BAB I
PENDAHULUAN

1.1 Latar Belakang

Random forest pertama kali dikenalkan oleh Breiman pada Tahun 2001. Dalam penelitiannya menunjukkan kelebihan *random forest* antara lain dapat menghasilkan error yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data *training* dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi *missing data* (Breiman, 2001). Penelitian sebelumnya tentang *random forest* dilakukan oleh (Sulaiman, 2011) melakukan penelitian tentang *web caching* dengan membandingkan akurasi klasifikasinya menggunakan metode CART, MARS, *random forest* dan *Tree Net*. Penelitian lain juga dilakukan (Dewi, 2001) tentang penerapan metode *random forest* dalam *driver analysis*. Pada Penelitian metode *ensemble* pada klasifikasi kemiskinan di Kabupaten Jombang diperoleh data bahwa *random forest* memberikan akurasi klasifikasi yang terbaik (Muttaqin, 2013). *Random forest* pun dimanfaatkan untuk memperkirakan cuaca dengan akurasi yang cukup baik oleh (Mujiasih, 2011) di Pusat Meteorologi Maritim dan Penerbangan BMKG.

Akan tetapi permasalahan umum yang sering terjadi pada saat mengimplementasikan random forest adalah waktu pemrosesan yang lama karena menggunakan data yang banyak dan membangun model *tree* yang banyak pula untuk membentuk *random trees* karena menggunakan *single processor*.

Oleh karena itu, diajukanlah rancangan *random forest* dengan *parallel computing*. *Parallel computing* yaitu penyatuan beberapa komputer atau *server* menjadi satu kesatuan sehingga dapat mengerjakan proses secara bersamaan ataupun secara simultan. *Parallel computing* membuat program maupun proses berjalan lebih cepat karena semakin banyak CPU yang digunakan (Baeney, 2014). *Parallel*

computing pernah digunakan untuk meningkatkan performa standar enkripsi tingkat lanjut (Vishal Pachori, 2012). Model *parallel computing* juga pernah diusulkan untuk simulasi dinamika populasi di demografi (Montañola-Sales, Onggo, Casanovas-Garcia, Cela-Espín, & Kaplan-Marcusán, 2016). Dalam penelitian tersebut membahas kekhasan simulasi dinamika populasi dengan menggunakan metode paralel diskrit dalam simulasi.

Modifikasi *Random forest* dengan *parallel computing* diharapkan akan memberikan kinerja yang lebih maksimal dan didapatkan hasil secara cepat, karena program bisa dieksekusi secara bersamaan dengan memanfaatkan semua *resource* di *processor*.

Random forest dengan *parallel computing* dalam penelitian ini akan diimplementasikan dalam bahasa R. R adalah bahasa pemrograman dan perangkat lunak untuk analisis statistika dan grafik. Bahasa R kini menjadi standar *de facto* di antara statistikawan untuk pengembangan perangkat lunak statistika, serta digunakan secara luas untuk pengembangan perangkat lunak statistika dan analisis data.

Studi kasus yang akan diambil dalam penelitian ini adalah penerapan *parallel random forest* untuk prediksi spesies bunga Iris, kualitas *wine* dan diagnosa diabetes wanita Pima Indian. Kasus-kasus tersebut diambil dari *dataset* yang telah tersedia di UCI *Machine Learning Repository*. UCI *Machine Learning Repository* adalah *website* yang berisi kumpulan *database*, teori domain, dan generator data yang digunakan oleh komunitas pembelajaran mesin (*machine learning*) untuk analisis empiris algoritma pembelajaran mesin. Arsip itu dibuat sebagai arsip ftp pada tahun 1987 oleh David Aha dan rekan-rekan mahasiswa pascasarjana di UC Irvine. (UCI, 2017)

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan di atas maka permasalahan yang akan diidentifikasi dalam penelitian ini adalah:

1. Bagaimana merancang r*andom forest* untuk *parallel computing*?
2. Bagaimana mengimplementasikan r*andom forest* untuk *parallel computing* dalam R?
3. Bagaimana meguji dan menganalisasi waktu komputasi *parallel Random forest* dalam memprediksi spesies bunga Iris, kualitas *Wine* dan kasus diabetes Pima Indian?

1.3 Batasan Masalah

Berdasarkan identifikasi masalah serta dengan mempertimbangkan banyak aspek seperti waktu, kemampuan peneliti dan kepentingan penelitian, maka permasalahan dibatasi pada hal-hal sebagai berikut:

1. Data yang digunakan adalah *dataset* yang diambil dari UCI yaitu *dataset* bunga Iris, diabetes wanita Pima Indian dan kualitas *Wine*.
2. Data yang diproses merupakan data yang tidak mengandung *noise*, sehingga tidak menangani proses *preprocessing*.
3. Perangkat dalam membuat aplikasi ini menggunakan RStudio dengan bahasa pemrograman R.
4. Pengujian *random forest* dengan *parallel computing* dilakukan menggunakan hanya sampai empat *processor* karena keterbatasan alat pengujian.

1.4 Tujuan Penelitian

Adapun tujuan yang hendak dicapai dari penelitian ini adalah:

1. Merancang r*andom forest* untuk *parallel computing*.
2. Mengimplementasikan *random forest* dengan *parallel computing* dalam bahasa R.
3. Menguji dan menganalisa waktu komputasi *parallel random forest* dalam prediksi spesies bunga Iris, kualitas *Wine* dan kasus diabetes wanita Pima Indian.

1.5 Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan banyak manfaat, antara lain sebagai berikut:

1. Memberikan alternatif kepada peneliti lain dalam pemilihan metode prediksi yang lebih cepat.
2. *Random Forest* dengan *parallel computing* dapat digunakan untuk memprediksi data yang lain.

1.6 Sistematika Pembahasan

Adapun sistematika penulisan penelitian ini dibagi kedalam lima bab, dan masing-masing bab terdiri dari beberapa sub bab, yaitu:

BAB I PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II KAJIAN PUSTAKA

Pada kajian pustaka diuraikan materi-materi yang berhubungan dengan penelitian. Materi ini mendasari penulis dalam melakukan penelitian. Materi yang disampaikan meliputi *machine learning*, algoritma *decision trees, random forest, parallel computing,* bahasa pemrograman R dan *parallel computing* di R.

BAB III METODE PENELITIAN

Bab ini menguraikan rancangan penelitian, metode yang digunakan dalam penelitian juga alat dan bahan yang digunakan untuk melakukan penelitian. Metode pengembangan perangkat lunak yang digunakan dalam penelitian ini adalah metode *waterfall*.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi uraian tentang bagaimana penelitian dilakukan, bagaimana hasil penelitian dan pembahasan terhadap hasil penelitian yang dilakukan.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari keseluruhan penelitian yang telah dilakukan, serta saran dari penulis untuk kegiatan penelitian selanjutnya terkait dengan topik yang sedang dibahas.