Kinerja Komparatif Optimasi Algoritma Naive Bayes dalam Klasifikasi Teks untuk Uji Klinis Kanker

Taslim[1], Susi Handayani[2], Fajrizal[3]
[1,2,3]Fakultas Ilmu Komputer Universitas Lancang Kuning
Pekanbaru, Indonesia
e-mail: [1]taslim@unilak.ac.id, [2]susi@unilak.ac.id, [3]fajrizal@unilak.ac.id

Diajukan: 17 Agustus 2023; Direvisi: 28 Agustus 2023; Diterima: 7 September 2023

Abstrak

Teknik klasifikasi teks dalam pemrosesan bahasa alami memegang peranan penting dalam mengelompokkan data digital ke dalam kategori yang telah ditentukan sebelumnya. Khususnya dalam bidang medis, klasifikasi teks klinis sangat penting untuk memahami dokumen medis, terutama teks klinis tentang kanker. Penelitian ini bertujuan untuk membandingkan kinerja tiga varian algoritma Naive Bayes *yaitu Multinomial, Bernoulli, dan Gaussian, pada data uji klinis kanker. Untuk mengoptimalkan kinerja algoritma, kami menggunakan pendekatan GridSearch dan cross-validation dengan k-fold (k=10). Pilihan algoritma memiliki pengaruh signifikan terhadap akurasi, presisi, recall, dan metrik kinerja lainnya. Melalui perbandingan varian* Naive Bayes*, kami dapat mengidentifikasi algoritma terbaik untuk dataset dan tugas klasifikasi teks klinis kanker. Hasil analisis menunjukkan bahwa Bernoulli* Naive Bayes *mencapai akurasi 0,79, presisi 0,88, dan recall 0,68. Sementara itu, Gaussian* Naive Bayes *mencapai akurasi 0,76, presisi 0,83, dan recall 0,65. Algoritma Multinomial* Naive Bayes *berhasil mencapai akurasi 0,80, presisi 0,84, dan recall 0,75. Penelitian ini memberikan panduan dalam memilih algoritma yang sesuai dengan tujuan dan prioritas klasifikasi. Hal ini dapat dikembangkan lebih lanjut dalam bahasa alami medis dan proses pengambilan keputusan medis. Melalui pengetahuan yang diperoleh dari penelitian ini, analisis teks medis dalam konteks klinis dapat dioptimalkan dengan lebih efektif.*

Kata kunci*: Klasifikasi teks,* Naive Bayes*, Gridsearch, Cross-validation, Uji klinis kanker.*

Abstract

The text classification technique in natural language processing plays a crucial role in categorizing digital data into predetermined categories. Especially in the medical field, clinical text classification is highly important for understanding medical documents, particularly clinical texts related to cancer. This study aims to compare the performance of three variants of Naive Bayes *algorithms – Multinomial, Bernoulli, and Gaussian – on clinical cancer test data. To optimize algorithm performance, we utilized the GridSearch approach and cross-validation with k-fold (k=10). The choice of algorithm significantly influences accuracy, presisi, recall, and other performance metrics. Through the comparison of* Naive Bayes *variants, we can identify the best algorithm for the dataset and the task of clinical cancer text classification. The results of the analysis indicate that Bernoulli* Naive Bayes *achieved an accuracy of 0.79, presisi of 0.88, and recall of 0.68. Meanwhile, Gaussian* Naive Bayes *reached an accuracy of 0.76, presisi of 0.83, and recall of 0.65. The Multinomial* Naive Bayes *algorithm successfully achieved an accuracy of 0.80, presisi of 0.84, and recall of 0.75. This research provides guidance in selecting an algorithm suitable for classification objectives and priorities. It can be further developed in the realm of medical natural language and clinical decision-making processes. Through the knowledge gained from this study, the analysis of medical texts in clinical contexts can be optimized more effectively.*

.

Keywords: *Text classification,* Naive Bayes*, Gridsearch, Cross-validation, Cancer clinical trials.*

1. Pendahuluan

Klasifikasi teks telah dikenal sebagai salah satu teknik dalam mengorganisir data digital[1] dan ini merupakan area paling vital dalam pemrosesan bahasa alami di mana data teks secara otomatis diurutkan ke dalam kumpulan kelas yang telah ditentukan sebelumnya. Aplikasi klasifikasi teks dalam pekerjaan

komersial seperti penyaringan spam, pengambilan keputusan, penggalian informasi dari data mentah, dan banyak aplikasi lainnya [2][1] dan umumnya dikerjakan dengan menggunakan algoritma *machine learning*[3]. Klasifikasi teks merupakan pekerjaan dasar dalam tugas pemrosesan bahasa alami yang bertujuan untuk mengatur dan mengklasifikasikan sumber teks, dan juga merupakan tautan utama untuk memecahkan masalah informasi teks[4]. Selama bertahun-tahun alat pemrosesan bahasa alami (NLP) telah berhasil digunakan untuk memproses informasi di berbagai bidang aplikasi. Salah satu bidang tersebut adalah kedokteran, di mana sebagian besar dokumen disajikan secara gratis dalam bentuk teks sebagai catatan perawatan primer, pemeriksaan fisik, laporan radiologi, catatan kemajuan, riwayat klinis, dan sebagainya[5].

Klasifikasi teks klinis merupakan masalah mendasar dalam pemrosesan bahasa alami medis[6]. Klasifikasi teks dalam domain kedokteran atau medis lebih menantang dibandingkan domain umum lainnya karena teks medis mencakup spesialisasi dan terminologi tingkat tinggi, dan untuk dapat mengidentifikasi *text* medis diperlukan pengetahuan ahli yag semantik dan terorganisir[7]. Data medis tekstual adalah sumber potensial untuk informasi klinis dan pengetahuan tersembunyi, dan analisis sistematis dari informasi ini dapat meningkatkan pemahaman dokter tentang tubuh manusia dan meningkatkan pengembangan pengetahuan kedokteran secara lebih baik[8]. Hal ini juga memberikan kemudahan pada dokter dalam pengambilan keputusan[9].

Salah satu data teks klinis yang memerlukan perhatian khusus adalah data klinis penyakit kanker karena penyakit ini merupakan penyebab utama kematian diseluruh dunia. Pada tahun 2020, sekitar 10 juta kematian atau hampir satu dari enam kematian disebabkan oleh penyakit kanker, sehingga menjadikannya penyebab kematian tertinggi di dunia. Kanker payudara, paru-paru, usus besar, rektus, dan prostat adalah jenis kanker yang paling umum[10]. Sampai hari ini belum ada terapi yang efektif untuk kanker dan menyebabkan kanker menjadi salah satu penyakit paling mematikan di dunia.. Penderita penyakit ini hanya bisa terhindar dari kondisi terburuk jika ditemukan pada stadium awal dan hanya terdapat sedikit peluang untuk bertahan hidup jika ditemukan pada tahap terakhir. Karena itu, diagnosis yang tepat dan dini sangat penting untuk pengobatan penyakit ini[11].

Ungkapan paling penting dalam uji klinis disebut kriteria kelayakan (layak dan tidak memenuhi syarat). Kriteria kelayakan adalah strategi utama untuk melakukan penyaringan target dalam sebuah uji klinis. Klasifikasi otomatis teks kriteria kelayakan uji klinis dengan menggunakan metode pembelajaran mesin meningkatkan efisiensi rekrutmen untuk mengurangi biaya riset klinis[12]. Karena persyaratan kelayakan untuk uji klinis biasanya ditulis dalam teks bebas, pemrosesannya memerlukan interpretasi komputer[13]

Beberapa penelitian telah dilakukan oleh para peneliti dalam mengklasifikasikan data klinis kanker. Jasmir melakukan penelitian mengenai klasifikasi teks data klinis kanker dengan menggunakan *deep neural networks* dan klasifikasi *fine-grained document*. Dalam penelitian ini, peneliti mengusulkan model klasifikasi baru dan mengevaluasi kinerja dengan berbagai algoritma *supervised learning*. Hasil dari penelitian ini menunjukkan bahwa *random forest* memberikan akurasi tertinggi, sedangkan metode *multilayer perceptron* memberikan akurasi terendah. Penelitian ini menyimpulkan bahwa *classifiers* dapat dilatih dan diuji menggunakan protokol uji klinis yang tersedia secara bebas[13]. Shang Gao menggunakan algoritma BERT, sebuah model pemrosesan bahasa alami yang terkenal, digunakan untuk mengklasifikasikan dokumen teks medis yang panjang. Peneliti mengajukan empat metode untuk mengadaptasi BERT agar sesuai dengan teks yang panjang dan membandingkannya dengan model-model sederhana. Hasil penelitian menunjukkan bahwa BERT seringkali memiliki hasil yang lebih rendah dibandingkan model sederhana ini dalam tugas mengklasifikasikan teks medis. Peneliti mengusulkan tiga strategi untuk meningkatkan hasil BERT, termasuk memecah dokumen panjang, menggunakan pengelompokan tertinggi, dan mengaplikasikan perhatian berlabel ganda. Mereka juga membandingkan strategi-strategi ini dengan model dasar dan mengevaluasi hasilnya pada dua set data yang berbeda[14].

Penelitian ini bertujuan untuk melakukan komparasi klasifikasi kriteria kelayakan uji klinis kanker dengan menggunakan salah satu algoritma klasifikasi yang populer yaitu Naïve Bayes[15] yang digunakan untuk memprediksi sampel yang tidak berlabel[16]. Naive Bayes sering digunakan sebagai *baseline* dalam klasifikasi teks karena cepat dan mudah diimplementasikan[17]. Dalam penelitian ini analisis komparatif klasifikasi teks dilakukan dengan tiga variasi algoritma Naïve Bayes yaitu multinomial Naïve Bayes, Gausian Naïve Bayes dan Bernoulli Naïve Bayes. Ketiga algoritma ini akan dianalisa berdasarkan metrik kinerja yaitu akurasi, presisi, *recall*, dan skor *F1 Score.*

Tiga algoritma Naive Bayes yaitu Bernoulli, Gaussian, dan Multinomial masing-masing memiliki kekuatan yang berbeda dan cocok untuk jenis data yang berbeda[18], yang membuat mereka menarik untuk dibandingkan dalam tugas klasifikasi teks seperti klasifikasi teks kanker. Bernoulli Naive Bayes yang bagus dalam menangani atribut boolean atau biner dimana algoritma ini bekerja dengan cara memodelkan

keberadaan atau ketiadaan fitur[19]. Multinomial Naive Bayes baik dalam menangani nilai diskrit. Algoritma Ini memodelkan jumlah frekuensi kemunculan kata dalam dokumen[20]. Gaussian Naive Bayes biasanya digunakan untuk data kontinu dan data dengan distribusi normal[21]. Membandingkan algoritma pada tugas klasifikasi teks kanker dapat membantu mengidentifikasi algoritma mana yang paling cocok untuk tugas spesifik[22][23], sehingga dapat memberikan pengetahuan tentang bagaimana distribusi dan karakteristik data mempengaruhi kinerja algoritma. Perbandingan ini dapat membantu dalam pemilihan algoritma yang tepat sehingga berpotensi meningkatkan akurasi dan efektivitas klasifikasi.

2. Metode Penelitian

Fokus utama dalam penelitian ini adalah untuk melakukan analisis kinerja tiga algoritma klasifikasi Naive Bayes yang telah terbukti efektif dalam tugas klasifikasi teks. Algoritma-algoritma ini meliputi Multinomial Naive Bayes, Gaussian Naive Bayes, dan Bernoulli Naive Bayes. Tujuan dari penelitian ini adalah untuk mengidentifikasi dan memahami perbedaan performa ketiga algoritma ini ketika dihadapkan pada tugas klasifikasi teks, khususnya dalam konteks kelayakan uji klinis kanker.

Dalam rangka mencapai tujuan tersebut, kami menggunakan sebuah *dataset* publik uji klinis kanker. *Dataset* ini diperoleh dari *platform* Kaggle. *Dataset* ini terdiri dari dua label kelas yang memiliki makna penting dalam analisis ini. Label kelas 0 mengindikasikan "*not eligible*" (tidak memenuhi syarat), sedangkan label kelas 1 mengindikasikan data yang "*eligible*" (memenuhi syarat). Penentuan label ini memberikan kerangka dasar bagi penilaian performa algoritma dalam memprediksi kelayakan peserta uji klinis.

Dalam melakukan eksperimen, kami mengalokasikan 80% dari total *dataset* sebagai data latih. Data latih ini digunakan untuk melatih algoritma-algoritma klasifikasi, mengajarkan mereka untuk mengenali pola-pola yang terkait dengan kriteria kelayakan data. Sisanya, yaitu 20% dari *dataset*, digunakan sebagai data uji untuk menguji kemampuan algoritma dalam membuat prediksi yang akurat. Dengan pendekatan ini, kami dapat mengukur performa relatif dari ketiga algoritma dan mengidentifikasi apakah ada perbedaan signifikan dalam kemampuan mereka dalam melakukan klasifikasi teks pada konteks uji klinis kanker. Proses pembersihan data meliputi tokenisasi dan lemmatisasi, untuk mengubah teks mentah menjadi bentuk yang lebih terstruktur.

Ekstraksi ciri dilakukan dengan menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) yang merupakan teknik statistik yang populer dalam pengambilan informasi dan pemrosesan bahasa alami yang berfungsi untuk mengukur signifikansi suatu istilah di dalam dokumen dalam kaitannya dengan korpus, atau kelompok, dokumen. Prosedur vektorisasi teks mengubah kata-kata dalam dokumen teks menjadi angka signifikansi. Metode ini bekerja dengan mengalikan *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) dari sebuah kata untuk menentukan skornya[24]. Adapun persamaan TF-IDF dapat dilihat pada persamaan berikut.

$$TF = \text{Frekuensi istilah muncul dalam dokumen} \tag{1}$$

$$IDF = \log\left(\frac{\text{jumlah dokumen dalam korpus}}{\text{jumlah dokumen dalam korpus berisi istilah tersebut}}\right) \tag{2}$$

$$TF - IDF \tag{3}$$

2.1. Multinomial Naïve Bayes

Multinomial Naive Bayes (MNB) merupakan algoritma klasifikasi yang berdasarkan pada teori Bayes dan digunakan untuk mengklasifikasikan data dengan fitur-fitur yang bersifat diskrit. Asumsi dasar dari MNB adalah bahwa fitur-fitur tersebut adalah independen ketika diberikan suatu kelas C, dan probabilitas fitur-fitur ini dihitung menggunakan aturan Bayes[25]. Dalam konteks klasifikasi teks, MNB sering diterapkan untuk mengklasifikasikan dokumen atau teks ke dalam kategori atau kelas yang sudah ditetapkan sebelumnya. Fitur-fitur yang digunakan dalam MNB dapat berupa kata-kata atau token-token yang muncul dalam teks. Asumsi independensi dalam MNB menyatakan bahwa setiap kata atau token

dalam dokumen dianggap independen dari kata atau token lainnya, walaupun pada kenyataannya, ada ketergantungan antara kata-kata dalam teks yang bisa terjadi.

2.2. Gausian Naïve Bayes

Gaussian Naive Bayes adalah metode klasifikasi yang memiliki aplikasi yang signifikan dalam analisis teks. Algoritma ini merupakan salah satu variasi dari Naive Bayes yang dirancang khusus untuk mengatasi data teks dengan fitur-fitur yang memiliki distribusi kontinu, seperti nilai numerik atau atribut berkelanjutan. Dalam konteks klasifikasi teks, setiap dokumen direpresentasikan sebagai vektor fitur, di mana setiap elemen vektor mewakili sebuah atribut atau fitur dalam teks. Gaussian Naive Bayes bekerja dengan asumsi bahwa fitur-fitur dalam setiap kelas mengikuti distribusi Gaussian (distribusi normal). Dengan kata lain, algoritma ini mengasumsikan bahwa setiap fitur memiliki rata-rata dan variasi yang dapat digunakan untuk menghitung probabilitas bahwa suatu dokumen termasuk dalam suatu kelas berdasarkan distribusi fitur-fiturnya[26]. Secara matematis, dalam Gaussian Naive Bayes, akan menghitung probabilitas posterior dari kelas C berdasarkan fitur-fitur x yang dimiliki oleh suatu dokumen. Persamaan umum untuk menghitung probabilitas posterior ini adalah:

$$P(C|x) = \frac{\mathrm{P}(x)|\mathrm{C}).\,\mathrm{P}(\mathrm{C})}{\mathrm{P}(x)} \tag{4}$$

Dalam analisis klasifikasi teks pada uji klinis kanker, konsep probabilitas memegang peran sentral dalam menginformasikan keputusan klasifikasi. Penggunaan probabilitas posterior (P(C|x)) bertujuan untuk menetapkan kelas berdasarkan fitur yang diamati. Sementara itu, probabilitas P(x|C) mengindikasikan kemunculan fitur dalam kelas dengan distribusi Gaussian, dan P(C) mewakili probabilitas prior dari kelas tersebut. P(x) menggambarkan probabilitas fitur dalam seluruh data. Integrasi probabilitas ini bertujuan untuk meningkatkan akurasi klasifikasi teks pada uji klinis kanker serta mendapatkan pemahaman yang lebih mendalam dari data yang ada. Dalam algoritma ini, probabilitas P(x|C) dihitung dengan memanfaatkan fungsi distribusi Gaussian.

$$P(x|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \,.\, e - \frac{z-\mu^2}{2\sigma^2} \tag{5}$$

Dimana μ adalah rata-rata dari fitur dalam kelas C, σ adalah deviasi standar, dan x adalah nilai fitur yang diamati. Integrasi konsep probabilitas dalam Gaussian Naive Bayes membantu dalam peningkatan kinerja klasifikasi teks dalam konteks uji klinis kanker.

2.3. Bernoulli Naïve Bayes

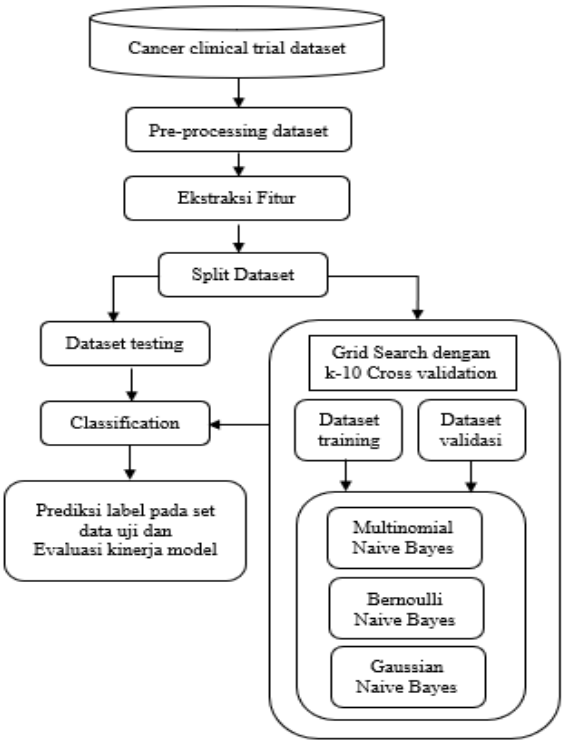
Terdapat kesamaan antara Bernoulli Naive Bayes dan model multinomial dalam konteks klasifikasi teks. Keduanya merupakan pendekatan yang populer untuk menangani tugas klasifikasi teks. Namun, terdapat perbedaan signifikan dalam pendekatan keduanya. Pendekatan Bernoulli Naive Bayes juga digunakan dalam klasifikasi teks, namun dengan fokus yang berbeda dari model multinomial. Pada dasarnya, Bernoulli Naive Bayes tertarik pada keberadaan atau ketiadaan suatu istilah dalam dokumen yang sedang dipertimbangkan. Pendekatan ini tidak memperhitungkan frekuensi istilah tersebut dalam dokumen, tetapi lebih fokus pada informasi biner apakah istilah tersebut hadir atau tidak hadir[27]. Dalam model Bernoulli Naive Bayes, setiap dokumen direpresentasikan sebagai vektor biner, di mana setiap elemen vektor mengindikasikan apakah suatu istilah hadir atau tidak hadir dalam dokumen. Oleh karena itu, perhitungan probabilitas dalam Bernoulli Naive Bayes dilakukan berdasarkan keberadaan atau ketiadaan istilah, bukan frekuensinya.

2.4. Optimasi

Selanjutnya dilakukan optimasi parameter menggunakan *GridSearchCV* untuk mencari parameter terbaik yang menghasilkan kinerja optimal untuk setiap algoritma. Performa algoritme dinilai dengan menggunakan metrik evaluasi standar seperti akurasi, presisi, *recall*, dan *F1 Score*. Pemanfaatan teknik *cross-validation* juga digunakan untuk mengurangi *overfitting* dan memberikan hasil yang stabil. Dalam penelitian ini, juga dipertimbangkan interpretasi hasil dari ketiga algoritme untuk memahami karakteristik dan keunggulan masing-masing algoritme dalam mengklasifikasikan data teks dalam domain data uji klinis kanker.

2.5. Alur logika Program

Logika program memberikan pandangan yang lebih rinci terhadap serangkaian tahapan dalam proses klasifikasi teks. Alur logika ini mencakup serangkaian tahapan yang penting dalam proses klasifikasi teks, dimulai dari pengumpulan dan preprocessing data hingga akhirnya mendapatkan hasil klasifikasi yang diinginkan. Alur logika program komparasi algoritma Naïve Bayes untuk klasifikasi teks dapat dilihat pada gambar 1 berikut.



Gambar 1. Alur logika *text classification*.

3. Hasil dan Pembahasan

Dalam bagian ini, kami akan memberikan paparan rinci tentang proses dan hasil dari setiap algoritma Naive Bayes yang kami teliti. Selain sekedar menghitung metrik performa, kami akan menggali lebih dalam untuk mengungkap solusi yang ditemukan serta temuan yang relevan dengan konteks latar belakang dan rumusan masalah penelitian.

3.1. Pembersihan Data dan Ekstraksi Fitur

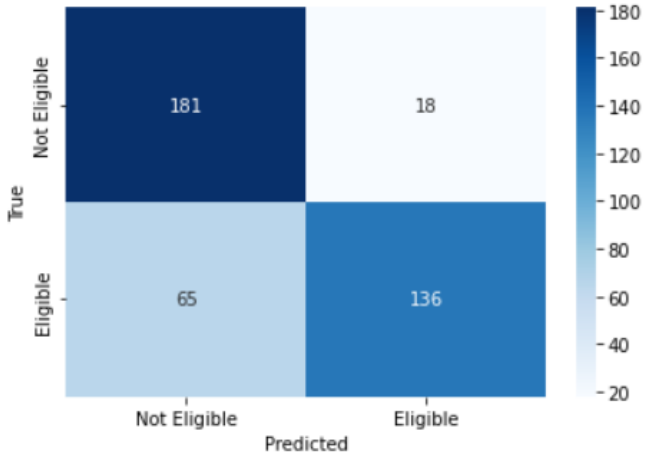
Pertama-tama, kami melakukan pembersihan data teks yang berasal dari *dataset* uji klinis kanker dengan menggunakan metode lemmatisasi. Langkah ini sangat penting karena membantu mereduksi dimensi fitur dan menghasilkan representasi tekstual yang lebih konsisten. Data kemudian dibagi menjadi data latih dan data uji dalam perbandingan 80:20. Proses ekstraksi fitur melibatkan dua pendekatan utama, yaitu CountVectorizer dan TF-IDF. CountVectorizer mengubah teks menjadi vektor biner sementara TF-IDF memberikan bobot pada kata-kata berdasarkan frekuensi kemunculannya dalam *dataset*.

3.2. Klasifikasi menggunakan Algoritma Bernaouli Naive Bayes (BNB)

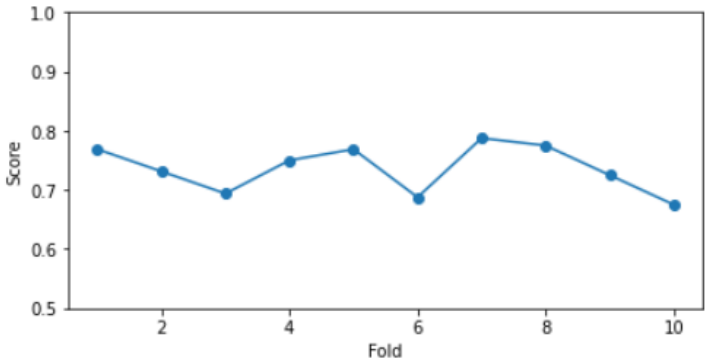
Hasil dari klasifikasi dengan menggunakan metode Bernoulli Naive Bayes menunjukkan variasi dalam nilai cross-validation di setiap iterasi, dengan rentang antara 0.675 hingga 0.7875. Rata-rata nilai cross-validation adalah 0.73625, menunjukkan bahwa model ini memiliki kinerja yang konsisten dalam berbagai pengujian. Dengan melakukan optimisasi parameter menggunakan *GridSearchCV*, kami berhasil mengidentifikasi parameter terbaik untuk model ini, yaitu {'alpha': 2.0}, yang menghasilkan akurasi *cross-validation* tertinggi sebesar 0.74. Pada tahap pengujian akhir dengan menggunakan data uji, model Bernoulli Naive Bayes berhasil mencapai akurasi sebesar 0.79.

Namun dalam pengujian ditemukan bahwa model ini memiliki tingkat presisi yang tinggi (0.88), menunjukkan bahwa mayoritas data yang diklasifikasikan sebagai positif benar-benar positif. Namun, tingkat *recall* (0.68) mengindikasikan bahwa kemampuan model dalam mendeteksi keseluruhan data positif masih perlu perbaikan. F1 *Score* (0.77), yang menggabungkan presisi dan *recall*, menunjukkan bahwa model ini memiliki keseimbangan yang baik dalam klasifikasi.

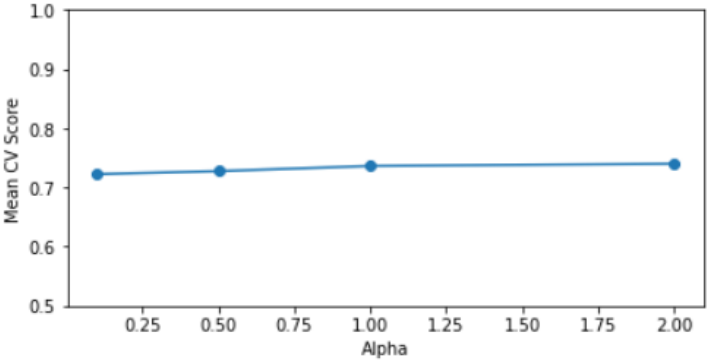
Pada analisis ini, kami menyimpulkan bahwa model Bernoulli Naive Bayes memiliki potensi yang baik untuk mengklasifikasikan teks dalam *dataset* uji klinis kanker. Kami melihat solusi untuk meningkatkan *recall* dengan penerapan strategi tertentu, dan temuan ini sangat sesuai dengan konteks penelitian kami. Untuk grafik *confusion matrix*, *cross-validation* dan *GridSearh* hasil pengujian dapat dilihat pada gambar 2,3 dan 4 berikut.



Gambar 2. *Confusion matrix* klasifikasi teks dengan BNB



Gambar 3. Grafik nilai *cross-validation* klasifikasi teks dengan BNB



Gambar 4. Grafik hasil *GridSearch* klasifikasi teks dengan BNB

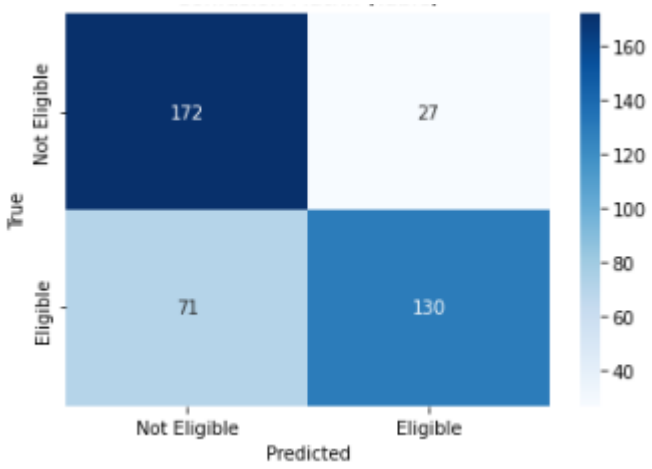
3.3. Klasifikasi menggunakan Algoritma Gaussian Naive Bayes (GNB)

Hasil analisis klasifikasi teks menggunakan algoritma Gaussian Naive Bayes menghadirkan temuan yang menarik dan relevan. Dalam tahap *Cross-Validation*, variasi akurasi mencakup kisaran 0.5125 hingga 0.65625 pada setiap lipatan validasi, dan rata-rata *Mean Cross-Validation Score* sekitar 0.595625. Variabilitas performa model ini dalam mengklasifikasikan data pada setiap iterasi validasi mencerminkan kompleksitas esensi permasalahan dalam klasifikasi teks uji klinis kanker.

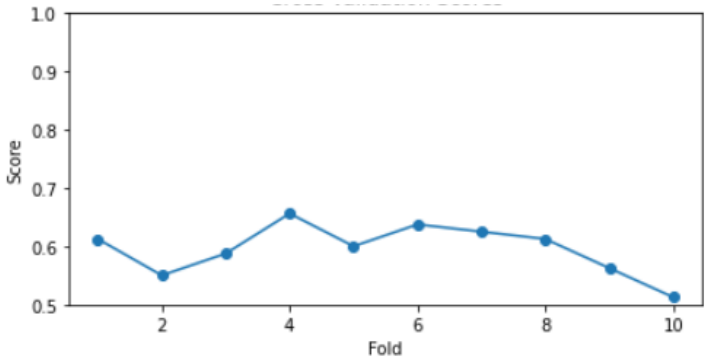
Dalam upaya mengejar optimalitas performa melibatkan *tuning hyperparameter* melalui pendekatan *GridSearchCV*. Hasilnya berhasil mengungkapkan bahwa parameter *'var_smoothing'* dengan nilai 0.01 menjadi pilihan terbaik, dan menghasilkan CV *Score* optimal sekitar 0.689375. Penerapan parameter ini secara signifikan meningkatkan kemampuan model dalam mengklasifikasikan data dengan lebih baik.

Ketika model diujicobakan pada data uji, model mencapai tingkat akurasi sebesar 0.76. Skor ini merefleksikan kompetensi model dalam mengenali dengan benar sekitar 76% data baru dengan tingkat tingkat presisi sebesar 0.83, yang menunjukkan bahwa model GNB memiliki kemampuan untuk memberikan prediksi yang akurat terhadap kelas yang ditargetkan. Meskipun demikian, tingkat *recall* dari model sebesar 0.65 menandakan bahwa model cenderung lebih baik dalam mengenali sebagian besar data yang sebenarnya positif. Namun, masih terdapat ruang untuk perbaikan dalam mendeteksi keseluruhan data positif.

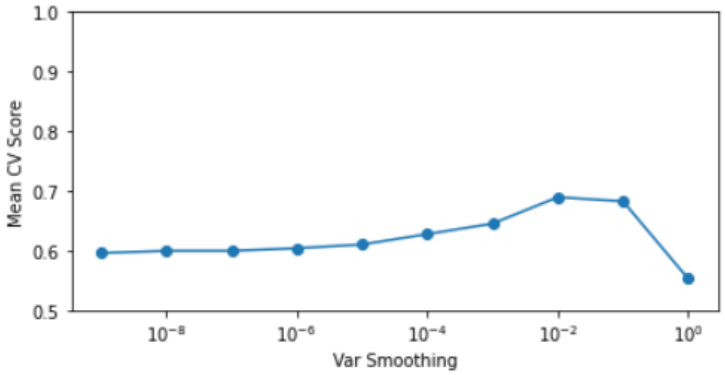
Melalui nilai F1 *Score* sebesar 0.73, terlihat bahwa model telah berhasil mencapai keseimbangan yang memadai antara *presisi* dan *recall* dalam melakukan klasifikasi teks pada data ini. Temuan ini menandakan pentingnya pengambilan kompromi antara mengurangi kesalahan positif dan negatif dalam konteks aplikasi ini. Solusi yang berhasil kami identifikasi, termasuk strategi *tuning* parameter dan pendekatan lebih cermat dalam memproses data positif, memiliki relevansi langsung dengan tantangan yang dihadapi dalam klasifikasi teks.*Confusion matrix, grafik nilai cross-validation* dan grafik hasil *GridSearch* dapat dilihat pada gambar 5, 6 dan 7 berikut.



Gambar 5. *Confusion matrix* klasifikasi teks dengan GNB



Gambar 6. Grafik nilai *cross-validation* klasifikasi teks dengan GNB

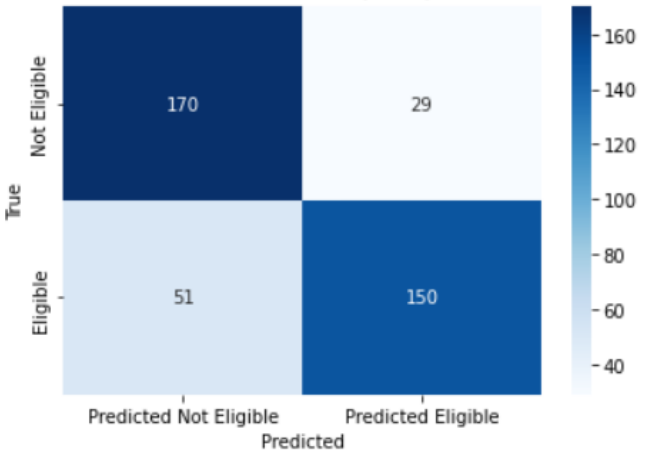


Gambar 7. Grafik hasil *GridSearch* klasifikasi teks dengan GNB

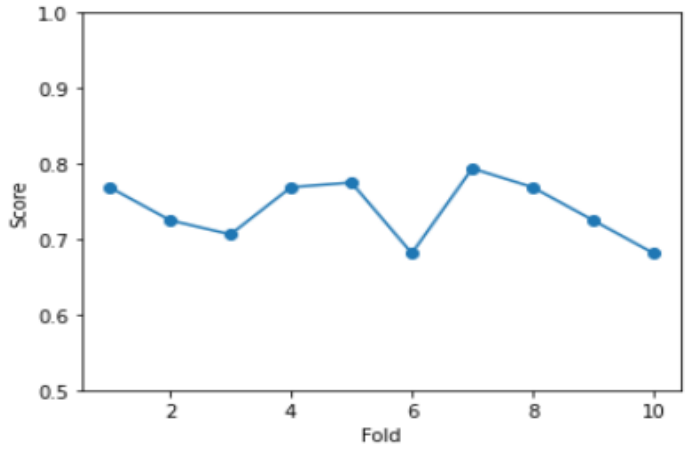
3.4. Klasifikasi menggunakan Algoritma Multinomial Naive Bayes (MNB)

Algoritma Multinomial Naive Bayes (MNB) bekerja berdasarkan pada teorema Bayes dengan asumsi independensi antara fitur. Dalam konteks klasifikasi teks, ini berarti bahwa keberadaan suatu kata dalam dokumen tidak mempengaruhi keberadaan kata lainnya. Melalui proses *Cross-Validation*, model ini menunjukkan konsistensi dengan akurasi yang berkisar antara 0.68125 hingga 0.79375 pada setiap lipatan validasi. Rata-rata nilai *Cross-Validation* mencapai sekitar 0.739375, mencerminkan konsistensi serta ketepatan model dalam melakukan klasifikasi terhadap data yang beragam.

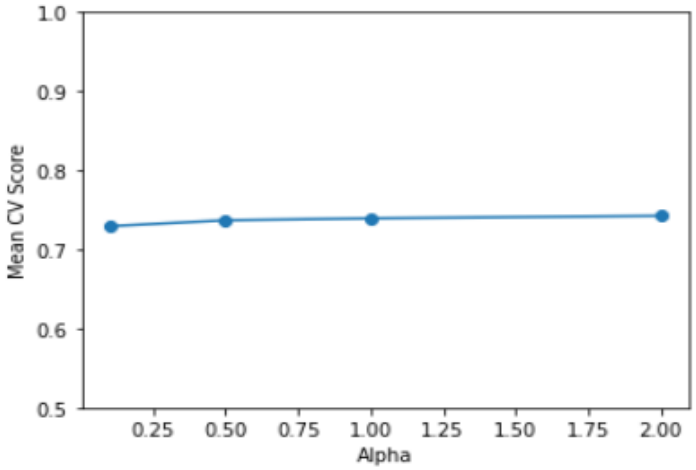
Selama tahap optimisasi *hyperparameter* dengan menggunakan *GridSearchCV*, parameter terbaik yang diidentifikasi adalah alpha dengan nilai 2.0. Nilai ini menghasilkan nilai *cross-validation* tertinggi sebesar 0.7425, menunjukkan kemampuan model dalam mengatasi variasi data. Pada tahap implementasi model pada data testing, model menunjukkan tingkat akurasi sekitar 0.80, mencerminkan kemampuan model dalam mengenali dengan benar sekitar 80% data baru. Selain itu, pengukuran presisi dengan nilai 0.84 menunjukkan kecakapan model dalam memberikan prediksi yang akurat terhadap kelas yang dituju. Model juga menunjukkan *recall* sebesar 0.75, menandakan kemampuan model dalam mengidentifikasi sebagian besar data positif dengan tepat. Nilai F1 *Score* sebesar 0.79 mencerminkan keseimbangan antara *presisi* dan *recall*, yang menandakan performa keseluruhan model yang baik dalam melakukan klasifikasi teks pada *dataset* ini. Grafik yang dihasilkan dari proses ini dapat memberikan visualisasi yang lebih baik tentang bagaimana model bekerja dan bagaimana performanya pada setiap tahap validasi silang dan pengujian. Dengan demikian, hasil ini menunjukkan bahwa algoritma Multinomial Naive Bayes dapat digunakan sebagai metode yang efektif untuk klasifikasi teks. *Confusion matrix*, grafik nilai *cross-validation* dan grafik hasil *GridSearch* dapat dilihat pada gambar 8, 9 dan 10 berikut.



Gambar 8. *Confusion matrix* klasifikasi teks dengan MNB



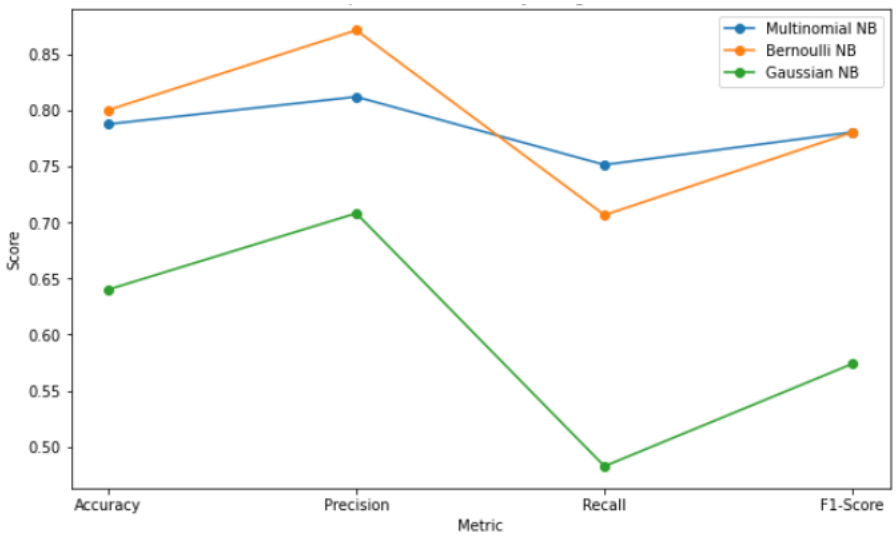
Gambar 9. Grafik nilai *cross-validation* klasifikasi teks dengan MNB



Gambar 10. Grafik hasil *GridSearch* klasifikasi teks dengan MNB

3.5. Komparasi hasil

Dalam menganalisis perbandingan hasil dari tiga algoritma klasifikasi teks Naive Bayes pada *dataset* uji klinis kanker, kami telah melakukan evaluasi terhadap performa masing-masing model. Dalam tulisan ini, kami akan memaparkan temuan yang signifikan dari perbandingan kinerja Model Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), dan Gaussian Naive Bayes (GNB) berdasarkan beberapa aspek, termasuk akurasi, presisi, dan *recall*. Adapun gambaran umum terhadap hasil perbandingan dapat dilihat pada gambar 11.



Gambar 11. Grafik komparasi hasil klasifikasi teks dengan MNB, BNB dan GNB

4. Kesimpulan

Hasil komparasi tiga algoritma klasifikasi teks dalam konteks *dataset* uji klinis kanker mengungkapkan berbagai temuan penting. Pertama, algoritma Bernoulli Naive Bayes (BNB) menunjukkan konsistensi dalam kinerja, dengan tingkat presisi yang tinggi, yang sangat berharga dalam menghindari kesalahan dalam mendiagnosis kasus positif kanker. Namun, tantangan terletak pada kemampuan model dalam mendeteksi secara keseluruhan kasus positif kanker, sehingga perlu meningkatkan *recall*. Kedua, algoritma Gaussian Naive Bayes (GNB) menghadapi variasi performa yang mengindikasikan kompleksitas dalam mengklasifikasikan data klinis kanker. Melalui penyetelan *hyperparameter*, performa model GNB dapat dioptimalkan, memberikan prediksi akurat, dan mencapai keseimbangan yang baik antara presisi dan *recall*. Terakhir, algoritma Multinomial Naive Bayes (MNB) menunjukkan konsistensi dalam kinerja dan kemampuan yang baik dalam beradaptasi dengan data yang beragam. Hasil optimisasi hyperparameter menghasilkan model dengan akurasi tinggi dan keseimbangan yang baik antara presisi dan *recall*. Hasil komparasi ini penting dalam konteks medis karena dapat digunakan sebagai dasar untuk memilih algoritma yang paling cocok untuk mendukung diagnosis atau prediksi kanker, dengan pemahaman yang mendalam tentang implikasi medis yang signifikan dari penggunaan algoritma ini dalam penanganan kanker di masa depan.

Daftar Pustaka

[1] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Eng. J.*, vol. 60, no. 3, hal. 3401–3409, 2021, doi: https://doi.org/10.1016/j.aej.2021.02.009.

[2] S. Ul, J. Ahamed, dan K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustain. Oper. Comput.*, vol. 3, no. February, hal. 238–248, 2022, doi: 10.1016/j.susoc.2022.03.001.

[3] B. Altınel dan M. C. Ganiz, "Semantic text classification: A survey of past and recent advances," *Inf. Process. Manag.*, vol. 54, no. 6, hal. 1129–1153, 2018, doi: https://doi.org/10.1016/j.ipm.2018.08.001.

[4] K. Zhang, "Study of text classification Natural Language Processing algorithms for four European areas' English dialects," in *2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, 2021, hal. 348–352, doi: 10.1109/CISAI54367.2021.00073.

[5] E. Hossain *et al.*, "Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review," *Comput. Biol. Med.*, vol. 155, hal. 106649, 2023, doi: https://doi.org/10.1016/j.compbiomed.2023.106649.

[6] L. Yao, C. Mao, dan Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," vol. 19, no. Suppl 3, 2019, doi: 10.1186/s12911-019-0781-4.

[7] G. Lan, M. Hu, Y. Li, dan Y. Zhang, "Contrastive knowledge integrated graph neural networks for Chinese medical text classification," *Eng. Appl. Artif. Intell.*, vol. 122, hal. 106057, 2023, doi: https://doi.org/10.1016/j.engappai.2023.106057.

[8] L. Cai, J. Li, H. Lv, W. Liu, H. Niu, dan Z. Wang, "Integrating domain knowledge for biomedical text analysis into deep learning: A survey," *J. Biomed. Inform.*, vol. 143, hal. 104418, 2023, doi: https://doi.org/10.1016/j.jbi.2023.104418.

[9] N. Chintalapudi, G. Battineni, M. Di Canio, G. G. Sagaro, dan F. Amenta, "Text mining with sentiment analysis on seafarers' medical documents," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, hal. 100005, 2021, doi: https://doi.org/10.1016/j.jjimei.2020.100005.

[10] WHO, "Cancer," 2022. https://www.who.int/news-room/fact-sheets/detail/cancer (diakses Feb 03, 2023).

[11] H. Kamel dan J. M. Al-tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," *2019 Int. Eng. Conf.*, hal. 165–170, 2019.

[12] K. Zeng, Y. Xu, G. Lin, L. Liang, dan T. Hao, "Automated classification of clinical trial eligibility criteria text based on ensemble learning and metric learning," vol. 21, no. Suppl 2, hal. 1–10, 2021.

[13] J. Jasmir, S. Nurmaini, R. F. Malik, dan D. Zaenal, "Text Classification of Cancer Clinical Trials Documents Using Deep Neural Network and Fine Grained Document Clustering," vol. 172, no. Siconian 2019, 2020.

[14] S. Gao *et al.*, "Limitations of Transformers on Clinical Text Classification," vol. 25, no. 9, hal. 3596–3607, 2021.

[15] S. Chen, G. I. Webb, L. Liu, dan X. Ma, "A novel selective Naïve Bayes algorithm," *Knowledge-Based Syst.*, vol. 192, hal. 105361, 2020, doi: https://doi.org/10.1016/j.knosys.2019.105361.

[16] D.-H. Vu, T.-S. Vu, dan T.-D. Luong, "An efficient and practical approach for privacy-preserving Naive Bayes classification," *J. Inf. Secur. Appl.*, vol. 68, hal. 103215, 2022, doi: https://doi.org/10.1016/j.jisa.2022.103215.

[17] F. Viegas *et al.*, "Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification," *Neurocomputing*, vol. 307, hal. 153–171, 2018, doi: https://doi.org/10.1016/j.neucom.2018.04.033.

[18] M. E. Ekpenyong *et al.*, "OPEN A hybrid computational framework for intelligent inter - continent SARS - CoV - 2 sub - strains characterization and prediction," *Sci. Rep.*, hal. 1–25, 2021, doi: 10.1038/s41598-021-93757-w.

[19] L. Ketsbaia, B. Issac, dan X. Chen, "Detection of Hate Tweets using Machine Learning and Deep Learning," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2020, hal. 751–758, doi: 10.1109/TrustCom50675.2020.00103.

[20] A. Roberson, "Applying Machine Learning for Automatic Product Categorization," *J. Off. Stat.*, vol. 37, no. 2, hal. 395–410, 3921, doi: doi:10.2478/jos-2021-0017.

[21] H. Ji, S. Huang, X. Lv, Y. Wu, dan Y. Feng, "Empirical Studies of a Kernel Density Estimation Based Naive Bayes Method for Software Defect Prediction ∗," no. 1, hal. 75–84, 2019.

[22] K. De Angeli, J. Doherty, A. Stroup, dan L. Coyle, "Deep Active Learning for Classifying Cancer Pathology Reports," 1803.

[23] A. Mascio *et al.*, "Comparative Analysis of Text Classification Approaches in Electronic Health Records," no. 1, hal. 86–94, 2020.

[24] A. Jalilifard *et al.*, "Semantic Sensitive TF-IDF to Determine Word Relevance in Documents," hal. 1–11, 2021.

[25] T. Parvin, M. M. Hoque, T. Parvin, dan M. M. Hoque, "ScienceDirect An Ensemble Ensemble Technique Technique to to Classify Classify Multi-Class Multi-Class Textual Textual Emotion Emotion," vol. 00, 2021.

[26] H. Kamel, D. Abdulah, dan J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," in *2019 International Engineering Conference (IEC)*, 2019, hal. 165–170, doi: 10.1109/IEC47844.2019.8950650.

[27] G. Singh, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," hal. 593–596, 2019.