PERBANDINGAN ALGORITMA *NAIVE BAYES* DENGAN ALGORITMA *K-NEAREST NEIGHBOR* UNTUK PREDIKSI PENYAKIT JANTUNG

[1]Agil Langga Dikayanto (1410651223)
[2] Agung Nilogiri, S.T., M.Kom.
Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember
agillangga95@gmail.com
agungnilogiri@unmuhjember.ac.id

ABSTRACT

Currently analysis of a data is very needed. One approach that can be done to analyze a data set is to classify data. Several classification methods are commonly used such as Artifficial Neural Network (ANN), Support Vector Machines (SVM), Decision Tree, Bayesian, and so on. The Naïve Bayes classification proved to have high accuracy and speed when applied to a large number of databases. Besides Naive Bayes, the K-Nearest Neighbor algorithm also has a high level of accuracy. Therefore the researcher wants to conduct a study by comparing the Naive Bayes algorithm and the K-Nearest Neighbor algorithm using heart disease diagnostic data taken from the public dataset provider UCI Machine Learning using Cross Validation testing techniques. From the whole test, the optimal K for the K-Nearest Neighbor algorithm that is neighboring distance is 7. The Naive Bayes algorithm produces the highest performance at 10-fold Cross Validation in the 4th test data folder with an accuracy of 90.00%, and a precision value of 86, 67%. The K-Nearest Neighbor Algorithm also produces the highest accuracy with a neighbor value of 7 in the 10-fold Cross Validation test with the highest accuracy in the 4th test data folder which is equal to 80.00%, and a precision of 90.00%. Based on the tests that have been done, the Naive Bayes algorithm is more accurate and better in the classification of heart disease than the K-Nearest Neighbor algorithm.

Keywords: *Naive Bayes Algorithm, K-Nearest Neighbor Algorithm, Analysis, Classification, Comparison, Cross Validation.*

PENDAHULUAN

Saat ini kebutuhan terhadap analisis suatu data sangat dibutuhkan. Pesatnya perkembangan data yang semakin tinggi mendorong untuk memanfaatkan data dalam penggalian informasi maupun pengetahuan. Salah satu pendekatan yang dapat dilakukan untuk menganalisis sekumpulan data yaitu dengan mengklasifikasi data. Klasifikasi termasuk dalam salah satu teknik data mining yang digunakan untuk membangun suatu model dari sampel data yang belum terklasifikasi untuk digunakan mengklasifikasi sampel data baru ke dalam kelas-kelas yang sejenis (Sartika & Sensuse, 2017).

Beberapa metode klasifikasi yang umum digunakan seperti *Artifficial Neural Network* (ANN), *Support Vector Machines* (SVM), *Decision Tree, Bayesian,* dan sebagainya. Selain beberapa metode tersebut, terdapat metode yang memiliki tingkat akurasi tinggi pada saat proses klasifikasi yaitu *Naïve Bayes*. Menurut Han dan Kamber dalam (Arisandy, 2017), klasifikasi *Naïve Bayes* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam basis data dengan jumlah yang besar. Penelitian yang dilakukan oleh Heru Sulistiono (2015) berjudul "*Kajian Penerapan Algoritma c4.5, Neural Network Dan Naïve Bayes Untuk Klasifikasi Mahasiswa Yang Bermasalah Dalam Registrasi*". Algoritma *Naive Bayes* memiliki tingkat akurasi yang paling tinggi dibandingkan dengan algoritma *Neural Network* dan *C4.5* dengan nilai persentase akurasi 93,58%. Selain *Naive Bayes*, algoritma *K-Nearest Neighbor* juga memiliki tingkat akurasi yang tinggi. Penelitian yang dilakukan oleh Mustakim (2016) berjudul "*Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa"* tingkat akurasi dari algoritma *K-Nearest Neighbor* sebesar 82%.

Pada tahun 2005 sedikitnya 17,5 juta atau setara dengan 30% kematian di seluruh dunia disebabkan oleh penyakit jantung. Menurut Badan Kesehatan Dunia (WHO) dalam (Oemiati & Rustika, 2015), 60% dari seluruh penyebab

kematian penyakit jantung adalah penyakit jantung koroner (PJK). Faktor gejala yang terdiagnosa sebagai penyakit jantung antara lain adalah jenis sakit dada (chest pain), tekanan darah tinggi (tresbps), kolesterol (chol), nilai tes EKG (resting electrodiagraphic "restacg"), denyut jantung (thalach) dan kadar gula (fasting blood sugar "FBS"), dan beberapa faktor lainnya yang mengidentifikasi bahwa seseorang mempunyai penyakit jantung (Rifai, 2013).

Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu (1) pembangunan model sebagai prototipe untuk disimpan sebagai memori dan (2) penggunaan model tersebut untuk melakukan pengenalan / klasifikasi / prediksi pada suatu objek data lain agar diketahui kelas mana objek data tersebut dalam model yang sudah disimpannya (Prasetyo, 2012).

Pengukuran Klasifikasi

Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Tabel 2.1 merupakan contoh matriks konfusi yang melakukan klasifikasi masalah biner (dua kelas), hanya ada dua kelas, yaitu kelas 0 dan 1. Setiap sel f_{ij} dalam matriks menyatakan jumlah *record*/data dari kelas i yang hasil prediksinya masuk ke kelas j. Misalnya, sel f_{11} adalah jumlah data dalam kelas 1 yang secara benar dipetakan ke kelas 1, dan f_{10} adalah data dalam kelas 1 yang dipetakan secara salah ke kelas 0 (Prasetyo, 2012).

Tabel 1 Matriks Konfusi untuk Klasifikasi Dua Kelas

F_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 0
Kelas asli (i)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Algoritma Naive Bayes

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam *Naive Bayes*, model yang digunakan adalah "model fitur independen" (Prasetyo, 2012).

Dalam *Bayes* (terutama *Naive Bayes*), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Contohnya, pada kasus klasifikasi hewan dengan fitur penutup kulit, melahirkan, berat, dan menyusui. Dalam dunia nyata, hewan yang berkembang biak dengan cara melahirkan dipastikan juga menyusui. Di sini ada ketergantungan pada fitur menyusui karena hewan yang menyusui biasanya melahirkan, atau hewan yang bertelur biasanya tidak menyusui. Dalam *Bayes*, hal tersebut tidak dipandang sehingga masing-masing fitur seolah tidak memiliki hubungan apa pun (Prasetyo, 2012).

Prediksi *Bayes* didasarkan pada *teorema Bayes* dengan formula umum sebagai berikut:

$$P(H|E) = \frac{P(E|H)\ x\ P(H)}{P(E)} \quad \text{.....(2.1)}$$

Algoritma K-Nearest Neighbor

K-Nearest Neighbor merupakan sebuah metode yang digunakan untuk klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran (Sumarlin, 2015).

Menurut (Sani, Zeniarja, & Luthfiarta, 2016) cara kerja dari KNN perlu adanya penentuan inputan berupa data latih, data uji dan nilai *K*. Kemudian mengurutkan data latih berdasarkan kedekatan jaraknya berdasarkan hitungan dari jarak data yang diuji dengan data latih. Setelah itu diambil dari *K* data latih teratas untuk menentukan kelas klasifikasi untuk kelas yang dominan dari *K* data latih yang diambil. Secara umum untuk mendefinisikan jarak antara dua objek *x* dan *y*, digunakan rumus jarak *Euclidean* (Mustakim & Oktaviani, 2016).

$$D(a,b) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \text{.....(2.2)}$$

Penyakit Jantung

Penyakit jantung koroner (PJK) adalah penyakit yang disebabkan adanya plak yang menumpuk di dalam arteri koroner yang mensuplai oksigen ke otot jantung. Penyakit ini termasuk bagian dari penyakit kardiovaskuler yang paling umum terjadi. Penyakit kardiovaskuler merupakan gangguan dari jantung dan pembuluh darah termasuk stroke, penyakit jantung rematik dan kondisi lainnya (Ghani, Dewi, & Novriani, 2016).

METODE PENELITIAN

Pengumpulan Data

Data yang peneliti gunakan dalam penelitian ini adalah dataset diagnosa penyakit jantung yang disediakan oleh (UCI) *Machine Learning University*, diunduh dari sits web http://archive.ics.uci.edu/ml/datasets/heart+disease. Data yang diperoleh sebanyak 300 yang diperiksa dan sebanyak 163 pasien terdeteksi sehat dan 137 terdeteksi sakit jantung.

Tabel 1 Atribut dan data penyakit jantung

No	Umur	Jenis Kelamin	Jenis Sakit Dada	Tekanan Darah	Kolesterol	Kadar Gula	Elektrokardiografi	Tekanan Jantung	Angina Induksi	oldpeak	slope	Scan Thallium	Hasil
1	63	male	angina	145	233	TRUE	hyp	150	FALSE	2,3	down	fix	buff
2	67	male	asympt	160	286	FALSE	hyp	108	TRUE	1,5	flat	norm	sick
3	67	male	asympt	120	229	FALSE	hyp	129	TRUE	2,6	flat	rev	sick
4	37	male	notang	130	250	FALSE	norm	187	FALSE	3,5	down	norm	buff
5	41	female	abnang	130	204	FALSE	hyp	172	FALSE	1,4	up	norm	buff
6	56	male	abnang	120	236	FALSE	norm	178	FALSE	0,8	up	norm	buff
7	62	female	asympt	140	268	FALSE	hyp	160	FALSE	3,6	down	norm	sick
8	57	female	asympt	120	354	FALSE	norm	163	TRUE	0,6	up	norm	buff
9	63	male	asympt	130	254	FALSE	hyp	147	FALSE	1,4	flat	rev	sick
10	53	male	asympt	140	203	TRUE	hyp	155	TRUE	3,1	down	rev	sick
11	57	male	asympt	140	192	FALSE	norm	148	FALSE	0,4	flat	fix	buff
12	56	female	abnang	140	294	FALSE	hyp	153	FALSE	1,3	flat	norm	buff
13	56	male	notang	130	256	TRUE	hyp	142	TRUE	0,6	flat	fix	sick
14	44	male	abnang	120	263	FALSE	norm	173	FALSE	0	up	rev	buff
15	49	male	abnang	130	266	FALSE	norm	171	FALSE	0,6	up	norm	buff
16	64	male	angina	110	211	FALSE	hyp	144	TRUE	1,8	flat	norm	buff
17	58	female	angina	150	283	TRUE	hyp	162	FALSE	1	up	norm	buff
18	58	male	abnang	120	284	FALSE	hyp	160	FALSE	1,8	flat	norm	sick
19	58	male	notang	132	224	FALSE	hyp	173	FALSE	3,2	up	rev	sick
20	60	male	asympt	130	206	FALSE	hyp	132	TRUE	2,4	flat	rev	sick

Preprocessing Data

Dari hasil *preprocessing data* terdapat beberapa *record* yang hilang atau rusak sebanyak 2 *record*. Oleh karena itu diperlukan sebuah teknik dalam *preprocessing data* (Rifai, 2013) yaitu:

a. *Data cleaning* bekerja membersihkan nilai kosong, tidak konsisten atau tupel kosong *(missing value dan noisy).*

b. *Data integration* menyatukan tempat penyimpanan (arsip) yang berbeda dalam satu arsip.

c. *Data reduction* jumlah atribut yang digunakan untuk data traning terlalu besar sehingga ada beberapa atribut yang tidak diperlukan dihapus.

Teknik Pengujian

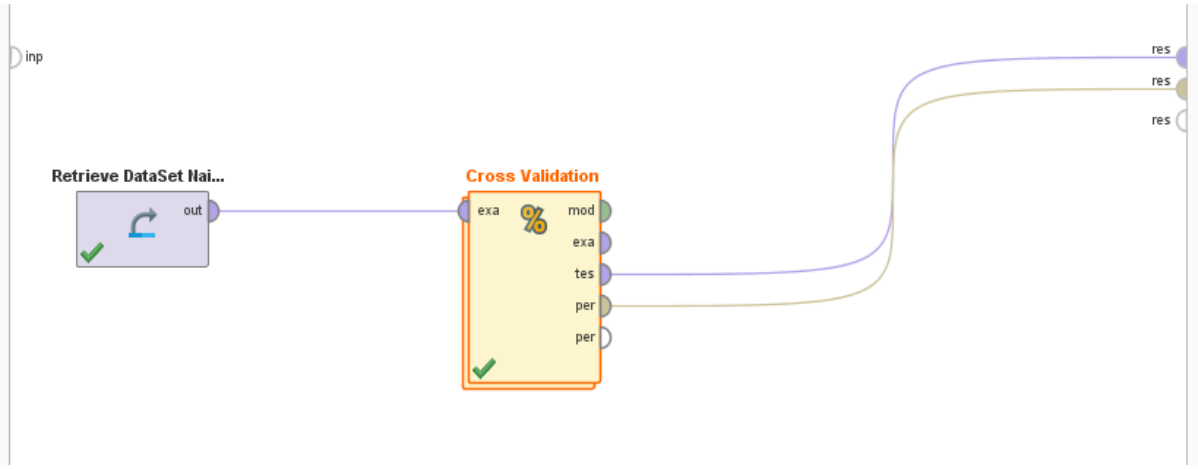
Teknik pengujian dalam penelitian ini yaitu dengan menggunakan pengujian *K-fold Cross Validation*. *K-fold Cross Validation* merupakan teknik untuk menghasilkan sebuah akurasi dengan cara membagi data set ke dalam data testing dan data training. Nilai *Cross Validation* yang digunakan pada pengujian yaitu *2-fold, 3-fold, 4-fold, 5-fold, 6-fold, dan 10-fold*.

HASIL DAN PEMBAHASAN

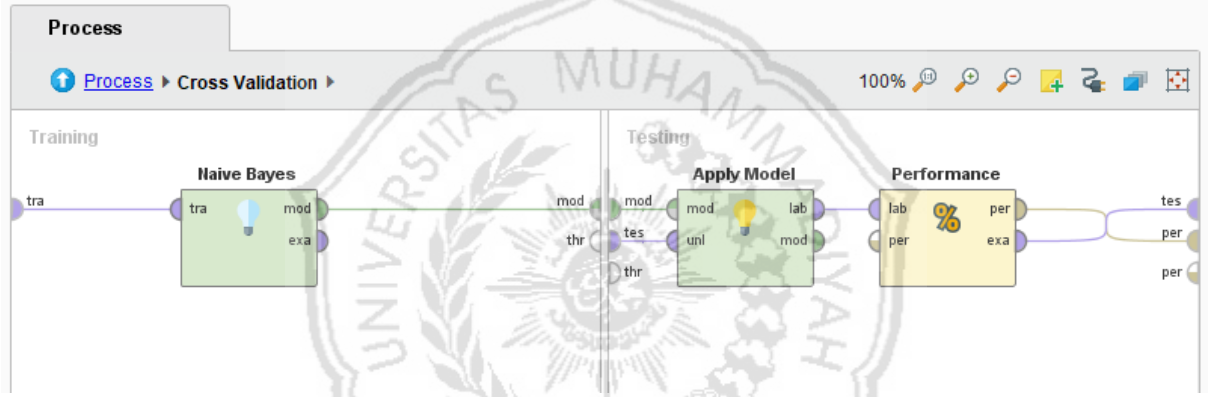
Data yang telah diperoleh, kemudian akan diolah untuk menghasilkan nilai klasifikasi yang nilai akhirnya yaitu berupa perbandingan Akurasi, dan Presisi dari algoritma *Naive Bayes* dan algoritma *K-Nearest Neighbor* pada dataset penyakit jantung.

Implementasi Algoritma *Naive Bayes* dan *K-Nearest Neighbor*.

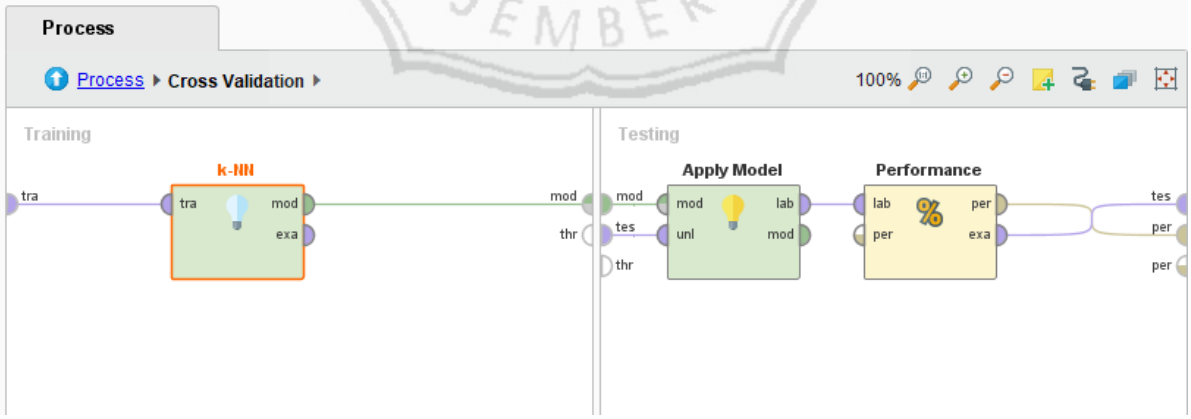
Pengujian dalam penelitian ini menggunakan teknik *Cross Validation* maka *output* dari *Retrieve Dataset* perlu dihubungkan dengan *port example* pada *Cross Validation.* Serta *port output test result* dan *performance* pada *Cross Validation* juga dihubungkan pada *port result* dari *output Process* untuk menampilkan hasil prediksi dan menampilkan hasil akurasi, presisi, dan recall dari pengujian dataset penyakit jantung.



Gambar 1 Tampilan teknik pengujian *Cross Validation*



Gambar 2 Proses dan Pengujian Algoritma *Naive Bayes*



Gambar 3 Proses dan Pengujian Algoritma *K-Nearest Neighbor*

Pengujian Algoritma *Naive Bayes*

Hasil dari pengujian dataset penyakit jantung dengan Algoritma *Naive Bayes* pada aplikasi *RapidMiner Studio* didapatkan hasil sesuai pada tabel berikut:

Tabel 2 Hasil Pengujian Algoritma *Naive Bayes*

Cross Validation	Akurasi	Presisi
K2	78,00%	77,78%
K3	85,00%	78,18%
K4	84,00%	82,35%
K5	86,67%	91,67%
K6	84,00%	85,71%
K10	90,00%	86,67%

Pengujian Algoritma *K-Nearest Neighbor*

Hasil dari pengujian Algoritma *K-Nearest Neighbor* pada dataset penyakit jantung menggunakan Aplikasi *RapidMiner Studio* mengugunakan teknik uji *Cross Validation* didapatkan hasil akurasi dan presisi dari masing-masing jumlah ketetanggaan sesuai dengan tabel berikut:

Tabel 3 Hasil Pengujian Algoritma *K-Nearest Neighbor*

K-Nearest Neighbor	Akurasi	Presisi
K=3	80,00%	78,57%
K=5	75,00%	93,33%
K=7	80,00%	90,00%
K=9	76,67%	81,82%

Selanjutnya peneliti melakukan perbandingan terhadap hasil pengujian dengan mengambil nilai akurasi tertinggi dari masing-masing algoritma dengan tujuan untuk mengetahui algoritma mana yang sesuai untuk diterapkan pada data penyakit jantung.

Tabel 3 Hasil Pengujian Perbandingan Algoritma *Naive Baye*s dengan Algoritma *K-Nearest Neighbor*

Algoritma	Perbandingan	
	Akurasi	Presisi
Naive Bayes	90,00%	86,67%
K-Nearest Neighbor	80,00%	90,00%

KESIMPULAN DAN SARAN

Kesimpulan

1. Berdasarkan dari seluruh pengujian yang telah dilakukan, didapatkan nilai *K* optimal untuk algoritma *K-Nearest Neighbor* yaitu dengan jarak ketetanggaan sebesar 7 dengan menggunakan teknik pengujian *10-fold Cross Validation* pada folder data pengujian ke-4 yang menghasilkan akurasi sebesar 80.00% dan presisi sebesar 90.00%.
2. Berdasarkan hasil pengujian yang dilakukan, algoritma *Naive Bayes* menghasilkna performa tertinggi pada *10-fold Cross Validation* pada folder data pengujian ke-4 dengan akurasi yaitu 90.00%, dan nilai presisi sebesar 86,67%. Algoritma *K-Nearest Neighbor* juga menghasilkan akurasi tertinggi dengan jumlah ketetanggaan sebesar 7 pada pengujian *10-fold Cross Validation* dengan akurasi tertinggi pada folder data pengujian ke-4 yaitu sebesar 80.00%, dan presisi sebesar 90.00%.
3. Berdasarkan pengujian yang telah dilakukan, didapatkan kesimpulan bahwa algoritma *Naive Bayes* lebih akurat dan lebih baik dalam klasifikasi penyakit jantung dibandingkan algoritma *K-Nearest Neighbor*.

Saran

1. Percobaan dengan menggunakan data penyakit yang berbeda dengan jumlah record, atribut dan parameter yang lebih banyak dan variatif agar hasil akurasi yang dihasilkan dapat berbeda dengan penelitian yang telah peneliti lakukan.
2. Percobaan dengan membandingkan algoritma selain *Naive Bayes* dan *K-Nearest Neighbor*, seperti Algoritma *Artificial Neural Network*, dan Algoritma *Decision Tree.*

DAFTAR PUSTAKA

Aprilla C, D., Aji Baskoro, D., Ambarwati, L., & Wayan Simri Wicaksana, I. (2013). *Belajar Data Mining dengan RapidMiner*. (R. Sanjaya, Ed.). Jakarta: Open Content Model.

Arisandy, D. L. (2017). *Analisis Perbandingan Algoritma Naive Bayes dan Algoritma C4.5 Untuk Klasifikasi Multi Data*.

Hospital Authority. (2016). Coronary Heart Disease Indonesian. *Smart Patien*.

Huda, N. M. (2010). *Aplikasi Data Mining Untuk*

Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus di Fakultas MIPA Universitas Diponegoro). Universitas Diponegoro.

Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Offset.

Medistra. (2017). Apakah Penyebab Dari Penyakit Jantung Koroner ? Retrieved April 11, 2018, from https://www.medistra.com/index.php?option=com_content&view=article&id=76

Mustakim, & Oktaviani, G. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. *Jurnal Sains, Teknologi Dan Industri*, *13*(2).

Novianti, A. G., & Prasetyo, D. (2017). Penerapan Algoritma K-Nearest Neighbor (K-NN) untuk Prediksi Waktu Kelulusan Mahasiswa. *Seminar Nasional APTIKOM*.

Nurhayati, H., & Nugroho, F. (2012). Implementasi Fuzzy Expert System Untuk Diagnosis Penyakit Jantung. *Implementasi Fuzzy Expert System Untuk Diagnosis Penyakit Jantung*.

Oemiati, R., & Rustika. (2015). Faktor Risiko Penyakit Jantung Koroner (PJK) Pada Perempuan (Baseline Studi Kohor Faktor Risiko PTM). *Buletin Penelitian Sistem Kesehatan*, *18*(1), 47–55.

Oktaviana, A. R. (2016). *Penerapan Data Mining Klasifikasi Pola Nasabah Menggunakan Algoritma C4.5 Pada Bank BRI Batang*. UniversitasDian Nuswantoro.

Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: Andi Offset.

Rifai, B. (2013). Algoritma Neural Network untuk Prediksi Penyakit Jantung. *Techno Nusa Mandiri*, *IX*(1), 1–9.

Rohman, A. (2012). Model Algoritma K-Nearest Neighbor (K-NN) untuk Prediksi Kelulusan Mahasiswa. *Dosen Jurusan Elektronika Fakultas Teknik Universitas Pandanaran Semarang*.

Samiadi, L. A. (2018). Penyakit Angina. Retrieved December 8, 2018, from https://hellosehat.com/penyakit/angina/

Sani, R. R., Zeniarja, J., & Luthfiarta, A. (2016). Penerapan Algoritma K-Nearest Neighbor pada Information Retrieval dalam Penentuan Topik Referensi Tugas Akhir. *Journal of Applied Intelligent System*, *1*(2).

Sartika, D., & Sensuse, D. I. (2017). Perbandingan Algoritma Klasifikasi Naive Bayes , Nearest Neighbour , dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian. *Jatisi*, *1*(2).

Sehat, D. (2017). Jantung Koroner – Penyebab, Gejala, Diagnosa, Pengobatan, dan Pencegahan. Retrieved April 11, 2018, from http://doktersehat.com/jantung-koroner/

Solichin, A. (2017). Mengukur Kinerja Algoritma Klasifikasi dengan Confusion Matrix. Retrieved December 8, 2018, from http://achmatim.net/2017/03/19/mengukur-kinerja-algoritma-klasifikasi-dengan-confusion-matrix/

Sumarlin. (2015). Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM. *Jurnal Sistem Informasi Bisnis*, *01*.

Wardhani, R. S. (2014). Aplikasi Sistem Fuzzy Untuk Diagnosa Penyakit Jantung Koroner (Coronary Heart Disease). *Universitas Negeri Yogyakarta*.

Wihardi, Y. (2013). K-Folds Cross Validation. Retrieved April 25, 2018, from http://blog.yayaw.web.id/riset/k-folds-cross-validation

Wurdianarto, S. R., Novianto, S., & Rosyidah, U. (2014). Perbandingan Euclidean Distante dengan Canberra Distance Pada Fce Recognition, *13*(1), 31–37.