JOURNAL OF MODERN SCIENCE
SPECIAL ISSUE
3/57/2024
www.jomswsge.com

DOI: doi.org/10.13166/jms/191165

KRZYSZTOF KRÓL
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0000-0002-0114-2794

TOMASZ SIDOR
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0009-0002-6711-8751

ANNA WIŚNIEWSKA
Vistula University in Warsaw, Poland
ORCID iD: orcid.org/0000-0003-0876-1763

EMANUEL JÓZEFACKI
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0000-0003-0078-6392

BARTŁOMIEJ BARTNIK
Graduate School of Business - National-Louis University, Poland
ORCID iD: orcid.org/0000-0003-2370-3326

ANALYZING MARKETING CAMPAIGN EFFECTIVENESS: A COMPARATIVE APPROACH USING TRADITIONAL AND ONLINE DATA ANALYSIS METHODS

ANALIZA SKUTECZNOŚCI KAMPANII MARKETINGOWYCH: PODEJŚCIE Z WYKORZYSTANIEM TRADYCYJNYCH I INTERNETOWYCH METOD ANALIZY DANYCH

Abstract

Advertising campaign analysis reports are considered an essential tool for marketing analytics. They are used to assess the effectiveness of the marketing activities carried out and to improve future activities. It is necessary to verify whether the actions taken – online and in the public space – align with the intentions and budget, whether they lead to achieving the objectives, and, if not, what the campaign errors are. Due to the ease of collecting and accessing data, analyzing online and social media advertising campaigns is a popular topic. With access to data on the number of clicks, the ad's reach, the number of interactions, and so on, one can move on to the next steps of analyzing the campaign to determine its effectiveness. Online marketing tools have a massive advantage over traditional media channels. When analyzing the results of advertising campaigns, it is necessary to approach the examination of the individual channels and then analyze which of them is the most profitable and in which to invest the most. However, traditional campaigns must be addressed in the analyses. Despite the limited data available, collecting relevant information and analyzing the traditional campaign is worth trying. In the case of conventional campaigns, we can mainly measure the amount of sales resulting from the campaigns. When dealing with an online campaign, we gain many additional indicators, such as the number of ad impressions, clicks, and conversions. In both cases, analysis tools may allow us to isolate factors that significantly influence the success or failure of a campaign and predict the effectiveness of a campaign with given characteristics.

Streszczenie

Raporty z analizą kampanii reklamowej są uważane za podstawowe narzędzie analityki marketingowej. Służą do oceny skuteczności przeprowadzonych działań marketingowych oraz do ulepszenia przyszłych działań. Ważne jest zweryfikowanie, czy podjęte działania – zarówno w sieci jak i w przestrzeni publicznej – są zgodne z zamierzeniami i budżetem, czy prowadzą do osiągnięcia założonych celów, a jeśli nie, to na czym polegają błędy w prowadzeniu kampanii. Ze względu na łatwość zbierania i dostępu do danych, popularnym zagadnieniem jest analiza kampanii reklamowych prowadzonych w Internecie i mediach społecznościowych. Mając dostęp do danych o liczbie kliknięć, zasięgu reklamy, liczbie interakcji i tak dalej, można przejść do kolejnych kroków analizy kampanii, by określić jej skuteczność. Narzędzia marketingu online mają w tej kwestii ogromną przewagę nad tradycyjnymi kanałami przekazu. Analizując wyniki kampanii reklamowych należy podejść do zbadania poszczególnych kanałów, a następnie przeanalizować, który z nich jest najbardziej opłacalny, i w który należy najwięcej inwestować. Nie można, jednakże w analizach

pominąć całkowicie kampanii prowadzonych w sposób tradycyjny. Mimo mniejszej dostępności do danych, warto włożyć wysiłek w zebranie odpowiednich informacji i przeanalizowanie również kampanii tradycyjnej. W przypadku kampanii tradycyjnych możemy mierzyć głównie wysokość sprzedaży wynikającą z przeprowadzonych kampanii. Mając do czynienia z kampanią internetową, zyskujemy wiele dodatkowych wskaźników, jak liczba wyświetleń reklamy, liczba kliknięć, liczba konwersji. W obu przypadkach narzędzia analizy mogą pozwolić na wyodrębnienie czynników istotnie wpływających na sukces lub porażkę kampanii, a także na przewidzenie skuteczności kampanii o danych cechach.

Keywords: *analysis of advertising campaigns, dataset in clustering, marketing analytics, RMSE, RMSLE*

Słowa kluczowe: *analiza kampanii reklamowych, zbiór danych w klastrach, analiza marketingowa, RMSE, RMSLE*

Introduction

Analyzing advertising campaigns is a vital tool in marketing to assess the effectiveness of promotional activities and to improve future strategies. It is essential to check whether the activities undertaken – both online and offline – are in line with the objectives and budget, whether they are delivering the expected results, and, if not, where the reason for campaign failures lies. The analysis of advertising campaigns, especially those conducted online and on social media platforms, is popular because of the ease of data acquisition. With information on clicks, reach, or interactions, the effectiveness of a campaign can be accurately assessed. Online marketing tools have a significant advantage over traditional communication channels, enabling detailed analysis and identification of the most cost-effective promotion channels (Borysiak, Wołowiec, Gliszczyński, Brych, Dluhopolskyi, 2022).

However, traditional advertising campaigns should be considered. Although data is less available, it is worth taking the time to collect and analyze it. For conventional campaigns, the leading indicator of success may be sales volume. For online campaigns, we have several additional metrics, such as the number of impressions or clicks, which allow us to assess the effectiveness of activities

more accurately. Regardless of the type of campaign, analysis will enable us to identify the key factors determining success or failure and to forecast the effectiveness of future promotional activities (Pate,l 2020; Yuan, 2019; Zheng).

The prepared functionality of the system makes it possible to carry out analyses of the effectiveness of both traditional and online advertising campaigns. In the following chapters, the selected data sets for both types of campaigns and the process of cleaning these data from deficiencies are described. Descriptive analyses were carried out on the data to provide a preliminary understanding of the factors influencing campaign effectiveness. Predictive and regression analyses will predict the success of the campaign and identify the variables that contribute most to the success or failure of the campaign (Sarstedt, 2014).

Research Methodology

A traditionally run campaign's objective is defined as follows: A particular restaurant chain plans to add a new item to its menu. However, it is still being determined which of three marketing campaigns will be used to promote the new product. The latest item is introduced at locations in several randomly selected markets to determine which promotion impacts sales most. A different promotion is used in each area, and weekly sales of the new item are recorded for the first four weeks.

The dataset for the restaurant chain's advertising campaign contains the following columns:

- MarketID: unique identifier of the market
- MarketSize: size of the market in terms of sales
- LocationID: unique identifier of the shop's location
- AgeOfStore: age of the shop in years
- Promotion: one of the three promotions that were tested
- Week: one of the four weeks during which the campaigns were run
- SalesInThousands: sales in thousands for the location, promotion, and week

The collection consists of a total of 548 observations.

The Facebook campaign dataset was selected to analyze online and social media campaigns. The dataset covers dates from 1 January 2020 to 30 September 2022. It contains information on 6723 campaigns. Variables appearing in the dataset include:

- impressions – number of impressions of a given ad
- frequency – the average number of impressions of a given advertisement by one user
- spend – the amount of money spent on a particular ad
- social_spend – a subset of spending related to users liking, commenting, and sharing content
- clicks – number of clicks
- reach – reach, i.e., the number of users who have viewed an ad at least once
- CPC – cost per click, the price an advertiser pays for each click on a link (quotient of the cost of a campaign by the number of clicks)
- CPM – cost per thousand (cost per mille), the price an advertiser pays for 1,000 ad impressions (cost divided by the number of impressions multiplied by 1,000)
- cpp – cost per pixel, the average amount spent on conversions from tracking pixels in adverts (a tracking pixel is a piece of code that allows data about a user's behavior to be collected and used to display tailored ads)
- ctr – click-through rate, the number of clicks divided by the number of impressions
- cost_per_inline_post_engagement – the cost of a user's interaction with an ad on social media (e.g., liking, sharing, commenting, etc.)
- cost_per_unique_click – the average amount spent per click by a unique user
- inline_post_engagement – number of times users interacted with an advertisement
- objective – objective, e.g., link clicks, interactions, page likes, and others
- optimisation_goal – in most cases, none, goals such as *Landing Page Views* and *Thruplay* appear

- actions-action_type – a type of action, e.g., interaction, add to cart, search, purchase, and others
- actions-value – the value of the action performed
- cost_per_action_type-action_type – analogous to actions-action_type
- cost_per_action_type-value – the cost of a given action incurred by the advertiser
- cost_per_unique_action_type-action_type – analogous to actions-action_type
- cost_per_action_type-action_type-value – analogous to cost_per_action_type-value, but calculated as an average value for unique users.

Data cleaning

The number of missing data in the datasets is checked using the ISNA () function of the Python language, derived from the Panda's library. The numbers of missing data are then summed in columns (Rachwał 2023). The dataset contains no data gaps for traditionally run campaigns. There are also no campaigns for which sales would be zero. The Facebook campaign dataset contained many missing data and required thorough cleaning. The dataset examined initially contained 53671 rows and 491 columns. After removing the empty columns, 79 columns remained in the collection. All columns with at least 30% missing values were also removed from the collection. In the remainder of the set, the most missing data (between 20 and 22%) was in the columns cpc, cost_per_inline_post_engagement, cost_per_unique_click, followed by 15. 24% in the cpm, cpp, and ctr columns, and almost 7% in the actions-action_type, actions-value, cost_per_action_type-action_type, cost_per_action_type-value, cost_per_unique_action_type-action_type, cost_per_unique_action_type-value columns. The inline_post_engagement column had a few missing data (less than 1% of the set). The percentages of missing data content in the columns that still need to be deleted are shown in Figure 1.

Figure 1. *Percentage of missing data in Facebook collection columns*

cost_per_unique_click	22.042720
cpc	22.041128
cost_per_inline_post_engagement	20.732794
cpp	15.240020
ctr	15.240020
cpm	15.240020
cost_per_unique_action_type-action_type	6.907748
cost_per_unique_action_type-value	6.907748
cost_per_action_type-value	6.887057
cost_per_action_type-action_type	6.887057
actions-value	6.887057
actions-action_type	6.887057
inline_post_engagement	0.904056

The missing columns cpp, ctr, and cpm appear where the number of impressions (impressions) and clicks (clicks) is zero. For this reason, these values have been filled in with zeros. In the cost_per_unique_click column, gaps appear where the number of clicks is zero – these gaps have been filled in with zeros (except for one row where the number of clicks was 1 – in this case, the gap was filled in by the value from the cpc column). Deficiencies in the CPC column only occur when the number of clicks is zero, so these were also filled in with zeros. Deficiencies in the columns containing action_type were filled by the phrase ‚no_action.' Missings in cost_per_inline_post_engagement, cost_per_unique_action_type-value, cost_per_action_type-value, and actions-value occur when inline_post_engagement (number of interactions) is zero. Hence, they have been filled out with zeros. Deficiencies in the cost_per_unique_action_type-action_type were filled in with values from the cost_per_action_type-action_type column. Missing columns in inline_post_engagement were filled by imputation using the k nearest neighbors method with k=10 (Beretta, 2016). In addition, despite the high percentage of missing data, conversion-related columns were left in the collection: conversions-action_type, conversions-value, cost_per_conversion-action_type, cost_per_conversion-value. Deficiencies in the numeric columns were filled by zeros and in the qualitative columns by ‚no_action.' It was decided to keep these columns in the dataset because conversion is essential to the parameters studied regarding campaign effectiveness. Another element of data cleaning was the analysis of the presence of duplicates. Removal of duplicate rows resulted in a dataset with 53602 rows.

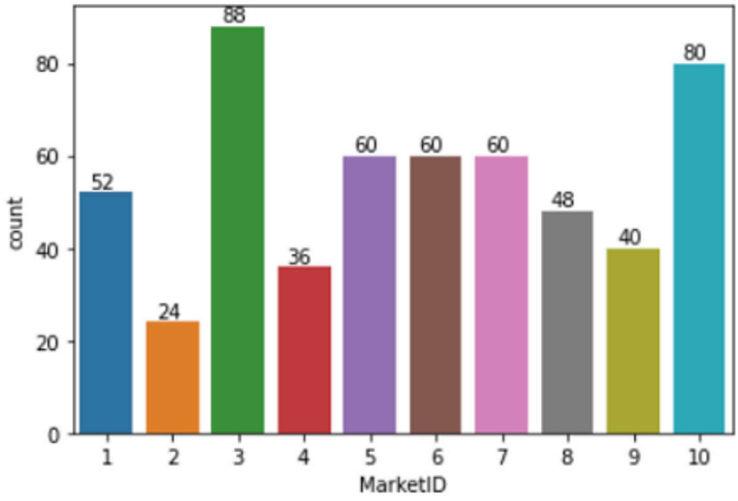
Implementation of appropriate analytical methods and development of predictive models

The following methods were used in the descriptive analysis of the collection(Kornacki, 2008):

- analysis of descriptive statistics for numerical variables, counts of occurrences of values for categorical variables,
- analysis of box plots for numeric variables, bar charts for categorical variables,
- chart analysis of the dependent variable by different values of the dependent variables.

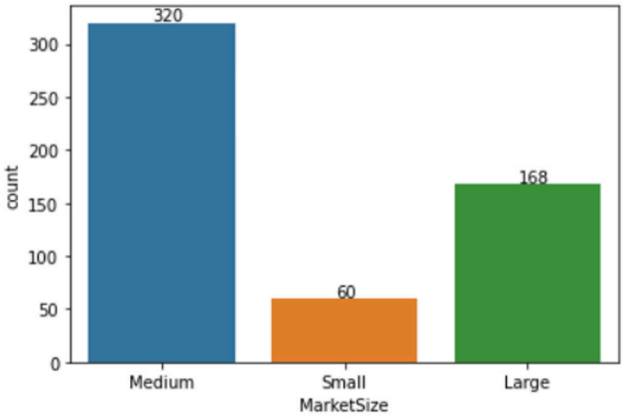
The figure shows a bar chart for the variable MarketID. We can see in it that ten different drawings appear in the set. Figures 3 and 10 have the most occurrences—88 and 80, respectively. Figures 5, 6, and 7 occur 60 times each. The fewest observations are for Figure 2—only 24 occurrences.

Figure 2. *Bar chart for the variable MarketID*



The figure shows a bar chart for the variable market size. Most drawings are in the dataset, and the average length is 320 observations. There are 168 large markets, and the lowest number of small markets is 60.

Figure 3. *Bar chart for the variable market size*



The figure shows a bar chart for the variable Promotion (type of campaign). Campaign types 2 and 3 appear 188 times in the dataset, while campaign 1 has 172 occurrences.

Figure 4. *Bar chart for the Promotion variable*

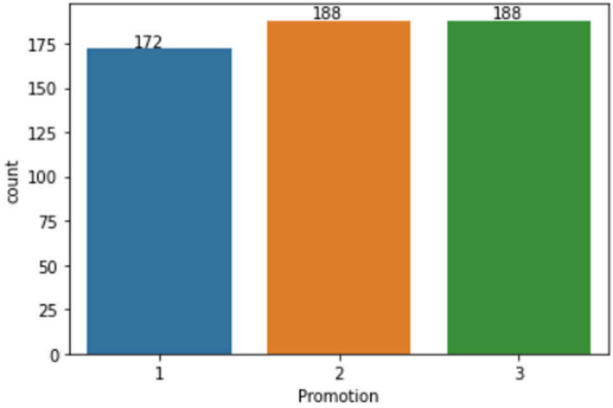


Figure 5 and Figure 6 show box plots of the variable denoting sales in thousands, broken down by the different levels of the categorical variable.

Figure 5. *Box plot of sales in thousands by market size*

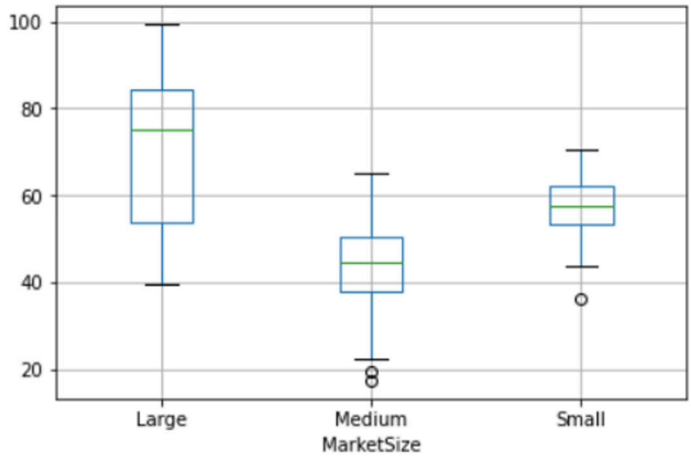


Figure 6. *Box plot of sales in thousands by campaign*

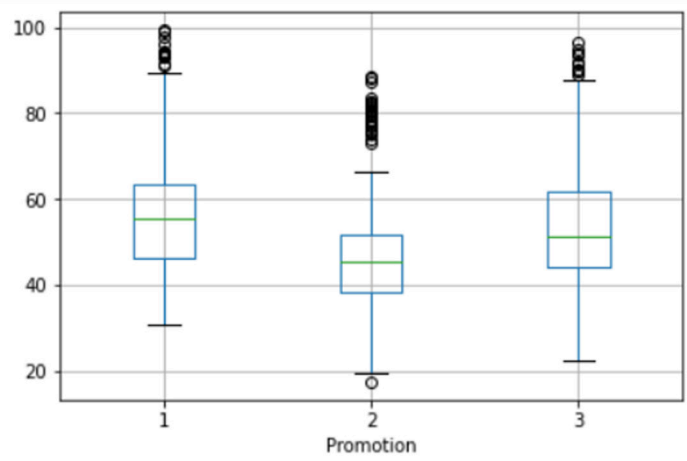


Figure 7 shows examples of the highest sales values during the campaign in restaurant chains.

To predict the success/failure of a given advertising campaign, the following predictive analysis methods were used:

- Decision tree
- Random forest
- Linear regression
- K nearest neighbors model (KNN)
- Support vector machine (SVM)
- Light GBM (improved decision tree algorithm)

The following methods will be used to assess the quality of the prediction from the model:

- regression model fit metrics, e.g. MSE, RMSE, RMSLE;
- Actual vs Predicted" graphs, comparing actual and predicted values.

Figure 7. *Table of the top ten sales during the campaign*

	MarketID	MarketSize	LocationID	AgeOfStore	Promotion	week	SalesInThousands
0	3	Large	218	2	1	1	99.65
1	3	Large	220	3	1	3	99.12
2	3	Large	209	1	1	4	97.61
3	3	Large	208	1	3	1	96.48
4	3	Large	209	1	1	2	96.01
5	3	Large	216	4	3	1	94.89
6	3	Large	210	19	1	3	94.43
7	3	Large	207	1	3	4	94.21
8	3	Large	202	8	1	4	94.17
9	3	Large	214	5	1	3	93.86

To assess the validity of the prediction, the datasets were split into a learning and a testing part. Each model used in the case was first trained on the learning part and then tested on a separate test part. Based on the metrics calculated for the test set, it is possible to assess the quality of model performance and comparisons between models.

Due to the construction of the datasets, where the dependent variable was in numerical form (in the case of the restaurant chain campaign, it is SalesInThousands, and in the case of the Facebook campaign, it is ctr), the models were in regression form. For this reason, the metric studied is the RMSE (root mean squared error), calculated from the formula.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

where n is the number of observations, y_i is the actual value of the dependent variable, and $\hat{y}_i$ is the value of the dependent variable predicted from the model. The RMSLE (root mean squared logarithmic error) metric was chosen for the traditional campaign dataset. This measure is used when the predictions have large deviations, which is the case for our collection, where sales values are reported in thousands. The formula defines RMSLE:

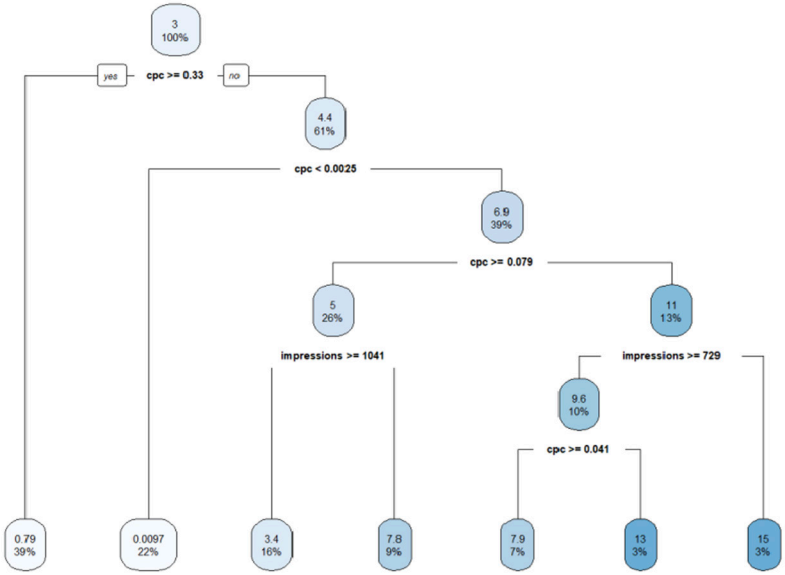
$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i + 1) - \log(y_i + 1))^2}$$

with designations identical to those of RMSE.

Table 1. *Comparison of predictive models for the Facebook collection*

Model	RMSE
Decision tree	0,52016
Random forest model	0,38545
Linear regression	0,91776
KNN	0,75860
SVN	0,93050

Figure 8. *Visualisation of the decision tree model for the Facebook collection*



The table in Figure 8 compares RMSE values for regression models on the Facebook set. The random forest (Cutler 2012) model had the lowest prediction error, so the ctr values predicted from this model were, on average, closest to the real ones. The following models were the decision tree and k nearest neighbors (Goyal 2014). The linear regression and support vector machine models performed the weakest. In their case, the RMSE is almost three times that of the random forest model. The LightGBM model was used and tested for the dataset analysis. A k-fold cross-validation with k=5 was used when learning the model.

In the case of an online campaign via social media, additional factors were taken into account that could influence interest in the online campaign. These are factors from three categories:

- weather data, e.g., temperature, precipitation, fog,
- data on social and cultural events, e.g., sports championships, elections, social campaigns,
- COVID-19 pandemic data, e.g. residency restrictions, number of new cases (Hale 2021, Roser 2022).

Figure 8 visualizes the divisions in the decision tree model for the Facebook set. The decision tree is the only one of the models used that can be visualized in such a simple way. For the other models, we will only be able to assess the validity of the variables (Loh, 2012). The validity of the variables in the Facebook campaign's decision tree model is shown below. The most relevant variables are CPC, displays (impressions), and account_name. Other essential variables are exit obstructions, number of new cases, average daily dewpoint (baritone), average pressure(SLP), and overall strength of obstructions (stringency index).

Conclusions

This paper comprehensively analyzes marketing campaigns by comparing traditional and online methods. It explores the effectiveness of various marketing strategies using data-driven analysis. Traditional campaigns are evaluated through sales metrics, while online campaigns use additional parameters like clicks, impressions, and engagement. The paper emphasizes integrating analytics in understanding campaign performance and improving future marketing strategies.

Traditional campaigns focus mainly on sales data, whereas online campaigns offer detailed insights through metrics like clicks, impressions, and engagement, providing a broader view of campaign effectiveness. Both traditional and online campaigns require extensive data cleaning and analysis. Online campaigns often need more data, while conventional campaigns have limited data availability. Applying various predictive models like decision trees, random forests, and regression models helps forecast campaign success, offering insights for better strategic planning. The most significant factors affecting campaign performance include cost per click (CPC), impressions, and user behavior data, which vary significantly between traditional and online campaigns. Factors like weather conditions, social and cultural events, and health crises like COVID-19 significantly impact online campaign performance. The research underscores the need for a comprehensive analysis of marketing campaigns, combining data from traditional and online sources to make informed decisions and enhance marketing strategy.

References

Beretta, L., Santaniello, A. (2016), Nearest neighbor imputation algorithms: a critical evaluation, *BMC Med. Inform. Decis. Mak*., t. 16, nr S3, s. 74.

Borysiak, O., Wołowiec, T., Gliszczyński, G., Brych, V, Dluhopolskyi, O. (2022). Smart Transition to Climate Management of the Green Energy Transmission Chain, *Sustainability*, 14, s. 11449.

Cutler, A., Cutler, D.R., Stevens, J.R. (2012). Random Forests. In: Zhang, C., Ma, Y. (eds) *Ensemble Machine Learning. Springer*, New York, NY.

Goyal, R., Chandra, P., Singh, Y. (2014). Suitability of KNN Regression in the Development of Interaction Based Software Fault Prediction Models, *IERI P*rocedia, t. 6, s. 15–21.

Hale, T., et al. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker, *Nat. Hum. Behav*., t. 5, nr 4, s. 529–538,

Koronacki, J., Ćwik, J. (2008). Statystyczne systemy uczące się, wydanie drugie. Warszawa: *Exit*.

Loh, W. (2011). Classification and regression trees WIREs Data Min. Knowl. Discov., t. 1, nr 1, s. 14–23.

Patel, E., Kushwaha, D. S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model *Procedia Comput. Sci.* 171, s. 158–167.

Rachwał, A., Popławska, E., Gorgol, I., Cieplak, T., Pliszczuk, D., Skowron, Ł., Rymarczyk, T. (2023). Determining the Quality of a Dataset in Clustering Terms *Applied Sciences* vol. 13, nr 5, s. 1-20.

Roser, M.(2022). Our World In Data, W*hat is the COVID-19 Stringency Index?* https://ourworldindata.org/metrics-explained-covid19-stringency-index

Sarstedt, M., Mooi, E. (2014). Regression Analysis w A Concise Guide to Market Research, Berlin, Heidelberg: Springer Berlin Heidelberg, s. 193–233.

Yuan, C., Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm *J*, t.2, nr 2, s. 226–235.

Zheng, C.-H., Yuan, L., Sha, W., Sun, Z.-L. (2014). Gene differential coexpression analysis based on weight correlation and maximum clique. *BMC Med. Inform. Decis. Mak*., t. 15, nr S3.