BAB II

LANDASAN TEORI

2.1. Data Science

Data science adalah disiplin ilmu yang berfokus pada pemahaman, analisis, dan pengelolaan data untuk mendapatkan wawasan yang berharga, membuat keputusan yang informasional, dan mengidentifikasi pola atau tren yang dapat diterapkan untuk tujuan bisnis atau penelitian [1]. Disiplin ini memanfaatkan berbagai metode ilmiah, algoritma, dan sistem informasi untuk menggali pengetahuan dari data yang besar dan kompleks. Ada beberapa elemen kunci dalam data science yang mencakup pengumpulan data, preprocessing, analisis statistik, dan pengembangan model prediktif.

Pertama-tama, pengumpulan data adalah langkah awal dalam proses data science. Ini melibatkan akuisisi data dari berbagai sumber, termasuk sensor, platform online, atau basis data tradisional. Setelah pengumpulan, data kemudian harus diproses dan dibersihkan, tahap yang dikenal sebagai preprocessing. Tujuan utama dari preprocessing adalah untuk memastikan data bebas dari kesalahan atau anomali dan siap untuk analisis lebih lanjut. Analisis statistik merupakan komponen utama dalam data science yang melibatkan ekstraksi makna dari data. Ini melibatkan penggunaan berbagai teknik seperti regresi, clustering, dan pengujian hipotesis untuk mengidentifikasi pola atau hubungan dalam data. Selain itu, data science juga mencakup pengembangan model prediktif, yang melibatkan pembuatan model matematis atau statistik untuk memprediksi hasil di masa depan berdasarkan data historis.

Data science memiliki aplikasi yang luas di berbagai industri, termasuk bisnis, kesehatan, finansial, dan penelitian ilmiah. Kemajuan teknologi dalam bidang ini telah memungkinkan organisasi untuk mengoptimalkan operasi mereka, membuat keputusan yang lebih cerdas, dan mengidentifikasi peluang baru. Dengan berkembangnya kompleksitas dan volume data yang dihasilkan oleh dunia digital, peran data science menjadi semakin krusial dalam membentuk masa depan inovatif dan efisien.

2.2. Machine Learning

Machine learning (ML) adalah cabang dari kecerdasan buatan yang fokus pada pengembangan algoritma dan model komputer yang memungkinkan sistem untuk belajar dari data [2] [3]. Berbeda dengan pendekatan tradisional yang mengandalkan pemrograman eksplisit, machine learning memanfaatkan kemampuan sistem untuk belajar pola dan informasi dari data masukan [4]. Ini mencakup sejumlah teknik seperti supervised learning, unsupervised learning, dan reinforcement learning. Dalam supervised learning, model dilatih menggunakan data yang sudah dilabeli, di mana input dan outputnya sudah diketahui [5]. Tujuannya adalah untuk membuat prediksi atau klasifikasi pada data baru yang belum terlihat. Sementara itu, unsupervised learning menghadapi data yang tidak dilabeli dan mencoba mengidentifikasi pola atau struktur di dalamnya. Clustering dan dimensionality reduction adalah contoh dari tugas yang dapat dipecahkan dengan unsupervised learning.

Reinforcement learning, di sisi lain, melibatkan sistem belajar melalui interaksi dengan lingkungan. Agent belajar dari tindakan yang diambilnya dan

mengoptimalkan keputusan yang diambil untuk mencapai tujuan tertentu. Hal ini sering diterapkan dalam konteks permainan, kontrol robot, dan navigasi otomatis. Penerapan machine learning merambah berbagai bidang, seperti pengenalan wajah, prediksi pasar keuangan, diagnosa medis, dan bahkan pengembangan mobil otonom. Keberhasilan model machine learning sangat bergantung pada kualitas dan representativitas data yang digunakan dalam pelatihan. Seiring dengan kemajuan teknologi, machine learning terus mengalami perkembangan untuk meningkatkan presisi, interpretabilitas, dan skala penerapannya dalam berbagai industri.

2.2.1. Model Reinforcement Learning

Model Reinforcement Learning (RL) adalah pendekatan dalam machine learning di mana agen belajar berinteraksi dengan lingkungannya dan mengambil keputusan untuk mencapai tujuan tertentu. Dalam RL, agen memperoleh pengetahuan melalui percobaan dan kesalahan, mendapatkan umpan balik positif atau negatif berdasarkan tindakannya. Salah satu konsep inti dalam RL adalah kebijaksanaan (policy), yang mewakili strategi agen untuk memilih tindakan berdasarkan keadaan lingkungan. Model RL terdiri dari beberapa elemen kunci, termasuk agen, lingkungan, keadaan, tindakan, kebijaksanaan, dan hadiah (reward). Agen merupakan entitas yang beroperasi di dalam lingkungan dan bertanggung jawab untuk membuat keputusan. Lingkungan mencakup konteks di mana agen beroperasi. Keadaan adalah representasi dari situasi atau kondisi lingkungan pada suatu waktu tertentu. Tindakan adalah langkah-langkah yang dapat diambil oleh agen untuk memengaruhi lingkungan. Kebijaksanaan adalah aturan atau strategi yang digunakan agen untuk memilih tindakan berdasarkan keadaan tertentu.

Pada dasarnya, proses RL melibatkan agen yang mengamati lingkungan, memutuskan tindakan, dan menerima umpan balik berupa hadiah atau hukuman. Tujuan agen adalah memaksimalkan jumlah hadiah yang diterima dari lingkungan seiring waktu. Algoritma dalam RL, seperti Q-learning dan metode berbasis kebijaksanaan seperti metode kebijaksanaan berbasis keuntungan (policy gradient methods), digunakan untuk melatih agen secara efektif dan mengoptimalkan strategi kebijaksanaannya. Model RL telah berhasil diterapkan dalam berbagai konteks, termasuk pengembangan permainan komputer, navigasi robot, dan pengelolaan sumber daya. Keterampilan adaptasi agen untuk beroperasi di lingkungan dinamis membuat RL menjadi pilihan yang kuat untuk masalah-masalah yang tidak dapat diselesaikan dengan metode tradisional machine learning. Perkembangan terbaru dalam RL mencakup integrasi dengan deep learning, dikenal sebagai deep reinforcement learning, yang meningkatkan kemampuan model untuk menangani masalah kompleks dan non-linear dengan representasi fitur yang lebih canggih.

2.2.2. Model Unsupervised Learning

Model Unsupervised Learning merupakan paradigma dalam machine learning yang berkaitan dengan pemrosesan data di mana algoritma diajarkan untuk mengidentifikasi pola atau struktur tanpa memerlukan anotasi label pada data pelatihan. Dalam unsupervised learning, tujuan utama adalah mengekstraksi informasi tersembunyi atau menciptakan representasi data yang bermakna tanpa panduan langsung. Terdapat beberapa pendekatan utama dalam unsupervised learning, termasuk clustering dan dimensionality reduction. Pada clustering,

algoritma berusaha untuk mengelompokkan data ke dalam kategori atau kelompok yang serupa berdasarkan kesamaan fitur atau karakteristik tertentu. K-means clustering dan hierarchical clustering adalah contoh teknik yang umum digunakan dalam konteks ini. Hasil dari proses ini adalah pembentukan kelompok yang dapat mengungkap struktur alami atau hubungan di antara titik data. Sementara itu, dimensionality reduction bertujuan untuk mengurangi jumlah fitur atau variabel dalam suatu dataset, tetapi dengan tetap mempertahankan sebanyak mungkin informasi yang relevan. Principal Component Analysis (PCA) adalah salah satu metode dimensionality reduction yang umum digunakan, di mana dimensi yang kurang penting dari data diidentifikasi dan dieliminasi, memungkinkan representasi data yang lebih ringkas.

Unsupervised learning digunakan dalam berbagai konteks, termasuk pengenalan pola, analisis klaster pelanggan, dan pemrosesan bahasa alami. Keberhasilannya tergantung pada kemampuan algoritma untuk menemukan struktur atau pola yang bermanfaat secara intrinsik dalam data, tanpa bimbingan label. Penerapan unsupervised learning terus berkembang, dan integrasinya dengan deep learning, seperti dalam autoencoders, semakin meningkatkan kapabilitasnya untuk mengolah dan memahami data yang semakin kompleks.

2.2.3. Model Supervised Learning

Model Supervised Learning adalah pendekatan dalam machine learning di mana algoritma dilatih menggunakan dataset yang sudah dilabeli, di mana setiap contoh data memiliki pasangan input dan output yang sesuai. Tujuan utama dari supervised learning adalah untuk membuat model yang dapat mempelajari

hubungan atau pola di antara input dan output, sehingga model dapat memberikan prediksi atau klasifikasi yang akurat untuk data baru yang belum pernah dilihat. Proses pelatihan dalam supervised learning melibatkan dua tahap utama: tahap pelatihan dan tahap pengujian. Selama tahap pelatihan, model belajar dari contoh-contoh yang sudah diberi label untuk menyesuaikan parameter internalnya sehingga dapat memberikan prediksi yang sesuai. Tahap pengujian kemudian melibatkan pengujian model pada data yang tidak pernah dilihat selama pelatihan untuk mengevaluasi kinerjanya. Metrik seperti akurasi, presisi, recall, dan F1-score sering digunakan untuk mengukur sejauh mana model dapat memberikan hasil yang benar.

Supervised learning memiliki berbagai aplikasi, termasuk klasifikasi dan regresi. Dalam klasifikasi, model memprediksi kategori atau label kelas tertentu untuk suatu input, sementara dalam regresi, model memprediksi nilai kontinu. Contoh penerapan supervised learning melibatkan pengenalan wajah, diagnosa medis, dan prediksi harga saham. Keunggulan utama dari supervised learning adalah kemampuannya untuk membuat prediksi atau keputusan yang akurat dengan menggunakan data yang sudah ada. Namun, ketergantungan pada dataset yang sudah dilabeli dapat menjadi tantangan, terutama dalam skenario di mana label data sulit atau mahal untuk diperoleh. Meskipun demikian, supervised learning tetap menjadi pendekatan yang sangat berguna dan banyak digunakan dalam berbagai domain, terutama dengan kemajuan teknologi yang memungkinkan pengumpulan dan anotasi data yang lebih efisien.

2.3. Model Klasifikasi

Model klasifikasi adalah salah satu bentuk pendekatan dalam machine learning yang fokus pada identifikasi dan alokasi suatu instance data ke dalam salah satu dari beberapa kategori atau kelas yang telah ditentukan [6] [7]. Dalam konteks ini, algoritma dilatih menggunakan dataset yang sudah dilabeli, di mana setiap contoh data memiliki kelas atau kategori yang telah ditetapkan sebelumnya. Tujuan utama dari model klasifikasi adalah untuk membuat suatu kebijaksanaan (policy) atau fungsi keputusan yang dapat memberikan prediksi kelas yang akurat untuk data baru yang belum pernah dilihat. Beberapa algoritma klasifikasi yang umum digunakan melibatkan pembelajaran dari data melibatkan Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), dan Logistic Regression. Decision Trees bekerja dengan cara membuat pohon keputusan yang membagi data berdasarkan fitur-fitur tertentu untuk mencapai hasil klasifikasi. SVM mencari batas keputusan yang optimal antara kelas-kelas yang berbeda, sementara k-NN memprediksi kelas suatu instance berdasarkan mayoritas kelas dari tetangga terdekatnya. Logistic Regression, meskipun namanya, sebenarnya digunakan untuk masalah klasifikasi, memprediksi probabilitas bahwa suatu instance termasuk dalam satu kelas tertentu.

Model klasifikasi memiliki berbagai aplikasi yang luas, termasuk pengenalan pola, identifikasi spam email, dan diagnosis medis [8]. Evaluasi performa model klasifikasi dapat dilakukan menggunakan metrik seperti akurasi, presisi, recall, F1-score, dan area di bawah kurva Receiver Operating Characteristic (ROC-AUC). Pengembangan model klasifikasi terus menjadi fokus penelitian dan aplikasi praktis

dalam berbagai industri, memungkinkan pengambilan keputusan yang lebih efisien dan akurat dalam berbagai konteks.

2.4. Algoritma Naïve Bayes

Algoritma Naïve Bayes adalah salah satu algoritma klasifikasi yang didasarkan pada teorema probabilitas Bayes, yang menggunakan asumsi "naïve" bahwa setiap fitur dalam data adalah independen satu sama lain, meskipun mungkin ada ketergantungan sebenarnya di antara mereka [9] [10]. Meskipun asumsi ini terkadang tidak realistis di dunia nyata, Naïve Bayes seringkali memberikan kinerja yang baik dan komputasi yang cepat, terutama dalam konteks klasifikasi teks dan analisis sentimen. Algoritma Naïve Bayes bekerja dengan memanfaatkan teorema Bayes untuk menghitung probabilitas kelas tertentu berdasarkan fitur-fitur yang teramati [11]. Rumus umumnya dinyatakan sebagai berikut.

$$\mathbf{P\ (A\ |\ B):}\ \frac{\boldsymbol{P\ (B\ |A)\ P\ (A)}}{\boldsymbol{P\ (B)}}$$

Information:

A	:	hipotesis data A (kelas tertentu)
B	:	data dengan kelas yang tidak diketahui
P (A | B)	:	Probabilitas hipotesis berdasarkan kondisi B
P (A)	:	Kemungkinan hipotesis A
P (B | A)	:	Probabilitas B ketika kondisi A
P (B)	:	Probabilitas

Terdapat beberapa variasi dari algoritma Naïve Bayes, yang paling umum termasuk Naïve Bayes Gaussian, Multinomial, dan Bernoulli, yang masing-masing cocok untuk jenis data tertentu. Naïve Bayes Gaussian umumnya digunakan untuk

data yang terdistribusi secara normal, sementara Naïve Bayes Multinomial lebih sesuai untuk data kategorikal, seperti data teks yang direpresentasikan sebagai model Bag-of-Words.

Algoritma Naïve Bayes sering diaplikasikan pada tugas klasifikasi teks, seperti kategorisasi dokumen atau analisis sentimen. Keunggulan utamanya terletak pada sederhananya, kemampuan untuk menangani dimensi fitur yang tinggi, dan kecepatan komputasi yang baik. Meskipun asumsi independensi fitur seringkali tidak sepenuhnya memenuhi realitas, Naïve Bayes tetap menjadi pilihan yang kuat dalam banyak aplikasi machine learning.

2.4.1. Uji Performa

Confusion Matrix adalah alat evaluasi yang umum digunakan dalam analisis performa model klasifikasi. Dalam konteks machine learning, Confusion Matrix membantu menggambarkan kinerja model dengan membandingkan prediksi yang dibuat oleh model terhadap nilai sebenarnya dari data yang diamati. Terdiri dari empat elemen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN), Confusion Matrix memungkinkan perhitungan metrik evaluasi seperti akurasi, presisi, recall, spesifisitas, dan F1-score. Dengan menyajikan informasi ini dalam bentuk matriks, Confusion Matrix memberikan wawasan yang jelas tentang kemampuan model dalam mengklasifikasikan data, membantu pemahaman yang lebih mendalam terkait kekuatan dan kelemahan model tersebut.

Tabel 2.3. 1. Confusion Matrix Metode Naïve Bayes

		Kelas Prediksi	
	Kelas	**Benar**	**Salah**
Kelas Atribut	Benar	True Positive (TP)	False Positive (FP)
	Salah	False Negative (FN)	True Negative (TN)

Dimana tabel ini berisi:

- TP (True Positive), yaitu jumlah data positif yang memiliki nilai benar.
- TN (True Negative), yaitu jumlah data negatif yang memiliki nilai benar.
- FN (False Negative), yaitu jumlah data negatif tetapi yang memiliki nilai salah.
- FP (False Positive), yaitu jumlah data yang positif tetapi yang memiliki nilai salah.

$$\boldsymbol{Acuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FN + FP}}$$

$$\boldsymbol{Presisi} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

2.5. Alat Bantu Program/Tools Pendukung

2.5.1. Orange

Orange adalah platform open-source yang kuat untuk analisis data dan machine learning, yang dirancang untuk mempermudah penggunaan dan

pemahaman algoritma-algoritma kompleks. Dikembangkan di University of Ljubljana di Slovenia, Orange menyediakan antarmuka grafis yang intuitif dan ramah pengguna, membuatnya cocok baik untuk pemula maupun para profesional di bidang data science. Salah satu fitur kunci dari Orange adalah visual programming, di mana pengguna dapat membangun alur kerja analisis data dan machine learning menggunakan blok-blok grafis yang dapat dihubungkan. Ini memungkinkan para pengguna untuk dengan mudah membuat dan menyesuaikan algoritma tanpa memerlukan pengetahuan mendalam tentang pemrograman. Selain itu, Orange menyediakan berbagai komponen visual untuk eksplorasi data, preprocessing, pemodelan, dan evaluasi, yang semuanya dapat diakses dan digunakan dalam satu antarmuka.

Platform ini juga menawarkan beragam algoritma machine learning termasuk klasifikasi, regresi, clustering, dan reduksi dimensi. Orange mendukung integrasi dengan Python, sehingga pengguna yang lebih berpengalaman dapat memanfaatkan kekuatan dan fleksibilitas bahasa pemrograman ini untuk mengembangkan algoritma kustom atau melakukan tugas analisis tambahan. Selain itu, Orange memiliki komunitas pengguna yang aktif yang berbagi tutorial, sumber daya, dan ekstensi tambahan, memperkaya pengalaman pengguna dan mendukung kolaborasi. Kehadiran visual programming dan fleksibilitas dalam penggunaan membuat Orange menjadi pilihan yang populer untuk pengajaran di bidang data science dan machine learning di berbagai institusi pendidikan. Secara keseluruhan, Orange memfasilitasi akses mudah ke teknik-teknik machine learning dan analisis data tanpa harus menyerah pada kemampuan atau kekuatan algoritma. Platform ini

terus berkembang dan menyediakan solusi yang efektif untuk berbagai kebutuhan analisis data.

Orange, sebuah platform open-source untuk analisis data dan machine learning, telah meraih popularitas besar di kalangan para profesional data science dan peneliti. Aplikasi Orange menemukan berbagai penggunaan yang luas dalam konteks machine learning. Salah satu aplikasi utamanya adalah di bidang klasifikasi dan prediksi, di mana pengguna dapat dengan mudah membangun model untuk memprediksi kategori atau label dari data baru menggunakan berbagai algoritma seperti Decision Trees, Support Vector Machines, dan Random Forests. Hal ini sangat bermanfaat dalam pengambilan keputusan berdasarkan pola data historis. Selain itu, Orange juga memainkan peran penting dalam analisis teks dan sentimen. Dengan alat analisis teksnya, pengguna dapat mengekstrak fitur dari dokumen dan menganalisis sentimen dari data teks, membuka pintu untuk pemahaman lebih dalam tentang respons pengguna, tinjauan produk, atau komentar di media sosial. Aplikasi ini memberikan wawasan berharga dalam strategi bisnis dan pemasaran.

Aplikasi Orange tidak hanya terbatas pada model machine learning, tetapi juga mencakup pengolahan data dan preprocessing. Dengan alat pemrosesan data, pengguna dapat membersihkan dan menyiapkan data sebelumnya, seperti mengatasi nilai yang hilang atau melakukan normalisasi, memastikan bahwa data yang digunakan untuk melatih model dalam kondisi yang optimal. Visualisasi dan interpretasi model juga menjadi kekuatan Orange. Pengguna dapat memvisualisasikan model dan hasil analisis data mereka, membuatnya lebih mudah untuk menginterpretasikan dan berkomunikasi temuan mereka kepada pemangku

kepentingan. Ini membantu memperjelas kompleksitas model dan memudahkan proses pengambilan keputusan.

Dalam dunia pendidikan, Orange telah menjadi alat yang sangat efektif untuk mengajarkan konsep-konsep machine learning. Antarmuka grafis yang intuitif memudahkan siswa dan pemula untuk memahami dasar-dasar machine learning tanpa harus menguasai keterampilan pemrograman yang rumit. Secara keseluruhan, Orange menawarkan keberagaman alat dan fungsionalitas yang membuatnya menjadi pilihan yang kuat dalam proyek-proyek machine learning dan analisis data, memungkinkan para pengguna untuk menjelajahi, memahami, dan menerapkan konsep-konsep ini dengan lebih mudah.

2.6. Metodologi Penelitian

2.6.1. Penelitian Terdahulu

Tabel 2.6. 1. Penelitian Terdahulu

Referensi Penelitian	1
Judul	Comparison Of The C.45 And Naive Bayes Algorithms To Predict Diabetes
Nama	Alam1)*, Divi Adiffia Freza Alana.2), Christina Juliane3)
Tahun	2023
Hasil	Metode Naive Bayes telah berhasil diimplementasikan secara efektif dalam mendeteksi penderita diabetes, menghasilkan tingkat akurasi sebesar

	90%. Dengan memanfaatkan probabilitas dan asumsi "naive" bahwa fitur-fitur yang diamati bersifat independen, model ini mampu melakukan klasifikasi dengan tepat berdasarkan informasi medis yang diberikan. Penerapan metode ini dalam konteks deteksi diabetes dapat memberikan kontribusi signifikan dalam membantu identifikasi dini dan manajemen penyakit, sehingga memungkinkan pemberian perawatan yang lebih efektif kepada individu yang berisiko atau telah terdiagnosis menderita diabetes alam [12].
Referensi Penelitian	2
Judul	Performance of Various Naïve Bayes Using GridSearch Approach In Phishing Email Dataset
Nama	Rizki Rahman1)*, Ferian Fauzi Abdulloh2)
Tahun	2023

Hasil	Penerapan metode Naive Bayes dalam membandingkan empat varian Naive Bayes untuk mengklasifikasikan email phishing telah membuktikan keefektifannya, dengan hasil akurasi mencapai sekitar 97%. Proses ini melibatkan tahap pra-pemrosesan data yang komprehensif, di mana email phishing dikumpulkan, dibersihkan, dan diubah menjadi fitur numerik yang sesuai. Keempat varian Naive Bayes yang dievaluasi mungkin melibatkan variasi dalam penanganan probabilitas dan asumsi independensi fitur. Tingkat akurasi yang tinggi ini menunjukkan bahwa metode Naive Bayes, dengan pendekatan pra-pemrosesan yang cermat, dapat menjadi pilihan yang sangat efisien dan andal dalam mengidentifikasi potensi ancaman phishing pada email [13].