NAMA : FAJRIN SARIFUDIN

NPM : 121055520121029

PRODI : INFORMATIKA

SEMESTER/KELAS : VI / INFO-6

MAPEL : DATA MINING

1. Apa definisi dari klasifikasi dalam konteks data mining?

Classification merupakan **metode yang digunakan apabila atributnya berupa numerik atau nominal, namun labelnya harus berupa nominal**. Pada metode data mining ini, dilakukan pengelompokan atau pengklasifikasian berdasarkan hubungan antara variabel kriteria dengan variabel target

2. jelaskan konsep dasar dari decision tree dan bagaimana cara kerja dalam klasifikasi nya

a. Konsep dari pohon keputusan adalah mengubah data menjadi decision tree dan aturan-aturan keputusan. Manfaat utama dari penggunaan decision tree adalah kemampuannya untuk mem- break down proses pengambilan keputusan yang kompleks menjadi lebih simple, sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan.

b. cara kerja decision tree ;

Mulai dari simpul akar, kita misalkan sebagai S, yang berisi dataset lengkap.

* Ambil atribut terbaik dalam dataset menggunakan Attribute Selection Measure (ASM). ASM yang bisa digunakan di antaranya Information Gain dan Gini Index
* Pisahkan himpunan S menjadi himpunan bagian yang berisi kemungkinan nilai untuk atribut terbaik.
* Buat simpul
* [decision](https://www.trivusi.web.id/2022/06/algoritma-decision-tree.html) tree yang berisi atribut terbaik.
* Buat simpul [decision](https://www.trivusi.web.id/2022/06/algoritma-decision-tree.html) tree baru secara rekursif menggunakan himpunan bagian dari kumpulan data yang dibuat pada langkah 3. Lanjutkan proses ini sampai tahap terakhir di mana kita tidak dapat mengklasifikasikan simpul lebih lanjut. Simpul ini yang menjadi simpul akhir atau disebut sebagai simpul daun (leaf node).

3. apa perbedaan antara supervised learning dan unsupervised learning

**Supervised Learning** adalah metode di mana algoritma dilatih menggunakan **data** berlabel, sementara **Unsupervised Learning** berfokus pada analisis **data** tanpa adanya label atau bimbingan manusia

4. Sebutkan dan jelaskan dua teknik pemangkasan (pruning) yang umum digunakan dalam decision tree.

1. Teknik Pre-Pruning

**Deskripsi:** Teknik ini melibatkan penghentian proses pembentukan pohon lebih awal, sebelum pohon mencapai ukuran maksimalnya. Pre-pruning menetapkan batasan tertentu yang harus dipenuhi agar pembentukan cabang baru dilanjutkan. Batasan ini dapat berupa batas kedalaman pohon, jumlah minimum sampel yang diperlukan untuk membuat sebuah node, atau jumlah maksimum node pada pohon.

* **Keuntungan:** Proses training lebih cepat dan lebih sederhana karena pohon tidak tumbuh terlalu besar.
* **Kekurangan:** Bisa jadi tidak optimal karena potensi informasi dari fitur-fitur yang lebih dalam tidak sepenuhnya dieksplorasi.

2. Pruning Pasca-Pemangkasan (Post-Pruning)

**Deskripsi:** Teknik ini dilakukan setelah pohon decision tree dibangun sepenuhnya. Pohon yang telah terbentuk kemudian dipangkas dengan menghapus cabang-cabang yang tidak memberikan kontribusi signifikan terhadap akurasi model. Beberapa metode yang digunakan dalam post-pruning termasuk Cost Complexity Pruning (atau Minimal Cost-Complexity Pruning), Reduced Error Pruning, dan lainnya.

* **Keuntungan:** Biasanya menghasilkan pohon yang lebih sederhana dan lebih mudah diinterpretasikan tanpa kehilangan banyak akurasi.
* **Kekurangan:** Memerlukan waktu lebih lama karena pohon harus dibangun terlebih dahulu dan kemudian dipangkas.

5. Bagaimana cara menangani data yang tidak seimbang dalam proses klasifikasi?

Beberapa teknik yang dapat digunakan untuk menangani data yang tidak seimbang meliputi:

1. Teknik Penyeimbangan Data

* + **Oversampling:** Meningkatkan jumlah sampel dalam kelas minoritas dengan menduplikasi sampel yang ada atau menggunakan teknik seperti SMOTE (Synthetic Minority Over-sampling Technique) untuk menciptakan sampel sintetis baru.
  + **SMOTE:** Teknik ini menghasilkan sampel sintetis dengan interpolasi antara sampel minoritas yang ada.
  + **Undersampling:** Mengurangi jumlah sampel dalam kelas mayoritas sehingga seimbang dengan kelas minoritas. Ini bisa dilakukan dengan memilih subset dari kelas mayoritas.
  + **Random Undersampling:** Mengurangi ukuran kelas mayoritas dengan memilih sampel secara acak.
  + **Cluster-based Undersampling:** Menggunakan teknik clustering untuk memilih sampel representatif dari kelas mayoritas.

2. Penyesuaian Algoritma

* + **Penggunaan Algoritma yang Dirancang untuk Ketidakseimbangan:** Beberapa algoritma machine learning, seperti Random Forests, dapat dimodifikasi atau digunakan dengan parameter khusus untuk menangani ketidakseimbangan.
  + **Pemberian Bobot pada Kelas:** Dalam beberapa algoritma, seperti Support Vector Machines atau Logistic Regression, Anda dapat memberikan bobot lebih pada kelas minoritas untuk memberikan dampak yang lebih besar pada model.
  + **Cost-sensitive Learning:** Menambahkan penalti yang lebih besar untuk kesalahan klasifikasi pada kelas minoritas.

3. Teknik Evaluasi Model

* + **Penggunaan Metrik Evaluasi yang Sesuai:** Menggunakan metrik yang lebih informatif daripada akurasi biasa, seperti:
  + **Precision, Recall, dan F1-Score:** Memberikan informasi yang lebih baik tentang performa model pada kelas minoritas.
  + **ROC Curve dan AUC (Area Under the Curve):** Menunjukkan kemampuan model untuk membedakan antara kelas.
  + **Confusion Matrix:** Memberikan gambaran detail tentang performa model pada tiap kelas.
  + **Precision-Recall Curve:** Berguna untuk dataset yang sangat tidak seimbang.

6. Apa itu overfitting dan bagaimana cara mencegahnya saat membangun model klasifikasi?

Overfiting terjadi ketika sebuah algoritme terlalu cocok atau bahkan persis dengan data pelatihannya, sehingga menghasilkan model yang tidak dapat membuat prediksi atau kesimpulan yang akurat dari data apa pun selain data

cara mencagah overfitting ;

1. Teknik Penyeimbangan Data

* + **Oversampling:** Meningkatkan jumlah sampel dalam kelas minoritas dengan menduplikasi sampel yang ada atau menggunakan teknik seperti SMOTE (Synthetic Minority Over-sampling Technique) untuk menciptakan sampel sintetis baru.
  + **SMOTE:** Teknik ini menghasilkan sampel sintetis dengan interpolasi antara sampel minoritas yang ada.
  + **Undersampling:** Mengurangi jumlah sampel dalam kelas mayoritas sehingga seimbang dengan kelas minoritas. Ini bisa dilakukan dengan memilih subset dari kelas mayoritas.
  + **Random Undersampling:** Mengurangi ukuran kelas mayoritas dengan memilih sampel secara acak.
  + **Cluster-based Undersampling:** Menggunakan teknik clustering untuk memilih sampel representatif dari kelas mayoritas.

2. Penyesuaian Algoritma

* + **Penggunaan Algoritma yang Dirancang untuk Ketidakseimbangan:** Beberapa algoritma machine learning, seperti Random Forests, dapat dimodifikasi atau digunakan dengan parameter khusus untuk menangani ketidakseimbangan.
  + **Pemberian Bobot pada Kelas:** Dalam beberapa algoritma, seperti Support Vector Machines atau Logistic Regression, Anda dapat memberikan bobot lebih pada kelas minoritas untuk memberikan dampak yang lebih besar pada model.
  + **Cost-sensitive Learning:** Menambahkan penalti yang lebih besar untuk kesalahan klasifikasi pada kelas minoritas.

3. Teknik Evaluasi Model

* + **Penggunaan Metrik Evaluasi yang Sesuai:** Menggunakan metrik yang lebih informatif daripada akurasi biasa, seperti:
  + **Precision, Recall, dan F1-Score:** Memberikan informasi yang lebih baik tentang performa model pada kelas minoritas.
  + **ROC Curve dan AUC (Area Under the Curve):** Menunjukkan kemampuan model untuk membedakan antara kelas.
  + **Confusion Matrix:** Memberikan gambaran detail tentang performa model pada tiap kelas.
  + **Precision-Recall Curve:** Berguna untuk dataset yang sangat tidak seimbang.

4. Strategi Lain

* + **Ensemble Methods:** Menggunakan teknik ensemble seperti bagging dan boosting yang seringkali lebih tahan terhadap data yang tidak seimbang. Misalnya, **Balanced Random Forest** atau **Adaptive Boosting (AdaBoost)**.
  + **Anomaly Detection Techniques:** Kadang-kadang, kelas minoritas dapat dianggap sebagai anomali atau outlier, dan teknik deteksi anomali dapat digunakan untuk mengidentifikasi sampel dari kelas minoritas.

7. Jelaskan bagaimana k-fold cross-validation digunakan dalam evaluasi model klasifikasi.

Validasi silang K-fold dalam validasi silang pembelajaran mesin adalah teknik yang ampuh untuk mengevaluasi model prediktif dalam ilmu data. Ini melibatkan pemisahan kumpulan data menjadi k subset atau lipatan, di mana setiap lipatan digunakan sebagai kumpulan validasi secara bergantian sedangkan k-1 lipatan sisanya digunakan untuk pelatihan. Proses ini diulang sebanyak k kali, dan metrik kinerja seperti akurasi, presisi, dan perolehan dihitung untuk setiap lipatan. Dengan merata-ratakan metrik ini, kami memperoleh perkiraan performa generalisasi model. Metode ini penting untuk penilaian model, pemilihan, dan penyesuaian hyperparameter, sehingga menawarkan ukuran efektivitas model yang andal. Dibandingkan dengan validasi silang satu kali keluar, yang menggunakan k sama dengan jumlah sampel, validasi silang K-fold efisien secara komputasi dan banyak digunakan dalam praktik.

8. **Apa saja metrik evaluasi yang dapat digunakan untuk mengukur kinerja model klasifikasi?**

**1. Akurasi (Accuracy)**

#### ****2. Confusion Matrix****

**3. Recall (Sensitivity atau True Positive Rate)**

**4. F1-Score**

**6. AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**

**7. AUC-PR (Area Under the Precision-Recall Curve)**

**8. Log Loss (Logarithmic Loss)**

**9. Specificity (True Negative Rate)**

**9.** **Bagaimana cara kerja metode ensemble seperti Random Forest dalam meningkatkan akurasi klasifikasi?**

**Metode ensemble, seperti Random Forest, meningkatkan akurasi klasifikasi dengan menggabungkan prediksi dari beberapa model untuk menghasilkan prediksi akhir yang lebih akurat dan stabil. Random Forest adalah salah satu metode ensemble yang populer dan efektif. Diantaranya **Pembentukan banyak decision tree, Pembelajaran fitur acak, kombinasi prediksi****

**10.** **Apa itu naive Bayes classifier dan dalam situasi apa algoritma ini paling efektif digunakan?**

**Naive Bayes classifier adalah kelompok algoritma klasifikasi yang didasarkan pada Teorema Bayes dengan asumsi kemandirian (independensi) yang kuat antara fitur-fitur. Meskipun asumsi ini jarang benar dalam praktik, algoritma ini sering bekerja dengan baik dan efisien dalam berbagai situasi.**

**algoritma naive bayes efektif digunakan dalam hal dibawah ini**

**A. Text Classification dan Spam Filtering:**

**B. Sistem Deteksi**

**C. Diagnosis Medis**

**D. Rekomendasi**