Available online at https://ejournal.almaata.ac.id/index.php/IJUBI
Indonesian Journal of Business Intelligence
Volume 5 | Issue 2 | December (2022)
IJUBI
Indonesian Journal ---- of ---- Business Intelligence

PENERAPAN METODE *DISCRETIZATION* DAN *ADABOOST* UNTUK MENINGKATKAN AKURASI ALGORITMA KLASIFIKASI DALAM MEMPREDIKSI PENYAKIT JANTUNG

Annisa Maulana Majid[1], Muhammad Najamuddin Dwi Miharja[2]
[1,2]Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa.
annisa.maulanamajid@pelitabangsa.ac.id , najamuddin.dwi@pelitabangsa.ac.id
Jalan Inspeksi Kalimalang Tegal Danas Arah Deltamas, Cikarang Selatan, Bekasi

Keywords:

Heart disease, Discretization, Adaboost, Decision Tree C45, KNN.

Abstract

The mortality rate caused by heart disease can be reduced if there is an accurate diagnosis early on. Previous research in predicting heart disease with a high level of accuracy has been carried out but produces little accuracy in the Decision Tree C4.5 and K-Nearest Neighbor (KNN) algorithms. For this reason, it is necessary to increase accuracy in order to produce accurate information. The purpose of this research is to improve the accuracy of the Decision Tree C4.5 and K-Nearest Neighbor (KNN) classification algorithms using the heart disease dataset from kaggle.com by applying the discretization technique and the ensemble method, namely adaboost. The results of this study with a single algorithm produced an accuracy of 89.17% in the Decision Tree and 84.68% in KNN, while the Decision Tree used a discretization technique and adaboots of 99.81% and KNN used a discretization technique and adaboots of 92.88%. The results show an increase in the classification algorithm using discretization and adaboots techniques.

Kata Kunci:

Penyakit Jantung Discretization, Adaboost, Decision Tree C45, KNN .

Abstrak

Angka kematian yang disebabkan oleh penyakit jantung dapat dikurangi jika ada diagnosa yang akurat sejak dini. Penelitian sebelumnya dalam memprediksi penyakit jantung dengan tingkat akurasi telah dilakukan namun menghasilkan akurasi yang kecil pada algoritma *Decision Tree C4.5* dan *K-Nearest Neighbor* (KNN). Untuk itu diperlukan adanya peningkatan akurasi agar menghasilkan keakuratan informasi. Tujuan penelitian ini adalah untuk meningkatkan akurasi dari algortima klasifikasi *Decision Tree C4.5* dan K-Nearest Neighbor (KNN) menggunakan data *heart disease dataset* dari kaggle.com dengan menerapkan teknik *discretization* dan metode *ensemble* yaitu a*daboost*. Hasil penelitian ini dengan algoritma tunggal menghasilkan akurasi sebesar 89,17% pada *Decision Tree* dan 84,68% pada KNN, sedangkan *Decision tree* menggunakan teknik *discretization* dan *adaboots* sebesar 99,81% dan KNN menggunakan teknik *discretizatio*n dan *adaboots* sebesar 92,88%. Hasil menunjukkan adanya peningkatan algortima klasifikasi menggunakan teknik *discretization* dan *adaboots.*

Pendahuluan

Penyakit jantung merupakan penyakit yang dapat menyebabkan kematian. Penyakit jantung merupakan penyakit dengan tingkat pravelensi terdiagnosis meningkat dari tahun 2013-2018 berdasarkan hasil Riset Kesehatan Dasar (Rikesdas) 2018. Prevalensi penyakit jantung tahun 2013 berdasarkan terdiagnosis oleh dokter sebesar 0,5 persen [1]. Pada tahun 2018 pravalensi penyakit jantung terdiagnosis dokter meningkat menjadi 1,5 persen [2]. Untuk itu diperlukannya penanganan secara dini untuk mencegah penyakit jantung dengan melakukan diagnosa awal. Penelitian tentang diagnosa penyakit jantung telah dilakukan menggunakan berbagai algoritma, diantaranya yaitu, Naïve Bayes, Decision Tree, K-Nearest

Neighbor (KNN), K-means, Support Vector Machine (SVM), dan lain-lain. Penelitian M. Salman Pathan dkk menghasilkan akurasi 69 % menggunakan algoritma KNN dan 61% menggunakan algoritma *Decision Tree* [3]. Penelitian Apriyanto Alhamad, dkk menghasilkan akurasi 70,93 % menggunakan algoritma KNN, 81,26% menggunakan algoritma *Decision Tree,* 80,25 menggunakan algoritma SVM dan 81,76 menggunakan algoritma *Naïve Bayes* [4]. Namun hasil menunjukkan bahwa tingkat akurasi Decision tree dan KNN memiliki tingkat akurasi rendah dengan perbedaan yang tidak signifikan untuk itu diperlukannya adanya penanganan pada *dataset* yang memiliki atribut kontinyu menggunakan teknik *discretization* dan diperlukan adanya peningkatan akurasi untuk memberikan hasil keputusan yang terbaik. Teknik *discretization* adalah salah satu teknik reduksi data yang paling mendasar,yang berfokus pada pemindahan atribut kontinu atau numerik ke atribut diskrit atau nominal dengan angka terbatas dengan interval [5]. *Adaptive Boosting* adalah kepanjangan dari *Adaboos*t memiliki keunggulan yaitu fokus dalam hal *misclasssified tuples* dan lebih tinggi tingkat akurasinya. Penelitian ini dilakukan dengan penggabungan teknik *discretization* dan *adaboost* untuk menghasilkan akurasi yang lebih baik. Teknik *discretization* untuk menangani *dataset heart disease* yang merupakan dataset dengan atribut numerik dan metode *adaboost* digunakan untuk meningkatkan akurasi algoritma klasifikasi dalam prediksi penyakit jantung.

Landasan Teori

1. Data Mining

Data mining merupakan proses yang dilakukan dengan menggali suatu nilai yaitu informasi dengan cara mengekstraksi dan menemukan pola penting atau menarik dari suatu basis data yang sebelumnya tidak diketahui jika dilakukan secara manual [6]. Data mining yaitu analisis yang dilakukan di dalam basis data untuk menemukan pengetahuan dalam bentuk pola atau relasi data valid yang disebut *Knowledge Discovery in Databases* (KDD) [7].

2. Decision Tree C4.5

Decision Tree C4.5 merupakan algoritma yang dikembang dari algoritma ID3. Pemilihan algoritma didasarkan pada *information gain* [8]. *Decision Tree* pada algoritma ID3 mengalami perbaikan dan diubah menjadi algoritma C4.5, salah satu perbaikan dari algoritma tersebut adalah dalam hal pemangkasan (*prunning*) [9]. *Decision Tree C4.5* merupakan algoritma sederhana sehingga pengguna mudah memahami dari arti *rule*/aturan yang dibentuk pada algoritma ini [10].

3. *K- Nearst Neighbor* (KNN)

Algoritma KNN adalah metode non-parametrik yang banyak digunakan untuk klasifikasi dalam pengenalan pola. Prinsip utama KNN adalah bahwa kategori titik data ditentukan sesuai dengan klasifikasi tetangga K terdekat [11]. KNN merupakan algoritma yang banyak digunakan dalam pembentukan pola dalam algoritma klasifikasi, namun berpengaruh pada sensitivitas ukuran k, sehingga dapat menurunkan tingkah akurasi [12].

4. Teknik *Discretization*

Discretization dapat meminimalkan jumlah interval tanpa kehilangan yang signifikan dari suatu *dataset*. Teknik *discritize* adalah salah satu teknik reduksi data yang paling mendasar,yang berfokus pada pemindahan atribut kontinu atau numerik ke atribut diskrit atau nominal dengan angka terbatas dengan interval [5]. *Discretization* adalah teknik reduksi data yang bertujuan memproyeksikan seperangkat nilai kontinu menjadi ruang diskrit dan terbatas [13]

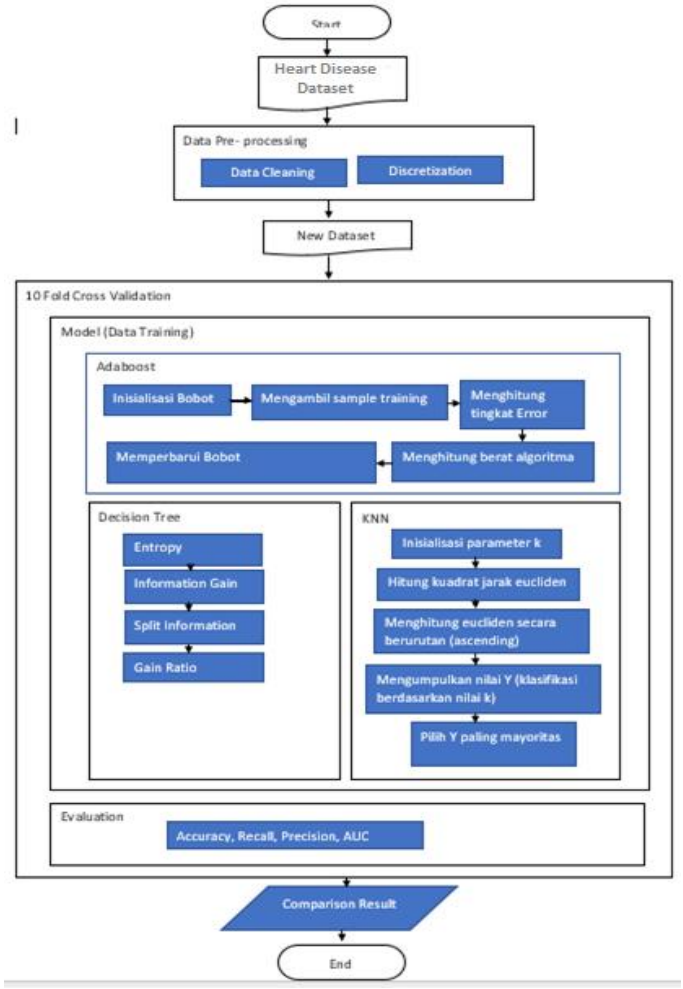
5. Metode *Adaboost*

Adaboost merupakan kepanjangan dari *adaptive boosting, adaboost* adalah teknik pemberian bobot terhadap klasifikasi lemah dan mengumpulkannya menjadi klasifikasi kuat [14]. Algoritma *adaboost* dari Freund dan Schapire merupakan algoritma penguat praktis pertama, dan tetap menjadi salah satu yang paling banyak digunakan dan dipelajari, dengan aplikasi di berbagai bidang [9]. Kelebihan dari *adaboost* sehingga sukses diterapkan yaitu teori yang ada dalam teknik

adaboost kuat, prediksi yang dihasilkan akurat, dan diimplementasikan dengan sederhana [15].

Metode

Penelitian ini mengunakan penerapan teknik *discretization* dan metode *adaboost* pada algoritma klasifikasi yaitu *Dicision Tree C4.5* dan KNN untuk meningkatkan akurasi dalam memprediksi penyakit jantung. Berikut tahapan dalam penelitian



Gambar 1. Tahapan Penelitian

1. *Dataset*

Sumber data yang digunakan penelitian ini berasal dari data publik yaitu *dataset heart disease* dari *kaggle.com* dengan jumlah data sebanyak 1025 data. *Dataset* tersebut terdiri dari 14 atribut, yaitu sebagai berikut :

1) *age*
2) *sex*
3) *chest pain type (4 values)*
4) *resting blood pressure*
5) *serum cholestoral in mg/dl*
6) *fasting blood sugar > 120 mg/dl*
7) *resting electrocardiographic results (values 0,1,2)*
8) *maximum heart rate achieved*
9) *exercise induced angina*
10) *oldpeak = ST depression induced by exercise relative to rest*
11) *the slope of the peak exercise ST segment*
12) *number of major vessels (0-3) colored by flourosopy*
13) *thal: 0 = normal; 1 = fixed defect; 2 = reversable defect*
14) *Class*

Pada *dataset* jumlah nilai positif ada 526 data sedangkan nilai negatif ada 499 data.

2. Teknik Pemodelan dan Validasi Data

Pada penelitian ini menggunakan *k-folds cross validation* sebagai metode untuk validasi dengan nilai k=10. Validasi dilakukan untuk menguji model algoritma yang digunakan. *K-folds cross validation* merupakan metode untuk mengetahui tingkat keberhasilan pada model algoritma dengan cara melakukan pengujian ulang atribut input yang acak, dalam metode ini data dibagi menjadi k subset secara acak, satu subset digunakan untuk data *testing* dan sisanya untuk data *training* [16]. Nilai k yang digunakan yaitu 5 atau 10, biasa disebut 10 *folds cross validation*, yaitu data dibagi menjadi 10 bagian, 90% untuk *training* dan 10% lainnya digunakan sebagai *testing*. Proses dilakukan berulang sampai dengan 10 kali atau 10 iterasi sampai semua *record* data mendapatkan bagian sebagai data testing [17]. Cara kerja *k-folds cross validation*, yaitu total data dibagi menjadi n bagian, iterasi atau *fold* ke 1, yaitu bagian ke 1 menjadi *testing*, bagian sisanya menjadi data *training*, kemudian hitung akurasi menggunakan persamaan berikut :

$$\text{Akurasi} = \frac{Jumlah\ klasifikasi\ benar}{Jumlah\ data\ uji} \times 100\ \%$$

Keterangan :

Jumlah klasifikasi benar : jumlah prediksi klasifikasi yang tepat

Jumlah data uji : jumlah dataset yang digunakan untuk testing

Pada *fold* ke 2, dimana bagian ke 2 yang menjadi *testing*, sisanya menjadi *training*, kemudian hitung akurasinya, proses tersebut berulang sampai mencapai *fold* ke -k. Hitung rata-rata dari semua nilai k, hasil akurasi

tersebut merupakan hasil akurasi akhir. Pada proses validasi dilakukan pembuatan model, dalam penelitian ini menggunakan metode e*nsembl*e dengan teknik *boosting*, yaitu *adaboost* dan menggunakan *decision Tree C4.5* dan KNN pada data *training.* Setelah itu dilanjutkan proses evaluasi dengan *confusion table* dan ROC *curve*. Hasil *confusion table* digunakan untuk menyajikan *accuracy, recall,* dan *precision* dalam algortima klasifikasi. *Accuracy* merupakan persentase antara nilai prediksi dengan nilai sebenarnya yang ada. *Recall* merupakan persentase nilai kinerja keberhasilan algoritma yang dipakai. *Precision* merupakan nilai akurasi dengan class yang telah diprediksi. Berikut merupakan *tabel confusion*:

Tabel 1. *Confusion Table*

Confusion matrix		Nilai Prediksi	
		Positive	*Negative*
Nilai Sebenarnya	*Positive*	TP	FP
	Negative	FN	TN

Rumus *Accuracy*: $Accuracy = \frac{TP+TN}{(TP+FP+TN+FN)}$

Rumus *Recall* : $Recall = \frac{TP}{(TP+FN)}$

Rumus *Precision* : $Precision = \frac{TP}{(TP+FP)}$

Keterangan :
TP : *True Positive*
FP : *False Positive*
TN : *True Negative*
FN : *False Negative*
Receiver Operating Characteristic (ROC) digunakan untuk evaluasi hasil akurasi dalam bentuk grafik. ROC merupakan kurva yang akan menghasilkan nilai *Area Under Curve* (AUC). AUC merupakan nilai akurasi area dibawah kurva yang dihasilkan oleh ROC [18]. Keakurasian nilai AUC dapat diklasifikasi menjadi 5 kelompok [19] antara lain, yaitu:

1) 0.90 – 1.00 = *Exellent Classification*
2) 0.80 – 0.90 = *Good Classification*
3) 0.70 – 0.80 = *Fair Classification*
4) 0.60 – 0.70 = *Poor Classification*
5) 0.50 – 0.60 = *Failure*

Hasil dan Pembahasan

Penelitian yang dilakukan yaitu pengujian 1 menggunakan algoritma klasifikasi *decision Tree C4.5*, pengujian 2 menggunakan teknik *discretization* dan metode *adaboost* dengan algoritma klasifiksi *decision Tree C4.5* , pengujian 3 menggunakan algoritma klasifikasi KNN, dan pengujian 4 menggunakan teknik *discretization* dan metode *adaboost* dengan algoritma klasifiksi KNN menghasilkan nilai *accuracy, recall, precision,* dan AUC. Berikut tabel hasil dari 4 pengujian yang telah dilakukan

Tabel 2. Hasil Pengujian

	Acc	*Recall*	*Precision*	AUC
Decision Tree	89,17%	94,70%	85,96%	0,953
Decision Tree +Discretize+ Adaboost	99,81%	100%	99,64%	0,700
KNN	84,68%	72,04%	97,60%	0,953
KNN+ Discretize+ Adaboost	92,88%	86,52%	99,57%	0.931

Berdasarkan hasil yang diperoleh menunjukkan bahwa adanya peningkatan nilai *accuracy, recall, precision,* dan nilai AUC. Pada algoritma *Decision Tree C4.5* dan KNN dengan menggunakan teknik *discrtize* dan metode *adaboost* meningkat dibandingkan jika hanya menggunakan satu teknik pembelajaran. Metode paling unggul diperoleh dari pengujian II dengan *accruracy* sebesar 99,81%.

Peningkatan tersebut terjadi karena adanya faktor perubahan atribut dari nominal menjadi interval dengan menggunakan teknik *diskritisasi* selain itu pemberian bobot pada algoritma tunggal dengan metode *adaboost* dapat meningkatkan akurasi algoritma klasifikasi

Kesimpulan dan Saran

Metode paling unggul dihasilkan dari pengujian menggunakan penggabungan 3 metode yaitu teknik *discretization*, metode *adaboost*, dan *decision tree C4.5* dalam mendiagnosa penyakit jantung. Hasil akurasi yang diperoleh yaitu 99,81% dan meningkat sebesar 10,64% dari hasil Decision Tree tunggal sebesar 89,17%. Penggabungan 3 metode yaitu teknik *discretization*, metode *adaboost*, dan

decision tree C4.5 dapat diterapkan untuk penelitian yang akan datang menggunakan dataset lain untuk meningkatkan akurasi.

Hasil pengujian yang telah dilakukan kedepannya dapat diimplementasikan dalam bentuk aplikasi untuk mendiagnosa penyakit jantung. Pengujian dapat menggunakan *dataset* lain untuk menguji seberapa akurat hasil yang diperoleh dalam pengujian. Pengembangan dengan metode atau teknik lain perlu dilakukan untuk menghasilkan akurasi yang lebih tinggi.

Ucapan Terima Kasih (*Acknowledgement*)

Kami ucapkan terima kasih kepada DPPM Universitas Pelita Bangsa yang telah mendukung secara finansial sehingga terlaksanala Peneitian ini. Kami juga mengucapkan terima kasih kepada seluruh pihak yang sudah embantu dalm terwujudnya penelitian ini.

Referensi

[1] [1] Kementrian Kesehatan Republik Indonesia, "Hasil Riskesdas 2013," *Expert Opin. Investig. Drugs*, vol. 7, no. 5, pp. 803–809, 2013, doi: 10.1517/13543784.7.5.803.

[2] Kementrian Kesehatan Republik Indonesia, "Laporan Riskesdas 2018 Nasional.pdf." p. 674, 2019.

[3] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthc. Anal.*, vol. 2, no. February, p. 100060, 2022, doi: 10.1016/j.health.2022.100060.

[4] A. Alhamad, A. I. S. Azis, B. Santoso, and S. Taliki, "Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 352, 2019, doi: 10.26418/jp.v5i3.37188.

[5] C. F. Tsai and Y. C. Chen, "The optimal combination of feature selection and data discretization: An empirical study," *Inf. Sci. (Ny).*, vol. 505, pp. 282–293, 2019, doi: 10.1016/j.ins.2019.07.091.

[6] R. T. Vulandari, *Data Mining : Teori dan Aplikasi Rapidminer*. Yogyakarta: Penerbit Gava Media, 2017.

[7] Suyanto, *Data Mining Untuk Klasifikasi dan Klastering Data*. Bandung: Informatika Bandung, 2017.

[8] S. Guggari, V. Kadappa, and V. Umadevi, "Non-sequential partitioning approaches to decision tree classifier," *Futur. Comput. Informatics J.*, vol. 3, no. 2, pp. 275–285, 2018, doi: 10.1016/j.fcij.2018.06.003.

[9] A. Nurzahputra and M. A. Muslim, "Peningkatan Akurasi Pada Algoritma C4.5 Menggunakan Adaboost Untuk Meminimalkan Resiko Kredit," 2017.

[10] N. E. I. Karabadji, I. Khelf, H. Seridi, S. Aridhi, D. Remond, and W. Dhifli, "A data sampling and attribute selection strategy for improving decision tree construction," *Expert Syst. Appl.*, vol. 129, pp. 84–96, 2019, doi: 10.1016/j.eswa.2019.03.052.

[11] Y. Guo, S. Han, Y. Li, C. Zhang, and Y. Bai, "K-Nearest Neighbor combined with guided filter for hyperspectral image classification," *Procedia Comput. Sci.*, vol. 129, pp. 159–165, 2018, doi: 10.1016/j.procs.2018.03.066.

[12] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A generalized mean distance-based k-nearest neighbor classifier," *Expert Syst. Appl.*, vol. 115, pp. 356–372, 2019, doi: 10.1016/j.eswa.2018.08.021.

[13] S. Ramírez-Gallego, S. García, and F. Herrera, "Online entropy-based discretization for data streaming classification," *Futur. Gener. Comput. Syst.*, vol. 86, pp. 59–70, 2018, doi: 10.1016/j.future.2018.03.008.

[14] S. Cheng, B. Liu, Y. Shi, Y. Jun, and B. Li, *Data Mining and Big Data*. 2015.

[15] E. Listiana and M. A. Muslim, "Penerapan Adaboost Untuk Klasifikasi Support Vector Machine Guna Meningkatkan Akurasi Pada Diagnosa Chronic Kidney Disease," no. 2015, pp. 35–40, 2017.

[16] M. A. Banjarsari, I. Budiman, and A.

Farmadi, "Penerapan K-Optimal Pada Algoritma Knn Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan Ip Sampai Dengan Semester 4," *Klik - Kumpul. J. Ilmu Komput.*, vol. 2, no. 2, pp. 159–173, 2016, doi: 10.20527/KLIK.V2I2.26.

[17] Indrayanti, D. Sugianti, and M. A. Al Karomi, "Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus," *SNATIF*, pp. 551–554, 2017.