BAB II

LANDASAN TEORI

A. Data

Menurut Arhami dan Nasir (2020 : 15), “Data merupakan fakta dan statistik yang telah dikumpulkan secara bersama-sama untuk digunakan dalam berbagai macam analisis atau dijadikan referensi-referensi dalam mendukung berbagai macam penelitian atau pendapat-pendapat.

Menurut Kenneth C. Laudon dan Jane P. Laudon (2017:44), data didefinisikan sebagai "aliran fakta mentah yang mewakili peristiwa yang terjadi di dalam organisasi atau lingkungan fisik sebelum mereka memiliki makna dan berguna bagi manusia".

Menurut Syafrizal Helmi Situmorang (2014:2) sifat data dapat dibagi menjadi dua yaitu:

1. Data Kualitatif yaitu data yang tidak berbentuk angka, data kualitatif mempunyai ciri tidak bisa dilakukan operasi matematika, seperti penambahan, pengurangan, perkalian, dan pembagian. misalnya: Kuesioner Pertanyaan tentang suasana kerja, kualitas pelayanan sebuah restoran atau gaya kepemimpinan, dsb.
2. Data Kuantitatif yaitu data yang berbentuk angka, misalnya: harga saham, besarnya pendapatan, dsb. Data kuantitatif bisa disebut sebagai data berupa angka dalam arti sebenarnya. Jadi. berbagai operasi matematika bisa dilakukan pada data kuantitatif.

Menurut Syafrizal Helmi Situmorang (2014:3) sumber data dapat dibagi menjadi dua yaitu:

1. Data Internal yaitu data dari dalam suatu organisasi yang menggambarkan keadaan organisasi tersebut. Misalnya suatu perusahaan: Jumlah karyawannya, jumlah modalnya, jumlah produksinya.
2. Data Eksternal yaitu data dari luar suatu organisasi yang dapat menggambarkan faktor–faktor yang mungkin mempengaruhi hasil kerja suatu organisasi. Misalnya:

daya beli masyarakat mempengaruhi hasil penjualan suatu perusahaan.

Menurut Syafrizal Helmi Situmorang (2014:3) cara memperoleh data dapat dibagi menjadi dua yaitu:

1. Data Primer (*primary data*) yaitu data yang dikumpulkan sendiri oleh perorangan/ suatu organisasi secara langsung dari objek yang diteliti dan untuk kepentingan studi yang bersangkutan yang dapat berupa interview, observasi.
2. Data Sekunder (*secondary data*) yaitu data yang diperoleh/ dikumpulkan dan disatukan oleh studi – studi sebelumnya atau yang diterbitkan oleh berbagai instansi lain. Biasanya sumber tidak langsung berupa data dokumentasi dan arsip – arsip resmi.

Menurut Syafrizal Helmi Situmorang (2014:3) waktu pengumpulan data dapat dibagi menjadi dua yaitu:

1. Data *cross section* adalah data yang dikumpulkan pada suatu waktu tertentu (*at a point of time*) untuk menggambarkan keadaan dan kegiatan pada waktu tersebut. Misalnya : data penelitian yang menggunakan kuesioner.
2. Data berkala (*time series*) adalah data yang dikumpulkan dari waktu ke waktu untuk melihat perkembangan suatu kejadian/kegiatan selama periode tersebut. Misalnya, perkembangan uang beredar, harga 9 macam bahan pokok, penduduk

B. Informasi

Menurut George (2017:4-6), "Informasi merupakan kumpulan data yang terorganisir dan diproses supaya memiliki nilai tambahan di luar nilai yang dimiliki oleh fakta-fakta secara individu". Informasi yang berkualitas berperan penting dalam pengambilan keputusan namun tidak semua data diproses menjadi informasi yang berkualitas".

Informasi yang berkualitas dapat dibedakan dengan karakteristik tersebut:

1. Mudah Diakses, informasi harus mudah diakses oleh pengguna-pengguna yang berkepentingan sehingga mereka dapat memperoleh informasi dalam format dan waktu yang tepat untuk memenuhi kebutuhan.
2. Akurat, informasi yang akurat merupakan informasi yang bebas dari kesalahan. Dalam beberapa kasus, informasi yang tidak akurat dihasilkan dari data yang tidak akurat yang dimasukkan ke dalam proses transformasi. Hal tersebut sering disebutnya sebagai sampah masuk dan sampah keluar.
3. Lengkap, informasi yang lengkap memuat semua fakta-fakta penting.
4. Ekonomis, informasi harus relatif ekonomis untuk diproduksi. Para pengambil keputusan harus selalu menjaga keseimbangan antara nilai dari informasi dan biaya yang dibutuhkan untuk memproduksi informasi tersebut.
5. Fleksibel, informasi yang fleksibel dapat digunakan untuk berbagai macam kegunaan.
6. Relevan, informasi yang relevan merupakan informasi yang penting untuk para pengambil keputusan.
7. Dapat Diandalkan, informasi yang dipercaya oleh para pengguna. Di banyak kasus, keandalan sebuah informasi tergantung dengan keandalan metode yang digunakan untuk mengumpulkan data. Di kasus lainnya, keandalan bergantung pada sumber informasi.
8. Aman, informasi harus aman dan dijauhkan dari akses pengguna-pengguna yang tidak berkepentingan.
9. Sederhana, informasi harus sederhana dan tidak kompleks. Informasi yang rumit dan detail biasanya tidak dibutuhkan.
10. Tepat Waktu, informasi yang tepat waktu dapat disajikan pada saat dibutuhkan.

11. Dapat Diverifikasi, informasi harus dapat diverifikasi dimana seseorang dapat mengecek apakah informasi tersebut benar dengan cara mengecek berbagai sumber untuk informasi yang sama.

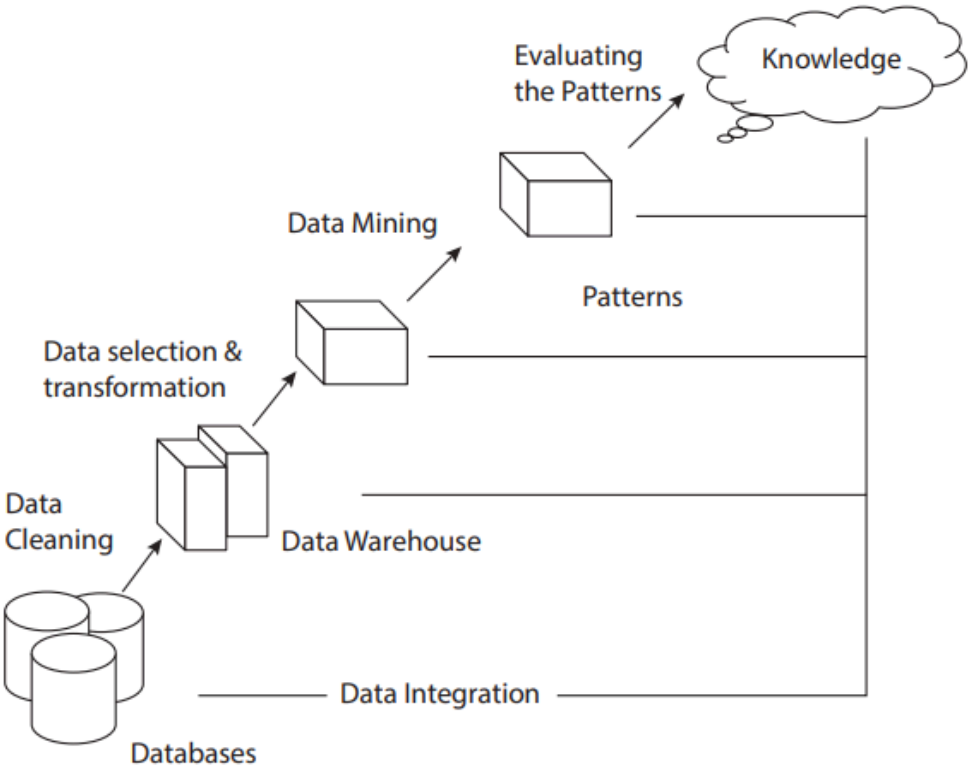
C. Data Mining

1. Pengertian Data Mining

Menurut Pan Ning-Tan (2019:24), "Data Mining adalah proses menemukan informasi yang berguna secara otomatis dalam repositori data besar. Teknik data mining digunakan untuk menyelidiki set data besar dengan tujuan menemukan pola yang baru dan bermanfaat yang mungkin tidak akan diketahui sebaliknya. Mereka juga memberikan kemampuan untuk memprediksi hasil dari observasi di masa depan, seperti jumlah yang akan dihabiskan oleh pelanggan di toko online atau toko fisik."

2. Proses *Data Mining*

Data mining adalah bagian integral dari penemuan pengetahuan dalam basis data (KDD), yang merupakan proses keseluruhan untuk mengubah data mentah menjadi informasi yang berguna, seperti yang ditunjukkan dalam Gambar 2.1. Proses ini terdiri dari serangkaian langkah, mulai dari pra-pemrosesan data hingga pemrosesan hasil data mining.



Gambar 2.1

Proses *data mining*

Sumber: Rohit Raja (2022:2)

Proses *Data Mining:*

a. Pembersihan data

Langkah ini dapat didefinisikan sebagai menghapus data yang tidak relevan. Menghapus data yang tidak relevan tidak lain adalah data yang tidak diinginkan dan dapat dihapus.

b. Integrasi data

Data dikumpulkan dari sumber yang heterogen dan diintegrasikan ke dalam sumber yang sama seperti gudang data.

c. Pemilihan & transformasi data

Setelah data dipilih, tugas selanjutnya adalah transformasi data.

d. Evaluasi pola

Evaluasi didasarkan pada beberapa ukuran; setelah langkah-langkah ini diterapkan, hasil yang diambil dibandingkan / dievaluasi secara ketat berdasarkan pola yang disimpan.

e. Representasi pengetahuan

Merepresentasikan data yang telah diproses ke dalam format yang dibutuhkan seperti tabel dan laporan

Menurut Pan Ning-Tan (2019 : 43), terdapat beberapa isu terkait data yang penting untuk keberhasilan data mining:

a. Tipe Data

Kumpulan data dapat berbeda dalam beberapa cara. Misalnya, atribut yang digunakan untuk menggambarkan objek data dapat berupa tipe yang berbeda, baik kuantitatif maupun kualitatif, dan kumpulan data sering memiliki karakteristik khusus; misalnya, beberapa kumpulan data berisi deret waktu atau objek dengan hubungan eksplisit satu sama lain. Tidak mengherankan, jenis data menentukan alat dan teknik mana yang dapat digunakan untuk menganalisis data. Memang, penelitian baru dalam data mining sering kali dipicu oleh kebutuhan untuk menyesuaikan area aplikasi baru dan jenis data baru yang muncul.

b. Kualitas Data

Data seringkali jauh dari kesempurnaan. Meskipun sebagian besar teknik data mining dapat mentoleransi beberapa tingkat ketidaksempurnaan dalam data, fokus pada pemahaman dan peningkatan kualitas data biasanya meningkatkan kualitas analisis yang dihasilkan. Masalah kualitas data yang sering perlu ditangani

mencakup adanya noise dan outlier; data yang hilang, inkonsisten, atau duplikat; dan data yang biasa atau, dengan cara lain, tidak mewakili fenomena atau populasi yang seharusnya dijelaskan oleh data tersebut.

c. Data Preprocessing

Seringkali, data mentah harus diproses agar menjadi lebih cocok untuk analisis. Sementara satu tujuan mungkin adalah untuk meningkatkan kualitas data, tujuan lainnya fokus pada modifikasi data agar lebih sesuai dengan teknik atau alat data mining yang ditentukan. Sebagai contoh, atribut yang bersifat kontinu, misalnya panjang, kadang perlu diubah menjadi atribut dengan kategori diskrit, misalnya pendek, sedang, atau panjang, agar dapat menerapkan teknik tertentu. Sebagai contoh lain, jumlah atribut dalam kumpulan data sering dikurangi karena banyak teknik yang lebih efektif ketika data memiliki jumlah atribut yang relatif kecil.

3. Tugas Inti *Data Mining*

a. Prediksi *(Predictive)*

Mengacu pada tugas membangun model untuk variabel target sebagai fungsi dari variabel penjelas. Ada dua jenis tugas pemodelan prediktif: klasifikasi, yang digunakan untuk variabel target diskrit, dan regresi, yang digunakan untuk variabel target kontinu.

b. Asosiasi *(Association)*

Digunakan untuk menemukan pola yang menggambarkan fitur-fitur yang sangat terkait dalam data. Pola yang ditemukan biasanya direpresentasikan dalam bentuk aturan implikasi atau himpunan bagian fitur. Karena ukuran ruang pencariannya yang eksponensial, tujuan analisis asosiasi adalah mengekstrak pola yang paling menarik dengan cara yang efisien.

c. Klastering *(Clustering)*

Clustering berusaha untuk menemukan kelompok pengamatan yang berkaitan erat sehingga pengamatan yang termasuk dalam cluster yang sama lebih mirip satu sama lain dibandingkan pengamatan yang termasuk dalam cluster lain. Pengelompokan telah digunakan untuk mengelompokkan kumpulan pelanggan terkait, menemukan wilayah lautan yang memiliki dampak signifikan terhadap iklim bumi, dan mengkompresi data.

d. Anomaly detection

Deteksi anomali adalah tugas mengidentifikasi observasi yang karakteristiknya berbeda secara signifikan dari data lainnya. Pengamatan seperti ini dikenal sebagai anomali atau outlier. Tujuan dari algoritma deteksi anomali adalah untuk menemukan anomali yang sebenarnya dan menghindari pemberian label yang salah pada objek normal sebagai anomali.

D. Machine Learning

Menurut Ian Goodfellow (2017 : 98), *"Machine learning* adalah suatu bentuk statistik terapan terapan dengan penekanan yang lebih besar pada penggunaan komputer untuk memperkirakan fungsi-fungsi yang rumit secara statistik"

Tujuan *machine learning* biasanya dijelaskan dalam bentuk bagaimana sistem *machine learning* harus memproses sebuah contoh. Contohnya adalah kumpulan fitur yang telah diukur secara kuantitatif dari beberapa objek atau peristiwa yang kita ingin proses oleh sistem pembelajaran mesin.

Menurut Andriy Burkov (2019:1), "Pembelajaran mesin adalah sub bidang ilmu komputer yang berkaitan dengan pembuatan algoritma yang, agar berguna, tergantung pada kumpulan contoh dari beberapa fenomena. Contoh-contoh ini dapat berasal dari alam, dibuat oleh manusia, atau dihasilkan oleh algoritma lain.".

Pembelajaran mesin juga dapat didefinisikan sebagai proses pemecahan masalah praktis dengan mengumpulkan kumpulan data, dan secara algoritmik membangun model

statistik berdasarkan kumpulan data tersebut. Model statistik tersebut diasumsikan dapat digunakan untuk memecahkan masalah praktis.

E. Klasifikasi

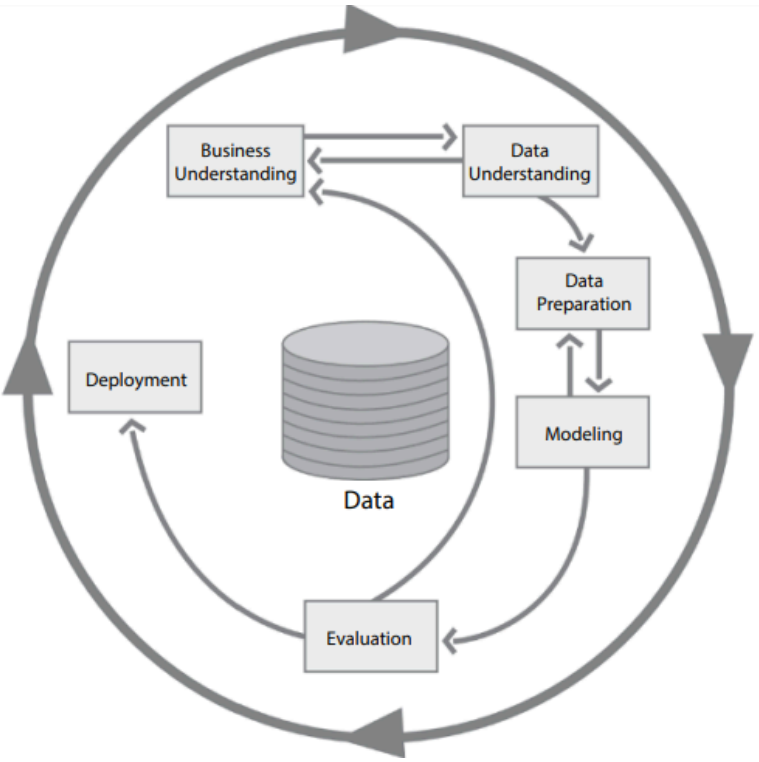
Menurut Parteek Bhartia (2019:65), "Klasifikasi adalah metode klasik yang digunakan oleh para peneliti pembelajaran mesin dan ahli statistik untuk memprediksi hasil dari sampel yang tidak diketahui. Metode ini digunakan untuk mengkategorikan objek (atau sesuatu) ke dalam sejumlah kelas diskrit".

Masalah klasifikasi dapat terdiri dari dua jenis, baik biner maupun multi kelas. Dalam klasifikasi biner, atribut target hanya dapat memiliki dua kemungkinan nilai. Sebagai contoh, tumor adalah kanker atau bukan, sebuah tim akan menang atau kalah, sentimen dari sebuah kalimat adalah positif atau negatif, dan seterusnya. Dalam klasifikasi multi kelas, atribut target dapat memiliki lebih dari dua nilai. Sebagai contoh, tumor dapat berupa kanker tipe 1, tipe 2, atau tipe 3.

F. Sirklus Penambangan Data CRISP-DM

Menurut Andres Fortino (2023:13), "CRISP-DM (Cross Industry Standard Process for Data Mining) seperti desain penilaian analisis informasi yang menggambarkan metode daur ulang yang biasanya digunakan oleh penggali data terstruktur untuk menangani spekulasi pasar".

Model referensi CRISP-DM (Cross Industry Standard Process for Data Mining) adalah proses yang berguna dan praktis untuk semua proyek. Pada gambar 2.2 menunjukan enam langkah penting pada proses penmabangan data yaitu:



Gambar 2.2

CRISP - DM *data mining cycle*

Sumber: Rohit Raja (2022:285)

Data mining cycle:

1. Langkah awal adalah mencoba untuk menunjukkan tanda-tanda perbaikan dengan memikirkan kebutuhan bisnis apa yang harus diekstraksi dari Data. Penyelidik perlu memahami apa yang benar-benar dibutuhkan klien dari sudut pandang bisnis. Klien sering kali memiliki beberapa tujuan dan batasan yang saling bertentangan yang harus diatur dengan tepat. Selain itu, tahap pemahaman bisnis terkait dengan mengkarakterisasi tujuan dan prasyarat khusus untuk Data mining. Konsekuensi dari tahap ini adalah definisi dari tugas dan penggambaran metodologi keras yang diatur untuk dicapai baik tujuan bisnis maupun tujuan Data mining. Ini juga menggabungkan pilihan yang mendasari peralatan dan strategi.

2. Langkah kedua adalah fase pemahaman data, dimulai dengan pengumpulan data awal untuk mendapatkan garis besar Data yang dapat diakses dan kualitasnya. Hal ini termasuk memeriksa apakah semua Data yang diperlukan (untuk memenuhi tujuan Data Mining) dapat diakses, serta menyusun pengaturan untuk mengetahui Data mana yang diperlukan. Untuk memulainya, pengguna menyajikan Informasi yang telah dikumpulkan, bersama dengan strukturnya, besarannya, beberapa garis besar substrat baru yang telah yang telah ditetapkan.
3. Pada tahap ini, Data siap untuk tindakan penggalian Data lebih lanjut. Kesiapan data adalah salah satu bagian yang paling penting dan sering kali membosankan dalam penggalian data. Memang, perencanaan Data, pada umumnya, memakan waktu 50-70% dari waktu dan tenaga suatu usaha. Pilihan bisnis bergantung pada pemeriksaan. Namun, jika Data tidak sesuai atau tidak memadai, investigasi Anda akan menjelaskan pilihan bisnis yang salah. Investigasi yang buruk menyiratkan pilihan bisnis yang tidak berdaya.
4. *Displaying* adalah pusat penjelasan dari tindakan penggalian data. Di sinilah penentuan dan pemanfaatan prosedur peragaan terjadi. Sebelum benar-benar membangun sebuah model, Anda biasanya memisahkan dataset ke dalam set pelatihan, pengujian, dan persetujuan. Pada saat itu, Anda membangun model pada set train. Beberapa prosedur yang menandakan menandai anggapan terbuka tentang Data, untuk kejadian, seluruh perilaku memiliki konstanta sirkulasi atau tidak ada kemampuan yang hilang dapat diterima.
5. Penilaian ini menjamin pemeriksaan yang pasti terhadap model Data yang dibuat dengan penugasan dan memilih Model yang paling tepat. Konsekuensi dari kemajuan masa lalu dinilai dengan menggunakan aturan bisnis yang dibangun pada awal usaha. Jadi tahap ini terkait dengan memeriksa apakah pengaturan Data

mining memenuhi masalah bisnis dan mencoba untuk menentukan apakah akan ada motif pasar dibalik kurangnya cetak biru.

6. Setelah perencanaan data, struktur model, dan konfirmasi model, model yang dipilih digunakan dalam tahap organisasi. Menciptakan sisa-sisa yang ideal biasanya bukan penampilan proyek. Terlepas dari apakah tujuannya adalah untuk memperluas Data pada Data, Data yang meningkat sekarang harus ditangani dan diperkenalkan kepada klien sehingga klien dapat memanfaatkannya tanpa masalah.

G. Pencemaran udara

1. Pengertian Pencemaran Udara

Menurut Peraturan Pemerintah Nomor 41 Tahun 1999 tentang Pengendalian Pencemaran Udara pasal 1 ayat 1, "Pencemaran udara adalah masuknya atau dimasukkannya zat, energi, dan/atau komponen lain ke dalam udara ambien oleh kegiatan manusia, sehingga mutu udara ambien turun sampai ke tingkat tertentu yang menyebabkan udara ambien tidak dapat memenuhi fungsinya".

Menurut Abhishek Tiwary dan Ian Williams (2019:1) Polusi udara didefinisikan sebagai "Keberadaan zat di atmosfer yang dapat menyebabkan efek buruk bagi manusia dan lingkungan".

Polusi (dalam pengertian umum) didefinisikan sebagai masuknya zat atau energi ke dalam lingkungan oleh manusia yang dapat menyebabkan bahaya bagi kesehatan manusia, kerusakan sumber daya hidup dan sistem ekologi, kerusakan struktur atau fasilitas, atau gangguan terhadap penggunaan lingkungan yang sah.

2. Sumber Pencemaran Udara

Pencemaran udara terbagi menjadi dua golongan, yaitu pencemar primer dan pencemar sekunder. Pencemar primer adalah polutan yang dipancarkan langsung ke atmosfer - misalnya, CO berasal langsung dari pembakaran bahan bakar fosil yang tidak sempurna pada kendaraan bermotor, dan SO2 dipancarkan dari pembangkit listrik dan pabrik-pabrik industri. Pencemar sekunder terbentuk di udara sebagai hasil

dari reaksi kimia dengan polutan lain dan gas-gas di atmosfer - misalnya, ozon dihasilkan oleh reaksi fotokimia di atmosfer (Abhishek Tiwary dan Ian Williams, 2019).

Sumber bahan pencemar udara primer dapat dibagi lagi menjadi dua golongan besar, yaitu (Abhishek Tiwary dan Ian Williams, 2019):

a. Sumber Alamiah *(Natural Sources)*

Beberapa kegiatan alam yang bisa menyebabkan pencemaran udara adalah aktivitas gunung berapi, kebakaran hutan, badai pasir, petir dan lain-lain.

b. Sumber Buatan Manusia *(Anthropogenic Sources)*

Sumber utamanya adalah pembakaran bahan bakar fosil untuk energi, terutama di pembangkit listrik dan kendaraan bermotor. Namun, ada banyak sumber yang tidak terkait dengan pembakaran, termasuk proses industri, pertambangan batubara, penggunaan pelarut rumah tangga dan industri, kebocoran gas alam di jaringan distribusi nasional, dan tempat pembuangan sampah. Sumber non-pembakaran sangat penting untuk VOC dan metana.

3. Jenis Bahan Pencemar Udara

Menurut Abhishek Tiwary dan Ian Williams (2019:6), pencemar udara dapat dibagi menjadi pencemar yang diatur dan tidak diatur, berdasarkan perlakuannya oleh badan-badan lingkungan hidup di Amerika Serikat, Uni Eropa, dan negara-negara besar di Asia. Pencemar udara utama yang diatur adalah:

a. Sulfur dioksida (SO2).
b. Nitrogen oksida (NOx).
c. Karbon monoksida (CO).
d. Logam berat seperti timbal (Pb), kadmium (Cd), dan platina (Pt).
e. Senyawa organik seperti benzena dan hidrokarbon aromatik polisiklik (PAH).

f. Oksidan fotokimia seperti ozon (O3) dan peroksiasetil nitrat (PAN).

g. Materi partikulat mencakup berbagai ukuran (biasanya antara 0,01 dan 100 µm) dan dapat terdiri dari bahan organik atau anorganik (atau campuran). Jenis materi partikulat yang dipantau meliputi total partikulat tersuspensi, asap, dan partikel dengan ukuran tertentu, seperti PM10 dan PM2.5.

H. Algoritma *Support Vector Machine* (SVM)

Menurut Mohammed J. Zaki et al. (2020:517), "SVM merupakan Metode klasifikasi berdasarkan diskriminan linier margin maksimum, yaitu tujuannya untuk menemukan hyperplane optimal yang memaksimalkan gap atau margin antar kelas."

Menurut Pang-Ning Tan, et al. (2019:478), "*Support Vector Machine* (SVM) adalah model klasifikasi diskriminatif yang mempelajari batas keputusan linear atau nonlinear pada ruang atribut untuk memisahkan kelas-kelas. Selain memaksimalkan pemisahan dua kelas, SVM menawarkan kemampuan regularisasi yang kuat, yaitu mampu mengontrol kompleksitas model untuk memastikan kinerja generalisasi yang baik".

Karena kemampuan uniknya untuk secara naluriah meregulasi pembelajarannya, SVM mampu mempelajari model yang sangat ekspresif tanpa menderita dari overfitting. Oleh karena itu, SVM telah menerima perhatian yang besar di komunitas pembelajaran mesin dan umumnya digunakan dalam beberapa aplikasi praktis, mulai dari pengenalan digit tulisan tangan hingga kategorisasi teks.

SVM memiliki akar yang kuat dalam teori pembelajaran statistik dan didasarkan pada prinsip minimisasi risiko struktural. Aspek unik lain dari SVM adalah bahwa ia mewakili batas keputusan hanya menggunakan subset dari contoh pelatihan yang paling sulit untuk diklasifikasikan, yang dikenal sebagai support vector. Oleh karena itu, ia adalah model diskriminatif yang hanya dipengaruhi oleh contoh pelatihan dekat dengan batas antara dua kelas. SVM secara luas diklasifikasikan menjadi dua jenis: SVM sederhana atau linier dan SVM kernel atau non-linier. Jenis-jenis SVM adalah sebagai berikut:

1. Linear SVM

SVM linier adalah pengklasifikasi yang mencari hyperplane pemisah dengan margin terbesar, oleh karena itu sering dikenal sebagai pengklasifikasi margin maksimal.

2. Non-Linear SVM

SVM jenis ini memiliki lebih banyak fleksibilitas untuk data non-linier karena dapat menambahkan lebih banyak fitur agar sesuai dengan hyperplane daripada ruang dua dimensi. Kernel SVM lebih disukai untuk tujuan klasifikasi data. Fungsi kernel yang paling umum digunakan adalah fungsi basis linier, polinomial, dan radial.

Algoritma SVM sangat efektif ketika kami mencoba menemukan hyperplane pemisah maksimum antara kelas-kelas berbeda yang tersedia di fitur target. SVM dapat digunakan untuk berbagai tugas, seperti klasifikasi teks, klasifikasi gambar, deteksi spam, identifikasi tulisan tangan, analisis ekspresi gen, deteksi wajah, dan deteksi anomali.

I. Algoritma *K-Nearest Neighbors*

Menurut Muhammad Arhami dan Muhammad Nasir (2020:96-98) “KNN merupakan salah satu algoritma untuk klasifikasi yang biasa digunakan dalam data mining. KNN juga masuk dalam kategori regresi yang juga dapat digunakan untuk memprediksi seperti halnya regresi”.

Ide dari metode k-nearest-neighbors adalah untuk mengidentifikasi k *records* dalam dataset pelatihan yang mirip dengan rekaman baru yang ingin kita klasifikasikan. Kami kemudian menggunakan catatan-catatan yang mirip (bertetangga) ini untuk mengklasifikasikan catatan baru ke dalam sebuah kelas, menetapkan catatan baru ke kelas yang dominan di antara tetangga-tetangga ini (Galit Shmueli 2018).

Nilai K merupakan suatu parameter yang merujuk kepada jumlah tetangga yang paling dekat dengan objek yang diprediksi kelasnya sehingga dapat ditentukan tetangga yang

mayoritas bagi suatu objek K memainkan peranan penting dalam menentukan keberhasilan model dan akurasi yang lebih baik. K juga merupakan batasan dalam setiap kelas. Berikut adalah langkah-langkah dari algoritma K-Nearest Neighbors (KNN):

1. Tentukan nilai k: Langkah pertama dalam algoritma KNN adalah menentukan nilai k, yaitu jumlah tetangga terdekat yang akan dipertimbangkan saat mengklasifikasikan suatu data baru.
2. Nilai k disesuaikan juga dengan jumlah data training yang ada dan sebaiknya nilai k diambil dalam jumlah ganjil seperti 1,3,5,7, … dan sebagainya, karena jika mengambil ganjil akan mudah dalam menentukan mayoritas dan minoritas kedekatan jarak antara record data uji yang diprediksi dengan record data latih.
3. Hitung jarak antara data baru dengan semua data pada dataset. Jarak yang paling umum digunakan adalah jarak Euclidean.
4. Urutkan jarak yang telah dihitung dan tentukan tetangga terdekatnya sesuai dengan k yang dipilih dan jarak minimumnya (dari yang terkecil ke terbesar) sehingga berdasarkan urutan tersebut akan didapatkan kategori dari data yang diprediksi.
5. Kategori data baru yang diprediksi akan masuk dalam kelas mayoritas sesuai dengan nilai k yang ditentukan sebelumnya.
6. Evaluasi model: Evaluasi performa model KNN menggunakan metrik seperti akurasi, presisi, recall, dan F1 score.

Algoritma KNN dapat digunakan untuk berbagai tugas seperti klasifikasi dan regresi. Namun, algoritma ini memiliki beberapa kelemahan seperti sensitivitas terhadap nilai k dan jarak yang digunakan, serta membutuhkan memori yang besar untuk menyimpan data latih.

J. Python

Menurut The Code Academy (2017), ”Python adalah salah satu dari sekian banyak bahasa pemrograman komputer yang semakin populer setiap harinya. Bahasa ini adalah bahasa tingkat tinggi umum yang mudah dipelajari dan bahkan lebih mudah digunakan. Gaya dan filosofinya menekankan pada keterbacaan dan kesederhanaan kode. Sintaksnya memudahkan programmer untuk menulis instruksi komputer dalam lebih sedikit baris kode dibandingkan dengan jumlah yang diperlukan untuk menulis instruksi serupa dalam bahasa lain seperti Java, C++, atau C#”.

Menurut Fabio Nelli (2015: 13-14), “Python adalah bahasa pemrograman serbaguna dan sangat portabel yang dapat ditafsirkan dan bersifat open-source. Bahasa ini dikenal dengan pendekatan berorientasi objek dan sifatnya yang interaktif, menjadikannya bahasa yang ramah pengguna. Kesederhanaan Python membuatnya mudah dipelajari dan digunakan, sementara sifatnya yang open-source mendorong kolaborasi dan inovasi. Selain itu, Python dapat dihubungkan dengan berbagai teknologi, meningkatkan kemampuan beradaptasi dan menjadikannya pilihan populer untuk berbagai aplikasi.”

Python adalah bahasa yang kaya tetapi sederhana pada saat yang sama, sangat fleksibel sehingga memungkinkan perluasan aktivitas pengembangan di banyak bidang pekerjaan (analisis data, ilmiah, antarmuka grafis, dll.).

K. Penelitian Kuantitatif

Menurut Umesh Kumar dan D,P. Kothari (2022 : 6), “Penelitian kuantitatif adalah penelitian yang berbasis pada variabel sementara penelitian kualitatif berbasis pada atribut. Penelitian kuantitatif didasarkan pada pengukuran atau penjumlahan fenomena yang sedang diteliti. Artinya, penelitian ini didasarkan pada data dan oleh karena itu lebih objektif dan populer”.

L. Penelitian Terdahulu

1. Analisis dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara di DKI Jakarta

Penelitian yang dilakukan oleh Syekh S A Umri, Muhammad S Firdaus, Aji Primajaya adalah Penelitian dengan menggunakan dataset Indeks Pencemar Udara pada wilayah DKI Jakarta dengan melakukan komparasi atau perbandingan antara lima algoritma yaitu SVM, *Decision Tree*, KNN, dan *Neural Network*

Backpropagation dengan menggunakan validasi 10-*fold cross validation* dan *feature selection* berupa *backward elimination* menghasilkan algoritma dengan performa terbaik yakni D*ecision Tree* dengan nilai akurasi sebesar 99.80%, nilai kappa yang hampir sempurna yakni 0.996, nilai RMSE terkecil dan di bawah 0.1 yakni 0.039, serta waktu yang dibutuhkan hanya 0.8 detik. Meskipun begitu, *Neural Network Backpropagation*, KNN, SVM, dan Naive Bayes juga masih dapat digunakan sebagai model klasifikasi yang baik karena mendapatkan nilai akurasi yang tinggi di atas 90% dan nilai kappa di atas 0.8.

2. Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma *Decision Tree* Dan *Support Vector Machine.*

Penelitian yang dilakukan oleh Adinda Inez Sang, Edi Sutoyo, dan Irfan Darmawan adalah penelitian menggunakan metode data mining yaitu klasifikasi karena metode ini dapat mengolah data parameter ISPU menjadi informasi yaitu tingkat kualitas udara per harinya dengan menggunakan algoritma Decision Tree dan Support Vector Machine (SVM). Hasil dari penerapan data mining untuk klasifikasi kualitas udara di DKI Jakarta yaitu algoritma Decision Tree memiliki performa yang lebih baik dengan rasio terbaik 90:10 dibandingkan dengan algoritma SVM dengan rasio terbaik 60:40 dan untuk melakukan klasifikasi kualitas udara di DKI Jakarta. Pada algoritma Decision Tree mendapatkan nilai Precision sebesar 99,02%, Recall 99,73%, F1-Measure 99,37%, Akurasi 99,40% dan pada algoritma SVM mendapatkan nilai Precision sebesar 95,82%, Recall 88,89%, F1-Measure 92,22% dan Akurasi 94,93%.

3. Prediksi Kualitas Udara Menggunakan Algoritma *K Nearest Neighbor.*

Penelitian yang dilakukan oleh Adinda Amalia , Ati Zaidiah , Ika Nurlaili Isnainiyah adalah memprediksi kualitas udara yang ada di DKI Jakarta berdasarkan data ISPU. Prediksi dilakukan dengan menggunakan teknik data mining dengan

metode klasifikasi. Algoritma yang berfungsi dalam melakukan prediksi yaitu K-Nearest Neighbor (KNN), dimana algoritma ini adalah algoritma yang mengklasifikasikan kelas objek baru dengan didasarkan pada tetangga terdekatnya. Data yang digunakan pada penelitian berjumlah 450 data, kemudian data tersebut dibagi 2 yakni data uji dan data latih. Penelitian ini juga melakukan evaluasi model algoritma yang meliputi nilai akurasi, presisi, recall, dan f-measure untuk setiap nilai K yang diujikan. Pengukuran ini bertujuan untuk mengetahui parameter yang optimal pada dataset yang digunakan. Adapun hasil yang diperoleh dari pengujian nilai K = 3 sampai K = 9, didapatkan bahwa nilai K = 7 mempunyai performa terbaik dengan akurasi tertinggi sebanyak 96%, presisi 92%, recall 95%, dan f-measure 93%.

4. Komparasi Metode *K-Nearest Neighbor* (KNN) Dengan *Support Vector Machine* (SVM) Terhadap Tingkat Akurasi Klasifikasi Kualitas Air

Penelitian yang dilakukan oleh Jauhari Maulani, dan Mayang Sari adalah perbandingan antara algoritma SVM dan KNN terhadap akurasi klasifikasi kualitas air. Berdasarkan hasil perhitungan performansi algoritma KNN dengan algoritma SVM, Dari hasil running di atas, menunjukan nilai akurasi SVM yaitu sekitar 69%, sedangkan hasil akurasi KNN bernilai 66%, sehingga hasil akurasi SVM lebih baik daripada KNN. Anda dapat menjalankan hasil prediksi biologis untuk melihat apakah kualitas air yang dapat Anda konsumsi sesuai standar. Hal ini sangat baik untuk prediksi dimana nantinya kesalahan prediksi baru meminimalkan data.

5. Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status

Penelitian yang dilakukan oleh Amri Danades, Devie Pratama, Dian Anggraini, dan Diny Anggriani adalah perbandingan tingkat akurasi pada klasifikasi kualitas air. Pengujian pertama menggunakan 10 F-C-V, membuktikan SVM lebih baik karena

nilai akurasinya lebih tinggi yaitu 92,40 % pada kernel Linear. Nilai rata-rata akurasi KNN hanya sebesar 71,28% pada K=7. Kernel dengan akurasi tertinggi pada SVM berada pada kernel Linear, sedangkan pada KNN berada pada K=7. Pengujian kedua membuktikan bahwa prediksi kualitas air dari 15 set data dengan menggunakan algoritma KNN hanya cocok dengan 6 data dan lainnya tidak cocok. Sedangkan prediksi algoritma SVM cocok dengan 12 data dan 3 lainnya tidak cocok. Hal ini membuktikan bahwa SVM lebih baik dibandingkan KNN.