SN Computer Science (2024) 5:87
https://doi.org/10.1007/s42979-023-02443-3

ORIGINAL RESEARCH

Identifying Public Tenders of Interest Using Classification Models: A Comparative Analysis

Yeersainth Figueroa-Gómez[1,2] · Ixent Galpin[2]

Received: 25 September 2023 / Accepted: 19 October 2023

Abstract

Public procurement processes related to tenders represent one of the most important financing strategies for companies in diverse sectors, as they present the opportunity to offer services, supplies and product sales to state entities. Given the high volume of public tenders in the Colombian Government SECOP II database, manually identifying tenders of interest can be a cumbersome and time-consuming process. In this work, we propose automating the identification of interesting tenders by training a supervised classification model. We manually label a sample of tenders published in the National Government open data platform, according to whether or not they are of interest to the Minuto de Dios University, and use them for model training. Several models are evaluated in order to select the best model for deployment, taking into account various metrics to determine best performance according to business needs. The best model is selected based on different analyses, comparing the application of data balancing techniques, the performance of the proposed models, and hyper-parameter settings measured against the test data.

Keywords Classification model · Tenders · Public procurement · Data balancing

Introduction

One of the main challenges faced by many entities is related to obtaining financial resources. An attractive option is to offer services or products aimed at entities in the public sector. The public sector is made up of state entities that contract their different requirements through public tenders. In Colombia, the National Public Procurement Agency–*Colombia Compra Eficiente* (ANCP–CCE), is the entity in charge of managing public bidding processes in Colombia through the SECOP II database. Its purported aims are to ensure greater efficiency, transparency and optimization of state resources [3]. SECOP II is a transactional virtual platform that allows both suppliers and bidders to carry out the contracting process on this platform [4].

Minuto de Dios University in Colombia is one, among many entities, that seeks to obtain state resources through public tenders. This entity must carry out an identification and validation process at the time of submitting the bidding offers that are published in the Colombia Compra Eficiente SECOP II platform. Carrying out this process entails significant effort due to the fact that it is a manual and repetitive process for each bid, which implies investing time and financial resources when identifying tenders of interest. Data related to public tenders is an important source of open data information, which generates a great opportunity for the development of analytics-oriented models. In Colombia, advances and research in the field of public procurement have been limited, and have focused mainly on descriptive data analysis. As such, it is postulated that venturing into the development of machine learning models presents a potentially efficient solution for the needs such entities.

There are a few private companies that offer services through consultation platforms to search for tenders. The

This article is part of the topical collection "Emerging Technologies in Applied Informatics" guest edited by Hector Florez and Marcelo Leon.

✉ Yeersainth Figueroa-Gómez
yeersainth.figueroa@uniminuto.edu

Ixent Galpin
ixent@utadeo.edu.co

[1] Parque Científico de Innovación Social-PCIS, Corporación Universitaria Minuto de Dios, Bogotá, Colombia

[2] Facultad de Ciencias Naturales e Ingeniería, Universidad de Bogotá-Jorge Tadeo Lozano, Bogotá, Colombia

Published online: 12 December 2023

most recognized companies are Licitaciones colombia.co[1] and Licitaciones al día,[2] which identify public financing opportunities. These provide tools that allow the identification of both public and private financing opportunities, using simple filters such as theme, geographic location and type of financing. However, none of these companies offer customized, bespoke models that adjust to the specific requirements of each entity. As such, these approaches do not significantly simplify the operational process for the search of tenders.

In this work, a comparative analysis between different classification models is carried out with the objective of identifying the best model that can classify tenders of interest for the Minuto de Dios University. In this analysis, five classification techniques are evaluated: (1) Logistic Regression, (2) Decision Trees (3) Random Forest, (4) Gradient Boosting, and (5) AdaBoost. For the data set analyzed, three different data balancing techniques are applied to determine the behavior of the models according to each case. The techniques used in this study are: (a) Random Under-Sampling, (b) NearMiss, and (c) Combined Advance Resampling SMOTE-Tomek Balancing.

The structure of this paper is as follows: "Related work" section presents a survey of work related to public procurement and identifying tenders of interest. The subsequent sections broadly follow the steps in the widely used CRISP-DM methodology [19]. "Business Understanding" section describes business needs, and "Data Understanding" section describes the underlying dataset obtained from the SECOP-II platform. "Data preparation" section focuses on the steps carried out for data preparation. The modeling process is described in "Modeling", and "Evaluation" sections presents the evaluation and results of the selected models to enable the selection of the best performing classification algorithm for deployment. "Discussion" section presents the discussion and detailed analysis of the results obtained, and finally "Conclusions" section concludes.

Related Work

The use of machine learning models for bidding was proposed over two decades ago. For example, in 2002 Skitmore et al. [17] presented a probability model to predict an economic proposal lower than that of the competitors from the estimation of probable values where the bidder did not participate. Loza et al. [13], in 2013, propose an approach that aims to estimate the number of bidders of tenders that are identified in public contracts, especially related to the European Commission procurement processes. Goswami et al. [6] propose an approach to classify documents associated with tenders from the Defence Research and Development Organisation website, applying the Naïve Bayes classification algorithm and using cross-fold validation.

Other relevant research is related to the identification of corruption risks in tenders. For example, the research conducted by Fazekas et al. [9] as part of the Digiwhist project proposes a complex tool which manages indicators and public data that works in collaboration with active civil society organizations, generating corruption alerts.

Previous work carried out in Colombia is relatively scant: Alvares et al. [7] propose a Big Data reference model which aims to manage tenders in Colombia through the use of tools such as Apache Hadoop and Apache Hive. This is highlighted by the Inter-American Public Procurement Network and the National Public Procurement Agency—*Colombia Compra Eficiente* [15], who indicate the need to strengthen the use of data and disruptive technologies because there is a low development of research and applications related to the field of public procurement.

Business Understanding

Public tenders are a contracting process whereby a state entity makes an open call, the purpose of which is to establish certain rules for interested parties to submit their tenders and award the most appropriate and best evaluated proposal [10]. In Colombia, this process is carried out through the SECOP II platform where the tenders that various public entities share to select suppliers that will carry out services, acquisitions or products are published and managed. At Minuto de Dios University, it is necessary to carry out a search and validation process to determine whether a particular bid is *interesting*, by which is meant that it is tailored specifically for opportunities that are aligned with the work themes that the university carries out. This implies investing considerable time and allocating human talent resources for the development of this process. The overall average number of tenders published is in the range of 30–40 tenders per day for each type of contract, however this value is relative, because it depends on other factors such as government policies and regulations. The large number of tenders that are published daily has as a consequence for the entity missing a considerable number of tenders that may be a funding opportunity.

In the process for the identification of tenders by an automated classification model, errors may occur. This most serious type of error occurs when a bid is classified as not interesting, but it actually turns out to be a bid of interest to the entity. This would represent the loss of an opportunity to acquire resources, and is referred to as a *false negative* in this paper. Another type of error is when a bid is identified

[1] https://www.licitacionescolombia.co/.

[2] https://www.licitacionesaldia.com/.

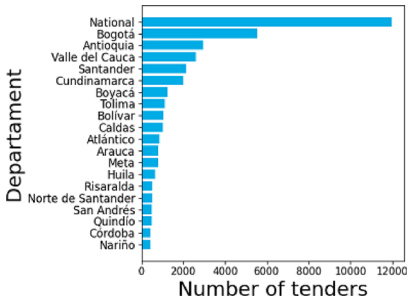

Fig. 1 Number of tenders by region in Colombia

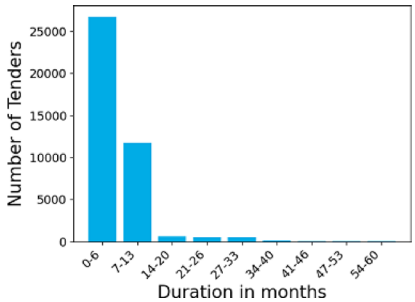

Fig. 2 Tender duration per month

as being of interest, when in fact it is not. Such an error results in time being wasted by employees of the entity, and is referred to as a *false positive* within the evaluation of a classification model. We note that such an error may be mitigated by being a little more specific in the identification of the tenders.

Data Understanding

The data source comes from the national open data platform of Colombia, which is open access,[3] where the bidding processes are updated daily and the history of different years is maintained. The dataset extracted for this work spans the period from 2018 to 2022, due to the fact that during this time the Minuto de Dios University began to carry out bidding processes.

The dataset is comprised of a total of 40,739 records and 59 attributes, where the following types of data were identified: 39 text type attributes, 13 numeric type attributes and 7 decimal numeric type attributes. A new attribute is added to this dataset, which labels the bidding records in two classes, in which 1 means *interesting* and 0 means *not interesting*. This is the output variable of the classification models to be developed.

In the exploratory analysis of the data, the key attributes to understand the status of the tenders were presented. Figure 1 presents a histogram showing the number of tenders per region in Colombia. As can be observed in the figure, by far the highest number of tenders are issued at a national level, with the Capital District of Bogotà, the capital, taking second place. Figure 2 presents a histogram showing the number of tenders according to their duration. As can be seen, most tenders are in the 0–6 month range.

Figures 3 and 4 show, respectively, the number of tenders by contracting modality as well as by type of contract,

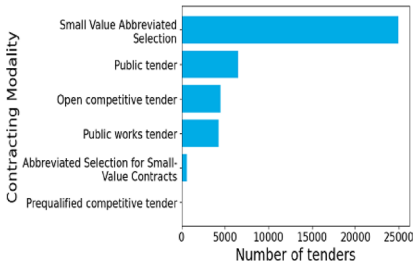

Fig. 3 Number of tenders by contracting modality

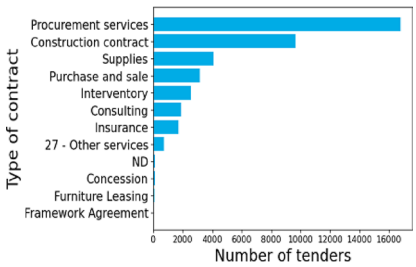

Fig. 4 Number of tenders by type of contracting

allowing us to identify under these two attributes where the largest number of tenders are classified.

The contracting modality refers to a specific procedure used to select the candidate in a public tenders taking into account a series of requirements. The type of contract refers to the specific form in which an agreement is established between the public entity and the selected contractor.

Data Preparation

From the exploration of the data set, it was possible to identify that most of the attributes are of textual, so they are attributes that cannot be used directly for model training. Table 1 shows the attributes deemed to be of interest as a result of business needs. In this case, they involve

[3] https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Procesos-de-Contrataci-n/p6dx-8zbt.

Table 1 Attribute selection for each tender

Attribute name	Attribute description	Data type
NIT	Unique ID of public entity	Integer
Invited_supplier_count	Number of invited suppliers	Integer
Suppliers_Expressed_Interest	Number suppliers that expressed interest	Integer
Tenders_Interest_count	Number tenders interest	Integer
Tender_month_duration	Number tender month duration	Real
Base_Price	Base price of tenders	Real
Department_Entity	Name department entity	Text
Recruitment_modality	Name recruitment modality	Text
Recruitment_type	Name recruitment type	Text
Procedure_Description	Tender description	Text

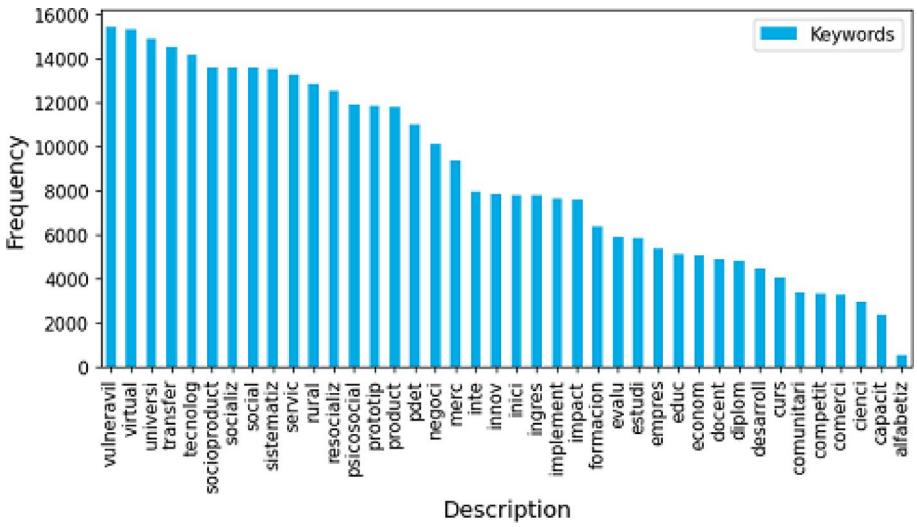

Fig. 5 Keyword selection

understanding of university internal processes when taking into account the identification of tenders that has been done previously.

The selected text-type attributes are defined as categorical variables of nominal type, because it is not possible to establish an order in their categories. Therefore, the *one-hot coding* technique [16] is carried out due to its easy integration when dealing with this type of categorical variables, using the get_dummies function of the Python Pandas library. Attribute conversion was performed for Department_Entity, Recruitment_modality and Recruitment_type.

For the attribute Procedure_Description which presents the object of a tender, it is a key attribute for the business, therefore the application of a text mining technique called TF-IDF, [8] which consists of measuring the relevance or weight of a word or term in a tender TF (term frequency) and then IDF value (inverse document frequency) for each term in the vocabulary. The TF and IDF values are then multiplied to obtain the final TF-IDF value.

A process of lemmatization and elimination of stop words that were not relevant to the model was also carried out.

After completing this process, a selection of 40 keywords was obtained, selecting those with the highest frequencies and of special interest in bid searches, as shown in Fig. 5.

Once the variables were converted for the development of the model, the new data set was integrated. The new data set contains a total of 97 variables and 40,739 records, which include the converted categorical variables, also the variable converted by text mining and the numerical type variables selected from the original data set.

In the new dataset, the number of records per class was calculated where each bid is labeled in one of the classes: Interesting = 1, Not interesting = 0. As per Fig. 6, it is observed that the dataset is extremely unbalanced with a total of 39,498 uninteresting tenders and 1,241 interesting tenders, thus requiring a data balancing technique.

Three different data balancing techniques are employed, which are subsequently used in the measurement of the models to be evaluated. The balancing techniques used are [14]:

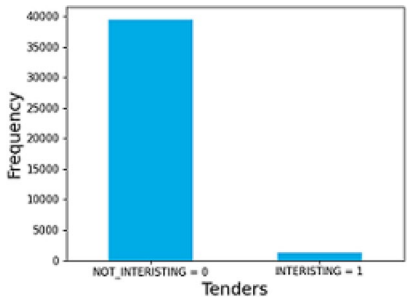

Fig. 6 Data distribution by label

1. Random Under-Sampling which eliminates random samples from the majority class (0) and equals the minority class (1).
2. NearMiss (*k*-neighbors) which balances the data by removing nearby samples from the majority class (0).
3. Combined Advance Resampling SMOTE-Tomek balancing where synthetic data are oversampled and removed with close losses under the Tomek method.

Modeling

The study emphasizes on training and analyzing different classification models in order to select the best model that can classify the largest number of tenders of interest, thus generating greater opportunities for resources for the Minuto de Dios University. The Python libraries Pandas, Numpy, Seaborn and Scikit-learn were used to develop the classification models. Source code for the Colab Notebook is available at: https://bit.ly/3Pfqmhw. In the modeling process, the following classification algorithms were selected:

- *Logistic Regression* Multivariate analysis algorithm [12], which is used when there is a dichotomous dependent variable is an attribute that has defined classes between zero and one, respectively, and a set of predictor or independent variables, which can be quantitative or categorical. Similarly, the concept of regression refers to the experimental law [12] that seeks to identify the relationship between correlated variables, which occurs when one variable is put in function of another or others.
- *Decision Trees* These are a data mining technique that performs binary partitions achieving predictions of future events [18], in classification problems are based on flowcharts, performs classification into groups or predict values of a dependent variable taking into account values of independent or predictor variables.
- *Random Forests* A machine learning algorithm consisting of a large number of decision trees [5], which have the same distribution for all trees in the forest but improves the accuracy and reliability of the model by avoiding model overfitting problems.
- *Adaptive Boosting* It is a supervised machine learning algorithm that works by training several weak learning models, including decision trees [11], performed sequentially, i.e., a classifier consisting of many classifiers called base classifiers where they are combined in a weighted voting process to produce a final classification. It is characterized by handling complex data which resists overfitting.
- *Gradient Boosting* Its implementation in classification problems is described as a weighted combination of multiple weak predictors [2]. It iteratively adjusts the data weights to directly reduce the prediction error so that it can improve the classification accuracy as opposed to a single model.

The data set was then divided into 60% for training, 20% for validation and hyperparameter adjustment, and 20% for model testing and evaluation.

Regarding the hyperparameter settings, the Randomized Search CV technique of the scikit-learn library was applied where the hyperparameter values are chosen randomly [1], applying cross-validation (Cross-Validation) with a total number of five iterations or folds for each model.

The modeling of the selected algorithms was performed for each of the three different data balancing techniques mentioned above, in order to identify different behaviors in the results of each model.

Evaluation

The results of the proposed models are presented below, taking into account the data set for testing and evaluation, previously considering the hyperparameter settings for each model. In this evaluation, accuracy, recall and F1 score were defined as analysis metrics.

Figure 7 shows the confusion matrix for each resulting model, where the data were balanced using the **Random Under-Sampling** technique, where the best performing model was the Random Forest Classifier.

According to Fig. 8, the results of the models with balanced data using the **NearMiss** technique are presented. The performance of the models resulted in Random ForestClassifier and GrandientBoostingClassifier show similar behavior.

Finally, Fig. 9 shows the results from the balanced data using the **SMOTE-Tomek** technique, where the best performing model was Random ForestClassifier which obtained better metrics for this balanced data set.

Table 2 presents a summary of the results obtained for each classification model, according to the data balancing

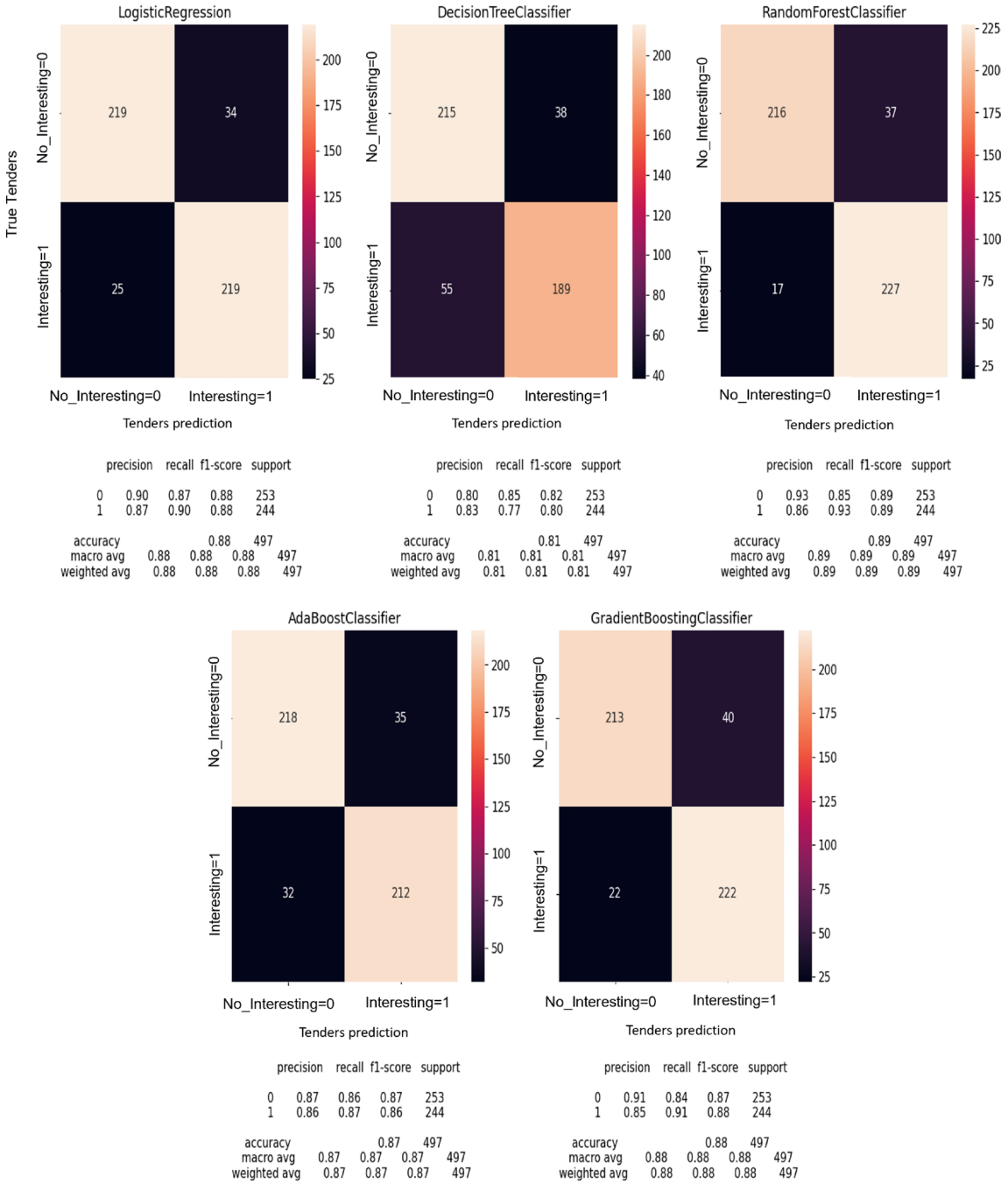

Fig. 7 Comparison of models—random under-sampling balanced data

technique applied and ordered from highest to lowest performance in each case.

Discussion

The proposed classification models show variations in their performance according to the three data balancing techniques applied. In the case of the Random Under-Sampling balancing technique, the models obtained a general accuracy margin of 85%, which means that these models have a relatively good performance when performing a random subsampling of the majority class, as a particular case, the model that obtained the best performance was Random Forest Classifier with an accuracy of 89%. In the second Near Miss balancing technique, the models that improved their performances were Random Forest Classifier, Gradient Boosting Classifier and AdaBoost Classifier which

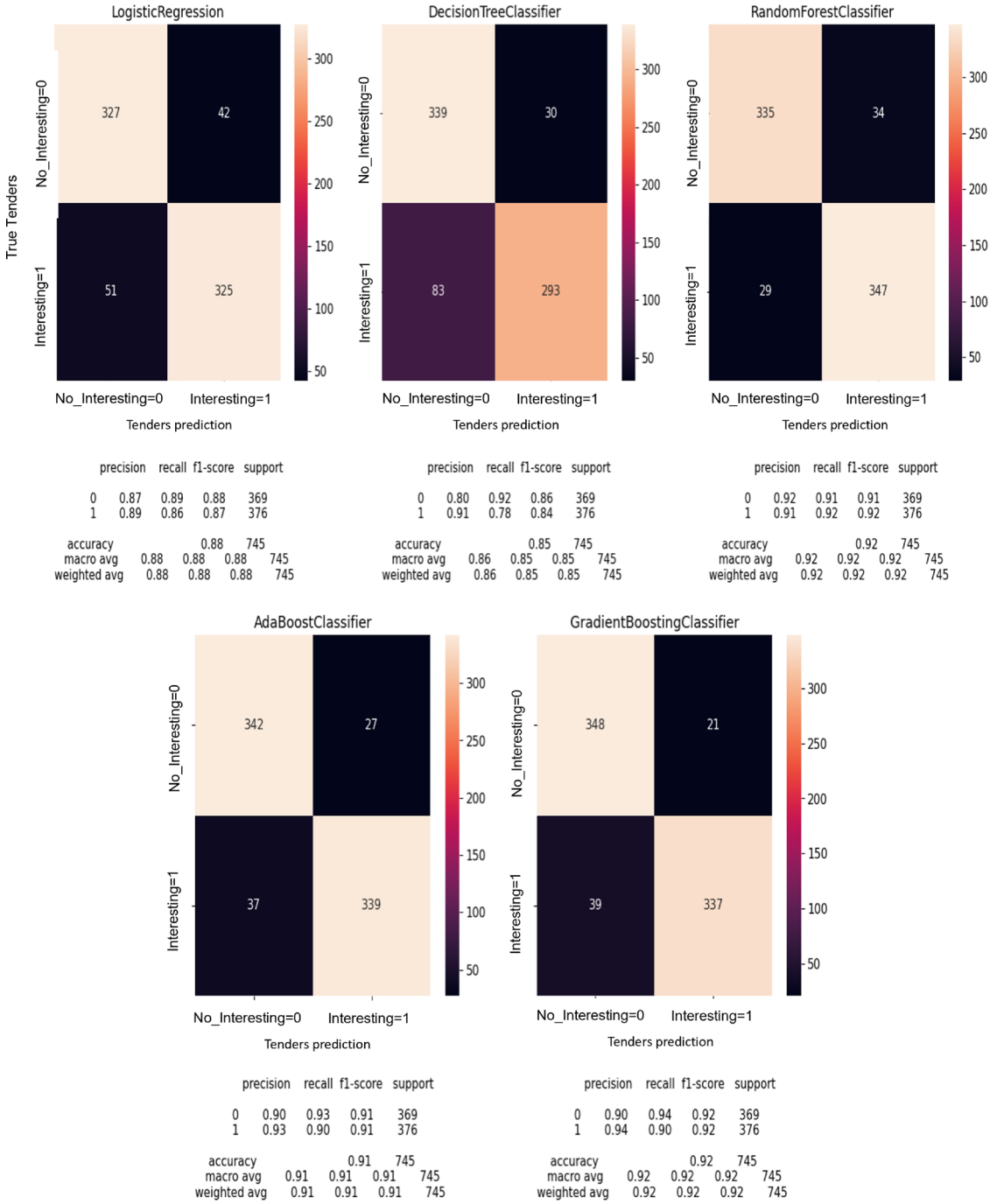

Fig. 8 Comparison of models—NearMiss Balanced Data

obtained an accuracy between 91% and 92%, which allows us to affirm that this balancing technique helps the models, improving their performance since the data set is balanced by removing those samples of the majority class that are closer to the minority class. Applying the SMOTE-Tomek balancing technique, the performance of all the models improved notably, achieving an overall average accuracy margin of 98% for all models, however the model with the best performance was again Random Forest Classifier which obtained an accuracy of 99%. Given that in some cases the results of the models presented similar performance metrics such as precision, recall and F1 Score, the false positive metrics of the confusion matrix were analyzed, which indicate those tenders of interest that the model failed to classify as

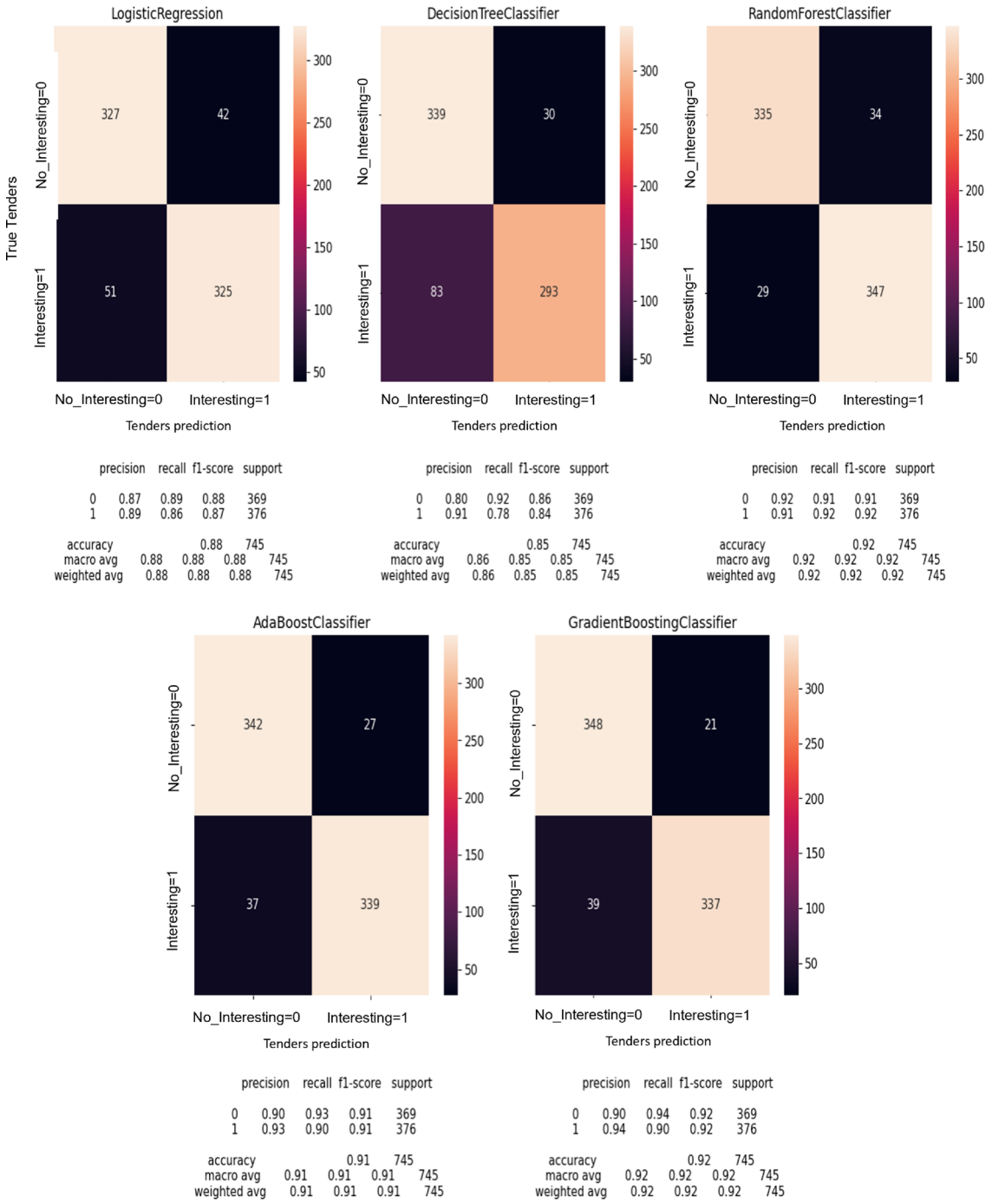

Fig. 9 Comparison of models—SMOTE-Tomek balanced data

interesting, this represents that by having a greater number of false positives the model is not adequate for the entity's objective despite obtaining good results in its metrics. This made it possible to rank the models evaluated in a top ranking from highest to lowest performance, taking into account the analysis offered by the confusion matrix. As can be seen, the model that obtained the best performance taking into account the balancing techniques and from the results of the evaluated metrics is Random Forest Classifier, which had the best performance.

When applying the SMOTE-Tomek data balancing technique, the models obtained an outstanding performance compared to other techniques, due to the use of a combined oversampling algorithm (SMOTE) generating synthetic data

Table 2 Results of models for the classification of public tenders

Data balancing technique	Top	Model	Evaluation metrics		
			Accuracy	Recall	F1 Score
Random Under-Sampling	1	Random Forest Classifier	0.89	0.93	0.89
	2	Gradient Boosting Classifier	0.88	0.90	0.88
	3	Logistic Regression	0.88	0.90	0.88
	4	AdaBoost Classifier	0.87	0.87	0.86
	5	Decision Tree Classifier	0.79	0.76	0.78
NearMiss	1	Random Forest Classifier	0.92	0.92	0.92
	2	Gradient Boosting Classifier	0.92	0.90	0.92
	3	AdaBoost Classifier	0.91	0.90	0.91
	4	Logistic Regression	0.88	0.86	0.87
	5	Decision Tree Classifier	0.85	0.78	0.84
SMOTE-Tomek	1	Random Forest Classifier	0.99	0.99	0.99
	2	Gradient Boosting Classifier	0.98	0.98	0.98
	3	AdaBoost Classifier	0.98	0.98	0.98
	4	Decision Tree Classifier	0.98	0.98	0.98
	5	Logistic Regression	0.96	0.95	0.96

for the minority class and (Tomek) that performs the elimination of close instances between classes achieving a more evident separation between classes in order to improve the accuracy of the models. Therefore, it can be inferred that the best balancing technique for the case of these models is SMOTE-Tomek, however, when performing data oversampling and the execution of each model, a greater computational effort was evidenced which has an impact on the processing time, which represents an important factor to consider in a possible deployment and development for the best model. It is recommended the importance of analyzing the SMOTE-Tomek technique in subsequent studies in order to check the behavior of the results with new data sets, which will allow to identify if under this technique no over-fits are identified. In short, the application of the models obtained positive results, which can be an input for further studies where the application of other types of algorithms for classification problems and the exploration of Text Mining techniques that can improve the efficiency and performance with respect to the challenge of classifying interesting tenders for the entity can be taken into account.

Conclusions

The comparative results between the proposed classification models allow us to affirm that the best classification model for this case was Random Forest Classifier, taking into account the different data balancing techniques used, the adjustment of hyperparameters and the performance metrics evaluated in each model. The data balancing technique that provided the best performance in most of the proposed models was SMOTE-Tomek, because it is a combined technique; however, it is important to consider the computational effort involved in developing a deployment of the best model.

Finally, it is recommended for further studies to analyze the behavior of other techniques for data balancing in order to identify overfitting or any other anomaly that may affect the performance of the models, including processing time. It is important for the development of further studies to explore different Text Mining techniques due to the characteristics of the data related to public tenders, which would allow to achieve models with greater adjustments and optimal performance against the business objectives of the entities.

Author Contributions YF-G contributed with conceptualization, methodology, coding, evaluation and paper writing. IG contributed with conceptualization, paper writing and supervision.

Funding Open Access funding provided by Colombia Consortium. Yeersainth Figueroa-Gómez received a grant from Minuto de Dios University in order to undertake Masters studies which resulted in this research.

Data Availability The data used in this article is open data and is publicly available at the following Colombian government website: https://www.datos.gov.co/Gastos-Gubernamentales/, https://www.SECOP-II-Procesos-de-Contrataci-n/p6dx-8zbt.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Human and Animal Rights This article does not contain any studies with human or animal participants.

Informed Consent There are no human participants in this article and informed consent is not applicable.

References

1. Andradóttir S. A review of random search methods. In: Handbook of simulation optimization; 2014. pp. 277–292.
2. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev. 2021;54:1937–67.
3. Colombia Compra Eficiente. Decreto 4170 de 2011. https://www.funcionpublica.gov.co/eva/gestornormativo/norma_pdf.php?i=44643. Last Accessed 2023/05/08.
4. Colombia Compra Eficiente. SECOP II. https://colombiacompra.gov.co/ciudadanos/preguntas-frecuentes/secop-ii. Last Accessed 2023/05/08.
5. Cutler A, Cutler DR, Stevens JR. Random forests. In: Ensemble machine learning: methods and applications; 2012. pp. 157–175.
6. Goswami S, Kapoor S, Bhardwaj P. Machine learning for automated tender classification. In: 2011 annual IEEE India conference. 2011. pp. 1–4. https://doi.org/10.1109/INDCON.2011.6139406
7. Limas Cano DS, Álvarez Villamizar LC, et al. Modelo de referencia para la gestión de procesos licitatorios en colombia usando analítica de datos.
8. Manning CD. An introduction to information retrieval. Cambridge University Press; 2009.
9. Mara Mendes MF. Digiwhist recommendations for the implementation of open public procurement data an implementer's guide. https://digiwhist.eu/wp-content/uploads/2017/04/digiwhist_implementers_guide.pdf. Last Accessed 2023/05/08.
10. Marcarelli G, Squillante M. A group-AHP-based approach for selecting the best public tender. Soft Comput. 2020;24(18):13717–24.
11. Margineantu DD, Dietterich TG. Pruning adaptive boosting. In: ICML, vol. 97. Citeseer; 1997. pp. 211–218.
12. McCullagh P. Generalized linear models. London: Routledge; 2019.
13. Mencıa EL, Holthausen S, Schulz A, Janssen F. Using data mining on linked open data for analyzing e-procurement information. In: Proceedings of the first DMoLD: data mining on linked data workshop at ECML/PKDD. 2013.
14. Parmar G, Gupta R, Bhatt T, Sahani G, Panchal BY, Patel H. A review on data balancing techniques and machine learning methods. In: 2023 5th international conference on smart systems and inventive technology (ICSSIT). IEEE; 2023. pp. 1004–1008.
15. Red interamericana de compras gubernamentales. analítica de datos sobre compras públicas. http://ricg.org/es/datos-regionales/analitica-de-datos-en-compras-publicas/. Last Accessed 2023/05/08.
16. Rocha Íñigo A. Codificación de variables categóricas en aprendizaje automático. 2020.
17. Skitmore M. Predicting the probability of winning sealed bid auctions: a comparison of models. J Oper Res Soc. 2002;53:47–56.
18. Von Winterfeldt D, Edwards W. Defining a decision analytic structure. Princeton: Citeseer; 2007.
19. Wirth R, Hipp J. Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, vol. 1. Manchester; 2000. pp. 29–39.