Xplore (2022), Vol. 11, No. 2, 168-177
DOI: https://doi.org/10.29244/xplore.v11i2.929

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Klasifikasi Status Keaktifan Siswa SMA di Jawa Barat Menggunakan Random Forest dengan SMOTE*

M. Itmamurohman[1], Pika Silvianti [2‡], and La Ode Abdul Rahman[3]

[123]Department of Statistics, IPB University, Indonesia
‡corresponding author: pikasilvianti@apps.ipb.ac.id

Abstract

The dropout rate in Indonesia has a higher percentage as education levels grow. The high school dropout rate in Indonesia is at 0.67%. West Java is the province with the highest high school dropout rate in the academic year 2017/2018. In the next academic year, the high school dropout rate in West Java decreased. The student who drop out of school was caused by various factors. This study examines important variables and classification performance that are generated by random forest. The number of dropout students is very small compared to the number of active students. The imbalance data is handled using SMOTE. Random forest with SMOTE is considered able to predict data classes better because it can increase sensitivity values and reduce errors in classifying dropout students as active students. Father's income, number of siblings, class, father's education level, and father's type of work are important variables that have a major influence in determining the active status of high school students in West Java.

Keywords: classification, drop out of school, important variable, random forest, SMOTE.

1. Pendahuluan

1.1 Latar Belakang

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia (Kemendikbudristek) menyatakan bahwa pada tahun ajaran 2017/2018 angka putus sekolah di Indonesia jenjang SD sebesar 0.13%, jenjang SMP sebesar 0.50%, dan jenjang SMA sebesar 0.67%. Angka putus sekolah terbesar terjadi pada jenjang SMA dan provinsi Jawa Barat memiliki angka siswa putus sekolah jenjang SMA sebesar 0.74%. Jawa Barat menjadi provinsi dengan angka putus sekolah jenjang SMA terbesar di Indonesia. Angka putus sekolah jenjang SMA di Jawa Barat pada tahun ajaran 2018/2019 turun menjadi 0.17%. Jika dibandingkan dengan tahun ajaran

* Received: Feb 2022; Reviewed: Feb 2022; Published: May 2022

sebelumnya, Provinsi Jawa Barat telah berhasil menurunkan angka putus sekolah jenjang SMA sebesar 0.57%. Angka partisipasi kasar dan murni Jawa Barat untuk jenjang SMA mengalami kenaikan dari 81.25% dan 60.64% menjadi 83.81% dan 64.61%. Penurunan angka putus sekolah dan kenaikan angka partisipasi tidak lepas dari upaya keras pemerintah Provinsi Jawa Barat dalam menanggulangi kasus putus sekolah.

Terdapat banyak faktor yang menyebabkan terjadinya kasus putus sekolah. Penelitian terkait faktor-faktor penyebab kasus putus sekolah pernah dilakukan oleh beberapa peneliti terdahulu. Kamsihyati (2016) menyatakan bahwa faktor penyebab anak putus sekolah di Desa Jangrana, Kecamatan Kesugihan, Kabupaten Cilacap adalah besarnya jumlah anak yang menjadi tanggungan orang tua, rendahnya ekonomi keluarga, rendahnya pendidikan di lingkungan tempat tinggal, dan turunnya minat anak untuk belajar karena lebih memilih bekerja. Menurut Sugianto (2017), faktor utama penyebab anak putus sekolah di Desa Bukit Lipai, Kecamatan Batang, Kabupaten Inderagiri Hulu adalah faktor ekonomi keluarga yang tidak mampu sehingga berakibat pada faktor-faktor non ekonomi lainnya seperti rendahnya pendidikan orang tua.

Pemodelan klasifikasi dilakukan pada kasus siswa putus sekolah jenjang SMA di Jawa Barat guna mengimplementasikan program penanggulangan kasus putus sekolah. Klasifikasi merupakan analisis statistika yang digunakan untuk menilai objek atau mengelompokkan suatu objek ke dalam kelompok tertentu berdasarkan nilai atributnya. Salah satu metode klasifikasi yang dapat digunakan adalah *random forest*. *Random forest* merupakan salah satu metode pohon gabungan yang memanfaatkan beberapa pohon tunggal untuk melakukan dugaan. Sartono dan Syafitri (2010) menyatakan bahwa metode pohon gabungan seperti *random forest* mampu bekerja lebih baik dan memberikan kinerja yang lebih tinggi akurasinya dibandingkan metode pohon tunggal. Salah satu yang menjadi kekurangan adalah visualisasi hasil dan interpretasi. Penelitian terdahulu terkait dengan *random forest* diantaranya adalah penelitian Albasia (2018) mengenai klasifikasi keberhasilan individu dalam melanjutkan pendidikan jenjang SMA di Banten menggunakan *Classification and Regression Tree* (CART) dan *random forest* dengan hasil bahwa random forest menghasilkan akurasi yang lebih baik dari CART. Zhou (2020) melakukan perbandingan *random forest* dan *decision tree* pada klasifikasi kecelakaan di perlintasan kereta api dengan hasil bahwa kinerja *random forest* lebih baik dari *decision tree*. Syukron *et al*. (2020) melakukan perbandingan *random forest* dan *XGboost* pada klasifikasi tingkat penyakit hepatitis C dengan hasil bahwa kinerja *random forest* lebih baik dari *XGboost*.

Jumlah siswa putus sekolah yang sangat sedikit dibandingkan jumlah siswa aktif sekolah menyebabkan data menjadi tidak seimbang. Kelas data yang objeknya lebih banyak disebut kelas mayor sedangkan lainnya disebut kelas minor. Permasalahan yang akan muncul pada kasus data tidak seimbang adalah kelas minor akan menghasilkan ketepatan prediksi yang rendah. Salah satu teknik yang dapat digunakan untuk menangani ketidakseimbangan data tersebut adalah *Synthetic Minority Oversampling Technique* (SMOTE). Penelitian terdahulu mengenai SMOTE diantaranya adalah penelitian Purnajaya dan Hanggara (2021) membandingkan beberapa teknik sampling untuk menangani ketidakseimbangan data pada klasifikasi

pasien covid-19 menggunakan *support vector machine* dengan hasil bahwa SMOTE menghasilkan kinerja yang lebih baik dari *Random Undersampling* (RUS), *Random Oversampling* (ROS), *Combine Over-Undersampling* (COUS), dan *Tomek Link*. Ubaya dan Juariah (2020) mengenai perbandingan kinerja *random forest* dengan RUS dan *random forest* dengan SMOTE dalam mengklasifikasikan data *spam twitter* dengan hasil SMOTE menghasilkan kinerja yang lebih baik dari RUS dalam menangani ketidakseimbangan data.

1.2 Tujuan

Tujuan penelitian ini adalah melakukan klasifikasi status keaktifan siswa SMA di Jawa Barat menggunakanmetode *random forest* dengan SMOTE, membandingkan kinerja metode *random forest* tanpa SMOTE dan metode *random forest* dengan SMOTE, dan mengidentifikasi peubah penjelas penting yang berhubungan dengan status keaktifan siswa SMA di Jawa Barat.

2. Metodologi

2.1 Bahan dan Data

Data yang digunakan dalam penelitian ini merupakan data seluruh siswa SMA di Jawa Barat tahun ajaran 2019/2020 periode Juli 2019. Data tersebut merupakan Data Pokok Pendidikan (Dapodik) yang diperoleh dari Sekretariat Direktorat Jenderal Pendidikan Dasar dan Menengah, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia (Kemendikbudristek). Terdapat 18 peubah yang digunakan pada penelitian ini, yaitu 1 peubah respon kategorik (Y) dan 17 peubah penjelas (X). Peubah respon yang digunakan terdiri dari 2 kategori yaitu putus sekolah dan aktif sekolah. Peubah penjelas terdiri dari 16 peubah kategorik dan 1 peubah numerik seperti pada Tabel 1.

Tabel 1 Daftar peubah

Kode	Peubah	Skala	Kode	Peubah	Skala
Y	Status Keaktifan	Nominal	X9	Jenis Pekerjaan Ibu	Nominal
X1	Jenis Kelamin	Nominal	10	Penghasilan Ayah	Ordinal
X2	Jenis Tinggal	Nominal	X11	Penghasilan Ibu	Ordinal
X3	Alat Transportasi	Nominal	X12	Jarak Tempat Tinggal	Ordinal
X4	Penerima KPS/PKH	Nominal	X13	Jumlah Saudara Kandung	Rasio
X5	Penerima KIP	Nominal	X14	Kelas	Ordinal
X6	Jenjang Pendidikan Ayah	Ordinal	X15	Jenis Beasiswa	Nominal
X7	Jenjang Pendidikan Ibu	Ordinal	X16	Jenis Prestasi	Nominal
X8	Jenis Pekerjaan Ayah	Nominal	X17	Jenis Sekolah	Nominal

2.2 Metode Penelitian

Tahapan analisis pada penelitian ini dijelaskan sebagai berikut:

1. Praposes data dengan menyeleksi data siswa yang tidak memiliki data lengkap.
2. Melakukan *cleaning* data dikarenakan data masih dalam bentuk data mentah.
3. Melakukan eksplorasi data untuk mengetahui gambaran umum data siswa.

4. Membagi data menjadi dua bagian yaitu 80% data latih dan 20% data uji.
5. Melakukan klasifikasi *random forest* tanpa SMOTE, tahapannya sebagai berikut:
   a. Membangun *random forest* tanpa SMOTE dengan menggunakan *m* peubah penjelas sebesar 2, 4, dan 8. Nilai *k* pohon yang dicobakan adalah 25, 50, 100, dan 200. Tahapan ini dilakukan sebanyak 100 kali ulangan.
   b. Menganalisis tingkat rataan misklasifikasi secara eksploratif untuk menentukan kombinasi *m* dan *k* yang digunakan sehingga mendapatkan *random forest* dengan rataan misklasifikasi terkecil.
   c. Mengevaluasi kinerja klasifikasi dengan melihat rataan nilai akurasi, sensitivitas, dan spesifisitasnya.
6. Melakukan klasifikasi *random forest* dengan SMOTE, tahapannya sebagai berikut:
   a. Melakukan penanganan data tidak seimbang pada data latih menggunakan teknik SMOTE dengan nilai *k*=5, *oversampling*=700 dan *undersampling*=1.
   b. Membangun *random forest* menggunakan kombinasi *m* dan *k* yang telah ditentukan pada tahapan 5b. Tahapan ini dilakukan sebanyak 100 kali ulangan.
   c. Mengevaluasi kinerja klasifikasi dengan melihat rataan nilai akurasi, sensitivitas, dan spesifisitasnya.
7. Membandingkan kinerja klasifikasi *random forest* tanpa SMOTE dan *random forest* dengan SMOTE.
8. Mengkaji peubah penjelas penting yang dihasilkan dengan mengurutkan nilai *Mean Decrease Gini* (MDG).

3. Hasil dan Pembahasan

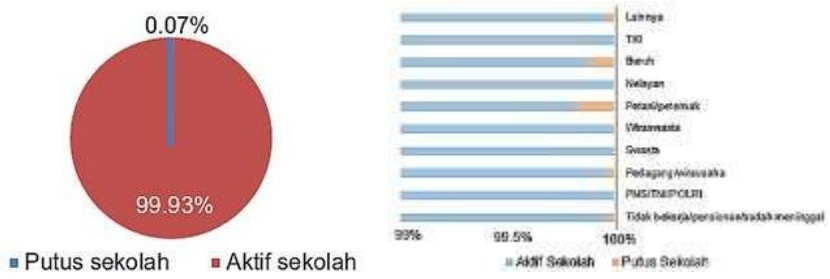
3.1 Praproses Data

Data siswa SMA di Provinsi Jawa Barat awalnya berjumlah 704587 amatan. Setelah ditelusuri lebih lanjut, terdapat 381660 amatan yang bernilai hilang. Seluruh amatan yang bernilai hilang merupakan amatan aktif sekolah dan amatan-amatan tersebut memiliki nilai hilang pada banyak peubah. Hal tersebut akan berdampak buruk pada proses klasifikasi apabila amatan dengan nilai hilang yang banyak ini tetap digunakan. Penanganan yang dilakukan pada data hilang adalah *complete case analysis* yaitu menghapus amatan-amatan tersebut dari dataset sehingga hanya amatan lengkap saja yang dianalisis.

3.2 Eksplorasi Data

Gambar 1 (a) menunjukkan persentase siswa putus sekolah sebesar 0.07% dan 99.93% sisanya merupakan siswa aktif sekolah. Jumlah amatan siswa putus sekolah sangat sedikit dibandingkan jumlah amatan siswa aktif sekolah. Perbedaan yang sangat ekstrim tersebut menunjukkan terdapat ketidakseimbangan pada data.

Terlihat pada Gambar 1 (b), mayoritas kasus putus sekolah terjadi pada siswa dengan ayah yang bekerja sebagai buruh dan siswa dengan ayah yang bekerja sebagai petani/peternak. Sangat sedikit kasus putus sekolah yang dialami siswa dengan ayah yang bekerja sebagai PNS/TNI/POLRI. Selain itu, tidak ada kasus putus sekolah pada siswa dengan ayah yang bekerja sebagai nelayan dan siswa dengan ayah yang bekerja sebagai TKI.



(a) (b)

Gambar 1 (a) Persentase status keaktifan dan (b) persentase status keaktifan berdasarkan jenis pekerjaan ayah

Terlihat pada Gambar 2 (a) kasus putus sekolah paling banyak terjadi pada siswa dengan ayah yang tidak bersekolah. Kasus putus sekolah paling sedikit terjadi pada siswa dengan ayah yang jenjang pendidikannya perguruan tinggi. Semakin rendah jenjang pendidikan ayah, maka semakin besar kecenderungan siswa mengalami putus sekolah. Gambar 2 (b) menunjukkan mayoritas kasus putus sekolah terjadi pada siswa dengan ayah yang berpenghasilan kurang dari satu juta. Kejadian putus sekolah paling banyak terjadi pada siswa dengan ayah yang berpenghasilan kurang dari 500 ribu. Persentase siswa putus sekolah memperlihatkan tren turun ketika penghasilan ayah semakin besar.

(a) (b)

Gambar 2 Persentase status keaktifan berdasarkan (a) jenjang pendidikan ayah dan (b) penghasilan ayah

Gambar 3 (a) menunjukkan bahwa rentang jumlah saudara kandung siswa putus sekolah adalah 0 sampai 6 orang. Kemudian nilai kuartil atas dan kuartil bawah menunjukkan bahwa mayoritas siswa putus sekolah memiliki jumlah saudara kandung sebanyak 2 sampai 3 orang. Pencilan bawah pada diagram kotak garis menandakan bahwa sangat sedikit siswa putus sekolah yang tidak memiliki saudara kandung. Selain itu, terdapat beberapa siswa putus sekolah yang memiliki jumlah saudara kandung lebih dari 4.

(a) (b)

Gambar 3 (a) diagram kota garis jumlah saudara kandung siswa putus sekolah (b) persentase kelas.

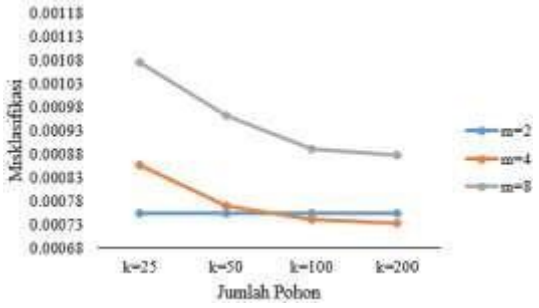
Gambar 3 (b) menunjukkan persentase kelas X, kelas XI, dan kelas XII pada siswa aktif sekolah tidak berbeda jauh. Persentase siswa yang mengalami putus sekolah pada tahun pertama atau kelas X ialah 14%. Kejadian putus sekolah paling banyak terjadi saat siswa berada di tahun kedua atau kelas XI yaitu sebesar 86%. Sebaliknya sangat sedikit kejadian putus sekolah terjadi saat siswa berada di tahun ketiga atau kelas XII. Hal tersebut menunjukkan bahwa siswa kelas XII memiliki kecenderungan yang lebih kecil mengalami putus sekolah.

3.3 Klasifikasi *Random Forest* tanpa SMOTE

Klasifikasi dilakukan menggunakan data latih dengan proporsi 80% dari keseluruhan data dan sisanya 20% untuk data uji yang digunakan untuk memberikan penilaian terhadap kinerja model. Menurut Musu et al. (2021), perbandingan 80:20 memberikan akurasi terbaik dibandingkan perbandingan lain seperti 70:30 dan 60:40. Persentase siswa putus sekolah dan siswa aktif sekolah pada data latih dan data uji sama yaitu masing-masing 0.07% dan 99.93%.

Random forest membangun kumpulan *k* pohon keputusan dengan memilih secara acak *m* peubah penjelas pada setiap proses pemisahannya. Menurut Breiman (2003), menggunakan $m=|\sqrt{d}|=4$ peubah penjelas akan memberikan hasil prediksi yang optimal, namun Breiman (2003) juga menyarankan agar mencoba $m=|\sqrt{\frac{d}{2}}|=2$ dan $m=|2\sqrt{d}|=8$ dengan *d* adalah banyaknya peubah penjelas pada data yaitu *d*=17. Banyaknya peubah penjelas yang digunakan yaitu *m*=2*,* *m*=4 dan *m*=8*,* sedangkan *k* pohon yang dibangun adalah 25, 50, 100, dan 200. Pemilihan kombinasi *m* dan *k* terbaik dilakukan dengan menganalisis nilai rataan misklasifikasi secara eksploratif. Misklasifikasi adalah peluang kesalahan dalam mengklasifikasikan suatu amatan ke dalam suatu kelas.

Breimen *et al* (2001) menyatakan bahwa tingkat kesalahan klasifikasi atau misklasifikasi akan konvergen menuju nilai tertentu saat ukuran *random forest* semakin besar. Hal tersebut terlihat pada Gambar 4, nilai misklasifikasi tidak banyak berubah atau konvergen saat ukuran pohon diperbesar sampai 200 pohon. Pada ukuran pohon *k*=200*,* nilai misklasifikasi yang dihasilkan peubah penjelas *m*=8 lebih besar dari peubah penjelas *m*=4 dan *m=2*. Kemudian nilai misklasifikasi yang dihasilkan peubah penjelas *m*=2 lebih besar dari peubah penjelas *m*=4. Berdasarkan rataan misklasifikasi yang dihasilkan, kombinasi ukuran peubah penjelas *m*=4 dan ukuran pohon *k*=200 dipilih karena menghasilkan rataan nilai misklasifikasi paling kecil.



Gambar 4 Perubahan rataan misklasifikasi

Kinerja *random forest* dalam mengklasifikasikan data dengan benar diprediksi menggunakan data uji. Ukuran kebaikan *random forestl* dapat dilihat dengan rataan

nilai akurasi, sensitivitas, dan spesifisitas. Berdasarkan Tabel 5, rataan akurasi *random forest* dalam mengklasifikasikan status keaktifan siswa SMA di Jawa Barat adalah 0.9993 yang berarti hampir seluruh siswa diklasifikasikan sesuai statusnya. Rataan sensitivitas yang dihasilkan sebesar 0.0188 yang artinya hanya 1.88% siswa putus sekolah tepat diklasifikasikan putus sekolah. Rataan spesifisitas yang dihasilkan sebesar 1 yang artinya seluruh siswa aktif sekolah tepat diklasifikasikan aktif sekolah. Jumlah amatan kelas mayor atau aktif sekolah yang sangat banyak dibandingkan kelas minor atau putus sekolah mengakibatkan hasil kinerja klasifikasi cenderung mengarah pada kelas mayor atau aktif sekolah.

Tabel 3 Rataan kinerja klasifikasi *random forest* tanpa SMOTE

Kinerja klasifikasi	Rataan
Akurasi	0.9993
Sensitivitas	0.0188
Spesifisitas	1

3.4 Klasifikasi *Random Forest* dengan SMOTE

Teknik SMOTE yang diterapkan merupakan kombinasi antara *oversampling* dan *undersampling*. *Undersampling* diterapkan pada data kelas mayor agar data sintetis yang dibangkitkan pada kelas minor tidak terlalu banyak. Chawla *et al* (2002) menerapkan teknik SMOTE yang dikombinasikan dengan *undersampling* dan nilai *undersampling*=1 menghasilkan kinerja klasifikasi yang optimal pada kelas minor (sesitivitas) maupun mayor (spesifisitas). Menurut Ubaya dan Juairiyah (2019), semakin seimbang data maka semakin baik kinerja yang dihasilkan.

Tabel 5 Proporsi amatan data latih

Kelas data	Tanpa SMOTE		Dengan SMOTE	
	n	Persentase	n	Persentase
Minor	185	0.07%	129500	50.00%
Mayor	250979	99.93%	129500	50.00%
Total	251164	100.00%	259000	100.00%

Pembangkitan data sintetis dilakukan dengan menggunakan *oversampling*=700, *undersampling*=1, dan *k*=5. Tabel 5 menunjukkan data kelas minor yang dibangkitkan sebanyak 700 kali sehingga menjadi 185x700=129500 amatan dan data kelas mayor direduksi agar jumlahnya sama dengan data kelas minor yaitu 129500 amatan. Perbandingan kelas mayor dan kelas minor menjadi seimbang (1:1). Selanjutnya, model klasifikasi dengan kombinasi ukuran peubah penjelas *m*=4 dan ukuran pohon *k*=200 dibangun menggunakan data latih yang telah diterapkan SMOTE.

Tabel 6 Rataan kinerja klasifikasi *random forest* dengan SMOTE

Kinerja klasifikasi	Rataan
Akurasi	0.9769
Sensitivitas	0.6415
Spesifisitas	0.9771

Kinerja klasifikasi *random forest* dengan SMOTE diukur menggunakan data uji yang sama dengan data uji *random forest* tanpa SMOTE. Tabel 6 menunjukkan kinerja klasifikasi *random forest* dengan SMOTE. Rataan nilai spesifitas *random forest* dengan SMOTE turun menjadi 0.9769, rataan sensitivitas yang diperoleh meningkat menjadi 0.6415, dan rataan nilai akurasi turun menjadi 0.9771. Penurunan nilai akurasi dan spesifisitas disebabkan oleh meningkatnya siswa aktif sekolah yang diklasifikasikan putus sekolah. Hasil prediksi *random forest* dengan SMOTE tidak lagi condong ke arah kelas mayor atau aktif sekolah, terbukti dengan nilai sensitivitas yang meningkat pesat dan menurunya nilai akurasi dan spesifisitas.

3.5 Perbandingan Kinerja Klasifikasi *Random Forest*

Tabel 7 Perbandingan kinerja klasifikasi *random forest*

Kinerja klasifikasi model	tanpa SMOTE	dengan SMOTE
Akurasi	0.9993	0.9769
Sensitivitas	0.0188	0.6415
Spesifisitas	1.0000	0.9771

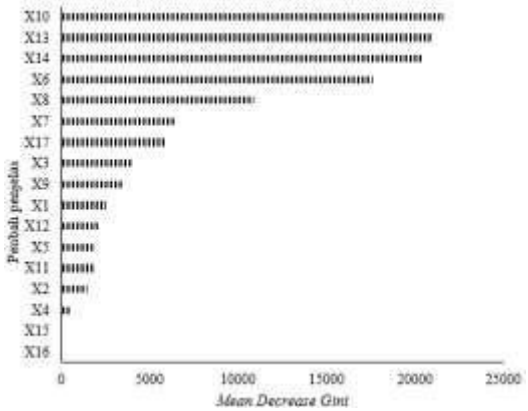
Kedua *random forest* yang telah dibangun, dibandingkan kinerjanya untuk menentukan yang terbaik. Tabel 7 menunjukkan bahwa nilai akurasi dan spesifisitas *random forest* dengan SMOTE sedikit lebih rendah, namun *random forest* tanpa SMOTE tidak dapat mengklasifikasikan kelas putus sekolah dengan baik. Hal tersebut dapat dilihat dari nilai sensitivitas yang sangat kecil yaitu 0.0188 sehingga kesalahan dalam mengklasifikasikan siswa putus sekolah sebagai siswa aktif sekolah sangat tinggi. Setelah diterapkan SMOTE, nilai sensitivitas meningkat menjadi 0.6415. Han dan Kamber (2006) menyatakan bahwa dalam kasus seperti ini sensitivitas menjadi prioritas dalam menentukan model terbaik. Kesalahan dalam mengklasifikasikan siswa aktif sekolah sebagai siswa putus sekolah memiliki risiko yang lebih kecil dibandingkan salah mengklasifikasikan siswa putus sekolah sebagai siswa aktif sekolah. Hal tersebut dikarenakan apabila banyak siswa putus sekolah yang diklasifikasikan sebagai siswa aktif sekolah akan berakibat fatal karena kebijakan pemerintah untuk menanggulangi kasus putus sekolah menjadi salah sasaran. *Random forest* dengan SMOTE ditentukan sebagai yang terbaik karena mampu meningkatkan nilai sensitivitas dan menurunkan kesalahan dalam mengklasifikasikan siswa putus sekolah sebagai siswa aktif sekolah.

3.6 Tingkat Kepentingan Peubah Penjelas

Random forest menghasilkan ukuran tingkat kepentingan pada masing-masing peubah penjelas (*X*) dengan melihat nilai *Mean Decrease Gini* (MDG). MDG merupakan rataan pengurangan nilai impuritas yang terjadi saat proses pemilahan pada pembentukan pohon tunggal klasifikasi. Semakin tinggi nilai MDG maka semakin tinggi tingkat kepentingan peubah penjelas tersebut. Peubah penting merupakan peubah yang memberikan pengaruh besar terhadap kategori respon yaitu status keaktifan siswa SMA di Jawa Barat.

Terlihat pada Gambar 8 *random forest* dengan SMOTE menghasilkan 5 peubah penjelas paling penting. Penghasilan ayah (X10) merupakan peubah terpenting dalam menentukan status keaktifan siswa SMA di Jawa Barat. Berdasarkan hasil eksplorasi,

terdapat banyak siswa putus sekolah dengan ayah yang berpenghasilan dibawah satu juta. Peubah paling penting kedua adalah jumlah saudara kandung (X13). Berdasarkan hasil eksplorasi, siswa putus sekolah mayoritas memiliki jumlah saudara kandung sebanyak 2 sampai 3 orang. Peubah kelas (X14) merupakan peubah ketiga yang memiliki tingkat kepentingan tinggi. Berdasarkan hasil eksplorasi, kasus putus sekolah lebih sering terjadi pada siswa kelas X dan kelas XI. Jenjang pendidikan ayah (X6) memiliki urutan keempat berdasarkan tingkat kepentingannya. Hasil eksplorasi menunjukkan mayoritas siswa putus sekolah memiliki ayah dengan jenjang pendidikan SMP/sederajat, SD/sederajat, dan tidak bersekolah. Sebaliknya, sangat sedikit siswa putus sekolah yang memiliki ayah dengan jenjang pendidikan di atas SMP/sederajat. Jenis pekerjaan ayah (X8) merupakan peubah terpenting kelima. Berdasarkan hasil eksplorasi, siswa putus sekolah mayoritas memiliki ayah yang bekerja sebagai buruh. Selain buruh, banyak juga siswa putus sekolah dengan ayah yang bekerja sebagai petani/peternak dan setelah ditelusuri memiliki penghasilan yang rendah.



Gambar 8 Tingkat kepentingan peubah penjelas *random forest* dengan SMOTE

4. Simpulan

Penelitian ini menunjukkan bahwa metode *random forest* dengan SMOTE dapat meningkatkan nilai sensitivitas dari 0.0188 menjadi 0.6415. Metode *random forest* dengan SMOTE bekerja lebih baik daripada metode *random forest* tanpa SMOTE dalam mengklasifikasikan status keaktifan siswa SMA di Jawa Barat karena memiliki nilai sensitivitas yang lebih besar. Nilai sensitivitas yang besar menunjukkan bahwa kesalahan dalam mengklasifikasikan siswa putus sekolah sebagai siswa aktif sekolah rendah. Metode *random forest* dengan SMOTE menghasilkan 5 peubah terpenting yang memiliki pengaruh besar dalam menentukan status keaktifan siswa SMA di Jawa Barat. Peubah-peubah tersebut yaitu penghasilan ayah, jumlah saudara kandung, kelas, pendidikan terakhir ayah, dan jenis pekerjaan ayah. Berdasarkan hasil eksplorasi, seorang siswa diklasifikasikan putus sekolah apabila penghasilan ayahnya kurang dari satu juta, memiliki jumlah saudara kandung 2 sampai 3 orang, berada di kelas X atau XI, pendidikan terakhir ayah SMP/sederajat atau sebelumnya, dan pekerjaan ayah sebagai buruh atau petani/peternak.

Daftar Pustaka

Albasia, MAY. (2018). Klasifikasi Keberhasilan Melanjutkan Pendidikan Jenjang SMA di Provinsi Banten dengan Metode CART dan Random Forest. Bogor(ID): IPB.

Breiman L. 2001. Random Forest. *Machine Learning*. 45:5-32.

Breiman L, Cutler A. (2003). Manual on Setting Up, Using, and Understanding Random Forest V4.0. California(US): University of California, Berkeley.

Chawla VN, Bowyer KW, Hall LO, Kegelmeyer WP. (2002). SMOTE: Synthetic minority oversampling technique*. Journal of Artificial Intelligence Research*. 16:321-357.

Han J, Kamber M. (2006). Data Mining: Concepts and Techniques. San Francisco(US): Elsevier.

Kamsihyati T, Sukino, Sakinah. (2016). Kajian faktor-faktor penyebab anak putus sekolah di Desa Jangrana Kecamatan Kesugihan Kabupaten Cilacap. Geo Edukasi.5(1):16-21.

Musu W, Ibrahim A, Heriadi. 2021. Pengaruh komposisi data training dan testing terhadap akurasi algoritma c4.5. Sistem Informasi dan Teknologi Informasi.10(1):186-195.

Purnajaya AR, Hanggara FD. (2021). Perbandingan performa teknik sampling data untuk klasifikasi pasien terinfeksi covid-19 menggunakan rontgen dada. *Journal of Applied Informatics and Computing*.5(1):37-42.

Sartono B, Syafitri UD. (2010). Metode pohon gabungan: solusi pilihan untuk mengatasi kelemahan pohon regresi dan klasifikasi tunggal.Forum Statistika dan Komputasi. 15(1):1-17.

Sugianto E. (2017). Faktor penyebab anak putus sekolah tingkat SMA di Desa Lipai Kecamatan Batang Cenaku Kabupaten Inderagiri Hulu. JOM FISIP. 4(2):1-17.

Syukron M, Santoso R, Widiharih T. (2020). Perbandingan metode SMOTE *random forest* dan SMOTE *XGboost* untuk klasifikasi tingkat penyakit hepatitis C pada imbalance data class. Jurnal Gaussian. 9(3):227-236.

Ubaya H, Juairiah RS. (2020). Performance of RUS and SMOTE Method on Twitter Spam Data Using Random Forest. *Journal of Physics: Conference Series*. 1050(2020)012130:1-8.

Zhou X et al. (2020). Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. *Reliability Engineering and System Safety*. 200:1-9.

[PSDPK] Pusat Data dan Statistik Pendidikan dan Kebudayaan. 2018. Ikhtisar Data Pendidikan Dasar dan Menengah Tahun 2017/2018. Jakarta (ID): Kemendikbudristek.

[PSDPK] Pusat Data dan Statistik Pendidikan dan Kebudayaan. 2019. Ikhtisar Data Pendidikan Dasar dan Menengah Tahun 2018/2019. Jakarta (ID): Kemendikbudristek.