JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 202–208
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.799

PERBANDINGAN MODEL RANDOM FOREST DAN XGBOOST UNTUK PREDIKSI KEJAHATAN KESUSILAAN DI PROVINSI JAWA BARAT

Adlina Khairunnisa
Prodi Statistika dan Sains Data, IPB University
Email: adlina.khairunnisa@bps.go.id

Abstrak

Jumlah kejahatan di Indonesia secara umum mengalami penurunan tetapi jumlah kejahatan terhadap kesusilaan mengalami peningkatan yang signifikan pada tahun 2020. Provinsi Jawa Barat menempati posisi ketiga dengan kejadian kejahatan kesusilaan tertinggi. Analisis klasifikasi ada atau tidaknya kejahatan kesusilaan diperlukan sebagai upaya pencegahan kejahatan kesusilaan. Masalah pada analisis klasifikasi adalah ketidakseimbangan data antar kelas pada peubah respon. Metode penanganan data tidak seimbang diantaranya undersampling, oversampling, SMOTE dan model ensemble. Data yang digunakan adalah data sekunder yang bersumber dari PODES 2018. Pemodelan menggunakan metode ensemble seperti XGBoost dan random forest dengan beberapa teknik penanganan data yang tidak seimbang seperti undersampling, oversampling, dan SMOTE. Penentuan threshold optimal digunakan untuk meningkatkan kinerja model. Model terbaik untuk mengklasifikasikan kejahatan kesusilaan di desa/kelurahan Provinsi Jawa Barat adalah XGBoost dengan SMOTE. Analisis peubah penting dengan menggunakan SHAP menyimpulkan bahwa adanya kejadian kejahatan kesusilaan di desa/kelurahan Jawa Barat memiliki karakteristik sebagai berikut: adanya pengedaran narkoba, kejahatan pencurian biasa, pencurian dengan kekerasan, penipuan, penganiayaan, perjudian, jumlah penginapan yang banyak dan jarak pub/diskotik yang dekat.
Kata Kunci: *Kesusilaan, Random Forest, SHAP, SMOTE, XGBoost.*

Abstract

The number of crimes in Indonesia has generally decreased but the number of sexual crimes has increased significantly in 2020. West Java Province ranks third with the highest incidence of sexual crimes. Classification analysis of the presence or absence of sexual crimes is required as an effort to prevent sexual crimes. The problem with classification analysis is the imbalance of data between classes on the response variable. Methods of handling imbalanced data such as under-sampling, oversampling, SMOTE and ensemble methods. The data used is secondary data sourced from PODES 2018. Modelling uses ensemble methods such as XGBoost and random forest with several imbalanced data handling techniques such as under-sampling, oversampling, and SMOTE. Optimal threshold determination is used to improve model performance. The best model for classifying sexual crimes in West Java Province is XGBoost with SMOTE. Analysis of important variables using SHAP concluded that there are incidences of sexual crimes in West Java that have the following characteristics: the presence of drug distribution, common theft, violent theft, fraud, abuse, gambling, a large number of hotels and close proximity of pubs/discotheques.
KeyWords: *Kesusilaan, Random Forest, SHAP, SMOTE, XGBoost.*

I. PENDAHULUAN

Salah satu tujuan SDGs 2030 yaitu mengurangi segala bentuk kekerasan dan angka kematian dimanapun. Kriminalitas adalah tindakan kejahatan yang merugikan seseorang karena menimbulkan trauma bagi korban dari tindak kejahatan. Kejadian kejahatan dapat mengakibatkan kondisi ekonomi yang buruk secara terus-menerus [1]. Pemerintah harus menambah pengeluaran untuk fasilitas keamanan baik menambah pos polisi maupun personil polisi.

Berdasarkan Statistik Kriminal 2021, jumlah kejahatan secara umum mengalami penurunan tetapi jumlah kejahatan terhadap kesusilaan di Indonesia meningkat signifikan pada tahun 2020. Jumlah kejahatan terhadap kesusilaan tahun 2019 sebanyak 5.233 kejadian naik menjadi 6.872 kejadian pada tahun 2020. Jumlah kejahatan terhadap kesusilaan meliputi kejadian perkosaan dan pencabulan. Provinsi Jawa Barat menempati posisi ketiga tertinggi kejadian kejahatan kesusilaan. Oleh karena itu, pemerintah memberikan perhatian khusus untuk mencegah, menangani segala bentuk kekerasan seksual, melindungi, dan memulihkan korban kekerasan seksual dengan menetapkan UU Nomor 12 Tahun 2022 terkait Tindak Pidana Kekerasan Seksual (TPKS). Untuk mengetahui ada atau tidaknya kejahatan kesusilaan di desa/kelurahan Jawa Barat sebagai upaya pencegahan tindakan kejahatan, diperlukan analisis klasifikasi.

Klasifikasi ada atau tidaknya kejahatan kesusilaan ditentukan berdasarkan karakteristik lingkungan di Provinsi Jawa Barat. Karakteristik lingkungan yang menunjukkan hubungan dengan kejadian kejahatan kesusilaan antara lain status sosial ekonomi suatu wilayah, jumlah penduduk pendatang, dan jumlah tindak kejahatan publik [2]. Akan tetapi, analisis klasifikasi memiliki permasalahan yaitu ketidakseimbangan data antar kelas pada peubah respon dimana jumlah salah satu kelas dari peubah respon jauh lebih banyak dibandingkan kelas lainnya. Hal ini mengakibatkan model cenderung condong terhadap kelas mayoritas dalam melakukan prediksi.

Ada beberapa metode penanganan data tidak seimbang antara lain *undersampling*, *oversampling*, dan pembangkitan data sintetis. Salah satu metode dengan pendekatan pembangkitan data sintetis yang bekerja lebih baik dari *undersampling* dan

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 202–208
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.799

oversampling adalah *Synthetic Minority Over Sampling* (SMOTE) [3]. SMOTE menerapkan metode penarikan contoh dalam melakukan modifikasi terhadap distribusi data antar kelas mayoritas dan kelas minoritas pada sekumpulan data *training* untuk menyeimbangkan jumlah data pada setiap kelas [3]. Pendekatan lain untuk meningkatkan akurasi klasifikasi pada data yang tidak seimbang adalah metode *ensemble* [4]. Metode *ensemble* dapat meningkatkan akurasi dengan membangun beberapa model klasifikasi dari data *training* kemudian menggunakan *majority vote* untuk prediksi akhir. Metode *ensemble* yang sering digunakan yaitu *Random Forest* (RF) yang menerapkan metode *bagging*. Selain itu, ada metode *ensemble* lainnya pengembangan dari *gradient boosting* yaitu *extreme gradient boosting* (XGBoost) yang dapat digunakan dalam regresi maupun klasifikasi. XGBoost dikenal memiliki kinerja yang baik dibandingkan metode *gradient boosting*.

Penelitian yang dilakukan [5] untuk klasifikasi kejahatan di Cina dengan pembelajaran mesin dengan model *Extreme Gradient Boosting* (XGBoost) menghasilkan tingkat akurasi yang tinggi. Perbandingan beberapa model pembelajaran mesin dilakukan oleh [6] untuk mengklasifikasikan subtipe kejahatan seksual di Denver. Model *Naïve Bayes* dengan *AdaBoost* (NB-AB) adalah model terbaik jika dibandingkan dengan *Logistic Regression*, *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), dan *Stochastic Gradient Descent* (SGD).

Pemodelan dengan pembelajaran mesin saja belum cukup untuk menjelaskan karakteristik kejahatan kesusilaan sehingga diperlukan interpretasi model. Pembelajaran mesin secara umum menghasilkan model yang sulit untuk dijelaskan dalam bentuk sederhana yang sering disebut dengan *black box* [7]. Salah satu metode untuk mengidentifikasi tingkat kepentingan peubah penjelas dalam model yaitu *Shapley Additive Explanations* (SHAP). Berdasarkan penjelasan di atas, penelitian akan mengkaji perbandingan metode XGBoost dan *random forest* dengan beberapa penanganan data tidak seimbang diantaranya *undersampling*, *oversampling*, dan SMOTE untuk mengklasifikasikan ada atau tidaknya kejadian kesusilaan di desa/kelurahan Provinsi Jawa Barat.

II. METODE

A. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder dari Badan Pusat Statistik (BPS) yang bersumber dari Survei Potensi Desa (PODES) 2018. Cakupan dari data yang digunakan adalah data sebanyak 5.957 desa/kelurahan di Provinsi Jawa Barat.

Tabel I: Daftar peubah yang digunakan

Nama Peubah	Skala Pengukuran	Keterangan
Kejadian perkosaan/kejahatan terhadap kesusilaan selama setahun terakhir	Nominal	Ada Tidak ada
Perkiraan jarak ke pos polisi/kantor polisi terdekat	Rasio	kilometer
Keberadaan lokasi berkumpul/mangkal anak jalanan (selain rumah singgah) di desa/kelurahan	Nominal	Ada Tidak ada
Keberadaan tempat mangkal gelandangan/pengemis di desa/kelurahan	Nominal	Ada Tidak ada
Keberadaan lokalisasi/lokasi/tempat mangkal Pekerja Seks Komersial (PSK) di desa/kelurahan	Nominal	Ada Tidak ada
Keberadaan permukiman di bantaran sungai	Nominal	Ada Tidak ada
Keberadaan permukiman kumuh (sanitasi lingkungan buruk, bangunan padat, dan sebagian besar tidak layak huni)	Nominal	Ada Tidak ada
Jarak pub/diskotik/tempat karaoke yang masih berfungsi	Rasio	kilometer
Jumlah hotel dan penginapan	Rasio	
Kejadian pencurian selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian pencurian dengan kekerasan selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian penipuan/penggelapan selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian penganiayaan selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian pembakaran selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian penyalahgunaan/peredaran narkoba selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian perjudian selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian pembunuhan selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian perdagangan orang selama setahun terakhir	Nominal	Ada Tidak ada
Kejadian korupsi selama setahun terakhir	Nominal	Ada Tidak ada

B. Prosedur Analisis

Prosedur analisis data yang dilakukan dalam penelitian ini adalah sebagai berikut:

1) Eksplorasi data untuk melihat gambaran umum peubah-peubah yang akan dianalisis.
2) Membagi data menjadi dua bagian yaitu data *training* 70% dan data testing 30%. Data *training* akan digunakan untuk pemodelan, sedangkan data *testing* digunakan dalam evaluasi model.
3) Peubah kategorik akan diubah menjadi peubah *dummy*.
4) Melakukan penyeimbangan data pada data *training* dengan metode *undersampling*, *oversampling*, dan *Synthetic Minority Oversampling Technique* (SMOTE) karena proporsi kategori ada kejadian kesusilaan lebih sedikit daripada proporsi kategori tidak ada kejadian kesusilaan.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 202–208
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.799

5) Setelah data *training* sudah seimbang, langkah selanjutnya adalah melakukan pemodelan *Random Forest* dan *Extreme Gradient Boosting* dengan *tuning hyperparamete*r menggunakan *10-fold cross validation* sampai diperoleh *hyperparameter optimal* dari masing-masing model.
6) Mengevaluasi kinerja model dengan melihat rata-rata *balanced accuracy*, sensitivitas, dan spesifisitas.
7) Pengukuran tingkat kepentingan peubah berdasarkan model terbaik yang terpilih dengan menggunakan SHAP. Tahapannya adalah menghitung nilai SHAP setiap desa/kelurahan per peubah penjelas berdasarkan model terbaik dan parameter yang optimal kemudian menentukan peubah penting dengan tingkat kepentingan yang tinggi dengan melihat SHAP *summary plot*.

C. Random Forest

Random Forest (RF) merupakan pengembangan dari algoritma *Classification and Regression Tree* (CART) dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection* [8]. RF dikembangkan dari metode bagging dengan melakukan pemilihan secara acak peubah penjelas untuk mengurangi korelasi antar pohon yang terbentuk [9]. Kelebihan dari metode ini adalah dapat mengatasi masalah overfitting, tidak sensitif terhadap pencilan dan dapat menghasilkan akurasi yang baik [10]. Tahapan klasifikasi dalam menggunakan RF pada gugus data yang terdiri atas n amatan dan p peubah penjelas [11]:

1) Tahapan *bootstrap*. *Bootstrap* adalah pengambilan contoh yang disertai dengan pengembalian. Pada tahap ini, ambil sebanyak n contoh acak dari data *training*.
2) Tahapan *random feature subset*. Pohon disusun berdasarkan data hasil *bootstrap* sebelumnya. Pada setiap proses pemisahan, pilih m peubah penjelas secara acak dimana $m < p$. Selanjutnya, lakukan pemisahan terbaik.
3) Ulangi langkah 1-2 sebanyak k kali sehingga diperoleh k buah pohon.
4) Lakukan penggabungan (agregasi) hasil prediksi k buah pohon dan menggunakan *majority vote* (mengambil *vote* terbanyak) untuk menentukan hasil prediksi akhir dari *k* pohon klasifikasi.

D. Extreme Gradient Boosting

Extreme gradient boosting (*XGBoost*) merupakan pengembangan dari *gradient boosting* [12]. XGboost membuat model baru berdasarkan sisa (*error*) dari model sebelumnya sehingga model baru yang diperoleh lebih akurat. XGBoost menggunakan teknik *boosting* yang secara iteratif membangun set model yang lemah pada subset data dengan meminimalkan nilai *mean square error* $(\hat{y} - y)$ dari model F dengan $\hat{y} = f(x)$ dan menimbang setiap prediksi yang lemah sesuai dengan kinerjanya. Prediksi akhir diperoleh dengan mengambil jumlah terbobot dari prediksi pohon keputusan. Algoritma dasar XGBoost yang meminimumkan fungsi kerugian dijelaskan pada Persamaan (1).

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i \hat{y}_i^{(t-1)} + f_i(x_i)\right) + \sum_{1=1}^{t} \Omega(f_i) \tag{1}$$

dengan $l\left(Y_i \hat{y}_i^{(t-1)}\right)$ adalah *loss function* untuk mengukur kesalahan prediksi dan f_i mengontrol kompleksitas untuk menghindari *overfitting*. Untuk menentukan pemisah sampul, digunakan skor *gain* yang diperoleh dari Persamaan (2).

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\epsilon I_L} g_i\right)^2}{\sum_{i\epsilon I_L} h_i + \lambda} + \frac{\left(\sum_{i\epsilon I_L} g_i\right)^2}{\sum_{i\epsilon I_L} h_R + \lambda} - \frac{\left(\sum_{i\epsilon I_L} g_i\right)^2}{\sum_{i\epsilon I_L} h_L + \lambda}\right] - \gamma \tag{2}$$

dengan nilai g_i adalah turunan pertama dan h_i adalah turunan kedua dari fungsi kerugian. λ adalah parameter regulasi dan γ adalah parameter penalti. Model XGBoost dapat memaksimalkan akurasi dengan menggunakan hyperparameter seperti *max_depth*, *random state*, *colsample_bytree*, dan lainnya.

E. Shapley Additive Explanations (SHAP)

Salah satu metode yang digunakan untuk menginterpretasikan model pembelajaran mesin adalah SHAP. Metode ini digunakan untuk menjelaskan prediksi individu berdasarkan nilai permainan *Shapley* [13]. SHAP bertujuan untuk menjelaskan prediksi suatu individu x dengan menghitung kontribusi dari masing-masing peubah. Nilai *Shapley* dapat menjelaskan prediksi global dan lokal. Nilai *Shapley* merupakan metode dalam permainan yang digunakan untuk menentukan hadiah kepada pemain secara adil dengan mempertimbangkan kontribusi pemain terhadap nilai yang diperoleh dalam permainan. Pemain dianalogikan sebagai peubah prediktor dan hadiah adalah nilai dugaan. Semua kemungkinan kombinasi peubah prediktor dengan peubah ke-j dan tanpa peubah ke-j dievaluasi untuk mendapatkan nilai *Shapley* (Persamaan (3)).

$$\Phi_j = \sum_{S \subseteq M(j)} \frac{|S|!(M - |S| - 1)!}{M!}(\nu(S \cup \{j\}) - \nu(S)), j = 1, \ldots, M \tag{3}$$

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 202–208
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.799

dengan ϕ_j adalah nilai kontribusi peubah ke- j (skor Shapley), $S \subseteq M$ adalah subset koalisi dari jumlah peubah, S adalah koalisi, M adalah jumlah peubah, $\nu S \cup j$ adalah prediksi yang menyertakan seluruh peubah, dan νS adalah prediksi tanpa ke-j.

Nilai SHAP menentukan urutan kontribusi dari peubah. Semakin besar nilai SHAP, peubah tersebut semakin penting. Nilai *Shapley* yang didapatkan selanjutnya digunakan untuk menghitung nilai SHAP. Perhitungan SHAP *feature importance* (FISHAP) dilakukan dengan merata-ratakan nilai *Shapley* dari seluruh pengamatan.

$$FI_{SHAP} = I_j = \frac{1}{n}\sum_{i=1}^{n}\left|\phi_j^{(i)}\right| \tag{4}$$

Nilai SHAP menentukan urutan kontribusi dari peubah. Semakin besar nilai SHAP, peubah tersebut semakin penting.

F. Ukuran Kebaikan Model

Ukuran kebaikan model digunakan untuk mengukur seberapa baik model dalam analisis klasifikasi. Perhitungan akurasi dan skor confusion matrix atau matriks konfusi adalah sebagai berikut:

1) Akurasi
   Akurasi adalah proporsi data yang diklasifikasikan dengan benar.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

2) Sensitivitas
   Sensitivitas adalah ukuran seberapa baik model dalam mengklasifikasikan data di kelas positif.

$$Sensitivitas = \frac{TP}{TP + FN} \tag{6}$$

3) Spesifisitas
   Spesifisitas adalah ukuran seberapa baik model mengklasifikasikan data di kelas negatif.

$$Spesifisitas = \frac{TN}{TN + FP} \tag{7}$$

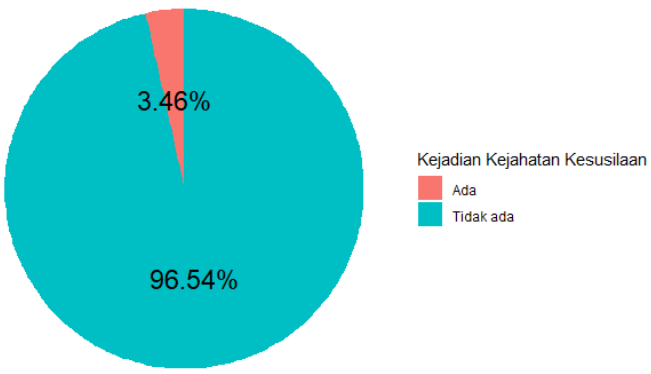
Akurasi kurang dapat menggambarkan tingkat prediksi pada data yang tidak seimbang karena prediksi akan cenderung pada kelas mayoritas. *Balanced accuracy* dapat digunakan untuk mengukur kinerja pada kelas tidak seimbang [14].

$$Balanced\ accuracy = \frac{1}{s}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) = \frac{1}{2}(Sensitivitas + Spesifisitas) \tag{8}$$

III. HASIL DAN PEMBAHASAN

A. Eksplorasi Data

Gambar 1 menunjukkan adanya masalah ketidakseimbangan kelas dimana persentase desa/kelurahan yang terdapat kejahatan kesusilaan lebih rendah daripada yang dikategorikan tidak ada kejahatan kesusilaan. Persentase desa/kelurahan yang terdapat kejahatan kesusilaan sebesar 3.46% dan desa/kelurahan yang tidak terdapat kejahatan kesusilaan sebesar 97.42%. Masalah data tidak seimbang ini perlu dilakukan penanganan dengan metode penyeimbangan data *undersampling*, *oversampling*, dan SMOTE.



Gambar 1: Pie Chart Proporsi Kejadian Kejahatan Kesusilaan

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 202–208
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.799

B. Pemodelan

Hasil perbandingan kinerja model *ensemble* dengan beberapa metode penanganan data tidak seimbang dan tanpa penanganan dapat dilihat pada Gambar 2. XGBoost menunjukkan kinerja model yang baik dengan penanganan data tidak seimbang yaitu *undersampling*, *oversampling* dan SMOTE. Hal tersebut ditunjukkan dengan nilai *balanced accuracy* yang cukup tinggi. Sementara itu, *Random Forest* menunjukkan kinerja yang baik dengan metode penanganan data tidak seimbang yaitu *undersampling* dan *oversampling*.

Evaluasi kinerja klasifikasi tidak hanya berdasarkan pada nilai *balanced accuracy* tetapi juga nilai sensitivitas dan spesifisitas. Data yang tidak seimbang mempengaruhi nilai tersebut. Nilai sensitivitas yang paling rendah yaitu nol terjadi pada model XGBoost dan *Random Forest* sebelum melalui proses penanganan data tidak seimbang. Nilai sensitivitas yang sangat rendah bahkan nol mengindikasikan proses klasifikasi cenderung condong ke kelas mayoritas dalam hal ini adalah tidak adanya kejadian kesusilaan di desa/kelurahan Provinsi Jawa Barat. Selain itu, nilai spesifisitas juga dapat dipertimbangkan untuk mengukur kebaikan model klasifikasi. Model yang baik sebaiknya memiliki nilai sensitivitas dan spesifisitas yang tinggi dan stabil.

Gambar 2: Perbandingan kinerja model ensemble dengan metode penanganan data tidak seimbang

Metode *ensemble* dapat meningkatkan nilai sensitivitas setelah melakukan penanganan data tidak seimbang. Akan tetapi, nilai tersebut berada di bawah nilai spesifisitas. Nilai sensitivitas dari model XGBoost dan *Random Forest* berkisar di antara 0.3 – 1.0 sedangkan nilai spesifisitasnya di antara 0.7 – 1.0. Karena distribusi peluang data tidak seimbang cenderung condong ke kelas mayoritas, ambang batas atau *threshold* 0.5 bukan pilihan terbaik [15]. Penentuan *threshold optimal* bertujuan untuk meningkatkan ukuran kinerja model baik *balanced accuracy*, sensitivitas, maupun spesifisitas suatu model. Pemilihan nilai *threshold optimal* dapat menggunakan *youden j index*. Indeks tersebut memberikan pembobot yang seimbang untuk nilai sensitivitas dan spesifisitas [16]. *Youden j index* yang tinggi menentukan *threshold* yang optimal.

Tabel II: Perbandingan kinerja model XGBoost dengan *threshold* optimal

Metode Penanganan Data Tidak Seimbang	Balanced Accuracy	Sensitivitas	Spesifisitas	*Threshold*
Undersampling	0.5000	1.0000	0.7705	1.00
Oversampling	0.7074	0.7591	0.7705	0.50
SMOTE	0.7130	0.6554	0.7705	0.49

Dari hasil evaluasi, model XGBoost akan digunakan sebagai model terbaik karena memiliki *balanced accuracy* lebih tinggi daripada *Random Forest*. Tabel II menunjukkan peningkatan nilai *balanced accuracy* pada XGBoost dengan metode penanganan data tidak seimbang yaitu *oversampling* dan SMOTE. Pemilihan nilai *threshold* optimal pada XGBoost dengan penanganan data tidak seimbang yaitu *undersampling*, menurunkan *balanced accuracy* yang semula 0.67 menjadi 0.50. Hal ini terjadi kemungkinan karena proses *undersampling* menyebabkan hilangnya sebagian data dari kelas mayoritas sehingga ada kemungkinan data tersebut adalah data yang berguna atau bahkan sangat penting dalam pembangunan model klasifikasi yang baik [17]. *Balanced accuracy* dihitung dari Persamaan (8) yaitu $\frac{1}{2}(0.6554 + 0.7705) = 0.7130$. Model yang paling baik untuk mengklasifikasikan kejahatan kesusilaan di Provinsi Jawa Barat adalah XGBoost dengan penanganan SMOTE dengan nilai *balanced accuracy* sebesar 0.7130.

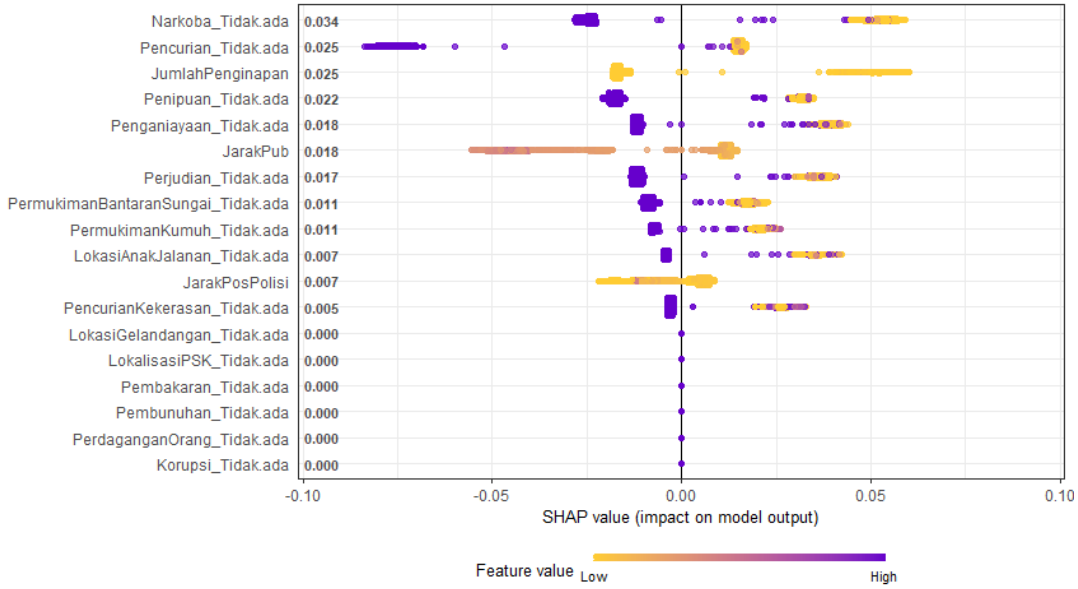
Hyperparameter yang optimal dalam pembentukan model XGBoost dengan SMOTE adalah $mtry = 1$, $ntrees = 270$, $treedepth = 14$, dan $learningrate = 0.002$. *Mtry* yaitu jumlah peubah yang digunakan sebagai *split* pada setiap pohon yang terbentuk. *N trees* (*number of trees*) adalah banyaknya pohon yang terbentuk secara sekuensial. *Tree depth* mengontrol kedalaman sebuah pohon. Pohon yang semakin dalam akan berisiko *overfitting*. *Learning rate* mengontrol algoritma supaya tidak terlalu cepat menuju *loss function minimum*. *Learning rate* berada di antara nilai 0-1, *learning rate* yang kecil dapat meningkatkan kinerja algoritma tetapi proses *training* akan berjalan lambat serta lebih rentan terjadinya *overfitting*.

C. SHAP Summary Plot

Setelah pembentukan model, interpretasi model dilakukan untuk melihat karakteristik peubah yang memiliki kontribusi terhadap klasifikasi kejahatan kesusilaan di Provinsi Jawa Barat. SHAP merupakan salah satu metode yang digunakan untuk menginterpretasikan model pembelajaran mesin. Tujuan SHAP adalah menghitung kontribusi setiap peubah sehingga mampu menjelaskan prediksi setiap individu [13]. Peubah paling penting mampu memprediksi kejadian pada peubah respon dan membedakan prediksi dengan baik antar kategori peubah. Pada peubah tertentu dihasilkan titik-titik amatan berwarna ungu (skor peubah tinggi) yang mengumpul di skor SHAP nol artinya peubah tidak memiliki kontribusi terhadap prediksi.

Peubah penting di urutan pertama adalah tidak adanya pengedaran narkoba di desa/kelurahan Jawa Barat dengan pola sebaran amatan berwarna ungu mengumpul pada daerah skor SHAP negatif. Tidak adanya pengedaran narkoba di desa/kelurahan Jawa Barat berkontribusi besar untuk berkurangnya kejadian kesusilaan di daerah tersebut. Hal ini juga terlihat pada peubah kejadian kejahatan pencurian biasa, pencurian dengan kekerasan, penipuan, penganiayaan, dan perjudian. Adanya tindak kejahatan

tersebut diprediksi menimbulkan tindakan kejahatan kesusilaan di Jawa Barat. Sementara itu, lokasi gelandangan, lokalisasi PSK, tindak kejahatan pembakaran, pembunuhan, perdagangan orang serta korupsi tidak berkontribusi terhadap ada atau tidaknya kejahatan kesusilaan di Jawa Barat.



Gambar 3: Urutan peubah penting dengan model XGBoost

IV. Simpulan

Evaluasi kinerja model untuk mengklasifikasikan kejadian kejahatan kesusilaan di desa/kelurahan Jawa Barat dengan model *ensemble* dan penanganan data tidak seimbang menunjukkan akurasi yang cukup baik terlihat dari nilai *balanced accuracy*. XGBoost dengan SMOTE menghasilkan kinerja yang lebih baik dalam proses klasifikasi dibandingkan *Random Forest*. Kinerja pengklasifikasian pada data tidak seimbang yang baik tidak hanya didasarkan pada *balanced* akurasi, melainkan ukuran ketepatan lainnya seperti sensitivitas dan spesifisitas. Akurasi yang bernilai sangat tinggi tidak berarti metode yang digunakan dalam mengklasifikasikan data sudah tepat. Masalah yang sering terjadi yaitu nilai akurasi yang dihasilkan sangat tinggi tetapi nilai sensitivitas yang dihasilkan sangat rendah, terutama pada kelas minoritas. Perbedaan nilai sensitivitas dan spesifisitas yang cukup jauh memerlukan penyesuain *threshold*. Penyesuain *threshold* dari 0.5 menjadi 0.49 dapat meningkatkan nilai *balanced accuracy* pada model XGBoost dengan SMOTE.

Analisis peubah penting menggunakan SHAP menyimpulkan adanya kejadian kejahatan kesusilaan di desa/kelurahan Jawa Barat memiliki karakteristik sebagai berikut adanya tindakan kejahatan pengedaran narkoba, kejahatan pencurian biasa, pencurian dengan kekerasan, penipuan, penganiayaan, perjudian, jumlah penginapan, jarak ke pub/diskotik, dan jarak ke pos polisi.

Pustaka

[1] L. Mauro and G. Carmeci, "A poverty trap of crime and unemployment," Review of Development Economics, vol. 11, no. 3, pp. 450–462, 2007.
[2] E. Gracia, A. López-Quílez, M. Marco, S. Lladosa, and M. Lila, "Exploring neighborhood influences on small-area variations in intimate partner violence risk: A bayesian random-effects modeling approach," International journal of environmental research and public health, vol. 11, no. 1, pp. 866–882, 2014.
[3] 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, 2002.
[4] J. Han, M. Kamber, and D. Mining, "Concepts and techniques," Morgan Kaufmann, vol. 340, pp. 94104–3205, 2006.
[5] X. Zhang, L. Liu, M. Lan, G. Song, L. Xiao, and J. Chen, "Interpretable machine learning models for crime prediction," Computers, Environment and Urban Systems, vol. 94, p. 101789, 2022.
[6] S. Parthasarathy and A. R. Lakshminarayanan, "Naïve Bayes–AdaBoost Ensemble Model for Classifying Sexual Crimes," in Data Intelligence and Cognitive Informatics, Springer, 2022, pp. 393–405.
[7] A. Apicella, F. Isgrò, R. Prevete, and G. Tamburrini, "Middle-level features for the explanation of classification systems by sparse dictionary methods," International Journal of Neural Systems, vol. 30, no. 08, p. 2050040, 2020.
[8] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.
[9] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer, 2009.
[10] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," International Journal of Computer Science Issues(IJCSI), vol. 9, Sep. 2012.
[11] B. Sartono and U. D. Syafitri, "Metode pohon gabungan: Solusi pilihan untuk mengatasi kelemahan pohon regresi dan klasifikasi tunggal," in Forum Statistika dan Komputasi, 2010, vol. 15, no. 1.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 202–208
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.799

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[13] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles." arXiv, 2018. doi: 10.48550/ARXIV.1802.03888.
[14] D. R. Velez et al., "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," Genetic Epidemiology: the Official Publication of the International Genetic Epidemiology Society, vol. 31, no. 4, pp. 306–315, 2007.
[15] G. King and L. Zeng, "Logistic regression in rare events data," Political analysis, vol. 9, no. 2, pp. 137–163, 2001.
[16] E. F. Schisterman, D. Faraggi, B. Reiser, and J. Hu, "Youden Index and the optimal threshold for markers with mass at zero," Statistics in medicine, vol. 27, no. 2, pp. 297–315, 2008.
[17] Y. Ma and H. He, "Imbalanced learning: foundations, algorithms, and applications," 2013.