UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume I Nomor 2 Tahun 2023 Halaman 67-73

ISSN (*Online*): 2985-475X

Comparison Between Naïve Bayes and K-Nearest Neighbor for DKI Jakarta Air Pollution Standard Index Classification

Nurdalia, Zilrahmi*, Dony Permana, Admi Salma
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: zilrahmi@fmipa.unp.ac.id

Submitted : 27 Desember 2022
Revised : 09 Januari 2023
Accepted : 13 Februari 2023

ABSTRACT

Using certain algorithms or procedures in accordance with knowledge or information, data mining is the process of extracting and searching for usable knowledge and information. Naive Bayes and K-Nearest Neighbor are the data mining classification techniques employed in this study. It is possible to categorize the DKI Jakarta air pollution standard index in 2021 based on six air pollutants, including dust particles (PM10), dust particles (PM2.5), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), and nitrogen dioxide by using the Naive Bayes and K-Nearest Neighbor methods. The test was conducted to ascertain the precision of forecasting the DKI Jakarta air pollution standard index in 2021 using the confusion matrix evaluation value. The best performance of the two methods is in the Naïve Bayes algorithm with high Naïve Bayes sensitivity values for all categories. Even though there are data in unbalanced categories, the Naïve Bayes algorithm shows good performance in accuracy, sensitivity, specificity.

Keywords: *Confusion Matrix, Data Mining, Naïve Bayes, KNN*

I. PENDAHULUAN

Indeks standar pencemaran udara (ISPU) merupakan suatu angka tanpa satuan yang menggambarkan keadaan kualitas udara ambien. Tujuan disusunnya ISPU untuk memberikan informasi kualitas udara ambien kepada masyarakat di waktu dan lokasi tertentu, serta sebagai bahan pertimbangan dalam melakukan upaya pengendalian pencemaran udara. Berdasarkan hasil pemantauan kualitas udara Amerika Serikat *Air Quality Index* (AQI) pada Juni 2022 DKI Jakarta menduduki ranking pertama kota paling berpolusi di Indonesia dan juga dunia dengan angka rentang mencapai 196 kategori tidak sehat. Dalam menentukan tingkat ISPU dapat dipermudah dengan proses klasifikasi *data mining*. *Data mining* adalah proses untuk mengumpulkan informasi yang berguna dengan menggunakan algoritma, metode, atau teknik yang sesuai dengan informasi yang dikumpulkan (Buulolo, 2020). *Naive Bayes* dan *K-Nearest Neighbor* (KNN) adalah dua metode klasifikasi yang digunakan dalam *data mining* (Adinugroho, 2018). *Naïve Bayes* memiliki akurasi yang sangat baik dengan perhitungan sederhana dalam bidang pembelajaran klasifikasi (Mustika *et al*, 2021:97). KNN mengklasifikasikan suatu objek dengan memperhatikan kelas yang paling dekat dengannya. (Prasetyo, 2014)

Perbandingan metode *Naïve Bayes* dan KNN pernah dilakukan oleh penelitian sebelumnya. Pada Tahun 2014, Putri dkk melakukan klasifikasi *Naïve Bayes* dan KNN pada analisis data status kerja di Kabupaten Demak. Berdasarkan penelitian ini, diketahui nilai akurasi pada klasifikasi *Naïve Bayes* yaitu sebesar 94%. Sedangkan nilai akurasi pada klasifikasi KNN dengan menggunakan nilai parameter *K* adalah 7 yaitu sebesar 96%. Selanjutnya Yusra dkk (2016) membandingkan metode klasifikasi *Naïve Bayes* dan KNN pada data tugas akhir mahasiswa jurusan Teknik Informatika. Dari penelitian ini, diketahui nilai akurasi pada klasifikasi *Naïve Bayes* yaitu sebesar 87%. Sedangkan nilai akurasi pada klasifikasi KNN dengan menggunakan nilai parameter *K* adalah 5 yaitu sebesar 84%.

Berdasarkan penelitian yang telah dilakukan sebelumnya, penelitian yang dilakukan untuk menentukan algoritma terbaik hanya berdasarkan evaluasi nilai akurasi, tanpa melakukan nilai evaluasi lainnya seperti nilai *sensitivity* dan *specificity* sebagai bahan pertimbangan untuk menentukan model algoritma terbaik. Maka pada penelitian ini peneliti membandingkan nilai evaluasi akurasi, *sensitivity* dan *specificity* untuk mendapatkan algoritma terbaik terhadap ISPU DKI Jakarta Tahun 2021. Paparan yang telah dibahas, penulis kemudian melakukan penelitian. yang berjudul **"Perbandingan *Naïve Bayes* dan *K-Nearest Neighbor* untuk Klasifikasi Indeks Standar Pencemaran Udara DKI Jakarta".**

DOI: 10.24036/ujsds/vol1-iss2/29

ISSN (*Online*): 2985-475X

II. METODE PENELITIAN

A. Algoritma *Naïve Bayes*

Naive Bayes didasarkan pada teorema Bayes, cepat dan akurat dan dapat digunakan untuk data berukuran besar. Langkah-langkah untuk menghitung metode algoritma *Naïve Bayes* (Jananto, 2013)

1. Mengitung kelas probabilitas untuk mencari nilai *prior probabilitas* dari data *training* pada persamaan (1),
   $P(H) = \frac{x_i}{n}$ (1)
   dengan x_i adalah nilai variabel ke-*I* dan n adalah ukuran sampel total
2. Hitung rata-rata dan standar deviasi masing-masing nilai variabel.
   Untuk menghitung nilai *mean* yaitu pada persamaan (2),
   $$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2)$$
   dengan μ merupakan nilai rataan dan x_i merupakan nilai variabel ke-*i*
   Untuk menentukan nilai standar deviasi, gunakan persamaan. (3),
   $$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}} \quad (3)$$
   dengan σ merupakan standar deviasi
3. Menghitung nilai *Densitas Gauss* yaitu pada persamaan (4),
   $$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\pi_{ij})^2}{2\sigma^2 ij}} \quad (4)$$
   dengan *P* merupakan peluang, X_i merupakan variabel *I* dan Y merupakan kelas yang dicari
4. Menghitung nilai *posterior probabilitas* yaitu pada persamaan (5),
   $$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (5)$$
   dengan X merupakan data dengan kelas yang tidak terklasifikasi, H merupakan data untuk hipotesis X yang merupakan kelas tertentu., P(H|X) merupakan peluang H dengan syarat X terjadi, P(H) merupakan kemungkinan yang dihipotesiskan H (probabilitas sebelumnya), P(X|H) merupakan X kelayakan ditentukan oleh keadaan pada hipotesis H dan P(X) adalah kemungkinan X

B. Algoritma *K-Nearest Neighbor*

Pertama yang harus dilakukan saat menggunakan KNN adalah berapa banyak K tetangga terdekat yang akan digunakan untuk mengklasifikasikan data uji. Nilai optimal untuk bilangan K adalah bilangan ganjil, seperti K = 1, 3, 5, dan seterusnya (Isman dkk, 2021). Menurut Nurjana dkk (2020) cara menghitung menggunakan algoritma KNN.

1. Menentukan K jumlah tetangga terdekat
2. Terhadap data sampel yang disediakan, hitung jarak Euclidean kuadrat setiap objek. dengan yaitu pada persamaan (6),
   $$D(x,y) = \sqrt{\sum_{k-1}^{n}(x_k - y_k)^2} \quad (6)$$
   dengan Jarak euclidean adalah D(x,y), x_k yaitu data latih, y_k yaitu data uji, *I* yaitu variabel data dan n merupakan informasi ukuran
3. Kemudian menggunakan objek dalam kelompok dengan jarak Euclidean terdekat.

C. Pengukuran Tingkat Akurasi Klasifikasi

Menurut Prasetyo (2012) karena tidak semua hasil dari kinerja klasifikasi pada data dapat 100% akurat, diperlukan metode pengukuran dalam model klasifikasi *confussion matrix*, yang ditunjukkan pada Tabel 1, digunakan untuk mengevaluasi model klasifikasi.

Tabel 1. *Confussion Matrix* Untuk Mengklasifikasikan Dua Kelas

f_{ij}		Kelas hasil prediksi (j)	
		Kelas=1	**Kelas=0**
Kelas asli (*i*)	**Kelas=1**	TP	FN
	Kelas=0	FP	TN

DOI: 10.24036/ujsds/vol1-iss2/29

ISSN (*Online*): 2985-475X

Keterangan:

a. TP (*True Positif*). *Observed Class* benar dengan hasil *predicted class* benar.
b. TN (*True Negatif*). *Observed Class* salah dengan hasil *predicted class* salah.
c. FP (*False Positif*). *Observed Class* salah dengan hasil *predicted class* benar.
d. FN (*False Negatif*). *Observed Class* benar dengan hasil *predicted class* salah.

Menurut Mustika dkk (2021) terdapat tiga *performance metrics* yang digunakan, yaitu: *accurasy, specificity, sensitivity.*

1) Akurasi (*Accuracy*)

Akurasi adalah proporsi prediksi akurat (termasuk positif dan negatif) untuk semua data. Nilai untuk akurasi dapat diperoleh. yaitu pada persamaan (7),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

2) Specificity

Specificity adalah dibandingkan dengan semua bukti negatif, kejujuran meramalkan hal yang negatif.. Nilai *specificity* dapat diperoleh yaitu pada persamaan (8),

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

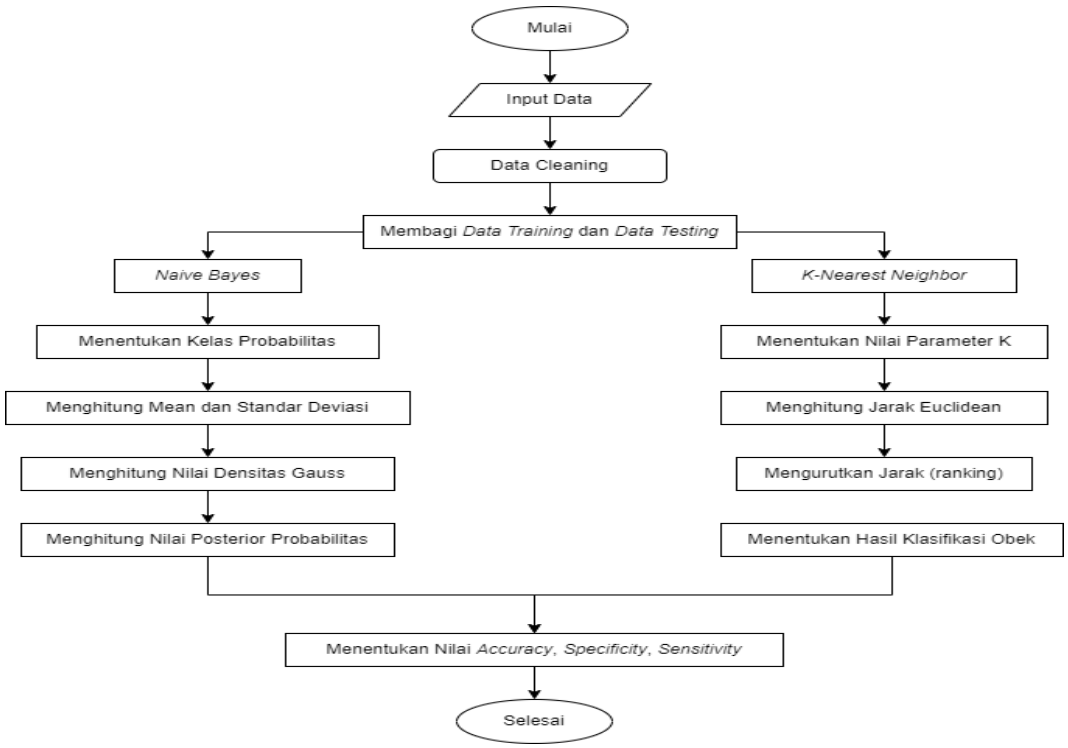
3) Recall (*Sensitivity*)

Recall adalah proporsi prakiraan positif akurat terhadap semua data positif akurat. Nilai ingat dapat dicapai. yaitu pada persamaan (9),

$$Sensitivity = \frac{TP}{TP+FN} \quad (9)$$

D. Sumber Data dan Teknik Analisis Data

Jenis penelitian berdasarkan permasalahan dan penelitian ini memiliki tujuan penelitian terapan. Pada penelitian ini data yang digunakan yaitu data sekunder, diantaranya *website* kaggle yaitu data ISPU DKI Jakarta Tahun 2021. Penelitian ini terdiri dari kandungan udara PM_{10}, $PM_{2,5}$, SO_2, CO, O_3, dan NO_2. Adapun langkah-langkah analisis dalam penelitian seperti yang terlihat pada Gambar 1.



Gambar 1. Diagram Alir Algoritma *Naïve Bayes* dan *K-Nearest Neighbor*

DOI: 10.24036/ujsds/vol1-iss2/29

ISSN (*Online*): 2985-475X

III. HASIL DAN PEMBAHASAN

A. Pembersihan data

Data penelitian ini adalah data ISPU DKI Jakarta Tahun 2021. Pada tahap ini dilakukan pembersihan terhadap data yang mengandung *missing value* akan dibersihkan dengan cara mencari nilai *mean*. Variabel yang yang mengandung *missing value* terdapat pada kandungan udara $PM_{2,5.}$

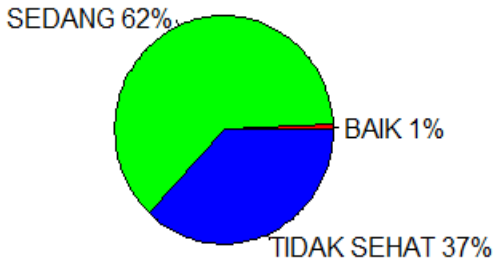
B. Analisis Statistika deskriptif

Berdasarkan gambaran data ISPU DKI Jakarta Tahun 2021 secara umum ditunjukkan pada Tabel 2.

Tabel 2. Variabel dalam Penelitian Deskriptif

Variabel	*Mean*	*Min*	*Max*	Standar Deviasi
Kandungan Udara PM_{10} ($\mu g/m^3$)	60,51	19	179	15,13
Kandungan Udara $PM_{2,5}$ ($\mu g/m^3$)	94,75	33	174	23,07
Kandungan Udara SO_2 ($\mu g/m^3$)	52,72	37	126	11,17
Kandungan Udara CO ($\mu g/m^3$)	15,39	7	47	5,84
Kandungan Udara O_3 ($\mu g/m^3$)	49,81	20	151	12,21
Kandungan Udara NO_2 ($\mu g/m^3$)	34,12	9	134	15,95

Berdasarkan Tabel 2, diperoleh informasi data kandungan udara DKI Jakarta Tahun 2021. Dari keseluruhan kandungan udara nilai rata-rata tertinggi suatu nilai dapat ditemukan di kandungan udara $PM_{2,5}$ yaitu sebesar 94,75, hal ini menunjukkan bahwa nilai tersebut melebihi nilai ambang batas artinya kondisi udara di DKI Jakarta rata-rata dikategorikan sedang. Selanjutnya dari keseluruhan kandungan udara memiliki nilai maksimum yaitu sebesar 179 artinya ada hari-hari tertentu di DKI Jakarta terdapat kandungan udara PM_{10} yang memiliki kategori yang tidak sehat. Selanjutnya dari keseluruhan kandungan udara Tahun 2021 memiliki nilai keragaman yang kecil terdapat pada kandungan udara CO artinya antar hari tidak terlalu berbeda. Selanjutnya berdasarkan kandungan udara DKI Jakarta Tahun 2021 terdapat kondisi udara DKI Jakarta yang dikategorikan baik, sedang dan tidak sehat. Secara keseluruhan kondisi udara DKI Jakarta pada kategori baik hanya sebesar 1%, hal ini menunjukkan kondisi udara DKI Jakarta sangat tercemar. seperti yang terlihat pada Gambar 2.



Gambar 2. Diagram Lingkaran Kandungan Udara DKI Jakarta Tahun 2021

C. Membagi Data latih dan Data uji

Sebelum melakukan analisis klasifikasi tehadap indeks standar pencemaran udara DKI Jakarta Tahun 2021. Langkah pertama adalah data dibagi mejadi data latih dan data uji yaitu data latih 80% yaitu 292 amatan dan data uji 20% yaitu 73 amatan.

DOI: 10.24036/ujsds/vol1-iss2/29

ISSN (*Online*): 2985-475X

D. Algoritma *Naïve Bayes*

Hasil kinerja algoritma *Naïve Bayes* dalam melakukan klasifikasi dilihat pada Tabel 3 *confusion matrix*.

Tabel 3. Hasil Klasifikasi *Naïve Bayes* Confusion Matrix

Prediksi	Aktual		
	Baik	Sedang	Tidak Sehat
Baik	1	1	0
Sedang	0	41	2
Tidak Sehat	0	3	25

Berdasarkan Tabel 3, nilai prediksi yang paling banyak benar terdapat kategori sedang yaitu 41 prediksi dan nilai prediksi yang benar paling sedikit terdapat pada kategori baik yaitu hanya 1 prediksi.

E. Algoritma *K-Nearest Neihbor*

Pengolahan pengklasifikasian menggunakan algoritma KNN untuk mencari nilai parameter K=1, 3, 5, dan 7 yang digunakan untuk mendapatkan nilai akurasi terbaik yang ditunjukkan pada Tabel 4.

Tabel 4. Evaluasi Akurasi Parameter KNN

Parameter *K*	Akurasi (%)
K=1	0,86
K=3	0,88
K=5	0,88
K=7	0,90

Berdasarkan Tabel 4 nilai akurasi yang menghasilkan nilai terbaik terdapat pada parameter *K*=7 yaitu nilai akurasi sebesar 90%. Untuk mengetahui hasil pengujian atau kinerja algoritma KNN dalam melakukan klasifikasi gunakan *confusion matrix* Tabel 5.

Tabel 5. Hasil Klasifikasi KNN *Confusion Matrix*

Prediksi	Aktual		
	Baik	Sedang	Tidak Sehat
Baik	0	0	0
Sedang	1	41	4
Tidak Sehat	0	4	23

Berdasarkan Tabel 5, nilai prediksi yang paling banyak benar terdapat kategori sedang yaitu 41 prediksi dan pada kategori baik tidak ada satupun nilai prediksi yang benar.

F. Perbandingan Nilai Akurasi Ketepatan Klasifikasi

Perbandingan nilai *accuracy*, *specificity*, *sensitivity* dengan menggunakan teknik klasifikasi *Naive Bayes* dan KNN terhadap ISPU DKI Jakarta Tahun 2021 ditunjukkan pada Tabel 6.

Tabel 6. Ketepatan Klasifikasi *Naïve Bayes* dan KNN

Nilai Akurasi *Naïve Bayes*	**Nilai Akurasi KNN**
0,9178	0,8767

Dilihat pada Tabel 6, dalam memprediksi ketepatan kandungan udara DKI Jakarta Tahun 2021 dengn memanfaatkan teknik klasifikasi KNN dan *Naive Bayes* dengan nilai parameter K=7, maka diperoleh tingkat *accuracy* yang tertinggi adalah metode *Naïve Bayes* yaitu sebesar 91% ini menunjukkan sebesar 91% keakuratan model *Naïve Bayes* dalam mengklasifikasikan dengan benar.

DOI: 10.24036/ujsds/vol1-iss2/29

ISSN (*Online*): 2985-475X

Tabel 7. Nilai Ketepatan Klasifikasi *Naïve Bayes* dan KNN

Metode	Kategori	*Sensitivity*	*Specificity*
Naïve Bayes	Baik	1	0,98
	Sedang	0,91	0,92
	Tidak Sehat	0,92	0,93
KNN	Baik	0	1
	Sedang	0,91	0,82
	Tidak Sehat	0,85	0,91

Dilihat pada Tabel 7, dalam memprediksi ketepatan kandungan udara DKI Jakarta Tahun 2021 memanfaatkan teknik tersebut klasifikasi *Naïve Bayes* dan KNN. Metode *Naïve Bayes* diperoleh nilai *sensitivity* yang menunjukkan prediksi untuk kategori baik 100%, untuk kategori 91% dan untuk kategori tidak sehat 92%. Selanjutnya nilai *specificity* yang menunjukkan prediksi untuk kategori baik 98%, untuk kategori sedang atau 93% dan untuk kategori tidak sehat 92%. Berdasarkan nilai tersebut dinyatakan dalam melakukan ketepatan memprediksi kandungan udara sangat baik.

Untuk metode KNN dengan nilai parameter K=7 diperoleh nilai *sensitivity* yang menunjukkan prediksi untuk kategori baik 0%, untuk kategori sedang 91% dan untuk kategori tidak sehat 85%. Selanjutnya nilai *specificity* yang menunjukkan prediksi untuk kategori baik 100%, untuk kategori sedang 82% dan untuk kategori tidak 91%. Berdasarkan hal tersebut dapat dinyatakan bahwa dalam melakukan ketepatan memprediksi kandungan udara kurang baik karena terdapat nilai *sensitivity* untuk kategori baik sebesar 0% maka dinyatakan bahwa sistem klasifikasi tidak bekerja dengan baik.

KESIMPULAN

Metode *Naive Bayes* dan KNN digunakan dalam penelitian ini dalam memprediksi indeks standar pencemaran udara DKI Jakarta Tahun 2021. Maka penelitian menunjukkan bahwa metode *Naive Bayes* lebih unggul dari algoritma KNN mengklasifikasikan ISPU DKI Jakarta Tahun 2021, dengan nilai s*ensitivity Naïve Bayes* yang tinggi untuk keseluruhan kategori meskipun terdapat data yang kategori yang tidak seimbang, maka algoritma *Naïve Bayes* menunjukkan performa yang baik dalam *accuracy, sensitivity, specificity*. Maka dengan ini penelitian diharapkan pemerintah dapat membantu Dinas Lingkungan Hidup Kota Jakarta dalam mengedukasi masyarakat tentang dampak negatif pencemaran udara agar dapat melakukan tindakan pencegahan saat beraktivitas di luar ruangan.

DAFTAR PUSTAKA

Adinugroho, S., dan Yuita, A. S. (2018). *Implementasi Data Mining Menggunakan Weka*. Malang: UB Press.

AK. (2020). "*Portal Direktorat Pengendalian Pencemaran Udara Ditjen Ppkl Klhk*", http://Menlhk.go.id, diakses pada tanggal 20 September 2022.

Buulolo, E. (2020). *Data Mining untuk Perguruan Tinggi*. Yogyakarta: Budi Utama.

Isman, Andani Ahmad, & Abdul Latief. (2021). Perbandingan Metode KNN Dan LBPH Pada Klasifikasi Daun Herbal. *Jurnal RESTI (Rekayasa Sistem Dan TeknologiInformasi)*, *5*(3), 557–564. https://doi.org/10.29207/resti.v5i3.3006.

Kaggle."Indeks Standar Pencemaran Udara DKI Jakarta Tahun 2021". Publikasi Kaggle diakses dari http://www.kaggle.com/,tanggal 15 Juli 2022.

Menlhk."Peraturan Mentri Lingkungan Hidup dan Kehutanan Republik Indonesia" Publikasi Menlhk diakses dari http://ditppu.menlhk.go.id/,tanggal 10 Agustus 2022.

Mustika, dkk. (2021). *Data Mining dan Aplikasinya*. Bandung: Widina Bhakti Persada Bandung.

DOI: 10.24036/ujsds/vol1-iss2/29

ISSN (*Online*): 2985-475X

Prasetyo, E. (2012). *Data Mining*: *Konsep dan Aplikasi Menggunakan Matlab,* Yogyakarta: Andi Offset.

Prasetyo, E. (2014). *Data Mining*: *Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

Putri, R. E., Suparti, & Rahmawati, R. (2014). Perbandingan Metode Klasifikasi Naive Bayes Dan K-Nearest Neighbour Pada Analisis Data Status Kerja di Kab.Demak. *Jurnal Gaussian*, *3*, 831–838.

Yusra. Olivita, D., & Vitriani, Y. (2016). Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode *Naïve Bayes* Classifer dan K-Nearest Neighbor. *Jurnal Sains, Teknologi dan Industri*, 14(1), 79-85

DOI: 10.24036/ujsds/vol1-iss2/29