Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

Analisis Prediksi Stunting pada Balita dengan Menggunakan *Random Forest* dan *Naïve Bayes*

Chintya Annisah Solin*, Putu Harry Gunawan

Fakultas Informatika, Universitas Telkom, Bandung, Indonesia
Email: [1,*]chintyaanisa@student.telkomuniversity.ac.id, [2]phgunawan@telkomuniversity.ac.id

Abstrak−Penelitian ini bertujuan untuk membandingkan performa algoritma Random Forest dan Naïve Bayes dalam memprediksi stunting pada balita menggunakan data dari Dinas Kesehatan Kabupaten Bekasi. Proses analisis diawali dengan pembersihan data, normalisasi, dan sampling menggunakan metode Adaptive Synthetic Sampling (ADASYN) untuk menangani ketidakseimbangan data, diikuti oleh validasi dengan Stratified K-Fold Cross Validation. Implementasi algoritma menunjukkan bahwa Random Forest memiliki akurasi tertinggi sebesar 89.62% dan F1-Score 89.09%. Naïve Bayes Gaussian menghasilkan akurasi 88.72% dan F1-Score 88.81%, sedangkan Naïve Bayes Bernoulli memiliki performa lebih rendah dengan akurasi 67.83% dan F1-Score 69.72%. Random Forest menunjukkan keunggulan dalam mengatasi noise dan data tidak seimbang, menjadikannya pilihan optimal untuk prediksi stunting. Sementara itu, performa Naïve Bayes dipengaruhi oleh karakteristik datanya, di mana variasi Gaussian lebih sesuai untuk data kontinu. Hasil penelitian ini memberikan wawasan bahwa pemilihan algoritma yang tepat, terutama pada data yang tidak seimbang, sangat penting untuk meningkatkan akurasi prediksi. Penelitian ini juga merekomendasikan perhatian lebih pada preprocessing data untuk memastikan kualitas prediksi yang optimal, khususnya pada kelas minoritas.

Kata Kunci: Stunting; Naïve Bayes; Random Forest; Adasyn; K-Fold

Abstract−This study aims to compare the performance of the Random Forest and Naïve Bayes algorithms in predicting stunting in toddlers using data from the Bekasi District Health Office. The analysis process begins with data cleaning, normalization, and sampling using the Adaptive Synthetic Sampling (ADASYN) method to handle data imbalance, followed by validation with Stratified K-Fold Cross Validation. The implementation of the algorithm shows that Random Forest has the highest accuracy of 89.62% and an F1-Score of 89.09%. Naïve Bayes Gaussian produces an accuracy of 88.72% and an F1-Score of 88.81%, while Naïve Bayes Bernoulli has a lower performance with an accuracy of 67.83% and an F1-Score of 69.72%. Random Forest shows advantages in overcoming noise and imbalanced data, making it an optimal choice for stunting prediction. Meanwhile, the performance of Naïve Bayes is influenced by the characteristics of the data, where the Gaussian variation is more suitable for continuous data. The results of this study provide insight that choosing the right algorithm, especially on imbalanced data, is very important to improve prediction accuracy. This study also recommends more attention to data preprocessing to ensure optimal prediction quality, especially for minority classes.

Keywords: Stunting; Naïve Bayes; Random Forest; Adasyn; K-Fold

1. PENDAHULUAN

Stunting merupakan salah satu permasalahan gizi kronis yang menjadi perhatian global, terutama di negara-negara berkembang, termasuk Indonesia. Masalah ini terjadi akibat kurangnya asupan gizi dalam jangka waktu panjang, yang berdampak pada terganggunya pertumbuhan fisik dan perkembangan anak[1]. Anak yang mengalami stunting umumnya memiliki tinggi badan lebih rendah dibandingkan anak seusianya, serta berisiko mengalami gangguan kognitif dan kesehatan di masa depan[2]. Situasi ini sangat mengkhawatirkan, terutama bagi anak-anak pada periode emas 1000 hari pertama kehidupannya (HPK), yang dihitung sejak masa kehamilan hingga anak berusia dua tahun. Periode HPK merupakan waktu yang sangat kritis karena pada masa inilah fondasi perkembangan otak, sistem imun, dan organ vital lainnya terbentuk.

Menurut data yang dikutip dari *upk.kemkes.go.id*, pemerintah melalui Kementerian Kesehatan mengumumkan bahwa prevalensi stunting di Indonesia mengalami penurunan yang cukup signifikan dalam beberapa tahun terakhir. Pada tahun 2021, angka prevalensi stunting berada di level 24,4%, dan berhasil menurun menjadi 21,6% pada tahun 2022. Penurunan ini merupakan hasil dari berbagai upaya pemerintah dan kolaborasi dengan berbagai pihak dalam program pencegahan stunting yang melibatkan edukasi gizi, pemberian makanan tambahan, hingga perbaikan layanan kesehatan ibu dan anak[3].

Secara lokal, pemerintah daerah juga terus menunjukkan komitmen dalam mengatasi masalah stunting. Salah satu contohnya adalah Kabupaten Bekasi, yang menargetkan penurunan prevalensi stunting hingga 14% pada tahun 2024, sebagaimana dikutip dari *prokopim.bekasikab.go.id*. Target ini cukup ambisius, mengingat penurunan angka stunting sebesar 3% per tahun yang konsisten dalam tiga tahun terakhir. Hal ini menunjukkan bahwa program pencegahan stunting di daerah ini telah memberikan hasil yang positif, sekaligus menjadi motivasi untuk terus meningkatkan kualitas program-program intervensi gizi dan kesehatan[4].

Meskipun berbagai upaya telah dilakukan dan hasilnya mulai terlihat, stunting masih menjadi ancaman serius bagi masa depan bangsa. Oleh karena itu, penelitian lebih lanjut mengenai stunting diperlukan untuk memahami dinamika prevalensinya, mengevaluasi efektivitas program yang telah berjalan, serta memprediksi tren di masa mendatang. Prediksi ini menjadi penting karena dapat membantu pemerintah dalam merancang kebijakan yang lebih tepat sasaran, terutama dalam menghadapi tantangan-tantangan baru seperti pandemi, perubahan iklim, atau krisis ekonomi yang dapat memengaruhi ketahanan pangan dan kesehatan masyarakat.

Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

Dalam era teknologi informasi, pemanfaatan kecerdasan buatan (AI) dan *machine learning* (ML) telah menjadi solusi inovatif dalam menangani berbagai permasalahan, termasuk dalam bidang kesehatan masyarakat. Teknologi ini memungkinkan analisis data yang lebih mendalam dan akurat, sehingga menghasilkan wawasan yang relevan untuk mendukung pengambilan keputusan. Dalam konteks penelitian stunting, algoritma ML dapat digunakan untuk melakukan prediksi dan klasifikasi berdasarkan dataset yang ada, seperti data antropometri anak, status gizi, kondisi kesehatan ibu, hingga faktor lingkungan.

Beberapa penelitian sebelumnya telah menunjukkan keberhasilan algoritma ML dalam memprediksi dan mengklasifikasikan data terkait stunting. Menurut Indah Pratiwi Putri dan rekan-rekannya. (2024),perbandingan antara tiga algoritma ML, yaitu *Naive Bayes*, *K-Nearest Neighbors* (KNN), dan *Random Forest*, menunjukkan bahwa *Random Forest* memberikan performa terbaik dengan akurasi 87,75%. Algoritma ini diikuti oleh KNN dengan akurasi 84,5%, dan *Naive Bayes* dengan akurasi 83,2%. Temuan ini menunjukkan bahwa *Random Forest* memiliki keunggulan dalam menangkap pola data yang kompleks, meskipun algoritma lain juga menunjukkan performa yang cukup baik [5]. Penelitian lain yang dilakukan oleh Fadellia Azahra dan rekan-rekannya memperkuat temuan ini. Mereka melaporkan bahwa model *Random Forest* mampu mencapai akurasi sebesar 97,88%, angka yang sangat tinggi dan menunjukkan potensi besar algoritma ini dalam menganalisis data stunting [6]. Hasil serupa juga ditemukan oleh Muhammad Ghiyaats Daffa, yang membandingkan algoritma *Random Forest*, KNN, dan *Boosted KNN*. Dalam penelitian tersebut, *Random Forest* kembali menunjukkan akurasi tertinggi sebesar 97,76%, dengan skor f1 sebesar 97,70%[7].

Namun, di balik keberhasilan algoritma *Random Forest*, terdapat pertanyaan menarik mengenai potensi algoritma lain yang lebih sederhana, seperti *Naive Bayes*. Algoritma ini memiliki beberapa varian, termasuk *Gaussian Naive Bayes* dan *Bernoulli Naive Bayes*, yang masing-masing memiliki kelebihan dalam menangani jenis data yang berbeda. *Gaussian Naive Bayes*, misalnya, dirancang untuk menangani data kontinu, sedangkan *Bernoulli Naive Bayes* lebih cocok untuk data biner. Dengan pendekatan yang berbeda ini, penelitian lebih lanjut diperlukan untuk mengevaluasi apakah salah satu varian *Naive Bayes* mampu bersaing atau bahkan melampaui performa *Random Forest* dalam konteks tertentu.

Penelitian ini bertujuan untuk membandingkan dua algoritma utama, yaitu *Naive Bayes* dan *Random Forest*, dalam melakukan prediksi stunting. Pada algoritma *Naive Bayes*, penelitian ini akan mengeksplorasi dua variannya, yaitu *Gaussian Naive Bayes* dan *Bernoulli Naive Bayes*. Perbandingan ini tidak hanya bertujuan untuk mencari model dengan akurasi terbaik, tetapi juga untuk memahami karakteristik masing-masing algoritma, termasuk kelebihan dan kekurangannya dalam konteks analisis data stunting.

Pemilihan algoritma *Random Forest* didasarkan pada kemampuannya untuk menangani data dengan jumlah variabel yang besar dan kompleksitas yang tinggi. Algoritma ini menggunakan pendekatan ensemble learning, di mana keputusan diambil berdasarkan hasil gabungan dari banyak pohon keputusan (*decision trees*). Keunggulan ini membuat *Random Forest* sangat efektif dalam mengatasi overfitting dan memberikan hasil yang stabil.

Di sisi lain, *Naive Bayes* menawarkan keunggulan dalam kesederhanaan dan efisiensi komputasi. Dengan asumsi independensi antar variabel, algoritma ini mampu memberikan hasil yang cukup baik meskipun dengan sumber daya komputasi yang terbatas. Hal ini menjadikan *Naive Bayes* sebagai pilihan yang menarik untuk diterapkan pada kasus-kasus di mana data yang tersedia relatif kecil atau ketika kecepatan pemrosesan menjadi prioritas.

Selain aspek teknis, penelitian ini juga diharapkan memberikan kontribusi praktis dalam upaya pencegahan stunting di Indonesia. Dengan memahami kekuatan dan kelemahan masing-masing algoritma, hasil penelitian ini dapat digunakan untuk mengembangkan sistem pendukung keputusan yang lebih efektif dalam memprediksi risiko stunting. Sistem ini nantinya dapat diintegrasikan dengan program-program pemerintah, seperti Posyandu, untuk memantau kondisi gizi anak secara real-time dan memberikan intervensi yang tepat waktu.

Kontribusi penelitian ini tidak hanya terbatas pada aspek teknis, tetapi juga mencakup dampak sosial yang lebih luas. Dengan prediksi yang lebih akurat, diharapkan program pencegahan stunting dapat lebih terarah, sehingga sumber daya yang tersedia dapat digunakan secara optimal. Pada akhirnya, penelitian ini bertujuan untuk mendukung visi Indonesia bebas stunting pada tahun 2045, sesuai dengan target pemerintah untuk menciptakan generasi emas yang sehat, cerdas, dan produktif.

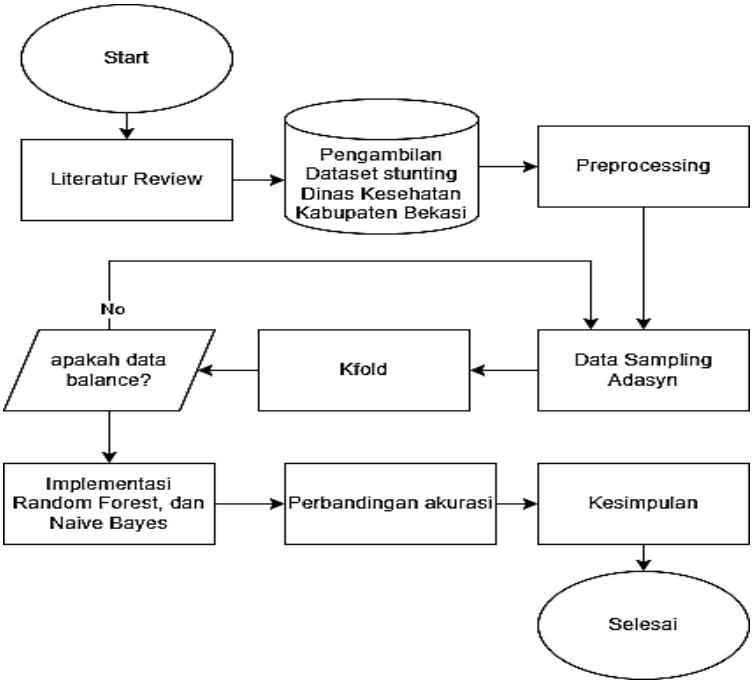
Dengan demikian, penelitian ini tidak hanya memberikan nilai tambah dalam ranah akademik, tetapi juga memiliki implikasi praktis yang signifikan. Kombinasi antara teknologi modern dan intervensi berbasis masyarakat diharapkan mampu menjadi solusi yang efektif dalam mengatasi masalah stunting di Indonesia.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan melalui serangkaian tahapan yang dirancang secara sistematis untuk memastikan validitas dan akurasi hasil yang diperoleh. Tahapan-tahapan tersebut mencakup langkah awal seperti literatur review hingga analisis akhir menggunakan algoritma yang telah dipilih. Penjelasan mendetail mengenai tahapan penelitian dapat dilihat pada Gambar 1 dan uraian berikut.

Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999



Gambar 1. Tahapan Penelitian

Berikut adalah penjelasan rinci mengenai tahapan-tahapan yang ada dalam penelitian ini, yang tercantum pada Gambar 1, mulai dari literatur review hingga kesimpulan akhir yang dihasilkan.

1. Literatur Review
   Tahapan pertama dalam penelitian ini adalah melakukan kajian pustaka atau literatur review. Proses ini bertujuan untuk memahami latar belakang masalah yang akan diteliti serta mencari metode dan algoritma yang relevan untuk digunakan dalam penelitian. Literatur yang dikaji mencakup jurnal, artikel ilmiah, laporan resmi, dan sumber terpercaya lainnya yang membahas topik stunting serta algoritma yang digunakan untuk klasifikasi data. Hasil dari literatur review ini menjadi dasar dalam menentukan langkah-langkah penelitian dan memberikan landasan teori yang kuat untuk analisis selanjutnya.
2. Pengambilan Dataset
   Dataset yang menjadi objek penelitian diambil dari Dinas Kesehatan Kabupaten Bekasi. Dataset ini berisi informasi terkait kondisi stunting pada anak, seperti usia, berat badan, tinggi badan, dan indikator lainnya yang relevan. Tahap ini sangat penting karena kualitas dan kelengkapan dataset akan sangat memengaruhi hasil analisis. Dataset yang diperoleh harus mencerminkan populasi yang sebenarnya agar hasil penelitian dapat menggambarkan kondisi di lapangan. Selain itu, proses ini juga melibatkan pengelolaan administrasi untuk memastikan data yang diambil sesuai dengan etika penelitian dan perlindungan privasi.
3. Preprocessing Data
   Setelah dataset diperoleh, langkah selanjutnya adalah melakukan preprocessing data untuk membersihkan data dari masalah-masalah yang dapat memengaruhi hasil analisis, seperti menghapus missing value, yaitu nilai yang hilang atau tidak terisi, yang dapat ditangani dengan menghapus data yang tidak lengkap atau mengisi nilai yang hilang menggunakan metode imputasi. Selain itu, nilai kosong (null) atau Not a Number (NaN) perlu dihapus atau diisi ulang agar dataset menjadi bersih. Selanjutnya, dilakukan normalisasi dan standarisasi data untuk memastikan semua variabel memiliki skala yang sama, yang penting agar algoritma seperti Naive Bayes dan Random Forest tidak terpengaruh oleh perbedaan skala variabel. Proses preprocessing ini memastikan dataset yang digunakan berkualitas tinggi, terstruktur, dan siap untuk tahap analisis berikutnya.
4. Data Sampling Menggunakan ADASYN
   Tahapan berikutnya adalah menangani ketidakseimbangan data dalam dataset menggunakan metode Adaptive Synthetic Sampling (ADASYN). Ketidakseimbangan data terjadi ketika jumlah data pada kelas tertentu, seperti stunting, jauh lebih kecil atau lebih besar dibandingkan kelas lainnya, seperti normal. Tujuan ADASYN adalah untuk menghasilkan data sintetis pada kelas minoritas agar distribusi data menjadi lebih seimbang. Hal ini penting karena ketidakseimbangan data dapat menyebabkan algoritma machine learning lebih condong memprediksi kelas mayoritas, sehingga akurasi untuk kelas minoritas menurun. Proses ADASYN bekerja dengan menganalisis distribusi data dan membuat sampel sintetis berdasarkan jarak dan kepadatan data pada kelas minoritas, menghasilkan dataset yang lebih seimbang dan representatif.
5. K-Fold Cross Validation
   Setelah data diseimbangkan, langkah selanjutnya adalah membagi data menjadi beberapa bagian (fold) menggunakan metode K-Fold Cross Validation. Proses ini bertujuan untuk memastikan bahwa data digunakan secara merata untuk pelatihan (training) dan pengujian (testing). Dalam cara kerja K-Fold, dataset dibagi menjadi

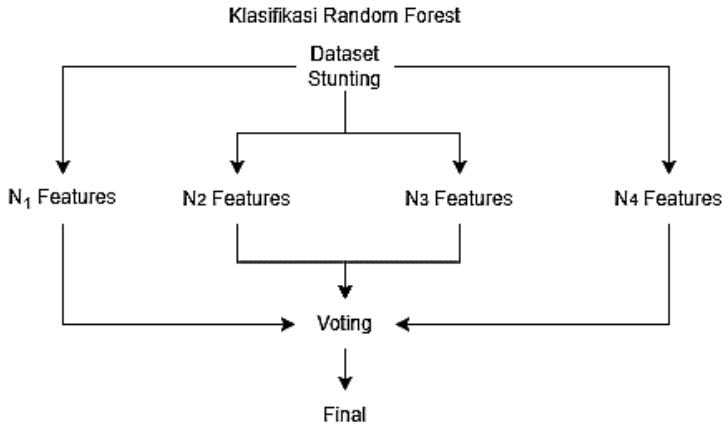
Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

K bagian yang sama besar. Dalam setiap iterasi, salah satu bagian digunakan sebagai data testing, sementara sisanya digunakan sebagai data training. Proses ini diulang sebanyak K kali sehingga setiap bagian data digunakan sebagai data testing sekali. Keuntungan dari K-Fold adalah memastikan bahwa seluruh data digunakan secara adil dalam proses pelatihan dan pengujian. Metode ini juga membantu mencegah overfitting dengan memastikan model diuji pada data yang beragam. Jika hasil K-Fold menunjukkan bahwa distribusi data masih belum seimbang, maka dilakukan sampling ulang untuk memperbaiki ketidakseimbangan tersebut.

6. Implementasi Algoritma
   Pada tahap ini, algoritma Random Forest dan Naive Bayes diterapkan pada dataset yang telah diproses. Kedua algoritma ini dipilih karena memiliki keunggulan masing-masing. Random Forest bekerja dengan membangun beberapa pohon keputusan dan menggabungkan hasilnya untuk membuat prediksi. Algoritma ini dikenal memiliki akurasi yang tinggi, toleransi terhadap data yang tidak seimbang, dan ketahanan terhadap overfitting. Sementara itu, Naive Bayes didasarkan pada teori probabilitas Bayes dan sering digunakan untuk klasifikasi. Dalam penelitian ini, dua varian Naive Bayes digunakan, yaitu Gaussian Naive Bayes untuk data kontinu dan Bernoulli Naive Bayes untuk data biner. Proses implementasi dilakukan dengan melatih data menggunakan kedua algoritma, dan hasil prediksi dari masing-masing model dicatat untuk evaluasi lebih lanjut.
7. Perbandingan Hasil Algoritma
   Setelah algoritma diterapkan, hasil dari kedua model dibandingkan untuk menentukan algoritma mana yang memberikan performa terbaik. Beberapa metrik yang digunakan untuk evaluasi meliputi akurasi, yang mengukur persentase prediksi yang benar dibandingkan dengan data sebenarnya; precision dan recall, yang digunakan untuk mengevaluasi kinerja model pada kelas tertentu, terutama jika dataset tidak seimbang; serta F1-Score, yang merupakan metrik gabungan yang mempertimbangkan precision dan recall, memberikan gambaran yang lebih lengkap tentang kinerja model. Dengan membandingkan hasil dari Random Forest dan Naive Bayes, penelitian ini diharapkan dapat memberikan rekomendasi algoritma yang paling sesuai untuk prediksi stunting berdasarkan dataset yang digunakan.
8. Kesimpulan
   Tahapan terakhir dalam penelitian ini adalah menarik kesimpulan dari seluruh proses yang telah dilakukan. Kesimpulan mencakup hasil akhir perbandingan performa kedua algoritma, algoritma mana yang memiliki akurasi terbaik untuk prediksi stunting, serta implikasi dari hasil penelitian terhadap upaya pencegahan stunting, seperti memberikan rekomendasi model yang efektif untuk analisis data stunting di masa mendatang. Selain itu, kesimpulan juga mencakup saran untuk penelitian selanjutnya, seperti eksplorasi algoritma lain atau penggunaan dataset yang lebih besar dan bervariasi untuk mendapatkan hasil yang lebih general.

2.2 *Random Forest*

Random Forest merupakan salah satu metode Machine Learning yang biasa digunakan untuk melakukan klasifikasi dan regresi dengan menghasilkan hasil akhir berupa pohon keputusan berdasarkan hasil voting yang dilakukan[8]. *Random Forest* merupakan jenis ensemble learning yang menggunakan metode bagging (*Bootstrap Aggregating*) untuk meningkatkan performa akurasi[9]. Algoritma ini memiliki keunggulan yaitu akurasi tinggi, mengurangi overfitting, toleransi terhadap noise yang disebabkan oleh data dan variabel yang tidak relevan sehingga hasil prediksi menjadi stabil karena variasi dalam dataset tidak terlalu mempengaruhi hasil akhir[10]. Berikut adalah contoh dari pengimplementasian *Random Forest*[11]:



Gambar 2. *Random Forest*

Pada gambar 1 diatas dataset Stunting akan memecah menjadi 4N Features yang didapatkan secara acak dari dataset Stunting dan menghasilkan 4 pohon keputusan (*Decision Tree*) tahapan selanjutnya setiap pohon keputusan yang dihasilkan akan memberikan prediksi untuk masing- masing class sesuai dengan inputan data. Setelah pohon

Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

keputusan menghasilkan prediksi selanjutnya algoritma Random Forest akan melakukan voting untuk menentukan prediksi akhir[12].

2.3 *Naïve Bayes*

Naïve Bayes adalah algoritma klasifikasi yang berdasarkan pada Teorema Bayes dengan asumsi independensi fitur. Dalam klasifikasi, Naïve Bayes menghitung probabilitas setiap kelas berdasarkan fitur-fitur yang ada, dengan memanfaatkan Teorema Bayes yang menghubungkan probabilitas bersyarat antara kelas dan fitur[13]. Algoritma ini mengasumsikan bahwa setiap fitur yang digunakan untuk mendeskripsikan observasi adalah independen, mengingat label kelas yang diberikan. Meskipun asumsi independensi ini dianggap 'naïve' atau sederhana, Naïve Bayes terbukti efektif dan sering digunakan dalam banyak aplikasi klasifikasi, terutama pada data besar dan teks[14]. Secara matematis, Naïve Bayes dapat dirumuskan dengan persamaan sebagai berikut:

$$P(A|B) = \frac{P(A|B)\,P(A)}{P(B)} \quad (1)$$

Keterangan:
P(A) = Peluang kejadian A
P(B) = Peluang kejadian B
P(A|B) = Peluang terjadinya peristiwa A, jika peristiwa B telah terjadi
P(B|A) = Peluang terjadinya peristiwa B, jika peluang A telah terjadi

Naive bayes yang digunakan dalam penelitian ini adalah naive bayes gaussian dan naive bayes bernoulli. *Naive Bayes Gaussian* (GNB) merupakan metode klasifikasi yang mengandalkan pendekatan probabilitas dan distribusi gaussian cocok dengan data kontinu dan Naive Bayes Bernoulli dengan data boolean[15].

2.4 *Evaluation Matrix*

Evaluasi matrik digunakan untuk mengukur performa dari suatu model algoritma berdasarkan tujuan tertentu. Dalam penelitian ini, evaluasi metrik digunakan untuk menilai seberapa baik model dalam memprediksi data dengan mengklasifikasikan data menjadi dua kategori, yaitu 'Stunting' dan 'Normal'. Evaluasi ini penting untuk mengetahui sejauh mana algoritma dapat mengidentifikasi dengan akurat kelas-kelas tersebut. Beberapa metrik yang umum digunakan dalam evaluasi klasifikasi termasuk akurasi, precision, recall, dan F1-score, yang memberikan gambaran lengkap mengenai kinerja model dalam memisahkan dua kelas tersebut. Berikut adalah evaluation matrix secara sistematis[16]:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Presisi = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 - Score = 2 * \left(\frac{presisi*recall}{presisi+recall}\right) \quad (5)$$

Rumus 2 sampai 5 merupakan rumus yang digunakan untuk metode evalusi metrik dalam menilai performa dari suatu model algoritma. TP (True Positive) adalah kasus yang diidentifikasi sebagai stunting, TN (True Negative) adalah kasus yang diindentifikasi sebagai normal atau tidak stunting, FP (False Positive) adalah kesalahan dalam memprediksi stunting dan FN (False Negative) adalah kesalahan dalam memprediksi kasus normal atau tidak stunting.

3. HASIL DAN PEMBAHASAN

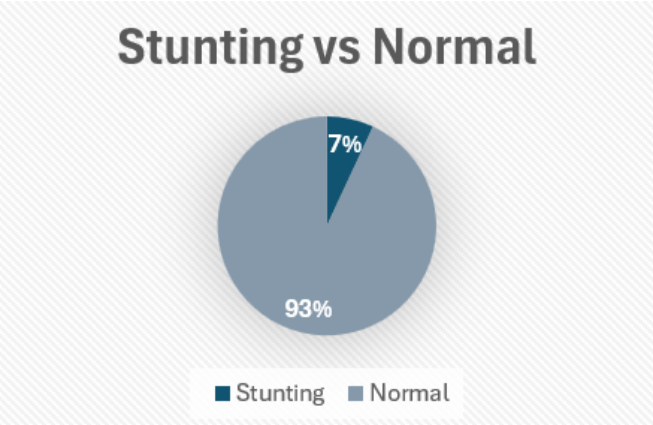
3.1 *Explore Data Analyst*

Dataset yang digunakan dalam penelitian ini adalah data Stunting yang berasal dari Badan Kesatuan Bangsa dan Politik Kab. Bekasi. Dataset ini memiliki data sebanyak 2255 baris data yang dikumpulkan pada bulan April 2024. Di dalam dataset ini, informasi yang ada berupa identitas anak, jenis kelamin, tanggal lahir, berat badan anak ketika lahir, tinggi badan anak ketika lahir, nama orang tua, puskesmas, posyandu, usia saat ukur, tanggal pengukuran, ZZ BB/U, TB/U, ZS TB/U, BB/TB, ZS BB/TB, dan lain-lain.

Dalam penelitian ini, fitur 'TB/U' mencakup kategori yang menggambarkan status tinggi badan anak, dengan nilai yang terbagi menjadi empat kategori, yaitu: 'Tinggi', 'Normal', 'Pendek', dan 'Sangat Pendek'. Sedangkan fitur 'ZS TB/U' merupakan hasil perhitungan skor Z-Score yang menunjukkan apakah anak tersebut mengalami stunting atau tidak, berdasarkan pedoman dan rumus yang dikembangkan oleh WHO untuk menilai status gizi anak. Gambar 3 menyajikan hasil eksplorasi data yang digunakan untuk menganalisis distribusi antara kondisi stunting dan normal dalam dataset. Analisis ini bertujuan untuk memahami pola distribusi data, apakah data stunting dan normal terdistribusi dengan baik, serta untuk memastikan apakah terdapat ketidakseimbangan pada dataset yang dapat

Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

memengaruhi kinerja model dalam proses klasifikasi lebih lanjut. Keseimbangan data yang baik sangat penting dalam mendapatkan hasil prediksi yang lebih akurat.

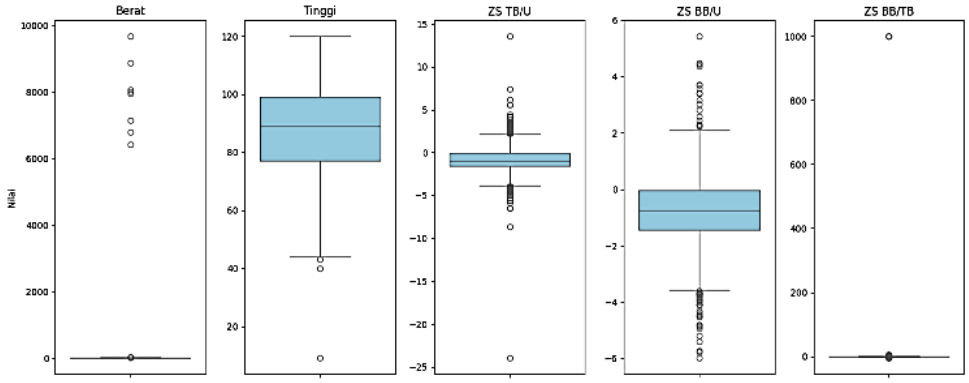


Gambar 3. Explore Data Analyst Stunting vs Normal

Berdasarkan Gambar 3, terlihat bahwa distribusi data antara kategori stunting dan normal masing-masing sebesar 7%. Distribusi ini menunjukkan adanya ketidakseimbangan data, yang dapat memengaruhi kinerja model prediksi, terutama jika algoritma yang digunakan sensitif terhadap distribusi data. Oleh karena itu, langkah-langkah seperti sampling atau penyesuaian distribusi data akan dilakukan untuk memastikan data seimbang sebelum diterapkan ke model. Selain itu, analisis eksplorasi ini juga bertujuan untuk mengidentifikasi potensi anomali atau outlier dalam data yang dapat memengaruhi proses preprocessing maupun hasil akhir dari model.

3.2 *Data Preprocessing*

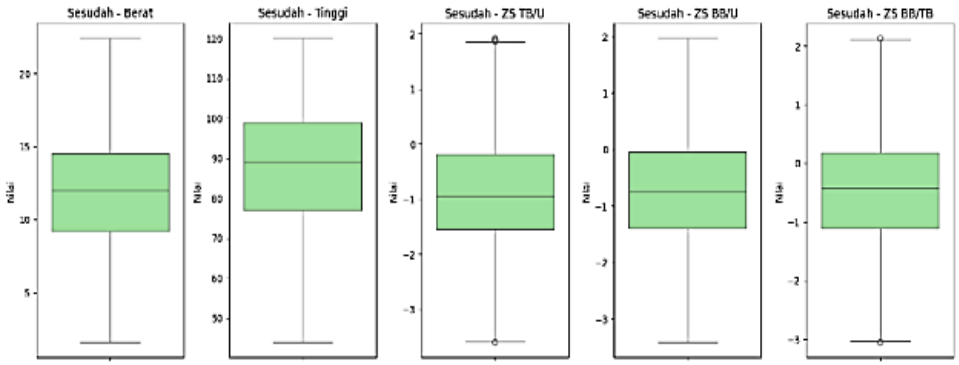
Data preprocessing yang dilakukan dalam penelitian ini melibatkan beberapa tahapan penting untuk mempersiapkan data sebelum digunakan dalam pelatihan model Machine Learning. Langkah pertama adalah pemilihan fitur yang relevan untuk melatih model, di mana hanya fitur-fitur yang dianggap signifikan untuk penelitian ini yang dipilih. Selain itu, proses pembersihan data juga dilakukan dengan menghapus nilai NaN dan menangani missing values agar dataset menjadi lebih konsisten dan siap diproses lebih lanjut. Selanjutnya, dilakukan pelabelan data untuk kategori stunting dan normal, yang kemudian digunakan untuk mendefinisikan variabel independen (X) dan dependen (Y). Meskipun terdapat 33 kolom dalam dataset, hanya 5 kolom yang dipilih sebagai fitur utama yang relevan dalam penelitian ini, yaitu: ‘Berat’, ‘Tinggi’, ‘ZS TB/U’, ‘ZS BB/U’, dan ‘ZS BB/TB’. Berikut ini adalah visualisasi persebaran data numerik setelah proses preprocessing, yang menggambarkan distribusi nilai pada masing-masing fitur yang digunakan.



Gambar 4. *Outliner Preprocessing*

Terlihat pada Gambar 4 bahwa persebaran data sebelum dilakukan preprocessing sangat tidak merata, dengan beberapa kategori data yang tampak lebih dominan dibandingkan yang lain. Ketidakseimbangan ini dapat memengaruhi kinerja model dalam memprediksi kelas yang minoritas, sehingga diperlukan langkah-langkah preprocessing untuk memperbaiki kondisi tersebut. Setelah proses preprocessing dilakukan, data yang awalnya tidak seimbang menjadi lebih merata dan lebih layak digunakan untuk pelatihan model. Berikut ini adalah visualisasi persebaran data setelah dilakukan preprocessing, yang menunjukkan distribusi data yang lebih seimbang dan siap untuk diproses lebih lanjut.

Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999



Gambar 5. *Outliner Preprocessing*

3.3 *Sampling Data dan K-Fold*

Untuk menangani data yang tidak seimbang (imbalanced), perlu dilakukan sampling data. Metode yang digunakan dalam sampling data adalah Adaptive Synthetic Sampling (ADASYN). Metode ini menghasilkan sampel data sintetis untuk data dalam kelas minoritas dalam dataset yang tidak seimbang[17][18]. Dengan menghasilkan data sintetis, distribusi data antara kelas mayoritas dan minoritas menjadi lebih seimbang, yang membantu meningkatkan kinerja model. Setelah dilakukan sampling data, proses selanjutnya adalah Stratified K-Fold Cross Validation (SKCV), di mana metode ini membagi data menjadi sejumlah fold secara merata untuk setiap fold-nya [19][20]. Dengan SKCV, setiap fold memiliki distribusi kelas yang serupa dengan dataset asli, yang memastikan bahwa model diuji pada data yang representatif. Proses ini nantinya akan digunakan dalam pengimplementasian model algoritma machine learning, seperti Random Forest dan Naive Bayes, untuk meningkatkan performa klasifikasi.

3.4 Implementasi *Random Forest*

Algoritma machine learning yang digunakan dalam penelitian ini adalah Random Forest, yang diterapkan menggunakan perpustakaan sklearn dengan Python. Data yang digunakan telah melalui proses fold dengan parameter n_estimator = 10 dan random_state = 42 untuk memastikan kestabilan dan keakuratan hasil prediksi. Random Forest sangat cocok digunakan untuk menangani data yang tidak seimbang, karena saat melakukan prediksi pada kelas minoritas, sering kali terdapat gangguan atau noise. Namun, Random Forest cenderung lebih tahan terhadap noise karena hasil prediksi yang dikeluarkan merupakan gabungan dari banyak pohon keputusan. Dengan demikian, keputusan akhir yang dihasilkan menjadi lebih stabil dan lebih akurat[21][22]. Berikut adalah hasil dari algoritma Random Forest yang diterapkan pada data yang telah melalui proses sampling dan k-fold, yang disajikan dalam tabel di bawah ini.

Tabel 1. Confussion Matrix Random Forest

		Predict Values	
		Stunting	Normal
Actual Values	Stunting	591	103
	Normal	41	653

Tabel 1. Confusion Matrix Random Forest menunjukkan hasil klasifikasi model terhadap data status stunting dan normal. Model ini berhasil mengklasifikasikan 591 data stunting dengan benar sebagai True Positive, yang menunjukkan bahwa model efektif dalam mendeteksi data stunting. Namun, terdapat 103 data stunting yang salah diklasifikasikan sebagai normal, yang merupakan False Negative. Di sisi lain, model berhasil mengklasifikasikan 653 data normal dengan benar sebagai True Negative, yang mengindikasikan bahwa model dapat dengan baik membedakan data normal. Namun, ada 41 data normal yang salah diklasifikasikan sebagai stunting, yang disebut sebagai False Positive. Hasil ini memberikan gambaran mengenai kinerja model dalam mengidentifikasi kedua kelas (stunting dan normal), serta menunjukkan bahwa meskipun model memiliki kinerja yang baik, masih terdapat kesalahan klasifikasi pada kedua kategori.

3.5 Implementasi *Naïve Bayes*

Berakar pada teorema Bayes, klasifikasi Naive Bayes adalah metode klasifikasi yang menggunakan probabilitas dan statistika untuk memprediksi kategori dari data yang belum diketahui. Dalam aplikasinya, Naive Bayes mengasumsikan bahwa setiap fitur bersifat independen, yang memungkinkan perhitungan probabilitas yang lebih sederhana dan efisien. Meskipun demikian, model ini sering kali memberikan hasil yang baik meskipun asumsi independensi tidak sepenuhnya terpenuhi dalam banyak kasus. Naive Bayes juga merupakan algoritma yang fleksibel, yang memungkinkan penggabungan dengan metode lain untuk meningkatkan kinerja dan membuat pendistribusian lebih merata dan seimbang[23][24]. Oleh karena itu, pendekatan ini sering digunakan dalam berbagai bidang, seperti pemrosesan bahasa alami dan analisis data. Di bawah ini, tabel menunjukkan prediksi dari kedua varian Naive Bayes

Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

yang digunakan dalam penelitian ini, untuk memberikan gambaran perbandingan kinerjanya[25]. Dibawah ini tabel prediksi dari kedua varian naïve bayes yang digunakan.

Tabel 2. Confussion Matrix Naïve Bayes Gaussian

		Predict Values	
		Stunting	Normal
Actual Values	Stunting	381	77
	Normal	17	358

Tabel 2 menunjukkan hasil klasifikasi yang diperoleh dari model Naïve Bayes Gaussian pada data dengan status stunting dan normal. Model ini berhasil mengklasifikasikan 381 data stunting dengan benar sebagai True Positive, sementara 77 data stunting salah diklasifikasikan sebagai normal, yang disebut sebagai False Negative. Untuk data dengan status normal, model berhasil mengidentifikasi 358 data dengan benar sebagai True Negative. Namun, model juga salah mengklasifikasikan 17 data normal sebagai stunting, yang disebut sebagai False Positive. Hasil klasifikasi ini memberikan gambaran mengenai kinerja model dalam membedakan antara kedua kategori tersebut, serta menunjukkan potensi area perbaikan pada klasifikasi data yang lebih sulit dibedakan.

Tabel 3. Confussion Matrix Naïve Bayes Bernoulli

		Predict Values	
		Stunting	Normal
Actual Values	Stunting	326	196
	Normal	72	239

Tabel 3 menggambarkan performa model Naïve Bayes Bernoulli dalam klasifikasi status stunting dan normal. Model ini berhasil mengklasifikasikan dengan benar sebanyak 326 data stunting sebagai True Positive, namun terdapat 196 data stunting yang salah diklasifikasikan sebagai normal, yang merupakan False Negative. Di sisi lain, model berhasil mengklasifikasikan 239 data normal dengan benar sebagai True Negative. Namun, terdapat 72 data normal yang salah diklasifikasikan sebagai stunting, yang disebut sebagai False Positive. Hasil ini menunjukkan bahwa meskipun model Naïve Bayes Bernoulli dapat mengidentifikasi sebagian besar data dengan benar, ada tantangan dalam membedakan data yang lebih sulit atau borderline antara kategori stunting dan normal.

3.6 Model Evaluation

Untuk mengevaluasi performa dari model yang diterapkan, dilakukan perbandingan antara algoritma Random Forest dan dua varian Naïve Bayes, yaitu Naïve Bayes Gaussian dan Naïve Bayes Bernoulli. Tabel 4 menyajikan hasil akurasi dan F1-Score dari ketiga model tersebut, yang menunjukkan seberapa baik masing-masing model dalam mengklasifikasikan data. Evaluasi ini memberikan gambaran tentang efektivitas setiap algoritma dalam menyelesaikan tugas klasifikasi, serta perbedaan performa di antara keduanya.

Tabel 4. Akurasi dan F1-Score Model

Model	Akurasi	F1-Score
Random Forest	89,62%	89,09%
Naïve Bayes Gaussian	88,72%	88,81%
Naïve Bayes Bernoulli	67,83%	69,72%

Pada Tabel 4, ditampilkan hasil perbandingan antara algoritma Random Forest dan dua varian Naïve Bayes, yaitu Naïve Bayes Gaussian dan Naïve Bayes Bernoulli. Hasil perbandingan menunjukkan bahwa algoritma Random Forest memiliki akurasi dan F1-Score tertinggi dibandingkan dengan Naïve Bayes, dengan memperoleh akurasi sebesar 89,62% dan F1-Score sebesar 89,09%. Sementara itu, jika dilihat dari hasil evaluasi Naïve Bayes Gaussian, model ini menunjukkan akurasi yang sangat baik yaitu 88,72% dengan F1-Score sebesar 88,91%, yang mengindikasikan kinerjanya yang relatif stabil. Sebaliknya, Naïve Bayes Bernoulli menunjukkan performa yang lebih rendah dengan akurasi sebesar 67,83% dan F1-Score sebesar 69,74%. Perbedaan performa ini bisa terjadi karena masing-masing varian Naïve Bayes memiliki keunggulan dan fokus penggunaannya yang berbeda. Naïve Bayes Gaussian cenderung lebih efektif pada data yang memiliki distribusi kontinu, sementara Naïve Bayes Bernoulli lebih cocok untuk data yang berbentuk biner atau kategorikal.

Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

4. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan dua algoritma, yaitu Random Forest dan Naïve Bayes, dalam mencari performa terbaik dengan nilai akurasi dan F1-Score yang maksimal, menggunakan data stunting yang diperoleh dari Dinas Kesehatan Kabupaten Bekasi. Sebelum melakukan perbandingan algoritma, dilakukan pengelolaan data untuk mengatasi ketidakseimbangan yang ada pada dataset. Data yang semula tidak seimbang kemudian diperbaiki menggunakan teknik Adasyn dan K-fold untuk memastikan kualitas dan keseimbangan data yang digunakan dalam proses pelatihan model. Setelah mengatasi ketidakseimbangan data, dilakukan perbandingan algoritma machine learning yang menghasilkan akurasi 89,62% dan F1-Score 89,09% untuk Random Forest. Penelitian ini juga membandingkan performa antara Naïve Bayes Gaussian (NBG) dan Naïve Bayes Bernoulli (NBB), dengan hasil akurasi NBG sebesar 88,72% dan F1-Score 81,81%, serta akurasi NBB sebesar 67,83% dan F1-Score 69,72%. Penulis berharap penelitian selanjutnya, khususnya dalam hal membandingkan algoritma menggunakan data yang tidak seimbang dengan fitur kelas minoritas, dapat lebih memperhatikan tahap preprocessing data. Terutama dalam cara-cara untuk menangani kelas minoritas agar data yang digunakan menjadi lebih layak dan optimal untuk proses pelatihan model.

REFERENCES

[1] H. Hatijar, "The incidence of stunting in infants and toddlers," *Jurnal Ilmiah Kesehatan Sandi Husada*, vol. 12, no. 1, pp. 224–229, 2023, doi: 10.35816/jiskh.v12i1.1019.

[2] N. D. Yanti, F. Betriana, and I. R. Kartika, "Faktor Penyebab Stunting Pada Anak: Tinjauan Literatur," *Real In Nursing Journal*, vol. 3, no. 1, pp. 1–10, 2020, doi: 10.32883/rnj.v3i1.447.

[3] D. Husnaniyah, D. Yulyanti, and R. Rudiansyah, "Hubungan tingkat pendidikan ibu dengan kejadian stunting," *The Indonesian Journal of Health Science*, vol. 12, no. 1, pp. 57–64, 2020, doi: 10.32528/ijhs.v12i1.4857.

[4] T. A. E. Permatasari, Y. Chadirin, T. S. Yuliani, and S. Koswara, "Pemberdayaan Kader Posyandu Dalam Fortikasi Pangan Organik Berbasis Pangan Lokal Sebagai Upaya Pencegahan Stunting Pada Balita," *Jurnal Pengabdian Masyarakat Teknik*, vol. 4, no. 1, pp. 1–10, 2021, doi: 10.24853/jpmt.4.1.1-10.

[5] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 257–265, Jan. 2024, doi: 10.57152/malcom.v4i1.1078.

[6] Fadellia Azzahra, N. Suarna, and Y. Arie Wijaya, "Penerapan Algoritma Random Forest Dan Cross Validation Untuk Prediksi Data Stunting," *Kopertip : Jurnal Ilmiah Manajemen Informatika dan Komputer*, vol. 8, no. 1, pp. 1–6, Feb. 2024, doi: 10.32485/kopertip.v8i1.238.

[7] M. G. Daffa and P. H. Gunawan, "Stunting Classification Analysis for Toddlers in Bojongsoang: A Data-Driven Approach," in *2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)*, IEEE, 2024, pp. 42–46. doi: 10.1109/ICoSEIT60086.2024.10497515.

[8] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis: Jurnal Ilmiah Ekonomi Dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.

[9] L. Ratnawati and D. R. Sulistyaningrum, "Penerapan random forest untuk mengukur tingkat keparahan penyakit pada daun apel," *Jurnal Sains dan Seni ITS*, vol. 8, no. 2, pp. A71–A77, 2020, doi: 10.12962/j23373520.v8i2.48517.

[10] A. A. Santika, T. H. Saragih, and M. Muliadi, "Penerapan Skala Likert pada Klasifikasi Tingkat Kepuasan Pelanggan Agen Brilink Menggunakan Random Forest," *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, vol. 11, no. 3, pp. 405–411, 2023, doi: 10.26418/justin.v11i3.62086.

[11] M. M. Mutoffar, M. Naseer, and A. Fadillah, "Klasifikasi kualitas air sumur menggunakan algoritma random forest," *Naratif: Jurnal Nasional Riset, Aplikasi dan Teknik Informatika*, vol. 4, no. 2, pp. 138–146, 2022, doi: 10.53580/naratif.v4i2.160.

[12] I. Kurniawan, D. C. P. Buani, A. Abdussomad, W. Apriliah, and R. A. Saputra, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin," *Jurnal Teknologi Informasi Dan Ilmu Komputer*, vol. 10, no. 2, pp. 421–428, 2023, doi: 10.25126/jtiik.20231026225.

[13] J. Pratama, F. Fauziah, and I. D. Sholihati, "Metode K-Nearest Neighbor Dan Naive Bayes Dalam Menentukan Status Gizi Balita," *Brahmana: Jurnal Penerapan Kecerdasan Buatan*, vol. 4, no. 2, pp. 214–221, 2023, doi: 10.30645/brahmana.v4i2.197.g196.

[14] A. F. Watratan and D. Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *Journal of Applied Computer Science and Technology*, vol. 1, no. 1, pp. 7–14, 2020, doi: 10.52158/jacost.v1i1.9.

[15] R. Ramadhani and R. Ramadhanu, "Metode Machine Learning untuk Klasifikasi Data Gizi Balita dengan Algoritma Naïve Bayes, KNN dan Decision Tree," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 15, no. 1, 2024, doi: 10.24176/simet.v15i1.10679.

[16] B. Rahman, F. Fauzi, and S. Amri, "Perbandingan Hasil Klasifikasi Data Iris menggunakan Algoritma K-Nearest Neighbor dan Random Forest: Comparison of Iris Data Classification Results using the K-Nearest Neighbor and Random Forest Algorithms," *Journal Of Data Insights*, vol. 1, no. 1, pp. 19–26, 2023, doi: 10.26714/jodi.v1i1.135.

[17] U. Ungkawa and M. A. Rafi, "Data Balancing Techniques Using the PCA-KMeans and ADASYN for Possible Stroke Disease Cases," *Jurnal Online Informatika*, vol. 9, no. 1, pp. 138–147, Jun. 2024, doi: 10.15575/join.v9i1.1293.

[18] C. G. Tekkali and K. Natarajan, "An advancement in AdaSyn for imbalanced learning: An application to fraud detection in digital transactions," *Journal of Intelligent & Fuzzy Systems*, vol. 46, pp. 11381–11396, 2024, doi: 10.3233/JIFS-236392.

[19] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers in Nanotechnology*, vol. 4, Aug. 2022, doi: 10.3389/fnano.2022.972421.

Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

[20] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042333.

[21] A. Nugroho and D. Harini, "Teknik Random Forest untuk Meningkatan Akurasi Data Tidak Seimbang," *JSITIK*, vol. 2, no. 2, 2024, doi: 10.53624/jsitik.v2i2.XX.

[22] Z. P. Agusta and Adiwijaya, "Modified balanced random forest for improving imbalanced data prediction," *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 1, pp. 58–65, Mar. 2019, doi: 10.26555/ijain.v5i1.255.

[23] Y. Yusnida Lase *et al.*, "Bulletin of Information Technology (BIT) Prediksi Dampak Pembelajaran Hybrid Learning Menggunakan Naive Bayes," vol. 4, no. 4, pp. 425–429, 2023, doi: 10.47065/bit.v3i1.

[24] N. S. Abd and D. A. Abdullah, "Diagnose of Chronic Kidney Diseases by Using Naive Bayes Algorithm," *Journal of Al-Qadisiyah for Computer Science and Mathematics*, vol. 13, no. 2, Jul. 2021, doi: 10.29304/jqcm.2021.13.2.819.

[25] I. Cholissodin *et al.*, "Development of big data app for classification based on map reduce of naive Bayes with or without web and mobile interface by RESTful API using Hadoop and spark," *Journal of Information Technology and Computer Science*, vol. 5, no. 3, pp. 302–312, 2020, doi: 10.25126/jitecs.202053233.