Jawaban Soal 1: Konsep Feature Selection

a. Apa yang dimaksud dengan Feature Selection dalam konteks Data Wrangling dan Machine Learning?

Feature Selection (Seleksi Fitur) adalah proses dalam *data wrangling* dan *machine learning* untuk memilih secara otomatis atau manual subset fitur (variabel atau kolom) yang paling relevan dan paling berkontribusi terhadap variabel target atau hasil prediksi. Tujuannya adalah untuk menyederhanakan model dengan mengurangi jumlah fitur input, sehingga dapat meningkatkan performa, mengurangi waktu training, dan membuat model lebih mudah diinterpretasikan tanpa mengorbankan akurasi secara signifikan.

b. Mengapa proses Feature Selection penting dalam analisis dan pemodelan data?

Proses ini sangat penting karena beberapa alasan fundamental:

1. **Menyederhanakan Model:** Model dengan lebih sedikit fitur lebih mudah untuk dipahami, dijelaskan, dan diinterpretasikan.
2. **Mengurangi Waktu Training:** Mengurangi jumlah fitur berarti mengurangi beban komputasi. Model akan berlatih lebih cepat, yang sangat krusial pada dataset yang sangat besar.
3. **Menghindari "Kutukan Dimensi" (Curse of Dimensionality):** Terlalu banyak fitur dapat menyebabkan ruang data menjadi sangat renggang, sehingga data yang ada tidak cukup untuk mendukung model yang andal. Hal ini dapat menurunkan performa model.
4. **Mengurangi Overfitting:** Dengan menghilangkan fitur yang tidak relevan atau *redundant* (berlebihan), kita dapat mengurangi *noise* dalam data. Ini membantu model untuk fokus pada sinyal yang sebenarnya dan melakukan generalisasi yang lebih baik pada data baru yang belum pernah dilihat sebelumnya, sehingga mengurangi risiko *overfitting*.
5. **Meningkatkan Akurasi Model:** Meskipun terkesan kontradiktif, menghilangkan fitur yang menyesatkan (misleading) dapat meningkatkan akurasi model secara keseluruhan.

Jawaban Soal 2: Perbandingan Metode Feature Selection

Berikut adalah perbandingan antara *Univariate Feature Selection* dan *Recursive Feature Elimination (RFE)* dalam format tabel perbandingan.

Aspek	**Univariate Feature Selection**	**Recursive Feature Elimination (RFE)**
Prinsip Kerja	Mengevaluasi setiap fitur secara independen terhadap variabel target menggunakan uji statistik	Bekerja secara iteratif. Metode ini melatih sebuah model, mengevaluasi pentingnya setiap fitur, lalu

	(misalnya, Chi-square, ANOVA, korelasi). Fitur dengan skor statistik tertinggi akan dipilih.	menghapus fitur yang paling tidak penting. Proses ini diulang hingga jumlah fitur yang diinginkan tercapai.
Ketergantungan Model	Tidak bergantung pada model (*model-agnostic*). Metode ini murni menggunakan statistik untuk menilai hubungan antar variabel.	Sangat bergantung pada model. Kinerja RFE ditentukan oleh model (*estimator*) yang digunakan untuk mengevaluasi fitur (misalnya, Random Forest, SVM, Regresi Linear).
Kelebihan	- Sangat cepat dan efisien secara komputasi. - Mudah dipahami dan diimplementasikan. - Baik sebagai langkah awal untuk penyaringan data.	- Mampu menangkap interaksi antar fitur, karena mengevaluasi fitur secara bersamaan dalam konteks sebuah model. - Umumnya menghasilkan subset fitur dengan performa yang lebih baik.
Kekurangan	- Mengabaikan interaksi antar fitur. Sebuah fitur mungkin tidak terlihat penting secara individu, tetapi menjadi sangat penting jika dikombinasikan dengan fitur lain. - Cenderung memilih fitur yang redundan jika mereka memiliki skor individu yang tinggi.	- Membutuhkan waktu komputasi yang lebih lama, karena harus melatih model berulang kali. - Hasilnya bisa tidak stabil jika model yang digunakan sensitif terhadap perubahan data.
Kapan Digunakan	- Sebagai langkah penyaringan awal pada dataset dengan ribuan atau jutaan fitur. - Ketika kecepatan komputasi menjadi prioritas utama. - Untuk mendapatkan pemahaman dasar yang cepat tentang hubungan setiap fitur dengan target.	- Ketika performa model adalah prioritas utama dan biaya komputasi bukan masalah besar. - Pada dataset di mana interaksi antar fitur diduga kuat memengaruhi hasil (misalnya, dalam diagnosis medis, keuangan). - Untuk menemukan subset fitur optimal yang bekerja paling baik untuk model tertentu.

Jawaban Soal 3: Studi Kasus & Analisis Hasil

Hasil Seleksi Fitur

1. Hasil dari Univariate Feature Selection (menggunakan `chi2`): Berdasarkan output kode yang telah saya buat pada file **Titanio_Yudista_24120500031_Univariate.ipynb**, 5 fitur dengan skor Chi-square tertinggi adalah:

- `area_worst` (Skor: 112598.43)
- `area_mean` (Skor: 53991.65)
- `area_se` (Skor: 8758.50)
- `perimeter_worst` (Skor: 3665.03)
- `perimeter_mean` (Skor: 2011.10)

2. Hasil dari Recursive Feature Elimination (RFE) dengan Random Forest: Berdasarkan output kode yang telah saya buat pada file **Titanio_Yudista_24120500031_RFE.ipynb**, 5 fitur yang terpilih dengan ranking 1 adalah:

- `concave points_mean`
- `radius_worst`
- `perimeter_worst`
- `area_worst`
- `concave points_worst`

Perbandingan Hasil

Apakah kelima fitur yang terpilih sama?

Tidak, kelima fitur yang terpilih oleh kedua metode tersebut tidak sama. Hanya ada **dua fitur** yang sama-sama terpilih oleh kedua metode, yaitu:

- `perimeter_worst`
- `area_worst`

Perbedaan ini menunjukkan bahwa kedua metode memiliki pendekatan yang fundamental berbeda dalam menilai "pentingnya" sebuah fitur.

Metode mana yang lebih cocok dan mengapa?

Menurut saya, metode **Recursive Feature Elimination (RFE) lebih cocok dan unggul** untuk digunakan pada dataset diagnosis kanker payudara ini. Berikut adalah alasan mendalam dari sudut pandang statistik dan interpretasi model:

1. **Dari Sudut Pandang Statistik (Penanganan Redundansi dan Interaksi):**
    - **Kelemahan Univariate:** Metode Univariate dengan `chi2` menilai setiap fitur secara terpisah. Hasilnya menunjukkan bahwa fitur-fitur yang terpilih (`area_worst`, `area_mean`, `area_se`, `perimeter_worst`,

`perimeter_mean`) semuanya sangat berkaitan dengan **ukuran (size)** tumor. Secara statistik, fitur-fitur ini sangat berkorelasi satu sama lain (mengalami **multikolinearitas**). Misalnya, `area_mean` pasti memiliki korelasi yang sangat tinggi dengan `perimeter_mean`. Metode ini hanya memilih fitur dengan skor statistik individu tertinggi tanpa mempertimbangkan bahwa informasi yang diberikan mungkin tumpang tindih atau redundan.

- **Kekuatan RFE:** RFE, dengan menggunakan Random Forest sebagai *estimator*, bekerja secara berbeda. Ia membangun model prediktif dan secara iteratif membuang fitur yang paling tidak berkontribusi pada akurasi model secara **keseluruhan**. Proses ini secara inheren mempertimbangkan **interaksi antar fitur**. RFE memilih `perimeter_worst` dan `area_worst` (metrik ukuran), namun ia juga memilih `concave points_mean` dan `concave points_worst` (metrik bentuk/tekstur) serta `radius_worst` (metrik ukuran lain). Ini menunjukkan bahwa setelah model mempertimbangkan fitur ukuran yang kuat, fitur yang memberikan informasi **baru dan berbeda** (seperti bentuk cekungan pada tumor) menjadi lebih berharga untuk meningkatkan performa model daripada sekadar menambahkan fitur ukuran lain yang redundan (`area_se` atau `perimeter_mean`).

2. **Dari Sudut Pandang Interpretasi Model dan Tujuan Akhir:**
   - Tujuan akhir dari analisis ini adalah untuk membangun model yang dapat mendiagnosis kanker secara akurat. Oleh karena itu, fitur yang paling "baik" adalah fitur yang secara **kolektif** memberikan daya prediksi tertinggi.

   - Univariate Selection hanya memberi tahu kita fitur mana yang secara individu memiliki hubungan statistik terkuat dengan diagnosis, tetapi tidak menjamin bahwa kombinasi 5 fitur tersebut adalah yang terbaik untuk sebuah model.

   - RFE secara langsung menjawab pertanyaan: "Jika saya hanya boleh menggunakan 5 fitur untuk membangun model Random Forest, manakah 5 fitur yang akan memberikan hasil terbaik?" Hasil dari RFE (`concave points`, `radius_worst`, dll.) adalah set fitur yang lebih beragam dan kuat secara prediktif. Fitur `concave points` (jumlah dan tingkat keparahan cekungan pada kontur tumor) secara klinis merupakan indikator penting keganasan, dan RFE berhasil mengidentifikasinya sebagai fitur kunci di samping metrik ukuran.

Kesimpulan Akhir:

Meskipun Univariate Selection lebih cepat secara komputasi, untuk kasus kritis seperti diagnosis medis, di mana akurasi dan kemampuan model untuk menangkap nuansa data sangat vital, **RFE adalah pendekatan yang secara konseptual lebih kuat dan lebih cocok.** Ia mampu mengatasi masalah multikolinearitas dan memilih satu set fitur yang tidak hanya relevan secara individu, tetapi juga kuat secara prediktif ketika digunakan bersama-sama dalam sebuah model.