SKRIPSI

APLIKASI METODE *RANDOM FORESTS* DALAM MEMPREDIKSI TINGKAT MORTALITA ASURANSI GRUP CACAT JANGKA PANJANG

Lucas Mangaratua

NPM: 2016710039

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2020

FINAL PROJECT

APPLICATION OF RANDOM FORESTS IN PREDICTING GROUP LONG TERM DISABILITY

Lucas Mangaratua

NPM: 2016710039

DEPARTMENT OF MATHEMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2020

LEMBAR PENGESAHAN

APLIKASI METODE *RANDOM FORESTS* DALAM MEMPREDIKSI TINGKAT MORTALITA ASURANSI GRUP CACAT JANGKA PANJANG

Lucas Mangaratua

NPM: 2016710039

Bandung, 29 Juli 2020

Menyetujui,

Pembimbing 1	Pembimbing 2
Dr. Julius Dharma Lesmono	**Felivia Kusnadi, M.Act.Sc.**
Ketua Tim Penguji	**Anggota Tim Penguji**
Farah Kristiani, Ph.D.	**Taufik Limansyah, M.T.**

Mengetahui,

Ketua Program Studi

Dr. Erwinna Chendra

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

APLIKASI METODE *RANDOM FORESTS* DALAM MEMPREDIKSI TINGKAT MORTALITA ASURANSI GRUP CACAT JANGKA PANJANG

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 29 Juli 2020

Meterai
Rp. 6000

Lucas Mangaratua
NPM: 2016710039

ABSTRAK

Pada skripsi ini akan dibahas mengenai prediksi Tingkat Mortalita dengan data Asuransi Grup Cacat Jangka Panjang milik *Society of Actuaries* (SOA) tahun 2017 menggunakan model aplikasi *Random Forests*. Metodologi yang dilakukan pada skripsi ini adalah membentuk model *Random Forests* lalu akan dibandingkan hasil tersebut dengan model *Decision Tree* pada laporan Asuransi Grup Cacat Jangka Panjang tahun 2017. Model *Random Forests* dibangun menggunakan program *Machine Learning* R, yaitu dengan membuat 500 himpunan bagian data dari data *training* (yang disebut data *bootstrapped*), sehingga masing-masing himpunan bagian dapat dibentuk *Decision Tree*-nya. Lalu dengan data *out-of-bag*, akan dibentuk plot *Random Forests* untuk melihat pergerakan galat terhadap masing-masing *tree*. Kemudian data *test* akan dimasukkan ke dalam 500 *tree*, sehingga menghasilkan nilai respon. Nilai prediksi tingkat mortalita merupakan rata-rata dari seluruh respon yang dihasilkan. Berdasarkan model yang diperoleh kemudian akan dihitung juga tingkat kecocokan model menggunakan *Mean Squared Errors*, *variable importance*, dan pengaruh parameter terhadap model. Berdasarkan hasil prediksi dan nilai *Mean Squared Errors* yang diperoleh, model *Random Forests* tidak memberikan hasil yang lebih baik untuk data Grup Cacat Jangka Panjang tahun 2017. Hal ini disebabkan karena data Grup Cacat Jangka Panjang tahun 2017 memiliki sangat banyak entri dengan jumlah prediktor yang tidak terlalu banyak.

Kata-kata kunci: Tingkat Mortalita, Asuransi Grup Cacat Jangka Panjang, *Decision Tree*, *Random Forests*, *Variable Importance*.

ABSTRACT

This final project will discuss the prediction of Mortality Rates with the Society of Actuaries (SOA) Group Long Term Disability 2017 data using the Random Forests application model. The method used in this final project is to form a Random Forests model and compare the result with the Decision Tree model on 2017 Group Long Term Disability report. The Random Forests model was developed using the Machine Learning R program, by creating 500 subsamples of data from the training data (called bootstrapped data), so that each subsamples of Decision Tree can be formed. Then with out-of-bag data, Random Forests plot will be formed to see the errors towards each tree. After that, the test data is entered into 500 trees, so that it produces a response value. The mortality rate prediction value is the average of all responses generated. Based on the model obtained, the fit of the model will be calculated using Mean Squared Errors, variables importance, and the influence of parameters on the model. Based on the predicted results and Mean Squared Errors obtained, the Random Forests model did not generate a better result for the 2017 Group Long Term Disability. This is because the 2017 Group Long Term Disability has a lot of entry with not too many predictors.

Keywords: Mortality Rate, Group Long Term Disability Insurance, Decision Tree, Random Forests, Variable Importance.

To my parents with love,
who never gave up on me.
Thank you for your unbiased support.

KATA PENGANTAR

Puji dan syukur kepada Tuhan Yesus Kristus atas segala berkat dan penyertaan-Nya yang melimpah, sehingga skripsi ini dapat diselesaikan. Skripsi yang berjudul "Aplikasi Metode *Random Forests* dalam Memprediksi Tingkat Mortalita Asuransi Grup Cacat Jangka Panjang" ini disusun sebagai salah satu syarat wajib untuk menyelesaikan studi Strata-1 Program Studi Matematika, Fakultas Teknologi Informasi dan Sains (FTIS), Universitas Katolik Parahyangan (UNPAR), Bandung. Penulis berharap skripsi ini dapat menjadi karya yang bermanfaat bagi setiap orang yang membacanya.

Penyusunan skripsi ini tidak luput dari hambatan dan kesulitan. Oleh karena itu, penulis ingin berterima kasih kepada :

- Orang tua penulis, mama Dumasary Simamora, dan papa Jhonny Munthe, yang tidak pernah lelah memberikan dukungan moral dan doa kepada penulis. Dan kedua kakak penulis Fenny Permatasari dan Debora Lusiana, yang sabar dan selalu memberi semangat. Tidak lupa kepada seluruh keluarga besar penulis yang selalu memberi motivasi.
- Bapak Dr. Julius Dharma Lesmono selaku dosen pembimbing 1 yang telah dengan sabar meluangkan waktu untuk membimbing penulis, memberi arahan, didikan, dan saran sehingga skripsi ini dapat terselesaikan tepat waktu.
- Ibu Felivia Kusnadi, M.Act.Sc. selaku dosen pembimbing 2 yang telah memberikan arahan, didikan, saran, dan bantuan di setiap proses penyusunan skripsi ini sehingga skripsi ini dapat terselesaikan dengan baik.
- Ibu Farah Kristiani, Ph.D. selaku dosen penguji yang telah memberikan arahan dan saran untuk perbaikan dan pengembangan skripsi ini.
- Bapak Taufik Limansyah, M.T. selaku dosen wali dan dosen penguji yang telah memberikan arahan dan nasihat kepada penulis selama perkuliahan dan dalam pengembangan skripsi ini.
- Bapak Liem Chin, M.Si. selaku koordinator skripsi atas segala saran dan bantuan yang diberikan.
- Seluruh dosen FTIS, terutama dosen Program Studi Matematika, terima kasih atas segala ilmu yang diberikan.
- Seluruh staf Tata Usaha dan karyawan FTIS. Terima kasih atas segala bantuan selama masa perkuliahan.
- Teman-teman kobra : Julius, Aldo, Rudi, Yonathan. Terima kasih untuk pengalaman bermain dan belajar sepanjang berkuliah bersama penulis.
- Semua teman-teman Matematika 2016. Terima kasih untuk seluruh pengalaman serta kebersamaan selama proses perkuliahan bersama penulis. *May success is our future hold.*
- Teman-teman Matematika angkatan 2014, 2015, 2017, 2018, yang tidak dapat disebutkan satu persatu.

- Teman-teman U10 : Michael, Emily, Erick, Sabrina, Nadya, Satrio, Grace, Joshua, Marie, Maxi, Rama. Terima kasih untuk dukungan motivasi, doa, dan canda tawa selama bersama dengan penulis.
- Semua pihak yang telah berjasa kepada penulis selama masa perkuliahan dan penyusunan skripsi.

Penulis menyadari bahwa skripsi ini masih memiliki banyak kekurangan dan jauh dari kata sempurna. Oleh karena itu, penulis mengharapkan masukkan saran dan kritik yang membangun dari para pembaca agar skripsi ini dapat menjadi lebih baik. Akhir kata semoga skripsi ini dapat bermanfaat dan dapat dikembangkan menjadi karya yang lebih baik lagi.

Bandung, Juli 2020

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Dalam menjalani hari-harinya, setiap individu dihadapkan oleh berbagai kejadian yang tidak menentu. Contohnya, cuaca yang berubah-ubah, nilai mata uang yang fluktuatif, serta penyakit yang muncul tidak menentu. Hal-hal ini cenderung merugikan dan dapat mengganggu kehidupan dari individu tersebut. Oleh karena itu, para ahli dalam berbagai bidang berusaha melakukan prediksi di masa yang akan datang, sehingga manusia dapat menghindari berbagai hal yang dapat merugikan dirinya. Dalam bidang meteorologi contohnya, setiap hari selalu disiarkan berita ramalan cuaca baik di stasiun televisi maupun surat kabar. Juga dalam bidang kesehatan, dokter (atau ahli kesehatan) dapat mendiagnosis penyakit tanpa harus melihat langsung adanya kerusakan organ dari pasien. Kedua ilustrasi ini merupakan contoh prediksi respons yang dilakukan lewat observasi gejala (prediktor) kasus serupa di masa lampau.

Peran matematika dalam hal ini adalah membentuk model aplikasi yang dapat digunakan dalam mempermudah prediksi. Namun, kendala muncul ketika bertambahnya variabel yang harus dipertimbangkan sehingga modelnya menjadi sangat kompleks untuk dikonstruksi. Untuk menyelesaikan permasalahan tersebut, para ahli cenderung menggunakan mesin sebagai alat teknologi modern untuk melakukan perhitungan yang akurat dalam prediksi. Dengan perkembangan komputer sains dan kecerdasan buatan yang semakin pesat, mesin telah memiliki performa yang lebih baik dari kalkulator. Sampai saat ini, mesin telah berkembang hingga mampu untuk mempelajari data dan melakukan prediksi yang terstruktur.

Contoh mesin yang dideskripsikan di atas dapat diklasifikasikan ke dalam *Machine Learning*. Berdasarkan [3], *Machine Learning* sering didefinisikan sebagai sistem yang dapat mempelajari data tanpa diprogram secara eksplisit. Secara khusus *Machine Learning* menyajikan algoritma yang mampu menyelesaikan persoalan data klasifikasi ataupun regresi. Terdapat beberapa contoh metode algoritma dalam *Machine Learning*, di antaranya adalah *Kernel*, *Nearest Neighbor*, *Boosting*, *Decision Tree*, dan *Random Forests*. Sebagai cabang dari *Decision Tree*, *Random Forests* memiliki keunggulan lebih yaitu tidak memerlukan *pruning* (pemangkasan) *tree*. Maka pada skripsi ini, *Random Forests* akan digunakan sebagai metode dalam pembentukan model.

Berdasarkan [4], *Random Forests* adalah *learning algorithm* yang secara acak membentuk dan menggabungkan *Decision Tree* berjumlah banyak ke dalam suatu himpunan *tree* yang disebut *forest*, kemudian menghasilkan rata-rata prediksi (regresi) atau keluaran kategori (klasifikasi) dari *forest* yang dikonstruksikan; sedangkan *Decision Tree* adalah *chart* bercabang dua, yang digunakan untuk membantu dalam pembuatan keputusan berupa regresi maupun klasifikasi. Metode *Random Forests* dilakukan dengan membentuk M buah *subsample* (*bootstrap sample*), dan masing-masing *subsample* akan dibentuk *Decision Tree*-nya. Pembentukan *subsample* diprogram secara acak dengan jumlah yang banyak untuk kemudian dihitung rata-rata nilai respons sebagai nilai prediksi.

Dengan perkembangan *machine learning* saat ini, banyak ahli matematika telah melakukan pemodelan *machine learning* untuk kasus di dunia kesehatan. Contohnya seperti model prediksi penyakit yang sering dilakukan menggunakan *decision tree*. Pada [5] telah dibahas bagaimana aplikasi model *decision tree* di bidang asuransi kesehatan dilakukan. Pemodelan dilakukan menggunakan

data Asuransi Grup Cacat Jangka Panjang (*Group Long Term Disability*) dari *Society of Actuary* (SOA) tahun 2008 [2] yang kemudian diekstraksi menjadi data *Group Long Term Disability* (GLTD) 2017 dalam mencari prediksi tingkat mortalita. Data berupa himpunan variabel prediktor, seperti durasi, umur, jenis kelamin, kategori cacat, kelompok pekerjaan serta indeks gaji bulanan, dan satu variabel respons berupa tingkat mortalita. Pada GLTD 2017 diperoleh hasil model yang rumit dan nilai galat yang lebih besar setelah dilakukan *pruning* (pemangkasan model *tree*). Oleh karena itu, model baru akan dibuat menggunakan *Random Forests* yang tidak membutuhkan *pruning* model *tree*. Terakhir, akan dihitung nilai galat dari nilai prediksi dengan data asli menggunakan *Mean Squared Errors*.

1.2 Rumusan Masalah

Dalam skripsi ini, rumusan masalah yang akan dikaji adalah :

1. Bagaimana menaksir tingkat mortalita secara regresif menggunakan metode *random forests*?
2. Bagaimana membentuk model prediksi tingkat mortalita secara regresif menggunakan metode *random forests*?
3. Bagaimana menaksir variabel terpenting bagi data mortalita dalam *random forests*?
4. Bagaimana menaksir nilai parameter optimum dalam *random forests*?
5. Bagaimana pengaruh perubahan parameter terhadap model?
6. Bagaimana perbandingan hasil prediksi model *decision tree* pada laporan GLTD 2017 dengan *random forests*?

1.3 Tujuan

Dari skripsi ini tujuan yang ingin dicapai adalah :

1. menaksir tingkat mortalita secara regresif menggunakan metode *random forests*;
2. membentuk model prediksi tingkat mortalita secara regresif menggunakan metode *random forests*;
3. menaksir variabel terpenting bagi data mortalita dalam *random forests*;
4. menaksir nilai parameter optimum dalam *random forests*;
5. mengetahui pengaruh perubahan parameter;
6. membandingkan hasil prediksi model *decision tree* pada laporan GLTD 2017 dengan *random forests*.

1.4 Batasan Masalah

Dari skripsi ini, terdapat batasan masalah, yaitu :

1. sampel *bootstrap* hanya akan dibentuk oleh program R;
2. diasumsikan kandidat *split* variabel tiap cabang hanya 1 (`mtry`=1);
3. data yang akan dimodelkan hanya data mortalita asuransi grup cacat jangka panjang.

1.5 Sistematika Pembahasan

Adapun sistematika penulisan skripsi ini adalah :

- BAB I PENDAHULUAN
  Bab ini berisi latar belakang masalah, rumusan masalah, tujuan, batasan masalah, serta sistematika pembahasan.
- BAB II LANDASAN TEORI
  Dalam bab ini dibahas definisi asuransi grup cacat, *machine learning*, metode *decision tree*, dan dasar perhitungan yang akan digunakan untuk perhitungan skripsi ini.
- BAB III *TREE-BASED MODEL* : *RANDOM FORESTS*
  Bab ini membahas model modifikasi, algoritma, data *out-of-bag*, *variable importance*, dan *tuning random forests* yang akan dilakukan secara regresif.
- BAB IV PENERAPAN *DECISION TREE*
  Bab ini membahas data yang digunakan, hasil laporan *Group Long Term Disability* tahun 2017, dan metode lanjutan dalam pengembangan model.
- BAB V KONSTRUKSI MODEL DAN ANALISIS HASIL
  Bab ini berisi langkah-langkah konstruksi model dalam mengolah data, serta analisa dari hasil yang diperoleh.
- BAB VI KESIMPULAN DAN SARAN
  Bab ini berisi kesimpulan analisis model yang telah dilakukan, sekaligus menjawab tujuan dari penulisan skripsi ini dan saran untuk pembahasan lebih lanjut mengenai topik ini.