BAB II
TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Penelitian analisis sentimen yang dilakukan oleh Suryono *et al*. (2018) dengan bertujuan untuk mengetahui tingkat kepercayaan masyarakat serta terbentuk citra kepada calon Gubenur Jawa Barat 2018-2023. Penelitian yang di lakukan menggunakan algoritma *Naïve Bayes* serta Support *Vector Machine*. Penentuan fitur menentukan hasil akurasi, dalam penentuan fitur seleksi digunakan *Genetic* Algorithma agar dapat meningkatan akurasi pengklasifikasian pada *Support Vector Machine* dan *Naive Bayes*. Perolehan Hasil pengujian data tweet mengenai calon gubernur jawa barat periode 2018-2023 model algoritma *Support Vector Machine* berbasis *Genetic* Algorithm adalah model algoritma terbaik dalam penelitian ini dan dapat memberikan hasil terbaik dalam pengujian dan pengklasifikasian analisis sentiment *tweet* calon gubernur jawa barat periode 2018-2023 dibandingkan dengan model algoritma *Naïve Bayes berbasis Genetic Algorithm (NB-GA)*. Di dalam penelitian yang dilakukan oleh Suryono terdapat persamaan penggunaan algoritma yang dilakukan oleh penulis dalam penelitiannya yaitu menggunakan Algoritma *Naïve Bayes* dan feature selection Algoritma *Genetika*.

Analisis sentimen selanjutnya dilakukan oleh Permadi (2020). Penelitian ini bertujuan unutk mengetahui peforma algoritma *Naïve Bayes Classifer* dalam melakukan klasifikasi sentimen bedasarkan komentar pengunjung restoran. Penelitian dilakukan dengan metode *Naïve Bayes Classifer* untuk mengklasifikasikan data komentar pengunjung restoran menjadi 2 kategori sentimen, yaitu positif dan negatif. Didasarkan pada hasil uji coba tingkat akurasi klasifikasi yang dilakukan oleh metode *Naïve Bayes Classifier* memiliki nilai tertinggi pada percobaan keempat dengan 66.79% dengan perbandingan

pembagian data yaitu 60% untuk data training dan 40% untuk data testing. Berdasarkan hasil pecobaan didapatkan fakta bahwa semakin banyak porsi data untuk data training maka tingkat akurasinya juga akan meningkat (Permadi, 2020). Di dalam penelitian yang dilakukan oleh Permadi terdapat perbedaan dalam penelitian yang dilakukan oleh penulis. Perbedaannya adalah di dalam penelitian yang dilakukan oleh Suryono menggunakan satu algoritma yaitu Algoritma *Naïve Bayes Classifer*, dan penelitian yang dilakukan itu menggunakan algoritma Naïve Bayes Classifer dan akan di optimasikan ke dalam algoritma *Genetika* untuk *feature selection* dengan tujuan untuk mencari nilai akurasi yang paling optimal .

Penelitian lainya yang dilakukan oleh Arif Rahman *et al*. (2021) yang bertujuan untuk melihat nilai akurasi yang tinggi terhadap aplikasi di google playstore. Pada penelitian ini menganalisis 4 aplikasi yang ada di playstore yaitu *Shoope, Ruangguru, Tokopedia dan Gojek*. Penelitian di mulai dari proses pengumpulan dataset, pengkelasan dataset, proses oversampling, proses pre-processing, proses algoritma *Naive Bayes* dan proses algoritma *Genetika*. Hasil nilai akurasi pada penelitian ini menyatakan bahwa nilai akurasi dari algoritma Genetika lebih tinggi dari nilai akurasi algoritma Naive Bayes, sehingga bisa dikatakan bahwa algoritma *Genetika* mampu meningkatkan kualitas darialgoritma *klasifikasi Naïve Bayes*. Pengambilan data dalam penelitian yang dilakukan oleh Rahman di ambil dari data aplikasi Playstore. Dan data yang akandi ambil oleh penulis dalam penelitiannya adalah data dari *Twitter*.

Penelitian yang dilakukan oleh Watrianthos *et al*. (2019). Penelitian ini bertujuan untuk menegetahui persepsi pengguna berdasarkan pengukuran kualitas layanan sehingga hasilnya dapat menjadi evaluasi bagi Traveloka dalam meningkatkan layanan. Studi menunjukkan bahwa selama opini publik periode ini menghasilkan sentimen negatif dengan nilai Vmap sebesar 0,31020 lebih besar dari sentimen positif dengan nilai 0,16132. Dengan menggunakan algoritma Naïve Bayes classifer. Hasil penelitian ini menghasilkan bahwa selama

periode ini opini publik menghasilkan negatif sentimen dengan nilai Vmap 0,31020 lebih besar dari positif sentimen dengan nilai 0,16132. Hasil negatif ini diperoleh mengenai aspek penggunaan aplikasi dan konsumen yang kecewa dengan tiket mahal harga. Perbedaan penelitian yang di lakukan oleh Ronal masih menggunakan satu algoritma yaitu Naïve Bayes Classifer. Data dalam penelitian di ambil dri review aplikasi Traveloka. Penelitian yang dilakukan oleh penulis menggunakan dua algoritma yaitu algoritma *Naïve Bayes Classifer* dan algoritma *Genetika*. Pengambilan data yg dilakukan penulis di ambil dari sentimen *Twitter* vaksinasi Covid-19.

Penelitian yang dilakukan Yulita *et a*l. (2021). Penelitian ini bertujuan untuk menganalisis pendapat tentang vaksinasi COVID-19 di Indonesia. Analisis dilakukan terhadap data 3780 tweet yang berkaitan vaksinasi dengan menggunakan algoritma Naïve Bayes Classifier. Hasil dari penelitian yang dilakukuan menghasilkan nilai akurasi sebesar 0,93(93%) dengan besar tweet senitimen positif sebanyak 60,3%, sementara jumlah tweet netral sebanyak 34,4% dan jumlah tweet negatif sebanyak 5,4%. Perbedaan penelitian yang dilakukan oleh yuliati dalam proses pelabelan yang dilakukan secara manual yang kemudian diverivikasi oleh balai bahasa sedangkan penelitian yang dillakukan oleh penulis dalam proses pelabelan data dilakukan menggunakan lexicon based di machine learning.

Penelitian selanjutnya yang dilakukan Puspasari & Subarkah (2022). Penelitian ini bertujuan untuk menghasilkan klasifikasi menggunakan model *naïve bayes.* Data yang digunakan dalam penelitian ini diambil dari beberapa vidio di yutub menggunakan menggunakan teknik data crawling menggunakan alat Coberry. Penelitian ini menunjukkan bahwa sentimen publik didominasi oleh sentimen negatif berdasarkan beberapa keraguan publik mengenai efek samping vaksin dan tindak lanjut pemerintah terkait pemulihan ekonomi negara.Perbedaan penelitian yang dilakukan oleh Puspita dalam proses pengambilan

data diambil dari youtube sedangkan penelitian yang dilakukan oleh penulis data yg digunakan diambil dari *twitter*.

B. Landasan Teori

1. Analisis Sentimen

Sentimen analisis atau bisa disebut juga *opinion mining*, adalah bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, sikap, dan emosi orang-orang terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya. *Opinion* atau pendapat adalah pusat dari semua aktifitas manusia karena merupakan pemberi pengaruh utama perilaku kita. Analisis sentimen dan Opinion miningterutama berfokus pada opini yang mengekspresikan atau menyiratkan sentimen positif atau negatif. Pada awal tahun 2000, analisis sentimen sudah mulai berkembang menjadi salah satu penelitian aktif dalam *naturallanguage processing*(NLP) (Kumala *et al.,* 2020).

Analisis sentiment merupakan salah satu cara untuk mengumpulkan pendapat orang banyak terhadap sesuatu seperti layanan public, isu, kinerja pemerintahan atau hal lain yang berkaitan. Analisis sentiment dapat digunakan sebagai salah satu cara untuk melakukan evaluasi terhadap layananyang telah diberikan. Analisis sentiment dapat dilakukan melalui berbagai cara salah satunya adalah dengan mengumpulkan pendapat orang banyakmelalui media social(Suryono *et al.,* 2018).

2. Vaksinasi Covid-19

Vaksin adalah produk biologi yang berisi antigen berupa mikroorganisme atau bagiannya atau zat yang dihasilkannya yang telah diolah sedemikian rupa sehingga aman, yang apabila diberikan kepada seseorang akan menimbulkan kekebalan spesifik secara aktif terhadap penyakit tertentu. Vaksinasi adalah proses di dalam tubuh, dimana seseorang menjadi kebal atau terlindungi dari suatu penyakit sehingga apabila suatu saat terpajan dengan penyakit tersebut maka tidak akan sakit atau hanya mengalami sakit ringan,

biasanya dengan pemberian vaksin. Vaksinasi COVID-19 adalah bagian penting dari upaya penanganan pandemi COVID19 yang menyeluruh dan terpadu meliputi aspek pencegahan dengan penerapan protokol kesehatan: menjaga jarak, mencuci tangan pakai sabun dan memakai masker (3M), vaksinasi COVID-19, dan 3T (Tes, Telusur, Tindak lanjut) (Sigalingging & Santoso, 2021). Ada berbagai macam vaksin yang masuk ke Indonesia dengan tingkat efikasi yang berbeda-beda tetapi memiliki tujuan yang sama. Macam-macam jenis vaksin diantaranya yaitu *sinovac, pfizer, vaksin covid Bio Farma, AztraZeneca, Sinophram, Moderna, Sputnik V, Jannsen,* dan *Convidecia*

3. Twitter

Twitter adalah sebuah media sosial dan layanan *microblogging* yang mengijinkan penggunanya untuk mengirimkan pesan *realtime*. Pesan ini populer dengan sebutan *tweet. Tweet* adalah sebuah pesan pendek dengan panjang karakter yang dibatasi hanya sampai 140 karakter. Dikarenakan keterbatasan karakter yang bisa dituliskan, sebuah *tweet* seringkali mengandung singkatan, bahasa slang maupun kesalahan pengejaan (Suharso, 2019).

4. Teks Mining

Text mining merupakan proses pengolahan data berupa teks dengan melalui beberapa tahapan untuk mendapatkan informasi penting. Data yang di proses adalah data yang tekstual dan sebelum melakukan *pemrosesan data mining* harus melakukan tahap *text preprocessing*. Tahap teks preprocessing terdiri dari *case folding, cleansing, tokenizing, stemming, dan stopword removal* (Pratama *et al.,* 2021).

Teks mining bertujuan menghasilkan informasi dari satu set dokumen. *Text Mining* mampu menghasilkan informasi melalui pemrosesan,

pengelompokan, dan analisis data-data tidak terstruktur dalam jumlah besar . *Teks mining* digunakan untuk mendapatkan informasi yang berguna dari serangkaian dokumen dengan sumber data pada teks yang memiliki format yang tidak terstruktur. Proses pengambilan informasi dalam teks mining dapat menghasilkan analisis perasaan yang secara emosional mengidentifikasi pernyataan jika positif atau negative. Objek teks mining merupakan dokumen tidak terstruktur atau semi terstruktur. Teks mining secara efektif mengekstrak informasi yang diperlukan dari sejumlah dokumen (Samsir *et al.,* 2021).

5. *Twitter* API (*Application Programming Interface*)

API Twitter atau *Application Programming Interface (API) twitter* adalah suatu program atau aplikasi yang disediakan oleh twitter untuk mempermudah developer lain dalam mengakses informasi yang ada di website *twitter*. Pendaftaran sebagai developer aplikasi twitter untuk menggunakan API twitter dapat dilakukan di lama *https://dev.twitter.com*. Setelah mendaftar developer akan mendapatkan c*onsumer key, consumer access, access token dan access token secret* yang akan digunakan sebagai syarat otentifikasi dari aplikasi yang akan kita bangun. Tujuan dari otentifikasi adalah untuk hak akses developer dalam mengunduh data yang ada di *twitter* (Sambodo *et al.,* 2016).

6. Crawling

Crawling adalah suatu teknik yang digunakan untuk mengumpulkan informasi yang ada dalam web (Saputra, 2017). *Crawling* bekerja secara otomatis, dimana informasi yang dikumpulkan berdasarkan kata kunci yang diberikan oleh peneliti. Untuk menerapkan teknik *twitter crawling* ini, pihak *Twitter* telah memberikan akses bagi pengguna untuk memanfaatkan *Twitter* API. Sehingga dengan memanfaatkan *Twitter API* tersebut, sehingga bisa

dengan mudah memperoleh data-data seperti *tweet,* data pengguna dan lainnya. Hasil data disimpan dalam sebuah file atau basis data.

7. Machine Learning

Machine learning merupakan aplikasi komputer dan algoritma matematika yang diambil dengan cara pembelajaran kata dan menghasilkan prediksi dimasa yang akan datang. Proses pembelajaran yang dimaksud adalah suatu usaha untuk memperoleh kecerdasan yang melalui tahap latihan (*training*) dan pengujian (testing). *Machine learning* terbagi menjadi 3 kategori yaitu:

a. *Supervised Learning*

Supervised learning merupakan metode klasifikasi dengan semua data diberikan label untuk mengklasifikasikan kelas yang tidak di kenal atau di ketahui. Dalam teknik ini dapat di kelompokan menjadi dua masalah yaitu klasifikasi dan regresi. Masalah klasifikasi adalah ketika variabel output berbentuk kategori atau pelabelan dan masalah regresi adalah ketika Variabel output merupakan nilai rill. Algoritma untuk klasifikasi yang digunakan dalam teknik ini yaitu *Super Vector Machine(SVM), Naïve Bayes Classifier(NBC), KNN, Tress Gradient Boosted(TGB), Random Tress(RT), dan Artificial Neural Networks(ANN).*

b. *Unsupervised Learning*

Unsupervised learning atau disebut dengan cluster karena dalam teknik ini tidak membutuhkan untuk diberikan label dalam data dan hasilnya tidak mengidentifikasi kelas yang telah di tentukan. Masalah dalam teknik ini dikelompokan lebih lanjut untuk maslah cluster dan asosiasi. Masalah Pengkelompokan (*clustering*) adalah untuk menemukan cluster yang ada didalam data dan masalah asosiasi

merupakan aturan yang menggambarkan sebagian besar data. Masalah dalam teknik ini biasanya digunakan dalam algoritma *k-means, Apriori, inedependent Subspace Analysis, dan BDSCAN.*

c. *Reinforcement Learning*

Reinforcement Learning merupakan teknik yang dinamis dimana konsepnya harus menyelesaikan tujuan tanpa pemberitahuan dari komputer secara eksplisit jika tujuan telah tercapai. Masalah dalam teknik ini diselesaikan dengan mempelajari pengalam baru melalui trial and error. Dalam teknik ini bedasarkan model pengambilan keputusan markov yang mencakup dua jenis yaitu metode berbasis model seperti algoritma *SARSA* yang dimana pertama kali mempelajari model, kemudian mendapatkan strategi yang optimal dari pengetahuanmodel tersebut. Dan metode kedua adalah Relevan model seperti algoritma *Temporal Difference* dan algoritma *Q-learning*.

8. NLTK

NLTK merupakan sebuah platform untuk membuat program Python yang berurusan dengan bahasa manusia . Menyediakan lebih dari 50 corpora dan *lexical resource seperti WordNet*, beserta dengan library pemrosesan teks yang sesuai untuk klasifikasi, tokenisasi, stemming, tagging, parsing, dan semantic reasoning. Hal ini membuat NLTK sangat cocok dan banyak digunakan sebagai tools untuk membuat program NLP(Setiawan *et al.,* 2019).

9. Algoritma *Naïve Bayes Multinomial*

Multinomial Naive Bayes merupakan sebuah metode yang bekerja dengan cara menghitung frekuensi setiap term pada dokumen. Sebagai contoh, frekuensi kata “jaringan” pada berita teknologi. Sehingga peran tokenisasi dalam *Multinomial Naive Bayes* ini sangat penting. Dalam *Multinomial Naive*

Bayes, dokumen urutan kejadian munculnya kata dalam dokumen tidak dipedulikan, jadi dokumen dianggap seperti "bag of words", sehingga setiap kata diolah menggunakan distribusi multinomial(Fauzi & Adinugroho, 2018).

Menurut (Amelia Rahman & Doewes, 2017) langkah klasifikasi *Naïve Bayes Multnomial* melalui beberapa tahap yaitu:

Langkah pertama yaitu menentukan probabilitas prior kelas c. Rumus menentukan probabilitas ditentukan pada persamaan (2.1).

$$P(c) = \frac{Nc}{N} \tag{2.1}$$

Keterangan:

Nc : Jumlah kelas c pada seluruh dokumenN

: Jumlah seluruh dokumen

Langkah selanjutnya yaitu menghitung probabilitas kata ke-n ditentukan dengan menggunakan teknik laplacian smoothing melalui persamaan (2.2):

$$P(tn \mid c) = \frac{count(tn, c) + 1}{count(c) + \mid V \mid} \tag{2.2}$$

Keterangan:

Count(tn,c) : Jumlah term tn yang ditemukan di seluruh data pelatihan dengan kategori c

count(c) : Jumlah term diseluruh data pelatihan dengan kategori c

V : Jumlah seluruh term pada data pelatihan

Sementara rumus Multinomial yang digunakan dengan pembobotan kata TF-IDF ditentukan pada persamaan (2.3).

K t rangan:
e e

$$P(tn \mid c) = \frac{Wtc + 1}{(\Sigma w' \text{€} vw' ct) + B'} \qquad (2.3)$$

Wct : Nilai pembobotan tfidf atau W dari term t di kategori c

$\sum W' \in V$W'ct : Jumlah total W dari keseluruhan term yang berada di Kategori c

B' : Jumlah W kata unik (nilai idf tidak dikali dengan tf) Pada seluruh dokumen

10. Algoritma Genetika

Algoritma *Genetik* merupakan prinsip dari genetika dan seleksi alam, yang pertma kali di kenalkan oleh seorang proffesor *University of Michigan, Ameka Serikat* yang bernama *John Holland* pada tahun 1970. Menurut buku(Muhammad, 2018) Algoritma genetika (GA) adalah metode Metaheuristik terinspirasi oleh proses seleksi alam. GA adalah proses yang terinspirasi oleh proses evolusi biologis berdasarkan teori evolusi Charles Darwin.

Algorima Genetika menurut (Wati, 2016)merupakan salah satu algoritma optimasi, yang diciptakan untuk meniru proses yang diamati dalam evolusi alam. Dalam proses Optimasi membutuhakan teknik *feature selection*. *Feature selection* adalah sebuah proses yang bisa digunakan pada machine learning dimana sekumpulan dari features yang dimiliki data digunakan untuk pembelajaran algoritma.

Langkah dalam penggunaan Algoritma Genetika menurut (Hermawanto, 2013)sebagai berikut:

a. Inisialisasi Populasi Awal

Tahap pertama yaitu inisialisai populasi awal dimana tahapan ini akan membentuk N kromosom, dan nilai gen secara random.

b. Evaluasi Nilai *Fitness*

Tahap selanjutnya yaitu evaluasi nilai *fitness*. Tujuan dari algoritma genetika yaitu mencari nilai fitness terbaik. Mencari nilai fitness dari setiap kromosom pada populasi pada penelitian ini menggunakan fungsi objektif, yang ditentukan pada persamaan (2.4).

$$f(x) = ((a + 2b + 3c + 4c + 5e) - 1) \tag{2.4}$$

c. Mencari Total Nilai *Fitness*

Mencari total nilai *fitness*, dapat dilihat pada persamaan (2.5).

$$totalfitnes\,s = f[1] + f[2] + f[3] + f[4] \tag{2.5}$$

d. Mencari Peluang untuk Masing-masing Kromosom

Dapat dilihat rumusnya pada persamaan (2.6).

$$P[i] = \frac{f[i]}{total} \tag{2.6}$$

e. Menghitung cumulative probability

Pada tahap ini nilai peluang akan dijumlahkan dengan peluang berikutnya, dengan rumus yang sudah ditentukan pada persamaan (2.7).

$$C[4] = P[1] + P[2] + P[3] + P[4] \tag{2.7}$$

f. Seleksi

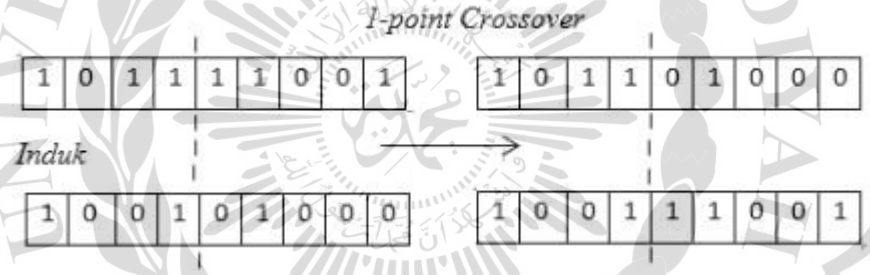
Proses seleksi menggunakan *roulete wheel*, dengan membangkitkan nilai R secara random. Terdapat beberapa metode dalam seleksi yangdapat digunakan yaitu, *roulette wheel selection, tournament selection* dan *rank-based selection.* Metode yang sering digunakan yaitu, seleksi *roulette*

wheel selection. Pada seleksi ini kromosom akan dipilih secara acak ditentukan dengan menghitung nilai kelayakan masing-masing

kromosom. Semakin besar nilainya maka semakin besar juga peluang kromosom tersebut untuk dipilih.

g. *Crossover*

Crossover (pindah silang) merupakan pertukaran antara kromosom sehingga akan membentuk kromosom baru dengan harapan lebih baik dari induknya. Proses ini semua kromosom mempunyai nilai gen secara random, nilai ≤ 0,25 akan dilakukan persilangan (Hermawanto, 2013). Teknik *crossover* ada 2 cara yaitu satu titik potong (*one point crossover*) dan dua titik potong (*n-point crossover*). Metode *Crossover* (pindah silang) yang sederhana yaitu metode pindah silang 1 titik. Dapat dilihat pada Gambar 2.1.



Gambar 2.1 Pindah Silang Satu Titik

h. Mutasi

Mutasi adalah proses penggantian gen dengan nilai terbaliknya. Gen 0 akan menjadi 1, dan gen 1 akan menjadi 0. Proses ini dilakukan secara acak pada lokasi gen tertentu pada individu yang dipilih untuk bermutasi. Sehingga akan menyebabkan terbentuknya kromosom baru (Muhammad, 2018).

i. Iterasi ke 2

Tahap ini akan menghitung nilai *fitness* kembali dengan fungsi objektif menggunakan kromosom baru yang telah di mutasi. Setelah tahap

ini selesai maka akan dilakukan perbandingan kromosom pada iterasi 1 dan iterasi 2. Sehingga akan mendapatkan nilai *fitness* terbaik dan solusi terbaik.

j. Etilisme

Siklus *algoritma genetika* diperbaiki oleh Zbigniew Michalewicz menambahkan proses *etilisme* pada tahap akhir. Etilisme adalah teknik yang digunakan untuk pertahankan individu yang memiliki nilai *fitness* tertinggi agar tidak mengalami kerusakan karena proses genetika bagi individu untuk bertahan hidup untuk generasi berikutnya.

11. Python

Python adalah bahasa pemrograman *interpretatif multiguna* dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. Python diklaim sebagai bahasa yang menggabungkan kapabilitas, kemampuan, dengan sintaksis kode yang sangat jelas, dan dilengkapi dengan fungsionalitaspustaka standar yang besar serta komprehensif (Syahrudin & Kurniawan,2018). Python mendukung multi paradigma pemrograman, utamanya, namun tidak dibatasi, pada pemrograman berorientasi objek, pemrograman imperatif, dan pemrograman fungsional. Salah satu fitur yang tersedia pada python adalah sebagai bahasa pemrograman dinamis yang dilengkapi dengan manajemen memori otomatis. Seperti halnya pada bahasa pemrograman dinamis lainnya, python umumnya digunakan sebagai bahasa script meski pada praktiknya penggunaan bahasa ini lebih luas mencakup konteks pemanfaatan yang umumnya tidak dilakukan dengan menggunakan bahasa script. *Python* dapat digunakan untuk berbagai keperluan pengembangan perangkat lunak dan dapat berjalan di berbagai platform sistem operasi. Salah satu notebook yang berbasis dalam bahasa pemograman python adalah *Google Colaboratory*, sehingga dalam penelitian ini menggunakan *Google Colaboratory* karena memberikan *machine learning* yang lebih banyak.

12. SMOTE (*Synthetic Minority Oversampling Technique*)

SMOTE merupakan teknik menyeimbangkan jumlah distribusi data sampel pada kelas minoritas dengan cara menyeleksi data sampel tersebut hingga jumlah data sampel menjadi seimbang dengan jumlah sampel pada kelas mayoritas (Kasanah *et al.,* 2019).