Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)

journal homepage: http://journal.lembagakita.org/index.php/jtik

Implementasi Metode Imputasi *Mean* dan *Single Center Imputation Chained Equation* (SICE) Terhadap Hasil Prediksi *Linear Regression* pada Data Numerik

Mario Rangga Baihaqi [1*], Tesa Nur Padilah [2], Mohamad Jajuli [3]

[1,2,3] Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang, Kabupaten Karawang, Provinsi Jawa Barat, Indonesia.*

a r t i c l e i n f o

Article history:
Received 5 May 2023
Received in revised form
10 July 2023
Accepted 12 September
2023
Available *online* October 2023

DOI:
https://doi.org/10.35870/jtik.v7i4.1169

Keywords:
Linear Regression; Mean;
SICE; RMSE.

Kata Kunci:
Regresi Linear; Mean; SICE;
RMSE.

a b s t r a c t

Data and information play an important role in all aspects of science, so data must be processed well through the process of data excavation or data mining. The excavation of patterns from data can be done using machine learning algorithms such as linear regression. However, in the process of extracting information from data, it can be less effective if there is a loss of value in a data. The purpose of this research is to implement the mean imputation and single center imputation chained equation (SICE) techniques against the linear regression algorithm. The data used in this research is numerical data. The root mean squared error (RMSE) value shows that the implementation of linear regression algorithm using the mean imputation technique results in better performance compared to the SICE imputation technique.

a b s t r a k

Data dan informasi memainkan peran penting dalam segala aspek ke ilmuan, sehingga data harus diolah dengan baik melalui proses penggalian data atau data mining. Penggalian pola dari data dapat dilakukan menggunakan Algoritma pembelajaran mesin seperti linear regression. Namun pada proses penggalian informasi dari data dapat menjadi kurang efektif jika terdapat kehilangan nilai pada suatu data. Tujuan penelitian ini adalah melakukan implementasi teknik imputasi mean dan single center imputation chained equation (SICE) terhadap Algoritma linear regression. Data yang digunakan pada penelitian ini merupakan data numerik. Nilai root mean squared error (RMSE) menunjukkan bahwa implementasi Algoritma linear regression menggunakan teknik imputasi mean menghasilkan performa yang lebih baik dibandingkan dengan teknik imputasi SICE.

* Author. Email: 1910631170094@student.unsika.ac.id [1*], tesa.nurpadilah@staff.unsika.ac.id [2], mohamad.jajuli@unsika.ac.id [3].

1. Latar Belakang

Data atau informasi memainkan peran penting dalam segala aspek salah satunya dalam pembangunan sosial-ekonomi dari tingkat individu sampai tingkat negara. Data yang relevan diperlukan untuk menjawab persoalan dalam segala bidang keilmuan sehingga data menjadi bagian penting dari program penelitian dan pengembangan, oleh karena itu data harus digali dan diolah dengan baik, penggalian data disebut dengan data mining [1]. Data *mining* merupakan proses penggalian informasi yang berguna dari kumpulan data serta mengekstraksinya agar menjadi informasi baru untuk pengambilan keputusan [2]. Pengolahan data diperlukan untuk ekstrasi informasi dari kumpulan data, pengolahan data disebut dengan data *preprocessing*. Data *preprocessing* dalam pembelajaran mesin dapat meningkatkan akurasi terutama dalam klasifikasi [3]. Beberapa informasi hilang dalam pemrosesan data, kehilangan informasi dalam sekumpulan data disebut dengan *missing values* [4].

Algoritma pembelajaran mesin meliputi *classifcation analysis*, *regression*, *data clustering*, *feature engineering*, *dimensionality reduction*, dan *reinforcement learning* secara efektif membangun sistem yang berbasis data [5]. Berdasarkan data dari Google Trends (2022) selama dua belas bulan terakhir, Algoritma *linear regression* memiliki jumlah ketertarikan terbanyak dengan nilai rata-rata sebesar 57, dibandingkan *decision tree* dengan nilai rata-rata sebesar 12 dan *k-nearst neighbors* dengan nilai rata-rata 13. Algoritma pembelajaran mesin umumnya tidak cukup kuat untuk memprediksi secara akurat dengan adanya *missing value* pada *dataset*. Data yang hilang menimbulkan elemen ambiguitas saat menganalisis data dan dapat memengaruhi prediksi secara statistik dan mengakibatkan kurangnya akurasi yang menyebabkan kesimpulan yang menyesatkan [6]. Terdapat tiga kategori untuk *missing value*: *missing at random* (MAR) merupakan *missing value* yang hanya dipengaruhi oleh variabel lengkap (diamati) dan bukan oleh karakteristik data yang hilang itu sendiri; *missing completely at random* (MCAR) merupakan *missing value* yang sama sekali tidak ada hubungan antara data yang hilang dan nilai lainnya dalam kumpulan data dan *missing data not at random* (MNAR) merupakan *missing value* dengan nilai-nilai yang hilang secara sistematis memiliki hubungan antara data yang hilang dengan data yang lain di dalam *dataset* [7].

Teknik imputasi merupakan salah satu cara yang dapat digunakan untuk menangani *missing value*. Teknik imputasi adalah metode penanganan *missing value* berdasarkan informasi yang tersedia pada *dataset* yang bertujuan memprediksi nilai yang valid untuk dimasukkan ke dalam *dataset* sebagai pengganti nilai yang hilang [9]. Pada penelitian yang membahas tentang teknik imputasi telah dilakukan oleh Jadhav, Pramod & Ramanathan (2019) tentang perbandingan performa metode imputasi untuk *numeric data*. Penelitian tersebut menggunakan delapan teknik imputasi, yaitu *mean imputation*, *median imputation*, *kNN imputation*, *predictive mean matching* (PMM), *Bayesian Linear Regression* (norm), *Linear Regression*, *non-Bayesian* (norm.nob), dan *Random Sample methods*. Penelitian dilakukan menggunakan *dataset* dari UCI *machine learning repository* dengan menggunakan metode imputasi *Normalized* RMSE (NRMSE) sebagai pembanding performa delapan teknik imputasi yang telah disebutkan. Hasil dari penelitian tersebut menunjukkan bahwa metode imputasi kNN yang memiliki performa paling baik dengan nilai NRMSE terkecil [2].

Hasil evaluasi teknik imputasi *multi variate* dengan *single variate* belum tentu akan menghasilkan sebuah hasil evaluasi yang sama jika menerapkan teknik imputasi ke dalam model prediksi. Pada penelitian Jadhav, Pramod & Ramanathan (2019) mengenai perbadingan teknik imputasi terhadap data numeric dengan judul *Comparison of Performance of Data Imputation Methods for Numeric Dataset* didapatkan hasil bahwa teknik imputasi *multi variate* menghasilkan performa yang lebih baik jika dibandingkan dengan teknik imputasi *single variate* penelitian menunjukkan bahwa metode teknik imputasi *kNN* memiliki performa yang lebih baik bila digunakan untuk data numerik dibandingkan teknik imputasi *mean* [2]. Pada penelitian teknik imputasi *SICE* yang dilakukan oleh Khan & Hoque (2020) menunjukkan bahwa hasil imputasi menghasilkan performa yang terbaik dibanding kan dengan teknik imputasi lain seperti *kNN, PMM* dan *BLR* dengan hasil *Root Mean Squared Error* (*RMSE*) sebesar 19,47 dibandingkan imputasi lain dengan nilai *RMSE* di atas 20 [10].

Pada penelitian Jajuli & Komarudin (2017) mengenai implementasi teknik imputasi *multi variate* dan *single variate* terhadap Algoritma *k-means* mendapatkan hasil bahwa prediksi *k-means* menggunakan teknik imputasi *single variate* memiliki performa lebih baik dibandingkan prediksi *k-means* dengan teknik imputasi *multi variate* yang ditunjukkan dengan nilai *Cubic Clustering Criterion* dan *Pseudo F* yang lebih tinggi untuk teknik imputasi *Mean* dibandingkan dengan teknik imputasi *Expectation Maximisation* [8]. Pada penelitian Ilham (2020) dengan judul *hybrid* metode *boostrap* dan teknik imputasi pada metode C4-5 untuk prediksi penyakit ginjal kronis. Teknik imputasi *multi variate* seperti *kNN* menunjukkan hasil yang lebih optimum dibandingkan dengan teknik imputasi *single variate* seperti modus terhadap pengklasifikasian Algoritma C4.5 [11].

Berdasarkan hasil dari penelitian sebelumnya, penelitian ini akan mencoba melakukan implementasi metode imputasi SICE dengan Algoritma *classification and regression tree* (CART) dan *mean* pada *data numeric* dengan *missing value* yang bersifat MCAR, dan selanjutnya akan dilakukan prediksi *regresi* menggunakan Algoritma *linear regression* pada data tersebut. Hasilnya, dari prediksi regresi yang telah dilakukan dengan imputasi *mean* dan prediksi dengan teknik imputasi SICE akan dievaluasi dengan metode *root mean squared error* (RMSE) dan kesalahan relatif untuk mengetahui apakah teknik imputasi *single variate* atau imputasi *multi variate* yang optimum untuk diterpakan pada kasus regresi.

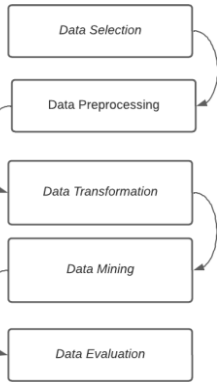
2. Objek dan Metode Penelitian

Objek Penelitian

Objek yang diteliti menggunakan dataset *Red Wine Quality dataset* yang berisikan tenang variasi anggur merah dan anggur putih dari negara Portugis, terdapat sebelas *independent variable* pada *dataset* ini seperti *fixed acidity, volatile acidity, citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol* dengan *dependent variable* menentukan kualitas dari suatu *wine*.

Metodologi Penelitian

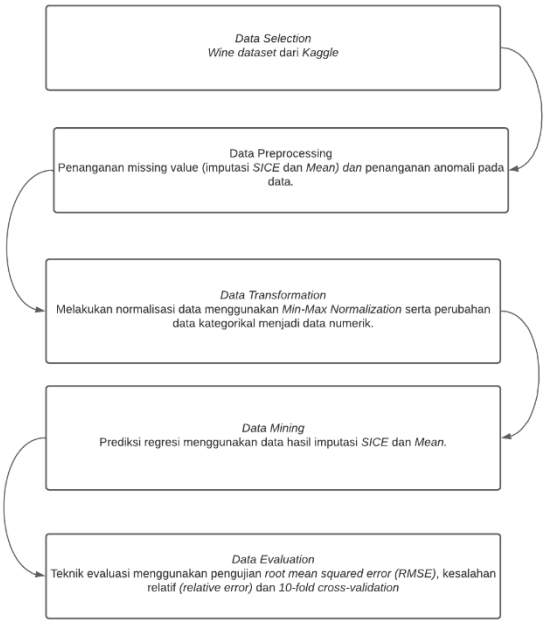
Metodologi penelitian pada penelitian ini menggunakan *Knowledge Discovery in Database* (*KDD*) dengan langkah-langkah *data selection*, *data preprocessing*, *data transformation*, *data mining*, dan *data interpretation* (*evaluation*). Proses *KDD* meliputi pemilihan data, penanganan data kotor dan memiliki kekurangan, pengubahan bentuk data, penggalian pengetahuan dari data, dan melakukan pengujian atau evaluasi terhadap data [2].



Gambar 1. Tahapan Metodologi Penelitian

Rancangan Penelitian

Pada penelitian ini penulis membuat alur rancangan penelitian seperti gambar di bawah.



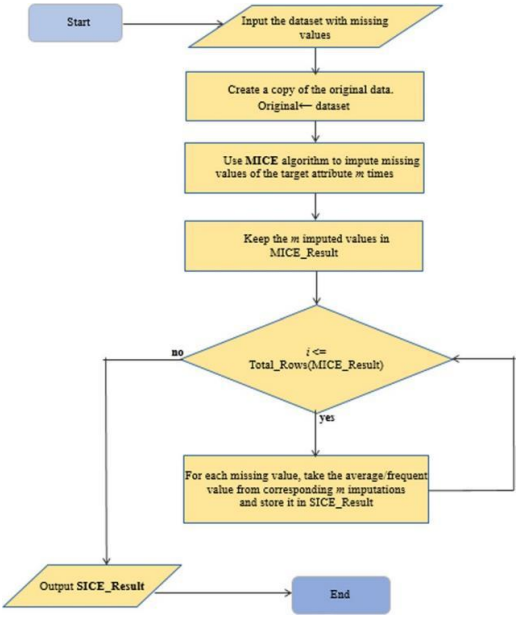
Gambar 2. Rancangan Penelitian

Data Selection

Data Red Wine Quality diunduh melalui *website* Kaggle.

Data Preprocessing

Algoritma SICE menunjukkan performa yang baik pada penerapan teknik imputasi tersebut pada data dengan *missing value* 10% dan data dengan tingkat *missing value* 20%-30% dapat meningkatkan *bias* pada data sehingga perlu dilakukan penanganan terhadap *missing value* pada data [10][8]. Penelitian ini melakukan pembangkitan *missing value* sebesar 10% dan 20%. Penanganan pada *missing value* dilakukan menggunakan teknik imputasi *mean* dan SICE, penanganan anomali meliputi penghapusan duplikasi data dan pembuangan *outliers* menggunakan *z-score*. Penanganan *missing value* menggunakan teknik imputasi *mean* dilakukan dengan cara menjumlahkan total data pada kolom dan membaginya dengan banyak data pada kolom [2]. Pada teknik imputasi *SICE*, *missing value* diatasi dengan menggunakan alur kerja di bawah:



Gambar 3. *Flowchart* Imputasi SICE

Data Transformation

Min-max normalization digunakan untuk menormalisasikan data, hasil dari normalisasi ini akan berupa perubahan nilai pada data menjadi nol hingga satu dengan representasi nilai sebenarnya. Penggunaan *min-max normalization* efektif menghasilkan performa yang lebih baik untuk kasus regresi terhadap data *wine quality* dibandingkan dengan metode lain seperti *statistical column normalization* dan *decimal scaling normalization* [15]. Pada proses ini juga dilakukan perubahan tipe data selain numerik menjadi data numerik sehingga data dapat diproses menggunakan Algoritma *linear regression*.

Data Mining

Tahap ini dilakukan penerapan model prediksi linear regression pada data yang telah dilakukan transformasi dan telah dilakukan imputasi SICE dan *mean*.

Data Evaluation

Model di evaluasi menggunakan pembagian *data* splitting. Dengan menggunakan data splitting dapat dilakukan evaluasi mengenai performa model terhadap hasil dari pelatihan model dan uji coba model. Rasio pembagian data yang sering digunakan merupakan 80:20, yang berarti 80% dari data digunakan untuk melatih model dan 20% dari data digunakan untuk pengujian model. Penerapan pembagian 80:20 dikenal dengan *pareto principle* tetapi aturan ini hanya aturan yang umum digunakan untuk praktisi pengembangan pembelajaran mesin [12]. Pengukuran model evaluasi menggunakan *data splitting* memiliki kemungkinan menghasilkan *output* dengan *bias* yang cukup tinggi, maka untuk mendukung pengukuran performa evaluasi model maka digunakan *k-fold cross-validation* popularitas *cross validation* tidak terlepas dari fakta bahwa secara konseptual memiliki improvisasi dibanding *data splitting*, nilai k sebesar 10 optimum dengan meminimalisir *bias* dan bahkan mendekati tidak *bias* [13][14]. Pada penelitian Chai & Draxler (2014) menunjukkan bahwa *RMSE* lebih tepat untuk merepresentasikan performa model dibandingkan *MAE*. Selain itu, RMSE memenuhi persyaratan ketidaksetaraan segitiga untuk metrik jarak [16]. Dengan demikian, evaluasi model *linear regression* pada penelitian ini akan menggunakan metode *data splitting* serta 10-*fold cross-validation* dengan teknik evaluasi *RMSE* dan kesalahan relatif untuk mendukung akurasi evaluasi.

3. Hasil dan Pembahasan

Hasil dan pembahasan pada penelitian ini dijelaskan pada poin-poin berikut:

Data Selection

Data yang digunakan diunduh melalui *website* Kaggle. Data berisikan kualitas dari *wine* dengan dua variasi anggur yaitu anggur merah dan anggur putih. Pada data tersebut berisikan *independent variable* seperti *fixed*

acidity, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates* dan *alcohol*, dengan *dependents variabel quality*. Data dideskripsikan pada tabel di bawah:

Tabel 1. Deskripsi Data Awal

No	Nama Kolom	Jumlah Data	Tipe Data
1	*Fixed Acidity*	1599 *non-null*	*Float64*
2	*Volatile Acidity*	1599 *non-null*	*Float64*
3	*Citric Acid*	1599 *non-null*	*Float64*
4	*Residual Sugar*	1599 *non-null*	*Float64*
5	*Chlorides*	1599 *non-null*	*Float64*
6	*Free Sulfur Dioxide*	1599 *non-null*	*Float64*
7	*Total Sulfur Dioxide*	1599 *non-null*	*Float64*
8	*Density*	1599 *non-null*	*Float64*
9	*pH*	1599 *non-null*	*Float64*
10	*Sulphates*	1599 *non-null*	*Float64*
11	*Alcohol*	1599 *non-null*	*Float64*
12	*Quality*	1599 *non-null*	*Int64*

Deskripsi data dengan pembangkitan *missing value* 10% dapat dilihat pada tabel di bawah:

Tabel 2. Deskripsi Data dengan *Missing Value* 10%

No	Nama Kolom	Jumlah Data	Tipe Data
1	*Fixed Acidity*	1417 *non-null*	*Float64*
2	*Volatile Acidity*	1410 *non-null*	*Float64*
3	*Citric Acid*	1399 *non-null*	*Float 64*
4	*Residual Sugar*	1412 *non-null*	*Float64*
5	*Chlorides*	1390 *non-null*	*Float64*
6	*Free Sulfur Dioxide*	1429 *non-null*	*Float64*
7	*Total Sulfur Dioxide*	1404 *non-null*	*Float64*
8	*Density*	1413 *non-null*	*Float64*
9	*pH*	1408 *non-null*	*Float64*
10	*Sulphates*	1418 *non-null*	*Float64*
11	*Alcohol*	1410 *non-null*	*Float64*
12	*Quality*	1599 *non-null*	*Int64*

Deskripsi data dengan tingkat *missing value* 20% dapat dilihat pada tabel di bawah:

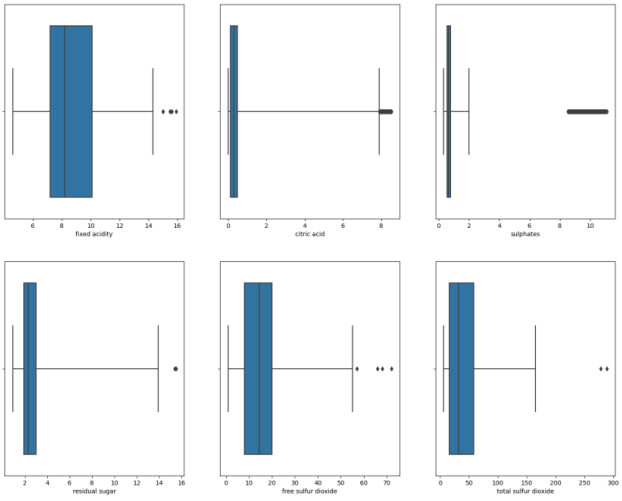
Tabel 3. Deskripsi Data dengan *Missing Value* 20%

No	Nama Kolom	Jumlah Data	Tipe Data
1	*Fixed Acidity*	1228 *non-null*	*Float64*
2	*Volatile Acidity*	1228 *non-null*	*Float64*
3	*Citric Acid*	1214 *non-null*	*Float 64*
4	*Residual Sugar*	1233 *non-null*	*Float64*
5	*Chlorides*	1172 *non-null*	*Float64*
6	*Free Sulfur Dioxide*	1260 *non-null*	*Float64*
7	*Total Sulfur Dioxide*	1191 *non-null*	*Float64*
8	*Density*	1231 *non-null*	*Float64*
9	*pH*	1223 *non-null*	*Float64*
10	*Sulphates*	1226 *non-null*	*Float64*
11	*Alcohol*	1225 *non-null*	*Float64*
12	*Quality*	1599 *non-null*	*Int64*

Data Preprocessing

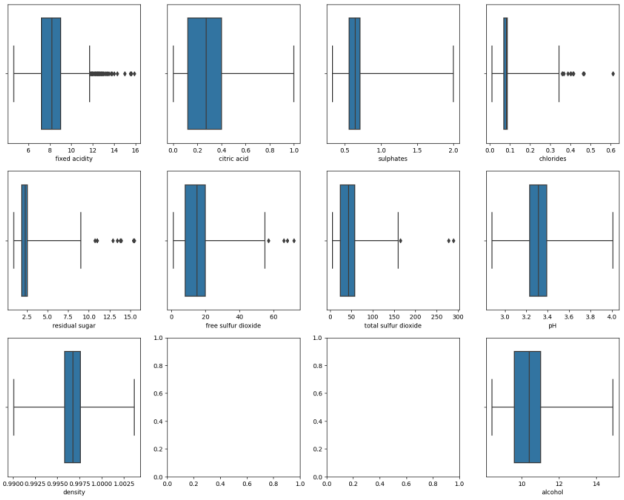
Imputaasi data dilakukan terhadap data yang terdapat *missing value*. Proses imputasi dilakukan menggunakan teknik imputasi SICE dan *mean* yang dilakukan dengan bahasa pemrograman Python. Pada imputasi SICE algoritma CART diimplementasikan menggunakan *library Scikit-Learn* yang terdapat di dalam *module Tree*. Ketika algoritma CART sudah dilakukan maka hasil dari imputasi CART disimpan ke dalam *array* yang kemudian akan dilakukan perhitungan *mean* terhadap *array* tersebut, hasil dari perhitungan *mean* di dalam *array* akan imputasi ke dalam atribut yang terdapat *missing value*, hasil dari imputasi tersebut merupakan

hasil dari teknik imputasi SICE. Data diimputasi dengan menggunakan nilai rata-rata (*mean*) terhadap atribut yang memiliki *missing value* menggunakan *library Scikit-Learn* dengan nama *module SimpleImputer*. Data yang telah diimputasi, kemudian dilanjutkan ke tahap penanganan anomali (*outlier*). Metode yang digunakan untuk menangani anomali menggunakan *z-score*. Batasan kurva distribusi normal adalah 3 dan -3, yang mengartikan jika terdapat data dengan nilai *z-score* lebih dari kurva yang telah ditentukan maka data tersebut dianggap *outlier* dan akan dilakukan pembuangan terhadap data tersebut. Visualisasi kolom yang memiliki *outlier* dapat dilihat pada *boxplot* di bawah:



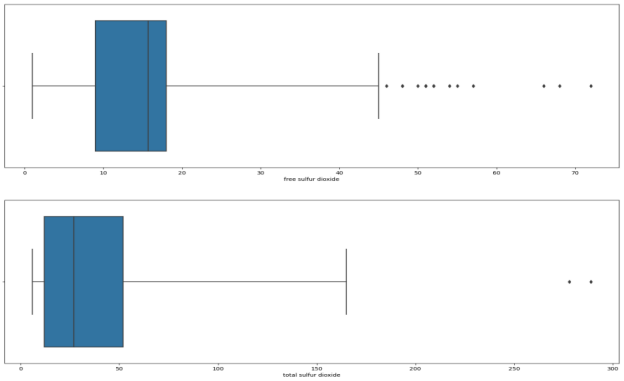
Gambar 4. Data *Outlier Missing Value* 10% SICE

Pada Gambar 4 *outlier* terdapat pada kolom *fixed acidity, citric acid, sulphates, residual sugar, free sulfur dioxide* dan *total sulfur dioxide*. Kolom *fixed acidity* memiliki *outlier* pada nilai lebih dari 14, *citric acid* memiliki *outlier* pada nilai lebih dari 8, *sulphates* memiliki *outlier* pada nilai lebih dari 2, *residual sugar* memiliki *outlier* pada nilai lebih dari 14, *free sulfur dioxide* memiliki *outlier* pada nilai lebih dari 55 dan *total sulfur dioxide* memiliki *outlier* pada nilai lebih dari 155.



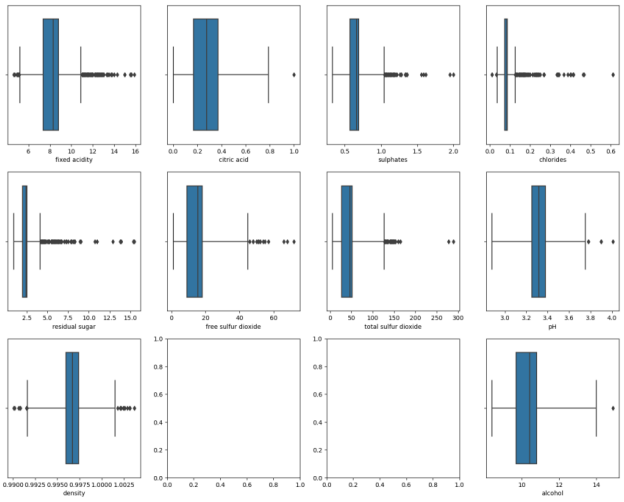
Gambar 5. Data *Outlier Missing Value* 10% *Mean*

Pada Gambar 5 menunjukkan persebaran data *outlier* pada data dengan *missing value* 10%. Kolom *fixed acidity* memiliki *outlier* dengan nilai lebih besar dari 12, kolom *chlorides* memiliki *outlier* dengan nilai lebih besar dari 0.35, kolom *residual sugar* memiliki *outlier* dengan nilai lebih besar dari 11, kolom *free sulfur dioxide* memiliki *outlier* dengan nilai lebih besar dari 50, kolom *total sulfur dioxide* memiliki *outlier* dengan nilai lebih besar dari 150.



Gambar 6. Data *Outlier Missing Value* 20% SICE

Pada Gambar 6 data dengan *outlier* berada pada kolom *free sulfur dioxide* dengan *outlier* berada pada nilai lebih dari 55 dan kolom *total sulfur dioxide* dengan nilai *outlier* lebih dari 155.



Gambar 7. Data *Outlier Missing Value* 20% *Mean*

Pada Gambar 7 data *outlier* berada pada kolom *fixed acidity* dengan nilai lebih dari 11, *citric acid* dengan nilai lebih dari 0.8, *sulphates* dengan nilai lebih dari 1, *chlorides* dengan nilai lebih dari 0.15, *residual sugar* dengan nilai lebih dari 3.15, *free sulfur dioxide* dengan nilai lebih dari 42, *total sulfur dioxide* dengan nilai lebih dari 125, pH dengan nilai lebih dari 3.8, *density* dengan nilai kurang dari 0.99 dan nilai lebih dari 1.0025 serta kolom *alcohol* dengan nilai lebih dari 14.

Data Transformation

Tahap ini merupakan proses penyeragaman data menggunakan *min-max normalization* dengan rentang 0-1. Tujuan dari hal tersebut agar data memiliki rentang nilai yang sama. Nomalisasi data dilakukan menggunakan library Scikit-Learn dengan nama *module Min-Max Normalization*. Normalisasi data ini dilakukan terhadap data dengan *missing value* 10% dan data dengan *missing value* 20%.

Data Mining

Pada tahap ini dilakukan percobaan memprediksi nilai menggunakan Algoritma *linear regression* pada bahasa pemrograman Python dengan *library Scikit-Learn*. Proses prediksi *linear regression* menggunakan Python diawali dengan melakukan *import library* untuk dapat mengakses *framework linear regression* yang berada di dalam *library Scikit-Learn*. Setelah proses *import* selesai maka dilakukan *model fitting* terhadap *dataset training* agar *model linear regression* dapat melakukan prediksi dengan akurat. Model *linear regression* digunakan pada data dengan *missing value* 10% dan 20% lalu dilakukan evaluasi menggunakan RMSE untuk mengetahui tingkat performa dari model *linear regression* yang telah dibuat. Pada model *linear regression* dengan teknik imputasi SICE dengan *missing value* sebesar 10% memiliki nilai *y-intercept* sebesar 5.58, sedangkan model *linear regression* dengan teknik imputasi mean dengan *missing value* 10% memiliki nilai *y-intercept* sebesar 5.37. Nilai koefisien untuk tiap atribut (*dependent variable*) dengan *missing value* 10% dapat dilihat pada Tabel 4 dan Tabel 5. Model *linear regeression* dengan *missing value* 20% dengan teknik imputasi SICE memiliki nilai *y-intercept* sebesar 5.23, untuk model *linear regression* dengan teknik imputasi *mean* memiliki nilai *y-intercept* sebesar 5.38. Nilai koefisien tiap atribut pada data dengan *missing value* 20% dapat dilihat pada Tabel 6 dan Tabel 7.

Tabel 4. Nilai Koefisien *Missing Value* 10% SICE

Nama Kolom	Nilai Koefisien
Fixed Acidity	0.07
Volatile Acidity	-0.06
Citric Acid	0.58
Residual Sugar	0.10
Chlorides	-0.07
Free Sulfur Dioxide	0.07
Total Sulfur Dioxide	-0.19
Density	-1.53
pH	-0.39
Sulphates	1.62
Alcohol	0.40

Nilai koefisien untuk data dengan imputasi *mean* dengan *missing value* sebesar 10% dapat dilihat pada tabel di bawah.

Tabel 5. Nilai Koefisien *Missing Value* 10% *Mean*

Nama Kolom	Nilai Koefisien
Fixed Acidity	0.19
Volatile Acidity	-0.87
Citric Acid	-0.12
Residual Sugar	0.05
Chlorides	-0.08
Free Sulfur Dioxide	0.24
Total Sulfur Dioxide	-0.29
Density	-0.31
pH	-0.41
Sulphates	1.26
Alcohol	1.10

Nilai koefisien untuk data dengan teknik imputasi SICE dengan *missing value* sebesar 20% ditunjukkan pada tabel di bawah.

Tabel 6. Nilai Koefisien *Missing Value* 20% SICE

Nama Kolom	Nilai Koefisien
Predicted Value	-
Fixed Acidity	0.49
Volatile Acidity	0.10
Citric Acid	0.13
Residual Sugar	-0.01
Chlorides	0.01
Free Sulfur Dioxide	-0.02
Total Sulfur Dioxide	-0.14
Density	-0.03
pH	-0.01
Sulphates	0.06
Alcohol	0.54

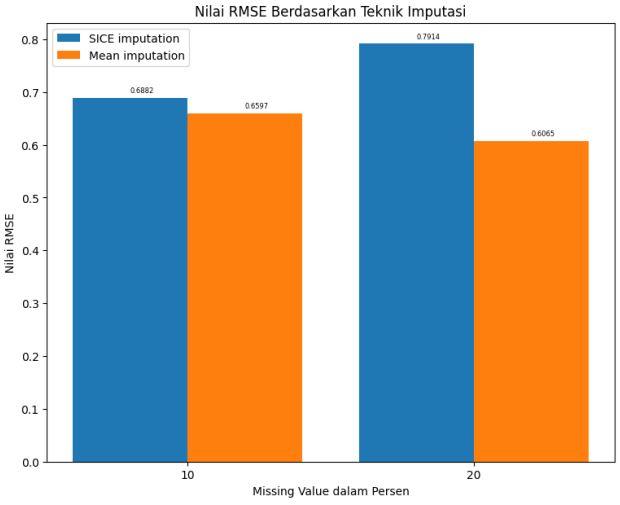
Nilai koefisien untuk data dengan teknik imputasi *mean* dengan *missing value* sebesar 20% ditunjukkan pada tabel di bawah.

Tabel 7. Nilai Koefisien *Missing Value* 20% *Mean*

Nama Kolom	Nilai Koefisien
Fixed Acidity	0.20
Volatile Acidity	-0.70
Citric Acid	0.05
Residual Sugar	0.11
Chlorides	-0.39
Free Sulfur Dioxide	0.14
Total Sulfur Dioxide	-0.25
Density	-0.39
pH	-0.23
Sulphates	1.13
Alcohol	1.09

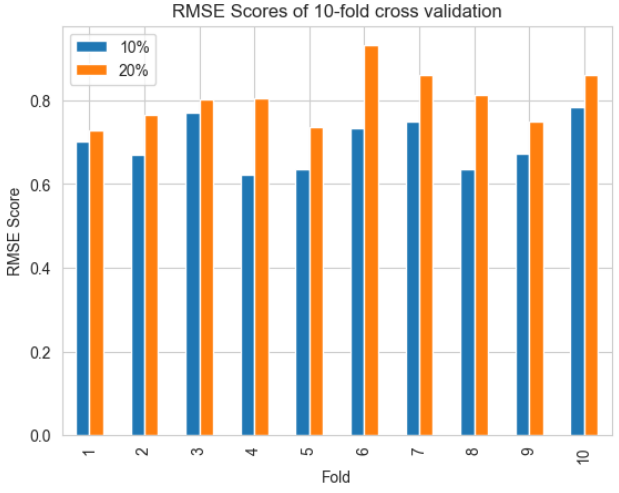
Data Evaluation

Setelah melewati tahap *data mining* model *linear regression* dilakukan evaluasi menggunakan teknik RMSE. Evaluasi RSME diimplementasikan terhadap hasil prediksi data dengan *missing value* sebesar 10% dan 20%. Evaluasi dilakukan untuk mengetahui efektifitas performa model *linear regression* terhadap data yang digunakan, berikut merupakan grafik hasil RMSE pada data dengan missing value 10% dan 20% dengan teknik imputasi SICE dan Mean menggunakan *data splitting*.



Gambar 8. Nilai RMSE *Linear Regression*

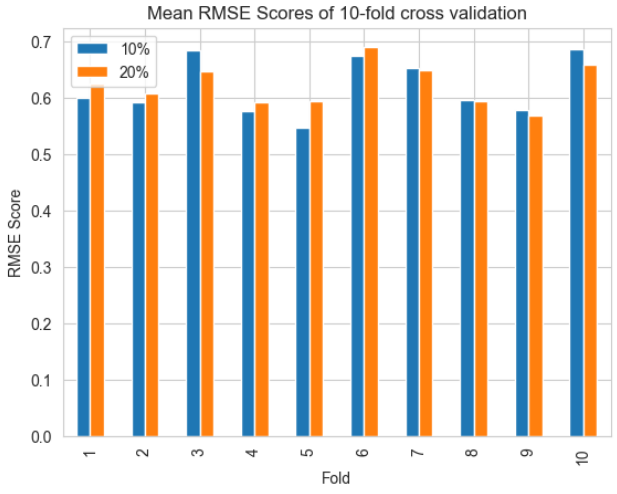
Nilai RMSE *linear regression* pada data dengan *missing value* 10% menggunakan teknik imputasi SICE sebesar 0.6882, sedangkan untuk *missing value* 20% sebesar 0.7914. Nilai RMSE *linear regression* pada data dengan *missing value* 10% menggunakan teknik imputasi *mean* memiliki nilai sebesar 0.6597, sedangkan untuk data dengan *missing value* 20% memiliki nilai sebesar 0.6065. Nilai evaluasi model linear regression menggunakan k-fold cross-validation dengan nilai k sebesar 10 pada data dengan missing value sebesar 10% dan 20% menggunakan teknik imputasi SICE ditunjukkan pada Gambar 9.



Gambar 9. Skor Evaluasi *RMSE* 10-*Fold Cross Validation* Imputasi *SICE*

Pada evaluasi model *linear regression* menggunakan *10-fold cross validation* untuk teknik imputasi *SICE* dengan tingkat *missing value* 10% mendapatkan nilai rata – rata *RMSE* sebesar 0,6973 dan skor rata – rata *RMSE*

sebesar 0,8 untuk tingkat *missing value* sebesar 20%. Evaluasi model *linear regression* menggunakan *10-fold cross validation* untuk teknik imputasi *mean* dapat dilihat pada Gambar 10.



Gambar 10. Skor Evaluasi *RMSE* 10-*Fold Cross Validation* Imputasi *Mean*

Pada evaluasi model *linear regression* menggunakan 10-*fold cross validation* untuk teknik imputasi mean dengan tingkat *missing value* 10% mendapatkan nilai rata – rata *RMSE* sebesar 0,619 dan skor rata – rata *RMSE* sebesar 0,623 untuk tingkat *missing value* sebesar 20%. Evaluasi 10-*fold cross-validation* menunjukkan bahwa untuk teknik imputasi *mean* memiliki hasil *RMSE* terbaik untuk tingkat *missing value* sebesar 10%, pada teknik imputasi *SICE* nilai *RMSE* terbaik didapatkan pada *missing value* sebesar 10%. Secara keseluruhan teknik imputasi *mean* lebih memiliki performa yang lebih baik dengan nilai *RMSE* lebih kecil baik pada *missing value* sebesar 10% dan 20% dibandingkan dengan teknik imputasi *SICE*.

Tabel 8. Tabel Kesimpulan Hasil Evaluasi

Teknik Imputasi	Missing Value	RMSE	Rataan RE
Mean (*Data Split*)	10%	0.6597	0.0934
	20%	0.6065	0.085
SICE (*Data Split*)	10%	0.6882	0.0984
	20%	0.7914	0.114
10-fold cross-validation (Mean)	10%	0.6192	-
	20%	0.6234	-
10-fold cross-validation (SICE)	10%	0.6973	-
	20%	0.8033	-

Pembahasan

Pada model *linear regression* imputasi *SICE* menggunakan *data splitting* dengan *missing value* 10% menghasilkan nilai *RMSE* sebesar 0.6882 dan nilai rata – rata kesalahan relatif sebesar 0,0984 serta nilai rata – rata RMSE menggunakan 10-*fold cross-validation* sebesar 0,6973. Model *linear regression* imputasi *SICE* menggunakan *data splitting* dengan *missing value* 20% mendapatkan nilai *RMSE* sebesar 0.7914 dan nilai rata – rata kesalahan relatif sebesar 0,114 serta rata – rata *RMSE* menggunakan 10-*fold cross-validation* sebesar 0,8033. Dari data tersebut dapat diketahui bahwa performa model *linear regression* dengan imputasi *SICE* menunjukkan hasil yang lebih baik dengan tingkat *missing value* sebesar 10% dengan nilai *RMSE* dan kesalahan relatif yang lebih kecil.

Model *linear regression* dengan imputasi *mean* menggunakan *data splitting* pada data dengan *missing value* 10% mendapatkan hasil *RMSE* dengan nilai 0.6597 dan nilai rata – rata kesalahan relatif sebesar 0,0934 serta rata – rata *RMSE* menggunakan 10-*fold cross-validation* sebesar 0,6192. Nilai *RMSE* menggunakan *data splitting* sebesar 0.6065 dan nilai rata – rata kesalahan relatif sebesar 0,0856 untuk data dengan *missing value* sebesar 20% serta nilai rata – rata *RMSE* menggunakan 10-*fold cross-validation* sebesar 0,6234. Berdasarkan data tersebut dapat diketahui bahwa model *linear regression* memiliki nilai *RMSE* dan kesalahan relatif terbaik dengan *missing value* pada data sebesar 20% berdasarkan *data splitting* dan nilai *RMSE* menggunakan 10-*fold cross-validation* menunjukkan bahwa *missing value* 10% menunjukkan hasil yang lebih baik, hal ini mungkin disebabkan oleh tingginya bias pada hasil evaluasi menggunakan *data splitting*.

Dari evaluasi *RMSE* berdasarkan *data splitting*, kesalahan relatif dan rata – rata nilai *RMSE* menggunakan 10-*fold cross-validation* model *linear regression* terhadap data dengan *missing value* menunjukkan bahwa model *linear regression* dengan teknik imputasi *mean* menunjukkan performa yang lebih baik dengan nilai *RMSE* menggunakan *data splitting*, kesalahan relatif dan nilai rata – rata *RMSE* menggunakan 10-*fold cross-validation* yang lebih kecil

jika dibandingkan dengan model *linear regression* dengan teknik imputasi *SICE*. Pada data dengan *missing value* 10% model *linear regression* dengan imputasi *mean* memiliki nilai *RMSE* 0.6597 menggunakan *data splitting*, 0,6192 untuk rata – rata *RMSE* berdasarkan 10-*fold cross-validation* dan nilai kesalahan relatif sebesar 0,0934. Model *linear regression* dengan imputasi *SICE* pada data dengan *missing value* 10% memiliki nilai *RMSE* 0.6882 menggunakan *data splitting*, 0,6973 untuk rata – rata *RMSE* menggunakan 10-*fold cross-validation* dan nilai kesalahan relatif sebesar 0,0986. Model dengan data *missing value* 20% menggunakan imputasi *mean* memiliki nilai *RMSE* 0.6065 menggunakan *data splitting*, 0,6234 untuk rata – rata nilai *RMSE* menggunakan 10-*fold cross-validation* dan nilai kesalahan relatif sebesar 0,0856. Model dengan imputasi *SICE* pada data dengan *missing value* 20% memiliki nilai *RMSE* 0.7914 menggunakan *data splitting*, 0,8033 untuk rata – rata nilai *RMSE* menggunakan 10-*fold cross-validation* dan nilai kesalahan relatif sebesar 0,114.

Hal ini juga menunjukkan bahwa teknik imputasi *single variate* memberikan performa yang lebih baik terhadap model *linear regression* untuk data *red wine quality* dibandingkan dengan teknik imputasi *multi variate*. Hasil ini disebabkan pada imputasi *missing value* tidak dilakukan penanganan *outlier* sehingga hasil imputasi *SICE* memiliki performa yang kurang optimum jika dibandingkan dengan hasil imputasi mean. Seperti yang dikatan Maniruzzaman (2018) pembelajaran mesin yang menggunakan data tanpa penanganan outlier akan mengurangi performa akurasi prediksi sehingga pada penelitian ini imputasi *SICE* memiliki performa yang kurang baik dikarenakan tidak dilakukan penanganan *outlier* sebelum teknik imputasi diterapkan.

4. Kesimpulan

Hasil dari evaluasi menggunakan *RMSE* untuk *data splitting* atau 10-*fold cross-validatoin* serta kesalahan relatif menunjukkan bahwa *model linear regression* dengan teknik imputasi *mean* memiliki hasil yang lebih baik dengan nilai RMSE sebesar 0.6597 dan nilai kesalahan relatif sebesar 0,0934 untuk *missing value* 10% menggunakan *data splitting* dan nilai rata – rata *RMSE* menggunakan 10-*fold cross-validation* sebesar 0.6192. Nilai RMSE sebesar 0.6065 dan nilai kesalahan relatif sebesar 0,0856 untuk data dengan *missing value* 20% teknik imputasi *mean* dengan menggunakan *data splitting* dan nilai rata – rata *RMSE* menggunakan 10-*fold cross validation* sebesar 0.6234 dibandingkan dengan model *linear regression* menggunakan teknik imputasi SICE yang memiliki nilai *RMSE* sebesar 0.6882 dan nilai kesalahan relatif sebesar 0,0986 untuk *missing value* 10% menggunakan *data splitting* dan nilai rata – rata *RMSE* 10-*fold cross-validation* sebesar 0.6973. Nilai RMSE 0.7914 serta kesalahan relatif sebesar 0,114 untuk *missing value* 20% menggunakan *data splitting* dan nilai rata – rata *RMSE* sebesar 0.8033 menggunakan 10-*fold cross-validation*.

5. Daftar Pustaka

[1] Jordanov, I., Petrov, N. and Petrozziello, A., 2018. Classifiers accuracy improvement based on missing data imputation. *Journal of Artificial Intelligence and Soft Computing Research*, *8*(1), pp.31-48. DOI: https://doi.org/10.1515/jaiscr-2018-0002.

[2] Jadhav, A., Pramod, D. and Ramanathan, K., 2019. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), pp.913-933. DOI: https://doi.org/10.1080/08839514.2019.1637138.

[3] Enders, C.K., 2022. *Applied missing data analysis*. Guilford Publications. Available at: https://lccn.loc.gov/2022009851.

[4] Afarisi, R., Tjandrasa, H. and Arieshanti, I., 2018. Perbandingan performa antara imputasi metode konvensional dan imputasi dengan algoritma mutual nearest neighbor. Jurnal Teknik Pomits, 2(1), pp.73-76. DOI: https://doi.org/10.12962/j23373539.v2i1.2735.

[5] Han, J., Kamber, M. and Pei, J., 2011. Data Mining: Concepts and techniques. Cambridge: Elsevier. DOI: https://doi.org/10.1016/C2009-0-61819-5.

[6] Susanti, P. and Azizah, N., 2019. Imputation of missing value using dynamic bayesian network for multivariate time series data. International conference on data and software engineering, 1, pp.1-5. DOI: https://doi.org/10.1109/ICODSE.2017.8285864.

[7] Das, D.D., Nayak, M. and Pani, S. K., 2019. Missing value imputation a review. International Journal of Computer Science and Engineering. 7(4), pp.548-558. DOI: https://doi.org/10.26438/ijcse/v7i4.548558.

[8] Jajuli, M. and Komarudin, O., 2017. Implementasi Metode Impusati Mean dan Expectation Maximisation terhadap Hasil Clustering k-Means Mahasiswa Pelamar Beasiswa Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang. SESIOMADIKA, 1, pp.19-27. Available at: http://pmat-unsika.eu5.org/Prosiding/4MohammadJajuli-SESIOMADIKA-2017.pdf.

[9] Young, W., Weckman, G. and Holland, W., (2011). A survey of methodologies for the treatment of missing values within datasets: limitations and benefits. Theoretical Issue in Ergonomics Science, 12(1), pp.16-43. DOI: https://doi.org/10.1080/14639220903470205

[10] Khan, S. and Hoque, A., 2020. SICE: an improved missing data imputation technique. Journal of Big Data, 7(37), pp.1-21. DOI: https://doi.org/10.1186/s40537-020-00313-w.

[11] Ilham, A., 2020. Hybrid Metode Boostrap dan Teknik Imputasi Pada Metode C4-5 untuk Prediksi Penyakit Ginjal Kronis. Statistika, 8(1), pp.43-51. Available at: https://garuda.kemdikbud.go.id/documents/detail/1767581.

[12] Joseph, V., 2022. Optimal Ratio for Data Splitting. Stat Anal Data Min: The ASA Data Sci Journal, 1, pp.531-538. DOI: https://doi.org/10.1002/sam.11583.

[13] Marcot, B. and Hanea, A., 2021. What is an Optimal Value of k in k-Fold Cross-Validation in Discrete Bayesian Network Analysis. Computational Statistics, 36(1), pp.2009-2031. DOI: https://doi.org/10.1007/s00180-020-00999-9.

[14] Berrar, D., 2018. Cross-Validation. Data Science Laboratory, 1, pp.542-545. DOI: https://doi.org/10.1007/s00180-020-00999-9.

[15] SinSomboonthong, S., 2022. Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification. International Journal of Mathematics and Mathematical Sciences, 1, pp.1-9. DOI: https://doi.org/10.1155/2022/3584406.

[16] Chai, T. and Draxler, R., 2014. Root mean squared error (RMSE) or mean absolute error (MAE)?. Geosci. Model Dev, 7(1), pp.1525-1534. DOI: https://doi.org/10.5194/gmdd-7-1525-2014.