Anomaly Detection in Public Procurements using the Open Contracting Data Standard

Conference Paper · April 2020

DOI: 10.1109/ICEDEG48599.2020.9096674

3 authors, including:

Julio Paciello
Universidad Nacional de Asunción
12 PUBLICATIONS **30** CITATIONS

SEE PROFILE

Anomaly Detection in Public Procurements using the Open Contracting Data Standard

Maria Elisabeth Kehler Niessen
Universidad Nacional de Asunción
Asunción, Paraguay
maelikehler@fpuna.edu.py

Julio Manuel Paciello Coronel
Universidad Nacional de Asunción
Asunción, Paraguay
julio.paciello@pol.una.py

Juan Ignacio Pane Fernandez
Universidad Nacional de Asunción
Asunción, Paraguay
jpane@pol.una.py

***Abstract*—The detection of anomalies in public procurements can enable the improvement of the quality of purchases, and consequently enable a better quality of life in the country through the correct use of public funds. In this paper, we use as a case study the public procurement data from Paraguay, available in the open data format of the Open Contracting Data Standard, to train an unsupervised learning model for anomaly detection based on the Isolation Forest algorithm. The Open Contracting Data Standard allows the developed technique to be replicated in other countries that implement the same data standard, thus achieving an interoperable solution. The resulting classification enables the scoring of contracts and procurement processes which can be used to identify anomalies and make it possible to obtain an intelligent sampling of the data. This can be utilized as a support in the task of the government in its role to regulate and control the public procurements. The effectiveness of the model is validated with local known anomalous procurement processes, which are: a) processes protested by entities involved in the contracting process, which were determined in favor of the protestant, and b) complaints about the contracting process from external entities with the possibility of anonymity. The results show an accuracy of over 90% in detecting these known anomalies as early as in the tender stage and during the contracting stage. Thus indicating a feasible approach for anomaly detection in public procurements.**

***Index Terms*—Artificial Intelligence, Data Mining, Open Data, Open Contracting, Anomaly Detection**

I. Introduction

Public procurements represents the purchases of the government. These, like all expenditures, are vulnerable to corruption and fraud. The United Nations Office on Drugs and Crimes indicates that the price of public contracts is increased by 10-25% because of corruption globally. Paraguay in 2018 obtained a score of 29/100 in the Corruption Perception Index of the Transparency International Organization, which places Paraguay among the 50 most corrupt countries. The public procurements in Paraguay in the last 10 years represented between 20-30% of the annual public expenditure.

The proper analysis of public procurements is crucial to reduce corruption and increasing the control of the investments and expenditures of the public sector. A notable difficulty in the analysis is due to the fact that the contracting processes are of long duration and agglomerate a large amount of data. In 2017 alone 16,738 public procurement processes were registered in Paraguay, where each process can include several contracts. Taking into account the number of processes and the entailed percentage of the expenditure of the country, the need of a thorough analysis of public procurements is notable.

This paper presents an application of the Information and Communication Technologies (ICT) with the purpose of proposing a tool that makes it possible to speed up the analysis of the public procurement processes. An anomaly detection model trained with the public procurement data from Paraguay is proposed. The model can intelligently obtain an anomaly score for the procurement processes, in order to obtain a subgroup, with the highest anomaly score, for a in depth analysis. The model is based on data in the format of an international data standard designed for application on public procurement, the Open Contracting Data Standard (OCDS). A transformation process is proposed to enable the application of machine learning algorithms to the data.

Two data sets were used for the experiment: a training data set and a validation data set. In order to validate the results some procurement processes where indicated as anomalous, the processes with protests in favor of the protestant or complaints. Where protests are claims by the people involved in the hiring process and complaints can be submitted by any entity with the option of identity protection. Due to the nature of the data, it was decided to train models according to the different phases of the contracting process, thus grouping together: a) the planning and tender phase; and b) the planning, tender, award and contracting phase, resulting in 2 trained models. In the results it was noted that over 90% of the known anomalies were detected in both resulting models.

The paper is organized as follows: in section II related works are presented; in section IV the methodology of the work is delineated; the obtained results are shown in

section V; and finally the conclusions are indicated in section VI.

II. State of Art

Auriol, Straub and Flochel presented evidence of fraud in public procurement and its impact on economic development, using the linear prediction method. The data used consists of almost 50,000 public procurement operations in Paraguay, covering the period from 2004 to 2007 [1]. They discovered that in Paraguay the main channel for corruption is the use of an exceptional purchasing mechanism, which is released from the restrictions required by the other mechanisms. Another form of fraud detected in public procurement is the division of contracts into lots, with amounts that can be published in an open manner in order to avoid competition and transparency. They also investigated the companies that obtained the highest number of awards and noted that the amounts were greater than those of the market, which is especially conspicuous in sectors where there is a lot of competition, such as the importation of miscellaneous goods.

Wang employed game theory, machine learning and statistical methods to detect possible cases of fraud in public procurement [2]. In his work he studied known fraud strategies through game theory and applied nonlinear Support Vector Machines (SVM) to public procurement data from the United States. The following variables are used: the amount of the contract, the number of offers, the number of employees of the supplier and the profit divided by the number of employees of the supplier. Logit Regression is applied to the output of the SVM model in the form of reverse engineering to study the obtained results.

OCDS collaborators compiled a series of red flags, which represent possible fraud indicators [3]–[6]. For some red flags a replicable calculation using fields represented in the Standard can be obtained and is being used in countries like Indonesia, Paraguay, Hungary and others. Examples include the time available for the presentation of offers, if the evaluation criteria vas published, if always the same supplier win or if payments where made that do not correspond to the contract.

Félix J. López-Iturriaga and Iván Pastor Sanz developed an early warning system of public corruption using neural networks taking the Spanish provinces as a case study [7]. The prediction of the model is based on economic factors, using a database with the most prominent cases of corruption in Spain. The model predicts, given certain macroeconomic and political determinants, the likelihood of corruption occurring in a certain time and region. The Self-Organizing Map (SOM) algorithm was used for the training of the model.

Vierci executed an exploratory data analysis using the data mining algorithms K-Modes and ROCK on the public procurement data from Paraguay in the CSV format [8]. The performances of K-Modes and ROCK were evaluated and in addition, a series of scripts for cleaning data entered by users, like city names, were developed.

III. Background

The OCDS is a common data model for the contracting processes that accompanies the process through five stages illustrated in the figure 1. It has been implemented to represent the public contracting data for over 30 countries to date. It's focus is to make the contracting process more transparent and enable a deeper analysis of the data [9].

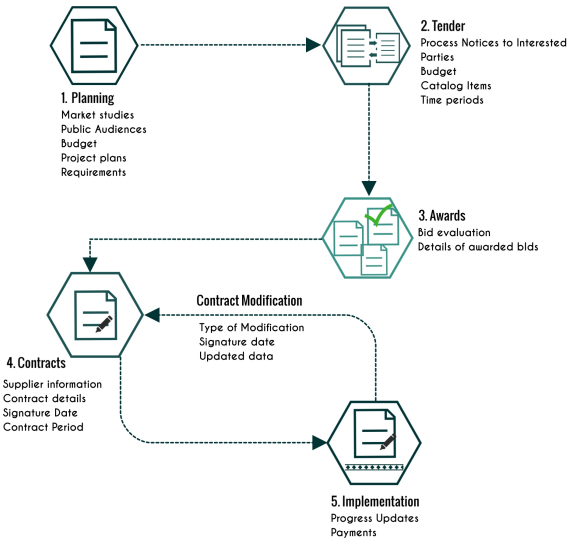

Fig. 1. Phases of the Contracting Process [10].

Planning is where the hiring process begins. During this phase the buyer compiles the requirements for the acquisition. Market studies and public hearings are carried out to determine the products needed, the budget available and to prepare the procurement plans.

During the tender phase the required technical specifications and the available budget are identified, as well as the deadlines for submission of offers, inquiries and awards. Interested companies, called suppliers, register their offer within the set period.

In the award the offers are evaluated and the winners are chosen. The award date, the amount awarded, and the awarded entity are recorded.

The contract phase formalizes the relationship between the buyer and the supplier. The standard includes the data of the contract period, the date of the signing of the contract, the amount of the contract, the status

of the contract during its execution, the supplier and the attached documents.

The implementation phase describes the settlement of what was agreed and formalized with the signing of the contract. Therefore it is fulfilled during the contract period. It includes the data of the transactions made between the buyer and the supplier, the goals achieved and the documents that are part of the implementation, which can be audit documents or evaluation reports.

During implementation, modifications to the contract of various types may arise, for example time or amount extensions. These modifications to the contracts register the date of signature of the modification, the new data of the contract and various documents, defined by the country where it is implemented.

IV. Methodology

One of the objectives of the analysis of public procurement data is that corruption and fraud can potentially be detected and prevented. Similarly, potentially anomalous patterns of the public procurement process can be found as they occur to verify if they represent possible problems. In addition, the transformation proposed for data in the OCDS format allows the implementation of algorithms to analyze them, which can be applied to the data of all countries that use the standard.

The proposition is to implement an open source tool [1] that obtains the public procurement data in the OCDS format, transforms it to numerical values and trains an unsupervised learning model to detect anomalies. The purpose of the mentioned tool is to perform a massive data analysis and assign an anomaly score to each procurement process, providing the possibility of obtaining an intelligently selected subset for an in-depth analysis to determine if it corresponds to an act of corruption or a deviation from the regulatory process of public procurement.

As was seen in the section II, there are several previous works that address the detection of anomalies in some stages of public procurement. This approach differs from the state of the art since it uses the public procurement data in the OCDS standard and applies the unsupervised learning algorithm Isolation Forest to obtain a model that can determine the ranking of data anomaly at the tender and contracting stages of the process.

[1] gitlab.com/MaEliK/anomaly_detection_public_procurements_py/

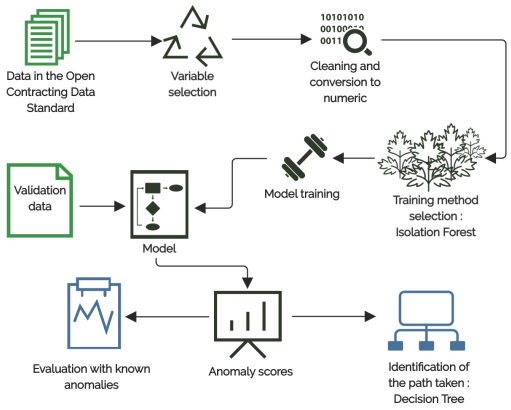

Fig. 2. Methodology of the proposed solution.

The figure 2 illustrates the methodology of the proposed solution.

The public procurement data in Paraguay is available publicly in the OCDS format from 2010 onward. The data was obtained from 2010 to 2018. To achieve the objective of training a learning model, the data must first be transformed into a more understandable and efficient format for analysis using machine learning algorithms. In the first step the most relevant variables were selected, for which the following factors were taken into account:

1) If the variable was in use.
2) If the content of the variable had a structure. Avoiding the use of free text or URLs.
3) Highly correlated or repeated variables were eliminated.
4) The data of the items were not used, the data of them not being complete and of more detailed level than the granularity defined for the project.
5) Variables of the state of the phases were also not used since the resulting model tended to identify minority states as anomalous given their scarcity.

Some variables were added taking into account the red flags compiled by OCDS collaborators [3], [4], [6]. Among them the periods between important dates, the percentage of increase in the amount of a contract modification and the percentage of time that elapsed from the beginning of the contract until the modification of the contract with respect to the duration of the contract.

It was noted that the data set to be used has the following types of data and a way of transforming each into numerical variables was determined:

- Money amounts:
  All amounts of money were converted to the numerical format corresponding to their amount in Guarani. So if the amount is in dollars the conversion rate corresponding to the date according to its stage is searched in the records of the *Banco Central del Paraguay (BCP)* and converted to Guarani.

- Dates:
  Of the variables that represent dates, only the month of the date was used, since the values for the dates may be too diverse to find a significant pattern between them. However, a new field was also created, the period between dates, which is a numerical value of the number of days between two dates. Another added field is the percentage of time that has passed since the beginning of a period, to another date, with respect to the end of the period.
- Collection of binary variables:
  For the collections of binary variables, the procedure is as follows: The collection of possible values are available in a dictionary; a binary array initialized with 0 is created; for each value in the dictionary that is part of the collection of categorical variables, the binary value is changed to 1, resulting in a binary number which is converted to numerical.
- Categorical variables:
  Categorical variables are handled as a binary array, where the options in the dictionary are the categorical options of the field and the result is the decimal number corresponding to the field according to its position in the binary array.
- Binary variables:
  For variables where the value can be only "Yes" or "No", a 1 is assigned if it is "Yes" and a 0 if the value is "No".
- Numerical variables:
  The numerical variables remain with their original value.
- Free text:
  Free text variables were converted to a unique hash value using the MD5 technique, which is obtained as a hexadecimal value and transformed to its corresponding decimal value. The resulting number is saved in a dictionary with its original value.
- Categorical variables stored as free text:
  Reviewing data of this type, it was perceived that especially city names had many variations. The scripts developed in [8] were adapted to clean the mentioned data and the process proceeds with the transformation of the data as free text.
- Identifiers:
  Some variables have an identifier and a name, description or other fields. In this case the identifiers proceeds with the process as a free text variable but the value of the corresponding name or description is maintained in the dictionary.

Analyzing the data, it is noticeable that there is little data in the form of electronic auction and this determines that the model identifies these auctions as anomalies, this leads to the use of only data in the form of physical delivery. In addition, it was determined to divide the data vertically into two groups according to the phases of the contracting process in order to generate a trained model according to the anomalies in each phase while also taking into account the previous phase. For the first division, the planning and tender phases were taken into account with the objective of detecting anomalies early before reaching the contracting phase. For the second division, the planning, tender, award and contracting phases were considered, where each record represents a contract and there may be several contracts per planning and tender.

The Isolation Forest algorithm is proposed for the training of the model, which is an anomaly detection technique designed by Liu and Ting [11]. It is based on a set of binary trees, called isolation forest, where each tree is created with a subset of data randomly selected from the data set. This technique is based on the assembled learning method called Random Forest. By dividing the data into random subsets, the probability of overfitting the model decreases, and the randomness frees it from bias in the selection of data and increases the processing speed. It is especially suitable for high dimensional data since it is not based on distance or density to determine clusters.

The Isolation Forest implementation of the *scikit-learn* package was used [12] in this experiment.

For the experiment, the model is trained with data representing 90% of the available data from 2010 to 2018. Subsequently, the model is validated with a random sample of 10% of the available data.

To validate the model, known anomalies were compiled which consists of procurement processes with:

1) **Protests that were resolved in favor of the protestant**, where protests are objections from any member participating in the procurement process.
2) **Complaints**, where complaints are objections by anyone, with the option of anonymity.

These sets are considered anomalous since the complaints to the contracting processes usually indicate an irregularity. The protests can be used to temporarily paralyze the hiring process to favor the protestant in some way, for this reason only the protests that were resolved in favor of the protestant can represent an irregularity or deviation from the process.

The selection of the parameters to be used, specifically the number of estimators and the maximum number of instances for each estimator of the Isolation Forest was determined empirically, testing with several values and observing the results obtained, where the results represent the number of anomalies detected in the totality of the data, and how many processes with protest in favor of the protestant or complaints were detected. The selection of the parameters was made to maximize the number of processes with protests in favor of the

protestant or with complaints detected as anomalous in the validation set.

Finally, the results were validated using a reverse process using decision trees to visualize the attributes used and most relevant in obtaining the anomalies. This was done by building a decision tree, using the KNIME framework, based on the anomaly score obtained by the isolation tree model. This anomaly score is classified as follows: all data with a score greater than or equal to 0 are considered "Normal", those less than 0 as "Anomalous" and those less than or equal to -0.1 as "Very Anomalous". This was determined by performing a quartile analysis of the distribution of the scores, were most of the scores where found in the range of 0.1 to -0.1.

V. Experimental Results

The training of the learning models is carried out with a set of data on public procurement in Paraguay in the OCDS format from 2010 to 2018, reaching 149,165 public procurement processes. To validate the anomalies detected, data sets provided by the body in charge, the *Dirección Nacional de Contrataciones Públicas (DNCP)* are used, which consist of public procurement processes with protests resolved in favor of the protestant or complaints.

The limitations of the experiment are the following:

1) Only metadata of the procurement process are used.
2) The data of the requirements, items and technical specifications are not used, as they are mostly found in unstructured documents.
3) Modifications of the contracts are not taking into account.
4) The model is based on finding processes with a different behaviour than the majority of procurement processes, which necessitates an posterior in depth manual review to analyse those procurement processes that resulted with a high anomaly score.

For training, a random sample of 90% of the data set is used, totaling 134,248 records and validated with the remaining 10% of the data representing 14,917 records. The processes with protests or complaints are distributed between the two data sets in the same manner. Two separate data sets were obtained according to the contracting phases, the first model using data of the planning and tender phases and the second model is trained with data of the planning, tender, award and contracting phases.

The results are presented according to the two data sets mentioned. The parameters to be used to train the models were obtained empirically. In order to contemplate variations in the results due to the random nature of the metric, 5 runs were made with each pair of parameters and the obtained averages were recorded.

Below is a table, for each model, which compares the results obtained for different values of parameters. The first 2 columns indicate the number of estimators and the sample size for each estimator respectively. The third column indicates the total number of anomalies detected in the training set. Columns 4 and 5 show how many of these protests in favor of the protestant (hereinafter, protests) and complaints were detected by the model as anomalous. In addition, the training time of the model is indicated. The first column of the validation set with title "Anomalies" indicates how many processes of the validation set were detected as anomalous and the columns with titles "Protests" and "Complaints" show how many protests and how many complaints respectively were detected as anomalous. The marked rows indicate the 3 best options according to the improvement in the detection of protests and complaints in the validation set.

For each model, a decision tree is constructed to obtain the approximation of the path taken by each procurement process to obtain its corresponding classification. For this the Konstanz Information Miner (KNIME) framework is used.

The experiments were performed with a seventh generation Intel i7 processor computer, 16 GB of RAM, an operating system Linux Debian 9.9 and using the programming language Python 3.5.3.

A. Planning and Tender Phases

In the table I it can be noted that with small samples of 20 units and 100 estimators the largest number of protests and complaints were detected in the validation set, a 90% detection of protests and 87% of complaint. Similar results were obtained with 50 estimators and 200 estimators, keeping the sample size at 20. The increase in sample size decreases the number of protests and complaints detected, as noted in the row of 20 estimators and a sample size of 100 units where 78% of the protests and 66% of the complaints are detected. These results are subject to little variation as the number of estimators increases. Training time increases with the increase in the number of estimators and also with the increase in the sample size, but still with 200 estimators and a sample size of 100 units for each estimator reaches only 110 seconds.

Set		Training				Validation		
Totals		93,373	437	223		10,375	56	37
Estimators	**Sample**	**Anomalies**	**Protests**	**Complaints**	**Time (sec)**	**Anomalies**	**Protests**	**Complaints**
20	20	38% (35744)	93% (407)	83% (186)	2,32s	38% (4002)	89% (50)	83% (31)
20	40	32% (29685)	89% (389)	78% (174)	3,62s	32% (3296)	85% (47)	74% (27)
20	60	28% (26383)	85% (374)	73% (164)	4,87s	28% (2949)	80% (45)	68% (25)
20	80	26% (24214)	83% (365)	60% (132)	6,18s	26% (2690)	80% (45)	66% (24)
20	100	21% (20122)	78% (341)	66% (147)	7,61s	21% (2225)	77% (43)	63% (23)
50	20	40% (37415)	97% (423)	89% (198)	10,08s	40% (4162)	90% (50)	87% (32)
50	40	31% (29467)	90% (393)	78% (174)	12,72s	31% (3249)	85% (48)	71% (26)
50	60	25% (23583)	86% (377)	72% (160)	15,41s	25% (2600)	84% (47)	69% (25)
50	80	27% (25085)	84% (368)	157% (71)	18,15s	27% (2795)	82% (45)	66% (25)
50	100	22% (20532)	81% (354)	67% (150)	20,90s	22% (2273)	80% (45)	65% (24)
80	20	41% (38465)	96% (418)	87% (195)	24,43s	41% (4274)	90% (50)	86% (32)
80	40	33% (30960)	93% (407)	82% (184)	28,24s	33% (3421)	88% (49)	77% (29)
80	60	28% (26104)	86% (377)	73% (162)	32,11s	39% (2911)	83% (47)	68% (25)
80	80	24% (22821)	83% (363)	70% (156)	36,17s	24% (2541)	82% (46)	66% (24)
80	100	22% (21056)	79% (344)	66% (146)	40,24s	23% (2343)	78% (44)	64% (24)
100	20	42% (39223)	97% (425)	90% (200)	44,64s	42% (4363)	92% (51)	90% (33)
100	40	31% (29179)	93% (405)	81% (181)	49,22s	31% (3213)	88% (49)	74% (27)
100	60	28% (26631)	88% (383)	76% (170)	54,17s	28% (2951)	84% (47)	69% (25)
100	80	24% (22684)	84% (367)	70% (155)	59,17s	24% (2517)	82% (46)	65% (24)
100	100	23% (21471)	81% (356)	68% (152)	64,18s	23% (2395)	80% (45)	64% (24)
200	20	41% (37872)	96% (418)	87% (195)	72,36s	41% (4212)	90% (50)	87% (32)
200	40	31% (29355)	92% (402)	82% (183)	81,45s	31% (3241)	88% (49)	74% (27)
200	60	26% (24621)	87% (382)	73% (162)	90,54s	26% (2751)	84% (47)	66% (24)
200	80	25% (23607)	84% (369)	70% (156)	100,22s	25% (2631)	82% (46)	64% (24)
200	100	23% (21073)	80% (351)	66% (148)	109,85s	23% (2339)	81% (45)	65% (24)

TABLE I
RESULTS OBTAINED TRAINING THE MODEL WITH THE DATA FROM THE PLANNING AND TENDER PHASES.

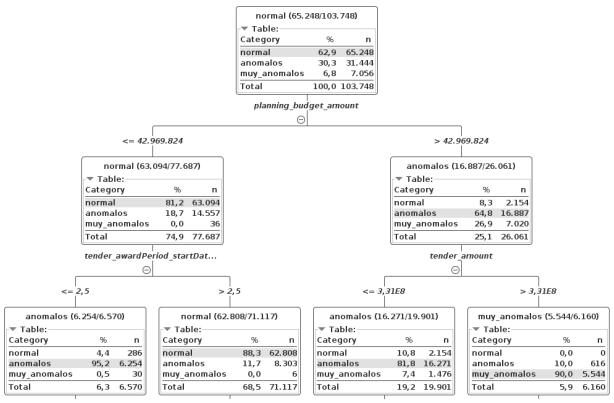

Fig. 3. Decision tree obtained for the data pf the planning and tender phases and the score from the trained isolation forest model.

The figure 3 illustrates a portion of the tree obtained using the framework KNIME with the data and the anomaly scores obtained with the model using 50 estimators and a sample size of 20 units. The first division corresponds to the amount of the planned budget, 75% rank in the left branch. These processes are cataloged according to the month of the scheduled start of the award period, those that were scheduled to start the first two months of the year have a marked tendency to qualify as anomalous. Those whose start is scheduled after the first 2 months of the year are mostly classified as normal, representing 69% of the total data.

B. Planning, Tender and Contracting Phases

The table II shows the results obtained for the data using the planning, tender, award and contract phases. It is notable that with 100 estimators and a sample size of 20 the best accuracy is obtained, 91% in the detection of protests and 92% in the detection of complaints. Keeping the estimators at 100 and increasing the sample size to 40 decreases the protests detected to 81% and the complaints to 76%. By increasing the number of estimators to 200 with the sample size by 20, the amount of protests and complaints detected remains almost equal. The execution time increases with the increase in the number of estimators and the sample size, however in the experiments performed it does not exceed 214 seconds.

The figures 4 and 5 illustrate the decision tree constructed from the results obtained for the model of 100 estimators and 20 processes for the sample of each estimator. The first variable of division of the tree is the amount of the award, 76% of the records have an amount less than or equal to 318,000,000 Guaranies. These are divided again according to the month of the date of signature of the contract, where those whose contract signature is after March are again divided according to the tender amount. Those that specify an amount greater than 124,000,000 Guaranies are again classified according the award month, where the data tends to be classified as anomalous if they are awarded in the last

Set		Training				Validation		
Totals		111,616	850	288	12,402	94	32	
Estimators	**Sample**	**Anomalies**	**Protests**	**Complaints**	**Time (sec)**	**Anomalies**	**Protests**	**Complaints**
20	20	38% (42257)	84% (716)	80% (231)	5,00s	38% (4699)	89% (84)	87% (25)
20	40	30% (33873)	76% (642)	76% (218)	7,47s	30% (3784)	82% (77)	79% (25)
20	60	22% (24277)	68% (581)	66% (191)	9,98s	22% (2727)	74% (70)	65% (21)
20	80	17% (19531)	64% (547)	61% (175)	12,80s	18% (2185)	70% (66)	64% (20)
20	100	21% (23253)	65% (553)	66% (190)	15,45s	21% (2593)	73% (68)	61% (20)
50	20	38% (42368)	89% (754)	84% (241)	20,22s	38% (4735)	91% (86)	89% (29)
50	40	30% (33391)	79% (671)	77% (223)	25,22s	30% (3735)	84% (79)	79% (25)
50	60	19% (20801)	69% (584)	67% (193)	30,28s	19% (2337)	74% (70)	64% (20)
50	80	19% (21479)	70% (594)	68% (195)	35,56s	19% (2422)	74% (70)	64% (21)
50	100	15% (16564)	61% (523)	59% (171)	40,88s	15% (1866)	68% (64)	56% (18)
80	20	37% (41646)	88% (746)	83% (240)	48,09s	38% (4674)	92% (87)	88% (28)
80	40	24% (27176)	78% (661)	76% (220)	55,72s	25% (3048)	83% (78)	76% (24)
80	60	21% (22985)	70% (598)	69% (199)	63,33s	21% (2586)	78% (74)	69% (22)
80	80	17% (18888)	66% (564)	65% (188)	71,27s	17% (2118)	75% (70)	64% (21)
80	100	15% (16331)	61% (519)	60% (171)	79,08s	15% (1831)	68% (64)	53% (17)
100	20	41% (45504)	87% (743)	85% (244)	87,67s	41% (5068)	91% (85)	92% (29)
100	40	26% (28764)	75% (642)	73% (212)	97,01s	26% (3198)	81% (76)	76% (24)
100	60	19% (21080)	69% (584)	68% (195)	106,53s	19% (2368)	76% (71)	69% (22)
100	80	19% (21118)	67% (568)	67% (192)	116,45s	19% (2368)	75% (71)	61% (19)
100	100	15% (16469)	60% (510)	58% (167)	126,30s	15% (1897)	66% (62)	52% (17)
200	20	32% (35936)	85% (727)	82% (236)	142,45s	32% (4000)	90% (85)	90% (29)
200	40	24% (26710)	76% (650)	76% (219)	160,04s	24% (2982)	83% (78)	80% (26)
200	60	19% (21520)	71% (605)	71% (204)	177,95s	19% (2407)	77% (73)	71% (23)
200	80	19% (20914)	67% (569)	68% (197)	196,23s	19% (2339)	75% (71)	64% (20)
200	100	15% (16945)	61% (516)	58% (167)	214,44s	15% (1908)	70% (65)	52% (17)

TABLE II
RESULTS OBTAINED FOR THE DATA OF THE PLANNING, TENDER, AWARD AND CONTRACTING PHASES.

month of the year (1% of the total data). In contracts with an tender amount of less than or equal to 124,000,000 Guaranies, a strong tendency to normal classification is noted, the data that followed this path represents 60% of the total data.

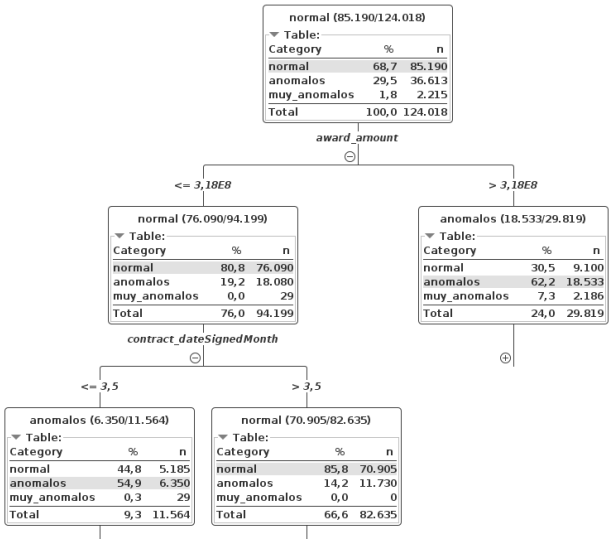

Fig. 4. Decision tree obtained for the data of the planning, tender, award and contracting phases and the score from the trained isolation forest model.

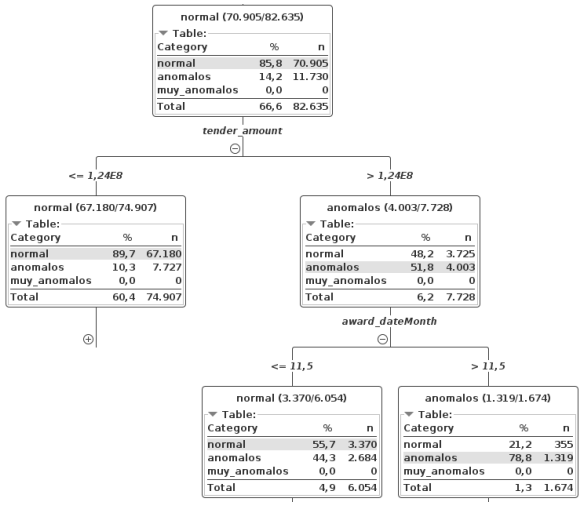

Fig. 5. Continuation of figure 4

VI. CONCLUSIONS

The results obtained show high accuracy in the detection of protests and complaints of the validation set used.

Analyzing the results obtained for the planning and tender phases it can be pointed out that already in this early phase of the procurement process 92% of the protests in favor of the protestant and 90% of the complaints can be detected. The model that includes the

planning, tender, awards and contracting phases detected more than 92% of the protests in favor of the protestant and 90% of the complaints. These models were trained in an unsupervised manner, without focus on these anomalies, however they managed to detect them with a high accuracy. It can be presumed that the models detects other anomalies that are still unknown.

According to the results, the most suitable parameters for training the models are small samples of 20 units with between 80 and 100 estimators. This indicates a preference for a bigger forest with small trees which suggests a proclivity for diversity in the model. In all experiments the training time was lower than 4 minutes with 111,616 records which suggests that the determination of the parameters to be used can be carried out in an automated process prior to training.

The decision trees proved to be an effective method of analyzing the obtained results and getting the approximate path taken by each procurement process to obtain the appropriate classification.

Utilizing data in the format of an international standard allows for the transformation, of the data of another country, to complete in a matter of days, while using data stored in an ad hoc manner would take weeks of analysing the data and determining the correct method of transforming the data.

Possible future projects could be the following:

1) Comparison between different anomaly detection techniques with the data conversion proposed in this paper.
2) Development of a model based on catalog items and their respective market prices.
3) Development of a model with supplier data, taking into account the red flags compiled by OCDS collaborators.
4) Development of a model for the implementation phase of the procurement.
5) Training of a supervised model with a marked list compiled by procurement experts.

REFERENCES

[1] E. Auriol, T. Flochel, and S. Straub, "Public Procurement and Rent-Seeking: The Case of Paraguay," TSE Working Papers 11-224, Toulouse School of Economics (TSE), Feb. 2011.

[2] Y. Wang, "Detecting fraud in public procurement," *Stony Brook Theses and Dissertations Collection*, 08 2016.

[3] O. C. Partnership. Available at https://www.open-contracting.org/resources/red-flags-integrity-giving-green-light-open-data-solutions/ (2019/06/27).

[4] C. Colombia, "Banderas rojas." Available at http://especiales.datasketch.co/contratos-colombia/banderas-rojas.html (2019/06/25).

[5] Czibik, M. Fazekas, B. Tóth, and T. János, "Toolkit for detecting collusive bidding in public procurement. with examples from hungary," *Corruption Research Center Budapest Working Paper Series CRC-WP/2014:02*, 01 2014.

[6] M. Fazekas and I. J. Tóth, "New ways to measure institutionalised grand corruption in public procurement," *U4 Brief*, 10 2014.

[7] F. Lopez-Iturriaga and I. Pastor Sanz, "Predicting public corruption with neural networks: An analysis of spanish provinces," *Social Indicators Research*, vol. 140, pp. 975–998, 11 2017.

[8] M. B. Vierci Codas, "Análisis exploratorio de datos públicos categóricos usando agrupación." Facultad Politécnica - UNA, 2018.

[9] O. C. Partnership, "The open contracting data standard." Available at https://www.open-contracting.org/data-standard/ (2019/02/05).

[10] O. C. Partnership, "Contracting process." Available at http://standard.open-contracting.org/latest/en/getting_started/contracting_process/ (2019/09/22).

[11] F. T. Liu, K. Ming Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery From Data - TKDD*, vol. 6, pp. 1–39, 03 2012.

[12] S. Learn, "Isolation forest." Available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html (2019/07/15).