

***CLUSTERING DATA MICRORNA KANKER PAYUDARA
MENGUNAKAN METODE K-MEANS***



SKRIPSI

**Disusun Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Komputer
pada Departemen Ilmu Komputer/Informatika**

Disusun oleh:

DICKY FABRO SAHARA

24010313120057

**DEPARTEMEN ILMU KOMPUTER/INFORMATIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO**

2019

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini:

Nama : Dicky Fabro Sahara

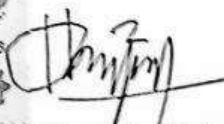
NIM : 24010313120057

Judul : *Clustering Data MicroRNA Kanker Payudara menggunakan Metode K-Means*

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka.

Semarang, 24 September 2019




Dicky Fabro Sahara
24010313120057

HALAMAN PENGESAHAN

Judul : *Clustering Data MicroRNA* Kanker Payudara menggunakan Metode *K-Means*

Nama : Dicky Fabro Sahara

NIM : 24010313120057

Telah diujikan pada sidang skripsi dan dinyatakan lulus pada tanggal 24 September 2019

Semarang, 27 September 2019

Mengetahui,

Ketua Departemen Ilmu Komputer/Informatika



Panitia Penguji Skripsi,

Ketua

Drs. Suhartono, M.Kom.

NIP. 195504071983031003

HALAMAN PENGESAHAN

Judul : *Clustering Data MicroRNA* Kanker Payudara menggunakan Metode *K-Means*

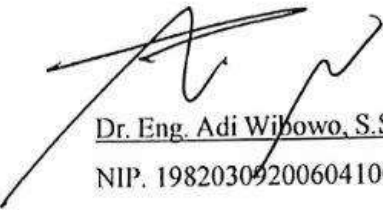
Nama : Dicky Fabro Sahara

NIM : 24010313120057

Telah diujikan pada sidang skripsi tanggal 24 September 2019

Semarang, 27 September 2019

Pembimbing,



Dr. Eng. Adi Wibowo, S.Si., M.Kom.

NIP. 198203092006041002

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI UNTUK KEPENTINGAN AKADEMIS

Sebagai *civitas* akademik Universitas Diponegoro, saya yang bertanda tangan di bawah ini:

Nama : Dicky Fabro Sahara

NIM : 24010313120057

Program Studi : Informatika

Departemen : Ilmu Komputer/Informatika

Fakultas : Sains dan Matematika

Jenis Karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan **Hak Bebas Royalti Non-eksklusif (*Non-exclusive Royalty Free Right*)** kepada Universitas Diponegoro atas karya ilmiah saya yang berjudul:

Clustering Data MicroRNA Kanker Payudara menggunakan Metode K-Means

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-eksklusif ini, Universitas Diponegoro berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan Skripsi saya tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenarnya.

Semarang, 24 September 2019

Yang Menyatakan



Dicky Fabro Sahara

24010313120057

ABSTRAK

Kanker payudara merupakan penyakit yang disebabkan oleh pertumbuhan sel yang abnormal dimana sel tersebut menyerang sel normal di jaringan sekitarnya. Peranan *MicroRNA* pada kanker telah menjadi perkembangan penting dalam studi kanker sejak penemuannya pada tahun 2002. *MicroRNA* merupakan indikator pengukuran dari tingkat keberadaan suatu kanker serta mempunyai peranan penting dalam deteksi dan diagnosis khususnya kanker payudara. Pada penelitian ini, dilakukan *clustering* data *MicroRNA* kanker payudara menggunakan metode *K-Means* dan mengetahui jumlah *cluster* terbaik serta pola pengelompokan yang dipengaruhi fitur pada data *MicroRNA*. Dari hasil penelitian ini, perbandingan pengujian nilai $k = 2$, nilai $k = 3$, nilai $k = 4$, dan nilai $k = 5$ menunjukkan bahwa nilai $k = 2$ merupakan *cluster* terbaik dengan nilai *silhouette* 0.454. Pola pengelompokan data menggunakan metode *K-Means* dipengaruhi oleh fitur ekspresi: *hsa-mir-10b*, *hsa-mir-21*, *hsa-mir-125b-1*, *hsa-mir-125b-2*, *hsa-mir-145*, *hsa-mir-200a*, dan *hsa-mir-200b*.

Kata Kunci: Kanker Payudara, *MicroRNA*, *K-Means Clustering*

ABSTRACT

Breast cancer is a disease caused by abnormal cell growth where the cells attack normal cells in the surrounding tissue. The role of MicroRNA in cancer has been an important development in cancer studies since its discovery in 2002. MicroRNA is an indicator of measurement of the level of the presence of a cancer and has an important role in the detection and diagnosis, especially breast cancer. In this study, clustering of breast cancer MicroRNA data was performed using the K-Means method and knowing the best number of clusters and grouping patterns affected by features in the MicroRNA data. From the results of this study, the comparison of testing the value of $k = 2$, the value of $k = 3$, the value of $k = 4$, and the value of $k = 5$ indicate that the value of $k = 2$ is the best cluster with a silhouette value of 0.454. The pattern of data grouping using the K-Means method is affected by the expression features: hsa-mir-10b, hsa-mir-21, hsa-mir-125b-1, hsa-mir-125b-2, hsa-mir-145, hsa-mir-200a, and hsa-mir-200b.

Keywords: Breast Cancer, MicroRNA, K-Means Clustering

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi dengan berjudul “*Clustering Data MicroRNA* Kanker Payudara menggunakan Metode *K-Means*” dengan baik dan lancar.

Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana strata satu pada Departemen Ilmu Komputer/Informatika Fakultas Sains dan Matematika Universitas Diponegoro. Dalam penyusunan skripsi ini penulis banyak mendapat bimbingan, bantuan, dan dukungan dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis menyampaikan terimakasih kepada:

1. Dr. Retno Kusumaningrum, S.Si., M.Kom., selaku Ketua Departemen Ilmu Komputer/Informatika;
2. Panji Wisnu Wirawan, S.T., M.T., selaku Koordinator Skripsi;
3. Dr. Eng. Adi Wibowo, S.Si., M.Kom., selaku Dosen Pembimbing Skripsi;
4. Semua pihak yang telah membantu kelancaran dalam penyusunan Skripsi, yang tidak dapat disebutkan satu persatu.

Pada penyusunan laporan ini penulis menyadari bahwa masih terdapat banyak kekurangan baik dari segi materi maupun segi penyampaian materi tersebut. Hal tersebut dikarenakan keterbatasan kemampuan dan pengetahuan dari penulis. Oleh karena itu, kritik dan saran yang bersifat membangun sangat penulis harapkan. Semoga laporan skripsi ini dapat bermanfaat bagi semua pihak.

Semarang, 24 September 2019

Dicky Fabro Sahara

DAFTAR ISI

	Hal
HALAMAN JUDUL	i
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PENGESAHAN	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI UNTUK KEPENTINGAN AKADEMIS	v
ABSTRAK.....	vi
ABSTRACT	vii
KATA PENGANTAR.....	viii
DAFTAR ISI	ix
DAFTAR GAMBAR.....	xii
DAFTAR TABEL	xiii
DAFTAR LAMPIRAN	xiv
BAB I PENDAHULUAN	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	2
1.3. Tujuan dan Manfaat.....	3
1.4. Ruang Lingkup	3
1.5. Sistematika Penulisan.....	3
BAB II LANDASAN TEORI.....	5
2.1. Tinjauan Pustaka	5
2.2. Kanker Payudara	6
2.2. <i>MicroRNA</i>	7
2.3. <i>K-Means Clustering</i>	9
2.4. Normalisasi Data	11
2.5. <i>Silhouette</i>	11

2.7.	Proses Pengembangan Perangkat Lunak.....	12
2.8.	Pengujian <i>Black Box</i>	16
2.10.	Bahasa Pemrograman <i>Python</i>	17
2.11.	<i>HyperText Markup Language</i>	18
2.12.	<i>Framework Web Flask</i>	19
2.13.	<i>Orange Tools</i>	20
BAB III METODOLOGI PENELITIAN		21
3.1.	Garis Besar Penyelesaian Masalah.....	21
3.1.1.	Pengumpulan Data	22
3.1.2.	<i>Data Normalization</i>	23
3.1.3.	Perhitungan <i>K-Means Clustering</i>	24
3.1.4.	Evaluation.....	32
3.2.	Deskripsi Umum Aplikasi	33
3.3.	Analisis.....	33
3.2.1.	Kebutuhan Fungsional.....	33
3.2.2.	Kebutuhan Non Fungsional.....	33
3.2.3.	Pemodelan Data.....	34
3.2.4.	Pemodelan Fungsional	34
3.3.	Desain.....	38
3.3.1.	Desain Fungsi	38
3.3.2.	Desain Antarmuka.....	40
BAB IV IMPLEMENTASI DAN PENGUJIAN		44
4.1.	Implementasi	44
4.1.1.	Lingkungan Implementasi.....	44
4.1.2.	Implementasi Data.....	44
4.1.3.	Implementasi Fungsi	45
4.1.4.	Implementasi Antarmuka	45
4.2.	Pengujian Sistem	48
4.2.1.	Rencana Pengujian	48
4.2.2.	Pelaksanaan Pengujian	49
4.2.3.	Evaluasi Pengujian	49
4.3.	Pengujian <i>K-Means Clustering</i>	49

4.3.1. Skenario Pengujian.....	49
4.3.2. Pembahasan Skenario Pengujian.....	50
4.3.3. Evaluasi Hasil Skenario Pengujian	57
BAB V PENUTUP	58
5.1. Kesimpulan.....	58
5.2. Saran.....	58
DAFTAR PUSTAKA.....	59
LAMPIRAN-LAMPIRAN	62
Lampiran 1. Data <i>MicroRNA</i>	63
Lampiran 2. Tabel Pengujian <i>Blackbox</i>	71
Lampiran 3. Implementasi Fungsi	72
Lampiran 4. Kartu Bimbingan Skripsi	76
Lampiran 5. Iterasi <i>Centroid</i>	78

DAFTAR GAMBAR

	Hal
Gambar 2.1. Model Waterfall.....	13
Gambar 3.1. Garis Besar Penyelesaian Masalah	21
Gambar 3.2. Nilai Silhouette dari Tools Orange	32
Gambar 3.3. Diagram Dekomposisi	35
Gambar 3.4. Data Context Diagram	35
Gambar 3.5. Data Flow Diagram Level 1	36
Gambar 3.6. DFD Level 2 Mengelola Normalisasi.....	37
Gambar 3.7. DFD Level 2 Mengelola Clustering	37
Gambar 3.8. Desain Fungsi Normalisasi	39
Gambar 3.9. Desain Fungsi Perhitungan K-Means	40
Gambar 3.10. Desain Antarmuka Halaman Beranda	41
Gambar 3.11. Desain Antarmuka Halaman Data	42
Gambar 3.12. Desain Antarmuka Halaman Data Normalisasi.....	43
Gambar 3.13. Desain Antarmuka Halaman Tabel Hasil Perhitungan	43
Gambar 4.1. Implementasi Antarmuka Beranda	46
Gambar 4.2. Implementasi Antarmuka Data	46
Gambar 4.3. Implementasi Antarmuka Normalisasi	47
Gambar 4.4. Implementasi Antarmuka Perhitungan	48
Gambar 4.5. Grafik Hasil Clustering dengan Nilai $k = 2$	52
Gambar 4.6. Grafik Hasil Clustering dengan Nilai $k = 3$	53
Gambar 4.7. Grafik Hasil Clustering dengan Nilai $k = 4$	55
Gambar 4.8. Grafik Hasil Clustering dengan Nilai $k = 5$	56
Gambar 4.9. Nilai Silhouette dari Tools Orange	57

DAFTAR TABEL

	Hal
Tabel 2.1. Tinjauan Pustaka	5
Tabel 2.2. SRS	13
Tabel 2.3. Notasi Pemodelan Fungsional	15
Tabel 3.1. Dataset MicroRNA	22
Tabel 3.2. Hasil Minimal dan Maksimal untuk Setiap Atribut	23
Tabel 3.3. Hasil Normalisasi	24
Tabel 3.4. Hasil <i>Euclidean Distance</i> Iterasi 1	27
Tabel 3.5. Hasil Alokasi Nilai Minimum ke Centroid terdekat Iterasi 1.....	28
Tabel 3.6. Hasil <i>Euclidean Distance</i> Iterasi 2	29
Tabel 3.7. Hasil <i>Euclidean Distance</i> Iterasi 3	29
Tabel 3.8. Hasil <i>Euclidean Distance</i> Iterasi 4	30
Tabel 3.9. Hasil Alokasi Nilai Minimum ke Centroid terdekat Iterasi 2.....	30
Tabel 3.10. Hasil Alokasi Nilai Minimum ke <i>Centroid</i> terdekat Iterasi 3	31
Tabel 3.11. Hasil Alokasi Nilai Minimum ke <i>Centroid</i> terdekat Iterasi 4	31
Tabel 3.12. Hasil Clustering	32
Tabel 3.13. Kebutuhan Fungsional.....	33
Tabel 3.14. Kebutuhan Non Fungsional	34
Tabel 4.1. Rencana Pengujian	48
Tabel 4.2. Pengelompokkan $k = 2$	51
Tabel 4.3. Pengelompokkan $k = 3$	52
Tabel 4.4. Pengelompokkan $k = 4$	54
Tabel 4.5. Pengelompokkan $k = 5$	55

DAFTAR LAMPIRAN

	Hal
Lampiran 1. Data <i>MicroRNA</i>	63
Lampiran 2. Tabel Pengujian <i>Blackbox</i>	71
Lampiran 3. Implementasi Fungsi	72
Lampiran 4. Kartu Bimbingan Skripsi	76
Lampiran 5. Iterasi <i>Centroid</i>	78

BAB I

PENDAHULUAN

Bab ini membahas latar belakang masalah, rumusan masalah, tujuan dan manfaat, serta ruang lingkup penelitian.

1.1. Latar Belakang

Kanker adalah pertumbuhan sel yang abnormal di mana sel tersebut menyerang sel normal di jaringan sekitarnya. Penyebab kanker paling umum adalah perubahan (mutasi) pada gen dalam sel. Kanker payudara adalah kanker yang menyerang jaringan payudara. Beberapa faktor risiko terjadinya kanker payudara, antara lain usia di atas 50 tahun, jenis kelamin perempuan, riwayat keluarga, keturunan, obesitas *post-menopause* dan terapi hormonal. Paparan terhadap estrogen terkait erat dengan etimologi kanker payudara. Estrogen merupakan hormon seks wanita yang dibutuhkan dalam sejumlah proses dalam tubuh. Mekanisme estrogen terhadap proses karsinogenik masih terlalu kompleks, namun terdapat beberapa bukti penelitian bahwa estrogen menyebabkan *proliferasi* sel payudara normal maupun ganas (Vogel, 2018). Mutasi pada gen *breast cancer 1* (BRCA1) dan *breast cancer 2* (BRCA2) bertanggung jawab terhadap 3-8% kasus kanker payudara. Kedua gen ini dipercaya berperan sebagai *tumor suppressor genes* yang berperan dalam mempertahankan integritas DNA dan regulasi transkripsi DNA (Suparman&Suparman, 2014). Ekspresi gen pada sebuah sel dikendalikan secara akurat dan didasarkan pada kondisi fisiologis pada tubuh dimana mereka berada. *MicroRNA* berukuran kecil memainkan peran penting dalam sebagian besar pengaturan sel (Sankar, et al., 2017).

MicroRNA (*MiRNA*) terdiri atas *RNA* endogen *non-coding* berukuran kecil, yaitu panjangnya 22-24 nukleotida. *MiRNA* bertindak sebagai *tuner* yang baik dari regulator gen pada eukariotik uniseluler, hewan, tumbuhan dan manusia. Secara umum, regulasi gen di dalam sel seluruhnya didasarkan pada jenis interaksi *MiRNA*-*mRNA* dan interaksinya dari satu sel ke sel lain bervariasi berdasarkan kondisi fisiologis (Sankar, et al., 2017). *MicroRNA* (*miRNA*) memiliki potensi besar sebagai *biomarker* tumor dan target terapi. Seiring perkembangan pesat dari teknologi eksperimental, penelitian ekspresi gen menjadi semakin terspesialisasi dan beragam. Struktur data yang

kompleks setelah membawa tantangan besar untuk identifikasi *biomarker* (Yang, et al., 2017). Dalam dunia kedokteran, *biomarker* adalah indikator yang dapat diukur dari tingkat keparahan atau keberadaan beberapa keadaan penyakit. Perubahan ekspresi *MiRNAs* dalam berbagai jenis *neoplasia* membentuk pola spesifik *MiRNA signature* untuk jenis kanker tertentu (Lan, et al., 2014). Perubahan ekspresi *MiRNA* pada kanker, menimbulkan dugaan bahwa *MiRNA* dapat berperan sebagai onkogen atau tumor supresor gen (Juwita, 2017). *MicroRNA* berperan sebagai modulator ekspresi gen sehingga dapat dijadikan sebagai kandidat diagnostik dan indikator *prognostik* serta target terapi yang potensial (Heneghan, et al., 2009). Oleh karena itu, *MiRNA* menjadi salah satu alternatif *biomarker* kanker yang sangat menjanjikan dimasa mendatang.

Clustering data microRNA adalah proses pengelompokkan data menjadi beberapa *cluster*. Pengelompokkan data tersebut berguna untuk mengidentifikasi pola data pada setiap *cluster*. Pembahasan mengenai *clustering* data MicroRNA pernah dilakukan di beberapa penelitian. Contoh penelitian tersebut yaitu penelitian yang dilakukan oleh (Yang, et al., 2017) yang menunjukkan bahwa hasil dari metode berbasis pengelompokan dapat mengidentifikasi *biomarker* kombinasi *MicroRNA* memiliki akurasi dan efisiensi tinggi.

Dalam penelitian digunakan algoritma *K-Means* untuk mengelompokkan data *MicroRNA* menjadi *k cluster* berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokkan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam *cluster* yang lain. Metode *K-Means* merupakan algoritma yang paling banyak digunakan karena secara konseptual metode k-means sederhana, dapat dimodifikasi, dan mudah diimplementasikan (Sopian, et al., 2016). Algoritma *K-Means* mencoba menemukan variabilitas yang berpengaruh pada jumlah *cluster*. Algoritma *K-Means* mengelompokkan *n* objek berdasarkan atribut menjadi *k* partisi, dimana $k < n$. Nilai *k* akan menentukan *cluster* pada data *MicroRNA* sedangkan *n* adalah sejumlah data *MicroRNA*.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka dapat dirumuskan permasalahan yaitu: Bagaimana membuat aplikasi *clustering* data *MicroRNA* dengan

menggunakan metode *K-Means* untuk mengetahui pengelompokan data dan analisis pola fitur yang berpengaruh?

1.3. Tujuan dan Manfaat

Tujuan dari penelitian ini adalah sebagai berikut:

1. Menghasilkan aplikasi untuk peng-*cluster*-an data penyakit kanker payudara menggunakan metode *K-Means Clustering*.
2. Menganalisis hasil kerja dari metode *K-Means* untuk *clustering* data penyakit kanker payudara berdasarkan jumlah *cluster*.
3. Menganalisis pola fitur yang berpengaruh pada data *MicroRNA* kanker payudara.

Manfaat dari aplikasi *Clustering Data MicroRNA Penyakit Kanker Payudara* berguna untuk mengelompokkan data *MicroRNA* sehingga dapat membantu menemukan pola pengelompokkan dalam *cluster*.

1.4. Ruang Lingkup

Ruang lingkup dalam menerapkan *clustering* data *MicroRNA* menggunakan metode *K-Means Clustering* adalah sebagai berikut:

1. Data berupa *dataset MicroRNA* yang mempunyai satuan integer.
2. Aplikasi dapat menampilkan pengelompokan data *MicroRNA* dengan metode *K-Means*.
3. Nilai k pada *K-Means* adalah 2, 3, 4, dan 5.
4. Bahasa pemrograman yang digunakan dalam mengembangkan aplikasi adalah *python*.

1.5. Sistematika Penulisan

Sistematika penulisan yang digunakan dalam penyusunan laporan skripsi ini terdiri dari 5 bab, yaitu Pendahuluan, Landasan Teori, Metodologi Penelitian, Implementasi dan Pengujian, serta Penutup.

BAB I PENDAHULUAN

Bab pendahuluan berisi latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup masalah, serta sistematika penulisan dalam pelaksanaan penelitian *Clustering Data MicroRNA* Kanker Payudara menggunakan Metode *K-Means*.

BAB II LANDASAN TEORI

Bab ini berisi tinjauan pustaka yang berhubungan dengan skripsi sebagai landasan untuk merumuskan dan menganalisa permasalahan pada skripsi ini. Tinjauan pustaka yang digunakan meliputi kanker payudara, *MicroRNA*, metode *K-Means Clustering*, metode pengembangan, permodelan analisis, metode pengujian, bahasa pemrograman, dan *tools* yang digunakan.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai tahapan yang dilakukan dalam penelitian tugas akhir. Tahapan tersebut meliputi garis besar penyelesaian masalah, pengumpulan data, normalisasi data, Perhitungan *K-Means*, evaluasi sistem, deskripsi umum aplikasi dan analisa kebutuhan sistem.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Bab ini membahas mengenai tahapan implementasi dan analisis dari skenario pengujian aplikasi maupun metode. Tahapan tersebut meliputi implementasi dari aplikasi, pengujian aplikasi, dan pengujian metode *K-Means Clustering*.

BAB V PENUTUP

Bab penutup berisi kesimpulan dari penelitian dan uraian yang telah dibahas pada bab sebelumnya dan saran untuk pengembangan sistem lebih lanjut.