In the name of God

Data analysis and

- *Hassan Ali Khojasteh Aliabadi*

- *Ghorban Karimi*

Comparison of LDA and Fast ICA methods using fourteen data analysis algorithms to develop a assessment risk management model for export declarations to combat illegal trade

- Customs risk management is in line with many international standards and principles as follows:
- • Revised Kyoto Convention to simplify and harmonize customs procedures;
- • World Customs Organization (SAFE) Security and Facilitation Standards;
- • EU Risk Management Framework;
- • Risk Management Guide of the World Customs Organization

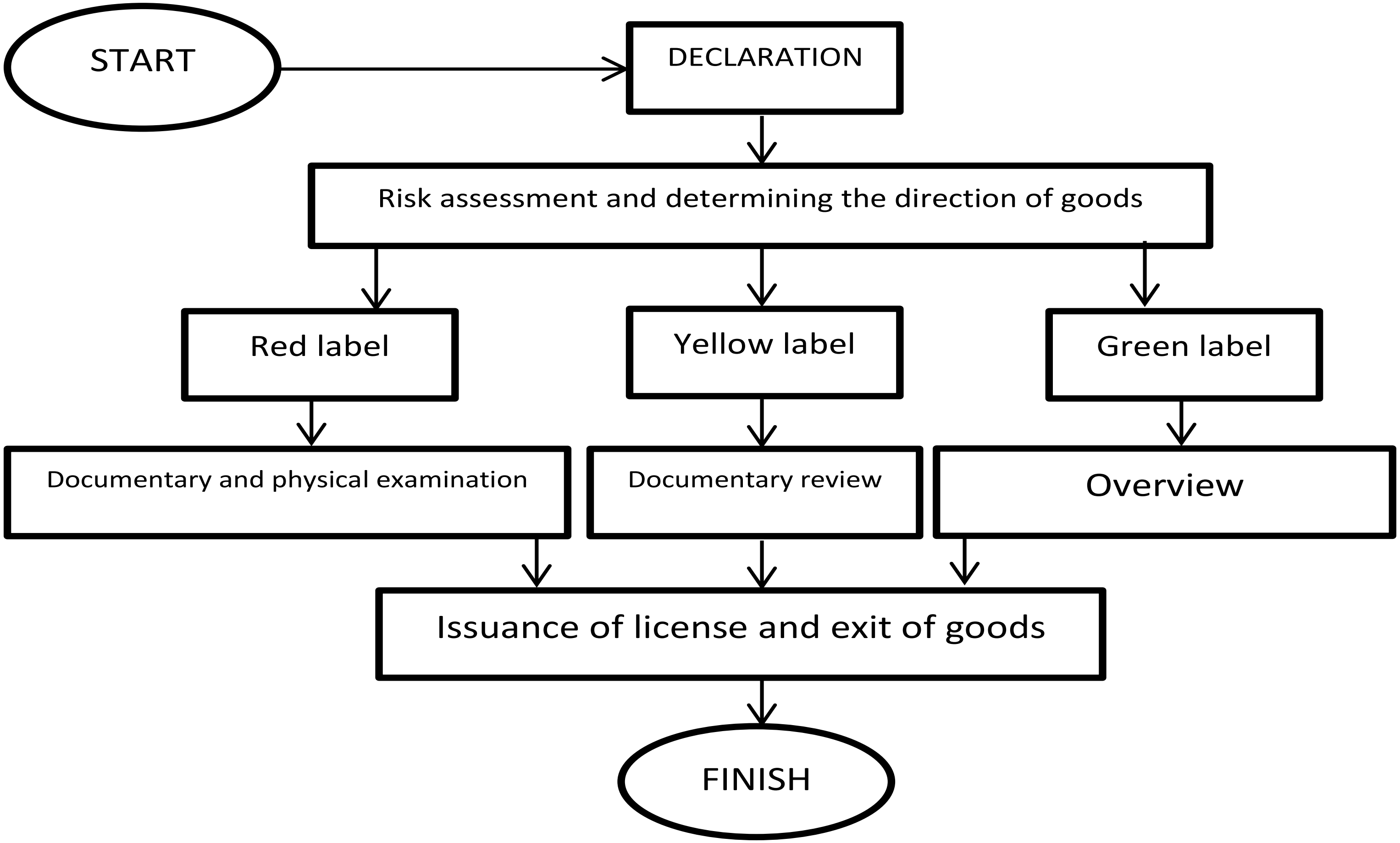
Also, by collecting data on a large scale and the speed of technological change and the need to analyze and process data as soon as possible, traditional systems are no longer able to access the information contained in this large database. Traditional statistical methods have lost their effectiveness today for two reasons. The first reason is the increase in the number of observations, and the second reason, which is more important, 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 computational technologies instead of manual and traditional analysis is felt more.

One of the best ways to extract behavioral patterns is to use data mining algorithms. Data mining algorithms use statistical methods and artificial intelligence to extract patterns in very large collections. Although data analytics is considered a technological breakthrough, so far there is only a limited understanding of how governments translate this potential.

- **Early forecasting and explanation system**
- Prediction systems help to make as accurate predictions as possible through consistent statistical perceptions between important variables. These techniques explicitly indicate which important risk factors determine the firm's large profits, and themselves provide information about the quality of the forecasts that facilitates the occurrence and estimation of risks. These techniques make it possible to change one unexplained variable or, in other words, another unpredictable one.

- **Channels of assessing and determining the level of risk**

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


START
DECLARATION
Risk assessment and determining the direction of goods
Red label
Yellow label
Green label
Documentary and physical examination
Documentary review
Overview
Issuance of license and exit of goods
FINISH

Risk management practices in goods formalities

- **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 an appropriate model for export risk management risk assessment consists of several steps, the general view of which is shown in Figure (2).

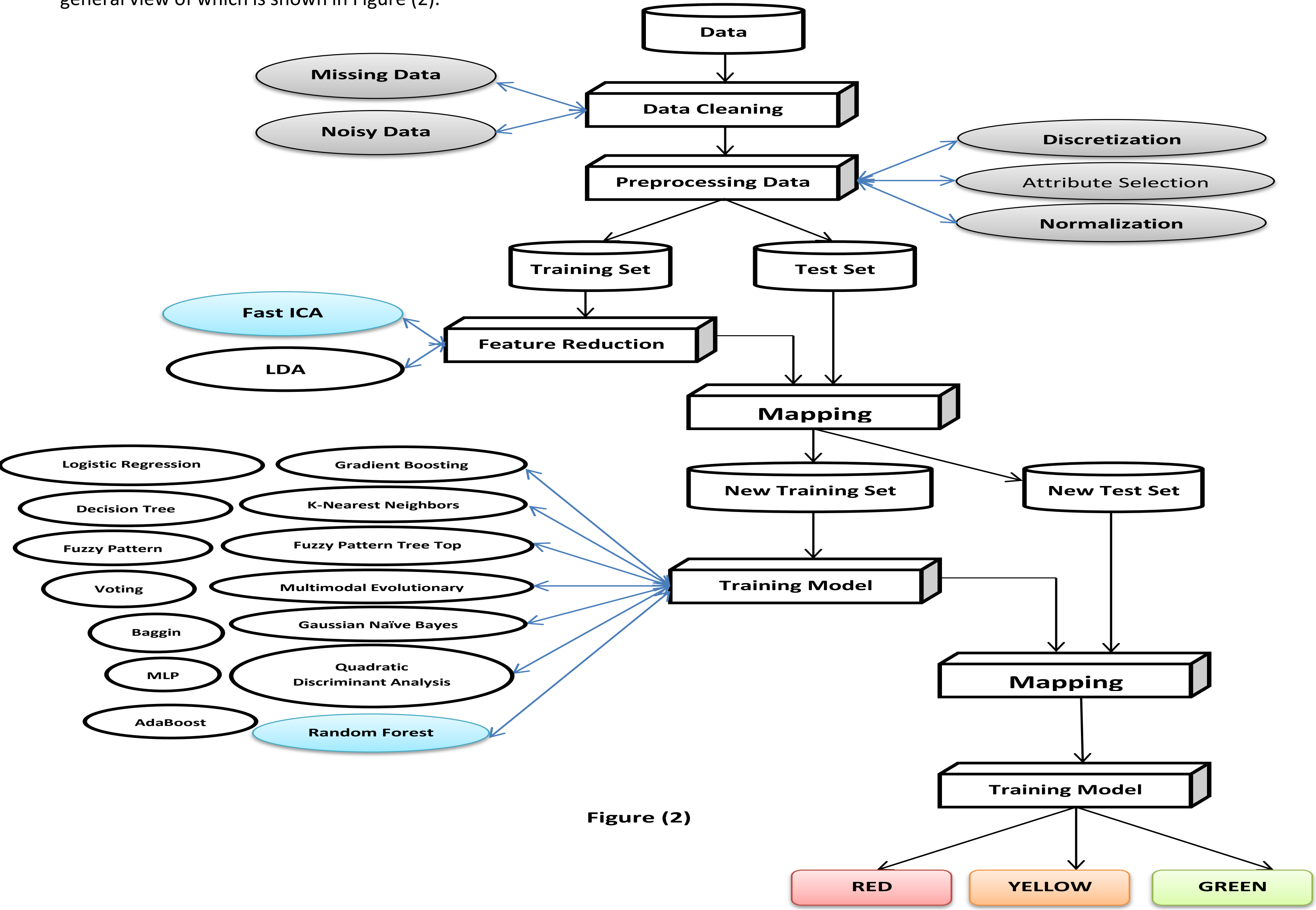


Figure (2)

- In the first stage, the data screening phase is performed. In fact, at this stage, junk data is detected and deleted. In the data set in this application, many attributes are incomplete, out of range, so in this step, this data is identified and deleted.

- In the next step, the discretization, extraction and evaluation of features are discussed. Much of the data in the export sector for all countries has non-quantitative or non-discrete data. In this regard, this type of feature is identified and discretization is performed on the data. It should be noted that the feature, like the tariff code, is extracted based on the HS coordinated code. At this stage, according to the available data set, the appropriate features are selected. Then the extracted features are normalized in the range of zero to one. It should be noted that a variety of normalization methods have been evaluated at this stage.

- The pre-processing stage of the test and experiment data set is selected. In the next step, the feature reduction method is used to extract effective features and evaluate these types of features according to the application of risk management, Linear Differential Analysis (LDA) and Fast Independent Component Analysis (Fast ICA). In these steps, a number of properties have been selected according to special vector diagrams. Equation (1) has been used to select the number of attributes.

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

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

At this stage, the test and training data sets are mapped to the same space according to the weights extracted from the training set. After applying the feature reduction methods, new training and testing datasets are created. In the next step, machine learning methods are used to manage risk in the training suite. At this stage, many methods such as methods based on decision trees, bagging, neural network, fuzzy, etc. have been used and a separate model has been created for each method. Finally, each model is used to evaluate the experimental set. It should be noted that the purpose of this research is data mining risk management on customs declarations. These types of declarations are divided into three categories according to the degree of risk: green, yellow and red. Each category represents their level of risk. In this regard, the models are trained and evaluated for three outputs.

- **Statistical information from different stages:**

In the first step, the data set contains 698,781 samples, which includes 20 attributes. After applying the data screening phase, removing the junk data, the set includes 691298 samples. In the preprocessing phase, after discretizing and removing similar features, the number of features changed to 16. In this step, 4 similar samples are removed. The number of samples in this data set is the number of risk samples with green color value equal to 250564, red color equal to 233460 and for yellow color equal to 207274. In the next step, the test and training set is divided into 20 to 80 ratios. Only 3 and 2 features were used for LDA and Fast ICA methods, respectively.

-Red label

In order to identify export labels with a red label with a high level of risk, customs needs to identify declarations that have led to violations in a previous period (for example, during the last twelve months) . For this purpose, 233460 declarations were identified. Declarations that contain violations such as miscalculation in order to evade payment of export duties, false statement, incorrect entry of tariff code, declaration of prohibited goods, under-declaration, over-declaration, abuse of exemptions, abuse of entrepreneurial rules (such as goods CKD & SKD), smuggling of goods, security issues, health issues, money laundering that have been detected by violations, virtual expertise, inspection, inspection and protection and auditing after the clearance of these violations.

-Yellow label

In order to identify export declarations with a yellow label with a medium level of customs risk, it is necessary to identify declarations that have been examined in the previous period on suspicion of violation, but their violation has not been proven. For this purpose, 207274 declarations were identified.

-Green label

In order to identify low-risk green-label export declarations, other declarations that have not been placed in either the high-risk or medium-risk levels in a previous period are reviewed. For this purpose, 250,564 declarations were identified

- **Linear differential analysis(LDA)**

The statistical method is to reduce the size of an issue and identify categories by maximizing the dispersion ratio between groups within groups. The linear diagnostic analysis approach is in fact similar to and borrowed from the method used by Ronald Fisher to determine the degree of differentiation between groups and as a basis for analysis of variance. For this reason, this analysis is sometimes called "linear differential analysis". Linear diagnostic analysis is very close to analysis of variance and regression analysis; In all three statistical methods, the dependent variable is modeled as a linear combination of other variables. However, the last two methods consider the dependent variable to be of the distance type, while linear differential analysis is used for the nominal or rank dependent variables. Therefore, linear differential analysis is more similar to logistic regression. Linear diagnostic analysis is also similar to principal component analysis and factor analysis; Both of these statistical methods 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 dimensions of the data. However, these methods have major differences: in linear differential analysis, class differences are modeled, while in principal component analysis, class differences are ignored.

- **Fast independent component analysis(Fast ICA)**

A way to find the underlying factors or components of multivariate data is signal-dependent. The fast independent component analysis method is actually an optimized method of the independent component analysis method. Convergence of results in this method is faster than the independent component analysis method. The fast independent component analysis method is based on a fixed point algorithm that has a known performance speed. This algorithm has been modified compared to conventional fixed point algorithms, which has resulted in higher performance. Also, this algorithm is similar to some neural algorithms, and is computationally simple and requires less computational memory.

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

-Accuracy:

The accuracy of a model on a test data set is the percentage of data in that set that is properly labeled by the model.

-Perturbation matrix:

It is a useful tool and with its help, the performance of the model can be observed to detect different class tuples. An ideal model places most of the tuples on the original diameter of the perturbation matrix, and it is desirable that all elements other than the original diameter of the matrix have a value of zero or close to zero.

-rock curves:

Rock curves are a useful visual tool for comparing several models. The rock curve plotted 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 below its curve, the higher the accuracy of the model and vice versa.

- **Findings of export declarations**
- **Findings from fourteen models using fast independent component analysis with three features**

Fourteen tests of research models in Table (3) using the method of rapid independent component analysis with three characteristics showed that the accuracy of risk prediction of the random forest and bagging model is 77% higher than other models in this method.

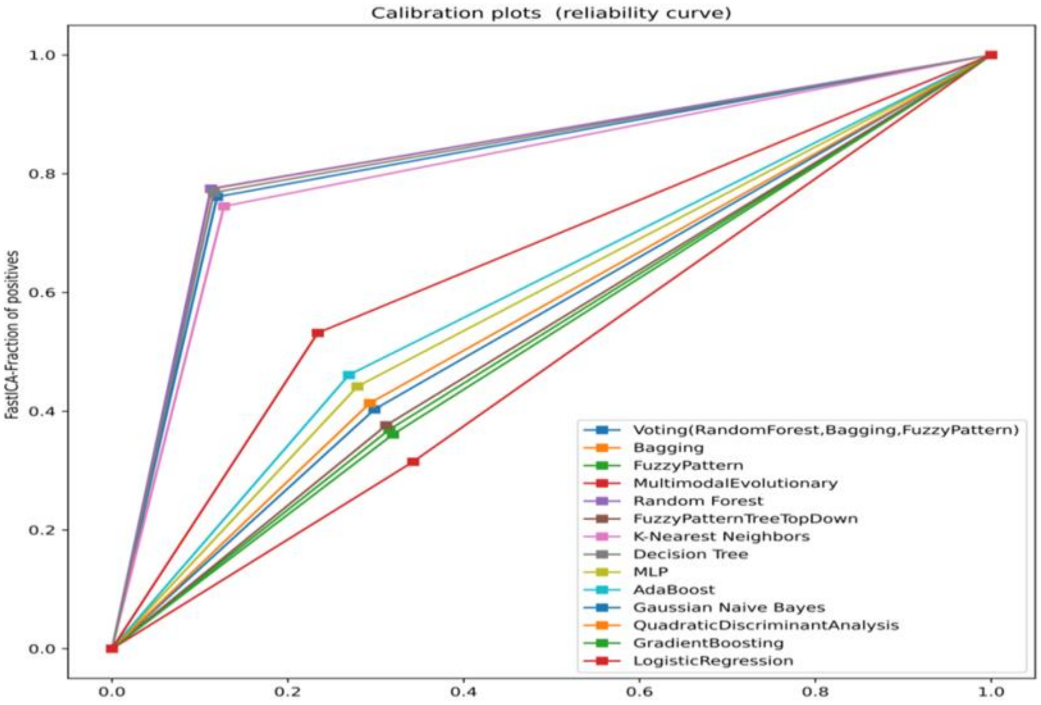
Table (3) Analysis of findings from the application of different algorithms with three characteristics

FastICA(3 com)

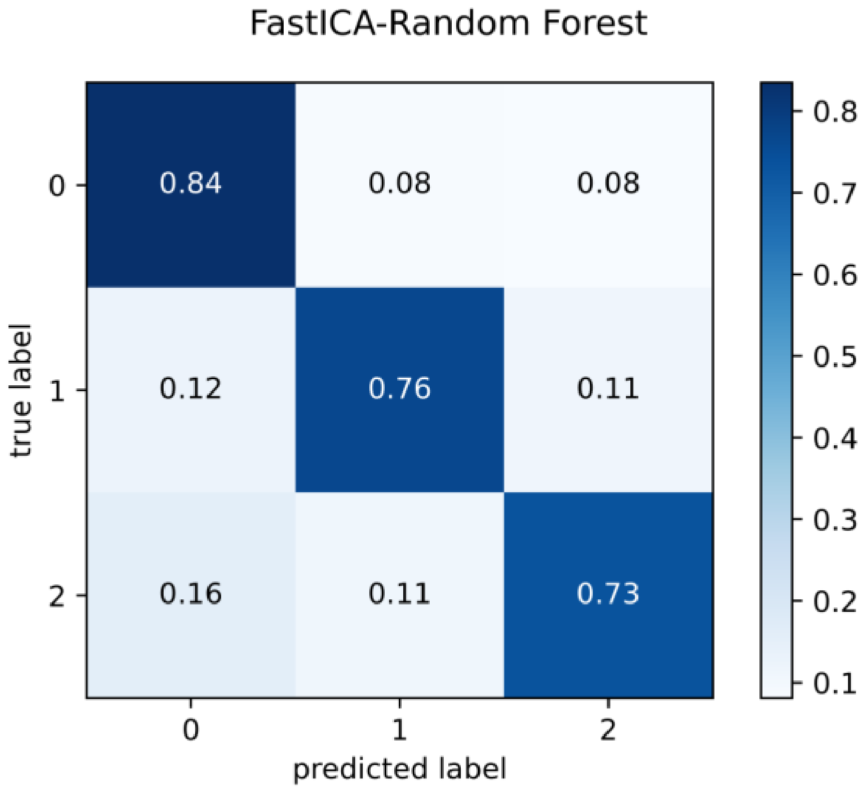
Method	Accuracy	macro avg	Label	precision	recall	f1-score
Voting(RandomForest,Bagging,FuzzyPattern)	76	76	0	73	79	76
			1	78	77	77
			2	77	73	75
Bagging	***77***	***77.5***	***0***	***72***	***84***	***77***
			1	***82***	***76***	***79***
			2	***79***	***73***	***76***
FuzzyPattern	37	34.5	0	31	9	14
			1	38	47	42
			2	37	51	43
MultimodalEvolutionary	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***
FuzzyPatternTreeTopDown	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
QuadraticDiscriminantAnalysis	41	41	0	43	23	30
			1	41	56	47
			2	42	42	42
GradientBoosting	36	21	0	0	0	0
			1	36	100	53
			2	0	0	0
LogisticRegression	53	53	0	59	39	47
			1	51	68	58
			2	53	50	51

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

- Finally, as shown in Figure (3), the rock curve compares the findings of the fourteen research models using a fast independent component analysis method with three characteristics. This model is higher than other models and bagging model.



- As can be seen in the stochastic matrix (1) of random forests in the principal component analysis method, the values of the main target variable, the red channel (label one), 76% of the data predicted in label one, actually belonged to label 1 (positive positive 1 or (positive True) and so does label zero with 84% and label two with 73%.



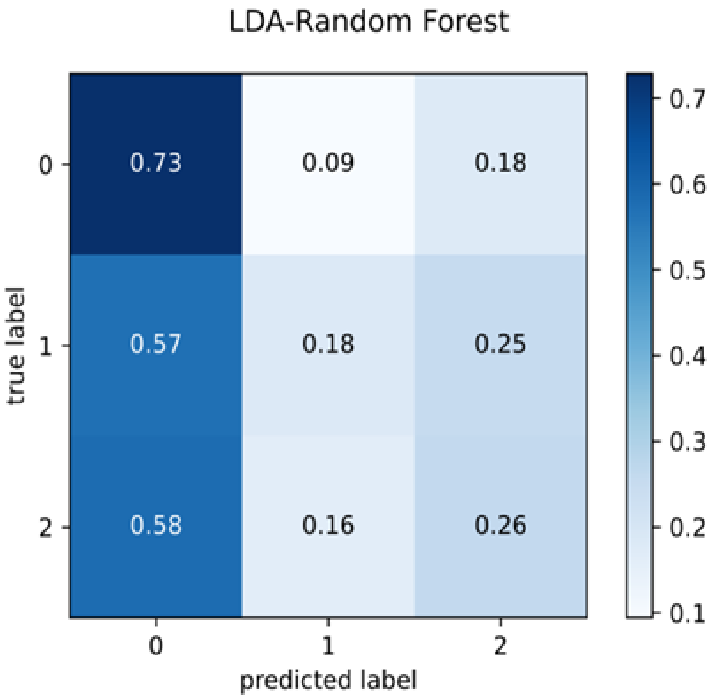
- **Findings from fourteen models using linear differential analysis with two features**

Fourteen tests of research models in Table (4) using the method of linear differential analysis with two characteristics showed that the accuracy of risk prediction of the random forest model is 37% higher than other models in this method.

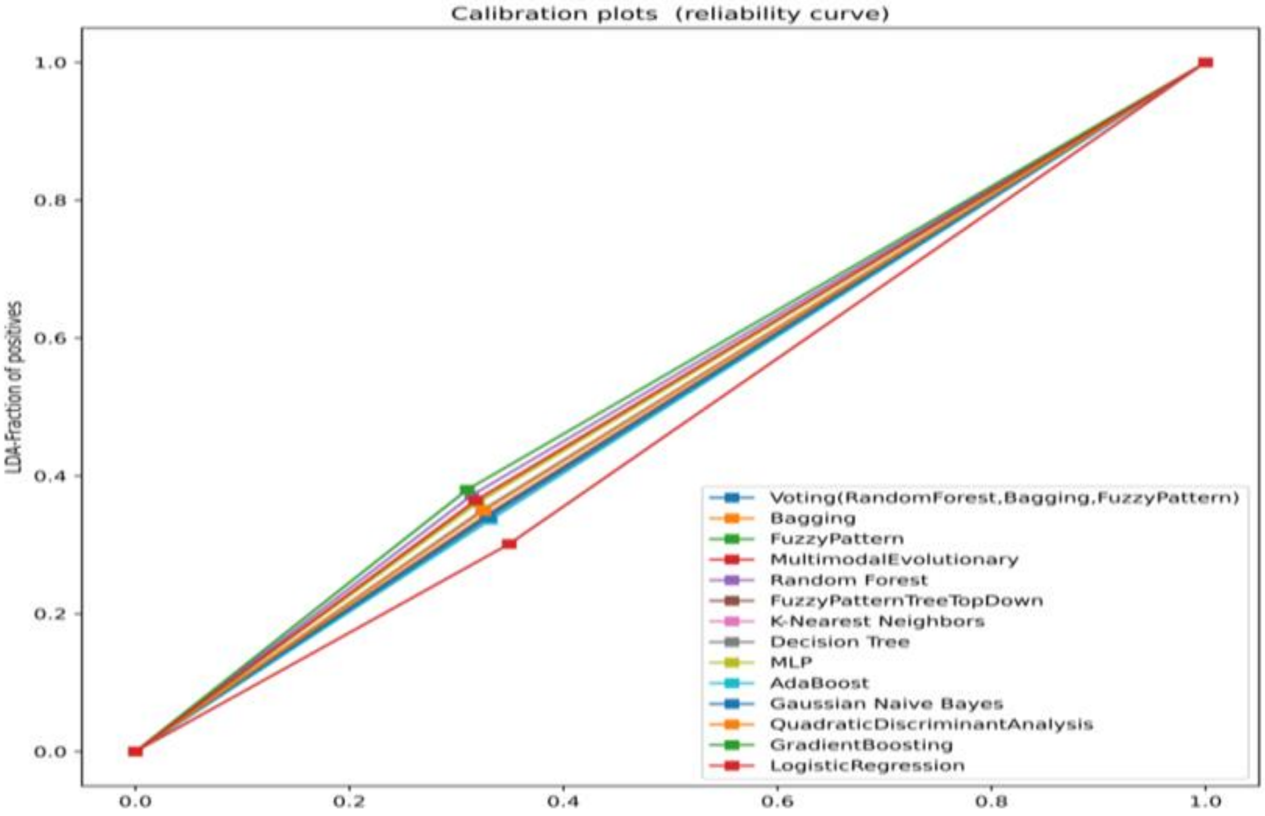
Table (4) Analysis of findings from the application of different algorithms with 2 features

LDA(2com)						
Method	Accuracy	macro avg	Label	precision	recall	f1-score
Voting(RandomForest,Bagging,FuzzyPattern)	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
FuzzyPattern	36	32	0	100	0	0
			1	36	100	53
			2	0	0	0
MultimodalEvolutionary	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***
FuzzyPatternTreeTop 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	100	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
QuadraticDiscriminantAnalysis	35	40.6	0	34	85	48
			1	100	0	0
			2	39	28	32
GradientBoosting	38	31.6	0	38	53	44
			1	0	0	0
			2	38	65	48
LogisticRegression	36	33.6	0	33	10	15
			1	36	89	52
			2	56	4	7

- In the linear differential analysis method, the accuracy of export declarations of 38% boosting gradient is higher than the model of random forests with 37% prediction accuracy. Predictive accuracy of 0% red class, 53% green class and 65% yellow class and due to lack of prediction in red class and due to forecast accuracy of 18% red class, 73% green class and 26% yellow class, the random forest model is a more suitable model and Is selected.
- As can be seen in the stochastic ambiguity matrix (2) of random forests in the linear differential analysis method, the values of the main target variable, the red channel (label one), 18% of the data predicted in label one, actually belonged to label 1 (positive positive value 1 or (positive True) and so does label zero with 73% and label two with 26%.



- As can be seen in Figure (3), the rock curve. Comparison of the results of the fourteen research models using the linear differential analysis method with two features, the larger area under the rock curve is the Boosting gradient model than other models, but due to the model preface Stochastic forests indicate the accuracy of the risk prediction of this model is higher than other models.



- **Comparison of findings and selection of the best model in export declarations**

Comparison of class prediction results in superior models in three main component analysis methods K-nearest neighborhood with 77% prediction accuracy and harmonic mean 77%, fast independent component analysis method of stochastic forest model with 77% prediction accuracy and harmonic mean 77.5% The linear differentiation of the stochastic forest model with 37% indicates the higher accuracy of the risk class prediction in the stochastic 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 levels of risk with an accuracy of 84% in the green class, 73% in the yellow class and 76% in the red class. Finally, since the accuracy of the stochastic forest model is higher than other models, so this model can be used as risk assessment in the customs organization in the proper identification of the risk category of export declarations as a suitable organizational knowledge.

- **Conclusion**

Using data mining techniques, the selection of declarations to be inspected can be significantly improved. Tests show that data mining techniques allow for effective targeting of declarations. The results of this study showed that the stochastic forest model was more accurate in identifying and determining the level and risk rating of export declarations than other models used. The rules resulting from this model are considered a hidden pattern in the Iranian customs database and can be used to predict high-level export declarations in the red, yellow and green channels and policies on declaration risk management. Applied for export In order to succeed in monitoring and targeting the declarations submitted to the customs to detect cases of risk and violation, it is necessary to manage the risk assessment by performing a series of basic data analysis in the form of data mining and developing an intelligent model for risk level forecasting. After forecasting and performing controls and obtaining the results, the correct results can be used to update and form more accurate risk profiles to improve the accuracy of the model forecast in future cases and as a tool that converts qualitative information into quantitative information. Intelligent risk management systems the capabilities of the automated risk assessment program provide customs with new opportunities such as faster detection and anticipation of the arrival of high-risk shipments.

END