BAB I

PENDAHULUAN

1.1 Latar Belakang

Menganalisis data dalam jumlah besar adalah salah satu pekerjaan yang menyulitkan bagi kebanyakan orang. Seiring dengan berkembangnya teknologi informasi, jumlah informasi yang tersimpan dalam *database* pendidikan akan terus menerus meningkat setiap tahunnya. *Database* yang besar tersebut menyimpan banyak informasi berharga berkaitan dengan data para siswa pada suatu Perguruan Tinggi. Permasalahan yang sulit muncul ketika akan menemukan informasi berharga yang tersembunyi dalam *database*, lalu mengidentifikasi serta mengklasifikasikannya untuk membangun sebuah model yang bertujuan untuk membantu level manajemen dalam mengambil keputusan (Pradeep dkk., 2015).

Data yang melimpah pada *database* pendidikan/akademik menyebabkan proses pengambilan informasi yang berdampak pada proses pengambilan keputusan yang lambat. Oleh karena itu, instansi pada perguruan tinggi harus mampu mengola semua data yang ada secara cepat dan efisien agar dapat menghasilkan informasi yang berkualitas sehingga berdampak kepada pengambilan keputusan yang tepat (Akbar dkk., 2017). Salah satu cara yang efektif untuk menganalisa data histori pendidikan dan serupa dalam mencari tren dan pola adalah melalui *business intelligence* dan teknik *data mining*, untuk membangun model dan kemudian mengekstrak pengetahuan (Witten dkk., 2011). Teknik *data mining* memerankan peran yang sangat penting dalam hal ini. Terdapat satu teknik *data mining* yang membahas mengenai penambangan data pendidikan atau sering disebut dengan *educational data mining* (EDM) (Pradeep dkk., 2015).

Mining data merupakan teknik penambangan data yang digunakan untuk menambang sekumpulan data menjadi informasi/pengetahuan yang baru dan bermanfaat dengan cara menganalisis repositori data yang besar untuk mengekstraksi pola-pola penting, asosiasi dan hubungan diantara semua variable/atribut pada sebuah data. EDM adalah salah satu disiplin

ilmu yang terlibat dengan pengembangan metode dan teknik, tidak hanya mengeksplorasi dan menganalisis data yang berasal dari konteks pendidikan tetapi juga untuk mengekstraksi informasi yang tersembunyi untuk memahami siswa dengan lebih baik melalui data yang tersedia (Gutierrez dkk., 2018).

Drop out merupakan salah satu isu yang paling umum pada dunia pendidikan yang membutuhkan perhatian lebih. Tingkat *drop out* yang tinggi dapat memberikan dampak buruk pada sebuah Perguruan Tinggi (Dharmawan dkk., 2018). Reputasi dari sebuah perguruan tinggi diukur berdasarkan persentasi lulusan pada perguruan tinggi dan bagaimana strategi yang dimiliki oleh perguruan tinggi untuk mempertahankan mahasiswanya agar tidak *drop out*. Prediksi awal mahasiswa yang berisiko *drop out* sangat penting dilakukan untuk menentukan tingkat keberhasilan strategi pada masing-masing perguruan tinggi (Gutierrez dkk., 2018).

Beberapa metode telah diterapkan untuk melakukan proses klasifikasi dan prediksi pada data pendidikan khususnya permasalahan tingkat *drop-out* pada perguruan tinggi seperti metode *naïve bayes, logistic regression, classification and regression tree* (CART), *decision tree* dan *random forest* (Gutierrez dkk., 2018; Devasia dkk., 2016; Kovacic, 2010; Shiratori, 2018; Ketui dkk., 2019). Pada penelitiannya, Kovacic (2010) menggunakan model klasifikasi statistik seperti *logistic regression* yang hanya menyediakan sejumlah variabel prediktor dan memberikan hasil akhir klasifikasi yang tidak signifikan sesuai dengan fakta asli. Sedangkan penggunaan metode klasifikasi pohon tunggal seperti *classification and regression tree* (CART) dan *decision tree* memberikan hasil yang buruk untuk dataset yang membutuhkan banyak variable prediktor (Gutierrez dkk., 2018), ketepatan dugaan yang lebih rendah dibandingkan dengan pembentukan model pada pohon *ensemble* (Bock dkk., 2010; Kocev dkk., 2013), efisiensi komputasi yang lebih rendah serta terdapat variasi bias (Li dan Zhang, 2010). Naïve bayes dan *decision tree* keduanya memiliki kelemahan lamanya waktu proses pembentukan model dan rendahnya tingkat akurasi penilaian, selain itu dengan sifat kesederhanaan yang dimiliki oleh kedua metode klasifikasi tersebut menarik banyak peneliti telah mengimplementasikannya dalam berbagai macam aplikasi yang tersedia (Rosandy, 2016).

Decision tree merupakan *banchmark* (patokan) awal dari proses pembentukan pohon pada algoritma *random forest*. Permasalahan rendahnya tingkat akurasi yang dihasilkan oleh model yang dibangun dengan algoritma *classification and regression tree* (CART) dan *decision tree* sebagai pengklasifikasi pohon tunggal dapat diatasi oleh algoritma *random forest* sebagai pengklasifikasi pohon acak karena algoritma *random forest* memiliki teknik *ensemble bagging* yang mampu menaikkan performa model klasifikasi (Zhou, 2012; Kumari dkk., 2018; Amrieh dkk., 2016). Selain itu, *random forest* juga dapat mengatasi permasalahan yang terdapat pada algortima *logistic regression* dengan mampu memberikan *error* yang lebih rendah pada hasil prediksi sehingga memberikan hasil prediksi yang lebih baik, mampu menangani jumlah data pelatihan yang besar dengan efisien, metode tersebut juga efektif dalam menangani data yang hilang (Breiman, 2001). Kelebihan lain dari algoritma *random forest* yaitu sudah dilengkapi dengan metode pemilihan fitur (metode seleksi atribut) selama proses eksekusi pemodelan pada algoritma pembelajaran (*machine learning*) (Jovic dkk., 2015). Dataset pada penelitian ini termasuk dataset dengan distribusi kelas pada data yang tidak seimbang, maka permasalahan kelas pada dataset yang tidak seimbang akan diselesaikan dengan salah satu metode *oversampling* yaitu SMOTE. Algoritma ini mampu menghasilkan *instance* sintetis daripada duplikat instance kelas minoritas, oleh karena itu masalah *overfitting* dapat dihidari (Kasanah dkk. 2019).

Berdasarkan penjelasan tentang beberapa algoritma klasifikasi pada *data mining*, maka dipilih algoritma *random forest* sebagai teknik *ensemble machine learning* pada penelitian ini yang didukung dengan penggunaan teknik penyeimbangan dataset yaitu SMOTE. Model yang dihasilkan kemudian diterapkan dalam pengembangan sistem prediksi *drop out*.

1.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah dapat menerapkan algoritma *random forest* yang merupakan kelanjutan dari algoritma *decision tree* yang merupkan patokan / *benchmark* dari pembentukan pohon klasifikasi untuk membuat model yang bertujuan memprediksi *drop-out*.

1.3 Manfaat Penelitian

Manfaat dari penelitian ini adalah munculnya nilai akurasi dari prediksi *drop out*, sehingga dapat meningkatkan probabilitas keberhasilan dalam masa studi untuk tahun-tahun berikutnya serta menjadikan *early warning* pada program studi/fakultas/perguruan tinggi untuk terus mengevaluasi kinerja mahasiswa sehingga permasalahan mahasiswa yang berpotensi *drop out* dapat diminimalisir.