BAB II

LANDASAN TEORI

2.1. Data Science

Data Science merupakan disiplin ilmu interdisipliner yang melibatkan pengumpulan, analisis, interpretasi, dan presentasi data untuk mendapatkan wawasan dan mendukung pengambilan keputusan. Bidang ini menggabungkan elemen-elemen dari statistika, matematika, ilmu komputer, dan domain pengetahuan khusus untuk mengungkap pola, tren, dan pengetahuan yang terkandung dalam data. Data Science berperan penting dalam mengubah cara kita memahami dunia, melakukan bisnis, dan mengatasi tantangan di berbagai sektor. Salah satu tahap awal dalam perjalanan Data Science adalah pengumpulan data. Ini melibatkan identifikasi, pembersihan, dan integrasi data dari berbagai sumber. Kemudian, analisis data menjadi fokus utama, dan statistika serta teknik pembelajaran mesin digunakan untuk mengeksplorasi dan menganalisis data tersebut. Data Science juga memanfaatkan visualisasi data untuk menyajikan informasi secara jelas dan mudah dimengerti.

Model prediktif adalah salah satu aspek kunci dari Data Science. Dengan menggunakan algoritma pembelajaran mesin, para praktisi Data Science dapat membangun model untuk membuat prediksi atau klasifikasi berdasarkan data historis. Pengembangan model ini memerlukan pemahaman mendalam tentang algoritma dan pemilihan fitur yang sesuai agar model dapat memberikan hasil yang akurat. Data Science juga melibatkan penggunaan teknologi dan alat analisis data seperti Python, R, dan platform seperti Apache Hadoop atau Spark. Penggunaan

teknologi ini memungkinkan analisis data yang cepat dan efisien, terutama ketika menghadapi volume data yang besar.

Keberhasilan Data Science tidak hanya terletak pada teknik dan algoritma yang digunakan, tetapi juga pada interpretasi hasilnya. Pengambilan keputusan yang tepat memerlukan pemahaman kontekstual dan kemampuan untuk mengkomunikasikan temuan secara efektif kepada pemangku kepentingan yang mungkin tidak memiliki latar belakang teknis. Data Science memiliki aplikasi yang luas, mulai dari prediksi pergerakan pasar keuangan, pengoptimalan rantai pasokan, hingga pengembangan obat baru dalam industri farmasi. Di era digital ini, Data Science juga menjadi kunci untuk memahami perilaku pengguna online, memberikan rekomendasi personalisasi, dan meningkatkan pengalaman pengguna.

Dengan terus berkembangnya teknologi dan ketersediaan data yang semakin besar, Data Science menjadi semakin penting dalam memberikan wawasan yang mendalam dan bernilai tambah di berbagai sektor. Menerapkan prinsip-prinsip Data Science dapat memberikan keunggulan kompetitif dan membantu organisasi membuat keputusan yang lebih informasional dan efektif.

2.2. Machine Learning

Machine Learning (ML), atau pembelajaran mesin, adalah cabang dari kecerdasan buatan yang memungkinkan sistem komputer untuk belajar dan meningkatkan kinerjanya tanpa di-program secara eksplisit. Konsep dasar di balik Machine Learning adalah memberikan kemampuan kepada sistem komputer untuk mengenali pola dalam data, membuat prediksi atau keputusan, dan secara otomatis meningkatkan kinerjanya seiring waktu berdasarkan pengalaman. Ada beberapa

jenis utama dari Machine Learning, termasuk supervised learning, unsupervised learning, dan reinforcement learning. Pada supervised learning, model dilatih menggunakan dataset yang berisi pasangan input dan output, sehingga model dapat memahami hubungan antara keduanya dan dapat digunakan untuk membuat prediksi pada data baru. Di sisi lain, unsupervised learning tidak memiliki output yang ditentukan sebelumnya, dan model diberi tugas untuk mengeksplorasi struktur dan pola dalam data tanpa bimbingan eksternal. Reinforcement learning melibatkan agen yang belajar melalui tindakan dan umpan balik dari lingkungan, dengan tujuan untuk memaksimalkan hadiah atau nilai tertentu.

Algoritma Machine Learning menggunakan berbagai teknik dan model untuk memproses data. Contoh algoritma termasuk Decision Trees, Support Vector Machines, Neural Networks, dan K-Nearest Neighbors. Masing-masing algoritma memiliki kelebihan dan kekurangan tergantung pada jenis tugas dan karakteristik data yang dihadapi. Keberhasilan Machine Learning sangat bergantung pada kualitas dan kuantitas data yang digunakan untuk pelatihan model. Data yang baik memberikan model kerangka kerja yang kokoh dan memungkinkan pembelajaran yang lebih efektif. Proses pelatihan melibatkan pengoptimalan parameter model untuk mencapai kinerja yang maksimal, yang dapat melibatkan tuning, validasi silang, dan seleksi fitur.

Aplikasi Machine Learning melibatkan berbagai industri, termasuk pengenalan wajah, diagnosa medis, analisis risiko keuangan, dan kendaraan otonom. Dalam konteks pengenalan pola, Machine Learning dapat membantu mengidentifikasi dan mengklasifikasikan objek dalam gambar atau suara,

sementara dalam bidang kesehatan, dapat digunakan untuk memprediksi penyakit atau merancang terapi yang disesuaikan. Meskipun Machine Learning telah mencapai kemajuan yang signifikan, masih ada tantangan yang perlu diatasi, termasuk interpretasi model, fairness, dan keamanan. Interpretabilitas model menjadi penting untuk memahami alasan di balik keputusan model, sementara fairness menjamin bahwa model tidak menghasilkan diskriminasi atau bias. Keamanan menjadi faktor kritis, terutama dalam konteks model yang digunakan untuk membuat keputusan penting atau sensitif.

Machine learning (ML) memiliki keunggulan dan kelemahan yang perlu dipertimbangkan dalam penerapannya. Salah satu keunggulan utama dari machine learning adalah kemampuannya untuk mengekstrak pola dan informasi dari data besar dengan kecepatan yang jauh lebih tinggi dibandingkan dengan metode konvensional. ML juga mampu memproses data yang sangat kompleks dan menemukan hubungan yang mungkin sulit diidentifikasi oleh manusia. Keunggulan lainnya adalah kemampuan machine learning untuk belajar dari pengalaman dan meningkatkan kinerjanya seiring waktu. Dengan menggunakan algoritma pembelajaran, sistem ML dapat menyesuaikan diri terhadap perubahan dalam lingkungan atau data masukan, sehingga meningkatkan akurasi dan efisiensinya.

Namun, machine learning juga memiliki beberapa kelemahan. Salah satu kelemahannya adalah ketergantungan pada data yang berkualitas baik. Model machine learning hanya sebaik data yang digunakan untuk melatihnya, dan jika data yang digunakan bermasalah atau tidak representatif, hasilnya mungkin tidak akurat atau dapat menyimpang. Selain itu, interpretasi model ML seringkali sulit dipahami

oleh manusia. Beberapa model, terutama yang kompleks seperti deep learning, dapat dianggap sebagai "black box" di mana sulit untuk menjelaskan secara tepat bagaimana suatu keputusan atau prediksi dibuat. Hal ini dapat menimbulkan kekhawatiran terkait dengan keamanan dan etika, terutama jika model tersebut digunakan dalam konteks yang memiliki dampak besar, seperti di bidang kesehatan atau hukum. Secara keseluruhan, meskipun machine learning memiliki keunggulan yang signifikan dalam mengolah data dan membuat prediksi kompleks, pemahaman yang baik tentang kelemahan dan tantangan yang terkait diperlukan agar dapat mengoptimalkan penerapannya dalam berbagai konteks.

2.2.1. Model Reinforcement Learning

Model Reinforcement Learning (RL) pada Machine Learning adalah paradigma pembelajaran mesin di mana agen belajar untuk berinteraksi dengan lingkungan sekitarnya dengan melakukan tindakan dan menerima umpan balik dalam bentuk hadiah atau hukuman. Tujuan utama dari Reinforcement Learning adalah untuk mengembangkan strategi atau kebijakan yang memungkinkan agen untuk memaksimalkan total hadiah yang diterima seiring waktu. Salah satu konsep kunci dalam RL adalah agen, entitas yang melakukan tindakan dalam lingkungan. Lingkungan, di sisi lain, adalah ruang di mana agen beroperasi dan di mana tindakan dan perubahan terjadi. Agar agen dapat membuat keputusan yang cerdas, terdapat fungsi nilai atau kebijakan yang harus dipelajari dan diperbarui sepanjang waktu. Fungsi nilai mengukur keuntungan jangka panjang dari suatu keadaan, sementara kebijakan menentukan tindakan yang harus diambil dalam keadaan tertentu.

Pembelajaran dalam Reinforcement Learning melibatkan dua proses utama: eksplorasi dan eksploitasi. Eksplorasi adalah ketika agen mencoba tindakan yang belum pernah diambil sebelumnya untuk mengeksplorasi lingkungan dan memahami lebih banyak tentang konsekuensi tindakan tersebut. Di sisi lain, eksploitasi adalah ketika agen memilih tindakan yang telah memberikan hasil positif di masa lalu berdasarkan pengetahuan yang sudah dimilikinya. Model Reinforcement Learning dapat diterapkan dalam berbagai konteks, termasuk robotika, permainan video, dan pengelolaan sumber daya. Misalnya, dalam pengembangan robotika, RL digunakan untuk mengajarkan robot bergerak dan berinteraksi dengan objek di sekitarnya. Dalam konteks permainan video, RL telah mencapai kesuksesan besar, terutama dengan pencapaian AlphaGo yang menggunakan RL untuk memenangkan pertandingan Go melawan pemain manusia berpengalaman.

Tantangan utama dalam RL termasuk menangani masalah eksplorasi-eksploitasi, mengelola trade-off antara kecepatan pembelajaran dan ketepatan, serta memastikan kebijakan yang dihasilkan aman dan etis. Selain itu, RL juga memerlukan pemodelan yang baik dari lingkungan dan sistem hadiah agar agen dapat belajar secara efektif. Meskipun masih ada beberapa tantangan yang perlu diatasi, Model Reinforcement Learning menjanjikan kemampuan untuk menghasilkan agen yang mampu mengambil keputusan cerdas dalam konteks dinamis dan kompleks. Dengan terus berlanjutnya penelitian dan pengembangan dalam RL, kita dapat mengharapkan aplikasi yang semakin luas di berbagai industri dan situasi kehidupan nyata.

2.2.2. Model Unsupervised Learning

Model Unsupervised Learning pada Machine Learning adalah paradigma pembelajaran mesin di mana algoritma dirancang untuk mengeksplorasi struktur dan pola dalam data tanpa memerlukan label atau output yang telah ditentukan sebelumnya. Dalam Unsupervised Learning, algoritma ditugaskan untuk mengidentifikasi hubungan dan karakteristik dalam dataset tanpa petunjuk eksternal, memungkinkan algoritma untuk mengeksplorasi dan menemukan pola yang mungkin tidak terlihat secara langsung. Salah satu tipe utama dari Unsupervised Learning adalah Clustering, di mana algoritma berusaha untuk mengelompokkan data ke dalam kategori atau cluster berdasarkan kemiripan atau karakteristik tertentu. Metode seperti K-Means Clustering dan Hierarchical Clustering digunakan untuk memisahkan data ke dalam kelompok yang sesuai. Ini berguna untuk mengidentifikasi pola alami atau segmentasi yang mungkin ada dalam data.

Dimensi Reduksi adalah konsep lain dalam Unsupervised Learning, yang bertujuan untuk mengurangi kompleksitas data dengan mempertahankan informasi yang paling relevan. Analisis faktor, analisis komponen utama (PCA), dan t-SNE (t-distributed stochastic neighbor embedding) adalah teknik yang digunakan untuk mencapai reduksi dimensi. Ini membantu dalam memahami struktur data yang lebih sederhana dan lebih mudah diinterpretasikan. Generative Models adalah jenis lain dari Unsupervised Learning yang berfokus pada pembangkitan data baru yang mirip dengan data pelatihan. Salah satu contoh paling terkenal adalah Variational Autoencoders (VAE) dan Generative Adversarial Networks (GAN), yang

digunakan untuk membuat data sintetis yang memiliki distribusi serupa dengan data pelatihan. Ini memiliki aplikasi luas dalam pembangkitan konten kreatif, seperti gambar atau teks.

Unsupervised Learning dapat diterapkan dalam berbagai bidang, termasuk pengolahan bahasa alami, pengelompokan konsumen, dan analisis anomali. Dalam konteks pengolahan bahasa alami, model dapat digunakan untuk mengelompokkan kata-kata yang sering muncul bersama atau mengidentifikasi topik dalam teks. Dalam pengelompokan konsumen, model dapat membantu mengidentifikasi kelompok pelanggan berdasarkan pola pembelian mereka. Analisis anomali bertujuan untuk mendeteksi perilaku atau data yang tidak biasa atau mencurigakan dalam suatu sistem. Tantangan utama dalam Unsupervised Learning melibatkan penentuan metrik evaluasi yang tepat dan interpretasi hasil. Tanpa label, mengevaluasi kualitas pembelajaran dapat menjadi lebih subjektif dan memerlukan pemahaman yang mendalam tentang domain masalah. Memahami interpretasi hasil juga menjadi tantangan, terutama ketika algoritma mengidentifikasi pola yang mungkin sulit dijelaskan dalam kata-kata manusia. Meskipun tantangan tersebut, Unsupervised Learning memiliki potensi untuk memberikan wawasan yang berharga dan menarik dari data yang belum terduga. Dengan terus berkembangnya teknik dan metode di bidang ini, Unsupervised Learning terus memainkan peran penting dalam mengeksplorasi dan memahami data tanpa bimbingan eksternal.

2.2.3. Model Supervised Learning

Model Supervised Learning pada Machine Learning adalah paradigma pembelajaran mesin di mana algoritma dilatih menggunakan dataset yang berisi

pasangan input dan output yang telah diberi label. Tujuan utama dari Supervised Learning adalah untuk mengembangkan model atau fungsi yang dapat memetakan input ke output dengan akurasi yang tinggi. Dalam konteks ini, input berperan sebagai fitur atau variabel independen, sedangkan output adalah variabel dependen atau target yang ingin diprediksi. Proses pelatihan dalam Supervised Learning dimulai dengan menyajikan dataset pelatihan ke model, di mana algoritma belajar untuk mengenali pola dan hubungan antara input dan output. Selama pelatihan, parameter model diubah sedemikian rupa sehingga output yang dihasilkan oleh model semakin mendekati nilai yang diinginkan. Tujuan akhirnya adalah menghasilkan model yang dapat membuat prediksi akurat terhadap data yang belum pernah dilihat sebelumnya.

Ada berbagai jenis tugas dalam Supervised Learning, termasuk klasifikasi dan regresi. Dalam klasifikasi, model bertugas untuk mengkategorikan input ke dalam kelas atau label yang telah ditentukan. Contoh tugas klasifikasi termasuk identifikasi email spam atau pengenalan digit tulisan tangan. Regresi, di sisi lain, melibatkan prediksi nilai kontinu berdasarkan input, seperti memprediksi harga rumah berdasarkan fitur-fitur tertentu. Algoritma populer dalam Supervised Learning termasuk Decision Trees, Support Vector Machines (SVM), Neural Networks, dan algoritma k-Nearest Neighbors (k-NN). Keberhasilan model Supervised Learning sangat bergantung pada kualitas dataset pelatihan dan pemilihan algoritma yang paling sesuai dengan karakteristik data.

Evaluasi model Supervised Learning melibatkan penggunaan metrik performa seperti akurasi, presisi, recall, F1-score, dan area di bawah kurva ROC

(AUC-ROC). Metrik ini memberikan gambaran tentang sejauh mana model dapat memprediksi dengan tepat dan efisien. Selain itu, validasi silang dan teknik evaluasi lainnya digunakan untuk memastikan bahwa model dapat menggeneralisasi dengan baik pada data yang belum pernah dilihat sebelumnya. Supervised Learning memiliki aplikasi yang luas dalam berbagai bidang, termasuk pengenalan wajah, pengolahan bahasa alami, analisis risiko keuangan, dan diagnosis medis. Dalam pengenalan wajah, model dapat dilatih untuk mengenali individu berdasarkan fitur-fitur wajah mereka. Dalam pengolahan bahasa alami, model dapat memahami dan merespons teks atau ucapan dengan akurasi tinggi. Dalam analisis risiko keuangan, model dapat digunakan untuk memprediksi kemungkinan gagalnya suatu investasi. Di bidang medis, model dapat membantu dalam mendiagnosis penyakit berdasarkan gejala dan data medis. Meskipun Supervised Learning telah mencapai banyak keberhasilan, tantangan yang dihadapi melibatkan pemilihan fitur yang tepat, mengelola ketidakseimbangan kelas, dan penanganan overfitting atau underfitting. Terus berkembangnya penelitian dalam Supervised Learning bertujuan untuk meningkatkan kemampuan model dalam mengatasi berbagai kompleksitas dan variasi data.

2.3. Model Klasifikasi

Model klasifikasi adalah suatu bentuk model dalam machine learning yang digunakan untuk memprediksi kategori atau kelas dari suatu data berdasarkan fitur-fitur yang diberikan. Tujuannya adalah untuk menghasilkan suatu fungsi atau model yang dapat memetakan input ke dalam kelas yang sesuai. Proses pembuatan model klasifikasi disebut dengan supervised learning, di mana model dilatih menggunakan

dataset yang telah diberi label. Algoritma klasifikasi beroperasi dengan mengidentifikasi pola atau karakteristik dalam data pelatihan dan membuat aturan atau kebijakan untuk memisahkan kelas-kelas yang berbeda. Beberapa contoh model klasifikasi termasuk Decision Trees, Support Vector Machines (SVM), Naive Bayes, dan Random Forests.

Pentingnya pemilihan algoritma yang sesuai dengan tugas klasifikasi tertentu tidak dapat diabaikan. Misalnya, Decision Trees efektif untuk tugas klasifikasi yang dapat dijelaskan secara hierarkis, sementara SVM dapat digunakan untuk menangani dataset yang tidak linier. Kinerja model klasifikasi juga dapat dipengaruhi oleh kualitas dan jumlah data pelatihan, serta keberadaan fitur-fitur yang relevan. Metrik evaluasi yang umum digunakan untuk mengukur kinerja model klasifikasi melibatkan akurasi, presisi, recall, F1-score, dan area di bawah kurva Receiver Operating Characteristic (AUC-ROC). Akurasi mengukur sejauh mana model dapat mengklasifikasikan data dengan benar, sementara presisi dan recall memberikan wawasan lebih rinci tentang performa model pada masing-masing kelas.

Model klasifikasi memiliki berbagai aplikasi di berbagai industri, termasuk pengenalan wajah, deteksi spam, analisis sentimen, dan diagnostik medis. Dalam pengenalan wajah, model dapat mengklasifikasikan wajah berdasarkan identitas individu. Dalam deteksi spam, model dapat memfilter email berdasarkan kontennya. Dalam analisis sentimen, model dapat mengklasifikasikan teks menjadi positif, negatif, atau netral. Di bidang medis, model klasifikasi dapat digunakan untuk mendeteksi penyakit berdasarkan gejala dan data medis. Penting untuk terus

meningkatkan dan mengoptimalkan model klasifikasi, baik melalui pemilihan fitur yang tepat, tuning parameter model, atau eksplorasi algoritma yang lebih canggih. Dengan pengembangan teknologi dan kemajuan dalam bidang machine learning, model klasifikasi terus menjadi alat yang kuat untuk mendukung pengambilan keputusan dan analisis data.

2.4. Algoritma Naïve Bayes

Algoritma Naïve Bayes adalah salah satu algoritma klasifikasi yang sederhana namun efektif dalam machine learning. Algoritma ini berdasarkan teorema probabilitas Bayes dan mengasumsikan kemerdekaan kondisional antar fitur, sehingga dinamakan "naïve" atau sederhana. Meskipun asumsi ini sering tidak terpenuhi dalam situasi dunia nyata, Naïve Bayes tetap menjadi pilihan yang baik untuk berbagai tugas klasifikasi karena kinerjanya yang baik dan sifatnya yang mudah diimplementasikan. Naïve Bayes bekerja dengan memanfaatkan probabilitas dan teorema Bayes untuk menghitung probabilitas bahwa suatu instance data termasuk dalam suatu kelas tertentu. Model Naïve Bayes diberi label menggunakan metode probabilistik, di mana setiap kelas dihitung probabilitasnya berdasarkan fitur-fitur yang teramati pada data pelatihan. Algoritma ini dapat diterapkan pada berbagai tugas klasifikasi, termasuk klasifikasi teks, klasifikasi spam email, dan pengenalan pola.

Keuntungan utama Naïve Bayes adalah kecepatan dan keefisienannya dalam melatih model, terutama pada dataset besar. Selain itu, Naïve Bayes juga memiliki ketahanan terhadap variasi fitur dan ketidakseimbangan kelas. Namun, asumsi kemerdekaan kondisional antar fitur dapat menjadi keterbatasan jika hubungan

antar fitur sangat berkorelasi. Model Naïve Bayes telah banyak diadopsi dalam aplikasi kehidupan sehari-hari. Dalam klasifikasi spam email, Naïve Bayes dapat mengidentifikasi apakah suatu email adalah spam atau bukan berdasarkan kata-kata kunci yang muncul di dalamnya. Dalam klasifikasi teks, Naïve Bayes dapat digunakan untuk memprediksi kategori atau sentimen dari suatu dokumen berdasarkan kata-kata yang terdapat di dalamnya.

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

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

Penerapan Naïve Bayes juga terlihat dalam sistem filtrasi informasi dan pengelompokan berita. Model ini dapat membantu dalam mengekstraksi pola dari berbagai sumber informasi dan memilah-milahnya ke dalam kategori yang sesuai. Meskipun sederhana, Naïve Bayes terus menjadi pilihan yang populer dalam pengenalan pola dan klasifikasi data karena keseimbangannya antara kinerja yang baik dan kemudahan implementasi.

2.4.1. Uji Performa

Confusion matrix atau matriks kebingungan adalah alat evaluasi kinerja yang penting dalam konteks klasifikasi dalam machine learning. Matriks ini memberikan gambaran rinci tentang sejauh mana model klasifikasi mampu memprediksi label target dengan benar dan mengidentifikasi ketidakakuratan yang mungkin terjadi. Matriks ini terdiri dari empat sel utama: true positive (TP), true negative (TN), false positive (FP), dan false negative (FN).

True positive (TP) merujuk pada jumlah kasus positif yang diprediksi dengan benar oleh model, sedangkan true negative (TN) mengindikasikan jumlah kasus negatif yang diprediksi dengan benar. Sebaliknya, false positive (FP) mencakup kasus negatif yang salah diprediksi sebagai positif, sementara false negative (FN) menggambarkan kasus positif yang salah diprediksi sebagai negatif.

Matriks kebingungan sangat bermanfaat dalam menghitung sejumlah metrik evaluasi model, seperti akurasi, presisi, recall (sensitivitas), dan nilai F1. Akurasi memberikan informasi tentang sejauh mana model secara keseluruhan dapat memprediksi dengan benar, sementara presisi mengukur tingkat ketepatan model dalam mengidentifikasi positif. Recall menggambarkan sejauh mana model dapat menangkap semua kasus positif yang sebenarnya. Dengan menggunakan confusion matrix, praktisi machine learning dapat mendapatkan wawasan yang lebih mendalam tentang kekuatan dan kelemahan model klasifikasi mereka. Hal ini memungkinkan penyesuaian strategi atau perubahan parameter untuk meningkatkan kinerja model. Oleh karena itu, confusion matrix tidak hanya memberikan gambaran tentang akurasi model, tetapi juga membantu

mengidentifikasi jenis kesalahan yang mungkin terjadi, memberikan dasar yang kuat untuk pengembangan dan evaluasi model yang lebih baik.

Tabel 2.4. 1. Confusion Matrix Metode Naive Bayes

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

Dimana tabel 1 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 analisis data visual yang mendukung kegiatan eksplorasi, pemodelan, dan visualisasi data [6] [7]. Dirancang untuk memudahkan pengguna yang memiliki berbagai latar belakang, dari pemula hingga ahli, Orange menyediakan antarmuka pengguna grafis yang intuitif untuk membangun dan

mengevaluasi model pembelajaran mesin. Platform ini memiliki sifat open-source, memungkinkan pengguna untuk mengakses dan mengubah sumber kode sesuai kebutuhan mereka. Salah satu fitur kunci Orange adalah fleksibilitasnya dalam menangani berbagai jenis tugas analisis data. Dengan beragam alat dan komponen visual, pengguna dapat melakukan eksplorasi data, preprocessing, dan pemodelan tanpa perlu pengetahuan pemrograman yang mendalam. Ini membuatnya menjadi alat yang ideal untuk para peneliti, pelajar, dan profesional yang ingin menggali potensi dari dataset mereka.

Orange mendukung berbagai algoritma pembelajaran mesin yang dapat digunakan untuk tugas-tugas seperti klasifikasi, regresi, dan pengelompokan. Platform ini juga menyertakan alat untuk validasi model, seleksi fitur, dan optimasi parameter, memberikan pengguna kebebasan untuk menyesuaikan dan meningkatkan model mereka sesuai kebutuhan. Fitur lain yang membuat Orange menonjol adalah kemampuannya dalam visualisasi data [8]. Pengguna dapat dengan mudah membuat grafik dan diagram interaktif untuk memahami distribusi data, pola, dan tren. Visualisasi ini membantu pengguna dalam merumuskan intuisi tentang data mereka dan membuat interpretasi yang lebih mendalam.

Orange tidak hanya menyediakan antarmuka pengguna yang ramah, tetapi juga mendukung ekosistem yang luas melalui add-on dan plugin. Ini memungkinkan pengguna untuk memperluas fungsionalitas platform sesuai dengan kebutuhan spesifik mereka [9]. Orange juga mendukung berbagai format data, sehingga pengguna dapat mengimpor dan mengolah data dari berbagai sumber dengan mudah. Dengan kombinasi kemudahan penggunaan, fungsionalitas yang

kuat, dan sifat open-source, Orange telah menjadi pilihan populer di kalangan praktisi data science dan peneliti. Platform ini terus berkembang seiring waktu, dengan pembaruan dan kontribusi komunitas yang berkelanjutan, menjadikannya alat yang dinamis dan relevan dalam ekosistem data science [10].

2.6. Metodologi Penelitian

2.7.1. Penelitian Terdahulu

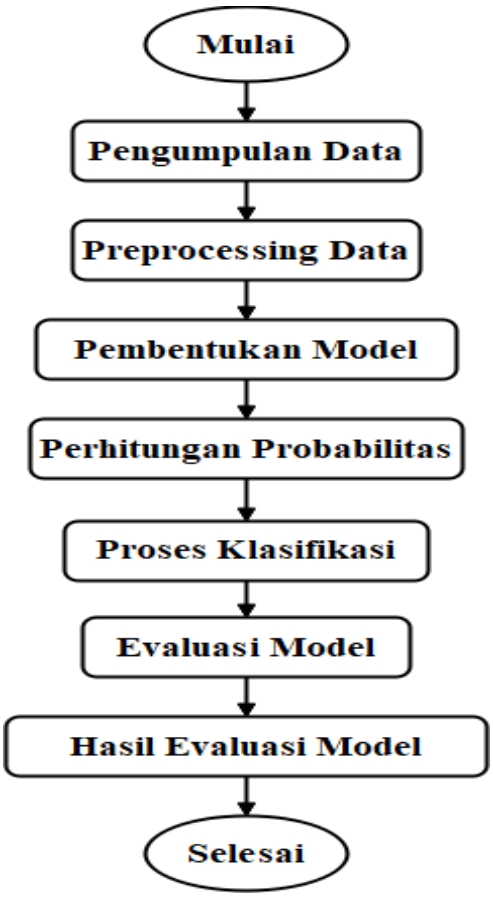
Referensi Penelitian	1
Judul	Darts: User-Friendly Modern Machine Learning for Time Series
Nama Penulis	Alexandre Gramfort
Tahun	2022
Hasil	Dalam penelitian kami, kami mengaplikasikan Darts1, sebuah pustaka pembelajaran mesin Python yang terfokus pada deret waktu, untuk menganalisis dan memprediksi pola dalam data deret waktu kami. Darts membuktikan kegunaannya dengan menyediakan beragam model, termasuk model klasik seperti ARIMA hingga jaringan saraf dalam yang lebih canggih. Kami memanfaatkan fungsionalitas perpustakaan yang modern, termasuk kemampuan mendukung rangkaian multidimensi, menyesuaikan model untuk beberapa rangkaian sekaligus, melatih pada kumpulan data besar, dan mengintegrasikan data eksternal untuk meningkatkan akurasi perkiraan kami. Desain API Darts yang cermat memastikan penggunaan yang mudah dan ramah pengguna, memungkinkan kami menerapkan model dengan efisien menggunakan metode umum seperti fit()/predict(), yang sejalan dengan pendekatan yang diterapkan dalam scikit-learn. Hal ini memberikan kerangka kerja yang kuat dan mudah digunakan untuk eksplorasi dan

	analisis data deret waktu dalam konteks penelitian kami [11].
Referensi Penelitian	2
Judul	Fairness in Machine Learning: A Survey
Nama Penulis	Simon Caton
Tahun	2021
Hasil	Penerapan teknologi Machine Learning dalam penelitian membutuhkan pertimbangan mendalam terkait dampaknya pada masyarakat, dengan perhatian khusus terhadap potensi implikasi sosial yang tidak terduga, seperti bias terhadap gender, etnis, dan/atau penyandang disabilitas. Dalam upaya untuk meningkatkan keadilan dalam Machine Learning, perusahaan dan peneliti harus menyadari kompleksitas dan tantangan dalam mengatasi bias. Artikel ini menyajikan tinjauan menyeluruh tentang berbagai aliran pemikiran dan pendekatan yang dapat diambil untuk memitigasi bias dan meningkatkan keadilan dalam konteks Machine Learning. Dengan mengorganisir metode ke dalam kerangka kerja metode pra-pemrosesan, dalam-pemrosesan, dan pasca-pemrosesan, artikel ini menyajikan 11 area metode yang diterima secara luas. Selain itu, diskusi tentang keadilan dalam konteks regresi, sistem rekomendasi, dan pembelajaran tanpa pengawasan turut diperkenalkan, disertai dengan referensi perpustakaan sumber terbuka yang tersedia saat ini. Meskipun banyak penekanan pada klasifikasi biner, artikel ini juga mencakup berbagai jenis model Machine Learning. Penutup artikel mengidentifikasi lima dilema dalam penelitian keadilan, mencerminkan tantangan terbuka yang memerlukan pemikiran dan tindakan lebih lanjut dalam mengembangkan solusi yang

	adil dan transparan dalam domain Machine Learning [12].
Referensi Penelitian	3
Judul	SENTIMEN ANALISIS PADA MEDIA SOSIAL MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER (NBC)
Nama Penulis	Muhamad Imam Syafii
Tahun	2023
Hasil	Dalam penelitian ini, penulis memfokuskan penerapan metode Naive Bayes untuk menganalisis sentimen pada Twitter, dengan menggunakan akun @Sandiuno sebagai sumber data. Twitter, sebagai platform microblogging, seringkali menjadi saluran bagi politisi atau tokoh publik untuk menyampaikan pesan dan kampanye menjelang pemilihan umum. Menghadapi kompleksitas bahasa yang digunakan di Twitter, termasuk penggunaan singkatan dan ejaan yang tidak standar, penelitian ini menggunakan proses text mining dan algoritma Naive Bayes Classifier untuk mengklasifikasikan 450 data tweet ke dalam kategori opini positif, negatif, dan netral. Pengujian dilakukan dengan berbagai skenario pembagian data training dan testing, dan hasil menunjukkan tingkat akurasi yang signifikan pada rasio 70:30, dengan akurasi mencapai 65,18%. Temuan ini menunjukkan potensi metode Naive Bayes dalam menganalisis sentimen di Twitter, meskipun dihadapkan pada variasi dan kompleksitas bahasa dalam data sosial media [13].
Referensi Penelitian	4
Judul	SISTEM PAKAR DIAGNOSA PENYAKIT KUCING PERSIA MENGGUNAKAN METODE NAIVE BAYES
Nama Penulis	Aditya Dwi Pratama
Tahun	2023

Hasil	Dalam penelitian ini, diterapkan metode Naïve Bayes Classifier untuk membuat sebuah sistem pakar dalam mendiagnosis penyakit pada kucing. Penyakit pada kucing memiliki tingkat risiko yang bervariasi, dan deteksi dini gejala dapat menjadi kunci untuk mencegah dampak serius, bahkan kematian. Keterbatasan pengetahuan mengenai penyakit kucing mendorong perlunya alat bantu seperti sistem pakar untuk memberikan diagnosis yang cepat dan tepat. Naïve Bayes Classifier, sebuah algoritma klasifikasi probabilitas, digunakan untuk membangun sistem ini dengan asumsi independensi antaratribut. Hasil penelitian menunjukkan bahwa sistem pakar ini memiliki tingkat akurasi sebesar 90.91%, memberikan kontribusi yang signifikan dalam membantu pemilik kucing dan masyarakat umum untuk lebih memahami penyakit dan gejala yang mungkin dialami oleh kucing mereka, serta mengambil tindakan preventif yang tepat [14].

3.1. Kerangka Penelitian



1. Pengumpulan Data

Pengumpulan data dalam text mining adalah langkah kritis yang melibatkan akuisisi, pengambilan, dan penyusunan sumber-sumber teks untuk analisis lebih lanjut. Proses ini mencakup pengumpulan dokumen dari berbagai sumber seperti situs web, database, dokumen teks, atau sumber-sumber lainnya yang relevan. Selain itu, metode pengumpulan data dapat mencakup teknik scraping untuk mengambil informasi dari halaman web atau penggunaan API untuk mendapatkan akses ke data yang diperlukan. Penting untuk memastikan bahwa data yang dikumpulkan representatif, berkualitas, dan relevan untuk tujuan analisis text mining agar hasilnya dapat memberikan wawasan yang berarti. Selain itu,

pengelolaan metadata, pemrosesan teks, dan normalisasi data juga merupakan aspek penting dalam pengumpulan data untuk memastikan bahwa data yang dianalisis dapat diolah secara efektif dalam tahap selanjutnya.

2. Preprocessing Data

Preprocessing data dalam text mining adalah fase kritis yang bertujuan untuk membersihkan, mengorganisir, dan mempersiapkan data teks sebelum dilakukan analisis lebih lanjut. Langkah-langkah dalam preprocessing melibatkan penghapusan karakter khusus, tanda baca, dan kata-kata umum yang tidak memberikan kontribusi signifikan terhadap pemahaman konten. Selain itu, proses normalisasi seperti stemming atau lemmatization digunakan untuk mengonversi kata-kata ke bentuk dasar mereka, sehingga mengurangi dimensi dan meningkatkan konsistensi dalam representasi teks. Pembagian teks menjadi token dan penghapusan stop words juga merupakan praktik umum untuk menyederhanakan struktur data. Preprocessing data text mining tidak hanya meningkatkan efisiensi analisis, tetapi juga membantu dalam meningkatkan kualitas dan relevansi hasil akhir, memastikan bahwa data yang diolah siap digunakan dalam model atau algoritma text mining dengan hasil yang lebih akurat dan bermakna.

3. Pembentukan Model

Pembentukan model dalam text mining adalah tahap kritis di mana representasi data teks yang telah melalui preprocessing diubah menjadi struktur yang dapat diinterpretasi oleh algoritma pembelajaran mesin. Langkah ini melibatkan pemilihan metode representasi teks, seperti model vektor term-frequency inverse document frequency (TF-IDF) atau embedding kata-kata menggunakan teknik seperti Word2Vec atau BERT. Setelah representasi data

ditentukan, model pembelajaran mesin, seperti klasifikasi, clustering, atau analisis asosiasi, diterapkan untuk mengidentifikasi pola atau hubungan dalam teks. Selain itu, tuning parameter model dan validasi lintas (cross-validation) digunakan untuk memastikan kinerja model yang optimal. Pemahaman domain dan penyesuaian model sesuai dengan kebutuhan analisis spesifik juga menjadi kunci dalam memastikan bahwa model yang terbentuk dapat memberikan hasil interpretatif dan prediktif yang maksimal dalam konteks text mining.

4. Perhitungan Probabilitas

Perhitungan probabilitas dalam text mining memiliki peran sentral dalam beberapa aspek, terutama ketika berhadapan dengan tugas seperti klasifikasi teks atau analisis sentimen. Dalam konteks ini, probabilitas sering digunakan untuk menilai sejauh mana sebuah dokumen atau kata tertentu terkait dengan kategori atau sentimen tertentu. Metode seperti Naive Bayes, yang berdasarkan pada perhitungan probabilitas kondisional, sering diterapkan untuk mengklasifikasikan teks. Pendekatan ini melibatkan pengukuran probabilitas dokumen berada dalam suatu kelas berdasarkan frekuensi kemunculan kata-kata atau fitur-fitur lainnya dalam kelas tersebut. Dengan demikian, perhitungan probabilitas memungkinkan model untuk membuat estimasi probabilitas kelas untuk setiap dokumen, membantu dalam pengambilan keputusan yang lebih informasional dan kontekstual dalam analisis teks.

5. Proses Klasifikasi

Proses klasifikasi dalam text mining merupakan tahap esensial di mana dokumen-dokumen teks dikategorikan ke dalam kelas atau kategori tertentu berdasarkan ciri-ciri atau fitur-fitur yang diambil dari teks tersebut. Metode

klasifikasi yang umum digunakan termasuk algoritma pembelajaran mesin seperti Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), dan Naive Bayes. Dalam konteks ini, representasi vektor dokumen, seperti TF-IDF atau embedding kata, sering digunakan untuk mewakili teks secara numerik sehingga dapat digunakan oleh algoritma klasifikasi. Selama proses klasifikasi, model yang telah dilatih dengan data pembelajaran mengidentifikasi pola dan hubungan antara fitur-fitur teks untuk memprediksi kategori atau label yang sesuai dengan dokumen baru. Validasi model dan tuning parameter penting untuk meningkatkan akurasi dan generalitas klasifikasi, sementara interpretasi hasil klasifikasi dapat memberikan wawasan yang berharga dalam pemahaman makna dan pola di dalam teks yang dianalisis.

6. Evaluasi Model

Evaluasi model dalam text mining merupakan langkah kritis untuk mengukur kinerja dan kehandalan model klasifikasi atau analisis teks. Metode evaluasi mencakup penggunaan berbagai metrik seperti akurasi, presisi, recall, dan F1-score, yang memberikan gambaran holistik tentang kemampuan model dalam mengklasifikasikan teks. Selain itu, matriks kebingungan (confusion matrix) digunakan untuk menyajikan hasil klasifikasi secara lebih rinci, memperlihatkan jumlah dokumen yang diklasifikasikan dengan benar dan yang salah untuk setiap kategori. Validasi lintas dan pengujian di luar sampel pelatihan (out-of-sample testing) juga menjadi langkah penting untuk mengevaluasi generalitas model terhadap data yang tidak terlibat dalam pembentukan model. Evaluasi model text mining yang komprehensif membantu memastikan bahwa model mampu

memberikan prediksi yang akurat dan berguna dalam konteks spesifik analisis teks, serta dapat digeneralisasikan dengan baik ke data teks yang belum pernah dilihat sebelumnya.

7. Hasil Evaluasi Model

Hasil evaluasi model dalam text mining memberikan wawasan penting tentang kualitas dan kinerja model dalam mengolah data teks. Akurasi model mengindikasikan sejauh mana model mampu mengklasifikasikan dokumen secara benar, sedangkan presisi dan recall memberikan informasi tentang kemampuan model dalam mengidentifikasi dan mengingat dokumen yang relevan. F1-score, sebagai ukuran kombinasi presisi dan recall, memberikan gambaran holistik tentang keseimbangan antara keduanya. Matriks kebingungan membantu menggambarkan performa model dengan rinci, memperlihatkan dokumen-dokumen yang diklasifikasikan dengan benar atau salah untuk setiap kategori. Evaluasi model yang baik tidak hanya mempertimbangkan akurasi global, tetapi juga kecocokan model terhadap tujuan dan kebutuhan analisis teks spesifik. Hasil evaluasi yang komprehensif memungkinkan penyesuaian model, tuning parameter, atau bahkan pemilihan metode yang lebih sesuai, untuk memastikan bahwa model text mining dapat memberikan hasil yang konsisten, akurat, dan berguna dalam mendukung pengambilan keputusan atau pemahaman konten teks secara lebih mendalam.