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.2021.1817

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 Penderita Penyakit Jantung Menggunakan Metode Naïve Bayes dengan Chi-Square untuk Pemilihan Atribut

Maftahatul Hakimah[1], Rani Rotul Muhima[2]

Jurusan Teknik Informatika Institut Adhi Tama Surabaya[1,2]
e-mail: hakimah.mafta@itats.ac.id

ABSTRACT

This study aims to determine whether the selection of heart disease dataset attributes can improve the Naïve Bayes Algorithm. The attribute selection is based on the attribute independence test for the response variable, namely the target. The attributes chosen are those that affect the response variable. The effect test here uses the chi-square test. There are two levels of significance used, namely 0.05 and 0.01. In all tests, Naïve Bayes with chi-square at a significance level of 0.01 can increase the accuracy and precision of the Naïve Bayes method by 1% and 5%, respectively. Meanwhile, Naïve Bayes without attribute selection showed the best performance on recall measurement compared to Naïve Bayes with chi-square.

Keywords: *Chi-square; Feature selection; Heart disease; Naïve Bayes.*

ABSTRAK

Penelitian ini bertujuan untuk mengetahui apakah pemilihan atribut *dataset* penyakit jantung dapat memperbaiki algoritma Naïve Bayes. Pemilihan atribut didasarkan pada uji independensi atribut terhadap variabel respons, yaitu target. Atribut yang dipilih adalah atribut yang berpengaruh terhadap variabel respons. Uji pengaruh di sini menggunakan uji Chi-square. Ada dua taraf signifikansi yang digunakan, yaitu 0,05 dan 0,01. Pada keseluruhan pengujian, Naïve Bayes dengan Chi-square pada taraf signifikansi 0,01 bisa meningkatkan akurasi dan presisi metode Naïve Bayes, masing-masing 1% dan 5%. Sedangkan Naïve Bayes tanpa pemilihan atribut menunjukkan kinerja terbaik pada pengukuran *recall* dibandingkan Naïve Bayes dengan Chi-square.

Kata Kunci: Chi-square; Naïve Bayes; Pemilihan fitur; Penyakit jantung.

PENDAHULUAN

Kematian akibat penyakit jantung menjadi urutan teratas penyebab kematian di seluruh dunia [1]. Kasus penyakit jantung mengalami pembengkakan pada tiga dasawarsa terakhir. Sedangkan angka kematian di periode tersebut juga meningkat dari angka 12,6 juta menjadi 18,6 juta [2]. Sebagai langkah pencegahan, perlu diketahui gejala-gejala penyakit jantung. Selain itu, pemeriksaan medis juga sangat diperlukan. Penelitian ini akan membantu memprediksi seseorang terkena penyakit jantung atau sehat menggunakan metode klasifikasi.

Naïve Bayes (NB) merupakan salah satu metode klasifikasi paling sederhana yang menerapkan konsep peluang. Metode Naïve Bayes memiliki asumsi bahwa setiap atribut pada *dataset* bersifat independen terhadap kelas yang diamati [3]. Penelitian-penelitian berikutnya melakukan modifikasi pada algoritma Naïve Bayes untuk mengatasi asumsi ini. Perbaikan Naïve Bayes dilakukan dengan melakukan uji hubungan antara atribut *dataset* dengan kelasnya. Atribut yang tidak memiliki relevansi terhadap kelas, akan dihapus atau diberi bobot yang kecil. Tahap ini kemudian dinamakan pemilihan atribut atau fitur data. Pemilihan atribut data dilakukan sebelum algoritma NB dijalankan. Metode NB yang telah diaplikasikan pada penentuan usia kelahiran bayi pada seorang pasien, dilakukan pemilihan fitur menggunakan *correlation based features selection*. Pemilihan fitur ini bisa menaikkan tingkat akurasi NB sebesar 2% [4]. Modifikasi NB dengan pemilihan fitur juga telah dilakukan untuk mendiagnosis penyakit hepatitis. Teknik pemilihan fitur yang digunakan ialah menerapkan algoritma optimasi Particle Swarm Optimization. Teknik ini bisa meningkatkan akurasi metode NB sebesar 5,15% [5]. Kinerja Naïve Bayes untuk klasifikasi penyakit kanker payudara dioptimalkan menggunakan *forward selection*. Teknik pemilihan fitur ini terbukti bisa meningkatkan akurasi Naïve Bayes [6].

Berdasarkan penelitian yang telah dijelaskan sebelumnya maka penelitian ini akan mengklasifikasi *dataset* penyakit jantung menggunakan algoritma Naïve Bayes dengan pemilihan fitur. Pemilihan fitur ini berdasarkan ada-tidaknya hubungan antara atribut data terhadap kelas sebagai variabel respons. Uji hubungan ini menggunakan Chi-square. Dengan Chi-square, fitur-fitur yang tidak memenuhi kriteria pengujian akan dihapus dan tidak dilibatkan dalam proses klasifikasi. Taraf signifikansi yang diperlukan pada pengujian ini akan dibuat berbeda untuk mengetahui perubahan akurasi pada hasil klasifikasi.

METODE

Tujuan penelitian ini akan dicapai dengan melakukan tahapan penelitian berikut ini.

Tahap 1. *Persiapan Data*. Data penyakit jantung yang dikaji pada penelitian ini merupakan data sekunder yang diunduh dari laman kaggle.com. Data terdiri dari 14 atribut. Terdapat 13 atribut yang dijadikan sebagai variabel prediktor dan 1 atribut sebagai variabel respons. Pada tahap ini, data kontinu akan diubah menjadi data diskrit. Selain itu, *dataset* akan dibagi menjadi data latih dan data uji.

Tahap 2. *Pemilihan Fitur*. Uji independensi bertujuan untuk mengetahui hubungan antara variabel prediktor dengan variabel respons. Uji independensi ini menggunakan uji Chi-square dengan hipotesis pengujiannya:

H_0 : Tidak ada hubungan antara atribut ke-*i* dengan variabel respons
H_1 : Ada hubungan antara atribut ke-*i* dengan atribut respons
dengan $i = 1, 2, \ldots, 13$, daerah penolakan $H_0 : \chi^2_{\text{Hitung}} > \chi^2_{\text{Tabel}(\alpha,\, df)}$, α adalah taraf signifikansi, serta *df* adalah derajat kebebasan. Nilai Chi-square hitung diberikan pada Persamaan 1 berikut [7].

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (1)$$

dengan χ^2 adalah nilai Chi-square, O_{ij} adalah nilai observasi baris ke-*i* kolom ke-*j* dan E_{ij} adalah nilai ekspektasi baris ke-*i* kolom ke-*j*. Dengan pengujian ini, atribut yang mempunyai nilai χ^2 Hitung < χ^2 Tabel akan dihilangkan pada algoritma NB. Taraf signifikansi yang digunakan adalah α = 0,01 dan α = 0,05.

Tahap 3. *Implementasi Algoritma NB*. Atribut terpilih pada tahap sebelumnya akan diproses oleh algoritma NB untuk menentukan kelas dari setiap atribut data. Algoritma dari NB diberikan berikut ini [8],[9].

a. Menghitung *prior probability* dari kelas yang diamati, yaitu $p(C_i)$; i = 0, 1
b. Menghitung *conditional probability* setiap atribut terhadap kelas yang diamati menggunakan Persamaan 2:

$$p(X_j|C_i) = \frac{p(C_i|X_j)\cdot p(X_j)}{p(C_i)} \quad (2)$$

dengan, $p(X_j)$ merupakan probabilitas atribut ke-*j* dan $p(C_i|X_j)$ merupakan probabilitas kelas ke-*i* berdasarkan kondisi atribut ke-*j*.

c. Menghitung *posterior probability* kelas terhadap atribut *dataset* menggunakan Persamaan 3:

$$p(C_i|X_1, X_2, \dots, X_j) = p(C_i) \times \prod p(X_j|C_i) \quad (3)$$

dengan, $p(C_i|X_1, X_2, \dots, X_j)$ merupakan probabilitas kelas ke-*i* berdasarkan kondisi data yang belum diketahui kelasnya.

d. Menentukan kelas dari objek data dengan Persamaan 4:

$$C(\boldsymbol{x}) = \arg\max p(C_i|X_1, X_2, \dots, X_j) \quad (4)$$

Tahap 4. *Evaluasi Hasil Klasifikasi*. Tahap evaluasi dilakukan dengan menguji model yang diperoleh pada data latih. Hasil klasifikasi NB dengan dan tanpa pemilihan atribut akan dibandingkan menggunakan *confusion matrix* [10].

Tabel 1. Confusion Matrix

Kelas Sebenarnya	**Kelas Prediksi**	
	Positif	**Negatif**
Positif	*True Positives*	*False Negatives*
Negatif	*False Positives*	*True Negatives*

Pengukuran kinerja metode NB berdasarkan Tabel 1 diberikan pada Persamaan 5.

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN}$$
$$Presisi = \frac{TP}{TP+FP} \quad (5)$$
$$Recall = \frac{TP}{TP+FN}$$

HASIL DAN PEMBAHASAN

Data Penelitian

Dataset penyakit jantung yang digunakan pada penelitian ini merupakan data sekunder yang diunduh dari tautan www.kaggle.com/ronitf/heart-disease-uci. *Dataset* yang diperoleh diberikan pada Tabel 2.

Dataset penyakit jantung tersebut terdiri dari 303 baris data. Atribut dengan tipe kontinu diubah menjadi diskrit dengan membagi setiap kategori menjadi beberapa kategori dan setiap kategori mempunyai interval yang sama. *Dataset* kemudian dibagi menjadi dua bagian yaitu *dataset* untuk pelatihan dan *dataset* untuk pengujian klasifikasi.

Tabel 2. Dataset yang digunakan.

No.	Kode Atribut	Tipe	*Min–Max*	Deskripsi
1	age	kontinu	29–77	Usia (tahun)
2	sex	diskrit	0–1	1 = Pria; 0 = Wanita
3	cp	diskrit	0–3	Tipe nyeri dada: 0 = *Asymptomatic;* 1 = *atypical angina;* 2 = *pain without relation to angina;* 3 = *typical angina*
4	trestbps	kontinu	94–200	*Resting blood pressure (mmHg on admission to hospital)*
5	chol	kontinu	126–564	Serum kolesterol (mg/dl)
6	fbs	diskrit	0–1	*Fasting blood sugar* >120 *mg/dl* (1 = *True*; 0 = *False*)
7	restecg	diskrit	0–2	Hasil pemeriksaan elektrokardiografi (0 = *Hypertrophy*; 1 = normal; 2 = abnormal pada T dan *segment* ST)
8	thalach	kontinu	71–202	Detak jantung maksimum
9	exang	diskrit	0–1	Exercise induced angina (1 = *Yes*; 0 = *No*)
10	oldpeak	kontinu	0–6,2	*ST depression induced by exercise relative to rest*
11	slope	diskrit	0–2	*The slope of the peak exercise ST segment* (0 = *descending;* 1 = *flat;* 2 = *ascending*)
12	ca	diskrit	0–3	*Number of major vessels colored by flouroscopy*
13	thal	diskrit	0–3	*Thallium scan* (3 = *Normal;* 6 = *fixed defect;* 7 = *reversable defect*)
14	target	diskrit	0–1	Diagnosis penyakit jantung (0 = *No*; 1 = *Yes*)

Proses Pelatihan Data

Pelatihan data dilakukan untuk memperoleh nilai *conditional probability* setiap atribut pada kelas masing-masing. Proses ini akan dicoba untuk jumlah *dataset* yang berbeda. Jumlah *dataset* pelatihan yang digunakan adalah 100, 150, 200, dan 240. Langkah berikutnya adalah pemilihan fitur dengan uji Chi-square. Nilai Chi-square setiap atribut terhadap variabel respons (target) menggunakan Persamaan 1. Tabel 3 berikut merangkum hasil pemilihan atribut berdasarkan kriteria pengujian pada Chi-square.

Tabel 3. Pemilihan Fitur

Jumlah *dataset*	Keputusan pada α = 0,05		Keputusan pada α = 0,01	
	Jumlah Atribut Terpilih	Atribut Keluar	Jumlah Atribut Terpilih	Atribut Keluar
100	7	age, sex, trestbp, chol, *fbs*, *restecg*	6	age, sex, trestbp, chol, *fbs*, *restecg*, thalach
150	11	*fbs*, *restecg*	10	trestbp, *fbs*, *restecg*
200	11	trestbp, *fbs*	8	age, trestbp, chol, *fbs*, *restecg*
240	10	trestbp, *fbs*, *restecg*	10	trestbp, *fbs*, *restecg*

Pemilihan atribut data menggunakan uji Chi-square mengakibatkan dimensi data berkurang. Dari pemilihan fitur ini, bisa dilihat bahwa taraf signifikansi α mempengaruhi banyaknya atribut yang terseleksi. Tabel 3 menunjukkan bahwa banyak atribut yang terseleksi pada jumlah *dataset* 100. Sedangkan jumlah dataset pelatihan selainnya, rata-rata atribut yang terseleksi sebanyak 3–5 atribut. Jika dilihat pada kode atributnya, *fbs* dan *restecg* selalu menjadi atribut yang terseleksi pada jumlah *dataset* dan taraf signifikansi yang ditentukan.

Langkah pelatihan berikutnya adalah mengimplementasikan algoritma Naïve Bayes menggunakan Persamaan 2–4. Hasil pelatihan data adalah nilai *conditional probability* untuk setiap atribut terpilih terhadap variabel target.

Proses Klasifikasi

Proses klasifikasi dilakukan pada data uji. Dengan beberapa dataset pelatihan tersebut, NB dengan atau tanpa pemilihan atribut akan diukur kinerjanya terhadap 30 data uji. Hasil *conditional probability* pada proses pelatihan diterapkan pada proses ini. Tabel berikut merangkum hasil perhitungan akurasi hasil klasifikasi NB tanpa dan dengan pemilihan atribut menggunakan Persamaan 5.

Tabel 4. Pengukuran Kinerja Metode Klasifikasi

Metode	Jumlah Dataset Pelatihan	Akurasi	Presisi	*Recall*
NB + α (0,01)	100	0,700	0,688	0,733
	150	0,800	0,800	0,800
	200	0,767	0,750	0,800
	240	0,833	0,813	0,867
	Mean	**0,775**	**0,763**	**0,800**
NB + α (0,05)	100	0,767	0,750	0,800
	150	0,767	0,750	0,800
	200	0,733	0,684	0,867
	240	0,833	0,813	0,867
	Mean	**0,767**	**0,750**	**0,800**
NB	100	0,767	0,722	0,867
	150	0,700	0,650	0,867
	200	0,733	0,684	0,867
	240	0,733	0,684	0,867
	Mean	**0,767**	**0,722**	**0,867**

Tabel 4 menunjukkan kinerja metode NB dengan pemilihan fitur dan tanpa pemilihan fitur. Kinerja NB dengan pemilihan fitur baik $\alpha = 0{,}01$ atau $\alpha = 0{,}05$ terbaik dicapai pada jumlah dataset pelatihan yang terbanyak. Berbeda dengan NB tanpa ada pemilihan fitur, jumlah dataset pelatihan terkecil menunjukkan kinerja terbaiknya. Secara keseluruhan, metode Naïve Bayes dengan taraf signifikansi 0,01 bisa meningkatkan akurasi sebesar 1% dan presisi sebesar 5% dari Naïve Bayes sebelumnya. Sedangkan untuk pengukuran *recall*, Naïve Bayes tanpa pemilihan atribut masih menunjukkan kinerja yang lebih baik daripada Naïve Bayes dengan pemilihan atribut.

KESIMPULAN

Metode Naïve Bayes dengan pemilihan atribut telah diimplementasikan pada *dataset* penyakit jantung. Pengujian metode Naïve Bayes dengan Chi-square untuk pemilihan atribut menunjukkan adanya peningkatan akurasi dan presisi masing-masing 1% dan 5% pada taraf signifikansi sebesar 0,01. Namun, untuk pengukuran *recall*, kinerja Naïve Bayes tanpa pemilihan atribut menunjukkan performansi terbaik dibandingkan dengan Naïve Bayes dengan Chi-square.

DAFTAR PUSTAKA

[1] F. DR. dr. Isman Firdaus Sp.JP (K), FIHA, FAPSIC, FAsCC, FESC, "Hari Jantung Sedunia (World Heart Day): Your Heart is Our Heart Too," 2019. http://www.inaheart.org /news_and_events/news/2019/9/26/press_release_world_heart_day_perki_2019#:~:text=Berdasarkan data Riset Kesehatan Dasar,di Indonesia menderita penyakit jantung. (accessed Apr. 30, 2021).

[2] G. Perkasa, "Penyakit Jantung Penyebab Kematian Utama di Dunia Artikel ini telah tayang di Kompas.com dengan judul 'Penyakit Jantung Penyebab Kematian Utama di Dunia', Klik untuk baca: https://lifestyle.kompas.com/read/2020/12/14/101607520/penyakit-jantung -penyebab-ke," *2019*.

[3] I. Rish, "An empirical study of the naive Bayes classifier," vol. 3, no. 22, pp. 4863–4869, 2001.

[4] Y. I. Kurniawan, T. Cahyono, Nofiyati, E. Maryanto, A. Fadli, and N. R. Indraswari, "Preprocessing Using Correlation Based Features Selection on Naive Bayes Classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 982, no. 1, pp. 0–8, 2020, doi: 10.1088/1757-899X/982/1/012012.

[5] T. S. N. Koeswara, M. S. Mardiyanto, and M. A. Ghani, "Penerapan Particle Swarm Optimization (Pso) Dalam Pemilihan Atribut Untuk Meningkatkan Akurasi Prediksi Diagnosispenyakit Hepatitis Dengan Metode Naive Bayes," *J. Speed – Sentra Penelit. Eng. dan Edukasi*, vol. 12, no. 1, pp. 1–10, 2020.

[6] L. W. Astuti, I. Saluza, and M. F. Alie, "Optimalisasi Klasifikasi Kanker Payudara Menggunakan Forward Selection pada Naive Bayes," vol. 11, no. 2, 2020.

[7] R. Efendi and L. Chairani, "Hubungan Sistem Pembelajaran Daring Di Era COVID-19 Terhadap Kesehatan Mental Guru SD : Uji Chi-Square dan Dependency Degree," pp. 608–615, 2020.

[8] J. Abellán and J. G. Castellano, "Improving the Naive Bayes classifier via a quick variable selection method using maximum of entropy," *Entropy*, vol. 19, no. 6, 2017, doi: 10.3390/e19060247.

[9] E. Manalu, F. A. Sianturi, and M. R. Manalu, "Penerapan Algoritma Naive Bayes Untuk Memprediksi Jumlah Produksi Barang Berdasarkan Data Persediaan dan Jumlah Pemesanan Pada CV. Papadan Mama Pastries," *J. Mantik Penusa*, vol. 1, no. 2, pp. 16–21, 2017, [Online]. Available: https://ezp.lib.unimelb.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ffh&AN=2008-10-Aa4022&site=eds-live&scope=site.

[10] P. Subarkah, E. P. Pambudi, and S. O. N. Hidayah, "Perbandingan Metode Klasifikasi Data Mining untuk Nasabah Bank Telemarketing," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 1, pp. 139–148, 2020, doi: 10.30812/matrik.v20i1.826.