JAMBURA JOURNAL OF MATHEMATICS
Jambura J. Math. Vol. 5, No. 1, pp. 83–94, February 2023
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jjom
DOI: https://doi.org/10.34312/jjom.v5i1.15869

Metode *AdaBoost* dan *Random Forest* untuk Prediksi Peserta JKN-KIS yang Menunggak

Ikhlasul Amalia Rahmi[1,*], Farit Mochamad Afendi[1], Anang Kurnia[1]

[1]*Departemen Statistika, Fakultas MIPA, IPB University, Bogor 16680, Indonesia*
[*]*Corresponding author. Email: essmalia@gmail.com*

ABSTRAK

Iuran peserta, pemberi kerja, dan/atau pemerintah merupakan salah satu hal terpenting dalam penyelenggaran Program Jaminan Kesehatan Nasional-Kartu Indonesia Sehat (JKN-KIS). Seluruh penduduk Indonesia diwajibkan untuk mengikuti program JKN-KIS yang terbagi menjadi empat jenis kepesertaan, salah satunya adalah Peserta Bukan Penerima Upah (PBPU) yang pembayaran iurannya dilakukan secara mandiri. Namun berdasarkan data Desember 2021 terdapat 60% dari peserta PBPU terlambat melakukan pembayaran iuran bulanan sampai menunggak. Pembayaran iuran yang tertunggak menyebabkan beberapa masalah termasuk pembayaran klaim hingga defisit. Penelitian ini memanfaatkan *big data* yang dimiliki BPJS Kesehatan dan *machine learning* berbasis pohon gabungan yaitu *AdaBoost* dan *random forest* untuk mendapatkan prediksi peserta menunggak. Hasil penelitian menunjukkan bahwa pembelajaran mesin berbasis pohon gabungan mampu memprediksi peserta PBPU menunggak dengan tingkat akurasi tinggi, dibuktikan dengan nilai AUC pada kedua model diatas 80%. Model *random forest* memiliki F_1-*score* dan nilai AUC lebih baik dibandingkan *AdaBoost* yaitu F_1-*score* 85,43% dan nilai AUC 87,20% dalam memprediksi peserta JKN-KIS yang menunggak pembayaran iuran.

Kata Kunci:
AdaBoost; Big Data; *Machine Learning*; Program JKN-KIS; *Random Forest*

ABSTRACT

The contribution of participants, employers, and/or the government is one of the most important things in the National Health Insurance Program-Healthy Indonesia Card (JKN-KIS) implementation. All Indonesian residents were required to participate in the JKN-KIS program which is divided into four types of participation, one of which is Non-Wage Recipient Participants (PBPU) whose contributions are paid independently. However, based on December 2021 data, 60% of PBPU participants were late in paying monthly until they were in arrears. Arrears in payment of contributions cause several problems, including payment of claims to deficits. This research utilized big data owned by the Healthcare and Social Security Agency (BPJS Kesehatan) and machine learning based on ensemble trees, namely AdaBoost and random forest to get the predictions of participants in arrears. The results showed that machine learning based on an ensemble tree was able to predict PBPU participants in arrears with high accuracy, as evidenced by the AUC values in both models above 80%. The random forest model has an F_1_score and the AUC value is better than the AdaBoost, namely the F_1_score of 85,43% and the AUC value of 87,20% in predicting JKN-KIS participants who are in arrears in payment of contributions.

e-ISSN: 2656-1344
Received: 16 August 2022 | Accepted: 13 November 2022 | Online: 1 February 2023

Keywords:
AdaBoost; Big Data; Machine Learning; JKN-KIS Program; Random Forest

Format Sitasi:

I. A. Rahmi, F. M. Afendi, and A. Kurnia, "Metode AdaBoost dan Random Forest untuk Prediksi Peserta JKN-KIS yang Menunggak", *Jambura J. Math.*, vol. 5, No. 1, pp. 83–94, 2023, doi: https://doi.org/10.34312/jjom.v5i1.15869

1. Pendahuluan

Machine learning atau pembelajaran mesin digunakan dalam mengembangkan model yang secara otomatis beradaptasi untuk mengidentifikasi pola yang kompleks dan tersembunyi dalam data, sehingga dapat membantu pengambil keputusan untuk memperkirakan dampak dari beberapa skenario yang masuk akal secara *real time*. Salah satu teknik *machine learning* adalah *supervised learning*, yaitu pembelajaran terawasi dimana hasil yang diharapkan sudah diketahui sebelumnya [1]. Metode *supervised learning* yang banyak digunakan adalah metode pohon keputusan, karena mudah dipahami di mana cara menentukan keputusan mirip dengan cara berpikir manusia [2]. Terdapat 2 macam metode pohon keputusan yaitu pohon tunggal dan pohon gabungan *(ensemble tree)*.

Ensemble tree merupakan pengembangan dari pohon tunggal yang menghasilkan beberapa pohon klasifikasi dari sekumpulan data dan membuat keputusan berdasarkan hasil gabungan prediksi masing-masing pohon [3]. Dua cara yang umum dilakukan pada teknik *ensemble tree* adalah *boosting* dan *bagging*, perbedaan dari kedua model tersebut adalah cara pembentukan pohonnya. Pembentukan pohon secara *boosting* dilakukan secara sekuensial, sedangkan *bagging* pembentukan pohon dilakukan secara paralel [4]. Salah satu metode *boosting* yang sering digunakan adalah *adaptive boosting* (*AdaBoost*) dan metode *bagging* yang sering digunakan adalah *random forest*.

Revolusi Industri 4.0 tentu tidak asing lagi, salah satu teknologi yang pilar utamanya adalah data yang besar, bahkan dalam jumlah yang tidak terbatas yang dikenal dengan *big data*. Hal ini menyebabkan analisis klasik kurang mampu dalam melakukan klasifikasi *big data* dengan baik [5]. Teknik *machine learning* menjadi solusi pada analisis *big data* untuk menangkap pola-pola tak linier, sehingga dapat memberikan informasi tambahan yang umumnya gagal ditangkap oleh pendekatan model linier klasik [6]. Analisis *big data* dengan bantuan *machine learning* mampu meningkatkan layanan dan memecahkan masalah pada berbagai sektor, salah satunya sektor kesehatan.

BPJS Kesehatan sebagai pengelola jaminan kesehatan terbesar di dunia, memanfaatkan *big data* untuk mengoptimalkan pelaksanaan program Jaminan Kesehatan Nasional - Kartu Indonesia Sehat (JKN-KIS), diantaranya mengurangi readmisi, meningkatkan efisiensi layanan kesehatan, dan meningkatkan kualitas layanan. Pada laporan *Business Intelligence* (BI) BPJS Kesehatan per Desember 2021, total cakupan peserta JKN-KIS sebanyak 235,7 juta peserta yang terbagi dalam empat jenis kepesertaan. Salah satunya adalah peserta Pekerja Bukan Penerima Upah (PBPU) yang melakukan pembayaran iuran setiap bulannya secara mandiri [7].

Berdasarkan laporan BI selama satu tahun, adanya pandemi covid-19 menunjukkan peningkatan peserta PBPU yang non aktif karena menunggak, sehingga semakin banyak peserta tidak terproteksi Program JKN-KIS. Banyak faktor yang mempengaruhi

rendahnya nilai penerimaan iuran peserta PBPU, salah satunya penelitian terkait faktor-faktor yang mempengaruhi kepatuhan peserta PBPU untuk membayar iuran [8]. Pada penelitian tersebut diperoleh pekerjaan, pengetahuan, pendapatan, dan jarak memiliki kaitannya dengan kepatuhan iuran JKN.

BPJS Kesehatan telah mengupayakan beberapa cara untuk meningkatkan nilai penerimaan iuran PBPU diantaranya dengan program donasi, *SMS blast*, telekolekting maupun pendaftaran peserta PBPU yang mewajibkan pembayaran melalui auto debit. Pengiriman *SMS blast* sebagai pengingat untuk melakukan pembayaran iuran, saat ini dikirimkan untuk seluruh peserta PBPU dengan nomor *handphone* valid pada *masterfile*. Pengiriman *SMS blast* tersebut perlu dilakukan kajian terlebih lanjut agar tepat sasaran, dengan mempelajari pola peserta PBPU menunggak diharapkan dapat memberikan prediksi peserta yang melakukan tunggakan pembayaran iuran pada bulan berikutnya.

Penelitian terkait prediksi peserta menunggak pembayaran telah dilakukan terhadap 1.079 pelanggan *Thailand Provincial Electricity Authority* (PEA) pada tahun 2020 [9]. Pada penelitian tersebut terdapat dua tahapan, yaitu pengelompokkan kelas pelanggan menggunakan algoritme *k-means* dan prediksi pelanggan; menggunakan lima model *machine learning* yaitu *logistic regression, decision tree, random forests, support vector machine* (SVM) dan *extreme gradient boosted* (XGBoost). Hasil penelitiannya menunjukkan *random forest* adalah model terbaik dengan *F1-Score* 97,93% dalam memprediksi kelas pelanggan yang menunggak pembayaran listrik. Sementara itu, metode *AdaBoost* banyak diimplementasikan dalam kasus optimasi dengan melibatkan bantuan algoritma seperti pada kasus klasifikasi penyakit diabetes dengan bantuan algoritma *Naive Bayes* [10], kasus penyakit stroke dengan bantuan algoritma C4.5 [11], juga pada kasus klasifikasi penyakit diabetes [12]. Adapun beberapa riset terbaru yang menggabungkan penggunaan metode *AdaBoost* dan *random forest* dapat ditemukan pada [13–15]. Hasil riset dengan kedua metode tersebut banyak dilakukan dalam berbagai kasus sehingga menarik untuk diterapkan pada kasus lain terutama pada kasus program JKN-KIS dengan memanfaatkan *big data* yang dimiliki oleh BPJS Kesehatan.

Penelitian terkait prediksi peserta yang menunggak pembayaran iuran program JKN-KIS belum pernah dilakukan dengan menggunakan *AdaBoost* dan *random forest*. Oleh karena itu, analisis *big data* dengan bantuan *machine learning* yang tepat menggunakan metode *ensemble tree* yaitu *AdaBoost* dan *random forest* menarik untuk dilakukan, dengan tujuan untuk mempelajari pola dan memprediksi peserta PBPU yang menunggak pembayaran iuran. Hasil prediksi peserta PBPU yang menunggak diharapkan mampu memberikan *insight* tambahan kepada BPJS Kesehatan dalam pencegahan peserta menunggak.

2. Metode

Data yang digunakan merupakan jenis data sekunder dari E-PPID (Elektronik-Pejabat Pengelola Informasi dan Dokumentasi) yang dikelola oleh BPJS Kesehatan, dengan posisi data Desember 2021 sebanyak 13.417.415 data peserta. Peubah respon (Y) yang digunakan yaitu peserta PBPU yang menunggak pembayaran iuran dan peserta PBPU tidak menunggak dengan 13 peubah penjelas (X) yang disajikan pada Tabel 1.

Prosedur analisis data pada penelitian ini menggunakan bantuan *software python*, tahap praproses data dilakukan untuk mengolah data mentah menjadi bentuk data yang lebih dipahami oleh sistem. Tahap pemodelan dilakukan untuk membentuk model, mendapatkan prediksi status peserta, dan melakukan evaluasi untuk mendapatkan

Tabel 1. Daftar peubah penelitian

Peubah	Keterangan	Tipe
Y	Status peserta; 0: Peserta tidak menunggak, 1: Peserta menunggak	Nominal
X1	Kabupaten/Kota peserta terdaftar; 514 Kabupaten/Kota	Nominal
X2	Kelas rawat peserta saat ini; 1: Kelas 1, 2: Kelas 2, 3: Kelas 3	Ordinal
X3	Riwayat mutasi naik kelas; 0: Tidak, 1: Ya	Nominal
X4	Riwayat mutasi turun kelas; 0: Tidak, 1: Ya	Nominal
X5	Usia peserta; 1: $<$ 21 Tahun, 2: 21 s.d 30 Tahun, 3: 31 s.d 40 Tahun, 4: 41 s.d 50 Tahun, 5: 50 s.d 60 Tahun, 6: $>$ 60 Tahun	Ordinal
X6	Jumlah tertanggung; 1: Tidak ada, 2: 1 anggota keluarga, 3: 2 atau lebih orang anggota keluarga	Nominal
X7	Kepemilikan COB atau asuransi lain; 0: Tidak, 1: Ya	Nominal
X8	Pembayaran melalui auto debit; 0: Tidak, 1: Ya	Nominal
X9	Riwayat kunjungan FKTP di bulan berjalan; 0: Tidak, 1: Ya	Nominal
X10	Riwayat kunjungan FKRTL di bulan berjalan; 0: Tidak, 1: Ya	Nominal
X11	Riwayat mutasi jenis kepesertaan; 0: Tidak, 1: Ya	Nominal
X12	Lama kepesertaan; 1: $<$ 1 Tahun, 2: 1 s.d 5 Tahun, 3: $>$ 5 Tahun	Nominal

model yang terbaik. Prosedur analisis data yang dilakukan adalah sebagai berikut:

1. **Praproses Data**
   (a) Persiapan data dengan import *big data* peserta JKN-KIS
   (b) *Exploratory data analysis*
   (c) Melakukan *feature engineering*
   Melakukan kategorisasi pada peubah usia, peubah jumlah tertanggung, dan peubah lama kepesertaan.
   (d) Melakukan sampling data
   Terdapat dua kategori hubungan keluarga yaitu peserta (P) dan anggota keluarga yang mencakup (suami/istri, anak, dan keluarga lain), di mana total kategori peserta (P) sebanyak $\pm$13 juta peserta PBPU Program JKN-KIS. Ukuran data atau populasi tersebut tergolong besar dan setelah dilakukan eksplorasi data terlihat bahwa objek penelitian termasuk homogen, sehingga pada penelitian ini digunakan sampling data agar menghemat waktu dan biaya. Sampling data yang digunakan sebanyak 50% dari keseluruan data dengan strata peubah penjelas X1 yaitu 514 kabupaten/kota dan strata peubah respon Y yaitu status peserta, total data yang digunakan untuk pemodelan 6.663.328 data peserta.
   (e) Membagi data dan penanganan ketidakseimbangan data
      i. Melakukan pembagian data menjadi 70% data latih dan 30% data uji.
      ii. Melakukan penanganan ketidakseimbangan data pada data latih menggunakan metode *syntetic minority oversampling* (SMOTE). SMOTE merupakan metode *oversampling* dengan membangkitkan data sintetik atau data buatan pada kelas minoritas berdasarkan algoritme k-tetangga terdekat, sehingga proporsi data antar kelas menjadi lebih seimbang [16].
2. **Pemodelan**
   (a) Membangun model klasifikasi *AdaBoost* dan *random forest* pada data latih dengan mencari *hyperparameter* yang optimal secara *grid search* dengan validasi silang lipat-5, yang bertujuan mengevaluasi kinerja model sebanyak lima pengulangan selama pencarian *hyperparameter* pada *grid* untuk setiap parameter.

(b) Parameter yang dilakukan optimasi adalah sebagai berikut:
   i. Model *AdaBoost*
      - *n_estimators :* jumlah pohon
      - *learning_rate :* nilai bobot kontribusi pada classifier
   ii. Model *random forest*
      - *n_estimators :* jumlah pohon
      - *max_features :* jumlah peubah penjelas yang dipertimbangkan saat mencari split terbaik
      - *min_samples_leaf:* Jumlah minimum sampel pada simpul daun

(c) Evaluasi dan validasi model
Evaluasi dan validasi model dilakukan untuk mengukur kebaikan kinerja model dalam mengklasifikasikan kelas dengan benar yang diukur menggunakan *confusion matrix*. *Confusion matrix* merupakan tabulasi silang antara data kelas positif dan kelas negatif yang masuk dalam kelas prediksi dan kelas aktual [17]. Pada penelitian ini, kelas positif adalah kelas status peserta menunggak dan kelas negatif adalah kelas tidak menunggak. Berdasarkan *confusion matrix* tersebut, dilakukan perbandingan nilai performa masing-masing model yang meliputi:
   i. Spesifisitas
   Menggambarkan proporsi data kelas negatif yang diprediksi ke kelas negatif, semakin baik kinerja model klasifikasi maka nilai spesifisitas mendekati 1
   ii. *Recall* atau sensitivitas
   Menggambarkan proporsi data kelas positif yang diprediksi ke kelas positif, semakin baik kinerja model klasifikasi maka nilai sensitivitas mendekati 1
   iii. *Precision* atau presisi
   Menggambarkan proporsi data kelas positif yang diklasifikasikan secara benar dibagi dengan total data yang diklasifikasi positif, semakin baik kinerja model klasifikasi maka nilai presisi mendekati 1
   iv. *F1-score*
   Perbandingan rata-rata dari nilai presisi dan nilai sensitivitas, di mana rentang nilai F_1adalah antara 0 hingga 1, semakin baik kinerja model klasifikasi maka nilai F_1 mendekati 1
   v. Nilai AUC
   AUC (*area under curve)* merupakan daerah di bawah kurva ROC. Nilai AUC memiliki rentang antara 0,5 sampai dengan 1. Interpretasi nilai AUC dapat diklasifikasikan menjadi lima bagian yang berbeda, yaitu 0,5 – 0,6 (akurasi salah), 0,6 – 0,7 (tingkat akurasi lemah), 0,7 – 0,8 (tingkat akurasi sedang), 0,8 – 0,9 (tingkat akurasi tinggi), dan 0,9 – 1 (tingkat akurasi sangat tinggi).

2.1. Adaptive Boosting (AdaBoost)

Boosting diperkenalkan oleh Freund dan Schapire tahun 1995 melalui algoritme *AdaBoost* dengan konsep dasar peningkatan bobot pengamatan yang salah klasifikasi [18]. Algoritme *AdaBoost* membangun model pohon gabungan secara sekuensial, yaitu pada setiap iterasi, bobot data dimodifikasi dengan tujuan mengoreksi data yang salah klasifikasi pada iterasi sebelumnya [4]. Data yang salah diklasifikasikan menerima

bobot lebih besar dibandings data yang benar. Label kelas ditentukan dari semua pengamatan model yang dibangun kemudian dipilih dengan *voting* dan prediksi data baru didasarkan pada bobot mayoritas, tahapan algoritme *AdaBoost* sebagai berikut:

1. Suatu pasangan contoh data (x_i, y_i) dengan $x_i \in X$, $y_i \in Y = 0,1$ dan $i = 1,2,\ldots,m$.
2. Jika bobot pada contoh ke-i dan iterasi ke-t dinyatakan dengan $D_t(i)$ maka bobot awal menggunakan Persamaan (1):

$$D_t(i) = \frac{1}{m}. \tag{1}$$

3. Untuk iterasi $t = 1,2,\ldots.T$ lakukan langkah berikut:
   (a) Suatu *weak learner* atau model lemah berikan bobot D_t
   (b) *Weak learner* mencari hipotesis $h_t : X \rightarrow \{0,1\}$ yang meminimumkan error dengan Persamaan (2):

$$\in_t = Pr_{i\sim D_t}\left[h_t(x_i) \neq y_i\right] = \sum_{i:h_t(x_i)\neq y_i} D_t(i). \tag{2}$$

   (c) Pilih

$$a_t = \frac{1}{2}\ln\left(\frac{1-\in_t}{\in_t}\right). \tag{3}$$

   (d) Perbaharui besar bobot dengan Persamaan (4):

$$D_{t+1}(i) = \frac{D_t(i)}{Zt} \times \begin{matrix} e^{-a_t}\ jika\ h_t(x_i) = y_i \\ e^{a_t}\ jika\ h_t(x_i) \neq y_i \end{matrix} \tag{4}$$

   dengan Zt adalah faktor normalisasi.
4. Persamaan (5) adalah *output* dari hipotesis akhir

$$H(x) = sign\left(\sum_{t=1}^{T} a_t h_t(x)\right). \tag{5}$$

2.2. Random forest

Random forest merupakan pemodelan klasifikasi pohon gabungan pengembangan dari pohon klasifikasi tunggal yang menerapkan *bootstrap aggregating* (*bagging*) dan *random feature selection* [19]. Cara kerjanya dengan membangun beberapa pohon klasifikasi secara paralel dan menentukan prediksi berdasarkan suara terbanyak (*majority vote*). Tahapan pembuatan model klasifikasi *random forest* dengan n observasi dan p peubah penjelas adalah sebagai berikut:

1. Proses *bootstrap* adalah menarik sampel acak berukuran n dengan pemulihan pada data latih.
2. Membangun pohon klasifikasi tunggal menggunakan data latih baru yang dihasilkan dari proses *bootstrap*. Pembangunan pohon klasifikasi dilakukan dengan menerapkan *random feature selection*, yaitu memilih peubah penjelas secara acak dengan $m < p$. Kemudian dari m peubah penjelas dipilih peubah penjelas terbaik sebagai pemisah dan dilanjutkan dengan pemisahan menjadi dua simpul baru. Proses ini terus berlanjut sampai ukuran minimum dari pengamatan pada

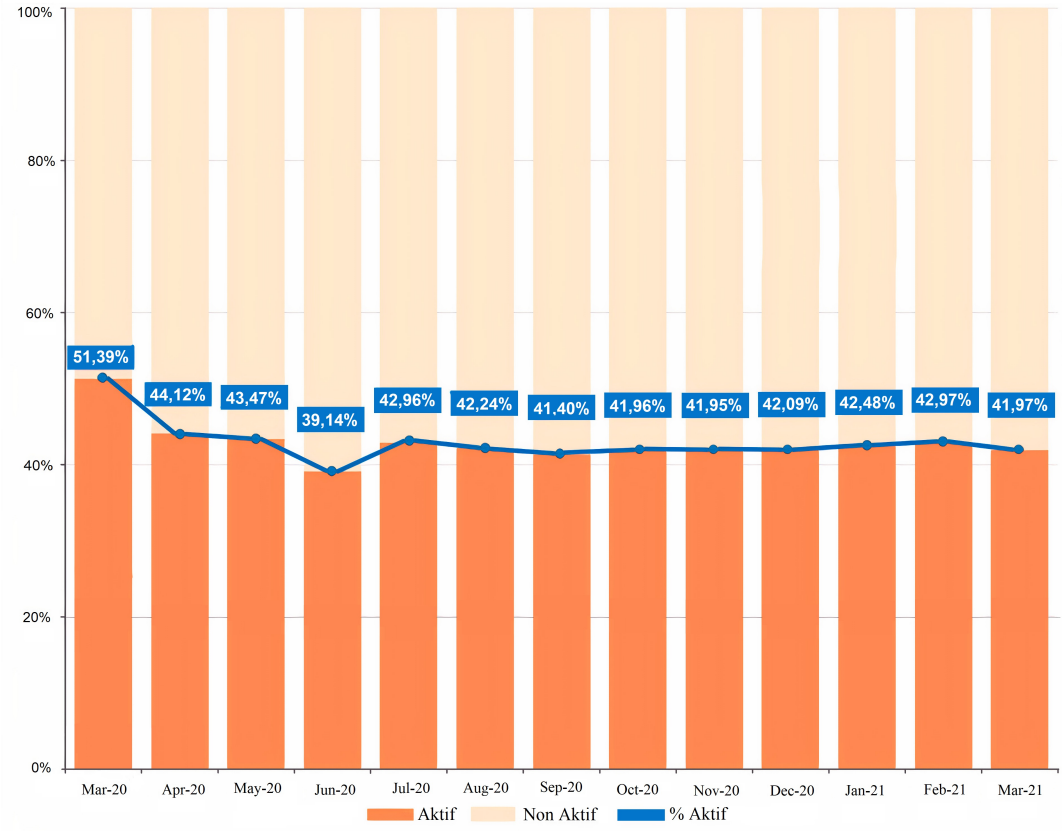
simpul tercapai. Nilai m yang direkomendasikan yaitu $\sqrt{p}$, namun perlu mengoptimalkan parameter terhadap nilai m untuk mendapatkan hasil terbaik [20].

3. Mengulangi tahapan 1 dan 2 sebanyak k kali untuk mendapatkan k pohon klasifikasi. Setiap pohon klasifikasi menghasilkan satu suara dan kelas klasifikasi ditentukan oleh suara terbanyak dari k buah suara.

3. Hasil dan Pembahasan

3.1. Praproses Data

Pandemi covid-19 yang terjadi di Indonesia mulai Maret 2020, berdampak pada penurunan persentase jumlah peserta yang aktif yang disajikan pada Gambar 1. Pada gambar tersebut terlihat bulan Maret 2020 persentase peserta yang aktif sebesar 51,39% dibandingkan dengan peserta yang non aktif, terjadi penurunan hingga 9,42% pada Maret 2021. Secara umum terlihat status peseta non aktif lebih besar dengan status aktif dan selama satu tahun pandemi covid-19 terjadi penurunan terhadap status peserta yang aktif setiap bulannya.



Gambar 1. Status keaktifan peserta PBPU selama satu tahun pandemi covid-19

Pada Gambar 1 terdapat dua kategori peserta yaitu peserta aktif (peserta tidak menunggak) dan peserta non aktif (peserta menunggak), yang pada penelitian ini digunakan sebagai peubah respon atau kelas peserta. Peserta dengan kelas tidak menunggak adalah peserta PBPU yang tidak memiliki bulan tunggakan, sedangkan kelas peserta menunggak adalah peserta PBPU yang tidak membayar iuran sampai

dengan akhir bulan berjalan. Berdasarkan data sampai dengan Desember 2021 perbandingan proporsi kelas peserta menunggak lebih besar yakni ±8.8 juta (66,06%) peserta, dibandingkan peserta tidak menunggak sebesar ±4.5 juta (33,94%).

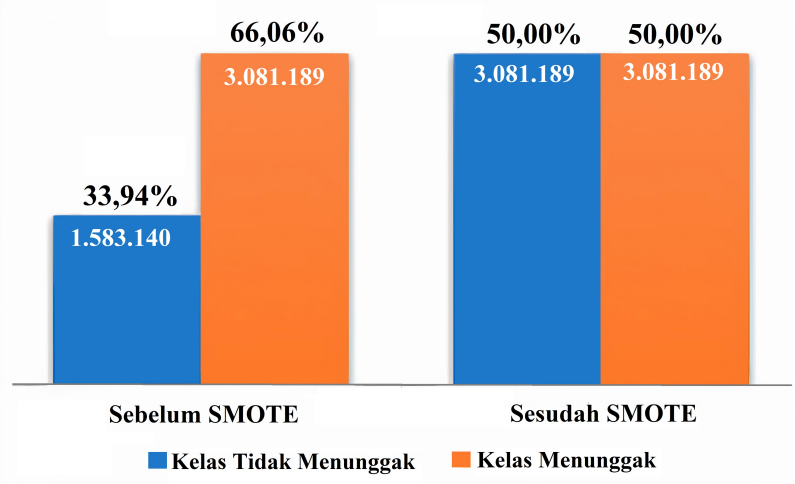
3.2. Pemodelan

Pada proses pemodelan digunakan 50% data sampling, dengan menerapkan *stratified random sampling* (pengambilan sampel acak berstrata) berdasarkan strata peubah penjelas X1 (kabupaten/kota peserta terdaftar) dan peubah respon Y (status peserta). Sampling tersebut memanfaatkan 514 kabupaten/kota peserta terdaftar dan status peserta. Sehingga sampling memuat semua kabupaten/kota, dengan jumlah sampel 50% dengan status menunggak dan 50% dengan status tidak menunggak pada masing-masing kabupaten/kota. Selanjutnya data hasil sampling tersebut dibagi menjadi 2 bagian yaitu 70% data latih dan 30% data uji, proporsi pembagian kelas pada data latih dan data uji dapat disajikan pada Table 2.

Tabel 2. Proporsi kelas data latih dan data uji

Kelas Peserta	Dataset peserta PBPU (hasil sampling)		
	Data Latih	Data Uji	(%)
Kelas Tidak Menunggak	1.583.140	678.438	33,94
Kelas Menunggak	3.081.189	1.320.561	66,06
Jumlah	4.664.329	1.998.999	

Proporsi kelas peserta pada Table 2 menujukkan adanya ketidakseimbangan kelas yang dapat berpengaruh terhadap hasil prediksi, karena menyebabkan model hanya baik ketika digunakan memprediksi kelas mayoritas. Oleh karena itu, dilakukan penanganan ketidakseimbangan kelas agar model yang dihasilkan lebih optimal menggunakan teknik SMOTE pada data latih. Gambar 2 merupakan diagram baris yang menyajikan perbedaan proporsi kelas peserta pada data latih sebelum dan sesudah dilakukan penanganan ketidakseimbangan kelas, terlihat proporsi data setelah dilakukan SMOTE menjadi seimbang, karena adanya proses pembuatan data sintetis pada kelas minoritas yaitu kelas tidak menunggak.



Gambar 2. Proporsi kelas peserta sebelum dan sesudah SMOTE pada data latih

Data latih yang telah dilakukan penanganan ketidakseimbangan kelas, selanjutnya

dilakukan proses pemodelan dengan peubah respon (status peserta) dan 11 peubah penjelas yaitu kelas rawat peserta, usia, jumlah tertanggung, kepemilikan asuransi lain, pembayaran melalui auto debit, riwayat mutasi naik kelas, riwayat mutasi turun kelas, riwayat kunjungan di FKTP, riwayat kunjungan di FKRTL, riwayat mutasi jenis kepesertaan, dan lama kepesertaan. Model klasifikasi yang digunakan dalam memprediksi peserta menunggak adalah *AdaBoost* dan *random forest*.

Pemodelan dilakukan pada data latih dengan menentukan *hyperparameter* tertentu secara *grid search,* yaitu melakukan pencarian secara menyeluruh terhadap parameter yang di ujikan. Optimasi *hyperparameter* dilakukan secara *grid search* menggunakan validasi silang lipat-5 yang berfungsi untuk mengevaluasi kinerja model sebanyak lima kali perulangan dalam proses *grid search* dari setiap parameter.

Pada pemodelan *AdaBoost*, *hyperparameter* yang ditentukan adalah *n_estimator* yaitu jumlah pohon dan *learning_rate* yaitu nilai bobot kontribusi. Pada pemodelan *random forest*, *hyperparameter* yang ditentukan adalah jumlah pohon (*n_estimator)*, *max_feature* yaitu jumlah peubah penjelas yang perlu dipertimbangkan saat mencari split terbaik, dan *min_sample_leaf* yaitu jumlah minimum sampel yang diperlukan untuk berada di simpul daun. Hasil optimasi *hyperparameter* terbaik yang telah dilakukan pada data latih secara *grid search* pada masing-masing model disajikan pada Tabel 3 dengan dan nilai mean AUC model *AdaBoost* sebesar 0,858991 dan nilai *mean* AUC model *random forest* sebesar 0,872070.

Tabel 3. Hasil optimasi *hyperparameter* secara *grid search*

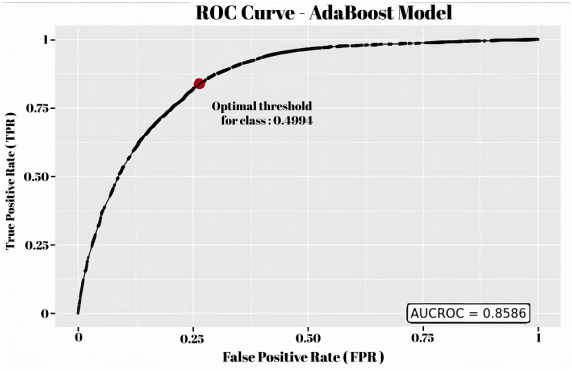
Model	*Hyperparameter*	Nilai *Hyperparameter*	*Hyperparameter* terbaik	*Mean* AUC
AdaBoost	*n_estimator*	50; 100; 350; 500	500	0,858991
	learning_rate	0,0001; 0,001; 0,01; 0,1; 1	1	
Random Forest	*n_estimator*	50; 100; 350; 500	500	0,872070
	max_feature	log2; auto; sqrt	Log2	
	min_sample_leaf	2; 8; 10; 20	20	

3.2.1. Hasil Pemodelan *AdaBoost*

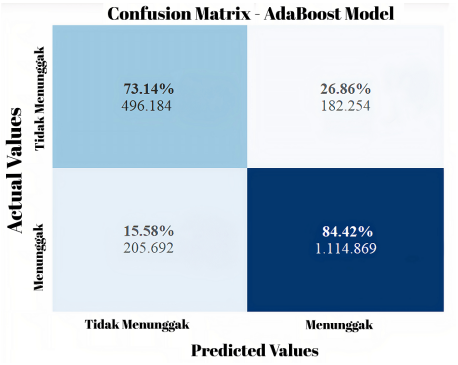
Berdasarkan model *AdaBoost* dengan *hyperparameter* terbaik, dilakukan evaluasi untuk mengukur kebaikan model pada data uji. Gambar 3.a menunjukkan kurva ROC antara nilai *false positive rate* (sumbu x) sebesar 0,268 dengan *true positive rate* (sumbu y) sebesar 0,844 diperoleh titik potong optimal adalah 0,499. Titik potong optimal digunakan untuk menentukan kelas peserta, jika hasil prediksi diatas 0,499, maka kelas peserta adalah kelas peserta menunggak selain itu kelas tidak menunggak. Berdasarkan kurva ROC diperoleh nilai AUC sebesar 0,8586 yang artinya hasil klasifikasi memiliki tingkat akurasi tinggi. Gambar 3.b menyajikan *confusion matrix* berdasarkan titik potong optimal, dengan nilai spesifisitas sebesar 73,14%, nilai sensitivitas sebesar 84,42%, nilai presisi sebesar 85,95% dan F_1-score sebesar 85,18%.

3.2.2. Hasil Pemodelan *Random forest*

Berdasarkan model *random forest* dengan *hyperparameter* terbaik, dilakukan evaluasi untuk mengukur kebaikan model pada data uji. Gambar 4.a menunjukkan kurva ROC antara nilai *false positive rate* (sumbu x) sebesar 0,252 dengan *true positive rate* (sumbu y) sebesar 0,842 diperoleh titik potong optimal adalah 0,5185 yang digunakan untuk



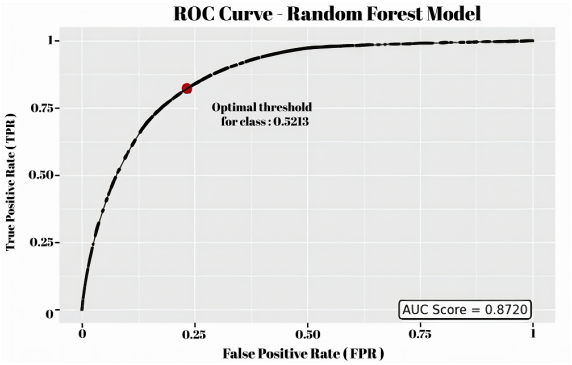
(a) Kurva ROC model *AdaBoost*



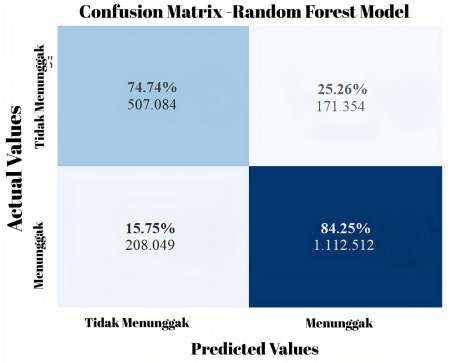
(b) *Confusion matrix*

Gambar 3. Hasil pemodelan *AdaBoost*

menentukan kelas peserta, jika hasil prediksi diatas 0,5185, maka kelas peserta adalah kelas peserta menunggak selain itu kelas tidak menunggak. Berdasarkan kurva ROC diperoleh nilai AUC sebesar 0,8720 yang artinya hasil klasifikasi memiliki tingkat akurasi tinggi. Gambar 4.b menyajikan *confusion matrix* berdasarkan titik potong optimal, diperoleh nilai spesifisitas sebesar 74,74% dan nilai sensitivitas 84,25, nilai presisi sebesar 86,65% dan F_1-score sebesar 85,43%.



(a) Kurva ROC model *Random forest*



(b) *Confusion matrix*

Gambar 4. Hasil pemodelan *Random forest*

3.3. Perbandingan Kinerja Pemodelan

Proses evaluasi hasil pemodelan dilakukan untuk membandingkan kebaikan kinerja antara kedua model, untuk mendapatkan model terbaik berdasarkan *confusion matrix* yang terbentuk. Perbandingan kebaikan model tersebut disajikan pada Tabel 4 yang terdiri dari persentase nilai presisi, sensitivitas, F1-Score dan Nilai AUC.

Tabel 4. Perbandingan kinerja model menggunakan *hyperparamater* terbaik

Model	Presisi (%)	Sensitivitas (%)	F_1-Score (%)	AUC (%)
AdaBoost	85,95	**84,42**	85,18	85,86
Random forest	**86,65**	84,24	**85,43**	**87,20**

Nilai presisi terbaik adalah model *random forest* sebesar 86,65% yakni persentase peserta PBPU yang sebenarnya menunggak dari keseluruan peserta PBPU yang diprediksi menunggak, sedangkan nilai sensitivitas terbaik adalah model *AdaBoost* sebesar 84,75%

yakni persentase peserta PBPU yang diprediksi menunggak dibandingkan keseluruan peserta PBPU yang sebenarnya menunggak. F_1-score terbaik adalah model *random forest* dengan nilai 85,43% yang artinya perbandingan rata-rata antara presisi dan sensitivitas mendekati nilai 100%. Evaluasi kebaikan model selanjutnya adalah nilai AUC, pada kedua model diperoleh hasil nilai AUC diatas 80%, yang artinya kedua model memiliki akurasi tinggi yaitu AUC rentang 0,8-0,9. Berdasarkan perbandingan nilai AUC tersebut diperoleh model klasifikasi terbaik adalah *random forest* dengan nilai AUC 87,20%.

4. Kesimpulan

Berdasarkan hasil penelitian dapat disimpulkan bahwa teknik *ensemble* mampu memprediksi peserta PBPU menunggak dengan tingkat akurasi tinggi, dibuktikan dengan nilai AUC pada kedua model diatas 80%. Model *random forest* memiliki F_1-*score* dan nilai AUC lebih baik dibandingkan *AdaBoost*. Pada proses pemodelannya, *random forest* membangun setiap pohon secara independen menggunakan sampel data secara acak, pengacakan ini membuat model menjadi lebih resisten dan mengurangi *overfitting* terhadap data latih.

Referensi

[1] J. Pustejovsky and A. Stubbs, "Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications." California: O'Reilly Media, 2012.

[2] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 10, pp. 74–78, oct 2018, doi: 10.26438/ijcse/v6i10.7478.

[3] B. Sartono, "Tinjauan terhadap Keunggulan Pohon Klasifikasi Ensemble untuk Memperbaiki Kemampuan Prediksi Pohon Klasifikasi Tunggal," *BIAStatistics*, vol. 9, no. 2, pp. 33–38, 2015, doi: 10.1234/bias.v9i2.57.

[4] T. Mildenberger, "Stephen Marsland: Machine learning. An algorithmic perspective," *Statistical Papers*, vol. 55, no. 2, pp. 575–576, may 2014, doi: 10.1007/s00362-012-0471-0.

[5] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification*, ser. Integrated Series in Information Systems. Boston, MA: Springer US, 2016, vol. 36, doi: 10.1007/978-1-4899-7641-3.

[6] S. Mullainathan and J. Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, may 2017, doi: 10.1257/jep.31.2.87.

[7] BPJS Kesehatan, "Cakupan Peserta Program JKN-KIS," 2021.

[8] F. V. Marpaung, M. Nyorong, and T. Moriza, "Factors Affecting the Compliance of National Health Insurance Participants Segment of Non-Wage Recipients in Paying the Contributions," *Journal La Medihealtico*, vol. 3, no. 3, pp. 171–179, may 2022, doi: 10.37899/journallamedihealtico.v3i3.656.

[9] P. Khansong, J. Karnjana, S. Laitrakun, and S. Usanavasin, "Customer Service Improvement based on Electricity Payment Behaviors Analysis using Data Mining Approaches," in *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*. IEEE, mar 2020, pp. 114–117, doi: 10.1109/ECTIDAMTNCON48261.2020.9090699.

[10] L. Pebrianti, F. Aulia, H. Nisa, and K. Saputra, "Implementasi Metode Adaboost untuk Mengoptimasi Klasifikasi Penyakit Diabetes dengan Algoritma Naive Bayes," *JUSTINDO: Jurnal Sistem dan Teknologi Informasi Indonesia*, vol. 7, no. 2, pp. 122–127, 2022, doi: https://doi.org/10.32528/justindo.v7i2.8627.

[11] N. D. Saputri, K. Khalid, and D. Rolliawati, "Comparison of Bagging and Adaboost Methods on C4.5 Algorithm for Stroke Prediction," *SISTEMASI*, vol. 11, no. 3, pp. 567–577, sep 2022, doi: 10.32520/stmsi.v11i3.1684.

[12] G. Abdurrahman, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Adaboost Classifier," *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)*, vol. 7, no. 1, pp. 59–66, mar 2022, doi: 10.32528/justindo.v7i1.4949.
[13] D. A. Jadhav, "An enhanced and secured predictive model of Ada-Boost and Random-Forest techniques in HCV detections," *Materials Today: Proceedings*, vol. 51, pp. 186–195, 2022, doi: 10.1016/j.matpr.2021.05.071.
[14] R. Natras, B. Soja, and M. Schmidt, "Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting," *Remote Sensing*, vol. 14, no. 15, p. 3547, jul 2022, doi: 10.3390/rs14153547.
[15] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach," 2022, pp. 219–244, doi: 10.1007/978-981-16-2164-2_19.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002, doi: 10.1613/jair.953.
[17] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, 2013, doi: 10.1007/978-1-4614-6849-3.
[18] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *ICML'96: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, 1996, pp. 148–156.
[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY: Springer New York, 2009, doi: 10.1007/978-0-387-84858-7.