NAMA : Karenina Octarisfa

NIM : 4122008

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.

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.

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.

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

6. Contoh algoritma unsupervised learning yang paling terkenal adalah….
   A. Decision Tree
   B. Linear Regression
   C. K-Means Clustering
   D. Logic Regression

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

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 acak
   D. Memperbarui posisi centroid berdasarkan rata-rata anggota klaster.

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.

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.

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

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.

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

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)

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

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

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.

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

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.

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.

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.

23. Pada regresi linear, 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

24. Pada dataset yang tidak seembang, metrik evaluasi yang kurang tepat digunakan adalah…
    A. F1-Score
    B. Recall
    C. Precision
    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.

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.

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.

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.

29. Regresi logistik (logistic regression) umumnya digunakan untuk….
    A. Analisis komponen utama (PCA)
    B. Pengelompokan (clustering).
    C. Regresi kontinu.
    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

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

JAWABAN:

Pentingnya Tahap Pengumpulan Data dan Pra Pemrosesan Data

- **Pengumpulan Data (Data Collection)**
  - Tahap awal dalam pembangunan model machine learning.
  - Data yang dikumpulkan menjadi dasar untuk belajar dan prediksi.
  - Kualitas dan kuantitas data mempengaruhi performa model.
  - Data berkualitas rendah bisa menyebabkan overfitting atau underfitting.

- **Pra Pemrosesan Data (Data Preprocessing)**
  - Dilakukan setelah pengumpulan data dan sebelum pelatihan model.
  - Tujuan untuk membersihkan dan mempersiapkan data.
  - Data tidak bersih dapat mengganggu pelatihan dan menghasilkan model akurat.

- **Contoh Proses Pra Pemrosesan Data:**
  - Pembersihan Data: Menghapus atau memperbaiki data hilang, duplikat, atau tidak konsisten.
  - Normalisasi dan Standarisasi: Mengubah skala fitur dengan Min-Max Scaling atau Z-score Normalization.
  - Encoding Kategori: Mengubah data kategori menjadi format numerik dengan One-Hot Encoding atau Label Encoding.
  - Pengurangan Dimensi: Mengurangi jumlah fitur dengan PCA (Principal Component Analysis).

- **Tantangan dalam Pra Pemrosesan Data:**
  - Data yang Hilang: Menangani data hilang bisa menjadi tantangan.
  - Data Tidak Seimbang: Kelas yang tidak seimbang dapat menyebabkan bias.
  - Variabilitas Data: Data berubah-ubah membuat model sulit belajar pola konsisten.
  - Waktu dan Sumber Daya: Proses dapat memakan waktu dan sumber daya, terutama untuk dataset besar.

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

JAWABAN:

Pemilihan Model dan Pelatihan & Pengujian

- Pemilihan Model
  - Pemilihan algoritma machine learning sesuai untuk masalah yang dihadapi.
  - Pertimbangan dalam pemilihan model:
  - Tipe Data: Data terstruktur atau tidak terstruktur, numerik atau kategori.
  - Kompleksitas Model: Model kompleks lebih akurat tetapi berisiko overfitting.
  - Waktu Pelatihan: Beberapa model memerlukan waktu pelatihan lebih lama.
  - Interpretabilitas: Model perlu dijelaskan baik atau tidak tergantung aplikasi.

- Pelatihan dan Pengujian
  - Setelah pemilihan model, pelatihannya dilakukan menggunakan data pelatihan.
  - Model belajar mengenali pola dan membuat prediksi dari data pelatihan.
  - Data pengujian digunakan untuk mengevaluasi kinerja model setelah pelatihan.

- Strategi Pembagian Data:
  - Data Pelatihan: Sekitar 70-80% dari total data untuk melatih model.
  - Data Validasi: Sekitar 10-15% untuk mengoptimalkan hyperparameter dan mencegah overfitting.
  - Data Pengujian: Sekitar 10-15% untuk mengukur kinerja akhir model.

- Pertimbangan dalam Pembagian Data:
  - Keseimbangan Kelas: Setiap subset data harus memiliki distribusi kelas seimbang.
  - Randomisasi: Pembagian data harus acak untuk menghindari bias.
  - Cross-Validation: Digunakan untuk menguji model pada berbagai subset data dan meningkatkan keandalan evaluasi.

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)

JAWABAN:

Proses Evaluasi Model Setelah Pelatihan

- Setelah model dilatih, evaluasi dilakukan untuk mengukur kinerja model dalam prediksi.
- Evaluasi menggunakan data pengujian yang tidak digunakan saat pelatihan.

Metrik Evaluasi

Untuk Klasifikasi:

- Akurasi: Persentase prediksi yang benar dari total prediksi.
- Precision: Proporsi prediksi positif yang benar dari semua prediksi positif.
- Recall (Sensitivitas): Proporsi prediksi positif yang benar dari semua data positif yang sebenarnya.
- F1 Score: Rata-rata harmonis dari precision dan recall, baik untuk ketidakseimbangan kelas.
- ROC-AUC: Area di bawah kurva ROC, mengukur kemampuan model dalam membedakan kelas positif dan negatif.

Untuk Regresi:

- Mean Absolute Error (MAE): Rata-rata selisih absolut antara nilai yang diprediksi dan nilai aktual.
- Mean Squared Error (MSE): Rata-rata kuadrat selisih antara nilai yang diprediksi dan nilai aktual.
- Root Mean Squared Error (RMSE): Akar kuadrat dari MSE, ukuran kesalahan dalam satuan data asli.
- R-squared (R^2): Proporsi variabilitas dalam data yang dijelaskan oleh model.

Fungsi Data Validasi dan Data Pengujian

- Data Validasi: Digunakan untuk mengoptimalkan hyperparameter dan tuning model, membantu memilih model terbaik dan mencegah overfitting.
- Data Pengujian: Digunakan untuk mengevaluasi kinerja akhir model, menunjukkan kinerja model pada data baru yang tidak terlihat.

Berikut contoh code python nya:

https://github.com/kareninarinne/ML_UTS_NO3/blob/main/logisticregression-uts.ipynb

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.

JAWABAN:

Tahapan dalam Siklus Pengembangan Machine Learning

- **Pengumpulan Data**: Mengumpulkan data yang relevan dari berbagai sumber untuk digunakan dalam pelatihan model.
- **Pembersihan Data**: Memproses dan membersihkan data untuk menghilangkan noise, mengatasi missing values, dan memastikan kualitas data.
- **Eksplorasi Data**: Menganalisis data untuk memahami pola, distribusi, dan hubungan antar variabel yang dapat mempengaruhi model.
- **Pemodelan**: Membangun model machine learning menggunakan algoritma yang sesuai berdasarkan data yang telah dipersiapkan.
- **Evaluasi Model**: Mengukur performa model menggunakan metrik yang relevan untuk memastikan model dapat memberikan prediksi yang akurat.
- **Deployment**: Memasukkan model yang telah dilatih ke dalam lingkungan produksi agar dapat digunakan oleh aplikasi atau sistem lain.

Apa yang Dimaksud dengan Deployment?

Deployment adalah proses penerapan model machine learning ke dalam lingkungan produksi, sehingga model tersebut dapat diakses dan digunakan oleh aplikasi lain untuk menghasilkan prediksi berdasarkan data baru. Ini mencakup integrasi model dengan sistem yang ada dan memastikan bahwa model dapat berfungsi dengan baik dalam skenario dunia nyata.

Tantangan dalam Deployment Model

- **Integrasi dengan Sistem yang Ada**: Memastikan model dapat berfungsi dengan baik dalam infrastruktur yang sudah ada tanpa mengganggu operasi yang sedang berjalan.
- **Pemantauan Performa**: Memantau kinerja model secara terus-menerus untuk mendeteksi masalah atau penurunan performa seiring waktu.
- **Penanganan Data Baru**: Mengelola dan memproses data baru yang masuk untuk memastikan model tetap relevan dan akurat.
- **Skalabilitas**: Memastikan bahwa model dapat menangani volume data yang besar dan permintaan yang tinggi tanpa mengalami penurunan kinerja.

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**: Diterapkan dalam industri keuangan untuk mendeteksi transaksi yang mencurigakan dan mencegah penipuan.

- **Analisis Sentimen**: Digunakan oleh perusahaan untuk menganalisis umpan balik pelanggan dan memahami persepsi publik terhadap produk atau layanan mereka.