SCIENCE AND INFORMATION TECHNOLOGY
SINTECH
JOURNAL

SINTECH JOURNAL | ISSN 2598-7305 | E-ISSN 2598-9642
Vol. 5 No 2 – Oktober 2022 | https://s.id/sintechjournal
DOI : https://doi.org/10.31598
Publishing : Prahasta Publisher | Manage by DRPM INSTIKI

Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke

Yufis Azhar[1], Aidia Khoiriyah Firdausy[2], Putri Juli Amelia[3]

[1,2,3]Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Malang

e-mail: yufis@umm.ac.id[1], aidiakhoiriyah@webmail.umm.ac.id[2], putrijuliaml@webmail.umm.ac.id[3]

Received : October, 2022	Accepted : October, 2022	Published : October, 2022

Abstract

Data mining is often called knowledge Discovery in Database (KDD). Data mining is usually used to improve future decision making based on information obtained from the past. For example for prediction, estimation, association, clustering, and description. Stroke is the second most deadly disease in the world according to WHO. The sufferer has an injury to the nervous system. Because of this, health experts, especially in the field of nursing, need special attention. Currently, the development of the Industrial Revolution Era 4.0 is collaborating in the fieldsof technology and health science so that it becomes something useful by using Machine Learning. There are so many benefits that are used in predicting several diseases that can be anticipated. In this study the dataset is dividedinto 2 parts, namely training data and testing data using split validation. Based on the results of the test that have been carried out in this study, the algorithm that has the highest accuracyvalue on balanced data is Logistic Regression with an accuracy rate of 75.65%, while for unbalanced data, the algorithm that has the highest accuracy results is Logistic Regression, Random Forest, SVM, and KNN with an accuracy rate of 98.63%. This testing process is carried out to identify stroke with data mining algorithms.
Keywords: *data mining, stroke, prediction*

Abstrak

"Data mining sering disebut Knowledge Dicovery in Database (KDD). Data mining biasanya digunakan untuk memperbaiki pengambilan keputusan dimasa yang akan datang berdasarkan informasi yang diperoleh dari masa lalu. Misalnya untuk prediksi, estimasi, assosiasi, clustering, dan deskripsi. Stroke adalah penyakit paling mematikan nomor dua di dunia menurut WHO. Penderitanya mengalami cedera pada system saraf. Karena hal inilah para pakar Kesehatan khususnya dibidang keperawatan memerlukan perhatian khusus. Saat ini perkembangan Era Revolusi Industri 4.0 yang berkolaborasi di bidang teknologi dan ilmu Kesehatan sehingga menjadi sesuatu yang bermanfaat dengan menggunakan Machine Learning. Banyak sekali manfaat yang digunakan dalam memprediksi beberapa penyakit yang dapat diantisipasi. Dalam penelitian dataset dibagi menjadi 2 bagian, yaitu data training dan data testing dengan menggunakan split validation. Berdasarkan hasil pengujian yang telah dilakukan pada penelitian ini, algoritma yang mempunyai nilai akurasi tertinggi yajitu Logistic Regression, Random Forest, SVM, dan KNN dengan tingkatb akurasi sebesar 98,63%. Proses pengujian ini dilakukan untuk mengindentifikasi penyakit stroke dengan data mining".
Kata Kunci: *data mining, stroke, prediksi*

1. PENDAHULUAN

Stroke adalah penyebab kematian dan kecacatan global utama. Diagnosis didasarkan pada karakteristik klinis dan pencitraan otak untuk membedakan antara stroke iskemik dan perdarahan intraserebral. [1] Penyedia layanan kesehatan di Indonesia mendiagnosis 43,1% kasus stroke yang melibatkan individu berusia

di atas 75 tahun dan 0,2% kasus yang melibatkan individu berusia antara 15 dan 24 [2]. Stroke adalah perubahan otak yang tiba-tiba, lebih dari 24 jam function.atau menyebabkan kematian.Ini mempengaruhi cara kerja otak secara lokal dan global [3].

Stroke merupakan salah satu penyakit paling umum di Indonesia dan salah satu penyebab kematian utama. Hal ini didasarkan pada data yang mencakup 41.590 kematian pada tahun 2014 dari sampel perwakilan warga negara Indonesia. Tenaga medis profesional dan tenaga terlatih melakukan otopsi verbal secara real time sesuai dengan pedoman WHO untuk setiap kematian ini [4]. Untuk membuat penilaian berdasarkan data klinis yang luas, teknik data mining sangat penting di bidang kesehatan. Penambangan data adalah tindakan menggunakan alat atau pendekatan tertentu untuk mencari pola atau informasi yang menarik dalam data yang dipilih. Penggunaan data mining diharapkan dapat memberikan informasi yang dapat digunakan untuk mencegah atau mengurangi penderitaan stroke di Indonesia dan di seluruh dunia, sehingga menurunkan jumlah korban stroke secara keseluruhan [5].

Orang-orang masih belum sepenuhnya memahami sifat dari kondisi stroke saat ini, dan banyak yang tidak menyadari tanda-tanda awal yang mungkin ada. Selain itu, kebanyakan orang enggan pergi ke rumah sakit hanya untuk menanyakan gejala yang mereka alami. Ini terus menjadi wabah, menyebabkan peningkatan pesat dalam frekuensi stroke dan menghantui kehidupan masyarakat. [6]

Ini mempromosikan banyak penelitian stroke, salah satunya menggunakan teknik berbasis komputer. Dengan bantuan algoritme tertentu, strategi ini dapat menangani kumpulan data besar untuk memberikan prediksi yang lebih cepat dan tepat. Untuk mengurangi kesalahan (perbedaan antara apa yang terjadi dan apa yang diproyeksikan), prediksi adalah tindakan memprediksi sesuatu secara metodis berdasarkan pengetahuan historis dan sekarang yang tersedia. [7]

Dataset bank sentral menunjukkan bahwa strategi dasar yang disarankan (algoritma K Nearest Neighbor dan Naïve Bayes), dengan nilai akurasi 89,58%, mengungguli pendekatan lain dengan data yang sama, menurut studi sebelumnya oleh Elma Zannatul Ferdousy et al.[8]

Tujuan dari penelitian yang telah dibuat adalah menggunakan teknik data mining, untuk memprediksi secara akurat jumlah pasien yang terkena penyakit stroke.

2. METODE PENELITIAN

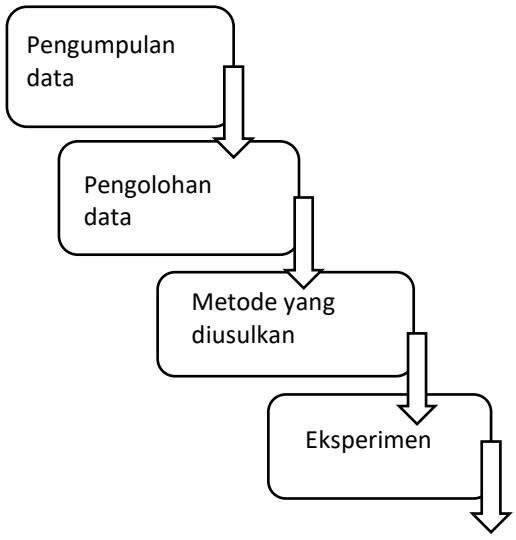
A. Penyakit Stroke

Stroke adalah serangan mendadak pada otak yang mengakibatkan gangguan aliran darah yang disebabkan oleh penyumbatan atau pecahnya satu atau lebih pembuluh darah otak, yang baik sebagian atau seluruhnya mengganggu fungsi otak. Sel-sel otak mulai mengalami kekurangan darah, oksigen, atau nutrisi sebagai akibatnya, dan mereka akhirnya dapat mati dengan cepat[9].

Organisasi Kesehatan Dunia melaporkan bahwa kelemahan wajah yang tiba-tiba dan bahkan mati rasa di lengan, kaki, atau di satu sisi tubuh adalah gejala yang paling sering dialami. Selain itu, fungsi sensorik tubuh hilang, dan sakit kepala yang sangat parah dirasakan, yang dapat menyebabkan pingsan atau tidak sadarkan diri. [10]

B. Data Mining

Data mining adalah kumpulan prosedur yang digunakan untuk menyelidiki nilai tambah dari kumpulan data dalam bentuk informasi yang belum ditemukan sebelumnya. Ada berbagai langkah dalam pipeline data mining, diantaranya [11]:



Evaluasi dan Validasi Hasil

1. Data Collection
   Data collection adalah proses mengumpulkan, mengukur, dan menganalisis berbagai jenis data dengan menggunakan beberapa metode. Tujuan utama data collection adalah untuk mengumpulkan data dan informasi yang paling dapat diandalkan serta menganalisis data untuk membuat keputusan nilai bisnis yang krusial. Setelah mengumpulkan informasi data masuk ke proses lain yaitu pembersihan dan pemrosesan data untuk digunakan oleh perusahaan.

2. Feature Extraction and Data Cleaning
   Feature Extraction adalah metode mengekstraksi karakteristik atau karakteristik dari suatu bentuk, dengan nilai yang diperoleh kemudian diperiksa untuk pemrosesan tambahan. Sedangkan pembersihan data adalah proses untuk memastikan keakuratan, konsistensi, dan kegunaan data yang terkandung dalam kumpulan data setelah data dikumpulkan, sangat penting untuk mengubah data ke format yang benar jika ada banyak jenis data yang tidak sesuai. untuk diproses. Ini akan memungkinkan data untuk diproses.

3. Analytical processing and algorithms
   Merancang analisis yang efisien dari data yang dihasilkan adalah langkah terakhir dalam proses data mining.

4. Preprocessing
   Pemrosesan data adalah proses mengubah data yang belum diproses menjadi format yang lebih mudah dipahami. Prosedur ini sangat penting karena data mentah seringkali tidak memiliki format standar. Selain itu, penambangan data tidak dapat menangani data mentah, sehingga prosedur ini harus diselesaikan untuk membuat yang berikut ini lebih sederhana.

5. Data Integration
   Salah satu metode untuk menggabungkan data milik korporasi dari beberapa sumber database adalah integrasi data.

6. Data Reduction
   Dengan mengintegrasikan atau menghapus data asing, reduksi data dapat digunakan untuk meminimalkan ukuran data.

7. Data Transformation
   Tujuan utama dari transformasi data adalah untuk mengubah skala pengukuran data asli menjadi sesuatu yang lain sehingga data tersebut dapat memenuhi asumsi analisis varians yang mendasarinya.

C. Decision Tree C4.5

Patokan untuk algoritma pembelajaran terawasi yang baru adalah metode C4.5, yang merupakan pengembangan dari algoritma ID3. Algoritma C4.5 bersifat prediktif dan digunakan untuk mengklasifikasikan atau mengelompokkan segmen. Sebuah pohon keputusan berfungsi sebagai dasar untuk prediksi algoritma C4.5. Daun pohon keputusan mewakili kelas atau segmen, sementara cabangnya menangani masalah kategorisasi. [12] Rasio perolehan pengotor digunakan untuk menilai semua karakteristik dalam pohon keputusan C4.5.Dengan menggunakan rumus, tentukan nilai gain sebagai berikut:

$$Gain(S,A) = Enthropy(S) - \sum_{i=1}^{n} - \frac{|Si|}{|S|} * Enthopy(Si)$$

1. S= Himpunan Kasus
2. A= Fitur
3. N= Jumlah partisi atribut A

4. |Si|= Proporsi Si terhadap S
5. |S|= Jumlah kasus dalam s

D. Logistic Regression

Tujuan dari analisis regresi adalah untuk memastikan bagaimana satu variabel mempengaruhi variabel lain. Model regresi linier langsung dalam bentuk [12] adalah model regresi yang paling sederhana.

$$Y = \beta 0 + \beta 1 + \varepsilon$$

Dimana :

1. Y= variable terikat (nilai yang diprediksi)
2. X= variable bebas
3. β0= konstanta
4. β1= koefisien regresi (nilai peningkatan ataupun penurunan)
5. ε= galat acak

E. Random Forest

Hutan Acak (RF) adalah kumpulan pohon keputusan di mana setiap pohon bergantung pada hasil sampel vektor acak terpisah yang diambil dari distribusi yang sama. Kekuatan RF terletak pada pemilihan karakteristik secara acak yang menghasilkan tingkat kesalahan yang relatif rendah.

F. Support Vector Machine (SVM)

Metode prediksi untuk klasifikasi dan regresi disebut Support Vector Machine (SVM). Dengan menggabungkan konsep kernel ke dalam ruang kerja berdimensi tinggi, SVM dirancang untuk memecahkan masalah non-linear. SVM mencakup instance klasifikasi yang dapat dipisahkan secara linier(), yang merupakan dasar dari pengklasifikasi linier.

G. K-Nearest Neighbours

Metode K-Nearest Neighbor digunakan untuk mengklasifikasikan objek berdasarkan data pembelajaran sebagai tetangga terdekat atau memiliki perbedaan nilai yang kecil dengan item tersebut. Dengan menggunakan karakteristik, data uji, dan data latih objek baru, algoritma ini mencoba mengklasifikasikannya .dalam kaitannya dengan karakteristik yang ditentukan oleh jarak Euclidean.Jika contoh pertama adalah (a1, a2, a3,..., an) dan contoh kedua adalah (b1, b2, b3,..., bn), maka persamaan berikut dapat digunakan untuk menentukan jarak antara keduanya:

$$d = \sqrt{(a1 + b1)2 + (a2 + b2)2 \ldots (an + bn)2}$$

Penjelasan rumus :

1. A: data uji, yang digunakan untuk mengevaluasi model yang dibangun menggunakan data latih.

2. B: Model terbaik dipilih menggunakan data pelatihan data ke-n: N

H. Naïve Bayes

Kategorisasi probabilitas berdasarkan Teorema Bayes disebut naive Bayes. Untuk membuat perhitungan yang diperlukan lebih mudah, Nave Bayes memperhitungkan dampak dari nilai atribut lainnya. rumus probabilitas Bayesian yang luas. [13]

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Keterangan:

1. $P(C_i|X)$: Probabilitas C_i jika diberi bukti X
2. $P(C_i)$: Probabilitas C_i tanpa memandang bukti apapun
3. $P(X|C_i)$Probabilitas X terjadi akan mempengaruhi C_i
4. P(X) : Probabilitas X tanpa memandang bukti apapun

3.HASIL DAN PEMBAHASAN

Metode yang akan digunakan untuk memprediksi data stroke dan mempelajari karakteristik pasien atau tes untuk mempermudah penelitian sehingga dapat berjalan dengan lancar dan metodis serta mencapai tujuan yang telah ditetapkan. Langkah-langkah berikut termasuk dalam tahapan penelitian ini yang diselesaikan:

A. **Analisis Data**

Data diperoleh dari situs website Kaggle sebagai acuan data dalam

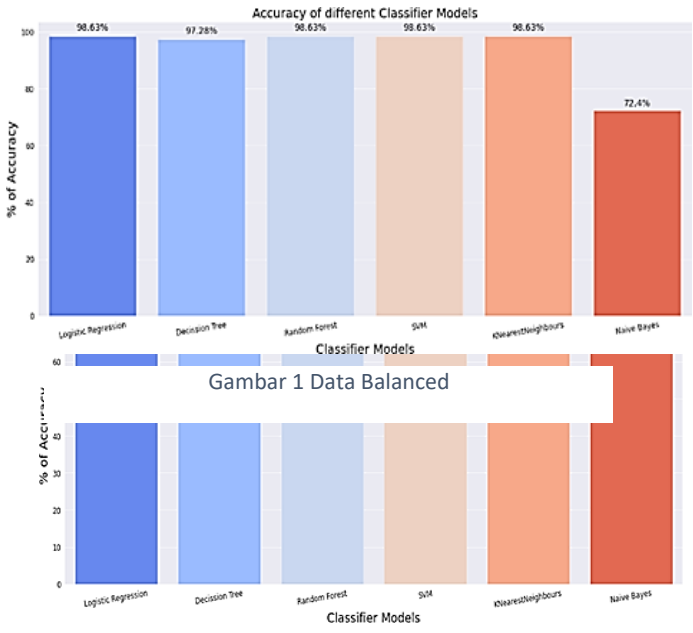
penelitian ini dan data diproses menjadi beberapa proses. Data Pasien ID, Jenis Kelamin, Usia, Hipertensi, Penyakit Jantung, Status Perkawinan, Jenis Pekerjaan, Tempat Tinggal, Glukosa, dan Indeks Massa yang dikumpulkan sebanyak 12 kolom yang dikumpulkan dibagi menjadi dua bagian: Data Testing adalah data yang digunakan untuk menguji algoritma tanpa memasukkan informasi diagnostik apa pun.Hasil akurasi Algoritma Data Mining tahapan dalam penelitian.

3.1 Hasil akurasi Algoritma Data Mining tahapan dalam penelitian

Table 1 Informasi Dataset

ID	Unique Identifier
GENDER	"Male", "Female", or "Other"
AGE	Age of the patient
HYPERTENSION	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
HEART-DISEASE	0 if the patient doesn't have any heart diseases, if the patient has a heart disease
EVER_MARRIED	"No" or "Yes"
WORK TYPE	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
RESIDENCE_TYPE	"Rural" or "Urban"
AVG_GLUCOSE-LEVEL	Average glucose level in blood
BMI	Body mass index
SMOKING_STATUS	"formerly smoked", "neversmoked", "smokes"or "Unknown"
STROKE	1 if the patient had a stroker or 0 if not

A. Pada eksperimen pertama, dilakukan uji coba data balanced dengan menggunakan algortima Logistic Regressiom dihasilkan akurasi 73.48%
B. Pada eksperimen kedua, dilakukan uji coba data balanced dengan menggunakan algoritma Decision Tree dihasilkan akurasi 68.70% sedangkan untuk data unbalanced sebesar 97.28%.
C. Pada eksperimen ketiga, dilakukan uji coba data balanced dengan menggunakan algoritma Random Forest dihasilkan akurasi 72.17% sedangkan untuk data unbalanced sebesar 98.63%.
D. Pada eksperimen keempat dilakukan uji coba data balanced dengan menggunakan algoritma Support Vector Machine (SVM) dihasilkan akurasi 76.52% sedangkan untuk data unbalanced sebesar 98.63%.
E. Pada eksperimen kelima, dilakukan uji coba data balanced dengan menggunakan algoritma K-Nearest Neighbours dihasilkan akurasi 72.61% sedangkan untuk data unbalanced sebesar 98.63%.
F. Pada eksperimen keenam, dilakukan uji coba data balanced dengan menggunakan algoritma Naïve Bayes dihasilkan akurasi 66.52% sedangkan untuk data unbalanced sebesar 72.40%. Dari hasil akurasi data mining dapat diringkas seperti dibawah unbalanced sebesar 98.63%.



Gambar 1 Data Balanced

Gambar 2 Data Unbalanced

4. KESIMPULA

Hasil penelitian algoritma yang mempunyai nilai akurasi tertinggi pada data yang diseimbangkan yaitu Logistic Regression dengan tingkat akurasi sebesar 75.65%, sedangkan pada data yang belum diseimbangkan, algoritma yang memiliki hasil akurasi tertinggi yaitu Logistic Regression, Random Forest, SVM, dan KNN dengan tingkat akurasi sebesar 98.63%. Sedangkan untuk nilai akurasi algoritma data mining terdapat dua data yaitu data balanced dimana nilai akurasi untuk Random Forest 72.17%, SVM 76.52%, KNN 72.61%. Pada data unbalanced dengan nilai akurasi Random Forest, SVM dan KNN memiliki nilai akurasi 98.63%. Karena berdasarkan pengecekan akurasi dengan confusion matrix diatas adalah kita dapat mengetahui perbandingan jumlah TRUE POSITIF, TRUE NEGATIF, FALSE POSITIF, dan FALSE NEGATIF dari kedua buah model. Berdasarkan studi kasus ini, model yang memprediksi lebih banyak pasien yang stroke (TRUE POSITIF) lebih baik karena artinya model dapat memprediksi kecendrungan pasien yang memiliki peluang besar mengidap stroke walaupun sebenarnya dia di diagnose belum atau tidak mengidap stroke.

PERNYATAAN PENGHARGAAN

Penulis berharap penelitian yang selanjutnya akan menggunakan metode pembelajaran mesin yang berbeda dengan yang digunakan pada penelitian ini untuk meningkatkan akurasi.

DAFTAR PUSTAKA

[1] A. Byna and M. Basit, "Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 9, no. 3, pp. 407–411, 2020, doi: 10.32736/sisfokom.v9i3.1023.

[2] U. Amelia *et al.*, "IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE (SVM) UNTUK PREDIKSI PENYAKIT STROKE DENGAN ATRIBUT BERPENGARUH," vol. III, pp. 254–259, 2022.

[3] R. S. Rohman, R. A. Saputra, and D. A. Firmansaha, "Komparasi Algoritma C4.5 Berbasis PSO Dan GA Untuk Diagnosa Penyakit Stroke," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 1, p. 155, 2020, doi: 10.24114/cess.v5i1.15225.

[4] R. E. Pambudi and F. I. Komputer, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C.45 1,2,3," vol. 16, no. x, pp. 221–226, 1978.

[5] D. Andri and M. Reza, "Penerapan Algoritma K-Nearest Neighbord Untuk Prediksi Kematian Akibat Penyakit Gagal Jantung," vol. III, no. 2020, pp. 105–112, 2022.

[6] F. Karim, G. W. Nurcahyo, and S. Sumijan, "Sistem Pakar dalam Mengidentifikasi Gejala Stroke Menggunakan Metode Naive Bayes," *J. Sistim Inf. dan Teknol.*, vol. 3, pp. 221–226, 2021, doi: 10.37034/jsisfotek.v3i4.69.

[7] F. Akbar, H. W. Saputra, and A. K. Maulaya, "Implementation of Decision Tree Algorithm C4 . 5 and Support Vector Regression for Stroke Disease Prediction Implementasi Algoritma Decision Tree C4 . 5 dan Support Vector Regression untuk Prediksi Penyakit Stroke," vol. 2, no. October, pp. 61–67, 2022.

[8] A. Puspitawuri, E. Santoso, and C. Dewi, "Diagnosis Tingkat Risiko Penyakit Stroke Menggunakan Metode K-Nearest Neighbor dan Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 4, pp. 3319–3324, 2019, [Online]. Available: e-issn: 2548-964X http://j-ptiik.ub.ac.id.

[9] S. L. Pratiwi and H. Cahyono, "View metadata, citation and similar papers at core.ac.uk," *PENGARUH Pengguna. PASTA LABU KUNING (Cucurbita Moschata) UNTUK SUBSTITUSI TEPUNG TERIGU DENGAN PENAMBAHAN TEPUNG ANGKAK DALAM PEMBUATAN MIE KERING*, vol. 1, no. 2, pp. 274–282, 2020.

[10] N. Azwanti and E. Elisa, "Analisis Pola Penyakit Hipertensi Menggunakan Algoritma C4.5," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 3, no. 2, pp. 116–123, 2019, doi: 10.30743/infotekjar.v3i2.944.

[11] E. I. Utami Indarto; Raharjo, Suwanto, "Prediksi Risiko Kematian Pasien Stroke Perdarahan Dengan Menggunakan

Teknik Klasifikasi Data Mining," *Inf. Interaktif*, vol. 5, no. Vol 5, No 2 (2020): Jurnal Informasi Interaktif, pp. 86–91, 2020, [Online]. Available: http://e-journal.janabadra.ac.id/index.php/informasiinteraktif/article/view/1172/790.

[12] Y. Tampil, H. Komaliq, and Y. Langi, "Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado," *d'CARTESIAN*, vol. 6, no. 2, p. 56, 2017, doi: 10.35799/dc.6.2.2017.17023.

[13] D. Prajarini, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit," *Informatics J.*, vol. 1, no. 3, p. 137, 2016, [Online]. Available: http://jurnal.unej.ac.id/index.php/INFORMAL/article/view/3424.