NAIVE BAYES ALGORITHM IN HS CODE CLASSIFICATION FOR OPTIMIZING CUSTOMS REVENUE AND MITIGATION OF POTENTIAL RESTITUTION

Hafizh Adam Muslim

Directorate General of Customs and Excise, Ministry of Finance, Indonesia
hafizh.adam@kemenkeu.go.id

The Directorate General of Customs and Excise, as a government revenue collector, must maximise import duty receipts each year. One common issue is the return of unpaid import duty and/or administrative punishments in the form of fines based on the objection judgement document. The Tax Court could help you minimise your gross receipts at the Customs Office. Data mining techniques are intended to provide valuable information regarding the HS Code classification technique, which can assist customs agents in determining duties and/or customs values. This study makes use of data from the Notification of Import of Goods at Customs Regional Office XYZ from 2018 to 2020. The Cross-industry Standard Process for Data Mining (CRISP-DM) model is used in this study, and the Naive Bayes Algorithm in Rapidminer 9.10 is used for data classification. According to the model, the calculation accuracy is 99.97 percent, the classification error value is 0.03 percent, and the Kappa coefficient is 0.999.

***Keywords*: Customs, HS Code, Data Mining, Naive Bayes, Rapidminer.**

I. INTRODUCTION

In the Indonesia's State Budget, tax revenues are classified into two sectors, namely domestic tax revenues and international trade tax revenues. One of the tax revenue posts, especially those from the international trade tax sector, is import duties. As a governmental revenue collector, the Directorate General of Customs and Excise must maximise import duty receipts each year.

Importation is done by submitting Import Notification Document (PIB) to the customs office. PIB is a notification document to customs by the importer on imported products through an electronic network, based on complementary customs documents and the self-assessment principle, which compels taxpayers to calculate, pay, and report taxes in compliance with the terms of the law. Rates and customs values are two of the most significant features of PIB. These two factors affect the amount of customs duty in goods import activities. The customs value is determined by inspecting the number, kind, and value of the items. Inspection of various types of commodities, in addition to identifying the customs value, can also determine the amount of import duty rates. Different types of items may be classified differently under the HS Code. Because import duty rates are decided by the HS Code, this discrepancy in HS Code allows for changes in import duty rates.[1]

The HS code is a 6-digit international numerical number used to designate and identify goods for international trade. In addition to the internationally recognised 6-digit number, each country can add more digits to the code to make it 8, 10, or 12-digit for tariff and statistical purposes. The process of finding the most exact description in the harmonised system (HS) for the commodities to be classed is known as HS Classification.[2] Correct classification is necessary by the government for three main reasons: 1. Duty, tax, and charge calculation 2. Determination of requisite permissions, licences, and certificates 3. Gathering of trade statistics Correct classification can help businesses speed up the customs clearing process by avoiding unnecessary noncompliance, which can result in shipping delays, extra inspections, fines, and other administrative penalties.[2]

If the results of the import document examination result in a failure to pay import duties, the Customs Official shall issue a Letter of Determination of Customs Value Tarifffs (SPTNP). The importer may register an objection with the Director General of Customs and Excise to the determination made by the customs and excise officer, one of which is about tarifffs and customs value for computing import duty, resulting in underpayment. If the objection is denied, the applicant has 60 days from the decision date to seek an appeal with the Tax Court. The Tax Court might still determine whether or not to allow or reject the objection. In the event that the objection decision is upheld and the Tax Court's decision is granted, the applicant may apply for a refund of the excess payment of import duty and/or administrative sanctions in the form of a fine, in the event that the applicant has paid off the invoice.

The issue that frequently arises is the return of overpaid import duty and/or administrative sanctions in the form of fines based on the objection decision document. The Tax Court may be a way to reduce gross receipts at the Customs Official. The use of data mining techniques is expected to provide useful information about the HS Code classification technique, which can assist Customs Officials in setting tarifffs and/or customs value. The proposed work is now focused on the unsupervised learning technique, specifically Naive Bayes classification, for categorising and predicting labeled data. The Naive Bayes approach has proven to be a tractable and efficient method for classification in multivariate analysis. [3] The purpose of using the model is to predict the HS Code according to the dataset and variable that has been defined by using the Naive Bayes algorithm It is expected that by mitigating these risks, it will optimize state revenues through the determination of appropriate tariffs and/or customs values on imported goods.

II. METHODE

Research plan is essential in order for the research to perform efficiently, methodically, and efficiently. The Cross-industry Standard Process for Data Mining (CRISP-DM) technique is used in this study. The CRISP-DM is a process model that serves as the foundation for a data science process. It is divided into six stages: business understanding; data understanding; data preparation; modeling; evaluation; and deployment. Figure 1 depicts the six stages of this research.

TABLE 1
Example Records of Dataset

Business understanding	The business condition should be examined in order to have an understanding of the available and required resources. One of the most crucial components of this phase is determining the data mining goal. First, the data mining kind (e.g., classification) and data mining success criteria should be explained (like precision). A needed project plan should be developed.
Data understanding	This step requires collecting data from data sources, exploring and summarising it, and ensuring data quality. To clarify, the user guide explains the data description task as involving statistical analysis and defining qualities and their collations.
Data preparation	Data selection should be done by setting inclusion and exclusion criteria. Cleaning data can help to improve data quality. Derived attributes must be produced based on the used model (specified in the first phase). Different approaches are available and model dependent for all of these processes.
Modeling	The data modelling process includes choosing a modelling technique, creating a test case, and modelling the model. All data mining techniques are applicable. In general, the decision is influenced by the business problem and the facts. What is more essential is how you explain your decision. For the model to be built, specific parameters must be set. It is appropriate to examine the model against evaluation criteria and select the best ones for assessment.
Evaluation	The outcomes are evaluated against the established business objectives during the evaluation phase. As a result, the results must be analysed and subsequent actions must be outlined. Another issue to consider is that the process as a whole should be examined.
Deployment	The user guide provides a general description of the deployment phase. It could be a final report or a piece of software. According to the user handbook, the deployment phase comprises of planning, monitoring, and maintenance.

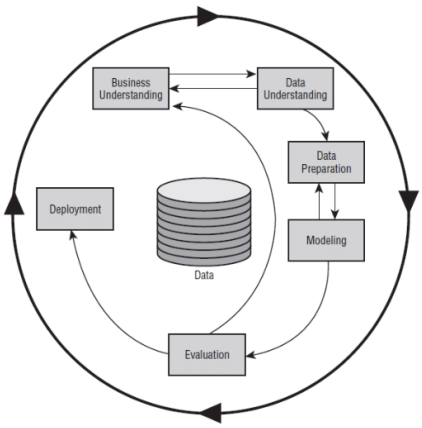

Fig. 1. CRISP-DM Diagrams

It was first published in 1999 to standardise data mining methods across sectors, and it has since become the most widely used approach for data mining, analytics, and data science initiatives. The advantages of adopting the CRISP-DM include reduced time and costs, as well as reduced expertise needs for data mining initiatives. Furthermore, utilizing CRISP-DM speeds up training, knowledge transfer, documentation, and the capture of best practises. [4] The following table 1 will explain the meaning of each stage in the CRISP-DM method. [4][5]

III. RESULT AND DISCUSSION

A. Business Understanding

The first stage in determining the problems that will be solved in order to attain the desired results. This research is limited to the Regional Customs Office XYZ for import activities in the 2018-2020 period. According to the data that has been obtained, it is known that there were objections filed with a total value of 294 trillion rupiah, of which 47 trillion rupiah with the status of the objection being accepted. This causes the loss of potential state revenue, which has a significant impact. It is hoped that by reflecting on historical data, a good decision will be made in the future. This research began with the need for a computer-based system for determining the initial risk of misclassification of goods on an import customs notification and the search for an initial risk determination method that can assist customs officials in determining the proper classification. The goal of data analysis is to discover relationships between attributes that influence the classification of goods and to make classification predictions that are accurate with the class grouping.

B. Data Understanding

The second stage is data preprocessing, this stage entails ensuring that the data to be processed is normal, complete, and consistent before it is included in the model. Classification means the process of grouping. In terms of data science, classification is then understood as the process of grouping data into several categories to make it easier to process and analyze. The purpose of this research is to categorise the HS Code based on historical data for submitting PIB and data on the determination of objections to a Customs Regional Office XYZ for the 2018–2020 timeframe. It is anticipated that by limiting these risks, it will maximise state income by determining suitable Tarifffs and/or customs values on imported items.

C. Data Preparation

The third stage is data preparation, with the selected data being attributes that are thought to be relevant in assessing Tarifffs and/or customs values up to the Tax Court's conclusion. This step encompasses all activities required to generate the final dataset. Existing data is frequently unstructured, and there is missing data. As a result, cleaning is required to increase data quality. This stage can be performed again and in various order. The data required includes the following components in Fig 2. The primary purpose of this phase is to increase the quality of the data provided so that we can ensure better model outputs in subsequent rounds. Data processing is one of the most time-consuming steps of the process model's execution. Because real data might be incomplete, noisy, and inconsistent, data preparation takes time. Before feeding the provided dataset to a model, there are numerous preprocessing processes. The management of null/missing values is one of the most critical. There are numerous approaches for dealing with null or missing values automatically.

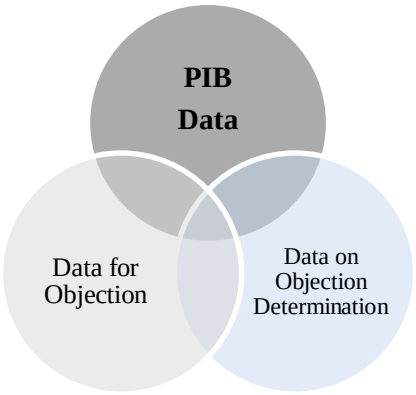

Fig. 2. Data Contain

Following the processing of data from the three raw data sources, a slice of data containing commodities in HS Code, kind of objection, resume, and decision is generated, as indicated in the table below:

TABLE 2
Example Records of Dataset

Commodity in HS Code	Type of Objection				Decision	Class
	Quantity	Type	Value	Tariff		
69072191 ceramic prouducts	x	x	√	x	completely refuse	Green
70139900 goods made of glass	x	x	x	√	accept all	Red
69072191 ceramic prouducts	x	x	√	x	accept all	Red
39261000 goods made of plastic	x	x	√	x	accept all	Red
87116011 motorcycle	√	√	√	√	set under-payment	Red
07031019 vegetables	x	x	x	√	accept all	Red
85022010 generator	x	√	√	x	accept all	Red
08081000 fruits	√	x	√	√	accept all	Red
87081090 car parts and accessories	x	x	√	x	partially reject	Yellow
85185090 sound system	x	x	√	√	partially reject	Yellow
87141090 bicycle parts and accessories	x	√	√	√	partially reject	Yellow
84314990 heavy equipment parts	x	x	√	x	set to add pay	Green
48201000 paper	√	x	√	√	completely refuse	Green
Etc.						

TABLE 3
Quantity of Dataset based of Class

Green	3480
Yellow	44
Red	559
Total	4083

TABLE 4
Quantity of Dataset based of Type of Objection

Type of Objection	Amount of data
Value	2.288
Tariff	1.459
Value, Tariff	231
Quantity, Value, Tariff	12
Quantity. Type, Value, Tariff	25
Quantity, Value	44
Type, Value	4
Type, Tariff	4
Type, Value, Tariff	3
Null	13
Total Data	4.083

TABLE 5
Quantity of Dataset based of HS Code

HS Code	Amount of data
08083000	268
07032090	138
08081000	79
11010019	79
07031019	74
08061000	64
64029990	51
28332100	49
10019999	46
08109010	41
62143090	32
there are 1270 HS Code Data contained in this study	

D. Modeling

Modeling is the fourth stage, which is the process of applying various data mining tools and determining ideal parameters or qualities. This stage also covers the analysis and assessment of the model set. The algorithm to be utilised is selected at the data mining approach selection phase.

The Naive Bayes algorithm is based on Bayes' Theorem. The theory states that if new facts occur, assumptions must change subjectively. When Naive Bayes is employed as the algorithm's foundation, it assumes that the presence of some variables in a class is unrelated to the presence of other variables.[6] The

Bayes theorem equation is shown in:

$$P(H|U) = \frac{P(U|H).P(H)}{P(U)}$$

where:

U	:	Data with unknown classes
H	:	The U data hypothesis is a specific class
P(H|U)	:	The probability of hypothesis H is based on condition U
P(H)	:	Hypothesis probability H
P(U|H)	:	U probability based on conditions
P(U)	:	U Probability

Fig 3. Bayes' Theorem

The Naive Bayes Classifier method will be divided into three classifications, NAMELY the Red, Yellow, and Green Categories, for the purpose of determining the HS Code, which will be chosen by the Customs officer. The red category is for the HS Code submitted with the decision to accept all and set to underpayment. The green category is for HS Codes submitted with a decision to complete refuse, set to add pay, and withdraw the application. The yellow category is for the HS Code submitted with a partially rejected decision. The following will explain further the calculation concept of the Naive Bayes algorithm. As an example of a case, how is the potential for determination of the HS Code 69072191 for Ceramic Product commodities with tariff errors.

the probability of the colors Green, Yellow, and Red:

P(Green) = 3480/4083 = 0.852
P(Yellow) = 44/4083 = 0.011
P(Red) = 559/4083 = 0.137

green chance of HS Code and Tariff error:

P(69072191 | Green) = 25/3480 = 0.007
P(Tariff | Green) = 1272/3480 = 0.366

yellow chance of HS Code and Tariff error:

P(69072191 | Yellow) = 0/44 = 0.000
P(Tariff | Yellow) = 8/44 = 0.182

red chance of HS Code and Tariff error:

P(69072191 | Red) = 5/559 = 0.009
P(Tariff | Red) = 179/559 = 0.320

Then determine what percentage of the HS Code is in green, yellow, and red.:

Green = P(G) x P(69072191|G) x P(Tariff |G)
= 0.852 x 0.007 x 0.366
= 0.218 %

Yellow = P(Y) x P(69072191|Y) x P(Tariff |Y)
= 0.011 x 0.0000 x 0.182
= 0.000 %

Red = P(R) x P(69072191|R) x P(Tariff |R)
= 0.137 x 0.009 x 0.320
= 0.039 %

So, based on the results of manual calculations using the Naive Bayes Algorithm, it was found that the highest percentage result was 0.218%, so it can be concluded that for HS Code 69072191, the tariff error is classified as Green, which means that the determination of the error on the HS Code has no potential for restitution.

Rapid Miner Studio 9.10 was utilised. The following datasets and models will be used in this study:

TABLE 6
Type of Dataset

Dataset	Type
PIB (id)	Integer
HS Code	Interger
Quantity	Binominal
Type	Binominal
Value	Binominal
Tariff	Binominal
Decision	Polynominal
Class (label)	Polynominal

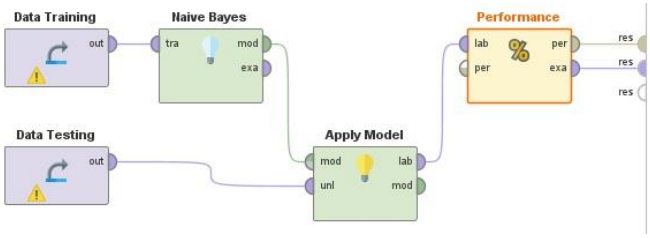

Fig. 4. Naive Bayes Modelling in Rapidminer 9.10

Naive Bayes is a low-variance, high-bias classifier that can develop a successful model even with a little data set. It is easy to use and computationally cheap. Text categorization, including spam detection,

sentiment analysis, and recommender systems, is a common use case. The key premise of Naive Bayes is that the value of any attribute is independent of the value of any other attribute given the value of the label (the class). The condition of independence greatly simplifies the computations required to construct the Naive Bayes probability model. Initially, this research will utilise a simple Naive Bayes model in general, however because the data lacks testing data, the data is split 80 %: 20 %. The Split Data operator accepts an ExampleSet as input and returns subsets of that ExampleSet via its output ports. The partitions option specifies the number of subsets (or partitions) and the relative size of each partition. The sum of the ratio of all partitions should be 1. The sampling type parameter decides how the examples should be shuffled in the resultant partitions. The Split Data operator can use several types of sampling for building the subsets. For the sampling type, automatic is selected, which means uses stratified sampling if the label is nominal, shuffled sampling otherwise.

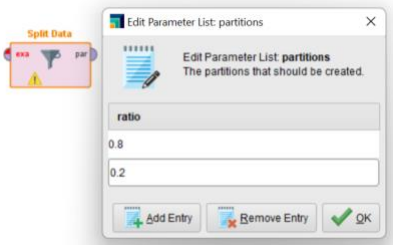

Fig. 5. Split Data Parameter

Cross Validation is used to see the level of accuracy of the Naive Bayes model used, the type of accuracy used is accuracy and the kappa coefficient. The Cross Validation Operator is a nested Operator. It has two subprocesses: a Training subprocess and a Testing subprocess. The Training subprocess is used for training a model. The trained model is then applied in the Testing subprocess. The performance of the model is measured during the Testing phase. This study uses the 10-Fold Cross Validation method. Number of folds specifies the number of folds (number of subsets) the ExampleSet should be divided into. Each subset has equal number of Examples. Also the number of iterations that will take place is the same as the number of folds. If the model output port is connected, the Training subprocess is repeated one more time with all Examples to build the final model. The automated mode uses stratified sampling per default. If it isn't applicable e.g. if the ExampleSet doesn't contain a nominal label, shuffled sampling will be used instead.

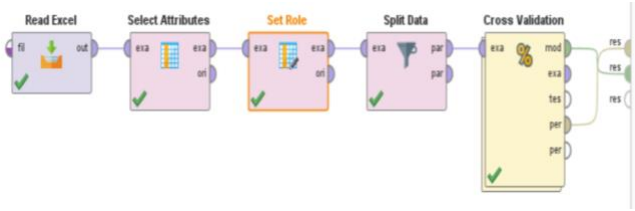

Fig. 6. Cross Validation Modelling in Rapidminer 9.10

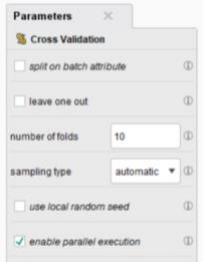

Fig.7. Cross Validation Parameter

The details of the model used in cross validation are as follows:

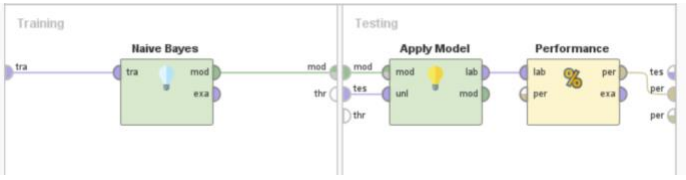

Fig. 8. Naive Bayes Modelling in Rapidminer 9.10

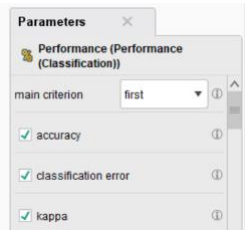

Fig. 9. Performance Parameter

Performance is used for statistical performance evaluation of classification tasks. This operator delivers a list of performance criteria values of the classification task. Accuracy is relative number of correctly classified examples or in other words percentage of correct predictions;classification_error

is relative number of misclassified examples or in other words percentage of incorrect predictions; and The kappa statistics for the classification. It is generally thought to be a more robust measure than simple percentage correct prediction calculation since it takes into account the correct prediction occurring by chance.

E. Evaluation

The fifth stage is evaluation. This phase involves a more complete evaluation of the set of models that have been implemented, as well as a review of the steps that have been taken.

TABLE 7
Accuracy Performance

accuracy: 99.97% +/- 0.10% (micro average: 99.97%)

	true GREEN	true RED	true YELLOW	class precision
pred. GREEN	2784	0	0	100.00%
pred. RED	0	446	1	99.78%
pred. YELLOW	0	0	34	100.00%
class recall	100.00%	100.00%	97.14%	

TABLE 8
Classification Error

classification_error: 0.03% +/- 0.10% (micro average: 0.03%)

	true GREEN	true RED	true YELLOW	class precision
pred. GREEN	2784	0	0	100.00%
pred. RED	0	446	1	99.78%
pred. YELLOW	0	0	34	100.00%
class recall	100.00%	100.00%	97.14%	

TABLE 9
Kappa Coefecient

kappa: 0.999 +/- 0.004 (micro average: 0.999)

	true GREEN	true RED	true YELLOW	class precision
pred. GREEN	2784	0	0	100.00%
pred. RED	0	446	1	99.78%
pred. YELLOW	0	0	34	100.00%
class recall	100.00%	100.00%	97.14%	

Based on the table, the confusion matrix is obtained in the Naive Bayes model and 10 fold Cross Validation. The table shows the following results:

- The number of Green data and detected true or true positive is 2784
- The number of data Red and detected is true or true positive is 446
- The number of Yellow data but detected is wrong or false negative is 34
- The number of Yellow data and detected true or true positive is 1

The calculation accuracy resulting from the modelling is 99.97%, the classification error value is 0.03%, and the Kappa coefficient is 0.999, indicating that the data mining technique used is very good.

F. Deployment

Deployment is not the end of data mining activities. Further research is also needed to develop this data mining model to be better. This research has resulted in a new pattern of knowledge from data classification using the Naive Bayes model of PIB, which aims to determine the category of HS Code that will be selected for determination, thus helping Customs Officers in making decisions. To support the need for effective and efficient data, a simple application system based on the Naive Bayes algorithm using the PHP programming language has also been created with the data processing flow as shown in the following flowchart:

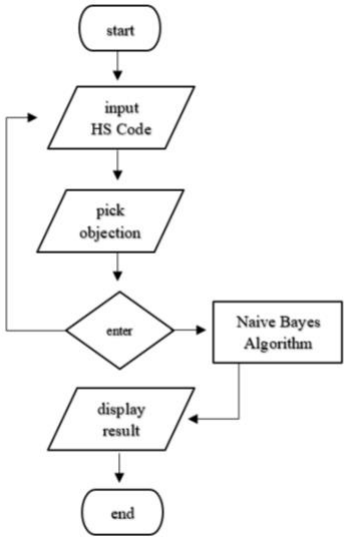

Fig.10. Flowchart of Application

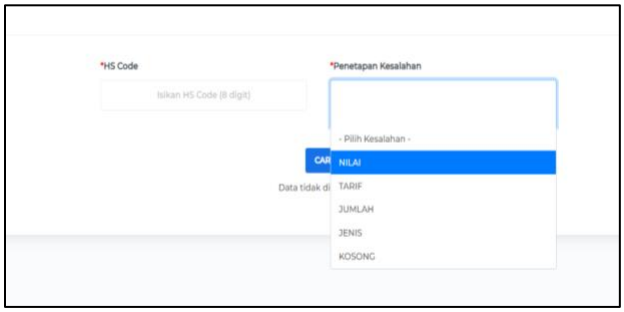

Fig. 11. Data Analytics Web Application

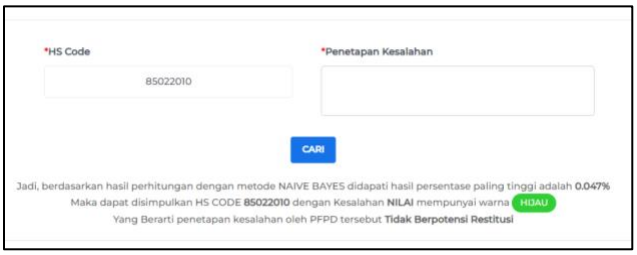

Fig. 12. HS Code 85022010 Classification

This application works by entering the HS Code manually in the search field. For example, if you want to search for a class category for HS Code 85022010 motorcycle commodity, then select the type of assignment error, with a database that works according to the rules of the Naive Bayes algorithm, the classification results for HS Code 85022010 motorcycle commodity will appear with a value of "Green," which means that the determination has no potential restitution. For the type of assignment error, you can choose more than one type and the expected classification results will appear.

IV. Conclusion

The level of confidence in the study's findings is high enough to be utilised as a classification consideration to estimate the risk of misclassification notices when compared to historical data on how frequently the HS Code has received objections, resulting in possible state revenue loss. The calculation accuracy resulting from the modelling is 99.97%, the classification error value is 0.03%, and the Kappa coefficient is 0.999, indicating that the data mining technique used is very good. This classification and rate forecast, however, is merely an initial risk evaluation and cannot be used as a direct reference for setting the HS Code. Further study is required, such as looking at supporting evidence and the direct appearance of goods. However, in the process of determining the right HS Code, it always requires professional adjustments from the Customs officer and is in accordance with applicable legal rules. HS Code data classification also includes many complex aspects that can continue to develop.

V. Acknowledgment

The author would like to express gratitude to the Directorate General of Customs and Excise for providing chances and information, both directly and indirectly, to enable the authors to conduct this research. Likewise, thanks to everyone who has assisted this research process begin with data collecting and continue to completion.

VI. References

[1] A. A. Irshadi and A. Wahyu Santoso, "PENGGUNAAN DATA MINING DALAM EKSTENSIFIKASI PENELITIAN ULANG," *J. Perspekt. BEA DAN CUKAI*, vol. 5, no. 2, pp. 218–132, Nov. 2021, doi: 10.31092/jpbc.v5i2.1305.

[2] L. Ding, Z. Z. Fan, and D. L. Chen, "Auto-categorization of HS code using background net approach," in *Procedia Computer Science*, 2015, vol. 60, no. 1, pp. 1462–1471. doi: 10.1016/j.procs.2015.08.224.

[3] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naive Bayes classification," *Comput. Oper. Res.*, vol. 135, Nov. 2021, doi: 10.1016/j.cor.2021.105456.

[4] 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, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[5] C. G. Skarpathiotaki and K. E. Psannis, "Cross-Industry Process Standardization for Text Analytics," *Big Data Res.*, vol. 27, Feb. 2022, doi: 10.1016/j.bdr.2021.100274.

[6] Y. I. Kurniawan, F. Razi, Nofiyati, B. Wijayanto, and M. L. Hidayat, "Naive bayes modification for intrusion detection system classification with zero probability," *Bull. Electr. Eng. Informatics*, vol. 10, no. 5, pp. 2751–2758, Oct. 2021, doi: 10.11591/eei.v10i5.2833.