BAB I
PENDAHULUAN

1.1 Latar Belakang

Klasifikasi adalah proses untuk menemukan sebuah model yang dapat membedakan label atau kelas pada data, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi label atau kelas yang belum diketahui sebelumnya. Klasifikasi biasanya digunakan untuk mendeteksi penipuan kartu kredit, diagnosa medis, dan lain-lain [1].

Terdapat permasalahan yang seringkali muncul pada klasifikasi, yaitu ketidakseimbangan kelas. Ketidakseimbangan kelas merupakan suatu keadaan dimana suatu kelas memiliki jumlah *instance* yang lebih besar dibandingkan dengan jumlah kelas lainnya. Kelas yang jumlah instance lebih besar disebut mayoritas, sedangkan kelas yang jumlah instance lebih kecil disebut minoritas. Algoritma klasifikasi yang dilatih dengan data dengan kelas tidak seimbang dapat menyebabkan kesalahan dalam melakukan klasifikasi. Algoritma klasifikasi akan menghasilkan akurasi yang tinggi pada kelas mayoritas, dan menghasilkan akurasi yang rendah pada kelas minoritas. Dengan kata lain, algoritma klasifikasi akan mengklasifikasikan kelas mayoritas saja dan cenderung mengabaikan kelas minoritas, bahkan kelas minoritas akan diklasifikasikan sebagai kelas mayoritas [2].

Terdapat beberapa pendekatan untuk mengatasi ketidakseimbangan kelas salah satunya adalah teknik *resampling*. Teknik *resampling* merupakan teknik yang meyeimbangkan jumlah *instance* suatu kelas pada *dataset* menggunakan beberapa metode sampling yaitu *oversampling*, *undersampling*, dan *hybrid* [1]. Teknik *resampling* merupakan sebuah solusi yang paling sering digunakan oleh para peneliti saat menghadapi kasus *dataset* yang tidak seimbang. Selain dapat merubah jumlah data suatu kelas pada *dataset*, teknik *resampling* dapat memperbaiki performa algoritma klasifikasi dalam mengklasifikasikan *dataset* yang tidak seimbang [3].

Pada penelitian ini, peneliti akan melakukan perbandingan antar teknik *resampling* pada klasifikasi *dataset* yang tidak seimbang. Perbandingan antar teknik *resampling* yang telah dilakukan sebelumnya hanya diujikan pada satu *dataset* saja. Tentunya hal itu akan menimbulkan suatu pertanyaan mengenai konsistensi performa teknik *resampling* seperti dapatkah teknik *resampling* tersebut dapat menyeimbangkan jumlah data pada suatu kelas, lalu setelah dilakukan *resampling* apakah performa algoritma klasifikasi dapat membaik dalam mengklasifikasikan *dataset* yang tidak seimbang.

Beberapa penelitian mengenai perbandingan antar teknik *resampling* ini belum ada yang menyimpulkan bahwa teknik *resampling* tersebut yang terbaik pada klasifikasi *dataset* yang tidak seimbang. Perbandingan teknik *resampling* selalu mendapatkan hasil yang berbeda, maksudnya adalah pada suatu penelitian teknik *resampling* tersebut memperoleh performa yang paling baik, namun pada penelitian lain performa teknik *resampling* tersebut mendapat performa yang kurang baik.

Maka dari itu untuk mengetahui konsistensi performa teknik *resampling* pada klasifikasi *dataset* yang tidak seimbang, peneliti akan menguji teknik *resampling* tersebut pada beberapa *dataset* dengan *imbalance ratio* yang berbeda-beda. Teknik *resampling* yang dipilih untuk penelitian ini adalah SMOTE dan *random undersampling*.

Peneliti memilih SMOTE (*Synthetic Minority Oversampling Technique*) karena metode ini lebih cerdas daripada *random oversampling*. Menurut (A. Fernández, dkk) SMOTE dianggap sebagai salah satu teknik yang paling berpengaruh dan menjadi pelopor bagi para komunitas riset dalam kasus ketidakseimbangan kelas. SMOTE melakukan pendekatan *oversampling* yang berbeda dibandingkan metode *oversampling* sebelumnya yaitu *random oversampling*. SMOTE akan membuat sebuah *instance* baru yang dinamakan *instance* sintetik (mirip seperti aslinya) alih-alih melakukan duplikasi pada kelas minoritas, sehingga dapat meminimalisir terjadinya *overfitting* [4].

Penelitian sebelumnya telah dilakukan oleh (Erlin, dkk) menggunakan metode SMOTE (*Synthetic Minority Oversampling Technique*) yang diuji menggunakan algoritma *random forest* pada *dataset* penyakit jantung. SMOTE dapat meningkatkan *sensitivity* dari yang sebelumnya (**85%**) menjadi (**91%**). Nilai *g-mean* pun meningkat dari sebelumnya (**90%**) menjadi (**94%**) yang berarti algoritma *random forest* memiliki kinerja yang baik dalam mengklasifikasikan kelas mayoritas dan minoritas. SMOTE dapat meminimalisir algoritma *random forest* mengalami *overfitting* dengan memperoleh akurasi data latih (**86%**) dan data uji (**90%**), dari yang sebelumnya akurasi data latih (**100%**) dan data uji (**87%**) [2].

Peneliti memilih *random undersampling* karena menurut (T. Hasanin, dkk) pada dapat mengurangi beban komputasi dan mempersingkat waktu dalam melakukan pengolahan data yang lebih besar [5]. Menurut (M. Bach, dkk) pada *random undersampling* telah dibuktikan oleh beberapa penelitian merupakan salah satu metode *resampling* yang lebih efektif. *Random undersampling* seringkali sangat sulit untuk diungguli dengan metode *undersampling* lainnya yang lebih canggih dan tak jarang mampu mengungguli metode oversampling [6].

Penelitian yang dilakukan oleh (J. Prasetya) membandingkan *random undersampling* dan *random oversampling* yang diuji menggunakan algoritma *naïve bayes* pada *dataset cervical cancer risk factors*. *Random undersampling* memperoleh nilai AUC lebih baik sebesar (**70%**) yang berarti klasifikasi cukup baik dibandingkan *random oversampling* dengan nilai AUC (**62%**) yang berarti klasifikasi buruk, dari yang sebelumnya nilai AUC hanya sebesar (**53%**) [7].

Penelitian yang dilakukan oleh (D. J. Dittman, dkk) membandingkan *random undersampling*, *random oversampling*, dan SMOTE pada yang diuji menggunakan algoritma *K-Nearest Neighbor* dan *Support Vector Machine* pada 6 *dataset* berbeda yaitu (*Brain Tumor*, *Chanrion* 2008, GSE20271, GSE3494-GPL96-ER, *Mulligan R-PD*, dan *Ovarian MAT*). Pada algoritma *k- nearest neighbor*, *random undersampling* memperoleh nilai AUC yang lebih baik pada 4 dataset, yaitu pada *dataset Brain*

Tumor (**89%**), *dataset* GSE20271 (**64%**, *dataset* GSE3494-GPL96-ER (**87%**), dan *Ovarian MAT* (**94%**). Pada algoritma *support vector machine*, *random undersampling* memperoleh nilai AUC yang lebih baik pada 4 *dataset*, yaitu pada *dataset Chanrion* 2008 (**81%**), *dataset* GSE20271 (**68%**), GSE3494-GPL96-ER (**89%**), dan dataset *Ovarian MAT* (**93%**). [8].

Penelitian ini akan diuji performanya dengan 3 *dataset* yang memiliki tingkat *imbalance ratio* yang berbeda dan 3 algoritma klasifikasi. Teknik *resampling* terbaik akan ditentukan berdasarkan persentase peningkatan tertinggi pada nilai *area under the curve* (AUC) dan *geometric mean* (*g-mean*), yang didapat dari hasil evaluasi dengan *confusion matrix* dengan menggunakan indikator *recall* (*sensitivity*) dan *specificity*.. Tujuan dari penelitian ini adalah untuk mengetahui konsistensi performa teknik *resampling* dalam memperbaiki performa algoritma klasifikasi dan menentukan teknik *resampling* mana yang terbaik pada klasifikasi *dataset* yang tidak seimbang.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah seperti yang dijelaskan diatas, peneliti merumuskan masalahnya sebagai berikut:

1. Bagaimana performa algoritma klasifikasi di setiap dataset dengan *imbalance ratio* yang berbeda setelah dilakukan *resampling*?
2. Teknik *resampling* mana yang paling baik dalam meningkatkan performa algoritma klasifikasi pada *dataset* tidak seimbang?

1.3 Batasan Penelitian

Berdasarkan identifikasi masalah diatas diperlukan pembatasan masalah supaya tidak terlalu melebar dari masalah diatas. Berikut batasan-batasan masalah pada penelitian kali ini:

1. Metode *resampling* yang digunakan yaitu *random undersampling* dan SMOTE (*Synthetic Minority Oversampling Technique*).

2. Dataset yang digunakan merupakan dataset tidak seimbang untuk klasifikasi *binary class*.
3. Algoritma klasifikasi yang digunakan yaitu *random forest*, *naïve bayes*, dan *k-nearest neighbor*.
4. Penelitian ini tidak akan membahas selain teknik *resampling* yang digunakan.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah diatas, tujuan yang ingin dicapai pada penelitian ini adalah:

1. Mengetahui performa *random undersampling* dan SMOTE terhadap algoritma klasifikasi pada setiap dataset dengan *imbalance ratio* yang bervariasi.
2. Mengetahui teknik *resampling* mana yang dapat digunakan pada berbagai tingkat *imbalance ratio* dan berbagai algoritma klasifikasi.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat, baik secara teoritis maupun praktis, diantaranya yaitu:

1. **Segi Teoritis:**
   a. Mengenalkan salah satu permasalahan yang terjadi pada klasifikasi *machine learning* yaitu ketidakseimbangan kelas pada *dataset*.
   b. Mengetahui cara mengatasi masalah ketidakseimbangan kelas pada *dataset*.
2. **Segi Praktis**:
   a. Memberikan rekomendasi metode *resampling* yang dapat menangani ketidakseimbangan kelas pada dataset untuk penelitian selanjutnya.

b. Menjadi referensi bagi para peneliti yang ingin mengembangkan penelitian mengenai ketidakseimbangan kelas pada *dataset*.

1.6 Sistematika Penulisan

Sistematika penulisan memuat uraian secara garis besar dan singkat untuk pada setiap bab:

1. BAB I PENDAHULUAN berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, dan manfaat penelitian.
2. BAB II TINJAUAN PUSTAKA berisi studi literatur dan dasar-dasar teori yang berhubungan dengan imbalanced dataset dan teknik resampling.
3. BAB III METODE PENELITIAN berisi tahapan-tahapan dilakukan, serta alat dan bahan yang dibutuhkan selama penelitan.
4. BAB IV HASIL DAN PEMBAHASAN berisi hasil penelitian yang telah dilakukan beserta pembahasan.
5. BAB V PENUTUP berisi kesimpulan selama penelitian dan saran untuk penelitian selanjutnya.