BAB II

LANDASAN TEORI

2.1. Data Science

Data Science adalah disiplin ilmu yang berkaitan dengan pengumpulan, analisis, interpretasi, presentasi, dan pengelolaan data untuk mendukung pengambilan keputusan. Ini melibatkan penerapan konsep dan teknik dari statistika, matematika, dan ilmu komputer untuk menggali wawasan atau pola yang terkandung dalam data. Tujuan utama dari Data Science adalah mengubah data mentah menjadi informasi yang berharga, yang dapat digunakan untuk membuat keputusan bisnis yang informasional. Pertama-tama, proses Data Science dimulai dengan pengumpulan data dari berbagai sumber. Ini bisa mencakup data terstruktur dari basis data relasional, data semi-struktur seperti format XML atau JSON, dan data tak terstruktur seperti teks, gambar, atau video. Setelah data dikumpulkan, langkah berikutnya adalah membersihkan dan merapikannya untuk memastikan kualitas dan keterandalan data.

2.2. Knowledge Discovery in Database

2.1.1. Data Mining

Data Mining adalah proses ekstraksi pola atau pengetahuan yang bermanfaat dari suatu set data yang besar, kompleks, dan sering kali tak terstruktur [1]. Tujuan utama dari Data Mining adalah untuk mengidentifikasi hubungan atau pola tersembunyi dalam data yang tidak dapat dengan mudah diungkap melalui analisis konvensional [2]. Dalam hal ini, teknik dan algoritma Data Mining digunakan untuk mengeksplorasi dan menganalisis data, mengungkapkan tren yang mungkin tidak

terlihat secara langsung. Salah satu aspek kunci dari Data Mining adalah penggunaan algoritma machine learning untuk mengenali pola dan relasi dalam data [3]. Algoritma ini dapat digunakan untuk melakukan klasifikasi, di mana data dibagi menjadi kelompok berdasarkan karakteristik tertentu, atau regresi, di mana hubungan antara variabel-variabel diukur dan diprediksi. Selain itu, teknik clustering juga sering digunakan untuk mengelompokkan data menjadi subset yang serupa berdasarkan karakteristik tertentu.

2.1.2. Database dan Data Processing

Database dan data processing (pengolahan data) adalah dua aspek kunci dalam dunia teknologi informasi yang saling terkait dan berperan penting dalam menyimpan, mengelola, dan memanfaatkan informasi. Pertama-tama, database adalah kumpulan data yang diorganisir dengan cara tertentu agar mudah diakses, dikelola, dan diperbarui. Tujuan utama database adalah menyediakan sarana untuk menyimpan data secara efisien dan memberikan akses yang cepat ke informasi yang diperlukan. Basis data dapat terdiri dari berbagai jenis data, termasuk teks, angka, gambar, atau suara, dan menggunakan struktur tertentu seperti tabel untuk menyusun dan mengelompokkan informasi.

2.1.3. Visualisation

Visualisasi data adalah teknik untuk menggambarkan informasi dan konsep melalui representasi visual, seperti grafik, diagram, atau peta. Tujuan utama dari visualisasi data adalah menyampaikan informasi secara jelas dan efektif agar dapat dengan mudah dipahami oleh pemirsa. Ini melibatkan penggunaan elemen visual seperti warna, bentuk, dan ukuran untuk menyampaikan makna dan pola dalam data

yang mungkin sulit dipahami dalam bentuk tabel atau angka saja. Dalam dunia yang semakin dibanjiri data, visualisasi data memainkan peran kunci dalam membantu individu dan organisasi untuk menggali wawasan yang berharga dari data mereka. Melalui grafik atau peta yang mudah dimengerti, pengguna dapat dengan cepat mengidentifikasi tren, anomali, atau pola yang dapat membimbing pengambilan keputusan atau memberikan pemahaman yang lebih mendalam tentang suatu topik.

2.1.4. Statistik

Statistik adalah cabang ilmu matematika yang berkaitan dengan pengumpulan, analisis, interpretasi, presentasi, dan pengorganisasian data. Tujuannya adalah untuk menyajikan data dalam bentuk yang bermakna, membantu dalam membuat keputusan yang informasional, dan memberikan wawasan tentang fenomena yang diamati. Statistik terbagi menjadi dua jenis utama: statistik deskriptif dan statistik inferensial. Statistik deskriptif berkaitan dengan cara menggambarkan dan merangkum data secara singkat. Ini melibatkan penggunaan ukuran pemusatan seperti rata-rata dan median, serta ukuran sebaran seperti deviasi standar dan rentang, untuk memberikan gambaran yang komprehensif tentang distribusi data. Grafik dan tabel juga sering digunakan dalam statistik deskriptif untuk menyajikan informasi secara visual.

2.1.5. Pattern Recognition

Pattern Recognition, atau Pengenalan Pola, adalah cabang ilmu yang berkaitan dengan identifikasi dan interpretasi pola dalam data. Tujuan utama dari Pattern Recognition adalah mengembangkan algoritma dan model untuk mengenali dan mengklasifikasikan pola yang mungkin tersembunyi dalam data yang kompleks

atau tak terstruktur. Ini mencakup penggunaan teknik-teknik dari berbagai disiplin ilmu, termasuk statistika, machine learning, dan ilmu komputer. Salah satu aplikasi utama Pattern Recognition adalah di bidang pengolahan citra dan pengenalan wajah. Dalam konteks ini, algoritma dapat diajarkan untuk mengenali wajah manusia atau objek tertentu dalam gambar atau video. Teknologi ini digunakan secara luas dalam keamanan, pengenalan wajah pada perangkat seluler, dan bahkan dalam industri kesehatan untuk mengidentifikasi pola medis pada citra radiologi.

Selain itu, Pattern Recognition juga digunakan dalam pengolahan suara untuk mengenali pola dalam sinyal audio, seperti pengenalan ucapan atau musik. Dalam aplikasi lain, seperti pengenalan tulisan tangan atau pengenalan karakter pada dokumen, teknik Pattern Recognition dapat digunakan untuk mengubah data yang bersifat tak terstruktur menjadi informasi yang dapat diinterpretasikan. Machine learning, khususnya dalam konteks deep learning, telah memberikan dorongan besar bagi kemajuan dalam bidang Pattern Recognition. Jaringan saraf tiruan dapat mempelajari pola yang kompleks dan mampu mengenali fitur yang sulit untuk diidentifikasi oleh algoritma tradisional. Ini memungkinkan aplikasi yang lebih canggih dalam pengenalan objek, klasifikasi, dan prediksi berbasis pola. Meskipun kemajuan yang signifikan telah dicapai dalam bidang ini, tantangan tetap ada, terutama ketika berhadapan dengan data yang tidak terstruktur atau saat diperlukan interpretasi kontekstual yang mendalam. Dengan terus berkembangnya teknologi, Pattern Recognition terus menjadi fokus penelitian dan pengembangan untuk memahami dan memanfaatkan pola dalam data dengan cara yang lebih cerdas dan adaptif.

2.3. Model Clustering

Model clustering merupakan metode analisis data yang digunakan untuk mengelompokkan objek atau data ke dalam kelompok-kelompok berdasarkan kemiripan fitur atau karakteristik tertentu [4]. Tujuan utama dari model clustering adalah mengidentifikasi pola atau struktur dalam data yang mungkin sulit ditemukan secara manual [5]. Salah satu model clustering yang umum digunakan adalah K-Means. Algoritma K-Means bekerja dengan menentukan pusat kelompok atau centroid, lalu mengelompokkan objek ke dalam kelompok yang memiliki centroid terdekat [6]. Hal ini membentuk partisi data ke dalam kelompok yang homogen dan meminimalkan varian di dalam kelompok tersebut.

Selain K-Means, model clustering lainnya adalah Hierarchical Clustering, yang membangun hirarki kelompok berdasarkan tingkat kemiripan antar-objek. Dengan pendekatan ini, objek yang lebih mirip satu sama lain dikelompokkan bersama dalam tingkat hirarki yang lebih tinggi. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) adalah model clustering yang fokus pada kepadatan objek dalam ruang fitur. DBSCAN dapat mengidentifikasi kelompok dengan bentuk dan ukuran yang beragam serta mengelola data noise.

Penerapan model clustering dapat ditemui dalam berbagai bidang, termasuk segmentasi pasar, analisis pola pembelian konsumen, pengelompokan dokumen teks, dan analisis citra. Keunggulan dari model clustering adalah kemampuannya untuk menghasilkan kelompok-kelompok yang tidak diketahui sebelumnya dan memberikan wawasan tentang hubungan antar-data. Meskipun ada berbagai model

clustering yang tersedia, pemilihan model yang sesuai dengan karakteristik data dan tujuan analisis sangat penting untuk mendapatkan hasil yang relevan dan bermakna.

Model clustering dalam data mining memiliki peran krusial dalam mengungkap struktur dan pola dalam dataset yang kompleks. Salah satu model clustering yang umum digunakan adalah K-Means. Dalam K-Means, data dibagi ke dalam kelompok-kelompok yang disebut cluster, dengan setiap cluster memiliki pusatnya sendiri. Algoritma ini bertujuan untuk meminimalkan varian dalam setiap cluster dan mengoptimalkan penempatan centroid. Kelebihan K-Means melibatkan kemudahan interpretasi hasilnya dan kemampuannya menangani besar dataset.

Model clustering Hierarchical Clustering membangun hirarki kelompok berdasarkan tingkat kemiripan antar-objek. Dengan pendekatan ini, objek yang mirip dikelompokkan bersama dalam tingkatan hirarki yang lebih tinggi. Hierarchical Clustering memberikan pemahaman yang baik tentang struktur hierarki dalam data, memungkinkan analisis pada berbagai tingkat resolusi. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah model clustering yang fokus pada kepadatan objek dalam ruang fitur. DBSCAN dapat mengidentifikasi kelompok dengan bentuk dan ukuran yang beragam serta mengelola data noise. Kelebihannya termasuk kemampuan untuk menangani kelompok dengan bentuk yang kompleks dan mampu menanggapi variasi kepadatan dalam dataset.

Penerapan model clustering dalam data mining sangat bervariasi, mulai dari analisis konsumen dan segmentasi pasar hingga pengelompokan dokumen teks dan identifikasi pola dalam citra. Dengan model clustering, data mining memungkinkan

penemuan pola tersembunyi dan relasi yang dapat memberikan wawasan berharga untuk pengambilan keputusan dan pemahaman yang lebih baik tentang struktur data yang ada. Pemilihan model clustering yang sesuai dengan karakteristik data dan tujuan analisis menjadi kunci untuk mendapatkan hasil yang akurat dan bermakna.

2.4. Algoritma K-Means

Algoritma K-Means Clustering adalah salah satu teknik clustering yang digunakan secara luas dalam analisis data dan machine learning [7]. Tujuan utama K-Means adalah mengelompokkan objek data ke dalam kelompok (cluster) berdasarkan kesamaan karakteristik. Algoritma ini mengasumsikan bahwa objek dalam suatu cluster memiliki kemiripan tinggi dan berbeda dengan objek di cluster lain. Proses klasifikasi dilakukan dengan menentukan pusat cluster (centroid) dan mengelompokkan objek ke dalam cluster yang memiliki centroid terdekat. Langkah awal K-Means melibatkan inisialisasi centroid secara acak atau menggunakan metode tertentu. Setelah itu, algoritma melakukan iterasi antara dua tahap utama: tahap atribusi dan tahap perbaruan centroid.

Pada tahap atribusi, setiap objek data ditempatkan ke dalam cluster yang memiliki centroid terdekat. Jarak antara objek dan centroid dihitung, dan objek ditempatkan dalam cluster dengan centroid terdekat berdasarkan perhitungan tersebut. Tahap perbaruan centroid melibatkan penghitungan pusat baru untuk setiap cluster berdasarkan rata-rata objek dalam cluster tersebut. Proses ini bertujuan untuk memperbarui lokasi centroid agar lebih merepresentasikan pusat sejati dari objek-objek dalam cluster.

Iterasi antara tahap atribusi dan perbaruan centroid terus dilakukan hingga tidak ada perubahan dalam pengelompokan atau hingga suatu kriteria konvergensi terpenuhi. Hasil akhirnya adalah pemisahan objek data ke dalam K cluster yang merepresentasikan pola dan kesamaan dalam dataset. Kelebihan algoritma K-Means mencakup kemudahan implementasi dan kecepatan komputasional yang tinggi. Meskipun demikian, K-Means memiliki beberapa kelemahan, seperti sensitivitas terhadap inisialisasi awal dan kinerja yang kurang optimal pada dataset yang memiliki bentuk cluster yang kompleks atau tidak teratur. Oleh karena itu, pemilihan jumlah cluster (K) dan pemilihan metode inisialisasi centroid memainkan peran penting dalam keberhasilan algoritma ini.

2.3.1. Uji Performa

Uji performa, atau pengujian kinerja, adalah proses evaluasi dan pengukuran efisiensi, keandalan, dan kinerja suatu sistem atau produk. Tujuan utama dari uji performa adalah untuk memastikan bahwa suatu sistem dapat beroperasi dengan baik dalam kondisi yang berbeda dan memenuhi persyaratan fungsional serta kebutuhan pengguna. Uji performa dapat diterapkan pada berbagai tingkatan, termasuk perangkat keras (hardware), perangkat lunak (software), dan sistem secara keseluruhan. Uji performa perangkat keras melibatkan penilaian terhadap respons suatu perangkat terhadap berbagai beban kerja atau penggunaan yang berbeda. Ini melibatkan pengukuran kinerja elemen-elemen seperti kecepatan pemrosesan, kapasitas penyimpanan, dan ketahanan fisik. Uji performa perangkat keras sangat penting untuk memastikan bahwa perangkat keras dapat menangani tugas-tugas yang dihadapi sehari-hari dan tetap andal selama penggunaan jangka panjang.

Uji performa perangkat lunak berkaitan dengan evaluasi kinerja suatu aplikasi atau sistem perangkat lunak. Ini mencakup pengujian fungsionalitas, respons waktu, skalabilitas, dan toleransi terhadap beban kerja yang tinggi. Uji performa perangkat lunak dapat membantu mengidentifikasi dan memperbaiki bug, memastikan kestabilan aplikasi, dan menilai sejauh mana aplikasi dapat beradaptasi dengan lingkungan yang dinamis. Uji performa sistem melibatkan pengujian elemen-elemen terpadu, seperti jaringan, server, dan aplikasi dalam suatu lingkungan produksi atau simulasi. Ini membantu dalam mengukur kemampuan sistem untuk menangani lalu lintas dan beban kerja yang beragam, serta memastikan bahwa performa keseluruhan sistem memenuhi kebutuhan pengguna dan spesifikasi yang telah ditetapkan. Penting untuk dicatat bahwa uji performa tidak hanya berfokus pada identifikasi masalah atau batasan, tetapi juga pada pengoptimalkan kinerja untuk mencapai tingkat efisiensi dan keandalan yang optimal. Dengan melakukan uji performa secara menyeluruh, organisasi dapat memastikan bahwa sistem atau produk yang dikembangkan memenuhi standar kualitas dan dapat memberikan pengalaman pengguna yang memuaskan.

2.5. Alat Bantu Program/Tools Pendukung

2.4.1. Orange

Orange adalah perangkat lunak sumber terbuka yang dirancang untuk analisis data visual dan pembelajaran mesin. Dikembangkan di University of Ljubljana, Slovenia, Orange menyediakan antarmuka pengguna grafis yang intuitif, memungkinkan pengguna dari berbagai latar belakang untuk mengakses dan menganalisis data tanpa perlu pengetahuan pemrograman yang mendalam. Salah

satu fitur utama Orange adalah keberagaman alat visual dan fungsionalitasnya, yang mencakup pembelajaran mesin, penggalian data, visualisasi, dan analisis statistik.Salah satu keunggulan Orange adalah antarmuka pengguna yang ramah pengguna dan intuitif. Dengan menggunakan pendekatan tata letak grafis, pengguna dapat membangun alur kerja analisis data dengan menggabungkan blok-blok atau widget yang mewakili berbagai operasi. Ini memberikan fleksibilitas dan kemudahan dalam merancang eksperimen dan mengatur proses analisis data

Orange menawarkan berbagai widget atau blok fungsi, yang mencakup alat-alat untuk pemrosesan data, visualisasi, ekstraksi fitur, dan pembelajaran mesin. Ini memungkinkan pengguna untuk menggabungkan berbagai teknik dan metode analisis dalam satu kerangka kerja, memungkinkan eksplorasi data yang holistik. Dalam konteks pembelajaran mesin, Orange menyediakan sejumlah algoritma pembelajaran mesin yang dapat diakses dan diterapkan dengan mudah. Ini termasuk pohon keputusan, regresi linier, clustering, dan metode-metode lainnya. Orange juga mendukung integrasi dengan bahasa pemrograman Python, memberikan pengguna kemampuan untuk menulis skrip tambahan jika diperlukan. Meskipun Orange umumnya digunakan oleh praktisi data science dan peneliti, keuniversalan dan antarmuka yang ramah pengguna membuatnya cocok untuk pemula yang ingin memahami analisis data dan pembelajaran mesin tanpa harus memiliki pengetahuan pemrograman atau matematika yang mendalam. Dengan komunitas yang aktif dan dukungan yang berkelanjutan, Orange terus berkembang sebagai perangkat lunak analisis data yang andal dan berdaya guna.

2.6. Metodologi Penelitian

2.5.1. Penelitian Terdahulu

Referensi Penelitian	1
Judul	Crop yield forecasting using data mining
Nama Penulis	Pallavi Kamath*, Pallavi Patil, Shrilatha S, Sushma, Sowmya S
Tahun	2021
Hasil	Di India, di mana sektor pertanian memegang peranan penting, memprediksi produksi pertanian menjadi tugas yang rumit dipengaruhi oleh faktor organik, ekonomi, dan musiman. Mengestimasi produksi pertanian merupakan tantangan besar untuk negara ini, terutama mengingat situasi populasi saat ini. Asumsi produksi tanaman yang dibuat jauh sebelumnya dapat membantu para petani melakukan perencanaan yang diperlukan, seperti penyimpanan dan pemasaran. Prediksi produksi tanaman melibatkan sejumlah besar data,

	menjadikannya kandidat ideal untuk metode data mining. Data mining adalah metode untuk mengumpulkan informasi yang sebelumnya tidak terlihat dari basis data yang luas. Data mining membantu dalam menganalisis pola dan karakteristik masa depan, memungkinkan perusahaan membuat keputusan yang terinformasi. Untuk wilayah tertentu, penelitian ini memberikan inspeksi cepat terhadap perkiraan hasil pertanian menggunakan pendekatan Random Forest [8].
Referensi Penelitian	2
Judul	The Application of Data Mining in Determining Timely Graduation Using the C45 Algorithm
Nama Penulis	Asro Pradipta[1], Dedy Hartama[2], Anjar Wanto[3], Saifullah[4], Jalaluddin[5]
Tahun	2019
Hasil	Penerapan data mining pada penelitian ini berkaitan dengan upaya meningkatkan angka kelulusan

	mahasiswa pada Universitas Simalungun Pematangsiantar, yang menjadi unsur penilaian akreditasi perguruan tinggi. Kelulusan tepat waktu, ditentukan oleh selesainya studi dalam waktu delapan semester atau empat tahun pada jenjang Strata 1, menjadi fokus utama. Standar kelulusan tepat waktu yang ditetapkan oleh BAN-PT mencapai 50%, dan tidak memenuhi standar tersebut dapat berdampak pada penurunan nilai akreditasi. Dalam menghadapi permasalahan ini, penelitian ini menggunakan Algoritma C4.5 untuk memprediksi kelulusan mahasiswa. Algoritma ini mengolah dataset berisi 150 data profil mahasiswa dengan label status kelulusan (benar atau terlambat) dan fitur seperti nama mahasiswa, jenis kelamin, status mahasiswa, dan IPK. Hasil dari algoritma C4.5 berupa model pohon keputusan yang dapat dengan

	mudah dianalisis, bahkan oleh orang awam. Model ini memberikan pemahaman tentang pola mahasiswa yang berpotensi lulus tepat waktu atau terlambat, sehingga langkah-langkah strategis dapat diambil untuk meningkatkan tingkat kelulusan secara efektif [9].
Referensi Penelitian	3
Judul	A clustering approach based on support vectors
Nama Penulis	Gaurav, Pawan Whig
Tahun	2022
Hasil	Kami menjelaskan teknik kernel unik untuk pengelompokan data berdasarkan deskripsi vektor dukungan data. Kernel mewakili proyeksi titik data dari ruang data ke ruang fitur berdimensi tinggi. Batas cluster ditentukan dalam ruang fitur sebagai bidang yang mewakili struktur geometris rumit dalam ruang data. Kami membangun metode

	pengelompokan dasar menggunakan representasi data geometris ini [10].
Referensi Penelitian	4
Judul	A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots
Nama Penulis	Xiaojuan Ran [1,2], Xiangbing Zhou [2,3,*], Mu Lei [4], Worawit Tepsan 1 and Wu Deng [2,5,*]
Tahun	2021
Hasil	Penelitian ini mengeksplorasi penerapan algoritma derau dalam konteks clustering, dimana algoritme derau digunakan untuk meningkatkan atribusi titik data dan hasil keluaran pengelompokan secara acak. Dalam upaya meningkatkan ketepatan dan efisiensi, penelitian ini memperkenalkan evaluasi derau sebagai komponen tambahan untuk secara otomatis menentukan jumlah klaster yang optimal untuk data tertentu, serta menginisialisasi pusat

	klaster. Pendekatan ini diharapkan dapat meningkatkan performa algoritme clustering, khususnya dalam hal inisialisasi klaster dan pemilihan jumlah klaster yang tepat untuk meminimalkan derau dan meningkatkan akurasi hasil pengelompokan [11].
Referensi Penelitian	5
Judul	ANALISIS DAN PENERAPAN ALGORITMA K-MEANS DALAM STRATEGI PROMOSI KAMPUS AKADEMI MARITIM SUAKA BAHARI
Nama Penulis	Tuti Hartati[1*], Odi Nurdiawan[2], Eko Wiyandi[3]
Tahun	2021
Hasil	Penelitian ini menerapkan metode K-Means dalam konteks analisis dan strategi promosi untuk Kampus Akademi Maritim Suaka Bahari. Metode K-Means digunakan sebagai alat analisis untuk mengelompokkan

	data terkait strategi promosi menjadi kelompok yang relevan. Dengan mengidentifikasi pola dan karakteristik dari data promosi, penelitian ini bertujuan untuk memberikan wawasan yang mendalam dalam perancangan strategi promosi yang lebih efektif dan terarah. Melalui pengaplikasian algoritma K-Means, penelitian ini berupaya menyederhanakan dan mengelompokkan informasi promosi sehingga memungkinkan pengambilan keputusan yang lebih cerdas dan berbasis data untuk meningkatkan efektivitas promosi kampus Akademi Maritim Suaka Bahari [12].
Referensi Penelitian	6
Judul	TINJAUAN PUSTAKA SISTEMATIS PADA DATA MINING: STUDI KASUS ALGORITMA K-MEANS CLUSTERING
Nama Penulis	Sekar Setyaningtyas[1)], Bangkit Indarmawan Nugroho[2)], Zaenul Arif[3)]

Tahun	2022
Hasil	Dalam penelitian ini, penerapan metode K-Means terfokus pada tinjauan pustaka sistematis dalam konteks data mining, dengan studi kasus pada algoritma K-Means Clustering. Metode ini digunakan untuk merinci secara sistematis kerangka kerja, perkembangan, dan aplikasi dari algoritma K-Means dalam literature terkait data mining. Tinjauan pustaka sistematis ini diharapkan dapat memberikan pemahaman mendalam tentang konsep, kelebihan, dan tantangan terkait penggunaan algoritma K-Means Clustering dalam konteks analisis data dan pengelompokan. Dengan demikian, penelitian ini bertujuan untuk memberikan kontribusi signifikan dalam memahami dan merinci peran serta aplikasi algoritma K-Means

	dalam kerangka kerja data mining secara umum [13].