e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 23 Agustus 2023

Penggunaan Metode NearMiss, SMOTE, dan Naïve Bayes untuk Klasifikasi Gangguan Tidur Berdasarkan Kualitas Tidur dan Gaya Hidup

Mohamad Reyhand Fatturrahman[1], Aliyah Kurniasih[2]
Program Studi Ilmu Komputer
Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ
JL TB Simatupang, Cilandak Timur, Kec. Ps. Minggu, Kota Jakarta Selatan 12560
mohamad.reyhand.f@students.esqbs.ac.id[1], aliyah.kurniasih@esqbs.ac.id[2]

Abstrak. Gangguan tidur merupakan permasalahan kesehatan yang melanda banyak individu global dan berpotensi memberikan dampak negatif terhadap kualitas hidup mereka. Dalam upaya mengidentifikasi jenis gangguan tidur yang dialami serta meramalkan jenis gangguan tidur yang belum teridentifikasi pada subjek pengamatan, penelitian ini mengadopsi metode klasifikasi *Naïve Bayes*. Pemilihan metode *Naïve Bayes* dipandang sesuai karena kemampuannya dalam melakukan pengklasifikasian dengan efisiensi tinggi. Keberhasilan penerapan *Naïve Bayes* terlihat dari berbagai aplikasi dunia nyata, seperti dalam klasifikasi dokumen dan deteksi spam. Penelitian ini juga menggunakan pendekatan *data mining* dan metode pembelajaran mesin untuk mengurai pola serta relasi antara kualitas tidur, gaya hidup, dan jenis gangguan tidur pada individu. Dengan algoritma *Naïve Bayes* sebagai landasannya, penelitian ini bertujuan menghasilkan hasil klasifikasi yang mendalam, memberikan pemahaman terkait gangguan tidur beserta faktor-faktor yang mempengaruhinya. Studi ini menerapkan metode klasifikasi dengan model *Naïve Bayes Bernoulli*, dengan akurasi sebelum dan sesudah penerapan *oversampling* serta undersampling berturut-turut mencapai 76% dan 79%. Hasil serupa juga dihasilkan dari metode *Naïve Bayes Complement* sebelum penerapan teknik yang sama, yaitu 79%. Meskipun oversampling berhasil meningkatkan akurasi model *Naïve Bayes Bernoulli* sebesar 7%, namun terdapat penurunan sebesar 8% dalam akurasi setelah melakukan undersampling jika dibandingkan dengan metode *Naïve Bayes Complement*.

Kata Kunci: Gangguan tidur, *Naïve Bayes*, *Oversampling*, *Undersampling*

1 Pendahuluan

Gangguan tidur adalah masalah kesehatan masyarakat yang utama, mempengaruhi jutaan orang di seluruh dunia. Mereka dapat menyebabkan berbagai masalah kesehatan, termasuk sakit kronis, obesitas, dan penyakit jantung. Gangguan tidur juga dapat berdampak signifikan pada kualitas hidup seseorang, memengaruhi kemampuannya untuk bekerja, belajar, dan bersosialisasi [1]. Ada banyak jenis gangguan tidur, dan dapat disebabkan oleh berbagai faktor, termasuk gaya hidup, stres, kecemasan, depresi, dan kondisi medis. Mendiagnosis gangguan tidur bisa jadi sulit, karena banyak gejala yang juga umum terjadi pada kondisi lain [2].

Dengan permasalahan tersebut menjadi penting untuk menemukan model yang dapat menggambarkan dan membedakan kelas data atau konsep dengan tujuan untuk memprediksi kelas yang belum diketahui dari objek pengamatan. Analisis diskriminan dan *regresi logistik* adalah metode klasifikasi yang umum digunakan dalam statistika. Namun, era data semakin populer, menunjukkan bahwa volume data yang luar biasa besar telah meningkat pesat, yang menghasilkan set data besar. Untuk mendapatkan informasi penting dari set data besar dan mengorganisasikannya menjadi pengetahuan yang terorganisir, diperlukan alat analisis yang kuat dan berguna.

Di sisi lain, seiring dengan pesatnya perkembangan teknologi kecerdasan buatan, muncul metode pembelajaran mesin. Metode ini adalah mesin yang memiliki kemampuan untuk belajar sendiri tanpa bantuan manusia dan dibangun berdasarkan disiplin ilmu lain seperti *data mining*, matematika, dan statistika. Dalam metode pembelajaran mesin, metode klasifikasi seperti *Classification and Regression Trees* (CART), *Random Forest*, *Naïve Bayes*, *Support Vector Machines* (SVM), dan lainnya sering digunakan [3].

Dalam penelitian sebelumnya, diketahui bahwa metode *Naïve Bayes* dapat digunakan untuk klasifikasi dan menghasilkan data dengan tingkat akurasi yang lebih tinggi daripada metode *k-nearest neighbors* (KNN). Hasil penelitian menunjukkan bahwa akurasi metode *Naïve Bayes* mencapai 78%, sementara KNN hanya mencapai 65%. Hal ini menunjukkan bahwa algoritma *Naïve Bayes* efektif untuk digunakan dalam tugas klasifikasi [4].

Algoritma *Naïve Bayes* dipilih untuk penelitian ini karena memiliki proses pengklasifikasian yang sangat cepat dibandingkan dengan pendekatan yang lebih canggih. Meskipun asumsi yang tampaknya terlalu sederhana, pengklasifikasi *Naïve Bayes* telah berhasil dalam banyak situasi dunia nyata, terkenal dengan klasifikasi dokumen dan memfilterkan spam. *Naïve bayes* hanya membutuhkan jumlah data pelatihan yang kecil untuk memperkirakan parameter yang diperlukan. Tujuan dari penelitian ini adalah untuk mendapatkan hasil klasifikasi yang dapat digunakan untuk mengidentifikasi pola atau hubungan antara kualitas tidur, gaya hidup, dan jenis gangguan tidur yang dialami oleh individu [5][6].

SMOTE adalah sebuah teknik yang digunakan untuk mengatasi ketidakseimbangan jumlah sampel antara kelas mayoritas dan kelas minoritas dengan cara menghasilkan data sintetis pada kelas minoritas. Teknik ini bekerja dengan memilih data sampel pada kelas minoritas dan membuat sampel baru yang mirip tetapi tidak identik dengan data yang ada. Dengan demikian, jumlah sampel pada kelas minoritas menjadi seimbang dengan jumlah sampel pada kelas mayoritas. Namun perlu diperhatikan pula *overfitting* pada data, karena label yang diduplikasi dan digunakan pada proses pelatihan [7].

Metode *NearMiss* memilih sampel dari kelas mayoritas berdasarkan jarak mereka ke sampel dari kelas minoritas. Jarak antara mereka dihitung dalam ruang fitur menggunakan jarak geometris. Metode *NearMiss* adalah salah satu dari tiga metode yang dapat digunakan. Dengan menggunakan metode ini, jumlah sampel yang diambil dari kelas mayoritas dapat diimbangi dengan jumlah sampel yang diambil dari kelas minoritas. Hal ini membantu menyeimbangkan distribusi data dan mengurangi bias yang mungkin terjadi karena ketidakseimbangan kelas [8].

Dalam penelitian ini, metode *Naïve Bayes* digunakan untuk klasifikasi gangguan tidur. Tujuannya adalah untuk menemukan pola dan hubungan antara kualitas tidur, gaya hidup, dan jenis gangguan tidur yang dialami oleh individu. Metode *Naïve Bayes* telah terbukti lebih efektif daripada *k-nearest neighbors* (KNN). Selain itu, teknik SMOTE menangani ketidakseimbangan dalam jumlah sampel antara kelas mayoritas dan kelas minoritas, sementara metode *NearMiss* memilih sampel dari kelas mayoritas berdasarkan jarak mereka dari sampel kelas minoritas untuk membantu menyeimbangkan distribusi data. Diharapkan bahwa penelitian ini akan memberikan hasil klasifikasi yang lebih akurat dan pemahaman yang lebih baik tentang gangguan tidur dan faktor-faktor yang mempengaruhinya dengan menggabungkan metode ini.

2 Tinjauan Pustaka

2.1 Gangguan Tidur

Gangguan tidur adalah kondisi di mana seseorang mengalami gangguan dalam jumlah, kualitas, atau waktu tidur. Gangguan tidur dapat dipengaruhi oleh berbagai faktor, baik yang bersifat medis maupun non-medis. Faktor non-medis meliputi jenis kelamin, pubertas, kebiasaan tidur, status sosioekonomi, keadaan keluarga, gaya hidup, dan lingkungan yang berhubungan dengan tidur yang terganggu [5]. Hampir semua orang pernah mengalami masalah tidur. 20%–40% orang dewasa mengalami masalah tidur setiap tahun, dan 17% di antaranya mengalami masalah serius. Jumlah gangguan tidur cenderung meningkat setiap tahun, sejalan dengan peningkatan usia dan berbagai penyebabnya. Menurut Kaplan dan Sadock, gangguan tidur menyertai sekitar empat puluh hingga lima puluh persen dari populasi usia lanjut. gangguan tidur jangka panjang (10–15%) disebabkan oleh gangguan psikiatri, ketergantungan obat, dan alkohol [5].

2.2 Penambangan Data

Penambangan data melibatkan serangkaian kegiatan yang meliputi pengumpulan, pembersihan, pemrosesan, analisis, dan penarikan wawasan yang berguna dari data. Penambangan data mencakup berbagai domain masalah, aplikasi, formulasi, dan representasi data yang ditemui dalam konteks aplikasi nyata. Oleh karena itu, istilah "penambangan data" digunakan secara umum untuk merujuk pada proses-proses ini yang bertujuan untuk memperoleh informasi berharga dari data.

Setiap aspek kehidupan modern dikomputerisasi oleh kemajuan teknologi baru, dan banjir data adalah hasil langsung dari kemajuan ini. Oleh karena itu, penting untuk memeriksa apakah seseorang dapat mengekstraksi wawasan yang sederhana dan dapat ditindaklanjuti dari data yang tersedia untuk tujuan aplikasi tertentu. Ini adalah tempat tugas penambangan data datang. Misalnya, data yang dikumpulkan secara manual dapat diambil dari berbagai sumber dalam format yang berbeda, tetapi entah bagaimana perlu diproses oleh program komputer otomatis untuk memperoleh pemahaman, karena data mentah mungkin sewenang-wenang, tidak terstruktur, atau bahkan dalam format yang tidak langsung cocok untuk pemrosesan otomatis. Untuk mengatasi masalah ini, analis penambangan data menggunakan jalur pemrosesan, yang mengumpulkan, membersihkan, dan mengubah data mentah menjadi format standar.

2.3 Klasifikasi

Klasifikasi merujuk pada proses pengelompokan atau pengkategorian data ke dalam kelas-kelas yang telah ditentukan sebelumnya berdasarkan nilai fitur khusus yang disebut sebagai label kelas. Dalam masalah penambangan data yang bersifat terarah (*supervised*), tujuan dari klasifikasi adalah untuk mempelajari hubungan antara fitur-fitur lain dalam data dengan label kelas yang telah ada.

Masalah klasifikasi dapat dipetakan ke versi spesifik dari masalah pendeteksian *outlier*, dengan memasukkan pengawasan pada versi terakhir. Sementara masalah deteksi *outlier* diasumsikan *unsupervised* secara *default*, banyak variasi masalah baik sebagian atau seluruhnya diawasi. Dalam deteksi *outlier* yang diawasi, beberapa contoh *outlier* tersedia. Dengan demikian, catatan data tersebut ditandai milik kelas langka, sedangkan catatan data yang tersisa milik kelas normal. Dengan demikian, masalah deteksi *outlier* yang diawasi dipetakan ke masalah klasifikasi biner, dengan peringatan bahwa label kelas sangat tidak seimbang [9].

Di sisi lain, masalah klasifikasi sering digunakan secara langsung sebagai alat yang berdiri sendiri di banyak aplikasi. Beberapa contoh aplikasi di mana masalah klasifikasi digunakan adalah sebagai berikut:

- Target pemasaran: Fitur tentang pelanggan terkait dengan perilaku pembelian mereka penggunaan model pelatihan.
- Deteksi intrusi: Urutan aktivitas pelanggan dalam sistem komputer mungkin digunakan untuk memprediksi kemungkinan gangguan.
- Deteksi anomali terawasi: Kelas langka dapat dibedakan dari normal kelas ketika contoh outlier sebelumnya tersedia..

2.4 Algoritma *Naïve Bayes*

Naïve Bayes adalah sebuah algoritma pembelajaran mesin yang digunakan dalam tugas klasifikasi, terutama dalam klasifikasi teks. Algoritma ini termasuk dalam keluarga algoritma pembelajaran generatif, yang berfokus pada pemodelan distribusi *input* dari suatu kelas atau kategori tertentu [10]. Teorema Bayes berguna untuk memperkirakan P(D|E) ketika sulit untuk memperkirakan P(D|E) secara langsung dari data pelatihan, tetapi probabilitas kondisional dan sebelumnya lainnya seperti P(E|D), P(D), dan P(E) dapat diperkirakan dengan lebih mudah. Secara khusus, teorema Bayes menyatakan sebagai berikut:

$$P(D|E) = \frac{P(E|D)P(D)}{P(E)}.$$

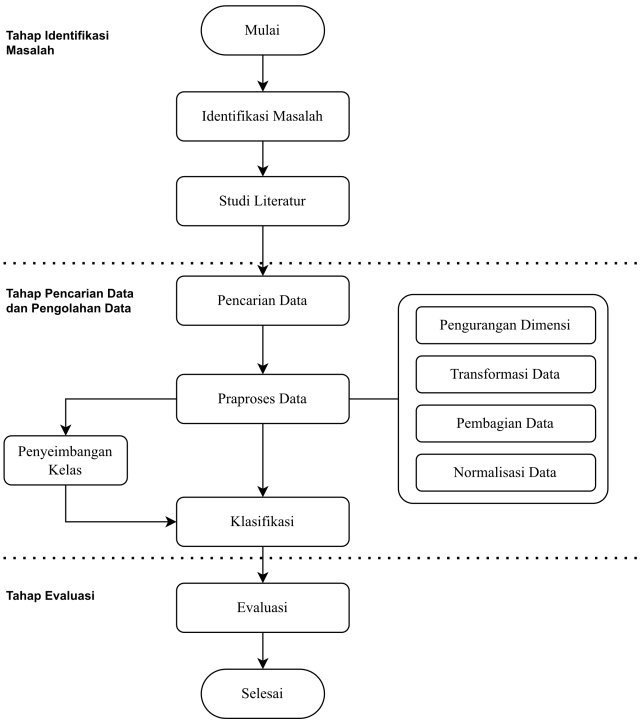
Keterangan:
X : Data kelas belum diketahui
H : Hipotesis data kelas spesifik
P(E) : Probabilitas data kelas belum diketahui
P(D) : Probabilitas hipotesis data kelas spesifik
P(E|D) : Probabilitas data
P(D|E) : Probabilitas posterior

2.5 SMOTE

Metode SMOTE adalah salah satu metode yang banyak digunakan untuk mengatasi ketidakseimbangan kelas dalam *dataset*. Cara kerja dari algoritma ini untuk setiap kasus minoritas, jumlah tetangga terdekat yang termasuk dalam kelas yang sama dihitung. Sebagian kecil dari mereka kemudian dipilih secara acak, tergantung pada tingkat *oversampling* yang diperlukan. Contoh data sintetik dibuat pada segmen garis yang menghubungkan contoh minoritas ke tetangga terdekatnya untuk setiap pasangan contoh tetangga yang dijadikan sampel. Di sepanjang garis, posisi contoh dipilih secara acak.

Data pelatihan termasuk contoh minoritas baru-baru ini, dan data tambahan digunakan untuk melatih pengklasifikasi. Algoritma SMOTE biasanya lebih akurat daripada metode *oversampling vanilla*. Metode ini menghasilkan wilayah keputusan yang lebih luas dari data yang di sampel ulang daripada situasi di mana hanya anggota kelas tertentu yang ditemukan dalam data pelatihan asli yang di sampel berlebihan [9].

3 Metode Penelitian



Gambar 1. Alur Penelitian

Identifikasi Masalah. Pada tahap ini, dilakukan pengamatan terhadap semua variabel yang dapat mempengaruhi gangguan tidur. Masalah pada penelitian ini adalah bagaimana data mining menggunakan algoritma *Naïve Bayes* dengan teknik *oversampling* SMOTE dan *undersampling NearMiss* untuk klasifikasi gangguan tidur.

Studi literatur. Pada tahap ini yaitu memberikan dukungan dalam menyelesaikan masalah penelitian dengan mengumpulkan buku dan jurnal yang relevan dengan judul penelitian. Sumber-sumber studi literatur yang digunakan berasal dari berbagai sumber yang berkaitan dengan topik klasifikasi atau pengolahan data menggunakan algoritma *Naïve Bayes*, SMOTE, dan *NearMiss*.

Pengumpulan Data. Pada tahap ini, dilakukan pencarian data untuk penelitian. Dalam penelitian ini, peneliti menggunakan dataset yang bersifat publik dari situs Kaggle. Dataset ini terdiri dari 374 data dengan 13 variabel, terdiri dari 8 fitur *input* dan 1 fitur untuk target klasifikasi model. Variabel fitur yang digunakan adalah *Sleep Duration*, *Quality of Sleep*, *Physical Activity Level*, *Stress Level*, *BMI Category*, *Blood Pressure*, *Heart Rate*, dan *Daily Steps*. Fitur target yang digunakan dalam data ini adalah *Sleep Disorder*, dengan label "*None*", "*Sleep Apnea*", dan "*Insomnia*".

Praproses Data. Tahap praproses data yaitu untuk membersihkan, mengubah, atau menyesuaikan data mentah sebelum dianalisis atau dimasukkan ke dalam model pembelajaran mesin. Tujuannya adalah untuk mempersiapkan data agar sesuai dengan persyaratan model dan meningkatkan kualitas hasil analisis. Tahap-tahap yang ada pada praproses penelitian ini sebagai berikut.

1. Pengurangan dimensi, untuk mengurangi jumlah fitur atau variabel dalam dataset. Hal ini dilakukan untuk mengatasi masalah *high dimensionality*, di mana terdapat banyak fitur yang mungkin tidak relevan atau saling berkorelasi tinggi, sehingga dapat mempengaruhi kinerja model.
2. Transformasi data, proses ini melibatkan perubahan skala atau bentuk data asli menjadi format yang lebih sesuai atau bermanfaat untuk analisis lebih lanjut.
3. Pembagian data, proses memisahkan dataset menjadi sub set yang berbeda untuk digunakan dalam pelatihan dan pengujian model. Biasanya, *dataset* dibagi menjadi data latih (*training data*) dan data uji (*testing data*).
4. Normalisasi data, proses mengubah skala data menjadi rentang atau distribusi yang lebih ter standarisasi. Ini dilakukan untuk menghilangkan perbedaan skala antara fitur-fitur dalam dataset.

Penyeimbangan Kelas. Proses untuk mengatasi ketidakseimbangan distribusi kelas dalam dataset. Ketidakseimbangan kelas terjadi ketika jumlah sampel dalam kelas positif dan negatif sangat tidak seimbang, yang dapat menyebabkan bias dalam klasifikasi. Teknik penyeimbangan kelas seperti *oversampling*, *undersampling*, atau kombinasi keduanya digunakan untuk memperbaiki ketidakseimbangan ini.

Klasifikasi. Tugas dalam pembelajaran mesin dimana model memprediksi kelas atau label dari sampel berdasarkan fitur-fitur yang ada. Model klasifikasi dilatih dengan menggunakan data latih yang telah diberi label dan kemudian digunakan untuk mengklasifikasikan sampel yang belum diketahui.

Evaluasi. Hal ini dilakukan untuk mengevaluasi sejauh mana model dapat memprediksi dengan akurat kelas atau nilai yang benar. Metrik evaluasi yang umum digunakan termasuk akurasi, *presisi*, *recall*, dan *f1-score*.

4 Hasil dan Pembahasan

4.1 Data

Tahap pertama dalam penelitian ini adalah pencarian data klasifikasi *Sleep Disorder*. Data yang digunakan dalam penelitian ini adalah dataset yang bersifat publik dan bersumber dari situs Kaggle, yang didapat dari (https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset). Dataset yang diberikan berjumlah 374 data yang terdiri dari 13 variabel yang terdiri dari 8 fitur input model dan 1 fitur target klasifikasi. Variabel fitur yang digunakan adalah *Sleep Duration*, *Quality of Sleep*, *Physical Activity Level*, *Stress Level*, *BMI Category*, *Blood Pressure*, *Heart Rate*, dan *Daily Steps*. Variabel target yang digunakan dalam dataset ini adalah *Sleep Disorder*, dengan kategori label "*None*", "*Sleep Apnea*", dan "*Insomnia*".

Tabel 1. Tabel Informasi Keseluruhan Data

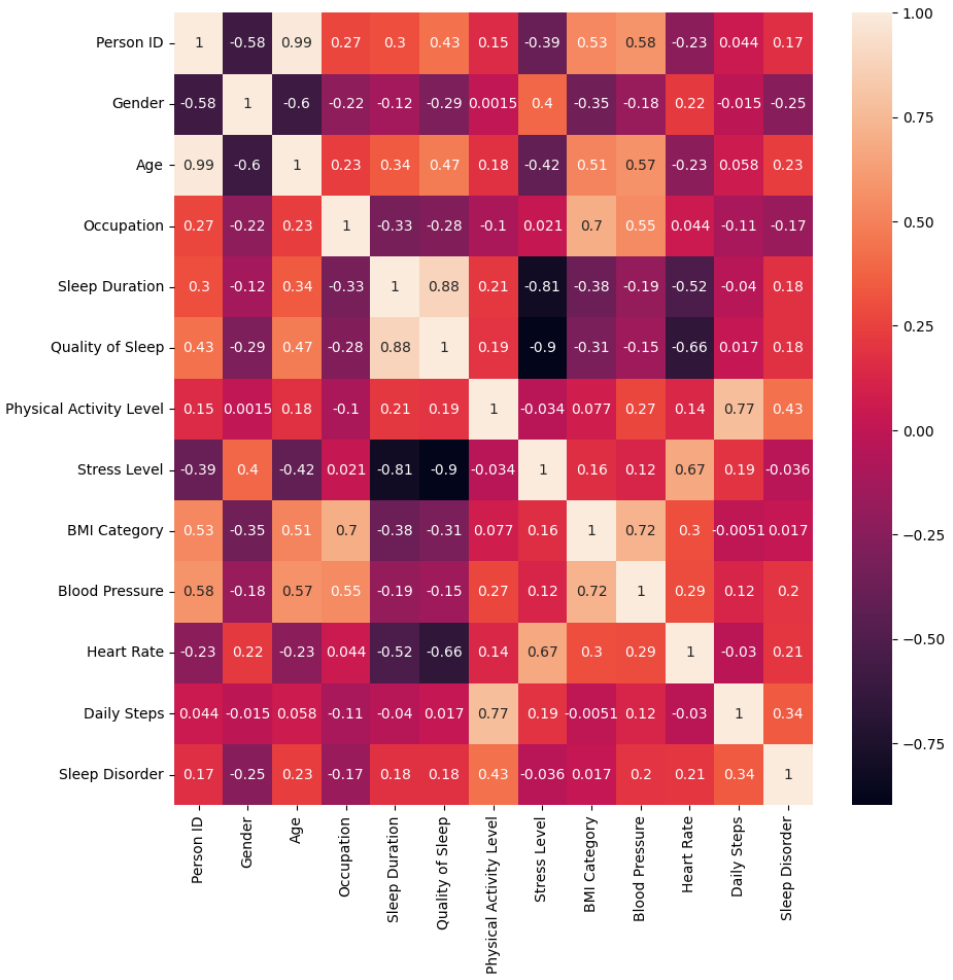
No	Variabel	Keterangan	Jenis	Nilai
1	*Person ID*	Identifikasi unik untuk setiap individu.	Numerikal	1 – 374
2	*Gender*	Menunjukkan jenis kelamin seseorang.	Kategorikal	*1. Male* *2. Female*
3	*Age*	Merupakan usia seseorang dalam tahun.	Numerikal	18 - 50
4	*Occupation*	Menunjukkan pekerjaan atau profesi seseorang.	Kategorikal	*1. Software Engineer* *2. Doctor* *3. Sales Representative* *4. Teacher* *5. Nurse* *6. Engineer* *7. Accountant* *8. Scientist* *9. Lawyer* *10. Salesperson* *11. Manager*
5	*Sleep Duration*	Merupakan jumlah waktu tidur yang dihabiskan oleh seseorang dalam satu malam.	Numerikal	6,1 – 8,2
6	*Quality of Sleep*	Menunjukkan seberapa baik kualitas tidur seseorang.	Numerikal	6 – 9
7	*Physical*	Tingkat aktivitas fisik seseorang	Numerikal	42 – 85

	Activity Level			
8	*Stress Level*	Menunjukkan seberapa tinggi tingkat stres seseorang	Numerikal	3 – 6
9	*BMI Category*	Mengacu pada kategori berat badan seseorang berdasarkan indeks massa tubuh (BMI)	Kategorikal	*1. Overweight* *2. Normal* *3. Obese* *4. Normal Weight*
10	*Blood Pressure*	Mengacu pada tekanan darah seseorang	Kategorikal	115/75, 115/78, 117/76, 118/75, 118/76, 119/77, 120/80, 121/79, 122/80, 125/80, 125/82, 126/83, 128/84, 128/85, 129/84, 130/85, 130/86, 131/86, 132/87, 135/88, 135/90, 139/91, 139/91, 140/90, 140/95, 142/92
11	*Heart Rate*	Merupakan jumlah detak jantung seseorang per menit	Numerikal	65 – 86
12	*Daily Steps*	Jumlah langkah yang diambil seseorang dalam satu hari	Numerikal	3000 – 10000
13	*Sleep Disorder*	Menunjukkan apakah seseorang mengalami gangguan tidur	Kategorikal	*1. None* *2. Sleep Apnea* *3. Insomnia*

4.2 Praproses Data

4.2.1 Pengurangan Dimensi

Proses ini merujuk pada proses mengurangi jumlah fitur atau variabel dalam *dataset*. Tujuannya adalah untuk menghilangkan fitur yang tidak relevan atau redundan, sehingga mengurangi kompleksitas dan dimensi data. Hal ini dilakukan untuk mengatasi masalah *high dimensionality*, di mana terdapat banyak fitur yang mungkin tidak relevan atau saling berkorelasi tinggi, sehingga dapat mempengaruhi kinerja model. Pada penelitian ini dilakukan penghapusan beberapa fitur, yaitu *Person ID*, *Gender*, *Age*, *Occupation*.



Gambar 2. Heatmap Correlation

4.2.2 Transformasi Data

Proses ini mengubah skala atau bentuk distribusi data asli menjadi bentuk yang lebih sesuai atau dapat diinterpretasikan dengan lebih baik. Tujuannya adalah untuk mengatasi masalah seperti skewness, heteroskedastisitas, atau perbedaan skala antar variabel. Pada tahap ini dilakukan transformasi data agar data yang digunakan dapat dilakukan pemrosesan. Pada tahap ini variabel fitur yang berjenis kategorikal akan diubah menjadi numerikal, yaitu fitur BMI Category, Blood Pressure, dan Sleep Disorder.

Tabel 2. Tabel perbandingan transformasi data

No	BMI Category	Blood Pressure	Sleep Disorder
	Data sebelum diubah		
1	Overweight	126/83	None
2	Normal	125/80	None
3	Normal	125/80	None
4	Obese	140/90	Sleep Apnea
5	Obese	140/90	Sleep Apnea
	Data setelah diubah		
1	3	11	1
2	0	9	1
3	0	9	1
4	2	22	2
5	2	22	2

4.2.3 Pemisahan Data

Proses ini merujuk pada proses memisahkan dataset menjadi sub set yang berbeda, biasanya data latih (*training data*) dan data uji (*test data*). Pembagian data dilakukan untuk melatih model pada *sub set data* latih dan menguji performa model pada sub set data uji. Pembagian ini penting untuk menghindari *overfitting* dan memastikan generalisasi yang baik pada data yang belum pernah dilihat sebelumnya. Pada penelitian ini, data latih sebesar 80% dan data uji sebesar 20%.

Tabel 3. Hasil Pembagian Data

Keterangan	Komponen X	Komponen Y
Data latih	(299, 8)	(299, 1)
Data uji	(75, 8)	(75,1

4.2.4 Normalisasi Data

Proses mengubah skala nilai variabel agar memiliki rentang atau skala yang serupa. Normalisasi umumnya dilakukan untuk menghindari bias yang disebabkan oleh perbedaan skala variabel. Salah satu metode normalisasi yang digunakan adalah *min-max scaling*, dimana nilai-nilai variabel dikonversi ke rentang 0-1 berdasarkan nilai minimum dan maksimum yang ada dalam dataset.

4.3 Penyeimbangan Kelas

Tahap ini dilakukan dengan dua data yang berbeda, data yang menggunakan teknik *oversampling* SMOTE dan data yang menggunakan teknik *undersampling NearMiss*. Metode SMOTE bekerja dengan cara menghasilkan sampel sintetis baru dari kelas minoritas dalam dataset dengan melakukan interpolasi linier antara sampel-sampel yang ada di kelas minoritas. Tujuan utama dari SMOTE adalah meningkatkan jumlah sampel dalam kelas minoritas agar seimbang dengan jumlah sampel dalam kelas mayoritas. Sedangkan metode *NearMiss* berfokus pada mengurangi jumlah sampel dari kelas mayoritas yang berada dekat dengan sampel-sampel dari kelas minoritas. Dengan demikian, kedua metode ini bertujuan untuk mengurangi ketimpangan antara kelas mayoritas dan kelas minoritas dalam *dataset*.

Tabel 4. Tabel perbandingan jumlah data latih sebelum dan sesudah *oversampling* dan *undersampling*

Jenis Data	0	1	2	Total
Data belum seimbang	57	178	64	374
Data setelah SMOTE	178	178	178	534
Data setelah NearMiss	57	57	57	171

Klasifikasi

Pada penelitian ini digunakan *library scikit-learn* untuk menerapkan metode klasifikasi *Naïve Bayes* dengan dua variasi: *Naïve Bayes Bernoulli* dan *Complement*. Untuk menerapkan *oversampling*, kami menggunakan metode SMOTE dan *undersampling*, kami menggunakan metode *NearMiss*. SMOTE digunakan untuk meningkatkan jumlah sampel dalam kelas minoritas dengan menciptakan sampel sintetis berdasarkan interpolasi linier dari sampel-sampel yang ada. Sedangkan *NearMiss* digunakan untuk mengurangi jumlah sampel dalam kelas mayoritas agar seimbang dengan kelas minoritas dengan cara memilih sampel-sampel mayoritas yang paling dekat dengan sampel-sampel minoritas.

Penelitian ini melakukan klasifikasi pada dua set data yang berbeda, yaitu data yang tidak seimbang dan data yang telah mengalami proses oversampling atau *undersampling*. Tujuan dari ini adalah untuk membandingkan kinerja model *Naïve Bayes Bernoulli* dan *Complement* pada kedua kondisi tersebut dan melihat apakah penggunaan metode oversampling atau *undersampling* dapat meningkatkan akurasi klasifikasi pada data yang tidak seimbang.

4.4 Evaluasi

Setelah melakukan proses klasifikasi, terdapat evaluasi untuk mengukur performa model yang telah dilakukan. Terdapat perbandingan hasil antara data yang tidak seimbang dan data yang telah seimbang menggunakan teknik SMOTE atau *NearMiss* dan metode *Naïve Bayes Bernoulli* dan *Complement* untuk kedua kondisi tersebut. Setelah melakukan evaluasi menggunakan metode *classification report*, kami membandingkan akurasi, presisi, f1-score, dan sensitivitas (recall) dari pengujian terhadap kedua data tersebut.

Pada data yang belum seimbang, dapat dilihat akurasi, presisi, f1-score, dan sensitivitas model *Naïve Bayes Bernoulli* dan *Complement*. Selanjutnya, kami melakukan evaluasi yang sama pada data yang telah seimbang menggunakan teknik SMOTE dan *NearMiss*. Dengan membandingkan hasil evaluasi tersebut, dapat dinilai apakah penggunaan teknik oversampling atau *undersampling* dapat meningkatkan performa model *Naïve Bayes Bernoulli* dan *Complement* pada data yang tidak seimbang.

Tabel 5. Hasil Evaluasi Model *Naïve Bayes Bernoulli* Sebelum *Oversampling* Dan *Undersampling* Menggunakan *Classification Report*

	Precision	*Recall*	*F1-score*	*Support*
0	0.55	0.80	0.65	20
1	0.90	0.85	0.88	41
2	0.86	0.43	0.57	14
Accuracy			0.76	75
Macro avg	0.77	0.69	0.70	75
Weighted avg	0.80	0.76	0.76	75

Tabel 6. Hasil Evaluasi Model *Naïve Bayes Complement* Sebelum *Oversampling* Dan *Undersampling* Menggunakan *Classification Report*

	Precision	*Recall*	*F1-score*	*Support*
0	0.61	0.85	0.71	20
1	0.91	0.73	0.81	41
2	0.86	0.86	0.86	14
Accuracy			0.79	75
Macro avg	0.79	0.81	0.79	75
Weighted avg	0.82	0.79	0.79	75

Tabel 7. Hasil Evaluasi Model *Naïve Bayes Bernoulli* Sesudah *Oversampling* Menggunakan *Classification Report*

	Precision	*Recall*	*F1-score*	*Support*
0	0.55	0.80	0.65	20
1	0.90	0.85	0.88	41
2	0.86	0.43	0.57	14
Accuracy			0.76	75
Macro avg	0.77	0.69	0.70	75
Weighted avg	0.80	0.76	0.76	75

Tabel 8. Hasil Evaluasi Model *Naïve Bayes Complement* Sesudah *Oversampling* Menggunakan *Classification Report*

	Precision	*Recall*	*F1-score*	*Support*
0	0.57	0.85	0.68	20
1	0.76	0.71	0.73	41
2	0.86	0.43	0.57	14
Accuracy			0.69	75
Macro avg	0.73	0.66	0.66	75
Weighted avg	0.73	0.69	0.69	75

Tabel 9. Hasil Evaluasi Model *Naïve Bayes Bernoulli* Sesudah *Undersampling* Menggunakan *Classification Report*

	Precision	*Recall*	*F1-score*	*Support*
0	0.52	0.85	0.64	20
1	0.91	0.78	0.84	41
2	0.86	0.43	0.57	14
Accuracy			0.73	75
Macro avg	0.76	0.69	0.69	75
Weighted avg	0.80	0.73	0.74	75

Tabel 10. Hasil Evaluasi Model *Naïve Bayes Complement* Sesudah *Undersampling* Menggunakan *Classification Report*

	Precision	*Recall*	*F1-score*	*Support*
0	0.81	0.85	0.83	20
1	0.81	0.93	0.86	41

2	0.86	0.43	0.57	14
Accuracy			0.81	75
Macro avg	0.83	0.74	0.75	75
Weighted avg	0.82	0.81	0.80	75

Tabel 11. Hasil Evaluasi Model *Naïve Bayes Bernoulli* dan *Complement* Sesudah *Oversampling* dan *Undersampling* Menggunakan *Classification Report*

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
BernaulliNB tanpa SMOTE	76%	77%	69%	70%
ComplementNB tanpa SMOTE	79%	79%	81%	79%
BernaulliNB dengan SMOTE	76%	77%	69%	70%
ComplementNB dengan SMOTE	69%	73%	66%	66%
BernaulliNB dengan NearMiss	73%	76%	69%	69%
ComplementNB dengan NearMiss	81%	83%	74%	75%

5 Kesimpulan dan Saran

Berdasarkan hasil evaluasi yang diberikan, terlihat bahwa nilai akurasi setelah penerapan *undersampling* menggunakan *NearMiss* pada *Complement Naïve Bayes* meningkat sebesar 2%, jika dibandingkan dengan penerapan *oversampling* menggunakan SMOTE pada *Bernoulli Naïve Bayes* dengan nilai akurasi turun sebesar 3%. Nilai akurasi model klasifikasi sebelumnya pada *Bernoulli* dan *Complement Naïve Bayes* sebesar 76% dan 79%, tetapi meningkat menjadi 73% dan 81% setelah *undersampling* menggunakan *NearMiss*. Dengan adanya peningkatan nilai akurasi dan metrik evaluasi lainnya, dapat disimpulkan bahwa undersampling menggunakan metode *NearMiss* memberikan hasil yang lebih baik dalam klasifikasi.

Rekomendasi untuk penelitian berikutnya adalah menggunakan metode lain selain *oversampling*, seperti *undersampling* atau kombinasi dari keduanya, guna mengatasi masalah ketidakseimbangan kelas yang ada. Selain itu, penting juga untuk melakukan pemilihan fitur yang lebih baik dengan menggunakan algoritma yang sesuai dengan tujuan penelitian ini.

Referensi

[1] Japardi. Dr. Iskandar, “Gangguan Tidur”.

[2] J. Keperawatan and P. Kesehatan Denpasar, “Faktor Yang Menyebabkan Gangguan Tidur (Insomnia) Pada Lansia I Nengah Sumirta AA Istri Laraswati.”

[3] P. R. Sihombing and I. F. Yuliati, “Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 2, pp. 417–426, May 2021, doi: 10.30812/matrik.v20i2.1174.

[4] R. Diki Nugraha, “Pentingnya Algoritma Naïve Bayes Sebagai Pengklasifikasi Data,” vol. 2, no. 1, 2023.

[5] A. Haryono, A. Rindiarti, A. Arianti, A. Pawitri, and Achmad Ushuluddin, “Prevalensi Gangguan Tidur pada Remaja Usia 12-15 Tahun di Sekolah Lanjutan Tingkat Pertama,” *Sari Pediatri*, vol. 11, 2009.

[6] N. Umar and M. Adnan Nur, “Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 585–590, Aug. 2022, doi: 10.29207/resti.v6i4.4179.

[7] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, A. Nikmatul Kasanah, U. Pujianto, T. Elektro, F. Teknik, and U. Negeri Malang, “Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN,” *JURNAL RESTI*, vol. 3, no. 3, pp. 196–201, 2017.

[8] A. Nurhopipah and C. Magnolia, “Perbandingan Metode Resampling Pada Imbalanced Dataset Untuk Klasifikasi Komentar Program MBKM,” *JUPIKOM*, vol. 1, no. 2, 2022.

[9] C. C. Aggarwal, *Data Mining*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.

[10] “What are Naïve Bayes classifiers?,” *https://www.ibm.com/topics/naive-bayes*.