International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Utilization of Data Analytics in Research of Customs Notification

Dicki Kurniawan[1], Hustinawaty[2]
[1, 2]Business Information System, Gunadarma University, Depok West Java, Indonesia-16424
Email Address: [1]dicki7kurniawan @ gmail.com, [2]ninaprasetya @ gmail.com

Abstract— *One of the main functions of the Directorate General of Customs and Excise is revenue collector through import duties, export duties, import taxes and excise taxes. Importers submit import documents with a self-assessment system, namely calculating, reporting, and paying the Import Duty owed themselves. Customs officers carry out checks if it is found that different customs tariffs and / or values will be subject to a corrective note which will have an impact on optimizing state revenue. Data Analytics has a predictive function of providing what might happen in the future based on data patterns using techniques such as machine learning. This can help customs and excise officers to help analyze import documents. This research method uses CRISP-DM and the classification process is carried out by comparing the attributes that are at risk of imported documents subject to corrective notes. As well as comparing the Logistic Regression, Decision Tree and Naive Bayes algorithms to find the best model to use to help analyze imported documents.*

Keywords— *Data, Analytics, Machine Learning, Classification, CRISP-DM.*

I. INTRODUCTION

One of the main functions of the Directorate General of Customs and Excise is revenue collector through import duties, export duties, import taxes and excise taxes. the process of importing goods by business entities, institutions, or individuals, whether in the form of legal entities or non-legal entities carrying out activities to import goods from abroad into the country, are called importer and are required to submit import declaration. The notification uses a self-assessment system that views the taxpayer as a legally responsible entity. The self-assessment system is to calculate, report, and pay import duty by yourself. In its application and to fulfill customs obligations, the importer is obliged to declare and shall be responsible for the notification of the import of goods from abroad into the country in the form of the import declaration of goods. The implementation of the self-assessment system is closely related to the authority of Customs and Excise Officers to carry out the document inspection process in goods import activities. Customs and Excise Officers make decisions on the results of the examination of import documents. Decisions include the calculation of import levies, stipulation and calculation of administrative sanctions. In the event that the inspection of documents on import of goods carried out by Customs and Excise Officers is found that the tariff and/or customs value is different from that notified by the importer which results in underpayment of import duty and tax, the officer will provide a corrective note to the import document. Based on this, the researcher wants to use data analytics to find out the attributes that affect the import documents that are hit by the corrective note.

Data analytics is a combination of skills including statistics, computer science, and mathematics. On-Line Analytical Processing (OLAP) has many variants that have the same goals but have slightly different emphasis on functions such as decision support systems (DSS), machine learning, data mining, data warehouse, or business intelligence. The field of data analysis practice is then given the term data analysis [6]. The difference between analysis and analytics is that data analysis understands information that happened in the past and what is happening, while data analytics understands why this happened and what is likely to happen in the future [4].

Machine Learning is an artificial intelligence that uses statistical techniques to generate an automatic model from a data set. Machine learning has the goal of giving computers the ability to learn and enabling computers to learn from data so that they can produce a model for carrying out the input-output process without using explicitly created program code. The learning process uses special algorithms called machine learning algorithms. one of the machine learning techniques is classification. The purpose of the classification process is similar to clustering by dividing a dataset into groups. in the classification can provide input to the machine group division algorithm or teach the machine how to divide the group. Whereas in clustering, we do not teach machines, but it will do the grouping by itself.

This study uses a classification method by comparing the logistic regression, decision tree, and naive bayes algorithm models. The best algorithm model will be used as a risk engine which will be integrated into the import application system of the Directorate General of Customs and Excise. From the description above, the formulation of the problem in this study is how to use data analytics to help customs and excise officers in analyzing import notification documents.

II. LITERATURE REVIEW

A. Data Analytics

Data Analytics is a combination of skills including statistics, computer science, and mathematics [3]. The Gatner Analytic Ascendancy Model (Gartner, March 2012) divides data analytics into four types based on the level of value and difficulty: (1). Descriptive Analytics, to get an overview of the data that has been collected. (2). Diagnostic Analytics, is carried out with the aim of finding the cause of the appearance of the data. (3). Predictive analytics, which aims to provide

Dicki Kurniawan and Hustinawaty, "Utilization of Data Analytics in Research of Customs Notification," *International Research Journal of Advanced Engineering and Science*, Volume 6, Issue 2, pp. 55-59, 2021.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

predictive results about something that will come. (4). Prescriptive Analytics, describes the actions that must be taken after knowing the problem or risk.

B. *Predictive Data Analytics*

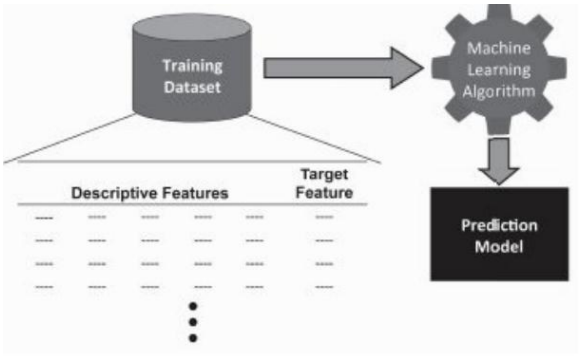
Data must be processed and analyzed to gain insight so that it can be used to make decisions. turning data into insights so that it is used for decision making is the task of data analytics. The progression from data to insight to decision is illustrated in Figure 1

Fig. 1. Data Analytics

Predictive data analytics aims to provide predictive results about something to come. The prediction in question is probability in nature and assigns a value to the unknown variable. Algorithm models are trained to make predictions based on historical data. To train the model can use machine learning [6].

C. *Machine Learning*

Machine Learning is an artificial intelligence that uses statistical techniques to generate an automatic model from a data set. Machine learning technique is to automatically learn the relationship between descriptive features and target features to create prediction models based on historical data. then use the model to make predictions. These two separate steps are shown in Figure 2 [6]:



(a) Learning a model from a set of historical instances

(b) Using a model to make predictions

Fig. 2. Machine Learning

D. *CRIPS-DM*

The following are the stages of the CRISP-DM method [7]:

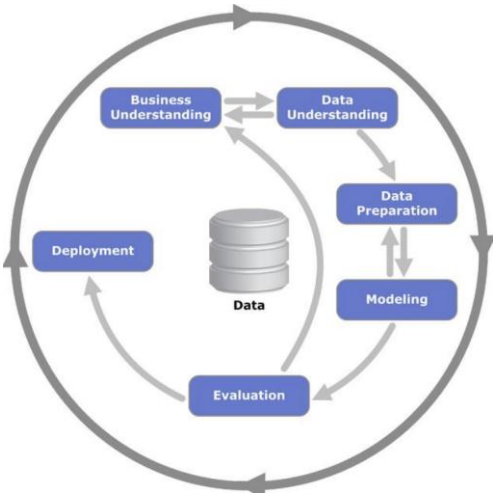


Fig. 3. Method CRISP-DM

1. *Business Understanding*
   The first phase is needed to determine business goals, problems, success criteria, assess situation and make a project plan.
2. *Data Understanding*
   Data understanding begins with collecting initial data, describe data and verify data quality. in this phase it is possible to return to the first phase because when understanding data it turns out that business goals are not possible from the available data so that it changes the goals of the project.
3. *Data Preparation*
   In this phase, prepare the final dataset from one source that will be used for modeling. This phase involves the process of cleansing data, integrate data and making changes to the format if necessary.
4. *Modeling*
   The modeling phase includes selecting the algorithm model, building the model and assessing the model.
5. *Evalution*
   At this stage, evaluate the model that has been made, whether it is in accordance with the original purpose.
6. *Deployment*
   in the last phase, it combines the models that have been made in the operational system and monitors the models that have been made in a certain period.

E. *Logistic Regression*

Logistic regression is a classification of outcomes for the relationship between (discrete/continuous) input features and the probability of a particular discrete output outcome. Logistic regression has binary regression for the dependent variable with a value between 1 and 0.

F. *Decision Tree*

Decision tree is a prediction model that uses a hierarchical structure or what is commonly called a tree structure. The process of using a decision tree to make predictions begins by testing the descriptive features at the root node of the tree. The results of this test determine the derivative of the root node

Dicki Kurniawan and Hustinawaty, "Utilization of Data Analytics in Research of Customs Notification," *International Research Journal of Advanced Engineering and Science*, Volume 6, Issue 2, pp. 55-59, 2021.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

which then needs to be downgraded the process. In principle, this test tests the value of the descriptive feature and decreases the level, then it is repeated until the process reaches the node [6].

G. Naïve Bayes

The Naive Bayes algorithm can make predictions or probabilities on the future based on past experiences. Naïve bayes model takes conditional independence to an extreme by assuming conditional independence between assigning all descriptive feature values given at the target level [9].

H. Confusion matrix

The confusion matrix provides classification details, at the top of the matrix are for prediction classes while on the left side of the matrix are the observed classes. TP and TN explained when the classifier was correct, while FP and FN explained when the classifier was wrong [9].

		Predicted Class	
		Normal	Attack
Actual Class	Normal	True Negative (TN)	False Positive (FP)
	Attack	False Negative (FN)	True Positive (TP)

Fig. 4. Confusion Matrix

Based on the confusion matrix table, the accuracy value can be calculated. Accuracy is describing how accurate the model is in classifying correctly: Accuracy formula = (TP + TN) / (TP + FP + FN + TN) [9].

III. Research Method

Research requires a research methodology so that work steps become more systematic. The steps taken by researchers in the use of data analytics in research of customs notification:

A. Methodology

The methodology we use in this study uses CRIPS-DM because in the 2014 KD Nuggets survey, the CRISP-DM methodology remains the top data analytics methodology. presented in the form of the following Figure:

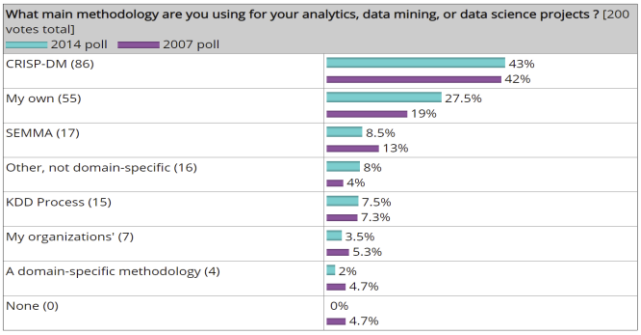


Fig. 5. Survey KD Nuggets

B. Data Collection Techniques

The data used in this study is to use data examples of import notification of goods at the Directorate General of Customs and Excise

C. Tools

Orange is an open source software for processing Data Analytics such as Rapidminer, WEKA, and others. Orange is better when it comes to visualization [8].

IV. Results and Discussion

A. Business Understanding

1. There is a potential for a decrease in state revenue due to errors in filling out the notification of imported goods due to the self-assessment process of filling in documents
2. Data mining modeling needs to be made to help customs officers to gain insight in making decisions to issue corrective notes
3. Customs Officials need to know the factors or attributes that are at risk of being subject to a corrective note.

B. Data Understanding

At this stage data collection and identification of each attribute:

TABLE 1. Dataset

No	Name	Data Type
1.	Company Status	Category
2.	Profile	Category
3.	Path	Category
4.	Facility	Category
5.	Contry of Origin	Category
6.	Correction Note	Category

In accordance with table 1. above:

1. Profile attribute shows the profile of the company's risk, the content of these attributes is Very high, High, Medium, and Low.
2. Path attribute is the result of the tracking profiling risk engine.
3. Facility Attribute consists of Free Trade Agreement (FTA) and Non-FTA.
4. Company status attribute shows the status of the company.
5. Correction note is a flag indicating that the document is subject to a correction note.
6. The number of records is 599.

C. Data Preparation

In the dataset, there are 6 records of missing data in the profile attribute. Orange tools have a preprocess feature to perform the cleansing process.

Dicki Kurniawan and Hustinawaty, "Utilization of Data Analytics in Research of Customs Notification," *International Research Journal of Advanced Engineering and Science*, Volume 6, Issue 2, pp. 55-59, 2021.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

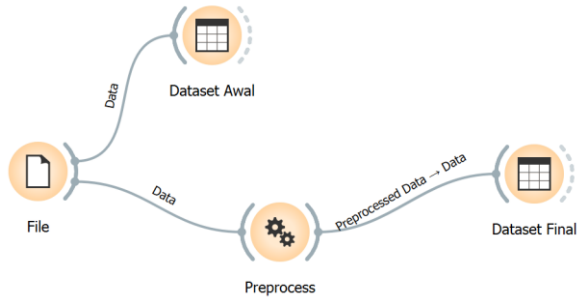


Fig. 6. Data Preparation

After the cleansing process and no missing data, the dataset can be used for modeling.

D. Modeling

As an algorithm model selection phase, this study compares three classification algorithms, namely Decision Tree, Naive Bayes, and Logistic Regression to compare the three algorithms.

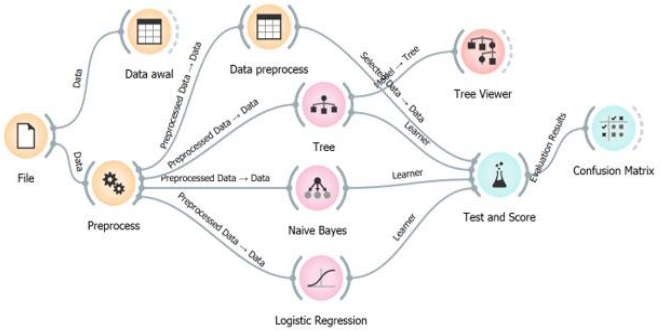


Fig. 7. Modelling

E. Evaluation

a. *Confusion matrix*

The Confusion matrix of each model can be calculated its Accuracy value, which describes how strong the model is in classifying correctly, with the following formula:

Accuracy = (TP+TN) / (TP+FP+FN+TN)

TABLE 2. Comparison of results

No.	Model	Result
1.	*Logistic Regression*	85.1%
2.	*Decision Tree*	85.1%
3.	*Naïve Bayes*	84.4%

In accordance with table 2 above comparison of results the results of the calculation of the accuracy value from the configuration matrix and it is known that the logistic regression and decision tree models have high results and can be used as a recommendation model for corrective notes in the import application system.

b. *Model Trials*

The logistic regression model was tested using testing data:

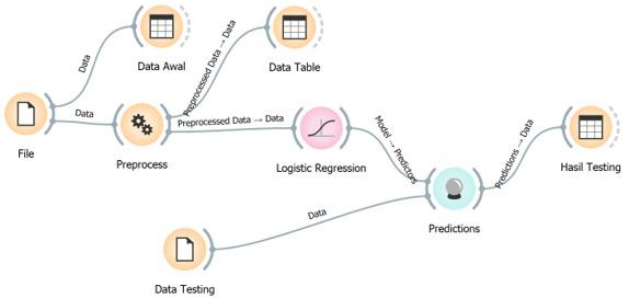


Fig. 8. Test Model

	Logistic Regression	gistic Regression (	gistic Regression (	Status Perusahaan	Profil	Penjaluran	Kode fasilitas	Negara Asal	Notul
1	1	0.116779	0.883221	IP - Importir ...	Very High	Merah	Non FTA	China	1
2	0	0.994759	0.00524145	Lainnya	Medium	Hijau	Non FTA	Singapore	0
3	0	0.974316	0.0256844	Lainnya	High	Hijau	FTA	China	0
4	0	0.998264	0.00173603	Lainnya	Medium	Hijau	FTA	Hong Kong	0
5	1	0.116779	0.883221	IP - Importir ...	Very High	Merah	Non FTA	China	1
6	0	0.993662	0.00633839	Lainnya	Medium	Hijau	Non FTA	Belgium	0
7	0	0.872144	0.127856	Lainnya	Very High	Hijau	FTA	China	0
8	0	0.527232	0.472768	IP - Importir ...	Very High	Merah	Non FTA	United Arab E...	0
9	1	0.277466	0.722534	IP - Importir ...	Very High	Merah	FTA	China	1
10	0	0.821433	0.178567	IP - Importir ...	Very High	Hijau	FTA	China	1
11	0	0.99748	0.00252015	Lainnya	Low	Hijau	FTA	Thailand	0
12	0	0.996706	0.00329355	MIIA	Low	Hijau	Non FTA	Singapore	0
13	0	0.872144	0.127856	Lainnya	Very High	Hijau	FTA	China	0
14	1	0.277466	0.722534	IP - Importir ...	Very High	Merah	FTA	China	1
15	1	0.116779	0.883221	IP - Importir ...	Very High	Merah	Non FTA	China	1
16	0	0.99363	0.0063697	Lainnya	Medium	Hijau	Non FTA	Japan	0
17	0	0.821433	0.178567	IP - Importir ...	Very High	Hijau	FTA	China	0
18	1	0.277466	0.722534	IP - Importir ...	Very High	Merah	FTA	China	1
19	0	0.635427	0.364573	Lainnya	Very High	Kuning	FTA	China	0
20	0	0.940654	0.0593455	Lainnya	Very High	Hijau	Non FTA	Hong Kong	0

Fig. 9. Data Testing Results

Based on the results of the testing data, it is known that there is only one document that should have been subject to a correction note but not a correction note.

F. Deployment

In this last stage, the implementation of the data analytics model is carried out into the import service application system so that the customs and excise officials automatically receive a notification in the system that the import notification document of the goods has the potential to be subject to a corrective note. After the model is implemented, a monitoring process is carried out on the recommendation output whether it is in accordance with the model that has been made. It is necessary to make the latest model at a certain period using the latest testing data in accordance with the CRISP-DM cycle which is always updated continuously.

V. Conclusions and Suggestions

Based on the research results the use of data analytics in research of customs notification:

a. The algorithm with the best accuracy is logistic regression with AUC 0.914.
b. The CRISP-DM method in classification can be used to obtain the best algorithm for predicting 95% correction notes.
c. In order to always update the tracking risk engine and company profile because these attributes are very influential due to the correction note.
d. Remodeling needs to be done by adding attributes that have a risk of being subject to a corrective note.

Dicki Kurniawan and Hustinawaty, "Utilization of Data Analytics in Research of Customs Notification," *International Research Journal of Advanced Engineering and Science*, Volume 6, Issue 2, pp. 55-59, 2021.

REFERENCES

[1] Myers, Ronin. 2019. *Data Management and Statistical Analysis Techniques.*
[2] Anggarwal, Charu C. 2015. *Data Mining the Textbook*. Springer Cham Heidelberg. New York.
[3] Myers, Ronin. 2019. Data Management and Statistical Analysis Techniques. ISBN 9781839473395.
[4] Park,David.2017.*https://www.eetimes.com/analysis-vs-analytics-past-vs-future.*
[5] Darono, Agung. 2020. *Data Analytics Dalam Administrasi Pajak Di Indonesia*. Vol.6, No.2.2020.
[6] Jhon D. Kelleher, Brian Mac Name, Aoifee D' Arcy. 2015. *Fundamentals of machine learning for predictive data analytics*. Massachusetts Institute of Technology.
[7] Chapman, Pete, et. al 2000. CRIPS-DM v.10 Step by Step Data Mining Guide. SPSS Inc.
[8] EmsarJanez.dkk.2013. Orange: Data Mining Toolbox in Python. The Journal of Machine Learning Research. 2349-2353.
[9] Han, J, Kamber, M, & Pei, J. 2012. Data Mining: Concept and Techniques, Third Edition. Waltham: Morgan Kaufmann Publishers