SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK I - Surabaya, 26 Juni 2021
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2021.1818

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Klasifikasi pada Dataset Penyakit Hati Menggunakan Algoritma Support Vector Machine, K-NN, dan Naïve Bayes

Citra Nurina Prabiantissa

Institut Teknologi Adhi Tama Surabaya
e-mail: citranurina@itats.ac.id

ABSTRACT

The liver has a crucial function, namely the metabolic center that functions to maintain the needs of the brain and as a blood filter from harmful substances that come from the intestines. Liver disease can arise due to liver function abnormalities, namely liver, hepatitis, liver cancer, liver cirrhosis, and other liver diseases. Complex liver diseases such as liver cancer require precise and accurate patient screening. This classification system helps doctors to determine liver disease in patients. This system is expected to reduce misdiagnosis in patients and doctors can perform treatment actions accurately. The first process is to clean the liver dataset by overcoming missing values and detecting outliers. After cleaning the data, the data is classified by three different algorithms. The classification algorithms are Naïve Bayes, KNN, and SVM. The performance of the three methods was compared to get the best method for the liver dataset, by determining the average of its accuracy, precision, recall, and F-measure. The results showed that from three algorithms, SVM has the best performance with an average of 82.36%.

Keywords: *Classification*; *KNN; Liver Disease*; *Naïve Bayes; SVM.*

ABSTRAK

Hati memiliki fungsi yang krusial, yaitu pusat metabolisme yang berfungsi untuk menjaga kebutuhan otak dan sebagai filter darah dari zat-zat yang berbahaya yang datang dari usus. Penyakit hati yang dapat timbul karena kelainan fungsi hati ini adalah liver, hepatitis, kanker hati, sirosis hati, dan penyakit hati lainnya. Penyakit hati kompleks seperti kanker hati membutuhkan *screening* pasien dengan tepat dan akurat. Sistem klasifikasi ini membantu dokter untuk menentukan penyakit hati pada pasien. Sistem ini diharapkan dapat mengurangi kesalahan

diagnosis pada pasien dan dokter dapat melakukan tindakan pengobatan dengan akurat. Proses pertama ialah dengan melakukan pembersihan *dataset* hati yaitu dengan mengatasi *missing value* dan mendeteksi *outlier*. Setelah pembersihan data, data diklasifikasi dengan tiga algoritma yang berbeda. Algoritma klasifikasi tersebut yaitu Naïve Bayes, KNN, dan SVM. Performa dari ketiga metode tersebut dibandingkan untuk mendapatkan metode yang terbaik untuk *dataset* hati, dengan cara menentukan nilai performanya yang berupa nilai *accuracy*, *precision*, *recall*, dan *F-measure*. Hasil penelitian menunjukkan bahwa dari ketiga algoritma, SVM memiliki rata-rata performa paling baik dengan akurasi sebesar 82,36%.

Kata kunci: Klasifikasi; KNN; Naïve Bayes; Penyakit Hati; SVM.

PENDAHULUAN

Organ hati mempunyai peranan penting pada tubuh manusia. Penelitian menunjukkan bahwa hati merupakan organ dengan berat mencapai 1,2 sampai 1,8 kg dari keseluruhan berat orang dewasa [1]. Hati memiliki fungsi yang krusial, yaitu pusat metabolisme, menjaga kebutuhan otak, dan sebagai filter darah dari zat-zat yang berbahaya yang datang dari usus.

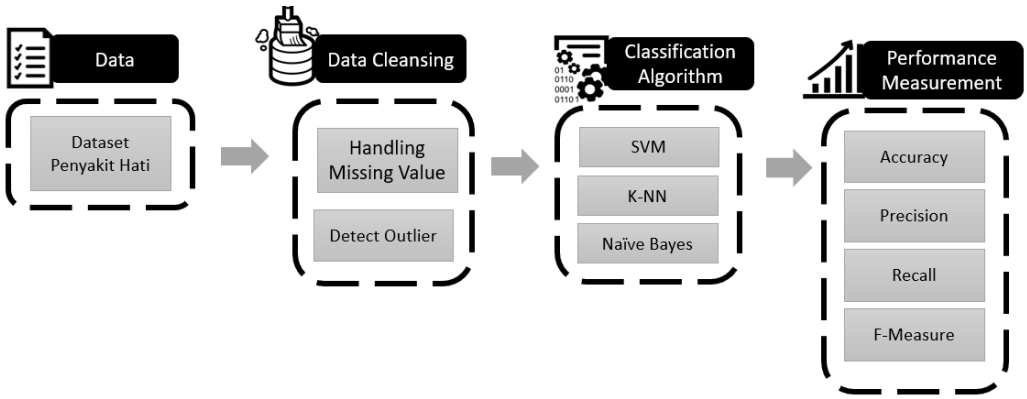
Ketidaksesuaian fungsi hati akan menimbulkan berbagai macam penyakit. Macam-macam penyakit hati yaitu liver, hepatitis, kanker hati, dan sirosis hati. Indonesia merupakan negara dengan 2,9 juta penduduk yang sekitar 0,6%-nya mengidap penyakit hepatitis. Kementerian Kesehatan sudah berupaya melakukan pencegahan dengan melakukan imunisasi HB 0–4 sejak bayi. Menurut data imunisasi dari Kementerian Kesehatan [2], terjadi peningkatan signifikan sampai melebihi target imunisasi, tetapi hal ini belum dapat mengubah angka pengidap penyakit hepatitis di Indonesia.

Gejala yang timbul pada penyakit hati dapat dibedakan menjadi dua, yaitu gejala klinis dan fisik [3]. Gejala klinis merupakan gejala yang dapat diketahui dengan melakukan *screening* tentang gejala yang dirasakan oleh pasien. Gejala fisik adalah gejala yang dapat diketahui dengan memeriksa tubuh pasien. Terkadang, dokter memiliki kendala jika penyakit hati tersebut sudah memiliki gejala yang kompleks sehingga dibutuhkan suatu sistem klasifikasi untuk menentukan seorang ini mengidap penyakit hati atau tidak. Sistem ini diharapkan dapat mengurangi kesalahan diagnosis pada pasien dan dokter dapat melakukan tindakan pengobatan dengan akurat.

Sistem klasifikasi ini bertujuan untuk membandingkan metode yang tepat pada *dataset* penyakit hati. *Dataset* tersebut memiliki atribut berbeda-beda sehingga penentuan klasifikasi dapat dilakukan secara akurat. Pada penelitian sebelumnya, ada beberapa solusi yang digunakan untuk menyelesaikan suatu permasalahan, yaitu dengan menerapkan algoritma klasifikasi. Salah satu penelitian tentang penentuan status gunung berapi menggunakan dua algoritma, yaitu KNN dan Naïve Bayes, menunjukkan bahwa KNN merupakan algoritma dengan performa terbaik [4]. Penelitian ini menggunakan penambahan satu algoritma klasifikasi lagi, yaitu Support Vector Machine (SVM). Metode-metode klasifikasi ini akan dibandingkan satu sama lain dengan melihat pada performa akurasinya.

METODE

Penelitian ini mempunyai beberapa proses yang berkesinambungan satu dengan yang lain. Proses dimulai dengan menyiapkan *dataset* penyakit hati, melakukan *data cleaning*, menggunakan algoritma klasifikasi, mengukur performa dari tiap algoritma, dan melakukan analisis dari hasil masing-masing performa algoritma klasifikasi. Gambar 1 menunjukkan serangkaian proses atau alur dari sistem yang akan dibuat.



Gambar 1. Diagram metode penelitian.

Data

Proses pertama pada diagram alur sistem Gambar 1 adalah data. Data perlu dipersiapkan dengan baik karena merupakan bagian penting yang mempengaruhi proses-proses selanjutnya. *Dataset* yang digunakan adalah dataset *heart disease* dari halaman web UCI Machine Learning Repository. Terdapat banyak *dataset* yang bisa digunakan dan diolah secara langsung. Data ini terdiri dari 14 *atrribute* dan 303 *instance*.

Data Cleansing

Data cleansing atau pembersihan data merupakan proses menganalisis data dengan cara mengubah atau menghapus data yang diakibatkan oleh kesalahan *input* (*human error*) dan kesalahan pada sistem [5]. *Data cleansing* ini penting sebagai langkah awal untuk seleksi data sehingga data yang terbentuk memiliki performa yang baik. Performa data yang baik akan mempengaruhi hasil dari penelitian yang dilakukan.

Teknik *data cleansing* dibagi menjadi dua proses, terdiri dari penanganan *missing value* dan deteksi *outlier*. *Missing value* ialah informasi yang hilang karena kesalahan sistem atau manusia atau ketidakmampuan responden saat memberikan jawaban pada saat survei [5].

Penanganan *missing value* dapat dilakukan dengan dua cara. Cara pertama yaitu menghapus keseluruhan baris dari nilai yang hilang. Kekurangan dari metode ini ialah jika data yang dimiliki sedikit, akan mempengaruhi keseluruhan performa dari data. Metode ini dapat digunakan apabila data yang dimiliki berjumlah banyak. Cara kedua untuk mengatasi *missing value* yaitu menghitung nilai pengganti (imputasi). Beberapa cara untuk menghitung nilai pengganti yaitu nilai yang merupakan bagian dari kumpulan data tersebut, misalnya angka 0 atau dengan menghitung nilai median dan rata-rata.

Teknik *data cleansing* yang kedua adalah deteksi *outlier*. Pada serangkaian data, terkadang memiliki data yang menyimpang dari data yang lain. Data ini adalah data *outlier* [6]. Nilai pada suatu data dapat dikatakan *outlier* jika memiliki nilai yang jauh berbeda dengan kelompoknya, contohnya jika nilai terlalu besar atau terlalu kecil.

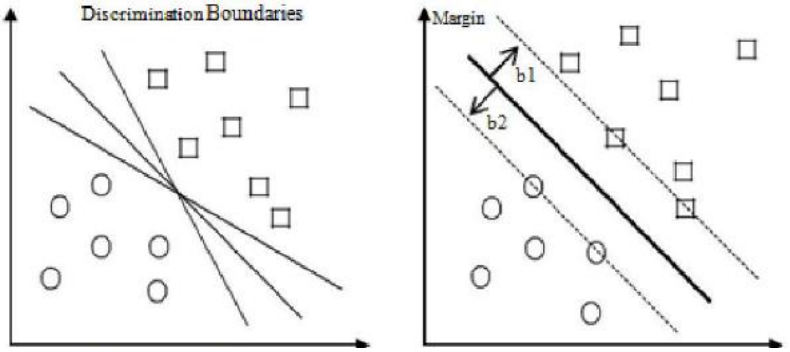
Algoritma Klasifikasi

Data yang sudah melalui proses *data cleansing* lalu diproses dengan algoritma klasifikasi. Berikut ini merupakan informasi dari ketiga algoritma yang digunakan.

1. *Support Vector Machine (SVM)*

   SVM adalah algoritma *supervised learning* yang digunakan untuk melakukan klasifikasi, regresi, maupun prediksi. SVM terbagi menjadi dua, yaitu SVM linear dan SVM non-linear. SVM linear adalah algoritma yang memiliki fungsi untuk klasifikasi permasalahan dengan

dua kelas sebagai usaha pencari *hyperplane* terbaik [7]. *Hyperplane* adalah garis pemisah antara dua kelas yang berbeda. Gambar 2 berikut ini merupakan ilustrasi dari *hyperplane*.



Gambar 2. *Hyperplane* pada SVM.

Sedangkan SVM non-linear menggunakan fungsi dari kernel ruang yang berdimensi tinggi. Jenis kernel SVM yang sering digunakan adalah polinomial, RBF, dan Sigmoid [8].

2. *K-Nearest Neighbor (k-NN)*

K-Nearest Neighbor (k-NN) ialah algoritma yang menggunakan klasifikasi data pembelajaran dengan cara menghitung jarak terdekat [9]. Berikut ini adalah tahapan dari algoritma ini.

a. Menentukan jumlah tetangga terdekat sesuai dengan jumlah K.
b. Menghitung jarak antara data *training* dan data *testing.*
c. Mengurutkan semua data berdasarkan jarak Euclidian terkecil dengan menggunakan Persamaan 1.

$$\text{Euc} = \sqrt{((x_1 - y_1)^2 + \ldots + (x_n - y_n)^2)} \quad (1)$$

dengan $x = x_1, x_2, \ldots, x_n$; $y = y_1, y_2, .., y_n$; dan nilai n sebagai nilai atribut.

d. Menentukan kelompok berdasarkan label terbanyak pada nilai K.

3. *Naïve Bayes*

Algoritma ini merupakan metode ketiga yang digunakan pada penelitian ini. Algoritma ini menggunakan probabilitas dengan pendekatan Bayesian [10]. Penggunaan algoritma Naïve Bayes menggunakan kombinasi dari probabilitas sebelumnya dan probabilitas bersyarat dengan menggunakan rumus yang digunakan untuk menghitung probabilitas yang mungkin [11]. Berikut ini perhitungan menggunakan algoritma Naïve Bayes.

$$P(H|X) = \frac{\text{P(H)P(X|H)}}{\text{P(X)}} \quad (2)$$

Performance Measurement

Pengukuran performa dari suatu sistem merupakan salah satu komponen penting. Pada penelitian ini, pengukuran performa menggunakan teknik *K-fold cross-validation*. *K-fold cross-validation* merupakan teknik untuk memvalidasi akurasi dari sebuah model menggunakan *dataset* yang digunakan. *Dataset* dibedakan menjadi data *testing* dan data *training*. Kemudian dilakukan eksperimen sejumlah K dan eksperimen dilakukan sesuai dengan partisi data yang dilakukan [12]. Pengukuran performa dilakukan dengan menghitung nilai *accuracy*, *precision*, *recall*, dan *F-measure.* Perhitungan dari empat cara tersebut adalah sebagai berikut.

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F-Measure} = 2 \times \frac{Precision\ x\ Recall}{Precision + Recall} \quad (6)$$

dengan TP adalah *true positive*, TN adalah *true negative*, FP adalah *false positive*, dan FN adalah *false negative* [4].

HASIL DAN PEMBAHASAN

Percobaan *dataset heart disease* yang menggunakan tiga algoritma memiliki hasil yang berbeda-beda. Algoritma klasifikasi yang digunakan mempengaruhi hasil *accuracy*, *precision*, *recall*, dan *F-measure*.

Tabel 2. Tabel performa algoritma klasifikasi.

Algoritma Klasifikasi	***Performance* (%)**				**Rata-Rata Performa (%)**
	Accuracy	***Precision***	***Recall***	***F-measure***	
Naïve Bayes	82,42	80,49	80,49	80,49	80,97
K-NN (K = 3)	68,13	65,00	63,41	64,20	65,18
K-NN (K = 5)	63,74	60,53	56,10	58,23	59,65
K-NN (K = 7)	63,74	61,76	51,22	56,00	58,18
SVM	**84,62**	**93,55**	**70,73**	**80,56**	**82,36**

Tabel 2 menunjukkan Naïve Bayes memiliki performa yang baik dengan menggunakan *dataset* penyakit hati yang memiliki rata-rata persentase 80,97%. Performa ini sangat baik jika dibandingkan dengan algoritma K-NN. Pada algoritma K-NN, dilakukan tiga skenario yang berbeda. Perbedaan terletak pada jumlah K yang digunakan. Pada algoritma KNN dengan jumlah K = 3 memiliki rata rata performa 65,18%, dengan jumlah K = 5 memiliki rata-rata performa 59,65%, dan dengan jumlah K = 5 memiliki rata rata performa 58,18%. Skenario ini dilakukan untuk melihat apakah jumlah K membuat performa menjadi lebih baik atau semakin menurun. Untuk algoritma SVM, performa rata-ratanya adalah 82,36% yang merupakan rata-rata paling baik di antara metode lainnya.

KESIMPULAN

Penelitian ini menggunakan beberapa proses pengolahan data sebelum menentukan algoritma yang paling baik. Kemudian dilakukan beberapa proses pada *dataset* seperti menghilangkan *missing value* dan menghilangkan *outlier*. Setelah data di-*cleansing*, dilanjutkan dengan menerapkan tiga algoritma yang berbeda. Algoritma Naïve Bayes menujukkan persentase yang baik, yaitu dengan rata-rata 80,97%. K-NN memiliki persentase yang tidak terlalu baik, walaupun sudah diberi tiga skenario yang berbeda. Semakin tinggi nilai K, semakin rendah persentase rata-ratanya. Dari penelitian ini, dapat disimpulkan bahwa algoritma SVM mempunyai performa terbaik di antara ketiga algoritma dengan hasil persentase 82,36%.

DAFTAR PUSTAKA

[1] A. G. Lazuardy, H. S. S. Kom, and M. Eng, "Proceeding SINTAK 2019," pp. 1–6, 2019.
[2] Profil Kesehatan Indonesia, *Kementrian Kesehatan Republik Indonesia (KEMENKES RI). 2019. Data dan Informasi Profil Kesehatan Indonesia 2018. Jakarta: Ditjen P2P, Kemenkes RI 2019.*, vol. 53, no. 9. 2019.
[3] A. Rosida, "Pemeriksaan Laboratorium Penyakit Hati," *Berk. Kedokt.*, vol. 12, no. 1, p. 123, 2016, doi: 10.20527/jbk.v12i1.364.

[4] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.

[5] M. Mukarromah, S. Martha, and I. Ilhamsyah, "Perbandingan Imputasi Missing Data Menggunakan Metode Mean Dan Metode Algoritma K-Means," *Bimaster*, vol. 4, no. 3, pp. 305–312, 2015, [Online]. Available: http://jurnal.untan.ac.id/index.php/jbmstr/article/view/12425/.

[6] R. Silvi, "Analisis Cluster dengan Data Outlier Menggunakan Centroid Linkage dan K-Means Clustering untuk Pengelompokkan Indikator HIV/AIDS di Indonesia," *J. Mat. "MANTIK,"* vol. 4, no. 1, pp. 22–31, 2018, doi: 10.15642/mantik.2018.4.1.22-31.

[7] S. Nurhayati, E. T. Luthfi, and U. Y. Papua, "Prediksi Mahasiswa Drop Out Menggunakan Metode Support Vector," *Prediksi menggunakan SVM*, vol. 3, no. 6, pp. 82–93, 2015.

[8] A. M. Puspitasari, D. E. Ratnawati, and A. W. Widodo, "Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 2, pp. 802–810, 2018.

[9] F. Liantoni, "Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor," *J. Ultim.*, vol. 7, no. 2, pp. 98–104, 2016, doi: 10.31937/ti.v7i2.356.

[10] F. Liantoni and H. Nugroho, "Klasifikasi Daun Herbal Menggunakan Metode Naïve Bayes Classifier Dan Knearest Neighbor," *J. Simantec*, vol. 5, no. 1, pp. 9–16, 2015.

[11] Yusra, D. Olivita, and Y. Vitriani, "Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor," *J. Sains, Teknol. dan Ind.*, vol. 14, no. 1, pp. 79–85, 2016.

[12] I. A. M. SUPARTINI, I. K. G. SUKARSA, and I. G. A. M. SRINADI, "Analisis Diskriminan Pada Klasifikasi Desa Di Kabupaten Tabanan Menggunakan Metode K-Fold Cross Validation," *E-Jurnal Mat.*, vol. 6, no. 2, p. 106, 2017, doi: 10.24843/mtk.2017.v06.i02.p154.

[13] Andras Janosi, M.D, William Steinbrunn, M.D, Matthias Pfisterer, M.D, dan Robert Detrano, M.D., Ph.D, Heart Disease Dataset : UCI Machine Learning Repository. Diakses pada : 1 April 2021. [Online]. Available : https://archive.ics.uci.edu/ml/datasets/Heart+Disease