BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Hendri Hartono, Alyauma Hajjah, Yulvia Nora Marlim (2023), Prodi Teknik Informatika, Fakultas Ilmu Komputer Institut Bisnis dan Teknologi Pelita Indonesia, dengan judul penelitian " Penerapan *Metode Naive Bayes Classifier* Untuk Klasifikasi Judul Berita" Berita merupakan media informasi utama di dunia. Beragamnya berita yang disajikan oleh media digital saat ini, mencakup berbagai aspek seperti olahraga, hiburan, politik, kesehatan, keuangan, teknologi, dan lain-lain. Dengan beragamnya berita, maka diperlukan pengelompokan berita tersebut, demi memudahkan masyarakat dalam mendapatkan informasi yang diinginkan. *Naïve Bayes* adalah metode *klasifikasi* dengan melakukan *preprocessing* pada judul berita, kemudian menghitung *probabilitas* setiap kelasnya. Kelas yang dipakai dalam metode ini adalah kategori berita. Kategori berita meliputi, Olahraga, Hiburan, Kesehatan, Politik, dan Teknologi. Dari 500 data latih yang dijadikan acuan untuk menghitung probabilitas, setelah data uji dimasukkan maka akan dihitung probabilitas setiap kata yang digunakan dan akan menghasilkan suatu kategori, dari 50 data yang diuji sebanyak 43 dokumen yang berhasil sesuai dengan kategori yang tepat yaitu sebesar 86% dan sebanyak 7 dokumen dengan kesalahan kategori sebesar 14%.

Rambut pada kepala adalah salah satu bagian terpenting bagi seseorang khususnya secara estetika. Kesehatan kulit kepala dan rambut adalah suatu

keadaan kulit kepala dan rambut dengan tidak adanya penyakit mengganggu seperti ketombe, rambut rontok, kering, berminyak, kusam, dan sulit disisir (Zaradiya Audrey Tritania 2023).

Kerontokan rambut dapat memengaruhi fungsi biologis rambut terhadap tubuh, apabila melebihi batas normal. Selain itu, kerontokan rambut yang dapat mengakibatkan kebotakan dan memengaruhi kepercayaan diri seseorang sehingga menjadi masalah yang sangat mengkhawatirkan (Gede Dikka 2023).

Setiap tahunnya lebih dari 36 juta orang meninggal karena Penyakit Tidak Menular (PTM) (63% dari seluruh kematian). Lebih dari 9 juta kematian yang disebabkan oleh penyakit tidak menular terjadi sebelum usia 60 tahun, dan 90% dari kematian "dini" tersebut terjadi di negara berpenghasilan rendah dan menengah. Secara Global PTM penyebab kematian nomor satu setiap tahunnya adalah penyakit kardiovaskuler. Penyakit Kardiovaskuler adalah penyakit yang disebabkan gangguan fungsi jantung dan pembuluh darah, seperti: penyakit coroner, penyakit gagal jantung atau payuh jantung, hipertensi dan stroke. Diperkirakan sebanyak 17,3 juta kematian disebabkan oleh penyakit kardiovaskuler. (Nugroho, 2023)

Pada penelitian berjudul "Penggunaan *Algoritma Naive Bayes* dan *Particle Swarm Optimization (PSO)* untuk Mendeteksi Stroke" membandingan antara dua metode untuk meningkatkan prediksi stroke otak berdasarkan algoritma Naive Bayes. Pertama, evaluasi kinerja Naive Bayes secara independen tanpa metode optimisasi tambahan. Hasil eksperimen menunjukkan bahwa model Naive Bayes memiliki tingkat akurasi sebesar 86,21%, menunjukkan kemampuan model

tersebut dalam memprediksi stroke otak tanpa adanya penyetelan tambahan. Namun, meskipun performa ini menunjukkan potensi, masih ada ruang untuk peningkatan lebih lanjut. (Hasibuan, 2024)

2.2 Landasan Teori

2.2.1 Rambut Rontok

Menurut dokter (Pittara, 2023) Rambut rontok adalah lepasnya rambut secara berlebihan. Kondisi ini dapat mengakibatkan penipisan rambut atau kebotakan, baik sementara maupun permanen. Rambut rontok juga bisa terjadi sedikit demi sedikit atau banyak secara tiba-tiba. Jumlah rambut normal adalah sekitar 100.000 helai, dan akan lepas atau rontok sekitar 50–100 helai setiap harinya. Hal ini merupakan kondisi yang normal, karena rambut yang rontok akan digantikan dengan rambut baru.

Pertumbuhan rambut normal dapat dibagi ke dalam 3 fase, yaitu fase anagen, fase katagen, dan fase telogen. Pada fase anagen atau fase pertumbuhan, rambut akan tumbuh dan bertahan selama 2–8 tahun. Selama 2–3 minggu, rambut akan memasuki fase katagen atau fase transisi. Pada fase ini, rambut tidak tumbuh secara aktif. Setelah itu, rambut akan memasuki fase telogen atau fase istirahat. Pada fase ini, rambut akan mengalami kerontokan dan akan diganti dengan rambut baru 2–3 bulan setelahnya. Jika fase pertumbuhan rambut ini terganggu, rambut akan rontok hingga bisa berujung pada kebotakan. Selain itu, jika rambut yang memasuki fase telogen lebih banyak dari normal, rambut juga bisa rontok secara berlebihan. Kondisi ini disebut sebagai *telogen effluvium.* Meskipun

rambut rontok lebih sering dialami remaja dan orang dewasa, anak-anak juga bisa mengalaminya.

a. Penyebab Rambut Rontok

Banyak faktor yang dapat menyebabkan siklus pertumbuhan rambut terganggu, hingga berakibat pada rambut rontok. Rambut rontok yang terjadi secara tiba-tiba dapat terjadi akibat berbagai faktor, yaitu:

- Penyakit autoimun, misalnya pada *alopecia areata*
- Efek samping kemoterapi
- Perubahan hormon, misalnya saat persalinan atau karena *polycystic ovary syndrome* (PCOS)
- Efek samping obat-obatan, seperti obat penekan sistem imun (*imunosupresan*), obat asam urat, dan obat tekanan darah tinggi.

Sementara itu, rambut rontok yang terjadi secara bertahap paling sering disebabkan oleh faktor genetik atau keturunan. Selain itu, pola makan yang tidak sehat, seperti kurang protein dan zat besi, juga dapat mengurangi kesuburan akar rambut, sehingga menimbulkan kerontokan. Rambut rontok juga bisa terjadi saat keramas, terutama jika cara keramas salah atau karena penggunaan sampo maupun produk perawatan rambut yang kurang cocok dengan rambut dan kulit kepala.

b. Gejala Rambut Rontok

Gejala rambut rontok tergantung pada penyebabnya. Gejala ini dapat muncul secara tiba-tiba atau bertahap. Beberapa gejala tersebut adalah:

- Penipisan rambut di puncak kepala (ubun-ubun)

- Pitak
- Penipisan rambut yang merata di kepala
- Rambut rontok di seluruh tubuh
- Rambut mudah patah

c. Diagnosis Rambut Rontok

Diagnosis rambut rontok diawali dengan tanya jawab terkait gejala yang dialami pasien, serta riwayat penyakit pasien dan keluarganya. Setelah itu, dokter akan melakukan pemeriksaan fisik pada rambut dan kulit kepala.

Pada pemeriksaan fisik, dokter akan menarik lembut rambut pasien untuk melihat seberapa banyak rambut yang rontok. Jika diperlukan, dokter dapat melakukan pemeriksaan penunjang, seperti:

- Tes darah, untuk mendeteksi kondisi yang menyebabkan rambut rontok
- Biopsi kulit kepala, untuk mendeteksi apakah terjadi infeksi yang dapat menyebabkan rambut rontok

d. Pengobatan Rambut Rontok

Penanganan rambut rontok tergantung pada penyebabnya. Pada rambut rontok yang terjadi akibat perubahan hormon saat persalinan, lebatnya rambut akan kembali normal dalam kurun waktu 6−9 bulan pasca melahirkan.

Pada rambut rontok yang terkait dengan stres, dokter akan menyarankan pasien untuk menjalani konseling dan psikoterapi. Sementara jika rambut rontok terjadi akibat status gizi yang kurang baik, maka dokter akan memberikan saran tambahan asupan gizi dan multivitamin.

Penanganan medis lain dapat dilakukan saat seseorang mulai merasa penampilannya terganggu akibat rambut rontok. Beberapa metode penanganan yang dapat dilakukan untuk mengatasi rambut rontok adalah:

- Pemberian obat oles kulit kepala yang mengandung *minoxidil*
- Pemberian obat minum yang mengandung finasteride atau spironolactone
- Penggunaan sampo khusus rambut rontok
- Cangkok atau transplantasi rambut, untuk mengatasi kebotakan akibat rambut rontok

e. Komplikasi Rambut Rontok

Rambut rontok yang tidak ditangani dapat mengganggu penderitanya dalam melakukan kegiatan sehari-hari. Penipisan rambut dan adanya pitak yang disebabkan kerontokan rambut dapat terlihat orang lain sehingga membuat penderitanya merasa malu. Jika kondisi tersebut dibiarkan, penderitanya dapat mengalami komplikasi berupa penurunan kepercayaan diri, gangguan kecemasan, hingga depresi.

f. Pencegahan Rambut Rontok

Kerontokan rambut tidak selalu dapat dicegah, terutama yang terkait dengan faktor keturunan. Akan tetapi, pencegahan rambut rontok bisa dimulai dari perawatan rambut dengan rangkaian sampo yang mengandung krim Argan dan esens alpukat yang membantu menguatkan dan menjaga rambut tetap sehat.

Selain perawatan rambut, ada beberapa upaya yang dapat dilakukan untuk menjaga kesehatan rambut agar tercegah dari kerontokan:

- Jangan sering mewarnai rambut.

- Lindungi rambut dari paparan sinar matahari secara langsung dengan memakai topi dan payung ketika cuaca terik.
- Sisir rambut dengan benar.
- Pilih produk perawatan rambut yang sesuai dengan jenis kulit kepala dan rambut.

2.2.2 Data Mining

Data Mining adalah proses menemukan informasi yang berguna secara otomatis dalam repositori data yang besar. Teknik data mining digunakan untuk menjelajahi kumpulan data yang besar untuk menemukan pola baru dan berguna yang mungkin tidak diketahui. Teknik ini juga memberikan kemampuan untuk memprediksi hasil pengamatan di masa depan, seperti jumlah yang akan dibelanjakan pelanggan di toko online atau toko fisik (Tan, P.-N., Steinbach, M., & Kumar, V. 2019).

Data Mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah data mining kadang disebut juga knowledge discovery. Data Mining merupakan proses ataupun kegiatan untuk mengumpulkan data yang berukuran besar kemudian mengekstraksi data tersebut menjadi informasi– informasi yang nantinya dapat digunakan (Andriyanto, 2022).

Langkah-langkah proses pelaksanaan data mining dalam tiga aktivitas adalah:

1. Eksplorasi Data, terdiri dari aktivitas pembersihan data, transformasi data, pengurangan dimensi, pemilihan ciri, dan lain-lain.
2. Membuat model dan Pengujian Validitas Model, merupakan pemilihan terhadap model-model yang sudah dikembangkan yang cocok dengan kasus yang dihadapi. Dengan kata lain, dilakukan pemilihan model secara kompetitif.
3. Penerapan model dengan data baru untuk menghasilkan perkiraan dari kasus yang ada. Tahapan ini meruapakan tahapan yang menentukan apakah model yang telah dibangun dapat menjawab permasalahan yang dihadapi.

Komponen-Komponen Data Mining Komponen-komponen utama dari proses klasifikasi adalah:

1. Kelas, merupakan varibael tidak bebas yang merupakan label dari hasil klasifikasi.
2. Prediktor, merupakan variabel bebas suatu model berdasarkan dari karakteristik atribut data yang diklasifikasikan, misalnya merokok, minum-minuman beralkohol, tekanan darah, status perkawinan, dan sebagainya.
3. Set Data Pelatihan, merupakan sekumpulan data lengkap yang berisi kelas dan prediktor yang dilatih agar model dapat mengelompokan ke dalam kelas yang tepat.
4. Set Data Uji, berisi data-data baru yang akan dikelompokan oleh model guna mengetahui akurasi dari model yang telah dibuat.

Data mining memiliki dua fungsi utama (muttaqin, 2023):

1. Deskriptif: Memahami data dengan mengidentifikasi pola atau karakteristik yang ada.

2. Prediktif: Menggunakan pola yang ditemukan untuk memprediksi nilai atau tren di masa depan.

Berikut adalah beberapa tujuan utama dari data mining:

1. Mengidentifikasi pola dan hubungan

   Tujuan utama data mining adalah untuk mengidentifikasi pola dan hubungan dalam data yang sebelumnya tidak diketahui. Dengan mengidentifikasi pola dan hubungan ini, organisasi dapat mengambil tindakan yang tepat dan efektif.

2. Memprediksi perilaku dan tren

   Data mining juga digunakan untuk memprediksi perilaku dan tren di masa depan berdasarkan data historis. Dengan memprediksi perilaku dan tren ini, organisasi dapat mengambil tindakan yang tepat dan mengambil keuntungan dari peluang yang muncul.

3. Menemukan informasi yang berguna

   Data mining juga digunakan untuk menemukan informasi yang berguna dari data yang sebelumnya tidak terlihat. Informasi ini dapat digunakan untuk memperbaiki proses bisnis, meningkatkan keefektifan organisasi, atau meningkatkan layanan yang diberikan.

4. Mendukung pengambilan keputusan

   Data mining juga dapat digunakan untuk mendukung pengambilan keputusan. Dengan mengumpulkan dan menganalisis data, organisasi dapat membuat keputusan yang lebih baik dan berdasarkan fakta.

5. Meningkatkan efisiensi dan efektivitas Tujuan utama Data Mining adalah untuk meningkatkan efisiensi dan efektivitas sebuah organisasi atau proses. Dengan mengumpulkan dan menganalisis data, organisasi dapat mengidentifikasi area yang memerlukan perbaikan dan mengambil tindakan untuk meningkatkan efisiensi dan efektivitas

Beberapa metode umum dalam data mining meliputi:

1. Klasifikasi (Classification)**:** Mengelompokkan data ke dalam kategori yang telah ditentukan.
2. Regresi (Regression)**:** Memprediksi nilai numerik berdasarkan variabel lain.
3. Klastering (Clustering)**:** Mengelompokkan data berdasarkan kesamaan tanpa label yang telah ditentukan.
4. Asosiasi (Association Rule Mining)**:** Menemukan hubungan antar item dalam dataset, seperti analisis keranjang belanja.
5. Deteksi Anomali (Anomaly Detection)**:** Mengidentifikasi data yang tidak biasa atau menyimpang dari pola umum.

2.2.3 Pengertian Klasifikasi

Klasifikasi ialah proses mengkategorikan atau memberi label data atau objek baru berdasarkan kualitas tertentu. Teknik klasifikasi melibatkan pemeriksaan variabel yang dihasilkan dari data yang ada. Tujuan dari klasifikasi adalah untuk mengantisipasi kelas yang dihasilkan dari item yang tidak diketahui. Ada tiga tahap klasifikasi yaitu konstruksi model, aplikasi model serta evaluasi. Pembuatan model melibatkan pembuatan contoh memakai data pelatihan yang telah memiliki

atribut serta kelas. informasi ini lalu digunakan untuk memilih kelas data atau objek baru. Data tersebut kemudian dievaluasi untuk melihat keakuratan yang didapatkan dari pengembangan dan penerapan model pada data baru. Salah satu metode *classifier* yang dapat digunakan adalah metode *Naive Bayes* yang sering disebut dengan *Naive Bayes Classifier (NBC)*. *Naive Bayes Classifier (NBC)* adalah salah satu metode klasifikasi dan statistik pengklasifikasi yang dapat memprediksi peluang untuk menjadi anggota kelas. Menurut *NBC*, nilai atribut grup tidak bergantung pada nilai karakteristik lainnya. Keunggulan *NBC* ialah dasar namun akurat. Prosedur kategorisasi data dibagi menjadi dua tahap. Langkah pertama adalah berlatih dengan contoh *(Training Example)*. Tahap kedua adalah proses pengkategorian data yang belum diketahui kelasnya (Angga Pebdika, Ruli Herdiana, Dodi Solihudin 2023).

Klasifikasi adalah pekerjaan yang menilai suatu objek data agar masuk kedalam kelas tertentu dari sejumlah kelas yang sudah ada. Klasifikasi dibagi menjadi dua pekerjaan utama yaitu membangun model sebagai prototype yang disimpan sebagai memori dan menggunakan model untuk melakukan pengklasifikasian prediksi pada objek data lain agar diketahui terdapat dikelas mana objek data yang disimpan (Putri, Suparti, & Rahmawati, 2014).

2.2.4 Naive Bayes

Algoritma *Naïve Bayes* menggunakan teknik percabangan matematis dengan mencari probabilitas terbesar dari sebuah klasifikasi berdasarkan frekuensi dari setiap klasifikasi terhadap data training, yang sering disebut sebagai teori probabilistik. Rumus perhitungan *Naïve Bayes* adalah sebagai berikut:

$$\mathrm{P}\left(\frac{X}{Y}\right) = \frac{P\left(\frac{Y}{X}\right) \mathrm{x}\, \mathrm{P(X)}}{\mathrm{P(Y)}}$$

Pada rumus tersebut, P(X) merupakan probabilitas awal dari hipotesis X atau seberapa besar kemungkinan hipotesis tersebut benar tanpa mempertimbangkan data yang ada. Selanjutnya, P(Y|X)P(Y|X)P(Y|X) adalah probabilitas Y berdasarkan kondisi dari hipotesis X, yang merepresentasikan seberapa besar kemungkinan data Y terjadi jika hipotesis X benar. Terakhir, P(Y)P(Y)P(Y) adalah probabilitas Y atau kemungkinan terjadinya data tanpa mempertimbangkan hipotesis (Ratnawati & Sulistyaningrum, 2020). Pendekatan ini memungkinkan algoritma untuk menghitung P(X|Y)P(X|Y)P(X|Y), yang adalah seberapa besar kemungkinan hipotesis X benar mengingat data Y. Nilai probabilitas ini kemudian digunakan untuk menentukan kelas yang paling mungkin terjadi pada data Y. Naive Bayes mengasumsikan bahwa semua fitur dalam data independen satu sama lain, yang menyederhanakan penghitungan namun tetap memberikan hasil yang efektif di berbagai aplikasi, termasuk prediksi risiko obesitas (Witata & Triloka, 2023).

Naïve Bayes atau multinomial *naïve bayes* merupakan metode yang digunakan untuk mengklasifikasikan sekumpulan dokumen. Algoritma ini memanfaatkan metode *probabilitas* dan *statistik* yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Metode NB menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap pengujian (klasifikasi). Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan

probabilitas prior bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan *term* yang muncul dalam dokumen yang diklasifikasi.

Dalam *naïve bayes,* kemungkinan dokumen d berada di *class* c dihitung sebagai berikut (Manning, 2009)[2][2]:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (2.1)$$

dimana $P(t_k|c)$ adalah *conditional probability* dari fitur t_k yang terdapat dalam dokumen dari *class* c. Dapat diartikan, $P(t_k|c)$ adalah ukuran berapa banyak kemunculan fitur t_k memberikan kontribusi bahwa c adalah *class* yang benar. $P(c)$ adalah *prior probability* dari dokumen yang terdapat di *class* c. Jika fitur dari sebuah dokumen tidak memberikan *evidence* yang jelas untuk sebuah *class* dibandingkan dengan *class* lainnya, maka fitur dengan *prior probability* tertinggi yang akan dipilih. Token dalam d ($t_1, t_2, \ldots, t_{n_d}$) merupakan bagian dari *vocabulary* yang digunakan untuk klasifikasi dan n_d adalah jumlah token tersebut dalam d.

Tujuan utama dalam klasifikasi teks adalah menemukan *best class* untuk sebuah dokumen. *Best class* dalam *naïve bayes* adalah yang paling mungkin atau *maximum a posteriori* (MAP) *class* c_{map} :

$$c_{map} = \arg max_{c \in C} \acute{P}(c|d) = \arg max_{c \in C} \acute{P}(c) \prod_{1 \leq k \leq n_d} \acute{P}(t_k|c) \quad (2.2)$$

dimana arg max adalah *argument maximum* dan untuk P ditulis $\acute{P}$ karena tidak diketahui nilai sebenarnya dari parameter $P(c)$ dan $P(t_k|c)$.

Pada Persamaan (2.2), banyak *conditional probability* yang dikalikan, satu untuk masing-masing posisi $1 \leq k \leq n_d$. Hal ini dapat mengakibatkan masalah

underflow. Masalah *underflow* adalah masalah yang muncul pada dokumen yang memiliki jumlah kata yang sangat besar. Hasil perkalian dari nilai-nilai *conditional probability* dari seluruh kata yang berjumlah sangat besar akan membuat variabel *score* bernilai sangat kecil. Nilai *score* yang sangat kecil dapat menimbulkan kesalahan saat dilakukan proses perbandingan. Oleh karena itu, lebih baik untuk melakukan perhitungan dengan menambahkan logaritma probabilitas daripada mengalikan probabilitas. *Class* dengan nilai probabilitas tertinggi masih yang paling mungkin. Oleh karena itu maksimalisasi yang sebenarnya dilakukan dalam kebanyakan implementasi dari *naïve bayes* adalah:

$$c_{map} = \arg max_{c \in C} [\log \acute{P}(c) + \sum_{1 \leq k \leq nd} \log \acute{P}(t_k|c)] \quad (2.3)$$

untuk menghitung nilai dari $\acute{P}(c)$ adalah sebagai berikut:

$$\acute{P} = \frac{N_c}{N} \quad (2.4)$$

(N_c adalah dokumen yang berada di *class* c dan N adalah jumlah dokumen) Diperkirakan *conditional probability* $\acute{P}(t|c)$ sebagai frekuensi relatif dari fitur t dalam dokumen-dokumen di *class* c dapat dihitung dengan persamaan:

$$\acute{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \quad (2.5)$$

(T_{ct} adalah jumlah kemunculan fitur t dalam *training* dokumen dari *class* c).

Persamaan (2.3) memiliki interpretasi yang sederhana. Setiap kondisi parameter log $P(t_k|c)$ adalah bobot yang menunjukkan seberapa baik indikator t_k untuk c. Demikian pula $\acute{P}(c)$ adalah bobot yang menunjukkan frekuensi relatif c. Hasil penjumlahan log *prior probability* dan bobot fitur adalah ukuran tentang

berapa banyak kemunculan yang ada untuk dokumen di *class c* dan Persamaan (2.3) memilih *class* dengan *evidence* terbanyak.

Persamaan (2.5) akan menimbulkan masalah baru apabila fitur tidak ditemukan dalam *training set*. Fitur yang tidak ditemukan menyebabkan masalah pembagian dengan nol (*devision by zero*). Untuk mengatasi hal tersebut maka digunakan *add-one* atau *Laplace Smoothing*, seperti tampak pada Persamaan (2.6).

$$\acute{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T'_{ct}+1)} = \frac{T_{ct}+1}{(\sum_{t' \in V} T'_{ct})+B'} \quad (2.6)$$

dimana $B = |V|$ adalah jumlah fitur dalam *vocabulary*

2.2.5 Klasifikasi

Pengujian ini dilakukan untuk mengetahui kinerja Algoritma *C4.5* dan *Naïve Bayes* dalam melakukan klasifikasi, peneliti menggunakan teknik validasi split validation melalui tool rapid miner untuk mengetahui apakah pengaruh variabel yang muncul baik. Selain itu dilakukan uji validitas data set dibagi menjadi 2 bagian yaitu data training dan data testing. Total dari data set adalah 30 data set dengan 6 variabel data yaitu pelayanan, kualitas produk, promosi, harga, fasilitas, dan hasil keputusan. Untuk mengetahui nilai akurasi dan AUC (Area Under Curve) ditentukan enam variabel untuk mengetahui apakah sudah sesuai atau belum. Perbandingan dilakukan melalui empat variabel yaitu pelayanan, kualitas produk, fasilitas, dan hasil keputusan dimana dilakukan validasi split validation dengan perbandingan yang berbeda antara data training dan data testing. Berikut ini adalah penjelasan dari mengenai 3 kali pengujian. (Lestari, 2020)

1. 70% data training dan 30% data testing
2. 80% data training dan 20% data testing
3. 90% data training dan 10% data testing

Penggunaan metode *naïve bayes* terdiri dari dua fase, yaitu fase pelatihan dan fase pengujian. Berikut ini adalah tahap-tahap pada fase pelatihan dan fase pengujian:

a. Fase Pelatihan (*training*)**:**

Fase pelatihan adalah sebagai berikut:

1. Ekstrak seluruh data dari seluruh dokumen dalam dokumen latih kemudian buat tabel representasi dokumen latih.
2. Hitung *prior probability* $P\ class$ untuk setiap *class* dengan menggunakan rumus pada persamaan (2.4).
3. Hitung *conditional probability* dari semua kata dan *class* dengan menggunakan rumus pada persamaan (2.6).

b. Fase Pengujian (*testing)*:

Fase pengujian adalah sebagai berikut:

1. Ekstrak seluruh data dari seluruh dokumen dalam dokumen latih.
2. Hitung bobot (*score*) dari dokumen *d* yang termasuk dalam *class C* dengan menggunakan rumus pada persamaan (2.2).
3. Prediksi *class* dokumen uji dengan cara memilih *class* yang memiliki skor terbesar berdasarkan persamaan (2.3).

2.2.6 Google Colabs

Google Colaboratory atau lebih dikenal sebagai *Google Colab* adalah layanan berbasis cloud dari Google yang memungkinkan pengguna untuk menulis dan mengeksekusi kode *Python* langsung dari browser. Google Colab dibangun di atas *Jupyter Notebook*, sehingga mendukung teks naratif, visualisasi data, dan kode yang dapat dijalankan dalam satu dokumen interaktif.

Layanan ini *gratis* dan sangat populer di kalangan peneliti, pelajar, dan praktisi data science, machine learning, dan AI karena memberikan akses gratis ke *GPU* dan *TPU***,** serta integrasi langsung dengan *Google Drive***.**

Fungsi dan Tujuan Google Colab:

1. Menjalankan eksperimen *Machine Learning (ML)* dan *Data Science* tanpa memerlukan instalasi software di komputer.
2. Membuat dan berbagi dokumen interaktif yang memuat kode, teks, gambar, dan visualisasi.
3. Mendukung kolaborasi real-time antar pengguna, seperti Google Docs.
4. Memanfaatkan komputasi awan dengan dukungan GPU/TPU untuk pelatihan model yang berat.

Fitur-Fitur Utama Google Colab:

1. Berbasis Jupyter Notebook

   Mendukung file `.ipynb` dengan format teks dan kode terpisah.
2. Akses GPU dan TPU Gratis

   Bisa menggunakan akselerator untuk ML, dengan keterbatasan runtime.
3. Berbasis Cloud

Tidak memerlukan instalasi lokal, cukup browser dan koneksi internet.

4. Terhubung ke Google Drive

   Mudah untuk menyimpan dan membuka file di Google Drive.

5. Kolaborasi Online

   Beberapa pengguna bisa mengedit notebook yang sama secara bersamaan.

6. Instalasi Library Custom

   Mendukung pip install langsung dalam notebook.

7. Dukungan Visualisasi

   Cocok untuk matplotlib, seaborn, Plotly, dll.

Manfaat Google Colab:

1. Belajar dan praktek Python tanpa install apapun.
2. Eksperimen cepat untuk model ML tanpa harus beli hardware mahal.
3. Berbagi eksperimen/catatan dengan tim atau publik.
4. Membuat laporan interaktif untuk presentasi data.

Kekurangan dan Batasan:

1. Runtime Terbatas

   Maksimal aktif selama 12 jam (versi gratis), lalu terputus.

2. Privasi Data

   Semua dijalankan di cloud, tidak cocok untuk data sensitif.

3. GPU Terbatas

   Penggunaan GPU gratis dibatasi (kadang harus antre).

4. Koneksi Internet Wajib

   Tidak bisa digunakan offline.

2.2.7 Kriteria Evaluasi

Untuk permasalahan dalam *classification*, kriteria evaluasi yang biasa digunakan adalah *precision, recall* dan *accuracy* (Sheu, 2008). maka *precision*, *recall* dan *accuracy* dapat dihitung menggunakan persamaan (2.7), (2.8) dan (2.9) berdasarkan data pada Tabel 2.1. Selain itu digunakan *F1-measure* sebagai rata-rata dari *precision* dan *recall* (Manning dkk., 2009) yang dapat dilihat pada persamaan (2.10).

Tabel 2.1 Tabel Penilaian

CLASSIFIER

	Positif	Negatif
Positif	a	b
Negatif	c	d

1. Precision

Dalam *binary classification*, *precision* dapat disamakan dengan *positive predictive value* atau nilai prediktif yang positif. Rumus *precision* adalah:

$$Precision = (\frac{d}{b+d}) \times 100\% \quad (2.7)$$

2. Recall

Recall adalah pengambilan data yang berhasil dilakukan terhadap bagian data yang relevan dengan *query*. Rumus *recall* adalah:

$$Recall = (\frac{d}{c+d}) \times 100\% \quad (2.8)$$

3. Accuracy

Accuracy adalah persentase dari total *e-mail* yang benar diidentifikasi. Rumus *Accuracy* adalah:

$$Accuracy = (\frac{a+d}{total}) \times 100\% \qquad (2.9)$$

4. F1-Measure

Skor *F1-measure* mencapai nilai terbaik pada 1 dan skor terburuk pada 0. Rumus menghitung F1-measure adalah:

$$F1 = \frac{2 x Precision\ x\ Recall}{Precision+Recall} \qquad (2.10)$$