

CAT SKB

ANALIS DATA ILMIAH AHLI PERTAMA

PREDIKSI SOAL

1. Dalam konteks analisis data ilmiah, urutan hirarki DIKW yang benar dari tingkat terendah ke tertinggi adalah...
 - A. Data - Informasi - Pengetahuan - Kebijakan
 - B. Informasi - Data - Pengetahuan - Kebijakan
 - C. Pengetahuan - Data - Informasi - Kebijakan
 - D. Data - Pengetahuan - Informasi - Kebijakan
 - E. Kebijakan - Pengetahuan - Informasi - Data
2. Prinsip data ilmiah yang menyatakan bahwa data harus dapat ditemukan, diakses, dioperasikan, dan digunakan kembali dikenal dengan prinsip...
 - A. AGILE
 - B. SMART
 - C. FAIR
 - D. SOLID
 - E. SWOT
3. Manakah operator Boolean yang digunakan untuk mempersempit hasil pencarian literatur ilmiah dengan menggabungkan dua kata kunci yang harus ada keduanya?
 - A. OR
 - B. NOT
 - C. AND
 - D. XOR
 - E. IF
4. Pengidentifikasi unik yang permanen untuk dokumen elektronik, yang sering digunakan untuk mengutip jurnal ilmiah secara presisi, disebut...
 - A. URL (Uniform Resource Locator)
 - B. IP Address
 - C. ISBN (International Standard Book Number)
 - D. DOI (Digital Object Identifier)
 - E. ISSN (International Standard Serial Number)
5. Langkah pertama yang paling krusial dalam menyusun rancangan kerja analisis data ilmiah adalah...

- A. Membersihkan data
 - B. Merumuskan pertanyaan penelitian atau hipotesis
 - C. Memilih algoritma machine learning
 - D. Membuat visualisasi data
 - E. Menulis laporan akhir
6. Sumber data sekunder dalam penelitian ilmiah merujuk pada...
- A. Data yang dikumpulkan sendiri melalui survei langsung
 - B. Data yang diperoleh dari hasil eksperimen laboratorium sendiri
 - C. Data yang sudah dikumpulkan dan diolah oleh pihak lain sebelumnya
 - D. Data yang dihasilkan dari simulasi komputer saat itu juga
 - E. Data yang bersifat rahasia dan tidak dipublikasikan
7. Tindakan tidak etis dalam ilmu data di mana peneliti memanipulasi data agar sesuai dengan hipotesis yang diinginkan disebut...
- A. Data Mining
 - B. Data Cleaning
 - C. Data Fabrication/Falsification
 - D. Data Warehouse
 - E. Data Science
8. Metadata adalah...
- A. Data yang berukuran sangat besar (Big Data)
 - B. Data yang menjelaskan tentang data lain (data about data)
 - C. Data yang rusak atau korup
 - D. Data hasil prediksi masa depan
 - E. Data yang berbentuk gambar atau video
9. Dalam siklus hidup data ilmiah, tahap preservasi bertujuan untuk...
- A. Menghapus data yang tidak lagi digunakan
 - B. Mengubah data menjadi grafik
 - C. Menjaga data agar tetap dapat diakses dan digunakan dalam jangka panjang
 - D. Mengumpulkan data dari responden
 - E. Menjual data kepada pihak ketiga
10. *Grey literature* atau literatur kelabu dalam penelusuran informasi ilmiah meliputi dokumen berikut, kecuali...

- A. Laporan teknis pemerintah
 - B. Skripsi dan Tesis
 - C. Prosiding konferensi
 - D. Artikel jurnal internasional bereputasi
 - E. Dokumen kebijakan internal
11. Aplikasi yang umum digunakan untuk manajemen referensi dan sitasi dalam penulisan karya ilmiah adalah...
- A. Adobe Photoshop
 - B. SPSS
 - C. Mendeley atau Zotero
 - D. Tableau
 - E. MySQL
12. Salah satu karakteristik "Big Data" yang menggambarkan kecepatan aliran data yang masuk dan harus diproses disebut...
- A. Volume
 - B. Variety
 - C. Velocity
 - D. Veracity
 - E. Value
13. Open Science adalah gerakan untuk membuat penelitian ilmiah dan datanya dapat diakses oleh...
- A. Hanya peneliti senior
 - B. Hanya penyandang dana
 - C. Hanya pemerintah
 - D. Semua tingkatan masyarakat (publik)
 - E. Hanya universitas tertentu
14. Dalam menyusun jadwal kerja analisis data, diagram yang umum digunakan untuk memetakan tugas terhadap waktu adalah...
- A. Diagram Venn
 - B. Diagram Gantt
 - C. Diagram Pie
 - D. Diagram Scatter
 - E. Diagram Batang

15. Reproducibility (keterulangan) dalam analisis data ilmiah berarti...
- A. Analisis menghasilkan kesimpulan yang berbeda setiap kali dijalankan
 - B. Analisis hanya bisa dijalankan oleh pembuat kode aslinya
 - C. Analisis dapat dijalankan ulang oleh orang lain dengan data yang sama dan menghasilkan hasil yang sama
 - D. Analisis menggunakan data yang berbeda untuk masalah yang sama
 - E. Analisis tidak memerlukan dokumentasi
16. Data yang berbentuk angka dan dapat diukur menggunakan operasi matematika disebut data...
- A. Kualitatif
 - B. Kuantitatif
 - C. Kategorikal
 - D. Nominal
 - E. Ordinal
17. Proses mengisi nilai yang hilang (*missing values*) dalam dataset dengan nilai rata-rata atau median disebut...
- A. Eliminasi
 - B. Imputasi
 - C. Normalisasi
 - D. Reduksi
 - E. Diskritisasi
18. Sebuah data pengamatan yang nilainya menyimpang sangat jauh dari pola umum kumpulan data lainnya disebut...
- A. Mean
 - B. Median
 - C. Outlier (Pencilan)
 - D. Modus
 - E. Kuartil
19. Teknik pengambilan sampel di mana setiap anggota populasi memiliki peluang yang sama untuk dipilih disebut...
- A. Purposive Sampling
 - B. Snowball Sampling
 - C. Quota Sampling

D. Simple Random Sampling

E. Convenience Sampling

20. Proses mengubah data kategori (misal: "Merah", "Biru", "Hijau") menjadi format angka (misal: 0, 1, 2) agar bisa diproses algoritma disebut...

A. Decoding

B. Encoding

C. Scaling

D. Clustering

E. Filtering

21. Skala pengukuran data yang memiliki nilai nol mutlak (misalnya: berat badan, tinggi badan) adalah...

A. Nominal

B. Ordinal

C. Interval

D. Rasio

E. Likert

22. Transformasi data menggunakan teknik Min-Max Scaling bertujuan untuk...

A. Mengubah distribusi data menjadi lonceng (Gaussian)

B. Mengubah skala data ke dalam rentang tertentu (misalnya 0 sampai 1)

C. Menghapus data duplikat

D. Mencari korelasi antar variabel

E. Mengubah data angka menjadi teks

Topik: Manajemen Data & Basis Data

23. Kumpulan data yang terorganisir dan saling berelasi yang disimpan secara elektronik disebut...

A. Spreadsheet

B. Harddisk

C. Basis Data (Database)

D. Folder

E. Algoritma

24. Dalam basis data relasional, kolom yang berfungsi sebagai identitas unik untuk setiap baris data disebut...

A. Foreign Key

- B. Primary Key
 - C. Candidate Key
 - D. Secondary Key
 - E. Alternate Key
25. Perintah SQL (*Structured Query Language*) yang digunakan untuk mengambil atau menampilkan data dari tabel adalah...
- A. INSERT
 - B. UPDATE
 - C. DELETE
 - D. SELECT
 - E. DROP
26. Struktur data berbentuk tabel yang terdiri dari baris dan kolom adalah ciri khas dari jenis database...
- A. NoSQL
 - B. Relasional (RDBMS)
 - C. Graph Database
 - D. Document Store
 - E. Key-Value Store
27. Redundansi data dalam manajemen basis data berarti...
- A. Data hilang karena virus
 - B. Data tersimpan secara terenkripsi
 - C. Adanya duplikasi data yang sama di beberapa tempat penyimpanan
 - D. Data tidak dapat diakses
 - E. Data memiliki hubungan yang kompleks
28. Format file teks sederhana yang memisahkan nilai data dengan tanda koma, yang umum digunakan untuk pertukaran data, adalah...
- A. PDF
 - B. DOCX
 - C. JPEG
 - D. CSV
 - E. MP3

29. Hubungan antara tabel "Mahasiswa" dan tabel "Mata Kuliah" di mana satu mahasiswa bisa mengambil banyak mata kuliah dan satu mata kuliah bisa diambil banyak mahasiswa adalah hubungan...
- A. One-to-One
 - B. One-to-Many
 - C. Many-to-One
 - D. Many-to-Many
 - E. Self-Reference
30. Proses periodik menyalin data ke lokasi penyimpanan sekunder untukantisipasi kehilangan data disebut...
- A. Restore
 - B. Recovery
 - C. Backup
 - D. Formatting
 - E. Defragmentation

Topik: Data Mining

31. Teknik data mining yang digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik tanpa adanya label kelas sebelumnya disebut...
- A. Klasifikasi
 - B. Regresi
 - C. Clustering (Klasterisasi)
 - D. Asosiasi
 - E. Prediksi
32. Manakah di bawah ini yang termasuk algoritma *Supervised Learning*?
- A. K-Means
 - B. Hierarchical Clustering
 - C. Linear Regression
 - D. Apriori
 - E. DBSCAN
33. *Market Basket Analysis* adalah contoh penerapan dari teknik data mining...
- A. Asosiasi
 - B. Klasifikasi
 - C. Prediksi

D. Deteksi Anomali

E. Time Series

34. Kondisi di mana model machine learning bekerja sangat baik pada data latih (*training data*) tetapi buruk pada data baru (*test data*) disebut...

A. Underfitting

B. Overfitting

C. Optimal

D. Balanced

E. Bias

35. Algoritma Decision Tree (Pohon Keputusan) umumnya digunakan untuk tugas...

A. Reduksi Dimensi

B. Klasifikasi dan Regresi

C. Klasterisasi saja

D. Asosiasi saja

E. Pembersihan data

36. Pembagian dataset menjadi *Training Set* dan *Testing Set* bertujuan untuk...

A. Menambah jumlah data

B. Mempersulit proses komputasi

C. Melakukan validasi dan evaluasi performa model

D. Menghapus outlier

E. Mengubah format data

37. *Text Mining* adalah proses menambang informasi dari data yang berbentuk...

A. Angka terstruktur

B. Gambar

C. Teks tidak terstruktur

D. Audio

E. Video

38. Ukuran pemusatan data yang paling tepat digunakan jika terdapat *outlier* (pencilan) ekstrem dalam data adalah...

A. Mean (Rata-rata)

B. Modus

C. Median

D. Standar Deviasi

E. Varians

39. Nilai statistik yang menunjukkan seberapa jauh sebaran data dari nilai rata-ratanya adalah...

A. Korelasi

B. Standar Deviasi

C. Persentil

D. Modus

E. Frekuensi

40. Jika koefisien korelasi (r) antara dua variabel bernilai -0.9, artinya...

A. Tidak ada hubungan antara kedua variabel

B. Terdapat hubungan positif yang kuat

C. Terdapat hubungan negatif yang kuat

D. Terdapat hubungan positif yang lemah

E. Terdapat hubungan negatif yang lemah

41. Dalam pengujian hipotesis, nilai p -value yang kurang dari 0.05 biasanya diinterpretasikan sebagai...

A. Menerima hipotesis nol (H_0)

B. Menolak hipotesis nol (H_0) / Signifikan secara statistik

C. Data tidak valid

D. Kesalahan perhitungan

E. Korelasi sempurna

42. Jenis grafik yang paling tepat untuk melihat distribusi frekuensi dari sebuah variabel numerik adalah...

A. Pie Chart

B. Scatter Plot

C. Histogram

D. Line Chart

E. Map Chart

43. *Scatter Plot* (Diagram Pencar) sangat berguna untuk melihat...

A. Proporsi kategori

B. Tren waktu

C. Hubungan atau korelasi antara dua variabel numerik

- D. Hirarki data
 - E. Lokasi geografis
44. Kesalahan logika di mana seseorang menganggap bahwa korelasi menyiratkan sebab-akibat (*Causation*) disebut...
- A. *Cum hoc ergo propter hoc*
 - B. *Ad hominem*
 - C. *Straw man*
 - D. *Confirmation bias*
 - E. *Sampling bias*
45. Analisis Regresi Linear Sederhana digunakan untuk...
- A. Mengelompokkan data
 - B. Memprediksi nilai variabel dependen berdasarkan satu variabel independen
 - C. Menghitung jumlah kategori
 - D. Membandingkan dua rata-rata
 - E. Mengurangi dimensi data
46. Grafik *Boxplot* (Diagram Kotak Garis) sangat efektif untuk menampilkan informasi berikut, kecuali...
- A. Median
 - B. Kuartil (Q1 dan Q3)
 - C. Outlier
 - D. Nilai Minimum dan Maksimum
 - E. Hubungan sebab akibat
47. Hipotesis Nol (H_0) biasanya menyatakan...
- A. Ada perbedaan yang signifikan
 - B. Ada pengaruh yang signifikan
 - C. Tidak ada perbedaan atau tidak ada pengaruh (status quo)
 - D. Alternatif dari kesimpulan
 - E. Kesalahan penelitian
48. Teknik analisis multivariat yang digunakan untuk mengurangi jumlah variabel menjadi beberapa faktor utama tanpa banyak mengurangi informasi disebut...
- A. ANOVA
 - B. Chi-Square

- C. PCA (Principal Component Analysis)
 - D. T-Test
 - E. Time Series Analysis
49. Kesalahan Tipe I dalam pengujian hipotesis terjadi ketika...
- A. Menolak H_0 padahal H_0 benar (False Positive)
 - B. Menerima H_0 padahal H_0 salah (False Negative)
 - C. Menolak H_0 saat H_0 salah
 - D. Menerima H_0 saat H_0 benar
 - E. Tidak mengambil keputusan
50. Visualisasi data yang buruk dan menyesatkan (*misleading*) seringkali disebabkan oleh...
- A. Penggunaan warna yang kontras
 - B. Sumbu Y yang tidak dimulai dari nol (truncated axis)
 - C. Pencantuman label yang jelas
 - D. Penggunaan judul yang deskriptif
 - E. Pemilihan tipe chart yang sesuai
51. Bahasa pemrograman yang paling populer dan memiliki ekosistem pustaka (library) terluas untuk analisis data ilmiah dan *machine learning* saat ini adalah...
- A. C++
 - B. Java
 - C. PHP
 - D. Python
 - E. HTML
52. Dalam publikasi ilmiah, jenis akses terbuka (*Open Access*) di mana penulis membayar biaya publikasi (APC) agar artikelnya dapat diakses gratis oleh pembaca segera setelah terbit disebut...
- A. Green Open Access
 - B. Gold Open Access
 - C. Diamond Open Access
 - D. Hybrid Open Access
 - E. Closed Access
53. Undang-Undang di Indonesia yang mengatur secara spesifik tentang perlindungan data pribadi subjek data adalah...
- A. UU ITE

- B. UU KIP (Keterbukaan Informasi Publik)
 - C. UU PDP (Pelindungan Data Pribadi)
 - D. UU Hak Cipta
 - E. UU Telekomunikasi
54. Tindakan mengutip karya orang lain tanpa menyebutkan sumber aslinya dalam sebuah laporan ilmiah disebut...
- A. Sitasi
 - B. Plagiarisme
 - C. Parafrase
 - D. Referensi
 - E. Kolaborasi
55. Variabel dalam penelitian yang nilainya dipengaruhi oleh variabel lain disebut...
- A. Variabel Independen (Bebas)
 - B. Variabel Dependen (Terikat)
 - C. Variabel Kontrol
 - D. Variabel Moderator
 - E. Variabel Intervening
56. Sistem pengenal penulis unik yang membedakan peneliti satu dengan peneliti lain yang memiliki nama sama (misalnya "J. Smith") adalah...
- A. DOI
 - B. Scopus ID
 - C. ORCID (Open Researcher and Contributor ID)
 - D. ISSN
 - E. PMID
57. Repositori data ilmiah adalah tempat penyimpanan terpusat untuk...
- A. Mengarsipkan perangkat lunak bajakan
 - B. Menyimpan, mengelola, dan membagikan dataset penelitian
 - C. Menulis blog pribadi
 - D. Menyimpan email korespondensi
 - E. Mengelola keuangan proyek
58. Proses ETL dalam *Data Warehousing* adalah singkatan dari...
- A. Extract, Transform, Load

- B. Enter, Transfer, Lock
 - C. Evaluate, Test, Launch
 - D. Extract, Test, Learn
 - E. Edit, Type, Load
59. Teknik mengubah variabel kategorikal menjadi beberapa kolom biner (0 dan 1) agar dapat digunakan dalam model regresi disebut...
- A. Label Encoding
 - B. One-Hot Encoding
 - C. Binning
 - D. Scaling
 - E. Parsing
60. Validitas data mengacu pada...
- A. Seberapa cepat data diproses
 - B. Seberapa besar ukuran file data
 - C. Seberapa akurat data tersebut mengukur apa yang seharusnya diukur
 - D. Seberapa konsisten hasil pengukuran jika diulang
 - E. Seberapa aman penyimpanan data
61. *Feature Engineering* adalah proses...
- A. Membeli perangkat keras komputer baru
 - B. Membuat fitur/variabel baru dari data mentah untuk meningkatkan performa model
 - C. Menghapus data yang rusak
 - D. Menginstal software analisis
 - E. Memvisualisasikan data
62. Jika sebuah dataset memiliki kelas target yang sangat tidak seimbang (misal: 95% kelas A dan 5% kelas B), teknik *resampling* yang mengurangi jumlah sampel kelas mayoritas disebut...
- A. Oversampling
 - B. Undersampling
 - C. Bootstrapping
 - D. Cross-validation
 - E. Random Sampling
63. Format pertukaran data yang ringan, mudah dibaca manusia, dan menggunakan struktur pasangan *key-value* adalah...

- A. XML
- B. JSON
- C. Binary
- D. SQL
- E. HTML

64. Diagram yang digunakan untuk merancang struktur basis data dengan menggambarkan entitas, atribut, dan hubungan antar entitas adalah...

- A. Flowchart
- B. ERD (Entity Relationship Diagram)
- C. Use Case Diagram
- D. Gantt Chart
- E. Pie Chart

65. Proses normalisasi basis data bertujuan utama untuk...

- A. Mempercepat proses input data
- B. Menghilangkan anomali data dan mengurangi redundansi
- C. Menggabungkan semua tabel menjadi satu
- D. Membuat tampilan antarmuka lebih menarik
- E. Mengenkripsi password

66. Jenis *Join* dalam SQL yang hanya mengembalikan baris data yang memiliki nilai yang cocok di kedua tabel yang dihubungkan adalah...

- A. LEFT JOIN
- B. RIGHT JOIN
- C. INNER JOIN
- D. FULL OUTER JOIN
- E. CROSS JOIN

67. Fungsi agregasi dalam SQL untuk menghitung nilai rata-rata dari sebuah kolom numerik adalah...

- A. SUM()
- B. COUNT()
- C. AVG()
- D. MAX()
- E. MIN()

68. Basis data NoSQL yang menyimpan data dalam bentuk dokumen (seperti JSON/BSON) contohnya adalah...
- A. MySQL
 - B. PostgreSQL
 - C. MongoDB
 - D. Oracle
 - E. Microsoft SQL Server
69. Klausa SQL yang digunakan untuk memfilter hasil *setelah* dilakukan pengelompokan (GROUP BY) adalah...
- A. WHERE
 - B. HAVING
 - C. ORDER BY
 - D. LIMIT
 - E. FROM
70. Tabel yang digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan prediksi model terhadap nilai sebenarnya (True Positive, False Positive, dll) disebut...
- A. Pivot Table
 - B. Confusion Matrix
 - C. ANOVA Table
 - D. Frequency Table
 - E. Correlation Matrix
71. Metrik evaluasi yang paling tepat untuk kasus klasifikasi dengan *imbalanced dataset* (data tidak seimbang) adalah...
- A. Accuracy
 - B. F1-Score
 - C. Mean Squared Error (MSE)
 - D. R-Squared
 - E. Euclidean Distance
72. Teknik *Dimensionality Reduction* (Reduksi Dimensi) yang memproyeksikan data ke dalam komponen utama untuk mempertahankan varians terbesar disebut...
- A. K-Means
 - B. PCA (Principal Component Analysis)

- C. Linear Regression
- D. Decision Tree
- E. Naive Bayes

73. Algoritma *K-Nearest Neighbors* (KNN) mengklasifikasikan data baru berdasarkan...
- A. Aturan if-then yang kompleks
 - B. Probabilitas statistik
 - C. Jarak terdekat dengan sejumlah K tetangga dalam data latih
 - D. Garis pemisah linear
 - E. Struktur pohon keputusan
74. Metode evaluasi model di mana dataset dibagi menjadi K bagian, dan setiap bagian digunakan sekali sebagai data uji (test) sementara sisanya sebagai data latih (train) disebut...
- A. Hold-out Validation
 - B. K-Fold Cross Validation
 - C. Random Subsampling
 - D. Bootstrap
 - E. Time Series Split
75. *Reinforcement Learning* adalah paradigma pembelajaran mesin yang berbasis pada konsep...
- A. Data berlabel
 - B. Pengelompokan data
 - C. Agen, Lingkungan, Aksi, dan Reward (Penghargaan)
 - D. Persamaan regresi
 - E. Analisis teks
76. *Ensemble Learning* yang menggabungkan banyak *Decision Tree* untuk meningkatkan akurasi dan mengurangi *overfitting* dikenal dengan nama...
- A. Random Forest
 - B. Logistic Regression
 - C. SVM (Support Vector Machine)
 - D. K-Means
 - E. Apriori
77. Deep Learning adalah cabang dari Machine Learning yang menggunakan arsitektur...
- A. Pohon Keputusan
 - B. Artificial Neural Networks (Jaringan Saraf Tiruan) dengan banyak lapisan

- C. Aturan Asosiasi
- D. Statistik Bayesian
- E. Logika Fuzzy

78. Dalam analisis deret waktu (*Time Series*), pola berulang yang terjadi dalam periode waktu tetap (misalnya penjualan meningkat setiap lebaran) disebut...
- A. Trend
 - B. Seasonality (Musiman)
 - C. Cyclical
 - D. Irregular
 - E. Stationary
79. Uji statistik non-parametrik yang digunakan untuk membandingkan dua kelompok independen ketika data tidak berdistribusi normal adalah...
- A. T-Test Independen
 - B. ANOVA
 - C. Mann-Whitney U Test
 - D. Paired T-Test
 - E. Chi-Square Test
80. Teknik analisis data eksploratif (EDA) bertujuan untuk...
- A. Membuat kesimpulan akhir
 - B. Memahami karakteristik data, pola, dan anomali sebelum pemodelan
 - C. Melaporkan hasil ke atasan
 - D. Menghapus database
 - E. Membuat website
81. *A/B Testing* sering digunakan dalam eksperimen data untuk...
- A. Menguji validitas kuesioner
 - B. Membandingkan performa dua varian (A dan B) untuk menentukan mana yang lebih baik
 - C. Menggabungkan dua dataset
 - D. Menghitung biaya proyek
 - E. Melakukan backup data
82. Distribusi probabilitas berbentuk lonceng simetris di mana Mean, Median, dan Modus berada di titik yang sama disebut...
- A. Distribusi Binomial

- B. Distribusi Poisson
 - C. Distribusi Normal (Gaussian)
 - D. Distribusi Eksponensial
 - E. Distribusi Uniform
83. Jika Anda ingin memvisualisasikan korelasi antar banyak variabel numerik sekaligus dalam satu gambar, grafik yang paling tepat adalah...
- A. Bar Chart
 - B. Heatmap (Peta Panas) Korelasi
 - C. Pie Chart
 - D. Area Chart
 - E. Gantt Chart
84. Metode statistik untuk membandingkan rata-rata lebih dari dua kelompok secara bersamaan adalah...
- A. T-Test
 - B. Z-Test
 - C. ANOVA (Analysis of Variance)
 - D. Korelasi Pearson
 - E. Regresi Logistik
85. *Data Storytelling* adalah kemampuan untuk...
- A. Membaca data dengan cepat
 - B. Mengomunikasikan wawasan data (insights) secara naratif dan visual agar mudah dipahami pengambil keputusan
 - C. Menulis kode program yang panjang
 - D. Menyimpan data dalam cerita fiksi
 - E. Membuat dongeng tentang komputer
86. Proses *Tokenization* dalam *Natural Language Processing* (NLP) adalah...
- A. Menerjemahkan bahasa
 - B. Memecah teks kalimat menjadi potongan-potongan kata atau frasa (token)
 - C. Menghitung jumlah huruf
 - D. Mengubah suara menjadi teks
 - E. Mengoreksi ejaan
87. Apa yang dimaksud dengan *Data Governance* (Tata Kelola Data)?

- A. Pemerintahan yang menggunakan data
 - B. Kebijakan, standar, dan prosedur untuk memastikan data dikelola dengan baik, aman, dan berkualitas di seluruh organisasi
 - C. Perangkat lunak antivirus
 - D. Hardware server pemerintah
 - E. Statistik nasional
88. Teknologi komputasi yang memungkinkan analisis data dilakukan melalui internet menggunakan sumber daya server jarak jauh (bukan di komputer lokal) disebut...
- A. Quantum Computing
 - B. Edge Computing
 - C. Cloud Computing
 - D. Mainframe
 - E. Embedded System
89. Dalam visualisasi data, prinsip "Data-Ink Ratio" yang dikemukakan Edward Tufte menyarankan agar...
- A. Menggunakan tinta sebanyak mungkin agar grafik terlihat penuh
 - B. Mengurangi elemen visual yang tidak perlu (dekorasi) untuk menonjolkan data inti
 - C. Selalu menggunakan latar belakang berwarna gelap
 - D. Menggunakan efek 3D pada semua grafik
 - E. Mencetak grafik di kertas tebal
90. API (Application Programming Interface) dalam konteks pengambilan data berfungsi sebagai...
- A. Database itu sendiri
 - B. Antarmuka yang memungkinkan perangkat lunak berkomunikasi dan bertukar data dengan perangkat lunak lain
 - C. Alat pembersih virus
 - D. Layar monitor
 - E. Harddisk eksternal
91. *Bias Sampling* terjadi ketika...
- A. Sampel yang diambil benar-benar acak
 - B. Sampel yang diambil tidak merepresentasikan populasi yang sesungguhnya
 - C. Jumlah sampel terlalu banyak
 - D. Peneliti tidak memiliki hipotesis

- E. Data dihitung dengan kalkulator
92. Perintah *Git* yang digunakan untuk menyimpan perubahan kode/skrip analisis ke dalam repositori lokal adalah...
- A. git pull
 - B. git clone
 - C. git commit
 - D. git push
 - E. git merge
93. Sebuah model regresi memiliki nilai R^2 (Koefisien Determinasi) sebesar 0.85. Ini berarti...
- A. Model memiliki akurasi 15%
 - B. 85% variasi dalam variabel dependen dapat dijelaskan oleh variabel independen dalam model
 - C. Terdapat kesalahan sebesar 85%
 - D. Model tidak layak digunakan
 - E. Korelasi antar variabel adalah negatif
94. *Dashboard* interaktif untuk pemantauan data secara *real-time* biasanya dibuat menggunakan *tools* seperti...
- A. Microsoft Word
 - B. Adobe Reader
 - C. Tableau atau Power BI
 - D. Notepad
 - E. Paint
95. *Root Mean Squared Error* (RMSE) adalah metrik evaluasi yang umum digunakan untuk masalah...
- A. Klasifikasi
 - B. Regresi
 - C. Klasterisasi
 - D. Asosiasi
 - E. NLP
96. Standar metadata yang umum digunakan untuk mendeskripsikan sumber daya digital (seperti judul, pencipta, subjek) adalah...
- A. ISO 9001
 - B. Dublin Core

- C. HTML5
 - D. TCP/IP
 - E. ASCII
97. Jenis data spasial yang merepresentasikan bentuk geografis menggunakan titik, garis, dan poligon disebut data...
- A. Raster
 - B. Vektor
 - C. Pixel
 - D. Voxel
 - E. Bitmap
98. Dalam analisis sentimen, komputer mengklasifikasikan teks (misal: ulasan produk) ke dalam kategori...
- A. Panjang, Pendek, Sedang
 - B. Positif, Negatif, Netral
 - C. Mahal, Murah
 - D. Lama, Baru
 - E. Lokal, Internasional
99. Konsep *Interoperabilitas* data berarti...
- A. Data hanya bisa dibuka di satu komputer
 - B. Kemampuan sistem atau perangkat lunak yang berbeda untuk bertukar dan menggunakan data tersebut
 - C. Data dilindungi password
 - D. Data tidak bisa diubah
 - E. Data memiliki ukuran kecil
100. Jika $p\text{-value} > 0.05$ (lebih besar dari taraf signifikansi), keputusan yang diambil adalah...
- A. Menolak H_0
 - B. Gagal menolak H_0 (Terima H_0)
 - C. Data salah
 - D. Hipotesis alternatif diterima
 - E. Ulangi penelitian

PEMBAHASAN

1. Jawaban: A

- **Pembahasan:** Hirarki DIKW adalah model yang menggambarkan struktur data menjadi kebijaksanaan. Urutannya adalah **Data** (fakta mentah) → **Informasi** (data yang diberi konteks) → **Pengetahuan** (informasi yang diberi makna/pengalaman) → **Kebijaksanaan** (penerapan pengetahuan untuk pengambilan keputusan).

2. Jawaban: C

- **Pembahasan:** Prinsip **FAIR** adalah standar global dalam pengelolaan data ilmiah agar dapat digunakan kembali oleh mesin dan manusia. FAIR singkatan dari *Findable* (Dapat ditemukan), *Accessible* (Dapat diakses), *Interoperable* (Dapat dioperasikan/digabungkan), dan *Reusable* (Dapat digunakan kembali).

3. Jawaban: C

- **Pembahasan:**
 - **AND:** Menggabungkan dua kata kunci, hasil harus memuat keduanya (mempersempit pencarian).
 - **OR:** Salah satu kata kunci muncul (memperluas pencarian).
 - **NOT:** Mengecualikan kata kunci tertentu.

4. Jawaban: D

- **Pembahasan:** **DOI** (*Digital Object Identifier*) adalah alamat unik permanen untuk artikel jurnal/dokumen digital. ISBN untuk buku, ISSN untuk seri majalah/jurnal, URL adalah alamat web biasa yang bisa berubah (link rot).

5. Jawaban: B

- **Pembahasan:** Dalam metode ilmiah, sebelum mengumpulkan atau membersihkan data, seorang analis harus terlebih dahulu **merumuskan pertanyaan penelitian atau hipotesis**. Ini menjadi panduan arah analisis.

6. Jawaban: C

- **Pembahasan:** Data sekunder adalah data yang sudah tersedia karena dikumpulkan oleh pihak lain sebelumnya (misal: data BPS, data satelit NASA). Data primer adalah data yang dikumpulkan sendiri (opsi A dan B).

7. Jawaban: C

- **Pembahasan:** **Fabrikasi** (membuat data palsu) dan **Falsifikasi** (memanipulasi/mengubah data agar sesuai keinginan) adalah pelanggaran etika penelitian berat. Data Mining dan Cleaning adalah proses teknis yang sah.

8. Jawaban: B

- **Pembahasan:** **Metadata** adalah "data tentang data". Contohnya: siapa penulisnya, kapan data dibuat, apa format filenya, dan apa satuan ukurannya. Metadata memudahkan pencarian dan pengelolaan data.

9. Jawaban: C

- **Pembahasan:** Preservasi data (*Data Preservation*) adalah serangkaian kegiatan untuk memastikan data tetap tersimpan aman, tidak rusak, dan dapat diakses dalam jangka waktu yang panjang (arsip).

10. Jawaban: D

- **Pembahasan:** *Grey Literature* adalah literatur yang tidak diterbitkan melalui jalur penerbit komersial formal. Contoh: Laporan teknis, skripsi, dokumen pemerintah. **Artikel jurnal bereputasi** termasuk *White Literature* (publikasi formal).

11. Jawaban: C

- **Pembahasan:** **Mendeley** dan **Zotero** adalah perangkat lunak *Reference Manager* yang populer untuk mengelola sitasi dan daftar pustaka secara otomatis.

12. Jawaban: C

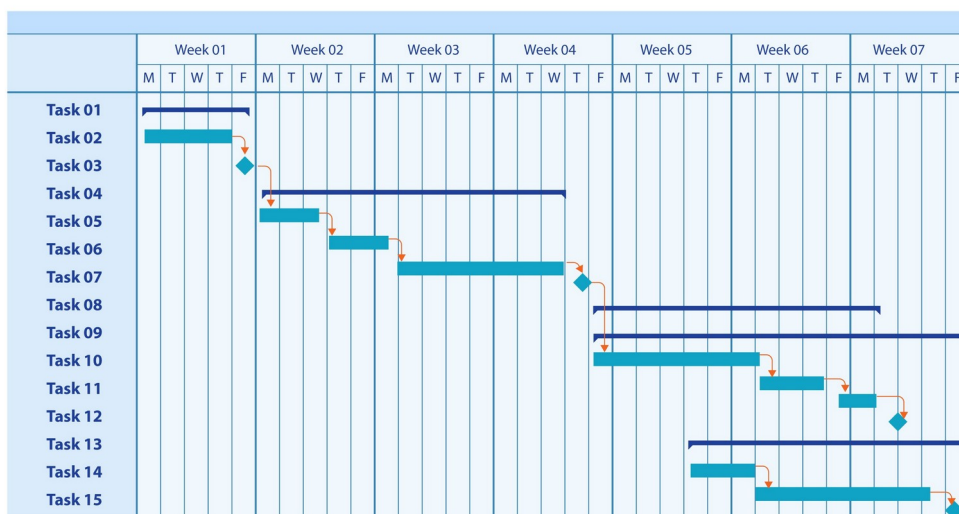
- **Pembahasan:** Konsep 3V/5V dalam Big Data:
 - **Velocity:** Kecepatan data masuk/diproses.
 - **Volume:** Besar ukuran data.
 - **Variety:** Beragam jenis data.

13. Jawaban: D

- **Pembahasan:** **Open Science** bertujuan membuat pengetahuan ilmiah, data, dan diseminasinya dapat diakses oleh semua tingkatan masyarakat (amatir maupun profesional) secara transparan dan gratis.

14. Jawaban: B

- **Pembahasan:** **Gantt Chart** adalah diagram batang horizontal yang memetakan jadwal tugas proyek terhadap waktu (kapan mulai, kapan selesai).



15. Jawaban: C

- **Pembahasan: Reproducibility** (Keterulangan) adalah prinsip utama sains modern. Artinya, jika analis lain menggunakan kode dan data yang sama dengan Anda, mereka harus mendapatkan hasil yang persis sama.

16. Jawaban: B

- **Pembahasan: Data Kuantitatif** adalah data numerik (angka) yang bisa dihitung/diukur. Data Kualitatif bersifat deskriptif (kata-kata).

17. Jawaban: B

- **Pembahasan: Imputasi** adalah teknik statistik untuk mengisi nilai yang kosong (*missing values*) dengan nilai pengganti yang masuk akal (seperti mean, median, atau modus) agar data bisa diproses.

18. Jawaban: C

- **Pembahasan: Outlier** (Pencilan) adalah data yang nilainya sangat berbeda (jauh lebih kecil atau jauh lebih besar) dari sebagian besar data lainnya dalam dataset.

19. Jawaban: D

- **Pembahasan: Simple Random Sampling** memberikan peluang yang sama bagi setiap anggota populasi untuk terpilih menjadi sampel, sehingga meminimalisir bias.

20. Jawaban: B

- **Pembahasan: Encoding** (seperti Label Encoding) mengubah data kategori (teks) menjadi angka agar bisa dipahami oleh mesin. Decoding adalah kebalikannya.

21. Jawaban: D

- **Pembahasan: Skala Rasio** memiliki nilai nol mutlak (contoh: berat 0 kg berarti tidak ada berat). Interval tidak memiliki nol mutlak (contoh: suhu 0 derajat Celcius bukan berarti tidak ada suhu).

22. Jawaban: B

- **Pembahasan: Min-Max Scaling** (Normalisasi) mengubah rentang data asli menjadi rentang spesifik, biasanya antara 0 hingga 1, dengan rumus:
$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

23. Jawaban: C

- **Pembahasan: Basis Data (Database)** adalah kumpulan data yang terorganisir, biasanya disimpan dan diakses secara elektronik dari sistem komputer.

24. Jawaban: B

- **Pembahasan: Primary Key** adalah kolom (atau kumpulan kolom) yang secara unik mengidentifikasi setiap baris dalam tabel. Tidak boleh ada nilai duplikat atau NULL dalam Primary Key.

25. Jawaban: D

- **Pembahasan:** **SELECT** adalah perintah DML (*Data Manipulation Language*) dalam SQL yang digunakan untuk *query* atau mengambil data dari database.

26. Jawaban: B

- **Pembahasan:** **RDBMS** (*Relational Database Management System*) menyimpan data dalam struktur tabel (baris dan kolom) yang saling berelasi. Contoh: MySQL, PostgreSQL. NoSQL biasanya tidak menggunakan tabel kaku.

27. Jawaban: C

- **Pembahasan:** **Redundansi data** adalah kondisi di mana data yang sama disimpan berulang-ulang di beberapa tempat, yang dapat menyebabkan pemborosan penyimpanan dan inkonsistensi data.

28. Jawaban: D

- **Pembahasan:** **CSV** (*Comma Separated Values*) adalah format file teks standar yang sangat ringan untuk menyimpan data tabular, di mana setiap kolom dipisahkan oleh koma.

29. Jawaban: D

- **Pembahasan:** Hubungan **Many-to-Many**. Satu mahasiswa bisa ambil banyak mapel, satu mapel bisa diambil banyak mahasiswa. Biasanya membutuhkan tabel perantara (*junction table*) dalam implementasinya.

30. Jawaban: C

- **Pembahasan:** **Backup** adalah proses penyalinan data untuk cadangan. *Restore* atau *Recovery* adalah proses mengembalikannya saat terjadi kerusakan.

31. Jawaban: C

- **Pembahasan:** **Clustering** (Klasterisasi) adalah metode *Unsupervised Learning* (tanpa label) untuk mengelompokkan data berdasarkan kemiripan. Klasifikasi adalah *Supervised Learning* (ada label kelas).

32. Jawaban: C

- **Pembahasan:** **Linear Regression** adalah *Supervised Learning* karena model dilatih menggunakan data yang memiliki target/label (variabel dependen). K-Means, Hierarchical Clustering, dan Apriori adalah *Unsupervised*.

33. Jawaban: A

- **Pembahasan:** **Asosiasi** (Association Rule Mining) digunakan untuk menemukan pola hubungan antar item, contohnya *Market Basket Analysis* ("Orang yang beli Roti biasanya juga beli Selai").

34. Jawaban: B

- **Pembahasan:** **Overfitting** terjadi ketika model "menghafal" data latih terlalu detail (termasuk *noise*-nya), sehingga tidak bisa melakukan generalisasi pada data baru.

35. Jawaban: B

- **Pembahasan: Decision Tree** adalah algoritma serbaguna yang bisa digunakan baik untuk **Klasifikasi** (target kategori) maupun **Regresi** (target angka kontinu).

36. Jawaban: C

- **Pembahasan:** Membagi data menjadi *Train/Test* penting untuk **evaluasi**. Kita melatih model dengan data *Train*, lalu menguji kehebatannya dengan data *Test* (data yang belum pernah dilihat model) untuk memastikan model valid.

37. Jawaban: C

- **Pembahasan: Text Mining** (Penambangan Teks) berfokus pada ekstraksi informasi dari data teks tidak terstruktur (artikel, email, media sosial).

38. Jawaban: C

- **Pembahasan: Median** (nilai tengah) adalah ukuran pemusatan yang *robust* (tahan) terhadap *outlier*. Mean (rata-rata) akan sangat terdistorsi jika ada nilai ekstrim.

39. Jawaban: B

- **Pembahasan: Standar Deviasi** mengukur seberapa menyebar data dari rata-ratanya. Semakin besar standar deviasi, semakin bervariasi datanya.

40. Jawaban: C

- **Pembahasan:** Nilai korelasi berkisar antara -1 hingga 1.
 - Mendekati 1: Positif kuat.
 - Mendekati -1: **Negatif kuat**.
 - Mendekati 0: Lemah/tidak ada hubungan.

41. Jawaban: B

- **Pembahasan:** Dalam statistik konvensional, jika $p\text{-value} < 0.05$ (taraf signifikansi), maka kita **Menolak H_0** . Artinya, hasil tersebut signifikan secara statistik (bukan kebetulan).

42. Jawaban: C

- **Pembahasan: Histogram** digunakan untuk melihat distribusi frekuensi data numerik (misal: sebaran umur penduduk). Bar chart untuk data kategori.

CHART TYPES



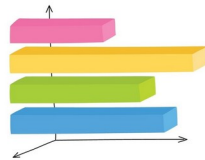
3-D Column



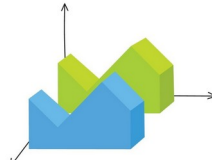
3-D Pie



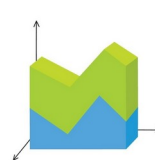
3-D Line



3-D Bar



3-D Area



3-D Area

43. Jawaban: C

- **Pembahasan:** **Scatter Plot** memetakan titik-titik data pada sumbu X dan Y untuk melihat pola hubungan atau **korelasi** antara dua variabel numerik.

44. Jawaban: A

- **Pembahasan:** Istilah latinnya *Cum hoc ergo propter hoc* ("bersama ini, maka karena ini"). Hanya karena dua hal berkorelasi (terjadi bersamaan), **belum tentu** ada hubungan sebab-akibat.

45. Jawaban: B

- **Pembahasan:** Regresi Linear Sederhana memodelkan hubungan antara satu variabel independen (X) dan satu variabel dependen (Y) untuk melakukan **prediksi**.

46. Jawaban: E

- **Pembahasan:** **Boxplot** menampilkan "Five-number summary": Minimum, Q1, Median, Q3, Maksimum, dan Outlier. Boxplot **tidak** bisa menunjukkan hubungan sebab akibat.

47. Jawaban: C

- **Pembahasan:** **Hipotesis Nol (H0)** adalah asumsi awal (default) yang menyatakan **tidak ada perbedaan** atau tidak ada pengaruh. Peneliti berusaha mencari bukti untuk menolak H0.

48. Jawaban: C

- **Pembahasan:** **PCA** (*Principal Component Analysis*) adalah teknik reduksi dimensi yang mengubah variabel-variabel asli yang mungkin berkorelasi menjadi sejumlah kecil variabel baru yang tidak berkorelasi (Principal Components).

49. Jawaban: A

- **Pembahasan:**
 - **Tipe I (False Positive):** Menolak H0 padahal H0 benar (Menghukum orang tidak bersalah).

- Tipe II (False Negative): Menerima H_0 padahal H_0 salah (Membebaskan orang bersalah).

50. Jawaban: B

- **Pembahasan: Truncated Axis** (Sumbu yang dipotong, tidak mulai dari 0) sering digunakan untuk memanipulasi persepsi visual, membuat perbedaan kecil terlihat sangat besar/dramatis. Ini adalah praktik visualisasi yang buruk (*misleading*).

51. Jawaban: D

- **Pembahasan: Python** adalah bahasa pemrograman paling populer untuk *Data Science* saat ini karena sintaksnya yang mudah dibaca dan memiliki ekosistem *library* yang sangat luas (Pandas, NumPy, Scikit-learn, TensorFlow).

52. Jawaban: B

- **Pembahasan:**
 - **Gold Open Access:** Penulis membayar (APC) agar artikel gratis bagi pembaca.
 - **Green Open Access:** Penulis mengarsipkan versi *preprint/postprint* di repositori (gratis).
 - **Diamond/Platinum:** Gratis bagi penulis dan pembaca (biasanya didanai lembaga).

53. Jawaban: C

- **Pembahasan: UU PDP (Pelindungan Data Pribadi)** atau UU No. 27 Tahun 2022 adalah payung hukum utama di Indonesia yang mengatur hak subjek data dan kewajiban pengendali data.

54. Jawaban: B

- **Pembahasan: Plagiarisme** adalah pengambilan karangan (pendapat dsb.) orang lain dan menjadikannya seolah-olah karangan (pendapat dsb.) sendiri tanpa mencantumkan sumber (sitasi).

55. Jawaban: B

- **Pembahasan: Variabel Dependen (Terikat)** adalah variabel yang dipengaruhi atau yang menjadi akibat, karena adanya variabel bebas (independen). Dalam rumus $Y = a + bX$, Y adalah dependen.

56. Jawaban: C

- **Pembahasan: ORCID (Open Researcher and Contributor ID)** menyediakan kode alfanumerik persisten untuk membedakan satu peneliti dengan peneliti lain secara unik, mengatasi masalah kesamaan nama.

57. Jawaban: B

- **Pembahasan: Repositori Data Ilmiah** (seperti Zenodo, Figshare, atau RIN) adalah infrastruktur untuk menyimpan, mengelola, dan mempublikasikan dataset penelitian agar sesuai prinsip FAIR.

58. Jawaban: A

- **Pembahasan:** **ETL** (*Extract, Transform, Load*) adalah proses inti dalam integrasi data: mengambil data dari sumber (*Extract*), membersihkan/mengubah formatnya (*Transform*), dan menyimpannya ke *Data Warehouse* (*Load*).

59. Jawaban: B

- **Pembahasan:** **One-Hot Encoding** mengubah variabel kategori (misal: "Merah", "Biru") menjadi kolom biner terpisah (Kolom "Is_Merah": 1/0, Kolom "Is_Biru": 1/0). Ini penting karena mesin tidak bisa memproses teks secara langsung.

60. Jawaban: C

- **Pembahasan:** **Validitas** berkaitan dengan ketepatan/akurasi (apakah alat ukur mengukur apa yang seharusnya diukur). **Reliabilitas** (opsi D) berkaitan dengan konsistensi.

61. Jawaban: B

- **Pembahasan:** **Feature Engineering** adalah seni menggunakan pengetahuan domain untuk mengekstrak fitur (variabel) baru dari data mentah guna meningkatkan kekuatan prediksi algoritma *machine learning*.

62. Jawaban: B

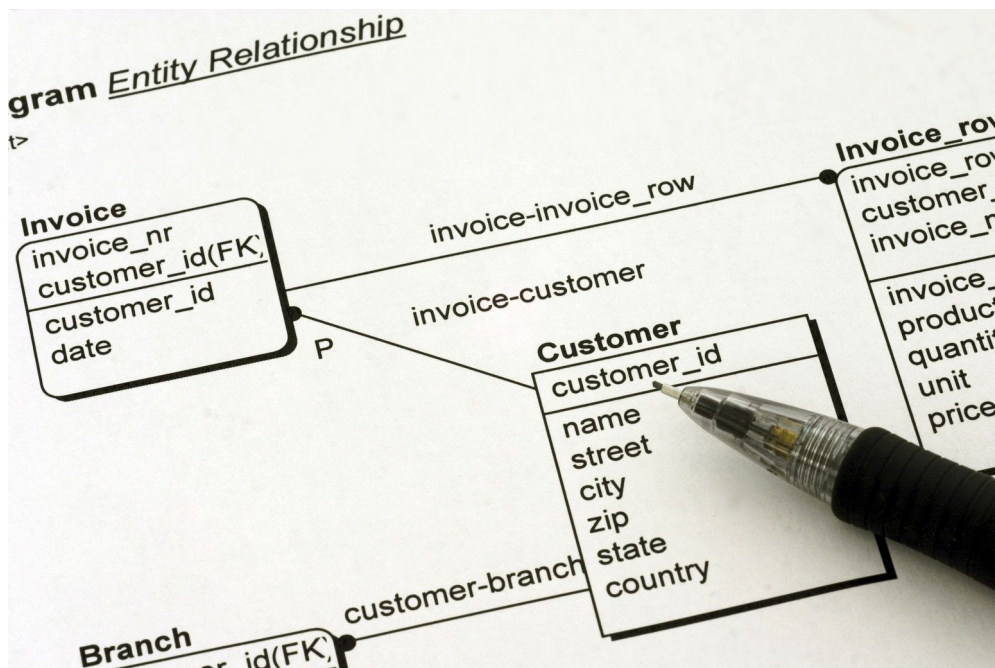
- **Pembahasan:**
 - **Undersampling:** Mengurangi jumlah data kelas mayoritas agar seimbang dengan minoritas.
 - **Oversampling:** Menambah/menduplikasi data kelas minoritas.

63. Jawaban: B

- **Pembahasan:** **JSON** (*JavaScript Object Notation*) adalah format pertukaran data yang sangat populer di web API karena ringan dan mudah dibaca manusia maupun mesin. XML lebih berat (*verbose*).

64. Jawaban: B

- **Pembahasan:** **ERD** (*Entity Relationship Diagram*) adalah model visual untuk menggambarkan struktur database, menunjukkan entitas (tabel), atribut (kolom), dan relasi (garis hubung).



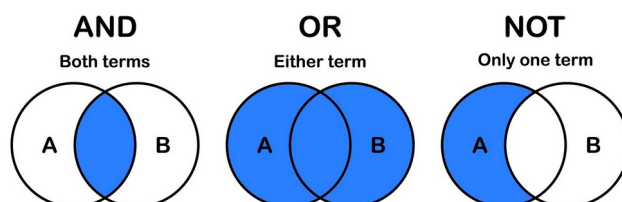
65. Jawaban: B

- **Pembahasan: Normalisasi** (1NF, 2NF, 3NF, dst) adalah teknik desain database untuk meminimalkan redundansi (duplikasi) dan mencegah anomali data (masalah saat insert/update/delete).

66. Jawaban: C

- **Pembahasan:**
 - **INNER JOIN:** Hanya data yang cocok di kedua tabel.
 - **LEFT JOIN:** Semua data kiri + yang cocok di kanan.
 - **RIGHT JOIN:** Semua data kanan + yang cocok di kiri.

BOOLEAN LOGIC



67. Jawaban: C

- **Pembahasan:** AVG() untuk rata-rata, SUM() untuk total, COUNT() untuk jumlah baris, MAX() nilai tertinggi, MIN() nilai terendah.

68. Jawaban: C

- **Pembahasan:** MongoDB adalah database NoSQL berbasis dokumen (*Document-oriented*) yang populer. MySQL, PostgreSQL, dan Oracle adalah database Relasional (SQL).

69. Jawaban: B

- **Pembahasan:** WHERE digunakan untuk memfilter baris *sebelum* pengelompokan. HAVING digunakan untuk memfilter hasil *setelah* pengelompokan (GROUP BY).

70. Jawaban: B

- **Pembahasan:** Confusion Matrix adalah tabel 2x2 (untuk biner) yang menunjukkan jumlah prediksi Benar Positif (TP), Benar Negatif (TN), Salah Positif (FP), dan Salah Negatif (FN).

71. Jawaban: B

- **Pembahasan:** Pada *Imbalanced Dataset*, Akurasi bisa menipu (model memprediksi kelas mayoritas terus menerus akan tetap punya akurasi tinggi). **F1-Score** (rata-rata harmonik Precision dan Recall) lebih objektif.

72. Jawaban: B

- **Pembahasan:** PCA adalah teknik statistik untuk mereduksi dimensi data dengan mengubahnya menjadi kumpulan variabel baru (Principal Components) yang saling tidak berkorelasi namun tetap mempertahankan varians data sebanyak mungkin.

73. Jawaban: C

- **Pembahasan:** KNN (*K-Nearest Neighbors*) bekerja dengan prinsip kesamaan jarak. Data baru diklasifikasikan mengikuti mayoritas label dari 'K' tetangga terdekatnya.

74. Jawaban: B

- **Pembahasan:** K-Fold Cross Validation membagi data menjadi K bagian (folds). Proses *training-testing* diulang K kali, sehingga semua data pernah menjadi data uji. Ini memberikan evaluasi model yang lebih akurat.

75. Jawaban: C

- **Pembahasan:** Reinforcement Learning belajar melalui interaksi *trial and error*. Agen melakukan Aksi di Lingkungan, lalu mendapatkan Reward (hadiah) atau penalti, untuk memaksimalkan total reward.

76. Jawaban: A

- **Pembahasan:** Random Forest adalah metode *Ensemble* (gabungan) yang terdiri dari banyak *Decision Tree*. Hasil akhirnya diambil dari suara terbanyak (voting) pohon-pohon tersebut, sehingga lebih stabil dan akurat.

77. Jawaban: B

- **Pembahasan: Deep Learning** menggunakan **Artificial Neural Networks** (Jaringan Saraf Tiruan) dengan banyak lapisan tersembunyi (*hidden layers*) untuk mempelajari representasi fitur secara otomatis dari data.

78. Jawaban: B

- **Pembahasan: Seasonality** (Musiman) adalah pola fluktuasi yang berulang secara teratur dalam periode waktu tertentu (misal: penjualan es krim naik setiap musim panas, penjualan baju naik setiap lebaran).

79. Jawaban: C

- **Pembahasan:** Jika data tidak normal, T-Test tidak valid. Alternatif non-parametriknya adalah **Mann-Whitney U Test** (untuk 2 sampel independen) atau Wilcoxon (untuk berpasangan).

80. Jawaban: B

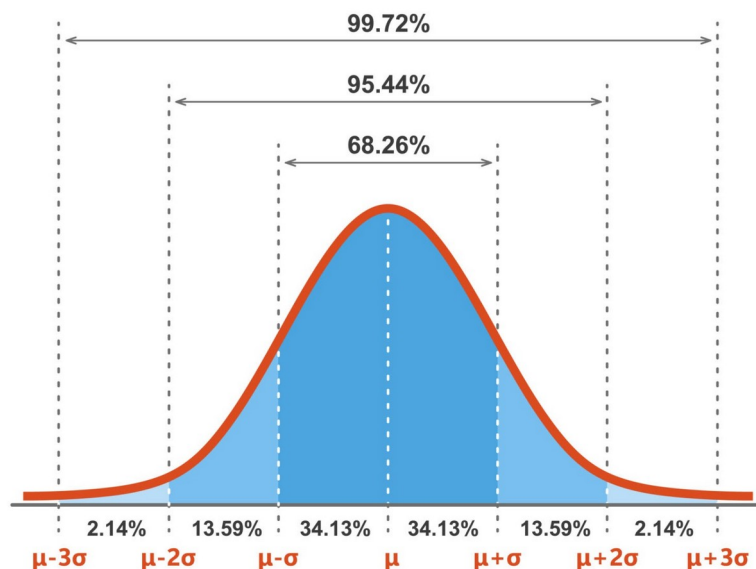
- **Pembahasan: EDA** (*Exploratory Data Analysis*) adalah langkah kritis awal untuk "berkenalan" dengan data (cek distribusi, korelasi, outlier) menggunakan statistik deskriptif dan visualisasi.

81. Jawaban: B

- **Pembahasan: A/B Testing** adalah eksperimen terkontrol (biasanya di web/marketing) di mana dua varian (A dan B) dibandingkan untuk melihat mana yang memberikan hasil (konversi) lebih baik.

82. Jawaban: C

- **Pembahasan: Distribusi Normal** (kurva lonceng) adalah distribusi terpenting dalam statistik. Sifat utamanya adalah simetris, dan Mean = Median = Modus.



83. Jawaban: B

- **Pembahasan: Heatmap Korelasi** menggunakan gradasi warna untuk menunjukkan kekuatan hubungan antar banyak variabel sekaligus, memudahkan identifikasi variabel mana yang saling berkorelasi kuat.

84. Jawaban: C

- **Pembahasan: ANOVA** (*Analysis of Variance*) digunakan untuk membandingkan rata-rata dari tiga kelompok atau lebih. T-Test hanya bisa untuk maksimal dua kelompok.

85. Jawaban: B

- **Pembahasan: Data Storytelling** adalah keterampilan lunak (*soft skill*) krusial bagi analis data untuk menerjemahkan hasil analisis yang rumit menjadi narasi yang mudah dipahami oleh pemangku kepentingan (*stakeholders*).

86. Jawaban: B

- **Pembahasan: Tokenization** adalah langkah awal NLP. Kalimat "Saya suka data" dipecah menjadi token ["Saya", "suka", "data"].

87. Jawaban: B

- **Pembahasan: Data Governance** memastikan data dikelola sebagai aset strategis organisasi, meliputi kualitas, keamanan, privasi, dan kepatuhan terhadap regulasi.

88. Jawaban: C

- **Pembahasan: Cloud Computing** (seperti AWS, Google Cloud, Azure) menyediakan daya komputasi dan penyimpanan via internet, memungkinkan analisis Big Data tanpa perlu memiliki superkomputer sendiri.

89. Jawaban: B

- **Pembahasan: Prinsip Data-Ink Ratio** menyarankan penghapusan tinta (elemen visual) yang tidak menyajikan informasi data (seperti *gridlines* tebal, latar belakang warna-warni, efek 3D) agar grafik bersih dan efisien.

90. Jawaban: B

- **Pembahasan: API** adalah jembatan komunikasi antar software. Bagi analis data, API sering digunakan untuk mengambil data dari platform pihak ketiga (misal: Twitter API, Google Maps API).

91. Jawaban: B

- **Pembahasan: Bias Sampling** (seperti *Selection Bias*) terjadi ketika metode pengambilan sampel menyebabkan karakteristik sampel berbeda secara sistematis dengan populasi, sehingga hasil analisis tidak bisa digeneralisasi.

92. Jawaban: C

- **Pembahasan:**
 - git commit: Menyimpan perubahan ke riwayat repositori lokal.
 - git push: Mengirim perubahan ke repositori remote (seperti GitHub).
 - git pull: Mengambil perubahan dari remote.

93. Jawaban: B

- **Pembahasan:** R^2 (**R-Squared**) atau Koefisien Determinasi menunjukkan proporsi varians variabel dependen yang dapat dijelaskan oleh model. Nilai 0.85 berarti 85% variasi data dapat dijelaskan oleh model (sangat baik).

94. Jawaban: C

- **Pembahasan:** **Tableau** dan **Power BI** adalah *tools Business Intelligence* (BI) terkemuka untuk membuat *dashboard* visualisasi data yang interaktif tanpa banyak *coding*.

95. Jawaban: B

- **Pembahasan:** **RMSE** mengukur rata-rata besarnya kesalahan prediksi pada model **Regresi** (data kontinu). Semakin kecil RMSE, semakin akurat modelnya. Untuk klasifikasi, digunakan Akurasi/F1-Score.

96. Jawaban: B

- **Pembahasan:** **Dublin Core** adalah skema metadata standar yang sederhana dan umum digunakan untuk mendeskripsikan berbagai sumber daya digital (perpustakaan digital, arsip web).

97. Jawaban: B

- **Pembahasan:**
 - **Vektor:** Merepresentasikan bumi dengan Titik, Garis, Poligon (peta jalan, batas wilayah).
 - **Raster:** Grid sel/piksel (citra satelit, foto udara).

98. Jawaban: B

- **Pembahasan:** **Analisis Sentimen** (*Opinion Mining*) bertujuan menentukan nada emosional di balik teks, umumnya diklasifikasikan menjadi sentimen Positif, Negatif, atau Netral.

99. Jawaban: B

- **Pembahasan:** **Interoperabilitas** adalah kunci dalam pertukaran data ilmiah. Sistem A harus bisa membaca dan memproses data dari Sistem B tanpa hambatan format atau struktur.

100. Jawaban: B

- **Pembahasan:** Jika $p\text{-value} > 0.05$ (taraf signifikansi), berarti tidak cukup bukti untuk menolak H_0 . Kesimpulannya adalah **Gagal Menolak H_0** (atau sering disebut Menerima H_0), yang berarti tidak ada efek/perbedaan signifikan.