International Journal of Research and Review
[Indexed, Peer-Reviewed Journal]
Volume 13; Issue: 6; June 2026
Website: www.ijrrjournal.com
Research Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

SMOTE-Based Supervised Learning Approaches for Import Cargo Routing Classification in Indonesian Customs

Affan Rafi Ardiansyah[1], Prajna Pramita Izati[2]

[1]Department of Accounting, State Financial Polytechnic STAN, Tangerang Selatan, Indonesia
[2]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

Corresponding Author: Prajna Pramita Izati

DOI: https://doi.org/10.52403/ijrr.20260622

ABSTRACT

This study compares the performance of several supervised learning methods in determining import lane status at KPPBC TMP B Teluk Bayur using variables such as CIF value, country of origin, importer status information, and HS Code. The evaluated methods include Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and Extreme Gradient Boosting (XGBoost), with SMOTE applied to address imbalanced data. The results show that all models are capable of classifying import documents into green lane and red lane categories. Among the models, XGBoost achieved the best overall performance with an accuracy of 96.3% and a balanced precision and recall value. Logistic Regression and SVM showed high recall values of 88.9%, indicating strong capability in detecting red lane cases, while Naive Bayes demonstrated lower overall performance. SHAP analysis revealed that CIF value, country of origin, importer status information, and HS Code were the most influential variables in determining import lane status. Overall, XGBoost can be considered the best model, although model selection should still depend on operational priorities between efficiency and risk detection capability.

Keywords: supervised learning, import lane classification, XGBoost, SHAP, customs risk management, SMOTE

INTRODUCTION

In the current era of globalization, international trade activities, especially imports, are one of the main supports in supporting the national economy (1). Import itself is an activity of entering goods into the customs area. The customs area includes the land area, waters, and airspace above it, as well as certain places in the Exclusive Economic Zone and the continental shelf in which the provisions of laws and regulations in the field of customs apply. KPPBC TMP B Teluk Bayur, as the front line in maintaining goods traffic in the West Sumatra region, is required to carry out the function of trade facilitator as well as community protector. The main challenge in the implementation of this function is how customs officers can speed up the import service process without neglecting the aspect of supervision of dangerous goods entering the customs area.

In carrying out its role as a Community Protector, the Directorate General of Customs and Excise uses a risk management system that divides imported commodities into several channels (lineup), namely Green Lane and Red Line. Green Lane is the process of service and supervision of the

expenditure of imported goods by conducting document research and physical examination of goods before the issuance of the Approval Letter for the Issuance of Goods (SPPB). Red Line is the process of servicing and supervising the expenditure of imported goods by conducting document research and physical examination of goods first before issuing the Approval Letter for the Issuance of Goods (SPPB). The determination of this path is vital. Misclassification can lead to two risks: under-protection (illegal goods escape) or over-regulation (logistics are hampered and warehouse costs are inflated) (2).

Currently, the determination of the status of the lineup is often based on the criteria of the importer's profile and the type of commodity that is static or using a rule-based system (3). Along with the ever-increasing volume of transaction data and increasingly complex patterns of breaches, these conventional systems are sometimes less flexible in capturing anomalies or new patterns hidden within historical data (4).

In the context of KPPBC TMP B Teluk Bayur, the main priority in import routing is security. This means that the system must be able to accurately identify high-risk shipments (Red Lines) to prevent the entry of illegal commodities or the flight of state revenues (5). Technically, system failure in detecting violations (false negatives) is much more dangerous than system errors that are too strict (false positives) (6).

Machine learning-based approaches, particularly supervised learning methods, offer promising solutions. This method allows the system to learn patterns from historical data that have been labeled and used to perform automatic classification (7). Several supervised learning algorithms such as Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and Extreme Gradient Boosting (XGBoost) have their own advantages in handling classification problems, both in terms of interpretability, accuracy, and the ability to handle complex and unbalanced data (8)(9)(10)(11).

However, there has been no study that specifically compares the performance of these algorithms in the context of determining the status of import pipelines at KPPBC TMP B Teluk Bayur. Each algorithm has different characteristics in handling data distribution, relationships between variables, and sensitivity to data imbalance, so comparative analysis is needed to determine the most optimal method in determining the alignment (12).

Based on this description, this study was conducted to analyze and compare the performance of several supervised learning methods in determining the status of import registration using variables of importer status, HS Code, CIF value, and country of origin. In addition, this study also considers the problem of data imbalance (Imbalance Data) as an important factor in model evaluation. The results of the study are expected to provide recommendations for the most effective and accurate methods, as well as contribute to the development of a more data-based and objective import route determination system in the KPPBC TMP B Teluk Bayur environment.

MATERIALS & METHODS

Types of Research

This research is quantitative research with *a machine learning* approach. A quantitative approach is used because this study focuses on numerical data processing as well as model performance measurement using statistical metrics. Meanwhile, *a machine learning* approach is used to build a classification model that is able to predict the status of import pipelines based on available historical data. The method used in this study is supervised learning, where the model is trained using data that already has a label in the form of an import tracking status, then used to predict new data.

Objects of Research

The object of this study is import document data processed at the Customs and Excise Supervision and Service Office (KPPBC) Intermediate Type B Customs in Teluk

Bayur. The data used reflects import activities that have gone through the process of determining the inspection route, so as to contain information related to the characteristics of imported goods and the status of the line determined by the customs system. The unit of analysis in this study is each imported document recorded in the dataset from 2021 to 2025. Each observation represents one import transaction that has attributes such as HS Code, CIF value, country of origin, and import path status (green lane and red lane). The selection of this research object is based on its relevance to the research objective, which is to analyze and compare *supervised learning methods* in determining the status of import tracing. The data used is considered representative because it comes from the real conditions of customs operations.

Types and Data Sources

The type of data used in this study is quantitative data, which is data that can be measured and processed numerically for the purposes of statistical analysis and *machine learning modeling*. The data in this study consists of a combination of numerical and categorical data that represent the characteristics of imported documents. Based on the source, the data used is secondary data, namely data obtained from related agencies without going through a direct collection process by researchers. The data in this study is sourced from official documents in the form of a dataset of import activities at the Customs and Excise Supervision and Service Office (KPPBC) of Intermediate Type B of Teluk Bayur.

The dataset contains imported historical data with a total of 2,163 observations. Each observation represents one import document that already has complete attributes, namely the office code (KD_KANTOR), import status information (KET_STATUS_IMP), HS Code, CIF (*Cost, Insurance, and Freight*) value, country of origin (NEG_ASAL), and import registration status (STATUS_JALUR).

The data used in this study has a good level of completeness, where no missing *value* is found in all variables. In addition, the data has gone through a recording process in the official customs system, so it has a high level of validity and reliability. The use of this document as the main data source is based on its suitability with the purpose of the research, which is to analyze and compare *supervised learning* methods in determining the status of import tracing. The historical data reflects real conditions in the process of determining import routes, thus allowing the built model to have high relevance to practices in the field.

Research Variables

The research variable is everything that is used as an object of observation in the research and has a variety of values. In this study, the variables are differentiated into dependent variables and independent variables used in the *supervised learning modeling process*

1. **Key Variables** is a variable used as the basis for the classification process of import routing status.

Table 1. Research Variables

Variables	Remarks	Data Type	Scale
X_1	Import Status Description (KET_STATUS_IMP)	Categorical	Nominal
X_2	HS Code (HS_CODE)	Categorical	Nominal
X_3	CIF Value (Cost, Insurance, and Freight)	Numerical	Ratio
X_4	Country of Origin (NEG_ASAL)	Categorical	Nominal

2. **Output Variable** is a variable used as a result of classification in the supervised learning model, namely Import Routing Status (STATUS_JALUR)

This variable shows the results of the determination of import paths classified into several categories, such as Green Lane (H) and Red Line (M). This variable is the main focus of the study because it represents the level of risk of an imported document.

Data Analysis Methods

The data analysis method in this study uses *a supervised learning* approach to classify the status of import rolling. The analysis was carried out by comparing several classification algorithms to obtain the best model based on the performance produced.

The stages of data analysis in this study are carried out systematically as follows:

1. **Data *Pre-processing***

a. Variable Coverage

In the initial stage, an examination of the completeness of the variables in the dataset was carried out. This study uses imported document data consisting of variables KET_STATUS_IMP, HS_CODE, CIF, NEG_ASAL, and STATUS_JALUR. Each observation is ensured to have a value on all of these variables so that it can be used in the classification process. Based on the results of the examination, the dataset has included all the required variables without any blank entries.

b. Missing Values Detection

Missing data checks are performed using data exploration functions in Python such as.isnull() or .info (). The results of the check show that there are no missing values in all variables in the dataset. Therefore, no data imputation process is required in this study.

c. Data Type Customization

Some variables such as HS_CODE and KD_KANTOR were adjusted to be categorical (object), even though the initial format was numerical. This is done because the variable represents a category, not a quantitative value that has a mathematical meaning.

d. Data Transformation and Encoding

In this study, categorical variables cannot be directly used by machine learning algorithms, so it is necessary to transform them into numerical forms. Target variables (STATUS_JALUR) are converted using encoding labels, while categorical feature variables such as KET_STATUS_IMP, HS_CODE, and NEG_ASAL are transformed using target encoding.

This approach was chosen because it is able to capture the relationship between categories and target variables in a more informative way than simple encoding methods.

e. Data Standardization (Data Scaling)

Standardization is done to ensure that all variables have a comparable scale. This is important because some algorithms such as Logistic Regression and SVM are sensitive to differences in data scale.

In this study, the standardization method (Z-score scaling) was used, which changed the distribution of data so that it had an average of close to zero and a standard deviation of one. Thus, no variables dominate the model learning process.

f. Handling *Imbalanced Data*

The dataset in this study has a class imbalance, where the amount of data on the green lane is much more than the red lane. To overcome this, the SMOTE (Synthetic Minority Over-sampling Technique) method was used on the training data.

This method produces synthetic data in a minority class so that the distribution of data becomes more balanced. The application of SMOTE is carried out only on training data to avoid data leakage.

2. Development of classification models

At this stage, a classification model was developed using several *supervised learning* algorithms to predict the status of import shipments. The model was built using trained data that has gone through *a preprocessing* process, including *encoding*, standardization, and handling of imbalanced data using SMOTE.

This study uses four classification algorithms, namely Logistic Regression, XGBoost, Support Vector Machine (SVM), and Naive Bayes. The selection of these four algorithms is based on the differences in the characteristics of each method, making it

possible to conduct a comprehensive performance comparison.

a. *Logistic Regression*

Logistic Regression is a linear classification algorithm used to model the relationship between independent variables and the probability of occurrence of a class. This model works by estimating the probability that an observation belongs to a particular class based on a linear combination of features (13). The Logistic Regression equation is:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

where $P(Y=1)$ is probability of an observation belonging to class 1, β_0 is intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are regression coefficients.

In this study, Logistic Regression is used as a baseline model because it is simple, easy to interpret, and has fairly good performance on relatively linear data (8). In addition, Logistic Regression remains one of the most widely used classification methods in machine learning research due to its computational efficiency and strong interpretability compared to more complex algorithms (14). The maximum iteration parameter is increased to ensure that the model convergence process runs optimally during training, especially when dealing with large-scale or imbalanced datasets (15).

b. XGBoost (*Extreme Gradient Boosting*)

XGBoost is an ensemble learning-based algorithm that uses boosting techniques to improve model performance. This algorithm works by building the model in stages, where each new model focuses on the mistakes made by the previous model (11). The general objective function of XGBoost can be written as:

$$\text{Obj}^{\{(t)\}} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

The prediction model in XGBoost is expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

In this study, XGBoost was used for its ability to handle complex data, including data with non-linear relationships and interactions between variables. Parameters such as the number of estimators, tree depth (max_depth), learning rate, and subsamples are adjusted to obtain optimal performance (8).

c. *Support Vector Machine* (SVM)

Support Vector Machine is a classification algorithm that works by looking for an optimal hyperplane that separates data between classes by maximum margin (9). In this study, a Radial Base Function (RBF) kernel was used which is able to capture non-linear patterns in the data.

The SVM hyperplane equation is:

$$f(x) = \omega^T x + b$$

The RBF kernel equation is:

$$K(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right)$$

SVMs are known to perform well on high-dimensional datasets and classification problems with complex boundaries (8), but are sensitive to data scale, so the standardization process is an important step before model training.

d. Naïve Bayes

Naive Bayes is a probabilistic algorithm based on Bayes' Theorem assuming independence between variables (10). Although this assumption is rarely perfectly fulfilled, Naive Bayes is still widely used due to its simplicity and efficiency in computing. The Bayes Theorem equation is:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

In this study, Gaussian Naive Bayes was used which is suitable for numerical data as a result of transformations. The Gaussian probability density function is:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(x_i-\mu_C)^2}{2\sigma_C^2}\right)$$

This model is used as a comparator because of its ability to handle small to medium-sized datasets (8).

3. Model Evaluation

After the model is trained, an evaluation is carried out to measure the performance of each algorithm in classifying the status of import queues. The evaluation was carried out using test data that was not used in the training process.

The evaluation metrics used include accuracy, precision, recall, F1-Score, ad AUC. In the context of this study, special attention was paid to the recall value for the red line class, as the class represented a high risk in import activities. A good model is expected to be able to minimize errors in detecting red lines.

4. Model Analysis (SHAP and Feature Importance)

To increase the interpretability of the model, an analysis was carried out on the contribution of each variable in the prediction process, especially in the XGBoost model. The method used is SHAP (SHapley Additive exPlanations), which is able to explain the contribution of each feature to the model's output (16).

The SHAP value can be expressed as:

$$\phi_i = \phi \sum_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

SHAP provides information about how much influence a variable has on increasing or decreasing the probability of a class. In addition, a feature importance analysis was carried out which showed the level of importance of each variable in the model. This analysis helps in understanding the key factors that influence the determination of the status of import threading.

5. Model Comparison

The final stage is carried out by comparing the performance of all models based on the evaluation metrics that have been obtained. This comparison aims to determine the best model that has a balance between accuracy and ability to detect minority classes. The model chosen is not only based on the highest accuracy value, but also considers the recall value and F1-score, especially for red line classes that have a higher level of risk.

RESULT

This study aims to compare the performance of several algorithms supervised learning in classifying the status of the import queue. Based on the results of data processing using Python, the results of the evaluation of four models were obtained, namely Logistic Regression, XGBoost, Support Vector Machine (SVM), then Naive Bayes. Evaluation is conducted using metrics accuracy, precision, recall, and F1-score, each of which provides a different picture of model performance, especially under unbalanced data conditions.

The results of the model evaluation are shown in the following Table 2:

Table 2. Model Evaluation Results

Model	Accuracy	Precision	Recall	F1-Score	AUC
XGBoost (Tuned)	0,9677	0,3333	0,5556	0,4167	0,9189
SVM (Tuned)	0,8129	0,0909	0,8889	0,1649	0,8737
Logistic Regression	0,7067	0,0597	0,8889	0,1119	0,8679
Naive Bayes (Tuned)	0,3972	0,0333	1,0000	0,0645	0,7854

From these results, it can be seen that there is a significant difference in performance between models. No single model excels across all metrics, which indicates a trade-off in model performance.

Imbalanced Data Analysis

One of the main characteristics of the dataset in this study is the imbalance in the amount of data between the green lane and red lane classes. The amount of green lane data is much more dominant than the red lane. This

condition causes the model to tend to be biased towards the majority class, so even though the accuracy value is high, the model's ability to detect the minority class is not necessarily good. This is evident in the XGBoost model which has high accuracy but relatively low recall.

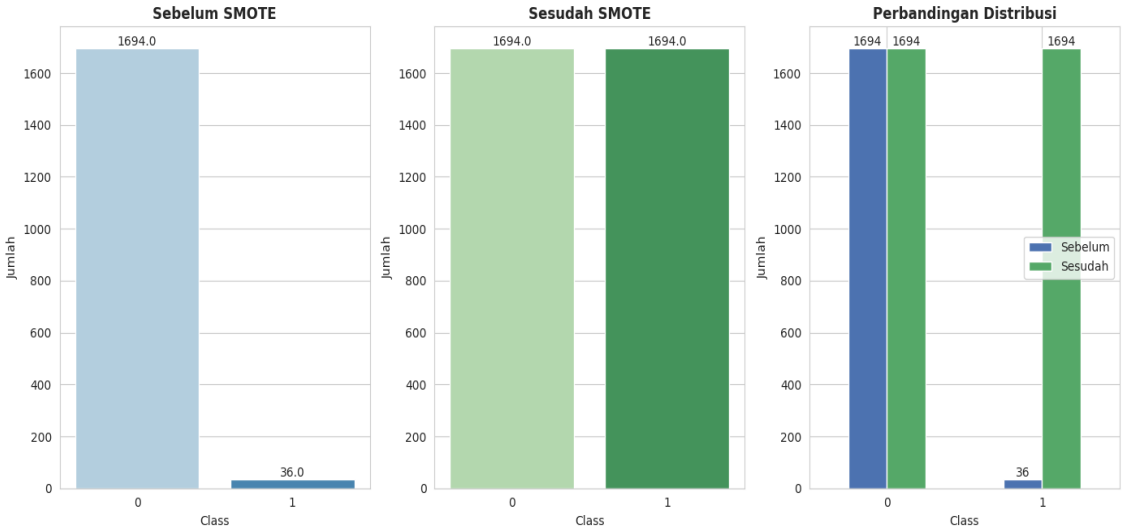

Figure 1. Differences in Class Distribution Before an After SMOTE

The use of the SMOTE method in this study aims to overcome this imbalance by adding synthetic data in minority classes. However, the results show that although SMOTE is helpful, the challenge of detecting minority classes remains not completely solved. The results showed that after the implementation of SMOTE, the amount of data in class 0 and class 1 became balanced, namely 1,694 observations each. This indicates that the SMOTE technique has succeeded in overcoming class imbalances by adding synthetic data to minority classes so that the distribution of data is balanced. This condition is expected to help the model in learning more fairly for both classes without bias towards the majority class.

Discussion per Classification Model

In the next stage, a model performance analysis was carried out in classifying the status of import pipelines using a supervised learning approach. This study uses four classification algorithms, namely Logistic Regression, XGBoost, Support Vector Machine (SVM), and Naive Bayes. The use of these algorithms aims to gain a more comprehensive understanding of the ability of each method to capture data patterns and to deal with the problem of imbalanced data contained in the dataset.

In contrast to unsupervised learning approaches such as clustering which focuses on grouping without labels, the supervised learning approach in this study utilizes target variables in the form of import routing status (green lane and red lane) as the basis for model learning. Thus, each algorithm is trained to recognize the pattern of relationships between independent variables, such as HS Code, CIF values, country of origin, and import status, to target variables that reflect the level of risk of an import activity.

The selection of the four algorithms is based on the differences in the characteristics and approaches of each method in conducting classification. Logistic Regression is used as a simple and easy-to-interpret linear model, so that it can provide a basic overview of the relationships between variables. Meanwhile, XGBoost was chosen for its ability to handle non-linear relationships and complex interactions between variables through a boosting-based ensemble learning approach. Furthermore, the Support Vector Machine (SVM) is used because of its ability to form optimal decision boundaries (optimal

hyperplane) that are able to effectively separate classes, especially on data with high dimensions and non-linear patterns. On the other hand, Naive Bayes is used as a probabilistic model that assumes independence between variables, so that it can provide different perspectives in understanding the distribution of data.

By combining the four algorithms, this study aims not only to determine the model with the best performance in general, but also to analyze the differences in the characteristics of each model in detecting minority classes (red lines) that have a higher level of risk. This is important because in the context of customs, errors in detecting red lines (false negatives) have more serious consequences than errors in classifying green lines as red lines (false positives).

In addition, the use of several models also allows comparative analysis related to the trade-off between accuracy, precision, and recall. Each model has a different tendency, with some models focusing more on improving overall accuracy, while others are more sensitive to detecting minority classes. Therefore, the results of each model are not only evaluated based on a single metric, but are thoroughly analyzed to understand their implications for operational efficiency and risk mitigation in the import routing process. With this approach, it is hoped that the research will be able to provide a more comprehensive and in-depth picture of the performance of various supervised learning methods, as well as produce model recommendations that are not only statistically optimal, but also relevant to be applied in the context of decision-making in the field of customs.

a. Logistic Regression Model

Table 3. Confusion matrix Logistic Regression Model

Actual \ Predicted	Green Lane (H)	Red Line (M)	Total
Green Lane (H)	298	126	424
Red Line (M)	1	8	9
Total	299	134	433

Table 4. Classification Report Logistic Regression Model

Classes	***Precision***	***Recall***	F1-Score	Support
Green Lane (H)	1,00	0,70	0,82	424
Red Line (M)	0,06	0,89	0,11	9
Accuracy			**0,71**	433
Macro Avg	0,53	0,80	0,47	433
Weighted Avg	0,98	0,71	0,81	433

Visualization of the results of Table 3. It can be seen in Figure 2 below.

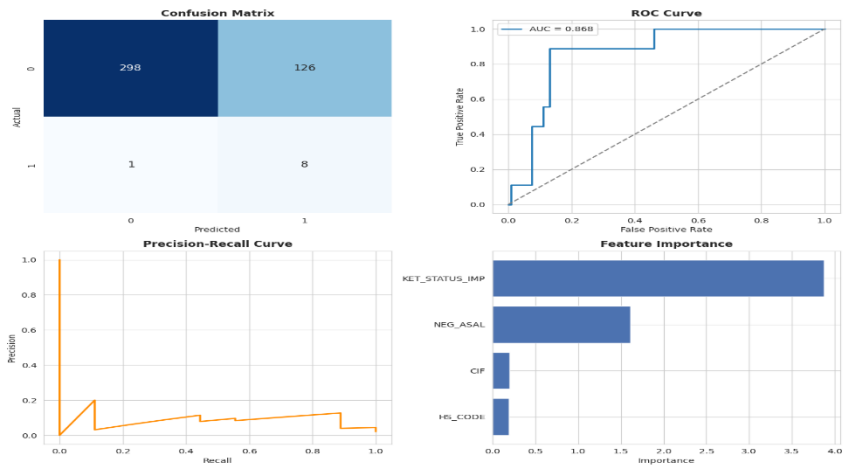

Figure 2. Confusion matrix, ROC Curve, Precision-Recall curve, and Feature Importance Model Logistic Regression

The Logistic Regression *model* shows classification performance with an accuracy rate of 70.67% based on Table 3, which indicates that the model is able to classify most of the data quite well. However, a more in-depth evaluation through *the confusion matrix* in Figure 2 shows that the model has characteristics that tend to be stronger in detecting minority classes, namely the red line. Out of a total of 9 red-track data, the model managed to identify 8 data correctly and only failed on 1 data, resulting in a high recall value of 88.9%. This suggests that the model has excellent sensitivity to risk, making it less likely that red line cases will be missed.

However, the model's ability to provide accurate predictions is still relatively low. This can be seen from the *precision* value of only 5.97%, which shows that most of the red line predictions produced are actually green lines. This condition is strengthened by the high number of false positives, which are as many as 126 cases, where the green lane data is erroneously classified as a red lane as a result, although the model is effective in detecting risks, operational efficiency is declining due to the increasing number of items to be inspected.

The significant difference between high *recall* values and low *precision* is also reflected in the relatively small F1-score value of 0.11 for the red line class. This shows that the balance between detection capabilities and prediction accuracy is still not optimal. This condition is inseparable from the characteristics of unbalanced data, where the amount of green lane data is much more dominant than the red lane data. The use of the SMOTE technique in this study did help increase the sensitivity of the model to minority classes, but on the other hand it also caused the model to become more aggressive in classifying the data as a red line.

Overall, Logistic Regression can be categorized as a risk-averse model, which is a model that prioritizes risk detection over efficiency. In the context of customs, this model has the advantage of minimizing errors in the form of false negatives, which are the most crucial errors. However, the high number of false positives suggests that this model is less efficient to be applied directly without further adjustments. Therefore, additional strategies such as threshold adjustments or combinations with other models are needed to achieve a balance between risk mitigation and operational efficiency.

XGBoost Model Analysis

Table 5. Confusion Matrix Results of XGBoost Model

Actual \ Predicted	Green Lane (H)	Red Line (M)	Total
Green Lane (H)	414	10	424
Red Line (M)	4	5	9
Total	418	15	433

Table 6. Table Classification Report XGBoost Model

Classes	***Precision***	***Recall***	F1-Score	Support
Green Lane (H)	0,99	0,98	0,98	424
Red Line (M)	0,33	0,56	0,42	9
Accuracy			**0,97**	433
Macro Avg	0,66	0,77	0,70	433
Weighted Avg	0,98	0,97	0,97	433

Visualization of the results of Table 5. It can be seen in the following Figure 3.

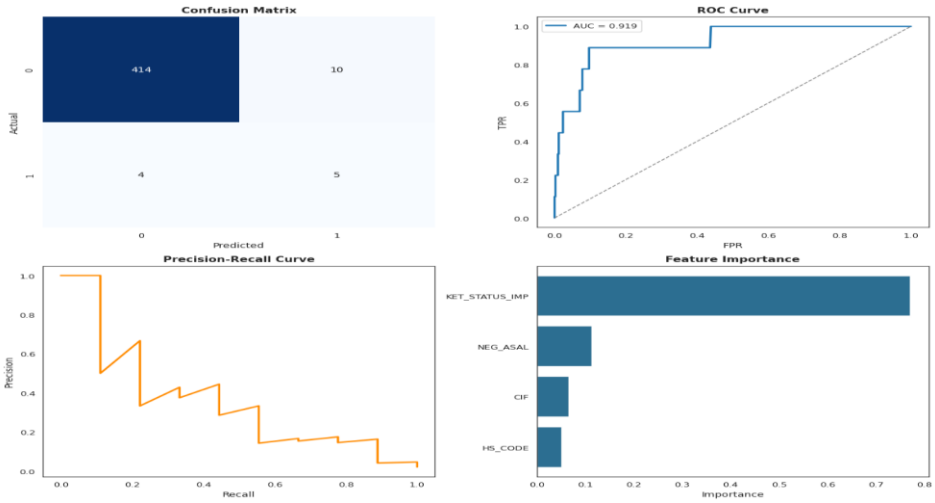

Figure 3. Confusion Matrix, ROC Curve, Precision-Recall curve, and Feature Importance XGBoost Model

The XGBoost model shows excellent performance with an accuracy rate of 96.7%, which is the highest value compared to other models. This value shows that the model is able to classify most of the data very accurately, especially in the majority class, i.e. the green path. This can be seen from the confusion matrix, where 414 out of 424 green path data were correctly classified, resulting in a very high recall value for the class (97%).

However, if you look deeper at the model's ability to detect red lanes as a minority class, the model's performance is still relatively limited. Out of a total of 9 red lane data, the model was only able to correctly identify 5 data, while the other 4 data were incorrectly classified as green lanes. This condition resulted in a recall value of 55.5%, indicating that more than half of the red line cases were not successfully detected by the model. These errors fall into the category of false negatives, which are the most crucial type of error in the context of customs because they can cause the goods to risk escaping without inspection.

On the other hand, the XGBoost model had a precision value of 33.3% for the red lane class, which was the highest value compared to other models in the study. This shows that when a model predicts a data as a red line, the probability of the prediction being true is relatively higher than that of other models. The low number of false positives (only 10 cases) also shows that the model is quite selective in predicting red lines, so that operational efficiency can be maintained.

The difference between the precision and recall values in the red lane class shows that there is a significant trade-off. The XGBoost model tends to be more conservative in classifying data as red paths, resulting in better accuracy, but with the consequent lower ability to detect all red line cases. This reflects XGBoost's characteristics as an ensemble learning-based model that seeks to optimize overall performance by minimizing global errors, rather than specifically focusing on minority classes.

The performance of this model is also inseparable from the influence of data imbalance (imbalanced data) which is still the main challenge. Even though the handling has been carried out using SMOTE, the very uneven distribution of data still affects the learning process of the model, so that the model is more likely to recognize patterns from the majority class.

In the context of customs operations, the XGBoost model has an advantage in terms of efficiency because it is able to minimize the number of unnecessary checks. However, the main drawback of this model lies in the high number of false negatives, which means that there is a risk that goods that should be in the red lane will be classified as green lanes. Therefore, the use of this model needs to be accompanied by additional strategies, such as

threshold adjustments or integration with other risk selection systems.

Overall, the XGBoost can be categorized as a model with high accuracy and high precision characteristics, but low recall for the minority class. This model is particularly suitable for use in conditions that emphasize operational efficiency, but needs to be further optimized when used in contexts that require a high level of risk detection

Support Vector Machine **(SVM) Model**

Table 7. Results of the SVM Model Confusion matrix

Actual \ Predicted	Green Lane (H)	Red Line (M)	Total
Green Lane (H)	344	80	424
Red Line (M)	1	8	9
Total	345	88	433

Table 8. Table Classification Report SVM Model

Classes	***Precision***	***Recall***	F1-Score	Support
Green Lane (H)	1,00	0,81	0,89	424
Red Line (M)	0,09	0,89	0,16	9
Accuracy			**0,81**	433
Macro Avg	0,54	0,85	0,53	433
Weighted Avg	0,98	0,81	0,88	433

The Support Vector Machine (SVM) model produced an accuracy rate of 81.3%, which indicates that the model is able to classify most of the data quite well. When compared to other models, SVM has a relatively balanced performance between risk detection capabilities and misclassification rates, although there are still some drawbacks to be noted.

Based on the confusion matrix, this model managed to correctly classify 344 out of 424 green lane data, and was able to detect 8 out of 9 red lane data. This resulted in a high recall value for the red line class of 88.9%, which indicates that the model has excellent sensitivity to minority classes. With only 1 false negative case, the risk of missing risky goods can be significantly minimized.

Nevertheless, the model still produces 80 false positive cases, which are conditions where the green line is classified as the red line. This causes the *precision* value for the red line class to be low, which is 9.1%. In other words, most of the red path predictions that the model generates are not actually red lines.

Visualization of the results of Table 7. It can be seen in the following Figure 4.

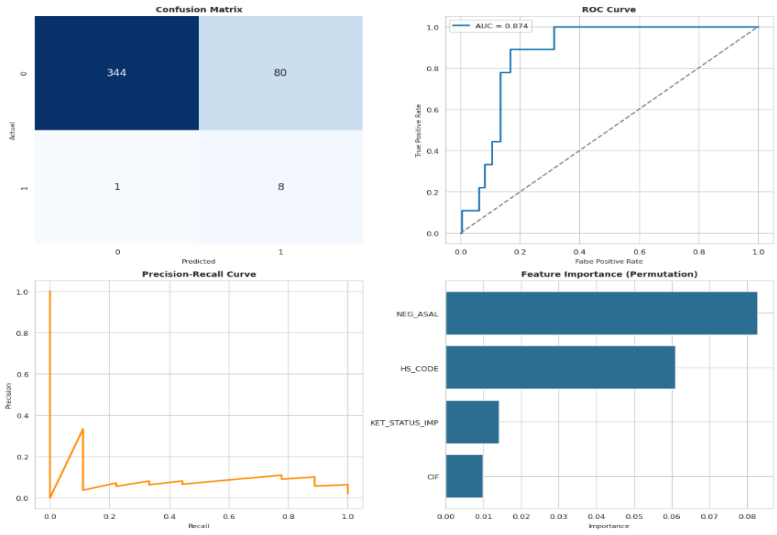

Figure 4. Confusion matrix, ROC Curve, Precision-Recall curve, and Feature Importance SVM

This condition shows a trade-off between the ability to recall and the accuracy of predictions. SVMs tend to be more aggressive in classifying data as red lines to ensure that risks are not missed. This causes the model to have similar characteristics to Logistic Regression, namely high recall but low precision, albeit with a slightly lower error rate.

In terms of concept, SVM works by forming a separator boundary (hyperplane) that maximizes the margin between classes. The use of the RBF kernel in this study allowed the model to capture non-linear patterns in the data. However, under unbalanced data conditions, the model tends to expand the boundaries of decisions in the direction of minority classes, resulting in more data being classified as red lines. This explains the high value of the recall as well as the increase in the number of false positives.

If associated with data characteristics, the use of SMOTE in the training process also affects the model's behavior. With the increasing amount of synthetic data in the red line class, models are becoming more sensitive to patterns associated with risk. However, this also causes the model to tend to overgeneralize the characteristics of the red line.

In the context of customs operations, the SVM model has an advantage in terms of risk mitigation, as it is able to detect almost all red line cases. This is very important to prevent potential violations and state losses. However, the weakness lies in efficiency, as a large number of false positives will increase the burden of inspecting goods. Thus, SVM can be categorized as a model that is in a compromising position between security and efficiency. This model is suitable for use in situations where risk detection remains a priority, but with moderate operational efficiency in mind.

Naive Bayes Model

Table 7. Results *of the Confusion matrix* of Bayes' Naïve Model

Actual \ Predicted	Green Lane (H)	Red Line (M)	Total
Green Lane (H)	163	261	424
Red Line (M)	0	9	9
Total	163	270	433

Table 10. Table *Classification Report* Model Naïve Bayes

Classes	*Precision*	*Recall*	F1-Score	Support
Green Lane (H)	1,00	0,38	0,56	424
Red Line (M)	0,03	1,00	0,06	9
Accuracy			**0,40**	433
Macro Avg	0,52	0,69	0,31	433
Weighted Avg	0,98	0,40	0,55	433

Visualization of the results of Table 9. It can be seen in Figure 5 below.

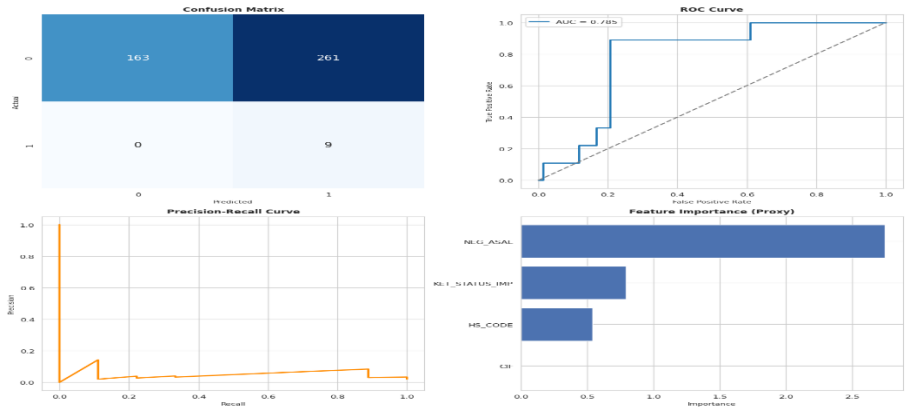

Figure 5. Confusion matrix, ROC Curve, Precision-Recall curve, and Feature Importance Naïve Bayes Model

The Naive Bayes model showed quite extreme performance compared to other models, with an accuracy rate of only 39.7%, which was the lowest score in the study. Despite this, the model has unique characteristics, especially in its ability to detect minority classes.

Based on the confusion matrix, the model managed to correctly classify all 9 red lane data, resulting in a recall value of 100%. This shows that the model has a very high ability to detect risks, not even a single case of red line being missed (false negative= 0). In the context of customs, this condition is theoretically ideal because the entire potential risk can be identified.

However, these advantages are followed by very significant weaknesses. The model produced 261 false positive cases, i.e. green lane data classified as red lanes. This amount is very large compared to the total green lane data, resulting in a very low *precision* value for the red lane class, which is only 3.3%. In other words, almost all of the model's red line predictions are inaccurate.

As a result of this high false positive, the overall *accuracy* value becomes very low. This suggests that while the model is capable of detecting the entire risk, it fails to properly classify the majority of data. The very small F1-score value (0.06 for the red lane) also indicates that the balance between recall and *precision* is not reached.

Conceptually, *Naive Bayes* works on the assumption that each feature is independent of each other. In practice, this assumption is rarely met, especially in imported data that has a relationship between variables, such as the relationship between HS Codes, CIF values, and countries of origin. The model's inability to capture the relationships between these variables causes the resulting probability distribution to be less accurate.

In addition, the use of SMOTE in the training process likely reinforces the model's bias against minority classes. With the increasing amount of synthetic red-path data, models become highly sensitive to the characteristics of those classes and tend to classify most of the data as red-paths.

In the context of customs operations, this model can be categorized as a very inefficient model. While there is no risk of being missed, the number of checks generated will be enormous as almost all goods are classified as red lines. This can cause overload on the inspection system and degrade overall operational performance. Thus, Naive Bayes can be categorized as a model with extreme sensitivity characteristics without error control, making it unsuitable for use as the main model in the import pathing system.

DISCUSSION

Based on the overall evaluation results, there is no single model that is perfectly superior in all assessment metrics. Each algorithm exhibits different performance characteristics as a consequence of the learning approach used as well as the unbalanced data conditions. Therefore, determining the best model in this study cannot be done by looking at just one indicator, such as accuracy, but must consider the balance between risk detection capabilities (recall), prediction accuracy (precision), and operational implications of misclassification. The XGBoost model shows the best performance in terms of accuracy (96.3%) and has a relatively higher level of precision than other models in detecting red lanes. This shows that the model is able to classify the overall data very well and be more selective in providing red line predictions. The low number of false positives is also an advantage because it can maintain operational efficiency, especially in reducing unnecessary inspection burdens. However, the limited recall value (44.4%) shows that this model is not optimal in detecting all red line cases, so there is still a significant risk of false negatives.

On the other hand, models such as Support Vector Machine (SVM) and Logistic Regression show better performance in terms of recall (88.9%), which means that it is able to detect almost all red line cases. This is very important in the context of customs because errors in the form of undetectable red lines

(false negatives) have a much more serious impact than other errors. However, both models have a weakness in low precision, which is reflected in the high number of false positives. This condition has the potential to reduce operational efficiency because it increases the number of goods that must be inspected even though it is not risky.

Meanwhile, the Naive Bayes model shows unbalanced performance with perfect recall (100%) but very low precision and accuracy. This suggests that the model is too sensitive to minority classes and is incapable of providing an overall accurate classification, making it less feasible to use in real implementations.

Taking all these aspects into account, the XGBoost model can be categorized as the best model in general because it is able to provide the most optimal performance in the context of a balance between accuracy and efficiency. However, if the main goal is to maximize risk detection, then models such as SVM or Logistic Regression are more appropriate alternatives because they have a higher level of sensitivity to red lines.

Thus, the selection of the best model in this study is not absolute, but rather depends on the priorities and objectives of use in the context of customs operations. If the system emphasizes more on safety and risk mitigation, then a high-recall model is more recommended. On the other hand, if operational efficiency is the top priority, then models with high accuracy and precision are more suitable. Therefore, the most ideal approach is to consider a combination of strategies, such as threshold adjustments or the integration of multiple models, in order to achieve the optimal balance between risk mitigation and efficiency in the import routing system.

In addition to evaluating the performance of the classification model, this study also analyzed the most influential variables in determining the status of import lines. This analysis aims to provide a deeper understanding of the factors that are the basis for model decision-making, so that the results of the research are not only predictive, but also interpretive and can be used as a consideration in policy-making in the field of customs.

Based on the results of interpretation using SHAP method on the XGBoost model, it was found that the most influential variables in determining the status of import routing in order were Cost, Insurance, and Freight (CIF), Country of Origin (NEG_ASAL), Importer Status Description (KET_STATUS_IMP), and HS Code. The sequence shows the relative contribution rate of each variable to the output of the model, where the variable with a higher SHAP contribution value has a greater influence on the classification process.

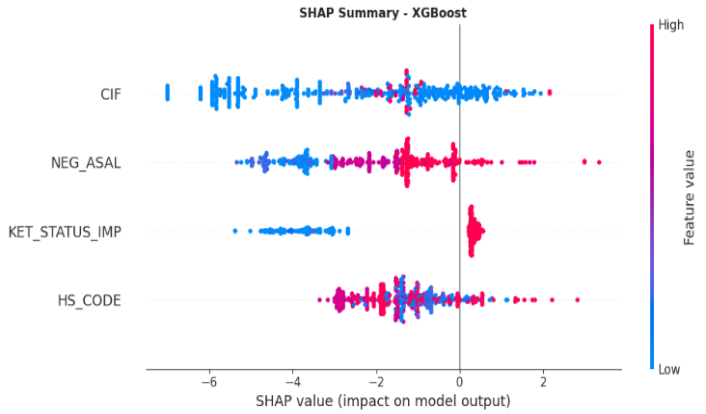

Figure 6. *SHAP Summary* XGBoost Model (Best)

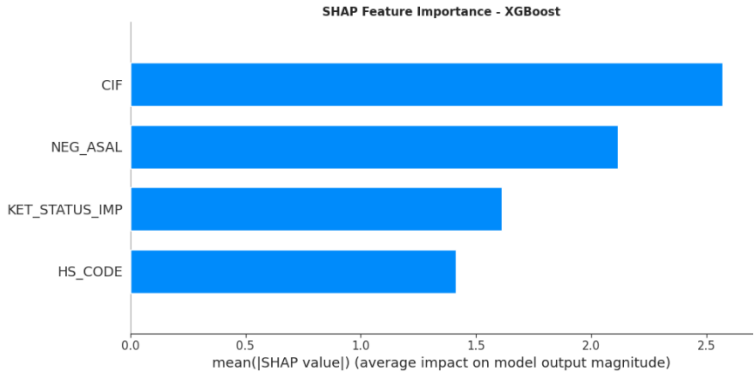

Figure 7. *SHAP Feature Importance* Model XGBoost (Best)

The CIF variable is the variable with the most dominant level of influence. This shows that the value of imported goods has a very significant role in determining the status of the line. Conceptually, CIF represents the total value of imported goods, so the higher the CIF value, the greater the potential risk inherent in the transaction. *The* XGBoost *model* is able to capture patterns that variation in CIF values correlate with a certain level of risk, so this variable becomes a key indicator in the classification process. The dominance of CIF in the SHAP summary plot also indicates that the economic value factor of transactions is the main determinant in determining the import route.

The Country of Origin (NEG_ASAL) variable ranks second in the level of influence. This shows that the country of origin of imported goods is an important factor that also influences model decisions. In customs practice, countries of origin are often associated with a certain risk profile based on historical data, trade policies, and international trade characteristics. The results of the SHAP analysis show that some countries of origin make a positive contribution to the likelihood of a transaction being classified as a red lane, while others tend to be associated with a green lane. This shows that the model is able to identify geographically-based risk patterns quite well.

Furthermore, the variable Importer Status Information (KET_STATUS_IMP) also has a significant influence on the classification results. This variable reflects the characteristics and profile of the importer, which in a customs risk management system is one of the important factors in determining the level of compliance and potential risk. *The* XGBoost model identifies that the status of certain importers has a different tendency to influence path setting, so these variables play a role in reinforcing the classification decisions generated by the model.

The HS Code variable is in fourth place, but it still makes a significant contribution to the classification process. The HS Code represents the type and classification of imported goods, which is directly related to the level of supervision in customs activities. Although the influence is not as large as the CIF variables, country of origin, and importer status, the HS Code still plays a role in helping the model recognize risk patterns based on commodity type. This shows that the model considers not only the value and origin aspects of the goods, but also the characteristics of the goods themselves in determining the status of the goods.

Overall, the results of this analysis show that *the* XGBoost *model* relies on a combination of several key variables in determining the status of the import queue. The dominance of CIF variables and countries of origin confirms that the aspect of transaction value and origin of goods is the main indicator in risk assessment. Meanwhile, the importer status information variable and HS Code function as supporting factors that enrich information in the classification process.

These findings suggest that the variables considered important by the model are in line with risk management principles in customs,

thereby increasing confidence in the results of the resulting model. Thus, the analysis of these variables not only provides an understanding of the working mechanism of the model, but also makes a practical contribution in supporting the development of a more effective, transparent, and data-based import routing system.

CONCLUSION

Based on the results of the research that has been conducted regarding the comparison of supervised learning methods in determining the status of import lines at KPPBC TMP B Teluk Bayur, several conclusions can be drawn as follows. The application of the supervised learning method has proven to be able to be used to classify the status of import routing by utilizing variables such as CIF values, country of origin, information on importer status, and HS Code. The built model is able to identify patterns in historical data and classify imported documents into green lane and red lane categories. The results of the comparison show that each model has different performance characteristics. The XGBoost model has the best performance in terms of accuracy, which is 96.3%, and is able to provide a relatively good balance between precision and recall. This model excels in operational efficiency because it is able to minimize overall misclassification. The Support Vector Machine (SVM) and Logistic Regression models show excellent ability to detect red lines with a high recall value, which is 88.9%. This suggests that both models are more sensitive to minority classes, making them more effective in detecting potential risks. However, the main drawback of both models lies in the low precision, which leads to a high number of false positives. The Naive Bayes model shows suboptimal performance with a low level of accuracy, despite having a very high recall value. This suggests that models tend to be overly sensitive to minority classes and are unable to provide a balance between precision and sensitivity. The results of the analysis using SHAP show that the most influential variables in determining the status of import routing in order are CIF, country of origin, information on importer status, and HS Code. These findings suggest that the model relies on transaction value factors and the origin of goods as key indicators in risk assessment, which is in line with risk management practices in customs. Thus, there is no single model that is absolutely superior in all aspects. The XGBoost model can be considered the best model in general, but the selection of the model should still be tailored to the intended use, whether it is more emphasis on operational efficiency or on risk detection capabilities.

Declaration by Authors

Acknowledgement: None
Source of Funding: None
Conflict of Interest: No conflicts of interest declared.

REFERENCES

1. Krugman PR., Obstfeld Maurice, Melitz MJ. International economics: theory & policy. Pearson; 2021.
2. Aigner D, Lovell CAK, Schmidt P. Formulation and estimation of stochastic frontier production function models. J Econom. 1977 Jul;6(1):21–37. doi:10.1016/0304-4076(77)90052-5
3. Sutton RS., Barto AG. Reinforcement learning: an introduction. The MIT Press; 2020. 526 p.
4. Han Jiawei, Kamber Micheline, Pei Jian. Data mining: concepts and techniques. Elsevier/Morgan Kaufmann; 2012. 703 p.
5. Widdowson D. Number 2 63 World Customs Journal. Vol. 14.
6. Provost Foster, Fawcett Tom. Data science for business. O'Reilly; 2013. 386 p.
7. Alpaydin Ethem. Introduction to machine learning. The MIT Press; 2020. 682 p.
8. James Gareth, Witten Daniela, Hastie Trevor, Tibshirani Robert. An introduction to statistical learning: with applications in R. Springer; 2021.
9. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep;20(3):273–97. doi:10.1007/BF00994018
10. Zhang H. The Optimality of Naive Bayes [Internet]. Available from: www.aaai.org

11. Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785–94. doi:10.1145/2939672.2939785
12. Haibo He, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009 Sep;21(9):1263–84. doi:10.1109/TKDE.2008.239
13. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002 Oct;35(5–6):352–9. doi:10.1016/S1532-0464(03)00034-0
14. Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. Nat Methods. 2018 Jan 3;15(1):5–6. doi:10.1038/nmeth.4551
15. Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIERGRISEL O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. Journal of Machine Learning Research [Internet]. 2011. Available from: http://scikit-learn.sourceforge.net.
16. Lundberg SM, Allen PG, Lee SI. A Unified Approach to Interpreting Model Predictions [Internet]. Available from: https://github.com/slundberg/shap

How to cite this article: Affan Rafi Ardiansyah, Prajna Pramita Izati. SMOTE-Based supervised learning approaches for import cargo routing classification in Indonesian customs. *International Journal of Research and Review*. 2026; 13(6): 219-235. DOI: *https://doi.org/10.52403/ijrr.20260622*

******