Penerapan K-Means *Clustering* pada *Imbalance dataset* Gejala Penyakit Jantung

Meida Cahyo Untoro[1,*], Leonard Rizta[2], Anugerah Perdana[3], Nazla Andintya Wijaya[4], Nestiawan Ferdiyanto[5]

Teknik Informatika, Institut Teknologi Sumatera, Indonesia

[1]cahyo.untoro@if.itera.ac.id

*corresponding author

INFO ARTIKEL

Sejarah Artikel
Diterima: 10 April 2022
Direvisi: 5 Mei 2022
Diterbitkan: 25 April 2023

Kata Kunci
Akurasi
Imbalance
Jantung
K-Means
Missclasification

ABSTRAK

Jantung merupakan organ yang sangat penting bagi tubuh kita, namun ternyata penyakit jantung merupakan penyakit paling mematikan di dunia. Penyebab penyakit jantung beragam, seperti pola hidup yang tidak sehat, mengkonsumsi makan yang mengandung tinggi kolestrol, merokok, lingkungan serta faktor keturunan atau genetik. Penyakit jantung tidak mengenal jenis kelamin dan usia, baik perempuan, laki-laki, balita, anak-anak bahkan orang dewasa dapat menderita penyakit jantung. Gejala penyakit jantung antara lain sakit dada, tekanan darah tinggi, hasil tes Elektrodiogram (EKG), denyut jantung, kadar gula dan kandungan kolesterol yang meningkat atau tidak normal. Minimnya data terkait penyakit jantung menyebabkan kesulitan untuk melakukan prediksi terhadap pasien yang menderita atau tidak. Data minim dapat diartikan sebagai data minoritas yang akan menyebabkan missclasifikation pada dataset utaman dan berdampak pada prediksi yang salah. Penelitian ini bertujuan memprediksi gejala penyakit jantung dengan menggunakan K-means *clustering* untuk mengelompokan kelas berpenyakit jantung atau minoritas dan kelas sehat atau kelas mayoritas. Hasil dari penelitian menunjukkan bahwa K-means memiliki nilai evaluasi *clustering* 0.5 dengan metode Elbow dan nilai akurasi 89%. Penerapan oversampling teknik pada dataset penyakit jantung memberikan peningkatan pada evaluasi elbow menjadi 0.8 dan 93% akurasi. Pada Penelitian ini dataset yang tidak seimbang memberikan nilai akurasi pada missclasification sebesar 4% dari dataset seimbang.

PENDAHULUAN

Jantung adalah organ penting bagi manusia karena berperan dalam sistem peredaran darah. Penyakit jantung merupakan kondisi saat jantung tidak bisa melakukan tugasnya dengan baik. Gangguan pada fungsi jantung dan pembuluh darah sering kerap kali dikenal sebagai kardiovaskular [1]. Penyakit kardiovaskular merupakan salah satu penyakit paling mematikan dan berbahaya di dunia. Menurut WHO pada tahun 2006, angka kematian penyakit ini sejumlah 17,5 juta termasuk pada penyakit jantung koroner, stroke, dan penyakit jantung rematik. Pada tahun 1990 jumlah kematian terus bertambah hingga 3 juta jiwa [2]. Menurut American Heart Association, pria lebih mungkin menderita penyakit kardiovaskular utama sebelum usia 60 tahun. Namun kenyataannya di Amerika, penyakit ini pun masih menjadi penyakit mematikan bagi wanita [3]. Sekitar 15 dari 1000 orang di Indonesia menderita penyakit jantung ((Riset Kesehatan Dasar (Riskesdas), 2018) [4].

Penyebab penyakit jantung cukup banyak seperti pola hidup yang tidak sehat dan makan-makanan yang mengandung tinggi kolestrol. Namun faktor lain seperti genetik dan

lingkungan juga turut berpengaruh. Faktor gejala penyakit jantung antara lain kolesterol, sakit dada (*chest pain*), denyut jantung, tekanan darah tinggi, nilai tes EKG, dan kadar gula [5]. Resiko penyakit jantung yang berbahaya ini menyebabkan banyak penderitanya tidak sadar apabila mengalami gejala penyakit jantung. Apabila tidak di deteksi sejak awal maka dapat menyebabkan komplikasi atau bahkan kematian. Deteksi penyakit jantung ini merupakan hal penting untuk mencegah kematian penderita. Salah satu cara untuk mendeteksi gejala penyakit jantung ini memanfaatkan data mining. Data mining dapat diterapkan dalam bidang ilmu kedokteran. Pada data mining terdapat banyak teknik seperti *clustering*, klasifikasi, asosiasi, dan prediksi [4].

Dalam deteksi gejala penyakit jantung, teknik yang digunakan penulis adalah *clustering*. *Clustering* memiliki cara kerja mencari dan mengelompokkan data yang mempunyai kesamaan atau kemiripan [6]. Dalam *clustering*, terdapat metode K-Means, metode ini mempunyai sifat efisien yang bertujuan membuat *cluster* objek berdasar atribut menjadi k partisi [7][8]. Pada penelitian yang dilakukan oleh Atik Nurmasani dan Yoga Pristyanto yang berjudul "Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class", para peneliti melakukan klasifikasi penyakit jantung dengan algoritma stacking dengan dataset yang tidak seimbang (*imbalance data*). Hasil dari penelitian ini adalah algoritma SVM mampu menghasilkan lebih baik nilai akurasi, AUC, TPR, TNR, dan G-Mean daripada single classifier lainnya pada Tree C4.5. Hal ini terjadi karena pada dataset terdapat data yang tidak seimbang (*imbalanced data*).

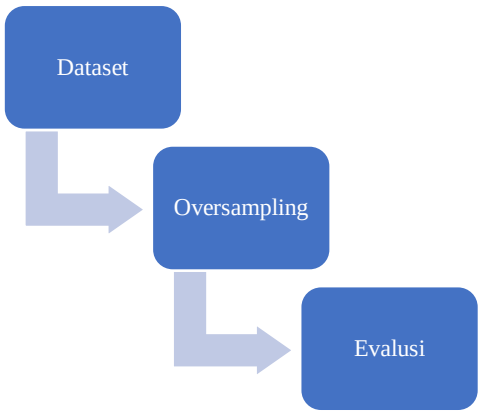
Berkaitan dengan banyaknya gejala dan faktor lain dalam pedeteksian penyakit jantung, maka diperlukan pengelompokkan jenis penyakit jantung. Dataset dari Kaggle menujukkan terdapat 14 variabel yang menjadi penentu gejala penyakit jantung. Dataset pada Kaggle ini juga bersifat *imbalance data*, maka diperlukan metode untuk membuat datanya seimbang yaitu *oversampling*. *Oversampling* adalah cara menangani permasalahan *imbalanced* dengan mendistribusikan data yang seimbang dengan replikasi data sintetik minoritas secara acak dengan melakukan iterasi [9]. Penelitian ini bertujuan untuk mengelompokkan dan membuat data sintetis pada dataset gejala penyakit jantung sesuai *cluster*-nya agar dapat memprediksi dengan tingkat akurasi yang representatif dan optimal.

METODE

Penelitian *imbalanced dataset* gejala penyakit jantung ini memiliki 3 tahap penyelesain (Gambar 1). Pada tahap pertama dilakukan pengambilan dataset dan preprosesing data. Pada tahap kedua dataset yang tidak seimbang akan dilakukan proses pemisahan antara data mayoritas dan minoritas. Data minoritas akan dijadikan master atau *parent* dalam pembuatan data sintetis dan terbentuk dengan jumlah data mayoritas dikurangi jumlah data minoritas. Tahap terakhir dari penelitian ini dilakukan evaluasi terhadap model yang telah dibentuk oleh data mayoritas, data minoritas serta data sintetis.

Dataset Penyakit Jantung

Penelitian dilakukan dengan menggunakan dataset dari kaggle Heart Disease UCI yang diunggah pada tahun 2019 dengan total 303 records (Tabel 1). Atribut pada dataset tersebut berisi 14 atribut, yaitu umur (age), jenis kelamin (sex), sakit dada (cp), tekanan darah istirahat (trestbps), kolesterol (chol), gula darah puasa (fbs), elektrokardiografi istirahat (restecg), detak jantung maksimum tercapai (thalach), angina (exang) yang diinduksi olahraga, depresi ST yang disebabkan oleh latihan relatif (oldpeak), kemiringan puncak latihan segmen ST (slope), jumlah pembuluh darah besar (0-3) diwarnai oleh flourosopy (ca), 3 = normal; 6 = cacat tetap; 7 = cacat reversibel (thal).



Gambar 1 Alur Penelitian

Tabel 1 Dataset Gejala Penyakit Jantung

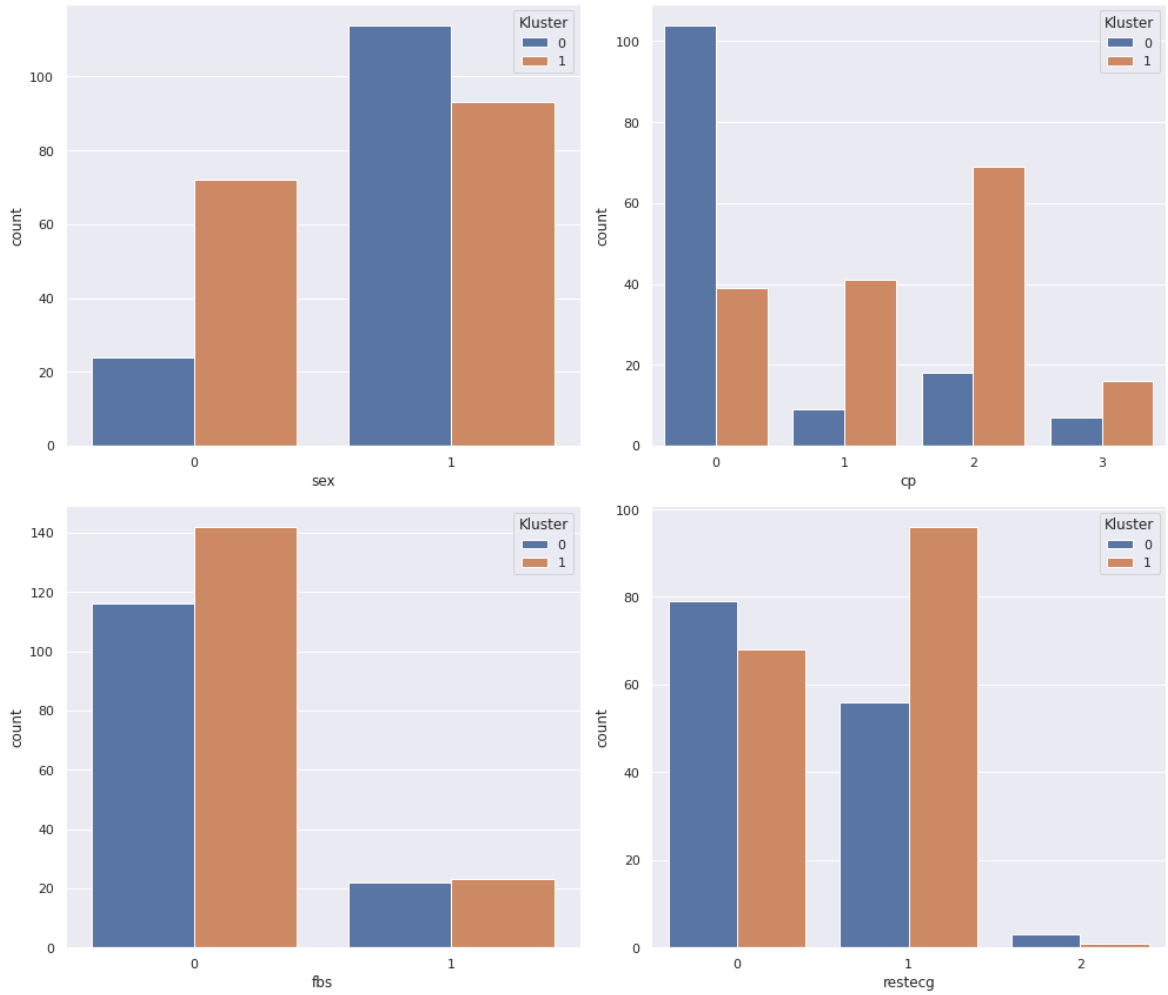
age	sex	cp	trestbps	chol	…	…	…	oldpeak	slope	ca	thal	target
63	1	3	145	233	…	…	…	2,3	0	0	1	1
37	1	2	130	250	…	…	…	3,5	0	0	2	1
41	0	1	130	204	…	…	…	1,4	2	0	2	1
56	1	1	120	235	…	…	…	0,8	2	0	2	1
…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…
56	0	1	140	294	…	…	…	0,5	1	0	2	1
44	1	1	120	263	…	…	…	1,6	2	0	3	1

Data *Preprocessing*

Pada tahap ini, dataset yang ada akan dilakukan *preprocessing* sebelum metode K-Means *Clustering* diterapkan dalam mengelola dataset tersebut. *Preprocessing* pada dataset ini meliputi penghapusan kolom yang tidak diperlukan dan pengisian nilai hilang (*missing values*) (Tabel 2). *Variable predictor* yang bersifat kategorik dan membuat plot secara bivariat, kemudian menerapkannya, hasil *Exploratory data analysis* (EDA) (Gambar2). Proses ini dilakukan untuk mencari insight dari dataset yang sudah di-*preprocessing* agar dataset bisa diterapkan untuk diolah memakai metode K-Means *Clustering*.

Tabel 2 Preposesing Gejala Penyakit Jantung

age	sex	cp	Trestbps	chol	…	…	…	oldpeak	slope	ca	thal	target
61	1	3	125	243	…	…	…	2,1	5	1	-	2
43	1	2	120	230	…	…	…	4,5	3	0	2	1
71	0	-	110	-	…	…	…	2,4	1	1	5	2
26	1	1	140	225	…	…	…	Null	2	Null	2	1



Gambar 2 Visualisasi data Exploratory data analysis

Evaluasi

Evaluasi bertujuan untuk mengetahui seberapa akurat model yang telah dibuat dapat memprediksi data baru dengan akurat. Evaluasi menggunakan metode *Accuracy*, *Precision*, *Recall*, dan F1 Score.

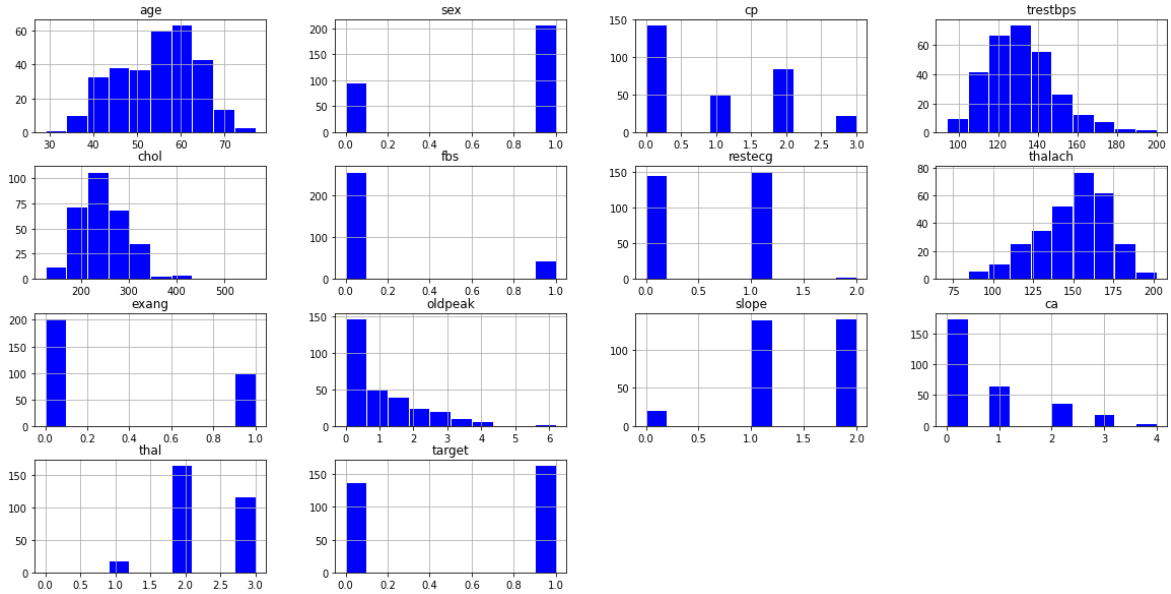
1. *Accuracy* merupakan rasio prediksi benar (Positif dan Negatif) dari keseluruhan data
   Accuracy = (TP + TN)/ (TP + FP +FN + TN)
2. *Precission* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif.
   Precission = (TP) / (TP + FP)
3. *Recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.
   Recall = (TP) / (TP + FN)
4. F1-Score merupakan perbandingan rata rata presisi dan *recall* yang dibobotkan
5. F1 Score = 2 * (*Recall***Precission*) /(*Recall* + *Precission*)

Untuk setiap nilai evaluasi yang semakin mendekati 1 maka semakin baik model yang dibuat dapat memprediksi data baru.

HASIL DAN PEMBAHASAN

Berdasarkan hasil visualisasi data, peneliti memproleh informasi bahwa untuk menentukan banyaknya yang terkena penyakit jantung disebabkan oleh sakit dada (cp), tekanan darah istirahat (trestbps), kolesterol (chol), gula darah puasa (fbs), elektrokardiografi

istirahat (restecg), detak jantung maksimum tercapai (thalach), angina (exang) yang diinduksi olahraga, depresi ST yang disebabkan oleh latihan relatif (oldpeak), kemiringan puncak latihan segmen ST (slope), jumlah pembuluh darah besar (0-3) diwarnai oleh flourosopy (ca), 3 = normal; 6 = cacat tetap; 7 = cacat reversibel (thal).



Gambar 3 Visualisasi Dataset Gejala Penyakit Jantung

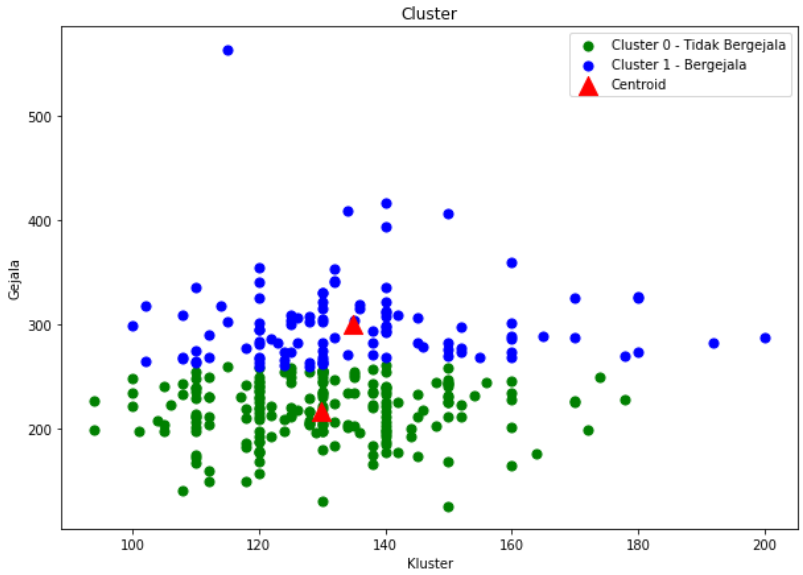
Berdasarkan hasil visualisasi data, dimendapat hasil bahwa jumlah dataset gejala penyakit jantung merupakan data yang tidak seimbang. Kelas pada dataset memiliki ratio imbalance 4:6, dimana kelas minoritas berjumlah 40% dari dataset dan 60% merupakan kelas mayoritas. Dataset yang tidak seimbang pada gejala penyakit memiliki resiko tinggi jika terjadi misscalasifiaction untuk penentuan prediksi diagnosis gejala penyakit. Kesalahan dalam klasifikasi akan berdampak pada penanganan dan pemberian tindakan kepada pasien yang dapat menimbulkan kelainan, keparahan dan juga bisa mengarah kematian. Diperlukan penanganan pada dataset yang tidak seimbang supaya model yang dibuat lebih optimal dan sesuai kelas yang ada. Penanganan *imbalance data* pada penelitian ini menerapkan K-means klustering dan SMOTE untuk mendapatkan data sintetis yang lebih representatif sehingga mengurangi kelasahan dalam melakukan klasifikasi data (**Algoritma 1 SMOTE**). Peneliti menerapkan membagi dataset menjadi dua bagian, data trainning dan data testing dengan rasio perbandingan 1:9. Ratio satu menyatakan data testing dan 9 data trainning.

Algoritma 1 SMOTE

```
from collections import Counter
from imblearn.over_sampling import SMOTE
# define oversampling strategy
SMOTE = SMOTE()
# fit and apply the transform
X, Y = SMOTE.fit_resample(X, Y)
# summarize class distribution
print("After oversampling: ",Counter(Y))
```

Clustering didapatkan menjadi dua atribut yaitu target dan gejala. Dari hasil visualisasi diketahui bahwa terdapat dua *cluster* yaitu *cluster* 0 yang artinya tidak terkena penyakit dengan angka centroid 130, ditunjukkan oleh warna biru. Selanjutnya terdapat *cluster* 1 yang artinya terkena penyakit jantung dengan angka centroid 135, ditunjukkan oleh warna merah

(Gambar 4). Hasil *clustering* dari beberapa data penyakit jantung yang ada di dataset (Tabel 3).



Gambar 4 Visualisasi Data Hasil *Clustering*

Tabel 3 Dataset dengan 2 kluster

No	**cp**	**trestbps**	**chol**	**…**	**…**	**…**	**…**	**…**	**…**	**ca**	**thal**	**target**	***cluster***
1	2	130	250	…	…	…	…	…	…	0	2	1	0
2	1	130	204	…	…	…	…	…	…	0	2	1	0
3	1	120	236	…	…	…	…	…	…	0	2	1	0
4	0	120	354	…	…	…	…	…	…	0	2	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
299	3	110	264	…	…	…	…	…	…	0	3	0	1
300	0	144	193	…	…	…	…	…	…	2	3	0	0
301	0	130	131	…	…	…	…	…	…	1	3	0	0
302	1	130	236	…	…	…	…	…	…	1	2	0	0

Evaluasi menggunakan Confusion Matrix dengan menerapkan kembali metode K-Means *Clustering* pada data training dan data testing (Tabel 4) serta terdapat hasil evaluasi untuk precision, recall dan F1-score pada Tabel 5. Hasil dari evaluasi menggunakan Confusion Matrix untuk menentukan *clustering* memikliki tingkat akurasi sebesar 93%.

Tabel 4 Confusion Matrix K-Means *Clustering*

	Y_Pred		
Y_Test	***cluster***	**0**	**1**
	0	39	0
	1	4	18

Tabel 5 Hasil Evaluasi oversampling

	precision	**recall**	**f1-score**	**support**
0	0.91	1	0.95	39
1	1	0.81	0.9	22
accuracy			0.93	61
macro avg	0.95	0.91	0.93	61
weighted vg	0.95	0.93	0.93	61

KESIMPULAN

Setelah penelitian ini dilakukan, maka peneliti mendapatkan hasil pengelompokkan data gejala penyakit jantung menggunakan K-Means dan oversampling adalah jumlah *cluster* yang optimal yaitu sebanyak dua *cluster*. *Cluster* 0 dengan nilai centroid 130 dan tidak memiliki gejala penyakit jantung 210 buah dan *cluster* 1 dengan nilai centroid 135 memiliki banyak gejala penyakit jantung 300 buah. Evaluasi dari *oversampling* memikliki tingkat akurasi sebesar 93%. Dengan pengelompokkan data tersebut, penderita yang mengalami gejala penyakit jantung dapat mendeteksi secara dini sehingga adanya pencegahan agar penyakit jantung tidak semakin memburuk.

REFERENSI

[1] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," Indonesian Journal of Data and Science, vol. 1 No 1, p. 6, 2020.
[2] W. N. Santosa and B. , "Penyakit Jantung Koroner dan Antioksidan," Jurnal Kesehatan dan Kedokteran, vol. 1, p. 98, 2020.
[3] F. Tahel, "PENERAPAN METODE *CLUSTERING* UNTUK MENGDIAGNOSA JENIS PENYAKIT GAGAL JANTUNG," Jurusan Sistem Informasi Fakultas Teknik Universitas Potensi Utama, 2018.
[4] A.Nurmasani and Y. Pristyanto, "ALGORITME STACKING UNTUK KLASIFIKASI PENYAKIT JANTUNG PADA DATASET IMBALANCED CLASS," Jurnal Pseudocode, vol. VIII No 1, pp. 21 - 22, 2021.
[5] A.Rifai, "ALGORITMA NEURAL NETWORK UNTUK PREDIKSI PENYAKIT JANTUNG," Jurnal Techno Nusa Mandiri, vol. X No 1, p. 1, 2013.
[6] M. C. Untoro, L. Anggraini, M. Andini, H. Retnosari and M. A. Nasrulloh, "Penerapan Metode K-Means *Clustering* Data COVID-19 di Provinsi," Teknologi: Jurnal Ilmiah Sistem Informasi, p. 60, 2021.
[7] C. Oktari, A. F. S, T. Safitri and S. Adi, "IMPLEMENTASI ALGORITMA K-MEANS UNTUK PENGELOMPOKKAN PENYAKIT YANG MENGIDAP JANTUNG KORONER," 2017.
[8] M. C. Untoro, L. Anggraini, M. Andini, H. Retnosari, and M. A. Nasrulloh, "Penerapan metode k-means *clustering* data COVID-19 di Provinsi Jakarta," Teknologi, vol. 11, no. 2, pp. 59–68, 2021, doi: 10.26594/teknologi.v11i2.2323.
[9] M. C. Untoro and J. L. Buliali, "Penanganan imbalance class data laboratorium kesehatan dengan Majority Weighted Minority Oversampling Technique," Jurnal Ilmiah Teknologi Sistem Informasi, p. 24, 2018.
[10] A.A. Z. Syahputra, A. D. Atika, M. A. Aslamsyah, M. C. Untoro and W. Yulita, "Pengelompokan Harga Ponsel Pintar berdasarkan Spesifikasi menggunakan Metode K-Means *Clustering*," Jurnal Teknik Informatika C.I.T, pp. 4 - 6, 2021.
[11] N. Lina and D. Saraswati, "Deteksi Dini Penyakit Jantung Koroner di Desa Kalimanggis dan Madiasari Kabupaten Tasikmalaya," Jurnal Warta LPM, 2020.