ISSN : 2407 - 3911

IMPLEMENTASI ALGORITMA *K-MEANS* DALAM PENGKLASTERAN MAHASISWA PELAMAR BEASISWA

Nurul Rohmawati W[1)], **Sofi Defiyanti**[2)], **Mohamad Jajuli**[3)]
[1),2),3)]*Teknik Informatika Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang*
Jl. H.S Ronggowaluyo Telukjambe Timur Karawang
nurul.rohmawati@student.unsika.ac.id, sofi.defiyanti@unsika.ac.id, mohamad.jajuli@staf.unsika.ac.id

Abstrak

Pengelompokan data pelamar beasiswa Bantuan Belajar Mahasiswa (BBM) dikelompokan menjadi 3 kelompok yaitu berhak menerima, dipertimbangkan dan tidak berhak menerima beasiswa. Pengelompokan menjadi 3 kelompok ini berguna untuk memudahkan dalam menentukan penerima beasiswa BBM. Algoritma k-means merupakan algoritma dari teknik clustering yang berbasis partisi. Teknik ini dapat mengelompokan data mahasiswa pelamar beasiswa.

Tujuan dari penelitian ini adalah untuk pengukuran kinerja algoritma, Pengukuran ini di lihat dari hasil cluster dengan menghitung nilai kemurnian (purity measure) dari masing – masing cluster yang di hasilkan. Data yang digunakan dalam penelitian ini adalah data mahasiswa yang mengajukan beasiswa kepada Fakultas Ilmu Komputer UNSIKA sebanyak 36 mahasiswa. Data akan diubah menjadi 3 dataset dengan format yang berbeda-beda, yakni data atribut kodifikasi sebagian, atribut kodifikasi keseluruhan dan atribut data asli. Nilai purity pada dataset data kodifikasi sebagian untuk hasil cluster algoritma k-means sebesar 61.11%. Pada dataset kodifikasi keseluruhan nilai purity hasil cluster algoritma k-means sebesar 80.56%. Dan untuk dataset data asli nilai purity hasil cluster algoritma k-means sebesar 75%. Maka dapat di simpulkan bahwa algoritma k-means lebih cocok digunakan pada dataset dengan format atribut yang dikodifikasi keseluruhan.

Kata kunci :

data mining, beasiswa, *clustering*, *k-means*, *purity measure*

Abstract

Data grouping scholarship applicants for Student Learning Assistance (BBM) grouped into 3 categories are entitled to receive, considered and not eligible to receive the scholarship. Grouping into 3 groups is useful to facilitate in determining scholarship recipients. K-means algorithm is an algorithm of clustering technique based partitions. This technique can categorize student data scholarship applicants.

The purpose of this research is to determine the algorithms for performance measurement, and measurement in view of the results of the cluster by calculating the value of purity (purity measure) of each - each cluster is generated. The data used in this research is data of students who apply for a scholarship to the School of Computer Science UNSIKA many as 36 students. The data will be converted into 3 datasets with different formats, ie attribute data codification in part, attributes and attribute the overall codification of the original data. Purity values in a dataset of data codification in part to the results of cluster k-means algorithm by 61.11%. At dataset codification overall value of purity results k-means cluster algorithm by 80.56%. And for the original data dataset purity value results k-means cluster algorithm by 75%. Then it can be concluded that the k-means algorithm is more suitable for use in datasets with formatting attributes that codified a whole

Keywords :

data mining, scholarship, clustering, k-means, purity measure

I. PENDAHULUAN

Salah satu sebab banyaknya mahasiswa mengajukan cuti akademik bahkan *dropout* yakni mengenai tingginya biaya perkuliahan yang mempengaruhi kelangsungan kegiatan belajarnya di sebuah instansi pendidikan tinggi. Beasiswa adalah bantuan yang di berikan kepada mahasiswa yang kurang mampu untuk memenuhi kewajibannya selama masa studinya. Pemberian beasiswa ini tentunya juga harus memperhatikan kriteria - kriteria tertentu sebelum di berikan kepada mahasiswa yang bersangkutan. Adapun kriteria ini tergantung pada ketentuan yang di tetapkan oleh pemberi beasiswa. Fungsi lain dari beasiswa ini juga sebagai penghargaan kepada mahasiswa berprestasi baik itu di dalam bidang akademik maupun non akademik. Pada penelitian ini beasiswa yang akan dibahas yaitu mengenai beasiswa BBM atau Bantuan Belajar Mahasiswa. Dimana beasiswa ini merupakan beasiswa yang disediakan untuk mahasiswa yang kurang mampu dan memliliki prestasi di bidang akademik maupun non akademik. Algoritma *k-means* dan *k-medoids dari t*eknik *clustering* dapat membantu dalam mengklasifikasi mahasiswa yang berhak menerima beasiswa, mahasiswa yang di pertimbangkan menerima dan mahasiswa yang tidak berhak menerima beasiswa.

Pada penelitian yang di lakukan oleh Noor Fitriana Hastuti, (2012) yaitu pemanfaatan algoritma *k-means clustering* untuk menentukan penerima beasiswa dengan membagi data mahasiswa menjadi 3 *cluster*. Berdasarkan hasil penelitiannya didapatkan nilai presisi dan ketetapan data yang tinggi dengan menggunakan metode *k-means clustering* (Hastuti, Saptono, & Suryani, 2012).

Berdasarkan penelitian tersebut akan dilakukan penelitian dengan menggunakan algoritma yang sama namun dengan format data yang berbeda - beda, sehingga dapat diketahui hasil *clustering* yang lebih baik (dari beberapa format atribut yang berbeda) untuk menentukan penerima beasiswa.

Adapun tujuan dari penelitian ini adalah membandingkan hasil *cluster* dari masing – masing format atribut dalam menentukan mahasiswa penerima beasiswa. Sehingga akan diketahui dari format atribut yang berbeda- beda, yang memiliki hasil *cluster* yang lebih baik.

II. PENELITIAN TERKAIT

II.1 Clustering

Menurut Baskoro *clustering* atau klusterisasi adalah salah satu alat bantu pada data mining yang bertujuan mengelompokkan objek-objek ke dalam *cluster* - *cluster*. *Cluster* adalah sekelompok atau sekumpulan objek - objek data yang similar satu sama lain dalam cluster yang sama dan disimilar terhadap objek-objek yang berbeda *cluster* (Nango, 2012).

II.2 Algoritma *K-Means*

Algoritma *k-means* adalah algoritma yang mempartisi data ke dalam *cluster* – *cluster* sehingga data yang memiliki kemiripan berada pada satu *cluster* yang sama dan data yang memiliki ketidaksamaan berada pada *cluster* yang lain. Sarwono mengemukakan secara lebih detail, algoritma *K-Means* adalah sebagai berikut:

1. Menentukan k sebagai jumlah kluster yang ingin di bentuk.
2. Membangkitkan nilai *random* untuk pusat *cluster* awal (*centroid*) sebanyak k.
3. Menghitung jarak setiap data *input* terhadap masing – masing *centroid* menggunakan rumus jarak *Euclidean (Euclidean Distance)* hingga ditemukan jarak yang paling dekat dari setiap data dengan *centroid.* Berikut adalah persamaan *Euclidian Distance*:

   $$d(xi, \mu j) = \sqrt{\sum (xi - \mu j)^2}$$

   Dimana :
   xi : data kriteria,
   μj : *centroid* pada *cluster* ke-*j*
4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan *centroid* (jarak terkecil).
5. Memperbaharui nilai *centroid.* Nilai *centroid* baru di peroleh dari rata-rata *cluster* yang bersangkutan dengan menggunakan rumus:

   $$\mu j(t+1) = \frac{1}{Nsj} \sum_{j \in sj} xj$$

   Dimana:
   $\mu j(t+1)$: *centroid* baru pada iterasi ke (t +1)
   N*sj* : banyak data pada *cluster Sj*.

6. Melakukan perulangan dari langkah 2 hingga 5, sampai anggota tiap *cluster* tidak ada yang berubah.

Jika langkah 6 telah terpenuhi, maka nilai pusat *cluster* (*μj*) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data (Sarwono).

II.3 Beasiswa

Beasiswa adalah bantuan yang di berikan kepada mahasiswa yang kurang mampu untuk memenuhi kewajibannya selama masa studinya. Bantuan ini biasanya berbentuk biaya atau ongkos yang harus di keluarkan oleh anak sekolah atau mahasiswa selama menempuh masa pendidikan di tempat belajar yang diinginkan.

Tujuan di berikannya beasiswa adalah:

1. Meningkatkan prestasi mahasiswa penerima baik kurikuler, ko-kurikuler, maupun ekstrakurikuler serta motivasi berprestasi bagi mahasiswa lain.
2. Mengurangi jumlah mahasiswa yang putus kuliah, karena tidak mampu membiayai pendidikan.
3. Meningkatkan akses dan pemerataan kesempatan belajar di perguruan tinggi.

Urutan prioritas penetapan mahasiswa penerima beasiswa BPP-PPA (Beasiswa BBM), sebagai berikut:

1. Mahasiswa yang memiliki keterbatasan kemampuan ekonomi.
2. Mahasiswa yang memiliki prestasi pada kegiatan ko/ekstra kurikuler (penalaran, minat dan bakat) tingkat internasional /dunia, Regional/Asia/Asean dan Nasional.
3. Mahasiswa yang mempunyai IPK paling tinggi.
4. Mahasiswa yang mempunyai SKS paling banyak dalam satu angkatan (DIKTI, 2013).

II.4 Purity Measure

Suwarsa (2013) mengemukakan bahwa *cluster* dikatakan murni (pure) semua objek dengan *class* yang sama berada pada *cluster* yang sama. Untuk mengukur tingkat akurasi *clustering* atau 'r', pengukuran nilai 'r' ini menggunakan persamaan berikut ini:

$$r = \frac{1}{n}\sum_{i=1}^{k} a_i$$

Dimana:

r : tingkat akurasi *clustering*

k : jumlah *cluster*

a_i : objek yang muncul didalam *cluster* C_i dan pada label *class* yang sesuai.

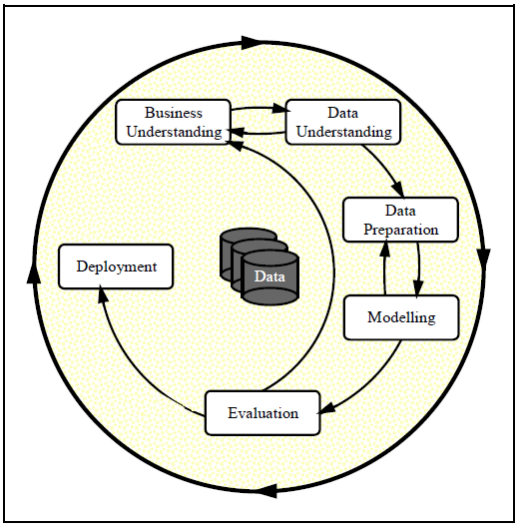
Semakin tinggi nilai r (semakin mendekati 1), semakin baik kulitas *cluster*. Sedangkan untuk menghitung *error cluster* atau 'e' seperti persamaan berikut ini:

$$e = 1 - r$$

Dimana r adalah nilai tingkat kemurnian *cluster* (Suwarsa, 2013).

III. METODA PENELITIAN

Metoda penelitian yang digunakan adalah Metodologi data mining CRISP-DM yang terdiri dari 6 tahap, karena penelitian ini bertujuan untuk membandingkan hasil *clustering*, maka tahapan CRISP-DM hanya sampai pada tahap ke 5. Adapun tahapannya sebagai berikut, pemahaman bisnis, pemahaman data, pengolahan data, pemodelan dan evaluasi.



Gambar 1. Model Crips-DM (Nisbet, Elder IV, & Liner, 2009)

IV. PEMBAHASAN

IV.1 Pemahaman Bisnis

Tujuan bisnis berdasarkan penjelasan mengenai fungsi dari beasiswa pada FASILKOM UNSIKA antara lain untuk membantu meringankan beban mahasiswa dalam menanggung biaya perkuliahan sehingga mengurangi jumlah mahasiswa yang putus kuliah karena tidak mampu membiayai pendidikan. Tujuan dari penelitian ini adalah membandingkan hasil *clustering* dari format data yang berbeda-beda, yang akan digunakan dalam mengklaster mahasiswa yang mengajukan beasiswa BBM. Kemudian dari hasil klasterisasi masing – masing format data tersebut akan di ketahui format data mana yang memiliki hasil *cluster* yang lebih baik sehingga dapat di ketahui mahasiswa yang tepat menerima beasiswa BBM berdasarkan *cluster* yang tepat.

IV.2 Pemahaman Data

Dari hasil pengumpulan data yang telah dilakukan diperoleh sebanyak 36 data mahasiswa yang mengajukan beasiswa. Kemudian dari data ini akan dipilih kriteria – kriteria yang dibutuhkan untuk masuk ketahap selanjutnya. Kriteria – kriteria ini yaitu, NPM, IPK, jumlah SKS yang sudah diambil, jumlah pendapatan orang tua dan jumlah tanggungan orang tua.

IV.3 Pengolahan Data

Dari data yang berhasil di kumpulkan, terdapat beberapa *missing value* pada kriteria penghasilan orang tua, kemudian *missing value* ini akan diisi dengan menggunakan teknik *mean imputation* atau diisi dengan nilai rata - rata dari kriteria penghasilan orang tua.

$$\bar{X} = \frac{Jumlah\ total\ atribut\ penghasilan\ orang\ tua}{Jumlah\ data}$$

$$\bar{X} = \frac{62208900}{36}$$

$$\bar{X} = 1728025$$

Jadi nilai rata – rata kriteria penghasilan orang tua adalah sebesar Rp. 1.728.025,-.

Pengkategorian kriteria penghasilan orang tua dibagi dengan jumlah tanggungan orang tua (dalam penelitian ini disingkat menjadi JP), dan pengkategorian kriteria SKS dengan cara mencari nilai *standard deviasi* dan *mean* dari masing – masing kriteria kemudian di kategorikan berdasakan tabel 1.

Tabel 1. Pengkategorian JP (Hastuti, Saptono, & Suryani, 2012)

Kategori	Kualifikasi	Kodifikasi
Kategori 4	JP ≤ $\bar{X}$ - S	4
Kategori 3	$\bar{X}$ - S < JP < $\bar{X}$	3
Kategori 2	$\bar{X}$ ≤ JP < $\bar{X}$ + S	2
Kategori 1	JP ≥ $\bar{X}$ + S	1

Setelah melakukan penghitung maka diketahui:

Mean JP ($\bar{X}$) : 1.025.394,4
Standar Deviasi JP (S) : 705.913,89
Kategori 4 : JP ≤ Rp. 319.480,5
Kategori 3 : Rp. 319.480,5 < JP < Rp. 1.025.394,4
Kategori 2 : Rp. 1.025.394,4 ≤ JP < Rp. 1.731.308,3
Kategori 1 : JP ≥ Rp. 1.731.308,3

Tabel 2. Pengkategorian SKS

Kategori	Kualifikasi	Kodifikasi
Kategori 5	SKS ≤ $\bar{X}$ - 2S	5
Kategori 4	$\bar{X}$ - 2S ≤ SKS < $\bar{X}$ - S	4
Kategori 3	$\bar{X}$ - S ≤ SKS < $\bar{X}$ + S	3
Kategori 2	$\bar{X}$ + S ≤ SKS < $\bar{X}$ + 2S	2
Kategori 1	SKS ≥ $\bar{X}$ + 2S	1

Setelah melakukan penghitung maka diketahui:

Mean SKS ($\bar{X}$) : 75,78
Standar Deviasi SKS (S) : 18,897
Kategori 5 : SKS ≤ 38
Kategori 4 : 38 < SKS < 56.89
Kategori 3 : 56.89 ≤ SKS < 94.67
Kategori 2 : 94.67 ≤ SKS < 113.56
Kategori 1 : SKS ≥ 113.56

Setelah dilakukan pengkategorian terhadap atribut SKS dan JP (penghasilan orang tua dibagi jumlah tanggungan orangtua), kemudian di buat *dataset* dengan nama *dataset* kodifikasi sebagian. Dan untuk membuat *dataset* kodifikasi keseluruhan untuk atribut IPK di kategorikan berdasarkan aturan pengambilan jumlah SKS berdasarkan IPK, dengan ketentuan seperti pada tabel 3.

Penelitian ini dilakukan untuk 3 jenis *dataset*, yakni *dataset* kodifikasi sebagian, *dataset* kodifikasi keseluruhan dan *dataset* data asli (atribut yang tidak di kategorikan).

Tabel. 3 Aturan pengambilan SKS berdasarkan IPK

Jumlah SKS	Rentang IPK	Kategori
24	3.00 - 4.00	1
21	2.50 - 2.99	2
18	2.01 - 2.49	3
15	1.90 - 2.00	4
12	< 1.49	5

IV.4 Pemodelan

Pemodelan *data mining* dalam penelitian ini dibuat dengan menggunakan perangkat lunak Rapidminer Studio 5. Pada aplikasi ini telah tersedia algoritma *clustering* berupa algoritma *k-means* dan *k-medoid*. Kedua algoritma sama - sama algoritma *clustering* berbasis partisi dijalankan dengan menggunakan aplikasi tersebut.

Algoritma *K-Means*

a. *Dataset* kodifikasi sebagian

Centroid akhir yang dihasilkan yakni sebagai berikut:

Tabel 4. *Centroid dataset* kodifikasi sebagian

	IPK	SKS	JP
Cluster 1	3.507	1.667	1.667
Cluster 2	3.471	3.1	1.5
Cluster 3	3.339	3.043	3.217

b. *Dataset* kodifikasi keseluruhan

Centroid akhir yang dihasilkan yakni sebagai berikut:

Tabel 5. *Centroid dataset* kodifikasi keseluruhan

	IPK	SKS	JP
Cluster 1	1	3.048	3.238
Cluster 2	1	2.769	1.538
Cluster 3	3	3	3

c. *Dataset* data asli

Centroid akhir yang dihasilkan yakni sebagai berikut:

Tabel 6. *Centroid dataset* data asli

	IPK	SKS	JP
Cluster 1	3.546	83.571	1,719,911.86
Cluster 2	3.330	73.5	3,000,000
Cluster 3	3.353	73.926	699,067.26

V. EVALUASI DAN PEMBAHASAN

V.1 Evaluasi

Dengan menggunakan persamaan pengujian *purity measure* (r) yang telah disebutkan pada poin 2 sebelumnya, untuk algoritma *k-means* dengan 3 format data yang berbeda-beda perbandingan nilai *purity* (r) pada *dataset* dengan atribut data dikodifikasi sebagian, data kodifikasi keseluruhan dan data asli disajikan pada tabel 7.

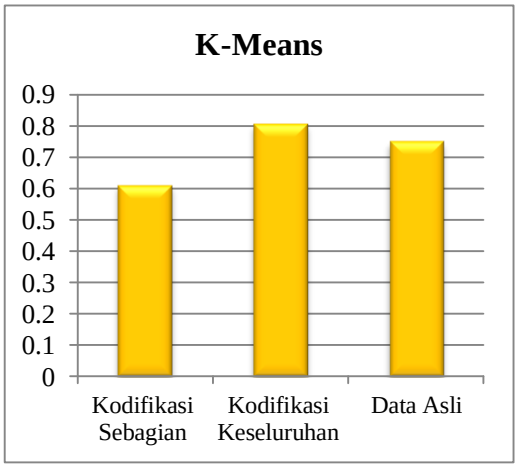
Tabel 7. Purity Measure algoritma k-means

Purity Measure (*r*)	
Dataset	*K-Means*
Kodifikasi sebagian	0.611
Kodifikasi Keseluruhan	0.806
Data Asli	0.750

Tabel 8 menunjukan perbandingan nilai *purity* dalam bentuk persentase.

Tabel 8. Nilai *Purity* Measure dalam persentase

Purity Measure (*r*)	
Dataset	*K-Means*
Kodifikasi sebagian	61.11%
Kodifikasi Keseluruhan	80.56%
Data Asli	75.00%



Gambar 1. Grafik perbandingan nilai *Purity Measure*

V.2 Pembahasan

Berdasarkan penghitung perbandingan nilai *purity measure* pada hasil *clustering* dari algoritma *k-*

means dengan *dataset* atribut kodifikasi sebagian diketahui sebesar 0.611 atau 61.11%. Dan untuk *dataset* data kodifikasi keseluruhan, hasil *cluster* algoritma *k-means* memiliki nilai *purity* (r) sebesar 0.806 atau 80.56%. Untuk *dataset* data asli yang pada penelitian ini mengandung *outlier*, diketahui nilai *purity* (r) untuk hasil *cluster* algoritma *k-means* sebesar 0.750 atau 75%.

Maka dapat disimpulkan bahwa untuk algoritma *k-means*, *dataset* dengan atribut data yang dikodifikasi keseluruhan memiliki hasil *cluster* yang lebih baik. Hal ini dikarenakan algoritma *k-means* menggunakan *mean* (rata - rata) sebagai pusat *cluster*nya *(centroid)*, juga *Euclidean* sebagai fungsi jarak untuk menghitung jarak kedekatan antar objek dengan *centroid*nya. Selain itu *dataset* yang dikodifikasi keseluruhan, persebaran nilai pada format *dataset* ini menjadi lebih sederhana, artinya tidak ada data yang bernilai ekstrim (*outlier*). Oleh karena itu algoritma *k-means* sangat cocok untuk mengelompokan data yang tidak mengandung *outlier*.

VI. KESIMPULAN DAN SARAN

VI.1 Kesimpulan

Membandingkan hasil *cluster* algoritma *k-means* berdasarkan hasil *clustering* dari masing - masing format *dataset* yang berbeda-beda (kodifikasi sebagian, kodifikasi keseluruhan dan data asli) dengan mengukur tingkat akurasi *clustering* yakni menghitung nilai *purity measure* dari hasil *cluster*nya. Semakin besar nilai *purity* (semakin mendekati 1) semakin baik kualitas *cluster* yang dihasilkan oleh suatu algoritma.

Berdasarkan penghitung perbandingan nilai *purity measure* pada hasil *clustering* dari algoritma *k-means* dengan format atribut *dataset* yang berbeda - beda (atribut data yang di kodifikasi sebagian, atribut data yang dikodifikasi seluruhnya dan atribut data asli). Diketahui nilai *purity* pada *dataset* data kodifikasi sebagian untuk hasil *cluster* algoritma *k-means* sebesar 0.611 atau 61.11%. Pada *dataset* kodifikasi keseluruhan nilai *purity* hasil *cluster* algoritma *k-means* sebesar 0.806 atau sebesar 80.56%. Untuk *dataset* data asli nilai *purity* hasil *cluster* algoritma *k-means* sebesar 0.750 atau 75%. Maka dapat disimpulkan bahwa tingkat akurasi *clustering* hasil *cluster* algoritma *k-means* berdasarkan nilai *purity measure*, *dataset* yang dikodifikasi keseluruhan lebih baik dari pada *dataset* yang di kodifikasi sebagian dan *dataset* data asli.

VI.2 Saran

Kriteria pembanding untuk algoritma *data mining* khususnya untuk metode *clustering* yang dilakukan pada penelitian ini hanya pada pengukuran hasil *clustering*. Untuk penelitian selanjutnya kriteria lain dapat ditambahkan misalnya kriteria *silhouette coefficient* dan *error ratio*.

REFERENSI

Arifin, S.Pd., M.Si., MOS, M. (2007). Rancang Bangun Sistem Informasi Koperasi (Studi Kasus pada Koperasi Pegawai Republik Indonesia"Teknik Sejahtera"). *Seminar Nasional Teknologi 2007 (SNT 2007)*, D-1.

DIKTI. (2013, Februari). Retrieved Januari 15, 2014, from Direktorat Jendral Pendidikan Tinggi: http://www.dikti.go.id

Hastuti, N. F., Saptono, R., & Suryani, E. (2012). Pemanfaatan Metode K-Means Clustering Dalam Penentuan Penerima Beasiswa. *Jurnal Informatika*.

Loekito, F. R., & Ayub, M. (2011). Aplikasi Pengelolaan Data Sistem Pelayanan Kesehatan ada Departement Kesehatan PT. Ateja Multi Industri. *Jurnal Informatika*, 145-146.

Nango, D. N. (2012). Penerapan Algoritma K-Means Untuk Clustering Data Anggaran Pendapatan Belanja Daerah di Kabupaten XYZ. *Skripsi*, 11-12.

Nisbet, R., Elder IV, J., & Liner, G. (2009). *Handbook of Statictical Analysis and Data Mining Applications.* Elsevier Inc.

Sarwono, Y. T. (n.d.). Aplikasi Model Jaringan Syaraf Tiruan Dengan Radial Basis Function Untuk Mendeteksi Kelainan Otak (Stroke Infark). *Jurnal Sistem Informasi*, 3-4.

Suwarsa, R. D. (2013). *Imlementasi K-Modes Pada*

Dlustering Data Kategori Menggunakan New Dissimilarity Measure. Malang: Universitas Brawijaya Malang.

Velmurugan, T. (2012). Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points. *Int.J. Computer Technology & Applications* , 1759.