Jurnal Ilmiah Sinus (JIS) Vol : 24, No. 2, Juli 2026
ISSN (Print) : 1693-1173 , ISSN (Online): 2548-4028

Klasifikasi Kategori Indeks dmft pada Data Pemeriksaan Gigi Menggunakan *Naïve Bayes* dengan Seleksi Fitur Kendall's Tau

Samsul Budiarto[1)] Fahmi Reza Ferdiansyah[2*)], Ramdan Nan Budiman[3)] Suharjanto utomo[4)]
[1,4)] Program Studi Teknik Informatika, Universitas Nurtanio
[2*)] Program Studi Informatika, Universitas Ekuitas
[3)] Program Studi Teknik Informatika, Institut Digital Ekonomi LPKIA
[1)]sambudiarto26@gmail.com [2*)]fahmi.reza@ekuitas.ac.id, [3)]ramdangates@gmail.com [4)]suharjanto.utomo@gmail.com

ABSTRACT

Dental caries severity in primary dentition can be summarized as ordinal dmft index categories. This study evaluates Gaussian Naïve Bayes for classifying these categories and explores Kendall's Tau-based feature selection. Kendall's Tau was selected as a nonparametric rank-based filter because the outcome is ordinal and the method does not require normally distributed variables, although its application to nominal predictors encoded as integers remains exploratory. Of 3,046 dental-examination records, 1,285 were retained after cleaning; 1,761 records (57.81%) were excluded mainly because responses were incomplete or could not be mapped consistently. A fixed 80:20 hold-out split was used. Accuracy increased descriptively from 66.15% with all columns in the initial implementation to 73.54% with six selected features. However, the majority-class baseline that always predicts the very-low category achieved 80.16% accuracy. The selected-feature model also produced a macro F1-score of 0.24, balanced accuracy of 23.07%, and zero recall for the very-high category. Its 95% Wilson confidence interval for accuracy was 67.83%-78.56%, which does not establish superiority over the baseline. Cross-validation was not available, feature selection preceded data splitting, and an identifier column was included in the initial full-column implementation. Therefore, feature selection improved performance only relative to the initial Naïve Bayes model, but the resulting classifier is not yet valid, balanced, or suitable for clinical screening.

Keywords: *dmft index category, dental caries, Kendall's Tau, Naïve Bayes, majority-class baseline, class imbalance*

I. PENDAHULUAN

Karies gigi merupakan masalah kesehatan mulut yang umum dan dapat menimbulkan nyeri, infeksi, gangguan makan, serta kehilangan gigi apabila tidak ditangani. Di Indonesia, gigi rusak atau berlubang dilaporkan sebagai masalah gigi dengan proporsi terbesar, yaitu 45,3% (Indonesia, 2018). Kondisi tersebut menunjukkan perlunya pengelolaan data pemeriksaan yang mampu mendukung identifikasi kelompok dengan pengalaman karies berbeda.

Pada gigi sulung, pengalaman karies diringkas menggunakan indeks decay, missing, filled teeth (dmft). Nilai dmft dapat dikelompokkan menjadi kategori sangat rendah, rendah, sedang, tinggi, dan sangat tinggi. Penelitian ini tidak melakukan diagnosis karies secara langsung, tetapi mengklasifikasikan kategori indeks dmft berdasarkan karakteristik pasien dan informasi keluhan yang tersedia.

Data pemeriksaan gigi umumnya memuat variabel kategorikal, nilai hilang, penulisan yang tidak konsisten, dan distribusi target yang tidak seimbang. Pada kondisi seperti ini, akurasi dapat memberikan kesan kinerja yang tinggi hanya karena model mengikuti kelas mayoritas. Karena itu, evaluasi perlu membandingkan model dengan *majority-class baseline* serta melaporkan macro F1, balanced accuracy, *recall* per kelas, dan confusion matrix.

Naïve Bayes dipilih karena sederhana, cepat, dan dapat digunakan sebagai baseline klasifikasi. Namun, Gaussian *Naïve Bayes* mengasumsikan fitur kontinu berdistribusi

Gaussian, sedangkan sebagian besar prediktor penelitian ini bersifat kategorikal. Oleh sebab itu, penerapannya diposisikan sebagai eksplorasi awal dan perlu dibandingkan dengan *Categorical Naïve Bayes* serta algoritma lain yang sesuai dengan data tabular kategorikal.

Penelitian karies terdahulu telah membandingkan berbagai algoritma pada data citra maupun data tabular. (Berdouses et al., 2015) membandingkan lima pengklasifikasi untuk karies oklusal; (Hung et al., 2019) menangani ketidakseimbangan data dalam prediksi karies akar; (Park et al., 2021) membandingkan model machine learning untuk early childhood caries; dan (Dey et al., 2024) membandingkan model machine learning dengan regresi pada outcome DMFT dan deft. Kajian sistematis (Reyes et al., 2022) juga menunjukkan heterogenitas data, target, dan prosedur validasi pada penelitian machine learning karies.

Seleksi fitur pada dental data mining sebelumnya menggunakan korelasi, GINI-mRMR, PCA, chi-square, dan feature importance (Alsubai, 2023; Berdouses et al., 2015; Kang et al., 2023). Metode-metode tersebut berguna, tetapi sebagian tidak secara khusus memanfaatkan sifat ordinal target. Kendall's Tau dipilih karena mengukur asosiasi monoton berbasis peringkat, tidak mensyaratkan normalitas, dan secara konseptual sesuai untuk outcome kategori dmft yang berurutan. Dibandingkan chi-square yang terutama menguji ketergantungan kategorikal tanpa arah ordinal, Kendall's Tau menyediakan koefisien asosiasi sekaligus p-value.

Keterbatasan literatur yang ditelaah adalah belum ditemukannya penerapan Kendall's Tau bersama *Naïve Bayes* untuk klasifikasi multikelas kategori indeks dmft pada data pemeriksaan gigi tabular. Selain itu, sejumlah penelitian lebih berfokus pada klasifikasi biner ada atau tidaknya karies dan belum selalu membandingkan hasil dengan baseline kelas mayoritas. Kebaruan penelitian ini adalah menguji kombinasi tersebut secara eksploratif sekaligus mengevaluasi dampak ketidakseimbangan kelas terhadap interpretasi hasil.

Tujuan penelitian adalah: (1) mengevaluasi Gaussian *Naïve Bayes* dalam mengklasifikasikan kategori indeks dmft; (2) membandingkan kinerja sebelum dan sesudah seleksi fitur Kendall's Tau; (3) membandingkan hasil dengan *majority-class baseline*; dan (4) mengidentifikasi keterbatasan yang berkaitan dengan kualitas data, ketidakseimbangan kelas, kesesuaian algoritma, dan prosedur validasi.

Kontribusi penelitian terletak pada penyajian alur pengolahan data pemeriksaan gigi dan evaluasi kritis terhadap peningkatan metrik model. Penelitian menunjukkan bahwa peningkatan relatif terhadap model awal tidak otomatis berarti model mengungguli baseline sederhana atau layak digunakan untuk skrining. Temuan ini memberikan dasar metodologis untuk pengembangan model dmft yang lebih valid dan seimbang.

II. TINJAUAN PUSTAKA

2.1 Karies Gigi dan Indeks dmft

Karies gigi merupakan penyakit multifaktorial, sedangkan indeks dmft merupakan ukuran pengalaman karies pada gigi sulung melalui penjumlahan komponen decay, missing, dan filled teeth. Kategori indeks dmft digunakan sebagai target ordinal dalam penelitian ini; karena itu, keluaran model harus ditafsirkan sebagai prediksi kategori skor, bukan diagnosis klinis langsung.

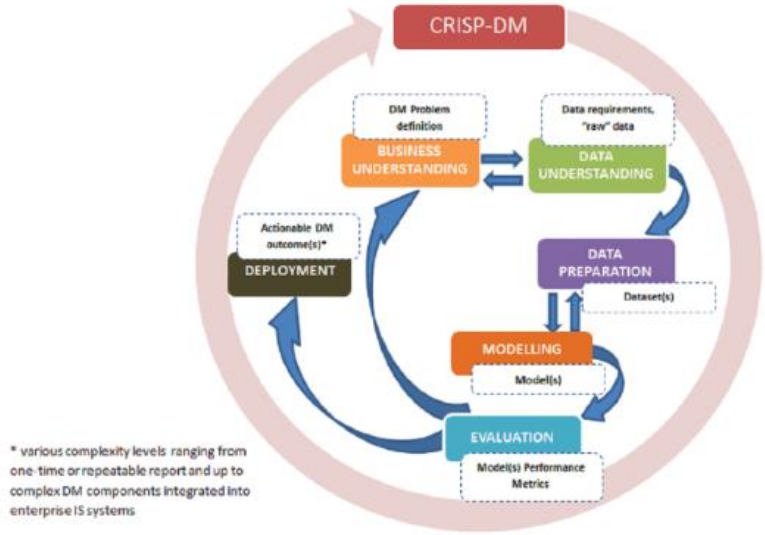
Tabel 1 Kategori indeks dmft yang digunakan dalam penelitian

Kode	Kategori	Rentang nilai dmft
1	Sangat rendah	< 1,2
0	Rendah	1,2-2,6
3	Sedang	2,7-4,4
4	Tinggi	4,5-6,5
2	Sangat tinggi	> 6,5

Kode pada Tabel 1 mengikuti urutan hasil LabelEncoder dalam implementasi penelitian. Kode tersebut hanya digunakan pada proses komputasi, sedangkan interpretasi hasil tetap mengacu pada nama kategori klinis.

2.2 Data Mining dan CRISP-DM

Penambangan data mengintegrasikan metode statistik, pembelajaran mesin, matematika, dan kecerdasan buatan untuk mengidentifikasi pola yang berharga dari kumpulan data. Kerangka Proses Standar Lintas Industri untuk Penambangan Data (CRISP-DM) menawarkan langkah-langkah pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan. Kerangka ini bisa disesuaikan dengan ciri-ciri domain penelitian, termasuk data layanan kesehatan.(Chapman et al., 2000; Plotnikova et al., 2021).



Gambar 1. Tahapan CRISP-DM yang digunakan sebagai kerangka penelitian

2.3 Algoritma Naïve Bayes

Naïve Bayes merupakan pengklasifikasi probabilistik yang menerapkan teorema Bayes dengan asumsi bahwa setiap fitur saling bebas secara kondisional terhadap kelas. Untuk vektor fitur X dan kelas Ck, probabilitas posterior dihitung menggunakan Persamaan (1). Model memilih kelas dengan probabilitas posterior terbesar.

$$P(C_k|X) = [P(C_k) \times \prod_i P(x_i|C_k)] / P(X) \qquad (1)$$

Gaussian *Naïve Bayes* mengasumsikan bahwa setiap fitur numerik pada masing-masing kelas mengikuti distribusi Gaussian. Asumsi ini paling sesuai untuk pengukuran kontinu, seperti usia dalam satuan tahun, tekanan darah, atau hasil pemeriksaan

laboratorium. Dalam penelitian ini, sebagian besar prediktor merupakan variabel kategorikal yang dikonversi menjadi kode bilangan. Kode tersebut berfungsi sebagai identitas kategori dan tidak otomatis memiliki jarak numerik yang bermakna, sehingga penerapan Gaussian *Naïve Bayes* diposisikan sebagai baseline komputasional yang bersifat eksploratif.

Untuk data diskret atau kategorikal, *Categorical Naïve Bayes* lebih sesuai karena mengestimasi probabilitas setiap kategori fitur secara langsung pada masing-masing kelas. (Murphy, 2022; Williams, 2024) menunjukkan bahwa representasi kategori dan pemilihan event model *Naïve Bayes* memengaruhi probabilitas posterior yang dihasilkan. Oleh karena itu, *Categorical Naïve Bayes* perlu digunakan sebagai model pembanding agar dapat diketahui apakah hasil penelitian dipengaruhi oleh seleksi fitur atau oleh ketidaksesuaian asumsi Gaussian terhadap kode kategori.

2.4 Seleksi Fitur Kendall's Tau

Kendall's Tau merupakan ukuran asosiasi berbasis peringkat yang membandingkan jumlah pasangan konkordan dan diskordan. Nilai tau berada pada rentang -1 sampai 1. Nilai yang mendekati 1 menunjukkan hubungan monoton searah yang kuat, nilai yang mendekati -1 menunjukkan hubungan berlawanan yang kuat, sedangkan nilai yang mendekati 0 menunjukkan hubungan monoton yang lemah.

$$\tau = (N^c - N^d) / [n(n - 1)/2] \qquad (2)$$

Pada penelitian ini, p-value < 0,05 digunakan sebagai kriteria penyaringan fitur. Pemilihan Kendall's Tau didasarkan pada sifat target dmft yang ordinal dan kebutuhan akan metode nonparametrik. Namun, signifikansi statistik harus dibaca bersama besar koefisien karena sampel yang relatif besar dapat menghasilkan p-value kecil meskipun hubungan sangat lemah. Untuk prediktor nominal yang dikodekan dengan LabelEncoder, tanda koefisien juga dipengaruhi urutan kode dan tidak boleh ditafsirkan sebagai arah hubungan klinis.

2.5 Ketidakseimbangan Kelas dan Evaluasi Model

Ketidakseimbangan kelas sering ditemukan pada penelitian prediksi karies. (Hung et al., 2019; Kang et al., 2023) melaporkan distribusi target yang timpang dan menggunakan teknik penyeimbangan pada data latih. Dalam kondisi tersebut, accuracy harus dibandingkan dengan *majority-class baseline*, yaitu prediksi sederhana yang selalu memilih kelas dengan jumlah anggota terbesar. Model yang tidak melampaui baseline tersebut belum memberikan manfaat diskriminatif yang memadai meskipun terlihat lebih baik daripada konfigurasi model lain.

Balanced accuracy merupakan rata-rata *recall* seluruh kelas dan sesuai digunakan ketika distribusi kelas tidak seimbang (Brodersen et al., 2010). Untuk lima kelas, prediksi yang hanya benar pada satu kelas memiliki *balanced accuracy* sekitar 0,20. Pada konteks skrining, *recall* kelas tinggi dan sangat tinggi juga penting karena false negative dapat menunda pemeriksaan lanjutan atau rujukan pasien yang membutuhkan perhatian lebih besar.

2.6 Penelitian Terdahulu

Tabel 2 Ringkasan penelitian terdahulu tentang klasifikasi karies, seleksi fitur, dan evaluasi data tidak seimbang

Peneliti	Objek penelitian	Hasil utama
(Berdouses et al., 2015)	Klasifikasi karies oklusal berbasis citra	Seleksi fitur berbasis korelasi dan perbandingan lima pengklasifikasi; Random Forest memberikan hasil terbaik.
(Kang et al., 2023)	Prediksi karies gigi berbasis data tabular	GINI dan mRMR digunakan untuk seleksi fitur sebelum GBDT; reduksi fitur meningkatkan kinerja beberapa model.
(Alsubai, 2023)	Prediksi karies berbasis fitur	PCA dan chi-square dipadukan dengan ensemble XGB, RF, dan ETC untuk meningkatkan prediksi.
(Hsieh & Cheng, 2024)	Klasifikasi enam kondisi gigi berbasis citra	Fusi fitur CNN diuji dengan SVM dan Naïve Bayes; SVM menghasilkan kinerja tertinggi.
(Kim et al., 2015)	Seleksi variabel kategorikal pada Naïve Bayes	Seleksi berbasis chi-square mempertahankan variabel relevan dan menyederhanakan model Naïve Bayes.
(Hung et al., 2019)	Prediksi karies akar berbasis data klinis	Data tidak seimbang ditangani dengan oversampling; evaluasi menunjukkan pentingnya sensitivitas kelas minoritas.
(Park et al., 2021)	Prediksi early childhood caries	Logistic regression, XGBoost, Random Forest, dan LightGBM dibandingkan pada data anak.
(Reyes et al., 2022)	Tinjauan sistematis machine learning karies	Menunjukkan heterogenitas target, data, algoritma, dan prosedur validasi pada penelitian diagnosis dan prognosis karies.
(Dey et al., 2024)	Prediksi DMFT dan deft pada data pediatrik	Model machine learning dibandingkan dengan regresi; kinerja dipengaruhi jenis outcome dan populasi.

Tabel 2 memperlihatkan bahwa penelitian karies telah menggunakan data citra, catatan klinis, dan data survei dengan beragam algoritma. Studi sebelumnya menegaskan tiga kebutuhan: pemilihan model yang sesuai dengan bentuk fitur, penanganan ketidakseimbangan kelas, dan validasi yang membandingkan beberapa algoritma. Metode seleksi fitur yang digunakan umumnya berupa korelasi, chi-square, PCA, GINI-mRMR, atau feature importance. Belum ditemukan penelitian pada tabel tersebut yang menggabungkan Kendall's Tau dan *Naïve Bayes* untuk klasifikasi multikelas kategori indeks dmft pada data tabular, sehingga penelitian ini mengisi celah tersebut secara eksploratif.

III. METODE PENELITIAN

3.1 Desain dan Sumber Data

Penelitian menggunakan desain observasional retrospektif dengan data sekunder hasil pemeriksaan gigi yang diperoleh dari Klinik Kedokteran Gigi Universitas Jenderal Achmad Yani (Unjani). Periode pengumpulan data belum tercantum dalam dokumen yang tersedia sehingga belum dapat dilaporkan secara rinci. Data awal terdiri atas 3.046 catatan yang mencakup karakteristik pasien, nilai dmft, keluhan, karakteristik keluhan, dan riwayat tindakan. Karakteristik dasar sampel, seperti distribusi usia dan jenis kelamin, juga belum tersedia secara lengkap dalam naskah ini sehingga keterwakilan populasi belum dapat dinilai secara menyeluruh. Informasi mengenai persetujuan atau pengecualian etik belum tercantum dalam dokumen sumber dan perlu diverifikasi kepada institusi pemilik data

sebelum publikasi. Variabel Name digunakan hanya untuk pemeriksaan duplikasi dan harus dikeluarkan sebelum seleksi fitur serta pemodelan.

Gambar 2. Contoh kondisi intraoral pada data pemeriksaan (dokumentasi penelitian, 2024)

3.2 Variabel Penelitian

Tabel 3 Variabel pada data pemeriksaan

Variabel	Jenis	Keterangan
Name	Identitas	Digunakan pada pemeriksaan duplikasi dan tidak termasuk model akhir
Job	Kategorikal	Pekerjaan responden
Age	Ordinal	Usia yang dikelompokkan menjadi bayi, anak-anak, remaja, dewasa, dan lansia
Sex	Kategorikal	Jenis kelamin
dmft	Target ordinal	Kategori sangat rendah sampai sangat tinggi
Complaints	Kategorikal	Ada atau tidaknya keluhan
Complaint_Time	Kategorikal	Waktu atau durasi keluhan
Complaint_Location	Kategorikal	Lokasi keluhan
Arise_Complaint	Kategorikal	Pola munculnya keluhan
How_Complaint	Kategorikal	Karakter atau cara keluhan dirasakan
Worsen_Complaint	Kategorikal	Faktor yang memperberat keluhan
Relieve_Complaints	Kategorikal	Faktor yang meredakan keluhan
Treatment	Kategorikal	Riwayat atau bentuk perawatan

3.3 Tahapan Pengolahan Data

Tahapan penelitian mengikuti CRISP-DM. Business understanding menetapkan kebutuhan untuk mengklasifikasikan kategori dmft dan mengevaluasi pengaruh seleksi fitur. Data understanding dilakukan melalui pemeriksaan jumlah data, tipe variabel, nilai kosong, duplikasi, dan konsistensi jawaban. Data preparation terdiri atas cleaning, integration, transformation, dan reduction.

Kriteria inklusi operasional adalah catatan yang memiliki nilai dmft yang dapat dikategorikan dan jawaban prediktor yang dapat dipetakan secara konsisten ke kategori penelitian. Catatan dikeluarkan apabila target dmft tidak tersedia, jawaban kunci kosong atau tidak dapat dipetakan, atau teridentifikasi sebagai duplikasi berdasarkan pemeriksaan identitas. Dari 3.046 catatan, 1.761 (57,81%) dikeluarkan dan 1.285 (42,19%) dipertahankan. Log pembersihan awal tidak mencatat jumlah eksklusi secara terpisah untuk missing target, missing predictor, duplikasi, dan inkonsistensi, serta tidak

menyediakan pola missingness per variabel. Penggunaan complete-case secara luas dapat menimbulkan bias seleksi apabila catatan yang lengkap berbeda secara sistematis dari catatan yang dikeluarkan; karena itu, generalisasi hasil dibatasi pada catatan yang lolos proses cleaning.

Pada tahap transformation, nilai dmft dikelompokkan berdasarkan rentang pada Tabel 1 dan usia dikelompokkan menjadi bayi (<5 tahun), anak-anak (5-9 tahun), remaja (10-18 tahun), dewasa (19-59 tahun), dan lansia (>=60 tahun). Variabel kategorikal dikonversi menjadi kode integer menggunakan LabelEncoder. Kode tersebut bukan ukuran kontinu dan tidak menunjukkan jarak setara antarkategori. Gaussian *Naïve Bayes* tetap digunakan sebagai baseline eksploratif sesuai implementasi awal, tetapi ketidaksesuaian asumsi ini menjadi alasan perlunya pembandingan dengan *Categorical Naïve Bayes*.

Pada implementasi awal, Kendall's Tau dihitung pada seluruh data sebelum pembagian data latih dan data uji. Prosedur ini berpotensi menyebabkan data leakage. Selain itu, kolom Name masih ikut dihitung pada audit korelasi dan masuk dalam konfigurasi 'seluruh variabel', meskipun pada definisi variabel dinyatakan hanya untuk pemeriksaan duplikasi. Hal tersebut merupakan inkonsistensi implementasi: Name bukan fitur klinis dan harus dikeluarkan sebelum seleksi fitur maupun pemodelan. Karena data mentah dan keluaran prediksi ulang tidak tersedia dalam naskah, angka model awal tetap dilaporkan secara transparan sebagai hasil implementasi awal yang memerlukan reanalisis. Prosedur yang benar adalah menghapus Name, membagi data secara terstratifikasi, lalu melakukan encoding dan seleksi fitur hanya pada data latih atau di dalam setiap fold.

3.4 Pemodelan dan Evaluasi

Data pada analisis awal dibagi menjadi 80% data latih dan 20% data uji menggunakan random_state 42, yaitu 1.028 data latih dan 257 data uji. Satu kali hold-out dipertahankan hanya untuk melaporkan ulang hasil yang telah tersedia. Stratified k-fold cross-validation belum dilakukan, sehingga tidak tersedia rata-rata, simpangan baku, atau interval variasi antar-fold. Ketidaktersediaan hasil cross-validation dilaporkan sebagai keterbatasan, bukan digantikan dengan estimasi yang tidak berasal dari data.

Model utama adalah Gaussian Naïve Bayes. Evaluasi juga memasukkan *majority-class baseline* yang selalu memprediksi kategori sangat rendah karena kategori tersebut merupakan kelas terbesar pada data uji. Baseline ini digunakan untuk menilai apakah kompleksitas model memberikan manfaat dibandingkan aturan prediksi paling sederhana. Analisis awal belum menjalankan *Categorical Naïve Bayes*, decision tree, random forest, ordinal logistic regression, atau boosting; karena itu, pengaruh seleksi fitur belum dapat dipisahkan dari karakteristik algoritma.

Metrik yang digunakan adalah accuracy, precision, *recall*, F1-score, macro F1, weighted F1, balanced accuracy, confusion matrix, dan interval kepercayaan 95% untuk accuracy menggunakan metode Wilson. *Balanced accuracy* dihitung sebagai rata-rata *recall* lima kelas. Perbedaan dua model pada data uji yang sama idealnya diuji dengan McNemar berdasarkan prediksi berpasangan, sedangkan ketidakpastian metrik pada validasi silang dapat dihitung dengan paired bootstrap. Keluaran prediksi per observasi tidak tersedia sehingga uji McNemar belum dapat dilakukan.

IV. HASIL DAN PEMBAHASAN

4.1 Hasil Data Understanding dan Cleaning

Setelah standardisasi dan eksklusi catatan tidak lengkap atau tidak konsisten, tersisa 1.285 data atau 42,19% dari data awal; 1.761 catatan atau 57,81% tidak digunakan. Karena jumlah eksklusi per penyebab dan pola missingness tidak tersedia, tidak dapat dipastikan apakah kehilangan data bersifat acak. Proporsi eksklusi yang besar menimbulkan potensi

bias seleksi dan membatasi keterwakilan sampel. Sumber data telah diidentifikasi sebagai Klinik Kedokteran Gigi Universitas Jenderal Achmad Yani (Unjani), tetapi periode pengumpulan, karakteristik peserta, dan informasi etik masih perlu dilengkapi agar validitas eksternal dan perlindungan subjek dapat dinilai.

Pada data uji, kategori sangat rendah berjumlah 206 dari 257 catatan (80,16%), sedangkan kategori sangat tinggi hanya lima catatan. Dengan distribusi ini, *majority-class baseline* yang selalu memprediksi sangat rendah memperoleh accuracy 80,16%. Nilai tersebut menjadi batas pembanding minimum yang harus dipertimbangkan sebelum menyatakan model memberikan perbaikan yang bermakna.

4.2 Hasil Seleksi Fitur Kendall's Tau

Tabel 4 Hasil korelasi Kendall's Tau terhadap kategori dmft

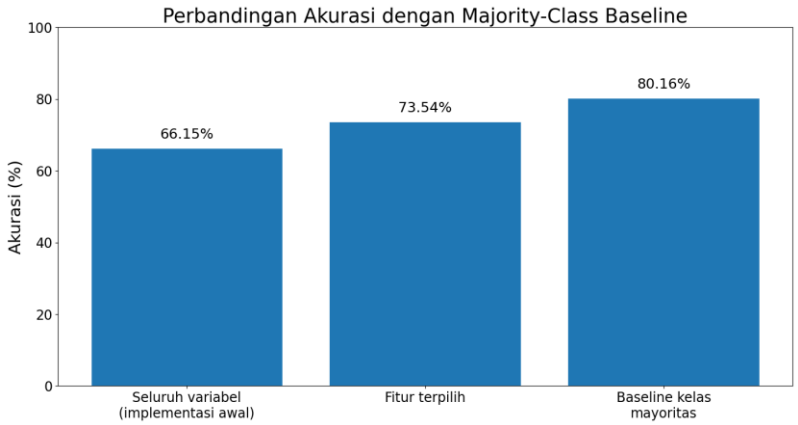
Variabel	Kendall's Tau	p-value	Keputusan
Name	0,026051	0,237161	Audit awal; identifier, bukan prediktor
Job	0,010122	0,707511	Tidak signifikan
Age	-0,104152	0,000057	Signifikan
Sex	0,006186	0,818630	Tidak signifikan
Complaints	-0,107741	0,000021	Signifikan
Complaint_Time	-0,011374	0,655847	Tidak signifikan
Complaint_Location	0,008678	0,738124	Tidak signifikan
Arise_Complaint	-0,090075	0,000468	Signifikan
How_Complaint	-0,082571	0,002199	Signifikan
Worsen_Complaint	-0,083768	0,001135	Signifikan
Relieve_Complaints	-0,070044	0,006378	Signifikan
Treatment	-0,023381	0,369381	Tidak signifikan

Enam fitur memiliki p-value < 0,05, yaitu Age, Complaints, Arise_Complaint, How_Complaint, Worsen_Complaint, dan Relieve_Complaints. Koefisiennya hanya berkisar dari -0,070044 sampai -0,107741, yang menunjukkan asosiasi monoton sangat lemah. P-value kecil dapat dipengaruhi ukuran sampel, sedangkan tanda pada fitur nominal dipengaruhi urutan kode LabelEncoder. Baris Name dipertahankan pada Tabel 4 hanya untuk menunjukkan audit implementasi awal; kolom tersebut merupakan identifier dan seharusnya dikeluarkan sebelum korelasi maupun pemodelan. Karena itu, hasil seleksi tidak boleh ditafsirkan sebagai hubungan sebab-akibat atau kekuatan klinis yang besar.

4.3 Hasil Pemodelan Naïve Bayes

Tabel 5 Perbandingan kinerja dengan *majority-class baseline*

Model (jumlah fitur)	Accuracy	Macro F1	Weighted F1	Balanced accuracy
Majority-class baseline (0)	80,16%	0,18	0,71	0,20
Seluruh kolom, implementasi awal (12; termasuk Name)	66,15%	0,24	0,68	Tidak tersedia
Enam fitur terpilih	73,54%	0,24	0,71	0,23



Gambar 3. Perbandingan akurasi model dengan *majority-class baseline*

Model enam fitur menghasilkan 189 prediksi benar dari 257 data uji atau accuracy 73,54%, sedangkan implementasi awal seluruh kolom menghasilkan 170 prediksi benar atau 66,15%. Peningkatan 7,39 poin persentase hanya menunjukkan perbaikan relatif terhadap model awal. *Majority-class baseline* menghasilkan 206 prediksi benar atau 80,16%, sehingga model enam fitur masih 6,62 poin persentase lebih rendah daripada baseline sederhana. Interval kepercayaan Wilson 95% adalah 60,16%-71,66% untuk implementasi awal, 67,83%-78,56% untuk model enam fitur, dan 74,85%-84,57% untuk baseline mayoritas.

Balanced accuracy model enam fitur yang dihitung dari *recall* per kelas adalah 23,07%, hanya sedikit di atas nilai 20% yang dihasilkan *majority-class baseline* pada masalah lima kelas. Macro F1 model sebesar 0,24 memang lebih tinggi daripada macro F1 baseline sekitar 0,18, tetapi weighted F1 keduanya sama-sama sekitar 0,71. Hasil tersebut menunjukkan adanya sedikit peningkatan pengenalan lintas kelas dibandingkan prediksi mayoritas, tetapi belum cukup untuk menghasilkan klasifikasi multikelas yang memadai. Hasil cross-validation tidak tersedia, sehingga kestabilan metrik pada pembagian data lain belum dapat dinilai.

Uji signifikansi berpasangan belum dapat dilakukan karena prediksi per observasi dari kedua model tidak tersedia. Tumpang tindih interval kepercayaan accuracy juga tidak dapat digunakan sebagai pengganti uji McNemar. Oleh sebab itu, klaim yang dapat dipertahankan hanya bahwa seleksi fitur berkaitan dengan peningkatan deskriptif dibandingkan implementasi awal, bukan bahwa model telah unggul secara statistik atau mengungguli baseline yang layak.

Tabel 6 Kinerja per kelas pada model dengan fitur terpilih

Kategori	Precision	*Recall*	F1-score	Support
Rendah	0,10	0,07	0,08	15
Sangat rendah	0,83	0,90	0,86	206
Sangat tinggi	0,00	0,00	0,00	5
Sedang	0,10	0,11	0,10	19
Tinggi	1,00	0,08	0,15	12

Recall kategori sangat rendah mencapai 0,90, sedangkan *recall* kategori sangat tinggi adalah 0,00. Kategori tinggi memiliki precision 1,00 tetapi *recall* hanya 0,08, yang berarti model hanya membuat sangat sedikit prediksi tinggi dan melewatkan sebagian besar kasus

aktual. Nilai tersebut menegaskan bahwa model belum menjalankan fungsi klasifikasi multikelas secara seimbang.

Tabel 7 *Confusion matrix* model dengan fitur terpilih

Aktual \ Prediksi	Rendah	Sangat rendah	Sangat tinggi	Sedang	Tinggi
Rendah	1	12	0	2	0
Sangat rendah	4	185	4	13	0
Sangat tinggi	1	2	0	2	0
Sedang	3	14	0	2	0
Tinggi	1	9	0	1	1

Confusion matrix menunjukkan 185 dari 206 data sangat rendah diklasifikasikan dengan benar, sementara sebagian besar data rendah, sedang, tinggi, dan sangat tinggi dialihkan ke kelas sangat rendah. Pola tersebut konsisten dengan bias menuju kelas mayoritas dan menjelaskan mengapa accuracy dapat terlihat cukup tinggi meskipun *balanced accuracy* dan macro F1 rendah.

4.4 Pembahasan

Perbandingan dengan *majority-class baseline* mengubah interpretasi hasil. Seleksi fitur meningkatkan accuracy dari 66,15% menjadi 73,54% relatif terhadap implementasi awal, tetapi model terbaik masih kalah dari baseline 80,16%. Dengan *balanced accuracy* 23,07% yang hanya sedikit di atas 20%, model belum menunjukkan kemampuan diskriminatif yang memadai pada seluruh kategori. Oleh karena itu, istilah 'memperbaiki kinerja' hanya berlaku dalam perbandingan internal dengan model awal dan tidak boleh digunakan untuk menyatakan bahwa model sudah efektif secara praktis.

Temuan ini sejalan dengan literatur dental machine learning yang menekankan pengaruh ketidakseimbangan kelas dan pilihan algoritma. Hung et al. (2019) dan Kang et al. (2023) menerapkan penanganan kelas tidak seimbang pada data karies, sedangkan Park et al. (2021) dan Dey et al. (2024) membandingkan beberapa model agar hasil tidak bergantung pada satu algoritma. Reyes et al. (2022) juga menunjukkan bahwa bukti kinerja model karies sangat heterogen dan membutuhkan validasi yang lebih konsisten. Dengan demikian, manfaat Kendall's Tau belum dapat dipisahkan dari kelemahan implementasi Gaussian *Naïve Bayes* dan komposisi data uji.

Implikasi klinis harus dinilai secara konservatif. Kegagalan mendeteksi kategori tinggi dan sangat tinggi menghasilkan false negative yang dapat menempatkan pasien berisiko ke kategori sangat rendah. Apabila model digunakan untuk menentukan prioritas pemeriksaan atau rujukan, kesalahan tersebut berpotensi menunda evaluasi dokter gigi. Karena itu, model tidak boleh digunakan sebagai alat skrining sampai sensitivitas kelas klinis penting meningkat, hasil tervalidasi pada populasi eksternal, dan ambang keputusan ditentukan bersama tenaga kesehatan gigi.

Keterbatasan algoritmik meliputi penggunaan LabelEncoder pada fitur nominal dan Gaussian *Naïve Bayes* pada kode kategori. *Categorical Naïve Bayes* lebih sesuai karena mengestimasi probabilitas kategori secara langsung. Selain itu, masuknya Name dalam implementasi awal seluruh kolom merupakan kesalahan karena identifier dapat menangkap pola administratif yang tidak dapat digeneralisasi. Konfigurasi tersebut harus dihitung ulang setelah Name dihapus agar baseline model internal dapat dibandingkan secara sah.

Seleksi fitur sebelum pembagian data menimbulkan risiko data leakage, sedangkan satu kali hold-out tidak mengukur variasi kinerja. Analisis lanjutan harus menempatkan penghapusan identifier, encoding, imputasi atau penanganan missing value, seleksi fitur,

dan resampling di dalam pipeline stratified k-fold cross-validation. Hasil perlu dilaporkan sebagai rata-rata dan simpangan baku atau interval kepercayaan, kemudian dikonfirmasi pada test set eksternal yang tidak digunakan untuk pemilihan fitur atau model.

Kehilangan 57,81% catatan dan tidak tersedianya rincian pola missingness membatasi validitas internal dan eksternal. Penelitian berikutnya perlu membuat diagram alur inklusi-eksklusi, mencatat jumlah data yang dihapus untuk setiap alasan, membandingkan karakteristik catatan yang disertakan dan dikeluarkan, serta melaporkan periode pengumpulan data, karakteristik sampel, persetujuan etik, dan proses de-identifikasi. Langkah ini diperlukan untuk menilai bias seleksi, keterwakilan populasi, dan perlindungan identitas pasien.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Seleksi fitur Kendall's Tau menghasilkan enam fitur dan meningkatkan kinerja Gaussian *Naïve Bayes* secara relatif terhadap implementasi awal, dari accuracy 66,15% menjadi 73,54%. Namun, accuracy tersebut masih lebih rendah daripada *majority-class baseline* 80,16%. *Balanced accuracy* hanya 23,07%, macro F1 tetap 0,24, dan *recall* kategori sangat tinggi adalah 0,00. Dengan demikian, seleksi fitur belum menghasilkan model klasifikasi kategori indeks dmft yang valid, seimbang, atau layak digunakan untuk skrining.

Kontribusi penelitian adalah menunjukkan bahwa evaluasi model pada data dmft yang tidak seimbang harus melampaui perbandingan antarkonfigurasi model dan mencakup baseline sederhana, metrik per kelas, serta pemeriksaan risiko false negative. Hasil juga mengidentifikasi masalah implementasi berupa masuknya identifier Name, ketidaksesuaian Gaussian *Naïve Bayes* untuk fitur kategorikal, data leakage, kehilangan data yang besar, dan tidak adanya cross-validation. Karena itu, seluruh hasil diposisikan sebagai temuan eksploratif yang memerlukan reanalisis sebelum klaim prediktif atau klinis dibuat.

5.2 Saran

Penelitian selanjutnya harus: menghapus Name sebelum analisis; melengkapi informasi periode pengumpulan data, karakteristik sampel, etik, dan de-identifikasi; membuat log inklusi-eksklusi serta pola missingness; menerapkan pipeline stratified k-fold cross-validation yang bebas leakage; dan membandingkan *Categorical Naïve Bayes*, Gaussian Naïve Bayes, ordinal logistic regression, decision tree, random forest, serta boosting. Penanganan ketidakseimbangan seperti class weighting, oversampling, undersampling, atau SMOTENC hanya boleh diterapkan pada data latih. Evaluasi wajib mencakup *majority-class baseline*, balanced accuracy, macro F1, *recall* kelas tinggi dan sangat tinggi, interval kepercayaan, serta uji berpasangan berdasarkan prediksi per observasi. Validasi eksternal diperlukan sebelum model dipertimbangkan untuk skrining.

DAFTAR PUSTAKA

Alsubai, S. (2023). Enhancing prediction of tooth caries using significant features and multi-model classifier. *PeerJ Computer Science*, *9*, e1631. https://doi.org/10.7717/peerj-cs.1631

Berdouses, E. D., Koutsouri, G. D., Tripoliti, E. E., Matsopoulos, G. K., Oulis, C. J., & Fotiadis, D. I. (2015). A computer-aided automated methodology for the detection and classification of occlusal caries from photographic color images. *Computers in Biology and Medicine*, *62*, 119–135.

https://doi.org/10.1016/j.compbiomed.2015.04.016

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The *balanced accuracy* and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition* (pp. 3121–3124). https://doi.org/10.1109/ICPR.2010.764

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.

Dey, P., Ogwo, C., & Tellez, M. (2024). Comparison of traditional regression modeling vs. AI modeling for the prediction of dental caries: A secondary data analysis. *Frontiers in Oral Health*, *5*, 1322733. https://doi.org/10.3389/froh.2024.1322733

Hsieh, S.-T., & Cheng, Y.-A. (2024). Multimodal feature fusion in deep learning for comprehensive dental condition classification. *Journal of X-Ray Science and Technology*, *32*(2), 303–321. https://doi.org/10.3233/XST-230271

Hung, M., Voss, M. W., Rosales, M. N., Li, W., Su, W., Xu, J., Bounsanga, J., Ruiz-Negrón, B., Lauren, E., & Licari, F. W. (2019). Application of machine learning for diagnostic prediction of root caries. *Gerodontology*, *36*(4), 395–404. https://doi.org/10.1111/ger.12432

Indonesia, K. K. R. (2018). *Laporan nasional Riset Kesehatan Dasar 2018*. Badan Penelitian dan Pengembangan Kesehatan.

Kang, I.-A., Njimbouom, S. N., & Kim, J.-D. (2023). Optimal feature selection-based dental caries prediction model using machine learning for decision support system. *Bioengineering*, *10*(2), 245. https://doi.org/10.3390/bioengineering10020245

Kim, M.-S., Choi, H., & Park, C. (2015). Categorical variable selection in *Naïve Bayes* classification. *The Korean Journal of Applied Statistics*, *28*(3), 407–415. https://doi.org/10.5351/KJAS.2015.28.3.407

Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. books.google.com. https://books.google.com/books?hl=en&lr=&id=wrZNEAAAQBAJ&oi=fnd&pg=PR27&dq=machine+learning&ots=LaoeOA7lOn&sig=jy3HSNWal93AACqhFjM_sZHmG44

Park, Y.-H., Kim, S.-H., & Choi, Y.-Y. (2021). Prediction models of early childhood caries based on machine learning algorithms. *International Journal of Environmental Research and Public Health*, *18*(16), 8613. https://doi.org/10.3390/ijerph18168613

Plotnikova, V., Dumas, M., & Milani, F. (2021). *Adapting the CRISP-DM data mining process: A case study in the financial services domain*. https://figshare.com/s/33c42eda3b19784e8b21

Reyes, L. T., Knorst, J. K., Ortiz, F. R., & Ardenghi, T. M. (2022). Machine learning in the diagnosis and prognostic prediction of dental caries: A systematic review. *Caries Research*, *56*(3), 161–170. https://doi.org/10.1159/000524167

Williams, C. K. I. (2024). *Naïve Bayes* classifiers and one-hot encoding of categorical variables. In *arXiv*. https://doi.org/10.48550/arXiv.2404.18190