# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Sebagai kompenen penting pada iklim bumi, hujan memiliki peran vital dalam kehidupan dan lingkungan (Pandey et al., 2024). Hal tersebut menunjukkan bahwa hujan memiliki pengaruh yang besar terhadap kehidupan dan aktivitas sehari-hari, maka dari itu hujan dengan frekuensi yang sangat tinggi dapat menyebabkan bencana alam seperti seperti banjir dan longsor. Hujan juga dapat merusak ekosistem dan menyebabkan masalah pada berbagai bidang, misalnya pada bidang pertanian, bidang infrastuktur, maupun bidang industri (Hassan et al., 2023). Dengan klasifikasi hujan yang akurat dapat membantu dalam pengelolaan sumber daya, pengambilan keputusan aktivitas sehari-hari dan juga dapat membantu masyarakat dalam mempersiapkan diri untuk meningkatkan kesiagaan dalam menghadapi masalah ataupun bencana alam yang dapat terjadi akibat hujan, sehingga dapat meminimalisir kerusakan properti dan mengurangi korban jiwa (Hassan et al., 2023; Latif et al., 2023). Oleh sebab itu, kemampuan untuk mengklasifikasikan hujan menjadi hal yang penting untuk diteliti lebih lanjut khususnya dalam hal keakuratan dan kinerjanya (Samsiah Sani et al., 2020).

Salah satu metode yang dapat digunakan untuk melakukan klasifikasi hujan adalah dengan *machine learning* (P. K. Das et al., 2024). *Machine learning* dapat membantu dalam mengidentifikasi pola hujan berdasarkan data historis dengan lebih akurat (Latif et al., 2023). Model klasifikasi yang akurat dipengaruhi oleh kualitas dan keseimbangan data yang digunakan saat proses pelatihan (Wiwaha et al., 2024). Namun dalam praktiknya terdapat masalah, yaitu masalah ketidakseimbangan kelas atau sering disebut dengan *imbalanced data* (Mohammed et al., 2020). *Imbalanced data* merupakan suatu keadaan dimana distribusi data yang ada tidak merata, hal ini ditandai dengan adanya suatu kelas yang memiliki jumlah sampel data yang lebih sedikit dibandingkan dengan kelas lainnya yang memiliki jumlah sampel data yang lebih banyak. Kelas dengan jumlah sampel yang lebih sedikit dikenal dengan istilah kelas minoritas sementara kelas dengan jumlah sampel yang lebih banyak dikenal dengan istilah kelas mayoritas. Distribusi data yang tidak merata ini mengakibatkan kurang terwakilinya kelas minoritas. Walaupun kurang terwakili, sampel data yang terdapat pada kelas minoritas memiliki tingkat urgensi yang tinggi (Asniar et al., 2022). Hal tersebut karena pada keadaan kelas yang tidak seimbang (*imbalanced*) mempengaruhi kinerja dari model sehingga mengakibatkan adanya kesalahan dalam klasifikasi (Fernández et al., 2018). Ketidakseimbangan kelas juga membuat model machine lerning bias terhadap kelas mayoritas, dengan kata lain lebih memihak pada kelas mayoritas (Asniar et al., 2022; Wiwaha et al., 2024). Sebagai contoh pada pemodelan klasifikasi hujan terdapat satu kelas yang memiliki hari tidak hujan yang termasuk dalam kelas mayoritas, sedangkan kelas lainnya yaitu hari hujan termasuk dalam kelas minoritas. Hal ini berarti jumlah data pada kelas yang memiliki hari tidak

hujan lebih banyak daripada kelas yang memiliki hari hujan. sehingga model pelatihan akan lebih banyak belajar dari kelas mayoritas dan kurang mendapatkan pembelajaran dari kelas minoritas. Dengan kata lain, model dapat menghasilkan kinerja yang baik dalam melakukan klasifikasi pada hari tidak hujan, namun kinerjanya dalam melakukan klasifikasi pada hari hujan kurang optimal. Karena kurang terwakilinya suatu kelas maka dapat memberikan dampak yang cukup signifikan selama proses pembelajaran. Kelas minoritas dianggap sebagai situasi yang jarang terjadi dan tidak umum, sehingga sulit untuk melakukan klasifikasi (Shamsudin et al., 2020). Prinsip pokok pada proses pembelajaran algoritma *machine learning* yaitu pembelajaran dilakukan dari data yang seimbang (de Vargas et al., 2023). Apabila keadaan data seimbang, maka setiap kelas akan memiliki porsi data yang sama dan tidak bias terhadap salah satu kelas sehingga model dapat memberikan kinerja yang baik dalam klasifikasi kelas mayoritas maupun kelas minoritas. Oleh sebab itu, masalah ketidakseimbangan kelas perlu ditangani guna meningkatkan kinerja model dalam melakukan klasifikasi kelas minoritas namun tetap menjaga kinerja model dalam melakukan klasifikasi kelas mayoritas (Asniar et al., 2022).

Terdapat dua tingkatan dalam menangani ketidakseimbangan kelas, yaitu tingkat data (*data level*) dan tingkat algoritma (*algorithm level*). Pada tingkat data, sebelum proses pembelajaran, tepatnya pada tahap preprocessing dilakukan teknik *undersampling* untuk mengurangi sampel mayoritas, *oversampling* untuk menambahkan sampel minoritas, atau kombinasi dari *oversampling* dan *undersampling*. Sebagai contoh oversampling, terdapat R*andom Oversampling, Adaptive Synthetic Sampling (ADASYN)*, dan *Synthetic Minority Oversampling Technique (SMOTE)*. Kemudian conrtoh *undersampling* Random Undesampling, Neighborhood Cleaning Rule (NCL), dan Tomek Links (Fernández et al., 2018). Sementara itu untuk tingkat algoritma, proses pembelajaran melalui algoritma yang telah ditingkatkan kinerjanya, seperti *cost-sensitive* dan metode *ensemble* (de Vargas et al., 2023).

Pada penelitian (Oswal, 2019) Oswal melakukan eksperimen dengan membandingkan kinerja antara metode *oversampling* dan metode *undersampling* dalam mengatasi ketidakseimbangan kelas. Penelitian tersebut menunjukkan adanya *overfitting* pada kinerja model saat menggunakan metode *oversampling*. *Overfitting* merupakan suatu keadaan dimana model memiliki kinerja yang sangat baik saat pelatihan namun memiliki kinerja yang kurang baik atau buruk saat pengujian. Selain itu, pada penelitian ini melakukan eksperimen dengan menggunakan beberapa model klasifikasi dan menunjukkan bahwa kinerja model dapat berbeda tergantung pada jenis metode yang digunakan dalam menyeimbangkan kelas, terdapat model yang memiliki kinerja yang baik menggunakan metode *oversampling* namun buruk saat menggunakan metode *undersampling* begitu pula sebaliknya. Selanjutnya dalam penelitian (Mohammed et al., 2020) yang juga melakukan perbandingan antara kinerja metode *oversampling*

dan *undesampling* menampilkan kinerja model yang baik pada penggunaan metode *oversampling* tapi menunjukkan adanya indikasi *overfitting* dan pada penggunaan metode *undersampling* menampilkan kinerja model yang tidak lebih baik bila dibandingkan dengan kinerja model dengan menggunakan metode *oversampling*.

Hal ini menunjukkan adanya pengaruh dari penggunaan teknik *resampling* yakni *oversampling* dan *undersampling* pada kinerja dari suatu model klasifikasi. Baik *oversampling* maupun *undersampling* memiliki potensi untuk meningkatkan kinerja model dengan menyeimbangkan distribusi data sebelum masuk dalam proses pelatihan. Namun terdapat pula keterbatasan dari penggunaan *oversampling* dan *undersampling*. Hal ini karena dengan hanya menggunakan *oversampling* sebagai satu-satunya pendekatan untuk mengatasi ketidakseimbangan kelas dapat menyebabkan *overfitting*. Selain itu, oversampling juga membuat wilayah kelas minoritas menjadi lebih luas sehingga dapat masuk ke wilayah kelas mayoritas dan berpeluang meningkatkan *overlap*. Sedangkan jika hanya menggunakan undersampling dengan mengurangi sampel pada kelas mayoritas dapat meningkatkan kemungkinan kehilangan data atau informasi penting pada dataset yang memiliki pengaruh signifikan terhadap proses pembelajaran model. Proses penghapusan pada undersampling dapat menyebabkan kelas mayoritas kurang merepresentasikan variasi dan pola aslinya. Situasi ini dapat berdampak pada kinerja model. (Asniar et al., 2022; Fernández et al., 2018; Pristyanto et al., 2018; Qian et al., 2014; Shamsudin et al., 2020; de Vargas et al., 2023).

Berdasarkan hal tersebut penelitian ini melakukan pendekatan SMOTE yang berperan sebagai metode *oversampling* dan Tomek links sebagai metode pembersihan (*cleaning method*). SMOTE menyeimbangkan distribusi kelas dengan cara menghasilkan sampel sintetis pada kelas minoritas berdasarkan tetangga terdekatnya, namun hal tersebut memiliki potensi untuk menghasilkan sampel sintetis pada wilayah yang tumpang tindih dengan kelas mayoritas (Fachrie et al., 2025). Disisi lain, Tomek links yang umumnya digunakan sebagai metode undersampling akan bertidak sebagai metode pembersihan pada penelitian ini. Tomek links sebagai metode undersampling bekerja dengan mendeteksi pasangan sampel dari dua kelas berbeda yang menjadi tetangga terdekat dan menghapus sampel dari kelas mayoritas. Selain digunakan sebagai metode undersampling, Tomek links tidak hanya untuk menyeimbangkan distribusi kelas tapi dapat pula digunakan sebagai metode pembersihan. Apabila tomek links bertindak sebagai metode pembersihan, maka akan menghapus sampel dari kelas mayoritas maupun kelas minoritas bukan hanya dari kelas minoritas saja. Pendekatan ini bertujuan untuk menghilangkan sampel yang bersifat ambigu dengan begitu dapat meningkatkan kualitas data dengan mengurangi overlap dan memperjelas batas keputusan. Tomek links berhubungan dengan tetangga terdekat antar sampel yang kemampuannya tergantung dari distribusi lokal data (Batista et al., 2004).

Baik SMOTE dan Tomek links melakukan modifikasi data dengan prosedur yang berbeda, sehingga perbedaan urutan penggunaan SMOTE dan Tomek links

sebagai metode pembersih berpotensi membuat karakteristik data yang berbeda (Carvalho et al., 2025; Jiang et al., 2025). Oleh karena itu, penelitian ini berfokus pada analisis integrasi SMOTE dan Tomek links dengan memanfaatkan Tomek links sebagai metode pembersih, penelitian ini juga melakukan analisis pengaruh urutan implementasi Tomek links sebelum dan sesudah SMOTE terhadap kinerja machine learning pada studi kasus klasifikasi hujan dengan keadaan kelas yang tidak seimbang.

### 1.2 Rumusan Masalah

Berikut merupakan rumusan masalah pada penelitian ini yang disusun berdasarkan latar belakang yang telah dipaparkan sebelumnya:

1. Bagaimana pengaruh SMOTE sebagai teknik *resampling* maupun Tomek links sebagai metode pembersih terhadap kinerja klasifikasi hujan?
2. Bagaimana pengaruh urutan penerapan SMOTE dan Tomek links sebagai metode pembersih terhadap kinerja model klasifikasi hujan?

### 1.3 Tujuan Penelitian

Berikut ini merupakan tujuan penelitian yang akan dicapai:

1. Menganalisis pengaruh SMOTE sebagai teknik *resampling* maupun Tomek links sebagai metode pembersih terhadap kinerja klasifikasi hujan.
2. Menganalisis pengaruh urutan penerapan SMOTE dan Tomek links sebagai metode pembersih terhadap kinerja model klasifikasi hujan.

### 1.4 Manfaat Penelitian

Berikut ini merupakan manfaat dari penelitian ini:

1. Menambah wawasan mengenai pentingnya penanganan ketidakseimbangan kelas dan pembersihan data serta memberi gambaran efektivitas penggunaan kedua metode tersebut.
2. Sebagai literatur bagi peneliti berikutnya yang ingin melakukan penelitian terkait ketidakseimbangan kelas.
3. Menambah referensi terkait penggunaan kombinasi teknik *resampling* dan pembersihan data dalam meningkatkan kinerja model.

### 1.5 Ruang Lingkup

1. Penelitian ini berfokus mengatasi ketidakseimbangan kelas dengan berdasarkan pendekatan *data level*.
2. Data yang digunakan pada penelitian ini adalah data curah hujan yang didasarkan pada *Bureau of Meteorology Australia* yang diperoleh dari *Kaggle*.

# BAB II
# TINJAUAN PUSTAKA

## 2.1 Tinjauan Pustaka

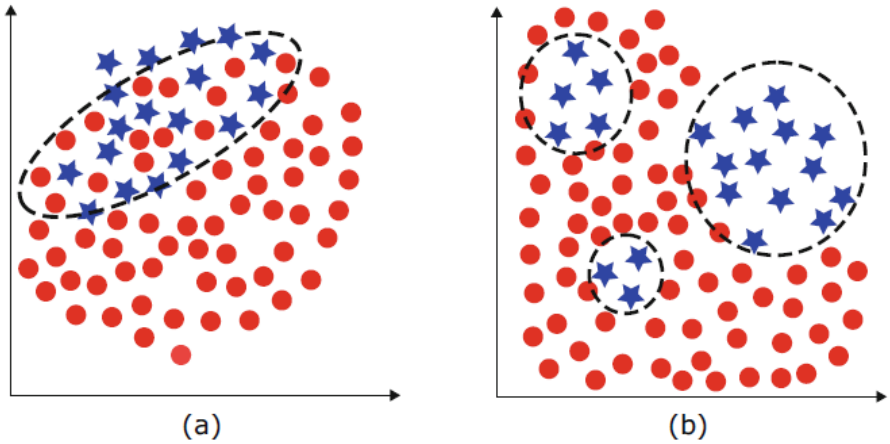
### 2.1.1 Hujan

Hujan adalah bentuk presipitasi yang terjadi secara bertahap yang diawali dengan proses penguapan air di permukaan bumi diiringi oleh pergerakan uap air yang menyebabkan terjadinya kondensasi dan pembentukan awan. Partikel-partikel air yang terdapat di dalam awan kemudian melalui proses mikrofisika, yakni penggabungan tetesan air atau pembentukan kristal es. Saat partikel air tersebut telah memperoleh massa dan ukuran yang cukup besar, maka gaya gravitasi menyebabkan partikel-partikel air jatuh ke permukaan bumi dalam bentuk hujan Hujan memiliki kontribusi terhadap kestabilan ekologi. Hal tersebut menunjukkan bahwa hujan sangat berpengaruh pada kehidupan makhluk hidup setiap harinya (Hassan et al., 2023). Kelembapan udara, suhu, dan kecepatan angin merupakan faktor-faktor yang saling berinteraksi satu sama lain yang mempengaruhi terjadinya hujan. Kelembapan udara merujuk pada ketersediaan uap air di atmosfer sebagai komponen utama dari pembentukan awan saat proses kondensasi. Suhu udara bertugas dalam mengelola kecepatan peguapan. Udara dengan suhu yang lebih tinggi atau hangat cenderung bergerak naik kelapisan atmosfer yang lebih tinggi dan memicu terbentuknya awan hujan. Kemuadian, kecepatan angin mempengaruhi penyebaran uap air sehingga dan pergerakan massa udara, hal inilah yang menjadi penentu lokasi dan intensitas hujan (Usman Saeed Khan et al., 2024).

Pada sektor pertanian, hujan menjadi salah satu faktor yang paling berpengaruh karena secara langsung berperan dalam ketersediaan air bagi tanaman. Selain itu, hujan juga dapat menjadi penentu waktu tanam atau penjadwalan panen dan pemilihan jenis taman serta manajemen irigasinya. Kurangnya informasi mengenai perkiraan hujan akan berdampak besar bagi sektor ini. Tingginya frekuensi hujan dapat mengakibatkan kerusakan pada tanaman dan erosi tanah jika tidak ada tindakan persiapan awal yang dilakukan. Selanjutnya hujan juga memiliki dampak yang besar bagi sistem drainase, hal ini karena hujan dengan frekuensi yang tinggi dapat melebihi batas atau kapasitas dari saluran drainase, sehingga menimbulkan genangan air atau bahkan banjir dan bencana lainnya (Pandey et al., 2024). Berdasarkan laporan WMO, bencana yang disebabkan oleh hujan seperti badai dan banjir termasuk dalam dua teratas dalam kategori sepuluh bahaya cuaca dan iklim dengan dampak terbesar. Adapun dampak yang dimaksud ialah banyaknya korban jiwa dan tingginya kerugian yang ditimbulkan (Wang et al., 2022).

### 2.1.2 Ketidakseimbangan Kelas (*Imbalanced Class*)

Ketidakseimbangan kelas atau *imbalanced class* merujuk pada kondisi ketika data yang terdapat pada suatu kelas tidak terdistribusi secara merata. Ketidakseimbangan kelas terjadi saat jumlah data yang mewakili suatu kelas lebih rendah dibandingkan dengan kelas yang lainnya. Maka dari itu, terdapat kelas yang kurang terwakili dalam dataset. Kelas yang kurang terwakili disebut dengan kelas minoritas sedangkan kelas lainnya yang memiliki jumlah data yang lebih banyak disebut dengan kelas mayoritas (Fernández et al., 2018).




**Gambar 1** (a) *Overlapping* (b) *Small disjunction* (Fernández et al., 2018)

Beberapa permasalahan berikut sering muncul saat terjadi ketidakseimbangan data:

1. *Overlapping*. Gambar 1 (a) mengilustrasikan ketika terjadi *overlapping* atau tumpang tindih dalam dataset, maka proses induksi aturan diskriminatif menjadi sukar. Hal tersebut menunjukkan bahwa model akan kesulitan dalam menemukan aturan yang membedakan antara kelas-kelas yang tumpang tindih tersebut sehingga menyebabkan terjadinya kesalahan dalam klasifikasi pada kelas minoritas (Fernández et al., 2018).
2. *Small disjunction*. Gambar 1 (b) mengilustrasikan *small disjunction* yang terjadi ketika suatu kelas minoritas berbentuk subkelas atau subkonsep. Apabila terjadi *small disjunction* maka dengan begitu kompleksitas dari suatu dataset akan bertambah karena jumlah data pada setiap subkelas tidak seimbang, sehingga dapat menyulitkan model dalam mengidentifikasi dan memahami pola pada kelas minoritas (Fernández et al., 2018)

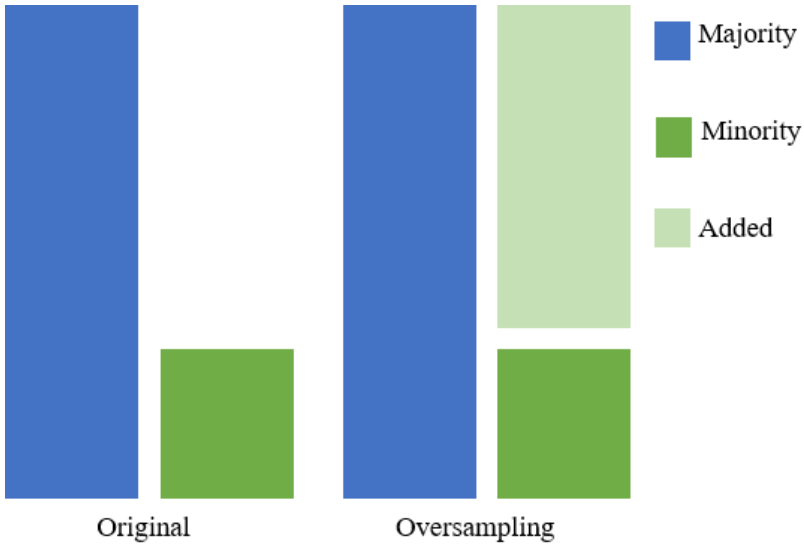
Permasalahan-permasalahan tersebut dapat memberikan dampak negatif pada model dalam mengidentifikasi kelas minoritas (Fahrudin et al., 2019). Kelas minoritas yang kurang terwakili cenderung diabaikan dan dianggap sebagai *noise.*

Akibatnya model pembelajaran bias terhadap kelas mayoritas. Dengan demikian, hal tersebut dapat menurunkan tingkat akurasi dan kehandalan dari model yang dilatih (de Vargas et al., 2023).

Ketidakseimbangan kelas dapat diatasi dengan pendekatan secara eksternal (*data level*) maupun secara internal (*algorithm level*) (Gosain & Sardana, 2017). Pendekatan secara eksternal atau yang dikenal juga dengan istilah *data level* akan memodifikasi data agar menghasilkan distribusi kelas yang lebih seimbang (Fernández et al., 2018). Pendekatan data level dilakukan pada tahap *preprocessing* (Wang et al., 2021). *Data level* memiliki tujuan untuk menyeimbangkan distribusi kelas dengan cara menambah sampel data kelas minoritas atau mengurangi sampel data dari kelas mayoritas dengan teknik *resampling*. Teknik *resampling* dapat dibagi menjadi dua, yaitu *oversampling* dan *undersampling* (Fernández et al., 2018). Sedangkan pada pendekatan *algorithm level* memodifikasi algoritma pembelajaran untuk dapat mengatasi ketidakseimbangan. Adapun contoh dari *algorithm level* yaitu *cost-sensitive* dan *ensemble method* (Fernández et al., 2018; Gosain & Sardana, 2017; de Vargas et al., 2023).

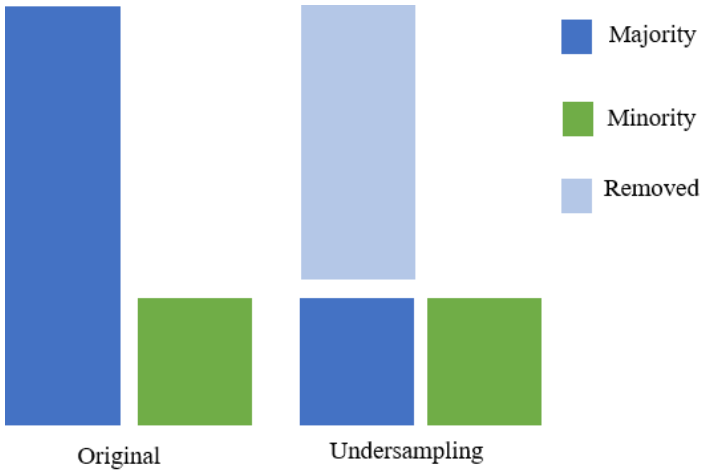
### 2.1.3 *Oversampling* dan *Undersamling*

Keadaan kelas yang tidak seimbang dapat diatasi dengan menggunakan teknik *resampling* atau yang lebih dikenal dengan istilah *oversampling* dan *undersampling*. Pada teknik *oversampling*, data sintesis dihasilkan dengan cara menambah data pada kelas yang kurang terwakili sehingga dapat mencapai keseimbangan atau dapat mendekati jumlah data pada kelas mayoritas (Fernández et al., 2018). Gambar 2 menunjukkan ilustrasi dari *oversampling*).




**Gambar 2** Ilustrasi *oversampling*

Jika *oversampling* mengatasi ketidakseimbangan kelas dengan menambahkan data sintesis pada kelas minoritas, teknik *undersampling* melakukan sebaliknya. *Undersampling* mengatasi ketidakseimbangan kelas dengan mereduksi atau mengeliminasi data yang terdapat pada kelas mayoritas yang tujuannya untuk menyeimbangkan jumlah sampel data pada tiap kelas (Fernández et al., 2018). Gambar 3 menunjukkan ilustrasi dari *undersampling*. Tetapi pada penelitian ini hanya berfokus pada *oversampling* dengan penambahan data sistesis menggunakan metode SMOTE serta Tomek links yang akan bertindak sebagai metode pembersih.




**Gambar 3** Ilustrasi *undersampling*

### 2.1.4 *Synthetic Minority Oversampling Technique* (SMOTE)

*Synthetic Minority Oversampling Technique* (SMOTE) merupakan salah satu pendekatan dari metode *oversampling*. SMOTE berkerja dengan cara memproduksi data sintetis dengan memanfaatkan karakteristik dari data yang telah ada sebelumnya, jadi SMOTE tidak memproduksi data baru dengan cara langsung melakukan duplikasi pada sampel kelas minoritas. Data sintetis ini dihasilkan melalui proses interpolasi di ruang fitur, yaitu dengan mengambil satu sampel minoritas yang dijadikan sebagai landasan, kemudian membentuk segmen garis ke beberapa sampel tetangga terdekatnya yang berasal dari kelas minoritas juga dan menghasilkan titik-titik baru di sepanjang garis tersebut. Proses tersebut bekerja berdasarkan nilai fitur bukan sekadar menambahkan baris data yang sama oleh karena itu, SMOTE dikenal juga berfokus pada *feature space* daripada *data space*. Pendekatan ini memiliki tujuan untuk menambahkan distribusi kelas minoritas dengan lebih realistis akibatnya model pun dapat mempelajari pola suatu data secara universal dan mengurangi bias terhadap kelas mayoritas (Fernández et al., 2018).

Berikut ini merupakan tahapan SMOTE (Chawla et al., 2002):

Input: T, N, k.
Tabel 1 menunjukkan penjelasan dari inputan yang digunakan pada SMOTE.

Output: (N/100) * T sampel sintetis kelas minoritas.

1. Jika N<100%:
   - Acak urutan sampel kelas minoritas.
   - Memilih hanya sebagian dari T sampel sesuai dengan nilai N%.
   - Menetapkan N=100 agar proses selanjutnya menjadi terstandarisasi.
2. Melakukan konversi N menjadi pengali bilangan bulat: Sebagai contoh, jika N=200, maka N=2, yang berarti dua sampel sintetis akan dihasilkan untuk setiap satu sampel asli.
3. Melakukan perulangan untuk setiap sampel kelas minoritas (sebanyak T):
   - Menentukan k tetangga terdekat untuk setiap sampel. Untuk k tetangga terdekat, dapat ditentukan dengan menggunakan persamaan mencari jarak terdekat yaitu dengan jarak Euclidean:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \quad (1)$$

Keterangan:
$x_i$ = Sampel ke-i dari kelas minoritas
$x_j$ = Sampel ke-j dari kelas minoritas
$M$ = Jumlah fitur atau atribut dataset
$m$ = Indeks fitur ke-m (m = 1, 2, … M)
$x_{im}$ = Nilai fitur ke-m pada sampel $x_i$
$x_{jm}$ = Nilai fitur ke-m pada sampel $x_j$

   - Untuk setiap N sampel sintetis yang akan dibuat:
     - Pilih secara acak satu dari k tetangga terdekat tersebut.
     - Hitung selisih nilai fitur antara sampel acuan dan tetangganya.
     - Kalikan selisih tersebut dengan bilangan acak antara 0 dan 1 untuk melakukan interpolasi, dengan persamaan sebagai berikut:

$$x_{sintetis} = x_i + \lambda(x_i - x_j) \quad (2)$$

     - Kemudian tambahkan hasil interpolasi tersebut ke nilai atribut sampel acuan guna membentuk nilai atribut sintetis.
     - Menambah indeks sampel sintetis (newindex++) dan mengurangi nilai N.

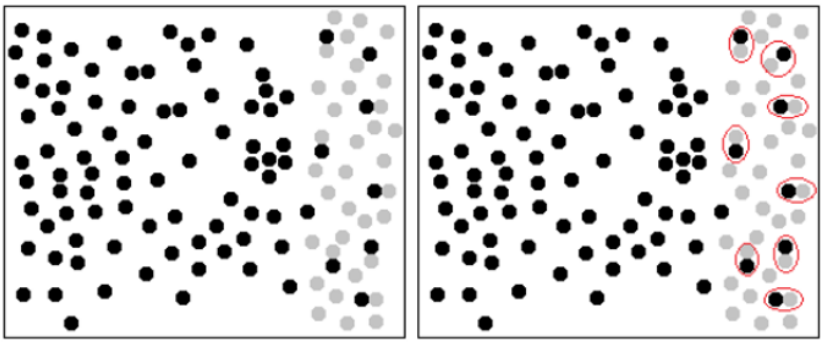
**Tabel 1** Inputan pada SMOTE

| Input | Keterangan |
|---|---|
| T | Jumlah sampel kelas minoritas |
| N | Jumlah sampel sintetis yang dihasilkan |
| k | Jumlah tetangga terdekat |

### 2.1.5 Tomek Links

Tomek Links merupakan salah satu metode yang digunakan pada permasalahan klasifikasi ketidakseimbangan kelas untuk mengidentifikasi pasangan sampel yang berada sangat dekat tetapi berasal dari kelas yang berbeda. Kemudian jarak antara kedua sampel tersebut dihitung menggunakan Euclidean Distance. Suatu pasangan data $(E_i, E_j)$ disebut sebagai Tomek Link apabila kedua sampel tersebut saling menjadi tetangga terdekat (*nearest neighbor*) dan tidak terdapat sampel lain $E_l$ yang memiliki jarak lebih dekat terhadap salah satu dari kedua sampel tersebut. Kondisi tersebut dapat dinyatakan sebagai: $d(E_i, E_l) < d(E_i, E_j)$ atau $d(E_j, E_l) < d(E_i, E_j)$ untuk setiap sampel $E_l$. Artinya, tidak ada sampel lain yang lebih dekat terhadap $E_i$ maupun $E_j$ dibandingkan jarak antara kedua sampel tersebut. Dengan kata lain, $E_i$ dan $E_j$ saling menjadi *nearest neighbor* satu sama lain meskipun berasal dari kelas yang berbeda. Dengan demikian, pasangan Tomek Link menunjukkan bahwa kedua sampel dari kelas berbeda berada pada area yang saling berdekatan sehingga berpotensi menyebabkan ambiguitas klasifikasi (Fernández et al., 2018).

Gambar 4 (a) menunjukkan keadaan original dataset atau data yang asli. Sedangkan Gambar 4 (b) menunjukkan *tomek Links* pada data. Jika terdapat dua sampel yang membentuk *tomek links*, maka salah satu dari kedua sampel tersebut adalah *noise* atau kedua sampel tersebut adalah *borderline*, yang mana sampel tersebut berada di dekat batas keputusan (*decision border*) yang mengakibatkan sulitnya untuk menentukan di kelas mana sampel tersebut seharusnya berada. *Tomek links* sering kali dimanfaatkan sebagai metode undersampling dengan mendeteksi pasangan data dari kelas yang berbeda yang merupakan tetangga terdekat lalu menghapus sampel data yang berasal dari kelas mayoritas. Hal tersebut bertujuan agar dapat mengurangi tumpang tindih antar kelas. Namun, *tomek links* juga dapat bertindak sebagai metode pembersih apabila menghapus kedua sampel yang membentuk pasangan *tomek links*. Dengan menghapus salah satu atau kedua sampel pada *Tomek links* maka dapat memperjelas batas antar kelas (Pereira et al., 2020).



(a) (b)

**Gambar 4** (a) Original Dataset (b) Tomek Links (Pereira et al., 2020)

### 2.1.6 *Confusion Matrix*

*Confusion matrix* adalah suatu cara untuk melakukan evaluasi terhadap kinerja dari model klasifikasi dengan menampilkan perbandingan antara kelas aktual dan kelas prediksi. Tabel 2 menunjukkan suatu *confusion matrix* pada klasifikasi biner.

**Tabel 2** *Confusion matrix*

| | Prediksi Positif | Prediksi Negatif |
|---|---|---|
| Kelas Positif | TP | FN |
| Kelas Negatif | FP | TN |

Berdasarkan informasi yang terdapat pada *confusion matrix*, dapat diturunkan berbagai metrik evaluasi untuk mengevaluasi kinerja model (Batista et al., 2004). Dibawah ini merupakan persamaan untuk melakukan proses pengujian dengan *confusion matrix*:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - Score = \frac{2(Precision * Recall)}{Precision + Recall} \quad (6)$$

Keterangan:

True Positive (TP) : Seluruh data yang termasuk kelas positif dan diprediksi benar sebagai kelas positif.

True Negative (TN) : Seluruh data yang termasuk kelas negatif dan diprediksi benar sebagai kelas negatif.

False Positive (FP) : Seluruh data yang termasuk kelas negatif tapi diprediksi sebagai kelas positif.

False Negative (FN) : Seluruh data yang termasuk kelas positif tapi diprediksi sebagai kelas negatif.

### 2.1.7 Kurva ROC-AUC

ROC atau *Receiver Operating Charactheristic* ialah kurva atau grafik yang digunakan untuk mengevaluasi kinerja suatu model dengan cara mengidentifikasi hubungan antara *true positive rate* dan *false positive rate* berdasarkan nilai dari ambang batas (*threshold*). Setiap variasi pada nilai ambang batas akan menghasilkan sejumlah titik pada ruang ROC, titik-titik tersebut dihubungkan menjadi kurva ROC yang memvisualisasikan hubungan antara kemampuan model mendeteksi kelas positif dan tingkat kesalahan prediksi positif yang keliru. Kurva ROC yang mendekati sudut kiri atas dari grafik memiliki makna bahwa model memiliki kinerja yang sangat baik dalam hal membedakan kelas positif dan juga kelas negatif. Sementara itu, AUC atau *Area Under the Curve* ialah luas wilayah yang terdapat di bawah kurva ROC, AUC merepresentasikan kurva ROC dalam nilai kuantitatif. Nilai tersebut menggambarkan kemampuan model untuk membedakan antara kelas positif dan kelas negatif. Secara statistik, nilai AUC ekuivalen dengan uji peringkat Wilcoxon. Oleh sebab itu, AUC banyak digunakan sebagai metrik evaluasi guna mengukur kinerja model klasifikasi (Batista et al., 2004).

## 2.2 Metode Penyelesaian Masalah

### 2.2.1 *State of The Art* Penelitian

**Tabel 3** *State of The Art*

| No. | Judul karya ilmiah, nama, tahun terbit dan penerbit | Objek dan permasalahan | Metode Penyelesaian | Kinerja | Korelasi |
|---|---|---|---|---|---|
| 1. | Judul: Analisis Penerapan Tomek Links | Objek: Ketidakseimbangan kelas | SMOTE dan *Tomek links* sebagai | | |

| No. | Judul karya ilmiah, nama, tahun terbit dan penerbit | Objek dan permasalahan | Metode Penyelesaian | Kinerja | Korelasi |
|---|---|---|---|---|---|
| | sebagai Metode Pembersih dan Smote untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Hujan<br><br>Penulis:<br>Nur Awalia. S. Z. Abidin | pada dataset hujan<br><br>Permasalahan: Terdapat ketidakseimbangan kelas pada kasus klasifikasi hujan yang dapat menyebabkan model bias terhadap kelas mayoritas sehingga berdampak pada kinerja model. | metode pembersih | | |
| 2. | Judul:<br>*Predicting Rainfall using Machine Learning Techniques*<br><br>Penulis:<br>Nikhil Oswal<br><br>Tahun:<br>2019<br><br>Penerbit:<br>arXiv | Objek:<br>Dataset hujan di Australia<br><br>Permasalahan: Penelitian ini menggunakan *machine learning* dengan mengeksplorasi ketidakseimbangan kelas untuk melakukan prediksi hujan, namun maih terdapat kekurangan yakni adanya indikasi *overfitting* | *Undersampling (Random Undersampling), Oversampling (SMOTE), Logistic Regression, Decision Tree, KNN, Rule Based, Ensemble* | • Pada undersampling dengan menggunakan Logistic Regression mencapai akurasi terbaik sebesar 84%<br>• Pada oversampling dengan menggunakan KNN mencapai akurasi terbaik sebesar 90% namun menunjukkan adanya indikasi *overfitting* | > |

| No. | Judul karya ilmiah, nama, tahun terbit dan penerbit | Objek dan permasalahan | Metode Penyelesaian | Kinerja | Korelasi |
|---|---|---|---|---|---|
| 3. | Judul: *Prediction of Rainfall Using Machine Learning*<br><br>Penulis: Deepika Mahajan, Sandeep Sharma<br><br>Tahun: 2022<br><br>Penerbit: IEEE | Objek: Dataset hujan di Australia<br><br>Permasalahan: Prediksi terjadinya hujan adalah tugas yang menantang karena pola hujan yang tidak konsisten dan adanya iklim yang bervariasi. | *Naïve Bayes*, SVM, *Random Forest*, KNN | Akurasi tertinggi dicapai dengan menggunakan Random Forest, yakni sebesar 85% | > |
| 4. | Judul: *Machine Learning-Based Rainfall Prediction: Unveiling Insight and Forecasting for Improved Preparedness*<br><br>Penulis: MD. Mehedi Hassan, Mohammad Abu Tareq Rony, MD. Asif Rakib Khan, Farhana Yasmin, Anindya Nag, Tazria Helal Zarin, Anupam Kumar Bairagi, Samah Alshathri, Walid El-Shafai | Objek: Dataset hujan di Australia<br><br>Permasalahan: Prediksi terjadinya hujan yang akurat dengan menggunakan *machine learning* masih menjadi tantangan karena sifat hujan yang kompleks. | *Naïve Bayes, Decision Tree*, SVM, *Random Forest, Logistic Regression*, ANN, LSTM | Akurasi terbaik diperoleh dengan menggunakan ANN, yakni sebesar 90% | > |

| No. | Judul karya ilmiah, nama, tahun terbit dan penerbit | Objek dan permasalahan | Metode Penyelesaian | Kinerja | Korelasi |
|---|---|---|---|---|---|
| | Tahun:<br>2023<br><br>Penerbit:<br>IEEE | | | | |
| 5. | Judul:<br>*Rainfall Prediction Enhancement Using SMOTE and Machine Learning Algorithms*<br><br>Penulis:<br>R Hemanth Kumar, Vishnu Prasad B, Himanshu Yadav, Manju Venugopalan<br><br>Tahun:<br>2024<br><br>Penerbit:<br>IEEE | Objek:<br>Ketidakseimbangan kelas pada dataset hujan<br><br>Permasalahan:<br>Prediksi hujan yang akurat diperlukan pada berbagai bidang. Prediksi hujan dengan menggunakan machine learning merupakan tantangan tersendiri karena terdapat permasalahan ketidakseimbangan kelas yang dapat mempengaruhi ketepatan model prediksi | SMOTE, Random Forest, Decision Tree, Perceptron, MLP KNN, Bagging, Adaboost, Gradient Boosting | Performa terbaik ditunjukkan oleh Random Forest dengan akurasi sebesar 89% | > |
| 6. | Judul:<br>*Leveraging Meteorological Data and Machine Learning for Improved Rainfall Forecasting in Australia*<br><br>Penulis: | Objek:<br>Dataset hujan di Australia<br><br>Permasalahan:<br>Hujan merupakan komponen penting pada sistem cuaca di Australia karena berpengaruh | Random Forest, XGBoost, Logistic Regression, KNN, ANN, Naïve Bayes | Kinerja model terbaik didapatkan dengan menggunakan Random Forest dan XGBoost, sebesar 86% | > |

| No. | Judul karya ilmiah, nama, tahun terbit dan penerbit | Objek dan permasalahan | Metode Penyelesaian | Kinerja | Korelasi |
|---|---|---|---|---|---|
| | Akash Das, K. M. Asieb Hasan<br><br>Tahun: 2024<br><br>Penerbit: IEEE | pada kondisi lingkungan, pertanian dan perekonomian. Namun prediksi hujan menggunakan *Machine Learning* masih menjadi tantangan. | | | |
| 7. | Judul: *Rainfall Prediction Using Logistic Regression and Random Forest Algorithm*<br><br>Penulis: Ritesh Pandey, Maneesh Upadhya, Manvendra Singh<br><br>Tahun: 2024<br><br>Penerbit: IEEE | Objek: Dataset hujan di Australia<br><br>Permasalahan: Prediksi hujan yang akurat diperlukan karena hujan memiliki dampak yang besar bagi kehidupan. Namun, geografi Australia yang beragam menjadikannya tantangan tersendiri dalam memprediksi hujan, terlebih lagi terdapat ketidakseimbangan kelas yang berpotensi membuat model bias | SMOTE, *Logistic Regression*, *Random Forest* | • Akurasi menggunakan Logistic Regression sebesar 77%<br>• Akurasi menggunakan Random Forest sebesar 84% | > |
| 8. | Judul: *Credit Card Fraud Detection Using Machine Learning Techniques: Dealing with* | Objek: Ketidakseimbangan kelas pada data penipuan kartu kredit | SMOTE, ROS+RUS, Random Forest, XGBoost, Voting, MLP | • Berdasarkan metrik evaluasi *precision*, kombinasi Random Forest | < |

| No. | Judul karya ilmiah, nama, tahun terbit dan penerbit | Objek dan permasalahan | Metode Penyelesaian | Kinerja | Korelasi |
|---|---|---|---|---|---|
| | *Imbalanced Data using Over-Sampling and Under-Sampling Methods*<br><br>Penulis:<br>Adil Hussain, Vineet Dhanawat, Ayesha Aslam, Noman Iqbal, Sajib Tripura<br><br>Tahun:<br>2024<br><br>Penerbit:<br>IEEE | Permasalahan:<br>Terdapat ketidakseimbangan kelas pada data penipuan kartu kredit. Total transaksi yang benar jauh lebih banyak dibandingkan transaksi penipuan sehingga sulit untuk mendeteksi kasus penipuan tersebut. | | dengan ROS+RUS memperoleh performa terbaik, dengan nilai 94%<br>• Berdasarkan metrik evaluasi *recall,* kombinasi XGBoost dengan SMOTE menunjukkan performa terbaik dengan nilai 88% | |
| 9. | Judul:<br>*Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques*<br><br>Penulis:<br>Al-Amain, Mirza Nadim Saad, Esrat Jahan, Rajon Bardhan<br><br>Tahun:<br>2021<br><br>Penerbit:<br>IEEE | Objek:<br>Ketidakseimbangan kelas pada data penyakit jantung<br><br>Permasalahan:<br>Terdapat ketidakseimbangan kelas pada data penyakit jantung yang mempengaruhi kinerja model dalam memprediksi penyakit | SMOTE, Logistic Regression, Naïve Bayes, Decision Tree, AdaBoost, SVM, ETC, GNB, SGD | Akurasi terbaik didapatkan dengan menggunakan ETC dengan nilai sebesar 92% | < |
| 10. | Judul:<br>*Enhanced Parkinson's Disease* | Objek:<br>Ketidakseimbangan kelas pada data | SMOTE, KNN, SVM, Logistic Regression, | Akurasi terbaik didapatkan dengan | < |

| No. | Judul karya ilmiah, nama, tahun terbit dan penerbit | Objek dan permasalahan | Metode Penyelesaian | Kinerja | Korelasi |
|---|---|---|---|---|---|
| | *Prediction Using Feature Selection, SMOTE, anda Machine Learning with Deep Learning Models*<br><br>Penulis:<br>Al-Amain, Mirza Nadim Saad, Esrat Jahan, Rajon Bardhan<br><br>Tahun:<br>2024<br><br>Penerbit:<br>IEEE | penyakit parkinson<br><br>Permasalahan: Terdapat ketidakseimbangan kelas pada data penyakit parkinson yang mempengaruhi kinerja model dalam memprediksi penyakit | Naïve Bayes, Decision Tree, AdaBoost, LSTM, PLSTM, MLP | menggunakan aloritma KNN dengan nilai sebesar 97% | |

Keterangan:

1. Analisis Penerapan Tomek Links sebagai Metode Pembersih dan Smote untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Hujan – Penelitian ini bertujuan mengintegrasikan tomek links yang berperan sebagai metode pembersih dan SMOTE guna mengatasi ketidakseimbangan kelas pada dataset serta menganalisis urutan penerapan kedua metode tersebut terhadap kinerja model.
2. *Predicting Rainfall using Machine Learning Techniques* – Penelitian ini bertujuan untuk melakukan prediksi curah hujan dengan mengeksplorasi tahap preprocessing dengan melakukan eksperimen teknik resampling, yaitu oversampling dan undersampling, serta mengeksplorasi model atau algoritma yang digunakan (Oswal, 2019). Celah dari penelitian ini yaitu penggunaan *oversampling* ataupun *undersampling* secara individu memprengaruhi kinerja dari model.
3. *Prediction of Rainfall Using Machine Learning* – Penelitian ini juga menggunakan machine learning untuk melakukan prediksi terjadinya hujan, namun pada penelitian ini tidak melakukan modifikasi data untuk menyeimbangkan kelas, dengan kata lain penelitian ini tetap menggunakan distribusi asli dataset tanpa

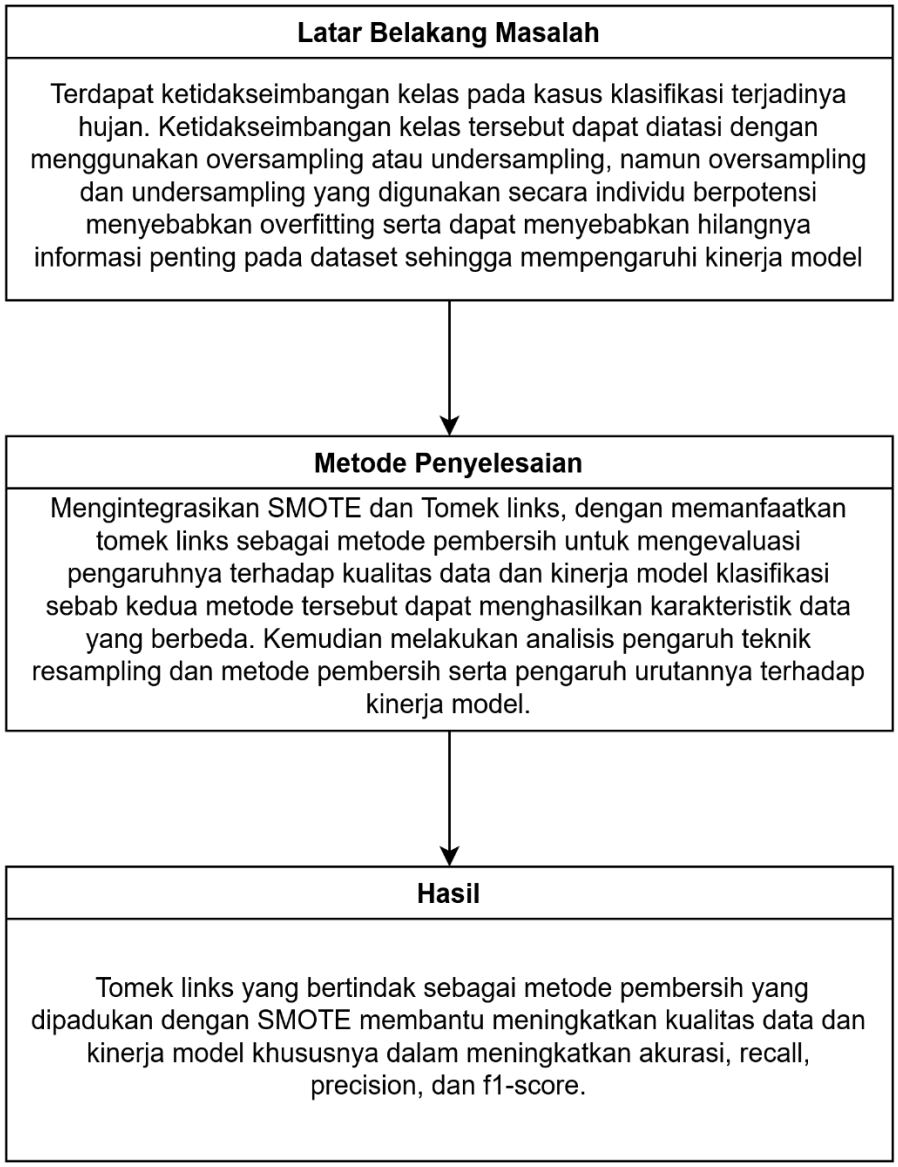
menerapkan teknik resampling, baik itu oversampling ataupun undersampling (Mahajan & Sharma, 2022).

4. *Machine Learning-Based Rainfall Prediction: Unveiling Insight and Forecasting for Improved Preparedness* – Penelitian ini menekankan bahwa pentingnya melakukan prediksi terjadinya hujan yang akurat agar dapat melakukan persiapan awal khususnya jika terjadi hujan ekstrem. Sama seperti penelitian sebelumnya, penelitian ini tetap menggunakan distribusi asli dari dataset tanpa mengatasi ketidakseimbangan kelas pada dataset. Adapun fokus utama dari penelitian ini adalah pengembangan dan evaluasi model prediksi hujan. Penelitian yang dilakukan oleh (Hassan et al., 2023) ini melakukan perbandingan antara tujuh algoritma *machine learning*, diantaranya *Naïve Bayes, Decision Tree*, *Support Vector Machine*, *Random Forest, Logistic Regression*, *Artificial Neural Network*, dan *Long Short-Term Memory*. Hal ini dilakukan untuk melihat kinerja model berdasarkan algoritma yang digunakan.
5. *Rainfall Prediction Enhancement Using SMOTE and Machine Learning Algorithms* – Penelitian ini ini menggunakan SMOTE untuk mengatasi distribusi data yang tidak seimbang kemudian melakukan komparasi antara berbagai algoritma *Machine Learning* untuk melihat performa dari model, penelitian ini juga melakukan *hyperparameter tuning* guna mendapatkan konfigurasi model terbaik (Hemanth Kumar et al., 2024).
6. *Leveraging Meteorological Data and Machine Learning for Improved Rainfall Forecasting in Australia* – Penelitian ini membandingkan antara beberapa algoritma *Machine Learning* dalam melakukan prediksi hujan. Penelitian ini menganalisis adanya ketidakseimbangan kelas pada dataset, namun tidak melakukan penerapan metode untuk menyeimbangkan kelas pada dataset (A. Das & Asieb Hasan, 2024).
7. *Rainfall Prediction Using Logistic Regression and Random Forest Algorithm* – Penelitian ini melakukan perbandingan antara algoritma Logistic Rgression dan Random Forest dalam melakukan prediksi hujan. Meski penelitian ini tidak berfokus pada penanganan dan pengelolaan ketidakseimbangan kelas, namun penelitian ini menggunakan SMOTE pada tahap *preprocessing* untuk mengatasi distribusi data yang tidak seimbang (Pandey et al., 2024). Perbedaan penelitian ini dengan penelitian yang penulis lakukan terletak pada fokus penelitian, penulis berfokus pada penerapan SMOTE dan *tomek links* sebagai metode pembersih kemudian mengidentifikasi efek dari penggunaan metode tersebut serta urutan penerapannya terhadap kinerja model.
8. *Credit Card Fraud Detection Using Machine Learning Techniques: Dealing with Imbalanced Data using Over-Sampling and Under-Sampling Methods* – Penelitian ini melakukan perbandingan antara beberapa *classifier* dengan melakukan pengujian pada metrik *precison* dan *recall*. Penelitian ini juga menganalisis kinerja metode penyeimbangan kelas SMOTE dan ROS+RUS pada setiap classifier. Hasil penelitian menjunjukkan adanya indikasi *trade-off precision* dan *recall* yang menunjukkan penanganan ketidakseimbangan kelas belum optimal (Hussain et al., 2024).

9. *Improving the Prediktion of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques* – Penelitian ini bertujuan untuk meningkatkan kinerja prediksi model dengan menggunakan SMOTE dan algoritma Machine Learning. Penelitian ini membandingkan kinerja model tanpa menggunakan SMOTE dan kinerja model dengan menggunakan SMOTE dengan berbagai algoritma Machine Learning, dan hasilnya menunjukkan dengan menggunakan SMOTE berhasil meningkatkan kinerja model (Ishaq et al., 2021).
10. *Enhanced Parkinson's Disease Prediction Using Feature Selection, SMOTE, anda Machine Learning with Deep Learning Models* – Penelitian ini menggunakan SMOTE untuk menangani ketidakseimbangan kelas lalu melakukan perbandingan antara beberapa algoritma *Machine Learning* untuk melihat kinerja model. Penelitian ini berfokus untuk memprediksi penyakit Parkinson secara tepat dengan mengevaluasi kinerja beberapa algoritma *Machine Learning* (Al-Amain et al., 2024).

Berdasarkan penelitian terdahulu pada tabel *State of The Art*, dapat dilihat bahwa hujan memiliki peran yang penting bagi kehidupan dan lingkungan namun klasifikasi hujan masih menjadi suatu kasus yang perlu diteliti lebih rinci lagi. Utamanya klasifikasi hujan dengan menggunakan *machine learning*, terdapat kondisi distribusi data yang tidak seimbang menjadikan model pembelajaran bias terhadap salah satu kelas, yaitu kelas mayoritas sehingga membuat model lebih banyak belajar pola dan karakteristik dari kelas mayoritas dan membuat model tidak mengenali kelas minoritas dengan baik. Permasalahan ketidakseimbangan kelas merupakan permasalahan yang tidak dapat dihindari terlebih pada beberapa kasus di dunia nyata, hal ini membuat penulis mengangkat permasalahan ini sebab implementasi teknik *resampling* (*oversampling* atau *undersampling*) yang digunakan secara individu, walaupun dapat menangani ketidakseimbangan kelas, tapi masih terdapat kekurangan seperti overfitting ataupun berpotensi kehilangan sampel yang memiliki informasi penting. Penelitian ini berfokus pada SMOTE dan tomek links yang berperan sebagai metode pembersih yang tidak hanya menyeimbangkan distribusi data, tetapi juga menganalisis efek penerapannya terhadap kinerja model pada kasus klasifikasi hujan. Pendekatan ini bertujuan untuk meningkatkan kualitas data dimana tomek links sebagai metode pembersih cenderung membersihkan sampel yang ambigu atau tumpang tindih di wilayah batas kelas.

## 2.3 Kerangka Pikir




**Gambar 5** Kerangka Pikir

Gambar 5 menunjukkan kerangka pikir dari penelitian ini, yang dimulai dengan adanya permasalahan ketidakseimbangan kelas pada kasus klasifikasi hujan. Kemudian dari permasalahan tersebut, ketidakseimbangan kelas ditangani dengan mengintegrasikan SMOTE dan *Tomek links*, dimana *Tomek links* akan bertindak sebagai metode pembersih sebelum masuk pada proses pelatihan.