Journal of Theoretical and Applied Information Technology
31st March 2022. Vol.100. No 6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPLEMENTATION OF DATA MINING ON CUSTOMS FALSE DECLARATION DETECTION

[1]ILHAM DEWANTORO, [2]TUGA MAURITSIUS
Information Systems Management Department, BINUS Graduate Program,
Master of Information Systems Management, Bina Nusantara University, Indonesia
E-mail: [1]ilham.dewantoro@binus.ac.id, [2] tmauritsus@binus.edu

ABSTRACT

Importation control in Indonesia is in need of an efficient method to oversee import declarations suspected of misleading information. This study offers a data mining method through Cross-Industry Standard Process for Data Mining (CRISP-DM) to provide a reliable method to predict false customs declarations. The method is tested using imports data from three major ports in Indonesia between 2019 and 2020. This study used decision trees algorithm as prior customs data mining study. The algorithm is compared with random forest and naïve bayes to detect false import declaration based on classification accuracy, precision, and recall scores. Testing results showed that random forest is the algorithm with the best classification accuracy, precision, and recall scores. Dataset preparation is very crucial since importation data has high cardinality and imbalance issues. This study used Synthetic Minority Oversampling Technique (SMOTE), normalization, and data transformation techniques to find a suitable dataset for a better prediction. The study finds that the model has a higher precision score than false declaration prediction based on whistleblower tips. This study is an addition to the limited amount of references pertaining to customs data mining.

Keywords: *Customs, False Declaration, Data Mining, CRISP-DM, SMOTE*

1. INTRODUCTION

Indonesian customs have developed a system to identify high-risk importation based on declaration channeling. Document verification and physical inspections of importation from medium and high-risk channels are conducted to ensure that import declarations represent actual conditions.

To assure the quality of Indonesian customs importation control, the internal audit office is involved in some of the physical inspections, especially after circumvention tips from whistleblowers. The number of false declarations found based on whistleblowers' tips is much higher than the number of false declarations found by routine inspections. However, there are still some occasions when the whistleblower tips are false. This condition requires a reliable method to improve the efficiency and accuracy of false declaration detection.

Data mining of historical data will be increasingly helpful for effective risk assessments and accurate decision-making. In business intelligence and analysis, data analytics is part of business intelligence and analytics technology related to data mining and statistical analysis [1].

Data mining can be utilized to perform predictive analysis. Several studies have used data mining as a basis for predicting an event. Also, in fraud detection, the data mining method is used for detecting fraud. Similarly, customs also face potential cases of fraud in declarations. Many studies have explored risk profiling with various data mining methods, such as statistical [2], classification [3], [4], and data assessing [5].

This study used CRISP-DM as a data mining method which consists of six stages, from business understanding stage to deployment stage. This paper also compares three algorithms, which are decision tree, random forest, and naive bayes, and evaluates them with classification accuracy, precision, and recall scores, considering misjudging a false declaration as truthful can be more harmful than misjudging a truthful declaration as false.

This study consists of five sections. The first section focuses on the introduction and business understanding. The second section focuses on the empirical study of data sourced. The third section focuses on the methodology in this study and the application of CRISP-DM. The fourth section discusses the study results, and the last section concludes the study.

2. LITERATURE REVIEW

2.1 Customs declaration and circumvention

The importation process started with the issuance of customs declaration by importers, which then transmitted in electronic forms via customs information system that records the information of importer identity, supplier identity, origin country, total value, total weight, total article, etc. Customs or import declarations have to state accurate information of products. Document and physical inspections are carried out to account for the declarations. Penalties are given to importers with false declarations.

The act of submitting a false declaration is similar as circumvention act. In this regard, the conditions for actionable circumvention is as below [6]:

1. Product modification, a slight modification of the product concerned to make it fall under customs codes which are normally not subject to the measures, provided that the modification does not alter its essential characteristics;
2. Parts imports to be assembled in the third country;
3. Transshipment circumvention, the consignment of the product subject to measures via third countries;
4. Lower duty rate circumvention, exportation through companies benefitting from a lower rate.

2.2 Data mining and CRISP-DM

Data mining has many definitions, e.g., the process of finding valuable chunks (valuable information) from a collection of raw materials (raw data) [7]. Data mining is also defined as a work that involves various disciplines and is repeatedly carried out in a structured manner to produce a mature work [8]. Data mining is also defined as a study to gain useful insights by conducting the process of collecting, cleaning, and analyzing datasets. [9].

This study used CRISP-DM as the data mining method, which includes six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

In prior customs data mining study, decision tree algorithms is used to establish customs risk and profile targeting [4]. In another customs data mining study, decision tree was also used to calculate classification based on cost in import to know the cost of committing a false negative or false positive error. The result of the general decision tree was compared to the decision tree with boosting to improve the accuracy of the prediction [3]. However, it needs to be compared to provide the best solution on false declaration detection. Since false declaration can be categorized as fraud, the comparison used algorithm on data mining research related to fraud detection. Random forest and naïve bayes were chosen since in the prior study of e-commerce crimes using the appropriate machine learning algorithms, random forest and naïve bayes exceeded the accuracy of decision tree [10]. Naïve bayes algorithm also exceeded decision tree on credit card scoring study [11].

This study compared three algorithms in the modeling stage: decision tree, random forest, and naïve bayes. Decision tree is one of the most popular modeling techniques in the data analysis industry. Decision tree as shown on *Figure* ***1*** is prediction model with hierarchies or tree structures and is most widely used in classification and prediction methods [12]..

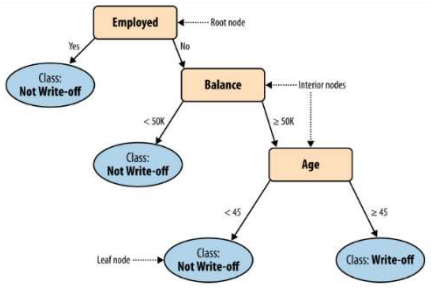

Figure 1: Simple classification tree [13]

Random forest is defined as a collection of decision trees where randomness has been explicitly incorporated into the model building process of each decision tree model [9]. The final result of random forest is often more accurate than the direct application of bagging on decision tree [9]. The architecture of random forest is as illustrated in *Figure 2*.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

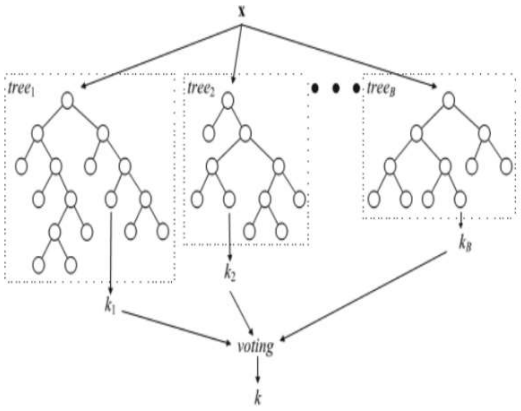

Figure 2: Architecture of random forest [10]

Naïve bayes is one of 10 data mining techniques often used in business applications [14]. Bayes classifiers are based on the bayes theorem for conditional probabilities. This theory quantifies the conditional probabilities of random variables (class variables), with general observations about the values of other sets of random variables (feature variables) [9]. As the equation (1), Naïve bayes goal is to determine posterior probability P(D|E) when E is observed as even and P (D) is the "prior" probability [9].

$$P(E) = \frac{P(E|D)P(D)}{P(E)} \qquad (1)$$

Evaluation is the step that ensures the model(s) has achieved the business objectives [15], technically correct and effective according to the data mining success criteria [16]. Most of the papers use metrics to evaluate the quality of trained models, and the metrics are also visualized to illustrate the results in each algorithm, e.g. confusion matrix [17].

This study used confusion matrix to evaluated the performance of models. It evaluated comparison between the number of prediction declaration and declaration are known to be truthful or false. The true positives (TP) is the number of false declarations true detected. The false positives (FP) is the number of actual false declaration predicted as truthful declarations. The true negatives (TN) is the number of truthful declarations detected correctly, and the number of false negatives (FN) is the number of missed truthful declarations.

The precision measures the proportion of TP among all positives (actual positive and prediction positive) or the percentage of truthful declarations among all predicted as true. The equation for the precision is given below :

$$Precision : \frac{TP}{TP+F} \qquad (2)$$

The recall measures the proportion of false declarations classified correctly. The equation for the recall is given below :

$$Recall : \frac{TP}{TP+FN} \qquad (3)$$

In this study, accuracy is the percentage of all transactions to the number of correct prediction for both false or truthful declaration. The accuracy equation is given below.

$$Classification\ Accuracy\ (CA): \frac{TP+F}{TP+FN+FP+TN} \qquad (4)$$

2.3 SMOTE.

Synthetic Minority Oversampling Technique or SMOTE is an algorithm that executes an oversampling approach to rebalance the original training set. The basis of the SMOTE procedure was to makes an interpolation among minority class instances which is able to increase the number of minority class instances by introducing new minority class examples, thereby the classifiers can improve its generalization capacity [18].

2.4 Related Work

There is a limited number of studies related to customs data mining, especially on the subject of false declaration detection, since any data related to the import export trade access tends to be sensitive [5]. Therefore this study serves as a new resource on customs data mining.

There are a few data mining methods used in customs studies. One study analysed data by statistically comparing the source data with other independent data. Comparisons are made of informations or evidence of known violations. Violations in the form of undervaluation can be identified by comparing prices between the source data and international trade data. Violation in the form of misclassification was identified by comparing the import data of the destination country based on the source data with the export data of the country of origin of independent data [2].

This study is supported by results of other studies which stated that imported data has high cardinality and is imbalanced. For these constraints data mining requires adjustment from the data side when the data is imbalanced [3]. As a solution, this study also incorporated SMOTE to find the influence imbalanced data has on data mining.

In order to complement previous studies, this study will describe the whole process of data mining,

therefore readers will be informed of critical steps in this study and the results.

Regarding algorithms applied in this study, the decision tree algorithms that were used to establish customs risk and profile targeting by previous study [4] *are taken as a method*. In other research, decision tree is also used to calculate classification based on cost in import so that it is known that the cost of committing a false negative or false positive error. The result of the general decision tree is compared to the decision tree with boosting to improve the accuracy of the prediction [3]. Another algorithm that used in similar research is SVM. It is used to detect fraud by combining with EasyEsemble algorithm [5].

As stated in previous paragraphs, past studies tended to use decision tree algorithms, therefore this study will present other algorithms as comparisons and eventually learn the best algorithm in customs data mining. Since circumvention patterns can be categorized as fraud, the comparison model will use data mining research related to fraud.

In the study of e-commerce crimes using the appropriate machine learning algorithms between Decision Tree, Naïve Bayes, Random Forest, and Neural Network, it is known that the accuracy of random forest is 95 percent, Naïve Bayes is 95 percent exceeds the accuracy of Decision tree is 91 percent [10]. For research detecting credit card fraud it is known that the results of Naïve Bayes prediction against negative predictive value reached 99.60%, then Decision Tree and random forest has a value that is not much different that reaches 97.59% and 97.53% [19]. Different results were obtained in Fraud Detection in Credit Card Data Using Machine Learning Techniques where random forest is the most accurate algorithm (99.95%) compared to linear regression and naïve bayes (91.16% and 89.35%) [20].

Based on comprehensive studies on comparison of algorithms used in financial fraud during 2009 to 2019, it is known that SVM, random forest, and naïve bayes become the three most commonly used algorithms in fraud prediction. Naïve bayes are known to have advantages as they are easy to implement and capable of scaling with datasets while random forest has the advantage of being fast during runtime and can work with unbalanced data. Different from the others, SVM has the drawback as being difficult to process datasets with high dimensions [21].

For those reasons, random forest and naïve bayes algorithms are fit to be used as comparisons. This study will focus on the comparisons to discover which algorithm is the best for detecting false customs declarations. The finding will be added values to the result of this study.

This study used CRISP DM as the most preferred methodology that has been used in data mining domain [22] and was applied in many prior studies. The discussion section of this study is presented in subsections according to each phase in CRISP DM. This is to help readers in mastering practical steps in running CRISP DM method.

3. METHODOLOGY

This study was conducted to provide a method that improves the effectiveness and efficiency of customs declaration control activities. The research methodology used is CRISP-DM framework as illustrated in Figure 3.

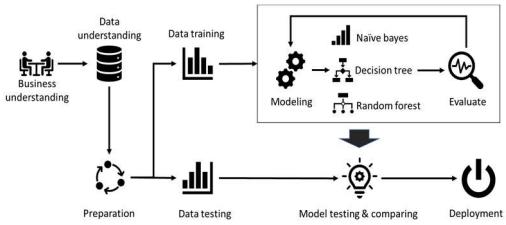

Figure 3 : Methodology

The implementation of research using the method of CRISP-DM stages follows:

3.1 Data understanding

Data understanding as the second stage of CRISP-DM will determine what data will be used in the data mining, based on the first stage's business understanding, which includes comprehension of business and circumvention acts. Data understanding stage includes the process of data collection and data interpretation.

This study used historical data on importation for the period 2019 to 2020. The data consists of import declaration in which documents had been verified and underwent physical inspection, whether it is of imports that passed the inspections or of imports that were given penalties.

This study used 239.465 records of declaration and article details, as many as 4.437.351 records. Data understanding is executed by selecting data that will be included in the modeling process. In this process, collected data is adjusted as the dataset model required.

3.2 Data Preparation

The data preparation process starts with data The data preparation process starts with data identification (describing), data format changes, and

merging data (transformation), feature and target determination (selection), and dataset cleansing.

Data identification is a process to find available data, including selecting data needed in the data transformation process. Data transformation is a process that involves merging data, aggregating records, deriving new attributes, sorting data, removing/altering missing values, data reduction, data numerosity reduction, and data dimensional reduction.

That is implemented according to the dataset model format. Subsequently, data cleansing is a method to format the missing value on a data set. The last step on data preparation is dividing the dataset into training dataset and validation dataset in the composition of 80% training dataset and 20% validation dataset.

The selection of data also weighs the distribution between false and truthful declarations to avoid bias on predictions. This study also used SMOTE to balance dataset between the truthful declaration and false declaration.

3.3 Modeling

This study used decision tree, random forest, and naïve bayes as data mining models. Training datasetwill be processed into these three algorithms. Tuning or adjusting the model is also carried out in this process to obtain the best results. In doing modeling calculations, one of the python-based open-source applications, Orange Data Mining, is commonly used to perform data mining calculations. Orange is an application that uses the python language to help in data visualization. Orange can also perform predictive modeling, analysis, subset selection, and empirical analysis, including performing tasks such as data manipulation and data transformation [19].

Each model is developed using 80% training data with repeated calculations using ten folds. Model validation is run by implementing the model on 20% remaining data to know whether the model has overfitting or underfitting characteristics.

3.4 Evaluation

The three models, decision tree, random forest, and naïve bayes, were evaluated by comparing the score of classification accuracy, precision, and recall. Classification accuracy evaluates the accuracy of true prediction on actual conditions while precision is essential to predict truthful declaration. Recall is also used to evaluate how good a model can predict false declaration among actual data. All of the models will test the actual declarations that are suspected false based on whistleblower tips. Model with the highest result indicates the best potential of deployment.

4. DISCUSSION

4.1. Dataset

The study used data from the internal audits office regarding import declaration of 2019 and 2020 in selected customs offices. The data selection was then narrowed down to import declarations of the red and yellow channels. This channeling system determines whether an import needs physical inspection after document verifications have been conducted.

The data table that used in this study are listed below:

- Header. The table contains data about the information associated with import identity submitted by the importer.
- Articles. The importer also submitted the articles information e.g. articles name, harmonized system (HS) code, weight, etc in this table.
- Payment. The information of tax and customs payment that must be paid by the importer.
- Flag. This table provides historical information of penalties given to importers with false declarations.
- Tip. Information of declaration that suspected false from whistleblower

This study only used data related to import business process and circumvention such as supplier identity, origin country, articles, tax payment, etc. Datamining evaluation will be tested with tip data in order to confirm whether data mining has better prediction accuracy.

Table 1 : Features

Features	Type	Description
X2	Categorical	origin country similarity
X3	Categorical	importer similarity
X4	Categorical	address similarity
X5	Categorical	transshipment usage
X6	Number	Container

X7	Number	Articles
X8	Number	count of two-digit hs code
X9	Number	count of four-digit hs code
X10	Number	count of six-digit hs code
X11	Number	count of eight-digit hs code
X12	Number	count of price
X13	Number	sum of packages
X14	Number	count of packages
X15	Number	count of articles declaration
X16	Number	origin country of articles
X17	Number	count of price code
X18	Number	count of articles netto
X19	Number	sum of articles value
X20	Number	Value
X21	Number	Netto
X22	Number	Bruto
X23	Number	Tax
Target		
Flag	Categorical	F for false declaration and NF for truthful declaration

4.2. Data Preparation

This process is about converting the data to a required format. Data preparation is a process that includes data selection, cleaning, transformation, aggregation, etc. This stage used Microsoft SQL Server Express version 15.0.2000.5 and SQL Server Management Studio version 15.0.18338.0.

As customs declaration has high cardinality data [5], the data preparation stage dominates time allocation in data mining compared with other stages [15]. This stage involves aggregating HS Codes after altering it into HS Codes with two digits, four digits, six digits, and eight digits. Besides aggregation of HS Codes of articles, several variables such as origin country, net weight, value, articles, etc., also need aggregation due to high cardinality data issues.

Data preparation also includes deriving new attributes to check anomalies on import declarations in similar data such as origin supplier country to origin sender country, supplier address to sender address, and whether import use transshipment or not. Import declaration, which uses transit ports, is categorized as transit, whereas the other is categorized as not transit.

Deriving new attributes is essential since some features have more than 100 variables. This can't be processed on tree-based algorithms. Decision tree creates a decision support tool in a tree with an internal node and require high computational power during the initial setup [20]. When a feature has 100 variables, it will create more than 100 nodes that exceed the processing ability of devices used in testing.

Table 1 describes 22 features applied for the data mining process after the data preparation process. After removing duplicates data, the data preparation process generated a dataset consisting of 237.212 records, 151.439 truthful declarations, and 85.773 records of false declarations as shows in *Figure 4*. That is including a tips data that will be used as a testing dataset in the evaluation stage. The ratio between truthful declarations and false declaration data is 36,16% to 63,84%.

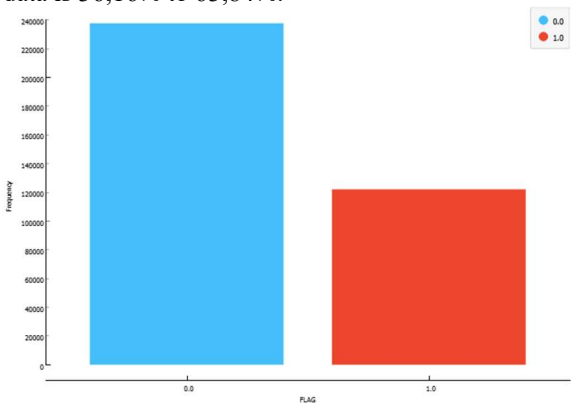

Figure 4 : Dataset distribution

Based on this dataset condition and the imbalance between the truthful declaration and false

declaration, another dataset is created to balance class distribution using SMOTE. This method is also applied in a previous study that has 70,4% truthful declarations and 29,6 false declarations [3]. After rebalance process using SMOTE, 302.630 datasets are distributed equally. Each truthful declaration and false declaration has 151.315 data, as shown in *Figure 5*.

Furthermore, the dataset also has a positive skew of class problems. That is the distribution of some features has a tail to the right side of the distribution [21]. So another dataset, dataset log 10, is made by calculating the feature with positive skew with log10(x) or log10(x+1) when data has 0 value.

Moreover, some data has a remarkably wide distribution while others have a narrow distribution on dataset log10. For this condition, another dataset is made after normalization with 1 as maximum interval and 0 as minimum internal. Six datasets format are used for this study, as detailed in this The last step of data preparation is splitting dataset to 80% training dataset and 20% validation dataset. The percentage of training dataset used in this study is about the same as previous studies, which used 70%-80% training dataset [3], [5]. 20% remaining data, used for validation dataset, will validate the model after developed on training dataset.

*Table **2**.*

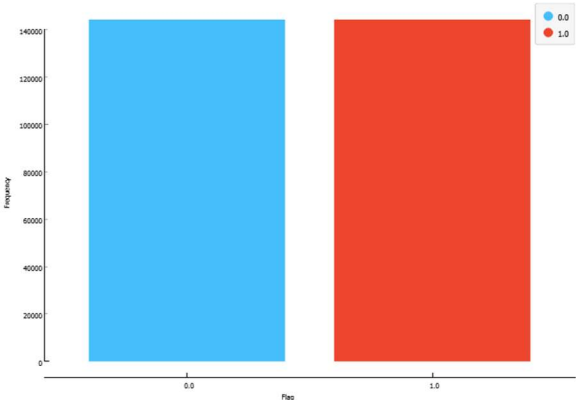

Figure 5: Dataset distribution with SMOTE

The last step of data preparation is splitting dataset to 80% training dataset and 20% validation dataset. The percentage of training dataset used in this study is about the same as previous studies, which used 70%-80% training dataset [3], [5]. 20% remaining data, used for validation dataset, will validate the model after developed on training dataset.

Table 2: Datasets

Dataset default
Dataset default SMOTE
Dataset log 10
Dataset log10 SMOTE
Dataset log10 normalization
Dataset log10 normalization SMOTE

4.3. Modeling

Training dataset is performed using three algorithms, naïve bayes, decision tree, and random forest. The results of these experiments are obtained from calculation of actual data and predicted results in confusion matrix.

This modeling used python Orange Data Mining version 3.30.1 widget. In experiments using naive bayes (NB), the model is developed using default parameters on Orange Data Mining since the tuning process is not applicable to naïve bayes. For the decision tree model, the model is developed using a Tree widget with default parameters and tuning parameters in experiments using decision tree, as shown in TABLE 3. The last modeling stage using random forest is performed using three variations on six datasets. First variation of random forest (RF) uses default parameters on random forest widget on orange data mining, which uses ten trees and ten attributes. Second variation of random forest uses ten trees with five attributes. Last variation of random forest uses 15 trees and five attributes as *Table **4** : Random ForestTable **4**.*

Table 3 : Decision Tree

Parameters	Default (Tree)	1st Tuning (Tree 50)	2nd Tuning (Tree 20)
Min. number of instances	2	3	2
Don't split subset smaller than	5	5	10
Limit the max tree depth	100	50	20

Table 4 : Random Forest

Parameters	Default (RF)	1st Tuning (RF 50)	2nd Tuning (RF 100)
Number of trees	10	50	100
Number of attributes considered at each split	10	10	10
Do not split subsets smaller than	10	10	10

Every model is processed with ten-fold validation of six data training. The result of training process as shown on *Table 6* and *Table 5*. The result of each experiment is validated with validation dataset to make sure that there is not any combination that has an overfitting or underfitting result as shown on *Table 7* and *Table 8*.

Table 5 : Non SMOTE Data Training Result

Model	Dataset	NON SMOTE		
		CA	Precision	Recall
NB	Default	0,65	0,52	0,53
NB	Log10	0,65	0,52	0,53
NB	Normalize	0,65	0,52	0,53
RF	Default	0,72	0,64	0,49
RF	Log10	0,72	0,64	0,50
RF	Normalize	0,72	0,64	0,49
RF 100	Default	0,73	0,68	0,48
RF 100	Log10	0,73	0,69	0,49
RF 100	Normalize	0,73	0,69	0,49
RF 50	Default	0,73	0,68	0,49
RF 50	Log10	0,73	0,68	0,49
RF 50	Normalize	0,73	0,68	0,49
Tree	Default	0,66	0,53	0,50
Tree	Log10	0,66	0,53	0,50
Tree	Normalize	0,66	0,53	0,49
Tree 20	Default	0,69	0,59	0,46
Tree 20	Log10	0,69	0,59	0,46
Tree 20	Normalize	0,69	0,59	0,47
Tree 50	Default	0,66	0,53	0,51
Tree 50	Log10	0,66	0,53	0,51
Tree 50	Normalize	0,66	0,53	0,51

Table 6 : SMOTE Data Training Result

Model	Dataset	SMOTE		
		CA	Precision	Recall
NB	Default	0,66	0,68	0,63
NB	Log10	0,64	0,65	0,59
NB	Normalize	0,63	0,65	0,58
RF	Default	0,77	0,79	0,73
RF	Log10	0,76	0,76	0,74
RF	Normalize	0,74	0,74	0,73
RF 100	Default	0,79	0,82	0,73
RF 100	Log10	0,78	0,79	0,76
RF 100	Normalize	0,76	0,77	0,75
RF 50	Default	0,78	0,82	0,73
RF 50	Log10	0,77	0,79	0,76
RF 50	Normalize	0,76	0,76	0,75
Tree	Default	0,72	0,73	0,69
Tree	Log10	0,70	0,71	0,68
Tree	Normalize	0,68	0,69	0,66
Tree 20	Default	0,74	0,76	0,70
Tree 20	Log10	0,70	0,71	0,69
Tree 20	Normalize	0,69	0,69	0,68
Tree 50	Default	0,72	0,72	0,70
Tree 50	Log10	0,70	0,70	0,69
Tree 50	Normalize	0,68	0,68	0,68

Table 7 : Non SMOTE Data Validation Result

Model	Dataset	NON SMOTE		
		CA	Precision	Recall
NB	Default	0,66	0,53	0,53
NB	Log10	0,65	0,52	0,54
NB	Normalize	0,65	0,52	0,53
RF	Default	0,72	0,65	0,50
RF	Log10	0,72	0,64	0,49
RF	Normalize	0,72	0,65	0,50
RF 100	Default	0,74	0,69	0,49
RF 100	Log10	0,73	0,68	0,49
RF 100	Normalize	0,74	0,69	0,49
RF 50	Default	0,73	0,69	0,49
RF 50	Log10	0,73	0,68	0,49
RF 50	Normalize	0,73	0,68	0,49
Tree	Default	0,66	0,54	0,49
Tree	Log10	0,66	0,54	0,49
Tree	Normalize	0,66	0,54	0,49
Tree 20	Default	0,69	0,60	0,47
Tree 20	Log10	0,69	0,59	0,45
Tree 20	Normalize	0,69	0,60	0,46
Tree 50	Default	0,66	0,54	0,51
Tree 50	Log10	0,66	0,53	0,51
Tree 50	Normalize	0,66	0,54	0,51

Table 8 : SMOTE Data Validation Result

Model	Dataset	SMOTE		
		CA	Precision	Recall
NB	Default	0,66	0,68	0,63
NB	Log10	0,64	0,65	0,58
NB	Normalize	0,63	0,65	0,58
RF	Default	0,77	0,79	0,73
RF	Log10	0,76	0,76	0,75
RF	Normalize	0,74	0,74	0,73
RF 100	Default	0,79	0,83	0,73

RF 100	Log10	0,78	0,79	0,77
RF 100	Normalize	0,76	0,77	0,75
RF 50	Default	0,78	0,82	0,73
RF 50	Log10	0,78	0,78	0,77
RF 50	Normalize	0,76	0,77	0,75
Tree	Default	0,72	0,73	0,70
Tree	Log10	0,71	0,71	0,69
Tree	Normalize	0,69	0,69	0,66
Tree 20	Default	0,74	0,77	0,70
Tree 20	Log10	0,71	0,70	0,71
Tree 20	Normalize	0,70	0,70	0,69
Tree 50	Default	0,72	0,72	0,71
Tree 50	Log10	0,71	0,70	0,71
Tree 50	Normalize	0,69	0,69	0,68

Referring to the validation dataset test, the result is not different when the model is applied to training dataset. It is a good sign that the model learning has no underfitting or overfitting characteristics. Overall the SMOTE dataset has a higher accuracy score than non-SMOTE dataset. The result shows that random forest 2nd tuning on default dataset with SMOTE has the highest CA score of 0,79. This combination also has the highest precision score of 0,83. This precision result means that 83% false declaration is truly predicted by the model. The random forest 2nd tuning is also the highest recall combined with dataset log10 with 0,77. It means that this model can truly predict 77% actual false declarations.

After the learning and validation process of three algorithms and each tuning, the random forest 2nd tuning is the algorithm that has the highest on all scores, accuracy, precision, and recall. The next stage is to evaluate these models to the real dataset to know the effectiveness of the model when predicting the false declaration.

4.4. Evaluation

In order to test which one of three model is applicable, they are evaluated in a real datasets or testing datasets. The number of testing datasets is 488 records of import declarations that are suspected as false declarations. The precision of this actual data is 0,72. The goal of evaluation process is to know the model that has a better precision of actual data precision. That model will be proposed to be used in predicting the false declaration to fulfill the business need.

Table 9 : Data Testing Result

Model	Default	Log10	Normalize
NB	0,83	0,83	0,82
RF	0,86	0,85	0,85
RF 100	0,90	0,87	0,87
RF 50	0,89	0,87	0,87
Tree	0,85	0,86	0,85
Tree 20	0,88	0,83	0,84
Tree 50	0,86	0,83	0,84

After evaluation, the result on *Table 9*. shows that random forest 2nd tuning has the highest score prediction on testing datasets. This precision score is higher than the model when previously tested on validation data because the actual test data was not randomly chosen from training or validation data. The random forest 2nd tuning on the default dataset with SMOTE has a 0,90 score of precision while the actual data has only 0,72. This precision is significantly better than the whistleblower tips method. This result refers to the confusion matrix of prediction widget on orange data mining while processed on random forest 2nd tuning as *Figure 6*

Confusion matrix for Rf 100 (showing number of instances-default SMOTE)

		Predicted		
		F	NF	Σ
Actual	F	256	68	324
	NF	30	94	124
	Σ	286	162	448

Confusion matrix for Rf 100 (showing number of instances-log10 SMOTE)

		Predicted		
		F	NF	Σ
Actual	F	277	47	324
	NF	41	83	124
	Σ	318	130	448

Confusion matrix for Rf 100 (showing number of instances-nofmalize SMOTE)

		Predicted		
		F	NF	Σ
Actual	F	277	47	324
	NF	41	83	124
	Σ	318	130	448

Figure 6 :Confusion Matrix

This study also found the most influential feature on data mining process as shown on *Figure 7*. Computation of feature importance was done on the random forest algorithm which has number of trees amounting to 100 as the algorithm with the highest precision score. Apparently the most influential feature in precision calculation was the total price in imports notifications (X12), followed by several other features, which were the quantities of containers (X6), the quantities packaging types (X13), the quantities of goods according to HS 4 digits (X9), and netto declared in imports documents (X21).

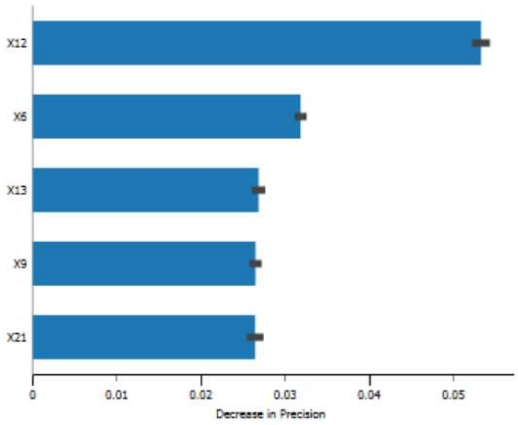

Figure 7 :Features Importance

These results provides a new insight for the organization to give more attention to a more extensive and thorough analysis of false declarations detection.

They can also be useful in further research on customs declaration data mining, particularly in exploring declaration validity value based on those variables, so that detection analysis could be conducted more extensively.

4.5. Deployment

The implementation of data mining in customs control started after importers input customs declarations to the customs system. Officers can assess false declarations using data mining and select which customs declaration is predicted to be truthful or false. The learning model will run continuously along with regular updates of two years period customs declaration data as training datasetparameters.

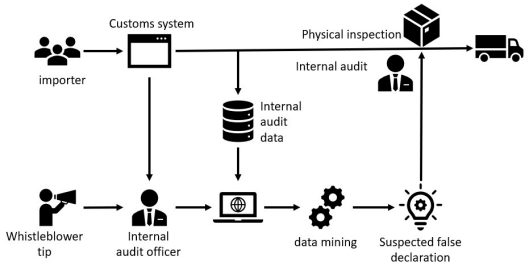

Figure 8 :Implementation Concept

The internal audit office assesses import declaration using 22 features to test the dataset on the model. If data mining prediction detects false declarations, the internal audit officer may arrange supervision activities to examine import declaration. But when the model detects whistleblower tips as a truthful declaration, the subsequent processing of import declaration may proceed as normal by customs risk management

5. CONCLUSION

The results of this study show that the data mining process using CRISP DM is a practical solution for an internal audit which needs it as a decision support system in predicting false declarations. This study also serves as a new resource on the currently limited amount of customs data mining literature.

The algorithms considered were decision tree, random forest, and naïve bayes using python Orange Data Mining. Datasets must be preserved as data preparation stages to solve high cardinality data issues. Some data needs aggregations, and others need particular formats. SMOTE is suitable for this study in resolving imbalanced import declarations data.

This study finds that random forest with 100 Trees, and 10 attributes considered at each split, also with no splitting subsets smaller than 10 set as the model with the highest accuracy and precision scores among three algorithms. This algorithm process on the normalized dataset has a 0,90 score of precision while the actual data has only a 0,72 score.

This study has some limitations in processing some variables with high cardinality issues. Future research should consider the use of leveling on some attributes e.g. country of origin, HS code, port, etc.

REFERENCES:

[1] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Q. Manag. Inf. Syst.*,

vol. 36, no. 4, 2012, doi: 10.2307/41703503.

[2] C. Chalendard, G. Raballand, and A. Rakotoarisoa, "The use of detailed statistical data in customs reforms: The case of Madagascar," *Dev. Policy Rev.*, vol. 37, no. 4, 2019, doi: 10.1111/dpr.12352.

[3] X. Zhou, "Data mining in customs risk detection with cost-sensitive classification," *World Cust. J.*, vol. 13, no. 2, 2019.

[4] B. Chermiti, "Establishing risk and targeting profiles using data mining: Decision trees," *World Cust. J.*, vol. 13, no. 2, 2019.

[5] J. Vanhoeyveld, D. Martens, and B. Peeters, "Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue," *Pattern Anal. Appl.*, vol. 23, no. 3, 2020, doi: 10.1007/s10044-019-00852-w.

[6] E. Vermulst, "EU Anti-Circumvention Rules : Do They Beat the Alternative ?," *SSRN Electron. J.*, 2015, doi: 10.2139/ssrn.2637796.

[7] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J. Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0030-3.

[8] F. Provost and T. Fawcett, *Data Science for Business. What You Need to Know About Data Mining and Data-Analytic Thinking*, vol. First Edit. 2013.

[9] C. C. Aggarwal, *Data Mining : The Text Book*, vol. 14, no. 3. 2015.

[10] A. Saputra and Suharjito, "Fraud detection using machine learning in e-commerce," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 9, 2019, doi: 10.14569/ijacsa.2019.0100943.

[11] E. D. Madyatmadja and M. Aryuni, "Comparative study of data mining model for credit card application scoring in bank," *J. Theor. Appl. Inf. Technol.*, vol. 59, no. 2, 2014.

[12] C. C. Lin, A. A. Chiu, S. Y. Huang, and D. C. Yen, "Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments," *Knowledge-Based Syst.*, vol. 89, 2015, doi: 10.1016/j.knosys.2015.08.011.

[13] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, 2013, doi: 10.1089/big.2013.1508.

[14] W. C. Lin, S. W. Ke, and C. F. Tsai, "Top 10 data mining techniques in business applications: a brief survey," *Kybernetes*, vol. 46, no. 7. 2017, doi: 10.1108/K-10-2016-0302.

[15] V. Plotnikova, M. Dumas, and F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Comput. Sci.*, vol. 6, 2020, doi: 10.7717/PEERJ-CS.267.

[16] IBM, "IBM SPSS Modeler CRISP-DM Guide," *IBM*, 2011.

[17] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, 2021, vol. 181, doi: 10.1016/j.procs.2021.01.199.

[18] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research*, vol. 61. 2018, doi: 10.1613/jair.1.11192.

[19] Y. Kültür and M. U. Çağlayan, "Hybrid approaches for detecting credit card fraud," *Expert Syst.*, vol. 34, no. 2, 2017, doi: 10.1111/exsy.12191.

[20] A. K. Rai and R. K. Dwivedi, "Fraud Detection in Credit Card Data Using Machine Learning Techniques," in *Communications in Computer and Information Science*, 2020, vol. 1241 CCIS, doi: 10.1007/978-981-15-6318-8_31.

[21] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Computer Science Review*, vol. 40. 2021, doi: 10.1016/j.cosrev.2021.100402.

[22] T. Mauritsius, A. S. Braza, and Fransisca, "Bank marketing data mining using CRISP-DM approach," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, 2019, doi: 10.30534/ijatcse/2019/71852019.