 ISSN:2302-8149
e-ISSN:2540-9719

Penanganan *Imbalance* Data pada Klasifikasi Kabupaten/Kota di Kawasan Timur Indonesia

Handling of Data Imbalance in Classification of Regencies/Municipalities in Eastern Indonesia

[1]Adham Malay Japany*, [2]Yuliagnis Transver Wijaya
[1,2]Politeknik Statistika STIS
Jalan Otto Iskandardinata No. 64C, Jatinegara, Jakarta Timur, Indonesia
*e-mail: 211910817@stis.ac.id

(***received*:** 2 Juni 2023, ***revised*:** 12 September 2023, ***accepted*:** 19 September 2023)

Abstrak

Ketidakseimbangan data antar kelas dapat menghasilkan prediksi yang salah dalam klasifikasi, sehingga dapat menimbulkan masalah dalam pengambilan keputusan. Kawasan Timur Indonesia (KTI) adalah salah satu daerah yang memiliki Indeks Pembangunan Manusia (IPM) di bawah IPM nasional, sehingga peningkatan potensi manusia dalam proses produksi di KTI harus difokuskan. Dalam pengkategorian kabupaten/kota di KTI terjadi *imbalanced* data. Hal ini menunjukkan pembangunan manusia antar wilayah di KTI masih belum merata. Untuk itu, dilakukan klasifikasi kabupaten/kota berdasarkan IPM ke dalam kategori tertentu secara akurat dan cepat. Hasil klasifikasi diharapkan dapat membantu pemerintah dalam menentukan langkah strategis kedepannya untuk meningkatkan kualitas SDM di KTI. Salah satu metode yang dapat menangani ketidakseimbangan data adalah *Synthetic Minority Over-sampling Technique* (SMOTE), dengan menggunakan tiga algoritma klasifikasi, yaitu *Support Vector Machine* (SVM), *K-Nearest neighbors* (KNN), dan *Random Forest* (RF). Didapatkan hasil bahwa dengan adanya penanganan *imbalance* data dan diterapkannya metode *k-fold cross validation*, ketiga algoritma menunjukkan peningkatan akurasi yang signifikan. Oleh karena itu, penanganan *imbalance* data terbukti mampu meningkatkan performa dari algoritma klasifikasi yang diterapkan.

Kata kunci: *Imbalance* Data, *Support Vector Machine*, *K-Nearest neighbors*, *Random Forest*, Kawasan Timur Indonesia

Abstract

Imbalance of data between classes can result in incorrect predictions in classification, which can cause problems in decision making. Eastern Indonesia (KTI) is one of the regions that has a Human Development Index (HDI) below the national HDI, so increasing human potential in the production process in KTI must be focused on. In the categorization of regencies/municipalities in KTI there is imbalanced data. This shows that human development between regions in KTI is still uneven. For this reason, a classification of regencies/municipalities based on HDI into certain categories is carried out accurately and quickly. The classification results are expected to help the government in determining future strategic steps to improve the quality of human resources in KTI. One method that can handle data imbalance is Synthetic Minority Over-sampling Technique (SMOTE), using three classification algorithms, namely Support Vector Machine (SVM), K-Nearest neighbors (KNN), and Random Forest (RF). It was found that with the handling of data imbalance and the application of the k-fold cross validation method, the three algorithms showed a significant increase in accuracy. Therefore, handling data imbalance is proven to be able to improve the performance of the applied classification algorithms.

Keywords: *Imbalance Data, Support Vector Machine, K-Nearest neighbors, Random Forest, Eastern Indonesia*

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Machine learning merupakan suatu bidang studi yang memberikan pembelajaran kepada mesin atau komputer tanpa diprogram secara eksplisit. Tujuannya yaitu mengajari mesin dari sekumpulan data yang ada, lalu hasil pembelajaran tersebut digunakan untuk menangani suatu data baru agar pengerjaannya lebih efisien [1]. *Machine learning* membantu kebutuhan manusia dalam membuat keputusan berdasarkan pengetahuan dan pola yang didapat dari data. Salah satu teknik dalam *machine learning* adalah klasifikasi. Menurut Elly (2015) dalam [2], klasifikasi adalah metode pembentukan pola untuk mengkategorikan *item* berdasarkan fitur yang digunakan. Nilai variabel prediktor dan variabel target diperlukan dalam algoritma klasifikasi untuk proses pembelajaran. Namun, data yang tidak seimbang atau *imbalance* data menjadi suatu masalah yang sering terjadi selama proses klasifikasi [3][4].

Ketidakseimbangan jumlah data antar kelas dapat menimbulkan kesalahan dalam hasil klasifikasi. Kesalahan mengklasifikasi dapat menyebabkan masalah yang serius dalam pengambilan keputusan dan kebijakan yang akan diambil selanjutnya. Selain itu, di berbagai penelitian menunjukkan performa algoritma klasifikasi menggunakan data yang tidak seimbang menjadi tidak bagus. Kelas mayoritas akan lebih akurat dibandingkan kelas minoritas [5], sehingga pengklasifikasian hanya membahas kelas mayoritas dan mengabaikan kelas minoritas [6]. Oleh karena itu, dalam klasifikasi dengan data yang tidak seimbang sebagian besar akan salah mengklasifikasikan kelas minoritas dan akan menghasilkan lebih banyak kesalahan dari sisi biaya, waktu, dan evaluasi risiko [7].

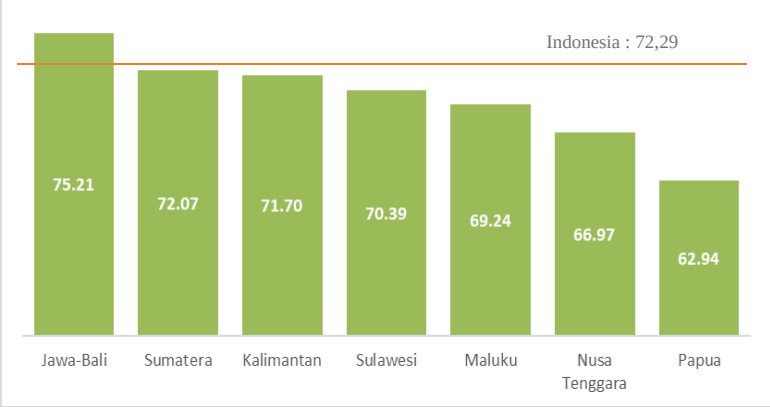
Salah satu metode yang dapat menangani *imbalance* data adalah *Synthetic Minority Over-sampling Technique* (SMOTE). Dengan menggunakan nilai sintesis melalui *sampling* ulang dari data kelas minoritas, didapatkan data hasil SMOTE yang lebih seimbang [8]. Adapun algoritma pengklasifikasian yang memanfaatkan *machine learning* dan diterapkan pada penelitian ini diantaranya *Support Vector Machine* (SVM), *K-Nearest neighbors* (KNN), dan *Random Forest* (RF). Penggunaan algoritma tersebut didasarkan pada algoritma terbaik yang didapatkan dari peneliti terdahulu, sehingga pada penelitian ini akan dilakukan perbandingan antar algoritma terbaik untuk melengkapi dari penelitian yang telah ada. Tidak sampai disitu saja, karena adanya permasalahan dalam *imbalance* data, maka pada penelitian ini akan membandingkan tiga algoritma klasifikasi sebelum dan sesudah dilakukan SMOTE.

Penerapan klasifikasi dilakukan berdasarkan indikator pembangunan manusia. Indeks Pembangunan Manusia (IPM) adalah alat untuk mengukur keberhasilan pembangunan manusia. IPM pertama kali diperkenalkan oleh United Nations Development Programme (UNDP) pada tahun 1990, dan dari waktu ke waktu dipublikasikan dalam laporan tahunan HDR [9]. IPM terdiri dari tiga aspek dasar yang digunakan untuk mengukur kualitas pembangunan manusia [10], diantaranya aspek pendidikan, kehidupan yang layak, dan aspek kesehatan. Indikator Harapan Lama Sekolah (HLS) dan Rata-rata Lama Sekolah (RLS) mencerminkan aspek pendidikan IPM. Kemampuan masyarakat untuk memenuhi kebutuhan dasar yang diperlukan, berdasarkan rata-rata pengeluaran per kapita disesuaikan, menunjukkan dari aspek kehidupan yang layak. Terakhir dari aspek kesehatan diwakili oleh indikator Angka Harapan Hidup (AHH).

Akan tetapi, aspek gender juga berpengaruh untuk mengukur keberhasilan daerah dalam meningkatkan kemampuan manusianya [2]. Gender adalah pembagian kedudukan, peran, kerja, dan pembagian tanggung jawab antara pria dan wanita. Pembagian ini telah dipandang menurut kepribadian alamiah yang melekat dalam diri setiap gender dan sejalan dengan adat istiadat, kepercayaan, aturan, atau kebiasaan masyarakat [11]. Selain dari sisi gender, tingkat kemiskinan juga menjadi hal yang perlu diperhatikan dalam mengukur pembangunan manusia [2]. Pada dasarnya, kemiskinan merupakan kondisi masyarakat yang tidak mampu menyanggupi kepentingan pokok manusia, baik untuk memenuhi kebutuhan pangan maupun non pangan, yang diukur melalui pendekatan pengeluaran. Ketika pengeluaran menyebar lebih luas di kelompok penduduk miskin, akan lebih banyak ketimpangan yang terjadi di daerah tersebut [12]. Indikator yang dapat mengukur dimensi gender yaitu Indeks Pemberdayaan Gender (IDG), dan untuk mengukur dimensi kemiskinan dapat direpresentasikan melalui Indeks Keparahan Kemiskinan (IKK) [2].

 ISSN:2302-8149
e-ISSN:2540-9719

Badan Pusat Statistik (BPS) sudah membagi IPM ke dalam empat kelompok, diantaranya IPM yang berkategori rendah (IPM < 60), kategori menengah ke bawah ($60 \leq$ IPM < 70), kategori tinggi ($70 \leq IPM < 80$), dan berkategori sangat tinggi ($IPM \geq 80$) [13]. Pencapaian suatu daerah di Indonesia dalam peningkatan kemampuan manusianya juga bergantung dari rancangan kerja dan kebijakan yang diambil pemerintah. Oleh karena itu, kebutuhan akan data menjadi sangat penting sebagai capaian dan evaluasi selanjutnya.



Sumber: BPS (2021), diolah

Gambar 1. IPM Berdasarkan Pulau Tahun 2021

Berdasarkan Gambar 1, pembangunan manusia di Indonesia sudah dapat dikategorikan tinggi yaitu sebesar 72.29. Namun, jika diliat berdasarkan keberadaan penduduk di suatu pulau, tingkat pembangunan manusianya pun berbeda-beda. IPM di Pulau Jawa dan Bali tergolong tinggi di atas IPM nasional, sedangkan di luar selain pulau tersebut masih berada di bawah IPM nasional seperti pada Pulau Sulawesi, Maluku, Nusa Tenggara, bahkan Pulau Papua, yang mana wilayah tersebut masuk ke dalam Kawasan Timur Indonesia (KTI).

Kemampuan dan potensi penduduk di Indonesia masih perlu untuk ditingkatkan dengan memfokuskan pada wilayah yang memiliki IPM jauh di bawah nasional seperti pada wilayah KTI. Berbagai program untuk mengedepankan kualitas manusia harus dilakukan secara tepat dan berkelanjutan. Untuk itu diperlukan suatu sistem keputusan yang dapat mengklasifikasikan IPM ke dalam kategori tertentu di masing-masing daerah secara akurat dan cermat.

Dengan demikian, dilakukannya penelitian ini untuk mengklasifikasikan kabupaten/kota di KTI tahun 2021 berdasarkan indikator pembangunan manusia dengan membandingkan hasil evaluasi sebelum dan sesudah diterapkannya SMOTE pada algoritma SVM, KNN, dan RF. Dengan adanya penelitian ini dapat mendukung keputusan pemerintah dalam menentukan langkah strategis kedepannya untuk meningkatkan kualitas SDM di KTI.

2 Tinjauan Literatur

Penerapan metode SVM pernah dilakukan oleh Yusharsah, Dur, dan Cipta [14] dalam mengklasifikasikan IPM di Sumatera Utara yang disusun dari empat indikator IPM. Penelitian ini menerapkan *k-fold cross validation* yang dibagi berdasarkan dua set data, yaitu data *training* dan *testing*. Selain itu juga menggunakan kernel *Radial Basic Function* (RBF) untuk mengklasifikasikan IPM menjadi kategori rendah dan kategori tinggi. Didapatkan hasil tingkat akurasi yang cukup baik sebesar 79.31%, dengan parameter C = 1, α = 0.25, ε = 0.1, γ = 0.5 dan λ = 0.5. Penelitian serupa juga pernah dilakukan oleh Darsyah [15] dengan menggunakan metode KNN untuk mengklasifikasikan IPM di Jawa Tengah. Tujuan dari penelitian tersebut adalah membandingkan tingkat akurasi antar penggunaan nilai k yang berbeda. Hasil yang didapat menunjukkan bahwa dengan k sebesar 5 dan 10 merupakan klasifikasi terbaik, karena memiliki tingkat akurasi, sensitivitas, dan spesivitas yang paling tinggi.

Selain itu, Mauludiyah [16] menggunakan algoritma *random forest* untuk klasifikasi IPM pada kabupaten/kota di Indonesia yang dibagi menjadi empat kategori, yaitu tinggi, sangat tinggi, rendah, dan sedang. Didapatkan hasil yaitu banyaknya pohon yang terbentuk (*ntree*) dan *random sample* yang

 ISSN:2302-8149
e-ISSN:2540-9719

terambil setiap percobaan (*mtry*) berturut-turut sebanyak 100 dan 2, karena memiliki nilai *error* yang paling kecil. Tingkat akurasi yang dihasilkan sebesar 93.69%, dengan variabel yang paling penting dalam peningkatan IPM adalah Pendapatan Per Kapita.

Penggunaan berbagai metode klasifikasi pernah dilakukan oleh Kemala dan Wijayanto [2]. Penelitian ini melakukan klasifikasi IPM di Indonesia menggunakan empat indikator berbeda, yaitu IDG, IKK, RLS, dan PPD. Algoritma klasifikasi yang diterapkan antara lain RF sebagai metode *Bagging*, serta algoritma *Naive Bayes*, *KNN*, dan C4.5 *Decision Tree* sebagai metode *non-ensemble*. Didapat hasil dari pemodelan melalui data *testing* bahwa metode terbaik adalah RF dengan mtry = 2 dan ntree = 500. Berbeda dengan Polat [17] yang membandingkan sebelas metode klasifikasi untuk mengklasifikasikan 100 negara di dunia sesuai dengan indikator dalam *Human Development Index* (HDI). Hasil penelitian menunjukkan bahwa metode klasifikasi terbaik adalah *Multilayer Perceptron* dengan tingkat akurasi tertinggi sebesar 88%. Selain itu, PDB per kapita US$ ditemukan sebagai variabel yang paling efektif dalam menentukan tingkat HDI suatu negara.

Beberapa peneliti telah membandingkan hasil klasifikasi dari data tidak seimbang dengan data yang telah dilakukan teknik *resampling*. Seperti dalam penelitiannya Haryawan dan Ardhana [18] yang menggunakan model klasifikasi SVM dengan dilakukan K-Means SMOTE dan SMOTE dan dibandingkan nilainya dengan aslinya. Didapatkan bahwa K-Means SMOTE menunjukkan performa nilai spesifisitas, sensitivitas, dan akurasi yang lebih tinggi dibandingkan SMOTE dan data asli, namun hasil evaluasi SMOTE lebih mendekati data aslinya. Pertiwi [8] membandingkan kinerja KNN menggunakan SMOTE dan KNN tanpa SMOTE untuk mengklasifikasi penyakit diabetes. Dari nilai akurasi yang didapat, penerapan SMOTE dapat meningkatkan akurasi tertinggi sebesar 8.25%. Hal ini juga diperkuat oleh penelitannya Mursianto dkk [19] yang mengklasifikasi prediksi hujan dengan membandingkan antara metode RF, XGBoost, SMOTE RF, dan SMOTE XGBoost. Hasil penelitian membuktikan bahwa penggunaan SMOTE dapat meningkatkan tingkat akurasi dan *recall* dari hasil klasifikasi.

Berdasarkan penelitian klasifikasi IPM yang pernah dilakukan sebelumnya, hampir semuanya menggunakan indikator pembentuk IPM, namun belum memasukkan indikator tambahan yang dapat memengaruhi pembangunan manusia. Selain itu, lokus penelitian masih dilakukan secara nasional atau hanya mengambil satu provinsi saja, untuk level provinsi belum memperhatikan wilayah yang memiliki IPM paling rendah. Selanjutnya dari penggunaan metode pengklasifikasian secara umum memiliki hasil metode terbaik yang serupa, perbandingan hasil SMOTE hanya menggunakan satu metode klasifikasi, namun belum ada yang membandingkan antar metode terbaik dan menggunakan SMOTE di setiap penelitian.

Fokus pada penelitian ini terdapat pada penambahan indikator pembangunan manusia yang dilihat dari aspek gender dan kemiskinan. Lokus penelitian yang diterapkan yaitu kabupaten/kota di wilayah KTI tahun 2021 dengan mempertimbangkan wilayah yang memiliki IPM di bawah nasional. Penggunaan metode klasifikasi dilakukan dengan perbandingan antar metode terbaik dari penelitian terdahulu yaitu metode SVM, KNN, dan RF, serta dibandingkan juga dengan hasil klasifikasi menggunakan SMOTE.

3 Metode Penelitian

Data sekunder digunakan dalam penelitian ini dengan sumber data dari Badan Pusat Statistik (BPS). Periode waktu penelitian ini pada tahun 2021, menyesuaikan dengan data terbaru yang dipublikasikan oleh BPS. Penelitian ini menganalisis 176 kabupaten/kota di Kawasan Timur Indonesia, yang terdiri dari 12 provinsi, meliputi Provinsi Nusa Tenggara Timur, Nusa Tenggara Barat, Sulawesi Selatan, Sulawesi Tengah, Gorontalo, Sulawesi Barat, Sulawesi Tenggara, Sulawesi Utara, Maluku Utara, Maluku, Papua, dan Provinsi Papua Barat. Seluruh dataset, *syntax R*, dan hasil SMOTE dapat diakses melalui https://github.com/Adhammalay16/DatasetSistemasi.git. Tabel 1 berikut menunjukkan variabel yang digunakan dalam penelitian ini.

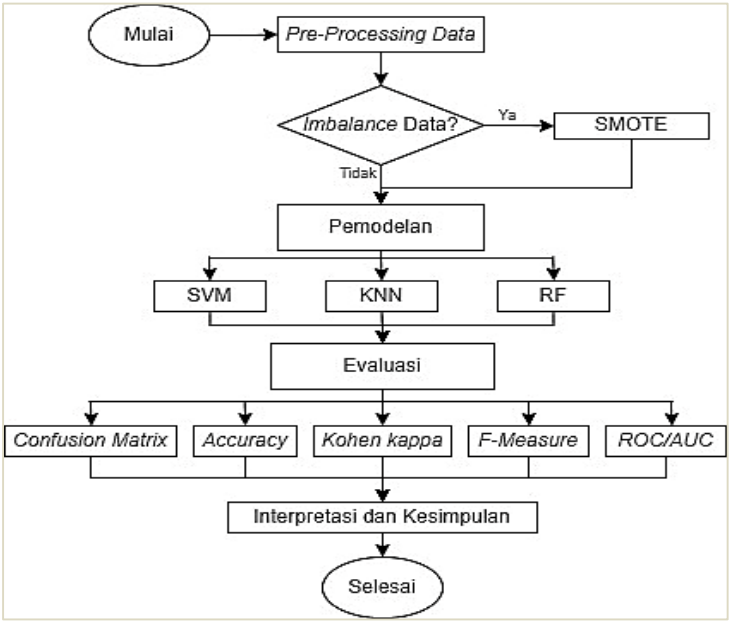
Tabel 1. Variabel Penelitian

Variabel	Keterangan Variabel	Tipe Data

 ISSN:2302-8149
e-ISSN:2540-9719

IPM	Indeks Pembangunan Manusia	Kategorik
RLS	Rata-Rata Lama Sekolah	Numerik
PENG	Pengeluaran Per Kapita Disesuaikan	Numerik
UHH	Umur Harapan Hidup	Numerik
HLS	Harapan Lama Sekolah	Numerik
IDG	Indeks Pemberdayaan Gender	Numerik
IKK	Indeks Keparahan Kemiskinan	Numerik

Adapun Gambar 2 berikut menunjukkan diagram alur yang merupakan tahapan penelitian ini.



Gambar 2. Diagram Alur Penelitian

A. *Pre-Processing* data

Pre-processing data adalah proses menyediakan data sebelum diproses melalui metode analisis yang digunakan dalam data mining [20]. Berikut ini adalah prosedur *preprocessing* data dalam penelitian ini [21].

1. Data *cleaning*. Pembersihan data yaitu melakukan pemeriksaan apakah terdapat *missing value* atau tidak pada setiap dataset yang digunakan. Jika terdapat data yang kosong maka dilakukan imputasi dengan menggunakan median atau mean dari atribut yang mengandung nilai yang hilang. Untuk data dengan sebaran normal dapat menggunakan mean, sedangkan jika menceng atau sebarannya tidak normal dapat menggunakan median.
2. Data *integration*. Integrasi yang dilakukan yaitu menggabungkan data dari sumber dataset atau *file* yang berbeda menjadi satu dataset yang lengkap, sehingga dapat digunakan dalam pemodelan.
3. Data *reduction*. Setelah data yang digunakan lengkap, langkah selanjutnya yaitu memeriksa ada tidaknya *outlier* yang berpengaruh dalam data. Dalam statistik, *outlier* adalah amatan yang jauh dari amatan lainnya. Outlier dapat terjadi karena adanya variasi dalam pengukuran atau dapat mengindikasikan kesalahan dalam proses imputasi data.
4. Data *Transformation*. Penggunaan atribut yang berbeda-beda menjadikan satuan data yang diperoleh berbeda pula. Transformasi menggunakan *z-score* diperlukan untuk mendapatkan data dengan rentang nilai yang sama. *Z-score* adalah metode normalisasi yang dihitung melalui nilai mean dan standar deviasi data.

   Persamaan *z-score* dapat dituliskan pada persamaan (1) berikut.

$$Z - score = (X - \mu)/\sigma \quad (1)$$

 ISSN:2302-8149
e-ISSN:2540-9719

Dengan keterangan:
X : Nilai amatan
μ : Rata-rata populasi
σ : Standar deviasi populasi

B. Penanganan *Imbalance* Data

Ketika penyebaran data antar kategori memiliki jumlah yang berbeda jauh atau timpang, maka akan terjadi ketidakseimbangan data. Hal ini akan memberikan pengaruh buruk dalam pemodelan menggunakan algoritma klasifikasi, dimana kelas minoritas seringkali diklasifikasikan sebagai kelas mayoritas dan menyebabkan salah klasifikasi [5]. Untuk menangani hal ini, dapat diterapkan metode SMOTE. Metode ini membangkitkan data untuk menambah dan menyeimbangkan kelas minoritas dengan melakukan *resampling* dari kelas minoritas menggunakan nilai tetangga terdekat.

Persamaan untuk menghitung data sintesis tersebut terdapat pada persamaan (2) berikut [22].

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \quad (2)$$

Dengan keterangan:
x_i : data yang akan di replikasi
x_{knn} : data dengan jarak terdekat dari x_i
δ : nilai acak berkisar antara 0 dan 1

C. Pemodelan

Pada tahapan ini dilakukan pembagian data seluruh observasi menjadi dua bagian, untuk pembentukan model menggunakan data *training*, sedangkan untuk menguji performa model yang dibentuk menggunakan data *testing* [20]. Metode *k-fold cross validation* digunakan dalam pembagian data dengan persentase sebanyak 80 persen untuk data pelatihan dan 20 persen untuk data pengujian.

K-fold cross validation merupakan teknik untuk memangkas data menjadi k subset dengan ukuran yang sama [23]. Setiap subset digunakan untuk pengujian data *testing* dan diulang sebanyak k kali, hasil akhir pengukuran adalah rata-rata dari k kali pengujian. Penelitian ini menetapkan jumlah k sebanyak 10, hal ini dilandaskan oleh temuan dari berbagai penelitian dan pembenaran teoritis yang telah dilakukan, menunjukkan bahwa metode *cross validation* dengan sepuluh perulangan merupakan metode yang paling efektif untuk mendapatkan performa terbaik [23][24][25]. Setelah kedua jenis dataset tersebut terbentuk, dilakukan *center* dan *scale* untuk mengurangi keragaman dalam data, kemudian dilanjutkan dengan pemodelan menggunakan tiga algoritma yang telah ditetapkan pada tujuan penelitian.

1. *Support Vector Machine* (SVM)

SVM merupakan algoritma klasifikasi *supervised* untuk mencari *hyperplane* (fungsi pemisah) dengan margin terbesar. *Hyperplane* terbaik adalah yang berada di tengah-tengah dua objek dari kedua kelas [26]. Persamaan *hyperplane* dapat dituliskan dalam persamaan (3) berikut.

$$W.X + b = 0 \quad (3)$$

Dengan keterangan:
W : bobot vektor $\{w_1, w_2, \dots, w_n\}$
n : jumlah atribut
b : skalar (bias)

Algoritma SVM umumnya diterapkan dalam klasifikasi data linear, akan tetapi SVM dapat juga diterapkan dalam kasus nonlinier secara efisien dengan memanfaatkan *kernel trick* untuk memperoleh *hyperplane* yang optimal dari dataset yang berbeda [27]. Secara matematis, fungsi kernel dituliskan pada persamaan (4).

$$K(X_i, X_j) = \phi(X_i) \times \phi(X_j) \quad (4)$$

Dalam pelaksanaannya, terdapat banyak keuntungan menggunakan *kernel trick*, salah satunya ketika menetapkan *support vector* cukup mengidentifikasi dari fungsi kernelnya saja, tanpa melihat bentuk fungsi dari non linear ϕ. Pada Tabel 2 terdapat macam-macam bentuk fungsi kernel beserta persamaannya [25].

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 2. Fungsi Kernel

Jenis Fungsi Kernel	Persamaan
Kernel polinomial *degree* ke-h	$K(X_i, X_j) = (X_i . X_j + 1)^h$
Kernel fungsi basis radial gaussian	$K(X\_i, X\_j) = e^{-\left\|X_i - X_j\right\|^2 / 2\sigma^2}$
Kernel Sigmoid	$K(X_i, X_j) = tanh(\kappa X_i . X_j - \delta)$

Setiap fungsi kernel menghasilkan pengklasifikasi nonlinier yang berbeda di ruang input. Namun dalam praktiknya, hasil dari akurasinya cenderung tidak berbeda signifikan antar kernel yang digunakan.

2. *K-Nearest neighbors* (KNN)

Pengklasifikasi dengan *nearest-neighbors* (tetangga terdekat) berlandaskan pada analogi pembelajaran, melalui perbandingan antara data *training* dan *testing* yang serupa. Data *training* dijelaskan oleh n atribut. Setiap data *training* mewakili dan disimpan dalam sebuah titik di ruang n-dimensi. Ketika terdapat data yang baru, pengklasifikasi KNN mencari ruang pola untuk k data *training* dengan jarak terpendek dari data yang baru. Dalam hal ini k data *training* adalah k "tetangga terdekat" dari data yang baru atau tidak diketahui [25].

Perhitungan jarak merupakan hal yang esensil dalam KNN untuk mengetahui objek mana saja dalam data *training* yang menjadi tetangga terdekat. Terdapat berbagai jenis jarak yang dapat digunakan, diantaranya yaitu jarak *Euclidean*, *Manhattan*, dan jarak *Minkowski*. Jarak dapat dihitung dengan rumus pada persamaan (5) berikut.

$$d(x_i, x_j) = \sqrt[g]{|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \cdots + |x_{in} - x_{jn}|^g} \quad (5)$$

Dengan keterangan:
x_i, x_j : dua set data yang akan dihitung jaraknya
$g = 1$, untuk menghitung jarak *Manhattan*
$g = 2$, untuk menghitung jarak *Euclidean*
$g = \infty$, untuk menghitung jarak *Chebysev*

Penentuan nilai k yang baik dapat dilakukan melalui eksperimen. Dimulai dengan k = 1, lalu dihitung tingkat kesalahan pengklasifikasinya menggunakan data *testing*. Proses ini akan terus diulang setiap kali ada penambahan jumlah k untuk memungkinkan satu tetangga lagi. Nilai k terbaik yaitu ketika tingkat kesalahan yang dihasilkan paling kecil [25].

3. *Random Forest* (RF)

Random forest merupakan suatu algoritma klasifikasi yang berasal dari gabungan pohon klasifikasi (*Classification and Regression Tree*) yang saling independen dan berasal dari sebaran yang sama melalui proses *voting* (jumlah terbanyak) [16]. Metode ini merupakan penyempurnaan metode *bagging*. *Bagging* atau *bootstrap aggregating* adalah metode dengan merata-ratakan suatu set prediktor (menggabungkan beberapa *decision trees*) untuk mendapatkan model terbaik. Adapun rumus untuk metode *bagging* terdapat dalam persamaan (6) [28].

$$f(x) = \sum_{m=1}^{M} \frac{1}{M} f_m(X) \quad (6)$$

Dengan keterangan:
f_m : pohon ke-m

Namun, ketika mengulang kembali algoritma yang sama pada himpunan data yang berbeda dapat menghasilkan prediktor yang sangat berkorelasi, selain itu tingkat keragaman data menjadi sulit untuk dikurangi. Untuk itu digunakan metode RF untuk meningkatkan akurasi. Pada algoritma RF, variabel dipilih secara acak untuk masuk ke dalam pohon. Dengan begitu, bentuk pohon dalam RF antar satu dan lainnya akan bervariasi, sehingga antar pohon semakin independen dan akurasinya pun akan meningkat. Korelasi yang kecil juga menyebabkan varians RF lebih kecil dari metode *bagging*.

 ISSN:2302-8149
e-ISSN:2540-9719

D. *Evaluasi*

1. *Confusion Matrix*, adalah metode untuk mengevaluasi kebaikan dari sebuah model klasifikasi menggunakan bantuan tabel. Dengan menggunakan *confusion matrix* dapat diperoleh berbagai kriteria model klasifikasi diantaranya akurasi, presisi, recall, dan *f-measure* [29]. Karena terdapat 4 kategori maka *confusion matrix* yang dihasilkan akan berukuran 4x4.

Adapun komponen yang terdapat dalam *confusion matrix* adalah sebagai berikut.

a. *True Positive* (TP)
   Sebuah nilai diperkirakan bernilai positif dan pada kenyataannya benar positif
b. *False Positive* (FP)
   Sebuah nilai diperkirakan bernilai positif namun pada kenyataannya bernilai negatif
c. *True Negative* (TN)
   Sebuah nilai diperkirakan bernilai negatif dan pada kenyataannya benar negatif
d. *False Negative* (FN)
   Sebuah nilai diperkirakan bernilai negatif namun pada kenyataannya bernilai positif

2. *Accuracy*, adalah metode evaluasi kualitas model klasifikasi yang paling umum dan sederhana, yaitu dengan menilai kinerja model dalam mengkorelasikan hasil pemodelan dengan atribut data yang digunakan. Nilai akurasi didapatkan dari proporsi antara jumlah observasi yang terklasifikasi secara benar dengan jumlah seluruh observasi [17].

Persamaan (7) adalah rumus untuk menghitung tingkat akurasi.

$$Akurasi = (TP + TN)/(TP + FP + TN + FN) \quad (7)$$

3. *Kohen kappa*, adalah alternatif untuk mengukur tingkat klasifikasi yang nilainya berkisar antara −1 hingga 1. Kappa digunakan sebagai ukuran yang sangat berguna untuk masalah *multiclass* karena lebih sederhana, dan untuk mengukur keakuratan pengklasifikasi sambil mengkompensasi keberhasilan acak [21].

Untuk menghitung Kohen kappa adalah dengan menggunakan *confusion matrix* yang didapatkan dari hasil klasifikasi. Secara khusus, ukuran kappa Cohen dapat diperoleh dengan menggunakan rumus pada persamaan (8) berikut [15].

$$kappa = \frac{n\sum_{i=1}^{C} x_{ii} - \sum_{i=1}^{C} x_{i.}x_{.i}}{n^2 - \sum_{i=1}^{C} x_{i.}x_{.i}} \quad (8)$$

Dengan keterangan:

x_{ii} : jumlah sel pada diagonal utama
n : jumlah sampel pada data
C : jumlah kategori
$x_{i.}$: jumlah baris
$x_{.i}$: jumlah kolom

4. *F-Measure*, adalah rata-rata harmonis dari presisi dan *recall*. Hal ini karena presisi dan *recall* tidak cukup untuk menggambarkan hasil perbandingan secara substansial [17]. Berikut adalah rumus untuk menghitung *f-measure* pada persamaan (9).

$$F = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

Dengan keterangan yaitu persamaan (10) dan (11):

$$precision = (TP)/(TP + FP) \quad (10)$$
$$recall = (TP)/(TP + FN) \quad (11)$$

5. *ROC/AUC*, adalah kurva untuk memvisualisasikan perbandingkan antar model-model klasifikasi. Di dalam kurva ROC terdapat sumbu vertikal, sumbu horizontal, dan garis diagonal. Sumbu vertikal menggambarkan *true positive rate*, sedangkan sumbu horizontal menggambarkan *false positive rate*. Ukuran akurasi model ditentukan dari luas daerah yang

 ISSN:2302-8149
e-ISSN:2540-9719

berada di bawah kurva ROC. Model klasifikasi dikatakan semakin akurat apabila luas daerah yang dihasilkan mendekati 1 [29].

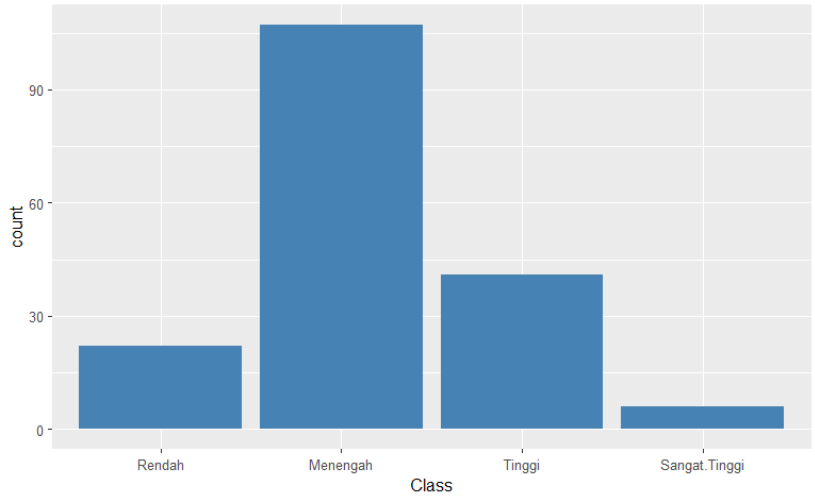
Berikut adalah kriteria untuk klasifikasi AUC.

- 0.90 - 1.00 = klasifikasi yang sangat baik
- 0.80 - 0.90 = klasifikasi yang baik
- 0.70 - 0.80 = klasifikasi yang wajar
- 0.60 - 0.70 = klasifikasi yang buruk
- 0.50 - 0.60 = klasifikasi yang gagal

4 Hasil dan Pembahasan

A. Gambaran Umum Pembangunan Manusia

Untuk melihat gambaran umum dari pembangunan manusia di KTI, dilakukan pengkategorian variabel IPM ke dalam empat kategori, yaitu kategori rendah, menengah, tinggi, dan sangat tinggi.



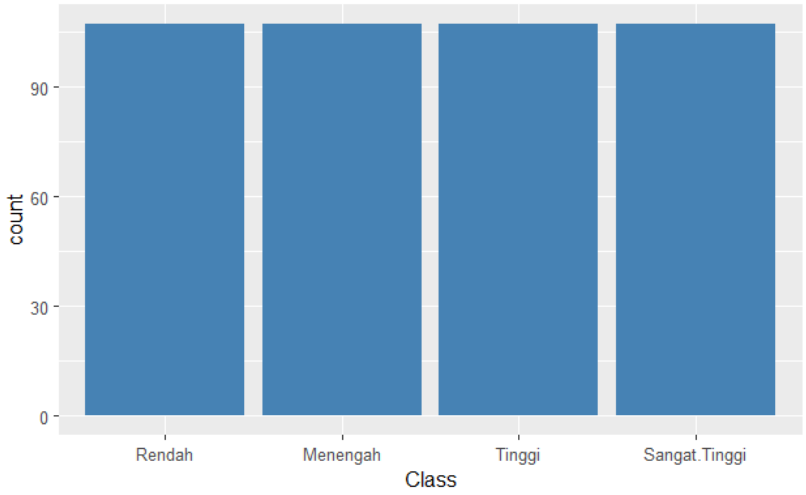
Sumber: BPS, 2021 (diolah)

Gambar 3. Distribusi Tingkat Pembangunan Manusia

Berdasarkan Gambar 3, tingkat pembangunan manusia di KTI masih didominasi oleh kabupaten/kota yang memiliki IPM berkategori menengah, yaitu sebanyak 107 kabupaten/kota. Terdapat 22 kabupaten/kota yang masih memiliki tingkat pembangunan manusia yang rendah, 44 kabupaten/kota yang berkategori tinggi, dan sedikit sekali berkategori sangat tinggi yang hanya sebanyak 6 kabupaten/kota. Hal ini menunjukkan pemerataan pembangunan manusia antar wilayah di KTI masih dirasakan oleh kabupaten/kota tertentu saja.

Dari hasil pengkategorian menunjukkan bahwa terjadi *imbalanced* pada dataset yang digunakan. Oleh karena itu, peneliti memutuskan untuk mengatasi ketidakseimbangan kelas dengan algoritma SMOTE.

 ISSN:2302-8149
e-ISSN:2540-9719



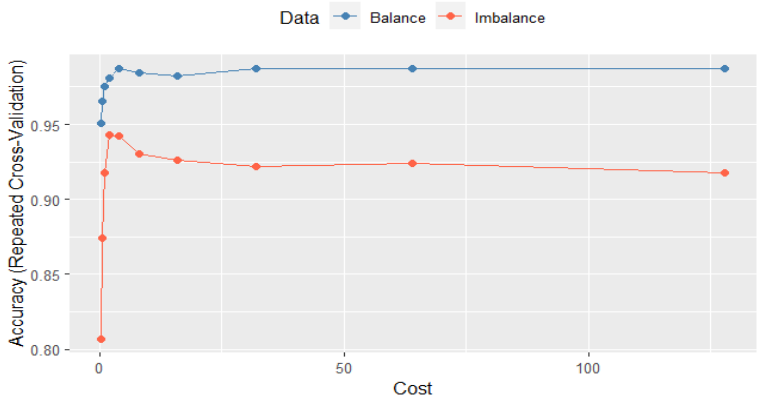
Sumber: Hasil Pengolahan

Gambar 4. Distribusi Pembangunan Manusia Setelah Dilakukan SMOTE

Berdasarkan Gambar 4 menunjukkan bahwa kelas dalam pembangunan manusia telah seimbang dengan *rasio imbalance* 100 persen. Dengan demikian, data yang digunakan sudah layak untuk dianalisis dengan algoritma klasifikasi. Untuk melihat keakuratan dari masing-masing algoritma akan dilakukan perbandingan antara data yang masih *imbalance* dengan data yang sudah *balance* melalui proses SMOTE.

B. Support Vector Machine (SVM)

Klasifikasi IPM yang pertama menggunakan SVM dengan parameter kernel *Radial Basis Function* (RBF). Tuning parameter (sigma) yang digunakan untuk data *imbalance* yaitu sebesar 0.2441, sedangkan untuk data yang sudah *balance* sebesar 0.4155. Parameter cost (C) yang digunakan untuk kedua data sebanyak 10, yaitu 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, dan 128. Adapun hasil yang didapat sebagai berikut.



Sumber: Hasil Pengolahan

Gambar 5. Perbandingan Tingkat Akurasi SVM

Berdasarkan Gambar 5 menunjukkan bahwa untuk data yang sudah *balance* memiliki tingkat akurasi yang lebih tinggi di setiap parameter C daripada data yang masih *imbalance*. Untuk *balance* data, tingkat akurasi tertinggi terjadi saat C sama dengan 32 yaitu sebesar 98.76 persen, sedangkan untuk *imbalance* data, tingkat akurasi tertinggi terjadi saat C sama dengan 2 yaitu sebesar 94.31 persen. Oleh karena itu, dengan dilakukannya penanganan *imbalance* data dapat meningkatkan akurasi untuk klasifikasi pembangunan manusia menggunakan SVM kernel RBF.

Dari Gambar 5 juga menunjukkan bahwa setiap adanya penambahan pada nilai C akan menghasilkan tingkat akurasi yang sama (konstan), sehingga dengan C sebesar 32 dan 2 sudah

 ISSN:2302-8149
e-ISSN:2540-9719

menunjukkan hasil yang paling optimal. Setelah didapatkan hasil pemodelan, dilanjutkan dengan memprediksi data *testing* dari model yang memiliki kriteria terbaik.

Tabel 3. *Confusion Matrix* Hasil Prediksi Model SVM

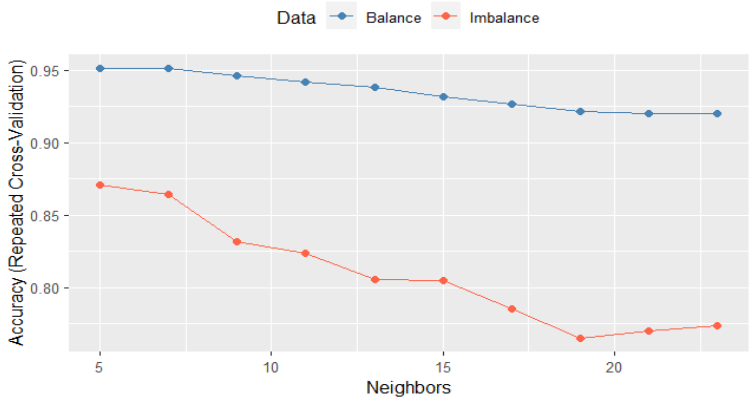
Data	Prediksi	Data *Testing*			
		Rendah	Menengah	Tinggi	Sangat Tinggi
Imbalance	Rendah	4	0	0	0
	Menengah	0	21	0	0
	Tinggi	0	0	8	0
	Sangat Tinggi	0	0	0	1
Balance	Rendah	21	0	0	0
	Menengah	0	19	0	0
	Tinggi	0	2	21	0
	Sangat Tinggi	0	0	0	21

Sumber: Hasil Pengolahan

Berdasarkan hasil prediksi yang disajikan pada Tabel 3, menunjukkan bahwa dari 34 data *testing imbalance*, seluruhnya terklasifikasi dengan tepat sesuai kategorinya. Hal ini dapat dilihat dari semua nilai berada pada diagonal utama dan nilai di luar diagonal utama memiliki angka 0. Dari 84 data *testing balance*, hanya 2 nilai yang kurang tepat hasil klasifikasinya yaitu pada nilai kategori IPM menengah yang diprediksi masuk ke dalam kategori IPM yang tinggi.

C. K-Nearest neighbors (KNN)

Hasil klasifikasi IPM dengan KNN menggunakan jumlah tetangga atau *neighbors* (k) sebanyak 10, yaitu 5, 7, 9, 11, 13, 15, 17, 19, 21, dan 23. Adapun hasil yang didapat sebagai berikut.



Sumber: Hasil Pengolahan

Gambar 6. Perbandingan Tingkat Akurasi KNN

Gambar 6 menunjukkan bahwa tingkat akurasi untuk *balance* data lebih tinggi di setiap tetangga (k) yang digunakan daripada *imbalance* data. Pada *balance* data, tingkat akurasi tertinggi terjadi saat jumlah k sama dengan 5 yaitu sebesar 95.17 persen, sedangkan untuk *imbalance* data tingkat akurasi tertinggi juga terjadi saat k sama dengan 5, namun akurasi yang didapat hanya sebesar 87.13 persen. Oleh karena itu, dapat dikatakan dengan dilakukannya penanganan *imbalance* data meningkatkan akurasi untuk klasifikasi pembangunan manusia menggunakan KNN.

Dari kedua data yang digunakan sama-sama menunjukkan bahwa, setiap adanya penambahan jumlah k akan menghasilkan tingkat akurasi yang semakin menurun, sehingga penggunaan jumlah k sebesar 5 sudah menunjukkan hasil yang paling optimal. Setelah didapatkan hasil pemodelan, dilanjutkan dengan memprediksi data *testing* dari model yang memiliki kriteria terbaik.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 4. *Confusion Matrix* Hasil Prediksi Model KNN

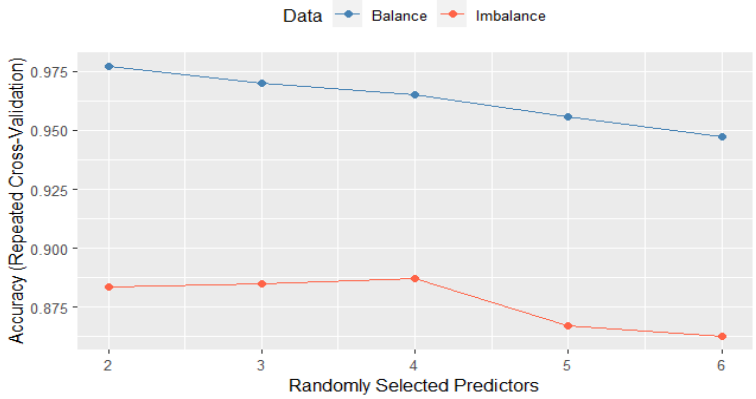
Data	Prediksi	Data *Testing*			
		Rendah	Menengah	Tinggi	Sangat Tinggi
Imbalance	Rendah	4	0	0	0
	Menengah	0	21	1	0
	Tinggi	0	0	7	0
	Sangat Tinggi	0	0	0	1
Balance	Rendah	21	0	0	0
	Menengah	0	14	0	0
	Tinggi	0	7	18	0
	Sangat Tinggi	0	0	3	21

Sumber: Hasil Pengolahan

Tabel 4 menunjukkan bahwa dari hasil prediksi menggunakan 34 data *testing imbalance*, sebanyak 33 data yang terklasifikasi dengan tepat dan hanya 1 yang terjadi salah klasifikasi, yaitu diprediksi masuk ke kategori menengah seharusnya masuk ke dalam kategori tinggi. Sementara itu, dari 84 data *testing balance* terdapat sebanyak 74 nilai yang terklasifikasi dengan tepat dan 10 nilai prediksi yang terjadi salah klasifikasi, yaitu sebanyak 7 nilai yang diprediksi masuk ke kategori tinggi seharusnya masuk ke dalam kategori menengah, dan terdapat 3 nilai yang diprediksi masuk ke kategori sangat tinggi namun sebenarnya masuk ke kategori tinggi.

D. Random Forest (RF)

Algoritma terakhir yaitu menggunakan *random forest* untuk mengklasifikasikan IPM dengan *mtry* sebanyak 5, yaitu 2, 3, 4, 5, dan 6. Peneliti tidak menggunakan 10-*fold* seperti algoritma lain dikarenakan *default* grid pada algoritma RF hanya sampai 5 parameter kompleksitas yang unik. Adapun hasil yang didapat sebagai berikut.



Sumber: Hasil Pengolahan

Gambar 7. Perbandingan Tingkat Akurasi RF

Tingkat akurasi RF yang disajikan dalam Gambar 7 menunjukkan perbedaan yang cukup signifikan antara *imbalance* data dengan *balance* data. Tingkat akurasi yang lebih baik di setiap *mtry* dimiliki oleh *balance* data, dengan akurasi tertinggi sebesar 97.74 persen yang terjadi pada saat *mtry* sama dengan 2. Berbeda dengan *imbalance* data yang hanya memiliki akurasi tertinggi sebesar 88.73 persen dan terjadi pada *mtry* sama dengan 4.

Saat *mtry* melewati 2 untuk *balance* data dan melewati 4 untuk *imbalance* data, terjadi penurunan yang signifikan. Hal ini terlihat dari nilai akurasi yang terus mengalami penurunan ketika terjadi penambahan *mtry*, sehingga masing-masing *mtry* yang digunakan sudah menunjukkan hasil yang paling optimal. Setelah didapatkan hasil pemodelan, dilanjutkan dengan memprediksi data *testing* dari model yang memiliki kriteria terbaik.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 5. *Confusion Matrix* Hasil Prediksi Model RF

Data	Prediksi	Data *Testing*			
		Rendah	**Menengah**	**Tinggi**	**Sangat Tinggi**
Imbalance	Rendah	4	0	0	0
	Menengah	0	21	0	0
	Tinggi	0	0	7	1
	Sangat Tinggi	0	0	1	0
Balance	Rendah	21	0	0	0
	Menengah	0	20	0	0
	Tinggi	0	1	20	0
	Sangat Tinggi	0	0	1	21

Sumber: Hasil Pengolahan

Hasil prediksi model RF pada Tabel 5, menunjukkan bahwa dari 34 data *testing imbalance*, hanya sebanyak 2 nilai yang terjadi salah klasifikasi, yaitu 1 nilai yang diprediksi masuk ke kategori tinggi, seharusnya masuk ke dalam kategori sangat tinggi. Sama halnya dengan *imbalance* data, pada data yang sudah *balance* terdapat 1 nilai yang diprediksi masuk ke kategori tinggi, seharusnya nilai tersebut masuk ke dalam kategori menengah, selain itu terdapat 1 nilai diprediksi masuk ke kategori sangat tinggi yang seharusnya masuk ke dalam kategori tinggi.

E. Evaluasi dan Perbandingan

Berdasarkan hasil analisis yang telah dilakukan menggunakan ketiga metode, dilakukan evaluasi dengan melihat dari nilai akurasi, kappa, *f-measure*, dan AUC. Berikut adalah ringkasan hasil evaluasi dari ketiga metode.

Tabel 6. Hasil Evaluasi Model Data *Testing*

Metode Evaluasi	*Imbalance*			*Balance*		
	SVM	**KNN**	**RF**	**SVM**	**KNN**	**RF**
Akurasi	1.0000	0.9706	0.9412	0.9762	0.8810	0.9762
Kappa	1.0000	0.9452	0.8927	0.9683	0.8413	0.9683
F-Measure	1.0000	0.9775	0.7188	0.9761	0.8790	0.9762
AUC	1.0000	0.9896	0.9062	0.9603	0.8452	0.9760

Sumber: Hasil Pengolahan

Pada Tabel 6 terlihat bahwa dengan menggunakan data *imbalance* menunjukkan algoritma SVM memiliki tingkat akurasi, kappa, *f-measure*, dan AUC yang paling tinggi yaitu sebesar 100 persen. Sementara itu, dengan data yang *balance* menunjukkan algoritma SVM dan RF memiliki tingkat akurasi dan kappa yang sama, tetapi dari hasil *f-measure* dan AUC yang paling tinggi terdapat pada algoritma RF. Jika dilihat dari proses pembentukan model dengan data *training*, ketiga algoritma menunjukkan peningkatan tingkat akurasi yang signifikan, sehingga penanganan dari *imbalance* data terbukti mampu meningkatkan performa dari algoritma klasifikasi yang diterapkan. Peningkatan tertinggi dari hasil prediksi terdapat pada pengukuran *f-measure* di algoritma RF yang meningkat sebesar 25.74 persen.

5 Kesimpulan

Kesimpulan yang dapat diambil dari penelitian yang telah dilakukan adalah, dengan adanya penanganan dari *imbalance* data dan diterapkannya metode *k-fold cross validation* didapatkan pemodelan yang paling optimum dari ketiga algoritma dan menunjukkan peningkatan akurasi yang signifikan. Hasil evaluasi performa ketiga algoritma untuk data *imbalance* menghasilkan bahwa algoritma SVM memiliki tingkat akurasi, kappa, *f-measure*, dan AUC yang paling tinggi, sedangkan

ketika dilakukan penanganan dengan SMOTE menunjukkan bahwa algoritma *random forest* memiliki tingkat *f-measure* dan AUC yang paling tinggi. Oleh karena itu, penanganan *imbalance* data terbukti mampu meningkatkan performa dari algoritma klasifikasi yang diterapkan dan algoritma RF melalui SMOTE mampu menghasilkan klasifikasi kabupaten/kota di KTI berdasarkan indikator pembangunan manusia dengan sangat baik daripada algoritma SVM dan KNN.

Adapun saran yang dapat diterapkan untuk penelitian selanjutnya yaitu dapat melakukan perbandingan dengan algoritma lainnya, agar didapatkan hasil evaluasi yang lebih baik dan akurat. Selain itu, dapat pula dibandingkan untuk klasifikasi pembangunan manusia dengan dua kategori.

Ucapan Terima Kasih

Pada kesempatan ini penulis ingin mengucapkan terima kasih kepada Bapak Yuliagnis Transver Wijaya S.ST, M.Sc., selaku dosen mata kuliah data mining atas segala waktu, bimbingan, dan ilmu yang diberikan kepada penulis sampai kepada penulisan dalam penelitian ini.

Referensi

[1] M. Batta, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 18, no. 8, hal. 381–386, 2018, doi: 10.21275/ART20203995.

[2] I. Kemala dan A. W. Wijayanto, "Perbandingan Kinerja Metode Bagging dan Non-Ensemble Machine Learning pada Klasifikasi Wilayah di Indonesia menurut Indeks Pembangunan Manusia," *J. Sist. dan Teknol. Inf.*, vol. 9, no. 2, hal. 269–275, 2021, doi: 10.26418/justin.v9i2.44166.

[3] M. C. Untoro dan J. L. Buliali, "Penanganan Imbalance Class Data Laboratorium Kesehatan dengan Majority Weighted Minority Oversampling Technique," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 4, no. 1, hal. 23–29, 2018.

[4] S. Mutmainah, "Penanganan Imbalance Data Pada Klasifikasi Kemungkinan Penyakit Stroke," *J. SNATi*, vol. 1, no. 1, hal. 10–16, 2021.

[5] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan KNN," *J. Inf. Syst. Dev.*, vol. 3, no. 1, hal. 44–49, 2018.

[6] H. Ali, M. Najib, M. Salleh, R. Saedudin, dan K. Hussain, "Imbalance class problems in data mining : A review Imbalance class problems in data mining : a review," no. April, hal. 1552–1563, 2019, doi: 10.11591/ijeecs.v14.i3.pp1552-1563.

[7] M. R. Longadge, M. S. S. Dongre, dan D. L. Malik, "Class Imbalance Problem in Data Mining : Review," *Int. J. Comput. Sci. Netw.*, vol. 2, no. 1, 2013.

[8] A. G. Pertiwi, "Perbandingan Kinerja Algoritma K-Nearest Neighbor Menggunakan SMOTE dan Algoritma K-Nearest Neighbor tanpa SMOTE dalam Diagnosis Penyakit Diabetes pada Data Tidak Seimbang." Semarang, 2019.

[9] D. Programme, *Human Development Report 1990*. United Nations Development Programme (UNDP), 1990.

[10] BPS, "Indeks Pembangunan Manusia," 2022. https://www.bps.go.id/subject/26/indeks-pembangunan-manusia.html

[11] BPS, "Gender," 2022. https://www.bps.go.id/subject/40/gender.html

[12] BPS, "Kemiskinan dan Ketimpangan," 2022. https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html

[13] BPS, *Indeks Pembangunan Manusia 2020*. Jakarta: Badan Pusat Statistik, 2020. [Daring]. Tersedia pada: https://www.bps.go.id/publication/2021/04/30/8e777ce2d7570ced44197a37/indeks-pembangunan-manusia-2020.html

[14] A. Yusharsah, S. Dur, dan H. Cipta, "Penerapan Metode Support Vector Machine dalam Klasifikasi Indeks Pembangunan Manusia di Sumatera Utara," *Math Educ. J.*, vol. 06, no. 01, hal. 12–19, 2022.

[15] M. Y. Darsyah, "Klasifikasi Indeks Pembangunan Manusia (IPM) Dengan Pendekatan K-Nearset Neighbor (KNN)," in *Seminar Nasional Pendidikan, Sains dan Teknologi Fakultas*

 ISSN:2302-8149
e-ISSN:2540-9719

Matematika dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang, 2020, no. October 2017, hal. 29–35.

[16] K. Mauludiyah, "Klasifikasi Indeks Pembangunan Manusia Kabupaten/Kota di Indonesia Menggunakan Metode Random Forest," 2020.

[17] E. Polat, "The Classification of Countries' Human Development Index Level Under Economic Inequality by Using Data Mining Classification Algorithms," *Rom. Stat. Rev.*, no. 4, hal. 27–44, 2021.

[18] C. Haryawan dan Y. M. K. Ardhana, "Analisa Perbandingan Teknik Oversampling SMOTE pada Imbalanced Data," *JIRE (Jurnal Inform. Rekayasa Elektron.*, vol. 6, no. 1, hal. 73–78, 2023.

[19] G. A. Mursianto, I. M. Falih, M. Irfan, T. Sakinah, dan D. S. Prasvita, "Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan," *Senamika*, vol. 2, no. 2, hal. 41–50, 2021.

[20] M. Fathurrahman dan N. Qisthi, "Klasifikasi Indeks Pembangunan Manusia (IPM) di Pulau Sumatera Pada Dataset Multi-Class Dengan Metode Artificial Neural Network (ANN)," in *Prosiding Seminar Nasional Fisika 7.0*, 2021, hal. 377–384.

[21] S. García, J. Luengo, dan F. Herrera, *Data Preprocessing in Data Mining*. Springer, 2015. doi: 10.1007/978-3-319-10247-4.

[22] A. N. Kasanah, Muladi, dan U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *J. Rekayasa Sist. dan Teknol. Inf.*, vol. 3, no. 2, hal. 196–201, 2019.

[23] D. A. Nasution, H. H. Khotimah, dan N. Chamidah, "Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma KNN," *J. Comput. Eng. Syst. Sci.*, vol. 4, no. 1, hal. 78–82, 2019.

[24] I. A. Nikmatun dan I. Waspada, "Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *J. SIMETRIS*, vol. 10, no. 2, hal. 421–432, 2019.

[25] J. Han dan M. Kamber, *Data Mining Concepts and Techniques - Second Edition*. San Francisco: Morgan Kaufmann, 2006.

[26] C. A. Pamungkas dan W. W. Widiyanto, "Klasifikasi Indeks Pembangunan Manusia di Indonesia Tahun 2022 dengan Support Vector Machine," *J. Ilm. Sist. Inf. dan Ilmu Komput.*, vol. 2, no. 3, hal. 139–145, 2022.

[27] F. Fauzi, "K-Nearset Neighbor (KNN) dan Support Vector Machine (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah," *J. MIPA*, vol. 40, no. 2, hal. 118–124, 2017.

[28] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. London: Massachusetts Institute of Technology, 2012.

[29] S. Pramana, B. Yuniarto, S. Mariyah, I. Santoso, dan R. Nooraeni, *Data Mining dengan R Konsep Serta Implementasi*. Jakarta: In Media, 2018.