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 III - Surabaya, 11 Maret 2023
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2023.4067

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

Perbandingan Metode Naive Bayes, K-NN dan Decision Tree Terhadap Dataset *Healthcare Stroke*

Adela Rizky Oktavyani, Aji Wicaksono, Alexandria Felicia Seanne, Astrid Dwi Karolin Nofana, Rakha Satria Putra, Muchamad Kurniawan

Institut Teknologi Adhi Tama Surabaya

e-mail: fgelicia@gmail.com

ABSTRACT

The purpose of this study is to compare several methods including Naive Bayes, KNN and Decision Tree. Where the data from the research is a data set of health reports for Stroke disease originating from kaggle.com, in this study the confusion matrix, precision, recall, accuracy, to f-measure will be measured and then the root mean square error of each method is also calculated, from the calculation the KNN method gets the highest accuracy of up to 95.20% so that it can be concluded that the d KNN classification method is better than the Naive Bayes and Decision Tree methods.

Keywords: *Naive Bayes; KNN; Decision Tree; Healthcare Stroke; Classification; Accuracy.*

ABSTRAK

Tujuan dari penelitian ini adalah membandingkan antara metode Naive bayes, KNN dan Decision Tree. Dimana data dari penelitian adalah dataset laporan kesehatan penyakit Stroke berasal dari kaggle.com, pada penelitian ini akan diukur confusion matrix, precision, recall, accuracy, hingga f-measure kemudian juga dihitung root mean square error dari tiap-tiap metode, dari perhitungan tersebut metode KNN mendapatkan accuracy tertinggi hingga 95,20% sehingga dapat

disimpulkan metode klasifikasi d KNN lebih baik dari metode Naive bayes maupun Decision Tree.

Kata kunci: Naive Bayes; KNN; Decision Tree; Healthcare Stroke; Klasifikasi; Akurasi.

PENDAHULUAN

Stroke adalah penyakit gangguan fungsional otak fokal maupun general secara akut, lebih dari 24 jam kecuali pada intervensi bedah atau meninggal, berasal dari gangguan sirkulasi serebral. Stroke selalu menyerang secara tiba-tiba. Stroke datang tanpa diundang, dan selalu mengagetkan siapa saja. Stroke dapat menyerang laki-laki maupun perempuan, kaya atau miskin, tua atau muda. Stroke menyerang tanpa pandang bulu. Selain itu rasio penyakit stroke dapat menyerang 1 diantara 6 orang di seluruh dunia. Kematian akibat stroke sangat tinggi hal itu dapat dijelaskan melalui berbagai penelitian epidemiologi dengan hasil bahwa kematian akibat stroke adalah berkisar antara 20% sampai dengan 25% [1].

Penyebab penyakit stroke ini telah menjadi beban bagi keluarga dan Negara. Kejadian stroke selalu meningkat dari tahun ketahun, di Negara eropa yaitu tercatat 650.000 penderita dan setiap 4 detik terjadi kasus kematian akibat stroke. Negara berkembang kejadian stroke berkisar antara 30 % - 70 % dengan stroke hemorrhagic dan non hemoragic. Indonesia insiden stroke diperkirakan 800- 1000 penderita setiap tahunnya dan merupakan Negara penyumbang insiden stroke terbesar di Negara Asia. Lampung di tahun 2013 prevalensi stroke 2,6 % , terdiagnosis oleh tenaga kesehatan 3,7 % sedangkan yang terdiagnosis hanya berdasarkan gejala ada 5,4 %.[2]

Sebelumnya Prayoga Permana et al. 2021 telah melakukan penelitian Analisis Perbandingan Algoritma Decision Tree, K – Nearest Neighbor, dan Naïve bayes untuk Prediksi Kesuksesan Start-up dengan hasil perbandingan antara algoritma Decision Tree, K – Nearest Neighbor, dan Naïve bayes, untuk melakukan klasifikasi terhadap 923 data startup, menunjukkan algoritma Decision Tree merupakan algoritma yang paling cocok untuk digunakan di antara algoritma K – Nearest Neighbor dan Naïve bayes. Hasil akurasi Decision Tree adalah sebesar 79,29%, sedangkan algoritma K – Nearest Neighbor dengan 66,69%, dan Naïve bayes dengan 64,21%. Selanjutnya untuk nilai presisinya, Decision Tree masil lebih unggul dengan nilai 78,99%, diikuti algoritma K – Nearest Neighbor dengan 55,13%, dan Naïve bayes 51,32%. Dari hasil performa recall, ternyata algoritma Naïve bayes menunjukkan hasil paling baik dengan 79,16%, sedangkan Decision Tree 56,27% dan K – Nearest Neighbor dengan 40,14%. Hasil pengujian Test juga menunjukkan algoritma Decision Tree adalah algoritma paling dominan di antara algoritma yang lain. Selain itu, faktor-faktor yang sangat mempengaruhi kesuksesan sebuah startup adalah age_first_funding_year, total_funding, serta relationship. Variabel age_first_funding_year, total_funding, serta relationship yang semakin besar maka semakin besar pula kesempatan sebuah startup tersebut akan sukses.[3]

Melihat penelitian yang telah dilakukan sebelumnya, dan mempertimbangkan kelebihan dan kekurangan masing-masing metode, maka kami memutuskan untuk melakukan perbandingan antara metode klasifikasi Decision Tree, Naïve bayes, dan K – Nearest Neighbor (K-Nearest Neighbor). Hal inilah yang juga menjadi keunikan dari penelitian ini, ketiga metode tersebut akan diimplementasikan kedalam *Healthcare Stroke* dataset sebanyak 5110 record penyakit stroke dengan 12 atribut atau variabel. Perbandingan ini dilakukan untuk menemukan algoritma terbaik yang dapat digunakan untuk melakukan klasifikasi terhadap *Healthcare Stroke*.

METODE

Berbagai istilah digunakan dalam penelitian ini untuk mendukung proses penelitian mempersiapkan evaluasi.

Naïve Bayes

Naïve Bayes adalah algoritma yang digunakan untuk mencari nilai kemungkinan maksimum untuk mengklasifikasikan data uji ke dalam kelas yang paling sesuai. Ada dua

tahapan dalam klasifikasi dokumen. Langkah pertama adalah pelatihan tentang dokumen yang kelasnya diketahui. Langkah kedua adalah klasifikasi dokumen yang tidak diketahui [4].

Konsep dasar yang digunakan oleh Bayes adalah teorema Bayes, yaitu suatu klasifikasi dengan cara menghitung nilai probabilitas. Klasifikasi dilakukan untuk menentukan kelas dari dokumen tersebut. Keuntungan dari pengklasifikasi Naïve bayes hanya membutuhkan sedikit data pelatihan untuk memperkirakan parameter (rata-rata dan varian variabel) [5]yang diperlukan untuk klasifikasi. Karena variabel independen diasumsikan, hanya varians dari variabel di setiap kelas yang harus ditentukan, bukan seluruh matriks kovarians [6].

KNN

K-nearest neighbor adalah metode klasifikasi objek berdasarkan data training yang paling dekat dengan objek. Data pelatihan diproyeksikan ke ruang multidimensi, di mana setiap dimensi mewakili fitur data. Ruang ini dibagi menjadi beberapa bagian sesuai dengan klasifikasi data pelatihan. Nilai k yang terbaik untuk algoritma ini bergantung pada data, pada umumnya nilai k yang tinggi akan mengurangi efek noise pada classifier, namun akan membuat batas antar kelas semakin kabur [7].

Decision Tree

Decision tree adalah metode klasifikasi data. Model pohon keputusan merupakan pohon yang terdiri dari simpul akar, simpul internal, dan simpul terminal. Sementara simpul akar dan dalam adalah variabel/fitur, simpul daun adalah label kelas. Saat melakukan klasifikasi, kueri data akan mengikuti simpul akar dan simpul dalam hingga mencapai simpul daun. Buat kueri label lapisan data berdasarkan label pada simpul dalam. Dalam pohon keputusan tradisional, data yang digunakan adalah data dengan nilai fitur yang telah ditentukan.

Beberapa langkah dari algoritma membuat pohon keputusan adalah:
1) menyiapkan data pelatihan. Data dikelompokkan ke dalam kelas-kelas tertentu dan menggunakan data pelatihan yang diambil dari data historis yang terjadi sebelumnya, 2) menghitung akar pohon. Nilai dasar berasal dari perolehan atribut tertinggi dan akan menjadi akar pertama setelah menghitung perolehan setiap atribut. [5].

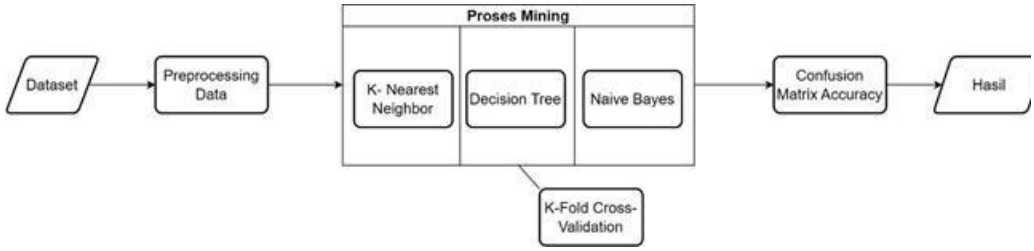
K-Fold Cross-Validation

Validasi yang digunakan dalam penelitian ini adalah k-fold cross-validation. Uji K-fold cross-validation secara acak membagi data menjadi k bagian yang masing-masing bagian memiliki nomor yang sama. Dalam hal ini, data dibagi menjadi k bagian secara bergantian menjadi data latih dan data uji hingga k iterasi. Tujuan dari K-fold cross-validation pada penelitian ini adalah untuk memastikan bahwa algoritma klasifikasi yang digunakan teruji lebih baik dan performa yang dihasilkan valid. Validasi silang atau estimasi rotasi adalah teknik validasi model yang digunakan untuk menilai bagaimana hasil analisis statistik digeneralisasikan ke data independen [6][7].

METODE

Metode penelitian digunakan sebagai acuan atau kerangka proses penelitian sehingga rangkaian proses dapat dilakukan secara sistematis dan terarah. Langkah pertama yang dilakukan adalah pengumpulan data. Data yang diperoleh berupa 5110 dataset penyakit stroke dengan 12 atribut. Langkah kedua dilakukan proses *preprocessing* data atau pengolahan data awal untuk mendapatkan data yang baik sebelum diolah menggunakan metode *K-Nearest Neighbor*, *Decision Tree*, dan *Naive Bayes*. Setelah *preprocessing* dilakukan maka dataset yang diperoleh adalah 5110 data penyakit stroke dengan 10 atribut. Langkah ketiga adalah proses mining dilakukan menggunakan metode *K-Nearest Neighbor*, *Decision Tree*, dan *Naive Bayes*. Teknik *K-Fold Cross Validation* digunakan untuk memvalidasi nilai akurasi ketiga metode yang diterapkan dengan membagi data secara acak dan mengelompokkan data tersebut sebanyak nilai

K pada *K-Fold* [5] dan tingkat akurasi dapat dilihat berdasarkan *Confusion Matrix*. Langkah terakhir adalah hasil pengujian dari metode *K-Nearest Neighbor*, *Decision Tree*, dan *Naive Bayes* akan dibandingkan untuk mengetahui metode terbaik dengan melihat tingkat akurasi paling tinggi. Dalam penelitian ini menggunakan *tools/software* Jupyter Notebook dan beberapa library diantaranya : sklearn, DecisionTreeClassifier, LabelEncoder, GaussianNB, KNeighborsClassifier, cross_val_score, dan Confusion Matrix.



Gambar 1. Alur Metode Penelitian

Pengumpulan Data

Penelitian ini menggunakan data penyakit stroke yang didapatkan dari *Kaggle* (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset). Data yang diperoleh sebanyak 5110 dataset penyakit stroke dengan 12 atribut atau variabel. Variabel yang digunakan antara lain adalah *id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke.*

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Gambar 2. Contoh Dataset

Preprocessing

Data yang digunakan dalam penelitian ini merupakan data yang belum siap untuk diolah. Data yang masih memiliki variabel yang tidak dibutuhkan, *missing value, noise,* dan data yang memiliki jenis data belum sesuai dengan penelitian yang dilakukan sehingga perlu dilakukan *preprocessing*. Penelitian ini memiliki kemiripan dengan penelitian sebelumnya [8] yang menggunakan data stroke sebagai dataset.

Dari hasil pengumpulan data sebelumnya diperoleh sebanyak 5110 dataset penyakit stroke dengan 12 atribut atau variabel. Namun tidak seluruhnya data tersebut digunakan. Proses *preprocessing* yang digunakan dalam penelitian ini antara lain seperti pada Tabel 3. 1

Tabel 1. *Preprocessing* Data

No	Kegiatan	Tujuan
1	Pembersihan Data dan Feature Extraction	Menghilangkan (menghapus) data yang kosong dan tidak lengkap untuk menghindari *missing value*. Serta digunakan untuk mengekstraksi karakteristik dari suatu bentuk, dengan nilai yang diperoleh kemudian diperiksa untuk pemrosesan tambahan.
2	Reduksi Data	Mengintegrasikan atau menghapus data asing, reduksi data dapat digunakan untuk meminimalkan ukuran data.

3	Transformasi Data	Digunakan untuk mengubah skala pengukuran data asli menjadi suatu yang lain sehingga data tersebut dapat memenuhi asumsi analisis variabel yang mendasarinya.

Proses Mining dan Validasi Nilai Akurasi

Proses mining menggunakan metode perhitungan *K-Nearest Neighbor*, *Decision Tree*, dan *Naive Bayes.* Dan validasi tingkat akurasi ketiga metode dilakukan dengan cara menambahkan *K-Fold Cross Validation* dengan melakukan 10 kali iterasi atau pengulangan (K=10 *fold*). Dalam 10 kali iterasi dibagi 10 subset data dimana data tersebut *Cross Validation* akan menggunakan 9 *fold* untuk pelatihan dan 1 *fold* untuk pengujian.

Evaluasi Tingkat Akurasi

Confusion Matrix digunakan untuk mengetahui perhitungan evaluasi model klasifikasi. *Confusion Matrix* merupakan alat pengukuran yang dapat digunakan untuk menghitung kinerja atau tingkat kebenaran proses klasifikasi. Serta menggunakan perhitungan nilai Accuracy, nilai Precision, nilai Recall, dan F1 Score..

HASIL DAN PEMBAHASAN

Berdasarkan pengujian dengan menggunakan 5110 data stroke yang telah dilakukan pada model maka didapatkan hasil sebagai berikut :

Tabel 2. Hasil Pengujian

	MSE	Akurasi	Precision	Recall	F1-score
NB	13,6	86	97	89	93
KNN	4,5	95	97	100	98
DT	5,97	94	96	98	97

Model *Naïve Bayes*

Pada Naïve bayes mendapatkan nilai MSE sebesar 13,6%, akurasi 86% dan pada confusion matrix mendapat kan nilai tertinggi pada precision sebesar 97%, recall sebesar 89% dan f1-score sebesar 93%.

Model *K-Nearest Neighbor*

Model ini menghasilkan nilai terbaik ketika menggunakan 7 n *neighbors*. Pada *K-Nearest Neighbor* mendapatkan nilai MSE sebesar 4,5%, akurasi 95% dan pada confusion matrix mendapat kan nilai tertinggi pada precision sebesar 96%, recall sebesar 100% dan f1-score sebesar 98%.

Model *Decision Tree*

Model ini menghasilkan nilai terbaik ketika menggunakan 7 max depth dan criterion ‘gini’. Pada DT mendapatkan nilai MSE sebesar 5,97%, akurasi 94% dan pada confusion matrix mendapat kan nilai tertinggi pada precision sebesar 96%, recall sebesar 98% dan f1-score sebesar 97%.

KESIMPULAN

Penelitian ini bertujuan untuk mendapatkan model yang memiliki nilai akurasi terbaik dalam melakukan klasifikasi 5110 data stroke. Berdasarkan hasil pengujian yang telah dilakukan,

model KNN menghasilkan nilai akurasi paling baik jika dibandingkan dengan model NB dan DT yakni sebesar 95%. Model KNN juga memiliki nilai MSE yang terkecil.

DAFTAR PUSTAKA

[1] B. D. Agustina, "Asuhan Keperawatan Gawat Darurat Pada Pasien Dengan Stroke Diinstalasi Gawat Darurat Rumah Sakit Umum Daerah sleman Yogyakarta," 2019.

[2] F. Susilawati, N. Hk, J. Keperawatan, and P. Tanjungkarang, "FAKTOR RESIKO KEJADIAN STROKE DI RUMAH SAKIT," 2018.

[3] A. Prayoga Permana, K. Ainiyah, and K. Fahmi Hayati Holle, "Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up," 2021. [Online]. Available: https://www.kaggle.com/manishkc06/startup-success-prediction.

[4] S. Tan, X. Cheng, Y. Wang, and H. Xu, "Adapting naive bayes to domain adaptation for sentiment analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5478 LNCS, no. June, pp. 362–374, 2009, doi: 10.1007/978-3-642-00958-7_31.

[5] L. C. van der Gaag and A. Capotorti, "Naive Bayesian Classifiers with Extreme Probability Features," 2018.

[6] and A. B. V. Narayanan, I. Arora, "Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model," *2013*.

[7] M. H. Abdurrahman, E. Suhartono, and E. Wulandari, "Deteksi Kualitas Kemurnian Susu Sapi Melalui Pengolahan Citra Digital Menggunakan Metode Scale Invariant Feature Transform Dengan Klasifikasi K-nearest Neighbor," *eProceedings of Engineering*, vol. 6, no. 2, pp. 3845–3852, 2019.

[8] D. Ulfatul, M. Rachmad, H. Oktavianto, and M. Rahman, "Perbandingan Metode K-Nearest Neighbor Dan Gaussian Naive Bayes Untuk Klasifikasi Penyakit Stroke Comparison Of K-Nearest Neighbor And Gaussian Naive Bayes Methods For Stroke Disease Classification," *Jurnal Smart Teknologi*, vol. 3, no. 4, pp. 2774–1702, 2022.