Credit Risk Management Prediction Using the Support Vector Machine (SVM) Algorithm

Iwan Setiawan[1], Evi Martaseli[2], Tugiman[3], Nizirwan Anwar[4], Mirfan[5], Panji Kuncoro Hadi[6(✉)], Imam Suhrawardi[7], and Hendry Gunawan[4]

[1] Nusa Putra University, Sukabumi 43152, Indonesia
iwansa@nusaputra.ac.id
[2] Muhammadiyah University, Sukabumi 43113, Indonesia
evimartaseli@ummi.ac.id
[3] Buddhi Dharma University, Banten 15115, Indonesia
[4] Esa Unggul University, Jakarta 11510, Indonesia
nizirwan.anwar@esaunggul.ac.id
[5] Handayani University, Makassar 90231, Indonesia
[6] PGRI University, Madiun 63118, Indonesia
[7] Politeknik Negeri Lampung, Lampung 35141, Indonesia

Abstract. Information sharing throughout the globe or universe has become a characteristic of social media. There has been a lot of research into the classification of sentiments. In this study, Twitter has been mined for unstructured Gofood Reviews data. It has been preprocessed to analyze the reviews' sentiment with polarity analysis, feature extraction with TF-IDF, and supervised learning with random forest. From June 1, 2022 to June 30, 2022, a total of 28763 tweets with the keyword "Gofood" were retrieved from Twitter. The data is processed by the Python programming language utilizing NLTK, Sastrawi for the Indonesian language, Textblob, TF-IDF, Random Forest Classification, and other algorithms. Twitter is a nearly limitless source for classifying text. This algorithm takes roughly five minutes to computer.

Keywords: Social-Media · Data Twitter · Random Forest Algorithm · Sentiment Classification · Polarity Analysis

1 Introduction

Loans are often needed by the community for various reasons. Some people take out loans to meet basic needs, start a business, or grow an ongoing business. Many companies or institutions provide loan facilities. However, banks that provide loans often experience problems with consumers who make loans, namely the inability to complete and repay loans with an agreed nominal value. As a result, the bank must work hard to collect so that all debts to it can be resolved. Due to the frequent repetition of this incident, finally, by utilizing technology, a way was found so that the incident could be anticipated, namely by predicting credit risk. Credit risk is a method to find out whether the borrower is a

I G. P. Suta Wijaya et al. (Eds.): MIMSE-I-C 2022, ACSR 102, pp. 195–206, 2022.
https://doi.org/10.2991/978-94-6463-084-8_18

good (good) or bad (bad) borrower, which if the borrower is a bad borrower, the bank can take anticipatory steps so that the loan that has been borrowed can still be resolved [1].

Data mining is done to get hidden information from the amount of data that is processed [1, 2]. The results of the data mining process will then be made machine learning models to be able to predict the data that has been collected. There are many algorithms that can be used to predict this credit risk case, one of which is SVM or Support Vector Machine [3, 4]. SVM is a classification algorithm by finding the optimal point (hyperplane) between classes [5]. One of the studies that used SVM as a classification algorithm showed good performance in classifying heart disease [4].

2 Methodology

2.1 Credit Risk

Credit risk is one of the important risks faced by banks [1]. There are several things that cause credit risk to occur, one of which is that it is too easy for banks to provide loans or invest in third parties without good supervision. The main cause of credit risk is that it is too easy for banks to provide loans and make investments because they are required to immediately take advantage of excess liquidity, so credit assessments are less accurate in anticipating various possible business risks financed by banks. Agency theory explains that debtors as third parties often ignore the interests of creditors (banks) in managing funds lent or invested by banks. The existence of debtor behavior that is detrimental and poses such risks causes banks to tend to be more careful in channeling loans/investments to customers/debtors. Banks tend to be more careful in channeling funds in terms of credit and investment when there is an increase in non-performing loans as reflected in the high ratio of Non-Performing Loans (NPL). The high and low NPL ratio reflects the size of the credit risk in banks.

The results of previous studies show that NPL has a negative effect on loans disbursed, which means that as NPL increases, lending decreases. This explains that the higher credit risk in banks can affect the amount of loans provided by banks to debtors (customers). This decrease in the number of loans granted can indirectly reduce the income on loans disbursed, so that it can reduce the level of bank profitability. Many non-performing loans can increase banking costs and then reduce bank profitability. Therefore, high credit risk can weaken the influence of credit on bank profitability. This is because credit risk can affect the level of prudence of banks in distributing loans to avoid non-performing loans, so that it can affect the profits earned and the level of bank profitability [6, 7].

On the other hand, credit risk indirectly encourages banks to always optimize intellectual capital capabilities so as not to cause large losses due to debtor behavior that is detrimental to the bank. The risks faced by banks have a positive relationship to intellectual capital. The existence of risk makes all elements in the bank's organization feel responsible for the risks faced, so that a formal structure is formed at the bank which has the duties and responsibilities to oversee the operation of the risk management system in the bank. Risk management, especially credit risk, has a positive effect on profitability. There is a relationship between credit risk management and bank profitability, where the more efficient and effective risk management will result in higher profits for the bank.

Based on the results of previous studies and the logical thinking that has been explained, the hypotheses that can be formulated in this study are as follows: H3: Credit risk weakens the effect of loans on profitability, and H4: Credit risk strengthens the influence of intellectual capital on profitability.

2.2 Python Programming Language

Python is an interpreted, high-level general-purpose programming language Python was developed by Guido van Rossum and originally made available in 1991; it designs philosophy places a significant emphasis just on readability of its own source code. Its object-oriented methodology and language design are intended to assist programmers in creating logical and understandable code for both little and big projects [8, 9]. Python has garbage collection and dynamic typing. It supports a number of programming paradigms, including functional, object-oriented, and structured (mostly procedural) programming. Python's extensive standard library has led to it being referred to as a "battery included" language. Python was developed as a replacement in the late 1980s. Python 2.0, which has been released in 2000, to the ABC language. Introduced features such as list comprehension and a garbage collection system with reference counting. Python 3.0, which was introduced in 2008, is a significant update to the language that is not entirely backwards compatible with Python 2. This means that a lot of Python 2 code cannot be run with Python 3 without changes. For many operating systems, there are Python translators available. The free and opensource reference implementation CPython is created and maintained by a large community of programmers worldwide. The Python Software Foundation, a non-profit, oversees and administers resources for Python and CPython development [10, 11].

2.3 Service Vector Machine (SVM)

Support Vector Machine (SVM) is one of the supervised learning algorithms for classification and regression analysis. The output of this algorithm is the optimal hyper-plane and maximizes the level between the two classes. The image below illustrates how SVM works (Fig. 1).

In the picture above, it is shown that two classes are represented by a red box for class -1 and a blue circle for class $+1$. In SVM, classification will be carried out by looking for the optimal dividing line (hyper-plane) between class -1 and class $+1$. The optimal hyper-plane can be found by measuring the margin between the two classes. In helping this, SVM has several kernels that can increase the accuracy obtained. In linear SVM, the separator is a linear function. The training data is represented by (x_i, y_i) and $x_i = \{x_1, x_2, ..., x_{iq}\}$ is an attribute (feature) set for the i-th training data. For $y_i \in \{-1, +1\}$ declares the class label. The definition of the equation of a separating hyperplane is written as:

$$w \cdot x_i + b = 0 \tag{1}$$

The data x_i which is divided into two classes, which belongs to class 1 (negative sample) is defined as a vector which satisfies the following inequality:

$$w \cdot x_i + b < 0 \; for \; y_i = -1$$

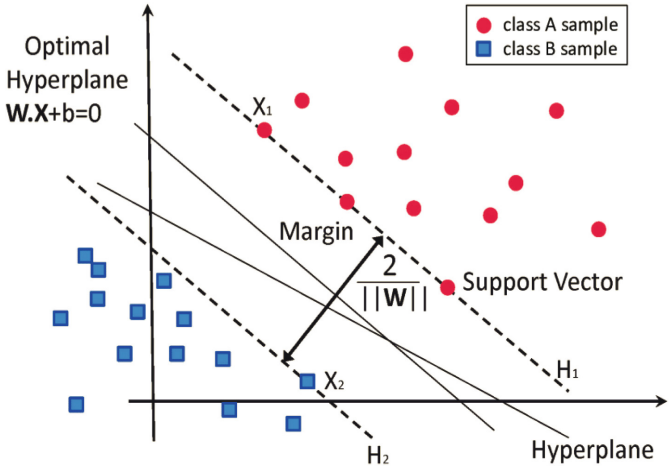

Fig. 1. How the SVM works which form a hyperplane as a separator between class −1 and class 1 [12]

Meanwhile, those belonging to class +1 (positive sample) meet the following inequalities:

$$w \cdot x_i + b < 0 \text{ for } y_i = +1$$

where:

x_i input data
y_i given label
w value of normal field
b position of x_i he plane relative to the center of the coordinate

Parameters w and b are the parameters to look for the value. If the data label is $y_i = -1$, then the delimiter becomes the following equation:

$$w \cdot x_i + b \leq -1 \tag{2}$$

If the data label is $y_i = +1$, then the delimiter becomes the following equation:

$$w \cdot x_i + b \geq +1 \tag{3}$$

The largest margin can be found by maximizing the distance between the boundary planes of the two classes and their closest point, which is $\frac{2}{|w|}$. . This is formulated as a Quadratic Programming (QP) problem, in which the goal is to identify the equation's minimum point by taking into account the following equation:

$$L(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i(y_i\left(\left(w^T x_i + b\right) - 1\right)\right) \tag{4}$$

$$L(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i y_i \left(w^T x_i + b\right) + \sum_{i=1}^{n} a_i \tag{5}$$

where a_i is the lagrange multiplier which is zero or positive ($a_i \geq 0$). The optimal value from the previous equation can be calculated by minimizing L to w, b and a can be seen in the following equation:

$$\frac{\partial L}{\partial a} = \sum_{i=1}^{n} a_i y_i \left(w^T x_i + b\right) - \sum_{i=1}^{n} a_i = 0 \tag{6}$$

2.4 Kernel Trick

SVM can be extended to draw a non-linear decision boundary by transforming the input from the original space to a high-dimensional space. Since the relationship between the input space and the transformation space is non-linear, the goal is to obtain a non-linear decision boundary [13, 14]. To improve the accuracy of the problem, SVM has a kernel trick that can help solve the problem of changing data into non-linear space. In general, some of the most commonly used kernel functions in SVM are as follows [14, 13]:

$$\text{Linear Kernel} \quad K\left(X_i, X_j\right) = X_i^T . X_j$$

$$\text{Polynomial Kernel} \quad K(X_i, X_j) = (X_i . X_j + 1)^h$$

$$\text{RBF} \quad K(X_i, X_j) = e^{-||X_i - X_j||^2 / 2\sigma^2}$$

$$\text{Sigmoid} \quad (X_i, X_j) = tanh(kX_i . X_j - \delta)$$

2.5 Python Programming Language

A identification or classification system must be capable to correctly classify all data sets in order to operate. However, it is impossible to disputed that a system's performance, that performs classification will not always be 100% correct. Therefore, the system performance must be measured. Generally, the way to measure classification performance is using a confusion matrix [13, 15]. Measurement of the effectiveness of a classification system is crucial. The system's performance is described using a performance classification system. in classifying data. One technique for evaluating a classification method's performance is the confusion matrix. The confusion matrix essentially provides data that contrasts the system's classification results with those of the classification that should be [13]. There are 4 (four) terms used to represent the outcomes of the categorization process when performance is measured using a confusion matrix. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative are the four terms (FN). The number of negative data that are accurately recognized is known as the True Negative

(TN) value, while false positives (FP) are instances of negative data that are mistakenly identified as positive. True Positive (TP), however, refers to positive data that has been accurately detected. The reverse of true positive, false negative (FN), refers to data that is positive but is mistakenly identified as negative. Accuracy values can be determined based on the values of True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). How accurately the system can correctly classify the data is shown by the accuracy value. In other words, the accuracy value compares the correctly classified data to the entire set of data. The precision value can be obtained by the equation below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Precision. It gauges true positive predictions. Equation 7 is used to calculate a model's precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

A sensitivity measure is this one. It is developed to evaluate how well a model predicts positive labels. It is determined by applying Eq. 8:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

F-Measure. Recall and precision (Eq. 9) are also considered in this metric. It can be thought of as a weighted average of recall and precision metrics, with values ranging from 0 (worst) to 1 (best). Equation 10 is used to calculate the F-measure.

$$\text{F} - \text{Measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

3 Result and Discussion

In processing data before making predictions, the data needs to be cleaned first so as not to damage the prediction results. One example of the need for data cleaning is if the data used has a Nan or Null value. Therefore, it is necessary to check to make sure this does not happen. The following are the results of checking the data used (Figs. 2 and 3).

From the results of checking, it can be ascertained that the data owned does not have a null value. That is, the data can be used for further processing. In addition to ensuring that the data does not have a null value, the data must also be balanced against the classes it has. It should be avoided if one class has more data quantity than the other class. In this study, it is necessary to balance the data in the data held so as not to damage the prediction results. As a result of discovering a class imbalance in the dataset, it is necessary to perform data reduction to rebalance the data and improve the accuracy of the resulting prediction model.

From the graph of the relationship between the home ownership column and the term against the label, it is certain that for home ownership there are more borrowers

```
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   loan_amnt       239310 non-null  int64
 1   funded_amnt     239310 non-null  int64
 2   term            239310 non-null  object
 3   home_ownership  239310 non-null  object
 4   annual_inc      239310 non-null  float64
 5   delinq_2yrs     239310 non-null  float64
 6   open_acc        239310 non-null  float64
 7   total_acc       239310 non-null  float64
 8   out_prncp       239310 non-null  float64
 9   total_pymnt     239310 non-null  float64
 10  acc_now_delinq  239310 non-null  float64
 11  credit_risk     239310 non-null  object
dtypes: float64(7), int64(2), object(3)
memory usage: 23.7+ MB
```

Fig. 2. Null-value check

```
0    12096
1    12096
Name: credit_risk, dtype: int64
```

Fig. 3. Data Balancing

who have pawn status who do not have credit difficulties. Uniquely, people who own houses with ownership status and more are at risk of non-performing loans. As for the term, more people are at risk of credit problems when taking a period of sixty months.

Of the two hundred thousand data used, the results of data balancing produce twenty thousand remaining data. That is, the data before the balancing process is very unbalanced with each other. The next process is to do EDA or Exploratory Data Analysis [16, 17]. The data analysis is carried out first by paying attention to categorical data on labels (Figs. 4 and 5). From the graph of the relationship between the home ownership column and the term against the label, it is confirmed that for home ownership there are more borrowers who have homes with mortgage status that do not have the risk of credit problems. Uniquely, people who own houses with ownership status and rent more are at risk of non-performing loans. As for terms, more people are at risk of having credit problems when taking a period of sixty months. While the results of visualizing the relationship between numerical data to the label, do not show any significant results or relationship between the two. The outliers contained in the graph are also not anomalies that can affect the model, but the shape of the data is indeed very diverse. Before carrying out the modeling process, it is necessary to understand the algorithm used requires different forms of data. For SVM, numerical data is needed because optimal hyperplane lines will be searched between classes. Therefore, it is necessary to convert categorical data into numerical form. In this process, a one-hot encoding process is carried out to convert it. The following are the results of the one-hot encoding performed (Fig. 6).

After processing the data so that the processed data can be optimally used as a model, the next step is to make the data as a model. The modeling process is carried out by utilizing the Sci-kit Learn library which has the SVM algorithm. From the processed data, it needs to be further divided into test data and training data. The data sharing process uses a ratio of 80:20 where 80% of the data will be used as training data, while

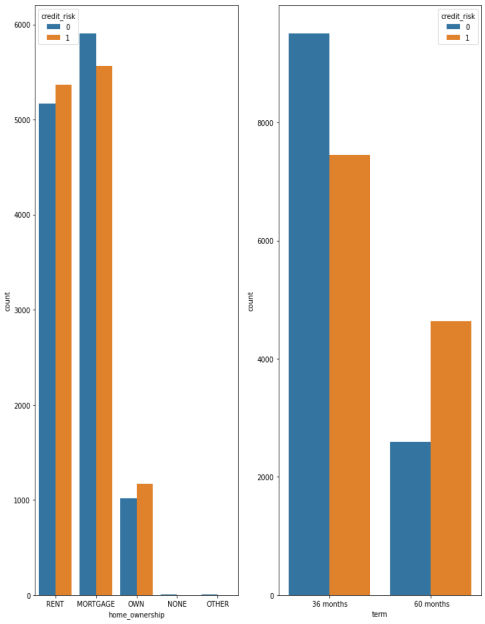

Fig. 4. EDA on Categorical Data

the remaining 20% will be used as test data. The following is the model fit process carried out.

Figure 7, shows the hyperparameters used in the SVM linear algorithm modeling process. The following table describes the confusion matrix of the predictive model used in this study. From Table 1, it can be concluded that the model can make accurate predictions, with only a few errors in predicting the negative class, where the model identifies the eight existing data as positive, not negative, classes. As for the AUC graph, the following are the model's predictions.

From Table 1, it can be concluded that the model can make accurate predictions, with only a few errors in predicting the negative class, where the model identifies the eight existing data as positive, not negative, classes. As for the AUC graph, the following are the model's predictions (Fig. 8).

4 Conclusion

From the results of the research that has been done, it can be concluded that the data that will be used as a model needs to be pre-processed to adjust the shape of the data to the algorithm used and so that it does not have noise or data that interferes with the modeling

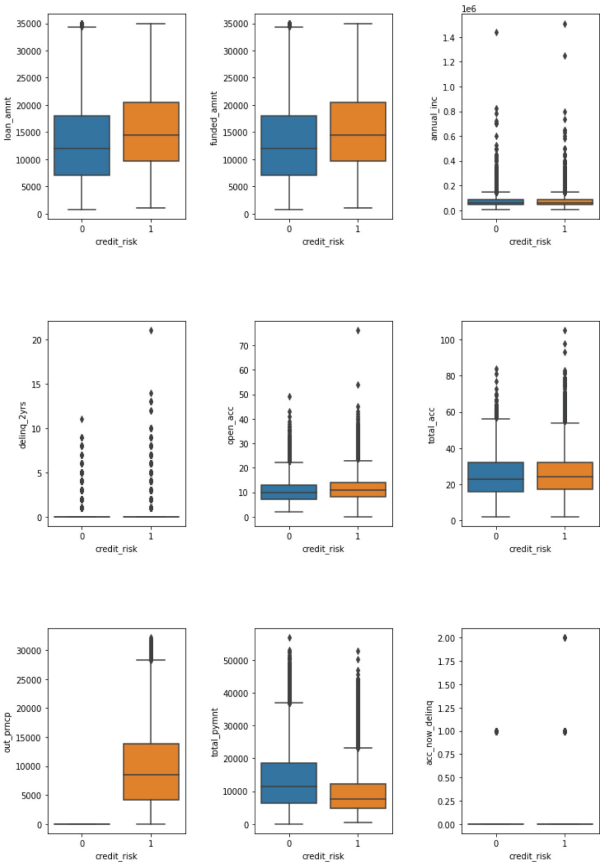

Fig. 5. EDA on Numerical Data

process, the SVM algorithm can be used as a model. Credit risk classification algorithm, the results of the SVM algorithm performance obtained are quite good by getting results above 90% for training data and test data. For further research, further experiments can be carried out by performing hyper-parameter tuning of the SVM algorithm, can try using other machine learning algorithms such as decision trees, KNN, and others, and can try using neural-network models.

```
One-hot encoding home_ownership:
 [[0. 1. 1. ... 0. 1. 0.]
 [1. 0. 1. ... 0. 0. 1.]
 [0. 1. 1. ... 0. 1. 0.]
 ...
 [0. 1. 1. ... 0. 1. 0.]
 [1. 0. 1. ... 0. 0. 1.]
 [1. 0. 1. ... 1. 1. 0.]]

One-hot encoding term:
 [[0. 1. 1. 0.]
 [1. 0. 0. 1.]
 [1. 0. 0. 1.]
 ...
 [1. 0. 0. 1.]
 [0. 1. 1. 0.]
 [0. 1. 1. 0.]]
```

Fig. 6. One-Hot Encoding

```
classifier = SVC(kernel = 'linear', random_state = 0)
classifier.fit(X_train, y_train)
```

Fig. 7. Model Fit

Table 1. Confusion Matrix model

		Class	
		Positive	Negative
Target	Positive	2433	0
	Negative	8	2398

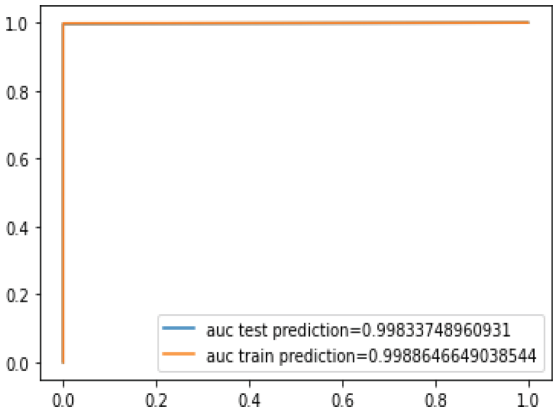

Fig. 8. AUC from Model

References

1. C. Nalini and T. Poovozhi: Data Mining Classification Technique Applied for Breast Cancer. Journal of Pure and Applied Mathematics 119(12), 10935–10945 (2018).
2. A. M. Widodo, N. Anwar, B. Irawan, and L. Meria: Data Mining Classification for Breast Cancer Prediction. Procedia Computer, (2019).
3. S. Amarappa and S. V. Sathyanarayana: Data classification using Support vector Machine (SVM), a simplified approach. International Journal Electron. Comput. Sci. Eng. 3, 435–445 (2011).
4. S. K. Sunori, D. K. Singh, A. Mittal, S. Maurya, U. Mamodiya, and P. K. Juneja.: Rainfall Classification using Support Vector Machine. In: Proc. 5th Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), pp. 433–437. I-SMAC 2021 (2021).
5. J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez: A comprehensive survey on support vector machine classification: Applications, challenges and trends. Journal Neurocomputin 408, 189–215 (2020).
6. A. Imani and A. Pracoyo: Analysis of The Effect of Capital, Credit Risk, and Liquidity Risk on Profitability in Banks. Journal Ilmu Manajemen Ekonomi 10(Juni), 44–50 (2018).
7. E. L. E. M. S. A. F. Rahman: The Effect of Loan and Intellectual Capital on Profitability with Credit Risk as Moderating. Journal Economy 15(2), 159–171 (2019).
8. J. R. Payne: Introduction to Computer Programming and Python. Python for Teenagers. Apress, (2019).
9. O. Embarak: Introduction to Data Science with Python. Data Analysis and Visualization Using Python. Apress, (2018).
10. V. Thangarajah: Python current trend applications-an overview. International Journal of Advance Engineering and Research Development 6(10), 6–12 (2019).
11. M. M. M. Fareez, V. Thangarajah, and S. Saabith: POPULAR PYTHON LIBRARIES AND THEIR APPLICATION DOMAINS. International Journal of Advance Engineering and Research Development 7(11), 18–26 (2020).
12. H. R. Baghaee, D. Mlakic, S. Nikolovski, and T. Dragicevic: Support Vector Machine-Based Islanding and Grid Fault Detection in Active Distribution Networks. IEEE Journal of Emerging and Selected Topics in Power Electronics 8(3), 2385–2403 (2020).
13. S. Visa, B. Ramsay, A. Ralescu, and E. Van Der Knaap: Confusion matrix-based feature selection. CEUR Workshop Proc., vol. 710, pp. 120–127 (2011).

14. K. D. Yonatha Wijaya and A. A. I. N. E. Karyawati: The Effects of Different Kernels in SVM Sentiment Analysis on Mass Social Distancing. Journal Elektron. Ilmu Komput. Udayana 9(2), 161 (2020).
15. I. Düntsch and G. Gediga: Confusion Matrices and Rough Set Data Analysis. Journal Phys. Conf. Ser. 1229(1), 12055 (2019).
16. C. Nicodemo and A. Satorra: Exploratory data analysis on large data sets: The example of salary variation in Spanish Social Security Data. Journal Bussiness Research Quarterly 25(3), 283–294 (2020).
17. C. Chatfield: Exploratory data analysis. Eur. Journal Oper. Res. 23(1), 5–13 (1986).