

**IMPLEMENTASI METODE *K-NEAREST NEIGHBOR* (KNN)
PADA DATA BIOINFORMATIKA**

(Studi Kasus: Klasifikasi *Gene Expression* pada Data *Maize Growth Stages*)

TUGAS AKHIR

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



Nurlina Tri Wulandari

14 611 161

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

2018

**HALAMAN PERSETUJUAN PEMBIMBING
TUGAS AKHIR**

Judul : Implementasi Metode *K-Nearest Neighbor* (KNN) pada
Data Bioinformatika (Studi Kasus: Klasifikasi *Gene
Expression* pada Data *Maize Growth Stages*)

Nama Mahasiswa : Nurlina Tri Wulandari

Nomor Mahasiswa : 14611161

**TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK
DIUJIKAN**

Yogyakarta, 31 Juli 2018



Menyetujui,
Dosen Pembimbing I

الجامعة الإسلامية
الإندونيسية

Menyetujui,
Dosen Pembimbing II

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

Husna Nugrahapraja, Ph.D.

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI METODE *K-NEAREST NEIGHBORS* (KNN)
PADA DATA BIOINFORMATIKA

Nama Mahasiswa : Nurlina Tri Wulandari

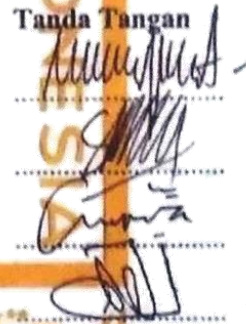
Nomor Mahasiswa : 14 611 161

TUGAS AKHIR INI TELAH DIUJIKAN
PADA TANGGAL 31 JULI 2018

Nama Penguji

1. Muryanto, S.P., M.Si
2. Muh. Hasan Sidiq Kurniawan, M.Sc
3. Husna Nugrahaprija, Ph.D
4. Dr. techn. Rohmatul Fajriyah, S.Si., M.Si

Tanda Tangan



Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam



(Prof. Riyanto, S.Pd., M.Si., Ph.D)

KATA PENGANTAR



Assalamu'alaikum Warahmatullaahi Wabarakaatuh

Alhamdulillah *alhamdulillah* *alhamdulillah*, segala puji bagi Allah SWT atas segala rahmat dan karunia-Nya yang telah dilimpahkan kepada penulis sehingga penulis dapat menyelesaikan penulisan tugas akhir ini yang berjudul **“Implementasi Metode K-Nearest Neighbors (KNN) pada Data Bioinformatika (Studi Kasus: Klasifikasi Gene Expression pada Data Maize Growth Stages)”** sebagai salah satu persyaratan yang harus dipenuhi dalam menyelesaikan jenjang strata satu di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia. Shalawat serta salam semoga selalu tercurah kepada Nabi Muhammad SAW serta para sahabat dan pengikutnya yang senantiasa menjaga keimanan dan keislamannya hingga akhir hayatnya.

Penyusunan tugas akhir ini tidak terlepas dari dukungan, bantuan, arahan, dan bimbingan dari berbagai pihak. Untuk itu dengan hati yang tulus penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam.
2. Bapak Dr. RB. Fajriya Hakim, M.Si. selaku Ketua Jurusan Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam yang juga telah banyak membantu dan memberi ilmu dan wawasan baru kepada penulis.
3. Ibu Dr. techn. Rohmatul Fajriyah, M.Si selaku dosen pembimbing I tugas akhir penulis, yang dengan sabar membimbing penulis untuk menyelesaikan tugas akhir ini. Semoga Allah memberikan kesehatan kepada Ibu dan ilmu yang diberikan dapat bermanfaat bagi penulis.
4. Bapak Husna Nugrahapraja, Ph.D selaku dosen pembimbing II tugas akhir penulis, terimakasih atas bimbingan dan kesabarannya sehingga penulis dapat menyelesaikan tugas akhir ini.

5. Kedua orang tua tercinta, Ibu dan Bapak serta kakak-kakak kandung yang selalu setia mendoakan, memberikan dukungan, kasih sayang, menemani serta memberikan motivasi kepada penulis selama menyelesaikan tugas akhir ini.
6. Seluruh staf pengajar Jurusan Statistika Universitas Islam Indonesia terimakasih atas ilmu yang diberikan.
7. Teman-teman bimbingan TA : Ika, Umami, Hawila, Husain, Himelda, Nanda, Arie, Maida, Leni, Shoddiq, Gilang, Aat dan Mas Rakhmat yang sudah sama-sama berjuang, saling mengingatkan dan memberi motivasi serta dorongan untuk menyelesaikan tugas akhir ini.
8. Teman-teman Kost Putri Muslimah Teratai 2 yang selalu memberi semangat untuk penulis selama menyelesaikan tugas akhir ini.
9. Teman-teman seperjuangan Statistika UII Angkatan 2014 sukses dan berjuanglah menuju hari esok yang lebih baik.
10. Semua pihak yang terlibat dalam penyusunan tugas akhir ini yang tidak bisa disebutkan satu per satu yang telah memberikan dukungan dan motivasi selama menyelesaikan tugas akhir ini.

Penulis menyadari bahwa tugas akhir ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran dari semua pihak yang bersifat membangun selalu penulis harapkan demi kesempurnaan tugas akhir ini. Semoga tugas akhir ini dapat bermanfaat. Akhir kata, semoga Allah SWT senantiasa meridhai segala usaha kita. *Aamiin yaa rabbal'alamiin.*

Wassalamu'alaikum Warahmatullaahi Wabarakaatuh

Yogyakarta, Juli 2018

Nurlina Tri Wulandari

DAFTAR ISI

HALAMAN JUDUL	ii
HALAMAN PERSETUJUAN PEMBIMBING	iii
HALAMAN PENGESAHAN	iv
KATA PENGANTAR	v
DAFTAR ISI.....	vii
DAFTAR GAMBAR.....	ix
DAFTAR TABEL.....	x
DAFTAR LAMPIRAN	xi
PERNYATAAN.....	xii
INTISARI.....	xiii
ABSTRACT	xiv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah	4
1.3 Batasan Masalah	4
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian.....	5
BAB II TINJAUAN PUSTAKA	6
2.1 Penelitian tentang metode <i>K-Nearest Neighbor</i> (KNN).....	6
2.2 Penelitian metode <i>K-Nearest Neighbor</i> (KNN) pada data Bioinformatika .	7
2.3 Penelitian tentang Jagung (<i>Maize</i>).....	10
BAB III LANDASAN TEORI.....	12
3.1 Bioinformatika.....	12
3.2 <i>Next Generation Sequencing</i> (NGS)	14
3.3 Ekspresi Gen	15
3.4 <i>Filtering</i>	16
3.5 Uji T Test	16
3.6 Jagung (<i>Zee mays</i>)	17

3.6.1 Taksonomi Jagung	17
3.6.2 Morfologi Tanaman Jagung	18
3.7 Analisis Deskriptif	19
3.8 <i>Machine Learning</i>	20
3.9 Klasifikasi	21
3.10 <i>K-Nearest Neighbors</i> (KNN).....	22
3.11 <i>Confusion Matrix</i>	23
3.12 <i>Kappa Statistics</i>	24
3.13 <i>Receiver Operating Characteristic</i> (ROC).....	25
BAB IV METODOLOGI PENELITIAN	27
4.1 Jenis dan Sumber Data	27
4.2 Tempat dan Waktu Penelitian	27
4.3 Variabel Penelitian	27
4.4 Metode Analisis Data	27
4.5 Tahapan Penelitian	28
BAB V HASIL DAN PEMBAHASAN.....	30
5.1 Analisis Deskriptif.....	30
5.2 <i>Filtering</i>	31
5.3 Pemilihan Nilai.....	33
5.4 Hasil Klasifikasi	34
5.5 <i>Variable Importance</i>	37
BAB VI KESIMPULAN DAN SARAN	39
6.1 Kesimpulan.....	39
6.2 Saran.....	40
DAFTAR PUSTAKA	41
LAMPIRAN	46

DAFTAR GAMBAR

Gambar 3.1 Struktur <i>gene expression</i>	16
Gambar 3.2 Tanaman jagung	18
Gambar 3.3 Kurva <i>Receiver Operating Characteristic</i> (ROC).....	25
Gambar 4.1 <i>Flowchart</i> penelitian	28
Gambar 4.2 <i>Flowchart</i> analisis data.....	29
Gambar 5.1 <i>Stripchart</i> Zm00001d027230 <i>expression value</i>	30
Gambar 5.2 Pemilihan nilai terbaik.....	33
Gambar 5.3 Kurva ROC	36
Gambar 5.4 Variabel yang paling berpengaruh	37

DAFTAR TABEL

Tabel 3.1 <i>Confusion matrix</i>	23
Tabel 3.2 Skala nilai Kappa.....	24
Tabel 3.3 Kualitas klasifikasi berdasarkan AUC.....	26
Tabel 5.1 Dataset <i>maize growth stages</i>	30
Tabel 5.2 <i>Gene expression data maize growth stages</i>	30
Tabel 5.3 Data <i>gene expression</i> yang akan diklasifikasikan.....	32
Tabel 5.4 Tabel nilai akurasi dan nilai Kappa	33
Tabel 5.5 Hasil prediksi data <i>testing</i>	34
Tabel 5.6 Tabel <i>confusion matrix</i>	34
Tabel 5.7 Keterangan variabel yang paling berpengaruh.....	37

DAFTAR LAMPIRAN

LAMPIRAN 1. <i>Script Input Data</i>	46
LAMPIRAN 2. <i>Script Data yang Digunakan</i>	46
LAMPIRAN 3. <i>Script untuk memanggil data, letak kromosom dan start</i>	46
LAMPIRAN 4. <i>Script Membuat Klasifikasi pada Data</i>	46
LAMPIRAN 5. <i>Script Filtering Data dan Feature Selection</i>	47
LAMPIRAN 6. <i>Script Deskripsi Data, Analisis KNN dan Hasil Akurasi</i>	48
LAMPIRAN 7. <i>Script Kurva ROC</i>	50
LAMPIRAN 8. <i>Script Variabel Importance</i>	50

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya 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 diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, Juli 2018



Nurlina Tri Wulandari

IMPLEMENTASI METODE *K-NEAREST NEIGHBOR* (KNN) PADA DATA BIOINFORMATIKA

(Studi Kasus: Klasifikasi *Gene Expression* pada Data *Maize Growth Stages*)

Nurlina Tri Wulandari

Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Islam Indonesia

INTISARI

*Perkembangan teknologi semakin hari semakin pesat dan mempengaruhi perkembangan pada bidang lainnya. Salah satu bidang yang berkembang karena pengaruh perkembangan teknologi yaitu bidang pertanian. Jagung adalah salah satu komoditi tanaman terpenting di pertanian, baik di dunia maupun di Indonesia. Hal ini dikarenakan jagung memiliki kandungan karbohidrat yang tinggi. Oleh karena itu, benih jagung yang digunakan harus bermutu tinggi, baik mutu genetik, fisik maupun fisiologinya dan berasal dari varietas unggul. Penelitian ini bertujuan untuk melakukan dan mengevaluasi klasifikasi *gen expression* pada tanaman jagung, yaitu pengklasifikasikan tahap pertumbuhan pada organ jagung. Metode yang digunakan adalah *K-Nearest Neighbor* (KNN). Berdasarkan penelitian diperoleh nilai *k* optimal yaitu 7 dengan nilai akurasi dari *confusion matrix* sebesar 85.7 dan ROC dengan nilai AUC sebesar 0.788. Hal ini berarti metode KNN dapat digunakan untuk mengklasifikasi tahap pertumbuhan pada organ jagung.*

Kata Kunci : Klasifikasi, *Gene Expression*, Jagung, *K-Nearest Neighbor* (KNN), AUC, *Variable Importance*

IMPLEMENTATION OF K-NEAREST NEIGHBOR (KNN) METHOD ON BIOINFORMATIC DATA

(Studi Case: Classification of *Gene Expression* on Data *Maize Growth Stages*)

Nurlina Tri Wulandari

*Departement of Statistics, Faculty of Mathematics and Natural Science
Islamic University of Indonesia*

ABSTRACT

Tecnological development increase rapidly day by day and affect to other fields. One field that develops due to the affect of technological developments is agricultural field. Maize is important agricultural commodities, both in the world and in Indonesia because maize has a high carbohydrate content. Therefore, the maize that used as the seed must be high quality, both genetic, physic and physiological qualities and derived from superior varieties. This study aims to conduct and evaluate the classification of gene expression in maize on classifying stages of growth in the maize organs. The method used is K-Nearest Neighbor (KNN). Based on research/study obtained the optimal is 7 with the accuracy value of confusion matrix 85.7 and ROC with AUC value 0.788. This means that KNN method can be used to classify the growth stages in the maize organs.

Keyword: *Classification, Gene Expression, Maize, K-Nearest Neighbor (KNN), AUC.*