NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Data-Driven Irregularity Detection in Portuguese Public Procurement

Manuel Salema Morgado Mendes Marreiros

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DATA-DRIVEN IRREGULARITY DETECTION IN PORTUGUESE PUBLIC PROCUREMENT

by

Manuel Salema Morgado Mendes Marreiros

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Bruno Miguel Pinto Damásio, PhD, Nova Information Management School

Flávio Luís Portas Pinheiro, PhD, Nova Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Manuel Salema Morgado Mendes Marreiros

Lisbon, July 2024

ACKNOWLEDGEMENTS

I am profoundly grateful to my parents for their endless support, love, and motivation, and I extend my heartfelt thanks to my colleagues and friends, whose inputs and encouragement have been invaluable throughout this journey.

I then want to express my gratitude to all the NOVA IMS community members and professors who have been very important in shaping my academic journey over the five years I spent studying in this university.

In particular, a word of appreciation to my advisors, professors Bruno Damásio and Flávio Pinheiro, for their support, wisdom and guidance throughout the development of this work.

Finally, thanks to NOVA IMS teaching assistant and PhD student Niclas Frederic Sturm who had the kindness to share the dataset of Portuguese Public Contracts he had collected and without which this project could not have been completed.

ABSTRACT

Despite being a vital governmental activity that directly affects the quality of life and well-being of a country's citizens, Public Procurement is also highly vulnerable to suffer from irregularities. Over time, procurement illicit behaviours are becoming more sophisticated and complex, posing ever so significant challenges for detection and prevention efforts. However, while that happens, data science and artificial intelligence methods are proving themselves to be powerful allies in the fight against procurement-related misconduct, due to the possibilities these methods provide of analysing very large datasets with few costs in less time and detecting patterns that a human cannot detect on his own. Leveraging these methods, this research aspires to address these challenges by implementing anomaly detection techniques tailored to a Portuguese procurement dataset of over two million records posted to Portal BASE between the years of 2009 and 2022, offering the possibility of enhancing the efficiency and effectiveness of irregularity detection processes. By integrating the anomaly score of each contract with red flags that capture hints of suspicious behaviours from the involved awarding and awarded entities, the study aims to construct a comprehensive methodology for prioritizing investigations, assigning a score of suspicion or risk to each contract, which can be used by human auditors to focus their attention and concentrate their efforts on a much smaller and more manageable set of contracts. Ultimately, this research seeks to contribute to the advancement of transparency, fair competition, integrity, and accountability within Public Procurement systems, thereby safeguarding the interests of citizens and promoting good governance practices.

KEYWORDS

Artificial Intelligence; Public Procurement; Anomaly Detection; Transparency; Red Flags; Clustering

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

1. Introduction 1
2. Literature Review 3
2.1 Machine Learning 4
2.2 Risk Indexes 5
2.3 Network Analysis 5
2.4 Summary of Related Work 7
3. Methodology 9
3.1 Data Retrieval from Portal BASE..9
3.2 Data Cleaning and Preprocessing 12
3.3 Data Exploration 15
3.4 Anomaly Detection at Contract Level...17
3.5 Creation and Calculation of Red Flags at Entity Level 21
3.5.1 Red Flag 1 – Percentage of participations in a public tender that has only 1 entity bidding...22
3.5.2 Red Flag 2 – Engaging in tenders with contract value altered during the execution of it...22
3.5.3 Red Flag 3 – Win share (much) greater than average of sector 23
3.5.4 Red Flag 4 – Often engaging in non-competitive procedures......................23
3.5.5 Red Flag 5 – Recurrent excessively fast conclusion of the contract 23
3.5.6 Red Flag 6 – Contract price (much) higher than sector average..................23
3.5.7 Red Flag 7 – High percentage of contracts celebrated between a contestant and an awarding entity 24
3.5.8 Red Flags suggested for Future Work 24
3.5.9 Red Flags Summary 25
3.6 Final Metric Calculation..26
3.7 Cluster Creation and Analysis 27
4. Results and Discussion..29
4.1 Anomaly Detection Results 29
4.2 Red Flags Results 32
4.3 Combined Results 33
4.4 Clustering Results 34
5. Conclusions...37
Bibliographical References...40
Appendix A – Public Procurement in Portugal..45
Appendix B – Main Procedure Types 46

Appendix C – Numeric variables distribution and correlation matrix before preprocessing 47
C1 - Histograms....47
C2 - Box Plots....47
C3 - Correlation Matrix....47
Appendix D – Numeric variables distribution after preprocessing....48
D1 - Histograms....48
D2 - Box Plots....48
Appendix E – Visualizations....49
E1 - Number Of Contracts Per Procedure Type....49
E2 - Expenditure Per Procedure Type....49
E3 - Number Of Contestants Per Procedure....49
E4 - Number of Winners Per Procedure....50
E5 - Final Price Distribution Per Procedure Type....50
E6 - Number of Contracts and Expenditure Per CPV Division....50
E7 - Expenditure Per District and Procedure Type....51
E8 - Number of Contracts and Expenditure per District per 10k Habitants....51
Appendix F – Auxiliar Results and Metrics....52
F1 - Kolmogorov-Smirnov Test Results....52
F2 - Clustering Metrics....52
Appendix G – Cluster Visualizations....53
G1 - t-SNE Visualization....53
G2 - UMAP Visualization....53

LIST OF FIGURES

Figure 1 - Methodology 9
Figure 2 - CPV Code Example 10
Figure 3 - Yearly Expenditure per Procedure Type 15
Figure 4 - Number of Celebrated Contracts 16
Figure 5 - Contract Expenditure 16
Figure 6 - SHAP Summary Plot for Anomaly Detection 29
Figure 7 - SHAP Force Plot of a Normal Contract 30
Figure 8 - SHAP Force Plot of an Anomalous Contract 31
Figure 9 - SHAP Feature Importance 31
Figure 10 - Number of Red Flags per Contracted Entity 32
Figure 11 - Number of Red Flags per Contracting Entity 33
Figure 12 - Contracts Ranked by Final Metric (Top 10) 34
Figure 13 - Cluster Profiles 35

LIST OF TABLES

Table 1 - Summary of Most Relevant Reviewed Articles 8
Table 2 - Data Dictionary 11
Table 3 - Descriptive Statistics of Metric Features 12
Table 4 - Number of Missing Values per Feature 13
Table 5 - Comparative Analysis of Anomaly Detection Algorithms 20
Table 6 - Red Flags Construction 25
Table 7 - Descriptive Statistics of Constructed Red Flags for the Awarded Side 25
Table 8 - Descriptive Statistics of Constructed Red Flags for the Awarding Side 26
Table 9 - Comparative Analysis of Clustering Algorithms 27
Table 10 - Cluster Evaluation Metrics 28
Table 11 - Anomalous Contracts Clusters 36

LIST OF ABBREVIATIONS AND ACRONYMS

AdC Autoridade da Concorrência (translates to *Competition Authority*)

AI Artificial Intelligence

CCP Código dos Contratos Públicos (translates to *Public Contracts Code*)

CPV Common Procurement Vocabulary

GDP Gross Domestic Product

IF Isolation Forest

IMPIC Instituto dos Mercados Públicos, do Imobiliário e da Construção (translates to *Institute of Public Markets, Real Estate and Construction*)

LOF Local Outlier Factor

ML Machine Learning

NIF Número Identificação Fiscal (translates to *Fiscal Number)*

OECD Organization for Economic Cooperation and Development

SHAP SHapley Additive exPlanations

TDC Tribunal de Contas (translates to *Court of Auditors)*

1. INTRODUCTION

According to the Organisation for Economic Co-operation and Development (OECD), Public Procurement is defined as the purchase with public money by governments and state-owned enterprises of goods, services and works, and it is an activity that should be conducted in a transparent and competitive manner to achieve the best possible value for money, foster essential synergies between public and private entities (Tadelis, 2012) and promote innovation (Bento et al., 2022). This activity accounted for 14% of European Union's GDP and 5.23% of Portugal's GDP in the year of 2022, a value that has been increasing year after year, as stated in *Relatório Anual da Contratação Pública em Portugal* (2022). According to the same report, more than one hundred eighty thousand public contracts were celebrated in Portugal in that year alone, another value that highlights the importance – but also the difficulty – of thoroughly analysing these contracts and ensuring their compliance with the regulations in force.

Despite being a vital government activity that affects the quality of life and wellbeing of citizens, Public Procurement is also one of the governmental activities most vulnerable to suffer from irregularities, due to the very large number of transactions that take place and the significant financial interests involved in them, but also because of factors such as the intricate nature of the process, the close collaboration between public officials and business entities, and the multitude of stakeholders influencing the process (OECD, 2009). These irregularities can lead to the loss of public funds through misallocations or higher expenses and lower quality of goods, which, in the long term, may jeopardize basic services such as healthcare, education, justice, science, security, and defence (Heggstad & Frøystad, 2011).

Thanks to the recent rise of E-Procurement, that has made contract-related data ever so available and is seen as a tool that can improve procurement in many ways (Lewis-Faupel et al., 2006), efforts in the realms of Data Science and Artificial Intelligence have been made to revolutionize how irregularities are detected, providing ways of analysing very large datasets with few costs in less time, detecting patterns that a human cannot detect on his own (Carneiro et al., 2020), and signalling malpractices quickly after they are consummated, thus allowing for greater cost savings (Gallego, 2021). Unlike traditional enforcement models that rely on periodic and retrospective audits to investigate, prosecute, and punish, data-driven approaches can identify issues much sooner, making corrective measures more effective (Baesens et al., 2015). However, there is still a lack of reliable indicators of irregular activity and concrete approaches to tackle the issue using the available data (Fazekas et al., 2016). In Portugal's case, despite the extensive data available on Portal BASE, its full utilization remains unrealized.

In that sense, this research asks the question of how can data science methods be used to detect and analyse patterns associated with irregularities in Public Procurement contracts, and focuses on implementing anomaly detection techniques at a contract level along with red flags at an entity level to identify such irregularities, this way creating a framework that can be used to assign a score of neediness to investigate to each contract, helping human auditors in the task of regulating and controlling the Public Procurement process.

It advances past studies by going beyond individual contract analysis, creating red flags at the entity level to integrate with the anomaly detection scores and that way derive a comprehensive metric representative not only of the contract, but also of the entities that partake in it. Additionally, a final layer of clustering over the identified anomalous contracts enables targeted actions for each group, allowing for even greater scrutiny. This methodology successfully identified in the period of 2017 to 2022 a subset of around eight thousand contracts exhibiting deviant values across multiple features, with both awarding and awarded entities displaying questionable behaviours. By providing a detailed and structured approach to anomaly detection, this research enhances the ability to pinpoint and address irregularities effectively in Portuguese Public Procurement.

This thesis is divided into four sections: literature review, methodology, results and discussion and conclusion. In the first section, the focus is on understanding the current landscape of applying AI and ML techniques to detect irregularities in Public Procurement. In the methodology, a detailed guide into the techniques, methods and tools that are going to be employed is provided. The obtained results and their implication within the study's context are then presented, and finally the main conclusions taken from this work are laid out, along with future research suggestions.

2. LITERATURE REVIEW

The purpose of this section is to understand the current landscape of applying AI and ML techniques to detected irregularities in Public Procurement, by summarizing the main results of a set of papers and articles. Some features of interest across all the reviewed literature consist of the disciplines involved, datasets used, techniques employed, and the methodologies and technologies adopted. To better understand the discussed studies, it is important to have some prior knowledge of how Public Procurement works and how it is regulated, which can be done by consulting Appendix A – Public Procurement in Portugal.

As digitalization becomes more common and the practice of publishing public contracts online is widespread throughout the world, recent years have been prolific in publications about this topic, with these getting each time more robust and technically advanced. Current investigations predominantly leverage machine learning, risk indexes, and graph networks to detect suspicious patterns and potential irregularities, with some of them achieving interesting results. Despite the different ways of tackling the problem, the goal in all of them is ultimately to increase the efficiency with which irregularity detection tasks are currently performed and achieve a degree of automatization that reduces the manual effort of the human auditors, allowing the person doing that job to target their focus on a smaller subset of contracts and not the entire, vast list of contracts.

Another common thing across research in this topic is the frequent use of unsupervised or semi-supervised learning techniques (Phua et al., 2010). This trend is largely driven by the challenges associated with creating a labelled dataset of irregular contracts, something that is often impractical due to several factors. One approach to create such dataset would be to categorize every contract associated with a company convicted of corruption as irregular, but this method is limited by the availability to the public of court records. Another approach might entail identifying individual irregular contracts through exhaustive examination of court records, yet this is labour-intensive and subject to the same accessibility issues. Additionally, one could label contracts as irregular based on complaints made against them, though this introduces subjectivity and potential bias.

Because of the substantial differences the three data-driven approaches have between them, a dedicated section will be allocated to explore the intricacies of each individually. However, before delving into the different methods employed to detect irregularities in Public Procurement, it is important to acknowledge that such advanced techniques have not always been available. Traditionally, the detection of irregularities relied on manual processes and audits conducted by human experts. These traditional methods involved the examination of procurement documents, financial records, and compliance reports, often requiring significant time and effort from auditors (Fazekas & Kocsis, 2020). Therefore, all the studies discussed below constitute an important step forward in the evolution of irregularity detection practices.

2.1 Machine Learning

Either through unsupervised learning methods that try to cluster together contracts or entities that possess similar characteristics, assigning them anomaly scores based on how different they are from the rest, or through supervised/semi-supervised techniques that utilize historical data to train classification and regression models, the applications of machine learning in Public Procurement irregularity detection are diverse and promising.

Niessen et al. (2020) exemplify the power of unsupervised learning through their use of the Isolation Forest algorithm to assign anomaly scores to procurement processes based on their intrinsic characteristics and deviant traits, harnessing both existing metadata features and innovative variables inspired by recognized red flags in the literature, such as the durations between critical contract dates or the percentage increase in contract modification amounts.

Torres-Berru and López Batista (2021) combine unsupervised and semi-supervised learning to identify dubious practices in tenders, such as altering selection criteria to favour certain companies and specific offers or prioritizing companies with prior experience and an history of employment. By using Self-Organizing Maps and K-Means clustering, they detect and group together anomalies in the presence, absence and weight of qualification parameters of procurement processes, and subsequently use Support Vector Machines to classify contracts based on the established clusters.

Gallego et al. (2021) take a supervised learning approach with Gradient Boosting Machines to develop early warning models, considering anomalous the contracts that included entities present in a database of vendors that had been fined due to a breach of contract, contract extensions, and contracts under investigation. By assigning risk scores to contracts based on features such as budget, duration, and vendor history, they identify contracts likely to lead to irregularities. Despite the already referred difficulties of gathering a labelled dataset of irregular contracts, this method showcases how historical data can be used to train models to predict future risks effectively, with other authors following similar approaches and testing out different models and configurations for the effect, from binary classification with tree-based methods like Random Forest (Aldana et al., 2022), to using a Multilayer-Perceptron for binary classification followed by multiclass classification (Bai & Qiu, 2023).

In the realm of deep learning, leveraging the notorious evolution this field has undergone in the recent past, significant strides have been made in applying natural language processing techniques and neural network models, such as Deep Neural Networks and Long Short-Term Memory Networks, in the detection of procurement irregularities (Rabuzin & Modrušan, 2019; Lima et al., 2020). These models, using techniques like TF-IDF for feature extraction, have demonstrated a good performance in classifying procurement-related texts and documents, making them a good complement for the more traditional ML models.

Finally, while the aforementioned methods aim to directly identify irregularities, there are alternative approaches that offer indirect yet insightful ways to detect anomalies in Public

Procurement. A notable example comes from Rodríguez et al. (2019) that propose an award price estimator that uses a Random Forest model to predict what would be the adequate price of a tender given its characteristics (duration, agency name, subtype code, CPV classification, or postal zone). With this approach, each tender is treated as a regression problem where the goal is to predict the award price. Even if the model is not directly detecting if a contract is irregular, it can be a good way of detecting cases in which the state entity is overpaying for a given product or service.

2.2 Risk Indexes

Indexes can play a very important role in reducing intricate and varied data to concise and understandable measurements. In what is probably the most remarkable application of risk indexes in this context, Fazekas et al. (2016) used public contracts of Hungary to develop an objective corruption index to which they called composite corruption risk index (CRI). To do so, they first identify three contract corrupt outcome indicators: single bidder, exclusion of all but one bidder and winner's share of issuer's contracts.

Then, the authors gather a broad array of elementary indicators, each associated to one of the three established corrupt outcomes. Using multiple regressions (logistic for the first two outcomes and linear for the third one), a link is made between likely corrupt inputs and outcomes. In this process, regression coefficients are derived with values between zero and one, representing the strength and direction of association between the various inputs and their corresponding outcomes. The component weights of the composite indicator are derived from these regression coefficients, which provide valuable insights into the extent to which each input influences the likelihood of a particular corrupt outcome occurring, and allow for different contracts to be compared over a common metric.

According to the authors, comparable datasets that enable the recreation of the CRI are available or can be constructed from public records in most developed countries, including all EU member states. However, this claim is not entirely accurate, as some variables used in the study, such as the "Relative Length of Eligibility Criteria," the "Weight of Non-Price Evaluation Criteria," or the indication of an "Annulled Procedure Re-Launched Subsequently," cannot be reconstructed using the metadata of Portuguese contracts.

2.3 Network Analysis

A network is a collection of nodes along with edges connecting them. In most cases, the nodes consist of the entities involved in the Public Procurement process, and the edges represent the relationships between the issuers and winners, often having a weighted value correspondent to the monetary worth of the contract, as well as a direction going from the issuer to the winner, making it a directed edge-labelled graph (Hogan et al., 2022).

Network analysis approaches sit on the basis that Public Procurement processes involve complex interactions and relationships among various stakeholders, and these interactions can be effectively captured and analysed using networks and network science methods. These methods have proven themselves to be particularly useful in uncovering cartels, which are entities that secretly cooperate to eliminate or limit competition and manipulate the procurement process to their advantage (Lyra et al., 2022), in what is called collusion.

To detect irregularities with network analysis, a very common approach is the use of community detection algorithms, used to locate communities within a graph, that is, sub-graphs with dense internal connectivity when comparing to external connectivity. Another frequently employed measurement is node centrality, that provides a way to quantify nodes based on their position and how many connections they have.

Carneiro et al. (2020) created a multi-layered architecture using graph databases and network analysis to detect irregularities in Public Procurement. They gathered data from legal documents and transactions (mostly from Portal BASE), storing it in a graph database, and then used a rule-based system to annotate nodes (representing stakeholders) and edges (representing the contracts) with indicators of irregularities. Network analysis algorithms were then applied to evaluate characteristics such as centrality and dispersion. Furthermore, a machine learning module trained on this annotated data helped automatically identify suspicious groups of nodes, showcasing how different domains can be interconnected to enhance irregularity detection capacities.

Lyra et al. (2021) applied network analysis to study co-bidding relationships between firms in Brazil. Opposingly to the work done in the previous study, this network did not capture interactions between state and private entities, but rather interactions between private entities bidding in processes together, with the similarity between the bidding patterns of two firms being captured with the Jaccard coefficient. Unsurprisingly, the network showed great relationships between firms that operated in a similar geographical space and firms that supplied similar types of services. Besides this expected effect, some firms also formed densely connected sub-graphs, which was seen as a potential risk of collusion between them. Analysis of single bidder contracts and average number of bidders within these communities provided additional evidence of suspicious activities and suggested conditions for coordination aimed at exerting control over specific markets. In another approach, Velasco et al. (2021) also focus their efforts on the entities bidding for tenders and the potential for collusion among them, looking at entities with similar patterns, concerted bids or common losers amongst different contests to identify potential causes for concern.

Wachs and Kertész (2019) presented a network-based framework to detect potential cartels in bidding markets by examining the patterns of firm interactions. Their method involved creating a network of firms based on their co-bidding behaviour, detecting interacting groups, and measuring their cohesion and exclusivity.

In a broader European context, Wachs et al. (2021) analysed over four million contracts between the years 2008 and 2016, comparing null models to Public Procurement networks composed of public bodies and winning firms with a weighted link between them representing the number of contracts celebrated by the two parties. Their findings highlighted a correlation between high market centralization and corruption risk, with centralized networks often indicating controlled markets. They also found that communities with high single-bidding rates were often indicative of irregular activities, suggesting a concentration of risk among central players rather than a dispersed distribution.

Studying countries individually, they concluded that Portugal showed a higher risk in the core of the market and a weaker tendency for risk to cluster, meaning that irregularity risks are more likely concentrated in major contracts or among central players, rather than being dispersed across different regions or smaller contracts. Sturm et al. (2023) use network statistics of centrality to study the impact of network position on tender earnings to further highlight this idea that the Portuguese Public Procurement market's dense core networks inhibit market access for peripheral firms, pointing the necessity for sector-specific regulations and enhanced entry opportunities for smaller enterprises to foster competition and economic benefits.

2.4 Summary of Related Work

The reviewed studies demonstrate the diverse methodologies employed to detect irregularities in Public Procurement, each with its own strengths and limitations. Machine learning models provide robust anomaly detection capabilities, which can be further enhanced in the presence of labelled data, risk indexes simplify complex data into actionable and objective metrics, and network analysis uncovers hidden relationships and patterns of collusion between entities. However, there is a notable gap in integrating these approaches into a unified framework that combines contract-level and entity-level analyses. While not always the case, it appears that ML and index-based approaches tend to focus more on the individual contracts, whilst graph networks provide a more holistic approach and focus the analysis on the awarding and awarded entities.

With this research gap in mind, the idea comes for a framework that can incorporate the strengths of these three methodologies. Such a framework would leverage the detailed anomaly detection of machine learning, the simplicity and clarity of risk indexes, and the entity-driven perspective of network analysis, providing a holistic view of the Public Procurement landscape. By merging these components, it becomes feasible to develop a comprehensive tool that not only detects irregularities but also offers insights into the underlying structures and relationships that facilitate the perpetration of those irregularities. This integrated approach would significantly enhance the efficiency of irregularity detection, allowing auditors to focus on the most suspicious contracts and entities, thereby improving transparency and accountability while also fostering fair competition in Public Procurement.

The following table summarizes the discussed studies that are most relevant to the proposed approach.

Table 1 - Summary of Most Relevant Reviewed Articles

Authors	Area	Dataset	Techniques Employed	Observations
Niessen, Paciello, Fernandez	ML (Unsupervised Learning)	~150000 contracts of Paraguay from 2010-2018	Isolation Forest	With unsupervised anomaly detection methods like IF, the authors were able to detect with high accuracy the contracts that were subject to protests and complaints. These results must be analysed carefully, as the complaints and protests are not definitive signs of irregular contracts.
Torres-Berru, López Batista	ML (Unsupervised and Semi-Supervised Learning)	~275000 contracts of Ecuador from 2010-2020	Self-Organizing Maps, K-Means, SVM, PCA	The main parameters involved in the determination of winning bidders were identified. Then, a group where the economic offer was not the main parameter was found and this cluster was used as a base for an anomaly detection model.
Rodríguez, Montequín, Fernández, Balsera	ML (Supervised Learning)	~58000 contracts of Spain from 2012-2018	Random Forest	The authors proposed an award price estimator that predicts what the expected price for a tender is as a regression problem.
Gallego, Rivero, Martínez	ML (Supervised Learning)	~2000000 contracts of Colombia from 2011-2015	LASSO Regression, GBM	Despite it being difficult to gather fraudulent contracts with labels, some approaches can be followed to do that and make it possible to implement supervised learning techniques in this problem.
Toth, Fazekas, King	Index	~53000 contracts of Hungary from 2009-2012	Logistic Regression, Linear Regression	By defining a set of corruption outcomes and corruption indicators, it is possible to derive weights that connect the two and, in that way, create a formula that can be applied to a number of contracts. Portuguese contract data does not allow for a direct replication of the index.
Carneiro, Veloso, Ventura, Palumbo, Costa	Network Analysis	Public contracts of Portugal taken from Portal BASE	Graph Network, Rule Engine, ML Module	It is possible to store contracts as relationships between nodes. Then, with a Rule Engine, it is possible to tag contracts that are not compliant with the established rules. These contracts can then be used as examples of irregular contracts in classification tasks to train ML models.
Lyra, Curado, Damásio, Bação, Pinheiro	Network Analysis	~85000 contracts of Brazil (State of Ceará) from 2015-2019	Graph Network, Jaccard Coefficient	Using the Jaccard coefficient to reveal significant connections between bidders, regional and service-based relationships emerged as densely connected sub-graphs, but some other companies also exhibited this connection between them, possibly indicating collusion. The analysis also identified a community with suspicious values in terms of participations in single bidder contracts and average number of bidders in each tender.
Wachs, Fazekas, Kertész	Network Analysis	~4000000 contracts of EU member states from 2008-2016	Graph Network	A network was created to map the connections between winning enterprises and governmental entities, identify important individuals, and evaluate the possibility of corruption. The study found a link between higher market centralization and a higher probability of corruption.
Sturm, Candia, Damásio, Pinheiro	Network Analysis	Public contracts of Portugal taken from Portal BASE	Graph Network, Jaccard Coefficient	Network science methods including network statistics of centrality and community detection were used to uncover the structural characteristics and competitive conditions of the Public Procurement market, concluding that firms in influential positions earn significantly more.

3. METHODOLOGY

The proposed approach comprises a combination of anomaly detection, red flag extraction, and clustering techniques. First, the contracts will be analysed and ranked according to their degree of anomaly across a set of features (how much they differ from the norm). Then, the companies on the awarded and awarding sides will be evaluated in terms of how many red flags they possess, and this information will be merged with the anomaly detection results to get a final ranking of the contracts in terms of the necessity to investigate them.

By integrating diverse methods from various domains, this approach combines the identification of specific indicators of potential deviant behaviours through anomaly detection with broader insights into possible illicit activities derived from red flag extraction. Clustering adds an extra layer of meaning and enhances these insights, facilitating the identification of coherent patterns in the dataset and allowing for personalized actions to be followed for each group.

This methodology is presented on a high-level in the next figure and described in greater detail right after it.

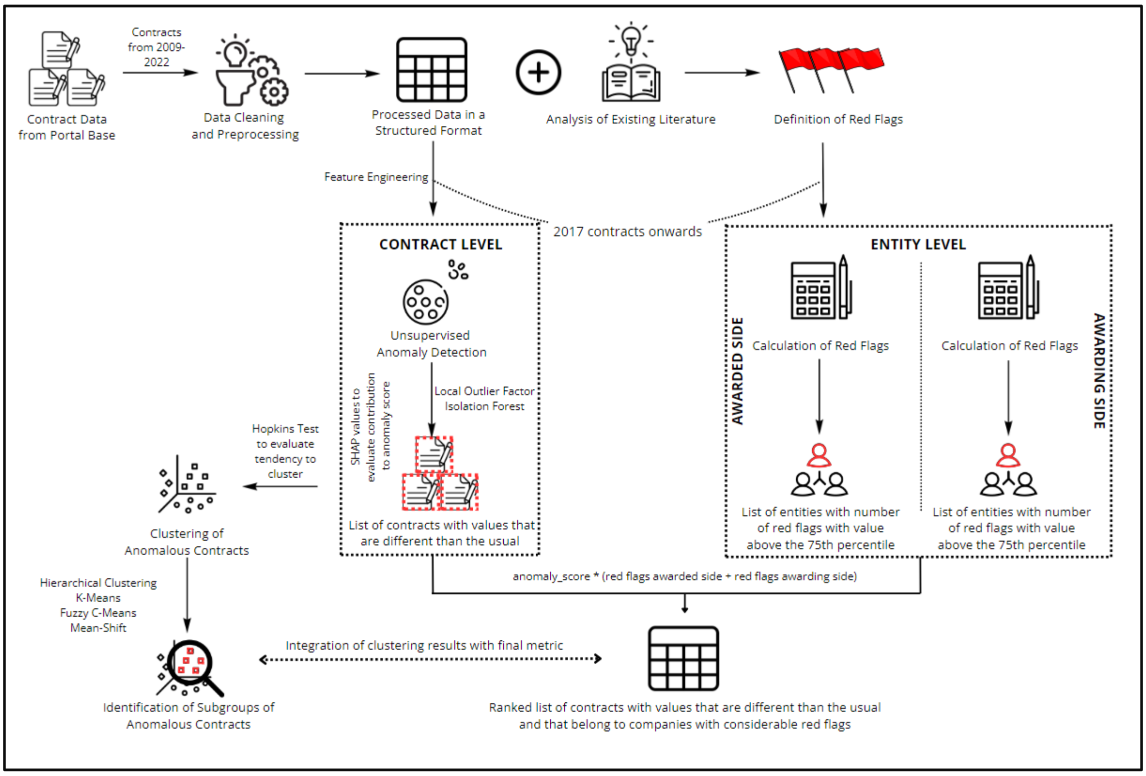


Figure 1 - Methodology

3.1 Data Retrieval from Portal BASE

The data used to conduct this study was extracted from Portal BASE as of the 15th march of 2023. This means that the more recent years, particularly 2022, have some contracts at fault, since their upload to Portal BASE has been done in a date posterior to the date of extraction. This situation must be taken into consideration when looking at the graphics from the exploratory data analysis, but it does not affect the remainder of the study, since no yearly partition is used in any of the subsequent steps. The collected data comprises:

- 2637060 observations between the years of 2009 and 2022, each corresponding to an entity applying to a contest to be the provider of a service or a good for a state entity, regardless of if they were the winner or not.
- 28 variables.

Each variable is described in Table 2. However, to fully comprehend the meaning of certain features, it is essential to clarify some key concepts related to Public Procurement.

Contracts can be divided regarding their nature and the number of competing entities into non-competitive and competitive procedures. In non-competitive procurement, a single company is directly contacted by the state entity to provide the service/good required. This is particularly useful in cases of emergency, where going through the entire and usual lengthy contracting process may not be possible because of time constraints, and there is an immediate need of getting the product delivered or the service complete. As for competitive procurement procedures, they involve an open process where multiple companies can apply, and it is the responsibility of the state entity to evaluate all the proposals and select the one that provides the best value for money. The dichotomy of competitive and non-competitive procedures can be further split into different types that are classified according to how the contestants are selected, the type of procedure and its specifications. The most relevant procedure types are listed and described in Appendix B – Main Procedure Types, according to the official information (Portal BASE, 2022).

The common procurement vocabulary (CPV) code, established by the European Commission in *Regulation 213/2008*[1], is a nine-digit standardized classification used to categorize the subject matter of public contracts across EU Countries. It works in a hierarchical way, meaning that digits add detail to the information progressively. The first two digits refer to the division, then the third one adds group context, the fourth refers to the class and the fifth to the category. Each of the last three digits add an additional degree of precision within each category, and the ninth digit serves for the verification of the preceding digits.

Division **72**: IT services: consulting, software development, Internet and support

Group 72-**3**: Data services

Class 72-3-**2**: Database services

Category 72-3-2-**1:** Added-value database services

Figure 2 - CPV Code Example

1 - https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008R0213

Every Public Procurement process is also categorized according to a contract type, which identifies the nature of the contract in terms of the purpose it was celebrated for: acquisition of movable goods, acquisition of services, concession of public works, concession of public services, public works contracts, rental of movable goods, partnership, and others.

Table 2 - Data Dictionary

Feature	Data Type	Description
contract_id	Integer	Contract unique identification number
name_contestants	Object	Name of a contestant that did not win the tender
nif_contestants	Object	Fiscal number of a contestant that did not win the tender
name_contracted	Object	Name of the contestant that won the tender
nif_contracted	Object	Fiscal number of the contestant that won the tender
contract_type	Object	Type of service provided or good acquisition associated to the contract
invitees	Object	Names and fiscal numbers of the invited entities (only procedures that allow invites)
centralized_procedure	Object	Whether it is a procedure to satisfy the needs of multiple entities or not
procedure_type	Object	Type of procedure followed for the contract adjudication
name_contracting_agency	Object	Name of the state entity that is adjudicating the contract
nif_contracting_agency	Object	Fiscal number of the state entity that is adjudicating the contract
publication_date	Object	Date in which the contract was published on Portal BASE
close_date	Object	Date in which the contract object was finalized (in case of a service) or delivered (in case of a good)
cpvs	Object	Common Procurement Vocabulary Code
execution_deadline	Integer	Number of days allowed until project completion
execution_location	Object	Country, District and Municipality where the contract was celebrated
initial_price	Float	Initial price of the contract
final_price	Float	Final price of the contract
signing_date	Object	Date in which the contract was signed
environmental_criteria	Boolean	Whether the procedure has environmental criteria or not
material_criteria	Boolean	Whether the procedure has material criteria or not
causes_deadline_change	Object	Reasons that caused a delay in the project (if applicable)
causes_price_change	Object	Reasons that caused a price change – increase or decrease – in the project (if applicable)
document_id	Float	Identification number of other documents associated to the procedure

document_name	Object	Name of the document associated to the procedure
contract_year	Integer	Year in which the contract was celebrated
num_winners	Integer	Number of winners in the procedure
award_per_winner	Float	How much money each winner gets. Obtained by dividing the final_price by the num_winners

It is possible to do a division of these features into numeric and non-numeric, with numeric features being those that represent quantities or measurements in a continuous way, and non-numeric features being the ones that refer to qualitative variables. The metric features, whose distribution can be seen in Appendix *C1* - Histograms and Appendix *C2 - Box* Plots, are *execution_deadline*, *initial_price*, *final_price*, *num_winners* and *award_per_winner*, and the remaining are non-metric features. From the metric features one can extract statistical descriptors that are helpful to understand how the data is distributed.

Table 3 - Descriptive Statistics of Metric Features

Feature	mean	std	min	25%	50%	75%	max
execution_deadline	1062.58	3.79e+05	0	30	181	365	5.09e+08
initial_price	128712.20	1.62e+06	0.01	2630	11904.03	46230.32	3.97e+08
final_price	138191.64	2.33e+06	0.01	2502	11568	45477.60	1.49e+09
num_winners	1.09	1.22	1	1	1	1	56
award_per_winner	133869.48	2.28e+06	0.003	2450	11268	44359.14	1.49e+09

Another thing that can be done with numeric variables is to analyse the correlations between them using the Spearman Correlation, which evaluates how well a monotonic function can explain the relationship between two variables (Appendix *C3 - Correlation* Matrix). Doing so depicts a correlation of 0.99 between the variable *final_price* and *award_per_winner*, which makes sense if we think that the latter is calculated by dividing the first one over the number of winners for each procedure. Besides that, *final_price* also appears to be correlated with *initial_price* at a degree of 0.7, which happens because the final price depends on the initial price and the two values are often the same if no alterations exist to the price of the contract during its execution.

3.2 Data Cleaning and Preprocessing

Data cleaning is a crucial part of any data science project and is one of the first things to be done. It entails analysing the data and correcting any mistakes or issues that it may have, making it ready and coherent for the subsequent steps. This step includes:

1. Handling missing values
2. Ensuring that no duplicate contracts exist

3. Handling outliers in metric features

4. Fixing data types and standardizing formats for money and dates

5. Fixing incorrect data and filtering out irrelevant data

6. Scaling numeric variables

Considering the missing values, eight variables had them: *name_contestants*, *invitees*, *centralized_procedure*, *close_date*, *causes_deadline_change*, *causes_price_change*, *document_id* and *document_name*.

Table 4 - Number of Missing Values per Feature

Feature	Missing Values
name_contestants	977557
invitees	2432959
centralized_procedure	365001
close_date	1841370
causes_deadline_change	2430846
causes_price_change	2315586
document_id	1343020
document_name	1343020
execution_location* *dictionary with missing values for the 3 keys	Country: 68162 Municipality: 334759 District: 376457

For the first feature, missing values were not an issue as they indicated non-competitive procedures without contestants. Similarly, the invitees feature had expected missing values since only certain procedures involve direct invitations. The *centralized_procedure* column was only relevant for specific multi-entity contracts, hence its occasional emptiness.

For the *close_date*, missing values were assumed to be due to a lack of caution when inserting the information, as every contract should have a close date. These were set to the *signing_date* plus the *execution_deadline*. Columns related to deadline or price change causes were also expected to have empty values since not all contracts undergo these modifications.

Missing *document_id* and *document_name* were not problematic as these variables were not used in the analysis, and their absence did not impact data quality. For the *execution_location*, while the country was set to "Portugal," filling in district and municipality accurately was not feasible and thus remained empty.

On step two, a check was made to verify the inexistence of duplicate records. During this check, many repeated *contract_id* showed up, but this does not constitute any problem, since each contract has an entry for every entity that applied to the respective procedure. In that regard, the same *contract_id* can appear many times, if the remaining information is not the same, which is the case.

Moving to step three, dealing with outliers is a sensitive topic on a dataset of this nature. In fact, due to the tight supervision and validation that these contracts are subject to, it is not expected that we find many outliers coming from human or system errors. For that reason, while the price related variables presented some extreme values with abnormal distances to other data points, the decision was not to correct them in bulk as outliers, but rather look at them individually in step number five. The *execution_deadline* variable, however, is a bit different and more prone to someone inserting a random or wrong number just to fill the field. Because of that, the variable was treated by replacing the top 0.1% of values with the corresponding value at the 99.9^{th} percentile.

On step four, all dates were converted into a datetime format, so that subsequent analyses can more easily extract information such as the specific month or day of the contract. As for the monetary variables, they were kept as floats to ensure the precision and exactness this format allows for.

Regarding the fifth step, some inconsistencies were found in the contract dates, as slightly more than 700 contracts had a *signing_date* posterior to the *close_date*. To avoid dropping them, and since only a small portion was affected by this issue, the decision was to alter the *close_date* so that it matched the *signing_date* plus the *execution_deadline*, following the same logic applied for the missing dates to ensure consistency. Some contracts also had errors in their price columns, which were detected by observing deviant points in the boxplots and histograms. To address this issue, a comparison to the original data source was made, which revealed that some values were, in fact, wrong. These discrepancies can likely be justified by mistakes at the time of the data extraction that were only detected and corrected online later, and they were fixed by altering the values to the ones currently shown on Portal BASE.

At this stage, it was also decided that when doing the analysis of the entities, their respective fiscal numbers would be used instead of the names. Although the names could be more intuitive, using numbers ensures an efficiency gain in processing data, as processing numbers is faster than processing text strings. Another reason to use the fiscal number instead of the name is because the fiscal number is constant through time, whereas the name changes a lot in some cases, which could jeopardize the analysis. An example of that is the pharmaceutical company Generis, which is always identified by the fiscal number 508107997, but presents several names that are incongruent such as: Generis, SA; GENERIS - FARMACÊUTICA, S. A.; Generis – farmacêutica; GENERIS - FARMACEUTICA SA; GENERIS FARMACEUTICA, SA.

As for scaling, this step is only done immediately before applying the anomaly detection algorithm that requires it (Local Outlier Factor) and then for the different clustering methods, as those are the only stages of this study that require data to be normalized. For those steps, it is crucial to do the normalization of the data, as the variables were measured in different scales, and so it is necessary to adjust the values to a common scale, since models assume that distances in different directions of the input space have equal importance. For the effect, the z-score standardization method was used.

3.3 Data Exploration

The Public Procurement landscape is not static: it evolves over the years. This evolution is characterized by fluctuations in the frequency of different procedure types, with some becoming more prevalent while others decline in frequency, reflecting shifts in government priorities, changes in regulations, advancements in technology, and evolving market conditions. For these reasons, an analysis of irregularities must consider this dynamism and employ time filters to focus on recent trends.

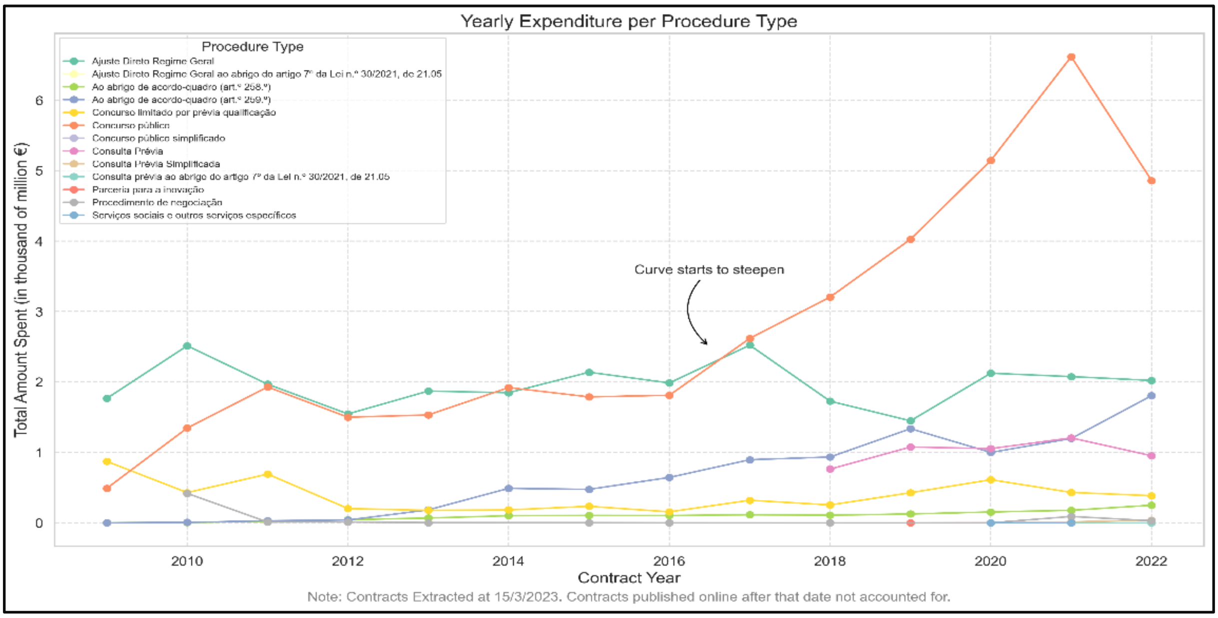


Figure 3 - Yearly Expenditure per Procedure Type

To provide a focused and relevant analysis, this study will concentrate on the last five years of contract data (2017-2022), and not the entire period starting from 2009. This contemporary period shows a clear trend in spending patterns and procedure types, with Public Tenders being on the rise. Thus, focusing only on it allows for a more accurate and up-to-date understanding of current practices and potential irregularities.

An analysis on a global scale, both in terms of the number of contracts and the total expenditure over the years, also justifies a greater focusing on the last five years. Particularly in what concerns the amount spent in public contracts, it seems to be a before and after 2017, with more money being spent each year ever since then.

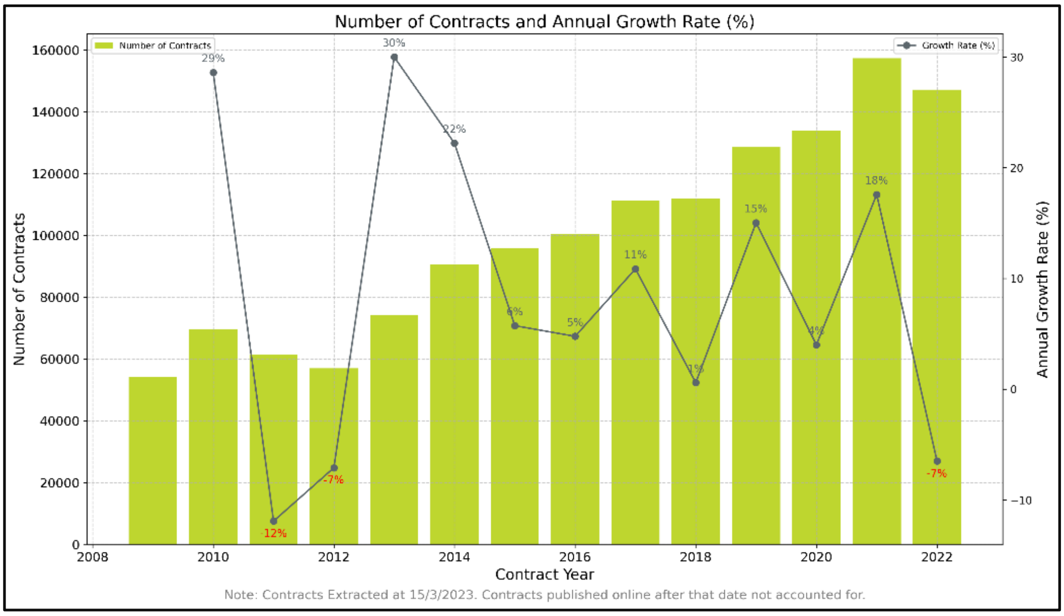


Figure 4 - Number of Celebrated Contracts

Moreover, there has been a steady increase in the number of public contracts that Portugal completed over the years. Besides the years 2011 and 2012, which were heavily affected by the economic crisis in Portugal, every other year has seen an increase of the number of contracts comparing to the previous year.

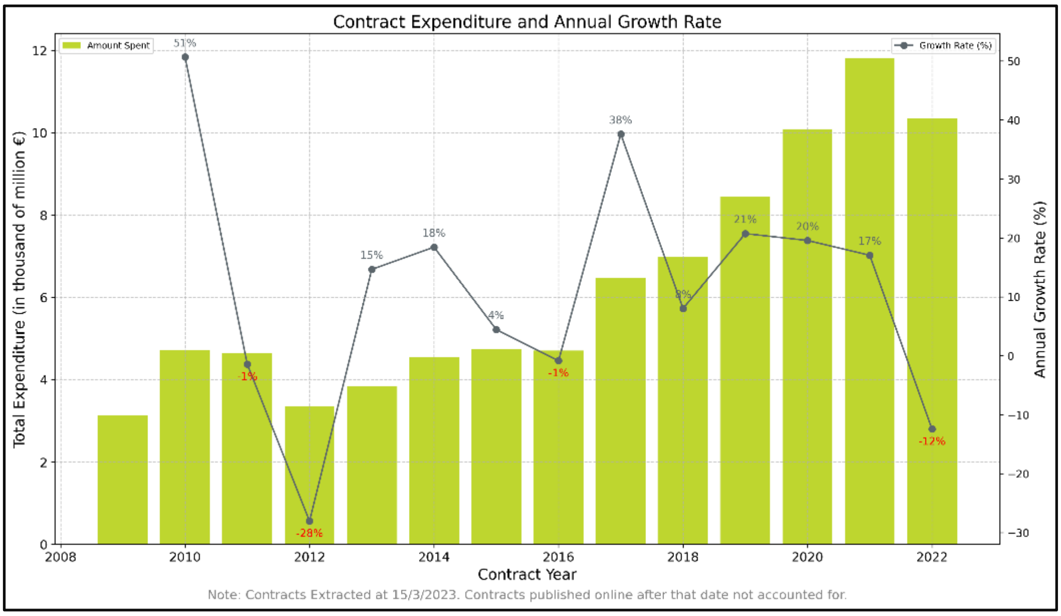


Figure 5 - Contract Expenditure

Regarding the amount of money involved in the contracts, the trend is not much different, with the only significant drop taking place in the year of 2012. The drop in the number of contracts seen in 2011 did not correspond to a drop in the total amount spent on those same contracts. What that means is that in 2011 there were less celebrated contracts than in 2010, but the celebrated contracts were of a higher value compared to the previous year.

3.4 Anomaly Detection at Contract Level

To identify the contracts that exhibit a different behaviour compared to the rest and rank them according to how anomalous they are, anomaly detection algorithms will be experimented over the almost eight hundred thousand contracts celebrated between the period of 2017 to 2022.

Anomaly detection algorithms are commonly employed for tasks in the fields of finance, cybersecurity or e-commerce, where there is a vast tray of transactions and logs, and it is usually important to be able to identify unusual patterns and abnormalities (Chandola et al., 2009). They are also helpful for cases like this one, since they can take contracts characteristics and separate apart those that have intrinsically different characteristics from the rest. They can belong both to the supervised and unsupervised learning domains, or even semi-supervised (Goernitz et al., 2013). Although their purpose is similar, which is to identify instances that deviate from the norm followed by most of the data, the way in which they operate is different. With supervised learning, the algorithm knows *a priori* which instances are anomalies and which aren't. In unsupervised learning, there is no label indicating what is an anomaly and what isn't, so the algorithm must be capable of making that distinction based on the characteristics of the data. It is useful for the cases in which anomalies are rare and unknown, which is what we have, thus being the method of choice.

When it comes to the exact algorithm, there are a lot of algorithms that can be used for anomaly detection (Agrawal & Agrawal, 2015). Some of these algorithms, like Local Outlier Factor, are more well suited to detect local anomalies, which consist of data points that deviate from the normal behaviour of their immediate neighbourhood or local region in the feature space. Others, like Isolation Forest, are better at detecting anomalies on a global scale, making it a more adequate method when we want to detect instances that behave abnormally when considering the entire dataset (Bouman et al., 2023). Not knowing exactly the nature of the anomalous contracts, the two algorithms will be experimented and tested, to ensure that anomalies are being captured both at a local and global level.

For these two anomaly detection algorithms, it is necessary to go through a feature selection process, since they only work with numeric variables and the presence of irrelevant features can negatively impact their performance.

In that regard, a difficult part of applying anomaly detection in a dataset of this nature lies in the fact that most features are categorical, and anomaly detection methods have more difficulties in dealing with this kind of variables since measuring distances between categorical observations is not as straightforward as it is with their quantitative counterparts. Identifying the statistical distribution followed by most observations, then considering the observations that do not follow that distribution, is a harder task with this kind of variables (Taha & Hadi, 2020), thus a choice being made to avoid categorical variables in the anomaly detection process.

Considering these limitations, some feature engineering was necessary to achieve a small yet representative of the contract number of features, listed below.

Award per Winner: This feature represents how much money each winning entity received. Contracts with exceptionally high award amounts per winner might signify anomalies.

$$award\ per\ winner = \frac{contract\ final\ price}{number\ of\ winners}$$

Number of Non-winning Bidders: This feature provides insight into the competitiveness of the bidding process. Contracts with an unexpectedly high number of non-winning bidders might be flagged as anomalous. For reference, Appendix E3 - Number Of Contestants Per Procedure outlines what is considered a normal range for the number of contestants per procedure.

$$number\ of\ non-winning\ bidders = \frac{number\ of\ bidders}{number\ of\ bidders - number\ of\ winning\ bidders}$$

Number of Winners: A feature that captures the number of winners for each contract. Anomalies could arise from contracts with an unusually high number of winners. For reference, Appendix E4 - Number of Winners Per Procedure outlines what is considered a normal range for the number of winners per procedure.

Price Difference: This feature captures alterations in contract values during their execution, a significant concern when analysing the contracts, as contracts with big changes in their price are often a cause for suspiciousness.

$$price\ difference = final\ contract\ price - initial\ contract\ price$$

Fast Conclusion: This feature might indicate contracts that were concluded unusually quickly or way after the expected, which could be indicative of anomalies.

$$fast\ conclusion = execution\ deadline\ days - actual\ contract\ duration$$

Price Ratio to Median Procedure Type Price: Represents the ratio of the *award_per_winner* for each contract to the median *award_per_winner* within its corresponding procedure type. For reference, Appendix E5 - Final Price Distribution Per Procedure Type contains the log-transformed final prices for different procedure types in the year 2022.

$$price\ ratio\ to\ procedure\ type = \frac{contract\ price}{median\ procedure\ type\ price}$$

Price Ratio to Median CPV Group Price: Represents the ratio of the *award_per_winner* for each contract to the median *award_per_winner* within its corresponding CPV (Common Procurement Vocabulary Code) group. For reference, Appendix E6 - Number of Contracts and Expenditure Per CPV Division outlines which divisions account for the most spendings.

$$price\ ratio\ to\ CPV\ Group = \frac{contract\ price}{median\ CPV\ group\ price}$$

Price Ratio to Median Contract Type Price: Represents the ratio of the *award_per_winner* for each contract to the median *award_per_winner* within its corresponding contract type.

$$price\ ratio\ to\ contract\ type = \frac{contract\ price}{median\ contract\ type\ price}$$

Time Between Contract Finish and Publication on Portal BASE: Once the object of the contract is concluded, it is mandatory for all contract-related information to be uploaded to Portal BASE. However, sometimes this information is not posted immediately, which may be cause for concern.

$$time\ between\ finish\ and\ publication = contract\ publication\ date - contract\ finishing\ date$$

Percentage of District Expenditures: This variable captures how much of the total district's expenditures a given contract represents in percentage. When the number of habitants grows, the amount spent per person on public contracts tends to decrease because procurement activity doesn't increase at the same rate as the population (Curado et al.). For reference, Appendix E7 - Expenditure Per District and Procedure Type and Appendix E8 - Number of Contracts and Expenditure per District per 10k Habitants, outline a breakdown of spendings per district.

$$percentage\ of\ district\ expenditures = \frac{contract\ price}{total\ expenditures\ of\ the\ district}$$

Once defined the group of contracts to be analysed and the ten features that will be used in that analysis, it is time to apply the set of algorithms. For the two algorithms, the chosen distance metric will be the Euclidean distance. Right below, a deeper dive into the specificities of each algorithm is made.

Local Outlier Factor (LOF) calculates the average distance of every data point to their k^{th} nearest neighbours, and then calculates the average distance of those neighbours to their respective neighbours. A point is seen as anomalous if it is a lot more far apart from its neighbours than they are from each other. An interesting property of LOF is that instead of merely classifying an instance as anomalous or not, it assigns to each object a degree of being an outlier, and this degree is what is called the local outlier factor (Breunig et al., 2000).

Isolation Forest (IF) is an ensemble of decision trees used for anomaly detection. It isolates data points by randomly picking a feature and then a random value for that feature within its limits. After repeating this process multiple times on different subsamples of the dataset, the average number of partitions it takes to isolate an observation is calculated over the multiple random trees. For anomalous observations, it is expected that this value is much smaller than for common observations, since they are easier to isolate from the remaining data points than those that are normal and therefore closer to one another. One advantage that this algorithm has over some others is that it is not as negatively affected by the introduction of irrelevant features (Emmott et al., 2016).

Despite some obvious differences between the two algorithms, one thing that will be kept the same for both is the contamination factor, set to 1%. The contamination factor is the parameter that allows us to specify the amount of contamination in the dataset, that is, the proportion of outliers/anomalies. Despite the possibility of applying some domain knowledge to estimate this value, it is impossible to know for sure what the true proportion is and set a threshold of what defines anomaly based on that (Perini et al., 2023). Setting it to 1% will make the algorithms identify approximately 1% of the data points with the most extreme characteristics as anomalies, and considering that the dataset has around eight hundred thousand different contracts, it means that around eight thousand will be returned, a decent base for human auditors to look further into without being a very big or very small number of contracts. The two algorithms are compared in terms of some of their properties in the table below.

Table 5 - Comparative Analysis of Anomaly Detection Algorithms

	Local Outlier Factor	**Isolation Forest**
Type	Density Based	Ensemble of DT
Speed	Slower than IF	Faster than LOF
Memory	Requires moderate memory space	Does not require much memory space
How well they scale to large datasets	Worse than IF	Better than LOF

To evaluate these two methods, a two sample Kolmogorov–Smirnov nonparametric test will be performed, comparing the distributions of each feature in the original population of contracts with the distributions of the same features in each of the two anomalous samples drawn from it (the one taken from LOF, and the one taken from IF). The higher the KS Statistic of each feature, the most its distribution deviates from the distribution of the original population, meaning that the algorithm did a good job at picking anomalous instances. Therefore, for each of the ten previously established features, the KS statistic between the distribution of the feature in the original population and the distribution of the feature in the two anomalous samples will be computed, and the selected method will be the one with the bigger number of features with a higher KS statistic.

Finally, once defined the algorithm that best identifies anomalous contracts and obtained those data points ranked in terms of anomaly score, SHAP (SHapley Additive exPlanations) will be used to explain why each point is considered anomalous and which features contributed to the classification made by the model. SHAP is a framework that is based on Shapley Values, a concept from game theory introduced in 2017 by Lundberg and Lee, and its purpose is to explain how much each variable contributes to the output of the model.

Applying SHAP is a pivotal step to finalize the anomaly detection process, as it offers information on the relative contributions of each attribute to the model's predictions, which aids in deciphering algorithmic behaviour and comprehending the traits of anomalies the model has detected. In a way, it is telling the answer to "why was this output produced from this input?", a key question to ask in every data science project.

Some important aspects that must be considered at every step are explainability and interpretability to ensure transparency, compliance with regulations, fairness, continuous improvement and trust. If a given contract is signalled as potentially interesting to further investigate, it is important to understand which factors caused that and what are the underlying reasons that made the contract be signalled. This kind of considerations are particularly important in sensitive topics like this one, where it is pivotal that the results of the model are trusted by all stakeholders including regulators, lawmakers, and the users that interact with the system.

3.5 Creation and Calculation of Red Flags at Entity Level

While the data on Portal BASE and the selected features for anomaly detection are useful for evaluating contract legitimacy, they do not provide a full context, particularly regarding the entities involved. For example, a single-bidder procurement process might not seem notable in isolation. However, if we observe that the same government and private entities frequently engage in single-bidder processes together, this could raise concerns.

To achieve a more nuanced understanding of the contract's legitimacy and the companies that partake in it, it is possible to identify several red flags that usually indicate reasons to concern at an entity level. A great source for this information was the work developed in Red Flags Listed by Project Cycle (International Anti-Corruption Resource Center, 2023), that combined into a single list divided by procurement stage detailed descriptions and examples for each red flag. In this context, a red flag refers to a hint, either direct or indirect, of possible illegal activity, identifying a chance for more examination of the contract and its entities that could ultimately bear out: not at all illegal; not illegal but lacking in quality service delivery, competitiveness, or value for money; illegal.

Entities will be eligible for red flag analysis if they meet specific criteria: awarded entities must have won at least three contracts, with at least one post-2020, and awarding entities must have issued at least three contracts, with at least one post-2020. Each red flag will be calculated accordingly, with some flags being specific to awarded entities and others applicable to both awarded and awarding entities. This distinction will be indicated for each flag with the keyword "Side".

The following red flags are an attempt to consolidate the existing literature and my own ideas adjusted to the Portuguese Public Procurement context to provide data metrics that can be used to assist a proactive assessment of contracts and entities, translating the existing contractual metadata features to more concrete features correlated to illegal behaviours. If successful in the creation of these metrics, they can not only be extremely useful instruments for identifying and averting any problems in Public Procurement in the early stages, but also to identify general flaws in the procurement ecosystem and suggest technological or policy adjustments to strengthen the system.

3.5.1 Red Flag 1 – Percentage of participations in a public tender that has only 1 entity bidding

Side: Both

In Public Tenders, being it a competitive procedure, it is expected that multiple enterprises apply, which encourages greater competition and ideally allows the state entity selecting the winner out of a pool of contestants to get better value for the money. When there is just a single bidder, questions concerning the transparency and equity of the bidding procedure are likely to arise. This lack of competition may lead to fewer options, more expensive prices, and a lower chance of getting the finest goods or services for the money. For that reason, while in some cases having only one entity bidding can be expected (in very specialized/niche markets, for instance), it is probably a good idea to scrutinize these scenarios. Furthermore, this red flag also serves to capture the cases in which other corruption techniques are used to restrict competition. In fact, Wittberg and Fazekas (2023) show that firms winning a higher share of single bidder contracts tend to be more profitable, as the lack of competition allows these firms to charge higher prices and reduces the pressure to innovate or cut costs.

3.5.2 Red Flag 2 – Engaging in tenders with contract value altered during the execution of it

Side: Both

The alteration of the contract value during the execution phase of a project (after the terms were agreed upon) can raise significant concerns due to its potential to unfairly disadvantage other proposals that were submitted based on the originally terms. For instance, we can image a scenario where the budget for a given service was initially 50.000€ and companies A and B both submit proposals to provide that service. Company A offers to provide it for 45.000€ and company B will do it for the budgeted 50.000€. In normal conditions where no other factors differentiate the two bids, company A is selected for the job based on its cost-effective proposal. However, as the project progresses, it becomes apparent that unforeseen complexities and additional requirements have arisen, which were not initially considered, and the contact value is adjusted to 55.000€. In this case, the company that made a better judgement of the money that had to be spent – company B – is undermined by a situation that penalizes their diligence. Negative values are acceptable in this flag as they indicate cases where the final price is lower than the initial price. Though rare, these situations must be considered.

3.5.3 Red Flag 3 – Win share (much) greater than average of sector

Side: Awarded

This red flag refers to circumstances where a specific bidder is selected as the winner of a tender with a success share that is higher than the normal for that sector – sector here being defined as the CPV group, given by the first three code digits. At the point when an organization reliably gets a success share a lot more prominent than the sector's normal, it can raise worries about the fairness of the procedures. Such a situation may indicate potential favouritism, unfair advantages, or anti-competitive behaviour that could undermine the integrity of the sector. To calculate this red flag, we must divide the number of times a given entity won a tender by the number of times that same entity was a contestant in a tender, and then compare the obtained value to the average/expected win rate of other companies in the same sector. Since companies can win contracts in multiple CPV groups, we first calculate each group's win share (winners divided by contestants). Then, we calculate each entity's win share in that group. Finally, we average each entity's win rate compared to the sector average, considering for that the number of procedures they entered in each group.

3.5.4 Red Flag 4 – Often engaging in non-competitive procedures

Side: Both

A non-competitive procedure in which the winning firm is directly contacted by the contracting authority can constitute a red flag on its own, as it can be seen as a way of running away from the normal competition of competitive procedures, since these procedures are sometimes less scrutinized and more easily manipulated.

3.5.5 Red Flag 5 – Recurrent excessively fast conclusion of the contract

Side: Awarded

When contracts are awarded, a specific deadline, measured in days, is stipulated for the winning party to complete the service or deliver the contracted good. If the contract is completed much faster than this date it can indicate that the proper due diligences were not considered or that there is a chance the bidding procedure was only ceremonial and that the winner was chosen in advance, which might be seen as favouritism or collusion.

3.5.6 Red Flag 6 – Contract price (much) higher than sector average

Side: Both

Considering that services or goods provided within a sector are usually relatively similar between them, it is expected that the prices also do not have great variation. In that sense, a price significantly above the sector average might suggest overvaluation or overpricing of services, raising concerns about inflated costs or potential financial irregularities.

3.5.7 Red Flag 7 – High percentage of contracts celebrated between a contestant and an awarding entity

Side: Both

Looking into the proportion of contracts awarded by state-entities can be an important indicator to understand if a significant proportion of contracts is consistently awarded to a particular contestant by the same entity. On the bidder side, if much of the revenue comes from a single source, this might indicate a conspicuous relationship with an issuer.

3.5.8 Red Flags suggested for Future Work

The red flags above constitute a good starting point when it comes to analysing suspicious behaviours of companies involved in the procurement landscape. However, more red flags exist that, if implemented correctly, can enrich the study and provide additional information. In that regard, the following red flags are a suggestion for future research in this field, as the data that was available for this study did not allow for the calculation of them.

Red Flag 8 – Participation in tenders with short period to submit bids

Side: Both

Giving bidders a short amount of time to submit their proposals might prevent them from finalizing the necessary preparation work that comes involved with doing such proposal, like managing bureaucratic requirements or conducting thorough research. Moreover, it might provide a few chosen bidders who receive early notice of the tender an unfair competitive advantage, and it can even be seen as a sign that the decision to award a given company was premeditated and that other offers were not properly analysed.

Red Flag 9 – Frequent withdrawal from competitive procedures during the tender phase

Side: Awarded

Entities that on a regular basis withdraw from competitive processes without detailing plausible reasons to do so may deserve a closer look, as an abrupt exist can be a possible sign of bid manipulation or collusion with other entities. While there may be legitimate reasons for withdrawal, such as unforeseen events or changes in the vision of the company, unexplained patterns of withdrawal are one more thing that warrants further investigation.

Red Flag 10 – Engaging in tenders with long eligibility criteria and prevalence of specific requirements

Side: Both

By including a lot of restrictions to the tender process, such as previous experience or specific equipment and infrastructure, the awarding entities can limit the companies who can participate and, in that way, exclude those entities they do not want to win even before they apply, which is contrary to the principles of equality and non-discrimination and increases the difficulties small and medium enterprises feel (Nicholas & Fruhmann, 2014).

3.5.9 Red Flags Summary

To further clarify the red flags and their significance, all explanations about the side they belong to, their meaning and their construction is centralized in Table 6.

Table 6 - Red Flags Construction

Red Flag	Name	Side	Construction
1	Percentage of participations in a public tender that has only 1 entity bidding	Both	$rf1 = \frac{tenders\ with\ 1\ bidder\ won}{tenders\ won}$
2	Engaging in tenders with contract value altered during the execution of it	Both	$rf2 = \frac{revenue\ from\ alterations}{total\ revenue\ without\ alterations}$
3	Win share (much) greater than average of sector	Awarded	$rf3 = \sum_{i=1}^{n} w_i(avg\ win\ share - avg\ sector\ win\ share)$ where n is the number of sectors an entity won a contract in, defined by the first three digits of CPV code, and w_i is the weight given by the number of times the entity was a contestant in that sector
4	Often engaging in non-competitive procedures	Both	$rf4 = \frac{non-competitive\ procedures\ won}{procedures\ won}$
5	Recurrent excessively fast conclusion of the contract	Awarded	$rf5 = avg(execution\ deadline\ days - actual\ contract\ duration)$
6	Contract price much higher than sector average	Both	$rf6 = \sum_{i=1}^{n} w_i \frac{avg\ contract\ price}{avg\ sector\ price}$ where n is the number of sectors an entity won a contract in, defined by the first three digits of CPV code, and w_i is the weight given by the number of times the entity won a contract in that sector
7	High percentage of contracts celebrated between a contestant and an awarding entity	Both	$rf7 = \frac{\max(procedures\ won\ with\ a\ state\ entity)}{procedures\ won}$

Table 7 contains some descriptive statistics helpful for understanding the red flags on the awarded side, and Table 8 contains that same information but for the awarding side.

Table 7 - Descriptive Statistics of Constructed Red Flags for the Awarded Side

Red Flag	mean	std	min	25%	50%	75%	max
1	0.48	0.39	0	0	0.5	1	1
2	-0.01	0.14	-0.89	-0.0009	0	0	12.3
3	0.17	0.22	-0.86	0.03	0.18	0.33	0.73
4	0.69	0.29	0	0.37	0.62	0.86	1
5	-11	48	-777	-17	-0.16	0	852
6	0.84	3	0	0.26	0.49	0.86	269
7	0.54	0.3	0.01	0.28	0.5	0.8	1

Table 8 - Descriptive Statistics of Constructed Red Flags for the Awarding Side

Red Flag	mean	std	min	25%	50%	75%	max
1	0.51	0.33	0	0.3	0.5	0.75	1
2	-0.01	0.09	-0.63	0	0	0	3.05
4	0.55	0.27	0	0.34	0.56	0.75	1
6	0.94	3.19	0	0.28	0.51	0.85	93.27
7	0.22	0.19	0.01	0.08	0.17	0.33	1

Once calculated the continuous values of the red flags for each entity, the final step is to see for each company on the awarded and awarding sides how many red flags they have with values above the 75^{th} percentile. This analysis helps identify companies that may be associated with a higher-than-average number of red flags, indicating potential irregularities in their contract activities, which is crucial information that will be later used in conjunction with the anomaly scores. As an example, a company in the awarded side will be signalled for the red flag number four (high ratio of non-competitive contracts) if their percentage of participation in these contracts is greater than 86% - the value of the 75^{th} percentile.

3.6 Final Metric Calculation

Only at this point we can merge the information extracted from the anomaly detection process and the calculated red flags, achieving a complete profile for each contract flagged as anomalous, including its anomaly score and the number of red flags for the involved awarded and awarding entities that exceed the 75^{th} percentile for each flag.

A comprehensive understanding of the contracts, the entities and their patterns is made possible by this metric, which combines more particular signs of possible irregularities in the red flags with overall contract anomaly, hopefully achieving a balance of different kinds of information that allows to capture the different kinds of illicit actions that might not be evident when considering each type of information separately. To merge all these information into a single metric that serves as an estimator of a contract's irregularity, the following formula will be used:

$$combined\ metric = anomaly\ score * (red\ flags\ above\ 75^{th}\ percentile\ awarded\ side^{*} + red\ flags\ above\ 75^{th}\ percentile\ awarding\ side)$$

*if there is more than one awarded entity, the average number of their red flags above the 75^{th} percentile will be used

Importantly, since the scores generated by the chosen anomaly detection methods will be negative for anomalous cases, the combined metric is inversely related to the need for investigation: the lower the value (or the larger the modulus), the higher the priority for examining the contract for potential irregularities.

3.7 Cluster Creation and Analysis

By clustering the identified anomalous contracts together, we can uncover different subgroups of contracts, and understand if any of these groups are linked to irregularity to a greater extent than others.

Before moving to the actual cluster creation, an importation step is to evaluate if the dataset has the necessary characteristics to create meaningful clusters. Clustering methods will always return clusters, but those clusters could just be arbitrary structures that agglomerate data points together with no reasoning (Cross & Jain, 1982). In that regard, to measure the clustering tendency of the created dataset, the Hopkins test will be used. This statistical test yields a score, which ranges between 0 and 1, with scores closer to 0 meaning that the data is not uniformly distributed and therefore clustering can be applied. Once the Hopkins test is passed it will be possible to identify trends, patterns and anomalies, thus allowing an even greater scrutiny over specific anomalous contract subgroups. To do so, this stage includes the experimentation of the different clustering techniques Hierarchical Clustering, K-Means and K-Medoids, Fuzzy C-Means and Mean-Shift.

Each algorithm has unique strengths and weaknesses, making it crucial to experiment with various algorithms and their parameters. The following table provides a high-level summary of these differences.

Table 9 - Comparative Analysis of Clustering Algorithms

	Hierarchical Clustering	**K-Means/K-Medoids**	**Fuzzy C-Means**	**Mean-Shift**
Type	Agglomerative (bottom-up) or Divisive (top-down)	Partitioning	Partitioning	Density-based
Speed	The slowest	The fastest	Slower than K-Means	Slower than Fuzzy C-Means
Memory	Requires a lot of memory space	Requires moderate memory space	Requires moderate memory space	Requires a lot of memory space
How well they can handle different cluster shapes	Struggles with irregular shapes	Struggles with non-convex shapes	Can handle irregular shapes	Suitable for irregular shapes
How well they scale to large datasets	Not well	Better than Hierarchical Clustering	Worse than K-Means	Similar to Fuzzy C-Means

After experimenting with different clustering techniques, it is also importation to perform the evaluation of those same clusters and their visualization in two dimensions – for which we will employ t-SNE and UMAP, with UMAP offering more stable results across multiple runs (Becht et al., 2018). This allows us to validate the created clusters and ensure that they make sense within the context of the problem. To evaluate the quality of the created clusters choosing a good evaluation metric is fundamental (Palacio-Niño & Berzal, 2019), so the following ones will be taken into consideration:

Table 10 - Cluster Evaluation Metrics

Metric	Description	Formula	Objective
Silhouette Score	The silhouette score measures the average distance between a data point and the other data points within the same cluster (intra-cluster distance) and the average distance between that same data point and the data points that belong to the nearest cluster.	$$s = \frac{b-a}{\max(a-b)}$$ Where a is the average distance between a data point and the other data points within the same cluster and b is the average distance between that same data point and the data points that belong to the nearest cluster	Max (between -1 and 1)
Calinski-Harabasz Index	The Calinski-Harabasz Index measures the inter-cluster dispersion against intra-cluster dispersion. Analysing this index helps in the task of determining how well the clusters are separated and how compact the clusters are.	$$CH = \frac{\left[\frac{\sum_{K=1}^{K} n_k \lVert c_k - c \rVert^2}{K-1}\right]}{\left[\frac{\sum_{K=1}^{K}\sum_{i=1}^{n_k} \lVert d_i - c_k \rVert^2}{N-K}\right]}$$ Where K is the number of clusters, N is the total number of observations, n_k and c_k are the number of points and centroid of the k^{th} cluster respectively, and c is the global centroid	Max (between 0 and no upper bound)
Davies-Bouldin Index	The Davies-Bouldin Index also uses the concepts of inter and intra cluster distances, but to calculate the similarity measure of each cluster with the cluster most similar to it.	$$DB = \frac{1}{K}\sum_{i=1}^{k} \max\left(\frac{\Delta(X_i)+\Delta(X_j)}{\delta(X_i,X_j)}\right)$$ Where K is the number of clusters, $\Delta(X_k)$ is the intracluster distance within the cluster X_k and $\delta(X_i,X_j)$ is the intercluster distance between those two clusters	Min (no bounds)

For each cluster, concrete characteristics of the contracts within it will be listed, leading to specific actions of investigation that can be taken to better understand what the main issues for that group of contracts are. Moreover, the average value of the previously calculated final metric will be obtained for each cluster, with which it will be possible to prioritize even more which are the contracts that the human auditors must investigate first.

4. RESULTS AND DISCUSSION

In this section the results obtained from the methods discussed in the methodology part will be presented and dissected. Although the different steps end up culminating in a common final analysis, they will first be discussed individually, allowing for a granular understanding of each element before a complete integration into the comprehensive final analysis. The results presented next result completely from the analysis of the data, lacking any kind of political bias or prior knowledge of the involved entities.

4.1 Anomaly Detection Results

Applying the two unsupervised anomaly detection algorithms – Local Outlier Factor and Isolation Forest – to the dataset of the eight hundred thousand contracts celebrated in Portugal between the years of 2017 to 2022 with a contamination factor of 1% resulted in two much smaller datasets with around eight thousand contracts each. To evaluate the results, the Kolmogorov-Smirnov test was applied to the output datasets, and that way it was determined that the best method was the Isolation Forest, since it was the one with the higher discrepancies in terms of distributions to the original dataset observed across the various features (notably higher KS Statistic values and the associated p-values closer to zero). The results of this statistical test can be consulted in Appendix *F1 - Kolmogorov-Smirnov Test* Results.

On a first look, the anomalous contracts immediately revealed some interesting patterns, like the fact that they all had very high prices relative to their CPV group, procedure type and contract type median prices, or the fact that most of them did not conclude in the expected date and had significant price differences in respect to the initially established price. To further understand their characteristics, SHAP values were used, starting with a summary plot to understand the impact of each feature on individual predictions.

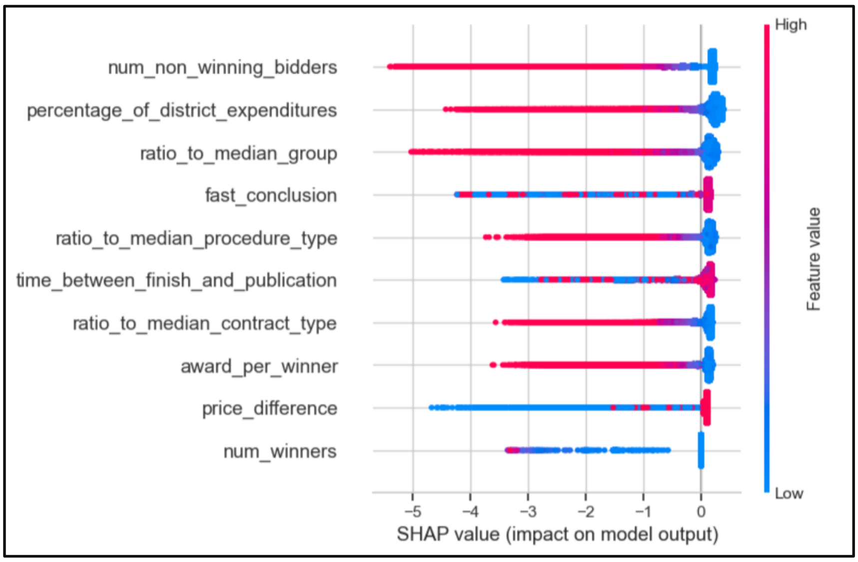


Figure 6 - SHAP Summary Plot for Anomaly Detection

In a plot of this kind, red indicates high values of the variable, whilst blue represents the smaller values. The lower the SHAP value on the x-axis, the greater the impact towards anomaly, since the decision function of the Isolation Forest method works in a way that the lower the assigned score, the bigger the anomaly. In that regard, one can conclude that:

- **Number of Non-Winning Bidders**: the greater the number of non-winning bidders, the more anomalous the contract. This indicates that most contracts have few bidders.
- **High Monetary Involvement**: contracts with a lot of money involved in them tend to be anomalous. High values of award per winner, percentage of district expenditures and the different ratios to the median prices of the CPV group, procedure type and contract type reveal a strong tendency towards the anomaly side.
- **Price Differences:** negative price differences, meaning that the contract value was altered to a smaller value that the initial one, are also associated with anomaly. This makes sense if we think that this behaviour is not very common and that changes in price usually make the contracts more expensive than they were, not cheaper.
- **Deviation from Expected Completion Times:** the contracts that finish way before or way after the expected time are seen as anomalous. The same can be said for the time it takes to post the contract to Portal BASE. Timelines out of the ordinary are a big reason for concern when investigating these contracts.
- **Number of Winners:** the number of winners in a contract does not offer much differentiation or insights into contract anomaly. However, a small fraction of contracts with high number of winners is linked with anomaly.

On a more detailed level, it is possible to analyse individual contracts and understand what features made them be classified as normal or anomalous. To do so, we use SHAP force plots. Below we analyse first a normal contract, and then one contract that was considered anomalous. In each plot, the red and blue "arrows" represent how much those features influence the final output of the model for the sample under study. A red "arrow" pushes for normality, whilst a blue "arrow" pushes for anomaly. The base value is the average of all output values of the model, meaning that if a given instance is above that value it will be considered normal, and if it is below that value it will be considered anomalous.

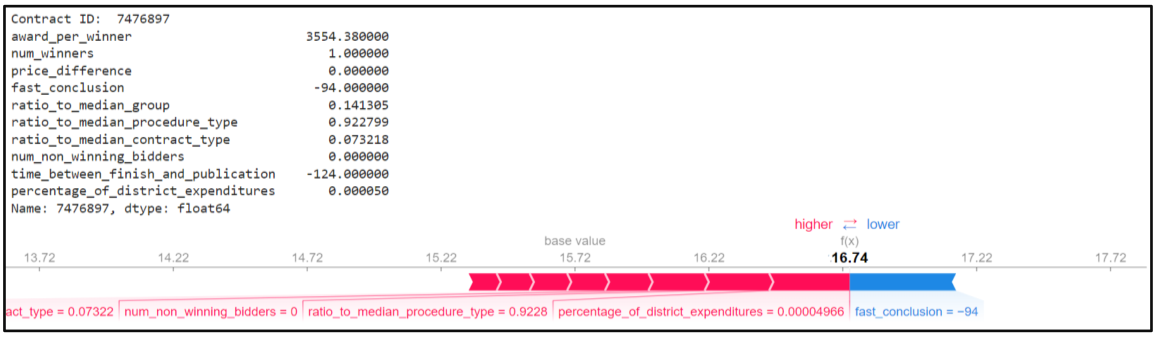


Figure 7 - SHAP Force Plot of a Normal Contract

In this first plot, everything is normal except for the *fast_conclusion* variable, that indicates that the contract was concluded 94 days after the expected. However, that variable alone is not sufficient to counterbalance all the other variables indicating normality.

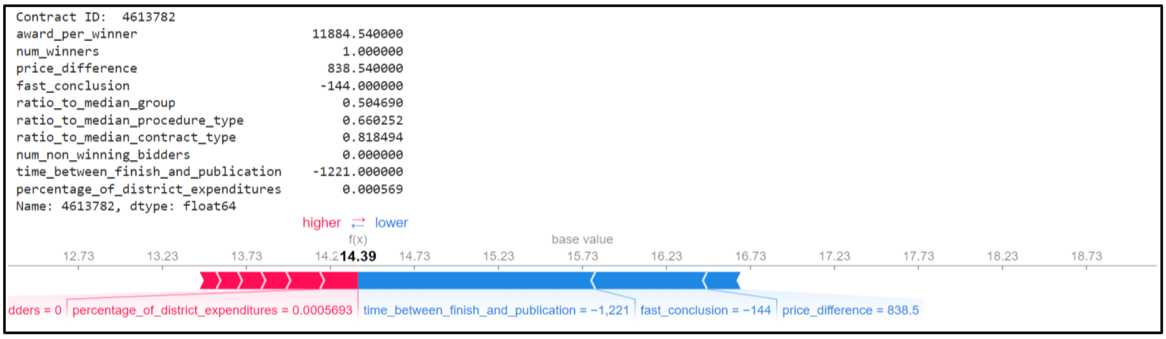


Figure 8 - SHAP Force Plot of an Anomalous Contract

In the second plot, we have more variables pushing the contract to the anomaly side, namely the time between the contract finished and the publication to portal BASE (it was posted 1221 days before its conclusion, which is more than 3 years), the price difference (800€ more than expected) and the difference between the expected and real conclusion dates (contract concluded 144 days after the deadline).

Every contract can be analysed with a SHAP force plot like the ones above, making it a great instrument to understand the factors that contribute to each single prediction.

Finally, SHAP can be used for to analyse feature importance. This analysis tells us how much each variable contributes to the model output on a more global scale, not looking into the individual predictions like the force plot or the feature values like the summary plot. In the plot below, the total of each feature's absolute Shapley values across the data is used to rank them, being the top features the ones that contribute the most to the model's output.

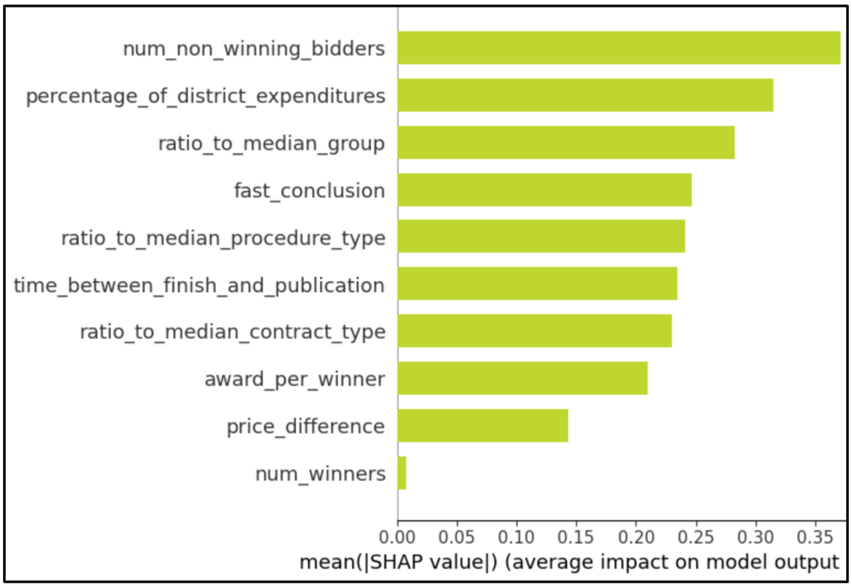


Figure 9 - SHAP Feature Importance

The least important feature (the one that offers less differentiation) appears to be the number of winners. This happens because the variable has the value one across almost all the contracts, since the number of contracts with multi winners is very reduced, as seen in the exploratory data analysis section. On the opposite side, the most important feature is the number of non-winning bidders in a contract, which is a variable that is highly indicative of the degree of competition that each contracts possesses.

To conclude, these anomalous contracts are characterized in terms of their procedure and contract types, district and CPV code. Out of the 7902 anomalous contracts, the most common procedure types are Public Tenders with 3556 instances, "Ao abrigo de acordo-quadro (art. º 259.º)" with 2344 instances and Direct Awards with 1237 instances. In terms of contract types, the most frequent are acquisition of movable goods with 3887 contracts, followed by concession of public services with 1941, and concession of public works with 1785. Geographically, the contracts are primarily concentrated in the districts of Lisbon (1666 contracts), Porto (923), Coimbra (568), Setúbal (430), and Aveiro (285). Finally, the top CPV divisions in terms of the number of anomalous contracts are 33 - Medical equipment, pharmaceuticals and personal care products with 2582 occurrences, 45 - Construction work with 1797 and 09 - Petroleum products, fuel, electricity and other sources of energy with 752.

4.2 Red Flags Results

In the red flags section, a total of 25249 companies on the awarded side and 3720 state entities on the awarding side were analyzed, as they were the ones that met the qualifying criteria. Once calculated the continuous value of each flag for each entity, the subsequent step entailed calculating for each entity how many flags they had above the 75th percentile. Despite it being an intermediary step that will later be integrated into the final metric calculation, the results of the red flags can be looked at on their own. In that regard, the next plots show for each side how many entities have each number of flags.

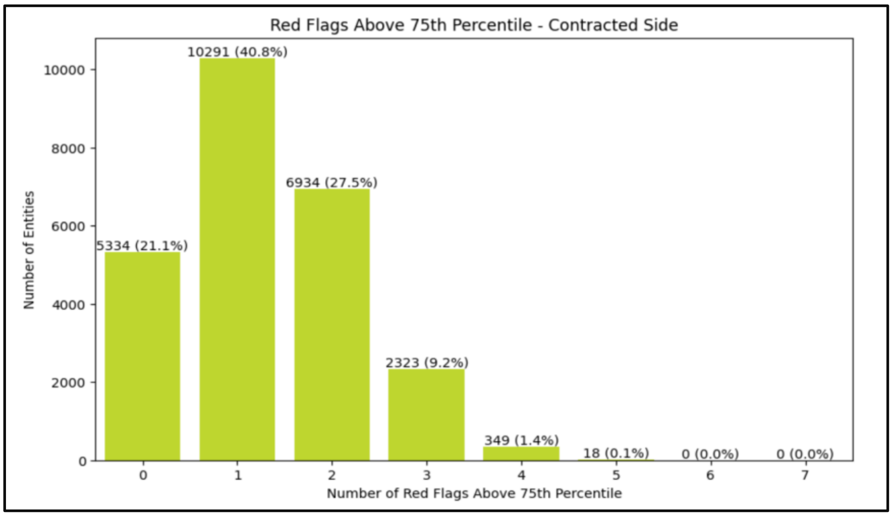


Figure 10 - Number of Red Flags per Contracted Entity

On the awarded side none of the companies have all the seven or even just six of the red flags with values above the 75th percentile. In fact, the percentage of companies that have as much as five red flags quite above the expected is still very small, and slightly bigger for four. The most common number of active red flags is just one, with almost 41% of the 25249 entities, which is a positive indicator for the healthiness of the Portuguese procurement landscape.

To validate the created red flags and see if they are representative of the misconducts that they are trying to capture, a list of entities that the *AdC* convicted of anti-competitive practices and collusion (cases opened between 2017 and 2022) was collected (Autoridade da Concorrência, 2024; Espírito Santo, 2022). The 19 private entities on that list, which belong to sectors such as security, health or communications, had on average 1.73 red flags with values above the 75th percentile, with the remaining non-convicted companies having on average 1.29 flags above that same percentile. The significance of this difference was tested with a non-parametric Mann-Whitney U Test, which resulted in a p-value of 0.028 that makes us reject the null hypothesis at the significance level of 0.05, supporting the hypothesis that the convicted entities have a higher average number of flags compared to the non-convicted entities.

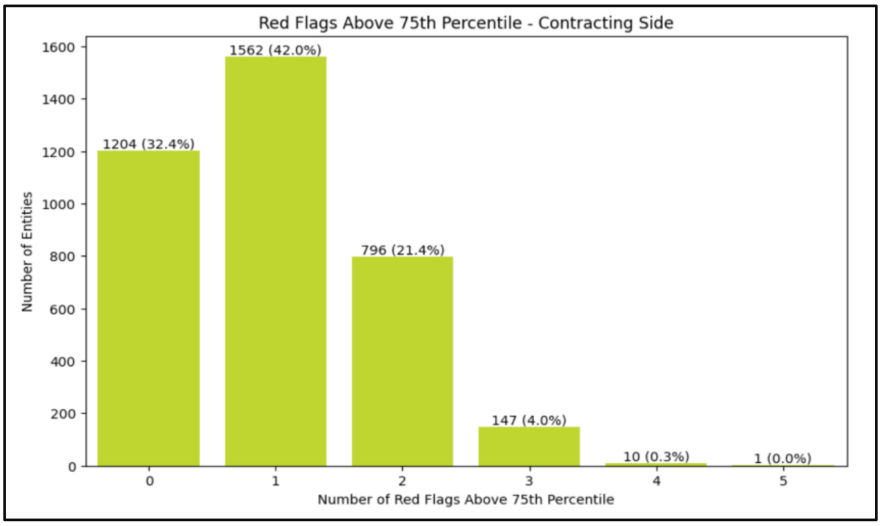


Figure 11 - Number of Red Flags per Contracting Entity

On the contracting side, the trend is not much different, with the bigger number of red flags having very few entities with a value above the 75th percentile, and around three fourths of the entities having as little as one or zero flags present.

4.3 Combined Results

The final step was to integrate the results of the anomaly detection and red flags stages into a single metric. This resulted in a final metric of suspiciousness for each contract, which can be used to rank the contracts in terms of neediness to investigate and that way help auditors in the task of narrowing down the contracts they must scrutinize.

Some of the contracts, despite having an anomaly score greater than zero, still exhibit a zero in the final metric. This happens because the entities involved in the awarded and awarding sides do not possess any red flag above the 75th percentile, so when we multiply the sum of red flags by the anomaly score, we end up with zero. Other contracts have entities with missing values for the number of red flags above the 75th, since those entities did not meet the necessary assumptions to be included in the red flag analysis. In those cases, only the anomaly score metric is used. The picture below shows the top 10 most suspicious contracts awarded in Portugal between 2017 and 2022, based on the defined metric.

contract_id	anomaly_score	nif_contracted	nif_contracting_agency	flags_above_75th_percentile_awarded	flags_above_75th_percentile_awarding	combined_metric
9635271	-0.157324	[500277443]	512012032	3.0	4.0	-1.101265
7333199	-0.153461	[500277443]	512012032	3.0	4.0	-1.074230
4672261	-0.141697	[500440131]	503933813	5.0	2.0	-0.991879
9400604	-0.153461	[503439800]	512012032	2.0	4.0	-0.920768
6763985	-0.150714	[500247480]	512012032	2.0	4.0	-0.904283
8743956	-0.145247	[502334967]	511284349	3.0	3.0	-0.871480
7622402	-0.172956	[514950994]	503933813	3.0	2.0	-0.864782
7375404	-0.170143	[514950994]	503933813	3.0	2.0	-0.850715
7300887	-0.141442	[512024979]	600087174	3.0	3.0	-0.848649
7642548	-0.139125	[502334967]	511284349	3.0	3.0	-0.834750

Figure 12 - Contracts Ranked by Final Metric (Top 10)

Framing these results with the existing literature on this topic is not easy, as the proposed approach has never been experimented in the exact same way, to the best of my knowledge. However, some methods, relying solely on anomaly detection, were able to obtain promising results in the field of detecting irregularities. In particular, the study from Niessen et al. (2020) is the one that most resembles this one, as they both rely on unsupervised learning and particularly in the Isolation Forest method. In fact, the two studies even share the fact that only the metadata of the procurement processes is used, and some of the chosen features are pretty much identical. Where the two studies diverge, is in the integration of red flags along with anomaly scores to derive a final metric which is done in this work. This added layer ensures that the entities involved are scrutinized comprehensively, and given the similarities between the two studies, it is to expect that the good results obtained in Paraguay can be replicated, if not enhanced, using the framework proposed here.

4.4 Clustering Results

Calculating the Hopkins statistic is crucial before creating clusters to assess the dataset's clustering tendency. This statistic for the anomalous contracts dataset was 0.1, validating the approach and confirming the data's suitability for clustering.

Once passed the Hopkins test and experimented the different clustering algorithms, Fuzzy C-Means was selected as it had the best metrics, which can be consulted in Appendix *F2 - Clustering* Metrics. Importantly, the number of clusters was set to four, which was the value that made most sense applying some domain knowledge, but it was also the one that several methods typically used to determine the number of clusters pointed to (Dendrogram from Hierarchical Clustering or the Elbow Method from K-Means).

To analyse the clusters constitution and find the root cause of why each contract is anomalous, we look into the mean values of each feature by cluster, and also the number of contracts that each of them have. In the left, where the standardized mean of each variable is plotted by cluster, there is also a line that corresponds to the non-anomalous contracts, helping clarify how each cluster's features deviate from the normal, putting the scaled values into perspective. One thing that immediately stands out is that all the anomalous contracts are more expensive that the normal ones, which can be inferred by comparing the four coloured lines to the dotted black line in the price-related variables.

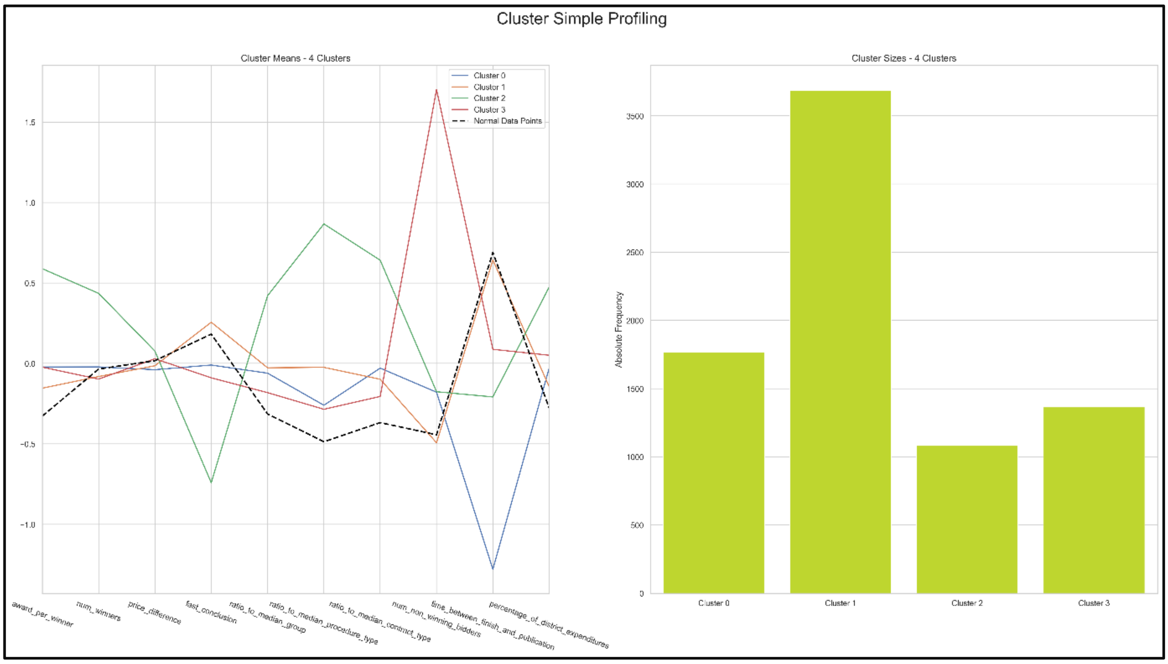


Figure 13 - Cluster Profiles

In terms of the intrinsic characterises of the clusters, Cluster 0 has a very low relative value of time between contract finishing and publication to portal BASE, meaning that these contracts are posted online before their completion.

As for cluster 1, relative to other clusters it has a smaller number of non-winning bidders, which might indicate that most contracts within this cluster are from non-competitive procedures. One thing that also corroborates that idea is the fact that it is the cluster with the smallest value of award per winner amongst the four anomalous clusters, something usually associated with non-competitive procedures. Moreover, contrasting with cluster 0, its contracts are posted online in a timely manner, with the line corresponding to this feature almost overlapping the one of normal contracts. Contracts within this cluster are also concluded a little bit quicker than expected.

Cluster 2 has the most expensive contracts, which is indicated not only by the award per winner variable, but also by the high price ratios comparing to the median prices of the procedure, contract and CPV group. These are also contracts with a high number of winners and contracts that finish after the time.

Cluster 3 is characterized by being the one with the biggest number of non-winning bidders, meaning that it is likely associated with competitive procedures. These contracts also get concluded after the expected, but not nearly as much as those in cluster 2.

All this information is condensed in the table below, alongside the average value of the calculated final metric and recommended investigation actions for each cluster.

Table 11 - Anomalous Contracts Clusters

Cluster	Contracts	Avg. Final Metric	Characteristics	Actions
Cluster 0 - Timeliness and Quality Assurance	1769	-0.163390	Contracts are posted online before being concluded	Understand why contracts are being published way before being concluded. Ensure due diligences were made and the output service/good has the required quality.
Cluster 1 - Rapid Conclusion	3685	-0.174292	Rapid Conclusion, Small Number of Non-Winning Bidders	Determine whether rapid conclusion times are justified or potentially indicative of irregularities (e.g., bypassing standard procedures). Investigate if fast conclusion times compromise quality or transparency of procurement processes.
Cluster 2 - Spending Efficiency and Delays	1083	-0.262130	High Number of Winners, High Money Spent, Contracts finish after the expected	Investigate the relationship between the winners and see if they participate in a lot of tenders together. Investigate whether these contracts represent potential cases of overspending or inefficiencies in resource allocation. Look into the causes that resulted in the execution delay. The fact that these are probably more complex services (due to their price) might justify some of the delays.
Cluster 3 - Competitive Bidding	1365	-0.168444	High Number of Non-Winning Bidders	If a contract has a high number of non-winning bidders, it can be assumed that it was a competitive procedure and that there was strong competition. It can be of interest to see if the chosen proposal was in the fact the best one across the different bids and if it was the one most aligned with the requirements.

Despite all the contracts analysed here having been deemed as irregular, cluster 2 is the one that is most linked with a high degree of suspiciousness, and therefore should be the first one to be investigated. Cluster 0 is the one with the lowest value, which is coherent with the fact that it is also the one with the smaller deviations to the group of normal contracts.

Finally, we visualize the multidimensional clusters in two dimensions, which is done using both t-SNE (Appendix *G1 - t-SNE* Visualization) and UMAP (Appendix *G2 - UMAP* Visualization). Looking into the graphics, one can see well-separated clusters in both, which is another positive indicator, alongside the obtained Hopkins score, concerning the quality of the clusters.

5. CONCLUSIONS

The objectives of this research were first to identify a set of contracts with anomalous values that deviate significantly from the remaining contracts, and secondly, to integrate that information with red flags calculated at the entity level to create a framework that can serve as a valuable tool in the task of narrowing down potential irregularities or areas of concern within the procurement processes.

To achieve this task, a dataset of over two million Portuguese public contracts from the years of 2009 to 2022 was used, with this dataset being broken down into a smaller one contemplating just the last five years of contracts to better capture recent procurement trends. Regarding the analysis of the contracts and the features, only their metadata was used, as the actual text of contracts and the documents themselves are considered unstructured data and would require more complex techniques to be incorporated. This prevents the use of rich information that would be useful for the task at hand at both the anomaly detection and red flags phases, making the incorporation of Natural Language Processing methods probably the biggest avenue for future researchers.

This reduced dataset that had slightly under eight hundred thousand contracts was object of several transformations, with a focus on feature engineering to create new features that capture various aspects of contract dynamics and possible anomalies.

Once treated the data, the unsupervised anomaly detection method Isolation Forest was used to assign an anomaly score to each contract, and SHAP values were calculated to understand the model outputs, with a focus on understanding which features most impacted the anomaly score and in what way they did so, as explainability is a core element in a project of this nature. Based on those anomaly scores, a fraction of around 8000 contracts was deemed as anomalous and became the new object of study.

Leveraging some of the created features and creating new ones at an entity level rather than an individual contract level, it was possible to capture information about the entities involved in the procedures, both on the state side that awards the contracts and on the private side that applies to and wins contracts. Once calculated the red flags for both sides, we then saw how many red flags each company had with a value above the 75^{th} percentile.

This information was after merged with the anomaly score, according to a formula, achieving a final combined metric that can be used to sort contracts in order of neediness to investigate. With this metric, auditors have a concrete way of knowing where to focus their attention and can concentrate their efforts of detecting irregularities on a much smaller and more manageable set of contracts, culminating in a more efficient allocation of resources.

To aid even further de human auditors, the anomalous contracts were separated into four clusters, each representing a subset of anomalous contracts with similar characteristics. For each cluster, investigation actions were then suggested in line with the cluster characteristics, allowing for targeted and efficient investigation for each individual contract, ultimately enhancing the effectiveness of irregularity detection and mitigation efforts.

This methodology provides a comprehensive data-driven way of tackling irregularities in Portuguese Public Procurement, holding significant implications for policymakers and other actors involved in this field. However, the implementation of such a system requires careful consideration of its real-world effects, costs, and benefits. First, its effectiveness must be tested, which can be done by comparing the rate of irregularities detected using this methodology with those found through traditional means. When doing so, it is important to consider the possibility for this framework to be calibrated and fine-tuned, giving weights to specific parameters that one wants to give greater focus to during the anomaly detection stage. This possibility was not duly explored in this work, as the idea was to create a generic metric, that can be implemented in several Public Procurement systems while being agnostic to the specific characteristics of each one. Then, technical obstacles related with data accessibility, quality, and interoperability among various procurement bodies must be addressed, as a difficulty would be the requirement for constant monitoring and adjustment to changing rules and procurement practices.

Eventual difficulties are more likely than not to arise, and resistance can also stem from a reluctance to change established procedures or from a scepticism towards new technologies. However, Public Procurement as a process has been through several reforms over the years, either in terms of legislation, factors that determine how the awards are decided or simply because the integration of new technologies to streamline the process necessarily ends up changing the ways in which it is conducted and how it is structured. This constant evolution reflects a clear goal of having a more transparent, efficient and fair process for everyone, and AI-based methods can be capable of improving the process and help meet the needs of citizens and government authorities in new ways. One such way is the one being addressed in this study: the reduction of irregularities in Public Procurement. And while now this is probably the most looked at aspect of AI in Public Procurement, there are many other applications that have tremendous potential. For example, with the recent developments of Generative AI, writing bid texts will be much easier for firms to do. For that reason, it can be expected that in the future there will be more firms bidding for more contracts. While this could lower barriers to entry the market for suppliers and in that way foster competitiveness even more, levelling the playing field in favour of small and medium enterprises, it also poses an increased challenge for governments who will presumably have a larger number of proposals to analyse, and creates the risk of suppliers submitting more speculative bids due to the lower effort required to elaborate them. In such context, a method like this one becomes even more useful.

Finally, the main assumption behind analysing Public Procurement processes is that irregularities can be detected and prevented. However, that is not necessarily the case, as some malpractices are still hard to track in the public sector, and data-driven methods may not always be effective in uncovering them. Instances of bribery and kickbacks, as well as collusion, consistently entail evading established controls and going “under the radar”. Therefore, relying solely on scrutinizing transaction data for patterns can be insufficient and auditors frequently rely on behavioural concepts and qualitative factors to gain insights beyond the confines of the available data, encompassing both what is present and what is conspicuously absent from the records. This perspective is lost when doing an analysis of this kind. Not only that, but things like an inferior quality in the deliverables, whether they are goods or services, cannot be addressed with a methodology relying solely on data.

After all the work that has been done, it can be difficult to accept that most procurement irregularities will keep going undetected, unreported, and therefore unmeasured. However, I believe that every attempt can contribute to the common objective of reducing them, not only because the work itself can be helpful in that regard, but also because it raises awareness and knowledge about the topic. Moreover, several domains should be engaged in this mission, from law, to finance and accounting, to technology. Every domain is essential for creating strong frameworks, putting in place efficient controls, and raising awareness in their own way about the repercussions of procurement irregularities.

On the technological side, and with structured data being ever as available, it is almost imperative that data-driven methods are massively employed in this task. Addressing irregularities in Public Procurement needs a holistic response incorporating detection and prevention, and that response will be much stronger if data is part of it.

BIBLIOGRAPHICAL REFERENCES

Agrawal, S., & Agrawal, J. (2015). Survey on Anomaly Detection using Data Mining Techniques. *Procedia Computer Science*, 60, 708–713. https://doi.org/10.1016/j.procs.2015.08.220

Aldana, A., Falcón-Cortés, A., & Larralde, H. (2022). A machine learning model to identify corruption in Mexico's public procurement contracts. *arXiv*. http://arxiv.org/abs/2211.01478

Autoridade da Concorrência. (2024). Processos com práticas restritivas de concorrência. https://extranet.concorrencia.pt/pesquisAdC/

Baesens, Bart, Van Vlasselaer, Veronique, & Verbeke, Wouter. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: A guide to data science for fraud detection*. Wiley. https://doi.org/10.1002/9781119146841

Bai, J., & Qiu, T. (2023). Automatic Procurement Fraud Detection with Machine Learning. *arXiv*. http://arxiv.org/abs/2304.10105

Becht, E., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). Evaluation of UMAP as an alternative to t-SNE for single-cell data. *Bioinformatics*. https://doi.org/10.1101/298430

Bento, N., Sousa, C., Trindade, P., Mamede, R. P., Fontes, M., & Alves, T. (2022). Robust relation between public procurement for innovation and economic development. *Economics Letters*, 211, 110241. https://doi.org/10.1016/j.econlet.2021.110241

Bouman, R., Bukhsh, Z., & Heskes, T. (2023). Unsupervised anomaly detection algorithms on real-world data: How many do we need?. *arXiv*. http://arxiv.org/abs/2305.00735

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. https://doi.org/10.1145/335191.335388

Carneiro, D., Veloso, P., Ventura, A., Palumbo, G., & Costa, J. (2020). Network Analysis for Fraud Detection in Portuguese Public Procurement. *Lecture Notes in Computer Science*, 390–401. https://doi.org/10.1007/978-3-030-62365-4_37

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Compuing Surveys*, 41. https://doi.org/10.1145/1541880.1541882

Cross, G. R., & Jain, A. K. (1982). Measurement of Clustering Tendency. *IFAC Proceedings Volumes*, 15(1), 315–320. https://doi.org/10.1016/S1474-6670(173365-2

Curado, A., Damásio, B., Encarnação, S., Candia, C., & Pinheiro, F. L. (2021). Scaling behavior of public procurement activity. *PLOS ONE*, 16(12), e0260806. https://doi.org/10.1371/journal.pone.0260806

Emmott, A., Das, S., Dietterich, T., Fern, A., & Wong, W.-K. (2016). A Meta-Analysis of the Anomaly Detection Problem. *arXiv*. http://arxiv.org/abs/1503.01158

Espírito Santo, M. (2022). Concorrência multa empresas de segurança em 41 milhões por cartel para concorrer a concursos públicos. *ECO*. https://eco.sapo.pt/2022/07/13/concorrencia-multa-empresas-de-seguranca-por-cartel-para-concorrer-a-concursos-publicos-em-41-milhoes/

Fazekas, M., Tóth, I. J., & King, L. P. (2016). An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research*, 22(3), 369–397. https://doi.org/10.1007/s10610-016-9308-z

Fazekas, M., & Kocsis, G. (2020). Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data. *British Journal of Political Science*, 50(1), 155–164. https://doi.org/10.1017/S0007123417000461

Gallego, J., Rivero, G., & Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting*, 37(1), 360–377. https://doi.org/10.1016/j.ijforecast.2020.06.006

García Rodríguez, M. J., Rodríguez Montequín, V., Ortega Fernández, F., & Villanueva Balsera, J. M. (2019). Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning. *Complexity*, 2019, 1–20. https://doi.org/10.1155/2019/2360610

Goernitz, N., Kloft, M., Rieck, K., & Brefeld, U. (2013). Toward Supervised Anomaly Detection. *Journal of Artificial Intelligence Research*, 46, 235–262. https://doi.org/10.1613/jair.3623

Heggstad, K. K., & Frøystad, M. (2011). The basics of integrity in procurement. https://www.u4.no/publications/the-basics-of-integrity-in-procurement.pdf

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J. E. L., Kirrane, S., Neumaier, S., Polleres, A., Navigli, R., Ngomo, A.-C. N., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2022). Knowledge Graphs. *ACM Computing Surveys*, 54(4), 1–37. https://doi.org/10.1145/3447772

Instituto dos Mercados Públicos, do Imobiliário e da Construção, I.P. (2023). *Relatório Anual Contratação Pública em Portugal 2022*. https://www.base.gov.pt/Base4/media/jv0fv3zp/relat%C3%B3rio-anual-da-contrata%C3%A7%C3%A3o-p%C3%BAblica_2022.pdf

International Anti-Corruption Resource Center. (2023). Red Flags Listed by Project Cycle. https://guide.iacrc.org/red-flags-listed-by-project-cycle/

Lewis-Faupel, S., Neggers, Y., Olken, B. A., & Pande, R. (2006). Can Electronic Procurement Improve Infrastructure Provision? Evidence From Public Works in India and Indonesia. *AMERICAN ECONOMIC JOURNAL: ECONOMIC POLICY*, 8(3), 258–283. https://doi.org/10.1257/pol.20140258

Lima, M., Silva, R., Lopes De Souza Mendes, F., R. De Carvalho, L., Araujo, A., & De Barros Vidal, F. (2020). Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. https://doi.org/10.18653/v1/2020.findings-emnlp.143

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv*. http://arxiv.org/abs/1705.07874

Lyra, M. S., Curado, A., Damásio, B., Bação, F., & Pinheiro, F. L. (2021). Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil) municipalities. *Applied Network Science*, 6(1). https://doi.org/10.1007/s41109-021-00418-y

Lyra, M. S., Damásio, B., Pinheiro, F. L., & Bacao, F. (2022). Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. *Applied Network Science*, 7(1), 83. https://doi.org/10.1007/s41109-022-00523-6

Nicholas, C., & Fruhmann, M. (2014). Small and medium-sized enterprises policies in public procurement: Time for a rethink?. *Journal of Public Procurement*, 14(3), 328–360. https://doi.org/10.1108/JOPP-14-03-2014-B002

Niessen, M. E. K., Paciello, J. M., & Fernandez, J. I. P. (2020). Anomaly Detection in Public Procurements using the Open Contracting Data Standard. 2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG), 127–134. https://doi.org/10.1109/ICEDEG48599.2020.9096674

OECD. (2009). OECD Principles for Integrity in Public Procurement. https://doi.org/10.1787/9789264056527-en

Palacio-Niño, J.-O., & Berzal, F. (2019). Evaluation Metrics for Unsupervised Learning Algorithms. *arXiv*. http://arxiv.org/abs/1905.05667

Perini, L., Buerkner, P., & Klami, A. (2023). Estimating the Contamination Factor's Distribution in Unsupervised Anomaly Detection. *arXiv*. http://arxiv.org/abs/2210.10487

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. *arXiv*. http://arxiv.org/abs/arXiv:1009.6119

Portal BASE. (2021). História do Base. https://www.base.gov.pt/Base4/pt/o-portal/historia-do-base/

Portal BASE. (2022). Características dos Procedimentos. https://www.base.gov.pt/Base4/pt/documentacao/caracteristicas-dos-procedimentos/

Rabuzin, K., & Modrušan, N. (2019). Prediction of Public Procurement Corruption Indices using Machine Learning Methods. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 333–340. https://doi.org/10.5220/0008353603330340

Rustiarini, N. W., T., S., Nurkholis, N., & Andayani, W. (2019). Why people commit public procurement fraud? The fraud diamond view. *Journal of Public Procurement*. https://doi.org/10.1108/JOPP-02-2019-0012

Sturm, N., Candia, C., Damásio, B., & Pinheiro, F. L. (2024). High Earnings through Firm Influence: The Role of Hierarchical Structures in Public Procurement. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4714647

Tadelis, S. (2012). Public procurement design: Lessons from the private sector. *International Journal of Industrial Organization*, 30(3), 297–302. https://doi.org/10.1016/j.ijindorg.2012.02.002

Taha, A., & Hadi, A. S. (2020). Anomaly Detection Methods for Categorical Data: A Review. *ACM Computing Surveys*, 52(2), 1–35. https://doi.org/10.1145/3312739

Torres-Berru, Y., & López Batista, V. F. (2021). Data Mining to Identify Anomalies in Public Procurement Rating Parameters. *Electronics*, 10(22), 2873. https://doi.org/10.3390/electronics10222873

Tribunal de Contas. (2023). Empreendimento de Obras Públicas: Relatório de auditoria 6/2023, 2.ª secção. https://www.tcontas.pt/pt-pt/ProdutosTC/Relatorios/RelatoriosAuditoria/Documents/2023/rel006-2023-2s.pdf

Velasco, R. B., Carpanese, I., Interian, R., Paulo Neto, O. C. G., & Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research*, 28(1), 27–47. https://doi.org/10.1111/itor.12811

Wachs, J., Fazekas, M., & Kertész, J. (2021). Corruption risk in contracting markets: A network science perspective. *International Journal of Data Science and Analytics*, 12(1), 45–60. https://doi.org/10.1007/s41060-019-00204-1

Wachs, J., & Kertész, J. (2019). A network approach to cartel detection in public auction markets. *Scientific Reports*, 9, 10818. https://doi.org/10.1038/s41598-019-47198-1

Wittberg, E., & Fazekas, M. (2023). Firm performance, imperfect competition, and corruption risks in procurement: Evidence from Swedish municipalities. *Public Choice*, 197(1–2), 227–251. https://doi.org/10.1007/s11127-023-01102-8

APPENDIX A – PUBLIC PROCUREMENT IN PORTUGAL

For the matter of storing contract-related information online, in January 2008, the initial iteration of the Public Procurement Portal (Portal BASE) was launched with the objective to centralize data concerning public contracts established within the country. This portal, which is overseen by *Instituto dos Mercados Públicos, do Imobiliário e da Construção* (IMPIC), established a digital environment where details regarding the initiation and fulfilment of public contracts are disclosed, thereby enabling their tracking and oversight. By storing all the Public Procurement contracts celebrated in the past years, this initiative enhanced the scrutiny over the celebrated contracts and facilitated the analysis of contract-related data. Not only that, it also provided a comprehensive dataset that can be utilized by researchers, policymakers, and stakeholders to identify trends, patterns, and insights within the Public Procurement landscape (Portal BASE, 2021).

Despite the clear upside that Portal BASE brought to the procurement landscape, there have been some concerns due to the quality of the data that is published there. Complaints have arisen about the careless or misleading insertion of information in certain contracts, leading to discrepancies and inaccuracies, which hinders the desired transparency and effectiveness of the portal, and jeopardizes the work of those who try to use such data to derive insights. This issue was brought to light in a more concrete way in mid-2023, when the Portuguese court of auditors concluded, in an audit of over a thousand public works projects in Portugal, that there was a lack of appropriate control over the reliability of public information, pointing out several deficiencies, such as errors in prices or problems with dates, like contracts with close dates who preceded the beginning dates (Tribunal de Contas, 2023).

In Portugal, besides the several regulatory instruments of European Union Law, to ensure well-structured and compliant Public Procurement procedures, laws are controlled with the help of CCP (*Código dos Contratos Públicos - Decree-Law no. 18/2008, of 29 January and subsequent changes*[2]). This regulation is obligatory for all contract arrangements and being compliant with it is a non-negotiable factor for the contract to be formalized. It outlines the respective roles and responsibilities of the involved parties, specifies the terms and conditions to be upheld during the contract's execution, and addresses the legal requirements that the awarded party must meet.

Regardless of the remarkable progress that has been made in terms of legislation and information availability, Public Procurement is still vulnerable to various risks and challenges that threaten its integrity and effectiveness, with illicit acts that lead to irregularities possibly being perpetrated by contractors external to the organisations, but also by people within the organisation if the right context is present, with this context involving many things that can be explained with four main vertexes: motivation (usually economical), opportunity, rationalizations to justify the act and capability to commit it (Rustiarini et al., 2019).

2 - http://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=2063&tabela=leis

APPENDIX B – MAIN PROCEDURE TYPES

Non-Competitive Procedures	Competitive Procedures
Direct Award (Ajuste Direto)	Prior Consultation (Consulta Prévia)
Simplified Direct Adjustment (Ajuste Direto Simplificado)	Public Tender (Concurso Público)
	Urgent Public Tender (Concurso Público Urgente)
	Tender Limited by Qualification (Concurso Limitado por Prévia Qualificação)

Direct Award is a procedure whereby the contracting authority directly invites an enterprise of its choice to submit a tender. This practice can only be followed in cases with specific conditions, namely the concession of public works with an estimated value below 30.000€, the rental or purchase of movable goods and acquisition of services with an estimated value below 20.000€, or any other contracts (excluding concessions for public works or public services and partnership contracts) with an estimated value below 50.000€. Other exceptions can be made for contracts to be directly awarded regardless of the amount of money involved, in what is called material criteria. Some of the assumptions that allow for this bypass of the regulations are public tenders that did not have anyone applying to them or every offer was excluded, urgent and compelling reasons and exclusive providers.

The simplified version of direct award, which was introduced along with other simplified procedures with the goal of accelerating Public Procurement and using community funds within the deadlines, waives any procedural formalities, and it is consummated when the responsible for the decision to contract approves the invoice presented by the invited entity.

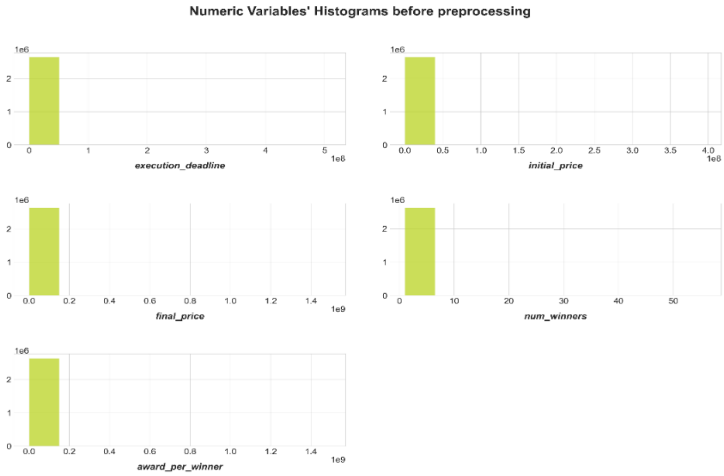
Public Tender is the default procedure of Public Procurement. In this procedure, entities start by submitting proposals, which means that there is no stage of evaluating the technical and/or financial capacity of the tenderers. Once all proposals are submitted and the established deadline for this stage ends, the state entity must evaluate them and pick the best one according to the established criteria.

Prior Consultation is a procedure in which the contracting authority directly invites at least three entities of its choice to present a proposal, with the possibility of negotiating with them the aspects of the performance of the contract to be celebrated.

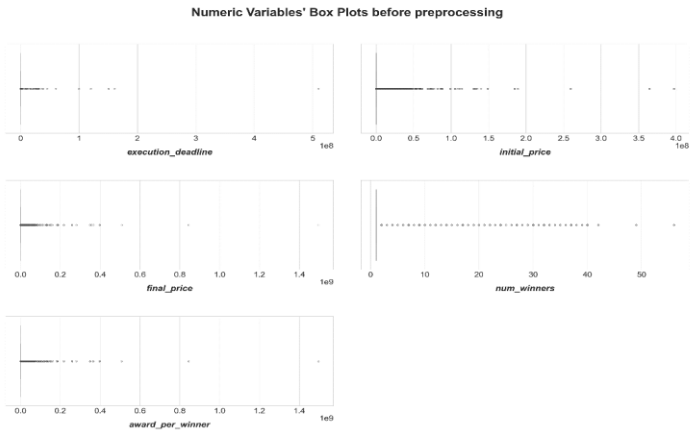
A Tender Limited by Qualification takes place when the value of the contract to be celebrated exceeds the European thresholds, and the procedure is characterized by being composed of two procedural phases: a first phase, where there is the presentation of applications and qualification of candidates; a second phase, where the analysis of proposals and adjudication take place.

APPENDIX C – NUMERIC VARIABLES DISTRIBUTION AND CORRELATION MATRIX BEFORE PREPROCESSING

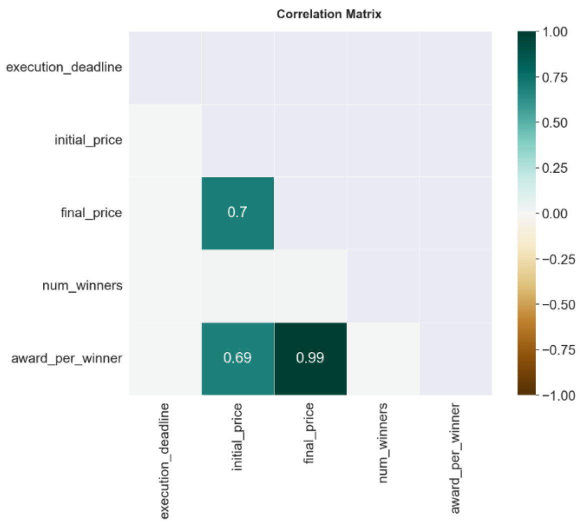
C1 - Histograms



C2 - Box Plots

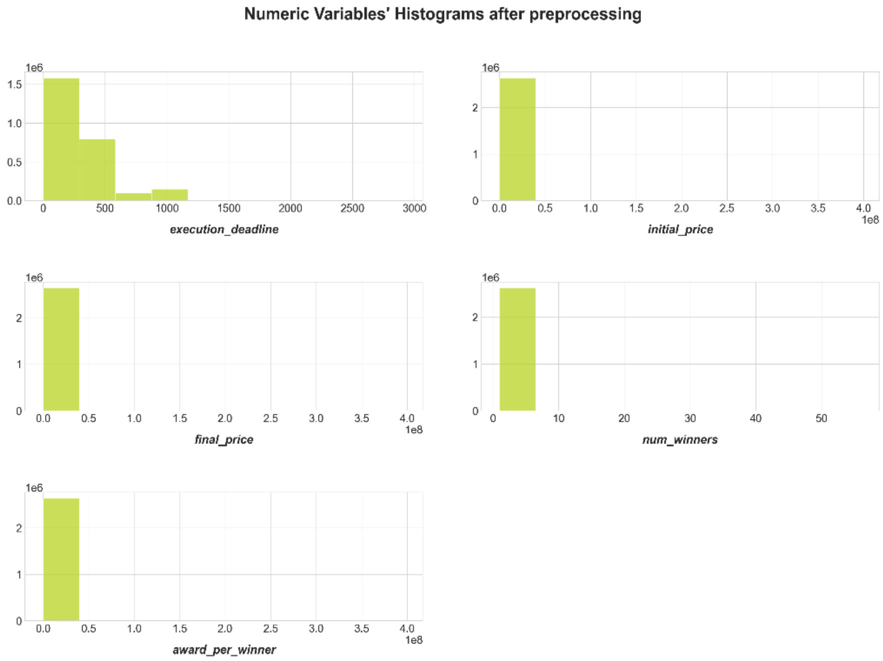


C3 - Correlation Matrix

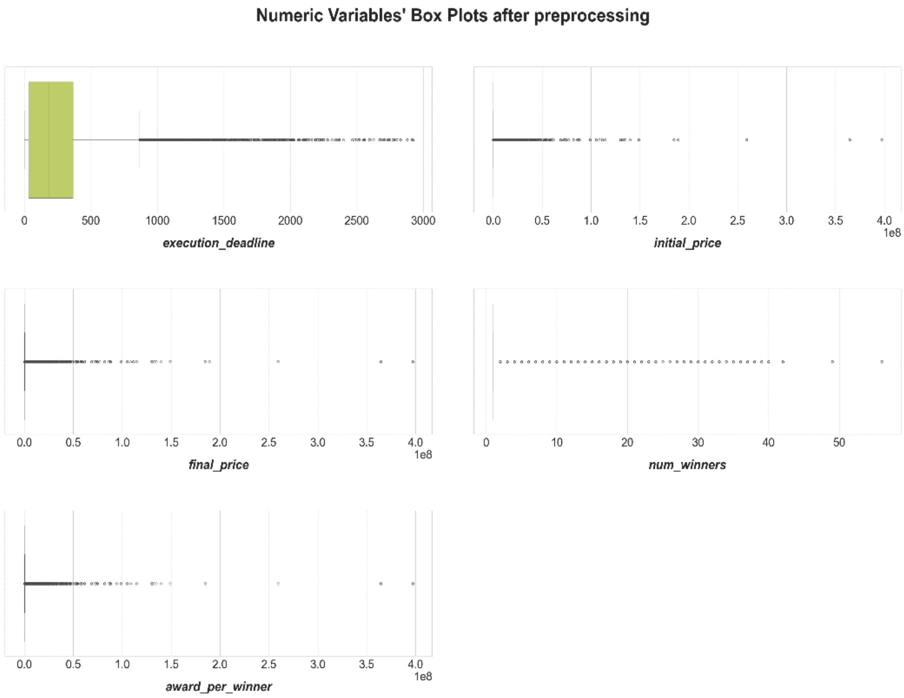


APPENDIX D – NUMERIC VARIABLES DISTRIBUTION AFTER PREPROCESSING

D1 - HISTOGRAMS

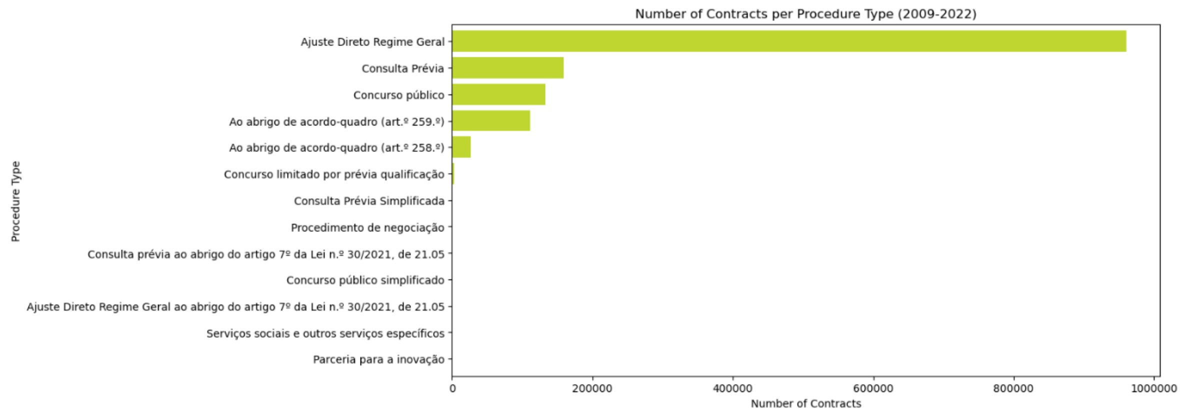


D2 - BOX PLOTS

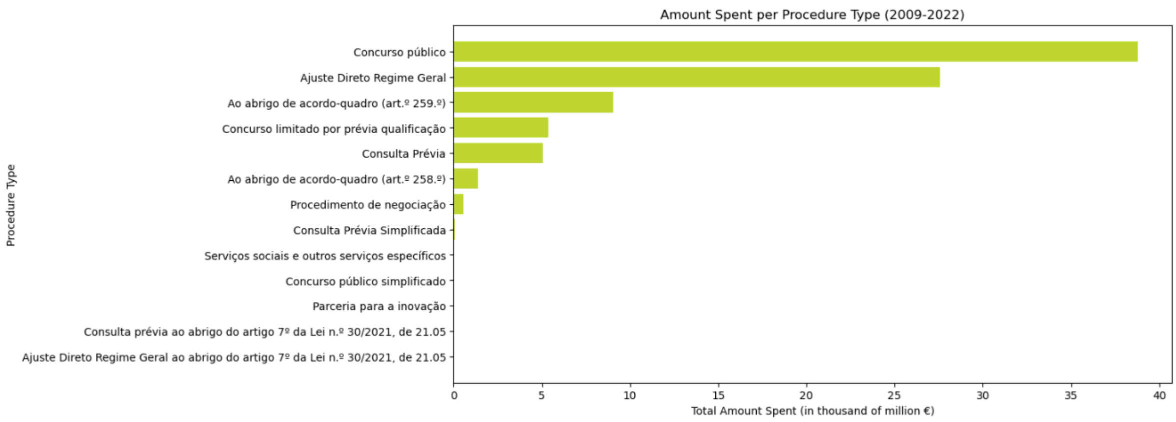


APPENDIX E – VISUALIZATIONS

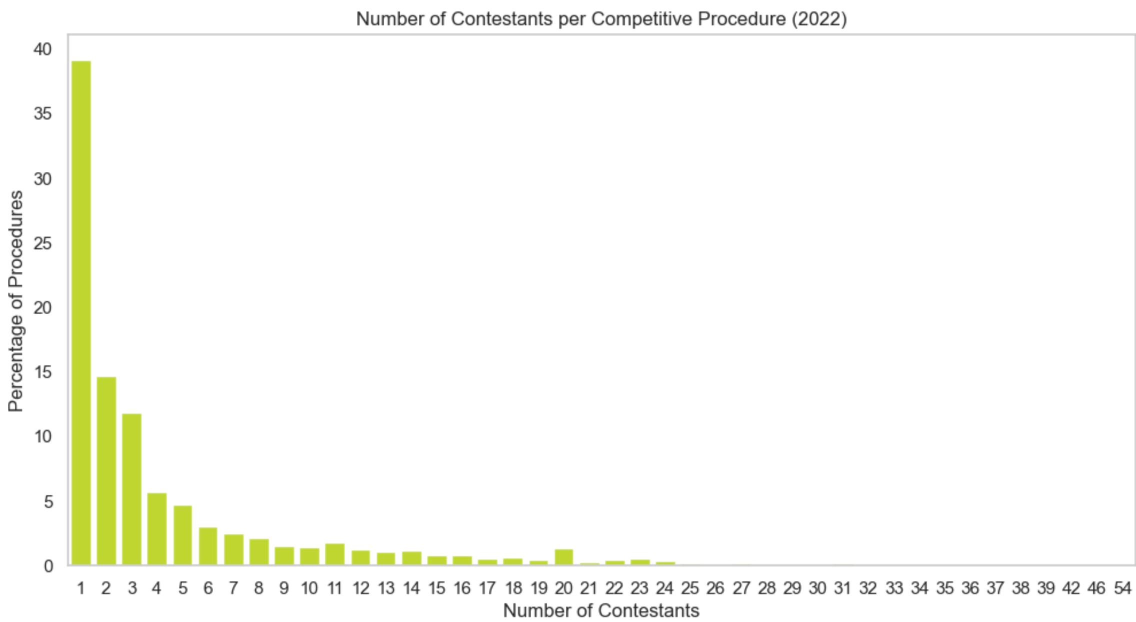
E1 - Number Of Contracts Per Procedure Type



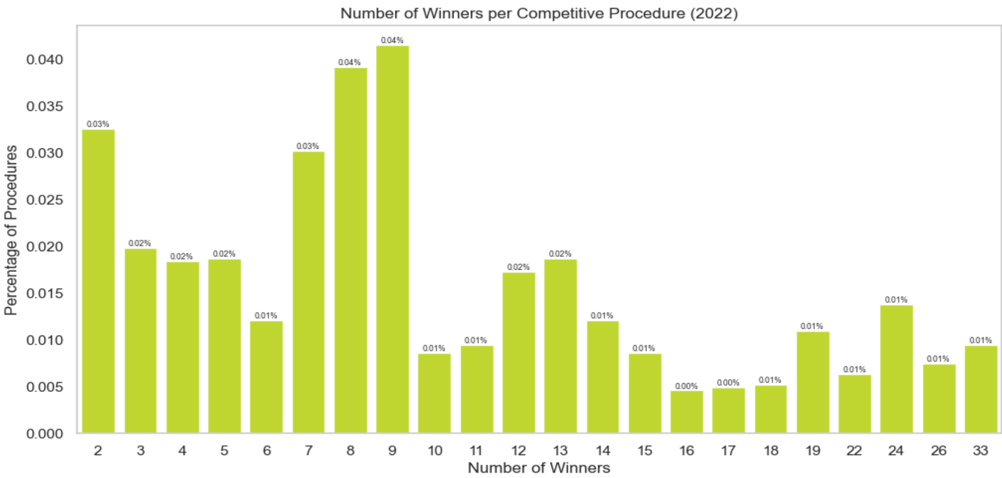
E2 - Expenditure Per Procedure Type



E3 - Number Of Contestants Per Procedure

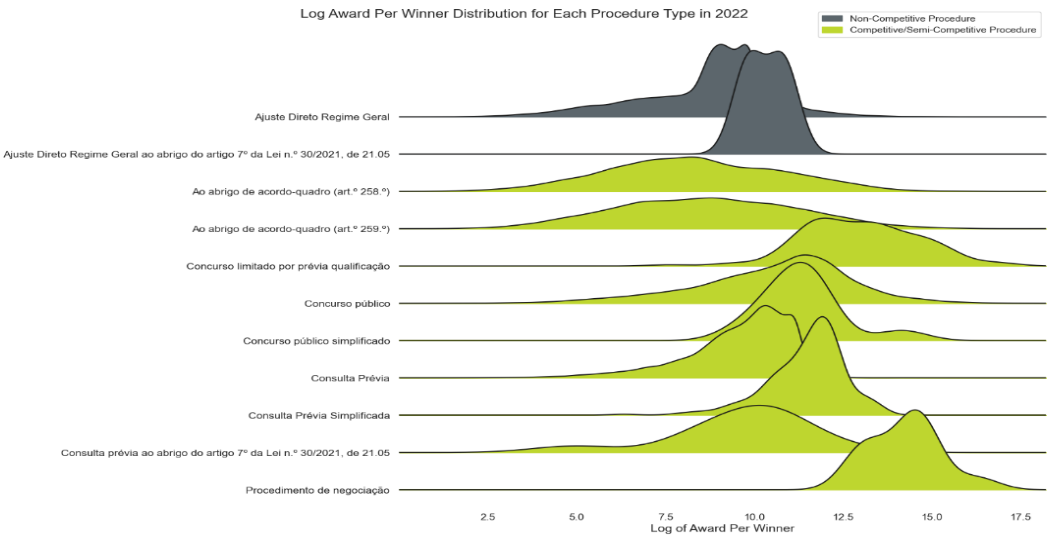


E4 - Number of Winners Per Procedure

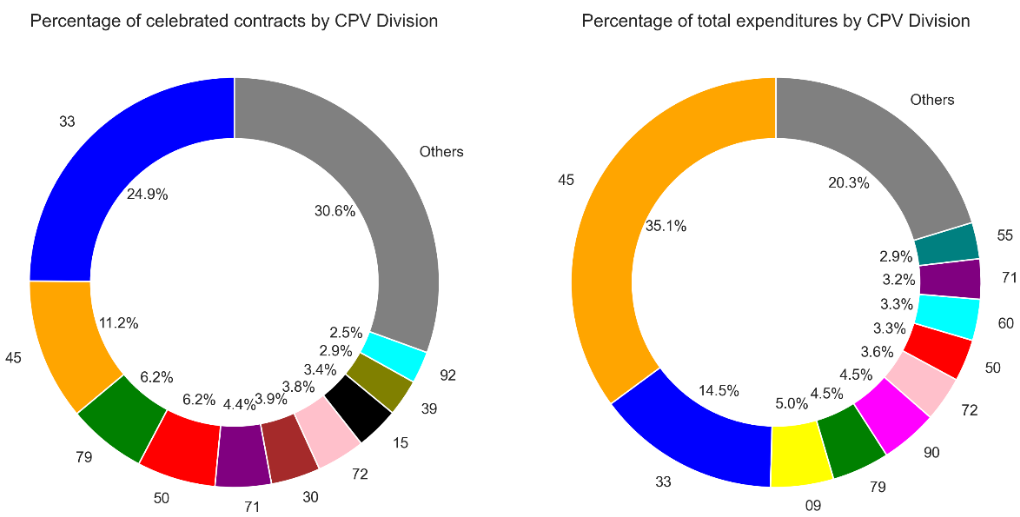


Note: Contracts with only 1 winner not included, as they were >99% of total number.

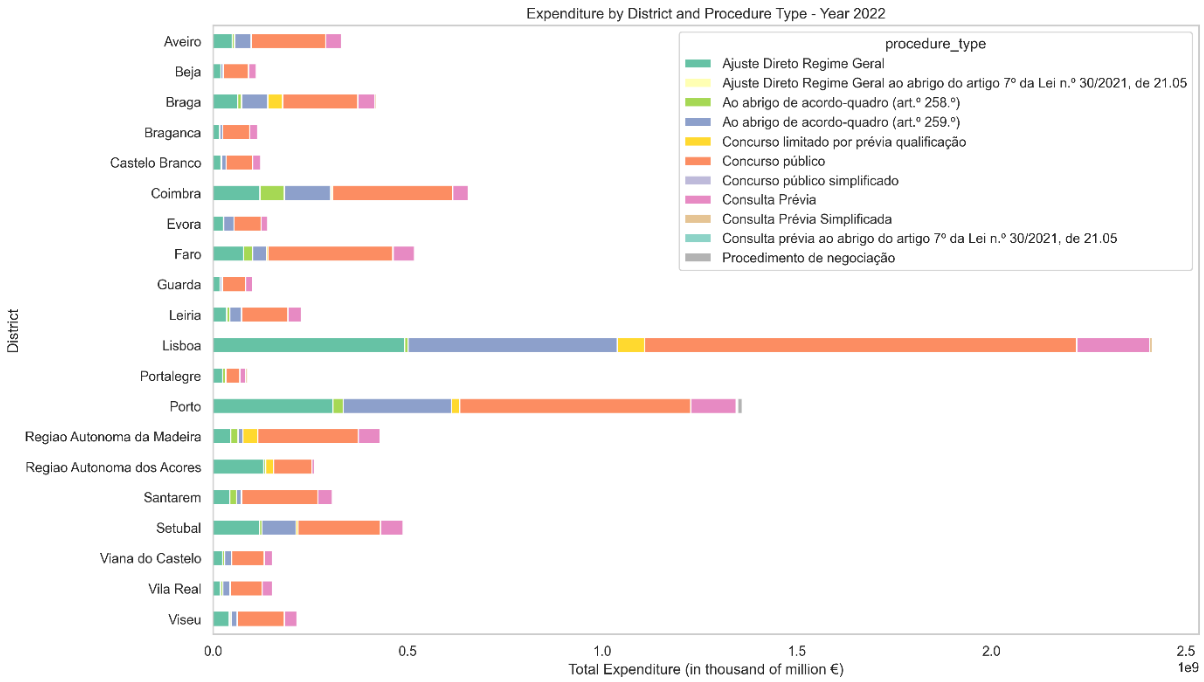
E5 - Final Price Distribution Per Procedure Type



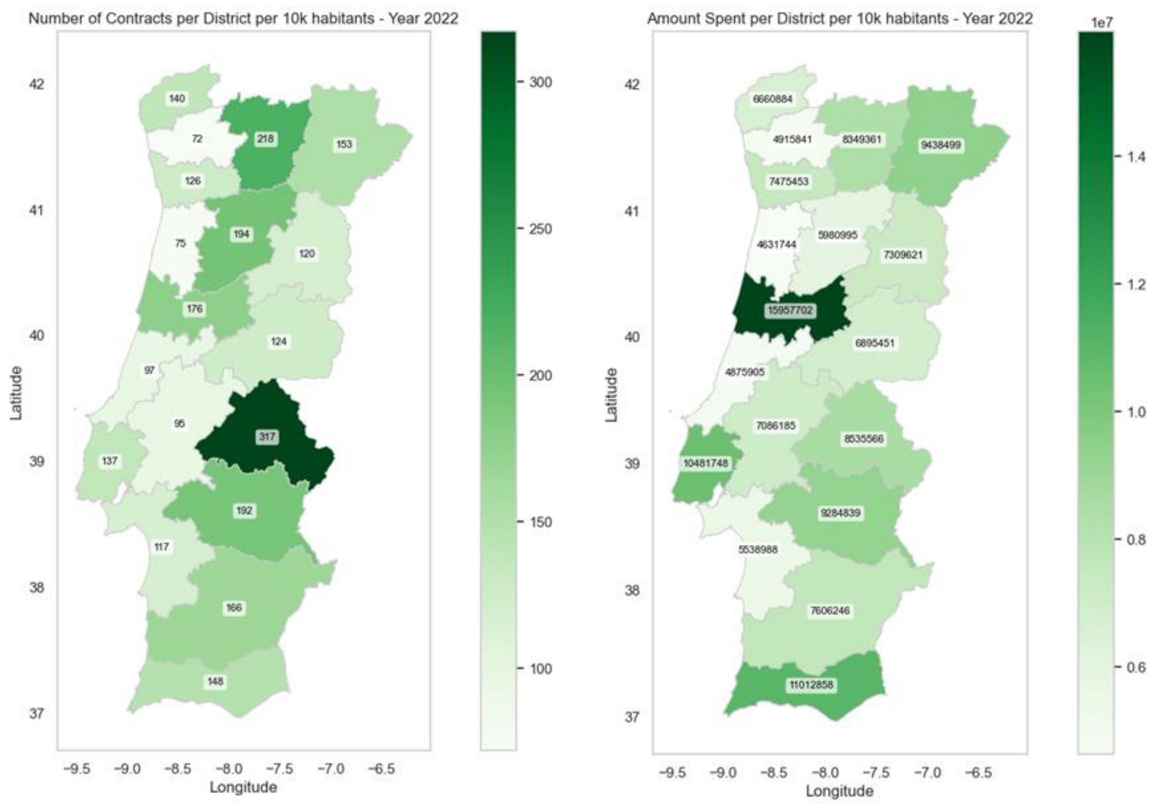
E6 - Number of Contracts and Expenditure Per CPV Division



E7 - Expenditure Per District and Procedure Type



E8 - Number of Contracts and Expenditure per District per 10k Habitants



APPENDIX F – AUXILIAR RESULTS AND METRICS

F1 - Kolmogorov-Smirnov Test Results

Feature	Isolation Forest	Local Outlier Factor
Award per Winner	0.96*	0.16*
Number of Winners	0.01	0.001
Price Difference	0.05*	0.23*
Fast Conclusion	0.06*	0.24*
Price Ratio to Median CPVS Group Price	0.89*	0.11*
Price Ratio to Median Procedure Type Price	0.90*	0.14*
Price Ratio to Median Contract Type Price	0.93*	0.13*
Number of Non-winning Bidders	0.33*	0.06*
Time Between Contract Finish and Publication on Portal BASE	0.42*	0.27*
Percentage of District Expenditures	0.73*	0.14*

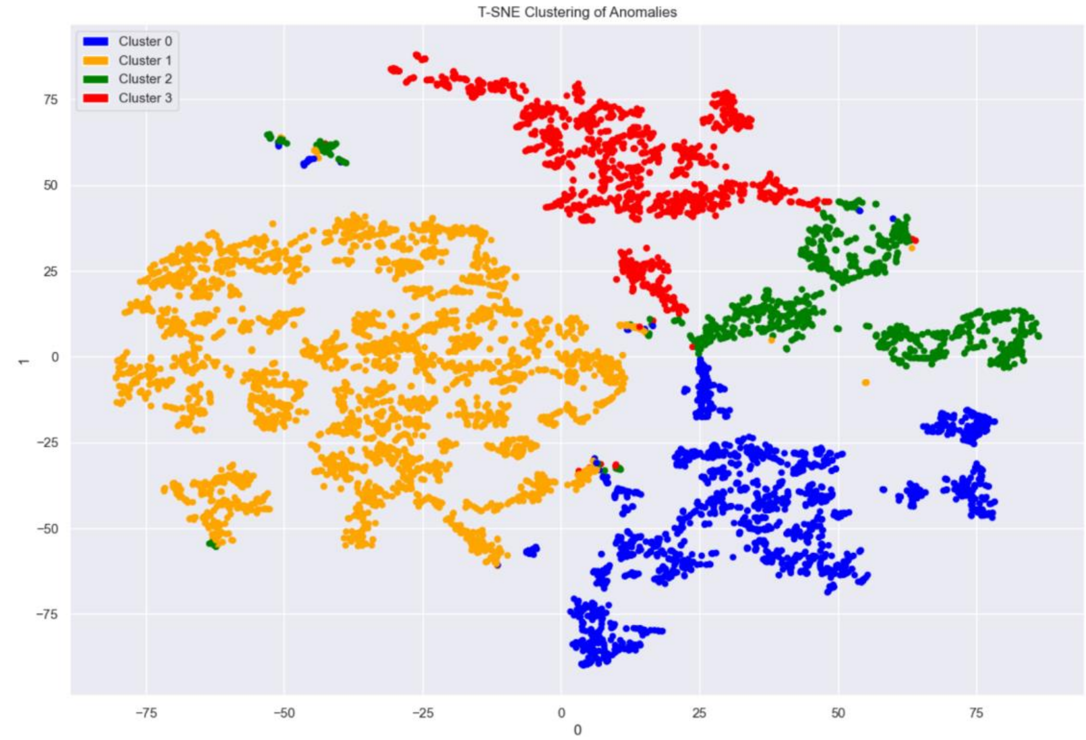
*Relevant p-value, most successful method

F2 - Clustering Metrics

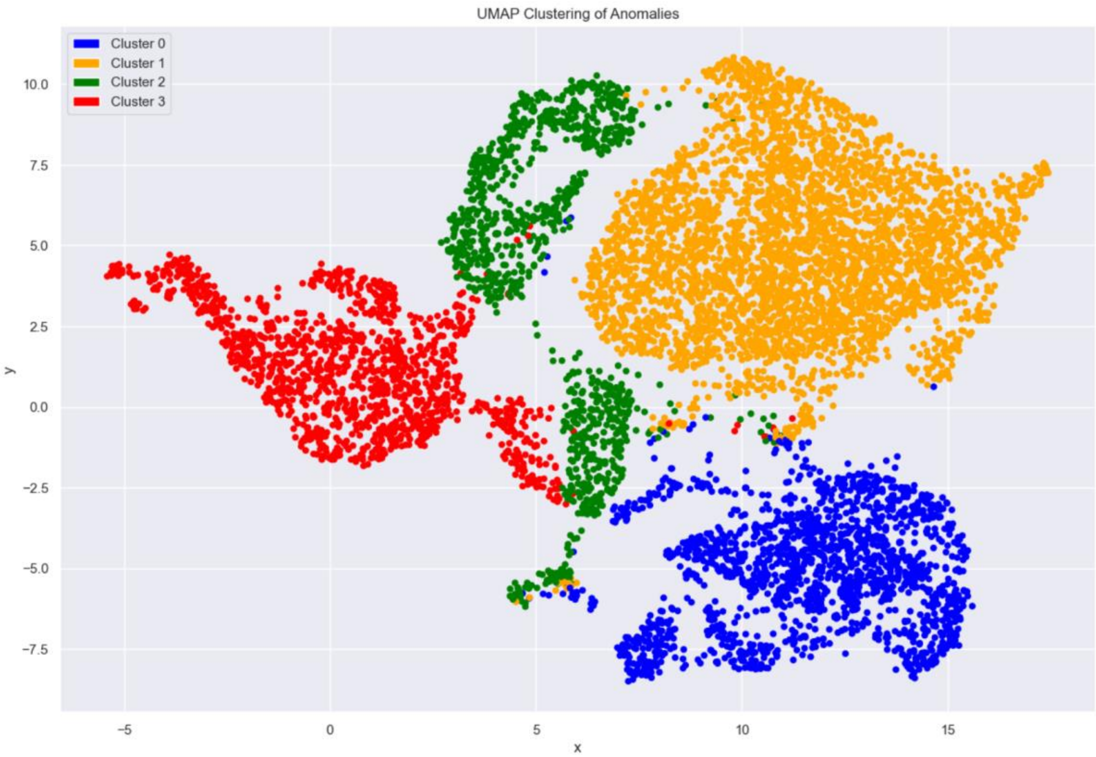
Method	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
Hierarchical Clustering	0.2468	1.268	1351.682
K-Means	0.3462	1.3008	1475.9631
Fuzzy C-Means	0.3512	1.250	1500.123
Mean-Shift	0.2548	2.3376	525.3772

APPENDIX G – CLUSTER VISUALIZATIONS

G1 - T-SNE VISUALIZATION



G2 - UMAP VISUALIZATION



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa