IMPLEMENTASI ALGORITMA *NAÏVE BAYES* UNTUK KLASIFIKASI PENDERITA PENYAKIT JANTUNG DI INDONESIA MENGGUNAKAN *RAPID MINER*

Donny Maulana[1)], Rezayadi Yahya[2)]

Program Studi Teknik Informatika Fakultas Teknik
Universitas Pelita Bangsa
donny.maulana@pelitabangsa.ac.id

Disetujui, 30 Desember 2019

Abstraksi

Penyakit jantung terjadi karena penyumbatan sebagian atau total dari suatu lebih pembuluh darah, akibat dari adanya penyumbatan maka dengan sendirinya suplai energi kimiawi ke otot jantung berkurang, sehingga terjadi gangguan keseimbangan antara suplai dan kebutuhan darah, penyakit jantung merupakan penyebab kematian tertinggi kedua setelah stroke. Penelitian ini bertujuan untuk mengetahui apakah teknik klasifikasi dengan penerapan algoritma *naive bayes* dapat digunakan untuk prediksi penyakit jantung, serta mendapatkan informasi mengenai *akurasi, presicion* dan *recall* yang didapat saat melakukan pengujian data pasien menggunakan algoritma *naive bayes.*Penelitian ini menggunakan teknik klasifikasi dan tahapan - tahapan pada data mining untuk klasifikasi data pasien yang menderita penyakit jantung dengan algoritma *Naïve Bayes* menggunakan *tool RapidMiner*, pengolahan data yang akan dijadikan *dataset* dalam penelitian ini. Dari data tersebut akan dibagi menjadi 80% data *training* dan 20% data *testing*. Hasil dari penelitian data menyatakan tingkat *accurcy* 70,00 %, *persicion 77,9* % dan *recall* 82.10 % putuskan dalam klasifikasi *naïve bayes.* Berdasarkan penelitian yang sudah dilakukan mendapatkan kesimpulan bahawa teknik klasifikasi dengan penerapan algoritma *Naïve Bayes* dapat digunakan untuk melakukan prediksi ada penyakit jantung.

Kata kunci : penyakit jantung, algoritma *Naive Bayes*, *RapidMiner*

Abstract

Disease occurs due to partial or total blockage of more than one blood vessel, as a result of the blockage, the chemical energy supply to the heart muscle decreases, resulting in a disruption of the balance between blood supply and demand, heart disease is the second highest cause of death after stroke This study aims to determine whether the classification technique with the application of the Naive Bayes algorithm can be used to predict heart disease, and to obtain information about the accuracy, precision and recall obtained when testing patient data using the Naive Bayes algorithm. This research uses classification techniques and stages. In data mining for data classification of patients suffering from heart disease with the Naïve Bayes algorithm using the Rapid Miner tool, data processing will be used as the dataset in this study. From this data will be divided into 80% training data and 20% testing data. The results of the data study stated that the accuracy rate was 70.00%, persicion 77.9% and recall was 82.10%, and decided in the naïve Bayes classification. Based on the research that has been done, it can be concluded that the classification technique with the application of the Naïve Bayes algorithm can be used to predict heart disease.

Keywords: *heart disease, Naive Bayes algorithm, Rapid Miner*

1. Pendahuluan

Jantung merupakan suatu organ otot berongga yang terletak dipusat dada yang menompa darah lewat pembuluh darah. Penyakit jantung di Indonesia merupakan penyakit nomor dua yang mendorong angka kematian yang cukup tinggi, sehingga sampai sekarang penyakit tersebut ditakuti oleh manusia. Penyakit jantung terjadi karena penyumbatan sebagian atau total dari suatu lebih pembuluh darah, akibat dari adanya penyumbatan maka dengan sendirinya suplai energi kimiawi ke otot jantung berkurang, sehingga terjadi gangguan keseimbangan antara suplai dan kebutuhan darah.

Data Organisasi Kesehatan Dunia *World Health Organization* (WHO) tahun 2015, menyebutkan lebih

dari 17 juta orang di dunia meninggal akibat penyakit jantung dan pembuluh darah. Lebih dari 75% kematian akibat penyakit jantung terjadi di negara berkembang yang berpenghasilan rendah sampai sedang. *Sample Registration System* (SRS) Indonesia tahun 2014 menunjukkan penyakit jantung merupakan penyebab kematian tertinggi kedua setelah stroke, yaitu sebesar 12,9% dari seluruh penyebab kematian tertinggi di Indonesia.

Tingginya faktor kematian akibat penyakit jantung dapat dicegah dan ditekan faktor resikonya. Kurangnya pengetahuan masyarakat tentang gejala penyakit jantung. Maka perlu dilakukan suatu langkah dini sebagai upaya penanganan dan pecegahan penyakit jantung. Hal ini bisa dihindari dengan memanfaatkan data-data pasien yang sudah tersimpan dalam basis data untuk dibuat suatu pola penetuan penyakit jantung dengan teknik komputasi cerdas sehingga masyarakat mengetahui faktor penyakit jantung.

Berdasarkan latar belakang tersebut, peneliti melakukan metode untuk skrining penyakit jantung melalui menerapkan data mining. Peneliti melakukan teknik klasifikasi menggunakan algoritma *Naïve Bayes* untuk melakukan pengolahan data penyakit jantung dengan menggunakan *RapidMiner* sehingga akan didapat hasil akurasi dari algoritma *Naïve Bayes* terhadap penyakit jantung.

2. Tinjauan Studi

2.1 Teknologi Informasi

Teknologi informasi adalah suatu teknologi yang digunakan untuk mengolah data, termasuk memproses, mendapatkan, menyusun, menyimpan, memanipulasi data dalam berbagai cara untuk menghasilkan informasi yang berkualitas, yaitu informasi yang relevan, akurat dan tepat waktu, yang digunakan keperluan pribadi, bisnis, dan pemerintahan dan merupakan informasi yang strategis untuk pengambilan keputusan (Sutabri, 2014).

2.2 Sistem Informasi

Menurut Krismaji (2015) Sistem informasi adalah cara-cara yang diorganisasi untuk mengumpulkan, memasukkan, dan mengolah serta menyimpan data, dan cara-cara yang diorganisasi untuk menyimpan, mengelola, mengendalikan, dan melaporkan informasi sedemikian rupa sehingga sebuah organisasi dapat mencapai tujuan yang telah ditetapkan.

2.3 Data Mining

Data Mining merupakan proses pengekstraksian informasi dari sekumpulan data yang sangat besar melalui penggunaan algoritma dan teknik penarikan dalam bidang statistik, pembelajaran mesin dan sistem manajemen basis data. Data mining adalah proses menganalisa data dari perspektif yang berbeda dan menyimpulkannya menjadi informasi-informasi penting yang dapat dipakai untuk meningkatkan keuntungan, memperkecil biaya pengeluaran, atau bahkan keduanya. Definisi lain mengatakan Data Mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam data berukuran besar. Dari beberapa definisi di atas dapat ditarik kesimpulan bahwa Data Mining merupakan proses ataupun kegiatan untuk mengumpulkan data yang berukuran besar kemudian mengekstraksi data tersebut menjadi informasi – informasi yang nantinya dapat digunakan.

2.4 Data Cleaning

Pada tahap data cleaning merupakan proses pembersihan dari data yang akan dipakai untuk penghapusan data dengan membuang *missing value*, duplikasi data, dan memeriksa inkonsistensi data dan memperbaiki kesalahan pada data. Proses pembersihan data dilakukan secara manual dengan bantuan *software spreadsheet.*

2.5 *Data Selection*

Data Selection merupakan proses pemilihan data dari sekumpulan data operasional yang ada sebelum masuk ke tahap mining data maupun informasi. Pada proses ini dilakukan pemilihan data atribut

atau variable yang relevan yang akan digunakan dalam penelitian. Karena tidak semua atribut yang terdapat dalam database dapat digunakan. untuk melakukan penelitian yaitu dengan memilih atribut-atribut yang akan digunakan dan menghilangkan atribut-atribut yang kurang relevan atau terdapat data yang rusak. Atribut atau variable yang digunakan dalam penelitian ini dapat dilihat pada tabel berikut:

Tabel 1. *Data Selection*

No	Atribut **Sebelum** *Selection*	No	Atribut **Sesudah** *Selection*
1	Usia	1	Usia
2	Jenis Kelamin	2	Jenis Kelamin
3	Tipe Sakit Dada	3	Tipe Sakit Dada
4	Tekanan Darah	4	Tekanan Darah
5	Kolesterol mg	5	Kolesterol mg
6	Gula Darah	6	Gula Darah
7	Elektrokardiografi	7	Detak Jantung Maksimal
8	Detak Jantung Maksimal		-
9	Sakit dada Ketika Olahraga		-
10	*Old peak*		-
11	Kemiringan Segmen		-
12	*Flourosopy*		-
13	Kondisi		-

2.6 *Data Transformation*

Tahap *Data Transformation* merupakan proses mengubah format data awal menjadi sebuah format data standar untuk proses pembacaan data dengan algoritma pada program maupun tool yang digunakan. dalam pengolahan dan dapat dihasilkannya hasil akurasi yang baik maka data harus di *transforming* kedalam bentuk data yang mudah di pahami. Pada penelitian ini data hasil unduhan diubah kedalam data kualitatif yang dapat mempermudah dalam pemodelan. Perubahan yang terjadi terhadap data yang sudah ditambang adalah sebagai berikut:

Tabel 2. *Data Transformation*

Atribut	Kelas
Usia	Dewasa
	Lansia
Jenis Kelamin	Laki-laki
	Perempuan
Tipe sakit dada	Angina pectoris
	Stable angina
	Unstable angina
	Prinzmetal's angina
Tekanan Darah	Rendah
	Normal
	Tinggi
Kolesterol	Normal
	Tinggi
Gula darah	Iya
	Tidak
Detak Jantung Maksimal	Normal
	Tidak Normal

2.7 *Metode Naive Bayes*

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya .

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas

jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. *Naive Bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan.

3. Desain Penelitian/Metodologi

3.1 Objek Penelitian

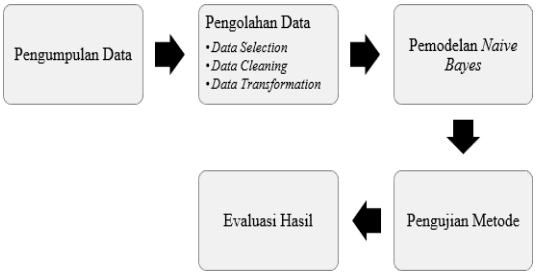
Penelitian secara umum dapat diartikan sebagai proses pengumpulan dan analisis data yang dilakukan secara sistematis dan logis untuk mencapai tujuan tertentu dan memperkaya pengetahuan itu sendiri oleh penemuan fakta dan wawasan yang tidak biasa, tahapan yang akan digunakan dalam melakukan prediksi terhadap data pasien yang menderita penyakit jantung dan penentuan atribut untuk mempermudah penelitian sehingga penelitian dapat berjalan dengan baik dan sistematis, serta memenuhi tujuan yang diinginkan.

3.2 Jenis Data Yang Digunakan

Data yang digunakan dalam penelitian ini merupakan data Sekunder dan kualitatif yang didapatkan dari proses unduh di peroleh dari situs di https://www.kaggle.com/.

3.3 Tahapan Penelitian

Dalam melakukan analisis dan mencari pola pada data pasien yang menderita penyakit jantung agar memudahkan penelitian dan dapat berjalan dengan sistematis dan memenuhi tujuan yang diinginkan maka dibuat langkah – langkah dalam tahapan penelitian yang akan dilakukan berikut :



Gambar 1. Tahapan Penelitian

Pada penelitian ini menggunakan teknik klasifikasi dan tahapan - tahapan pada data mining untuk klasifikasi data pasien yang menderita penyakit jantung algoritma *Naïve Bayes*, pengolahan data dan yang akan dijadikan dataset dalam penelitian ini Dari data tersebut akan dibagi menjadi 80% *data training* dan 20% *data testing*. Memisahkan data menjadi training dan testingset dimaksudkan agar model yang diperoleh nantinya memiliki kemampuan generalisasi yang baik dalam melakukan klasifikasi data. Data training atau trainingset adalah bagian dataset yang dilatih untuk membuat klasifikasi atau menjalankan fungsi dari sebuah algoritma sesuai dengan tujuannya masing-masing. *Data testing* atau *test set* adalah bagian dataset yang digunakan untuk melihat keakuratan atau performa dalam hasil.

3.4 Pengumpulan Data

Pengumpulan data cara atau teknik yang dapat digunakan oleh peneliti untuk mengumpulkan data. Atribut atau variabel adalah sifat atau nilai dari suatu objek, atau kegiatan yang mempunyai variasi tertentu yang ditetapkan oleh peneliti untuk dipelajari kemudian ditarik kesimpulannya, teknik pengumpulan data yang digunakan dalam penelitian dalam melakukan analisis dan mencari pola agar memudahkan penelitian dan dapat berjalan dengan sistematis dan memenuhi tujuan yang diinginkan.

3.5 Pengolahan Data

Pengolahan data pada penelitian ini menggunakan teknik klasifikasi pada data mining untuk memprediksi pasien yang menderita penyakit jantung dengan algoritma *Naïve Bayes*. Data yang akan dijadikan dataset dalam penelitian ini adalah data pasien penderita penyakit jantung data tersebut akan dibagi menjadi 80% *data training* dan 20% *data testing*. Memisahkan data menjadi training dan testing set dimaksudkan agar model yang diperoleh nantinya memiliki kemampuan generalisasi yang baik dalam melakukan klasifikasi data. Data training atau training set adalah bagian dataset yang dilatih untuk membuat prediksi atau menjalankan fungsi dari sebuah algoritma sesuai dengan tujuannya masing- masing. *Data testing* atau *test set* adalah bagian dataset yang digunakan untuk melihat keakuratan atau performa dari suatu data.

3.6 Pemodelan

Pemodelan pada penelitian ini dilakukan dengan data mining teknik klasifikasi algoritma *Naïve Bayes*. Teknik ini dipilih karena merupakan metode yang umum dipakai pada penelitian data mining untuk mengklasifikasi atau mengenali data- data yang dipelajari terutama pada prediksi penderita penyakit jantung. Algoritma yang akan diterapkan pada penelitian ini adalah *Naïve Bayes* . Algoritma ini merupakan algoritma yang sudah mapan dan banyak diimplementasikan pada teknik klasifikasi. Selain itu algoritma ini memiliki kelebihan yaitu berupa akurasinya yang baik dalam menangani sebuah dataset yang diolah.

3.7 Pengujian dan Validasi

Pengujian metode dilakukan dengan maksud mengetahui hasil perhitungan yang dianalisa dan mengukur metode serta algoritma yang digunakan apakah berfungsi dengan baik atau tidak. Proses pengujian menggunakan *tool rapidmener* dan melihat data apakah sesuai dengan hasil yang diperoleh melalui *tool* tersebut. Sedangkan validasi metode dan algoritma *Naïve Bayes* dilakukan dengan mengukur hasil *accuracy, percision* dan *recall* dan dapat dihitung dengan menggunakan *Confusion Matrix* sebagai berikut :

Nilai *accuracy* dihitung dengan cara menjumlah data benar yang bernilai positif (*True Positive*) ditambah dengan nilai Negatif (*True Negative*) dibagi dengan jumlah data benar yang bernilai positif *(True Positive*), Negatif (*True Negative*) dan ditambah dengan data salah yang bernilai positif (*False Positif*), Negatif (*False Negative*). *Accuracy* didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual.

4. Hasil Dan Pembahasan

4.1 Data Uji

Penelitian ini adalah pengujian terhadap algoritma naïve bayes, untuk mengetahui algoritma yang mana yang akan mendapatkan hasil nilai accuracy, precision, dan recall serta prediksi yang dapat digunakan untuk mengetahui pasien yang memiliki penyakit atau tidak memiliki penyakit jantung. Sumber data sebagai objek pada penelitian ini adalah data public yang terdapat situs https://www.kaggle.com/&https://archive.isuuci.edu/ (Universitas California,Invene). Data yang digunakan dalam penelitian ini terdiri dari atribut atau variabel seperti: Usia, Jenis kelamin, tipe sakit dada, tekanan darah, kolesterol, gula darah dan detak jantung maksimal.

4.2 Split Validation

Split Validation merupakan teknik validasi yang membagi data menjadi dua bagian, sebagian data training dan sebagian data testing. Data yang sudah disiapkan untuk klasifikasi dibagi menjadi dua menggunakan teknik sampling random untuk data training (80%) dan data testing (20%). Contoh perhitungan untuk pengambilan data testing adalah sebagai berikut :

Jumlah data keseluruhan (N)	= 500
Jumlah data testing	= 20% x 500 = 100
Jumlah sample (n)	= 100

Interval sampling (k) =N/n

= 500/100= 5

Unsur pertama yang diambil untuk data testing (D) = 1

Dataset akan dibagi menjadi data training dan data testing. Proses training dan testing dilakukan sebanyak 5 kali secara berulang- ulang. Pada iterasi ke-1, partisi $D1$ disajikan sebagai data testing dan partisi sisanya digunakan secara bersamaan dan berurutan sebagai data training. Iterasi kedua, subset $D1$,$D2$,…,Dk akan dites pada $D2$, Iterasi ketiga, subset $D1$,$D2$, D3 … ,Dk akan dites pada $D3$, dan selanjutnya hingga $D5$). Dari hasil diatas diperoleh data testing sebanyak 100 data, maka sisanya dijadikan data training sebanyak 500 – 100 = 400 data.

Tabel 3. Data Testing

Usia	Jenis Kelamin	Tipe Sakit Dada	Tekanan Darah	Kolesterol	Gula Darah	Detak Jantung Maksimal
Lansia	Laki-laki	pectoris angina,	Tinggi	Tinggi	Iya	Normal
Lansia	Laki-laki	unstable angina,	Tinggi	Tinggi	Tidak	Tidak Normal
Lansia	Laki-laki	prinzmetal's angina,	Rendah	Tinggi	Tidak	Normal
Lansia	Perempuan	unstable angina,	Normal	Tinggi	Tidak	Normal
Dewasa	Laki-laki	pectoris angina,	Rendah	Normal	Tidak	Tidak Normal
Lansia	Perempuan	prinzmetal's angina,	Tinggi	Tinggi	Iya	Normal
Lansia	Laki-laki	prinzmetal's angina,	Tinggi	Tinggi	Tidak	Normal
Lansia	Laki-laki	unstable angina,	Tinggi	Tinggi	Tidak	Tidak Normal
Lansia	Laki-laki	prinzmetal's angina,	Tinggi	Tinggi	Iya	Tidak Normal

Data Dos tersebut disimpan dalam format excel workbook yang selanjutnya diubah menjadi data.frame dengan perintah read.excel. Berikut Ini adalah data uji atau data testing untuk di olah ke dalam *tools RapidMiner*.

4.3 Hasil

Pembahasan dilakukan guna untuk mendapatkan nilai accuracy, precision, untuk algoritma naïve bayes, memprediksi, klasifikasi dos , recall. Pengertian akurasi adalah tingkat kedekatan hasil prediksi dengan hasil fakta. Presisi adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Serta recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

Dari data sampel yang diutarakan yaitu sebanyak 100 data , kemudian hasil dari data penyakit jantung menyatakan tingkat ke *accuracy, recall* dan *persicion naïve bayes*, tingkat *accuracy* 70,00 %, Percision 77,9 % dan *Recall* 82.10 % putuskan dalam klasifikasi prediksi *naïve bayes.* Model *model for label attribute Prediksi Naïve Bayes*

Training %	Testing %	Accuracy	Precision	*Recall*
(80%)	(10%)	70%	77,9%	82.10%

Gambar 2. Hasil Pengujian

5. Kesimpulan

Berdasarkan hasil pengujian dengan menggu nakan algoritma *naive bayes* dapat diambil suatu kesimpulan sebagai berikut :

1. Teknik klasifikasi dengan penerapan algoritma *Naïve Bayes* dapat digunakan untuk melakukan prediksi ada penyakit jantung atau tidak ada penyakit jantung.
2. Hasil dari penelitian data menyatakan tingkat *accuracy* 70,00%, *Percision* 77,9 % dan *Recall* 82.10% diputuskan dalam klasifikasi *naïve bayes*.

Daftar Pustaka

R. Pramunendar, I. Dewi, and H. Asari, "*Penentuan Prediksi Awal Penyakit Jantung Menggunakan Algoritma Back Propagation Neural Network dengan Metode Adaboost*," Semantik, vol. 2013, no. November, pp. 298–304, 2013.

Kemenkes RI, "*Situasi kesehatan jantung*," Pus. data dan Inf. Kementeri. Kesehat. RI, p. 3, 2014, doi: 10.1017/CBO9781107415324.004.

C. Science, L. A. Nurjanah, and D. S. Noviyanti, *"Klasifikasi Penyakit Diabetik Retinopathy dengan Metode Naïve Bayes pada Citra Retina,"* Comput. Sci. ICT, vol. 4, no. 1, pp. 978–979, 2018.

A. Rahmawati, D. Wintana, and S. Suhada, *"KLASIFIKASI NAÏVE BAYES UNTUK MENDIAGNOSIS PENYAKIT PNEUMONIA PADA ANAK BALITA (STUDI KASUS : UPTD PUSKESMAS SUKARAJA SUKABUMI),"* vol. 06, no. 03, pp. 241– 253, 2019.

A. Febrianto and P. Handayani, *"Implementasi Metode Model View Controller (MVC) Dalam Rancang Bangun Website SMK Yayasan Bakti Prabumulih,"* Paradig. - J. Komput. dan Inform., vol. XXI, no. 1, pp. 1–8, 2019, doi: 10.31294/p.v20i2.

N. Usman and G. Setiawan, "*Nurdin Usman, Konteks Implementasi Berbasis Kurikulum, Grasindo,Jakarta,2002,hal70 Guntur Setiawan, Impelemtasi dalam Birokrasi Pembangunan ,Balai Pustaka,Jakarta,2004,hal39 7,"* pp. 7–18, 2002.

S. Susanto and D. Suryani, *"Pengantar Data Mining,"* 2010.

S. Agarwal, Data mining: Data mining concepts and techniques. 2014.

D. Hukuman, T. Kepatuhan, S. Baru, P. Di, and P. Pesantren, *"Fakultas Psikologi Universitas Islam Negeri (Uin) Malang Universitas Islam Negeri (Uin) Malang,"* 2014.

A. Saleh, *"Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga,"* Creat. Inf. Technol. J., vol. 2, no. 3, pp. 207–217, 2015.

L. D. Benley and N. Damayanti, "5 Bab

2. Landasan Teori," pp. 5–22, 2008.

Hidayatul and dkk, *"Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes,"* J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 9, pp. 2546–2554, 2018.

R. Darmojo, *"Kecenderungan Meningkatnya Penyakit Jantung Di Indonesia,"* Bul. Penelit. Kesehat., vol. 21, no. 4 Des, 1993, doi: 10.22435/bpk.v21i4Des.367.

B. Rifai, *"Algoritma Neural Network Untuk Prediksi,"* Techno Nusa Mandiri, vol. IX, no. 1, pp. 1–9, 2013.

Susandra, *"Modul Panduan Microsoft Excel*," 2010.