Int. J. Nonlinear Anal. Appl. 15 (2024) 7, 309–324
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.27923.3767

Comparison of three LDA, PCA and ICA fast methods using fourteen data analysis algorithms to develop a risk assessment management model for export declarations to deal with illegal trade in Iran customs

Hassan Ali Khojasteh Aliabadi[a], Saeed Daei-Karimzadeh[b,*], Majid Iranpour Mobarakeh[c], Farsad Zamani Boroujeni[d]

[a] *Department of Public-Financial Management, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran*

[b] *Department of Economics, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran*

[c] *Faculty of Computer Engineering and Information Technology, Payame Noor University, Tehran, Iran*

[d] *Faculty of Engineering, Department of Computer, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran*

(Communicated by Zakieh Avazzadeh)

Abstract

Risk assessment is the main component of risk management, therefore, developing a suitable data analysis model is particularly important in customs. The purpose of this research is to use data mining techniques to develop an intelligent model for timely prediction of the risk level of export declarations in customs and as a result to prevent irreparable damages. Data mining techniques have been used in this research considering the data-oriented statistical population. The statistical data of the cross-border trade system of the Iranian customs is 698,781 data of the export declaration of the entire customs of the country of Iran for the year 2019-2020. Using Python programming language, feature reduction and effective feature extraction were performed after data preprocessing and preparation, with three methods of principal component analysis, linear differential analysis, and fast independent component analysis. Then for the predictive modelling of fourteen classification algorithms, three methods of principal component analysis (PCA), linear discriminant analysis (LDA) and fast independent component analysis (Fast ICA) were used and eighty percent of the training data were used. After training the models, forty-two different models were extracted. For testing, the obtained models were tested with twenty percent of the data. The test results of the models were compared with standard metrics to evaluate the efficiency of the models and the model obtained from the random forest algorithm with the fast independent component analysis method with three features was selected as the best model for predicting and determining the risk level of export declarations in customs.

Keywords: risk, risk assessment, risk management, data mining, customs declaration
2020 MSC: 91B05

[*] Corresponding author

Email addresses: khojaste1390@yahoo.com (Hassan Ali Khojasteh Aliabadi), saeedkarimzade@yahoo.com (Saeed Daei-Karimzadeh), iranpour@pnu.ac.ir (Majid Iranpour Mobarakeh), f.zamani@khuisf.ac.ir (Farsad Zamani Boroujeni)

Received: June 2023 *Accepted:* August 2023

1 Introduction

For the export procedure in the customs, the most significant risks are misdeclaration of goods in areas such as financial violations related to value, amount of tariff code (HS), country of origin and drug trafficking [6]. The increase in the exchange rate and the decrease in the value of the national currency in Iran has led to an increase in the purchasing power of foreign buyers in Iran abuse by criminals and a greater chance of illegal entry of foreign capital. Also, with the start of the war between Russia and Ukraine in 2022 and the decrease in the supply of food items such as wheat (flour) and edible oil in the world and the increase in their global prices and the supply of these goods with subsidies in the country of Iran for the poor, it will increase the desire of the offenders. It has increased the smuggling of subsidized goods to other countries. Unfortunately, during the past years and the spread of war in Iran's neighbouring countries and the destruction of existing infrastructure in these countries, their interest in illegally obtaining Iran's agricultural genetic reserves, which are prohibited from export, has increased more and more. Also, goods such as gasoline, diesel kerosene and other petrochemicals derived from oil due to the subsidies available in them to help the poor in Iran and the price of these goods being lower than the FOB price in the Persian Gulf cause abuse by criminals and increase in illegal exports and smuggling. Therefore, the application of legal and preventive controls to prevent smuggling and customs violations and to protect the security of society against transnational organized crimes is felt more than ever [3]. On the other hand, full inspection of large shipments is almost impossible [15] and customs delays caused by inspections hurt exporters' foreign sales. These effects are more severe for time-sensitive goods [14]. Therefore, customs control methods should be based on the use of effective modern methods of combating fraud and promoting legitimate trade [1]. Therefore, the need to use risk management tools, especially in organizations such as customs, which is responsible for collecting export duties and fighting violations, is strongly felt due to the sensitivity of the profession [2]. Customs risk management is in line with many international standards and principles as follows:

- Revised Kyoto Convention for Simplification and Harmonization of Customs Procedures;
- Security and Facilitation of Global Trade (SAFE) standards of the World Customs Organization;
- EU risk management framework;
- World Customs Organization's risk management guide [20].

In the country of Iran, more than 2000 export declarations are prepared daily in the EPL comprehensive system of customs affairs, and it causes the accumulation of extensive data in the databases of the customs of Iran. Because databases store all the data of customers' business dealings, it is possible to prepare a suitable model for risk assessment and predicting the performance of customers in the future by examining them [10]. Unfortunately, at present, the customs managers use only a part of the abundant information available, and the processing of the information is done according to the selectivity approach, and the criteria used are numerically few, and the analysis of each criterion is often in the form of two answers (yes or no). This system causes its performance to decrease and on the other hand, it increases intrusive inspections. While the detection rate of violations resulting from these inspections is extremely low. Therefore, such a function indicates that the classification was not intelligent and indicates the inefficiency resulting from the incomplete statistical analysis and the inability to effectively use the information available by the system [18, 20].

Also, with the collection of data on a massive scale the speed of technological changes and the need to analyze and process data as soon as possible, traditional systems are no longer able to access the information hidden in these massive databases. Traditional statistical methods have lost their effectiveness today for two reasons. The first reason is the increase in the number of observations, and the second and more important reason is the increase in the number of variables related to an observation. When the scale of data and work on them are higher than human capabilities, the need for computing technologies instead of manual and traditional analysis is felt more [11].

One of the best methods for extracting behavioural patterns is the use of data mining algorithms. Data mining algorithms use statistical methods and artificial intelligence to extract patterns in very large collections [10]. Although data analytics is considered a technological advance, so far there is only a limited understanding of how governments can translate this potential [17]. Due to the large number of customs exchanges and multiple risks, risk analysis has not been done enough to help identify crimes in customs. New challenges must be developed to solve this problem [21]. These techniques can enable customs to detect schemes for certain crimes, from smuggling trade to false declarations, and help them identify potentially related crimes (tax avoidance or concealment of counterfeit money transactions) that require heavy investigation or auditing. become [15]. The use of risk management systems in customs affairs through information refinement and risk assessment is the most important means to advance global trade and identify and deal with high-risk goods and persons, and it allows customs to make accurate and appropriate decisions on the basis and resources focus on identifying and dealing with high-risk individuals [18].

2 The difference between risk management and risk assessment

The risk assessment stage is the most important pillar of risk management, so the more accurate the results of this stage, the more reliable the risk management process can be [9]. To avoid any kind of mistake, the difference between the concepts of risk management and risk assessment should be stated. Risk management includes the systematic application of management procedures and operations to collect the necessary information to better describe and understand risk, while risk assessment includes all the processes of risk identification, analysis, evaluation and prioritization. Risk analysis is one of the basic parts of the risk assessment system, which provides valuable information for making decisions and determining how to face the risk [5].

3 Risk evaluation and prioritization

Risk assessment and prioritization are required to take advantage of pre-defined profiles that mark new interactions. It is necessary to implement the selectivity criterion to control and identify high-risk interactions and direct declarations to appropriate control channels by customs. As a general rule, merchandise targeting techniques today rely on innovative methods based on the use of complex IT systems and related software, which previously relied exclusively on the experiences, judgments, and insights of employees. Automatic analysis methods, in addition to control purposes, reduce the possibility of corruption and violation to the minimum possible and also prevent employees from exercising discretion to select items for control. These software and systems collect all the necessary relevant data enter the risk analysis equation and provide interpretation results. In addition, the obtained data and results can be reused in future analyses and thereby save considerable time [19].

4 Evaluation channels and determination of risk level

For this purpose, the customs clearance routes are defined as follows:

- **Red channel**: a statement placed in this direction has a very high level of risk. In addition to the document inspection, a physical inspection is also carried out, which means that all control measures have been carried out in this direction and in addition to the fact that the desired documents are fully examined the goods have also been physically inspected and the accuracy of the information contained in the declaration has been carefully examined is placed.
- **Yellow channel**: The yellow channel conveys the concept that the existing declarations have an average risk level. It is subjected to documentary inspection but there is no need for physical inspection and direct inspection of the goods and as it is known this route is prioritized before the red route. In addition sampling is also done in this direction if necessary.
- **Green channel**: The declarations that are placed in this way seem to have no danger for the customs and have a very low risk level. Therefore regarding the declarations of this route only a brief review is sufficient.

5 Cases regarding violations in the export declaration in Iran's customs affairs law

Article 107 of the Law on Customs Affairs reads:

If during the examination of the declaration or the inspection of the export goods, it is found that funds less than the prescribed amount have been declared or deposited in addition to receiving the difference at the discretion of the head of customs a fine of 5% to 50% of the difference will be charged. Note - Whenever an unrealistic value is declared more than ten percent less or more in the export declaration in order to create illegal facilities and benefits for the owner of the goods, between ten and one hundred percent (10% to 100%) will be charged for the difference in value.

Important points:

- The value of the export goods is determined by the customs.

- The fine is charged according to the difference in the export value.

When is the value of the export goods in the export declaration lower or higher than the actual value in the declaration by the owner of the goods?

A- The export value is sometimes mentioned more than the actual value:

- In a situation where there is no foreign exchange obligation for exporters, mentioning an unrealistic value for using export facilities such as export bonus is a license to import more goods.

B- The export value is sometimes lower than the actual value:

- In a situation where there is a foreign exchange obligation for the exporters, mentioning an unrealistic value in order to return the export currency less, and selling the remaining currency in the open market or currency smuggling.

6 Risk management based on information processing (smart)

According to the customs strategy in the 21st century, risk management based on information processing refers to a system based on feedback learning loops. This system allows customs to integrate risk-related activities by taking advantage of its past learnings, and by using information technology systems and necessary software, it will transform it from a merely responsive organization to a forward-looking organization with complex forecasting capabilities. Intelligent risk management, along with other capacities of the organization, is very dependent on creating a connection between risk assessment and available information, and by providing the necessary support through intelligent analysis, it plays a complementary role to the operational levels in identifying possible risks [8].

7 Early prediction and explanation system

Forecasting systems help to obtain as accurate forecasts as possible through statistical inferences between important variables. These techniques clearly show which important risk factors determine the company's big profits and provide information about the quality of forecasts that facilitate the occurrence and estimation of risks. These techniques make the changes of an unexplained or in other words unpredictable variable possible [6].

8 Research method

The theoretical framework or conceptual model of this research is derived from the model of Daimler and Chrysler and Fayad et al. The proposed method in this research to develop a suitable management model for export declaration risk assessment includes several steps, the overview of which is shown in diagram 1.

8.1 Data collection and understanding stage

8.1.1 Data collection stage

The data of this research includes 698,781 data related to the information of export declarations related to the period of 2019-2020, which includes twenty characteristics, extracted from the cross-border trade system of Iranian customs.

8.1.2 Data understanding stage

One of the important factors in building a model is the correct selection of features. In order to segment statements based on risk, the first and most important step is to identify risk factors. In order to identify the influential features and better understand the data and form a risk profile in this research, identification is done in two phases.

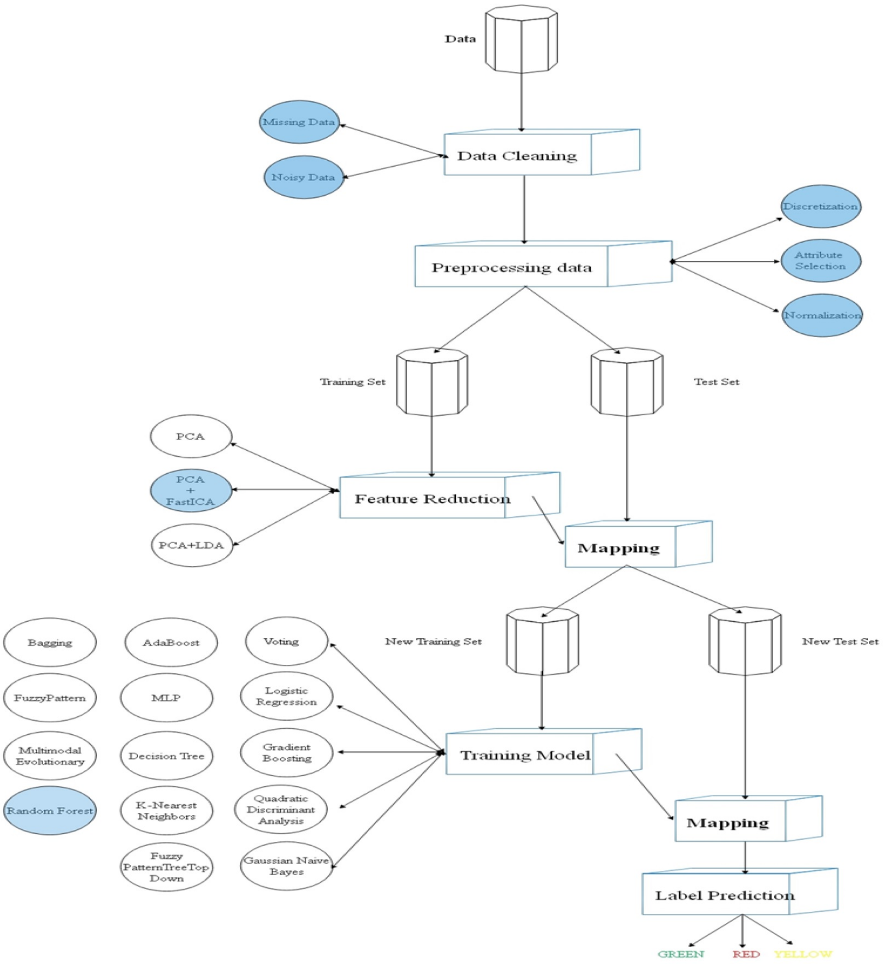


Figure 1: Research Implementation

8.1.3 Studying and reviewing existing research

To be successful in monitoring and correctly targeting declarations submitted to customs to discover cases of risk and violation, it is necessary to perform a series of preliminary data analysis tasks in the form of data mining. To do this, it is necessary to examine the characteristics and specifications of the statements made and to examine the details of the statements and the results of inspections in a past period when violations were committed in them [19]. In this section, based on the available studies, the issues raised are divided into revenue groups (import fee collection), money laundering, documents, customs procedures, transportation issues, special and demographic issues, history of the goods owner and declarant. The cases that can be mentioned as a potential risk or risk in the customs and hinder the implementation of the regulations include changing the tariff, reducing the value to avoid paying taxes and duties, changing the certificate of origin, entering non-sanitary and non-standard goods, and such cases. Laporte forms a risk profile for six different criteria based on the level of violation including transport agents, tariff book, certificate of origin, health and safety, importer/exporter and customs procedure [12]. Commercial frauds include misrepresenting the value of goods, misrepresenting the tariff of goods, misrepresenting the amount of goods, misrepresenting the country of origin or destination, tax violations, and correct deposit of guarantees.

8.1.4 Survey of experts

Considering the importance of the research, the researcher tried to be personally present in all the interviews and used the opinions of respected customs experts and managers to improve the quality. As it was said, the subject of study in this research is the Customs Organization of the Islamic Republic of Iran, so it has been tried to refer to the managers, advisors and deputies of the organization at different levels and use their opinions according to the different topics and stages and the required information. Also, to make the identified factors as productive as possible, an effort was made to use the points of view of customs experts outside the customs organization. The most important factors including value, tariff, country of origin, number of goods and customs procedure were identified. Finally, due to the nature of data mining and the fact that there may be a specific pattern in other variables of the goods declaration, therefore, the risk profile considered in this research includes the titles of the variables used in the declarations, the same information required to complete the international declaration form known as the administrative unit document in The window system is the cross-border trade unit of Iranian customs [7]. For this reason, all available information is used, i.e. the contents of all verified declarations and the results of inspections during a reference period [12]. The risk profile is according to under table.

Table 1: Variables in the Customs Declaration before Data Preparation

Type of transport	**Input customs**	**Product Type**	**The declarant**	**Product owner**
Value components	Customs procedure	Country of origin	Total invoice amount	Country of the transaction
Terms of delivery	The value of the product pen	Eight-digit tariff code	Source of input rights	Customs value of goods item
Input rights	payment method	Number of units of goods	Gross weight	net weight

8.2 Data preparation step

In the book "Data Preparation for Data Mining", Dorian Peel estimates the time required for data preparation to be 60% of the total time of the data mining process [13]. Data mining is a critical exploratory process, so data must be defined correctly and appropriately for this important practice. The main condition in choosing a suitable subset of high-dimensional data is that this subset, despite being small, also has the characteristics of the original data [9]. Data preparation takes place in three stages: data cleaning, data preprocessing, feature reduction, and effective feature extraction.

8.2.1 Cleaning (refining) data

In this application, the data is not stored well for several reasons, the most important of which are the following:

- Human error due to forgetting to fill a field.
- Data was lost when manually migrating from the old database.
- There was a programming or system error.

In this thesis, the examples that included this type of data, that is, the field related to that feature was incomplete, that example was removed from the training and test set.

Among this amount of data, all the data have the required quality and are not capable of modification and are not qualified to enter the final model.

Factors such as defective data, the presence of unrelated data, data whose main fields are missing or incorrectly recorded, duplicate data, outliers (some statistical methods are sensitive to the presence of such values and may produce unstable results).

Therefore, after applying the data screening phase, it is inevitable to remove extraneous and noisy data.

Table 2: Data Set after Data Preparation

<table>
<tr><th>Method</th><th>Number of features</th><th colspan="3">Number of samples</th></tr>
<tr><td>Data</td><td>20</td><td colspan="3">698781</td></tr>
<tr><td>Data after Data Cleaning</td><td>16</td><td colspan="3">691298</td></tr>
<tr><td>Data after Data Transformation</td><td>16</td><td colspan="3">691298</td></tr>
<tr><td rowspan="3">Number of samples in each class</td><td rowspan="3">16</td><td colspan="3">691298</td></tr>
<tr><td>Red channel</td><td>Yellow channel</td><td>Green Channel</td></tr>
<tr><td>233460</td><td>207274</td><td>250564</td></tr>
</table>

Reducing attributes and extracting effective attributes

According to the application of risk management, also, three methods of PCA principal component analysis, LDA linear differential analysis and fast independent fast ICA component analysis are used to reduce the attribute to extract effective attributes and evaluate these types of attributes.

Principal Components Analysis (PCA)

PCA is a multivariate statistical analysis that selects a smaller number of factors as the principal components from among the primary factors so that several insignificant data are removed. In the first extracted basic component, the maximum amount of data scatter is in the entire data set, i.e. the first component is correlated with at least several variables. The second extracted component has two important features: this component considers the largest data set not calculated by the first component, and it is not correlated with the first component. In other words, regardless of the previous component, by passing from the initial component to the final components, each component describes less variance. It means that the first principal component always describes the maximum amount of variance and the last components describe the least variance; thus, such information will not be lost by deleting the last components [10].

Liner Differential Analysis (LDA)

LDA is a statistical method to reduce the dimensions of an issue and identify categories by maximizing the ratio of scatters between groups to scatters within groups. The linear diagnostic analysis approach is similar to and borrowed from the method used by Ronald Fisher to determine the degree of differentiation in groups. It became the basis for variance analysis and thus is sometimes called ”linear differential analysis”. The linear diagnostic analysis is very close to variance and regression analyses; all three statistical methods model the dependent variable as a linear combination of other variables. However, variance and regression analyses take the dependent variable as an interval one, while linear differential analysis is used for nominal or ordinal dependent variables. Therefore, linear differential analysis is more similar to logistic regression. The linear diagnostic analysis is also similar to principal components analysis and factor analysis which are used to linearly combine variables in a way that best describes the data. A major application of both of these methods is to reduce the number of data dimensions. However, these methods differ significantly: in linear differential analysis, class differences are modelled, while in principal component analysis, class differences are ignored [14].

Fast Independent Component Analysis (Fast ICA)

This is an efficient algorithm to detect underlying factors or components of multivariate data dependent on signal separation. Fast ICA is an optimized method of independent component analysis. This method shows a faster convergence of results than the independent component analysis method. Fast ICA is based on a fixed point algorithm that has a known performance speed. Compared to conventional fixed point algorithms, this algorithm has been modified to provide higher performance. It is also similar to some neural algorithms, and is computationally simple, and requires less computational memory [9].

In the proposed method, PCA method is used to evaluate the attribute. The number of attributes has not been reduced for this purpose and the same 16 attributes have been included. For LDA and Fast ICA methods, only 2 and 3 attributes were used, respectively. In these steps, a number of properties have been selected according to special vector diagrams. Equation is also used to select the number of attributes.

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{f}\lambda_j} \geq 0.98 \tag{8.1}$$

in this relation, k is the number of selected properties, f is the total number of properties, and λ is the values of the special vector.

Table 3: LDA and Fast ICA Methods

Method	**Number of features**	**Number of samples**	
Data after Feature Reduction	PCA+LDA, 2 Feature	test 20%	training 80%
	PCA + Fast ICA, 3 Feature		
Sample segmentation	16		

Discretization

Many of the data in the declarations section have non-quantitative or non-discrete data. In this regard, this type of features is recognized and discretization is done on the data.

Normalization

Also, the types of normalization methods are evaluated and the extracted features are normalized in the range of zero to one. Data normalization is a method of uniformizing the range of values related to different variables and is also known as data scaling. If the unit of measurement of the studied variables is diverse, the data can be scaled using normalization methods. There are different methods for data normalization. In this treatise, one of the most famous methods called MinMaxNormalization is used.

In this method, each of the data can be converted into an arbitrary interval. The general MinMaxNormalization formula for converting data to the interval between a and b is as the following equation:

$$Z = (b - a)\frac{x - \min(x)}{[\max(x) - \min(x)]} + a \tag{8.2}$$

in this relationship, x is the characteristic that must be normalized, $min(x)$ and $max(x)$ are the minimum and maximum values for characteristic x, respectively. a and b are the normalization interval which is considered in this treatise for $a = 0$ and $b = 1$.

8.3 Data modelling

According to the required result, which is risk prediction, and the most important purpose of the classification is to provide a model for prediction, in this research, out of fourteen predictive data mining algorithms, such as decision trees, Gaussian simple Bayes, fuzzy logic, random forests, logistic regression, k-nearest neighbor, Ada boost, gradient boosting, bagging, neural network, hybrid bimodal, voting, multi-model and fuzzy tree are used from top to bottom. At this stage, supervised learning tools are used. Also, the process of building the model is a two-step process, in the first step, the model is built with the help of the training data set, where the class label of all its samples is known. This stage is known as the learning or training stage. In the second step, the obtained model is validated or tested with the help of the test data set in which the class labels are usually unknown. In fact, the evaluation of the model is calculated according to how many classes of experimental data samples are correctly estimated [9].

8.3.1 Modelling with training data

This stage is known as the learning or training stage. In this section, by random sampling, 80% of the data was allocated and determined for training. According to the machine learning methods for risk management in the training set, fourteen algorithms are used and a separate model is made for each algorithm. It should be noted that at this stage, various combinations of principal component analysis, linear differential analysis, and fast independent component analysis are used. It is noteworthy that, as mentioned before, according to the level of risk, the declarations are divided into three categories: green with the label number zero, red with the label number one, and yellow with the label number two, each category representing their level of risk. In this regard, the trained models are for the three mentioned outputs.

8.3.2 Validation (testing) of models

In fact, the evaluation of the model is calculated according to how many of the test data samples the class is correctly estimated. The obtained models are tested with the help of the test data set where the class labels are usually unknown. Therefore, the validity and predictive power of the models are tested by experimental data without class labels (green, yellow and red) which have not been encountered so far.

In this section, according to random sampling, 20% of the data is allocated for the test. For this reason, the experimental data are considered as data for training models.

8.3.3 Model evaluation and selection

After building a model, many questions may arise. Questions like:

- With more than one model, how can you choose the best among them?
- How accurate is the model for predicting the risk class of the statement?
- What is the accuracy of a model and how is its value estimated?
- How to get a reliable estimate of the model?

Since the presented method is measured in terms of validity in any research, therefore, in this research, which is a data-oriented model with an observer, the validation method is performed in the following ways.

- **Metrics to evaluate the efficiency of models**

  Common metrics for evaluating models include accuracy or detection ratio, sensitivity or recall, clarity, precision, F_1 and F_β, which are summarized in Table ago of evaluation metrics.

- **Accuracy**: The accuracy of a model on a test data set is the percentage of data in this set that are correctly labeled by the model.

- **Recall**: The accuracy measure is considered as a correct measure. The percentage of data that is labeled as positive (true) is actually classed as positive (true), but it says nothing about the data that is true and labeled as false.

- **Sensitivity**: It is a measure of completeness. The percentage of correct data that is correctly classified but does not say anything about the incorrect data that is labeled as correct. Explaining that accuracy and sensitivity metrics have an inverse relationship with each other, and an increase in one may cause a decrease in the other. Therefore, the solution to using them is to use the F1-score measure.

- **The combination of sensitivity and accuracy**: F measure is the harmonic mean of two measures of accuracy and sensitivity.

Table 4:

Formula	**Measure**
$(TP + TN)/(P + N)$	Recognition Rate, accuracy
$(FP + FN)/(P + N)$	Error rate
TP/P	Sensitivity, recall, true positive rate
TN/N	Specificity, true negative rate
$TP/(TP + FP)$	Precision
$(2 \times Precision \times recall)/(Precision + recall)$	F_1, F, F-score
$((1 + \beta^2) \times Precision \times recall)/(\beta^2 \times Precision + recall)$	F_β

In the definition of metrics:

TP correctly labeled positive samples,

TN correctly labeled negative samples,

FP falsely labeled positive samples,

FN falsely labeled negative samples,

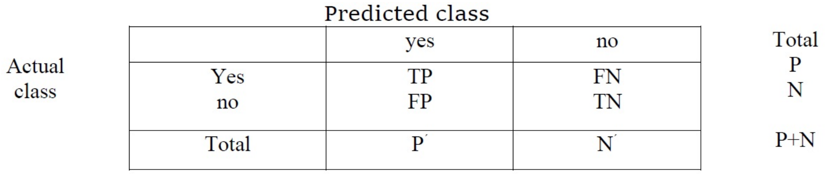
P labeled positive samples and

N are labeled negative samples. These terms are summarized in a matrix called disturbance matrix.

- **Perturbation matrix**

  It is a useful tool and with its help you can see the performance of the model to recognize tuples of different classes. An ideal model places most of the tuples on the main diameter of the perturbation matrix, and it is desirable that the rest of the elements except the main diameter of the matrix have a value of zero or close to zero. Therefore, ideally, the values of FN and FP are close to zero.

Table 5: The disturbance matrix with the display of the sum of positive and negative tuples



		Predicted class		
		yes	no	Total
Actual class	Yes	TP	FN	P
	no	FP	TN	N
	Total	P'	N'	P+N

- **Comparison of models based on rock curves**

  Rock curves are a useful visual tool for comparing multiple models. A rock curve drawn for a model shows the relationship between the true positive ratio and the false positive ratio. The farther the rock curve of a model is from the diagonal line and the larger the surface area under the curve, the higher the accuracy of the model and vice versa [11].

9 Findings of export declarations

9.1 Findings from fourteen models using principal component analysis method with 16 features

The test of fourteen research models in table number six using principal component analysis method with 16 features showed that the risk prediction accuracy of k-nearest neighbor model is 77% higher than other models in this method.

Table 6: Analysis of findings from the application of different algorithms with 16 features

PCA-16Features						
Method	**Accuracy**	**macro avg**	**Label**	**precision**	**recall**	**f1-score**
Voting (Random Forest, Bagging, Fuzzy Pattern)	74	74	0	72	73	72
			1	76	78	77
			2	75	72	73
Bagging	70	70	0	69	67	68
			1	72	75	74
			2	69	58	69
Fuzzy Pattern	30	29	0	30	100	46
			1	0	0	0
			2	0	0	0
Multimodal Evolutionary	36	35.5	0	37	27	31
			1	35	39	37
			2	35	40	37
Random Forest	75	75	0	73	74	73
			1	76	79	77
			2	76	73	74
Fuzzy Pattern Tree Top Down	38	30.5	0	37	37	37
			1	38	74	50
			2	0	0	0
K-Nearest Neighbors	77	77	0	71	81	76
			1	82	75	78
			2	78	74	76
Decision Tree	53	53	0	56	45	48
			1	52	64	60
			2	52	49	60
MLP	52	54.5	0	55	70	61
			1	96	8	14
			2	47	82	60
AdaBoost	52	52	0	51	49	50
			1	54	58	56
			2	52	50	51
Gaussian Naive Bayes	36	24.5	0	0	0	0
			1	36	100	53
			2	0	0	0
Quadratic Discriminant Analysis	36	24.5	0	0	0	0
			1	36	100	53
			2	0	0	0
Gradient Boosting	40	40	0	41	33	37
			1	40	50	44
			2	41	36	39
Logistic Regression	59	59	0	60	54	57
			1	58	66	62
			2	60	57	58

In the method of analysis of the main components of export declarations, the K-nearest neighbor model is chosen compared to other models due to its higher accuracy of 77%.

As can be seen in the uncertainty matrix number one of the k-nearest neighbor in the principal component analysis method, the values of the main target variable, i.e. the red channel (label one), 75% of the data predicted in label one, really belonged to label 1 (correct positive rate 1 or positive True) and in the same way, the label zero works with 81% and the label two works with 74%.

As it can be seen in figure number one of the rock curve, the comparison of the results of the fourteen research models using the principal component analysis method with the sixteen features of the larger surface under the k-

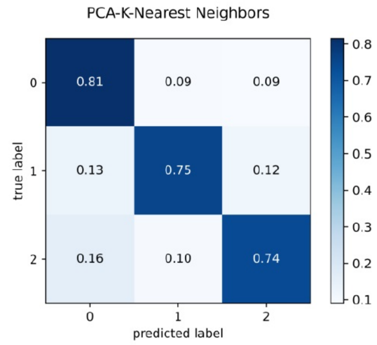


Figure 2: Ambiguity matrix number (1)

nearest neighbor rock curve model compared to other models shows a higher accuracy of predicting the risk level this model is from other models.

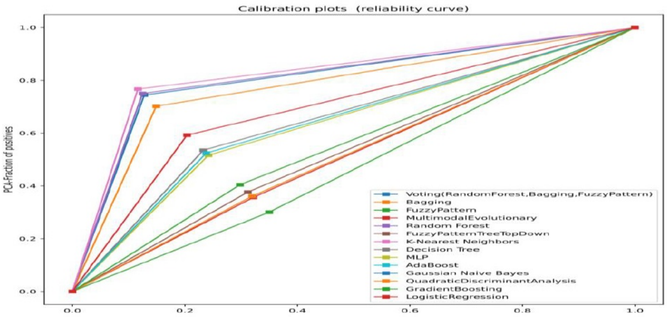


Figure 3: rock curve

9.2 Findings from fourteen models using fast independent component analysis method with three features

The test of fourteen research models in table number seven using the fast independent component analysis method with three characteristics showed that the risk prediction accuracy of the random forests and bagging model is 77% higher than other models in this method.

In the fast independent component analysis method of export declarations, two models of random forests and bagging are chosen with 77% due to their higher accuracy.

Finally, as it can be seen in figure number five of the rock curve, the comparison of the results of the fourteen research models using the fast independent component analysis method with the three features of the larger surface under the rock curve of the random forest model compared to other models shows the accuracy of risk prediction in this. The model is higher than other models and the bagging model.

As can be seen in the uncertainty matrix number six of random forests in the principal component analysis method, the values of the main target variable, i.e. the red channel (label one), 76% of the data predicted in label one, really belonged to label 1 (correct positive rate 1 or (positive True) and in the same way, the label zero works with 84% and the label two works with 73%.

9.3 Findings from fourteen models using the method of linear differential analysis with two characteristics

The test of fourteen research models in table number eight using the method of linear differential analysis with two features showed that the risk prediction accuracy of the random forest model is 37% higher than other models in this method.

In the method of linear differential analysis of export declarations, the accuracy of gradient boosting is 38% higher than the random forest model with a prediction accuracy of 37%. The prediction accuracy of 0% of the red class,

Table 7: Analysis of findings from the application of different algorithms with three characteristics

PCA+Fast ICA(3 com)						
Method	**Accuracy**	**macro avg**	**Label**	**precision**	**recall**	**f1-score**
Voting (Random Forest, Bagging, Fuzzy Pattern)	76	76	0	73	79	76
			1	78	77	77
			2	77	73	75
Bagging	77	77.5	0	72	84	77
			1	82	76	79
			2	79	73	76
Fuzzy Pattern	37	34.5	0	31	9	14
			1	38	47	42
			2	37	51	43
Multimodal Evolutionary	31	26	0	30	55	39
			1	33	41	37
			2	0	0	0
Random Forest	77	77.5	0	72	84	77
			1	82	76	79
			2	79	73	76
Fuzzy Pattern Tree Top Down	38	37.5	0	41	26	32
			1	37	50	43
			2	36	35	36
K-Nearest Neighbors	75	75	0	69	77	73
			1	79	74	76
			2	75	73	74
Decision Tree	77	77	0	71	84	77
			1	81	75	78
			2	79	72	75
MLP	44	44	0	44	38	40
			1	43	59	50
			2	46	34	39
AdaBoost	46	46	0	50	32	39
			1	46	62	53
			2	44	41	43
Gaussian Naive Bayes	40	40	0	44	22	29
			1	39	56	46
			2	41	40	40
Quadratic Discriminant Analysis	41	41	0	43	23	30
			1	41	56	47
			2	42	42	42
Gradient Boosting	36	21	0	0	0	0
			1	36	100	53
			2	0	0	0
Logistic Regression	53	53	0	59	39	47
			1	51	68	58
			2	53	50	51

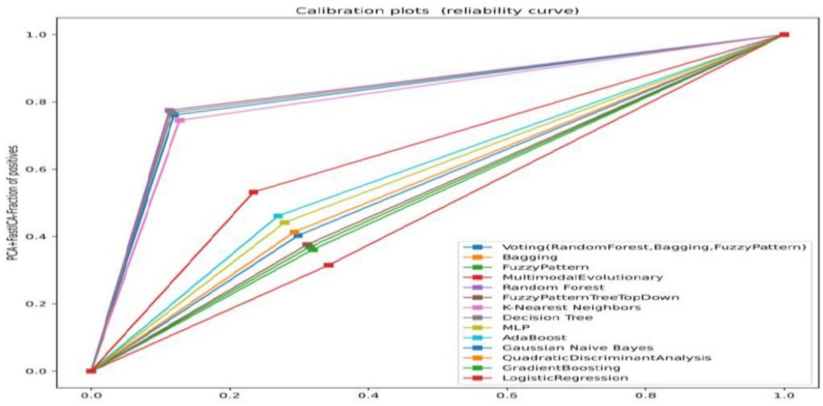


Figure 4: rock curve

53% of the green class and 65% of the yellow class and due to the lack of prediction in the red class and due to the prediction accuracy of 18% of the red class, 73% of the green class and 26% of the yellow class, the random forest

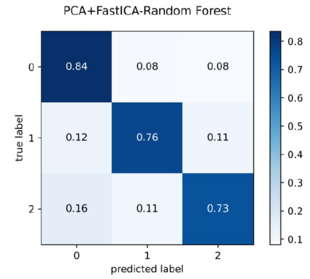


Figure 5: Ambiguity matrix number (2)

Table 8: Analysis of findings from the application of different algorithms with 2 features

PCA+LDA2com						
Method	**Accuracy**	**macro avg**	**Label**	**precision**	**recall**	**f1-score**
Voting (Random Forest, Bagging, Fuzzy Pattern)	34	31.6	0	42	0	0
			1	35	67	46
			2	33	29	31
Bagging	36	36.3	0	35	77	48
			1	44	15	23
			2	37	22	28
Fuzzy Pattern	36	32	0	100	0	0
			1	36	100	53
			2	0	0	0
Multimodal Evolutionary	30	26.3	0	30	100	46
			1	0	0	0
			2	60	0	100
Random Forest	37	37.6	0	35	73	47
			1	44	18	25
			2	38	26	31
Fuzzy Pattern Tree Top Down	34	36	0	33	7	11
			1	92	0	0
			2	34	96	51
K-Nearest Neighbors	34	28.3	0	27	2	4
			1	39	2	3
			2	34	97	50
Decision Tree	35	34.6	0	34	82	48
			1	42	16	23
			2	36	14	21
MLP	36	22.6	0	0	0	0
			1	36	100	53
			2	14	0	0
AdaBoost	34	37	0	10	0	0
			1	34	11	17
			2	34	88	49
Gaussian Naive Bayes	34	28.3	0	33	1	2
			1	31	0	0
			2	34	99	51
Quadratic Discriminant Analysis	35	40.6	0	34	85	48
			1	100	0	0
			2	39	28	32
Gradient Boosting	38	31.6	0	38	53	44
			1	0	0	0
			2	38	65	48
Logistic Regression	36	33.6	0	33	10	15
			1	36	89	52
			2	56	4	7

model is a more suitable model and is selected. As can be seen in the uncertainty matrix number seven of random forests in the linear differential analysis method, the values of the main target variable, i.e. the red channel (label one) are 18% of the data predicted in label one, they really belonged to label 1 (True positive rate) and in the same way label zero works with 73% and label two with 26%.

As it can be seen in figure number eight of the rock curve the comparison of the results of the fourteen research models using the linear differential analysis method with two characteristics the area under the rock curve of the

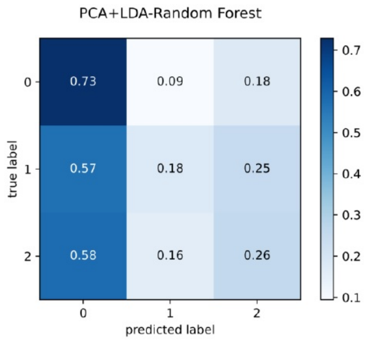


Figure 6: Ambiguity matrix number (3)

gradient boosting model is larger than other models but because of the prediction of the model Random forests indicate the risk prediction accuracy of this model is higher than other models.

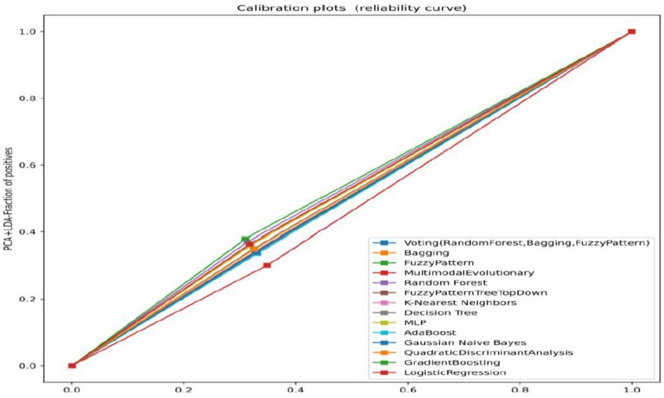


Figure 7: rock curve

9.4 Comparison of findings and selection of the best model in export declarations

Comparison of class prediction results in top models in three main component analysis methods of K-nearest neighbor model with 77% prediction accuracy and 77% harmonic mean, The fast independent component analysis method of the random forest model with 77% prediction accuracy and the harmonic mean of 77.5% and the linear differential analysis method of the random forest model with 37% indicate the higher prediction accuracy of the risk class in the random forest model in the fast independent component analysis method. In other words, with the help of the model, we are able to classify export declarations in one of the three risk levels with an accuracy of 84% in the green class, 73% in the yellow class, and 76% in the red class. Finally, considering that the accuracy of the random forest model is higher than other models so this model can be used for the management of risk assessment in the customs organization in the proper diagnosis of the risk class of export declarations as a suitable organizational knowledge.

10 Conclusion

Using data mining techniques can significantly improve the selection of declarations to be inspected.

The conducted tests show that data mining techniques allow effective targeting of declarations. The results of this research showed that the random forest model has a higher accuracy than other used models in detecting and determining the level and risk rating of export declarations. The rules resulting from this model are considered as a hidden pattern in the Iranian customs database and it is possible to use this model to predict export declarations with a high level of risk in the red, yellow and green channel and to make policies regarding declaration risk management applied for export. In order to be successful in monitoring and correctly targeting declarations submitted to customs to discover risk and violation cases there is a need to manage risk assessment by performing a series of basic data analysis tasks in the form of data mining and developing a smart model for predicting the level of risk. After forecasting and conducting controls and obtaining results, the correct results can be used to update and form more accurate risk profiles in order to increase the accuracy of model forecast estimation in future cases and it is considered as a tool that converts qualitative information into quantitative information. Intelligent risk management systems, the capabilities of the automatic risk assessment program, provide customs with new opportunities such as faster identification and prediction of the arrival of high-risk shipments. The most important purpose of the selected model is to provide the wasted financial revenues of the government from the place of salaries and input duties and with different sub-objectives as follows:

- Prevent tax evasion and entry fees and access to lost government revenues,
- Preventing the sleep of capital in customs,
- Facilitate and accelerate trade and contribute to economic growth,
- Prevent smuggling, money laundering and organized crime,
- Facilitate and accelerate the supply chain of goods,
- Increase controls, if necessary,
- Targeted regulation of goods inspection programs,
- Formation of risk profiles and analysis of risk information,
- Identify potential security threats.

10.1 Discussion

Although, there is a very good discourse in favor of the adoption of risk management practices by customs at the highest levels of decision-making by government, there are few studies on the principles of implementation and application of customs risk management [17]. However, from a research point of view, there seem to be two main reasons: First, few theories have been explained and formulated to embed risk management in the body of the customs organization and to institutionalize it. Second, the data available in the field of customs risk management do not show much desire to share information with researchers because traders may misuse the published data and statistics in line with their illegal greed and calculated avoidance of customs inspections. Therefore, despite the importance of risk management for customs, related techniques have not been fully studied empirically and theoretically [19]. But Since the concept of risk-based management is applicable in almost every business and government field, many experiences can be shared with the subject of customs [4].

In an article entitled ”Amending the association rules of customs crime information based on repeated motives”, using the association rules, introduced a way to discover the knowledge used in customs crime data. They came with a set of rules of knowledge to predict and target specific hazard conditions [21]. According to [13], the decision tree is the most appropriate data mining technique. Using two techniques of self-organizing network and k-means algorithms, conducted a study entitled ”Comparison of two data mining methods in risk-based clustering of vehicle insurance customers (Case study: Mellat Insurance Company)”. The results showed that the outputs of the two systems are very close to each other. The customers were clustered based on risk and the strengths and weaknesses of each technique were reported [9]. In another study, in an article entitled ”Investigating the usefulness of data mining techniques in detecting fraud in financial statements” examined a sample of 202 companies, including 101 fraudulent companies and 101 non-fraudulent ones. The results of applying data mining techniques such as multilayer feed forward neural networks, support vector machines, genetic programming, group data modeling, logistic regression, and probabilistic neural network indicated the superiority of probabilistic neural networks without feature extraction in the discovery of the fraudulent financial statements. In the case of including feature extraction, genetic programming and possible neural networks performed with almost equal accuracy [16]. In a paper entitled ”Risk management systems: the use of data mining in the customs of developing countries” examined several simple statistical regression methods, logit, and probit in Senegal. They indicated that using statistical risk indexing methods, 96.6% of declaration violations were detected by inspecting only 20.6% of them. Therefore, it can be claimed that the number of intrusive inspections can be reduced by 80% [12].

References

[1] V. Afanasieva, L. Ivanov, and D. Yanushkevych, M*odern approaches to risk management and their use in customs*, Traektoriâ Nauki Path Sci. **3** (2017), no. 4, 1–6.

[2] S. Ali Asghari, F. Ahmadi Abkenar, and A. Shah Bahrami, *Identification of systemic and business risks and risk management in customs*, Nat. Conf. Organ. Risk Manag., Tehran, Center for Productivity and Human Resources Studies, 2015.

[3] M. Arabi, *Strategic Planning of the Customs of the Islamic Republic of Iran*, Tehran, Nil Publications, 2004. [In Persian]

[4] J. Biljan and A. Trajkov, *Risk management and Customs performance improvements: The case of the Republic of Macedonia*, Proc.-Soc. Behav. Sci. **44** (2012), 301–313.

[5] G. Fayyad, P. Piatestsky-Shapiro, and P. Symth, *From data mining to knowledge discovery in databases*, Al Magazin **17** (1996), 37–54.

[6] W. Gleissner and T. Berger, *Einfach Lernen! Risikomanagement*, Retrieved November, 2009.

[7] R. Goli, *Explanation and Analysis of the Customs Affairs Law and its Executive Regulations*, Tehran University Route, 2018.

[8] P. Hanafizadeh, *Comparison of two data mining methods in segmenting car body insurance customers based on risk (Case study: Mellat Insurance Company)*, Ind. Manag. Stud. **30** (2013), 77–97.

[9] M. Hayati, M. Atai, R. Khalokakaei, and A. Sayadi, *Risk assessment and ranking in the supply chain using taxonomic analysis method (Case study: Isfahan Steel Complex)*, J. Operat. Res. Appl. **40** (2013), no. 1, 85–103.

[10] A. Jamaat and F. Asgari, *Credit risk management in the banking system with a data mining approach*, Quart. J. Quant. Stud. Manag. **11** (2010), 115–126.

[11] H. Jiawei and K. Micheline, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[12] B. Laporte, *Risk management systems: Using data mining in developing countries customs administrations*, World Customs J. **5** (2011), no. 1, 17–29.

[13] D.T. Larose and C.D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, 2014.

[14] C.V. Martincus, J. Carballo, and A. Graziano, *Customs*, J. Int. Econ. **96** (2015), 119–137.

[15] Y. Okazaki, *Implications of big data for customs -how it can support risk management customs*, J. WCO Res. Paper **39** (2017), 1–24.

[16] P. Ravisankar, V. Ravi, G.R. Rao, and I. Bose, *Detection of financial statement fraud and feature selection using data mining techniques*, Decision Support Syst. **50** (2011), 491–500.

[17] B. Rukanova, Y.-H. Tan, M. Slegt, M. Molenhuis, B.V. Rijnsoever, J. Migeotte, M.L.M. Labare, K. Plecko, B. Caglayan, G. Shorten, O.V.D. Meij, and S. Post, *Identifying the value of data analytics in the context of government supervision: Insights from the customs domain*, Gover. Inf. Quart. **38** (2021), 101496.

[18] M. Shishechiha, *Risk Management in Customs Affairs*, Tehran, Basic Science Development Publications, 2015. [In Persian]

[19] M. Yousefi, *Comparative Study of Customs Risk Management*, Tehran, Dara Publications, 2016. [In Persian]

[20] M. Yousefi, *Modern Customs Programs in the 21st Century*, Tehran, Dara Publications, 2016. [In Persian]

[21] B.-B. Zehero, E. Soro, Y. Gondo, P. Brou, and O. Asseu, *Elicitation of association rules from information on customs offences on the basis of frequent motives*, Engineering **10** (2018), no. 9, 588–605.