INTENSIF, Vol.7 No.1 February 2023
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v7i1.18283

Comparing Data Mining Classification for Online Fraud Victim Profile in Indonesia

Received:
7 July 2022
Accepted:
1 January 2023
Published:
10 February 2023

[1]Sunardi, [2]Abdul Fadlil, [3*]Nur Makkie Perdana Kusuma
[1,2] Departement of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3] Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
E-mail: [1]sunardi@mti.uad.ac.id, [2]fadlil@uad.ac.id, [3]nur2008048034@webmail.uad.ac.id
*Corresponding Author

Abstract— Classification is one of the most often employed data mining techniques. It focuses on developing a classification model or function, also known as a classifier, and predicting the class of objects whose class label is unknown. Categorizing applications include pattern recognition, medical diagnosis, identifying weaknesses in organizational systems, and classifying changes in the financial markets. The objectives of this study are to develop a profile of a victim of online fraud and to contrast the approaches frequently used in data mining for classification based on Accuracy, Classification Error, Precision, and Recall. The survey was conducted using Google Forms, which is an online platform. Naive Bayes, Decision Tree, and Random Forest algorithms are popular models for classification in data mining. Based on the sociodemographics of Indonesia's online crime victims, these models are used to classify and predict. The result shows that Naïve Bayes and Decision Tree are slightly superior to the Random Forest Model. Naive Bayes and Decision Tree have an accuracy value of 77.3%, while Random Forest values 76.8%.

Keywords— Data Mining; Online Fraud Victims' Profile; Naïve Bayes; Random Forest

Corresponding Author:

Nur Makkie Perdana Kusuma,
Master Program of Informarmatics,
Universitas Ahmad Dahlan, Yogyakarta,
Email: nur2008048034@webmail.uad.ac.id

I. INTRODUCTION

Cyber fraud is a crime that uses the Internet for business and trade purposes. It no longer relies on the business of an honest conventional company. In principle, internet fraud is the same as traditional fraud. The only difference is in action, namely using electronic systems (computers, Internet, telecommunication devices). Online fraud is included in the crime group of abuse of information technology in Computer Related Fraud[1]–[3].

The emergence of cybercrime forms rules in cyberspace, known as cyber law. Cyberlaw summarizes the laws related to Internet use, which vary by jurisdiction from country to country. In Indonesia, cyberlaw is listed in the Law of the Republic of Indonesia Number 11 of 2008, which regulates Information and Electronic Transactions. This regulation has been applied since the emergence of online crime cases in Indonesia.

Cybercrime cases that often occur in Indonesia are cases of online fraud. This crime occurred by utilizing Instant Messenger (IM). Sunardi[3] stated that online fraud cases in Indonesia were dominated by Instagram and WhatsApp social media. Along with the many attacks from cybercriminals, several software companies, especially anti-virus and malware makers, are always trying to cover gaps in the existing system such as two-step verification facilities (Google Mail, Yahoo Mail, Steam, and others).

The increase in online fraud cases in Indonesia is not caused by the application's lax security measures but by its users. In Indonesia, online account security is considered something that is not important, either e-mail accounts or other social media accounts, seen from passwords that are not unique and not secure. These gaps can be a loophole for cybercriminals, especially with social engineering.

Social engineering has not received special attention from internet users in Indonesia. Phishing is one of the social engineering techniques that IM users do not understand. According to Marshal[4], in the evaluation model for online crime, there are six factors can estimate the likelihood value of an attack from a cyber-criminal. One of the six factors is Victim Expertise (Ve), which shows how much expertise or knowledge the victim has. The smaller the value of Ve, the value of the possibility to be attacked becomes greater.

Authorities will be able to focus more intently on investigating other sources by focusing their search by narrowing it with the aid of profiling hackers. Cyber profiling of cybercrime perpetrators has been widely implemented[5]–[7]. It creates profile information of cybercriminals, such as the characteristics of the perpetrators, motives, backgrounds, behaviour patterns, and sociodemographic data based on cybercrimes[8]–[10]. Profiling of victims of cybercrime is intended to facilitate the targeting of information dissemination and carry out prevention efforts[11]–[13].

Cyber-attacks can happen and succeed due to a lack of security awareness. In order to secure people and businesses, it is necessary to assess cyber-security knowledge. Recent research has focused on how well-informed Indonesians are about cybercrime, particularly mobile malware and cloud computing, with publications released between 2012 and 2021. On the other hand, existing research has several gaps, such as 1). Existing research focuses on a particular place, group, or background, which does not reflect national understanding. 2) Theoretical models and the method of computing the sample size of the individuals have not been explored in current approaches. 3) The number of people who utilize the Internet has skyrocketed.

The preceding list demonstrates a significant research gap in Indonesia when it comes to measuring cybercrime awareness. It is critical to perform a survey to assess awareness using a theoretical model and recruit active technology users from various backgrounds and geographies, with a sufficient sample size[14].

There are several techniques to assess a person's knowledge of cybercrime[15]. The following important factors are the focus the research[16], [17]: 1) The subject's Internet usage habits and how he or she interacts with technology on digital gadgets. 2) Identifying current patterns and how individuals deal with everyday security activities and what they think about cybercrime. 3) Observing how people behave when confronted with (or will be confronted with) a cyber-crime occurrence.

Saroha[18] states that profiling cyber criminals will help the authorities narrow the search scope and focus on searching other available sources intensively. Technology is indeed the primary defence against cyber-attacks. A better understanding of the psychological, criminological, and sociological aspects can provide input on protection efforts and catch cyber criminals before the distance gets further.

Innab, Al-Rashoud, Al-Mahawes et al.[19] investigated the present state of phishing email knowledge and training among 116 personnel in Riyadh's government and business sectors. This study mainly focused on Saudi Arabians who had not worked in the IT field. Employee demographics, administrative data, employee understanding of the idea of email phishing, and awareness of the organization's phishing protection were all included in the survey. Because awareness and anti-phishing training were found to be at an all-time low, the study concluded that increasing awareness through anti-phishing training programs is critical. Employees, in particular, should get proper electronic email phishing awareness training, as email is the most convenient way to execute phishing assaults.

Reep-van den Bergh & Junger[20] researched in Europe by categorizing six types of cybercrimes such as online shopping fraud, online fraud banking/payment, other cyber fraud (such as advanced fee fraud), cyber threats/harassment, malware, and hacking. This study analyses the percentage based on the number of victims of cybercrimes that occurred in Europe based on six categories in the period from 2009 to 2016.

Alotaibi, Furnell, Stengel et al.[21] used a quantitative online-based survey with 629 individuals to investigate cyber security knowledge among Saudi people (70% male, 30% female). Although the participants had strong IT expertise, the study revealed that their understanding of cybercrime, cyber security measures, and the role of government and companies in maintaining the integrity of online information were limited.

Zayid & Farah[22] surveyed 132 undergraduate students with an information technology background in the Alnamas area of Saudi Arabia, finding that 15% of the participants had experienced a cybercrime, 80.7% were interested in receiving training to improve their knowledge, and 69.6% of cybercrimes occurred through social media, with 57% of them occurring through Facebook.

By creating an online questionnaire and then disseminating it among 161 subjects, Arachchilage & Love[23], conducted a study to determine whether conceptual or procedural knowledge positively affects computer awareness. They found that positive effects were obtained when they applied both conceptual and procedural knowledge to prevent further phishing risks.

Abawajy[24] investigated information security awareness distribution options, such as interactive videos, internal training session courses, screen savers, emails, and social media, to increase end-user knowledge and behaviour, specifically with phishing attacks. They looked at the effectiveness of text-based, game-based, and video-based techniques. The trials involved 60 people showed that video was the most popular technique, followed by texting. Ahmed, Kulsum, Azad et.al.[25] employed online and offline questionnaires of Bangladeshi citizens to assess their cyber-security awareness. According to the report, the degree of knowledge is insufficient. A significant majority of individuals are ignorant of conventional cyber security standards.

Alzubaidi[16] conducted a study that focused on measuring the level of cyber security awareness in Saudi Arabia, in terms of cyber security, awareness level, and incident reporting, through an online questionnaire with 1230 participants. The questionnaire results showed that 31.7% used public Wi-Fi to access the Internet, 51% used their personal information to create passwords, 32.5% did not know about phishing attacks, 21.7% had been victims of cybercrimes. In comparison, only 29.2% of them reported a crime, reflecting their level of awareness.

One of the most often used data mining methods is classification, which focuses on creating a classification model or function, also known as a classifier, and predicting the class of objects

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v7i1.18283

whose class label is unknown. Pattern identification, medical diagnosis, spotting flaws in business systems, and categorizing financial market movements are a few examples of categorization applications[26][27].

For the classification experiments, RapidMiner has been used to classify the product using Naïve Bayes, Decision Tree, and Random Forest. RapidMiner is a data science software platform capable of integrating data preparation, machine learning, deep learning, text mining, and predictive analytics into a single environment. The software supports all aspects of the machine learning process, including data preparation, result visualization, model validation, and optimization, and is used for business and commercial applications as well as research, education, training, rapid prototyping, and application development. RapidMiner is based on an open-source architecture[28], [29].

Rapidminer is a comprehensive software with visual workflow design and full automation so there is no need to do any coding for data mining tasks. In evaluating the performance of the clustering algorithm in profiling; the percentage of accuracy, precision, and recall rate of the classification was measured using RapidMiner software. RapidMiner is a data mining platform that enables focuses on machine learning and data mining[30]–[32]. In this study, RapidMiner is used to classify and predict data using the Naïve Bayes, Decision Tree, and Random Forest.

The Naive Bayes is a probabilistic machine learning model that is used for classification tasks based on the Bayes theorem. Naïve Bayes is a supervised learning algorithm based on Bayes theorem which is used to solve classification problems by following a probabilistic approach[33]. Naïve Bayes was put forward by the British scientist Thomas Bayes, which predicts future opportunities based on previous experience so it is known as Bayes' theorem[34], [35]. The Random Forest has a large number of decision trees that can do classification independently, and the most voted class is regarded as the model's forecast. The accuracy of random forest can be increased since it uses the categorization capability of multiple trees. The crucial point, however, is to create low-correlated decision trees within the random forest. Otherwise, the errors of the separate decision trees can accumulate and cause incorrect classification[17], [33], [36]. In machine learning, image processing, and pattern recognition, Decision Tree is one of the most powerful techniques that are frequently used. Conceptual rules are considerably simpler to create when building a neural network of connections between nodes than numerical weights. Data mining frequently uses the Decision Tree classification model. Each tree is made up of branches and nodes. Each node represents a feature in the classification category, and each subset specifies a value the node may accept[37]–[39].

Many studies use naïve Bayes algorithms, decision trees, and random forests as models for classification or profiling[40]–[42]. And a lot of studies compare the performance of the three models. The three algorithms proved capable of performing their duties as a classification model. these studies do not have the same conclusion as to which is the best of the three models. The capabilities of the three models are based on the type of data used.

Kusumarini's[43] research states that the random forest algorithm outperforms naïve Bayes and decision trees based on accuracy. In contrast to Shamala's[27] research, the Decision Tree is the best algorithm for profiling bank customer data in terms of accuracy. Still, random forest excels in precision, and Naïve Bayes excels in recall percentage. However, in Putu's[44] research, Naïve Bayes is superior in terms of precision.

This study aims to create an online fraud victim profile and to understand which method is the best model for cybercrime profiling based on accuracy, classification error, precision, and recall. The classification algorithms are Naive Bayes, Decision Tree, and Random Forest. These models are applied to classify and predict based on the sociodemographic of online crime victims in Indonesia[3], [29]. This paper is organized into 4 sections, where "Introduction" section contains an introduction and the Literature review of previous research papers and studies conducted by other researchers worldwide. "Research Method" section comprises of Methodology that has been adopted in this paper. The "result and discussion" section describes the findings obtained based on the analysis using the method used in this paper and provides a discussion of these results. And the last part is the Conclusion section.

II. RESEARCH METHOD

There are five stages in this research, as shown in Figure 1. The study begins with data collection, data cleaning, analysis for profiling using the Naïve Bayes, Decision Tree, and Random Forest models, comparing the performance of the three models, and drawing conclusions based on the research results.

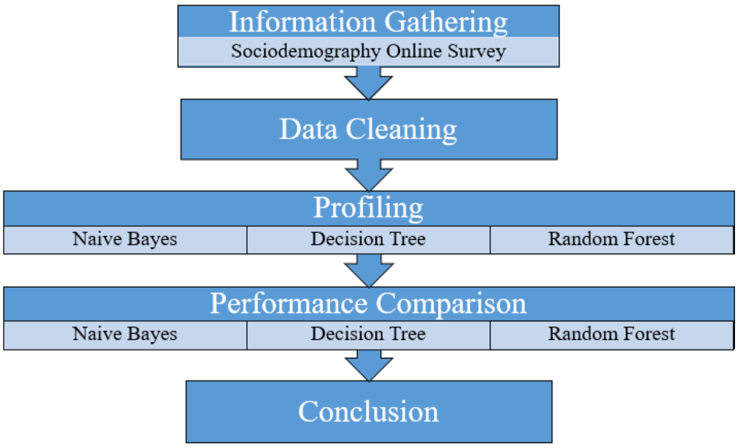


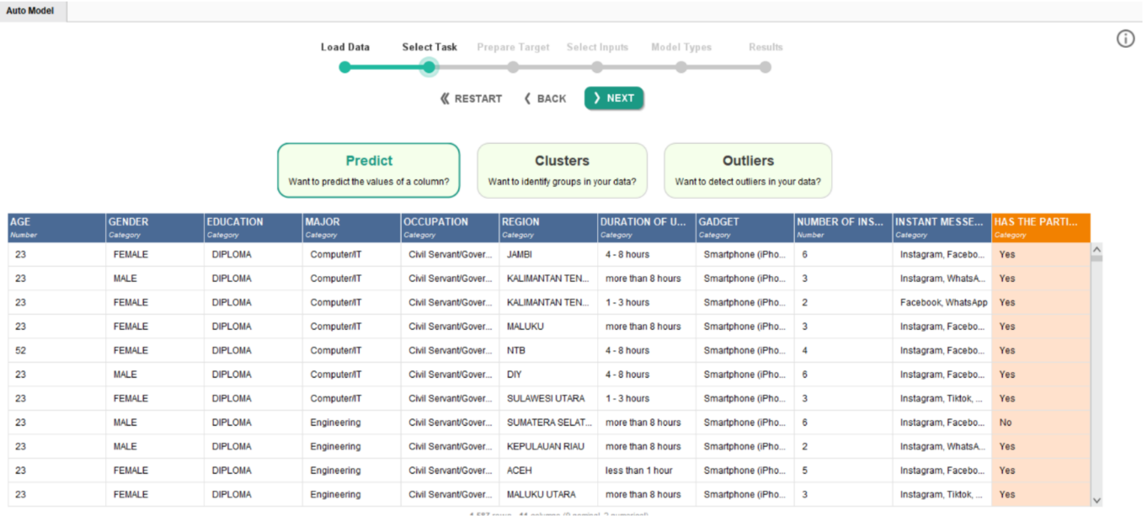
Figure 1. RESEARCH FLOWCHART

The first step in this research is data collection using an online questionnaire. The survey was conducted using Google Forms, which is an online platform. The findings were recorded in a local database for later analysis, and anonymous participation was permitted to protect data confidentiality. Participants were asked if they agreed or disagreed with taking part in the survey during the pre-processing phase. If they approve, they might access the questionnaire by connecting into Google forms with their Google accounts. They were only given one opportunity to submit their responses. Following this, all replies saved to the local hard drive to be processed, and the findings could be further analysed.

This participants in this research are Indonesian IM users who's currently active using the internet. Participants were chosen with simple random sampling for this experiment from December 2021 to March 2022. Participants were given a link to a Google Form to fill out survey regarding the sociodemographic characteristic of online fraud victims. Snowball sampling techniques were used to identify research subjects. As a result, the number of data source samples will larger, similar to how a snowball grows larger over time. The snowball sampling approach is picked so, if data from one source is still missing, we may rely on information from other sources.

The next step is data cleaning. by using the features of RapidMiner, tabulated data is cleaned so that it meets the requirements to enter the classification stage. From 1982 data from online surveys, 295 data were found that could not be processed due to missing values. Meanwhile, all socio-demographic variables can be used to analyse the classification model based on the category.

At the third stage, RapidMiner provides convenience with the presence of an AutoModel which automatically conducts data training (data training) along with data analysis. As much as 60% of the total data will automatically be used as training data. Figure 1 shows the initial steps in conducting the analysis process with RapidMiner. In this step, the researcher uses the attribute [HAS THE PARTICIPANT EVER EXPERIENCED ONLINE CRIME?] as a label to predict other correlated attributes. In addition to analysing the Naïve Bayes, Decision Tree, and Random Forest model, RapidMiner is also used to perform text mining for the Instant Messenger attributes [21], [22], [32].


Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT
Predict
Want to predict the values of a column?
Clusters
Want to identify groups in your data?
Outliers
Want to detect outliers in your data?

AGE Number	GENDER Category	EDUCATION Category	MAJOR Category	OCCUPATION Category	REGION Category	DURATION OF U... Category	GADGET Category	NUMBER OF INS... Number	INSTANT MESSE... Category	HAS THE PARTI... Category
23	FEMALE	DIPLOMA	Computer/IT	Civil Servant/Gover...	JAMBI	4 - 8 hours	Smartphone (iPho...	6	Instagram, Facebo...	Yes
23	MALE	DIPLOMA	Computer/IT	Civil Servant/Gover...	KALIMANTAN TEN...	more than 8 hours	Smartphone (iPho...	3	Instagram, WhatsA...	Yes
23	FEMALE	DIPLOMA	Computer/IT	Civil Servant/Gover...	KALIMANTAN TEN...	1 - 3 hours	Smartphone (iPho...	2	Facebook, WhatsApp	Yes
23	FEMALE	DIPLOMA	Computer/IT	Civil Servant/Gover...	MALUKU	more than 8 hours	Smartphone (iPho...	3	Instagram, Facebo...	Yes
52	FEMALE	DIPLOMA	Computer/IT	Civil Servant/Gover...	NTB	4 - 8 hours	Smartphone (iPho...	4	Instagram, Facebo...	Yes
23	MALE	DIPLOMA	Computer/IT	Civil Servant/Gover...	DIY	4 - 8 hours	Smartphone (iPho...	6	Instagram, Facebo...	Yes
23	FEMALE	DIPLOMA	Computer/IT	Civil Servant/Gover...	SULAWESI UTARA	1 - 3 hours	Smartphone (iPho...	3	Instagram, Tiktok, ...	Yes
23	MALE	DIPLOMA	Engineering	Civil Servant/Gover...	SUMATERA SELAT...	more than 8 hours	Smartphone (iPho...	6	Instagram, Facebo...	No
23	MALE	DIPLOMA	Engineering	Civil Servant/Gover...	KEPULAUAN RIAU	more than 8 hours	Smartphone (iPho...	2	Instagram, WhatsA...	Yes
23	FEMALE	DIPLOMA	Engineering	Civil Servant/Gover...	ACEH	less than 1 hour	Smartphone (iPho...	5	Instagram, Facebo...	Yes
23	FEMALE	DIPLOMA	Engineering	Civil Servant/Gover...	MALUKU UTARA	more than 8 hours	Smartphone (iPho...	3	Instagram, Tiktok, ...	Yes

Figure 2. AUTOMODEL USING RAPIDMINER

For the fourth stage, the three classification models will be compared based on the confusion matrix obtained from the results of each model. Formulas 1 to 4 show how to calculate performance based on the Accuracy, Precision, Classification Error, and Recall. Calculations use True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values obtained from each of the three models' confusion matrix tables.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

$$Classification\ Error = \frac{FP+FN}{TP+TN+FN+FP} \qquad (3)$$

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

III. RESULT AND DISCUSSION

Based on the survey, there are a total 1587 of participants’ data that can be processed. First they are cleaned to remove inappropriate or missing data (missing value). There are 1220 respondents were found to have been victims of online fraud via IM and 367 had never been victims of online fraud. The characteristics are divided into age, gender, education, occupation, regions, duration of using the internet in a day, the number of IM installed, gadgets, and IM media

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v7i1.18283

used. The data was then processed using the RapidMiner Auto Model and analysed using the Naïve Bayes, Decision Tree, and Random Forest models.

This part divides the data description into sections: descriptive data from participants' sociodemographic and analysis data from the classification models. Figure 2 indicates that females are more victims of online fraud than males. It clearly shows that respondents who were victims of online fraud were primarily women, with 726 participants.

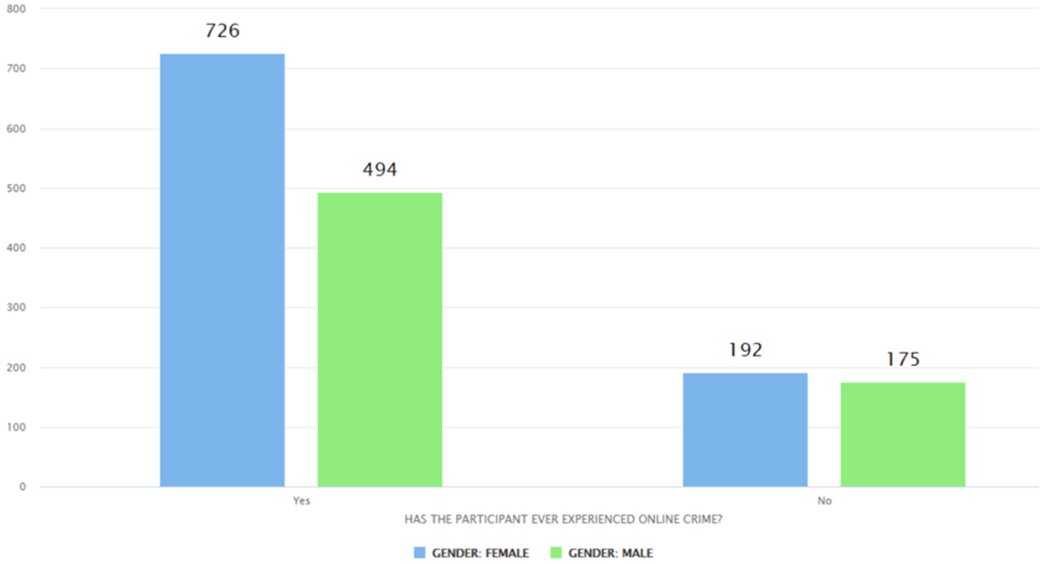


Figure 3. COMPARISON OF THE NUMBER OF PARTICIPANTS BASED ON WHETHER OR NOT THEY HAVE BEEN VICTIMS OF ONLINE FRAUD GROUPED BY GENDER

Regarding the sociodemographic data description based on the age of the respondents, Figure 3 shows that the youngest respondent is 17 years old, while the oldest is 73. It shows that the age of participants who became victims of online crimes spread from the youngest age of 17 years, and the oldest of 73 years. Figure 3 displays the percentages for each age range. The data shows that the most dominant age is in the range of 23 to 28 years with 35%. Followed by 17 to 22 years old with 33%, 29 to 33 years with 13%.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v7i1.18283

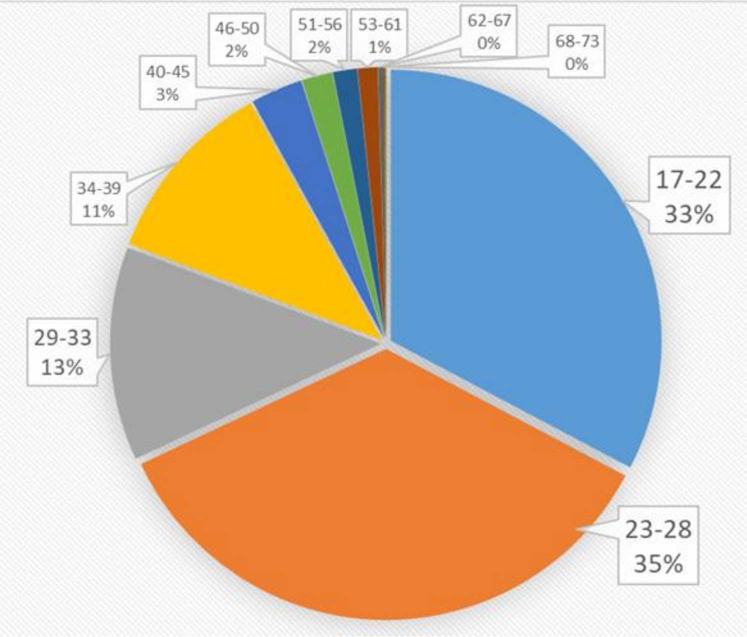


Figure 4. PERCENTAGE OF AGE

The education level responses were distributed as follows: 517 with a High School Diploma, 350 with or working toward a Bachelor/Undergraduate Degree, 247 with an Associate's Degree/Diploma, and 106 with a Postgraduate Degree. Table 1 shows the description along with the percentage.

Table 1. PARTICIPANTS' SOCIODEMOGRAPHIC BASED ON EDUCATION

Value	Count	Percentage (%)
High School Graduate	517	42.38
Bachelor's Degree/Under	350	28.69
Associate's Degree/Diploma	247	20.25
Post Graduate	106	8.69

Figure 4 displays the subjects from all regions of Indonesia; the highest data is Special Region of Yogyakarta with 260, followed by Central Java Province with 155. It can also be seen in Figure 4 that the lowest is Banten. The data of participants' sociodemographic based on regions in Indonesia is distributed well in 34 provinces.

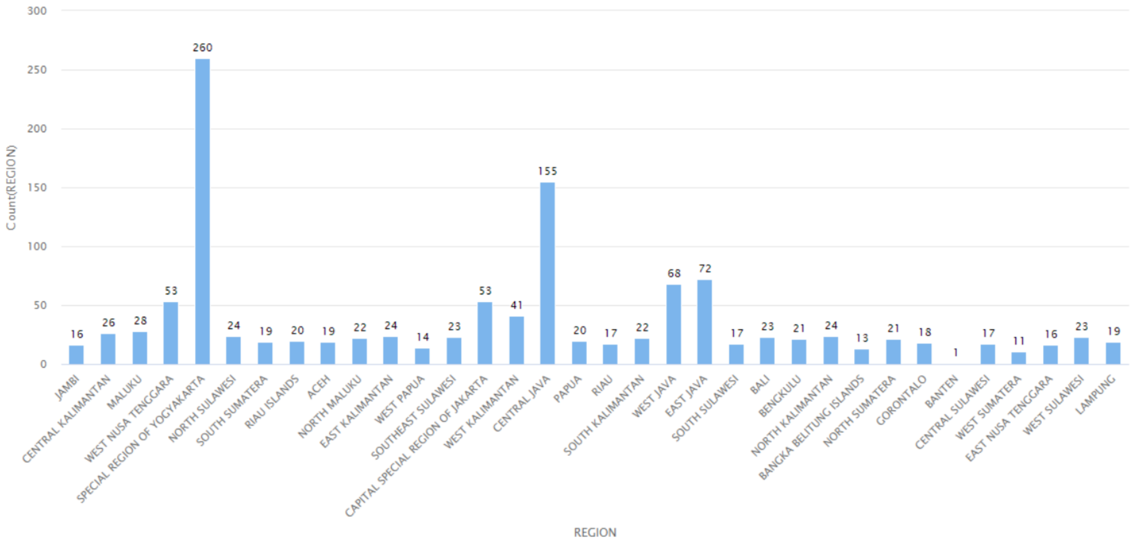


Figure 5. PARTICIPANTS' SOCIODEMOGRAPHIC BASED ON REGIONS

The majors of the field study are Management, Computer Science/IT, Engineering, Social Science, Education, Health and Medicine, and Language/Literature. Table 2 shows the responses were distributed into 537 in Management, 483 chose Others, 69 in Computer Science/IT, 53 in Engineering, 46 in Social Science, 28 Education majors, 3 Medicine and Health majors, and 1 in Language/Literature major.

Table 2. PARTICIPANTS' SOCIODEMOGRAPHY BASED ON MAJOR

Value	Count	Percentage (%)
Management	537	44.02
Other	483	39.59
Computer/IT	69	5.66
Engineering	53	4.34
Social Science	46	3.77
Education	28	2.29
Health/Medicine	3	0.25
Language/Literature	1	0.08

The subjects were asked about their occupation. Table 3 shows the occupation distributed among 607 Student/College Students, 316 Private Employees, 113 in Civil Servant/Government, 97 Entrepreneur, 55 Teacher/Lecturer/Instructor, and 32 others. It shows that most online fraud victims' sociodemographic based on occupation is Student/College Student.

Table 3. PARTICIPANTS' SOCIODEMOGRAPHY BASED ON OCCUPATION

Value	Count	Percentage (%)
Student/College Student	607	49.75
Private Employee	316	25.90
Civil Servant/Government	113	9.26
Entrepreneur	97	7.95
Teacher/Lecturer/Instructor	55	4.51
Others	32	2.62

We asked subjects about how often they access the Internet in a day. The answers were distributed to 398 (30,45%) accessing the Internet for more than 8 hours, 417 (31.91%) for 4 to 8 hours, 249 (19.81%) accessing the Internet for 1 to 3 hours, and 146 (11.17%) for less than an hour. The majority of online fraud victims' sociodemographic based on spending time on the Internet is 4 to 8 hours a day. These values are represented in Table 4.

Table 4. PARTICIPANTS' SOCIODEMOGRAPHY BASED ON DURATION OF USING INTERNET

Value	Count	Percentage (%)
More than 8 hours	398	30.45
4 – 8 hours	417	31.91
1 – 3 hours	259	19.82
Less than 1 hour	146	11.17

The sociodemographic section on statistics ends with a discussion of the duration of internet use in a day. The following discussion is the result of the Naïve Bayes, Decision Tree, and Random Forest. This study uses participants' data as a reference in making prediction models using Naïve Bayes, Decision Tree, and Random Forest. RapidMiner has the advantage of the AutoModel feature so that more than one models can be run simultaneously. Table 5 shows the results of online fraud victim profile.

Table 5. ONLINE FRAUD VICTIM PROFILE

Attribute	Prediction
Age	26.0
Number of Instant Messengers Owned	3.7
Gender	Female
Education	High School
Major	Management
Occupation	Student/College Student
Region	Special Region of Yogyakarta
Duration of Using the Internet in One Day	More than 8 hours
Gadget	Smartphone (iPhone/Android)
Instant Messenger	Instagram, Facebook, WhatsApp

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v7i1.18283

The results of a survey of 1220 participants who have been victims of online fraud using text mining with RapidMiner are shown in Table 6. Social media that victims often use are Instagram and WhatsApp. Table 7 shows that victims often use a gadget like a smartphone, regardless of whether it is an Android or Smartphone.

Table 6. INSTANT MESSENGER THAT OFTEN USED BY ONLINE FRAUD VICTIM

Instant Messenger	**Number of Instant Messenger used by Victim**
Instagram	699
Whatsapp	691
Facebook	483
Telegram	339
Tiktok	217
Twitter	141
Line	96
Others	21
Michat	17
Snapchat	4

Table 7. GADGET THAT OFTEN USED BY ONLINE FRAUD VICTIM

Gadget	**Number of Gadegt used by Victim**
Smartphone (android/iphone)	726
Laptop	157
Computer	36
Tablet	21
Others	10
Android	726

Tables 8, 9, and 10 show the Confusion Matrix results for each model. This Confusion Matrix can calculate accuracy, precision, and recall with the Formula (1) for Accuracy, Formula (2) for Precision, Formula (3) for Classification Error, and Formula (4) for Recall.

Table 8. CONFUSION MATRIX MEASUREMENT ON NAÏVE BAYES MODEL

Actual Class	**Prediction Class**	
	Positive	**Negative**
Positive	350 (TP)	0 (FN)
Negative	103 (FP)	0 (TN)

Table 9. CONFUSION MATRIX MEASUREMENT ON DECISION TREE MODEL

Actual Class	**Prediction Class**	
	Positive	**Negative**
Positive	350 (TP)	0 (FN)
Negative	103 (FP)	0 (TN)

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v7i1.18283

Table 10. CONFUSION MATRIX MEASUREMENT ON RANDOM FOREST MODEL

Actual Class	Prediction Class	
	Positive	Negative
Positive	348 (TP)	0 (FN)
Negative	103 (FP)	0 (TN)

Regarding the performance of the three classification models, Table 11 shows the comparisons between Naïve Bayes, Decision Tree, and Random Forest Model by Accuracy, Classification Error, Precision, and Recall. The comparison of the performance of the three algorithms for classification in data mining in Table 11 shows the implementation of the three models for profiling online fraud victims in this study.

Performance is measured by the percentage of Accuracy, Classification Error, Precision, and Recall. Based on Table 11, it can be said that Naïve Bayes and Decision Tree are slightly superior to the Random Forest Model. Naive Bayes and Decision Tree have an accuracy value of 77.3%, while Random Forest values 76.8%. Table 10 shows that the True Positive values in the random forest model are fewer than the True Positive values in the Naïve Bayes model and the Decision Tree. This outcome indicates that the Random Forest model with default mode from RapidMiner cannot recognize "YES" data from the profiling data set in this study. There is a difference of two points compared to the results of testing by the other two models. Even so, with an accuracy value of around 77%, the three models are can create profiles of online crime victims in Indonesia.

Table 11. COMPARISONS BETWEEN NAÏVE BAYES, DECISION TREE, AND RANDOM FOREST MODEL

Model	Performance			
	Accuracy	Classification Error	Precision	Recall
Naïve Bayes	77.3%	23.7%	77.3%	100%
Decision Tree	77.3%	23.7%	77.3%	100%
Random Forest	76.8%	23.2%	77.2%	99.4%

IV. CONCLUSION

In this study, it can be concluded that the three models: Naïve Bayes, Decision, and Random Forest, can classify and predict online fraud victims in Indonesia based on sociodemographic data from participants. The models are compared based on accuracy, classification error, precision, and recall indicators. The results from RapidMiner show that the Naive Bayes and Decision Tree model is slightly more significant than the Random Forest model. Naive Bayes and Decision Tree have an Accuracy percentage of 77.3% and a Recall percentage of 100%. In comparison, Random Forest has an Accuracy percentage of 76.8% and a Precision percentage of 99.4%.

REFERENCES

[1] A. A. Gillespie and S. Magor, "Tackling online fraud," ERA Forum, vol. 20, no. 3, pp. 439–454, 2020, doi: 10.1007/s12027-019-00580-y.

[2] N. P. Singh, "Online Frauds in Banks with Phishing," J. Internet Bank. Commer., vol. 12, no. 2, pp. 1–28, 2007, [Online]. Available: http://eprints.utm.my/8136/.

[3] Sunardi, A. Fadlil, and N. M. P. Kusuma, "Implementasi Data Mining dengan Algoritma Naïve Bayes untuk Profiling Korban Penipuan Online di Indonesia," vol. 6, pp. 1562–1572, 2022, doi: 10.30865/mib.v6i3.3999.

[4] A. M. Marshal, Digital Forensics Digital Evidence in Criminal Investigations, 1st ed. Wiley-Blackwell, 2009.

[5] E. R. Leukfeldt, "Phishing for suitable targets in the Netherlands: Routine activity theory and phishing victimization," Cyberpsychology, Behav. Soc. Netw., vol. 17, no. 8, pp. 551–555, 2014, doi: 10.1089/cyber.2014.0008.

[6] R. Ahmad and R. Thurasamy, "A Systematic Literature Review of Routine Activity Theory's Applicability in Cybercrimes," J. Cyber Secur. Mobil., vol. 11, no. 3, pp. 405–432, 2022, doi: 10.13052/jcsm2245-1439.1133.

[7] J. Hawdon, M. Costello, T. Ratliff, L. Hall, and J. Middleton, "Conflict Management Styles and Cybervictimization: Extending Routine Activity Theory," Sociol. Spectr., vol. 37, no. 4, pp. 250–266, 2017, doi: 10.1080/02732173.2017.1334608.

[8] E. I. B. C. Tompsett, A. M. Marshall, and N. C. Semmens, "Cyberprofiling: Offender profiling and geographic profiling of crime on the internet," Work. 1st Int. Conf. Secur. Priv. Emerg. Areas Commun. Networks, 2005, vol. 2005, pp. 22–25, 2005, doi: 10.1109/SECCMW.2005.1588290.

[9] M. M. Hassan, "Customer Profiling and Segmentation in Retail Banks Using Data Mining Techniques," Int. J. Adv. Res. Comput. Sci., vol. 9, no. 4, pp. 24–29, 2018, doi: 10.26483/ijarcs.v9i4.6172.

[10] K. K. Sindhu and B. B. Meshram, "Digital Forensics and Cyber Crime Datamining," J. Inf. Secur., vol. 03, no. 03, pp. 196–201, 2012, doi: 10.4236/jis.2012.33024.

[11] Angkasa, "Legal Protection for Cyber Crime Victims on Victimological Perspective," SHS Web Conf., vol. 54, p. 08004, 2018, doi: 10.1051/shsconf/20185408004.

[12] B. K. Mamade and D. M. Dabala, "Exploring The Correlation between Cyber Security Awareness, Protection Measures and the State of Victimhood: The Case Study of Ambo University's Academic Staffs," J. Cyber Secur. Mobil., vol. 10, no. 4, pp. 699–724, 2021, doi: 10.13052/jcsm2245-1439.1044.

[13] S. R. Sebastian, B. P. Babu, and S. R. Sebastian, "Are we cyber aware ? A cross sectional study on the prevailing cyber practices among adults from Thiruvalla , Kerala," vol. 10, no. 1, pp. 235–239, 2023, doi: 10.18203/2394-6040.ijcmph20223550.

[14] A. Kigerl, "Routine Activity Theory and the Determinants of High Cybercrime Countries," Soc. Sci. Comput. Rev., vol. 30, no. 4, pp. 470–486, 2012, doi: 10.1177/0894439311422689.

[15] T. Van Nguyen, "Cybercrime in Vietnam: An analysis based on routine activity theory," Int. J. Cyber Criminol., vol. 14, no. 1, pp. 156–173, 2020, doi: 10.5281/zenodo.3747516.

[16] A. Alzubaidi, "Measuring the level of cyber-security awareness for cybercrime in Saudi Arabia," Heliyon, vol. 7, no. 1, p. e06016, 2021, doi: 10.1016/j.heliyon.2021.e06016.

[17] A. Alzubaidi, "Cybercrime Awareness among Saudi Nationals: Dataset," Data Br., vol. 36, p. 106965, 2021, doi: 10.1016/j.dib.2021.106965.

[18] R. Saroha, "Profiling a Cyber Criminal," Int. J. Inf. Comput. Technol., vol. 4, no. 3, pp. 253–258, 2014.

[19] N. Innab, H. Al-Rashoud, R. Al-Mahawes, and Wauood Al-Shehri, "Evaluation of the Effective Anti-Phishing Awareness and Training in Governmental and Private

Organizations in Riyadh," 2018 21st Saudi Comput. Soc. Natl. Comput. Conf., pp. 1–5, 2018, doi: 10.1109/NCG.2018.8593144.

[20] C. M. M. Reep-van den Bergh and M. Junger, "Victims of cybercrime in Europe: a review of victim surveys," Crime Sci., vol. 7, no. 1, 2018, doi: 10.1186/s40163-018-0079-3.

[21] F. Alotaibi, S. Furnell, I. Stengel, and M. Papadaki, "A survey of cyber-security awareness in Saudi Arabia," 2016 11th Int. Conf. Internet Technol. Secur. Trans., pp. 154–158, 2016, doi: 10.1109/ICITST.2016.7856687.

[22] E. I. M. Zayid and N. A. A. Farah, "A study on cybercrime awareness test in Saudi Arabia - Alnamas region," 2017 2nd Int. Conf. Anti-Cyber Crimes, pp. 199–202, 2017, doi: 10.1109/Anti-Cybercrime.2017.7905290.

[23] N. A. G. Arachchilage and S. Love, "Security awareness of computer users: A phishing threat avoidance perspective, Computers in Human Behavior," Comput. Human Behav., vol. 38, no. 304–312, p. 161, 2014, doi: 10.1016/j.chb.2014.05.046.

[24] J. Abawajy, "User preference of cyber security awareness delivery methods," Behav. Inf. Technol. - Behav. IT., vol. 33, pp. 1–12, 2012, doi: 10.1080/0144929X.2012.708787.

[25] N. Ahmed, U. Kulsum, I. Bin Azad, A. S. Z. Momtaz, M. E. Haque, and M. S. Rahman, "Cybersecurity awareness survey: An analysis from Bangladesh perspective," p. 111, 2017, doi: 10.1109/R10-HTC.2017.8289074.Abstract.

[26] M. Norouzi, A. Souri, and M. S. Zamini, "Behavioral Malware Detection," vol. 2016, pp. 20–22, 2016.

[27] S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, "Customer profiling using classification approach for bank telemarketing," Int. J. Informatics Vis., vol. 1, no. 4–2, pp. 214–217, 2017, doi: 10.30630/joiv.1.4-2.68.

[28] M. Server, R. Excel, T. Rapidminer, and R.-M. Value, "Analysis of classification algorithms with rapidminer," pp. 517–520.

[29] Dr.J.Arunadevi, S.Ramya, and M. R. Raja, "A study of classification algorithms using Rapidminer," Int. J. Pure Appl. Math., vol. Volume 119, no. 12, pp. 15977–15988, 2018.

[30] N. Baharun, N. F. M. Razi, S. Masrom, N. A. M. Yusri, and A. S. A. Rahman, "Auto Modelling for Machine Learning: A Comparison Implementation between RapidMiner and Python," Int. J. Emerg. Technol. Adv. Eng., vol. 12, no. 05, pp. 15–27, 2022, doi: 10.46338/ijetae0522.

[31] J. F. Andry and H. Hartono, "Analysis and Prediction Supermarket Sales with Data Mining using RapidMiner Analysis and Prediction Supermarket Sales with Data Mining using RapidMiner," no. January, 2022.

[32] Rapidminer, "What's New in RapidMiner Server 9," no. September, 2020, [Online]. Available: https://docs.rapidminer.com/9.2/server/releases/changes-9.2.0.html.

[33] G. Michael, "Knowledge Based System for Predicting Cyber Crime Patterns Using Data Mining," J. Crit. Rev., vol. 7, no. 10, pp. 2043–2053, 2020.

[34] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with Naive Bayes - Which Naive Bayes?," 3rd Conf. Email Anti-Spam - Proceedings, CEAS 2006, no. January, 2006.

[35] Y. K. Putra, Fathurrahman, and M. Sadali, "Comparison of Pso-Based Naive Bayes and Naive Bayes Algorithm in Determining the Feasibility of Bumdes Credit," J. Phys. Conf. Ser., vol. 1539, no. 1, 2020, doi: 10.1088/1742-6596/1539/1/012030.

[36] G. Oh, J. Song, H. Park, and C. Na, "Evaluation of Random Forest in Crime Prediction: Comparing Three-Layered Random Forest and Logistic Regression," Deviant Behav., vol. 00, no. 00, pp. 1–14, 2021, doi: 10.1080/01639625.2021.1953360.

[37] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. De Carvalho, and A. A. Freitas, "A survey of evolutionary algorithms for decision-tree induction," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 42, no. 3, pp. 291–312, 2012, doi: 10.1109/TSMCC.2011.2157494.

[38] Anuradha and G. Gupta, "A self explanatory review of decision tree classifiers," Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2014, no. June, 2014, doi: 10.1109/ICRAIE.2014.6909245.

[39] H. Hauska and P. Swain, "The Decision Tree Classifier : Design and Potential Hans Hauska," no. June, 2014.

[40] B. Çiğşar and D. Ünal, "Comparison of Data Mining Classification Algorithms Determining the Default Risk," Sci. Program., vol. 2019, 2019, doi: 10.1155/2019/8706505.

[41] R. Sharma, S. N. Singh, and S. Khatri, "Data mining classification techniques - Comparison for better accuracy in prediction of cardiovascular disease," Int. J. Data Anal. Tech. Strateg., vol. 11, no. 4, pp. 356–373, 2019, doi: 10.1504/IJDATS.2019.103756.

[42] L. Marlina, M. lim, and A. P. Utama Siahaan, "Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)," Int. J. Eng. Trends Technol., vol. 38, no. 7, pp. 380–383, 2016, doi: 10.14445/22315381/ijett-v38p268.

[43] A. I. Kusumarini, P. A. Hogantara, M. Fadhlurohman, and N. Chamidah, "Perbandingan Algoritma Random Forest, Naïve Bayes, Dan Decision Tree Dengan Oversampling Untuk Klasifikasi Bakteri E. Coli," no. April, pp. 792–799, 2021.

[44] I. P. Wibina, K. Gumi, and A. Syafrianto, "Perbandingan Algoritma Naïve Bayes dan Decision Tree Pada Sentimen Analisis," vol. 1, pp. 1–15, 2022.