Classification of Potential Blood Donors Using Artificial Neural Networks and Alternating Least Squares

Andreas Handojo
Informatics Department
Petra Christian University
Surabaya, Indonesia
handojo@petra.ac.id

Tanti Octavia
Industrial Engineering Department
Engineer Profesion Education Department
Petra Christian University
Surabaya, Indonesia

Widya Arditanti
Informatics Department
Petra Christian University
Surabaya, Indonesia

Abstract— **Blood has an important role in human life. Lack of blood can cause illness and even death. To meet the needs of human blood, it requires blood donors from other people. The problem that occurs is that everyone can only donate blood once every three months. The problem is that not all blood donors routinely donate blood every three months. Therefore, it is important for the Red Cross to be able to classify active and passive donors, especially when there is an urgent need for blood. A method is needed that can be used to provide recommendations for active donors who have a high probability of donating blood so that they can do so immediately. So that the Red Cross can conduct the selection of potential donors more effectively and so that there are not many expired blood stocks. In this study, the classification of potential donors was carried out using an Artificial Neural Network and Alternating Least Squares. The results from the list of potential donor recommendations are then tested using the Mean Reciprocal Rank. The test results show a result of 0.0052186103057361, with an average time of 127. With the Artificial Neural Network model that uses two dense layers of 1024 nodes, 10 epochs, and a sigmoid as the activation function, the training accuracy is 98.90% and the testing accuracy is 99.20%.**

Keywords—blood donors, classification, artificial neural network, alternating least squares, mean reciprocal rank

I. INTRODUCTION

As population growth increases, the need for blood bags increases. According to World Health Organization (WHO) standards, Indonesia needs 5.1 million blood bags per year, which is 2% of Indonesia's total population of approximately 270 million people [1]. Meanwhile, Indonesia is only able to obtain 4.1 million blood bags per year [2]. So that the available blood bags are only about 16 blood bags per year per 1000 people. This requires special attention considering the need for data bags, which often appear suddenly, such as in the event of an accident, childbirth, surgery, or natural disaster. These conditions can occur suddenly and require the availability of sufficient blood bags to meet the patient's urgent needs.

Until now, there has been no adequate technology to replace human blood. Therefore, the need for blood is obtained through blood donor activities. Blood donation is the process of voluntarily giving blood for the purposes of blood transfusions for other people in need [3]. The organization that acts as the organizer of blood donation activities is the Red Cross. People may donate blood approximately once every three months, with a maximum of five donors in one year [4]. Not all blood donors routinely donate every three months. Therefore, it is important for the Red Cross to be able to classify active and passive donors. Especially if there is an emergency case that requires blood donors quickly. With this donor classification, it can also be useful for the Red Cross to invite certain donors to carry out blood donation activities when supplies of blood bags are running low. Considering that blood supplies are often unstable because the number of donors is uncertain, blood donation is a voluntary activity [5].

In general, the Red Cross consistently sends invites through SMS, email, WhatsApp, etc. to all blood donors to invite donors who have previously donated their blood. Messages are delivered every 60 days after a blood donor's donation. The problem is that there is still no donor classification system to find suitable potential donors to meet the demand for blood at hand, which makes it difficult for the Red Cross to send the invitation to the right donors that are needed. This is very crucial, especially when there is an emergency. For example, based on blood type, rhesus, age, and other factors. The right invitation given to the right donor at the right time is very important, considering that if a certain type of blood stock is excessive at a certain time, then that blood supply will be wasted. This happens because the blood bag stock has an expiration date. The storage time for this type of blood bag varies greatly, from 5 days to 35 days.

There have been several studies conducted to classify blood donors, such as using the neural network and decision tree methods [6], the k-nearest neighbors and logistic regression methods [7], the Naive Bayes method, the k-nearest neighbors, and the decision tree C4.5 [8]. From these researches, it was found that the neural network algorithm has better accuracy, precision, and recall values for carrying out the classification process compared to other algorithms. However, this study only classifies donors based on a survey or physical condition, such as age, weight, hemoglobin, blood pressure, and others. There is still no research that focuses on efforts to classify donors with the aim of selecting potential active donors to be able to donate blood at a certain time.

To answer this problem, this research was carried out with the aim of classifying donors (active or inactive) using the Artificial Neural Network (ANN) method. Then the list of classifying donors is ranked in order to obtain the best donor recommendation order using the Alternating Least Squares (ALS) method. The results will be tested using the Mean Reciprocal Rank (MRR) method. In this way, the Red Cross can contact a number of active donors according to the needs that occur more effectively. This will be able to reduce the amount of expired blood.

II. Literature Review

There are several researches that try to look for potential blood donors using different methods, such as Atmaja et al. [5], who conducted a research to predict whether a donor has the potential to become a permanent donor by using the physical data of the donor. In this research, the method used is the C4.5 algorithm, using a Java-based open-source data mining application. The result of this research is that from the decision tree obtained from the data mining process, it can be concluded that private employees over the age of 26 are the most generous donors.

Prabeja et al. [7] conducted research to identify traits in order of importance that influence a person's decision to donate blood through a survey in India. This research classifies students in India as both donors and non-donors. This research compares the k-nearest neighbor and logistic regression methods. The result of this study is the classification of students in India as donors and non-donors by comparing k-nearest neighbors and logistic regression. From these two methods, it is found that the k-nearest neighbor produces better results than logistic regression, with an F1-Score of 75.61%. The factors that influence the decision of students in India to donate are the time it takes to donate blood, the intimacy of friends or family members when donating blood, the availability of a familiar place for blood donation, and the feeling of being valued.

Wahono & Riana [8] conduct research to compare the naive Bayes algorithm, k-nearest neighbors, and decision tree C.45 to predict whether someone will donate blood based on the donor's physical condition, such as gender, amount of hemoglobin, blood pressure, and others. The methods used in this research are the naive Bayes classifier (NBC) algorithm, k-nearest neighbors, and the C.45 decision tree. The results of this study show that the decision tree algorithm C4.5 obtained the highest accuracy value of 93.83%, better than NBC (accuracy value of 85.15%) and KNN (accuracy value of 84.10%). The decision tree algorithm C4.5 also has the highest Area Under Curve value (0.987), followed by NBC (0.927) and KNN (0.816).

Boonyanusith & Jittamai [6] conducted research to find out the factors that influence potential donors to donate blood using survey data. The methods used in this research are an artificial neural network and a decision tree. The result of this research is that using ANN is better than using the decision tree, where the ANN model can classify donors into the blood donor group with precision and recall of 81.7% and 88% and classify donors into the non-donor group with precision and recall of 53.8% and 41.6%.

Handojo et al. [9] try to predict potential blood donors who can attend blood donation activities using a support vector machine to give donors at a certain location and at a certain time with an F1 Score of 85%.

From the research that has been conducted, it appears that there has been no research specifically aimed at classifying potential donors based on current needs. For example, specificity of the type of blood needed (blood type and rhesus), gender of the donors, when donors last donated, location of the donor, and so on. This will be very useful for the Red Cross in estimating the right number of donors to invite to meet the specific blood needs that occur at that time. For example, if you need 100 bags of blood with type A and 80 bags with type B, the Red Cross can calculate how many invitations are distributed, and the invitations will be given to anyone to meet those needs. So, for example, to meet the need for blood type A, donor invitations were distributed to 110 potential donors from blood group A who met the predetermined criteria for potential donors (out of a total of 300 donors who had blood type A). In this way, the amount of blood obtained will not be excessive.

III. Research Method

The process of this research is divided into three main parts (Fig.1): the process of classifying the types of donors, the process of sorting donors based on the priorities set, and the process of broadcasting invitations to priority donors. The classification of donor types is carried out using the Artificial Neural Network (ANN) method. Classification is done using historical data on blood donor transactions obtained from the Red Cross. Then, a list of donor recommendations was formed in the form of a ranking order using the Alternating Least Square (ALS) algorithm. The list of potential donor recommendations will be used by the Red Cross to determine potential donors who will be sent invitations to donate blood.

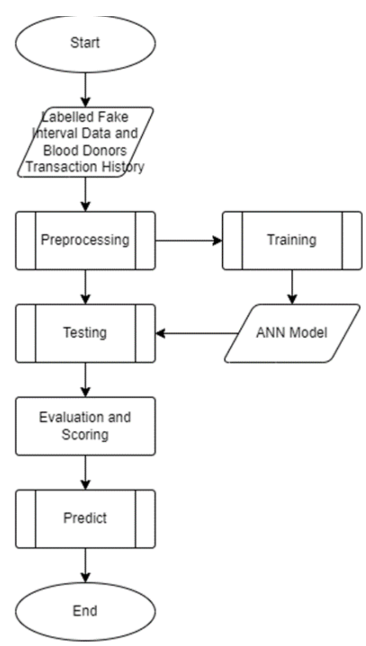

Fig. 1. Donors Classification using ANN

In the preprocessing process, dummy interval data is labeled as input for this process. The data is divided into two groups randomly, where 80% is used as training data and the remaining 20% is used as testing data. The results of the division will be returned to the initial process as input for the next process. The training process begins by using training data as input. The data is divided into two parts, where the first part is the parameter and the second part is the label. Parameters are the columns that determine the labeling, while the label is the classification or class value of the existing parameter values. After the parameters and labels are defined, the data will be used as training data for the creation of the ANN model.

The model generated from this training process consists of 60 nodes on the input layer and 5 nodes on the output layer. The 60 nodes in the input layer represent the donor interval for 60 months or 5 years of the observation period (the column is 1 for donors and 0 for non-donors), while the 5 nodes in the output layer represent 5 categories that will be used to classify donor types. In the output layer, five columns from each category of donor type will produce the percentage of classification weight. Each weight percentage will be averaged and normalized at a later stage.

The testing process is used to see how well the ANN model that was created performs using training data. Which is obtained from the output of the preprocessing process, and the ANN model, which is obtained from the output of the training process. Existing testing data will be divided into two categories: parameters and labels. Parameters will be predicted using the existing ANN model, and the prediction results for the classification of donor types will be formed. The prediction results are then calculated for accuracy and compared to the actual data or labels that have been formed previously.

The prediction process is used to generate predictions of donor class and type based on donor transaction history data. This process begins by using the ANN model that was generated in the previous process and blood donor transaction history data. This data is then changed to fit the form of the input model so that the data can be predicted using the ANN model that has been formed and produces classified donors. The input for using this model is blood donor transaction history data that has gone through the transformation stage into 60 columns that describe the donor interval for 60 months, or 5 years of observation. The output of this model is in the form of five columns that describe the percentage for each category of donors, and then each value will be averaged and normalized to get the final donor category results.

The ranking list for donor recommendations is generated using the ALS algorithm. Parameter rating for data from the classification of donor types from ANN The data will go through the preprocessing stage and then enter the training stage, where the ALS model will be formed. After the ALS is implemented and the model is formed, the model will be evaluated using the Mean Reciprocal Rank (MRR) method by comparing the recommendation results with donor transaction data the following month.

IV. Implementation and Testing

The implementation of the potential donor classification process is carried out using six years of data (2017–2022) with a total of 330342 (Table 1). The data is then divided into 80% training data and 20% testing data. This data was obtained from the Surabaya Red Cross.

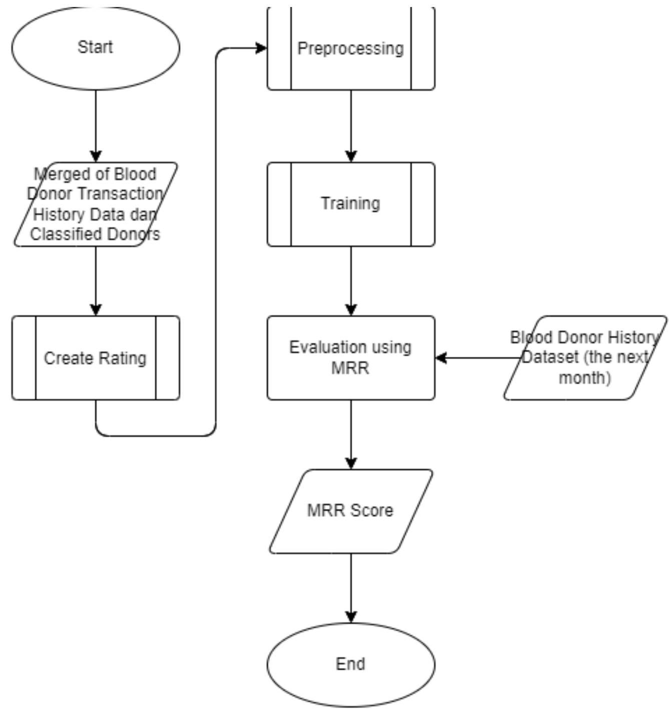

Fig. 2. Donors Recommendation

TABLE I. Blood Donors Data Training and Testing

Years	Number of Data
2017	55430
2018	61638
2019	77904
2020	65470
2021	65150
2022	4750
Total Data	330342
Training Data	264274
Testing Data	66068

Another test is also done by looking at the MRR value of ALS. To find the best model, testing the classification of donor types is carried out, starting with testing the layer configuration, epoch, and activation function. The test begins with finding the layer configuration that can produce the highest accuracy value with other patented parameters (Table 2). The patent parameters used for this test are the use of dense layers, the number of epochs of 10, and the sigmoid activation function. After testing the layer configuration, a dense layer consisting of 2 layers with

1024 nodes for both layers, produces the highest accuracy value compared to other layer configurations.

Donor prediction is done using the ANN algorithm. Prediction model testing is done by looking for layer configurations to get the best accuracy value. Based on the tests that have been carried out, the highest accuracy value for training and testing models is achieved using two layers, consisting of 32 nodes for the first layer and 128 nodes for the second layer. The accuracy obtained from the model is 74.69% for training and 75.57% for testing. The time it takes to predict is 12 seconds. Therefore, this model is used in donor prediction.

TABLE II. LAYER CONFIGURATION TESTING

Number of Layer	Layer Configuration	Accuracy Training	Accuracy Testing	Time Elapse (s)
1	Dense (128)	75.80%	71.20%	3
	Dense (512)	87.70%	83.20%	2
2	Dense (64) Dense (128)	77.30%	76.00%	3
	Dense (128) Dense (128)	81.10%	81.20%	2
	Dense (512) Dense (128)	92.10%	93.20%	2
	Dense (1024) Dense (128)	95.40%	94.80%	2
	Dense (512) Dense (512)	96.50%	96.40%	2
	Dense (1024) Dense (1024)	**98.70%**	**99.20%**	**3**
3	Dense (128) Dense (512) Dense (1024)	88.80%	89.60%	2

The next test is to find the number of epochs that produce the best accuracy value. The layer configuration used for this test is constant, in accordance with the previous best test results. Meanwhile, the activation function constant used is sigmoid. After testing the number of epochs, using 10 epochs produced the highest accuracy value compared to other epochs (Table 3 and Fig.3).

TABLE III. NUMBER OF EPOCH TESTING

Layer Configuration	Epoch	Training Accuracy	Testing Accuracy	Time Elapse (s)
Dense (1024)	5	73.20%	85.60%	5
	10	**98.90%**	**99.20%**	**3**

The next test is the activation function. This test is carried out with the aim of finding the activation function that can produce the best accuracy value. Layer configuration parameters and the number of epochs used were obtained from the previous test. After testing the activation function, using sigmoid as the activation function produced the highest accuracy value compared to other layer configurations (Table 4).

TABLE IV. ACTIVATION FUNCTION TESTING

Layer Configuration	Epoch	Activation Function	Training Accuracy	Testing Accuracy	Time Elapse (s)
Dense (1024)	10	**Sigmoid**	**98.90%**	**99.20%**	**3**
		ReLU	74.70%	75.20%	5
		Hard Sigmoid	99.95%	97.60%	3

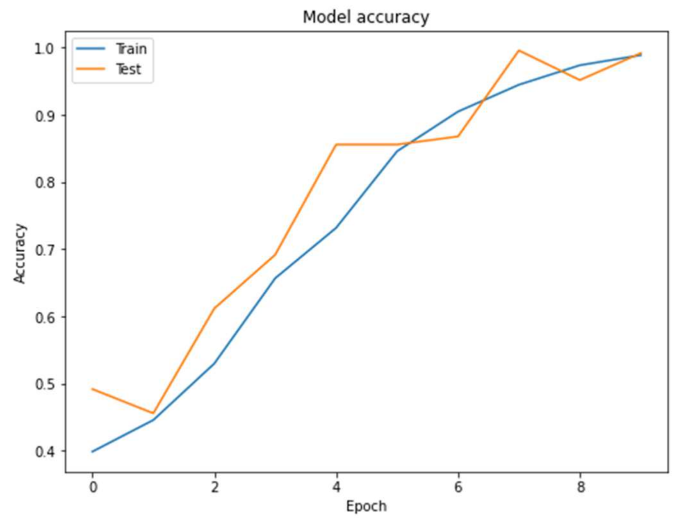

Fig. 3. Test Results for Number of Configuration Epoch 1024

TABLE V. MRR TESTING FOR DONOR RECOMMENDATION TEST RESULT

Testing Number	MRR	Time Elapse (s)
1	0.003860809431240	174
2	0.008562887244257	137
3	0.004338391625711	82
4	0.003679190171968	102
5	0.006838957392494	119
6	0.003551439900427	109
7	0.004762345886123	154
8	0.004570983992334	154
9	0.003549190171298	102
10	0.008471907241509	137
Average	0.005218610305736	127

The average MRR value obtained from ten experiments on the outcomes of active donor recommendations using MRR was

0.0052186103057361, with an average time of 127 seconds (Table 5). With the highest MRR value of 0.008562887244257 (137 seconds) and the lowest MRR of 0.003549190171298 (102 seconds).

V. Conclusion

Supplies of blood bags are very valuable for people in need. If stored for a long time, the supply of blood bags will expire, considering that stored blood has an expiration date of five to 35 days. Meanwhile, the supply of blood from donors is uncertain because not all donors donate blood regularly. The Red Cross needs to classify active and inactive donors in order to make the invitations given to donors to donate blood more efficient. This is very important, in addition to avoiding excessive stocks. It is also important if there is an urgent need. This study classified donors as active or inactive using the ANN method. The results of the classification are then ranked using the ALS method. With these results, it is hoped that the Red Cross will be able to provide invitations more precisely and effectively. From the test results using MRR, the average MRR value is 0.0052186103057361, with an average time of 127 seconds. With the ANN model that uses two dense layers of 1024 nodes, 10 epochs, and a sigmoid as an activation function. This model produces a training accuracy of 98.90% and a testing accuracy of 99.20%.

Acknowledgment

This research was supported by Petra Christian University Research Institutions and Community Service, especially Supply Chain Research Group. We also thank Red Cross Surabaya staff for valuable data, discussion, and feedback.

References

[1] Ministry of Health of the Republic of Indonesia, "Indonesia Health Profile Data and Information 2018, "in Ministry of Health of the Republic of Indonesia (2019).

[2] Ministry of Health of the Republic of Indonesia, Indonesia Needs 5.1 Million of Blood Bags per Year, 2017, in http://www.depkes.go.id/article/view/17071100002/indonesia-butuh-darah-5-1-juta-kantong-pertahun.html, accessed on 21 January 2023

[3] C.A. Sugianto and T.M. Zundi, Mobile-Based Blood Donor Application Design at PMI Bandung Regency [Rancang Bangun Aplikasi Donor Darah Berbasis Mobile di PMI Kabupaten Bandung], KOPERTIP: Scientific Journal of Informatics and Computer Management, vol. 1, no. 1, 2017, 11-18. doi:10.32485/kopertip.v1i15

[4] F. Nurin, Not only benefits, blood donation can also bring side effects if it's frequent [Tidak Cuma Manfaat, Donor Darah Juga Bisa Membawa Efek Samping Jika Keseringan], hellosehat, 2021. In https://hellosehat.com/kelainan-darah/efek- samping-donor-darah, accessed on 21 January 2023

[5] K.J. Atmaja, I.B.G. Anandita, and N.K.C. Dewi, Application of Data Mining to Predict Potential Blood Donors to Become Permanent Donors Using Decision Tree Method C.45. S@CIES Journal, vol. 7, no. 2, pp. 101-108, 2017. doi:10.31598/sacies.v7i2.284, 2017

[6] W. Boonyanusith, and P. Jittamai, Blood Donor Classification Using Neural Network and Decision Tree Techniques. Proceedings of the World Congress on Engineering and Computer Science, 2012

[7] Pabreja, K., & Bhasin, A. (2021). A Predictive Analytics Framework for Blood Donor Classification. International Journal of Big Data and Analytics in Healthcare, 6(2), 1–14. doi:10.4018/ijbdah.20210701.oa1

[8] H. Wahono and D. Riana, Predicting Potential Blood Donor Candidates Using Naïve Bayes Algorithm, K-Nearest Neighbors and C4.5 Decision Tree [Prediksi Calon Pendonor Darah Potensial Dengan Algoritma Naïve Bayes, K-Nearest Neighbors dan Decision Tree C4.5], JURIKOM (Computer Research Journal),vol 7, no. 7, 2020. doi:10.30865/jurikom.v7i1.1953.

[9] A. Handojo, A. N. Purbowo, G. B. Karaeng and T. Octavia, Predicting Potential Blood Donors Who Can Attend Blood Donation Activities using a Support Vector Machine, International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Indonesia, 2022, pp. 351-355, doi: 10.1109/ISRITI56927.2022.10053032.