

ALGORITMA K-NEAREST NEIGHBOR DENGAN MENGGUNAKAN METODE EUCLIDEAN DISTANCE UNTUK MEMPREDIKSI KELULUSAN UJIAN NASIONAL BERBASIS DESKTOP PADA SMA NEGERI 12 TANGERANG

Pandu Yuli Santoso¹⁾, Dewi Kusumaningsih²⁾

Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur
Jl. Raya Ciledug, Petukangan Utara, Kebayoran Lama, Jakarta Selatan 12260
Telp. (021) 5853753, Fax. (021) 5866369
E-mail : ¹⁾panduschief@gmail.com ²⁾ dewi.kusumaningsih@budiluhur.ac.id

Abstrak

SMA Negeri 12 Tangerang merupakan salah satu dari banyaknya SMA Negeri di Indonesia yang akan melaksanakan Ujian Nasional yang diselenggarakan oleh pemerintah. Untuk melakukan persiapan Ujian Nasional, maka dari itu pihak sekolah mengadakan ujian Tryout. Ujian Tryout ini berfungsi untuk mempersiapkan para siswa/i SMAN Negeri 12 Tangerang dalam menghadapi Ujian Nasional. Dalam ujian Tryout ini terdiri dari mata pelajaran Bahasa Indonesia, Matematika, Bahasa Inggris, dan mata pelajaran siswa/i sesuai jurusan masing-masing. Maka dari itu ujian Tryout biasa dijadikan sebagai tolak ukur para siswa/i untuk menghadapi Ujian Nasional dengan baik. Dari hasil ujian Tryout tersebut para guru dapat melihat kesiapan para siswa/i-nya dalam menghadapi Ujian Nasional yang akan berlangsung. Siswa/i yang nilai ujian Tryout-nya tidak melampaui standar yang ditentukan, maka siswa/i dapat berkonsultasi. Maka dari itu pihak sekolah tidak boleh terlambat dalam membuat solusi yang tepat bagi siswa/i-nya. Namun dalam hasil Tryout tersebut, guru masih kesulitan dalam memprediksi kelulusan Ujian Nasional setiap tahunnya. Hal ini dikarenakan banyaknya siswa/i sehingga memerlukan waktu yang cukup lama, kurangnya tingkat efektifitas dalam memprediksi, dan masih menggunakan cara yang manual. Dengan adanya masalah tersebut, maka Algoritma K-Nearest Neighbor dengan metode pengukuran jarak Euclidean Distance dirasa cukup membantu dalam mengatasi masalah efisiensi waktu dan efektifitas prediksi kelulusan Ujian Nasional. Atribut yang digunakan dalam prediksi dan klasifikasi ini ialah Nasional yaitu nilai Tryout. Dengan menggunakan nilai Tryout siswa/i 5 tahun terakhir berjumlah 1140 data. Setelah data tersebut diproses menghasilkan rata-rata akurasi sebesar 88.58%. Nilai K terbaik adalah 7 rata-rata akurasi 89.126%, presisi 90.62%, dan recall sebesar 97.61%. Maka dari itu dapat disimpulkan dari pengujian yang telah dilakukan dengan menggunakan dataset siswa/i SMA Negeri 12 Tangerang sistem dapat memprediksi dan mengklasifikasikan dengan baik dan cepat.

Kata kunci : Algoritma K-Nearest Neighbor, Euclidean Distance, Prediksi

1. PENDAHULUAN

Pada era ini perkembangan teknologi yang pesat telah banyak memberikan banyak manfaat dan kemajuan di berbagai bidang. Teknologi sudah menjadi kebutuhan yang sangat diperlukan bagi umat manusia dalam menunjang segala aktifitas agar lebih efisien. Dalam era teknologi yang berkembang ini pun, segala sesuatu yang belum pernah terjadi dapat diprediksi menggunakan teknologi yang ada walaupun tingkat akurasi tidak mencapai 100%.

SMAN 12 Tangerang merupakan salah satu instansi pendidikan yang akan melaksanakan Tryout sebelum Ujian Nasional. Tryout tersebut merupakan salah satu cara guru untuk membantu para siswa/i dalam mempersiapkan Ujian Nasional yang akan diselenggarakan oleh pemerintah. Nantinya bagi siswa/i yang tidak melampaui nilai BSNP (Badan Standar Nasional Pendidikan) dapat berkonsultasi dengan guru SMAN 12 Tangerang.

Permasalahan yang ada pada SMAN 12 Tangerang dalam memprediksi kelulusan siswa/i masih dilakukan dengan cara yang manual, yaitu masih menggunakan *Microsoft Excel*. Dengan

menggunakan cara manual ini data yang akan dikelola tidak maksimal dan mengurangi efisiensi serta efektifitas dari proses prediksi yang dilakukan.

Maka dari itu tujuan penulisan ini, penulis berniat untuk membangun suatu sistem yang dapat digunakan untuk memprediksi kelulusan Ujian Nasional siswa/i SMAN 12 Tangerang dengan mengimplementasikan ilmu *data mining* menggunakan pendekatan klasifikasi. Perlunya data mining karena adanya sejumlah besar data yang dapat digunakan untuk menghasilkan informasi dan *knowledge* yang berguna. Informasi dan *knowledge* yang didapat tersebut dapat digunakan untuk mengetahui suatu pola dalam suatu data yang banyak, terlebih lagi besarnya kebutuhan untuk mengubah data tersebut menjadi informasi yang berguna [1].

Algoritma yang akan digunakan untuk pengklasifikasian data tersebut yaitu, algoritma *K-Nearest Neighbor*. Pemilihan algoritma klasifikasi ini dikarenakan algoritma *K-Nearest Neighbor* sudah teruji dan mudah diimplementasikan [2], dan juga salah satu metode yang memiliki kinerja yang intensif ketika diberikan *training dataset* yang benar [3].

Algoritma K-Nearest Neighbor (K-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut [4]. Metode pengukuran jarak yang digunakan pada penulisan tugas akhir ini menggunakan metode pengukuran jarak terdekat *Euclidean Distance*. Diharapkan sistem yang dibangun menggunakan algoritma *K-Nearest Neighbor* dengan metode pengukuran jarak terdekat *Euclidean Distance* ini dapat meningkatkan efisiensi dan efektifitas dalam memprediksi kelulusan Ujian Nasional siswa/i SMAN 12 Tangerang.

2. LANDASAN TEORI

2.1 Metode Klasifikasi dan Prediksi

Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (fase *training*), dimana algoritma klasifikasi dibuat untuk menganalisa data *training* lalu direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi [3]. Proses klasifikasi didasarkan pada empat komponen [2]:

a. Kelas

Variabel dependen yang berupa kategorikal yang merepresentasikan 'label' yang terdapat pada objek. Contohnya: resiko penyakit jantung, resiko kredit, *customer loyalty*, jenis gempa.

b. Predictor

Variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Contohnya: merokok, minum alkohol, tekanan darah, tabungan, aset, gaji.

c. Training dataset

Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor*.

d. Testing dataset

Berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi.

Prediksi menurut [5] bertujuan untuk dapat melakukan pemetaan dari setiap himpunan variabel ke setiap target, yang kemudian menggunakan model tersebut untuk menentukan nilai target pada himpunan baru berdasarkan data-data yang sudah ada. Terdapat dua model prediksi yang dikatakan oleh [5], yaitu klasifikasi dan regresi. Dimana klasifikasi digunakan untuk variabel target yang bersifat diskrit, sedangkan regresi untuk variabel target yang bersifat kontinu.

2.2 Dataset

Data dalam terminologi statistika adalah kumpulan objek dengan atribut-atribut tertentu, dimana objek tersebut adalah individu berupa data dimana setiap data memiliki sejumlah atribut [6].

2.3 Algoritma K-Nearest Neighbor

K-Nearest Neighbor (KNN) termasuk kelompok *instance-based learning*. Algoritma ini juga merupakan salah satu teknik *lazy learning*. KNN dilakukan dengan mencari kelompok k objek dalam data *training* yang paling dekat (mirip) dengan objek pada data baru atau data *testing* [7]. Adapun pendekatan yang dilakukan oleh metode *K-Nearest Neighbor* ini adalah dengan mengklasifikasikan objek atau *query instance* baru kedalam suatu kelas berdasarkan kedekatan atribut atau fitur dengan "K" buah tetangganya pada data pelatihan [2]. Data pelatihan sebelumnya akan membentuk vektor yang diproyeksikan pada ruang dimensi, dimana masing-masing vektor merepresentasikan atribut atau fitur dari suatu data dan kemudian ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi yang telah dilakukan terhadap data pelatihan [4].

Alur kerja dari algoritma *K-Nearest Neighbor* menurut [8] adalah sebagai berikut:

- Tentukan nilai K, sebagai parameter ketetanggaan terdekat.
- Hitung jarak antara *query instance* dan semua data pelatihan.
- Urutkan berdasarkan jarak terkecil.
- Tentukan label kelas *query instance* berdasarkan label mayoritas pada K objek data latih terdekat.

2.4 Euclidean Distance

Euclidean Distance mengukur kedekatan antara dua buah objek yang digambarkan sebagai pengukuran garis lurus atau secara langsung [3]. Metode pengukuran ini cocok apabila diimplementasikan terhadap 14 data yang memiliki nilai atribut bersifat numerikal, khususnya dengan atribut kontinu [2]. *Euclidean Distance* dihitung dengan persamaan di bawah ini:

$$d_{Euclidean}(x, y) = |x - y| = \sqrt{\sum_i^n (x_i - y_i)^2}$$

Sebagai contoh, terdapat 2 buah objek yang akan diukur tingkat kedekatannya dengan menggunakan *Euclidean Distance*. Dimana 2 buah objek tersebut adalah sebagai berikut:

$$X = [8, 2, 4, 3]$$

$$Y = [1, 6, 7, 5]$$

Maka, jarak antara objek X dan Y jika dihitung dengan persamaan adalah:

$$\begin{aligned} (x, y) &= \sqrt{(8-1)^2 + (2-6)^2 + (4-7)^2 + (3-5)^2} \\ &= \sqrt{49 + 16 + 9 + 4} \\ &= \sqrt{84} \\ &= 9,17 \end{aligned}$$

Maka dari itu berarti tingkat kedekatan X dan Y adalah 9,17.

2.5 Confusion Matrix

Confusion Matrix merupakan salah satu metode yang umum digunakan sebagai metode evaluasi

untuk mengetahui kerja model klasifikasi yang didasarkan kepada tingkat akurasi yang dihasilkan [4].

Confusion Matrix memberikan nilai kinerja model klasifikasi berdasarkan perhitungan objek yang diklasifikasikan dengan benar dan objek yang diklasifikasikan dengan salah [2]. Yang mana hasil perhitungan di gambarkan kedalam sebuah tabel, seperti yang di gambarkan pada tabel 1

Tabel 1 :Tabel Confusion Matrix

		Predicated Class		Total
		YES	NO	
Actual Class	YES	TP	FN	P
	NO	FP	TN	N
Total		P'	N'	P + N

Setelah data uji dimasukkan ke dalam *confusionmatrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity (recall)*, *specificity*, *precision* dan *accuracy*. *Sensitivity* digunakan untuk membandingkan jumlah TP terhadap jumlah *record* yang positif sedangkan *specificity* adalah perbandingan jumlah TN terhadap jumlah *record* yang negatif. Untuk menghitung *sensitivity*, *specificity*, *precision*, *accuracy* berikut ini [3] :

$$sensitivity = \frac{TP}{P} \quad specificity = \frac{TN}{N}$$

$$precision = \frac{TP}{TP + FP}$$

$$accuracy = sensitivity \frac{P}{(P + N)} + specificity \frac{N}{(P + N)}$$

Keterangan:

- P (*Positives*) : Jumlah kelas positif (YES)
- N (*Negatives*) : Jumlah kelas negatif (NO)
- TP (*True Positives*) : Jumlah kelas positif yang terklasifikasi dengan benar.
- TN (*True Negatives*) : Jumlah tupel negatif yang terklasifikasi dengan benar.
- FP (*False Positives*) : Jumlah tupel positif yang terklasifikasi dengan tidak benar.
- FN (*False Negatives*) : Jumlah tupel negatif yang terklasifikasi dengan tidak benar.

3. APLIKASI DAN PERANCANGAN

3.1 Spesifikasi Basis Data

Pada Spesifikasi Basis Data ini akan menjelaskan tabel siswa yang digunakan untuk menampung atribut data kelulusan siswa/i SMAN 12 Tangerang. Atribut yang digunakan dapat dilihat pada tabel 2 berikut:

Tabel 2 : Tabel Atribut Data

No	Nama Field	Tipe Data	Panjang	Keterangan
1.	tahun_ajaran	VarChar	10	Tahun Ajaran
2.	nis	VarChar	9	NIS
3.	nama	VarChar	50	Nama
4.	Jenis_kelamin	Char	1	Jenis Kelamin
5.	mtk1	Double	5	Nilai <i>Tryout</i> Pertama Matematika
6.	bindo1	Double	5	Nilai <i>Tryout</i> Pertama Bahasa Indonesia
7.	bing1	Double	5	Nilai <i>Tryout</i> Pertama Bahasa Inggris
8.	ipa1	Double	5	Nilai <i>Tryout</i> Pertama IPA (Rata-Rata)
9.	mtk2	Double	5	Nilai <i>Tryout</i> Kedua Matematika
10.	bindo2	Double	5	Nilai <i>Tryout</i> Kedua Bahasa Indonesia
11.	bing2	Double	5	Nilai <i>Tryout</i> Kedua Bahasa Inggris
12.	Ipa2	Double	5	Nilai <i>Tryout</i> Kedua IPA (Rata-Rata)
13.	hasilto	Char	5	Keterangan Hasil Ujian <i>Tryout</i>
14.	hasilun	Char	5	Keterangan Hasil Ujian Nasional

3.2 Sumber Data

Data merupakan salah satu komponen utama dalam kegiatan Prediksi dan Klasifikasi, karena data dapat dikatakan baik apabila dapat mewakili keadaan suatu objek yang sedang diteliti dan sesuai dengan keadaan yang ada.

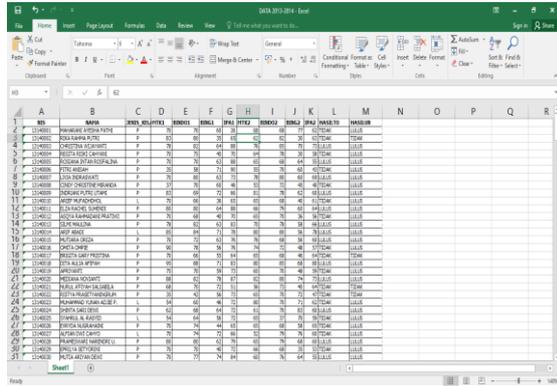
Pengumpulan data ini dilakukan dengan memperoleh data primer dan sekunder. Data Primer diperoleh dari data hasil Ujian *Tryout* dan hasil Ujian Nasional SMAN 12 Tangerang. Dataset yang dipergunakan adalah data hasil Ujian *Tryout* dan hasil Ujian Nasional selama 5 tahun terakhir, tabel 3 berikut:

Tabel 3: Data Penelitian

No	Tahun Ajaran	Jumlah Data
1	2013/2014	266
2	2014/2015	246
3	2015/2016	237
4	2016/2017	211
5	2017/2018	180

3.3 Persiapan Data

Dataset yang didapatkan dari SMAN 12 Tangerang, format data harus disesuaikan dengan yang ada pada sistem yang dibuat agar dapat melakukan uji coba sistem.



Gambar 1 : Data Pengujian

Tahap-tahap yang dilakukan untuk mengubah dataset yang didapat bisa digunakan untuk pengujian adalah:

- File harus berekstensi .xls atau .xlsx.
- Memformat ulang urutan dan nama kolom pada dataset menjadi NIS, NAMA, JENIS_KELAMIN, MTK1, BINDO1, BING1, IPA1, MTK2, BINDO2, BING2, IPA2, HASILTO, HASILUN.
- Menempatkan nama kolom yang dimulai dari sel A1.

3.4 Pembersihan Data

Pada proses ini dilakukan pembersihan data terhadap data yang dianggap *outliner* atau data yang mengandung *noise*. Data yang dianggap sebagai *outliner* yaitu:

- Data Nilai *Tryout* yang bernilai 0 atau *null*.
- Duplikasi data

3.5 Pemilihan Data

Tujuan proses ini adalah memilih atribut yang tepat untuk proses analisa data. Pemilihan atribut ini merupakan hasil diskusi dengan Guru Bidang Kurikulum SMAN 12 Tangerang yang mendapatkan keputusan terkait atribut yang dipakai untuk kelulusan Ujian Nasional SMAN 12 Tangerang. Atribut yangdidapat yaitu:

- Nilai *Tryout* pertama Matematika
- Nilai *Tryout* pertama Bahasa Indonesia
- Nilai *Tryout* pertama Bahasa Inggris
- Nilai *Tryout* pertama Ilmu Pengetahuan Alam (rata-rata)
- Nilai *Tryout* kedua Matematika
- Nilai *Tryout* kedua Bahasa Indonesia
- Nilai *Tryout* kedua Bahasa Inggris
- Nilai *Tryout* kedua Ilmu Pengetahuan Alam (rata-rata)
- Keterangan Hasil *Tryout*
- Keterangan Hasil Ujian Nasiona

3.6 Transformasi Data

Proses ini dilakukan perubahan data ke bentuk yan sesuai dengan algoritma klasifikasi yang digunakan. Yang sebelumnya diketahui bahwa algoritma *K-Nearest Neighbor* tidak cocok dilakukan untuk atribut yang bertipe kategorikal, jadi perlu dilakukan transformasi data pada atribut hasil *Tryout* menjadi atribut bertipe numerikal seperti tabel 4 dibawah ini.

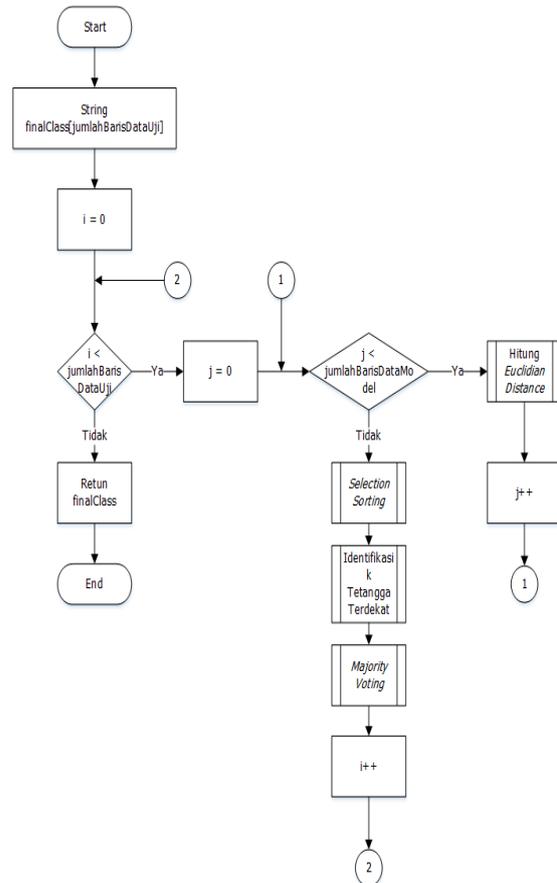
Tabel 4 : Transformasi Data

Kategorikal Hasil Tryout	Hasil Transformasi
LULUS	1
TIDAK	0

Selain atribut tersebut, nilai hasil ujian *Tryout* pertama dan nilai *Tryout* kedua juga di tranformasikan menjadi nilai sesuai kebijakan yang ada. Yaitu dengan cara menentukan nilai rata-rata dari nilai hasil ujian *Tryout* pertama dan nilai hasil ujian *Tryout* kedua seperti di bawah ini.

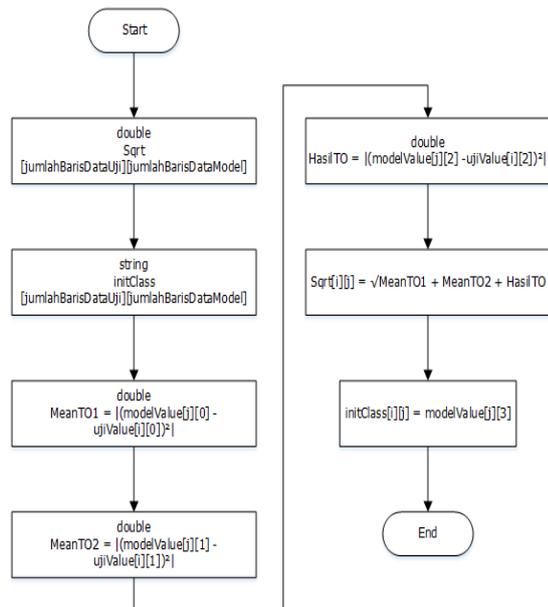
Rata-rata Nilai Ujian *Tryout* pertama dan kedua:
 $(MTK1 + BINDO1 + BING1 + IPA1)/4$
 $(MTK2 + BINDO2 + BING2 + IPA2)/4$

3.7 Flowchart K-Nearest Neighbor



Gambar 2 : Flowchart K-Nearest Neighbor

3.8 FlowChart Euclidean Distance



Gambar 3 : Flowchart Euclidean Distance

4. HASIL DAN PEMBAHASAN

4.1. Tampilan Layar Menu Utama

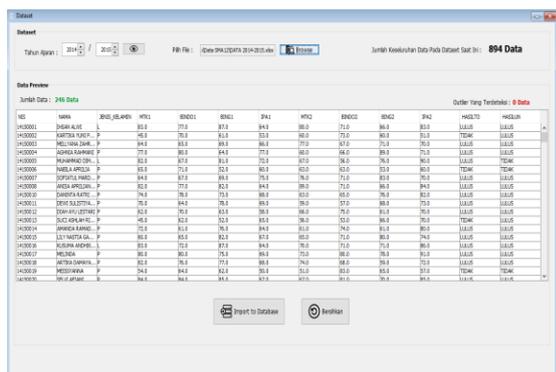
Dalam *form* ini terdapat beberapa menu yang mempunyai fungsi yang berbeda-beda yaitu, menu dataset, menu, menu klasifikasi.



Gambar 4: Tampilan Layar Menu Utama

4.2. Tampilan Layar Dataset

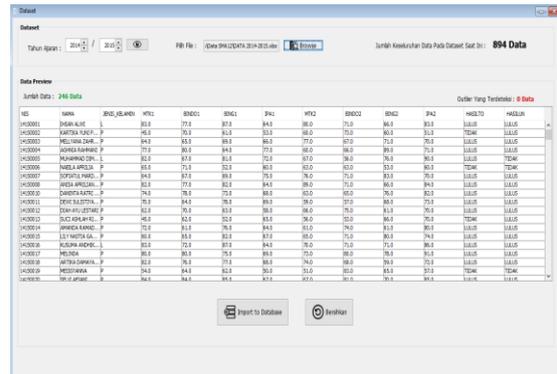
Form dataset berfungsi untuk meng-import dataset yang berekstensi *.xls atau *.xlsx kedalam database yang ingin digunakan sebagai dataset uji dan dataset model.



Gambar 5 : Tampilan Layar Dataset

4.3. Tampilan Layar Evaluasi

Form evaluasi terdapat fitur *Single Testing* dan *Multi Testing* berfungsi untuk pengujian serta mencari nilai k yang memiliki akurasi terbaik.



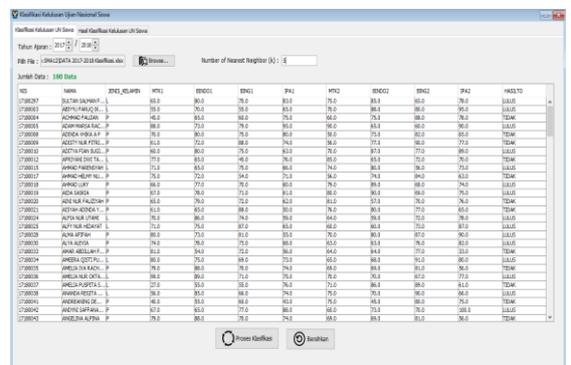
Gambar 6 : Tampilan Layar Evaluasi Tab Dataset Pengujian



Gambar 7 : Tampilan Layar Evaluasi Tab Hasil Evaluasi

4.4. Tampilan Layar Klasifikasi

Form Klasifikasi ini berfungsi untuk melakukan prediksi dan klasifikasi kelulusan Ujian Nasional menggunakan dataset yang belum memiliki keterangan Hasil UN dengan dataset Model yang sudah di-import sebelumnya.



Gambar 8 : Tampilan Layar Klasifikasi

4.5. Hasil Uji Coba Program

Pada tahap ini merupakan hasil pengujian program yang bertujuan untuk mengevaluasi parameter nilai(k) yang dipilih. Pengujian kali ini akan dilakukan 2 tipe yaitu *Single Testing* dan *Multi Testing*.

a. Pengujian Multi Testing

Pengujian *multi testing* ini menggunakan nilai *Tryout 5* tahun terakhir sebagai dataset model yang akan dibandingkan dengan nilai *Tryout 1* tahun terakhir sebagai dataset uji. Nilai k yang digunakan adalah 5, 7, 9, 11, 13.

Pengujian *multi testing* pertama dataset yang digunakan adalah dataset tahun ajaran 2013/2014, 2014/2015, 2015/2016, 2016/2017 sebagai dataset model dan 2017/2018 sebagai dataset uji.

Tabel 5 : Multi Testing Pertama

K	Accuracy	Recall	Precision
5	88.33%	97.95%	88.88%
7	87.77%	98.63%	87.87%
9	85.55%	96.59%	87.11%
11	86.11%	97.95%	86.74%
13	84.44%	96.59%	86.06%

Pengujian *multi testing* kedua dataset yang digunakan adalah dataset tahun ajaran 2013/2014, 2014/2015, 2015/2016, 2017/2018 sebagai dataset model dan 2016/2017 sebagai dataset uji.

Tabel 6 : Multi Testing Kedua

K	Accuracy	Recall	Precision
5	79.14%	93.64%	83.07%
7	79.62%	94.21%	83.16%
9	79.14%	94.21%	82.74%
11	79.14%	94.21%	82.74%
13	72.72%	94.21%	81.50%

Pengujian *multi testing* ketiga dataset yang digunakan adalah dataset tahun ajaran 2013/2014, 2014/2015, 2016/2017, 2017/2018 sebagai dataset model dan 2015/2016 sebagai dataset uji.

Tabel 7 : Multi Testing Ketiga

K	Accuracy	Recall	Precision
5	91.56	99.06	92.20
7	91.56	99.06	92.20
9	91.13	99.06	91.81
11	92.40	99.53	92.64
13	91.98	99.53	92.24

Pengujian *multi testing* keempat dataset yang digunakan adalah dataset tahun ajaran 2013/2014, 2015/2016, 2016/2017, 2017/2018 sebagai dataset model dan 2014/2015 sebagai dataset uji.

Tabel 8 : Multi Testing Keempat

K	Accuracy	Recall	Precision
5	92.27%	96.52%	95.27%
7	93.08%	97.39%	95.31%
9	93.90%	98.69%	93.47%
11	94.30%	99.13%	95%
13	94.30%	99.56%	94.62%

Pengujian *multi testing* kelima dataset yang digunakan adalah dataset tahun ajaran 2014/2015, 2015/2016, 2016/2017, 2017/2018 sebagai dataset model dan 2013/2014 sebagai dataset uji.

Tabel 9 : Multi Testing Kelima

K	Accuracy	Recall	Precision
5	93.98%	98.79%	94.96%
7	93.60%	98.79%	94.59%
9	93.60%	98.79%	94.59%
11	93.60%	98.79%	94.23%
13	93.60%	98.79%	94.59%

b. Pengujian Single Testing

Pengujian *single testing* ini menggunakan nilai *Tryout 5* tahun terakhir sebagai dataset model yang akan dibandingkan dengan nilai *Tryout 1* tahun terakhir sebagai dataset uji. Nilai k yang digunakan adalah 7.

Pengujian *single testing* pertama ini yang digunakan sebagai dataset uji adalah tahun ajaran 2017/2018 dan dataset model tahun ajaran 2013/2014, 2014/2015, 2015/2016, 2016/2017.

Confusion Matrix

	Predicted LULUS	Predicted TIDAK	Class Recall
Actual LULUS	145	2	98,64 %
Actual TIDAK	20	13	39,39 %
Class Precision	87,88 %	86,67 %	

*Accuracy
Overall Accuracy is 87,78 %

Gambar 9 : Single Testing Pertama

Pengujian *single testing* kedua ini yang digunakan sebagai dataset uji adalah tahun ajaran 2016/2017 dan dataset model tahun ajaran 2013/2014, 2014/2015, 2015/2016, 2017/2018.

Confusion Matrix

	Predicted LULUS	Predicted TIDAK	Class Recall
Actual LULUS	163	10	94,22 %
Actual TIDAK	33	5	13,16 %
Class Precision	83,16 %	33,33 %	

*Accuracy
Overall Accuracy is 79,62 %

Gambar 10 : Single Testing Kedua

Pengujian *single testing* ketiga ini yang digunakan sebagai dataset uji adalah tahun ajaran 2015/2016 dan dataset model tahun ajaran 2013/2014, 2014/2015, 2016/2017, 2017/2018.

Confusion Matrix			
	Predicted LULUS	Predicted TIDAK	Class Recall
Actual LULUS	213	2	99,07 %
Actual TIDAK	18	4	18,18 %
Class Precision	92,21 %	66,67 %	

***Accuracy**
Overall Accuracy is 91,56 %

Gambar 11 : Single Testing Ketiga

Pengujian *single testing* keempat ini yang digunakan sebagai dataset uji adalah tahun ajaran 2014/2015 dan dataset model tahun ajaran 2013/2014, 2015/2016, 2016/2017, 2017/2018.

Confusion Matrix			
	Predicted LULUS	Predicted TIDAK	Class Recall
Actual LULUS	224	6	97,39 %
Actual TIDAK	11	5	31,25 %
Class Precision	95,32 %	45,45 %	

***Accuracy**
Overall Accuracy is 93,09 %

Gambar 12 : Single Testing Keempat

Pengujian *single testing* kelima ini yang digunakan sebagai dataset uji adalah tahun ajaran 2013/2014 dan dataset model tahun ajaran 2014/2015, 2015/2016, 2016/2017, 2017/2018.

Confusion Matrix			
	Predicted LULUS	Predicted TIDAK	Class Recall
Actual LULUS	245	3	98,79 %
Actual TIDAK	14	4	22,22 %
Class Precision	94,59 %	57,14 %	

***Accuracy**
Overall Accuracy is 93,61 %

Gambar 13 : Single Testing Kelima

5. KESIMPULAN

5.1. Kesimpulan

Melalui Proses pengerjaan dan pengujian dalam penelitian ini, maka dari itu dapat disimpulkan beberapa hal, yaitu:

- Algoritma *K-Nearest Neighbor* dapat diimplementasikan dalam sistem ini.
- Waktu pengujian yang relatif cepat dibandingkan dengan pengujian manual sebelumnya.
- Proses prediksi dan klasifikasi terbilang cukup efisien dan efektif.
- Nilai (k) harus bernilai ganjil untuk menghindari munculnya hasil jumlah jarak yang sama pada proses pengklasifikasian.
- Perbedaan tingkat akurasi tidak terlalu jauh antar nilai (k).

Nilai akurasi rata-rata sistem ini terbilang cukup besar, yaitu 88.58% dengan nilai k terbaik adalah k=7.

5.2. Saran

Beberapa saran yang dapat diberikan untuk pengembangan sistem lebih lanjut dengan harapan dapat menghasilkan penelitian yang lebih baik kedepannya, sebagai berikut:

- Dapat mengimplementasikan algoritma klasifikasi lain yang lebih bervariasi untuk proses prediksi dan klasifikasi.
- Dapat menggunakan metode pengukuran jarak seperti, *Manhattan Distance*, *Cosine Distance*, *Correlation Distance*, *Hamming Distance*, dan lainnya.
- Menggunakan nilai k yang lebih bervariasi untuk mengetahui perbedaan hasil pengujian.
- Menggunakan dataset yang lebih banyak lagi, agar dapat mengetahui ketepatan sistem dalam memprediksi serta mengklasifikasikan dataset yang ada.
- Mengoptimalkan kinerja sistem agar lebih cepat lagi dalam pemrosesan dataset.
- Menambahkan pengklasifikasian serta prediksi jurusan perguruan tinggi melalui bobot nilai dan peminatan siswa/i.

6. DAFTAR PUSTAKA

- Obbie Kristanto, "Penerapan Algoritma Klasifikasi Data Mining Id3 Untuk Menentukan Penjurusan Siswa SMAN 6," pp. 6–7, 2013.
- F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. 2011.
- M. K. J.Han, J.Pei, *Data Mining: Concepts and Techniques*, vol. 3. 2012.
- Nursalim, Suprapedi, and H.himawan, "Klasifikasi Bidang Kerja Lulusan Menggunakan Algoritma K-Nearest Neighbor," *Tekno. Inf.*, vol. 10, no. April, pp. 31–43, 2014.
- K. U. S. bin Lukman and A. Labellapansa, "Analisa Nilai Lamda Model Jarak Minkowsky Untuk Penentuan Jurusan SMA (Studi Kasus di SMA Negeri 2 Tualang)," *J. Tek. Inform. dan Sist. Inf.*, vol. 1, pp. 163–171, 2015.
- C. Y. Sillueta, "IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI KELULUSAN Perpustakaan Universitas Budi Luhur," vol. 1, no. Huda 2010, 2016.
- X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.
- Kurniawan and P. Sukma, "Perancangan Data Mining untuk Analisis Kriteria Nasabah Kredit yang Potensial dan Manfaatnya untuk Customer Relationship Management Perbankan," no. 11, pp. 155–174, 2015.