





Large Language Models in Civic Education on the Supervision and Risk Assessment of Public Works

Joaquim J. C. M. Honório¹^a, Paulo C. O. Brito¹^b, J. Antão B. Moura²^c
and Nazareno F. Andrade²^d

¹Graduate Program in Computer Science, Federal University of Campina Grande (UFCG), Brazil

²Systems and Computing Department, Federal University of Campina Grande (UFCG), Brazil

Keywords: Civic Education, Large Language Models, Machine Learning, Public Works.

Abstract: The Public Administration spends an estimated 13 trillion USD annually worldwide, of which approximately 20% is allocated to public works. Despite strict rules, unfinished works for legal reasons, including corruption, are not atypical, negatively impacting the region's economy, culture, and society. Civic awareness about this problem may help reduce such losses. This study investigates the use of Large Language Models (LLM) and Retrieval-Augmented Generation (RAG) to support civic education on risks in public works. While LLMs interpret and create human language, RAGs combine text production with access to other external data, allowing contextualized responses. Here, we evaluate how these technologies can facilitate the population's understanding of technical information about public works. To this end, we initially create and evaluate 4 Machine Learning models for risk prediction of public work failure, using data from real public works. We provide a failure estimate for each contracted work based on the most efficient model. These data and others related to government development and risk processes are accessed and presented to the user through a web support system. Tests with 35 participants indicate a significant improvement in citizens ability to understand complex aspects related to risks and contracts of public works.


1 INTRODUCTION


In the current situation of increasing complexity of public funded, government managed projects, citizen education becomes very relevant to monitoring and assessing costs, quality, and effectiveness of said projects. Of particular interest here are the so called "public works". The participation of the community as an observer is important to properly monitor project activities, and to ensure that the community's needs are met and that the public administration is transparent and effective (Twizeyimana & Andersson, 2019).


Public Administrations worldwide procure – i.e., contract the acquisition, of – goods, services and works from companies to the yearly tune of an estimated 13 trillion US dollars¹. Corruption and


mismangement of procurement contracts, however, may cause up to 30% of this estimate to be lost, with procurement contracts be responsible for 57% of all bribery cases in the Organization for Economic Cooperation and Development (OECD) countries². These large numbers suggest an important problem in terms of supporting the Public Administration to manage procurement contracts and society at large to monitor more effectively, by forecasting risk of contract failure in special.

Information about public works projects is frequently made available via government transparency portals, public databases, and independent organizations. Aside from particular contract information, it is common to give statistical estimations, frequently collected using Machine Learning (ML) approaches. These estimates, aimed

^a <https://orcid.org/0000-0001-5746-2108>

^b <https://orcid.org/0000-0002-0913-7586>

^c <https://orcid.org/0000-0002-6393-5722>

^d <https://orcid.org/0000-0001-5990-9495>

¹ The Global Value of Public Procurement: <https://spendnetwork.com/13-trillion-the-global-value-of-public-procurement/>

² OCDE - Highlights Reforming Public Procurement: <https://www.oecd.org/gov/public-procurement/public-procurement-progress-report-highlights.pdf>.

at both the general public (e.g.: citizens) and regulatory agencies, can include projections of costs (Bayram & AlJibouri, 2016) (Barros, Marcy, & Carvalho, 2018), contractual additions (Gallego, Rivero, & Martínez, 2021), bid-rigging (Huber & Imhof, 2019), and execution time (Titirla & Aretoulis, 2019), among other things, as well as the risk of project failure owing to legal rulings (Sun & Sales, 2018) (Gallego, Rivero, & Martínez, 2021).

Several studies have investigated applying machine learning to assess risks in public works. However, non-specialists – such as members of the general population – may need help to use these strategies in practice (Yang, Suh, Chen, & Ramos, 2018). While robust, machine learning frequently allows for complicated outcomes their complexity nature requires specialist expertise for interpretation, making citizen participation and comprehension difficult. Chat assistants can help overcome this barrier, improving a lay person’s understanding of technical issues (Pérez, Daradoumis, & Puig, 2020).

Large Language Models (LLM) are extensively trained language models that can comprehend and generate human language, trained with a large volume of textual data (Kasneci, et al., 2023). Retrieval-augmented generation (RAG) is an extension of the models mentioned above that enables the incorporation of the capability to obtain and use information from external source (Lewis, et al., 2020). By being combined with RAG, LLM’s replies may include up-to-date external information beyond the range of the data they were trained in.

In everyday life, LLMs offer support in various applications, facilitating access to information in different areas (Kasneci, et al., 2023). In education, these tools pave the way for the creation of tutors specialized in specific areas. They can be adapted to offer personalized support in different disciplines, for example, to answer multiple-choice questions about code (Jaromir, Arav, Christopher, & Majd, 2023), for medical education (Kung, et al., 2023), to give feedback to students (Wei, et al., 2023), among others. Despite efforts, these models are yet to be applied to the context of public works to support civic education.

Despite advances in research developed around ML in public works management, there are limitations in results obtained so far. Firstly, there are few efforts published in academic circles aimed at predicting risks in public works, most of which are directed at the entire area of government procurement. Furthermore, there is a lack of academic efforts specifically aimed at the practical application of these technologies so that they are

accessible to the non-specialized population. This scenario highlights the need for methods that enable technical information to be made available in a more accessible language, allowing greater understanding and community participation in supervising, and monitoring public works.

This work aims to support civic education by making information about complex topics, such as predictions generated by machine learning algorithms more easily accessible to the non-technical, general public. To achieve this objective, a model for estimating risks in public works is developed based on ML techniques to predict the risk of project failure. Simultaneously, a chat assistant model is developed that utilizes LLM and RAG to transform and contextualize this technical information into a language easily comprehensible and accessible to non-technical citizens. Thus, this article addresses the following research questions that are derived from the above objectives:

- Is it feasible to develop a specialized ML model for assessing and predicting risks in public works projects?
- Are LLM models capable of assisting non-experts in understanding information about public works?
- How does the integration of external information (e.g., processes, machine learning models, data on public works) through RAG affect the accuracy of the information provided?

This article contributes to Civic Education (and also, to applications of Artificial Intelligence – AI), by promoting transparency and citizen participation through simplifying complex information about public works. To date and to the best of our knowledge, this is the first article to:

- Evaluate the performance of LLM and RAG models, including their limitations, in simplifying complex information about public works.
- Provide an overview of the development flow to describe complex information in an accessible manner, contributing to transparency and civic education.
- Present a risk prediction model for estimating a public work’s failure risk.

To facilitate replication of the contents of this article, all used instruments and procedures are made available according to practices of the Open Science Framework (<https://osf.io/7byd9/>) in the external repository.

2 FUNDAMENTALS AND RELATED WORK

2.1 Problem Definition

In the development of a ML model for risk assessment in public works, a detailed set of project characteristics is crucial. Let $P = \{p_1, p_2, \dots, p_n\}$ be the complete set of available projects. A given project, p_i , where $i = 1, 2, \dots, n$, is described by a feature vector \vec{x}_{p_i} that belongs to the risk space R^d , where d represents the variables such as cost, duration, dimension, type of infrastructure, among others.

The main function of the ML model, denoted by M , is to map this feature vector to a risk estimate. Mathematically, this is represented by the function $M: R^d \rightarrow [0,1]$, which generates an estimated probability of failure or a degree of risk, \hat{R}_{p_i} , for each project p_i . The aim of the model is to fine-tune this mapping function to minimize the prediction error, quantified by the sum of squared differences between the observed real risks, R_{p_i} , and the model's estimates, \hat{R}_{p_i} . This goal is expressed by the objective Equation 1.

$$\min_M \sum_{i=1}^n (R_{p_i} - \hat{R}_{p_i})^2 \quad (1)$$

Where n is the total number of projects evaluated. This minimization process is crucial to ensure that the model makes accurate predictions about the risks associated with different infrastructure projects.

An assistant, A , is a chatbot that functions to transform the quantitative outputs of the model M into clear and understandable textual explanations. The model A can be described by the function $A: \hat{R}_{p_i} \times R^d \times \mathcal{D} \rightarrow \mathcal{T}$, where \mathcal{D} is a set of external data and \mathcal{T} is the space of all possible explanatory texts. For each project p_i , A generates an explanatory text T_{p_i} , which is a function of the risk estimates \hat{R}_{p_i} , the project's features \vec{x}_{p_i} , and the contextual data from \mathcal{D} . The efficacy of the chat assistant is evaluated based on the accuracy, clarity, and relevance of the textual explanations T_{p_i} , by presenting a random subset S of selected projects for assessment. Here, S is mathematically defined as a specific collection of elements, chosen from a larger set P of all possible projects, such that $S \subseteq P$.

2.2 Machine Learning Model for Risk Prediction

In recent years, the application of data mining and machine learning (ML) techniques has gained significant attention in improving the accuracy around public procurement. Previous research in public works has focused on estimating cost (Titirla & Aretoulis, 2019), duration of projects (Titirla & Aretoulis, 2019), preventing collusion in bidding (Huber & Imhof, 2019) and predicting risk, inefficiencies (Gallego, Rivero, & Martínez, 2021). In parallel, ML studies have addressed risk estimation and anomaly detection in all types of government procurement, including works (Domingos, Carvalho, Carvalho, & Ramos, 2016) (Ivanov & Nesterov, 2019) (Sun & Sales, 2018).

One widely adopted framework in data mining projects is the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, known for its systematic and structured approach. Studies have showcased its effectiveness in guiding the prediction process and providing actionable insights for decision-makers (Schröer, Kruse, & Gómez, 2021).

The methodology includes the steps illustrated in Figure 1, starting with *Business Understanding*, where relevant data for public works is collected and prepared. The next step, *Data Understanding* defines project objectives and aligns data analysis with these goals. In *Data Preparation*, data is cleaned, and key features are selected for prediction. The *Modelling* step involves selecting machine learning algorithms and splitting the data for training and testing. *Evaluation* assesses model performance using metrics like F1-score and ROC-AUC, with adjustments made as needed. Finally, *Deployment* implements resulting models in real-world scenarios, integrates them into systems.

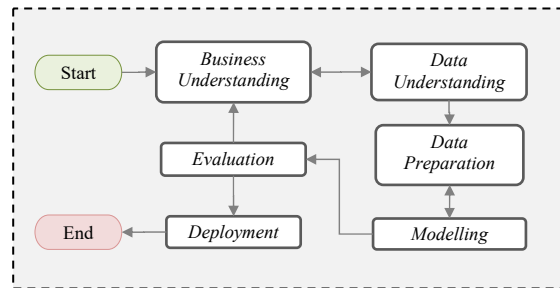


Figure 1: Overview of the CRISP-DM Methodology for Risk Model Development.

Within the *Modelling* step of CRISP-DM, various machine learning algorithms have been utilized to predict public contracting outcomes.

Notably, Random Forest (RF) (Auret & Aldrich, 2012) algorithms have proven to be robust and accurate, capable of handling complex data structures and capturing nonlinear relationships. Artificial neural networks (ANN) (Hond, Asgari, & Jeffery, 2020) with their ability to learn complex patterns, have also shown promise in forecasting contracting outcomes. Additionally, Stochastic Gradient Boosting (SGB) (Bentéjac, Csörgő, & Martínez-Muñoz, 2021), known for its ensemble learning technique, has demonstrated its effectiveness in improving prediction accuracy. Logistic regression (LR), a classical statistical method (Christodoulou, et al., 2019), has been utilized as well, offering interpretability and model transparency.

Evaluation metrics play a vital role in assessing the performance of prediction models. Studies in this field have employed various evaluation metrics to gauge the accuracy and reliability of the models. Here, adopted metrics are accuracy, precision, recall, F1-Score, and area under the ROC curve.

Precision (Equation 2) is a metric that measures the proportion of correctly predicted positive instances out of all instances classified as positive. In the context of public works contract failure prediction, precision indicates the model's ability to accurately identify contracts that are likely to fail. A higher precision suggests that the model has a lower rate of false positive predictions, minimizing the chances of wrongly flagging contracts as failures.

$$P = \frac{\sum TP}{\sum TP + \sum FP} \quad (2)$$

Where true positives (TP) represent the instances correctly identified as contract failures. These are cases where the prediction aligns with the actual outcome, indicating that the model accurately identified a failing contract. True negatives (TN) denote instances correctly identified as successful contracts, where the model correctly predicts the absence of failure. On the other hand, false positives (FP) occur when the model incorrectly predicts a contract failure that succeeds. False negatives (FN) represent cases where the model fails to identify an actual contract failure, incorrectly classifying it as a successful contract.

Recall (Equation 3), also known as sensitivity or true positive rate, evaluates the proportion of correctly predicted positive instances out of all actual positive instances. In the context of contract failure prediction, recall assesses the model's ability to capture and identify all contracts that are failing.

A higher recall indicates that the model can effectively identify a larger portion of failing contracts.

$$R = \frac{\sum TP}{\sum TP + \sum FN} \quad (3)$$

F1-Score (Equation 4) is a composite metric that combines precision and recall into a single value. It considers both false positives and false negatives and provides a balanced measure of the model's performance.

$$F1 = \frac{\sum P * \sum R}{\sum P + \sum R} \quad (4)$$

ROC-AUC is a metric commonly used in binary classification tasks and evaluates the model's ability to discriminate between positive and negative instances at various probability thresholds. It plots the true positive rate against the false positive rate and calculates the area under the resulting curve. A higher ROC-AUC score indicates better discrimination ability and overall performance of the contract failure prediction model.

By employing these metrics, we aim to comprehensively evaluate the performance of the public works contract failure prediction models proposed in this article. The adopted metrics provided a robust set of measurements, capturing different aspects such as false positives, false negatives, overall accuracy, and discriminative ability of the models in predicting contract failures.

2.3 Large Language Models

AI is increasingly evident in various domains and stands out in education. Among the most notable technological advances today is LLM, which emphasizes OpenAI's ChatGPT³. Generative AI encompasses the creation of novel synthetic content for diverse tasks (García-Peñalvo & Vázquez-Ingelmo, 2023). These models, including Generative Pre-trained Transformers (GPT), are based on deep neural networks, which allows them to understand and generate text effectively (Leippold, 2023). GPT-3.5 Turbo⁴, with its 175 billion parameters, has outstanding capacity compared to previous models.

Despite the gains brought through LLM, it presents restrictions. One of the most significant is "hallucination" caused by returning generic or

³ ChatGPT: <https://chat.openai.com/>

⁴ GPT-3.5 Turbo: <https://platform.openai.com/docs/model-index-for-researchers>

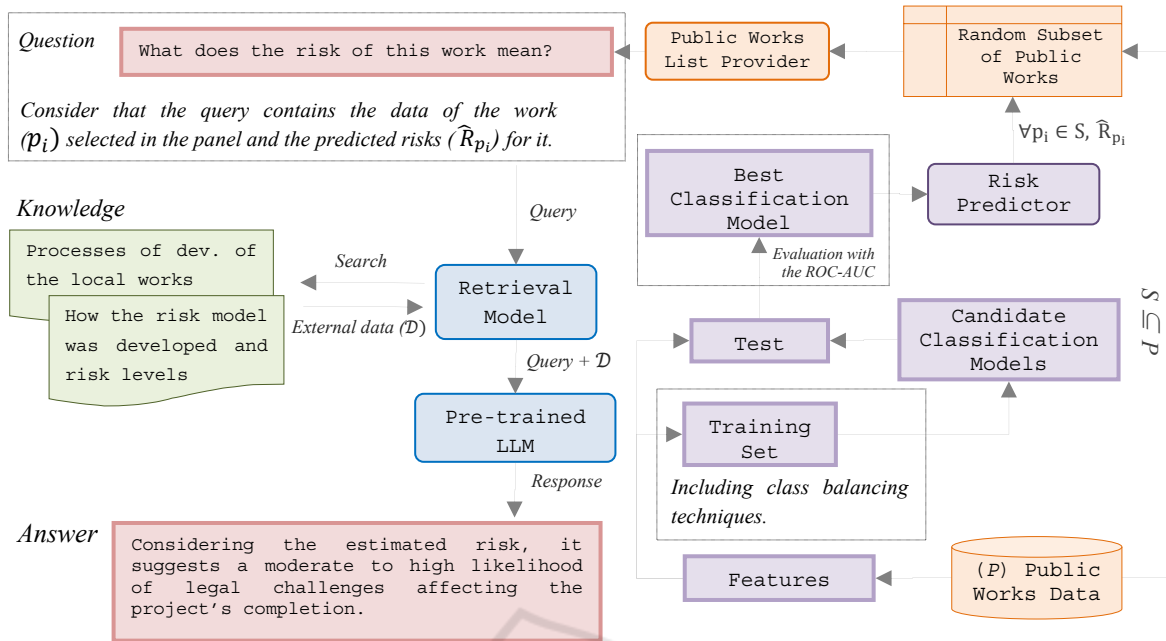


Figure 2: Flowchart depicting the entire process from risk prediction for each project in the study set to the question-and-answer mechanism of the chat assistant.

fictitious information, which can lead to a loss of credibility and trust among users (Yuan, et al., 2023). Furthermore, although advanced, these models still struggle with challenges related to understanding contexts, which can result in responses that do not capture what was asked by the user.

The potential of the LLM in education has been studied for different purposes. Researchers evaluated the effectiveness of GPT-3 in programming tests with code snippet questions (Jaromir, Arav, Christopher, & Majd, 2023) (Savelka, Agarwal, Bogart, Song, & Sakr, 2023). Another study evaluated GPT-3 in medical exams, achieving or approaching approval without specialized training, suggesting its potential in medical education and clinical decisions (Kung, et al., 2023). GPT has also been evaluated in automated educational feedback systems, showing that it generates detailed, coherent responses aligned with instructor assessments, potentially helping to develop students’ learning skills (Wei, et al., 2023). Research has also explored how English learners perceive and use ChatGPT outside of the classroom, finding that LLM’s perceived usefulness has had positive perceptions for its actual use in informal English learning (Guangxiang & Chaojun, 2023). Other researchers have proposed an architectural model aimed at LLM

research, with practical use in education (Gonzalez, 2023).

Despite advances in integrating LLMs into various educational areas, there is a significant gap in the literature in terms of a lack of studies focused on using LLMs to address complex topics in civic education, such as processes and statistics. Thus, this work adds to those in the literature by exploring the effectiveness of these models to enrich teaching and learning on socially relevant topics such as citizens’ awareness of risks of public works projects running afoul of the law, intended quality and costs or of social interests.

3 MATERIALS AND METHODS

3.1 Methodology

In this study, we evaluate the use of LLM and RAG to teach complex concepts in civic education, focusing on risk estimates of public works failure. Initially, we develop and evaluate four different ML models, selecting the most efficient one based on the ROC-AUC metric. The selected ML model is then integrated with the chat assistant to enhance system-generated responses. Figure 2 presents an overview of the structure used.

As part of the methodology, a questionnaire⁵ was applied to measure different aspects of the user's experience with the chat assistant which makes use of the risk estimate produced by the selected ML method. The questionnaire was designed to assess the users' understanding of the concepts presented, the accuracy of the delivered information, the ease of use and the practicality of applying the information in real-world scenarios. The questionnaire also included questions that ranged from demographic information to specific questions about public works concepts - such as risk. Users received the questionnaire digitally after participating in an interactive session with the chat assistant, during which they explored information about public works and risk prediction. This process ensured that users' responses reflected their practical experiences using the system. All interaction between the user and the system was conducted entirely in Portuguese.

We employed a combination of research methods to analyze both quantitative and qualitative descriptive data gathered from evaluation by users. To assess participants perspectives quantitatively we applied techniques to evaluate their responses on five-point Likert scales. Additionally, we conducted a qualitative analysis using content analysis methodology, where writers assessed the feedback. This approach aimed to explore factors that might not be adequately captured by the data gathered from the sample.

Combining quantitative and qualitative approaches gave us a more comprehensive understanding of user perceptions. Although Likert scales helped us measure values, analyzing the final answers provided qualitatively made it possible to interpret the quantitative data more deeply. This combination of quantitative methods improved our research, contributing to understanding how chat assistants can be applied effectively in civic education.

3.2 Data Set

In the development of both the Risk Estimation Model and the Chat Assistant, we utilized data from

Brazilian public works, spanning the period from 2010 to 2019. The selection of this timeframe allows for considerable completeness of the available data during these years. By encompassing a decade of public contracting records, this study ensures a comprehensive and robust analysis of the dynamics and patterns within the Brazilian public administration regarding works procurement. In total, four databases were used:

- **Companies' Registers:** this database, provided by the Federal Revenue Service in Brazil (in Portuguese, Receita Federal⁶), has data about companies' registries, partners, investors, registration status, the national classification of economic activities and other data about legal entities in Brazil. This base is available in a database format, making it difficult to use directly. To solve this difficulty, the database was configured in a local environment.
- **Status of Public Works Contracts:** this database provided by Tramita⁷ consists of a system for internal processing and managing electronic documents/processes. It is possible to access it from the start of procurement until it is archived. For research on public works, the data gathered in Tramita is essential, mainly for collecting information regarding accurate contract resolutions (whether a work contract has been successfully completed or terminated for administrative reasons). Although Tramita is a system that does not require internal credentials to access processes and documents, there is no means to access and download the complete dataset of contract terminations. To circumvent such limitation, contractual data through 2019 was made available by auditors with special access to the tool.
- **Contract Attributes:** through the SAGRES⁸ application it is possible to access information such as budget execution, bidding, administrative procurements, registration information, and personnel payrolls of related jurisdictional units.

⁵ Text Assistant Survey Questionnaire: https://osf.io/7byd9/?view_only=7e2f643daa5b4d09a68b7284e5429159

⁶ Companies in Brazil: <https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica--cnpj>

⁷ System for electronic processes: <https://tramita.tce.pb.gov.br/tramita/pages/main.jsf>.

⁸ Electronic reporting system SAGRES: <https://tce.pb.gov.br/sagres-online>

- **Public Works Attributes:** Through the databases from the Geo-PB⁹ and Painel de Obras¹⁰, it is possible to access information regarding the public works carried out in the state of Paraíba in Brazil. The data in this system gathers information about geo-references, contract number, financial amounts which were paid, type, and built area, among other details of a public work.

During the cleaning process, incorrect, incorrectly formatted, duplicated, and incomplete data were removed from the data set. After cleaning, the data were integrated into one file, resulting in a sample with 643 works.

3.3 Risk Prediction Model

Four supervised machine learning algorithms were evaluated: Artificial Neural Networks (ANN), Random Forest (RF), Logistic Regression, and Stochastic Gradient Boosting (SGB). The training for each was conducted using Scikit-learn¹¹ library in Python, streamlining the creation and adjustment of machine learning models by merely changing specific parameters. The data sets were divided in chronological order to prevent information leakage between training and testing phases. For each model, cross-validation was conducted with five subsets, considering metrics such as Precision, Recall, F1-Score, and ROC-AUC.

In building the ANN, the primary parameters set were size and decay. The size refers to the number of neurons in the hidden layer, with values of 1, 3, and 5, while decay, a regularization parameter, was set at 0, 0.1, and 0.0001. For the RF models the number of predictors was randomly selected for each branch of the trees at each node, with values of 2, 14, and 26. The development of SGB involved a number of trees ranging between 50, 100, and 150, and depth values of 1, 2, and 3.

In this study, 10 attributes from relevant literature were selected, chosen for their importance and applicability to the subject. The selection was based on bibliographic reviews and consultations with experts, aiming to ensure that each attribute contributed significantly to the study. The attributes used in each algorithm were:

- **Total Bids Won:** The total number of bids won by the company (contractor).

- **Total Bids Disputed:** Total number of bids (including those won) the company participated in.
- **Number of Previous Failures:** Total number of failures in the conclusion of the work.
- **Bid Type:** Modality of public procurement process, classified into six types: competitive bidding, invitation, price taking, contest, trading floor, and auction.
- **Number of Activities:** The number of company activities according to the National Classification of Economic Activities (CNAE), indicating the activities performed by the company.
- **Age of the Company (Years):** The time elapsed from the creation of the company until the conclusion of the contract.
- **Pending Submission of Information to Control Agencies:** History of pending registrations, such as lack of monitoring, lack of georeferencing, outdated estimates, among others.
- **Number of Districts Served:** The number of districts in which the contractor has worked.
- **Value of the Work:** Total amount destined to the execution of the work.
- **Duration:** Time estimated for the development of the work.

After defining and evaluating the algorithms' hyperparameters, the performances of the four classification models were assessed and compared using the original data sample and two artificial balancing methods (undersampling and oversampling).

3.4 Chat Assistant

A user interface developed in React¹² (see Figure 3) was implemented to conduct the empirical evaluation of the use of LLM and RAG, which connects to an API built in Flask Python¹³. This API is responsible for the functions: returning responses generated by the LLM/RAG model, carrying out the calculation of risks associated with public works, and providing information relevant to the work that the user is viewing on the interface (see Figure 2, which presents an overview of the architecture used).

⁹ Information's about public works: <https://tce.pb.gov.br/noticias/geo-pb-ferramenta-de-controle-de-obras-e-servicos-de-engenharia>

¹⁰ Information's about public works: <https://paineldeobras.tce.pb.gov.br/>

¹¹ Scikit-learn: <https://scikit-learn.org/stable/>

¹² React: <https://react.dev/learn>

¹³ Flask Python: <https://flask.palletsprojects.com/en/3.0.x/>

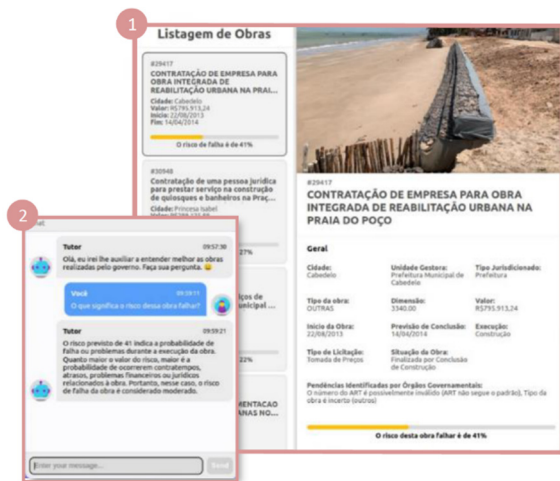


Figure 3: Demonstration of the chat assistant interface presented to the user. (1) data provided to the user regarding the works, including risk estimates; (2) chat that allows the user to ask questions, including about elements presented in item 1.

The chat assistant was built using the LLM algorithm, specifically GPT 3.5 Turbo. We chose this algorithm because it offers a more affordable solution compared to GPT-4 but is still effective in generating coherent responses. To utilize this algorithm, we integrated it with the OpenAI API¹⁴. This integration was implemented to enhance the experience of experiment participants by ensuring faster response times through resources. We relied on the Langchain¹⁵ library to integrate the user interface and specific system functionalities.

For the responses generated by the system to be relevant to the context of public works education, it was necessary to create a prompt. The prompt created, as shown in Figure 4, was designed to direct the way the system processes and answers the questions. As it can be seen, it was developed to act as a tutor on the topic of public works, having detailed information about the work being presented to the user.

A range of external data was made available for the models. The first refers to information about the risk prediction model, including data on the input variables, the methodology used in its construction, and the variables that have the most significant impact on the calculated risk results obtained through the measurement of the importance of each of them through Random Forest. This integration

¹⁴ OpenAI API: <https://platform.openai.com/>

¹⁵ Langchain: https://python.langchain.com/docs/get_started/introduction

was added to ensure that when users inquire about specific aspects of the risk model, the assistant, through RAG, can provide accurate and detailed answers. In addition, information was also made available on local public works development processes, specifically the processes used in Brazil.

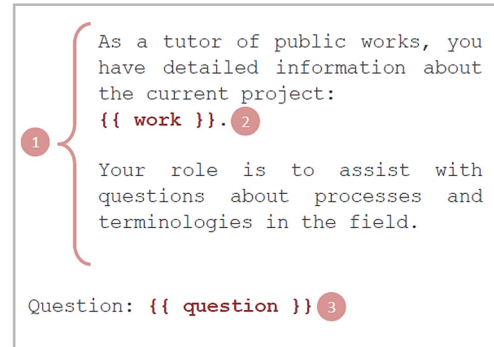


Figure 4: Representation of the prompt used in the chat assistant: (1) Describes the task that an AI should perform; (2) Information about the work p_i and \hat{R}_{p_i} displayed in the interface; (3) User-entered question.

4 RESULTS AND DISCUSSION

4.1 Risk Prediction Model

In a first experiment, the data used for training the classification models were not balanced, so the contracts considered as high risk corresponded to 37% of the data.

The Random Forest, Stochastic Gradient Boosting, and Neural Networks models were the least effective in distinguishing between the classes, having a ROC-AUC close to or equal to 0.5. The small amount of available data is considered to have compromised the learning process of the models due to the attempt to minimize classification errors, resulting in classifying all being low risk.

After class balancing through the undersampling technique on the training set, the resulting sample had 192 data items, 96 for each class. It can be seen from Table 1 that, in this second experiment, the most effective model, with a ROC-AUC of 0.64, was Random Forest, correctly classifying 19 contracts as high risk and with 9 false positives. Despite the Random Forest model having a higher ROC-AUC, the Neural Networks model correctly classified 24 contracts as high risk, with only 4 false positives. Finally, Logistic Regression was the model with the worst performance among the others, where the classifier labelled all contracts as high

Table 1: Comparison of machine learning model performances with different data balancing methods. The Random Forest model under Oversampling, highlighted for its top ROC-AUC score, is the best performer.

Balancing method	Model	Precision	Recall	F1-Score	ROC-AUC
Original Data Sample	RF	-	1	0	0.5
	ANN	0.32	0.24	0.27	0.54
	LR	0.29	0.19	0.22	0.66
	SGB	-	1	0	0.5
Undersampling	RF	0.68	0.38	0.48	0.64
	ANN	0.86	0.27	0.41	0.61
	LR	0.26	1	0.41	0.5
	SGB	0.33	0.65	0.44	0.62
Oversampling	RF	0.42	0.63	0.51	0.75
	ANN	0.71	0.39	0.50	0.66
	LR	0.64	0.39	0.48	0.65
	SGB	0.47	0.39	0.43	0.67

risk. It is believed that this performance was due to insufficient data from the balancing technique used.

In the third and last experiment, class balancing was performed through the oversampling technique applied to the training set, resulting in a sample with 1094 data items, 547 for each class. Finally, as with balancing through oversampling, the Logistic Regression was the model with the worst performance when the ROC-AUC metric was observed (see Table 1). Despite the low efficiency when compared to the other models, the model correctly classified 18 high-risk contracts, but just like the Neural Networks model, it obtained a high value of false negatives compared to Random Forest.

Among the three experimental scenarios observed for the risk modelling of public works procurement, the classifiers that best differentiated the high and low-risk classes were, based on the ROC-AUC metric, Random Forest with Oversampling technique, Random Forest with undersampling technique, and Logistic Regression through the original sample without any balancing treatment.

The oversampling technique, aimed at increasing the number of records in the class with lower frequency until the base is balanced, together with Random Forest, was the treatment with the highest ROC-AUC among the others, having a confidence interval ranging from 0.64 to 0.87 (95% CI). Next, the Logistic Regression algorithm proved to have better results when used with the original sample (95% CI: 0.46 - 0.81). Finally, the undersampling technique, also with the Random Forest algorithm, was the approach with the lowest results with a ROC-AUC of 0.64 (95% CI: 0.57 - 0.72). All the results presented can be seen in Figure 5.

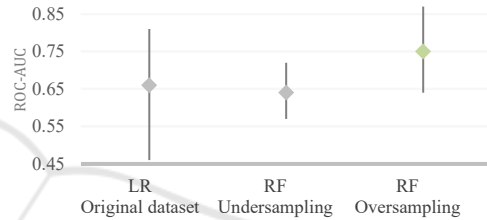


Figure 5: Comparison between the algorithms with the best results, and their data balancing techniques.

It is possible to use machine learning-based classifiers to reveal risks in public works, considering that the ROC-AUC values are higher than the threshold value of 0.5 (random classifier). Despite such threshold, as there is an overlap between the confidence intervals, there is no significant difference between the approaches used to indicate the most effective approach. In the best-case scenario, Random Forest with Oversampling will provide better results than the others. In contrast, in the worst-case scenario, the Logistic Regression with the original balancing may be less effective than the other approaches and, in the best case, superior to Random Forest with undersampling. By measuring the importance of each of the Random Forest with oversampling input variables, it was possible to identify that the 5 most important variables for the result were: value of the work, dimension, number of activities of the contracted company, pending issues detected by government agencies and estimated duration.

The above observations come with limitations. Generalizing our findings faces the threat of data poor representativeness, as public data we used may not reflect the full range of projects or adequately capture the diversity of contexts and conditions in other countries.

4.2 Chat Assistant

The research used questionnaires to evaluate LLM within the scope of civic education and supervision of public works. Quantitative and qualitative analyses were based on responses to a questionnaire administered to 35 individuals.

The average age of the individuals was 38 years old with ages ranging from 19 to 54 years. In terms of gender distribution there were 19 males, 14 females and 2 participants who identified as other. When it comes to education most participants (16) had completed high school while 14 had a university undergraduate degree and 5 held a postgraduate degree. This indicates that the sample used in this study exhibits a range of educational levels. The inclusion of ages and educational backgrounds showcases the diversity, within this participant group.

We also included some questions to gauge participants understanding of AI. The results showed that 20 participants had no knowledge about it; 10 had a basic understanding; 4 had moderate knowledge; and only one demonstrated advanced knowledge. Alongside AI knowledge we also assessed their familiarity with public works. Most participants categorized their knowledge of works as basic (18) while 16 indicated they had no knowledge and only one said it had “advanced knowledge”.

Figure 6 presents users’ experiences with the chat assistant, evaluated using a 5-level Likert scale. The graph highlights that 21 participants strongly agree with the clarity and ease of understanding of the chat assistant’s responses. Ease of use also scored highly, with 34 participants agreeing or strongly agreeing with its ease of use. For understanding public works, 32 participants gave ratings that ranged from neutral to strongly agree. As can be seen, only a tiny portion of the responses

demonstrate disagreement, which suggests strong user approval of the chat assistant’s features.

A better understanding of users’ responses was possible through the joint analysis of quantitative (Likert questions) and qualitative (open questions) approaches. This analysis made it possible to identify that participants with no or basic knowledge of AI tended to focus on general questions about public works or superficial questions about risk. In contrast, participants with more advanced or moderate knowledge of AI used external data through RAG more effectively, especially to understand complex processes such as risk estimation in public works. This group demonstrated to take full advantage of RAG’s advanced capabilities, leveraging its potential for more complex and detailed tasks (e.g., how the models were built, algorithms used, and the importance of each variable).

It was also found that some users rated the clarity and precision of the system’s responses as low (strongly disagree or neutral). This classification occurred especially in situations where they were looking for unavailable external information, such as details about tenders that a specific company had already participated in. The lack of availability of this data in the databases consulted by RAG led to the provision of more generic responses, affecting users’ perception of satisfaction with the system. These same participants highlighted that it would be indifferent for them to implement such functionality in government systems.

Analysis of the responses demonstrates that the RAG and the LLM have distinct but complementary roles in disseminating information and civic education. RAG, with its ability to access and integrate information from external databases, proved to be particularly efficient on two fronts: the first was to provide details regarding simple information presented to the user during the iteration

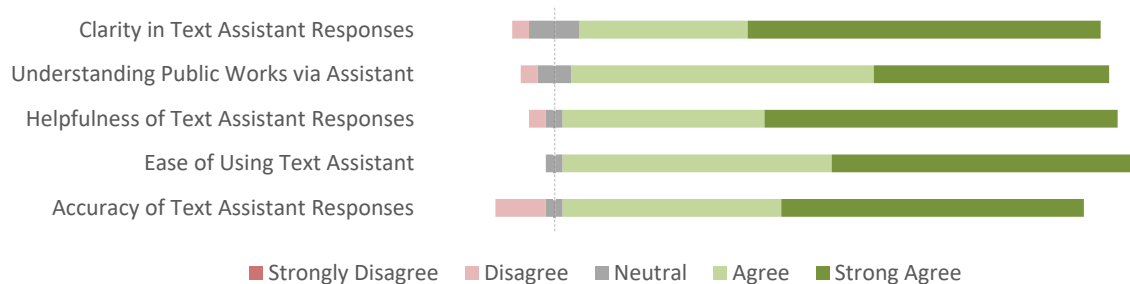


Figure 6: Graph depicting participant ratings of the chat assistant on five aspects: response accuracy, ease of use, helpfulness, understanding of public works, and response clarity, using a 5-point Likert scale from 'Strongly Disagree' to 'Strongly Agree'.

with the public work, such as the value of work and the type of bidding. The second regarded resolving more detailed questions with complex subjects, such as risks estimation. Despite these benefits, the effectiveness of the RAG is limited by the availability and scope of information in the databases to which it has access. The lack of specific information, such as details about companies involved in public works projects, resulted in less satisfactory responses for some users.

Finally, it was observed that, using LLM in conjunction with the developed prompt, it was possible to transform technical and complex information into understandable explanations. By presenting specialized concepts in a more accessible way, it was possible to contribute to easing the comprehension of public works. Such a contribution can improve civic education and promote greater understanding among the population regarding the processes inherent to the management and supervision of public works projects.

5 CONCLUSIONS

Considering the inherent complexities surrounding purchases and contracts made by government agencies, civic education becomes essential to enable an understanding of how government resources are being spent. Based on this, the investigation reported in this article focused on developing and evaluating the use of LLM and RAG to simplify the presentation of concepts related to public works and their contracting process and risk estimation carried out through AI's machine learning (ML). The goal was to make these advanced ideas easily understandable to citizens and encourage their informed involvement.

This study initially focused on developing a predictive model for estimating the risk of public work failures, using ML methods. Based on ROC-AUC values, the Random Forest algorithm coupled with the Oversampling technique for data balancing, was the most efficient approach among the 4 evaluated, evidenced by the model's ability to differentiate between high and low risk contracts.

In addition to the risk prediction model, this study also evaluated the applicability of LLM and RAG to civic education in the context of public works projects. To carry out this assessment, questionnaires were applied to people of different ages and educational levels. Through quantitative and qualitative analysis of the data, it was possible to verify that: (i) there was a strong approval for the

use of chat assistant resources, where most participants reported that this resource could be available in government tools; (ii) participants with less knowledge of AI tended to focus on general questions about public works, while those with moderate or advanced knowledge more effectively used external data to understand risk estimation processes; and (iii) in specific cases, the lack of external data affects user satisfaction due to the generation of generic responses.

Based on these findings, it is possible to conclude that it is feasible to develop risk predictors to predict whether public work will be halted due to court decisions. To this end, the Random Forest algorithm proved to be the most efficient among the algorithms analyzed. Furthermore, LLM models may also help understand complex information related to the risk prediction model and public works. Finally, integrating external information through RAG enabled the user to have accurate answers on different topics. However, it is necessary to make it clear to target users to avoid searches involving non-existent data, causing inaccurate answers.

Thus, this study emphasizes the significance of AI in educational methods and fostering greater civic engagement and understanding of the complexities associated with the supervision and execution of public works.

ACKNOWLEDGEMENTS

This work was carried out with support from the National Council for Scientific and Technological Development (CNPq), Brazil. We extend our gratitude to Jair Guedes for prior discussions and support on LLMs, and to the anonymous reviewers for their constructive criticism.

REFERENCES

- Auret, L., & Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, 35, (pp. 27-42). doi:<https://doi.org/10.1016/j.mineng.2012.05.008>
- Barros, L. B., Marcy, M., & Carvalho, M. T. (2018). Construction cost estimation of Brazilian highways using artificial neural networks. *International Journal of Structural and Civil Engineering Research*, pp. 283-289.
- Bayram, S., & A. S. (2016). Efficacy of estimation methods in forecasting building projects' costs.

- Journal of construction engineering and management* v. 142, p. 142.
- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54, (pp. 1937-1967). doi:<https://doi.org/10.1007/s10462-020-09896-5>
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110, pp. 12-22.
- Domingos, S. L., Carvalho, R. N., Carvalho, R. S., & Ramos, G. N. (2016). Identifying it purchases anomalies in the Brazilian Government Procurement System using deep learning. *15th IEEE International Conference on Machine Learning and Applications (ICMLA)* doi: 10.1109/ICMLA.2016.0129, (pp. 722-727).
- Gallego, J., Rivero, G., & Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting* 37.1, pp. 360-377.
- García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI.
- Gonzalez, B. (2023). Smart Surveys: An Automatic Survey Generation and Analysis Tool. In *Proceedings of the 15th International Conference on Computer Supported Education - Volume 2: CSEDU*, (pp. 113-119). doi:10.5220/0011985400003470
- Guangxiang, L., & Chaojun, M. (2023). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching*, (pp. 1-14). doi:10.1080/17501229.2023.2240316
- Hond, D., Asgari, H., & Jeffery, D. (2020). Verifying Artificial Neural Network Classifier Performance Using Dataset Dissimilarity Measures. *19th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 115-121). IEEE. doi:10.1109/ICMLA51294.2020.00027
- Huber, M., & Imhof, D. (2019). Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization* 65, pp. 277-301.
- Ivanov, D., & Nesterov, A. (2019). Identifying bid leakage in procurement auctions: Machine learning approach. *Proceedings of the 2019 ACM Conference on Economics and Computation.*, (pp. 69-70). doi:<https://doi.org/10.1145/3328526.3329642>
- Jaromir, S., Arav, A., Christopher, B., & Majd, S. (2023). Large Language Models (GPT) Struggle to Answer Multiple-Choice Questions About Code. *International Conference on Computer Supported Education*.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, (p. 103).
- Kung, T. H., Cheatham, M., A., M., Sillos, C., De Leon, L., Elepaño, C., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2).
- Leippold, M. (2023). Thus spoke GPT-3: Interviewing a large-language model on climate finance. *Finance Research Letters*. doi:<https://doi.org/10.1016/j.frl.2022.103617>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, (pp. 9459-9474).
- Pérez, J. Q., Daradoumis, T., & Puig, J. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education* 28.6, (pp. 1549-1565).
- Savelka, J., Agarwal, A., Bogart, C., Song, Y., & Sakr, M. (2023). Can generative pre-trained transformers (gpt) pass assessments in higher education programming courses? In *Proceedings of the 28th Annual ACM Conference on Innovation and Technology in Computer Science Education*, (pp. 117-123). doi:<https://doi.org/10.1145/3587102.3588792>
- Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, (pp. 526-534). doi:<https://doi.org/10.1016/j.procs.2021.01.199>
- Sun, T., & Sales, L. J. (2018). Predicting public procurement irregularity: An application of neural networks. *Journal of Emerging Technologies in Accounting* 15.1, pp. 141-154.
- Titirla, M., & Aretoulis, G. (2019). Neural network models for actual duration of Greek highway projects. *Journal of Engineering, Design and Technology* 17.6, pp. 1323-1339.
- Twizeyimana, J. D., & Andersson, A. (2019). The public value of E-Government—A literature review. *Government information quarterly*, pp. 167-178.
- Wei, D., Jionghao, L., Hua, J., Tongguang, L., Yi-Shan, T., Dragan, G., & Guanliang, C. (2023). Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. *IEEE International Conference on Advanced Learning Technologies (ICALT)* doi: 10.1109/ICALT58122.2023.00100.
- Yang, Q., Suh, J., Chen, N. C., & Ramos, G. (2018). Grounding interactive machine learning tool design in how non-experts actually build models., (pp. 573-584).
- Yuyan, C., Qiang, F., Yichen, Y., Zhihao, W., Ge, F., Dayiheng, L., . . . Yanghua, X. (2023). Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, (pp. 245-255). doi:<https://doi.org/10.1145/3583780.3614905>.