

**APLIKASI KLASIFIKASI *K-NEAREST NEIGHBOR* (K-NN)
PADA PASIEN KANKER PARU-PARU**

(Studi Kasus : Klasifikasi Stadium Pasien Kanker Paru-paru Berdasarkan
Gene Expression Data Amerika Serikat, GSE10072)

TUGAS AKHIR

**Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana
Jurusan Statistika**



Disusun oleh :

Ines Riantika

13611192

**JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2018**

**HALAMAN PERSETUJUAN PEMBIMBING
TUGAS AKHIR**

Judul : Aplikasi Klasifikasi *K-Nearest Neighbor* (K-NN)
pada Kanker Paru-paru
(Studi Kasus : Klasifikasi Stadium Pasien Kanker
Paru-paru Berdasarkan *Gene Expression* Data
di Amerika, GSE10072)

Nama Mahasiswa : Ines Riantika

Nomor Mahasiswa : 13611192



TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN
Yogyakarta, 03 Januari 2018

Pembimbing

(Dr.techn. Rohmatul Fajriyah, S.Si, M.Si)

HALAMAN PENGESAHAN
TUGAS AKHIR

APLIKASI KLASIFIKASI *K-NEAREST NEIGHBOR (K-NN)*
PADA PASIEN KANKER PARU-PARU
(Studi Kasus : Klasifikasi Stadium Pasien Kanker Paru-paru Berdasarkan
Gene Expression Data di Amerika Serikat, GSE10072)

Nama Mahasiswa : Ines Riantika

Nomor Mahasiswa : 13611192

TUGAS AKHIR INI TELAH DIUJIKAN
PADA TANGGAL 13 FEBRUARI 2018

Nama Penguji


1. Fitria Dyah Ayu Suryanegara, M.Sc., Apt
2. Ayundyah Kesumawati, M.Si
3. Dr.techn.Rohmatul Fajriyah, S.Si, M.Si

Tanda Tangan

.....
.....
.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam


(Drs. Allywar., M.Sc., P.hD)

KATA PENGANTAR



Assalamu'alaikum, wr, wb.

Alhamdulillah rabbil'alamin, Puji Syukur saya panjatkan kehadiran Allah SWT atas berkat rahmat dan hidayah-Nya sehingga saya dapat menyelesaikan penulisan skripsi ini yang berjudul “**Aplikasi Klasifikasi *K-Nearest Neighbor* (K-NN) Pada Pasien Kanker Paru-paru**”. Untuk memperoleh gelar sarjana di Jurusan Statistika dapat terselesaikan tanpa hambatan yang berarti. Shalawat serta salam semoga selalu tercurah kepada Nabi Muhammad SAW.

Pada kesempatan ini penulis mengucapkan terimakasih kepada dosen pembimbing penulis Dr.tench.RohmatulFajriyah.S.Si.,M.Si. yang selalu memberikan dukungan, dan saran dalam menyelesaikan penelitian ini. Trimakasih penulis berikan terhadap pihak-pihak lainnya :

1. Bapak Drs. Allwar, M.Sc.,Ph.D, selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia
2. Bapak Dr. RB Fajriya Hakim, M.Si., selaku ketua prodi statistika, beserta jajarannya.
3. Kedua Orang tua tercinta, Ahmad Kamildan ibu Maisarah serta Kakak-kakak dan saya, Mutia Elvarina, Azlan Romil, Fikri Aldi dan Afifah Azmi yang selalu memberikan doa, dukungan, kasih sayang yang menjadi semangat terbesar untuk menyelesaikan skripsi ini.
4. Teman seperjuangan Rizka Wulandari, Ditha Runita, Mega Ayu Lestari, dan Suci Eka Purnamasari yang memberikan semangat dan doronganya ketika mnyusun skripsi ini.
5. Seluruh teman-teman penulis Mayda Rosa Devina, Novita Kharimatunnisa, Novia Nigsih, Siti Amelia Mubarokkah, Nevy Tarigan,Yayi Diyah dan Intan Anggraini yang memberikan saran dan doronganya selama penulisan.

Semoga Allah SWT membalas segala kebaikan mereka dengan segala anugrah, rahmat, dan Hidayah-Nya. Penulis menyadari sepenuhnya bahwa Tugas Akhir ini masih jauh dari sempurna, oleh karena itu segala kritik dan saran yang sifatnya membangun selalu penulis harapkan. Semoga Tugas Akhir ini dapat bermanfaat bagi penulis khususnya dan bagi semua yang membutuhkan umumnya. Akhir kata, semoga Allah SWT selalu melimpahkan rahmat serta hidayah-Nya kepada kita semua, Amin amin ya robbal alamiin.

Wassalamu'alaikum Wr.Wb.

Yogyakarta, 16 Februari 2017

Penulis,

(Ines Riantika)

DAFTAR ISI

HALAMAN PERSETUJUAN PEMBIMBING	Error! Bookmark not defined.
HALAMAN PENGESAHAN	Error! Bookmark not defined.
KATA PENGANTAR	ii
DAFTAR ISI	v
DAFTAR TABEL	vii
DAFTAR GAMBAR	viii
DAFTAR LAMPIRAN	ix
PERNYATAAN	Error! Bookmark not defined.
INTISARI	xi
ABSTRACT	xii
BAB IPENDAHULUAN	Error! Bookmark not defined.
1.1 Latar Belakang.....	Error! Bookmark not defined.
1.2 Rumusan Masalah	Error! Bookmark not defined.
1.3 Tujuan Penulisan	Error! Bookmark not defined.
1.4 Manfaat Penelitian.....	Error! Bookmark not defined.
1.5 Batasan Masalah.....	Error! Bookmark not defined.
BAB II TINJAUAN PUSTAKA	Error! Bookmark not defined.
BAB III LANDASAN TEORI	Error! Bookmark not defined.
3.1 Bioinformatika.....	Error! Bookmark not defined.
3.2 <i>Microarray</i>	Error! Bookmark not defined.
3.3 <i>Affymetrix</i>	Error! Bookmark not defined.
3.4 <i>Gene Expression</i>	Error! Bookmark not defined.
3.5 <i>Preprocessing</i>	Error! Bookmark not defined.
3.6 <i>Filtering</i>	Error! Bookmark not defined.
3.7 <i>Feature Selection</i>	Error! Bookmark not defined.
3.8 Kanker	Error! Bookmark not defined.
3.9 <i>Kanker Paru -paru</i>	Error! Bookmark not defined.

3.10	<i>K- Nearest Neighbor</i>	Error! Bookmark not defined.
3.11	<i>Cross Validation</i>	Error! Bookmark not defined.
3.12	<i>Confusion Matrix</i>	Error! Bookmark not defined.
3.	<i>Receiver Operating Characteristic (ROC)</i>	Error! Bookmark not defined.
BAB IV METODOLOGI PENELITIAN		Error! Bookmark not defined.
4.1	Data.....	Error! Bookmark not defined.
4.2	Variabel Penelitian	Error! Bookmark not defined.
4.3	Metode Analisis Data	Error! Bookmark not defined.
4.4	Alat dan Cara Organisir Data	Error! Bookmark not defined.
BAB V PEMBAHASAN		Error! Bookmark not defined.
5.1	Deskripsi Data	Error! Bookmark not defined.
5.2	<i>Proses Data Microarray</i>	Error! Bookmark not defined.
5.3	Pemilihan nilai <i>k</i>	Error! Bookmark not defined.
5.4	Hasil Klasifikasi	Error! Bookmark not defined.
BAB VI PENUTUP		Error! Bookmark not defined.
6.1	Kesimpulan.....	Error! Bookmark not defined.
6.2	Saran	Error! Bookmark not defined.
DAFTAR PUSTAKA		Error! Bookmark not defined.

DAFTAR TABEL

Nomor	Judul	Halaman
3.1	<i>Gene ExpressionDataset</i>	14
3.2	<i>Type dan subtype kanker paru-paru</i>	18
3.3	<i>Table cofusion matrix</i>	21
3.4	<i>Skala nilai Kappa</i>	22
3.5	Kualitas klasifikasi berdasarkan AUC	23
5.1	<i>Gene expression data</i>	28
5.2	<i>Tabel nilai akurasi dan kappa</i>	31
5.3	<i>Hasil prediksi test data</i>	32
5.4	<i>Confusion matrix</i>	33
5.5	<i>Annotation varaible importance</i>	35

DAFTAR GAMBAR

Nomor	Judul	Halaman
3.1	<i>Kontribusi Ilmu Terapan Lainnya Pada Bidang Bioinformatik</i>	7
3.2	<i>Jenis Microarrays</i>	10
3.3	<i>Affymetrix Microarray</i>	11
3.4	<i>Proses Affymetrix</i>	12
3.5	<i>Proses preprocessing</i>	14
3.6	<i>Kanker di Amerika</i>	16
3.7	<i>Kanker paru-paru pada pria di Amerika</i>	17
5.1	<i>Bar plot jenis kelamin pasien kanker paru-paru</i>	26
5.2	<i>Bar plot jaringan pasien kanker paru-paru</i>	26
5.3	<i>Bar plot status merokok pasien kanker paru-paru</i>	27
5.4	<i>Bar plot stadium pasien kanker paru-paru</i>	27
5.5	<i>Box plot raw data GSE10072</i>	29
5.6	<i>Box plot preprocessing data GSE10072</i>	29
5.7	<i>Gen yang terpilih</i>	30
5.8	<i>Pemilihan nilai k terbaik</i>	31
5.9	<i>Kurva ROC</i>	34
5.10	<i>Variable importance</i>	35

DAFTAR LAMPIRAN

- Lampiran 1.** *Script* pemanggilan dan *preprocessing*
- Lampiran 2.** Pembuatan *expression set*
- Lampiran 3.** *Filtering* dan *feature selection*
- Lampiran 4.** Penggabungan data fenotip dan gene expression dan *split* data
- Lampiran 5.** Deskripsi data, analisis K-NN dan hasil akurasi
- Lampiran 6.** Gen yang paling berpengaruh dan anotasi.
- Lampiran 7.** Hasil *remove filtering*, struktur data, dan hasil nilai *k* terbaik
- Lampiran 8.** Hasil klasifikasi K-NN dan ROC

PERNYATAAN

Dengan ini penulis menyatakan bahwa Tugas Akhir ini tidak terdapat karya orang lain yang sebelumnya digunakan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi, dan tidak terdapat karya yang pernah diterbitkan oleh orang lain, kecuali semua yang dijadikan referensi yang telah disebutkan dalam daftar pustaka.

Yogyakarta, 16 Februari 2016



[Handwritten Signature]
Penulis

**APLIKASI KLASIFIKASI *K-NEAREST NEIGHBOR(K-NN)* PADA *GENE
EXPRESSION MICROARRAY***

(Studi Kasus :Klasifikasi Stadium Pasien Kanker Paru-paru Berdasarkan *Gene
Expression* Data Di Amerika, GSE10072)

Oleh : Ines Riantika

Program Studi Statistika Fakultas MIPA

Universitas Islam Indonesia

INTISARI

Kanker adalah suatu penyakit yang disebabkan oleh tidak terkontrolnya pertumbuhan dan penyebaran sel yang tidak normal. Kanker adalah penyebab kematian terbesar didunia,bahkan pada tahun2017 di estimasikan penderita baru kanker di Amerika Serikat mencapai 1.688.780 jiwa. Salah satu jenis kanker yang paling banyak diderita oleh warga Amerika adalah jenis kanker paru-paru dengan tingkat kematian pria pada tahun 2017 mencapai 155.870 jiwa. Stadium pada kanker biasanya dilakukan untuk melakukan bagaimana pengobatan yang disarankan untuk pasien. Maka dari itu pada penelitian tersebut dilakukan klasifikasi menggunakan data *gene expression* untuk mengklasifikasikan stadium kanker paru-paru *non small cell lung cancer(NSCLC)* untuk membedakan jaringan tumor stadium awal dan jaringan tumor stadium akhir dengan menggunakan $k = 9$ didapatkandidapat hasil akurasi dari *confusion matrix* sebesar 0.848, dan ROC dengan nilai AUC 0.9306 yang juga diartikan bahwa klasifikasi baik untuk data ini.

Kata kunci : Klasifikasi , *Gene expression*, Kanker, *K-nearest neighbor*

**APPLICATION CLASSIFICATION OF K-NEAREST NEIGHBOR
(K-NN) ON GENE EXPRESSION MICROARRAY**

(Case Study: Classification of Stage Patients of Lung Cancer In America Using
Gene Expression Data GSE10072)

Studi Kasus : Klasifikasi Stadium Pasien Kanker Paru-paru Berdasarkan *Gene
Expression* Data Di Amerika, GSE10072

By : Ines Riantika

Supervised by : Dr. Tench. Rohmatul Fajriyah, S.Si., M.Si.

ABSTRACT

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cell. Cencer is the biggest cause of death in the world. Estimated numbers of new cancer cases for 2017 at United States reached 1,688,780 people. One type cancer with hight death rate in 2017 is lung cncer with 155.870 people. Staging in cancer is usually for given the recommended treatment for the patient. Therefore, in this study, based of gene expression data for classify the stages of lung cancer of non-small cell lung cancer (NSCLC) for early-stage tumor tissue and late-stage tumor tissue using $k = 9$. it is found confusion matrix with 0.848, and ROC with AUC value 0.9306 which is also interpreted good K-NN classification.

Keyword : Classifiaction , Gene expression, Cancer, K-nearest neighbor

