COMPUTER METHODS IN MATERIALS SCIENCE

2022, vol. 22, no. 4, 211–228
https://doi.org/10.7494/cmms.2022.4.0791

# Rule modeling of ADI cast iron structure for contradictory data

Artur Soroczyński*, Robert Biernacki, Andrzej Kochański

Warsaw University of Technology, Institute of Manufacturing Technologies, Warsaw, Poland.

## Abstract

Ductile iron is a material that is very sensitive to the conditions of crystallization. Due to this fact, the data on the cast iron properties obtained in tests are significantly different and thus sets containing data from samples are contradictory, i.e. they contain inconsistent observations in which, for the same set of input data, the output values are significantly different.

The aim of this work is to try to determine the possibility of building rule models in conditions of significant data uncertainty. The paper attempts to determine the impact of the presence of contradictory data in a data set on the results of process modeling with the use of rule-based methods. The study used the well-known dataset (Materials Algorithms Project Data Library, n.d.) pertaining to retained austenite volume fraction in austempered ductile cast iron. Two methods of rule-based modeling were used to model the volume of the retained austenite: the decision trees algorithm (DT) and the rough sets algorithm (RST).

The paper demonstrates that the number of inconsistent observations depends on the adopted data discretization criteria. The influence of contradictory data on the generation of rules in both algorithms is considered, and the problems that can be generated by contradictory data used in rule modeling are indicated.

**Keywords:** rule modeling, contradictory data set, uncertainty, data preparation, decision tree, rough set theory

## 1. Introduction

Ductile iron, in all its varieties, is the fastest-growing replacement for other construction materials in many different applications for the parts of machines working under heavy loads, e.g. agricultural, automotive, mining ones (Barbosa et al., 2015; Colin García et al., 2021; Kochański et al., 2015; Wieczorek et al., 2022). Searching for substitutes forces one to answer the following question: can the new material meet the requirements of the one used so far? In such situations, the best solution is to use a model for forecasting material properties as a function of given process parameters or selecting the ranges of process parameters as a function of the desired product properties.

Building a generalized model for nodular cast iron and for ausferritic cast iron that allows the prediction of properties in a wide range of variability of input parameters faces a number of problems resulting from the nature of the modeled material itself. Due to the nature of cast iron, the obtained test results are highly uncertain or contradictory. This also happens when the cast iron melts are carried out under laboratory conditions.

In many publications, the research results presented indicate the sensitivity of cast iron. Even slight changes

*Corresponding author: artur.soroczynski@pw.edu.pl
ORCID ID's: 0000-0002-5935-3690 (A. Soroczyński), 0000-0002-2294-5879 (R. Biernacki), 0000-0002-1060-1390 (A. Kochański)

in the implemented process, such as a slight difference in the rate of heat dissipation, may cause significant differences in the measured values (Dal Corobbo & Arias, 2009; Wohlfahrt et al., 2010). The work by Dal Corobbo & Arias (2009) presents the results of hardness measurements of a ring with an internal diameter of ⌀390 mm and an external diameter of ⌀520 mm. Average hardness measurements of castings made of the same alloy on the inner and outer surfaces differed from 13 HV to 38 HV, i.e. from 4.5% to 13%. The influence of heat dissipation rate significantly influences the size of the graphite particles (Wohlfahrt et al., 2010). The samples cut from the Y test castings differed significantly in the size of graphite – they measured 25 μm and 32 μm in the lower and upper parts of the casting. In the case of alloys with a clear tendency to segregation, such as in work (Heydarzadeh et al., 2004), cast in massive molds, the average strength measured in the upper part and in the lower part of the casting in the selected case was 1275 MPa and 1475 MPa (reading from the chart included in the publication), i.e. it was higher by approx. 15%. The second reason for such an uncertainty of the data is the nature and manner of carrying out the measurement itself. This obtains, for example, in the case of measuring the elongation A5, in which a non-metallic inclusion, carbide precipitation or bubbles of unobservable size can significantly reduce the result value. Examples of such scattered measurements can be found in (Olofsson et al., 2011), where at the hardening temperatures of 250°C, 300°C, 350°C and 400°C and during 1 h, elongation in the following ranges was recorded: 1 – 5%; 3 – 4%; 5 – 10.5% and 7.5 – 10%, and for the 2 h time: 1 – 1.5%; 3 – 4.5%; 3 – 7% and 4 – 9%. In turn, the measurement of impact toughness is strongly determined, among others, by: the place where the samples were cut (Szykowny et al., 2010), the orientation of the samples (Chawla et al., 2008) and the method of sample preparation (Nobuki et al., 2010). The paper by Szykowny et al. (2010) shows the influence of the distance of the notch from the top surface of the mold. According to Chawla et al. (2008), the direction of cutting out the impact specimens is important, as crack propagation always starts with the graphite spheroid, but continues along the austenite-ferrite boundary. It was shown in the work by Nobuki et al. (2010) that despite the much lower sensitivity of cast iron, as compared to steel, to the notch shape in impact samples, it has, nevertheless, a significant influence.

The effect of the sensitivity of cast iron and the uncertainty of measurements is that the data sets containing the chemical compositions of the melts, parameters of the smelting and heat treatment process, pouring parameters and the properties of the obtained castings, both in the case of those derived from laboratory experiments and those from industrial research, contain numerous contradictions (Kochański et al., 2012).

The fact that there are contradictions in a set, even if they are numerous, does not make it impossible to build a model. The properties of ductile iron have been repeatedly modeled with the use of soft mathematical models, but prediction models were usually used such as: multiple linear regression, artificial neural networks, support vector machine, projection pursuit regression (Kochański et al., 2012; Perzyk & Kochański, 2001; Perzyk et al., 2015; Rodríguez-Rosales et al., 2022; Wilk-Kołodziejczyk et al., 2018). Less frequently, work was undertaken on property modeling with the use of rule-creating tools based on the theory of fuzzy sets and decision trees (Kochański et al., 2013, 2014; Perzyk & Soroczyński, 2008, 2019; Perzyk et al., 2011). It seems that an important reason for the use of contradictory algorithms such as, e.g., artificial neural networks for sets is their greater ability to generalize, i.e. to ignore or eliminate the influence of contradictory observations on the model. The fact that rule modeling is less resistant to contradictory data should not lead to the abandonment of this type of modeling. It has a number of advantages, as well as the significant advantage of the easy interpretation of the operation of the model itself.

Therefore, the aim of the work was to try to determine the possibility of building rule models in conditions of significant data uncertainty, and then to assess the predictive capabilities of such models. An addional aim was to indicate the risks present in such models.

## 2. Austempered ductile iron database

In this paper, the authors decided to indicate the risks arising from the uncertainty of the data in the set used for rule models, such as those built with the use of e.g. decision trees and the theory of raw sets. For this purpose, a generally available set was used, containing the chemical compositions of melts and the amount of residual austenite in the structure of cast iron after heat treatment (Yescas et al., 2001). The data file is available online on the MAP website (Materials Algorithms Project Data Library, n.d.).

The collection was created as a database of research results from many scientific works. As described, the collection contains 1910 observations from research results and experiments published in peer-reviewed scientific journals. The data in the file was saved in 12 (11 + 1) columns containing:

– chemical composition – 6 elements: Carbon, Silicon, Manganese, Molybdenum, Nickel, Copper [wt%];

– four heat treatment parameters: austenitising temperature [°C]; austenitising time [min], austempering temperature [°C], austempering time [min];
– retained austenite volume fraction [%];
– the reference (source of data).

Table 1 shows the correlation coefficients between the input variables and Figure 1 shows selected pairs of variables. The highest value of the correlation coefficient was observed for the pair: austempering temperature – the amount of residual austenite, and the lowest for the pairs molybdenum – austempering time and carbon – molybdenum. They are respectively: 0.350 and 0.0008 and 0.0036. The table also shows five pairs with a correlation coefficient modulus in the range 0.25–0.28. The above may prove the independence of the input variables, but it is not sufficient proof.

The graphs of the pairs of variables shown in Figure 1 illustrate the even distribution of observations in terms of the variability of parameters.

**Table 1.** Correlation coefficients of the input and output variables

| | C | Si | Mn | Mo | Ni | Cu | *T* austenit. | *t* austenit. | *T* austemp. | *t* austemp. | Ret_ aust |
|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 1.0000 | −0.1109 | −0.1127 | −0.0036 | −0.1296 | −0.2760 | 0.1531 | −0.0113 | −0.0304 | −0.0337 | −0.0589 |
| Si | – | 1.0000 | −0.2594 | −0.0727 | −0.0385 | 0.2488 | 0.1045 | −0.0599 | 0.0502 | 0.0086 | 0.0840 |
| Mn | – | – | 1.0000 | −0.0872 | −0.1286 | −0.2572 | −0.1078 | −0.0084 | −0.0798 | 0.1002 | 0.0891 |
| Mo | – | – | – | 1.0000 | −0.0287 | −0.0539 | −0.1130 | 0.1999 | −0.0599 | 0.0008 | −0.1042 |
| Ni | – | – | – | – | 1.0000 | 0.2650 | 0.0314 | −0.1699 | −0.0483 | 0.0217 | 0.1030 |
| Cu | – | – | – | – | – | 1.0000 | 0.0408 | −0.0297 | 0.0638 | 0.0227 | 0.1511 |
| *T* austenit. | – | – | – | – | – | – | 1.0000 | 0.0258 | 0.0102 | −0.0309 | 0.2547 |
| *t* austenit. | – | – | – | – | – | – | – | 1.0000 | −0.1259 | −0.1507 | −0.0987 |
| *T* austemp. | – | – | – | – | – | – | – | – | 1.0000 | −0.0252 | 0.3500 |
| *t* austemp. | – | – | – | – | – | – | – | – | – | 1.0000 | −0.1999 |
| Ret_aust | – | – | – | – | – | – | – | – | – | – | 1.0000 |

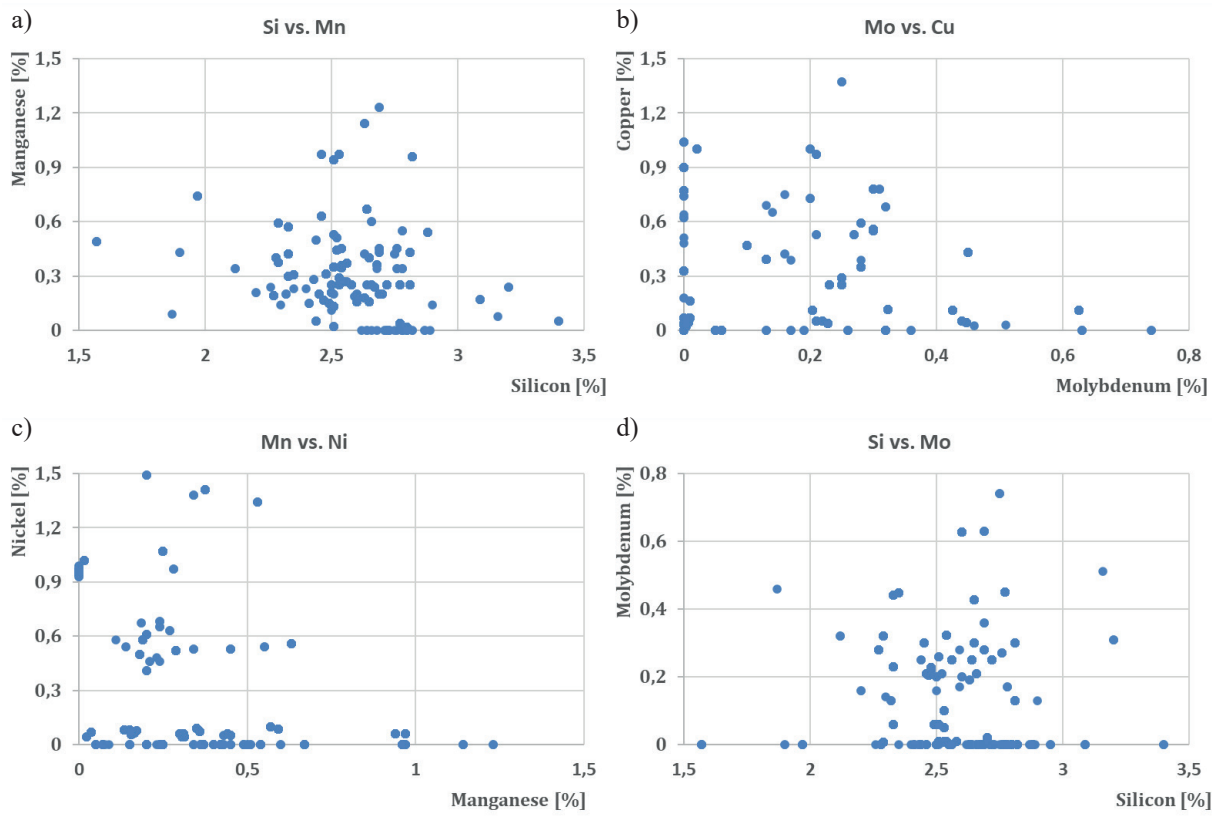



**Fig. 1.** Distribution of variability of pairs of input variables: a) silicon – manganese; b) molybdenum – copper; c) manganese – nickel; d) silicon – molybdenum

In order to confirm this observation, tests of independence were performed for pairs of variables. Figure 2 shows the distribution of the variability of the pair of variables with the highest correlation coefficient, Autemperization temperature vs Residual austenite content, equal to 0.35.

For a pair of variables, the null hypothesis of the independence of variables was verified using the $\chi 2$ test (chi-square test). An alternative hypothesis was the dependence of the variables in the discussed set. Figure 3 presents a contingency table for the variables of austempered temperature and retained austenite.

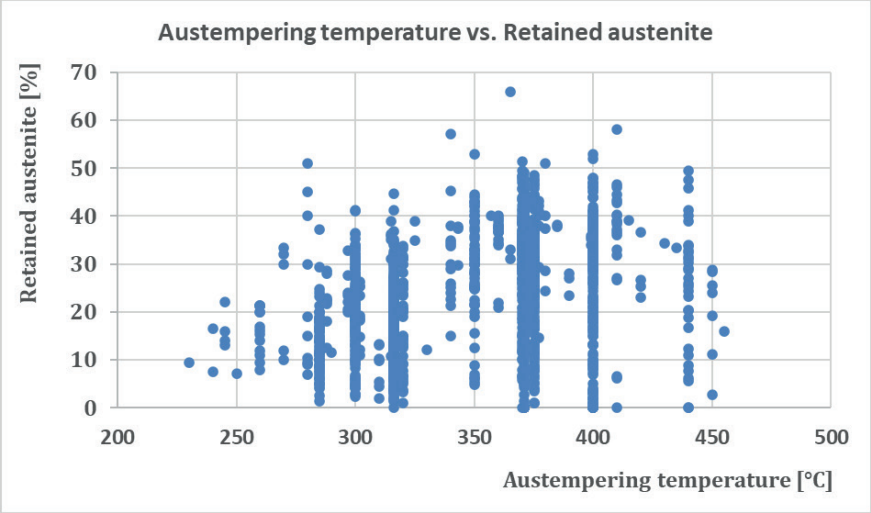



**Fig. 2.** Distribution of variability of pairs of input variables: austempering temperature – retained austenite

| Retained austenite (equal – width discretization) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **11** | 0.0 | 0.1 | 0.7 | 0.9 | 2.0 | 7.7 | 0.3 | 0.0 | 3.4 | 0.0 | 0.2 |
| | **10** | 0.2 | 0.4 | 0.5 | 4.0 | 9.4 | 0.3 | 2.2 | 0.4 | 9.5 | 0.2 | 5.1 |
| | **9** | 0.4 | 0.9 | 2.4 | 4.9 | 15.6 | 12.7 | 4.3 | 2.9 | 2.9 | 0.3 | 0.0 |
| | **8** | 1.0 | 2.3 | 9.3 | 16.8 | 38.7 | 0.4 | 26.1 | 5.7 | 17.4 | 1.6 | 3.1 |
| | **7** | 1.1 | 0.1 | 10.7 | 3.4 | 14.8 | 4.4 | 1.3 | 14.8 | 3.5 | 1.1 | 0.4 |
| | **6** | 1.1 | 2.6 | 7.5 | 0.3 | 7.5 | 2.0 | 10.2 | 3.7 | 3.7 | 1.1 | 1.1 |
| | **5** | 0.0 | 3.5 | 0.6 | 5.5 | 13.1 | 0.5 | 1.0 | 5.3 | 7.0 | 0.0 | 0.2 |
| | **4** | 1.0 | 0.2 | 7.7 | 33.6 | 18.5 | 4.4 | 14.4 | 11.8 | 4.1 | 0.9 | 0.2 |
| | **3** | 1.1 | 3.1 | 21.9 | 0.1 | 40.2 | 7.3 | 10.0 | 2.5 | 14.4 | 0.9 | 0.7 |
| | **2** | 9.0 | 0.2 | 26.6 | 0.0 | 10.6 | 0.1 | 15.2 | 1.3 | 0.9 | 0.5 | 0.2 |
| | **1** | 0.4 | 1.0 | 0.2 | 0.0 | 0.0 | 2.3 | 0.1 | 3.7 | 17.2 | 0.4 | 0.3 |
| **639.3** | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
| | | Austempered temperature (equal – width discretization) | | | | | | | | | | |

**Fig. 3.** Contingency table for pairs of variables: austempered temperature and retained austenite

The value of the test statistic is 639.3 (shown in Figure 3, in the lower left corner), which supports the rejection of the hypothesis of independence of variables in the study set.

The analysis of the data gathered in the base (Yescas et al., 2001) showed that a number of observations are contradictory. Examples of such data are shown in Table 2 and in Table 3. The contradictions concern both obtaining different values of the output variable (amount of retained austenite) for the same values of the input variables, and the same value of the output variable (amount of austenite) for different values of one input variable (austempering time). The first situation is shown in the example of pairs of observations 1151 and 1166, as well as 63 and 72 (in this case, the differences in the chemical composition are negligibly small, within the limits of the measurement error) from the set, as shown in Table 2. The second type of contradiction, in which for pairs differing in the value of one input variable, the same output variable value was recorded, is shown in Table 3 on the example of observations 1803 and 1804 and 1572, 1577 and 1580.

**Table 2.** Examples of contradictory observations recorded in the collection; contradictory data of the first type (Yescas et al., 2001)

| No. of observ. | C | Si | Mn | Mo | Ni | Cu | Austenitising temp./time | | Austempering temp./time | | Amount of retained austenite |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | % | % | % | % | % | % | °C | min | °C | min | % |
| … | … | … | … | … | … | … | … | … | … | … | … |
| 1151 | 3.81 | 2.54 | 0.347 | 0.323 | 1.545 | 0.113 | 871 | 120 | 316 | 840 | **34.6** |
| 1166 | 3.81 | 2.54 | 0.347 | 0.323 | 1.545 | 0.113 | 871 | 120 | 316 | 840 | **20.3** |
| … | … | … | … | … | … | … | … | … | … | … | … |
| 63 | 3.49 | 2.68 | 0.00 | 0.00 | 0.96 | 0.00 | 900 | 60 | 375 | 240 | **26.2** |
| 72 | 3.46 | 2.73 | 0.00 | 0.00 | 0.95 | 0.00 | 900 | 60 | 375 | 240 | **19.1** |

**Table 3.** Examples of contradictory observations recorded in the set of type II contradictory data (Yescas et al., 2001)

| No. of observ. | C | Si | Mn | Mo | Ni | Cu | Austenitising temp./time | | Austempering temp./time | | Amount of retained austenite |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | % | % | % | % | % | % | °C | min | °C | min | % |
| … | … | … | … | … | … | … | … | … | … | … | … |
| 1572 | 3.80 | 2.77 | 0.037 | 0.00 | 0.07 | 0.33 | 950 | 120 | 300 | **5** | 23.3 |
| 1577 | 3.80 | 2.77 | 0.037 | 0.00 | 0.07 | 0.33 | 950 | 120 | 300 | **120** | 23.3 |
| 1580 | 3.80 | 2.77 | 0.037 | 0.00 | 0.07 | 0.33 | 950 | 120 | 300 | **720** | 23.3 |
| … | … | … | … | … | … | … | … | … | … | … | … |
| 1803 | 3.50 | 2.65 | 0.40 | 0.30 | 1.60 | 0.55 | 871 | 120 | 260 | **90** | 21.4 |
| 1804 | 3.50 | 2.65 | 0.40 | 0.30 | 1.60 | 0.55 | 871 | 120 | 260 | **120** | 21.3 |

In the set, 34 pairs of observations of the first type of contradiction were found, i.e. different values of the output variable with the same values of the input variables. In the case of 10 pairs of observations, the difference in the measured content of retained austenite exceeded 15% of the measured value. This type of contradiction accounts for 2% of the recorded observations. 206 pairs were identified in the set involving observations with a contradiction of the second type, i.e. differing in the value of one input variable with the same output value. This contradiction concerned only three variables: austenitization temperature, austempering temperature and austempering time and occurred in 18, 20 and 168 cases respectively. In total, this accounts for 10.7% of registered observations. It should be noted that the two types of contradictions mentioned do not characterize all of the possible contradictions present in the dataset.

## 3. Data preparation

According to Grzegorzewski & Kochański (2019, p. 29): "Obtained raw data should not be included into essential inference without their cross-examination. This step, called data preprocessing, is necessary to examine whether data are genuine or faked, to detect possible measurement errors, recording errors and outliers, to test validity of prior information, to get rid of irrelevant or redundant information and so on". The position of the data preparation task in the whole analysis process is shown in Figure 4.

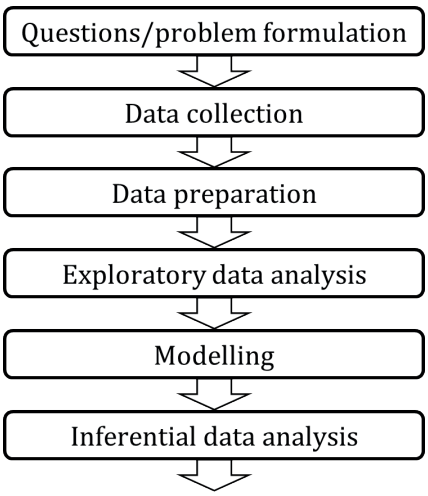



**Fig. 4.** Structure of data analysis (Grzegorzewski & Kochański, 2019)

The collected data was subjected to minimal preparation, which did not affect the inconsistencies in the dataset. As the aim was to analyze the impact of data contradiction on the rule model, it was decided that only outliers should be removed from the set. The distributions of the input and output variables were analyzed. Figure 5 shows two exemplary distributions: austenitization temperature and the amount of retained austenite. Tukey's tests were performed for the analyzed variables. The analysis of the distribution of variables showed that outliers were detected in two variables: the input variable, i.e. the austempering time, and the output variable, i.e. the amount of retained austenite. In the case of the first variable (austempering time), the observations with times shorter than 5 and longer than 4,320 minutes (corresponding to a heat treatment exceeding 72 hours) were removed from the set, as shown in Figure 6 (values indicated by green arrows).

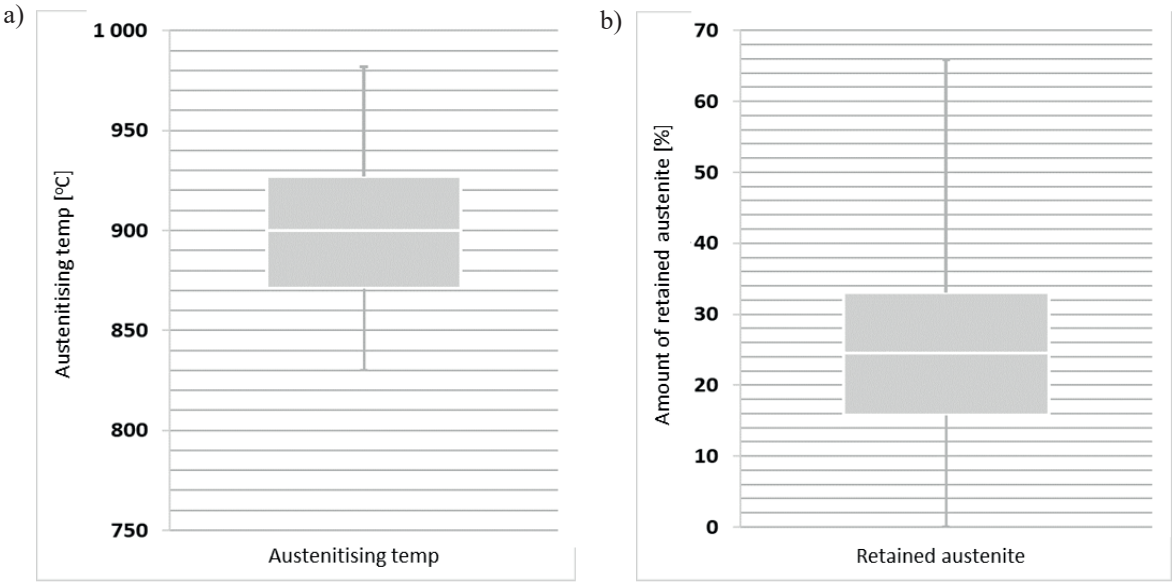



**Fig. 5.** Distribution of variables: a) austenitization temperature; b) amount of retained austenite in the original dataset

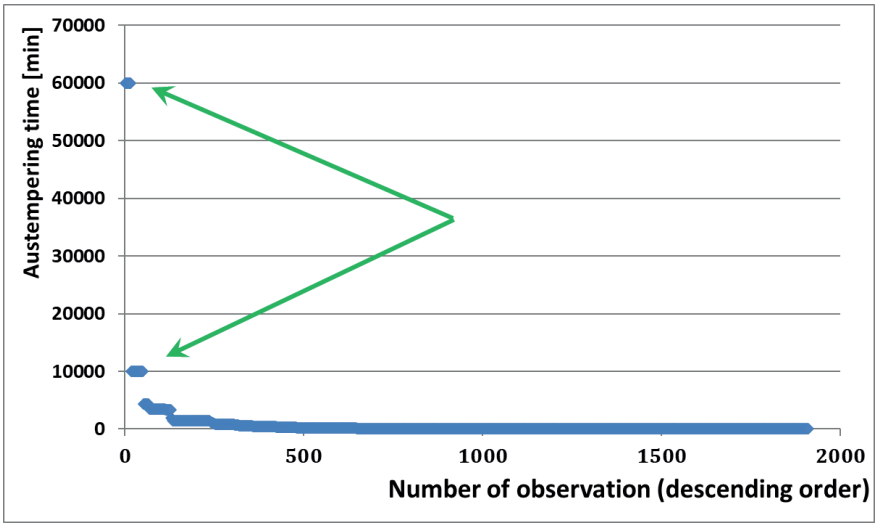
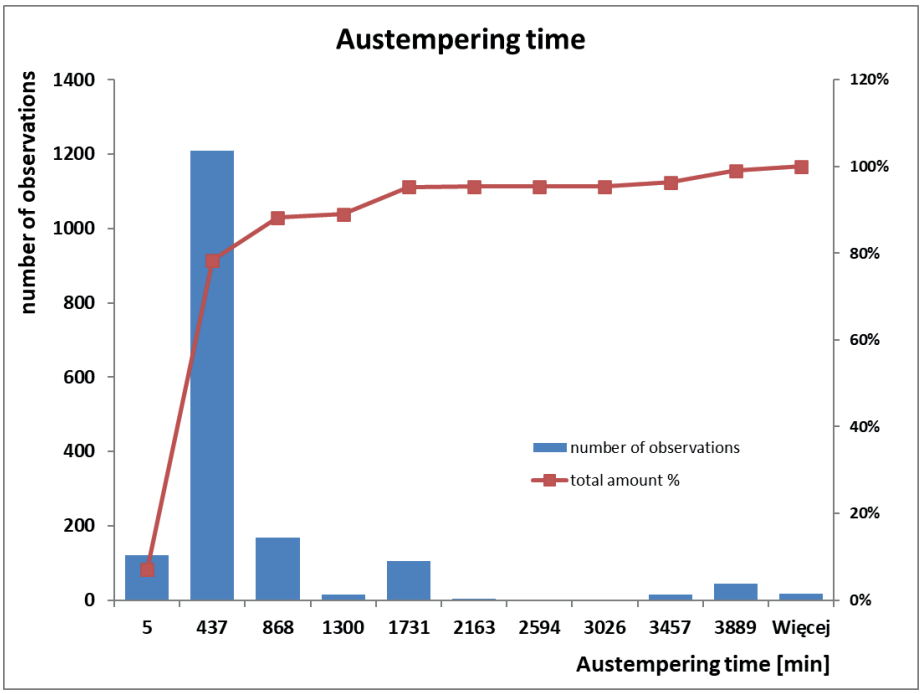



**Fig. 6.** Time of austempering in order from the largest to the smallest (green arrows indicate outliers, removed from the dataset in further analyzes)

After removing outliers for a given austempering time from the set of data, logarithmic normalization was performed, changing the variability range from the range 5–4320, to the range 0.699–3.635. The distribution of the original data showed a huge data density for times ranging from 5 to 435 [min], as shown in Figure 7a. After nonlinear logarithmic normalization, a uniform distribution was obtained, as shown in Figure 7b. The Shapiro–Wilk test performed on the transformed data did not allow for the rejection of the hypothesis with a normal distribution (for significance $\alpha = 0.05$).

a)




b)

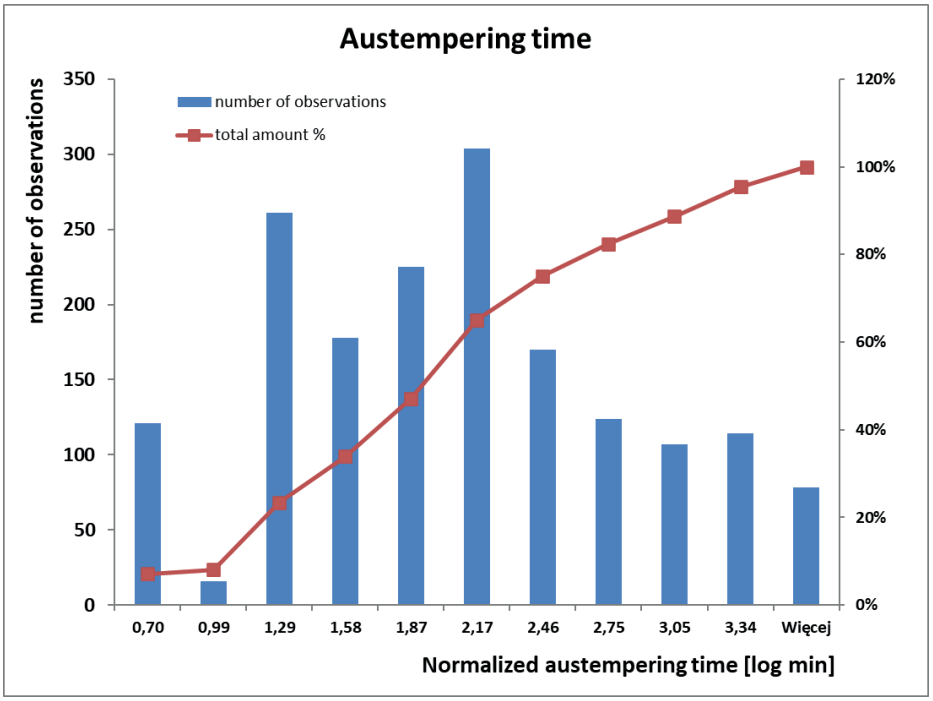



**Fig. 7.** Distribution of the variable austempering time: a) originally data [min]; b) after logarithmic normalization

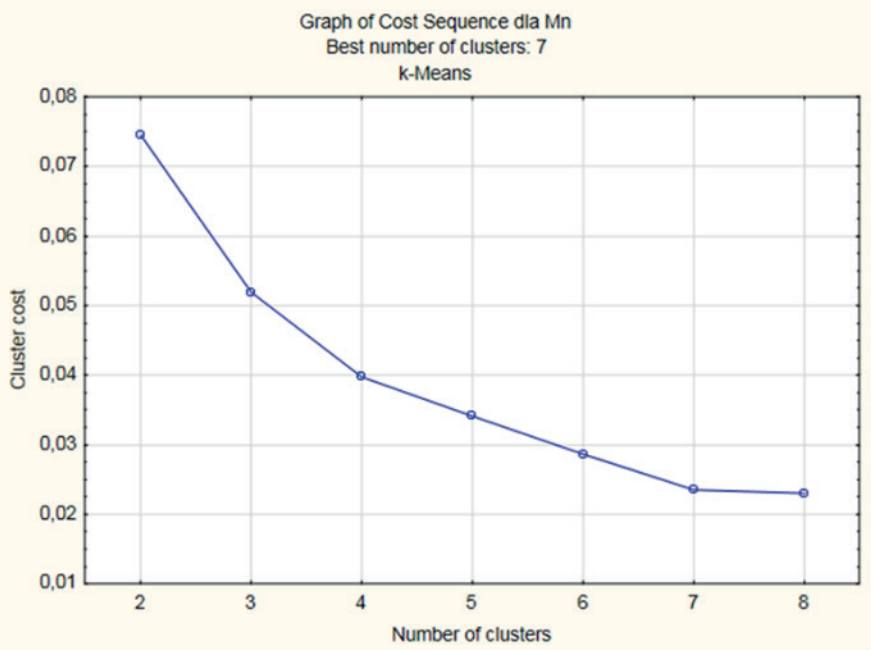
## 4. Discretization methods

All input and output variables were discretized in the dataset. Three methods of data discretization were used in the research. Two of them are the most commonly used discretization methods: equal width (*W*) and equal depth (*D*). The third, equally widely used, uses the *k*-means (*k*) cluster analysis algorithm. All methods require the number of intervals to be arbitrarily indicated. In the first two methods, discretization was made with the assumption of division into three, five and seven intervals. They were used to split the output variable. However, in the *k*-means method, using the cost function being the mean value of the observation distance from the centroid, discretization was performed for the suggested number of intervals. Examples of the split cost functions for the variables: Mn and Mo are shown in Figure 8. For the manganese

content in accordance with the adopted criterion, the number of intervals equal to 7 was chosen as the optimal, as shown in Figure 8a, while for molybdenum, the value of 5 was chosen, as shown in Figure 8b. The *k*-means method was used to divide both the input and output variables.

Table 4 presents a summary of the optimal number of intervals for the input parameters.

a)




b)

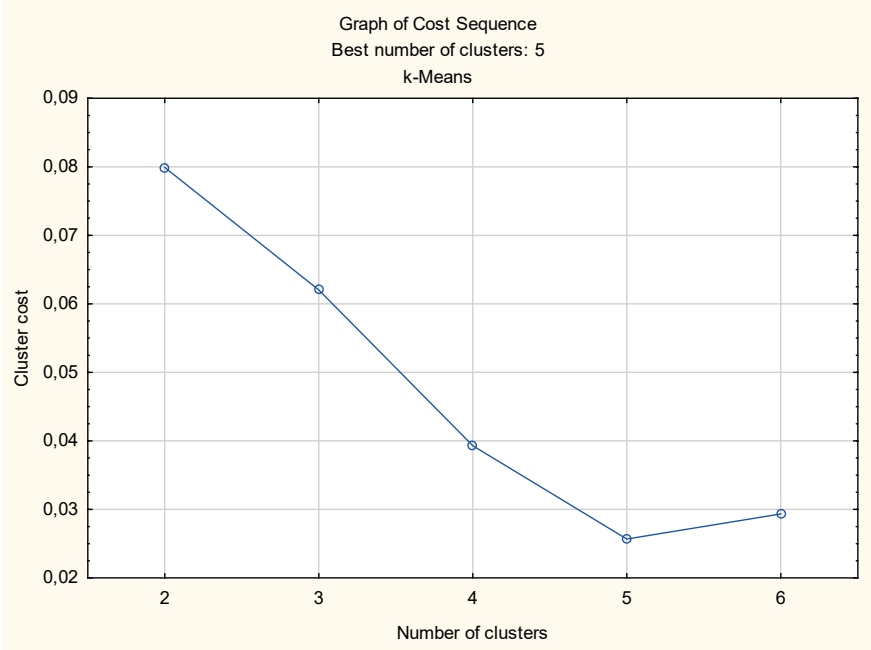



**Fig. 8.** Cost function for dividing the content of: a) manganese for the number of discretization intervals from 2 to 8; b) molybdenum for the number of intervals from 2 to 6

**Table 4.** Discretization of input signals using the *k*-means method according to the cost function

| | Elements | | | | | | Process parameters | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C | Si | Mn | Mo | Ni | Cu | austenitization | | austempering | |
| | | | | | | | temp. | time | temp. | time |
| No. of classes | 6 | 3 | 7 | 5 | 4 | 6 | 6 | 6 | 4 | 8 |

For the output variable, i.e. the amount of retained austenite, the division into 11 intervals was established for the adopted selection method (cost curve), as shown in Figure 9.

After the discretization of the input variables, the first type of contradiction was revealed (different values of the output variable for the same set of input data). For the analyzed set, examples of such contradictions resulting from data discretization are presented in Table 5. The left part of the table shows the values of the input variables discretized and the column *Retained_austenite* shows the values of the output variable, i.e. the amount of the residual austenite. The right part of the table shows the classes to which the output variable has been assigned, depending on the selected discretization method.

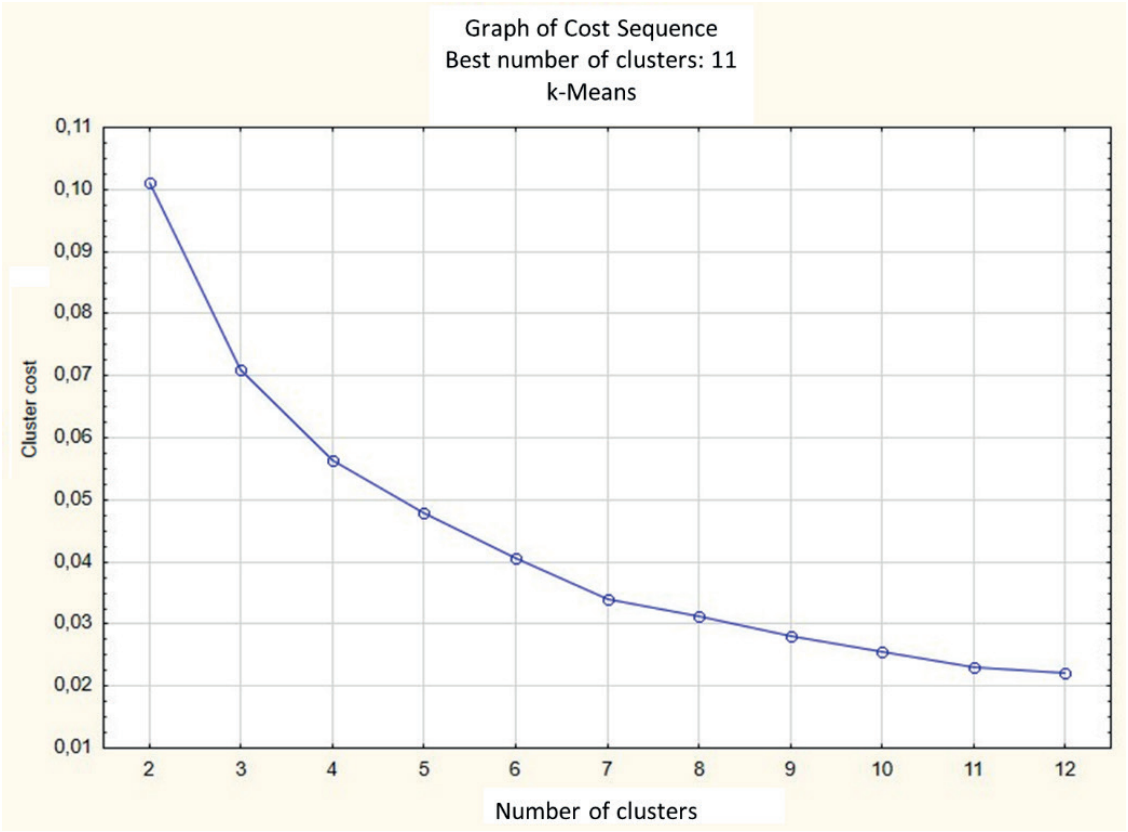



**Fig. 9.** Cost curve for the output variable – amount of restained austenite

**Table 5.** Selected records after discretization creating uncertainty

| No. | *k*-means optimal division (individual for each parameter) | | | | | | | | | | Ret_ aust | Equal width | | | Equal depth | | | *k*-means | | | *k*-means |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C | Si | Mn | Mo | Ni | Cu | *T* austenit. | *T* austenit. | *T* austemp. | log (*t* austemp.) | | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 | 11 |
| 1 | 6 | 1 | 5 | 4 | 4 | 1 | 2 | 4 | 1 | 8 | 16.4 | 1 | 2 | 3 | 1 | 2 | 2 | 1 | 2 | 2 | 3 |
| 2 | 6 | 1 | 5 | 4 | 4 | 1 | 2 | 4 | 1 | 8 | 20.3 | 2 | 3 | 4 | 2 | 2 | 3 | 2 | 2 | 3 | 4 |
| 3 | 6 | 1 | 5 | 4 | 4 | 1 | 2 | 4 | 1 | 8 | 21.6 | 2 | 3 | 4 | 2 | 2 | 3 | 2 | 3 | 3 | 4 |
| 4 | 6 | 1 | 5 | 4 | 4 | 1 | 2 | 4 | 1 | 8 | 22.6 | 2 | 3 | 4 | 2 | 3 | 3 | 2 | 3 | 3 | 4 |
| 5 | 6 | 1 | 5 | 4 | 4 | 1 | 2 | 4 | 1 | 8 | 34.6 | 3 | 4 | 5 | 3 | 4 | 6 | 3 | 4 | 5 | 8 |
| 6 | 6 | 1 | 5 | 4 | 4 | 1 | 2 | 4 | 1 | 8 | 35.1 | 3 | 4 | 5 | 3 | 5 | 6 | 3 | 4 | 5 | 8 |

Table 6 shows the numerical description of the data uncertainty in the database for the discretization of the output variable the content of residual austenite using the *k*-means method with the number of intervals equal to 11. After removing outliers, 1,698 observations remained in the database, including 1,161 observations

distinguishable in terms of input. A unique, single value for output occurred for 1,098 observations, 929 of which were distinguishable. This means that 64.66% of the observations are consistent observations.

**Table 6.** The number of occurrences in the database of contradictory and non-contradictory observations

| No. of predicted classes | No. of cases | No. of observations | Percentage of observations [%] |
|---|---|---|---|
| 1 | 929 | 1098 | 64.66 |
| 2 | 188 | 446 | 26.27 |
| 3 | 43 | 150 | 8.83 |
| 4 | 1 | 4 | 0.24 |
| 05–11 | 0 | 0 | 0 |

There were contradictions in 35.34% of the observations in the database. Two different classes of output occurred in more than 26% of the cases (see cases with number 1151 and 1166 in Table 2). For 43 observations, the digitized baseline variable was assigned to three classes (e.g., in the case shown in Table 5). In the case of one set of input variables, the output variable assumed four different values. No cases were found with an output variable assigned to 5 or more classes.

Table 7 presents a summary of the impact of all selected methods of discretization of the output variable on the set contradiction. For the analyzed dataset, in the corresponding divisions into 3, 5 and 7 classes, discretization of equal width is the least contradictory.

**Table 7.** The number of output classes covered by the distinguishable input record for all output variable discretization methods

| Discretization methods | | The number of output classes covered by the distinguishable input record | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Equal width | 3 | 1053 | 104 | 4 | – | – | – | – | – | – | – | – |
| | 5 | 1004 | 144 | 13 | 0 | 0 | – | – | – | – | – | – |
| | 7 | 982 | 153 | 26 | 0 | 0 | 0 | 0 | – | – | – | – |
| Equal depth | 3 | 1,039 | 116 | 6 | – | – | – | – | – | – | – | – |
| | 5 | 1001 | 142 | 17 | 1 | 0 | – | – | – | – | – | – |
| | 7 | 948 | 185 | 27 | 1 | 0 | 0 | 0 | – | – | – | – |
| *k*-means | 3 | 1050 | 104 | 7 | – | – | – | – | – | – | – | – |
| | 5 | 987 | 158 | 16 | 0 | 0 | – | – | – | – | – | – |
| | 7 | 958 | 172 | 31 | 0 | 0 | 0 | 0 | – | – | – | – |
| | 11 | 929 | 188 | 43 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## 5. Rule modelling

Two different methods of rule modeling were used in the work. The first one uses the decision tree algorithm (DT) and the second one is based on the rough set theory (RST). In order to compare the effectiveness of the methods, in both cases, modeling was performed using all previously discussed discretization methods. Both selected methods of rule-based modeling are commonly known and used, but, importantly, they work in a different way.

Decision trees are non-parametric classification models, built on the basis of data collected in a set, thanks to the division of the set into smaller subsets. Starting from the entire set, it is divided in such a way that the obtained subsets of the values of the classes of decision variables are as homogeneous (preferably identical) as possible. The division is done using one of the variables called the breaking variable. The partitioning procedure is repeated for the resulting subsets, resulting in a tree-like structure of the model. The points of division are called nodes. The subsets that are no longer subdivided, called leaves, are the results of the classification (the decision class which predominates in such a subset decides). Generally, tree construction requires a limited degree of detail. This is done either in the tree-building phase (e.g. by imposing a minimum number of examples on a node) or in the procedure of simplifying an already generated, excessively large tree (the so-called pruning). There are many algorithms for generating (inducing) decision trees, differing in terms of the criterion for assessing the homogeneity of classes when divided at a node and the criterion of the degree of detail, i.e. tree expansion. Note that any tree path from the root to the leaf of the tree can be written as a logical rule. Decision

trees also allow for the assessment of the relative significance of attributes, based on the so-called cleanliness of the divisions in the nodes. The greater the increase in class homogeneity in the subsets of data obtained as a result of division with the use of a given variable, the greater its relative importance.

Generating rules using the rough set theory requires that not only the result variables, but also the attributes assume discrete values, i.e. nominal or ordinal values. Each distinguishable example is essentially a rule. The set of rules obtained in this way can usually be reduced or the rules appearing in it can be simplified (i.e. the conditional part may be shortened). This is done by deleting the attributes that do not contribute anything to the classification, i.e. after the omission of which the rule still indicates the same class of the result value (for all records present in the database). The rules are assessed primarily in terms of the uniqueness of the classification expressed by the so-called reliability of the rule. This parameter is defined as the ratio of the number of examples in which there is a given combination of attribute values and at the same time a given output class to the number of all examples in which this combination of attribute values occurs (i.e. also those in which the output class is different). Another rule evaluation parameter is the number (share) of examples corresponding to a given rule in the training set, called rule support. If it is not possible to obtain rules with 100% likelihood from the data set, then less explicit rules are used, usually assessed on the basis of a combination of reliability and support or other, more complex criteria. The rough set theory also makes it possible to easily assess the significance of individual attributes, based on the assessment of reducing the uniqueness of the classification by omitting a given attribute in all rules.

Models DTs were obtained using Statistica ver. 13.3 commercial software package. Various splitting conditions, stopping criteria and pruning parameters were tried out. RST procedure, oriented at generation of a full set of rules, was written by the present authors with a somewhat similar approach as used in the Explore algorithm (Stefanowski & Vanderpooten, 2001).

The differences in the operation of both algorithms were revealed in the rules obtained for the modeled set. The operation of the decision tree algorithm makes it possible to build rules in which the same variable is taken into account many times. This means that the same variable can be used in several nodes as a split criterion. An example of such a rule, obtained for the analyzed set, is shown in Figure 10. In the rule described by nodes 1, 2, 5, 15, *austenitization_temperature* (T_aust) was used twice as the division criterion.

Rule modeling for contradictory sets with the use of two modeling methods showed two significantly different solutions in terms of the quality of the obtained rules.

For the division of the output variable into three classes, the shortest rules were obtained for modeling using RST. For an exit class of 2, eight rules with two arguments were obtained. All these rules, presented in Table 8, had 100% support. For example:

If Mn = 2 and T_austenit = 5 than Ret_aust = 2

However, for decision trees, the shortest rule consisted of three arguments:

If T_austemp ≤ 1,5 and T_austenit ≤ 2,5 and Ni ≥ 2,5 than Ret_aust = 2

**Table 8.** A set of binary rules with 100% support (discretization methods: equal width, 3 classes, W3)

| C | Si | Mn | Mo | Ni | Cu | Austenitization | | Austempering | | Ret_aust (W3) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | temp. | time | temp. | log(time) | |
| – | 3 | – | – | – | – | – | – | 1 | – | 2 |
| – | – | – | – | – | – | 1 | – | – | 1 | 2 |
| – | – | – | – | – | – | 1 | – | – | 2 | 2 |
| 2 | – | – | – | – | – | – | – | – | 1 | 2 |
| – | – | 2 | – | – | – | 5 | – | – | – | 2 |
| – | – | – | – | – | – | – | 3 | – | 2 | 2 |
| – | – | – | – | – | 4 | – | – | – | 7 | 2 |
| 1 | – | – | – | – | – | – | – | – | 4 | 2 |

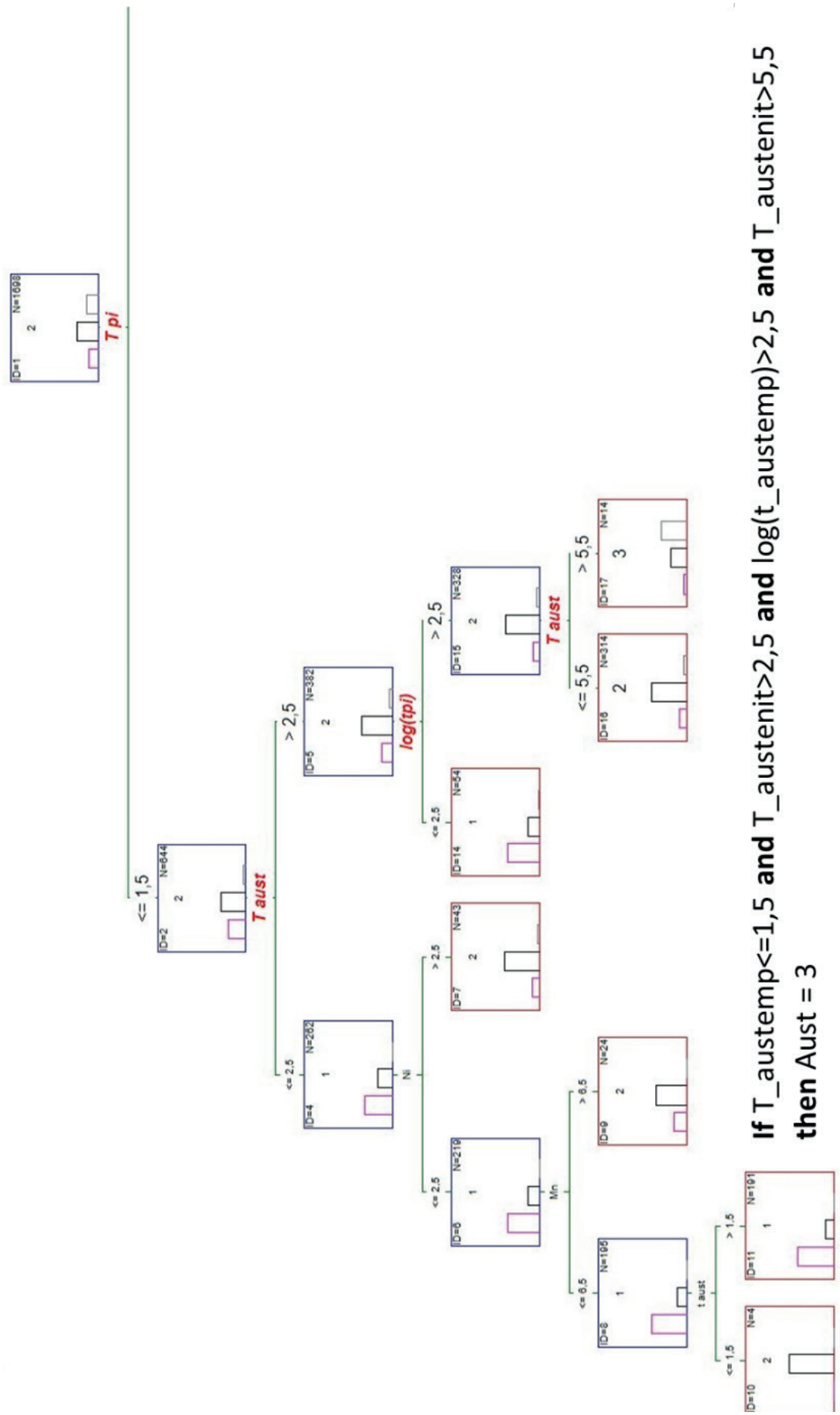



**Fig. 10.** Decision tree DT rules

Figure 11 shows the results obtained from the decision tree and rough set theory rules when dividing the output variable into three classes.

In the case of discretization of the output value into three intervals, the levels of correct predictions are close to each other and fluctuate around 70% for the decision tree (see Fig. 11a). This happens regardless of whether the discretization was done using the equal width, equal depth or *k*-means method. The differences in the proportion of correct predictions cannot be directly explained by the size of the individual classes (shown in Table 9) or the number of conflicting observations in the class (shown in Figure 12).

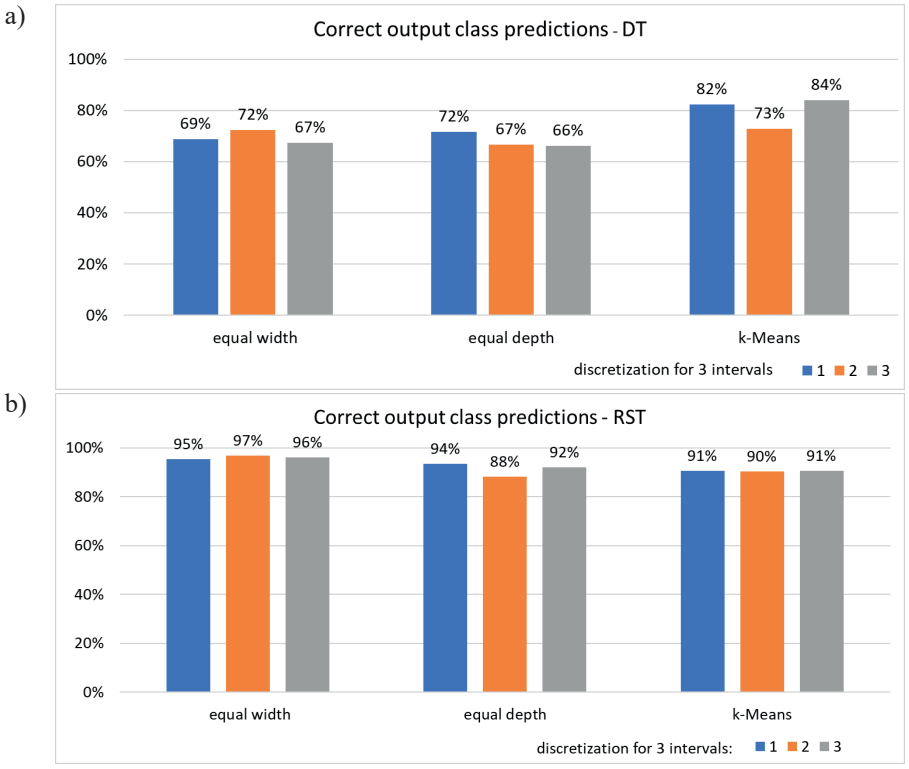



**Fig. 11.** Correct class prediction rules for dividing the output variable into 3 classes by: a) DT; b) RST

**Table 9.** The number of observations in each class for the three methods of discretization of the set into 3 classes

| | | Discretization method | | |
|---|---|---|---|---|
| | | equal width | equal depth | *k*-means |
| No. of classes | 1 | 391 | 566 | 431 |
| | 2 | 837 | 566 | 687 |
| | 3 | 470 | 566 | 580 |

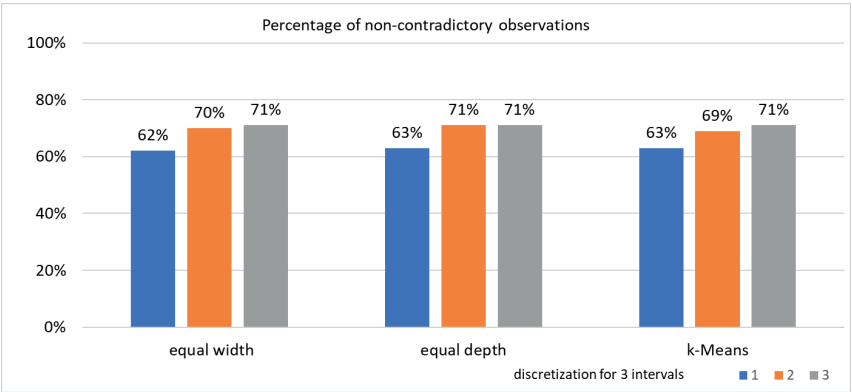



**Fig. 12.** Percentage of non-contradictory observations in individual classes for three discretization methods (for 3 classes)

Table 10 presents the results of the prediction distribution of models based on DT and RST on the example of a set with the division of the output variable into 3 classes. The results clearly show that the model based on RST makes more accurate predictions, which translates into the number of correct predictions (diagonal values marked in bold). This is due to the fact that the decision tree model induces rules with greater coverage but less accuracy, while the model based on RST makes more precise predictions at the expense of the number of induced rule sets.

**Table 10.** Prediction accuracy table for DT and RST on the example of a set divided into 3 classes – equal width

| | | Predicted | | | | | |
|---|---|---|---|---|---|---|---|
| | | RST | | | DT | | |
| | | 1 | 2 | 3 | 1 | 2 | 3 |
| Observed | 1 | **373** | 17 | 1 | **269** | 103 | 19 |
| | 2 | 47 | **783** | 7 | 98 | **600** | 139 |
| | 3 | 5 | 39 | **426** | 21 | 64 | **385** |

In the case of RST modeling, the percentage of correct predictions is higher and fluctuates around 90% (see Fig. 11b).

Figure 13 shows the results of prediction by decision tree rules for the remaining discretization methods, i.e. for the division into 5 and 7 classes by methods of equal width, equal depth and *k*-means. Compared to the division into three classes, there is a significant variation in the number of correct predictions. The maximum value of correct predictions occurred for class 4, divided into 5 classes using the equal width method, and amounts to 78%. On the other hand, the lowest values of correct predictions, oscillating around 20%, were observed for all the right methods with the division into 7 classes and for the equal width method with the division into 5 classes. Such low values were not observed when dividing by methods of equal depth and *k*-means into 5 classes. For these methods, the lowest values exceed 45%, and the mean value of correct predictions exceeds 50%.

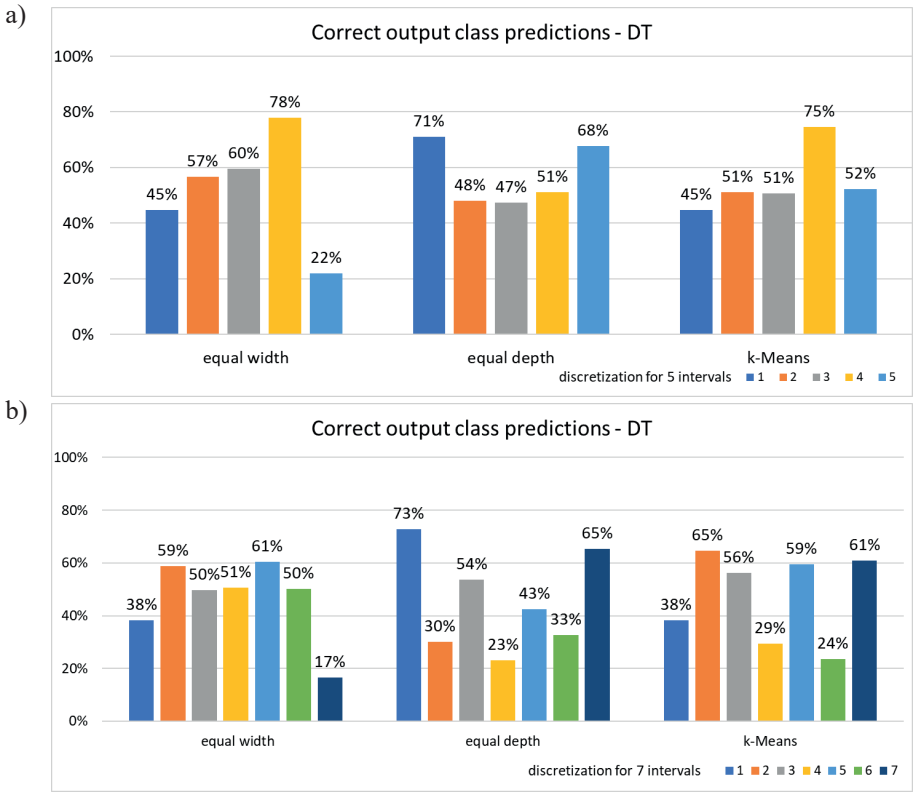



**Fig. 13.** Correct class prediction with decision tree rules (DT) for dividing the output variable into 5 (a) or 7 (b) classes using three discretization methods

Figure 14 shows the predictions of DT and RST models for the set of 11 *k*-means. There is a visible advantage of the RST model over the DT model in terms of the correctness of prediction of the output variable. The lowest value of predicting the output variable using the RST model was obtained for class 9, and it was 71%, while for the DT model it was 0% for class 10. This means that using the DT model no prediction indicating class 10 was obtained. This transpired despite the fact that there were 57 observations for class 10 in the discretized dataset. In the case of predicting the remaining classes, the results were twice as bad for

the DT model. The exception is class 11, where the predictions are at a comparable level and amount to 88% for the RST method and 80% for the DT method.

The summary results for all sets of the correctness of predicting the output variable for all sets are presented in Table 11. The clearly better model is one based on the rough sets theory. It is characterized by a smaller mean prediction error for each class in each set compared to the DT model. For the DT model, no correlation was observed between the correctness of the prediction and the degree of inconsistency. For RST rules, a trend was detected that the more inconsistencies in the data set, the worse the model's prediction. In addition, note that the RST model based on rule evaluation measures does not omit any class. Such a phenomenon was observed for the DT model.

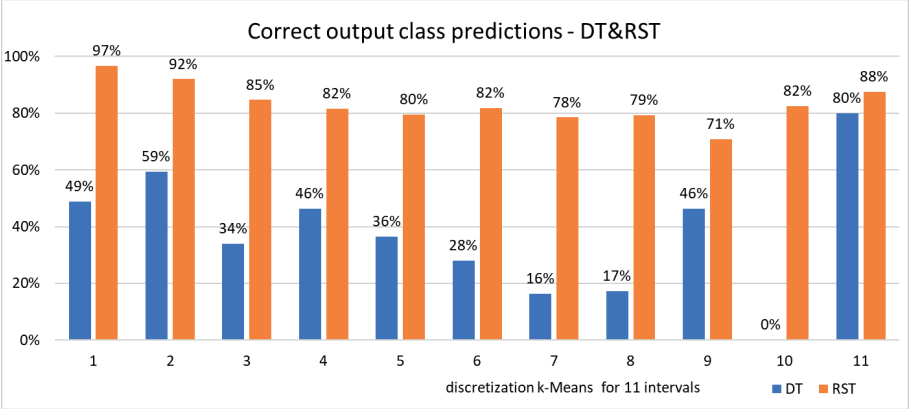



**Fig. 14.** Correct class prediction with RST and DT rules for 11 *k*-means

**Table 11.** Correct predictions of rule models for data sets for individual classes of the output variable [%]

| Model | Output classes | Name of the data set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | equal width | | | equal depth | | | *k*-means | | | |
| | | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 | 11 |
| DT | 1 | 69 | 45 | 38 | 72 | 71 | 73 | 67 | 45 | 38 | 49 |
| | 2 | 72 | 57 | 59 | 67 | 48 | 30 | 66 | 51 | 65 | 59 |
| | 3 | 82 | 60 | 50 | 73 | 47 | 54 | 84 | 51 | 56 | 34 |
| | 4 | – | 78 | 51 | – | 51 | 23 | – | 75 | 29 | 46 |
| | 5 | – | 22 | 61 | – | 68 | 43 | – | 52 | 59 | 36 |
| | 6 | – | – | 50 | – | – | 33 | – | – | 24 | 28 |
| | 7 | – | – | 17 | – | – | 65 | – | – | 61 | 16 |
| | 8 | – | – | – | – | – | – | – | – | – | 17 |
| | 9 | – | – | – | – | – | – | – | – | – | 46 |
| | 10 | – | – | – | – | – | – | – | – | – | 0 |
| | 11 | – | – | – | – | – | – | – | – | – | 80 |
| RST | 1 | 95 | 88 | 95 | 97 | 97 | 98 | 96 | 95 | 93 | 97 |
| | 2 | 94 | 91 | 91 | 88 | 89 | 85 | 92 | 89 | 87 | 92 |
| | 3 | 91 | 89 | 85 | 90 | 82 | 85 | 91 | 84 | 87 | 85 |
| | 4 | – | 89 | 87 | – | 87 | 78 | – | 89 | 82 | 82 |
| | 5 | – | 85 | 88 | – | 90 | 85 | – | 86 | 85 | 80 |
| | 6 | – | – | 85 | – | – | 79 | – | – | 82 | 82 |
| | 7 | – | – | 83 | – | – | 86 | – | – | 84 | 78 |
| | 8 | – | – | – | – | – | – | – | – | – | 79 |
| | 9 | – | – | – | – | – | – | – | – | – | 71 |
| | 10 | – | – | – | – | – | – | – | – | – | 82 |
| | 11 | – | – | – | – | – | – | – | – | – | 88 |

## 6. Summary

In the conducted research, a data set containing industrial-type data was used, having all the features of such a set, including a significant level of uncertainty resulting from contradictory data. The conducted research showed that, despite the presence of contradictory data in the data set, rule-based modeling is possible, however, the obtained models may be characterized by a high number of erroneous predictions. This is shown in Figure 15, and the summary of numerical data is presented in Table 12. The presented summary results show that:

- no statistically significant difference was observed in the case of contradictory sets in the impact of the discretization method (equal width, equal depth, *k*-means) on the quality of the rules generated;
- the RST method turned out to be significantly better than the rules obtained by the decision tree method;
- in the RST method, a correlation was observed between the degree of class contradiction and the quality of the rules.

In the case of the analyzed database, the higher quality of RST rule predictions results from:

- generating shorter rules – the shortest RST rules contained two attributes, while the decision tree rules had three;
- a higher number of correctly predicted classes of the output function – for all types of discretization the correct predictions exceeded 70% for RST, while for decision trees the average predictions oscillated around 50%;
- ability to predict all classes of output functions – RST rules provided for all classes of output functions regardless of the type of discretization, while in the case of discretization of 11 classes using the *k*-means method, decision trees omitted one class.

In addition, decision trees generate rules that are either unsupported or inconsistent with the data, as also shown in (Perzyk & Soroczyński, 2019). The paper shows that contradictory data in rule-based modeling does not lead to a weak or incorrectly predicting model, but to the lack of a model for specific cases (meaning no predictions at all, for specific ranges of output variability).

It should be noted that a significant impact on the quality of the predicted rules was observed in both modeling methods, both by the discretization method and the number of classes. It is therefore important to work on defining the minimum number of decision classes to keep the rules short, making it easier to analyze the process.

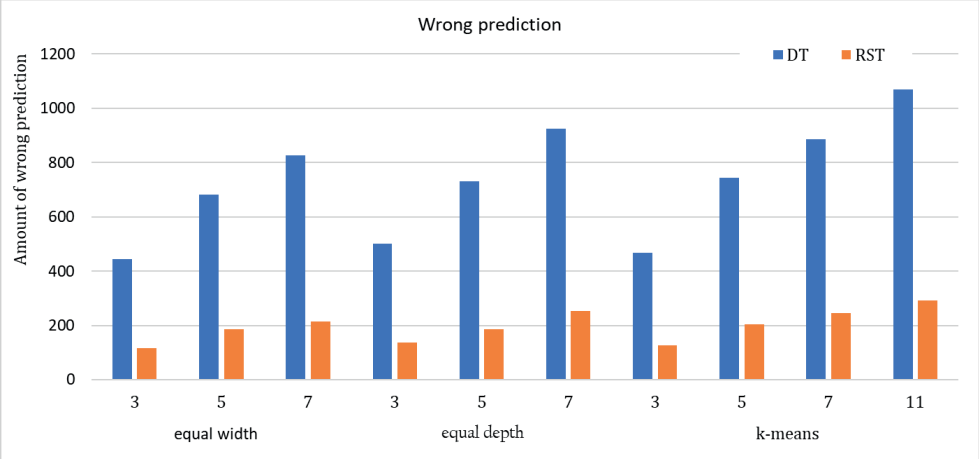



**Fig. 15.** Number of cases with erroneous predictions obtained by both rule generation methods

**Table 12.** Number of observations with error predictions for all output function discretization methods

| Modeling method | | Name of the dataset – dicretization method | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | equal width | | | equal depth | | | *k*-means | | | |
| | | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 | 11 |
| DT | medium error [%] | 26 | 40 | 49 | 29 | 43 | 54 | 28 | 44 | 52 | 63 |
| | no. of erroneous | 444 | 682 | 826 | 500 | 731 | 924 | 468 | 743 | 885 | 1070 |
| RST | medium error [%] | 7 | 11 | 13 | 8 | 11 | 15 | 7 | 12 | 14 | 17 |
| | no. of erroneous | 116 | 186 | 215 | 138 | 187 | 254 | 127 | 203 | 245 | 293 |

## 7. Conclusions

The analysis of the DT and RST models built using a dataset containing a significant percentage of uncertain data showed that in the case under study: RST generates shorter rules than DT and DT bypasses the classes in the set and may not create a rule even if the class has a significant percentage in the database.

In order to generalize the presented conclusions, it is necessary to conduct similar analyzes using generated synthetic databases and industrial databases containing a significant percentage of uncertain data.

## References


Barbosa, P.A., Costa, É.S., & Guesser, W.L., & Machado, Á.R. (2015). Comparative study of the machinability of austempered and pearlitic ductile irons in drilling process. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *37*(1), 115–122. https://doi.org/10.1007/s40430-014-0161-z.

Chawla, V., Batra, U., Puri, D., & Chawla, A. (2008). To study the effect of austempering temperature on fracture behaviour of Ni-Mo Austempered Ductile Iron (ADI). *Journal of Minerals & Materials Characterization & Engineering*, *7*(4), 307–316. https://doi.org/10.4236/jmmce.2008.74024.

Colin García, E.; Cruz Ramírez, A., Reyes Castellanos, G., Chávez Alcalá, J.F., Téllez Ramírez, J., & Magaña Hernández, A. (2021). Heat treatment evaluation for the camshafts production of ADI low alloyed with Vanadium. *Metals*, *11*(7), 1036. https://doi.org/10.3390/met11071036.

Dal Corobbo, M., Arias, S. (2009). *Evaluation of impact properties of austempered ductile iron*. [Master's thesis, Department of Materials and Manufacturing Technology, Chalmers University of Technology]. Printed by Chalmers Reproservice Göteborg, Sweden.

Grzegorzewski, P., & Kochański, A. (2019). Data preprocessing in industrial manufacturing. In P. Grzegorzewski, A. Kochański, J. Kacprzyk (Eds.), *Soft Modeling in Industrial Manufacturing* (pp. 75–88). Springer Cham. https://doi.org/10.1007/978-3-030-03201-2_3.

Heydarzadeh Sohi, M., Nili Ahmadabadi, M., & Bahrami Vahdat, A. (2004). The role of austempering parameters on the structure and mechanical properties on heavy section ADI. *Journal of Materials Processing Technology*, *153–154*, 203–208. https://doi.org/10.1016/J.JMATPROTEC.2004.04.308.

Kochański, A., Perzyk, M., & Kłębczyk, M. (2012). Knowledge in imperfect data. In C. Ramirez (Ed.), *Advances in Knowledge Representation* (pp. 181–210). InTech. https://doi.org/10.5772/37714.

Kochański, A., Soroczyński, A., & Kozłowski, J. (2013). Applying rough set theory for the modeling of austempered ductile iron properties. *Archives of Foundry Engineering*, *13*(Special Issue 2), 70–73.

Kochański, A., Grzegorzewski, P., Soroczyński, A., & Olwert, A. (2014). Modeling of austempered ductile iron using discrete signals. *Computer Methods in Materials Science*, *14*(3), 190–196.

Kochański, A., Krzyńska, A., Chmielewski, T., & Stoliński, A. (2015). Comparison of austempered ductile iron and manganese steel wearability. *Archives of Foundry Engineering*, *15*(Special Issue 1), 51–54.

Materials Algorithms Project Data Library (n.d.). *Data Library MAP_DATA_ADI_RETAINED-AUSTENITE*. http://www.phase-trans.msm.cam.ac.uk/map/data/materials/adiret.html.

Nobuki, T., Hatate, M., & Shiota, T. (2010). Notch effects on impact and bending characteristics of spheroidal graphite and compacted vermicular graphite cast irons with various matrices. *Key Engineering Materials*, *457*, 392–397. https://doi.org/10.4028/www.scientific.net/KEM.457.392.

Olofsson, J., Larsson, D., & Svensson, I.L. (2011). Effect of austempering on plastic behavior on some austempered ductile iron alloys. *Metallurgical and Materials Transactions A*, *42*(13), 3999–4007. https://doi.org/10.1007/s11661-011-0796-7.

Perzyk, M., & Kochański, A.W. (2001). Prediction of ductile cast iron quality by artificial neural networks. *Journal of Materials Processing Technology*, *109*(3), 305–307. https://doi.org/10.1016/S0924-0136(00)00822-0.

Perzyk, M., & Soroczynski, A. (2008). Comparison of selected tools for generation of knowledge for foundry production. *Archives of Foundry Engineering*, *8*(4), 163–166.

Perzyk, M., & Soroczyński, A. (2019). Assessment of selected tools used for knowledge extraction in industrial manufacturing. In P. Grzegorzewski, A. Kochański, J. Kacprzyk, J. (Eds.), *Soft Modeling in Industrial Manufacturing* (pp. 27–41). Springer Cham. https://doi.org/10.1007/978-3-030-03201-2_5

Perzyk, M., Biernacki, R., Kochanski, A., Kozlowski, J., & Soroczynski, A. (2011). Applications of data mining to diagnosis and control of manufacturing processes. In K. Funatsu (Eds.). *Knowledge-oriented applications in data mining* (pp. 147–166). InTech. https://doi.org/10.5772/13282.

Perzyk, M., Kochański, A., Kozłowski, J., & Myszka, D. (2015). Control of ductile iron austempering process by advanced data driven modeling. In *71st World Foundry Congress: Advanced Sustainable Foundry (WFC 2014). 19–21 May 2014, Bilbao, Spain*. https://www.scopus.com/record/display.uri?eid=2-s2.0-84928901998&origin=resultslist&sort=plf-f.

Rodríguez-Rosales, N.A., Montes-González, F.A., Gómez-Casas, O., Gómez-Casas, J., Galindo-Valdés, J.S., Ortiz-Cuellar, J.C., Martínez-Villafañe, J.F., García-Navarro, D., & Muñiz-Valdez, C.R. (2022). Statistical data-driven model for hardness prediction in austempered ductile irons. *Metals*, *12*(4), 676. https://doi.org/10.3390/met12040676.

Stefanowski, J., & Vanderpooten, D. (2001). Induction of decision rules in classification and discovery-oriented perspectives. *International Journal of Intelligent Systems*, *16*(1), 13–27. https://doi.org/10.1002/1098-111X(200101)16:1<13::AID-INT3>3.0.CO;2-M.

Szykowny, T., Ciechacki, K., Skibicki, A., & Sadowski, J. (2010). The effect of microstructure of low-alloy spheroidal cast iron on impact strength. *Archives of Foundry Engineering*, *10*(Special Issue 1), 75–80.

Wieczorek, A.N., Wójcicki, M., Drwięga, A., Tuszyński, W., Nuckowski, P.M., & Nędza, J. (2022). Abrasive wear of mining chain drums made of austempered ductile iron in different operating modes. *Materials*, *15*(8), 2709. https://doi.org/10.3390/ma15082709.

Wilk-Kołodziejczyk, D., Regulski, K., Giętka, T., Gumienny, G., Jaśkowiec, K., & Kluska-Nawarecka, S. (2018). The selection of heat treatment parameters to obtain austempered ductile iron with the required impact strength. *Journal of Materials Engineering and Performance*, *27*(11), 5865–5878. https://doi.org/10.1007/s11665-018-3714-y.

Wohlfahrt, M., Oberwinkler, C., Tunzini, S., Rauscher, A., Prida Caballero, R., de la, & Eichlseder, W. (2010). The role of sampling position on fatique of austempered ductile iron. *Procedia Engineering*, *2*(1), 1337–1341. https://doi.org/10.1016/j.proeng.2010.03.145.

Yescas, M.A., Bhadeshia, H.K.D.H., & MacKay, D.J. (2001). Estimation of the amount of retained austenite in austempered ductile irons using neural networks. *Materials Science and Engineering: A*, *311*(1–2), 162–173. https://doi.org/10.1016/S0921-5093(01)00913-3.