Jurnal EDU ELEKTROMATIKA (JEE)
ISSN: 2747-0784 (p); xxxxxx (e)
Vol 5,No. 1, Juni 2024

PRE PROCESSING K-NN METHOD USING K-MEANS OPTIMISATION FOR IMMUNOTHERAPY CLASSIFICATION IN SKIN CANCER DISEASE

PRE PROCESSING METODE K-NN MENGGUNAKAN OPTIMASI *K-MEANS* UNTUK KLASIFIKASI *IMMUNOTHERAPY* PADA PENYAKIT KANKER KULIT

[1a]Ary Akhyar, [1a]Zainal Abidin M.Kom, [1a]Bijanto M.Kom
[1a]Program Studi Informatika, STTP
e-mail : aryart1604@gmail.com, zainal.frsd@yahoo.co.id, bijanto@sttp.ac.id

Abstrack *The existence of data outliers in the dataset can cause low accuracy in the classification process. Outliers contained in the dataset can be removed or deleted at the pre-processing stage of the classification algorithm. A method that is often used to detect data outliers is the boxplot. A boxplot is a diagram based on a summary of five numbers, namely the first quartile (Q1), the median or second quartile (Q2), the third quartile (Q3), the minimum value and the maximum value. In this study, pre-processing of the K-NN method will be carried out using K-Means optimization for the classification of immunotherapy in skin cancer. This study uses a dataset of immunotherapy in the treatment of skin cancer, totaling 90 instances, with 8 attributes and 2 classes. The results of this study obtained the highest accuracy rate of 97.76% with an error of 2.24% with the K-NN classification using a 5-fold cross validation scheme. This study also obtained a recall value of 100% with a precision value of 97.57%.*

Keywords*: immunotherapy dataset, boxplot, K-Means, K-NN, outliers, cross validation.*

Abstrak Adanya data *outliers* pada dataset dapat menyebabkan rendahnya hasil akurasi pada proses klasifikasi. *Outlier* yang terdapat pada *dataset* dapat dihilangkan atau dihapus pada tahapan *pre processing* algoritma klasifikasi. Sebuah metode yang sering digunakan untuk mendeteksi data *outliers* yaitu *boxplot. Boxplot* merupakan sebuah diagram yang dibuat dengan berlandaskan pada ringkasan lima angka yaitu kuartil pertama (*Q1*), median atau kuartil kedua (*Q2*), kuartil ketiga (*Q3*), nilai minimum, dan nilai maksimum. Dalam penelitian ini akan melakukan *pre processing* metode K-NN menggunakan optimasi *K-Means* untuk klasifikasi *immunotherapy* pada penyakit kanker kulit. Kajian ini menggunakan *dataset immunotherapy* pada pengobatan kanker kulit yang berjumlah 90 *instance*, dengan 8 atribut dan 2 kelas. Hasil dari penelitian ini didapatkan tingkat akurasi tertinggi sebesar 97,76% dengan *error* sebesar 2,24% dengan klasifikasi K-NN menggunakan skema 5-*fold cross validation*. Penelitian ini juga mendapatkan nilai *recall* 100% dengan nilai *precission* sebesar 97,57%.

Kata kunci : *immunotherapy dataset, boxplot, K-Means*, K-NN, *outliers, cross validation.*

PENDAHULUAN

Pada setiap manusia pastinya dalam menjaga kesehatan adalah hal yang paling penting bagi setiap manusia, dengan tubuh yang sehat kita dapat melakukan segala hal aktifitas dan produktifitas sehari-hari (Fazriansyah et al., 2020). Kesehatan merupakan suatu keadaan yang dinamis, dipengaruhi faktor genetik, lingkungan dan pola hidup sehari-hari seperti makan, minum, kerja, dan istirahat

yang Kanker merupakan beban besar bagi masyarakat di negara-negara yang kurang berkembang secara ekonomi (Sulaiman et al., 2019). Terjadinya kanker meningkat karena pertumbuhan dan penuaan penduduk, serta peningkatan prevalensi faktor risiko seperti merokok, kelebihan berat badan, kurangnya aktivitas fisik, dan perubahan pola reproduksi yang terkait dengan urbanisasi dan pembangunan ekonomi (Widowati & Mudahar, 2009). Kanker merupakan salah satu penyakit utama penyebab kematian di dunia. Pada 2012 diperkirakan terdapat 14 juta kasus baru kanker dan 8,2 juta kematian akibat kanker di dunia (Binabar & Ivandari, 2017). Penggunaan tembakau merupakan faktor penyebab kematian pada kanker secara umum (Sung et al., 2021). Hal ini mempengaruhi perkembangan pada penelitian terbaru dalam menemukan obat baru dengan bahan alami kini banyak diteliti untuk pengobatan penyakit kanker terutama jenis kanker kulit yang sekarang banyak dialami (Pamuji & Ramadhan, 2021).

Diagnosis kanker kulit merupakan salah satu cabang dari bidang kesehatan yang digunakan untuk mendeteksi adanya kelainan pada kulit (Maryam & Ariono, 2022). Kanker kulit dapat diklasifikasikan dalam tiga tipe terbanyak yaitu *karsinoma sel basal*, *karsinoma sel skuamosa*, dan *melanoma maligna* (Supriyatna & Mustika, 2018). Salah satu jenis terapi non pembedahan adalah *immunotherapy*. Metode pengobatan ini menggunakan imun cytokine memiliki efek samping yang besar dan metode ini terus dikembangkan (Rahmadi et al., 2020).

Algoritma K-Means dan K-Nearst Neighbor merupakan algoritma data mining yang umum digunakan. Dalam penelitian ini algoritma K-Means digunakan sebagai Optimasi hasil dari pendeteksian outlier pada dataset, sedangkan algoritma K-Nearst Neighbor digunakan sebagai metode klasifikasi. Penelitian ini dilakukan dengan tujuan untuk menggabungkan lebih dari satu algoritma, serta dapat meminimalisir adanya outlier dalam memprediksi keberhasilan metode immunotherapy pada penyakit kanker kulit.

Pada penelitian sebelumnya menunjukkan bahwa *Immunotherapy* memiliki tingkat keberhasilan yang lebih baik dibandingkan dengan metode *Cryotherapy ,*dari hasil penelitian yang dilakukan Fahime Khozeimeh, *et al* (2017) dengan judul *Intralesional immunotherapycompared to cryotherapy in the treatment of warts* , menyatakan bahwa hasil yang diperoleh melalui metode immunotherapy dapat memberikan efek positif dengan tingkat respon yang lebih tinggi yaitu 76,7% pasien sembuh total selain itu, terbukti menjadi pengobatan yang memiliki respons terapeutik yang tinggi, jumlah sesi pengobatan yang diperlukan lebih sedikit dan resikoefek samping yang lebih rendah. Sementara hanya 57,7% pasien sembuh dengan metode cryotherapy (Khozeimeh, Jabbari Azad, et al., 2017), (Khozeimeh, Alizadehsani, et al., 2017).

METODE PENELITIAN

Tahapan penelitian ini meliputi persiapan data, *pre processing* data, optimasi data hasil *pre processing*, klasifikasi data, dan evaluasi.

1. Persiapan Data

Data yang digunakan pada penelitian ini yaitu Immunotherapy mengenai pengobatan kanker kulit. Dataset tersebut merupakan data publik atau data sekunder yang bukan berasal langsung dari peneliti tetapi dari sumber lain (Khozeimeh, Alizadehsani, et al., 2017). Dataset yang digunakan dalam penelitian ini diperoleh dari UCI Repository of machine learning dataset

(Khozeimeh, Alizadehsani, et al., 2017). Data immunotherapy yang digunakan memiliki 8 atribut, dengan satu atribut sebagai kelas dan 7 atribut sebagai fitur.

Tabel 1. *Atribut Immunotherapy Dataset*

No	Fitur atau Atribut	Deskripsi
1	*Sex*	Jenis Kelamin Pasien (41 Laki-laki, 49 Wanita)
2	*Age*	Usia Pasien (15-56)
3	*Time*	Waktu berlalu sebelum perawatan (*month*, 0-12)
4	*Number_of_Warts*	Angka Kanker Kulit (1-19)
5	*Type*	Jenis Kanker Kulit (1-*Papilloma*, 2-*Plantar*, 3-*Both*).
6	*Area*	Luas Permukaan Kanker Kulit (mm^2, 6-900)
7	*Induration_Diameter*	Diameter Indurasi Tes Awal (mm, 5-70)
8	*Result_of_Treatment*	Hasil Pengobatan (*Yes/No*)

Atribut kelas label dataset memiliki dua nilai, yaitu *yes* sebagai berhasil dan *no* sebagai gagal. Tabel 1 menyatakan daftar atribut-atribut, tipe data, dan deskripsi atribut pada dataset tersebut. Untuk alur keseluruhan dalam perancangan sistem penelitian ini ditunjukan pada gambar 1.

2. Pre Processing Data

Tujuan dari *pre processing* adalah mendeteksi serta menghapus *oulier* pada dataset. Metode yang digunakan dalam *pre processing* ialah metode *boxplot*. *Boxplot* merupakan salah satu metode yang umum digunakan untuk mendeteksi outlier data berdasarkan 5 ukuran yaitu nilai minimum, kuartil pertama *(Q1),* kuartil kedua atau median *(Q2)*, kuartil ketiga *(Q3)*, dan nilai maksimum (Soemartini, 2007).

Berikut dijelaskan langkah detail pada proses metode *boxplot* untuk deteksi dan penghapusan *outlier* dari dataset *immunotherapy*. Proses pertama dimulai dengan menentukan alokasi dari nilai minimum, *Q1, Q2, Q3*, dan nilai maksimum pada dataset. Langkah kedua adalah menentukan nilai *IQR (Inter Quatile)* yang akan digunakan untuk mencari nilai batas atas yang dinyatakan pada persamaan 2, sedangkan bawah pencilan (*outlier*) pada persamaan 3.

Langkah ketiga adalah mencari nilai batas atas dan batas bawah pencilan berdasarkan nilai *IQR* yang didapat seperti yang ditunjukan pada persamaan 1.

$$Rumus\ (IQR) = Q3 - Q1 \quad (1)$$

Langkah keempat adalah menentukan data outlier, untuk menentukan ada data outlier atau tidak yaitu jika data pada masing-masing atribut lebih kecil dari nilai batas bawah yang sudah ditentukan maka akan dinyatakan ada data *outliers*, dan data pada masing-masing atribut lebih besar dari nilai batas atas yang sudah ditentukan maka dinyatakan terdapat data *outliers*.

$$Rumus\ (BBP) = Q1 - (1.5 \times IQR) \quad (2)$$

$$Rumus\ (BAP) = Q3 + (1.5 \times IQR) \quad (3)$$

3. Optimasi Data

Pada tahap optimasi penelitian ini menggunakan metode *k-means* dengan tujuan untuk mengoptimalkan data hasil data sebelum proses klasifikasi. Tahapan ini mengelompokkan data hasil dari metode *boxplot* untuk mendeteksi *outliers* yang kemudian akan didapatkan data baru hasil dari pengolahan menggunakan algoritma *k-means*. Langkah pertama proses optimasi data dimulai dengan menentukan jumlah *cluster* sesuai kelas pada dataset. Langkah kedua adalah mencari pusat *cluster* atau *centroid* menggunakan *k-means*. Algoritma *k-means* ini dilakukan dengan menentukan *k* titik pusat *cluster* secara acak. Data dikelompokkan berdasarkan jarak data ke setiap pusat *k cluster*. Langkah untuk menemukan *cluster* dan pengelompokkan tersebut diulang sampai nilai pusat *cluster* tidak lagi berubah.

$$de = \sqrt{(xi - si)^2 + (yi - ti)^2} \quad (4)$$

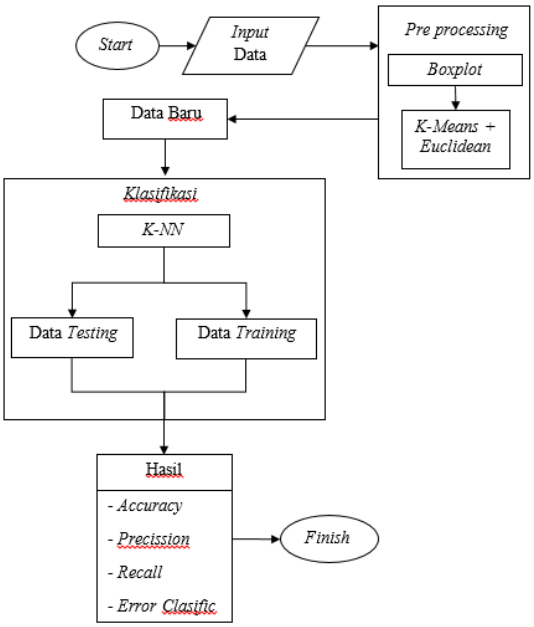
Langkah ketiga adalah menghitung jarak semua *instance* dengan masing-masing pusat *cluster* yang dihasilkan langkah kedua menggunakan *matriks* jarak. *Matriks* jarak yang digunakan adalah *euclidean* yang dinyatakan pada persamaan 4. Langkah keempat adalah menemukan kelas label baru berdasarkan jarak terdekat dari masing-masing *instance* dengan pusat *cluster*.

4. Klasifikasi Data Dan Evaluasi

Tahapan klasifikasi merupakan tahapan yang utama dalam proses analisis data. Pada bagian ini akan menggunakan algoritma klasifikasi K-NN pada *training* data yang telah diolah dari tahapan *pre processing* dan *clustering* dengan menggunakan pengujian *k-fold cross validation*.

Langkah pertama dalam tahapan ini yaitu membagi data dengan skema *k-fold cross validation*. Dalam penelitian ini mengusulkan 5-*fold cross validation,* karena dalam buku yang ditulis oleh Hastie *et al* (2008), dengan model pengujian *k*=5 atau 10 dapat digunakan untuk memperkirakan tingkat kesalahan yang terjadi, sebab data *training* pada setiap *fold* cukup berbeda dengan data *training* yang asli (Fajar Rodiyansyah & Winarko, 2013). Oleh karena itu, Secara keseluruhan 5 atau 10-*fold cross validation* sama-sama direkomendasikan dan disepakati bersama (Fajar Rodiyansyah & Winarko, 2013) (Utomo & Mesran., 2020). Artinya *dataset* akan dibagi menjadi 5 bagian total keseluruah data. Satu bagian (data *testing*) digunakan untuk pengujian dan sisanya digunakan untuk data *training*. Langkah-langkah pada tahapan klasifikasi dan evaluasi meliputi pembagian dataset menjadi data latih 90 % dan 10 % untuk data uji, pemrosesan data uji dengan algoritme klasifikasi kNN dengan parameter k=5, dan pengukuran performa model dalam bentuk *confusion matrix*. Performa yang diukur adalah akurasi, *error, recall,* dan *precission*.

Jurnal EDU ELEKTROMATIKA (JEE)
ISSN: 2747-0784 (p); xxxxxx (e)
Vol 5,No. 1, Juni 2024



Gambar 1. Perancangan Metode yang Diusulkan

HASIL DAN PEMBAHASAN

Identifikasi jumlah data dari masing-masing kelas yang dimiliki dilakukan setelah penyiapan dataset. Dataset memiliki jumlah data 90 instance. Distribusi data berdasarkan atribut kelas yang ditunjukan pada Tabel 2, yaitu *yes* sejumlah 71 dan *no* sejumlah 19. Langkah pertama dari tahapan *pre processing* yaitu dengan metode yang diusulkan menggunakan *boxplot* untuk mendeteksi adanya *outlier* pada dataset. Untuk mendeteksi *outlier*, metode *boxplot* menghitung berdasarkan 5 ukuran statiskik meliputi nilai minimum, *Q1, Q2, Q3*, dan nilai maksimum.

Tabel 2. Jumlah Distribusi Data Tiap Kelas

No	Kelas	*Result*	Jumlah Data
1	*Yes*	1	71
2	*No*	0	19
Total data			90

Langkah kedua menentukan alokasi dari 5 nilai statistik tersebut, kemudian mencari nilai *IQR* yang akan digunakan untuk menentukan nilai dari batas atas dan batas bawah pencilan atau *outlier*. Setelah itu, menghitung untuk mencari nilai batas atas (BAP) dan batas bawah (BBP) berdasarkan nilai *IQR* yang sudah ditentukan seperti pada Tabel 3.

Tabel 3. Hasil Perhitungan IQR, BAP, BBP, dan BAPj

Attribut	Min	*Q1*	*Q2*	*Q3*	*Max*	*IQR*	BBP	BAP	BAPj
Age	15	15,5	21	27,5	35	12	-2,5	45,5	51,5
Time	1,75	5,25	6,62	8,62	12	3,37	0,18	13,68	15,37
NOW	1	2	5	9,5	19	7,5	-9,25	20,75	24,5
Area	6	27.5	48.5	92	900	64,5	-69,25	188,75	221
Diameter	2	5,5	7	8	70	2,5	1,75	11,75	13

Langkah terakhir dari metode prapemrosesan yang diusulkan adalah menghapus data-data outlier yang ditemukan. Data yang dibersihkan dengan menggunakan metode boxplot berdasarakan hasil dari batas atas dan batas bawah yang sudah ditentukan seperti pada Tabel 3.

Tahapan pengelompokan yang bertujuan untuk mengoptimalkan data sebelum proses klasifikasi. Tahapan ini mengelompokkan data hasil dari metode *boxplot* untuk mendeteksi *outlier* yang kemudian akan didapatkan data baru hasil dari pengolahan menggunakan algoritma *k-means*. Pengelompokan menggunakan algoritma *k-means* dalam perhitungannya menggunakan Euclidean seperti pada persamaan 4. Langkah pertama dari proses pengelompokan adalah menentukan jumlah cluster disesuaikan dengan jumlah kelas pada data. Langkah kedua, menentukan pusat cluster atau centroid pada data secara acak.

Langkah ketiga adalah menghitung jarak objek dengan *centroid*. Untuk menghitung jarak antara objek dengan pusat *cluster (centroid)* dapat dilakukan dengan menggunakan beberapa pendekatan. Pada penelitian ini digunakan rumus *euclidean distance*. Dataset optimasi dengan menggunakan metode k-means dan evaluasi *matriks* jarak *euclidean* menghasilkan dataset seperti pada Tabel 4.

Tabel 4. Distribusi Jumlah Data dari Hasil Optimasi K-Means

No	Kelas	Data Awal	Data Baru
1	Yes	71	84
2	No	19	6
Total data		90	90

Pada proses selanjutnya yaitu klasifikasi. Dalam penelitian ini adalah dengan menggunakan data baru yang dihasilkan dari proses tahapan optimasi menggunakan k-means. Langkah pertama, membagi data dengan skema *k-fold cross validation* yaitu k=5 dengan parameter 3, 5, 7, dan 9 untuk perbandingan k-optimal.

Langkah kedua, menghitung pengujian berdasarkan confusion matrix untuk mencari nilai akurasi, *error classification, recall,* dan *precission* dari masing–masing 5 pengujian menggunakan cross validation.

KESIMPULAN

Berdasarkan penelitian, implementasi dan pengujian, maka dapat diambil kesimpulan sebagai berikut :

Dari hasil pengujian menyimpulkan bahwa penggabungan dengan dua algoritma yaitu k-means dan K-NN menunjukkan bahwa akurasi terbaik terkait dengan metode pengobatan immunotherapy dataset sebesar 97,76% dan hasil lain pada tabel 5.

Tabel 5. Hasil Pengujian

Parameter	***Accuracy***	***Error***	***Precission***	***Recall***
k=3	0,9776	0,0224	0,9757	1
k=5	0,9776	0,0224	0,9757	1
k=7	0,9554	0,0446	0,9548	1
k=9	0,9444	0,0556	0,9444	1

DAFTAR PUSTAKA

Binabar, S. W., & Ivandari. (2017). Optimasi Parameter K pada Algoritma KNN untuk

Deteksi Penyakit Kanker Payudara. *IC-Tech*, *XII*(2), 11–18.

Fajar Rodiyansyah, S., & Winarko, E. (2013). Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *7*(1), 13. https://doi.org/10.22146/ijccs.3048

Fazriansyah, A., Azis, M. A., & Yudhistira. (2020). Analysis Of Neural Network Clasification Algorithm To Know The Success Level Of Immunotherapy. *Jurnal Techno Nusa Mandiri*, *17*(1), 57–62.

Khozeimeh, F., Alizadehsani, R., Roshanzamir, M., Khosravi, A., Layegh, P., & Nahavandi, S. (2017). An expert system for selecting wart treatment method. *Computers in Biology and Medicine*, *81*(January), 167–175. https://doi.org/10.1016/j.compbiomed.2017.01.001

Khozeimeh, F., Jabbari Azad, F., Mahboubi Oskouei, Y., Jafari, M., Tehranian, S., Alizadehsani, R., & Layegh, P. (2017). Intralesional immunotherapy compared to cryotherapy in the treatment of warts. *International Journal of Dermatology*, *56*(4), 474–478. https://doi.org/10.1111/ijd.13535

Maryam, M., & Ariono, H. W. (2022). Sistem Pakar Pengklasifikasi Stadium Kanker Serviks Berbasis Mobile Menggunakan Metode Decision Tree. *Jurnal Kajian Ilmiah*, *22*(3), 267–278. https://doi.org/10.31599/jki.v22i3.1368

Pamuji, F. Y., & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy. *Jurnal Teknologi Dan Manajemen Informatika*, *7*(1), 46–50.

Rahmadi, M., Kaurie, F., & Susanti, T. (2020). Uji Akurasi Dataset Pasien Pasca Operasi Menggunakan Algoritma Naïve Bayes Menggunakan Weka Tools. *JURIKOM (Jurnal Riset Komputer)*, *7*(1), 134. https://doi.org/10.30865/jurikom.v7i1.1761

Soemartini. (2007). *Pencilan (Outlier)*.

Sulaiman, F. H., Yulianti, K., & Serviana, H. (2019). Model Matematika Terapi Kanker Menggunakan Kemoterapi, Imunoterapi Dan Biochemotherapy. *Jurnal EurekaMatika*, *7*(1), 1–10.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). *Global Cancer Statistics 2020 : GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*. *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Supriyatna, A., & Mustika, W. P. (2018). Komparasi Algoritma Naive bayes dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil. *Jurnal Sains Komputer & Informatika*, *2*(2), 152–161.

Utomo, D. P., & Mesran. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma*, *4*(2), 437–444. https://doi.org/10.30865/mib.v4i2.2080

Widowati, L., & Mudahar, H. (2009). *Ujiaktivitas ekstrak etanol 50% umbi keladi tikus (typhoniumflagelliforme (lood) bi) terhadap sel kanker payudara mcf-7 in vitro*. *XIX*, 3–8.