BAB I

PENDAHULUAN

1.1 Latar Belakang

Seleksi fitur merupakan salah satu bidang penelitian dalam *data mining* untuk dataset yang memiliki atribut yang relatif banyak. Seleksi fitur adalah salah satu faktor yang paling penting yang dapat mempengaruhi tingkat akurasi klasifikasi karena jika dataset berisi sejumlah fitur, dimensi dataset akan menjadi besar hal ini membuat rendahnya nilai akurasi klasifikasi. Hal ini bertujuan untuk mereduksi data yang ada sehingga dapat mempercepat waktu komputasi. Tidak hanya mempercepat waktu komputasi, seleksi fitur juga dapat meningkatkan akurasi klasifikasi dari algoritma yang digunakan (Dony dkk, 2017).

Sistem deteksi intrusi adalah salah satu area penelitian yang penting dalam jaringan dan keamanan komputer. Beberapa penelitian telah dilakukan untuk deteksi intrusi menggunakan data mining berbasis klasifikasi untuk menentukan paket data yang terekam apakah masuk dalam kategori serangan atau bukan. *Framework* sistem deteksi intrusi jaringan berbasis *naïve bayes classifier* telah diusulkan dan dibandingkan dengan pendekatan jaringan syaraf tiruan, hasil menunjukkan bahwa metode yang diusulkan menghasilkan tingkat deteksi yang lebih tinggi namun menghasilkan *false positive* yang cukup tinggi (Mrutyunjaya, 2007).

Dataset NSL-KDD merupakan dataset yang dapat digunakan sebagai pembanding berbagai metode klasifikasi untuk deteksi intrusi. Dataset ini memiliki dimensi yang cukup tinggi dengan 41 fitur. Seleksi fitur merupakan

salah satu proses penting untuk mereduksi dataset dengan menghilangkan fitur yang tidak penting dari dataset NSL-KDD tidak semua fitur yang ada dalam dataset tersebut memiliki pengaruh terhadap kelas atau label. Oleh karena itu menghilangkan atribut yang tidak penting dengan kelas label merupakan hal penting yang harus dilakukan untuk meningkatkan kinerja *classifier*. Tujuan utama dari seleksi fitur adalah memilih fitur yang penting dan menghapus fitur yang tidak penting dan kurang penting terhadap kelas label guna meningkatkan kinerja *classifier* yaitu meningkatkan akurasi dan mengurangi waktu komputasi.

Salah satu teknik yang dapat mengolah informasi dari kumpulan data yang besar menggunakan *data mining*. Pada penelitian sebelumnya, sudah terdapat beberapa pemanfaatan data mining menggunakan metode tertentu untuk meningkatkan kinerja klasifikasi pada suatu dataset. Seperti penelitian yang dilakukan oleh Mohamad Fajarianditya Nugroho dan Setyoningsih Wibowo (2017) melakukan penelitian dengan dataset status kelulusan mahasiswa fakultas ilmu komputer UNAKI semarang menggunakan *fiture forward selection* untuk pemilihan atribut dan pengklasifikasian menggunakan algoritma *naive bayes.* Hasil penelitiannya menunjukan bahwa *forward selection* berbasis *naive bayes* lebih akurat dan efektif dalam mengklasifikasikan status kelulusan mahasiswa dengan hasil akurasi 97,14% dan termasuk dalam kategori "*excellent classification*". *Forward selection* juga memerlukan waktu komputasi yang relatif lebih pendek dibandingkan dengan *backward elimination* maupun dengan *stepwise selection.* Terdapat pula penelitian yang dilakukan oleh M. Rudi Fanani (2020) dimana penelitian yang dilakukan oleh M. Rudi Fanani mengambil topik implementasi Algoritma *Naïve Bayes* Berbasis *Forward Selection* Untuk

Memprediksi Bimbingan Konseling Siswa. Penentuan rekomendasi bimbingan dan konseling kurang akurat dikarenakan harus berdasarkan pengetahuan dan pengolahan banyaknya data yang ada sehingga diperlukan sebuah perhitungan yang menerapkan metode prediksi dengan teknik data mining. Penerapan algoritma *naïve bayes* berbasis *forward selection* terbukti berhasil meningkatkan hasil akurasi yang relatif tinggi mencapai 94,84% dengan membuang beberapa fitur yang tidak relevan. Mark A. Hall dan Geoffrey Holmes dalam penelitiannya menyajikan perbandingan patokan metode *fiture selection* dari beberapa algoritma klasifikasi. Dan menyimpulkan bahwa metode *forward selection* sangat cocok untuk *Naïve Bayes.*

Forward selection adalah pendekatan *wrapper* yang sering digunakan dalam seleksi fitur, *forward selection* mampu menghapus fitur yang tidak relevan, mengembangkan dan menambah kualitas data, serta meningkatkan performa dan akurasi model. *Forward selection* dilakukan dengan cara memasukkan *variabel independen* yang signifikan masuk ke dalam model satu persatu hingga tidak ada lagi variabel independen yang signifikan untuk dimasukkan ke dalam model (Ladha and Deepa, 2011).

Berdasarkan hal tersebut, maka penulis melakukan penelitian pada dataset NSL-KDD untuk menentukan atribut yang relevan dan meningkatkan kinerja akurasi menggunakan *fiture forward selection* dengan klasifikasinya menggunakan algoritma *naïve bayes*.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah diatas, maka rumusan masalah yang dijadikan sebagai fokus penelitian ini yaitu seleksi fitur pada dataset NSL-KDD menggunakan *forward selection* berbasis algoritma *naïve bayes* untuk menentukan fitur apa saja yang memiliki pengaruh signifikan atau relevan terhadap waktu komputasi untuk prediksi serangan atau bukan serangan.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan dari penelitian ini adalah untuk menentukan fitur-fitur dari dataset NSL-KDD yang memiliki pengaruh signifikan terhadap waktu komputasi prediksi serangan atau bukan serangan berdasarkan seleksi fitur *forward selection* berbasis algoritma *naïve bayes* dan *decision tree*.

1.4 Batasan Masalah

Untuk menjaga fokus dari penelitian ini, maka batasan masalah dari penelitian ini adalah data yang digunakan sebanyak 125.973 *record* dengan 41 atribut data yang diunduh dari *CIC (Catalonia Independence Corpus)* repository melalui *http://205.174.165.80/CICDataset/NSL-KDD/* yaitu dataset NSL-KDD deteksi intrusi sistem keamanan jaringan.

1.5 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini adalah dapat membantu meningkatkan kualitas perangkat lunak dengan memberikan peringatan dini jika ada serangan dan membantu lebih cepat dalam melakukan komputasi untuk bisa meningkatkan keamanan sistem deteksi intrusi.