Comparison of Algorithms for Classification of Financial Intelligence Reports

Roberto Zaina(✉), Douglas Dyllon Jeronimo de Macedo, Moisés Lima Dutra, Vinicius Faria Culmant Ramos, and Gustavo Medeiros de Araujo

Federal University of Santa Catarina, Florianópolis, Brazil
rzaina@gmail.com, {douglas.macedo,moises.dutra,v.ramos, gustavo.araujo}@ufsc.br

Abstract. This work shows the application of machine learning algorithms to decide whether Financial Intelligence Reports analyzed by the Brazilian Federal Police should be investigated or archived. We explain how the suspicious financial operations found in the Financial Intelligence Reports are a crucial piece of information to combat money laundering crimes. We depict the processing of such reports, which are often used to initiate a police investigation. When that is not the case, the reports are archived. In this work, we propose using machine learning to analyze these reports. We trained and used three classification algorithms: Decision Tree, Random Forest, and KNN. The results show that most reports should be archived. While Decision Tree and Random Forest indicated that about 2/3 of the reports should be archived, KNN indicated that about 4/5 of them should be archived. In the end, this work shows the feasibility of automating the analysis of the Financial Intelligence Reports by the Brazilian Federal Police, despite the need for more adjustments and tests to improve the accuracy and precision of the models developed.

Keywords: Financial intelligence report · Artificial intelligence · Machine learning

1 Introduction

1.1 Financial Intelligence Report (FIR)

Money laundering is a crime that consists of operations to give a lawful appearance to resources that come from illegal activities [1]. This crime aims to introduce money from criminal activities into the legal, financial circuits, through complex operations that promote the untying of its illicit origin [2]. Such activity harms society in many ways, mainly because of the damage it causes to the economic order and the administration of justice. “Dirty money” in the economy causes distortions in financial markets, causes fluctuations in stock exchanges, and harms legitimate businesses that do not rely on the easy supply of illicit money [3].

A. L. Pinto and R. Arencibia-Jorge (Eds.): DIONE 2022, LNICST 452, pp. 204–217, 2022.
https://doi.org/10.1007/978-3-031-22324-2_16

In money laundering investigations in Brazil, the essential information to be analyzed is financial and patrimonial, such as suspicious financial transactions, banking transactions and tax declarations. Suspicious financial transactions are contained in documents called Financial Intelligence Reports (FIR), produced by the Brazilian Financial Intelligence Unit (FIU) and disseminated mainly to criminal prosecution agencies, such as the Brazilian Judicial Police and Public Prosecutors.

One of the agencies that receive FIR in Brazil is the Federal Police (FP). After an initial analysis by a significant sector, the FP distributes the reports to police stations throughout the country for further analysis, this time with more context and depth. After this second analysis, the action for each FIR is decided. It can be of 4 types: 1) to forward it; 2) to archive it; 3) to investigate it; 4) to keep it under analysis.

To forward means the FIR is sent to another agency because the facts narrated in the report are not the competence of the Federal Police. For example, when it turns out that the suspected operations of the FIR are related to crimes of state competence, such as loan sharking, which the state police should investigate.

The archiving happens when there is insufficient evidence to start a formal investigation by the FIR analysis. The FIR is stored in a database for future access and can be unarchived whenever relevant new information emerges.

To investigate means that the FIR presents evidence of federal crimes. A formal investigation is necessary to be held, which will allow the FP to use other investigative diligence, including those that need judicial authorization.

Finally, keeping the FIR under analysis means that assessing the operations described in the report has not yet been completed. Thus, there is no decision whether the FIR will be forwarded, archived, or whether the FP will start a formal investigation.

It is important to emphasize that the decision of forwarding, archiving, or investigating is complex and laborious. It depends on analyzing many elements related to suspicious financial transactions and the people mentioned in the FIR. Aspects such as the number of people and companies mentioned, the sum of the values of operations, locations, and characteristics of transactions, among others, can be considered. Furthermore, the FIR is usually a very extensive and complex document, which cites dozens of people and companies and contains hundreds or thousands of suspicious financial transactions that involve, in many cases, millions of Reais (The Brazilian currency in suspicious values).

This paper proposes to train machine learning algorithms to assist in the FIR-based decision-making of investigating or archiving. This proposal intends to contribute to the automation of the FIR analysis process and the improvement of the efficiency of the investigation work of money laundering crimes by the Brazilian Federal Police.

2 Background

2.1 Artificial Intelligence and Machine Learning

Artificial intelligence is the ability of machines to solve a particular problem on their own, without any human intervention, that is, that the solutions are not programmed directly in the system, but rather that the data is interpreted by the artificial intelligence tool, which produces a solution by itself [4]. Artificial intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs, and intelligence is the computational ability to achieve goals in the world [5].

Machine learning is a subfield of artificial intelligence that trains a machine with specific data and changes when exposed to new data. According to these authors, machine learning focuses on extracting information from large data sets and then detects and identifies patterns using various statistical measures to improve their ability to interpret new data and produce more effective results. Among the most used statistical measures in machine learning are clustering, association, regression and classification [4].

Since our proposal aims to assist in the FIR-based decision-making of investigating or archiving, we sought efficient algorithms for this purpose. We obtained the results presented in this paper by using the Decision Tree, Random Forest, and K-Nearest Neighbors (KNN) algorithms.

2.2 Decision Tree Algorithm

Decision Tree is a predictive model that can represent classification and regression measures, composed of tree decisions and consequences [6]. Classification trees are used to sort an object in a set of classes based on its variables, as shown in Fig. 1.

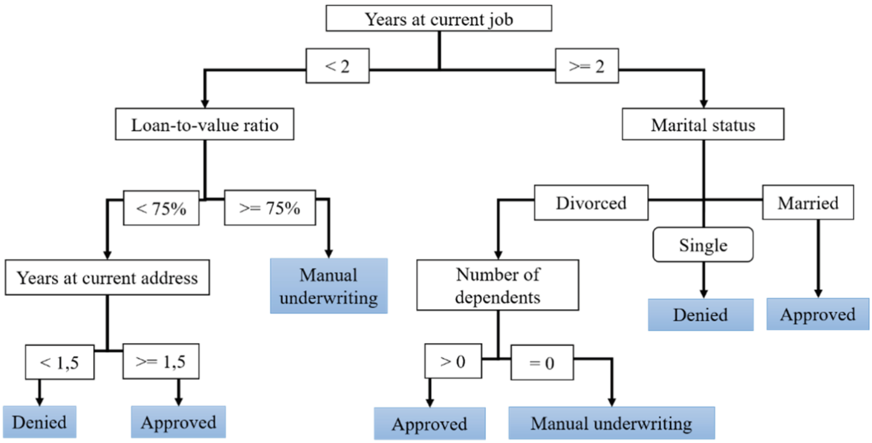

Fig. 1. Example of Decision Tree. Adapted from MAIMON and ROKACH [5]

In this example, a Decision Tree was trained to decide on real estate mortgage lending to choose an "Approved," "Denied," or "Manual underwriting" class. For example, at least two years old in the current job, a married person would have approved his/her mortgage. On the other hand, a person under two years of current employment, with a loan less than 75% of the property's value, with less than 1.5 years of current residence, would have the mortgage denied.

For the present study, Fig. 2 shows, hypothetically, what is expected of applying the Decision Tree algorithm to classify an FIR.

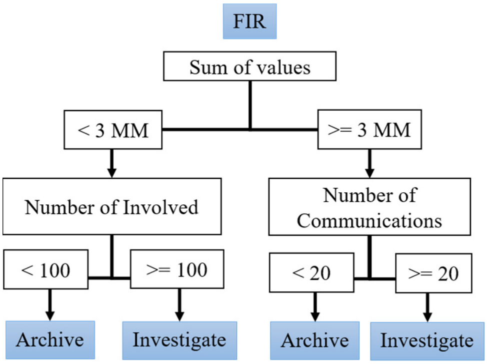

Fig. 2. A hypothetical example of applying Decision Tree to an FIR

In the example above, an FIR with a sum of R$ 3 million with at least 20 communications will be investigated, while one with a sum less than R$ 3 million and less than 100 involved will be archived.

2.3 Random Forest Algorithm

Random Forest consists of tree-structured classifiers, where variables are randomly distributed and identical. Each tree registers a vote to decide the most popular class (most voted) [7]. Random Forest algorithm uses multiple Decision Trees to determine the most votes class, besides having better accuracy compared to a single "tree" [8], as shown in Fig. 3.

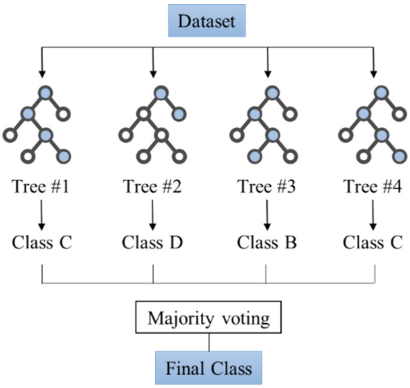

Fig. 3. Random Forest example. Adapted from KIRASICH et al. [7]

In this example, the Random Forest algorithm was trained and processed using 4 Decision Trees, with variables randomly positioned, and each tree (#1, #2, #3, and #4) resulted in a class (C, D, B, and C, respectively). The final class chosen was the one with most of the results, class C.

Figure 4 exemplifies the application of the Random Forest algorithm to determine the action to be taken with an FIR.

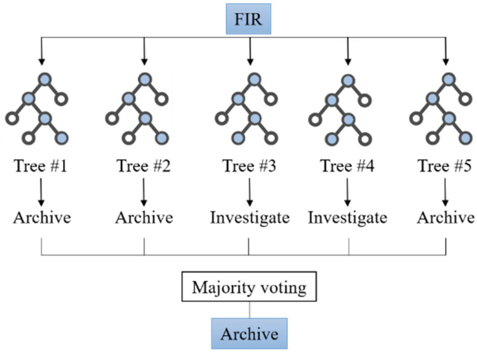

Fig. 4. Hypothetical example of applying Random Forest to a FIR

By Random Forest, defining whether an FIR will be archived or investigated will depend on the decision-making process. In this work, the decision that appeared the most was to "Archive."

2.4 K-Nearest Neighbors (KNN) Algorithm

KNN is based on the closest grouping of elements with the same characteristics and decides the class category in its "k" nearest neighbors [9]. That "k" value depends on the dataset's size and the type of classification problem. Figure 5 below exemplifies the application of KNN.

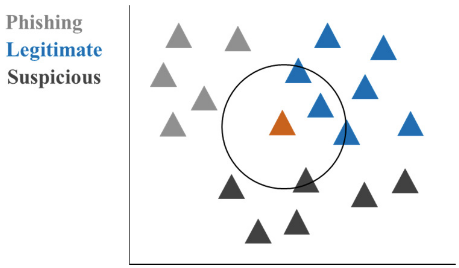

Fig. 5. Example of K-Nearest Neighbors (KNN). Adapted from Altaher [8]

In this example, to sort whether a site (the orange triangle) should be framed as "Phishing," "Legitimate," or "Suspicious," we must check the class type that has the most neighbors nearby. When considering the three nearest neighbors, the site would be set to "Legitimate."

By applying the KNN algorithm to classify an FIR, we expect it to behave as shown in Fig. 6.

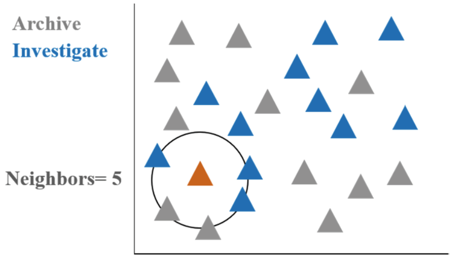

Fig. 6. A hypothetical example of KNN application in an FIR

When using the KNN algorithm to indicate whether an FIR will be archived or investigated, the class with more than "k" nearest neighbors is checked. In this example above, if we consider the nearest five neighbors, the "Investigate" class has more occurrences than the "Archive" class and, therefore, it is chosen.

3 Methodology

The using of Decision Tree, Random Forest, and K-Nearest Neighbors (KNN) algorithms to classify Financial Intelligence Reports as "investigate" or "archive" comprises four tasks (Fig. 7): i) Data Collection; ii) Pre-processing; iii) Learning; and iv) Application.

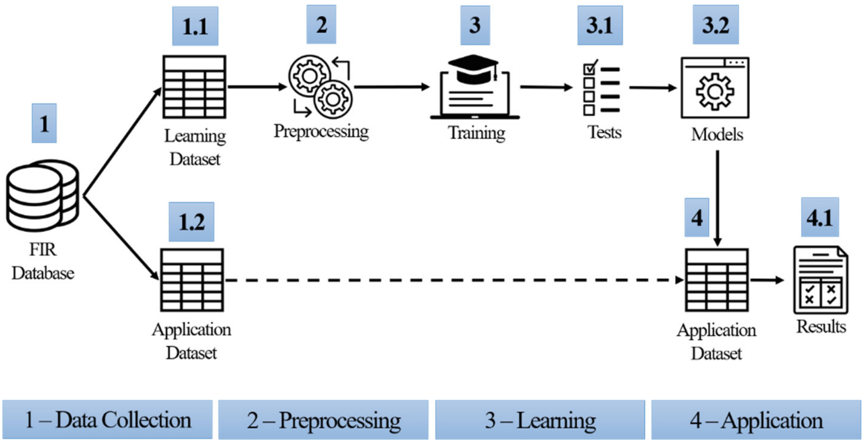

Fig. 7. Methodology for FIR classification

During step 1, we collect the FIR's data from an FP's internal database and store them in two datasets: Learning and Application. In step 2, we process the learning data to eliminate variables with high correspondence and low importance. Next, we run the Decision Tree, Random Forest, and KNN algorithms to train, test, and generate the models (step 3). Finally, in the last step, we apply the models to the Application dataset to get the model's final decision on whether to archive or investigate the FIR.

4 Development

4.1 Data Collection

We accessed a database of 1,427 Financial Intelligence Reports and extracted the following information from each report: "FIR," "NumComm," "SumVal," "NumInvolv," "NumPEPs," "NumEmpl," and "Status." These variables mean:

- FIR: a unique identifier code for the FIR;
- NumComm: number of communications found in the FIR; communication is a set of transactions conveyed to the Financial Intelligence Unit (FIU) by a bank or an investment broker, for example, whenever they identify suspicious financial transactions;
- SumVal: the sum of the values of all operations of the FIR;
- NumInvolv: number of individuals and legal entities involved in the operations of the FIR;
- NumPEPs: number of politically exposed persons involved in FIR operations, such as heads of state or government, politicians, and high-ranking government officeholders;
- NumEmpl: number of government servants involved in the FIR's operations;
- Status: FIR status in the Federal Police, which may be "archived," "investigated," "under analysis," "forwarded," "not registered," etc.

Below, we can check the status of the 1,427 reports collected in the database.

It is worth noticing that the "Status" field is the decision to be predicted by the models, i.e., it is the "target variable." With that in mind, we separated the records into two datasets:

1. FIRs registered as "investigated" or "archived";
2. Application: FIRs "Under analysis" (yet to predict whether to investigate or archive).

We discarded the "Other" FIRs as they did not present relevant information about their progress. After this disposal, we distributed the remaining 1,366 reports according to Table 2:

Table 1. FIR's according to "Situation"

Status	Amount of FIRs
Archived	390
Investigated	183
Under analysis	793
Other	61
Total	1.427

Table 2. 3 Distribution between Learning and Application Tables.

Dataset	Amount of FIRs
Learning	573
Application	793
Total	1.366

4.2 Learning Table Pre-processing

The Learning dataset, composed of 573 FIRs, had the following initial composition of variables (Table 3).

Table 3. Learning dataset variables

Variable	Type
FIR	–
NumComm	Predictor Variable
SumVal	Predictor Variable
NumInvolv	Predictor Variable
NumPEPs	Predictor Variable
NumEmpl	Predictor Variable
Status	Target Variable

The data from the Learning dataset were uploaded and pre-processed. The first processing was verifying the data types and missing data. Then, we applied functions to ascertain the occurrence of variables with high correspondence and variables with low importance, which could negatively affect the learning of the models.

We used the Pearson method to check the correspondence between the predictor variables; Table 4 shows the results.

Table 4. Correspondence between variables (Pearson method)

Variable	NumComm	SumVal	NumInvolv	NumPEPs	NumEmpl
NumComm	–	0,16	0,45	0,47	0,22
SumVal	0,16	–	0,17	0,09	0,09
NumInvolv	0,45	0,17	–	0,59	0,62
NumPEPs	0,47	0,09	0,59	–	0,36
NumEmpl	0,22	0,09	0,62	0,36	–

We cannot see highly matched variables (above 0.9) in the table above, so no variable was dropped from the dataset.

We used two methods to check for low-importance variables: "chi-square" and "f_classif". Table 5 shows the results.

Table 5. Variable Importance

Variable	chi-square	f_classif
NumComm	7015.59	2.10
SumVal	137,654,803,192.23	2.72
NumInvolv	5893.51	3.59
NumPEPs	1.80	0.25
NumEmpl	30.65	1.57

We can notice that in both the "chi-square" and the "f_classif" method, the variables "NumPEPs" and "NumEmpl" are of low importance when compared to the rest. Therefore, we decided to eliminate these variables not to harm the model. Table 6 shows the remaining dataset variables.

Table 6. Learning dataset after preprocessing

Variables	x/y	Type
NumComm	x	Predictive
SumVal	x	Predictive
NumInvolv	x	Predictive
Status	y	Target

4.3 Learning

The Learning process through Decision Tree, Random Forest, and K-Nearest Neighbors algorithms comprise three stages: training, testing, and model generation. Learning data was separated into training (80%) and testing (20%), as shown in Table 7.

We used the following Python-based classification algorithms/APIs during the training stage.

- DecisionTreeClassifier (package sklearn.tree);
- RandomForestClassifier (package sklearn.ensemble);
- KNeighborsClassifier (package sklearn.neighbors).

Table 7. Learning dataset's training and test data

Dataset	Percentage	Amount
Training	80%	458
Test	20%	115

After the training, we tested the models. Decision Tree presented an accuracy of 0.60, Random Forest an accuracy of 0.70, and K-Nearest Neighbors an accuracy of 0.63. Figure 8 shows a comparative chart.

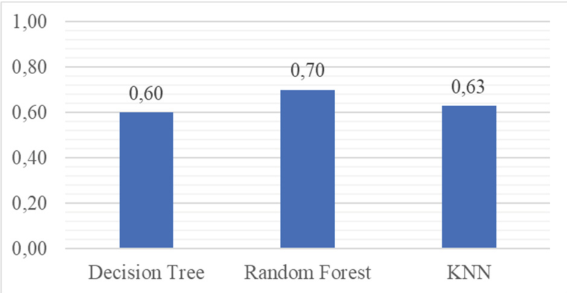

Fig. 8. Accuracy of the tested classification algorithms

In addition to accuracy, another valuable method for evaluating a model training is the confusion matrix, a table generated for a classifier on a data set and used to describe the classifier's performance [10], as shown in Fig. 9.

	Predicted: No	**Predicted: Yes**
Actual: No	TN	FP
Actual: Yes	FN	TP

Fig. 9. Confusion matrix. MARIA NAVIN and PANKAJA [9]

Description of table terms:

- True Positives (TP) - both prediction and actual are yes.
- True Negatives (TN) - prediction is no, and actual is yes.
- False Positives (FP) - prediction is yes, and actual is no.
- False Negatives (FN) - prediction is no and actual is no.

For the present study, the confusion matrix indicates the "True Investigated," "False Investigated," "True Archived," and "False Archived" FIRs, as shown in Fig. 10.

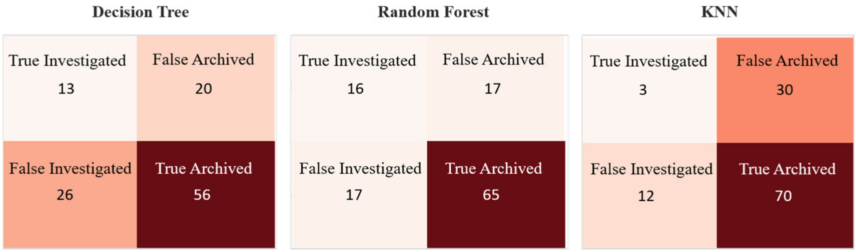

Fig. 10. Confusion matrix the tested classification algorithms

We can see above that the tested algorithms produced a lot of "True Archived" (bottom right), which shows that they got most of the predictions right when deciding to "archive." However, they also produced many "False Investigated" (bottom left), i.e., the model was wrong in many predictions when deciding to "investigate". Random Forest presented the best performance between hits and misses, reaching almost 80% of the "archived" and half of the "investigated."

4.4 Application

We applied the three adequately trained and tested classification models on the 793 FIRs of the Application dataset to determine whether the FIR should be investigated or archived. The Decision Tree recommended that 291 reports be investigated and 502 archived (Fig. 11).

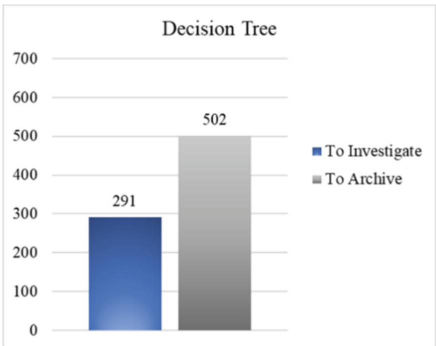

Fig. 11. Decision Tree algorithm recommendation

Random Forest recommended that 271 reports be investigated and 522 archived (Fig. 12).

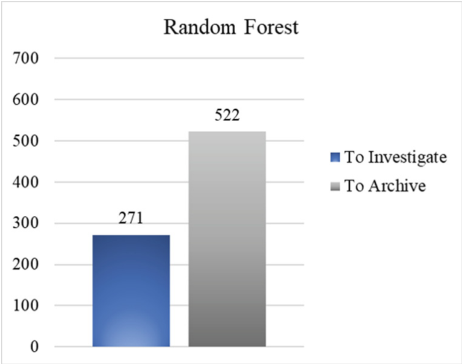

Fig. 12. Random Forest algorithm recommendation

Finally, KNN recommended that 165 reports be investigated and 628 archived (Fig. 13).

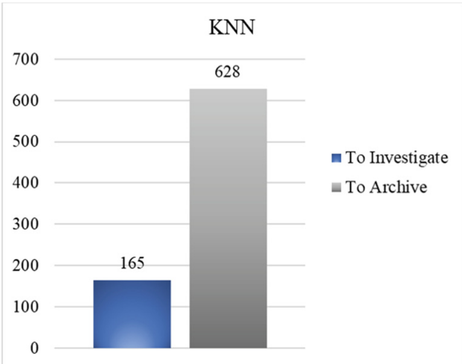

Fig. 13. KNN algorithm recommendation

5 Results

When applying the three algorithms in 793 Financial Intelligence Reports, they recommended that most reports be archived, as shown in Fig. 14.

We can notice that while Decision Tree and Random Forest indicated that about 2/3 of the reports should be archived and 1/3 investigated, KNN recommended that about 4/5 be archived and only 1/5 investigated.

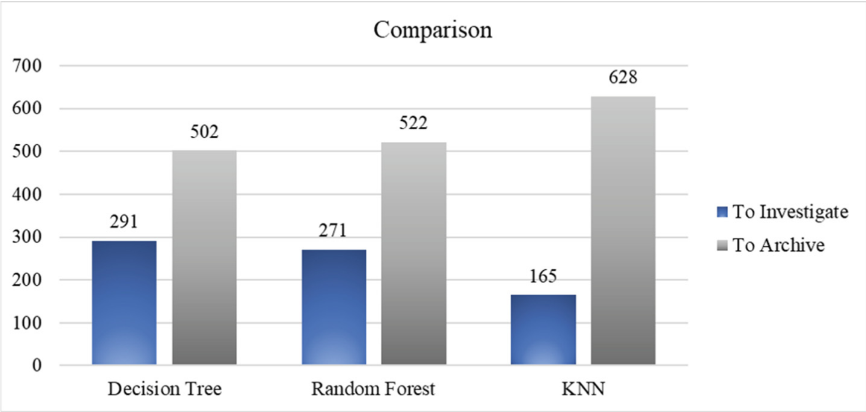

Fig. 14. Comparison of classification models

6 Final Considerations

Money laundering is a crime of complex investigation that usually involves a large set of financial and patrimonial data, including Financial Intelligence Reports (FIR). An FIR is a document generated by the Brazilian Financial Intelligence Unit and disseminated to intelligence agencies for its investigation.

One of those intelligence agencies is the Federal Police, which, after an initial analysis, distributes the reports to police stations throughout the country. Then, they carry out a new analysis to decide if the FIR will be archived or investigated.

This decision to investigate or archive involves the analysis of several variables, such as the number of stakeholders involved, locations, transaction features, etc. This work proposed and applied a machine learning model to process FIRs using the Decision Tree, Random Forest, and K-Nearest Neighbors classification algorithms.

We collected data from 1,427 Financial Intelligence Reports, of which 573 had already been investigated or archived, and 793 had no decision yet. We separate these reports into training and application datasets. In summary, we used the 573 "already labeled" FIRs to train the model and then applied this model to the 793 pending-decision FIRs.

Applying the model in these 793 Financial Intelligence Reports recommended that most of them be archived. While the Decision Tree and Random Forest algorithms recommended that about 2/3 of the reports be archived, the KNN algorithm recommended that 4/5 be archived. As in any statistical analysis, there are advantages and disadvantages in the application of each algorithm, which will vary according to the methodologies used, the computational intensity and complexity of the algorithm [11].

However, these are still early results, which depend on more data, tests, and model calibrations to improve the accuracy and precision of the Learning step. In the next steps of this research, we intend to increase the amount of training data, consider new variables, and test new algorithms, such as Logistic Regression, Support Vector Machine, and Artificial Neural Networks.

References

1. Zaina, R., Ramos, V., Araujo, G.: Triagem Automatizada de Relatórios de Inteligência Financeira. En T.M.R. Dias (Ed.), Informação, Dados e Tecnologia. Advanced Notes in Information Science, volume 2. Tallinn, Estonia: ColNes Publishing, (2022). https://doi.org/10.47909/anis.978-9916-9760-3-6.115
2. Oliveira, T.B.: O bem jurídico-penal no crime de lavagem de dinheiro. Revista Esmat, [S.l.], vol. 4, no. 4, pp. 269–299 (2012)
3. Aras, V.: Sistema nacional de combate à lavagem de dinheiro e de recuperação de ativos. Revista Jus Navigandi, ano 12, no. 1411. Teresina (2007)
4. Das, K., Behera, R.N.: A survey on machine learning: concept, algorithms and applications. Int. J. Innov. Res. Comput. Commun. Eng. **5**(2), 1301–1309 (2017)
5. Mccarthy, J.: What is artificial intelligence. Stanford University (2007)
6. Maimon, O.Z., Rokach, L.: Data mining with decision trees: theory and applications. World scientific (2014)
7. Breiman, L.: Random Forests. Machine Learning, 45, 5–32, Kluwer Academic Publishers. Netherlands (2001)
8. Kirasich, K., Smith, T., Sadler, B.: Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review, vol. 1, No. 3, Article 9 (2018)
9. Altaher, A.: Phishing Websites Classification using Hybrid SVM and KNN Approach. Int. J. Adv. Comput. Sci. Appl. (ijacsa), **8**(6) (2017)
10. Maria Navin, J.R., Pankaja, R.: Performance analysis of text classification algorithms using confusion matrix. Int. J. Eng. Tech. Res. (IJETR) **6**(4), 75–78 (2016)
11. Daizadeh, I.: Trademark and patent applications are structurally near-identical and cointegrated: implications for studies in innovation. Iberoamerican J. Sci. Measure. Commun. **1**(2) (2021). https://doi.org/10.47909/ijsmc.33