

Komparasi Algoritma Neural Network, K-Nearest Neighbor Dan Naive Baiyes Untuk Memprediksi Pendoror Darah Potensial

Wahyu Eko Susanto¹⁾ Dwiza Riana²⁾

¹⁾Manajemen Informatika

Akademik Manajemen Informatika dan Komputer BSI Yogyakarta (AMIK BSI Yogyakarta)
wahyueko851@gmail.com

²⁾Program Pascasarjana Magister Ilmu Komputer

Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri)
dwiza@bsi.ac.id

Abstract – To be able to maintain a minimum stock of blood transfusion, donate their blood, potential donor returned must be known, since the blood results transfusion can no longer be used after 42 days. During this time in predicting potential donors donate their blood again produces different accuracy on some algorithms of classification by using dataset are different. So it is not yet known where the dataset with algorithm suitable for predictions. Need to find out and distinguish between potential blood donors donate their blood again and what not, need to be built so that the blood donor unit can take the decision to keep the blood stock to keep it secure. In this study performed comparisons of Neural Network Algorithm, K-Nearest Neighbor and Naïve Bayes data is applied to the donors and blood donor transaction data with RFMTC dataset and PMI dataset. From the test results by measuring the performance of these three algorithms when applied to both of the dataset test using the Confusion Matrix and ROC Curves, it is known that neural network algorithm with dataset RFMTC has the value of the highest accuracy. So obtained the use if neural network algorithm with dataset RFMTC which fits in this research to applied to the prediction of potential blood donors.

Keyword: Blood Donors, Neural Network, Naïve Bayes, K-NN, RFMTC.

1. PENDAHULUAN

Pemenuhan kebutuhan darah sangat penting untuk meningkatkan kualitas pelayanan kesehatan dan menyelamatkan nyawa seseorang (Kementerian Kesehatan Republik Indonesia, 2014). Idealnya ketersediaan darah yang diperuntukkan untuk donor adalah 2,5% dari Jumlah Penduduk. Indonesia pada tahun 2013 terdapat kekurangan sebanyak 2.476.389 kantong darah (Pusat Data dan informasi Kementerian Kesehatan RI, 2014)

Mengingat darah adalah salah satu media yang digunakan untuk penularan berbagai penyakit, maka sebaiknya darah transfusi berasal dari donor resiko rendah. Menurut rekomendasi *World Health Organization* (WHO) darah beresiko rendah yang aman dan berkualitas adalah darah yang berasal dari pendonor sukarela, daripada berasal dari donor pengganti (Kementerian Kesehatan Republik Indonesia, 2014). Donor sukarela adalah pendonor yang memberikan darahnya dimana darah tersebut tidak diperuntukkan bagi orang tertentu sebagai pengguna darah yang didonorkan sedangkan yang maksud dengan donor pengganti ialah pendonor yang memberikan darahnya ditujukan untuk orang tertentu yang telah ditunjuk oleh pendonor (*American Cancer Society*, 2014). Namun minimnya kesadaran masyarakat untuk menjadi pendonor sukarela hal ini menjadi salah satu faktor yang mengakibatkan persediaan darah di

Unit Donor Darah (UDD) menjadi rendah (Pusat Data dan informasi Kementerian Kesehatan RI, 2014).

Dilain sisi menurut *American Red Cross*, darah hasil donor tidak lagi dapat digunakan setelah 42 hari. Satu-satunya cara untuk memenuhi permintaan adalah dengan memiliki sumbangan rutin dari sukarelawan yang sehat (Darwiche, Feuilloy, Bousaleh, & Schang, 2010). Untuk menjawab tantangan tersebut beberapa penelitian telah dilakukan untuk menentukan pendonor darah yang potensial mendonorkan darahnya kembali.

Penelitian mengenai donor darah potensial antara lain diteliti oleh Mahmud, Soekotjo dan Santoso (2014) menggunakan algoritma C4.5 dan Fuzy Tahani dengan menggunakan variabel golongan darah, jenis kelamin, usia, berat badan, tekanan darah, kadar HB (*hemoglobin*), dan kadar HTC (*hematocrit*) adapun target variabel outputnya adalah atribut Donor Lagi yang berisikan nilai 'ya' atau 'tidak' berdasarkan nilai ambang batas jumlah donor, bernilai 'ya' apabila jumlah donor ≥ 2 dan bernilai 'tidak' apabila jumlah donor < 2 . Pada penelitian ini menghasilkan akurasi rata-rata sebesar 89,64% dengan menggunakan dataset yang berasal dari PMI Kota Malang.

Penelitian lainnya juga dilakukan oleh Darwiche, Feuilloy, Bousaleh dan Schang (2010) pada penelitian ini mereka menggunakan algoritma MLP dan SVM dan metode yang

digunakan adalah RFMTC (*Recency, Frequency, Monetary, Time, Churn Probability*) yang dikembangkan oleh I-Cheng et al merupakan pengembangan metode RFM (*Recency, Frequency, Monetary*). Dataset yang digunakan berasal dari *UCI Machine Learning* adapun variabel yang digunakan adalah kebaruan (*Recency*), frekwensi (*Frequency*), jumlah darah yang diambil (*Monetary*), lama menjadi pendonor (*Time*), dan variabel targetnya adalah menyumbang atau tidak menyumbang (*Donation or non donation*) berdasarkan apakah pada bulan maret 2007 pendonor tersebut melakukan donor atau tidak. Dari penelitian ini dihasilkan *sensitivity* 65.8% dan *specificity* 78.2% dari 600 data *training* dan 148 data *testing*.

Fais A, Aditya D, Mulya I, Ramadien dan Sani (2015) melakukan penelitian klasifikasi pendonor menggunakan metode *naïve bayies clasifier*. Adapun variabel yang digunakan dalam penelitian ini adalah *hemoglobin* (HB), tensi atas, tensi bawah, berat badan, umur, jenis kelamin, riwayat penyakit menular, interval donor serta mempunyai variabel target pendonor dan non pendonor. Dalam penelitian ini digunakan dataset sebanyak 400 *record* yang dibagi menjadi 350 *record* sebagai data *training* dan 50 *record* sebagai data *testing*. Hasil yang didapatkan dalam penelitian ini adalah keakuratan rata-rata yang didapatkan dari pengujian ini sebesar 74% dari 4 percobaan pengujian.

Dengan melihat perbedaan akurasi dari penelitian terdahulu maka untuk menjawab permasalahan dalam penelitian ini dicariilah dataset dan algoritma yang tepat dalam menentukan pendonor darah potensial di UDD PMI Kabupaten Bantul dengan cara membandingkan penerapan dataset RFMTC dan penerapan dataset donor darah dari PMI, pada Algoritma *Neural Network*, *K-Nearest Neighbor* dan *Naïve Baiyes*. agar di ketahui tingkat akurasi yang paling tinggi pada masing-masing dataset dan masing-masing algoritma dalam menentukan pendonor darah yang potensial mendonorkan darahnya kembali

2. LANDASAN/KERANGKA PEMIKIRAN

2.1 Donor Darah

Donor darah (*Donor*, 2012) adalah proses pengambilan darah dari seseorang secara sukarela untuk disimpan dibank darah untuk kemudian dipakai pada transfusi darah. Sedangkan Berdasarkan PP NO 7 Tahun 2011 Pendonor Darah adalah orang yang menyumbangkan darah atau komponennya

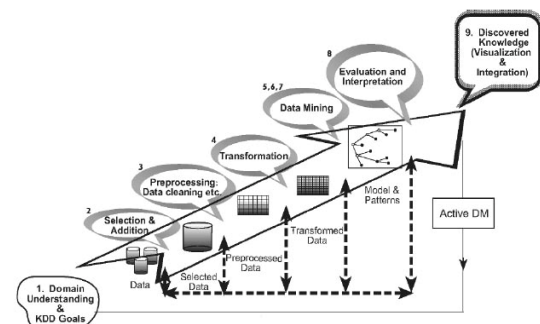
kepada pasien untuk tujuan penyembuhan penyakit dan pemulihan kesehatan.

Adapun syarat yang diatur oleh PMI Kabupaten Bantul agar seseorang dapat mendonorkan darahnya ialah dengan memenuhi ketentuan sebagai berikut (Nurfajar, 2015) :

- Sehat.
- Umur 17 – 60 tahun.
- Berat badan minimal 50 kg.
- Kadar *hemoglobin* 12,5 g/dl – 18 g/dl.
- Tidak berpenyakit yang bisa ditularkan melalui darah.
- Bukan pecandu alkohol dan narkoba.
- Bagi wanita, tidak sedang haid, hamil atau menyusui.
- Jarak terakhir donor 3 bulan

2.2 KDD (*Knowledge Discovery in Databases*)

Knowledge Discovery in Databases (Maimon and Rokach, 2010) adalah proses yang terorganisir untuk mengidentifikasi valid, baru, berguna, dan pola yang dapat dimengerti dari kumpulan data besar dan kompleks. Dimana *Data Mining* (DM) merupakan bagian dari proses *Knowledge Discovery in Database* (KDD). Adapun *Data Mining* sendiri didefinisikan (Witten, 2011) sebagai proses penemuan pola dalam data. Dimana dengan menggunakan data mining ini kita dapat menemukan pola tersebut dengan cara menganalisa data yang sudah ada dalam *database*



Sumber: (Maimon and Rokach, 2010)

Gambar 1. KDD Proses

2.3 K-Nearest Neighbor

Algoritma *Nearest Neighbor* menurut Prasetyo (2010) merupakan Algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) satu data dengan data yang lain. Pengenalan pola wilayah dengan menggunakan algoritma "*k-Nearest Neighbor*" (*k-NN*) merupakan metode klasifikasi, dimana objek baru diberi label berdasarkan objek yang terdekat (Gorunescu, 2011).

Rumus untuk menghitung kedekatan antara objek pada algoritma k-Nearest Neighbor (Kusrini dan Luthfi, 2009) adalah:

$$\text{similarity}(T, S) = \frac{\sum_{i=1}^n f(T_i, S_i) \times w_i}{w_i}$$

Keterangan :

- T : Kasus baru
- S : Kasus yang ada dalam penyimpanan
- n : Jumlah variabel dalam tiap kasus
- i : Variabel individu antara 1 sampai dengan n
- f : Fungsi *similarity* variabel *i* antara kasus T dan S
- w : Bobot yang diberikan pada variabel ke-*i*

2.4 Neural Network

Neural Network menurut Prasetyo(2012) merupakan suatu konsep rekayasa pengetahuan dalam bidang kecerdasan buatan yang di desain dengan mengadopsi system syaraf manusia, yang pemrosesan utamanya adalah otak. Pada Jaringan Saraf Tiruan (JST) pemrosesan informasi mempunyai karakteristik sama dengan jaringan saraf biologis (Siang, 2009)

Terdapat beberapa algoritma dalam *Neural Network* yang dapat digunakan untuk mencari pola dalam proses pelatihannya (Prasetyo, 2012) antarlain *Preceptor*, *Backpropagation*, *Self-Organizing Map (SOM)*, *Delta*, *Associative Memory*, *Learning Vector Quantization* dan sebagainya.

Backpropagation (Prasetyo, 2012) adalah salah satu *Multilayer Perceptron* yang merupakan pengembangan dari *preceptorn* untuk mengatasi kekurangan-kekurangan pada *perceptorn* salah satunya karena ketidak mampuannya untuk bekerja dengan baik pada data yang tidak dapat dipisahkan secara linear.

2.5 Naive Bayes

Naïve Bayesian Classifier (Abidin, 2015) Adalah metode pengklasifikasian yang berdasarkan probabilitas dan Teorema *Bayesian* dengan asumsi bahwa setiap variabel *X* bersifat bebas (*independence*).

Hubungan antara klasifikasi, korelasi hipotesis dan bukti dengan klasifikasi pada *naïve bayes* (Prasetyo, 2012) adalah label kelas yang menjadi target pemetaan dalam klasifikasi *bayes* merupakan hipotesisnya, dan fitur-fitur yang menjadi inputan kelas tersebut adalah buktinya. Adapun formula umum teorima *bayes* adalah sebagai berikut :

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

- P(H|E) Probabilitas akhir bersyarat (*conditional probability*) suatu hipotesis H terjadi jika diberikan bukti (*evidence*) E terjadi.
- P(H) Probabilitas awal (*priori*) hipotesis H terjadi tanpa memandang bukti apapun
- P(E|H) Probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis H
- P(E) Probabilitas awal (*priori*) bukti E terjadi tanpa memandang hipotesis/bukti yang lain

2.5 Variabel Klasifikasi Donor Darah

1. Dataset RFMTC

Variabel yang dimiliki dataset RFMTC (*Recency, Frequency, Monetary value, Time since first purchase and Churn probability*) merupakan modifikasi dari teknik RFM yang digunakan untuk meramalkan perilaku konsumen dimasa yang akan datang, RFM didasarkan pada tiga variabel sederhana, yaitu *Recency of purchase, Frequency of purchase*, dan *Monetary value of purchase*, (Aviliani, Sumarwan, Sugema, dan Saefuddin, 2011)

Y, I-Cheng, Y, King-Jang, and T, Tao-Ming merupakan orang yang melakukan modifikasi sehingga terciptalah RFMTC yang digunakan untu meramalkan perilaku pendonor darah apakah pendonor tersebut termasuk klasifikasi yang potensial mendonorkan darahnya kembali atau tidak, Akurasi dari RFMTC ini menurut *I-Cheng dkk*, lebih tinggi daripada RFM, Adapun penjelasan dari variabel RFMTC tersebut (Darwiche, Feuilloy, Bousaleh, and Schang, 2010) adalah *Recency* Jumlah bulan sejak terakhir menyumbangkan darah. *Frequency* adalah Jumlah berapakai donor, *Monetary* adalah Jumlah darah yang disumbangkan dalam c.c., *Time* yaitu Jumlah bulan sejak pertama menyumbangkan darah, *Donation or non Donation* yaitu Variabel biner yang merepresentasikan apakah pendonor mendonorkan darahnya kembali atau tidak, 1 menyatakan mendonorkan darah, 0 menyatakan tidak menyumbangkan darah.

2. Dataset PMI

Variabel yang terdapat pada dataset PMI yang digunakan dalam penelitian ini adalah variabel yang di catatat oleh petugas dari seorang pendonor ketika melakukan donor darah di UDD PMI Kabupaten bantul, Adapun variabel yang digunakan dalam penelitian ini adalah jenis kelamin, golongan darah, pekerjaan, usia, tensi, berat badan, *hemoglobin*.

Jenis Kelamin atau seks adalah perbedaan fisik yang secara kodrat membedakan laki-laki dan perempuan (Udau, 2013). Selain itu pengertian jenis kelamin (seks) merupakan pembagian dua jenis kelamin yang ditentukan secara biologis, yaitu bahwa pria mempunyai penis (zakar) serta memproduksi sperma, sedangkan wanita memiliki alat reproduksi seperti memiliki rahim. Jenis kelamin secara permanen tidak dapat berubah dan merupakan ketentuan biologis (Sudarma, 2008).

Golongan darah merupakan sebuah pengklasifikasian darah berdasarkan kehadiran atau ketidakhadiran dari substansi antigen yang menempel pada permukaan sel darah merah (Nofiansyah dan Rochmawati, 2014). Dalam proses transfusi darah golongan darah ini diperlukan untuk mencocokkan golongan darah pendonor dan golongan darah dari pasien. Adapun jenis A, B, AB, dan O merupakan jenis penggolongan darah manusia yang dikenal didunia kedokteran (Melati, Passarella, Primartha, dan Murdiansyah, 2011).

Definisi dari Pekerjaan adalah kegiatan yang dilakukan dan mendapatkan upah, Pekerjaan seseorang akan menggambarkan aktivitas dan kesejahteraan ekonomi yang akan didapatkan (Wadud, 2012).

Menurut Chaniago dalam (Widyasih, Mubin, dan Hidyati, 2014) Umur adalah lamanya waktu hidup yaitu terhitung sejak lahir sampai dengan sekarang, Penentuan umur dilakukan dengan menggunakan hitungan tahun. Menurut psikolog dari Universitas Indonesia yang bernama Dra. Ny. Jos Masdani menyatakan bahwa usia lansia merupakan kelanjutan dari usia dewasa. Adapun kedewasaan dapat dibagi menjadi empat bagian, yaitu (Nugroho, 2006):

- 1) Fase Iuventus (20 Tahun – 40 Tahun)
- 2) Fase Verilitas (40 Tahun -50 Tahun)
- 3) Fase Prasenium (55 Tahun – 65 Tahun)
- 4) Fase Senium (Usia 65 Tahun hingga tutup usia)

Tensi atau tekanan darah adalah daya yang diperlukan agar darah dapat mengalir di dalam pembuluh darah dan beredar mencapai semua jaringan tubuh manusia (Moniaga, Pangemanan, dan Rampengan, 2013), Tekanan darah melibatkan dua pengukuran yaitu sistolik dan diastolik, tergantung apakah otot jantung berkontraksi (*sistole*) atau berelaksasi di antara denyut (*diastole*).

Klasifikasi tekanan darah atau tensi untuk usia 18 tahun atau lebih seperti yang diberikan oleh the *Seventh Report of the Joint National Committee on Prevention Detection, Evaluation and Treatment of High Blood Pressure* (JNC 7) di bagi menjadi 4 kategori yaitu (Kementrian

Kesehatan Republik Indonesia, 2014) (1) Normal yaitu Tekanan darah *systolic* < 120 dan *diastolic* < 80, (2) Prehipertensi yaitu Tekanan darah *systolic* 120-139 dan *diastolic* 80-89, (3) Hipertensi Level 1 yaitu Tekanan darah *systolic* 140-159 dan *diastolic* 90-99, (4) Hipertensi level 2 yaitu Tekanan darah *systolic* ≥ 160 dan *diastolic* 100 atau > 100.

Menurut Cipto Surono mendefinisikan berat badan adalah ukuran tubuh dalam sisi beratnya yang ditimbang dalam keadaan berpakaian minimal tanpa perlengkapan apapun. (Dinkes Lumajang, 2013)

Hemoglobin (Hb) adalah protein kompleks yang ada dalam eritrosit mengandung zat besi dan berwarna merah (Fa'rifah dan Purhadi, 2012). Kadar *Hemoglobin* normal dinyatakan pada wanita antara 12,1 sampai 15,1 g/dl dan untuk pria 13,8 sampai 17,2 g/dl (*Low hemoglobin*, 2014).

2.6 Pengujian Metode Klasifikasi Data Mining

Sebagai langkah untuk menguji algoritma yang telah digunakan dalam penelitian ini penulis menggunakan pengujian dengan metode *Confusion Matrix*, dan kurva ROC (*Receiver Operating Characteristic*).

a. Confusion Matrix

Pada Tabel 2.16 di tampilkan tabel *Confusion Matrix* dimana terdapat empat sel yang menyusunnya, dimana keempat sel tersebut diberi label TP, FP, FN, dan TN. Jika pada data hanya terdiri dari dua kelas, maka kelas yang satu dianggap sebagai positif dan yang lainnya negative. (Bramer, 2007).

Untuk menentukan nilai *accuracy* digunakan rumus sebagai berikut :

$$\text{accuracy} = (TP + TN) / (P + N)$$

Keterangan :

TP = Jumlah *true positive*

TN = Jumlah *true negative*

P = *Total Instances positive*

N = *Total Instances negative*

FP = Jumlah *false positive*

FN = Jumlah *false negative*

b. Kurva ROC

Receiver Operating Characteristics (ROC) merupakan teknik untuk memvisualisasikan, mengatur dan memilih pengklasifikasi berdasarkan performa. Pada kurva ROC terdapat dua dimensi kurva, nilai *true positive* (TP) ditempatkan pada sumbu Y dan nilai *false positive* (FP) ditempatkan pada sumbu X (Olson dan Delen, 2008).

AUC pada kurva ROC memiliki tingkat nilai diagnosa yaitu (Gorunescu, 2011):

- 1) Akurasi 0,90 – 1,00 = *excellent classification*
- 2) Akurasi 0,80 – 0,90 = *good classification*
- 3) Akurasi 0,70 – 0,80 = *fair classification*
- 4) Akurasi 0,60 – 0,70 = *poor classification*
- 5) Akurasi 0,50 – 0,60 = *failure*

2.7 Tinjauan Studi

Penelitian mengenai donor darah potensial antara lain diteliti oleh Mahmud, Soekotjo dan Santoso (2014) menggunakan algoritma C4,5 dan Fuzy Tahani dengan menggunakan variabel golongan darah, jenis kelamin, usia, berat badan, tekanan darah, kadar HB (*hemoglobin*), dan kadar HTC (*hematocit*) adapun target variabel outputnya adalah variabel Donor Lagi yang berisikan nilai 'ya' atau 'tidak' berdasarkan nilai ambang batas jumlah donor, bernilai 'ya' apabila jumlah donor ≥ 2 dan bernilai 'tidak' apabila jumlah donor < 2 . Pada penelitian ini menghasilkan akurasi rata-rata sebesar 89,64% dengan menggunakan dataset yang berasal dari PMI Kota Malang.

Penelitian lainnya juga dilakukan oleh Darwiche, Feuilloy, Bousaleh dan Schang (2010) pada penelitian ini mereka menggunakan algoritma MLP dan SVM dan metode yang digunakan adalah RFMTC (*Recency, Frequency, Monetary, Time, Churn Probability*) yang dikembangkan oleh I-Cheng et al merupakan pengembangan metode RFM (*Recency, Frequency, Monetary*). Dataset yang digunakan berasal dari *UCI Machine Learning* adapun variabel yang digunakan adalah kebaruan (*Recency*), frekwensi (*Frequency*), jumlah darah yang diambil (*Monetary*), lama menjadi pedonor (*Time*), dan variabel targetnya adalah menyumbang atau tidak menyumbang (*Donation or non donation*) berdasarkan apakah pada bulan maret 2007 pedonor tersebut melakukan donor atau tidak. Dari penelitian ini dihasilkan *sensitivity* 65,8% dan *specificity* 78,2% dari 148 data *testing*.

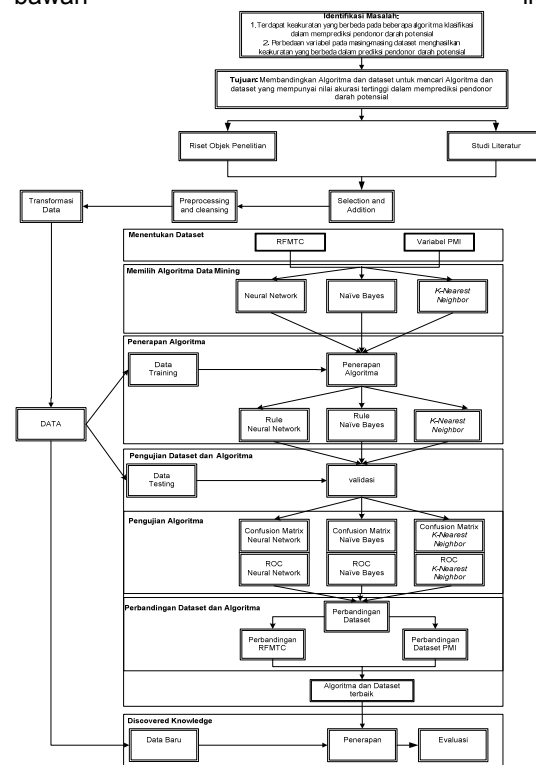
Fais A, Aditya D, Mulya I, Ramadien dan Sani (2015) melakukan pula penelitian klasifikasi pendonor menggunakan metode *naïve baiyes clasifier*. Adapun variabel yang digunakan dalam penelitian ini adalah *hemoglobin* (HB), tensi atas, tensi bawah, berat badan, umur, jenis kelamin, riwayat penyakit menular, interval donor serta mempunyai variabel target pendonor dan non pendonor. Dalam penelitian ini digunakan dataset sebanyak 400 *record* yang di bagi menjadi 350 *record* sebagai data *training* dan 50 *record* sebagai data *testing*. Hasil yang didapatkan

dalam penelitian ini adalah keakuratan rata-rata yang didapatkan dari pengujian ini sebesar 74% dari 4 percobaan pengujian.

Dalam penelitian ini akan dibandingkan dataset RFMTC dengan dataset PMI dalam penentuan pendonor darah yang potensial untuk mendonorkan darahnya kembali, serta data yang digunakan dalam penelitian ini adalah data *real* yang ada di PMI Kabupaten Bantul, Kemudian dicari algoritma yang paling akurat untuk digunakan pada kedua dataset tersebut dengan cara membandingkan tiga algoritma yaitu *Neural Network*, *K-Nearest Neighbor* dan *Naive Bayes* tujuannya agar mendapatkan algoritma dengan akurasi terbaik dari dataset yang diuji, untuk memprediksi pendonor darah yang potensial mendonorkan darahnya kembali.

2.8 Kerangka Pemikiran

Sebagai proses awal analisa data dilakukan pada awal penelitian ini dan untuk memberikan kemudahan dibuat suatu kerangka pemikiran yang dapat dilihat pada Gambar 2 di bawah ini.



Gambar 2. Kerangka pemikiran

Untuk penelitian ini data diperoleh melalui data pendonor yang melakukan donor darah di UDD PMI Kabupaten Bantul. Melalui pengamatan data tersebut dan dari studi literature baik dari jurnal maupun buku,

ditentukan variabel yang berhubungan dengan pendonor darah. Dengan menggunakan teknik *data mining* dengan pendekatan Algoritma *neural network*, *k-nearest neighbor* dan *naive bayes* serta menggunakan program aplikasi aplikasi *rapid miner*. Kemudian diukur dengan *confusion matrix* dan kurva *ROC*. Sehingga melalui proses tersebut akan terlihat suatu pola pendonor darah apakah pendonor tersebut termasuk pendonor potensial atau bukan dan tingkat keakuratan hasil dari tiga algoritma pada dua dataset yang kita uji dapat diketahui, untuk kemudian dipilih algoritma dan dataset terbaik sebagai dasar membangun sistemnya.

3. METODOLOGI PENELITIAN

Penelitian ini adalah penelitian eksperimen. Penelitian ini bertujuan untuk mencari algoritma dan dataset yang paling tepat untuk digunakan untuk mengklasifikasikan pendonor sukarela yang potensial mendonorkan darahnya kembali dan yang tidak. Variabel target dari penelitian ini adalah donor lagi yang mempunyai *class* “Ya” dan “Tidak”. Data yang digunakan dalam penelitian ini berasal dari data pendonor yang melakukan donor darah di UDD PMI Kabupaten Bantul tahun 2012 sampai 2014. Data pendonor tersebut akan dianalisa untuk kemudian ditarik kesimpulan algoritma serta dataset mana yang tepat untuk digunakan dalam pengklasifikasian.

3.1 Variabel Dataset

Variabel yang digunakan berkaitan dengan pendonor darah sukarela yang potensial mendonorkan darahnya kembali, pada penelitian ini akan digunakan dua dataset yaitu menggunakan dataset RFMTC dan dataset PMI. Adapun variabel dari masing-masing dataset yang digunakan adalah sebagai berikut :

Tabel 1. Variabel Dataset PMI

Variabel	Keterangan
Jkel	Jenis Kelamin
Levelumur	Tingkat umur
Pekerjaan	Pekerjaan
Goldarah	Golongan darah
Tensi	Tekanan darah
Beratbadan	Berat badan
KadarHB	Kadar Hemoglobin

Tabel 2. Variabel Dataset RFMTC

Variabel	Keterangan
Recency	Jarak waktu dengan donor terakhir
Frequency	Jumlah total donor
Monetary	Jumlah darah yang didonorkan
Time	Jarak waktu sejak donor pertama

Populasi dari penelitian ini adalah 15.201 data yang merupakan hasil olahan dari database UDD PMI Kabupaten Bantul. Untuk

mendapatkan data yang baik, dari 15.201 data yang didapatkan dari UDD PMI Kabupaten Bantul dilakukan *preprocessing*, sebanyak 15.201 *record* direduksi dengan menghilangkan duplikasi data sehingga didapatkan 6874 data sebagai dataset PMI. Dan jumlah *record* sebagai dataset RFMTC terdapat 9.680 data dan didapatkan 824 data setelah dihilangkan duplikasi datanya

Untuk pembentukan model dengan dataset RFMTC dan dataset PMI, sebagai data training diambil 700 *record* dan data testing diambil 100 *record*, sebagai data baru diambil sebanyak 24 *record*. Untuk menentukan pembagian sampel data *training* dan data *testing* dibagi dengan menggunakan metode *Systematic Random Sampling*.

Langkah awal yang dilakukan dalam *Systematic Random Sampling* adalah menentukan interval sampel. Adapun cara mendapatkan Interval sampel dengan membagi jumlah populasi dengan jumlah sampel yang akan diambil (Eriyanto, 2007).

$$\text{Interval sampel} = \frac{\text{Jumlah Populasi}}{\text{Jumlah sampel}}$$

Selanjutnya menentukan *record* pertama dari sampel. Jika *record* pertama kita beri simbol x dan jarak antar sampel kita beri simbol i , maka *record* selanjutnya adalah :

$$\begin{aligned} \text{record pertama} &= x \\ \text{record kedua} &= x + i \\ \text{record ketiga} &= x + 2i \text{ dan seterusnya} \\ &\text{hingga unsur ke-}n \end{aligned}$$

3.2. Metode Pengumpulan Data

Dalam penelitian ini metode pengumpulan data yang digunakan adalah metode pengumpulan data sekunder, yaitu dengan cara mengunduh data yang telah tersedia di *database* sistem informasi donor darah UDD PMI Kabupaten Bantul. Sampel dari data penelitian ini adalah data profile pendonor darah dan data hasil transaksi donor darah yang didapatkan dari UDD PMI Kabupaten Bantul.

3.3. Analisis Data

1. Domain Understanding dan KDD Goals

Tahap pertama dari KDD yaitu *Domain Understanding* dan *KDD Goals* dimana pada tahapan ini adalah mencari persoalan-persoalan yang berkaitan dengan klasifikasi pendonor darah yang potensial mendonorkan darahnya kembali, yang dilanjutkan dengan menentukan rumusan masalah serta menganalisa kebutuhan untuk dapat memecahkan masalah tersebut.

2. Selection and Addition

Pada tahapan *Selection and Addition* data yang didapatkan dari database PMI dianalisa agar dapat ditentukan variabel apa saja yang diperlukan berdasarkan analisa kebutuhan yang diperoleh. Adapun variabel yang digunakan ada di tabel 1 untuk variabel dataset PMI dan tabel 2 variabel dataset RFMTC.

3. Preprocessing Data Cleaning

Pada tahap ini akan dilakukan pembersihan data. Data yang didapat untuk penelitian ini sebanyak 15.147 record. Untuk penggunaan metode dengan dataset PMI pada penelitian ini digunakan 7 variabel adapun variabel dataset PMI yang digunakan terdapat pada tabel 1. Sedangkan yang menggunakan dataset RFMTC digunakan 4 variabel adapun variabel yang digunakan ada pada tabel 2. Setelah dihilangkan duplikasi dan *null value*, 6874 record dapat digunakan untuk pembuatan model dengan dataset PMI dan terdapat 824 record untuk dataset RFMTC. Karena akan dilakukan pengujian akurasi algoritma klasifikasi antara kedua dataset, maka penggunaan jumlah data antara dataset PMI dan dataset RFMTC disamakan yaitu dengan menggunakan 824 data sebagai dataset, adapun pemilihan data sampel menggunakan *Systematic Random Sampling*.

4. Transformation

Data yang digunakan dalam penulisan ini beberapa masih dalam bentuk angka, oleh karenanya data ditransformasikan ke dalam klasifikasi atau kelompok berdasar interval untuk pengujian algoritma *K-Nearest Neighbour* dan *Naive Bayes* sedangkan untuk pengujian menggunakan algoritma *Neural Network*, tetap dalam bentuk angka.

Pilihan interval kelas tergantung pada jumlah kelas untuk distribusi tertentu dan ukuran data. Interval kelas diusahakan harus dengan ukuran yang sama. Prof. Sturge telah memberikan rumus berikut untuk menentukan ukuran interval kelas (Jain dan Aggarwal, 2010):

$$\text{size of Class Interval } i = \frac{\text{Large Value} - \text{Smallest Value}}{1 + 3,322 \log N}$$

Keterangan :

N = Total Frekuensi

i = Ukuran Kelas Intervalence

5. Evaluation and Interpretation

Pada tahap ini adalah tahap pengujian dilakukan dengan menggunakan 700 data training data dan 100 data testing untuk dataset

RFMTC dan dataset PMI dengan melihat hasil akurasi pada proses klasifikasi pada dataset PMI maupun dataset RFMTC menggunakan algoritma *Neural Network*, *K-Nearest Neighbor* dan *naive Bayes*. Serta dievaluasi menggunakan metode *confusion matrix* dan kurva ROC.

6. Discovered Knowledge

Setelah pembentukan model, dilakukan analisa, dan pengukuran, selanjutnya pada tahap ini diterapkan model yang paling akurat dari algoritma dan dataset yang diuji dengan menggunakan data baru sebagai dasar pembuatan aplikasi yang dibuat berdasarkan algoritma dan dataset terbaik

4. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk melakukan klasifikasi pendonor darah yang potensial mendonorkan darahnya kembali pada unit donor darah dan melihat akurasi algoritma *Neural network*, *K-Nearest Neighbor*, dan *Naive Bayes* dengan menggunakan dataset PMI dan dataset RFMTC, kemudian menganalisa akurasi ketiga algoritma dengan membandingkan ketiga algoritma tersebut pada masing-masing dataset. Sehingga didapatkan dataset dan algoritma yang paling cocok digunakan untuk melakukan klasifikasi pendonor darah yang akan mendonorkan darahnya kembali atau tidak.

4.1 Pengujian Confussion Matrix

1. Pengujian Dataset RFMTC

Dari Tabel *confussion matrix* dari masing-masing algoritma pada pengujian dengan dataset RFMTC, menghasilkan perbandingan nilai *accuracy* pada masing-masing algoritma pada Tabel 3.

Tabel 3. Nilai *accuracy* dataset RFMTC

Algoritma	Akurasi
<i>Neural network</i>	80%
<i>K-Nearest Neighbor</i>	71%
<i>Naive Bayes</i>	71%

Dari Tabel 3 diatas, dapat kita lihat perbandingan nilai *accuracy* pada masing-masing algoritma pada dataset RFMTC menunjukkan hasil bahwa algoritma *neural network* memiliki nilai akurasi 80%. Selanjutnya algoritma *naive Bayes* dengan akurasi 71%. Sedangkan algoritma *K-Nearest neighbor* mempunyai akurasi sama dengan *naive Bayes* sebesar 71%.

Sehingga dari hasil pada Tabel 3 diatas didapatkan bahwa algoritma *neural network* yang paling cocok untuk digunakan untuk

klasifikasi pendonor darah potensial dengan dataset RFMTC dibandingkan dengan algoritma *K-Nearest neighbor* dan *naïve bayes*. Dengan tingkat akurasi 80%.

2. Pengujian Dataset PMI

Sedangkan dari tabel *confussion matrix* dari masing-masing algoritma pada pengujian dengan dataset PMI, Berikut perbandingan nilai *accuracy* pada masing-masing algoritma pada dataset tersebut yang disajikan pada Tabel 4.

Tabel 4. Nilai *accuracy* dataset PMI

Algoritma	Akurasi
<i>Neural network</i>	65%
<i>K-Nearest Neighbor</i>	52%
<i>Naïve Bayes</i>	56%

Dari Tabel 4 diatas, dapat kita lihat perbandingan nilai *accuracy* pada masing-masing algoritma pada dataset PMI, menunjukkan hasil bahwa algoritma *neural network* memiliki tingkat akurasi yang lebih tinggi dengan akurasi 65% .Tingkat akurasi yang paling tinggi diantara ketiga algoritma yang lainnya pada dataset PMI dimiliki oleh algoritma *neural network* .

Sehingga dari hasil pada Tabel 4 diatas didapatkan bahwa algoritma *neural network* yang paling memungkinkan untuk digunakan sebagai klasifikasi pendonor darah potensial dengan dataset PMI dibandingkan dengan algoritma *K-Nearest neighbor* dan *naïve bayes*.

Dengan menggunakan *Confusion Matrix* pada kedua dataset dengan tiga algoritma, dapat dilihat perbandingan akurasi kedua dataset dengan tiga lagoritma tersebut pada Tabel 5:

Tabel 5. Perbandingan Akurasi

Metode	Akurasi	
	Dataset RFMTC	Dataset PMI
<i>Neural network</i>	80%	65%
<i>K-Nearest Neighbor</i>	71%	52%
<i>Naïve Bayes</i>	71%	56%

Dari hasil akurasi pada Tabel 6 algoritma *neural network* dengan dataset RFMTC memiliki nilai akurasi yang paling tinggi dengan akurasi 80%, dibandingkan dengan yang lainnya sehingga algoritma *neural network* dan dataset RFMTC dapat digunakan sebagai dasar untuk melakukan klasifikasi pendonor yang potensial menyumbangkan darahnya kembali.

4.2 Pengujian Curva ROC

Adapun hasil evaluasi dengan kurva ROC pada pengujian dengan dataset RFMTC dan dataset PMI, hasilnya dapat dilihat pada Tabel 5, sebagai berikut:

Tabel 6. Komparasi nilai ROC

Algoritma	Dataset RFMTC	Datase t PMI
<i>Neural network</i>	0,851	0,605
<i>K-Nearest Neighbor</i>	0,500	0,500
<i>Naïve Bayes</i>	0,751	0,566

Berdasarkan komparasi nilai ROC pada Tabel 5 pada dataset RFMTC, nilai pada *Neural network* memiliki tingkat ROC paling tinggi yaitu 0,851 dengan tingkat nilai diagnosa *good classification*. Sedangkan nilai ROC yang dihasilkan berdasarkan komparasi nilai ROC pada dataset PMI, nilai pada *Neural network* juga memiliki tingkat ROC paling tinggi yaitu 0,744 dengan tingkat nilai diagnosa *fair classification* diantara algoritma yang lainnya. Sedangkan jika di bandingkan antara dataset RFMTC dan dataset PMI , maka dataset RFMTC dengan algoritma *Neural Network* yang mempunyai nilai paling tinggi.

Berdasarkan perbandingan akurasi yang dihasilkan dengan pengujian *Confusion Matrix* dan perbandingan nilai ROC maka dataset RFMTC dengan algoritma *Neural Network* lah yang mempunyai nilai paling tinggi. Kemudian ketika algoritma dan dataset telah terpilih diterapkan pada data baru, didapatkan akurasi sebesar 83,33% untuk data baru

.Tabel 7. Hasil Akurasi Data Baru

Table View Plot View			
accuracy: 83.33%			
	true 0	true 1	class precision
pred 0	10	2	83.33%
pred 1	2	10	83.33%
class recall	83.33%	83.33%	

3. KESIMPULAN

Kesimpulan yang dapat diambil dari penelitian mengenai pendonor yang potensial mendonorkan darahnya kembali yang dilakukan di UDD PMI Kabupaten Bantul adalah:

1. Penelitian ini membandingkan dua dataset yaitu dataset RFMTC dan dataset PMI serta tiga algoritma yaitu *neural network*, *k-nearest neighbor* dan *naïve bayes*, untuk mendapatkan dataset dan algoritma yang tepat dalam memprediksi pendonor darah yang potensial mendonorkan darahnya kembali pada UDD PMI Kabupaten Bantul.
2. Dalam penelitian ini pembuatan model dilakukan menggunakan dataset PMI dan dataset RFMC serta algoritma *neural network*, *k-nearest neighbor* dan *naïve bayes* menggunakan data pendonor yang melakukan donor darah pada tahun 2012 sampai 2014. Model yang dihasilkan dari

masing-masing dataset, dikomparasi untuk mengetahui algoritma dan dataset yang paling baik dalam memprediksi pendonor darah potensial. Untuk mengukur kinerja ketiga algoritma tersebut digunakan metode pengujian *Confusion Matrix* dan Kurva ROC, diketahui bahwa algoritma *neural network* dengan dataset RFMTC memiliki nilai akurasi paling tinggi, diikuti oleh metode *Naïve bayes* dengan dataset RFMTC dan *k-nearest neighbor* dengan dataset RFMTC, dataset PMI dengan algoritma *neural network*, *k-nearest neighbor* dan *naïve bayes* menempati urutan terbawah.

3. Algoritma *neural network* dengan dataset RFMTC mempunyai akurasi yang cukup baik dengan nilai akurasi data *testing* 80%, dibandingkan dengan algoritma *K-Nearest Neighbor* dan *Naïve Bayes* maupun ketiga algoritma tersebut dengan variabel PMI. Sehingga algoritma *neural network* dengan dataset RFMTC pada penelitian ini dapat direkomendasikan untuk melakukan prediksi pendonor yang potensial mendonorkan darahnya kembali.

Adapun saran untuk penelitian selanjutnya mengenai prediksi pendonor darah potensial agar penelitian ini menjadi lebih berkembang yaitu :

1. Penelitian pendonor darah potensial ini dapat dikembangkan lagi dengan lebih spesifik pada masing-masing golongan darah sehingga didapatkan prediksi pendonor darah potensial menurut golongan darah dan menurut *rhesus*.
2. Membandingkan algoritma maupun dataset yang terpilih menggunakan algoritma yang lainnya agar dapat diketahui algoritma yang terbaik untuk memprediksi pendonor darah potensial.
3. Penelitian ini dapat dikembangkan dengan melakukan optimasi pada variabel dengan menggunakan *feature selection* sehingga dapat menentukan variabel yang paling berpengaruh terhadap akurasi hasil.

4. DAFTAR PUSTAKA

- [1] Abidin, T. F., & Subianto, M. (2015). <http://www.informatika.unsyiah.ac.id/>. Retrieved 12 5, 2014, From <http://www.informatika.unsyiah.ac.id/Tfa/Dm/Dm-Praktikum-Naive-Bayesian.Pdf>
- [2] Akthar, F., & Hahne, C. (2012). *Rapid Miner 5 Operator Reference*. Dortmund: Rapid-I GmbH.
- [3] American Cancer Society. (7. Oktober 2014). *Blood Transfusion and Donation*. 20. January 2015: <http://www.cancer.org>.
- [4] Aviliani, Sumarwan, U., Sugema, I., & Saefuddin, A. (2011). Segmentasi Nasabah Tabungan Mikro Berdasarkan Recency, Frequency, Dan Monetary : Kasus Bank BRI. *Finance and Banking Journal*, 97.
- [5] Billett, H. (26. February 2015). *Chapter 151 Hemoglobin and Hematocrit*. 26 February 2015: <http://www.ncbi.nlm.nih.gov/books/NBK259/>
- [6] Bramer, M. (2007). *Principles of Data Mining*. Verlag London: Springer.
- [7] Darwiche, M., Feuilloy, M., Bousaleh, G., & Schang, D. (2010). *Prediction of blood transfusion donation*. *IEEE*, 978.
- [8] Dinkes Lumajang. (20 November 2013). *berat-badan : dinkeslumajang.or.id*. 3 Maret 2015: <http://dinkeslumajang.or.id/berat-badan/>
- [9] Eriyanto. (2007). *Teknik Sampling Analisis Opini Publik*. Yogyakarta: LKIS.
- [10] Fais A, S. N., Aditya D, M., & Mulya I, S. (2015). *Klasifikasi Calon Pendonor Darah Dengan Metode Naive Bayes Classifier*. Malang: Universitas Brawijaya
- [11] Fa'rifah, R., & Purhadi. (2012). Analisis Survival Faktor-Faktor yang Mempengaruhi Laju Kesembuhan Pasien Penderita Demam Berdarah Dengue (DBD) di RSU Haji Surabaya dengan Regresi Cox. *Jurnal Sains Dan Seni Its*, 271-276.
- [12] Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques*. Verlag Berlin Heidelberg: Springer.
- [13] Institut Teknologi Sepuluh November. (28 November 2014). *Open Content*. 18 January 2015 <http://oc.its.ac.id/ambilfile.php?idp=77>
- [14] Jain, T.R & Aggarwal, S.C . (2010). *Statistics for BBA*. New Delhi: VK Indiana Enterprises.
- [15] Jeffheaton. (2008, September 14). *The Number Of Hidden Layers*. Retrieved December 18, 2014, From <http://www.Heatonresearch.Com/>: <http://www.Heatonresearch.Com/Node/707>
- [16] Kementerian Kesehatan Republik Indonesia. (24 Juni 2014). *Darah Untuk Menyelamatkan Ibu*. 6 Desember 2014: <http://www.depkes.go.id/article/print/201406250001/darahuntukmenyelamatkanibu.html>
- [17] Kusriani, & Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Offset.

- [18] Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. New York: Springer.
- [19] Melati, E., Passarella, R., Primartha, R., & Murdiansyah, A. (2011). Desain dan Pembuatan Alat Pendeteksi Golongan Darah Menggunakan Mikrokontroler. *Jurnal Generic*, 48-54.
- [20] Moniaga, V., Pangemanan, D., & Rampengan, J. (2013). Pengaruh Senam Bugar Lansia Terhadap Tekanan Darah Penderita Hipertensi Di BPLU Senja Cerah Paniki Bawah. *Jurnal e-Biomedik (eBM)*, 785-789.
- [21] Nofiansyah, & Rochmawati, D. H. (2014). Hubungan Antara Golongan Darah Dengan Perilaku Kekerasan pada Pasien Gangguan Jiwa di Rumah Sakit Jiwa Daerah Dr. Amino Gondohutomo Semarang. *Prosiding Konferensi Nasional II PPNI Jawa Tengah 2014*, 213-221.
- [22] Nugroho, W. (2006). *Komunikasi Dalam Keperawatan Gerontik*. Jakarta: EGC.
- [23] Nurfajar, L. (24. Januari 2015). Personal Interview.
- [24] Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset.
- [25] Pusat Data dan informasi Kementrian Kesehatan RI. (2014). *Situasi Donor Darah di Indonesia*. Jakarta Selatan: Pusat Data dan informasi Kementrian Kesehatan RI.
- [26] Siang, J. J. (2009). Jaringan Syaraf Tiruan & Pemrogramannya menggunakan Matlab. Yogyakarta: Andi Offset.
- [27] Sihombing, M. (2010). Hubungan Perilaku Merokok, Konsumsi Makanan/Minuman, dan Aktivitas Fisik dengan Penyakit Hipertensi pada Responden Obes Usia Dewasa di Indonesia. *Maj Kedokt Indon*, 406-412.
- [28] Sudarma, M. (2008). *Serologi Untuk Kesehatan*. Jakarta: Salemba Medika.
- [29] Uda, U. (2013). Pemahaman Orang Tua Tentang Gender dalam Menerapkan Pola Asuk Kepada Anak Remaja di Desa Long Payau. *eJournal sosiatri*, 72-84.
- [30] Wadud, M. (2012). Hubungan Umur dan Pekerjaan Ibu Dengan Kejadian Hyperemesis Gravidarum di Instalasi Kebidanan Rumah Sakit Muhammadiyah Kota Palembang Tahun 2012. S. 1-16.
- [31] Widiasih, E., Mubin, M., & Hidyati, E. (2014). Persepsi Masyarakat Terhadap Pelayanan BPJS Di RSI Kendal. *Prosiding Konferensi Nasional II PPNI Jawa Tengah 2014*, (S. 274-279).
- [32] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning and Tools*. Burlington: Morgan Kaufmann Publisher.
- [33] Yuniar, R. J., Rahadi S, D., & Setyawati, O. (2013). Perbaikan Metode Prakiraan Cuaca Bandara Abdulrahman Saleh dengan Algoritma Neural Network Backpropagation. *Jurnal EECCIS*, 65-70.
- [34] Yunus, M., Dahlan, H. S., & Santoso, P. B. (2014). SPK Pemilihan Calon Pendonor Darah Potensial dengan Algoritma C4.5 dan Fuzzy Tahani. *Jurnal EECCIS*, 47-54.
- [35] Zailani, A. U. (2014). Pengujian Model Neural Network Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Kanker Payudara. Pamulang: [Http://www.unpam.ac.id/](http://www.unpam.ac.id/).