P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

CHURN ANALISIS PADA DATA PELANGGAN TELEKOMUNIKASI MENGGUNAKAN ENSEMBLE LEARNING

Muthia Nadhira Faladiba[1*], Rizqi Haryastuti[2]

[1] Program Studi Statistika, Fakultas MIPA, Universitas Islam Bandung
*Email Korespondensi: muthia.nadhira@unisba.ac.id
[2] PT Visionet Internasional (OVO)
Email: rharyastuti197@gmail.com

ABSTRACT

Intense competition in broadband services will create high opportunities for consumers to switch providers, such as conditions that arise in competition for SMS, telephone, and internet services. The churn rate is the percentage of consumers who stop subscribing to the service. Ideally, this churn percentage is only 5% – 10%, and if it exceeds this figure, it indicates the company's inability to retain customers. A high churn rate indicates a decline in the cellular operator's market share and affects the company's revenue. Based on these problems, it is necessary to analyze the churn behavior of broadband subscribers to determine the dissatisfaction factors of cellular telecommunications consumers. Then predictions are made for customers who tend to churn from provider companies and determine the characteristics of churn and stay customers. The ensemble method is used to detect churn, which consists of several methods, including random forest, boosting, and super learner. Random Forest is proven to produce the best classification method with an excellent ability to predict customer churn, which is 80.1%, with an average usage time of 3 years.
Keywords: churn, ensemble, random forest, super learner, classification.

ABSTRAK

Persaingan layanan *broadband* yang semakin ketat akan menimbulkan peluang tinggi pelanggan untuk berpindah *provider* seperti kondisi yang timbul pada persaingan layanan SMS, telepon dan internet. *Churn rate* merupakan persentase banyaknya konsumen yang yang berhenti menggunakan jasa atau berhenti berlangganan. Idealnya, persentase churn ini berada pada angka 5% – 15% saja, jika melebihi angka tersebut mengindikasikan ketidakmampuan perusahaan untuk mempertahankan konsumen. Tingkat *churn* yang tinggi mengindikasikan penurunan pangsa pasar dan berpengaruh pada pendapatan perusahaan. Berdasarkan permasalahan tersebut maka diperlukan analisa akan perilaku *churn* pelanggan *broadband* untuk mengetahui faktor ketidakpuasan konsumen telekomunikasi selular. Maka dilakukan prediksi untuk pelanggan yang memiliki kecenderungan *churn* dari perusahaan *provider* dan menentukan karakteristik pelanggan *churn* dan *stay*. Metode *ensemble* digunakan untuk mendeteksi *churn*, yang terdiri dari beberapa metode diantarannya *random forest, boosting* dan *super learner*. *Random Forest* terbukti menghasilkan metode klasifikasi terbaik dengan kemampuan menduga pelanggan *churn* sangat bagus yaitu sebesar 80.1% dengan waktu rata – rata pemakaian 3 tahun.
Kata kunci: Churn, Ensemble, Random Forest, Super Learner, Klasifikasi.

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

1. PENDAHULUAN

Telekomunikasi selular berperan penting dalam kehidupan masyarakat dan perekonomian nasional terutama pada era digital seperti saat ini. Industri telekomunikasi selular berkembang pesat sejak Undang-Undang Nomor 36 Tahun 1999 tentang Telekomunikasi diberlakukan. Merujuk pada laporan e-Conomy Sea 2022, ekonomi digital Indonesia bernilai 77 miliar USD pada 2022 dan diprediksi menyentuh angka 130 miliar USD pada 2025, dengan *e-commerce* (perdagangan elektronik) sebagai pendorong utama (Google et al., 2020). Hal ini menunjukkan Indonesia sebagai pasar *e-commerce* terbesar di Asia Tenggara. Masa depan bisnis komunikasi seluler yang diprediksi terus meningkat tentu memberikan pengaruh besar bagi perusahaan jasa telekomunikasi di Indonesia. Terdapat 11 perusahaan seluler di Indonesia menimbulkan persaingan yang semakin ketat pada industri telekomunikasi seluler (Rizal, 2017).

Meningkatnya persaingan pada industri telekomunikasi membuat prediksi *churn rate* pelanggan menjadi sangat penting karena perusahaan telekomunikasi harus meningkatkan *Costumer Relationship Management* (CRM) yaitu dengan mempertahankan pelanggan yang sudah ada (Idris & Khan, 2012) (Miguéis et al., 2012). Prediksi *churn rate* pelanggan berpotensi membantu perusahaan telekomunikasi untuk mengidentifikasi pelanggan dengan potensi tinggi untuk berhenti berlangganan (Y. et al., 2022). Untuk mempertahankan pelanggan, perusahaan dapat mengevaluasi situasi dan merancang paket layanan yang sesuai bagi pelanggan. Beberapa faktor menjadi penyebab utama perilaku *churn*, diantaranya faktor ketidakpuasan konsumen dan faktor kondisi situasional (penentuan harga, ketidak nyamanan layanan, masalah etika, penarikan perhatian dari pesaing, dan situasi tak disadari). Studi lain dilakukan untuk menganalisis faktor *churn* pada pelanggan seluler prabayar di negara India, menginvestigasi bahwa jangkauan jaringan, penyelesaian keluhan, kecepatan internet dan layanan berbasis teknologi adalah faktor utama yang menyebabkan churn pelanggan (Rajeswari & Ravilochanan, 2014). Penulis menyimpulkan bahwa penyedia layanan harus fokus pada kualitas layanan dan jangkauan jaringan untuk mempertahankan pelanggan.

Loyalitas pelanggan memberikan kepercayaan kepada pelanggan dengan penyedia layanan mereka dan mendorong mereka untuk menggunakan layanan secara lebih konsisten dan dengan demikian mencegah pelanggan untuk mengalihkan loyalitas mereka ke pesaing lain (Mahajan et al., 2017). Berdasarkan permasalahan tersebut maka diperlukan analisa akan perilaku *churn* pelanggan broadband untuk mengetahui faktor ketidakpuasan konsumen, faktor kondisi situasional, dan *Switching Cost* yang dipersepsikan oleh pelanggan jasa *broadband* telekomunikasi selular. Informasi mengenai faktor-faktor tersebut akan bermanfaat untuk menentukan berbagai kebijakan strategis maupun rekomendasi yang sesuai dengan karakteristik perilaku *churn* pelanggan *broadband*.

Beberapa penelitian telah dilakukan dalam melalukan prediksi *churn*. (Ullah et al., 2019) melakukan prediksi pada dua data set telekounikasi dan menghasilkan metode *random forest* terbaik dalam memprediksi dengan akurasi sebesar 89.59%. Selanjutnya *churn rate* pada data telekomunikasi menggunakan metode regresi logistik menghasilkan akurasi sebesar 85.24% (Jain et al., 2020). (Ebrah & Elnasir, 2019) Membandingkan beberapa metode yaitu *naïve bayes, support vector machine* dan *decision tree* untuk dapat memprediksi *churn* pada dua data set dan metode *support vector machine* menghasilkan nilai akurasi tertinggi yaitu sebesar 87% (data set 1) dan 99% (data set 2).

Sebagian besar penelitian tersebut berfokus pada peningkatan model prediktif menggunakan algoritma *machine learning* seperti *decision tree, support vector machine*, dll. Namun, hasil prediksi belum terlalu baik karena masih ada pengklasifikasian yang salah. Sebagai alternatif metode yang dapat digunakan dalam memprediksi tingkat *churn* adalah

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

metode *ensemble*. (Mung & Phyu, 2020) melakukan perbandingan metode *ensemble* dan *non ensemble* pada data kanker serviks mendapatkan hasil berupa nilai akurasi 99.89 % dengan metode *Boosting* menggunakan SVM sebagai *base-classifier*, dan nilai akurasi 98.59 % dengan menggunakan metode *Bagging* dan *Decision Tree* sebagai *base-classifier*. Sehingga menyimpulkan metode *ensemble* menghasilkan hasil lebih baik.

Dari hal tersebut maka rumusan masalah dalam penelitian ini adalah "*Churn* Analisis Pada Data Pelanggan Telekomunikasi Menggunakan *Ensemble Learning*". Tujuan dari penelitian ini adalah untuk melakukan prediksi untuk pelanggan yang memiliki kecenderungan *churn* (berhenti berlangganan) dari perusahaan telekomunikasi dan menentukan karakteristik pelanggan yang melakukan churn atau menetap (*stay*).

2. METODOLOGI

Data yang digunakan adalah data pelanggan suatu perusahaan telekomunikasi. Langkah awal dalam proses analisis ini adalah menyiapkan data terlebih dahulu agar menjadi data yang berkualitas. Kualitas dari data masukan akan mempengaruhi hasil klasifikasi yang didapatkan. *Data preparation* merupakan tahapan mereduksi dimensi dari data, mengidentifikasi serta mengatasi data hilang (*missing value*), mengidentifikasi pencilan, menangani data *noise*, mengoreksi data yang tidak konsisten, dan sebagainya.

2.1. Imbalanced Data

Imbalanced Class atau data dengan kelas tidak seimbang merujuk pada situasi dimana keberadaan masing-masing kelas jumlahnya timpang, atau dengan kata lain kurang proporsional. Kasus data dengan kelas tidak seimbang perlu ditangani, sebab jika tidak, akan berdampak pada hasil prediksi dimana model gagal memprediksi amatan dari kelas minoritas (yang lebih sedikit jumlahnya) dengan benar meskipun mempunyai akurasi yang tinggi (Sun et al., 2009). Dengan kata lain, model hanya dapat memprediksi kelas mayoritas, sedangkan yang diinginkan tentu sebuah model yang dapat memprediksi dengan benar, baik itu kelas mayoritas maupun minoritas.

Undersampling merupakan salah satu penanganan data tidak seimbang dengan menggunakan pendekatan level data, bekerja dengan menggunakan semua amatan dari kelas minoritas dan mereduksi kelas mayoritas secara acak hingga sebanyak kelas minoritas, sedemikian sehingga kelas minoritas dan mayoritas mempunyai jumlah amatan yang sama.

2.2. Random Forest

Random Forest merupakan pengembangan dari metode CART (*Classification and Regression Tree*), yaitu dengan menerapkan metode *bootstrap aggregating (bagging)* dan *random feature selection* (Breiman, 2001). Berbeda dengan metode CART, metode ini merupakan metode pohon gabungan secara acak dimana banyaknya pohon yang dihasilkan membentuk suatu hutan (*forest*) sehingga analisis dilakukan pada kumpulan pohon tersebut. Algoritma random forest dengan gugus data berukuran n dengan peubah penjelas sebanyak q yaitu sebagai berikut.

1. Penarikan contoh acak dilakukan dengan pemulihan berukuran n dari gugus data. Tahap ini disebut tahap *bootstrap*.
2. Dengan menggunakan contoh *bootstrap*, pohon dibangun hingga mencapai ukuran maksimum. Hal ini dilakukan dengan menerapkan *random feature selection* pada setiap proses pemilihan pemisah, yaitu m peubah penjelas yang dipilih secara acak dengan $m < q$, lalu pemisah terbaik dipilih berdasarkan m peubah penjelas tersebut. Tahap ini disebut tahap *random sub-setting*.

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

3. Langkah 1 dan 2 dilakukan sebanyak L kali sehingga diperoleh L pohon keputusan. *Random forest* memprediksi respon suatu amatan dengan cara menggunakan semua hasil prediksi L pohon keputusan. Pada kasus klasifikasi digunakan teknik suara terbanyak (*majority vote*) untuk menentukan hasil suatu prediksi, yaitu kategori yang paling banyak muncul sebagai hasil prediksi dari L pohon klasifikasi.

2.3. Boosting

Serupa dengan *random forest, boosting* juga memprediksi model berdasarkan gabungan dari pohon keputusan, namun setiap pohon keputusan dibangun menggunakan informasi dari pohon keputusan sebelumnya. Boosting tidak melibatkan pengambilan contoh secara *bootstrap* (Lemmens & Croux, 2006). Sebagai gantinya setiap pohon keputusan dicocokkan dengan versi modifikasi dari data set.

Algoritma *boosting* dengan gugus data berukuran n dan peubah penjelas sebanyak q yang memiliki k kelas, yaitu sebagai berikut.

1. Tentukan bobot awal setiap pengamatan, yaitu $W_i = \frac{1}{n}, i = 1, \dots, n$
2. Untuk setiap iterasi ke-m, $m = 1, \dots, M$
   a. Susun pohon tunggal dengan bobot sebesar W_i
   b. Hitung tingkat kesalahan klasifikasi
   c. Hitung nilai ε_m
   d. Tentukan bobot baru untuk setiap pengamatan $W_i = W_i \varepsilon_m$ untuk yang diduga tidak tepat, untuk yang diduga tepat maka bobotnya tetap.
3. Prediksi akhir adalah kelas *k* yang memiliki nilai terbesar.

2.4. Super Learner

Super Learner adalah pendekatan berbasis *cross-validation* (CV) untuk menggabungkan prediksi dari beberapa metode dan meminimalkan *loss function*, seperti kesalahan prediksi, *negative-log-likelihood*, atau *rank loss* yang kemudian menghasilkan prediksi yang setidaknya sama baiknya dengan beberapa metode terbaik yang dimasukkan (Lee et al., 2022). Manfaat superlearner yaitu:

1. *Super Learner* memungkinkan Anda untuk menyesuaikan model *ensemble* dengan hanya menambahkan algoritma.
2. Seperti yang sudah Anda baca sebelumnya, *Super Learner* gunakan validasi silang, yang secara inheren digunakan untuk memperkirakan risiko untuk semua model.
3. *Super Learner* membuat *ensemble* menjadi efisien dengan secara otomatis memperkirakan bobot dari *ensemble*.
4. *Super Learner* secara otomatis menghapus model yang tidak berkontribusi pada kekuatan prediksi *ensemble*, ini membuat bebas untuk bereksperimen dengan banyak algoritma.

2.5. Prosedur Analisis

Analisis data dalam penelitian ini menggunakan perangkat lunak statistika R dan RStudio. Adapun *package* yang diperlukan: *caret, xgboost* dan *randomForest.* Selain itu dibutuhkan paket *classInt* dan *discretization* untuk melakukan diskretisasi pada peubah numerik dan menghitung nilai *weight of evidence* (WOE). Berikut adalah tahapan analisis data:

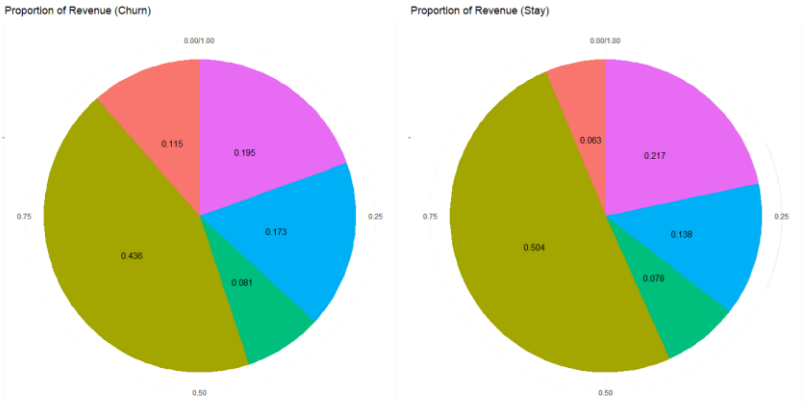
P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

1. Melakukan eksplorasi terhadap data berupa Visualisasi terhadap peubah respon, peubah prediktor dan membuang data amatan yang memiliki nilai *length of stay* lebih dari pencilan minor.
2. Melakukan *feature engineering*, yakni membuat peubah baru
3. Melakukan *undersampling* pada data *training*.
4. Melakukan pemodelan klasifikasi dengan menggunakan *cross validation* dengan *base learner random forest*, *boosting*, dan *RUS Boost* dengan k-fold sebesar 5 dan iterasi sebanyak 3.
5. Melakukan pemodelan klasifikasi dengan menggunakan *cross validation* dengan *super learner random forest, ranger, XGBoost, bagging,* dan *Bayes GLM.*
6. Membandingkan *accuracy*, *sensitivity*, *specificity* dan rataan gemotrik dari setiap metode.
7. Memilih metode terbaik berdasarkan hasil dari *balanced accuracy* dari prediksi sisa data.

3. HASIL DAN PEMBAHASAN

3.1. Eksplorasi Data

Dari total 90000 pelanggan, tercatat 61398 orang bertahan dengan provider tersebut (68.22%). Hal ini menunjukkan bahwa tingkat *Churn* pelanggan tidak begitu rendah. Namun tetap harus dicermati sebagai upaya pencegahan peningkatan *churn rate*. Pendapatan perusahaan provider pada satu bulan tersebut didominasi oleh para pelanggan yang melakukan isi ulang. *Recharge revenue* memberikan kontribusi sebesar Rp 2.569.740.000,00 atau berkisar 49.1% dari total pemasukan yang ada. Angka tersebut menunjukkan bahwa pelanggan memiliki kebutuhan yang besar dalam telekomunikasi. Hal tersebut dikarenakan pengguna pasca-bayar hanya perlu melakukan pengisian ulang jika paket layanan yang telah dibeli sebelumnya untuk satu bulan tersebut telah habis.



Gambar 1. Gambaran proporsi *revenue* pada status *Churn* dan *Stay*

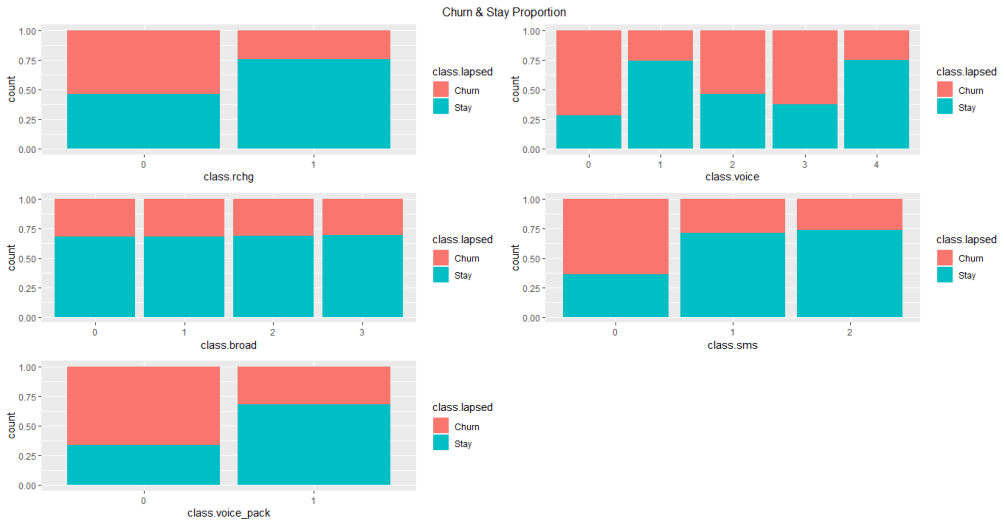
Selain itu perusahaan ini didominasi oleh pengguna layanan telepon, yang dibuktikan dengan *voice package revenue* dan *voice revenue* yang menduduki kontributor peringkat kedua dan ketiga, tanpa memandang pelanggan yang *Churn* atau *Stay*. Namun kontributor terkecil pada kedua kelas tersebut berbeda. Pada pelanggan yang *Churn*, pemasukan layanan SMS hanya berkisar 8.1% atau sebesar Rp 82.603.383,00 dari pemasukan. Sedangkan pada pelanggan *Stay*, pemasukan terkecil adalah layanan internet atau *broadband*, sebesar 6.3% atau sebesar Rp 264.486.932,00. Tentunya kita beranggapan bahwa pelanggan yang *Churn* adalah pelanggan memiliki nilai *length of stay* atau *los* yang relatif lebih kecil atau sedikit dibanding pelanggan *Stay*.

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

Tabel 1. Ringkasan peubah *length of stay* (hari)

	Min	Q1	Median	Q3	Maks	Rata-rata
Churn	19	265	715	1786	3225	1044
Stay	19	384	1078	2223	3250	1311

Ringkasan statistik pada Tabel 1 terdapat 33.40% pelanggan berstatus *Churn*. Berdasarkan data sebulan terakhir, pelanggan yang *Churn* memiliki waktu rata-rata untuk bertahan sepanjang 1044 hari atau 2.9 tahun. Sedangkan waktu paling sedikit untuk *stay* adalah 19 hari dan paling panjang adalah 3225 hari.



Gambar 2. Proporsi *Churn* dan *Stay* berdasarkan kategori layanan

Berdasarkan diagram batang untuk setiap peubah kategori layanan, didapatkan perbedaan tingkat *Churn* dan *Stay* yang cukup berarti pada kategori dalam layanan *rchg*, *voice*, *sms*, dan *voice_package*.

Tabel 2. Karakteristik Pelanggan *Churn* dan *Stay* Berdasarkan Layanan

Layanan	*Churn*	*Stay*
Recharge	Pelanggan yang tidak melakukan isi ulang atau *rchg_0* (56.85%).	Pelanggan yang melakukan isi ulang atau *rchg_1* (74.52%).
Voice	▪ Pelanggan yang tidak memiliki riwayat pada layanan telepon atau *voice_0* (72.04%). ▪ Pelanggan yang tidak membeli layanan telepon dan melakukan *misscall* saja atau *voice_2* (56%). ▪ Pelanggan yang membeli layanan telepon tapi melakukan *misscall* saja atau *voice_3* (62.92%).	▪ Pelanggan beruntung yang tidak membeli layanan telepon tapi dapat melakukan panggilan atau *voice_1* (72.83%). ▪ Pengguna aktif layanan telepon (74%).
Broadband	Tidak ada karakter khusus yang mewakili *Churn* maupun *Stay*, karena proporsi di setiap kategori sama besar.	
SMS	▪ Pelanggan yang tidak memiliki riwayat pada layanan SMS atau *sms_0* (64.53%).	▪ Pelanggan beruntung yang tidak membeli layanan SMS tapi dapat menggunakan layanannya atau *sms_1* (69.09%).

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

		▪ Pengguna aktif layanan SMS atau *sms_2* (67.83%).
Voice Package	Pelanggan beruntung yang tidak membeli layanan paket telepon tapi dapat menggunakan layanannya atau *voice_pack_0* (66.73%).	Pengguna aktif layanan paket telepon atau *voice_pack_1* (66.83%).

3.2. Hasil Analisis Data

Seperti langkah-langkah yang telah dijelaskan dalam prosedur analisis data, setelah dilakukan eksplorasi terhadap masing-masing peubah, kemudian dilakukan transformasi, serta penambahan peubah atau biasa disebut dengan *feature engineering*. Data dengan peubah-peubah yang baru selanjutnya dikenai beberapa metode klasifikasi, dimana sebelumnya dilakukan *penanganan imbalanced* pada data.

Cross-validation (cv) dengan *base learner* merupakan metode *cross validation* secara manual dimana fungsi dibangkitkan bukan dari *package*. Pada tahapan ini sesuai dengan hasil klasifikasi '*class*' pada bagian eksplorasi, maka '*class*' dijadikan sebagai *future engineering*. Selanjutnya dilakuan analisis menggunakan *cv base learner* dengan tiga metode klasifikasi yaitu *random forest, boosting, dan rus-boost*. Ukuran ketepatan klasifikasi dapat dilihat pada tabel 2.

Tabel 3. Ukuran Ketepatan Klasifikasi

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Geometric Mean*
Random Forest	0.75370	0.72059	0.78682	0.75319
Boosting	0.75863	0.73000	0.78726	0.75824
RUS-Boost	0.72989	0.66926	0.79052	0.72816

Berdasarkan tabel 2 dapat dilihat bahwa dari ketiga metode ini metode *boosting* memiliki rataan geometri tertinggi dibandingkan metode lainnya. Metode *boosting* memberikan ketepatan prediksi sebesar 75.8%. dapat dilihat bahwa kemampuan model dalam menduga pelanggan '*churn*' yaitu 78.7% dan kemampuan model dalam menduga pelanggan '*stay*' sebesar 73%. Dengan demikian dapat disimpulkan bahwa model terbaik pada *cross validation* dengan base learner yaitu *boosting.*

Selain dengan CV secara manual (*base learner*), dilakukan juga CV pada 5 metode sekaligus dengan package: "*SuperLearner*" yang lebih efisien untuk digunakan karena mampu mengklasifikasikan data dari beberapa metode secara bersamaan. Hasil *cross validation* dari *super learner* dapat dilihat pada tabel 3.

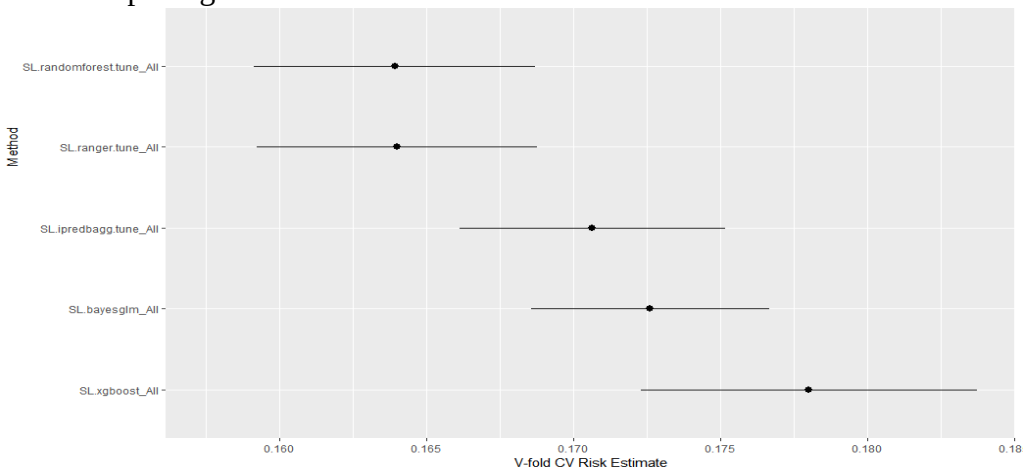
Tabel 4. Keluaran CV *Super Learner*

algoritma	**ave**	**Se**	**min**	**max**
Ranger	0.16401	0.0024358	0.15825	0.17015
Bagging	0.17065	0.0023058	0.16380	0.17643
Xgboost	0.17801	0.0029132	0.16976	0.18357
Random forest	0.16393	0.0024363	0.15799	0.16999
Bayes GLM	0.17261	0.0020627	0.16607	0.17743

Dengan melihat nilai *ave* (*average risk*), *cv super learner* menghasilkan metode terbaik yaitu *random forest* atau *Ranger*. *Ranger* sendiri adalah *package* untuk melakukan *random forest* dengan versi yang lebih cepat dan efisien. Dalam CV *super learner* dapat digunakan untuk melihat faktor resiko masing-masing metode, semakin kecil resiko artinya metode tersebut semakin baik. Berdasarkan gambar 3, dapat dilihat bahwa nilai estimasi resiko yang

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

dihasilkan, dapat dilihat bahwa rata-rata resiko terendah dimiliki oleh metode *random forest*. Kemudian dilanjutkan dengan melihat risiko setiap metode klasifikasi untuk memastikan metode tersebut dapat digunakan dalam analisis.



Gambar 3. Plot resiko *super learner*

Salah satu kemudahan *super learner* yaitu dalam hal memilih metode mana dapat digunakan, jika model tidak cocok atau tidak berkontribusi banyak, maka akan berbobot ke nol atau nilai *coeffiecient*-nya akan nol.

Tabel 5. Risiko pemilihan *base learner*

Method	Risk	Coef
SL.ranger.tune_All	0.1638571	0.1087025
SL.ipredbagg.tune_All	0.1706497	0.1502327
SL.xgboost_All	0.1742856	0.2129698
SL.randomforest.tune_All	0.1637506	0.3121086
SL.bayesGLM_All	0.1726527	0.2159865

Dalam analisi kali ini, dari 5 metode klasifikasi yang dicoba, semua algoritma dapat digunakan untuk data, tetapi random forest memiliki nilai *ave* tertinggi dibandingkan yang lain. Semakin kecil nilai resiko (*risk*), maka akan semakin baik metode itu untuk digunakan. Metode dengan nilai *coef* terbesar dan risk terkecil yang akan dipilih, dan untuk kasus ini, metode *random forest* merupakan metode terbaik yang dapat digunakan untuk mengklasifikasikan churn seorang pelanggan.

Langkah selanjutnya yaitu memilih metode terbaik dari dua tipe *cross validation tersebut*. Metode klasifikasi yang terpilih pada *cross validation* baik *base learner* maupun *super learner* dimodelkan ke data sampling dan dilakukan prediksi terhadap 81000 pelanggan untuk membandingkan hasil dari kedua cara tersebut.

Tabel 6. *Confussion Matrix* Prediksi

Model	***Type of cross-validation***	***Accuracy***	***Sensitivity***	***Specificity***	***Balance accuracy***
Random Forest	*Super learner*	0.7599	0.7393	0.8087	0.7740
Boosting	*Base learner*	0.7461	0.7296	0.7850	0.7573

Berdasarkan tabel 5 dapat dilihat bahwa *balance accuracy* tertinggi yaitu metode *random forest* dengan nilainya sebesar 0.7740. Metode random forest mampu memberikan ketepatan prediksi sebesar 75.99%. Dapat dilihat juga bahwa kemampuan model dalam

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

menduga pelanggan '*churn*' yaitu 80.1% dan kemampuan model dalam menduga pelanggan '*stay*' sebesar 73.9%. Dengan demikian dapat disimpulkan bahwa metode klasifikasi terbaik untuk data telekomunikasi status churn pelanggan yaitu *random forest* pada *cross validation* dengan *super learner*.

4. SIMPULAN

Berdasarkan pembahasan maka dapat disimpulkan bahwa metode klasifikasi terbaik yaitu *random forest* dengan menggunakan *cross validation* pada *super learner*. Nilai Akurasi yang diperoleh yaitu 75.99 % dengan kemampuan menduga pelanggan '*churn*' sangat bagus yaitu sebesar 80.1%, sedangkan dalam menduga '*stay*' bagus yaitu sebesar 73.9%. *Super learner* mampu memberikan hasil yang lebih bagus dan kecepatan komputasi yang lebih baik.

Secara umum, pelanggan yang *Churn* adalah pelanggan yang tidak memiliki riwayat transaksi atau *revenue* pada layanan *voice, SMS,* dan *voice_package*. Namun beberapa diantaranya masih memiliki rekam kegiatan dengan beberapa layanan tersebut, dimungkinkan pelanggan dengan karakter ini mendapatkan promo atau bonus karena telah membeli produk layanan yang lain.

5. DAFTAR PUSTAKA

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Ebrah, K., & Elnasir, S. (2019). Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms. *Journal of Computer and Communications*, *07*(11), 33–53. https://doi.org/10.4236/jcc.2019.711003

Google, temasek, & bain & company. (2020). *E-Conomy SEA 2020 Report*. https://economysea.withgoogle.com/

Idris, A., & Khan, A. (2012). Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers. *2012 15th International Multitopic Conference (INMIC)*, 23–27. https://doi.org/10.1109/INMIC.2012.6511498

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, *167*, 101–112. https://doi.org/10.1016/j.procs.2020.03.187

Lee, S., Nguyen, N., Karamanli, A., Lee, J., & Vo, T. P. (2022). Super learner machine-learning algorithms for compressive strength prediction of high performance concrete. *Structural Concrete*, suco.202200424. https://doi.org/10.1002/suco.202200424

Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*, *43*(2), 276–286. https://doi.org/10.1509/jmkr.43.2.276

Mahajan, V., Misra, R., & Mahajan, R. (2017). Review on factors affecting customer churn in telecom sector. *International Journal of Data Analysis Techniques and Strategies*, *9*(2), 122. https://doi.org/10.1504/IJDATS.2017.085898

Miguéis, V. L., Van den Poel, D., Camanho, A. S., & Falcão e Cunha, J. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*, *39*(12), 11250–11256. https://doi.org/10.1016/j.eswa.2012.03.073

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 5, No. 1, Januari 2023
Artikel ke-5, Halaman: 45-54

Mung, P. S., & Phyu, S. (2020). Ensemble Learning Method for Enhancing Healthcare Classification. *Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering*. 2020 the 10th International Workshop on Computer Science and Engineering. https://doi.org/10.18178/wcse.2020.02.024

Rajeswari, P. S., & Ravilochanan, P. (2014). Churn Analytics on Indian Prepaid Mobile Services. *Asian Social Science*, *10*(13), p169. https://doi.org/10.5539/ass.v10n13p169

Rizal, Y. (2017). Evaluasi Strategi Pengembangan Jaringan Telekomunikasi dengan Blue Ocean Strategy. *Jurnal Telekomunikasi dan Komputer*, *6*(1), 45. https://doi.org/10.22441/incomtech.v6i1.1148

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719. https://doi.org/10.1142/S0218001409007326

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999

Y., N. N., Ly, T. V., & Son, D. V. T. (2022). Churn prediction in telecommunication industry using kernel Support Vector Machines. *PLOS ONE*, *17*(5), e0267935. https://doi.org/10.1371/journal.pone.0267935