

BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Menurut (Ardika Putri, 2023), Sampah merupakan hasil dari kegiatan makhluk hidup yang dapat terdiri dari zat padat, cair, gas, dan zat organik maupun anorganik. Sampah ini dapat terbagi menjadi sampah yang dapat terurai kembali dan sampah yang tidak dapat diurai kembali yang tidak memberikan manfaat dan dibuang manusia ke lingkungan.

2.1.1 Jenis-jenis dan Karakteristik Sampah

Menurut Hadiwiyoti (1983, dikutip dalam (Rantunuwu, 2020)), Sampah terbagi menjadi beberapa jenis dan kriteria di antaranya asal sampah. Komposisi sampah, bentuk sampah, lokasi sampah, terjadinya sifat dan jenisnya.

2.1.2 Sumber Sampah

Sumber sampah yaitu tempat dimana sampah tersebut dihasilkan dari berbagai aktivitas dan fenomena yang menghasilkan sampah. Pemahaman terkait sumber sampah penting untuk mengetahui pengelolaan sampah yang baik dan berpotensi menghasilkan nilai ekonomis. Menurut buku yang ditulis oleh (Tarigan & Mariana, 2023), sumber sampah dapat bersumber dari pemukiman, industri, perkantoran, dan pusat perbelanjaan

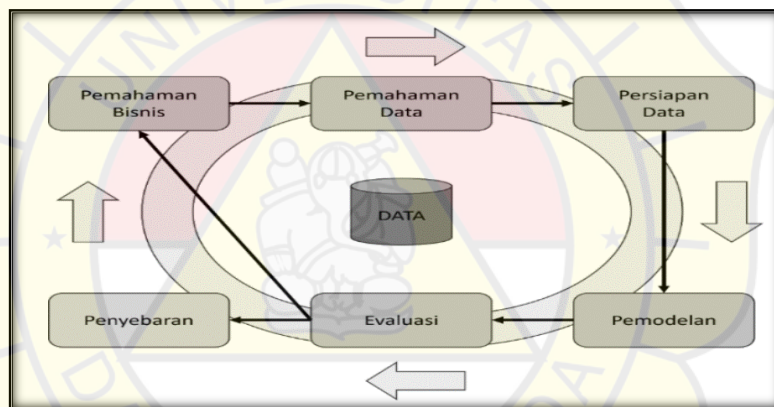
2.1.3 Data Mining

Data mining adalah proses untuk menemukan pola-pola yang tersembunyi dalam data dengan tujuan untuk memahami dan menganalisis data, serta dapat

menjelaskan suatu fenomena tertentu. Alur kerja data mining terdiri dari beberapa langkah diantaranya pengumpulan data, pembersihan data, transformasi, agregasi, pemodelan, analisis prediktif, visualisasi, dan penyebaran hasil (Lasfeto, 2021).

2.1.4 CRISP DM

Menurut (Asyraf & Prasetya, 2024) metode CRISP-DM adalah kerangka kerja yang sering digunakan dalam analisis data di data mining, terdiri dari enam tahapan yaitu pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi.



Gambar 2. 1 Siklus CRISP-DM Sumber (Asyraf & Prasetya, 2024)

1. Pemahaman Bisnis

Proses ini melibatkan pengidentifikasian masalah yang dihadapi. Selain itu, proses pemahaman bisnis juga digunakan untuk menentukan tujuan bisnis, menganalisis situasi dan kondisi saat penelitian dilakukan, serta merumuskan tujuan penelitian yang akan diselesaikan dengan menggunakan teknik data mining (Dhewayani et al., 2022) .

2. Pemahaman Data

Pada tahap ini, data yang telah terkumpul dianalisis untuk memastikan kelengkapan dan keakuratan informasi. Selain itu, tahap pemahaman data juga digunakan untuk mempelajari data dan mengidentifikasi masalah yang terkait (Hidayati et al., 2020) .

3. Persiapan Data

Pada tahap berikutnya, data yang telah dipahami sebelumnya dipersiapkan untuk proses penambangan data. Data tersebut akan diubah menjadi format digital, kemudian dilakukan pembersihan dan pengolahan untuk memastikan kualitasnya. Proses ini bertujuan agar data siap digunakan dalam pembuatan model menggunakan algoritma pohon keputusan (Asyraf & Prasetya, 2024) .

4. Pemodelan

Menurut (Dhewayani et al., 2022) , Pemodelan atau modeling adalah tahap penerapan algoritma yang digunakan untuk mencari, mengidentifikasi, dan menghasilkan pola yang akan diterapkan pada data penelitian.

5. Evaluasi

Selanjutnya yaitu tahap evaluasi, dimana menurut (Hidayati et al., 2020) , evaluasi merupakan tahap untuk menafsirkan hasil dari proses data mining yang dihasilkan selama tahap pemodelan.

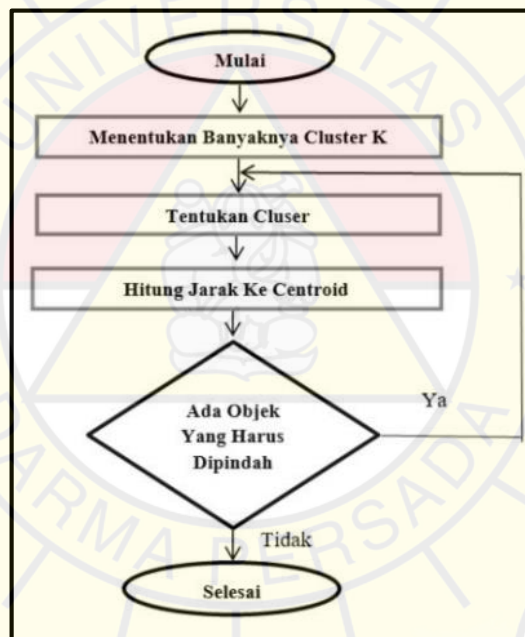
6. Penyebaran (implementasi data)

Tahap terakhir, hasil dari model dan analisis data mining akan diterapkan dalam konteks bisnis yang sesungguhnya. Model yang telah dibuat akan digunakan dalam situasi bisnis nyata, menyajikan temuan analisis, atau bahkan diintegrasikan

ke dalam sistem yang relevan untuk mendukung pengambilan keputusan (Asyraf & Prasetya, 2024) .

2.1.5 Algoritma *K-Means*

Algoritma *K-Means* adalah metode *clustering* tanpa *supervise* yang digunakan untuk mengelompokkan data tanpa label menjadi sebuah kelompok yang menandakan jumlah kelompok yang ingin dicapai sebagai hasil akhir dari proses pengelompokkan (Lasfeto, 2021).



Gambar 2. 2 Alur *K-Means* Sumber (Prastiwi et al., 2022)

Bedasarkan penelitian yang dilakukan oleh (Pramitasari & Nataliani, 2021) langkah-langkah untuk mengolah data dengan algoritma clustering diantaranya :

1. Memasukkan data.
2. Menentukan jumlah *cluster*

3. Mengambil data secara acak sesuai dengan jumlah klaster yang akan digunakan untuk pusat klaster (*centroid*).
4. Menghitung jarak antara data dengan pusat *cluster*. Dibawah ini merupakan beberapa metode yang digunakan:

a. *Euclidean distance*

Berdasarkan penelitian yang dilakukan oleh (Wahyu Pribadi et al., 2022), rumus untuk menentukan jarak dengan metode *Euclidean distance* adalah sebagai berikut:

$$d(x, y) = |x - y| = \sqrt{\sum_{i=1}^N (x_i - y_i)^2} \quad (1)$$

Keterangan:

d = jarak antara x dan y

x = data pusat *cluster*

y = data pada atribut

i = setiap data

n = jumlah data

x_i = data pada pusat *cluster* ke- i

y_i = data pada setiap data ke- i

b. *Manhattan distance*

Rumus untuk menentukan jarak dengan *Manhattan distance* menurut (Wahyu Pribadi et al., 2022)

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (2)$$

Keterangan:

d = jarak antara x dan y

n = jumlah data

x = data pusat *cluster*

x_i = data pada pusat *cluster* ke- i

y = data pada atribut

y_i = data pada setiap data k

i = setiap data

c. *Minkowski distance*

Rumus untuk menentukan jarak dengan *Minkowski distance* menurut (Putrasyah et al., 2023)

$$d(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (3)$$

Keterangan:

d = jarak antara x dan y

n = jumlah data

x = data pusat custer

x_i = data pada custer ke i

y = data pada atribut

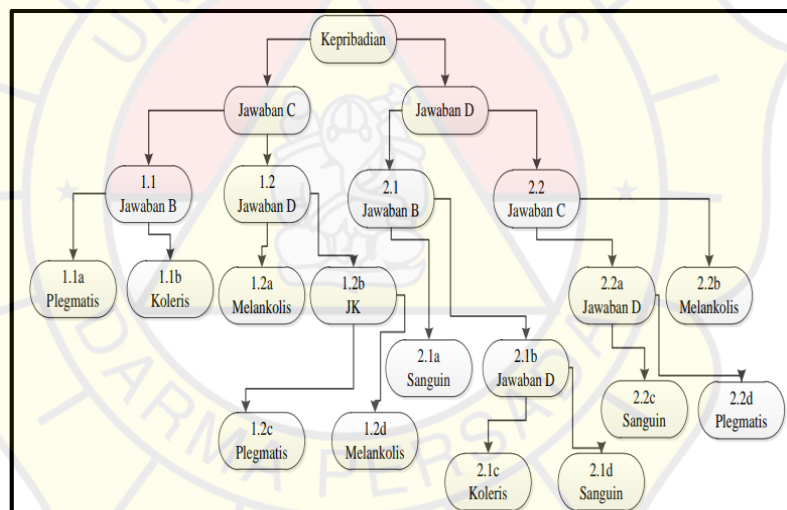
y_i = sata pada setiap data ke i

i = setiap data

p = power

2.1.6 Algoritma *Decision Tree*

Decision Tree adalah sebuah metode klasifikasi yang memiliki struktur menyerupai *flowchart* berbentuk seperti pohon, dimana setiap simpul mewakili sebuah pengujian atribut, kemudian setiap cabang menunjukkan hasil dari pengujian tersebut dan daun menggambarkan kelas atau distribusi kelas. Manfaat utama dari penggunaan *Decision Tree* adalah kemampuannya untuk menyederhanakan proses pengambilan keputusan yang rumit, sehingga mempermudah pengambilan keputusan serta lebih cepat menghasilkan solusi untuk masalah yang dihadapi (Fauziningrum & Suryaningsih, 2021).



Gambar 2. 3 Alur *Decision Tree* Sumber (Ferdinan Leo Simanjuntak et al., 2021)

Pada algoritma *decision tree*, mengukur dan memilih atribut yang paling informatif untuk membagi data pada setiap langkah pemisahan dalam pohon keputusan dapat menggunakan *Entropy* dan *Information Gain*. Menurut penelitian (Ferdinan Leo Simanjuntak et al., 2021) rumus dari *Entropy* dan *Information Gain* diantaranya :

a. Rumus *Entropy*

$$Entropy(S) = -\frac{w}{n} * \log_2 \left(\frac{w}{n} \right) - \frac{x}{n} * \log_2 \left(\frac{x}{n} \right) - \frac{y}{n} * \log_2 \left(\frac{y}{n} \right) - \frac{z}{n} * \log_2 \left(\frac{z}{n} \right) \quad (4)$$

Keterangan :

S = Entropi atribut

y = total data dari tipe Melankolis

w = total data dari tipe Sanguinis

z = total data dari tipe Plegmatis

x = total data dari tipe Koleris

n = jumlah total keseluruhan data

b. Rumus *Information Gain*

$$\begin{aligned} \text{Gain} = & \text{entropy}(S) - \frac{nL}{n} * \left(-\frac{nLw}{nL} \right) * \log_2 \left(\frac{nLw}{nL} \right) - \frac{nLx}{nL} * \log_2 \left(\frac{nLx}{nL} \right) - \frac{nLy}{nL} * \\ & \log_2 \left(\frac{nLy}{nL} \right) - \frac{nLz}{nL} * \log_2 \left(\frac{nLz}{nL} \right) + \frac{nP}{n} * \left(-\frac{nPw}{nP} \right) - \frac{nPy}{nP} \left(\frac{nPw}{nP} \right) - \frac{nPx}{nP} * \log_2 \left(\frac{nPx}{nP} \right) - \\ & \frac{nPy}{nP} \log_2 \left(\frac{nPz}{nP} \right) - \frac{nPz}{nP} * \log_2 \left(\frac{nPz}{nP} \right) \end{aligned} \quad (5)$$

Keterangan:

n = total jumlah keseluruhan data

nLz = total data bertipe Plegmatis

nL = total jumlah data laki-laki

dengan jenis kelamin laki-laki

nP = total jumlah data untuk jenis

nPw = total data bertipe Sanguinis

kelamin perempuan

dengan jenis kelamin perempuan

nLw = total data bertipe Sanguinis

nPx = total data bertipe Koleris

dengan jenis kelamin laki-laki

dengan jenis kelamin perempuan

nLx = total data bertipe Koleris

nPy = total data bertipe Melankolis

dengan jenis kelamin laki-laki

dengan jenis kelamin perempuan

nLy = total data bertipe Melankolis

nPz = total data bertipe Plegmatis

dengan jenis kelamin laki-laki

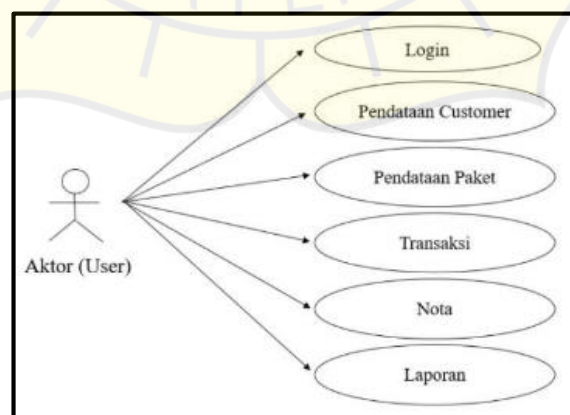
dengan jenis kelamin perempuan

2.1.7 Pemodelan Sistem UML

Menurut penelitian (Abdillah, 2021), *Unified Modeling Language* (UML) adalah bahasa pemodelan perangkat lunak yang digunakan sebagai alat untuk merancang rancangan perangkat lunak. UML dapat digunakan untuk berbagai tujuan, seperti visualisasi, spesifikasi, konstruksi, dan dokumentasi bagian-bagian dari sistem dalam perangkat lunak.

2.1.7.1 Usecase Diagram

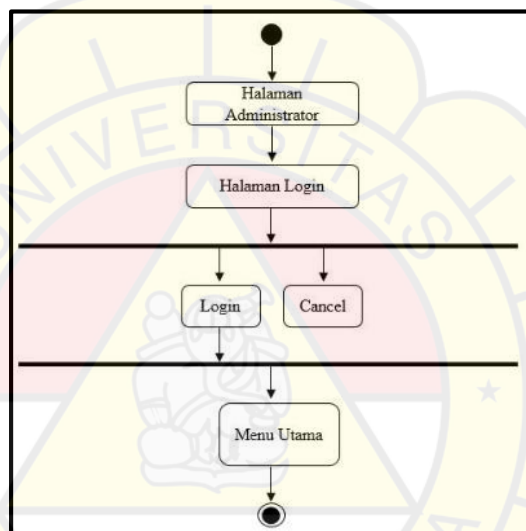
Use Case Diagram menggambarkan hubungan antara pengguna dan sistem secara umum. Dibawah ini merupakan contoh *usecase* diagram pada penelitian (Abdillah, 2021), dimana pengguna atau aktor yang berinteraksi dengan sistem, dalam hal ini adalah admin dari usaha persewaan. Admin memiliki akses penuh terhadap sistem dan dapat mengelola berbagai fitur yang tersedia. Sementara itu, *customer* atau pemilik usaha tidak memiliki hak akses langsung ke sistem, namun admin akan memberikan laporan sebagai hasil keluaran dari sistem perangkat lunak.



Gambar 2. 4 Contoh *usecase* diagram Sumber (Abdillah, 2021)

2.1.7.2 Activity Diagram

Activity diagram menggambarkan langkah-langkah dalam proses yang dapat dijalankan oleh sistem, dengan fokus pada tiap sub bagian (objek). Diagram ini kemudian menjelaskan bagaimana setiap fitur objek akan diimplementasikan dalam kode program. Objek-objek tersebut akan saling terhubung melalui koneksi yang mengarah ke database untuk menyimpan atau mengelola data yang diperlukan.



Gambar 2. 5 Contoh *usecase* diagram sumber (Abdillah, 2021)

2.1.8 Software dan Pemrograman Terkait

Pada bagian ini akan dibahas mengenai perangkat lunak dan bahasa pemrograman yang digunakan dalam pengolahan data dan pengembangan model

2.1.8.1 Google Colab Editor

Dalam pengembangan model berbasis data, menggunakan platform pemrograman sangat penting untuk memastikan proses analisis dan pengujian algoritma berjalan dengan baik. Menurut (Fadil Danu Rahman et al., 2024) *Google Colab* adalah platform berbasis web yang memungkinkan pengguna menulis dan menjalankan kode *Python* secara fleksibel langsung melalui browser, platform ini

sangat cocok untuk kebutuhan *machine learning* dan analisis data, sehingga mempermudah proses pengembangan model secara praktis.

2.1.8.2 Bahasa Pemrograman Python

Python merupakan bahasa pemrograman berbasis skrip yang bersifat berorientasi objek. Bahasa ini dapat digunakan untuk mengembangkan perangkat lunak dan kompatibel dengan berbagai sistem operasi. Saat ini, Python menjadi salah satu bahasa yang banyak digunakan di bidang data *science* dan analisis, terutama karena dukungannya terhadap berbagai pustaka (*library*) yang menyediakan alat untuk analisis data, pembelajaran mesin, pengolahan data, serta visualisasi data (Edwardo, 2018, dikutip dalam (Koretsky, 2023)).

Pustaka-pustaka yang membantu Python melakukan analisis data diantaranya:

1. NumPy

Menurut (Talia et al., 2024), NumPy adalah sebuah library Python yang dirancang untuk mendukung operasi numerik secara efisien. Library ini menawarkan struktur data berupa array serta berbagai fungsi matematika untuk mempermudah dan meningkatkan kecepatan dalam pengolahan data numerik.

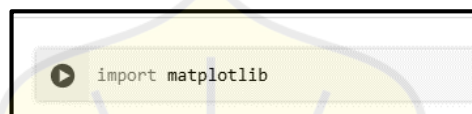
Dibawah ini merupakan implementasi Numpy:



Gambar 2. 6 Implementasi Library Numpy

2. Matplotlib

Matplotlib merupakan pustaka Python yang digunakan untuk membuat visualisasi data, khususnya grafik dan plot. Pustaka ini banyak diterapkan dalam analisis data untuk menghasilkan visualisasi yang jelas dan menarik. Dibawah ini merupakan implementasi Matplotlib:



Gambar 2. 7 Implementasi Library Matplotlib

3. Pandas

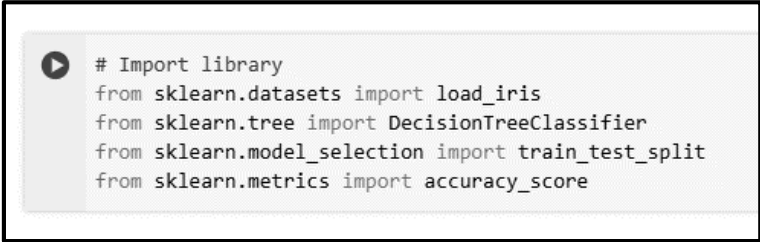
Menurut (Isa Albanna & R. Tri hadi laksono, 2022), Pandas adalah pustaka Python yang dirancang untuk mengolah data secara efisien. Dengan sifatnya yang fleksibel, serbaguna, dan berbasis open source, Pandas menjadi alat yang dapat diandalkan dan banyak digunakan untuk berbagai kebutuhan dalam pengelolaan dan analisis data.



Gambar 2. 8 Implementasi Library Pandas

4. Scikit Learn

Menurut (Suarsa et al., 2021), scikit-learn adalah library Python yang populer untuk machine learning, menyediakan berbagai algoritma dan alat untuk tugas seperti klasifikasi, regresi, dan clustering. Dibawah ini merupakan contoh implementasi penggunaan Scikit Learn:

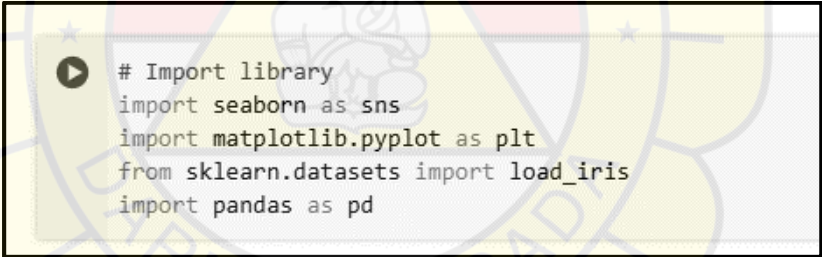
A screenshot of a code editor window with a light gray background. It contains Python code for importing Scikit-Learn libraries. The code is as follows:

```
# Import library
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Gambar 2. 9 Implementasi Scikit Learn

5. Seaborn

Menurut (Harahap et al., 2021), seaborn adalah library Python yang digunakan untuk memproses, membuat visualisasi data, dan menyesuaikan data. Library ini membantu membuat grafik statistik yang mudah dibaca dan menarik, serta bekerja dengan baik bersama Pandas dan Matplotlib untuk analisis data. Dibawah ini merupakan contoh implementasi Seaborn:

A screenshot of a code editor window with a light gray background. It contains Python code for importing Seaborn, Matplotlib, and Pandas, along with the Scikit-Learn dataset loader. The code is as follows:

```
# Import library
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
import pandas as pd
```

Gambar 2. 10 Implementasi Seaborn

2.1.8.3 Streamlit

Python Streamlit, kerangka kerja aplikasi *open-source* yang dirancang untuk merampingkan pembuatan aplikasi data (Choudhary & Jain, 2019), seperti yang dikutip dalam (Maulana & Laksana, 2024). Dengan menggunakan Streamlit, pengembang dapat dengan cepat membangun aplikasi interaktif yang menampilkan visualisasi data tanpa perlu menulis banyak kode.

2.2 Tinjauan Literatur/Kajian Penelitian Terdahulu

Pada bagian ini, akan menjelaskan mengenai tinjauan pustaka atau kajian penelitian terdahulu yang berhubungan dengan topik penelitian ini. Tinjauan pustaka bertujuan untuk memperdalam pemahaman mengenai penelitian-penelitian yang telah dilakukan sebelumnya dalam bidang yang sama atau terkait, Berikut adalah rangkuman penelitian pertama yang relevan dengan topik ini.

2.2.1 Paper 1

Paper pertama yaitu penelitian yang berjudul “Penerapan Algoritma Klasterisasi dan Klasifikasi pada Tingkat Kepentingan Sistem Pembelajaran di Universitas Terbuka” yang ditulis oleh (Hidayati et al., 2020) dipublikasikan pada tahun 2020 dan penelitian ini termasuk dalam jurnal klasifikasi SINTA 4.

2.2.1.1 Tujuan Penelitian

Tujuan penelitian tersebut adalah untuk mengklaster dan mengklasifikasi data hasil kuesioner tingkat kepentingan sistem pembelajaran Universitas Terbuka dan untuk memahami penilaian terhadap masing-masing mahasiswa pada sistem pembelajaran yang diberikan.

2.2.1.2 Metodologi Yang Digunakan

Metodologi yang digunakan dalam penelitian ini yaitu metode K-medoids clustering yang diterapkan pada proses datamining dan langkah langkah mengikuti tahapan CRISP-DM.

2.2.1.3 Temuan Utama

Temuan utama pada penelitian ini meliputi:

- a. Penelitian ini berhasil mengelompokkan data tingkat kepentingan sistem pembelajaran di Universitas Terbuka menjadi dua klaster yaitu klaster 0 dengan 273 data dan klaster 1 dengan 97 data.
- b. Menemukan algoritma klasifikasi yang tinggi dari beberapa algoritma diantaranya Naïve Bayes, k-NN, dan C4.5, diantara ketiga algoritma tersebut Naïve Bayes menghasilkan akurasi tertinggi sebesar 72,70%.

2.2.1.4 Kelemahan Penelitian

Kelemahan dari penelitian ini adalah rendahnya akurasi model. Meskipun akurasi *Naïve Bayes* sebesar 72,70%, ini masih belum cukup untuk aplikasi yang lebih luas dalam bidang pendidikan, artinya model tersebut masih belum optimal saat digunakan. Selain itu, keterbatasan jumlah data yang digunakan menjadi kendala, dimana penelitian ini merekomendasikan agar studi selanjutnya melibatkan data yang lebih banyak untuk meningkatkan akurasi.

2.2.2 Paper 2

Paper kedua yaitu penelitian berjudul “Klasifikasi Timun Segar dan Busuk Menggunakan K-Means Clustering” yang ditulis oleh (Saputra et al., 2024). Penelitian ini diterbitkan dalam *Journal of Education Research* dan dikategorikan sebagai jurnal dengan klasifikasi SINTA 4.

2.2.2.1 Tujuan Penelitian

Tujuan penelitian tersebut yaitu untuk membangun sebuah model klasifikasi yang baik menggunakan algoritma *K-Means* untuk membedakan timun segar dan timun busuk berdasarkan karakteristik fisiknya.

2.2.2.2 Metodologi Yang Digunakan

Metodologi yang digunakan akan melibatkan beberapa langkah sistematis untuk menerapkannya ke dalam algoritma *K-Means* diantaranya pengumpulan data, pemilihan atribut, penerapan algoritma dan evaluasi hasil.

2.2.2.3 Temuan Utama

Temuan utama dari penelitian ini meliputi:

- a. Metode *K-Means* efektif dalam mengklasifikasikan timun segar dan busuk dengan tingkat akurasi yang tinggi.
- b. Penggunaan teknik pemrosesan untuk memisahkan kecerahan dan warna serta tekstur dan bentuk, model yang digunakan dapat secara akurat mengidentifikasi kondisi timun.
- c. Model berhasil mengklasifikasikan 49 citra timun dengan benar.

2.2.2.4 Kelamahan Penelitian

Penelitian ini memiliki beberapa kelemahan yaitu Pertama, jumlah data yang digunakan pada pengujian sangat terbatas. Keterbatasan jumlah data ini dapat mempengaruhi kemampuan daya prediksi model. Kedua, penelitian ini terbatas pada metode ekstraksi fitur sederhana seperti warna, tekstur, dan ukuran dari citra timun.

2.2.3 Paper 3

Paper ketiga yaitu studi yang berjudul “K-Means Clustering Untuk Klasifikasi Standar Kualifikasi Pendidikan Dan Pengalaman Kerja Guru Smk” yang ditulis oleh (Wibowo & Habanabakize, 2022) dipublikasikan pada tahun 2022 dalam Jurnal Dinamika Vokasional Teknik Mesin dan klasifikasi penelitian ini termasuk dalam klasifikasi jurnal SINTA 4.

2.2.3.1 Tujuan Penelitian

Penelitian tersebut bertujuan untuk mengklasifikasikan seluruh provinsi di Indonesia dalam bentuk *cluster* berdasarkan kesamaan karakteristik kualifikasi guru serta mengklasifikasikan provinsi yang memenuhi standar kualifikasi guru berdasarkan Permendikbud.

2.2.3.2 Metodologi Yang Digunakan

Metodologi yang digunakan yaitu pendekatan kuantitatif dengan metode K-Means. Metode tersebut dipilih karena umum digunakan untuk mengelompokkan karakteristik kesamaan dengan jumlah data yang besar dan komputasi yang cepat.

2.2.3.3 Temuan Utama

Temuan utama dari penelitian ini adalah metode K-Means berhasil mengelompokkan provinsi-provinsi di Indonesia menjadi tiga kluster berdasarkan kualifikasi pendidikan dan pengalaman kerja guru SMK.

2.2.3.4 Kelemahan Penelitian

Pada penelitian ini ditemukan beberapa kelemahan diantaranya:

- a. Penggunaan data sekunder dari statistik pendidikan yang belum sepenuhnya menggambarkan kondisi terbaru di lapangan.
- b. Hanya dua indikator yang digunakan untuk mengukur kualitas guru SMK yaitu, kualifikasi pendidikan minimal S1 dan pengalaman kerja minimal 4 Tahun.
- c. Hasil *clustering* pada penelitian ini mungkin sulit untuk diterapkan ke dalam konteks pendidikan di seluruh Indonesia kaeran adanya perbedaan budaya, ekonomi dan kebijakan.

2.2.4 Paper 4

Paper keempat yaitu studi yang berjudul “Clustering Sampah Yang Dihasilkan Oleh Masyarakat Kota Surabaya Dengan Menggunakan Metode K-Means” yang ditulis oleh (Ardika Putri, 2023) dipublikasikan pada tahun 2023 dalam Jurnal senopati dan klasifikasi penelitian ini termasuk dalam klasifikasi jurnal SINTA 5.

2.2.4.1 Tujuan Penelitian

Penelitian ini berfokus pada pengelompokan wilayah di Kota Surabaya berdasarkan tingkat produksi sampah tertinggi dan terendah menggunakan metode clustering K-means. Dengan pendekatan ini, diharapkan penelitian dapat memberikan rekomendasi penanganan sampah yang lebih efektif untuk wilayah-wilayah tersebut.

2.2.4.2 Metodologi Yang Digunakan

Penelitian ini menggunakan metodologi clustering K-means, di mana prosesnya meliputi pengumpulan data tentang jenis, sumber, dan berat sampah di Kota Surabaya. Data yang diperoleh kemudian diproses menggunakan software RapidMiner untuk mengelompokkan wilayah berdasarkan tingkat produksi sampah tertinggi dan terendah.

2.2.4.3 Temuan Utama

Temuan utama dari penelitian ini adalah mendapatkan hasil identifikasi dua *cluster* berdasarkan Produksi sampah di Surabaya, diantaranya:

- a. *Cluster 1*, terdiri dari 11 distrik dengan produksi sampah tertinggi di Surabaya
- b. *Cluster 2*, terdiri dari 636 distrik yang merupakan penghasil sampah terendah.

2.2.4.4 Kelemahan Penelitian

Penelitian ini memiliki batasan terkait data yang digunakan, yang kemungkinan tidak mencakup semua jenis atau sumber sampah sehingga hasil pengelompokan mungkin tidak sepenuhnya akurat dalam mewakili keseluruhan populasi. Selain itu, metode K-means yang diterapkan juga memiliki kelemahan, seperti tantangan dalam menentukan jumlah kluster yang optimal serta ketergantungan pada pemilihan awal centroid, yang dapat berdampak pada hasil akhir analisis.

2.2.5 Paper 5

Paper kelima yaitu studi yang berjudul “Penerapan Metode Principal Component Analysis (PCA) Pada Klasifikasi Status Kredit Nasabah Bank Sumsel

Babel Cabang KM 12 Palembang Menggunakan Metode Decision Tree” Menurut B. Charbuty and A. Abdulazeez dalam (Sartika & Saluza, 2022) dipublikasikan pada tahun 2022 dalam Jurnal generic dan klasifikasi penelitian ini termasuk dalam klasifikasi jurnal SINTA 5.

2.2.5.1 Tujuan Penelitian

Penelitian ini bertujuan untuk menganalisis status pembayaran nasabah, mempertimbangkan berbagai faktor seperti riwayat pembayaran kredit sebelumnya. Selain itu, penelitian ini mengevaluasi bagaimana penerapan PCA dapat meningkatkan keakuratan klasifikasi. Hasilnya diharapkan memberikan pemahaman lebih dalam tentang dampak PCA pada indikator kinerja seperti akurasi, presisi, recall, dan F1-score dalam klasifikasi status kredit

2.2.5.2 Metodologi Yang Digunakan

Penelitian ini menggunakan pendekatan metodologi yang melibatkan beberapa langkah kunci. Tahap pertama adalah kajian literatur, yang bertujuan untuk mengumpulkan informasi mengenai konsep data mining, metode klasifikasi, proses pre-processing, Principal Component Analysis (PCA), algoritma Decision Tree, serta metrik evaluasi seperti precision, recall, dan F1-score. Kemudian, perumusan masalah dilakukan dengan melakukan observasi di Bank Sumsel Babel Cabang KM 12 Palembang untuk memahami mekanisme penilaian status kolektibilitas kredit nasabah. Data yang dikumpulkan terkait status kolektibilitas kredit nasabah selanjutnya dianalisis, di mana data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian. Proses analisis ini bertujuan untuk mengevaluasi dampak PCA pada performa klasifikasi dengan algoritma Decision Tree.

2.2.5.3 Temuan Utama

Penggunaan Principal Component Analysis (PCA) terbukti secara signifikan meningkatkan efektivitas klasifikasi dengan metode Decision Tree dalam mengevaluasi status kredit nasabah di Bank Sumsel Babel.

2.2.5.4 Kelemahan Penelitian

Penelitian ini memiliki beberapa keterbatasan, salah satunya adalah ukuran sampel yang kecil, yakni hanya 49 data kredit nasabah, yang mungkin tidak cukup untuk merepresentasikan populasi yang lebih besar. Ukuran sampel yang terbatas ini bisa memengaruhi ketepatan dan konsistensi model klasifikasi yang dibangun. Di samping itu, penelitian ini hanya mempertimbangkan 11 variabel independen, yang mungkin belum mencakup seluruh faktor yang dapat memengaruhi status kredit nasabah, sehingga model yang diperoleh mungkin kurang menyeluruh.