www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

Arch. Min. Sci. **65** (2020), 1, 117-127

Electronic version (in color) of this paper is available: http://mining.archives.pl

DOI 10.24425/ams.2020.132710

JANUSZ RUSEK[1*]

THE POINT NUISANCE METHOD AS A DECISION-SUPPORT SYSTEM BASED ON BAYESIAN INFERENCE APPROACH

The article attempts to transfer information from the *Point Nuisance Method* (PNM) used in Poland in the issue of protection of buildings in mining areas, to the system of inference based on *Bayesian* formalism. For this purpose, all possible combinations occurring in PNM were selected. The number of numerically generated patterns was 6,718,464 cases. Then, based on *Python* package *Scikit-Learn*, a classification model was created in the form of the *Naïve Bayes Classifier* (NBC). The effectiveness of three methods used to build this type of decision-support system was analysed, from which the *Categorical Multinomial Naive Bayes* (CMNB) approach was finally selected. With the created classifier, its properties were verified in terms of quality of classify and generalization. For this purpose a general approach was used, analysing the level of accuracy of the model in relation to training and teaching data, and detailed, based on the analysis of the confusion matrix. Additionally, the operation of the created classifier was simulated to determine the optimal *Laplace* smoothing parameter α. The article ends with conclusions from the carried out calculations, in which an attempt was made to answer the question concerning potential reasons for incorrect classification of the created CMNB model. The discussion ends with a reference to the planned research, in which, among other things, the use of more complex *Bayesian belief networks* (BBN) is planned.

Keywords: Naïve Bayes, Resistance of Buildings, mining area, reliability, Bayes inference framework, surface deformations

1. Introduction

In the case of forecasting negative effects from mining exploitation on building structures constituting the development of the area, their resistance to the effects of continuous deformation

[1] AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY, FACULTY OF MINING SURVEYING AND ENVIRONMENTAL ENGINEERING, DEPARTMENT OF ENGINEERING SURVEYING AND CIVIL ENGINEERING, AL. MICKIEWICZA 30, 30-059 KRAKOW, POLAND

* Corresponding author: rusek@agh.edu.pl

is assessed. Both the resistance of the object and the predicted intensity of mining impacts are random [1-6]. Therefore, in order to give the scope of negative impacts from mining exploitation, it is necessary to determine the risk of potential mining damage. According to the theory of reliability of structures, the risk is presented in probabilistic notation [7,8]. On the basis of knowledge of probability distributions describing the resistance category of a given object and the predicted impacts from mining exploitation, it is possible to determine the level of structural reliability [1-3].

The so-called *Building Resistance Point Method* (BRPM) used so far in Poland, also known as the *Point Nuisance Method* (PNM), gives sharp limits of permissible mining impacts, which the analyzed structure can carry without safety hazard [2,9,10]. Although it is an effective tool in the case of resistance assessment for a large number of buildings constituting the development of a given mining area, from the point of view of structural reliability, it does not allow for a full comparative analysis with the predicted mining impacts.

The introduction of probabilistic notation in the methodology of assessment of resistance of buildings would allow to determine the quantitative level of threat of a large group of buildings on the negative influence of forecasted continuous ground deformations caused by mining exploitation. Attempts to describe the influence of random dispersion of continuous deformation indices taking into account their influence on surface development were made both with probabilistic methods in [11,12] and with methods based on fuzzy set theory [13]. However the problem of including dispersion in the assessment of resistance for a large number of buildings remains unsolved [1,3].

In this paper an attempt is made to present the PNM used in Poland in the form of *Naïve Bayes classifier* (NBC). This approach does not change the relations occurring in the original method, but it makes them more detailed by the probability value for a given indication. Generally, the original information contained in the point method was transferred to the *Bayesian inference framework* [14,15]. Such implementation represents the state-of-art approach to the field of protection of mining areas, which can be further improved on the basis of new observation data.

An analogous research methodology was used in the paper [16], in which an attempt was made to assess the technical condition of buildings located in the mining area of *Legnica-Glogow Copper District* (LGCD) using the *Bayesian* classifier. A similar approach, using *probabilistic neural networks* (PNNs), was also applied in the assessment of the risk of mining damage in the event of high-energy mining tremors [17] and in the problem of the risk of damage to the underground infrastructure network resulting from the deformation of the continuous area [18]. Research is currently underway on the application of the *Bayesian Belief Network* (BBN) [14] to these issues. The key here is the defined damage index for buildings, which is a measure of possible mining damage [19]. In all cases, the main purpose of the analyses was to give the probability of an event consisting in the occurrence of damage of a given intensity or mining damage. In this paper, the use of *Bayes's* classifier will allow to give an estimated probability of occurrence of a given category of building's resistance, depending on the structural features of a given building and its technical condition.

In the next part of the work, according to [2], *nuisance categories* (NC) will be used as a measure of building resistance.

2. The main principles of the nuisance point method

The *Point Nuisance Method* (PNM) considers information on structural and material solutions, geometry and technical condition [2]. This information is obtained mainly on the basis of in-situ inventory and analysis of available documentation. As a result, for a given building, the number of points is determined, which is the basis for classifying the object into the so-called *nuisance category* (NC). Thus, apart from the indication of a given nuisance category, there is no other measure to make the indication more precise, as the points do not provide any interpretable information for the construction engineer. The decision about belonging to a given category is sharp and does not allow to take into account the uncertainty. This situation is particularly evident in the case of buildings whose number of points reaches the limits between the categories.

In order to be able to analyse situations where there is uncertainty, as to whether a given building belongs to the relevant category, it was decided to introduce an additional measure that can be interpreted in an engineering sense. Such situations occur when a building reaches a value expressed by the number of points at the intersection of two categories. In view of the work of [2], which postulated the possibility to interpret the PNM on the basis of the principles of the reliability analysis of the structure, the value of probability was taken as a more precise measure of classification. On the other hand, given that in the PNM approach all variables are independent of each other, it was decided to adopt *Naive Bayesian Classification* approach (NBC).

3. Research methodology

The basic assumption of the applied method was the best possible transformation of the information contained in the PNM to the *Bayesian* classifier (NBC) structure. For this purpose, a complete set of all possible combinations for PNM and the corresponding number of points was generated. Each variable has been divided into a number of categories corresponding to the adopted version of PNM [2]. Finally, a dataset of 12 input variables and one output variable with a total number of 6,718,464 cases was generated. It was then divided into a training and test set (in proportions: 0,8 : 0,2), which were used to learning and verification the quality of the created classifier in relation to the correctness of classification and generalization of acquired knowledge. Then, for the set of combinations selected in this way, the number of corresponding points from the PNM method was calculated. This was the basis for determining the *nuisance category* (NC) for each case (**K0**, **K2**, **K3**, **K4**, **K5**). Thus, the set of input variables was extended with a categorized output variable, which was decision variable for analysed problem. The scale of categories number for particular input variables of the PNM is presented in Table 1.

4. Methodological basis of the naive bayesian classification

The *Naive Bayes Classifier* (NBC) is the simplest method of inference, based on the probabilistic notation [14,15,20]. In the issues related to data mining and machine learning, the NBC method is considered independently of the Bayesian belief networks. However, it can be considered to be a special, and at the same time the simplest, form of *Bayesian Belief Network* (BBN) [21]. This results from the assumption that all input variables from the set $X = \{x_i\}_{i=1}^{n}$ included in

TABLE 1

Number of categories for all variables of the PNM [2]

LENGTH OF BUILDING	SHAPE OF THE BUILDING	BUILDING FOUNDATION	FOUNDATION SOIL	BUILDING STRUCTURE		EXISTING BUILDING STRUCTURE REINFORCEMENTS	TECHNICAL STATE	
6	6	3	3	Foundations	4	4	Natural Wear	3
				Basement Walls	3		Demages	4
				Ceiling Of The Lowest Floor	4			
				Lintels	3			
				Other Structures Elements	3			

the description of a specific decision problem, are conditionally independent of each other with respect to the output decision variable y. The decision variable y is expressed separately as a set of all states (decision classes) that are assigned to it $y=\{c_k\}_{k=1}^{m}$. On the other hand, the individual variables x_i are represented by a set of assigned states (categories) in which they can be observed $x_i=\{a_i^l\}_{l=1}^{q_i}$. The consequence of such a situation is the possibility of expressing the NBC structure with the following equation (1) [20], whose graphic interpretation has been presented in Figure 1.

$$P(y|x_1,\ldots,x_n)=\frac{P(y)\prod_{i=1}^{n}P(x_i|y)}{P(x_1,\ldots,x_n)} \tag{1}$$

In the context of the operation of a classifier, it is assumed that:

$$P(y|x_1,\ldots,x_n)\propto P(y)\prod_{i=1}^{n}P(x_i|y) \tag{2}$$

Such a statement allows to finally indicate the decision class c_k for a given set of input variables $\{x_i\}_{i=1}^{n}$. Generally, this procedure is based on a ranking that can be described by a relationship (3).

$$c_k=\arg\max_{y}P(y)\prod_{i=1}^{n}P(x_i|y) \tag{3}$$

For this reason, there is the possibility of a broader insight into the classification result generated by the model. Ranking is based on the levels of conditional probabilities for each category of decision variable. Thus, the model includes all levels of probability of obtaining all possible states of the decision variable.

In the case of fitting the NBC model from discrete observed data, the problem boils down to finding optimal parameters $\theta_y = \{\theta_{y1},...,\theta_{yn}\}$ describing conditional probabilities between the input variables and the decision variable $\theta_y = \{\theta_{y1},...,\theta_{yn}\}$; $\theta_{yi} = P(x_i|y)$. For this purpose, the MLE (Maximum Likelihood Estimation) or more generally MAP (Maximum a Posteriori Estimation) procedure is used [14]. For discrete data, a set of parameters θ_y is sought, corresponding to the adoption of multinomial conditional probability distributions. In general, the *Dirichlet* or multinomial *Bernoulli* distribution is very often used [21]. However, in the course of research, using the *Sickit-Learn* package, an approach based on the adoption of categorical distribution for each input variable was applied [20]. The choice of this method was made after the results of preliminary studies, in which NBC classifiers were also built using the above mentioned approaches. But, in case of *Dirichlet* and multinomial *Bernoulli* approach much worse classification results obtained. From that reason they were abandoned in further studies.

As a consequence of the categorical approach used, the probability of category $a_i^l = t$ in input variable x_i given class c_k is estimated as [20]:

$$P\left(x_i = a_i^l = \hat{a} \,|y = c_k ; \alpha\right) = \frac{N_{\hat{a}ic_k} + \alpha}{N_{c_k} + \alpha n_i} \tag{4}$$

Where:

$Z = \{1,...,m\}$ — index set of the samples,

m — as the number of samples,

$N_{\hat{a}ic_k} = \left|\{z \in Z | x_{iz} = a_i^l = \hat{a},\ y_z = c_k\}\right|$ — is the number of times category $\hat{a}$ appears in the samples x_i, which belong to class c_k of decision output variable y,

$N_{c_k} = \left|\{z \in Z | y_z = c_k\}\right|$ — is the number of samples with class c_k of decision output variable y,

α — is a smoothing parameter resulting from adopted Laplace method of smoothing categorical data [14],

n_i — is the number of available categories states of variable x_i.

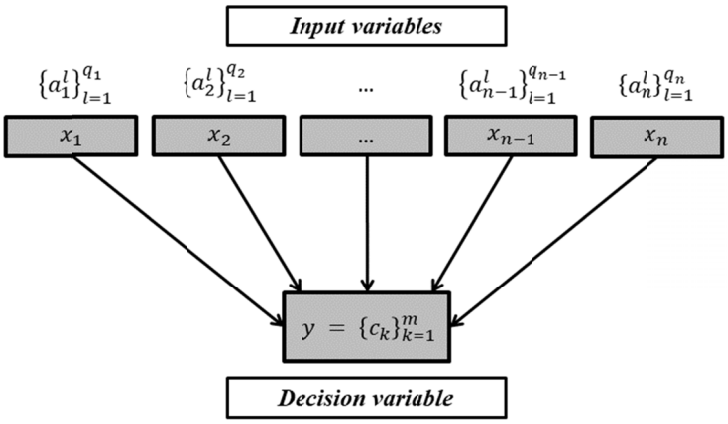

Fig. 1. Schematic diagram of the structure of the Naive Bayes Classifier (NBC).
Source: own study

5. Study results

Before the commencement of the numerical analysis, the generated database had been divided into a training set (5,374,770 cases) and a test set (1,343,693 cases).With the training and test sets separated, the model of the *Naive Bayes Classifier* was built by performing calculations in the *Python* programming language with use a *Scikit-Learn* package [20]. *The Categorical Multinomial Naive Bayes* (CMNB) approach was used. In addition to CMNB other methods dedicated to creating the *Bayes Naive* classifier have also been tested [20]. Unfortunately, for both *Multinomial Naive Bayes* (MNB) and *Bernoulli Naive Bayes* (BNB), the level of correctly classified patterns has been less than 50% both for training and testing sets. Therefore, it has been decided to abandon their use in further research.

Using the CMNB approach, the estimated probability levels depend on the *Laplace* smoothing parameter α (see formula 4). In order to test how the choice of the value of this parameter affects the quality of the classifier, calculations were made in two variants. In the first variant an α value of 0,1 was used. In the second variant an α value of 1 was used.

As a result of the analyses carried out, it was found that a lower value of the α parameter, and thus a weaker smoothing of categorized data, results in improved classification results. This is particularly evident when analysing the confusion matrix. In the case of the parameter $\alpha = 1$, misclassified cases for a given category cover only those categories that are in the immediate neighbourhood. While the adoption of the parameter $\alpha = 1$ contributes to spreading the classification errors to the further distant categories. This effect is shown in Table 2, where the confusion matrices are set up separately for training and test sets. The values of the elements of the confusion matrix have been given in [%] and referred to the total number of cases of a given category for prediction.

TABLE 2

Indices of accurate classifications for the model (in percentage)

Training set (5,374,770 cases)											
		$\alpha = 0,1$					**$\alpha = 1,0$**				
		True category from PNM									
		K0	K1	K2	K3	K4	K0	K1	K2	K3	K4
Predicted category NBC	K0	98,23	1,79	0,00	0,00	0,00	98,23	1,77	0,00	0,00	0,00
	K1	45,30	52,06	2,64	0,00	0,00	45,30	52,06	2,64	0,00	0,00
	K2	0,00	17,92	81,21	0,86	0,00	0,00	17,91	81,23	0,86	0,00
	K3	0,00	0,00	28,83	71,14	0,01	0,00	0,00	28,72	71,27	0,01
	K4	0,00	0,00	0,00	56,68	39,23	0,00	4,73	1,20	54,80	39,26
Testing set (1,343,693 cases)											
		$\alpha = 0,1$					**$\alpha = 1,0$**				
		True category from PNM									
		K0	K1	K2	K3	K4	K0	K1	K2	K3	K4
Predicted category NBC	K 0	98,19	1,81	0,00	0,00	0,00	98,23	1,79	0,00	0,00	0,00
	K 1	45,22	52,15	2,64	0,00	0,00	45,30	52,15	2,63	0,00	0,00
	K 2	0,00	17,95	81,16	0,89	0,00	0,00	17,94	81,17	0,89	0,00
	K 3	0,00	0,00	28,54	71,44	0,01	0,00	0,00	28,42	71,56	0,01
	K 4	0,00	0,00	0,00	56,95	43,05	0,00	5,35	1,34	52,13	41,19

The effect of data smoothing is also visible during simulating of model. By marginalizing the input variables, the probability levels of occurrence of a given category was predicted. The results are presented in the domain of points according to PNM range values. The simulation includes all the cases (6,718,463) numerically generated at the beginning of research without division into training and test set. The results of such simulations, given separately for different values of α parameter, are presented in Figure 2.

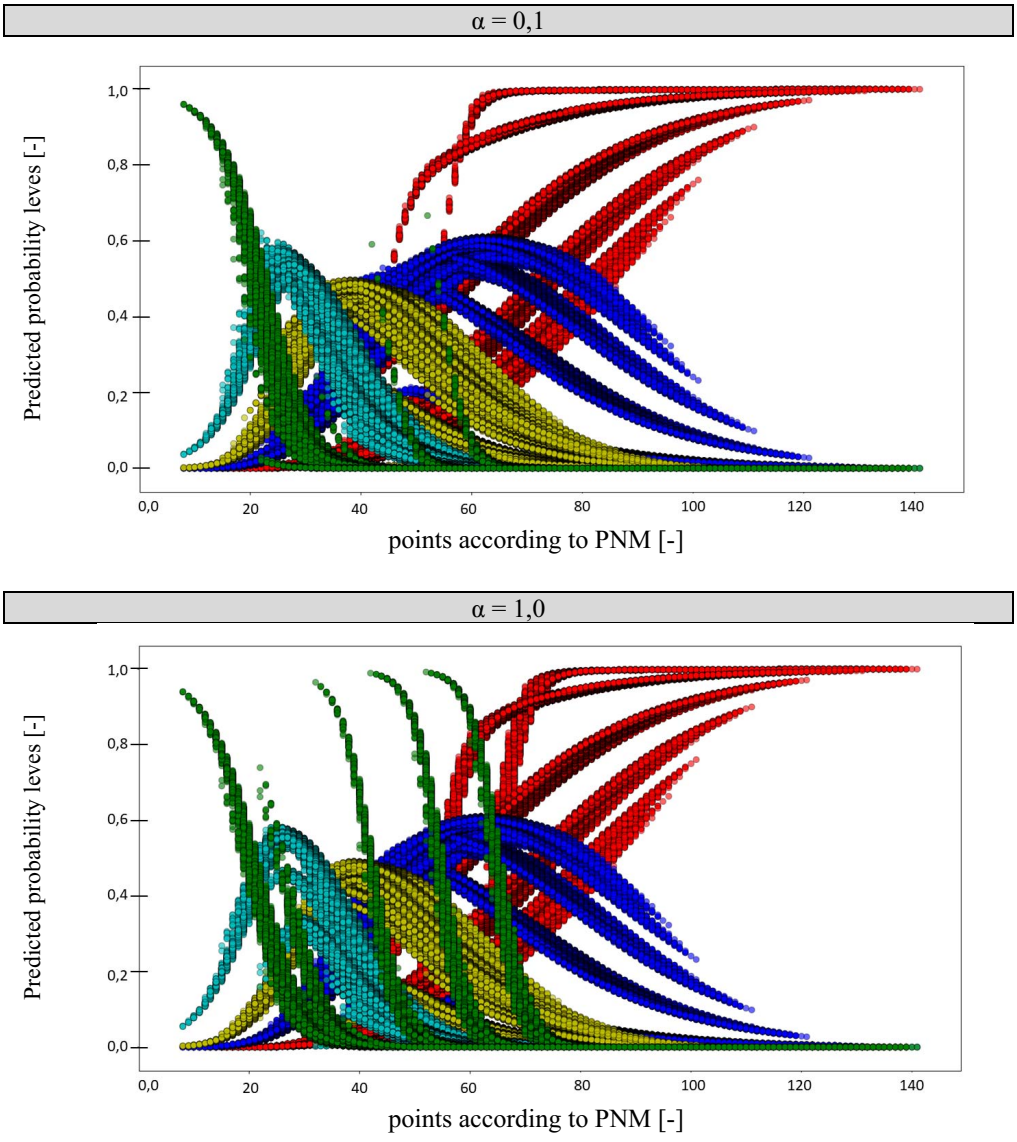

Fig. 2. Simulated probability values for the individual nuisance categories for the smooth parameter $\alpha = 0,1$ and $\alpha = 1,0$ (*colour labels of predicted categories:* ***K4*** *– green,* ***K3*** *– cyan,* ***K2*** *– yellow,* ***K1*** *– blue,* ***K0*** *– red*)

As can be seen on Fig. 2, the greatest influence of smoothing is visible for category **K4**. For the parameter $\alpha = 1$, the values of relatively high probabilities cover categories **K3** and **K2**, and even category **K1**. Taking into account, according to (3), that the values of these probabilities are the basis for determining the final classification result, it may cause significant distortions in determining the actual category for a given building (even different by 3 categories).

On the basis of these conclusions, it has been decided that, in the framework of these studies, the target model will be a classifier which values of conditional probabilities will be estimate for the value of parameter $\alpha = 0{,}1$.

Table 3 compares the results of studies on the accuracy of the model classification for the training set and the test set for the CMNB approach used. The obtained classification model (CMNB) has a high degree of accuracy in relation to the learning data (80,45% of correctly classified models – Tab. 3). Almost identical level of accuracy was obtained for the test set (80,44% of correctly classified models – Tab. 3). This demonstrates good fitting quality of the model, as well as no effect of overfitting [15].

TABLE 3

Indices of accurate classifications for the model (in percentage)

The percentage of correctly classified cases for the training set (5,374,770 cases)	The percentage of correctly classified cases for the test set (1,343,693 cases)
80,45 %	80,44 %

Finally, assuming the model for the parameter $\alpha = 0{,}1$, based on the results of the simulations presented in Figure 3, an attempt was made to function estimate the distributions for each category. The distribution of data for individual categories obtained in the result of CMNB model simulation has been estimated with two functions. Categories: **K0** and **K4** were approximated by *logistic* functions (5). On the other hand categories: **K1**, **K2** and **K3** were approximated using *Gauss* functions (6). Table 4 lists the estimated parameters for each of the approximated functions. The results of the estimation are shown in Figure 5.

$$\frac{L}{1+e^{-k(x-x_0)}} \tag{5}$$

$$a \cdot e^{-\frac{(x-x_0)^2}{2\sigma^2}} \tag{6}$$

Where: a, x_0, σ, L, k — estimated parameters

It should be noted, however, that both the results of CMNB classifier prediction and estimated function for obtained distributions, they do not represent probability density distributions in classical terms. They should be understood in the context of Bayesian inference framework as belief levels.

However, with appropriate calibration in accordance with the requirements of the *probability theory* for functions representing probability distributions, such estimated functions may be included in the reliability analysis for buildings in mining areas. This requires further verification of the model based on in-situ data. Collecting actual data will allow to tune up the

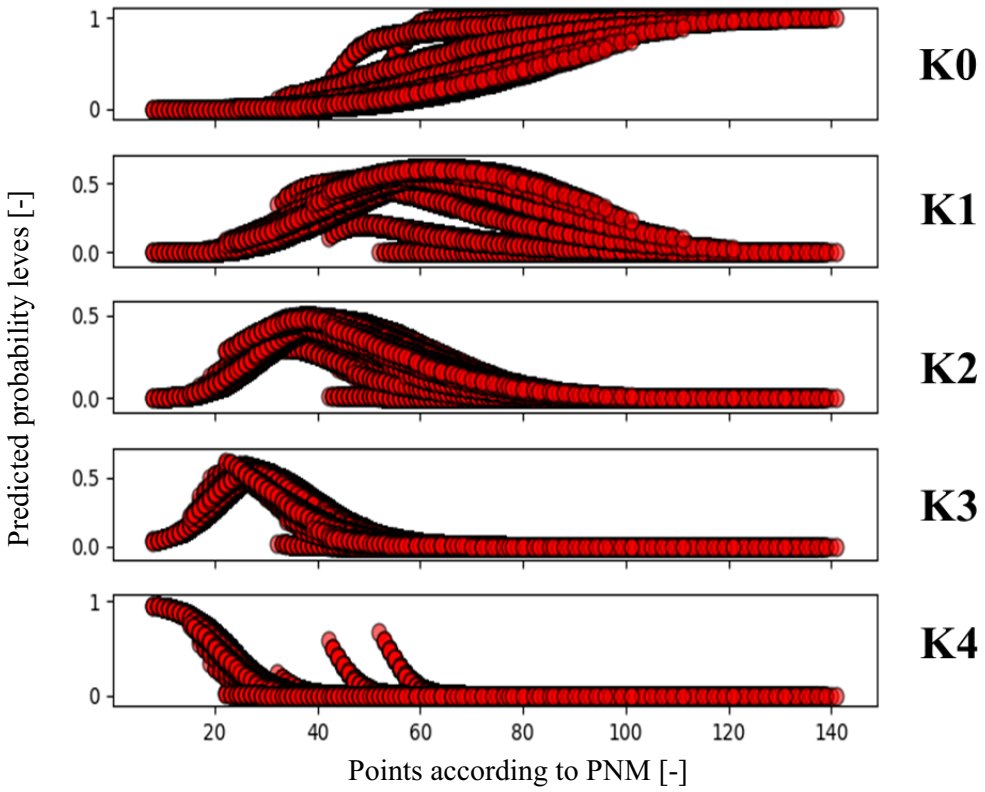

Fig. 4. Simulated probability values for the individual nuisance categories

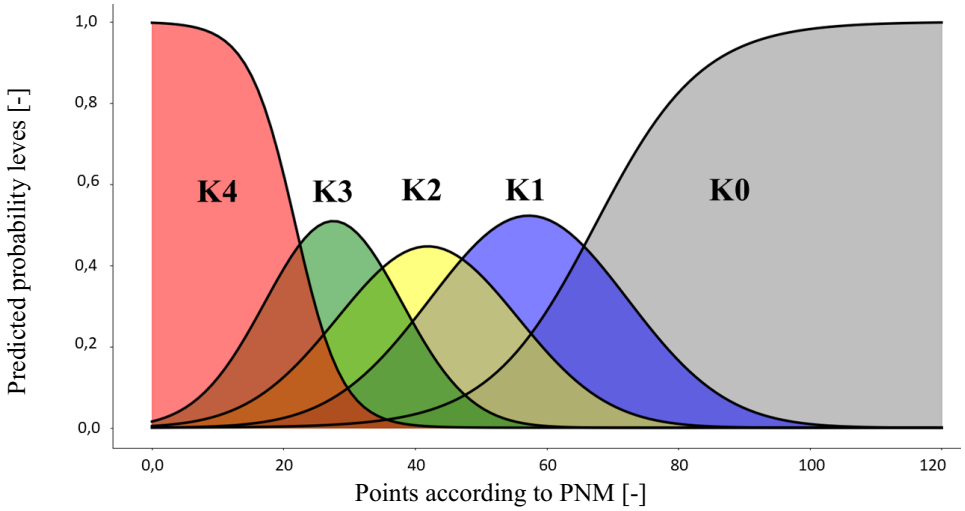

Fig. 5. Estimated distribution of belief levels for each category of nuisance according to PNM

model parameters and obtain more accurate classification results. This is in line with the *Bayesian* theory, according to which the created model CMNB represents the basic *prior knowledge* resulting from expert findings.

TABLE 4

Lists of estimated parameters of the approximated functions for each category

K0		**K1**		**K2**		**K3**		**K4**	
***Logistic* – (4)**		***Gauss* – (5)**		***Gauss* – (5)**		***Gauss* – (5)**		***Logistic* – (4)**	
L	1,0	a	0,52	a	0,45	a	0,51	L	0,99
x_0	67,36	x_0	57,26	x_0	41,92	x_0	27,63	x_0	21,89
k	0,12	σ	15,01	σ	13,70	σ	10,38	k	–0,28

6. Summary and conclusions

The results of the carried out research confirm the possibility of transferring the information contained in the *Point Nuisance Method* (PNM) to the structure of the decision-support system based on the *Bayesian* framework of reasoning. During the analysis it was found that in case of categorized data, the best method to create such a decision-support system is *Categorical Multinomial Naive Bayes* (CMBN) approach. Finally, a classifier with a very high level of nuisance accuracy to the training and testing set was obtained. It also presents how the probability distribution of indications for individual nuisance categories (**K0**, **K1**, **K2**, **K3**, **K4**) is shaped, which was referred to points from the PNM. The influence of *Laplace* α smoothing parameter, which is used in CMBN method for estimation of probability values, is shown. Finally, the value of this parameter was determined at the $\alpha = 0.1$.

The results obtained indicate, however, that there are still some distortions in terms of classify, which can be seen by analyzing the results presented in the form of a confusion matrix. The reason for this is that the distributions of the estimated probability values corresponding to the individual nuisance categories overlap. This leads to situations where, for a given case, not two adjacent but as many as four possible nuisance categories should be considered. This is, for example, the case for a range of points from 15 to 25, cf. Figure 5. The reason for this may be, first of all, the lack of specification of the information contained in the PNM. Taking into account that it covers 6,718,464 combinations, it should be concluded that it has not been sufficiently verified since its creation in the 1960s. Therefore, with the created CMBN model, the information contained therein can be considered as *prior knowledge*. Thus, in accordance with *Bayesian* model fitting, it is possible to further improve its classification quality on the basis of real *in-situ* data. The second reason for some inaccuracies may be the number of analyzed combinations, which may not correspond with reality. Therefore, the set of all combinations should be analysed in terms of their possible occurrence in construction practice.

A third reason may be the adoption of mutual independence between individual input variables. For this reason, the planned studies foresee the testing of more advanced *Bayesian belief networks* (BBN).

However, despite the above indications, the proposed methodology may lead in future to a solution to the problem of describing mining impacts on buildings based on the reliability of building structures.

Acknowledgments

The research was carried out under the R&D subsidy: 16.16.150.545

References

[1] J. Kwiatek, *Probabilistyczna ocena niezawodności obiektów budowlanych na terenach górniczych*. WUG Bezp. Pr. Ochr. Śr. w Gór. **6**, 14-16 (2006).

[2] J. Kwiatek, *Obiekty budowlane na terenach górniczych*. Główny Instytut Górnictwa, 2007.

[3] J. Kwiatek, *Ocena niezawodności budynków na terenach wstrząsów górniczych*. Gór. Geol. **2**, 5, 121-131 (2010).

[4] J. Kwiatek at others, *Ochrona obiektów budowlanych na terenach górniczych*. Wydaw. GIG Katowice, 1997.

[5] E. Popiołek, *Ochrona terenóww górniczych*. Wydawnictwa AGH, 2009.

[6] K. Tajduś, *New method for determining the elastic parameters of rock mass layers in the region of underground mining influence*. Int. J. Rock Mech. Min. Sci. **46**, 8, 1296-1305, (2009).

[7] A.S. Nowak, K.R. Collins, *Reliability of structures*. CRC Press, 2012.

[8] M. Lemaire, *Structural reliability*. John Wiley & Sons, 2013.

[9] M. Kawulok, *Szkody górnicze w budownictwie*. Wydawnictwa Instytutu Techniki Budowlanej, 2010.

[10] W. Mika, L. Chomacki, L. Słowik, *Zasady oceny odporności budynków na ciągłe deformacje terenu*. Przegląd Gór. **73** (2017).

[11] J. Ostrowski, *Deformacje powierzchni a zagrożenie uszkodzeniami budynków na terenach górniczych w ujęciu probabilistycznym*. AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, 2006.

[12] J. Ostrowski, A. Ćmiel, *The use of a logit model to predict the probability of damage to bullding structures in mining terrains*. Arch. Min. Sci. **53**, 2, 161-182 (2008).

[13] A. Malinowska, *Fuzzy logic-based approach to building damage risk assessment considering the social and economic value*. Gospod. Surowcami Miner. **24** (2008).

[14] K.P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[15] C.M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[16] J. Rusek, K. Firek, *Assessment of technical condition of prefabricated large-block building structures located in mining area using the naive bayes classifier*. Int. Multidiscip. Sci. GeoConference SGEM Surv. Geol. Min. Ecol. Manag. **2**, 109-116 (2016).

[17] M. Witkowski, J. Rusek, *Wykorzystanie probabilistycznych sieci neuronowych do wyznaczania ryzyka powstania szkód w budynkach poddanych wstrząsom górniczym*. Przegląd Gór. **73**, 1, 44-47 (2017).

[18] J. Rusek, *Support vector machines and probabilistic neural networks in the assessment of the risk of damage to water supply systems in mining areas*. 2016.

[19] K. Firek, *Proposal for classification of prefabricated panel building damage intensity rate in mining areas*. Arch. Min. Sci. **54**, 3, 467-479 (2009).

[20] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*. J. Mach. Learn. Res. **12**, 2825-2830 (2011).

[21] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques. MIT press, 2009.