Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru

Anak Agung Aryasatya Daniswara[1], I Kadek Dwi Nuryana[2]

[1,2] Jurusan Teknik Informatika Fakultas Teknik Universitas Negeri Surabaya
[1]anak.19050@mhs.unesa.ac.id
[2]dwinuryana@unesa.ac.id

***Abstrak*— Dalam era digital, jumlah data di bidang pendidikan semakin kompleks. Preprocessing data penting dalam analisis data mining untuk membersihkan, mengubah format, dan mempersiapkan data agar lebih mudah dan akurat. Artikel ini menjelaskan tahap preprocessing data untuk data penilaian per kelas per mahasiswa. Tahapannya mencakup pembersihan data, transformasi data, seleksi fitur, encoding variabel kategorikal, dan pengurangan dimensi. Setelah preprocessing, data siap digunakan dalam analisis data mining seperti clustering, klasifikasi, dan prediksi. Implementasi preprocessing data berdampak positif pada kualitas pendidikan dan pengambilan keputusan yang lebih cerdas. Studi kasus digunakan untuk analisis data penilaian per kelas per mahasiswa dengan teknik data mining dan preprocessing data. Hasilnya memberikan wawasan bermanfaat untuk bidang pendidikan.**

***Kata Kunci*— Data mining, preprocessing data, pendidikan, pembersihan data, transformasi data.**

I. PENDAHULUAN

Dalam era digital, jumlah data yang dihasilkan semakin besar dan kompleks, terutama dalam konteks pendidikan. Data-data tersebut mencakup penilaian, hasil ujian, dan informasi lain yang dapat memberikan wawasan penting bagi pengambilan keputusan di bidang pendidikan. Namun, sebelum data tersebut dapat diolah lebih lanjut dan digunakan untuk keperluan analisis atau pemodelan, perlu dilakukan tahap preprocessing data.

Preprocessing data merupakan proses penting dalam analisis data mining yang bertujuan untuk membersihkan[1], mengubah format, dan mempersiapkan data agar lebih mudah dan akurat dalam proses analisis. Dalam artikel ini, akan dijelaskan mengenai tahapan preprocessing data untuk data penilaian per kelas per mahasiswa dalam konteks pendidikan. Data tersebut mencakup berbagai fitur seperti email peserta, peran peserta, kategori Diklat PPG, nilai penilaian, dan lainnya.

Tahapan preprocessing data yang akan dijelaskan meliputi:

1. Pembersihan Data: Tahap ini melibatkan identifikasi dan penanganan data yang hilang (null), outlier, atau tidak valid[3]. Data yang hilang dapat dihapus atau diisi dengan nilai yang sesuai. Outlier dapat dihapus atau ditangani dengan metode statistik seperti trimming atau winsorizing. Pada tahap ini, juga akan dilakukan penghapusan data yang memiliki duplikasi.
2. Transformasi Data: Beberapa atribut dalam dataset mungkin perlu diubah skala atau bentuk distribusinya untuk memenuhi persyaratan analisis atau pemodelan[3]. Normalisasi dan standarisasi adalah teknik yang umum digunakan untuk mengubah skala atribut. Selain itu, transformasi logaritmik dapat digunakan untuk mengubah distribusi atribut yang cenderung mengikuti pola eksponensial.
3. Seleksi Fitur: Tahap ini melibatkan identifikasi atribut yang paling informatif dan relevan untuk tujuan analisis[3]. Pilihan fitur yang baik dapat meningkatkan efisiensi dan kualitas model. Atribut yang tidak relevan atau memiliki korelasi yang tinggi dengan atribut lain dapat dihapus dari dataset.
4. Encoding Variabel Kategorikal: Jika dataset mengandung variabel kategorikal[4], perlu dilakukan encoding untuk mengubahnya menjadi bentuk numerik agar dapat digunakan dalam algoritma pembelajaran mesin. Teknik yang umum digunakan adalah one-hot encoding dan label encoding.
5. Pengurangan Dimensi: Tahap ini bertujuan untuk mengurangi kompleksitas data dengan mengurangi jumlah atribut atau fitur yang digunakan dalam analisis[4]. Teknik yang umum digunakan termasuk Analisis Komponen Utama (PCA) untuk data numerik dan Analisis Diskriminan Linier (LDA) untuk pemodelan kelas.

Setelah proses preprocessing data selesai dilakukan, data siap untuk digunakan dalam analisis data mining. Data tersebut dapat diaplikasikan pada teknik-teknik data mining seperti clustering, klasifikasi, atau prediksi untuk mendapatkan wawasan baru yang dapat membantu pengambilan keputusan di bidang pendidikan.

Melalui tahap preprocessing data yang sistematis dan teliti, data penilaian per kelas per mahasiswa dapat diolah menjadi informasi berharga yang dapat meningkatkan kualitas dan efektivitas proses pembelajaran, mendukung pengambilan keputusan, dan memberikan pemahaman lebih baik tentang performa mahasiswa dan program pendidikan secara keseluruhan[3]. Data mining menjadi alat penting dalam menganalisis data besar dan kompleks dalam pendidikan, dan preprocessing data merupakan fondasi yang kuat dalam menjalankan proses ini[1].

II. METODE PENELITIAN

Untuk mengimplementasikan dan menguji tahapan preprocessing data dan proses data mining pada penilaian per kelas per mahasiswa, metode penelitian yang cocok adalah Studi Kasus. Metode ini memungkinkan peneliti untuk menyelidiki fenomena atau proses tertentu secara mendalam dalam situasi dunia nyata. Dalam konteks ini, studi kasus akan

memungkinkan peneliti untuk menganalisis data penilaian per kelas per mahasiswa secara lengkap dan mendalam.

Langkah-langkah dalam metode studi kasus sebagai berikut:

1. Pengumpulan Data: Tahap pertama dalam metode penelitian ini adalah pengumpulan data[4]. Data penilaian per kelas per mahasiswa yang diperlukan untuk studi kasus ini dapat diperoleh dari pangkalan data PPG yang relevan. Data ini harus dikumpulkan dengan teliti dan harus mencakup semua atribut yang relevan yang telah dijelaskan sebelumnya, seperti email peserta, course, penilaian proses pm, penilaian produk pm, dan nilai pendalaman materi. Dalam proses pengumpulan data ini, pastikan bahwa dataset yang digunakan lengkap dan akurat untuk mendukung analisis yang valid dan representatif.
2. Identifikasi Tujuan Penelitian: Setelah data terkumpul, langkah selanjutnya adalah mengidentifikasi tujuan penelitian secara jelas dan spesifik[4]. Misalnya, tujuan penelitian dapat berfokus pada pengelompokan peserta berdasarkan penilaian mereka, atau memprediksi kinerja akademik berdasarkan variabel-variabel tertentu dalam dataset. Dengan menetapkan tujuan yang jelas, peneliti akan dapat mengarahkan proses analisis data dengan tepat dan menghasilkan temuan yang relevan dan bermanfaat.
3. Preprocessing Data: Setelah tujuan penelitian ditetapkan, proses preprocessing data dapat dilakukan. Tahapan preprocessing ini akan membersihkan, mentransformasi, dan mempersiapkan data agar siap digunakan dalam proses data mining[4]. Proses preprocessing ini melibatkan langkah-langkah seperti penanganan data yang hilang, penghapusan duplikasi, normalisasi data, dan pengubahan variabel kategorikal menjadi bentuk numerik yang dapat diolah dalam algoritma data mining. Proses preprocessing yang baik akan memastikan bahwa data siap untuk tahap selanjutnya, yaitu penerapan data mining.
4. Penerapan Data Mining: Setelah data telah melewati tahapan preprocessing, langkah selanjutnya adalah menerapkan teknik data mining pada dataset[4]. Teknik data mining yang dapat diaplikasikan termasuk k-means clustering untuk mengelompokkan peserta ke dalam kelompok yang serupa berdasarkan penilaian mereka, atau decision tree untuk memprediksi kinerja akademik berdasarkan atribut-atribut tertentu dalam dataset. Jika diperlukan, teknik lain seperti association rule mining atau neural networks juga dapat diuji. Penerapan data mining ini bertujuan untuk mengidentifikasi pola dan hubungan di antara variabel dalam dataset, sehingga dapat diperoleh wawasan dan informasi yang berharga.
5. Evaluasi dan Interpretasi Hasil: Setelah proses data mining selesai, hasilnya dievaluasi untuk mengukur kualitas dan relevansinya terhadap tujuan penelitian yang telah ditetapkan[4]. Evaluasi dilakukan dengan menggunakan metrik-metrik yang sesuai untuk setiap teknik data mining yang digunakan. Hasil yang ditemukan diinterpretasikan untuk memahami makna dan implikasi dari pola atau informasi yang diidentifikasi. Selama tahap ini, peneliti juga dapat mengidentifikasi temuan menarik, tren, atau pola yang dapat menjadi landasan untuk pengambilan keputusan lebih lanjut.
6. Pengambilan Keputusan: Hasil dari analisis data mining digunakan sebagai dasar untuk pengambilan keputusan yang lebih baik dan informasi yang lebih akurat[4]. Temuan yang ditemukan dapat digunakan untuk mengidentifikasi tren, memahami performa mahasiswa, atau menyusun strategi perbaikan dalam pendidikan. Misalnya, dengan mengelompokkan peserta berdasarkan penilaian mereka, pihak pendidik dapat memberikan perhatian lebih pada kelompok mahasiswa yang membutuhkan dukungan tambahan.
7. Implementasi dan Tindak Lanjut: Langkah terakhir adalah mengimplementasikan hasil dari analisis data mining ke dalam tindakan nyata dalam konteks pendidikan[4]. Hasil analisis dapat digunakan untuk mengoptimalkan strategi pembelajaran, meningkatkan kualitas program pendidikan, atau mengevaluasi efektivitas metode pengajaran. Implementasi dan tindak lanjut ini menjadi bagian penting dalam siklus penggunaan data mining untuk meningkatkan kualitas pendidikan dan pengambilan keputusan yang lebih cerdas.

Studi kasus dalam metode penelitian ini memungkinkan peneliti untuk menyajikan temuan secara konkret dan mendalam, serta memberikan wawasan yang berarti bagi bidang pendidikan. Dengan menggunakan teknik data mining dan preprocessing data, penelitian ini dapat memberikan kontribusi penting dalam pemahaman dan pengelolaan data penilaian dalam konteks pendidikan.

III. HASIL DAN PEMBAHASAN

Pada penelitian ini, data yang digunakan untuk analisis cluster adalah data dalam format Excel. Namun, terdapat banyak bagian data yang tidak diperlukan dan dapat mengganggu proses cluster. Oleh karena itu, dilakukan proses preprocessing data sebelum analisis cluster untuk memastikan data yang digunakan sudah sesuai dan berkualitas. Proses preprocessing data dalam analisis cluster merupakan tahap yang krusial dalam memastikan data yang digunakan berkualitas, relevan, dan siap untuk diolah lebih lanjut. Tahap-tahap dalam preprocessing data memiliki peran penting dalam menghilangkan hambatan dan memastikan hasil analisis yang akurat dan bermakna. Berikut ini adalah lebih lanjut tentang setiap tahap dalam proses preprocessing data pada penelitian ini:

A. Penghapusan Data Null

Penghapusan data null (nilai kosong) adalah langkah pertama dalam proses preprocessing. Pada tahap ini, dilakukan pengecekan terhadap data untuk mencari apakah ada baris data

yang memiliki nilai null. Jika ditemukan baris data dengan nilai null, baris tersebut akan dihapuskan dari dataset. Penghapusan data null bertujuan untuk menghindari potensi distorsi atau bias dalam hasil analisis akibat adanya data yang hilang. Dengan menghapus baris data yang memiliki nilai null, data yang digunakan dalam analisis menjadi lebih lengkap dan konsisten.

```
[ ]  data = data.dropna()
```

Gambar 1. Dropna

```
[ ]  data.isnull().sum()

Email                                    0
Nama                                     0
Course                                   0
Penilaian Proses Pm                      0
Penilaian Produk Pm                      0
Penilaian Kehadiran Pm                   0
Nilai Pendalaman Materi                  0
Penilaian Proses Ppp                     0
Penilaian Produk Ppp                     0
Penilaian Perangkat Ppp                  0
Penilaian Kehadiran Ppp                  0
Penilaian Praktik Ppl                    0
Nilai Perancangan Perangkat Pembelajaran 0
Penilaian Produk Ppl                     0
Penilaian Perangkat Ppl                  0
Penilaian Proses Ppl                     0
Max Nilai Ujian Komprehensif             0
Keaktifan Mengajar                       0
Cara Mengajar                            0
Hubungan Guru Murid                      0
Dedikasi                                 0
Kedalaman Materi                         0
Kompetensi Sosial                        0
Adaptibilitas                            0
Konsistensi                              0
Kepribadian                              0
dtype: int64
```

Gambar 2. Info Null

B. Penghapusan Data Duplikasi

Langkah selanjutnya adalah melakukan penghapusan data duplikasi atau kembar. Proses ini bertujuan untuk mengecek apakah ada data yang muncul lebih dari satu kali dalam dataset. Jika ditemukan data yang sama atau duplikat, salah satu data tersebut akan dihapus untuk mencegah perhitungan ganda yang dapat mengganggu hasil analisis. Penghapusan data duplikasi penting untuk memastikan bahwa setiap entitas dalam dataset hanya dihitung sekali, sehingga analisis lebih akurat dan valid.

```
[ ]  data = data.drop_duplicates()
```

Gambar 3. Drop Duplicate

```
[ ]  data.duplicated().sum()

     0
```

Gambar 4. Info Duplicate

C. Normalisasi Data

Setelah tahap penghapusan data null dan duplikasi, langkah selanjutnya adalah melakukan normalisasi data. Proses normalisasi ini dilakukan untuk mengubah skala data sehingga nilai-nilai dalam dataset memiliki rentang yang serupa. Normalisasi data sangat penting dalam analisis cluster karena membantu menghindari dominasi atribut dengan skala besar sehingga hasil cluster tidak didominasi oleh satu variabel saja. Dengan normalisasi, data akan lebih seimbang dan proses analisis cluster akan lebih efektif dan efisien.

```
# Melakukan normalisasi data
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)
```

Gambar 5. Normalized

D. Pengecekan Informasi Data

Pengecekan informasi data merupakan langkah untuk mengetahui berapa dimensi data dan apa jenis tipe data yang ada dalam dataset. Dengan mengetahui informasi ini, peneliti dapat mempersiapkan strategi analisis yang sesuai dengan tipe data yang ada. Misalnya, data kategorikal memerlukan pendekatan berbeda dalam analisis dibandingkan dengan data numerik. Pengecekan informasi data membantu peneliti untuk memahami karakteristik data yang akan diolah dan menentukan pendekatan analisis yang tepat.

E. Mengubah Email Menjadi ID

Proses ini melibatkan pengubahan parameter email menjadi ID unik. Tujuan dari langkah ini adalah untuk menyederhanakan data sehingga setiap email diwakili oleh ID yang unik. Dengan mengubah email menjadi ID, data menjadi lebih ringkas dan proses analisis lebih efisien. Selain itu, mengubah email menjadi ID juga membantu melindungi privasi data, karena ID tidak mengandung informasi pribadi seperti email.

```
[14] import re

     def casefolding(text):
       #membuat semua huruf menjadi huruf kecil
       text = text.lower()
       #menghilangkan huruf spesial
       text = re.sub(r'[@,.?#]', '', text)
       #menghilangkan URL
       text = re.sub(r'https?://\S+|www\.\S+', '', text)
       #untuk menghilangkan huruf
       text = re.sub(r'[a-zA-Z]', ' ', text)
       #menghilangkan angka dan perhitungan
       #text = re.sub(r'[-+]?[0-9]+', '', text)
       #menghilangkan non text
       text = re.sub(r'[^\w\s]','', text)
       #menghilangkan space
       text = text.strip()
       return text
```

```
[15] raw_sample = data['Email'].iloc[5]
     case_folding = casefolding(raw_sample)

     print('Raw data\t\t: ', raw_sample)
     print('Case folding\t\t: ', case_folding)

     Raw data            :  201900664115@guruku.id
     Case folding        :  201900664115
```

Gambar 6. Email To Id

F. Pengubahan Data Menjadi Parameter

Pada tahap ini, dilakukan pengubahan data menjadi parameter dengan memberikan bobot angka pada setiap course dan kepribadian yang ada dalam dataset. Pengubahan ini membantu dalam proses modeling data, karena data course dan kepribadian akan berbentuk angka setelah diberikan parameter. Hal ini mempermudah dalam perhitungan dan analisis data lebih lanjut. Pengubahan data menjadi parameter juga memungkinkan untuk memasukkan atribut kualitatif ke dalam analisis cluster.

G. Reduksi Dimensi

Reduksi dimensi adalah proses menggabungkan beberapa variabel menjadi satu untuk mengurangi kompleksitas data. Pada penelitian ini, dilakukan reduksi dimensi pada parameter penilaian proses pm dan penilaian produk pm dengan menghitung rata-rata dari kedua parameter tersebut. Selanjutnya, hasil rata-rata ini dijadikan kolom baru sebagai rata-rata penilaian. Reduksi dimensi membantu dalam mengurangi kompleksitas data dan memperoleh fitur yang lebih relevan dalam analisis cluster.

```
[ ]  # Membuat DataFrame dari parameter_map
     df = pd.DataFrame(parameter_map.items(), columns=['Course', 'Parameter'])

     # Menampilkan DataFrame
     print(df)
```

```
                                                Course  Parameter
0                 114 - 746 - 1 - Kelas 001 Bahasa Jawa          0
1              114 - 562 - 1 Gel 2 - Kelas 001 Seni Rupa          1
2             114 - 156 - 2 - Kelas 002 Bahasa Indonesia          2
3    114 - 097 - 2 - Kelas 009 Ilmu Pengetahuan Ala...          3
4    114 - 097 - 2 - Kelas 001 Ilmu Pengetahuan Ala...          4
..                                                 ...        ...
137  114 - 027 - 1 Gel 2 - Kelas 002 Pendidikan Gur...        137
138    114 - 514 - 3 - Kelas 001 Teknik Telekomunikasi        138
139  114 - 027 - 1 Gel 2 - Kelas 003 Pendidikan Gur...        139
140                  114 - 859 - 3 - Kelas 001 Kuliner        140
141          114 - 839 - 3 - Kelas 001 Teknik Perkapalan        141

[142 rows x 2 columns]
```

Gambar 7. Parameter Course

```
[ ]  # Membuat DataFrame dari parameter_map
     df = pd.DataFrame(parameter_map_2.items(), columns=['Kepribadian', 'Parameter'])

     # Menampilkan DataFrame
     print(df)
```

```
  Kepribadian  Parameter
0     Sanguin          1
1  Melankolis          2
2     Koleris          3
3   Plegmatis          4
```

Gambar 8. Parameter Kepribadian

H. Menjadikan Hasil Menjadi Dataset

Setelah seluruh proses preprocessing data selesai, data dihasilkan dalam bentuk parameter berupa ID, parameter_course, parameter_kepribadian, dan rata-rata nilai. Data tersebut kemudian disimpan dalam format CSV sebagai dataset yang siap untuk digunakan dalam analisis cluster. Penyimpanan dalam format CSV memudahkan penggunaan data di berbagai platform analisis dan memastikan data dapat diakses dengan mudah.

Dengan melakukan proses preprocessing data seperti dijelaskan di atas, data yang digunakan dalam analisis cluster sudah siap dan sesuai untuk menghasilkan temuan dan informasi penting dalam penelitian ini. Proses preprocessing merupakan langkah krusial dalam analisis data dan membantu memastikan kualitas data sebelum memasuki tahap analisis lebih lanjut. Hasil analisis cluster dari data yang telah diolah akan memberikan wawasan yang bermanfaat dan dapat digunakan sebagai dasar untuk pengambilan keputusan yang lebih baik dalam konteks penelitian ini..

UCAPAN TERIMA KASIH

Terima kasih yang sebesar-besarnya disampaikan kepada semua pembaca dan pihak yang telah berkontribusi dalam penulisan artikel ini. Tanpa dukungan dan partisipasi dari anda, artikel ini tidak akan terwujud.

REFERENSI

[1] Aggarwal, C. C. (2015). Data mining: the textbook (Vol. 1). New York: springer.
[2] Dista, T. M., & Abdulloh, F. F. (2022). Clustering Pengunjung Mall Menggunakan Metode K-Means dan Particle Swarm Optimization. Jurnal Media Informatika Budidarma, 6(3), 1339-1348.
[3] Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.
[4] Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.
[5] Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.
[6] Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2019). A survey of optimization methods from a machine learning perspective. IEEE transactions on cybernetics, 50(8), 3668-3681.