

C5.0 Algorithm Implementation On Web-Based Software and Usability Evaluation

¹Ricky Wijaya, ²Deny Jollyta*

^{1,2}Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Institut Bisnis dan Teknologi Pelita Indonesia

Jl. Jend. Ahmad Yani No.78-88 Pekanbaru, Riau, Indonesia

*e-mail: deny.jollyta@lecturer.pelitaindonesia.ac.id

(received: 28 Januari 2021, revised: 12 April 2021, accepted: 24 April 2021)

Abstrak

Perangkat lunak merupakan alat bantu yang memudahkan pengguna dalam pengolahan data dengan cepat dan tepat. Para pengambil keputusan membutuhkan alternatif perangkat lunak yang dapat digunakan setiap saat dengan teknik klasifikasi data algoritma C5.0 sesuai kriteria yang diinginkan. Namun perangkat lunak yang ada umumnya terdiri dari sejumlah teknik dan belum dapat digunakan secara online. Sebagai salah satu algoritma klasifikasi yang populer dalam ilmu data *mining*, C5.0 dapat memberikan hasil yang lebih baik. Penelitian bertujuan untuk membangun perangkat lunak yang dapat melakukan klasifikasi data menggunakan algoritma C5.0 berbasis *web*. Perangkat lunak dapat digunakan oleh siapa saja, terutama para pengambil keputusan. Penelitian ini juga dilengkapi dengan pengujian perangkat lunak *usability* sebelum digunakan. Hasil pengujian memperlihatkan bahwa perangkat lunak yang dibangun dapat diterima dengan nilai *usability* 76,892% dan berada pada predikat Baik. Diharapkan melalui penelitian ini, dapat memberikan alternatif perangkat lunak yang mampu menyelesaikan masalah klasifikasi menggunakan algoritma C5.0.

Kata kunci: perangkat lunak, klasifikasi, algoritma c5.0, *usability*.

Abstract

Software is a tool that makes it easy for users to process data quickly and precisely. Decision makers need an alternative software that can be used at any time with the C5.0 algorithm data classification technique according to the desired criteria. However, the existing software generally consists of a number of techniques and cannot be used online. As one of the popular classification algorithms in data mining science, C5.0 can provide better results. This study aims to build software that can classify data using the web-based C5.0 algorithm. Software can be used by anyone, especially decision makers. This research is also complemented by testing Usability software before used. The test results showed that the software built can be accepted with a Usability value of 76.892% and is in the Good predicate. It is hoped that through this research, it can provide alternative software that is able to solve classification problems using the C5.0 algorithm.

Keywords: *software, classification, c5.0 algorithm, usability*

1 Introduction

Data mining is a way to describe the knowledge contained in large-volume databases [1]. The various parsing techniques include classification. The prediction process for an object class whose class label is unknown can be seen from the discovery of a classification model that is able to explain and differentiate data classes [2].

The C5.0 algorithm is the classification that has helped many users in solving problems such as bank customer credit [3] and system recommendations [4]. In several studies, the C5.0 algorithm was run with several application tools, such as R language [5], RapidMiner [6], [7] and SPSS [8]. These applications support the processing of classified data to produce a decision tree. Generally, the

software used includes a number of techniques. Some companies or users find it confusing and cannot be used at any time to communicate.

Based on some of these studies, very few run the C5.0 algorithm in self-designed software to produce knowledge in the form of decision trees and rules. This study aims to build a software that can run the web-based C5.0 algorithm. This is based on the idea that there are more and more problems that require a classification solution, such as the Covid-19 pandemic. The classification can help the government produce information about the status of Covid-19 sufferers based on symptoms. Apart from that, the addition of web and usability testing software, provides an opportunity that this software can be used by various users such as organizations or governments. Usability testing is important to do to get software performance measures [9] and many software have been tested using this technique [10], [11]. However, classification does not only use the C5.0 algorithm, depending on the form of the problem being solved [12], [13].

2 Literature Review

The C5.0 algorithm has been used to classify many cases in various fields and various applications. Research [5] classified 15 factors that influence on-time graduation for students, including gender, regional origin, entry status, number of credits and GPA, parent's occupation and so on. The criteria were tested on the student data of a college using the R programming. The R results show that the 6th semester GPA, 6th semester SKS, 4th semester GPA, gender, 2nd semester GPA, high school type, regional origin, and 4th semester GPA are the selected factors that influence on time graduation from a student.

In research [6], testing of product certification data for SNI mark users on bottled drinking water uses the Rapid Miner application. This application generates 7 rules that form the basis of classification. In addition, Rapid Miner was also successful in solving the cleaning service selection problem at PT. ISS Indonesia Medan. The criteria used are education, height, weight, and experience. RapidMiner provides the greatest Gain value on the experience criterion.

Based on the explanation above, it shows that the application used in solving the C5.0 Algorithm case cannot be done online. The diversity of users and information needs in real time raises different interests, thus enabling the development of software specifically designed for a particular interest.

But as a consequence, the software created must be tested. The purpose of being web based is to make it easier for interested users in an organization to classify data. Software can be tested with Usability. In research [10], usability successfully tested a smart academic system based on user experience and a web-based shortest route search system [11].

This study is to develop software to classify data using a web-based C5.0 Algorithm with good and interactive standards. It is still possible to do this because there are still companies or users who want it so that it can be an alternative to the classification process.

3 Research Method

This research was conducted following a framework designed to facilitate the achievement of objectives. It is shown in Figure 1.

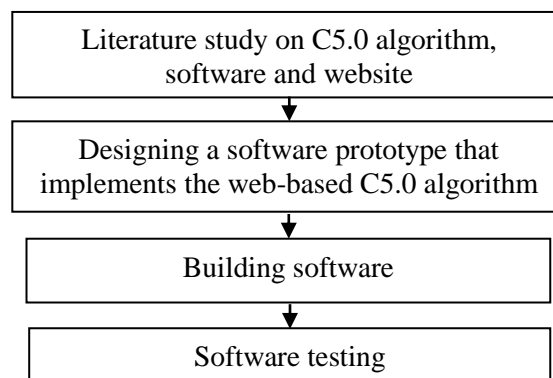


Figure 1. Research Stages

The step is started from studying about the C5.0 algorithm and its implementation. C5.0 algorithm is developed by Ross Quinlan in 1987 [1]. The principle of this algorithm is to produce a decision tree based on highest information gain with the following equation [14]:

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i) \tag{1}$$

Where:

Info (D) or Entropy is the information needed to classify class labels, whereas pi is a non-zero probability with a random tuple in D. To generate Entropy based on attribute A, equation (2) is used.

$$Info_A(D) = \sum_{j=1}^y \frac{|D_j|}{D} x Info(D_j) \tag{2}$$

To get the information gain from partition A, equation (3) is used.

$$Gain(A) = Info(D) - Info(D_j) \tag{3}$$

Where:

Gain (A) states the number of branches that will be obtained on A.

The next step is designing and building a software based on web. The software design used Hypertext Preprocessor (PHP) and enter equations (1), (2) and (3) to obtain the desired classification. For software testing, the Covid-19 dummy data was used. The data was obtained from the explanation of health experts generally at a hospital in Pekanbaru. Data is shown in Table 1. Data is classified to get status of monitored person (ODP), monitored patient (PDP) dan people without symptoms (OTG).

Table 1. Dummy Data

ID	Symptoms	Observation of Airway	Travel History	History of Contact with Patient	Patient Status
1	Fever, Cough, Shortness of Breath, Sore Throat	Distrubed	Abroad Affected	Yes	PDP
2	Fever, Cough, Batuk, Shortness of Breath, Sore Throat	Distrubed	Abroad Affected	No	PDP
3	Fever, Cough, Shortness of Breath, Sore Throat	Distrubed	Abroad Affected	No	PDP
....					
8	Fever, Cough, Shortness of Breath, Sore Throat	Not Distrubed	Abroad Affected	No	ODP
...					
14	Anosmia	Not Distrubed	Domestic Red Zone	No	OTG
15	Anosmia	Not Distrubed	Nothing	No	OTG
16	Fever, Cough, Shortness of Breath, Sore Throat	Not Distrubed	Abroad Affected	No	ODP
...					
115	Fever, Cough, Shortness of Breath	Distrubed	Domestic Red Zone	No	PDP

The last step is testing the software using usability, because software generally provides a number of features or menus to make it easier for users [15]. In [11] and [10] research, explained the five components of usability which are used as a measure of software success, namely the system is easy to operate and understand (learnability), the speed of the system can help the user (efficiency),

the system is easy to learn, so if it is not used in the long term, the user still able to easily operate (memorability), the system has a minimum error rate (error) and it means that the user is satisfied to use it and feels helped by this system (satisfaction).

The usability test is carried out in several stages [11]:

Stage I : determine the initial value of the usability component with equation (4)

$$\text{Initial Value} = Pn/T \quad (4)$$

Where: Pn is the Likert scale score

T is the number of respondents

Stage II : determines the percentage of usability components with equation (5)

$$\text{Usability Component} = \frac{\text{Total Score}}{\text{Maximum Score}} \times 100\% \quad (5)$$

Stage III : determine the usability value of the software with equation (6)

$$\text{Usability}\% = \frac{\text{Total Component Percentage}}{5} \quad (6)$$

Usability testing was carried out before implementing online by distributing questionnaires to 30 respondents. Respondents filled out a questionnaire consisting of 14 questions using a five-value Likert scale, as shown in Table 2.

Table 2. Likert Scale

Information	Scale
Very Poor	1
Poor	2
Enough	3
Good	4
Very Good	5

Source: [16]

In Table 2, the scale shows the description of each questionnaire answer. 1 is the lowest scale while 5 is the highest.

4 Results and Analysis

To be able to use software that has been designed, users can login by entering a username and password as in Figure 2. Next, the user is directed to the main menu which consists of Dashboard, Training Data, Calculation Results, Decision Tree and Sign Out.

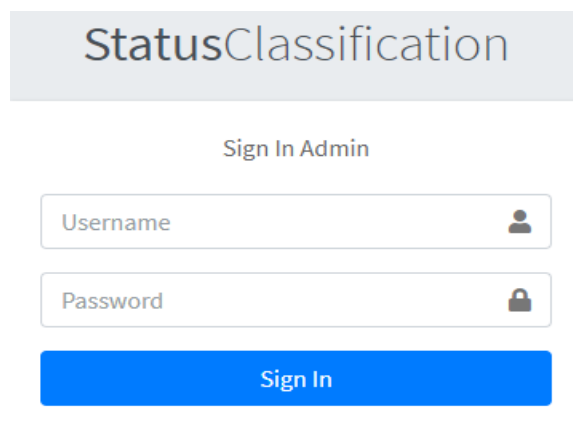


Figure 2. Login Display

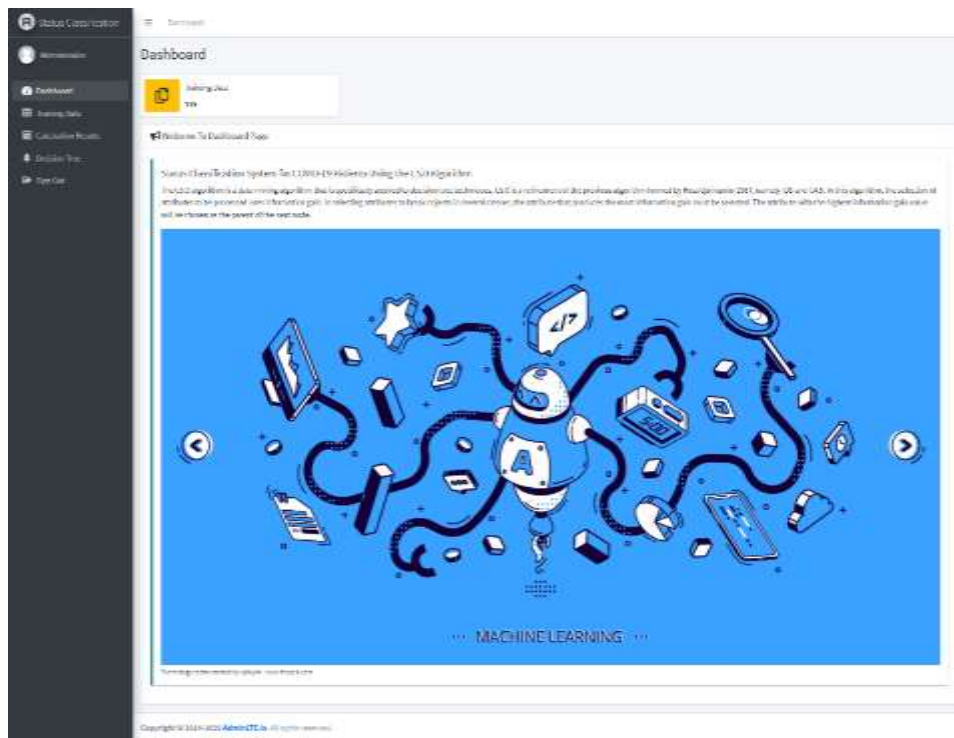


Figure 3. Dashboard Display

Figure 3 is a software dashboard display that is made. Below the dashboard writing, there is a display that shows the amount of data that has been previously entered. If the data does not yet exist or data needs to be taken from outside (import data), a menu of Training Data has been prepared to carry out these activities. The display of the Training Data menu is in Figure 4.

ID	Symptoms	Shortness of Airway	Travel History	History of Contact with Patient	Patient Status	Action
1	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
2	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
3	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
4	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
5	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
6	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
7	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
8	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
9	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
10	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
11	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
12	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
13	Fever, Cough, Shortness of Breath, Sore Throat	Disrupted	Abroad (France)	Yes	RDP	[Edit] [Delete]
14	Asymptomatic	Not Disrupted	Domestic (Not Done)	No	DDP	[Edit] [Delete]
15	Asymptomatic	Not Disrupted	Domestic (Not Done)	No	DDP	[Edit] [Delete]

Figure 4. Data Training Display

After the data is prepared according to the C5.0 algorithm testing needs, the user can carry out the classification process through the Decision Tree menu. To provide users convenience, this software provides a Calculation Result menu for confirmation before the classification process is carried out.

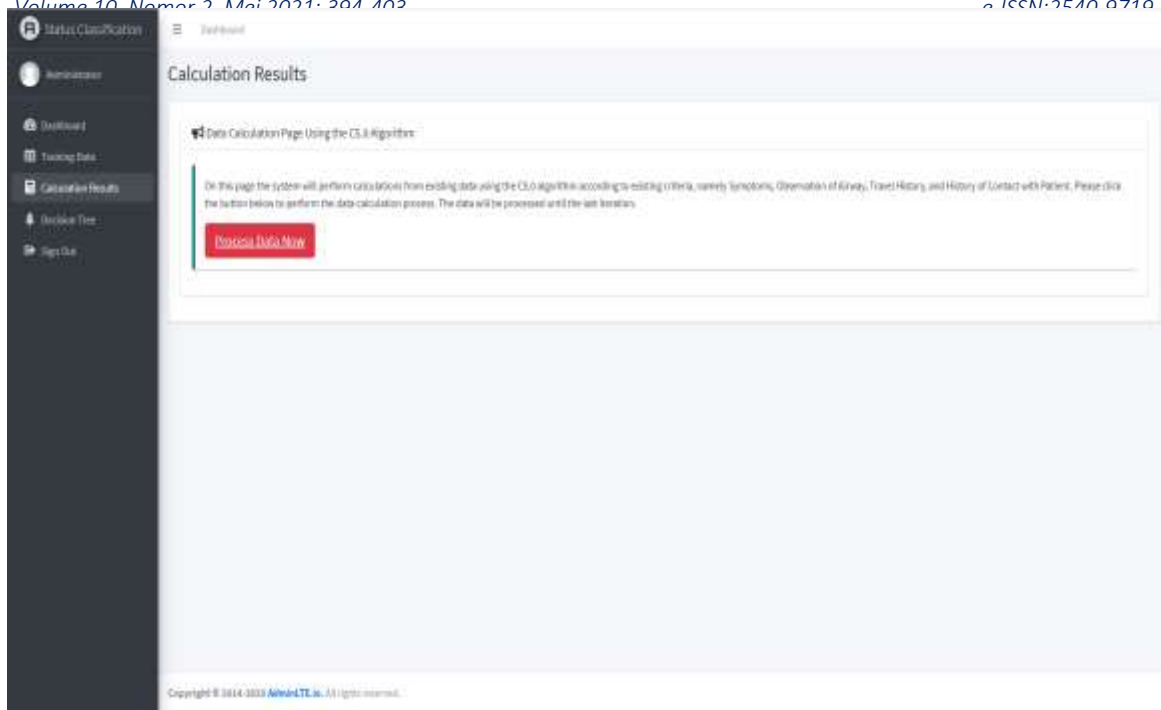


Figure 5. Confirmation Display

The confirmation display shown in Figure 5 aims to ensure the user that the selected data will be tested using the C5.0 algorithm. If there is a data error, the user can return to the previous menu.

Next is the classification process using the C5.0 algorithm by selecting the Process Data Now button. The calculation results will appear in the Decision Tree menu.

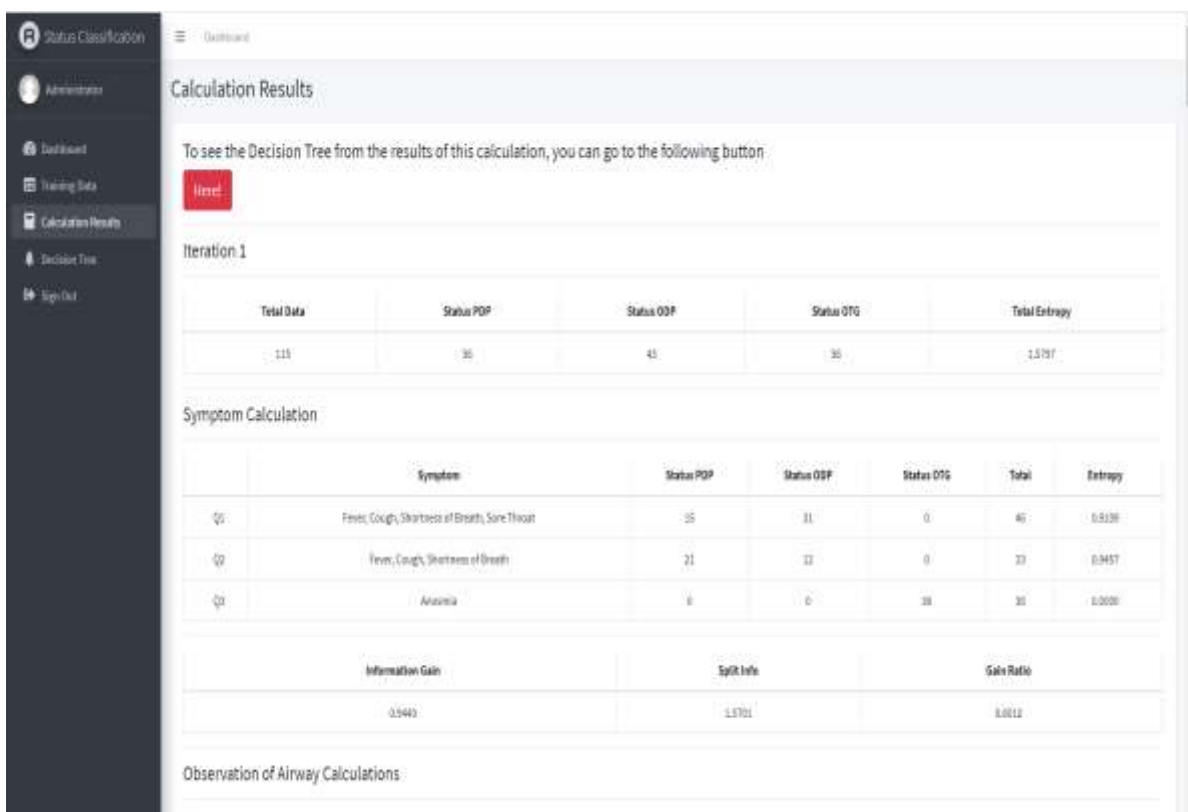


Figure 6. C5.0 Algorithm Calculation Results

Figure 6 is the calculation of the C5.0 Algorithm which is processed using equations (1), (2), and (3). The results are shown in Figure 7.

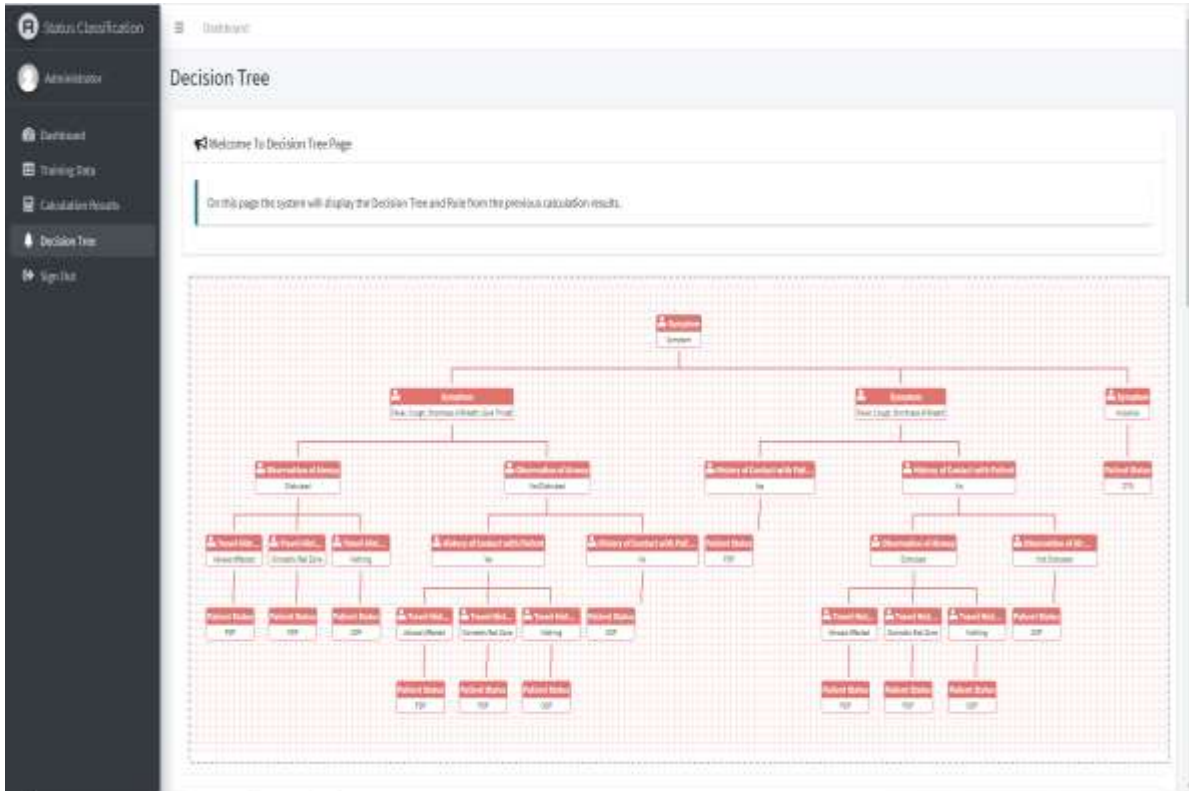


Figure 7. Decision Tree Structure

Figure 7 is a decision tree form of the Covid-19 dummy data classification. The calculation results show that the symptom criterion is the first level (root node) of the structure and becomes the basis for the formation of the next level (leaf node).

The classification process is shown in its entirety, as shown in Figure 6, starting from the search for the Root Node to the leaf node which later becomes a decision tree by referring to equation (1), (2) and (3). Decision tree logic, built automatically using Array. Criteria that can form the next level of leaf nodes as well as criteria that only have one value (positive or negative), can directly form a decision tree as shown in Figure 7. Based on the decision tree, there are rules that shape the desired knowledge as a result of extracting Covid-19 data as shown in Figure 8.



Figure 8. Classification Rules

The classification rules generated by web-based software automatically become knowledge about the characteristics of Covid sufferers with ODP, PDP and OTG status. There are 13 rules. Details can be seen in Figure 9 below.

R1	: If Symptoms = Fever, Cough, Shortness of Breath, Sore Throat ^ Observation of Airway = Disturbed ^ Travel History = Affected Abroad, then Patient Status = PDP
R2	: If Symptoms = Fever, Cough, Shortness of Breath, Sore Throat ^ Observation of Airway = Disturbed ^ Travel History = Domestic Red Zone, then Patient Status = PDP
R3	: If Symptoms = Fever, Cough, Shortness of Breath, Sore Throat ^ Observation of Airway = Disturbed ^ Travel History = None, then Patient Status = ODP
R4	: If Symptoms = Fever, Cough, Shortness of Breath, Sore Throat ^ Observation of Airway = Not Disturbed ^ History of Contact with Patient = Yes ^ Travel History = Infected Abroad, then Patient Status = PDP
R5	: If Symptoms = Fever, Cough, Shortness of Breath, Sore Throat ^ Observation of Airway = Not Disturbed ^ History of Contact with Patients = Yes ^ Travel History = Domestic Red Zone, then Patient Status n = PDP
R6	: If Symptoms = Fever, Cough, Shortness of Breath, Sore Throat ^ Observation of Airway = Not Disturbed ^ History of Patient Contact = Yes ^ Travel History = None, then Patient Status = ODP
R7	: If Symptoms = Fever, Cough, Shortness of Breath, Sore Throat ^ Observation of Airway = Not Disturbed ^ History of Contact with Patient = None, then Patient Status = ODP
R8	: If Symptoms = Fever, Cough, Shortness of Breath ^ History of Contact with Patient = Yes, then Patient Status = PDP
R9	: If Symptoms = Fever, Cough, Shortness of Breath ^ History of Contact with Patient = None ^ Observation of Airway = Disturbed ^ Travel History = Affected Abroad, then Patient Status = PDP
R10	: If Symptoms = Fever, Cough, Shortness Breath ^ History of Contact with Patient = None ^ Observation of Airway = Disturbed ^ Travel History = Domestic Red Zone, then Patient Status = PDP
R11	: If Symptoms = Fever, Cough, Shortness of Breath ^ History of Contact with Patients = None ^ Observations Salu Breath = Disrupted ^ Travel History = None, then Patient Status = ODP
R12	: If Symptoms = Fever, Cough, Shortness of Breath ^ History of Contact with Patients = None ^ Observation of Airway = Not Disturbed, then Patient Status = ODP
R13	: If Symptom = Anosmia, then Patient Status = OTG

Figure 9. All Rules

Software testing is done by distributing questionnaires then processed according to the Usability testing stage. To get the initial value for each Usability component, equation (4) is used with the results as in Table 3.

Table 3. Initial Value of Each Component

Usability Component	Initial Value
Learnability	3.65
Efficiency	4.36
Memorability	4.20
Error	4.34
Satisfaction	3.22

Furthermore, the number of respondents who filled out the questionnaire was calculated based on the Likert scale value. Based on the filled questionnaire data, there were 8 respondents who filled in the value 5, 11 respondents filled in the value 4, 8 respondents filled in the value 3, 3 respondents filled in the value 2 and no respondent filled the questionnaire with a value of 0. To get the maximum score, the number of respondents who fill in according to value, multiplied by the number of respondents. Referring to the equation (5), the percentage of each component of Usability in Table 4.

Table 4. Percentage Usability Component

Usability Component	Percentage
Learnability	76.10
Efficiency	81.53
Memorability	78.78
Error	78.02
Satisfaction	70.03

To be able to provide The predicate of the assessment is based on the percentage of each component according to Table 5, then dividing the percentage interval for the five scales used:

Table 5. Percentage of Likert Scale

Interval	Predicate
81% - 100%	Very Good
61% - 80%	Good
41% - 60%	Enough
21% - 40%	Poor
0% - 20%	Very Poor

Provisions Table 5 shows that the software assessment for each component is in a different predicate, where only the Efficient component is considered Very Good by the respondent because it has a percentage value of 81.53%. The other components are in the Good predicate.

The last step of calculating Usability is obtaining the Usability value of the software itself. Referring to equation (6), the Usability value of the software designed is 76.892%. This means that respondents rated the web-based software that applies the C5.0 algorithm is Good.

5 Conclusion

Overall, the software produced from this study can assist users in classifying the Covid-19 dummy data using the C5.0 algorithm. This is indicated by the usability results of 76.892% and a good level. The various menus in the software are very easy to understand and execute by the user. The software can process data well with the right results. In addition, a menu is provided to retrieve data from outside the software to make it easier for users to change data and add classification criteria as needed. However, when viewed from the satisfaction component, the software still has to be equipped with an attractive appearance and menus that better support the C5.0 algorithm, such as the process of turning and simplifying the rules of the decision tree. This is what makes satisfaction get the lowest score, as shown in table 4. For this reason, the limitations of the software being built can still be developed in order to produce better and more accurate knowledge or information.

Reference

- [1] D. Jollyta, W. Ramdhan, and M. Zarlis, *Konsep Data Mining dan Penerapan*, Pertama. Yogyakarta: Deepublish, 2020.
- [2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques (2nd edition)*, vol. 54, no. Second Edition. 2006.
- [3] S. PANG and J. GONG, "C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks," *Syst. Eng. - Theory Pract.*, vol. 29, no. 12, pp. 94–104, 2009, doi: 10.1016/s1874-8651(10)60092-0.
- [4] S. D. Jadhav and H. P. Channe, "Efficient Recommendation System Using Decision Tree Classifier and Collaborative Filtering," *Int. Res. J. Eng. Technol.*, vol. 3, no. 8, pp. 2113–2118, 2016.
- [5] V. Rahmayanti, Y. Azhar, and A. E. Pramudita, "Penerapan algoritma C5.0 pada analisis faktor-faktor pengaruh kelulusan tepat waktu mahasiswa Teknik Informatika UMM," *J. Repos.*, vol. 1, no. 2, pp. 131–140, 2019, doi: 10.22219/repositor.v1i2.545.
- [6] M. A. Manurung, "Implementasi Data Mining Algoritma C5 . 0 Dalam Sertifikasi Produk

- Pengguna Tanda SNI Pada Air Minum Dalam Kemasan (Studi Kasus : Balai Riset dan Standardisasi Industri Medan),” *J. Comput. Syst. Informatics*, vol. 1, no. 3, pp. 199–206, 2020.
- [7] R. P. Padang, “Implementasi Data Mining Algoritma C5 . 0 Dalam Memprediksi Penerimaan Cleaning Service (Cs) Pada Pt Iss Indonesia Medan,” *Majalah Ilmiah INTI*, vol. 6, pp. 304–309, 2019.
- [8] I. Kurniawan and R. A. Saputra, “Penerapan Algoritma C5.0 Pada Sistem Pendukung Keputusan KelayakanPenerimaan BerasMasyarakat Miskin,” *J. Inform.*, vol. 4, no. 2, pp. 236–240, 2017.
- [9] M. E. Brown and D. L. Hocutt, “Learning to Use, Useful for Learning: A Usability Study of Google Apps for Education,” *J. Usability Stud.*, vol. 10, no. 4, pp. 160–181, 2015, [Online]. Available: <http://www.upassoc.org>.
- [10] R. Andriani, “Evaluasi User Experience Dengan Pendekatan Usability Testing Pada Sistem Informasi Smart Academic,” *Sistemasi*, vol. 9, no. 3, pp. 372–386, 2020, doi: 10.32520/stmsi.v9i3.633.
- [11] Y. N. Marlim, D. Jollyta, and F. Saputra, “Analisis Sistem Jalur Terpendek Menggunakan Algoritma Djikstra dan Evaluasi Usability,” *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 1, pp. 54–60, 2020, doi: 10.26418/jp.v6i1.37627.
- [12] D. Jollyta, M. Wahyudi, H. T. Sihotang, S. Wahyuni, M. Zarlis, and S. Efendi, “Open Access Decision Tree and Chi-Square Analysis to Determine Student ’ s Criteria Choosing Study Program,” *Am. J. Eng. Res.*, vol. 7, no. 10, pp. 12–20, 2018, [Online]. Available: https://www.researchgate.net/profile/Deny_Jollyta/publication/336846142_Ddecision_Tree_and_Chi_Square_Analysis_to_Determine_Student’s_Criteria_Choosing_Study_Program/links/5db65c9d4585155e270b54a2/Decision-Tree-and-Chi-Square-Analysis-to-Determine-Students.
- [13] B. Satrian and G. Gusrianty, “Penerapan Algoritma K-NN Untuk Klasifikasi Gamers Usia Sekolah,” *J. Mhs. Apl. Teknol. Komput. dan Inf.*, vol. 2, no. 1, pp. 19–23, 2020.
- [14] Q. Zhang, J. Zhang, Z. Chen, M. Zhang, and S. Li, “A New Stock Selection Model Based on Decision Tree C5.0 Algorithm,” *J. Invest. Manag.*, vol. 7, no. 4, pp. 117–124, 2018, doi: 10.11648/j.jim.20180704.12.
- [15] L. Setiyani, *Rekayasa Perangkat Lunak [Software Engineering]*, no. May. Karawang: Jatayu Catra Internusa Email, 2018.
- [16] A. Joshi, S. Kale, S. Chandel, and D. Pal, “Likert Scale: Explored and Explained,” *Br. J. Appl. Sci. Technol.*, vol. 7, no. 4, pp. 396–403, 2015, doi: 10.9734/bjast/2015/14975.