Analisis dan Komparasi Metode Naive Bayes dan Logistic Regression dengan Seleksi Variabel Berbasis Genetic Algorithm untuk Prediksi *Software Defect*

TESIS

Diajukan Sebagai Salah Satu Syarat Untuk Menyelesaikan
Program Strata Dua (S2) Magister Komputer

OLEH :

DICKY SURYA DWI PUTRA

3712101182

PROGRAM STUDI TEKNIK INFORMATIKA

PROGRAM PASCA SARJANA (S2) MAGISTER KOMPUTER

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER ERESHA

JAKARTA

2012

PERSETUJUAN TESIS

Nama	:	Dicky Surya Dwi Putra
NPM	:	3712101182
Konsentrasi	:	Software Engineering
Judul tesis	:	Analisis dan Komparasi Metode Naive Bayes dan Logistic Regression dengan Seleksi Variabel Berbasis Genetic Algorithm untuk Prediksi *Software Defect*

Telah disetujui untuk diseminarkan pada Sidang Tesis pada Program Pasca Sarjana (S2) Magister Komputer, Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer Eresha.

Jakarta, September 2012

Pembimbing Utama | Pembimbing Pendamping

(Romi Satria Wahono, B. Eng., M. Eng.) (Dr. Rufman Iman Akbar E., SE, MM, M.Kom.)

Mengetahui :

Direktur Progam Pasca Sarjana

(Dr. Rufman Iman Akbar E., SE, MM, M.Kom.)

PERNYATAAN KEASLIAN TESIS

Nama	: Dicky Surya Dwi Putra
NPM	: 371 210 1182
Konsentrasi	: *Software Engineering*
Judul tesis	: Analisis dan Komparasi Metode Naive Bayes dan Logistic Regression dengan Seleksi Variabel Berbasis Genetic Algorithm untuk Prediksi *Software Defect*

Dengan ini saya menyatakan bahwa dalam Tesis ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan Strata 2 di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Jakarta, September 2012

(Dicky Surya Dwi Putra)

Dicky Surya Dwi Putra, 3712101182

Analisis dan Komparasi Metode Naive Bayes dan Logistic Regression dengan Seleksi Variabel Berbasis Genetic Algorithm untuk Prediksi *Software Defect*; dibawah bimbingan Romi Satria Wahono, B.Eng., M.Eng dan Dr. Rufman Iman Akbar E., MM, M.Kom.

103 + xiii hal / 43 tabel / 38 gambar / 1 lampiran / 30 pustaka (1997 – 2011)

ABSTRAK

Permintaan akan *software* yang berkualitas sangat berkembang pesat pada dunia global saat ini. Maka dari itu diperlukan *software-software* yang tidak memiliki kesalahan / error (*defect*). Kebanyakan kesalahan disebabkan oleh kesalahan manusia. Banyak peneliti lain melakukan pembuatan model untuk pengembangan *software*, yang nantinya akan digunakan untuk para pengembang agar mengikuti model yang diusulkan saat membuat *software*. Model yang diusulkan diupayakan agar dapat menghasilkan *software* tanpa *defect*.Model yang saat ini paling baik dalam melakukan prediksi *software defect* adalah Naive Bayes dan Logistic Regression. Ada juga penelitian lain yang dilakukan untuk meningkatkan hasil akurasi yang ada dapat dilakukan pemilihan variabel yang digunakan. Genetic Algoritma akan diterapkan untuk pemilihan variabel pada metode Naive Bayes dan Logistic Regression. Setelah itu akan dilakukan uji statistik dengan *T-Test*, model mana yang paling baik untuk menghasilkan tingkat akurasi paling tinggi dalam prediksi *software defect*. Hasil akurasi yang diperoleh membuktikan bahwa Logistic Regression dengan Genetic Algorithm lebih tinggi dibandingkan Naive Bayes dengan Genetic Algorithm dengan persentase akurasi 86,47%, tetapi pada hasil performa yang diperoleh membuktikan bahwa Naive Bayes dengan Genetic Algorithm lebih unggul dibandingkan Logistic Regression dengan performa 34,77. Dan, hasil uji statistic terhadap kedua model tersebut meyimpulkan bahwa Logistic Regression dengan Genetic Algorithm lebih baik untuk melakukan prediksi *software defect*.

Kata kunci: *Software Defect,Naive Bayes,Logistic Regression, Genetic Algorithm*

Dicky Surya Dwi Putra, 3712101182

Analisys and Comparation Naive Bayes and Logistic Regression Method with Variabel Selection Base On Genetic Algorithm for Software Defect Prediction; under the guidances of Romi Satria Wahono, B.Eng., M.Eng and Dr. Rufman Iman Akbar E., MM, M.Kom.

103 + xiii pages / 43 tables / 38 images / 1 enclosure / 30 references (1997 – 2011)

ABSTRACT

The demand for quality software has been tremendously growing globaly in the world. Therefore, need software without fault / error (*defect*). Mostly error come from human error. A lot of researchers design model for software development, wisely use for developer develop software with model references. The model hope can create software without defect. Nowadays, good model used to software defect prediction are Naive Bayes and Logistic Regression. And another researcher use variabel selection to increase accuracy at prediction. Genetic Algorithm will use to variabel selection at Naive Bayes and Logistic Regression method. Finally, T-Test used to obtain statistic result, which model can concluded higher accuracy at software defect prediction. The result of accuracy show Logistic Regression with Genetic Algorithm higher than Naive Bayes with Genetic Algorithm with accuracy 86.47%, but the result of performance show Naive Bayes with Genetic Algorithm more good than Logistic Regression with Genetic Algorithm with value 34,77. And, the result of statisctic from two models concluse Logistic Regression with Genetic Algorithm more capable to software defect prediction.

Kata kunci: *Software Defect,Naive Bayes,Logistic Regression, Genetic Algorithm*

KATA PENGANTAR

Dengan memanjatkan puji syukur kehadiran Tuhan Yang Maha Esa yang telah melimpahkan segala rahmat dan hidayahnya kepada penulis, sehingga tersusunlah tesis yang berjudul **"**Analisis dan Komparasi Metode Naive Bayes dan Logistic Regression dengan Seleksi Variabel Berbasis Genetic Algorithm untuk Prediksi *Software Defect*". Tesis tersebut melengkapi salah satu persyaratan yang diajukan dalam rangka menempuh ujian akhir untuk memperoleh gelar Magister Komputer (M.Kom.) pada Program Pasca Sarjana (S2), Program Studi Teknik Informatika di Sekolah Tinggi Manajemen Informatika dan Komputer Eresha

Penulis menghaturkan penghargaan dan ucapan terima kasih yang sebesar - besarnya kepada yang terhormat :

1. Bapak Ir. Damsiruddin Siregar, MMT, selaku Ketua STMIK Eresha
2. Bapak Dr. Rufman Iman Akbar E., SE, MM, M.Kom, selaku Direktur Pasca Sarjana STMIK Eresha yang sekaligus menjadi dosen pembimbing pendamping yang telah banyak membantu dalam penulisan thesis ini.
3. Bapak Didik Setiyadi, M.Kom, selaku Puket I STMIK Eresha
4. Bapak Bobby Reza, S.Kom., MM, selaku Puket III STMIK Eresha
5. Bapak Romi Satria Wahono, B.Eng., M.Eng, selaku dosen pembimbing utama yang telah banyak membantu memberikan ide dan saran-saran dalam penulisan tesis ini.
6. Bapak/Ibu dosen STMIK Eresha yang telah memberikan ilmunya.
7. Rekan-rekan perjuangan Indah, Susanto H, Desiyanna, Rino, dan Edy dalam mengerjakan thesis ini hingga perjuangan terkahir.
8. Rekan-rekan mahasiswa angkatan 37 (pusat) STMIK Eresha yang telah berjuang bersama dalam perkuliahan

Akhir kata mohon maaf atas kekeliruan dan kesalahan yang ada dalam tesis ini, baik yang disengaja maupun tidak disengaja dan berharap semoga tesis ini dapat memberikan manfaat bagi khasanah pengetahuan teknologi informasi di Indonesia.

Penulis

DAFTAR ISI

Hal.

DAFTAR TABEL

Hal.

DAFTAR GAMBAR

Hal.

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Sebagian besar software yang dibuat dan dikembangkan oleh para developer dibuat untuk keperluan sebuah perusahaan atau institusi. Software tersebut dibuat dengan tujuan agar membantu perusahaan atau institusi agar proses kinerja dapat berjalan dengan cepat, tepat, efektif serta efisien. Seperti yang ditulis(Fenton & Pfleeger, 1997) dan dikutip (Gayatrri, Nickolas, & Reddy, 2010, hal. WCECS 2010) bahwa, atribut-atribut pada kualitas software adalah reliability, functionality, fault proneness, reusability, dan comprehensibility. Fault proneness adalah probabilitas kesalahan yang terdapat pada software (Pai & Dugan, 2007, hal. 675). Fault proneness merupakan salah satu atribut dalam menilai software yang menjadi perhatian karena dapat menjadi alat ukut kesalahan / error (defect). Perusahaan atau institusi pasti memerlukan software yang tanpa kesalahan agar investasinya di bidang teknologi informasi tidak menjadi sia-sia. Jumlah defect pada software dapat digunakan untuk mengukur kualitas pengembang software dan mengatur proses software (Song, *et al*, 2006, hal. 69). Iker Gondra menyatakan bahwa permintaan pada software yang berkualitas sangat berkembang pesat beberapa tahun ini (Gayatrri, Nickolas, & Reddy, 2010, hal. WCECS 2010). Untuk itu diperlukan software-software yang tidak ada defect, setidaknya dengan intensitas kesalahan yang sangat sedikit.

Prediksi defect software menjadi topik penelitian yang penting pada lingkup Software Engineering selama lebih dari 30 tahun (Song, *et al*, 2011, hal. 1). Pengembangan software yang besar dan sistem yang rumit merupakan tantangan (Lessmann, *et al*, 2008, hal. 485). Berbagai cara dilakukan pada saat pembuatan software agar software dapat berjalan sesuai proses dan berfungsi sesuai dengan fungsinya. Tapi dari cara-cara itu, tidak menutup kemungkinan adanya kesalahan pada software yang telah dibuat. Salah satu caranya dengan melakukan prediksi software pada saat proses pengembangan software. Prediksi defect software diharapkan dapat mengurangi adanya kesalahan pada software.

Banyak peneliti melakukan eksperimen dan penelitian pada prediksi defect software (Hall, *et al*, 2011, hal. 1), (Lessmann, *et al*, 2008, hal. 485), (Song, *et al*,

2011, hal. 1),(Gayatrri, Nickolas, & Reddy, 2010, hal. WCECS 2010),(Menzies, Greenwald, & Frank, 2007, hal. 2),(Pai & Dugan, 2007, hal. 675), dan yang lainnya. Awalnya beberapa peneliti melakukan prediksi defect software dengan menggunakan 1 metode data mining dan mereka menghasilkan kesimpulan yang cukup akurat. Seperti yang dilakukan oleh Ganesh J. Pai dengan menggunakan metode Bayesian Network, Joao A. Duares dengan menggunakan teknik G-SWIFT, Tang W. Dengan metode k-means, Khoshgoftaar dengan Classification Trees, dan yang lainnya. Setelah itu, peneliti lain melakukan eksperimen dengan menggunakan beberapa metode pada beberapa dataset, hasilnya sangat tidak relevan dengan peneliti dengan 1 metode. Eksperimen dengan banyak metode pada beberapa dataset menghasilkan prediksi yang beragam. Tidak ada metode yang paling akurat pada semua dataset (Lessmann, *et al*, 2008, hal. 486). Begitu pula pernyataan Qinbao Song, bahwa pemilihan metode yang tepat pada dataset yang berbeda-beda, proses evaluasi, dan proses penentuan keputusan sangatlah penting (Song, *et al*, 2011, hal. 2). Maka dari itu belum ada hasil akurasi yang sangat tepat pada prediksi defect software.

Eksperimen yang dilakukan oleh pada peneliti yang menggunakan beberapa metode pada beberapa dataset menggunakan model dan teknik yang berbeda-beda. Ada yang melakukan seleksi atribut ataupun klasifikasi atribut, kemudian melakukan komparasi terhadap metode yang digunakan, dan kemudian evaluasi hasilnya. Model penelitian yang dilakukan oleh Tim Menzies dan rekannya pada tahun 2007 dengan melakukan seleksi atribut dengan InfoGain, kemudian melakukan komparasi ROC Curve pada metode Naive Bayes, OneR dan J48, dan akhirnya menilai performa dengan Quartile Charts of Performance Deltas (Menzies, Greenwald, & Frank, 2007, hal. 9). Model penelitian yang dilakukan oleh Stefan Lessmann dan rekannya pada tahun 2008 dengan melakukan klasifikasi atribut dengan Binary Classifier, kemudian melakukan komparasi AUC pada 22 metode (Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, Naive Bayes, Bayesian Networks, Least-Angle Regression, Relevance Vector Machine, k-Nearest Neighbor, K-Star, Multi-Layer Perceptron 1, Multi-Layer Perceptron 2, Radial Basis Function Network, Support Vector Machine, Lagrangian SVM, Least Squares SVM, Linear Programming, Voted Perceptron, C4.5 Decission Tree, Classification and Regression Tree, Alternating

Decission Tree, Random Forest, dan Logistic Model Tree), dan akhirnya melakukan benchmarking dengan diagram Demsar (Lessmann, *et al*, 2008, hal. 491). Model penelitian yang dilakukan oleh Gayatri dan rekannya pada tahun 2010 dengan melakukan klasifikasi atribut dengan Decission Tree, kemudian melakukan seleksi atribut dengan Decission Tree Induction, kemudian melakukan komparasi AUC pada 18 metode (J48, BFTree, Random Forest, Classification and Regression Tree, Naive Bayes, Logistic Regression, Multi Layer Perceptron, Radial Basis Function, SMO, IBK, K-Star, CvR, Ensemble, VFI, DTNB, JRip, PART, dan Conjuctive Part), dan akhirnya pengukuran performa dengan Mean Absolute Error (MAE) dan Root Mean Squared Error (RMSE) (Gayatrri, Nickolas, & Reddy, 2010, hal. WCECS 2010). Model penelitian yang dilakukan oleh Qinbao Song dan rekannya pada tahun 2011 dengan melakukan seleksi atribut dengan Decission Tree, kemudian melakukan komparasi AUC pada 3 metode (Naive Bayes, OneR, J48), dan akhirnya melakukan pengukuran performa dengan Wilcoxon Signed-Rank Test (Song, *et al*, 2011, hal. 9). Model penelitian yang dilakukan oleh Tracy Hall dan rekannya pada tahun 2011 dengan feature selection kemudian melakukan komparasi AUC pada 12 metode (C4.5, Decission Tree, LOC, Linear Regression, Logistic Regression, Naive Bayes, Neural Network, OSB, Random Forest, SBPH, Tree Disc, dan Support Vector Machine), dan akhirnya melakukan pengukuran performa dengan F-Measure, Precision, dan Recall (Hall, Beecham, Bowes, D., & Counsell, 2011, hal. 2). Model yang sudah digunakan oleh para peneliti tersebut dapat menjadi pembelajaran bagi peneliti yang lain.

Pemilihan variabel atau yang sering didengar dengan Feature Selection sering digunakan dalam lingkup data mining yang berfungsi untuk meningkatkan akurasi. Salah satu metode yang digunakan adalah metode Genetic Algorithm. Genetic Algorithm proses menggabungkan metodologi evaluasi yang secara heuristik/natural (Anbarasi, Anupriya, & Iyengar, 2010, p. 5372). Tujuan dari pemilihan variabel adalah mengidentifikasi variabel yang sama pentingnya dalam dataset, kemudian membuang variabel lain yang nilainya tidak relevan dan berlebihan (Maimon & Rokach, 2010, p. 84). Pemilihan variabel dapat mengurangi pemakaian data, hal ini memungkinkan lebih efektif dalam operasi yang lebih cepat dari beberapa algoritma data mining. Dengan adanya pemilihan

variabel membuat metode lebih cepat dan lebih efektif karena tidak menggunakan variabel yang tidak relevan dan berlebihan. Terlebih lagi hasil dengan pemilihan variabel memungkinkan dapat meningkatkan akurasi dalam pengklasifikasian data.

Dari semua model yang telah diteliti, belum ada model yang menghasilkan akurasi yang sangat tepat pada prediksi *Software Defect*. Meskipun dengan proses model yang berbeda, tetap saja belum ada model yang dapat menjadi acuan untuk prediksi defect software. Dataset yang mereka gunakan juga belum sepenuhnya sama. Penelitian oleh Tim Menzies dan rekannya pada tahun 2007 menggunakan NASA dataset dari Promise Repository, sedangkan penelitian oleh Stefan Lessmann dan rekannya, Khoshgoftaar dan rekannya, Qinbao Song dan rekannya menggunakan NASA dataset dari MDP Repository (Shepperd, *et al*, 2011, hal. 1). Penggunaan dataset yang berasal dari repository yang berbeda juga dapat menghasilkan akurasi yang berbeda.

Dari semua hasil penelitian yang sudah dilakukan untuk menghasilkan model yang paling tepat untuk prediksi defect software, penelitian Tracy Hall dan Qinbao Song yang menghasilkan akurasi lebih baik dibandingkan yang lainnya. Model Tracy Hall dengan metode Naive Bayes dan Logistic Regression serta model Qinbao Song dengan metode Naive Bayes, keduanya mendapatkan penggunaan metode yang berbeda pada hasil mereka yang dianggap mendominasi prediksi defect software masing-masing. Kedua metode tersebut menjadi panutan dalam penelitian yang akan dilakukan, tentunya akan ditambahkan optimisasi dan penggunaan dataset yang sudah dibersihkan. Penelitian ini akan menggunakan metode Naive Bayes dan Logistic Regression yang akan dilakukan pemilihan variabel dengan Genetic Algorithm (GA) pada keduanya dan juga penggunaan NASA dataset dari repository yang telah dibersihkan oleh Martin Shepperd.

1.2 Permasalahan Penelitian

1.2.1 Identifikasi Masalah

Dalam penelitian ini menghasilkan beberapa masalah:

1. Model dalam prediksi defect software yang sudah dilakukan penelitian lain dengan metode Naive Bayes dan Logistic Regression menghasilkan

hasil yang baik. Naive Bayes dan Logistic Regression berguna dalam pembetukan model baru untuk melakukan prediksi tetapi keduanya masih memiliki kekurangan pada jumlah variabel yang cukup banyak. Untuk itu, peneliti mengusulkan model baru dengan menggunakan metode tersebut yang kemudian dioptimisasi dengan Genetic Algorithm.

2. Prediksi pada dataset dengan jumlah variabel yang banyak perlu dilakukan pemilihan variabel. Genetic Algorithm digunakan pada variabel di dataset untuk menentukan pemilihan variabel. Variabel yang digunakan pada eksperimen ini adalah variabel yang sudah ditentukan dengan Genetic Algorithm.

1.2.2 Ruang Lingkup Masalah

Penelitian ini dilakukan dengan proses eksperimen aplikasi data mining pada dataset yang mencakup pada *software defect* agar hasil prediksi dapat lebih efisien, akurat dan efektif. Metode-metode yang digunakan adalah metode Naive Bayes dan Logistic Regression, serta yang digunakan untuk hasil yang optimal dan efisien dengan metode Genetic Algorithm.

1.2.3 Rumusan Masalah

Pertanyaan penelitian (*research questions*) pada penelitian ini adalah: Algoritma manakah yang paling akurat dan performa terbaik antara metode Logistic Regression dan Naive Bayes yang dilakukan pemilihan variabel dengan Genetic Algorithm pada NASA dataset dapat menghasilkan model pada prediksi *Software Defect*?

1.3 Tujuan dan Manfaat Penelitian

1.3.1 Tujuan Penelitian

Penelitian ini bertujuan untuk melakukan analisis dan komparasi metode Logistic Regression dan Naive Bayes yang telah dioptimisasi dengan Genetic Algorithm untuk pemilihan variabel pada NASA dataset yang telah dibersihkan untuk menghasilkan model dengan hasil lebih akurat dan performa yang baik pada prediksi defect software.

Untuk menentukan model dengan metode mana yang paling akurat dalam melakukan prediksi *Software Defect* adalah dengan menggunakan uji statistik *T*-

Test. T-Test digunakan saat hasil dari rata-rata dalam melakukan prediksi sudah didapat, kemudian dibandingkan pada setiap dataset dan metode yang digunakan.

1.3.2 Manfaat Penelitian

Penelitian ini dapat bermanfaat. Yakni:

1. Bagi pengembang, untuk menghasilkan software yang berkualitas. Dengan adanya model dan prediksi ini, software yang dihasilkan tanpa defect ataupun jumlah defectnya yang sedikit.
2. Bagi perusahaan, dapat melakukan prediksi defect software untuk menghasilkan akurasi yang lebih akurat dibandingkan penelitian lain yang sebelumnya sudah dilakukan. Dengan meningkatnya akurasi, penelitian selanjutnya dapat lebih akurat lagi.
3. Akurasi yang dihasilkan dapat bermanfaat bagi lingkup Software Engineering. Perkembangan minat penggunaan software dan komputerisasi dapat makin berkembang dengan adanya software-software yang berkualitas, tentunya dengan tingkat defect yang sedikit.

Penelitian ini memberikan kontribusi sebagai berikut:

1. Menerapkan optimisasi dengan Genetic Algorithm untuk pemilihan variabel pada metode Logistic Regression dan Naive Bayes untuk menghasilkan akurasi lebih tinggi.

2. Melakukan analisis dan komparasi Logistic Regression dengan Naive Bayes yang dioptimisasi dengan Genetic Algorithm.

1.4 Sistematika Penulisan

Sistematika penulisan pada peneltian ini adalah sebagai berikut :

Bab I Pendahuluan

Bab I berisi penjelasan tentang latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian dan kontribusi serta sistematika penulisan.

Bab II Landasan Teori dan Kerangka Pemikiran