Laporan Akhir Tugas MK. Data Mining (KOM332), Semester Genap 2021/2022

Klasifikasi Keselamatan Pasien yang Dirawat di Rumah Sakit dengan Menggunakan *Random Forest*

Nadira Nazla (G64190035), Tio Ramadhan (G64190045), Rizal Mujahiddan (G64190069), Muhammad Akmal Khairurrahman (G64190110)
Kelompok: 5, Kelas Paralel: 1

ABSTRAK

Kualitas pelayanan kesehatan di rumah sakit yang komprehensif dan responsif merupakan hal yang penting bagi keselamatan pasien. Maka penulis ingin mengetahui faktor apa saja yang paling berpengaruh dalam tingkat kematian pasien di rumah sakit, serta seberapa sesuai model Random Forest dengan permasalahan tersebut, serta memprediksi keberlangsungan hidup pasien di rumah sakit. Data yang penulis gunakan pada penelitian ini diambil dari Kaggle yang berjudul "Patient Survival Predictions". Dari hasil eksplorasi data tersebut, penulis melakukan preprocessing dengan menghilangkan Missing Value, Melakukan Feature Engineering, Feature Selection, serta Balancing Class. Perangkat lunak yang digunakan pada penelitian ini adalah R dan R-Studio. Dalam mengatasi Missing Value, penulis menghilangkan variabel tertentu yang lebih dari 10%, serta imputasi variabel datang yang kurang dari 10%. Feature Engineering, dilakukan perataan nilai yang memiliki min dan max lalu dilakukan pemfaktoran dari nilai kategorik tersebut. Pada Feature Selection, dilakukan seleksi dengan multikolinearitas. Dengan bantuan model regresi logistik, dihilangkan variabel atau atribut yang menyebabkan singularitas. Balancing Class dilakukan agar algoritma optimal dan seimbang. Klasifikasi *Random Forest* menghasilkan 500 trees, dan diperoleh variabel yang paling berpengaruh terhadap klasifikasi keselamatan pasien. Pemodelan menggunakan random forest mendapatkan akurasi sebesar 86.62%. hal ini membuktikan bahwa pemodelan random forest cukup tepat digunakan dalam permasalahan ini.

Kata Kunci : Keselamatan, Pasien, Random Forest

PENDAHULUAN

Keselamatan pasien merupakan suatu hal yang sangat penting dalam sebuah pelayanan kesehatan di rumah sakit. Maka dalam mencapai dan menjaga kualitas pelayanan rumah sakit dipelukan tindakan yang komprehensif dan responsif dari kejadian yang tidak diinginkan (KTD). agar kejadian serupa tidak terulang kembali, resiko KTD dapat diminimalkan bahkan dicegah dengan memperhatikan keselamatan pasien. Di Amerika, hasil studi keselamatan pasien pada akhir tahun 1990-an menemukan angka 3,9% dan 2,7% pada angka kejadian yang tidak diinginkan (KTD) pada pasien rawat inap (Brennan 1991).

Berdasarkan alasan yang telah dipaparkan di atas serta rumusan masalah yang telah penulis teliti tersebut, penulis ingin mengetahui faktor apa yang paling terpengaruh dalam tingkat kematian pasien yang dirawat, lalu mengetahui seberapa sesuai model random forest dengan permasalahan tersebut serta mengetahui seberapa banyak pohon keputusan yang paling sesuai dalam suatu random forest pada permasalahan tersebut

Pemodelan dengan algoritma *Random Forest* ini berguna untuk memprediksi keberlangsungan hidup pasien selama rawat inap di rumah sakit (*survive or death*). Dengan adanya solusi ini kita dapat melakukan langkah awal sebagai bentuk pencegahan tingkat kematian pasien rawat inap di rumah sakit.

Latar Belakang

Keselamatan pasien merupakan suatu hal yang sangat penting dalam sebuah pelayanan kesehatan di rumah sakit. Hingga studi-studi terkini. Di Amerika, hasil studi keselamatan pasien pada akhir tahun 1990-an menemukan angka 3,9% dan 2,7% angka kejadian yang tidak diinginkan (KTD) pada pasien rawat inap(Brennan 1991). Dua puluh tahun kemudian, pengukuran dengan Global Trigger Tool menunjukkan bahwa KTD meningkat 10 kali lipat (menjadi 32%) (Classen 2001).

Penulis melakukan pemodelan dengan algoritma random forest untuk memprediksi keberlangsungan hidup pasien selama rawat inap di rumah sakit (survive or not). Dengan adanya solusi ini kita dapat melakukan langkah awal sebagai bentuk pencegahan tingkat kematian pasien rawat inap di rumah sakit.

Tujuan

1. Mengetahui atribut apa yang paling terpengaruh dalam tingkat kematian pasien yang dirawat
2. Mengetahui seberapa cocokkah model *random forest* dengan permasalahan tersebut
3. Mengetahui seberapa banyak pohon keputusan yang paling sesuai dalam suatu *random forest* pada permasalahan tersebut

Ruang Lingkup

Ruang lingkup dalam penelitian ini ialah set data dari MIT's GOSSIS (*Global Open Source Severity of Illness Score*) *initiative* (2021), yaitu data yang menekankan tentang kondisi kronis diabetes. Metode yang digunakan pada penelitian ini adalah *random forest*. Selain itu, digunakan *Caret package* dan *ROSE package* pada R.

Manfaat

Manfaat penelitian ini adalah membuat suatu model yang dapat mengklasifikasi keselamatan pasien yang dirawat di rumah sakit sehingga ke depannya, dapat dilakukan pencegahan tingkat kematian pasien rawat inap di rumah sakit..

TINJAUAN PUSTAKA

Data Mining

Data Mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar (Larose D, T., 2005).

Data mining mengolah data mentah yang tersimpan pada basis data sehingga menghasilkan informasi yang dapat digunakan. Data mentah merupakan data yang akan disimpan sebagai bukti dokumentasi. Pengolahan data mentah dapat dimanfaatkan untuk menemukan informasi baru yang dibutuhkan. Data mining menelusuri data pada database

untuk membangun model dan menggunakannya untuk mengenali pola data lain yang tidak tersimpan dalam basis data.

Klasifikasi

Klasifikasi adalah suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan menggunakan hasilnya untuk memberikan sejumlah aturan.

Missing Value

Missing value adalah informasi yang tidak tersedia untuk sebuah objek atau data. Missing value terjadi karena informasi untuk sesuatu tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada. Missing value pada dasarnya tidak bermasalah bagi keseluruhan data, apalagi jika jumlahnya hanya sedikit, misal hanya 1 % dari seluruh data. Namun jika persentase data yang hilang tersebut cukup besar, maka perlu dilakukan pengujian apakah data yang mengandung banyak missing tersebut masih layak diproses lebih lanjut ataukah tidak.

Random Forest

Metode Random Forest merupakan metode yang dapat meningkatkan hasil akurasi, karena untuk setiap node dilakukan secara acak. Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari root node, internal node, dan leaf node dengan mengambil atribut dan data secara acak sesuai ketentuan yang berlaku. Root node merupakan akar dari pohon keputusan. Internal node adalah simpul percabangan, dimana node ini mempunyai output minimal dua dan hanya ada satu input. Sedangkan leaf node merupakan simpul terakhir yang hanya memiliki satu input dan tidak mempunyai output. Pohon keputusan dimulai dengan cara menghitung nilai entropy sebagai penentu tingkat ketidakmurnian atribut dan nilai information gain (Schouten, K., F. Frasincar, and R. Dekker, 2016). Untuk menghitung nilai entropy digunakan rumus:

$$Entropy\,(Y) \;=\; -\sum_i p(c|Y)\,log_2\,p(c|Y)$$

Dimana Y adalah himpunan kasus dan p(c|Y) merupakan proporsi nilai Y terhadap kelas c.

$$Information\,Gain\,(Y,\,a) \;=\; Entropy\,(Y) \;-\; \sum_{v \epsilon Values(a)} \frac{|Y_v|}{|Y_a|}\,Entropy\,(Y_v)$$

Dimana Values(a) merupakan semua nilai yang mungkin dalam himpunan kasus a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya adalah semua nilai yang sesuai dengan a.

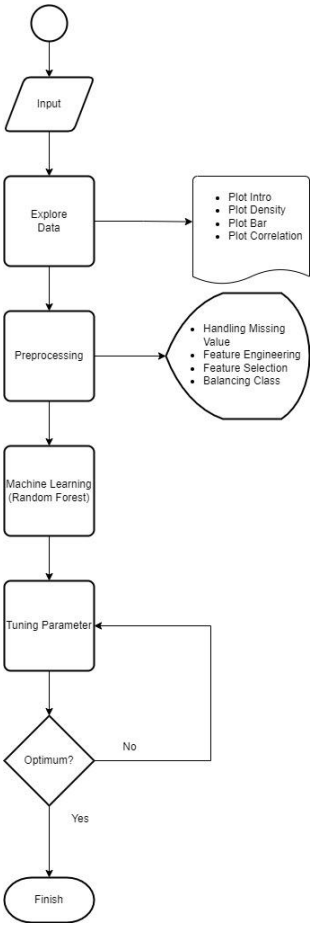
METODE

Data

Data yang digunakan adalah data yang bersumber dari kaggle yang berjudul **Patient Survival Prediction** yang dikirim oleh Mitisha Agarwal. Data ini menjelaskan mengenai hubungan antara gejala penyakit baik tidak berbahaya maupun berbahaya pada pasien rumah sakit dengan tingkat kematian pada pasien tersebut. Data tersebut biasanya diperoleh dengan keperluan GOSSIS (GLOBAL OPEN SOURCE SEVERITY OF ILLNESS SCORE) Consortium dan didapat dari pihak GOSSIS tersebut. Jadi Mitisha

Agarwal tersebut melakukan penggabungan data - data. Data tersebut terdiri banyak atribut yang menyatakan kandungan suatu kadar suatu enzim maupun hormon seperti insulin dan juga atribut yang lain mengenai kesehatan tersebut.

Tahapan Kegiatan



Penulis mengeksplorasi data tersebut dengan melihat data secara keseluruhan. mulai dari ukuran data , melihat *missing value* pada kolom tertentu, melihat keseluruhan data yang memiliki *missing value* (setiap satu sel data), hingga melihat banyaknya atribut yang diskrit dan kontinu. Kemudian, melihat density plot, untuk melihat distribusi pada masing masing kolom. Mengapa melihat distribusi data?, Dikarenakan untuk menentukan metode apa yang digunakan imputasi (apakah dengan menggunakan mean atau median). Plot bar dilihat untuk melihat seberapa pengaruh variabel independen terhadap variabel respon dikarenakan variabel respon bersifat kategorik. plot kategorik untuk melihat korelasi seluruh variabel agar mengetahui apakah ada suatu multikolinearitas.

Preprocessing yang digunakan banyak sekali, dan ini berdasarkan dari eksplorasi yang telah dilakukan. preprocessing yang dilakukan oleh penulis adalah sebagai berikut:

1. Mengatasi *Missing Value*
2. Melakukan *Feature Engineering*
3. Melakukan *Feature Selection*
4. Menyeimbangkan kelas target (*Balancing Class*)

Untuk mengatasi missing value, penulis menghilangkan attribut yang persentase *missing value* melebihi 10%. maka variable X dibuang. Kemudian, melakukan imputasi

median dikarenakan kebanyakan data numerik cenderung menjulur ke kiri maupun ke kanan (*Skewed distribution*), sedangkan untuk data kategorik lebih baik diimputasi dengan modus saja.

Untuk melakukan *Feature Engineering* , penulis hanya melakukan perata-rataan dengan nilai yang memiliki min dan max. dan kemudian dengan nilai kategorik tersebut, dilakukan pemfaktoran data di kategori.

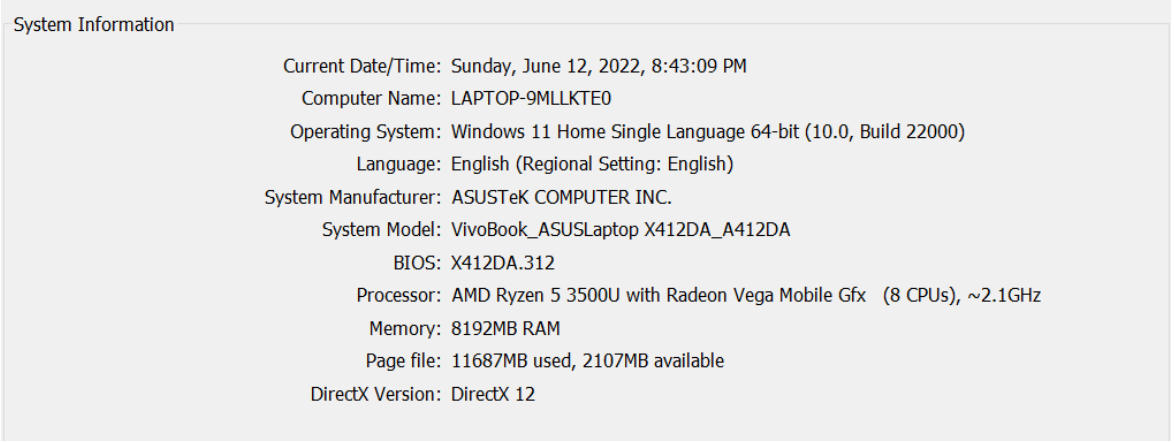
Untuk melakukan *Feature Selection* , penulis melakukan seleksi dengan multikolinearitas (dibuktikan ada multikolinearitas dengan bantuan plot korelasi). Dengan bantuan model regresi logistik, maka jika diperhatikan masih ada singularitas, maka dihilangkan variabel yang menyebabkan singularitas tersebut.

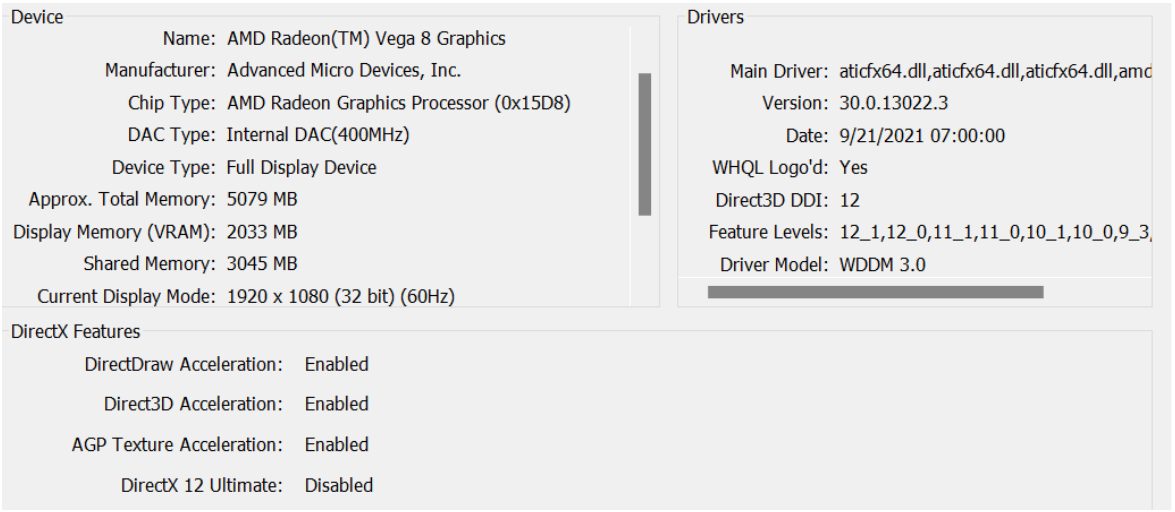
Untuk menyeimbangkan kelas target (*Balancing Class*), ini dilakukan dikarenakan setiap target memiliki nilai yang tidak seimbang. Untuk konsep cara menyeimbangkan kelas target, mengikuti konsep yang sudah tertera di dokumentasi library ROSE tersebut.

Lingkungan Pengembangan

Perangkat Lunak yang digunakan adalah menggunakan R. R itu adalah suatu bahasa pemrograman dan juga environment untuk grafik dan perhitungan statistika seperti grafik qq plot dengan uji kenormalannya . Untuk penggunaan GUI-nya. Penulis menggunakan R Studio dikarenakan ada sistem autocomplete syntax, bisa melihat objek atau variabel yang sudah dideklarasikan, mudah untuk mensetting directory dan mengakses file directory secara cepat dan tepat, membuat grafik yang mudah untuk diakses, dan yang paling penting adalah open source dan gratis.

Perangkat Keras yang digunakan adalah sebagai berikut

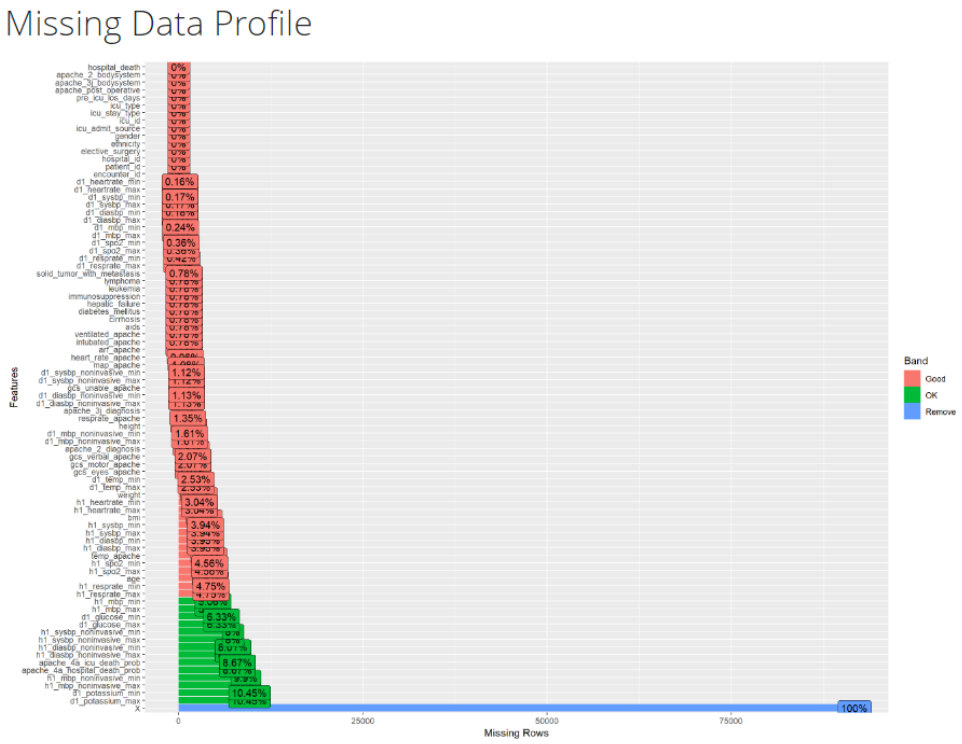




HASIL DAN PEMBAHASAN

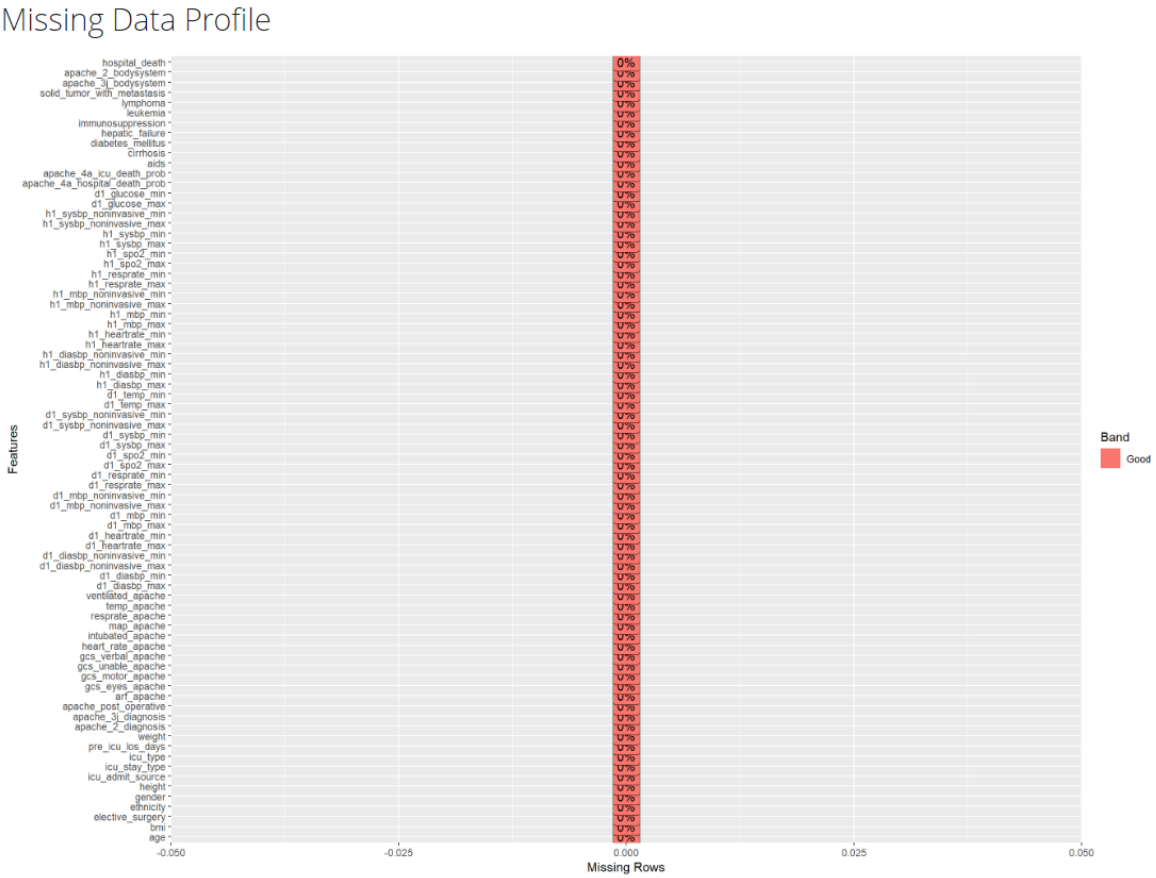
Sebelum dilakukan Preprocessing , data tersebut memiliki identitas sebagai berikut:

Name	Value
Rows	91,713
Columns	85
Discrete columns	7
Continuous columns	77
All missing columns	1
Missing observations	283,190
Complete Rows	0
Total observations	7,795,605
Memory allocation	37.1 Mb



Jika diperhatikan, banyak variabel yang memiliki missing value . Kemudian, untuk langkah pertama, Penulis menghilangkan variabel tertentu yang lebih dari 10% , dan juga imputasi variabel yang datanya kurang dari 10% . Untuk pemilihan jenis imputasinya , penulis menggunakan median untuk numerik dikarenakan distribusinya bersifat *skewed*, sedangkan untuk kategorik digunakan modus dikarenakan pada umumnya kategorik di imputasi dengan modusnya. Setelah dilakukan ini. maka hasilnya akan seperti ini

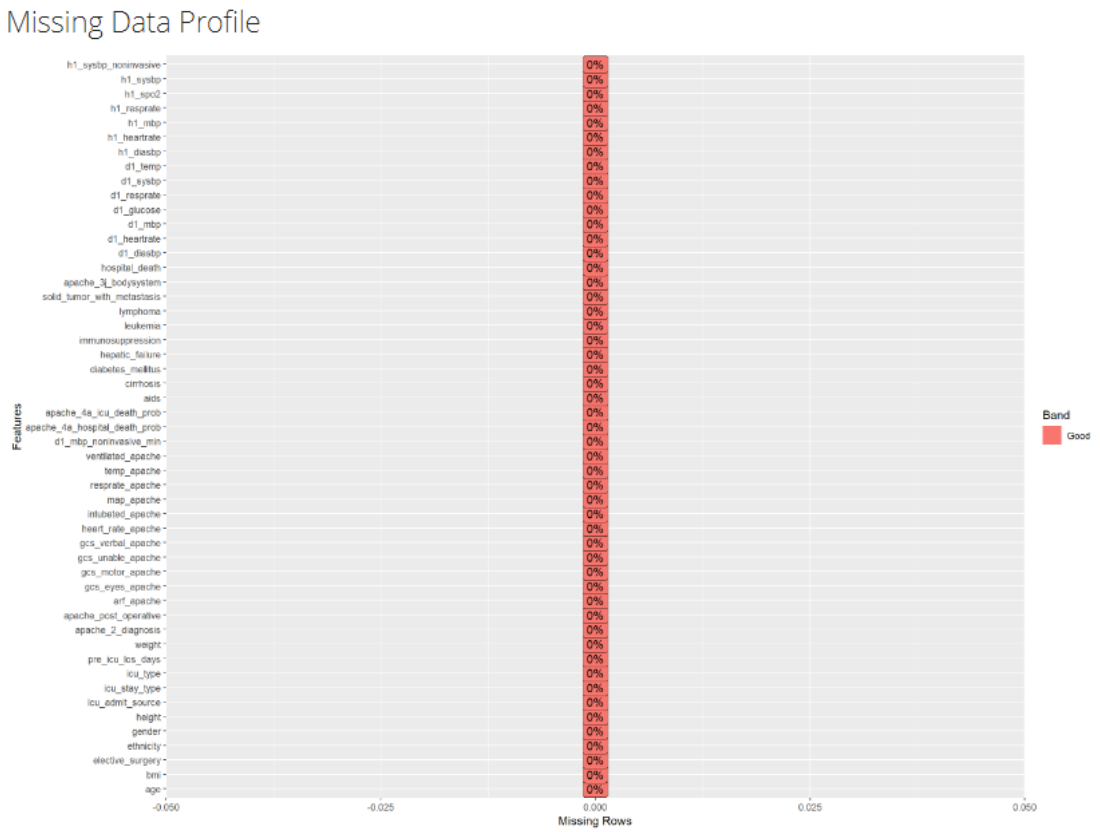
Name	**Value**
Rows	91,713
Columns	78
Discrete columns	23
Continuous columns	55
All missing columns	0
Missing observations	0
Complete Rows	91,713
Total observations	7,153,614
Memory allocation	41 Mb



Setelah mengatasi missing value, penulis melakukan *Feature Engineering* ,akan dilakukan perataan dengan nilai yang memiliki min dan max. dan kemudian dengan nilai kategorik tersebut, dilakukan pemfaktoran data di kategori. Setelah dilakukan ini. maka hasilnya akan seperti ini

Name	**Value**
Rows	91,713
Columns	58
Discrete columns	23
Continuous columns	38
All missing columns	0
Missing observations	0
Complete Rows	91,713
Total observations	5,319,354

Memory allocation	30.5 Mb



Selanjutnya penulis melakukan *Feature Selection* , penulis melakukan seleksi dengan multikolinearitas (dibuktikan ada multikolinearitas dengan bantuan plot korelasi). Dengan bantuan model regresi logistik, maka jika diperhatikan masih ada singularitas, maka dihilangkan variabel/atribut yang menyebabkan singularitas tersebut. Setelah hal ini dilakukan, maka hasilnya akan seperti ini

Name	**Value**
Rows	91,713
Columns	51
Discrete columns	22
Continuous columns	29
All missing columns	0
Missing observations	0
Complete Rows	91,713

Total observations	4,677,363
Memory allocation	25.9 Mb

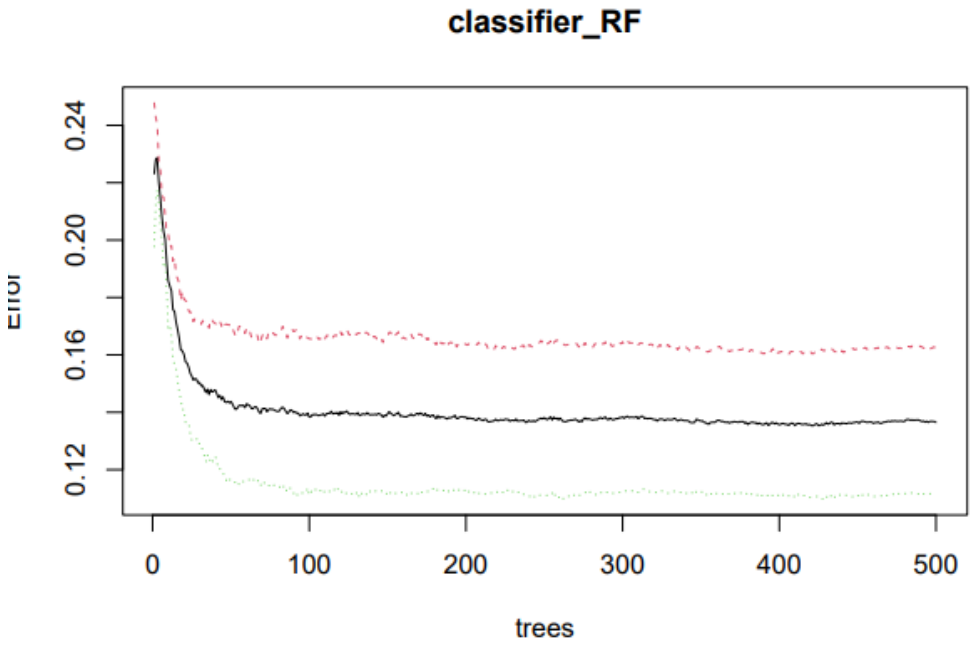
tahap terakhir dalam *preprocessing* yang kamu lakukan ialah *Balancing Class,* Untuk menyeimbangkan kelas target (*Balancing Class*), hal ini dilakukan dikarenakan setiap target memiliki nilai yang tidak seimbang. Untuk konsep cara menyeimbangkan kelas target, mengikuti konsep yang sudah tertera di dokumentasi library ROSE.

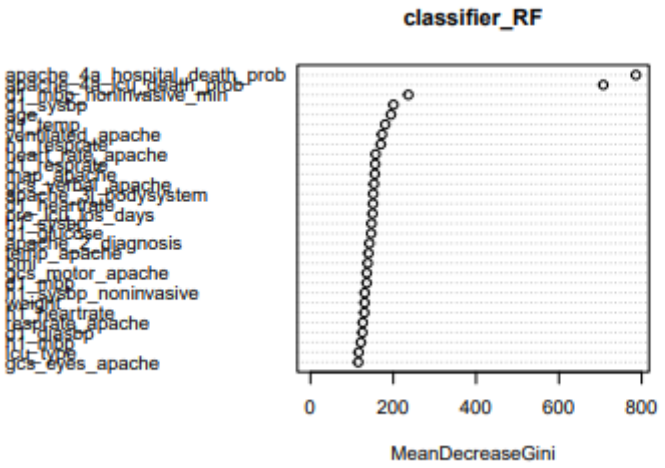
Sebelum *balancing class* :

Setelah dilakukan *balancing class*

Dapat kita lihat setelah dilakukan *balancing class*, atribut *class hospital_death* lebih seimbang daripada sebelumnya. Hal ini penting sebelum menggunakan algoritma *random forest* karena tujuan akhir penulis adalah melatih model yang dapat digeneralisasi dengan baik untuk semua kelas yang mungkin dengan asumsi penulis memiliki kumpulan data biner dengan jumlah sampel yang sama.

Setelah dilakukan preprocessing, maka data dapat digunakan dalam klasifikasi keselamatan pasien yang dirawat di rumah sakit dengan menggunakan *Random Forest.* penulis membagi data training dan data test dengan proporsi 80% untuk data training dan 20% untuk data test. lalu melakukan *model evaluation.* seperti hasil yang dibawah ini





Dihasilkan sebanyak 500 trees dan diperoleh ***apache_4a_hospital_death_prob***, *dan* ***apache_4a_icu_Death_prob*** memiliki pengaruh yang paling kuat terhadap klasifikasi keselamatan pasien yang dirawat inap dirumah sakit dengan perubahan signifikan errornya terjadi antara 1-70 pohon.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Survived Death
##   Survived     1310   176
##   Death         256  1426
##
##                Accuracy : 0.8636
##                  95% CI : (0.8512, 0.8754)
##     No Information Rate : 0.5057
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7271
##
##  Mcnemar's Test P-Value : 0.0001442
##
##             Sensitivity : 0.8365
##             Specificity : 0.8901
##          Pos Pred Value : 0.8816
##          Neg Pred Value : 0.8478
##              Prevalence : 0.4943
##          Detection Rate : 0.4135
##    Detection Prevalence : 0.4691
##       Balanced Accuracy : 0.8633
##
##        'Positive' Class : Survived
##
```

Dari *confusion matrix* yang diperoleh didapatkan akurasi sebesar 86.36 % , dengan True-Positive sebesar 1310 record, False-Positive sebesar 176 record, True-Negative sebesar 256 record, dan False-Negative sebesar 1426 record.. Pemilihan model mengacu pada proses pemilihan model yang tepat yang sesuai dengan data. Ini dilakukan dengan menggunakan matriks evaluasi tes. Hasil dari data pengujian diteruskan kembali ke hyperparameter tuner untuk mendapatkan hyperparameter yang paling optimal.

Description: df [4 x 6]

ntreeku <dbl>	mtryku <dbl>	accuracyku <dbl>	precisionku <dbl>	recallku <dbl>	f1ku <dbl>
100	7	0.8636364	0.8795181	0.8390805	0.8588235
200	7	0.8655303	0.8835801	0.8384419	0.8604194
500	7	0.8661616	0.8852901	0.8378033	0.8608924
1000	7	0.8655303	0.8846154	0.8371648	0.8602362

4 rows

Diperoleh sebanyak 500 trees dengan nilai akurasi yang paling tinggi diantara pohon keputusan lainnya dengan nilai akurasi sebesar **86.62%**

KESIMPULAN DAN SARAN

Berdasarkan hasil pemodelan menggunakan random forest untuk Klasifikasi Keselamatan Pasien yang Dirawat di Rumah Sakit, diperoleh :

1. Atribut yang paling berpengaruh dalam tingkat keselamatan pasien yang dirawat dirumah sakit ialah atribut ***apache_4a_hospital_death_prob***, ***apache_4a_icu_Death_prob***
2. Pemodelan menggunakan random forest untuk Klasifikasi Keselamatan Pasien yang Dirawat di Rumah Sakit mendapatkan akurasi sebesar **86.62%**. hal ini membuktikan bahwa pemodelan random forest cukup tepat digunakan dalam permasalahan ini.
3. Banyak pohon keputusan yang paling sesuai dalam pemodelan menggunakan *random forest* pada permasalahan ini didapatkan sebesar 500 pohon keputusan dengan akurasi sebesar **86.62%**
4. Jika diperhatikan, variabel yang berpengaruh adalah *death probability*. Disini ada suatu hal yang bisa disimpulkan bahwa kemungkinan peluang tersebut mendekati hasil akhir tersebut. Kemudian, hasil dari *random forest* tersebut kurang baik jika terlalu banyak pohon maupun terlalu sedikit pohon. Hal tersebut dikarenakan jika sedikit pohon menjadi ***underfitting***. Dan jika banyak pohon menjadi ***overfitting***. Bisa diperhatikan juga bahwa akurasi, precision recall, dan f1 memiliki nilai yang sama tinggi pada masing masing seakan-akan relatif. Metrik apapun yang dipakai akan menghasilkan metrik yang sama tinggi dengan metrik yang lain jika dibandingkan dengan parameter yang berbeda.

DAFTAR PUSTAKA

Brennan TA. and Leape LL.1991. *Adverse events, negligence in hospitalized patients: Results from the Harvard Medical Practice Study. Perspect Healthc Risk Manag*(US).11(2): 2-8.

Han J, Kamber M. 2006. Data Mining Concepts and Techniques Second Edition. San Fransisco (US): Morgan Kaufmann Publisher.

James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. 1st ed. 2013, Corr. 7th printing 2017 edition. Springer; 2013.

Janecek AGK, Gansterer WN, Demel MA, Ecker GF. (2008). *On the relationship between feature selection and classification accuracy*. JMLR: Workshop and Conference Proceedings, 4, 90-105.

S. Cohen, N. Dagan, N. Cohen-Inger, D. Ofer and L. Rokach, "ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models," in IEEE Access, vol. 9, pp. 91584-91592, 2021, doi: 10.1109/ACCESS.2021.3091622.

Schouten, K., F. Frasincar, and R. Dekker. An Information gain-Driven Feature Study for Aspect-Based Sentiment Analysis. Natural Language Processing and Information Systems, 2016: p. 48-59.