UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Towards a Methodology for Analyzing Public Procurement Data from Kapak's Database

.

Ilter Sthefano Ulloa Miranda

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 18 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

**HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA**

Towards a Methodology for Analyzing Public Procurement Data from Kapak's Database

Ilter Sthefano Ulloa Miranda

Nombre del profesor, Título académico	**Daniel Riofrío, Ph. D**

Quito, 18 de diciembre de 2024

© DERECHOS DE AUTOR

Nombres y apellidos:	Ilter Sthefano Ulloa Miranda
Código:	00215689
Cédula de identidad:	1750312181
Lugar y fecha:	Quito, 18 de diciembre de 2024

ACLARACIÓN PARA PUBLICACIÓN

UNPUBLISHED DOCUMENT

RESUMEN

Los sistemas de contratación pública generan enormes cantidades de datos complejos y heterogéneos que contienen información valiosa para la detección de riesgos de corrupción. Sin embargo, analizar estos datos presenta desafíos significativos debido a su naturaleza no estructurada, almacenamiento fragmentado y complejidades técnicas. Esta investigación introduce un novedoso pipeline de datos diseñado para optimizar el análisis de datos de contratación pública del Sistema Oficial de Contratación Pública (SOCE) de Ecuador, basándose en el proyecto Kapak, una iniciativa que utiliza tecnologías de big data para mejorar la transparencia y la rendición de cuentas en el sistema de contratación pública del país.

El proyecto Kapak mantiene una base de datos integral de procesos de contratación mediante su rastreador web automatizado, que extrae datos del SOCE. No obstante, la complejidad de los datos recopilados —incluyendo archivos USHAY codificados en base64, documentos fragmentados y datos JSON dispersos— genera importantes barreras para su análisis. Nuestro pipeline aborda estos desafíos críticos al proporcionar una solución automatizada para la reconstrucción de archivos, decodificación y consolidación de datasets.

Para demostrar la efectividad del pipeline, aplicamos técnicas de análisis de texto a documentos de contratación de la modalidad de Subasta Inversa Electrónica (SIE). Utilizando vectorización TF-IDF y similitud coseno, analizamos similitudes entre procesos de contratación de alto riesgo, revelando patrones potenciales que podrían indicar riesgos de corrupción. Nuestro análisis se centró particularmente en las especificaciones técnicas, documentos de licitación e interacciones de preguntas y respuestas entre las partes interesadas.

Esta investigación contribuye al campo al proporcionar un pipeline generalizable que transforma datos complejos y sin procesar en datasets listos para análisis. Esto permite a los investigadores y organismos de supervisión enfocarse en desarrollar y aplicar métodos analíticos avanzados, en lugar de enfrentarse a los desafíos de preprocesamiento de datos. El diseño modular de esta metodología facilita la integración de técnicas de PLN más sofisticadas, como modelos basados en transformers u otros enfoques de última generación, sentando las bases para un análisis más completo de riesgos de corrupción en los sistemas de contratación pública. Al aprovechar la infraestructura existente de Kapak, este pipeline mejora la capacidad del proyecto para servir como una herramienta poderosa para la supervisión de contrataciones y la prevención de la corrupción.

Palabras clave: Contratación pública, detección de riesgo de corrupción, Big data, SOCE, Subasta Inversa Electrónica (SIE), procesamiento de lenguaje natural (PLN), Pipeline de datos, Metodología, Transparencia, Kapak.

ABSTRACT

Public procurement systems generate vast amounts of complex, heterogeneous data that holds valuable insights for corruption risk detection. However, analyzing this data presents significant challenges due to its unstructured nature, fragmented storage, and technical complexities. This research introduces a novel data pipeline designed to streamline the analysis of procurement data from Ecuador's Sistema Oficial de Contratación Pública (SOCE), building upon the Kapak project—an initiative that leverages big data technologies to enhance transparency and accountability in Ecuador's public procurement system.

The Kapak project maintains a comprehensive database of procurement processes through its automated web crawler, which extracts data from SOCE. However, the complexity of the data—including base64-encoded USHAY files, fragmented documents, and scattered JSON data—creates significant barriers to performing analysis. Our pipeline addresses these critical challenges by providing an automated solution for file reconstruction, decoding, and dataset consolidation.

To demonstrate the pipeline's effectiveness, we used text analysis techniques on procurement documents from the Subasta Inversa Electrónica (SIE) modality. Using TF-IDF vectorization and cosine similarity, we analyzed similarities among high-risk procurement processes, revealing potential patterns that could indicate corruption risks. Our analysis focused on technical specifications, bidding documents, and question-answer interactions between stakeholders.

This research contributes to the field by providing a generalizable pipeline that transforms raw, complex procurement data into analysis-ready datasets. This enables researchers and oversight bodies to focus on developing and applying advanced analytical methods rather than dealing with data preprocessing challenges. The methodology's modular design facilitates the integration of more sophisticated NLP techniques, such as transformer-based models or other state-of-the-art approaches, laying the groundwork for more comprehensive corruption risk analysis in public procurement systems. By building upon Kapak's existing infrastructure, this pipeline enhances the project's capability to be a powerful tool for procurement oversight and corruption prevention.

Key words: Public procurement, corruption risk detection, Big data, SOCE, Electronic Reverse Auction (SIE), natural language processing (NLP), Data Pipeline, Methodology, Transparency, Kapak.

TABLE OF CONTENT

TABLE OF FIGURES

INTRODUCTION

Public procurement in Ecuador plays a critical role in government operations, with significant implications for economic integrity and public trust. **Sistema Oficial de Contratación Pública del Ecuador (SOCE)** is Ecuador's central e-procurement system, overseeing diverse processes such as open bidding, restricted bidding, negotiated procedures, and electronic reverse auctions [1]. Given its scale and complexity, SOCE is vulnerable to corruption risks, which have become increasingly important to address, particularly in light of recent global crises [2]. The **COVID-19 pandemic** highlighted the urgent need for transparency and accountability in public procurement systems worldwide. Governments faced immense pressure to secure critical supplies, exposing weaknesses in procurement processes and revealing a heightened risk of corruption [2]. Irregularities such as price gouging, preferential supplier treatment, and opaque contract awards were reported globally, particularly in countries with fragile governance structures [3]. The economic impact of such procurement-related corruption is substantial, with global public procurement estimated to be a $13 trillion market. Studies suggest that corruption inflates procurement costs by 20% to 25%, and in some regions, such as Latin America and the Caribbean, these losses amount to approximately 4.4% of GDP [2]. In Ecuador, around 200,000 public contracts were awarded in 2020 alone, but only a small portion has been formally investigated for irregularities [2]. SOCE, as the primary hub for managing these contracts, remains at the forefront of efforts to enhance procurement integrity. Among the most significant procedures managed within SOCE are **Subasta Inversa Electrónica (SIE)**, an electronic reverse auction, and **Giro Específico del Negocio (GEN)**, a direct procurement method used under specific circumstances [1]. Each of these modalities is governed by distinct regulatory methodologies and faces unique corruption risks.

Leveraging Big Data for Transparency

The digitalization of public procurement systems, such as SOCE, has resulted in the accumulation of large volumes of data, much of which is unstructured and highly complex [1]. This data presents a valuable opportunity for corruption monitoring if appropriately analyzed [3]. Big data technologies, combined with text analysis techniques, allow for the systematic examination of procurement processes, enabling the identification of patterns and anomalies indicative of fraud [4]. For example, sentiment analysis applied to procurement documents can help uncover bias or favoritism toward certain suppliers, providing a more nuanced understanding of potential corruption risks [4].

The Kapak Project: Overview and Limitations

To address these challenges, the **Kapak** project was developed to enhance transparency and accountability within Ecuador’s public procurement system [5]. Kapak leverages big data technologies to focus on the SIE and GEN modalities, providing a methodology for detecting corruption risks through the analysis of procurement data. The project comprises several key components:

- **Web Crawler:** Automatically gathers data from SOCE, including contract details, supplier information, and tender documents.
- **Data Lakehouse:** Serves as a central repository for storing and managing structured, semi-structured, and unstructured procurement data.
- **Web Portal:** Provides a public interface displaying corruption risk indicators, enabling citizens and stakeholders to monitor procurement activities.
- **Matrix of Indicators:** A set of 122 predefined corruption risk indicators, developed in consultation with legal and procurement experts, is used to evaluate procurement processes across various dimensions, including pricing, competition, and contract modifications.

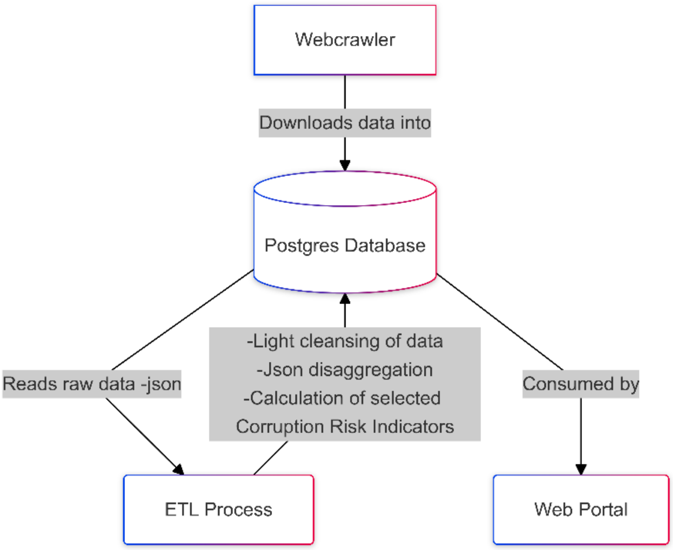


Figure 1: Kapak architecture

While Kapak offers significant potential for improving procurement transparency, it also faces notable limitations. Currently, it operates primarily as a **retrospective tool**, analyzing completed procurement processes, which limits its ability to prevent corruption in real-time. Furthermore, its scope is restricted to only two procurement modalities—SIE and GEN—representing just a subset of all SOCE-managed activities. Additionally, the efficacy of Kapak is heavily dependent on the quality and completeness of the data it analyzes; any inaccuracies or inconsistencies within SOCE data can impair the accuracy of its risk assessments.

Future Directions

Future enhancements to Kapak aim to broaden its scope to include additional procurement modalities and integrate real-time analysis capabilities. These improvements would enable more proactive monitoring and detection of corruption risks, providing stakeholders with timely insights to prevent irregularities before contracts are awarded. However, there is still more groundwork to be done, such as establishing a methodology to develop, analyze, and iterate datasets and algorithms in a reproducible and fast way.

RELATED WORK

Public Procurement Data Analysis Approaches

Corruption in public procurement is a significant global issue due to its economic impact, leading to inefficiencies, inflated costs, and eroded public trust. A key challenge in procurement analysis is the variety and complexity of data sources involved. While some studies focus on structured data like bid prices and dates, comprehensive corruption detection requires analyzing both structured and unstructured data sources. Numerous studies have explored data mining and machine learning techniques to detect and prevent corruption in procurement processes. These methods typically involve analyzing large datasets to identify patterns and anomalies indicative of corrupt practices such as bid-rigging, favoritism, and price fixing [6]. Supervised learning techniques, which rely on labeled datasets of known corruption cases, have been effectively utilized in several contexts to detect such irregularities. However, the challenge of acquiring accurately labeled data remains a critical obstacle in many countries.

To address data complexity challenges, unsupervised learning methods have emerged as promising alternatives, as they can identify outliers across varied procurement data types without requiring labeled training sets. Araujo et al [7], demonstrated this by applying one-class SVMs to detect anomalous bids within a dataset of 626 procurement proposals. Other approaches have explored graph databases to model complex relationships between contractors, contracts, and authorities, facilitating the detection of collusion and conflicts of interest, though these approaches often require careful data preprocessing and structuring [3]. These studies highlight the broad range of methodologies being employed across various countries to combat corruption in public procurement, each with different data handling requirements and capabilities.

Text Analysis in Procurement Data

While structured procurement data provides valuable insights, a significant portion of procurement information exists in unstructured text format –from tender specifications to bid documents and communications. This presents unique challenges for analysis, as procurement texts often contain domain-specific terminology and complex legal language. Studies applying NLP to procurement have had to address these challenges through specialized preprocessing and domain adaptation approaches. While there is limited literature directly applying techniques such as TF-IDF and cosine similarity in the public procurement context, Natural Language Processing (NLP) holds considerable promise for analyzing textual data within procurement processes.

NLP has been employed to extract insights from unstructured data, including tender documents, contracts, and communication records, to detect hidden patterns indicative of corruption risks. For instance, sentiment analysis has been used to identify biases or favoritism in public procurement, particularly in interactions between suppliers and contracting entities [4]. In Ecuador, Torres-Berru et al [1]applied sentiment analysis to procurement-related questions and responses, revealing instances where responses appeared to favor specific bidders, suggesting potential corruption. These applications of NLP highlight the importance of examining textual data in procurement, where subjective biases and hidden patterns may not be evident through the analysis of structured data alone.

Data Processing Architectures and Tools

Existing procurement analysis tools vary significantly in their ability to handle different data types and volumes. Many early initiatives focused on small, structured datasets that were relatively straightforward to analyze. For example, several studies in Ecuador analyzed samples of less than 1,000 procurement processes using basic statistical methods [2]. In contrast, newer initiatives like Kapak process millions of documents, combining structured procurement records with unstructured text data from multiple sources [5]. This requires more sophisticated data processing architectures, including data lakehouses and specialized preprocessing pipelines.

Other countries have implemented innovative tools for public procurement oversight with varying data processing capabilities. Brazil’s decision support system integrates graph theory, clustering, and regression analysis to detect collusion and conflicts of interest in procurement processes, representing a middle-ground approach that handles moderate data complexity through integrated analysis of multiple structured data sources [8]. While these systems have proven effective, they largely depend on structured data and predefined indicators. This focus limits their ability to uncover more subtle forms of corruption hidden within unstructured data, such as tender documents or communication records. The variation in data handling capabilities across different tools and methodologies highlights the need for a more generalized methodology for procurement data analysis that can effectively process both structured and unstructured data at scale.

RESEARCH OBJECTIVES

1. **Formulation of a Question-Driven Analytical Methodology:** Establish a comprehensive methodology that can aid in creating a base path to analyze public procurement data from the Kapak database in a structured and reproducible way.

2. **Development of a Generalized Data Cleaning Pipeline:** Develop a data cleaning pipeline that aids the methodology. By addressing issues like missing data, segmented files, and disaggregated JSONs. The goal is to create a scalable pipeline capable of supporting diverse analyses of public procurement data.

3. **Empirical Validation through Case Studies:** Demonstrate the methodology and pipeline's effectiveness by applying it to case studies from the Kapak dataset.

DATA SOURCES AND COLLECTION

As both the methodology and pipeline will reference aspects, or specific data that is present in the Kapak database, it is necessary to understand the database, it's sources and challenges.

Public Procurement Data in Ecuador

The primary data source for this research is Ecuador's public procurement system, SOCE (Sistema Oficial de Contratación Pública), managed by SERCOP (Servicio Nacional de Contratación Pública del Ecuador). The focus of the analysis is on procurement processes from the Subasta Inversa Electrónica (SIE) for the year 2022, although the data pipeline is flexible and allows for the selection of other years or process types.

The raw data consists of several types of documents that detail different stages of procurement processes, including:

- **USHAY Files:** These compressed files contain documents and detailed information about each procurement process in an XML format. The XML file is generated by SOCE's "Módulo Facilitador," a PHP application used by contracting entities to create new procurement processes and by suppliers to submit proposals. The XML content summarizes the procurement process and lists the associated files or documents. However, the XML is initially encoded in Base64, and some tags require additional decryption and processing before the information can be fully utilized.
- **Pliegos (Specifications):** These documents, that are part of the .ushay files, outline the technical requirements, terms, and conditions of the procurement contracts.
- **Questions and Clarifications:** These are records of interactions between bidders and contracting entities, which provide critical insights into the procurement process.

- **Contract Summaries:** These summarize the final terms of awarded contracts, including details about the suppliers and contract amounts.

- **Other data:** The SOCE portal also includes other type of data, that in some cases is only available at certain process phases. This data includes but is not limited to publication date, bid submission deadlines, contract award date.

The Kapak database, sourced from SOCE, contains both structured and unstructured data:

- **Structured Data:** Organized in database tables, this includes details on each SIE process (e.g., contract ID, contracting entity, supplier, and contract value), stored in tables such as soce_sie_ic_proceso and soce_descripcion.

- **Unstructured Data:** Primarily in document form, this includes information embedded within USHAY files and JSON files containing the "questions and clarifications" data, which require specialized extraction and text processing techniques to make them analyzable.

Challenges in Data Collection and Processing

Several challenges arose in the extraction, cleaning, and processing of the data, which required customized solutions:

- **Understanding the Kapak Database:** Before conducting any analysis, it was essential to gain a thorough understanding of the database design and its limitations. The Kapak database contains tables for storing raw information, tables for processed data, and numerous temporary tables and views used for preprocessing data ingested by the web crawler. As a result, not all tables are fully normalized, and some do not consistently adhere to the use of the primary key, contract_id, or treat it as the same data type across all instances. For instance, the figure of a fragment of the Database diagram shows how only a handful of tables are actually related with contract_id:

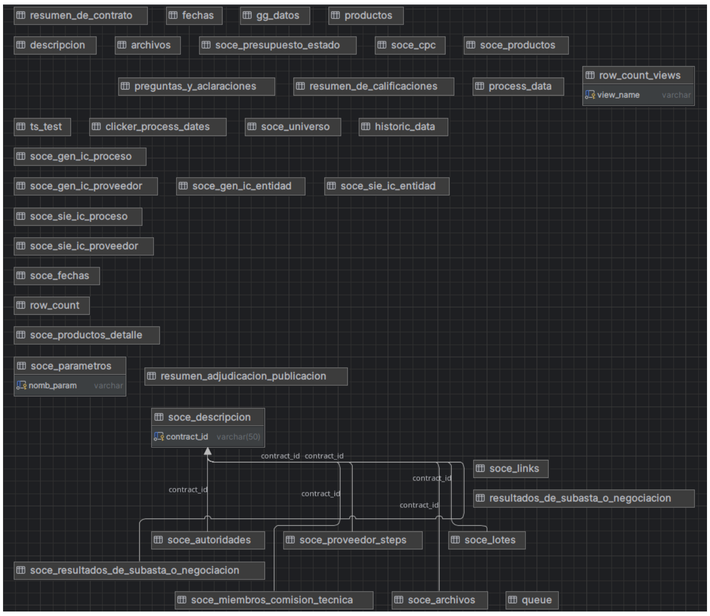


Figure 2: Fragment of Database diagram

- **File Reconstruction from the Kapak Database:** The Kapak project database, which uses PostgreSQL, includes a table called soce_links containing various details about files or documents from the SOCE portal. Among these is a bytea column that stores the binary data of the actual file. However, the database does not include columns for the file's name or extension, meaning this information must be retrieved from the SOCE portal to complete the file reconstruction process.

- **Data Extraction from Unstructured Sources:** Extracting usable information from the SOCE's USHAY compressed files required the use of the 7zip utility, and a specialized pipeline to handle fragmented files or repeated parts since the soce_links table contains duplicate references to files, due to the unexpected SOCE generation of different links per visit to a file hosted there.

- **Data Cleaning and Standardization:** The dataset contained inconsistent formats, missing values, and disaggregated JSON files, all of which required specialized handling. Each data type was processed to extract text from various formats, including PDF, Word, older Word formats, and Excel.

- **Data Decoding and Decryption:** The XML file in the USHAYs is encoded in Base64, and certain tags related to numeric quantities, such as monetary amounts, are encrypted using the Advanced Encryption Standard (AES). To handle this, it was necessary to reverse-engineer the "Módulo Facilitador" PHP code in order to decipher the encryption. This process was automated to efficiently decode and decrypt all the XML files in the sample.

- **Data Volume:** The large volume of data, particularly in the soce_links table, which stores downloaded procurement files, posed storage and processing challenges. Efficient handling of these large datasets required advanced data storage and optimization strategies.

Indicators for Corruption Risk

In collaboration with legal experts, a matrix of corruption risk indicators was developed to assess the effectiveness of SOCE in mitigating corruption risks [5]. Each indicator represents a potential red flag that could signal corrupt practices, and these are based on established literature and an analysis of public procurement corruption risks.

The calculation of these indicators provides valuable insights within the Kapak database, as they can serve as target classes for supervised learning algorithms. However, a significant limitation of the Kapak database is its incomplete information for calculating all corruption risk indicators across every procurement process. This issue stems from technical constraints of the web crawler, which can struggle to consistently access and extract data from all SOCE modules. For example, the crawler may experience timeouts or errors when trying to access certain URLs, resulting in incomplete data extraction, particularly from sections such as "Questions and Clarifications." These inconsistencies hinder the calculation of some indicators and limit the overall comprehensiveness of the dataset.

THE NEED FOR A DATA ANALYSIS METHODOLOGY

Public procurement data analysis presents unique challenges that require a systematic methodological approach. While traditional model-driven or rule-based approaches like procurement indicators have provided valuable insights, they often struggle to capture the full complexity of modern purchasing processes [6]. As procurement generates increasingly large volumes of structured and unstructured data across multiple systems and formats, there is a clear need for more sophisticated data-driven methodologies [1]. The complexity and volume of procurement data creates several key challenges that necessitate a structured analytical methodology. First, procurement data often exhibits high dimensionality and heterogeneity, combining numerical data (e.g. prices, quantities), text data (e.g. tender descriptions, specifications), and categorical data (e.g. procurement types, vendor categories) [7]. Second, procurement processes generate data across multiple stages and systems, requiring careful integration and cleaning approaches. Third, procurement data may contain various biases and quality issues that need to be systematically addressed [4]. The need for such methodological rigor is particularly acute in procurement analytics, where findings directly impact public spending decisions. As noted by Ortiz-Prado et al. [2], analysis of procurement data requires careful consideration of data quality, statistical validity, and potential biases to generate reliable insights. A comprehensive methodology ensures that analyses meet scientific standards while remaining practically applicable for procurement decision-making.

KEY ASPECTS OF AN EFFECTIVE ANALYSIS METHODOLOGY

A data-driven methodology helps address procurement analysis challenges through several key components. As demonstrated by Lu et al. [9], an effective analytical methodology must incorporate:

Data Quality and Preprocessing

The methodology must provide systematic approaches for assessing and ensuring data quality. This includes procedures for:

- Data cleaning and standardization.
- Handling missing or incomplete data.
- Integration of heterogeneous data sources.
- Detection and treatment of outliers and anomalies.

Validation and Testing

Multiple validation approaches are essential for ensuring reliable results [8]:

- Statistical hypothesis testing.
- Cross-validation across different datasets.
- Domain expert review and validation.
- Performance metrics assessment.

Documentation and Reproducibility

Clear documentation is critical for ensuring reproducibility and knowledge transfer [5]:

- Detailed methodology documentation.
- Clear statement of assumptions and limitations.

- Code and data accessibility where possible.
- Documentation of parameter choices and thresholds.

Integration of Methods

The methodology should support integration of multiple analytical approaches [1]:

- Statistical analysis.
- Machine learning techniques.
- Network analysis.
- Text mining and natural language processing.

Additionally, a robust methodology must be flexible enough to accommodate different procurement contexts while maintaining analytical rigor. This includes supporting both exploratory analysis for hypothesis generation and confirmatory analysis for hypothesis testing. The methodology should also facilitate the incorporation of domain knowledge and expertise while maintaining objectivity in the analytical process. The effectiveness of such a methodology can be assessed through several criteria:

- Reliability and reproducibility of results.
- Ability to detect meaningful patterns and anomalies.
- Practical applicability in procurement decision-making.
- Scalability to different procurement contexts.
- Support for continuous improvement and learning.

By incorporating these key aspects, a procurement data analysis methodology provides the foundation for reliable, reproducible analytics while remaining adaptable to different analytical needs and contexts.

ANALYTICAL METHODOLOGY

This section presents a systematic methodology for conducting research using public procurement data from Ecuador’s SOCE system through the Kapak project’s data pipeline. The methodology guides researchers through a structured decision-making process while maintaining flexibility for diverse research objectives.

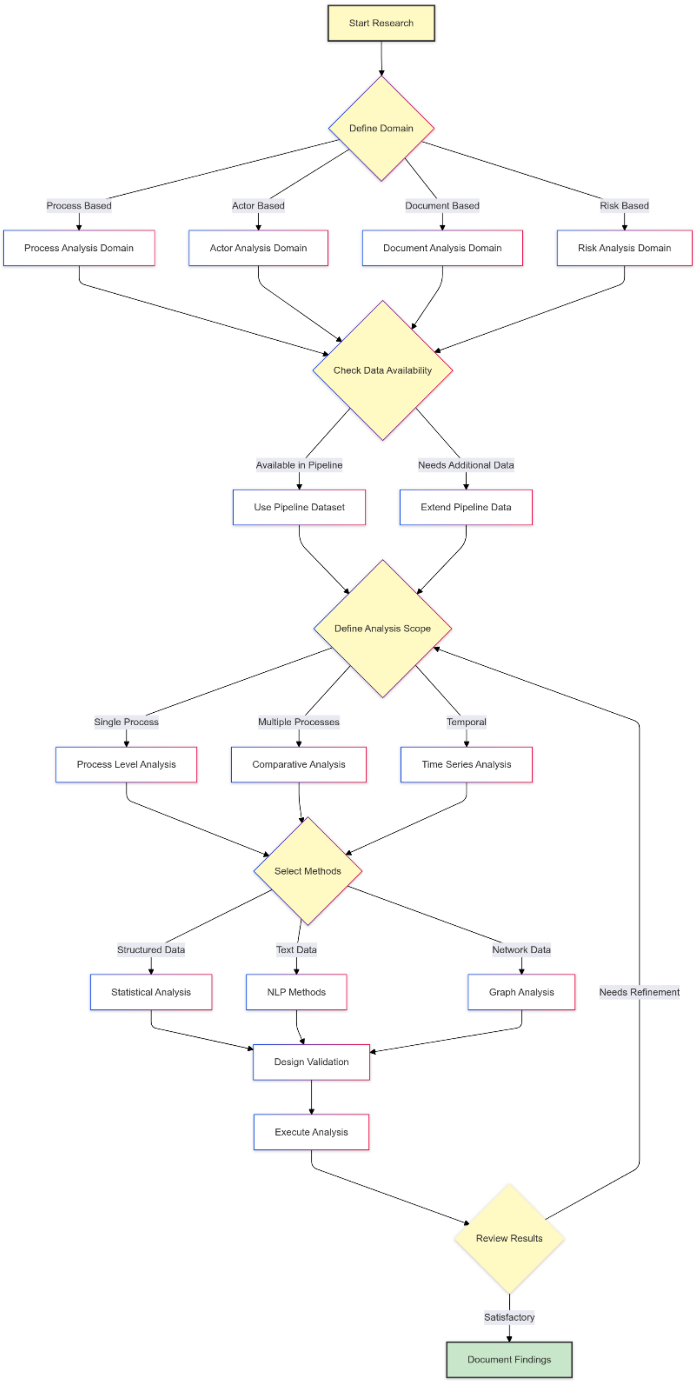


Figure 3: Analytical Methodology flowchart

Domain Identification

The first step involves identifying the primary domain of investigation. Procurement analysis can be approached through several key perspectives:

- **Process-Based Analysis:** Focuses on procurement procedures, their characteristics, and outcomes. Researchers can leverage structured data from the pipeline's cleaned dataset, particularly columns prefixed with sie_ic_ and sd_ which contain process metadata, status information, and temporal data.

- **Actor-Based Analysis:** Examines relationships and patterns among procurement stakeholders (contracting entities, suppliers, officials). The cleaned dataset provides this information through columns such as sd_entidad, sie_ic_proveedor, and related fields documented in SOCE's "GUIA DE NAVEGACION" manual.

- **Document-Based Analysis:** Investigates procurement documentation content and evolution. Researchers can utilize the pipeline's document reconstruction and extraction capabilities, accessing the files located in the file_path column.

- **Risk-Based Analysis:** Studies corruption risk patterns and indicators. The dataset includes several risk indicators for Electronic Reverse Auction (SIE in Spanish) processes (sie_ic_indicador_01 through sie_ic_indicador_27), detailed in the SIE "Fichas Metodológicas" documentation.

Research Question Formulation

After identifying the domain, researchers should formulate specific questions considering:

1. **Data Availability:** Review the pipeline's column documentation to confirm data availability. For data not in the cleaned dataset, consult Kapak's documentation for additional SOCE table mappings.

2. **Temporal Scope:** Consider the time period covered by available data and its relevance to the research question.

3. **Analysis Granularity:** Determine whether the research requires process-level, entity-level, or system-level analysis.

4. **Indicator Relevance:** For risk-related research, consult the methodological files to understand indicator calculations and implications. For example, sie_ic_indicador_04 flags potentially biased language in questions and answers, while sie_ic_indicador_09 addresses competition levels.

Data Requirements Assessment

Once research questions are formulated, assess data requirements against available sources:

1. **Primary Pipeline Data:** Begin with the cleaned dataset, which provides standardized, processed procurement information.

2. **Extended Kapak Data:** For requirements beyond the cleaned dataset, consult Kapak's database schema and SOCE documentation to identify relevant tables and relationships.

3. **Document Content:** Evaluate whether the research requires text extraction from procurement documents, available through the pipeline's document processing capabilities.

4. **Data Quality:** Review data completeness, especially for critical fields identified in the research questions.

Methodology Selection

Select appropriate analytical methods based on data characteristics:

- **Structured Data Analysis:** For numerical and categorical data from the cleaned dataset, particularly useful for risk indicator analysis.

- **Text Analysis:** For document content and communication records, leveraging the pipeline's text extraction capabilities.

- **Network Analysis:** For investigating relationships between procurement actors and processes.

- **Temporal Analysis:** For studying patterns and changes over time using process timestamps and modification records.

Validation Approach

Design appropriate validation strategies:

1. **Statistical Validation:** For quantitative analyses, particularly when working with risk indicators.

2. **Expert Review:** Especially important for qualitative findings or pattern identification.

3. **Cross-Reference:** Compare findings against known cases or established procurement patterns.

4. **Limitations:** Document data gaps or quality issues that might affect conclusions.

This methodology provides a structured approach while maintaining flexibility for diverse research objectives. Researchers should iterate through these steps as needed, refining their approach based on initial findings and data availability. The methodology's effectiveness relies heavily on understanding both the pipeline's capabilities and the underlying procurement system documentation.

Additional Considerations for Implementation:

- **Public Procurement Processes:** Researchers must first acquire a fundamental understanding of public procurement in Ecuador. This includes familiarity with the procurement process flow, which involves multiple stages, key actors, and relevant terminology. Furthermore, it is essential to comprehend the various procurement modalities such as SIE and GEN.

- **Kapak Pipeline Familiarity:** It is crucial for researchers to understand the Kapak pipeline's data processing. In this way, they will gain knowledge of the ready to work data to use in further analysis.

- **Iterative Refinement:** The analytical approach should be iteratively refined throughout the research process. Researchers must be prepared to adjust their methods and strategies as new insights emerge or as limitations in the available data become evident. This iterative approach enhances the robustness and accuracy of the analysis over time.

- **Compliance with Ethical Standards:** Ensuring compliance with ethical standards is paramount in any research endeavor. Researchers must guarantee that data usage adheres to privacy laws and follows established ethical guidelines. This includes ensuring informed consent where applicable and safeguarding sensitive information to maintain the integrity of the research process.

- **Software Environment:** To facilitate the documentation and sharing of the analytical process, researchers should employ appropriate software tools such as Jupyter Notebooks or RStudio. These platforms not only enable effective code execution but also allow for comprehensive documentation. However, Jupyter notebooks are preferred since the project codebase is already in Python.

DATA PIPELINE FOR PROCUREMENT ANALYSIS

The complexities of procurement data, as highlighted in previous sections—particularly in the Kapak database and the SOCE system—often require researchers to dedicate significant effort to data preparation, leaving less time for actual analysis. While ensuring data quality, completeness, and consistency is critical, it is equally important that these preparatory tasks do not eclipse the broader analytical objectives. To address this challenge, a systematic and well-structured data pipeline can effectively balance the demands of data preparation and analysis.

Building a robust data pipeline for handling procurement data entails leveraging tools and methodologies designed for data extraction, transformation, and loading (ETL) processes. In this section, we examine some widely used tools—DBT (Data Build Tool), Apache Spark, and Apache Airflow—and evaluate their suitability for the specific requirements of this data pipeline. Additionally, we discuss the advantages of beginning with a Python-based approach, which provides a flexible foundation for data processing while allowing for future integration with more sophisticated tools as needs evolve.

Tools for Big Data Pipelines

Several tools are commonly used to manage data processing pipelines, each offering distinct features for different stages of ETL:

1. **DBT (Data Build Tool)**: DBT is widely used for data transformations within data warehouses, leveraging SQL-based transformations. It offers features like version control for transformations and incremental data modeling. However, its capabilities are limited to SQL-based operations on data already in the database, making it unsuitable for complex preprocessing tasks such as data extraction, decoding, or decompression.

2. **Apache Spark**: Spark is a powerful distributed data processing methodology that supports large-scale data analytics and parallel data processing. Its features include:
   - **Distributed Computing**: Enables data processing across multiple nodes, handling large datasets efficiently.
   - **DataFrame API**: Facilitates high-level data manipulation.
   - **Parallel Operations**: Supports complex transformations, such as decompression and decryption.
3. **Apache Airflow**: Airflow excels in workflow orchestration, providing features for scheduling and managing complex data pipelines through Directed Acyclic Graphs (DAGs). It allows for custom tasks using Python, as well as built-in monitoring and logging capabilities.

Limitations of DBT and Its Unsuitability for the Current Pipeline

While DBT is a valuable tool for transforming data in data warehouses, its limitations make it unsuitable for the current pipeline's needs:

- **Lack of Support for Complex Data Extraction**: The pipeline requires tasks like making HTTP requests, handling compressed files, and processing XML data, which DBT cannot perform.
- **Data Transformation Limitations**: DBT's SQL-based approach restricts it from performing advanced data processing tasks such as base64 decoding and XML decryption.

Given these limitations, DBT is not a viable option for this project.

Using Apache Spark and Airflow: Benefits and Challenges

While Apache Spark and Airflow offer powerful capabilities for data processing and orchestration, they present certain challenges:

- **Development Overhead**: Implementing a full-fledged pipeline with Spark and Airflow would require significant development effort, including configuring the infrastructure, writing custom code for orchestration, and handling task dependencies.

- **Initial Complexity**: Setting up and optimizing a distributed system from the start may add unnecessary complexity, especially if the current requirements can be met with simpler tools.

Starting with a Python-Based Pipeline

Given the limitations of DBT and the complexities of Spark and Airflow, it is more efficient to initially implement the pipeline in Python. Python provides the flexibility to build the pipeline incrementally while leveraging existing libraries for tasks such as:

- **Data Extraction and Preprocessing**: Libraries like requests, 7zip, and *xml.etree.ElementTree* can handle tasks such as making HTTP requests, decompressing files, and parsing XML data.

- **Data Transformation and Integration**: Python's built-in capabilities allow for custom transformations and data integration with databases like PostgreSQL.

By starting with Python, the pipeline can be built more quickly and tested iteratively. Later, as data processing needs scale, the Python logic can be integrated into a more robust setup with Apache Spark and Airflow, minimizing rework.

PIPELINE OVERVIEW

The data pipeline processes procurement information from the Ecuadorian public procurement system (SOCE), designed to transform raw data into clean, structured datasets for analysis. It follows an Extract, Transform, Load (ETL) process involving data acquisition, file reconstruction, and detailed data extraction to generate analysis-ready datasets.

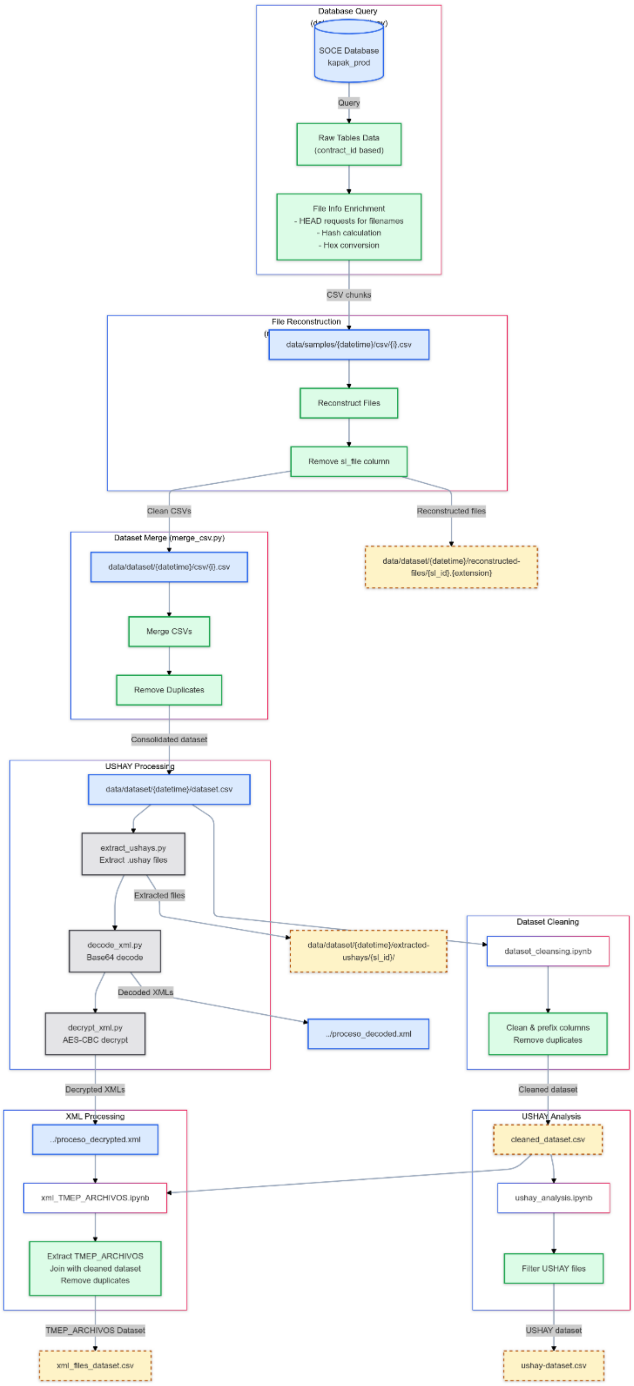


Figure 4: Pipeline Diagram

PIPELINE STAGES

1. **Data Acquisition and Filtering**

The pipeline starts by querying relevant tables in the Kapak database (kapak_prod), focusing on those where contract_id serves as a unique key. Key tables include soce_sie_ic_proceso, soce_descripcion, resumen_de_contrato, preguntas_y_aclaraciones, resultados_de_subasta_o_negociación, soce_fechas, and soce_links. Data is filtered to target records from 2022 but could easily target other years. With particular attention given to potential "ushays" (electronic bid files) and "pliegos" (bidding documents).

2. **File Information Enrichment**

Since the soce_links table lacks filename and extension data, a HEAD request is made to each file's URL (sl_link) to retrieve these details. The content of each file (sl_file) is hashed for unique identification, and its binary content is converted to hexadecimal for storage efficiency. The enriched information is saved in individual CSV chunk files in the data/samples/{datetime}/csv directory.

3. **File Reconstruction and Dataset Consolidation**

CSV chunks are read, and the hexadecimal representation of the file content is converted back to its original binary form to reconstruct the files. The processed CSV chunks are updated by removing the sl_file column to avoid redundancy and saved in the directory data/dataset/{datetime}/reconstructed_files/. All processed chunks are then merged into a consolidated dataset (data/dataset/{datetime}/dataset.csv), with duplicates removed.

4. **Extraction of ‘.ushay‘ Files**

The pipeline identifies ‘.ushay‘ files (electronic bid files) and extracts them using the 7z utility into folders named after their corresponding sl_id. Multiple-part ‘.ushay‘ files are handled individually, with the extracted parts stored in data/dataset/{datetime}/extracted-ushays/{sl_id}. Note that the current implementation does not reassemble multi-part files, potentially resulting in incomplete data.

5. **Dataset Cleaning and Preparation**

The consolidated dataset undergoes cleaning: column names are prefixed based on their source tables (e.g., sl_ for soce_links), and redundant or empty columns are removed. Date columns associated with internal database record creation are excluded, keeping only relevant SOCE-related timestamps. Originally, this dataset consisted of 55 columns: 20 numerical, 1 boolean, and 34 textual. A feature extraction and processing step was applied to enhance the dataset, generating additional, more usable columns. The processed version includes 73 columns: 25 numerical, 33 boolean, and 15 textual. These transformations could improve the dataset’s utility for analysis, enabling richer insights. These cleaned datasets are saved as data/dataset/{datetime}/cleaned-dataset.csv and data/dataset/{datetime}/cleaned-fe-dataset.csv

6. **’ushay’ Dataset Creation**

A subset of the cleaned dataset, focusing specifically on ‘.ushay‘ files, is created. This dataset, stored at data/dataset/{datetime}/ushay-dataset.csv, captures essential procurement process documents, facilitating further analysis.

7. XML files Handling

Each process has an XML associated with it inside the .ushay file or files named proceso.xml. However, the contents of the tags in this XML file are all encoded in base 64 and some tags related to prices or amounts are further encrypted by the SOCE's Modulo facilitador software written in PHP. In specific, these tags are encrypted using the Advanced Encryption Standard (AES) in Cipher Block Chaining (CBC) mode. This section outlines the decoding and decryption process required to make the XML content usable for analysis.

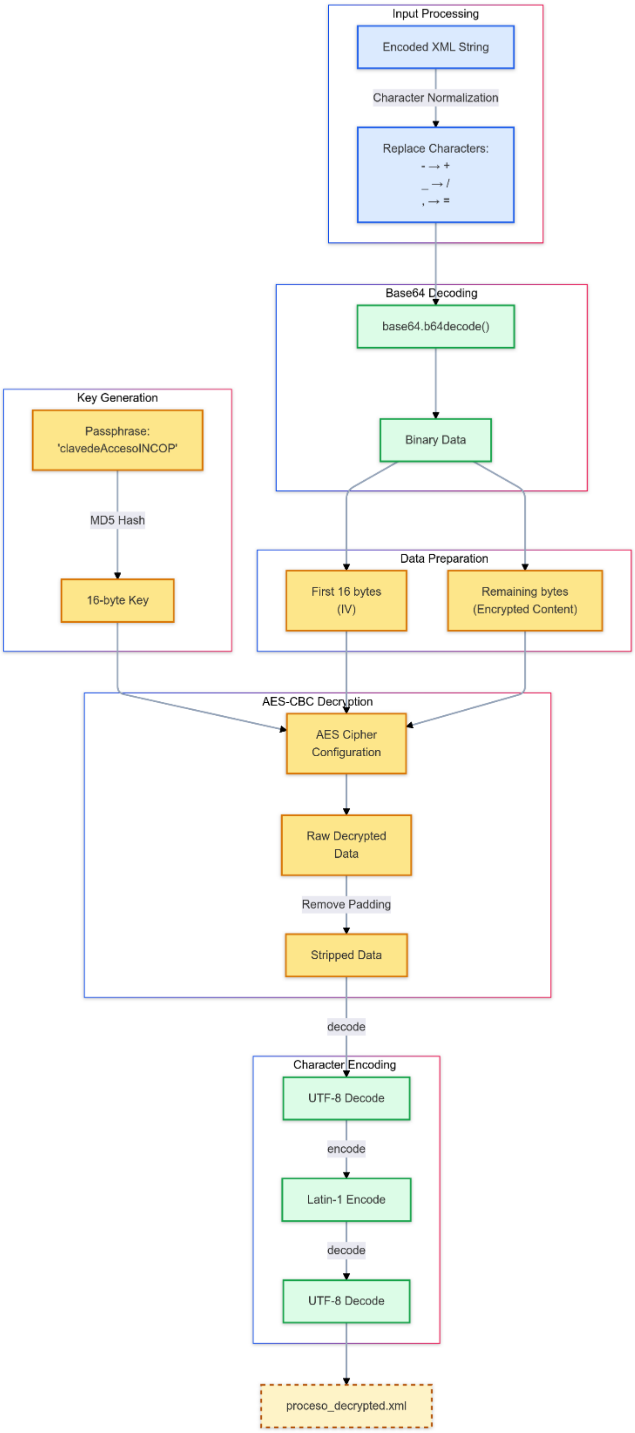


Figure 5: XML's files handling

a. **Base64 Decoding and Character Normalization**: Initially, encoded XML strings utilize non-standard Base64 characters (− , _ , ,) as replacements for standard symbols (+ , / , =). These are reverted, and the content is then decoded from Base64 to extract the original binary data into a new file called proceso_decoded.xml.

b. **AES Decryption Setup**: The decryption key is derived by computing the MD5 hash of the passphrase "clavedeAccesoINCOP", producing a 16-byte key. The first 16 bytes of the decoded data function as the initialization vector (IV), with the subsequent bytes representing the encrypted content.

c. **Decryption Execution**: Using the derived key and IV, the AES algorithm in CBC mode decrypts the ciphertext. Post-decryption, padding is removed to restore the original plaintext.

d. **Handling Character Encoding**: Given differences between PHP-based encryption and Python's decryption environment, an additional encoding normalization step is employed. This involves decoding the decrypted output as UTF-8, re-encoding as Latin-1, and final conversion back to UTF-8 to ensure correct character representation.

The following Python code details the implementation:

```python
import hashlib
from Crypto.Cipher import AES
import base64

def decrypt(plaintext, from_php_encryption=True):
    plaintext = plaintext.strip()
    if not plaintext:
        return plaintext

    # Derive the decryption key
    key = hashlib.md5('clavedeAccesoINCOP'.encode('utf-8')).digest()

    # Revert character replacements and decode from Base64
    plaintext = plaintext.replace('-', '+').replace('_',
'/').replace(',', '=')
```

```
    decoded_data = base64.b64decode(plaintext)

    # Extract the IV and encrypted content
    iv, encrypted_content = decoded_data[:16], decoded_data[16:]

    # Decrypt using AES in CBC mode
    cipher = AES.new(key, AES.MODE_CBC, iv)
    decrypted_data = cipher.decrypt(encrypted_content).rstrip(b'\0')

    # Additional decoding for PHP-encrypted data
    if from_php_encryption:
        decrypted_data = decrypted_data.decode('utf-8',
errors='ignore').encode('latin-1', errors='ignore').decode('utf-8')

    return decrypted_data.strip()
```

This procedure effectively converts the protected XML data into a readable format as a new file called proceso_decrypted.xml.

8. XML-Based Dataset Creation

The pipeline then extracts content from the <TMEP_ARCHIVOS> tag within each ‘.ushay‘ file’s decrypted XML, where individual documents are listed under <Row> sub-tags. Each document is treated as a separate entry in a new dataset, which is then merged with the cleaned dataset using sl_id to add procurement process details. Duplicate files are removed based on hash values, and the resultant dataset enables analysis of associated documents.

GENERATED DATASETS

The pipeline produces several datasets at various stages, each serving distinct analytical purposes:

- **Raw Processed Data**:

  data/samples/{datetime}/csv/{i}.csv

  Contains chunks of processed data with file information and hexadecimal file content, along with contract details.

- **Reconstructed Files**:

  data/dataset/{datetime}/reconstructed-files/{sl_id}.{extension}

  Stores original files, reconstructed from the raw data and organized by sl_id.

- **Consolidated Dataset**:

  data/dataset/{datetime}/dataset.csv

  Represents the merged and deduplicated data from all CSV chunks.

- **Extracted '.ushay' Files**:

  data/dataset/{datetime}/extracted-ushays/{sl_id}

  Contains contents extracted from '.ushay' files, organized by sl_id.

- **Cleaned Dataset**:

  data/dataset/{datetime}/cleaned-dataset.csv

  The primary dataset with cleaned and prefixed columns, emphasizing relevant procurement data for comprehensive analysis.

- **’ushay’ Subset Dataset**:

  data/dataset/{datetime}/ushay-dataset.csv

  Focuses solely on ‘.ushay‘ files, capturing procurement documents necessary for analysis.

- **XML-Based Dataset**:

  Information extracted from the <TMEP_ARCHIVOS> tag in ‘.ushay‘ XML files, linked to procurement processes.

COLUMN OVERVIEW

The **Enhanced Kapak Datasets** includes a range of columns designed to capture detailed information on public procurement processes in Ecuador, enhancing data quality and usability for analysis. The dataset comprises various tables, each with distinct columns representing different aspects of the procurement process, such as contract details, indicators of risk, descriptive information, questions and clarifications, and file metadata. We should note that each column includes the prefix of the origin table for traceability.

soce_links (sl_)

- **sl_id**: A sequential identifier generated for each file as it is added to the soce_links table.
- **sl_contract_id**: Connects the file to its associated contract, linking it with other relevant information in the dataset.
- **sl_link**: The URL pointing to the file's location on the SOCE portal, extracted via the web crawler.
- **sl_datos**: A JSON field that includes details about the user who uploaded the file, such as the company name and upload date.
- **sl_descripcion_del_archivo**: The original file name provided by the user.
- **sl_content**: The text content extracted from the downloaded file using OCR technology (Tesseract library), as part of the web crawling and file download process.
- **sl_flags**: A boolean value initially set to FALSE, indicating that files are assumed not to be malicious. If malware is detected, the value is set to TRUE to flag and potentially exclude the file from future processing.

- **sl_selected**: This boolean value indicates whether the file was selected for inclusion in specific data samples.
- **sl_request_date**: The date when the file was downloaded by the web crawler.

soce_sie_ic_proceso (sie_ic_)

- **sie_ic_anio**: The year the contract process began.
- **sie_ic_mes**: The month the contract process began.
- **sie_ic_fecha_publicacion**: The date on which the contract was made public.
- **sie_ic_estado_del_proceso**: The current status of the contract (e.g., "Finalizado" or "Adjudicado - Registro de Contratos").
- **sie_ic_ruc**: The tax identification number (RUC) of the awarded supplier.
- **sie_ic_proveedor**: The name of the awarded supplier.
- **sie_ic_monto_presupuestado**: The original budget amount for the contract.
- **sie_ic_monto_contrato**: The final contract amount.
- **sie_ic_monto_adjudicacion**: The awarded contract amount.
- **sie_ic_indicador_01 to sie_ic_indicador_27**: A series of indicators designed to capture potential risk factors and irregularities in procurement processes.
  - **sie_ic_indicador_01**: Flags potentially restrictive delivery timeframes. A value of 0 indicates compliance with regulations, while 1 suggests an unusually short delivery period.

- **sie_ic_indicador_04**: Detects potentially accusatory language in questions, with keywords signaling possible manipulation (1), while their absence gives a 0.
- **sie_ic_indicador_05**: Indicates whether formal complaints have been filed (1) or not (0).
- **sie_ic_indicador_06**: Flags modifications in procurement documents (1 for modifications, 0 otherwise).
- **sie_ic_indicador_09**: Reflects the number of bidders, where a lower count may indicate limited competition.
- **sie_ic_indicador_11**: Highlights cases with only one bidder (1), potentially signaling low competition.
- **sie_ic_indicador_15**: Signals whether eligibility verification was properly conducted (1 if not, 0 otherwise).
- **sie_ic_indicador_17**: Identifies potential overpricing by comparing unit prices against historical data.
- **sie_ic_indicador_19**: Checks if the contract was published within the legally mandated timeframe.
- **sie_ic_indicador_22**: Indicates processes where only one qualified bidder led to direct negotiation.
- **sie_ic_indicador_25**: Flags processes declared "unsuccessful" without proper justification.

  - **sie_ic_indicador_26**: Identifies cases with supplementary contracts, which may bypass standard procedures.
  - **sie_ic_indicador_27**: Flags contract amendments, which may indicate unfair alterations.
- **sie_ic_promedio**: The average risk score derived from the individual indicator scores, serving as an overall measure of risk for the contract.

soce_descripcion (sd_)

- **sd_codigo**: The entity-assigned code for the procurement process.
- **sd_entidad**: The name of the contracting entity.
- **sd_autoridades**: A JSON field listing the authorities involved, further detailed in the soce_autoridades table.
- **sd_descripcion**: A general description of the contract, which often duplicates information from sd_objeto_de_proceso.
- **sd_tipo_compra**: The type of purchase, such as "Bien" (Good) or "Servicio" (Service).
- **sd_forma_de_pago**: The payment method specified in the contract.
- **sd_comision_tecnica**: Indicates the presence of a technical commission (TRUE/FALSE). Further details are stored in the soce_miembros_comision_tecnica table.
- **sd_miembros_comision_tecnica**: A JSON field listing the commission members, if applicable.
- **sd_plazo_de_entrega**: The contractually specified delivery timeframe.

- **sd_objeto_de_proceso**: A description of the procurement objective.
- **sd_estado_del_proceso**: The process's status at the time of data collection.
- **sd_vigencia_de_oferta**: The period during which submitted offers remain valid.
- **sd_tipo_de_adjudicacion**: The type of award, typically "Subasta Inversa Electrónica" for reverse auctions.
- **sd_tipo_de_contratacion**: The specific procurement method used.
- **sd_funcionario_encargado_del_proceso**: The official responsible for overseeing the process.
- **sd_presupuesto_referencial_total_sin_iva**: The total budget for the contract, excluding VAT. If unavailable, it may read "No Disponible."
- **sd_variacion_minima_de_la_oferta_durante_la_puja**: Minimum acceptable bid variation during the auction.

resumen_de_contrato (rc_)

- **rc_resumen_de_contrato**: A JSON summary of the contract, detailing key aspects not explicitly listed in other columns.

preguntas_y_aclaraciones (pya_)

- **pya_preguntas_y_aclaraciones**: A JSON field containing questions and clarifications related to the procurement. This data is structured in the soce_preguntas_y_aclaraciones table in the testing environment but not in production.

resultados_de_subasta_o_negociacion (rsn_)

- **rsn_resultados_de_subasta_o_negociacion**: A JSON field summarizing auction or negotiation outcomes.

File Information

- **filename**: The original file name extracted from the URL.
- **extension**: The file extension (e.g., .pdf, .ushay).
- **hash**: A unique hash value of the file content for duplication detection.
- **file_path**: The local storage path of the downloaded file.

This comprehensive column structure supports diverse analytical tasks aimed at assessing procurement practices and identifying potential risks of corruption.

EMPIRICAL VALIDATION

To validate our data pipeline and analytical methodology, we conducted an analysis following the structured approach outlined in the Analytical Methodology section. This section demonstrates how the methodology guides research from initial question formulation through analysis execution.

Domain Identification

We identified two primary domains for our validation study:

- **Document-Based Analysis:** Investigating the relevance and relationships between procurement documents and process characteristics.
- **Risk-Based Analysis:** Examining the predictability of risk levels in procurement processes using process characteristics.

Research Question Formulation

Based on these domains, we formulated specific research questions:

1. **Document Relevance Question:** How relevant are different procurement documents to the process-specific questions and answers? This question aims to identify which documents contain the most pertinent information for understanding procurement specifics.
2. **Risk Classification Question:** Can process characteristics predict the risk category of a procurement process? This question explores whether supervised and unsupervised learning approaches can effectively categorize procurement processes by their risk levels.

Data Requirements Assessment

We assessed our data needs against the pipeline's output:

1. **Primary Data Requirements:**
   - Process identifiers (contract_id, sd_codigo).
   - Document content.
   - Questions and answers (pya_preguntas_y_aclaraciones).
   - Risk indicators (sie_ic_indicador_* columns).
   - Average risk score (sie_ic_promedio).
2. **Data Availability:** All required data was available in the cleaned dataset and associated document files, requiring no additional data from the Kapak database, but only filtering and little processing of JSONs for the Questions and answers.
3. **Data Quality:** We then verified completeness of the data, by ensuring all processes had full data on all 3 generated tables as defined afterwards.
   - Document content extraction.
   - Question-answer records.
   - Risk indicator calculations and process characteristics.

As the foundation of this research lies in the flexible and robust data pipeline developed for processing and integrating public procurement data. The base pipeline played a critical role in simplifying the generation of datasets needed to answer the research questions.

The pipeline's structured workflow allowed for the automatic extraction, transformation, and loading (ETL) of procurement data from diverse SOCE sources, producing a series of standardized datasets. Among these datasets, cleaned_dataset.csv serves as the primary dataset, encompassing essential information drawn from various Kapak project tables. By consistently applying ETL processes, the pipeline ensured high data quality and ease of expansion, enabling rapid experimentation with different analytical techniques.

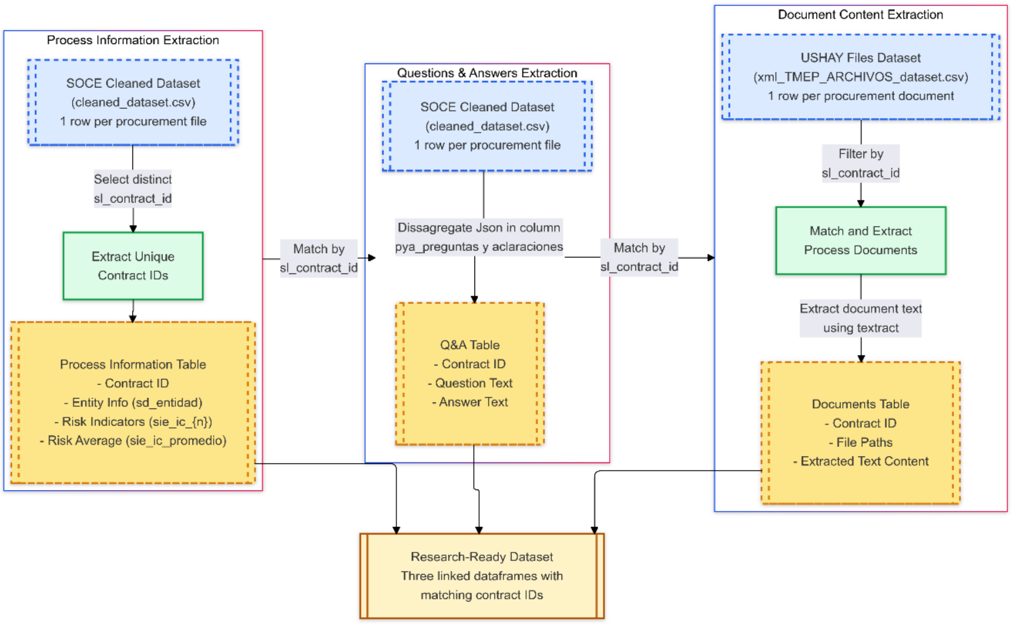


Figure 6: Dataset expansion

Filtering cleaned_dataset.csv *for Document-Level Analysis*

Starting from cleaned_dataset.csv, we created three interconnected datasets:

The **process info table** consolidated essential procurement information for each contract:

- Contract identifiers (sl_contract_id, sd_codigo).
- Descriptive information (sd_entidad, sd_objeto_de_proceso, rc_resumen_de_contrato).
- Risk indicators (indicador columns).
- Average indicator score (sie_ic_promedio).

Questions and answers data was extracted by:

- Selecting one row per contract_id from cleaned_dataset.csv.
- Parsing the JSON in pya_preguntas_y_aclaraciones column.
- Disaggregating into individual rows containing contract_id, question text, and answer text.

File-level details were obtained from xml_TMEP_ARCHIVOS_dataset.csv by:

- Filtering to match contract_ids in process_info_table.
- Retaining metadata like file paths and names from .ushay files.
- Ensuring each document was properly linked to its procurement process.

Extracting and Integrating Textual Content

Textual content extraction was also simplified by the base pipeline's organization of file data. The pipeline reconstructed files from .ushay archives, enabling easy access to documents for text extraction. Using a custom function and the textract library, text was extracted from supported file formats. The integration of this textual data was made efficient by the pipeline's prior standardization of file paths and metadata, allowing for quick merging into the existing datasets.

The decision not to perform text extraction in the base pipeline itself avoided data redundancy while maintaining the original documents. Instead, text extraction was performed only when needed for specific analyses, such as generating TF-IDF vectors for cosine similarity calculations.

Standardizing and Refining the Datasets

The base pipeline's consistent naming conventions and data organization facilitated the final steps of standardizing and refining the datasets for analysis. Column names were aligned, redundant information was removed, and new datasets were saved in a format ready for further processing. This approach not only ensured data consistency but also allowed for quick adaptation of the datasets to different stages of the research.

Methodology Selection

Based on our data characteristics and research questions, we selected the following methods. While there are many alternatives for this kind of analysis, we opted for these approaches due to their interpretability and demonstrated effectiveness. As the focus of this work is on the methodology and pipeline itself, we proceeded with these established methods.

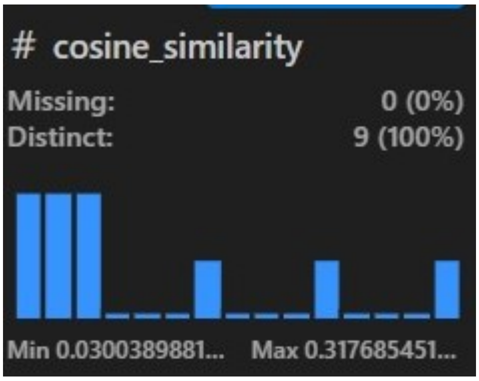
1. **Document Relevance Analysis:**
   - TF-IDF vectorization for document content.
   - Cosine similarity calculation between document vectors and question-answer vectors.
   - Ranking of documents based on similarity scores.
2. **Risk Classification Analysis:**
   - Supervised approach using Gradient Boosting Trees.
   - Unsupervised approach using K-means clustering.
   - Feature engineering from process characteristics.
   - Comparison of classification performance between approaches.

Analysis Execution and Results

It is important to note that the focus of this work lies in the development of the methodology and data pipeline for public procurement analysis, rather than in the detailed exploration of specific research questions. However, to demonstrate the methodology's utility and the pipeline's efficiency in enabling rapid analysis, we present preliminary results from two ongoing research questions being explored within the Kapak project.

Document Relevance Analysis

The results showed that relevant documents tended to have higher ranking (cosine similarity scores) with the set of questions and answers for the processes, especially when compared to templates and generic legal documents. However, this relationship is contingent on the nature of the questions themselves. When questions focus on specific details about goods or services requirements, the similarity scores effectively identify the most relevant technical documents. Showing that within each process most documents are not relevant, with cosine similarity scores below 0.1, while a few ones considered relevant could be used for more thorough analysis, as shown in Figure 7.



cosine_similarity

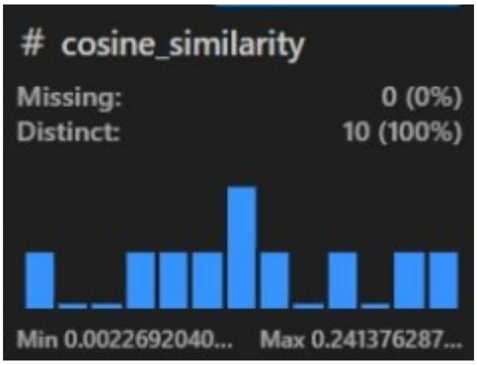
Missing: 0 (0%)

Distinct: 9 (100%)

Min 0.0300389881... Max 0.317685451...

Figure 7: Distribution of document relevance scores for technical requirement questions, showing clear separation between relevant and non-relevant documents

Conversely, when questions primarily address administrative concerns such as timelines or pricing complaints, the distinction becomes less clear. In these cases, even standardized legal templates showed elevated similarity scores, comparable to those of process-specific documents, as evidenced in Figure 8. This suggests that document relevance assessment should consider the question context and may require different approaches for different types of procurement inquiries.



cosine_similarity

Missing: 0 (0%)

Distinct: 10 (100%)

Min 0.0022692040... Max 0.241376287...

Figure 8: Distribution of document relevance scores for administrative questions, showing no clear distinction between document types

Risk Classification Analysis

While the complete analysis of risk classification remains an ongoing research effort within the Kapak project, preliminary results demonstrate the pipeline's capability to support both supervised and unsupervised learning approaches. The Gradient Boosting Trees classifier showed promising discriminative ability across risk categories, particularly for extreme cases, as evidenced by the multiclass ROC curves in Figure 9.

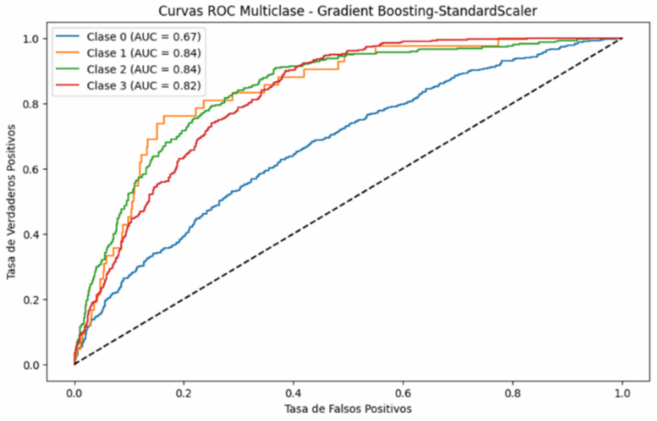


Figure 9: Multiclass ROC curves for Gradient Boosting Trees classifier during training phase

The unsupervised K-means clustering approach revealed interesting patterns that align with the original risk categorizations. Figure 10 shows a t-SNE visualization of the process characteristics, with the top panel colored by original risk categories and the bottom panel showing the K-means clustering results. The comparison suggests particularly effective clustering of medium and very high-risk processes, supported by a silhouette score of 0.85.

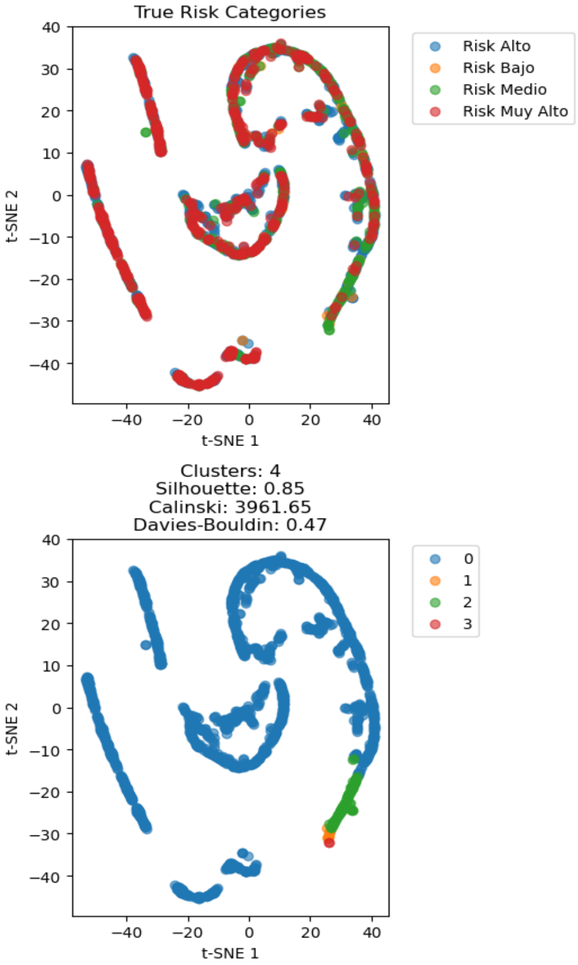


Figure 10: t-SNE visualization comparing original risk categories (top) with K-means clustering results (bottom)

These preliminary results, while not the focus of this work, demonstrate how the developed methodology and pipeline facilitate rapid exploration of research questions in public procurement analysis. The ability to quickly prepare and analyze data using different approaches highlights the practical utility of the pipeline, even as the detailed investigation of these specific research questions continues within the broader Kapak project.

Proposed Validation Approaches

While our current analysis demonstrates the methodology's applicability, several validation approaches could strengthen the findings:

1. **Statistical Validation:**
   - Cross-validation schemes for both supervised and unsupervised approaches.
   - Performance metric selection and evaluation.
   - Statistical significance testing of classification results.
2. **Expert Validation:**
   - Review of identified relevant documents by procurement specialists.
   - Assessment of risk classification results by anti-corruption experts.
   - Verification of findings against known corruption cases.
3. **Methodological Validation:**
   - Comparison of different classification algorithms.
   - Evaluation of feature importance and selection.
   - Testing of different risk category thresholds.
   - Objective comparison between supervised and unsupervised approaches.

Thus, this empirical validation demonstrates how the analytical methodology can guide research from initial question formulation through analysis execution. The results suggest that the methodology effectively supports both document-centric and risk classification analyses while maintaining flexibility for different analytical approaches. Also, the relevance of the developed pipeline is shown in how straightforward it is to prepare the data needed to solve the research questions.

RESULTS AND FUTURE DIRECTIONS

Methodology Validation and Outcomes

Our empirical validation demonstrates the effectiveness of both the analytical methodology and data pipeline in supporting procurement analysis. The methodology successfully guided research from initial question formulation through analysis execution, while the pipeline enabled efficient data preparation and processing. Key outcomes include:

- **Document Analysis Efficiency**: The pipeline's modular architecture facilitated rapid document processing and analysis, with clear separation between relevant and non-relevant documents (cosine similarity scores < 0.1 for non-relevant documents).

- **Risk Assessment Capabilities**: Initial results from both supervised and unsupervised approaches show promise in risk classification, with the Gradient Boosting Trees classifier achieving meaningful discrimination across risk categories and K-means clustering (silhouette score: 0.85) effectively identifying high-risk process patterns.

- **Scalability and Adaptability**: The pipeline's design proved capable of handling diverse data types and analysis requirements, from document reconstruction to text analysis and machine learning applications.

Methodological Contributions

The research makes several key contributions to the field of procurement analysis:

- **Standardized Processing Pipeline**: Development of a reproducible approach to handling complex procurement data, including solutions for file reconstruction, decoding, and dataset consolidation.

- **Analytical Methodology**: Creation of a structured methodology for procurement analysis that balances rigorous academic requirements with practical applicability.

- **Technical Solutions**: Novel approaches to handling USHAY files and encrypted content, providing a foundation for future procurement data analysis efforts.

Future Research Directions

Our research suggests several key areas for future development and enhancement:

Pipeline Enhancement and Integration

- **Kapak Web Crawler Integration**:
  - Implementation of real-time XML decryption and file processing during the crawling phase.
  - Integration of data cleaning and validation procedures directly within the crawler’s ETL process.
  - Development of automated quality checks for incoming procurement data with automated quality scoring mechanisms.
- **Technology Modernization**:
  - Migration to Apache Spark for distributed processing capabilities.
  - Implementation of Apache Airflow for robust workflow orchestration.
  - Design of a horizontally scalable architecture to support future growth.
  - Enhancement of real-time processing capabilities for immediate data availability.

Domain and Analysis Extension

- **Research Methodology Improvements**:
  - Development of standardized question templates for common analysis scenarios.
  - Establishment of feasibility validation criteria for new analysis approaches.
- **Methodological Advancement**:
  - Development of standardized validation approaches.
  - Creation of comprehensive documentation for methodology extensions.
  - Integration of expert validation procedures.

CONCLUSIONS

This research presents a comprehensive methodology for analyzing public procurement data, with particular focus on Ecuador’s SOCE system through the Kapak project. The developed pipeline successfully addresses key challenges in procurement data analysis, including data extraction, preprocessing, and integration. Our empirical validation demonstrates the methodology’s effectiveness in supporting both document-based and risk-based analysis while maintaining flexibility for diverse analytical approaches.

The methodology’s modular design and standardized methodology provide a foundation for future research in procurement analysis. While initial results are promising, particularly in document relevance assessment and risk classification, there remain significant opportunities for enhancement through the integration of advanced processing capabilities and analytical methods.

The primary contribution of this work lies in providing a structured, reproducible approach to procurement data analysis, enabling researchers and oversight bodies to focus on analytical insights rather than data preparation challenges. As public procurement systems continue to generate increasing volumes of complex data, methodologies like this will become increasingly valuable for ensuring transparency and accountability in public spending.

REFERENCES

[1] Y. Torres-Berru and V. F. Lopez Batista, “Data mining to identify anomalies in public procurement rating parameters,” *Electronics*, vol. 10, no. 22, p. 2873, 2021.

[2] E. Ortiz-Prado, R. Fernandez-Naranjo, Y. Torres-Berru, R. Lowe, and I. Torres, “Exceptional prices of medical and other supplies during the COVID-19 pandemic in Ecuador,” *The American Journal of Tropical Medicine and Hygiene*, vol. 105, no. 1, pp. 81–87, 2021.

[3] Dhurandhar, B. Graves, R. Ravi, G. Maniachari, and M. Ettl, “Big data system for analyzing risky procurement entities,” in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1741–1750.

[4] Y. Torres-Berru, V. F. Lopez-Batista, and L. Conde Zhingre, “A data mining approach to detecting bias and favoritism in public procurement,” *Intelligent Automation & Soft Computing*, vol. 36, no. 3, pp. 3501–3516, 2023.

[5] M. Fortuny, E. Guerrero, D. Riofrío, and F. Simon, “Towards smart citizen control in public procurement: Ecuador’s case study,” in *2023 Ninth International Conference on eDemocracy eGovernment (ICEDEG)*, 2023, pp. 1–6.

[6] N. Modrušan, L. Mršić, and K. Rabuzin, “Review of public procurement fraud detection techniques powered by emerging technologies,” *International Journal of Industrial Engineering and Management*, vol. 12, no. 2, 2021.

[7] H. R. F. Araujo, P. F. Leite, J. J. C. M. Honorio, I. M. L. Souza, D. W. Albuquerque, and D. F. S. Santos, “Detection of anomalous proposals in governmental bidding processes: A machine learning-based approach,” in *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2024, pp. 69–70. [Online]. Available: https://tce.pb.gov.br/

[8] M. Ayala-Chauvin, F. Aviles-Castillo, and J. Buele, “Exploring the landscape of data analysis: A review of its application and impact in Ecuador,” *Computers*, vol. 12, no. 7, p. 146, 2023.

[9] J. Lu, Z. Yan, J. Han, and G. Zhang, “Data-driven decision-making (D3M): Methodology, methodology, and directions,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 4, pp. 286–296, 2019.