BAB I
PENDAHULUAN

1.1 Latar Belakang

Dewasa ini, data telah menjadi salah satu aset terpenting bagi kelangsungan hidup suatu perusahaan. Kemajuan ilmu data menghasilkan sebuah profesi baru yaitu ilmuwan data atau biasa disebut *data scientist*. Salah satu fokus para *data scientist* di dunia adalah menemukan suatu pola terstruktur yang terkandung dalam data. Untuk mengidentifikasi pola tersebut, salah satu strategi yang dapat digunakan ialah *machine learning*.

Machine learning didefiniskan sebagai seperangkat teknik dan alat yang memungkinkan komputer berpikir dengan menciptakan algoritma matematis berdasarkan akumulasi data (Landau, 2016). *Machine learning* memungkinkan komputer memiliki kemampuan untuk belajar tanpa perlu diprogram secara eksplisit (Samuel, 1959). Dengan kemampuan untuk belajar dari data, *machine learning* mengkaji berbagai algoritma yang dapat membuat keputusan atau prediksi berdasarkan proses *learning* yang telah dilakukan. Algoritma-algoritma tersebut diklasifikasikan berdasarkan hasil yang diinginkan dan dikelompokkan menjadi tipe-tipe algoritma umum yaitu *supervised*, *unsupervised*, *semi-supervised* dan *reinforcement learning* (Ayodele, 2010). Dalam *supervised learning*, salah satu cara melakukan prediksi adalah dengan proses regresi. Regresi merupakan suatu proses statistik untuk mengukur hubungan antara *dependent variable* atau *response* dan *independent variable* atau *predictor* (Fumo & Biswas, 2015). Dalam model regresi linier, dihasilkan suatu *response* nilai riil tunggal yang diprediksi berdasarkan *predictor* dengan menggunakan persamaan linier. Regresi digunakan untuk memprediksi suatu nilai dari data yang telah dikumpulkan, sebagai contoh, memprediksi jumlah penjualan berdasarkan harga produk atau memprediksi jumlah buah yang dihasilkan berdasarkan curah hujan.

Untuk membuat model prediksi pada kegiatan regresi, dapat digunakan algoritma *Gradient Descent* (GD). GD (Cauchy, 1847) merupakan algoritma

optimasi orde pertama untuk mencari nilai minimum lokal dari suatu fungsi. GD terus berkembang sepanjang waktu, bahkan hingga saat ini banyak peneliti yang mengembangkan optimasi pada GD untuk meningkatkan kinerjanya. Dalam hal banyaknya data yang diproses, *Mini-Batch Gradient Descent* (MBGD) (Cotter, dkk, 2011) memakai sebagian data untuk diproses dan *Stochastic Gradient Descent* (SGD) (Bottou, 2010) maupun *Stochastic Average Gradient Descent* (SAGD) (Schmidt, dkk, 2013) hanya memakai satu data yang dipilih secara acak untuk diproses. Dalam hal meningkatkan kecepatan *learning*, *Momentum Gradient Descent* (MGD) (Qian, 1999) dan *Accelerated Gradient Descent* (AGD) (Nesterov, 1983) memberi kecepatan tambahan pada proses *learning*. Adapun suatu optimasi GD yang bisa menyesuaikan proses *learning* yang adaptif seperti *Adagrad* (Duchi, dkk, 2011), *Adadelta* (Zeiler, 2012), *RMSprop* (Ruder, 2016) dan *Adam* (Kingma & Ba, 2015). Dengan berbagai variasi ini, GD banyak digunakan dalam berbagai implementasi seperti komputasi paralel (Zinkevich, dkk, 2010).

Penelitian ini berfokus untuk melanjutkan penelitian tentang implementasi metode berbasis *Gradient Descent* dan variasinya dalam *R Package* (Handian, dkk, 2016). Penelitian tersebut mengimplementasikan sebanyak 10 metode, yaitu *GD, MBGD, SGD, SAGD, MGD, AGD, Adagrad, Adadelta, RMSprop,* dan *Adam* ke dalam *R Package* yang telah berhasil diunggah dan dipublikasikan ke *repository Comprehensive R Archive Network* (CRAN). CRAN merupakan portal jaringan *ftp* dan *web server* yang menyimpan kode dan dokumentasi bahasa R paling mutakhir di seluruh dunia. Berdasarkan eksperimen yang dilakukan untuk memprediksi faktor kompresibilitas gas diketahui bahwa hasil prediksi dan nilai akurasi dari setiap metode pada penelitian sebelumnya berhasil didapatkan. Hasil prediksi tersebut dinilai cukup baik dengan nilai *RMSE* yang secara rata-rata sangat minimum, meskipun tidak ada satupun metode yang dapat menjawab tepat semua prediksi. Selain itu, perbandingan juga berhasil dilakukan dan menghasilkan metode terbaik yaitu *AGD*. Metode *AGD* memiliki nilai *RMSE* rata-rata terkecil dibanding metode lainnya dan proses learning-nya dinilai sangat cepat dibanding beberapa metode lainnya. Meskipun *AGD* merupakan metode terbaik secara rata-rata, tetapi nilai *RMSE* minimum dicapai oleh *SAGD* dan nilai

maksimum dicapai oleh *Adam*. Selain 10 metode yang telah diimplementasikan pada penelitian sebelumnya, masih ada beberapa metode lain yang merupakan variasi dari *GD* itu sendiri. Maka dari itu, dalam penelitian ini ditambahkan beberapa metode lain yang dapat mempercepat dalam proses konvergensi yaitu *Semi-Stochastic Gradient Descent* (S2GD) (Papamakarios, 2014), *Stochastic Variance Reduced Gradient* (SVRG) (Johnson & Zhang, 2013), *Stochastic Recursive Gradient Algorithm* (SARAH) (Nguyen, dkk, 2017), dan *Stochastic Recursive Gradient Algorithm+* (SARAH+) (Nguyen, dkk, 2017).

Bahasa R dipilih karena digunakan khusus untuk menangani kasus data analisis dan statistik yang memanfaatkan data yang besar (Ihaka & Gentleman, 1996). R merupakan proyek berlisensi GNU GPL-2 yang bisa digunakan, didistribusikan maupun dikembangkan secara gratis dibawah lisensi tersebut. R memiliki suatu portal jaringan atau *repository* resmi bernama *Comprehensive R Archive Network* (CRAN) yang bisa diakses untuk mengunduh *intepreter* dan *package* lain, dan mengunggah *package* pada *repository* tersebut.

Untuk menguji dan mengetahui hasil dari penelitian ini diperlukan suatu *dataset*. Data yang digunakan adalah data densitas volume gas CO_2 (Kennedy, 1954) untuk memprediksi nilai faktor kompresibilitas gas. Data tersebut memiliki jumlah sebanyak 2.110 baris, memiliki rentang parameter tekanan dari 500 sampai 1.400 *bars*, dan rentang suhu dari 0 sampai 1.000 derajat *celcius*. Dengan rentang tersebut, data ini dapat digunakan untuk memprediksi nilai faktor kompresibilitas gas CO_2 dengan skala parameter yang lengkap pada tekanan dan suhu yang ditentukan. Dengan menggunakan hasil penelitian sebelumnya, dapat diketahui bagaimana perbandingan antara implementasi variasi *Gradient Descent* pada penelitian sebelumnya dengan variasi *Gradient Descent* pada penelitian ini.

1.2 Rumusan Masalah

Permasalahan yang bisa didapatkan dari latar belakang penelitian di atas yaitu:

1. Bagaimana mengembangkan *R package* dengan mengimplementasikan metode *SVRG, SSGD, SARAH* dan *SARAH+*?

2. Berapa tingkat akurasi yang dihasilkan oleh metode *SVRG*, *SSGD*, *SARAH*, dan *SARAH+* pada kasus perhitungan faktor kompresibilitas gas CO_2 ?
3. Bagaimana membandingkan metode berbasis GD yang paling baik berdasarkan nilai akurasi galat terkecil dan waktu eksekusi pembangunan model?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai dari penelitian ini yaitu:

1. Mengembangkan *R package* yang mengimplementasikan beberapa algoritma berbasis *gradient descent* yaitu *SVRG, SSGD, SARAH, SARAH+* dan mempublikasikannya ke *CRAN*.
2. Mengetahui hasil prediksi dan tingkat akurasi yang dihasilkan oleh metode *SVRG, SSGD, SARAH* dan *SARAH+* pada kasus perhitungan faktor kompresibilitas gas CO2.
3. Analisis perbandingan metode GD berdasarkan nilai akurasi galat terkecil menggunakan *root-mean-square-error (RMSE)* dan waktu eksekusi pembangunan model.

1.4 Batasan Masalah

Batasan masalah yang terdapat pada penelitian ini yaitu:

1. Data yang digunakan merupakan data yang didapat dari hasil penelitian yang dilakukan (Kennedy, 1954).
2. Data yang digunakan hanya data tentang perilaku gas CO_2 pada suhu dan tekanan tertentu.

1.5 Manfaat Penelitian

Dengan adanya penelitian ini, diharapkan dapat memberikan manfaat sebagai berikut.

1. Membantu praktisi, peneliti ataupun pengguna *R package* untuk memprediksi masalah analisis regresi menggunakan 14 metode, yaitu *GD, MBGD, SGD, SAGD, MGD, AGD, Adagrad, Adadelta, RMSprop, Adam, SVRG, SSGD, SARAH* dan *SARAH+*.

2. Dengan dilakukannya penelitian ini, maka akan didapatkan hasil perbandingan antara metode *GD, MBGD, SGD, SAGD, MGD, AGD, Adagrad, Adadelta, RMSprop, Adam, SVRG, SSGD, SARAH* dan *SARAH+* untuk memprediksi nilai faktor kompresibilitas CO_2 dan akurasi dari masing-masing metode, sehingga bisa ditentukan metode terbaik untuk menyelesaikan permalasalahan.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut.

BAB I PENDAHULUAN

Bab ini berisi latar belakang penelitian yang berisi pengantar *machine learning* yang menjadi landasan dalam pengembangan *R Package* yang menangani kasus regresi menggunakan *gradient descent* dan penjelasan mengenai penelitian yang dilakukan sebelumnya. Selanjutnya bab ini berisi rumusan masalah penelitian, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II KAJIAN PUSTAKA

Pada kajian pustaka akan diuraikan materi-materi yang berhubungan dengan penelitian. Bab ini berisi teori dan konsep terkait dalam penelitian seperti penjelasan tentang *machine learning*, regresi, *gradient descent*, bahasa pemrograman R, dan faktor kompresibilitas gas.

BAB III METODOLOGI PENELITIAN

Bab ini berisi langkah-langkah penelitian yang diilustrasikan dengan skema desain penelitian, metode penelitian yang terdiri dari studi literatur dan proses pengembangan perangkat lunak, dan alat maupun bahan penelitian yang digunakan.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi hasil pengumpulan data, perancangan *R Package*, perancangan eksperimen dan simulasinya, pembahasan perbandingan hasil eksperimen dan perbandingan nilai RMSE untuk semua metode dengan simulasi *R package*.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan yang merupakan jawaban dari masalah pada penelitian, serta berisi saran yang dapat menjadi rujukan untuk penelitian selanjutnya.