NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Anomaly Detection in Portuguese Public Procurement Contracts

An Embedding-Based Machine Learning Approach

Ana Rita Saraiva da Silva

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

Anomaly Detection in Portuguese Public Procurement Contracts

An Embedding-Based Machine Learning Approach

by

Ana Rita Saraiva da Silva

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

Supervised by

Flávio Pinheiro, PhD, NOVA Information Management School

Bruno Damásio, PhD, NOVA Information Management School

July, 2025

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.

Aveiro, 15/07/2025

ACKNOWLEDGEMENTS

I would like to firstly express my gratitude to professors Flávio Pinheiro and Bruno Damásio for accepting me to develop this project under their supervision. I deeply appreciate all the assistance that they offered me. Their insights combined with constructive criticism allowed me to solve challenges and motivated me to see things from a different perspective. I would also like to thank the support provided by the projects funded by the Portuguese Foundation for Science and Technology (FCT) 2024.07378.IACDC and 2024.07601.IACDC, which are led by the professors mentioned above.

The guidance and feedback provided by Elsa Camuamba and Niclas Sturm were also very valuable throughout the development of this work. Their genuine interest in the topic significantly enriched the quality of the study and contributed meaningfully to its direction.

I wish to acknowledge my parents for always maintaining their trust in me. Their faith in my abilities and their persistent encouragement have been a source of strength. I appreciate all the affection and support which they have constantly provided me.

To my partner, thank you for your care throughout this journey. You took sincere interest in this thesis and your emotional support helped me to stay on track.

To my friends and colleagues who formed part of this experience with me, I am thankful for your company. The laughter we shared, complemented by the meaningful moments we enjoyed, enriched my life. The friendships made at NOVA IMS along with the ones created throughout my academic journey tremendously transformed me. Regardless of where life takes us, I will carry our memories with me forever.

Lastly, a heartfelt thank you to the wonderful professors who taught me throughout the years, which deserve my greatest gratitude. I truly appreciate your dedication and the endless wisdom provided. Your passion has inspired me, and I am grateful for everything I learned from you.

ABSTRACT

Public procurement is an essential aspect of public administration and one of the functions that help fulfill the population's needs, although it is prone to irregularities. With the evolution of digital infrastructures, many governments have adopted electronic procurement systems that automate workflows, improve transparency, and increase access to contract information. These advancements make it possible to employ techniques that analyze large volumes of procurement data to uncover unusual patterns and behaviors. Considering that, this study uses real-world data extracted from the Portuguese public procurement platform BASE, covering the years 2020 to 2024, to explore the application of unsupervised machine learning algorithms, particularly clustering, for anomaly detection. A key innovation in this research is the creation of text embeddings, a technique rarely applied in this field. The resulting model aims to enhance fair competition and good governance by identifying deviations from normative procurement processes. Additionally, it provides auditors and policymakers with the tools to effectively identify potential irregularities.

KEYWORDS

Public Procurement; Anomaly Detection; Unsupervised Learning; Embeddings; Clustering

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
BGMM	Bayesian Gaussian Mixture Model
BERT	Bidirectional Encoder Representations from Transformers
CPC	Code of Public Contracts
CPV	Common Procurement Vocabulary
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
E-Procurement	Electronic Procurement
EU	European Union
GDP	Gross Domestic Product
GMM	Gaussian Mixture Model
GPT	Generative Pre-trained Transformer
LaBSE	Language-agnostic BERT Sentence Embedding
ML	Machine Learning
NLP	Natural Language Processing
OECD	Organisation for Economic Cooperation and Development
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection

1. INTRODUCTION

1.1. CONTEXT AND MOTIVATION

Public procurement represents one of the primary forms of government spending that seeks to deliver works, goods, or services (García Rodríguez et al., 2022) in order to achieve specific goals, fulfill requirements, and perform activities necessary to obtain desired policy outcomes (Prier & McCue, 2009). It has effects in multiple industries, including transportation, communication, healthcare, and defense, by fostering innovation, regional development, and establishing partnerships with suppliers (Axelsson & Torvatn, 2017). Beyond these direct impacts, this activity also acts as a forward policy for improving government spending, lowering public costs, and stimulating economic activity (Harink, 1999).

Public procurement is vastly relevant since it accounts for a growing share of the government public spending over the last decade across countries that belong to the Organisation for Economic Co-operation and Development (OECD), increasing from 11.8% of Gross Domestic Product (GDP) in 2007 to 12.9% in 2021. Additionally, in 2021, total government procurement spending represented 27.8% of their total government expenditure (OECD, 2024a).

However, despite its growing economic importance, this field faces serious challenges due to its vulnerability to irregularities such as fraud, corruption, and collusion. These issues undermine governance and economic efficiency, cause significant financial losses, and erode public trust in government institutions (OECD, 2019). Such irregularities include a wide range of bribery, embezzlement, collusion, abuse of office, favoritism, misappropriation, and nepotism (Padhi & Mohapatra, 2011), with corruption alone estimated by the United Nations Office on Drugs and Crime, to raise the price of public contracts by an average of 10 to 25% worldwide (UNODC, 2013).

Considering these factors, is essential to understand the anomalies present in public procurement and analyze the approaches and tactics that scholars have attempted to study in the direction of determine the reasons behind fraudulent activities. In response to these challenges, and supported by improvements in information access and digital systems, a new trend has emerged where electronic procurement (e-procurement) systems are used to enhance transparency (Lyra et al., 2022).

As an early adopter of e-procurement in Europe (Costa et al., 2013), Portugal implemented digital systems to modernize the field, which turns the country into a particularly compelling case of study. However, regardless technological advancements, concerns about institutional quality (Mélon & Spruk, 2020) and corruption persist (Transparency International, 2025).

Building on the digitalization of procurement, the application of advanced data science techniques enables the identification of irregular patterns that may signal potential

misconduct or inefficiencies, providing a tool that can be used to improve oversight and decision-making.

1.2. Problem Definition and Research Focus

Research on public procurement grew severely after 2010, but it was not until 2019 that the field began to more frequently incorporate machine learning (ML) techniques, as observed in Annex A. However, embedding-based anomaly detection remains largely underexplored, especially in Portugal, where existing studies primarily focus on policy analysis and transparency mechanisms. While such techniques have been successfully applied in other domains, their adaptation to procurement has only recently gained traction. Notable contributions have emerged from countries like Brazil (Hott et al., 2023) and Finland (Rahman, 2022), offering valuable insights that inform the methodological approach of this study.

To adress this gap, the present study employs embedding-based ML methods to detect anomalies in public procurement contracts in Portugal. The research is structured around the following research question and objectives:

Research Question:

- How effective are embedding-based ML models in detecting anomalies in Portuguese public procurement contracts?

Research Objectives:

- To collect and preprocess a comprehensive dataset of public procurement contracts using recent data extracted from Portugal's e-procurement system.
- To represent contract data using embedding techniques, transforming textual and structured information into dense vector representations that preserve semantic relationships.
- To apply unsupervised anomaly detection methods in order to identify clusters of typical contracts and isolate potential anomalies.
- To define and interpret cluster profiles, and analyze the distinguishing characteristics of anomalous contracts in comparison to typical procurement patterns.

1.3. Document Structure

This study offers a systematic overview of the theories, methods, and empirical results concerning anomaly detection in public procurement in Portugal, where each section is structured in the following way:

Section 2 - Literature Review: reviews and analyzes existent research while addressing the concepts, issues, and findings related to the domain of public procurement, embeddings and anomaly detection, emphasizing the research gaps addressed in this study.

Section 3 - Methodology: describes the approach and methods that have been implemented to achieve the objectives and the outcomes of this study. It details the overall methodology, including data collection, exploration, and preprocessing steps. Additionally, it defines the techniques and ML algorithms used for anomaly detection, as well as the metrics employed to measure the results.

Section 4 - Results and Discussion: presents the findings of the study, including a performance evaluation. It discusses the implications of the results, providing understanding into the factors that contribute to procurement anomalies, highlighting the effectiveness of the models employed.

Section 5 - Conclusions and Future Works: summarizes the main conclusions and contributions of this work. It describes the limitations of the research and suggest possible directions for future investigation.

2. LITERATURE REVIEW

2.1. THE STRATEGIC ROLE OF PUBLIC PROCUREMENT

Public procurement is a fundamental pillar of public service supply, that when well managed, is able to encourage public trust and influence the quality of citizens life. Moreover, beyond its functional role, it is increasingly employed by governments as a strategic tool to achieve policy objectives such as innovation, environmental sustainability, job creation, and supporting small to medium sized enterprises (OECD, 2019).

Government spending in this field encompasses various areas, including healthcare, environmental protection, public order, and economic affairs, where effective governance in these areas reinforces public services. For instance, healthcare spending among OECD nations increased from 29.3% to 31.9% of procurement expenditure in 2021, primarily due to increased purchases of health related goods during the COVID-19 pandemic. Other significant categories of public spending include economic affairs (16.4%), education (10.7%), defense (9.9%), and social protection (9.8%) (OECD, 2022).

This form of procurement also shapes the development of countries by allowing many nations to use contracts in order to improve their conditions and stimulate economic firm growth. It has been shown that gaining access to government contracts plays an important role in driving firm expansion across both the short and medium term (Ferraz et al., 2015). This trend has been particularly strong in developed economies, where public infrastructure spending enabled business expansion and development opportunities (Ortiz-Ospina & Roser, 2016).

The aims of public procurement contracts are diverse and include:

1. *Advancing Market Competition*

   Public contracting serves important purposes within the context of developing and sustaining competitive markets by removing barriers that limit competition. It achieves strategic objectives through the incentive management of suppliers, and the development of strong supplier relationships. These factors are critical to achieve enduring procurement efficiency and value (Caldwell et al., 2005).

2. *Attaining Cost Efficiency*

   The introduction of competitive tendering to public service delivery often leads to significant cost savings without compromising quality of service. For instance, a study from the United Kingdom indicated that such practices were able to reduce costs by approximately 20% (Domberger & Jensen, 1997).

3. *Forecasting Innovation*

   Public procurement is applied more often to stimulate innovation by influencing the demand side of the market. Governments apply targeted purchasing to encourage

suppliers to develop new solutions, which promotes industrial diversification and transformative innovation. This approach helps to align procurement with key industrial, and technological policy goals (Uyarra et al., 2020).

4. *Promoting Sustainability*

Procurement policies advance the social and environmental sustainability agenda, particularly the EU public procurement directives. These require contracting authorities to take life cycle costing into account in their purchases, conduct supply chain mapping and monitor breaches of environmental and social standards (Andhov et al., 2020).

5. *Adress Social Goals*

Procurement policies have also been used to promote employment and social inclusion. For instance, the Northern Ireland Unemployment Pilot Project highlighted how public contracts created job opportunities in public service and construction projects (Erridge, 2007). Additionally, construction firms were increasingly required to demonstrate support for community engagement and disadvantaged groups, often through partnerships with social enterprises (Loosemore, 2016).

6. *Balancing Discretion and Accountability*

An effective procurement system has to balance the discretion given to public officials with mechanisms that ensured transparency and accountability. This balance is required to guarantee efficient allocation of public funds and to prevent the procurement process from being overly influenced by political connections (Baltrunaite et al., 2018).

2.2. PUBLIC PROCUREMENT IN THE EUROPEAN UNION AND PORTUGAL

Public procurement in Portugal, as in the rest of the European Union (EU), is governed by a numerous of international, EU, and national legal frameworks. In Portugal, the legal framework is centered on the Code of Public Contracts (CPC), which establishes public contracts as an agreement between public authorities and private contractors, by acting as the contracting entity. This type of agreement aim to accomplish specific legal outcomes such as the purchase of goods and services or the execution of public works (Carneiro et al., 2020).

Historically, before the introduction of EU procurement directives, cross-border participation in public tenders was limited, where only around 2% of contracts were awarded to non-national companies, reflecting fragmented procurement markets and significant inefficiencies. This situation put in place the development of internal market rules to improve the efficient use of public resources and increase competitiveness. This allows governments to obtain high-quality goods and services at competitive prices while promoting cross-border participation (European Parliament, 2024).

Subsequently, the European Commission created a unified legal structure that aimed at improving transparency, equal treatment, fraud, and administrative burdens (Curado et al., 2021). Nowadays, public procurement not only facilitates the delivery of services but also represents more than 16% of the EU's GDP, indicating its importance from an economic perspective (European Parliament, 2024).

The EU's regulatory framework for public procurement has evolved over the years, with Directive 2014/24/EU replacing earlier Directive 2004/18/EC. This updated directive ensures fair and transparent procurement processes throughout the member states, maintaining fundamental principles of the Treaty on the Functioning of the EU, such as free movement of goods, freedom of establishment, and provision of services. It allows modern procedures such as competitive dialogue and electronic procurement to be introduced, alongside strategic objectives like social inclusion and environmental sustainability (Official Journal of the European Union, 2014).

Portuguese law alignes with these directives through a threshold-based system that defines applicable procedures. These procedures can be of different types. For example, the open procedure allows any interested operator to submit tenders, while restricted and competitive negotiation procedures limit participation to pre-qualified candidates. Competitive dialogue deals with complex, undefined requirements through pre-arranged conversations with selected participants. Innovation partnerships enable exceptional collaborative development, and exceptional cases allow contracts to be awarded without prior publication.

The tenders are awarded on the basis of the "most economically advantageous tender" criterion, which considers not only price but also other factors including quality, life cycle costs, innovation, social and environmental aspects (European Parliament, 2024). In support of transparency, Portuguese contracting authorities are under a formal obligation to disclose the evaluation criteria and their relative importance to the public in advance (Mateus et al., 2010).

Portugal was the first country in the EU to fully implement e-procurement by mandating electronic methods for open, restricted, and negotiated tenders as of the 1st of November 2009 (Costa et al., 2013). This achievement was facilitated with the use of the previously created BASE portal, that acts as a national repository for procurement data by consolidating information from the Portuguese Official Journal and certified electronic platforms. Since 2012, BASE began publishing all contracts with performance information, further increasing transparency and accountability (European Commission, 2017; IMPIC, 2023).

E-procurement is broadly defined as "the automation of the procurement process" (Vaidyanathan & Devaraj, 2008). It enables members within supply chains to procure products and services over the Internet (Pollock & Williams, 2009) while assisting in the detection of problematic areas and the formulation of streamlined procurement processes (Panayiotou et al., 2004). This technology allows to enhance supplier-buyer relationships and decision-

making processes while lowering transaction costs, with best practices including the use of paperless systems and automated approvals (Walker & Brammer, 2012; Costa et al., 2013).

However, some scholars argue against e-procurement's use as a tool to reduce corruption (Heeks, 1998), where overall, its effects are mixed. For example, while e-procurement strengthened law compliance in the Netherlands and improved administration efficiency in Denmark. In contrast, a similar reform in Portugal coincided with a drop in institutional quality and a rise in corruption and regulatory concerns (Mélon & Spruk, 2020). These findings underscore the need to examine the Portuguese case more closely.

According to the 2024 Corruption Perceptions Index, which rates 180 countries on public sector corruption from 0 (highly corrupt) to 100 (very clean), Portugal received a score of 57/100, positioning itself below the EU average of 62 (Transparency International, 2025). This is a score that highlights procurement integrity issues, as it can be seen in Annex B. Key challenges include limited resources for law enforcement and prosecution, court delays hindering dispute resolution, and emergency procurement measures during the COVID-19 pandemic that, despite parliamentary scrutiny, risked transparency due to their expedited nature (European Commission, 2021). Additionaly, public-private partnership models, especially in infrastructure projects, were noted for raising questions concerning cost efficiency (Macário et al., 2015).

To address these encounters, in alignment with initiatives like the Europe 2020 Strategy, the Portuguese Competition Authority prioritized procurement in an attempt to improve the allocation of resources. Other strategies included conducting impact assessments to improve legislative processes and the overall quality of legislation, although their effects were still under evaluation (Ferreira Gomes & Rodrigues, 2014).

In economic terms, public procurement constituted of 10.3% of Portugal's GDP in 2021, which was below the OECD-EU average of 14.9%, as shown in Appendix A. However, it still remained important for economic activity, presenting an increase since 2019 largely attributed to the Recovery and Resilience Facility, which encouraged public investment. Efficient procurement is thus essential to public administration, service delivery, but also to promote sustainable and inclusive growth (OECD, 2024b).

2.3. INTEGRITY RISKS AND IRREGULARITIES IN PRACTICE

Despite its potential benefits, public procurement, faces some distinct difficulties in its functions. The integrity of the field is defined as the ethical management of public funds, resources, assets, and authority, ensuring their use aligns with institutional purposes and ultimately serves the public good (Azmi & Rahman, 2015). Any deviation from this principle is considered unethical, suspicious or potentially criminal. Such breaches of ethics can occur through any stage of the procurement process, including tender design and documentation, contract awarding, execution and post-implementation (Schuster & Merjan, 2016). The

enduring presence of ethically questionable practices erodes public trust in government institutions and their integrity, undermining procurement systems (Chen et al., 2024).

An additional layer of complexity is that public sector procurement often differs significantly from private sector practices. This divergence requires the creation of new skills and approaches for efficient procurement management. Notably, the public sector continues to face substantial risks of fraud, corruption, and conflicts of interest, which are issues that are less common or managed differently in private business (Arlbjørn & Freytag, 2012).

In reality, a variety of procurement irregularities disrupts equitable competition, increases project costs, and reduces the quality of goods and services provided. These irregularities include:

1. *Collusion and Bid Rigging*

   Collusion takes place when competing firms work together to control pricing or the terms of bidding. A particularly harmful example of this is bid rigging, where firms agree, often through either bid rotation or market division, to predetermine contract outcomes. These strategies create inflated costs, reduce competition, and diminish market efficiency (Carbone et al., 2024).

   Price fixing is often done in parallel with bid rigging and includes coordinated practices such as complementary bidding, where firms submit purposely non-competitive bids to ensure the selection of a preferred winner (Kei Kawai et al., 2022).

   The Portuguese Competition Authority highlights public procurement as being particularly susceptible to collusive activity, stating that such practices increase costs while lowering the quality of services and restrict innovation (Autoridade da Concorrência, 2022).

   The most common forms of bid rigging observed in Portugal include (Autoridade da Concorrência, 2024):

   - Bid Rotation – Competitors take turns winning contracts.
   - Cover Bidding – Firms submit deliberately high bids to simulate competition.
   - Bid Suppression – Some companies agree not to submit bids.
   - Market Allocation – Markets are divided among firms based on customer type, region, or service.
   - Subcontracting – Losing firms are compensated through subcontracting agreements with the winning bidder.

2. *Fraud, Corruption, and Conflicts of Interest*

Fraud in public procurement usually involves the act of deception, such as false invoicing, cost inflation, and misrepresentation of qualifications, to gain an unfair advantage. Such practices have negative consequences on procurement performance and highlight the need for adequate internal controls (Matthew et al., 2013). Subcontracting schems are also extensive, where primary contractors collude with predetermined subcontractors or set up fictitious vendors to inflate project costs without delivering services. Frequently, these strategies mask profit sharing among conspiratorial firms, with repetitive subcontracting among the same entities signaling collusion (G. L. Albano et al., 2006).

Unlike fraud, corruption usually requires collusion by insiders, which can occur in the form of bribery or kickbacks, where officials taeke advantage of their positions of authority for personal gain. Corruption accounts for approximately 5% of the total value of public procurement, which is about 14% of EU's GDP (Modrušan et al., 2021).

Conflicts of interest occurs when procurement officials have personal or financial relationships with contractors, which leads to biased judgment and non-competitive contract awards. This greatly restricts access to a competitive market and creates unequal competition (Treisman, 2000). Effective procurement integrity relies on transparency laws, disclosure mandates, and strict supervision of the official-contractor relationship (Andvig & Fjeldstad, 2001). Moreover, research suggests that companies with high corruption risk scores tend to have greater profitability and stronger political connections, highlighting the need for more sophisticated procurement integrity safeguards (Fazekas et al., 2016).

3. *Restrictive Procurement Practices*

Restrictive procurement practices obstruct fair competition and transparency, resulting in inefficiencies, inflated costs, and limited supplier engagement. One common practice is unjustified sole sourcing, where authorities skip competitive bidding without a valid reason. This reduces accountability and simultaneously increases the risk of corruption and favoritism (Palguta & Pertold, 2017).

Similarly, direct awards permit contracts to be issued without competitive tendering, which can result in the selection of less competitive or more expensive suppliers (Tas, 2023). Another practice is scope manipulation, that involves designing contracts just under regulatory thresholds to avoid triggering open bidding requirements (Palguta & Pertold, 2017), often leading to contract concentration among a small group of firms and diminishes market competition (Hoekman & Onur Taş, 2024).

Striking a balance between discretion and oversight remains a challenge. While discretion in procurement can add flexibility, it also raises issues such as favoritism and reduce productivity among contractors (Baltrunaite et al., 2018).

4. *Non-Compliance with Procurement Regulations*

Failure to comply with procurement regulations undermines the integrity and efficiency of public procurement. Firms with prior misconduct frequently perpetuate violations of procurement laws, indicating a lack of sufficient enforcement mechanisms (Kistler et al., 2024).

Evidence suggests that compliance outcomes improves when organizations offer incentives, ethical behavior programs, and adequate trainings on procurement processes (Gelderman et al., 2010). However, the enforcement of EU directives remain inconsistent. In practice, local governments often show bias towards local competing suppliers, which represents a clear breach of non-discrimination clauses, weakening the regulatory framework's intent (Martin et al., 1999).

5. *Operational and Market Inefficiencies*

Operational inefficiencies diminish the procurement system's operational performance and adversely affect project delivery, value, and inclusion. One prominent example is the significant variance in bid prices, frequently caused by vague specifications and different levels of competence among bidders. These discrepancies indicate insufficient planning as well as poor communication during the procurement design phase.

Some contractors use unbalanced bidding techniques, such as front-end loading, where inflated costs are charged for initial stages to guarantee high paying contracts early on. These approaches distort financial plans and strain budgets. Similarly, errors in estimating quantities are taken advantage of, resulting in excessive costs, delays, and extended schedules (Hyari & Alamayreh, 2023).

A high frequency of change orders serves as another indicator of poor planning and insufficient project scoping. Such changes result from weak initial management and lead to budgetary inefficiencies (Ibbs et al., 2001). Bureaucratic delays also interrupt procurement schedules, especially within engineering, procurement, and construction settings, where encountering gaps in planning and design quality result in prolonged timelines (Herweg & Schmidt, 2020).

2.4. Anomalies and Their Impact in Procurement Contracts

The concept of anomalies was first introduced by Grubbs (1969), who defined an outlying observation as "one that appears to deviate markedly from other members of the sample in which it occurs". Broadly, anomalies are commonly described as nonconformities in data as unusual patterns that diverge from the expected behavior. Considering this, these irregularities can generally be divided into three types. A point anomaly, which occurs when a singular data instance deviates from a dataset's pattern. A contextual anomaly, that arises when a data point is considered anomalous in one context but normal in another. Lastly, a

collective anomaly, that refers to a group of data instances that, while individually unremarkable, exhibited unusual behavior as a group compared to the overall dataset (Ahmed et al., 2016).

In the context of public procurement, governments aim to achieve optimal cost-benefit outcomes, balancing price with quality. However, contracts are known for its vulnerability to irregularities (Lyra et al., 2022). Therefore, to preserve transparency, integrity, and efficiency in procurement processes, anomaly detection is of greatest importance.

These techniques, also known as outlier detection, play a critical role in uncovering irregular patterns that may indicate issues such as fraud or corruption. By identifying data points that deviate from expected norms, these methods contribute to more equitable and accountable procurement practices.

Their pratical application is relevant in identifying fraud. Financially wasteful activities of this nature are rampant during public crises, such as the COVID-19 pandemic (Dikmen & Çiçek, 2023), where because of their rapid nature and reduced oversight, emergency procurement processes are highly vulnerable to fraudulent manipulation.

Applications of anomaly detection have been widespread, spanning industries from credit card fraud to cybersecurity intrusion monitoring. Which in the case of public procurement, it enables the quality of purchases to be improved and sophisticated the developmental impact of the government's spending (Prasad et al., 2009).

Ultimately, identifying and addressing anomalies further advance the efficiency in the use of public resources and improve the public service delivery system (Niessen et al., 2020).

2.5. LEVERAGING NATURAL LANGUAGE PROCESSING

In public procurement, Natural Language Processing (NLP) techniques can be particularly useful for uncovering hidden patterns in textual datasets such as tender documents, contracts, and information on awarded entities. Algorithms such as clustering and topic modelling allow to discover and extract valuable insights from large datasets that contain unstructured data. This is particularly important, as many government procurement portals lack well structured data across several features. As a result, there is a sustained need for automation of procurement data analysis using ML methods tailored for this porpuse (Hott et al., 2023).

NLP encompasses a range of techniques that enables machines to interpret human language. These can include text classification (assigning text to predefined categories), named entity recognition (identifying entities such as people, organizations, locations, and dates), relation extraction (detecting semantic relationships), and terminology extraction (identifying domain-specific terms). Collectively, these methods are able to contribute to building structured

knowledge bases and improved information retrieval, even in data environments dominated by fragmented or ambiguous language.

Although NLP research traces back to the 1960s (Grishman & Sundheim, 1996), real-world adoption remains limited, being often constrained by differences in evaluation focus, long development timelines, and the need for interpretability as well as ongoing maintenance (Suganthan et al., 2015; Zhang et al., 2023).

One of the key components of many NLP methods is word representation, specifically through embedding words into lower dimensions by using neural networks. This technique creates dense vectors for each word, allowing semantically similar words to be mapped close together in vector space (Chiu et al., 2016). However, procurement data presents a unique challenge of integrating numbers with textual data.

Taken this into consideration, the ability to understand and manipulate numbers in word and digit form, referred to as numeracy, is essential to understand in this context. According to Wallace et al. (2019), while many NLP models treat numbers as standard tokens, embeddings such as ELMo, BERT and GloVe, demonstrate varying degrees of capturing numeracy. For example, algorithms like GloVe and Word2Vec can represent magnitude accurately for numbers up to 1,000, while character-level models such as ELMo outperform word and sub-word level models in this context. This highlights that embedding-based approaches can preserve accuracy in numerical interpretation within procurement analysis, but only until a certain extend.

Despite the general underutilization of NLP in procurement, its potential is evident, particularly in areas like anomaly and fraud detection. Embedding-based approaches have already been applied successfully to detect anomalies in other domains, including anomaly detection in images (Zavrtanik et al., 2021), networks (Yu et al., 2018), and medical fraud cases (Johnson & Khoshgoftaar, 2021). These successes indicate that similar methods can be adapted for identifying irregularities in public procurement datasets.

2.6. Review of Anomaly Detection Techniques

Detecting anomalies in procurement data remains challenging due to the varied nature of irregularities. In particular, the complexity of collusion among firms further complicates detection efforts, as firms utilize complex schemes to conceal their illicit agreements (García Rodríguez et al., 2022)

The ongoing digitization of public procurement has improved data accessibility, creating new opportunities for analytical techniques. While numerous researchers focused on employing statistical and ML techniques in this domain, the application of embedding-based approaches is still limited. One of the main unresolved challenges is the the lack of sufficient labeled data, which restricts the utility of supervised learning approaches. As a result, unsupervised and

semi-supervised anomaly detection methods have gained prominence because of their ability to identify irregularities without relying on predefined labels.

This section reviews several influential studies showcasing a variety of methodologies used to detect anomalies, highlighting how different analytical strategies have been applied across diverse contexts.

2.6.1. Corruption, Collusion and Fraud Detection in Public Procurement

In terms of a more classical approach, Fazekas et al. (2016) conducted extensive research on detecting corruption within public procurement, where one prominent study analyzed procurement data from Hungary to find indicators of corruption. This study created a Composite Corruption Risk Index (CRI) that used 13 predictive indicators, particularly in contexts involving competition and seasoned bidders. Some of the most important indicators included receiving only one bid per tender, exclusion of all but one bid in the evaluation, and a high "Winner's Share of Issuer's Contracts", which indicated recurring award of contracts to a small number of firms. The CRI's credibility was established through the observation of the outcomes providing attributes such as increased firm profits, contracts with a value higher than the estimate, and politically connected or companies registered as offshore.

Fazekas & Kocsis (2020) conducted a separate study that analyzed the institutional characteristics that influence the risk of corruption, focusing on bureaucratic quality in 212 European regions. From a dataset consisting of over 1.4 million procurement contracts and a survey of 18,000 public officials, it was evident that regions with stronger meritocratic bureaucracies demonstrated lower corruption levels. Their model suggested that improving bureaucratic meritocracy could decrease the prevalence of single-bid tenders by 7.5 to 11.6 percentage points, which could result in savings of €13 to 20 billion per year (in 2010 prices). This further solidified the role governance structures play in reducing corruption.

Westerski et al. (2021) created a rule-based model using 48 fraud indicators that could be understood and explained. Each procurement transaction was assigned a score according to these indicators, achieving a precision of 67.1%. The system was effectively put into operation as a component of real-world supervision workflows in Singapore, which illustrates the usefulness of interpretable rule-based systems in low data environments.

A supervised learning case was introduced by García Rodríguez et al. (2022) which studied collusive bidding from an international perspective, including countries like Brazil, Italy, Japan, Switzerland, and the United States. They defined cartel-like behavior with criteria such as identical bid amounts, submission of bids in a predictable order, and award contracts at prices significantly above the market. The study evaluated the performance of over 11 ML classifiers and 7 statistical screening variables, including coefficient of variation, skewness, and relative bid distance. Among the models tested, Extra Trees, Random Forest, and AdaBoost were shown to perform best, which confirmed the effectiveness of collusion detection, when labeled data is available.

2.6.2. Unsupervised and Semi-Supervised Anomaly Detection

In terms of the use of unsupervised learning, for instance, Niessen et al. (2020) used Isolation Forests on Paraguay's procurement data structured under the Open Contracting Data Standard. This approach computed anomaly scores for each procurement activity, which were subsequently tested against known irregularities for validation, including those that involved formal protests or complaints.

Network analysis emerged as another powerful tool for uncovering hidden relationships in procurement systems. As reported in the systematic literature review conducted by Lyra et al. (2022), using the PRISMA methodology, 48 studies were found that techniques such as implemented cluster analysis, community detection, and centrality measures, while including degree and betweenness centrality. These approaches were instrumental in revealing indirect associations involving bidders and authorities. However, multiple studies warned against an overdependence on network parameters, advocating for the integration of bidder behavior metrics to enhance the effectiveness of fraud detection in procurement networks.

In the context of fraud detection, Modrušan et al. (2021) analyzed indicators like short bid deadlines, early bid acceptance, high rates of bid rejection, and connections to known politically exposed persons or blocklisted individuals. However, they pointed out a significant gap in the research which is the lack of available data with explicit indicators of fraud. As a result, in many cases "suspicious" or "risky" labels were applied, that limited algorithmic validation in semi-supervised learning approaches.

Torres-Berru and Batista (2021) attempted to solve these issues by implementing a multi-phase anomaly detection model based on procurement data from Ecuador. Their pipeline combined Self-Organizing Maps to evaluate the impact of a particular feature, K-Means for clustering, and Support Vector Machines with PCA for a semi-supervised classification. The model was able to achieve more than 90% accuracy, revealing persistent irregularities, especially in consultancy contracts where the lowest bidder frequently did not win the contract.

2.6.3. Embedding-Based Representation and Analysis

In recent years, embedding-based techniques gained traction as a more scalable and automated way to detect procurement anomalies. Unlike more traditional approaches that relied on predetermined indicators, embeddings utilize representation learning to capture the semantic and structural patterns within text and graph data.

For instance, Hott et al. (2023) applied BERTopic for topic modeling in Brazilian procurement records. Their pipeline included preprocessing, dimensionality reduction through UMAP, and clustering using HDBSCAN. The use of the domain-adapted language model LiBERT-SE improved topic coherence, demonstrating the effectiveness of tailored embedding models in this context.

Similarly, Rahman (2022) explored procurement anomaly detection using multilingual embedding techniques. inBERT was used for Finnish language documents, while sBERT and RoBERTa Large were applied to English texts and translations. The study employed a standard pipeline involving dimensionality reduction with UMAP and clustering using HDBSCAN, preceded by targeted preprocessing steps such as the removal of numerals, geographic names, and other noise inducing elements. These refinements aimed to enhance clustering quality and better capture semantic patterns within procurement records.

Pastor Sanz (2022) utilized graph embeddings with Spanish procurement data collected during the COVID-19 pandemic. By generating embeddings of the procurement network using Node2Vec, the study clustered contracts based on their corruption risk levels. This was achieved through a two-stage process combining Self-Organizing Maps and Ward's hierarchical clustering. The model's validity was supported by comparing its outputs with media reported irregularities, confirming its relevance for real-world anomaly detection.

While nascent in procurement, embedding techniques had already been widely used in other domains. For example, in large social networks, Hu et al. (2016) introduced a method for detecting structurally inconsistent nodes using a clustering-based embedding model. This approach retained a high level of interpretability by linking each embedding dimension to specific network regions.

Yu et al. (2018) proposed a framework called NetWalk, which was intended for anomaly detection in dynamic networks. It used DeepWalk and Node2Vec for initial embedding, then applied autoencoders and clustering for real-time anomaly detection. NetWalk surpassed the most advanced benchmarks and demonstrated high efficiency and accuracy.

Johnson and Khoshgoftaar (2021), explored the use of embeddings in the healthcare sector in order to detect detecting Medicare fraud. Provider specialties were represented using GloVe, Med-W2V, and the more domain specialized HcpsVec and RxVec. These embeddings were assessed with classification models, while t-SNE visualization techniques revealed meaningful clusters that aligned with the provider types.

In contrast to private procurement, public procurement practices are still mostly underexplored. Current methodologies often rely on predefined indicator-based systems to identify hidden and multidimensional structures within unlabelled data. Consequently, there is a growing demand for innovative approaches that enable data-driven automation across various procurement scenarios. To address this gap, this study investigates embedding techniques, known for their ability to capture complex structural and semantic relationships, as a means to advance unsupervised anomaly detection in public procurement, with a particular focus on the Portuguese context.

3. METHODOLOGY

3.1. OVERVIEW

This research implements an unsupervised ML approach to anomaly detection utilizing public procurement datasets, having as primary focus the evaluating the effectiveness of embeddings in identifying anomalies through clustering techniques.

Due to the main challenge of scarcity of labels, that makes training supervised models based on labeled instances of known anomalies nearly impossible (Jin et al., 2023), a unsupervised learning approach is adopted to support pattern recognition and semantic understanding (Chandola et al., 2009). Specifically, by applying text embeddings to capture the semantic meaning of unstructured fields in procurement contracts, while also integrating structured information.

To process this heterogeneous data, a modeling strategy is employed that accommodates mixed data types. Unstructured text fields, categorical attributes, and boolean indicators are transformed into dense vector representations using trained language models. Numerical variables, however, are preprocessed separately and concatenated later on with these embeddings in order to create a combined feature representation. Dimensionality reduction techniques are applied in order to enhance computational efficiency and minimize noise. Finally, clustering algorithms are subsequently used to uncover potential anomalies that may indicate irregularities within the procurement process.

The methodology is illustrated in the figure below, with further detailed information being provided in the following sections of this chapter.

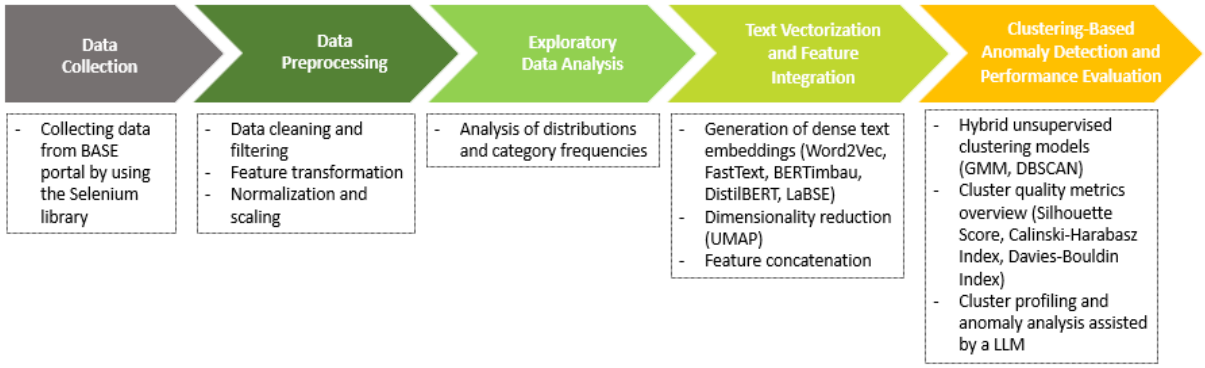


Figure 1 – Methodology Overview

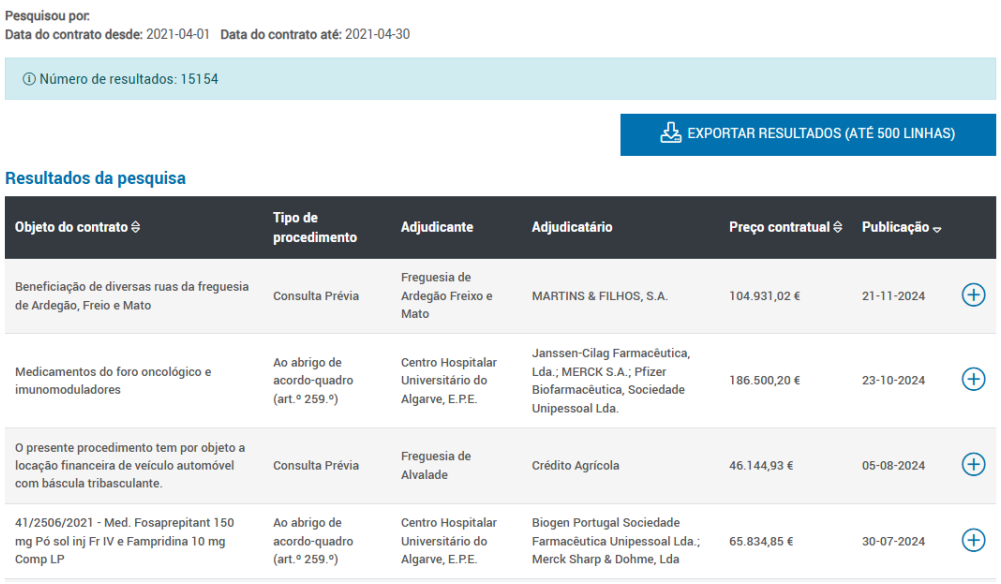
3.2. DATA COLLECTION

The data used in this investigation was obtained from the BASE portal (www.base.gov.pt) for the period 2020 to 2024. This platform serves as a public contracts repository system in Portugal, which has the purpose to gather all the relevant information on contracts awarded under the CPC and make it available for monitoring and oversight purposes.

The system is administered by IMPIC - Instituto dos Mercados Públicos, do Imobiliário e da Construção, the regulator of public contracts. In addition to its regulatory functions, IMPIC is responsible to perform some statistical reporting to the European Commission, as stated in Article 472 of the CPC that concerns public contracts that involve supplies, services, works and concessions (Portal BASE, 2024).

For the data collection process, to facilitate large-scale data extraction, a web scraping approach was employed using the Selenium Python library. Because of the considerable number of available contracts and the time-consuming nature of the work, data was collected gradually. Contracts were filtered and scraped by month, based on the contract date, which is defined as the date on which the final party signed the agreement.

An initial exploration of the platform showed that all contracts were listed in descending order, with the newest contracts prioritized at the top of the list. Each contract entry included a hyperlink to a dedicated detail page that contain more granular information presented in tabular format. This information is illustrated in the following figures.



Pesquisou por:
Data do contrato desde: 2021-04-01 **Data do contrato até:** 2021-04-30

ⓘ Número de resultados: 15154

EXPORTAR RESULTADOS (ATÉ 500 LINHAS)

Resultados da pesquisa

Objeto do contrato	Tipo de procedimento	Adjudicante	Adjudicatário	Preço contratual	Publicação
Beneficiação de diversas ruas da freguesia de Ardegão, Freio e Mato	Consulta Prévia	Freguesia de Ardegão Freixo e Mato	MARTINS & FILHOS, S.A.	104.931,02 €	21-11-2024
Medicamentos do foro oncológico e imunomoduladores	Ao abrigo de acordo-quadro (art.º 259.º)	Centro Hospitalar Universitário do Algarve, E.P.E.	Janssen-Cilag Farmacêutica, Lda.; MERCK S.A.; Pfizer Biofarmacêutica, Sociedade Unipessoal Lda.	186.500,20 €	23-10-2024
O presente procedimento tem por objeto a locação financeira de veículo automóvel com báscula tribasculante.	Consulta Prévia	Freguesia de Alvalade	Crédito Agrícola	46.144,93 €	05-08-2024
41/2506/2021 - Med. Fosaprepitant 150 mg Pó sol inj Fr IV e Fampridina 10 mg Comp LP	Ao abrigo de acordo-quadro (art.º 259.º)	Centro Hospitalar Universitário do Algarve, E.P.E.	Biogen Portugal Sociedade Farmacêutica Unipessoal Lda.; Merck Sharp & Dohme, Lda	65.834,85 €	30-07-2024

Figure 2 – Layout of the Contracts List Provided by BASE Portal

Informação detalhada

Data da publicação	21-11-2024
Tipos de contrato	Empreitadas de obras públicas
Nº do acordo quadro	Não aplicável.
Descrição do acordo quadro	Não aplicável.
Tipologia da medida especial	-
Tipo de procedimento	Consulta Prévia
Descrição	Consulta Prévia 01/2021 - Ardegão, Freixo e Mato
Fundamentação	Artigo 19.º, alínea c) do Código dos Contratos Públicos
Fundamentação para recurso ao ajuste direto (se aplicável)	Não aplicável
Regime	Código dos Contratos Públicos (DL 111-B/2017)
Critérios materiais	-
Entidades adjudicantes	Freguesia de Ardegão Freixo e Mato (510832865)

Figure 3 – Layout of the Detailed Contract Information Page on BASE Portal

To automate the scraping process, the HTML structure of both the listing and detail pages was examined. A Selenium-based script was created to perform multiple actions in an systematic manner. The script launches a Google Chrome browser, navigates to the BASE portal website with specific search filters (e.g. December 2020), and automatically clicks on the detail links for each listing to open the contract pages in new tabs.

From the tables of Detailed Information and Contract Execution, semi-structured data was collected. These tables contain fields such as contract type, publication and contract dates, procedure types, contracting and contracted entities, contract values, execution timelines, and other pertinent attributes. After retrieving all the contracts from a page, the scraper uses the arrow buttom to continue through the paginated results.

The last page of every month, however, is not reachable through the navigation arrow and can only be accessed by selecting the page number. As a result, the last contract entries for each month were manually located. All acquired information was held in a Python dictionary and later exported to a Pandas DataFrame for additional preprocessing.

The initial dataset compiled through this method includes 902,767 records across 35 variables, all in Portuguese. To facilitate understanding, both feature names and values used or created were translated into English throughout this document. However, for clarity regarding the original terminology, a glossary of common used procurement terms is offered in Appendix B.

Table 1 – Number of Public Procurement Contracts in Portugal per Year

Year	Number of records
2020	149,727
2021	177,161
2022	173,292
2023	188,568
2024	214,019

The previous table displays the total number of procurement contracts captured for the years 2020 to 2024. The data suggest a strong upward trend with an approximate 43% increase on the number of over this period. Despite a slight decrease in 2022, the overall trend points to increasing procurement activity. Where the peak in 2024 may reflect changes in administrative practices, with more contracts being signed.

3.3. DATA PREPROCESSING

One of the first steps in this study involved cleaning the dataset to identify and correct any issues present in the original data. This ensured coherence and readiness for the subsequent stages of analysis by improving overall data quality. A more detailed description of the steps taken in this specific analysis includes the following:

1. Handling missing values and dropping irrelevant features
2. Fixing data types
3. Removing duplicates
4. Feature enginnering and correlation analysis
5. Resetting the index
6. Normalization and scaling features

In the first step, it needed to be taken into consideration that every field was saved initially as text , regardless of their actual data type, since Selenium does not inherently interpret content types during scraping. In this phase inconsistent representations of missing values were noticed, where spaces, dots, hyphens, and asterisks were frequently due to manual data entry practices. As a result, 27 out of 35 features were containing missing values.

To improve data quality and reduce noise, irrelevant features were removed. For example, the feature intended to capture justifications for direct award procedures was erased since it had missing values across all records. Likewise, fields that contained only hyperlinks, such as references to official announcements or supporting documents were also dropped since their content was not extracted during scraping and could not contribute to the analysis.

Several indicators such as the use of environmental or material criteria were noted to follow a pattern where relevant entries contained a specific keyword or phrase, while all others were left blank. These variables were accordingly encoded as true or false, allowing them to be treated as pseudo-boolean features while remaining compatible with embedding techniques.

For conceptual numeric features like contract price and total effective price, missing values were filled using median imputation. Additionally, boolean indicators were created to flag the presence of missing values, including contract closure dates. This allowed missingness itself to be treated as a potential anomaly signal, rather than being artificially corrected. Often, these missing values reflected contracts that had not yet been finalized.

Several descriptive fields, for example those indicating the type of special measure or termination reasons, had a significant number of empty entries. In these cases, missing information was filled with Not applicable, following the terminology convention used by the original data source. For other text features, such as the ones related to entities or location of execution of the contract, missing data was replaced as Not informed to clearly distinguish between irrelevant and absent information.

In the second step, data types were corrected. Monetary values were converted to numerical values, dates were transformed into a consistent structure suitable for chronological comparisons, and certain count-based features were converted initially to discrete numbers. While some attributes could be treated as categories, they were kept in textual form to support the application of embedding techniques later in the analysis.

In the third step, records were found and removed for exact duplicate entries using the identification number of the contract. Cases that appeared nearly identical but differed in key details such as the supplier or contract amount were retained, as they referred to distinct procurement agreements. After this stage, this unique identifier used during data extraction was discarded, as it no longer served an analytical purpose.

In the fourth step, feature engineering began with the transformation of the variable related to the Common Procurement Vocabulary (CPV) code, which is a classification system for public procurement contracts, developed by the European Commission under Regulation 213/2008. Each CPV code consists of nine characters and is organized hierarchically, where the first 2 characters represent one of 45 major divisions. Given this hierarchical structure, the codes were shortened to their first two digits to enable a broader procurement subjects.

Division	**42**00000-0	Industrial machinery
Group	42**5**0000-1	Cooling and ventilation equipment
Class	425**3**000-0	Parts of refrigerating and freezing equipment and heat pumps
Category	4253**100**-7	Parts of refrigerating equipment

Figure 4 – Example of the CPV Code Structure

Date related information, such as contract dates, publication dates, and closure dates, was used to create a simple indicator showing whether these events occurred on a weekend. This indicator followed the same textual encoding format used for previous boolean style variables, ensuring compatibility with embedding methods.

There is a strong emphasis on retaining as many text features as possible, not only due to the use of embeddings but also because of other factors. First, as mentioned previously on the literature review by Wallace et al. (2019), numeracy in the embeddings can only be captured to a certain extend. Second, numerical features appeared to introduce excessive noise into

the anomaly detection process, resulting in too many false positives. This issue is illustrated in Appendix C, where test trials on a small sample (1/12 of the dataset) revealed how numerical features negatively impacted model performance. These trials formed the foundation of the finalized methodology.

To address this challenge, more detailed numeric breakdowns of date information, such as day, month, or year components, were intentionally avoided. Additionally, some features that would be classified as numeric were transformed and treated as text. Only those numerical features believed to have a better chance of drawing meaningful inferences about possible anomalies were considered. As a result, most original date fields were excluded, except for newly created indicators which reflect whether certain key dates fell on weekends or in some cases if they are missing.

Afterwards, as illustrated below, the relationship between numerical features was analyzed using a Spearman correlation matrix.

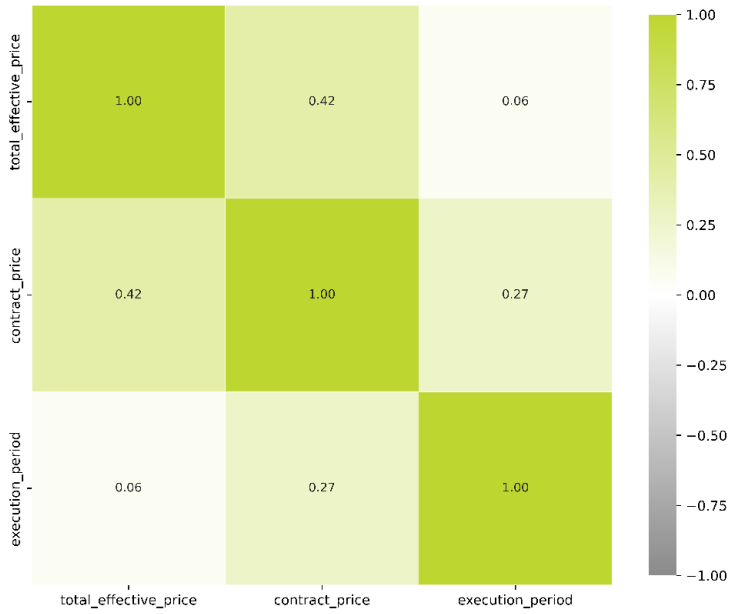


Figure 5 – Spearman Correlation Matrix for Numerical Features

Notably, the highest correlation is observed between contract price and total effective price, showing a moderate correlation of 0.42. However, this level of association does not indicate sufficient redundancy to justify removing either variable.

The contract execution period was transformed into a text and grouped into broader duration categories to simplify interpretation, as follows:

- Short duration: less than 100 days
- Média duração: between 100 and 300 days
- Long duration: between 301 and 600 days
- Very long duration: exceeding 600 days

In the fifth step, the dataset's index was reset and a separate copy was saved. This copy would serve the purpose of tracing the detected anomalies to the contract records later on while ensuring that the index remains unchanged throughout the process.

In the final step, the few numerical features remaining, related to contract price and total effective price, were adjusted using the Yeo-Johnson Power Transformation to reduce skewness and approximate a normal distribution (Yeo & Johnson, 2000). This was applied because the original distributions were largely left-skewed, as illustrated below:

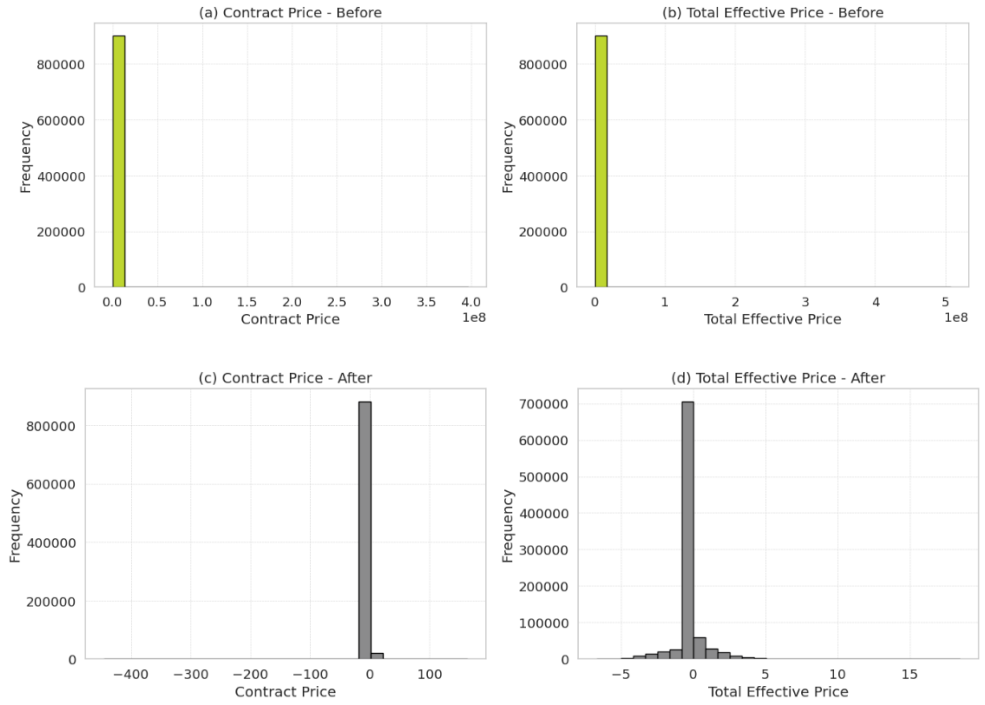


Figure 6 – Numerical Features Before and After Yeo-Johnson Transformation

It can be seen that the transformation slightly improved the distribution of the features, so for that reason, it was maintained even if it was not a major difference.

Subsequently, all numerical variables were standardized by applying StandardScaler, which ensures that each feature has a mean of zero and a standard deviation of one (Kappal, 2019). Scaling is significant in maintaining the comparability of features for distance and projection-based models. This standardization is given by the following formula:

$$z = \frac{x-\mu}{\sigma} \qquad (2)$$

Where z is the standardized value, x is the original value, μ is the mean of the feature and σ is the standard deviation of the feature.

Upon finalizing all the initial preprocessing stages, the dataset utilized for this study presented 902,627 records. To gain a deeper understanding of the final 32 features used in this project, the table below was created, where features are organized in a logical order, beginning with the original variables and concluding with six engineered boolean indicators used for flagging specific conditions.

Table 2 – Data Description

Feature	Conceptual Data Type	Transformed Data Type	Description	Example
Contract Price	Decimal Number	Decimal Number	Initial contract cost agreed upon at signing	6,012.50 €
Total Effective Price	Decimal Number	Decimal Number	Final contract cost at completion	7,520.30 €
Framework Agreement Number	Identifier	Text	Unique code that identifies the framework agreement, which can be used by multiple related contracts	4641215
CPV	Categorical	Text	Common Procurement Vocabulary code	33
Contract Type	Categorical	Text	General category of the contract	Acquisition of movable goods
Procedure Type	Categorical	Text	Type of procurement procedure used	Under framework agreement (Art. 259)
Regime	Categorical	Text	Legal or administrative regime governing the contract	CPC (DL111-B/2017) and Law No. 30/2021, of 21.05
Special Measure Typology	Categorical	Text	Specifies if the contract falls under any special procurement regime	Recovery and Resilience Plan (RRP) – Article 6 of Law No. 30/2021
Justification	Categorical	Text	Explanation provided by the contracting entity for selecting its procurement approach	Article 259 of the CPC
Justification no Written Contract	Categorical	Text	Describes the lack of a formal written contract	Article 95, No. 1, paragraph b), leasing or acquisition of movable property or acquisition of services under a public supply contract
Execution Period	Categorical	Text	Planned contract duration in days	Short duration
Execution Location	Categorical	Text	Country, District and Municipality where the contract was celebrated	Portugal, Évora, Estremoz

Notice	Categorical	Text	Refers an associated public announcement of the contract	Contract under review! Status: in rectification
Framework Agreement Description	Text	Text	Description of the related framework agreement	CP 2020/84 - Framework Agreement for the supply of wheelchairs to SNS institutions and services
Description	Text	Text	General description of the contract	Q7/2535/2021 Acquisition of wheelchairs
Contract object	Text	Text	Subject matter of the contract	Acquisition of Wheelchairs for SNS Institutions and Services, Goal i1.09 – Modernize Equipment, under SPMS CPAs
Contracting Entities	Text	Text	Entities responsible for launching and managing the contract	Regional Health Administration of Alentejo (503148768)
Competing Entities	Text	Text	Bidders that submitted proposals and did not win the tender	F.S.A. Digital Med, Ltd. (510908497), Anastácio Saldanha PLC. (505804441), Epjmédica (506820513)
Awarded Entities	Text	Text	Entity that awarded the contract	Ergométrica - Support, Orthopedic and Elevation Products, Ltd. (502111216)
Observations	Text	Text	Additional notes related to the contract	The proposed quantities were adjusted to the packaging units
Contract Termination Cause	Text	Text	Reason provided for premature termination of the contract	Full execution of the contract
Causes of Deadline Changes	Text	Text	Reason for changes in the final execution date of the contract	Project redesign needed
Causes of Price Changes	Text	Text	Reasons for increases or decreases from the initial contract price	Supply of additional equipment
Material Criteria	Boolean	Text	Checks if the contract follows material criteria	True / False
Centralized Procedure	Boolean	Text	Checks if the contract satisfy the needs of multiple entities	True / False
Environmental Criteria	Boolean	Text	Checks if the contract is sustainable	True / False
Missing Contract Price	Boolean	Text	Checks if the contract price is missing	True / False
Missing Total Effective Price	Boolean	Text	Checks if the contract final price is missing	True / False

Missing Contract Closure Date	Boolean	Text	Checks if the contract closure date is missing	True / False
Publication Date Falls on a Weekend	Boolean	Text	Checks if the publishing date of each contract on the portal was on a weekend	True / False
Contract Date Falls on a Weekend	Boolean	Text	Checks if the date when the contract was signed was on a weekend	True / False
Contract Closure Date Falls on a Weekend	Boolean	Text	Checks if the date in which the contract object was finalized was on a weekend	True / False

3.4. EXPLORATOTY DATA ANALYSIS

The Exploratory Data Analysis was conducted on raw but clean data to gain a better understanding of the overall landscape of public procurement. This analysis focused on visualizing distributions and category frequencies, which is particularly important for contextualizing the anomalies identified later since it provides insights about the regular structure of the contracts.

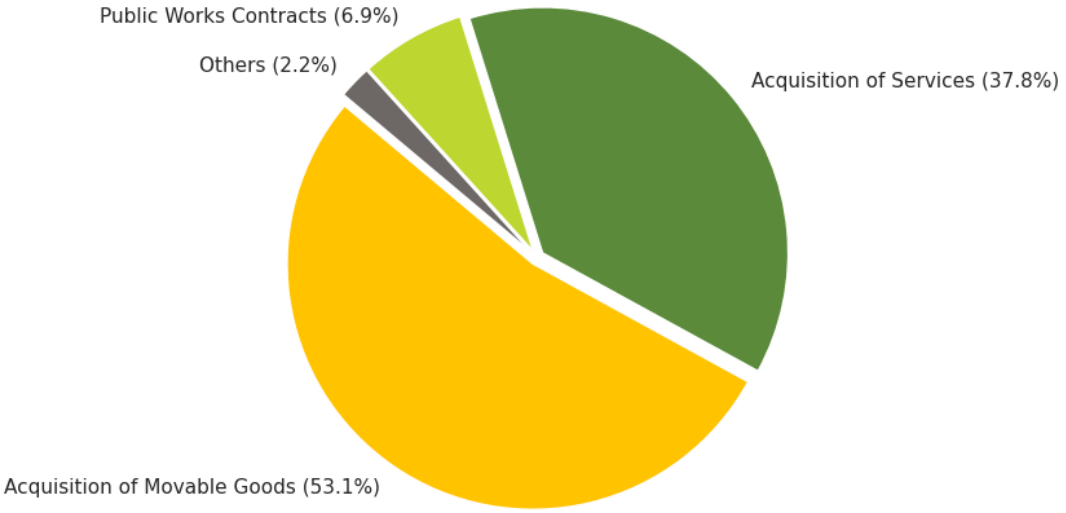


Figure 7 – Most Common Contract Types

The chart above reveals the most common contract types. Where it is noticeable that Acquisition of Movable Goods accounts for the majority of contracts, representing 53.1% of the dataset, highlighting its significant role in public procurement. This is followed by Acquisition of Services, which makes up 37.8% and Public Works Contracts that represent a smaller share of 6.9%. Although other contract types exist, they have been grouped under

Others, which represent 2.2% of the total data. This distribution underscores a strong emphasis on operational and service related needs in procurement practices.

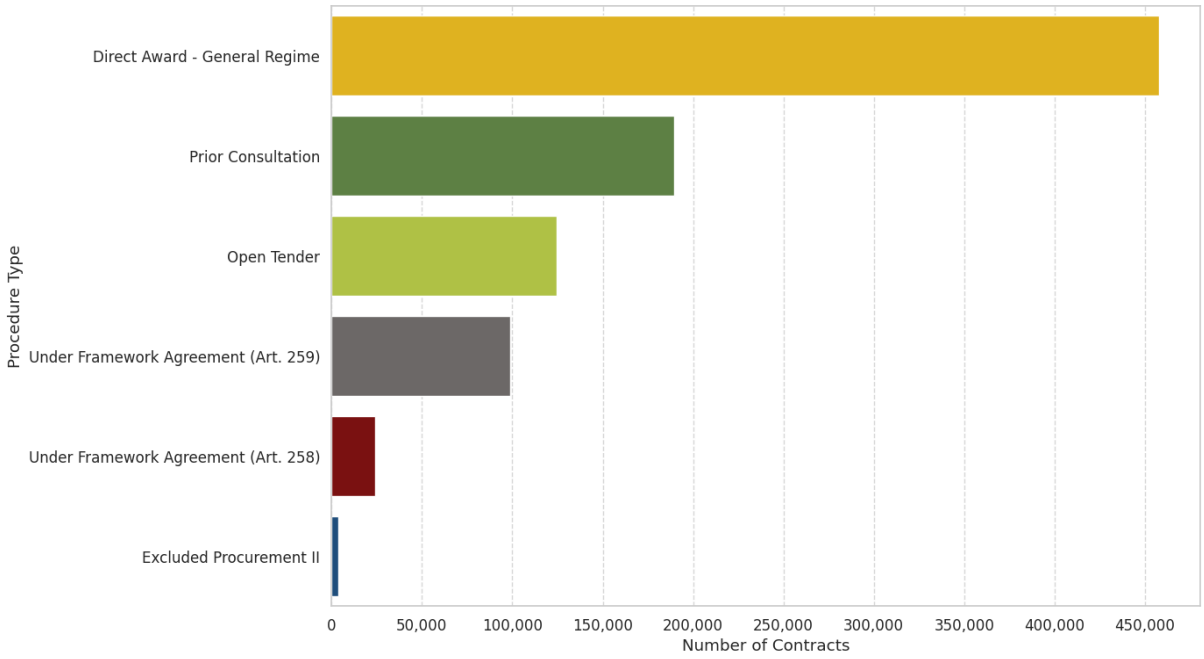


Figure 8 – Most Common Procedure Types

In terms of frequent procedure types, Direct Award - General Regime is by far the most commonly used, accounting for the highest number of contracts, which surpasses 450,000. It is followed by Prior Consultation, with just under 200,000 contracts, and Open Tender, which has around 120,000, also representing significant portions of the dataset. Notably, from Under Framework Agreement (Art. 258) onward, the number of contracts drops sharply, showing that the following procedure types have minimal representation. These results highlight a clear preference for simplified and standardized procurement methods in public contracting, with a strong concentration in a few dominant procedure types.

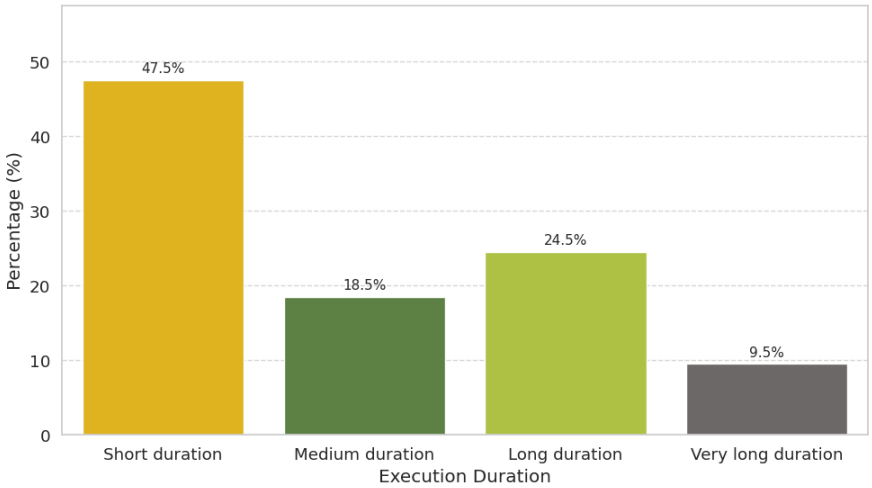


Figure 9 – Percentage Distribution of Execution Period Categories

The pattern for the distribution of execution periods shows that short duration contracts represent the largest share of 47.5%, which indicates a strong tendency toward quick term engagements in public procurement. Medium duration and long duration contracts also occur with notable frequency, accounting for 18.5% and 24.5% respectively, reflecting a balanced use of moderately timed projects. However, contracts of very long duration are the least common, making up only 9.5% of the total. This may suggest that projects requiring extensive timelines are either less typical in public sector operations or subject to stricter controls and planning procedures.

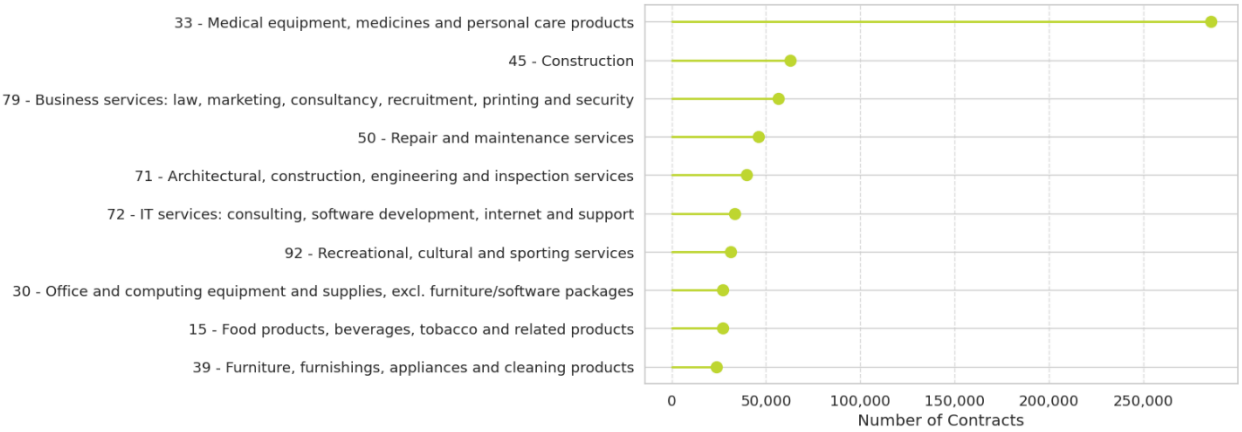


Figure 10 – Most Frequent CPV's in Contracts

An analysis of the most frequent CPV divisions reveals that the code 33—Medical Equipment, medicines, and personal care products, far surpasses all others in contract frequency, exceeding 250,000 contracts. The next most most common nine divisions range from roughly somewhere between 25,000 to 65,000 contracts, highlighting a strong concentration of procurement activity in the healthcare and medical sector.

3.5. Text Vectorization and Feature Integration

3.5.1. Generation of Text Embeddings

In public procurement, much of the available information exists in unstructured textual format, where elements like descriptions, and details of the awarded entities are usually captured in natural language. These text features are crucial to the detection of anomalous behavior since they can signal unusual patterns, inconsistencies or suspicious formulations.

However, ML models do not work with raw text. Unlike numbers, text has to go through transformation to a structured format that retains meaning, which is where text embeddings prove to be useful. Text embeddings transform words, sentences, or documents into high-dimensional numerical vectors. In this space, elements of text that are semantically identical are placed closer and dissimilar elements are further apart (Mikolov et al., 2013), making embeddings powerful tools for capturing textual meaning for modeling.

Sparse vector techniques were excluded intentionally due to their inherent limitations. Sparsity in representation occurs within extremely high dimensional spaces (H. Zhang et al., 2024), thus creating a challenge for storage and retrieval, particularly with large datasets. Additionally, sparse vectors merely indicate word presence or frequency, lack semantic nuance, being vulnerable to vocabulary mismatches and reducing robustness when encountering previously unseen words.

Taking that into consideration, this study employs 5 embedding models adapted to Portuguese: Word2Vec, FastText, BERTimbau, DistilBERT, and LaBSE. Each model transforms input text into dense high-dimensional vectors with fixed-length that capture semantic relationships in the data.

Word2Vec, was implemented using the Skip-gram variant that predicts surrounding words from a target word. It is a non-contextual embedding method, assigning the vector to a word regardless of context (Church, 2017). While simplistic, it tends to perform acceptably on shorter texts. Word2Vec's requires prior preprocessing, such as tokenization, lowercasing, lemmatization, and removal of stopwords, with the exception of the word No which was preserved for its contextual significance in legal discourse.

FastText represents words as collections of character n-grams (Church, 2017). This capability improves dealing with rare words, typos, and morphological variants, which is particularly relevant for Portuguese. Like Word2Vec, FastText remains non-contextual in nature yet captures subword information which aids in tolerance out-of-vocabulary terms.

BERTimbau is a Bidirectional Encoder Representations from Transformers (BERT) based model pre-trained on Brazilian Portuguese corpora (Yao et al., 2020). It produces embeddings relatively to the context of the sentence, and thus, they reflect the meaning of the given words in the specific context. This is crucial for legal texts as many words, such as the portuguese word for Value, has multiple interpretations. Despite the rich contextual understanding BERTimbau offers, it has a higher computational cost, especially on longer or more complex texts.

LaBSE and DistilBERT were also utilized to obtain sentence-level embeddings. DistilBERT is a more efficient version of BERT, which underwent knowledge distillation, allowing it to be much more efficient than BERT while retaining most of its accuracy (Sanh et al., 2019). It provides a blend of efficiency and representational power, making it useful for large scale anomaly detection. LaBSE was selected for its robustness in capturing nuanced justifications and contract clauses because it excels at generating meaningful sentence embeddings across multiple languages (Feng et al., 2020).

All embedding pipelines followed a consistent workflow: preprocessing or loading text, generating embeddings, and processing in batches to optimize resource usage. To avoid redundant computation, embeddings were stored as parquet files, which support efficient storage and preserve DataFrame indices for alignment.

For each feature, generated embeddings were saved with uniquely prefixed column names to prevent conflicts. After all features were processed, the embeddings were concatenated horizontally. This stacked representation is then saved for later use.

3.5.2. Dimensionality Reduction Techniques

As mentioned previously, the vectors produced by the embedding models are high-dimensional, which increases the risk of overfitting and incurs computational costs. To resolve these issues, the non-linear Uniform Manifold Approximation and Projection (UMAP) algorithm was applied for dimensionality reduction.

All embeddings underwent an L2 normalization to unit length prior to UMAP processing. This step is crucial as UMAP uses cosine distance as its primary metric. Normalization ensures that cosine similarity can be treated as a directional measure, improving the likelihood that important semantic distances will be preserved during projection.

UMAP is particularly useful at reducing the dimensionality of data while preserving local neighborhoods and retaining more of the global data structure compared to other techniques (McInnes et al., 2018). It first constructs a graph based on high-dimensional relationships using the 15 nearest neighbors and then seeks an optimized low-dimensional representation.

In this study, UMAP was applied to the text embeddings which were originally of varying dimensionalities, where existed 3,000 features for Word2Vec, 9,000 for FastText, and 23,040 for BERTimbau, DistilBERT, and LaBSE. The output dimensionality was empirically tuned and set to 50 dimensions to strike a balance between preserving important semantic structure and noise reduction. To ensure a fair comparison, UMAP was fitted a randomly selected subset of 100,000 records across all embedding types due to memory constraints. This controlled setup allows differences in performance to be attributed to the embeddings themselves rather than varying input volumes.

Earlier experiments using Principal Component Analysis (PCA) with 50 dimensions were also conducted. PCA is an example of a linear method that attempts to capture global structure but lacks local, non-linear semantic arrangements (Kadappa & Negi, 2016). Unfortunately, PCA didn't seem appropriate for representing these text embeddings, as it can be shown in Appendix D. A reason for this situation is likely associated that these embeddings depend more on local relationships to identify subtle deviations than on global variance.

To maintain consistency across data types, the reduced embeddings were adjusted using StandardScaler, which is consistent with the approach taken for the numerical features.

3.5.3. Integration with Numerical Features

Apart from the use of textual features, certain structured numerical variables such as the contract price and the total effective price, were considered for anomaly detection. These

features were joined with the previously created reduced UMAP embeddings to create a unified feature vector for each record.

With this this integrated representation to create feature spaces, the model is allowed to capture both semantic outliers (e.g., justifications that are out of the ordinary) and deviations in quantitative metrics (e.g., pricing that is out of proportion).

3.6. CLUSTERING-BASED ANOMALY DETECTION AND PERFORMANCE EVALUATION

Anomaly detection encompasses a variety of techniques to isolate procurement contracts with anomalous behaviors. In this context, clustering-based and density-based methods are adopted for anomaly detection, which aim to reveal structures and uncover anomalous behaviors without any prior input (Ester et al., 1996). Although difficult to validate exhaustively, this approach of an unsupervised ML approach is well-suited to the characteristics of Portuguese public procurement data, which lacks labeled instances of anomalies.

3.6.1. Hybrid Strategy and Label-Free Detection

This study employs a clustering-based anomaly detction approcah using two complementary algorithms: Gaussian Mixture Models (GMM) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), that function in a integrated semantic and numerical feature space.

GMM analyzes dense, inlier regions of data using a probabilistic approach. The GMM model assumes that the data has been generated by a mixture of Gaussian distributions, and it computes the likelihood of each data point with respect to the overall data distribution (Reynolds, 2009). Points in high-density regions receive higher log-probability values, whereas points in low-density regions may contain subtle outliers.

For this use case, contracts with log-probability values in the lowest 5% of the distribution were considered anomalous. Log-probabilities are useful due to their improved numerical stability. A higher value means the contract aligns closely with expected data patterns, while a lower value suggests the contract is less likely to be under the model and may be an anomaly.

Additionally, small clusters were defined as those whose sizes fall below the 35th percentile, meaning that they belong to the smallest 35% of all clusters when sorted by size. These small clusters potentially represent less common or more unusual groups of contracts compared to the larger, more typical clusters.

As a complement to GMM, DBSCAN was used, which is a non-parametric density-based clustering technique that is particularly effective in identifying outliers in datasets that do not have globular cluster shapes, since it doesn’t assume any data distribution. DBSCAN classifies points as core, border, or noise based on the local point density (Ester et al., 1996). Its primary strength is identifying outliers as low-density noise points that do not belong to any dense

region. The hyperparameters of neighborhood radius and minimum points required to form a dense region were set empirically to control the level of detail in the clusters while maintaining the approach's ability to detect outliers. DBSCAN is effective at filtering out contracts that are isolated or grouped sparser than what the dominant patterns would suggest.

The combination of GMM and DBSCAN offers a stratified approach to anomaly detection. GMM captures subtle, intra-cluster anomalies by evaluating the log-probability of data points within clusters, identifying deviations from expected cluster behavior. In contrast, DBSCAN is applied globally to detect more severe, density-based outliers, flagging data points that do not belong to any dense region as noise.

Anomalies are defined as records that exhibit low log-probability within their assigned GMM cluster and are simultaneously flagged as noise by DBSCAN. This situation is made possible because GMM and DBSCAN generate different cluster structures, where some anomalous points can coincide with the GMM clusters and others not.

This intersection approach was preferred over using a union strategy because anomalies flagged by both methods tend to be more distinct and consistent, making them easier to inerpret while reducing false positives. Since GMM and DBSCAN rely on different criteria, and may focus on different features when identifying outliers, using the union of their results could lead to a mixed set of anomalies with little in common, making interpretation more difficult.

3.6.2. Performance Evaluation

Because this study is done in an unsupervised manner, the use of traditional supervised metrics such as precision, recall, and F1-score could not be applied. Instead, the internal clustering metrics shown below and qualitative interpretability were used to assess performance.

Table 3 – Cluster Quality Evaluation Metrics

Metric	Description	Formula	Range
Silhouette Score	Measures how similar a point is to its own cluster compared to other clusters. Higher scores indicate better defined clusters.	$s = \frac{b - a}{\max(a, b)}$ Where a is the average distance between a point and all the other points in the same cluster and b is the smallest average distance between that point and all the points in the closest different cluster.	[-1, 1]
Calinski-Harabasz Index	Evaluates the ratio of dispersion between clusters to dispersion of points inside each cluster. Higher values suggest dense, well separated clusters.	$CH = \frac{TR(B_k)}{TR(W_k)} \times \frac{n - k}{k - 1}$ Where n is the total number of samples, $k\ is\ the$ number of clusters, $TR(B_k)$ is the trace of the between-cluster dispersion matrix and $TR(W_k)$ is the trace of the within-cluster dispersion matrix.	[0, ∞[

Davies-Bouldin Index	Computes the average similarity between each cluster and its most similar one, by comparing how close points are inside a cluster to how far apart the clusters are from each other. Lower values indicate better clustering.	$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$ Where k is the number of clusters, σ_i is the average distance of all points in cluster i to the cluster centroid c_i and $d(c_i, c_j)$: distance between centroids of clusters i and j	[0, ∞[

A visual examination was also performed to assess the GMM clusters and the anomalies detected by the models through the use of dimensionality reduction projections. In this case, UMAP was selected as the primary visualization method, as it preserves both local and global structures more effectively than t-distributed Stochastic Neighbor Embedding (t-SNE), while faithfully reflecting the geometry of the clusters (Marx, 2024). Although t-SNE is particularly strong at maintaining local neighborhood relationships, it tends to distort global distances and produce inconsistent layouts across runs. However, since it can still provide useful insights t-SNE visualizations were also generated for selected cases and included in a appendix for further comparative purposes.

Lastly, an in-depth review of some of the most relevant GMM clusters and flagged anomalous contracts was performed by matching the index of the results dataset with the actual human readable contracts that exist before the embedding generation phase.

To support the analysis of these elements, it was necessary to summarize patterns in contract types, procedures, and other relevant fields using Generative Pre-trained Transformer (GPT), a large language model (LLM) developed by OpenAI. GPT is based on a decoder-only Transformer architecture, trained on vast amounts of general text corpora. This artificial intelligence (AI) model generates coherent, context-aware summaries by predicting the most likely next words in a given prompt, while being able to produce a human-like writing output (Adhikari & Dhakal, 2023). In this study, the model GPT-4o Mini was prompted with the total contract data in order to extract and express their semantic commonalities and differences in natural language.

While manual inspection was applied selectively, the majority of the analysis of flagged anomalous contracts and cluster profiling was guided by the output of the LLM. These elements were generated using a structured prompt design that incorporated both quantitative summaries (e.g., mean, standard deviation, minimum, and maximum of contract prices) and representative textual records from each cluster. Numerical features such as contract price and total effective cost were statistically described and included alongside examples of contract descriptions and other text fields. This combined input was then used to prompt the LLM, enabling it to generate rich, context-aware summaries of typical behavior within each selected cluster. The LLM also assisted in explaining why certain flagged contracts deviated from these profiles, offering semantic interpretations of anomalies.

By using this approach, the process enabled a more structured, interpretable, and scalable examination of potential irregularities in public procurement data. Targeted human review was mainly focused in adding additional information for the analysis by comparing specific cluster profiles against the overall dataset, while also selectively verifying the reliability of the outputs provided by the LLM.

4. RESULTS AND DISCUSSION

4.1. EMBEDDING EVALUATION AND CONFIGURATION STRATEGY

To ensure the effectiveness of anomaly detection in public procurement data, careful selection of the embedding type and clustering configuration was fundamental. Five embedding techniques were evaluated: Word2Vec, FastText, BERTimbau, DistilBERT, and LaBSE. Each was assessed through GMM clustering using cluster counts ranging from 50 to 100.

The following table presents the clustering quality metrics Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index, used to compare performance.

Table 4 – Clustering Quality Metrics by Embedding Type and Number of GMM Clusters

Embedding Type	Number of Clusters	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
Word2Vec	50	0.3449	78,934.76	1.2471
	60	0.2819	79,860.59	1.1722
	70	0.2575	78,698.48	1.2444
	80	0.2570	80,186.13	1.1490
	90	0.2655	83,204.83	1.2612
	100	0.2628	84,850.73	1.1622
FastText	50	0.5349	111,447.31	1.1501
	60	0.4916	114,527.66	1.5018
	70	0.4784	114,584.68	1.4240
	80	0.4728	111,956.62	1.5163
	90	0.4541	110,326.82	1.7452
	100	0.4313	103,666.38	2.0625
BERTimbau	50	0.4333	125,010.27	1.1044
	60	0.4388	132,413.18	1.0970
	70	0.4175	128,941.33	1.1561
	80	0.3776	126,480.63	1.3155
	90	0.3419	121,272.81	1.4444
	100	0.3112	124,645.26	1.6659
DistilBERT	50	0.4848	134,868.91	1.2806
	60	0.4309	133,028.51	1.3089
	70	0.4087	136,641.60	1.3060
	80	0.3821	129,789.01	1.5084
	90	0.3600	137,575.55	1.5116
	100	0.3470	133,013.62	1.4540

LaBSE	50	0.4945	121,395.49	0.9743
	60	0.4862	121,466.49	1.1426
	70	**0.4772**	**123,690.20**	**1.1453**
	80	0.4722	130,455.99	1.2215
	90	0.4721	126,682.48	1.3503
	100	0.4090	125,334.58	1.5589

In contrast to the strongest performing embeddings, Word2Vec consistently yielded the weakest clustering results, particularly on the Silhouette Score and Calinski-Harabasz Index, indicating low intra-cluster cohesion and poor inter-cluster separation. However, selecting the most suitable embedding type among the other models was not straightforward. Since each of them tended to perform well in one specific metric while underperforming in others, although their overall results were relatively strong. Taking this into consideration, a balanced strategy was adopted in order to favor the overall clustering quality and robustness, allowing for minor tradeoffs when supported by improvements in other key metrics.

For Word2Vec, the 60 cluster configuration offered a reasonable compromise, with modest improvements in Calinski-Harabasz and Davies-Bouldin Indexes despite a slight drop in Silhouette Score performance. FastText and BERTimbau displayed strong metric alignment at 50 and 60 clusters, respectively, while DistilBERT performed best at 50 clusters, where compactness and separation peaked.

LaBSE, although achieved its highest Silhouette Score and lowest Davies-Bouldin Index at 50 clusters, was ultimately configured with 70 clusters to introduce greater granularity in the representation of the data. This adjustment, even though it was influenced by practical analytical goals, it remains statistically comprehensive. The 70 cluster setup still demonstrated strong clustering quality, with a Silhouette Score of 0.4772, a high Calinski-Harabasz Index of 123,690.20 and a low Davies-Bouldin Index of 1.1453. These metrics suggest that even with increased granularity, allowing finer distinctions between patterns of behavior, the clusters retained compactness and separation. This finer resolution is particularly advantageous in the context of anomaly detection, as it enables the model to capture more subtle deviations from typical patterns while maintaining structural coherence across clusters.

Ultimately, the configuration of 70 clusters using LaBSE embeddings was chosen as the optimal setup. This decision strikes the best balance between statistical performance, interpretability, and operational feasibility. It meets the objective of deriving an adequate representation of the data, one that reveals its fundamental structure and supports effective anomaly detection. The corresponding visualization of the optimal clustering for this embedding is shown in the following figure.

Figure 11 – LaBSE UMAP Projection with 70 GMM Clusters

To give an exhaustive comparison, visualizations for the other embedding types are provided in Appendix E, where each is marked with its optimal number of clusters, as underlined in Table 4. These supplementary visualizations reinforce the robustness of the selection and offer additional insight into the behavior of alternative embeddings without undermining the main analysis.

All UMAP visualizations utilized in this study were adjusted to preserve the local structure while increasing the spacing between clusters. For cluster separation projections, a color palette with 20 colours was used. However, due to the large number of clusters, some colors appear more than once. This repetition is a known limitation in discrete palettes for dense visualizations, being something that should be considered when interpreting the plots. While larger palettes exist, they often introduce many visually similar shades, which can reduce interpretability by making it harder to distinguish between clusters. Therefore, a smaller, clearer palette was selected to maintain visual clarity and avoid confusion.

4.2. Detection of Outcomes and Model Performance

To assess the performance of clustering and anomaly detection, both GMM and Bayesian Gaussian Mixture Models (BGMM) were executed on the LaBSE embeddings.

Regarding GMM with 70 clusters, to detect anomalous procurement records, the bottom 5% of data points that were flagged based on their likelihood under the GMM distribution presented 45,132 detected anomalies. Furthermore, clusters containing 3,243 data points or fewer, which correspond to the 35th percentile of cluster sizes were considered small clusters, existing 25 in total. This combination of likelihood-based and density-based approaches

provided a strong hybrid framework, enabling more meaningful and interpretable visualizations.

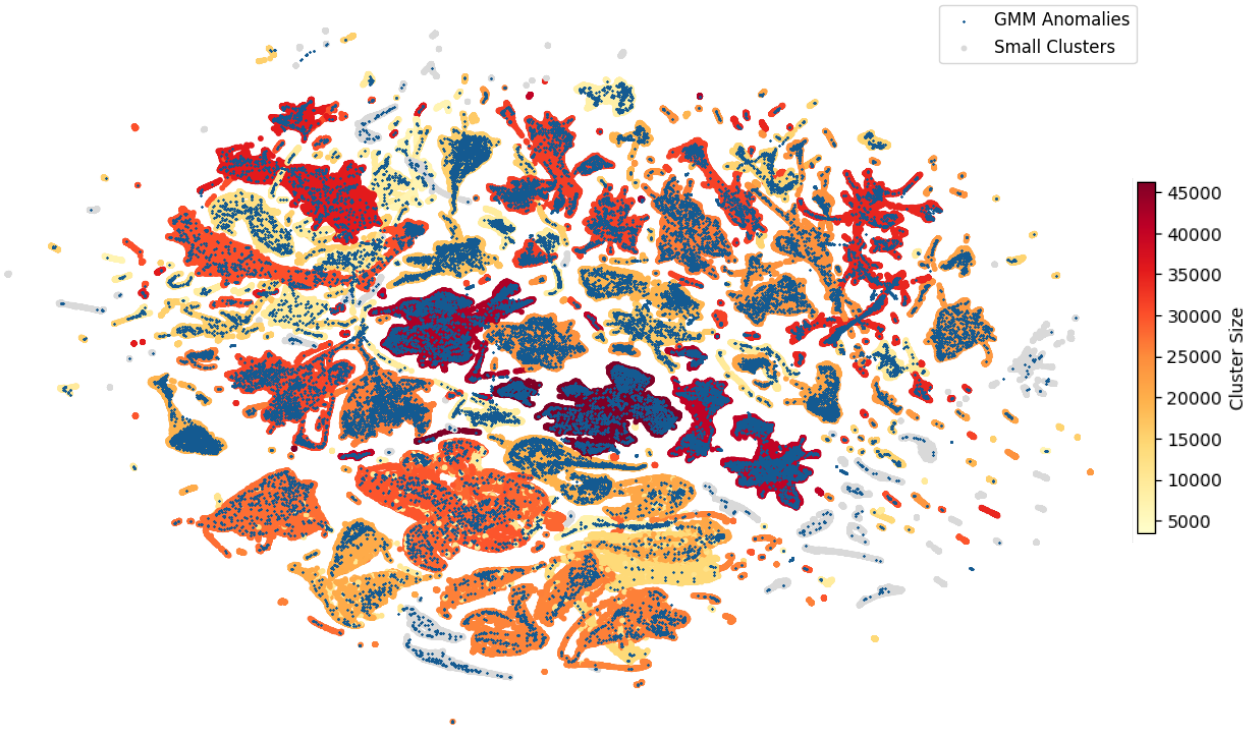


Figure 12 – UMAP Projection with 70 GMM Clusters and its Detected Anomalies

The previous figure visualizes the GMM clusters along with the detected anomalies, where it is noticeable that the anomalies detected by the model are overall well distributed around the feature space. In this type of visualization, an intensity color scale is used to represent cluster sizes that depend on the existent number of data points per cluster, where dark red indicates large clusters, orange corresponds to medium-sized clusters, and yellow represents smaller clusters. However, the clusters represented in yellow are not necessarily the smallest ones, since those are distinctly highlighted in light grey.

Although both BGMM and GMM detected the same number of anomalies, BGMM demonstrated significantly less consistent clustering performance. This is reflected in its evaluation metrics where even though Calinski-Harabasz Index remained reasonable at 117,388.33, the Silhouette Score was notably low at 0.1414, and the Davies-Bouldin Index was considerably high at 4.5994. These scores indicate that BGMM produced less compact and less well-separated clusters compared to GMM with 70 clusters.

A key issue with BGMM lies in its adaptive approach to estimate how many clusters it should create. In this case, it used the maximum allowed value of 150 clusters, which is more than double the number identified by GMM. This likely over-segmentation suggests that BGMM may have struggled to consolidate similar patterns, leading to fragmented and less interpretable cluster structures.

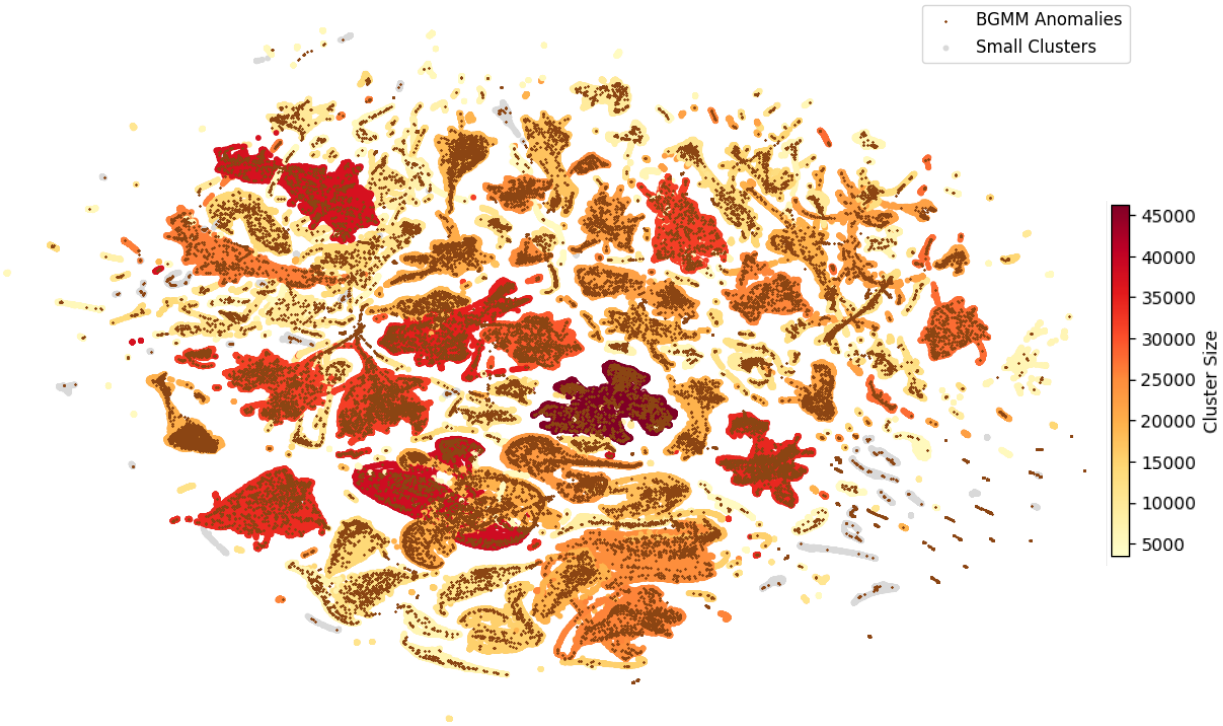


Figure 13 – UMAP Projection with 150 BGMM Clusters and its Detected Anomalies

As shown in the previous figure, BGMM produced a much larger number of lower density clusters compared to GMM.

Regarding the detection of small clusters, both GMM and BGMM applied the same threshold criterion. However, due to the significantly larger number of clusters generated by BGMM, this threshold corresponded to a lower absolute value of 1,075 data points or fewer, resulting in a total of 53 small clusters. In contrast, GMM, as previously mentioned, identified only 25 small clusters. This means that GMM's more conservative clustering approach yielded fewer small clusters, which were generally easier to interpret and less susceptible to overfitting on sparse or noisy data.

Taking this into consideration, while BGMM provides theoretical benefits such as automatic component selection and flexibility, the empirical results clearly favored GMM in terms of stability, clarity, and clustering quality. For this reason, GMM was the selected model.

Additionaly, in an effort to improve anomaly detection, GMM was run in parallel with DBSCAN on the LaBSE embeddings, where the anomalies identified uniquely by DBSCAN are shown in the following figure.

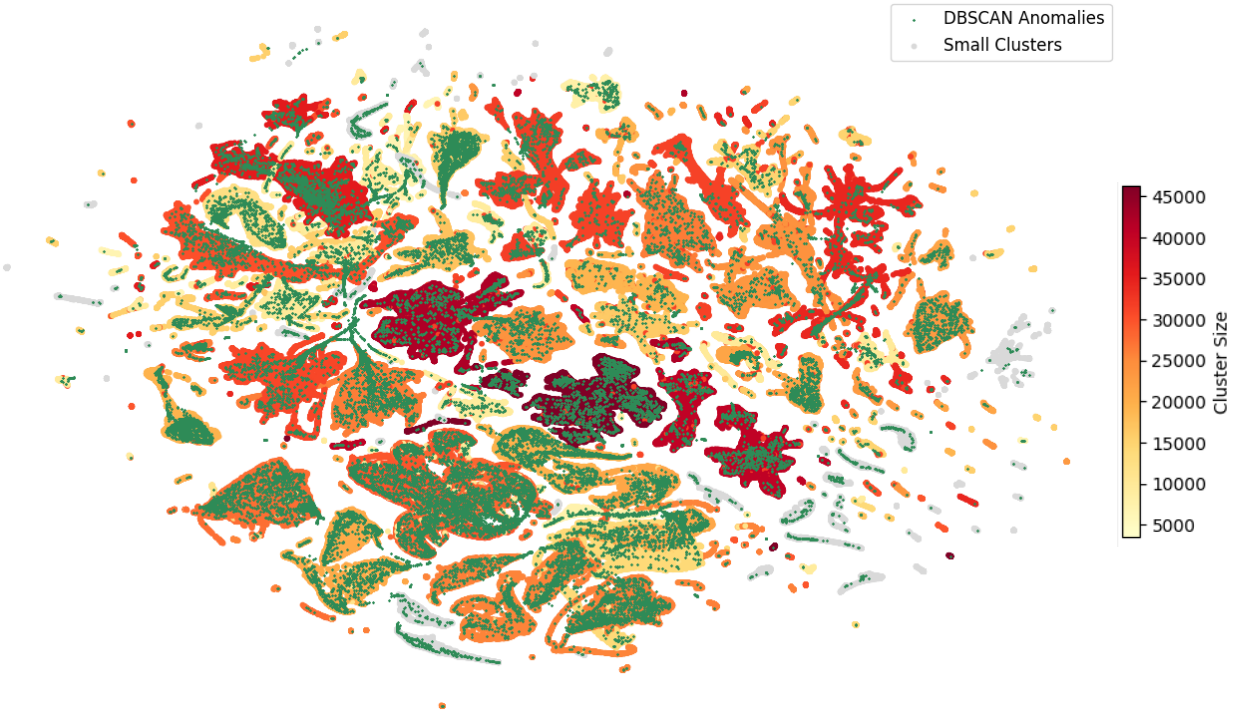


Figure 14 – UMAP Projection with 70 GMM Clusters and DBSCAN Detected Anomalies

While GMM had detected 45,132 anomalies on its own, it was found that DBSCAN independently identified 41,494 anomalies, being majority of them not flagged by GMM. When considering the total of anomalies detected by the intersection of both techniques a total of 19,005 data points are encountered, which for the data in consideration, this corresponds to an anomaly detection rate of 2.11%.

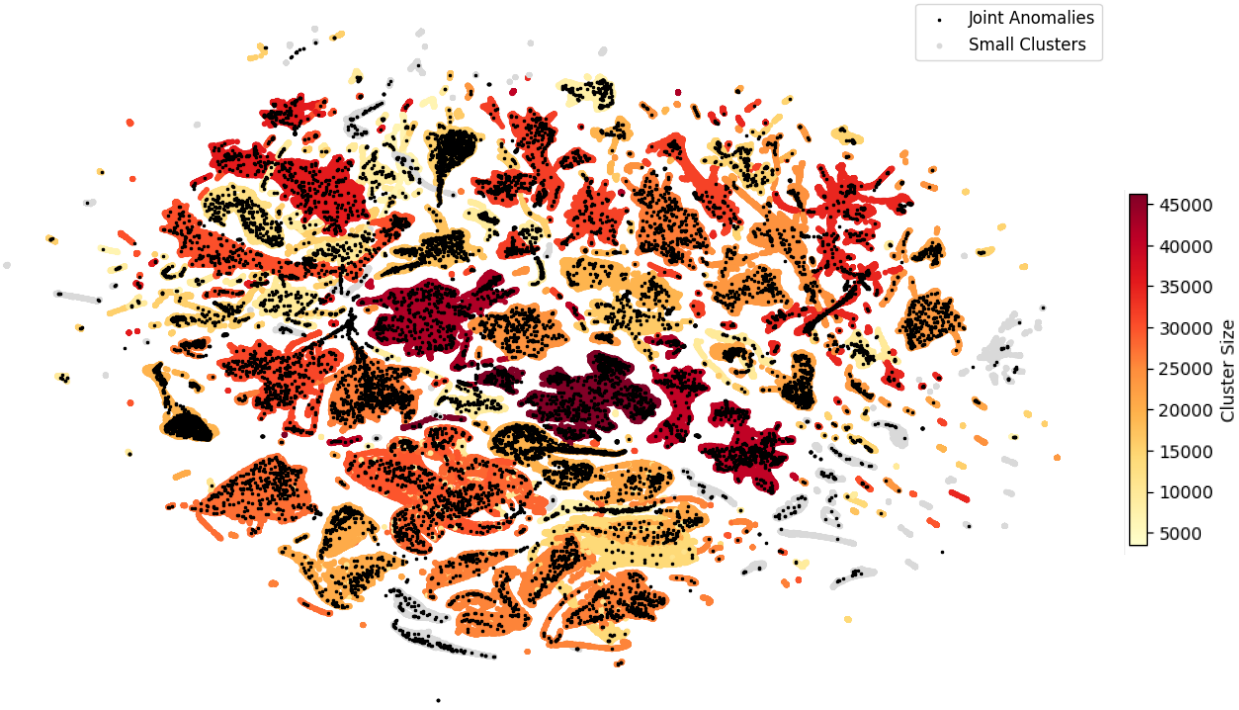


Figure 15 – UMAP Projection for the GMM and DBSCAN Anomaly Detection Model

By observing the previous figure, the visual output from DBSCAN, in conjunction with the GMM clusters and their respective anomalies, reveals a significant degree of overlap in the anomalies detected by both methods. This alignment underscores the strength of the hybrid detection strategy and demonstrates how combining density-based and probabilistic approaches can provide a more comprehensive identification of anomalous behavior in the data.

The techniques adopted for anomaly detection are further illustrated using additionally t-SNE visualizations, which are placed in Appendix F for reference. These visualizations help highlight subtle structure in local neighborhoods that may not be as apparent in UMAP projections.

4.3. VISUALIZING DATA IRREGULARITIES

From the following visualizations, it can be observed that the procurement dataset contains recurring and distinct patterns where some of them can potentially be classified as anomalies. The visual inspection serves as an important complementary step to the algorithmic anomaly detection, allowing to intuitively check of where and how these outliers manifest within the broader data landscape.

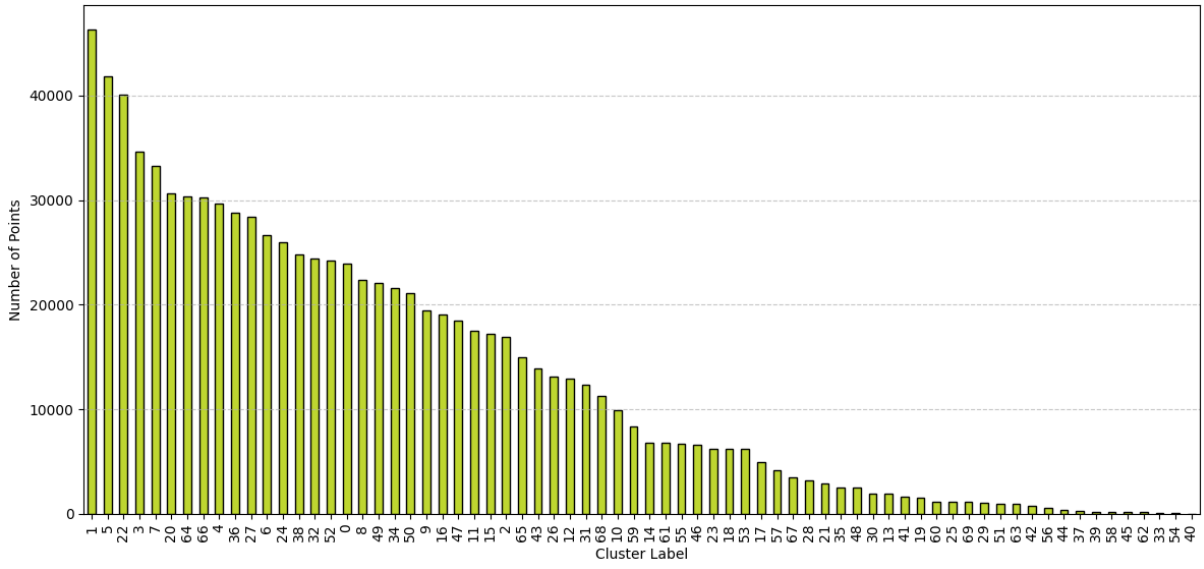


Figure 16 – Cluster Size Distribution

For instance, when analysing the distribution of data points across clusters in a descending hierarchy, it is noticeable that a few clusters contain a large share of the total data points, while many others have relatively less. However, the overall skewness of the data disribution does not present to be too extreme. The three largest clusters contain over 40,000 records each, marking a slightly higher volume from the rest. In contrast, a significant number of clusters have fewer than 10,000 data points.

This suggests substantial diversity in the dataset. While large clusters may reflect recurring patterns, smaller ones could correspond to less frequent or specialized contracts, not all of which are necessarily anomalous.

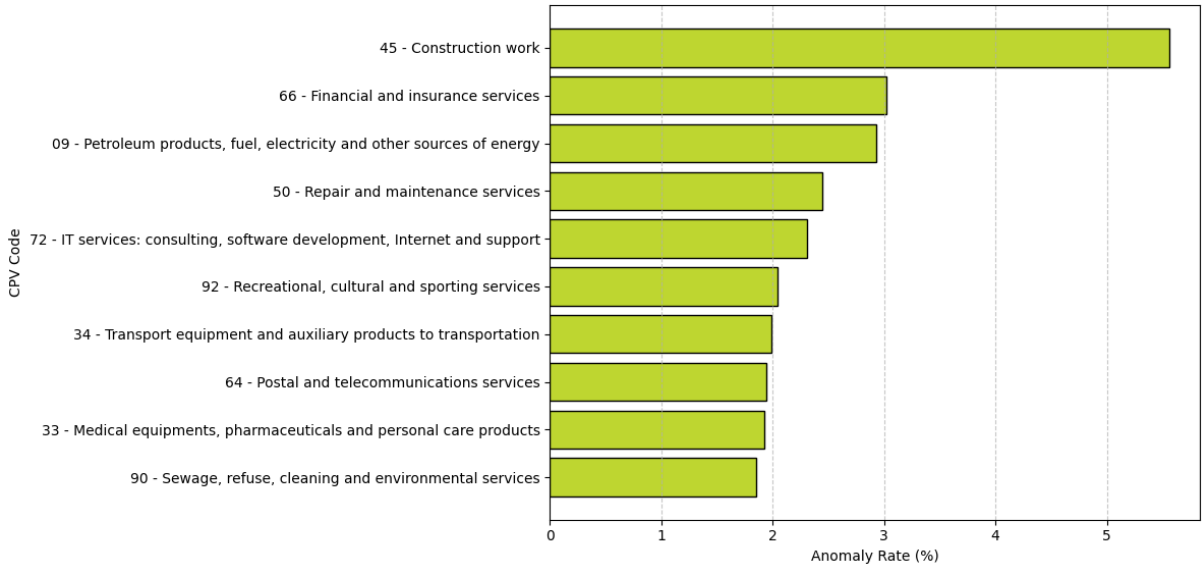


Figure 17 – Top 10 CPV Codes by Anomaly Rate

In terms of the CPV codes which are more frequently associated with anomalies code 45 - Construction work stands out by presenting an anomaly rate exceeding 5%. This is considerably higher when compared to other procurement categories. The construction sector often involves high value contracts and complex legal frameworks, which may increase the likelihood of inconsistencies or irregularities. Additionaly, the presence of various other CPV codes among the top ten indicates that anomalies are not confined to a single procurement type but are distributed across multiple sector divisions.

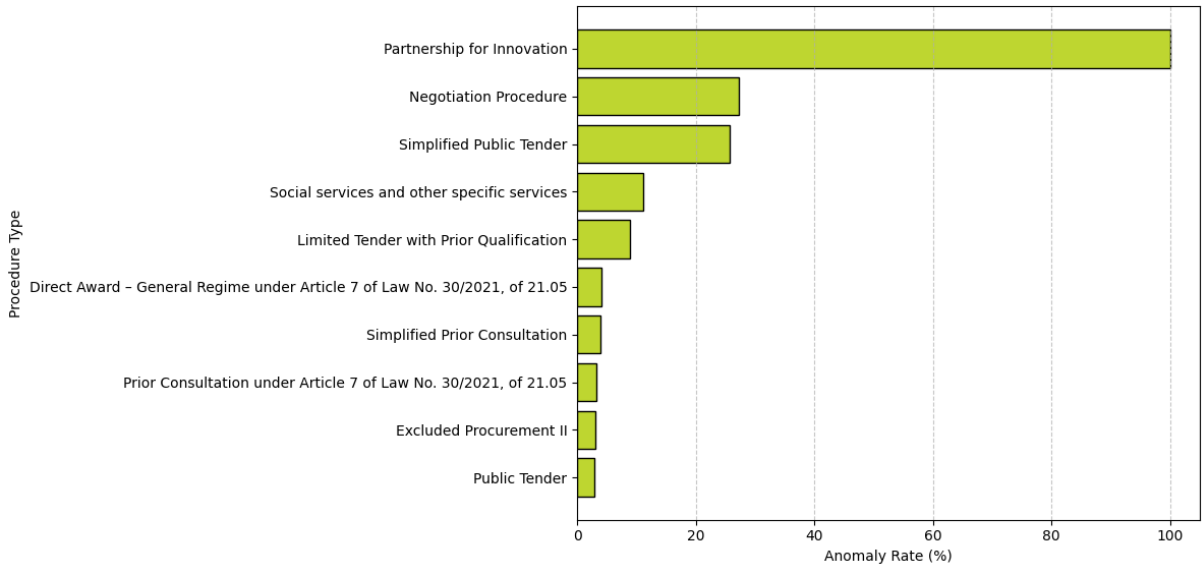


Figure 18 – Top 10 Procedure Types by Anomaly Rate

Regarding procedure types with the greatest rates of anomaly occurrence, it is noticeable that some categories exhibit particularly severe anomaly rates. Most notably, Partnership for Innovation, have 100% of the contracts were flagged as anomalous. This means every contract under this procedure type was flagged as an anomaly, likely due to the category containing only a single record. Such high anomaly rates can often arise when rare or specialized procedures differ significantly in structure, value, or participants.

Other procedure types that showed elevated anomaly rates include Negotiation Procedure and Simplified Public Tender, both presenting rates around 25%. These values suggest that a substantial portion of contracts in these categories exhibit characteristics that deviate from the typical data distribution, potentially reflecting irregular procurement behavior, or risk-prone conditions. Unlike Partnership for Innovation, these categories contain a higher volume of records, making their anomaly rates more statistically meaningful.

4.4. CLUSTER-LEVEL ANOMALY ANALYSIS

This section provides a deeper analysis of the results obtained from clustering and anomaly detection using GMM and DBSCAN. Special focus is placed on the relationship between cluster size and the concentration of anomalies, as this highlights some of the weaknesses and strengths of the use of the intersection of anomalies from both models.

Table 5 – Top 10 Clusters by Highest Anomaly Percentage

GMM Cluster	Total Points	Anomaly Percentage Combined
66	30,227	46.72
59	8,356	30.28
28	3,192	2.63
30	1,959	2.25
14	6,768	1.88
60	1,183	1.61
21	2,868	1.60
55	6,754	1.57
19	1,561	1.47
42	784	1.40

The table above contains the 10 GMM clusters with the highest anomaly percentages based on combined anomalies from both GMM and DBSCAN. Remarkably, Clusters 66 and 59 are notable for presenting a big discrupency in anomaly percentages when compared to the following clusters. Cluster 66 presents a 46.72% anomaly rate and also a high number of 30,227 data points, being followed by Cluster 59 with 30.28% anomalous occurennce, which seems to be a bit smaller.

Considering also their variable size, these clusters contribute disproportionately to the total number of anomalies and are priority candidates for further examination.

Table 6 – Anomaly Percentage of the 10 Smallest Clusters

GMM Cluster	Total Points	Anomaly Percentage Combined
40	1	0.00
54	96	0.00
33	109	0.00
62	144	0.00
45	150	0.00
58	156	0.00
39	187	0.00
37	262	0.00
44	403	0.00
56	584	0.68

To complement this analysis, the previous table provides an overview of the 10 smallest clusters in the dataset and their respective anomaly percentages. All of these clusters contain fewer than 600 records, with several below 200, making them statistically fragile for likelihood-based estimation through GMM alone.

Almost none of these small clusters were flagged to present anomalies in the combined detection approach because anomalies were only considered when both GMM and DBSCAN agreed to identify them. Since GMM did not flag these clusters, they were ultimately excluded. This illustrates that data points in sparse, low-density regions may be overlooked if they do not meet the requirements of GMM, even if DBSCAN alone might have flagged some of them as shown in Appendix G. This trade-off reflects a deliberate choice to favor confidence over coverage in the anomaly detection strategy.

Interestingly, it is only from the 10^{th} smallest cluster onward that any anomalies are detected, and even then, only at a moderate rate. Those clusters situate themselves close to the average of combined anomalies per cluster which corresponds to 1.57% of the total contracts. In total, it exists 11 clusters that do not present anomalies, meaning that other than 2 other clusters, majority of non-anomalous clusters have a lower density of data points. This supports the notion that low-density clusters are less likely to be flagged under the adopted criteria.

The fact that the the smallest cluster presents only 1 point gives a potential interest on its analysis, even if it doesn't present any anomaly on the combined approach used.

Table 7 – Anomaly Percentage of the 10 Biggest Clusters

GMM Cluster	Total Points	Anomaly Percentage Combined
1	46,291	0.25
5	41,861	0.13
22	40,055	0.29
3	34,628	0.30
7	33,272	0.12
20	30,672	0.26
64	30,311	0.07
66	30,227	46.72
4	29,706	0.20
36	28,771	0.39

In contrast, the previous table shows the 10 largest clusters in the dataset, each containing more than 28,770 points. The anomaly rates in these large clusters present usually a small variation, remaining between 0.07% and 0.39%, with the exception of Cluster 66 that has an anomaly rate of 46.72%. This means that the 8th biggest cluster is also the one that presents the highest anomaly percentage, as previously shown on Table 5.

It is noticed that irregularities may reside even within dominant segments of the data, reinforcing the necessity for their comprehensive coverage in anomaly detection.

4.5. Behavioral Deviations and Notable Patterns

The analysis of clusters and their associated anomalies was conducted with the assistance of a LLM after aligning the indices from the anomaly detection output with the corresponding preprocessed records. This step was essential because, after dimensionality reduction the embedded feature space does no longer contain a direct correspondence to the original input features. Thus, a purely quantitative analysis of determining which original features contributed to the anomaly becomes not possible.

Instead, an attributed-based analysis was employed to locate patterns that might explain what flagged the anomalies.

4.5.1. High Anomaly Clusters Analysis

Clusters with the highest anomaly percentages were given special attention, as they often reveal distinctive patterns that diverge significantly from the general behavior observed in the dataset.

These type of clusters can be described by:

- *1st Highest Anomaly Cluster*

This cluster is primarily characterized by the presence of Public Works Contracts as a contract type, which account for 10.99% of the normal contracts, but nearly double in the anomalous group at 21.46%.

Regarding procedure types, Direct Award - General Regime appears in 54.51%, of regular cases, while its presence in anomalous contracts is of 46.11%, suggesting that more competitive or complex procedures may dominate the flagged records. Additionally, 13.04% of these flagged contracts cite Article 24, No. 1, paragraph c) of the CPC as their legal justification, which is a less commonly used clause in the normal contracts.

CPV code distribution is mainly dominated by code 33 - Medical equipments, pharmaceuticals and personal care products and 45 - Construction work, accounting for 31.28% and 21.75% respectively of all anomalous entries, an unusually high share relative to other clusters.

In terms of execution timelines, the majority of contracts fall into short durations, with 52.3% of the normal group of contracts presenting this, compared to 45.9% in the atypical one. Although material and environmental criteria are largely absent, in both cases that isn't unsual. However, 44% of the anomalous contracts are marked as having a justification for not reducing the contract to writing, which when compared the number of times that happen in this cluster in general to all contracts of the dataset, it is noticed this only represents 3.35%.

Geographically, the execution location for most contracts, both anomalous and normal, is centered around Lisbon and Porto, which is not uncommon across the dataset, but the relative concentration may still suggest a regional pattern.

Notably, the average contract price among anomalous entries is 520,547.54€, which is much higher than the 40,934.37€ observed in non-anomalous contracts, a contrast that reinforces the idea that larger contracts are more prone to be flagged.

Key factors for anomaly detection in this cluster may be related to the high contract values, slight change in contract type, and use of less common legal justifications.

- *2nd Highest Anomaly Cluster*

This cluster is characterized by a high concentration of simplified procedures in public procurement. Among the normal contracts, the most common procedure type is Direct Award - General Regime, accounting for 75.58% of the cases, while in anomalous cases this value drops to 70.87%. The legal justification most frequently cited is Article 20, No. 1, paragraph d) of the CPC, appearing in 45.18% of the normal contracts and 46.72% on the detected anomalies.

The most common CPVs among the anomalous contracts include code 79 - Business services: law, marketing, consulting, recruitment, printing and security, which represents 13.68%, and 92 - Recreational, cultural and sporting services that stands for 11.74%. While the normal contracts show a similar distribution, with slight variations in the percentages.

Additionally, 76.21% of these normal contracts indicate no presence of material criteria, suggesting a lower reliance on formal evaluation standards, where that value rises in atypical contracts to 80.99%.

Regarding missing or incomplete data, both normal and anomalous contracts exhibit high proportions of fields filled with Not applicable, such as competing entities representing 64.95% in regular contracts and 63.88% for anomalous, with this cluster in general only representing 1.15% of the total dataset.

On average, normal contracts have lower values, with a mean contractual price of 20,836.96€ and mean effective total price of 19,464.55€. However, the average contract values of anomalies are significantly higher, being 40,319.11€ for the contractual price and 336,000.57€ for the total effective price. Showing a standard deviation in price much higher of 206,633.83€ for contract price and of 11,077,239.42€ for total price on anomalous cases, indicating that these contracts are less uniform and potentially riskier.

This shows that the main reason for the anomalous contracts being flagged in this cluster being are associated with price differences and not so much to textual features, since they present variations but not too significant.

4.5.2. Smallest Cluster Analysis

The smallest cluster was analyzed due to its markedly different density compared to the rest of the dataset, consisting of only a single contract. The purpose of this analysis is to determine whether the anomaly detection by DBSCAN alone is justified based on the contract's characteristics, or if it was merely flagged due to its extreme underrepresentation.

This contract refers to a Public Service Concession, which makes this contract only correspond to 0.08% of total contracts, which is not considerable. It is awarded through a Public Tender process and governed by the CPC (DL 111-B/2017), like big portion of procurement contracts in general. The legal justification cited is Article 20, No. 1, paragraph b) of the CPC and it is associated to the CPV code 60 - Transport services (excluding. Waste transport).

The contract duration is classified as very long, which is the least frequent duration of the dataset. Its effective total price is 9,349.20€, but presents a negative contractual price of -173,664€, which is highly irregular and potentially indicative of a data entry error or a complex concession model involving revenue sharing.

In this case, a negative contractual price has contributed to the contract being flagged as an anomaly by DBSCAN, showing its flagging was indeed plausible, although it was not jointly flagged by both models.

4.5.3. Large Clusters Analysis

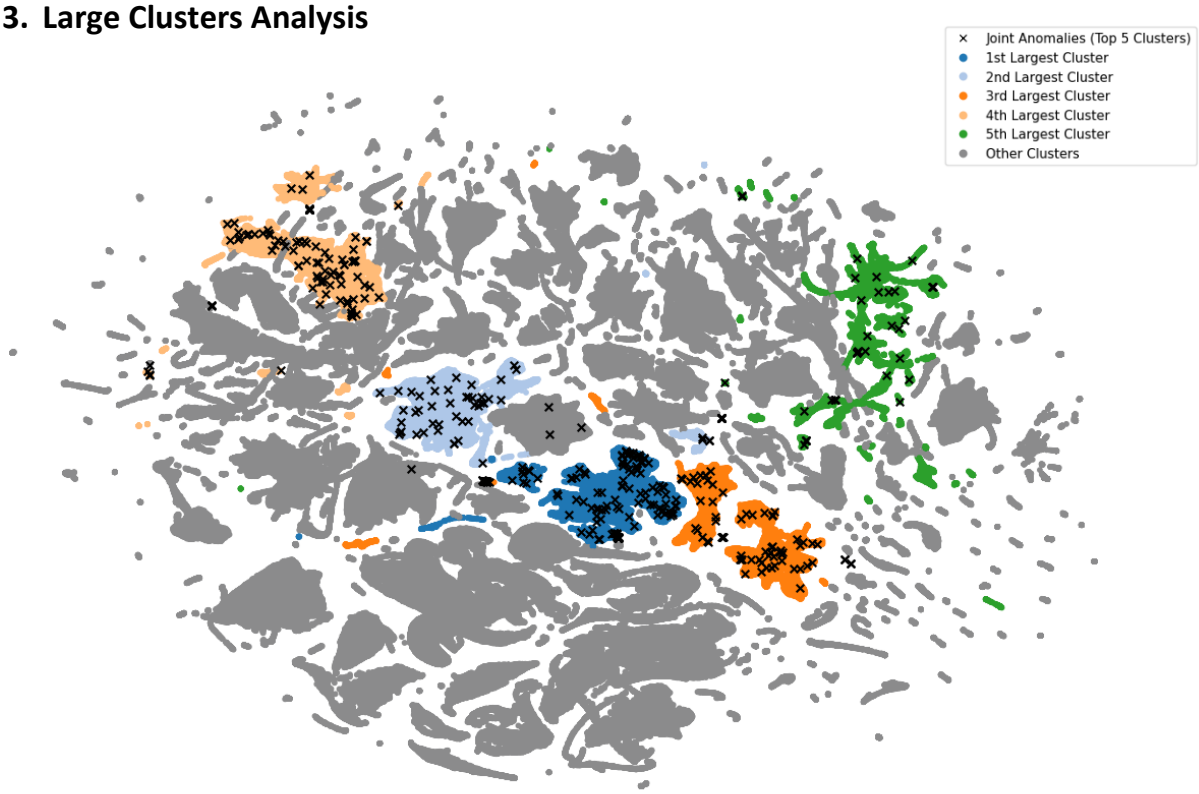


Figure 19 – Highlight of the 5 Largest Clusters Detected by the Anomaly Model

The previous figure presents the 5 largest clusters identified using GMM, with joint anomalies highlighted. This visualization emphasizes the distinct separation between clusters, offering a qualitative perspective on how the most populated segments of the dataset are distributed.

While certain anomalies may appear spatially separated in the UMAP projection, owing to its non-linear dimensionality reduction, they all are from the top five clusters in the original high-dimensional embedding space. These visualizations are intended to provide a human-interpretable representation, as 2 dimensions projections are easier to see and understand.

Analyzing these large clusters is particularly important because they represent the majority of the dataset, which makes their interpretation crucial for understanding the overall procurement behavior. Moreover, their anomaly rates are relatively low, which helps establish a baseline for what is typically considered a more “normal” anomaly percentage.

The descriptions of the largest clusters are the following:

- *1st Largest Cluster*

This cluster consists predominantly of public contracts related to the Acquisition of Services, with 98.87% of records falling under this contract type. It accounts for 13.42% of the total data within this feature category.

Among the normal contracts, the dominant procedure type is Direct Award - General Regime, which is present in 80.71% of cases. Notably, this cluster represents 8.16% of all contracts awarded under this procedure type, which constitutes a significant share when considering it is the most frequently used procedure in the dataset.

However, this procedure appears less often in anomalous contracts, where its presence drops to 73.5%. In contrast, Public Tenders, which are atypical for this cluster, appear in 16.24% of anomalous contracts. Additionally, Prior Consultation, the second most common procedure among normal contracts, accounts for 12.49%. These shifts suggest that anomalies may result from attempts to apply more formal procurement processes in contexts typically governed by more informal procedures.

The most frequently cited justification is Article 20, No. 1, paragraph d) of the CPC, appearing in 74.49% of the contracts, which indicates a relatively consistent legal basis across the cluster. However, its use is slightly less common among the detected anomalies, where it accounts for 63.25%, while Article 20, No. 1, paragraph a) of the CPC is cited in 13.68% of anomalous cases, being significantly more than in the normal group, suggesting a potential deviation in legal reasoning or justification patterns.

In terms of division classification, CPV code 79 - Business services: law, marketing, consulting, recruitment, printing and security is the most common and accounts for 20.05% of the normal contracts, while there is a slight increase for anomalous ones being represented with 28.21%. Execution timelines in the normal group are typically long, with 34.12% of contracts categorized under this duration range. However, the anomalous contracts present themselves as 31.62% being of very long duration.

Atypical points have an average contractual price of 236,887.29€, a significant contrast to the 25,535.78€ observed among normal contracts. This large difference, suggests the presence of contracts with extreme or inconsistent pricing.

In this case, anomalies in this cluster may be flagged due to inflated contract prices, longer duration execution periods, unusual legal justifications, and deviations from expected procurement procedures.

- *2nd Largest Cluster*

This cluster is also predominantly composed of Acquisition of Services contracts, with 98.77% of the normal contracts falling under this category, similarly, to the anomalous group that has 98.11%. Where this cluster corresponds to 12.12% of the total data points for this contract type.

The prevailing procedure type among the normal contracts is Prior Consultation, accounting for a significant 94.50%. This indicates a preference for relatively simplified procurement methods, being this procedure 20.86% of the total dataset and the text that most distincts this cluster from the others. However, for anomalous contracts this procedure type only appears 68.91% and additionally Excluded Procurement II appears also 15.09%, which usually doesn't appear in normal contracts.

In terms of legal justification this cluster relies heavily on the Article 20, No.1, paragraph c) of the CPC for normal contracts, while anomalies show a lower value of 52.83%. Additionally, the cluster presents a high concentration of CPV code 71 - Architectural, construction, engineering and inspection services that represent 19.29% and code 79 - Business services: law, marketing, consulting, recruitment, printing and security that has 18.54% of these contracts.

Regarding execution periods, 27.36% of normal contracts are categorized as having very long durations. This percentage rises to 41.51% among anomalous contracts, which is a notable increase, as very long durations are typically the least common in the dataset in general. This suggests that greater execution complexity or higher project risks may be contributing to the classification of anomalies.

In terms of geographic distribution, 12.79% of contracts list only Portugal as the execution location. However, since this characteristic appears in 7.10% of the overall dataset, it means that this cluster presents this execution location more than usual.

Financially, anomalies are characterized by a notably higher average contractual price of 278,435.06€ compared to just 35,432.27€ in the normal group. Despite this disparity, effective price remained uniformly low across both groups at 9,349.20€. This disparity in pricing could indicate either cost inflation, scope discrepancies, or documentation issues.

Anomalous records are likely flagged in this case due to rare procedure types, the increase on execution periods to be of very long durations, and excessive contract prices.

- *3rd Largest Cluster*

This cluster is primarily composed of contracts related to the Acquisition of Movable Goods. Within the normal group, 91.63% of contracts fall under this category, reflecting a typical trend in public procurement. The cluster itself represents 7.65% of all records associated with this contract type. However, only 75.42% of anomalous contracts are associated with this contract category, while an additional 11.86% are related to Public Works Contracts, a notable shift that indicates greater diversity in contract types among flagged records.

Majority of the regular contracts are predominantly processed using the Prior Consultation procedure, which accounts for 63.85%, indicating a reliance on simplified and less competitive tendering, while anomalous records rely 54.24% on Direct Award - General Regime.

In terms of justification, it is more varied for anomalous contracts, with a considerable proportion of 20.34% referring to Article 24, No. 1, paragraph e), sub-paragraph ii) of the CPC, which does not appear in the normal contracts. This may indicate a more complex or less commonly used justification for the procurement process. Additionally, a very small share of the normal contracts of 3.67% are marked of being relied with material criteria, while at 13.56% the irregular contracts show a higher percentage, suggesting that those are more subject to exceptions or special conditions.

This cluster has a high representation of the CPV code 33 - Medical equipments, pharmaceuticals and personal care products which represents 24.07% and code 15 - Food, beverages, tobacco and related products that belongs to 12.46% of all the contracts in this cluster.

Numerically, the average contractual price in normal contracts is relatively moderate at 49,000.08€, and the total effective price averages on 9,335.44€. However, anomalies are characterized by significantly higher values with their average for contract price being 1,075,857.55€, which is more than twenty times the normal average, and their average total effective price is also higher at 11,101.64€.

The contracts that were classified as anomalous are likely due to their high average contractual price and the diversity in contract types and procedures. The presence of less common legal justifications and a higher rate of material criteria being met also contributed to this classification. So far this was the cluster that was noticed higher percentage of variations in most of the fields.

- *4th Largest Cluster*

This cluster is primarily composed of contracts classified under the Acquisition of Movable Goods, representing 99.92% of the contracts within the cluster and accounting for 7.21% of this contract type in the overall dataset.

All contracts in this cluster are processed through framework agreements, predominantly under Article 259 of the CPC. Among anomalous contracts, 87.38% follow this procedure, while an additional being 12.62% associated with Article 258 of the CPC, which is a procedure not observed in normal contracts, suggesting exceptions or special cases in procurement handling.

The distinction is further reinforced by the legal justifications, which largely align with the cited articles and exhibit a clear procedural pattern. Additionally, 11.65% of anomalous contracts include explicit references to framework agreement numbers or

descriptions, an element completely absent in the normal group. This implies that anomalies may involve customized or uniquely defined agreements not typically present in standard records.

In terms of procurement content, 99.6% of the contracts in this cluster are assigned to the CPV code 33 - Medical equipments, pharmaceuticals and personal care products, showing a highly uniform, specialized segment of procurement, where the code in this cluster constitutes 12.08% of the all dataset.

Most noticeably, the average contract price of anomalous contracts reaches 1,089,774.39€, massively surpassing the 54,163.48€ mean of normal contracts, indicating potential irregularities, or at minimum, atypically large procurement operations. However, these contracts typically do not vary in effective price, with all contracts, both normal and anomalous, showing a fixed total effective value of 9,349.20€.

In this case, the combination of higher contract price values, slightly less common legal justifications and procedure types, and additional framework specification details likely contributed to these contracts being flagged as anomalies.

- *5th Largest Cluster*

This cluster is composed entirely of contracts categorized as Acquisition of Movable Goods, appearing in 100% of both normal and anomalous records. As such, the contract type is defining but not differentiating, having this cluster representing 6.90% of the overall dataset for this contract type.

Most contracts in the cluster follow the Direct Award - General Regime procedure type, with 99.65% among normal contracts and 97.44% among anomalies, which corresponds to 7.24% of the total data for this type. However, divergence emerges in the legal justifications cited. Anomalous contracts more frequently use Article 24, No. 1, paragraph e), sub-paragraph iii) of the CPC corresponding to 28.21%, compared to 13.2% in normal contracts, signaling more specific or less common justifications.

A slight variation is also observed in the application of material criteria. While 92.51% of normal contracts apply these criteria, the rate drops to 84.62% in anomalous ones. Typically, material criteria are not applied as frequently in public procurement, but this cluster represents 15.07% of all contracts in the dataset that utilize such feature highlighting that this is as an intrinsic characteristic of this group.

Differences also appear in justifications for not reducing contracts to writing. Among anomalies, 43.59% cite Article 95, No. 1, paragraph c), compared to just 18.36% in normal contracts. This signals the use of less conventional documentation practices in anomalous entries.

In terms of award entities, anomalous contracts show greater concentration. For instance, institutions such as Hospital de Braga and ARTUR SALGADO, S.A. each account for 10.26% of the anomalous cases, suggesting a recurrent pattern among a limited number of contracting bodies.

The cluster is consistent in its use of CPV code 33 - Medical equipment, pharmaceuticals, and personal care products, which appears in 90.81% of the contracts.

Geographically, anomalous contracts show higher concentrations in Porto and Braga, suggesting a potential regional procurement pattern that may differ from broader distributions observed in other clusters.

Most notably, anomalous contracts have an average contractual price of 209,856.09€, which is drastically higher than the 6,953.17€ average for normal contracts. Despite this, the total effective price remains fixed across both groups.

Anomalies in this cluster are mainly driven by significantly higher contractual prices, increased use of uncommon legal justifications, and more frequent references to Article 95 to justify not reducing contracts to writing. There is also a noticeable concentration of awarding entities, along with a regional focus in Porto and Braga.

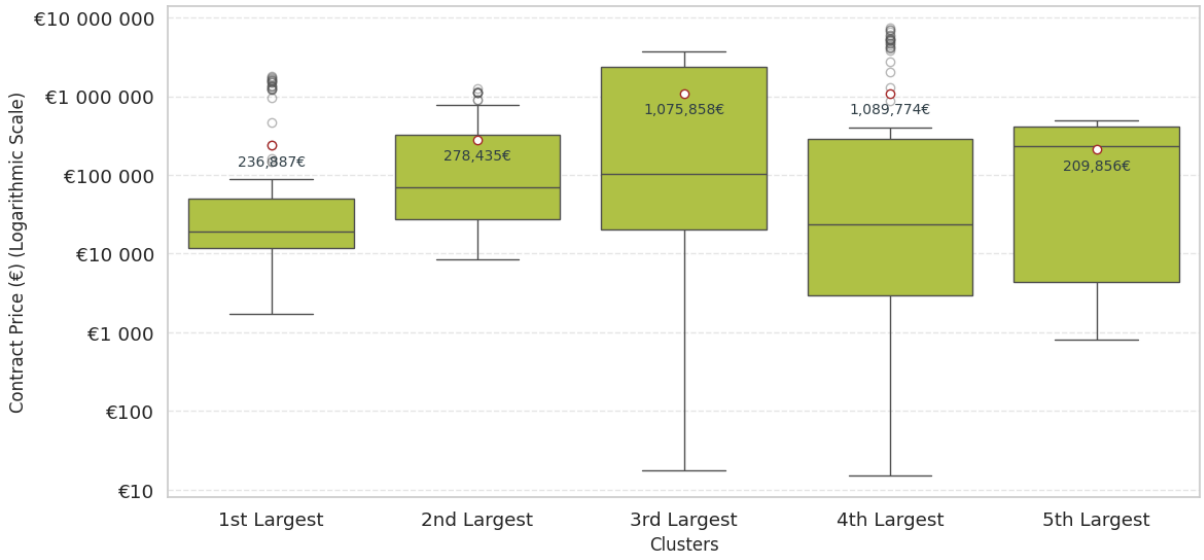


Figure 20 – Contract Price Variations for Anomalous Contracts in the 5 Largest Clusters

Since high deviations in contract prices are a common occurrence across all clusters, the boxplots above aim to clearly illustrate the distribution of contract prices for anomalous contracts within the 5 largest GMM clusters.

Each box represents the interquartile range, with the median shown as a horizontal line and the average indicated by a brown dot, annotated with its exact value below. A logarithmic

scale is used to accommodate the presence of extremely high contract values while preserving the readability of the full distribution.

The plot reveals that the 3rd and 4th largest clusters contain the most financially significant anomalies, both in terms of their high average contract prices and the broad spread of values in general. In contrast, the 1st , 2nd, and 5th clusters, while still anomalous, show less extreme price dispersion.

Here, the large gap between the average and median in several clusters highlights the impact of high value outliers, which are a key factor in identifying the distribution of financial anomalies. By contrast, the 5th largest cluster shows a closer alignment between the mean and median, indicating a more concentrated distribution of anomalous prices. However, despite this tighter spread, its overall price level is still not the lowest. When combined with the previously discussed characteristics of this cluster relative to normal contracts, this emphasizes that even less dispersed anomaly contracts can still involve substantially high value contracts.

4.5.4. Remarks on Detected Anomalies

All clusters analysed that present anomalies show a consistent pattern among their anomalous contracts, where regardless of sector, anomalies consistently exhibit extreme inflation in average contractual prices.

Although some textual deviations may appear more subtle, anomalous contracts often diverge from commonly used legal justifications, such as Article 20, No. 1, paragraph d) of the CPC, by instead citing less typical or more obscure clauses. While normal contract clusters generally rely on simplified procurement procedures (e.g., Prior Consultation or Direct Award - General Regime), anomalous contracts are more likely to involve formal or less commonly used procedures, potentially indicating attempts to introduce additional complexity or reduce scrutiny.

Textual deviations were observed to be more pronounced in the largest clusters compared to the highest anomaly ones, showing that larger data volumes allow for clearer comparison between typical and atypical patterns.

In general, the clusters analysed exhibit a higher CPV concentration on code 33 - Medical equipments, pharmaceuticals and personal care products and code 79 - Business services: law, marketing, consulting, recruitment, printing.

Contracts often present missing data for contract closure date and effective price having all of these values being around 90% or higher. Additionally, reasons for price or deadline changes also present a share of 80% or higher for this to happen. Shared traits like non-weekend contract activity, Lisbon focused contract execution (except for the 5th largest cluster), and

fixed effective prices appear often across clusters and are not useful for distinguishing anomalies.

The absence of this key data, however, is not necessarily concerning, as many contracts in the dataset are still ongoing or recently signed and have not yet reached their closing stage. However, using non-finalized contracts is important because it reflects the most current procurement activity, ensuring the models are applied to real-time data where potential risks can be detected early. Waiting only for finalized contracts would limit the ability to identify anomalies proactively, potentially delaying oversight or corrective measures.

In summary, anomalies across all clusters share deviations that together form a recognizable risk pattern. Although steps were taken to minimize overemphasis on numerical features, contract price remains the dominant factor driving anomaly detection.

Although supervised methods might offer clearer interpretability, the absence of labels limits their application. Therefore, anomaly detection should be seen as a diagnostic tool, valuable for flagging cases for further review, but not as definitive evidence of irregular behavior.

Crucially, being flagged as an anomaly does not necessarily imply wrongdoing or misconduct. These records simply differ statistically from others based on the chosen features and modeling approach. Further contextual and domain-specific analysis is essential to assess whether an anomaly reflects a data quality issue, an unusual but valid case, or a potential ilegallity.

5. CONCLUSIONS AND FUTURE WORKS

5.1. CONCLUSIONS

This thesis investigates the application of embedding-based ML models in detecting anomalies in the public procurement contracts of Portugal. It integrated previously preprocessed data, combining dense vector representations of text features derived from embedding techniques with numerical variables. The research used GMM and DBSCAN within the framework of unsupervised learning to detect anomalous contracts in the procurement processes.

The results demonstrate that embedding-based models uncover nuanced anomalies that traditional numerical methods may overlook. Procurement contract descriptions often contain subtle semantic patterns that text embeddings can effectively capture. When combined with structured numerical data, these embeddings enable the detection of anomalies that reflect both linguistic and quantitative deviations. Clustering algorithms revealed meaningful partitions in the data and exposed different types of anomalies across both high anomaly and large clusters.

From an analytical perspective, this study demonstrates that embedding-based representations, along with clustering techniques improves the identification of potential procurement irregularities. While unsupervised models lack definitive labels for validation, model reliability was assessed through Silhouette Scores and other cluster validity metrics. Moreover, methods like UMAP or t-SNE enable some degree of visual inspection, which verifies the interpretability and internal consistency of the clusters. Together, this combination of assessments increases the confidence in the anomalies detected.

Additionally, it was observed that GMM and DBSCAN detect anomalies through different mechanisms, and even when only one of the models flags a contract as anomalous, it does not necessarily mean the contract lacks anomalous characteristics. Each model captures distinct aspects of the data, highlighting complementary anomaly pattern.

A in-depth analysis considering different types of clusters revealed subtle deviations by analyzing feature comparisons to more typical clusters. Within these clusters, the model identifies anomalies such as excessively high contract price averages and non-standardized textual descriptions relatively to the rest of the cluster.

5.2. WORK CONTRIBUTIONS

Meaningful contributions arise in both academic knowledge and practical applications. Academically, it adds to the growing body of literature on NLP and ML applications in public policy and governance by demonstrating the feasibility of integrating NLP techniques with unsupervised learning for anomaly detection in real-world legal and financial data. The methodological pipeline developed can be adaptable across various domains and geographies, enhancing its potential as a foundation for further scholarly exploration. Beyond

technical insights, this work fosters awareness and expands understanding of procurement irregularities. Where each analytical effort, even when constrained by limited irregularity reporting, helps highlight systemic vulnerabilities within public procurement.

Practically, this work contributes to enhancing transparency and efficiency in public procurement monitoring by demonstrating how automated anomaly detection can be applied to large scale contract datasets. The developed methodology enables systematic identification of contracts that deviate from normative patterns. This process can assist oversight bodies by flagging contracts that warrant closer human inspection, especially when manual review of the entire dataset is infeasible. In this sense, the research takes a concrete step toward more data-driven oversight mechanisms. Furthermore, the insights gained from clustering analysis provide valuable information about recurring irregularities, which can inform regulatory adjustments and supporting efforts to standardize procurement practices.

5.3. LIMITATIONS

Certain limitations arise primarily from a lack of computational resources. Processing textual data into embeddings, as well as working with high-dimensional representations, requires considerable memory, processing power, and storage. In some cases, the available hardware was limited. Such constraints necessitated the purchase of a Google Colab Pro subscription to access more powerful computing resources, particularly for memory intensive tasks and faster processing. Running times were improved significantly by the use of GPUs when available and storage requirements were met through the purchase of a Google Drive subscription.

These limitations affected scaling during experimentation and extensive hyperparameter optimization, being large scale tests not performed. To determine the optimal approach, multiple trials were performed on a small sample size of the final dataset size. In any case, these limitations are overcome by the fact that the work illustrates the feasibility and potential of embedding-based anomaly detection in public procurement data.

5.4. FUTURE WORKS

There are many directions that could expand the scope of this work. One important path is to apply an embedding-only approach, excluding numerical features, to assess whether deviations in textual features become more pronounced when not overshadowed by strong numerical signals such as contract price. Although price features constitute an important feature of public contracts this could provide deeper insights into the semantic characteristics of procurement anomalies, even if this means analysing text and numerical features separatly.

Future efforts might also look into applying the same methodological framework to procurement datasets from other countries, which would allow for cross-national comparisons and explore more systemic issues or best practices.

Another valuable direction involves incorporating domain specialists, such as auditors, legal experts, and procurement officers, into the evaluation process. Their feedback would enhance the model's real-world relevance and improve the accuracy and interpretability of the flagged anomalies. Building upon these findings, future work could also center on creating tools that enable real-time or periodic monitoring of public procurement data through an anomaly detection pipeline. This would require embedding the detection models into lenses where users from oversight institutions could review, contextualize, and respond to flagged anomalies. Such development would require not only improved interface design but also closer alignment with institutional workflows and compliance criteria, which could advance the proactively transparent public procurement systems.

An innovative and relatively unexplored direction for future work could additionally be the use of multi-agent LLM systems, where different AI agents are assigned specific expert roles, such as auditor, legal analyst, or procurement specialist, in order to collaboratively analyze and explain flagged anomalies. This approach could serve as a viable alternative if the development of a fully integrated procurement interface proves impractical in real-world applications. Each agent would function by being attributed a specific role conditioned instance of a large language model, either prompted or fine-tuned to mimic the reasoning style, vocabulary, and priorities of its respective domain. Rather than relying on a single model to provide an explanation, these agents would contribute distinct perspectives and engage in structured multi-turn dialogues, such as deliberation protocols or guided critiques, to reveal conflicting interpretations, regulatory ambiguities, or alternative assessments of risk. This simulated multi-expert discussion would resemble the interdisciplinary review processes found in real-world oversight bodies and could generate richer, more transparent anomaly explanations. Such a system would enhance not only the interpretability of AI-driven decisions but also serve as a valuable testbed for collaborative decision-making between humans and AI. To function effectively, the interactions between AI agents could be managed through a coordination layer that handles turn-taking, dialogue structure, and the sharing of evidence. This setup could also include access to external knowledge sources, such as legal corpora or regulatory databases, to support more informed reasoning. By integrating role-based agent reasoning into anomaly detection pipelines, this approach could significantly increase both the credibility and institutional trust in automated systems used for public procurement oversight.

BIBLIOGRAPHICAL REFERENCES

Adhikari, S., & Dhakal, B. (2023). Revolutionizing Natural Language Processing with GPT-based Chatbots: A Review. *Technical Journal*, *3*(1), 109–120. https://doi.org/10.3126/tj.v3i1.61943

Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, *55*, 278–288. https://doi.org/10.1016/j.future.2015.01.001

Andhov, M., Caranta, R., Stoffel, T., Grandia, J., Janssen, W. A., Vornicu, R., Czarnezki, J. J., Gromnica, A., Tallbo, K., Martin-Ortega, O., Mélon, L., Edman, Å., Göthberg, P., Nohrstedt, P., & Wiesbrock, A. (2020). Sustainability Through Public Procurement: The Way Forward – Reform Proposals. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3559393

Andvig, J. Christopher., & Fjeldstad, O.-Helge. (2001). *Corruption : a review of contemporary research*. Chr. Michelsen Institute, Development Studies and Human Rights. http://hdl.handle.net/11250/2435853

Arlbjørn, J. S., & Freytag, P. V. (2012). Public procurement vs private purchasing: Is there any foundation for comparing and learning across the sectors? *International Journal of Public Sector Management*, *25*(3), 203–220. https://doi.org/10.1108/09513551211226539

Autoridade da Concorrência. (2022). *AdC sanciona sete empresas por participação em cartel em concursos públicos no setor da vigilância e segurança*. https://www.concorrencia.pt/pt/artigos/adc-sanciona-sete-empresas-por-participacao-em-cartel-em-concursos-publicos-no-setor-da

Autoridade da Concorrência. (2024). *Combate ao conluio na contratação pública*. https://www.concorrencia.pt/pt/combate-ao-conluio-na-contratacao-publica

Axelsson, B., & Torvatn, T. (2017). Public Purchasing in an Interactive World. In *No Business is an Island: Making Sense of the Interactive Business World* (pp. 173–194). Emerald Group Publishing Ltd. https://doi.org/10.1108/978-1-78714-549-820171010

Azmi, K. S. A. , & Rahman, A. A. L. A. (2015). E-Procurement: A Tool to Mitigate Public Procurement Fraud in Malaysia? In Dr Carl Adams (Ed.), *Proceedings of The 15th European Conference on eGovernment* (pp. 361–368). Academic Conferences International Limited.

Baltrunaite, A., Giorgiantonio, C., Mocetti, S., & Orlando, T. (2018). Discretion and Supplier Selection in Public Procurement. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3210748

Caldwell, N., Walker, H., Harland, C., Knight, L., Zheng, J., & Wakeley, T. (2005). Promoting competitive markets: The role of public procurement. *Journal of Purchasing and Supply Management*, *11*(5–6), 242–251. https://doi.org/10.1016/j.pursup.2005.12.002

Carbone, C., Calderoni, F., & Jofre, M. (2024). Bid-rigging in public procurement: cartel strategies and bidding patterns. *Crime, Law and Social Change*, *82*(2), 249–281. https://doi.org/10.1007/s10611-024-10142-0

Carneiro, D., Veloso, P., Ventura, A., Palumbo, G., & Costa, J. (2020). Network Analysis for Fraud Detection in Portuguese Public Procurement. In C. , Analide, P. , Novais, D. , Camacho, & H. Yin (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2020: Vol. 12490 LNCS* (pp. 390–401). Springer, Cham. https://doi.org/10.1007/978-3-030-62365-4_37

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Reference Format*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chen, F., van Dalen, J., & Wynstra, F. (2024). The grey side of procurement: Measuring the prevalence of questionable purchasing practices. *Journal of Purchasing and Supply Management*, *30*(3), 100922. https://doi.org/10.1016/j.pursup.2024.100922

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to Train Good Word Embeddings for Biomedical NLP. *BioNLP 2016 - Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 166–174. https://doi.org/10.18653/V1/W16-2922

Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162. https://doi.org/10.1017/S1351324916000334

Costa, A. A., Arantes, A., & Valadares Tavares, L. (2013). Evidence of the impacts of public e-procurement: The Portuguese experience. *Journal of Purchasing and Supply Management*, *19*(4), 238–246. https://doi.org/10.1016/j.pursup.2013.07.004

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

Dikmen, S., & Çiçek, H. G. (2023). Fighting Against Corruption and Bribery in Public Procurements During the Covid-19 Pandemic. In R. W. , McGee & S. Benk (Eds.), *The*

Ethics of Bribery: Theoretical and Empirical Studies (pp. 309–328). Springer, Cham. https://doi.org/10.1007/978-3-031-17707-1_18

Domberger, S., & Jensen, P. (1997). Contracting out by the public sector: theory, evidence, prospects. *Oxford Review of Economic Policy*, *13*(4), 67–78. https://doi.org/10.1093/oxrep/13.4.67

Erridge, A. (2007). Public Procurement, Public Value and The Northern Ireland Unemployment Pilot Project. *Public Administration*, *85*(4), 1023–1043. https://doi.org/10.1111/j.1467-9299.2007.00674.x

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD-96*. https://cdn.aaai.org/KDD/1996/KDD96-037.pdf

European Commission. (2017). *European Semester Thematic Factsheet: Fight Against Corruption*. https://commission.europa.eu/system/files/2018-06/european-semester_thematic-factsheet_fight-against-corruption_en_0.pdf

European Commission. (2021). *2021 Rule of Law Report: Country Chapter on the rule of law situation in Portugal*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021SC0723

European Parliament. (2024). *Public procurement contracts*. https://www.europarl.europa.eu/erpl-app-public/factsheets/pdf/en/FTU_2.1.10.pdf

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

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

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT Sentence Embedding. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

Ferraz, C., Finan, F., & Szerman, D. (2015). *Procuring Firm Growth: The Effects of Government Purchases on Firm Dynamics*. https://doi.org/10.3386/w21219

Ferreira Gomes, A., & Rodrigues, A. S. (2014). Enhancing efficiency in public procurement in Portugal : an overview of the relevant competition issues. *Revista de Concorrência e Regulação, Ano V, Número 19*, 181–212.

https://www.concorrencia.pt/sites/default/files/imported-magazines/CR19_-_Antonio_Ferreira_Gomes_-_Ana_Sofia_Rodrigues.pdf

G. L. Albano, P. Buccirossi, G. Spagnolo, & M. Zanza. (2006). Preventing Collusion in Procurement: A Primer. In Nicola Dimitri, Gustavo Piga, & Giancarlo Spagnolo (Eds.), *Handbook of Procurement*. Cambridge University Press. https://doi.org/10.1017/CBO9780511492556

García Rodríguez, M. J., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E. D., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, *133*, 104047. https://doi.org/10.1016/j.autcon.2021.104047

Gelderman, K., Ghijsen, P., & Schoonen, J. (2010). Explaining Non-Compliance with European Union Procurement Directives: A Multidisciplinary Perspective. *JCMS: Journal of Common Market Studies*, *48*(2), 243–264. https://doi.org/10.1111/j.1468-5965.2009.02051.x

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6. *COLING '96: Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, 466–471. https://doi.org/10.3115/992628.992709

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *11*(1), 1–21. https://doi.org/10.1080/00401706.1969.10490657

Harink, J. H. A. (1999). *Excelling with e-procurement: The Electronic Highway to Competitive Advantage* . Alphen aan den Rijin, Holland: Samson.

Herweg, F., & Schmidt, K. M. (2020). Procurement with unforeseen contingencies. *Management Science*, *66*(5), 2194–2212. https://doi.org/10.1287/mnsc.2019.3290

Hoekman, B., & Onur Taş, B. K. (2024). Discretion and public procurement outcomes in Europe*. *European Journal of Political Economy*, *82*, 102525. https://doi.org/10.1016/j.ejpoleco.2024.102525

Hott, H. R., Silva, M. O., Oliveira, G. P., Brandão, M. A., Lacerda, A., & Pappa, G. (2023). Evaluating Contextualized Embeddings for Topic Modeling in Public Bidding Domain. In M. C. , Naldi & R. A. C. Bianchi (Eds.), *Intelligent Systems. BRACIS 2023.: Vol. 14197 LNAI* (pp. 410–426). Springer, Cham. https://doi.org/10.1007/978-3-031-45392-2_27

Hu, R., Aggarwal, C. C., Ma, S., & Huai, J. (2016). An Embedding Approach to Anomaly Detection. *2016 IEEE 32nd International Conference on Data Engineering, ICDE 2016*, 385–396. https://doi.org/10.1109/ICDE.2016.7498256

Hyari, K. H., & Alamayreh, T. (2023). Unbalanced bidding in construction projects: a contractors' perspective. *International Journal of Construction Management*, *23*(12), 2058–2066. https://doi.org/10.1080/15623599.2022.2035498

Ibbs, C. W., Wong, C. K., & Kwak, Y. H. (2001). Project Change Management System. *Journal of Management in Engineering*, *17*(3), 159–165. https://doi.org/10.1061/(ASCE)0742-597X(2001)17:3(159)

Jin, H., Raghavan, K., Papadimitriou, G., Wang, C., Mandal, A., Deelman, E., & Balaprakash, P. (2023). *Self-supervised Learning for Anomaly Detection in Computational Workflows*. https://arxiv.org/pdf/2310.01247

Johnson, J. M., & Khoshgoftaar, T. M. (2021). Medical Provider Embeddings for Healthcare Fraud Detection. *SN Computer Science*, *2*(4), 1–15. https://doi.org/10.1007/s42979-021-00656-y

Kadappa, V., & Negi, A. (2016). A theoretical investigation of feature partitioning principal component analysis methods. *Pattern Analysis and Applications*, *19*(1), 79–91. https://doi.org/10.1007/s10044-014-0390-x

Kappal, S. (2019). Data Normalization using Median & Median Absolute Deviation (MMAD) based Z-Score for Robust Predictions vs. Min-Max Normalization. *London Journals Press*, *19*(4). https://doi.org/10.13140/RG.2.2.32799.82088

Kei Kawai, Jun Nakabayashi, & Daichi Shimamoto. (2022). *A Study of Bid-rigging in Procurement Auctions: Evidence from Indonesia, Georgia, Mongolia, Malta, and state of California* (w30271; NBER Working Paper Series). http://www.nber.org/papers/w30271

Kistler, J. T., Sharma, L., Jayaram, J., & Eckerd, S. (2024). Does history really repeat itself? An empirical investigation of recurring misconduct violations in public procurement. *Journal of Purchasing and Supply Management*, *30*(1), 100893. https://doi.org/10.1016/j.pursup.2023.100893

Loosemore, M. (2016). Social procurement in UK construction projects. *International Journal of Project Management*, *34*(2), 133–144. https://doi.org/10.1016/j.ijproman.2015.10.005

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*. https://doi.org/10.1007/s41109-022-00523-6

Macário, R., Ribeiro, J., & Costa, J. D. (2015). Understanding pitfalls in the application of PPPs in transport infrastructure in Portugal. *Transport Policy*, *41*, 90–99. https://doi.org/10.1016/j.tranpol.2015.03.013

Martin, S., Hartley, K., & Cox, A. (1999). Public Procurement Directives in the European Union: A Study of Local Authority Purchasing. *Public Administration*, *77*(2), 387–406. https://doi.org/10.1111/1467-9299.00159

Marx, V. (2024). Seeing data as t-SNE and UMAP do. *Nature Methods*, *21*(6), 930–933. https://doi.org/10.1038/s41592-024-02301-x

Mateus, R., Ferreira, J. A., & Carreira, J. (2010). Full disclosure of tender evaluation models: Background and application in Portuguese public procurement. *Journal of Purchasing and Supply Management*, *16*(3), 206–215. https://doi.org/10.1016/j.pursup.2010.04.001

Matthew, K., Patrick, K., & Denise, K. (2013). The effects of fraudulent procurement practices on public procurement performance. *International Journal of Business and Behavioral Sciences*, *3*(1). https://www.researchgate.net/profile/Patrick-Kakwezi/publication/236208095_The_Effects_of_Fraudulent_Procurement_Practices_on_Public_Procurement_Performance/links/55b05fd008aeb923991723c3/The-Effects-of-Fraudulent-Procurement-Practices-on-Public-Procurement-Performance.pdf

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://arxiv.org/pdf/1802.03426

Mélon, L., & Spruk, R. (2020). The impact of e-procurement on institutional quality. *Journal of Public Procurement*, *20*(4), 333–375. https://doi.org/10.1108/JOPP-07-2019-0050

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://arxiv.org/pdf/1301.3781

Modrušan, N., Mršić, L., & Rabuzin, K. (2021). Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *12*(2). https://doi.org/10.14569/IJACSA.2021.0120272

Niessen, M. E. K., Paciello, J. M., & Fernandez, J. I. P. (2020). Anomaly Detection in Public Procurements using the Open Contracting Data Standard. *2020 7th International Conference on EDemocracy and EGovernment, ICEDEG 2020*, 127–134. https://doi.org/10.1109/ICEDEG48599.2020.9096674

OECD. (2019). *Reforming Public Procurement: Progress in Implementing the 2015 OECD Recommendation* (OECD Public Governance Reviews). OECD. https://doi.org/10.1787/1de41738-en

OECD. (2022). *Public procurement*. https://www.oecd.org/en/topics/public-procurement.html

OECD. (2024a). *Public procurement*. https://www.oecd.org/en/topics/public-procurement.html

OECD. (2024b). *Strengthening Oversight of the Court of Auditors for Effective Public Procurement in Portugal: Digital Transformation and Data-driven Risk Assessments*. https://doi.org/10.1787/35aeab1e-en

OECD Data Explorer. (2024). *Size of public procurement - Government at a glance indicators, 2023 edition [Dataset]*. OECD Data Explorer. https://data-explorer.oecd.org/?fs[0]=Topic%2C1%7CPublic%20governance%23GOV%23%7CPublic%20procurement%20and%20infrastructure%23GOV_PPR%23&pg=0&fc=Topic&bp=true&snb=4

Official Journal of the European Union. (2014). *Official Journal of the European Union*, *57*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2014:094:TOC

Ortiz-Ospina, E., & Roser, M. (2016). *Government Spending*. https://ourworldindata.org/government-spending

Padhi, S. S., & Mohapatra, P. K. J. (2011). Detection of collusion in government procurement auctions. *Journal of Purchasing and Supply Management*, *17*(4), 207–221. https://doi.org/10.1016/j.pursup.2011.03.001

Palguta, J., & Pertold, F. (2017). Manipulation of Procurement Contracts: Evidence from the Introduction of Discretionary Thresholds. *American Economic Journal: Economic Policy*, *9*(2), 293–315. https://doi.org/10.1257/pol.20150511

Panayiotou, N. A., Gayialis, S. P., & Tatsiopoulos, I. P. (2004). An e-procurement system for governmental purchasing. *International Journal of Production Economics*, *90*(1), 79–102. https://doi.org/10.1016/S0925-5273(03)00103-8

Pastor Sanz, I. (2022). A New Approach to Detecting Irregular Behavior in the Network Structure of Public Contracts. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4215466

Portal BASE. (2024). Portal BASE. https://www.base.gov.pt/Base4/pt/o-portal/base/

Prasad, N. R., Almanza-Garcia, S., & Lu, T. T. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *14*(1), 1–22. https://doi.org/10.1145/1541880.1541882

Prier, E., & McCue, C. P. (2009). The implications of a muddled definition of public procurement. *Journal of Public Procurement*, *9*(3/4), 326–370. https://doi.org/10.1108/JOPP-09-03-04-2009-B002

Rahman, D. (2022). *Clustering of Neural Document Embeddings for Machine Generation of Search Extension Terms in Finnish in the Public Procurement Domain*. https://helda.helsinki.fi/server/api/core/bitstreams/5b9c2f1f-0063-4873-ab3f-1a857df51676/content

Reynolds, D. (2009). Gaussian Mixture Models. *Encyclopedia of Biometrics*, 659–663. https://doi.org/10.1007/978-0-387-73003-5_196

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. https://arxiv.org/pdf/1910.01108

Schuster, I., & Merjan, S. (2016). *Assessment Report of corruption risks in public procurement in the Republic of Moldova*. https://www.undp.org/moldova/publications/assessment-report-corruption-risks-public-procurement-republic-moldova

Scopus. (2025). *Search results for documents related to public procurement and machine learning*. Elsevier. https://www.scopus.com

Suganthan, P. G. C., Sun, C., Gayatri, K. K., Zhang, H., Yang, F., Rampalli, N., Prasad, S., Arcaute, E., Krishnan, G., Deep, R., Raghavendra, V., & Doan, A. (2015). Why Big Data industrial systems need rules and what we can do about it. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, *2015-May*, 265–276. https://doi.org/10.1145/2723372.2742784

Tas, B. K. O. (2023). Bunching below thresholds to manipulate public procurement. *Empirical Economics*, *64*(1), 303–319. https://doi.org/10.1007/s00181-022-02250-4

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

Transparency International. (2025). *Corruption Perceptions Index 2024*. https://www.transparency.org/en/cpi/2024

Treisman, D. (2000). The causes of corruption: a cross-national study. *Journal of Public Economics*, *76*(3), 399–457. https://doi.org/https://doi.org/10.1016/S0047-2727(99)00092-4

UNODC. (2013). *Guidebook on anti-corruption in public procurement and the management of public finances*. https://www.unodc.org/documents/corruption/Publications/2013/Guidebook_on

_anti-corruption_in_public_procurement_and_the_management_of_public_finances.pdf

Uyarra, E., Zabala-Iturriagagoitia, J. M., Flanagan, K., & Magro, E. (2020). Public procurement, innovation and industrial policy: Rationales, roles, capabilities and implementation. *Research Policy*, *49*(1), 103844. https://doi.org/10.1016/j.respol.2019.103844

Walker, H., & Brammer, S. (2012). The relationship between sustainable procurement and e-procurement in the public sector. *International Journal of Production Economics*, *140*(1), 256–268. https://doi.org/10.1016/j.ijpe.2012.01.008

Wallace, E., Wang, Y., Li, S., Singh, S., & Gardner, M. (2019). Do NLP Models Know Numbers? Probing Numeracy in Embeddings. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 5307–5315. https://doi.org/10.18653/v1/d19-1534

Westerski, A., Kanagasabai, R., Shaham, E., Narayanan, A., Wong, J., & Singh, M. (2021). Explainable anomaly detection for procurement fraud identification—lessons from practical deployments. *International Transactions in Operational Research*, *28*(6), 3276–3302. https://doi.org/10.1111/itor.12968

Yao, T., Zhai, Z., & Gao, B. (2020). Text Classification Model Based on fastText. *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Information Systems, ICAIIS 2020*, 154–157. https://doi.org/10.1109/ICAIIS49377.2020.9194939

Yeo, I.-K., & Johnson, R. A. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

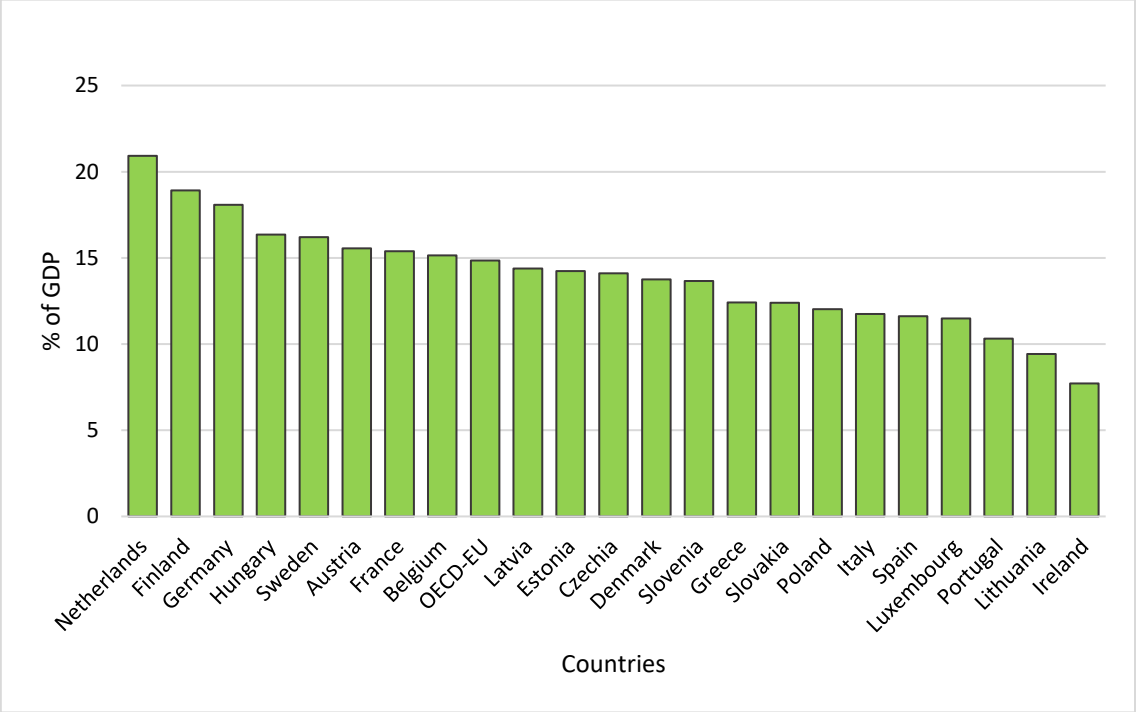
Yu, W., Cheng, W., Aggarwal, C. C., Zhang, K., Chen, H., & Wang, W. (2018). NetWalk: A flexible deep embedding approach for anomaly detection in dynamic networks. *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2672–2681. https://doi.org/10.1145/3219819.3220024

Zavrtanik, V., Kristan, M., & Skočaj, D. (2021). DRÆM - A discriminatively trained reconstruction embedding for surface anomaly detection. *Proceedings of the IEEE International Conference on Computer Vision*, 8310–8319. https://doi.org/10.1109/ICCV48922.2021.00822

Zhang, H., Liu, J., Zhu, Z., Zeng, S., Sheng, M., Yang, T., Dai, G., & Wang, Y. (2024). *Efficient and Effective Retrieval of Dense-Sparse Hybrid Vectors using Graph-based Approximate Nearest Neighbor Search*. https://arxiv.org/pdf/2410.20381

Zhang, Z., Jasaitis, T., Freeman, R., Alfrjani, R., & Funk, A. (2023). *Mining Healthcare Procurement Data Using Text Mining and Natural Language Processing - Reflection From An Industrial Project*. https://doi.org/10.48550/arXiv.2301.03458

APPENDIX A – GENERAL GOVERNMENT PROCUREMENT SPENDING AS A PERCENTAGE OF GDP FOR 2021



Data source: OECD Data Explorer (2024)

APPENDIX B - GLOSSARY OF PROCUREMENT TERMS (PORTUGUESE - ENGLISH)

B1 – CONTRACT TYPES

Portuguese Term	English Translation
Aquisição de Bens Móveis	Acquisition of Movable Goods
Aquisição de Serviços	Acquisition of Services
Empreitadas de Obras Públicas	Public Works Contracts

B2 – PROCEDURE TYPES

Portuguese Term	English Translation
Ao abrigo de acordo-quadro (Art. 258.º)	Under Framework Agreement (Art. 258)
Ao abrigo de acordo-quadro (Art. 259.º)	Under Framework Agreement (Art. 259)
Ajuste Direto Regime Geral	Direct Award - General Regime
Concessão de Serviço Público	Public Service Concession
Concurso Público	Open Tender
Concurso Público Simplificado	Simplified Public Tender
Consulta Prévia	Prior Consultation
Contratação Excluída II	Excluded Procurement II
Parceria para a Inovação	Partnership for Innovation
Procedimento de Negociação	Negotiation Procedure

B3 – LEGAL BASIS AND SPECIAL MEASURES

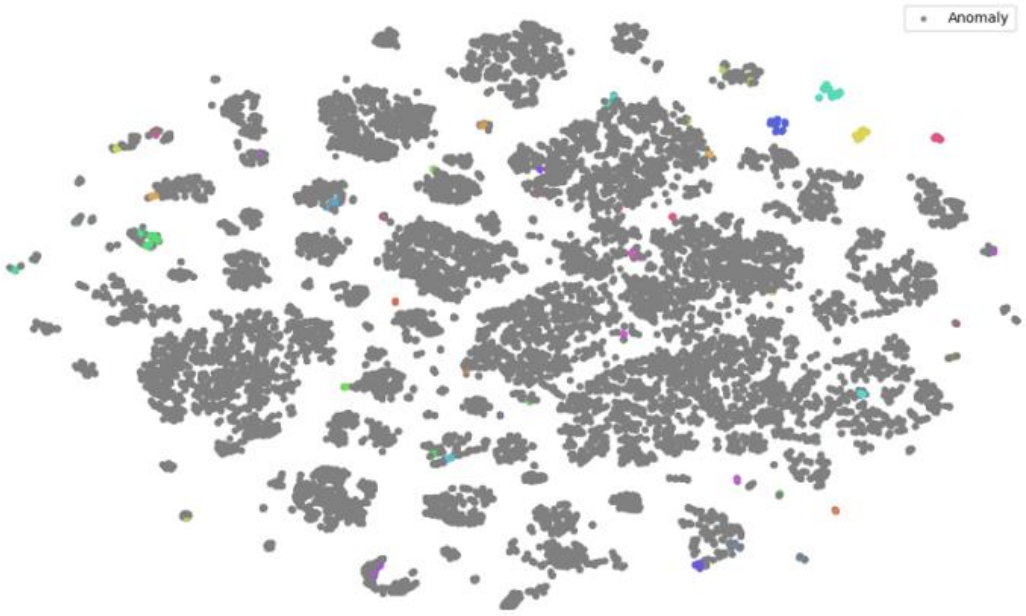
Portuguese Term	English Translation
Artigo 20.º, n.º 1, alínea a) do Código dos Contratos Públicos	Article 20, No. 1, paragraph a) of the CPC
Código dos Contratos Públicos (DL 111-B/2017) e Lei n.º 30/2021	CPC (DL111-B/2017) and Law No. 30/2021
Plano de Recuperação e Resiliência (PRR) – Artigo 6.º da Lei n.º 30/2021	Recovery and Resilience Plan (RRP) – Article 6 of Law No. 30/2021

B4 – CPV CODES

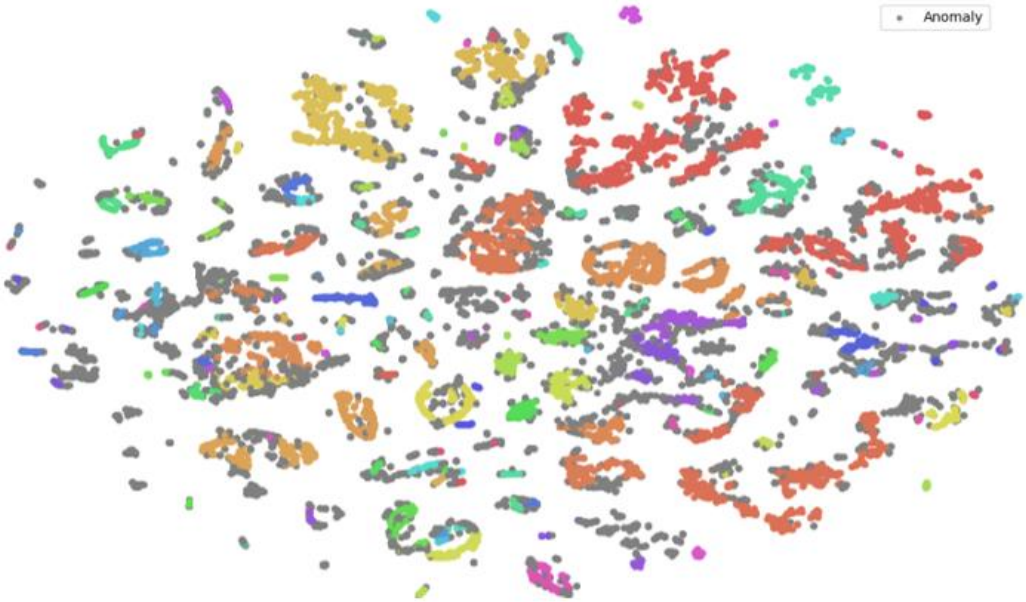
CPV Code	Portuguese Term	English Translation
33	Equipamento médico, medicamentos e produtos para cuidados pessoais	Medical equipment, pharmaceuticals and personal care products
45	Construção	Construction work
60	Serviços de transporte (exceto transporte de resíduos)	Transport services (excluding waste transport)
71	Serviços de arquitectura, construção, engenharia e inspecção	Architectural, construction, engineering and inspection services
79	Serviços a empresas: direito, comercialização, consultoria, recrutamento, impressão e segurança	Business services: law, marketing, consulting, recruitment, printing and security
92	Serviços recreativos, culturais e desportivos	Recreational, cultural and sporting services

APPENDIX C - PRESENTATION OF NUMERICAL NOISE THROUGH ANOMALIES DETECTED USING GMM AND DBSCAN, ILLUSTRATED BY T-SNE VISUALIZATIONS

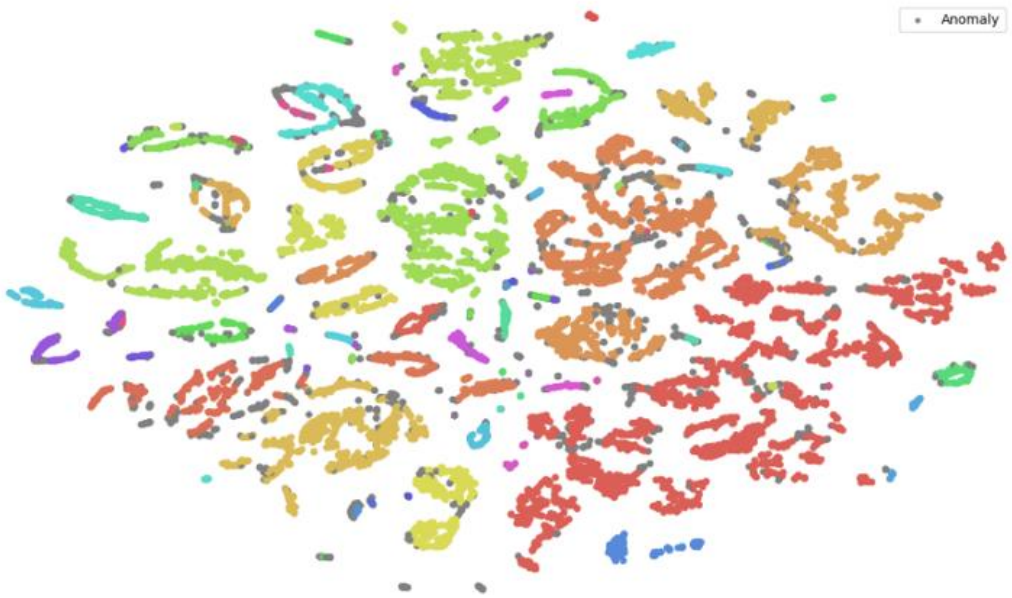
C1 – Utilization of All Original Numerical Features, Including Transformed Date Features (Days, Months, Years, and Weekend Indicators)



C2 - Utilization of All Original Numerical Features, Including Transformed Date Features Limited to Weekend Indicators



C3 - Utilization of Contract Price and Total Effective Price, with Execution Period Converted to Text



APPENDIX D – DIMENSIONALITY REDUCTION TECHNIQUES FOR 50 FEATURES

Embedding Type	Dimensionality Reduction Technique	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
Word2Vec	PCA	0.0146	5773.17	3.8244
	UMAP	0.3449	78934.76	1.2471
FastText	PCA	0.0475	10148.12	3.4795
	UMAP	0.5349	111447.31	1.1501
BERTimbau	PCA	0.0062	6191.24	3.7562
	UMAP	0.4333	125010.27	1.1044
DistilBERT	PCA	0.0077	6411.19	4.0049
	UMAP	0.4848	134868.91	1.2806
LaBSE	PCA	0.0153	7572.62	3.4569
	UMAP	0.4945	121395.49	0.9743

APPENDIX E - UMAP VISUALIZATIONS WITH GMM CLUSTERING BY EMBEDDING TYPE

E1 - WORD2VEC (60 CLUSTERS)

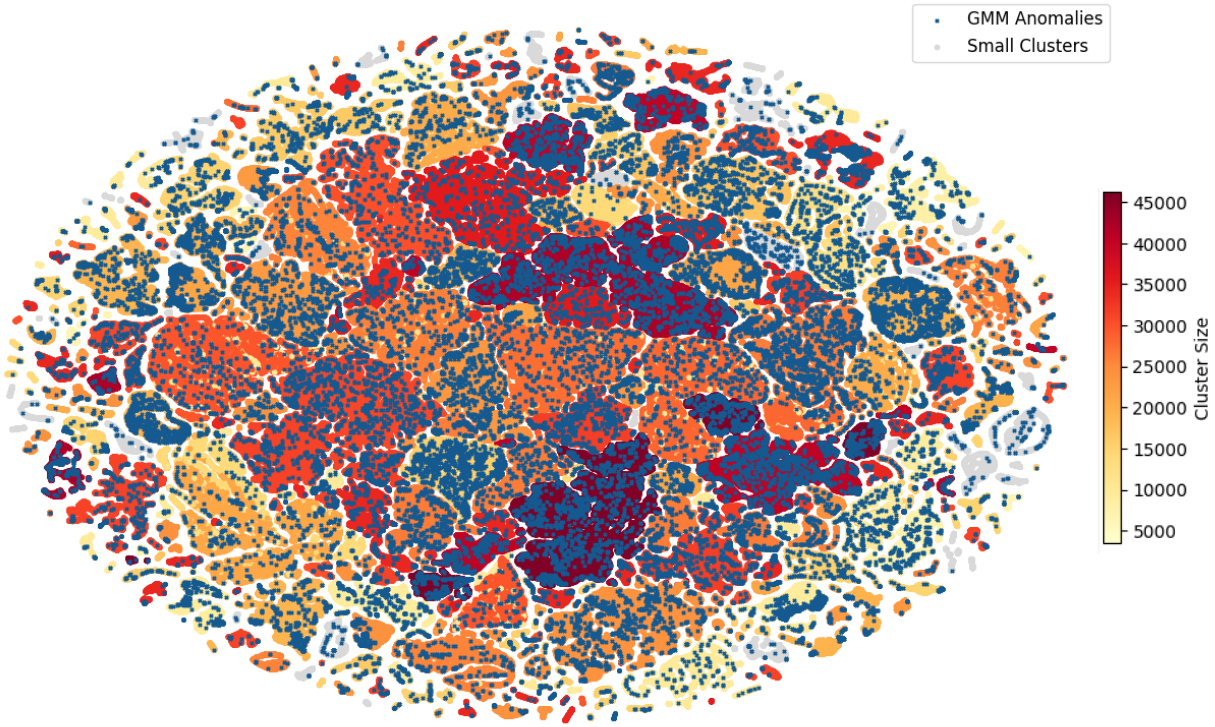
E2 - FASTTEXT (50 CLUSTERS)

E3 - BERTimbau (60 Clusters)

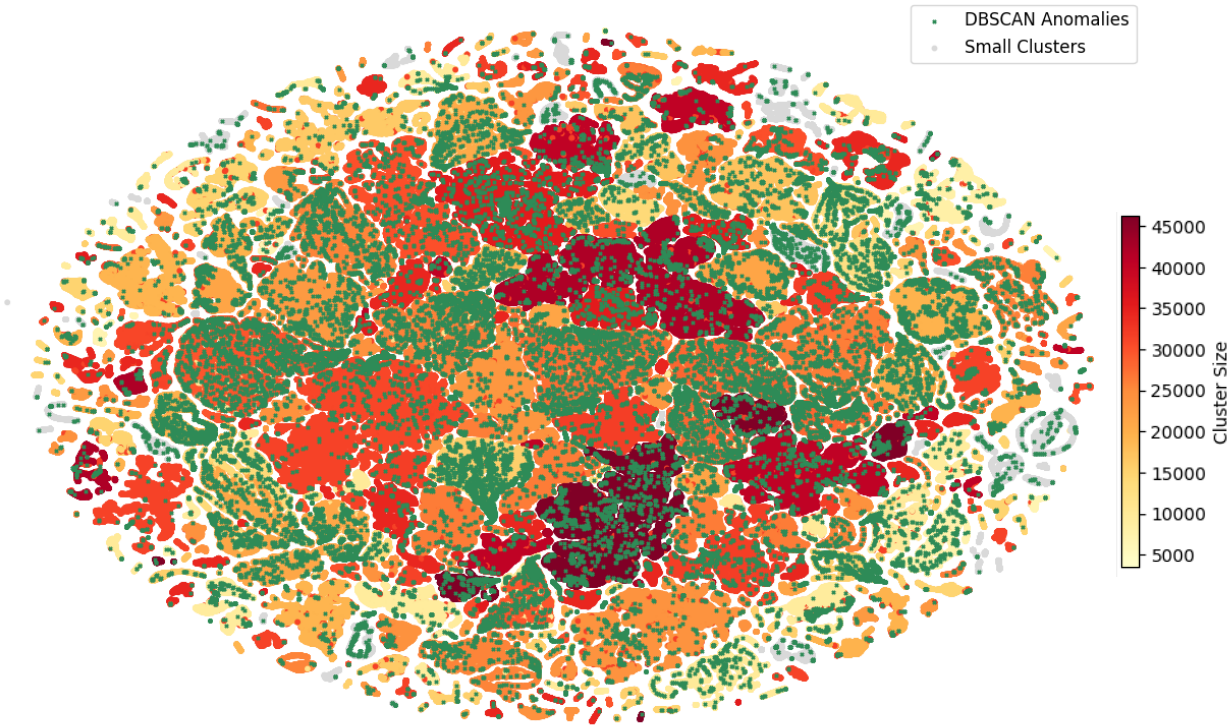
E4 - DistilBERT (50 Clusters)

APPENDIX F - COMPLEMENTARY T-SNE VISUALIZATIONS FOR LABSE

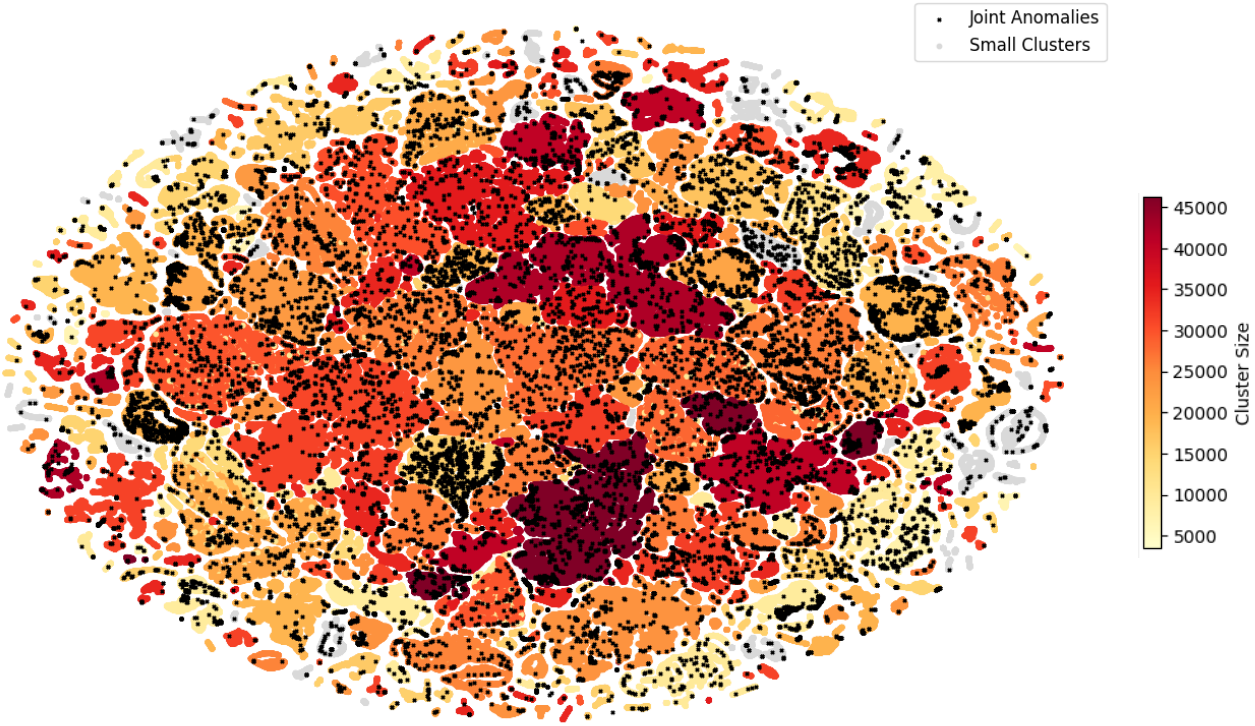
F1 - GMM ANOMALY DETECTION MODEL (70 CLUSTERS)



F2 - DBSCAN ANOMALY DETECTION MODEL (70 CLUSTERS)



F3 – GMM AND DBSCAN ANOMALY DETECTION MODEL



APPENDIX G - ANOMALY PERCENTAGE OF THE 10 SMALLEST CLUSTERS BY MODEL

GMM Cluster	Total Points	Anomaly Percentage GMM	Anomaly Percentage DBSCAN
40	1	0.00	100.00
54	96	0.00	0.00
33	109	0.00	3.67
62	144	0.00	0.69
45	150	0.00	0.00
58	156	0.00	2.56
39	187	0.00	0.00
37	262	0.00	1.15
44	403	0.00	0.74
56	584	0.86	6.68

APPENDIX H - ETHICS COMMITTEE REPORT

Project No.: **DSCI2025-4-259152**

Project Title: **Anomaly Detection in Portuguese Public Procurement Contracts: An Embedding-Based Machine Learning Approach**

Principal Researcher: **Ana Rita Saraiva da Silva**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 29/04/2025.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

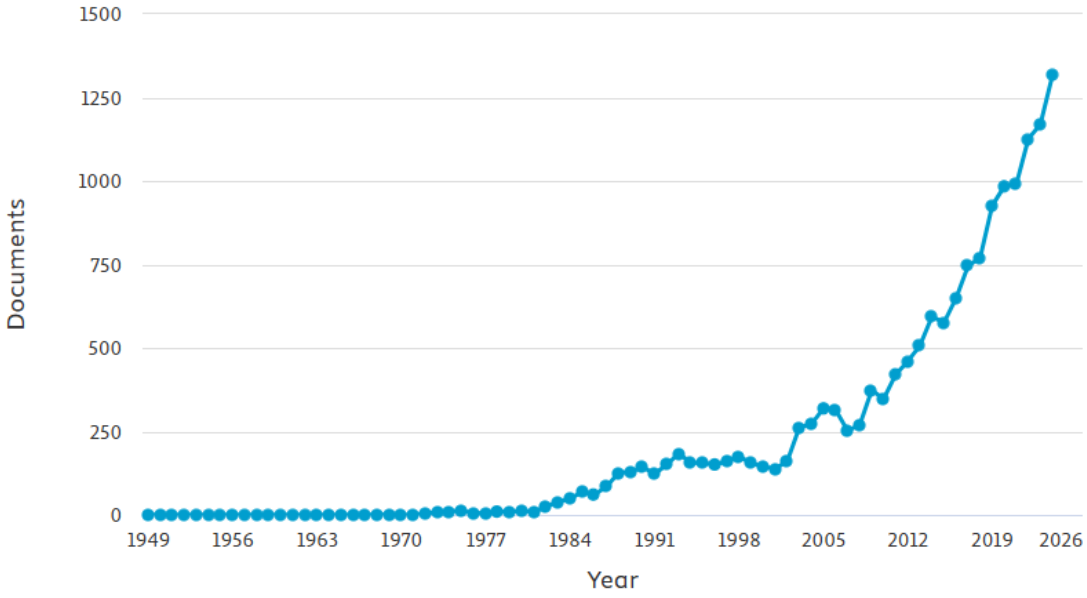
The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of

- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 29/04/2025
NOVA IMS Ethics Committee
ethicscommittee@novaims.unl.pt

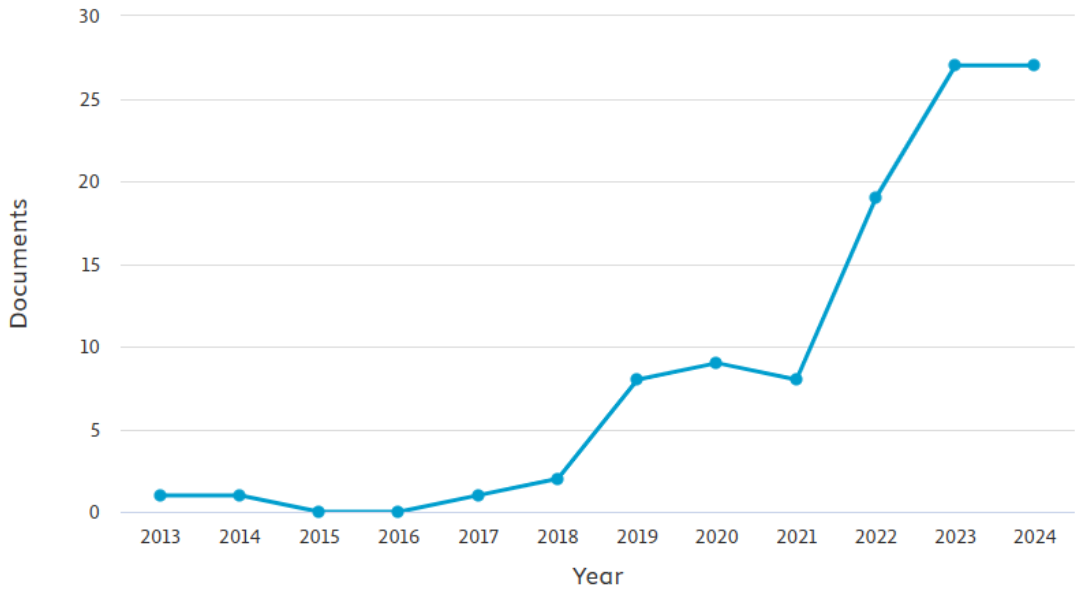
ANNEX A – PUBLIC PROCUREMENT RESEARCH TREND ACCORDING TO SCOPUS

A1 - NUMBER OF DOCUMENTS BY YEAR RELATED TO PUBLIC PROCUREMENT



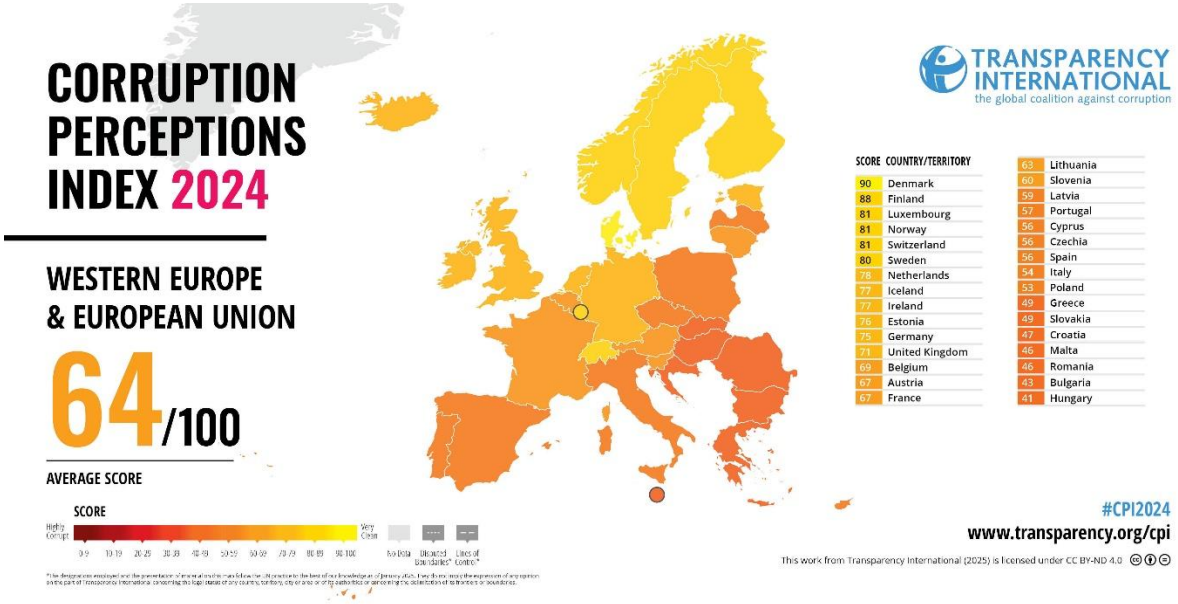
Source: Scopus (2025). Retrieved June 8, 2025.

A2 - NUMBER OF DOCUMENTS BY YEAR RELATED TO PUBLIC PROCUREMENT AND MACHINE LEARNING



Source: Scopus (2025). Retrieved June 8, 2025.

ANNEX B – CORRUPTION PERCEPTIONS INDEX 2024 IN EUROPE



Source:Transparency International (2025). Retrieved June 6, 2025.

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

Universidade Nova de Lisboa