BAB I
PENDAHULUAN

A. LATAR BELAKANG

Kualitas produk merupakan suatu hal yang penting dalam menentukan pemilihan suatu produk oleh konsumen. Produk yang ditawarkan harus produk yang benar-benar teruji dengan baik kualitasnya, mengingat yang diutamakan oleh konsumen adalah kualitas dari produk itu sendiri dan salah satu kriteria dari produk berkualitas adalah tidak adanya kecacatan pada produk itu sendiri. Konsumen akan lebih memilih produk yang mempunyai kualitas lebih baik dibandingkan dengan produk lain yang mempunyai kualitas lebih rendah. Kualitas produk adalah faktor kunci untuk menilai kemampuan produksi dan meningkatkan kompetensi Kao et al., (2017). Selain itu memiliki kualitas produk yang baik akan mendapatkan keuntungan dalam bentuk pembelian kembali/ulang oleh konsumen yang telah mengetahui kualitas produk tersebut. Untuk mengontrol dan meningkatkan kualitas produk banyak teknik yang diimplementasikan ke dalam proses pembuatan. Prediksi kualitas menjadi salah satu cara yang paling efektif untuk mengendalikan dan meningkatkan kualitas manufaktur dan telah dikembangkan menggunakan berbagai teknik *data mining* seperti pada penelitian yang dilakukan oleh Bai et al., (2017) kemudian menghasilkan pernyataan bahwa semakin besar ukuran sampelnya, semakin baik performanya. Lalu penelitian Kao et al., (2017) mengusulkan kombinasi analisis komponen utama dengan klasifikasi dan algoritma penambangan aturan asosiasi untuk mengekstrak fitur yang mewakili hubungan beberapa *workstation*, memprediksi kualitas produk akhir, dan menganalisis akar penyebab cacat produk.

Proses produksi di suatu perusahaan, pada umumnya menghasilkan produk cacat dalam jumlah yang sedikit dibanding produk yang berkualitas baik. Hal ini menyebabkan formulasi model prediksi kualitas biasanya diterapkan pada dataset yg tidak seimbang atau *imbalanced*. Ketidakseimbangan data merupakan suatu keadaan dimana distribusi kelas data tidak seimbang, jumlah kelas data (*instance*) yang satu lebih banyak atau sedikit dibandingkan dengan jumlah kelas data yang lainnya menurut Ali et al., (2015). Untuk kelompok kelas data yang lebih sedikit dikenal dengan sebutan kelompok minoritas (*minority*), sedangkan kelompok kelas data yg banyak dari yang lain disebut kelompok mayoritas (*mayority*).

Penelitian-penelitian yang dijelaskan sebelumnya oleh Bai et al., (2017) dan Kao et al., (2017) diterapkan tanpa memperhatikan permasalahan ketidakseimbangan data. *Imbalanced* dataset menyulitkan metode klasifikasi

dalam melakukan proses *machine learning*. Hampir semua algoritma klasifikasi seperti *Naïve Bayes, Random Forest*, C 4.5 dan lainnya menunjukkan performa yang buruk ketika di uji coba pada data dengan kelas yang tidak seimbang. Metode-metode yang telah disebutkan di atas tidak dilengkapi dengan kemampuan untuk menangani masalah ketidakseimbangan kelas data.

Pada proses yang dilakukan pada data yang tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas Gu et al., (2016). Perbedaan ini yang menjadikan suatu indikator performa klasifikasi yang buruk. Pada *imbalanced* data itu menunjukkan ketidakakuratan dikarenakan ada kelas mayoritas mendominasi yang menjadikan kelas minoritas tidak terdeteksi. Pada beberapa kasus justru kelas minoritas lebih penting dan menarik untuk diidentifikasi daripada kelas mayoritas. Misalnya pada cacat produk semikonduktor, dengan mengurangi jumlah cacat pada produk yang dihasilkan dapat meningkatkan kualitas semikonduktor itu sendiri yang tidak terlihat secara fisik namun berkurangnya fungsi, seperti dislokasi pada *halov perovskite* yang pada intinya menurunkan kemampuan menahan energi yang berasal dari cahaya dalam bentuk elektron Martialay, (2019). Hal tersebut dinilai sangat penting untuk mengetahui seberapa basar cacat yang ada pada kualitas produk semikonduktor tersebut meskipun jumlah kasusnya jauh lebih sedikit dibandingkan yang tidak cacat.

Pada penelitian sebelumnya telah banyak metode yang telah dilakukan untuk menyelesaikan permasalahan ketidakseimbangan data ini mulai dari pendekatan dengan teknik-teknik level data/*preprocessing*, pendekatan algoritmik hingga kombinasi antar keduanya. Penelitian Siringoringo, (2018), menyebutkan bahwa menerapkan metode SMOTE untuk menangani ketidakseimbangan kelas dengan metode *k-Nearest Neighbor* (kNN) untuk melaksanakan fungsi klasifikasi menghasilkan rata-rata nilai performa *G-Mean* sebesar 81,0% sedangkan rata-rata nilai performa *F-Measure* yakni 81,8% untuk skema SMOTE+kNN dibandingkan dengan skema kNN yang rata-rata nilai performa G-Mean yaitu 53,4% dan rata-rata nilai performa F-Measure yaitu 38,7%. Penelitian selanjutnya dari Saifudin et al., (2015) yang menggunakan FSMOTE untuk menangani ketidakseimbangan kelas dan Naïve Bayes untuk mengklasifikasikan yang menyebutkan bahwa kombinasi tersebut merupakan model pendekatan level data/*preprocessing* terbaik. Adapun yang menggunakan model pendekatan level data/*preprocessing* SMOTE dengan melakukan *resampling* ulang sampel kelas minoritas 600% yang memiliki performa terbaik menurut penelitian Karlık et al., (2016).

Metode *Synthetic Minority Over-Sampling Technique (SMOTE)* merupakan metode yang populer, yang sering 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). Penelitian Saifudin et al., (2015) menghasilkan bahwa model FSMOTE + Naïve Bayes merupakan model pendekatan level data terbaik pada prediksi cacat software karena nilai sensitivitas dan *G-Mean* model FSMOTE + Naïve Bayes meningkat secara signifikan, sedangkan model *Random Oversampling + Naïve Bayes* dan *Random Undersampling + Naïve Bayes* tidak meningkat secara signifikan. Begitupun pada penelitian (Kaur & Gosain, 2018) menyebutkan bahwa pendekatan *Synthetic Minority Over-Sampling Technique (SMOTE)* lebih baik dibandingkan *Random Undersampling* (RUS).

Seperti pada umumnya data kualitas, pada kasus produksi semikonduktor, yang direpresentasikan oleh dataset SECOM McCann Michael & Johnston, (2008), fenomena ketidakseimbangan data juga terjadi. Hal ini menyebabkan kesulitan jika akan melakukan formulasi model prediksi. Pada kasus ini terdapat 1567 data, dimana 1463 data merupakan data positif dan sisanya 104 merupakan data negatif/produk cacat. Karakteristik data tersebut cenderung memiliki lebih banyak data positif dengan *imbalance ratio* 1 : 14,067.

Penerapan teknik-teknik *preprocessing* level data untuk mengatasi ketidakseimbangan data seperti *Random Oversampling (ROS), Random Undersampling (RUS), dan Synthetic Minority Over Sampling Technique (SMOTE).* Seperti apa yang telah diulas pula sebelumnya, hanya terbukti efektivitasnya pada dataset tertentu. Dalam kasus manufaktur semikonduktor yang melibatkan data atribut numerik dengan jumlah atribut yang sangat banyak dengan kelas nominal masih perlu dikaji efektivitasnya. Selain itu, penerapan metode yang tidak tepat juga dapat menimbulkan pada performansi yang buruk seperti salah satunya akurasi model prediksi, sehingga diperlukan juga kajian lebih mendalam untuk menguji akurasi model yang dihasilkan.

B. RUMUSAN MASALAH

Seperti yang telah dijelaskan dalam latar belakang, formulasikan model prediksi kualitas dengan algoritma klasifikasi akan menghadapi masalah akurasi akibat ketidakseimbangan data. Pada kasus produksi semikonduktor yang direpresentasikan pada dataset SECOM, formulasi model prediksi juga menghadapi masalah-masalah sebagai berikut:

1. Belum diketahuinya teknik *preprocessing* yang efektif dalam mengatasi ketidakseimbangan data pada dataset dengan atribut numerik dan kelas nominal.

2. Hasil performansi yang beragam di setiap parameter yang diujikan pada penerapan teknik yang dilakukan ini menjadikan formulasi model prediksi kualitas ini bersifat kompleks.

C. RUANG LINGKUP PENELITIAN

Ruang lingkup penelitian ini terfokus pada :

1. Bidang manufaktur, yang khususnya di bagian prediksi kualitas pada data produk semikonduktor.
2. Pendekatan level data dan level algoritmik untuk permasalahan ketidakseimbangan data.

D. TUJUAN PENELITIAN

Berdasarkan permasalahan yang dibahas pada rumusan masalah, tujuan yang ingin dicapai pada peneltitian ini adalah sebagai berikut:

1. Menentukan teknik *preprocessing* yang tepat untuk digunakan pada masalah ketidakseimbangan data pada dataset dengan atribut numerik dan kelas nominal;
2. Menentukan kombinasi teknik *preprocessing* dan algoritma klasifikasi yang menghasilkan model dengan akurasi tinggi pada kasus *imbalanced dataset* dengan atribut numerik dan kelas nominal; dan
3. Memformulasikan model prediksi kualitas pada kasus manufaktur semikonduktor yang memiliki *imbalanced dataset* dengan atribut numerik dan kelas nominal.

E. SISTEMATIKA PENULISAN

Sistematika penulisan penelitian ini disusun dalam 5 bab, dimana di tiap bab tersebut akan dibagi lagi menjadi sub-bab yang akan dibahas secara terperinci. Berikut merupakan sistematika dari masing-masing bab dan keterangan singkatnya :

1. BAB 1 Pendahuluan

   Pada bab ini akan dibahas tentang gambaran umum penelitian, diantaranya latar belakang penelitian mengenai *imbalanced dataset* pada produk semikonduktor yang akan dikaji pada penelitian ini; rumusan masalah yang mengenai akurasi pada formulasi model prediksi kualitas; tujuan dari penelitian ini adalah menentukan teknik *preprocessing* yang tepat untuk digunakan pada masalah yang ada di penelitian ini; dan sistematika penulisan yang menjadi format acuan dalam penyusunan penelitian ini.
2. BAB 2 Studi Literatur

   Bagian ini berisi tentang berbagai teori penunjang pada penelitian ini. Adapun teori-teori yang disajikan mengenai kualitas produk; prediksi

kualitas; *imbalanced data*; metode *resampling/preprocessing*; algoritma pengklasifikasian; dan pengukuran performa untuk teknik yang telah diujikan.

3. BAB 3 Metodologi Penelitian
   Pada bab ini berisi tentang metode yang akan digunakan pada penelitian ini, yang didalamnya berisi mengenai langkah penelitian (*flowchart*); data *understanding* yaitu data yang digunakan pada penelitian; data *preparation*; *modelling*; evaluasi; dan analisis.
4. BAB 4 Hasil dan Pembahasan
   Pada bab ini menjelaskan tentang hasil dari sebuah penelitian yang telah dilakukan. Hasil penelitian yang akan ditampilkan seperti hasil dari pengolahan data *preprocessing*; perbandingan hasil validasi performa; penentuan metode terbaik; model prediksi; dan analisis.
5. BAB 5 Penutup
   Berisi kesimpulan dan saran. Dalam bagian ini menyimpulkan hasil penelitian yang telah dilakukan dan sesuai dengan permasalahan penelitian. Setelah hasil penelitian disimpulkan, peneliti juga harus mampu memberikan saran yang operasional berdasarkan temuan penelitian.