1. Pendahuluan

Twitter merupakan layanan yang memiliki kemudahan untuk saling berbagi informasi secara *real-time*. Di dalam Twitter terdapat fitur *retweet* yang merupakan bentuk penyebaran informasi dan sarana untuk berpartisipasi dalam sebuah percakapan yang meyebar, dimana seseorang dapat memposting ulang *tweet* yang ditulis oleh pengguna lain saat menemukan *tweet* yang menarik untuk dibagikan kepada pengikutnya [1]. Penyebaran *tweet* tidak hanya menyampaikan suatu informasi ke pengguna lain saja, tetapi bisa dengan memvalidasi dan mengomentari isi *tweet* [2]. Penyebaran informasi di Twitter tidak merata, banyak informasi yang dibagikan di jejaring sosial, tetapi dari banyaknya informasi yang dibagikan hanya informasi tertentu yang dapat menyebar lebih luas dibandingkan informasi lainnya [1]. Artinya, dari banyaknya *tweet* yang dibagikan tersebut ada yang mendapat *retweet* dan ada yang tidak mendapat *retweet*, oleh karena itu bagaimana membangun model yang dapat memprediksi apakah suatu *tweet* mendapat kelas *retweet* atau tidak, apa faktor yang mempengaruhi *tweet* tersebut di *retweet* dan apa metode yang dapat mengatasi ketidakseimbangan data.

Suh dkk. meneliti sejumlah fitur yang mempengaruhi retweetabilitas *tweet*, peneliti membangun model *retweet* dengan langkah-langkah yang sederhana. Peneliti mengumpulkan fitur konten (URL, *hashtags* dan *mention*) selain itu terdapat fitur kontekstual (jumlah *tweet*, jumlah *following*, jumlah *followers*, jumlah *tweet favorit* dan usia akun) dari 74 juta *tweet* tersebut kumpulan data digunakan untuk mengidentifikasi faktor yang sangat terkait dengan tingkat *retweet*. Dari penelitian tersebut ditemukan bahwa dari fitur konten, URL dan *hashtags* memiliki hubungan yang kuat dengan *retweetability*. Pada fitur kontekstual jumlah *following*, jumlah *followers*, dan usia akun mempengaruhi retweetabilitas. Sementara jumlah *tweet* sebelumnya tidak dapat memprediksi retweetabilitas *tweet* pengguna [1].

Tugas akhir ini bertujuan untuk merancang suatu model menggunakan algoritma pohon keputusan jenis CART untuk memprediksi apakah suatu *tweet* masuk ke dalam kelas *retweet* atau tidak menggunakan fitur berbasis pengguna, berbasis waktu dan berbasis konten, serta mengatasi ketidakseimbangan data dengan *oversampling* dan *undersampling*. Metode CART bisa digunakan pada himpunan data yang jumlahnya besar dan variabel yang banyak untuk menentukan aturan-aturan yang kompleks. Penelitian tentang analisis algoritma CART pernah dilakukan oleh Zimmerman dkk untuk memprediksi *influenza* pada pasien perawatan primer, hasilnya algoritma CART memiliki sensitivitas yang baik dan *Negative Prediction Value* (NPV) yang tinggi, tetapi *Positive Prediction Value* (PPV) yang rendah untuk mengidentifikasi *influenza* pada pasien rawat jalan ≥ 5 tahun. Sehingga, bagus untuk mengidentifikasi kelompok yang tidak memerlukan tes atau antivirus dan memiliki kinerja prediktif yang baik untuk *influenza* [3].