BAB I
PENDAHULUAN

1.1 Latar Belakang

Ketidak seimbangan data sering terjadi dalam pengambilan sebuah informasi pada minat tertentu. Dicirikan dengan adanya jumlah salah satu kelas jauh lebih banyak dibandingkan jumlah kelas lainnya. Kasus ketidak seimbangan data juga berpengaruh dalam pengambilan keputusan. Penelitian sebelumnya [1] menunjukkan bahwa kasus ini berpengaruh pada manajemen risiko dan diagnosis medis karena prediksi yang salah tentang tidak adanya penyakit kanker (negatif palsu) akan menimbulkan kematian, sedangkan prediksi yang salah tentang adanya penyakit kanker (positif palsu) akan menimbulkan kecemasan dan penanganan medis yang tidak diperlukan. Salah satu cara untuk menangani kasus ketidak seimbangan data ini yaitu menggunakan metode *sampling*.

Metode *sampling* yang dapat digunakan untuk mengatasi ketidak seimbangan data meliputi: *Random Under Sampling method (RUS)* dan *Random Over Sampling method* (*ROS*) [2]. Kedua metode ini dapat dilakukan dengan memperlakukan data minoritas dan data mayoritas secara berbeda. Pada penelitian yang pernah dilakukan sebelumnya [3] menggunakan metode *Random Under Sampling (RUS)* dapat menangani kasus ketidak seimbangan data pada prediksi cacat *software*, dimana terdapat data yang memiliki jumlah mayoritas akan dikurangi jumlahnya hingga mencapai akurasi yang tepat sesuai dengan data yang memiliki jumlah minoritas. Metode ini sangat tepat

dilakukan apabila data yang berjumlah minoritas adalah data yang memiliki kelas negatif, sehingga untuk menampilkan data yang bernilai benar maka data mayoritas yang memiliki kelas positif harus dikurangi jumlahnya menggunakan metode *Random Under Sampling (RUS*).

Penanganan yang berbeda dengan menggunakan metode *Random Over Sampling (ROS)*. Hasil eksperimen yang pernah dilakukan Alejandro Moreo [4] bisa meningkatkan keakuratan pada klasifikasi teks biner dimana contoh-contoh yang memiliki kelas negatif memiliki jumlah yang jauh lebih banyak dibandingkan dengan contoh-contoh yang memiliki kelas positif. Sehingga untuk menampilkan data yang bernilai benar yaitu dengan cara menduplikasi atau memperbanyak jumlah data yang memiliki kelas positif hingga mencapai keakuratan yang tepat menggunakan metode *Random Over Sampling (ROS)*. Metode ROS sangat berkaitan dengan penelitian yang dilakukan, karena data yang digunakan memiliki kelas positif jauh lebih sedikit dibandingkan dengan kelas yang bernilai negatif, sehingga untuk mendapatkan tingkat akurasi yang tepat maka digunakan metode *Random Over-Sampling (ROS)*.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah diatas, maka dapat dirumuskan masalah yaitu nilai akurasi terbaik didapatkan sebelum dilakukan *balancing* atau setelah dilakukan *balancing* dengan menerapkan metode *Random Over-Sampling*?

1.3 Batasan Masalah

Agar penelitian dan penulisan tidak menyimpang, maka hanya difokuskanpada beberapa pokok permasalahan dalam penerapan metode *Random Over- Sampling* ini. Dataset yang diambil berupa teks berbahasa inggris dan memiliki perbandingan kelas yang sangat jauh, maka dari itu untuk mendapatkan akurasi terbaik pencarian akurasi akan dilakukan sebelum menerapkan metode *Random Over-Sampling* dan setelah menerapkan metode *Random Over-Sampling*.

1.4 Maksud dan Tujuan Penelitian

Adapun tujuan yang ingin dicapai dari kegiatan penelitian ini adalah untuk mendapatkan akurasi data terbaik dengan menerapkan metode *Random Over-Sampling*.

1.5 Manfaat Penelitian

Setiap hasil penelitian pada prinsipnya harus memiliki nilai manfaat, adapun manfaat dari penelitian ini yaitu untuk meningkatkan pengetahuan tentang penerapan metode *Random Over-Sampling* pada data teks.

1.6 Metode Penelitian

Metode penelitian yang dilakukan yaitu pengambilan data dari dataset public yang bisa diakses melalui *platform* seperti Kaggle, dengan data yang sudah diklasifikasikan menjadi *multi-class* data, setelah itu dilakukan *pre-*

processing data. Metode yang dilakukan yaitu dengan menerapkan algoritma klasifikasi KNN dengan tanpa melakukan *oversampling* menggunakan ROS lalu dibandingkan dengan melakukan *oversampling* menggunakan ROS, dan setelah dilakukan perbandingan kemudian dilakukan *testing*.

1.7 Sistematika Penulisan

Secara garis besar laporan tugas akhir ini dibagi dan disusun dalam 5 bab, yaitu:

BAB I: PENDAHULUAN

Berisi tentang latar belakang dari pembuatan tugas skripsi, tujuan dan manfaat dari pembuatan skripsi, rumusan masalah dan batasan masalah dari pembuatan skripsi, metodologi penelitian dan sistematika penulisan dari skripsi.

BAB II: LANDASAN TEORI

Berisi tentang penjelasan secara teoritis yang akan mengarah ke proses pembuatan skripsi.

BAB III: METODE PENELITIAN

Bab ini berisi tentang perancangan dari penerapan metode yang digunakan dalam pembuatan skripsi.

BAB IV: HASIL DAN PEMBAHASAN

Bab ini berisi tentang pembahasan dari pembuatan skripsi dan hasil yang sudah dibuat.

BAB V: PENUTUP

Bab ini berisi tentang kesimpulan dan saran dari hasil penelitian yang telah dilakukan.