ANALISIS METODE *SMOTE, TOMEK LINKS,* DAN *HYBRID SMOTE+TOMEK LINKS* TERHADAP KLASIFIKASI *NAIVE BAYES* UNTUK MENGATASI DATA TIDAK SEIMBANG PADA DIAGNOSA PENYAKIT TUBERKULOSIS

(Tesis)

Oleh

NAFLAH FAULINA
NPM 2227031011

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

ABSTRACT

ANALYSIS OF SMOTE, TOMEK LINKS, AND HYBRID SMOTE+TOMEK LINKS METHODS ON NAIVE BAYES CLASSIFICATION FOR ADDRESSING IMBALANCED DATA IN TUBERCULOSIS DIAGNOSIS

By

NAFLAH FAULINA

Naive Bayes classification is a method based on Bayes' theorem, utilizing probabilistic and statistical techniques. In practice, many datasets exhibit imbalanced class distributions. A common issue with imbalanced data is that the classifier tends to predict the class with the larger data composition. As a result, this leads to high prediction accuracy for the majority class in the training data but poor prediction accuracy for the minority class. One resampling technique to address this issue involves using oversampling, undersampling, or a combination of both. The aim of this study is to balance the training data by applying SMOTE (Synthetic Minority Oversampling Technique), Tomek Links, and a hybrid approach combining both methods. The performance of the Naive Bayes classifier on the original imbalanced data is compared with its performance on the balanced data in diagnosing tuberculosis at Mayjend HM Ryacudu Kotabumi Hospital. The results show that the hybrid approach, combining SMOTE for oversampling and Tomek Links for undersampling, demonstrates the best performance in balancing the data and improving the accuracy of the Naive Bayes model. Specifically, the hybrid method achieved an average accuracy of 93%, an average sensitivity of 88%, an average specificity of 96%, an average False Positive Fraction (FPF) of 4%, and an average False Negative Fraction (FNF) of 12%.

Key words: Naive Bayes classification, SMOTE, Tomek Links, SMOTE+Tomek Links, Tuberculosis

ABSTRAK

ANALISIS METODE *SMOTE, TOMEK LINKS,* DAN *HYBRID SMOTE+TOMEK LINKS* TERHADAP KLASIFIKASI NAIVE BAYES UNTUK MENGATASI DATA TIDAK SEIMBANG PADA DIAGNOSA PENYAKIT TUBERKULOSIS

OLEH

NAFLAH FAULINA

Klasifikasi naive bayes adalah metode yang didasarkan pada teorema Bayes dengan metode probabilitas dan statistik. Dalam penerapannya banyak data yang ditemui memiliki distribusi tidak seimbang di setiap kelasnya. Permasalahan yang sering terjadi pada data tidak seimbang, klasifikasi cenderung memprediksi kelas yang memiliki komposisi data lebih besar. Akibatnya dihasilkan hasil akurasi prediksi yang baik pada kelas data *training* yang mayoritas, sedangkan akan dihasilkan akurasi prediksi yang buruk pada data *training* yang minoritas. Salah satu teknik *resampling* dapat dilakukan untuk menanganinya yaitu dengan metode *oversampling*, *undersampling*, dan gabungan keduanya. Tujuan dari penelitian ini adalah mengaplikasikan metode *Synthetic Minority Oversampling Technique* (SMOTE), Tomek Links, dan *hybrid* SMOTE+Tomek Links untuk mengatasi data tidak seimbang dan membandingkan kinerja (performa) klasifikasi naive bayes pada data original dan data yang telah seimbang dalam diagnosa penyakit Tuberkulosis di RSD. Mayjend HM Ryacudu Kotabumi. Hasil yang diperoleh *hybrid* antara teknik *oversampling* dengan SMOTE dan teknik *undersampling* dengan Tomek Links menunjukkan kinerja terbaik dalam menyeimbangkan data dan meningkatkan akurasi model naive bayes, dengan akurasi rata-rata mencapai 93%, *sensitivity* rata-rata 88%, *specificity* rata-rata 96%, rata-rata *False Positive Fraction(*FPF*)* 4%, dan rata-rata *False Negative Fraction (*FNF*)* 12%.

Kata Kunci: Klasifikasi Naive Bayes, SMOTE, Tomek Links, SMOTE+Tomek Links, Tuberkulosis

ANALISIS METODE *SMOTE, TOMEK LINKS*, DAN *HYBRID SMOTE+TOMEK LINKS* TERHADAP KLASIFIKASI NAIVE BAYES UNTUK MENGATASI DATA TIDAK SEIMBANG PADA DIAGNOSA PENYAKIT TUBERKULOSIS

Oleh

NAFLAH FAULINA

Tesis

Sebagai Salah Satu Syarat untuk Mencapai Gelar

MAGISTER MATEMATIKA

Pada

Program Studi Magister Matematika

Fakultas Matematika Dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

Judul	:	**ANALISIS METODE *SMOTE, TOMEK LINKS*, DAN *HYBRID SMOTE+TOMEK LINKS* TERHADAP KLASIFIKASI *NAIVE BAYES* UNTUK MENGATASI DATA TIDAK SEIMBANG PADA DIAGNOSA PENYAKIT TUBERKULOSIS**
Nama Mahasiswa	:	Naflah Faulina
NPM	:	**2227031011**
Program Studi	:	**Magister Matematika**
Jurusan	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

Menyetujui,

1. Komisi Pembimbing

Dr. Khoirin Nisa, S.Si., M.Si.
NIP. 19740726 200003 2 001

Ir. Warsono, M.S., Ph.D.
NIP. 19630216 198703 1 003

2. a.n. Ketua Program Studi Megister Matematika

Prof. Dr. Asmiati, S.Si., M.Si.
NIP. 19760411 200012 2 001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Khoirin Nisa, S.Si., M.Si.** .....

Sekretaris : **Ir. Warsono, M.S., Ph.D.** .....

Penguji
Bukan Pembimbing : 1. **Prof. Drs. Mustofa Usman, MA., Ph.** .....

2. **Dr. Aang Nuryaman, S.Si., M.Si.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 19711001 200501 1 002

3. Direktur Program Pascasarjana

Prof. Dr. Ir. Murhadi, M.Si.
NIP. 19640326 198902 1 001

4. Tanggal Lulus Ujian Tesis: **07 Agustus 2024**

PERNYATAAN TESIS MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	: **Naflah Faulina**
Nomor Pokok Mahasiswa	: **2227031011**
Program Studi	: **Magister Matematika**
Jurusan	: **Matematika**

Dengan ini menyatakan bahwa tesis saya yang berjudul, **"ANALISIS METODE *SMOTE, TOMEK LINKS,* DAN *HYBRID SMOTE+TOMEK LINKS* TERHADAP KLASIFIKASI NAIVE BAYES UNTUK MENGATASI DATA TIDAK SEIMBANG PADA DIAGNOSA PENYAKIT TUBERKULOSIS "** adalah hasil pekerjaan saya sendiri. Semua tulisan yang tertuang dalam tesis ini telah mengikuti kaidah karya penulisan ilmiah Universitas Lampung. Apabila kemudian hari terbukti bahwa tesis ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 7 Agustus 2024

Penulis,

Naflah Faulina

NPM. 2227031011

RIWAYAT HIDUP

Penulis bernama Naflah Faulina, lahir di Kotabumi pada tanggal 21 April 1998, dan merupakan anak kedua dari tiga bersaudara dari pasangan Bapak Surachman dan Ibu Yunwinarni.

Penulis menenempuh pendidikan Sekolah Dasar di SD Islam Ibnu Rusyd Kotabumi pada tahun 2004-2010, selanjutnya pada tahun 2010-2013 penulis melanjutkan pendidikan Sekolah Menengah Pertama di SMPN 07 Kotabumi dan tahun 2013-2016 penulis melanjutkan Sekolah Menengah Atas di SMAN 02 Bandar Lampung.

Pada Tahun 2017 penulis melanjutkan pendidikan Strata Satu (S1) di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

KATA INSPIRASI

"Sesungguhnya bersama kesulitan ada kemudahan"
(QS. Al-Insyirah : 6)

"Bila kau cemas dan gelisah akan sesuatu, masuklah ke dalamnya sebab ketakutan menghadapinya lebih menganggu daripada sesuatu yang kau takuti sendiri."
(Ali bin Abi Thalib)

PERSEMBAHAN

Alhamdulillah hirobbil'alamin,
Puji dan syukur tiada hentinya terpanjatkan kepada Allah SWT atas ridhonya sehingga penulis dapat menyelesaikan tesis ini. Saya persembahkan karya ini untuk:

Kepada kedua orang tuaku yang selalu memberikan doa dan dukungan terus menerus kepadaku. Terimakasih atas semuanya, orang tuaku benar benar luar biasa.

Kepada dosen-dosen Pembimbing dan Pembahas yang telah sangat sabar dalam membimbing dan memberikan masukan, ide-ide yang membangun sehingga dapat menyelesaikan tesis ini.

Sahabat tercinta, terimakasih atas keceriaan, doa dan semangat yang telah diberikan.

Almamater kebanggaan, Universitas Lampung.

SANWACANA

Puji syukur penulis panjatkan kehadirat Allah SWT. atas segala limpahan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan tesis yang berjudul **"Analisis Metode *SMOTE, Tomek Links,* Dan *Hybrid SMOTE+Tomek Links* Terhadap Klasifikasi Naive Bayes Untuk Mengatasi Data Tidak Seimbang Pada Diagnosa Penyakit Tuberkulosis"** sebagai syarat untuk memperoleh gelar Magister Matematika di Universitas Lampung.

Penulis menyadari bahwa dalam penulisan tesis ini tidak terlepas dari bimbingan dan bantuan berbagai pihak. Dengan selesainya penyusunan tesis ini, penulis mengucapkan terima kasih kepada :

1. Dr. Khoirin Nisa, S.Si., M.Si., selaku Pembimbing I yang telah memberikan waktu, tenaga, dan ilmu pengetahuan, serta senantiasa memberikan saran dan kritik yang membangun kepada penulis selama penyusunan tesis ini.
2. Ir. Warsono, M.S., Ph.D., selaku Pembimbing II yang telah memberikan waktu, tenaga, dan ilmu pengetahuan, serta senantiasa memberikan saran dan kritik yang membangun kepada penulis selama penyusunan tesis ini.
3. Prof. Drs. Mustofa Usman, MA., Ph.D. selaku Penguji I yang telah bersedia memberikan kritik dan saran serta evaluasi kepada penulis sehingga dapat lebih baik lagi.
4. Dr. Aang Nuryaman, S.Si., M.Si., selaku Penguji II yang telah bersedia memberikan kritik dan saran serta evaluasi kepada penulis sehingga dapat lebih baik lagi.
5. Dr. Khoirin Nisa, S.Si., M.Si., selaku Dosen Wali yang selalu membimbing, mendukung dan memberikan semangat kepada penulis.

6. Bapak dan Ibu dosen Program Studi Magister Matematika yang telah memberikan ilmu dengan ikhlas dan sabar selama penulis menyelesaikan studi di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Dan pihak-pihak lain yang tidak dapat penulis sebutkan satu persatu.

Semoga penulisan tesis ini dapat bermanfaat bagi ilmu pengetahuan. Penulis menyadari bahwa penulisan tesis ini masih jauh dari sempurna dan memiliki kekurangan. Oleh karena itu, penulis mengharapkan sumbangan pemikiran dari pembaca berupa kritik dan saran yang membangun guna menyempurnakan penulisan tesis ini.

Bandar Lampung, 7 Agustus 2024

Penulis

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Klasifikasi naive bayes adalah metode yang didasarkan pada teorema Bayes dari ilmuwan Inggris Revered Thomas Bayes dengan metode probabilitas dan statistik, yaitu cara memprediksi peluang di masa depan berdasarkan hasil pengalaman di masa sebelumnya. Berdasarkan teorema bayes, klasifikasi telah dikembangkan lebih lanjut oleh para peneliti dalam *machine learning* (Han et al., 2012). Metode naive bayes mengasumsikan bahwa efek dari suatu atribut pada kelas tertentu adalah independen dari atribut-atribut lainnya, yang dikenal sebagai asumsi independensi kondisional kelas. Meskipun metode naive bayes sederhana, metode ini memiliki banyak kelebihan, termasuk kecepatan, efisiensi, dan performa yang baik pada berbagai tugas klasifikasi. Oleh karena itu, naive bayes tetap menjadi salah satu algoritma yang sering digunakan dalam *machine learning* termasuk klasifikasi teks, diagnosis medis, dan manajemen kinerja sistem (Domingos & Pazzani, 1997).

Kasus yang ditemukan di dunia nyata adalah banyak data yang ditemui memiliki distribusi tidak seimbang di setiap kelasnya. Jenis data ini dikenal sebagai data tidak seimbang, yaitu kondisi ketidakseimbangan dalam jumlah data *training* antara dua kelas yang berbeda, salah satu kelasnya merepresentasikan jumlah data yang sangat besar *(majority class)* sedangkan kelas yang lainya merepresentasikan jumlah data yang sangat kecil *(minority class)* (Sastrawan et al., 2010). Permasalahan yang sering terjadi pada data tidak seimbang, klasifikasi cenderung memprediksi kelas yang memiliki komposisi data lebih besar. Akibatnya dihasilkan hasil akurasi prediksi yang baik pada kelas data *training* yang mayoritas, sedangkan akan dihasilkan akurasi prediksi yang buruk pada data

training yang minoritas (Sain & Purnami, 2015). Salah satu metode untuk mengatasi data tidak seimbang adalah melakukan teknik *resampling*. Teknik *resampling* adalah teknik *preprocessing* yang menyamakan distribusi kelas data secara algoritmik untuk meningkatkan *imbalance ratio* dan mengurangi efek distribusi kelas tidak seimbang dalam proses pembelajaran *machine learning*. Teknik *resampling* dapat dilakukan dengan metode *oversampling*, *undersampling*, dan gabungan keduanya (*hybrid*).

Oversampling bekerja pada kelas minoritas, kelas minoritas akan direplikasi sampai jumlah observasinya relatif sama dengan kelas mayoritas. Contoh teknik *oversampling* diantaranya adalah *Synthetic Minority Oversampling Technique* (SMOTE) yang menyeimbangkan data dengan membuat *instance* sintetis untuk kelas minoritas. Sedangkan *undersampling* mengurangi jumlah pengamatan dari kelas mayoritas untuk membuat kumpulan data menjadi seimbang. Tomek links merupakan metode *undersampling* yang menghapus data dari kelas mayoritas yang memiliki karakteristik yang serupa. Teknik *hybrid* adalah metode untuk mengatasi data tidak seimbang dengan cara menggabungkan metode *oversampling* dan *undersampling* sekaligus. Dengan mengkombinasikan dua metode tersebut, sebuah data set diharapkan tidak akan mengalami hilangnya informasi yang terlalu banyak (efek negatif dari *undersampling*) dan tidak mengalami *overfitting* (efek negatif dari *oversampling*). Salah satu teknik *hybrid* adalah SMOTE+Tomek Links.

Data tidak seimbang dalam bidang kesehatan adalah masalah umum. Di bidang kesehatan, klasifikasi naive bayes diterapkan secara luas untuk berbagai tujuan, termasuk prediksi penyakit, penilaian risiko, dan perkiraan hasil. Kemudahan dan efisiensinya membuatnya sangat cocok untuk aplikasi medis di mana pengambilan keputusan cepat berdasarkan model probabilistik. Berkenaan dengan situasi Tuberkulosis (TBC) di Indonesia, pada tanggal 2 Januari 2024, diperkirakan terdapat sekitar 1.060.000 kasus TBC. Data tahun 2023 menunjukkan penemuan sebanyak 792.404 kasus TBC, yang menunjukkan peningkatan setiap tahunnya. Menghadapi tingginya angka kasus TBC, didirikan gerakan TOSS TBC di

Indonesia, sebuah pendekatan yang bertujuan untuk menemukan, mendiagnosis, mengobati, dan menyembuhkan pasien TBC dengan tujuan utama menghentikan penularan penyakit ini di masyarakat.

Adapun penelitian mengenai klasifikasi dilakukan Asha et al., pada tahun 2011 membandingkan kinerja klasifikasi menggunakan pendekatan *machine learning* klasifikasi pada data Tuberkulosis. Model-model klasifikasi dilatih menggunakan data aktual yang dikumpulkan dari sebuah rumah sakit untuk memprediksi dua kategori utama: Tuberkulosis Paru (PTB) dan Tuberkulosis Retroviral (RPTB) yang terkait dengan AIDS. Hasil penelitian menunjukkan bahwa SVM (*Support Vector Machine*) unggul di antara klasifikasi, sedangkan *random forest* juga mencapai akurasi tinggi 99,14% untuk kedua jenis klasifikasi (Asha et al., 2011). Sedangkan beberapa penelitian sebelumnya yang telah dilakukan terkait mengatasi data tidak seimbang menggunakan teknik *resampling* yaitu, penelitian yang dilakukan oleh Tyagi et al. pada tahun 2020 mengevaluasi beberapa teknik *oversampling* dan *undersampling*, untuk mengatasi ketidakseimbangan data dan performa klasifikasi algoritma *K- Nearest Neighbors, Neural Networs,* dan *Support Vector Machines* pada beberapa dataset tidak seimbang. Pendekatan ADASYN memberikan hasil terbaik dalam menyeimbangkan data, tetapi secara keseluruhan performa klasifikasi terbaik didapat oleh teknik *undersampling* NCL (Tyagi & Mittal, 2020). Sedangkan penelitian oleh Sastrawan et al. pada tahun 2010 yang dilakukan pada penelitian ini adalah mengetahui bagaimana pengaruh metode *combine sampling* dalam klasifikasi K-Nearest Neighbor yang digunakan terhadap akurasi prediksi data *churn* dengan melakukan penghitungan akurasi model *churn prediction* yang dinyatakan dalam bentuk *lift curve, top decile* dan *gini coefficient* serta *f-measure* untuk hasil evaluasi yang berbeda terhadap dataset sebagai data *churn* dan sebagai data tidak seimbang (Sastrawan et al., 2010).

Berdasarkan beberapa penelitian di atas, peneliti akan menggunakan metode SMOTE , Tomek Links, dan SMOTE+Tomek Links terhadap klasifikasi *naive bayes* untuk mengatasi data tidak seimbang pada diagnosa penyakit Tuberkulosis di RSD. Mayjend HM Ryacudu Kotabumi. Faktor-faktor yang menunjang diagnosa Tuberkulosis adalah sebagai berikut: jenis kelamin, usia, status perokok, Indeks Massa Tubuh (IMT), riwayat penyakit TB di keluarga terdekat, dan Hasil Tes Cepat Molekuler (TCM).

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah mengaplikasikan metode *Synthetic Minority Oversampling Technique* (SMOTE), *Tomek Links*, dan *hybrid SMOTE+ Tomek Links* untuk mengatasi data tidak seimbang dan membandingkan kinerja (performa) klasifikasi *naive bayes* pada data original dan data yang telah seimbang dalam diagnosa penyakit Tuberkulosis di RSD. Mayjend HM Ryacudu Kotabumi.

1.3 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah untuk menambah wawasan tentang penerapan metode *Synthetic Minority Oversampling Technique* (SMOTE), *Tomek Links*, dan *hybrid SMOTE+ Tomek Links* untuk mengatasi data tidak seimbang dan memberikan informasi tentang kinerja terbaik klasifikasi *naive bayes* pada data original dan data yang telah seimbang dalam diagnosa penyakit Tuberkulosis di RSD. Mayjend HM Ryacudu Kotabumi.

II. TINJAUAN PUSTAKA

2.1 *Machine Learning*

Tahun 1959, Arthur Samuel pertama kali menggunakan istilah *machine learning*. Arthur Samuel mengatakan bahwa *machine learning* adalah bidang ilmu komputer yang memberikan kemampuan kepada komputer untuk dapat belajar tanpa diprogram secara explisit. Selain Samuel, menurut Mitchell (1997), juga memberikan sebuah definisi ringkas dan jelas mengenai ML dimana *machine learning* adalah satu program komputer yang dikatakan telah melakukan pembelajaran dari pengalaman *E* (*Experience*) terhadap tugas *T* (*Task*) dan mengukur peningkatan kinerja *P* (*Performance Measure*), jika kinerja Tugas *T* diukur oleh kinerja *P*, maka meningkatkan pengalaman *E*. Dari definisi ini Mitchel dapat dikatakan sebuah aplikasi *machine learning* memiliki 3 komponen yaitu *Task*, *Performance Measure*, dan *Experience*.

Machine learning dapat dikelompokkan berdasarkan bagaimana cara belajar sehingga dapat melakukan tugasnya. Pembagian *machine learning* berdasarkan cara belajarnya dibagi menjadi tiga kelompok yaitu : *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Menurut Kotsiantis (2007), jika contoh berlabel yang diketahui (sesuai output benar) maka *learning* disebut dengan *supervised*. Hal tersebut berbeda dengan *unsupervised learning*, yakni contoh tidak berlabel. Sedangkan, *reinforcement learning* mempunyai ide bahwa harus mengatasi tujuan tanpa adanya notifikasi dari komputer secara jelas jika tujuan tersebut telah tercapai. *Supervised learning* bertujuan untuk memprediksi hasil berdasarkan input, sedangkan tujuan *unsupervised learning* adalah menjelaskan hubungan, serta pola diantara data-data input.

2.2 Klasifikasi

Metode klasifikasi atau dapat juga disebut metode *supervised* merupakan proses memisahkan kelas data berdasarkan data yang ada untuk menentukan kelas data target. Klasifikasi digunakan dalam memprediksi kategori label kelas berdasarkan model yang telah dibangun dengan kumpulan data latih dan label kelas yang dapat mengklasifikasikan data pengujian yang baru (Jadhav & Channe, 2016).

Proses klasifikasi didasarkan pada empat komponen (Gorunescu, 2011):

1. Kelas atau label kelas adalah variabel terikat dari model yang merupakan variabel kategorik yang mewakili suatu label untuk objek setelah klasifikasi.
2. Prediktor adalah variabel bebas yang mewakili karakteristik untuk model data yang diklasifikasikan.
3. *Training* data set berisi kumpulan data yang berisi nilai dari dua komponen sebelumnya (kelas dan prediktor) yang digunakan untuk melatih model agar mengenali kelas, berdasarkan prediktor yang sudah ada.
4. *Testing* data set berisi data baru untuk diklasifikasikan dengan model yang telah dibuat, juga untuk mengukur tingkat akurasi klasifikasi, sehingga dapat dilihat efektivitas kinerja dari model klasifikasi.

Menurut Gorunescu (2011), dalam *machine learning*, *supervised learning* merupakan teknik yang digunakan untuk menyimpulkan suatu fungsi dari data *training*. Tujuan dari *supervised learning* adalah untuk memprediksi nilai (output) dari fungsi tersebut untuk setiap objek atau sampel baru (input) setelah proses pelatihan selesai. Teknik klasifikasi, sebagai metode prediktif, adalah salah satu contoh teknik *supervised learning*, dengan asumsi adanya sekelompok instance yang diberi label untuk setiap kategori objek. Secara ringkas, sebuah proses klasifikasi dicirikan oleh:

1. Input: Dataset *training* yang berisi objek-objek dengan atribut-atribut, di mana salah satu atribut tersebut adalah label kelas.

2. Output: Model *(classifier)* yang menetapkan label khusus untuk setiap objek (mengklasifikasikan objek ke dalam satu kategori) berdasarkan atribut-atribut lainnya.
3. *Classifier* tersebut digunakan untuk memprediksi kelas dari objek-objek baru yang belum diketahui. Dataset pengujian juga digunakan untuk menentukan akurasi model tersebut.

2.3 Teorema Bayes

Teorema bayes diperkenalkan oleh ilmuwan Inggris Revered Thomas Bayes, antara tahun 1702 dan 1761, yaitu cara memprediksi peluang di masa depan berdasarkan hasil pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes (Han et al., 2012).

Menurut Mitchell (1997), teorema Bayes menyatakan bagaimana kita dapat menghitung probabilitas dari suatu hipotesis B diberikan data A. Teorema ini memberikan cara untuk memperbarui probabilitas awal (*prior*) menjadi probabilitas setelah mengamati data (*posterior*) dengan memperhitungkan seberapa baik hipotesis tersebut memprediksi data yang diamati (*likelihood*).

Menurut Bain & Engelhardt (1992), peluang bersyarat dari kejadian *A* bersyarat *B*. Didefinisikan sebagai berikut:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad , P(B) > 0$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad , P(A) > 0$$

Kemudian $P(A \cap B) = P(B \cap A)$ dengan menggunakan aturan perkalian, menghasilkan :

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

$$P(B|A) = \frac{P(B)\,P(A|B)}{P(A)} \qquad (2.1)$$

Misalkan $B_1, B_2, \dots, B_n$ merupakan suatu partisi di dalam ruang sampel *S* dengan $P(B_i) \neq 0$ untuk $i = 1,2, \dots, n$ dan $P(B_1) + P(B_2) + \cdots + P(B_n) = 1$ serta misalkan terdapat kejadian sembaran *A*, dimana

$A = (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n)$ dan $(B_i \cap B_j) = \emptyset$ untuk setiap $i \neq j$ serta $(A \cap B_1), (A \cap B_2), \dots, (A \cap B_n)$ saling lepas, maka berlaku:

$$\mathrm{P}(A) = P\big((A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n)\big)$$

$$\mathrm{P}(A) = P\big((A \cap B_1) + (A \cap B_2) + \cdots + (A \cap B_n)\big)$$

$$\mathrm{P}(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

$$\mathrm{P}(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_n)P(A|B_n)$$

$$\mathrm{P}(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i). \qquad (2.2)$$

Dimana $P(A)$ disebut hukum peluang total.

Dengan menggunakan hukum peluang total Persamaan (2.2), aturan Bayes kemudian dapat dinyatakan dengan berlandasan pada definisi peluang bersyarat Persamaan (2.1) untuk menghasilkan $P(B_i|A)$ adalah jika *S* suatu ruang sampel dan $\{B_1, B_2, \dots, B_n\}$ kejadian yang tidak menenggang (saling terpisah) dengan peluang prior $P(B_1), P(B_2), \dots, P(B_n)$ dengan $P(B_i) \neq 0$. Untuk $i = 1,2, \dots, n$ dan jika kejadian B terjadi, maka peluang posterior dari B_i dengan syarat A telah terjadi adalah:

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)} \qquad (2.3)$$

dengan,

$P(B_i|A)$: peluang terjadinya B_i berdasarkan kondisi *A*

$P(B_i)$: peluang terjadinya B_i

$P(A|B_i)$: peluang terjadinya A berdasarkan kondisi pada hipotesis B_i.

$P(A)$: peluang terjadinya *A* (peluang total).

2.4 Klasifikasi Naive Bayes

Klasifikasi naive bayes adalah pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang didasarkan oleh ilmuwan Inggris Revered Thomas Bayes, antara tahun 1702 dan 1761, yaitu cara memprediksi peluang di masa depan berdasarkan hasil pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes (Han et al., 2012).

Naive Bayes adalah metode klasifikasi yang mengaplikasikan prinsip Teorema Bayes untuk menghitung probabilitas kelas dari data baru, dengan asumsi bahwa fitur-fitur dalam data adalah independen satu sama lain diberikan kelas (Murphy, 2012)

Algoritma Naive Bayes adalah algoritma klasifikasi yang didasarkan pada aturan Bayes. Dapat dituliskan dengan kata-kata sederhana sebagai berikut:

$$Posterior = \frac{Prior \times Likelihood}{Evidence} \tag{2.4}$$

Probabilitas *posterior*, dalam konteks masalah klasifikasi, dapat diartikan sebagai: "Berapa peluang suatu objek tertentu masuk ke dalam kelas *i* untuk nilai fitur yang diamati?" Contoh yang lebih konkrit adalah: "Berapa peluang seseorang menderita tuberkulosis?".

Notasi umum dari probabilitas *posterior* dapat ditulis sebagai berikut:

$$P(C_k|x_i) = \frac{P(C_k)\, P(\, x_i|C_k)}{P(x_i)} \tag{2.5}$$

Sehingga didapatkan rumus naive bayes sebagai berikut:

$$P(C_k|x_i) = P(C_k)\, P(\, x_i|C_k)$$
$$= \; P(C_k)\,.P(x_1|C_k).\,P(x_2|C_k)\,\ldots P(x_n|C_k)$$
$$= P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \qquad (2.6)$$

Untuk menghitung peluang *prior* yang merupakan peluang kelas C_k muncul sebelum sampel masuk dapat dihitung menggunakan persamaan sebagai berikut:

$$P(C_k) = \frac{N_{C_k}}{N} \qquad (2.7)$$

dengan,

N_{C_k} = Jumlah sampel dari kelas C_k

N = Jumlah semua sampel

Untuk menghitung peluang *likelihood* $P(x_i|C_k)$ terdapat dua aturan:

a. Jika data dari atribut x_i merupakan data bertipe kategorik maka nilai $P(x_i|C_k)$ adalah jumlah kejadian di mana fitur x_i terjadi pada kelas C_k dibagi dengan jumlah total kejadian pada kelas C_k.

$$P(x_i|C_k) = \frac{n(x_i \cap C_k)}{n(C_k)} \qquad (2.8)$$

b. Jika data dari atribut x_i bertipe kontinu maka untuk mencari nilai $P(x_i|C_k)$ diasumsikan mengikuti distribusi Normal dengan parameter mean μ dan standar deviasi σ sebagai berikut:

$$g(x, \mu\, , \sigma) = \frac{1}{\sqrt{2\pi\, \sigma^2}}\; e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2.9)$$

sehingga , $P(x_i|C_k) = g\left(x_i, \mu_{C_k}, \sigma_{C_k}\right)$. dengan, $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$

dan $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}$

Hasil klasifikasi naive bayes ditentukan dengan melihat nilai $P(C_k|x_i)$ paling besar dari setiap variabel x_i (Han et al., 2012).

2.5 Fenomena Data Tidak Seimbang

Suatu fenomena data tidak seimbang adalah kondisi ketidakseimbangan dalam jumlah data *training* antara dua kelas yang berbeda, salah satu kelasnya merepresentasikan jumlah data yang sangat besar (*majority class*) sedangkan kelas yang lainya merepresentasikan jumlah data yang sangat kecil (*minority class*) (Sastrawan et al., 2010).

Ketidakseimbangan kelas adalah salah satu faktor paling berpengaruh dalam kinerja prediksi klasifikasi. Pengklasifikasi cenderung akan membuat model pembelajaran bias yang memiliki akurasi prediksi yang buruk atas kelas minoritas dibandingkan dengan kelas mayoritas. Ini karena sebagian besar pembelajaran pengklasifikasi, seperti *naive bayes, decision tree, backpropagation neural network, support vector machines,* dan lainnya dirancang berdasarkan dengan asumsi bahwa distribusi kelas relatif seimbang (Zheng & Jin, 2020).

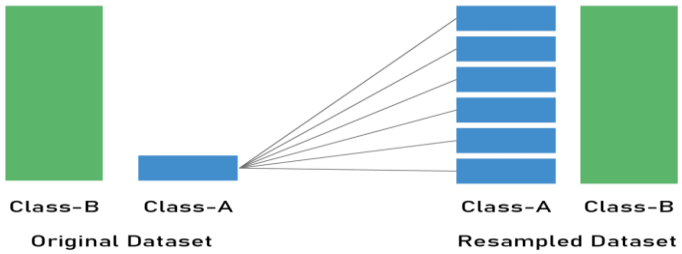
Terdapat tiga pendekatan untuk menangani dataset tidak seimbang, yaitu pendekatan pada level data, level algoritmik, dan menggabungkan metode. Pendekatan pada level data mencakup berbagai teknik *resampling* dan sintesis data untuk memperbaiki kecondongan distribusi kelas data latih. Pada tingkat algoritmik, metode utamanya adalah menyesuaikan operasi algoritma yang ada untuk membuat pengklasifikasi agar lebih konduktif terhadap klasifikasi kelas minoritas. Pada pendekatan algoritma dan ensemble memiliki tujuan yang sama, yaitu memperbaiki algoritma pengklasifikasi tanpa mengubah data, sehingga dapat dianggap ada dua pendekatan saja, yaitu pendekatan level data dan pendekatan level algoritma. Dengan membagi menjadi dua pendekatan dapat mempermudah fokus objek perbaikan, pendekatan level data difokuskan pada pengolahan awal data, sedangkan pendekatan level algoritma difokuskan pada perbaikan algoritma atau menggabungkan (Yap et al., 2014).

2.6 Teknik *Resampling*

Teknik *resampling* adalah salah satu teknik *preprocessing* di mana distribusi data diseimbangkan kembali untuk mengurangi efek distribusi kelas tidak seimbang dalam proses pembelajaran. Teknik resampling menyamakan distribusi kelas secara algoritmik untuk meningkatkan *imbalance ratio* dan mengurangi efek distribusi kelas tidak seimbang dalam proses pembelajaran *machine learning*. Teknik *resampling* dapat dilakukan dengan metode *oversampling*, *undersampling*, dan gabungan keduanya (Tyagi & Mittal, 2020).

2.6.1 *Oversampling*

Oversampling adalah metode pembangkitan data kelas minoritas agar mendekati atau sama dengan kelas mayoritas. Dalam *oversampling* ini, kelas minoritas akan direplikasi sampai jumlah observasinya relatif sama dengan kelas mayoritas (Chawla et al., 2002). Ilustrasi metode *oversampling* dapat dilihat pada Gambar 1 berikut.



Gambar 1. Ilustrasi *Oversampling*

Contoh teknik *oversampling* adalah *Synthetic Minority Over-sampling Technique* (SMOTE), *Adaptive Synthetic Sampling* (ADASYN), dan *Random Over Sampling* (ROS).

2.6.1.1 *Synthetic Minority Oversampling Technique (SMOTE)*

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode yang populer diterapkan dalam rangka menangani ketidak seimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas (Siringoringo, 2018).

SMOTE merupakan salah satu metode *oversampling* yaitu teknik pengambilan sampel untuk meningkatkan jumlah data pada kelas positif (minoritas) dengan cara mereplikasi jumlah data pada kelas positif (minoritas) secara acak sehingga jumlahnya sama dengan data pada kelas negatif (mayoritas). Algoritma SMOTE pertama kali ditemukan oleh (Chawla et al., 2002) pendekatan ini bekerja dengan membuat *synthetic* data, yaitu data replikasi dari data minor. Metode SMOTE bekerja dengan mencari *K-Nearest Neighbors* (ketetanggaan data). Teknik ini bekerja dengan mengelompokkan data berdasarkan tetangga terdekat. Tetangga terdekat dipilih berdasarkan jarak *euclidean* antara kedua data.

Misalkan diberikan dua data dengan *p* dimensi yaitu:

$$x^T = [x_1, x_2, \dots, x_p] \text{ dan } y^T = [y_1, y_2, \dots, y_p] \tag{2.10}$$

maka jarak *euclidean* $d(x, y)$ antara kedua vektor data adalah sebagai berikut,

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \tag{2.11}$$

sedangkan *synthetic* data dilakukan dengan menggunakan persamaan berikut:

$$x_{syn} = x_i + (x_{knn} - x_i) \times \beta, i = 1,2, \dots, n \tag{2.12}$$

dengan,

x_{syn} : data hasil replikasi.

x_i : data yang akan direplikasi.

x_{knn} : data yang memiliki jarak *euclidean* terdekat dari data yang akan direplikasi.

β : bilangan random antara 0 sampai 1.

Pembangkitan data buatan yang berskala numerik berbeda dengan kategorik. Data numerik diukur jarak kedekatannya dengan jarak *Euclidean* sedangkan data kategorik, *nearest neighbor* dihitung menggunakan versi modifikasi dari *Value Difference Metric* (VDM) yang diajukan oleh Cost dan Salzberg pada Tahun 1993. Matriks mendefinisikan jarak antara nilai fitur yang sesuai untuk vektor fitur yang dibuat (Chawla et al., 2002). Jarak δ antara dua nilai fitur yang sesuai didefinisikan sebagai berikut:

$$\Delta(A,B) = \sum_{i=1}^{N} \delta(V_{1i}, V_{2i}) \tag{2.13}$$

dengan,

$\Delta(A,B)$: jarak antara amatan A dengan amatan B

N : banyaknya variabel independen

$\delta(V_{1i}, V_{2i})$: jarak antara amatan A dan B untuk setiap variabel yang dihitung

Untuk menentukan jarak antar amatan A dan B untuk setiap variabel maka digunakan persamaan:

$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right| \tag{2.14}$$

dengan,

n : banyaknya kategori pada varibel ke-i

C_{1i} : banyaknya kategori-1 yang termasuk pada variabel ke-i

C_{2i} : banyaknya kategori-2 yang termasuk pada variabel ke-i

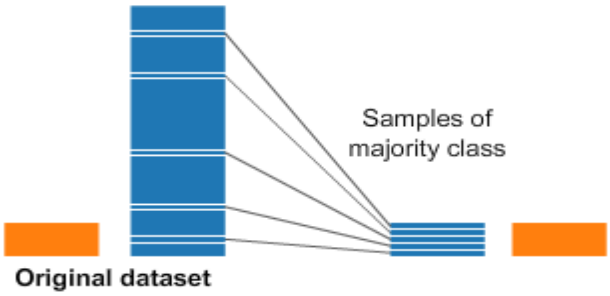
C_1 : banyaknya kategori ke-1 terjadi

C_2 : banyaknya kategori ke-2 terjadi

Sedangkan untuk menghasilkan vektor fitur kelas minoritas baru (*synthetic*), dapat menciptakan nilai fitur baru dengan mengambil suara mayoritas dari vektor fitur yang sedang dipertimbangkan dan k- tetangga terdekatnya (Chawla et al., 2002).

2.6.2 *Undersampling*

Undersampling bekerja pada kelas mayoritas dengan mengurangi jumlah pengamatan dari kelas mayoritas untuk membuat kumpulan data menjadi seimbang (Drummond & Holte, 2003). Ilustrasi metode *undersampling* dapat dilihat pada Gambar 2 berikut ini.



Gambar 2. Ilustrasi *Undersampling*

Contoh teknik *undersampling* adalah Tomek Links, *Random Under Sampling* (RUS), *Edited Nearest Neighbors* (ENN) , dan *One Side Selection* (OSS).

2.6.2.1. *Tomek Links*

Menurut (Tomek, 1973) *tomek links* adalah pasangan contoh dari kelas yang berbeda yang merupakan tetangga terdekat satu sama lain. Dalam konteks ini, sepasang contoh disebut *Tomek Link* jika jarak antara kedua contoh tersebut adalah yang terkecil dibandingkan dengan jarak contoh tersebut ke contoh lain dari kelas yang berbeda. Metode ini digunakan untuk mengidentifikasi dan menghapus contoh yang membingungkan, terutama dalam dataset yang tidak seimbang, guna meningkatkan kualitas data dan performa model klasifikasi.

Tomek links merupakan salah satu metode *undersampling*, yang diperkenalkan oleh Tomek pada Tahun 1976. Metode ini bekerja dengan menghapus data kelas negatif (mayoritas) yang merupakan kasus *borderline* atau yang memiliki

kesamaan karakteristik. *Tomek links* dapat digunakan sebagai mencari contoh yang merupakan *tomek links* menggunakan *1-NN* untuk dataset yang diberikan. Untuk mengurangi ketidakseimbangan, contoh kelas mayoritas yang terlibat dalam *tomek links* dihapus. *tomek links* dapat digunakan sebagai teknik *undersampling* atau sebagai langkah pembersihan setelah pemrosesan (Pereira et al., 2020).

Menurut (Batista et al., 2003), misalkan terdapat a dan b dimana $\delta(a, b)$ adalah jarak *euclidean* antara a dan b. Jika a dan b masuk ke dalam kelas yang berbeda dan tidak terdapat observasi lain misalnya c, maka a dan b disebut observasi *tomek links* sedemikian rupa sehingga

$$\delta(a, c) < \delta(a, b) \text{ atau } \delta(b, c) < \delta(b, a). \quad (2.15)$$

Dengan kata lain, a dan b termasuk ke dalam *tomek links* apabila:

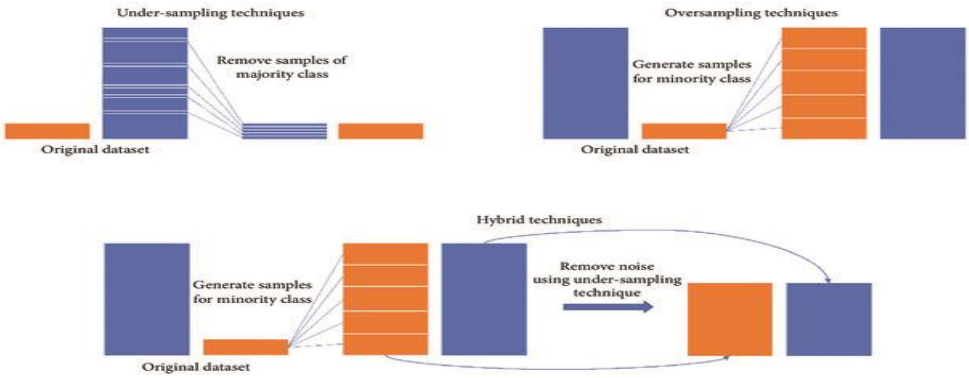
1. lingkungan terdekat a adalah b,
2. lingkungan terdekat b adalah a,
3. a dan b berada pada kelas yang berbeda.

2.6.3 *Hybrid Oversampling* dan *Undersampling*

Teknik *oversampling* dan *undersampling* dapat digabungkan untuk mengurangi *noise* dataset yang dihasilkan oleh *oversampling* dengan *undersampling* sebagai metode pembersih. Teknik *hybrid* adalah metode untuk mengatasi data tidak seimbang dengan cara menggabungkan metode *undersampling* dan *oversampling* sekaligus. Dengan mengkombinasikan dua metode tersebut, sebuah data set diharapkan tidak akan mengalami hilangnya informasi yang terlalu banyak (efek negatif dari *undersampling*) dan tidak mengalami *overfitting* (efek negatif dari

oversampling). Teknik *hybrid* yang telah dikembangkan dan cukup familiar yaitu: SMOTE+Tomek Links dan SMOTE+ENN (Batista et al., 2004).

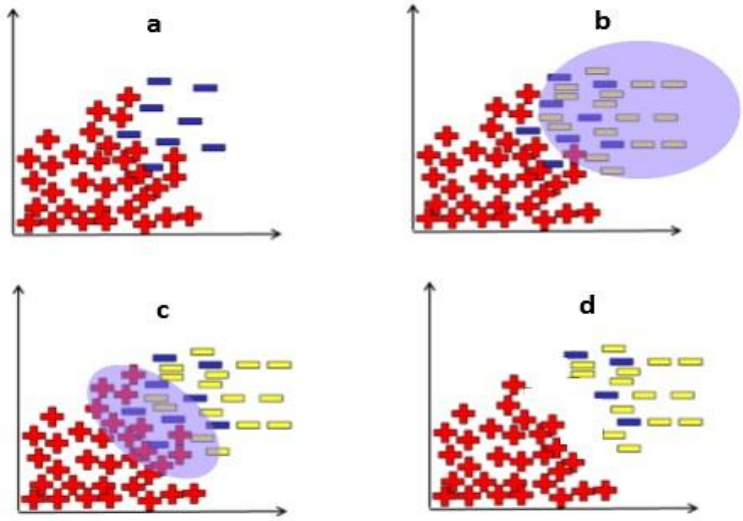
Dimana SMOTE berguna untuk melakukan *oversample* pada kelas minoritas, sedangkan Tomek Links digunakan untuk melakukan *undersample* pada kelas mayoritas.



Gambar 3. Kombinasi *Oversampling*dan *Undersampling*

2.6.3.1. *SMOTE+Tomek Links*

SMOTE+Tomek Links melakukan *sampling* terhadap data dengan SMOTE dan *cleaning* dengan *tomek links* (Batista et al., 2004). Metode ini merupakan metode kombinasi antara SMOTE dan *tomek link* ssebagai metode pembersihan data. Cara kerja *tomek links* adalah dengan menghapus data minor ataupun mayor yang memiliki kesamaan karakteristik. Untuk setiap data, jika satu tetangga yang paling dekat memiliki kelas label yang berbeda dengan data tersebut maka kedua data akan dihapus karena dianggap sebagai *noise* atau *misclassify*.



Gambar 4. Ilustrasi SMOTE+Tomek Links

Ilustrasi langkah-langkahnya dapat dilihat pada Gambar 4. Data contoh pada Gambar 4(a) akan di *oversampling* dengan metode SMOTE sehingga menghasilkan data dengan karakteristik seperti Gambar 4(b). Kemudian di Gambar 4(c) memperlihatkan metode *tomek links* bekerja dengan pengecekan setiap tetangga terdekat untuk tiap data. Apabila ditemukan tetangga yang memiliki kelas label berbeda, maka kedua data itu akan dihapus dari data *training* sampai menghasilkan data *training* yang bersih dari *noise* seperti pada Gambar 4(d).

2.7 *K-Fold Cross Validation*

K-fold cross validation adalah sebuah metode proses validasi untuk memperkirakan kinerja dari model pembelajaran mesin atau *machine learning*. Kerja dari *k-fold cross validation* adalah data pertama dipartisi menjadi *k* atau segmen yang berukuran sama (atau hampir sama). Selanjutnya dilakukan sejumlah *k-fold cross validation* dengan masing-masing validasi menggunakan data partisi ke-*k* sebagai data *testing* dan menggunakan sisa partisi lainnya sebagai data *training*, tahap selanjutnya menghitung rata-rata akurasi dari *k-fold cross validation* yang digunakan sebagai validasi final (Olson & Delen, 2008).

2.8 Evaluasi Kinerja (*Performance*) Klasifikasi

Confusion Matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan menunjukkan perbandingan antara prediksi yang dibuat oleh model dan hasil sebenarnya. *Confusion matrix* membantu dalam memahami dan menginterpretasikan kesalahan yang dibuat oleh model klasifikasi serta menghitung berbagai metrik evaluasi seperti akurasi, *sensitivity*, *specificity*, *False Positive Fraction* (FPF), dan *False Negative Fraction* (FPF) (Sullivan, 2018).

Tabel 1. *Confusion matrix*

	Aktual		Total
Prediksi	Positif (A)	Negatif ($\bar{A}$)	
Positif (P^+)	a	b	$a+b$
Negatif (P^-)	c	d	$c+d$
Total	$a+c$	$b+d$	n

dengan,

1. *a* adalah jumlah subjek yang mengidap penyakit tersebut dan hasil prediksinya positif.
2. *b* adalah jumlah subjek yang tidak mengidap penyakit tersebut dan hasil prediksinya positif.
3. *c* adalah jumlah subjek yang mengidap penyakit tersebut dan hasil prediksinya negatif.
4. *d* adalah jumlah subjek yang tidak mengidap penyakit tersebut dan hasil prediksinya negatif.

Kinerja klasifikasi dievaluasi dengan memperhatikan ukuran-ukuran sebagai berikut:

a. Akurasi

   Akurasi adalah metode yang didasari tingkat kedekatan antara nilai prediksi dengan nilai sebenarnya. Akurasi adalah hasil dari penjumlahan nilai

diagonal dibagi dengan jumlah total keseluruhan data dan selajutnya dikalikan 100% (Gorunescu, 2011). Rumus untuk menghitung akurasi klasifikasi sesuai persamaan berikut.

$$Akurasi = \frac{P(P^+ \cap A) + P(P^- \cap \bar{A})}{P(P^+ \cap A) + P(P^- \cap \bar{A}) + P(P^+ \cap \bar{A}) + P(P^- \cap A)}$$

$$Akurasi = \frac{\frac{a}{n} + \frac{d}{n}}{\frac{a}{n} + \frac{d}{n} + \frac{b}{n} + \frac{c}{n}} = \frac{\frac{a+d}{n}}{\frac{a+d+b+c}{n}} = \frac{a+d}{n} \times 100\% \quad (2.16)$$

b. *Sensitivity*

Sensitivity juga disebut *true positive fraction* dan didefinisikan sebagai *probability* pasien yang sakit dinyatakan positif (Sullivan, 2018). Rumus untuk menghitung *sensitivity* klasifikasi sesuai persamaan berikut.

$$Sensitivity = True\ Positive\ Fraction = P(P^+|A)$$

$$Sensitivity = \frac{P(P^+ \cap A)}{P(A)} = \frac{\frac{a}{n}}{\frac{a+c}{n}} = \frac{a}{a+c} \times 100\% \quad (2.17)$$

c. *Specificity*

Specificity juga disebut *true negatif fraction* dan didefinisikan sebagai *probability* pasien yang tidak sakit dinyatakan negatif (Sullivan, 2018). Rumus untuk menghitung *specificity* klasifikasi sesuai persamaan berikut.

$$Specificity = True\ Negative\ Fraction = P(P^-|\bar{A})$$

$$Specificity = \frac{P(P^- \cap \bar{A})}{P(\bar{A})} = \frac{\frac{d}{n}}{\frac{b+d}{n}} = \frac{d}{b+d} \times 100\% \quad (2.18)$$

d. *False Positive Fraction* (FPF)

False Positive Fraction (FPF) adalah $1 - specificity$, yang menunjukkan peluang hasil prediksi positif padahal penyakit atau kondisi yang diperiksa sebenarnya tidak ada.

$$False\ Positive\ Fraction = P(P^+|\bar{A})$$

$$FPF = \frac{P(P^+ \cap \bar{A})}{P(\bar{A})} = \frac{\frac{b}{n}}{\frac{b+d}{n}} = \frac{b}{b+d} \times 100\% \quad (2.19)$$

e. *False Negative Fraction* (FNF)

False Negative Fraction (FNF) adalah $1-sensitivity$, yang menunjukkan peluang hasil prediksi negatif padahal penyakit atau kondisi yang diperiksa sebenarnya ada.

$$False\ Negative\ Fraction\ = P(P^{-}|A)$$

$$FNF = \frac{P(P^{+} \cap A)}{P(A)} = \frac{\frac{c}{n}}{\frac{a+c}{n}} = \frac{c}{a+c} \times 100\% \qquad (2.20)$$

2.9 Tuberkulosis

Tuberkulosis yang biasa disingkat TB merupakan penyakit menular yang disebabkan oleh bakteri yang dapat menyerang paru dan organ lainnya yaitu *mycobaterium tuberculosis, mycobacterium africanum, mycobacterium bovis, mycobacterium Leprae* (Kemenkes RI, 2011).

Mendiagnosa TB dapat dilakukan pada penderita TB dewasa yaitu, pemeriksaan sputum, x-ray dan tes tuberkulin. Namun, pemeriksaan yang paling baik adalah pemeriksaan sputum. Hal ini dikarenakan x-ray hanya menggambarkan ketidaknormalan pada paru-paru yang dapat terjadi karena sebab lain. Begitu juga dengan tes tuberkulin yang hanya dapat mengindikasikan pernah tidaknya seseorang terinfeksi dengan kuman Tuberkulosis. Pemeriksaan sputum dilakukan tiga kali yang dikenal dengan istilah SPS. Pemeriksaan sputum berfungsi untuk menegakkan diagnosis, menilai keberhasilan pengobatan dan menentukan potensi penularan. Pemeriksaan dahak untuk penegakan diagnosis dilakukan dengan mengumpulkan tiga spesimen dahak yang dikumpulkan dalam dua hari kunjungan yang berurutan berupa Sewaktu Pagi Sewaktu (SPS).

a. S (sewaktu): dahak dikumpulkan pada saat suspek TB datang berkunjung pertama kali. Pada saat pulang, suspek membawa sebuah pot dahak untuk mengumpulkan dahak pagi pada hari kedua.
b. P (pagi): dahak dikumpulkan di rumah pada pagi hari kedua, segera setelah bangun tidur. Pot dahak dibawa dan diserahkan sendiri kepada petugas.

c. S (sewaktu): dahak dikumpulkan pada hari kedua, saat menyerahkan dahak pagi. Pengambilan tiga spesimen dahak masih diutamakan dibanding dengan dua spesimen dahak mengingat masih belum optimalnya fungsi sistem dan hasil jaminan mutu eksternal pemeriksaan laboratorium (Kemenkes RI, 2011).

Faktor-faktor yang menunjang diagnosa TB adalah sebagai berikut:

1. Jenis kelamin

   Pada kasus MTB+ menurut data jenis kelamin pada laki- laki lebih tinggi dari pada perempuan yaitu hanpir 5,1 kali dibandingkan kasus MTB+ pada perempuan. Disparitas paling tinggi antara laki-laki dan perempuan diantaranya terjadi di Sumatera Utara, dengan kasus dua kali lipat dari kasus pada perempuan (Anggraeni et al., 2015)

2. Usia

   Berdasarkan data Riskesdas 2018 bahwa kelompok umur kasus baru yang ditemukan banyak terdapat pada kelompok usia 65-74 tahun sebanyak 1% atau 40.180 kasus.

3. Status perokok

   Merokok dapat mengganggu efektifitas sebagian mekanisme pertahanan respirasi atau pernapasan. Asap rokok dapat menurunkan pergerakan silia dan merangsang pembentukan mukus, sehingga akan terjadi penimbunan mukosa dan peningkatan resiko pertumbuhan bakteri termasuk kuman micobacterium tuberculosis yaitu kuman penyebab TB paru, sehingga dapat menimbulkan infeksi. Penelitian yang telah dilakukan oleh Rustono, 2008 melaporkan bahwa memiliki kebiasaan merokok beresiko 2,56 kali lebih besar terkena TB paru jika dibandingkan dengan yang tidak pernah merokok (Rusnoto, 2016)

4. Indeks Massa Tubuh (IMT)

   Nilai IMT diperoleh dari perbandingan antara berat badan (kg) dan tinggi badan kuadrat (m) seperti pada rumus berikut:

$$IMT = \frac{\text{berat badan (kg)}}{\text{tinggi badan } (m)^2} \tag{2.21}$$

5. Riwayat penyakit TB di keluarga terdekat

   Pada waktu batuk atau bersin, penderita TB paru menyebarkan kuman ke udara dalam bentuk droplet atau percikan dahak yang mengandung kuman yang dapat bertahan diudara sampai beberapa jam pada suhu kamar. Jika droplet terhirup kedalam saluran pernafasan, dan masuk ke dalam tubuh lainnya melalui sistem peredaran darah, sistem saluran limfe, saluran nafas, atau penyebaran langsung ke bagian-bagian tubuh lainnya maka orang tersebut akan terinfeksi olek kuman Mycobaterium Tuberculosis. daya penularan seseorang ditentukan oleh banyaknya kuman yang dikeluarkan oleh parunya atau konsentrasinya driplet dalam udara, sehingga daya tahan tubuh yang rendah menjadi faktor seseorang mudah terserang penyakit TB paru, diantaranya gizi buruk atau HIV/AIDS (Dziri et al., 2024).

6. Hasil Tes Cepat Molekuler (TCM)

   Pemeriksaan TCM dengan Xpert MTB/RIF merupakan metode deteksi molekuler berbasis *nested real-time* PCR untuk diagnosis TB. Primer PCR yang digunakan mampu mengamplifikasi sekitar 81 bp daerah inti gen rpoB MTB kompleks, sedangkan probe dirancang untuk membedakan *sekuen wild type* dan mutasi pada daerah inti yang berhubungan dengan resistansi terhadap rifampisin (Kemenkes RI, 2017).

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester genap tahun ajaran 2023/2024 di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder dari bulan Januari-Desember Tahun 2023 yaitu, data rekam medis diagnosa penyakit tuberkulosis yang diperoleh dari RSD. Mayjend HM Ryacudu Kotabumi. Data yang digunakan adalah data 1033 pasien yang melakukan pemeriksaan kesehatan diagnosa tuberkulosis. Data penelitian ini memiliki 6 variabel sebagai fitur yaitu: jenis kelamin (X_1), usia (X_2), status perokok (X_3), Indeks Massa Tubuh (IMT) (X_4), riwayat penyakit tuberkulosis di keluarga terdekat (X_5), dan hasil Tes Cepat Molekuler (TCM) (X_6) dengan variabel target diagnosa (Y). Daftar variabel penelitian ini disajikan pada Tabel 2 sebagai berikut.

Tabel 2. Variabel Penelitian Rekam Medis Diagnosa Penyakit Tuberkulosis

No.	Variabel	Deskripsi	Jenis	Keterangan
Kelas Fitur				
1.	X_1	Jenis Kelamin	Kategorikal	0 = Perempuan 1 = Laki-laki
2.	X_2	Usia	Numerik	Tahun
3.	X_3	Status Perokok	Kategorikal	0 = Tidak beresiko (jika pasien tidak perokok) 1 = Beresiko (jika pasien perokok)
4.	X_4	Indeks Massa Tubuh	Numerik	$IMT = \frac{\text{berat badan (kg)}}{\text{tinggi badan } (m)^2}$
5.	X_5	Riwayat Penyakit TB Di Keluarga Terdekat	Kategorikal	0 = Tidak ada anggota keluarga yang menderita TB 1 = Ada anggota keluarga yang menderita TB
6.	X_6	Tes Cepat Molekuler	Kategorikal	0 = Hasil tes menunjukkan MTB tidak terdeteksi 1 = Menunjukkan MTB terdeteksi
Kelas Target				
7.	Y	Diagnosa	Kategorikal	0= Tidak terdiagnosa TB 1= Terdiagnosa TB

3.3 Metode Penelitian

Penelitian ini menggunakan bantuan *software* Rstudio dengan versi 4.2.2.

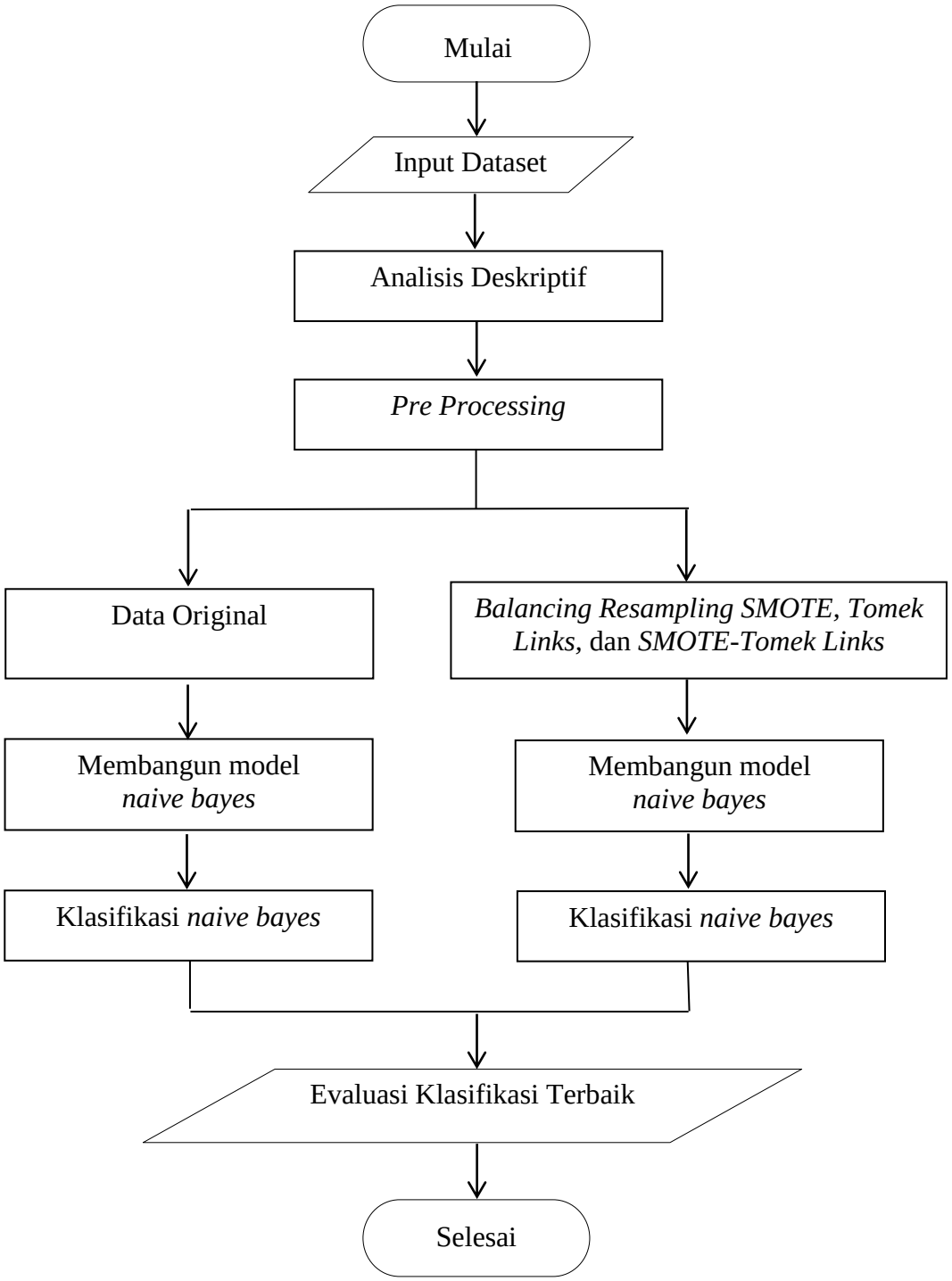
3.3.1 Analisis Data

Penelitian ini menggunakan metode SMOTE, Tomek Links, dan SMOTE+Tomek Links terhadap klasifikasi Naive Bayes untuk mengatasi data tidak seimbang. Adapun langkah-langkah yang dilakukan dalam penelitian ini sebagai berikut:

1. melakukan analisis deskriptif.

2. melakukan *preprocessing* data, yaitu :
    a. pengkodean variabel kategorik.
    b. melakukan pembagian data *training* dan data *testing* dengan perbandingan 80:20.
    c. dalam data *training* gunakan 5-*fold cross validation.*
3. melakukan proses klasifikasi *naive bayes* dengan data original (tidak diseimbangkan).
    a. membangun model *naive bayes* pada data *training*.
    b. melakukan klasifikasi menggunakan data *testing* .
    d. melakukan evaluasi model menggunakan *confusion matrix* pada data *testing.*
4. menyeimbangkan data menggunakan SMOTE.
    a. menentukan jumlah data kelas mayoritas dan kelas minoritas.
    b. menentukan persentase SMOTE yang digunakan (N%) dengan cara (jumlah data kelas mayoritas/jumlah data kelas minoritas) $\times$ 100% .
    c. untuk data kategorikal menentukan k-*nearest neighbour* dengan jarak *value difference metric* dan data numerik menggunakan jarak *euclidean* terdekat dari setiap data minoritas untuk membangkitkan data sintesis.
    d. menentukan data sintetis berdasarkan suara mayoritas dari fitur yang sedang dipertimbangkan dan k- *nearest neighbour* nya.
5. menyeimbangkan data menggunakan Tomek Links.
    a. menentukan jumlah data kelas mayoritas dan kelas minoritas.
    b. melakukan pengecekan setiap data dari kelas yang berbeda dengan menggunakan jarak *value difference metric* untuk data kategorikal dan data numerik menggunakan jarak *euclidean*. Apabila ditemukan sepasang data yang memiliki kelas label berbeda dan merupakan kasus Tomek Links, maka data dari kelas mayoritas akan dihapus dari data *training.*

6. menyeimbangkan data menggunakan *hybrid* SMOTE+Tomek Links.
    a. menentukan jumlah data kelas mayoritas dan kelas minoritas.
    b. meningkatkan jumlah sampel pada kelas minoritas menggunakan SMOTE.
    c. identifikasi Tomek Links pada data hasil SMOTE.
7. melakukan proses klasifikasi *naive bayes* dengan data yang diseimbangkan.
    a. membangun model *naive bayes* pada data *training*.
    b. melakukan klasifikasi menggunakan data *testing.*
    c. melakukan evaluasi model menggunakan *confusion matrix* pada data *testing.*
8. membandingkan hasil evaluasi pada data yang tidak diseimbangkan dan data yang diseimbangkan.



Gambar 5. Diagram Alir Penelitian.

KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan yaitu menyeimbangkan data *training* dengan mengaplikasikan metode SMOTE, Tomek Links, dan *hybrid* SMOTE+ Tomek Links untuk membandingkan kinerja (performa) klasifikasi naive bayes pada data original dan data seimbang dalam diagnosa penyakit Tuberkulosis di RSD. Mayjend HM Ryacudu Kotabumi diperoleh bahwa kombinasi antara teknik *oversampling* dengan SMOTE dan teknik *undersampling* dengan Tomek Links menunjukkan kinerja terbaik dalam menyeimbangkan data dan meningkatkan akurasi model naive bayes, dengan akurasi rata-rata mencapai 93%, *sensitivity* rata-rata 88%, *specificity* rata-rata 96%, rata-rata *False Positive Fraction(*FPF*)* 4%, dan rata-rata *False Negative Fraction (*FNF*)* 12%.

5.2 Saran

Hasil penelitian ini membuka peluang bagi penelitian lebih lanjut untuk mengeksplorasi kombinasi teknik penyeimbangan lainnya atau penerapan teknik ini pada algoritma klasifikasi lain untuk meningkatkan performa model. Selain itu, pengujian pada dataset yang lebih besar dan lebih beragam dapat memberikan wawasan tambahan tentang efektivitas teknik ini dalam berbagai konteks medis.

DAFTAR PUSTAKA

Anggraeni, S. K., Raharjo, M., & Nurjazuli. (2015). Hubungan Kualitas Lingkungan Fisik Rumah dan Perilaku Kesehatan dengan Kejadian TB Paru di Wilayah Kerja Puskesmas Gondanglegi Kecamatan Gondanglegi Kabupaten Malang. *Jurnal Kesehatan Masyarakat*, *3*(1), 559–568.

Asha, T., Natarajan, S., & Murthy, K. N. B. (2011). Effective Classification Algorithms to Predict the Accuracy of Tuberculosis-A Machine Learning Approach. *Effective Classification Algorithms to Predict the Accuracy of Tuberculosis-A Machine Learning Approach*, *9*(7), 89–94.

Bain, L. J., & Engelhardt, M. (1992). *Introduction To Probability and Mathematical Statistics*.

Batista, Bazzan, A. L. C., Monard, M.-C., Batista, G. E. A. P. A., & Monard, M. C. (2003). Balancing Training Data for Automated Annotation of Keywords: a Case Study. *In: Proceedings of the Second Brazilian Workshop on Bioinformatics*, *4*(1), 35–43.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A Study Of The Behavior Of Several Methods For Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(2), 321–357. https://doi.org/10.1613/jair.953

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, *29*(2–3), 103–130. https://doi.org/10.1023/a:1007413511361

Drummond, C., & Holte, R. C. (2003). Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. *Physical Review Letters*, *91*(3).

Dziri, S., Marin, J., Quagliaro, P., Genestet, C., Dumitrescu, O., Carbonnelle, E., & Billard-Pomares, T. (2024). Optimization of Mycobacterium Tuberculosis DNA Processing Prior to Whole Genome Sequencing. *Tuberculosis*, *148*(4), 102543. https://doi.org/10.1016/j.tube.2024.102543

Gorunescu, F. (2011). Data Mining Concepts, Model and Techniques. In *Springer*.

Han, J., Kambe, M., & Pe, J. (2012). Data Mining Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Jadhav, S. D., & Channe, H. P. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR)*, *5*(1), 1842–1845. https://doi.org/10.21275/v5i1.nov153131

Kemenkes RI. (2011). Pedoman Nasional Pengendalian Tuberkulosis. In *Chemotherapy* (Vol. 52, Issue 1). https://doi.org/10.1159/000090244

Kemenkes RI. (2017). Petunjuk Teknis Pemeriksaan TB Menggunakan Tes Cepat Molekuler. In *Kemenkes RI*.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. https://doi.org/10.1007/s10751-016-1232-6

Mitchell, T. M. (1997). Machine Learning. In *McGraw-Hill Science*. https://doi.org/10.1007/978-3-031-17922-8_9

Murphy, K. (2012). *Machine Learning A Probabilistic Perspective*. The MIT Press.

Olson, D. L., & Delen, D. (2008). Advanced Data Mining Techniques. In *Springer*. https://doi.org/10.1007/978-3-540-76917-0

Pereira, R. M., Costa, Y. M. G., & Silla, C. N. (2020). MLTL: A Multi-Label Approach For The Tomek Link Undersampling Algorithm. *Neurocomputing*, *383*, 95–105. https://doi.org/10.1016/j.neucom.2019.11.076

Rusnoto. (2016). Hubungan Riwayat Penyakit Tb Anggota Keluarga Dan Kondisi Rumah Dengan Terjadinya Penyakit TB Paru Pada Pasien Di Wilayah Kerja Puskesmas Ngemplak. *The 3rd University Research Colloquium 2016*, 348–353.

Sain, H., & Purnami, S. W. (2015). Combine Sampling Support Vector Machine for Imbalanced Data Classification. *Procedia Computer Science*, *72*, 59–66. https://doi.org/10.1016/j.procs.2015.12.105

Sastrawan, A. S., Studi, P., Informatika, T., Studi, P., Komputasi, I., Sains, F., Teknologi, I., Telekomunikasi, J., & Batu, T. B. (2010). Analisis Pengaruh Metode Combine Sampling Dalam Churn Prediction Untuk Perusahaan Telekomunikasi. *Seminar Nasional Informatika 2010 (SemnasIF 2010) UPN*, *1*(1), 14–22.

Siringoringo, R. (2018). Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor. *Jurnal ISD*, *3*(1), 44–49.

Sullivan, L. M. (2018). *Essentials of Biostatistics in Public Health*. Jones & Bartlett Learning.

Tomek, I. (1973). An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics, SMC*, *6*(6), 448–453.

Tyagi, S., & Mittal, S. (2020). Sampling Approaches For Imbalanced Data Classification Problem In Machine Learning. *Lecture Notes in Electrical Engineering*, *597*(7), 209–221. https://doi.org/10.1007/978-3-030-29407-6_17

Yap, B. W., Rani, K. A., Abd Rahman, H. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. *Proceedings Data and Information Engineering*, 13–22. https://doi.org/10.1007/978-981-4585-18-7_2

Zheng, W., & Jin, M. (2020). The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study. *Communications in Computer and Information Science*, *20*(1), 3–17. https://doi.org/10.1007/s42979-020-0074-0