Tugas perbaikan
Bab 3

Nama : muhammad makhlufi makbullah

Kelas : TK 45 01

NIM : 1103210171

Bab 3 dari buku **"Introduction to Machine Learning with Python"** yang berjudul **"Unsupervised Learning and Preprocessing"** membahas jenis pembelajaran mesin yang tidak memerlukan label target (y) seperti dalam pembelajaran terawasi. Fokus utama adalah mengidentifikasi pola tersembunyi dalam data atau mengubah data menjadi representasi yang lebih berguna untuk analisis lebih lanjut.

Berikut adalah ringkasan konsep yang dibahas dalam bab ini:

---

1. Unsupervised Learning

- **Definisi**: Algoritma unsupervised learning bekerja tanpa label target. Tujuannya adalah untuk memahami struktur internal data.
- **Tujuan**:
    - **Clustering (Pengelompokan)**: Membagi data menjadi kelompok-kelompok berdasarkan kesamaan.
    - **Dimensionality Reduction (Reduksi Dimensi)**: Mengurangi jumlah fitur dalam data sambil tetap menjaga informasi penting.

---

2. Clustering

- **K-Means**:

    - Mengelompokkan data menjadi sejumlah cluster (k).
    - Data yang serupa dikelompokkan bersama dalam cluster yang sama.
    - Algoritma berulang untuk meminimalkan jarak dalam cluster.
    - Contoh penggunaan: Segmentasi pelanggan.
- **Agglomerative Clustering**:
    - Algoritma hirarkis yang membangun cluster dari bawah ke atas.
    - Berguna untuk data dengan struktur hierarkis.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**:
    - Mengelompokkan data berdasarkan kepadatan.
    - Cocok untuk dataset dengan bentuk cluster yang tidak biasa dan data dengan noise.

---

3. Dimensionality Reduction

- **PCA (Principal Component Analysis)**:
    - Teknik untuk mereduksi dimensi data sambil mempertahankan sebanyak mungkin varians.
    - Berguna untuk memvisualisasikan data berdimensi tinggi dalam 2D atau 3D.
- **t-SNE (t-Distributed Stochastic Neighbor Embedding)**:
    - Teknik untuk visualisasi data berdimensi tinggi.
    - Menjaga hubungan lokal antar data dalam dimensi rendah.

---

4. Preprocessing

- **Standardization (Standarisasi)**:
    - Menyesuaikan data sehingga setiap fitur memiliki rata-rata 0 dan standar deviasi 1.

    - Penting untuk algoritma seperti SVM dan PCA yang sensitif terhadap skala fitur.

- **Normalization (Normalisasi)**:
    - Memetakan data ke rentang tetap, seperti [0, 1].
    - Berguna untuk algoritma berbasis jarak seperti k-NN.
- **One-Hot Encoding**:
    - Mengubah data kategori menjadi format numerik biner.
    - Contoh: Kolom warna dengan nilai "merah", "biru" menjadi kolom biner terpisah.
- **Imputation**:
    - Mengisi nilai yang hilang dalam data.
    - Contoh: Mengisi nilai kosong dengan rata-rata, median, atau nilai yang paling sering muncul.
- **Feature Scaling**:
    - Menyesuaikan skala semua fitur dalam data, seringkali dilakukan dengan normalisasi atau standarisasi.

---

5. Visualisasi

- **Visualisasi dalam Clustering**:
    - Membantu memahami pola dalam data setelah clustering.
    - Teknik visualisasi: PCA, t-SNE.
- **Evaluasi Clustering**:
    - **Silhouette Score**: Mengukur kualitas clustering berdasarkan jarak intra-cluster dan inter-cluster.
    - **Elbow Method**: Menentukan jumlah cluster optimal dalam K-Means.

---

6. Implementasi

- Algoritma unsupervised learning biasanya digunakan dalam eksplorasi data awal untuk:
    - Mencari pola tersembunyi.
    - Mengurangi dimensi data untuk mempercepat model pembelajaran terawasi.
    - Membuat representasi fitur baru.

penjelasan singkat dari sub-bab Bab 3 **"Unsupervised Learning and Preprocessing"** berdasarkan poin-poin yang disebutkan:

---

1. Types of Unsupervised Learning

- **Clustering**: Mengelompokkan data ke dalam beberapa grup (cluster) berdasarkan kesamaan.
- **Dimensionality Reduction**: Mengurangi jumlah fitur (dimensi) data sambil mempertahankan informasi penting.
- **Feature Extraction**: Mengubah data ke dalam representasi yang lebih informatif untuk analisis lebih lanjut.

---

2. Challenges in Unsupervised Learning

- **Tidak adanya label target** membuat evaluasi hasil lebih sulit.
- Kesulitan dalam menentukan jumlah cluster atau dimensi yang optimal.
- Beberapa algoritma sensitif terhadap parameter awal atau noise.

---

3. Preprocessing and Scaling

- Preprocessing adalah langkah penting untuk membersihkan dan menyiapkan data.

- Scaling diperlukan untuk menyamakan skala fitur sehingga algoritma berbasis jarak, seperti k-NN dan SVM, dapat bekerja dengan baik.

---

4. Different Kinds of Preprocessing

- **Scaling**: Menstandarisasi atau menormalisasi data untuk menyamakan skala fitur.
- **Imputation**: Mengisi nilai yang hilang.
- **One-Hot Encoding**: Mengubah data kategori menjadi representasi biner.

---

5. Applying Data Transformations

- Transformasi data dapat membantu meningkatkan performa model dengan membuat data lebih seragam.
- Contoh: **Log transform** untuk menyelesaikan distribusi miring, atau **polynomial transform** untuk menambah non-linearitas.

---

6. Scaling Training and Test Data the Same Way

- **Penting** untuk menerapkan transformasi (seperti scaling) yang sama pada data pelatihan dan data uji untuk mencegah bias model.
- **StandardScaler** dan **MinMaxScaler** adalah alat umum untuk scaling.

---

7. The Effect of Preprocessing on Supervised Learning

- Preprocessing dapat secara signifikan meningkatkan akurasi dan efisiensi algoritma supervised learning.
- Contoh: PCA untuk reduksi dimensi dapat mempercepat pelatihan model.

---

8. Dimensionality Reduction, Feature Extraction, and Manifold Learning

- **Dimensionality Reduction**: Mengurangi jumlah fitur untuk menyederhanakan data.
- **Feature Extraction**: Mengidentifikasi kombinasi fitur yang lebih informatif.
- **Manifold Learning**: Teknik non-linear untuk mengungkapkan struktur tersembunyi dalam data berdimensi tinggi.

---

9. Principal Component Analysis (PCA)

- Teknik reduksi dimensi yang memproyeksikan data ke dalam beberapa komponen utama.
- Berguna untuk visualisasi dan mempercepat proses analisis.

---

10. Non-Negative Matrix Factorization (NMF)

- Teknik feature extraction untuk data yang hanya memiliki nilai positif.
- Berguna untuk aplikasi seperti analisis dokumen atau rekomendasi.

---

11. Manifold Learning with t-SNE

- t-SNE adalah teknik visualisasi yang sangat efektif untuk data berdimensi tinggi.
- Menjaga hubungan lokal antar data, menghasilkan plot yang intuitif.

---

12. Clustering

- Proses mengelompokkan data berdasarkan pola dan kesamaan.

---

13. k-Means Clustering

- Algoritma clustering populer yang membagi data menjadi **k** cluster.

- Menggunakan centroid untuk menentukan jarak.

---

14. Agglomerative Clustering

- Clustering hirarkis yang membangun cluster dari data individu (bawah ke atas).
- Dapat divisualisasikan menggunakan **dendrogram**.

---

15. DBSCAN

- Algoritma clustering berbasis kepadatan.
- Tidak memerlukan jumlah cluster (k) yang telah ditentukan.
- Efektif untuk dataset dengan noise atau bentuk cluster yang tidak biasa.

---

16. Comparing and Evaluating Clustering Algorithms

- **Silhouette Score**: Mengukur seberapa baik data dikelompokkan.
- **Adjusted Rand Index (ARI)**: Membandingkan hasil clustering dengan ground truth (jika ada).
- Tidak ada pendekatan universal; pilih algoritma berdasarkan kebutuhan dataset.

---

17. Summary of Clustering Methods

- **k-Means**: Cepat, cocok untuk cluster berbentuk bulat.
- **Agglomerative**: Berguna untuk struktur hirarkis.
- **DBSCAN**: Cocok untuk cluster berbentuk kompleks dan data dengan noise.

---

18. Summary and Outlook

- **Unsupervised learning** sangat penting dalam eksplorasi data dan preprocessing.
- Reduksi dimensi, clustering, dan preprocessing membentuk fondasi untuk pipeline machine learning yang kuat.
- Meskipun sulit dievaluasi, algoritma unsupervised membuka wawasan baru yang tidak dapat ditemukan dengan pendekatan supervised.