Application of Business Intelligence for Customs Declaration: A Case Study in Indonesia

Canrakerta[1]**, Achmad Nizar Hidayanto**[2]**, Yova Ruldeviyani**[2]

[1] Ministry of Finance, Indonesia
[2] Faculty of Computer Science, Universitas Indonesia, Indonesia

Corresponding author e-mail: canrakerta@kemenkeu.go.id

Abstract. The customs declaration made by the self-assessment needs to be re-examined by the document inspector. Customs declaration may have potential errors to determine even intentionally or not. However, the re-examination by the document inspector has not been optimal. This study uses business intelligence to address this problem by providing analysis capabilities for document inspectors. This research focuses on developing a data warehouse with the Kimball methodology. The results of research in the form of data warehouse design. The data warehouse will be used later for dashboards, OLAP, and data mining in order to detect fraud. The data mining algorithms used are firmness tree, support vector machine, neural network, and several ensemble methods. The results of the data mining process show that SMOTE has a significantly increased sensitivity score than non-SMOTE technique.

1. Introduction

The Directorate of General Customs and Excise (DGCE) is a part of the Ministry of Finance (MoF) which has the function of formulating and implementing policies to conduct supervision, law enforcement, services, and also to optimize revenues in customs and excise fields based on the Law in Force [1]. In 3 years, the number of services to import documents always increases. The growth in import documents needs to be an interest in DGCE. The reason for the interest is the value of state revenue and requires more effective supervision to maintain the time of completion of goods at the port.

Import documents are submitted by self-assessment to DGCE based on the international agreement of The General Agreement on Tariffs and Trade (GATT). This process helps the acceleration of the import process from the goods enter the customs area until the gate. However, the self-assessment process has potential errors. For example, errors in defining the classification of goods and customs value intentionally or not. This problem may result in a lack of state revenue as well as the missed of prohibited and restricted commodity goods.

DGCE needs to recheck the import documents submitted through an examination mechanism conducted by the document inspector. However, the examination carried out by the document inspector has not been as expected. There is still a difference in the determination that occurs between the document inspector, which causes reduced revenue collection and restricted commodity could be entered Indonesian area. This condition is very crucial because the total value of the examination mechanism over the past five years has reached trillions of Rupiah. Based on interviews conducted, this happened because of differences in analytical skills possessed by each document inspector. The lack of analytical skills occurs because there are no tools that can provide the analysis capabilities of the document inspector.

The application of Business Intelligence (BI) will be used as terminology to solve the problem of the lack of tools that can help document inspector in conducting analysis. In its application, the research will focus on developing a data warehouse for analysis when conducting import document research. The data warehouse design is also closely related to the development of end-user access tools in the data warehouse architecture proposed by Connolly and Begg [2]. Development of end-user access tools, including user visualization through dashboards, OLAP, and data mining utilization. The research question of this study is how the implementation of the data warehouse design and data mining approach can provide information needs for document inspector.

Through the implementation of the data warehouse, the document inspector can examine import documents more efficiently. The problem of document inspection is not optimal so far could be minimized. It also increases state revenue on taxes on imported goods and makes security on goods even better. The objectives of this research include (1) creating a data warehouse design for document inspection; (2) using data warehouse to do document inspection; (3) create a visualization such as dashboard and OLAP into an information system; and (4) conducting a data mining process to create model for detection of import documents that indicated as fraud documents.

2. Literature Study

2.1. Data Warehouse

Sharda, Delen, Turban, Aronson, and Liang declared that a data warehouse is a gathering place for data produced to support decision making [3]. They also said that the data warehouse is a collection of current or past data that has a unique attraction for the top leadership. Generally, the data used is structured data for analysis needs. Related to BI, the term of the data warehouse is commonly used as a gathering place for data and can be used as material for analysis.

Sharda, Delen, Turban, Aronson, and Liang explained that two approaches could take in developing a data warehouse [3]. The first approach, according to Bill Inmon, developing a data warehouse can be done with a top-down mechanism that adopts a traditional relational database to build Enterprise-wide Data Warehouse (EDW) [4]. The second approach, according to Ralph Kimball, developing a data warehouse can be done with a bottom-up mechanism that created dimensional modelling or commonly also known as the data mart approach [5].

In designing the data warehouse, some dimensional models can be developed based on the star scheme, snowflake scheme, and constellation scheme models. Based on these models, Kimball and Ross explained that there were four critical decisions in designing dimensional modelling, namely business process selection, grain determination, dimension table identification, and fact table identification [6].

2.2. Data Mining

According to Han, Kamber, and Pei, all types of data can be used in data mining, as long as the data has a meaning that is relevant to our purpose [7]. The most basic types of data that commonly used as a source of the data mining process are an operational database, data warehouse, and transactional data.

According to Kotu and Deshpande, problems that are solved by data mining generally can be categorized into two learning methods, namely supervised and unsupervised learning [8]. The supervised learning method utilizes data that already has labels to determine labels on new data. Meanwhile, unsupervised learning methods are used to see hidden patterns of data that do not have a label.

Data mining is one of the end-user access tools in data warehouse architecture. Data mining is also one of the technological approaches to building BI in an organization, so the mining process in this research is carried out to assist the analysis process of document examiners. The mining process carried out in this study uses classification techniques. Classification is a data mining approach that uses data source as data training and data test to produce a model that can provide predictions [8].

The classification approach in this study was carried out to predict the fraud of import documents. Historical data of import documents which has been verified by the document inspector was used as a data source. The resulting prediction in the form of a model or pattern can be used by the document inspector to determine the attributes that need attention.

Han et al. explained that the process of data mining is iterative and consists of several steps [7]. These steps are (1) data cleaning (to eliminate noise and inconsistent data); (2) data integration (to make a combination of data from various data sources owned); (3) data selection (to select data relevant to the analysis to be conducted); (4) data transformation (according to the needs of the mining process); (5) data mining (the application of mining techniques to generate insights based on the data they have); (6) pattern evaluation (to identify insights gained from the mining process); and (7) knowledge presentation (to describe and present knowledge acquired to users).

It is necessary to do a sampling of a dataset to avoid bias condition in the classification model. Besides, sampling will also affect the accuracy of the results obtained [7]. There are several sampling and validation methods that can be used, such as the holdout method and k-fold cross-validation. K-folds cross-validation utilizes the initial dataset to be divided as much as k value. The partition of input data will be used as a test set, while the other will be used as a training set. The difference between the holdout and random subsampling methods is that each sampling is used at the same time for training, but only once for testing. Figure 1 explains the k-fold cross-validation approach, which divided the dataset into several k iterations.

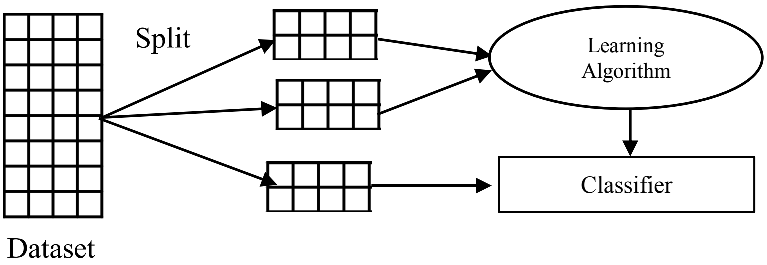

Figure 1. K-Fold Cross-Validation.

There is a case where the number of classes in a dataset is not balanced. That condition causing the resulting accuracy to be irrelevant. For example, in the case of credit card fraud, the class of fraud will undoubtedly be significantly less when compared to not fraud. Dutta, Dutta, and Raahemi use oversampling approach to solve that problem [9]. Oversampling techniques that commonly use is Synthetic Minority Oversampling Technique (SMOTE). SMOTE adds several tuples to the minority class with the k-nearest neighbour's approach.

Confusion Matrix and Receiver Operator Characteristics (ROC) used for model evaluation. The confusion matrix is a table that is often used to assess the performance of a classification model in test data. According to Kotu and Deshpande, to see the performance of a classification model, the confusion matrix is the best way [8]. We can suppose that the confusion matrix divided into two class labels Y or N. The Y value represents the right condition besides the N value represents the false condition. The accuracy of classification can we see into four possibilities as table 1. Several measurements can be done based on the results of the confusion matrix dimensions, i.e., sensitivity, specificity, precision, recall, and accuracy. According to Kotu and Deshpande, ROC Curves are arranged based on the values of True Positive (TP) and False Positive (FP) [8]. In a graph, we can describe the TP value in the vertical side while the FP value in the horizontal side.

Table 1. Confusion Matrix

Correct Classification	Classified as	
	+	-
+	True positives	False negatives
-	False positives	True negatives

3. Research Method

In this study, there are two methodologies used. The first methodology is designing the data warehouse; second, the methodology for data mining. The methodology in designing a data warehouse will use the Kimball method. Meanwhile, the methodology for data mining will use Knowledge Discovery Database (KDD). The entire steps of the study shown in table 2.

Table 2. Step of Reasearch Process

Step	Output
Problem identification	Research question
Literature study	Theoritical framework
Data collection	Business requirement
Data warehouse analysis and design	Dimensional modelling, ETL process, and visualization
Data Mining	Model of fraud document
Conclusion	Evaluation and suggestion

4. Results and Discussion

4.1. Business Requirement

In designing a data warehouse, it is necessary to find the information needed. The list of information needed is obtained based on the results of observations and interviews outlined in table 3.

Table 3. Requirement Lists

Req Code	Requirement
Req-01	Comparison of the use of Free Trade Agreement (FTA) Tariff and Most Favored Nation (MFN) Tariff based on the period and service office
Req-02	Import growth for goods that have quota based on period, types of goods, and service office
Req-03	Correction level, correction value, and correction appeal based on the importer, service office, country of origin, and period
Req-04	The level of compliance of notification of import documents with physical goods based on service office, importer, and period
Req-05	The average payment per teus based on importer, service office, and period
Req-06	The percentage of import document hit rates through the red line based on service office, types of goods, and period
Req-07	Percentage of the suitability of port of unloading with indentor's area of goods based on importer and period
Req-08	Percentage of countries of origin of goods based on the type of goods, importer, service office, and period
Req-09	Percentage of flags and modes of transportation based on importer, service office,

Req Code	Requirement
	and period
Req-10	The percentage of the carrier based on the importer, service office, and period
Req-11	Percentage of the seller based on importer, service office, and period
Req-12	Percentage of the indentor based on importer, service office, and period
Req-13	Comparison of invoices, transportation costs and insurance costs based on importer, service office, types of goods and period
Req-14	Percentage of completeness of description of goods based on importer, service office and period
Req-15	The price range of goods based on the HS code, type of goods, service office, type of services, and period
Req-16	Percentage of use of a broker for customs services based on the importer, service office, and period
Req-17	Number of packages and weight of goods based on the importer, type of packaging, type of goods, service office, and period
Req-18	The number of suspended based on the importer, suspend type, service office, and period
Req-19	Audit report based on importer and period

4.2. Dimensional Modelling

Overall information needed can be obtained based on existing data sources at the DJBC organization. Among these data sources come from the import for use, bonded zone imports, e-commerce, manifest, audit, registration, reference, and treasury. Based on the information needed, the activity of determining grain, identifying the dimension table, and identifying the fact table can be set out in table 4 and table 5. Dimensional modeling used in this study is constellation data model.

Table 4. Determine Grain

Grain	Req Code
Total of FTA Tariff and MFN Tariff based on period and office	Req-01
Total of good that have quota requirement based on period, type of goods, and office	Req-02
Total of correction document, amount, and correction appeal based on importer, office, country origin, and period	Req-03
Total of compliance based on office, type of goods, importer, and period	Req-04, Req-06
Average of payment per teus based on importer, office, and period	Req-05
Total of suitability between port and indentor’s area based on importer and period	Req-07
Total of customs declaration based on country of origin, flag, modes of transportation, carrier, seller, indentor, broker, type of goods, importer, office, and period	Req-08, Req-09, Req-10, Req-11, Req-12, Req-16
Agerage of invoices, transport costs, and insurance costs based on importer, office, type of goods, and period	Req-13
Total of brand column that occupied, type column that occupied, and length of description of goods based on importer, office, and time	Req-14
Maximum price, average price, and minimum price based on HS Code, type of	Req-15

Grain	Req Code
goods, office, type of services, and period	
Total of packaging and weight based on importer, type of packaging, type of goods, office, and period	Req-17
Total of suspend that occurred based on importer, suspend type, office, and period	Req-18
Total of Audit report based on importer and period	Req-19

Table 5. Identification Dimension Table and Fact Table

Dimension	Fact
Office	Tariff, Quota, Correction, Compliance of goods, Payment per teus, Customs declaration, Cost Insurance Freight (CIF), Description of goods, Price, Packaging, Suspends
Period	Tariff, Quota, Correction, Compliance of goods, Payment per teus, Suitability of port, Customs declaration, Cost Insurance Freight (CIF), Description of goods, Price, Packaging, Suspends, Audit
Type of Services	Price
Type of Goods	Quota, Customs declaration, Price, Packaging
Importer	Correction, Compliance of goods, Payment per teus, Suitability of port, Customs declaration, Cost Insurance Freight (CIF), Description of goods, Packaging, Suspends, Audit
Mode of Transportation	Customs declaration
Carrier	Customs declaration
Seller	Customs declaration, Correction
Indentor	Customs declaration
Broker	Customs declaration
Type of Packaging	Packaging
Suspend type	Suspends

4.3. Extraction, Transformation, Loading (ETL) Process

To ensure each stage of data integration based on existing data sources leading to a data warehouse is using the ETL process. ETL systems used in this research based on store procedure and jobs functions.

4.4. Visualization

Data warehouse results can utilize for the needs of end-user access tools. In this research, the development of information systems divided into three components, i.e., dashboard, OLAP, and the use of data mining. The examples of dashboard and OLAP results shown in Figure 2.

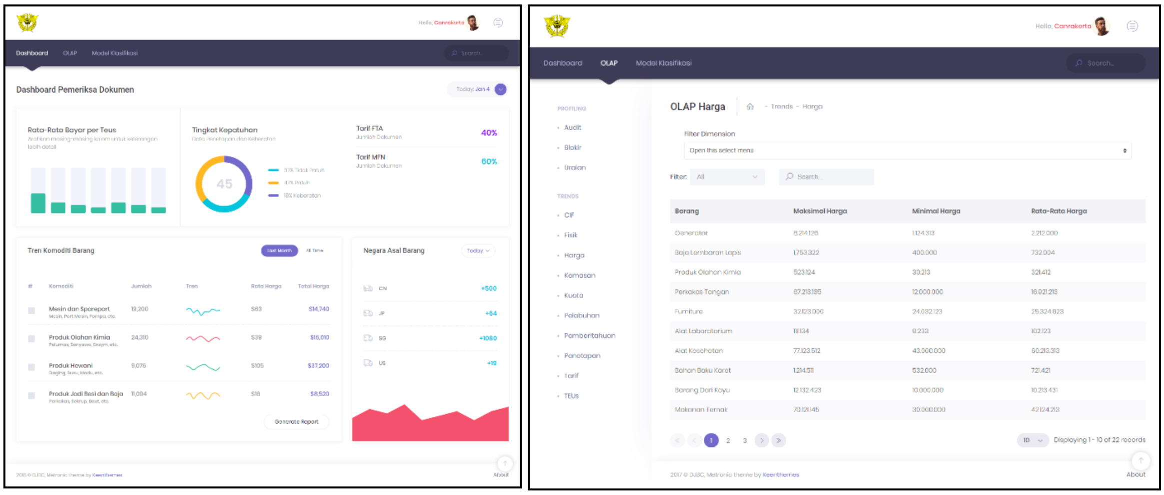

Figure 2. Dashboard and OLAP Visualization.

The list of OLAP requirements displayed based on several fact tables. It consists of tariff cube, quota cube, correction cube, suitability of port cube, compliance of goods cube, payment per teus cube, customs declaration cube, CIF cube, description of goods cube, price cube, packaging cube, suspends cube, and audit cube.

4.5. Data Mining Process

The use of data mining in this research conducted to find a fraud model from import documents. Data source characteristics and attributes used in the data mining process explained in table 6 and table 7.

Table 6. Data Source Characteristics

Attributes type	Nominal
Number of tuples	7204
Number of attributes	8 (the class attribute not included)
Missing value	-
Data Source	Data set for import document that declared by the importer in Tanjung Perak Service Office with yellow lines and red lines for January 2018 until March 2018

Table 7. List of Attribute

#	**Attributes**	**Attributes type**
1	FTA_TARIFF	Nominal
2	QUOTA	Nominal
3	COMPLIANCE_OF_GOODS	Nominal
4	DESCRIPTION_OF_GOODS	Nominal
5	RESTRICTED_GOODS	Nominal
6	COUNTRY_OF_ORIGIN	Nominal
7	SUITABILITY_OF_PORTS	Nominal
8	PAYMENT_PER_TEUS	Nominal

Based on the characteristics of the data, this research do the data transforms that described as follows:

- Conduct groupings of data values on the FTA_TARIF attribute, QUOTA attribute, COMPLIANCE_OF_GOODS attribute, and RESTRICTED_GOODS attribute. Grouping is done by changing the 0 value to "NO" and other than that to "YES"
- Conduct an oversampling approach with the SMOTE technique in the minority class. The SMOTE technique is carried out using the Rapidminer tool

Modeling is done using SVM, C4.5, NN, Random Forest, and Gradient Boosted Tree algorithms. The entire algorithm is carried out and is experimental by utilizing the method of sampling and validation with k-fold cross validation with a value of k = 10.

In this research, there are also class conditions that are not balanced. Comparison of the number of data for the class label "fraud" is 1490, while for the class label "not fraud" is 5714. The class label "not fraud" has a presentation of about 79% of the total amount of data, so the approach was taken using the SMOTE technique.

Table 8. Confusion Matrix for Non-SMOTE

Model	TP	FP	TN	FN
				Non-SMOTE
SVM	243	225	5489	1247
C4.5	195	139	5575	1295
Neural Network	1163	2472	3242	327
Random Forest	255	157	5557	1235
Gradient Boosted Trees	1137	2369	3345	353

Table 9. Confusion Matrix for SMOTE

Model	TP	FP	TN	FN
				SMOTE
SVM	4533	2584	3130	1181
C4.5	4931	2546	3168	783
Neural Network	4918	2696	3018	796
Random Forest	4751	2356	3358	963
Gradient Boosted Trees	5029	2742	2972	685

Table 10. Evaluation for Non-SMOTE

Model	Acc.	Pre.	Sen.	Spe.
				Non-SMOTE
SVM	79.57%	52.71%	16.31%	96.06%
C4.5	80.09%	57.52%	13.09%	**97.57%**
Neural Network	61.15%	32.05%	**78.05%**	56.74%
Random Forest	**80.68%**	**61.91%**	17.11%	97.25%
Gradient Boosted Trees	62.22%	32.44%	76.31%	58.54%

Table 11. Evaluation for SMOTE

Model	Acc.	Pre.	Sen.	Spe.
				SMOTE
SVM	67.05%	63.93%	79.33%	54.78%
C4.5	70.87%	65.96%	86.30%	55.44%
Neural Network	69.44%	64.63%	86.07%	52.82%
Random Forest	**70.96%**	**66.89%**	83.15%	**58.77%**
Gradient Boosted Trees	70.01%	64.73%	**88.01%**	52.01%

Table 12. AUC

Model	AUC	
	Non-SMOTE	*SMOTE*
SVM	0.645	0.710
C4.5	0.560	0.740
Neural Network	0.719	0.760
Random Forest	0.717	**0.776**
Gradient Boosted Trees	**0.724**	0.765

Evaluation of the data mining model described in tables 8, 9, 10, 11, and 12. Based on the evaluation, there are several things as follow:

- Non-SMOTE approach resulting sensitivity values almost into low level, while SMOTE approach makes sensitivity values for the whole classifier increased. It happens because the oversampling conducted can provide broader knowledge on the resulting model so the minority class or positive class, in this case, becomes better.
- The expected output from the classification techniques in this study is how the model or pattern for determining a fraud document can be identified. The sensitivity results using SMOTE approach are above 80%, except for SVM. The sensitivity results in unbalanced class cases are more appropriate than accurate results. It tells us the model that more suitable to use by document inspector is the model with SMOTE approach.

5. Conclusion

Schema drafting for a data warehouse based on information needed resulted in 12 dimension tables and 13 fact tables. The entire table is modelled using the fact schema approach. The ETL process itself is carried out on a staging database by creating scripts in-store procedures that can be run every day. Meanwhile, for visualization needs, end-user access tools were developed, including the establishment of OLAP, dashboards, and the use of data mining.

Utilization of the data mining process is carried out to create a model of import documents that can be used by document inspectors. In this research, the data source has an unbalanced class condition that requires an oversampling approach with the SMOTE technique. Besides, sensitivity values in the case of unbalanced class labels also need to be considered because they avoid overfitting that occurs based on the results of accuracy. Based on evaluation results, the Gradient Boosted Trees algorithm has the highest sensitivity value. This research should be optimized by trying to add the attributes, discretization of attributes, or other methods that can be used to improve accuracy.

Through the development of a data warehouse for import documents, there is another initiative that can be future research as an example is audit targetting using a classification approach. The audit target selection can help the organization determine a more optimal audit targetting with their limitation. The sample of limitation condition is the number of employees is not as much as the number of importers. It makes the customs clearance process can be carried out not only during clearance but also at post clearance.

References

[1] Ministry of Finance of the Republic of Indonesia, "Peraturan Menteri Keuangan Nomor 234 Tahun 2015 tentang Organisasi dan Tata Kerja Kementerian Keuangan," p. 1032, 2015
[2] T. Connolly and C. Begg, Database Systems: A Practical Approach to Design, Implementation, and Management, Global Edition, 6th Editio. Pearson, 2015
[3] R. Sharda, D. Delen, E. Turban, J. E. Aronson, and T.-P. Liang, Business Intelligence and Analytics: Systems for Decision Support. 2015
[4] I. Abramson, Data Warehouse : The Choice of Inmon versus Kimball. 2004
[5] W. H. Inmon, Building the Data Warehouse. Wiley, 2005
[6] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Editio. Wiley, 2013
[7] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012
[8] V. Kotu and B. Deshpande, Predictive Analytics and Data Mining. 2015
[9] I. Dutta, S. Dutta, and B. Raahemi, "Detecting financial restatements using data mining techniques," Expert Syst. Appl., vol. 90, pp. 374–393, 2017

Acknowledgement

This research was supported by Universitas Indonesia under the PIT 9 Grant 2019 of "Optimization of Digital Business Contribution trough Platform-based Project Management Model" (No: NKB0014/UN2.R3.1/HKP.05.00/2019). We would like to express our gratitude to the Faculty of Computer Science and the Directorate of Research and Community Engagement, Universitas Indonesia.