RABIT : Jurnal Teknologi dan Sistem Informasi Univrab
Volume 9 No. 2 | Juli 2024 : Hal : 242-251

ISSN CETAK : 2477-2062
ISSN ONLINE : 2502-891X

PENANGANAN *MISSING VALUES* DAN PREDIKSI DATA TIMBUNAN SAMPAH BERBASIS *MACHINE LEARNING*

[1] Anisa Widianti, [2]Irfan Pratama
[1,2] Program Studi Sistem Informasi, Teknologi Informasi, Universitas Mercu Buana Yogyakarta
[1,2]. Jl. Jembatan Merah 84-C, Gejayan, Yogyakarta.
E-Mail : [1] anisawidianti77@gmail.com, [2] irfanp@mercubuana-yogya.ac.id

ABSTRAK
Permasalahan peningkatan jumlah sampah seiring dengan bertambahnya jumlah penduduk dan aktivitas manusia menjadi tantangan serius dalam pengelolaan sampah di Jawa Tengah. Dalam upaya menanggulangi peningkatan sampah perlu dilakukannya prediksi sampah dimasa depan guna mengetahui tindakan yang tepat untuk mengurangi sampah tersebut. Namun dalam proses prediksi terdapat salah satu hambatan utama dalam penelitian prediksi sampah yaitu banyaknya data yang kosong. Hal itu dapat mengurangi akurasi model prediksi, sehingga diperlukannya penanganan *missing value* terlebih dahulu. Dalam menangani *missing value* penelitian ini menggunakan tiga metode yaitu *Mean Imputation, Interpolation dan KNN Imputer*.Alasan menggunakan tiga metode penanganan *missing value* untuk mengetahui dan menganalisis pengisian *missing value* mana yang cocok dan sesuai dengan dataset timbunan sampah. Setelah data terisi semua dengan penanganan *missing value* diatas, selanjutnya dapat melakukan perhitungan nilai prediksi dengan menggunakan tiga model prediksi yaitu *Random Forest, Gradient Boost*ing, dan KNN, penerapan metodologi diawali dengan pengumpulan data, *prepocessing* data, peneraan model prediksi dan yang terakhir evaluasi. Dari hasil penelitian, urutan performa terbaik pengisian *missing value* dan model prediksi berdasarkan nilai RMSE terkecil adalah *KNN Imputer* dengan model *Gradient Boosting* (RMSE = 0.188), *Mean Imputation* dengan model *Random Forest* (RMSE = 0.349), dan *Interpolation* dengan model *Gradient Boosting* (RMSE = 0.543). Pendekatan paling efektif dalam penelitian ini adalah *KNN Imputer* yang digabungkan dengan model *Gradient Boosting*, memberikan nilai RMSE terendah.

Kata Kunci: Prediksi, *Random Forest, Gradien Boosting, KNN, Interpolate, KN-Imputasi, Mean*.

ABSTRACT
The increasing waste accumulation due to population growth and human activities poses a serious challenge to waste management in Central Java. To mitigate this issue, predicting future waste levels is crucial for implementing effective waste reduction measures. However, a significant hurdle in waste prediction research is the presence of missing data, which can reduce prediction model accuracy. Therefore, this study employs three methods—Mean Imputation, Interpolation, and KNN Imputer—to handle missing values. These methods are evaluated to determine the most suitable approach for the waste dataset. Once the data is imputed, predictions are made using three models: Random Forest, Gradient Boosting, and KNN, following a methodology involving data collection, preprocessing, model application, and evaluation. The research findings indicate that the most effective combination, yielding the lowest RMSE values, involves KNN Imputer paired with Gradient Boosting (RMSE = 0.188), followed by Mean Imputation with Random Forest (RMSE = 0.349), and Interpolation with Gradient Boosting (RMSE = 0.543). This approach underscores the efficacy of KNN Imputation alongside Gradient Boosting for accurate waste prediction modeling.

Keyword: *Prediction, Random Forest, Gradient Boosting, KNN, Interpolation, KNN-Imputation, Mean*

.

PENDAHULUAN

Sampah adalah bahan atau objek yang tidak lagi dibutuhkan oleh masyarakat oleh karena itu sering dibuang [1], karena banyaknya masyarakat yang membuang sampah dan pengelolaan sampah yang kurang efisien sehingga Sampah menjadi masalah besar di Indonesia, terutama di kota-kota besar seperti yang ada di Provinsi Jawa Tengah yang masih mengalami kesulitan dalam menangani masalah ini secara efektif. Hal ini menyebabkan ketidakseimbangan lingkungan, seperti polusi air, udara, dan tanah. Berdasarkan data tahun 2019, jumlah sampah di Jawa Tengah mencapai 5,7 juta ton per tahun atau 15.671 ton per hari. Selain itu, pada tahun 2022, Jawa Tengah dinobatkan sebagai

DOI : https://doi.org/10.36341/rabit.v9i2.4789

provinsi penghasil sampah terbanyak di Indonesia, dengan jumlah mencapai 5,76 juta ton [2]..

Pertumbuhan populasi yang pesat, tingkat urbanisasi yang tinggi, serta aktivitas industri yang intensif telah menyebabkan peningkatan signifikan *volume* sampah yang dihasilkan. Manajemen sampah yang tidak efisien dapat berdampak serius pada lingkungan dan kesehatan masyarakat. Oleh karena itu, diperlukan pendekatan yang lebih maju dan terarah dalam mengelola sampah untuk meminimalkan dampak *negatif* tersebut, sehingga diperlukannya prediksi timbunan sampah di provinsi Jawa Tengah Memperkirakan tingkat timbulan sampah menjadi sangat penting karena hasil prediksi ini dapat digunakan oleh pihak berwenang untuk mengambil tindakan proaktif dalam menyelesaikan masalah sampah di masa depan. Hal ini termasuk penyediaan tempat penampungan sementara (TPS), petugas kebersihan, sarana dan prasarana pengangkutan yang memadai, anggaran biaya operasional, hingga persiapan tempat pembuangan akhir (TPA) atau lahan baru [3]. Fasilitas dan infrastruktur yang memadai akan berdampak pada efektivitas proses pengelolaan sampah yang baik [4]. Namun dataset yang tersedia terdapat data yang masih kosong atau NAN sehingga hal tersebut dapat menghambat proses prediksi timbunan sampah. Pengolahan data yang tidak lengkap merupakan tantangan umum dalam analisis data dan pembelajaran mesin. Salah satu solusi yang sering digunakan adalah imputasi nilai yang hilang atau mengisi nilai *missing value* [5].

Missing value dalam dataset dapat mengurangi jumlah data yang dapat digunakan untuk proses prediksi, sehingga dapat menurunkan akurasi hasil prediksi. Metode imputasi seperti *Mean Imputation* dan *KNN Imputer* digunakan untuk menggantikan data yang hilang dengan nilai pengganti yang dihitung menggunakan teknik statistik. *Mean Imputation* mengganti nilai hilang dengan rata-rata nilai yang ada, sementara *KNN Imputer* menggunakan algoritma *K-Nearest Neighbors* untuk memperkirakan nilai yang hilang berdasarkan kemiripan data [6]. Selain dua metode pengisian *missing value* diatas ada satu lagi metode pengisian *missing value* yaitu *interpolation*. *Interpolation* dapat digunakan untuk memperkirakan nilai yang hilang dalam suatu dataset berdasarkan nilai-nilai yang ada di sekitarnya [7].

Dengan menggunakan tiga metode pengisian *missing value* diatas diharapkan mendapatkan nilai yang hilang dengan memperhatikan data-data yang sudah ada dan bersifat spesifik dengan data-data tersebut. Tujuan penelitian ini mencari *misisng value* dengan tiga metode untuk melihat diantara ketiga *missing value* yang digunakan dapat mengetahui manakan *missing value* yang sesuai dengan dataset pada penelitian ini. Untuk menguji ketiga metode pengisian *missing value* diatas penelitian ini menggunakan tiga model prediksi secara langsung yaitu pertama *Random Forest* (RF) adalah metode yang efektif untuk memprediksi data dalam skala besar. Teknik ini menggabungkan berbagai pohon keputusan yang telah dilatih menggunakan sampel data, sehingga memungkinkan untuk mengatasi kompleksitas dan kebisingan dalam data [8]. Kedua, *Gradient boosting* adalah algoritma klasifikasi dalam pembelajaran mesin yang menggunakan ensemble dari pohon keputusan untuk memprediksi nilai tertentu [9]. Sedangkan yang terakhir, *K-Nearest Neighbor* (KNN) adalah algoritma pembelajaran mesin yang berfokus pada pembelajaran *nonparametrik* dan metode *lazy learning* [10]. Ide dasar K-NN adalah menemukan jarak

DOI : https://doi.org/10.36341/rabit.v9i2.4789 *Corresponding Author* : Anisa Widianti

terpendek antara data yang dievaluasi dengan data terdekatnya [11]. Dalam melakukan penelitian ini, tinjauan literatur dari studi sebelumnya dengan dataset yang serupa sangat penting sebagai referensi yang kuat. Studi-studi ini menjadi dasar untuk memahami metodologi, tantangan, dan hasil terkait prediksi dan pengelolaan sampah.

Penelitian yang dilakukan oleh [12] Penelitian ini menyoroti efektivitas metode *Backpropagation Neural Network* (BPNN) dalam memprediksi jumlah sampah di Kota Magelang, dengan hasil menunjukkan nilai *Mean Squared Error* (MSE) terbaik sebesar 0,00013 pada Kelurahan Potrobangsan. Dengan menggunakan parameter jaringan 30-7-1 dan 1000 epoch, penelitian ini menunjukkan bahwa BPNN mampu memberikan prediksi yang akurat. Normalisasi data dalam tahap pra-proses terbukti krusial dalam mengurangi *error* dan meningkatkan akurasi prediksi.

Penelitian yang dilakukan [13] Penelitian ini mengkaji prediksi tingkat timbulan sampah selama lima tahun mendatang di TPA Ganet, Kota Tanjungpinang, dengan hasil utama menunjukkan bahwa meskipun jumlah penduduk diproyeksikan meningkat dari 246.483 jiwa pada tahun 2023 menjadi 274.883 jiwa pada tahun 2027, jumlah timbulan sampah justru diprediksi menurun menjadi 29.283 ton/tahun pada tahun 2027. Penelitian ini penting karena prediksi yang akurat mengenai timbulan sampah dapat membantu pihak berwenang mengambil tindakan proaktif dalam menangani masalah sampah. Penelitian ini juga menemukan bahwa tidak ada kebutuhan penambahan armada pengangkut sampah pada tahun 2027, meskipun ada peningkatan jumlah penduduk. Hasil ini menekankan pentingnya peran serta masyarakat dalam mengikuti kebijakan pengelolaan sampah yang ditetapkan oleh pemerintah dan pengelola TPA Ganet untuk memastikan pengelolaan sampah yang efektif dan berkelanjutan di masa depan.

Penelitian yang dilakukan oleh [14] Penelitian ini mengevaluasi timbunan sampah di Pasar Ujungberung, Kota Bandung, dengan hasil utama menunjukkan bahwa rata-rata timbulan sampah per kios adalah 0,464 kg/hari, dengan variasi signifikan tergantung pada jenis barang yang dijual. Kios makanan jadi menghasilkan sampah tertinggi (3,16 kg/hari), sementara kios hasil bumi menghasilkan sampah terendah (0,02 kg/hari). Studi ini juga menemukan bahwa plastik, organik, dan kertas adalah jenis sampah dominan di pasar ini. Variasi timbulan sampah dipengaruhi oleh luas kios, jumlah pekerja, dan jam operasional. Meskipun Pasar Ujungberung memiliki fasilitas TPS yang memadai, lokasi strategisnya menyebabkan TPS digunakan juga oleh masyarakat umum, sehingga kapasitasnya tidak mencukupi. Data dari penelitian ini penting untuk memperbaiki perencanaan pengelolaan sampah di pasar tradisional lainnya dengan memfokuskan pada kerjasama antara pedagang dan petugas kebersihan untuk mencapai pengelolaan sampah yang lebih efektif dan efisien.

Berdasarkan penelitian sebelumnya menunjukkan pentingnya prediksi dan pengelolaan sampah yang efektif. Meskipun jumlah penduduk meningkat, pengelolaan sampah yang baik dapat mengatasi tanpa perlu penambahan armada pengangkut. Keseluruhan studi ini menggarisbawahi pentingnya metode prediksi yang akurat, sehingga dapat menghasilkan nilai prediksi yang akurat serta pentingnya peran serta masyarakat dalam pengelolaan sampah untuk mencapai lingkungan yang bersih dan berkelanjutan. Yang membedakan penelitian ini dengan penelitian prediksi timbunan sampah sebelumnya adalah penelitian ini akan

DOI : https://doi.org/10.36341/rabit.v9i2.4789 *Corresponding Author* : Anisa Widianti

melakukan perbandingan tiga penanganan *missing value* dengan menerapkan tiga model prediksi. Tujuan penelitian ini diharapkan dapat melakukan proses penanganan *missing value* dan model prediksi serta mengetahui perbandingan nilai RMSE setiap penanganan *missing value* dengan menerapkan tiga model prediksi *machine learning*, memberikan landasan yang kokoh bagi pengembangan strategi pengelolaan sampah yang lebih efektif dan berkelanjutan. Dengan pemahaman yang lebih baik tentang pola dan tren dalam pembentukan timbunan sampah, selain itu diharapkan upaya mitigasi dan pengurangan dampak *negatif* dari sampah dapat dilakukan secara efisien dan tepat sasaran dan juga dapat memberikan pengetahuan dan menjadi landasan untuk melakukan penelitian dengan dataset serupa.

METODE

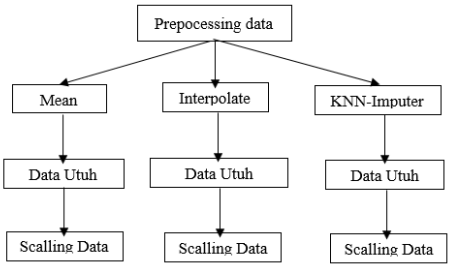
Alur penelitian untuk mendapatkan perbandingan RMSE setiap model dari tiga penanganan *missing value* secara sistematis terdiri dari proses pengumpulan data, *Prepocessing* data, Penerapan model dan yang terakhir melakuakan evaluasi yang akan digambarkan pada Gambar 1.

Gambar 1. Metodologi data yang digunakan

Pengumpulan Data

Dalam penelitian ini, data yang digunakan diperoleh melalui sumber data sekunder dengan metode studi literatur. Pengumpulan data dilakukan dengan mencari informasi pada website resmi pemerintah melalui SIPSN (Sistem Informasi Pengelolaan Sampah Nasional), yang dapat diakses secara langsung dan legal oleh siapa pun melalui link berikut: https://sipsn.menlhk.go.id/sipsn/public/data/ti Selain itu, dilakukan pencarian referensi dari berbagai jurnal atau buku dengan topik yang sama. Pengambilan dataset dilakukan pada tanggal 10 Mei 2024. Atribut yang terdapat dalam penelitian ini terdiri dari data jumlah timbunan sampah dari tahun 2019 hingga 2023 di 34 Kabupaten/Kota di Jawa Tengah yang bersifat *numerik* dan daftar Kabupaten/ Kota di Jawa Tengah

Preprocessing Data



Gambar 2. Preprocessing Data

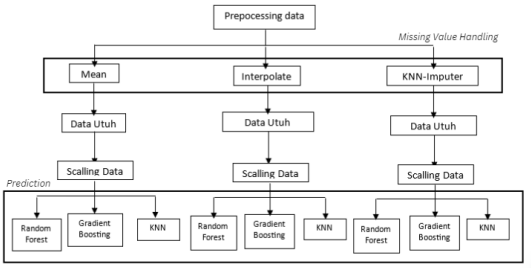
Preprocessing data adalah tahapan krusial dalam proses data mining karena tidak semua data atau atribut dalam dataset digunakan dalam analisis. Langkah ini dilakukan untuk memastikan bahwa data yang digunakan sesuai dengan kebutuhan analisis [15]. Preprocessing data bertujuan untuk mengurangi ukuran data, menemukan hubungan antar data, menormalkan data, menghapus outlier, dan mengekstrak fitur data [16]. Setelah mendapatkan dataset yang dibutuhkan untuk melakukan penelitian ini. Tahap berikutnya adalah proses *prepocessing* data, pada proses ini perlu dilakukannya pembersihan data (data *cleaning*), terdapat dua langkah penting. Pertama, dilakukan konversi kolom tahun ('2019', '2020', '2021', '2022', '2023') ke tipe data *numerik* untuk memastikan konsistensi data dalam pengolahan lebih lanjut. Langkah berikutnya adalah penanganan nilai-nilai yang hilang menggunakan tiga metode pengisian: *mean, interpolasi*, dan *KNN-imputer*. Hal ini penting untuk memastikan dataset bebas dari kekurangan yang dapat mempengaruhi analisis akhir.

Selanjutnya, dilakukan transformasi data dengan memisahkan kolom 'Kabupaten/Kota' untuk fokus pada analisis data numerik, serta melakukan *scaling* data guna normalisasi dan

DOI : https://doi.org/10.36341/rabit.v9i2.4789 *Corresponding Author* : Anisa Widianti

peningkatan kinerja model *machine learning*. Pendekatan ini secara keseluruhan bertujuan untuk memperbaiki kualitas dataset agar memberikan hasil analisis yang lebih akurat dan andal dalam konteks strategi pengelolaan sampah yang efektif dan berkelanjutan.

Penerapan Model



Gambar 3. Penerapan Model

Setelah dilakukannya *preprocessing* dataset, data tersebut dapat diproses dengan model yang diinginkan. Model yang digunakan untuk melakuakn prediksi timbunan sampah pada riset ini yaitu *Random Forest, Gradien Boosting* dan *KNN* dengan masing-masing pengisian *missing value handling.*

Evaluasi

Evaluasi merupakan tahap terakhir pada penelitian, setelah dilakukannya penerapan model sehingga mendapatkan nilai RMSE masing-masing model dengan penerapan penanganan *missing value* yang berbeda, Dapat membuat kesimpulan dari hasil RMSE tersebut dan dapat mengetahui model dan penanganan *missing value* yang paling baik dengan dataset serupa. *Root Mean Square Error* (RMSE) adalah teknik evaluasi yang umum digunakan untuk menilai kesalahan model dalam memprediksi data kuantitatif. RMSE mengukur seberapa jauh titik-titik data menyebar dari garis *regresi linear*, memberikan informasi tentang seberapa dekat data berkumpul di sekitar garis tersebut [17]. Semakin kecil nilai RMSE, semakin tinggi tingkat akurasi model dalam memprediksi data [18]. Untuk rumus RMSE dapat dilihat pada Persamaan (1) dibawah.

$$\text{RMSE} = \left(\frac{\sum(\text{Yi}-\text{Yi})}{\text{n}}\right)^{1/2} \qquad (1)$$

HASIL

Pengumpulan Data

Penelitian ini dimulai dengan mencari dataset yang dibutuhkan. Data yang dibutuhkan untuk studi ini mencakup informasi mengenai jumlah timbunan sampah di provinsi Jawa Tengah. Data ini diperoleh dari situs resmi SIPSN (Sistem Informasi Pengelolaan Sampah Nasional). Dataset yang digunakan mencakup 34 Kabupaten/Kota di Jawa Tengah, dengan rincian jumlah timbunan sampah yang dikumpulkan dari tahun 2019 hingga 2023. Dataset dapat dilihat di Tabel 1 dibawah ini

Tabel 1. Dataset Asli

No	Kabupaten/Kota	2019	2020	2021	2022	2023
1.	Kab. Cilacap	333228	343019	344409	347055	
2.	Kab Banyumas			195357	195964	197758
3.	Kab Purbalingga			184585	186120	
4.	kab. Kebumen	148386			169013	
5.	Kab. Purworejo	104874	105146	105420	105694	117432
6.	Kab. Wonosobo	115411	128352	132496	133682	127485,25
7.	Kab Magelang		248530	248800		241767
8.	Kab. Boyolali	97052	105094	106159	106781	

Preprocessing Data

Berdasarkan Tabel 1 Diatas dataset yang ingin digunakan untuk proses prediksi masih banyak data yang kosong atau bernilai NAN. Untuk melakukan prediksi tentu diperlukannya tahap *prepocessing* data untuk mengelola data

a. *Cleaning* data

Pertama, kolom tahun ('2019', '2020', '2021', '2022', '2023') akan diubah menjadi tipe data *numerik* untuk memastikan data bersifat *numerik* dan dapat mengubah bentuk data menjadi *numerik*.

Kedua yaitu menangani nilai-nilai yang hilang (*missing values*) dalam dataset. Karena dataset yang digunakan mengandung beberapa data yang kosong, proses pengisian *missing value* menjadi sangat penting. Untuk menangani masalah ini, penelitian ini akan menerapkan tiga metode pengisian *missing value* yang berbeda, yaitu *mean, interpolasi (interpolat*e), dan *KNN-imputer*. Dengan pendekatan ini, kami bertujuan untuk memperbaiki kualitas dataset sehingga dapat memberikan hasil yang lebih akurat dalam melakuakan prediksi.

1. Pengisian *missing value* dengan menggunakan *Mean*

Metode imputasi ini (Mean) menghitung rata-rata dari nilai-nilai yang ada pada setiap atribut data yang tidak hilang, kemudian mengganti

nilai-nilai yang hilang pada setiap atribut secara *independen* dari atribut lainnya. Cara ini hanya dapat diterapkan pada data *numerik*. Nilai atribut yang kosong ditandai dengan simbol NaN, kemudian NaN diganti dengan nilai rata-rata yang dihitung dengan membagi total nilai dalam satu kolom dengan jumlah nilai yang ada dalam kolom tersebut [5]. Hasil pengisian *missing value* menggunakan *mean* dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pengisian missing value menggunakan mean

No	Kabupaten/Kota	2019	2020	2021	2022	2023
1.	Kab. Cilacap	333228	343019	344409	347055	**341927,75**
2.	Kab Banyumas	**196359,67**	**196359,67**	195357	195964	197758
3.	Kab Purbalingga	**185352,5**	**185352,5**	184585	186120	**185352,5**
4.	kab. Kebumen	148386	**158699,5**	**158699,5**	169013	**158699,5**
5.	Kab. Purworejo	104874	105146	105420	105694	117432
6.	Kab. Wonosobo	115411	128352	132496	133682	127485,25
7.	Kab Magelang	**246365,99**	248530	248800	**246365,99**	241767
8.	Kab. Boyolali	97052	105094	106159	106781	**103771**

Pada tabel 2, data yang tercetak tebal menunjukkan hasil dari proses imputasi *missing value* menggunakan metode *mean*. Data yang hilang telah diisi dengan nilai rata-rata dari data yang ada pada kolom yang bersangkutan, sehingga memungkinkan analisis lebih lanjut tanpa kehilangan informasi penting.

2. Pengisian *missing value* dengan menggunakan *Interpolate*

Penanganan missing value kedua yaitu interpolasi, Interpolasi adalah sebuah teknik dalam matematika yang digunakan untuk memperkirakan nilai di antara titik-titik data yang diketahui. Dengan menggunakan interpolasi, kita dapat mengisi celah dalam data atau memperoleh perkiraan nilai di antara titik-titik yang telah diamati. Proses interpolasi melibatkan pemilihan metode interpolasi yang sesuai, persiapan data yang akurat, perhitungan nilai interpolasi, evaluasi hasil, dan penggunaan hasil interpolasi untuk berbagai tujuan. Meskipun interpolasi dapat memberikan perkiraan yang berguna, penting untuk diingat bahwa hasilnya mungkin kurang akurat jika data tidak cukup representatif. Oleh karena itu, interpolasi harus digunakan dengan hati-hati dan disesuaikan dengan konteks dan kebutuhan spesifik. Hasil pengisian *missing value* mengunakan interpolasi dapat dilihat pada tabel Tabel 3.

Tabel 3. Hasil Pengisian Missing Value Menggunakan Interpolate

No	Kabupaten/Kota	2019	2020	2021	2022	2023
1.	Kab. Cilacap	333228	343019	344409	347055	**197758**
2.	Kab Banyumas	**271614**	**283550,75**	195357	195964	197758
3.	Kab Purbalingga	**210000**	**224082,5**	184585	186120	**170982,67**
4.	kab. Kebumen	148386	**164614,25**	**145002,5**	169013	**144207**
5.	Kab. Purworejo	104874	105146	105420	105694	117432
6.	Kab. Wonosobo	115411	128352	132496	133682	**179599**
7.	Kab Magelang	**106231,5**	248530	248800	**120231,5**	241767
8.	Kab. Boyolali	97052	105094	106159	106781	**239408**

Pada Tabel 4, terdapat nilai yang tercetak tebal menunjukkan hasil dari pengisian *missing value* menggunakan metode interpolasi. Interpolasi merupakan teknik yang digunakan untuk memperkirakan nilai yang hilang berdasarkan pola atau tren dari data yang ada di sekitarnya. Dalam hal ini, nilai-nilai yang hilang telah diisi dengan menggunakan interpolasi, yang memungkinkan untuk memanfaatkan informasi yang tersedia dengan lebih baik dan menjaga kontinuitas data.

Proses ini penting untuk memastikan bahwa analisis statistik dan pemodelan yang dilakukan pada dataset ini dapat memberikan hasil yang lebih akurat dan konsisten, karena interpolasi membantu mengurangi dampak dari *missing value* dengan memberikan estimasi yang lebih realistis daripada metode imputasi yang lebih sederhana.

3. Pengisian *missing value* dengan menggunakan *KNN-Imputer*

Dengan menggunakan metode imputasi KNN (*KNN Imputation*), kita akan menghitung menggunakan data terdekat (*k-nearest neighbors*) untuk menemukan k sampel terdekat yang digunakan sebagai pengganti nilai yang hilang. Pendekatan ini memungkinkan kita untuk mengisi elemen yang hilang berdasarkan kemiripan dengan data sekitarnya. Namun, metode ini hanya dapat diterapkan pada data *numerik* [19]. Hasil pengisian *missing value* menggunakan *KNN Imputer* dapat dilihat pada Table 4.

Tabel 2. Hasil Pengisian Missing Value Menggunakan KNN Imputer

No	Kabupaten/Kota	2019	2020	2021	2022	2023
1.	Kab. Cilacap	333228	343019	344409	347055	**315156**
2.	Kab Banyumas	**142521,8**	**163913,6**	195357	195964	197758
3.	Kab Purbalingga	**146338,2**	**163913,6**	184585	186120	**182317,6**

DOI : https://doi.org/10.36341/rabit.v9i2.4789 *Corresponding Author* : Anisa Widianti

4.	kab. Kebumen	148386	**151839,8**	**167030**	169013	**161829**
5.	Kab. Purworejo	104874	105146	105420	105694	117432
6.	Kab. Wonosobo	115411	128352	132496	133682	**125398,6**
7.	Kab Magelang	**204963,8**	248530	248800	**234703**	241767
8.	Kab. Boyolali	97052	105094	106159	106781	**104426**

Pada Tabel 4, data yang tercetak tebal menunjukkan hasil dari pengisian *missing value* menggunakan metode *KNN Imputer*. *KNN Imputer*, atau *K-Nearest Neighbors Imputer*, adalah teknik yang memperkirakan nilai yang hilang berdasarkan nilai-nilai yang paling mirip dalam dataset. Dalam proses ini, *KNN Imputer* mencari sejumlah titik data terdekat (tetangga) yang memiliki nilai lengkap dan menggunakan nilai-nilai tersebut untuk menghitung estimasi dari *missing value*. Dengan menggunakan *KNN Imputer*, nilai-nilai yang hilang dalam dataset telah diisi dengan cara yang lebih akurat dibandingkan metode imputasi yang lebih sederhana seperti *mean atau median*. Metode ini mempertimbangkan hubungan dan pola dalam data, sehingga menghasilkan estimasi yang lebih realistis dan meminimalkan distorsi pada analisis selanjutnya. Proses ini sangat penting untuk memastikan kualitas data tetap tinggi, memungkinkan analisis statistik dan pemodelan yang lebih akurat. Dengan data yang lebih lengkap dan terisi secara logis, hasil yang diperoleh dari analisis dan model prediktif akan lebih dapat diandalkan dan berguna untuk pengambilan keputusan.

b. Transformation

Selanjutnya melakukan data transformasi dengan memisahkan kolom 'Kabupaten/Kota', yang bertujuan agar dapat lebih fokus pada analisis data numerik tanpa gangguan dari data kategorikal. Langkah ini memungkinkan kita untuk melakukan operasi statistik, pengisian *missing value* dan *machine learning* dengan lebih efisien

c. *Scalling* data

Setelah proses pengisian *missing value* dilakukan, langkah berikutnya adalah melakukan *preprocessing scaling data*. Proses *scaling* data ini bertujuan untuk menormalisasi data, yaitu mengubah nilai-nilai dalam dataset yang awalnya memiliki rentang besar menjadi skala antara 0 dan 1. Normalisasi ini sangat penting dalam algoritma *machine learning* karena membantu dalam mengurangi skala perbedaan antar variabel, sehingga model dapat belajar lebih efektif dan efisien.

Penerapan Model dan Evaluasi

Tahap selanjutnya setelah dilakukan proses prepocessing data dan juga data sudah siap di proses yaitu penerapan model. Penelitian ini akan menggunakan tiga model prediksi *Random Forest, Gradient Boosting dan KNN* setiap masing masing pengisian *missing value*. Hasil dari penerapan model dan *preprocessing* di atas berupa nilai RMSE dari setiap model dan visualisasi data prediksi serta data aktualnya. Dengan nilai RMSE tersebut, dapat menentukan metode penanganan *missing value* dan model mana yang lebih baik digunakan jika dataset yang dimiliki serupa dengan dataset dalam penelitian ini. Selain itu, melalui penelitian ini, akan didapatkan nilai prediksi untuk setiap tahun dengan menggunakan ketiga model prediksi yang berbeda. Berikut hasil RMSE masing masing model prediksi dapat dilihat pada Tabel 5.

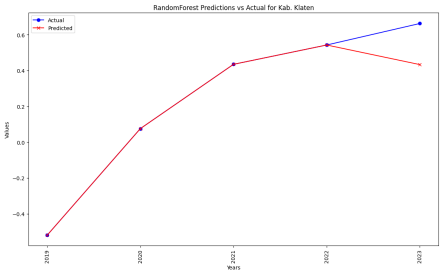
Tabel 3. Hasil Rata Rata RMSE setiap Missing valua dan Prediksi

No	Metode Missing Value	Prediksi	Rata-Rata RMSE
1.	Mean	Random Forest	0.349
		Gradient Boosting	0.353
		KNN	0.381
2.	Interpolate	Random Forest	0.607
		Gradient Boosting	0.543
		KNN	0.569
3.	KNN- Imputer	Random Forest	0.205
		Gradient Boosting	**0.188**
		KNN	0.308

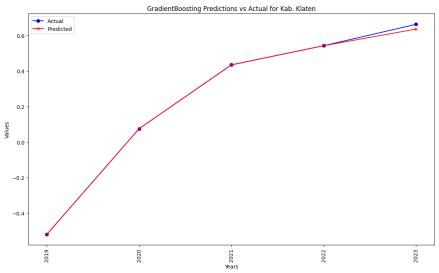
Evaluasi kinerja model prediksi terhadap data sampah di provinsi Jawa Tengah menunjukkan bahwa metode pengisian *missing value* dan model prediksi yang digunakan memiliki pengaruh signifikan terhadap hasil yang diperoleh. Dalam penelitian ini, kombinasi metode *KNN-Imputer* dengan model *Gradient Boosting* menghasilkan nilai RMSE terendah, menunjukkan bahwa kombinasi ini paling efektif dalam menangani missing value dan

DOI : https://doi.org/10.36341/rabit.v9i2.4789 *Corresponding Author* : Anisa Widianti

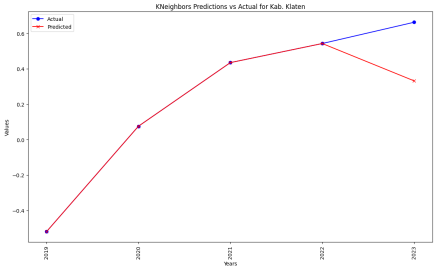
menghasilkan prediksi yang akurat. Sementara itu, *Random Forest* juga menunjukkan performa yang baik, terutama ketika menggunakan metode interpolasi dan *mean*. Di sisi lain, model *KNN* secara konsisten menghasilkan nilai RMSE tertinggi di berbagai metode pengisian *missing value*, menunjukkan bahwa model ini mungkin kurang cocok untuk dataset ini tanpa penyesuaian lebih lanjut. Oleh karena itu, pemilihan metode pengisian *missing value* yang tepat dan model prediksi yang sesuai sangat penting untuk memastikan akurasi dan keandalan hasil analisis. Berikut visualisasi dari data *actual* dan prediksi yang telah di lakukan gambar dibawah.



Gambar 4. Random Forest prediksi di Kab. Klaten



Gambar 5. Gradient Boosting prediksi Kab. Klaten



Gambar 6. KNN Prediksi di Kab Klaten

Gambar menunjukkan bahwa model *Gradient Boosting* menjadi model terbaik dalam memprediksi dataset yang digunakan untuk penelitian ini. Karena dapat dilihat perbandingan tiga visualisasi diatas nilai *actual* dan nilai prediksi yang dimiliki *Gradient Boosting* menghasilkan nilai yang tidak jauh berbeda dibandingkan dengan model yang lain.

KESIMPULAN

Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa pemilihan metode pengisian *missing value* dan model prediksi memiliki dampak signifikan terhadap hasil akhir dalam prediksi jumlah timbunan sampah di provinsi Jawa Tengah. Tiga metode imputasi yang diterapkan, yaitu *mean, interpolasi,* dan *KNN-Imputer*, menunjukkan performa yang berbeda-beda dalam kombinasi dengan tiga model prediksi yang berbeda: *Random Forest, Gradient Boosting*, dan KNN. Pertama, untuk metode imputasi menggunakan *mean*, model *Random Forest* memberikan nilai RMSE terendah sebesar 0.349, menunjukkan kemampuannya dalam menangani data yang sudah diimputasi dengan rata-rata. Namun, interpolasi menunjukkan hasil yang kurang baik, terutama pada model *Random Forest* dengan nilai RMSE 0.607, menunjukkan bahwa interpolasi mungkin tidak cocok untuk kasus ini tanpa penyesuaian lebih lanjut. Kedua, interpolasi menunjukkan hasil yang baik dengan nilai RMSE yang rendah saat menggunakan metode *Gradient Boosting*, yaitu 0.543, menunjukkan kemampuannya dalam menangani data yang diisi dengan interpolasi. Model KNN menunjukkan performa yang cukup baik dengan RMSE 0.569, tetapi masih lebih rendah daripada interpolasi. Ketiga, *KNN-Imputer* menunjukkan performa yang paling baik di antara semua metode imputasi, terutama dengan model *Gradient Boosting* yang memiliki RMSE terendah sebesar 0.188. Hasil ini menunjukkan bahwa *KNN-Imputer* efektif dalam mengisi nilai yang hilang berdasarkan data terdekat dengan dataset yang serupa dengan penelitian ini. Secara keseluruhan, model *Gradient Boosting* menonjol sebagai model terbaik dalam memprediksi jumlah timbunan sampah di Kabupaten/Kota Jawa Tengah dengan akurasi yang tinggi, terutama saat dipasangkan dengan metode *KNN-Imputer* untuk pengisian *missing value*. Kombinasi ini menunjukkan potensi besar

DOI : https://doi.org/10.36341/rabit.v9i2.4789
Corresponding Author : Anisa Widianti

untuk digunakan dalam pengambilan keputusan yang lebih baik terkait manajemen sampah di Kabupaten/Kota Jawa Tengah. Hasil penelitian ini juga menyoroti pentingnya memilih metode imputasi yang sesuai dengan karakteristik data dan model prediksi untuk mendapatkan hasil prediksi yang paling akurat.

DAFTAR PUSTAKA

[1] N. Trisnawati, Y. N. E. Putri, N. T. Rahma, E. M. Sari, and A. T. Yulinda, "Pelatihan Daur Ulang Sampah Botol Plastik Menjadi Celengan Di Desa Air Hitam Kabupaten Mukomuko," *J. Ilm. Mhs. Kuliah Kerja Nyata*, vol. 2, no. 1, pp. 153–159, 2022, doi: 10.36085/jimakukerta.v2i1.2542.

[2] F. Novitasari and W. Nurharjadmo, "Implementasi Strategi Dinas Lingkungan Hidup dalam Pengelolaan Sampah di Kabupaten Sukoharjo pada Masa Pandemi Covid-19 Febrianti Novitasari, Wahyu Nurharjadmo," *J. Mhs. Wacana Publik*, vol. 3, no. 1, pp. 104–118, 2023.

[3] Gunawansyah, R. H. Laluma, and A. Prasetya, "Prediksi Volume Dan Ritasi Pengelolaan Sampah," *J. Techno-Socio Ekon.*, vol. 15, no. 1, pp. 49–60, 2022.

[4] C. W. Wardani, "Analisa Kelayakan Fasilitas Dan Sistem Pengelolaan Tempat Pembuangan Akhir (TPA) Benowo Surabaya," *Rekayasa Tek. Sipil*, vol. 10, no. 2, pp. 1–11, 2022, [Online]. Available: https://ejournal.unesa.ac.id/index.php/rekayasa-teknik-sipil/article/view/48990

[5] F. Yulian Pamuji, Ahmad Rofiqul Muslikh, Rizza Muhammad Arief, and Delviana Muti, "Komparasi Metode Mean dan KNN Imputation dalam Mengatasi Missing Value pada Dataset Kecil," *J. Inform. Polinema*, vol. 10, no. 2, pp. 257–264, 2024, doi: 10.33795/jip.v10i2.5031.

[6] M. I. Ananda, "Multivariate Forecasting Harga Daging Ayam dan Sapi Melibatkan Faktor Cuaca, Ekonomi, dan Kesehatan Menggunakan GRU Multivariate Forecasting of Chicken and Beef Prices Involving Weather, Economic, and Health Factors Using GRU," *J. Ilmu Komput. dan Argi-Informatika*, vol. 10, no. 1, pp. 111–120, 2023, [Online]. Available: https://jurnal.ipb.ac.id/index.php/jika

[7] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest," *J. Komput. Terap.*, vol. 7, no. 1, pp. 24–32, 2021, doi: 10.35143/jkt.v7i1.4618.

[8] R. Risanti, "Analisis Model Prediksi Cuaca Menggunakan Support Vector Machine, Gradient Boosting, Random Forest, Dan Decision Tree," vol. XII, pp. 119–128, 2024, doi: 10.21009/03.1201.fa18.

[9] I. W. S, J. Triloka, and R. E. Badri, "Prediksi Potensi Penjualan Leopard Gecko Pada Snowy Gecko Farm Menggunakan Kajian Algoritma K-NN dan Naïve Bayes," *Semin. Nas. Has. Penelit. dan Pengabdi. Masy. 2023*, vol. 1, no. Leopard Gecko, pp. 208–217, 2023.

[10] D. Safitri, S. S. Hilabi, and F. Nurapriani, "Analisis Penggunaan Algoritma Klasifikasi Dalam Prediksi Kelulusan Menggunakan Orange Data Mining," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 8, no. 1, pp. 75–81, 2023, doi: 10.36341/rabit.v8i1.3009.

[11] M. R. A. Prasetya, A. M. Priyatno, and Nurhaeni, "Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining," *J. Inf. dan Teknol.*, vol. 5, no. 2, pp. 52–62, 2023, doi: 10.37034/jidt.v5i2.324.

[12] M. Shofwan Khamid *et al.*, "Prediksi Jumlah Sampah Kelurahan Menggunakan Neural Network Backpropagation," *J. Inf. Syst. Res.*, vol. 5, no. 2, pp. 713–721, 2024, doi: 10.47065/josh.v5i2.4825.

[13] V. A. Simbolon, Tarisa, and H. Horiza, "Prediksi Tingkat Timbulan Sampah 5 Tahun Mendatang (2023-2027) di TPA Ganet Kota Tanjungpinang," *Sulolipu*

Media Komun. Sivitas Akad. dan Masy., vol. 23, no. 2, pp. 303–310, 2023, doi: 10.32382/sulo.v23i2.105.

[14] M. Chaerul and T. P. Dewi, "Al-Ard: Jurnal Teknik Lingkungan Al-Ard: Jurnal Teknik Lingkungan Analisis Timbulan Sampah Pasar Tradisional (Studi Kasus: Pasar Ujungberung, Kota Bandung)," *Al-Ard J. Tek. Lingkung.*, vol. 5, no. 2, pp. 98–106, 2020, [Online]. Available: http://jurnalsaintek.uinsby.ac.id/index.php/alard/index

[15] R. R. Rerung, "Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk," *J. Teknol. Rekayasa*, vol. 3, no. 1, p. 89, 2018, doi: 10.31544/jtera.v3.i1.2018.89-98.

[16] N. Sunanto and G. Falah, "Penerapan Algoritma C4.5 Untuk Membuat Model Prediksi Pasien Yang Mengidap Penyakit Diabetes," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 7, no. 2, pp. 208–216, 2022, doi: 10.36341/rabit.v7i2.2435.

[17] A. S. B. Karno, "Prediksi Data Time Series Saham Bank BRI Dengan Mesin Belajar LSTM (Long ShortTerm Memory)," *J. Inform. Inf. Secur.*, vol. 1, no. 1, pp. 1–8, 2020, doi: 10.31599/jiforty.v1i1.133.

[18] A. A. A. Purnamaswari, I. K. G. D. Putra, and I. M. S. Putra, "Komparasi Metode Neural Network Backpropagation dan Support Vector Machines dalam Prediksi Volume Sampah TPA Suwung," *JITTER J. Ilm. Teknol. dan Komput.*, vol. 3, no. 1, pp. 853–861, 2022, [Online]. Available: https://ojs.unud.ac.id/index.php/jitter/article/view/83024/43066

[19] X. Xu, W. Chen, and Y. Sun, "Over-sampling algorithm for imbalanced data classification," *J. Syst. Eng. Electron.*, vol. 30, no. 6, pp. 1182–1191, 2019, doi: 10.21629/JSEE.2019.06.12.