


Fuel Classification in Electronic Tax Documents

Yúri Faro Dantas de Sant'Anna¹^a, Mariana Lira de Farias^{2,3}^b, Methanias Colaço Júnior^{2,3}^c,
Daniel Oliveira Dantas^{2,3}^d and Max Castor Rodrigues Junior³^e

¹*Centro de Informática, Universidade Federal de Pernambuco, Recife, PE, Brazil*

²*Departamento de Computação, Universidade Federal de Sergipe, São Cristóvão, SE, Brazil*

³*Centro Universitário Estácio de Sergipe, Aracaju, SE, Brazil*

Keywords: Supervised Learning, Invoice, Text Classification, Naive Bayes.

Abstract: The Tax on the Circulation of Goods and Services (Imposto sobre Circulação de Mercadorias e Serviços, ICMS), a responsibility of the federative units, is the main Brazilian tax collection resource. One way to collect this tax is through a product's weighted average price to the end consumer (preço médio ponderado ao consumidor final, PMPF) of a product. The PMPF is the only resource for charging state fees for the fuel segment, so if improperly calculated, it can lead to losses both in the collection of public funds and in the evolution of prices practiced by merchants. The objective of this work is to make a comparative analysis of classification algorithms used to calculate the PMPF of fuels in the state of Sergipe to select the most appropriate technique. This system circumvented deficiencies present in the previously applied simple random sampling methodology. The naive Bayes algorithm was considered the most effective approach due to its high accuracy and feasibility of application in a real-life scenario.

1 INTRODUCTION

The Tax on the Circulation of Goods and Services (Imposto sobre Circulação de Mercadorias e Serviços, ICMS) is the revenue with the highest volume of collection in all Brazilian states (Rezende, 2009). Regulated by the Kandir Act (Brasil, 1996). Its collection is based on the product category and the current state legislation, defining the rate and the calculation base used for each economic segment or product.


The weighted average price to the end consumer (preço médio ponderado ao consumidor final, PMPF) is a value that reflects the average price used by merchants to final consumers (Santo, 2021) and aims to facilitate the review and monitoring of the ICMS collection. According to Queiroz (Queiroz et al., 2014), this is an essential factor in the reduction of fraud and tax evasion since the ICMS of the entire chain of these products will be collected only by the producer so that the calculation base of the transactions is determined once and by a single actor. It is important to empha-


size that the PMPF is obtained by calculations made through the product's final price and can be mapped from market research, tax inspections to taxpayers, and the issued receipts.


The consumer receipt (Nota Fiscal de Consumidor Eletrônica, NFC-e) is a digital document issued and stored electronically that aims to document transactions of movement of goods or services rendered (Brasil SPED, 2016). Within this document, various resources (fields) aim to map the various characteristics of the products, in addition to the values relating to their emission (product price, tax, freight, etc).


The Mercosur Common Nomenclature (Nomenclatura Comum do Mercosul, NCM) is a classification of goods present in receipts that maps goods into affinity groups. The Department of Finance (Secretaria da Fazenda, SEFAZ) may obtain the PMPF of fuels through a sample of NFC-e filtered by the codes that represent them.


The classification in the NCM is an assignment of the taxpayers and has a declaratory character. Thus, there is no legal penalty for NCM errors or omissions in tax documents. Therefore, in the calculation of PMPF, fuel receipts with incorrect or missing codes, as well as receipts of other products wrongly classified, can be taken into account.

^a  <https://orcid.org/0000-0002-3527-6862>

^b  <https://orcid.org/0009-0007-3113-2849>

^c  <https://orcid.org/0000-0002-4811-1477>

^d  <https://orcid.org/0000-0002-0142-891X>

^e  <https://orcid.org/0000-0003-0392-6696>

In addition to the problem generated by the filling of the NCM, the calculation of the PMPFs is done in a sampling manner, and their average is calculated purely statistically. The value is subject to the variations caused by the selection of the different samples. Due to these problems, several states have sought strategies for a more accurate calculation of their PMPFs, avoiding the bias of selecting a sample that does not faithfully represent the reality of prices, limited to receipt filled with the code relevant to the fuel type and subject to non-fuel receipt.

The use of pattern recognition techniques, especially classification algorithms, has the potential to eliminate the problems mentioned above. According to Jarude (Jarude, 2020), using these techniques can lead to significant gains in the efficiency of the services provided by the tax administration.

The objective of this study is to select and evaluate algorithms to classify invoices into fuel classes. The class is used to calculate the average product price in a dynamic real-life scenario such as the one found at SEFAZ in Sergipe. The classifier must be able to identify the fuel class from the textual field containing the product descriptions in the invoices, thus circumventing the problems in the NCM classification.

This study is organized as follows: Section 2 presents related works and relevant references for the development of this project. Section 3 contains the methodology used in conducting this study. In Section 4 are the development steps of the classifier. In Section 5, the results are discussed. Finally, Section 6 presents the conclusions and possible future works.

2 RELATED WORKS

The study of Batista (Batista et al., 2018) aims to automatically classify NCM codes based on product descriptions contained in NFC-e. Using the naive Bayes algorithm, the invoices were classified into two NCM classes. Batista used three datasets with different difficulties, simple, medium, and complex, obtaining accuracies of 98%, 90%, and 83% respectively.

Dias (Dias and Júnior, 2022) used classification committees to identify products sold in Rio Grande do Norte state based on the product description field of a document similar to NFC-e. Different committee architectures were used so that it was possible to compare their robustness. The bagging architecture obtained the best performance.

The work of Madeira (Madeira, 2015) applied data analysis and mining techniques to identify invoices issued incorrectly based on the description of the services provided. A system was developed with

the k-means algorithm, where the NFC-e from the tax subgroup code 07.19.04 (consulting engineering services) previously pre-classified were used, using the naive Bayes and stochastic gradient descent algorithms.

The present work differs from others due to the need to generate a classifier that recognizes and groups products recognized as fuels, not just categories, within an authentic and comprehensive mass of tax documents. This objective is achieved through a comparative analysis between the tested algorithms and selecting the methodology that obtained the best results.

3 METHODOLOGY

The proposed methodology can be divided into four steps: planning and selecting algorithms, data collection and generation of databases, comparison of algorithms, and analysis of results.

The first step consisted of a literature review to find algorithms that could present themselves as possible solutions to the proposed problem. The techniques listed as applicable to our problem were the naive Bayes algorithm, classification with a support vector machine, K-nearest neighbors, random forests, and decision trees.

In the second step, three datasets were created, two for training and one for testing the classifier. According to Dönmez (Dönmez, 2013), the number of inputs used for training is crucial for the efficacy of classifying algorithms. Therefore, an eight-day set of invoice items was used. Approximately 500,000 NFC-e are issued per day in the state.

The first training dataset is called the Natural dataset. It contains product descriptions that are contained in fuel invoices that are issued over a day. It has two columns: product description, which is the classifier input, and fuel class, which is the variable to predict.

According to Purohit (Purohit et al., 2015), it is common in text classification problems to use keywords that discriminate a particular set instead of using the full texts, which may contain noise or terms irrelevant to data classification. Therefore, a Keyword dataset containing standard terms found in invoice descriptions was also created to verify the applicability of keywords and a possible gain in the performance of the fuel classifier with their use. Similarly to the Natural dataset, it consists of two columns: the keyword, which represents the input data of the classifier, and the fuel class, the target variable to predict.

Furthermore, the Test dataset was created with the

description of the products contained in the invoices issued over the remaining seven days. More details about the datasets are given in Subsection 4.2.2. Finally, the algorithms were evaluated using a Monte Carlo method based on four quality metrics: accuracy, sensitivity, precision, and kappa coefficient. Each step will be detailed in the following sections.

The Python programming language in version 3.7 and an Oracle database management system (SGDB) were used to develop the classifiers. The language has libraries with extensive documentation and applicability in actual cases, such as the consolidated *scikit-learn* used for classifier training and the evaluation methodology (Scikit-learn, 2022).

4 FUEL CLASSIFIER DEVELOPMENT

This section will explain the process of developing the experiments for constructing the SEFAZ-SE fuel classifier. The experimental process was based on that presented by Wohlin (Wohlin et al., 2012).

4.1 Objective Definition

The objective definition of this study was formalized using the GQM (goal, question, metric) approach (Caldiera and Rombach, 1994). Our study aims to create a functional fuel classifier that can categorize electronic invoices from the fuel industry. This classification allows calculating the PMPF for each product category. Experiments were conducted using the selected algorithms to determine the most effective technique based on their accuracy, sensitivity, precision, and kappa coefficient.

4.2 Planning of Experiments

To identify the most effective classifier, the algorithms were trained using two datasets: the Natural database and the Keyword database. The Test dataset was used for testing. The algorithms used, the process of creating the datasets, and details about the experiments will be detailed below.

4.2.1 Algorithms

Five algorithms were compared to find a promising solution to solve the classification problem. They all use their canonical structure, widely explored to solve problems like this (Duda et al., 2001) (Kubat, 2017). The algorithms used were naive Bayes, KNN, SVC, random forest, and decision tree.

4.2.2 Datasets

Three datasets were developed, two for training (Natural and Keyword) and a Test dataset. The two training datasets were created to verify the potential benefit of utilizing frequent terms from fuel descriptions in the training, instead of complete descriptions, in the classifier's performance. All three datasets have two columns: one with the product class and another with the text to be classified.

The Natural dataset is composed of product descriptions from the day of NFC-e. It has two columns: product class and product description. Table 2 shows an example of this base.

On the other hand, the Keyword dataset is composed of terms frequently used to describe the fuels present in the SEFAZ-SE database. This dataset has two columns: product class and frequent terms. Table 3 illustrates an example of records from this database. It is important to note that there are thousands of contributors to the fuel segment alone in a database of this size. Each contributor can use different ways to describe their products. Standard word detection was done through database queries and empirical mapping with the help of the audit team responsible for monitoring this segment.

The Test dataset contains descriptions of products present in NFC-e issued in one week. This dataset has two columns: product class and the product description. The variable to be predicted by the classifier is the product class. It can assume one of the following eight categories: REGULAR GASOLINE, GASOLINE WITH ADDITIVES, DIESEL OIL S10, DIESEL OIL S500, VEHICULAR NATURAL GAS, AVIATION KEROSENE, LPG (liquefied petroleum gas), and IGNORED.

In a production environment, several non-fuel products are misplaced under NCM categories different from theirs. These occurrences also need to be identified by the fuel classifier. Therefore, the IGNORED category has been created, a collection of products often incorrectly placed under NCMs fuel categories.

The three datasets were labeled manually, and on account of the eight-day volume of invoices, duplicate terms were removed to reduce the datasets and facilitate the labeling step. Therefore, the Natural dataset contains 1285 records, while the Keyword dataset contains 207 record, and the Test dataset 2499 records. Table 1 shows the number of records per product class.

To create the Natural and Test datasets, the invoices were filtered based on NCM codes, in which it was possible to find products from the desired group and not just the NCMs that are indicated for these

Table 1: Distribution of examples by class.

Product class	Number of records		
	Keyword dataset	Natural dataset	Test dataset
FUEL ALCOHOL (ETHANOL)	13	46	35
VEHICULAR NATURAL GAS	8	12	4
GASOLINE WITH ADDITIVES	8	54	36
REGULAR GASOLINE	8	42	38
PREMIUM GASOLINE	8	1	1
LPG	12	25	33
IGNORED	124	1047	2310
DIESEL OIL S10	8	50	35
DIESEL OIL S500	8	7	6
AVIATION KERESONE	8	1	1

Table 2: Natural dataset example.

Product class	Data example
IGNORED	OLEO LUBRAX TURBO 15W40
IGNORED	ALCOOL LIQ BRILUX 70 500ML
REGULAR GASOLINE	GASOLINA C COMUM (B1)
REGULAR GASOLINE	GASOLINA COMUM B6
REGULAR GASOLINE	GASOLINA TIPO C Bico
GASOLINE WITH ADDITIVES	GASOLINA ADITIVADA VPOWER BICO 15
GASOLINE WITH ADDITIVES	GASOLINA ADITIVADA V POWER BICO 13
GASOLINE WITH ADDITIVES	GASOLINA PETROBRAS GRID B7
VEHICULAR NATURAL GAS	GNV GAS NATURAL
VEHICULAR NATURAL GAS	GAS NATURAL VEICULO-GNV
DIESEL OIL S10	DIESEL EVOLUX S-10 B3
DIESEL OIL S10	OLEO DIESEL B S10 ADITIVADO PETROBRAS GRID B1
DIESEL OIL S500	OLEO DIESEL BS 500 ADITIVADO
DIESEL OIL S500	OLEO DIESEL B S500 B9
LPG	GLP BOTIJAO 13 KG
LPG	GLP VASILHAME SGB 13KG
AVIATION KEROSENE	JET A1 NAO TABELADO - LI

Table 3: Keyword dataset example.

Product class	Data example
IGNORED	LUBRAX TURBO
IGNORED	ALCOOL BRILUX
REGULAR GASOLINE	GASOLINA COMUM
REGULAR GASOLINE	GASOLINA TIPO C
GASOLINE WITH ADDITIVES	GASOLINA V-POWER
GASOLINE WITH ADDITIVES	GASOLINA PETROBRAS GRID
VEHICULAR NATURAL GAS	GNV
VEHICULAR NATURAL GAS	GAS NATURAL VEICULO
DIESEL OIL S10	OLEO DIESEL S10 COMUM
DIESEL OIL S10	OLEO DIESEL BS10
DIESEL OIL S500	EXTRA DIESEL BS 500
DIESEL OIL S500	OLEO DIESEL S500
LPG	GLP 13KG
LPG	GLP 13KG
AVIATION KEROSENE	JET A-1 NAO TABELADO - LI

Table 4: Statistical tests.

Test	p-Value
Kruskall-Wallis	0
Friedman	0
Wilcoxon	0.0388

Table 5: Ranking of algorithms according to accuracy (A).

Algorithm	A
Naive Bayes - KW	0,9996
KNN (1 neighbor) - KW	0,9869
SVC - KW	0,9836
KNN (3 neighbor) - KW	0,9817
KNN (2 neighbor) - KW	0,9784
Random forest - KW	0,9772
Decision tree - KW	0,9759
Random forest - NDS	0,9719
KNN (8 neighbor) - KW	0,9695
KNN (6 neighbor) - KW	0,9686
Decision tree - NDS	0,966
KNN (7 neighbor) - KW	0,9658
KNN (4 neighbor) - KW	0,9638
KNN (5 neighbor) - KW	0,9522
SVC - NDS	0,9462
Naive Bayes - NDS	0,5655
KNN (1 neighbor) - NDS	0,2937
KNN (2 neighbor) - NDS	0,2501
KNN (4 neighbor) - NDS	0,2275
KNN (3 neighbor) - NDS	0,2241
KNN (5 neighbor) - NDS	0,214
KNN (6 neighbor) - NDS	0,2072
KNN (7 neighbor) - NDS	0,1926
KNN (8 neighbor) - NDS	0,1862

products. It should be noted that there are cases in which fuels are linked with incorrect NCM codes. Therefore, these codes were also considered in this process.

The NCM codes used were: From group 27 (mineral fuels, mineral oils and products of their distillation; bituminous materials and mineral waxes), 22071090 (Neutral alcohol), 22072019 (drinks, alcoholic liquids, and kinds of vinegar - Undenatured ethyl alcohol, with an alcohol content by volume equal to or greater than 80%; ethyl alcohol and spirits, denatured, with any alcohol content - Ethyl alcohol and spirits, denatured, with any alcohol content) and 84812090 (Nuclear reactors, boilers, machines, apparatus and mechanical instruments, and parts thereof - Taps, valves and similar devices, for pipes, boilers, reservoirs, vats and other containers - Valves for hydraulic or pneumatic oil transmissions).

Table 6: Classification of algorithms according to precision (P).

Algorithm	P
Naive Bayes - KW	0,9997
KNN (1 neighbor) - KW	0,9914
Random forest -KW	0,9912
SVC - KW	0,9908
KNN (3 neighbors) - KW	0,9884
KNN (4 neighbors) - KW	0,9873
Decision tree - KW	0,9868
KNN (2 neighbors) - KW	0,9867
KNN (5 neighbors) - KW	0,9861
KNN (2 neighbors) - NDS	0,986
KNN (1 neighbor) - NDS	0,986
KNN (3 neighbors) - NDS	0,9858
KNN (5 neighbors) - NDS	0,984
KNN (6 neighbors) - NDS	0,9834
KNN (8 neighbors) - KW	0,9832
KNN (6 neighbors) - KW	0,9828
KNN (8 neighbors) - NDS	0,9826
Random forest - NDS	0,9824
KNN (7 neighbors) - NDS	0,9821
Decision tree - NDS	0,98
SVC - NDS	0,9797
KNN (4 neighbors) - NDS	0,9782
KNN (7 neighbors) - KW	0,9762
Naive Bayes - NDS	0,9607

4.3 Experiments

Initially, the databases went through preprocessing steps in order to increase the quality of the classification. Terms related to fuel pump numbers were removed from the note descriptions using regular expressions (BICO [0-9]+ and B[0-9]+). To adapt the inputs to the algorithms, the vectorization technique was applied, which consists of converting texts into matrices of terms.

It is important to emphasize that, for each algorithm, two models were trained: one with the Natural dataset (identified with the name of the algorithm and the acronym NDS) and the other with the Keyword dataset (identified with the name of the algorithm and acronym KW).

For the execution and validation of the results, the Monte Carlo (Besag and Diggle, 1977) evaluation method was used, where up to 20% (the exact value is chosen randomly) of the Test dataset was removed at each iteration. A total of 100 iterations were performed. At each iteration, the evaluation metrics mentioned above were calculated.

In the Test step, comparisons of the metrics were made. The mean, median, and maximum values of

Table 7: Ranking of algorithms according to sensitivity (S).

Algorithm	S
Naive Bayes - KW	0,9996
KNN (1 neighbor) - KW	0,9869
SVC - KW	0,9836
KNN (3 neighbors) - KW	0,9817
KNN (2 neighbors) - KW	0,9784
Random forest - KW	0,9772
Decision tree - KW	0,9759
Random forest - NDS	0,9719
KNN (8 neighbors) - KW	0,9695
KNN (6 neighbors) - KW	0,9686
Decision tree - NDS	0,966
KNN (7 neighbors) - KW	0,9658
KNN (4 neighbors) - KW	0,9638
KNN (5 neighbors) - KW	0,9522
SVC - NDS	0,9462
Naive Bayes - NDS	0,5655
KNN (1 neighbor) - NDS	0,2937
KNN (2 neighbors) - NDS	0,2501
KNN (4 neighbors) - NDS	0,2275
KNN (3 neighbors) - NDS	0,2241
KNN (5 neighbors) - NDS	0,214
KNN (6 neighbors) - NDS	0,2072
KNN (7 neighbors) - NDS	0,1926
KNN (8 neighbor) - NDS	0,1862

the 100 iterations were calculated. Tables 5, 6, 7 and 8 show the average results for each metric.

In order to verify the distribution of the accuracies, statistical tests were applied. Three non-parametric tests were used: Kruskal-Wallis, Friedman, and Wilcoxon. The results of the statistical tests are in Table 4.

5 RESULTS

Evaluation metrics allow us to analyze how correct a model is in its predictions (Han et al., 2011). We evaluated the performance of the proposed classifiers with four evaluation metrics: accuracy (A), precision (P), sensitivity (S), and kappa coefficient (K) (Han et al., 2011).

Three approaches, trained using the Keyword dataset, have very close values: naive Bayes, KNN with one neighbor, and SVC. By analyzing the results obtained, shown in the tables below, it was possible to answer the research question of this project. The three statistical tests show p-values lower than the significance level. Therefore, it can be concluded that there is a significant difference between the results of these techniques.

Table 8: Ranking of algorithms according to the kappa coefficient (K).

Algorithm	K
Naive Bayes - KW	0,9972
KNN (1 neighbor) - KW	0,9102
SVC - KW	0,8948
KNN (3 neighbors) - KW	0,8767
KNN (2 neighbors) - KW	0,8586
Decision tree - KW	0,8545
Random forest - KW	0,8522
Random forest - NDS	0,8337
Decision tree - NDS	0,8031
KNN (8 neighbors) - KW	0,7932
KNN (6 neighbors) - KW	0,7929
KNN (4 neighbors) - KW	0,7791
KNN (7 neighbors) - KW	0,7694
SVC - NDS	0,7214
KNN (5 neighbors) - KW	0,7163
Naive Bayes - NDS	0,1969
KNN (1 neighbor) - NDS	0,1097
KNN (2 neighbors) - NDS	0,0995
KNN (3 neighbors) - NDS	0,0944
KNN (4 neighbors) - NDS	0,0933
KNN (5 neighbors) - NDS	0,0906
KNN (6 neighbors) - NDS	0,0893
KNN (7 neighbors) - NDS	0,0862
KNN (8 neighbors) - NDS	0,0851

Table 9: Evolution of Collection in the months of July and August.

Year	Fuel revenue	Evolution
2019	52 million BRL	1%
2018	51 million BRL	-13%
2017	42 million BRL	0,007%
2016	48 million BRL	-4%

With the information obtained in the section above and the metrics listed in the tables, it is possible to verify that applying the naive Bayes algorithm is the most appropriate option for the proposed problem. Using the dataset with keywords significantly increased the metrics of the models trained with it.

Given the promising results, the classifier was officially implemented in May 2019. The initial tax payments, fully computed by the classification system, were executed in July 2019. Table 9 shows the increase in revenue within the fuel sector compared to the corresponding periods in previous years. To mitigate the seasonality impact, this analysis assessed the revenue evolution between July and August. An increase in revenue was observed during a month historically characterized by a decrease or stagnation.

6 CONCLUSIONS AND FUTURE WORK

This study proposed the development of an automatic tool for classifying fuel prices to replace the statistical approach previously used by SEFAZ in the state of Sergipe. Five commonly applied text classification techniques were studied, evaluated, and compared. Upon completing algorithm execution and evaluation, it became evident that the naive Bayes classification algorithm was the most efficient in addressing the proposed problem and forming the developed tool.

After implementation, continuous evaluation, and successful use, it was concluded that the system exhibits high reliability and effectiveness. Consequently, the system was adopted by the tax auditor team responsible for the fuel sector. Its use has significantly improved the accuracy and speed of calculating the averages used for the PMPF. It is worth noting that the results of classifications performed in a real-life scenario were audited and approved by the gas station union in Sergipe.

The success achieved in implementing the fuel classifier highlights the potential of applying this pattern recognition algorithm in tax scenarios. The results indicate that the tool may function in a broader scope, although there is no guarantee that the high degree of assertiveness obtained will be maintained if applied to products from other economic segments.

Potential future work may involve extending classification algorithms to other tax segments. The results underscore the possibility of employing some of these techniques to formulate tax guidelines, a fiscal resource that monitors the prices of specific products for tax collection, price monitoring, and price transparency for the end consumer.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES), Finance Code 001.

REFERENCES

- Batista, R. d. A., Bagatini, D. D., and Frozza, R. (2018). Classificação automática de códigos NCM utilizando o algoritmo naïve bayes. *iSys-Brazilian Journal of Information Systems*, 11(2):4–29.
- Besag, J. and Diggle, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Applied Statistics*, 26(3).
- Brasil (1996). Lei complementar nº 87, de 13 de setembro de 1996. Available at: http://www.planalto.gov.br/ccivil/_03/leis/LCP/Lcp87.htm. Last accessed: 8 jun, 2022.
- Brasil SPED (2016). NFC-e. Available at: <http://sped.rfb.gov.br/pagina/show/1519>. Last accessed: 8 jun, 2022.
- Caldiera, V. R. B. G. and Rombach, H. D. (1994). The goal question metric approach. *Encyclopedia of software engineering*, pages 528–532.
- Dias, E. R. F. and Júnior, J. C. X. (2022). Classificação automática de produtos comercializados por órgãos públicos do Rio Grande do Norte através de comitê de classificadores. *Research, Society and Development*, 11(9):e29211931836–e29211931836.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, New York, 2nd edition.
- Dönmez, P. (2013). Introduction to Machine Learning. *Natural Language Engineering*, 19(2):285–288.
- Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Jarude, J. N. D. M. (2020). O estado da arte da fiscalização tributária federal e o uso de inteligência artificial. *Publicações de Parceiros da Enap — Finanças Públicas*.
- Kubat, M. (2017). *An Introduction to Machine Learning*. Springer International Publishing, Gewerbestrasse 11, 6330 Cham, Switzerland, 2nd edition.
- Madeira, R. d. O. C. (2015). *Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais*. PhD thesis, FGV EMAP.
- Purohit, A., Atre, D., Jaswani, P., and Asawara, P. (2015). Text classification in data mining. *International Journal of Scientific and Research Publications*.
- Queiroz, J. V., Lima, N. C., Oliveria, S. V. W. B. d., Martins, E. S., and Oliveira, M. M. B. d. (2014). Considerações tributárias do combustível etanol hidratado. *Revista de Administração e Ciências Contábeis do IDEAU*.
- Rezende, F. (2009). ICMS: Como era, o que mudou ao longo do tempo, perspectivas e novas mudanças. Available at: https://efaz.fazenda.pr.gov.br/sites/default/arquivos_restritos/files/migrados/File/Forum_Fiscal_dos_Estados/FFEB_Caderno_n_10.pdf. Last accessed: 8 jun, 2022.
- Santo, E. (2021). Nota de esclarecimento. Available at: [https://sefaz.es.gov.br/Media/Sefaz/Not/C3/A Dcias/Nota%20de%20esclarecimento%20sobre%20combust/C3/ADveis%20\(1\).pdf](https://sefaz.es.gov.br/Media/Sefaz/Not/C3/A Dcias/Nota%20de%20esclarecimento%20sobre%20combust/C3/ADveis%20(1).pdf). Last accessed: 8 jun, 2022.
- Scikit-learn (2022). Scikit-learn: Machine Learning on Python — SVM — scores and probabilities. Available at: <https://scikit-learn.org/stable/modules/svm.html#scores-and-probabilities>. Last accessed: 8 jun, 2022.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.