STUDIA Z AUTOMATYKI I INFORMATYKI
TOM 42 – 2017

Anna Kasperczuk[*], Agnieszka Dardzińska[†]

LOGISTIC REGRESSION METHODS APPLICATION IN MEDICAL INFORMATION SYSTEMS

Keywords: cancer, medical database, classification, logistic regression, ROC

1. Introduction

The problems of classifying diseases diagnoses are often encountered in medicine. Making a correct diagnosis requires knowledge and many years of experience. In difficult clinical situations, when invasive testing is contemplated, it may be useful to use predictive models of the fastest statistical methods. Correctly sampled models are sentences from useful knowledge, discovery of complex relationships and explanations of observed trends, which may be the basis for differentiation. For a data interpretation and modelling, a number of tests are used to adapt their choice to the solve medical problems [1, 13].

Breast cancer is the leading type of women cancer. It accounts for 25% of all cases. During the last twenty years, it resulted in almost 1.7 million cases and 520,000 deaths [13]. It is more common in developed countries and is almost 130 times more common with women than men. Therefore, it becomes extremely important to look for methods that can provide support for doctors in the diagnosis and early detection of disease. In this case, statistic methods seems to be helpful.

We show how we can use statistical methods to build a system, which helps to build decision-making system [8]. In this paper, we conduct our tests on medical database. For the purpose of this paper we use medical information system. It becomes extremely important to find such traits among patients that have the greatest impact on the occurrence of cancer [12].

In practice, we often find ourselves in situations, where the dependent variables we are measuring are zero-one, assuming the value 0 – lack of a phenomenon and 1 - occurrence of a phenomenon (concrete behaviour, consent to something, disclosure of attitudes, opinions, etc.). Both the general linear model and the linear regression analysis do not apply in the dichotomous, nominal dependent variable. In this situation, we are forced to apply nonlinear analyses. The regression model used for this type of dependent variable is logistic regression. The article presents the application of the binomial logistic regression model in experimental research [11].

2. Material and methods

In this paper, we have used the dataset taken from UCI Machine Learning Repository. Data set includes 669 instances and contains the following variables:

[*] Białystok University of Technology, Department of Biocybernetics and Biomedical Engineering, Wiejska 45C Str., 15-351 Białystok, e-mail: a.kasperczuk@pb.edu.pl

[†] Białystok University of Technology, Department of Biocybernetics and Biomedical Engineering, Wiejska 45C Str., 15-351 Białystok, e-mail: a.dardzinska@pb.edu.pl

- Sample code number: id number,
- Clump Thickness: 1–10,
- Uniformity of Cell Size: 1–10,
- Uniformity of Cell Shape: 1–10,
- Marginal Adhesion: 1–10,
- Single Epithelial Cell Size: 1–10,
- Bare Nuclei: 1–10,
- Bland Chromatin: 1–10,
- Normal Nucleoli: 1–10,
- Mitoses: 1–10,
- Class: (0 – for benign, 1 – for malignant).

We focus on creating the best classification algorithm for given information system using logistic model. Classification method is based on finding a mapping of data in a set of predefined classes [6]. Based on the contents of the database the model is constructed and used to classify new objects in the database or a deeper understanding of the existing division of objects on predefined classes. Classification methods find series of applications, for example: recognition of financial market trends, automatic object recognition in images of large databases, decision support granting bank loans [2, 6].

The main purpose of the classification is to build a formal model called the classifier. The input in the classification process is a training set of tuples (examples, observations, samples), which lists the descriptive attributes (descriptors) and the class label attributes. The result of the classification process can be obtained by a model (classifier), which assigns each tuple of decision attribute based on the values of others [5].

In the classification, first we describe a set as a predetermined class. Each tuple is assumed to belong to a predefined class as determined by a class label attribute, the set of tuples called training set are used for model construction. The model can be represented as decision tree, classification rule, or mathematical formula. It helps in analysis and description of future data trends and unknown objects. It estimates the accuracy of the constructed model by using certain test cases. Test sets are always independent of the training sets. In this work, we apply supervised classification [4, 9].

One of the most popular classification methods is logistic regression. The first work on the application of logistic functions was made at the end of nineteenth century in a statistical environment dealing with the description of demographic features. The complete logistic regression model was developed in 1972. This was done by D.J. Finney in *Probit analysis* [14]. This statistical method is applicable where the dependent variable is measured on a nominal scale and assumes two values coded as 0 - no occurrence of the considered phenomenon and 1 - occurrence of the phenomenon [10]. There is also a modified version of the classic logistic regression used for multi-dependent dependent variables - called polynomial logistic regression.

The best known and simplest method for testing the significance of differences between groups for categorical variables - test χ^2 - is primarily used in four-column tables. The χ^2 test can also be used for multi-field tables, but its result is difficult to interpret in such a situation (the significance of the test is for the whole table - that is, all values of both variables). Moreover, it often happens that one or two in a table cell multipole determine the statistical significance χ^2, while in other fields remain similar numbers. It does not meet the requirements of researchers planning more complex experiments than involve one binary variable independent variable. An indispensable method is to allow for a comprehensive analysis of the model, thus taking into account several independent variables, not necessarily of the same type.

These requirements are met by logistic regression, which is a mathematical model that we can use to describe the influence of one or more independent variables on a dichotomous dependent variable. It allows inclusion of quantitative variables (measured on an interval scale) and qualitative variables (measured on a nominal scale) in the model. The terms of application of this calculation method are much less restrictive than the General Linear Model. In addition to the aforementioned dichotomy of the dependent variable, sufficient logistic regression is required for the use of logistic regression. The sample size must be greater than $10(k+1)$ where k is the number of independent variables.

Logistic regression is a frequently used statistical method for classification problems when the variable is explained on a dichotomous scale. This means that the predictive model of logistic regression determines the probability of one of two possible outcomes: pathologies or no pathologies. Logistic regression is based on a specific way of expressing probability, called the odds ratio (OR). The opportunity quotient determines the probability as the ratio of probability of success to probability of failure.

$$S(A) = \frac{p(A)}{p(\neg A)} = \frac{p(A)}{1 - p(A)} \tag{1}$$

It can be calculated using the above formula. The main advantage of the odds ratio, in comparison to conventional probability, is that the odds ratio assumes values in the range $(0, +\infty)$ for p range from 0 to 1, and the logarithm value of the field $(-\infty, +\infty)$. This means that we can use regression methods not limited to a range $(0,\ 1)$ such as linear regression to estimate the log of chance in a regression model. $S(A)$ determines the odds ratio for malignancy among patients with known risk factors for cancer and for pathological survival among patients who do not have this risk. For tumour diagnostics, the odds ratio greater than 1.0 indicates that the risk of hyperplasia or altered cells among patients with increased risk of OR is increased, while $S(A) < 1.0$ allows for exclusion of pathology [10].

Definition 1 (Odds ratio)**.** *For the odds ratio, a 95% confidence interval is given, the span is based on the number of patients in the study group. The odds ratio can also be calculated taking into account the division of the respondents into two separate groups using the following formula:*

$$OR_{AxB} = \frac{S(A)}{S(B)} = \frac{p(A)}{1 - p(A)} \div \frac{p(B)}{1 - p(B)}. \tag{2}$$

We interpret them as follows:

- If $OR > 1$, then in the first group the occurrence of the event is more likely.
- If $OR < 1$, the event occurs in the second group more likely.
- If $OR = 1$, then the event is equally likely in both classes.

Definition 2 (Likelihood function)**.** *We derive the form of the reliability function* L *for logistic regression. The explanatory variable* Y *is binary and for single observation* i *and occurs:*

$$Y_i|X_i = \begin{cases} 1 & \text{with probability } \ p(X_i), \\ 0 & \text{with probability } \ 1 - p(X_i), \end{cases} \tag{3}$$

$$L(X_i, \beta) = P(Y_i = 1|X_i)^{Y_i} \cdot P(Y_i = 0|X_i)^{1-Y_i} = p(X_i)^{Y_i} [p\,(X_i)]^{1-Y_i}. \tag{4}$$

Definition 3 (Logistic function)**.** *Transformation function on the logarithm of the probability of chance is called* logit*:*

$$\text{logit}(P) = \ln \frac{p}{1-p} = \ln(p) - \ln(1-p), \tag{5}$$

where

$$p = \frac{\mathrm{e}^{\text{logit}(p)}}{1 + \mathrm{e}^{\text{logit}(p)}} = \frac{1}{1 + \mathrm{e}^{\text{logit}(p)}}. \tag{6}$$

The logistic model is based on the function $f(z)$ *shown in the following Figure 1:*

$$f(z) = \frac{\mathrm{e}^{z}}{1 + \mathrm{e}^{z}} = \frac{1}{1 + \mathrm{e}^{-z}}. \tag{7}$$

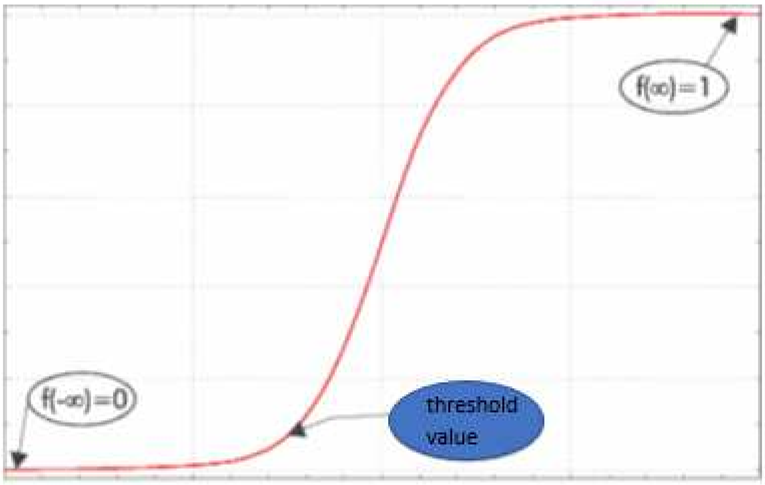

Fig. 1. The form of logistic function [10]

From the graph of the logistic function, it can be concluded that feature changes are minimal before the function reaches a threshold, and then rises sharply until it reaches a value that does not significantly affect the probability of the event. In this situation, the graph remains at a level close to 1. The predictive model predicts the probability of malignancy with the probability of benign tumours. The constructed model also allows us to observe which of the tested independent variables can influence the dependent variable explained on a dichotomous scale. Conditional probability that the dependent variable Y takes a value 1 for the values of the independent variables $x_0,\ x_1,\ldots,x_k$ can be described using a following pattern:

$$P\left(Y=1|x_0,\,x_1,\ldots,x_k\right)=\frac{\mathrm{e}^{\left(a_0+\sum_{i=1}^{k}\ a_i x_i\right)}}{1+\mathrm{e}^{\left(a_0+\sum_{i=1}^{k}\ a_i x_i\right)}},\tag{8}$$

where a_i, for $i=0\ldots k$ are regression coefficients, while $x_1,\ x_2,\ldots,x_3$ represent independent variables. The independent variable in logistic regression model can be quantitative or qualitative. Estimates are calculated using the most reliable method. The greater reliability of a model, the more likely it is that the variable will appear in the sample, which means better matching model to the data [14].

3. Methods for estimating parameters and testing hypotheses

ML — maximum likelihood – in the linear regression model, the regression constant and the regression coefficient are estimated by the least squares method. This method does not apply to logistic regression because of the lack of linearity of the dependent variable distribution. Logistic regression coefficients are estimated using the most reliable method. The calculation algorithm of such method is based on multiple estimates of all regression coefficients, so as to maximize the probability of obtaining the results obtained in the trial [10]. The calculation formula takes into account the combined probability for the criterion cases (the dependent variable was 1) and no criteria cases (those for which the dependent variable was 0). This method of estimating parameters requires very complex calculations (means multiplying the probability coefficients for different parameter values until the maximum product is reached — maximum reliability), so the relevant statistical software is now used for this purpose [6].

LR — likelihood ratio — using a quotient of reliability, we answer the question whether a model containing variable (independent) variables will give us better predictability (i.e., behavior of the test), than the model without the variable. Calculating this factor means comparing two values of reliability statistics, namely its specific form, i.e. logarithmised reliability statistics multiplied by $-2(-2\ln likelihood)$.

The formula for the likelihood ratio is:

$$LR=-2\ln L_1-(-2\ln L2).\tag{9}$$

Where:

- $\ln L_1$ logarithm of the reliability function for the full model,
- $\ln L_2$ logarithm of the reliability function for the reduced model.

The distribution of the values of the confidence quotient is consistent with χ^2 distribution with so many degrees of freedom, how many variables differed the full model from the reduced model.

Wald Test — is used to test the statistical significance of each coefficient (b) in the model. A Wald test calculates a Z statistic. This Z value is then squared, yielding a Wald statistic with a χ^2 distribution. The calculation formula for the Wald coefficient is very simple and is based on the already established regression coefficients and their standard errors. We obtain Wald's statistic by dividing the parameter estimate at variable X by the standard error of this estimate (denoted by SE) [10]:

$$Wald(Z) = \frac{\beta_1}{\mathrm{SE}_\beta}. \tag{10}$$

Hosmer-Lemeshow Goodness of Fit Test — evaluates the goodness-of-fit by creating 10 ordered groups of subjects and then comparing the actual number in each group (observed) to the number predicted by the logistic regression model (predicted). Thus, the test statistic is a χ^2 statistic with a desirable outcome of non-significance, indicating that the model prediction does not significantly differ from the observed [13].

Definition 4 (ROC curve — Receiver Operating Characteristic)**.** *In this work, we also use the ROC analysis to verify the model. ROC curves are a statistical tool used to determine the accuracy of the classification [3, 15] as a statistical method initially were part of the "Signal Detection Theory", which was established during the Second World War with the aim of analysing radar images.*

Upon receipt of the radar picture, the operator must decide whether the signal on the monitor represents an enemy object, an ally, or if it is a disturbance in the system. The theory of signal detection was aimed at determining the operator's ability to differentiate these signals and was named Receiver Operating Characteristics. In the 1970s, the theory of signal detection was used as a statistical method in medicine.

Today, ROC curves are a frequently used method to describe the diagnostic accuracy of a test. They are used for the comparison of diagnostic tests, analyses whether the test is effective in distinguishing between different populations, and determining the cut-off value.

In medicine, a common problem is the need to compare different types of diagnosis, therapy, or diagnostic tests. For example, how do you determine which diagnostic procedure is better for distinguishing a sick person from healthy? How do you decide whether adding a new test or a treatment to an existing treatment plan has a positive or negative effect on the condition and progression of the disease? Analysis of similar issues can be done using ROC curves. ROC curves describe the whole range of the classifier's work and allow for a meaningful comparison of the results from the various classifiers. The advantage of using ROC curves is their independence from the units and scales used in the study and the ability to use categorical data (not just binary).

The curve construction is based on varying threshold values across the test area. The curve outlines the sensitivity versus specificity for the entire range of decision thresholds. Sensitivity is represented on the axis of Oy and is a true positive fraction, calculated on the basis of data from patients with a given pathology. On the other hand, the specificity value is represented on the Ox axis, and is a false positive fraction, calculated on the basis of the data in the subgroup of healthy patients.

We can use the AUC (Area Under Curve), which is the surface area under the ROC curve to obtain more accurate classification performance. The area under the ROC curve is the probability that a classifier will give a higher profile to chance of randomly selected instance from the respective groups, and no chance of randomly selected group, in which it is known that the requested data does not exist. The AUC includes a description of the detection precision across the entire operating range of the system. An AUC of 0.5 can be described as a random action, and a value of 1.0 is an ideal indicator. This means that a curve closer to the upper left corner shows greater diagnostic accuracy [7, 8].

The commonly used quality description of diagnostic can be described as follows:

- $0.9 - 1.0$ — very good,
- $0.8 - 0.9$ — good,
- $0.7 - 0.8$ — satisfactory,
- $0.6 - 0.7$ — average,
- $0.5 - 0.6$ — insufficient.

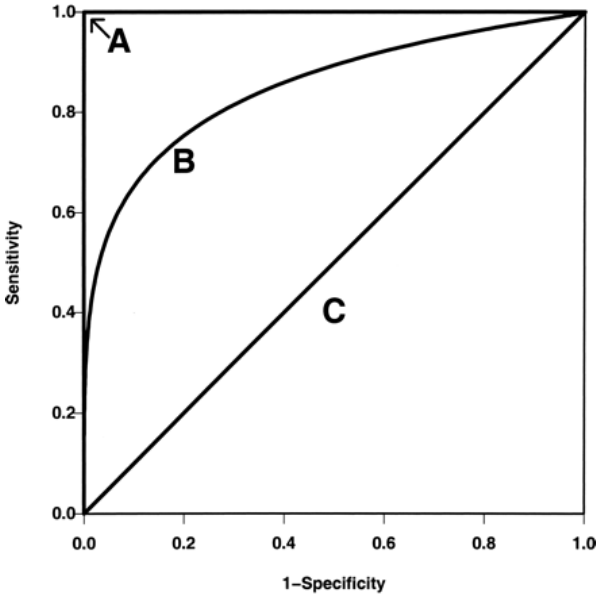

Fig. 2. Hypothetical ROC curves [8]

Three hypothetical ROC curves (Fig. 2) representing the diagnostic accuracy of the gold standard (lines A; AUC= 1) on the upper and left axes in the unit square, a typical ROC curve (curve B; AUC= 0.85), and a diagonal line corresponding to random chance (line C; AUC= 0.5).

The AUC calculation is performed by computer programs. The surface area is calculated most often by two methods [10]:

- the non-parametric method is based on the construction of trapezoids under the curve to estimate the area under the graph,
- the parametric method is based on the maximum likelihood estimator to match the continuous lines to the data plot.

4. Results and discussion

In this part of the analysis we calculate how independent variables may affect the dependent variables: Class. Table 1 and Figure 3 show the distribution of relapse.

To verify which of the predictors have an impact on affects the type of cancer, logistic regression methods can be applied. We select the best logistic regression model for our data. The quality of the model is assessed on the basis of two criteria: compliance with logistic regression and ease of medical interpretation, which makes it practical to use. Various methods of variable selection have been developed, however we applied some of them, the most informative from our data point of view.

The first method selected explanatory variables based on medical knowledge and intuition. Unfortunately, the disadvantage of this approach is the risk of incorrectly mapping dependencies and over-extension of the model. The consequences of using this method can also be the burden of estimating coefficients or standard errors.

Another method tested statistical significance by performing tests based on, among other things, Wald$Z$ statistics or Deviance statistics. We then reject these explanatory variables, for which the *p*-value of the test is greater than the determined significance level. This level is higher than usually used in testing hypotheses of 5%. Hosmer and Lemeshow advise the choice $\alpha = 25\%$.

We used the step method to select variables. First, we include all predictors in the model. The normality of distributions of the variables tested was assessed using the Shapiro-Wolf test. To detect the significance of differences between compared groups of patients we used the median test and the Mann-Whitney or Kruskal-Wallis quantitative characteristics and uniformity test for qualitative features. In the first part of the study, a statistical analysis was carried out in order to investigate which of the analysed parameters have a statistically significant influence on the probability of malignant cancer in the examined group of pats. For each of the tested features, the probability level *p* was calculated. When *p* is below 0.05 it can be stated that the tested feature is statistically significant. Statistical analysis was carried out in the whole group of studied patients.

For classification algorithm, we used cross validation option with 10 folds. 10-folds cross validation works as follows:

- divide dataset into 10 parts (folds),
- hold out each part in turn,
- average the results,
- each data point used once for testing, 9 times for training.

Tab. 1. Distribution of relapse

Class	Distribution	
	Population	Percentage
benign	458	68.98
malignant	206	31.02

Received model is statistically significant: $\chi^2(13) = 55.69; p < 0.001$. It explains 26% of the observed variance (R — $squared\ Nagelkerke = 0.260$). Hosmer and Lemeshow test showed that the data fit to the model: $\chi^2(8) = 8.52; p < 0.384$.

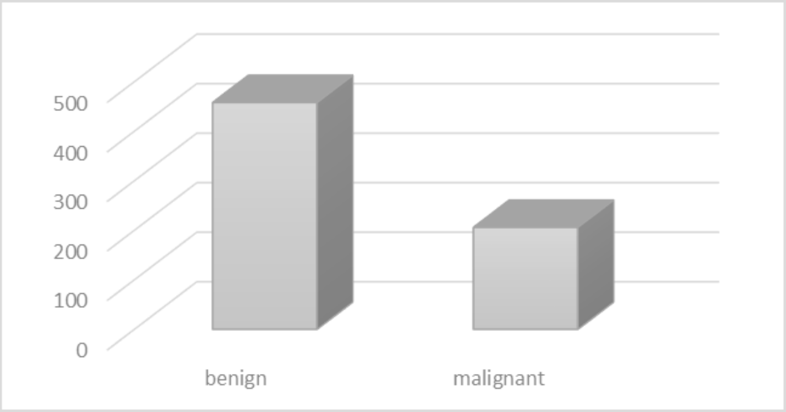

Fig. 3. The distribution of the class attribute

Table 2 and Figure 4 show the classification results.

Tab. 2. Confusion matrix

Observed effects	Expected effects	
	benign	malignant
benign	439	19
malignant	29	212

The model correctly classifies about 93.13% of the observed events (651 instances). The value of incorrectly classified instances is equal to 6.87% (48 instances).

Due to many variables appeared to be significant predictors in the regression model we once again did a logistic regression analysis using stepwise method (Forward Wald method).

The model proved to be statistically significant: $\chi^2(2) = 40.01$; $p < 0.001$. It explained 25.8% of the observed variance (R — $squared\ Nagelkerke = 0.258$). Hosmer and Lemeshow tests showed that the data fit to the model: $\chi^2(3) = 3.61$; $p = 0.202$. Table 3. shows the parameters for the predictors entered into the model. The analysis in two steps showed that the variables in Table 3. were statistically significant predictors in the defining the type of cancer (class 4).

We also should look at the following parameters we received (Han and Kamber 2006):

- TP Rate — rate of instances correctly classified as a given class,
- FP Rate — rate of instances falsely classified as a given class,
- Precision — proportion of instances that are truly of a class divided by the total instances classified as that class,

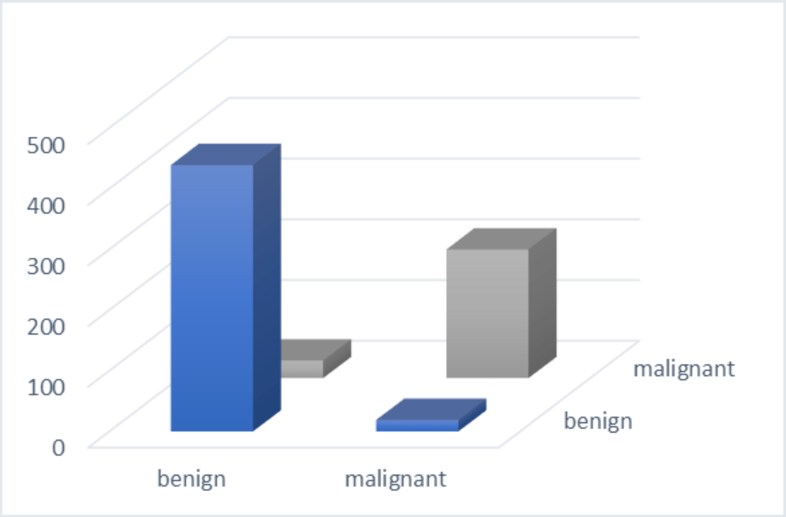

Fig. 4. The distribution of observed and expected effects

Tab. 3. Parameter for the model — Forward Wald Method

Variable	Parameters				
	Exp(B)	Wald Statistic	95% CI of Exp(B) Low	95% CI of Exp(B) High	Level of Significance
constant	26.3498	66.69096	20.0	32.67	0.000000
Clump_Thickness (3)	-6.3206	10.31703	-10.2	-2.46	0.001318
Clump_Thickness (4)	-4.9976	6.99083	-8.7	-1.29	0.008193
Clump_Thickness (5)	-4.9021	6.85759	-8.6	-1.23	0.008827
Clump_Thickness (6)	-4.9946	8.63392	-8.3	-1.66	0.003300
Uniformity_of_Cell_Size (1)	-10.2472	11.17242	-16.3	-4.24	0.000830
Uniformity_of_Cell_Size (2)	-7.8144	9.13735	-12.9	-2.75	0.002504
Uniformity_of_Cell_Size (3)	-6.2301	8.24159	-10.5	-1.98	0.004094
Marginal_Adhesio (5)	-7.6423	25.71713	-10.6	-4.69	0.000000
Marginal_Adhesio (6)	-5.7055	8.88999	-9.5	-1.95	0.002867
Bare_Nuclei (1)	-5.7191	10.43872	-9.2	-2.25	0.001234
Bare_Nuclei (5)	-4.6182	4.35399	-9.0	-0.28	0.036922
Bland_Chromatin (1)	-13.1932	10.84947	-21.0	-5.34	0.000988
Bland_Chromatin (2)	-7.6565	19.47955	-11.1	-4.26	0.000010
Bland_Chromatin (3)	-8.0774	20.69868	-11.6	-4.60	0.000005
Bland_Chromatin (4)	-5.9190	10.10132	-9.6	-2.27	0.001482

- Recall — proportion of instances classified as a given class divided by the actual total in that class,
- F-Measure — general indicator of quality of the model
- ROC Curve (ROC Area) — the graph from Figure 2 shows three ROC curves representing excellent, good, and worthless tests plotted on the same graph. The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question. Accuracy is measured by the area under the ROC curve,
- Kappa Statistic — it is a measure of conformity between the proposed allocation instance of the class and the actual, which is about the overall accuracy of the model.

For the all of the classification parameters we received great results (Table 4), what indicates we have obtained a very good classification model.

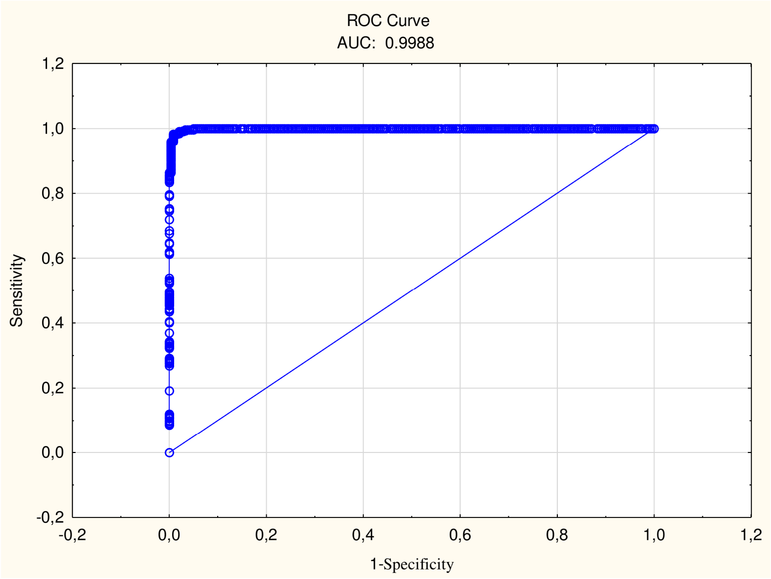

Fig. 5. ROC Curve

Changes in sensitivity and specificity when shifting the limit value for the features and model used are shown in ROC curve (Receiving Operating Characteristic Curve) and field under the curve (AUC). Clinical data was used for model construction. ROC curve presented in Figure 5 was obtained from Statistica software.

5. Conclusions

In this paper, we conducted classification analysis in breast cancer database using selected statistical methods. We built logistic regression model for dependent variable Class (informing about type of a cancer). It becomes important to find such traits, which have the greatest

Tab. 4. Classification parameters

Factor	Classification algorithm
	Logistic
Kappa Statistic	0.847
TP Rate	0.931
FP Rate	0.093
Precision	0.931
Recall	0.931
F-Measure	0.931
ROC Area	0.998

impact on type of a cancer cells. In this work, we use logistic regression method to build models showing what variables have the greatest impact on the dependent variable or are the most linked to it. The obtained results are interesting and they indicate, that considered methods can be useful in diagnosis and treatment of patients with breast cancer.

Acknowledgements

Research was performed as a part of projects MB/WM/8/2016 and financed with the use of funds for science of the Ministry of Science and Higher Education.

References

[1] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. In *Int J Med Informatics*, pages 81–97, 2008.

[2] R. Bouckaert. *Naive Bayes Classifiers That Perform Well with Continuous Variables*, volume 3339. Lecture Notes in Computer Science, 2004.

[3] R. Centor and G. Keightley. Receiver operating characteristics (roc) curve area analysis using the roc analyser. *Proc 13th Ann Symp Computer Application in Medical Care*, pages 222–262, 1989.

[4] A. Dardzinska. *Action Rules Mining*. Springer, 2013.

[5] A. Dardzinska and A. Romaniuk. Incomplete distributed information systems optimization based on queries. *Advances in Swarm and Computational Intelligence*, 9142:265–274, 2015.

[6] J. Deogun, V. Raghavan, and H. Sever. Rough set based classification methods and extended decision tables. *International Workshop on Rough Sets and Soft Computing*, pages 301–309, 1994.

[7] W. Frawley and G. Piatetsky-Shapiro. Knowledge discovery in databases. *An overview, Knowledge Discovery in Databases*, pages 1–27, 1991.

[8] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann Publishers, 2006.

[9] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *ACM SIGMOD International Conference on Management of Data*, pages 1–12, 2000.

[10] D. Hosmer and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2013.

[11] A. Kasperczuk. Selected logistic regression methods in medical database. *Badania i Rozwój Młodych Naukowców w Polsce: Nauki techniczne i inżynieryjne*, 8:87–94, 2017.

[12] A. Kasperczuk and A. Dardzinska. Comparative evaluation of the different data mining techniques used for the medical database. *Acta Mechanica et Automatica*, 10(3), 2016.

[13] B. Modan, E. Ron, and G. Lerner. Cancer incidence in a kohort of infertile women. *Am J Epidemiol*, (147):1038–1042, 1998.

[14] A. Stanisz. *Przystępny kurs statystyki z wykorzystaniem programu STATISTICA PL na przykładach z medycyny*. StatSoft Polska, Kraków, 2000.

[15] W. Zatonski. *Nowotwory złośliwe w Polsce w 2002 roku.* Centrum Onkologii — Instytut im. Marii Skłodowskiej-Curie, Warszawa, 2004.

ABSTRACT

Data analysis methods are widely used to solve many different problems. Intensive development in the field of knowledge discovery in database is a response to a sharp increase in the amount of information in electronic format. Classification is one of the main steps in data analysis. Taking into consideration medical data, it becomes extremely important to obtain knowledge containing valuable information about the patients with serious illness, e.g. cancer. In this paper we mainly focused on logistic regression model parameters. We built model, which shows the impact of predictors on the dependent variable. It will help to create the medical knowledge base as a next step.

ZASTOSOWANIE METOD REGRESJI LOGISTYCZNEJ W MEDYCZNYCH SYSTEMACH INFORMACYJNYCH

STRESZCZENIE

Metody analizy danych są szeroko stosowane w rozwiązywaniu problemów z różnych dziedzin. Intensywny rozwój w zakresie odkrywania wiedzy w bazach danych jest odpowiedzią na gwałtowny wzrost ilości informacji w formie elektronicznej. Klasyfikacja jest jednym z kluczowych etapów analizy danych. Biorąc pod uwagę dane medyczne, niezwykle ważne staje się pozyskanie wiedzy zawierającej cenne informacje na temat pacjentów objętych np. chorobami nowotworowymi. W artykule przedstawiono metodykę budowy modelu regresji logistycznej pacjentów z chorobą nowotworową się głównie na parametrach modelu regresji logistycznej. Zbudowaliśmy model, który pokazuje wpływ czynników predykcyjnych na zmienną zależną. Pomoże to stworzyć medyczną bazę wiedzy jako kolejny krok.

Received: 2017-11-03
Accepted: 2017-12-05