JURNAL

TECHNO NUSA MANDIRI

Journal of Computing and Information Technology

As an Accredited Journal Rank 4 based on **Surat Keputusan Dirjen Risbang SK Nomor 85/M/KPT/2020**

Vol. 19. No. 1 March 2022

ISSN: 1978-2136 (Printed)

ISSN: 2527-676X (Online)

Publisher:
Lembaga Penelitian dan Pengabdian Masyarakat Universitas Nusa Mandiri
Jl. Jatiwaringin Raya No. 02 RT 08 RW 013 Kelurahan Cipinang Melayu
Kecamatan Makassar Jakarta Timur 13620 Phone: 021 28534471
http://ejournal.nusamandiri.ac.id/index.php/techno/index

Vol. 19 No. 1 March 2022
P-ISSN: 1978-2136 | E-ISSN: 2527-676X
DOI: 10.33480/techno/v19i1

Vol. 19 No. 1 September 2022
P-ISSN: 1978-2136 | E-ISSN: 2527-676X
DOI: 10.33480/techno/v19i1

EDITORIAL BOARD

Vol. 19 No. 1 September 2022
P-ISSN: 1978-2136 | E-ISSN: 2527-676X
DOI: 10.33480/techno/v19i1

Editorial Address : Kampus Universitas Nusa Mandiri Tower Jatiwaringin
Jl. Jatiwaringin Raya No. 02 RT 08 RW 013 Kelurahan Cipinang Melayu
Kecamatan Makassar Jakarta Timur 13620

Website : http://ejournal.nusamandiri.ac.id/index.php/techno

Editorial Email : jurnal.techno@nusamandiri.ac.id

Vol. 19 No. 1 March 2022
P-ISSN: 1978-2136 | E-ISSN: 2527-676X
DOI: 10.33480/techno/v19i1

PREFACE

Editor of TECHNO(Journal of Computing and Information Technology), said praise and gratitude to the presence of Allah S.W.T, creator of the universe who mastered knowledge as wide as heaven and earth, for the abundance of grace and gifts that have been given to TECHNO editors to publish TECHNO Vol. 19, No. 1 March 2021, which is used by lecturers, researching, and professionals as a medium or media to publish publications on the findings of research conducted in each semester.

TECHNO is published 1 (one) year for 2 (two) times at the end of each semester, TECHNO editors receive scientific articles from the results of research, reports / case studies, information technology studies, and information systems, which are oriented to the latest in science and information technology in order to be a source of scientific information that is able to contribute to the increasingly complex development of information technology.

The editor invited fellow researchers, scientists from various tertiary institutions to make scientific contributions, both in the form of research results and scientific studies in the fields of management, education, and information technology. The editors really expect input from readers, information technology professionals, or those related to publishing, for the sake of increasing the quality of journals as we all hope.

The editor hopes that the scientific articles contained in the TECHNO scientific journal will be useful for academics and professionals working in the world of management, education, and information technology

Chief Editor

Vol. 19 No. 1 September 2022
P-ISSN: 1978-2136 | E-ISSN: 2527-676X
DOI: 10.33480/techno/v19i1

Vol. 19 No. 1 March 2022
P-ISSN: 1978-2136 | E-ISSN: 2527-676X
DOI: 10.33480/techno/v19i1

TECHNO (Journal Of Computing And Information Technology)

TABLE OF CONTENTS

DECISION SUPPORT SYSTEM FOR PURCHASING OF MIRRORLESS CAMERA USING WEIGHTED PRODUCT METHOD

Ayu Manik Martawiharjo[1*]; Melan Susanti[2]; Mari Rahmawati[3]

[1]Sistem Informasi, [2]Teknik Informatika
Universitas Nusa Mandiri
www.nusamandiri.ac.id
[1*)] ayuman11162196@nusamandiri.ac.id; [2]melan.msu@nusamandiri.ac.id

[3]Sistem Informasi Akuntansi
Universitas Bina Sarana Informatika
www.bsi.ac.id
[3]mari.mrw@bsi.ac.id
(*) Corresponding Author

Abstract— *A mirrorless camera is a camera that does not have a mirror or a pentaprism with the size and workpiece of a compact camera, but has an equivalent capability to a DSLR camera. There are several mirrorless camera manufacturers widely known in the market, among others: Canon, Sony, Fujifilm, Nikon, Olympus, and Panasonic with the advantages of each manufacturer's specifications highlighted to enhance the attractiveness of consumers. With many types of mirrorless cameras in the market, many consumers are still confused in choosing which mirrorless camera is right and suited to their needs. Therefore, it takes a decision support system for the selection of mirrorless cameras using the weighted product method that can generate decisions about mirrorless cameras that comply with the selection of consumer criteria. The criteria used in this study are price, sensor size, megapixel, maximum ISO, and LCD. The results of this study show that the alternative mirrorless camera which has the highest value is the Olympus PEN E-PL9 camera with a value of 0.148.*

Keywords: *Mirrorless Camera, Weighted Product, Decision Support System.*

Abstrak — Kamera *mirrorless* merupakan kamera yang tidak memliki cermin atau pentaprisma dengan ukuran dan cara kerja seperti kamera saku, namun memiliki kemampuan yang setara dengan kamera DSLR. Ada beberapa produsen kamera *mirrorless* yang cukup dikenal secara luas di pasaran antara lain: Canon, Sony, Fujifilm, Nikon, Olympus dan Panasonic dengan keunggulan masing-masing spesifikasi yang ditonjolkan oleh produsen agar dapat meningkatkan daya tarik konsumen. Dengan banyaknya jenis kamera *mirrorless* dipasaran, banyak konsumen yang masih bingung dalam memilih kamera *mirrorless* mana yang tepat dan sesuai dengan kebutuhannya. Untuk itu, dibutuhkan suatu sistem pendukung keputusan untuk pemilihan kamera *mirrorless* menggunakan metode *weighted product* yang dapat menghasilkan keputusan tentang kamera *mirrorless* yang sesuai dengan pemilihan kriteria konsumen. Kriteria-kriteria yang digunakan dalam penelitian ini adalah harga, ukuran sensor, megapiksel, maksimum ISO dan LCD. Hasil dari penelitian ini menunjukkan bahwa alternatif kamera *mirrorless* yang memiliki nilai tertinggi adalah kamera Olympus PEN E-PL9 dengan nilai 0,148.

Kata Kunci: Kamera *Mirrorless*, *Weighted Product*, Sistem Pendukung Keputusan

INTRODUCTION

A mirrorless camera is a camera that does not have a mirror and an optical viewfinder with image quality equivalent to that of a *Digital Single Lens Reflex* (DSLR) camera (Hermawan 2016). Currently, camera sales are quite fast in the market, especially mirrorless cameras and DSLR cameras (Fajar 2019).

There have been many consumers who prefer Mirrorless cameras over DSLR cameras (Susmikawati and Sunarti 2017). The rapid number of mirrorless camera products on the market has made some consumers confused about choosing the right camera according to their needs. (Fajar 2019). In general, consumers buy mirrorless cameras with high specifications, but their use is only a recreational hobby. In fact, by using it as a recreational hobby, consumers can choose a mirrorless camera with medium specifications. (Gani, Kridalaksana, and Arifin 2019). This confusion is one of the problems in choosing a mirrorless camera (Gani et al. 2019). Consumers need to pay attention to this problem so that they don't make the wrong purchase of a camera that will later harm them. (Putra et al. 2018)

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v17i1.1722

Based on the explanation above, this research creates a decision support system for purchasing a mirrorless camera using a weighted product.

The purpose of this study is to apply the weighted product method, with the creation of a decision support system that is expected to help consumers who want to buy a mirrorless camera according to their needs.

Table 1. Related research

No	Topic/Title	Author	Results/Conclusion
1	Decision Support System for High School Scholarship Recipients Using Fuzzy Multiple Attribute Decision-Making Model Weighted Product	T. Hidayatulloh, S. Suhada, E. Nursyifa, and L. Yusuf	In this case, SMAN 1 Cicurug Sukabumi organizes scholarships for outstanding students to continue their studies at SMAN 1 Cicurug Sukabumi. The criteria used in determining the scholarship are the average value of report cards, the average value of diplomas, number of dependents of parents, parents' income, and areas of expertise with the weight of each criterion. The results obtained from the 5 alternatives used are those that have the highest value in the ranking, namely the 2nd alternative (Hidayatulloh et al. 2018).
2	Design and Application of Weighted Product Method in Decision Support System for Purchase of Laptop	Nur Sumarsih	The purpose of this study is to assist consumers in choosing a laptop based on predetermined criteria. The criteria used include price, RAM capacity, type of processor, hard drive capacity, and VGA (Sumarsih 2019).
3	Implementation of the Weighted Product Method in the Decision Support System for Choosing the Best Electronic Money Service	M. Wasti, S. Hartini, and Rinawati,	This research was conducted to assist users in choosing the best electronic money. Because electronic money is easier to use than conventional money. The criteria used in this study were comfort, safety, promotion, speed, and convenience. The alternatives used are OVO, Go-Pay, and Dana. Based on the results obtained in the calculation using the weighted product, the first alternative recommendation for electronic money is that OVO has a final V vector value of 0.394. (Wasti, Hartini, and Rinawati 2019).

MATERIALS AND METHOD

Purchasing

Purchasing is an effort to procure goods or services that are used in the company to be managed by themselves, for the benefit of production or resale (Indrajani 2015).

According to (Indrajani 2015) the function of purchasing is responsible for determining the selected contribution in the procurement of goods, issuing purchase orders to selected contributors, and obtaining information about the price of goods.

Decision Support System

Decisions are the result of the process of choosing the best option from several available alternatives. To get the best choice, we will try to devote all the thoughts and activities necessary to the decision-making process (Diana 2018).

According to (Kusrini 2017) the stages of decision making are as follows:

a. Identify the problem
b. Choose a problem-solving model
c. Collect the data needed to implement the decision model
d. Implementing the model
e. Evaluating the positive side of the chosen alternative
f. Implement the chosen solution

Decision support systems are used to support management in carrying out analytical work in less structured conditions and with unclear criteria (Kusrini 2017).

In (Simangunsong and Sinaga 2019) A decision support system is a system that has been designed and can be implemented to support decisions that have been agreed upon in the selection of an object.

From the explanation above, it can be concluded that a decision support system is a system that can assist and provide decision results in processing data or information for its users.

Characteristics and Capabilities of Decision Support System

A decision support system has several characteristics. According to (Sari 2018) characteristics of decision support systems, among others:

1. Support all organizational activities,
2. Support interactive decisions,
3. It is constant and can be used repeatedly,
4. Data and models are the main components,
5. Using external and internal data,
6. Some of the models used are quantitative models.

Weighted Product

The decision-making technique of several alternatives available in Multi-Attribute Decision Making (MADM) is the definition of Weighted Product. The method evaluates several alternatives against a set of criteria or attributes, where each attribute is independent of one another (MZ 2018).

According to (Nofriansyah and Defit 2017) The weighted product method is a method that uses multiplication in connecting attribute ratings, then each attribute rating is raised to the power of the available attribute weights.

The basic concept of normalization of the Weighted Product (WP) method is to find S_i by raising the criterion value on an alternative with a weighted value for each criterion owned. After that, the value of S_i will be used to find the value of V_i by dividing the value of S_i by $\sum S_i$, so that it will produce a value for each alternative (Maruloh et al. 2020).

The steps in completing the Weighted Product (WP) method according to (Diana 2018) are as follows:

a. Identify the criteria that will be used as a reference in decision-making.
b. Each criterion is determined by initial weight and normalizes or corrects the weights to produce a value of wj= 1 where 1, 2, …, n is the number of alternatives and Σ_{wj} is the total number of weighted values.

$$W_j \frac{w_j}{\sum w_j} \quad \text{.....(1)}$$

c. Determine the vector value (S)

$$S_i = \prod_{j=1}^{n} X_{ij}\, wj \quad \text{.....(2)}$$

d. Determine the vector value (V)

$$V_j = \frac{\prod_{j=1}^{n} X_{ij} wj}{\prod_{j=1}^{n} (X_{ij*}) wj} \quad \text{.....(3)}$$

Vector V is an alternative preference that will be used for ranking by dividing each number of vector values S by the number of all vectors S

Likert Scale

The Likert scale is a psychometric scale that is commonly used for questionnaires and is the most widely used scale in research in the form of surveys (Nofriansyah and Defit 2017). In this scale, the rating ranges from 1 to 5 with several scale formats in table 2.

Table 2. Positive Statement Likert Scale

Very important	5
Important	4
Quite important	3
Unimportant	2
Very unimportant	1

Source: (Nofriansyah and Defit 2017)

Research Stages

Below are the stages in the research as follows:

1. Problem Identification
   The first step in this research is to identify the problem that will be used as the subject of the research discussion, namely determining what are the criteria that influence consumers in choosing a mirrorless camera and how to apply the weighted product method to a decision support system in choosing a mirrorless camera.
2. Literature Study
   The next step is a literature study by studying and understanding theories related to the object of research derived from books, journals, and previous research studies and will then be used as a theoretical study in the preparation of this thesis report.
3. Data Collection
   At this stage, the authors collect data in the form of information about the selection of mirrorless cameras using observations and interviews with sellers and prospective buyers of mirrorless cameras as objects of research.
4. Data Analysis
   Analysis of the mirrorless camera selection data from this study used the weighted product method which was carried out by collecting data from a questionnaire about the selection of a mirrorless camera and the results of the analysis to obtain information that must be concluded.
5. Results of Data Analysis
   After the data analysis phase of the purchase of a mirrorless camera using the weighted product method, an analysis result is produced which is the result of a research process carried out.
6. Conclusions and Suggestions
   This stage is the last stage of the description of the research process by concluding the results of

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v17i1.1722

the research on decision support systems in purchasing mirrorless cameras and providing suggestions for developing further research to be even better in purchasing mirrorless cameras using the weighted product method.

Research Instruments

According to (Zai, Mesran, and Buulolo 2017) the human instrument or referred to as a qualitative researcher, is used to determine the focus of research, select informants as data sources, collect data, assess data quality, interpret data and draw conclusions from their research.

The research instrument used for this study which serves as a data collection tool is a questionnaire. The research instrument is used to measure the value of the criteria being analyzed, thus the number of instruments to be used for research will depend on the number of criteria to be studied.

In this study the criteria used are:

1. Price, basically the price is the most important benchmark for most potential buyers.
2. Sensor size relates to the ability to capture light and determine whether or not the photos taken are good. Therefore the size of the camera sensor is the most important part in considering the purchase of a camera.
3. Megapixels are the main thing in a camera and are an important consideration in choosing a camera because the size of the megapixels on the camera affects the quality of the resolution in the photos taken.
4. International Organization for Standardization (ISO) maximum, used in low light conditions. The higher the International Organization for Standardization (ISO) is used, the more sensitive the sensor will be so that the images or photos captured by the camera are brighter.
5. Liquid Crystal Display (LCD), serves to display the results of photoshoots with several system settings on the camera. Liquid Crystal Display (LCD) is also important. Because the small or wide Liquid Crystal Display (LCD) makes consumers interested in choosing a camera

As for the alternative product choices from mirrorless camera shops, they are as follows: Fujifilm X-T20, Fujifilm A-X5, Canon EOS M100, Nikon Z50, Sony A6400, Olympus Pen E-PL9, Panasonic Lumix GF10, and Panasonic Lumix GF9. In observing the respondents, the measurement scale used is the Likert scale, which will get answers in the form of strongly agree, agree, neutral, disagree, and strongly disagree.

Table 2. Positive Statement Likert Scale

Very important	5
Important	4
Quite important	3
Unimportant	2
Very unimportant	1

Source: (Nofriansyah and Defit 2017)

Table 3. Negative Statement Likert Scale

Very important	1
Important	2
Quite important	3
Unimportant	4
Very unimportant	5

Source: (Nofriansyah and Defit 2017)

Research Methods

The author obtains data by conducting direct research in a systematic and standard procedure to obtain good and correct data with the following data collection techniques:

a. Observation Method (Observation)
   Observations were made directly to prospective mirrorless camera buyers and gave a questionnaire about the decision to choose a mirrorless camera to be studied so that they got the materials needed. According to (Sugiyono 2019) a questionnaire is a list of questions or statements made based on indicators of research variables given to respondents.
b. Interview Method (Interview)
   The author obtains data and examines the truth of the information and data by conducting questions and answers directly with the camera seller as the object of research.
c. Library Method (Literature)
   The literature study is looking for various reference books sourced from journals, articles, literature books, and the internet later to support the completeness of the formulation and comparison materials related to the problems to be discussed.

Population and Research Sample

An object or subject that has certain quantities and characteristics determined by researchers to study and then draw conclusions is a population (Zai et al. 2017). According to (Zai et al. 2017) population is not just people, but objects and other things. And not just the amount that is in the object or subject but includes all the characteristics possessed by the object or subject.

In this study, the population taken is consumers or potential buyers of mirrorless cameras. Sampling in this study using a technique or method of random sampling. The sample is selected from the population elements at random, where every member of the population has the same rights to be sampled.

In this study, samples were taken as many as 30 respondents, namely prospective buyers of mirrorless cameras, to represent the population as a whole.

Data Analysis Method

In this study, we will use Weighted Product (WP) analysis were in determining a decision using multiplication to link the attribute rating, where the rating of each attribute is raised first by the weight of the attribute in question.

According to (Frieyadie and Fariati 2019) argues that the Weighted Product (WP) method is a quantitative decision-making method.

From these criteria then fix the weights first. The weight of the criteria used as a test is obtained from the results of the questionnaire where the respondent chooses the level of importance according to the appropriate needs in selecting a mirrorless camera, then normalization of the weight or weight improvement is carried out by determining the vector S, namely the value of each alternative, this calculation is carried out where the data to be analyzed is carried out. multiplied by the previous one must be raised to the power of the weight of each criterion. After each vector S gets a value, the next step is to determine the value of the vector V used for alternative ranking. After the calculation using the V vector is complete, the next step is to enter all the calculation results into the table according to the highest value of the V vector value, then the calculation results will show the ranking of the V vector values from the largest to the smallest so that the best alternative recommendation for choosing a mirrorless camera is obtained based on the highest value vector V.

RESULTS AND DISCUSSION

Research Data

At this stage, the authors collect the mirrorless camera selection data needed to perform calculations using the weighted product method. The following are criteria that are used as a reference in choosing a mirrorless camera using the weighted product method:

Table 4. Criteria

Criteria	Symbol
Price	C1
Sensor Size	C2
Megapiksel	C3
Maksimum ISO	C4
LCD	C5

From the table, a level of importance for criteria is determined based on the weight value for each criterion with a weight value of 1 to 5, this weighting refers to the Likert scale in Tabel 6.

Tabel 6. Value Weight

Statement	Weight
Very important	5
Important	4
Quite important	3
Unimportant	2
Very unimportant	1

Weighted Product Step
Determining Alternative

Table 7. Mirrorless Camera Data

No	Mirrorless Camera	Specification					Code
		Price	Sensor Size	MP	ISO	LCD	
1	Fujifilm X-T20	Rp. 10.350.000	APS-C	24	51200	3.00”	A1
2	Fujifilm X-A5	Rp. 7.550.000	APS-C	24	12800	3.00”	A2
3	Canon EOS M100	Rp. 5.500.000	APS-C	24	25600	3.00”	A3
4	Nikon Z50	Rp. 15.500.000	APC-C	20	51200	3.02”	A4
5	Sony A6400	Rp. 13.900.000	APC-C	24	102400	3.00”	A5
6	Olympus Pen E-PL9	Rp. 10.500.000	Four Thirds	16	25600	3.00”	A6
7	Panasonic Lumix GF10	Rp. 6.900.000	Four Thirds	16	25600	3.00”	A7
8	Panasonic Lumix GF9	Rp. 4.500.000	Four Thirds	16	25600	3.00”	A8

Table 4 is the first step in determining the alternative to be used in the calculation. In this study, 8 samples of mirrorless camera data will be used.

A. Determining Improvement Criteria Weight

Determine the weight of preference or determine the level of importance based on the level of importance of each criterion. The following is the weight value given by the respondents, namely:

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v17i1.1722

Table 7. Respondent Input

Criteria	Value
Price	3
Sensor Size	3
Megapiksel	3
Maksimum ISO	2
LCD	2

Next, the weights will be corrected first with an initial weight of W = (3, 3, 3, 2, 2), where W is the weight of each criterion that the respondent gives. The following is the calculation for the improvement of the criteria as follows:

1) Price Criteria. W_1 = 0,230
2) Sensor Size Criteria. W_2 = 0,230
3) Megapixel Criteria. W_3 = 0,230
4) Maksimum ISO Criteria. W_4 = 0,154
5) LCD Criteria. W_5 = 0,154

The following are the results of the calculation of the improvement in the criteria weights:

Table 8. Criteria Weight Improvement

Criteria	Value	Weight
Price	3	0,230
Sensor Size	3	0,230
Megapiksel	3	0,230
Maksimum ISO	2	0,154
LCD	2	0,154

B. Determining the Weight of Each Alternative

The next step is to give weighting criteria for each mirrorless camera data contained in table 7 Mirrorless Camera Data. The following is the weight of the criteria for each mirrorless camera, namely:

Table 9. Criteria Weight of Each Mirrorless Camera

Criteria	Alternatives							
	A1	A2	A3	A4	A5	A6	A7	A8
C1	2	3	4	1	1	2	3	4
C2	3	3	3	3	3	4	4	4
C3	3	3	3	3	3	5	5	5
C4	2	4	3	2	1	3	3	3
C5	4	4	4	3	4	4	4	4

C. Calculating Vector S

After getting the results of the calculation of the improvement value of the criterion weights, then the next step is to calculate the vector S where this calculation will be multiplied but before it is raised with the weight of each criterion. With the weight as a positive power for the favorable criteria and negative weight for the cost criteria. The following is the calculation of the vector S, namely:

1) Alternative Mirrorless Camera A1
   S_1 = 1,947
2) Alternative Mirrorless Camera A2
   S_2 = 1,972
3) Alternative Mirrorless Camera A3
   S_3 = 1,765
4) Alternative Mirrorless Camera A4
   S_4 = 2,183
5) Alternative Mirrorless Camera A5
   S_5 = 2,050
6) Alternative Mirrorless Camera A6
   S_6 = 2,491
7) Alternative Mirrorless Camera A7
   S_7 = 2,269
8) Alternative Mirrorless Camera A8
   S_8 = 2,122

The following are the results of the calculation of the vector S, namely:

Table 10. Vector S Calculation

Alternative	Value
A1	1,947
A2	1,972
A3	1,765
A4	2,183
A5	2,050
A6	2,491
A7	2,269
A8	2,122

D. Determine the Vector V

After getting the vector S value, the next step is to determine the alternative ranking of mirrorless cameras by dividing the vector V value used for ranking for each alternative by the total value of all S vector alternative values. The following is the vector V calculation, namely:

1) V_1 = 0,116
2) V_2 = 0,117
3) V_3 = 0,105
4) V_4 = 0,130
5) V_5 = 0,122
6) V_6 = 0,148
7) V_7 = 0,135
8) V_8 = 0,126

E. Final Value Obtained

After the calculation using vector V is complete, the next step is to enter all the calculation results into the table according to the highest value of vector V, then the highest value will be obtained as the recommended value in Table 12.

Table 12. Result Value

Alternative	Value	Ranking
A6	0,148	1
A7	0,135	2
A4	0,130	3
A8	0,126	4
A5	0,122	5
A2	0,117	6
A1	0,116	7
A3	0,105	8

So the results of the calculation of choosing a mirrorless camera using the weighted product method state that the first rank is the A6 alternative with a value of 0.148, namely the Olympus PEN E-PL9 mirrorless camera. The second is the A7 alternative with a value of 0.135, namely the Panasonic Lumix GF10 mirrorless camera. The third is the A4 alternative with a value of 0.130, namely the Nikon Z50 mirrorless camera. The fourth is the A8 alternative with a value of 0.126, namely the Panasonic Lumix GF9 mirrorless camera. The fifth is the A5 alternative with a value of 0.122, namely the Sony A6400 mirrorless camera. The sixth is the A2 alternative with a value of 0.117, namely the Fujifilm X-A5 mirrorless camera. The seventh is the A1 alternative with a value of 0.116, namely the Fujifilm X-T20 mirrorless camera. And the last eighth ranking is the A3 alternative with a value of 0.105, namely the Canon EOS M100 camera.

CONCLUSION

From the results of the research discussion on the decision support system for choosing a laptop using the weighted product method, the authors can conclude that in using the weighted product method, criteria are needed to be considered, the criteria that have been determined are price, sensor size, megapixel, maximum ISO and LCD. Meanwhile, to build a mirrorless camera purchase decision support system using the weighted product method, the first step is to determine the criteria and alternatives for mirrorless cameras to be compared, then the data will be calculated using the weighted product method.

REFERENCE

Diana. 2018. *Metode Dan Aplikasi Sistem Pendukung Keputusan*. Yogyakarta: Deepublish Publisher.

Fajar. 2019. "Kamera DSLR vs Kamera Mirrorless? Temukan Perbedannya Di Sini!"

Frieyadie and Fariati Fariati. 2019. "Penerapan Metode Weighted Product Sebagai Pendukung Keputusan Seleksi Karyawan Baru Pt. Hi-Lex Indonesia." *Jurnal Riset Informatika* 2(1):9–16.

Gani, Arman, Awang Harsa Kridalaksana, and Zainal Arifin. 2019. "Analisa Perbandingan Metode Simple Additive Weighting (SAW) Dan Weight Product (WP) Dalam Pemilihan Kamera Mirrorless." 14(2).

Hermawan. 2016. "Apa Sih Kamera Mirrorless Itu?"

Hidayatulloh, Taufik, Satia Suhada, Eva Nursyifa, and Lestari Yusuf. 2018. "Pengambilan Keputusan Penerima Beasiswa Sma Menggunakan Fuzzy Multiple Atribute Decision Making Model Weighted Product." *Jurnal Pilar Nusa Mandiri* 14(2):247.

Indrajani. 2015. *Database Design*. Jakarta: Elex Media Komputindo.

Kusrini. 2017. *Konsep Dan Aplikasi Sistem Pendukung Keputusan*. 1st ed. Yogyakarta: Andi Publisher.

Maruloh, Mohammad Darussalam, Mochamad Nandi Susila, and Wahyudia. 2020. "Sistem Penunjang Keputusan Seleksi Karyawan Terbaik PT. Golden Living Indonesia Dengan Metode Weighted Product." VI(1):81–88.

MZ, Yumarlin. 2018. "Prototype Sistem Pendukung Keputusan Pemilihan Kamera Digital." *Jurnal Informasi Interaktif* 3(1):95–103.

Nofriansyah, Dicky and Sarjon Defit. 2017. *Multi Criteria Decision Making (MCDM) Pada Pendukung Keputusan*. Cetakan Pe. Yogyakarta: Penerbit Deefpublish.

Putra, Guntur Maha, Novica Irawati, Sistem Informasi, and Stmik Royal. 2018. "Analisis Pemilihan Handphone Rekomendasi Dengan Metode Weighted Product." Pp. 199–204 in *Seminar Nasional Royal (SENAR) 2018*. Vol. 9986. Medan: STMIK ROYAL.

Sari, Febriana. 2018. *Metode Dalam Mengambil Keputusan*. Yogyakarta: Deepublish.

Simangunsong, P. B. N. and Soni Bahagia Sinaga. 2019. "Sistem Pendukung Keputusan Pemilihan Dosen Berprestasi." 58.

Sugiyono. 2019. *Metode Penelitian Kuantitatif, Kualitatif, Dan R&D*. edited by Sutopo. Bandung: Alfabeta.

Sumarsih, Nur. 2019. "Perancangan Dan Penerapan Metode Weighted Product Dalam Sistem Pendukung Keputusan Pembelian Laptop." 1(4):207–10.

Susmikawati, Y. and S. Sunarti. 2017. "Pengaruh Country Of Origin Terhadap Perceived Quality Dan Minat Beli Konsumen (Studi Pada Calon Konsumen Yang Berminat Membeli Kamera Mirrorless Fujifilm X-Series Di Kota Malang)." *Jurnal Administrasi Bisnis S1 Universitas Brawijaya* 49(2):88–95.

Wasti, Melati, Sari Hartini, and Rinawati. 2019. "Implementasi Metode Weighted Product Dalam Sistem Pendukung Keputusan

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v17i1.1722

Pemilihan Layanan Uang Elektronik Terbaik." *Jurnal Teknika* 11(2):1131–37.

Zai, Yosa'aro, Mesran, and Efori Buulolo. 2017. "Sistem Pendukung Keputusan Untuk Menentukan Buah Rambutan Dengan Kualitas Terbaik Menggunakan Metode Weighted Product (WP)." *Media Informatika Budidarma (MIB)* 1(1):8–11.

Vol. 19, No. 1 March 2022 | DOI: https://doi.org/10.33480/techno.v19i1.1722

ANALYSIS OF CUSTOMER SERVICE QUALITY BASED ON SERVQUAL (SERVICE QUALITY) AND ANALYTIC HIERARCHY PROCESS (AHP) METHODS

Regina Puteri Laurichela[1]; Cepi Cahyadi[2]

[1]Information Systems Study Program
Universitas Nusa Mandiri
www.nusamandiri.ac.id
regina11162876@nusamandiri.ac.id; cepi.ccd@nusamandiri.ac.id

Abstract— *The purpose of conducting this study is to identify the attributes of service quality, the difference between expectations and perceptions (gaps) of each attribute with the Servqual method, and the priority of improvement recommendations with the Analytical Hierarchy Process method. This research is a descriptive research. The collection of data used is with a simple random. The population in this study is consumers who use services in The Healthy Catering. The sample took 80 respondents. The research instruments used are the identification of servqual instruments and AHP instruments. The method of data collection is by observation, library study, interview, and questionnaire dissemination. Data analysis techniques that will be conducted are validity, reliability, Servqual (Service Quality), and Analytic Hierarchy Process (AHP) tests. From the results of this study showed that 1) obtained six highest gap values including E1 with a gap value of -0.375, R2 with gap value -0.350, R4 with gap value -0.338, E3 with gap value -0.275, R1 with gap value is -0,263, and RV1 with gap value is -0,150. 2) weighted Servqual calculation result shows improvement priority for attributes with high gaps, it is E1, R4, E3, RV2, R2, and R1.*

Keywords: *Service Quality, Consumer, Service Quality, Analytic Hierarchy Process.*

Intisari— Tujuan dilakukannya penelitian ini adalah untuk mengidentifikasi atribut-atribut kualitas pelayanan, nilai selisih antara harapan dan persepsi (gap) dari masing – masing atribut dengan metode Servqual, dan prioritas rekomendasi perbaikan dengan metode Analytical Hierarchy Process. Penelitian ini adalah jenis penelitian deskriptif. Pengumpulan data yang digunakan adalah dengan acak sederhana. Populasi dalam penelitian ini adalah konsumen yang menggunakan jasa pada The Healthy Catering. Sampel yang diambil sebanyak 80 orang responden. Instrumen penelitian yang digunakan adalah identifikasi instrument servqual dan instrumen AHP. Metode pengumpulan data dengan observasi, studi pustaka, wawancara dan penyebaran kuesioner. Teknik analisis data yang akan dilakukan adalah dengan uji validitas, reliabilitas, metode Servqual (Service Quality) dan Analytic Hierarchy Process (AHP). Dari hasil penelitian ini menunjukkan bahwa 1) didapatkan 6 nilai gap tertinggi diantaranya E1 dengan nilai gap -0,375, R2 dengan nilai gap -0.350, R4 dengan nilai gap -0.338, E3 dengan nilai gap -0.275, R1 dengan nilai gap adalah -0,263, dan RV1 dengan nilai gap adalah -0,150 2) hasil perhitungan Servqual terbobot menunjukkan prioritas perbaikan untuk atribut dengan gap tinggi yaitu E1, R4, E3, RV2, R2, dan R1.

Kata kunci: Kualitas Layanan, Konsumen, *Service Quality*, *Analytic Hierarchy Process.*

INTRODUCTION

Culinary is one of the most promising industries. This is because the culinary sector can sustain market share, and its product, such as food, satisfies one of the three basic human needs. Catering is a prevalent culinary business among the public. Catering is a prevalent business that employs many people, as it is believed to have good prospects and produces promising returns. According to Anggraini (2021), the community has a variety of catering businesses, including catering for company personnel, catering for transportation, catering for celebrations/parties, special catering, traditional catering, and catering for hospitals.

A catering entrepreneur must consider many factors to ensure that the catering business they are building succeeds, including the quality of the food delicacy, the variety of the menu supplied to customers, the performance of serving staff, and the cleanliness of the food and equipment used (Riadikemas, 2021). These factors are summed up in one significant factor: the quality of catering services. Catering service quality is determined by the food's quality, the catering staff's skill, and standard work procedures (Syafitr & Herlawati,

Vol. 19, No. 1 March 2022 | DOI: https://doi.org/10.33480/techno.v19i1.1722

2016). The primary focus for service providers to ensure that the services they provide are of high quality and consumer-oriented is on achieving excellent service quality.

This study uses the Service Quality method and the Analytic Hierarchy Process, which can determine the quality criteria that must be further improved by a service provider where the gap between expectations and perceptions is used as the basis for measuring (Sikumbang, 2017).

The Servqual method will be developed to assist consumers in locating high-quality services across five servqual dimensions, namely Reliability, Responsiveness, Tangibles, Assurance, and Empathy. (2017) (Supartiningsih). This method allows for identifying the gap between consumer expectations and desires. Meanwhile, AHP is utilized to aid in weighting each criterion's sub-criteria (tangibles, reliability, responsiveness, assurance, and empathy) through pairwise comparison analysis. The research findings are supposed to provide insight into service criteria that need to be improved in order for the organization to increase service quality.

Numerous scholars have researched service quality analysis using both servqual and AHP methods (Ammarapala, 2017); (Mahmudi, 2021); (Putri et al., 2020); (Zaidiah et al., 2021); (Efendi et al., 2019).

Serving consumers is a phenomenon that frequently manifests itself in acquiring high-quality service, such as that provided by catering services, The Healthy Catering. Consumer complaints frequently include the server's performance. Consumers express their dissatisfaction with the waiter's politeness, friendliness, alertness, and speed in giving service (Simanjuntak et al., 2018). Then there is the presentation and food quality. The issues that occur may be something that The Healthy Catering specializes in solving by providing the highest quality service. As a result, to optimize performance and service quality, a service provider requires a function capable of incorporating the "voice of customer" while prioritizing consumer expectations and desires. By determining the quality of service concerning consumer desires, this research is supposed to provide companies with solutions for service improvement.

According to Loveloc (Fatihudin & Firmansyah, 2019), a service is a service that service providers provide to service users. The process is not implemented through physical products, but through intangible services that typically do not result in ownership of any production sources. Kotler and Armstrong (in Indrasari, n.d.) define service quality as the features and characteristics that a service or product as a whole possesses that can directly or indirectly meet consumer needs. Meanwhile, consumer satisfaction, as defined by Philip Kotler and Kevin Lane Keller, is the feeling that consumers have when they compare the performance (results) of the products or services they receive to their expectations for the performance obtained, both in terms of pleasure and disappointment (Indrasari, n.d.).

The servqual and AHP methodologies were used to research the Thai toll road authority controlled by the Expressway Authority of Thailand (EXAT). This study aims to determine the quality of toll road services and identify the most critical service criteria. The SERVQUAL approach is combined with a gap analysis model to evaluate service expectations and perceptions, as well as the Analytic Hierarchy Process (AHP) method to determine the relative importance of different service criteria. The research contributes to a complete understanding of toll road users' expectations and how to maximize their satisfaction by minimizing detected gaps. Additionally, the findings advise which service characteristics EXAT should prioritize to maintain the best V toll road services standards. (Ammarapala, 2017).

The research conducted by Putri (2020), entitled "Assessment of Customer Satisfaction with Service Quality X Using the Servqual and AHP Methods," was also conducted to ascertain customer satisfaction with service quality and the order of priority for planned service quality X enhancements. Customer satisfaction was positive at a value of 4.025, indicating that customers are satisfied with X's service quality.

Syafitri & Herlawati (2016) conducted another study on the same topic, entitled "Assessment of Longe Digital Service Quality Using Servqual Methods and Analytical Hierarchy Process." This research intends to assist the Bank in establishing the highest priority areas for service quality improvement, specifically those where service quality is still judged unacceptable by the Bank's service users. Similar to this study, this one employs the Sevqual method to ascertain the size of the perception gap between actual and perceived customer expectations, as well as the AHP method to embed the Servqual dimension variable attribute in terms of attribute importance. The study's findings suggest which service variable attributes are least capable of satisfying clients and their relative importance in terms of improvement.

Then, research on the integration of servqual and AHP was conducted to evaluate Dekranasda's service quality of Rembang Regency. This research aimed to evaluate the quality of customer service at the Dekranasda of

Vol. 19, No. 1 March 2022 | DOI: https://doi.org/10.33480/techno.v19i1.1722
|

Rembang Regency, using the analytical Hierarchy Process (AHP) to weight each dimension and criterion and the Servqual analysis method to ascertain consumer perceptions and expectations across the five servqual dimensions. The study determined that Dekranasda needed to enhance 10 service criteria to improve the institution's service quality (Mahmudi, 2021).

The Healthy Catering is a subsidiary of PT. Al EHSAN AMS provides food and beverage services, including catering to factory employees. At the moment, the company employs approximately 360 people. The company is based in South Bekasi, Bekasi City.

This research was conducted in order for researchers to identify service quality attributes in terms of five service quality dimensions, the value of the difference between expectations and perceptions (gaps) for each attribute using the Servqual method, prioritize recommendations for improvement using the Analytical Hierarchy Process, and then provide recommendations for attributes with a large gap between expectations and perceptions.

Based on the problem formulation, it can be inferred that the following are the outputs of this study's hypothesis:

H0 = a) Five dimensions of service quality variables do not affect customer service at The Healthy Catering. b) The Servqual and AHP methods generate different recommendation rating values.

H1 = a) Five dimensions of service quality variables: reliability, tangible, empathy, assurance, and responsiveness positively affect customer service at The Healthy Catering. b) The Servqual and AHP methods generate the same recommendation rating values.

MATERIALS AND METHODS

This research is a descriptive evaluation study (E, Barlian). The following steps are included in the implementation of this research: 1) planning the research, 2) identifying problems, 3) developing and assembling research instruments (questionnaires), 4) collecting data, 5) testing data, 6) processing and analyzing data, and 7) drawing conclusions. This study employs two instruments, the servqual, and the AHP. The researchers collected data using four methods: interviews, observation, distribution of questionnaires, and literature review.

The population is a category of items or persons that share similar characteristics and qualities, based on the criteria established by the researcher for subsequent study. Berlian (2016) asserted that the population could take the form of creatures or people, or it can take the form of all objects and items contained within the population. The population for this study is the consumer of The Healthy Catering service, as determined by a simple random sampling technique. To estimate the minimum sample size, the researchers used the Slovin formula as described in Suryani's book (Suryani & Hendryadi, 2016), namely:

$$n = \frac{N}{1+Ne^2} = \frac{360}{1+360(0.1)^2} = 78.2608$$

As a result, it can be determined that this study's sample consisted of 80 respondents. The validity, reliability, gap value analysis, and weighting using the analytical hierarchy process (AHP) will all be used to analyze the data in this research.

RESULTS AND DISCUSSION

1. Validity Test

In this study, the validity of the R_{count} value of each attribute with R_{table} was determined. This study suggests that the R_{count} value for all the attributes examined is 22, with the expectations and perceptions column indicating a greater value than R_{table}, indicating that the survey's questions are valid.

2. Reliability Test

Cronbach Alpha was used in this study utilizing SPSS software and yielded a result of 0.812 and 0.854 for the expectation and perception attributes, respectively. These findings indicate that the instrument employed in this study is reliable or trustworthy, as the value achieved is more than 0.6.

3. Recapitulation of the Calculation of the Service Quality Gap

The service quality gap value of The Healthy Catering was computed by subtracting from the mean score of perception and expectation that consumers have filled in. The gap value is shown in Table 1. A negative score indicates that the service results are of low quality and high satisfaction. In contrast, a positive value indicates that the service results are high quality and high satisfaction.

Table 1. Recapitulation of the Calculation of the Service Quality Gap

No	Dimension	Mean Value		Gap	Note
		Perception	Expectation		
1	Tangibles	**3.671**	**3.557**	**0.114**	high quality and high satisfaction
2	Reliability	**3.585**	**3.715**	**-0.130**	low quality and low satisfaction

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI: https://doi.org/10.33480/techno.v19i1.1722

3	Responsiveness	**3.571**	**3.471**	**0.100**	high quality and high satisfaction	
4	Assurance	**3.747**	**3.556**	**0.191**	high quality and high satisfaction	
5	Emphaty	**3.571**	**3.750**	**-0.179**	low quality and low satisfaction	

Source: (Laurichela, 2021)

4. Analytical Hierarchy Process (AHP)

The next step is to use the Analytic Hierarchy Process (AHP) to assign weights to the research attribute gaps. The data used to establish the weight of each attribute is based on responses to a weighting questionnaire completed by respondents of The Healthy Catering's management, specifically the President Director, Operations Manager, and Human Resources Manager.

5. Calculation of Weighting

The weight assigned to each criterion is calculated by comparing the relative importance of the criteria (pairwise comparison). Based on the distribution of weighting questionnaires to management, pair comparison values are calculated. Table 2 displays the results of the calculations for the comparison matrix between criteria.

Table 2. Matrix of comparison among criteria

	Tangibles	Reliability	Responsiveness	Assurance	Emphaty
Tangibles	1.000	1.442	0.585	0.143	0.693
Reliability	0.693	1.000	0.822	0.131	1.442
Responsiveness	1.710	1.216	1.000	0.598	0.693
Assurance	7.000	7.612	1.671	1.000	5.593
Emphaty	1.442	0.693	1.442	0.179	1.000
Total	**11.846**	**11.964**	**5.520**	**2.051**	**9.422**

Source: (Laurichela, 2021)

The matrix normalization calculation is then performed by dividing each element by the total element.

Table 3. Normalization of the paired comparison matrix

	Tangibles	*Reliability*	*Responsiveness*	*Assurance*	*Emphaty*	*Eigen Vektor*
Tangibles	0,084	0,121	0,106	0,070	0,074	0,091
Reliability	0,059	0,084	0,149	0,064	0,153	0,102
Responsiveness	0,144	0,102	0,181	0,292	0,074	0,158
Assurance	0,591	0,636	0,303	0,487	0,5944	0,522
Emphaty	0,122	0,058	0,261	0,087	0,106	0,127
Total	**1**	**1**	**1**	**1**	**1**	**1**

Source: (Laurichela, 2021)

6. Consistency Test of Pairwise Comparison Matrix Among Criteria at the Consistency Stage

In order to find the maximum lamda value, a consistency test is run. Tables 2 and 3 are examples of this. The eigence vector is multiplied by the total of each alignment comparison matrix member. The multiplication results are then summed to get a maximum lambda of 5.43318. The matrix includes 5 primary criteria, which result in the following consistency index (CI) value:

$$CI = \frac{\lambda max - n}{(n-1)} = \frac{5.43318 - 5}{(5-1)} = 0.108295$$

CR = CI / RI
= 0.108295 / 1,12
= 0.0966

The computation results obtain a CR value of 0.1 (10%), indicating that the assessment preference is consistent.

7. Consistency Test of Pairwise Comparison Matrix Among Sub-criteria

The following is the eigenvector calculation in the comparison matrix among sub-criteria on the Empathy criteria.

Table 4. Pairwise comparisons with normalized eigen vectors

1.000	3.302	1.747	X	0.509	=	1.527
0.303	1.000	0.275		0.126		0.379
0.572	3.634	1.000		0.365		1.094

Source: (Laurichela, 2021)

The previous multiplication results are then combined to provide 3.05833.

$$CI = \frac{\lambda max - n}{(n-1)} = \frac{3.05833 - 3}{(3-1)} = 0.029165$$

CR = CI / RI
= 0.029165 / 0.58
= 0.053027272

The assessing preference is consistent because the CR < 0.1 (10%). When the Consistency Ratio (CR) for each comparison matrix is calculated, it is evident that the CR value for each comparison matrix among sub-criteria for each criterion is < 0.1). These findings suggest that management's assessment of each sub-criteria is consistent,

Vol. 19, No. 1 March 2022 | DOI: https://doi.org/10.33480/techno.v19i1.1722

implying that the respondent's assessment is compatible with actual conditions.

8. Global Weight Calculation

According to the analysis's results, the Assurances criteria have the highest weight value at 0.522197516, while Tangibles criteria have the lowest weight value of 0.090826882.

9. Calculation of Weighted Servqual Value

Attribute Weighted Servqual Value A1 = 0.200 x 0.027093539 = 0.005418708.. According to the analysis's results, the Responsiveness Dimension has the highest average weighted gap, with a weighted mean of 0.0075806, followed by the Tangibles Dimension's weighted mean gap of 0.00177888, and then the Reliability Dimension's weighted mean gap of -0.0029413, the Assurance Dimension's weighted mean gap of -0.0057629, and last, the Empathy Dimension's weighted mean gap of -0.0117

Attribute values with gap values greater than 0 indicate that the service qualities provided were sufficient to meet consumer expectations. However, these outcomes still need to be enhanced to meet consumer expectations truly. Meanwhile, a positive gap indicates that consumer expectations are lower than perceptions. Consumer expectations have not been met by the negative quality gap. It also indicates that the service provider has met consumer expectations. As a result, management should maintain or improve service performance on these qualities. AHP is used to calculate the weight for each criterion. Finally, the weighted gap value is calculated by multiplying the gap value of each characteristic by each weight. The weighted gap value with the most significant attribute gap value is shown in Table 5.

Table 5. Highest Weighted Servqual Value

No	Variable	Sub Criteria Weight	GAP	Weighted Gap	Priority
1	E1	0.0645849	-0.375	-0.024219	1
2	R2	0.0060114	-0.350	-0.002103	5
3	R4	0.0511680	-0.338	-0.017269	2
4	E3	0.0462595	-0.275	-0.012721	3
5	R1	0.00520818	-0.263	-0.00136	6
6	RV2	0.02129848	-0.150	-0.003194	4

Source: (Laurichela, 2021)

The authors provide the following recommendations for improvements to The Healthy Catering based on the results of the weighted gap analysis assessment:

a. E1(Waitress serves with a greeting, smile, and friendly and attentive).

b. R4 (Delivery of catering food on time).

c. E3 (Truthfulness in prioritizing the interests of consumers).

d. RV2 (Employees are willing to answer complaints and suggestions from consumers).

e. R2 (Consumers can request certain foods).

f. R1 (Using environmentally friendly packaging)

CONCLUSION

Researchers can draw several conclusions based on the research findings, including the following: 1) Obtained five criteria based on the dimensions of service quality, namely Reliability, Empathy, Tangibles, Assurance, and Responsiveness; 2) The gap between perception and expected value was calculated using the Servqual method, with the six highest gap values being E1 (-0.375), R2 (-0.350), E3 (-0.275), and R1 (-0.263) attributes; 3) obtained the weight value for each criterion and sub-criteria calculated by the AHP method; 4) The weighted Servqual calculation shows the results in the form of priority improvements for attributes that have a large gap; 5) The priority in repair from the highest priority to the lowest are E1, R4, E3, RV2, and R2 attributes.

The researchers recommend that 1) businesses implement the recommendations for improving service attributes in this study to improve service quality at The Healthy Catering; 2) that subsequent researchers develop research by adding analysis to other gaps, not limited to the gaps studied in this study, or only limited to the gap between consumer expectations and perceptions.

REFERENCE

Ammarapala, V. (2017). Servqual Model and Analytic Hierarchy Process on the Expressway Service Quality Assessment. *Panyapiwat Journal*, 122–136.

Anggraini, D. A. (2021). *Peluang Usaha Katering*. https://economy.okezone.com/read/2021/01/11/455/2342616/peluang-usaha-

Vol. 19, No. 1 March 2022 | DOI: https://doi.org/10.33480/techno.v19i1.1722

katering-agar-dapat-cuan?page=2

Berlian, E. (2016). *Metodelogi Penelitian Kualitatif & Kuantitatif*. Sukabina Press.

Efendi, M., Harianto, W., & Nugraha, D. A. (2019). *Penerapan metode servqual dan ahp sebagai analisis kualitas pelayanan terhadap kepuasan konsumen bengkel akena malang*. *x*(x), 1–9.

Fatihudin, D., & Firmansyah, A. (2019). *Pemasaran Jasa (Strategi, Mengukur Kepuasan & Loyalitas Pelanggan)* (Pertama). Deepublish.

Indrasari, D. M. (n.d.). *Pemasaran dan Kepuasan Pelanggan* (Pertama). Unitomo Press.

Mahmudi, A. A. (2021). Integrasi servqual dan ahp untuk evaluasi kualitas layanan dekranasda. *SAINTEKBU: Jurnal Sains Dan Teknologi*, *13*(01), 8–18.

Putri, A. E. R., Harianto, W., & Aziz, A. (2020). *PENILAIAN KEPUASAN PELANGGAN TERHADAP KUALITAS LAYANAN X DENGAN METODE SERVQUAL DAN ANALYTICAL HIERARCHY PROCESS*. *2*(3), 202–208.

Riadikemas. (2021). *Kiat Sukses Manangkap Bisnis Usaha Katering*.

Sikumbang, E. D. (2017). Analisa Tingkat Kepuasan Pelanggan Dengan Metode Fuzzy Servqual. *Jurnal Teknik Komputer AMIK BSI (JTK)*, *III*(1), 37–43.

Simanjuntak, L. S., Sagala, J. R., & Gea, A. (2018). Implementasi Sistem Pendukung Keputusan Dengan Metode AHP Dalam Menentukan Tingkat Kepuasan Pelanggan. *Jurnal Armada Informatika*, *2*(1), 76–88. https://doi.org/10.36520/jai.v2i2.34

Supartiningsih, S. (2017). Kualitas Pelayanan an Kepuasan Pasien Rumah Sakit: Kasus Pada Pasien Rawat Jalan. *Jurnal Medicoeticolegal Dan Manajemen Rumah Sakit 10.18196/Jmmr.2016*, *6*(1), 9–15. https://doi.org/10.18196/jmmr.6122

Suryani, & Hendryadi. (2016). *Metode Riset Kuantitatif* (p. 194). Prenadamedia Group.

Syafitr, L. S., & Herlawati. (2016). *Penilaian Kualitas Pelayanan Digital Lounge Menggunakan Metode Servqual Dan Analitycal Hierarchy Process*. *3*(1), 73–84.

Zaidiah, A., Astriratma, R., & Seta, H. B. (2021). *ANALISIS KUALITAS LAYANAN E-LEARNING DENGAN METODE SERVICE QUALITY (SERVQUAL) DAN ANALYTICAL HIERARCHY PROCESS (AHP)*. *23*(1), 46–59.

Vol. 19, No. 1 Maret 2022 | DOI: https://doi.org/10.33480/techno.v19i1.2980

ANALYSIS SENTIMENT ON THE ACCEPTANCE OF CPNS 2021 ON TWITTER SOCIAL MEDIA USING TEXTBLOB

Widi Astuti[1]; Elly Firasari[2] ; F. Lia Dwi Cahyanti[3]; Fajar Sarasati[4] ;

[1,4] Bisnis Digital; [2,3]Sistem Informasi;
[1,2,3,4] Universitas Nusa Mandiri, Jakarta, Indonesia
[1,2,3,4] www.nusamandiri.ac.id
widiastuti.wtu@nusamandiri.ac.id[1]; elly.efa@nusamandiri.ac.id[2] ; flia.fdc@nusamandiri.ac.id[3];
fajar.fss@nusamandiri.ac.id[4];

Abstract— *Information technology is developing rapidly with the development of hardware and hardware developed by the world's largest companies. These advances have a significant impact on human life. Many jobs in human life use the help of technology. data mining technology, one of which is used in the field of research. Data mining extracts valuable information by analyzing the presence of certain patterns or relationships from large amounts of data. Indonesian government agencies routinely organize the recruitment and selection of Candidates for Civil Servants (CPNS). Almost every year the government opens CPNS formations and it is never empty of applicants. Prospective Civil Servants (abbreviated as CPNS) are employees who have just passed the first stage of the selection test for Candidates for Civil Servants. Minister for Administrative Reform and Bureaucratic Reform (PANRB) Tjahjo Kumolo said the government would eliminate the recruitment of prospective civil servants (CPNS) in 2020. In order to find out the enthusiasm of the people in Indonesia, this study took data from reviews on the Twitter application with the topic of CPNS Acceptance. 2021. Using Textblob sentiment analysis, the sentiment class used was positive, negative and neutral sentiment class, and the data mining method used was Logistic Regression.*

Keywords *: Data mining, CPNS acceptance, Textblob, SVM, Logistic Regression.*

Intisari— Teknologi informasi berkembang pesat dengan pengembangan perangkat keras dan perangkat keras yang dikembangkan oleh perusahaan terbesar di dunia. Kemajuan tersebut memberikan dampak yang signifikan terhadap kehidupan manusia. Banyak pekerjaan dalam kehidupan manusia menggunakan bantuan teknologi. teknologi data mining salah satunya yang digunakan dalam bidang penelitian. Data mining mengekstrak informasi berharga dengan menganalisis keberadaan pola atau hubungan tertentu dari sejumlah besar data. Instansi pemerintah Indonesia secara rutin menyelenggarakan rekrutmen dan seleksi Calon Pegawai Negeri Sipil (CPNS). Hampir tiap tahun pemerintah membuka formasi CPNS dan tidak pernah sepi pelamar. Calon Pegawai Negeri Sipil (disingkat CPNS) adalah pegawai yang baru lulus tes seleksi penerimaan Calon Pegawai Negeri Sipil tahap pertama. Menteri Pendayagunaan Aparatur Negara dan Reformasi Birokrasi (PANRB) Tjahjo Kumolo mengatakan pemerintah akan meniadakan rekrutmen calon pegawai negeri sipil (CPNS) tahun 2020. Dalam rangka ingin mengetahui antusiasme masyarakat di indonesia penelitian ini mengambil data dari ulasan-ulasan di

Vol. 19, No. 1 Maret 2022 | DOI https://doi.org/10.33480/techno.v19i1.2980

aplikasi Twitter dengan topik Penerimaan CPNS 2021. Menggunakan sentimen analisis Textblob kelas sentimen yang digunakan kelas sentimen positif , negatif dan netral, dan metode data mining yang digunakan adalah Logistic Regression.
Kata Kunci: Klasifikasi, Penerimaan CPNS, Textblob, Logistic Regression.

PENDAHULUAN

Instansi pemerintah di Indonesia secara berkala menyelenggarakan rekrutmen dan seleksi Calon Pegawai Negeri Sipil (CPNS). Hampir tiap tahun pemerintah membuka formasi CPNS dan tidak pernah sepi pelamar. Calon Pegawai Negeri Sipil (disingkat CPNS) adalah pegawai yang baru lulus seleksi penerimaan calon pegawai negeri sipil tahap pertama. Calon PNS tidak memenuhi kewajiban 100% gaji PNS. Anda akan menerima persentase 80% berdasarkan CPNSSK yang ditentukan berdasarkan hukum yang berlaku di Indonesia. Namun, pada saat seleksi akreditasi CPNS 2018, hanya 128.000 dari 1.700.000 pendaftar yang lolos dinilai gagal, sehingga tidak memenuhi jumlah individu yang layak menduduki jabatan pemerintahan. Kami masih memiliki beberapa tes seleksi kompetisi lapangan (SKB), sehingga kami telah sepenuhnya lulus seleksi CPNS oleh masing-masing institusi. (Nurafni Kusumawardhani, 2019). Selain itu, proses seleksi CPNS di Indonesia juga dinilai sangat buruk, dengan pendaftaran yang kompleks dikombinasikan dengan pilihan tradisional daripada berdasarkan keahlian dan kompetensi secara keseluruhan, sehingga mengakibatkan korupsi, kolusi dan nepotisme. (Sulaiman, 2021).

Indonesia merupakan pengguna aktif media social, berdasarkan We are social dalam tahun 2020, Indonesia mencapai nomor 175,4 Juta orang pengakses internet. Dan Twitter adalah galat satu platform media yg paling poly pada pakai sang warga Indonesia. Menurut asal berdasarkan We are Social & Hootsuite tahun 2020, twitter menepati urutan ke lima pada kategori yang paling sering di gunakan dengan jumlah presentasi 56% setelah Youtube, Whatsapp, Facebook dan Instagram (We Are Social dan Hootsuite, 2020). Berdasarkan Twitter merupakan Sebuah situs web yang menyediakan layanan microblogging online yang memungkinkan pengguna untuk berbagi konten yang saat ini dibatasi hingga 280 karakter. (Informatika, 2018). Fitur tweet yang sering digunakan untuk menuliskan pemikiran, serta opini oleh para pengguna platform yang dapat digunakan penulis dalam pengolahan informasi yang sangat berguna. Adanya perbedaan pendapat tersebut sangat penting dan juga merupakan salah satu pengaruh terpenting pada perilaku manusia untuk mendapatkan hasil dari emosi.

Dengan latar belakang tersebut, penulis membahas tentang analisis opini publik tentang penerimaan CPNS di Twitter di media sosial. Hal ini sangat menarik untuk ditelaah dan menemukan opini publik tentang antusiasme masyarakat di tengah pandemi yang sedang berlangsung, apakah reaksinya positif, negatif atau netral.

BAHAN DAN MODEL

A. Teknik Pengumpulan Data

Ada beberapa cara untuk mengumpulkan data:

1. Crawling data
   Crawling data adalah proses mengambil atau mengunduh data dari server Twitter dalam format data pengguna dan tweet menggunakan antarmuka pemrograman aplikasi (API) Twitter. (Brata Mas Pintoko, 2018). Objek pada penelitian ini yaitu Penerimaan CPNS 2021 yg sedang berlangsung pada Indonesia. Total lebih menurut 3500 data.

2. *Preprocessing*
   Pada tahap ini dilakukan tindak lanjut dari data yang di tarik dengan merapikan data, sortir data, membersihkan data dari duplikat sampai data siap di gunakan untuk proeses seanjutnya. Data yang dibersihkan adalah data yang Masih banyak simbol dan kata-kata yang tidak perlu (Brata Mas Pintoko, 2018).
3. Sentimen Analysis menggunakan *Textblob*
   Sentiment Analisis adalah metode untuk memahami dan mengekstraksi data opini secara otomatis dan memproses data tekstual untuk menangkap emosi yang terkandung dalam opini tersebut. Pengelompokan data analisis sentimen ini memiliki beberapa pendapat dan jawaban, termasuk positif, negatif, dan netral. (Nova Tri Romadloni, Imam Santoso, 2019). Analisis sentimen juga dapat disebut sebagai suatu proses untuk menemukan suatu makna dari perilaku, opini, pandangan, dan emosi dari suatu teks, perkataan, *tweets*, dan *database* dengan sumber dari *Natural Language Processing* (NLP) (Hernikawati, 2021).
   Sedangkan Textblob adalah salah satu library Python2 dan Python3 yang bisa Anda gunakan untuk mengolah data teks. Textblob mudah diakses dan dapat digunakan untuk pembuatan prototipe cepat (Hernikawati, 2021).
4. Modelling
   Pemodelan dilakukan dengan tujuan untuk menguji keakuratan prediksi sistem berdasarkan data model yang dibuat. (Andini, 2021). klasifikasi dengan menggunakan algoritma Logistic Regression. Hasil dari proses klasifikasi kemudian dibandingkan untuk mengetahui algoritma mana yang terbaik akurasinya.

Saat mengembangkan metode penelitian eksperimental menggunakan CRISPDM, metode penelitian standar yang digunakan dalam penambangan data, terdiri dari enam tahap: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, dan evaluasi. , Dan ada langkah-langkah penyebaran. (Binsar, 2020). Model yang digunakan dalam penelitian ini antara lain:

1. *Business Understanding*
   Pada tahap ini dilakukan pemahaman terhadap objek penelitian yang dilakukan dengan menemukan informasi melalui media social twitter tentang Penerimaan CPNS 2021 yang sedang berlangsung di Indonesia mengekspresikan berbagai macam pendapat baik negatif maupun positif pada tweet pengguna media sosial.
2. *Data Understanding*
   Pada tahap data understanding adalah proses memahami data yang digunakan sebagai bahan yang diselidiki dan memungkinkan Anda untuk melanjutkan ke langkah berikutnya, pra-pemrosesan. Pada fase ini proses pengambilan data mentah berjalan sesuai dengan atribut yang dibutuhkan. Data yang digunakan dalam survei ini berasal dari kolom kicauan di media sosial Twitter. Di bawah ini adalah Tabel 1 dari data yang diambil selama fase ini.

Tabel 1. Proses Data Understanding

Date	***Tweet***
04 Juli 2021	'Seleksi CASN Kemenparekraf/ Baparekraf mulai masa pendaftaran 2-21 Juli 2021 #CPNS2021 https://t.co/KF2uRLbamu',Disporapar Jateng #JatengGayeng
05 Juli 2021	'Buruan! Pendaftaran CASN Pesisir Selatan Tahun 2021 Telah Dibuka https://t.co/ajZp0f7I9c',Harian Haluan

Vol. 19, No. 1 Maret 2022 | DOI https://doi.org/10.33480/techno.v19i1.2980

Date	*Tweet*
06 Juli 2021	'LINK Pendaftaran CASN Kemenparekraf Tahun 2021 Dengan 186 Jumlah Kebutuhan Formasi Dari Berbagai Lulusan https://t.co/xQnT22pWqq',Portal Kudus - Pikiran Rakyat
07 Juli 2021	'Pendaftaran Seleksi CASN Kemenag Dibuka Hingga 21 Juli 2021 https://t.co/prEsA9OiBs',radarbangsa.com

Sumber: (Astuti, 2022)

3. Data Preparation
   Tahap data preparation merupakan tahap dengan proses penyiapan data digunakan dalam data dibersihkan dan Siap dapatkan digunakan dalam penelitian. Dalam text mining tahapan awal yang akan dilakukan adalah tahap text preprocessing. Berikut merupakan langkah yang dilakukan dalam text preprocessing:
4. Modelling
   Dalam sesi Modelling ini hendak dicoba metode pengklasifikasian informasi yang sangat akurat. Buat menyamakan ataupun mengkomparasi, pada riset ini hendak digunakan algoritma SVM serta Logistic Regression.
5. Evaluation
   Model yang terbentuk akan diuji menggunakan confusion matrixyang akan mengetahui tingkat akurasi. Confusion matrixakan menggambarkan Hasil akurasi berkisar dari prediksi positif positif, prediksi positif palsu, prediksi negatif positif, dan prediksi negatif palsu. Akurasi dihitung dari semua prediksi yang benar (baik prediksi positif maupun negatif). Dibandingkan dengan seluruh data testing. Semakin tinggi nilai akurasi, semakin baik pula model yang dihasilkan. Pengujian juga diukur dengan menggunakan ROC Curve. ROC Curve akan menggambarkan kelas positif dalam bentuk kurva. Pengujian dilakukan dengan menghitung nilai area under the curve (AUC), dan semakin tinggi nilai kurva AUC dan ROC maka akan semakin baik model klasifikasi yang terbentuk.
6. Deployment
   Sesi ini merupakan sesi terakhir dari CRISP DM dan merupakan hasil dari semua tahapan yang sebenarnya digunakan sebelumnya. Maknanya merupakan melaksanakan suatu bersumber pada pengetahuan yang didapatkan dari aktivitas mining terhadap informasi. Pelaksanaan dalam riset ini hendak dibesarkan dengan PHP serta MySQL.

HASIL DAN PEMBAHASAN

a. Proses @#Annotation removal adalah proses untuk menghilangkan teks yang memiliki anotasi @ dan #. Proses preprocessing ini dilakukan dengan menggunakan python yaitu menggunakan regular expression perintah re.sub("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," ", text). Sehingga tweet dengan mention user dan hashtag hilang. Tabel 2 menunjukan perbandingan teks sebelum dan sesudah dilakukannya proses @#Annotation Removal.

Tabel 2. Perbandingan teks sebelum dan sesudah proses @Annotation Removal

@#Annotation Removal	
Data sebelum	Data sesudah
#'@himynameisokky Pertanyaan seputar Pendaftaran CASN 2021 agar disampaikan melalui WA Helpdesk 087884112321',BKPSDM_PURWAKARTA	pertanyaan seputar pendaftaran casn agar disampaikan melalui wa helpdesk,BKPSDM_PURWAKARTA,

Sumber: (Astuti, 2022)

b. Tokenization: Regexp

Vol. 19, No. 1 Maret 2022 | DOI: https://doi.org/10.33480/techno.v19i1.2980

Proses Tokenization merupakan data tweet pada setiap kata akan dipisahkan berdasarkan berdasarkan spasi yang ditemukan(Brata Mas Pintoko, 2018). Tokenization peneliti gunakan untuk memisahkan kata atau huruf dari tanda baca dan simbol (Eka Rini Yulia, 2021). Regexp adalah proses untuk menghilangkan tanda baca dan angka sehingga hasilnya adalah kata. Proses preprocessing ini dilakukan dengan menggunakan library Regexp di python. Tabel 3 menunjukan perbandingan teks sebelum dan sesudah dilakukannya proses Tokenization: Regexp.

Tabel 3. Hasil Proses Tokenization

Tokenization: Regexp	
Data sebelum	Data sesudah
'@Kota_Tangerang Hallo min, mau tanya. Untuk menanyakan terkait pendaftaran CASN kemana ya? Adakah nomor yang bisa dihubungi. Terima kasih'"	tangerang hallo min mau tanya untuk menanyakan terkait pendaftaran casn kemana ya adakah nomor yang bisa dihubungi terima kasih

Sumber: (Astuti, 2022)

c. Indonesian Stemming

Proses Indonesian Stemming adalah Konversi data tweet dari preposisi ke kata dasar (Brata Mas Pintoko, 2018). Proses preprocessing ini dilakukan dengan menggunakan library python Sastrawi. Tabel 4 menunjukkan perbandingan teks sebelum dan sesudah proses stemming bahasa Indonesia..

Tabel 4. Proses Indonesian Stemming

Indonesian Stemming	
Data sebelum	Data sesudah
Buat SobatBKN, yang punya banyak pertanyaan seputar pendaftaran CASN 2021 pada portal SSCASN, akan ada sesi tanya jawab live pada akun Instagram BKN BKNgoidoffcial, di waktu yang tertera berikut Jangan sampai kelewatan ya	buat sobatbkn, yang punya banyak tanya putar daftar casn 2021 pada portal sscasn, akan ada sesi tanya jawab live pada akun instagram bkn bkngoidoffcial, di waktu yang tera ikut jangan sampai lewat ya

Sumber: (Astuti, 2022)

d. *Indonesian Stop Word Removal*

Proses Indonesian Stop word removal adalah Proses menghilangkan kata-kata umum yang biasanya terlihat dalam jumlah banyak dan dianggap tidak berarti. Beberapa contohnya adalah dg, rt, with, ny, d, klo, if, amp, let, make, say, no, ga, krn, nya, ya, si, know, no, uh, for, yes, And soon. Proses prapemrosesan ini dilakukan menggunakan pustaka stopword Python NLTK. Tabel 5 menunjukkan perbandingan teks sebelum dan sesudah prosedur penghapusan stopword bahasa Indonesia. Tabel 5. Hasil Stop Word Removal

Indonesian Stop Word Removal	
Data sebelum	Data sesudah
BKNgoid halo selamat siang admin BKN saya mau nanya untuk pendaftaran akun seleksi CASN untuk tempat lahir	@bkngoid halo selamat siang admin bkn nanya pendaftaran akun seleksi casn lahir ktp dgn ijazah berbeda. ktp

Vol. 19, No. 1 Maret 2022 | DOI https://doi.org/10.33480/techno.v19i1.2980

saya antara ktp dgn ijazah berbeda. Ktp depok ijazah bogor. Kira-kira bagaimana ya solusinya? Terima kasih, Tyas Murti Lestari	depok ijazah bogor. solusinya? terima kasih, tyas murti lestari

Sumber : (Astuti, 2022)s

e. N Chars Filter

Proses N Chars Filter yaitu untuk menghapus kata yang kurang dari satu suku kata. Tabel 6 Menampilkan perbandingan teks sebelum dan sesudah proses penyaringan N-karakter.

Tabel 6. Hasil Nchars Filter

N Chars Filter	
Data sebelum	Data sesudah
b'Segera Daftar! Pemkab Pessel Buka Pendaftaran CASN Tahun 2021 https://t.co/ZLsD8Q4oDr',BERITAMINANG	segera daftar pemkab pessel buka pendaftaran casn tahun,BERITAMINANG

Sumber: (Astuti, 2022)

F. Stop Word Removal

Pada tahap ini, proses penghilangan stopword adalah untuk mengecek apakah stopword tersebut mengandung sebuah kata. Proses ini menghilangkan istilah-istilah yang tidak berarti dan tidak relevan. Istilah yang diambil dari tingkat tokenisasi diperiksa dalam daftar stopword. Jika sebuah kata dimasukkan dalam daftar stopword, kata itu tidak akan diproses.

G. Textblob Analysis Sentiment

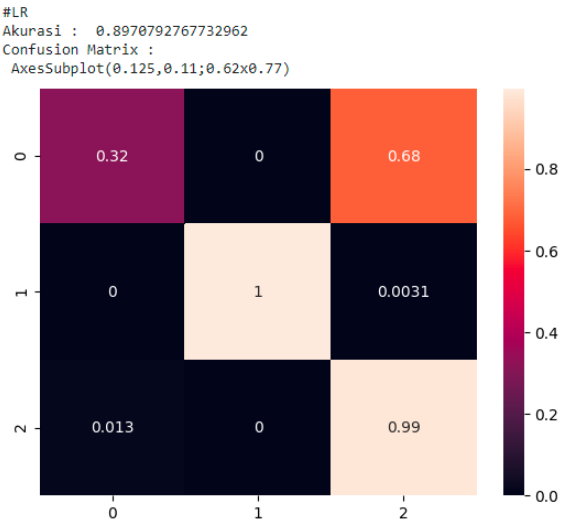
Pada proses ini yaitu menghitung sentiment dari setiap tweet untuk memproses data tekstual. Pada tahap ini menggunakan teknik Natural Language Processing (NLP) dan data tweet umumnya tergolong kedalam nsupervised Learning. Diperlukan NLP untuk mengidentifikasi opini dan sentimen dan mengklasifikasikannya menjadi label positif, negatif, atau netral. Library python yang digunakan untuk identifikasi analisis sentimen pada penelitian ini adalah Textblob. Dan dari masing-masing tweet tersebut ditetapkan polaritasnya apakah bermakna positif, negatif, atau netral.

Tabel 7. Klasifikasi Identifikasi Analisis Sentimen *Textblob*

Sentiment	Count	Percentage
Positive	1.527	42,5%
Neutral	1.597	44,4%
Negative	471	13,1%
Total	3.595	100%

Sumber: (Astuti, 2022)

Gambar 1. Hasil Klasifikasi Identifikasi Analisis Sentimen *Textblob* dibawah ini.



Gambar 1. Hasil Klasifikasi Logistic regression

Sumber: (Astuti, 2022)

Akurasi yang dicapai adalah 89,70% data analisis sentimen positif, negatif, dan netral terkait Asumsi CPNS 2021. 32% data tweet negatif sesuai dengan ekspektasi. Data tweet negatif yang termasuk dalam prediksi positif adalah 68%, dan data tweet negatif yang diprediksi netral adalah 0%. Tweet

negatif berisi 0,01% data tweet positif, 99n data tweet positif sesuai ekspektasi, 100% data tweet netral sesuai ekspektasi, dan 0,0031% tweet netral yang masuk positif.

Kesimpulan

Hasil penelitian ini telah ditemukan kesimpulan berdasarkan data yang telah di dapatkan dari media sosial twitter mengenai Penerimaan CPNS 2021 pada masyarakat indonesia adalah sebagai berikut 42,5% masyarakat bersentimen Positif, dan 44,4% bersentimen Netral dan 13,1% bersentimen Negatif. Kesimpulan ini di buktikan dengan keakuratan pengujian metode data mining Logistic Regression dengan akurasi sebesar 89,70%. Dari hasil tersebut dapat diartikan bahwa masyarakat indonesia cenderung bersentimen positif terhadap berita penerimaan CPNS.

REFERENSI

Andini, G. R. dan R. (2021). Analisis Respon Masyarakat Pada Platform Media Sosial Twitter Terhadap Tokoh Politik, Jenderal TNI (PURN.) Gatot Nurmantyo. *Jurnal Ilmiah Akuntansi Dan Keuangan*, *4*(Vol. 4 No. 2 (2021): FairValue : Jurnal Ilmiah Akuntansi dan Keuangan), 12. https://doi.org/https://doi.org/10.32670/fairvalue.v4i2.650

Astuti, W. et all. (2022). ANALYSIS SENTIMENT ON THE ACCEPTANCE OF CPNS 2021 ON TWITTER SOCIAL MEDIA USING TEXTBLOB, *17*(1), 1–6.

Binsar, T. M. dan F. (2020). Cross-Industry Standard Process for Data Mining (CRISP-DM).

Brata Mas Pintoko, K. M. L. (2018). Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naive Bayes Classifier. e-proceeding of Engineering.

Eka Rini Yulia, K. S. (2021). Implementasi Particle Swarm Optimization (PSO) pada Analysis Sentiment Review Aplikasi Trafi Menggunakan Naive Bayes (NB). Jurnal Teknik Komputer AMIK BSI.

Hernikawati, D. (2021). Kecenderungan Tanggapan Masyarakat Terhadap Vaksin Sinovac Berdasarkan Lexicon Based Sentiment Analysis The Trend of Public Response to Sinovac Vaccine Based on Lexicon Based Sentiment Analysis. *Jurnal Ilmu Pengetahuan Dan Teknologi Komunikasi*, *23*(1), 21–31.

Informatika, K. K. dan. (2018). *Memaksimalkan Penggunaan Media Sosial dalam Lembaga Pemerintah*. jakarta: Dorektorat Jenderal Informasi dan Komunikasi Publik, Kementerian Komunikasi dan Informatika.

Nova Tri Romadloni, Imam Santoso, S. B. (2019). Perbandingan Metode Naive Bayes , KNN, dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line. Jurnal IKRA-ITH Informatika.

Nurafni Kusumawardhani, A. dan R. L. (2019). Problematika Seleksi CPNS 2018 dalam Pengangkatan CPNS yang tidak Memenuhi Passing Grade. Civil Service.

Sulaiman, dan D. R. (2021). Opini Peserta Terhadap Transparansi Penerimaan CPNS Melalui Metode Computer Assisted Test (CAT). Jurnal Eksos.

We Are Social dan Hootsuite. (2020). Indonesian Digital Report 2020.

SENTIMENT ANALYSIS ON TWITTER SHOPEECARE USING NAIVE BAYES, ADABOOST, AND SVM (EVOLUTION) ALGORITHM COMPARATIVE METHODS

Rizky Nugraha Pratama [1*]; Ghina Amanda Kamila [2]; Kresna Lazani T [3]; Ilham Fauzi [4]; Muhammad Reynaldo Oktaviano [5]; Dedi Dwi Saputra [6]

Program Studi Sistem Informasi[1,2,3,4,5,6]
Universitas Nusa Mandiri[1,2,3,4,5,6]
https://nusamandiri.ac.id/
11213113@nusamandiri.ac.id [1*]; 11212986@nusamandiri.ac.id [2]; 11213091@nusamandiri.ac.id [3]; 11212960@nusamandiri.ac.id [4]; 11213088@nuamandiri.ac.id [5]; dedi.eis@nusamandiri.ac.id [6]

Abstract *— The growth of Indonesian e-commerce is increasing along with the growth of internet use in Indonesia. In 2015, there were 92 million internet users in Indonesia. One of the popular online shopping platforms in Indonesia is Shopee. The role of social media also does not escape the role of e-commerce players by utilizing one of them, namely social media twitter. the amount of customer enthusiasm in tweeting and Retweeting existing content, made us decide to research about Sentiment analysis on twitter social media accounts: Shopeecare uses the smote NB, ADboost, and SVM comparison methods. From the data, the comparison results from the test experiments used the Smote + Naive Bayes, Smote + Naive Bayes + Adaboost, and Smote + SVM models. It is known that the Accuracy, Precision, AUC values of the Smote + SVM algorithm are higher than other algorithms, namely Accuracy 76.24%, Precision 75.65%, AUC 0.822. From the results of the algorithm comparison, it shows that the algorithm in determining the sentiment of the complaint and not complaint analysis is better than other algorithms.*

Keywords *: Shopee, Twitter, Sentiment Analysis, Naïve Bayes, Adaboost, SVM.*

Intisari— Pertumbuhan e-commerce Indonesia meningkat seiring dengan tumbuhnya penggunaan internet di Indonesia. Pada tahun 2015, terdapat 92 juta pengguna internet di Indonesia. Salah satu platform belanja online yang popular di Indonesia adalah Shopee. peran media sosial juga tidak luput dari para pelaku e-commerce dengan memanfaatkan salah satunya yaitu media sosial twitter. banyaknya antusiasme pelanggan dalam men-*tweet* dan Retweet konten yang ada, membuat kami memutuskan untuk meneliti tentang Analisis sentimen pada akun media sosial twitter: *Shopeecare* menggunakan metode komparasi SMOTE NB, ADboost, dan SVM. Dari data hasil perbandingan dari percobaan pengujian menggunakan model Smote + Naive Bayes, Smote + Naive Bayes + Adaboost, dan Smote + SVM. Diketahui bahwa nilai *Accuracy, Precision,* AUC dari algoritma Smote + SVM lebih tinggi dari algoritma lainnya, yaitu *Accuracy* 76.24%, *Precision* 75.65%, AUC 0.822. Dari hasil komparasi algoritma menunjukan bahwa algoritma tersebut dalam menentukan sentimen analisis complain dan not complaint lebih baik daripada algoritma lain.

Kata Kunci : Shopee, Twitter, Sentimen Analisis, Naïve Bayes, Adaboost, SVM.

INTRODUCTION

Referring to data from a study entitled The Opportunity of Indonesia initiated by TEMASEK and Google, Indonesia's e-commerce growth is increasing along with the growth of internet use in Indonesia. In 2015, there were 92 million internet users in Indonesia. In 2020, it is predicted that Indonesian internet users will increase to 215 million users(Harahap 2018). Of the total number of internet users, in 2015, there were 18 million online shoppers in Indonesia. By 2025, 119 million people are predicted to become online shoppers in Indonesia. So it's not surprising that this increase will increase the value of the Indonesian e-commerce market. TEMASEK and Google predict that the value of the Indonesian e-commerce market will reach $81 billion by 2025 (Zaenudin 2017).

One of the most popular online shopping platforms in Indonesia is Shopee. Reporting from Idn Times, according to Ipsos, a global market research company based in Indonesia released the latest research results related to competition in the

e-commerce industry during the end of 2021. Based on the survey results, among the three main e-commerce players in Indonesia , namely Tokopedia, Shopee, and Lazada, it is known that Shopee ranks first(Perdana 2022).

Shopee is expanding its market in Indonesia by providing an easy shopping experience. The role of social media also does not escape the role of e-commerce players including shopee in marketing and informing their products and services. According to the daily compass (Syam 2022). Youtube, Whatsapp, Facebook, Instagram, TikTok and Twitter are the six most commonly used social media in Indonesia. Twitter is one of the most used social media because the features offered by twitter make it very easy for its users to spread information to other users. The like and retweet feature is one of twitter's mainstay features that is useful for spreading information to other users. Not only that, the trending topic feature is also the most widely used feature to see hot news that is being discussed, this feature works by summarizing keywords or hashtags that are being widely used by users in a region or throughout the world.

Based on previous research via twitter The results of the study for the Naïve Bayes Classification Algorithm for Sentiment Analysis of the Shopee Application resulted in an accuracy value of 71.50% and an AUC (Area Under Curve) Value of 0.500 (Masripah and Utami 2020). by looking at the amount of customer enthusiasm in tweeting and Retweeting existing content, we decided to research about sentiment analysis on twitter social media accounts: Shopeecare uses NB, Adaboost, and SVM comparison methods. This study is expected to find out the results of the percentage of accuracy between complaints or non-complaints contained in shopeecare user tweets.

MATERIALS AND METHODS

In this section, the system design in this study will be explained. The method was carried out through several stages. The data research that we used about 3 comparison methods, There's namely:
1. SMOTE + NB
2. SMOTE + NB + AdaBoost
3. SMOTE + SVM (Evolution)
The research flow can be explained through the diagram in Figure 1.

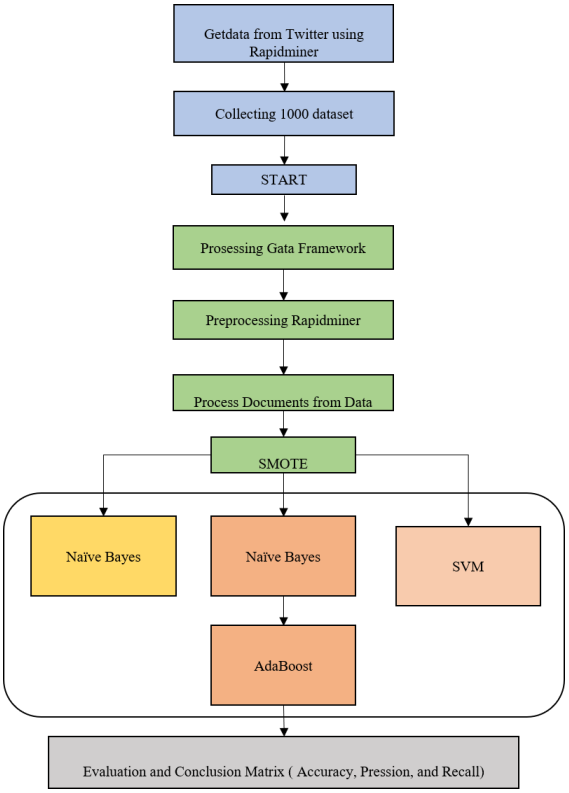


Figure 1. Research flow

As we can see about that flow we should take the data and collecting about 1000 dataset. After that the data should be analyze to complain and not complain. Next we can start the pre processing GataFramework and use the method of Smote combine of Naïve Bayes, AdaBoost, and SVM. Which is that three methods will be explain in this part.

1. Dataset

The data used in this study is data from the Twitter Shoppe Care Indonesia (https://twitter.com/ShopeeCare). The data used in this study are all the tweet data in the shopeecare account which contains reviews, criticisms, Services and comments on products and promos they were campaign in 2022. For the application we use is Rapidminer version 9.10.1. The Dataset Connection flow can be explained through the Figure 2.

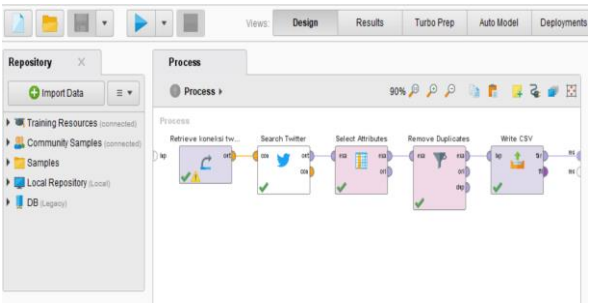


Figure 2. Dataset Connection

We took amounted about 1000 data which is tweets and commented are in Indonesian language and we follow that direction in the rapidminer application to gain all tweets and changes the following data obtained from twitter shopeecare. Firstly we should retrive connection with the link and username twitter that we want to analyzed. The Account twitter was @Shopeecare and after that we select the attribute data were setting in the currently date and filtered if the data contains duplicates tweets. Then we pick a location data where we want to saved. For the checking, we can click the Play button, it's for viewing the result if the following command works correctly. So finaly in the excel worksheet we should Analyze that to labeling sentences if contains complain or not complain. After that Process the Dataset is ready to the next Processing GataFrameworks part.

Table 1. Dataset Twitter Shopeecare

No	Text	Status
651	@ShopeeCare Min, akun dibatasi caranya biar kembali normal gimana?	not complaint
652	@ShopeeCare min, kalo akun dibatasi itu bisa kembali normal nggak? Kalo bisa caranya gimana?	not complaint
653	Selamat pagi @ShopeeID @ShopeePay_ID @ShopeeCare ,kenapa saya tdk bisa melakukan pembayaran via shopeepay ya? Pdhal saldo masih cukup, terima kasih https://t.co/0JOZIepIRZ	complaint
654	@ShopeeCare min cek dm	not complaint
655	@ShopeeCare Baca dm min	not complaint

2. Processing Gataframework

Gataframework is a framework for Indonesian text mining preprocessing that provides Indonesian stopword removal, Indonesian stemming, regular expressions(Riska Aryanti, Atang Saepudin, Eka Fitriani, Rifky Permana 2019), transform: URL removal and annotation removal. Within this framework, preprocessing stages can be performed on individual text and uploaded files. To use them, you can use the following link www.gataframework.com. Below is the Shopeecare Twitter data table processed by Gataframework first step is parting the dataset into a 50 data because the server is limited and then we should upload that excel file. The next part we separate the sentences with unneeded words, using tools :

1. @Anotation removal to remove text that has @annotations, such as: @shopee_ID, @shopeecare @shoppe_pay or refer to another account.
2. *transformation:remove URL* to remove links contained in the tweet itself, such as : http://t.co//0JOZIepIRZ, or other inserted link sources.
3. *Regexp to remove punctuation marks, Such as : ?!,.*
4. *indonesian stemming* is the process of turning the word affixed into a base word. Such as: how to Be how to be, How to be a Way, limited to being a limit.
5. *Indonesian stopword removal* is the process of removing unimportant words, such as : this, that is, if, the way.

For more details please see table 2.

Table 2. Shopeecare twitter dataset after *Gataframework* process

No	Regexp	Indonesian Stemming	Indonesian Stop word removal
651	min akun dibatasi caranya biar kembali normal gimana	min akun batas cara biar kembali normal gimana	min akun batas biar normal gimana
652	min kalo akun dibatasi itu bisa kembali normal nggak kalo bisa caranya gimana	min kalo akun batas itu bisa kembali normal nggak kalo bisa cara gimana	min kalo akun batas normal nggak kalo gimana
653	selamat pagi kenapa saya tdk bisa melakukan pembayaran via shopeepay ya pdhal saldo masih cukup terima kasih	selamat pagi kenapa saya tdk bisa laku bayar via shopeepay ya pdhal saldo masih cukup terima kasih	selamat pagi tdk laku bayar via shopeepay pdhal saldo terima kasih
654	min cek dm	min cek dm	min cek dm
655	baca dm min	baca dm min	baca dm min

3. Preprocessing Rapidminer

Rapidminer is software for data processing. Using data mining principles and algorithms, Rapidminer extracts patterns from large data sets by combining statistical methods, artificial intelligence and databases(Afifah Cahayani Adha

Vol. 19, No.1 Maret 2022 | DOI https://doi.org/10.33480/techno.v19i1.3086

2019). Preprocessing is the process of converting raw data into a form that is easier to understand.

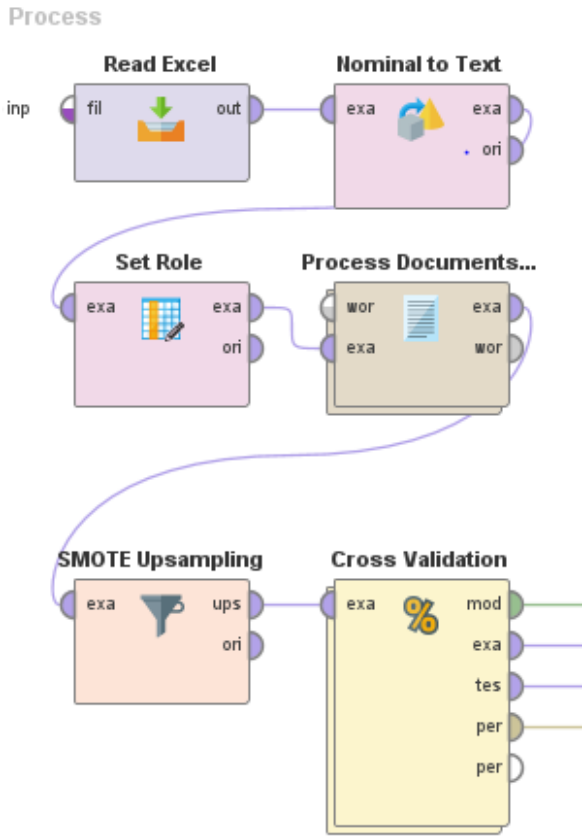


Figure 3. Preprocessing Rapidminer

In figure 3 shows the flow of the preprocessing process on the rapid miner, from this stage describes the steps starting from reading the dataset file, nominal to text is used to change the nominal attribute type to text, the role set is used to change the role of one or more attributes, this process document from data is used to extract information from documents with structured content, SMOTE Upsampling and coss validation.

4. Process Documents from Data

a. Transform Cases

In the transform cases stage, the goal of this stage is to change all letters in the dataset to all lowercase or all capital letters. In this study, the dataset will be converted to lowercase because the majority of the text is written, most of which are lowercase(Jaka 2015).

Table 3. Transform Cases

No	Document	Transform Cases
45	Beli baju di Solo Jateng, kirim ke Grobogan jawa tengah tapi kok sampai Kalimantan tengah di aplikasi nya	beli baju di solo jateng, kirim ke grobogan jawa tengah tapi kok sampai kalimantan tengah di aplikasi nya

b. Tokenize

The tokenize process serves to remove punctuation marks, symbols and non-letter characters in each dataset document(Riska Aryanti, Atang Saepudin, Eka Fitriani, Rifky Permana 2019). All unnecessary characters will be discarded including redundant white space and all punctuation marks.

Table 4. Tokenize

No	Document	Tokenize
8	sistemnya emang sering ngawur udah seringkali kejadian begini dan pasti ngeles 'setelah kami cek tidak ada kendala di sistem kami ya...' log kalian ampas	sistemnya emang sering ngawur udah seringkali kejadian begini dan pasti ngeles setelah kami cek tidak ada kendala di sistem kami ya log kalian ampas

c. Filter Tokens (by Length)

Filter tokens is the process of taking important words from the tokens (Langgeni, Baizal, and W 2010)In this process, words that have a certain length will be deleted.

d. Stemming

In the process of grouping words into groups that have the same basic word and performing transformations for the weighting process by calculating the presence or absence of a word in the document(Riska Aryanti, Atang Saepudin, Eka Fitriani, Rifky Permana 2019), with the aim that all words that have been selected as tokens in the previous stage will be converted into basic words.

Figure 4. Stemming

e. Stopword Removal

Stopword removal is the removal of irrelevant words, such as conjunctions and others(Riska Aryanti, Atang Saepudin, Eka Fitriani, Rifky Permana 2019), for example for, will, so, between, which are words that do not have their own meaning if they are removed and words related to adjectives related to sentiment analysis. At this stage will refine the previous dataset.

Table 5. Stopword Removal

No	Document	Stopword Removal
4	selamat pagi kenapa saya tdk bisa laku bayar via shopeepay ya pdhal saldo masih cukup terima kasih	selamat pagi tdk laku bayar via shopeepay pdhal saldo terima kasih

5. SMOTE

The Synthetic Minority Oversampling Technique (SMOTE) method is a popular method for dealing with imbalances. The SMOTE method is a development of the oversampling method, which is a technique that synthesizes new samples from minority classes to balance the dataset by resampling minority class samples(Kasanah, Muladi, and Pujianto 2019).

6. Cross Validation

Cross validation is a method used to get the best model. The explanation of the method used by this research is described as follows.

a. Naive Bayes

Naive Bayes is a method based on the Bayes theorem proposed by British scientist Thomas Bayes, this method belongs to a simple probabilistic classification algorithm that calculates a set of probabilities by adding up the frequencies and combinations of values from a given dataset. This algorithm assumes that object attributes are independent(Iswanto et al. 2021). The probabilities involved in generating the final estimate are calculated as the sum of the frequencies from the decision table. Here is the Bayes theorem equation(Iswanto et al. 2021):

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad \text{.....(1)}$$

Description :

X :Data with unknown class

H :Hypothesis data with a certain class

P(H|X) :Probability of hypothesis H based on condition X

P(H) :Hypothesis probability H

P(X) :Probability X

P(X|H) :Probability of X based on the conditions on the hypothesis H

b. AdaBoost

Adaboost or Adaptive Boosting was first introduced by Yoav Freund and Robert Schapire. AdaBoost algorithm itself is an acronym for Adaptive Boosting, this algorithm is widely applied to prediction models in data mining. The whole point of the AdaBoost algorithm is to give more weight to improper observations (weak classification)(Zulhanif 2015). To build a model in the supervised learning algorithm using 2 variables, namely the independent variable and the target variable. Adaboost utilizes boosting to improve predictor accuracy. Adaboost and its variants have been successfully applied to several fields (domains) due to their strong theoretical basis, accurate predictions, and great simplicity. The steps in the Adaboost algorithm are as follows:

1. Input : A collection of research samples on the label $\{(x_i, y_i), \ldots, (x_n, y_n)\}$ a component learn algorithm, the number of cycles T.
2. Initialize : Weight of a training sample $W_i^1 = 1/N$, for all $i = 1, \ldots, N$
3. Do for : $t = 1, \ldots, T$
4. Use component learn algorithms to train a classification component h_t on the training weight sample.
5. Count his training errors on h_t: $\varepsilon_t = \sum_{i=1}^{n} W_i^t, y_i \neq h_t(x_i)$
6. Assign weights to classifier components $h_t = a_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

Vol. 19, No.1 Maret 2022 | DOI https://doi.org/10.33480/techno.v19i1.3086

7. Training sample weight update $w_i^{t+1} \frac{w_i^t \exp\{-a_t y_i h_t(x_i)\}}{C_t}, i = 1, \dots, NC_t$ is a normalization constant.

8. Output:

$$f(x) = sign\,(\sum_{t=1}^{T} a_t\, h_t(x))$$

c. Support Vector Machine(Evolutionary)

SVM have been around since the 1960s, and this approach makes SVM a promising new method for classifying linear and nonlinear data. An SVM is an algorithm that transforms the original training data into higher dimensions in a new dimension using a non-linear mapping, and then searches for a linearly optimal hyperplane separator (i.e., the "" that separates tuples from one class to another decision boundary"). For sufficiently high dimensions, using a suitable non-linear mapping, the data from the two classes can always be separated by a hyperplane. SVM finds this hyperplane using support vectors ("basic" training tuples) and edges (defined by support vectors) (Muslim et al. 2018). The initial steps of an SVM algorithm are as follows(Ritonga and Purwaningsih 2018):

a) Training data is used for the training process. The training process uses SVM Multi-Class. Each classification model is trained by using the entire data, to find solutions to problems.

b) After training and producing a classification model, the next step is to test the classification model with data testing to determine the accuracy of the classification model.

c) After the model is generated from the classification, testing is carried out on data other than the dataset, testing using the resulting data.

d) Using the results of the implementation above, the researchers then analyzed and discussed the classification results generated by the Support Vector Machine (SVM) method.

7. Evaluasi

This study compares the accuracy level between the Smote + Naive Bayes, Smote + Naive Bayes + AdaBoost and Smote + SVM algorithm approaches which are evaluated using accuracy, precision, recall and AUC. The explanation of the evaluation is described as follows(Maragoudakis, Fakotakis, and Kokkinakis 2004)

a) Accuracy

Accuracy is an evaluation based on the number of correct prediction proportions. Accuracy can be measured by the following equation:

$$a = \frac{tp}{tp+ fp} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(2)$$

b) Precision

Precision is the level of accuracy between the information requested by the user and the answer given by the system. The precision formula is as follows:

$$p = \frac{tp}{tp+ fp} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(3)$$

c) Recall

Recall is the success rate of the system in retrieving information. The recall formula is as follows:

$$r = \frac{tp}{tp+ fn} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(4)$$

d) AUC

AUC is used to measure discriminatory performance by estimating the probability of the output that has been obtained from a randomly selected sample from a positive or negative population, the greater the AUC value, the stronger the resulting classification. Since AUC is part of the unit area of a square, the resulting value will always be the same as it generates, between 0.0 and 1.0(Wisdayani, Indah Manfaati Nur 2019)

Berikut panduan untuk mengklasifikasikan keakuratan diagnosa menggunakan AUC(Riska Aryanti, Atang Saepudin, Eka Fitriani, Rifky Permana 2019):

1. 0.90-1.00= excellent classification;

2. 0.80-0.90 = good classification.
3. 0.70-0.80 = fair classification.
4. 0.60-0.70 = poor classification.
5. 0.50-0.60 = failure clasification.

RESULTS AND DISCUSSION

1. Smote and Naive Bayes

The test results of the Smote and Naive Bayes algorithms carried out in this study were to measure the performance of Accuracy, precision, recall and AUC from the results of training and testing datasets that had gone through the data pre-processing process. The following are the results of testing the Smote and Naive Bayes algorithms. As in the table below.

Table 6. Accuracy Smote and Naive Bayes

	true Complaint	true Not_Complaint	class precision
pred. Complaint	261	41	86.42%
pred. Not_Complaint	242	462	65.62%
class recall	51.89%	91.85%	

Table 6 shows that the Smote and Naive Bayes algorithms produce an Accuracy value of 71.87%.

Table 7. Precision Smote and Naive Bayes

	true Complaint	true Not_Complaint	class precision
pred. Complaint	261	41	86.42%
pred. Not_Complaint	242	462	65.62%
class recall	51.89%	91.85%	

Table 7 shows that the Smote and Naive Bayes algorithms produce a Precision value of 65.73%.

Table 8. Recall Smote and Naive Bayes

	true Complaint	true Not_Complaint	class precision
pred. Complaint	261	41	86.42%
pred. Not_Complaint	242	462	65.62%
class recall	51.89%	91.85%	

Table 8 shows that the Smote and Naive Bayes algorithms produce a recall value of 91.84%.

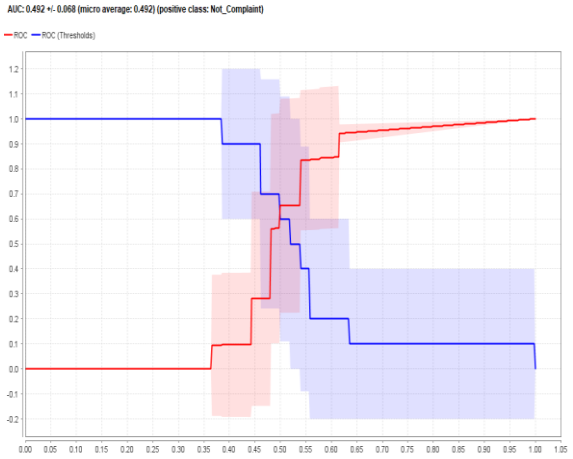


Figure 5. AUC Smote and Naive Bayes

Figure 5 shows that the calculation results on the ROC curve, depicting the ROC curve for the Smote and Naive Bayes algorithms. The ROC curves of the Smote and Naive Bayes algorithms have an AUC of 0. 492. The curve illustrates that the prediction of not complaint 242 is considered a review complaint, the prediction of complaint 261 is included in the review complaint, while the prediction of not complaint 462 is included in the review not complaint and the prediction of complaint 41 is included in the review not complaint . These results show that the Smote Naive Bayes algorithm method is included in the category of failure clasification.

2. Smote, Naive Bayes and Adaboost

The test results of the Smote, Naive Bayes and AdaBoost algorithms carried out in this study are to measure the performance of Accuracy, precision, recall and AUC from the results of training and testing datasets that have gone through the data pre-processing process. The following are the results of testing the Smote and Naive Bayes algorithms. As in the table below.

Table 9. Accuracy Smote, Naive Bayes and AdaBoost

	true Complaint	true Not_Complaint	class precision
pred. Complaint	261	41	86.42%
pred. Not_Complaint	242	462	65.62%
class recall	51.89%	91.85%	

Table 9 shows that the Smote, Naive Bayes and AdaBoost algorithms produce an Accuracy value of 71.87%

Vol. 19, No.1 Maret 2022 | DOI https://doi.org/10.33480/techno.v19i1.3086

Table 10. Precision Smote, Naive Bayes and AdaBoost

	true Complaint	true Not_Complaint	class precision
pred. Complaint	261	41	86.42%
pred. Not_Complaint	242	462	65.62%
class recall	51.89%	91.85%	

Table 10 shows that the Smote, Naive Bayes and AdaBoost algorithms produce a Precision value of 65.73%.

Table 11. Recall Smote, Naive Bayes and AdaBoost

	true Complaint	true Not_Complaint	class precision
pred. Complaint	261	41	86.42%
pred. Not_Complaint	242	462	65.62%
class recall	51.89%	91.85%	

Table 11 shows that the Smote, Naive Bayes and AdaBoost algorithms produce a recall value of 91.84%.

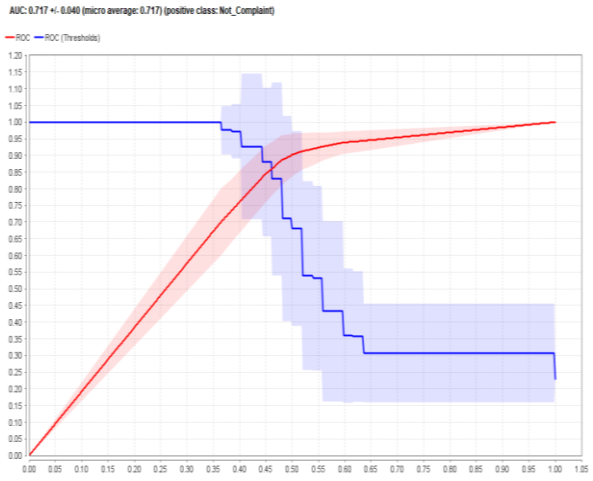


Figure 6. AUC Smote, Naive Bayes and AdaBoost

Figure 6 shows that the calculation results on the ROC curve, depicting the ROC curve for the Smote, Naive Bayes and AdaBoost algorithms. The ROC curves of the Smote, Naive Bayes and AdaBoost algorithms have an AUC of 0.717. The curve illustrates that the prediction of not complaint 242 is considered a review complaint, the prediction of complaint 261 is included in the review complaint, while the prediction of not complaint 462 is included in the review not complaint and the prediction of complaint 41 is included in the review not complaint . These results show that the Smote, Naive Bayes and AdaBoost algorithm method is included in the category of fair clasification.

3. Smote and Support Vector Machine (Evolutionary)

The test results of the Smote and SVM algorithms carried out in this study are to measure the performance of Accuracy, precision, recall and AUC from the results of training and testing datasets that have gone through the data pre-processing process. The following are the results of testing the Smote and Naive Bayes algorithms. As in the table below.

Table 12. Accuracy Smote and SVM

	true Complaint	true Not_Complaint	class precision
pred. Complaint	376	112	77.05%
pred. Not_Complaint	127	391	75.48%
class recall	74.75%	77.73%	

Table 12 shows that the Smote and SVM algorithms produce an Accuracy value of 76.24%.

Table 13. Precision Smote and SVM

	true Complaint	true Not_Complaint	class precision
pred. Complaint	376	112	77.05%
pred. Not_Complaint	127	391	75.48%
class recall	74.75%	77.73%	

Table 13 shows that the Smote and SVM algorithms produce a Precision value of 75.65%.

Table 14. Recall Smote and SVM

	true Complaint	true Not_Complaint	class precision
pred. Complaint	376	112	77.05%
pred. Not_Complaint	127	391	75.48%
class recall	74.75%	77.73%	

Table 14 shows that the Smote and SVM algorithms produce a recall value of 77.72%.

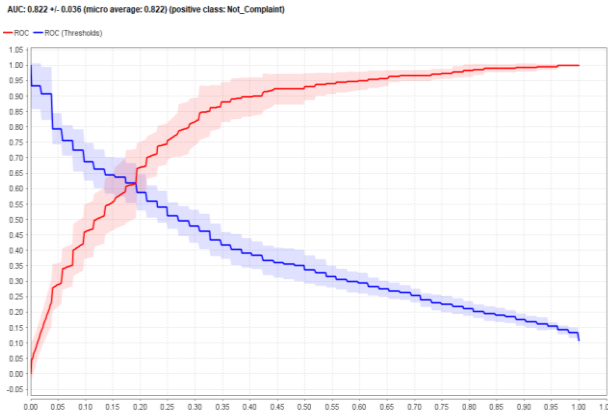


Figure 7. AUC Smote and SVM

Figure 7 shows that the calculation results on the ROC curve, depicting the ROC curve for the Smote and SVM algorithms. The ROC curves of the Smote and SVM algorithms have an AUC of 0.822. The curve illustrates that the prediction of not complaint 127 is considered a review complaint, the prediction of complaint 376 is included in the review complaint, while the prediction of not complaint 391 is included in the review not complaint and the prediction of complaint 112 is included in the review not complaint. These results show that the Smote and SVM algorithm method is included in the category of good clasification.

4. Results

The following is a comparison of the experimental results using the Smote + Naive Bayes, Smote + Naive Bayes + Adaboost, and Smote + SVM models.

Table 15. Comparison of Algorithms

Algoritma	Accuracy	Precision	Recall	AUC
Smote + Naive Bayes	72.87%	65.73 %	91.84%	0.492
Smote + Naive Bayes + AdaBoost	71.87%	65.73 %	91.84%	0.717
Smote+SVM (Evolution)	76.24%	75.65 %	77.72%	0.822

Based on Table 15 above, it can be seen that the Accuracy, Precision, AUC values of the Smote + SVM algorithm are higher than other algorithms, namely Accuracy 76.24%, Precision 75.65%, AUC 0.822. The results of the comparison of the algorithms show that the algorithm in determining the sentiment analysis of complaints and not complaints is better than other algorithms. The comparison of the Smote + Naive Bayes, Smote + Naive Bayes + Adaboost, and Smote + SVM model algorithms is depicted graphically in Figure 8.

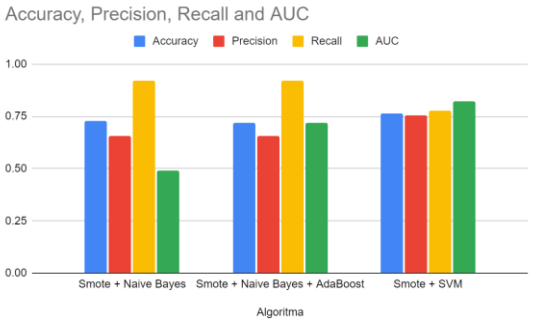


Figure 8. Comparison of Algorithms

CONCLUSION

From the comparison data from testing experiments using the Smote + Naive Bayes, Smote + Naive Bayes + Adaboost, and Smote + SVM models. It is known that the Accuracy, Precision, AUC values of the Smote + SVM algorithm are higher than other algorithms, namely Accuracy 76.24%, Precision 75.65%, AUC 0.822. The results of the comparison of the algorithms show that the algorithm in determining the sentiment analysis of complaints and not complaints is better than other algorithms.

REFERENCE

Afifah Cahayani Adha. 2019. "APLIKASI PENCARIAN HUBUNGAN ANTAR POKOK BAHASAN PADA AYAT AL-QURAN MENGGUNAKAN ALGORITMA APRIORI." Universitas Islam Negeri Sultan Syarif Kasim Riau.

Harahap, Dedy Ansari. 2018. "Perilaku Belanja Online Di Indonesia: Studi Kasus." *JRMSI - Jurnal Riset Manajemen Sains Indonesia* 9(2):193–213. doi: 10.21009/jrmsi.009.2.02.

Iswanto, Hery, Erni Seniwati, Yuli Astuti, and Dina Maulina. 2021. "Comparison of Algorithms on Machine Learning For Spam Email Classification." *IJISTECH (International Journal of Information System and Technology)* 5(4):446. doi: 10.30645/ijistech.v5i4.164.

Jaka, Aris Tri. 2015. "Preprocessing Text Untuk Meminimalisir Kata Yang Tidak Berarti Dalam Proses Text Mining." *Informatika UPGRIS* 1:1–9.

Kasanah, Anis Nikmatul, Muladi Muladi, and Utomo Pujianto. 2019. "Penerapan Teknik SMOTE Untuk Mengatasi Imbalance Class Dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 3(2):196–201. doi: 10.29207/resti.v3i2.945.

Langgeni, Diah Pudi, Z. K. Abdurahman Baizal, and

Yanuar Firdaus A. W. 2010. "Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection." *Seminar Nasional Informatika 2010* 2010(semnasIF):1–10.

Maragoudakis, Manolis, Nikos Fakotakis, and George Kokkinakis. 2004. "A Bayesian Model for Shallow Syntactic Parsing of Natural Language Texts." *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004* 847–50.

Masripah, Siti, and Lila Dini Utami. 2020. "Algoritma Klasifikasi Naïve Bayes Untuk Analisa Sentimen Aplikasi Shopee." *Swabumi* 8(2):114–17. doi: 10.31294/swabumi.v8i2.8444.

Muslim, M. A., S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah. 2018. "Optimization of C4.5 Algorithm-Based Particle Swarm Optimization for Breast Cancer Diagnosis." *Journal of Physics: Conference Series* 983(1):22–27. doi: 10.1088/1742-6596/983/1/012063.

Perdana, Hana Adi. 2022. "Survei: Shopee Rajai Pasar E-Commerce Indonesia Di 2021." *Idn Times* 1–8.

Riska Aryanti, Atang Saepudin, Eka Fitriani, Rifky Permana, Dede Firmansyah Saefudin. 2019. "Komparasi Algoritma Naive Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Pengguna Busway." *Jurnal Teknik Komputer* V No. 2Agu. doi: 10.31294/jtk.v4i2.

Ritonga, Alven Safik, and Endah Supeni Purwaningsih. 2018. "Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding)." *Ilmiah Edutic* 5(1):17–25.

Syam, Ali Musri. 2022. "Pemanfaatan Twitter Sebagai Sarana Informasi _ KKN Tematik UPI - Kompasiana."

Wisdayani, Indah Manfaati Nur, Rochdi Wasono. 2019. "Perbandingan Algoritma K-Nearest Neighbor Dan Naive Bayes Untuk Klasifikasi Tingkat Keparahan Korban Kecelakaan Lalu Lintas Di Kabupaten Pati Jawa Tengah." *Universitas Muhammadiyah Semarang*.

Zaenudin, Ahmad. 2017. "Profil Konsumen Belanja Online Di Indonesia." *Tirtoid*.

Zulhanif. 2015. "Algoritma AdaBoost Dalam Pengklasifikasian." Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS.

COMPARING ALGORITHM FOR SENTIMENT ANALYSIS IN HEALTHCARE AND SOCIAL SECURITY AGENCY (BPJS KESEHATAN)

Asyharudin[1]; Novi Kusumawati[2]; Ulfah Maspupah[3]; Destia Sari R.F.[4]; Amir Hamzah[5]; Duwik Lukito[6]; Dedi Dwi Saputra[7]

Information Systems
Nusa Mandiri University
https://nusamandiri.ac.id/
job.adin.gmail.com[1]; novikusumawati703@gmail.com[2]; ulfahmaspupah83@gmail.com[3]; destiafadhillah19@gmail.com[4]; amirhamzah.jkt@gmail.com[5]; duwiklukito09@gmail.com[6]; dedi.eis@nusamandiri.ac.id[7]

Abstract— *Twitter is a social media that can be used to express opinions and exchange information quickly with individuals and institutions such as the Healthcare and Social Security Agency (BPJS Kesehatan). Every word that a Twitter user utters has meaning and stellar emotion. This meaning can be reached through the process of sentiment analysis. Sentiment analysis is the process of understanding and classifying emotions such as positive or negative or complaining or not complaining. This study classifies tweet data related to BPJS Health services into two classifications, namely complain and no complain. Using 1,000 data from Twitter written on the BPJS Kesehatan Twitter account. In text mining, to build a classification, the transform case, tokenize, token filter by length, stemming and stopword techniques are used. Gataframework is used to assist the preprocessing and cleansing process. Rapidminer was used to create sentiment analysis in comparing three different classification methods of the Twitter data. The method used is the Nave Bayes algorithm and the Naïve Bayes algorithm with the addition of a Synthetic Minority Over-sampling Technique (SMOTE) feature and the Naïve Bayes algorithm with an SMOTE feature that is optimized with Adaboost. The Naïve Bayes algorithm is added with the SMOTE feature which is optimized with Adaboost to get the best value with an accuracy value of 69.11%, precision 69.93%, recall 68.89% and AUC 0.770.*

Keywords: *Text Mining, Naïve Bayes, Adaboost, classification, Sentiment Analysis.*

Intisari— *Twitter salah satu media sosial yang bisa digunakan untuk menyampaikan opini dan bertukar informasi dengan cepat kepada individu maupun kepada institusi seperti Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan. Setiap kata yang diutarakan pengguna Twitter memiliki makna dan emosi tersirat. Makna tersebut bisa dipahami melalui proses sentimen analisis. Sentimen analisis merupakan proses memahami dan mengelompokkan emosi seperti positif atau negatif maupun complain atau no complain. Penelitian ini mengklasifikasikan data tweet yang berkaitan dengan layanan BPJS Kesehatan menjadi dua klasifikasi yaitu complain dan no complain. Menggunakan 1.000 data dari Twitter yang ditulis di akun Twitter BPJS Kesehatan. Pada text mining untuk membangun klasifikasi digunakan teknik transform case, tokenize, token filter by length, stemming serta stopword. Gataframework digunakan untuk membantu proses preprocessing dan cleansing. Rapidminer digunakan untuk menciptakan sentimen analisis dalam membandingkan tiga metode klasifikasi yang berbeda dari data Twitter tersebut. Metode yang digunakan adalah, algoritma Naïve Bayes dan algoritma Naïve Bayes ditambahkan feature Synthetic Minority Over-sampling Technique (SMOTE) serta algoritma Naïve bayes ditambahkan feature SMOTE yang di optimasi dengan Adaboost. Algoritma Naïve Bayes ditambahkan feature SMOTE yang di optimasi dengan Adaboost mendapatkan nilai terbaik dengan nilai accurasy 69.11%, precission 69.93%, recall 68.89% dan AUC 0,770.*

Kata Kunci: *Text Mining, Naïve Bayes, Adaboost, Klasifikasi, Sentimen Analis.*

INTRODUCTION

The Indonesian Internet Service Providers Association (APJII) conducted a survey in 2016. There are around 132.7 million internet users in Indonesia (a significant increase from 88 million

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3167

users in 2014). Of this number, 97.4% (129.2 million) are users who use the internet to access social media. The five social media with the most users are Facebook, Instagram, Youtube, Google Plus, and Twitter(Deviyanto & Wahyudi, 2018).

Twitter is a social media that can be used to express opinions and exchange information quickly to individuals and to institutions such as the Healthcare and Social Security Agency (BPJS Kesehatan). The opinion conveyed to BPJS Kesehatan is very important to improve the quality of services. Improving the quality of services at BPJS Kesehatan is very important in order to increase satisfaction with the community in obtaining good and quality health services. BPJS Kesehatan is a legal entity created to be able to organize insurance programs for health(Puspita & Widodo, 2021). Health is a state of health both physically, mentally, spiritually and socially that enables everyone to live socially and economically productive lives. while health efforts are every activity to maintain and improve health carried out by the government and or the community(Suprapto & Malik, 2019).

Every word in the opinion expressed by Twitter users has an implied meaning and emotion. This meaning can be understood through the process of sentiment analysis. Sentiment analysis is the process of understanding and classifying emotions such as positive or negative or complain or no complain.

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, measure, and study affective states and subjective information. Sentiment analysis is widely applied to customer voice materials such as survey reviews and responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. With the advent of deep language models, such as RoBERTa, more difficult data domains can also be analyzed, for example, news texts where writers usually express their opinions/sentiments less explicitly(*Sentiment Analysis*, n.d.)

Based on the background that has been described, the authors are interested in conducting research with the title "Comparing Algorithm for sentiment analysis in Healthcare and Social Security Agency (BPJS Kesehatan)".

In order for the problem being reviewed to be more focused and achieve the predetermined targets, problem boundaries must be given, including:

a. The object used in this research is tweet data from BPJS Kesehatan twitter users in April 2022.
b. The tweet that will be used is the tweet sentence that uses Indonesian only.
c. The algorithm that will be used for classification in this research is Adaboost and NBC (Naive Bayes Classifier).
d. In this research, the stemming and stopword processes are only for Indonesian words.

MATERIALS AND METHODS

A. Data Collection Techniques

There are several ways to collect data in this research:

1) Data Analysis

Data analysis is a data processing process that aims to find useful information so that it can be used as a basis for decision making as a solution to solve a problem(Kurniasari, 2021). The data used in this study were 1,000 Indonesian-language tweets on Twitter containing the opinions of the Indonesian people on BPJS Kesehatan services. The data is selected manually, namely by selecting tweet sentences that are in Indonesian and do not contain images. The selected data is then stored in excel form. The data in this study consisted of two types, namely training data and test data. For the purposes of training data, the data that has been collected is then categorized manually to assess the sentiment in the tweet, which is included in the complain or no complaint category.

Table 1. Kind of Sentiment

Description	*Sentiment*		
	Complain	***No Complain***	***Grand Total***
Total	485	515	1.000

From the table above, there are 485 complaint data and 515 no-compliance data.

2) Text Processing Analysis

Text processing is a process of extracting, processing, organizing information by analyzing the relationship, the rules that exist in semi-structured or unstructured textual data. To be more effective in the processing process, data transformation steps are carried out into a format that is easy for user needs. This process is called text processing. Once in a more structured form with the above process, the data can be used as a data source that can be processed further. The stages for text processing consist of tokenizing, feature normalization, case folding and stopword removal(Sudiantoro et al., 2018).

B. Research Methods

The research method used is to collect tweet data using the Crawling method from Twitter. Data was taken randomly as many as 1,000 tweets in Indonesian with the keyword BPJS Kesehatan.

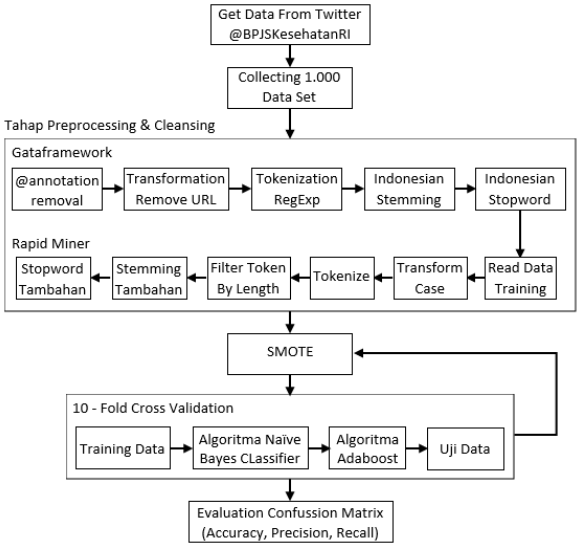


Figure 1. Research Methods

Based on the Figure 1 above, the research process begins with collecting data from Twitter using Rapid Miner, the data taken in this study is 1,000 data with the keyword BPJS Kesehatan, the data that has been collected is stored in excel format. After the data collection process is complete, the next step is to label each data complaint or no complain. After labeling the data, the next process is the preprocessing and cleansing stages. This preprocessing and cleansing stage uses two tools, Gataframewok and Rapid Miner. Gataframework is used to perform the first stage of preprocessing, in the first preprocessing stage the processes carried out are @Annotation Removal, Remove URL, Regexp, Indonesian Stemming and Indonesian stopword. In Rapid Miner, the processes carried out are Transform case, Tokenize, Filter Token By Length, additional Stemming and additional stopwords. The last stage in this research is the process of implementing the Naive Bayes algorithm.

RESULTS AND DISCUSSION

A. Research Stages

1) Types Of Research

Sentiment analysis is used to determine the sentiment or polarity of a text whether it is Extremely positive, positive, neutral, negative, Extremely negative. Usually sentiment analysis is applied to text data of public opinion on an object, for example a review of an e-commerce product, a review of a film and comments on social media(Prima et al., 2022). The opinion is in the form of a tweet which will later become a news spread on the Twitter timeline. Each of these opinions is very important for improving the quality of BPJS Kesehatan services to the community, so that people can get good and quality services.

2) Data Collection

The data collection process was carried out using the Twitter API for the Rapid Miner application with the Query "@BPJSKesehatanRI" for the period April 2022 with 1,000 data.

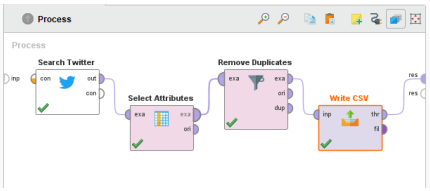


Figure 2. Twitter Data Collection Stage

Based on Figure 2 above, the Search Twitter Operator is used to connect Rapid Miner with Twitter to collect as much as 1,000 data with the keyword BPJS Health. The data collected from Twitter is only taken from the Text Column using the Select Attributes Operator, in this Text Column it contains tweets from Twitter users. To avoid duplicate tweet data, the Remove Duplicates Operator is used. The data that has been selected based on the Text Column and the duplicate data has been removed then the data saved into csv format using the Write CSV Operator.

3) Data Labeling

Data labeling is an advanced stage from the previous stage where calculations will be carried out polarity of the reviews that have been taken, so can produce two categories, namely labels(Herlinawati et al., 2020).

The data that has been collected is then given a sentiment label (Complain or No Complain) using VADER (Valence Aware Dictionary And Sentiment Reasoner). VADER is a glossary and tool for performing sentiment analysis depending on the exclusive standardized law to manifest sentiment on social media. VADER is an open source tool that is completely free. Combines word setting considerations and degree qualifications(America & States, 2021).

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3167

Table 2. Data Labeling Stage

Text	Sentiment
@BPJSKesehatanRI min, apa kartu bpjs kesehatan harua di cetak dulu untuk bisa mendapatkan pelayanan faskes? Apa boleh kita nunjukin kartu virtual di aplikasi saja?	*No_Complain*
@BPJSKesehatanRI @julio_airlangga Manfaatnys adalah pemasukan cuma^2 tanpa kewajiban mencover biaya rs bagi anggota yg kesulitan bayar tepat waktu.	*No_Complain*
@BPJSKesehatanRI Min ni gabisa2 Swafotonya? https://t.co/c83xRcjyif	*Complain*
@BPJSKesehatanRI Mau nanya kenapa BPJS nggak bisa di aktifkan lewat mobile JKN ...???	*Complain*
@BPJSKesehatanRI Apaan pandawa malah jawab mohon maaf terus	*Complain*

4) @Annotation Removal
Sometimes in a tweet a user embeds or tags another user's username using the @xxxxx notation. In this process, the stages of removing the username on each tweet are carried out.

Table 3. @Annotation Removal Stage

Before	After
@BPJSKesehatanRI Min ni gabisa2 Swafotonya? https://t.co/c83xRcjyif	Min ni gabisa2 Swafotonya? https://t.co/c83xRcjyif

5) Remove URL
In a tweet there are usually several URL links (Uniform Resource Locator) entered by the user. This URL is usually included to provide more detailed information because of the limitation of a tweet that is only 280 words. In this process, the process of removing the URL is carried out.

Table 4. Remove URL Stage

Before	After
Min ni gabisa2 Swafotonya? https://t.co/c83xRcjyif	Min ni gabisa2 Swafotonya?

6) Tokenization RegExp
In this process, the procedure for removing punctuation marks on a tweet is carried out, among others. , : " " " ' ? !, etc. In this process, every word contained in the document will be collected and then the punctuation marks, symbols or anything that is not a letter will be removed(Utami, 2018).

Table 5. Tokenization RegExp Stage

Before	After
Min ni gabisa2 Swafotonya?	Min ni gabisa Swafotonya

7) Indonesian Stemming
Stemming is a process to get the basic word from the original word in a sentence. The original word can contain affixes that are separated based on certain rules, for example the word makanan, dimakan, memakan which has the same root word, namely makan(Wardana et al., 2019).

8) Indonesian Stopword
This stage is the process of eliminating certain words in a tweet that are considered meaningless (stopwords). Basically stopword is a list of words in a language. Stopwords tend to be omitted in research related to text mining because stopwords are used repeatedly in a sentence so that stopwords are omitted so that research can focus more on words that are more important. Examples of stop words in Indonesian include yang, dan, di, dari, etc.
The essence of stopwords is to remove words that have low information value or that have no relevance to the content of the document(Hendra & Fitriyani, 2021).

9) Transform Case
In writing a tweet there are several forms of letters used by users, both uppercase and lowercase letters. At this stage all existing letters are converted to lowercase.

Table 6. Transform Case Stage

Before	After
Min ni gabisa Swafotonya	min ni gabisa swafotonya

10) Tokenize
In the tokenize process, the tokenization process is carried out in words or cutting a sentence into word for word.
The tokenization process is the process of separating a series of characters based on each word that composes them or space characters, and it is possible to delete word characters at the same time(Sari et al., 2021).

Table 7. Tokenize Stage

Before	After
Min ni gabisa Swafotonya	min ni gabisa swafotonya

11) Filter Token By Length
This filter token by length is a very interesting function, with this function we can filter tokens of a certain length, the length attribute of a

parameter needs to be specified at the minimum length and maximum length, where we can determine whether a token with the minimum length to the maximum range will stay in document or not(Kalra & Aggarwal, 2018).

In this research the minimum number of characters is 3 and the maximum number of characters is 25.

Table 8. Filter Token By Length Stage

Before	After
min	min
ni	gabisa
gabisa	swafotonya
swafotonya	

B. Implementation of the Naive Bayes Algorithm

The Naive Bayes algorithm is one of the classification techniques algorithms with probability and statistical methods proposed by British scientist Thomas Bayes, which predicts future opportunities based on past experience and is known as Bayes' theorem. The theorem is combined with Naive where it is assumed that the conditions between attributes are independent. Naive Bayes classification assumes that the presence or absence of certain characteristics of a class has nothing to do with the characteristics of other classes(Nofitri & Irawati, 2019).

The results of the model testing carried out are classifying tweets complaining and tweeting no complaints using the Maive Bayes algorithm, the Naive Bayes Algorithm is added with the Synthetic Minority Over-sampling Technique (SMOTE) feature and the Naive Bayes Algorithm is added with the Synthetic Minority Over-sampling Technique (SMOTE) feature which is optimized with Adaboost.

The Naive Bayes Classifier method is used to categorize, namely to see the opinion or tendency of opinion on a problem or object by someone, whether it tends to be in the category of complaint or no complaint. The data that has gone through the text processing process will then go through the classification stage using the Naïve Bayes Classifier to find out whether the data is in the positive category or the negative category.

1) Implementation of Naive Bayes Algorithm Only

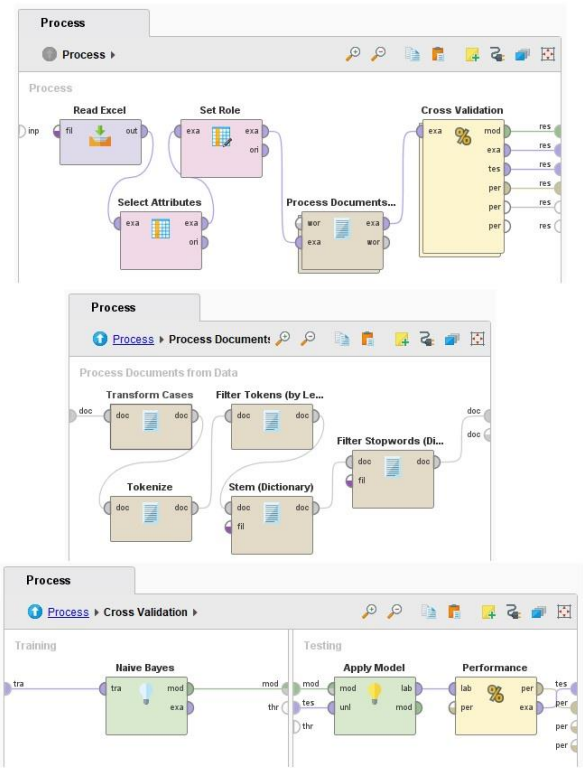


Figure 3. Nave Bayes Algorithm Implementation Only

Based on Figure 3 above, at this implementation stage the data that has gone through the preprocessing and cleansing stages are imported into Rapid Miner using the Read Excel operator. The imported data is then processed using the Process Documents operator, in the Process Documents operator there are several processes including Transform case, Tokenize, Filter Token By Length, additional Stemming and additional stopwords. In Figure 3 above, the implementation process only uses Naive Bayes.

The implementation using only the Naive Bayes algorithm gets the result:

Table 9. Implementation results using only the Naive Bayes algorithm

Description	*Accuracy*	*Precision*	*Recall*	*AUC*
Result	71.68%	77.37%	61.17%	0.745

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3167

2) Implementation of Naïve Bayes Algorithm Plus SMOTE Features

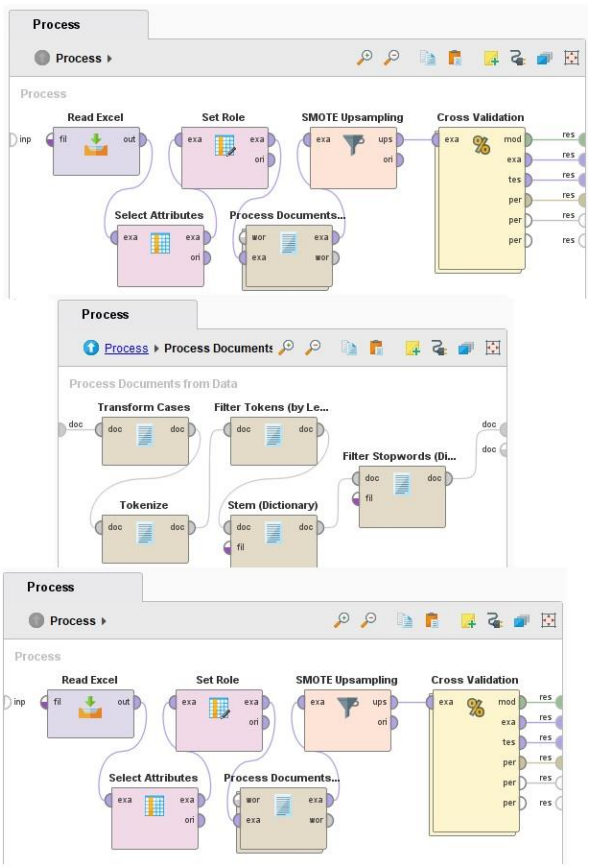


Figure 4. Implementation of Naive Bayes Algorithm Added SMOTE feature

Based on Figure 4 above, at this implementation stage the data that has gone through the preprocessing and cleansing stages are imported into Rapid Miner using the Read Excel operator. The imported data is then processed using the Process Documents operator, in the Process Documents operator there are several processes including Transform case, Tokenize, Filter Token By Length, additional Stemming and additional stopwords. In Figure 3 above, the implementation process uses Naive Bayes added the SMOTE feature.

Implementation using the Naïve Bayes algorithm added the SMOTE feature to get the following results:

Table 10. Implementation results using the Naïve Bayes algorithm added the SMOTE feature

Description	*Accuracy*	*Precision*	*Recall*	*AUC*
Result	73.27%	80.24%	61.76%	0.755

3) Implementation of Naïve Bayes & Adaboost Algorithm Plus SMOTE Features

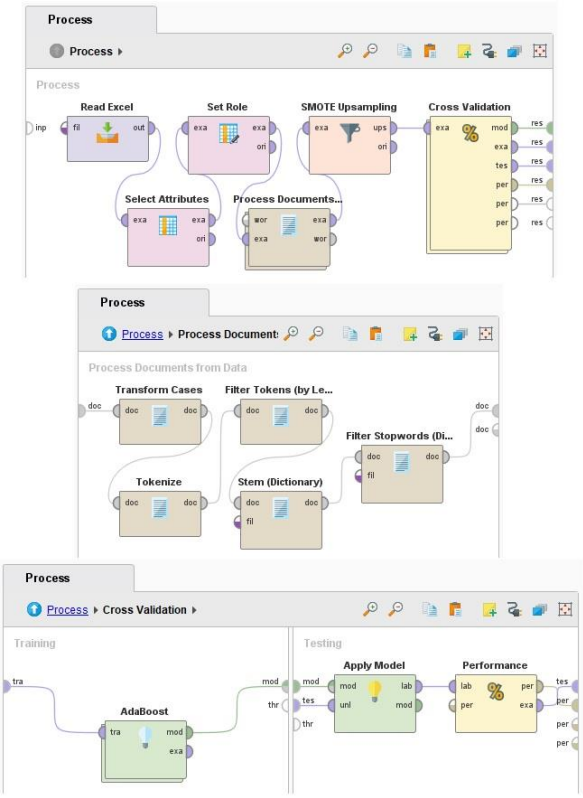


Figure 5. Implementation of Naïve Bayes & Adaboost Algorithm Plus SMOTE Features

Based on Figure 5 above, at this implementation stage the data that has gone through the preprocessing and cleansing stages are imported into Rapid Miner using the Read Excel operator. The imported data is then processed using the Process Documents operator, in the Process Documents operator there are several processes including Transform case, Tokenize, Filter Token By Length, additional Stemming and additional stopwords. In Figure 3 above, the implementation process uses Naive Bayes and Adaboost added the SMOTE feature.

Implementation using the Naïve Bayes and Adaboost algorithm added the SMOTE feature got the following results:

Table 11. Implementation Results Using Naïve Bayes Algorithm and Adaboost Plus SMOTE Features

Description	*Accuracy*	*Precision*	*Recall*	*AUC*
Result	69.11%	69.93%	68.89%	0.770

CONCLUSION

The results of this research indicate that the Naive Bayes Algorithm when added with the Synthetic Minority Over-sampling Technique (SMOTE) feature which is optimized with Adaboost

produces accuracy: 69.11%, precision: 69.93%, recall: 68.89% and AUC: 0.770. This reaserch also uses the Naïve Bayes algorithm without adding the SMOTE feature that produce accuracy: 71.68%, precision: 77.37%, recall: 61.17% and AUC: 0.745. Meanwhile, the Naïve Bayes algorithm added with the SMOTE feature produces accuracy: 73.27%, precision: 80.24%, recall: 61.76%, and AUC: 0.755. Based on the results of this research, it can be concluded that the Nave Bayes Algorithm with SMOTE features added which is optimized using Adaboost is a better classification to use than the Nave Bayes Algorithm with SMOTE features and Nave Bayes Algorithm without SMOTE features.

REFERENCE

America, N., & States, U. (2021). *Survey of Twitter Viewpoint on Application of Drugs by VADER Sentiment Analysis among Distinct Countries*. *March 2021*.

Deviyanto, A., & Wahyudi, M. D. R. (2018). Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *3*(1), 1. https://doi.org/10.14421/jiska.2018.31-01

Hendra, A., & Fitriyani, F. (2021). Analisis Sentimen Review Halodoc Menggunakan Naï¨ve Bayes Classifier. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(2), 78–89. https://doi.org/10.14421/jiska.2021.6.2.78-89

Herlinawati, N., Yuliani, Y., Faizah, S., Gata, W., & Samudi, S. (2020). Analisis Sentimen Zoom Cloud Meetings di Play Store Menggunakan Naïve Bayes dan Support Vector Machine. *CESS (Journal of Computer Engineering, System and Science)*, *5*(2), 293. https://doi.org/10.24114/cess.v5i2.18186

Kalra, V., & Aggarwal, R. (2018). Importance of Text Data Preprocessing & Implementation in RapidMiner. *Proceedings of the First International Conference on Information Technology and Knowledge Management*, *14*(January), 71–75. https://doi.org/10.15439/2017km46

Kurniasari, D. (2021). *Analisis Data Adalah: Mengenal Pengertian, Jenis, Dan Prosedur Analisis Data*. https://www.dqlab.id/analisis-data-adalah-mengenal-pengertian-jenis-dan-prosedur-analisis-data

Nofitri, R., & Irawati, N. (2019). Analisis Data Hasil Keuntungan Menggunakan Software Rapidminer. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *5*(2), 199–204. https://doi.org/10.33330/jurteksi.v5i2.365

Prima, J., Sistem, J., Komputer, I., No, V., Banjarnahor, J., Indra, E., & Sinurat, S. H. (2022). *ANALISIS PERBANDINGAN SENTIMEN CORONA VIRUS DISEASE- 2019 (COVID19) PADA TWITTER MENGGUNAKAN METODE LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE (SVM)*. *5*(2).

Puspita, R., & Widodo, A. (2021). Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS. *Jurnal Informatika Universitas Pamulang*, *5*(4), 646. https://doi.org/10.32493/informatika.v5i4.7622

Sari, S., Khaira, U., Pradita, P., & Tri, T. S. (2021). … Beauty Shaming Di Media Sosial Twitter Menggunakan Algoritma SentiStrength: Sentiment Analysis Against Beauty Shaming Comments on Twitter Social Media …. *Indonesian Journal of …*, *1*(1), 71–78. https://journal.irpi.or.id/index.php/ijirse/article/view/55%0Ahttps://journal.irpi.or.id/index.php/ijirse/article/download/55/24

Sentiment Analysis. (n.d.). Retrieved June 30, 2022, from https://en.wikipedia.org/wiki/Sentiment_analysis

Sudiantoro, A. V., Zuliarso, E., Studi, P., Informatika, T., Informasi, F. T., Stikubank, U., & Mining, T. (2018). Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naive Bayes Classifier. *Dinamika Informatika*, *10*(2), 398–401.

Suprapto, S., & Malik, A. A. (2019). Implementasi Kebijakan Diskresi Pada Pelayanan Kesehatan Badan Penyelenggara Jaminan Kesehatan (Bpjs). *Jurnal Ilmiah Kesehatan Sandi Husada*, *7*(1), 1–8. https://doi.org/10.35816/jiskh.v7i1.62

Utami, L. D. (2018). Komparasi Algoritma Klasifikasi Pada Analisis Review Hotel. *Jurnal Pilar Nusa Mandiri*, *14*(2), 261. https://doi.org/10.33480/pilar.v14i2.1023

Wardana, H. K., Swanita, I., & Yohanes, B. W. (2019). Sistem Pemeriksa Pola Kalimat Bahasa Indonesia berbasis Algoritme Left-Corner Parsing dengan Stemming. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *8*(3), 211. https://doi.org/10.22146/jnteti.v8i3.515

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3059

WEBSITE-BASED CERTIFICATE MANAGEMENT INFORMATION SYSTEM DESIGN IN TRAINING AND CONSULTANT DIVISION

Muhammad Salbiyath [1] ; Daning Nur Sulistyowati [2]

Information Systems
Nusa Mandiri University
www.nusamandiri.ac.id
[1] 11207134@nusamandiri.ac.id ; [2] daningnur.dgs@nusamandiri.ac.id

Abstract — Certification is a determination given by a professional organization to a person to show that a person has competence, able to do a specific job or task. The management of certificates at PT Markco International is still carried out in a simple manner using Microsoft Office applications, starting from managing personal data, passing person status, numbering and issuing certificates. The design of this Certificate Management Information System is intended to provide a better solution, namely a web-based information system that can convert the currently running manual system into a computerized system using a database. Waterfall software development method and data collection techniques by conducting observations and literature studies are applied in the development of this program. This design discusses the process of managing personal data, registration agencies and certificates. As a result, the certificate management program that has been created can manage related information such as data on persons, agencies, certificate categories, registrations and certificates. This proves that the Certificate Management Information System Design at PT Markco International can manage the certificate management process well.

Keywords: information system, certificates, website-base certificates, training and consultant division

Abstract — Sertifikasi merupakan suatu penetapan yang diberikan oleh suatu organisasi profesional terhadap seseorang untuk menunjukkan bahwa orang tersebut memiliki kompetensi, mampu untuk melakukan suatu pekerjaan atau tugas spesifik. Pengelolaan sertifikat pada PT Markco International masih dilakukan secara sederhana menggunakan aplikasi Microsoft Office mulai dari pengelolaan data person, status kelulusan person, penomoran dan penerbitan sertifikat. Perancangan Sistem Informasi Pengelolaan Sertifikat ini dimaksudkan untuk memberikan solusi yang lebih baik yaitu sistem informasi berbasis web yang dapat mengkonversi sistem manual yang saat ini berjalan ke dalam sistem terkomputerisasi dengan menggunakan databse. Metode pengembangan perangkat lunak Waterfall dan teknik pengumpulan data dengan melakukan observasi dan studi pustaka diterapkan dalam pengembangan program ini. Perancangan ini membahas proses pengelolaan data person, instansi, kategori sertifikat, registrasi dan sertifikat. Sebagai hasilnya, program pengelolaan sertifikat yang telah dibuat dapat mengelola informasi terkait seperti data person, instansi, kategori sertifikat, registrasi dan sertifikat. Hal ini membuktikan bahwa Perancangan Sistem Informasi Pengelolaan Sertifikat Pada PT Markco International dapat mengelola proses pengelolaan sertifikat dengan baik.

Kata Kunci: sistem informasi, sertifikat, website sertifikat, divisi training and consultant

INTRODUCTION

The high human need for information encourages the development of technology and information to be very rapid. At this time to get information is very easy, with the help of the internet we can find information very quickly and easily through the website. Website is a collection of web pages that have been published on the internet network and have a domain/URL (Uniform Resource Locator) that can be accessed by all internet users by typing the address. (., Ibrahim, and Ambarita 2018) . This is made possible by the existence of World Wide Web (WWW) technology ". Web- based applications are one form of information technology development that helps make it easier for organizations to do various things,

so that organizations can grow and develop quickly. But at this time there are still many organizations that have not taken advantage of the development of information technology, one of which is Markcert.

Markcert is a brand owned by PT Markco International which is engaged in training and consultant. The process of managing and issuing certificates conventionally at the Bali Province LPMP raises problems including longer processing time and processes, the difficulty of recapitulating the certificates that have been issued, the certificate distribution process is still done manually, allowing data loss if the computer is damaged because the storage is still on the local computer. (Arya et al. 2021) . Markcert has a problem where the management of person data, person graduation status, numbering and certificate issuance is still done simply by using Microsoft Office applications. There are two types of certificates issued by Markcert, namely certificates that are valid for life and certificates that are valid for a certain period. The process of manually issuing certificates takes a long time, then the absence of a system to verify the authenticity of certificate ownership, of course, allows certificate forgery to occur (Samala and Fajri 2021) .

MATERIALS AND METHODS

Markcert is a brand of PT. Markco International and MarkCert are transformations of AFNOR Indonesia in 2020 which have experience in the fields of training, mentoring (consultation), and conformity assessment (certification). Markcert has two types of certificates, namely training certificates and competency certificates. The training certificate is issued after the training has been completed, while the competency certificate is issued after all stages of testing are carried out and the person is declared competent.

TNE and PCP divisions provide registration files to the IT Subdivision. Registration files contains the participant's personal data then the IT Division saves the participant data into the Microsoft Excel application. After the data is saved, the IT division asks for confirmation from the TNE and PCP divisions. The certificate that has been made is then submitted to the TNE and PCP Division for validation, if it is appropriate, the certificate will be distributed to participants. If there is an error in the certificate, it will be returned to the IT division for correction. The system runs on PT. Markco International, especially the Markcert brand, in this certificate management system is still manual with procedures which can be seen in Figure 1.

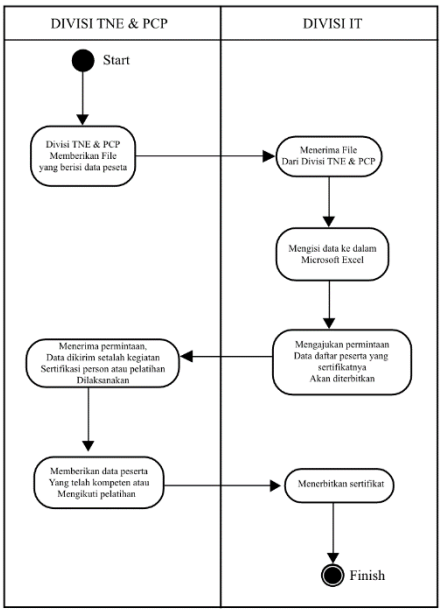


Figure 1 Certificate Making Process

This certificate management system development model uses a web -based system design consisting of:

1. Software Requirements Analysis

The certificate management application requires data such as person, graduation status, schema and other information that is recorded into a single database .

2. Design (Design)

application design uses Unfied Modeling Language (UML) and the diagrams used are Use Case Diagrams . While the database design himself using Entity Relational Diagram (ERD) and Logical Relational Diagrams (LRS).

3. Code Generation

The stage of data translation or problem solving that has been designed into a particular programming language. This application uses PHP, HTML and CSS programming languages which are used to compile the layout of the designs created and the database used is MySQL.

4. Testing

Is the testing phase of the software that was built. Testing is carried out in accordance with the desired system work. Blackbox testing is carried out to ensure whether all the performance of the system is running well.

5. Support

This is the final stage where the software is finished. At this stage ensure that all systems run properly according to user needs and can also provide changes or add new features according to user requests.

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3059

RESULTS AND DISCUSSION

Analysis

At this stage, it contains the requirements specification (system requirements) of a web - based certificate management system which is divided into 2 (two), namely:

1. Admin Page
2. Supervisor Page

On the admin page the access obtained includes:

A.1 Manage Certificate Categories
A.2 Manage Agencies
A.3 Manage Person
A.4 Manage Schema
A.5 Manage Registration
A.6 Manage Certificates

As for the access supervisor page, the user management is obtained.

Use Case Diagrams

use case diagram is a diagram that shows the relationship or user interaction with the system being developed (Samala and Fajri 2021) .

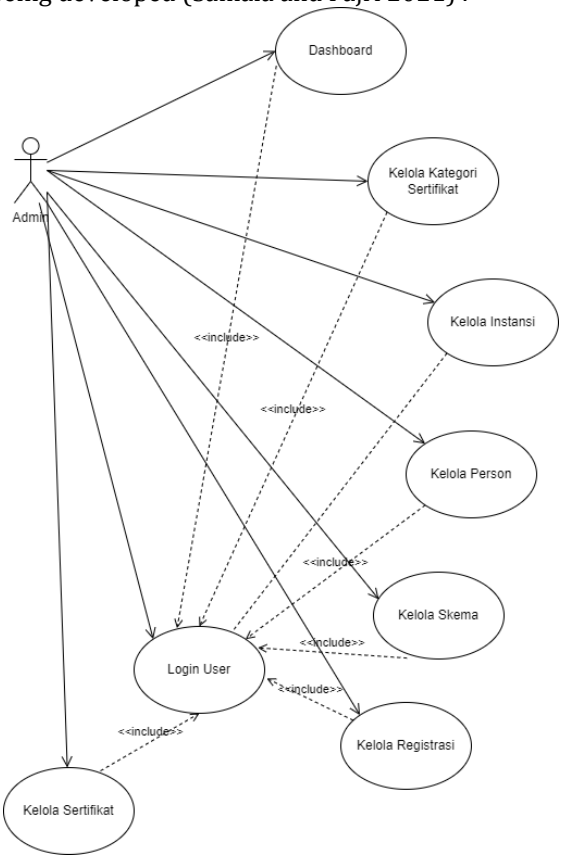


Figure 2 Use Case Diagram Admin

Use case diagram Figure 2 admin can manage certificate categories, agencies, persons, schemas, registrations, certificates online through the website which includes viewing data, adding data and changing data.

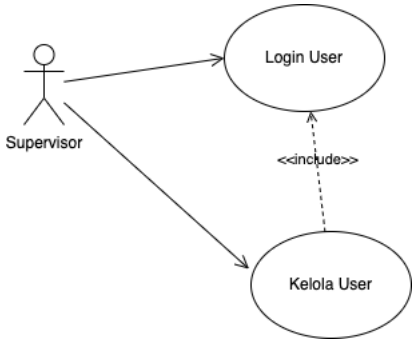


Figure 3 Use Case Diagram Supervisor

Use case diagram Figure 3 supervisor can manage users including viewing user data, adding user data and changing user data.

Activity Diagram

activity diagram is a diagram that describes the flow of work or activities from a system or business process or menu that is in the software (Anna et al. 2018) .

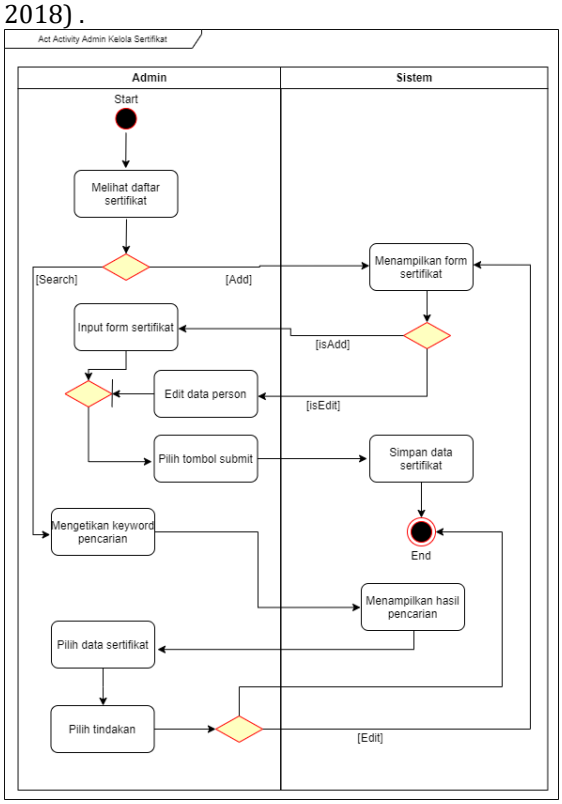


Figure 4 Activity Diagram

Figure 4 is a description of the flow of each process carried out by the admin to manage certificate data in the system.

Design

Software design is the stage of doing design that focuses on making programs including data structures, architecture, user interfaces and coding. The software requirements that have been analyzed at the requirements analysis stage are translated

into a design representation so that the program at the next stage can be implemented (Anna et al. 2018) ,

At the design stage contains an explanation of the database design and software architecture design of the designed system.

1. Entity Relationship Diagram

Entity Relationship Diagram (ERD) is information that is created, used, and stored in a business system that is shown in the form of images or diagrams (Trisyanto 2018) .

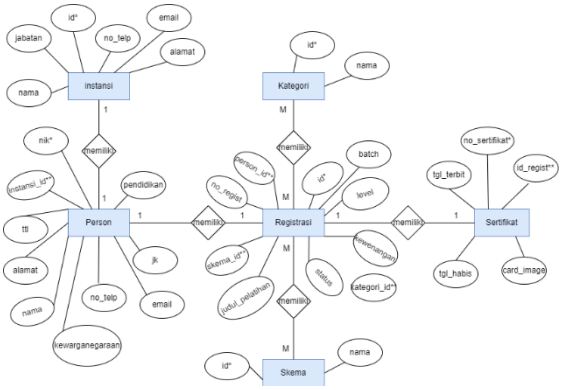


Figure 5 Entity Relationship Diagram

2. Logical Record Structure

Logical Record Structure (LRS) is a representation that comes from the structure of the records in the table where the tables were formed from the results of the set between entities in the entity relationship diagram that has been transformed as an LRS form (Anna et al. 2018) .

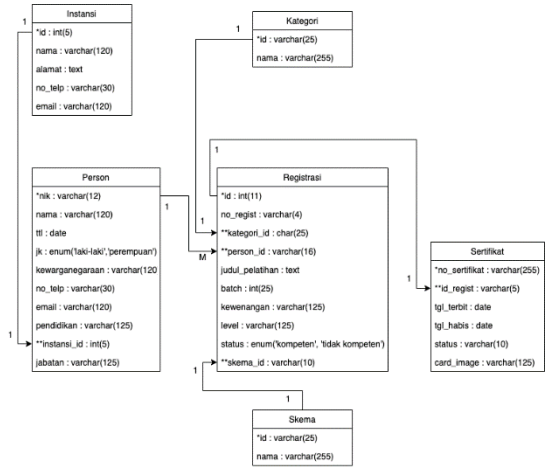


Figure 6 Logical Record Structure

3. Class Diagram

Class diagrams are classes that are defined to build the system. There are attributes and operations or methods in a class, variables owned by a class are attributes , while methods or operations are functions that are owned by a class (Heriyanto 2018) .

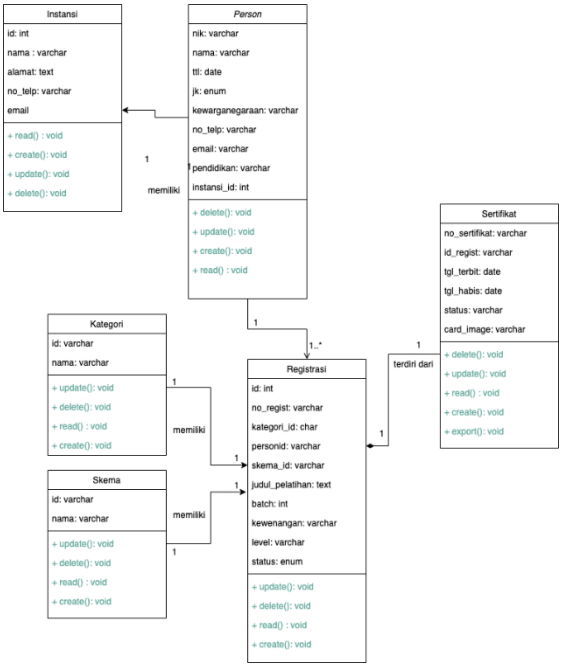


Figure 7 Class Diagram

4. Sequence Diagrams

Sequence diagrams are interactions between objects that occur in and around the system (including displays , users and so on) in the form of messages depicted against time described through pictures. (Anna et al. 2018) .

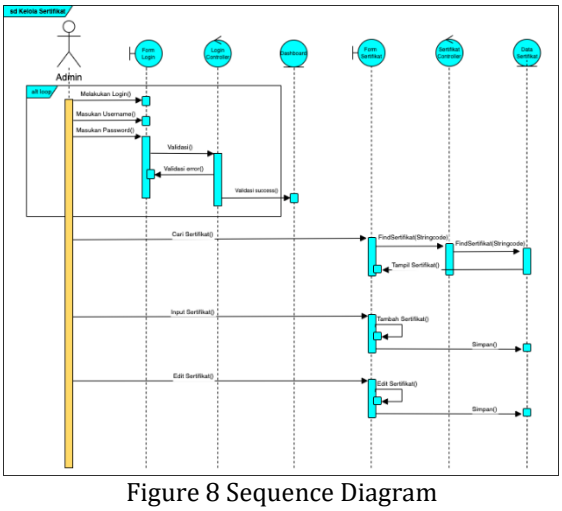


Figure 8 Sequence Diagram

5. Component Diagram

Component diagram is a system dependency component or software on existing components and managing organization, component diagrams are related to class diagrams (Cordeaux 1877) .

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3059

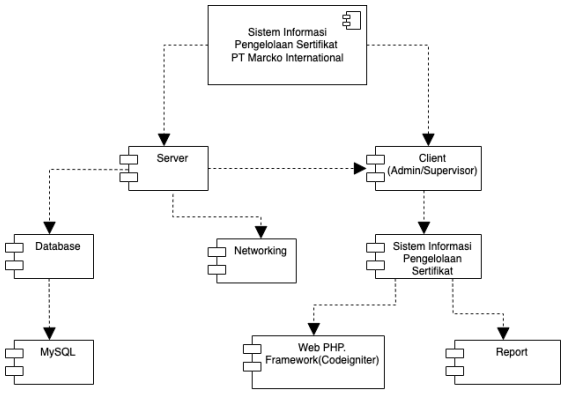


Figure 9 Component Diagram

6. Deployment Diagram

Deployment diagram is a diagram that describes how the components are arranged in detail on the system infrastructure (Anna et al. 2018) .

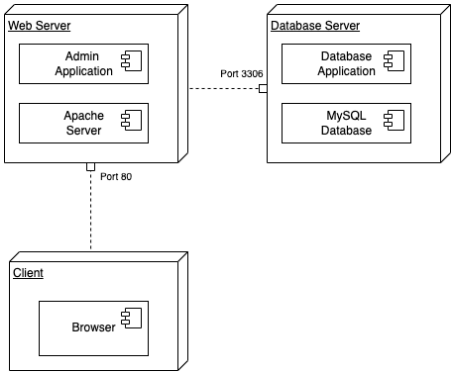


Figure 10 Deployment Diagram

Implementation

Implementation of the designed system can be seen as follows.

1. Dashboard

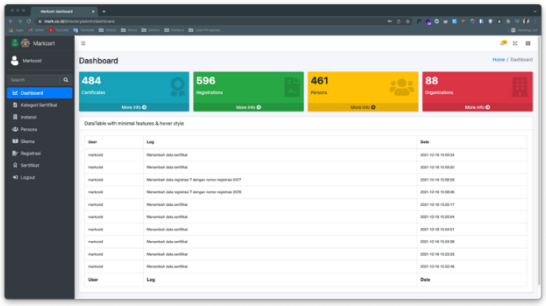


Figure 11 Dashboard Display

Figure 11 is a user interface dashboard display, serves to display general information about the amount of data that already exists in the database.

2. Manage Certificate Categories

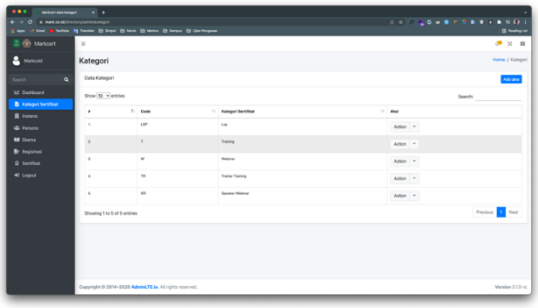


Figure 12 Certificate Category Menu Display

Figure 12 is a user interface display for the category management menu, which functions to manage certificate categories.

3. Manage Agencies

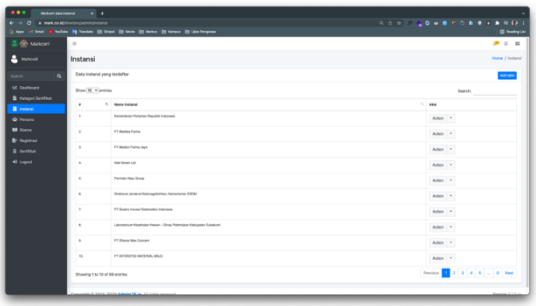


Figure 13 Agency Menu Display

Figure 13 is a user interface display for the management of the institution, serves to manage agency data

4. Manage Person

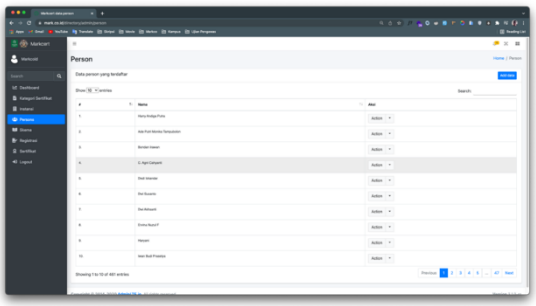


Figure 14 Display Menu Person

Figure 14 is the user interface display for the manage person menu, which functions to manage the person data.

5. Manage Registration

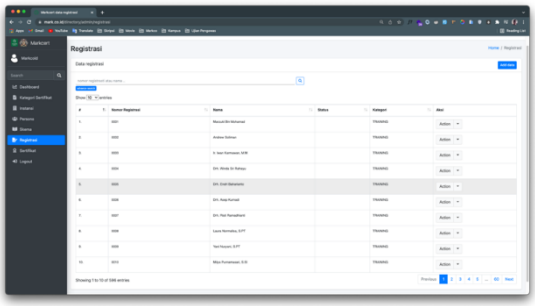


Figure 15 Display of the Registration Menu

Figure 15 is a user interface menu display manage registration, serves to manage registration data.

6. Manage Certificates

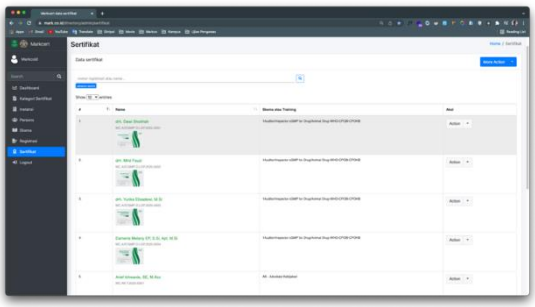


Figure 16 Certificate Menu Display

Figure 16 is a user interface display for the certificate management menu, which functions to manage certificate data

Testing

Blackbox testing is carried out to ensure whether all the performance of the system is running well.

Table 1 Testing the Login Menu

No	Test scenario	test case	Expected results	Test result	Conclusion
1	Clear all login forms then click login	Username : (blank) Password : (blank)	Access is denied by the system, a notification appears “ Username and password required ”	According to expectations	Valid
2	Fill in the Username while the password is empty	Username: (admin) Password: (blank)	Access is denied by the system, a message " password required " appears.	According to expectations	Valid
3	Fill in the password while the Username is empty	Username: (blank) Password: (1234 56)	Access is denied system, a notification appears “ Username required ”	According to expectations	Valid
4	Fill out the form completely but the password entered does not match	Username: (admin) Password: (wrong)	Access is denied by the system, a notification " wrong password " appears	According to expectations	Valid
5	Fill out the form completely but the User name entered does not match	Username: (wrong) Password: (1234 56)	Access is denied by the system, a notification “ User not found ” appears.	According to expectations	Valid
6	Fill out the form completely, the User name and password are entered according ly	Username: (admin) Password: (1234 56)	Access is accepted by the system, the user is redirected to the dashboard page	According to expectations	Valid

Table 2 Testing the Registration Menu

No	Test scenario	Test case	Expected results	Test result	Conclusion
1	Empty all forms add registration then click submit	Registration category: (Empty) Full Name : (Empty) Scheme or Training: (Empty) Batch : (Empty) Title of training: (Empty) Level : (Empty) Status: (Empty) Authority: (Empty)	Access is denied system, a notification appears “(Each field is empty) is required”	According to expectations	Valid
2	Filling Field batch with letters not with numbers	Batch: (two)	Access is denied by the system, a notification " Batch should be numeric " appears.	According to expectations	Valid
3	Fill in the registration form completel	Registration category: LSP Full Name:	Access is accepted by the system, a notificatio	According to expectations	Valid

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3059

	y and correctly	Agung Scheme or Training: BRCGS Batch: 1 Title of training: (Blank) Level: Interme diate Status: Passed Authorit y: Auditor	n appears "Registrati on successfull y added"		

Table 3 Testing the Certificate Menu

No	Test scenario	Test case	Expected results	Test result	Conclusion
1	Clear all the add certificate forms then click submit	Categor y: (blank) Registr ation Numb er: (blank) Date of issue: (Blank)	Access is denied system, a notificatio n appears "(Each Field is empty) is required "	Accord ing to expect ations	Valid
2	Fill in an unregiste red registrati on number	Registr ation Numb er : LSP01 92	Access denied system, pops up notificatio n "Number not found"	Accord ing to expect ations	Valid
3	Fill in the registrati on form completel y and correctly	Catego ry: LSP Registr ation Numb er: LSP19 28 Releas e date: 14/12 /2021 Expira tion date: 15/12 /2023 Status: Compe tent Busine ss card: name.j pg	Access is accepted by the system, a notificatio n appears "Certificate added successfull y"	Accord ing to expect ations	Valid

CONCLUSION

Based on the results of the discussion in this study, it can be concluded that the existence of a website-based certificate management information system can provide convenience for employees in managing data related to certificate issuance so that the process becomes more effective and efficient and the existing data can be backed up neatly and efficiently. systemized.

REFERENCE

. Nofyat, Adelina Ibrahim, and Arisandy Ambarita. 2018. "Sistem Informasi Pengaduan Pelanggan Air Berbasis Website Pada Pdam Kota Ternate." *IJIS - Indonesian Journal On Information System* 3(1). doi: 10.36549/ijis.v3i1.37.

Abdullah, Rohi. 2016. *Easy & Simple Web Programing*. Jakarta: PT Elex Media Komputindo.

Affandi, Much Irsyad, and Hudan Eka Rosyadi. 2019. "Perancangan Aplikasi Toko Online Al-Ihsan Berbasis Php & Mysql." *Seminar Nasional Sistem Informasi (SENASIF)* 3(1):2164–69.

Anna, Anna, Nurmalasari Nurmalasari, and Angelina Ella Yusnita. 2018. "Rancang Bangun Sistem Informasi Akuntansi Penerimaan Dan Pengeluaran Kas Pada Kantor Camat Pontianak Timur." *Jurnal Khatulistiwa Informatika* 6(2):107–18. doi: 10.31294/khatulistiwa.v6i2.153.

Arya, I. Komang, Ganda Wiguna, Ida Bagus, Gede Sarasvananda, Stmik Stikom Indonesia, Jl Tukad, Pakerisan No, and Kec Denpasar Sel. 2021. "Sosialisasi Penggunaan Sistem Informasi Sertifikat Elektronik Pada Lembaga Penjamanin Mutu Pendidikan Provinsi Bali Menurut Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia No 26 Tahun 2020 Tentang Organisasi Dan Tata Kerja Unit Pelaksana Te." 2(1):42–49.

Cordeaux, John. 1877. "'Wicks' of the Mouth." *Notes and Queries* s5-VII(159):37. doi: 10.1093/nq/s5-VII.159.37-a.

Edukom. n.d. *Pengenalan Internet*. Tangerang: Loka Aksara.

Hendri, Hendri, Dony Oscar, and Rachman Komarudin. 2020. "Implementasi Waterfall Model Pada Sistem Informasi Penyewaan Tanah Makam Pada Tpu Perwira." *Jurnal Infortech* 2(2):211–16. doi: 10.31294/infortech.v2i2.9214.

Heriyanto, Yunahar. 2018. "Perancangan Sistem Informasi Rental Mobil Berbasis Web Pada PT.APM Rent Car." *Jurnal Intra-Tech* 2(2):64–77.

Hirmawan, A., M. P, and D. Azizah. 2016. "ANALISIS SISTEM AKUNTANSI PENGGAJIAN DAN PENGUPAHAN KARYAWAN DALAM UPAYA MENDUKUNG PENGENDALIAN INTERN (Studi Pada PT.Wonojati Wijoyo Kediri)." *Jurnal Administrasi Bisnis S1 Universitas Brawijaya* 34(1):189–96.

Marisa, Fitri. 2017. *Web Programming (Client Side and Server Side)*. Yogyakarta: Deepublish.

Marlina, Masnur, and Muh. Dirga.F. 2021. "Aplikasi E-Learning Siswa Smk Berbasis Web." *JURNAL SINTAKS LOGIKA Vol.* 1(1):2775–412.

Maulana, Alief, Muhammad Sadikin, and Arief Izzuddin. 2018. "Implementasi Sistem Informasi Manajemen Inventaris Berbasis Web Di Pusat Teknologi Informasi Dan Komunikasi – BPPT." *Setrum : Sistem Kendali-Tenaga-Elektronika-Telekomunikasi-Komputer* 7(1):182. doi: 10.36055/setrum.v7i1.3727.

Nugraha, Ade Chandra. 2022. "Penerapan Teknologi Blockchain Dalam Lingkungan Pendidikan." *Produktif : Jurnal Ilmiah Pendidikan Teknologi Informasi* 4(1):302–7. doi: 10.35568/produktif.v4i1.386.

Rerung, Rintho Rante. 2018. *Pemograman Web Dasar*. Yogyakarta: Deepublish.

Retno, Sujacka, Novia Hasdyna, Mutasar Mutasar, and Rozzi Kesuma Dinata. 2020. "Algoritma Honey Encryption Dalam Sistem Pendataan Sertifikat Tanah Dan Bangunan Di Universitas Malikussaleh." *INFORMAL: Informatics Journal* 5(3):87. doi: 10.19184/isj.v5i3.20804.

Samala, Agariadne Dwinggo, and Bayu Ramadhani Fajri. 2021. "Rancang Bangun Aplikasi E-Sertifikat Berbasis Web Menggunakan Metode Pengembangan Waterfall." *Jurnal Teknik Informatika* 13(2):147–56. doi: 10.15408/jti.v13i2.16470.

Simangunsong, Agustina, and Manajemen Informatika. 2018. "Sistem Informasi Pengarsipan Dokumen Berbasis Web." *Jurnal Mantik Penusa* 2(1):11–19.

Trisyanto. 2018. *Analisis & Perancangan Sistem Basis Data*. Surabaya: Garuda Mas Sejahter.

INTEGRATION OF FUZZY LOGIC METHOD AND COCOMO II ALGORITHM TO IMPROVE PREDICTION TIMELINESS AND SOFTWARE DEVELOPMENT COST

Neneng Rachmalia Feta[1*]; Deki Satria[2]; Fitria[3]

Information Systems and Technology[1,2,3]
Bank Rakyat Indonesia Institute of Technology and Business
http://bri-institute.ac.id/
nenengrachmaliafeta@gmail.com[1*], deki.satria@bri-institute.ac.id[2], fitria.fitrib@gmail.com[3]

Abstract—*This study discusses improving the prediction of timeliness and cost of software development using the Constructive Cost Model II (COCOMO II) method and the application of Fuzzy Logic. And aims to obtain accurate time and cost prediction estimates on software development projects to obtain maximum cost results for a software development project. This study utilizes an adaptive fuzzy logic model to improve the timeliness of software development and cost estimates. Using the advantages of fuzzy set logic and producing accurate software attributes to increase the prediction of the time and price of software development. The fuzzy model uses the Two-D Gaussian Membership Function (2-D GMF) to make the software attributes more detailed in terms of the range of values. In COCOMO I, NASA98 data set; and four data projects from software companies in Indonesia were used to evaluate the proposed Fuzzy Logic COCOMO II, commonly known as FL-COCOMO II. Using the Mean of Magnitude of Relative Error (MMRE) evaluation technique, after experimenting using MF1, MF2, and MF3 treatments, the results obtained that MF1 with a value of 65.51 is a better treatment than MF2 with a value of 92.74 and MF3 with a value of 163.36 because the MF1 value has the smallest MMRE value among other treatments. While at the minimum MRE points, MF2 has the smallest value, namely 0%, compared to MF1 with a value of 0.02% and MF3 with a value of 70.12%.*

Keywords: COCOMO II, Fuzzy Logic, Software Cost Estimation, Gaussian Membership Function (2-D GMF).

Abstrak—*Penelitian ini membahas mengenai peningkatan prediksi ketepatan waktu dan biaya pengembangan perangkat lunak dengan menggunakan metode Constructive Cost Model II (COCOMO II) dan penerapan Logika Fuzzy. Serta bertujuan untuk memperoleh perkiraan prediksi ketepatan waktu dan biaya yang akurat pada proyek pengembangan perangkat lunak, sehingga dapat memperoleh hasil biaya yang maksimal untuk sebuah proyek pengembangan perangkat lunak. Penelitian ini menggunakan model logika fuzzy adaptif untuk meningkatkan ketepatan waktu pengembangan perangkat lunak dan perkiraan biaya. Menggunakan keuntungan dari logika himpunan set fuzzy serta menghasilkan atribut perangkat lunak yang akurat untuk meningkatkan prediksi ketepatan waktu dan biaya pengembangan perangkat lunak. Dua-Dimensi Gaussian Membership Function (2-D GMF) digunakan dalam model fuzzy untuk membuat atribut perangkat lunak lebih rinci dalam hal rentang nilai. Pada COCOMO I, NASA98 set data; dan empat proyek data dari perusahaan perangkat lunak di Indonesia digunakan dalam evaluasi yang diusulkan Fuzzy Logic COCOMO II atau yang biasa disebut FL-COCOMO II. Dengan menggunakan teknik evaluasi Mean of Magnitude of Relative Error (MMRE), setelah dilakukan percobaan menggunakan perlakuan MF1, MF2, dan MF3 didapatkan hasil bahwa MF1 dengan nilai 65,51 merupakan perlakuan yang lebih baik dibandingkan dengan MF2 dengan nilai 92,74 dan MF3 dengan nilai a nilai 163,36 karena nilai MF1 memiliki nilai MMRE paling kecil diantara perlakuan lainnya. Sedangkan pada titik MRE minimum, MF2 memiliki nilai terkecil yaitu 0%, dibandingkan dengan MF1 dengan nilai 0,02% dan MF3 dengan nilai 70,12%.*

Kata Kunci: *COCOMO II, Fuzzy Logic, Estimasi Biaya Perangkat Lunak, Gaussian Membership Function (2-D GMF).*

INTRODUCTION

Many software development project failures occur because the project is completed more than the planned cost and schedule and is a significant problem for software project managers. Poor forecasting causes projects to exceed budget and schedule and, in many cases, causes software development projects to fail. Between 30% and 40% of software, projects are ultimately completed outside budget or schedule, and more projects are canceled or fail (Pospieszny et al., 2018). Among the many reasons for failure, inaccuracy in

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3037

software estimates has been identified as the root cause of a high percentage of losses in software development (Singal et al., 2020).

Software cost prediction defines organizations' techniques and procedures to estimate bid proposals, project planning, and probability estimation (Christina & Banumathy, 2019). Thus, prediction accuracy is an important and significant issue for executives, managers, technical staff, and practitioners who perform or rely on cost prediction (Parwita et al., 2017). Unfortunately, accurate estimation of software development costs also challenges software engineering researchers due to the continued lack of precise estimates (Singal et al., 2020). The reasons that software cost estimation is complex and error-prone include (Christina & Banumathy, 2019):

a. Software cost estimation requires a significant amount of cost to be done correctly;
b. This process is often carried out in a hurry, with no cost calculations required to make estimates;
c. Experience is required to develop estimates, especially for large projects, and;
d. Human prediction

Some software development cost prediction models can be classified as algorithmic and non-algorithmic models. The model algorithm is based on a statistical analysis of historical data (Indra & Aqlani, 2018) (Subandri & Sarno, 2017), for example, Software Life Cycle Management (SLIM) (Kholed Langsari et al., 2018) and Constructive Cost Model (COCOMO) (Yadav, 2017). Non-algorithmic techniques based on new approaches such as Parkinson's, Expert Judgment, Price to-Win, and Machine Learning Approaches (K Langsari & Sarno, 2017b) (Sinha & Gora, 2021). Machine Learning Approaches represent facts from the human mind (K Langsari & Sarno, 2017a), for example, rule induction, fuzzy systems, genetic algorithms, artificial neural networks, bayesian networks, and evolutionary computing. These five approaches are classified into Soft Computing Groups. The importance of approximation from algorithmic and non-algorithmic techniques is discussed further in the following sections.

The well-known cost estimation algorithmic models are Boehm's COCOMO I and II (Singal et al., 2020), Albrecht's Function Point; and Putnam's SLIM (Kholed Langsari et al., 2018). This model requires input and accurate estimation of specific attributes, such as Source Line Of Code (SLOC), number of users, interface, complexity, etc. This is not easy to obtain in the early stages of software development. However, the formulas and calculations of this model are easy to understand and can also provide fast estimates compared to non-algorithmic models. In addition, the attributes and relationships used to estimate software development costs may change over time, and/or differ in the software development environment (Yadav, 2017). The limitations of algorithmic models lead to the exploration of non-algorithmic models.

In 1990, a non-algorithmic model was born and was proposed for software cost estimation. Software researchers have turned their attention to new approaches based on soft computing approaches such as artificial neural networks, fuzzy logic, and genetic algorithms. Fuzzy Logic (FL) offers a robust linguistic representation representing inaccuracies in input and output models and provides a more knowledge-based approach to constructing an effective model. Research shows that using FL can lead to better performance in reducing the inaccuracy of input and output parameters (K Langsari & Sarno, 2017b).

Gray and Macdonnell, in the study of Singal et al., compare popular techniques in software cost estimation, such as regression techniques, function point analysis (FPA), fuzzy logic, and artificial neural networks. Their research shows that the fuzzy logic model has produced better performance than other models (Singal et al., 2020). They introduced an application of fuzzy logic for cost estimation, which was used as a development tool, FUzzy logic SOftware MEasuring (FULSOME) (Iqbal & Sang, 2021), to assist software managers in decision making. The FULSOME model selects two critical variables: the complexity adjustment factor and the point function mismatch. Then triangular membership functions are defined for small, medium, and large interval sizes, complexity, and software effort.

The research of Raza, Fei, and Liu attempted to apply fuzzy logic to the algorithmic cost estimation model in dealing with uncertainty and imprecision problems in the model (Raza, 2019). They proposed a software fuzzy size model for COCOMO I. This study found an unusual input fuzzy rule linguistic variable when setting the determinate value for the software attribute measure in COCOMO I because an accurate estimate of Kilo Delivered Source Instructions (KDSI) could not do before starting the project.

Riyanarto applies fuzzy modelling techniques to COCOMO I and the Point Function Model (Sarno et al., 2015). Murad and Goyal et al. investigated the application of fuzzy logic to Effort Multipliers (EM) among COCOMO I (Murad et al., 2021) (Goyal et al., 2015). Also, fuzzification of COCOMO I without considering the adjustment factor, so they introduced F-COCOMO. They gave the software a new measure of COCOMO I, and the coefficients associated with the development mode

were assigned to a fuzzy set. In another study, Kumar et al. in Bedi's research applied fuzzy logic to the Manpower Buildup Index (MBI) of the Putnam estimation model based on 64 different rules. Bedi's research results show that fuzzy logic can be effectively applied to software cost estimation (Bedi & Singh, 2017).

Fuzzy logic has also been applied to non-algorithmic models to overcome model uncertainty. For example, Indra and Aqlani proposed a combination of estimation from the fuzzy logic model with the analogy technique (Indra & Aqlani, 2018). Analogy estimation is one of the expert-based classification techniques, and this is a type of Case-based Reasoning (CBR) method (Zhang, 2019). In addition, a fuzzy analogy for software cost estimation has also been applied to web-based software.

In summary, fuzzy logic has been applied to algorithmic and non-algorithmic cost estimation models to achieve better estimation results. However, there is still a lot of uncertainty about what technique is used to look at the type of estimation problem (Raza, 2019). Therefore, choosing between different approaches is a difficult decision that requires the support of a well-defined evaluation method to demonstrate each estimation technique and apply it to estimation problems (Singal et al., 2020).

For decades, accurate software cost prediction has been essential for software development projects. However, inaccurate estimates in leading the project can exceed the budget and schedule, and even in many cases, the project can be stopped completely (Indra & Aqlani, 2018). The ability to accurately estimate development time, costs, labour, and new methodological changes can replace older software development cost prediction models. Therefore, an accurate software development cost prediction model will be needed in software development project management.

This study proposes an effective Fuzzy Decision Tree model for embedding in COCOMO II to overcome the ambiguity and uncertainty of software attributes, resulting in more accurate estimation results. The steps are taken to apply the estimate: strategic planning, feasibility studies, system specifications, evaluation of supplier proposals, and software development project planning (Baiquni et al., 2017).

MATERIALS AND METHODS

The research method describes the approach to calculating software cost predictions with COCOMO II and the Fuzzy Logic approach and will then discuss the effect of Fuzzy Logic on COCOMO II or what we call FL-COCOMO II.

A. Model COCOMO II

In Yadav's research, the COCOMO I model is a regression-based estimation software cost prediction model developed by Boehm in 1981 (Yadav, 2017) and is considered the best known and the most reasonable model among all traditional cost estimation models. The COCOMO I was the most stable model at the time. One of the problems with using COCOMO I at the time was that it was incompatible with the development environment of the late 1990s. Therefore, in 1997, Boehm developed it into COCOMO II to solve most of the problems of COCOMO I (Baiquni et al., 2017). Figure 1 shows the software schedule, cost, and labour estimation formulas and processes in COCOMO II. Equation (1) shows the exact formula regarding Efforts or efforts that can be made. Equation (2) discusses the formula regarding the software scheduling process. Equation (3) shows the formula for estimating labour staff in software development. Equation (4) shows the formula for calculating software development costs (K Langsari & Sarno, 2017b). COCOMO II includes several software attributes such as 17 Effort Multipliers (EMs), 5 Scale Factors (SFs), Software Size (Software Size), and effort estimates used in the Post Architecture Model COCOMO II (Putri et al., 2017). Descriptions of the 17 EM and 5 SFS based on their numerical values and productivity ranges are shown in Table 1 (Baiquni et al., 2017).

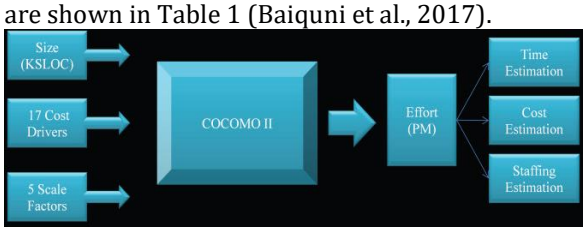


Figure 1. Software schedule, cost, and labour estimation process in COCOMO II (Putri et al., 2017)

$$Effort_{PM} = Ax[Size]^{B+0.01x\sum_{j=1}^{5} SF_j} x[i = 1] \; 17 \; \Pi \; EM \quad \text{.....(1)}$$

$$Schedule_{Months} = Cx\,(Effort)^{D+0.2x\,0.01x \sum_{j=1}^{5} SF_j} \quad \text{.....(2)}$$

$$Average\ staffing_{People} = \frac{Effort}{Schedule} \quad \text{.....(3)}$$

$$\text{COST} = Effort\ x\ (Payment_{Month}) \quad \text{.....(4)}$$

A = 2.94; B = 0.91; C = 3.67; D = 0.28

Size: Software Size (SLOC) (K Langsari & Sarno, 2017b)

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3037

The ambiguity and uncertainty of software attributes can impact software estimates. Thus, accurate software attributes produce special software estimates that software project managers and organizations most desire.

Table 1. COCOMO II EMs Range

No	Effort Multiplier	Range
1	Required software reliability (RELY)	0.82-1.26
2	Database size (DATA)	0.90-1.28
3	Product complexity (CPLX)	0.73-1.74
4	Developed for reusability (RUSE)	0.95-1.24
5	Documentation match to life-cycle needs (DOCU)	0.81-1.23
6	Execution time constraint (TIME)	1.00-1.63
7	Main storage constraint (STOR)	1.00-1.46
8	Platform volatility (PVOL)	0.87-1.30
9	Analyst capability (ACAP)	1.42-0.71
10	Programmer capability (PCAP)	1.34-0.76
11	Personnel continuity (PCON)	1.29-0.81
12	Applications experience (APEX)	1.22-0.81
13	Platform experience (PLEX)	1.19-0.85
14	Language and tool experience (LTEX)	1.20-0.84
15	Use of software tools (TOOL)	1.17-0.78
16	Multisite development (SITE)	1.22-0.80
17	Required development schedule (SCED)	1.43-1.00

Table 1 represents the variables or factors that affect software estimates. In determining the factors that influence software attributes, it can be done by reviewing the existing literature based on previous studies related to software estimation. For example, regarding the COCOMO II Architecture model, these parameters were determined based on the use of COCOMO'81 and the experience of a group of senior software cost analysts.

The value of SFs is based on the premise that they are a significant source of exponential variation in project effort and product variation.

B. Fuzzy Logic (FL)

In 1965, in Kholed's research, Lotfi Zadeh officially developed the multi-value theory and introduced the term fuzzy into the engineering literature (Kholed Langsari et al., 2018). Fuzzy Logic (FL) started using the concept of fuzzy set theory. Fuzzy theory is a class theory with unclear boundaries and will be seen as an extension of the set of a classical theory (Kaur et al., 2018). The membership $\mu A(x)$ of element X of the classical set A, as part of the universe X, is defined by equation (5), as follows:

$$\mu A(x) = \begin{cases} 1 \text{ if } x \in A \\ 2 \text{ if } x \notin A \end{cases} \quad \text{..... (5)}$$

A system based on FL has a direct relationship with fuzzy concepts (such as fuzzy sets, linguistic variables, etc.) and fuzzy logic. The popular fuzzy logic systems can be categorized into three types: Pure fuzzy logic systems, Takagi and Sugeno fuzzy systems, and fuzzy logic systems with fuzzified and defuzzifiers (Kaur et al., 2018). Since most engineering applications produce crisp data as input and expect crisp data as output, the last type is the most widely used type of fuzzy logic system. The first fuzzy logic system with a fuzzifier and defuzzifier was proposed by Mamdani and has been successfully applied to various industrial processes and consumer products (Kaur et al., 2018). The three main steps of using fuzzy logic in a model are:

Step 1: Fuzzification: convert input crips to fuzzy set.

Step 2: Fuzzy Rule-Based System: Fuzzy logic system using fuzzy IF-THEN rules; Fuzzy Inference Engine: After all the input crips values are fuzzified into the respective linguistic values, the inference engine accesses the fuzzy rule base to obtain linguistic values for intermediate and output linguistic variables.

Step 3: Defuzzification: converts the fuzzy output into crisp output (Raza, 2019).

C. Fuzzy Logic COCOMO II (FL-COCOMO II)

FL-COCOMO II is based on COCOMO II and Fuzzy Logic. COCOMO II includes a set of software attribute inputs: 17 EM, 5 SFS, 1 SS, and one output, estimated effort. The FL-COCOMO II architecture is shown in Figure 2.

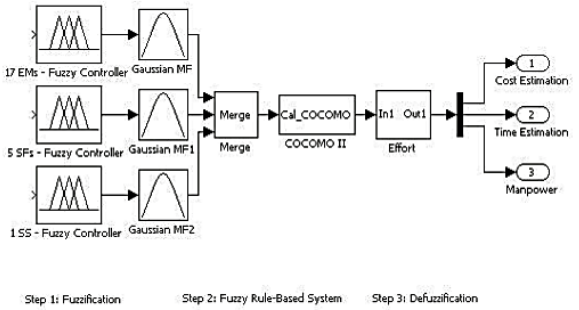


Figure 2. FL-COCOMO II architecture includes: 17 EMs, 5 SFs, 1 SS, and estimated effort

In COCOMO II, the effort is expressed as Person Months (PM). Determination of the effort required by a software development project based on the size of the software project is Kilo Source Lines of Code (KSLOC). Traditionally, software effort estimation problems have relied on single (numeric) values of EMs, SFs, and software sizes given as software attributes to estimate effort. However, software size can be calculated based on previously developed software similar to the current one (especially at the start of the project). Although the truth and accuracy of such estimates are minimal, it is fundamentally important to recognize the situation and with a technique that can evaluate the associated imprecision that is in the final cost estimate.

The software size can be determined using fuzzy sets in EMs, SFs, and attributes by dividing possible values rather than using fixed values. Generally, this distribution form is represented in the form of a fuzzy set. It is essential to clarify that ambiguity and uncertainty at the input level of COCOMO II produce uncertainty at the output level (Singal et al., 2020). Converting software attributes for each fuzzy set can increase software attributes' accuracy, resulting in very accurate estimates. On the other hand, inaccurate input estimates can lead to less detailed effort estimates. The overlapping membership functions of bell-shaped cost drivers and trapezoidal cost drivers change the fuzzy model to a more precise model.

In addition, it is possible that when using the membership function bell-shaped cost drivers and trapezoidal cost drivers, there are several attributes assigned to the maximum level of compatibility instead of being transferred to the lowest degree. To avoid this linearity, it is proposed to use better functions, membership functions of bell-shaped cost drivers and trapezoidal cost drivers, to represent the model's inputs. 2-D GMF is represented by equation (6) as follows:

$$\mu A_i(x) = Gaussian\,(x, c_i, \sigma_i) = e^{\frac{-(x-c_i)^2}{2\,\sigma_i 2}} \text{.....(6)}$$

Where c_i is the midpoint of i^{th} of the fuzzy set and σ_i is the width of i^{th} fuzzy set (Baiquni et al., 2017).

Applying fuzzy logic to COCOMO II to construct FL-COCOMO II is described as follows. The three main processes in FL-COCOMO II are Fuzzification, Fuzzy Rule-Based/Fuzzy Inference Engine, and Defuzzification. Software attributes in COCOMO II are converted to fuzzy variables based on the fuzzification process with terms and conditions Extra Low (XL), Very Low (VL), Low (L), Nominal (N), High (H), Very High (VH)), and Extra High (XH) were defined for the 11 software attributes (17 EMs, 5 SFs, and 1 SS) assigned to each software attribute. The fuzzy sets corresponding to various language-related values for each software attribute are defined using 2-D GMF. For example, fuzzification of Applications Experience (AEXP) EM is based on the membership function bell-shaped cost drivers and trapezoidal cost drivers function, using the Fuzzy Inference System tool in the MATLAB software; the definitions are shown in Table 2 and Figure 3. Fuzzy Inference System (FIS) is a fuzzy tool in the MATLAB software used in the fuzzification, fuzzy calculations, fuzzy rules, and defuzzification process of FL-COCOMO II. FIS supports the Mamdani fuzzy method and the Sugeno fuzzy method. FLCOCOMO II is based on the Sugeno fuzzy system, which is more accurate than the FIS Mandani method.

Table 2. Applications Experience (AEXP) EM Description

Effort Multiplier: Applications Experience (AEXP)						
Descriptors	2 months	6 months	1 year	3 years	6 years	-
Effort Multipliers	1.42	1.19	1	0.85	0.71	-
Rating	VL	L	N	H	VH	EX

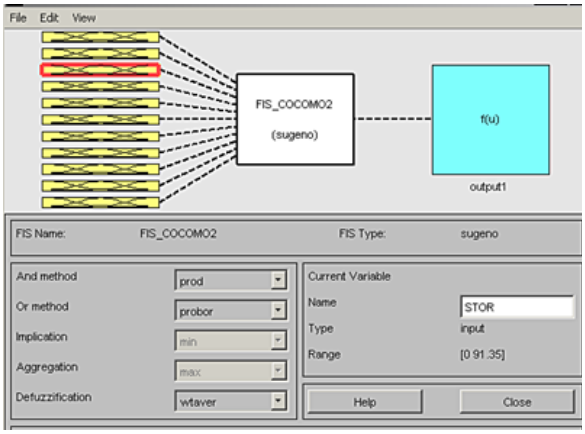


Figure 3. Fuzzification Process in MATLAB

In the first step, all software attributes of COCOMO II are converted in response to the fuzzy set and its variables (***FuzzyEMij***) instead of using fixed values of ***EMij***. ***FuzzyEMij*** is calculated using equation (7). The original ***EMij*** value and the membership function bell-shaped cost drivers and trapezoidal cost drivers μ are defined for various fuzzy sets related to EMs, SFs, and SS. This process helps to reduce the ambiguity and uncertainty of software attributes at this level.

$$\boldsymbol{FuzzyEMij} = \boldsymbol{F}\left(\mu_{A1}^{V1}, \ldots\ \mu_{Ai}^{Vi}, \boldsymbol{EM}_{i1} \ldots \boldsymbol{EM}_{ij}\right) \text{.....(7)}$$

For convenience, F is taken as a linear function, where μ_{A1}^{V1} the membership function of the fuzzy set Aj is related to the controlling value Vi, as shown in equation (8).

$$\boldsymbol{FuzzyEMij} = [\boldsymbol{j} = 1]\, k_i \sum \mu_{A1}^{V1} * \boldsymbol{EM}_{ij} \text{ .....(8)}$$

The fuzzy rules for FL-COCOMO II are defined through linguistic variables in the fuzzification process. It is important to note that the fuzzy rules are adapted to all functions of the degree of precision, according to the test and the characteristics of the project. Fuzzy rules are defined based on the "AND" and "OR" relationships or their combination between the input variables, as shown below:

Aturan Fuzzy:
IF TOOL is Low THEN effort is Low
IF PCAP is Very Low THEN effort is Very High
IF RUSE is Nominal THEN effort is Nominal

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3037

IF DATA is Very High THEN effort is Very High
. . . (Bedi & Singh, 2017)

The number of rules defined for FL-COCOMO II is more than 193 based on input variables. In applying fuzzy rules from FL-COCOMO II, the FIS tool in the MATLAB software is used, as shown in Figure 4. The process carried out is to input the fuzzy rules. Then, the fuzzy input rules are adjusted to The Range of COCOMO II Ems in Table 1.

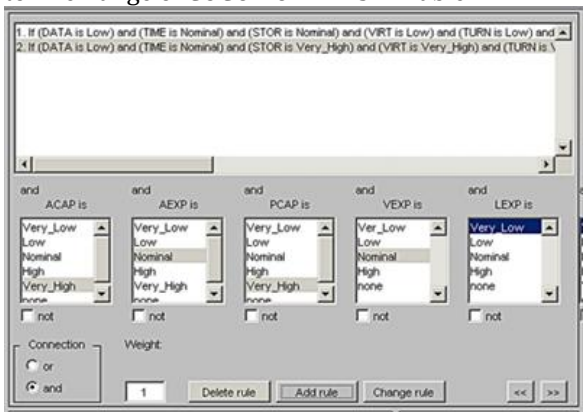


Figure 4. The process of applying fuzzy rules in MATLAB

The last step is the defuzzification of the effort variable using defuzzification techniques such as Mean of Maximum (MOM), Center of Area (COA), and First of Maximum (FOM). Defuzzification of the "Effort" output is carried out using the Mean of Maximum (MOM) technique because the results obtained are more accurate when compared to other defuzzification techniques.

RESULTS AND DISCUSSION

Data set is a collection of data used in software development project data (Bedi & Singh, 2017). Two data sets are used in the research: NASA data consisting of 93 project data collected from 2000 to 2016 and project data as many as 60 data collected from 2008 to 2018. But from 2008 to 2018 data, some data overlap. Before calculation, the data set is converted into an Excel file for easy analysis.

FL-COCOMO II was evaluated through two sources; namely, the data used for testing the fuzzy method is the NASA data group which is 102 project data from 2000 to 2018, and data for 15 types of COCOMO cost drivers, number of lines of program code, project type, and actual effort. Project development.

A. Dataset 1

COCOMO data set I-Boehm (Singal et al., 2020) was the first researcher to look at software engineering from an economics point of view and its relationship to model cost estimation from the dataset. The COCOMO I dataset includes 63 project history data. Data is available at http://openscience.us/.

B. Dataset 2

NASA93 data set NASA93 data set includes 93 project data from NASA, which is the data element measurement benchmarking manager. This data is available at http://openscience.us/

C. Metode Evaluasi

Enterprises in the COCOMO model are described as Person Months (PM). PM is the effort required by a person or group to complete a project. This study uses the Magnitude of Relative Error (MRE) as shown in equation (9) and the Mean Magnitude of Relative Error (MMRE) formula shown in equation 10 to evaluate research results (Suherman et al., 2020). In addition to the two formulas, the study also looked at the maximum, minimum, median, and standard deviation values of the MRE.

$$MRE = \frac{Actual\ Effort - Predictive\ Effort}{Actual\ Effort} \quad \text{.....(9)}$$

$$MMRE = \frac{1}{N}\sum_{i}^{N} \mathrm{MRE}_i \quad \text{.....(10)}$$

The 11 variables with effort multipliers each have a value from 0.70 to 1.46, multiplied by the cost drivers. Effort multipliers were studied by Boehm in 1981 after project regression analysis in the COCOMO I data set. However, not all cost drivers are defined in fuzzy sets because the cost drivers are only an ordinary description (Tahir & Adil, 2018). For example, the cost drivers are RELY, CPLX, MODP, and TOOL (Murad et al., 2021). As for the number of lines of program code in this study using the Kilo Size Line Of Code (KSLOC).

Next, determine the membership function for the fuzzification process. The membership function used is taken from the journal Cost Model Using Fuzzy Logic (Goyal et al., 2015). The Fuzzy Logic Toolbox assists the creation of this membership function in MATLAB. The membership function in the journal is a trapezoidal membership function which can be seen in Figure 5. The membership function graph depicts the range of each cost driver. Membership function points are different for each cost driver. The treatment with the trapezoidal membership function is the MF1 treatment.

This study will compare the use of the trapezoidal membership function with the bell-shaped membership function. For the treatment of MF2 using a bell-shaped membership function. The bell-shaped membership function is obtained with the help of MATLAB by changing the trapezoidal so that the bell-shaped points and function graphs are obtained. The bell-shaped membership function is

made by looking at the interval approach, similar to the trapezoidal membership function. However, the number of points used in the bell-shaped is different from the trapezoidal one. The trapezoidal membership function uses four points to create a graph, while the bell-shaped one only requires three points.

The last treatment, namely MF3, combines trapezoidal and bell-shaped membership functions. This treatment aims to find the best model. A bell-shaped membership function is used for the 'nominal' category, while the 'very low', 'low', 'high', and 'very high' types use a trapezoidal membership function. For MF3, the 'nominal' category was chosen for the cost drivers to be converted into a bell-shaped function, considering that the 'nominal' type has a wide interval. The data entered is processed with fuzzy logic and will be calculated using COCOMO II calculations. The results of COCOMO II calculations can be seen in Table 3. After the results of the estimated effort are obtained for each treatment, the accuracy will be measured.

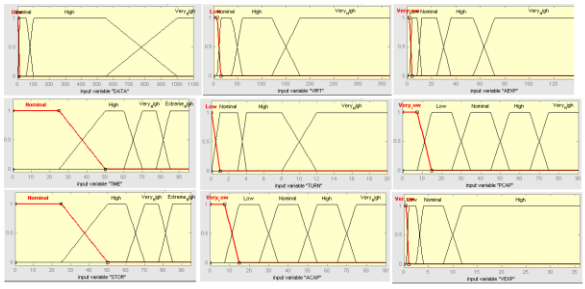


Figure 5. Trapezoidal membership function cost drivers

Table 3. COCOMO II calculation results

No	Cost Driver	Kategori	Effort Multipliers
1	RELY	High	1.15
2	DATA	Low	0.94
3	CPLX	High	1.15
4	TIME	Nominal	1.00
5	STOR	Nominal	1.00
6	VIRT	Low	0,87
7	TURN	Low	0,87
8	ACAP	Nominal	1,00
9	AEXP	Nominal	1,00
10	PCAP	Nominal	1,00
11	VEXP	Nominal	1,00
12	LEXP	Nominal	1,00
13	MODP	High	0.91
14	TOOL	Nominal	1,00
15	SCED	Low	1.08

The estimated results of the project development effort will be compared with the actual project effort. To measure the accuracy of the proposed business estimates, this study uses MRE. MRE looks at how close the estimated business results are to the actual effort. The smaller the MRE value, the better the business forecast results. The graph of the comparison of the MRE values for each treatment can be seen in Figure 6. This study used a threshold value of 20%. From the graph of the MRE value in the MF1 treatment, only 32 projects met the threshold value. While in MF2, there are 17 projects, and in MF3, there are no projects that meet the threshold. But most of the projects do not meet the specified threshold.

The smaller the MRE value, the closer the software effort estimate is to the actual business value. To measure the accuracy of the data set, the average MRE (MMRE) value was used for the three treatments. The results of the three treatments can be seen in Table 5. The MMRE value of the MF1 treatment is the most minor compared to other treatments. Then the MF1 treatment has a better predictive result. The maximum value and standard deviation of MRE in the MF1 treatment were also better than in MF2 and MF3 treatments. The project belongs to the semi-detached type from the table above, with the coefficient values shown in Table 4.

Number of Program Code Lines (KLSOC) = 25.9
Actual effort of project development = 117.6

For example, from the sample project data in Table 3 above, the business value will be calculated using the COCOMO II model as follows:

Table 4. Coefficient Score

No	Coefficient	Score
1	Coefficient A	3
2	Coefficient B	1.12

Effort Adjustment Factor (EAF) =
1.15 * 0.94 * 1.15 * 1.00 * 1.00 * 0.87 * 0.87 * 1.00 * 1.00 * 1.00 * 0.91 * 1.00 * 1.08 = 0.925.

Effort Adjustment Factor (Effort) =
EAF * Coefficient A * KSLOC $^{\text{Coefficient B}}$ =
$0.925 * 3 * 25.9^{1.12}$ =
106.208
(Person−Months)

Table 5. Evaluation Results on Treatment of MF1, MF2, MF3

No	Treatment	MMRE (%)	Min MRE (%)	Max MRE (%)	MRE Standard Deviation (%)
1	MF1	65.51	0.02	1014.53	132.30
2	MF2	92.74	0.00	1039.62	139.55
3	MF3	163.36	70.12	5316.87	526.91

In comparing treatments from the MMRE value, it was found that MF1 has the smallest value, then MF2 with a difference of 27.23%, and MF3 with the most significant MMRE value can be seen in Figure 6. The graph depicts the accuracy level of

MF1, which is higher than the other two treatments. Because if the MMRE value is taller, the difference in the estimated results will be further away from the actual value of the business.

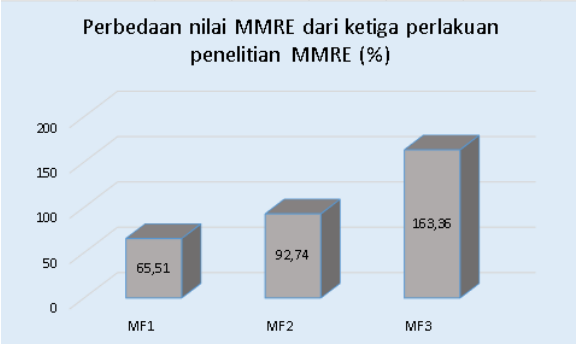


Figure 6. Graph of differences in MMRE values of the three treatments

The MF1 treatment had a minimum MRE value of 0.02%, but the MF2 treatment had a smaller MRE minimum value of 0%. The minimum MRE value percentage comparison shows that MF2 can reach the data better, even though MF1 has good results in the overall MRE calculation.

CONCLUSION

After experimenting using MF1, MF2, and MF3 treatments, the results obtained that MF1 with a value of 65.51 is a better treatment than MF2 with a value of 92.74 and MF3 with a value of 163.36 because the MF1 value has the smallest MMRE value among other treatments. While at the minimum MRE points, MF2 has the smallest value, namely 0%, compared to MF1 with a value of 0.02% and MF3 with a value of 70.12%. It can be concluded that the trapezoidal membership function provides better accuracy than using a bell-shaped. But the bell-shaped has the smallest MRE value. The rapid growth of technology causes more and more new methods and logic that effectively solve a problem. Future research is expected to collaborate with NASA data-providing institutions to test the business forecast model. The new study is also likely to provide more recent data.

REFERENCES

Baiquni, M., Sarno, R., Sarwosri, & Sholiq. (2017). Improving the accuracy of COCOMO II using fuzzy logic and local calibration method. *2017 3rd International Conference on Science in Information Technology (ICSITech)*, 284–289. https://doi.org/10.1109/ICSITech.2017.8257126

Bedi, R. P. S., & Singh, A. (2017). Software Cost Estimation using Fuzzy Logic Technique. *Indian Journal of Science and Technology*, *10*(3). https://doi.org/10.17485/ijst/2017/v10i3/109997

Christina, M. A., & Banumathy, C. (2019). Software cost estimation using neuro fuzzy logic Framework. *International Journal of Research in Engineering, Science and Management*, *2*(1), 219–224.

Goyal, M. V, Satapathy, S. M., & Rath, S. K. (2015). Software project risk assessment based on cost drivers and Neuro-Fuzzy technique. *International Conference on Computing, Communication & Automation*, 823–827. https://doi.org/10.1109/CCAA.2015.7148487

Indra, M., & Aqlani, Z. (2018). Comparative Analyisis of Software Cost Estimation Project using Algorithmic Method. *Engineering Software Requirements*, *1*(1), 17–27.

Iqbal, N., & Sang, J. (2021). Fuzzy Logic Testing Approach for Measuring Software Completeness. *Symmetry*, *13*, 604. https://doi.org/10.3390/sym13040604

Kaur, I., Narula, G. S., Wason, R., Jain, V., & Baliyan, A. (2018). Neuro fuzzy—COCOMO II model for software cost estimation. *International Journal of Information Technology*, *10*(2), 181–187. https://doi.org/10.1007/s41870-018-0083-6

Langsari, K, & Sarno, R. (2017a). Optimizing COCOMO II parameters using particle swarm method. *2017 3rd International Conference on Science in Information Technology (ICSITech)*, 29–34. https://doi.org/10.1109/ICSITech.2017.8257081

Langsari, K, & Sarno, R. (2017b). Optimizing effort and time parameters of COCOMO II estimation using fuzzy multi-objective PSO. *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 1–6. https://doi.org/10.1109/EECSI.2017.8239157

Langsari, Kholed, Sarno, R., & Sholiq. (2018). Optimizing time and effort parameters of COCOMO II using fuzzy Multi-objective Particle Swarm Optimization. *Telkomnika (Telecommunication Computing Electronics and Control)*, *16*(5), 2199–2207. https://doi.org/10.12928/TELKOMNIKA.v16i5.9698

Murad, M. A., Abdullah, N. A. S., & Rosli, M. M. (2021). Software Cost Estimation for Mobile Application Development-A Comparative Study of COCOMO Models. *2021 IEEE 11th International Conference on System Engineering and Technology, ICSET 2021 - Proceedings*.

https://doi.org/10.1109/ICSET53708.2021.9612528

Parwita, I. M. M., Sarno, R., & Puspaningrum, A. (2017). Optimization of COCOMO II coefficients using Cuckoo optimization algorithm to improve the accuracy of effort estimation. *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, 99–104. https://doi.org/10.1109/ICTS.2017.8265653

Pospieszny, P., Czarnacka-Chrobot, B., & Kobylinski, A. (2018). An effective approach for software project effort and duration estimation with machine learning algorithms. *Journal of Systems and Software*. https://doi.org/10.1016/j.jss.2017.11.066

Putri, R. R., Sarno, R., Siahaan, D., Ahmadiyah, A., & Rochimah, S. (2017). Accuracy Improvement of the Estimations Effort in Constructive Cost Model II Based on Logic Model of Fuzzy. *Advanced Science Letters*, *23*, 2478–2480. https://doi.org/10.1166/asl.2017.8767

Raza, K. (2019). Fuzzy logic based approaches for gene regulatory network inference. *Artificial Intelligence in Medicine*, *97*, 189–203. https://doi.org/10.1016/j.artmed.2018.12.004

Sarno, R., Sidabutar, J., & Sarwosri. (2015). Improving the accuracy of COCOMO's effort estimation based on neural networks and fuzzy logic model. *2015 International Conference on Information & Communication Technology and Systems (ICTS)*, 197–202. https://doi.org/10.1109/ICTS.2015.7379898

Singal, P., Kumari, A. C., & Sharma, P. (2020). Estimation of Software Development Effort: A Differential Evolution Approach. *Procedia Computer Science*, *167*(2019), 2643–2652. https://doi.org/10.1016/j.procs.2020.03.343

Sinha, R. R., & Gora, R. K. (2021). Software effort estimation using machine learning techniques. In *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-981-15-5421-6_8

Subandri, M. A., & Sarno, R. (2017). Cyclomatic Complexity for Determining Product Complexity Level in COCOMO II. *Procedia Computer Science*, *124*, 478–486. https://doi.org/10.1016/j.procs.2017.12.180

Suherman, I. C., Sarno, R., & Sholiq. (2020). Implementation of Random Forest Regression for COCOMO II Effort Estimation. *2020 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 476–481. https://doi.org/10.1109/iSemantic50169.2020.9234269

Tahir, F., & Adil, M. (2018). An Empirical Analysis of Cost Estimation Models on Undergraduate Projects Using COCOMO II. *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 1–5. https://doi.org/10.1109/ICSCEE.2018.8538361

Yadav, R. (2017). OPTIMIZED MODEL FOR SOFTWARE EFFORT ESTIMATION USING COCOMO-2 METRICS WITH FUZZY LOGIC. *International Journal of Advanced Research in Computer Science*, *8*, 121–125. https://doi.org/10.26483/ijarcs.v8i7.4113

Zhang, L. (2019). The Research on General Case-Based Reasoning Method Based on TF-IDF. *2019 2nd International Conference on Safety Produce Informatization (IICSPI)*, 670–673. https://doi.org/10.1109/IICSPI48186.2019.9095927

COMPARATIVE CLASSIFICATION OF LUNG X-RAY IMAGES WITH CONVOLUTIONAL NEURAL NETWORK, VGG16, DENSENET121

Muhammad Ilham Prasetya [1*)] ; Yuris Alkhalifi [2] ; Rifki Sadikin [3] ; Yan Rianto [4]

[1,3,4] Computer Science
[1,3,4] Nusa Mandiri University
[1,3,4] www.nusamandiri.ac.id
[1] 14002329@nusamandiri.ac.id[*]), [3] rifki.rdq@nusamandiri.ac.id, [4] yan.yrt@nusamandiri.ac.id

[2] Computer Technology
[2] Bina Sarana Informatika University
[2] www.bsi.ac.id
[2] yuris.yak@bsi.ac.id

Abstract *— Lungs are one of the organs of the human body, and lung tissue will ultimately affect human abilities. The respiratory system exchanges oxygen and carbon dioxide in the blood. Problems that often occur are polluted air quality, many bacteria that attack the lungs, and lung disease that can cause shortness of breath, mobility difficulties, and hypoxia so that if not detected immediately, it can cause death. In this regard, this study aims to compare the classification of normal lungs with those of those suffering from Cardiomegaly. The preparation of this dataset contributes to improving the quality of the disease classification system on X-ray images. CNN, VGG 16, and DenseNet methods were chosen as classification methods to ensure performance and determine which method is the best for classifying Lung Diseases. It can be concluded that by using the DenseNet121 model, X-Ray images in this research dataset get an accuracy of 67.06%. For the VGG16 model, it gets an accuracy of 68.94%, and for the CNN model, it gets the highest accuracy of 80.54%.*

Keywords: *Chest X-Ray, CNN, DenseNet121, VGG16*

Abstract — Paru-paru merupakan salah satu organ tubuh manusia, dan jaringan paru-paru pada akhirnya akan mempengaruhi kemampuan manusia. Sistem pernapasan menukar oksigen dan karbon dioksida di dalam darah. Masalah yang sering terjadi adalah kualitas udara yang tercemar, banyak bakteri yang menyerang paru-paru, dan penyakit paru-paru dapat menyebabkan sesak napas, kesulitan mobilitas, dan hipoksia, sehingga jika tidak segera terdeteksi dapat menyebabkan kematian. Sehubungan dengan hal tersebut, tujuan penelitian yaitu komparasi klasifikasi paru-paru normal dengan paru-paru yang menderita Cardiomegaly. Penyusunan dataset ini sebagai bentuk kontribusi dalam meningkatkan kualitas sistem klasifikasi penyakit pada citra X-ray. Metode CNN, VGG 16 dan DenseNet dipilih sebagai metode klasifikasi guna memastikan kinerja dan metode tersebut mana yang paling terbaik untuk melakukan klasifikasi Penyakit Paru – Paru. Dapat disimpulkan bahwa dengan menggunakan model DenseNet121, citra X-Ray pada dataset penelitian ini mendapatkan akurasi sebesar 67,06%, untuk model VGG16 mendapatkan akurasi sebesar 68,94% dan untuk model CNN mendapatkan akurasi tertinggi yakni sebesar 80,54%.

Keywords*: Lung X-Ray, CNN, DenseNet121, VGG16*

PRELIMINARY

Technological developments in image processing are currently beneficial in the health or medical world, especially for detecting a human diseases. (Cholik, 2021)(Hadianti & Riana, 2018). Currently, there are many studies that explain image processing in the world of Health to detect a disease, such as skin disease (Triyono et al., 2021) breast cancer (Hilaliyah, 2021), cervical cancer (Agustyawati et al., 2021). , eye disease (Indraswari et al., 2022), Alzheimer's disease (Phiadelvira, 2021), Dental Abscess Disease (Setiaji et al., 2018) and many more. From several studies that have been mentioned, it is stated that current technological developments are very clearly useful and help experts to process disease detection through images. Image processing plays a role as image processing using a computer, becoming a better quality image. One of the image processing operations is object recognition from a digital image. An essential process in recognizing objects that are presented visually or in the form of prints is segmentation (Gao et al., 2018)

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3010

The lungs are one of the main organs in the human body that function for the respiratory system or the respiratory process (Idatin Nikmah et al., 2019). Problems that often occur in the respiratory system are polluted air quality, so many bacteria attack the lungs, lung disorders can cause sufferers to have difficulty breathing, have a problem doing activities, and lack of oxygen so that if it is not detected quickly, it can cause death (Jaisakthi et al., 2019).

Research by (Baltruschat et al., 2018) conducted comparisons of deep learning on approaches to X-ray classification. This study used the extended model of the ResNet-50 architecture. In a systematic evaluation, using 5-fold re-sampling and multi-label loss functions, comparing the performance of different approaches to pathology classification with ROC statistics, and analyzing differences between classifiers using rank correlation. Overall, the researchers observed a fair amount of spread in the performance achieved and concluded that ResNet-38 X-ray-only, integrating non-image data yielded the best overall results. Next, a class activation map is used to understand the classification process, and a detailed analysis of the impact of non-image features is provided.

Another research by (Bharati et al., 2020) introduces that Detecting Lung Disease in a Timely is very important. In a lot of image processing and image modeling, in this study, Various forms of deep learning existing techniques including *convolutional neural network* (CNN), *vanilla neural network*, *visual geometry group based* neural network (VGG), and capsule network, are applied for the prediction of lung disease. Therefore, the researcher proposes a new deep *hybrid learning framework* combining VGG, data augmentation, and spatial transformer network (STN) with CNN. This new *hybrid* method is referred to here as VGG Data STN with CNN (VDSNet). As an implementation tool, with a complete and sampled data set, VDSNet outperforms existing methods in terms of several metrics including precision, gain, F0.5 score, and validation accuracy. For the entire dataset case, VDSNet showed a validation accuracy of 73%. At the same time, vanilla grey, vanilla RGB, hybrid CNN, and VGG, as well as the modified capsule network, had accuracy values of 67.8%, 69%, 69.5%, and 63.8%, respectively. Full data and samples, VDSNet outperforms existing methods in terms of several metrics including precision, gain, F0.5 score, and validation accuracy. For the entire dataset case, VDSNet shows a validation accuracy of 73%. At the same time, vanilla Grey, vanilla RGB, hybrid CNN and VGG, and the modified capsule network have accuracy values of 67.8%, 69%, 69.5%, and 63.8%, respectively. Each. When a sample data set is used instead of a complete data set, VDSNet requires substantially less training time at the expense of slightly lower validation accuracy.

The study was (Chamveha et al., 2020)conducted to calculate *the cardiothoracic ratio* (CTR) of *Chest X-Ray* by applying a U-Net-based *deep learning model* with VGG16 to extract lung and heart masks from chest X-rays, which segmented the U-Net model image based on pixels individually. High-speed accuracy across a wide range of segmentation using an end-to-end network of *encoder-decoders containing encoders* that perform feature extraction into output. As well as the area of the heart mask obtained. The results were obtained with a success rate of 76.5% accuracy.

As for the research that will be carried out in this study, a comparison of the classification of normal lungs with lungs suffering from *Cardiomegaly will be carried out*. The preparation of this dataset contributes to improving the quality of the *Cardiomegaly* classification system on X-ray images. The model that will be used in this study is to use the CNN, VGG16, and DenseNet121 models as a classification method to ensure performance and determine which way is the best for classifying lung disease.

MATERIALS AND METHODS

This stage is the stage that describes the method. The following are the methods or stages that will be carried out in this research.

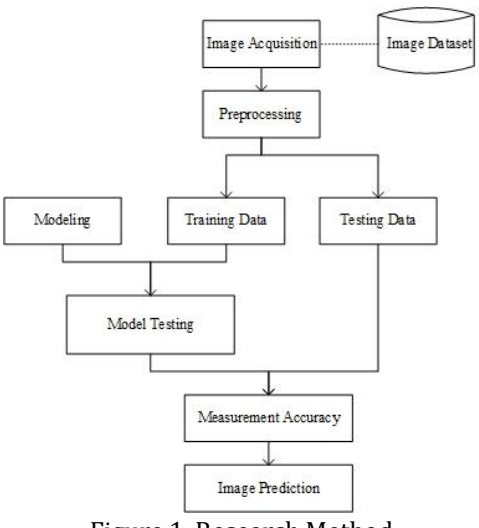


Figure 1. Research Method

Dataset

The dataset in this study was taken from the Kaggle dataset (National Institutes of Health Chest X-Ray Dataset, 2018) by taking 5826 image data by experimenting with 2 classes, namely the

Cardiomegaly class and the No Finding class. The number of images from each category can be seen in Table 1.

Table 1
Number of Images from Each Class

No	Class	Number of Images
1	Cardiomegaly	3050 Image
2	No Finding	2776 Image

As for the example image of each of these classes can be seen in Figure 2.

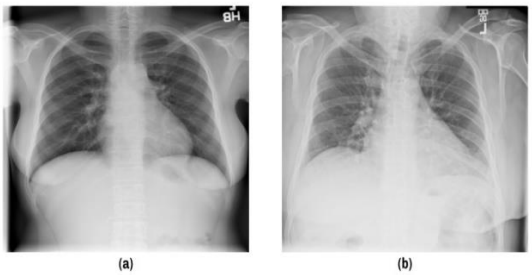


Figure 2. (a) *No finding class image* (b) *Cardiomegaly class image*

Preprocessing

At this stage, *preprocessing is carried out* to correct the data used by suppressing unwanted distortion and some important unwanted features. With a dataset of 5826 For image scaling, this research uses an image with a size of 32x32 Pixels by maintaining three-channel colors. As for the distribution of image data, the dataset will be divided into three parts, namely training data, validation data, and testing data, with details as shown in Table 2.

Table 2
Dataset

20% Testing Data	80%	
	Training Data	Data Validation
586 Image	5192 Image	50 Images

The dataset that has been preprocessed will then be randomized with *random_state* = 101, which can be seen in Figure 3.

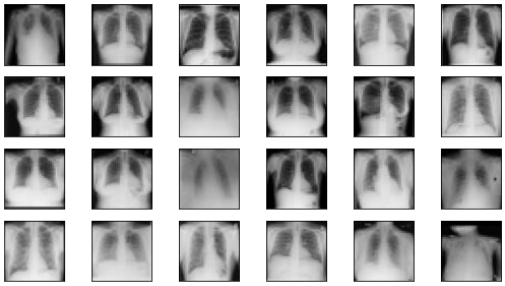


Figure 3. Image of Dataset after *Preprocessing* with *random_state*

Modeling

1. CNN models

In the CNN model, the stages of the model are made, as shown in Figure 4.

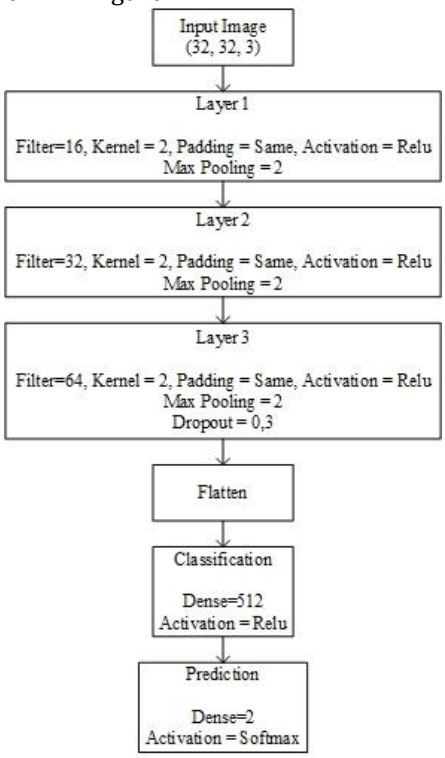


Figure 4. CNN Model

2. VGG16 and DenseNet121 . models

In this model, the stages of the model are made, as shown in Figure 5.

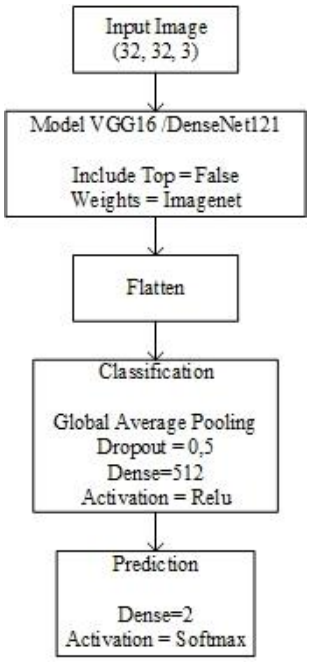


Figure 5. VGG16 and DenseNet121. Models

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3010

Testing Stage

In this section, the data has been processed into the tested model. After being tested with the model, the accuracy value, *loss value, and the Confusion Matrix* truth table are obtained. After testing the DenseNet121 and VGG16 architectures to get accuracy in the classification of lung diseases after getting the results, a comparison is made with the test model. After comparing the accuracy of the three models, image prediction will also be made using the model with the highest accuracy.

RESULTS AND DISCUSSION

Convolutional Neural Network Model

After the data is entered into the CNN model, the accuracy and loss values in the training data and valid data can be seen in Figures 6 and 7.

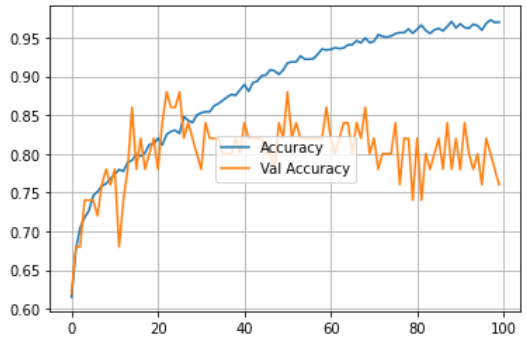


Figure 6. CNN Accuracy Value

The higher the value, the better the experimental results for the accuracy value. From Figure 6 above, it is known that the accuracy value of the training data (blue) tends to increase in each epoch. Still, the accuracy value of the validation data (yellow) increases to epoch 10; the accuracy tends to rise and fall until the last epoch. And for the accuracy value at the end of the epoch, the accuracy value is 96.99% for the training data, and the accuracy value is 76% for the validation data.

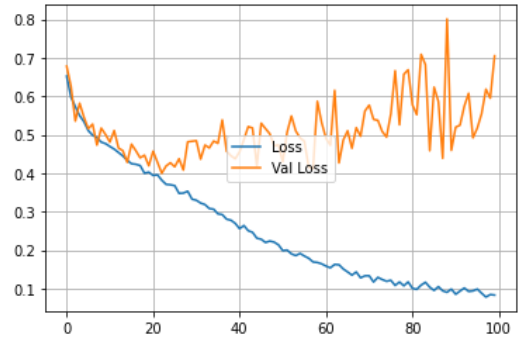


Figure 7. CNN *Loss Value*

For the loss value, the lower the value, the better the experimental results. From Figure 7 above, it is known that the *loss value* in the training data (blue) tends to decrease in each *epoch*, but the lost *weight* in the validation data (yellow) decreases until *epoch* 20. The *loss value* tends to rise and fall until the last *epoch*. And for the *loss value* at the end of the *epoch, the loss* value is 0.0830 for the training data, and the accuracy value is 0.7050 for the validation data.

Then from the CNN model on the *epoch value,* which has the lowest *loss value* that was successfully stored, namely at the 23rd *epoch with a value of 0.3997, the next step is to test the data on the testing data.* After the data testing process, the accuracy value was 80.54%.

Transfer Learning Model

DenseNet 121

After the data is entered into the DenseNet121 model, the accuracy and *loss values* in the training data and valid data are obtained, as shown in Figures 8 and 9.

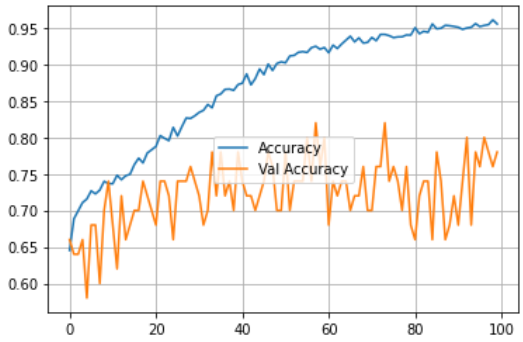


Figure 8. DenseNet121 Accuracy Value

From Figure 8 above, it is known that the accuracy value of the training data (blue color) tends to increase in each *epoch. Still, the* accuracy value of the validation data (yellow color) tends to rise and fall until the last *epoch*. And for the accuracy value at the end of the *epoch,* the accuracy value is 95.54% for the training data, and the accuracy value is 78% for the validation data.

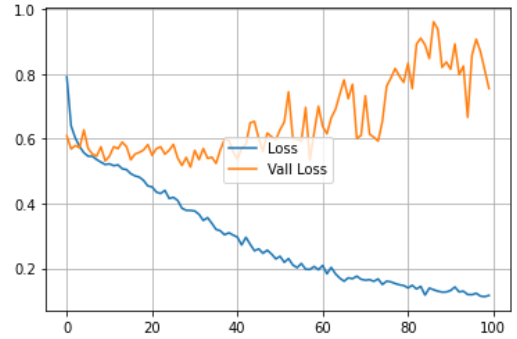


Figure 9. DenseNet121 *Loss Value*

From Figure 9 above, it is known that the *loss value* in the training data (blue) tends to decrease in each *epoch*, but the *loss value* in the validation data (yellow) tends to rise and fall until the last *epoch*. And for the *loss value* at the end of the *epoch, the loss* value is 0.1165 for the training data, and the accuracy value is 0.7555 for the validation data.

Then from the DenseNet121 model on the *epoch value with* the lowest *loss value* that was successfully stored, namely the 97th *epoch with a value of 0.0019, the next step is to test the data-on-data testing.* After the data testing process, the accuracy value was 67.06%.

VGG16

After the data is entered into the VGG16 model, the accuracy and *loss values* in the training data and valid data can be seen in Figures 10 and 11.

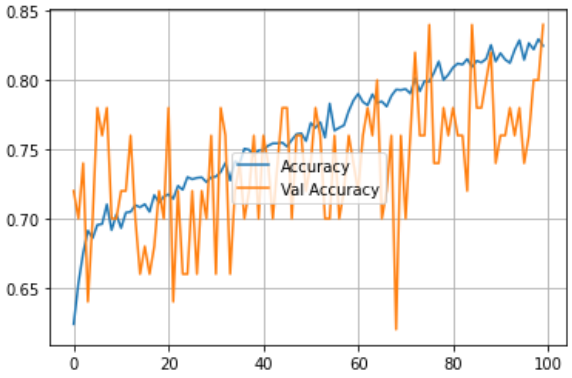


Figure 10. VGG16. Accuracy Value

From Figure 10 above, it is known that the accuracy value of the training data (blue color) tends to increase in each *epoch. Still, the* accuracy value of the validation data (yellow color) tends to rise and fall until the last *epoch*. And for the accuracy value at the end of the *epoch,* the accuracy value is 82.48% for the training data, and the accuracy value is 84% for the validation data.

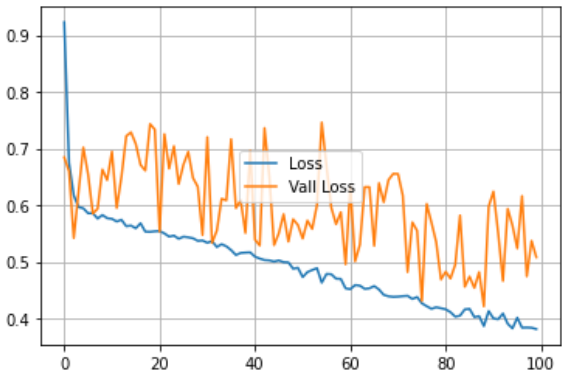


Figure 11. VGG16. *Loss Value*

From Figure 11 above, it is known that the *loss value* in the training data (blue) tends to decrease in each *epoch*, but the *loss value* in the validation data (yellow) tends to rise and fall until the last *epoch*. And for the *loss value* at the end of the *epoch, the loss* value is 0.3815 for the training data, and the accuracy value is 0.5082 for the validation data.

Then from the VGG16 model, the *epoch value with* the lowest *loss value* was successfully saved, namely at the 89th *epoch with a value of 0.4216. The next step is to test the data-on-data testing.* After the data testing process, the accuracy value is 68.94%.

Comparison Value Accuracy

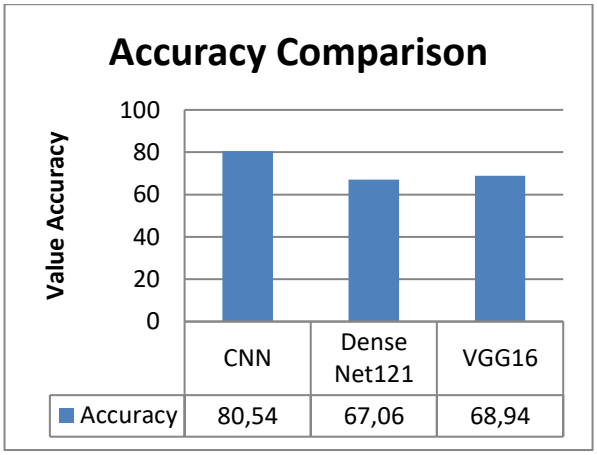


Figure 12. Comparison Graph of Accuracy Values

From Figure 12 above, it is known that the accuracy value of the testing data for the CNN model is 80.54%, the DenseNet121 model is 67.06%, and the VGG16 model is 68.94%.

Image Prediction

The X-Ray image will be predicted using the best model among the three proposed models, namely the CNN model. The results of image prediction can be seen in Figure 13.

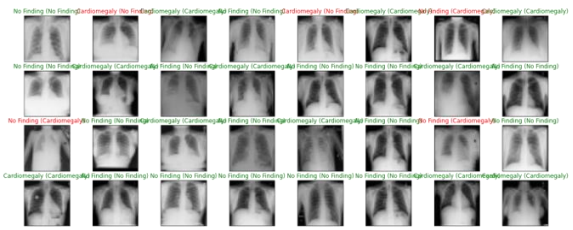


Figure 13. Prediction of X-Ray Image

In Figure 13, it is known that the results of image prediction using the CNN model, from a total of 32 images that were predicted, 27 images were predicted correctly, and five images that were expected to be wrong.

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3010

CONCLUSION

This study proposes Chest X-ray to classify lung diseases by conducting several research stages starting from image input, preprocessing stage, conducting data modeling using Conventional CNN, VGG16 and DenseNet121, then entering the data training stage, data testing, model testing, measurement accuracy and the last is the stage of predicting the image. so as to get the accuracy of each - each model test. It can be concluded that by using the DenseNet121 model, X-Ray images in this research dataset get an accuracy of 67.06%, for the VGG16 model it gets an accuracy of 68.94% and for the CNN model it gets the highest accuracy of 80.54%.

REFERENCE

Agustyawati, D. N., Fauzi, H., & Pratondo, A. (2021). PERANCANGAN APLIKASI DETEKSI KANKER SERVIKS MENGGUNAKAN METODE CONVOLUTIONAL NEURAL NETWORK APPLICATION DESIGN OF SERVIKS CANCER DETECTOR BASED USING CONVOLUTIONAL NETWORK. *E-Proceeding of Engineering*, *08*(04).

Setiaji, R. A., Hidayat, B., & Suhardjo. (2018). SINTESIS PENELITIAN DETEKSI PENYAKIT ABSES PADA GIGI MANUSIA MELALUI CITRA PERIAPIKAL RADIOGRAF DOMAIN SPASIAL. *EProceedings of Engineering*, *05*(03).

Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2018). Comparison of deep learning approaches for multi-label chest X-Ray classification. *ArXiv*, 1–10.

Bharati, S., Podder, P., & Mondal, M. R. H. (2020). Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked*, *20*. https://doi.org/10.1016/j.imu.2020.100391

Chamveha, I., Promwiset, T., Tongdee, T., Saiviroonporn, P., & Chaisangmongkon, W. (2020). Automated cardiothoracic ratio calculation and cardiomegaly detection using deep learning approach. *ArXiv*, 1–11. https://arxiv.org/abs/2002.07468

Cholik, C. A. (2021). *Perkembangan Teknologi Informasi Komunikasi ICT dalam Berbagai Bidang*. *2*(2), 2746–1209.

Gao, M., Bagci, U., Lu, L., Wu, A., Buty, M., Shin, H. C., Roth, H., Papadakis, G. Z., Depeursinge, A., Summers, R. M., Xu, Z., & Mollura, D. J. (2018). Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, *6*(1), 1–6. https://doi.org/10.1080/21681163.2015.1124249

Hadianti, S., & Riana, D. (2018). Segmentasi Citra Bemisia Tabaci Menggunakan Metode K-Means. *Seminar Nasional Inovasi Dan Tren (SNIT)*.

Hilaliyah, P. K. (2021). *Deteksi Dini Kanker Payudara Pada Citra Histopatologi Menggunakan Metode Convolution Neural Network (CNN)*.

Idatin Nikmah, Z., Kurnia Aditya, S., & Gustri Wahyuni, E. (2019). Aplikasi Web Untuk Pendeteksi Penyakit Paru-Paru Menggunakan Metode Certainty Factor. *Seminar Nasional Informatika Medis*.

Indraswari, R., Herulambang, W., & Rokhana, R. (2022). Deteksi Penyakit Mata Pada Citra Fundus Menggunakan Convolutional Neural Network (CNN) Ocular Disease Detection on Fundus Images Using Convolutional Neural Network (CNN). *TechnoCom*, *21*(2), 378–389. https://www.kaggle.com/datasets/jr2ngb/cataractdataset

Jaisakthi, S. M., Mirunalini, P., Thenmozhi, D., & Vatsala. (2019). Grape leaf disease identification using machine learning techniques. *ICCIDS 2019 - 2nd International Conference on Computational Intelligence in Data Science, Proceedings*. https://doi.org/10.1109/ICCIDS.2019.8862084

National Institutes of Health Chest X-Ray Dataset. (2018). *NIH Chest X-rays*. Kaggle. https://www.kaggle.com/nih-chest-xrays/data

Phiadelvira, B. Y. (2021). *Klasifikasi Kanker Serviks berdasarkan Citra Kolposkopi menggunakan Convolutional Neural Network (CNN) Model Alexnet*.

Triyono, L., Nur, A., Thohari, A., Hestiningsih, I., & Yobioktabera, A. (2021). KLASIFIKASI PENYAKIT KULIT MENGGUNAKAN METODE CONVOLUTIONAL NEURAL NETWORK. *Sentrikom*, *04*(01), 37–44.

FINAL GRADE PREDICTION MODEL BASED ON STUDENT'S ALCOHOL CONSUMPTION

M. Rangga Ramadhan Saelan [1*]; Siti Masturoh [2]; Taopik Hidayat [3]; Siti Nurlela [4]; Risca Lusiana Pratiwi [5]; Muhammad Iqbal [6]

[1,3]Sains Data, [2,4,5]Sistem Informasi, [6]Teknik Informatika
[1,2,3,4,5]Universitas Nusa Mandiri, [6]Universitas Bina Sarana Informatika
[1,2,3,4,5]www.nusamandiri.ac.id, [6]www.bsi.ac.id
rangga.mgg@nusamandiri.ac.id[1*], siti.uro@nusamandiri.ac.id[2], taopik.toi@nusamandiri.ac.id[3], siti.sie@nusamandiri.ac.id[4], risca.ral@nusamandiri.ac.id[5], iqbal.mdq@bsi.ac.id[6]

Abstract— *The influence of alcohol consumption and several other factors that are estimated to have a role on the level of learning performance of adolescents who are still in school, in this work, research is carried out on public data that has been obtained. This work trains a model on a dataset that provides information about student grades from the first year to the final year. In this work, we will focus on the final year mark in Portuguese as a label, with several variables as factors that influence the final grade as a reference for learning performance. on a group of students who are the target of research. Using machine learning techniques by training several models to predict the final grade as a reference for student learning performance through the final grade results obtained. By training several machine learning models to predict final year grades (G3) from Portuguese lessons by doing a comparative method comparing the Support Vector Regressor (SVR) and Random Forest (RF) models. All models have hyperparameters that must be adjusted using Cross Validation. The best models for predicting G3 are SVR and RF, and have a mean absolute error (MAE) of about 2.24 and 2.25, respectively. Through the MAE plot, the SVR and RF models work well. However, By analyzing the distribution of errors made by both models. Through this work, it can be concluded that SVR is more balanced, i.e. has a better ratio between underestimation and overestimation, while RF performs better on outliers.*

Keywords*: Model, MAE, Parameter, Machine, Learning, Value*

Intisari— Pengaruh konsumsi alcohol dan dan beberapa faktor lainnya yang diperkirakan memiliki peran terhadap tingkat kinerja belajar remaja yang masih bersekolah, pada pekerjaan ini dilakukan penelitian terhadap data publik yang telah didapatkan. Pekerjaan ini melatih model terhadap dataset yang menyediakan informasi mengenai nilai-nilai pelajar dari tahun pertama hingga tahun akhir, Pada pekerjaan ini akan berfokus kepada nilai tahun akhir bahasa portugis sebagai label, dengan beberapa variable sebagai faktor yang berpengaruh terhadap nilai akhir yang menjadi acuan kinerja belajar pada sekumpulan pelajar yang menjadi sasaran penelitian. Menggunakan teknik *machine learning* dengan melatih beberapa model untuk memprediksi nilai akhir sebagai acuan kinerja belajar pelajar melalui hasil nilai akhir yang didapatkan. Dengan melatih beberapa model *machine learning* untuk memprediksi nilai tahun akhir (G3) dari pelajaran bahasa portugal dengan melakukan metode komparatif membandingkan model *Support Vector Regressor* (SVR) dan *Random Forest* (RF). Semua model memiliki hyperparameter yang harus disesuaikan menggunakan *Cross Validation*. Model terbaik untuk memprediksi G3 adalah SVR dan RF, dan memiliki *mean absolute error* (MAE) masing-masing sekitar 2,24 dan 2,25. Melalui plot MAE, model SVR dan RF bekerja dengan baik. Tetapi, Dengan menganalisis distribusi kesalahan yang dibuat oleh kedua model. Melalui pekerjaan ini, dapat disimpulkan bahwa SVR lebih seimbang, yaitu memiliki rasio yang lebih baik antara nilai yang diremehkan dan ditaksir terlalu tinggi, sementara RF berkinerja lebih baik pada *outlier*.

Kata Kunci*:* Model, MAE, Parameter, *Machine, Learning*, Nilai

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3056

INTRODUCTION

The seriousness, totality and focus of adolescents in the learning period at school are the main problems at this time to organize the younger generation for the future. On the other hand, alcohol consumption needs to get great attention from the relevant authorities, from the government, social institutions, and families. This is because alcohol consumption is currently not only a target for consumption among adults, but also a target for teenagers who are still relatively young where they should focus more on their learning period. Thus, with the increasing percentage of consumption of alcohol by adolescents who are studying this will become a disorder or obstacle in the process of receiving knowledge gained from educational institutions. In 2016, the National Institutes of Health reported that 26% of 8th graders, 47% of 10th graders, and 64% of 12th graders had experience consuming alcoholic beverages (Palaniappan et al., 2017).

Data mining (DM) has been applied in the field of education, and is an emerging interdisciplinary research field also known as Educational Data Mining (EDM). One of the goals of the EDM is to better understand how to predict the academic performance of students given personal, socio-economic, psychological and other environmental attributes. Another goal is to identify factors and rules that affect the academic outcomes of education (Satyanarayana & Nuckowski, 2016).

To determine the effect of alcohol consumption and several other factors that are estimated to have a role in the level of learning performance of adolescents who are still in school, a study is currently being carried out on public data that has been obtained using machine learning techniques by training several models to predict the final value as a reference for student learning performance. Alcohol consumption is reported by 85% of students. About 70% of first- and third-year students and 47% of sixth-year students are motivated by socializing with peers. Alcohol consumption is widespread in those who engage in physical activity (93%) and live with family (89%) (Freire et al., 2021). Of course, there is a correlation between the level of alcohol consumption and environmental factors and so on, therefore it is necessary to conduct research to find out the relationship between one factor and another, as well as the influence of certain factors that play a very large role in the level of alcohol consumption by students so that it has an impact on learning performance in the form of final grade results (G3). Alcohol consumption can be influenced by several factors such as heredity/habits, environment, mental health. Alcohol consumption in learners has a long-term effect on students' brain and learning performance (Nur'artavia, 2017). Alcoholic beverages themselves Ethanol is a psychoactive ingredient that can reduce the consciousness of its consumers (Wijaya, 2016). In particular, students who have consumed excessive alcohol tend to have difficulties related to brain memory and the ability to focus. Alcohol addiction can be influenced by several factors such as genetics, social environment, and mental health. Alcohol consumption at a young age (learners) has a long-term effect on students' brain and academic performance (Sagala & Tampubolon, 2018).

The purpose of the following work is to train some machine learning models so that later it can be used to predict the final Portuguese grades of students attending two schools in Portugal. In the Portuguese undergraduate system, the value is between 0 and 20. By analyzing exploratory data (EDA) from the available data by focusing on the label, namely the end-of-year value (G3). Then in this work will train several models so that the best model will be known to later be able to predict (G3). Then the last goal is to analyze the distribution of errors committed by the two models, and the latter gives us a discriminant to choose one of the two models.

Some machine learning models will be trained on the training dataset, this is done to get the best model. This related research has been carried out by several research literacy, including research conducted by sagala et al (Sagala & Tampubolon, 2018) "This study aims to apply and perform performance analysis of data mining algorithms to predict alcohol consumption and analyze related factors in intermediate level students. The stages carried out are data pre-processing, feature selection, classification, and model evaluation. At the preprocess stage, some features are transformed into appropriate shapes to facilitate the classification process. The results showed that the classification model built using Naïve Bayes had the highest accuracy value using the 5 best features of the Gain Ratio. In addition, the use of the feature selection method is able to improve the performance of all classifiers in general."

Further research conducted by Pisutaporn et al. (Pisutaporn et al., 2018) "In this paper, we have studied student alcohol consumption and identified factors that have a significant impact on students' alcohol consumption. It was found that men tend to have more alcohol consumption rates than women. The high rate of going out with friends leads to high alcohol consumption. In addition, students who have a short weekly study

time, no school support for additional education and do not have the desire to go on to higher education are more likely to have more alcohol consumption. We also found that the random forest algorithm performed better than the decision tree algorithm for this classification problem. Finally, we have confirmed the negative relationship between the level of alcohol consumption and the value of students."

The purpose of this study is to train several machine learning models to predict the final year value of the Portuguese language by conducting a comparative method comparing the Support Vector Regressor (SVR) and Random Forest (RF) models so that the best model will be obtained to predict.

Literature Review

Machine Learning

ML is defined as a scientific field that gives machines the ability to learn without being strictly programmed (Liakos et al., 2018).

Machine learning (ML) is used to teach machines how to handle data more efficiently. Many industries are applying machine learning to extract relevant data. The purpose of machine learning is to learn from data. A lot of research has been done on how to make machines learn on their own without being explicitly programmed. Many mathematicians and programmers apply several approaches to find a solution to this problem that has a very large data set (Batta, 2020).

Support Vector Regression

The purpose of the SVR is to find a linear regression equation suitable for all sample points and minimize the total sample point variance of this regression hyperplane. here is an example of a training set

$E = \{(x_i, y_i) | i = 1,2,..n\}, x_i \in R^n, y_i \in R$. Function $f(x_i)$ investigated on R^n, in such a way, that $y_i = f(x_i)$, and there is always value y which is appropriate for each input x. Such a function $f(x_i)$ called the regression function, and $f(x_i)$ can be described as follows.

$$f(x_i) = \omega . \emptyset (x_i) + b, \quad (1)$$

where $\omega \in R^n$ is *Weight Vector*, $\emptyset(x_i)$ is a nonlinear mapping that serves to map data from space R^n to the higher feature space, and b is biased. Equation (1) can be attached to all sample points with precision.

$$|y_i - [\omega . \emptyset(x_i) + b]| \leq \varepsilon, \; i = 1,2, \ldots n \quad (2)$$

Because there is a certain installation error, the slack variable $(\xi_i, \xi_i *)$ and penalty parameter C is introduced. Regression adjustment issues are turned into optimization problems (Tang et al., 2022).

Random Forest

Random Forest (RF) is an algorithm that uses recursive binary separation methods to reach the final node in a tree structure based on classification and regression trees (Yoga Religia et al., 2021)

Derived from the theory of learning ensembles, RF combines several individual Decision Trees (DTs). Due to simplification and nonparametric behavior, classification and regression trees (CART) are commonly used as DT in RF Each DT relies on a random bootstrap dataset (Liu et al., 2021).

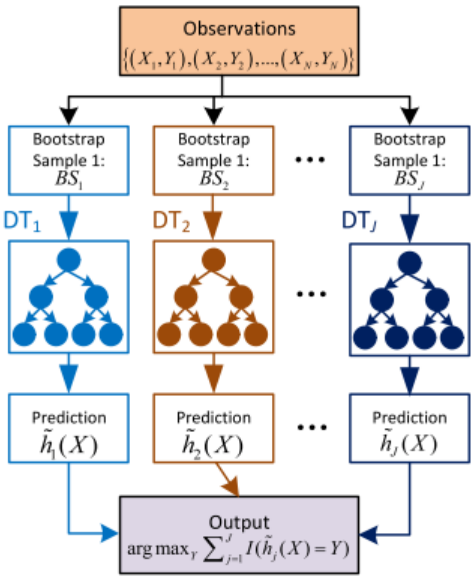


Figure 1. Classification Structure of Random Forest (Liu et al., 2021)

Breiman in 2001 introduced the RF algorithm by showing several advantages including being able to produce relatively low errors, good performance in classification, being able to overcome large amounts of training data efficiently, and effective methods for estimating missing data.

The main purpose of the RF training stage is to build many uncorrelated DT. To reduce variance associated with classification, an overlapping sampling solution named 'bagging' was adopted in RF (Liu et al., 2021).

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3056

METHOD

1. Data Collection

Data source is retrieved through a public site *https://www.kaggle.com/datasets/uciml/student-alcohol-consumption*, This dataset is from portugal country by P. Cortez and A. Silva. Data publik ini memiliki 33 atribut dengan jumlah record sebanyak 649 responden. Data were obtained in a survey of students of mathematics and Portuguese language courses in high school. It contains a lot of interesting social, gender and study information about students.

2. Research Methods

The cross-industry standards for data mining (CRISP DM) process is a framework for translating business issues into data mining tasks and implementing data mining projects independent of the application area and technology used (Huber et al., 2019).

CRISP-DM with a few general steps and more, tasks for each step: *business understanding, data understanding, data preparation, data modelling, results evaluation and deployment* (Cazacu & Titan, 2020).

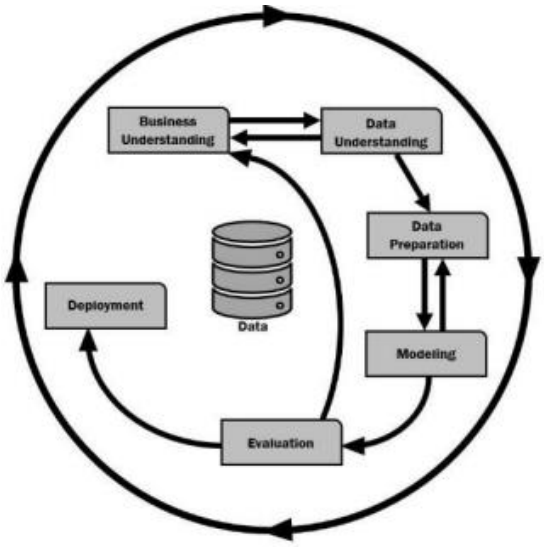


Figure 2. CRISP-DM (Cazacu et al., 2020)

1) Business Understanding

This stage explains the purpose of the research carried out. Exploratory data analysis (EDA) of the considered data set is provided. In particular, EDA is mainly focused on labels, but other insights from the data are also taken into account. The purpose of this study is to train several machine learning models to see the best model in predicting students' Portuguese scores in the final year (G3) of The Portuguese language, by training the Support Vector Regressor (SVR) and Random Forest (RF) model models so that later the best model will be obtained to predict the final value based on alcohol consumption by students.

2) Data Understanding

The following is a prospectus of all features into the dataset, explaining in detail each feature so that it can be understood and make it easier for research to be carried out.

Table 1. Feature description of the dataset

Atribute	***Record***
School	:(binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
Sex	:(binary: 'F' - female or 'M' - male)
Age	:(numeric: from 15 to 22)
Address	:address type (binary: 'U' - urban or 'R' - rural)
Famsize	:(binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	:(binary: 'T' - living together or 'A' - apart)
Medu	:mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	:father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	:mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	:father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Reason	:reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
Guardian	:(nominal: 'mother', 'father' or 'other')
Traveltime	:(numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
Studytime	:(numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
Failures	:(numeric: n if 1<=n<3, else 4)
Schoolsup	:support (binary: yes or no)
Famsup	:family support(binary: yes or no)

Paid	:(Math or Portuguese) (binary: yes or no)
Activities	:extra-curricular(binary: yes or no)
Nursery	:attended nursery school (binary: yes or no)
Higher	:wants to take higher education (binary: yes or no)
Internet	:at home (binary: yes or no)
Romantic	:(binary: yes or no)
Famrel	:family relationships (numeric: from 1 - very bad to 5 - excellent)
Freetime	:(numeric: from 1 - very low to 5 - very high)
Gout	:(numeric: from 1 - very low to 5 - very high)
Dalc	:(numeric: from 1 - very low to 5 - very high)
Walc	:(numeric: from 1 - very low to 5 - very high)
Health	:(numeric: from 1 - very bad to 5 - very good)
Absences	:(numeric: from 0 to 93)
G1	:first period grade (numeric: from 0 to 20)
G2	:second period grade (numeric: from 0 to 20)
G3	:final grade (numeric: from 0 to 20, output/ target

Source: (Saelan et al., 2020)

3) *Data Preparation*

Making preparations to make the data ready to be propses at the modeling stage is important, including: trying to eliminate missing values, changing the names of some column values so that they can make plot visualization cleaner, making plots for categorical and numerical data. Before training some machine learning models to predict the final value (G3), it begins by defining the dataset, namely the x feature and the y label. In the current study, the final value (G3) becomes label data (y) so that the work will refer to a trained model that is later considered good for predicting label values. After several plots of each feature, it was discovered that the average score of portuguese (G3) was influenced by the alcohol consumption of 'Walc' and 'Dalc' learners, the higher level of alcohol consumption was associated with lower learning performance.

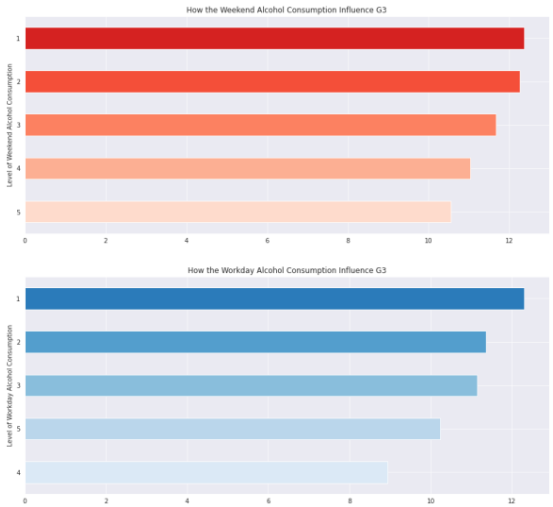


Figure 2. Walc and Dalc plot against G3 (Source: research, 2022)

In addition to the correlation of alcohol consumption with the final value, there are also some of the most influential features and it is considered important to predict the final value (G3).

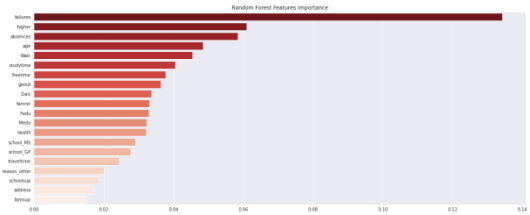


Figure 3. Feature Importance (Source: Research, 2022)

According to the EDA, it can be seen that there are ten most important features for the prediction of the value of G3 are: *'failures', 'higher', 'Absences', 'Age', 'Walc', 'StudyTime', 'freetime', 'goout', 'Dalc', 'famrel'*.

4) *Modeling*

Adjustment of the hyperparameter by using Cross Validation (CV) to obtain the optimal model at the time of training, so that when getting analysis with optimal adjustments can start calculating the predicted value and MAE by training the model that has been formed. The models to be trained to predict label values are the Random Forest and Support Vector Regressor.

5) *Evaluation*

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3056

Based on the evaluation of the test, the predicted value and MAE of the trained model are obtained. The metrics used to evaluate the trained model are *Mean Absolute Error* (MAE).

6) *Deployment*

If needed, from the model that has been generated, it is tested using new data and re-evaluated for data accuracy. Using generated and represented models or knowledge representation processes.

DISCUSSION

The dataset taken is a collection of data that provides information about student values from the first-year end but in this work will focus on the final year values of the Portuguese language as label data, with several variables that can function as factors that affect the final score which is a reference for learning performance in a group of students who are the target of research.

The dataset was taken from a public site related to alcohol consumption by a group of students. Data is divided into 2 types: categorical data and Numerical data. Before processing by training the model against the data, the data is neated first, if there is empty data, or duplicates so that it will be ready to be processed.

Train multiple models to predict the final class (G3) value. Start by defining the dataset, i.e. features (x) and labels (y). For that it is necessary to preprocess the data. Specifically, convert the categorical data into dummy variables, by using a one-hot encoder. Then divide the dataset into 80% training data and 20% test data from all tests. All models have hyperparameters that must be adjusted. To tune this hyperparameter using cross validation.

Table 2. Description of the model hyperparameter set

MODEL	HYPERPARAMETER	MEANING
SVR	Kernel	the type of kernel to be used in the algorithm
	C	regularization parameters. The power of regularization is inversely proportional to C. It must be absolutely positive
	Epsilon	specifies an epsilon tube in which there is no penalty associated in the training loss function with the predicted points within the distance of epsilon from the actual value.
	Gamma	kernel coefficients for 'rbf', 'poly' and 'sigmoid
RF	n_estimators	number of decision trees used
	max_depth	maximum depth of one decision tree
	max_features	number of random features to consider on each split

Source: (Yang & Shami, 2020)

To form an optimal model it is necessary to make adjustments to the hyperparameter. The best parameters obtained are as follows:

Table 3. Set Hyperparameters

Model	Hyperparameters	Value
SVR	Kernel	Rbf
	Gamma	Scale
	Epsilon	0.0001
	C	2
RF	n_estimator	87
	max_feature	0.2
	max_depth	10

After obtaining the best set of hyperparameters, it can be tested the model and it can be seen how the model performs its best.

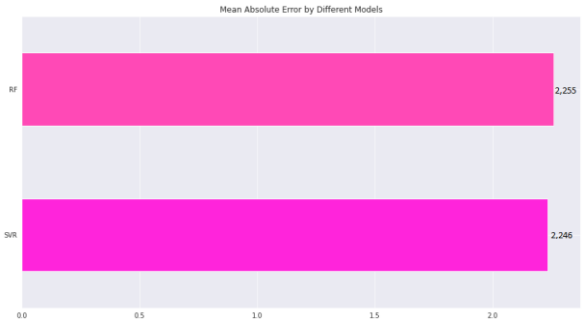


Figure 3. MAE SVR and RF

Source: (research, 2022)

To better understand how this model works, a plot is created between test and prediction data. So that the spread statistics can be known

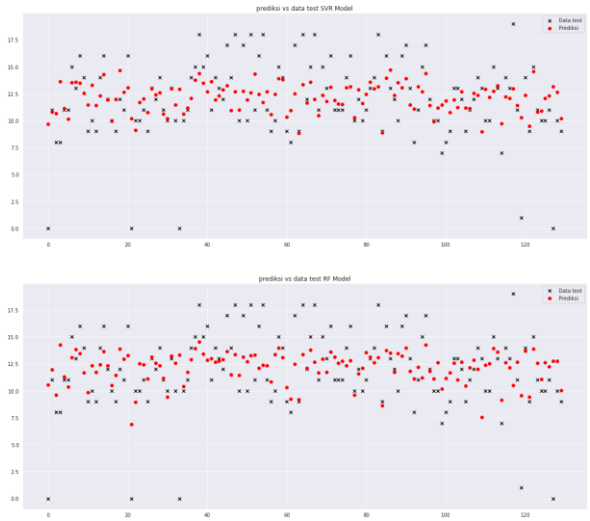


Figure 4. Plot Predictions against Test Data
Source: (research, 2022)

Based on the plot, both models tend to give predictions of G3 close to the average value of the latter. thus, both models perform well enough to predict values close to the average value, but perform poorly to predict values that are far from the latter, that is, the best student scores and the worst student scores. In addition, models tend to overestimate values below the average value and underestimate values above the average value.

Plots containing MAE associated with different models show that the model is RF and SVR works fine. Although the second performs better than the first, these two MAEs are very similar. To determine the best performance between the two models, errors distribution checks are carried out by these models.

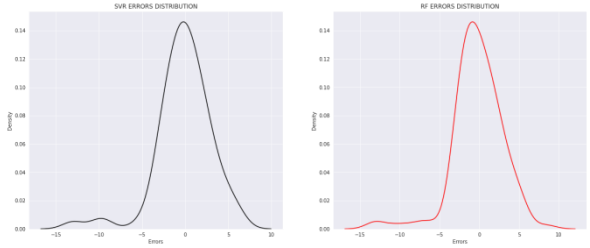


Figure 5. Errors Distribution
Source: (research, 2022)

Errors Distibution SVR is more symmetrical around the main peak and the latter is closer to 0, while the RF fault distribution has a less wide secondary peak. In other words, SVR tends to be more balanced in the ratio of below/exaggerated value with respect to RF, whereas RF performs better than SVR in outliers. Therefore, the choice between SVR and RF depends on what is the main focus: having a more balanced model or a model that performs better on the outlier.

CONCLUSION

At the end of this study, several conclusions were obtained stating that in predicting the final value (G3) of the Portuguese language SVR will be better than RF, looking at the two models trained by SVR and RF, the best model for predicting the final value of G3 is the Support Vector Regressor (SVR) by having an absolute mean error (MAE) of about 2.24 each, while Random Forest has a MAE of 2.25; By analyzing the distribution of errors made by both models, it can be concluded that the SVR is more balanced, that is, it has a better ratio between underestimated and overestimated values; then Based on the provided EDA, the most important feature for G3 predictions is *'failures', 'higher', 'Absences', 'Age', 'Walc', 'StudyTime', 'freetime', 'goout', 'Dalc', 'famrel'*. In reality some variables that were not previously considered the most important clearly affect the final value as shown in the EDA. However, based on several results of the plot analysis of each feature on the final value, Dalc and Walc have the highest influence on the final value (G3).

REFERENCES

Batta, M. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJ*, *9*(1), 381–386. https://doi.org/10.21275/ART20203995

Cazacu, M., & Titan, E. (2020). Adapting CRISP-DM for Social Sciences. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *11*(2sup1), 99–106. https://doi.org/10.18662/brain/11.2sup1/97

Freire, B. R., de Castro, P. A. S. V., & Petroianu, A. (2021). Alcohol consumption by medical students. *Revista Da Associacao Medica Brasileira*, *66*(7), 943–947. https://doi.org/10.1590/1806-9282.66.7.943

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors (Switzerland)*, *18*(8), 1–29. https://doi.org/10.3390/s18082674

Liu, K., Hu, X., Zhou, H., Tong, L., Widanage, W. D., & Marco, J. (2021). Feature Analyses and Modeling of Lithium-Ion Battery

Manufacturing Based on Random Forest Classification. *IEEE/ASME Transactions on Mechatronics*, *26*(6), 2944–2955. https://doi.org/10.1109/TMECH.2020.3049046

Nur'artavia, M. R. (2017). Karakteristik Pelajar Penyalahguna Napza Dan Jenis Napza Yang Digunakan Di Kota Surabaya. *The Indonesian Journal of Public Health*, *12*(1), 27. https://doi.org/10.20473/ijph.v12i1.2017.27-38

Palaniappan, S., A Hameed, N., Mustapha, A., & Samsudin, N. A. (2017). Classification of Alcohol Consumption among Secondary School Students. *JOIV : International Journal on Informatics Visualization*, *1*(4–2), 224. https://doi.org/10.30630/joiv.1.4-2.64

Pisutaporn, A., Chonvirachkul, B., & Sutivong, D. (2018). Relevant factors and classification of student alcohol consumption. *2018 IEEE International Conference on Innovative Research and Development, ICIRD 2018*, *May*, 1–6. https://doi.org/10.1109/ICIRD.2018.8376297

Saelan, M. R. R., Sahputra, D. A., Widiastuti, W., & Gata, W. (2020). Komparasi Algoritma Klasifikasi untuk Prediksi Minat Sekolah Tinggi Pelajar pada Students Alcohol Consumption. *Jurnal Sains Dan Informatika*, *6*(2), 120–129. https://doi.org/10.34128/jsi.v6i2.236

Sagala, N., & Tampubolon, H. (2018). Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi Alkohol Siswa. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *4*(2), 98. https://doi.org/10.23917/khif.v4i2.7061

Satyanarayana, A., & Nuckowski, M. (2016). *Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance Ashwin Satyanarayana, Mariusz Nuckowski*. https://academicworks.cuny.edu/ny_pubs

Tang, F., Wu, Y., & Zhou, Y. (2022). Hybridizing Grid Search and Support Vector Regression to Predict the Compressive Strength of Fly Ash Concrete. *Advances in Civil Engineering*, *2022*. https://doi.org/10.1155/2022/3601914

Wijaya, P. A. (2016). Faktor-Faktor Yang Mempengaruhi Tingginya Konsumsi Alkohol Pada Remaja Putra Di Desa Keramas Kecamatan Blahbatuh Kabupaten Gianyar. *Jurnal Dunia Kesehatan*, *5*(2), 15–23. https://media.neliti.com/media/publications/76931-ID-faktor-faktor-yang-mempengaruhi-tingginy.pdf

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yoga Religia, Agung Nugroho, & Wahyu Hadikristanto. (2021). Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 187–192. https://doi.org/10.29207/resti.v5i1.2813