

**ANALISIS KINERJA ALGORITMA DATA MINING K-NEAREST  
NEIGHBOR (KNN) DAN NAIVE BAYES UNTUK KLASIFIKASI  
KELUHAN MASYARAKAT TERHADAP SOLUSI  
PENANGANAN SAMPAH**

SKRIPSI

Diajukan untuk Memenuhi Sebagian Syarat Guna Memperoleh  
Gelar Sarjana Matematika dalam Ilmu Matematika Murni



**Oleh : RAHYAN ELENA MAHATIARA**

**NIM : 1908046003**

PROGRAM STUDI MATEMATIKA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI WALISONGO SEMARANG  
2023

**ANALISIS KINERJA ALGORITMA DATA MINING K-NEAREST  
NEIGHBOR (KNN) DAN NAIVE BAYES UNTUK KLASIFIKASI  
KELUHAN MASYARAKAT TERHADAP SOLUSI  
PENANGANAN SAMPAH**

**SKRIPSI**

Diajukan untuk Memenuhi Sebagian Syarat Guna Memperoleh  
Gelar Sarjana Matematika dalam Ilmu Matematika Murni



**Oleh : RAHYAN ELENA MAHATIARA**

**NIM : 1908046003**

**PROGRAM STUDI MATEMATIKA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI WALISONGO SEMARANG  
2023**

## PERNYATAAN KEASLIAN NASKAH

Yang bertanda tangan dibawah ini :

Nama : Rahyan Elena Mahatiara

NIM : 1908046003

Jurusan/Program Studi : Matematika

Menyatakan bahwa skripsi yang berjudul :

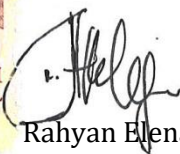
### **ANALISIS KINERJA ALGORITMA DATA MINING K-NEAREST NEIGHBOR (KNN) DAN NAIVE BAYES UNTUK KLASIFIKASI KELUHAN MASYARAKAT TERHADAP SOLUSI PENANGANAN SAMPAH**

Secara keseluruhan adalah hasil penelitian/karya saya sendiri, kecuali bagian tertentu yang dirujuk sumbernya.

Semarang, 21 Juni 2023

Saya yang menyatakan,



  
Rahyan Elena Mahatiara

1908046003

## Halaman Pengesahan



KEMENTERIAN AGAMA  
UNIVERSITAS ISLAM NEGERI WALISONGO  
**FAKULTAS SAINS DAN TEKNOLOGI**  
Jl. Prof. Dr. Hamka Ngaliyan Semarang  
Telp.024-7601295 Fax.7615387

### PENGESAHAN

Naskah skripsi berikut ini :

Judul : **Analisis Kinerja Algoritma Data Mining K-Nearest Neighbor dan Naive Bayes Untuk Klasifikasi Keluhan Masyarakat Terhadap Solusi Penanganan Sampah**

Penulis : Rahyan Elena Mahatiara

NIM : 1908046003

Jurusan : Matematika

Telah diajukan dalam sidang tugas akhir oleh Dewan Penguji Fakultas Sains dan Teknologi UIN Walisongo dan dapat diterima sebagai salah satu syarat memperoleh gelar sarjana dalam Ilmu Matematika.

Semarang, 27 Juni 2023

### DEWAN PENGUJI

Ketua Sidang,

**Yolanda Norasia, M.Si.**  
NIP. 199409232019032011

Penguji Utama I,

**Any Muanalifah, M.Si, P.hd**  
NIP. 198201132011012009

Pembimbing I,

**Ariska Kurnia Rachmawati, M.Sc**  
NIP. 198908112019032019

Sekretaris Sidang,

**Ariska Kurnia Rachmawati, M.Sc**  
NIP. 198908112019032019

Penguji Utama II,

**Dinni Rahma Oktaviani, M.Si**  
NIP. 199410092019032017

Pembimbing II,

**Eva Khoirun Nisa, M.Si.**  
NIP. 198701022019032010

## NOTA DINAS

Semarang, 23 Juni 2023

Yth. Ketua Program Studi Matematika

Fakultas Sains dan Teknologi

UIN Walisongo Semarang

*Assalamualaikum warrahmatualli wabarakatuh*

Dengan ini diberitahukan bahwa saya telah melakukan bimbingan, arahan dan koreksi naskah skripsi dengan :

Judul : Analisis Kinerja Algoritma Data Mining K-Nearest Neighbor (Knn) Dan Naive Bayes Untuk Klasifikasi Keluhan Masyarakat Terhadap Solusi Penanganan Sampah

Peneliti : Rahyan Elena Mahatiara

NIM : 1908046003

Program Studi : Matematika

Saya memandang bahwa naskah skripsi tersebut sudah dapat diajukan kepada Fakultas Sains dan Teknologi UIN Walisongo Semarang untuk diajukan dalam Sidang Munaqasyah.

*Wassalamualaikum warrahmatullahi wabarakatuh*

Pembimbing I



**Ariska Kurnia R. M.Sc**

NIP. 198908112019032019

## NOTA DINAS

Semarang, 23 Juni 2023

Yth. Ketua Program Studi Matematika

Fakultas Sains dan Teknologi

UIN Walisongo Semarang

*Assalamualaikum warrahmatualli wabarakatuh*

Dengan ini diberitahukan bahwa saya telah melakukan bimbingan, arahan dan koreksi naskah skripsi dengan :

Judul : Analisis Kinerja Algoritma Data Mining K-Nearest Neighbor (Knn) Dan Naive Bayes Untuk Klasifikasi Keluhan Masyarakat Terhadap Solusi Penanganan Sampah

Peneliti : Rahyan Elena Mahatiara

NIM : 1908046003

Program Studi : Matematika

Saya memandang bahwa naskah skripsi tersebut sudah dapat diajukan kepada Fakultas Sains dan Teknologi UIN Walisongo Semarang untuk diajukan dalam Sidang Munaqasyah.

*Wassalamualaikum warrahmatullahi wabarakatuh*

Pembimbing II



**Eva Khoirun Nisa, M.Si**

NIP. 198701022019032010

## ABSTRAK

Judul : **ANALISIS KINERJA ALGORITMA DATA MINING K-NEAREST NEIGHBOR (KNN) DAN NAIVE BAYES UNTUK KLASIFIKASI KELUHAN MASYARAKAT TERHADAP SOLUSI PENANGANAN SAMPAH**

Peneliti : Rahyan Elena Mahatiara

NIM : 1908046003

Sampah merupakan masalah utama yang terdapat disetiap daerah, sehingga peneliti akan mengangkat sampah menjadi objek untuk penelitian. Dalam penelitian ini akan membandingkan metode yang terdapat dalam data mining yaitu *K-Nearest Neighbor* (KNN) dengan *Naive Bayes* pada aplikasi media sosial *twitter* dengan tujuan mengetahui klasifikasi keluhan masyarakat terhadap penanganan sampah dan membandingkan tingkat keakuratan pada kedua metode tersebut. Pengambilan data yang digunakan dalam penelitian ini menggunakan *software Rapid Miner* dengan *hashtag* sampah dengan skala waktu 05 Agustus hingga 22 Agustus 2022, data yang terambil dalam proses ini sejumlah 500 data. Setelah dilakukannya *crawling data* dilanjut dengan proses *cleaning data*, dalam proses ini data dibersihkan dari *noise* sehingga data berubah menjadi 149 data dengan data berupa *sentiment* positif sejumlah 70 data dan *sentiment* negatif sejumlah 79 data dan pembagian rasio *data* dilakukan oleh *split data* dengan perbandingan 80:20, dengan presentasi 80% data *training* dan 20% data *testing*, maka dengan itu data *training* memiliki *sentiment* positif sebanyak 52 data dan *sentiment* negatif sebanyak 67 data dan data *testing* memiliki *sentiment* positif sebanyak 18 data dan *sentiment* negatif sebanyak 12 data. Kemudian data tersebut diolah menggunakan 2 metode yaitu KNN dan *Naive Bayes* dengan bantuan *google colab* dengan bahasa pemrograman *python*,

berdasarkan analisis kata yang sering muncul pada *tweet* adalah “kotor” dan “bau” dan diperoleh data dengan menggunakan algoritma *Naive Bayes* memiliki tingkat keakurasian lebih baik dibanding dengan algoritma KNN yaitu memiliki tingkat keakurasian 93% sedangkan dengan menggunakan KNN memiliki tingkat keakurasi 83%.

**Kata kunci :** sampah, *Naive Bayes*, *K-Nearest Neighbor* (KNN)



## KATA PENGANTAR

Bismillahirrahmanirrahim

Puji syukur peneliti panjatkan kehadirat Allah SWT yang sudah memberikan rahmat serta karunianya sehingga peneliti dapat menyelesaikan skripsi dengan baik dan tepat waktu. Sholawat serta salam peneliti sanjungkan kepada Nabi Muhammad SAW, keluarga, sahabat beliau yang telah memberikan pencerahan dan pembelajaran bagi kita semua sehingga kita dapat merasakan nikmat iman dan islam atas kemuliaan ilmu dengan pengharapan semoga dapat memberi syafaat di hari kiamat kelak, Aamiin yaa Robbal 'Alamiin.

Penulisan skripsi ini yang berjudul **“ANALISIS KINERJA ALGORITMA DATA MINING K-NEAREST NEIGHBOR (KNN) DAN NAIVE BAYES UNTUK KLASIFIKASI KELUHAN MASYARAKAT TERHADAP SOLUSI PENANGANAN SAMPAH”** bertujuan agar memenuhi persyaratan memperoleh gelar sarjana dalam menyelesaikan pendidikan pada program studi S.1 Matematika Fakultas Sains dan Teknologi Universitas Islam Negeri Walisongo Semarang.

Pada kesempatan kali ini perkenankanlah peneliti untuk mengucapkan rasa terimakasih terhadap semua pihak yang telah membantu dalam penyusunan skripsi ini karena

dalam penulisan skripsi ini peneliti menemui halangan dan kesulitan, namun berkat bimbingan dan dukungan dari berbagai pihak, sehingga penulisan skripsi ini terselesaikan dengan baik. Ucapan berterima kasih ini peneliti sampaikan dengan segala kerendahan hati dan rasa hormat kepada :

1. Prof. Dr. Imam Taufiq, M.Ag, selaku Rektor UIN Walisongo Semarang beserta Wakil Rektor I,II dan III UIN Walisongo Semarang .
2. Dr.Ismail SM, M.Ag, selaku Dekan Fakultas Sains dan Teknologi UIN Walisongo Semarang, beserta Wakil Dekan I,II dan III Fakultas Sains dan Teknologi UIN Walisongo Semarang.
3. Emy Siswanah, M.Sc, selaku ketua Prodi Matematika Fakultas Sains dan Teknologi UIN Walisongo Semarang.
4. Ahmad Aunur Rohman M.Pd, selaku Sekretaris Prodi Matematika Fakultas Sains dan Teknologi UIN Walisongo Semarang.
5. Ariska Kurnia Rachmawati, M.Sc, selaku Dosen pembimbing I yang telah menyenggangkan waktu untuk memberikan bimbingan, dukungan, memotivasi serta memberi penjelasan sehingga skripsi ini dapat terselesaikan.
6. Eva Khoirun Nisa, M.Si, selaku Dosen pembimbing II yang telah menyenggangkan waktu untuk memberikan

- dukungan, semangat, memotivasi serta memberi penjelasan sehingga skripsi ini dapat terselesaikan.
7. Bapak/Ibu Dosen serta Staf Fakultas Sains dan Teknologi UIN Walisongo Semarang atas pelajaran ilmu dan moral serta pelayanan sewaktu masa kuliah hingga penyusunan skripsi ini selesai.
  8. Terspesial kedua orang tua peneliti tercinta bapak Suyanto dan ibu Budi Rahayu yang senantiasa mendidik dengan ikhlas, memberi dukungan penuh terhadap peneliti dan tak lelah mendoakan pengharapan yang tulus dan baik untuk anaknya.
  9. Kakak peneliti Rahyan Tata Widya Pitaloka yang tanpa henti memberikan semangat dan dukungan untuk meraih dan mewujudkan cita-cita.
  10. Keluarga besar peneliti yang selalu memberikan dukungan dan nasihat untuk menyelesaikan skripsi ini, semoga Allah memberikan balasan dalam setiap langkah kita, Aamiin yaa Robbalalamin.
  11. Teman-teman Program studi Matematika, terkhusus angkatan 2019 yang berjuang bersama kurang lebih 4 tahun dan selalu memberikan semangat dan motivasi, terutama pada sahabat peneliti Diah Ayu Anggraini.

12. Teman-teman KKN MIT 15 posko 25 yang bersama selama 45 hari selalu memberikan dukungan untuk cepat menyelesaikan skripsi ini.
13. Terkhusus untuk pendamping saya selama pembuatan skripsi ini Muhammad Rifa'i yang menjadi *support system* selama penulisan dan menjadi *mood booster* terbaik untuk dapat menyelesaikan skripsi ini.

Peneliti menyadari bahwa terdapat banyak kekurangan dalam penataan skripsi ini, maka dari itu peneliti sangat membutuhkan kritik dan saran dari pembaca sebagai bahan memperbaiki penulisan peneliti di kemudian hari. Kritik dan saran dapat dikirim melalui email [rahyan\\_1908046003@student.walisongo.ac.id](mailto:rahyan_1908046003@student.walisongo.ac.id).

Semarang, 21 Juni 2023

Rahyan Elena Mahatiara

NIM. 1908046003

## DAFTAR ISI

PERNYATAAN KEASLIAN NASKAH.....	iii
Halaman Pengesahan .....	iv
NOTA DINAS.....	v
NOTA DINAS.....	vi
ABSTRAK .....	vii
KATA PENGANTAR.....	ix
DAFTAR ISI.....	xiii
DAFTAR GAMBAR.....	xvi
DAFTAR LAMPIRAN .....	xviii
BAB I PENDAHULUAN.....	1
1. Latar Belakang Masalah .....	1
2. Rumusan Masalah.....	9
3. Tujuan Penelitian.....	10
4. Manfaat Penelitian.....	10
5. Batasan Masalah .....	11
BAB II KAJIAN PUSTAKA.....	12
1. <i>Data Mining</i> .....	12
2. Data .....	14
3. Persiapan Data .....	15

4.	Normalisasi Data.....	19
5.	Klasifikasi .....	20
6.	<i>Naive Bayes</i> .....	21
7.	<i>K-Nearest Neighbor</i> .....	30
8.	<i>Confusion Matrix</i> .....	31
9.	ROC ( <i>Receiver Operating Characteristic</i> ) .....	34
10.	AUC ( <i>Area Under Curve</i> ) .....	37
11.	Keluhan Masyarakat Solusi Penanganan Sampah di Indonesia.....	38
12.	Literatur Terdahulu.....	40
BAB III METODE PENELITIAN.....		45
1.	Jenis Penelitian.....	45
2.	Sumber Data Penelitian.....	45
3.	Variabel Penelitian .....	46
4.	Teknik Analisis Data.....	46
BAB IV HASIL DAN PEMBAHASAN.....		50
1.	Sumber Data.....	50
2.	Crawling Data.....	51
3.	<i>Cleaning Data</i> .....	53
4.	Pre-Processing Data .....	57
5.	<i>Split Data</i> .....	61
6.	<i>Normalization Data dan SMOTE</i> .....	62

7. Modelling.....	64
8. Evaluasi.....	69
BAB V KESIMPULAN DAN SARAN .....	72
1. Kesimpulan.....	72
2. Saran .....	74
DAFTAR PUSTAKA.....	75
LAMPIRAN.....	108
DAFTAR RIWAYAT HIDUP .....	132

## DAFTAR GAMBAR

Gambar 2. 1 kurva ROC (D. Gunawan et al., 2022) .....	36
Gambar 2. 2 Kurva AUC (Latifa, 2018).....	38
Gambar 3. 1 Alur penelitian .....	47
Gambar 4. 1 Tampilan pencarian twitter dengan kata kunci sampah .....	50
Gambar 4. 2 fitur “ <i>search twitter</i> ” pada Rapid Miner ....	51
Gambar 4. 3 Hasil atribut dari <i>crawling</i> data .....	52
Gambar 4. 4 Hasil setelah pemilihan atribut.....	53
Gambar 4. 5 Import data ke <i>google colab</i> .....	54
Gambar 4. 6 <i>output</i> fitur <i>lowercase</i> .....	54
Gambar 4. 7 hasil output removal annotation.....	55
Gambar 4. 8 hasil <i>output tokenizing</i> .....	56
Gambar 4. 9 hasil <i>output stemming</i> .....	57
Gambar 4. 10 pelabelan pada <i>sentiment</i> .....	57
Gambar 4. 11 Pemisahan antara teks dengan label .....	58
Gambar 4. 12 <i>output</i> TF-IDF .....	59
Gambar 4. 13 pengurutan nilai terhadap kata .....	60
Gambar 4. 14 penyeleksian kata dengan pemilihan fitur sebesar 100 .....	61
Gambar 4. 15 <i>output proses split data</i> .....	62
Gambar 4. 16 <i>plot</i> hasil proses SMOTE .....	63
Gambar 4. 17 skala nilai k.....	64
Gambar 4. 18 penentuan nilai k.....	65
Gambar 4. 19 <i>confusion matrix</i> KNN .....	65
Gambar 4. 20 perhitungan akurasi menggunakan google colab.....	66
Gambar 4. 21 <i>confusion matrix Naive Bayes</i> .....	68



Gambar 4. 22 <i>output</i> akurasi menggunakan google colab .....	69
Gambar 4. 23 <i>output</i> kurva ROC & AUC KNN .....	70
Gambar 4. 24 <i>output</i> kurva ROC & AUC KNN .....	71

## DAFTAR LAMPIRAN

Lampiran 1 Hasil <i>crawling data</i> .....	108
Lampiran 2 Hasil dari pemilihan atribut kelas .....	110
Lampiran 3 Hasil data sesudah melalui proses <i>cleaning data</i> .....	112
Lampiran 4 Hasil data setelah melakukan proses <i>pre-processing data</i> .....	113
Lampiran 5 <i>script cleaning data</i> .....	115
Lampiran 6 proses <i>pre-processing data</i> dan <i>modelling</i> .....	119
Lampiran 7 panduan <i>script naive bayes</i> .....	131
Lampiran 8 panduan script K-Nearest Neighbor .....	131

# BAB I

## PENDAHULUAN

### 1. Latar Belakang Masalah

Statistika dan statistik memiliki karakteristik yang berbeda. Statistik mempunyai sifat sebagai data yang bisa digunakan untuk menyatakan kumpulan fakta, data statistik berupa bilangan yang disusun dalam tabel atau diagram yang menggambarkan suatu permasalahan (Fitriatien, 2017). Statistika menurut Mustafidah & Giarto (2021) merupakan suatu bidang ilmu pengetahuan yang membahas tentang proses pengumpulan data, proses pengolahan data, proses analisis data, penarikan kesimpulan, dan pembuatan kebijakan atau keputusan yang cukup kuat alasannya berdasarkan data dan fakta yang benar.

Statistik dengan data mining memiliki hubungan keterikatan sebagaimana di dalam teknik pengolahan data mining terdapat proses menganalisis dan rekognisi informasi dari berbagai *database* besar yang bermanfaat dengan memanfaatkan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* (Mardi, 2017). Data mining merupakan tahap penting dalam proses

perancangan pengetahuan dalam *database* yang menghasilkan pola atau model yang bermanfaat dari data. Istilah *database* berbeda dengan data mining, *database* lebih mengutamakan pada keseluruhan proses menemukan pengetahuan yang berguna dari data, sedangkan data mining mengacu untuk menemukan pola baru dari banyak data dalam *database* dengan berfokus pada algoritma untuk mengekstrak pengetahuan yang berguna (Silwattananusarn & KulthidaTuamsuk, 2012).

Data mining dapat dikelompokkan berdasarkan tugas yang akan dilakukan menurut buku yang berjudul "Discovering Knowledge in Data: *An Introduction to Data Mining*" yang ditulis oleh Larose (2014) yaitu deskripsi, estimasi, prediksi, klasifikasi, pengklasteran, asosiasi. Model klasifikasi merupakan model yang banyak digunakan dalam penelitian keakurasian untuk pengelompokkan dengan ciri variabel bersifat kategori, fungsi klasifikasi tersebut sesuai dengan sifat data yang akan diuji, sehingga model klasifikasi digunakan dalam penelitian ini. Sebagai contoh, penggolongan pendapat seseorang dapat diklasifikasikan dalam tiga kategori, yaitu pendapat positif, pendapat netral, dan pendapat negatif (Gunadi & Sensuse, 2016).

Algoritma klasifikasi sudah banyak dipakai untuk mengakomodasi dalam pengolahan data seperti *Random Forest*, *Decision Tree*, *Naïve Bayes*, *K-Means* dan juga *Support Vector Machine (SVM)* (Novianti et al., 2022). Metode di atas yang diaplikasikan dalam penelitian ini adalah perbandingan keakurasian mengenai metode *Naive Bayes* dan *K-Nearest Neighbor (KNN)*, Pengklasifikasian *Naive Bayes* termasuk dalam keluarga pengklasifikasian *probabilistik* berdasarkan teori *bayes*. Fitur utama dari klasifikasi ini adalah asumsi bahwa semua variabel independen bersyarat yang menjadi alasan untuk menyebutkan *naive*. Parameter klasifikasi *Naive Bayes* dapat dipelajari secara terpisah, lebih sederhana dan lebih cepat (Ilić et al., 2022) dan memiliki keakuratan lebih baik dibanding metode *data mining* lainnya (Hozairi; et al., 2021).

Pengklasifikasian *K-Nearest Neighbor (KNN)* merupakan metode *data mining* yang cara kerjanya dengan mencari kelompok k objek berdasarkan data *training*. Keakuratan algoritma KNN ini sangat dipengaruhi oleh ada atau tidaknya karakteristik yang tidak berhubungan, atau ketika bobot fitur tersebut tidak sesuai dengan relevansinya terhadap klasifikasi (Nuqoba & Djunaidy, 2014). Algoritma KNN mempunyai beberapa

kelebihan yaitu pelatihan sangat cepat, sederhana dan mudah dipelajari, tahan terhadap data *training* yang memiliki tidak beraturan, dan efektif jika data *training* besar. Namun algoritma ini juga mempunyai kekurangan yaitu nilai  $k$  bias, komputasi kompleks, *limit* memori, dan mudah tertipu dengan atribut yang tidak relevan (Rosso, 2019).

Literatur terdahulu mengenai perbandingan algoritma *data mining* *Naive Bayes* dan KNN yaitu penelitian yang pertama diteliti oleh penelitian pertama dari Rachmawati et al., (2022) yang berjudul *Comparison and Prediction of Data Mining Models to Determine the Classification of Family Planning Program User Status*. Penelitian tersebut membandingkan dari 4 metode yang terdapat dalam *data mining* yaitu *Decision Tree*, *Naive Bayes*, *Logistic Regression* dan *Gradient Boosted Trees*, dalam penelitian tersebut membahas hasil survei aparat Desa Mangunharjo tahun 2020 mengenai keberhasilan program KB dengan tujuan mengklasifikasikan status pengguna KB pada daerah tersebut. Hasil dalam penelitian tersebut dilihat dari hasil uji t dan kurva AUC, dalam penelitian tersebut menunjukkan bahwa metode *Decision Tree* lebih unggul dari 3 metode lainnya dengan perolehan akurasi mencapai 92,4% dan nilai AUC 0,939.

Penelitian yang kedua ditulis oleh Syarifuddin (2020) yang berjudul Analisis Sentimen Opini Publik Mengenai Covid-19 Pada *Twitter* Menggunakan Metode *Naive Bayes* dan KNN, yang menjelaskan tentang penyelarasan dan pandangan baru mengenai suatu isu dalam *twitter* yang memiliki kecenderungan opini masyarakat terhadap *twitter* condong positif dengan klasifikasi metode *Naive Bayes* lebih akurat dengan nilai akurasi 63,21% daripada klasifikasi KNN dengan nilai akurasi 58,94%.

Literatur ketiga dari Tempola et al. (2018) dengan judul Perbandingan Klasifikasi Antara KNN dan *Naive Bayes* Pada Penentuan Status Gunung Berapi Dengan *K-Fold Cross Validation*, yang membahas tentang membandingkan metode KNN dan *Naive Bayes* pada data aktivitas status gunung berapi yang terdapat di Indonesia dan memperoleh rata-rata akurasi sistem ketika memakai KNN sebesar 63,68% dan *standar deviasi* sebesar 7,47%, sedangkan ketika diterapkan memakai *Naive Bayes classifier* dihasilkan rata-rata akurasi sistem sebesar 79,71% dan *standar deviasi* sebesar 3,55%, dengan kesimpulan dengan menerapkan *Naive Bayes classifier* akurasi sistem lebih unggul dibanding KNN.

Oleh karena itu peneliti akan menguji apabila metode *Naive Bayes* menggunakan cara *split validation* untuk penentuan akurasi tetap akan lebih unggul dari pada KNN dengan objek yang dipakai dalam penelitian ini adalah pembahasan mengenai solusi penanganan sampah karena dalam observasi yang dilakukan peneliti penanganan sampah masih banyak kelalaian dan proses penanganan yang lambat. Permasalahan sampah saat ini menjadi suatu hal yang membutuhkan perhatian khusus karena sampah-sampah yang dibiarkan saja akan menimbulkan dampak negatif bagi lingkungan (Axmalia & Asti Mulasari, 2020). Sampah merupakan suatu permasalahan lingkungan yang sangat besar dalam segala aspek manapun karena mengakibatkan dampak masalah pada ekonomi, politik dan sosial budaya.

Adanya Tempat Pembuangan Akhir (TPA) pada suatu tempat yang letaknya kurang dari 1 kilometer menimbulkan berbagai masalah bagi masyarakat, sesuai dengan fakta lapangan di TPA daerah Kendal, nyatanya peraturan pemerintah nomor 18 tahun 2012 bahwa jarak pemukiman dengan tempat pembuangan akhir harus lebih dari 1 kilometer. Dampak yang terjadi apabila sampah di TPA tidak dikelola dengan baik maka akan menyebabkan pencemaran yang ditimbulkan dari



tumpukan sampah tersebut (Axmalia & Asti Mulasari, 2020). Maka dari itu penanggulangan terhadap pengelolaan sampah yang baik dan benar oleh pengangkut sampah merupakan salah satu usaha dalam mengurangi pencemaran lingkungan dan timbulnya penyakit bagi masyarakat. (Fajariani et al., 2022). Mengubah sampah menjadi suatu hal yang bernilai positif membutuhkan penanganan yang efisien, strategis dan cerdas (Wahyudi & Nawafilaty, 2020),

Informasi yang diperoleh dari Kementerian Lingkungan Hidup dan Kehutanan Direktorat Jenderal Pengelolaan Sampah, Limbah dan B3 Direktorat Penanganan Sampah, sampah di Indonesia mencapai 26.443.235,59 ton per tahun masalah ini dikarenakan kurangnya kesadaran masyarakat terhadap lingkungan namun permasalahan sudah mengakibatkan korban kehilangan nyawa. Sebagaimana tercantum dalam Q.S. Ar-Rum ayat 41 :

ظَهَرَ الْفَسَادُ فِي الْبَرِّ وَالْبَحْرِ بِمَا كَسَبَتْ أَيْدِي النَّاسِ لِيُذِيقَهُمْ بَعْضَ الَّذِي عَمِلُوا  
لَعَلَّهُمْ يَرْجِعُونَ

Artinya : Telah tampak kerusakan di darat dan di laut disebabkan karena perbuatan tangan manusia; Allah menghendaki agar mereka merasakan sebagian dari

(akibat) perbuatan mereka, agar mereka kembali (ke jalan yang benar).

Ayat tersebut dijelaskan dalam tafsir Al-Mu'tabar dengan menegaskan bahwa Allah SWT menghendaki agar perbuatan-perbuatan manusia yang tidak bertanggung jawab yang telah merusak lingkungan di muka bumi yang telah diciptakanNya mendapatkan apa yang telah mereka perbuat agar mereka kembali ke jalan yang benar dengan menjaga kesesuaian perilakunya dengan fitrahnya (Permata, 2022), dalam ayat tersebut dapat menjadi dalil tentang kewajiban melestarikan lingkungan hidup bagi manusia yang mempunyai akal pikiran, sebab terjadinya berbagai macam bencana juga karena adanya ulah manusia yang merusak alam tanpa diimbangi dengan upaya pelestarian (Santoso, 2014).

Penanganan sampah di Indonesia sering dikaitkan dengan media sosial sebagai solusi penanganan sampah yang dapat dilihat melalui bentuk dukungan saling bekerja sama mengelola sampah (Suryani, 2014). Media sosial yang dapat digunakan untuk analisis data salah satunya adalah *twitter*, hal itu dikarenakan dalam data *tweet* merupakan contoh sumber data *real-time* (Mualana & Redjeki, 2016). Sumber data yang akan

diambil dalam penelitian adalah data yang berasal dari aplikasi *twitter* tentang solusi penanganan sampah di Indonesia. Aplikasi *twitter* merupakan aplikasi media sosial yang sering digunakan masyarakat dalam menceritakan keluhan kesah kehidupan sehari-hari (Girnanfa & Susilo, 2022). Sehingga pengambilan data melalui aplikasi *twitter* dinilai sangat efektif untuk mengambil data keluhan masyarakat terhadap penanganan sampah.

Penelitian yang akan dilaksanakan berdasarkan latar belakang diatas adalah Analisis Kinerja Algoritma Data Mining K-Nearest Neighbor Naive Bayes Untuk Klasifikasi Keluhan Masyarakat Terhadap Solusi Penanganan Sampah.

## **2. Rumusan Masalah**

Berdasarkan latar belakang yang telah diuraikan sebelumnya, maka peneliti merumuskan permasalahan yang akan dikaji sebagai berikut :

- a. Bagaimana klasifikasi keluhan masyarakat terhadap penanganan sampah dari aplikasi *twitter* menggunakan metode *K-Nearest Neighbor*?

- b. Bagaimana klasifikasi keluhan masyarakat terhadap penanganan sampah dari aplikasi *twitter* menggunakan metode *Naive Bayes*?
- c. Bagaimana perbandingan dari kedua metode KNN dan *Naive Bayes* berdasarkan tingkat keakurasian untuk menganalisis keluhan masyarakat terhadap penanganan sampah dari aplikasi *twitter*?

### **3. Tujuan Penelitian**

- a. Untuk mengetahui bagaimana klasifikasi metode KNN terhadap penanganan sampah dari aplikasi *twitter*
- b. Untuk mengetahui bagaimana klasifikasi metode *Naive Bayes* terhadap penanganan sampah dari aplikasi *twitter*
- c. Untuk membandingkan tingkat keakurasian metode KNN dan metode *Naive Bayes* terhadap penanganan sampah dari aplikasi *twitter*.

### **4. Manfaat Penelitian**

- a. Memberikan pengetahuan mengenai analisis data mining menggunakan metode KNN terhadap solusi penanganan sampah kepada pembaca.
- b. Memberikan pengetahuan mengenai analisis data mining menggunakan metode *Naive Bayes*

terhadap solusi penanganan sampah kepada pembaca.

- c. Memberikan tujuan diskusi komunitas data *science* dan pihak-pihak lain yang berkaitan dari hasil output dari penelitian.
- d. Rujukan penelitian bagi peneliti data mining.

## 5. Batasan Masalah

Luasnya cangkupan penulisan, maka peneliti membatasi ruang lingkup pembahasan agar peneliti lebih fokus pada pembahasan yang ada

- a. Menggunakan data training hasil *crawling* twitter dengan tweet yang merujuk pada *hashtag* tentang sampah pada tanggal 5 sampai 25 Agustus 2022.
- b. Pengambilan data diambil dengan *tweet* berbahasa Indonesia.

## BAB II

### KAJIAN PUSTAKA

#### 1. *Data Mining*

*Data mining* mulai digunakan pada tahun 1990-an untuk membantu menentukan analisis tanpa adanya data dan hipotesis awal data *fishing* atau pengerukan data yang telah digunakan sebelumnya. Teknik yang terdapat dalam *Data mining* adalah pengumpulan, ekstraksi, analisis dan metode statistik, istilah data mining disebut juga dengan *knowledge discovery* (Osman, 2019).

Istilah dalam *data mining* diantaranya adalah *knowledge discovery* atau *pattern recognition*, istilah *knowledge discovery* mengacu pada penemuan pengetahuan yang tepat yang masih tersembunyi di dalam *data base*, *pattern recognition* lebih mengacu pada pengenalan pola yang tepat yang masih tersembunyi di dalam *data base* (Nabila et al., 2021). *Data mining* merupakan sebuah langkah yang melibatkan satu atau lebih teknik pembelajaran komputer (*machine learning*) dengan tujuan menjabarkan dan mengekstraksi pengetahuan (*knowledge*) secara otomatis (Syukri Mustafa et al., 2018).

*Data mining* merupakan ilmu yang saling berkaitan dengan ilmu lain diantaranya adalah *database system*, *data warehousing*, statistik, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi dan *data mining* juga didukung dengan ilmu lain seperti *neural network*, pengenalan pola, *spatial data analysis*, *image database*, *signal processing* (Meilani & Susanti, 2016).

Data yang terdapat pada *data mining* merupakan data yang berjumlah atau berkapasitas besar, *data mining* berfungsi sebagai teknik penelusuran data untuk membangun sebuah model menjadi informasi yang bermanfaat (Manurung & Hasugian, 2019) dan untuk mendapatkan informasi aktual dengan cara menentukan pola atau aturan tertentu yang berasal dari data (Farid et al., 2022). Pengelompokan yang dilakukan dalam data memiliki fungsi yaitu untuk mengerti pola secara *universal* yang berasal dari data yang bertujuan untuk dilakukannya proses selanjutnya yang berguna sebagai pendukung dan tujuan akhir tertentu (Utomo & Purba, 2019). Sumber data yang diambil dalam proses *data mining* dapat berupa *data base*, *warehouse*, Web, *repositori* dan informasi lainnya, atau dari data yang disambungkan ke sistem secara dinamis (Reza Noviansyah et al., 2018).

## 2. Data

Data adalah kumpulan beberapa fakta didasari pengukuran. Pengambilan keputusan yang sempurna apabila hasil dari penarikan kesimpulan berdasarkan pada pengamatan data dan fakta yang akurat, dapat disimpulkan bahwa data merupakan kejadian yang menggambarkan kenyataan atau bentuk data yang belum tercampur apapun dan dilanjutkan dengan beberapa proses agar mendapatkan informasi (Putry & Sari, 2022).

Data dibagi menjadi dua menurut jenis variabelnya yaitu data numerik dan data kategori. Data numerik adalah data yang bersifat kuantitatif yang nilainya berbentuk angka, sedangkan data kategorik merupakan sekumpulan kategori dan setiap nilai mewakili beberapa kategorik dengan kata lain data kategorik adalah data kualitatif yang berbentuk tidak beraturan(Putri et al., 2022)

Algoritma umumnya memiliki dua data diantaranya adalah :

### a. *Data Training*

*Data training* berfungsi untuk menciptakan model (Nasution et al., 2019), data yang dipakai yaitu



data yang sudah melalui proses *labeling* secara manual oleh peneliti yang akan dibedakan makna dari perkata dan makna yang didapat dipakai untuk masukkan dalam proses *training* (Rahman Isnain et al., 2021), atau dapat disimpulkan bahwa data *training* merupakan pembentukan fungsi atau pemetaan  $y = f(x)$ , dimana  $y$  adalah data *testing* sedangkan  $x$  adalah *data testing*.

b. *Data Testing*

*Data testing* merupakan sebuah pengujian dalam bentuk yang telah dibentuk dengan data lainnya dengan tujuan untuk mengetahui tingkat akurasi dari model tersebut (Nasution et al., 2019)

### 3. **Persiapan Data**

Proses persiapan data memerlukan adanya proses pembersihan atau data *cleaning*, data *integration* dan data *reduction*.

- a. data *cleaning* berfungsi untuk menghapus *noise* dan untuk mengganti data yang tidak terorganisasi menjadi data yang terorganisasi atau dengan kata lain mengganti teks menjadi *term indeks* yang akan diproses dengan beberapa langkah-langkah, yaitu :

- 1) *Transform Cases* : proses ini berfungsi menyalurkan semua teks menjadi huruf kecil karena untuk menjauhi adanya masalah pada saat *tokenizing*.
  - 2) *Removal annotation* : proses ini berfungsi menghilangkan *annotation* yang sering ditemui pada *tweet*, proses ini dilakukan karena *annotation* merupakan *noise* dan tidak mempunyai arti.
  - 3) *Tokenizing* : proses ini berfungsi membagi kalimat menjadi beberapa kata yang sering disebut dengan token.
  - 4) *Filtering* : proses ini berfungsi mendapatkan kata yang mempunyai arti dan menghapus kata yang tidak mempunyai arti dari hasil proses *tokenizing*.
  - 5) *Stemming* : proses ini berfungsi menghapus kata imbuhan dari hasil sebelumnya hingga menjadi bentuk kata dasar, dengan kata lain proses *stemming* adalah sebuah langkah pencarian kata dasar yang berasal dari kata *derivative* (Garbian Nugroho et al., 2016)
- b. *Pre-processing data* adalah suatu proses yang terdapat dalam *data mining* yang prosesnya adalah mengolah data mentah dengan cara mengeliminasi

data yang tidak sesuai dengan *machine learning* (Luqyana et al., 2018). *Pre-processing data* memiliki beberapa tahap yaitu :

- 1) *Feature engineering* : proses yang terdapat dalam *machine learning* yang bertujuan untuk mengekstraksi fitur dari data mentah menjadi sebuah model prediktif yang bertujuan untuk meningkatkan performa model(Nurdin et al., 2020).
- 2) *Feature extrcation* : suatu proses mengubah suatu *tweet* menjadi suatu nilai yang berbentuk *vector* agar *tweet* dapat diklasifikasi ke dalam kelas-kelas yang telah dibuat. *Feature extracktion* yang digunakan penulis adalah TF-IDF dengan karakteristik pembobotan setiap kata yang terdapat dalam *sentiment* dengan persamaan

$$W_{ij} = tf_{ij} \times \log\left(\frac{D}{df_j}\right) \quad (2.1)$$

Dimana :

$W_{ij}$  = bobot *term* ( $t_j$ ) terhadap dokumen ( $d_i$ )

$tf_{ij}$  = jumlah kemunculan *term* ( $t_j$ ) dalam dokumen ( $d_i$ )

$D$  = jumlah semua dokumen yang ada dalam data

*df* = banyaknya kata yang dicari pada sebuah dokumen (Fitriyani & Arifin, 2020)

- 3) *Feature selection* : suatu langkah yang dapat memaksimalkan kinerja *clasifier* dengan mengurangi atribut yang kurang tepat yang berfungsi untuk meningkatkan akurasi (Somantri & Apriliani, 2018).

Persiapan data dapat menggunakan dua metode yaitu *cross validation* dan *split validation*. *K-fold validation* merupakan sebuah metode yang membagi semua ruang sampel ke dalam sekelompok *k dataset*, dimana *dataset* mempunyai probabilitas untuk menjadi data *testing* (Rhomadhona & Permadi, 2019). Namun pada penelitian ini menggunakan metode *split validation*, metode *split validation* merupakan sebuah metode yang seluruh data *record* dan atribut digunakan dengan tujuan akan memiliki data set yang sesuai dengan asumsi-asumsi yang telah ditentukan. Data set yang akan ditentukan dibagi menjadi dua yaitu data *training* dan data *testing* dengan perbandingan dapat menggunakan 60:40, 70:30, 75:25, 80:20, 90:10, dsb. (Rahman et al., 2018). Setiap pengolahan data mining memiliki *split validation* yang berbeda-beda untuk memiliki hasil yang akurat (Witten et al., 2008).

#### 4. Normalisasi Data

Proses normalisasi data merupakan tahap yang dilakukan sebelum ke pemodelan *data mining*, proses *normalization* atau normalisasi adalah sebuah tahap yang menskalakan nilai atribut dari data pada rentang tertentu (Nasution et al., 2019).

Proses normalisasi data diperlukan dikarenakan sering sekali pada dataset terdapat rentang nilai yang berbeda – beda pada setiap atribut. Perbedaan rentang nilai yang cukup jauh dari atribut-atribut yang ada menyebabkan tidak berfungsinya secara optimal peranan atribut pada dataset. Maka sekiranya perlu dilakukan perubahan data berupa normalisasi data dengan menggunakan beberapa cara seperti *min- max* dan *cosine normalization* (Whendasmoro & Joseph, 2022).

- a. *Min-max normalization* adalah sebuah metode normalisasi yang prosesnya dengan melakukan tranformasi linier terhadap data mentah sehingga mendapatkan keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses. Metode *min-max normalization* dapat menggunakan persamaan sebagai berikut

$$\text{normalized } x = \frac{\text{minRange} + x - \text{minValue} - \text{minRange}}{\text{maxValue} - \text{minValue}} \quad (2.2)$$

- b. *Z-score normalization* adalah proses normalisasi berdasarkan mean (nilai rata-rata) dan standar deviasi dari data, proses ini berfungsi apabila tidak diketahui nilai aktual minimum dan maksimum data. Proses *Z-score normalization* dapat menggunakan persamaan sebagai berikut

$$\text{nilai baru} = \frac{\text{nilai lama} - \text{mean}}{\text{standar deviasi}} \quad (2.3)$$

(Nasution et al., 2019).

## 5. Klasifikasi

Klasifikasi merupakan salah satu metode yang terdapat dalam *data mining*, Klasifikasi merupakan proses menelaah suatu objek data untuk menentukan kelas tertentu dalam data dari seluruh kelas yang tersedia (Ariyanti & Iswardani, 2020), dengan tujuan guna meramalkan kelas dari suatu objek yang belum diketahui sebelumnya (Nasution et al., 2019) dan memiliki target variabel kategori (Reza Noviansyah et al., 2018), metode klasifikasi dapat disebut dengan metode *supervised learning* karena didalam klasifikasi memerlukan suatu pembelajaran data sebelumnya guna memperoleh hasil data baru (Ayudhitama & Pujiyanto, 2020) dan memiliki tujuan untuk mencari hubungan

antara data *training* dengan data *testing* (Hendrian, 2018).

Cara kerja teknik klasifikasi adalah mengenali suatu objek data untuk mengelompokkan ke dalam kelas berdasarkan kesamaan karakteristiknya untuk menguji ketepatan hasil yang akan diperoleh dari data atau dapat dikenali dengan data *testing*. Klasifikasi membuat model berdasarkan *data training* yang ada kemudian memakai model tersebut untuk pengklasifikasian pada data baru (Utomo & Mesran, 2020) Setelah dilakukan pembuatan model dan penerapannya, maka harus terdapat adanya proses evaluasi guna mengetahui tingkat akurasi dari pembangunan dan penerapan model berdasarkan data baru (Nasution et al., 2019). Variabel yang digunakan dalam proses klasifikasi adalah variabel dengan ciri kategori (Ardiansyah & Walim, 2018).

## 6. *Naive Bayes*

Sebelum kita menjelaskan lebih lanjut tentang Naive Bayes, terlebih dahulu akan diperkenalkan beberapa definisi dasar dan notasi tentang probabilitas sebagai erikut.

Misalkan kita notasikan  $S$  adalah ruang sampel dan  $A, B$  adalah subset dari  $\Omega$  yang kita sebut sebagai kejadian.

Berikut ini didefinisikan probabilitas dari suatu kejadian  $A$ .

### Definisi

Diberikan ruang sample  $S$  dan kejadian  $A$  subset dari  $S$ . Probabilitas dari  $A$  didefinisikan sebagai sebuah fungsi dari angka yang menyatakan seberapa banyak kemungkinan kejadian  $A$  bisa terjadi dan biasanya dinotasikan dengan  $P(A)$  dan dirumuskan sebagai berikut:

$$P(A) = \frac{n(A)}{n(S)}$$
 dimana  $n(A)$  adalah banyaknya anggota dari kejadian  $A$  dan  $n(S)$  adalah banyaknya anggota yang ada diruang sampel  $S$ .

### Dengan

1.  $0 \leq P(A) \leq 1$
2.  $P(\emptyset) = 0$  dan  $P(S) = 1$ .
3. Jika  $A$  dan  $B$  adalah subset dari  $S$  dan  $A, B$  adalah dua kejadian yang saling independen maka  $P(A \cup B) = P(A) + P(B)$ .

Algoritma *Naive Bayes* merupakan algoritma yang didasari oleh Teorema *Bayes* pada pertengahan abad ke-18 dan dikemukakan oleh ilmuwan Inggris bernama Thomas Bayes, Algoritma *Naive Bayes* merupakan salah satu algoritma yang dipakai untuk proses klasifikasi *machine learning* (winata, 2022), dan algoritma *Naive Bayes* merupakan metode klasifikasi populer dan masuk



dalam sepuluh algoritma terbaik dalam data mining (Mukminin & Riana, 2017).

Algoritma ini berfungsi untuk memprediksi sebuah *probabilitas* di masa yang akan datang berdasarkan dengan data *training* sehingga disebut sebagai teorema *Bayes* dan mengasumsikan bahwa algoritma *Naive Bayes* memiliki atribut yang *independent* (Iskandar & Suprpto, 2013). Teorema *Bayes* juga merupakan pendekatan statistik yang mendasari pada pengidentifikasian dalam pola (*pattern recognition*) (Dhany & Izhari, 2019). Teorema *Bayes* adalah sebuah metode yang cocok dalam mesin pembelajaran sesuai data *training* dengan memakai probabilitas bersyarat sebagai mulanya.

Terlebih dahulu akan kita berikan definisi dari suatu probabilitas kejadian bersyarat sebagai berikut.

#### Definisi 2.4 (Probabilitas Bersyarat)

Diketahui ruang sampel  $S$  dan  $A, B \in S$ . Probabilitas kejadian  $B$  jika kejadian  $A$  terlebih dahulu terjadi dapat dinyatakan dengan

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (2.4)$$

Dimana  $P(B|A)$  = Probabilitas B terjadi jika diketahui A terjadi,  $P(A \cap B)$  = Probabilitas terjadinya kejadian A sekaligus kejadian B dan  $P(A)$  = Probabilitas kejadian A dengan  $P(B) \neq 0$ .

Metode ini juga dapat menghasilkan sebuah estimasi parameter menggunakan cara penggabungan informasi dari sampel dan informasi yang sudah didapat sebelumnya.

Teorema *Naive Bayes* diasumsikan bahwa kondisi antar atribut saling bebas sehingga ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lain.

Teorema 2.5 (persamaan *Bayes*). Misalkan  $A$  subset dari ruang sampel  $S$  dengan  $P(A) > 0$ . Kemudian untuk  $B$  subset  $S$ .

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \quad (\text{Han et al., 2012})$$

Bukti. Dari definisi 2.4 dipunyai

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Dimana pertidaksamaan kedua kembali menggunakan definisi 2.2 dan teorema 2.5

Catatan 2.6 seperti dalam hukum probabilitas total. Hal tersebut dapat diperluas dalam  $k$  peristiwa. Misalkan  $B_1, \dots, B_k$  subset  $\Omega$  untuk  $\Omega$  dan semua  $i \geq 1$ ,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^k P(A|B_i)P(B_i)}$$

Contoh 2.7 (kasus bayi kembar). Sekitar 0,3% dari semua kelahiran menghasilkan kasus kembar identik dan 0,7% dari semua kelahiran menghasilkan kembar fraternal. Kembar identik adalah kelahiran kembar dengan memiliki jenis kelamin yang sama, sedangkan kembar fraternal adalah kelahiran tidak identik atau kembar dengan memiliki jenis kelamin yang berbeda.

Ditunjukkan bahwa  $A$  merupakan angka kelahiran kembar identik dengan jenis kelamin perempuan dan ditunjukkan bahwa  $B$  merupakan angka kelahiran kembar fraternal, sehingga diketahui bahwa  $P(B) = 0,7, P(A|B) = 0,25, P(B^c) = 0,3$  dan  $P(A|B^c) = 0,5$ , oleh karena itu, dihasilkan

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^c)P(B^c)}$$

$$= \frac{(0,25)(0,7)}{(0,25)(0,7) + (0,5)(0,3)} = \frac{7}{13}$$

Sehingga diketahui apabila probabilitas terjadinya kelahiran angka kembar fraternal bila terjadi kelahiran kembar identik

Algoritma *Naive Bayes* merupakan algoritma dengan pengklasifikasian probabilitas yang sederhana dengan menghitung seluruh probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai yang berasal dari *dataset* yang diberikan berdasarkan dengan pengasumsian *independent* (Ling, 2022).

Kelebihan dalam menggunakan metode *Naive Bayes* adalah dalam melakukan pengolahan hanya memerlukan data *training* berjumlah kecil untuk menghitung parameter *mean* dan varian dari variabel yang digunakan sebagai klasifikasi (Lestari et al., 2022) dan memiliki kecepatan saat mengaplikasikan ke dalam *database* dengan data berjumlah besar (Muslehatin et al., 2017).

Langkah untuk menyelesaikan metode *Naive Bayes* yaitu :

- a. Memilah data *training* sesuai dengan *split data* atau *k-fold validation*
- b. Menghitung jumlah dan probabilitas pada data *training*
  - 1) Apabila terdapat data yang numerik, maka harus menemukan nilai *mean* dan standar deviasi dari masing-masing parameternya yang menyatakan data bernilai angka.

#### Definisi 2.6 (mean)

Mean adalah nilai rata-rata dari beberapa data

$$\mu = \sum_{i=1}^n \frac{x_i}{n} \text{ atau } \mu = \frac{x_1+x_2+x_3+x_n}{n} \quad (2.8)$$

Dimana  $\mu = \textit{mean}$  hitung,  $x_i =$  nilai sampel sampai ke- $i$  dan  $n =$  jumlah sampel. Dan untuk mencari standar deviasi dapat menggunakan persamaan 2.7 seperti dibawah ini :

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} \quad (2.9)$$

dimana :

$\sigma =$  standar deviasi

$\mu = \textit{mean}$  hitung

$x_i$  = nilai sampel sampai ke- $i$

$n$  = jumlah sampel

- 2) Apabila data tidak bernilai numerik maka nilai probabilitas dihitung setiap kategori yang sama dengan jumlah data dari kategori yang sama selanjutnya dibagi sesuai data pada kategori yang didapat.
- c. Menilai probabilitas fitur di setiap kelas dengan menghitung jumlah data yang terdapat pada kategori yang sama kemudian dibagi dengan jumlah data yang terdapat pada kategori tersebut.
- d. Proses selanjutnya adalah mencari nilai *probabilitas* untuk fitur data *testing* yang memiliki data bersifat *numerik* atau angka. Persamaan untuk mencari proses tersebut adalah dengan mencari nilai distribusi *Gaussian*.

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \times e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \quad (2.10)$$

Keterangan :

$x_i$  = nilai sampel kelas ke- $i$

$\sigma$  = standar deviasi

$\mu$  = *mean* hitung

- e. Menghitung probabilitas akhir untuk setiap kelas dengan menginputkan seluruh data nilai distribusi *gaussian* yang ada ke dalam satu kelas yang sama. Misalkan terdapat  $n$  kelas, yaitu  $V_1, V_2, \dots, V_N$ . Untuk sebuah *tuple* masukan  $X$ , *Naive Bayes classifier* memprediksi bahwa *tuple*  $X$  termasuk ke dalam kelas  $V_i$  jika dan hanya jika

$$P(X|Kelas) = P(V_1|Kelas) \times P(V_2|Kelas) \times P(V_3|Kelas) \times P(V_4|Kelas) \times P(V_n|Kelas) \quad (2.11)$$

- f. Langkah selanjutnya adalah menghitung probabilitas akhir dengan melalui perhitungan nilai probabilitas akhir kelas ke dalam rumus *Naive Bayes Classifier* seperti halnya persamaan nomer (2.5).
- g. Langkah terakhir yang dapat dilakukan dengan mencari nilai probabilitas satu kategori dengan jumlah nilai semua kategori.

$$P(Kelas) = \frac{P(Kelas|X)}{P(X|Kelas) + P(X|Kelas)} \quad (2.12)$$

Dimana :

$P(Kelas|X)$  = probabilitas kelas akan terjadi apabila  $X$  terjadi,  $P(X|Kelas)$  = probabilitas  $X$  terjadi apabila Kelas terjadi (Imandasari et al., 2019).

## 7. *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor* (KNN) termasuk kedalam metode yang memanfaatkan algoritma *supervised learning, supervised learning* memiliki tujuan untuk mendapatkan pola baru pada data dengan menghubungkan pola data sebelumnya dengan pola data yang terbaru (Liantoni, 2015). Algoritma KNN merupakan algoritma klasifikasi non parametrik yang sangat efisien. Metode KNN adalah sebuah metode algoritma yang memperhitungkan jarak objek yang saling berdekatan, jarak tersebut dapat ditentukan menggunakan *formula* jarak *Euclidean* (Zhang et al., 2022). Metode klasifikasi KNN memiliki tujuan untuk mengklasifikasikan sebuah objek baru menurut data *training* dan atribut (Limbong et al., 2022).

Langkah-langkah dalam menentukan jarak *euclidean* tersebut dapat dipaparkan sebagai berikut :

- a. Menentukan nilai  $k$  ( nilai tetangga terdekat) yang akan digunakan,  $k$  berguna untuk melakukan klasifikasi data baru, nilai  $k$  dipertimbangkan berdasarkan banyaknya data yang ada dan ukuran dimensi atau atribut yang dibentuk oleh data.



- b. Menghitung kedekatannya dengan persamaan jarak *Euclidean* terhadap data *training*

$$D(x, y) = \|x - y\|^2 = \sqrt{\sum_{j=1}^N |x - y|^2} \quad (2.13)$$

Dengan keterangan  $x$  adalah sampel data atau data *training* dan  $y$  adalah data uji atau data *testing*

- c. Mengurutkan hasil jarak yang telah dicari, dengan ketentuan dari nilai yang tertinggi ke terkecil.
- d. Menghitung jumlah pada kelas dengan menurut  $k$  terdekat.
- e. Kelas dengan hasil mayoritas akan menjadi kelas baru dari data *testing*. (Erdiansyah et al., 2022).

## 8. **Confusion Matrix**

*Confusion Matrix* adalah tabel yang memiliki 4 kombinasi yang berbeda dari nilai prediksi dan nilai aktual (Heydarian et al., 2022). *Confusion matrix* merupakan metode yang berfungsi untuk menghitung keakurasian pada *data mining*, proses yang dilakukan dalam metode ini adalah *recall*, *precision*, *accuracy* dan *rate galat*. Evaluasi model klasifikasi digunakan untuk pengujian perkiraan objek yang benar dan salah (Dewi, 2016).

*Confusion Matrix* bermuat informasi mengenai klasifikasi aktual dan prediksi dilakukan oleh klasifikasi sistem. Kinerja sistem seperti itu biasanya diperkirakan menggunakan data dalam matriks. Tabel berikut menunjukkan matriks *confusion* untuk classifier dua kelas (Santra & Christy, 2012).

Tabel 1: Tabel *Confusion Matrix*

		Nilai Aktual	
		Positive	Negative
Nilai Prediksi	Positive	TP	FP
	Negative	FN	TN

Sumber : (Han Jiawei et al., 2012)

Dimana TP (*true positive*) adalah nilai yang benar-benar positif dan diprediksi positif, FP (*false positive*) adalah nilai yang sebenarnya negatif tetapi diprediksi positif, FN (*false negative*) adalah nilai yang sebenarnya positif tetapi diprediksi negatif dan TN (*true negative*) adalah nilai yang sebenarnya negatif dan diprediksi negatif (Samsir et al., 2021).

Berdasarkan tabel 1 maka didapat definisi-definisi, sebagai berikut

Definisi 2.12 (akurasi)

Akurasi adalah ukuran atau presentase dalam menentukan tingkat keakurata data yang dihitung

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad (2.14)$$

Definisi 2.13 (*precision*)

*Precision* adalah kecocokan antara bagian data yang diambil dengan informasi yang dibutuhkan

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2.15)$$

Definisi 2.14 (*recall*)

Recall adalah penggambaran keberhasilan model dalam menemukan kembali sebuah informasi

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2.16)$$

Definisi 2.15 (*F-Measure*)

*F-Measure* adalah perhitungan perbandingan rata-rata *precision* dan *recall* yang dibobotkan

$$\text{F - measure} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (2.17)$$

Han et al., 2012)

## 9. ROC (*Receiver Operating Characteristic*)

Kurva ROC (*Receiver Operating Characteristic*) awalnya dikembangkan untuk menentukan antara sinyal (hasil positif benar) dan *noise* (hasil positif palsu) ketika menganalisis sinyal pada layar radar selama Perang Dunia II (Nahm, 2022). Korelasi analisis sensitivitas dan spesifisitas pengukuran ditampilkan dalam bentuk kurva yang disebut dengan kurva ROC (*Receiver Operating Characteristic*) dan Kurva ROC adalah kurva untuk mengevaluasi potensi ambang nilai *median* untuk setiap *dependent* (Cortez et al., 2022). Kurva ROC adalah metode analisis yang direpresentasikan sebagai grafik dan untuk memeriksa dan menafsirkan suatu hasil tes diagnostik dimana tes yang diterapkan harus diketahui (Ozdemir & Algin, 2022) dan perlu diklasifikasikan ke dalam salah satu kategori dikotomis yang jelas. Analisis ROC adalah tes penting yang bertujuan untuk menilai akurasi tes kuantitatif atau kinerja diskriminasi di seluruh jajaran variabel di bawah desain eksperimental (Susanti et al., 2022). Namun, karena banyak hasil tes disajikan sebagai variabel kontinu atau ordinal, nilai referensi (nilai batas) untuk diagnosis harus ditetapkan.

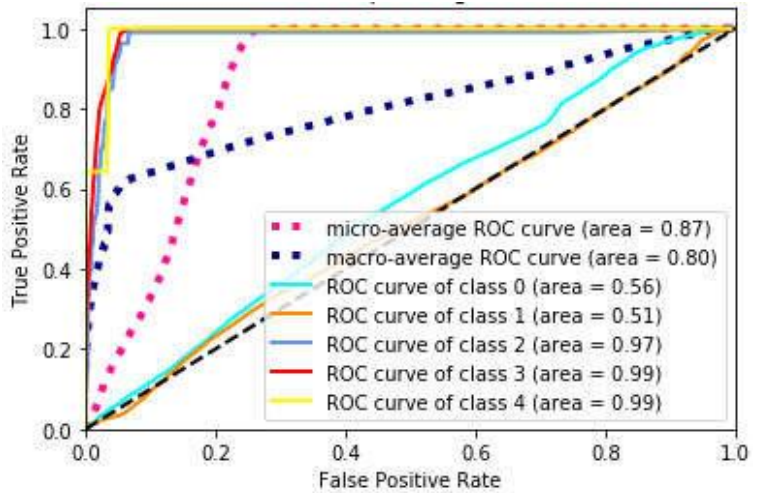
Lebih tepatnya kurva ROC digunakan untuk dapat menentukan kehadiran sesuatu berdasarkan nilai *cut-off* (Nahm, 2022). Diagnosis yang sebenarnya yaitu metode terbaik yang digunakan dan diterima untuk memberikan hasil yang akurat (Ozdemir & Algin, 2022) dan optimal dari skala tersebut (Naz et al., 2022). Kriteria untuk menentukan baik tidaknya tingkat akurasi dapat dilihat dari kecepatan, kehandalan, *skalabilitas* dan *interpretasi* (Dewi, 2016).

Kurva ROC melihat keakurasian dan mempertimbangkan klasifikasi secara nyata, ROC menunjukkan *confusion matrix*. Kurva ROC dilihat dari hasil nilai yang telah didapatkan pada perhitungan dengan *confusion matrix*, yaitu antara *false positive rate* dengan *true positive rate* (Wanika Siburian & Elvina Mulyana, 2019). ROC memiliki grafik dua dimensi dengan *false positive* untuk menjadi garis horizontal dan *true positive* menjadi garis vertikal.

Hasil penerapan kurva ROC dipresentasikan dengan kurva AUC. ROC mempunyai tingkat nilai diagnosa yaitu:

- a. Akurasi bernilai 0,91 - 1,00 = akurasi sangat baik
- b. Akurasi bernilai 0,81 - 0,90 = akurasi baik

- c. Akurasi bernilai  $0,71 - 0,60 =$  akurasi cukup
- d. Akurasi bernilai  $0,61 - 0,50 =$  akurasi kurang
- e. Akurasi bernilai  $0,51 - 0,40 =$  akurasi sangat kurang (Andriani, 2013).

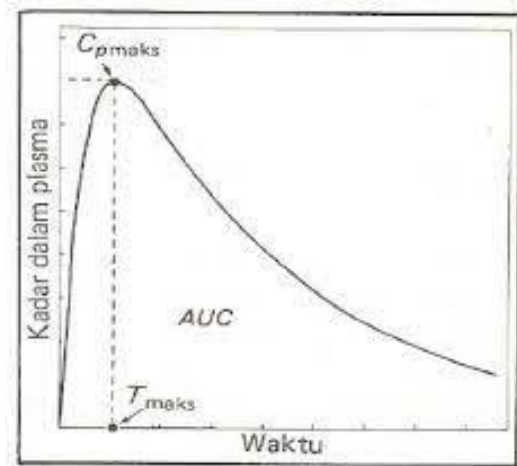


Gambar 2. 1 kurva ROC (D. Gunawan et al., 2022)

Kurva diatas berfungsi untuk mempresentasikan kinerja algoritma, algoritma klasifikasi yang bernilai bagus adalah kurva yang semakin mendekati sudut kiri atas, apabila kurva mendekati diagonal 45 derajat dari ruang ROC maka semakin tidak akurat klasifikasi tersebut (D. Gunawan et al., 2022).

## 10. AUC (*Area Under Curve*)

Kurva AUC (*Area Under Curve*) merupakan kurva yang cara kerjanya menghitung luasan dari kinerja rata-rata kinerja *classifier* (Yang et al., 2021) dan termasuk metode yang populer untuk mengevaluasi, memvisualisasikan, mengatur dan memilih pengklasifikasian berdasarkan kinerja mode (Mukminin & Riana, 2017) dan mengklasifikasikan dengan baik pada distribusi kelas seimbang dan tidak seimbang, AUC juga berfungsi untuk merangkum hubungan antara tingkat positif benar dan salah dari pengklasifikasi biner, untuk ambang keputusan yang berbeda (Brzezinski & Stefanowski, 2017). Untuk mempresentasikan baik buruknya hasil dari data yang diukur dapat dilihat dari *Area Under Curve* (AUC) yang terdapat pada kurva ROC yang sebelumnya didapat (Indrayuni, 2019).



Gambar 2. 2 Kurva AUC (Latifa, 2018)

## 11. Keluhan Masyarakat Solusi Penanganan Sampah di Indonesia

Sampah menurut Kamus Besar Bahasa Indonesia adalah suatu barang atau benda yang dibuang dikarenakan sudah tidak terpakai lagi dan sebagainya dan sampah menurut pasal 1 Undang-undang Nomor 18 Tahun 2008 tentang Pengelolaan Sampah yaitu sisa dari kegiatan sehari-hari manusia dan atau proses alam yang berbentuk padat (Ihsyaluddin & Mane, 2022).

Sampah merupakan masalah yang sering terjadi dan harus dihadapi oleh seluruh masyarakat,



tidak dapat dipungkiri setiap tahun volume sampah meningkat dengan diiringi oleh pertumbuhan manusia yang pesat dan meningkatnya ekonomi membuat manusia semakin menambah pola konsumerisme masyarakat (Suryani, 2014).

Jumlah penduduk Indonesia dari tahun ketahun mengalami kenaikan, jumlah penduduk Indonesia pada tahun 2022 mencapai angka 275.361.267 jiwa per Juni, meningkatnya jumlah penduduk di Indonesia mengakibatkan jumlah sampah di Indonesia juga meningkat, jumlah sampah yang dihasilkan tiap hari di Indonesia mencapai 175.000 ton perhari atau 0,7 kg per orang (Oktara, 2022).

Akibat yang ditimbulkan oleh sampah sangatlah beragam seperti membahayakan kesehatan, keselamatan, berkurangnya kenyamanan dan membuat keterbatasan lahan dan masih banyak lagi masalah yang ditimbulkan oleh sampah (Mahyudin, 2017), dampak yang timbul inilah yang membuat pemerintah menyerukan aksi penanganan persoalan sampah dan limbah searah pada diterapkannya *Sustainable Development Goals* tujuan 12.5, yaitu diharap pada tahun 2030 setiap negara akan

substansial harus menerapkan 3R yaitu *reuse*, *reduce* dan *recycle* pada produksi limbah agar bisa menjamin pola produksi dan pola konsumsi yang tidak terputus dan pemerintah berharap pada tahun 2025 sampah akan berkurang 30% dan sampah yang ditanggulangi harus mencapai angka 70% (Verawati & Verawati, 2022).

## **12. Literatur Terdahulu**

Penelitian mengenai metode *Naive Bayes* dan *K-Nearest Neighbor* (KNN) memang sudah banyak dilakukan salah diantaranya adalah penelitian pertama dari Rachmawati et al., (2022) yang berjudul *Comparison and Prediction of Data Mining Models to Determine the Classification of Family Planning Program User Status*. Penelitian tersebut membandingkan dari 4 metode yang terdapat dalam *data mining* yaitu *Decision Tree*, *Naive Bayes*, *Logistic Regression* dan *Gradient Boosted Trees*, dalam penelitian tersebut membahas hasil survei aparat Desa Mangunharjo tahun 2020 mengenai keberhasilan program KB dengan tujuan mengklasifikasikan status pengguna KB pada daerah tersebut. Hasil dalam penelitian tersebut dilihat dari

hasil uji t dan kurva AUC, dalam penelitian tersebut menunjukkan bahwa metode *Decision Tree* lebih unggul dari 3 metode lainnya dengan perolehan akurasi mencapai 92,4% dan nilai AUC 0,939.

Penelitian kedua dari Syarifuddinn (2020) yang berjudul Analisis Sentimen Opini Publik Mengenai Covid - 19 pada Twitter Menggunakan Metode *Naïve Bayes* dan KNN. Penelitian tersebut berfokus pada perbandingan hasil klasifikasi metode *Naive Bayes* dengan metode KNN dengan memanfaatkan 1098 opini dengan kata kunci Covid-19 yang diambil menggunakan API *twitter*. Hasil yang didapat dalam penelitian ini metode *Naive Bayes* memiliki nilai akurasi yang lebih tinggi dari pada metode KNN, dengan nilai akurasi sebesar 63,21% sedangkan nilai akurasi yang diperoleh menggunakan metode KNN adalah 58,1%.

Penelitian ketiga ditulis oleh Tempola et al. (2018) dengan judul Perbandingan Klasifikasi Antara KNN Dan *Naive Bayes* pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation, penelitian tersebut bertujuan untuk mengetahui tingkat keakurasian mengenai status gunung berapi yang

diklasifikasikan menggunakan 3 status yaitu normal, waspada dan siaga dengan menggunakan metode *Naive Bayes* dan KNN, dan untuk validasi data menggunakan *k-fold validation*. hasil yang diperoleh dalam penelitian tersebut ketika menggunakan KNN memiliki tingkat keakurasian sebesar 63,68% dan standar deviasi sebesar 7,47 sedangkan ketika menggunakan *Naive Bayes* memiliki tingkat keakurasian sebesar 79,91% dan standar deviasi sebesar 3,55%.

Penelitian terdahulu yang keempat adalah penelitian dari Kiran et al., (2018) yang berjudul *Credit card fraud detection using Naive Bayes model based and KNN classifier*, penelitian tersebut memiliki tujuan untuk meningkatkan ketepatan, akurasi dan meningkatkan fleksibilitas dari algoritma *Naive Bayes* dan KNN dengan objek dataset kartu kredit yang sama sehingga apabila terdapat transaksi penipuan dapat teridentifikasi dengan cepat dan terarah. Hasil dalam penelitian ini metode *Naive Bayes* memiliki keakuratan lebih baik untuk mendeteksi penipuan kartu kredit yaitu sebesar 95% sedangkan metode KNN memiliki tingkat keakuratan sebesar 90% dengan catatan apabila mendeteksi dengan

menggunakan satu algoritma saja maka penelitian tersebut tidak efisien karena setiap algoritma memiliki kelebihan dan kekurangannya masing-masing, sehingga untuk mendapatkan hasil akurat harus menggabungkan kedua metode tersebut agar memiliki hasil yang akurat dan efisien.

Penelitian terdahulu yang terakhir adalah penelitian dari Irfan et al. (2018) berjudul *Comparison of Naive Bayes and K-Nearest Neighbor methods to predict divorce issues*, pada penelitian irfan et al memiliki tujuan untuk memprediksi terjadinya perceraian dengan dua metode algoritma data mining untuk memperoleh hasil yang maksimal, efektif dan akurat dengan membandingkan metode algoritma *Naive Bayes* dan KNN, metode pengembangan perangkat lunak yang digunakan peneliti adalah dengan memanfaatkan model *prototype* karena memiliki tingkat ruang lingkup yang kecil dan merupakan penelitian baru. Hasil yang diperoleh dalam penelitian ini mengungkapkan bahwa algoritma *Naive Bayes* lebih unggul untuk tingkat keakurasiannya dari pada KNN dengan hasil 72,5% untuk algoritma *Naive Bayes* dan 57,5% untuk

algoritma KNN dengan menggunakan 20 data *testing*  
dan 130 data *training*.

## **BAB III**

### **METODE PENELITIAN**

#### **1. Jenis Penelitian**

Jenis pendekatan yang digunakan pada penelitian ini adalah pendekatan kuantitatif dengan membandingkan dua perlakuan dalam suatu parameter atau beberapa parameter dalam waktu bersamaan yaitu untuk mengetahui perbedaan tingkat akurasi yang dihasilkan dalam metode algoritma *Naive Bayes* dan *K Nearest Neighbor* (KNN) dengan penelitian yang bersifat deskriptif.

#### **2. Sumber Data Penelitian**

Sumber data yang diambil dalam penelitian bersifat data sekunder karena dalam penelitian menggunakan sumber data yang berasal dari *sentiment* di *twitter* dengan *hashtag* penanganan sampah di Indonesia. Untuk mendapatkan dataset peneliti memanfaatkan *software* RapidMiner yang telah terhubung dengan *twitter* untuk mengambil data dari tanggal 05 Agustus 2022 hingga 22 Agustus 2022.

### 3. Variabel Penelitian

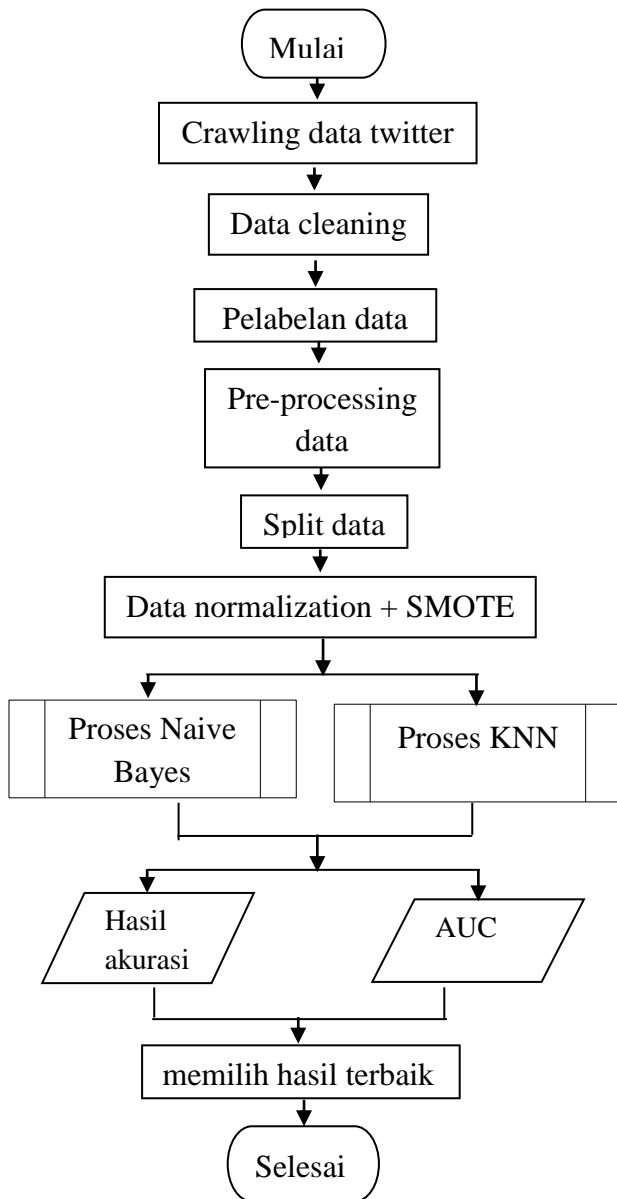
Pendekatan data mining yang dilakukan dalam penelitian ini adalah teknik klasifikasi. Penggolongan variabel dalam penelitian ini dibagi menjadi dua kelas yaitu:

- a. sentimen positif yaitu *sentiment* yang memiliki sifat membangun dan tidak memiliki unsur kebencian
- b. sentimen negatif yaitu *sentiment* yang memiliki unsur kebencian, mengkritik dan bersifat menjatuhkan. (B. Gunawan et al., 2018)

### 4. Teknik Analisis Data

Secara umum langkah-langkah yang terdapat dalam penelitian memiliki proses yang terstruktur dimulai dari *input* data hingga *output* data. Dalam penelitian akan digunakan *split data* sebesar 80:20 untuk perbandingan data *training* dan data *testing*. Gambaran umum penjelasan proses penelitian dijelaskan dalam gambar 3.1





Gambar 3. 1 Alur penelitian

Penjelasan mengenai alur penelitian diatas yaitu

1. *Crawling* data

*Crawling* data adalah sebuah istilah proses pengumpulan data pada *twitter*. Cara kerja tahap ini adalah mengunduh data berupa *user* atau *tweet* dari *server twitter* dengan memakai bantuan "*search twitter*" pada *software Rapid Miner*. Proses *crawling* data pada penelitian ini menggunakan #sampah dengan lingkup wilayah negara Indonesia, pada proses ini akan diberikan label *sentiment positive*, *sentiment negative* pada setiap *tweet* secara manual atau dapat dinamakan dengan *polarity*.

2. *Cleaning Data* yang terdiri dari berbagai proses seperti *transform cases*, *removal annotation*, *tokenizing*, *filtering*, *stemming*, proses ini dilakukan dengan menggunakan *google colab* dengan bahasa pemograman *python*.

3. Pelabelan

Proses pelabelan merupakan proses penilaian *sentiment* positif dan *sentiment* negatif terhadap teks.

4. *Pre-processing data* terdiri dari *feature engineering*, *feature extrcation*, *feature selection*.

5. *Split data*

Pada penelitian ini *split data* yang digunakan adalah perbandingan 80:20, dengan 80% untuk data *training*

dan 20% untuk data *testing*. Pembagian ini bertujuan untuk memperoleh hasil akurasi yang tinggi.

6. *Normalization data* dan SMOTE
  - a. Proses *normalization* atau normalisasi yang digunakan dalam penelitian adalah metode *min-max normalization* dengan perubahan nilai numerik dalam dataset ke skala 0 dan 1 dengan menerapkan persamaan 2.2
  - b. Proses SMOTE merupakan tahap *balancing* terhadap label *sentiment*, karena data *training* dan data *testing* pada penelitian ini tidak seimbang maka dilakukannya proses SMOTE untuk mendapatkan data yang baik dengan proses mengimbangi kelas yang diakibatkan oleh *overfitting*.
7. *Modelling data* yang digunakan dalam penelitian ini adalah metode algoritma *data mining* Naive Bayes dan KNN.
8. Evaluasi

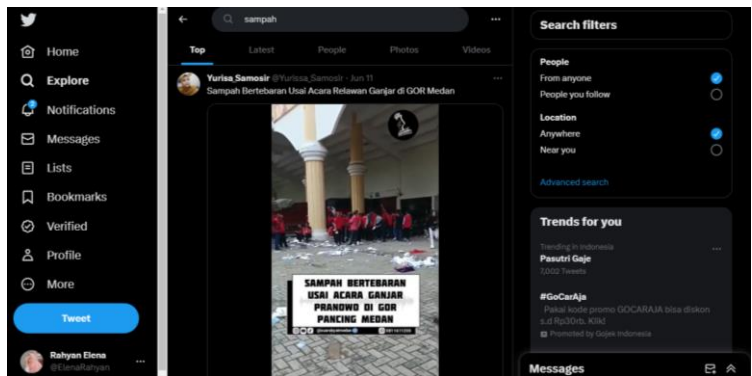
Evaluasi yang digunakan dalam pengujian data yaitu hasil perhitungan akurasi dalam perhitungan *confusion matrix*, kurva ROC, luasan AUC dengan menentukan klasifikasi *data mining* yang terbaik antara KNN dan *Naive Bayes*.

## BAB IV

### HASIL DAN PEMBAHASAN

#### 1. Sumber Data

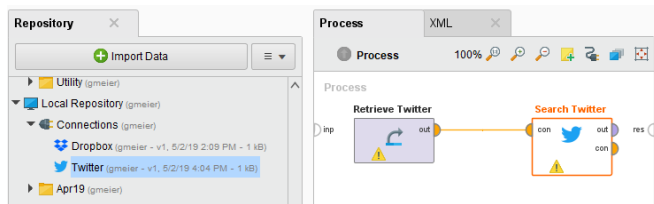
Sumber data dalam penelitian ini adalah data sekunder yang didapat melalui *sentiment-sentiment* masyarakat dengan hastag sampah yang diambil dalam bentuk teks yang merupakan *tweet* dari pengguna *twitter*. Data yang diambil dalam periode waktu 5-25 Agustus 2022, batasan dalam pencarian hastag sampah ini adalah dengan menggunakan bahasa Indonesia. Penelitian ini diperkuat dengan adanya jurnal terdahulu dan buku-buku pendukung.



Gambar 4. 1 Tampilan pencarian twitter dengan kata kunci sampah

## 2. Crawling Data

*Proses crawling* data dalam penelitian ini dengan melakukan pengambilan data dari *twitter* dengan menggunakan bantuan *software Rapid Miner* yang disimpan dalam bentuk *file csv*, dalam menggunakan bantuan *Rapid Miner* peneliti memanfaatkan fitur “*search twitter*” yang telah disambungkan pada akun *twitter* melalui koneksi *twitter* di *software Rapid Miner*.



Gambar 4. 2 fitur “*search twitter*” pada Rapid Miner

*Crawling* data yang diambil dalam penelitian berbentuk teks yang merupakan tweet dari pengguna *twitter*, hasil yang didapat dari *crawling* data berjumlah 500 *tweet*. Tahap *crawling* data menghasilkan 12 atribut yang berisi :

- a. *Created\_at* yaitu waktu yang dihasilkan saat pengguna mengunggah *tweet*.
- b. *From\_user* yaitu nama pengguna *twitter*.

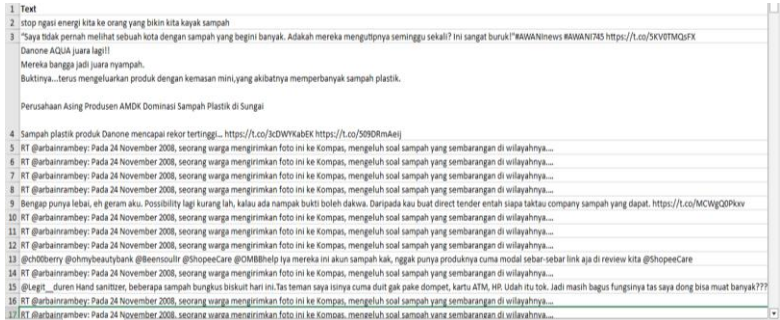
- c. *From-user-id* yaitu nomor id dari pengguna *twitter*.
- d. *To-user* yaitu pengaturan publikasi dari pengguna *twitter*.
- e. *To-user-id* yaitu nomor id yang terdapat dalam *tweet*.
- f. *Language* adalah bahasa yang digunakan ketika membuat *tweet*.
- g. *Source* adalah link menuju ke *tweet*.
- h. *Text* adalah isi *tweet* yang diunggah.
- i. *Geo-location* merupakan letak dari pengguna *twitter*.
- j. *Retweet-count* merupakan jumlah *retweet* pada *tweet* tersebut.
- k. *Id* merupakan nomor id *twitter*.

Seperti pada gambar dibawah ini

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	Created-A	From-User	From-User-To-User	From-User-To-User	To-User-Id	Language	Source	Text	Geo-Local	Geo-Local	Retweet	Retweet-Id		
1	#####	zaza	14149490916504412	12	-1	in	<a href="#"	stop ngasi energi kita ke orang			95,0	1635853715625345025		
2	#####	????Astro	497665754		-1	in	<a href="#"	"Saya tidak pernah melihat seb			18,0	1636037790126460929		
3	#####	Ary Prasef	435004096		-1	in	<a href="#"	Danone AQUA juara lagi!!Mere			22,0	1635540083326746624		
4	#####	Sultan Set	160368336716929433		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636188102707535872		
5	#####	beruangla	47976736		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636188072688910337		
6	#####	kacang	102142620046654668		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			0	163618806576723970		
7	#####	zulfi	756631692		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187993882128384		
8	#####	Syafiq	285908413		-1	in	<a href="#"	Bengap punya lebal, eh geram			0	1636187991705255937		
9	#####	-	251087525		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187973749444608		
10	#####	eliteight	106581645679164620		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187889104195585		
11	#####	Sky   busy	109956160931843276		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187800071725057		
12	#####	Aya	15797048e ch00berry	147451508		in	<a href="#"	@ch00berry @ohmybeautybar			0	1636187779502845952		
13	#####	zi	139599046738112103		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187729485778946		
14	#####	En-En	142456122 Endah_en	142456125		in	<a href="#"	@Legitt_duren Hand sanitizer,			0	1636187715833315328		
15	#####	thv.97	134153969360787865		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187688859762688		
16	#####	fia	11009463464317706		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187638561640449		
17	#####	Mayaya n	124402985		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187520215175170		
18	#####	xc 7D kok	126808301 PRISAI		-1	in	<a href="#"	@PRISAI @itxyongs disaat sam			0	163618750606771394		
19	#####	aku siapa	2703568842		-1	in	<a href="#"	RT @bdngfless: Euy, rewass suar			2145,0	1636187458999316480		
20	#####	????????	103386905148162048		-1	in	<a href="#"	RT @Onlyzza: stop ngasi energi			95,0	1636187448010227712		
21	#####	ayah 1 an	45820026		-1	in	<a href="#"	RT @arainrambey: Pada 24 Nc			902,0	1636187337284788224		
22	#####	-jomblok	1959269221		-1	in	<a href="#"	Nvadar iduo belum bener neur			0	1636187312681013248		
23		Search Twitter												

Gambar 4. 3 Hasil atribut dari *crawling* data

Untuk mempermudah dalam proses penganalisisan data, maka peneliti hanya menggunakan atribut *text* saja untuk proses lanjutan.



Gambar 4. 4 Hasil setelah pemilihan atribut

### 3. *Cleaning Data*

Tahap selanjutnya adalah tahap *cleaning data*, data yang didapat setelah melakukan tahap *cleaning data* dengan menyeleksi duplikasi dengan proses manual adalah berjumlah 149 tweet berupa *sentiment* positif dan *sentiment* negatif. Tahapan ini bertujuan agar sistem komputer lebih mengenali bentuk data set. Selain itu, tahapan ini dapat mengubah data yang tidak tersusun menjadi data yang tersusun rapi. Dalam *cleaning data* terdapat 4 proses yaitu *Transform Cases*, *Removal Annotation*, *Tokenizing*, dan *Stemming*. Pemrosesan dalam tahap ini dilakukan melalui *google colab* dengan bahasa pemrograman *Python*. Langkah

ini diawali dengan mengimportkan data dan package yang berguna untuk pembersihan data.

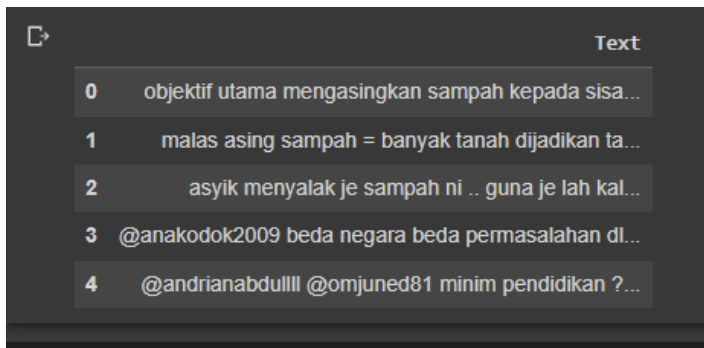


	Created-At	From-User	Text
0	2023-03-16 09:03:13	Zaidi Tuah	Objektif utama mengasingkan sampah kepada sisa...
1	2023-03-14 14:03:08	Zaidi Tuah	Malas asing sampah = banyak tanah dijadikan ta...
2	2023-03-14 14:02:47	Ahmad Zaim	Asyik menyalak je sampah ni .. Guna je lah kal...
3	2023-03-14 13:19:48	*hanya seorang kuli™??	@anakodok2009 Beda negara beda permasalahan dl...
4	2023-03-14 13:23:05	mbahimu salto	@andrianabdullll @omjuned81 Minim pendidikan ?...

Gambar 4. 5 Import data ke *google colab*

a. *Transform Cases*

*Transform case* merupakan proses pemerataan huruf dari huruf kapital menjadi huruf kecil atau sebaliknya. Pada penelitian ini data set diubah menjadi huruf kecil semua dengan menggunakan fitur *lowercase*.



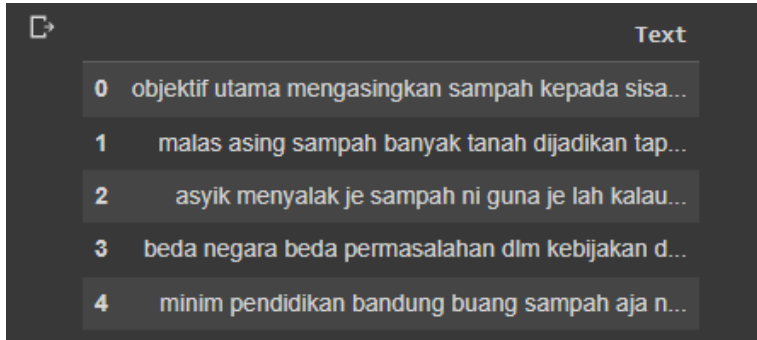
	Text
0	objektif utama mengasingkan sampah kepada sisa...
1	malas asing sampah = banyak tanah dijadikan ta...
2	asyik menyalak je sampah ni .. guna je lah kal...
3	@anakodok2009 beda negara beda permasalahan dl...
4	@andrianabdullll @omjuned81 minim pendidikan ?...

Gambar 4. 6 *output* fitur *lowercase*



b. *Removal Annotation*

Pada proses ini menghapus atribut yang tidak dibutuhkan dan tidak memiliki makna, seperti *hashtag*, *mention*, *retweet*, *whitespace*, dll.



Gambar 4. 7 hasil output removal annotation

c. *Tokenizing*

Selanjutnya proses tokenisasi adalah proses memecahkan kalimat menjadi potongan kata atau token untuk mengetahui asal munculnya kata.

	Text
0	[objektif, utama, mengasingkan, sampah, sisa, ...
1	[mala, ase, sampah, tanah, dijadikan, tapak, p...
2	[asyik, menyalak, je, sampah, ni, je]
3	[beda, negara, beda, permasalahan, dlm, kebij...
4	[minim, pendidikan, bandung, buang, sampah, aj...
5	[ak, cm, sampah]
6	[mjawapan, standard, sampah]
7	[kayanya, sadarkan, deh, tuh, kaya, gak, mikir...
8	[aqua, mempertahankan, posisi, klasemen, industri...
9	[aqua, suruh, telen, aja, tuh, sampah, plastik...
10	[aqua, tuh, emang, gak, otak, sampah, plastik,...
11	[gara, <sup>2</sup> , sampah, plastik, aqua, aqua, ei, ken...
12	[gimana, ngelak, aqua, teriakin, negara, tuh, ...
13	[gak, tau, <sup>2</sup> , aqua, mengubur, sampah, plastik,...
14	[samp, danon, ngelak, sampah, plastik, ketawa,...
15	[ya, sampah, plastik, aqua, samp, dimana]

Gambar 4. 8 hasil *output tokenizing*

d. *Stemming*

*Stemming* berfungsi sebagai penghapus kata imbuhan yang ada pada hasil sebelumnya hingga menjadi bentuk kata dasar. Berikut merupakan hasil pemrosesan *stemming*.

	Text
0	objektif utama mengasingkan sampah sisa baki sisa kitar mengurangkan kebergantungan tapak pelupusan sampah tapak pelupusan dibuka buruk nati jahnya a
1	mala ase sampah tanah dijadikan tapak pelupusan sampah asingkanlah sampah sisa dikitar sisa baki
2	asyik menyalak je sampah ni je
3	beda negara beda permasalahan dlm kebijakan tata kotanya tetanggaku gak sampah area umum pdhl selisih jam perjalanan
4	minim pendidikan bandung buang sampah aja numpuk pinggir jalan wkwkw kolam masjid nyetem goblok
5	ak em sampah
6	mjawapan standard sampah
7	kayanya sadarkan dieh tuh kaya gak mikir lingkungan dimana sampah plastik tuh * sampah plastik botol gela yg mendominasi
8	aqua mempertahankan posisi klase industri nyampah kek sengaja banget pengen bikin sampah plastik kemas
9	aqua suruh telen aja tuh sampah plastik gelasannya emang gila nya bikin lingkungan emosi aturan size up kaya tulisan * persekongkolan nih lihat aqua
10	aqua tuh emang gak otak sampah plastik aqua tanggung aja udah polutan ehk pakek aqua gela cebol
11	gara * sampah plastik aqua aqua ei kenal golongan negara segudang sampah
12	gimana ngelak aqua teriakin negara tuh gak jujur sampah plastik
13	gak tau * aqua mengubur sampah plastik nya tersembunyi
14	samp daron ngelak sampah plastik ketawa dalam hati sampah aqua datanya meningkat
15	ya sampah plastik aqua samp dimana
16	sampah si berantakan kali aja koleksi httpstcoomgymw
17	yg ngetik nyinyir udah buang sampah tempatnya
18	dage busuk dalam kulka sampah pembalut ga dibungku kulka kotor gaada bersihin tai ga disirem cucian diredem berhari
19	kumpulan sampah sampah mengganggu
20	sampah makanannya bau
21	wkt cuti krja pomp jawa timur nambah ilmu agama pake sandal eiger tidur masjid lantai sandal ditaruh didekat ta baju sandal aman pakai pa penutup sandal yg a

Gambar 4. 9 hasil *output stemming*

## 4. Pre-Processing Data

### a. Feature Engineering

Proses ini dilakukan saat data sudah melakukan *cleaning* dengan cara mengekstraksi fitur dari data mentah menjadi sebuah model prediktif dengan menggunakan proses menggantikan pelabelan dengan angka 0 diasumsikan sebagai sentimen negatif dan angka 1 diasumsikan sebagai sentimen positif.

```
↳ ['negatif' 'positif']
   [0 1]
```

Gambar 4. 10 pelabelan pada *sentiment*

Setelah dilakukannya proses *feature engineering* data akan berubah yang semula data berupa data kategori dan berubah menjadi data numerik.

```
x
0      objektif utama mengasingkan sampah kepada sisa...
1      malas asing sampah  banyak tanah dijadikan tap...
2      beda negara beda permasalahan dlm kebijakan da...
3      minim pendidikan  bandung buang sampah aja num...
4      reduce reuse recycle merupakan maksud penanganan...
...
144    pembangunan incinerator untuk optimalkan penan...
145    masa pegawai perusahaan besar gak paham alur p...
146    udah sering bikin macet jalur puncak ditambah ...
147    prediksi kami tpa sarimukti sudah overload pen...
148    memprihatinkan mau isinya rumput atau apa tete...
Name: Text, Length: 149, dtype: object

y
0      1
1      0
2      0
3      0
4      1
...
144    1
145    0
146    0
147    0
148    0
Name: Label, Length: 149, dtype: int64
```

Gambar 4. 11 Pemisahan antara teks dengan label

*b. Feature Extraction*

Proses *feature extraction* dipenelitian ini memanfaatkan metode TF-IDF yaitu pembobotan yang dipakai untuk mengetahui korelasi kata (*term*) terhadap kalimat pada masing-masing kata. Sebuah kata akan semakin besar bobotnya ketika seringnya kata tersebut muncul dalam kalimat, dan berlaku sebaliknya dengan mengaplikasikan persamaan 2.1.

	10	12	2024	3r	70	abis	abong	acara	ada	adakah	...	wilayahnya	wisata	wskw	wkdw	ya	yakni	yang	yayaan	yg	zero
0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0
1	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0
2	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.137879	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0
3	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.287744	0.0	0.0	0.000000	0.0	0.0	0.0
4	0.0	0.0	0.0	0.237044	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.105298	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
144	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0
145	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0
146	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0
147	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0
148	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0

149 rows x 1043 columns

Gambar 4. 12 *output* TF-IDF

c. *Feature Selection*

*Feature selection* adalah suatu proses yang bertujuan untuk memilih fitur yang berpengaruh dan mengesampingkan fitur yang tidak berpengaruh, proses ini dilakukan didalam *mechine learning*. Dalam proses ini nilai data tabular dari TF-IDF akan diubah menjadi *array* agar tidak terjadi galat saat proses seleksi fitur dijalankan. Apabila hasil dari nilai semakin tinggi maka hal tersebut juga mempengaruhi hasil fiturnya.

	nilai	fitur
992	3.077399	tpst
86	2.501218	bau
44	2.465307	aqua
556	1.978282	mengganggu
710	1.912666	pengelolaan
...	...	...
25	0.000124	aku
366	0.000075	juga
302	0.000065	harus
595	0.000044	minim
91	0.000028	beberapa

1043 rows x 2 columns

Gambar 4. 13 pengurutan nilai terhadap kata

Setelah mengetahui hasil dari nilai maka akan ditentukan fitur yang terbaik dan menyeleksi fitur yang memiliki nilai terendah, pemilihan fitur yang dipilih adalah 100 fitur, pemilihan ini dipilih untuk menyempurnakan akurasi pada model,

	3r	ada	agar	aja	apa	aqua	baik	bakti	bangun	bau	...	tapi	teman	tempatnya	terpadu	thrif
0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
1	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.137879	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
3	0.000000	0.000000	0.0	0.190787	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
4	0.237044	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
144	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
145	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
146	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
147	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
148	0.000000	0.000000	0.0	0.000000	0.258218	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

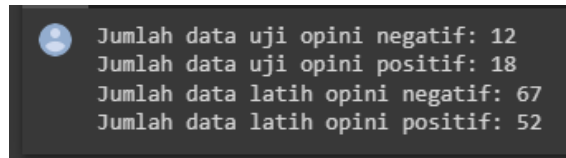
149 rows x 100 columns

Gambar 4. 14 penyeleksian kata dengan pemilihan fitur sebesar 100

### 5. Split Data

Pemilihan besar pembagian antara data *training* dan data *testing* sangat penting dalam proses penelitian ini karena dalam pemilihan besar split data yang mengakibatkan tingkat keakurasian yang tinggi pula, maka dengan ini pada peneliti menggunakan split data sebesar 80:20 dengan 80% untuk jumlah data *training* dan 20% untuk data *testing*. Data ini terdiri dari 149 data yang terdiri dari 2 atribut yaitu *text* dan label. Label dibagi menjadi 2 *sentiment* yaitu negatif (0) dan positif (1). Data dengan *sentiment* negatif sebanyak 79 data sedangkan untuk *sentiment* positif sebanyak 70 data. Selanjutnya data ini dibagi menjadi dua data secara acak dengan data *training* sebesar 80% dan data *testing*

sebesar 20%. Dari *split* data ini diperoleh data *training* sebesar 119 data dengan data beropini negatif sebesar 67 data dan data beropini positif sebesar 52 data. Sedangkan pada data *testing* keseluruhan terdapat 30 data dengan data *bersentiment* negatif sebesar 12 data dan data *bersentiment* positif sebesar 18 data.

A screenshot of a terminal window with a dark background and light blue text. On the left side, there is a small circular icon with a blue and white gradient. The text in the terminal is as follows:

```
Jumlah data uji opini negatif: 12  
Jumlah data uji opini positif: 18  
Jumlah data latih opini negatif: 67  
Jumlah data latih opini positif: 52
```

Gambar 4. 15 *output proses split data*

## 6. ***Normalization Data dan SMOTE***

### a. Normalization data

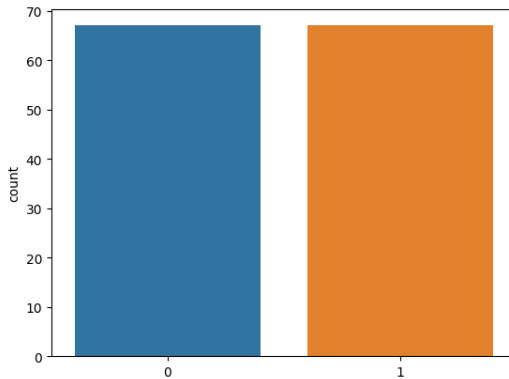
Data yang telah dibagi selanjutnya perlu dilakukan *scaling data* atau normalisasi merupakan proses untuk mengganti nilai numerik dalam dataset ke skala umum, skala umum pada *scaling data* berada pada rentang 0 dan 1, tanpa mendistorsi perbedaan dalam rentang nilai agar data yang digunakan nantinya tidak memiliki penyimpangan yang besar terhadap atribut lainnya. Adapun metode yang digunakan dalam mengatasi penyimpangan ini adalah *MinMaxScaler*



yang mana merupakan *class* dari *sklearn* dengan memanfaatkan persamaan 2.2.

b. Sampling Data

Pada label data terlihat mengalami ketidakseimbangan kelas dikarenakan *sentiment* negatif berjumlah 67 dan *sentiment* positif berjumlah 52 sehingga perlu untuk dilakukan proses sampling data pada data *training* terlebih dahulu. Pada penelitian ini, peneliti melakukan proses sampling data dengan metode SMOTE.

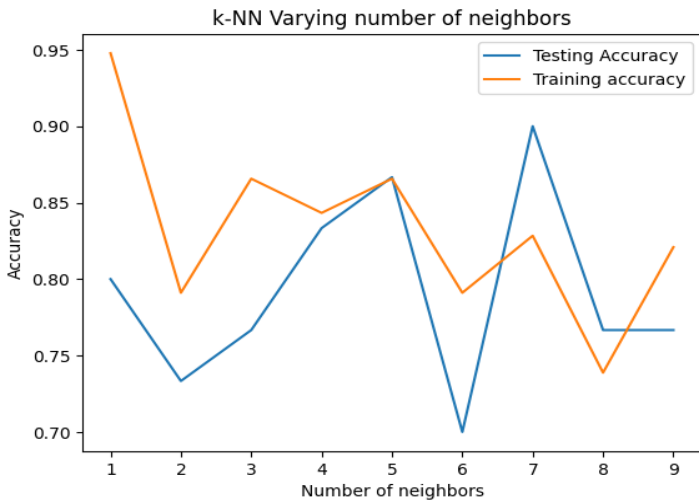


Gambar 4. 16 *plot* hasil proses SMOTE

## 7. Modelling

### a. Klasifikasi dengan *K-Nearest Neighbor*

Parameter yang digunakan pada *K-Nearest Neighbor* ini yaitu *n\_neighbors*, *metric='euclidean'* dimana nilai *k* diambil berdasarkan grafik perbandingan akurasi data *testing* dengan data *training* dengan skala nilai *k* = 1-9 yang akan diinterpretasikan ke dalam gambar 4.17, setelah menentukan skala maka langkah yang dilakukan selanjutnya adalah mencari tetangga terdekat dengan formula jarak *euclidean* seperti pada *persamaan 2.11*



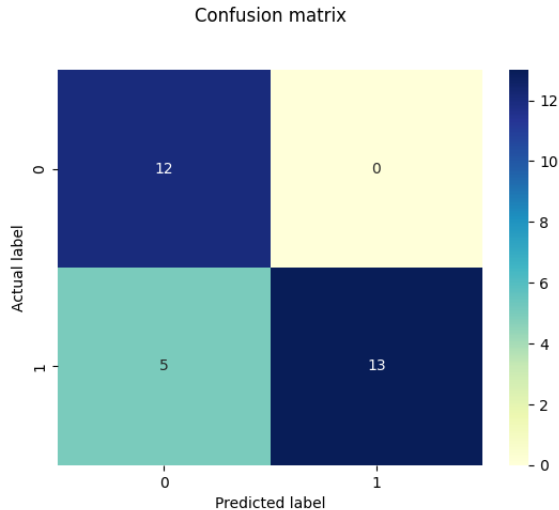
Gambar 4. 17 skala nilai *k*

Berdasarkan grafik di atas dapat dilihat bahwa kondisi *fitting* terbaik ketika nilai  $k=4$ .

```
KNeighborsClassifier  
KNeighborsClassifier(metric='euclidean', n_neighbors=4)
```

Gambar 4. 18 penentuan nilai  $k$

Berdasarkan hasil eksperimen yang telah peneliti lakukan ini menggunakan algoritma klasifikasi *K-Nearest Neighbor* diketahui bahwa perhitungan akurasi dengan memanfaatkan *confusion matrix* yang didapat sebagai berikut :



Gambar 4. 19 *confusion matrix* KNN

Perhitungan Akurasi pada Data *Testing*

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \\ &= \frac{12 + 13}{12 + 13 + 5 + 0} \times 100\% \\ &= 83.3\% \end{aligned}$$

Perhitungan manual dapat dikoreksi dengan perhitungan dalam google colab sehingga menghasilkan

	precision	recall	f1-score	support
0	0.71	1.00	0.83	12
1	1.00	0.72	0.84	18
accuracy			0.83	30
macro avg	0.85	0.86	0.83	30
weighted avg	0.88	0.83	0.83	30

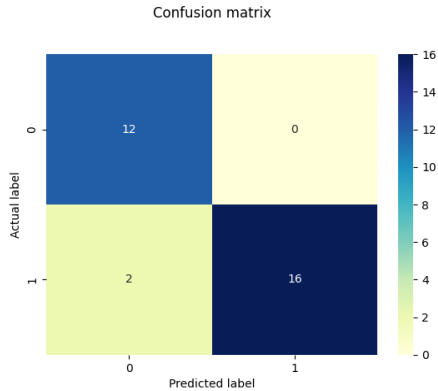
Gambar 4. 20 perhitungan akurasi menggunakan google colab

Sehingga dapat disimpulkan bahwa hasil perhitungan manual dan perhitungan menggunakan bantuan google colab adalah sama yaitu memiliki akurasi 83%

b. Klasifikasi dengan *Naïve Bayes*

Metode naïve bayes yang peneliti gunakan yaitu *Gaussian Naive Bayes* yang mana merupakan algoritma dasar berdasarkan penerapan teorema bayes sesuai dengan teorema 2.5. Penerapan teorema bayes memanfaatkan tahap *preprocessing* karena tahap *gaussian Naive Bayes* membutuhkan data berbentuk numerik.

Selanjutnya mencari probabilitas menggunakan metode algoritma *Naive Bayes*. Berdasarkan hasil eksperimen yang telah peneliti lakukan ini menggunakan algoritma klasifikasi *Naïve Bayes* diketahui bahwa model juga telah berhasil membedakan data ke dalam kelas klasifikasi. Diketahui bahwa perhitungan akurasi dengan memanfaatkan *confusion matrix* yang didapat sebagai berikut :



Gambar 4. 21 *confusion matrix Naive Bayes*

Perhitungan Akurasi pada Data *Testing*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$= \frac{12 + 16}{12 + 16 + 2 + 0} \times 100\%$$

$$= 93.3\%$$

Perhitungan manual dapat dikoreksi dengan perhitungan dalam google colab sehingga menghasilkan

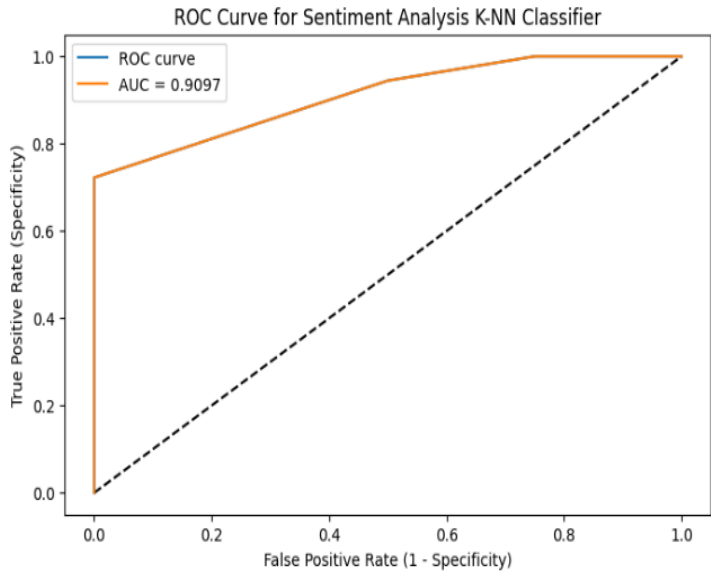
	precision	recall	f1-score	support
0	0.86	1.00	0.92	12
1	1.00	0.89	0.94	18
accuracy			0.93	30
macro avg	0.93	0.94	0.93	30
weighted avg	0.94	0.93	0.93	30

Gambar 4. 22 *output* akurasi menggunakan google colab

## 8. Evaluasi

Kurva ROC dan AUC merupakan sebuah perhitungan kinerja pada klasifikasi data mining untuk mengetahui keakurasian pada suatu model klasifikasi data mining. Pada kurva ROC apabila garis mendekati 1 maka akurasi yang diperoleh sangat tinggi, hal ini didukung dengan nilai yang didapat oleh kurva AUC yang berada pada bawah garis.

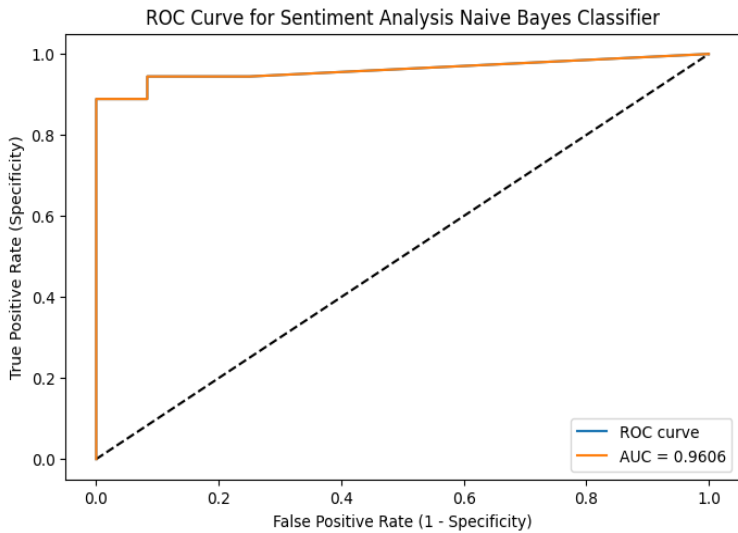
Kurva AUC yang didapat oleh klasifikasi data mining KNN adalah 0,9097, nilai tersebut apabila dimasukkan kedalam penilaian kurva ROC dengan dipresentasikan kurva AUC memiliki performa yang sangat baik.



Gambar 4. 23 output kurva ROC & AUC KNN

Sedangkan kurva AUC yang didapat oleh klasifikasi data mining *Naive Bayes* adalah 0.9606. dan nilai tersebut apabila dimasukkan kedalam penilaian kurva ROC dengan dipresentasikan kurva AUC memiliki performa yang sangat baik.





**Gambar 4. 24** output kurva ROC & AUC KNN

Berdasarkan perbandingan kedua algoritma yang telah dilakukan, maka dapat disimpulkan bahwasanya algoritma *Naive Bayes* merupakan algoritma terbaik dalam menyelesaikan proses analisis *sentimen* ini. Hal ini dibuktikan dengan hasil akurasi dan nilai AUC pada algoritma *Naive Bayes* yang lebih unggul dari algoritma *K-Nearest Neighbor* yaitu 93.3% untuk nilai akurasi dan 0.9606 pada nilai AUC.

## BAB V

### KESIMPULAN DAN SARAN

#### 1. Kesimpulan

Berdasarkan pembahasan hasil pengujian maka dapat diambil kesimpulan sebagai berikut:

- a. Berdasarkan analisis dari metode klasifikasi KNN, kata yang sering muncul pada tweet tentang keluhan masyarakat tentang sampah adalah “kotor” dan “bau” dengan data berjumlah 149 *tweet*. Data digolongkan menjadi 2 label yaitu *sentiment* positif dan *sentiment* negatif. Data diolah menggunakan *split data* sebesar 80:20 dengan 80% data *training* dan 20% data *testing* sehingga data *training* berubah menjadi 119 data dengan *sentiment* negatif sebesar 67 data dan *sentiment* positif sebesar 52 data dan data *testing* berubah menjadi 30 data dengan data ber *sentiment* negatif sebesar 12 dan data ber *sentiment* positif sebesar 18 data. Pemodelan KNN ini menghasilkan akurasi yang diperoleh dari metode penelitian tersebut adalah 83% dengan nilai kurva AUC sebesar 0,9097.
- b. Berdasarkan analisis dari metode klasifikasi *Naive Bayes*, kata yang sering muncul pada tweet tentang

keluhan masyarakat tentang sampah adalah “kotor” dan “bau” dengan data berjumlah 149 *tweet*. Data digolongkan menjadi 2 label yaitu *sentiment* positif dan *sentiment* negatif. Data diolah menggunakan *split data* sebesar 80:20 dengan 80% data *training* dan 20% data *testing* sehingga data *training* berubah menjadi 119 data dengan *sentiment* negatif sebesar 67 data dan *sentiment* positif sebesar 52 data dan data *testing* berubah menjadi 30 data dengan data ber *sentiment* negatif sebesar 12 dan data ber *sentiment* positif sebesar 18 data. Pemodelan *Naive Bayes* ini menghasilkan akurasi yang diperoleh dari metode penelitian tersebut adalah 93% dengan nilai kurva AUC sebesar 0,9606.

- c. Keakurasian metode K-Nearest Neighbor mendapatkan akurasi sebesar 0.8333 % atau 83 % sedangkan dengan metode Naive Bayes didapat akurasi sebesar 0,9333% atau 93% . Nilai tersebut didapatkan dari perhitungan akurasi dengan menggunakan TP, TN, FP, dan FN yang telah didapatkan dari hasil *confusion matrix*. Dari hasil akurasi tersebut maka membuktikan bahwa kedua metode layak digunakan namun metode *Naive Bayes*

memiliki keakurasian lebih tinggi daripada metode KNN.

## **2. Saran**

Ada beberapa hal yang peneliti sarankan untuk pengembangan penelitian selanjutnya:

- a. Penelitian ini dapat diganti menggunakan topik yang berbeda dengan mencari topik masalah yang spesifik.
- b. Penelitian ini dapat dikembangkan dengan metode klasifikasi data mining lainnya dengan tingkat *split data* yang berbeda pula.

## DAFTAR PUSTAKA

- Andriani, A. (2013). Seminar Nasional Teknologi Informasi dan Komunikasi 2013 (SENTIKA 2013) SISTEM PENDUKUNG KEPUTUSAN BERBASIS DECISION TREE DALAM PEMBERIAN BEASISWA STUDI KASUS: AMIK “BSI YOGYAKARTA.” *Seminar Nasional Teknologi Informasi Dan Komunikasi 2013 (SENTIKA 2013)*, 163–169.
- Ardiansyah, D., & Walim. (2018). *ALGORITMA C4.5 UNTUK KLASIFIKASI CALON PESERTA LOMBA CERDAS CERMAT SISWA SMP DENGAN MENGGUNAKAN APLIKASI RAPID MINER* / Ardiansyah / *Jurnal Inkofar*. *Jurnal Inkofar*. <http://politeknikmeta.ac.id/meta/ojs/index.php/inkofar/article/view/29/45>
- Ariyanti, D., & Iswardani, K. (2020). Teks Mining untuk Klasifikasi Keluhan Masyarakat Pada Pemkot Probolinggo Menggunakan Algoritma Naïve Bayes. *Jurnal IKRA-ITH Informatika*, 4(3), 125–132.
- Axmalia, A., & Asti Mulasari, S. (2020). Dampak Tempat Pembuangan Akhir Sampah (TPA) Terhadap Gangguan Kesehatan Masyarakat. *Jurnal Kesehatan Komunitas*, 6(2), 171–176.  
<https://doi.org/10.25311/KESKOM.VOL6.ISS2.536>
- Ayudhitama, A. P., & Pujiyanto, U. (2020, February). *View of*

*ANALISA 4 ALGORITMA DALAM KLASIFIKASI LIVER MENGGUNAKAN RAPIDMINER*. Jurnal Informatika Polinema.

<http://jip.polinema.ac.id/ojs3/index.php/jip/article/view/274/234>

Brzezinski, D., & Stefanowski, J. (2017). Prequential AUC: properties of the area under the ROC curve for data streams with concept drift. *Knowledge and Information Systems*, 52(2), 531–562.

<https://doi.org/10.1007/S10115-017-1022-8/FIGURES/8>

Cortez, S., Mcnerney, K., Arbelaez, A. M., Zdonczyk, A., Tychsen, L., & Reynolds, M. (2022). 412 Cortisol cut off point to diagnose adrenal insufficiency (AI) using a monoclonal antibody immunoassay. *Journal of Clinical and Translational Science*, 6(s1), 80–80.

<https://doi.org/10.1017/CTS.2022.239>

Dewi, S. (2016). *View of KOMPARASI 5 METODE ALGORITMA KLASIFIKASI DATA MINING PADA PREDIKSI KEBERHASILAN PEMASARAN PRODUK LAYANAN PERBANKAN*. Jurnal Techno Nusa Mandiri.

<http://ejournal.nusamandiri.ac.id/index.php/techno/article/view/218/194>

Dhany, H. W., & Izhari, F. (2019, July). *View of ANALISIS*

*ALGORITHMS SUPPORT VECTOR MACHINE DENGAN  
NAIVE BAYES KERNEL PADA KLASIFIKASI DATA.*

JURNALTEKNIK DAN INFORMATIKA .

<https://journal.pancabudi.ac.id/index.php/Juti/article/view/675/639>

Erdiansyah, U., Lubis, A. I., & Erwansyah, K. (2022, January).

*Komparasi Metode K-Nearest Neighbor dan Random  
Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan*

*Penyakit Kutil | Erdiansyah | JURNAL MEDIA*

*INFORMATIKA BUDIDARMA. JURNAL MEDIA*

*INFORMATIKA BUDIDARMA. [\*\[budidarma.ac.id/index.php/mib/article/view/3373/239\]\(http://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/3373/239\)\*](http://ejurnal.stmik-</a></i></p></div><div data-bbox=)*

*0*

Fajariani, R., Vidyaningrum, D. U., & Haryati, S. (2022).

*PENGGUNAAN ALAT PELINDUNG DIRI DAN KELUHAN  
PENYAKIT KULIT PADA PETUGAS PENGANGKUT*

*SAMPAH. *Jurnal Ilmu Kesehatan Masyarakat*, 18(2), 91–*

*98. <https://doi.org/10.19184/IKESMA.V18I1.26881>*

Farid, M., Wibowo, S., Puspitasari, N. F., & Satya, B. (2022).

*PENERAPAN DATA MINING DAN ALGORITMA NAÏVE  
BAYES UNTUK PEMILIHAN KONSENTRASI MAHASISWA*

*MENGGUNAKAN METODE KLASIFIKASI. *Journal of**

*Information System Management (JOISM)*, 3(2), 39–45.

<https://doi.org/10.24076/JOISM.2022V3I2.680>

Fitriatien, S. R. (2017). PENGANTAR STATISTIKA UNTUK PENELITIAN: SUATU KAJIAN. *Buana Pendidikan: Jurnal Fakultas Keguruan Dan Ilmu Pendidikan Unipa Surabaya*, 13(23), 47–53.

<https://doi.org/10.36456/BP.VOL13.NO23.A450>

Fitriyani, F., & Arifin, T. (2020). PENERAPAN WORD N-GRAM UNTUK SENTIMENT ANALYSIS REVIEW MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (STUDI KASUS: APLIKASI SAMBARA). *Sistemasi: Jurnal Sistem Informasi*, 9(3), 610–621.

<https://doi.org/10.32520/STMSI.V9I3.954>

Garbian Nugroho, D., Herry Chrisnanto, Y., & Wahana, A. (2016). ANALISIS SENTIMEN PADA JASA OJEK ONLINE MENGGUNAKAN METODE NAÏVE BAYES. *Prosiding Seminar Sains Nasional Dan Teknologi*, 1(1).

[https://www.publikasiilmiah.unwahas.ac.id/index.php/PROSIDING\\_SNST\\_FT/article/view/1526](https://www.publikasiilmiah.unwahas.ac.id/index.php/PROSIDING_SNST_FT/article/view/1526)

Girnanfa, F. A., & Susilo, A. (2022). *View of Studi Dramaturgi Pengelolaan Kesan Melalui Twitter Sebagai Sarana Eksistensi Diri Mahasiswa di Jakarta*. *Journal of New Media and Communication*.

<https://journal.sinergiinstitute.com/index.php/JNMC/article/view/2/5>

Gunadi, G., & Sensuse, D. I. (2016). ... analysis terhadap data



penjualan produk buku dengan menggunakan algoritma apriori dan frequent pattern growth (fp-growth): studi kasus percetakan pt. Gramedia. *Telematika MKOM*, 118–132.

<https://journal.budiluhur.ac.id/index.php/telematika/article/view/164%0Ahttps://journal.budiluhur.ac.id/index.php/telematika/article/download/164/158>

Gunawan, B., SastyPratiwi, H., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, 4(2), 113–118.

<https://doi.org/10.26418/JP.V4I2.27526>

Gunawan, D., Riana, D., Ardiansyah, D., Akbar, F., & Alfarizi, S. (2022). Komparasi Algoritma Support Vector Machine Dan Naïve Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Calon Gubernur Jabar 2018-2023. *Jurnal Teknik Komputer AMIK BSI*, 6(1), 121–130.

<https://doi.org/10.31294/jtk.v4i2>

Hendrian, S. (2018). Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan. *Faktor Exacta*, 11(3).

<https://doi.org/10.30998/FAKTOREXACTA.V11I3.2777>

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). *MLCM: Multi-Label Confusion Matrix*. IEEE Acces.

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9711932>

Hozairi,, Anwari,, & Alim, S. (2021). IMPLEMENTASI ORANGE DATA MINING UNTUK KLASIFIKASI KELULUSAN MAHASISWA DENGAN MODEL K-NEAREST NEIGHBOR, DECISION TREE SERTA NAIVE BAYES. *Network Engineering Research Operation*, 6(2), 133–144.

<https://doi.org/10.21107/NERO.V6I2.237>

Ihsyaluddin, & Mane, A. (2022, July). *View of KESADARAN LINGKUNGAN DALAM PENGELOLAAN SAMPAH DI PANTAI NIRWANA KOTA BAUBAU*. *Urnal Green Growth and Manajemen Lingkungan*.

<http://journal.unj.ac.id/unj/index.php/jgg/article/view/26419/12717>

Ilić, M., Srdjević, Z., & Srdjević, B. (2022). Water quality prediction based on Naïve Bayes algorithm. *Water Science and Technology*, 85(4), 1027–1039.

<https://doi.org/10.2166/WST.2022.006>

Imandasari, T., Irawan, E., Perdana Windarto, A., & Wanto, A. (2019). Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, 1(0), 750–761.

<http://tunasbangsa.ac.id/seminar/index.php/senaris/article/view/81>

- Indrayuni, E. (2019). Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes. *Jurnal Khatulistiwa Informatika*, 7(1), 29–37. <https://doi.org/10.31294/JKI.V7I1.5740>
- Irfan, M., Uriawan, W., Ramdhani, M. A., & Dahlia, I. A. (2018). Comparison of Naive Bayes and K-Nearest Neighbor methods to predict divorce issues. *IOP Conf. Series: Materials Science and Engineering*. <https://doi.org/10.1088/1757-899X/434/1/012047>
- Iskandar, D., & Suprpto, Y. K. (2013). Perbandingan akurasi klasifikasi tingkat kemiskinan antara algoritma C4.5 dan Naïve Bayes Clasifier. *JAVA Journal of Electrical and Electronics Engineering*, 11(1), 14–17.
- Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D., & Sharma, M. (2018). Credit card fraud detection using Naïve Bayes model based and KNN classifier. *International Journal of Advance Research*, 4(3), 44–48. [www.IJARIIIT.com](http://www.IJARIIIT.com)
- Latifa, N. (2018). *PROFIL FARMAKOKINETIK FLAVONOID EKSTRAK DAUN PEPAYA PADA PLASMA DARAH TIKUS*.
- Lestari, F. A., Efrizoni, L., Ali, E., & Rahmiati. (2022, June). *View of Sistem Klasifikasi Pengaduan Masyarakat Pada BPJS Ketenagakerjaan Menggunakan Algoritma Naïve Bayes Berbasis Mobile*. Building of Informatics, Technology and Science (BITS). <http://ejurnal.seminar->

- id.com/index.php/bits/article/view/1685/1107
- Liantoni, F. (2015). Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor. *Ultimatics : Jurnal Teknik Informatika*, 7(2), 98–104. <https://doi.org/10.31937/TI.V7I2.356>
- Limbong, J. J. A. ., Sembiring, I., & Hartomo, K. D. (2022, April). *Analisis Klasifikasi Sentimen Ulasan pada E-Commerce Shopee Berbasis Word Cloud dengan Metode Naive Bayes dan K-Nearest Neighbor | Limbong | Jurnal Teknologi Informasi dan Ilmu Komputer*. Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIK). <https://jtiik.ub.ac.id/index.php/jtiik/article/view/4960/pdf>
- Ling, H. (2022). Teaching Design of Mathematics Application Based on Naive Bayes. *Mathematical Problems in Engineering*, 2022, 1–6. <https://doi.org/10.1155/2022/7244001>
- Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2(11), 4704–4713. <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/3051>
- Mahyudin, R. P. (2017). *KAJIAN PERMASALAHAN*

*PENGELOLAAN SAMPAH DAN DAMPAK LINGKUNGAN DI TPA (TEMPAT PEMROSESAN AKHIR) | Mahyudin | Jukung (Jurnal Teknik Lingkungan). Jurnal Teknik Lingkungan. <https://ppjp.ulm.ac.id/journal/index.php/jukung/article/view/3201/2745>*

Manurung, E., & Hasugian, P. S. (2019). *DATA MINING TINGKAT PESANAN INVENTARIS KANTOR MENGGUNAKAN ALGORITMA APRIORI PADA KEPOLISIAN DAERAH SUMATERA UTARA | Journal Of Informatic Pelita Nusantara. Journal Of Informatic Pelita Nusantara. <https://ejournal.pelitanusantara.ac.id/index.php/JIPN/article/view/608/0>*

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, 2(2), 213–219. <https://doi.org/10.22202/EI.2016.V2I2.1465>

Mardianti, S., Zidny Naf, M., Hidayatulloh, I., & Teknologi, F. (2018). EKSTRAKSI TF-IDF N-GRAM DARI KOMENTAR PELANGGAN PRODUK SMARTPHONE PADA WEBSITE E-COMMERCE. *SEMNAS TEKNOLOGI ONLINE*, 6(1), 1-2-79. <https://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/2061>

Meilani, B. D., & Susanti, N. (2016). *APLIKASI DATA MINING*

UNTUK MENGHASILKAN POLA KELULUSAN SISWA  
DENGAN METODE NAÏVE BAYES. *Network Engineering  
Research Operation*, 1(3), 182–189.

<https://nero.trunojoyo.ac.id/index.php/nero/article/view/27>

Mualana, R., & Redjeki, S. (2016). *ANALISIS SENTIMEN  
PENGGUNA TWITTER MENGGUNAKAN METODE  
SUPPORT VECTOR MACHINE BERBASIS CLOUD  
COMPUTING | Maulana | Jurnal TAM (Technology  
Acceptance Model)*. *Jurnal TAM (Technology Acceptance  
Model)*.

[https://ojs.stmikpringsewu.ac.id/index.php/JurnalTam/  
article/view/57/57](https://ojs.stmikpringsewu.ac.id/index.php/JurnalTam/article/view/57/57)

Mukminin, A., & Riana, D. (2017). Komparasi Algoritma C4.5,  
Naïve Bayes Dan Neural Network Untuk Klasifikasi  
Tanah. *Jurnal Informatika*, 4(1), 21–31.

[https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/v  
iew/1002](https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/1002)

Muslehatin, W., Ibnu, M., & Mustakim. (2017, May).  
*PENERAPAN NAÏVE BAYES CLASSIFICATION UNTUK  
KLASIFIKASI TINGKAT KEMUNGKINAN OBESITAS  
MAHASISWA SISTEM INFORMASI UIN SUSKA RIAU |  
Muslehatin | Seminar Nasional Teknologi Informasi  
Komunikasi dan Industri*. <http://ejournal.uin->

- suska.ac.id/index.php/SNTIKI/article/view/3276/2158
- Mustafidah, H., & Giarto, W. G. P. (2021). Aplikasi Berbasis Web untuk Analisis Data Menggunakan Korelasi Bivariat Pearson. *Sainteks*, 18(1), 39–50.  
<https://doi.org/10.30595/SAINTEKS.V18I1.10564>
- Nabila, Z., Isnain, A. R., Permata, & Abidin, Z. (2021). *ANALISIS DATA MINING UNTUK CLUSTERING KASUS COVID-19 DI PROVINSI LAMPUNG DENGAN ALGORITMA K-MEANS / Nabila | Jurnal Teknologi dan Sistem Informasi*. Jurnal Teknologi Dan Sistem Informasi (JTSI).  
<http://jim.teknokrat.ac.id/index.php/sisteminformasi/article/view/868/355>
- Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, 75(1), 25–36.  
<https://doi.org/10.4097/kja.21209>
- Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *CESS (Journal of Computer Engineering, System and Science)*, 4(1), 78–82.  
<https://jurnal.unimed.ac.id/2012/index.php/cess/article/view/11458>
- Naz, I., Bano, Z., & Anjum, R. (2022). Construct and criterion validity of adjustment scale for adults using the

- correlation and Receiver-Operating Characteristics Analysis. *Rawal Medical Journal*, 47(1), 89–93.
- Novianti, N., Zarlis, M., & Sihombing, P. (2022, April). *Penerapan Algoritma Adaboost Untuk Peningkatan Kinerja Klasifikasi Data Mining Pada Imbalance Dataset Diabetes* / Novianti / *JURNAL MEDIA INFORMATIKA BUDIDARMA*. *JURNAL MEDIA INFORMATIKA BUDIDARMA*. <http://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/4017/2676>
- Nuqoba, B., & Djunaidy, A. (2014). ALGORITMA PREDIKSI OUTLIER MENGGUNAKAN BORDER SOLVING SET. *Jurnal Informatika Mulawarman*, 9(3), 10.
- Nurdin, A., Anggo, B., Aji, S., Bustamin, A., & Abidin, Z. (2020). PERBANDINGAN KINERJA WORD EMBEDDING WORD2VEC, GLOVE, DAN FASTTEXT PADA KLASIFIKASI TEKS. *Jurnal Tekno Kompak*, 14(2), 74–79. <https://doi.org/10.33365/JTK.V14I2.732>
- Oktara, B. (2022, June). *View of HUBUNGAN SIKAP DENGAN PERILAKU MASYARAKAT DALAM PENGELOLAAN SAMPAH*. *Jurnal Ilmiah Wijaya*. <https://jurnalwijaya.com/index.php/jurnal/article/view/161/152>
- Osman, A. S. (2019). *View of Data Mining Techniques: Review*.



Al-Madinah International University Malaysia.

<http://ojs.mediu.edu.my/index.php/IJDSR/article/view/1841/717>

Ozdemir, S., & Algin, A. (2022). Interpretation Of the Area Under the Receiver Operating Characteristic Curve. *Experimental and Applied Medical Science*, 3(1), 310–311. <https://doi.org/10.46871/eams.2022.35>

Permata, A. D. (2022). *Dimensi Ekologi Dalam Penafsiran Alquran Surah Ar-Rum Ayat 41 Dan Al-A'raf Ayat 56 (Studi Kitab Tafsir AlMisbah Karya Muhammad Quraish Shihab)*.

Purba, E., Purba, B., Syafli, A., Khairad, F., Damanil, D., Siagian, V., Ginting, A. M., Arfandi, & Ernanda, R. (2021). *Metode Penelitian Ekonomi* (R. Watrianthos (ed.)). Yayasan Kita Menulis.

[https://books.google.co.id/books?hl=id&lr=&id=5DE0EAAAQBAJ&oi=fnd&pg=PA59&dq="+Penelitian+ini+melakukan+kuantitatif+komparatif+yang+memiliki+fungsi+untuk+membandingkan+dua+perlakuan+dalam+suatu+parameter+atau+beberapa+parameter+dalam+waktu+bersamaan&ots](https://books.google.co.id/books?hl=id&lr=&id=5DE0EAAAQBAJ&oi=fnd&pg=PA59&dq=)

Putri, H. N., Retno, D., & Saputro, S. (2022). Clustering Data Campuran Numerik dan Kategorik Menggunakan Algoritme Ensemble Quick RObust Clustering using linkS

(QROCK). *PRISMA, Prosiding Seminar Nasional Matematika*, 5, 716–720.

<https://journal.unnes.ac.id/sju/index.php/prisma/article/view/54590>

Putry, N. M., & Sari, B. N. S. (2022, September). *KOMPARASI ALGORITMA KNN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELLITUS | Putry | EVOLUSI : Jurnal Sains dan Manajemen*. Jurnal Sains Dan Manajemen .

<https://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/view/12514/5403>

Rachmawati, A. K., Saleh Saleh, M., & Ramdani, M. N. (2022). Comparison and Prediction of Data Mining Models to Determine the Classification of Family Planning Program User Status. *Indonesian Journal of Mathematics Education*, 4(2), 66–73.

Rahman Isnain, A., Indra Sakti, A., Alita, D., & Satya Marga, N. (2021). Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm. *Jdmsi*, 2(1), 31–37. <https://t.co/NfhnmJtXw>

Rahman, M. A., Hidayat, N., & Supianto, A. A. (2018). *Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)*.

2(12), 6346–6353. <http://j-ptiik.ub.ac.id>

Reza Noviansyah, M., Rismawan, T., & Marisa Midyanti, D. (2018). PENERAPAN DATA MINING MENGGUNAKAN METODE K-NEAREST NEIGHBOR UNTUK KLASIFIKASI INDEKS CUACA KEBAKARAN BERDASARKAN DATA AWS (AUTOMATIC WEATHER STATION) (STUDI KASUS: KABUPATEN KUBU RAYA). *Coding Jurnal Komputer Dan Aplikasi*, 6(2), 48–56.  
<https://jurnal.untan.ac.id/index.php/jcskommipa/article/view/26672>

Rhomadhona, H., & Permadi, J. (2019). Klasifikasi Berita Kriminal Menggunakan Naïve Bayes Classifier (NBC) dengan Pengujian K-Fold Cross Validation. *Jurnal Sains Dan Informatika*, 5(2), 108–117.  
<https://doi.org/10.34128/JSI.V5I2.177>

Rosso, G. A. (2019). Milton. *William Blake in Context*, September, 184–191.  
<https://doi.org/10.1017/9781316534946.021>

Samsir, S., Ambiyar, A., Verawardina, U., Edi, F., & Watrianthos, R. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 5(1), 157–163.  
<https://doi.org/10.30865/MIB.V5I1.2580>

- Santoso, I. B. (2014). Deteksi Obyek Nyata (Pada Lingkup : Visualisasi dan Deteksi Obyek Nyata pada Lingkungan Hidup). *MATICS*, 6(2), 59–64.  
<https://doi.org/10.18860/MAT.V6I2.2597>
- Santra, A. K., & Christy, C. J. (2012). *Genetic Algorithm and Confusion Matrix for Document Clustering*. IJCSI International Journal of Computer Science.  
<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.2710&rep=rep1&type=pdf>
- Silwattananusarn, T., & KulthidaTuamsuk, A. (2012). Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 2(5), 13–24.  
<https://doi.org/10.48550/arxiv.1210.2872>
- Somantri, O., & Apriliani, D. (2018). Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 5(5), 537–548.  
<https://doi.org/10.25126/JTIK.201855867>
- Suryani, A. S. (2014a). Peran Bank Sampah Dalam Efektivitas Pengelolaan Sampah (Studi Kasus Bank Sampah Malang). *Aspirasi*, 5(1), 71–84.

<https://dprexternal3.dpr.go.id/index.php/aspirasi/article/view/447/344>

Suryani, A. S. (2014b, June). *PERAN BANK SAMPAH DALAM EFEKTIVITAS PENGELOLAAN SAMPAH (STUDI KASUS BANK SAMPAH MALANG) | Suryani | Aspirasi: Jurnal Masalah-masalah Sosial*. Jurnal Masalah-Masalah Sosial. <http://jurnal.dpr.go.id/index.php/aspirasi/article/view/447>

Susanti, L., Utomo, S. W., & Takarina, N. D. (2022). Estimating the Nanobubble Aerated System and Stocking Density Effects on Oxygen Consumption and Survival of *Litopenaeus vannamei* (Boone, 1931) Postlarvae 8 Using Receiver Operating Characteristic (ROC) Analysis. *International Journal on Advanced Science, Engineering and Information Technology*, 12(1), 270–277. <https://doi.org/10.18517/IJASEIT.12.1.15323>

Syarifuddin, M. (2020). ANALISIS SENTIMEN OPINI PUBLIK MENGENAI COVID-19 PADA TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN KNN. *INTI Nusa Mandiri*, 15(1), 23–28. <https://doi.org/10.33480/INTI.V15I1.1347>

Syukri Mustafa, M., Rizky Ramadhan, M., & Thenata, A. P. (2018). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma

Naive Bayes Classifier. *Creative Information Technology Journal*, 4(2), 151–162.

<https://citec.amikom.ac.id/main/index.php/citec/article/view/106>

Tempola, F., Muhammad, M., & Khairan, A. (2018a, October).

*Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation | Tempola | Jurnal Teknologi Informasi dan Ilmu Komputer*. PERBANDINGAN KLASIFIKASI ANTARA KNN DAN NAIVE BAYES PADA PENENTUAN STATUS GUNUNG BERAPI DENGAN K-FOLD CROSS VALIDATION.

<https://jtiik.ub.ac.id/index.php/jtiik/article/view/983/pdf>

Tempola, F., Muhammad, M., & Khairan, A. (2018b).

Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 5(5), 577–584.

<https://doi.org/10.25126/JTIK.201855983>

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode

Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 4(2), 437–444.

<https://doi.org/10.30865/MIB.V4I2.2080>

Utomo, D. P., & Purba, B. (2019). Penerapan Datamining pada Data Gempa Bumi Terhadap Potensi Tsunami di Indonesia. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, 1(0), 846–853.

<http://tunasbangsa.ac.id/seminar/index.php/senaris/article/view/91>

Verawati, P., & Verawati, P. (2022). KEBIJAKAN EXTENDED PRODUCER RESPONSIBILITY DALAM PENANGANAN MASALAH SAMPAH DI INDONESIA MENUJU MASYARAKAT ZERO WASTE. *JUSTITIA : Jurnal Ilmu Hukum Dan Humaniora*, 9(1), 189–197.

<https://doi.org/10.31604/justitia.v9i1.189-197>

Wahyudi, W. E., & Nawafilaty, T. (2020). Pendampingan Pemuda Investasi Sampah Berbasis Media Sosial di Desa Jugo, Sekaran, Lamongan. *Jurnal Abdimas Berdaya : Jurnal Pembelajaran, Pemberdayaan Dan Pengabdian Masyarakat*, 1(02), 73–81.

<https://doi.org/10.30736/JAB.V1I02.24>

Wanika Siburian, V., & Elvina Mulyana, I. (2019). Prediksi Harga Ponsel Menggunakan Metode Random Forest. *Annual Research Seminar (ARS)*, 4(1), 144–147.

<https://seminar.ilkom.unsri.ac.id/index.php/ars/article/view/1992>

- Whendasromo, R. G., & Joseph, J. (2022). Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN. *JURIKOM (Jurnal Riset Komputer)*, 9(4), 872.  
<https://doi.org/10.30865/jurikom.v9i4.4526>
- winata, P. A. (2022). *View of KLASIFIKASI NAIVE BAYES KEPARAHAN TRAUMA PASIEN MENGGUNAKAN DATA NEURO COGNITIVE DAN DATA PHYSIOLOGIC DENGAN PYTHON*. Eminar Nasional Matematika, Geometri, Statistika, Dan Komputas.  
<https://jurnal.unej.ac.id/index.php/prosiding/article/view/33500/11662>
- Witten, I. H. ., Frank, E., & Hall, Ma. A. . (2008). Data Mining. In *Encyclopedia of Ecology, Five-Volume Set*.  
<https://doi.org/10.1016/B978-008045405-4.00153-1>
- Yang, Z., Xu, Q., Bao, S., He, Y., Cao, X., & Huang, Q. (2021). When All We Need is a Piece of the Pie: A Generic Framework for Optimizing Two-way Partial AUC. *PMLR*, 139, 139.
- Zhang, C., Zhong, P., Liu, M., Song, Q., Liang, Z., & Wang, X. (2022). Hybrid Metric K-Nearest Neighbor Algorithm and Applications. *Mathematical Problems in Engineering*, 1–15. <https://doi.org/10.1155/2022/8212546>



## LAMPIRAN

### Lampiran 1 Hasil *crawling data*

	Created-At	From-User	From-User-id	To-User	To-User-id	Language	Source	Text	Geo-Location-Latitude	Geo-Location-Longitude	Retweet-Count	Id	
1	2023-03-15 11.01.33	zaza	1414949091650441220		-1	in	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	stop ngasi energi kita ke orang yang bikin kita kayak sampah			95,0	1635853715625345025	
2	2023-03-15 23.12.59	????Astro AWANI????	497665754		-1	in	<a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>	"Saya tidak pernah melihat sebuah kota dengan sampah yang begini banyak. Adakah mereka mengujinya seminggu sekali? ini sangat buruk!" #AWANINews #AWANITAS https://t.co/SKVOTMQtFX			18,0	1636037790126460929	
3	2023-03-14 14.15.17	Ary Prasetyo	435004096		-1	in		Danone AQUA juara lagi!! Mereka bangga jadi juara nyampah. Buktinya...terus mengeluarkan produk dengan kemasan mini, yang akibatnya memperbanyak sampah plastik. Perusahaan Asing Produsen AMDK Dominasi Sampah Plastik di Sungai			22,0	1635540083326746624	
4	2023-03-16 09.10.17	Sultan Sehn??	1603683367169284336		-1	in	<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	Sampah plastik produk Danone mencapai rekor tertinggi... https://t.co/3cDWYKbEK					
5	2023-03-16 09.10.10	beruanglaut	47976736		-1	in	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	RT @arbalrambey: Pada 24 November 2008, seorang warga mengirimkan foto ini ke Kompas, menguuh soal sampah yang sembarangan di wilayahnya....			902,0	1636188102707535872	
6	2023-03-16 09.10.10	beruanglaut	47976736		-1	in	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	RT @arbalrambey: Pada 24 November 2008, seorang warga mengirimkan foto ini ke Kompas, menguuh soal sampah yang sembarangan di wilayahnya....			902,0	1636188072688910337	

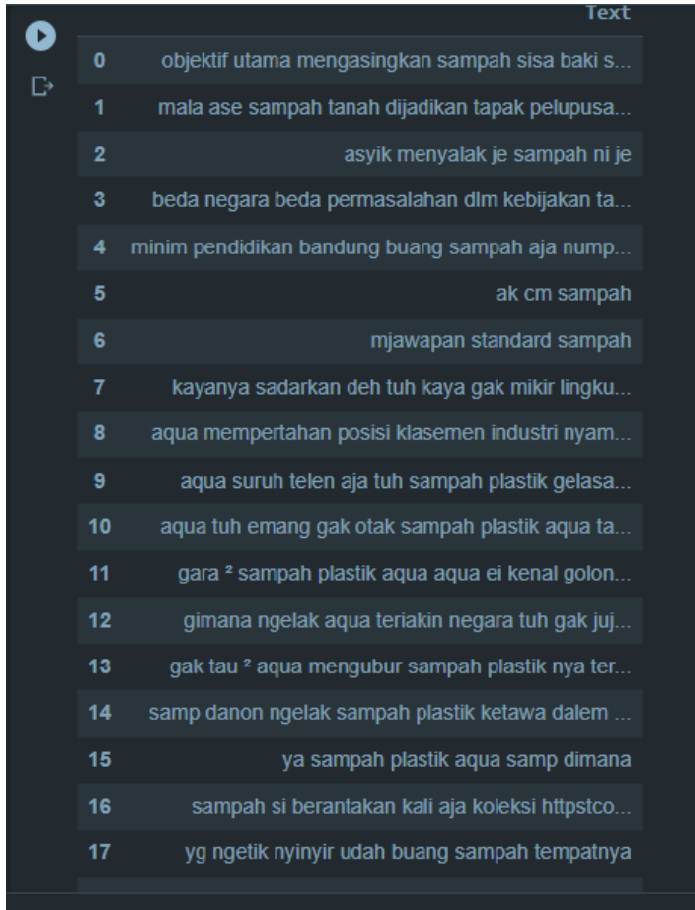
	Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Text	Geo-Location-Latitude	Geo-Location-Longitude	Retweet-Count	Id
1												
495	2023-03-14 12.59.46	kadal	1575913572454256640		-1	in	<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	RT @bdngfess: Euy, rawas suami aku balik dalam keadaan kaya gini bercucur pisan ua Allah???? ditebas sama samurai ku gengobig di jalan Paso...			2145,0	1635521080025845760
496	2023-03-14 12.59.44	Hotticlassleekyzzz	1533773404720877574		-1	in	<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	RT @exquisiteme_: @tanyanfess Tau ga cowok paling digustung di dunia: yg gampang bergetar hatinya km cewe lain. For Anne, maybe this is s...			798,0	1635521071335215105
497	2023-03-14 12.59.40	IDCorner????	168469708		-1	in	<a href="https://idcorner.co.id" rel="nofollow">IDCorner.Co.Id</a>	Dimas LH Tanggapi Tak Angkut Sampah Pinggir Tol Ini: Bukan Bagian Pelayanan - <a href="https://t.co/r18IGd9PF">https://t.co/r18IGd9PF</a>			.0	1635521053232623616
498	2023-03-14 12.59.26	au ah	803984354850418688		-1	in	<a href="http://twitter.com/download/phone" rel="nofollow">Twitter for iPhone</a>	RT @bdngfess: Euy, rawas suami aku balik dalam keadaan kaya gini bercucur pisan ua Allah???? ditebas sama samurai ku gengobig di jalan Paso...			2145,0	1635520994864680961
499	2023-03-14 12.59.19	Siti Aisyah	1546998799847276544	jennief	1097020578257485826	in	<a href="https://mobile.twitter.com" rel="nofollow">Twitter Web App</a>	@jennief: @mainimasjambu03 @tanyakand Bukan tersinggung masalahnya diluar pulau Jawa juga banyak sampah seperti ini, kenapa harus Jawa doang yg ditebut? Kalau mau gua bisa sebutin daerah mana aja, tapi gw gak mau rasis kek lu . Kok bisa otaknya gak bisa buat mikir hal seade			.0	1635520967400386560
500	2023-03-14 12.59.18	????????????????_mentioned after dm please -!!!	1305736164188405760		-1	in	<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	RT @Tukang_wlso: @bdngfess @RESTABES_BDG Mohon untuk diamankan pak, tembak ditempat saja pelakunya, walaupun anak dibawah umur, tembak aja....			80,0	1635520961561923584
501	2023-03-14 12.59.05	kumparan	759692754985242625		-1	in	<a href="https://tvri1.com/" rel="nofollow">tvri1.com</a>	Prilly Latuconsina tunjukkan aksi pungut sampah di pantai, sisinggung rencana terkait program Generasi Peduli Bumi. #womensupdate #update #woman #text <a href="https://t.co/R77H18irt">https://t.co/R77H18irt</a>			.0	1635520906897559557

Lampiran 2 Hasil dari pemilihan atribut kelas

<b>Text</b>	<b>Label</b>
objektif utama mengasingkan sampah kepada sisa baki dengan sisa kitar semula adalah untuk mengurangkan kebergantungan kita kepada tapak pelupusan sampah semakin banyak tapak pelupusan dibuka semakin buruk natijahnya kepada a	positif
malas asing sampah banyak tanah dijadikan tapak pelupusan sampah asingkanlah sampah anda kepada sisa boleh dikitar semula dan sisa baki	negatif
beda negara beda permasalahan dlm kebijakan dan tata kotanya tetanggaku misalnya hampir gak ada sampah pun di area umumpdhl cuma selisih jam lebih perjalanan	negatif
minim pendidikan bandung buang sampah aja numpuk pinggir jalan wkwkw kolam masjid malah buat nyelem goblok	negatif
reduce reuse recycle merupakan maksud penanganan sampah yang terdiri dari 3	positif

unsur mengurangi menggunakan ulang dan mendaur ulang sampah juga dikenal sebagai 3r	
ridwan kamil menyebut penyediaan lahan tempat pembuangan sampah tps terpadu yang ditargetkan selesai di 70 titik penanganan lahan kritis hingga kualitas air sejauh ini masih menjadi pekerjaan besar untuk segera dituntaskan	positif
dkg sulteng gelar pelatihan pengelolaan dan penanganan sampah plastik di laut	Positif
kolaborasi seluruh masyarakat diperlukan untuk memutus mata rantai permasalahan sampah plastik di laut indonesia	Positif
sampah cekungan bandung butuh penanganan serius bpkp cekban urun rembug	Negative
penanganan sampah di indonesia masih sangat mengkhawatirkan	Negative
:	:
aqua mau mempertahankan kan posisi atas klasemen industri paling nyampah kek sengaja banget pengen bikin masalah sampah plastik lewat kemasan kecil	Negative

### Lampiran 3 Hasil data sesudah melalui proses *cleaning data*



The image shows a screenshot of a text editor window with a dark background. The title bar at the top right says "Text". On the left side, there are two icons: a play button and a document icon with an arrow. The main area contains 18 lines of text, each starting with a number from 0 to 17. The text is partially obscured by a dark overlay, but the visible parts are as follows:

Line Number	Text
0	objektif utama mengasingkan sampah sisa baki s...
1	mala ase sampah tanah dijadikan tapak pelupusa...
2	asyik menyalak je sampah ni je
3	beda negara beda permasalahan dlm kebijakan ta...
4	minim pendidikan bandung buang sampah aja nump...
5	ak cm sampah
6	mjawapan standard sampah
7	kayanya sadarkan deh tuh kaya gak mikir lingku...
8	aqua mempertahankan posisi klasemen industri nyam...
9	aqua suruh telen aja tuh sampah plastik gelasa...
10	aqua tuh emang gak otak sampah plastik aqua ta...
11	gara ² sampah plastik aqua aqua ei kenal golon...
12	gimana ngelak aqua teriakin negara tuh gak juj...
13	gak tau ² aqua mengubur sampah plastik nya ter...
14	samp danon ngelak sampah plastik ketawa dalem ...
15	ya sampah plastik aqua samp dimana
16	sampah si berantakan kali aja koleksi httpstco...
17	yg ngetik nyinyir udah buang sampah tempatnya

#### Lampiran 4 Hasil data setelah melakukan proses *pre-processing data*

	3r	ada	agar	aja	apa	aqua	baik	bakti	bangun	bau	...	tapi	teman	tempatya	terpadu	thrifl
0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
1	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.137879	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
3	0.000000	0.000000	0.0	0.190787	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
4	0.237044	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
144	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
145	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
146	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
147	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
148	0.000000	0.000000	0.0	0.000000	0.258218	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

149 rows x 100 columns

3	apa	aqua	baik	bakti	bangun	bau	...	tapi	teman	tempatny	terpadu	thrift	tpst	tuh	untuk	upaya	utama
)	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.112096	0.0	0.155269
)	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000
)	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000
7	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000
)	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000
-	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
)	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.256733	0.0	0.000000
)	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000
)	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000
)	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000
)	0.258218	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000

## Lampiran 5 *script cleaning data*

```
from google.colab import drive
drive.mount('/content/drive')

import numpy as np
import pandas as pd

import string

import nltk

nltk.download('stopwords')

# mencari pola kata spesifik yang dicari/mendeteksi pola
# tertentu dalam string

import re

# melakukan klasifikasi sentimen dari suatu kalimat/tweet

from textblob import TextBlob

# Hilangkan notif karena migrasi dari Python 2.7.x ke 3.7.1

import warnings

warnings.filterwarnings('ignore')

# Membaca dataset dari file excel

dataset =
pd.read_excel('/content/drive/MyDrive/datasampah.xlsx')

dataset.head()
```



```

# ambil record Text
dataClean = pd.DataFrame(dataset[["Text"]])
dataClean.tail()

# konversi ke lowercase
dataClean["Text"] = dataClean["Text"].apply(lambda x: "
".join(x.lower() for x in x.split()))
dataClean.head()

# hilangkan angka
dataClean["Text"] = dataClean["Text"].str.replace('[0-9]+','')
dataClean.head()

# Remove link
def cleanmentions(kata):
    kata = re.sub(r'@[A-Za-z0-9]+',' ', kata)
    kata = re.sub(r'https?:\:\/\/\S+',' ', kata)
    kata = re.sub(r'(?:\@|http?:\:\/\/|https?:\:\/\/|www)\S+',' ',
kata)
    kata = re.sub(r'_',' ', kata)

    return kata

# Remove mentions
dataClean["Text"] = dataClean["Text"].apply(cleanmentions)

```

```

dataClean.head()
# Hilangkan Punctuation >> eg : @, #, etc
dataClean['Text'] = dataClean['Text'].str.replace('[^\w\s]','')
dataClean.head()
# Tokenize, Stopwords, Stemming
from nltk.tokenize import TweetTokenizer
from nltk.corpus import stopwords
stopwords = stopwords.words('indonesian')
#import Stemmer
ps = nltk.PorterStemmer()
# hilangkan stopWords
dataClean['Text'] = dataClean['Text'].apply(lambda x: ".join(x for x in x.split() if x not in stopwords)")
dataClean.head()
from nltk.tokenize import TweetTokenizer
from nltk.corpus import stopwords
stopwords = stopwords.words('indonesian')
#import Stemmer
ps = nltk.PorterStemmer()
# tokenize tweets
def tokenizing(tweet):

```

```

tokenizer = TweetTokenizer(preserve_case = False,
reduce_len = True strip_handles = True)

tweet_tokens = tokenizer.tokenize(tweet)

tweets_clean=[]

for word in tweet_tokens:

    if (word not in stopwords and word not in
string.punctuation):

        #stemming

        stem_word = ps.stem(word)

        #append word stemmed

        tweets_clean.append(stem_word)

return tweets_clean

dataClean['Text'] = dataClean['Text'].apply(lambda x:
tokenizing(x))

dataClean.head(20)

def join(Text):

    Text = " ".join([char for char in Text])

    return Text

dataClean['Text'] = dataClean['Text'].apply(lambda x: join(x))

dataClean.to_excel("dataClean.xlsx")

dataClean.head(20)

```

## Lampiran 6 proses *pre-processing data* dan *modelling*

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
from scipy import stats
from mlxtend.preprocessing import minmax_scaling
warnings.filterwarnings("ignore")
import pickle
data_path = '/content/drive/MyDrive/dataClean.xlsx'
# Mengimport dataset

data = pd.read_excel(data_path)
data.head()
data.info()
data_review = data[['Text']]
data_review.head()
```

```

data["Label"].value_counts()

print('Total Jumlah Data:', data.shape[0], 'data\n')

print('terdiri dari (label):')

print('-- [1] negatif\t:', data[data.Label == 'negatif'].shape[0],
'data')

print('-- [2] positif\t:', data[data.Label == 'positif'].shape[0],
'data')

height = data['Label'].value_counts()

labels = ['negatif', 'positif']

y_pos = np.arange(len(labels))

plt.figure(figsize=(5,3), dpi=100)

plt.ylim(0,60)

plt.title('jumlah opinion tweet', fontweight='bold')

plt.xlabel('opinion', fontweight='bold')

plt.ylabel('Jumlah tweet', fontweight='bold')

plt.bar(y_pos, height, color=['deepskyblue', 'royalblue',
'skyblue'])

plt.xticks(y_pos, labels)

plt.show()

from sklearn.preprocessing import LabelEncoder

LE = LabelEncoder()

# Convert feature 'Label'

```

```
data['Label'] = LE.fit_transform(data['Label'])
print(LE.classes_)
print(np.sort(data['Label'].unique()))
print("")
```

# Pisahkan kolom feature dan target

```
X = data['Text']
```

```
y = data['Label']
```

```
X
```

```
y
```

```
'''
```

Convert a collection of raw documents to a matrix of TF-IDF features

[https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_extraction.text.TfidfVectorizer.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)

```
'''
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
tf_idf = TfidfVectorizer(ngram_range=(1,1))
```

```
tf_idf.fit(X)
```

```
x_tf_idf = tf_idf.transform(X)
```

# Melihat Jumlah Fitur

```
print(len(tf_idf.get_feature_names_out()))
```

```

# Melihat fitur-fitur apa saja yang ada di dalam corpus
print(tf_idf.get_feature_names_out())

# Melihat matriks jumlah token

# Data ini siap untuk dimasukkan dalam proses pemodelan
(machine learning)

X_tf_idf = tf_idf.transform(X).toarray()

X_tf_idf

# Melihat matriks jumlah token menggunakan TF IDF, lihat
perbedaannya dengan metode BoW

# Data ini siap untuk dimasukkan dalam proses pemodelan
(machine learning)

data_tf_idf = pd.DataFrame(X_tf_idf,
columns=tf_idf.get_feature_names_out())

data_tf_idf

with open('tf_idf_feature.pickle', 'wb') as output:

    pickle.dump(X_tf_idf, output)

# Mengubah nilai data tabular tf-idf menjadi array agar dapat
dijalankan pada proses seleksi fitur

X = np.array(data_tf_idf)

y = np.array(y)

from sklearn.feature_selection import SelectKBest

from sklearn.feature_selection import chi2

# Ten features with highest chi-squared statistics are selected

```

```

chi2_features = SelectKBest(chi2, k=100)
X_kbest_features = chi2_features.fit_transform(X, y)

# Reduced features

print('Original feature number:', X.shape[1])

print('Reduced feature number:', X_kbest_features.shape[1])

# chi2_features.scores_ adalah nilai chi-square, semakin
tinggi nilainya maka semakin baik fiturnya

data_chi2 = pd.DataFrame(chi2_features.scores_,
columns=['nilai'])

data_chi2

# Menampilkan fitur beserta nilainya

feature = tf_idf.get_feature_names_out()

data_chi2['fitur'] = feature

data_chi2

# Mengurutkan fitur terbaik

data_chi2.sort_values(by='nilai', ascending=False)

# Menampilkan mask pada feature yang diseleksi

# False berarti fitur tidak terpilih dan True berarti fitur
terpilih

mask = chi2_features.get_support()

mask

```



```

# Menampilkan fitur-fitur terpilih berdasarkan mask atau
nilai tertinggi yang sudah dikalkulasi pada Chi-Square

new_feature = []

for bool, f in zip(mask, feature):

    if bool:

        new_feature.append(f)

    selected_feature = new_feature

selected_feature

tf_idf.vocabulary_

# Lihat vocab yang dihasilkan oleh TF_IDF

# tf_idf.vocabulary_

kbest_feature = {} # Buat dictionary kosong

for (k,v) in tf_idf.vocabulary_.items(): # Iterasi untuk
mengulangi vocab yang dihasilkan TF_IDF

    if k in selected_feature:          # Cek apakah fitur termasuk
k fitur yang diseleksi

        kbest_feature[k] = v          # Jika iya, simpan fitur
tersebut pada dictionary kosong diatas

kbest_feature

# Menampilkan fitur-fitur yang sudah diseleksi

# Beserta nilai vektornya pada keseluruhan data untuk
dijalankan pada proses machine learning

```

```

# Hanya k fitur yang terpilih sesuai parameter k yang
ditenentukan sebelumnya

data_selected_feature = pd.DataFrame(X_kbest_features
columns=selected_feature)

data_selected_feature

with open('kbest_feature.pickle', 'wb') as output:
    pickle.dump(kbest_feature, output)

#membagi data training dan data testing
#membagi untuk testing 20%

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test =
train_test_split(X_kbest_features, y, test_size = 0.2,
random_state = 42)

negatif_test = (y_test == 0).sum()
positif_test = (y_test == 1).sum()

negatif_train = (y_train == 0).sum()
positif_train = (y_train == 1).sum()

print('Jumlah data uji opini negatif:', negatif_test)
print('Jumlah data uji opini positif:', positif_test)
print('Jumlah data latih opini negatif:', negatif_train)
print('Jumlah data latih opini positif:', positif_train)

from sklearn.preprocessing import MinMaxScaler

```

```

sc = MinMaxScaler()
X_test = sc.fit_transform(X_test)
X_train = sc.fit_transform(X_train)
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state = 0)
X_train, y_train = sm.fit_resample(X_train, y_train)
sns.countplot(x = y_train)

#import KNeighborsClassifier
from sklearn.neighbors import KNeighborsClassifier
#Setup arrays to store training and test accuracies
neighbors = np.arange(1,10)
train_accuracy = np.empty(len(neighbors))
test_accuracy = np.empty(len(neighbors))
for i, k in enumerate(neighbors):
    #Setup a knn classifier with k neighbors
    knn = KNeighborsClassifier(n_neighbors=k)
    #Fit the model
    knn.fit(X_train, y_train)
    #Compute accuracy on the training set
    train_accuracy[i] = knn.score(X_train, y_train)

```

```

#Compute accuracy on the test set
test_accuracy[i] = knn.score(X_test, y_test)

#Generate plot
plt.title('k-NN Varying number of neighbors')
plt.plot(neighbors, test_accuracy, label='Testing Accuracy')
plt.plot(neighbors, train_accuracy, label='Training accuracy')
plt.legend()
plt.xlabel('Number of neighbors')
plt.ylabel('Accuracy')
plt.show()

#Setup a knn classifier with k neighbors

knn =
KNeighborsClassifier(n_neighbors=4,metric='euclidean')

#Fit the model

knn.fit(X_train, y_train)

#let us get the predictions using the knn classifier we had fit
above

y_pred_knn = knn.predict(X_test)

# Simpan model knn hasil traning

from joblib import dump
dump(knn, filename='model_knn.joblib')

```

```

#import NaiveBayesClassifier
from sklearn.naive_bayes import GaussianNB
#Setup naive bayes classifier
NaiveBayes = GaussianNB()
#let us get the predictions using the naive bayes classifier we
had fit above
y_pred_nb = NaiveBayes.predict(X_test)
# Simpan model naive bayes hasil traning
from joblib import dump
dump(NaiveBayes, filename='model_nb.joblib')
from sklearn.metrics import confusion_matrix
y_pred_knn = knn.predict(X_test)
cnf_matrix = confusion_matrix(y_test, y_pred_knn)
p = sns.heatmap(pd.DataFrame(cnf_matrix), annot=True,
cmap="YlGnBu", fmt='g')
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
from sklearn.metrics import classification_report
akurasi = classification_report(y_test, y_pred_knn)
print(akurasi)

```

```

# Drawing the ROC Curve

from sklearn.metrics import (confusion_matrix,
precision_recall_curve, auc,

                                roc_curve, recall_score, classification_report,
f1_score,

                                precision_recall_fscore_support)

fpr, tpr, thresholds =
roc_curve(y_test, knn.predict_proba(X_test)[:,1])

roc_auc = auc(fpr, tpr)

plt.plot(fpr, tpr, label="ROC curve")

plt.plot([0,1], [0,1], 'k--')

plt.plot(fpr, tpr, label='AUC = %0.4f' % roc_auc)

plt.title('ROC Curve for Sentiment Analysis K-NN Classifier')

plt.xlabel('False Positive Rate (1 - Specificity)')

plt.ylabel('True Positive Rate (Specificity)')

plt.legend()

plt.gcf().set_size_inches(8,5)

plt.show()

from sklearn.metrics import confusion_matrix

y_pred_nb = NaiveBayes.predict(X_test)

cnf_matrix = confusion_matrix(y_test, y_pred_nb)

```

```

p = sns.heatmap(pd.DataFrame(cnf_matrix), annot=True,
cmap="YlGnBu", fmt='g')

plt.title('Confusion matrix', y=1.1)

plt.ylabel('Actual label')

plt.xlabel('Predicted label')

from sklearn.metrics import classification_report

akurasi = classification_report(y_test, y_pred_nb)

print(akurasi)

# Drawing the ROC Curve
from sklearn.metrics import (confusion_matrix,
precision_recall_curve, auc,
roc_curve, recall_score, classification_report,
f1_score,
precision_recall_fscore_support)

fpr, tpr, thresholds =
roc_curve(y_test, NaiveBayes.predict_proba(X_test)[:,-1])

roc_auc = auc(fpr, tpr)

plt.plot(fpr, tpr, label="ROC curve")

plt.plot([0,1], [0,1], 'k--')

plt.plot(fpr, tpr, label='AUC = %0.4f' % roc_auc)

plt.title('ROC Curve for Sentiment Analysis Naive Bayes
Classifier')

plt.xlabel('False Positive Rate (1 - Specificity)')

plt.ylabel('True Positive Rate (Specificity)')

```

```
plt.legend()
```

```
plt.gcf().set_size_inches(8,5)
```

```
plt.show()
```

Lampiran 7 panduan *script naive bayes*

[https://colab.research.google.com/github/ShinyQ/Analisis-Sentimen-Kebijakan-Vaksinasi-COVID-19-Pemerintah-Naive-Bayes-Classifier/blob/main/Tugas\\_Besar\\_WGTIK.ipynb#scrollTo=C4i9CpbSyPm](https://colab.research.google.com/github/ShinyQ/Analisis-Sentimen-Kebijakan-Vaksinasi-COVID-19-Pemerintah-Naive-Bayes-Classifier/blob/main/Tugas_Besar_WGTIK.ipynb#scrollTo=C4i9CpbSyPm) (diakses pada tanggal 11 Desember 2022)

Lampiran 8 panduan script K-Nearest Neighbor

[https://colab.research.google.com/github/teliofm/Minerando/blob/master/Scikit\\_Learn\\_KNN\\_Best\\_Practices.ipynb](https://colab.research.google.com/github/teliofm/Minerando/blob/master/Scikit_Learn_KNN_Best_Practices.ipynb) (diakses pada tanggal 23 September 2022)



## **DAFTAR RIWAYAT HIDUP**

### **1. Identitas Diri**

Nama Lengkap : Rahyan Elena Mahatiara  
Tempat, Tanggal Lahir : Brangsong, 21 Juni 2001  
Alamat : Jalan Nyai Sentono RT 07/RW  
03 Brangsong, Kec. Brangsong,  
Kab. Kendal  
Nomor HP : 0899577436  
Email : [rhyneln21@gmail.com](mailto:rhyneln21@gmail.com)

### **2. Riwayat Pendidikan**

Pendidikan Formal :

- a. SD Negeri 01 Brangsong
- b. SMP Negeri 01 Brangsong
- c. SMA Negeri 2 Kendal

Pendidikan Non Formal :-

Semarang, 21 Juni 2023

Rahyan Elena Mahatiara

NIM. 1908046003