NAMA : Ulfah Nuraini

NIM :4122026

KELAS : PAGI

UJIAN TENGAH SEMESTER MACHINE LEARNING

SOAL PILIHAN GANDA!

1. Pernyataan manakah yang benar tentang **pembelajaran terawasi (supervised learning)**?
   A. Data latihnya memiliki label kelas atau nilai target.
   B. Model belajar tanpa menggunakan label data.
   C. Mengandalkan interaksi agen dan lingkungan.
   D. Berfokus pada penguatan (reinforcement) dari hasil.
   Jawab: **A**. Data latihnya memiliki label kelas atau nilai target.

2. Manakah contoh tugas yang termasuk dalam **pembelajaran tanpa pengawasan (unsupervised learning)**?
   A. Klasifikasi email spam.
   B. Pengelompokan (clustering) data pelanggan berdasarkan karakteristik.
   C. Prediksi harga saham.
   D. Permainan catur oleh agen komputer.
   Jawab : **B.** Pengelompokan (clustering) data pelanggan berdasarkan karakteristik.

3. Manakah pernyataan yang benar tentang **pembelajaran penguatan (reinforcement learning)**?
   A. Memerlukan data pelatihan berlabel.
   B. Agen belajar melalui interaksi dengan lingkungan untuk menentukan kebijakan.
   C. Fokus pada pengelompokan data.
   D. Menggunakan fungsi aktivasi untuk pembelajaran.
   Jawab : **B.** Agen belajar melalui interaksi dengan lingkungan untuk menentukan kebijakan.

4. Pohon keputusan (**decision tree**) dapat digunakan untuk tugas...
   A. Klasifikasi data.
   B. Klasterisasi data.
   C. Regresi.
   D. Klasifikasi dan regresi.
   Jawab: **D.** Klasifikasi dan regresi.

5. Tujuan utama **regresi linear** adalah...
   A. Mengelompokkan data.
   B. Memprediksi label kelas.
   C. Memprediksi nilai kuantitatif kontinu.
   D. Mengurangi dimensi data.
   Jawab: **C.** Memprediksi nilai kuantitatif kontinu.

6. Contoh algoritma unsupervised learning yang paling terkenal adalah….
   A. Decision Tree
   B. Linear Regression

C. K-Means Clustering
D. Logic Regression
Jawab: **C.** K-Means Clustering

7. Perbedaan utama antara tugas klasifikasi dan regresi adalah….
   A. Klasikasi memprediksi label kategori sedangkan regresi memprediksi nilai kontinu
   B. Klasifikasi memprediksi nilai kontinu sedangkan regresi memprediksi kategori
   C. Keduanya memprediksi nilai kontinu.
   D. Keduanya memprediksi label kategori
   Jawab: **A.** Klasifikasi memprediksi label kategori sedangkan regresi memprediksi nilai kontinu

8. Berikut adalah langkah utama dalam algoritma **K-Means**, kecuali….
   A. Menentukan jumlah klaster (K) yang diinginkan.
   B. Menginisialisasi pusat klaster (centroid) secara acak.
   C. Membagi data latih menjadi beberapa subset acal
   D. Memperbarui posisi centroid berdasarkan rata-rata anggota klaster.
   Jawab : **C.** Membagi data latih menjadi beberapa subset acak

9. Kriteria yang biasa digunakan dalam pohon keputusan untuk memilih atribut terbaik adalah….
   A. Indeks Gini
   B. Fungsi aktivasi ReLU
   C. Gradient Descent.
   D. Jarak Euclidean.
   Jawab : **A.** Indeks Gini

10. **Overfitting** terjadi ketika….
    A. Model terlalu sederhana sehingga gagal menangkap pola utama data.
    B. Model terlalu rumit sehingga mengikuti noise dalam data latih.
    C. Dataset yang digunakan memiliki terlalu sedikit fitur.
    D. Data latih tidak dibagi dengan benar.
    Jawab : **B.** Model terlalu rumit sehingga mengikuti noise dalam data latih.

11. Principal Component Analysis (PCA) termasuk algoritma….
    A. Pembelajaran terawasi (supervised)
    B. Pembelajaran tidak terawasi (unsupervised).
    C. Pembelajaran penguatan (reinforcement)
    D. Optimasi (optimization).
    Jawab: **B.** Pembelajaran tidak terawasi (unsupervised).

12. Dalam pembelajaran penguatan (reinforcement leaming), **reward** atau hadiah adalah….
    A. Data latih berlabel yang diberikan kepada agen
    B. Sinyal yang menunjukkan seberapa baik tindakan agen.
    C. Algoritma yang digunakan dalam proses pembelajaran.
    D. Fungsi aktivasi pada jaringan syaraf.
    Jawab: **B.** Sinyal yang menunjukkan seberapa baik tindakan agen.

13. Tujuan utama pemisahan data mmjadi set pelatihan, validasi, dan pengujian adalah….

A. Melatih model, menyetel hyperparameter, dan mengukur kinerja model.
B. Mencegah model terlalu bergantung pada data latih.
C. Menambah variasi data pelatihan.
D. Menghapus fitur yang tidak relevan
Jawab : **A.** Melatih model, menyetel hyperparameter, dan mengukur kinerja model.

14. Dalam matriks kebingungan (conhasion matric), presisi (precision) didefinisikan sebagai…
A. TP / (TP+FN)
B. TP / (TP+FP)
C. TN / (TN+FP)
D. TN / (TN+FN)
Jawab : **B.** TP / (TP+FP)

15. Normalisasi Min Max pada data biasanya mengubah nital fitur ke rentang…
A. -1 hingga 1
B. -∞ hingga +∞
C. 0 hingga 1
D. 0 hingga 100
Jawab : **C.** 0 hingga 1

16. Algoritma K-Nearest Neighbors (KNN) termasuk jenis pembelajaran….
A. Tanpa pengawasan (unsupervised).
B. Penguatan (reinforcement).
C. Semi-terawasi (semi supervised).
D. Terawasi (supervised).
Jawab : **D.** Terawasi (supervised).

17. Indeks Gini dalam pohon keputusan digunakan untuk….
A. Mengukur variansi error dalam node.
B. Mengukur kemumian (purity) suatu node setelah split.
C. Menetapkan learning rate pada setiap cabang
D. Menentukan jumlah fitur yang digunakan.
Jawab: **B.** Mengukur kemurnian (purity) suatu node setelah split.

18. Dalam regresi dengan regularisası L1 (Lasso), biasanya diperoleh….
A. Banyak koefisien parameter menjadi nol (solusi jarang).
B. Semua koefisien parameter menjadi nol.
C. Tidak ada koefisien yang diubah.
D. Hanya satu fitur yang dipilih.
Jawab : **A.** Banyak koefisien parameter menjadi nol (solusi jarang).

19. Dalam Q-learning (reinforcement learning), **Q-table** digunakan untuk....
A. Menyimpan nilai-nilai tindakan (action values) untuk pasangan state-action.
B. Mengelompokkan state berdasarkan reward.
C. Menghitung fungsi biaya (loss).
D. Memodelkan interaksi agen-lingkungan
Jawab : **A.** Menyimpan nilai-nilai tindakan (action values) untuk pasangan state-action.

20. Kurva ROC (Receiver Operating Maracteristic) digunakan untuk….
    A. Menentukan threshold optima ara otomatis.
    B. Memvisualisasikan trade off antara true positive rate dan false positive rate.
    C. Memilih fitur yang paling informatif.
    D. Mengukur kesalahan absolut pada data regresi.
    Jawab: **B.** Memvisualisasikan trade-off antara true positive rate dan false positive rate.

21. Untuk mengatasi overfitting, salah satu cara yang benar adalah...
    A. Mengurangi jumlah fitur secara drastis.
    B. Meningkatkan jumlah terasi pelatihan tanpa perubahan lainnya.
    C. Menerapkan teknik dropout (pada neural network).
    D. Menambah jumlah data pelatihan atau menerapkan regularisasi.
    Jawab : **D.** Menambah jumlah data pelatihan atau menerapkan regularisasi.

22. Teknik *cross-validation* (misalnya K-Fold) digunakan untuk….
    A. Mempercepat proses pelatihan model.
    B. Meningkatkan ukuran dataset pelatihan.
    C. Mengurangi jumlah fitur.
    D. Mendapatkan estimasi kinerja model yang lebih akurat.
    Jawab : **D.** Mendapatkan estimasi kinerja model yang lebih akurat.

23. Pada regresi Iinear, metode gradient descent digunakan untuk…..
    A. Menginisialisasi bobot secara acak.
    B. Menentukan bobot yang meminimalkan fungsi loss.
    C. Menghitung metrik evaluasi MSE
    D. Menentukan arsitektur model terbaik
    Jawab : **B.** Menentukan bobot yang meminimalkan fungsi loss.

24. Pada dataset yang tidak seembang, metrik evaluasi yang kurang tepat digunakan adalah…
    A. F1-Score
    B. Recall
    C. Precision
    D. Akurasi
    Jawab: **D.** Akurasi
25. ‘Curse of dimensionality' dalam pembelajaran mesin mengacu pada fenomena….
    A. Bertambahnya jumlah fitur meningkatkan kinerja model tanpa batas.
    B. Kinerja model cenderung stabil seiring bertambah data
    C. Semakin tinggi dimensi fitur, data menjadi sangat jarang di ruang fitur.
    D. Algoritma menjadi lebih cepat pada data berfitur tinggi.
    Jawab : **C.** Semakin tinggi dimensi fitur, data menjadi sangat jarang di ruang fitur.

26. Berbeda dengan supervised learning, dalam **reinforcement learning**
    A. Setiap data pelatihan dilabell
    B. Model hanya memproses data statis
    C. Model dioptimalkan untuk tugas klasifikasi.
    D. Agen belajar melalui interaksi dengan lingkungan berdasarkan reward tanpa label eksplisit.

Jawab : **D.** Agen belajar melalui interaksi dengan lingkungan berdasarkan reward tanpa label eksplisit.

27. Normalisasi Min-Max, pada data, biasa menggunakan rumus: (x-min)/(max-min), menghasilkan nilai di antara….
    A. -1 sampai 1.
    B. 0 sampai 100.
    C. -∞ hingga +∞
    D. 0 sampai 1.
    Jawab : **D.** 0 sampai 1.

28. One-hat encoding pada data kategorikal dilakukan agar...
    A. Mengurangi jumlah data fitur.
    B. Data kategorikal dapat langsung digunakan Nam model.
    C. Setiap kategori diwakili sebagai vektor biner
    D. Data menjadi lebih mudah dipisahkan.
    Jawab : **C.** Setiap kategori diwakili sebagai vektor biner
29. Regresi logistik (logistic regression) umumnya digunakan untuk….
    A. Analisis komponen utama (PCA)
    B. Pengelompokan (clustering).
    C. Regresi kontinu.
    D. Klasifikasi biner dengan probabilitas
    Jawab : **D.** Klasifikasi biner dengan probabilitas
30. Salah satu kelebihan utama pohon keputusan adalah….
    A. Modelnya mudah diinterpretasikan dan divisualisasikan.
    B. Hanya cocok untuk hubungan linier sederhana.
    C. Selalu mencapai akurasi tinggi tanpa tuning.
    D. Membutuhkan data pelatihan yang sangat besar
    Jawab : **A.** Modelnya mudah diinterpretasikan dan divisualisasikan.

SOAL ESSAY!

1) Jelaskan pentingnya tahap pengumpulan data (data collection) dan pra pemrosesan data (data preprocessing) dalam pembangunan model machine learning
    - Berikan contoh proses pra pemrosesan data yang umum dilakukan dan tantangan yang mungkin ditemui pada tahap ini

    Jawab :
    1. Pengumpulan Data:
    Pengumpulan data adalah langkah awal untuk menyediakan informasi yang akan digunakan untuk melatih model machine learning.
    - Model ML hanya bisa belajar dari data yang tersedia.
    - Data yang tidak relevan atau berkualitas buruk akan membuat model tidak akurat, bahkan menyesatkan.
    - Tanpa data yang cukup, model tidak akan mampu menangkap pola atau hubungan yang dibutuhkan.

2. Prapemrosesan Data:

Prapemrosesan adalah proses membersihkan dan menyiapkan data sebelum dimasukkan ke algoritma ML.

- Tujuan:
- Menghilangkan noise, duplikasi, dan data tidak konsisten.
- Menangani data kosong (missing values).
- Mengubah data mentah ke format yang dapat dimengerti oleh mesin.

- Contoh Prapemrosesan Data yang Umum Dilakukan

Langkah	Penjelasan
Cleaning	Menghapus nilai kosong, outlier, atau data duplikat
Encoding	Mengubah data kategorikal menjadi numerik (misal: "Laki-laki" → 0, "Perempuan" → 1)
Normalization / Scaling	Mengubah rentang data agar sebanding, contohnya min-max scaling (0–1)
Handling Missing Values	Mengisi nilai kosong dengan rata-rata, median, atau menghapus baris/kolom
Feature Selection	Memilih fitur (kolom) yang paling relevan untuk model
Text Preprocessing	Untuk data teks: menghapus stopword, stemming, tokenizing
Balancing Data	Menyeimbangkan jumlah data antar kelas (jika dataset tidak seimbang)

- Tantangan dalam Pengumpulan & Prapemrosesan Data

Tantangan	Penjelasan
Data Tidak Lengkap	Banyak data hilang (missing values) yang bisa memengaruhi hasil
Noise dan Outlier	Data tidak wajar atau error sensor membuat model bias
Data Tidak Terstruktur	Data mentah seperti teks, gambar, atau log perlu diproses lebih dalam
Skala dan Format Berbeda	Data dari berbagai sumber tidak konsisten dalam satuan atau format
Overfitting karena Pembersihan Berlebihan	Terlalu banyak menghapus bisa menyebabkan hilangnya pola penting
Imbalanced Data	Kelas data yang tidak seimbang membuat model condong ke mayoritas
Waktu dan Biaya	Proses cleaning dan preprocessing memerlukan waktu dan sumber daya besar.

2) Jelaskan tahapan pemilihan model (model selection) dan pelatihan (training) & Pengajian (testing) dalam pipeline machine learning. Sertakan kriteria apa yang perlu dipertimbangkan Ketika memilih model serta strategi pembagian data pelatihan, valIdasi, dan pengujian.

Jawab:

Tahapan Pemilihan Model (Model Selection)

- Identifikasi Masalah: Tentukan jenis masalah yang ingin diselesaikan (klasifikasi, regresi, clustering, dll).

- Pemilihan Algoritma: Pilih algoritma yang sesuai berdasarkan karakteristik data dan tujuan analisis. Misalnya, untuk klasifikasi, bisa menggunakan Decision Trees, Random Forest, atau Neural Networks.
- Evaluasi Model: Gunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score untuk menilai performa model.
- Tuning Hyperparameter: Lakukan pencarian hyperparameter untuk mengoptimalkan kinerja model. Ini bisa dilakukan dengan teknik seperti Grid Search atau Random Search.

Pelatihan (Training)

- Penggunaan Data Training: Model dilatih menggunakan data training untuk belajar pola dari data tersebut.
- Validasi Model: Selama pelatihan, gunakan data validation untuk mengevaluasi model secara berkala dan mencegah overfitting.
- Penyimpanan Model Terbaik: Simpan model dengan performa terbaik berdasarkan hasil evaluasi pada data validation.
- Pengujian (Testing)
- Evaluasi Akhir: Setelah pelatihan selesai, gunakan data test untuk mengukur kinerja model secara objektif.
- Analisis Hasil: Tinjau hasil pengujian untuk memastikan model dapat generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya.

Kriteria Pemilihan Model

- Kompleksitas Model: Pertimbangkan kompleksitas model dan kemampuannya untuk menangkap pola dalam data.
- Waktu Pelatihan: Evaluasi waktu yang dibutuhkan untuk melatih model, terutama jika dataset besar.

- Kemampuan Generalisasi: Pastikan model tidak hanya baik pada data training tetapi juga dapat berfungsi dengan baik pada data baru.
- Interpretabilitas: Beberapa aplikasi memerlukan model yang mudah dipahami dan dijelaskan.
- Strategi Pembagian Data
- Pembagian Umum: Pembagian data yang umum adalah 80% untuk training, 10% untuk validation, dan 10% untuk testing. Ini memberikan keseimbangan antara pelatihan dan evaluasi.
- Stratified Sampling: Untuk dataset yang tidak seimbang, gunakan stratified sampling untuk memastikan proporsi kelas yang sama di setiap set.
- K-Fold Cross-Validation: Teknik ini membagi data menjadi K bagian, melatih model K kali, dan menggunakan setiap bagian sebagai data validasi satu kali. Ini membantu dalam mendapatkan estimasi yang lebih stabil dari kinerja model.
- Penyegaran Data: Secara berkala, kumpulkan data baru untuk memperbarui set validasi dan test agar tetap relevan dengan kondisi saat ini.

3) Jelaskan bagaimana proses evaluasi model dilakukan setelah pelatihan selesai. Sebutkan beberapa metrik evaluasi yang relevan untuk klasifikasi maupun regresi, dan jelaskan fungsi dari data validasi dan data pengujian.
    - Buatkan contoh code python untuk Machine Learning Model dalam Algoritma Regresi logistik (logistic regression)

Jawab :

Proses Evaluasi Model Setelah Pelatihan

- Penggunaan Data Pengujian: Setelah model dilatih, data pengujian digunakan untuk mengevaluasi kinerja model. Data ini tidak pernah digunakan selama proses pelatihan, sehingga memberikan gambaran yang lebih akurat tentang bagaimana model akan berfungsi di dunia nyata.
- Analisis Hasil: Hasil dari prediksi model dibandingkan dengan label sebenarnya dalam data pengujian untuk menghitung metrik evaluasi yang relevan.
- Identifikasi Kelemahan: Proses evaluasi juga membantu dalam mengidentifikasi kelemahan model, seperti kelas yang tidak terdeteksi dengan baik atau kesalahan prediksi yang sering terjadi.

Metrik Evaluasi untuk Klasifikasi

1. Akurasi: Mengukur proporsi prediksi yang benar dari total prediksi yang dibuat. Cocok untuk dataset seimbang.
2. Presisi: Mengukur seberapa banyak prediksi positif yang benar dibandingkan dengan seluruh prediksi positif. Penting dalam konteks di mana kesalahan positif lebih merugikan.
3. Recall (Sensitivitas): Mengukur kemampuan model untuk menemukan semua instance kelas positif. Berguna untuk memastikan bahwa model tidak melewatkan kasus penting.
4. F1-Score: Rata-rata harmonik antara presisi dan recall, memberikan keseimbangan antara keduanya, terutama pada dataset yang tidak seimbang.
5. Confusion Matrix: Menyajikan jumlah prediksi yang benar dan salah untuk setiap kelas, memberikan gambaran lebih mendalam tentang kinerja model.

Metrik Evaluasi untuk Regresi

1. Mean Absolute Error (MAE): Mengukur rata-rata kesalahan absolut antara prediksi dan nilai sebenarnya. Memberikan gambaran yang jelas tentang seberapa jauh prediksi dari nilai aktual.
2. Mean Squared Error (MSE): Mengukur rata-rata kuadrat dari kesalahan, memberikan penalti lebih besar untuk kesalahan yang lebih besar.
3. R-squared: Mengukur proporsi variabilitas dalam data yang dapat dijelaskan oleh model. Nilai mendekati 1 menunjukkan model yang baik.

Fungsi Data Validasi dan Data Pengujian

- Data Validasi: Digunakan selama proses pelatihan untuk mengoptimalkan hyperparameter dan mencegah overfitting. Data ini membantu dalam menilai kinerja model secara berkala dan melakukan penyesuaian yang diperlukan.

- Data Pengujian: Digunakan setelah pelatihan selesai untuk evaluasi akhir model. Data ini memberikan gambaran tentang seberapa baik model dapat generalisasi pada data baru yang tidak pernah dilihat sebelumnya, memastikan bahwa model siap untuk diterapkan di dunia nyata.

4) Jelaskan tahapan dalam siklus pengembangan machine leaming.
   - Apa yang dimaksud dengan deployment, serta tantangan apa saja yang dapat dihadapi saat memasukkan model ke dalam lingkungan produksi? Berikan contoh aplikasi deployment model yang umum.

   Jawab :

   Tahapan dalam Siklus Pengembangan Machine Learning

   - Pengumpulan Data

   Mengumpulkan data dari berbagai sumber yang relevan untuk masalah yang ingin diselesaikan.

   - Prapemrosesan Data

   Membersihkan data dari noise dan outlier.

   Mengubah format data agar sesuai untuk analisis.

   Melakukan rekayasa fitur untuk meningkatkan kualitas data.

   - Pelatihan Model

   Memilih algoritma yang sesuai untuk model.

   Melatih model menggunakan data pelatihan.

   Mengoptimalkan hyperparameter untuk meningkatkan performa model.

   - Evaluasi Model

   Menggunakan data pengujian untuk mengevaluasi kinerja model.

   Menggunakan metrik seperti akurasi, presisi, recall, dan F1-score.

   - Deployment

   Menempatkan model yang telah dilatih ke dalam lingkungan produksi agar dapat digunakan oleh aplikasi atau sistem lain.

   Deployment adalah proses mengimplementasikan model machine learning ke dalam lingkungan produksi sehingga model dapat memberikan prediksi atau analisis terhadap data baru secara real-time. Ini melibatkan integrasi model dengan aplikasi yang ada dan memastikan bahwa model dapat berfungsi dengan baik dalam kondisi nyata.

Tantangan dalam Deployment Model ke Lingkungan Produksi

1. Integrasi dengan Sistem yang Ada

Memastikan model dapat berfungsi dengan sistem perangkat lunak yang sudah ada tanpa mengganggu operasi yang berjalan.

2. Pemantauan Performa Model

Memantau kinerja model secara berkelanjutan untuk memastikan bahwa model tetap akurat dan relevan seiring waktu.

3. Penanganan Data Baru

Mengelola dan memproses data baru yang masuk untuk memastikan model dapat memberikan prediksi yang akurat.

4. Reproduksibilitas

Memastikan bahwa model dapat direproduksi dengan hasil yang konsisten ketika diberikan data yang sama.

5. Kesesuaian Lingkungan

Memastikan bahwa lingkungan tempat model di-deploy memiliki semua dependensi dan konfigurasi yang diperlukan.

Contoh Aplikasi Deployment Model yang Umum

- Sistem Rekomendasi

Digunakan oleh platform e-commerce untuk merekomendasikan produk kepada pengguna berdasarkan perilaku dan preferensi mereka.

- Deteksi Penipuan

Digunakan oleh lembaga keuangan untuk mendeteksi transaksi yang mencurigakan dan mencegah penipuan.

- Analisis Sentimen

Digunakan oleh perusahaan untuk menganalisis umpan balik pelanggan dan memahami sentimen terhadap produk atau layanan mereka.

- Chatbot

Menggunakan model NLP untuk memberikan respons otomatis kepada pengguna dalam aplikasi layanan pelanggan.