Application of blood donor routine detector using K-Nearest neighbors

IOP Conf. Series: Earth and Environmental Science **293** (2019) 012042 doi:10.1088/1755-1315/293/1/012042

Application of blood donor routine detector using K-Nearest neighbors

Y Nurdiansyah[1,*], P Pandunata[1], N D Prasetyo[1], A Trihartono[2], F G Putrianti[3], and F Wijayanto[4]

[1]Faculty of Computer Science, Universitas Jember (UNEJ), Jl. Kalimantan 37, Jember 68121, Indonesia
[2]Center for Research in Social Sciences and Humanities (C-RiSSH), Universitas Jember (UNEJ), Jl. Kalimantan 37, Jember 68121, Indonesia
[3]University of Sarjanawiyata Tamansiswa, Jl. Kusumanegara No.157, Yogyakarta 55165, Indonesia
[4]Institute for Computing and Information Science, Radbound University Comenisulaan 4, 6525 HP Nijmegen The Netherlands

*Corresponding author: yanuar_pssi@unej.ac.id

Abstract. The application of blood donor routine detection using the K-nearest Neighbors method at UTD PMI (*Unit Transfusi Darah, Palang Merah Indonesia*, Indonesian Red Cross, Hereafter PMI) is a system that aims to provide the class status of the donors aimed at facilitating the employee in selecting the right donor to be recontacted. The K-Nearest Neighbors method is used to calculate the distance of the test data with the training data. The K-nearest Neighbors method is chosen because it has a simple algorithm, works based on the shortest distance from the test data sample to the training data sample to determine the distance. The creation of this system is built by adapting the waterfall model. The objective is that the system is able to implement the K-nearest Neighbors method to help determine the class status of the donors and that the system can send SMS gateway to the donors that donor information can be publicly published. The test calculation uses full train full set testing on all training data and the result of K5=84 %, K7=88 %, K9=87 %, K11=82 %, K13=77 %, K15=75 %. The highest result is on the K7.
Keywords: Indonesian Red Cross, SMS gateway, waterfall model

1. Introduction

This paper examines the application of blood donor routine detection using the K-nearest Neighbors method at UTD PMI (*Unit Transfusi Darah, Palang Merah Indonesia*, Indonesian Red Cross, Hereafter PMI) at Jember District, East Java Province, Indonesia. Practically, even though Jember District is one of the advanced cities in East Java, Indonesia, in the PMI of Jember District the process of recording is still traditional and manual. The PMI still uses Microsoft Excel applications in data collecting. Regarding, this research aims at offering methods that will significantly help the efficiency and effectiveness of the process of recording donors.

The PMI is a neutral and independent organization of which activity and objective are to help humanity voluntarily. PMI prioritizes those who seek immediate help for the safety of their lives. The objective of PMI is to facilitate and help fellow human beings regardless of the causes, groups,

nationalities, skin color, gender, language, religion or belief in times of peace and war. One of the activities often held by PMI is the blood transfusion unit.

Blood transfusion constitutes one of the basic activities held by PMI because via blood transfusion a patient who needs blood or is undergoing blood loss can be helped. The blood transfusion process requires a bloodstock based on the blood type of the patient. Every donor who would like to donate at PMI can help a patient and community in general [1].

Recent data shows that the demand for blood by both medical institutions and individuals increases significantly. This significant increase in blood demand requires PMI to implement accurate action regarding the availability of bloodstock. One means is to conduct blood donation in various strategic places, such as schools, government offices, or city centers [2]. Such methods certainly have disadvantages, one of which is a large amount of cost incurred. The said cost is in terms of transportation and consumption, while the availability of the donors is considered not to be optimal.

The solution to the issue is to contact the donors who have previously registered that they can make another blood donation [3]. Contacts can be made via telephone or SMS to the previous donors to provide reconfirmation and make a returning donor [4, 5]. However, it raises another issue that is the cost incurred to contact the donors. The result of the conducted survey shows that the total SMS sent is more than 3 000 SMS per month with 125 to 320 SMS per day, meaning that the total cost incurred ranges from IDR 750 000 to IDR 1 920 000 per month. This total cost is indeed not balanced out with the number of the donors reminded since the number of the donor making a returning donor is merely 60 %.

On the other hand, such issue can be resolved by a systematic approach, namely opening new member registration; of which purpose is to compare the data of the new registrants with the previous donors to determine the criteria if the person is a routine donor. The method used is the K-Nearest Neighbors (KNN) method [6, 7]. It is considered to be suitable because the method includes data classification of a high level of consistency by calculating the proximity between the new case and the old case based on weight matching. This algorithm is considered to be more effective in conducting large data training and can produce more accurate data.

2. Research method

Study on application of routine detection using the K-nearest Neighbors method at UTD PMI Regional Jember used the System Development Life Cycle (SDLC) waterfall [8, 9, 1. The waterfall model is a systematic and sequential model starting from the level and progress of the system to the analysis, design, code, test, and maintenance.

The development of this system used the waterfall model because it was built based on the information system needs. The built system was a small scale system; the human resources (HR) building the system consists of one person, and the system is adjusted based on its users. The implementation of the K-nearest Neighbors method begins with training data input, data training labeling, k, and data testing [11–15]. The implementation of the KNN method is presented in figure 1.

Figure 1 illustrates the implementation of the KNN method to the system starting from the data attribute input and followed by calculating the distance to determine the *Euclidean distance*. Afterward, the distance list is sorted starting from the smallest data. The distance list of distances was sorted according to the specified k. The collected data according to the K is determined its majority class. The result of the data classification is identified. The system development consists of an analysis of needs and system design.

2.1 Analysis of needs

Based on the waterfall system development method, the initial step is the analysis. The analysis was performed on the study object to obtain all needs of the built system including functional needs and non-functional needs.

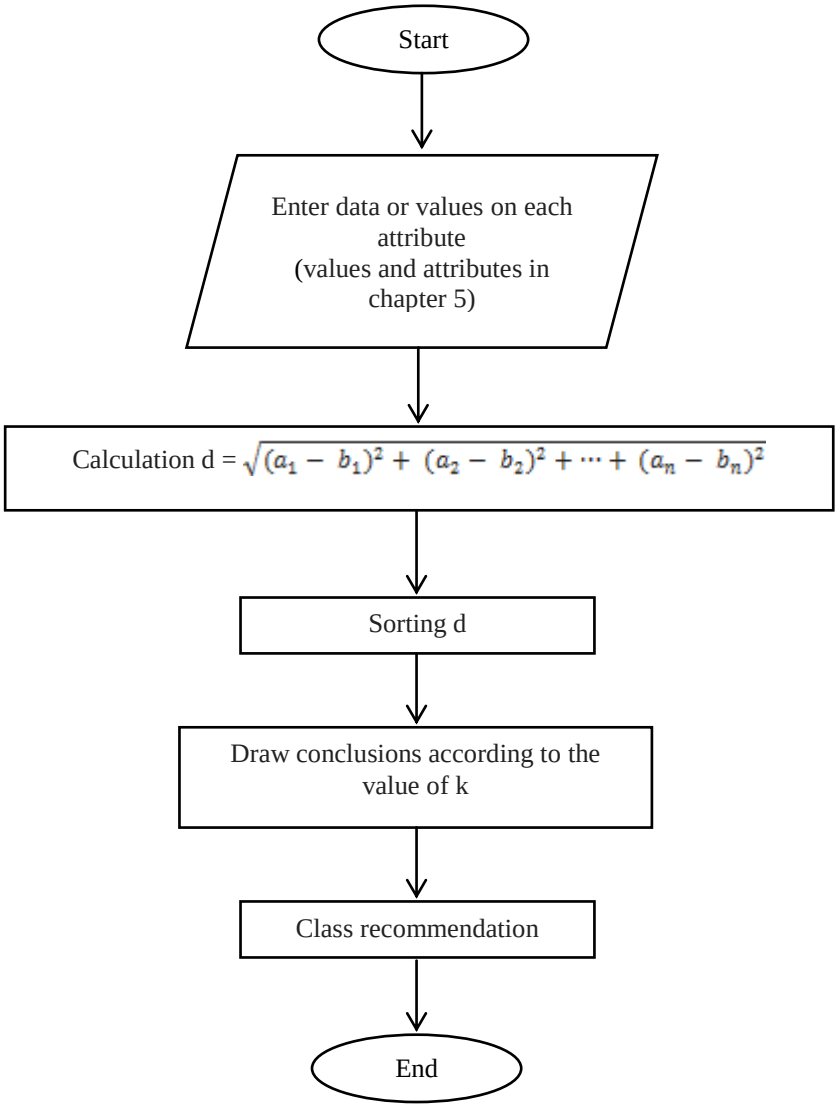

Figure 1. Flowchart of the KNN implementation to the system.

The application of blood donor detection by implementing the K-nearest Neighbors method is an application to classify if a donor is a routine blood donor. The system requires the input of the donor's personal data to be used as reference including address, status, gender, age, and occupation. The classification algorithm based on the KNN method works on the aforementioned input and is compared with the previously available dataset. This system is expected to increase the efficiency of the SMS sending process since the donor status is identified regarding the routine estimation of whether or not the donor is to make a return donation. The objective of the system development is to reduce SMS Gateway financing and improve SMS efficiency because it can only send messages to donors categorized as a routine donor monthly.

2.2. System design

The system design includes a business process and uses case diagram [16].

2.2.1. Business process. A business process is a diagram encompassing the data needed by the system. There are several components including input data, such as the donor data and SMS. The input data is processed into output data that is the donor list and SMS list. The media used is a website, and the

objective is that the system is able to provide a classification of the donors to be contacted during blood donor. The Business Process can be seen in figure 2.

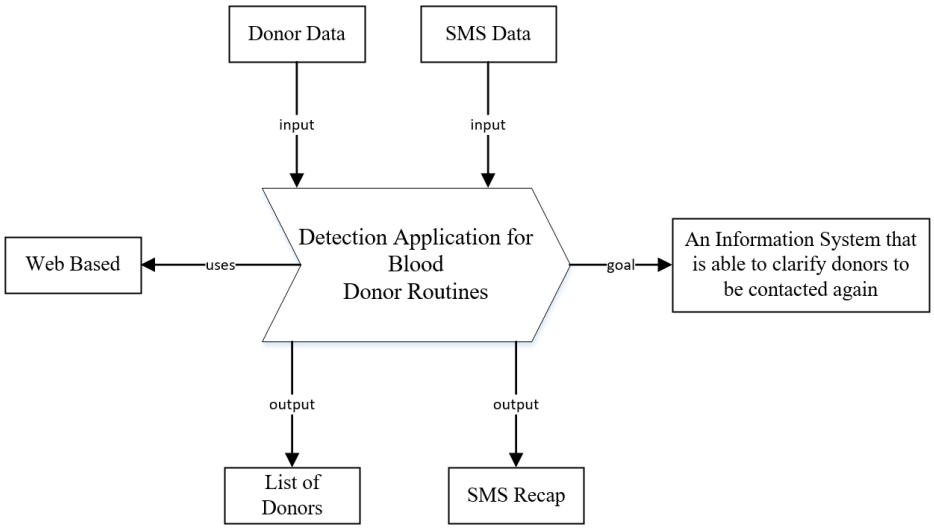

Figure 2. Business process

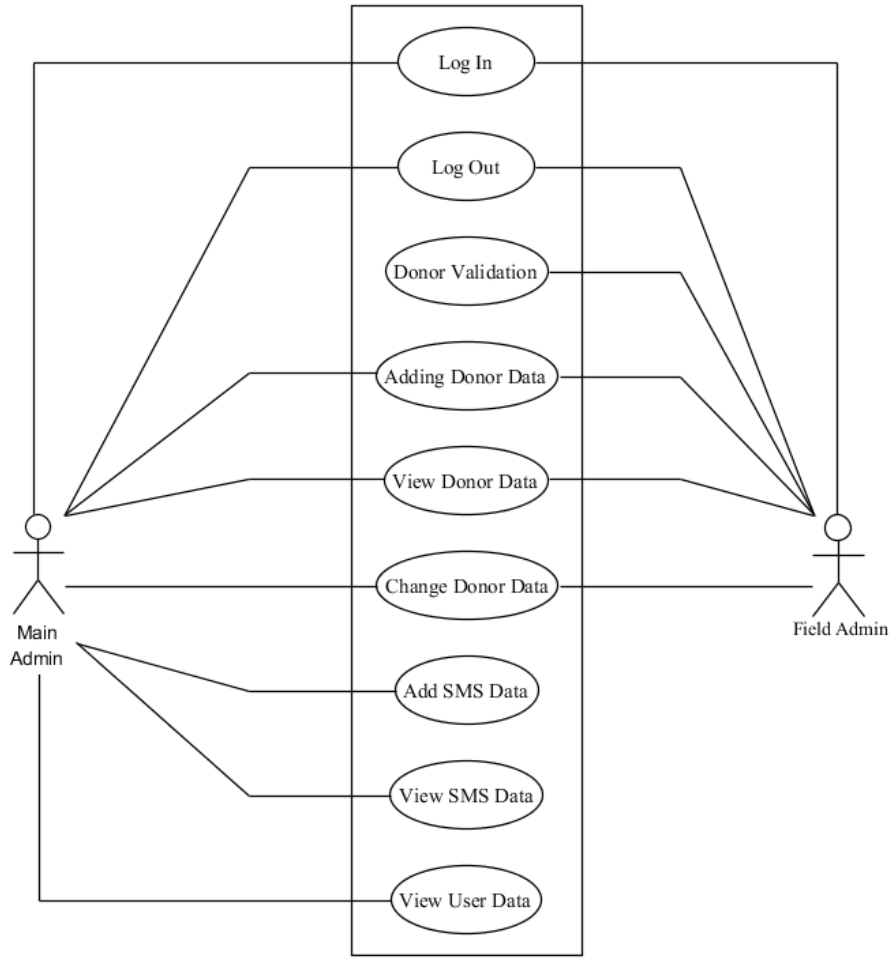

Figure 3. Use case diagram.

2.2.2. Use case diagram. Use case diagram is used to describe the interaction between the actor and the system. The use case diagram identifies the interaction of the actor with the system in accordance with the access rights owned by the actor or user. The use case diagram consists of two actors and nine features. The main actor is the field admin containing eight features. The second actor is the field admin containing six features. Use case diagram is presented in figure 3.

3. Discussion

3.1 Calculation of K-Nearest neighbors

The process of calculating the K-nearest Neighbors method is used to determine the class status of the donors. The status later explains if the donor is a routine donor; from such a message is sent to the donor to make a return donor. The implementation of K-nearest Neighbors on the system is based on the shortest distance of the test data to the training data. Searching for the shortest distance at K-nearest Neighbors usually uses Euclidean distance. Before calculating the distance, the test data and training data are given the weight of the predetermined criteria. K-Nearest Neighbors identifies the k in the shortest data order to determine the class status. The following is the steps of the K-nearest Neighbors algorithm.

Step 1. Determine the value of K = 7.

Step 2. Input data to be tested. For example, a donor named Sujiwo Tejo. His gender, date of birth, domicile, occupation and marital status are male, February 22, 1996, outside Jember, student, and single, respectively.

Step 3. Weighing of the test data.

Gender, age, marital status, domicile and occupational in the test data weigh 2, 1, 2, 2, 1. Hence, the weight of the first training data is 2, 1, 1, 1, 1.

Step 4. Calculating the Euclidean distance

The following is the calculation of Euclidean distance [17] in equation (1) in the test data with the first training data:

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2} \quad (1)$$

In which $a = a_1, a_2, .. a_n$ and $b = b_1, b_2, .. b_n$ represents n the attribute value of the two records and d is the distance.

$$d = \sqrt{((2)-(2))^2 + ((1)-(1))^2} + ((2.5)-(1))^2 + ((4)-(1))^2 + ((1)-(1))^2 \qquad d = 3.35$$

The distance between the test data and the first training data is 3.35. The Euclidean Distance (ED) calculation to the 100th data can be seen in table 1. Gender, marital status, address, occupation, age, routine and non-routine, are shortened into G, MS, Add, Occ, A, R, and NR, respectively.

Table 1. The result of Euclidean distance.

No	G	MS	Add	Occ	A	Class	E.D
1	2	1	1	1	1	R	3.35
2	1	1	4	3	1	NR	2.69
3	1	1	1	1	1	R	3.50
4	2	2.5	4	2	1	NR	1.00
5	2	2.5	1	2	1	R	3.16
6	2	1	4	1	1	R	1.50
7	2	2.5	4	1	1	NR	0.00
8	2	2.5	4	1	1	NR	0.00
9	2	1	4	1	1	R	1.50
10	2	2.5	4	1	1	NR	0.00
11	2	2.5	4	1	1	NR	0.00
12	2	2.5	4	3	1	NR	2.00
13	2	1	1	1	1	R	3.35
14	1	1	4	1	1	NR	1.80

Continue on next page.

Table 1. Continued.

No	G	MS	Add	Occ	A	Class	E.D
15	1	1	4	1	1	NR	1.80
16	2	1	4	1	1	R	1.50
17	2	1	4	1	1	NR	0.00
18	1	2.5	4	2	3	R	2.45
19	1	2.5	4	2	2	NR	1.73
20	1	1	4	1	1	NR	1.80
21	2	2.5	4	1	4	NR	3.00
22	2	2.5	4	1	1	NR	0.00
23	2	2.5	4	1	4	NR	3.00
24	2	2.5	4	2	3	NR	2.24
25	2	1	4	1	1	NR	1.50
26	2	2.5	4	1	3	R	2.00
27	2	2.5	4	1	1	NR	0.00
28	2	2.5	4	1	1	NR	0.00
29	2	1	1	2	1	R	3.50
30	1	1	4	3	1	NR	2.69
31	2	2.5	4	3	4	R	3.61
32	2	1	1	3	1	R	3.91
33	2	2.5	4	1	2	R	1.00
34	2	2.5	1	3	2	R	3.74
35	2	2.5	4	2	4	R	3.16
36	2	2.5	4	2	4	R	3.16
37	2	2.5	4	1	4	NR	3.00
38	2	2.5	4	2	4	R	3.16
39	2	2.5	4	2	4	NR	3.16
40	2	2.5	4	2	4	NR	3.16
41	2	2.5	4	2	4	R	3.16
42	2	2.5	4	1	1	NR	0.00
43	2	2.5	4	2	4	NR	3.16
44	2	1	4	1	1	R	1.50
45	2	2.5	4	2	2	R	1.41
46	2	2.5	4	1	2	R	1.00
47	2	2.5	4	2	2	R	1.41
48	1	2.5	4	2	2	NR	1.73
49	2	1	4	1	1	NR	1.50
50	2	2.5	4	2	3	R	2.24
51	2	2.5	4	3	3	R	2.83
52	1	2.5	4	2	2	NR	1.73
53	1	2.5	4	2	4	NR	3.32
54	2	2.5	4	2	4	R	3.16
55	2	2.5	4	1	1	NR	0.00
56	2	2.5	4	2	2	NR	1.41
57	2	1	4	2	1	R	1.80
58	2	2.5	4	2	3	NR	2.24
59	2	2.5	4	1	2	NR	1.00
60	2	2.5	4	2	4	NR	3.16
61	2	1	4	2	1	NR	1.80
62	1	1	4	1	1	NR	1.80
63	2	2.5	4	3	3	R	2.83
64	2	2.5	4	2	2	R	1.41
65	2	2.5	4	2	4	R	3.16
66	2	1	4	1	1	NR	1.50
67	2	2.5	4	3	3	R	2.83

Continue on next page.

Table 1. Continued.

No	G	MS	Add	Occ	A	Class	E.D
68	1	2.5	4	2	2	NR	1.73
69	2	2.5	4	3	3	R	2.83
70	2	2.5	4	2	4	NR	3.16
71	2	2.5	4	2	4	NR	3.16
72	2	2.5	4	2	4	R	3.16
73	2	2.5	4	2	4	R	3.16
74	2	1	1	2	1	R	3.50
75	2	2.5	4	3	4	NR	3.61
76	2	2.5	4	3	3	R	2.83
77	2	2.5	4	2	4	NR	3.16
78	2	2.5	4	2	4	NR	3.16
79	2	2.5	4	3	3	NR	2.83
80	2	2.5	4	3	2	R	2.24
81	1	2.5	1	3	4	R	4.80
82	2	1	1	1	1	NR	3.35
83	2	2.5	1	3	4	R	4.69
84	2	2.5	1	2	2	NR	3.32
85	2	2.5	1	2	2	R	3.32
86	1	2.5	1	1	1	NR	3.16
87	1	1	1	1	1	NR	3.50
88	1	1	1	1	1	NR	3.50
89	2	1	1	1	1	NR	3.35
90	2	2.5	1	1	2	R	3.16
91	1	1	1	1	1	NR	3.50
92	2	1	1	2	1	R	3.50
93	1	1	4	3	1	NR	2.69
94	1	1	4	2	1	R	2.06
95	2	2.5	4	3	2	R	2.24
96	2	1	1	2	1	R	3.50
97	2	2.5	4	2	2	NR	1.41
98	2	1	4	1	1	R	1.50
99	1	2.5	4	3	3	NR	3.00
100	1	2.5	4	4	4	R	4.36

Step 5. The count data were sorted and taken based on K=7 as in table 2.

Table 2. The result of sorting based on K

No	JK	SM	A	P	A	Class	E.D
7	1	2.5	4	2	1	1	NR
8	2	2.5	4	2	1	1	NR
10	2	2.5	4	2	1	1	NR
11	2	2.5	4	2	1	1	NR
17	2	2.5	4	2	1	1	NR
27	2	2.5	4	2	1	1	NR
28	2	2.5	4	2	1	1	NR

Step 6. Drawing a conclusion.

Based on the data sorting according to Euclidean distance, the data retrieval process is based on K, and the majority class is listed as in table 2. It can be concluded that the classification result for the test donor under the name of Sujiwo Tejot is categorized as not routine. Selecting the Euclidean distance is for determining routine classes and non-routine classes. Giving values for attributes with category values is done by looking at the total indicators of each of the supporting attributes used in calculating the k-nearest neighbors. For example, gender attributes consist of women and men. The total indicator

of this attribute is "2", then the determination of values "1" and "2" is given for female and male indicators of the sex attribute. Weight "2" is given to men based on the results of interviews where Male are more donors than Female.

3.2. The result of system testing

This test used 20 training data of which class status is hidden into the test data (full train full set). The K values used in the testing are K5, K7, K9, K11, K13, and K15. In determining the value of K, so far there are no rules to determine it. A high K value will reduce the level of accuracy in data classification. So, this research chose k5 to k15 and used some of this "k" to test as many as 100 data to produce a level of accuracy that is considered good (90 %). The accuracy value will depend on the size of the data and / or the value of K, so the accuracy of the data will differ in results. The test result can be seen in table 3, of which the terms of true, false, true percentage and false percentage are shortened into T, F, TP, and FP, respectively. The calculation to determine the True Percentage (TP) is in equation (2).

$$TP = \frac{Correct\ Amount}{Number\ of\ Sample\ Data} \times 100\,\% \quad (2)$$

Table 3. System testing

Value	K7	K9	K11	K13	K15
T	88	87	82	77	75
F	12	13	18	23	25
TP	88 %	87 %	82 %	77 %	75 %
FP	12 %	13 %	18 %	23 %	25 %

Based on table 3, the accuracy value of K5, K7, K9, K11, K13 and K15 is 84 %, 88 %, 87 %, 82 %, 77 %, 75 %, respectively. The highest true percentage value is obtained at K7 of 88 % followed by K9, K5, and K11 of 87 %, 84 % and 82 %, respectively; and the true percentage value of K13 and K15 of 77 % and 75 %.

4. Conclusion

The implementation of the K-Nearest Neighbors method in the blood donor detection application is able to determine the classification result based on the parameters used with 88 % accuracy in K7. The accuracy of K is calculated using the full train full set calculation or class status hiding of all selected test data. Accuracy testing is conducted using K5, K7, K9, K11, K13, and K15. The accuracy testing obtains the highest true percentage value at K7 of 88 % followed by the true percentage values of K9, K5, and K11 of 87 %, 84 %, and 82 %; and the true percentage values of K13 and K15 of 77 % and 75%. The accuracy value depends on the data and/or the K value that the data accuracy varies according to the result of each data and the K value.

The development of the blood donor detection application using the K-Nearest Neighbors method uses the waterfall model. The data were collected from observation and interview at UTD PMI Jember. The system design phase implements the Unified Modeling Language (UML) modeling designed using the Object Oriented Programming (OOP) concept. The making of the detection application for blood donor routine also includes the SMS Gateway feature used to provide message to the donors. SMS Gateway reduces the cost incurred to send message done by UTD PMI Jember that it can improve the efficiency of sending message, since it only sends message to the routine donors.

Further system development is expected to add more training data by considering donor history data in determining the class. So that it can improve the accuracy of the calculation results on this system. The addition of training data is needed because the more training data used, the more training data is compared to or calculated with the test data.

References

[1] Quéniart A 2013 Blood donation within the family: The transmission of values and practices *Transfusion* **53**(5) 151S–156S
https://onlinelibrary.wiley.com/doi/full/10.1111/trf.12474

[2] Kowsalya V, Vijayakumar R, Chidambaram R, Srikumar R, Reddy E P, Latha S, Fathima I G and Kumar C K 2013 A study on knowledge, attitude and practice regarding voluntary blood donation among medical students in Puducherry, India *Pak. J. Biol. Sci.* 2013 **16**(9) 439–42
https://www.ncbi.nlm.nih.gov/pubmed/24498809

[3] Schlumpf K S, Glynn S A and Schreiber GB 2008 Factors influencing donor return *Transfusion* **48**(2) 264–72
https://www.ncbi.nlm.nih.gov/pubmed/18005325

[4] Mousavi F, Tavabi A A, Golestan B, Ammar-Saeedi E, Kashani H, Tabatabaei R and Iran-Pour E 2011 Knowledge, attitude and practice towards blood donation in Iranian population *Transfusion Med.* 2011 **21** 308–17
https://www.ncbi.nlm.nih.gov/pubmed/21696474

[5] Gillespie T W and Hillyer C D 2002 Blood donors and factors impacting the blood donation decision *Transfusion medicine Reviews* **16** 115–17
https://www.ncbi.nlm.nih.gov/pubmed/11941574

[6] Mucherino A, Petraq J P and Panos M P 2009 *k-Nearest Neighbor Classification* (New York: Springer) pp 83–106
https://link.springer.com/chapter/10.1007/978-0-387-88615-2_4

[7] Zhang Z, 2016 Introduction to machine learning: K-nearest neighbors *Ann. Transl. Med.* **4**(11) 218
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916348/

[8] Leau Y B, Loo W K, Tham W Y and Tan S F 2012 Software development life cycle AGILE vs traditional approaches *IPCSIT* **37** 162–67
https://pdfs.semanticscholar.org/69b1/9ddc8a578f4c63d1dfe15252a465ee12fe5d.pdf

[9] McMurtrey M 2013 A case study of the application of the systems development life cycle (SDLC) in 21st century health care: Something old, something new? *Journal of the Southern Association for Information Systems* **1**(1) 15–25
https://www.researchgate.net/publication/269662156_A_Case_Study_of_the_Application_of_the_Systems_Development_Life_Cycle_SDLC_in_21st_Century_Health_Care_Something_Old_Something_New

[10] Nurdiansyah Y, Wijayanto F and Firdaus F 2018 The design of e-commerce system in the shrimp paste industry using the method of Structured Analysis and Design Technique (SADT) to increase marketing *MATEC Web Conf.* **164**(2018) 01049
https://doi.org/10.1051/matecconf/201816401049

[11] Gorunescu F 2011 *Data Mining : Concepts, Models and Techniques* (Springer: Verlag Berlin Heidelberg) pp 1–10
https://www.springer.com/in/book/9783642197208

[12] Balasubramanian T and Umarani R 2012 An analysis on the impact of fluoride in human health (dental) using clustering data mining *Technique Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering* (Salem, Tamilnadu: Prime) pp. 221–23
https://ieeexplore.ieee.org/document/6208374/

[13] Molovic B and Milovic M 2012 Prediction and decision making in health care using data mining *IJPHS* **1**(2) 69–78
https://www.iaescore.com/journals/index.php/IJPHS/article/view/4593/3465

[14] Nurdiansyah Y F, Muharrom N and Firdaus F Implementation of winnowing algorithm based K-gram to identify plagiarism on file text-based document *MATEC Web Conf.* **164**(2018) 01048
https://doi.org/10.1051/matecconf/201816401048
[15] Nurdiansyah Y S, Bukhori and Hidayat R, Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method *J. Phys. Conf. Ser.* **1008** 012011
https://iopscience.iop.org/article/10.1088/1742-6596/1008/1/012011/meta
[16] Mäkinen V 2003 *Analysis of Use Case Approaches to Requirements Engineering* Master thesis (Jyväskylä: Department of Computer Science and Information Systems, University of Jyväskylä) pp 35–39
https://jyx.jyu.fi/dspace/bitstream/handle/123456789/12463/G0000231.pdf?sequence
[17] Bramer M 2007 *Principle of Data Mining: Undergraduate Topics in Computer Science* (London: Springer Verlag) pp 35–36
https://www.springer.com/gp/book/9781447173069