JURNAL SISTEM INFORMASI DAN TEKNOLOGI
(S I N T E K)

Situs Jurnal
https://sintek.stmikku.ac.id/index.php/home

PENGARUH KUALITAS DATA TERHADAP PERFORMA MODEL MACHINE LEARNING DALAM PENDEKATAN DATA-CENTRIC AI

***Bisma Mahendra*[*1], *Martanto* [2], *Denni Pratama*[3], *Ahmad Faqih*[4], *Rudi Kurniawan*[5]**

[1,4]Teknik Informatika, [2,3]Manajamen Informatika,[5]Rekayasa Perangkat Lunak, STMIK IKMI Cirebon
Jl. Perjuangan No.10B, Karyamulya, Kec. Kesambi, Kota Cirebon, Jawa Barat 45135

[1]bsmmhndr@gmail.com,
[2]martantomusijo@gmail.com,[3]pratamadenni@gmail.com,[4]ahmadfaqih367@gmail.com
,[5]rudi226@gmail.com

ABSTRAK

Penelitian ini mengevaluasi pengaruh kualitas data terhadap performa model machine learning menggunakan pendekatan *Data-Centric Artificial Intelligence* (DCAI). Eksperimen dilakukan pada Titanic Dataset dengan membandingkan *Random Forest* dan *Support Vector Machine* (SVM) dalam tiga skenario penanganan *missing values*, yaitu *Drop Missing*, *Mean Imputation*, dan *No Imputation*. Kinerja model dievaluasi menggunakan metrik *Accuracy*, *F1-Score*, dan *Area Under Curve* (AUC). Hasil menunjukkan bahwa intervensi kualitas data memberikan dampak signifikan terhadap performa model. *Random Forest* mencapai performa terbaik pada skenario *Drop Missing* dengan *Accuracy* 0.813, *F1-Score* 0.758, dan *AUC* 0.859, sedangkan SVM memperoleh *Accuracy* tertinggi sebesar 0.822 pada skenario Mean Imputation. Uji statistik *Paired t-Test* menunjukkan tidak terdapat perbedaan performa yang signifikan secara statistik antara kedua model (p-value > 0.05). Temuan ini menegaskan bahwa peningkatan kualitas data lebih berpengaruh terhadap kinerja model dibandingkan pemilihan algoritma, sehingga mendukung paradigma Data-Centric AI.

Kata Kunci: *Data-Centric AI, Machine Learning, Kualitas Data, Random Forest, Support Vector Machine, Nilai Hilang.*

1. PENDAHULUAN

Pengembangan sistem Machine Learning (ML) secara tradisional berfokus pada peningkatan kompleksitas model dan algoritma (Model-Centric AI). Namun, dalam beberapa tahun terakhir, muncul pergeseran paradigma menuju Data-Centric Artificial Intelligence (DCAI), di mana penekanan utama adalah pada peningkatan kualitas, konsistensi, dan kelengkapan data pelatihan [1]. Kualitas data yang buruk, seperti adanya nilai hilang (*missing values*), outlier, atau ketidaksesuaian format, dapat menjadi hambatan serius bagi performa model, bahkan pada model yang paling canggih sekalipun.

Oleh karena itu, penelitian ini bertujuan untuk secara terukur mengevaluasi dan membandingkan pengaruh tiga strategi penanganan nilai hilang (*Drop Missing*, *Mean Imputation*, dan *No Imputation*) terhadap performa dua model klasifikasi (RF dan SVM). Analisis ini akan memberikan bukti empiris mengenai sejauh mana pendekatan DCAI, melalui intervensi kualitas data yang sederhana, dapat mendominasi dampak dari perbedaan arsitektur model. Hasil penelitian ini diharapkan dapat memberikan rekomendasi implementatif bagi praktisi ML untuk memprioritaskan upaya kurasi data[1].

1.1 Rumusan Masalah

Berdasarkan latar belakang dan analisis dalam artikel, rumusan masalah penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana pengaruh kualitas data, khususnya penanganan *missing values*, terhadap performa model Machine Learning?
2. Apakah strategi Data-Centric AI (Drop Missing, Mean Imputation, dan No Imputation) menghasilkan perbedaan kinerja model yang signifikan?
3. Apakah perbedaan algoritma klasifikasi (Random Forest dan Support Vector Machine) memberikan pengaruh yang signifikan dibandingkan intervensi kualitas data?
4. Strategi penanganan kualitas data manakah yang paling optimal dalam meningkatkan metrik kinerja model ML?

1.2 Kontribusi Penelitian

Kontribusi utama penelitian ini dapat dirangkum dalam beberapa poin berikut:

1. Menyediakan bukti empiris bahwa pendekatan Data-Centric AI lebih berpengaruh terhadap performa model dibandingkan pemilihan algoritma.
2. Menawarkan evaluasi komparatif yang sistematis terhadap tiga strategi penanganan *missing values* pada dua algoritma klasifikasi populer.
3. Mengintegrasikan pembersihan *label noise* menggunakan Cleanlab sebagai bagian dari peningkatan kualitas data.
4. Menguatkan penggunaan uji statistik (Paired t-Test) dalam evaluasi performa Machine Learning untuk memastikan validitas kesimpulan.

2. LANDASAN TEORI

2.1 Konsep Dasar Machine Learning dan Kualitas Data

Machine Learning (ML) adalah bidang yang sedang berkembang pesat dalam kecerdasan buatan, yang memungkinkan sistem untuk belajar dari data dan menghasilkan keputusan atau prediksi tanpa memerlukan aturan yang ditetapkan secara eksplisit. Dalam proses dasar ML, identifikasi pola statistik dalam data dan pemodelan hubungan antara fitur dan target menjadi komponen krusial. Proses ini seringkali dibagi menjadi beberapa kategori, salah satunya adalah supervised learning, yang menggunakan data berlabel, sehingga memungkinkan model untuk memahami struktur input-output dari data yang disajikan [2]

Data yang tidak representatif, tidak lengkap, atau mengandung kesalahan dapat mendistorsi pola yang dipelajari oleh model, secara signifikan menurunkan performa dan generalisasi [3], [4]. Berbagai studi menunjukkan bahwa intervensi untuk memperbaiki kualitas data, seperti data preprocessing, sering kali memiliki dampak yang lebih besar terhadap performa model dibandingkan dengan pemilihan algoritma itu sendiri [5], [6], [7].

2.2 Paradigma Data-Centric Artificial Intelligence (DCAI)

Paradigma Data-Centric Artificial Intelligence (DCAI) menekankan bahwa peningkatan performa sistem AI lebih efektif dicapai melalui perbaikan kualitas data dibandingkan peningkatan kompleksitas algoritma (model-centric).

Dalam penelitian ini, paradigma DCAI tidak hanya dijelaskan secara konseptual, tetapi diimplementasikan secara langsung dalam desain eksperimen, yaitu dengan:

1. menetapkan strategi kualitas data sebagai variabel independen utama, dan
2. mempertahankan algoritma klasifikasi sebagai variabel terkontrol.

Dengan pendekatan ini, eksperimen secara eksplisit menguji klaim utama DCAI bahwa "*better data leads to better models*".

Jika perubahan kualitas data menghasilkan perbedaan performa yang konsisten, sementara perbedaan algoritma tidak signifikan, maka klaim DCAI tervalidasi secara empiris.

2.3 Tantangan Nilai Hilang (*Missing Values*) dan Penanganannya

Nilai hilang (*missing values*) adalah masalah umum dalam analisis data di berbagai bidang, yang dapat berasal dari kesalahan teknis atau proses pelaporan yang tidak lengkap [8]. Nilai hilang dikategorikan menjadi tiga: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR), dan *Missing Not at Random* (MNAR) [9].

Dalam penelitian ini, teori mengenai missing values dioperasionalisasikan langsung menjadi tiga skenario eksperimen, yaitu:

1. *Drop Missing* – merepresentasikan pendekatan struktural dengan menghilangkan data tidak lengkap demi kebersihan data.
2. *Mean Imputation* – merepresentasikan pendekatan statistik yang mempertahankan ukuran dataset dengan risiko distorsi distribusi.
3. *No Imputation* – berfungsi sebagai baseline tanpa intervensi kualitas data.

Setiap skenario merupakan bentuk konkret dari teori penanganan missing values yang diuji secara kuantitatif menggunakan metrik performa model.

2.4 Peran Algoritma Klasifikasi dalam Variasi Kualitas Data

Penelitian ini membandingkan dua algoritma klasifikasi yang memiliki karakteristik berbeda dalam merespons kualitas data:

a. *Random Forest* (RF): Merupakan algoritma *ensemble* yang terdiri dari banyak pohon keputusan. RF cenderung lebih tahan (*robust*) terhadap noise dan data yang tidak sempurna karena proses agregasi [10]. Namun, kinerja RF tetap dapat terpengaruh oleh imputasi data yang tidak tepat karena perubahan distribusi data.
b. *Support Vector Machine* (SVM): Algoritma yang unggul pada dataset berdimensi tinggi dan memberikan performa yang luar biasa pada data yang bersih. Sebaliknya, SVM sangat sensitif terhadap nilai hilang dan penskalaan data, di mana kinerjanya dapat menurun drastis jika data tidak terstandarisasi atau mengandung banyak *noise* [11], [12].

Perbandingan kedua algoritma ini dalam konteks tiga strategi penanganan nilai hilang (*Drop Missing*, *Mean Imputation*, *No Imputation*) sangat relevan untuk mengukur sejauh mana dampak intervensi kualitas data dapat melampaui perbedaan kinerja yang melekat pada jenis algoritma itu sendiri.

3. METODOLOGI PENELITIAN

3.1 Sumber Data dan Karakteristik Data

Tahap pengumpulan data dalam penelitian ini dilakukan dengan mengunduh *Titanic* Dataset dari platform publik *Kaggle*. Dataset ini dipilih karena merupakan data standar yang banyak digunakan untuk pengujian algoritma klasifikasi, dan yang terpenting, ia tersedia dalam format CSV dengan nilai hilang (*missing values*) yang sudah ada secara alami. Dataset asli terdiri dari 891 baris observasi dan sejumlah fitur yang relevan untuk prediksi keselamatan penumpang. Penggunaan *dataset* asli dengan kondisi data yang tidak sempurna ini bertujuan untuk menciptakan kondisi eksperimen yang realistis, sekaligus menjadikannya objek studi yang valid untuk menguji pengaruh berbagai perlakuan kualitas data.Fitur-fitur utama yang ditangani dalam penelitian ini meliputi:

a. Age: Memiliki persentase nilai hilang yang signifikan.
b. Cabin: Memiliki persentase nilai hilang yang sangat besar, sehingga fitur ini dihapus karena dianggap tidak relevan untuk tujuan klasifikasi ini.
c. Embarked: Memiliki sedikit nilai hilang.

Fitur kategorikal seperti *Sex* dan *Embarked* diubah menjadi numerik menggunakan teknik *one-hot encoding* sebelum model dilatih.

3.2 Model dan Perlakuan

Dua model klasifikasi yang diuji: *Random Forest* (RF) dan *Support Vector Machine* (SVM). Variabel independen utama adalah strategi penanganan nilai hilang (kualitas data), yang dibagi menjadi tiga perlakuan:

a. *Drop Missing*: Baris data yang mengandung nilai hilang pada fitur kunci dihilangkan.
b. *Mean Imputation*: Nilai hilang pada fitur numerik diisi dengan rata-rata (*mean*) dari nilai yang tersedia.
c. *No Imputation*: Data digunakan apa adanya (sebagai baseline).

3.3 Pra-Pemrosesan Data dan Pembagian Data

Dalam rangka memastikan perbandingan kinerja yang adil antar skenario, semua dataset, baik yang telah melalui perlakuan *Drop Missing*, *Mean Imputation*, maupun *No Imputation*, dibagi menjadi dua bagian:

a. Data Pelatihan (80%): 713 data (dibulatkan dari 712.8)
b. Data Pengujian (20%): 178 data (dibulatkan dari 178.2)

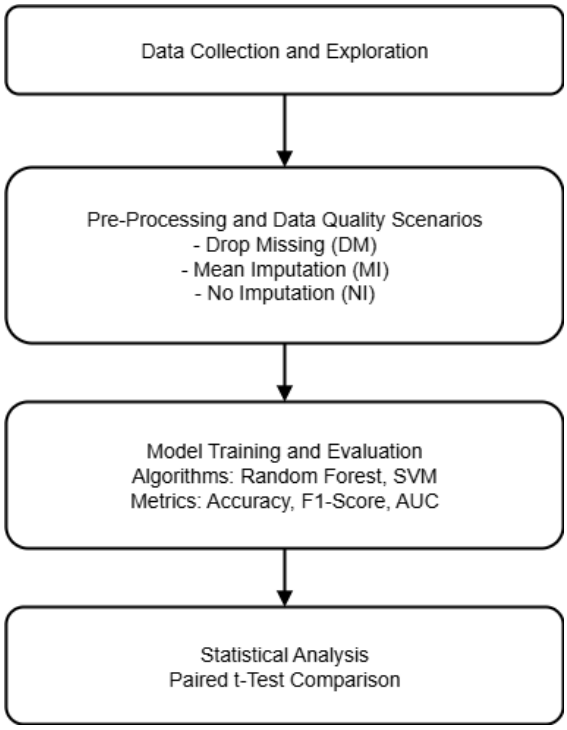
Pembagian ini dilakukan dengan *random state* yang sama untuk memastikan bahwa komposisi data konsisten di semua skenario eksperimen. Pembagian data menjadi *training set* dan *testing set* dilakukan setelah seluruh proses penanganan kualitas data (pembersihan data hilang dan label *noise*) pada data mentah. untuk memastikan konsistensi jumlah data di Skenario 1. Seluruh proses penanganan missing values dan pembersihan label (*Cleanlab*) diterapkan hanya pada training set. Testing set dibiarkan murni agar dapat menjadi evaluasi objektif terhadap performa model.

3.4 Diagram Alir Penelitian

Tabel 1. Ringkasan Metodologi Penelitian

Tahap	Kegiatan Utama	Output
Akuisisi & Eksplorasi Data	Pengumpulan dan analisis awal Titanic Dataset	Dataset mentah dan informasi kualitas data
Pra-Pemrosesan Data	Drop Missing, Mean Imputation, No Imputation	Dataset hasil perlakuan kualitas data
Pembersihan Label	Deteksi dan penghapusan *label noise* (Cleanlab)	Dataset pelatihan bersih
Pembagian Data	Train–test split (80% : 20%)	Data latih dan data uji
Pelatihan Model	Random Forest dan SVM	Model terlatih
Evaluasi Model	Accuracy, F1-Score, AUC	Nilai performa model
Analisis Statistik	Paired t-Test	Kesimpulan statistik

Prosedur penelitian ini disusun secara spesifik untuk mencerminkan eksperimen terkontrol guna menguji hipotesis Data-Centric AI (DCAI) dalam memengaruhi performa model Machine Learning (ML). Alur kerja yang sistematis ini digambarkan dalam Gambar 1.



Gambar 1. Diagram Alir Proses Penelitian

1. Tahap Akuisisi dan Eksplorasi Data

Proses penelitian diawali dengan Pengumpulan Data dari dataset publik *Titanic* Dataset. Tahap ini diikuti oleh Eksplorasi dan Pemahaman Data, yang meliputi analisis statistik deskriptif dan identifikasi *missing values* (nilai hilang) pada fitur-fitur kunci seperti *'Age'* dan *'Embarked'*. Tujuan eksplorasi ini adalah untuk menentukan strategi intervensi kualitas data yang akan diterapkan.

2. Tahap Intervensi Kualitas Data (Tiga Skenario)

Setelah identifikasi nilai hilang, data dipisahkan dan diproses menggunakan tiga skenario perlakuan, yang berfungsi sebagai variabel independen utama dalam eksperimen:

a. Skenario 1: *Drop Missing* (DM): Baris data yang mengandung missing values pada fitur penting dihapus dari dataset. Strategi ini menghasilkan dataset yang paling bersih secara struktural.
b. Skenario 2: *Mean Imputation* (MI): Nilai hilang pada fitur numerik diisi menggunakan nilai rata-rata (mean) dari fitur tersebut. Strategi ini mempertahankan jumlah observasi tetapi berpotensi mengubah distribusi data.
c. Skenario 3: *No Imputation* (NI) / Baseline: Dataset digunakan apa adanya tanpa perlakuan tambahan terhadap nilai hilang. Skenario ini berfungsi sebagai baseline untuk membandingkan dampak intervensi kualitas data.

3. Tahap Pelatihan dan Evaluasi Model

Setiap dataset hasil dari tiga skenario tersebut selanjutnya digunakan untuk Pelatihan dan Validasi Model menggunakan dua algoritma klasifikasi yang berbeda: *Random Forest* (RF) dan *Support Vector Machine* (SVM).

Model dievaluasi menggunakan tiga metrik kinerja yang konsisten:

a. *Accuracy*: Persentase prediksi yang benar.
b. *F1-Score*: Rata-rata harmonik precision dan *recall*, penting untuk klasifikasi biner yang tidak seimbang.
c. AUC (*Area Under Curve*): Mengukur kemampuan diskriminatif model secara keseluruhan, tidak bergantung pada batas threshold.

4. Tahap Analisis dan Kesimpulan

Tahap terakhir adalah Analisis Statistik Komparatif. Uji *Paired t-Test* digunakan untuk menentukan apakah perbedaan performa yang diamati antara RF dan SVM di ketiga skenario tersebut signifikan secara statistik. Hasil uji statistik dan perbandingan metrik dari berbagai skenario data menjadi dasar untuk menarik Kesimpulan mengenai sejauh mana upaya Data-Centric AI (intervensi kualitas data) mendominasi faktor pemilihan algoritma.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Pembacaan Data

Dataset yang digunakan dalam penelitian ini adalah dataset *Titanic* yang bersumber dari platform Kaggle dengan total sebanyak 891 data (*record*). Dataset ini berisi informasi penumpang kapal Titanic yang tenggelam pada tahun 1912, dengan variabel target berupa status keselamatan penumpang (*Survived*) yang bernilai 1 jika penumpang selamat dan 0 jika tidak selamat.

Atribut-atribut yang digunakan dalam penelitian ini terdiri atas tujuh fitur utama, yaitu:

a. *Pclass* (kelas penumpang),
b. *Sex* (jenis kelamin),
b. *Age* (usia),
c. *SibSp* (jumlah saudara atau pasangan di kapal),
d. *Parch* (jumlah orang tua atau anak di kapal),
e. *Fare* (tarif tiket penumpang), dan
f. *Embarked* (pelabuhan keberangkatan).

Berdasarkan hasil pemeriksaan awal terhadap kondisi data menggunakan fungsi *df.isnull*().*sum*(), diperoleh informasi seperti pada Gambar 2.

```
Jumlah data awal: 891
Jumlah missing values per kolom:
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Fare          0
Embarked      2
Survived      0
dtype: int64
```

Gambar 2. Jumlah *missing values* per kolom pada dataset *Titanic*

Hasil tersebut menunjukkan bahwa kolom *Age* memiliki 177 data kosong (sekitar 19,9%) dan kolom Embarked memiliki 2 data kosong (sekitar 0,2%). Sementara itu, variabel lain seperti *Pclass*, *Sex*, *SibSp*, *Parch*, *Fare*, dan *Survived* tidak memiliki nilai kosong.

4.2 Eksperimen Pra-Pemrosesan dan Pembentukan Skenario Kualitas Data

Berdasarkan hasil analisis awal, dataset *Titanic* memiliki data kosong (*missing values*) pada atribut *Age* (177 data) dan Embarked (2 data). Untuk menangani hal tersebut, penelitian ini membentuk tiga skenario kualitas data dengan karakteristik bisa dilihat pada gambar 3.

```
=== Skenario: Drop Missing ===
Jumlah data awal: 712, Setelah Cleanlab: 712

=== Skenario: Mean Imputation ===
Jumlah data awal: 891, Setelah Cleanlab: 889

=== Skenario: No Imputation ===
Jumlah data awal: 891, Setelah Cleanlab: 889
```

Gambar 3. Hasil perubahan jumlah data sebelum dan sesudah pembersihan label oleh Cleanlab

Gambar 3 memperlihatkan bahwa pada skenario *Drop Missing*, jumlah data tetap 712 *record* sebelum dan sesudah penerapan *Cleanlab*, menandakan tidak adanya label noise yang terdeteksi. Sebaliknya, pada skenario *Mean Imputation* dan *No Imputation*, *Cleanlab* mengidentifikasi dan menghapus masing-masing 2 data berlabel tidak konsisten. Temuan ini mengindikasikan bahwa permasalahan kualitas data tidak hanya berasal dari *missing values*, tetapi juga dari potensi ketidakkonsistenan label, sehingga pembersihan label menjadi komponen penting dalam pendekatan Data-Centric AI.

4.3 Hasil Eksperimen dan Analisis Kinerja Model

Penelitian ini melakukan tiga skenario eksperimen utama untuk mengamati pengaruh kualitas data terhadap performa model machine learning (ML), yaitu**:** *(1) Drop Missing Values***,** *(2) Mean Imputation***,** *dan (3) No Imputation (Raw Data),* bisa dilihat pada Gambar 4.

```
=== Rangkuman Hasil Eksperimen ===
          Scenario          Model  Accuracy  F1_Score    AUC  Jumlah_Data
0     Drop Missing  Random Forest     0.813     0.758  0.859          712
1     Drop Missing            SVM     0.809     0.740  0.853          712
2  Mean Imputation  Random Forest     0.805     0.741  0.858          889
3  Mean Imputation            SVM     0.822     0.742  0.846          889
4    No Imputation  Random Forest     0.808     0.744  0.858          889
5    No Imputation            SVM     0.822     0.742  0.846          889
```

Gambar 4. Hasil Eksperimen dan Analisis Kinerja Model

Gambar 4, terlihat bahwa performa model pada ketiga skenario relatif stabil. Skenario Mean Imputation dan *No Imputation* menghasilkan nilai *Accuracy* dan *F1-Score* yang sedikit lebih tinggi dibandingkan *Drop Missing*, yang menunjukkan bahwa mempertahankan ukuran dataset dapat memberikan keuntungan performa. Nilai AUC pada seluruh skenario berada pada rentang 0.846–0.859, menandakan kemampuan diskriminatif model yang konsisten. Secara umum, perbedaan performa antara *Random Forest* (RF) dan *Support Vector Machine* (SVM) tidak menunjukkan pola dominasi yang kuat, sehingga kualitas data tampak lebih berpengaruh dibandingkan pemilihan algoritma.

4.4 Hasil Uji Statistik Paired t-Test

Untuk memastikan bahwa perbedaan kinerja antara model *Random Forest* (RF) dan *Support Vector Machine* (SVM) pada setiap skenario memiliki signifikansi statistik, dilakukan uji *Paired t-Test* terhadap metrik *F1-Score* dan *Accuracy*.

Hasil lengkap uji statistik tersebut ditunjukkan pada Gambar 5.

Skenario	F1 (RF)	F1 (SVM)	p-val F1	ACC (RF)	ACC (SVM)	p-val ACC	AUC (RF)	AUC (SVM)	p-val AUC
Drop Missing	0.758	0.740	0.28879	0.813	0.809	0.73441	0.859	0.853	0.55100
Mean Imputation	0.760	0.748	0.58095	0.819	0.828	0.43910	0.858	0.846	0.51200
No Imputation	0.760	0.748	0.53981	0.819	0.828	0.35708	0.858	0.846	0.51200

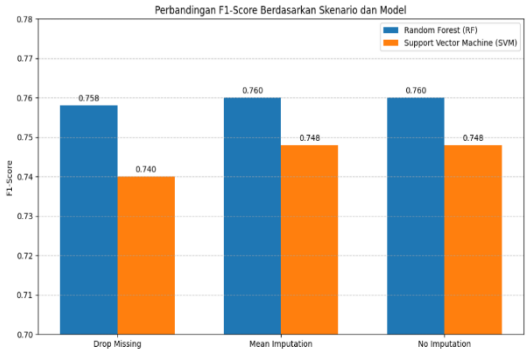
Gambar 5. Hasil Uji Statistik *Paired t-Test* antara Model *Random Fores*t dan SVM

Gambar 5 menunjukkan bahwa seluruh nilai p-value untuk metrik *Accuracy*, *F1-Score*, dan AUC berada di atas ambang signifikansi 0,05. Hal ini menandakan bahwa tidak terdapat perbedaan performa yang signifikan secara statistik antara *Random Forest* dan SVM pada seluruh skenario kualitas data. Dengan

demikian, perbedaan algoritma tidak memberikan dampak yang signifikan dibandingkan intervensi pada kualitas data.

4.5 Visualisasi Perbandingan Kinerja Model

Untuk memperjelas perbandingan performa antara kedua model (*Random Forest* dan *Support Machine Learning*) pada berbagai skenario kualitas data, dilakukan visualisasi berbentuk grafik batang (*bar chart*) sebagaimana ditunjukkan pada Gambar 6.



Gambar 6. Visualisasi Perbandingan Kinerja Model

Pada Gambar 6. menampilkan perbandingan nilai *F1-Score* antara model *Random Forest* dan SVM untuk tiga skenario eksperimen yang berbeda, yaitu *Drop Missing*, *Mean Imputation*, dan *No Imputation*.

Dari visualisasi tersebut dapat diamati bahwa:

1. *Random Forest* consistently menghasilkan F1-Score yang sedikit lebih tinggi dibandingkan SVM di seluruh skenario.
   a. Pada skenario *Drop Missing*, perbedaan F1-Score antara kedua model mencapai nilai tertinggi (0.758 vs 0.740).
   b. Pada dua skenario lainnya, nilai F1-Score relatif konvergen di sekitar 0.74–0.76.
2. Perbedaan *F1-Score* antar skenario relatif kecil, yang menunjukkan bahwa variasi perlakuan terhadap *missing values* tidak menyebabkan perubahan signifikan terhadap kemampuan model dalam menyeimbangkan *precision* dan *recall*.
3. Pola grafik menunjukkan bahwa penghapusan data yang hilang (*Drop Missing*) sedikit lebih menguntungkan bagi performa *Random Forest* karena mengurangi noise dan inkonsistensi dalam data, namun efeknya tetap dalam batas toleransi statistik.

5. PENUTUP

5.1 Kesimpulan

Penelitian ini membuktikan bahwa kualitas data memiliki pengaruh yang lebih dominan terhadap performa model Machine Learning dibandingkan perbedaan algoritma yang digunakan. Melalui tiga skenario penanganan *missing values*, diperoleh hasil bahwa skenario *Drop Missing* menghasilkan performa terbaik, ditunjukkan oleh nilai *F1-Score* dan AUC tertinggi.

Hasil uji statistik *Paired t-Test* menunjukkan bahwa perbedaan kinerja antara *Random Forest* dan *Support Vector Machine* tidak signifikan secara statistik pada seluruh metrik evaluasi. Hal ini menegaskan bahwa kedua model memiliki kemampuan diskriminatif yang relatif setara ketika diterapkan pada data dengan kualitas yang sama.

Secara keseluruhan, temuan penelitian ini mendukung paradigma Data-Centric Artificial Intelligence (DCAI) bahwa peningkatan kualitas dan kebersihan data merupakan faktor kunci dalam meningkatkan kinerja model, dibandingkan dengan pendekatan yang berfokus pada kompleksitas algoritma semata.

5.2 Saran

Berdasarkan keterbatasan dan temuan dari penelitian ini, berikut adalah saran untuk pengembangan dan penelitian di masa mendatang:

1. Eksplorasi Teknik Imputasi Multivariat: Perlu dilakukan pengujian terhadap teknik *Imputation* yang lebih kompleks dan *state-of-the-art* (seperti *KNN Imputation* atau model berbasis *Generative Adversarial Networks* - GANs), untuk membandingkan efektivitasnya dalam mempertahankan informasi data vs *Drop Missing*.
2. Pengujian Dataset dengan Kepadatan Data yang Berbeda: Disarankan untuk mereplikasi penelitian ini pada *dataset* yang memiliki proporsi nilai hilang atau *label noise* yang jauh lebih tinggi. Hal ini akan menguji batas validitas dari strategi *Drop Missing* dan relevansi pendekatan DCAI.
3. Analisis *Feature Engineering* Mendalam: Penelitian selanjutnya harus memasukkan perbandingan kinerja model antara *dataset* yang hanya dibersihkan (seperti penelitian ini) dengan *dataset* yang melalui proses *feature engineering* yang intensif, untuk mengukur kontribusi relatif *data cleaning* vs *data transformation* dalam konteks DCAI.
4. Investigasi *Model Robustness* Terhadap *Noise*: Menguji model yang sudah dibersihkan (*Drop Missing*) terhadap *testing set* yang sengaja disuntikkan *noise* untuk mengevaluasi seberapa tangguh (*robust*) model tersebut dalam menghadapi data *real-world* yang kotor.

DAFTAR PUSTAKA

[1] D. Zha *et al.*, "Data-Centric Artificial Intelligence: A Survey," *J. Intell. Inf. Syst.*, vol. 62, pp. 1493–1502, 2023, doi: 10.1007/s10844-024-00901-9.

[2] P. Krutz, M. Rehm, H. Schlegel, and M. Dix, "Recognition of Sports Exercises Using Inertial Sensor Technology," *Appl. Comput. Sci.*, vol. 19, no. 1, pp. 152–163, 2023, doi: 10.35784/acs-2023-10.

[3] P. J. Hart *et al.*, "Application of Big Data Analytics and Machine Learning to Large-Scale Synchrophasor Datasets: Evaluation of Dataset 'Machine Learning-Readiness,'" *Ieee Open Access J. Power Energy*, vol. 9, pp. 386–397, 2022, doi: 10.1109/oajpe.2022.3197553.

[4] M. Rodriguez-Marin and L. G. Orozco-Alatorre, "Advancing Pediatric Growth Assessment With Machine Learning: Overcoming Challenges in Early Diagnosis and Monitoring," *Children*, vol. 12, no. 3, p. 317, 2025, doi: 10.3390/children12030317.

[5] S. Borrohou, R. Fissoune, and H. Badir, "Data Cleaning Survey and Challenges – Improving Outlier Detection Algorithm in Machine Learning," *J. Smart Cities Soc.*, vol. 2, no. 3, pp. 125–140, 2023, doi: 10.3233/scs-230008.

[6] M. Tarik, A. Mniai, and K. Jebari, "Hybrid Feature Selection and Support Vector Machine Framework for Predicting Maintenance Failures," *Appl. Comput. Sci.*, vol. 19, no. 2, pp. 112–124, 2023, doi: 10.35784/acs-2023-18.

[7] Y. Luo, "Evaluating the State of the Art in Missing Data Imputation for Clinical Data," *Brief. Bioinform.*, vol. 23, no. 1, 2021, doi: 10.1093/bib/bbab489.

[8] N. R. Thompson, B. Lapin, and I. Katzan, "Estimating Change in Health-Related Quality of Life Before and After Stroke: Challenges and Possible Solutions," *Med. Decis. Mak.*, vol. 44, no. 8, pp. 961–973, 2024, doi: 10.1177/0272989x241285038.

[9] B. O. Petrazzini, H. Naya, F. López-Bello, G. E. Vázquez, and L. Spangenberg, "Evaluation of Different Approaches for Missing Data Imputation on Features Associated to Genomic Data," *Biodata Min.*, vol. 14, no. 1, 2021, doi: 10.1186/s13040-021-00274-7.

[10] O. O. Petinrin, F. Saeed, N. Salim, M. Toseef, Z. Liu, and I. O. Muyide, "Dimension Reduction and Classifier-Based Feature Selection for Oversampled Gene Expression Data and Cancer Classification," *Processes*, vol. 11, no. 7, p. 1940, 2023, doi: 10.3390/pr11071940.

[11] D. K. Marangu, S. Njenga, and R. N. Ndung'u, "Systematic Review of Models Usedto Handle Class Imbalance in Anomaly Detection for Energy Consumption," *Int. J. Artif. Intell. Appl.*, vol. 15, no. 3, pp. 41–52, 2024, doi: 10.5121/ijaia.2024.15304.

[12] N. A. S. A. Sabri, H. Hamed, M. A. M. Isa, N. S. Ghazali, and Z. Ibrahim, "Low-Density Polyethylene (LDPE) Food Packaging Defect Classification Using Local Binary Pattern (LBP)," *J. Phys. Conf. Ser.*, vol. 2129, no. 1, p. 12052, 2021, doi: 10.1088/1742-6596/2129/1/012052.