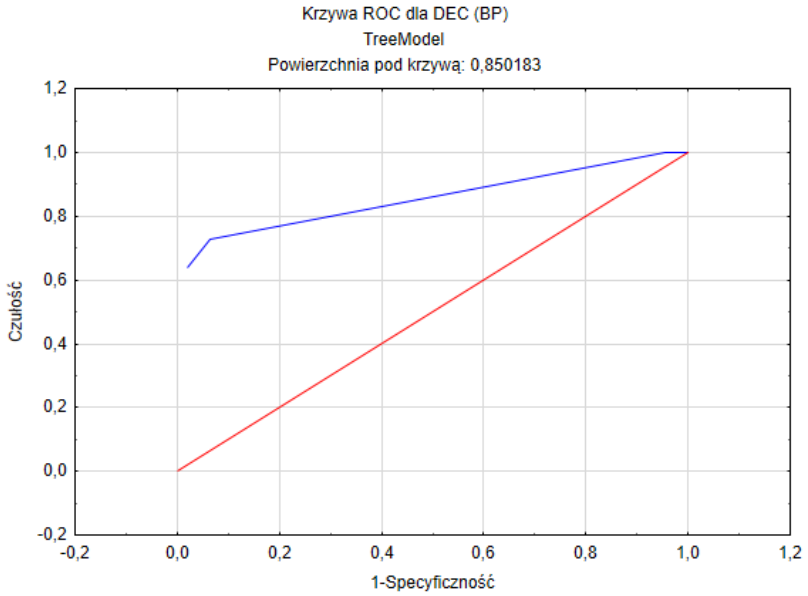
Rys. 7. Wykres przyrostu dla decyzji „B5"

W oprogramowaniu [8] występuje jeszcze jedno narzędzie pozwalające ocenić jakość powstałego modelu predykcyjnego. Jest nim wykres skumulowanej wartości współczynnika przyrostu (*lift chart*), który jest graficznym podsumowaniem użyteczności modelu do przewidywania wartości zmiennej zależnej.

Przykładowy wykres wartości przyrostu dla klasy „B5" przedstawia rys. 7. Wykresy tego typu zostały stworzone dla wszystkich decyzji podiagnostycznych. Na osi (y) wyświetlane są wartości przyrostu, czyli wielokrotności względem linii odniesienia, a na

osi (x) wpisane są wartości „percentyl”, czyli intuicyjnie mówiąc, percentyl jest wielkością, poniżej której padają wartości zadanego procenta próbek.

Z wykresu tego można odczytać, że biorąc np. 10% przypadków najpewniej zaklasyfikowanych do klasy „B5” (o najwyższych prawdopodobieństwach klasyfikacyjnych) otrzymamy próbkę, która zawiera ok. 1,128 razy więcej przypadków niż gdyby wybór był losowy. Powstały wykres tworzony jest dla zmieniających się zbiorów, które zawierają zwiększającą się liczbę przypadków o największym prawdopodobieństwie trafienia do klasy, wynikającym z tworzonego modelu tak, że następny zawiera poprzedni (w tym sensie wykres jest skumulowany).



Rys. 8. Wykres krzywej „ROC” dla decyzji „BP”

Kolejnym narzędziem służącym ocenie działania zbudowanego modelu jest krzywa „ROC” (Receiver Operating Characteristic), która ma wiele zastosowań, w szczególności służy ona do oceny jakości modelu przewidującego przynależności do różnych klas. Bardzo użytecznym wskaźnikiem jakości powstałego modelu jest pole pod tą krzywą „ROC”: im jest ono większe, tym zbudowany model jest lepszy. Przykładową taką krzywą dla klasy „BP” przedstawia rys. 8. Jak widać, powierzchnia pod krzywą wynosi 0,850183, co oznacza, że zbudowany model w zakresie tej klasy jest dobry. Krzywe typu „ROC” zostały opracowane dla wszystkich możliwych rodzajów zmiennych zależnych w modelu, czyli dla decyzji podiagnostycznych.

5. Podsumowanie

W artykule podjęto próbę zaprojektowania oraz zbudowania modelu drzewa decyzyjnego według metody C&RT dla zapłonników typu KW-4, które stosowane są w nabojach artyleryjskich kalibru od 57 mm do 152 mm. Cel określony na początku artykułu został w pełni osiągnięty. Określono niezbędne parametry do zbudowania prawidłowego drzewa, które w efekcie finalnym doprowadziły do zbudowania i wyboru najlepszego modelu drzewa klasyfikacyjnego. Dzięki specjalistycznemu oprogramowaniu komputerowemu [8] zbudowano model drzewa, którego najlepszą postać (nr 3) automatycznie wybrało to oprogramowanie. Wykonana procedura przycinania drzewa, ograniczyła jego wielkość i przedstawiła postać tego drzewa, które jest najbardziej optymalne dla założonych na początku predyktorów drzewa. Proces przycinania jest realizowany zgodnie z zasadą: lepiej mieć mniejszy model, ponieważ jest on bardziej uniwersalny w praktyce.

Dzięki zaprojektowaniu i zbudowaniu naszego drzewa klasyfikacyjnego, mamy teraz możliwość dokonywania szybkiej oceny dla nowych badanych partii zapłonników typu KW-4. Wdrożenie tego drzewa decyzyjnego do procesu oceny wymaga podłączenia komputerów z nowymi wynikami badań do oprogramowania [8], w którym zainstalowany jest nasz zaprojektowany i zbudowany model ocenowy. Otrzymane z badań nowe wartości predyktorów umożliwiają zaprojektowanemu modelowi drzewa szybką ocenę tych nowych partii zapłonników, poprzez automatyczne wyznaczenie decyzji podiagnostycznej. Oczywiście, zastosowanie zbudowanego drzewa klasyfikacyjnego w praktyce wymaga zgody na implementację tego narzędzia. Z praktycznego jednak punktu widzenia, zastosowanie tego narzędzia wydaje się więc koniecznością, ponieważ zbudowane drzewo klasyfikacyjne automatycznie wskazuje prawidłową decyzję podiagnostyczną dla nowych badanych partii zapłonników. Unika się w ten sposób najsłabszego ogniwa w procesie ocenowym, a mianowicie elementu ludzkiego.

Wydaje się, że sztuczna inteligencja, która szturmem wchodzi do większości dziedzin naukowych, będzie stosowana również do oceny wyników badań elementów specjalnych obiektów technicznych, a takimi są przecież zapłonniki typu KW-4. Podobne drzewa klasyfikacyjne można zaprojektować i zbudować dla innych badanych rodzajów zapłonników artyleryjskich. Warunkiem niezbędnym jest posiadanie bazy danych, która będzie zawierała wyniki przeprowadzonych dotychczasowych diagnostycznych badań laboratoryjnych.

6. Literatura

1. Amunicja wojsk lądowych. Wydawnictwo MON, Szefostwo Służby Uzbrojenia i Elektroniki, 1985.
2. Breiman L., Friedman J.H., Olsen R.A., Stone C.J.: Classification and Regression Trees. Chapman & Hall, 1984.
3. Karty badań laboratoryjnych zapłonników typu KW-4. Archiwum Wojskowego Instytutu Technicznego Uzbrojenia (WITU).
4. Koronacki J., Ćwik J.: Statystyczne systemy uczące się. Akademicka Oficyna Wydawnicza Exit, Warszawa 2008.
5. Łapczyński M., Demski T.: Data mining – metody predykcyjne. Statsoft Polska – materiały szkoleniowe z kursu, Kraków 2019.
6. Metodyka badań diagnostycznych amunicji – Indeks N-5001b – Archiwum WITU, 1985.
7. Sprawozdania z badań środków bojowych – Archiwum WITU.
8. Statystyka 13.3 PL, oprogramowanie komputerowe. Statsoft Polska 201.