Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

PENERAPAN TEKNIK IMPUTASI K-MEANS TERHADAP PERFORMA HASIL KLASIFIKASI ALGORITMA NAÏVE BAYES

Ahmad Khusaeri

Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
Jl. HS Ronggowaluyo Karawang, Indonesia

*e-mail: ahmad.khusaeri@fasilkom.unsika.ac.id

Abstrak

Data merupakan salah satu komponen terpenting dalam melakukan sebuah penelitian. Ketersediaan data dapat memudahkan penelitian yang akan dilakukan. Dalam penelitian di berbagai bidang membutuhkan data yang lengkap. Namun kenyataannya adalah selalu ada beberapa komponen data yang tidak lengkap atau dikenal dengan istilah Missing Value. Penyebab terjadinya Missing Value karena informasi tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada. Salah satu proses yang digunakan dalam menentukan serta menetapkan nilai dalam mengganti Missing Value disebut dengan teknik imputasi. Pada Option Test dengan menggunakan k-fold cross validation dengan fold sebesar 10 menghasilkan nilai akurasi tertinggi adalah dengan melakukan penanganan Missing Value dengan menghapus data sebesar 0,985 dengan Missing Value sebesar 10%. Dari total data 136, 2 data salah diprediksi dan 134 data berhasil diprediksi dengan benar. Dari ketiga metode, nilai akurasi paling tinggi sebesar 0,985 dengan penanganan Missing Value dilakukan dengan menghapus data dengan tingkat Missing Value sebesar 10%. Adapun presisi dan Recall sebesar 0,984 dan 0,985. Sedangkan dengan Option Test percetage split menghasilkan pengolahan data dengan penanganan Missing Value dengan menghapus data menghasilkan nilai akurasi tertinggi sebesar 1 dengan nilai Recall dan presisi pun sebesar 1. Dari 44 data, semua data berhasil diprediksi dengan benar. Dari beberapa hasil pengolahan data dari data hasil imputasi menunjukkan bahwa nilai akurasi tertinggi berada pada data hasil penganan Missing Value dengan cara menghapus data.

Kata kunci: Imputasi, K-Means, Mean, Missing Value, Naïve Bayes.

Abstract

Data is one of the most important components in conducting a study. The availability of data can facilitate the research that will be conducted. In research in various fields, it requires complete data. But the reality is that there are always some incomplete data components or known as Missing Values. The cause of Missing Value is because information about objects is not given, is difficult to find, or indeed the information does not exist. One of the processes used in determining and determining the value of replacing Missing Value is called the imputation technique. In the Option Test using k-fold cross validation with a fold of 10, the highest accuracy value is done by handling the Missing Value by deleting data of 0.985 with a Missing Value of 10%. From 136 total data, 2 data were incorrectly predicted and 134 data were correctly predicted. Of the three methods, the highest accuracy value of 0.985 with the handling of the Missing Value is done by deleting data with a level of Missing Value of 10%. The precision and Recall are 0.984 and 0.985. Whereas the percetage split Option Test produces data processing by handling Missing Value by deleting data resulting in the highest accuracy value of 1 with even Recall value and precision of 1. From 44 data, all data were successfully predicted correctly. From the results of processing data from imputation data, it shows that the highest accuracy value is in the data value of Missing Value by deleting data.

Keywords: *Imputation, K-Means, Mean, Missing Value, Naïve Bayes.*

Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

1 Pendahuluan (or Introduction)

Data merupakan salah satu komponen terpenting dalam melakukan sebuah penelitian (Giriputra, 2016). Ketersediaan data dapat memudahkan penelitian yang akan dilakukan. Dalam penelitian di berbagai bidang membutuhkan data yang lengkap. Data yang sesuai dengan kebutuhan penelitian akan memudahkan dalam proses analisis data, serta memiliki ketepatan di dalam hasil pengolahan data. Namun kenyataannya adalah selalu ada beberapa komponen data yang tidak lengkap atau dikenal dengan istilah Missing Value. Penyebab terjadinya Missing Value karena informasi tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada [1]. Selain itu, Missing Value juga dapat disebabkan oleh berapa hal lain yaitu responden yang tidak memberikan jawaban karena informasi bersifat rahasia, kesalahan pada pengumpulan data seperti pertanyaan yang terlewat dan kesalahan memasukkan data[2]. Dampak dari hilangnya beberapa bagian data adalah akan mempengaruhi tingkat akurasi dari pemrosesan data [3]. Selain itu, Missing Value dapat menyebabkan pendugaan parameter menjadi tidak efisien karena berkurangnya ukuran data [4]. Metode untuk menangani Missing Value telah banyak yang dikembangkan. Salah satu proses yang digunakan dalam menentukan serta menetapkan nilai dalam mengganti Missing Value disebut dengan teknik imputasi.

K-Means (KM) adalah salah satu algoritma pengelompokan data yang dapat digunakan untuk melakukan imputasi pada Missing Value. Algoritma KM mengelompokkan data (klasterisasi) berdasarkan titik pusat klaster (Centroid). Algoritma K-Means merupakan metode pengelompokkan data non hirarki, jumlah kelompok yang akan dibentuk sudah ditentukan dan diketahui terlebih dahulu jumlahnya. Tujuan dari algoritma K-Means adalah mengelompokkan data sesuai dengan kemiripan data dalam satu klaster dan meminimalisasi kemiripan data antar klaster [5]. Data mining merupakan sebuah proses ekstraksi data untuk mendapatkan sebuah informasi yang belum diketahui sebelumnya pada suatu data [6]. Data mining juga merupakan bagian dari Knowledge Discovery in Database (KDD) yang merupakan proses ekstraksi informasi yang bermanfaat, tidak diketahui informasi sebelumnya dan tersembunyi.

Naïve Bayes adalah algoritma klasifikasi probabilitas sederhana yang berdasarkan pada teorema Bayes, asumsi bebas yang kuat. Naïve Bayes juga merupakan algoritma klasifikasi yang utama pada data mining dan banyak diterapkan dalam masalah klasifikasi dikehidupan sehari-hari karena memiliki performa klasifikasi yang tinggi. Algoritma Naïve Bayes juga memiliki beberapa keunggulan seperti mudah serta biaya perhitungan kecil, dapat menangani data missing, memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar[7].

Dari beberapa penelitian sebelumnya, maka penelitian ini akan membahas terkait penerapan teknik imputasi terhadap algoritma. Adapun teknik imputasi yang akan diterapkan adalah dengan menggunakan algoritma K-Means. Algoritma K-Means akan menangani Missing Value dengan level 10%, 20% dan 30%. Sedangkan algoritma yang digunakan dalam proses data mining adalah dengan menggunakan algoritma Naïve Bayes. Jadi penelitian ini akan membahas mengenai Penerapan Teknik Imputasi K-Means Terhadap Performa Hasil Klasifikasi Algoritma Naïve Bayes.

2 Tinjauan Literatur (or Literature Review)

Missing Value adalah suatu kondisi dimana data tidak ada atau data hilang. Terdapat tiga tipe Missing Value berdasarkan mekanisme yaitu *Missing Completely at Random (MCAR)*, *Missing at Random (MAR)*, *Not Missing at Random (NMAR)* [8]. Imputasi adalah proses yang digunakan untuk menentukan dan menetapkan nilai pengganti untuk *Missing Value*. Metode imputasi menjadi penting dalam situasi dimana dataset lengkap dibutuhkan untuk analisis. Berikut ini adalah Ada beberapa metode imputasi yaitu imputasi secara manual, imputasi dengan konstanta global, imputasi dengan metode konvensional dan imputasi dengan suatu model prediksi. Metode *K-Means* adalah salah satu metode dalam fungsi *Clustering* atau pengelompokan. Algoritma *K-Means* merupakan algoritma klasterisasi yang mengelompokkan data berdasarkan titik pusat klaster (*Centroid*) terdekat dengan data. Tujuan dari *K-Means* adalah pengelompokkan data dengan memaksimalkan kemiripan data dalam satu klaster dan meminimalkan kemiripan data antar klister [9]. *Elbow method* adalah metode yang

Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51
P-ISSN: 2746-5985
e-ISSN: 2797-0930

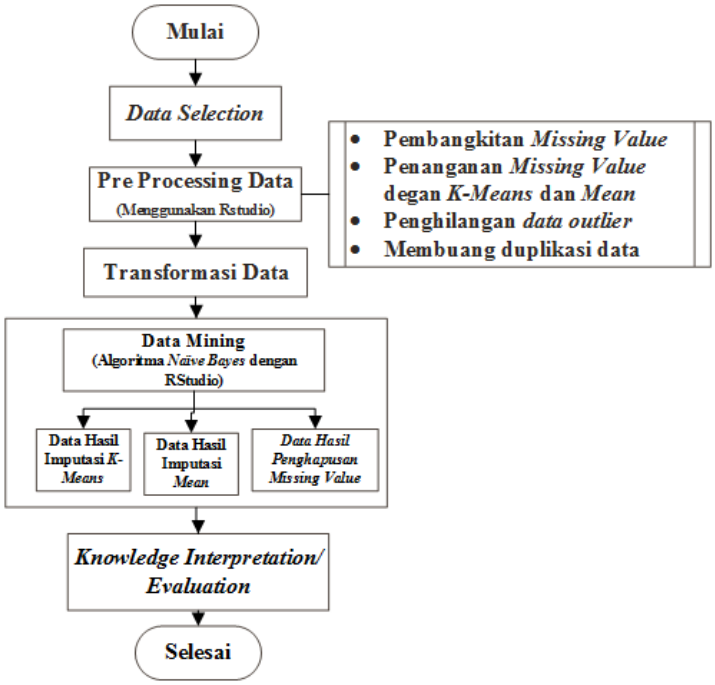
DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

digunakan untuk menentukan jumlah cluster yang tepat pada sebuah dataset[10]. Metode *elbow* digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik. Tahapan algoritma metode *Elbow* dalam menentukan nilai k pada *K-Means* adalah menginisialisasi awal nilai k, menaikan nilai k, menghitung hasil sum of square error dari tiap nilai k, analisis hasil sum of square error dari nilai k yang mengalami penurunan secara drastic, cari dan tetapkan nilai k yang berbentuk siku[11].

Teorema *bayes* adalah perhitungan statistik dengan menghitung probabilitas kemiripan kasus lama yang ada dibasis kasus dengan kasus baru. Teorema *bayes* memiliki tingkat akurasi yang tinggi dan kecepatan yang baik ketika diterapkan pada *database* yang besar[12]. Keuntungan penggunaan *Naive Bayes* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian[13]. *Confusion matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Evaluasi dengan *confusion matrix* menghasilkan nilai akurasi, presisi dan *Recall*. Akurasi dalam klasifikasi adalah persentase ketepatan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi[14]. Presisi atau *confidence* adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. *Recall* atau sensitivitas adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar[15].

3 Metode Penelitian (or Research Method)

Penelitian yang dilakukan menggunakan metodologi penelitian *Knowledge Discovery in Database* (KDD). Tahapan dari metodologi penelitian menggunakan KDD adalah *data selection, preprocessing data, transformation data, data mining* dan *knowledge interpretation/evaluation*.



Gambar 1 Tahapan Penelitian

Berikut ini adalah tahapan penelitian yang dilakukan adalah sebagai berikut:

a. Data Selection
   Penelitian ini menggunaan dataset iris data yang berasal dari *UCI Machine Learning.* Adapun atribut yang digunakan adalah *Sepal Length*, *Sepal Width*, *Petal Length* dan *Petal Width.* Iris dataset memiliki jumlah data sebesar 150 data.

Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51
P-ISSN: 2746-5985
e-ISSN: 2797-0930
DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

b. Pre-Processing Data
Pada proses *Pre-Processing Data* dilakukan beberapa hal untuk persiapan data. Sebelum melakukan penanganan *Missing Value*, yang pertama dilakukan adalah melakukan pembangkitan *Missing Value* sebesar 10%, 20% dan 30%. Adapun kegiatan selanjutnya yang dilakukan saat *pre-processing data* adalah melakukan penanganan *Missing Value* sebesar 10%, 20% dan 30% menggunakan teknik imputasi *K-Means* dan *Mean* dengan *tools* RStudio. Setelah itu melakukan pengecekan dan penanganan terhadap *data Outlier*. Melakukan pengecekan dan penanganan terhadap duplikasi data, standarisasi data dan reduksi data.

c. Data Transformation
Pada tahap ini dilakukan perubahan tipe data untuk memudahkan saat proses pengolahan data. Perubahan tipe data yang dilakukan adalah dengan merubah data numerik menjadi data kategorikal sesuai kebutuhan saat pengolahan data dengan menggunakan algoritma *Naïve Bayes*.

d. Data Mining
Pada tahap ini, dilakukan proses pengolahan data menggunakan algoritma klasifikasi yaitu algoritma *Naïve Bayes*. Data yang diolah merupakan data hasil teknik imputasi dengan *K-Means*, *Means* dan *Missing Value* dihapus. Pengolahan data dilakukan menggunakan *tools* Rstudio.

e. Knowledge Interpretation/ Evaluation
Pada tahap ini adalah hasil dari proses *mining data* diinterpretasikan dalam bentuk grafik. Adapun evaluasi dilakukan dengan menggunakan pengukuran tingkat akurasi, *precision* dan *Recall*.

4 Hasil dan Pembahasan (or Results and Analysis)

4.1 Data Selection

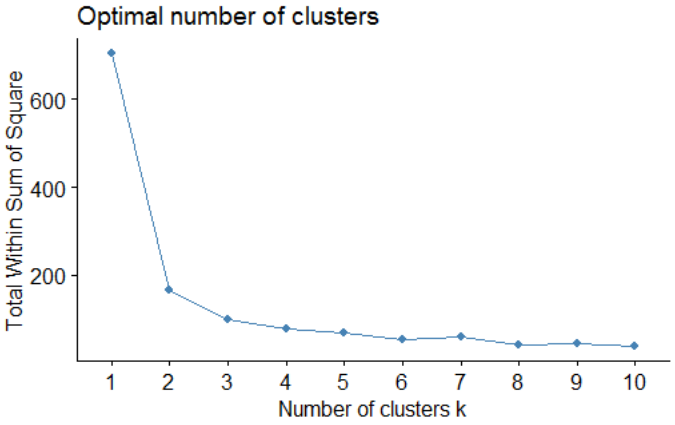
Data yang digunakan dalam penelitian ini adalah dengan menggunakan data iris yang bersumber pada *UCI Machine Learning*. Data iris memiliki 150 data yang terdiri dari 4 atribut. Atribut yang digunakan dalam data iris adalah *Sepal Length* (SL), *Sepal Width* (SW), *Petal Length* (PL) dan *Petal Width* (PW). Data awal yang digunakan merupakan dataset yang lengkap atau tidak memiliki *Missing Value*. Oleh karena itu, pada tahapan ini dilakukan proses pembangkitan *Missing Value* sebesar 10%, 20% dan 30%. Pembangkitan *Missing Value* dilakukan secara acak menggunakan *tools* kutools yang terdapat pada excel.

4.2 Preprocessing Data

Pada tahap ini dilakukan penanganan terhadap *mising value* di data iris. Penanganan *Missing Value* dilakukan dengan 3 penanganan. Penanganan *Missing Value* pertama dilakukan dengan menggunakan teknik imputasi *K-Means*. Penanganan *Missing Value* yang kedua dilakukan dengan menggunakan metode *Mean*. Penanganan yang ketiga adalah dengan menghapus data yang memiliki *missings value*.

Penanganan *Missing Value* diterapkan pada data yang memiliki *Missing Value* masing – masing sebesar 10%, 20% dan 30% dengan menggunakan 3 metode yaitu teknik imputasi *K-Means*, *Mean* dan penghapusan *Missing Value*. Tahapan pertama yang dilakukan dalam menangani *Missing Value* dengan *K-Means* adalah dengan melakukan penentuan banyaknya *cluster* (k). Penentuan banyaknya cluster dilakukan menggunakan metode *elbow* dengan menggunakan *tools RStudio*. Hasil yang diperoleh dengan menggunakan metode *elbow* menunjukkan bahwa K optimum yang digunakan adalah k=2. Adapun hasil dari penentuan k-optimum dengan menggunakan metode *elbow* sesuai dengan Gambar 2.

Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51
P-ISSN: 2746-5985
e-ISSN: 2797-0930
DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765



Gambar 2 Penentuan K-Optimum

Setelah menentukan jumlah *cluster* (k), yaitu menentukan *Centroid* awal (Co). Selanjutnya menghitung jarak masing – masing *Centroid* yang digunakan sebagai acuan dalam mengelompokkan setiap objek berdasarkan jarak yang terdekat. Selanjutnya adalah melakukan iterasi dengan menentukan *Centroid* selanjutnya hingga kelompok *cluster* tidak berubah.

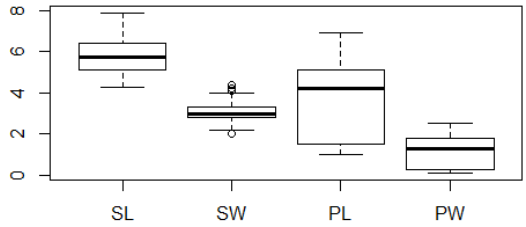
```
Cluster means:
        SL       SW       PL        PW
1 6.137500 2.741250 4.983750 1.5700000
2 4.548571 3.051429 1.371429 0.6028571
```

Gambar 3 Nilai Tengan Klaster

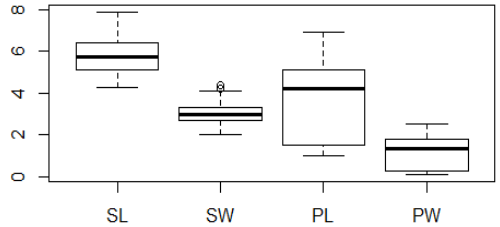
Setelah didapat *cluster* dari setiap data, selanjutnya masukkan nilai *Centroid* yang terakhir dari setiap variabel ke dalam data yang memiliki *Missing Value*. Nilai *Centroid* yang dimasukkan ke dalam data yang memiliki *Missing Value* , disesuaikan dengan nilai *cluster* data yang memiliki *Missing Value*. Setelah mengisi *Missing Value* dengan menggunakan *K-Means*, selanjutnya mengisi *Missing Value* dengan menggunakan rata – rata (*Mean*). Penanganan *Missing Value* dengan *Mean* adalah memasukkan rata-rata tiap atribut kedalam *Missing Value*.

Setelah melakukan penanganan *Missing Value* dengan *K-Means* dan *Mean*, penanganan selanjutnya adalah dengan menghapus baris data yang memiliki *Missing Value*. Setelah semua *Missing Value* ditangani dengan 3 metode yaitu *K-Means, Mean* dan penghapusan *Missing Value*, selanjutnya adalah melakukan pengecekan data *Outlier* dari setiap data yang telah mendapatkan penanganan *Missing Value* dengan 3 metode tersebut.

Pengecekan data *Outlier* pertama dilakukan pada data hasil imputasi dengan *K-Means*. Pada data hasil imputasi *K-Means* sebesar 10%, 20% dan 30% ditemukan secara berturut - turut 4 buah data, 2 buah data dan 3 buah data yang mengalami *Outlier*. Adapun visualisasi data *Outlier* dengan menggunakan *box plot* sesuai pada Gambar 4.8 sampai 4.10.

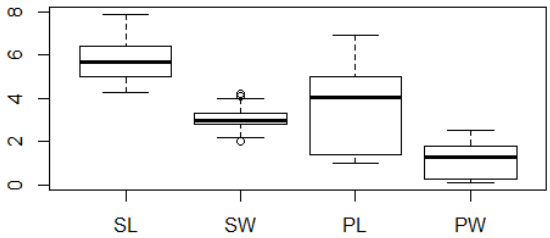


Gambar 4 Data Outlier Imputasi Missing Value dengan K-Means 10%



Gambar 5 Data Outlier Imputasi Missing Value dengan K-Means 20%

Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51
P-ISSN: 2746-5985
e-ISSN: 2797-0930
DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765



Gambar 6 Data Outlier Imputasi Missing Value dengan K-Means 30%

Selanjutnya adalah melakukan pengecekan data *Outlier* terhadap data hasil imputasi menggunakan *Mean* dan melakukan penghapusan *Missing Value*. Adapun jumlah data *Outlier* per dataset sesuai dengan tabel 4.11.

Tabel 1 Jumlah Data Outlier

Penanganan *Missing Value*	**Besaran *Missing Value***		
	***Missing Value* 10%**	***Missing Value* 20%**	***Missing Value* 30%**
K – *Means*	4 data	2 data	3 data
Mean	4 data	4 data	3 data
Data dihapus	4 data	12 data	2 data

Setelah diketahui jumlah *Outlier* dari setiap data, langkah selanjutnya adalah dengan melakukan penanganan data *Outlier* dengan cara menghapus data yang memiliki *Outlier*.

4.3 Transformasi Data

Pada tahap transformasi data dilakukan perubahan tipe data dari data numerik menjadi data kategorikal. Perubahan tipe data ini dilakukan untuk memudahkan dalam proses klasifikasi data. Perubahan tipe data dilakukan dengan menggunakan metode kuartil data. Adapun contoh perhitungan dengan metode kuartil dengan menggunakan data imputasi *K-Means* dengan *Missing Value* sebesar 10% dengan atribut *Sepal Length* dan data (n) sebanyak 150 data sebagai berikut.

Tabel 2 Perhitungan Kuartil Data Atribut SW Hasil Imputasi K-Means 10%

Kuartil 1	**Kuartil 2**	**Kuartil 3**
K1 = ¼ (n+1)	K2 = ½ (n+1)	K3 = ¾ (n+1)
K1 = 0,25 (150 + 1)	K2 = 0,5 (150 + 1)	K3 = 0,75 (150 + 1)
K1 = 0,25 (151)	K2 = 0,5 (151)	K3 = 0,75 (151)
K1 = 37,75	K2 = 75,5	K3 = 113,25

Setiap atribut diklasifikasikan menjadi 3 kategori. Hasil dari pengkategorian menggunakan kuartil data dapat dilihat pada tabel 4.16. Salah satu hasil dari transformasi data dengan menggunakan kuartil data dapat dilihat sesuai dengan tabel 4.17 dan 4.18.

Tabel 3 Kategori Iris Data

Imputasi	**MV**	**Kategori**	***Sepal Length***	***Petal Length***	**Kategori**	***Sepal Width***	***Petal Width***
		Pendek	< 5,1	< 1,5	**Rendah**	< 2,8	< 0,3
	10%	**Sedang**	5,1 – 6,4	1,5 – 5,1	**Sedang**	2,8 – 3,3	0,3 – 1,8
		Tinggi	> 6,4	> 5,1	**Tinggi**	> 3,3	> 1,8
K-Means		Pendek	< 5,1	< 1,5	Rendah	< 2,7	< 0,3
	20%	Sedang	5,1 – 6,4	1,5 – 5,1	Sedang	2,7 – 3,3	0,3 – 1,8
		Tinggi	> 6,4	> 5,1	Tinggi	> 3,3	> 1,8
	30%	**Pendek**	< 5,1	< 1,5	**Rendah**	< 2,8	< 0,3

Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51
P-ISSN: 2746-5985
e-ISSN: 2797-0930
DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

		Sedang	5,1 – 6,4	1,5 – 5,0	**Sedang**	2,8 – 3,2	0,3 – 1,8
		Tinggi	> 6,4	> 5,0	**Tinggi**	> 3,2	> 1,8
	10%	Pendek	< 5,1	< 1,6	Rendah	< 2,8	< 0,3
Mean	**20%**	Sedang	5,1 – 6,4	1,6 – 5,1	Sedang	2,8 – 3,3	0,3 – 1,8
	30%	Tinggi	> 6,4	> 5,1	Tinggi	> 3,3	> 1,8
Hapus	**10%**	**Pendek**	< 5,1	< 1,5	**Rendah**	< 2,8	< 0,3
Missing	**20%**	**Sedang**	5,1 – 6,4	1,5 – 5,1	**Sedang**	2,8 – 3,3	0,3 – 1,8
Value	**30%**	**Tinggi**	> 6,4	> 5,1	**Tinggi**	> 3,3	> 1,8

Tabel 4 Data Numerik

No	*Sepal Length*	*Sepal Width*	*Petal Length*	*Petal Width*	Class
1	5,1	3,5	1,4	0,2	Iris-setosa
2	4,9	3,0	1,4	0,2	Iris-setosa
3	4,7	3,2	1,3	0,2	Iris-setosa
4	4,6	3,1	1,5	0,2	Iris-setosa
5	5,9	3,6	1,4	0,2	Iris-setosa

Tabel 5 Data Kategorikal

No	*Sepal Length*	*Sepal Width*	*Petal Length*	*Petal Width*	Class
1	Sedang	Tinggi	Pendek	Rendah	Iris-setosa
2	Pendek	Sedang	Pendek	Rendah	Iris-setosa
3	Pendek	Sedang	Pendek	Rendah	Iris-setosa
4	Pendek	Sedang	Pendek	Rendah	Iris-setosa
5	Sedang	Tinggi	Pendek	Rendah	Iris-setosa

4.4 Data Mining

Pada tahapan ini dilakukan penambangan data. *Data mining* dilakukan pada dataset yang telah melalui tahapan *preprocessing data* dan transformasi data. Dataset yang akan diolah berjumlah 9 dataset dengan rincian sesuai dengan tabel 4.19.

Tabel 6 Banyaknya Dataset

Penanganan *Missing Value*	**Besaran** *Missing Value*		
	***Missing Value* 10%**	***Missing Value* 20%**	***Missing Value* 30%**
K – *Means*	1 dataset	1 dataset	1 dataset
Mean	1 dataset	1 dataset	1 dataset
Data dihapus	1 dataset	1 dataset	1 dataset

Pada tahapan *data mining*, sebelum melakukan pengolahan data terlebih dilakukan pembagian dataset dengan menggunakan beberapa metode. Metode yang dilakukan dalam *splitting data* adalah dengan menggunakan *k-fold cross validation* dan *Percentage Split*. Metode *k-fold cross validation* akan menggunakan *fold* sebesar 10 *folds*. Sedangakan metode percantage split akan melakukan pembagian dataset sebesar 60%, 70% dan 80% . Setelah dilakukan *splitting data*, maka tahapan selanjutnya adalah proses *data mining* dengan menggunakan algoritma *naïve bayes*.

Tahapan yang pertama adalah dengan melakukan *data mining* pada dataset iris data hasil teknik imputasi menggunakan algoritma *K-Means* dengan besaran *Missing Value* sebesar 10% 20% dan 30% menggunakan *Option Test k-fold cross validation.* Pengolahan data menggunakan algoritma *naïve bayes* dengan *Option Test k-fold cross validation* menghasilkan hasil yang beraneka ragam. *Option Test* dengan *k-fold cross validation* menggunakan *fold* sebesar 10.

Hasil tersebut menunjukkan bahwa tingkat akurasi dari penanganan *Missing Value* dengan menggunakan metode *K-Means*, memiliki nilai akurasi tertinggi sebesar 0,973 dengan nilai presisi sebesar 0,975 dan nilai *Recall* sebesar 0,972. Data yang memiliki tingkat akurasi paling tinggi di data

Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51
P-ISSN: 2746-5985
e-ISSN: 2797-0930
DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

hasil penganangan *Missing Value* dengan menggunakan *K-Means* memiliki *Missing Value* sebesar 20%. Total data yang berhasil diprediksi sebanyak 148 data. Adapun data yang berhasil diprediksi dengan benar yaitu sebanyak 144 data sedangkan 4 data lainnya salah diprediksi

Hasil dari penanganan *Missing Value* dengan menggunakan metode *Mean*, memiliki nilai akurasi yang paling tinggi sebesar 0,973 dengan presisi sebesar 0,975 dan *Recall* sebesar 0,973. Total data yang berhasil diprediksi sebanyak 150 data, 146 data berhasil diprediksi dengan benar sedangkan 4 data salah diprediksi. Sedangkan penanganan *Missing Value* dengan menghapus data memiliki nilai akurasi paling tinggi sebesar 0,985 dengan *Missing Value* sebesar 10%. Total data keseluruhan yang berhasil diprediksi sebanyak 136 data, data yang salah diprediksi sebanyak 2 data dan 134 data berhasil diprediksi dengan benar.

Dari ketiga metode yang dilakukan dalam proses penanganan *Missing Value*, nilai akurasi yang paling tinggi saat pengolahan data dengan *spliting data* menggunakan *k-fold cross validation* terdapat pada data hasil penanganan *Missing Value* dengan cara menghapus data. Nilai akurasi tertinggi pada *k-fold cross validation* sebesar 0,985 dengan tingkat *Missing Value* sebesar 10%. Adapun presisi dan *Recall* sebesar 0,984 dan 0,985. Setelah melakukan pengolahan data dengan menggunakan *Option Test k-fold cross validation*, selanjutnya melakukan pengolahan data dengan menggunakan *Option Test Percentage Split*. Besaran *Percentage Split* yang dilakukan sebesar 60%, 70% dan 80%.

Hasil dari probabilitas atribut *Petal Width* menunjukkan bahwa kemungkinan terbesar *class* atribut PW yang keluar sebesar 1 dengan *class target iris-versicolor* dengan kategori sedang. Hasil dari pengolahan data menggunakan *Option Test Percentage Split* menunjukkan data yang tidak jauh berbeda. Pengolahan data hasil teknik imputasi *K-Means* menunjukkan bahwa nilai akurasi yang tertinggi pada data imputasi *K-Means* terdapat pada data yang memiliki *Missing Value* sebesar 30% dan *Percentage Split* sebesar 70% dengan nilai akurasi sebesar 0,977, presisi sebesar 0,972 dan *Recall* sebesar 0,981. Pada hasil tersebut menunjukkan bahwa dari 44 data, 43 berhasil diprediksi dengan benar dan 1 data diprediksi salah.

Hasil pengolahan data imputasi dengan *Mean* menunjukkan bahwa nilai akurasi tertinggi sebesar 0,983. Adapun nilai presisi dan *Recall* sebesar 0,987 dan 0,981. Nilai akurasi tebesar terdapat pada *Percentage Split* sebesar 60% dengan data yang memiliki *Missing Value* sebesar 10% , 20% dan 30%. Hasil tersebut menunjukkan bahwa dari total 60 data, 59 data berhasil diprediksi dengan benar sedangkan 1 data diprediksi dengan salah.

Hasil pengolahan data dengan penanganan *Missing Value* dengan menghapus data menghasilkan nilai akurasi tertinggi sebesar 1 dengan nilai *Recall* dan presisi pun sebesar 1. Dari 44 data, semua data berhasil diprediksi dengan benar. Dari beberapa hasil pengolahan data dari data hasil imputasi menunjukkan bahwa nilai akurasi tertinggi berada pada data hasil penganan *Missing Value* dengan cara menghapus data.

4.5 Knowledge Interpretation/Evaluation

Hasil dari pengolahan data dengan teknik imputasi *K-Means* , *Mean* dan penghapusan *Missing Value* dengan pengolahan data menggunakan algortima *naïve bayes* dengan *Option Test k-fold cross validation* dengan 10 *fold* dan *Option Test Percentage Split* dengan nilai *percentage* sebesar 60%, 70% dan 80% divisualisasikan dengan menggunakan grafik. Hasil pengolahan data menggunakan *Option Test k-fold cross validation* dengan *fold* sebesar 10 sesuai pada tabel 4.20.

Tabel 7 Hasil Naive Bayes dengan K-Fold Cross Validation

Penanganan *Missing Value*	**Besaran *Missing Value***	**Akurasi**	***Confussion Matrix*** **Rata - Rata Presisi**	**Rata – Rata *Recall***
K-Means	10%	0,897	0,974	0,972
	20%	0,973	0,975	0,972
	30%	0,965	0,969	0,965
Mean	10%	0,973	0,975	0,973
	20%	0,973	0,975	0,973

Jurnal Manajemen Informatika Jayakarta
Volume 5, Nomor 1,Februari 2025: halaman 41-51
P-ISSN: 2746-5985
e-ISSN: 2797-0930
DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

	30%	0,973	0,975	0,973
	10%	0,985	0,984	0,985
Data dihapus	20%	0,950	0,954	0,951
	30%	0,954	0,954	0,956

Hasil dari pengolahan data menggunakan *Naive Bayes* dengan menggunakan *Option Test k-fold cross validation* menghasilkan nilai akurasi, presisi dan *Recall* tertinggi pada penanganan *Missing Value* menggunakan penghapusan data dengan besaran *Missing Value* sebesar 10% dengan akurasi 0,985, presisi 0,984 dan *Recall* 0,985, sedangkan dengan penanganan *Missing Value* dengan cara *K-Means* dan *Mean* cenderung stabil dengan nilai berkisar 0,9.

Hasil grafik tersebut menunjukkan bahwa nilai *Recall* tertinggi berada pada metode penghapusan *Missing Value* dengan besaran *Missing Value* sebesar 10% dengan nilai *Recall* sebesar 98,4 %. Selanjutnya adalah grafik pengolahan data dengan *Option Test Percentage Split*.

Tabel 8 Hasil Pengolahan Naive Bayes dengan Option Test Percentage Split

CM	*Percentage Split*	*K-Means*			*Mean*			Data Dihapus		
		10%	20%	30%	10%	20%	30%	10%	20%	30%
Akurasi	**60%**	0,966	0,966	0,965	0,983	0,983	0,983	1	1	1
	70%	0,95	0,955	0,977	0,977	0,977	0,977	1	1	1
	80%	0,96	0,966	0,965	0,966	0,966	0,966	1	1	1
Avg. Presisi	**60%**	0,969	0,972	0,962	0,987	0,987	0,987	1	1	1
	70%	0,95	0,966	0,972	0,981	0,981	0,981	1	1	1
	80%	0,97	0,972	0,969	0,966	0,966	0,966	1	1	1
Avg. *Recall*	**60%**	0,965	0,964	0,968	0,981	0,981	0,981	1	1	1
	70%	0,95	0,950	0,981	0,979	0,979	0,979	1	1	1
	80%	0,96	0,969	0,969	0,974	0,974	0,974	1	1	1

Hasil dari pengolahan data mengsgunakan algoritma *Naive Bayes* dengan menggunakan *Option Test Percentage Split* menghasilkan nilai akurasi terbaik sebesar 1 dengan penanganan *Missing Value* menggunakan cara menghapus data. Adapun data yang digunakan adalah dengan menggunakan data dengan tingkat *Missing Value* sebesar 10%, 20% dan 30%. Sedangkan hasil yang lainnya cenderung stabil dengan besaran nilai kisaran 0,9.

5 Kesimpulan (or Conclusion)

Berdasarkan penelitian yang telah dilaksanakan, dapat diambil beberapa kesimpulan sebagai berikut:

1. Penanganan *Missing Value* dapat dilakukan dengan melakukan metode penanganan dengan cara teknik imputasi. Dalam hal ini dilakukan 3 metode yang dilakukan dalam proses penanganan *Missing Value*. Metode pertama adalah dengan menggunakan algoritma *K-Means*. Algoritma *K-Means* menangani *Missing Value* dengan cara membuat kluster berdasarkan banyak nya klaster optimum yang telah dicek terlebih dahulu. Setelah itu, nilai yang memiliki *Missing Value* diisi dengan *Centroid* terakhir hasil klaster. Metode kedua adalah dengan mengisi berdasarkan rata-rata setiap atribut. Metode yang ketiga adalah dengan melakukan proses penghapusan data yang memiliki *Missing Value*. Ketiga metode ini dapat diterapkan pada 3 jenis data yang memiliki masing – masing nilai *Missing Value* sebesar 10%, 20% dan 30%.
2. Penanganan *Missing Value* dengan 3 metode memiliki dampak terhadap pengolahan data dengan menggunakan algoritma *naïve bayes*. Hal ini dapat terlihat dari hasil akurasi setiap data dari pengolahan data yang menggunakan algoritma *naïve bayes* dengan metode yang berbeda. Akurasi dari setiap data melebihi nilai 90% hasil dari pengolahan data menggunakan data hasil teknik imputasi dan menggunakan beberapa *Option Test*. Teknik imputasi *K-Means* memiliki nilai akurasi tertinggi dengan menggunakan *Option Test k-fold cross validation* sebesar 0,973 pada data yang memiliki *Missing Value* sebesar 20% dan 0,977 dengan menggunakan *Option Test Percentage*

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

Split dengan *Missing Value* sebesar 30% dengan besaran *Percentage Split* 70%. Sedangkan pada *Option Test* dengan menggunakan *K-Fold Cross Validation* dengan *fold* sebesar 10 menghasilkan nilai akurasi tertinggi adalah dengan melakukan penanganan *Missing Value* dengan menghapus data sebesar 0,985 dengan *Missing Value* sebesar 10%. Dari total data 136, 2 data salah diprediksi dan 134 data berhasil diprediksi dengan benar. Dari ketiga metode, nilai akurasi paling tinggi sebesar 0,985 dengan penanganan *Missing Value* dilakukan dengan menghapus data dengan tingkat *Missing Value* sebesar 10%. Adapun presisi dan *Recall* sebesar 0,984 dan 0,985. Sedangkan dengan *Option Test percetage split* menghasilkan pengolahan data dengan penanganan *Missing Value* dengan menghapus data menghasilkan nilai akurasi tertinggi sebesar 1 dengan nilai *Recall* dan presisi pun sebesar 1. Dari 44 data, semua data berhasil diprediksi dengan benar. Dari beberapa hasil pengolahan data dari data hasil imputasi menunjukkan bahwa nilai akurasi tertinggi berada pada data hasil penganan *Missing Value* dengan cara menghapus data.

Referensi (Reference)

[1] Miraati Laia, "Analisis Kinerja Algoritma K-Nearest Neighbor Imputation (KNNI) Untuk Missing Value Pada Klasifikasi Data Mining," *J. Informatics, Electr. Electron. Eng.*, vol. 2, no. 3, pp. 92–98, 2023, doi: 10.47065/jieee.v2i3.891.

[2] A. S. Arifianto, K. Dewi Safitri, K. Agustianto, and I. G. Wiryawan, "Pengaruh Prediksi Missing Value pada Klasifikasi Decision Tree C4.5," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 4, pp. 779–786, 2022, doi: 10.25126/jtiik.2022944778.

[3] A. R. Aziz, B. Warsito, and A. Prahutama, "Pengaruh Transformasi Data Pada Metode Learning Vector Quantization Terhadap Akurasi Klasifikasi Diagnosis Penyakit Jantung," *J. Gaussian*, vol. 10, no. 1, pp. 21–30, 2021, doi: 10.14710/j.gauss.v10i1.30933.

[4] I. Eldiyana, E. Nurlaelah, and N. Herrhyanto, "Estimasi Missing Data dengan Metode Multivariate Imputation by Chained Equations (MICE) untuk Membentuk Persamaan Regresi Linier Berganda," *J. EurekaMatika*, vol. 9, no. 1, pp. 95–106, 2021, doi: 10.17509/jem.v8i1.25750.

[5] D. Indriani, B. Irawan, and A. Bahtiar, "Penerapan Algoritma K-Means Clustering Untuk Menentukan Persediaan Stok Barang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 182–187, 2024, doi: 10.36040/jati.v8i1.8322.

[6] S. Darma, Y. Yusman, and J. Hendrawan, "Analisis Data Tingkat Kehadiran Pegawai dengan Menggunakan Clustering K-Means Pada Dinas Pekerjaan Umum dan Penataan Ruang Kabupaten Langkat," *J. Minfo Polgan*, vol. 13, no. 1, pp. 1105–1116, 2024, doi: 10.33395/jmp.v13i1.13958.

[7] N. Widiastuti, A. Hermawan, and D. Avianto, "Implementasi Metode Naïve Bayes Untuk Klasifikasi Data Blogger," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, no. 3, pp. 985–994, 2023, doi: 10.29100/jipi.v8i3.3713.

[8] 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.

[9] Y. Febriani, Y. P. Sari, and D. Octaria, "Metode K-Means Cluster Untuk Mengelompokkan Kota/Kabupaten di Sumatera Selatan Berdasarkan Produksi Ikan Air Tawar," *Sainmatika J. Ilm. Mat. dan Ilmu Pengetah. Alam*, vol. 18, no. 2, p. 175, 2021, doi: 10.31851/sainmatika.v18i2.6722.

[10] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 14, no. 2, pp. 277–288, 2023, doi: 10.24176/simet.v14i2.9630.

[11] M. Guntara and N. Lutfi, "Optimasi Cacah Klaster pada Klasterisasi dengan Algoritma KMeans Menggunakan Silhouette Coeficient dan Elbow Method," *JuTI "Jurnal Teknol. Informasi,"* vol. 2, no. 1, p. 43, 2023, doi: 10.26798/juti.v2i1.944.

[12] R. R. Adhitya, Wina Witanti, and Rezki Yuniarti, "Perbandingan Metode Cart Dan Naïve Bayes

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v5i1.1765

Untuk Klasifikasi Customer Churn," *INFOTECH J.*, vol. 9, no. 2, pp. 307–318, 2023, doi: 10.31949/infotech.v9i2.5641.

[13] A. H. Hasugian, R. A. Putri, and M. A. Simatupang, "Penerapan Algoritma Klasifikasi Naïve Bayes Untuk Analisis Sentimen Tentang Pemindahan Ibu Kota Negara," *J. Sci. Soc. Res.*, vol. 4307, no. 2, pp. 635–644, 2024, [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR

[14] M. F. Rizalno, A. Johar, and F. F. Coastera, "Analisis Prediksi Masa Studi Mahasiswa Menggunakan Metode Decision Tree Dengan Penerapan Algoritme Cart (Classification and Regression Trees) (Studi Kasus Data Alumni Fakultas Teknik Universitas Bengkulu)," *Rekursif J. Inform.*, vol. 10, no. 1, pp. 96–106, 2022, doi: 10.33369/rekursif.v10i1.21362.

[15] I Putu Agus Aryawan, I Nyoman Purnama, and Ketut Queena Fredlina, "Analisis Perbandingan Algoritma Cnn Dan Svm Pada Klasifikasi Ekspresi Wajah," *J. Teknol. Inf. dan Komput.*, vol. 9, no. 4, pp. 399–408, 2023, doi: 10.36002/jutik.v9i4.2545.