electronics

MDPI

Article

Data Mining to Identify Anomalies in Public Procurement Rating Parameters

Yeferson Torres-Berru [1,2,3,*] and Vivian F. Lopez Batista [1]

[1] Department of Computer Science and Automatics, University of Salamanca, 37008 Salamanca, Spain; vivian@usal.es
[2] Departamento de Investigación, Instituto Tecnologico Superior Sudamericano, Loja 1101608, Ecuador
[3] Escuela de Ingeniería en Tecnologías de la Información, Universidad Internacional del Ecuador, Loja 1101608, Ecuador
* Correspondence: yeferson.torres11@gmail.com

Abstract: The awarding of public procurement processes is one of the main causes of corruption in governments, due to the fact that in many cases, contracts are awarded to previously agreed suppliers (favouritism); for this selection, the qualification parameters of a process play a fundamental role, seeing as due to their manipulation, bidders with high prices win, causing prejudice to the state. This study identifies processes with anomalies and generates a model for detecting possible corruption in the assignment of process qualification parameters in public procurement. A multi-phase model was used (the identification of anomalies and generation of the detection model), which uses different algorithms, such as *clustering* (K-Means), Self-Organizing map (SOM), Support Vector Machine (SVM) and Principal Component Analysis (PCA). SOM was used to determine the level of influence of each rating parameter, K-Means to create groups by clustering, semi-supervised learning with SVM and PCA to generate a model to detect anomalies in the processes. By means of a case study, four groups of processes were obtained, highlighting the presence of the group "null economic offer" where the values for the economic offer do not exceed 1%, and a greater weight is given to other qualification parameters, which include direct contracting. The processes in this cluster are considered anomalous. Following this methodology, a semi-supervised learning model is built for the detection of anomalies, which obtains an accuracy of 95%, allowing the detection of procedures where the aim is to benefit a particular supplier by means of the qualification assignment parameters.

Keywords: corruption; public procurement; self-organizing map; support vector machine; machine learning; data mining

check for updates

Citation: Torres-Berru, Y.; López Batista, V.F. Data Mining to Identify Anomalies in Public Procurement Rating Parameters. *Electronics* **2021**, *10*, 2873. https://doi.org/10.3390/electronics10222873

Academic Editors: Amir Mosavi

Received: 15 October 2021
Accepted: 19 November 2021
Published: 22 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Favouritism in the state sector is the natural human propensity to privilege friends, relatives and any close and trustworthy person in a public procurement process [1]. When a public purchase is made to favour an entity or company with preliminary agreement with the contracting entity, the bidder with the best offer is not being awarded the contract; in this bad practice, usually in the winner's qualification parameters, lower scores are established than those required for the economic offer, therefore, in order to complete the remaining score, the contracting entity includes additional parameters which privilege a particular participant. In this sense, the economic offer is not the decisive parameter; instead, new technical parameters are used, allowing the bidder with the higher price to win the process [2]; another bad practice is to focus the procedures on bidders who have previously worked with the institutions, requesting previous work experience, excluding new (inexperienced) bidders [3,4]. It is also usual to establish in the section "Other Parameters", specific conditions and requirements with high scores that only an agreed bidder can satisfy, ensuring the disqualification of the rest of the proposals; the "technical specifications" of the object of the tender are not in accordance with the needs and functions stipulated in the object of the contract, with the aim of directing the procedures to a supplier. Favouritism is also based on the characteristics of the staff that will be part of the project, and that only the agreed company possesses; thus, in the parameter "Compliance with specifications", a certain age, title, experience in a specific area are included, without a legal justification to support such requests. In other words, favouritism causes less purchasing power for the public institution, higher prices that have an impact on the quality of the product and generate unfair competition.

In Ecuador, the Public Procurement System (SERCOP) is in responsible for promoting access to and use of public information, increasing transparency, combating fraud and corruption that could originate from bad practices in public procurement. In 2017, 5.8 billion dollars were transacted through public procurement portal or 19.6% of the general state budget and 5.8% of the Gross Domestic Product (GDP). The participation by government sector was distributed mainly in state administration (28.5%), autonomous municipal governments (21.2%) and public agencies (18%). In 2019, public procurement accounted for 17% of the state's general government budget [5], also showing in Figure 1 a decrease in public investment from 2011 to date.

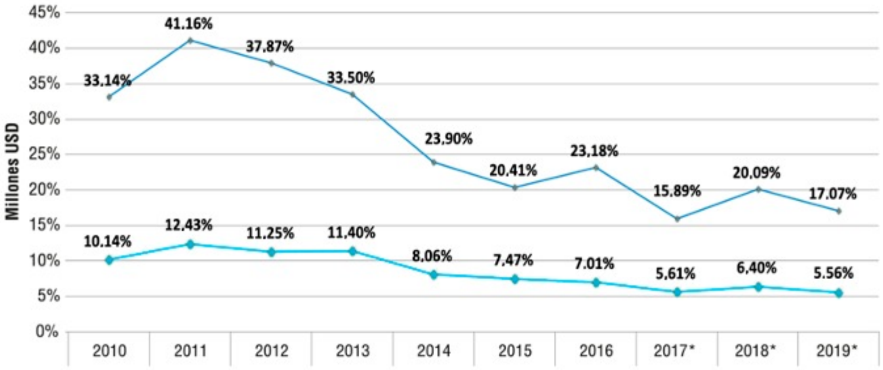


Figure 1. Percentage of public procurement of millons in relation to the General State Budget (public procurement portal Ecuador).

SERCOP hosts documents in PDF format for each contracting process, where data on the specifications are stored: Terms of Reference (TDR), invitations to suppliers, offers submitted, and observations, and in summary, all the documentation generated by the purchase. The types of procurement processes carried out by state entities and available in the SERCOP database are as follows:

- Execution of works.
- Purchase of products and services.
- Consultancy contracting

As part of the information for each process, in parallel to the qualification parameters, the following conditions are considered to evaluate the relationship of each purchase executed, and all the conditions (Table 1) identified are important and must comply with the execution of the contracting by the public entity; therefore, their importance is highlighted:

Table 1. Conditions considered of each process.

Condition	Description
Timeline of the procedure	It emphasises important dates in the process.
Duration of the offer	Item used to determine the number of days the process will remain in effect.
Purchase price	Is the price of the process (purchase), which the institution lists on the public procurement portal.
Type of purchase	The classification used by the institution for the purchase carried out can be: goods, consultancy, work, insurance and service.
Recruitment Types	It is the method used to contract the acquisition is classified in: bidding, quotation, special publication, short list and direct contracting.
Payment method	The forms of payment are: advance payment, remaining value of the contract and at the end of the contract.
Status of the process	Is the state in which the contracting process is currently running two general statuses are obtained: **correct** (to be awarded, awarded, finalised and in execution) and **not executed** (unilaterally terminated, terminated by mutual agreement, cancelled and deserted).

As a technological resource and with the objective of discovering favouritism, Data Mining (DM) has a fundamental role to contribute with its tools and methods to find hidden information in the massive volumes of data [6]. The use of this technique in public procurement is used as a critical tool, facilitating the monitoring of information, as well as the control of contracting processes [7]. Applying DM, it was established that in Sweden 58% of time the bidder who submits the lowest bid is not the winner of the process [8]; in Paraguay [9], using data from 4 years and 47,615 procurement processes, this study estimates, through the construction of a mathematical model, the correlation between the companies and their possibility of obtaining a contract, detecting the existence of a previous relationship between the supplier and the contracting entity, which produces corruption when the procurement is made. SALER [10] applying DM, analyses contracts and groups them by contract object, procurers, amount, number of contracts and total contract amount, determining characteristics of groups with corrupt practices and their relationship to a risk index for each process. The study conducted by Kehler [11] evaluates anomalies in public contracts using Isolation Forest algorithm [12] based on the modifications undergone by the contracts during the process to determine the corruption originated by these modifications to benefit a particular supplier.

With this background, the hypothesis is proposed: it is possible to develop a composed model to identify processes with anomalies in public procurement qualification parameters. The main objective of the work is to generate a model to identify patterns in the awarding of qualifications to public procurement contracts through the use of data mining techniques and then predict contracts where anomalies exist based on the reviewed data with the use of unsupervised learning techniques.

To simplify the reading of this document, after this introduction, Section 2 describes the data, models and techniques used; Section 3 presents the main results obtained, divided into two sub-sections: Section 3.2 related to unsupervised learning and Section 3.3 to semi-supervised learning, as techniques to validate the hypothesis. In the final part, the conclusions of the study are provided.

2. Methodology

After analysing the various approaches existing in the current literature on favouritism that attempt to provide an answer to the problem posed, this section details the proposal of the present work, designed to test the hypothesis based on the CRISP-DM methodology for data mining [13]. In the literature review, it was found that most of the published works use supervised learning, as contracts with price anomalies are labelled [14]. About 79% of the research corresponds to detection and 21% to prediction. This is not the case in Ecuador, which still lacks labelled data; therefore, in the initial phase of the research, we decided to use unsupervised learning techniques to detect anomalous patterns in contracts.

2.1. Data Set Description

As Ecuador's public procurement does not have an open data website, a web scraping technique was applied [15] on the data provided on the website of the SERCOP (https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/) (accessed on 11 November 2021). Through this technique, the information is obtained on public processes from 2010 until 2020, as well as the documents (attachments) of each process.

We approach our research through an experiment using publicly available datasets (https://bit.ly/PametersCorruption) (accessed on 11 November 2021), and the parameters for the qualification of bids were evaluated in 275,730 public procurement contracts in Ecuador. A total of 21 numeric parameters were assessed to determine the winner of each process, which is detailed in Table 2. The rating parameters vary according to the process, and they are all considered for the evaluation without excluding for the subsequent evaluation of the impact of each parameter on the final score, in addition to the fundamental aspects of the process such as the following: the type of purchase, status of the process and type of procurement.

Table 2. Parameters for qualification of an individual process.

Parameters	Description
General experience	Experience of the bidder in the general domain.
Specific experience	Experience in specific projects in the area of sourcing.
Similar works	Number of similar projects executed by the supplier.
Subcontracting	The supplier is able to partially subcontract the execution of the project.
Financial ratios	The solvency and debt ratios of the participating companies are assessed.
Methodology, Work Plan	Parameters for evaluating the bidder's presentation of the project.
Supply date	Estimated delivery date stated in the offer.
Economic offer	Value submitted by the bidder
Proposed team	Characteristics of the team that executes the work.
Inclusion parameters	They aim to include people and companies with disabilities.
Instruments equipment	Referring to models and brands of the products available.
Specification compliance	Technical product specifications and characteristics
Technical guarantee	Technical product guarantee.
National partnership	Priority is given to international suppliers who partner with local producers.
National SMEs	Priority to national micro-enterprises
Local participation	Priority to suppliers from the place of purchase
Ecuadorian participation	Priority to national companies
Bonus awarded by lottery	Bonus awarded by lot in case of a tie between bidders
Other qualification parameters	Defined by the procuring entity
Variable scoring	According to the requirements presented.
Technology transfer	Added value to processes that are born as a technology transfer from educational institutions.

2.2. Data Pre-Processing

For qualitative data review, it employs a technique proposed by Chu [16] which helps to find errors in the data and to scale or normalise them for use. Firstly, the data set is processed, eliminating erroneous values corresponding to processes with qualification parameters that had errors, as these must add up to a value equal to 100% and in some cases had lower values, such as 98% or bigger, such as 105%.

Using the pandas tool (https://pandas.pydata.org/) (accessed on 11 November 2021), the missing values are replaced, assigning 0 to the null fields, since the same qualification parameters are not met in all the processes. Finally, the data obtained are scaled. As this is unsupervised learning, it is decided to use the entropy measure for each attribute, so that more variability can be obtained. Therefore, the data for the 21 rating parameters are normalised.

2.3. Proposed System

The followed methodology and techniques are summarised in Figure 2. The process is initiated when data are collected from SERCOP, and once the data retrieved on public procurement are processed through web scraping, they are analysed through a multi-phase methodology, which uses different machine learning algorithms for the detection and prediction of favouritism in public procurement such as: *clustering* (K-Means), Self-Organizing Map (SOM), Support Vector Machine (SVM) and Principal Component Analysis (PCA). Following an analysis of different techniques for *clustering*, the K-Means algorithm was chosen [17] to group the data according to the type of recruitment. Leveraging the advantages of class visualisation provided by the SOM was used to identify the impact of every variable, and to compare the *clusters* obtained from K-Means based on distance and density, it can be used to analyse the data for possible *clusters* [18]. This allows for the identification of the clusters where the contracts with possible anomalies are located and is the input for the construction model.

Finally, with a semi-supervised learning model, anomaly detection is performed using PCA and SVM.

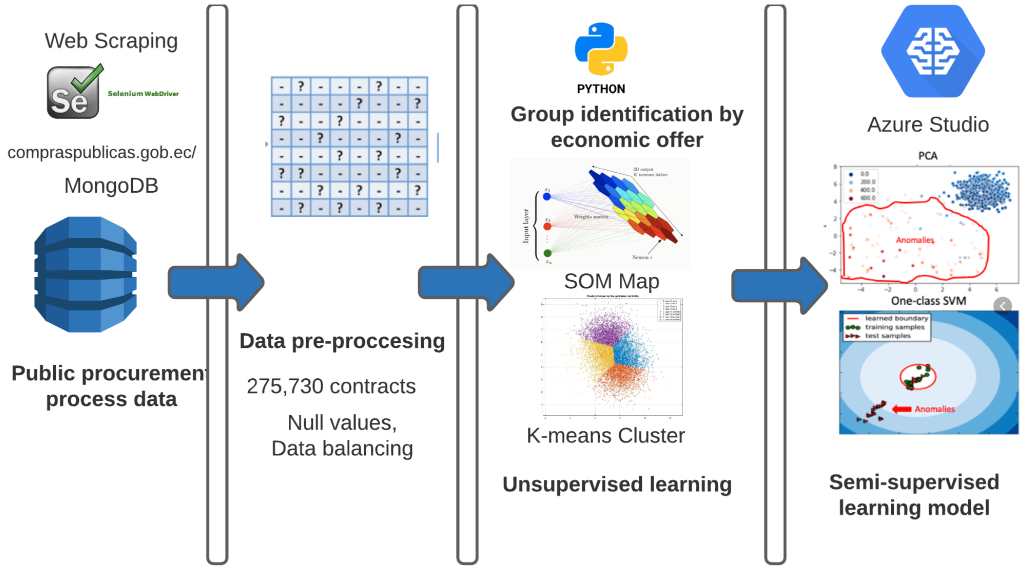


Figure 2. Diagram of the proposed methodology for find anomalies in qualification parameters.

2.4. Training and Learning Phase

The choice of learning algorithms is justified based on the number of data in the data set, the number of parameters to evaluate; starting from metrics such as the Clustering Accuracy (ACC) and the Normalised Mutual Information (NMI), based on the work of [19], it is understood that most of the unsupervised feature selection methods (filter, wrapper or hybrids) require the specification of hyper-parameters such as the number of features, number of clusters or other parameters inherent to the feature selection technique used by each method, and the quality of the feature extraction of data directly affects the detection performance of SVM. Describing the autocorrelation among data is an important factor that affects the fault detection performance [20]. The use of machine learning techniques to classify public procurement processes according to their qualification parameters and generate the detection model is described.

2.4.1. Self-Organizing Maps

In Table 2, as many as 21 parameters are evaluated to determine the winner of a public process, but these parameters are not repeated in all processes; therefore, it is necessary to determine the main parameters common in most processes, which is why the SOM maps [21] were chosen. A rectangular topology was implemented, consisting of 10 input rows and 10 input columns [18]. The Gaussian neighbor′ is selected, and the quality of the SOM map is influenced by the initial weights of the training map [17] we chose random. Finally, the number of training iterations is set to 1000, and finally, two types of metrics Quantification and topographic error were taken into consideration for the evaluation of SOM maps.

2.4.2. Clustering Algorithm

As described in Section 2.3, it is necessary to identify the processes with anomalies in the ratings, which is why clustering is used in combination with SOM maps. The K-Means clustering algorithm makes it possible to analyse data and find groups within that data using some kind of similarity measure, such as Euclidean distance. No one metric of universal similarity works for all cases [22] (depending on the problem itself). Therefore, starting at eight different centroids and using the elbow technique, the optimal number of clusters was determined ($k = 4$), and metrics such as ACC and NMI were evaluated. Once the cluster with anomalies was identified, semi-supervised learning was applied to detect anomalies in public procurement processes.

2.4.3. Support Vector Machine

SVM classifies the data, if the data are linearly separable, SVM classifies it linearly for the training and identification of anomalies with SVM, and the contracts of the groups where the economic offer has a greater weight in determining the winner are considered as normal (class 1), and data that are different can be predicted as anomalies (class 2).

When this version of the algorithm is applied, we use the property [23] *nu*, which allows us to control the balance among the outliers and normal cases, and therefore assigns nu = $[1e-3, 1e-2, 1e-1, 1]$, while the parameter affecting the number of iterations used, when optimising the model, is taken as $epsilon$ = $[1e-4, 1e-3, 1e-2]$. The optimal hyperplanes for machine learning are then determined using a *Hyper-parameters*,the model is trained and evaluated using the ROC and *accuracy* metrics. The values of the minimum and maximum metrics are [0.9, 0.97] equivalent to a very good test.

2.4.4. Principal Component Analysis

The accuracy of PCA-based anomaly detection depends on a good choice of principal components, which is achieved with the use of SOM Maps being the main characteristic for the choice of the algorithm. Distance metrics are applied to identify the cases that represent anomalies; therefore, they are used with a range of parameters (*rank*) and *oversampling* of [2, 4, 6, 8, 10]. Finally, the model is trained using the Score Model and ROC; for this method, 80% of the data is used for training and 20% for testing.

2.5. Tools

For the programming job, the Python language is used with the Selenium test environment. https://selenium-python.readthedocs.io (accessed on 1 November 2021) and Jupyter (https://jupyter.org/) (accessed on 1 November 2021), in addition to libraries Stick-learn (https://sklearn.org/) for the application of the machine learning algorithms and [24] the Python libraries Minisom https://pypi.org/project/MiniSom/ (accessed on 1 November 2021) and Sompy https://github.com/sevamoo/SOMPY (accessed on 1 November 2021) were used for Self-Organizing Maps .

It also uses the *machine learning* service provided by AZURE (https://studio.azureml.net) for training and testing data sets, due to the size of the data evaluated. It is assessed using the metrics: ROC curves, *accuracy*, *precision*, *FScore*n and *Recall*. The ROC curve shows the ratio between false positives and false negatives.

3. Experimental Results

To build the case study, information was retrieved considering the URL of the purchase process as input, fields such as: description, dates, products, qualification parameters, invitations, documents and questions from the suppliers. Each section was extracted according to its equivalent identification (tag) in HTML through scraping and stored in a non-relational database (MongoDB).

3.1. System Implementation

Figure 3 details the two main phases that composed the developed model, starting with the identification of contracts with anomalies using unsupervised learning with *K-Means* once the internal validation of the cluster was accomplished, and the following results are obtained: four groups, of which in in three, the economic offer is expected to determine the winner of the process and in one not; at the same time, the main parameters that have the greater influence on the determination of the winner of the process are evaluated with the use of *SOM maps*; therefore, two types of contracts are identified: regular contracts and contracts with anomalies.

With the identification of the groups and the influence of the variables on the rating, the following is required for the second phase of the model the detection of anomalies with the use of SVM and PCA; in the second phase, training is performed with the metrics described in the methodology to avoid overtraining, and the model is evaluated with data not present in the model (in this case, 2021 data). Therefore, it is suggested that the accuracy of the model obtained is between 85% and 97%.

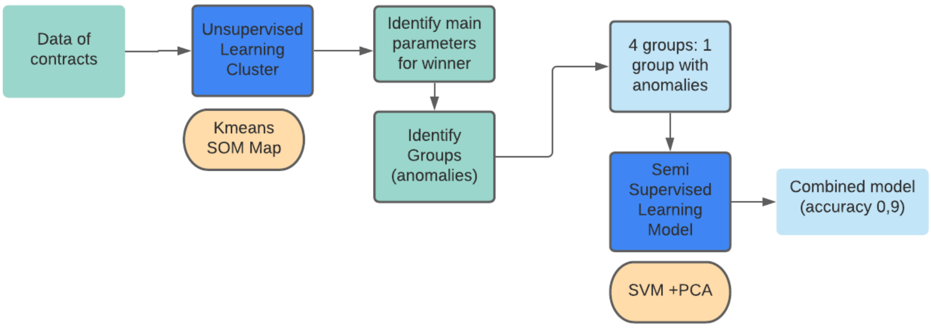

Data of contracts
Unsupervised Learning Cluster
Kmeans SOM Map
Identify main parameters for winner
Identify Groups (anomalies)
4 groups: 1 group with anomalies
Semi Supervised Learning Model
SVM +PCA
Combined model (accuracy 0,9)

Figure 3. Detailing two main phases that composed the model.

3.2. Unsupervised Learning Cluster

Using the SOM, the main parameters influencing the process rating and their influence on the cluster classification are identified.

Figure 4 shows the influence of each rating parameter on the cluster, with those in blue having the least influence and the colour scale representing the greatest influence; therefore, the main rating parameters found by using SOM Maps are: economic offer, specification compliance, other qualification parameters, general experience, specific experience, proposed team, technical guarantee, instruments and equipment and similar works.

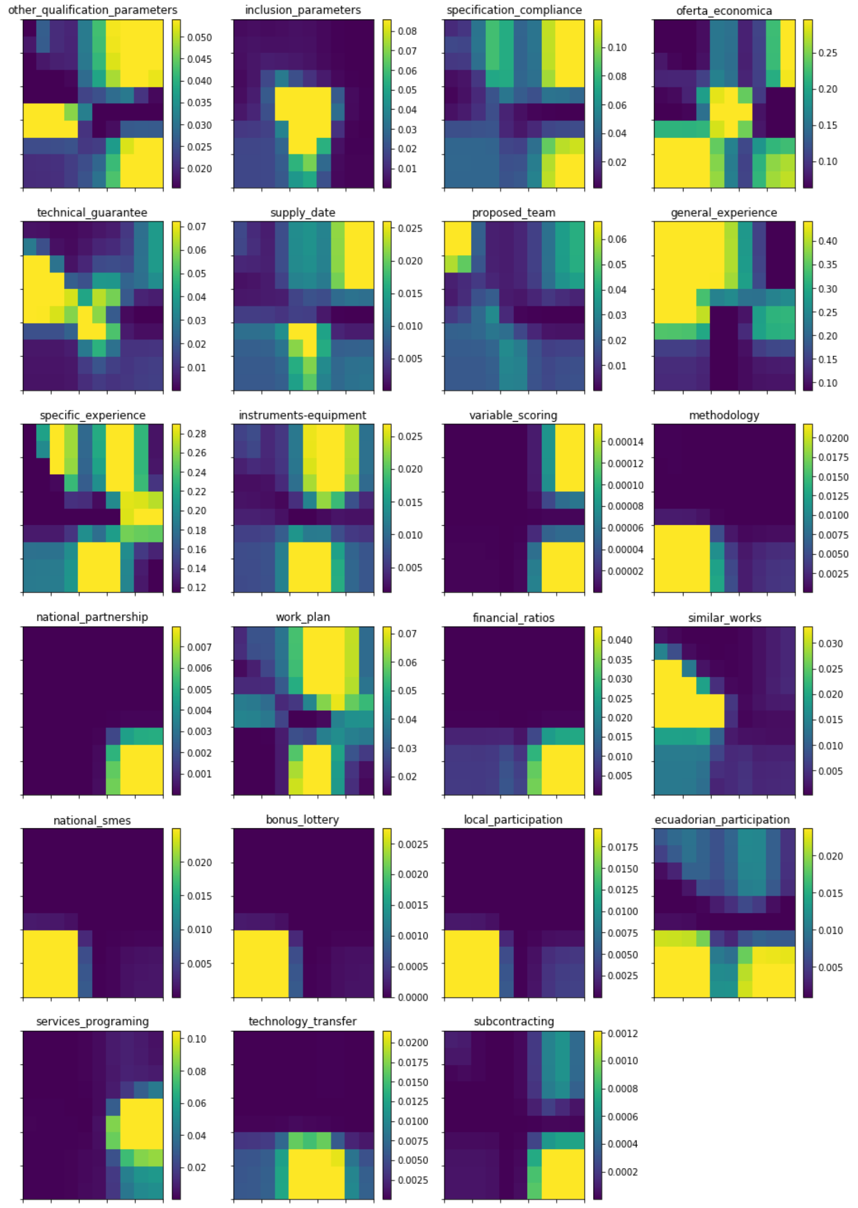
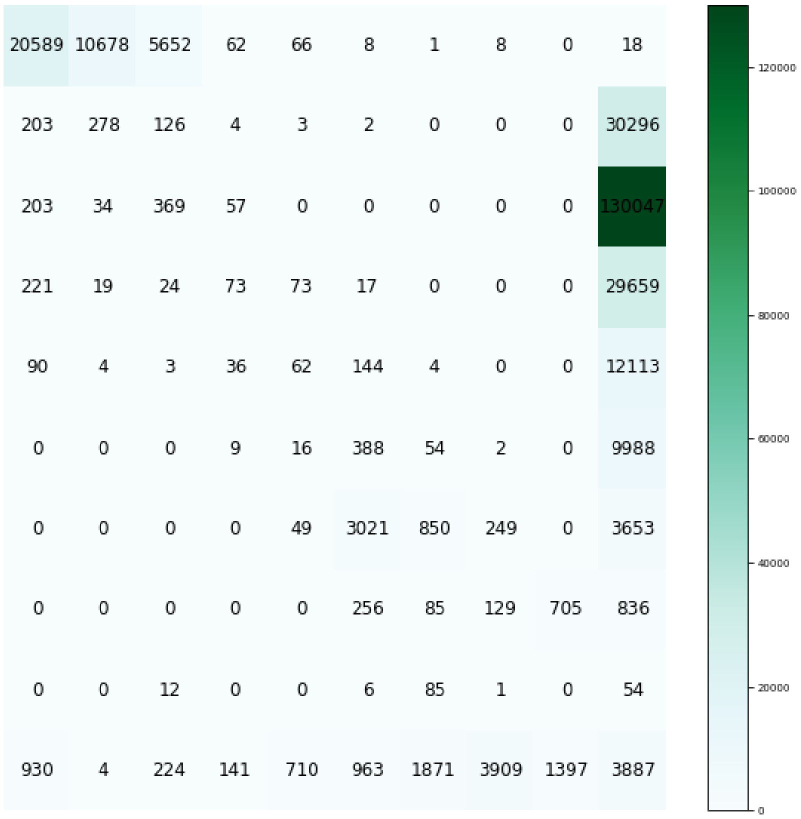


Figure 4. Evaluation of influence of qualification parameters with SOM.

A heat map (Figure 5) shows the assignment of the processes to each cluster represented in green, light blue, orange and red for each cluster, and the dark-blue values represent a small number of elements and are assigned to the nearest cluster. A colour scale from zero (withe) to 60,000 (dark green) represents the number of elements associated with the cluster.



20589	10678	5652	62	66	8	1	8	0	18
203	278	126	4	3	2	0	0	0	30296
203	34	369	57	0	0	0	0	0	130047
221	19	24	73	73	17	0	0	0	29659
90	4	3	36	62	144	4	0	0	12113
0	0	0	9	16	388	54	2	0	9988
0	0	0	0	49	3021	850	249	0	3653
0	0	0	0	0	256	85	129	705	836
0	0	12	0	0	6	85	1	0	54
930	4	224	141	710	963	1871	3909	1397	3887

Figure 5. Heat map of distances to individual clusters.

Figure 6 shows the evolution of the quantisation and topographic error with 1000 iterations, observing that from iteration 600 it stabilises and reaches optimal values for the model, obtaining a quantisation error of 0.2878 and a topographic error of 0.30796, ensuring in this way a correct reliability of the maps.

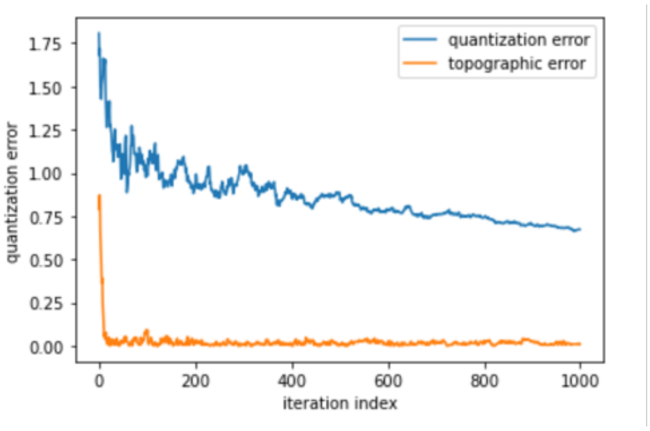


Figure 6. Iterations index.

By applying the K-Means algorithm with four centroids, four different clusters were obtained. Table 3 shows the 12 main characteristics associated with the variables related to the type of purchase. For example, *general experience* is predominant in the *cluster 3*, *specific experience* is predominant in the *cluster 4*, and other qualification parameters and specification compliance are predominant in *cluster 1*. The last row details the number of records (processes) belonging to each *cluster*.

Taking into consideration the state of the process, it can be classified as follows: correct or non-executed, the percentage of non-executed processes was 4.71% in cluster 1, 15.80%

in cluster 2, 39.93% in cluster 3 and 26.69% in cluster 4.0%. It is therefore determined that: in the cluster 1, the number of non-executed processes is under the average, and compliance with specifications and the economic offer have a greater influence. In cluster 2, the number of non-executed processes is equal to the average and is more influenced by the economic offer and an equal distribution among the other variables. The cluster 3 is below the average number of non-executed processes and is more influenced by overall experience and economic offer. Finally, at the cluster 4, the number of non-executed processes is above average, and general experience, specific experience and the work plan are more influential. This indicates that cluster 4 is the cluster with "anomalies".

Table 3. *Clusters* K-Means.

Parameter	1	2	3	4
Instruments-equipment	0.18%	0.21%	1.63%	4.94%
Specification compliance	46.66%	1.61%	5.22%	3.09%
Other qualification parameters	2.55%	7.16%	4.71%	5.61%
Specific experience	1.83%	10.73%	2.66%	51.03%
Similar works	0.27%	0.78%	1.09%	0.20%
General experience	2.12%	4.37%	46.46%	16.36%
Economic offer	33.99%	47.43%	18.63%	0.45%
Proposed team	2.68%	5.75%	2.93%	1.12%
Technical guarantee	2.22%	3.72%	3.42%	0.59%
Supply date	5.21%	5.85%	2.79%	0.50%
Methodology and work plan	0.17%	0.95%	6.00%	12.70%
Number of records	81,261	71,959	34,141	88,358

Figure 7 shows the influence of the six main qualification parameters, which are related to the economic offer. It can be seen graphically, the null participation of the Economic offer in cluster 4, a moderate involvement in the cluster 1, high participation in cluster 2 and weak participation in cluster 3. Therefore, for a better understanding for the reader, in the next sections, the clusters are renamed based on the influence of the economic offer and are as follows: Cluster 1 = Moderate economic offer, Cluster 2 = High economic offer, Cluster 3 = Low economic offer, Cluster 4 = Null economic offer

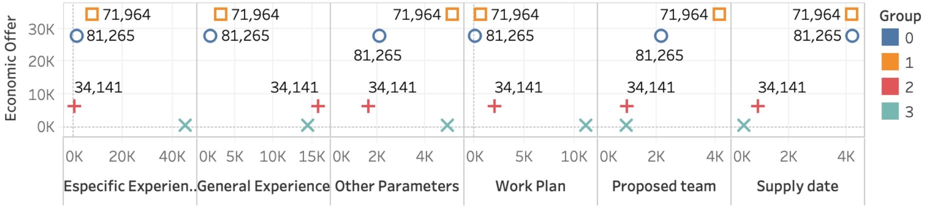


Figure 7. Relationship of the Economic offer to other qualifying parameters.

3.3. Cluster Analysis for Process Variables Not Involved in Purchasing Qualifications

The clusters obtained are matched with the type of purchase made and the type of procurement with which the process was performed.

With respect to the relationship between the rating parameters and the type of purchase made, Table 4 shows that the *"Moderate economic offer" cluster*, the Economic Offer rating is higher for the purchase of products and services, and as highlighted in the table, the compliance with technical specifications is higher for the purchase of services (specifications are usually given for products).

In the *"High economic offer" cluster* , the predominant procurement of products, services and works, with a high percentage is given to Economic offer in all processes; however, in works and services processes, a high value is given to experience between 15% and 10%, respectively, and in the procurement of services a value of 11% is assigned to other parameters. *"Low economic offer"* cluster purchase of services, products and consultancy predominates, in the respective order of the main qualification parameters, the General experience, Economic offer and to a lesser extent the specific experience. Finally, in the *"Null economic offer" cluster*, the purchase of consultancy and services predominates, with a high influence of the parameters of qualification of specific experience, general experience and compliance with specifications.

Table 4. *Clusters* related type of purchase.

Type of Purchase	**Equipment**	**Technical Guarantee**	**Work Plan**	**Supply Date**	**Proposed Team**	**Other Parameters**	**General Experience**	**Specification Compliance**	**Specific Experience**	**Economic Offer**	**Number of Process**
"Moderate Economic offer" Cluster											
Product	0.1%	2.7%	0.0%	5.7%	3.4%	2.0%	1.24%	46.8%	0.9%	35%	46,928
Consultancy	0.4%	1.1%	1.4%	4.6%	3.5%	2.7%	4.6%	47.3%	7.2%	23.4%	102
Work	0.2%	1.7%	0.1%	4.1%	7.6%	2.1%	2.5%	43.3%	3.6%	28.6%	388
Assurance	0.0%	0.4%	0.1%	0.2%	0.5%	8.6%	4.6%	34.2%	3.8%	33.2%	823
Service	0.2%	1.3%	0.3%	4.7%	1.5%	3.1%	3.3%	47.3%	3.0%	32.7%	29,187
"High Economic offer" Cluster											
Product	0.2%	8.9%	2.3%	11%	10.5%	8.9%	2.2%	1.79%	2.9%	46.8%	23,024
Consultancy	0.9%	1.6%	3.0%	10.2%	8.9%	16.0%	1.9%	1.02%	8.5%	37.6%	59
Work	0.0%	0.0%	0.0%	0.4%	2.0%	1.9%	5.5%	0.3%	17.7%	47.6%	14,814
Assurance	0.0%	0.2%	0.0%	0.0%	0.8%	11.8%	9.1%	2.4%	15.1%	50.4%	2,313
Service	0.3%	2.3%	0.5%	7.0%	5.7%	10.4%	4.0%	2.2%	10.5%	47.3%	20,373
"Low Economic offer" Cluster											
Product	0.2%	6.9%	0.3%	4.8%	5.0%	2.4%	45.5%	6.34%	0.4%	23.4%	9,932
Consultancy	5.4%	0.0%	22.5%	0.0%	0.0%	9.8%	48.5%	0.24%	8.2%	0.2%	5,799
Work	1.1%	1.3%	0.5%	4.3%	3.7%	1.3%	42.6%	7.58%	2.2%	23.8%	384
Assurance	0.0%	1.2%	0.0%	0.8%	2.0%	14.3%	32.0%	11.8%	4.4%	19.6%	72
Service	0.3%	3.0%	0.7%	2.0%	2.9%	3.4%	46.4%	7.10%	0.9%	25.7%	14,563
"Null Economic offer" Cluster											
Product	0.6%	4.6%	0.8%	3.1%	7 %	3.3%	9.9%	**17.8**%	48.0%	1.3%	4,866
Consultancy	5.7%	0.0%	15.0%	0.0%	0.0%	5.8%	18.0%	0.05%	51.7%	0.0%	49,523
Work	0.9%	0.8%	0.8%	1.6%	7.3%	1.9%	6.6%	7.10%	45.7%	11.5%	458
Assurance	0.0%	1.1%	0.0%	1.1%	2.9%	18.6%	10.3%	13.8%	37.6%	9.7%	34
Service	1.3%	2.7%	1.9%	2.6%	5.3%	5.4%	8.6%	17.0%	47.6%	2.3%	9,895

The type of procurement performed influences the qualification parameters in Table 5; therefore, we observe that in the cluster "Moderate Economic offer" special publication processes predominate with 93.8% of the total number of processes in this cluster and 47.35% impact of compliance with specifications as a qualification parameter, while in the cluster "High economic offer", the quotation and special publication processes predominate. In the cluster "Low Economic offer", we have only direct contracting and special publication processes, with the latter being predominant. Finally, cluster "Null Economic offer" contains direct contracting processes and special publication highlighting the influence of specific experience reaching up to 60% of the total qualification of the process.

Table 5. *Clusters* Type of purchase.

Recruitment Type	Equipment	Technical Guarantee	Work Plan	Supply Date	Proposed Team	Other Parameters	General Experience	Specification Compliance	Specific Experience	Economic Offer	Number of Process
"Moderate Economic offer" Cluster											
Quotation	0.2%	1.6%	0.1%	1.7%	1.3%	1.9%	3.0%	30.7%	3.0%	28.9%	2,423
Bidding	0.0%	0.4%	0.0%	0.3%	0.5%	8.2%	4.6%	34.8%	3.8%	33.2%	1,105
Special	0.1%	2.2%	0.1%	5.4%	2.7%	2.4%	2.0%	47.4%	1.6%	34.3%	76,216
"High Economic offer" Cluster											
Quotation	0.1%	0.1%	0.0%	0.4%	1.6%	1.9%	6.5%	1.5%	16.3%	48.0%	27,233
Bidding	0.1%	0.1%	0.0%	0.5%	1.4%	2.9%	6.4%	0.8%	22.8%	49.0%	5,950
Assurance bidding	0.0%	0.2%	0.0%	0.1%	0.8%	12.0%	9.3%	2.4%	14.7%	50.1%	2,913
Special	0.2%	7.3%	1.8%	11.0%	10.0%	11.0%	1.9%	1.7%	4.0%	46.5%	35,570
"Low Economic offer" Cluster											
Direct contracting	5.6%	0%	22.9%	0.0%	0.0%	9.9%	48.6%	0.0%	8.1%	0.0%	8,120
Special	0.3%	4.6%	0.6%	3.5%	3.7%	3.0%	46.2%	6.8%	0.8%	24.8%	24,800
"Null Economic offer" Cluster											
Direct contracting	5.9%	0.0%	16.0%	0.0%	0.0%	6.0%	18.1%	0.0%	50.3%	0.0%	61,722
Short List	5.2%	0.0%	9.0%	0.0%	0.0%	2.5%	17.2%	0.0%	60.9%	0.0%	8,960
Special	1.1%	3.3%	1.8%	2.6%	5.9%	4.6%	9.3%	17.0%	47.8%	1.9%	15,536

3.4. Semi-Supervised Learning Model

As previously described in the cluster called "Null Economic offer", processes with anomalies are identified. For the detection of anomalies, the processes associated with the clusters are defined as "normal", where the economic indicator is respected as a preponderant factor for the qualification and determination of the winner of the process. For semi-supervised learning, a training data set (80%) and a test data set (20%) are separated. As detailed in the methodology, a semi-supervised learning model is applied using SVM and PCA that can be applied in the evaluation of the regression model and for the detection of anomalies in the processes. As metrics to evaluate the success of the applied algorithms, we use: ROC curves, where the blue line represents SVM and the red line PCA, which allows us to evaluate the influence of each technique on the model Figure 8. Analysing the results, we have that the precision of the model is (0.9%) and *accuracy* is (0.92%), indicating an acceptable detection rate.

We can observe that the semi-supervised learning model applying SVM and PCA can be applied in the evaluation of the regression model and for the detection of anomalies in the processes. Table 6 indicates the evaluation metrics for each technique in the detection of anomalies model.

Table 6. Models metrics.

Technique	Precision	Recall	AUC	Accuracy	FScore
SVM	0.95%	0.92	0.90	0.92	0.96
PCA	0.92%	0.89	0.91	0.90	0.93

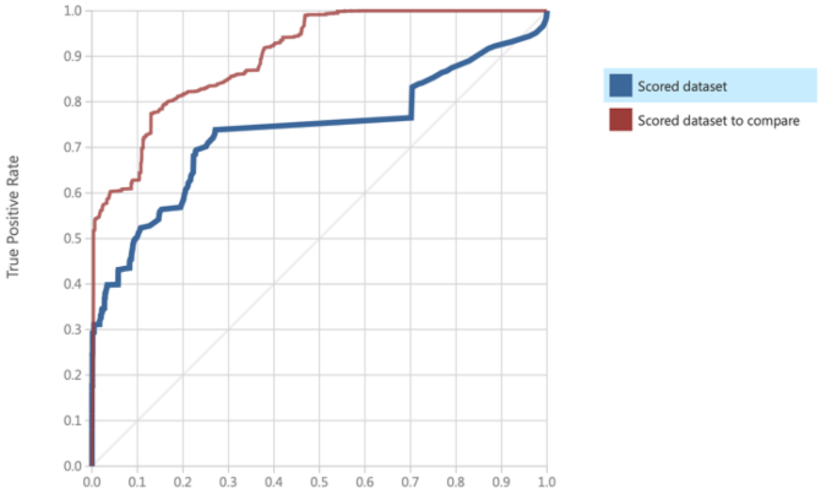


Figure 8. Semi-supervised models evaluation.

4. Discussion and Conclusions

With the experimental work, we have been able to verify the different phases of the proposed methodology to identify processes with anomalies and generate the corruption detection model. With the SOM algorithm, the main parameters involved in the qualification of winning bidders in a public procurement were identified. The K-Means algorithm allowed the identification of the three main groups where Economic Offer represented the main scoring parameter and also a group, *"Null Economic offer" Cluster*, where only 0.45% of the total rating was considered out of 100%. In this group, "other parameters" were evaluated with the greatest weight, with direct contracting, shortlisting and special publications predominating. Regarding the type of purchase, most of the purchases in this cluster are "Consultancies". It is therefore concluded that 88,358 (equivalent to 32.11%) of the processes evaluated could present anomalies in the evaluation parameters for the adjudication of contracts.

Based on the findings ("Null Economic Offer" cluster) obtained from the use of unsupervised learning, an anomaly detection model based on SVM and PCA was developed, obtaining results higher than 90% reliability; therefore, we can verify the hypothesis that guides this research.

The results of the application of the model created, in the case study, allow us to be optimistic. We consider, that through the use of data mining, anomalies can be identified, and new corruption cases can be detected. Specifically, in the definition of qualification parameters in a public procurement process which does not consider the Economic offer and causes prejudice to the government, permitting one to indicate in which cases the qualification parameters are correctly established and in which cases they are not. Experimental results are in concordance with the work of Hyytinen et al. [8], since the municipalities have the highest number of cases with anomalies in the qualification of contracts. The bidder with the lowest economic offer does not win but presents better results in terms of evaluation metrics, due to the machine learning techniques used. It also shows a difference in results with the SALER platform [10] which, while considering various parameters such as relationships between companies, does not rank contracts by the value of the economic offer in the qualification. The model is consistent and demonstrates what the previously reviewed literature points out [2], in that in order to favour certain suppliers, the contracting entity lowers the qualification of the economic offer so that the supplier with certain "special" conditions wins the process and not the provider that submits the most beneficial offer for the state. This research shows that with the use of data mining techniques, this model can be applied in several countries because in each public procurement process, qualification parameters are established to determine the winner, considering that the most important thing is to identify the processes with anomalies in the qualification, in order to adjust the model. This work represents a breakthrough in corruption research with technological tools in Latin America because as already defined in [14], there has been no progress except for in three countries.

To continue with the present work, it is important to determine the present findings with the SERCOP portal, in addition to providing a base of processes with anomalies in their qualification, new techniques for supervised learning RandomForest, Convultional networks, etc., or new combined models can be tested to determine future anomalies, such as those of cluster 4.

5. Future Work

As a future line of work, it is intended to integrate the *deep learning* in the methodology with natural language processing for the classification of contractors and relations with entities, evaluating award times. In addition, it is planned to build a *framework* that evaluates, detects and helps in the prediction of favouritism in public procurement processes.

Author Contributions: Conceptualization, Y.T.-B. and V.F.L.B.; methodology, Y.T.-B. and V.F.L.B.; formal analysis, Y.T.-B.; investigation, Y.T.; writing—original draft preparation, Y.T.; writing—review and editing, Y.T.-B. and V.F.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available in https://bit.ly/PametersCorruption (accessed on 1 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bramoullé, Y.; Goyal, S. Favoritism. *J. Dev. Econ.* **2016**, *122*, 16–27. [CrossRef]
2. Martinez Fernandez, J.M. Transparencia Versus Corrupción en la Contratación pública. Medidas de Transparencia en Todas las Fases de la Contratación Pública Como Antídoto Contra la Corrupción. 2015. Available online: https://dialnet.unirioja.es/servlet/dctes?codigo=50035 (accesed on 1 October 2021)
3. Cordova Vinueza, J.; Vaca Ojeda, P.; Hernandez Jaramillo, M. *Las Compras Gubernamentales como Política Pública*; Servicio Nacional de Contratación Pública-SERCOP: Quito, Ecuador, 2015; pp. 43.
4. Dávid-Barrett, E.; Fazekas, M. Grand corruption and government change: an analysis of partisan favoritism in public procurement. *Eur. J. Crim. Policy Res.* **2020**, *26*, 411–430. [CrossRef]
5. Servicio Nacional de Contratacion Publica del Ecuador. Análisis Anual de Contratación Pública. 2020. Available online: https://portal.compraspublicas.gob.ec/sercop/wp-content/uploads/2020/01/analisis_anual_2019_2.pdf (accesed on 1 October 2021)
6. Hermawati, F.A. Data Mining . *Min. Massive Datasets* **2005**, *2*, 5–20. [CrossRef]
7. Ferreira, I.; Camões, P.J.; Cunha, S.; Amaral, L.A. Electronic platforms and transparency in public procurement. In Proceedings of the 30th International Business Information Management Association Conference, IBIMA 2017-Vision 2020: Sustainable Economic Development, Innovation Management, and Global Growth, Madrid, Spain, 8–9 November 2017; Volume 2017, pp. 3898–3906.
8. Hyytinen, A.; Lundberg, S.; Toivanen, O. Politics and Procurement: Evidence from Cleaning Contracts. *SSRN Electron. J.* **2011**, 233; [CrossRef]
9. Auriol, E.; Straub, S.; Flochel, T. Public Procurement and Rent-Seeking: The Case of Paraguay. *World Dev.* **2016**, *77*, 395–407. [CrossRef]
10. Alzate, C.; Monreale, A.; Assem, H.; Bifet, A.; Sandra Buda, T.; Caglayan, B.; Drury, B.; García-Martín, E.; Gavaldà, R.; Kramer, S.; et al. *SALER: A Data Science Solution to Detect and Prevent Corruption in Public Administration*; Springer: London, UK, 2019; pp. 103–117. [CrossRef]
11. Kehler, M.E.K.; Paciello, J.; Fernandez, J.I.P. Anomaly Detection in Public Procurements using the Open Contracting Data Standard; In Proceedings of the 2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG), Buenos Aires, Argentina, 22–24 April 2020.
12. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the IEEE International Conference on Data Mining, ICDM, Sorrento, Italy, 15–19 December 2008; pp. 413–422. [CrossRef]
13. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*; Springer: London, UK, 2000; pp. 29–39.
14. Torres Berru, Y.; López Batista, V.F.; Torres-Carrión, P.; Jimenez, M.G. Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review. In *Communications in Computer and Information Science*; Springer: London, UK, 2020; Volume 1194, pp. 254–268. [CrossRef]

15. Saurkar, A.V.; Gode, S.A. An Overview On Web Scraping Techniques And Tools. *Int. J. Future Revolut. Comput. Sci. Commun. Eng.* **2018**, *4*, 363–367.
16. Chu, X.; Ilyas, I.F.; Krishnan, S.; Wang, J. Data cleaning: Overview and emerging challenges. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*; Association for Computing Machinery: New York, NY, USA, 2016; Volume 26, pp. 2201–2206. [CrossRef]
17. Akinduko, A.A.; Mirkes, E.M. Initialization of Self-Organizing Maps: Principal Components Versus Random Initialization. A Case Study. 2012. Available online: http://xxx.lanl.gov/abs/1210.5873 (accesed on 1 October 2021).
18. Ultsch, A.; Mörchen, F. *ESOM-Maps: Tools for Clustering, Visualization, and Classification with Emergent SOM*; Technical Report Dept. of Mathematics and Computer Science, University of Marburg: Marburg, Germany, 2005; pp. 1–7.
19. Solorio-Fernández, S.; Carrasco-Ochoa, J.A.; Martínez-Trinidad, J.F. A review of unsupervised feature selection methods. *Artif. Intell. Rev.* **2020**, *53*, 907–948. [CrossRef]
20. Guo, J.; Li, T.; Li, Y. SVM Based on Gaussian and Non-Gaussian Double Subspace for Fault Detection. *IEEE Access* **2021**, *9*, 66519–66530. [CrossRef]
21. Kohonen, T. The Self-Organizing Map. *Proc. IEEE* **1990**, *78*, 1464–1480. [CrossRef]
22. Kapil, S.; Chawla, M. Performance evaluation of K-means clustering algorithm with various distance metrics. In Proceedings of the 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, ICPEICES 2016, Delhi, India, 4–6 July 2016. [CrossRef]
23. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques; *Informatica* **2007**, *160*, 3–24.
24. Vettigli, G. MiniSom, a minimalistic and Numpy based implementation of the Self Organizing Maps Giuseppe. *J. Open Source Softw.* **2021**, 1–2. Available online: http://xxx.lanl.gov/abs/1806.02199 (accessed on 1 October 2021).