

BAB I

PENDAHULUAN

1.1 Latar Belakang

Menemukan suatu pola terstruktur yang terkandung dalam suatu data telah menjadi fokus para *data scientist* di dunia dan *data science* merupakan suatu *trend* teknologi yang populer saat ini (IEEECS, 2016). Salah satu strategi untuk mengidentifikasi pola dalam data adalah dengan menggunakan *machine learning*. *Machine learning* merupakan cabang ilmu komputer yang berfokus pada algoritma yang bisa belajar dari data (Kohavi & Provost, 1998) dan diimplementasikan pada mesin atau komputer untuk belajar memecahkan masalah tanpa terprogram dengan jelas (Mitchell, 1997), sehingga *machine learning* akan membuat keputusan berdasarkan data yang telah dilakukan *learning*. Terdapat banyak jenis algoritma yang bisa melakukan *learning* terhadap data dan algoritma tersebut dikelompokkan menjadi *supervised*, *unsupervised*, *semi-supervised* dan *reinforcement learning* (Ayodele, 2010). Salah satu kegiatan yang termasuk dalam *supervised learning* adalah regresi. Regresi merupakan suatu proses statistik untuk mengukur hubungan antara *dependent variable* atau *response* dan *independent variable* atau *predictor* (Fumo & Biswas, 2015). Model dari regresi adalah sebuah persamaan linier yang menghasilkan suatu *response* nilai riil tunggal yang didapatkan berdasarkan *independent variable* atau *predictor*. Regresi biasanya digunakan untuk memprediksi suatu nilai keluaran dari data nyata yang dikumpulkan seperti menentukan tinggi anak perempuan berdasarkan tinggi ibu mereka dan menentukan panjang ikan berdasarkan umur mereka (Weisberg, 2014).

Gradient descent (GD) merupakan algoritma yang bisa digunakan untuk membuat model prediksi pada kegiatan regresi. *GD* (Cauchy, 1847) merupakan algoritma optimasi orde pertama untuk mencari nilai minimum lokal dari suatu fungsi. *GD* terus berkembang sepanjang waktu, sehingga banyak peneliti mengembangkan optimasi pada *GD* untuk meningkatkan kinerjanya. Dalam hal

banyaknya data yang diproses, *Mini-Batch Gradient Descent (MBGD)* memakai sebagian data untuk diproses (Cotter, Shamir, Srebro, & Sridharan, 2011) dan

Stochastic Gradient Descent (SGD) maupun *Stochastic Average Gradient Descent (SAGD)* hanya memakai satu data yang dipilih secara acak untuk diproses (Bottou, 2010) (Schmidt, Le Roux, & Bach, 2013). Dalam hal meningkatkan kecepatan *learning*, *Momentum Gradient Descent (MGD)* dan *Accelerated Gradient Descent (AGD)* memberi kecepatan tambahan pada proses *learning* (Qian, 1999) (Nesterov, 1983). Adapun suatu optimasi *GD* yang bisa menyesuaikan proses *learning* yang adaptif seperti *Adagrad* (Duchi, Hazan, & Singer, 2011), *Adadelata* (Zeiler, 2012), *RMSprop* dan *Adam* (Kingma & Ba, 2015). Dengan berbagai variasi ini, *GD* banyak digunakan dalam berbagai implementasi seperti komputasi paralel (Zinkevich, Weimer, Smola, & Li, 2010).

Penelitian ini berfokus untuk melanjutkan penelitian tentang implementasi metode berbasis *GD* untuk memprediksi faktor kompresibilitas gas menggunakan bahasa *R* (Riza, Nasrulloh, Junaeti, Zain, & Nandiyanto, 2016). Namun dalam penelitian itu, metode yang diimplementasikan hanya ada empat yaitu *GD*, *MBGD*, *SGD*, dan *SAGD*. Metode yang diimplementasikan tersebut hanya metode yang memanfaatkan perubahan jumlah data. Maka dari itu, penelitian ini akan melanjutkan penelitian tersebut dengan menambahkan metode lain yaitu *MGD*, *AGD*, *Adagrad*, *Adadelata*, *RMSprop*, dan *Adam*. Metode-metode tersebut akan dibandingkan dan dipilih yang terbaik berdasarkan tingkat akurasi pada uji kasus yang diajukan. *R* dipilih karena digunakan khusus untuk menangani kasus data analisis dan statistik yang memanfaatkan data yang besar (Ihaka & Gentleman, 1996). *R* merupakan perangkat yang banyak digunakan untuk kasus analisis data, *data mining* dan *data science* berdasarkan survei (KDNuggets, 2016) dan melebihi perangkat lain seperti *MATLAB*, *Python*, *SAS*, dan *SPSS*. Selain itu, *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 *CRAN* yang bisa diakses untuk mengunduh *intepreter* dan *package* lain. *R Package* yang dikembangkan akan diunggah pada *repository* tersebut.

Untuk menguji *R Package* yang dikembangkan agar bisa memiliki nilai guna maupun nilai jual, suatu studi kasus yang memiliki tingkat yang tinggi

Dendi Handian, 2017
gradDescentR 2.0 : IMPLEMENTASI METODE GRADIENT DESCENT DAN VARIASINYA DALAM R PACKAGE
 Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu

diperlukan dalam pengujiannya. Penelitian ini memakai data densitas volume gas CO₂ (Kennedy, 1954) untuk memprediksi nilai faktor kompresibilitas gas. Data tersebut memiliki jumlah sebanyak 2110 baris, memiliki rentang parameter tekanan dari 500 sampai 1400 *bars* dan rentang suhu dari 0 sampai 1000 derajat *celcius*. Dengan rentang tersebut, data ini dapat digunakan untuk memprediksi nilai faktor kompresibilitas gas CO₂ dengan skala parameter yang lengkap pada tekanan dan suhu yang ditentukan.

1.2 Rumusan Masalah

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

1. Bagaimana mengembangkan *R package* dengan mengimplementasikan metode *GD*, *MBGD*, *SGD*, *SAGD*, *MGD*, *AGD*, *Adagrad*, *Adadelta*, *RMSprop* dan *Adam*?
2. Berapa tingkat akurasi yang dihasilkan oleh metode *GD*, *MBGD*, *SGD*, *SAGD*, *MGD*, *AGD*, *Adagrad*, *Adadelta*, *RMSprop* dan *Adam* pada kasus perhitungan faktor kompresibilitas gas CO₂ ?
3. Manakah metode terbaik hasil simulasi yang telah dibandingkan berdasarkan tingkat akurasi?

1.3 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 CO₂ pada suhu dan tekanan tertentu.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dari penelitian ini yaitu:

1. Mengembangkan *R package* yang mengimplementasikan beberapa algoritma berbasis *gradient descent* yaitu *GD*, *MBGD*, *SGD*, *SAGD*,

MGD, *AGD*, *Adagrad*, *Adadelta*, *RMSprop*, *Adam* dan mempublikasikannya ke *CRAN*.

2. Mengetahui hasil prediksi dan tingkat akurasi yang dihasilkan oleh metode *GD*, *MBGD*, *SGD*, *SAGD*, *MGD*, *AGD*, *Adagrad*, *Adadelta*, *RMSprop*, dan *Adam* pada kasus perhitungan faktor kompresibilitas gas CO₂.
3. Membandingkan dan memilih metode berbasis *GD* paling baik berdasarkan nilai akurasi galat terkecil menggunakan *root-mean-square-error (RMSE)* dan waktu eksekusi pembangunan model.

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 10 metode, yaitu *GD*, *MBGD*, *SGD*, *SAGD*, *MGD*, *AGD*, *Adagrad*, *Adadelta*, *RMSprop* dan *Adam*.
2. Dengan dilakukannya penelitian ini, maka akan didapatkan hasil perbandingan antara metode *GD*, *MBGD*, *SGD*, *SAGD*, *MGD*, *AGD*, *Adagrad*, *Adadelta*, *RMSprop* dan *Adam* untuk memprediksi nilai faktor kompresibilitas CO₂ dan akurasi dari masing-masing metode, sehingga bisa ditentukan metode terbaik untuk menyelesaikan permasalahan.

1.6 Sistematika Penelitian

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

Bab ini berisi teori dan konsep terkait dalam penelitian seperti penjelasan tentang *machine learning*, analisis regresi linier, *gradient descent*, bahasa pemrograman *R* dan faktor kompresibilitas gas.

BAB III METODOLOGI

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* dan perbandingan *tools* yang memiliki kegunaan yang sama dengan *R Package* yang dikembangkan pada penelitian ini.

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.