BAB I
PENDAHULUAN

1.1 Latar Belakang

Data merupakan kumpulan informasi yang diperoleh dari pengamatan di mana dapat berupa angka, lambang, atau sifat (Jollyta dkk, 2020). Perkembangan akumulasi data saat ini sudah meningkat pesat tetapi pengolahannya belum maksimal. Banyak kebutuhan informasi yang harus dipenuhi tetapi tidak dapat diperoleh dengan mudah karena volume data yang sangat besar sehingga perlu adanya metode yang dapat mengorganisir dan mengklasifikasikan dokumen secara otomatis untuk mempermudah pencarian informasi yang relevan dengan kebutuhan, oleh karena itu peran *data mining* sangat diperlukan (Nurhayati dkk, 2019).

Data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Beynon-Davies, 2004). *Data mining* menganalisa data yang cenderung terus membesar dan teknik terbaik yang digunakan kemudian berorientasi kepada data berukuran sangat besar untuk mendapatkan kesimpulan dan keputusan yang paling layak (Elmande & Widodo, 2012). Model *data mining* dibuat berdasarkan jenis pembelajaran (*learning*) yaitu *supervised* dan *unsupervised* (Nengsih, 2019).

Fungsi pembelajaran *supervised* dapat digunakan untuk memprediksi suatu nilai atau membuat penggolongan (kelas-kelas) dari dokumen berdasarkan pada kelas kelompok yang sudah diketahui sebelumnya. Sementara fungsi pembelajaran *unsupervised* digunakan untuk mencari struktur instrinsik, relasi dalam suatu data yang tidak memerlukan kelas atau label sebelum dilakukan proses pembelajarannya atau dengan kata lain dokumen yang digunakan belum terlihat struktur kelompoknya (Darujati & Gumelar, 2012). Contoh algoritma pembelajaran *supervised* yaitu Naïve Bayes untuk klasifikasi dan contoh algoritma *unsupervised* yaitu k-means *clustering* dan apriori *association rules* (Ginantra dkk., 2021).

Salah satu proses penting dalam *data mining* adalah klasifikasi data. Pada persoalan klasifikasi, sejumlah kasus (data sampel) yang dimiliki akan diprediksi menjadi beberapa kelas yang ada pada data sampel tersebut. Setiap entitas (objek)

data memiliki banyak atribut, di mana masing-masing atribut memiliki satu dari beberapa kemungkinan nilai (Rizki & Amijaya, 2019). Hanya ada satu atribut yang disebut atribut target dan atribut yang lain disebut atribut prediktor. Klasifikasi biasa digunakan untuk segmentasi pelanggan, pemodelan bisnis, analisa kartu kredit, dan masih banyak pengaplikasiannya (Siregar & Puspabhuana, 2017). Contohnya, para dokter spesialis ingin memprediksi tingkat keganasan sebuah kanker berdasarkan tanda-tanda fisik yang dialami oleh penderita kanker (Mardi, 2016).

Terdapat berbagai metode klasifikasi, diantaranya analisis diskriminan, *decision tree*, regresi logistik, dan beberapa metode dengan pendekatan program komputasi seperti *Artificial Neural Network* (ANN), Naive Bayes, *Support Vektor Machine* (SVM), *Classification Adaptive Regression Tree* (CART), dan sebagainya (Witten, Frank, & Hall, 2011). Regresi logistik merupakan salah satu metode klasifikasi yang sering digunakan.

Kelebihan metode klasifikasi dengan menggunakan regresi logistik dapat melihat pengaruh beberapa variabel bebas yang bersifat numerik atau kategorik terhadap variabel terikat yang bersifat kategorik sebagai dasar untuk mengklasifikasikan pengamatan (Kutner, et al., 2005). Regresi logistik juga menjadi metode klasifikasi yang tangguh dengan memberikan ambang (*threshold*) probabilitas dan mencakup masalah klasifikasi *multiclass* (Karsmakers, et al., 2007). Pengklasifikasian umumnya menggunakan regresi logistik biner atau klasifikasi dengan dua kemungkinan. Namun ada kalanya pengklasifikasian lebih dari 2 kemungkinan atau *multiclass classification* sehingga harus menggunakan regresi logistik multinomial jika kategori respon merupakan nominal lebih dari 2 kategori, sementara untuk kategori respon bertingkat (ordinal) digunakan pendekatan regresi logistik ordinal (Ramadhini, 2018).

Pada kasus klasifikasi *multiclass* dengan sejumlah *dataset* besar sering ditemukan kondisi di mana himpunan data tidak seimbang (*imbalanced data*). Data tidak seimbang merupakan kondisi data yang tidak berimbang antar kelas data satu dengan kelas data yang lain. Kelas data yang lebih banyak disebut dengan kelas mayoritas sedangkan kelas data lainnya yang lebih sedikit disebut kelas minoritas.

Permasalahan data tidak seimbang ini mengakibatkan mesin *classifier* akan memprioritaskan untuk memprediksi kelas data yang banyak (mayoritas) dibandingkan dengan kelas minoritas sehingga akurasi prediksi lebih baik untuk data *training* kelas mayoritas sedangkan untuk data *training* kelas minoritas memiliki akurasi prediksi yang kurang baik (Chawla dkk, 2002). Permasalahan data tidak seimbang ini terjadi pada berbagai bidang antara lain klasifikasi teks, diagnosa medis, deteksi tumpahan minyak dari pencitraan satelit, deteksi penipuan kartu kredit, data medis kanker, dan lain-lain (Aminullah, 2021).

Permasalahan data tidak seimbang ini dapat diatasi dengan berbagai metode yang terbagi menjadi 3 kategori yaitu dengan pendekatan algoritma, pendekatan *data preprocessing*, dan pendekatan seleksi fitur (Heranova, 2019). Tentunya masing-masing dari teknik tersebut memiliki kelebihan dan kekurangannya (Sulasih, 2015). Beberapa pengembangan metode regresi logistik telah dilakukan untuk meningkatkan ketepatan klasifikasi pada data tidak seimbang.

Pengembangan regresi logistik ini diantaranya dengan pendekatan kernel ada Maalouf dan Trafalis (2010) mengembangkan metode *Rare Event Weighted Kernel Logistic Regression* (RE-WKLR) untuk data berukuran kecil sampai sedang. Rahayu (2012) juga mengembangkan metode dengan pendekatan kernel yaitu *AdaBoost Newton Truncated Regularized Weighted Kernel Logistic Regression (AB-WKLR)* dan *Adaboost NTR Weighted Regularized Logistic Regression (AB-WLR)*. Selanjutnya pendekatan non kernel juga dikembangkan oleh Maalouf dan Siddiqi (2014) yaitu metode *Rare Event-Weighted Logistic Regression (RE-WLR)* untuk klasifikasi data berskala besar dan menghasilkan performansi yang lebih baik dibandingkan *Truncated-Regularized Iteratively Re-Weighted Least Squares (TR-IRLS).* Pada tahun 2017 Maalouf membandingkan metode *Rare Event Weighted Logistic Regression (RE-WLR)* dengan metode *Truncated Regularized Prior Correction (TR-PC)* untuk mengklasifikasikan data set besar dengan kasus data tidak seimbang dan menyimpulkan bahwa RE-WLR menghasilkan performansi yang lebih baik.

Metode *Rare Event Weighted Logistic Regression (RE-WLR)* merupakan salah satu metode pengembangan dari regresi logistik di mana metode tersebut

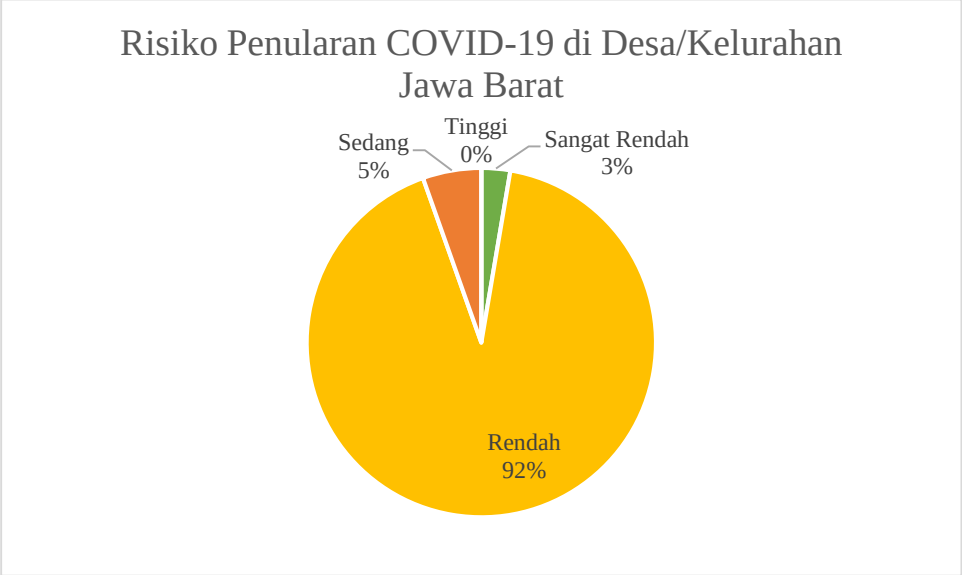
menerapkan regularisasi, *weighting* (pembobotan), dan *bias correction* (bias terkoreksi) yang digunakan untuk mengatasi permasalahan pada kasus tidak seimbang dengan data yang berukuran besar (Maalouf & Siddiqi, 2014).

Pada penelitian sebelumnya Sulasih (2016) mengkaji dan menerapkan metode RE-WLR untuk mengklasifikasikan desa tertinggal di Provinsi Jawa Timur dengan variabel respon dan prediktor yang digunakan terbatas pada klasifikasi biner. Dalam penelitiannya diperoleh hasil bahwa metode RE-WLR memberikan hasil kinerja klasifikasi yang lebih baik dibandingkan dengan metode TR-IRLS. Selanjutnya, Triasmoro (2018) melakukan penelitian dengan pendekatan *prior correction* dan membandingkan metode TR-PC dengan RE-WLR untuk mengklasifikasi tingkat kesejahteraan di Provinsi Bali dan diperoleh hasil bahwa metode RE-WLR menghasilkan akurasi ketepatan klasifikasi yang lebih baik dibandingkan dengan metode TR-PC. Pada kedua penelitian tersebut model regresi logistik yang digunakan adalah $p_i = \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}$.

Pada penelitian ini, peneliti tertarik untuk menerapkan metode *Rare Event Weighted Logistic Regression (RE-WLR)* untuk mengklasifikasikan kasus *multiclass* pada data tidak seimbang karena dinilai memiliki performansi yang lebih baik dibandingkan dengan metode penanganan data tak seimbang lainnya pada penelitian-penelitian terdahulu yang telah disebutkan dan model yang digunakan disederhanakan menjadi $p_i = \frac{1}{1+e^{-x_i^T \beta}}$ dengan batas $0 \leq p_i \leq 1$ yang diterapkan pada studi kasus klasifikasi tingkat risiko penularan COVID-19 di Provinsi Jawa Barat. COVID-19 merupakan wabah penyakit jenis baru yang ditemukan pertama kali di Wuhan, China (Dong, Mo, Hu, & al, 2020) dan dinyatakan sebagai "*Public Health Emergency of International Concern*" pada tanggal 30 Januari 2020, kemudian pada tanggal 11 Maret 2020 WHO meningkatkan status keadaannya menjadi pandemi untuk seluruh dunia termasuk Indonesia (Nikos, Konstantinos, & Betrand, 2020).

Jawa Barat merupakan provinsi dengan penduduk terpadat di Indonesia yang menyumbang sekitar 16,5% kasus terkonfirmasi COVID-19 dengan risiko penularan di Desa/Keluarahan Provinsi Jawa Barat pada tanggal 26 Desember 2021

tercatat sebanyak 159 Desa/Kelurahan berisiko sangat rendah (hijau), 5.475 Desa/Kelurahan berisiko rendah (kuning), 323 desa berisiko sedang (oranye), dan tidak ada Desa/Kelurahan berisiko tinggi (merah) (PIKOBAR, 2022).



Gambar 1.1 Presentase Tingkat Risiko Penularan COVID-19 di Jawa Barat
Sumber: https://pikobar.jabarprov.go.id/transmission-potential

Berdasarkan latar belakang tersebut, penelitian ini akan menerapkan metode *Rare Event-Weighted Logsitic Regression (RE-WLR)* dalam kasus klasifikasi tingkat risiko penularan COVID-19 di Jawa Barat karena perbedaan yang cukup jauh antar kelasnya atau tingkat *rarity* yang mencapai 92% pada proporsi tingkat berisiko rendah, 5% pada proporsi tingkat berisiko sedang, dan 3% pada proporsi tingkat berisiko sangat rendah penularan COVID-19 sehingga kasus ini termasuk ke dalam kasus *multiclass* dengan data tidak seimbang.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang sebelumnya, maka masalah pada penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana prosedur RE-WLR untuk mengklasifikasikan kasus *multiclass* dengan data tidak seimbang?
2. Bagaimana hasil ketepatan metode klasifikasi RE-WLR dalam pengklasifikasian kasus *multiclass* dengan data kasus klasifikasi tingkat risiko penularan COVID-19 di Desa/Kelurahan Provinsi Jawa Barat pada bulan Desember 2021?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan sebelumnya, maka tujuan dalam penelitian ini adalah:

1. Menjelaskan prosedur metode RE-WLR untuk mengklasifikasikan kasus *multiclass* dengan data tidak seimbang.
2. Menganalisis hasil ketepatan metode klasifikasi RE-WLR dalam pengklasifikasian kasus *multiclass* untuk data kasus klasifikasi tingkat risiko penularan COVID-19 di Desa/Kelurahan Provinsi Jawa Barat pada bulan Desember 2021.

1.4 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Secara Teoritis

   Penelitian ini menjelaskan bagaimana cara mengaplikasikan metode RE-WLR dalam mengklasifikasikan kasus *multiclass* data tidak seimbang yang akan menghasilkan performansi lebih baik dibandingkan dengan metode TR-IRLS dan hasil penelitian ini memperkuat teori yang menyatakan bahwa metode RE-WLR akan menghasilkan performansi lebih baik dibandingkan dengan metode TR-IRLS dalam mengklasifikasikan kasus *multiclass* data tidak seimbang. Secara praktis penelitian ini bermanfaat memberikan penjelasan bagaimana cara mengatasi permasalahan dan penanganan kasus *multiclass* dengan data tidak seimbang menggunakan metode RE-WLR pada studi kasus pengklasifikasikan tingkat risiko penularan COVID-19 di Desa/Kelurahan Provinsi Jawa Barat pada bulan Desember 2021.

2. Secara Praktis

   Penelitian ini bermanfaat memberikan penjelasan bagaimana cara mengatasi permasalahan dan penanganan kasus *multiclass* dengan data tidak seimbang menggunakan metode RE-WLR pada studi kasus pengklasifikasikan tingkat risiko penularan COVID-19 di Desa/Kelurahan Provinsi Jawa Barat pada bulan Desember 2021 dan hasil penelitian ini diharapkan dapat menambah wawasan tentang pentingnya penanganan COVID-19 yang sesuai dengan tingkat penularan risiko COVID-19 di Desa/Kelurahan Provinsi Jawa Barat.