Predicting Potential Blood Donors Who Can Attend Blood Donation Activities using a Support Vector Machine

Andreas Handojo
Informatics Department, Faculty of Industrial Technology, Petra Christian University
Indonesia
handojo@petra.ac.id

Anita Nathania Purbowo
Informatics Department, Faculty of Industrial Technology, Petra Christian University
Indonesia

Gita Berliany Karaeng
Informatics Department, Faculty of Industrial Technology, Petra Christian University
Indonesia

Tanti Octavia
Industrial Engineering Department, Faculty of Industrial Technology, Petra Christian University
Indonesia
tanti@petra.ac.id

Abstract— **Lack of blood will be fatal for the human body. Current technology has not been able to produce human blood, therefore blood donors from other people are needed. Because of this need, the Red Cross organized blood donation activities to obtain blood supplies. Blood donation is given by people voluntarily, therefore it is difficult to predict how much blood supply will be obtained in an organized blood donation activity. A system is needed to predict the number of potential donors so that the supply is sufficient. This study will classify potential blood donors and predict the number of blood donors who have the possibility to attend a blood donation activity held certain location at a certain time by using a support vector machine. With this prediction, the Red Cross can predict in advance how many bags of blood can be obtained before carrying out blood donation activities at a certain location at a certain time. In this way, the Red Cross will be able to find the right place and date to obtain maximum blood donation. The dataset used is data on blood donations carried out by donors at the Indonesian Red Cross in 2015 - 2019 with a total data of 53708 donors. From this study, it was found that the application made was able to classify potential donors and predict donors who could be present to give donors at a certain location and at a certain time with an F1 Score of 85%.**

Keywords- Support Vector Machine; Blood Donor Potential; Prediction; Red Cross; Indonesia

I. INTRODUCTION

Blood is very important for the human body. Lack of blood can cause the body to become sick and risk death. Until now, blood can only be produced by the human body, no technology has been found to be able to produce blood [1]. Therefore, if a person lacks blood, that person needs additional blood or what is often called a blood transfusion. Blood transfusion activities are carried out by transfusing from a blood bag into the patient's body. These blood bags are obtained through blood donations from other people in blood donation activities. Generally, blood donations are given voluntarily through the Blood Transfusion Unit. Currently, Indonesia is only able to obtain 4.1 million blood bags/per year [2, 3], this is still far from the World Health Organization/WHO standard, which is at least 2% of the total population [4]. Where Indonesia has a population of about 255 million people (approximately at least 5.1 million bags of blood). This is one of the major causes of death due to lack of blood, such as during childbirth, which amounted to 28%. Of course, this must be overcome by providing adequate blood reserves provided [5].

Blood donation activities are generally carried out by the government and Red Cross organizations or in Indonesia known as the Indonesian Red Cross (PMI). Blood donation activities are served every day at PMI headquarters in several cities. In addition, blood donation activities are also held at certain events organized by PMI, the government, the private sector, community organizations, religious organizations, and others. Event organizers can contact PMI to ask PMI to hold blood donation activities at the event being held. According to a report from the Ministry of Health of the Republic of Indonesia, there are 420 blood transfusion units (220 units managed by PMI and 200 managed by the government) distributed in 34 provinces [6].

Generally, these blood donation activities are announced to prospective donors using banners on the streets, so that they post the activity on the official PMI website and social media (Facebook / Instagram). This makes it difficult for information to be distributed considering that information is given generally without any purpose to donors who have the potential to donate blood. Remembering that not do it every day someone can donate. The donation can only be done for approximately 56 days (to do a whole blood donor) after someone has done a donor activity. Where after donating, the body must reproduce the blood that has been removed. Therefore, a system is needed to be able to invite and remind donors to donate again after passing the required blood reproduction time. So that the stock of blood bags can still be available for those in need.

A system is needed that can predict potential blood donors who have a high probability of contributing to donating blood in a blood donation activity that is carried out in a particular location. Based on the prediction data of potential blood donors, the distribution of information/invitations to blood donors will be distributed specifically. So, it is hoped that blood donation activities can run more effectively. In addition, the system is also expected to be able to predict how many chances of obtaining blood bags from the blood donation activities that are held.

This study will use the Support Vector Machine method to classify blood donors into potential donors and non-potential donors to donate blood at blood donation activities at certain locations at certain times. The classification of potential/non-potential donors aims to predict how many donors will attend and donate blood if the Red Cross carries out a blood donation activity at a certain location at a certain time. With this, it is hoped that the Red Cross can determine the most appropriate location and time to carry out blood donation activities to obtain maximum results.

II. Support Vector Machine

Support Vector Machine (SVM) is a method of pattern recognition that has received a lot of attention. SVM is an attempt to find the optimal hyperplane that is used to separate the two classes [7]. The optimal hyperplane between the two classes is obtained by measuring the hyperplane margin to find its maximum point. The hyperplane in the d-dimensional room is an affline subspace d-1 which will divide the vector space into two different class parts [8]. The margin is the distance between the closest hyperplane of each class, while the closest pattern is called a support vector. Two conditions can be solved by SVM, data conditions that can be separated linearly and data that cannot be separated linearly.

The main purpose is to separate l training set examples with vector data x_i and class labels y_1 $(x_1,y_1),\ldots(x_l, y_{(l)}) \in R^n$ X $\{-1,1\}$ by searching weight vector $w \in R^n$ and $b \in R$ from hyperplane:

$$H: \mathbb{R}^n \longrightarrow \{-1, 1\} \quad (1)$$

$$x \longmapsto sign(w.x + b) \quad (2)$$

With very many possible margins. So in the case of linearly separable data, hyperplane (w,b). Optimization problem solving:

$$\begin{cases} minimize\ \frac{1}{2}\parallel w \parallel^2 \\ subject\ to: (w.x_i + b) \geq 1\ (i = 1, \ldots, \ell) \end{cases} \quad (3)$$

Then the maximum margin of the hyperplane is

$$\gamma = \frac{1}{2}\parallel w \parallel^2 \quad (4)$$

The algorithm is used for data that can be separated linearly but can be used to generalize to data that cannot be separated linearly with the use of non-negative slack variables. ξ_i [9]. Resultant problems are minimized:

$$\begin{cases} \frac{1}{2}\parallel w \parallel^2 + C\sum_{i=1}^{n}\xi_i \\ subject\ to: (w.x_i + b)y_i \geqslant 1 - \xi_i\ \xi_i > 0\ \forall i \end{cases} \quad (5)$$

After fixing this problem with positive Lagrangian multipliers, the optimization problem is maximized:

$$\begin{cases} \sum_{i=1}^{\ell}\propto_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\propto_i\propto_j y_i\ , y_j K(x_i, x_j) \\ subject\ to: 0 \leq \propto_i \leq C\ and\ \sum_{i=1}^{\ell}\propto_i y_i = 0 \end{cases} \quad (6)$$

After doing the mapping, in a higher dimension, the function for solving the problem becomes:

$$\begin{cases} \sum_{i=1}^{\ell}\propto_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\propto_i\propto_j y_i\ , y_j K(x_i, x_j) \\ subject\ to: 0 \leq \propto_i\ and\ \sum_{i=1}^{\ell}\propto_i y_i = 0 \end{cases} \quad (7)$$

Where K will design kernel functions. The hyperplane solution has the following functions:

$$h(x) = \sum_{i,j=1}^{\ell}\propto_i K(x_i, x_j) + b \quad (8)$$

For further implementations, we will use the most commonly used kernels such as linear kernels, polynomial kernels, and radial bases [10].

Linear Kernel:

$$K(x_i, x_j) = x^{\Upsilon}x_i \quad (9)$$

Polynomial Kernel:

$$K(x_i, x_j) = (x^{\Upsilon}x_i + 1)^d \quad d > 1 \quad (10)$$

Radial Kernel:

$$K(x_i, x_j) = \exp\left(\frac{-\parallel x^T - x_i \parallel^2}{2\sigma^2}\right) \quad \sigma > 0 \quad (11)$$

III. Blood Donor Classification with SVM

The prediction process is carried out with 4 main processes, starting from the process for selecting parameters, sorting the dataset according to parameters, the classification, and the prediction process using SVM, and the process for grouping the classification results. As can be shown in Fig. 1.

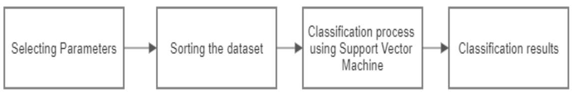

Figure 1. Blood Donor Potential using SVM

In the first process, the available parameters will be selected to determine the shape of the dataset that will be used in the SVM process later. The parameters that need to be inputted by the user are the location where potential blood donors will be predicted, the time the activity will be carried out, and the time to see the range of data to be taken.

The dataset selection is done by querying the database which results in a new dataset. Based on these parameters, the new dataset to be formed has several attributes. The first attribute is the time interval between the first time the donor donates blood to the date the event will be held. The second attribute is the time interval between the last time the donor donated blood and the event date. The third attribute is the frequency of donors making blood donations. The fourth attribute is the frequency of donors making blood donations at that location. The fifth attribute is the distance between the last time the donor donated blood at that location and the event date. The sixth attribute is the amount of blood donated by donors (in cc). The seventh attribute is the ID of the donor, and the last attribute is the donor's blood group. These eight attributes will be used to predict potential donors.

In addition to selecting the dataset for testing data, sorting out the dataset for training data is also needed. For selecting the dataset, you will use the data in the previous event at the inputted location. Based on the event data, a new dataset will be formed which has the same attributes as the testing data. However, in the training data there are additional attributes, namely attributes that contain information on whether the donor made blood donations at the event.

After the data training and data testing have been formed, the next step is to run SVM using the two datasets. To run SVM is done using the library from Sklearn. Because the form of data from the dataset that is owned cannot be separated linearly, SVM will use the Radial Base Kernel to get faster and optimal results compared to other kernels. After the SVM is executed, the results of the training data will be managed to display. Data will be grouped according to blood type, potential donors or not, and other forms of data as needed.

This prediction and filtering feature serve to display locations that have the number of potential blood donors who meet the input requirements. The system will only display the location of activities that have potential donors according to the requests that have been entered by the user based on blood type. On the maps section, markers will only appear in locations that meet the requirements. When the marker is pressed, it will display potential donor information based on blood group. The parameters needed to use this feature are the minimum number of each blood group, the date of the event, and the time span of the data to be processed.

This research uses a dataset containing data from donors who routinely donate blood to Red Cross Indonesia (Fig. 2). The dataset used as training data is a dataset containing donor history data at a location. There are 9 attributes owned by the dataset as parameters for running SVM. The first attribute, “id_user” is the user id number registered in the database. The second attribute, “gol_blood” is the blood group owned by the user. The third attribute, “first” is the period of time for the donor to donate blood for the first time (in months). The fourth attribute, “last” is the distance from the last time the donor donated blood. “Attribute freq” is the number of donors ever done by the donor. The “freq_loc” attribute is the number of donors that have been done by the donor at the location. The “last_loc” attribute is the time interval since the last time the donor made a blood donation at the location. The attribute "cc" is how much blood has been donated.

	id_user	gol_darah	first	last	freq	freq_loc	last_loc	cc
0	1	A	31	3	14	14	3	4900
1	2	B	28	4	12	12	4	4200
2	3	AB	65	1	16	16	1	5600
3	4	O	10	2	4	4	2	1400
4	5	A	64	4	20	19	4	7000
5	6	B	32	2	10	9	2	3500
6	7	AB	36	4	8	8	4	2800
7	8	O	28	2	13	10	2	4550
8	9	A	31	1	10	8	1	3500
9	10	B	42	3	13	11	3	4550

Figure 2. Example Training Dataset

This dataset will run on the SVM. The dataset shows attribute values that indicate that the donor routinely donates blood. For testing the routine donor dataset is divided into two parts, data testing for donor data who routinely conduct blood donation activities at that location and the dataset for donors who rarely donate blood at that location.

The first research test was carried out by entering data on blood donors who donate to a location on a regularly. This is done with the aim of knowing whether the application made can recognize regular donors as potential donors. The "target" column shows whether the user is potential or not. The value "1" in the target column indicates potential donors and "0" for non-potential.

The results of the prediction (Fig. 3) show that almost all donors are potential donors. Non-potential donors were obtained because the last donor was not within 2 months. So, it can be concluded that the predictions obtained can identify routine donors (potential donors) as well as non-potential donors (because the distance between the last donors has not been within 2 months to be able to donate blood again).

The second research test was carried out by entering data on blood donors who routinely donate blood but rarely donate at that location. This is done with the aim of knowing whether the The prediction results of donors that routinely donors as non-potential donors for blood donation activities carried out at certain locations.

	id_user	gol_darah	first	last	freq	freq_loc	last_loc	cc	target
0	1	A	31	3	14	14	3	4900	1
1	2	B	28	4	12	12	4	4200	1
2	3	AB	65	1	16	16	1	5600	0
3	4	O	10	2	4	4	2	1400	1
4	5	A	64	4	20	19	4	7000	1
5	6	B	32	2	10	9	2	3500	1
6	7	AB	36	4	8	8	4	2800	1
7	8	O	28	2	13	10	2	4550	1
8	9	A	31	1	10	8	1	3500	0
9	10	B	42	3	13	11	3	4550	1

Figure 3. The prediction results of donors that routinely donate their blood at a certain location

Prediction results (Fig. 4) show that the application can recognize donors who rarely donate blood in that location. Where the prediction results show that most of the datasets are not potential donors. This is because the frequency of donors who donate at that location is relatively small compared to the total frequency of donors who have been made by donors. This indicates that the donor frequency parameter at the site also affects the prediction results. Even though a donor regularly donates blood, if the donor does not donate blood at that location, the donor will not be a potential donor at that location. So it can be concluded that the prediction can recognize routine donors but rarely donate blood at certain locations as non-potential donors.

	id_user	gol_darah	first	last	freq	freq_loc	last_loc	cc	Predict
0	1	A	49	4	15	3	6	5250	0
1	2	B	22	0	11	1	10	3850	0
2	3	AB	59	3	14	3	16	4900	0
3	4	O	47	3	11	0	14	3850	0
4	5	A	40	2	19	2	38	6650	0
5	6	B	17	1	4	1	4	1400	1
6	7	AB	46	4	14	3	24	4900	0
7	8	O	35	1	17	1	10	5950	0
8	9	A	32	2	15	3	4	5250	0
9	10	B	2	2	1	0	2	350	0
10	11	AB	15	3	6	0	60	2100	0
11	12	O	13	4	3	0	60	1050	0
12	13	A	30	2	14	1	10	4900	0
13	14	B	17	1	8	2	10	2800	0
14	15	AB	6	0	3	0	60	1050	0
15	16	O	19	4	5	0	60	1750	0
16	17	A	52	4	16	2	14	5600	0
17	18	B	35	3	8	2	2	2800	0
18	19	AB	49	1	16	1	10	5600	0
19	20	O	12	2	5	1	6	1750	0

Figure 4. The prediction results of donors that routinely donate their blood on but not in that location

The third research test was carried out by entering data on blood donors who did not routinely donate blood. This is done with the aim of knowing whether the application made can recognize non-routine donors as non-potential donors.

Prediction results (Fig. 5) show that the application can recognize that all of these donors are not potential donors. This proves that the time and frequency parameters affect the results of the predictions.

	id_user	gol_darah	first	last	freq	freq_loc	last_loc	cc	Predict
0	1	A	31	5	3	0	60	1050	0
1	2	B	60	8	5	1	36	1750	0
2	3	AB	13	5	1	0	60	350	0
3	4	O	51	10	4	0	60	1400	0
4	5	A	42	10	4	1	32	1400	0
5	6	B	51	12	5	1	41	1750	0
6	7	AB	59	8	5	1	24	1750	0
7	8	O	23	12	2	0	60	700	0
8	9	A	42	12	5	1	17	1750	0
9	10	B	32	11	3	0	60	1050	0
10	11	AB	24	10	3	0	60	1050	0
11	12	O	16	11	2	0	60	700	0
12	13	A	46	7	4	1	46	1400	0
13	14	B	13	11	1	0	60	350	0
14	15	AB	14	11	2	0	60	700	0
15	16	O	16	9	2	0	60	700	0
16	17	A	16	4	1	0	60	350	0
17	18	B	31	6	4	1	8	1400	0
18	19	AB	56	6	5	1	34	1750	0
19	20	O	38	4	3	0	60	1050	0

Figure 5. The prediction results of donors that non routinely donate their blood

From the test results, it can be concluded that the parameters entered in the SVM can produce predictive results obtained capable of classifying potential/non-potential donors in a blood donor activity carried out at a certain location at a certain time. The next test is to test the predicted results obtained compared to donors who come and donate blood to activities carried out at certain locations at certain times.

Test the prediction results, it is done by using precision, recall, and F1 score formulas. The results obtained can be seen in Table 1. It can be seen that the acquisition of precision was 91%, the recall was 83%, and the F1 score was 85%. So, it can be concluded that the prediction results obtained are quite feasible to use.

TABLE I. PREDICTION TESTING

Classification	Value
Precision	91%
Recall	83%
F1-Score	85%

From the test results, it was found that around 85% of the predicted potential donors came and made a donation. With these results, it can be concluded that the prediction process is quite accurate.

IV. CONCLUSION

Blood is very important for human life. Until now technology has not been found to make human blood. So, if someone lacks blood then that person needs blood transfusions obtained from other people. Because of this need, the Red Cross carries out blood donation activities to obtain blood supplies from donors. We need a system that can predict how much blood can be obtained if the Red Cross carries out blood donation activities at a location at a certain time. So that the Red Cross can maximize the amount of blood supply that can be obtained at each blood donation activity.

This study will use the Support Vector Machine method to classify blood donors into potential donors and non-potential donors to donate blood at blood donation activities at certain locations at certain times. The classification of potential/non-potential donors aims to predict how many donors will attend and donate blood if the Red Cross carries out a blood donation activity at a certain location at a certain time.

From the test results, it can be concluded that the parameters entered in the SVM can produce predictive results obtained capable of classifying potential/non-potential donors in a blood donor activity carried out at a certain location at a certain time. From the test results, it was found that around 85% of the predicted potential donors came and made a donation. With these results, it can be concluded that the classification and prediction processes are quite accurate.

Improvements in future research can be done by adding other parameters to increase the accuracy of the predictions obtained. For example, using the classification of active/non-active donors or ranking of donors so that it can be seen how likely it is that an active donor will attend to donate blood. Another improvement can also be to enter a certain time period parameter to determine the activity of a donor based on the history of donors that have been carried out in soon (e.g. the last 2-3 years). So, if a donor is not actively donating blood within that period, the donor can be categorized as a non-potential donor.

ACKNOWLEDGMENT

This research was supported by the Indonesia Directorate of Research and Technology of Higher Education and Petra Christian University Supply Chain Research Group. We also thank UTD PMI Surabaya staff for valuable data, discussion, and feedback.

REFERENCES

[1] Alodokter, “Blood Transfusion, Here's What You Should Know - Transfusi Darah, Ini yang Harus Anda Ketahui,” 2018, Retrieved May 17, 2021, from https://www.alodokter.com/transfusi-darah-ini-yang-harus-anda-ketahui

[2] Kementrian Kesehatan Republik Indonesia, “Indonesia Needs 5.1 Million of Blood Bags per Year,” 2017, Retrieved June, 11, 2021, from http://www.depkes.go.id/article/view/17071100002/indonesia-butuh-darah-5-1-juta-kantong-pertahun.html.

[3] Jakarta Globe,” Indonesia Faces Worrying Lack of Blood Donors,” 2017, Retrieved June, 11, 2021, from https://jakartaglobe.id/context/indonesia-faces-worrying-lack-blood-donors.

[4] World Health Organization, “Improving access to safe blood products through local production and technology transfer in blood establishments,” WHO Library Cataloguing-in-Publication Data, 2016

[5] Adi, P. “Getting Healthier with Blood Donation - Directorate General of State Wealth” - Makin Sehat dengan Donor Darah - Direktorat Jenderal Kekayaan Negara, 2016, Retrieved May 17, 2021, from https://www.djkn.kemenkeu.go.id/2013/kilasperistiwa/makin-sehat-dengan-donor-darahabout. (n.d.).

[6] Ministry of Health of the Republic of Indonesia,”Indonesia Health Profile Data and Information 2018,” Ministry of Health of the Republic of Indonesia, 2019

[7] Andrew, A. M. 2001. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Kybernetes. https://doi.org/10.1108/k.2001.30.1.103.6

[8] Selvaraj, P., Sarin, A., & Seraphim, B.I, “Forecasting System for Donation of Blood Using SVM Model,” International Journal for Research in Applied Science & Engineering Technology, 2022, vol 10 issue V

[9] Nugroho, E. B., Furqon, M. T., & Hidayat, N, “Classification of Blood Donors Using the Support Vector Machine Method, Journal of Information Technology and Computer Science Development” - Klasifikasi Pendonor Darah Menggunakan Metode Support Vector Machine, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(10), 2018, pp. 3860–3864.

[10] Nugroho, A. S., Witarto, A. B., & Handoko, D., “Support Vector Machine, Theory and Its Application in Bioinformatics,” Wiley Interdisciplinary Reviews: Computational Statistics, 2011. https://doi.org/10.1002/wics.149