Individual Risk Classification of Crime Groups using Ensemble Classifier Method

Ardhito P. Anggana, Amalia Zahra
Computer Science Department, BINUS Graduate Program
Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

Abstract*—The most significant challenge for humanity worldwide to crime, especially terrorist attacks, should be considered. Determining the priority scale for anticipating individual terrorist groups is not easy and will significantly affect work activities and subsequent decision-making measures. Priority scale determination decisions should be made carefully so team members cannot choose the desired priority target. Determining the exact priority scale for a target can be influenced by several factors, such as desire factors and ability factors, using Dataset Intelligence. This research aims to find out the ability of each target and pattern to be carried out. Based on this problem, the study used the K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), and Ensemble Bagging methods. Each of these algorithms has its characteristics; This classification technique can group priority targets according to their similarities, abilities, and desires. The value of each method used can be used as a reference to determine the correct group information for officers to determine the next steps. The study obtained a maximum accuracy of 70.25% using the Ensemble Bagging-Backward Elimination-K-Nearest Neighbor (KNN) classification method using 20 features. The results showed tests conducted and final analysis and conclusions based on accuracy and recall performance. The precision performance revealed that the Ensemble Bagged KNN, more precisely than KNN, Naïve Bayes, Decision Tree, and Bagging Naïve Bayes and Bagging Decision Tree. The KNN Bagging ensemble model can add accuracy, map individuals, and detect who should be intensely monitored based on predictive results.

***Keywords*—Priority scale; data mining; classification; ensemble classifier method**

I. Introduction

The crime of terror is one of the most challenging threats to the global community. Its heterogeneous and complex nature has fostered an increasing interest in the scientific community, primarily to inform policy-oriented measures. The recent availability of large data sets, the diffusion of powerful machines, and advances in mathematical modeling techniques have contributed to the development of several approaches to the study of terrorism [1]. However, it is alarming that the sheer abundance of data makes it nearly impossible for authorities to examine every person, conversation thread, or social media post to classify whether they are linked to terrorism or contain elements of terrorist activity [2].

Forms of criminal acts spread in the corridors of terrorism can be intimidation and threats, murder, persecution, bombings, detonation, arson, kidnapping, hostage-taking, and piracy. The impact of these forms of terror is very diverse, including the onset of panic, feelings of fear/intimidation, worry, loss of property, incision, and even death. Current hidden or missing link prediction models based on network analysis models rely on machine learning techniques to improve model performance in terms of prediction accuracy and computing power [3]. In addition, with the increasing use of computerized systems to track crimes, computer data analysts have begun helping law enforcement officers and analysts to speed up the crime resolution process [4].

Determining parameters is one of the most vital things in the division based on desire and ability factors to overcome how the data mining process can classify the individual risk of criminal groups to determine priority scales and predict the priority scale in each individual as well as how the comparison of trial results from the model used can predict the priority scale. Determining these parameters is not easy for indicated people and will affect the person's activities indicated and included in the subsequent investigation. Trend analysis is a challenging task because crime data relies heavily on timing. Any data collected around criminal behavior and crime types can repeatedly change during the investigation [5]. In this study, the priority scale in this grouping was divided into three priority scales. With the use of priority scales, we can find out the abilities and desires of the person, indicating whether to commit a criminal act or not.

This research will develop previous research [6]. It presents the best machine learning models that can be used on terrorism-related data to predict terrorist groups most accurately. The decision to take a subset of data analysis was to help overcome the limitations. However, this is still possible to collect, record, and process data using the Decision Tree, Naïve Bayes, K-Nearest Neighbour, and Ensemble Bagging approaches and use optimal classification in analyzing individual data. Datasets using intelligence data with data used amounted to 1088 data with 21 attributes. Of the 1088 data, are people indicated to be committing a crime of terror or related to the act. However, in this study, researchers wanted to determine which attributes affect and do not affect the process with the four algorithms mentioned earlier in the feature selection process. So that later it can be used to form a reliable model in knowing the patterns of individuals who have the possibility of entering the Green, yellow, and orange priority scale using the Ensemble Classifier Method Risk Classification Model [7] because it is theoretically and empirically proven to provide much better performance than Single Learner [8], [9]. This research is limited to RapidMiner as a tool, data using intelligence data, not doing the data imbalance data process.

II. Theory

In this study, individual risk in the context of terrorism provides a unique opportunity to holistically consider risk factors rather than the individual critical factors often given in analysis [10]. Furthermore, understanding the modus operandi of each terrorist group provides a vast advantage to counter-terrorist institutions so that the necessary steps can be taken first to address the threat posed by those groups [6]. The machine learning approach can solve problems by finding a suitable algorithmic model and is better at generating predictive values from an input variable [7], and has four categories that are generally applied to the concept of data mining [11] is supervised learning, unsupervised learning, semi-supervised learning, and active learning.

In this study, the main categories of Machine Learning used utilizing existing data to perform classification [12]. Such models to perform introduction/classification/prediction is used in crime analysis [13]. When solving problems, no algorithm that provides the desired quality is proposed to use a composition or ensemble algorithm [14]. Data input for classification is a collection of records [15], where x is a set of attributes and y is a particular attribute. Classification models are helpful for descriptive encodings for distinguishing objects from different classes [16] and predictive modeling to predict class labels from unknown records [17]. Classification algorithms will produce patterns or rules that can be used to predict classes. Some of them are Naïve Bayes Classifier[18], Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN), and Random Forest Classifier.

The decision tree [14] is one of the exciting classification algorithms for taking measurements using a tree structure consisting of a collection of decision nodes connected by branches from the decision root to the leaf node to produce new decisions until they finally find the correct decision (leaf node) [19]. Data testing is conducted at each decision node[9] to separate datasets into subsets based on data homogeneity. Generally, the Decision Tree method used in modeling is Decision Tree CART. The CART algorithm [7] can build classification and regression modeling using the Gini Index [20] for the attribute selection process. Criteria determine the model of the decision tree formed, which is measured using the formula Entropy.

$$H = \sum P_{k^{log_2(P_k)}} \quad (1)$$

K-Nearest Neighbor (KNN), One classification method that can train the model without using parameters (non-parametric) [21] by classifying the object with the most vote values of each predefined object. The technique that can be used to measure the distance between two points or tuples of them is the Euclidean distance technique. Let us say point X is $\mathrm{X_1} = (\mathrm{x_1}, \mathrm{x_2}, \dots, \mathrm{x_n})$, Then point Y is $\mathrm{Y_1} = (\mathrm{y_1}, \mathrm{y_2}, \dots, \mathrm{y_n})$. Then the measurement formula used.

$$dist\ (d) = \sqrt{(x_1 - y_1)^2 + (x_1 - y_1)^2 + \dots + (x_n - y_n)^2} \quad (2)$$

The Naive Bayes algorithm is a supervised learning approach used for classifying to predicting target variables. Generally, classification techniques predict labeled classes based on attributes by looking for significant correlations between input and output variables. Naïve Bayes can be a simple probabilistic classifier that can build modeling on large datasets without estimating complex parameters [22]. The basic formula used in Bayes' theory.

$$P = \left(Y \mid X = \frac{P\ (Y) \times P\ (\ Y \mid X)}{P\ (X)} \right) \quad (3)$$

Ensemble Classifier Method is a diverse concept of modeling methods used to solve the problem of base learners by developing and combining a set of hypotheses to correct training data weaknesses using a single-learners approach [7]. This method is generally used in classification to build a model, including bagging, boosting, and stacking by reducing errors and optimizing accuracy results to be better than the base classifier itself [7]. The basic formula used in Ensemble Classifier's theory is.

$$F(x) = c_0 + \sum_{m=1}^{M} c_m T_m\ (x) \quad (4)$$

In this study, the algorithm used as a meta-classifier in the Ensemble Classification method is a Bagging algorithm. This simple ensemble meta-algorithm learning method helps reduce variance and improve the prediction and stability of feature selection [23]. By attaching each Ensemble basic learners to a subset of instances, by size n, tasted with repetition of instances n available. As a result, base learners will have a low statistical correlation, improving the Ensemble's predictive performance [24] and discarding error change segments [25] where individual algorithm errors are compensated reciprocally [26]. Furthermore, it is based on the idea of training multiple classifiers (primary) on the same sample, and the combination of its predictions conforms to some rules for new testing objects. Thus making it possible to collectively obtain more complex models than each model separately [14].

III. Related Work

This research [1] uses the Global Terrorist Database (GTD) to learn to forecast the perpetrators of terrorist attacks and provided data on the types of attacks, targets, and weapons in addition to location, year, and other attributes using the Random Forest, Decision Tree and Gradient Boosting methods. Research [6] Presents the best machine learning models that can be used on terrorism-related data to predict the most accurate terrorist groups responsible for attacks based on historical data in India by modeling the behavior of terrorist groups using machine learning algorithms such as J48, IBK, Naive Bayes and Ensemble Voting approaches. In research [27], Create a framework for terrorist attacks that predicts the use of the Global Terrorism Database (GTD). The research approach assumes that textual features may influence the enhanced ability of classifiers to predict the types of terrorist attacks. Fitur text is extracted and represented using text representation techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (Bow), and Word Embedding (W2vec), Extracted later combined with data set features. The results showed that combining textual features with key features improved prediction accuracy significantly in research [28] using hypothesis tests and regression models. From a practice perspective, exploring the characteristics identified in patterns can lead to prevention strategies, such as changes in the physical or systemic environment. On research

[29] presents new insights into groups and target intruders using data mining algorithms by proposing a framework using historical data to train machine learning classifiers and predict intruder groups and attack types based on selected features using J48 and IBK algorithms. Research [30] uses terrorist event predictions from the Global Terrorism Database (GTD) with support vector machine (SVM), naive Bayes(NB), and linear regression(LR) techniques. Two feature selection methods, including Minimal-redundancy maximal-relevancy (mRMR) and maximum relevance (Max-Relevance), are used to improve classification accuracy. On research [31], Determine whether members of different organized crime groups cooperate using intelligence from the Canadian province of Alberta, which centers on criminals and criminal groups involved in different types of crimes in multiple locations. Bayesian techniques are used to extract multilevel network analytical frameworks and random graph models to uncover determinants of criminal collaboration between groups.

IV. Research Method

Over the past 20 years, terrorism has become a critical influencing factor in international politics and is now marked by increased terrorist attacks across international borders [30]. The increase in cases of terror crimes in the country itself is of particular concern to institutions, especially officers, because terror crimes affect the country’s stability and harm the community. Therefore, to know the list of priority scales of targets under investigation should be seen from the reasons factors that indicate the priority scale of the list of supervised members and impact the amount of security stability and comfort of the entire community.

In this study, priority scale indicator predictions were designed using discrete methods described to define the target field by studying its features to identify problems [28] using a supervised learning approach [5]. The agreements to be used are supervised classification learning, including Machine Learning Decision Tree, KNN, Naïve Bayes, and Bagging, and building a predictive model from which results are interpreted [6].

Based on Fig. 1, the first stage is to determine the background and formulation of the problem to be raised using the study of the research literature that has been done to validate the urgency of the problem raised. The next step is from the background and formulation of the above problems and then re-conducted literature studies to determine the purpose of the research and the scope of the research and deepen the sentiment analysis model that will be offered as a solution. The next step is to collect the data used from the intelligence source of the investigation process, which is a collection of datasets from each terror target located in the West Java region obtained from the data collection process taken from Raw Data CDR, Raw Data Medsos, Raw Data Surveillance, Raw Data Funding, and Raw Data BAP. The dataset consists of 1088 data with 21 attributes. Crime datasets have inherent geographic features where all data in the dataset is not distributed randomly [5]. To divide by data dimensions, we analyzed the dataset to find multiple attributes by selecting the most promising feature attributes [27], [32], potentially contributing to identifying the perpetrators [1]. From the attribute-giving techniques in this study, datasets are divided between desire patterns and abilities.

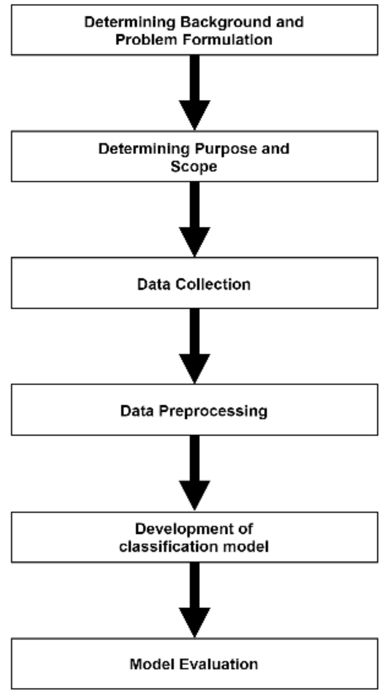

Fig. 1. The Flow of the Prediction System.

The desired stage pattern is based on the existence of intention and motivation, while the ability stage pattern consists of unique network patterns and patterns. It starts with the collection or retrieval of datasets, and then data is explored to determine inputs and outputs in dataset training and testing processes. The desire stage pattern is based on the existence of intention and motivation, while the ability stage pattern consists of unique patterns and network patterns. The attributes of these parameters are generated values from this priority scale which will be used as a result of research. The next step is to do data preparation which is done in several stages so that in the end, data can be used at the next stage. These stages include selecting and selecting data, transformation, and cleaning. The next step in building the classification will be through the stages that must be done in sequence and correctly. The steps are:

- Classification models are built for all three datasets using machine learning classification algorithms such as KNN, Naïve Bayes, and Decision Tree.
- Validation models are built for each algorithm to test base learner algorithms. The accuracy of the classifier can be further improved by using the Ensemble classifier.
- Validation models are built for each algorithm to test algorithms using Ensemble classification.
- The feature selection process aims to select subsets that can optically characterize the original data [30].

- In this study, classification algorithms were used to apply an ensemble approach and the use of attributes based on the best results of the Feature Selection process.
- Then the performance of the ensemble method is compared to the individual model.

After building a classification model, the stage that will be carried out in the evaluation stage. The stages are described below:

- Conducting a mode evaluation process using validation data tested with data testing then determines the value of Recall Rate, Precision Rate, and Accuracy Rate using confusion matrix measurement technique [27] where accuracy is single-handed, there is a significant difference between the number of Green, yellow and orange labels. Acquisition and accuracy can be used as criteria for classifier evaluation. This parameter is related to True Positive and False Positive (TP/FP), which refers to the number of positive predictions of right/wrong and True Negative and False Negative for the number of negative predictions of right/wrong (TN/FN). Confusion Matrix can measure machine learning performance in classification [11].
- Comparing the results with the other four methods (Decision Tree, KNN, Naïve Bayes, and Bagging).

A. *Data Preparation*

Data preparation will be done in several stages to obtain data used in the next stage eventually. These stages include Selection of data sampling, selection data, transformation, and cleaning with the priority scale sample dataset, and localized data system for pre-processed stages. Dataset used consists of 1088 Records with 23 Attributes. The primary purpose of the data is to find targets based on the risk of justice involvement in classification into three class groups, others referring to the green class, yellow class, and orange class. Then, perform pre-processed data and table determination. Separate data into training data and data testing 70:30 [7]. Use training data to train predictive models built on machine learning methods such as Decision Tree, Naïve Bayes, and K-Nearest Neighbour [19]. Finally, use data training to train validation models by performing k-fold cross validation processes [33] (k-fold =10) to get validation accuracy.

B. *Model Development*

Classification will go through the stages using machine learning classification algorithms such as KNN, Naïve Bayes, and Decision Tree. The Decision Tree Algorithm belongs to the family of trees used to generate decision trees. Naive Bayes is a probabilistic classifier belonging to the Bayes family. Furthermore, KNN is a lazy learning algorithm that implements the nearest K-neighbor algorithm. Validation models are built for each algorithm to test the base learner algorithm. The accuracy of the classifier can be further improved by using ensemble classifiers. Validation models are built for each algorithm to test the algorithm using Ensemble classification. The use of feature selection process because the feature aims to select a subset that can optically characterize the original data [30]. Apply the ensemble approach and attributes based on the best results of the Feature Selection process. The three classification algorithms, KNN, Naïve Bayes, and Decision Tree, were given as inputs because the algorithm results had been analyzed earlier. Then the ensemble method performance was compared to the individual model.

Acquisition and accuracy can be used as criteria for classification evaluation. This parameter is associated with True Positive and False Positive (TP/FP), which refers to the number of true positive and false positive predictions, and True Negative and False Negative for true/false-negative predictions (TN/FN). Confusion Matrix can measure machine learning performance in classification [11]. Then the validation process is done using feature selection [33] to find which attributes can be used and which are not used by testing their accuracy. Furthermore, the determination of predictive accuracy values is based on the results of the cross-validation process. The results of the Feature Selection process determine the eight methods to be used based on the score used as which input can achieve the highest accuracy value and then compare the results with four other methods (Decision Tree, Naïve Bayes, K-Nearest Neighbour, and Bagging). Then the study concentrates on comparing predictive results to test data based on the results of feature training and then testing with data testing into the form of graphs or comparison tables.

The parameters and values of this priority scale results are then used as a dataset. Table I broadly describes the data obtained.

C. *Confusion Matrix*

For evaluation, we used the Confusion Matrix, as shown in Table II, to measure the accuracy of the classifier by calculating the ratio between the correctly predicted result and the number of samples. In this study, we will measure the level of accuracy, precision, and recall.

The explanations in Table II are:

- True Positive (TP) is the sum of one TRUE class that can be correctly predicted in the TRUE class.
- True Negative (TN) is the number of one FALSE class that can be predicted correctly in the FALSE class.
- False Positive (FP) is a condition where the TRUE class whose prediction is wrong in the FALSE class, while.
- False Negative (FN) is where the conditions in the FALSE class are predicted incorrectly in the TRUE class.
- The standard formula for calculating the degree of accuracy, precision, and recall is based on the confusion matrix as shown in Equations 1-3.

$$Accuracy\ Rate = \frac{(TP+TN)}{TP+TN+FP+FN} \tag{5}$$

$$Precision\ Rate\ (p) = \frac{TP}{(TP+FP)} \tag{6}$$

$$Recall\ Rate\ (r) = \frac{TP}{(TP+FN)} \tag{7}$$

TABLE I. RESEARCH DATASET

Type Variable	Variable	Description
Attribute	Mobile phone password	There is a password (No password=0, There is a password=1)
	Mobile phone encrypted	Mobile phone starts encrypted (Unencrypted=0, Encrypted=1)
	Mobile phone off/on	The mobile phone starts to die (Mobile phone off=0, Mobile on=1
	Radical site access	Access radical sites (Not accessing radical sites =0, accessing radical sites =1)
	The meeting is getting more intense	Frequent meetings (Not attending meetings =0, Joining meetings =1)
	Get away from the network	Remove from the network / lone wolf (Not removing = 0, Removing yourself = 1)
	Counter Surveillance (SV)	Under surveillance (Unsupervised=0, Supervised =1)
	Purchase of Materials and Weapons	Making purchases of materials (Not Making illegal purchases =0, Buying illegal goods=1)
	Visiting the prison	Visiting the prison (Not visiting the prison =0, visiting the prison =1)
	Passport creation	Make a passport (Not make a passport =0, Create a passport =1)
	Withdrawal of large amounts of funds	Withdrawing vast amounts of funds (Not withdrawing funds =0, Withdrawing funds=1)
	Have the essential ability to make bombs	Has a bombing base (Has no base=0, Has a base=1)
	Personal funding capabilities	Have a permanent or non-permanent job (Unemployment=0, Work=1)
	Active Target Network	Active activities (Inactive=0, Active=1)
	Active Network Training activities	Frequent training activities (Not taking training =0, Taking training =1)
	Network Funding Capabilities	Frequently collecting network funds (Not raising funds=0, Raising funds =1)
	Permission from leader	Frequent visits to the leader of the organization (Never visited = 0, Visited = 1)
	Status background	Have a personal status background (Has no background=0, has background=1)
	Family background	Having the involvement of family members as network actors (Not having =0, Having=1)
	Group	Group network
	Region	Network Region
Label	Green	If there is no desire and no ability
	Orange	If there is a desire and no ability
	Yellow	If there is no desire and there is the ability
ID	Initials Name	The name of each individual

TABLE II. CONFUSION MATRIX

Correct Classification	Classified as	
	Positive (+)	Negative (-)
Positive (+)	True Positive (TP)	False Negative (FN)
Negative (-)	False Positive (FP)	True Negative (TN)

V. RESULT AND DISCUSSION

In this study, we researchers will explain and discuss the study's results following the methods discussed in the previous chapter. The flow in this chapter will be presented in the flowchart:

Based on Fig. 2, an analysis of business needs will be carried out so that the data mining built can meet the needs of the company's goals. Unit XYZ is also unable to determine with certainty and estimate each target managed in developing the investigation process. In practice, it is often difficult to determine the priority scale of each target. To ensure that priority scales can run effectively and efficiently, a strategy that considers the appropriate priority scale is needed to support the acceleration of handling the monitored targets. In this study, we used a dataset based on the investigation process taken from 2020 to 2021 for an investigation with Multivariate characteristics, with characteristic Attributes Polynominal and Integer, consisting of 3 Classes, 1088 Records, and 23 Attributes. Then the process using Retrieve Operator will upload an Object in the form of sample data from the Repository. After that, adding a Subprocess, this Operator that will combine other operators for the Preprocess stage will handle the Attributes of the data sample that the Retrieve Operator has loaded. The pre-processing stage is to convert nominal to numeric, normalize and replace missing values. Because there are still incorrect Attributes in writing, the MAP function maps a specific value of the selected attribute to be changed to a new value.

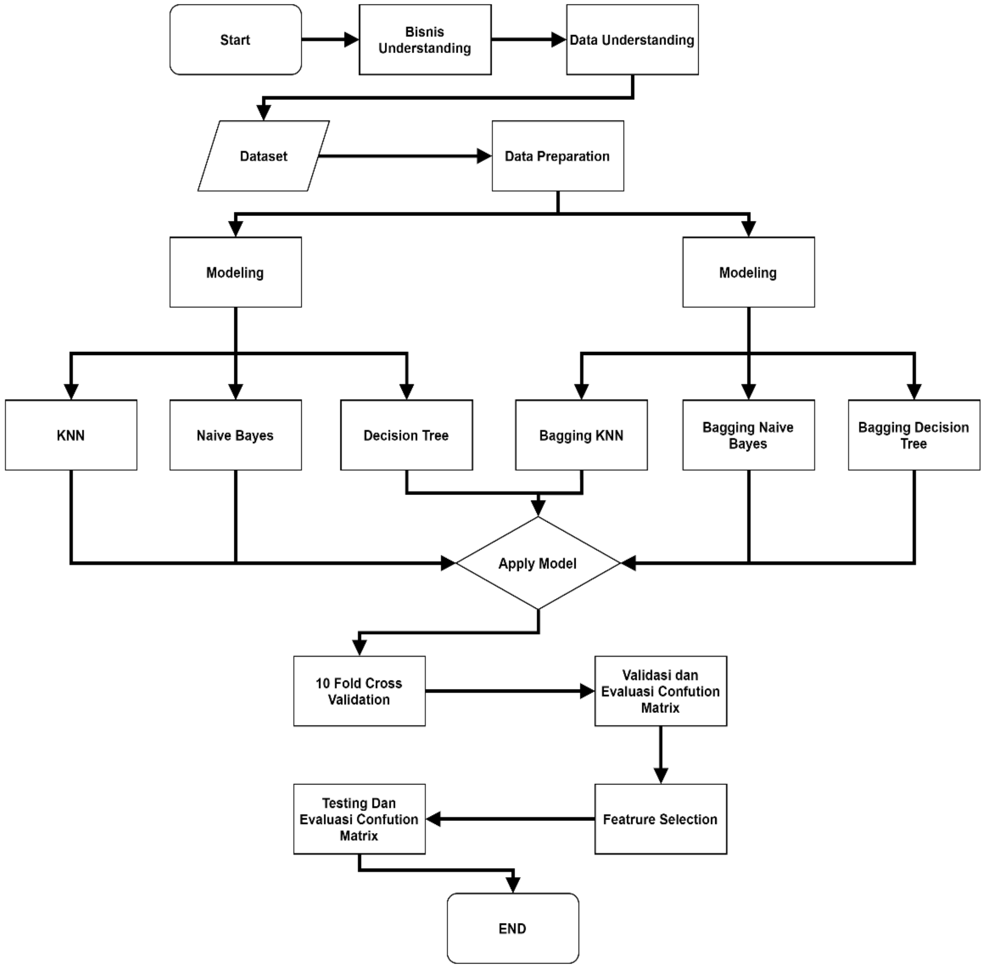

Fig. 2. Flowchart Result and Discussion.

Once the MAP process is complete, the next stage is Select Attributes. This Operator provides different filters to facilitate Attribute selection by selecting a subset of Attributes from Example Set and removing others. Only selected Attributes are sent to the output port. The rest was removed from Example Set. The next stage performs a Role Set to convert regular Attributes into Labels. This model calculation can be done by applying Set Role to turn the Priority Scale into a label. The next stage is to do the Split Data process based on the error score results. From this process, we are splitting the data 70:30. Next, separate the training data used to perform the validation test and the test data used to perform the test.

The Validation Test process tests several models using the Cross-Validation method. The accuracy of each model can be compared. The models tested were KNN, Naïve Bayes, and Decision Tree. The same Ensemble Bagging Model also uses three models to be tested: Bagging KNN, Bagging Naïve Bayes, and Bagging Decision Tree. The six models were chosen because they can process data available on RapidMiner. The validation testing process uses some training data. Validation accuracy results can be seen in Table III.

The next stage is the tuning process to get the best parameter value for feature selection and data mining. The feature selection results are carried out on the training data, consisting of 21 input variables, using all the Attributes of the Dataset. The Feature Selection process uses the Forward Selection, Backward Elimination, and Optimize Selection parameters and produces two results from the Feature Selection process. The result of the first process is the feature selection process using single learner, KNN, Naïve Bayes, and Decision Tree. While the following result is the results of Feature Selection using the Bagging model that uses the Bagging KNN model, Bagging Naïve Bayes, and Bagging Decision Tree. Feature Selection results show that the Single Learner model obtained the highest result in the Decision Tree model using the Forward Selection-Decision Tree, getting 73.24% with 6 Attributes.

TABLE III. VALIDATION ACCURACY

Validation Model	Accuracy
KNN	71.26%
Naïve Bayes	65.88%
Decision Tree	70.35%
Bagging Decision Tree	69.68%
Bagging Naïve Bayes	66.01%
Bagging KNN	71.65%

Furthermore, the highest results in the Naïve Bayes model using Backward Elimination-Naïve Bayes get 71.79% with 13 Attributes. Moreover, lastly, the highest results obtained in the KNN model using Backward Elimination-KNN get 72.45% with 19 Attributes. The results of the Feature Selection get the result that bagging models obtained the highest results in bagging decision tree models using Forward Selection Bagging-Decision Tree get results 73.89% with 6 Attributes. Furthermore, the highest results were obtained in the Bagging Naïve Bayes model using Optimize Selection Bagging-Naïve Bayes got a result of 67.08% with 9 Attributes. Moreover, lastly, the highest results obtained in the Bagging KNN model using Backward Elimination Bagging KNN get 72.30% with 20 Attributes.

As shown in Table IV, the tuning process of the Single Learner model then produces the three best performance data, as seen below:

As shown in Table V, tuning the Bagging model produces the highest accuracy results for each model judging from the performance data for the following process. From the results of accuracy taken the best of each model as seen in Table V.

The processes presented in Fig. 3 are to conduct an Accuracy Testing process with Attributes based on the results of the Tuning process. The testing process uses models such as images with tuning parameters. The dataset used as input are divided into two separate parts.

The training process uses a dataset of 762 data with attributes that have been filtered according to the tuning process. The dataset used to produce the expected model is a dataset with 326 data and Attributes according to the tuning results in the testing process. In the testing model, there are two inputs, namely mods derived from the output of the training data model and Unl in the Apply Model, which comes from the output of the filter dataset and normalization. Accuracy Testing results have two results; the first is the process of Accuracy Testing using a single learner model that uses the Backward Elimination-KNN model using 19 Attributes, Backward Elimination-Naïve Bayes model uses 13 Attributes. The Forward Selection-Decision Tree model uses 6 Attributes. The following result is accuracy testing Ensemble Classifier using Bagging model that uses Bagging-Backward Elimination-KNN model using 20 Attributes, Bagging-Optimize Selection-Naïve Bayes model uses 9 Attributes, and Bagging-Forward Selection-Decision Tree model uses 6 Attributes. The Accuracy Testing results get results that for the Single Learner model obtained results for the KNN model using Backward Elimination get a result of 69.63%, with 19 Attributes.

Next is the Naïve Bayes model using Backward Elimination which produces 39.88% with 13 Attributes. Moreover, lastly, the Decision Tree model using Forward Selection gets 65.95% with 6 Attributes. The Accuracy Testing results get results that the Ensemble-Bagging model obtained results for the Bagging-KNN model using Backward Elimination get a result of 70.25% with 20 Attributes. Furthermore, the Naïve Bayes model using Optimize Selection results in 39.88% with 9 Attributes. Moreover, lastly, the Decision Tree model using Forward Selection gets 68.10% with 6 Attributes.

The study used a data testing process tested using Feature Selection tuning validation data to see the accuracy of models in each class. By calculating the accuracy of some test data, the classification effectiveness can be known. This study uses 762 training data and 326 test data. The analysis compared and discussed three of the highest and different general classification performances, namely Backward Elimination K-Nearest Neighbor (KNN), Forward Selection Decision Tree (DT), and Elimination Naive Bayes (NB), as shown in Table VI and Table VII of the highest and other general classification Bagging performance bagging Backward Elimination K-Nearest Neighbor (KNN), Bagging Optimize Selection Naïve Bayes (NB), and Bagging Forward Selection Decision Tree (DT).

TABLE IV. SUMMARY FEATURE SELECTION SINGLE LEARNER

Backward Elimination			
	KNN	Naïve Bayes	Decision Tree
accuracy	72.45%	71.79%	71.80%
Forward Selection			
	KNN	Naïve Bayes	Decision Tree
accuracy	72.44%	66.80%	73.24%
Optimize Selection			
	KNN	Naïve Bayes	Decision Tree
accuracy	71.51%	66.80%	72.56%

TABLE V. SUMMARY FEATURE SELECTION BAGGING MODEL

Backward Elimination Bagging			
	KNN	Naïve Bayes	Decision Tree
accuracy	72.30%	66.93%	72.06%
Forward Selection Bagging			
	KNN	Naïve Bayes	Decision Tree
accuracy	72.31%	66.67%	73.89%
Optimize Selection Bagging			
	KNN	Naïve Bayes	Decision Tree
accuracy	71.78%	67.08%	73.76%

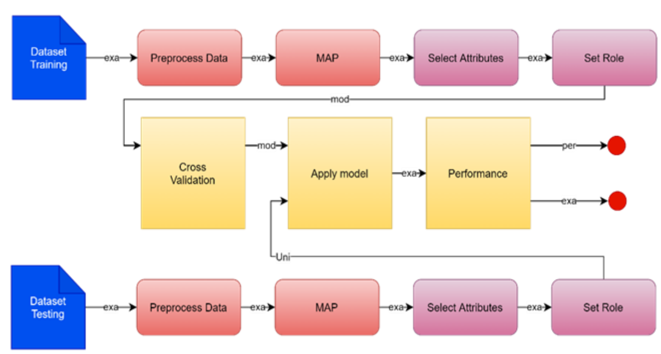

Fig. 3. Process Testing Flow.

TABLE VI. COMPARISON OF SINGLE LEARNER PREDICTION RESULTS

Predictive Actual	**Backward Elimination KNN**			**Backward Elimination NB**			**Forward Selection DT**		
	Green	Orange	Yellow	Green	Orange	Yellow	Green	Orange	Yellow
Green	1	0	0	0	0	0	1	0	0
Orange	0	60	27	1	130	193	0	66	45
Yellow	1	71	166	1	1	0	1	65	148
Total	2	131	193	2	131	193	2	131	193

TABLE VII. COMPARISON OF ENSEMBLE BAGGING PREDICTION RESULTS

Predictive Actual	**Bagging Backward Elimination KNN**			**Bagging Optimize Selection NB**			**Bagging Forward Selection DT**		
	Green	Orange	Yellow	Green	Orange	Yellow	Green	Orange	Yellow
Green	1	0	0	0	0	0	1	0	2
Orange	0	58	23	1	130	193	0	51	21
Yellow	1	73	170	1	1	0	1	80	170
Total	2	131	193	2	131	193	2	131	193

The first prediction modeling was done according to Table VI. It uses the Backward Elimination K-Nearest Neighbor method to predict an accuracy rate of 0. 6963, the precision rate is 0.7957, and the recall rate is 0.6060. The following prediction model uses the Naïve Bayes Backward Elimination method and predicted accuracy of 0.3988; the precision rate was 0.1337, and the recall rate of prediction was 0. 3308. The following prediction model used the Forward Selection Decision Tree method and obtained a prediction accuracy of 0. 6595, the precision rate is 0.7621, and the recall rate of the prediction is 0.5902. For modeling, the second prediction is made according to Table VII. They are bagging the Backward Elimination K-Nearest Neighbor method that predicts an accuracy rate of 0.7025, the precision rate is 0.8043, and the recall rate is 0.6079. The following prediction model using the Bagging-Optimize Selection method Naïve Bayes obtained a prediction accuracy of 0.3988; the precision rate was 0.1337, and the recall rate of prediction was 0. 3308. The following prediction model used the Bagging- Forward Selection Decision Tree method and obtained a prediction accuracy of 0. 6810, the precision rate is 0.5730, and the recall rate of the prediction is 0.5900.

Of the several models tested, it is known that the one with the highest accuracy is Bagging-Backward Elimination-KNN, as seen in Table VIII.

Ensemble Bagging Classification Method to try to get better predictive accuracy. In this research, it is necessary to build a predictive model using Ensemble Bagging methods such as machine learning meta-algorithms to improve classification in terms of the stability and accuracy of classifications. It also reduces variance and helps avoid overfitting. The results were higher for Ensemble Bagging testing than testing using the single learner model. Of the several models tested, it is known that the one with the highest accuracy is Bagging-Backward Elimination-KNN, as seen in Fig. 4.

TABLE VIII. EVALUATION OF PREDICTION RESULTS

Classification Model	**Accuracy**	**Precision**	**Recall**
Backward Elimination-KNN	69.63%	79.57%	60.60%
Backward Elimination-Naïve Bayes	39.88%	13.37%	33.08%
Forward Selection-Decision Tree	65.95%	76.21%	59.02%
Bagging-Forward Selection-Decision Tree	68.10%	57.30%	59.00%
Bagging-Optimize Selection-Naïve Bayes	39.88%	13.37%	33.08%
Bagging-Backward Elimination-KNN	70.25%	80.43%	60.79%

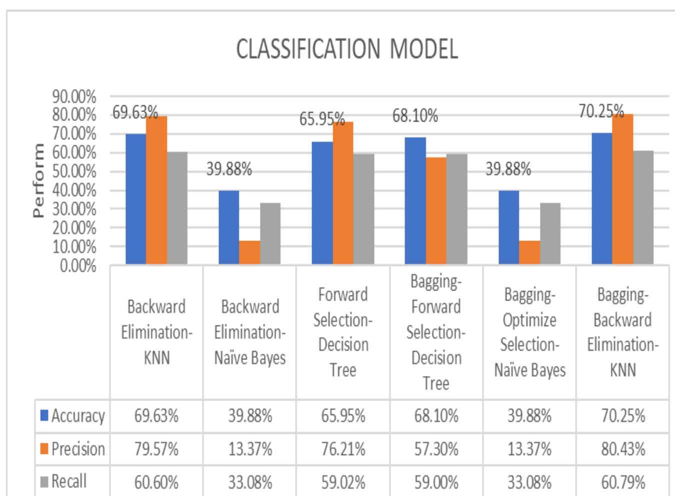

Fig. 4. Testing Performance.

VI. CONCLUSION

The experiments show how historical data or priority scale calculation patterns can be learned through data mining and generate new knowledge to predict future possibilities more accurately with other methods because the role of the priority scale is a picture of the consistency of individual targets in a behavior. The research shows that Bagging-KNN using Backward Elimination can be used, and the accuracy is 70.25%. Attributes that are considered to be no effect of Bagging KNN using Backward Elimination only one is There Is a Password. Influential attributes in determining the priority scale are Access to radical sites, Visiting prisons, Counter SV, Mobile Off, Permission from Amir, Active target network, Active Network training activities, Groups, Network Funding Capabilities, Personal funding capabilities, Family background, Background status, Removing from the network, Having basic bomb-making capabilities, Starting encrypted, Purchasing Materials and weapons, Making passports, Withdrawals in large numbers, increasingly intense Meetings and Territories. This research has discussed how historical data or priority scale calculation patterns can be learned through data mining and generate new knowledge to predict future possibilities more accurately with other methods. This study also compares the classification model of the previous ensemble research [6]. The Ensemble model is shown to improve the classification model with the research earlier model but uses an intelligence dataset instead of a dataset derived from GTD and uses the Ensemble method from previous studies using the J48, Naïve Bayes, IBK, and ensemble models using VOTE using the same model, only the VOTE Ensemble model use the same model. They are replaced using the Ensemble Bagging model. After implementing the research, it is known that the Bagging-KNN Ensemble Model using Backward Elimination can be used, and the accuracy reaches 70.25%. These results showed that the Ensemble model carried out against an existing model, Ensemble bagging KNN Backward Elimination, can increase the accuracy value by 0.62%.

Data mining with classification methods can predict the Priority Scale of each terrorist target. Using these prediction results, analysts can carry out an effective priority scale process in handling terrorism crimes so that the increase in cases of dissociation can be prevented and handled earlier. There are still many shortcomings in this study. For further work, we recommend increasing the number of variations of correlative features and large datasets so that it will help to improve the better performance of this study, namely external assessment features. In addition, more research is needed on feature selection grid methods so that each feature is more significant and very optimal for use in classification modeling.

REFERENCES

[1] M. ALfatih, C. Li, and N. E. Saadalla, “Prediction of Groups Responsible for Terrorism Attack Using Tree Based Models,” in Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Wuhan Hubei China, Jul. 2019, pp. 320–324.

[2] H. A. H. Al-Sukhni, M. M. Saudi, and A. Ahmad, “A Review of Web Classifier Approach with Possible Research Direction to Detect Islamic Terrorists,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, p. 8, 2018.

[3] M. Lim, A. Abdullah, and N. Jhanjhi, “Performance optimization of criminal network hidden link prediction model with deep reinforcement learning,” Journal of King Saud University - Computer and Information Sciences, p. S1319157819308584, Jul. 2019.

[4] S. V. Nath and S. Nath, “Crime Pattern Detection Using Data Mining,” p. 4.

[5] M. Farsi, A. Daneshkhah, A. H. Far, O. Chatrabgoun, and R. Montasari, “Crime Data Mining, Threat Analysis and Prediction,” in Cyber Criminology, H. Jahankhani, Ed. Cham: Springer International Publishing, 2018, pp. 183–202.

[6] V. T. Gundabathula, Department of Computer Science, Christ University, Bengaluru - 560029, Karnataka, India, V. Vaidhehi, and Department of Computer Science, Christ University, Bengaluru - 560029, Karnataka, India, “An Efficient Modelling of Terrorist Groups in India Using Machine Learning Algorithms,” Indian Journal of Science and Technology, vol. 11, no. 15, pp. 1–10, Apr. 2018.

[7] N. Hutagaol and S. Suharjito, “Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education,” Adv. sci. technol. eng. syst. j., vol. 4, no. 4, pp. 206–211, 2019.

[8] P. O. Oketch, “An evaluation of hybrid machine learning classifier models for identification of terorist groups in the aftermath of an attack,” p. 146.

[9] R. Sathyaraj and S. Prabu, “A hybrid approach to improve the quality of software fault prediction using Naïve Bayes and k-NN classification algorithm with ensemble method,” p. 14.

[10] I. Zafar, I. Y. Wuni, G. Q. P. Shen, S. Ahmed, and T. Yousaf, “A fuzzy synthetic evaluation analysis of time overrun risk factors in highway projects of terrorism-affected countries: the case of Pakistan,” International Journal of Construction Management, pp. 1–19, Aug. 2019.

[11] M. A. Jabbar and S. Suharjito, “Fraud Detection Call Detail Record Using Machine Learning in Telecommunications Company,” Adv. sci. technol. eng. syst. j., vol. 5, no. 4, pp. 63–69, Jul. 2020.

[12] F. Yuan, L. Lu, and Q. Zou, “Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms,” Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, vol. 1866, no. 8, p. 165822, Aug. 2020.

[13] S. R. Bandekar and C. Vijayalakshmi, “Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India,” Procedia Computer Science, vol. 172, pp. 122–127, 2020.

[14] I. Lavrov and J. Domashova, “Constructor of compositions of machine learning models for solving classification problems,” Procedia Computer Science, vol. 169, pp. 780–786, 2020.

[15] D. Gibert, C. Mateu, and J. Planes, “HYDRA: A multimodal deep learning framework for malware classification,” Computers & Security, vol. 95, p. 101873, Aug. 2020.

[16] R. Victoriano, A. Paez, and J.-A. Carrasco, “Time, space, money, and social interaction: Using machine learning to classify people’s mobility strategies through four key dimensions,” Travel Behaviour and Society, vol. 20, pp. 1–11, Jul. 2020.

[17] M. M. Mastoli, “Machine Learning Classification Algorithms for Predictive Analysis in Healthcare,” vol. 06, no. 12, p. 5, 2019.

[18] C. Frank, A. Habach, R. Seetan, and A. Wahbeh, “Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis,” Adv. sci. technol. eng. syst. j., vol. 3, no. 2, pp. 184–189, Mar. 2018.

[19] Y. C. Widiyono and S. M. Isa, “Utilization of Data Mining to Predict Non-Performing Loan,” Adv. sci. technol. eng. syst. j., vol. 5, no. 4, pp. 252–256, 2020.

[20] F. Lopes, J. Agnelo, C. A. Teixeira, N. Laranjeiro, and J. Bernardino, “Automating orthogonal defect classification using machine learning algorithms,” Future Generation Computer Systems, vol. 102, pp. 932–947, Jan. 2020.

[21] G. N. Srivastava, A. S. Malwe, A. K. Sharma, V. Shastri, K. Hibare, and V. K. Sharma, “Molib: A machine learning based classification tool for the prediction of biofilm inhibitory molecules,” Genomics, vol. 112, no. 4, pp. 2823–2832, Jul. 2020.

[22] U. Pujianto, T. Widiyaningtyas, D. D. Prasetya, and B. Romadhon, “Penerapan algoritma naïve bayes classifier untuk klasifikasi judul

skripsi dan tugas akhir berdasarkan Kelompok Bidang Keahlian," TEKNO, vol. 27, no. 1, p. 79, Jul. 2019.

[23] R. Siva Subramanian and D. Prabha, "Customer behavior analysis using Naive Bayes with bagging homogeneous feature selection approach," J Ambient Intell Human Comput, vol. 12, no. 5, pp. 5105–5116, May 2021.

[24] V. G. T. da Costa, S. M. Mastelini, A. C. P. de L. F. de Carvalho, and S. Barbon, "Making Data Stream Classification Tree-Based Ensembles Lighter," in 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, Brazil, Oct. 2018, pp. 480–485.

[25] S. Kabiraj, L. Akter, M. Raihan, N. J. Diba, E. Podder, and Md. M. Hassan, "Prediction of Recurrence and Non-recurrence Events of Breast Cancer using Bagging Algorithm," in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, Jul. 2020, pp. 1–5.

[26] R. O. Alabi et al., "Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer," International Journal of Medical Informatics, vol. 136, p. 104068, Apr. 2020.

[27] M. Abdalsalam, C. Li, A. Dahou, and S. Noor, "A Study of the Effects of Textual Features on Prediction of Terrorism Attacks in GTD Dataset," vol. 29, no. 2, p. 16, 2021.

[28] G. Singer and M. Golan, "Identification of subgroups of terror attacks with shared characteristics for the purpose of preventing mass-casualty attacks: a data-mining approach," Crime Sci, vol. 8, no. 1, p. 14, Dec. 2019.

[29] S. Nazir, M. A. Ghazanfar, N. R. Aljohani, M. A. Azam, and J. S. Alowibdi, "Data analysis to uncover intruder attacks using data mining techniques," in 2017 5th International Conference on Information and Communication Technology (ICoIC7), Melaka, Malaysia, May 2017, pp. 1–6.

[30] H. Mo, X. Meng, J. Li, and S. Zhao, "Terrorist event prediction based on revealing data," in 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(, Beijing, China, Mar. 2017, pp. 239–244.

[31] J. A. Coutinho, T. Diviák, D. Bright, and J. Koskinen, "Multilevel determinants of collaboration between organised criminal groups," Social Networks, vol. 63, pp. 56–69, Oct. 2020.

[32] F. N. Yacoub, M. Mamdouh, K. F. Kassem, M. Torki, and M. S. Abougabal, "Towards Terrorist Groups Prediction in Middle East and North Africa," p. 7.

[33] O. Somantri, S. Wiyono, and D. Dairoh, "K-Means Method for Optimization of Student Final Project Theme Classification Using Support Vector Machine (SVM)," SJI, vol. 3, no. 1, pp. 34–45, Jun. 2016.