Establishing risk and targeting profiles using data mining: Decision trees

Bassem Chermiti

Abstract

The application of technology and the computerisation of management processes in customs administrations have undoubtedly accelerated the processes related to data storage.

In this context, customs administrations possess vast amounts of data on trade and financial flows. Data mining tools can be effective in analysing huge reams of data. Data mining consists of understanding, preparing, modelling and analysing data using different techniques, such as machine learning. Many of these techniques have advanced predictive analytical capacities that can ultimately lead to improved analytical capabilities in risk management.

The Chi-square Automatic Interaction Detector (CHAID) decision tree method was selected for the purposes of this paper to determine the customs risk factors associated with import declarations recorded in the customs clearance system. The CHAID method is also used to create risk profiles and predict non-compliant customs declarations based on established rules.

1. Introduction

In contemporary society, data processing has the potential to improve the quality of management decisions. Moreover, analyses of economic activity over recent decades has shown the impact of technology (such as the computerisation of almost all trading processes), which has led to increased storage of a considerable amount of data.

Customs administrations have a large amount of data on trade and financial flows. However, the quantity of data available is less important than what the administrations do with it. Only robust analyses can make this data useful and usable in the decision-making processes.

Cognisant of this importance, the World Customs Organization (WCO) dedicated the year 2017 to data analysis and used the slogan ‘Data analysis for Effective Border Management’. The aim was to encourage member countries to further promote their efforts and activities in a vital and indispensable sector of the customs modernisation process: data collection and analysis.

Data mining enables the examination and exploitation of stored data. It has developed as a multidisciplinary approach capable of analysing a large amount of data and identifying significant models.

Data mining has developed very rapidly with the support of technologies and sciences, such as statistics, artificial intelligence and machine learning. Indeed, there is a wide range of data mining methods that can be used to solve multiple issues, some of which have been used for risk management purposes in the customs domain (Geourjon, Laporte, Coundoul, & Gadiaga, 2012).

In recent years, machine learning and artificial intelligence have garnered significant interest and popularity, particularly in the financial field, with the expectation that their introduction will result in an improvement in analytical capabilities, including risk management.

In the customs context, risk management is one of the key measures contained in the World Trade Organization's (WTO) Trade Facilitation Agreement (TFA), and in the WCO's Revised Kyoto Convention (RKC). In operational terms, customs risk management is an effective way to handle trade flows with a guarantee of trade fluidity as only the most risky goods are targeted. The objective is to obtain controls adapted to the risk profiles that are determined.

In order to isolate risky shipments, several methods have been implemented by customs that are largely based on the exchange of information, in-depth analysis of fraud trends and the review of available information on traffic flows and trade patterns.

In this paper, we will use machine learning as a tool for identifying and analysing customs risks. We will also use the decision tree method, which is particularly suitable for data mining and enables the development of predictive models.

The CHAID (Chi-Square Interaction Detector) method is one of the most common decision tree algorithms and features among the most popular data mining techniques. Among the many researchers who have used this method are Ôcal, Ercan and Kadioglu (2015), who used the CHAID method to predict the financial crisis, and Koyuncugil and Ozgulbas (2017), who used it for financial profiling and operational risk detection.

This method is used to uncover new information in large databases, to detect unspecified interactions between variables, and to create predictive models.

Our study was conducted using data obtained from a partner customs administration in the framework of a seminar organised by the WCO to highlight the utility of data analysis in the customs domain.

This study is composed of two sections. The first section focuses on data mining, particularly machine learning, to leverage stored information and the decision tree method as explored through the CHAID algorithm, which we will apply in the context of customs risk management.

The second section focuses on the empirical study of data sourced from the partner customs administration in an effort to develop a predictive model for detecting non-compliant import declarations and risk profiles based on detected risk factors.

2. Data mining, machine learning and decision trees

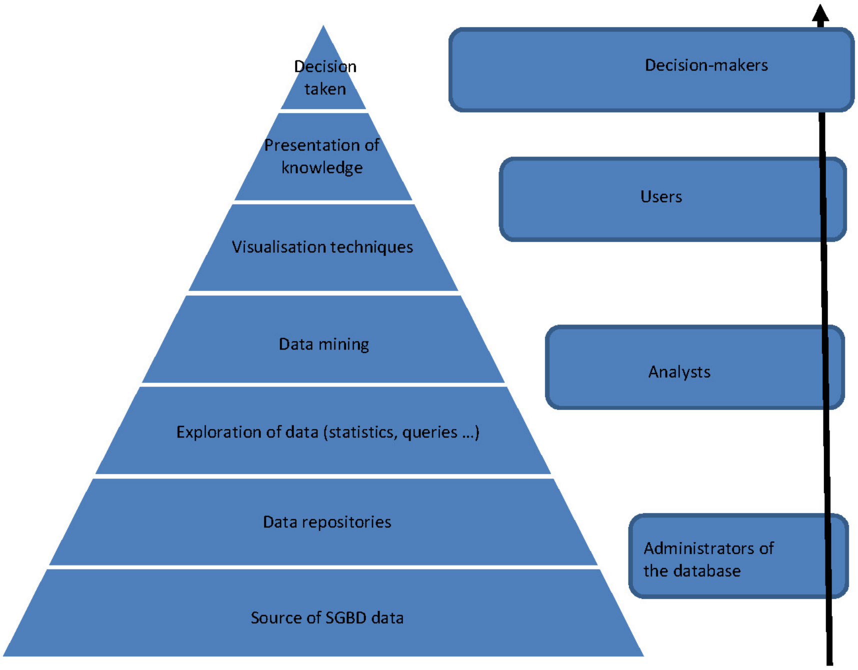
2.1 Data mining

'Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge driven decisions' (Thearling, 2003).

Data mining is a combination of computer and statistical techniques designed to perform an exploratory data analysis. These techniques have been applied in several areas, such as fraud detection, tourism, industry, marketing, finance and customs. From the flow of raw information to underpinning relevant decision-making, data mining is an appropriate tool that enables data analysis.

In this context, the pyramid below outlines the responsibilities of the actors in the decision-making process. The database administrator designs, manages and administers database management systems and data warehouses, while analysts extract information by running queries and using various techniques, including data mining, to assign meaning to the data stored in the databases. The results are intercepted by users, who in turn share the information with decision makers to allow for optimal decision-making.

Figure 1: Information flow and participants



Source: Rakotomalala Course, 'Introduction to Data Mining', University Lumière Lyon 2.

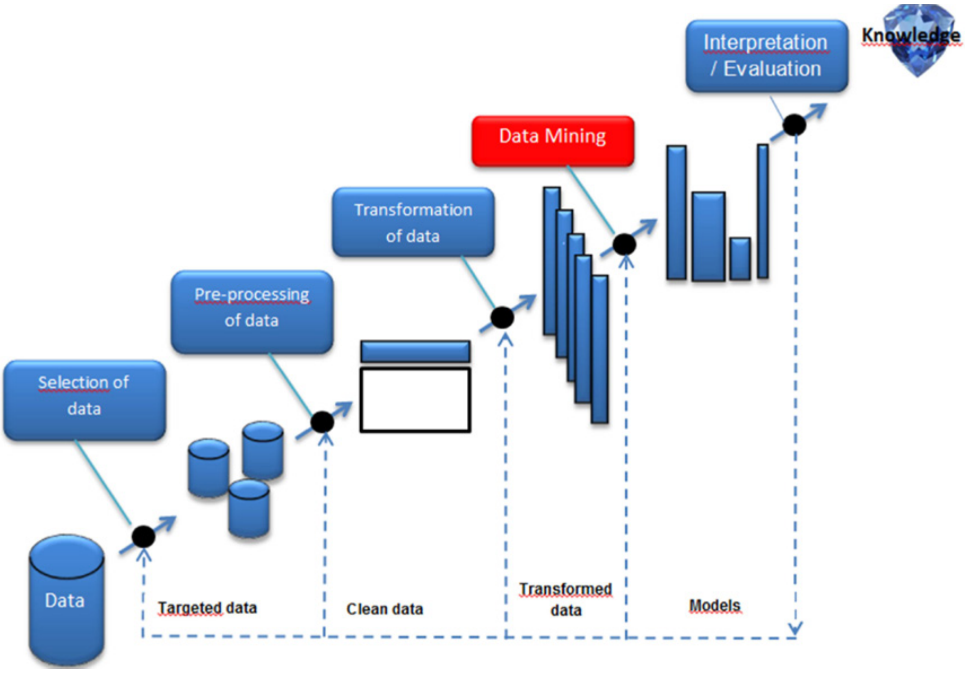
Data mining is therefore at the centre of the process of uncovering information contained in databases. According to Fayyad, Piatetsky-Shapiro and Smyth (1996), this process begins with the selection of data, which is a two-step process: first, it is crucial to develop and understand the application domain and the existing information; second, a set of target data is created from which the discovery will be made. The result is the acquisition of target data.

Data pre-processing is the second step in this process and involves processing noisy and missing data. The goal is therefore to ensure that the process model of information discovery in databases produces accurate results. The legacy of this step is the acquisition of cleaned or pre-processed data.

The transformation of data represents the final step. This is the final stage of data processing before data analysis techniques are applied. It consists of finding useful attributes by applying dimension reduction and transformation methods and finding an invariant representation of the data; the result is transformed data. This phase is followed by the most important step, data mining, which consists of choosing the appropriate algorithms or methods, and matching the specific objectives with these methods (regression, classification, trees, grouping, etc.) in order to find and apply appropriate data mining prototypes. This final step results in models.

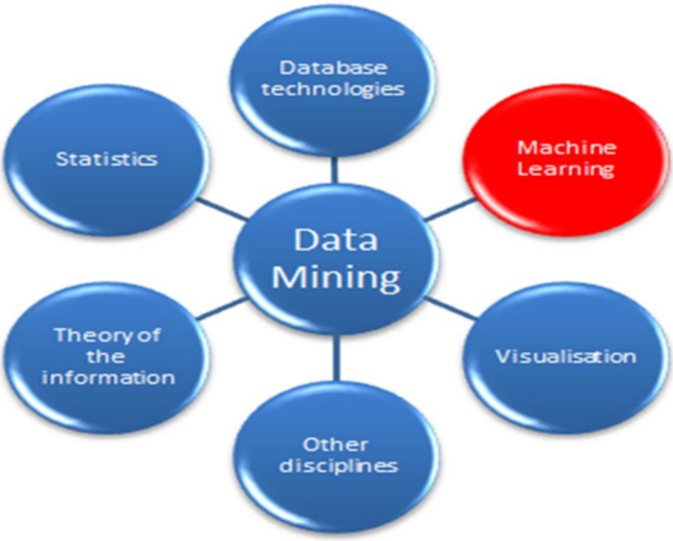
This process ends with an interpretation and evaluation phase, which outlines the information uncovered. If the result obtained is not considered significant, a new iteration is necessary. The interpretation and evaluation phase can also feature visualisations of the extracted models. The information then needs to be consolidated by integrating it into the performance system, or simply documenting it and reporting it to the appropriate units. This step can include checking and resolving any potential conflicts with information previously created. The outcome is knowledge.

Figure 2: Process of knowledge discovery Fayyad, Piatetsky-Shapiro and Smyth (1996)



Note also that data mining differs from conventional statistical techniques in its artificial intelligence capacity. It uses several techniques and technologies, such as statistics, machine learning techniques and SQL query language.

Figure 3: Data mining techniques



3. Machine learning and decision trees

3.1 Machine learning

Machine learning is a method used in data mining. It consists of algorithms that analyse a set of data in order to deduce rules constituting new knowledge and to analyse new situations.

This method is capable of analysing vast volumes of data, while providing in-depth predictive analysis. Perhaps as a result, machine learning and artificial intelligence have received unparalleled attention in recent years.

However, administrations that possess a large amount of data clearly need powerful analytical tools to manage that data. Machine learning is widely regarded as a technique that can provide this analytical power to model complex, non-linear relationships.

Machine learning includes a range of analytical tools that can be classified as ‘supervised’ and ‘unsupervised’ learning tools. Supervised machine learning involves the creation of a statistical model to predict or estimate a result based on one or more inputs (in our case this article predicts the non-compliance of a customs declaration registered in the IT system of the partner administration in accordance with several variables or risk factors).

In unsupervised learning, a set of data is analysed with no dependent variables to estimate or predict. Instead, data is analysed to show patterns and structures in a dataset.

Machine learning can also be a particularly powerful tool when it comes to forecasting. Certainly, because of its ability to process a large dataset and its computational power, machine learning is closely associated with the ‘Big data revolution’.

According to Thearling (2003), the most used technique in data mining is decision trees, which are tree-like structures representing sets of decisions that generate rules for classifying a dataset. Specific decision tree methods include classification, regression (CART) and automatic detection of interactions with chi-square (CHAID). There are also artificial neural networks that represent another data mining technique that enables complex problem solving by adjusting weighting coefficients in a learning phase. There are also genetic algorithms—an optimisation technique that uses processes such as genetic combination, mutation and natural selection in a model that is centred on the concepts of evolution. An additional method, nearest neighbour, groups each record into a set of data based on a combination of the most similar classes of k records contained in a group of historical data. This technique is also called k-nearest neighbour (k_NN).

In the context of this study, we use the decision tree technique, specifically the CHAID method.

3.2 Decision trees

The decision tree is a non-parametric supervised learning method used for classification purposes and for the development of predictive algorithms. The objective in using decision trees, in this article, is to create a model that predicts the value of a target variable by learning simple decision rules derived from the characteristics of the data.

Decision trees are constructed by seeking, through the successive fragmentation of the training set, partitions in the space of the optimal predictors capable of predicting the modality of the response variable. Each rupture is done in accordance with the values of a predictor. During the first step, all the predictors are tested in order to identify which are best. Then the process is repeated at each new node until a stop criterion is satisfied. The determination of the best rupture at each node is made in accordance with a local criterion. The choice of criterion is the main difference between the various existing methods of tree induction. Among the most frequently used criteria are:

- Shannon's entropy, applicable to all types of explanatory variables. This measure is used by Quillan in C4.5 and C5.0 to measure uncertainty:

  $$Entropy\ (node_{\ t}) = \sum_{i=1}^{k} - f_i \log_2 f_i$$

  Where $f_i, (i = 1, \ldots, p)$ are the relative frequencies in the node t of k classes to predict.

- The CART algorithm produces binary decision trees and applies the Gini index, called quadratic entropy, to select the explanatory variables of any type.

  $$Gini\,(node_{\ t}) = \sum_{i=1}^{k} f_i(1 - f_i) = 1 - \sum_{i=1}^{k} f_i^{\ 2}$$

- The CHAID method, outlined below.

3.3 The CHAID method

CHAID is one of the oldest classification tree methods and was proposed by Kass (1980). Unlike other tree algorithms, the CHAID method can build non-binary decision trees.

CHAID modelling is an exploratory data analysis method used to study the relationships between a dependent measure and a large set of possible predictor variables that may interact with one another. The dependent measure may be qualitative (nominal or ordinal) or quantitative.

For qualitative variables, a series of chi-square analyses is performed between dependent and predictive variables.

For quantitative variables, variance analysis methods are used when intervals (disaggregation) are optimally determined for independent variables to maximise the ability to explain a dependent measure in terms of components of variance (Thearling, 2003).

The Chi-square test value is calculated using the formula:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - previj)}{previj}$$

$$\text{or } previj = \frac{n_{i.} \times n_{.j}}{n}$$

- n_{ij} is the number in the box marked by row i and column j
- $n_{i.}$ is the marginal size of line i
- $n_{.j}$ is that of column j
- n is the total number (the size of the population)
- previj: the expected strength for the cross-analysis field in row i and column j.

The theoretical test is then applied with a level of significance chosen by the user to check whether there is independence between the crossing of the two variables.

The decision tree consists of:

- Root node: The root node contains the dependent or target variable. For example, in our case, we want to predict the non-conformity of a declaration according to details such as the origin of the goods, the quantity and the declared value, the risk of non-conformity is the target variable and the factors remaining are the predictor variables.
- Parent nodes: The algorithm divides the target variable into two or more categories. These categories are also known as initial nodes.

- Child nodes: The categories of independent variables that are below the parent categories in the CHAID tree.
- Terminal nodes: The last categories of the CHAID analysis tree. In the CHAID analysis tree, the category that has a major influence on the dependent variable comes first, and the least important category comes last.

The process of building a CHAID decision tree is divided into three parts:

1. Select the relevant independent variables from the input variables. The first variable selected to divide the data is the variable with the lowest p-value and therefore the most strongly associated with the dependent variable. By applying the hypothesis test, if the value p is equal to or less than the level of significance α predefined, then the alternative hypothesis, which suggests a dependency between the variables, is accepted. Otherwise, the node is considered as the terminal node. The construction of the tree stops when the p-values of all observed independent variables are greater than a certain fractionation threshold. In the case of a quantitative dependent variable, an ANOVA F test is used to compare the means of the dependent variable for each of the categories of the explanatory variable used for the separation.

2. Merge the pairs of independent variable values that are the least different from the dependent variable. We use a distributional equivalence test of Khi-2 for this purpose. If the value p obtained is greater than a certain threshold of fusion, the algorithm merges particular categories without statistically significant differences.

3. Search for a new merge pair until the pairs, for which the value p is less than the defined level of significance α, are not identified.

In principle, because of the nature of data mining analysis, the application of the CHAID method requires the use of large samples.

Decision trees, and especially the CHAID method, have many advantages, including that they:

- are non-parametric; no assumptions on the distribution of the data are postulated
- can handle a large number of variables and they allow automatic selection of relevant variables
- can handle large volumes of heterogeneous data (categorical or numerical), resulting in reduced computing times
- are robust against outliers and offer a solution for missing data; input quality issues can be detected thus avoiding the construction of an invalid model based on poor quality data
- provide results that are visual and simple to interpret: the shape of a CHAID tree is intuitive, it can be expressed in the form of a set of explicit rules, in fact the paths summarising decisions transcribed in the form of rules (if ... therefore) are understandable and therefore easily interpretable
- are easy to integrate into existing IT processes due to their high level of automation and the ease of translating decision models into SQL for relational database deployment.

The contribution of this article is manifested by the application of the CHAID algorithm in the context of customs risk management in the detection of non-compliant revenue declarations.

4. Decision trees and the creation of customs risk profiles

4.1 Risk management and control channels

To deal with operational risks, the partner customs administration—like other customs administrations—is making efforts to improve its ability to target high-risk shipments. Indeed, this administration has two entities in charge of risk analysis: a private company that is the result of a public-private partnership and the Intelligence and Risk Analysis Service (SRAR). The private company uses a tool called Profiler, which is based on a scoring method to determine the risks, and then passes the profiles to the SRAR service that adds its own profiles for the purpose of transmitting them daily to the offices.

Regardless of the efficiency of the selectivity system of the partner customs administration, in this document we will use the CHAID algorithm to determine a predictive model to predict the 'revenue' risk associated with an import operation to detect risk factors and develop risk profiles.

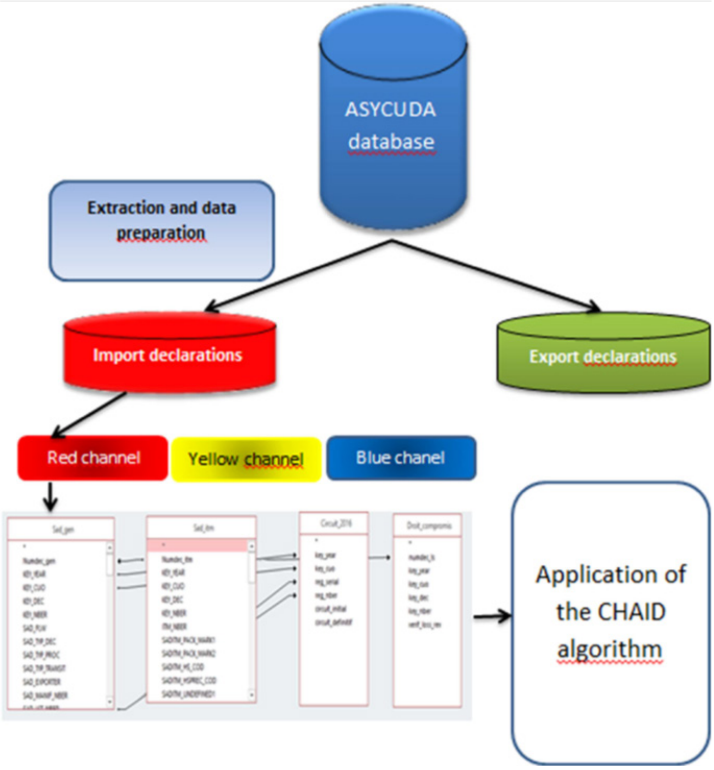
4.2 Modelling

To apply the CHAID procedure, we will present the data, variables, assumptions and specifications.

4.2.2 Data

The study was carried out with data extracted from the customs information system of the partner administration. We have established our study base on the import declarations registered in the partner customs clearance system for the year 2016.

Figure 4: Diagram of the flow of data



The first step is to select the data; this step decides what data will be used for the analysis. The data is collected in the form of four tables:

1. Sad_gen: Declaration header table
2. Sad_item: Table lines of declarations
3. Channel_2016: Declaration channels table
4. Laws_compromised: Laws compromised table.

The customs selectivity system of the partner administration, like that of many customs administrations around the world, directs all customs declarations to the blue, yellow or red channel. Import declarations in 2016 are set out in Table 1.

Table 1: Assignment of import declarations per channel

Channel	**Number of declarations**	**Percentage %**
Blue	31,494	25.31
Yellow	65,679	52.79
Red	27,253	21.90
Total	124,426	100.00

In our study, we will focus on the red channel as we have feedback on the compliance levels associated with these declarations because of physical inspections. It should be noted that the number of declarations found to be non-compliant within the red circuit is estimated at 1,440 declarations (5.28%).

4.2.2 Variables

The CHAID algorithm is developed on the basis of two groups of variables: the target variable (non-conforming declaration) and the predictor variables that will explain the target variable. These variables constitute the risk factors that may explain the non-conforming of a declaration.

Predictive variables or risk factors are presented Table 2.

Table 2: List of variables contained in the model

Variables	Type of variable
Reporting firm code	Chain
Consignee code	Chain
Type of customs procedure	Chain
Container (yes or no)	Chain
Customs procedure applied	Chain
Type of payment	Chain
Mode of transport	Chain
Delivery method	Chain
Active methods of transport	Chain
Last shipment country	Chain
Export country code	Chain
Nationality of the mode of transport	Chain
Place of unloading	Chain
Country of origin	Chain
Currency of invoice	Chain
Total value	Numerical
Total taxes	Numerical
Total gross weigh	Numerical
Total net weight	Numerical
Value by kilo of declaration	Numerical
Tax burden declaration	Numerical
HS2	Chain
HS4	Chain
HS6	Chain
HS8	Chain
Gross weight article	Numerical
Net weight article	Numerical
Value article	Numerical
Taxes per article	Numerical
Value per kilo of article	Numerical
Tax burden article	Numerical

4.2.3 Assumptions and specifications of the model

In the database received from the partner administration, several components related to the non-conforming of customs declarations are not communicated or are unavailable for reasons such as litigation or disputes.

This lack of information leads to the following assumptions:

H1. Any declaration that has been subject to an additional liquidation is deemed non-compliant.

H2. If a line of the declaration is found to be non-compliant, then all lines of the declaration are non-compliant and the additional amounts associated with this declaration count as many times as the lines of the declaration.

H3. A line of the declaration is treated as a declaration.

H4. The study covers all the import declarations oriented towards the red channel.

H5. A 10 per cent risk threshold is set in order to consider a declaration as non-compliant.

The specifications of the model are[1]:

S1. Number of observations per parent node is set at 1000 observations.

S2. Number of observations per child node is fixed at 500 observations.

S3. The Pearson Chi-square test is chosen and a common level of significance ($\alpha = 0.05$) for the division of the nodes and the merger of the categories of independent variables is determined.

S4. To evaluate the model we proceeded to cross-validation (k-fold cross-validation). We selected k =10 as a value, which is very common in the field of applied machine learning (Vercellis, 2009). Cross-validation consists of dividing the sample into k sub-samples of approximately the same size. The trees are generated by excluding the data from each sub-sample. The first tree is based on all observations except those in the first sub-sample, these data are the training samples and the sub-sample is the validation or test sample. The second tree is based on all observations, except those of the second sub-sample, and so on.

The overall accuracy is calculated as the arithmetic mean of k individual accuracies. The advantage of this method is that all observations are used for both learning and validation, and that each observation is used for validation precisely one time.

4.3 Results and interpretations

The results of the CHAID procedure (shown in Table 3 and Figure 6) indicate that the model created contains five levels of tree depth, for a total of 145 nodes, of which 99 are terminals. In addition, out of a total of 31 independent variables, the final model contains 15 variables, while the other variables are not statistically significant from a compliance perspective.

Table 3: Results

Total number of nodes	**145**
Number of terminal nodes	99
Depth	5
Number of variables included in the model	15

Figure 5: Overview of the decision tree

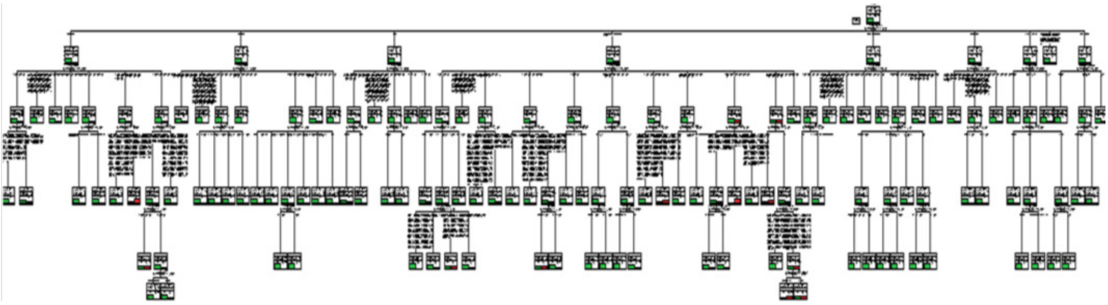
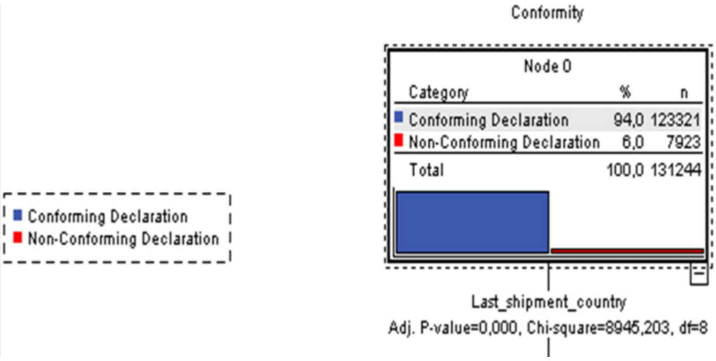


Figure 6: Root node



Conformity

Node 0

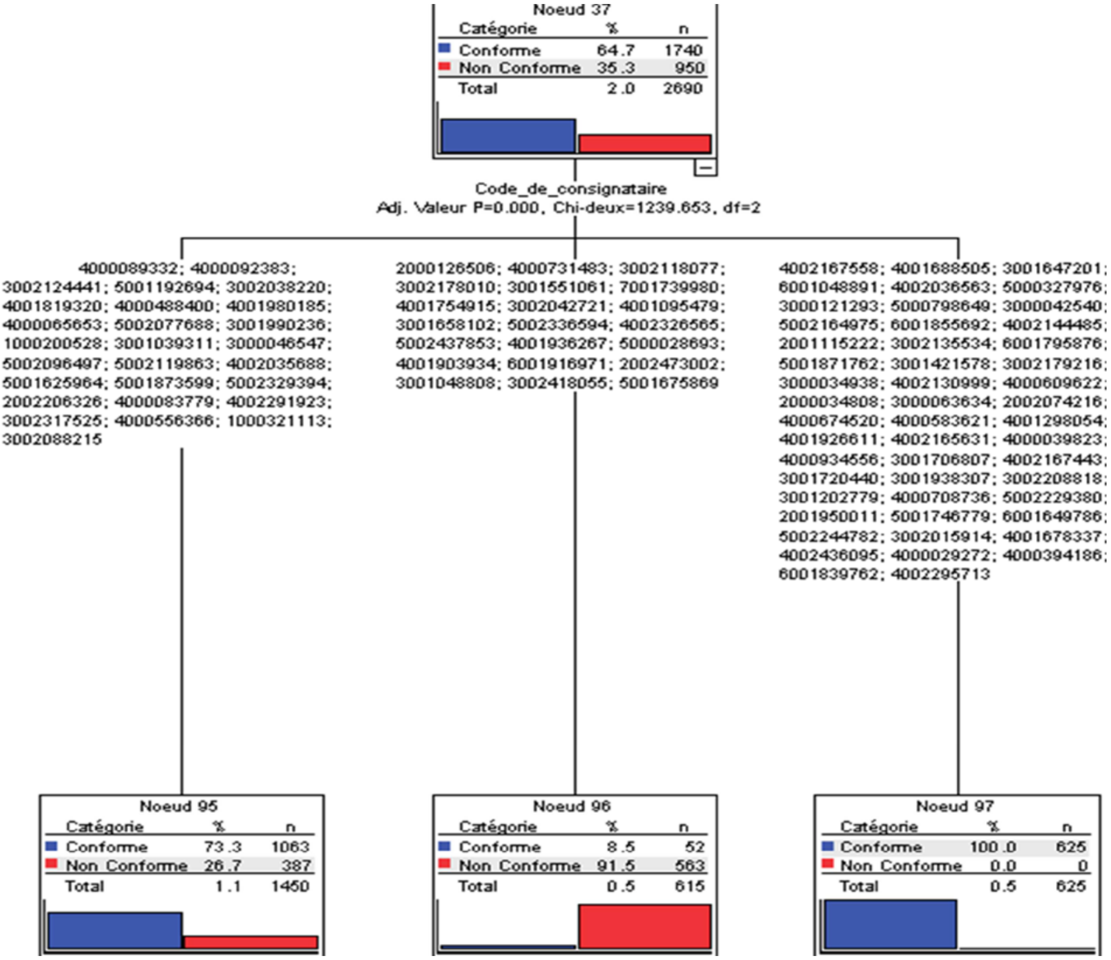
Category	%	n
🟦 Conforming Declaration	94,0	123321
🟥 Non-Conforming Declaration	6,0	7923
Total	100,0	131244

🟦 Conforming Declaration
🟥 Non-Conforming Declaration

Last_shipment_country
Adj. P-value=0,000, Chi-square=8945,203, df=8

Figure 7 shows the root node that contains 131,244 lines of red-oriented declarations including 7923 non-conforming lines (6% of all lines).

Figure 7: Part of the decision tree



Referring to the decision tree, the last shipment country had the most significant effect on the conformance of a reporting line, which means that it is the variable most strongly associated with the dependent variable and that it has the most power in the division of observations into groups. In other words, for the observed data, this variable has the greatest potential to differentiate and classify the lines of clearance declarations into two groups (compliant and non-compliant), the statistical significance of the variable was determined, with α = 0.05 using the following values: ($\chi 2 = 8945.203$, df = 8, p value = 0.000). As the first discriminator, it divides the root node, that is, a number of 131,244 declaration lines into eight groups containing different categories of the variable 'last shipment country'. The second-best predictor is the 'reporting firm code' for some parent nodes. The discrimination variables used, other than the variables mentioned above, are:

Consignee code	Country of origin	Total value
Net weight total	Method of payment	Article tax burden
Tax burden article	Nationality of the methods of transport	Article net weight
Mode of delivery	Customs procedure applied	Article value
Taxes per article	Declaration tax burden	

The table below (Table 4) shows the gain values that provide information about target categories (non-compliant reports). This table is only available if one or more target categories are specified. In our example, there is only one target category (non-compliant declaration). Only one gains table for the nodes is generated.

Table 4: Gains of the nodes

Node	Node		Gain		Response %	Index %
	N	Percentage %	N	Percentage %		
96	615	0.5	563	7.1	91.5	1516.4%
62	534	0.4	406	5.1	76.0	1259.4%
98	723	0.6	501	6.3	69.3	1147.9%
143	1,095	0.8	685	8.6	62.6	1036.3%
92	594	0.5	328	4.1	55.2	914.7%
. . .						
18	18,766	14.3	0	0.0	0.0	0.0%
41	7,560	5.8	0	0.0	0.0	0.0%
25	4,259	3.2	0	0.0	0.0	0.0%
94	4,045	3.1	0	0.0	0.0	0.0%
82	3,423	2.6	0	0.0	0.0	0.0%

The column ‘Gain N’ represents the number of observations in each terminal node of the target category. The gain percentage is the ratio of the number of observations of the target category to the total number of observations of this modality, namely the number and percentage of observations displaying the lines of non-conforming declarations in our study.

This table, therefore, contains all the terminal nodes. For each of the terminal nodes, there is a decision rule. A decision rule is an expression in the form ‘If condition 1 ... Then ...’.

As indicated above, the terminal node 96 contains 615 lines of declarations, that is, 0.5 per cent of the total lines of declarations; the node 62 contains 534 lines.

The gain for node 96 is 563 lines of declarations, representing 7.1 per cent of all earnings and a response percentage of 91.5 per cent. The response percentage is calculated by dividing the gain by the total number of observations per node. This percentage is interpreted as follows:

- If a declaration line has the same characteristics of the lines belonging to node 96, then there is a 91.5 per cent probability that this line belongs to a non-conforming declaration.
- However, the node 18 contains 18,766 lines of the declarations, that is to say 14.3 per cent of the total lines of declarations, with a response rate equal to 0 per cent. Otherwise, the physical control of the declarations containing these lines of declarations did not result in anything.

The model development process is not complete until its performance has been evaluated. Tables 5 and 6 present basic information on the performance of the developed model in terms of accuracy and predictive potential.

Table 5: Risk (Response rate >=10%)

Method	**Estimate**	**Standard error**
Re-substitution	0.180	0.002
Cross validation	0.206	0.002

Table 6: Classification (Response rate >=10%)

Observed	**Predicted**		
	Conforming	**Non Conforming**	**Percentage correct**
Compliant	107,786	15,535	87.4%
Non-compliant	812	7,111	89.8%
Overall percentage	82.7%	17.3%	87.5%

The results in the classification table (Table 6) show that the model correctly ranks 87.5 per cent of the reporting lines, taking into account a risk ratio of 10 per cent.

- The percentage of correct classifications of the lines of the conforming declarations is equal to 87.4 per cent.
- The percentage of correct classifications of the lines of non-compliant declarations is equal to 89.8 per cent.

Table 5 presents the risk of prediction as a percentage of observations classified incorrectly. The risk estimate is 0.180; the risk of misclassification of a reporting line is approximately 18 per cent, while the average risk of misclassification using cross-validation is 20.6 per cent.

Note that the risk estimate and the overall correct classification rate are no longer moving in the same direction. With an overall correct classification rate of 87.5 per cent, the risk estimate should be 0.12—that is due to the cost of misclassifying high-risk declarations (10% risk ratio) that makes the interpretation less obvious.

In general, the best measure of the model's performance is not its raw accuracy, but its usefulness and effectiveness in achieving the main purpose for which it was created in order to solve a specific problem.

In addition, by grouping the lines by declaration, the model selects 16.70 per cent of the total import declarations directed toward a red channel to detect 80.35 per cent of the total non-conforming declarations in this channel.

However, by applying the model on all import declarations, the model targets 16,227 in total, including 11,072 from the yellow channel, 603 from the blue channel and 4,552 from the red channel, thus the physical control rate decreases from 21.90 per cent to 13.04 per cent.

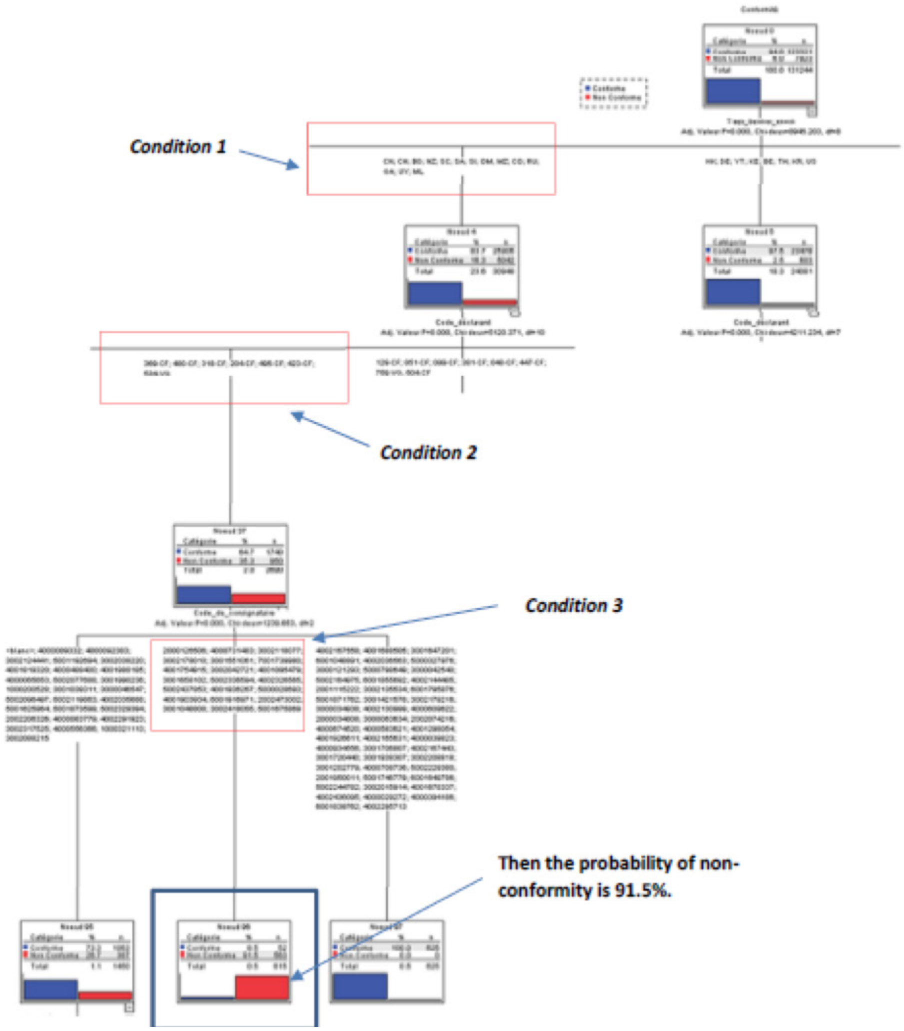
In terms of forecasting, considering a risk rate of 10 per cent, the number of targeting rules in this case is 32. Other rules that include terminal nodes can be used to direct declarations to other control channels (blue and yellow).

A problem may arise in the case where a declaration is unclassified, then it may be considered because of a lack of prior information. 'IF' rule 1 'OTHERWISE' rule 2 'OTHERWISE' rule 3 '... OTHERWISE classified by default (red channel)'.

4. 4 Creation of risk profiles

The paths of a decision tree are represented as 'if-then' rules. For example, 'if condition 1 and condition 2... and condition k occur, then result j occurs'. This is illustrated in Figure 8.

Figure 8: Example that shows how a rule from the decision tree is interpreted

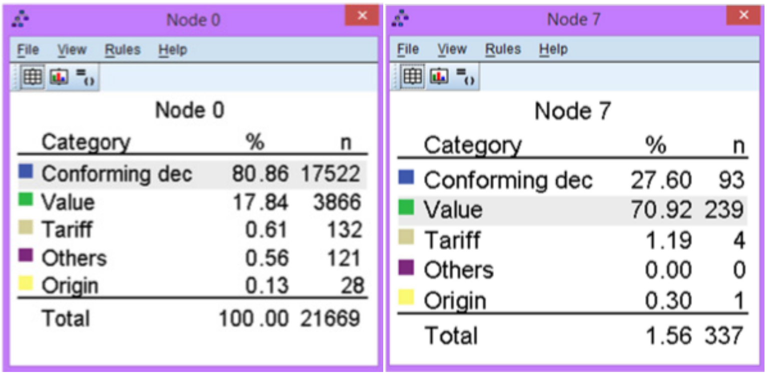


As mentioned in the previous section, reading the decision tree is easy. Indeed, a simple vertical reading of the tree going from the root node to the terminal node allows for rules to be extracted, as shown in the Figure 8 above.

The rules can be translated as SQL instructions and can easily be entered into the customs clearance system.[2] Appendix A provides an example of the way the rules or risk profiles can be written.

Moreover, the CHAID method's ability to generate non-binary trees is particularly interesting. Risk profiles can be determined depending on the nature of the offence, and of course depending on the information available in customs databases, as shown in Figure 9.

Figure 9: Example risk profiles



Node 0

Category	%	n
🟦 Conforming dec	80.86	17522
🟩 Value	17.84	3866
🟫 Tariff	0.61	132
🟪 Others	0.56	121
🟨 Origin	0.13	28
Total	100.00	21669

Node 7

Category	%	n
🟦 Conforming dec	27.60	93
🟩 Value	70.92	239
🟫 Tariff	1.19	4
🟪 Others	0.00	0
🟨 Origin	0.30	1
Total	1.56	337

5. Conclusion

In this paper, we presented a supervised learning method and applied it to customs data. We used the CHAID algorithm to develop a risk model.

This method is considered among the best techniques for creating visual and easily interpretable models, allowing for the establishment of a series of rules, and consequently creating risk profiles and classifying customs declarations of goods using such profiles.

We sought to extract the most relevant risk factors, such as the country of last shipment and the declarant, evaluated the model by the classical 'cross-validation' method, and showed how easy it is to derive rules from the results obtained.

The application of this algorithm can be extended to solve other problems. However, this decision tree method does contain some limitations, sometimes related to the large number of categories that renders them complex and illegible. Furthermore, it is very difficult to identify an optimal decision tree.

Acknowledgements

I would like to thank Mr Thomas Cantens, Head of the WCO's Research Unit; Mrs Antsa Rakotoarisoa of the customs administration of Madagascar; and Mr Christopher Grigoriou, Mr Mourad Arfaoui and Rachel McGauran.

References

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery: an overview.* AI Magazine, 17(3), 37–54.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). *Advances in knowledge discovery and data mining*. Cambridge: AAAI Press.

Financial Stability Board. (2017). *Artificial intelligence and machine learning in financial services: market development and financial stability implications.* Retrieved from http://www.fsb.org/wp-content/uploads/P011117.pdf

Geourjon, A-M, Laporte, B., Coundoul, O., & Gadiaga, M. (2012). *Contrôler moins pour contrôler mieux: L'utilisation du data mining pour la gestion du risque en douane.*

Gilbert, R. (2010). *CHAID and earlier supervised tree methods*. Cahiers du département d'économétrie, Faculté des sciences économiques et sociales, Université de Genève.

Huang, C-S, Lin, Y-J, Lin, C-C. (2008). Implementation of classifiers for choosing insurance policy using decision trees: a case study. *WSEAS Transactions on Computers, 10*(7), 1679–1689.

IBM. (2012). IBM SPSS Decision Trees 21. Retrieved from http://www.sussex.ac.uk/its/pdfs/SPSS_Decision_Trees_21.pdf

Janine, O. (1999). *Neural networks versus CHAID*. A White paper from smart FOCUS.

Koyuncugil A. S., & Ozgulbas N. (2017). *Financial profiling for detecting operational risk by data mining*. Retrieved from http://www.aabri.com/OC09manuscripts/OC09117.pdf

Koyuncugil A. S. & Ozgulbas N. (2009). Risk modeling by CHAID decision tree algorithm. *ICCES, 11*(2), 39–46.

Kumar, D., Krishna, V., Suresh, K., Jnaneswari, Ch., Lakshmi Kranthi, G., Anilpatro. (2012). Discovering hidden values in data warehouse with predictive data mining. *International Journal of Computer Science and Information Technologies*, *3*(4), 4794–4797.

Mitchell, T. M. (1997). *Machine learning*. Boston: WCB/McGraw-Hill.

Neville, P. G. (1999). *Decision trees for predictive modeling*. SAS Institute Inc.

Öcal, N., Ercan, M. K., & Kadioglu, E. (2015). Predicting financial failure using decision tree algorithms: an empirical test on the manufacturing industry at Borsa Istanbul. *International Journal of Economics and Finance, 7*(7), 189–206.

Rakotomalala, R. (n.d.) *Introduction to data mining.* Retrieved from http://eric.univlyon2.fr/~ricco/cours/slides/Introduction_au_Data_Mining.pdf

Rakotomalala, R. (1997). *Arbres de Décision*. Laboratoire ERIC Université Lumière Lyon 2.

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: theory and applications*. New Jersey: World Scientific.

Thearling, K. (2003). *An introduction to data mining: discovering hidden value in your data warehouse*. Retrieved from http://akira.ruc.dk/~bulskov/undervisning/E2003/data_mining.pdf

Thearling, K. (2003). *An overview of data mining techniques.* Retrieved from http://akira.ruc.dk/~bulskov/undervisning/E2003/data_mining.pdf

Tufféry, S. (2011). *Data mining and statistics for decision making*. Chichester: John Wiley.

Udrea, C., Fadila, B., Darmont, J., & Boussaid, O. (2004). *Intégration efficace de méthodes de fouille de données dans les SGBD*. ERIC – Université Lumière Lyon 2.

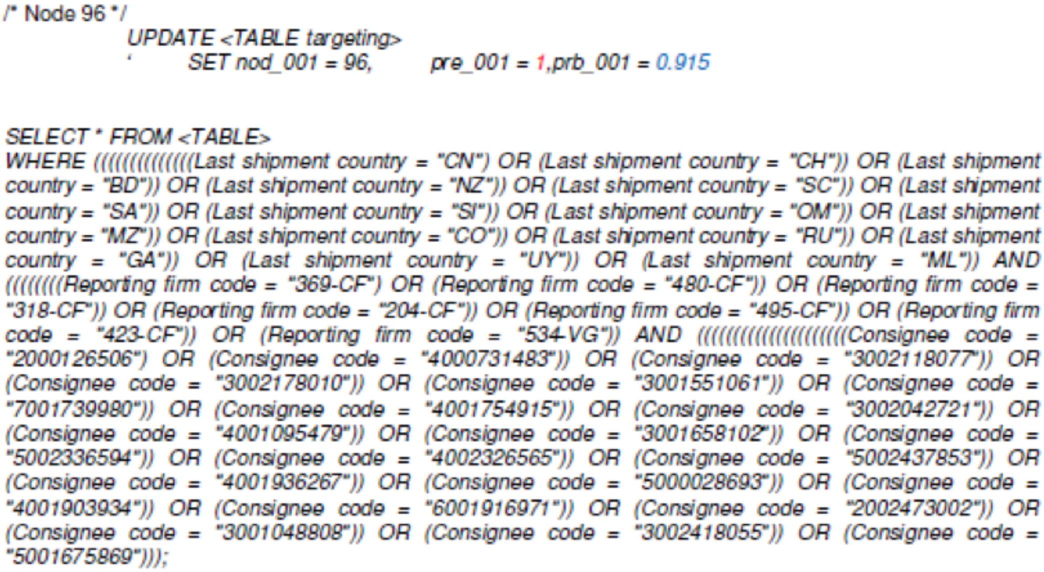
Vercellis, C. (2009). *Business intelligence: data mining and optimization for decision making*. John Wiley & Sons.

Notes

[1] We used the SPSS V 21.00 programme, which allows for the implementation of a decision tress process—the CHAID method. However, it is also possible to use open source software such as R or Python.

[2] SPSS software enables the extraction of rules in the form of SQL instructions.

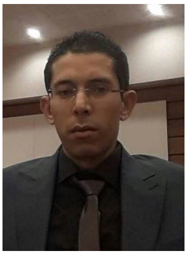
Appendix A



```
/* Node 96 */
            UPDATE <TABLE targeting>
            '      SET nod_001 = 96,      pre_001 = 1,prb_001 = 0.915


SELECT * FROM <TABLE>
WHERE ((((((((((((((Last shipment country = "CN") OR (Last shipment country = "CH")) OR (Last shipment
country = "BD")) OR (Last shipment country = "NZ")) OR (Last shipment country = "SC")) OR (Last shipment
country = "SA")) OR (Last shipment country = "SI")) OR (Last shipment country = "OM")) OR (Last shipment
country = "MZ")) OR (Last shipment country = "CO")) OR (Last shipment country = "RU")) OR (Last shipment
country = "GA")) OR (Last shipment country = "UY")) OR (Last shipment country = "ML")) AND
((((((((Reporting firm code = "369-CF") OR (Reporting firm code = "480-CF")) OR (Reporting firm code =
"318-CF")) OR (Reporting firm code = "204-CF")) OR (Reporting firm code = "495-CF")) OR (Reporting firm
code = "423-CF")) OR (Reporting firm code = "534-VG")) AND ((((((((((((((((((((((Consignee code =
"2000126506") OR (Consignee code = "4000731483")) OR (Consignee code = "3002118077")) OR
(Consignee code = "3002178010")) OR (Consignee code = "3001551061")) OR (Consignee code =
"7001739980")) OR (Consignee code = "4001754915")) OR (Consignee code = "3002042721")) OR
(Consignee code = "4001095479")) OR (Consignee code = "3001658102")) OR (Consignee code =
"5002336594")) OR (Consignee code = "4002326565")) OR (Consignee code = "5002437853")) OR
(Consignee code = "4001936267")) OR (Consignee code = "5000028693")) OR (Consignee code =
"4001903934")) OR (Consignee code = "6001916971")) OR (Consignee code = "2002473002")) OR
(Consignee code = "3001048808")) OR (Consignee code = "3002418055")) OR (Consignee code =
"5001675869")));
```

Bassem Chermiti



Bassem Chermiti is a Central Inspector of Financial Services, Head of Selectivity and Risk Analysis Unit at the General Directorate of Tunisian Customs. He holds a university degree in finance, a diploma of higher specific studies in IT applied to management from Tunisian Higher Institute of Management and a diploma of higher specific studies of public finance from the Institute of Customs and Tax Economics Kolea Algeria. His work focuses on the use of data analysis techniques in the context of customs risk management.