**1. Jelaskan perbedaan antara Populasi, Observasi, dan Sampel !**

Populasi : keseluruhan objek pengamatan atau pun penelitian yang akan digunakan sebagai bahan penelitian dan mempunyai karateristik tertentu, lengkap dan juga jelas. Populasi bisa juga dikatakan sebagai totalitas dari semua objek yang hendak diteliti atau sebagai unit objek yang menggambarkan data yang dimiliki. Contoh Populasi : semua email yang dikirim tahun lalu oleh karyawan di perusahaan besar, BigCorp

Observasi : kegiatan penelitian dengan memahami knowladge dari sebuah fenomena berdasar knowledge dan gagasan sebelumnya, digunakan untuk mendapat informasi yang dibutuhkan sebagai penelitian selanjutnya. Dengan kata lain merupakan proses komputasi atau analisis yang dilakukan dari data yang dimiliki. Contoh : daftar hal: nama pengirim, daftar penerima, tanggal pengiriman, teks email, jumlah karakter dalam email, jumlah kalimat dalam email, dan lamanya waktu hingga balasan pertama

Sampel : potongan kecil dari poppulasi yang digunakan untuk objek pengamatan karena dianggap sebagai perwakilan populasi tersebut, hasilnya akan digunakan sebagai penentu kesimpulan terhadap populasi tersebut. Contoh : buat daftar semua karyawan dan pilih 1/10 orang secara acak, lalu ambil semua email yang pernah mereka kirim

Perbedaan :

Populasi : Terdiri atas objek atau subjek yang mempunyai kualitas dan ciri khusus dan telah ditetapkan oleh peneliti untuk dapat dianalisis dan didapatkan sebuah kesimpulan. Populasi adalah bagian besar dari sampel dan sama sama merupakan objek penelitan.

Sampel : Sebagian kecil dari populasi penelitian tersebut. Sampel adalah yang menjadi objek penelitian secara langsung.

**2. Sebutkan tahapan-tahapan EDA !**

Exploratory Data Analysis  ( EDA) merupakan metode eksplorasi data dengan menggunakan teknik aritmatika sederhana dan teknik grafis dalam meringkas data pengamatan yang memungkinkan analyst memahami isi data yang digunakan, mulai dari distribusi, frekuensi, korelasi dan lainnya.

Langkah pada EDA :

a) Buat pertanyaan tentang data. Membuat rumusan masalah terhadap data yang akan diteliti.

b) Lakukan pencarian untuk mendapat jawaban dengan memvisualisasikan, mengubah, dan memodelkan data.

c) Gunakan apa yang dapat pelajari untuk memperbaiki pertanyaan terkait data dan / atau menghasilkan pertanyaan baru.

Contoh EDA pada klasifikasi nilai ujian dalam bentuk histogram

**3. Jelaskan perbedaan antara Data Mining, Data Science, dan Data Engineering !**

Data Mining adalah proses pengambilan pola pada data yang akan di proses dengan hasil berupa berupa informasi yang berguna. Informasi diperoleh dari sebuah proses yang akan mengidentifikasi dan mengekstraksi informasi dari suatu database besar. Proses tersebut berupa artificial intelligence, teknik statistik, ilmu matematika, machine learning, dan lain sebagainya.

Fungsi data mining :

* Mendapat pola dari data yang dianalisis dengan variabel untuk memprediksi variabel lain
* Mendapat ciri khusus dari data
* Mendapat model dan fungsi untuk penggambaran kelas atau konsep dari data
* Mendapat hubungan antara nilai atribut dari beberapa data

Data Science adalah bidang yang mempelajari data terutama data kuantitatif dengan melakukan ekstrak data agar dapat disaring untuk mendapatkan data asli dan menghasilkan produk data yang jelas.

Tujuan :

* Memberikan saran untuk client atas hasil analisis data
* Menunjukkan prioritas yang harus dilakukan
* Membantu dalam pencapaian target dengan memproses data

Data Engineering adalah data yang digunakan untuk mengembangkan dan membangun design arsitektur manajemen data, serta  memelihara/memonitor infrastruktur data di perusahaan.

Perbedaan Data Mining dengan Data Science

|  |  |
| --- | --- |
| Data Mining | Data Science |
| Berbicara mengenai trend tentang dataset | Berbicara mengenai bidang studi apapun tentang data (analitik data, statistik, data mining, dll) |
| Fokus ke bisnis proses dari data yang ada | Fokus ke bidang ilmu mengenai data |
| Membuat data lebih berguna | Membuat data center di dalam suatu organisasi |
| Menghasilkan pola | Menghasilkan berbagai macam tentang data dan sangat bervariatif |
| Tujuannya menentukan trend | Tujuannya analisis dan prediksi |
| Orang yang mempunyai data dan data statistika | Seseorang yang membutuhkan pengetahuan machine learning, programming dan data, dan grafik visualisasi |
| Bagian dari data science | Multi disiplin |
| Type datanya banyak yang terstruktur | Type data lebih tidak terstruktur |

**4. Proses Data Science**

**a. Jelaskan tiap langkah-langkah inti dalam proses Data Science !**

1. Real Word

Di dalam kehidupan sehari-hari banyak orang yang sibuk dengan berbagai kegiatan. Dari berbagai kegiatan tersebut , dapat diambil sebuah data untuk langkah selanjutnya.

2. Pengumpulan Data Mentah → Pemilahan Data (Membersihkan data)

Dalam pengolahan data, data yang akan diolah pastinya merupakan data mentah seperti log, catatan, atau email. Kemudian data akan dipilah dan dianalisis sehingga mendapatkan format data yang sesuai dalam. Proses ini biasanya menggunakan Python, atau R.

3. Analisis Data Eksploratorium (EDA)

Setelah mendapat dataset, langkah selanjutnya adalah melakukan semacam EDA. Ketika proses melakukan EDA, biasanya data akan terdeteksi karena duplikat, nilai yang hilang, pencilan yang absurd, dan data yang tidak benar-benar dicatat atau tidak dicatat dengan benar sehingga harus kembali melakukan pengumpulan lebih banyak data.

4. Model

Selanjutnya, proses perancangan model dengan menggunakan beberapa algoritma seperti k-tetangga terdekat (k-NN), regresi linier, Naif Bayes, atau yang lainnya. Model yang dipilih tergantung pada jenis masalah yang akan coba diselesaikan, tentu saja, yang bisa berupa masalah klasifikasi, masalah prediksi, atau masalah deskripsi dasar.

5. Komunikasi

Setelah pemrosesan data sebelumnya,maka dapat melakukan penafsiran, pemvisualisasian, pelaporan, atau pengkomunikasian hasil. Komunikasi dapat berupa pelaporan hasil kepada atasan atau orang yang membutuhkan data.

6. Build Data Product

Tujuannya untuk membangun atau membuat prototipe "produk data"; mis., klasifikasi spam, atau algoritma peringkat pencarian, atau sistem rekomendasi. Disini adalah kunci yang membuat ilmu data istimewa dan berbeda dari statistik yaitu bahwa produk data ini akan dimasukkan kembali ke real word, kemudian pengguna berinteraksi dengan produk itu, dan yang menghasilkan lebih banyak data, yang menciptakan umpan balik.

**b. Sebutkan peranan Data Scientist dalam langkah-langkah inti pada proses Data Science !**

Pada kehidupan sehari-hari , data scientist menggunakan data untuk mengajukan pertanyaan terkait data apa yang perlu direkam atau dikumpulkan dengan kata lain untuk merumuskan hipotesis. Contohnya Manusia melakukan kegiatan pada bidang biologi, keuangan, internet, kedokteran, sosiologi, olimpiade.

Pada proses pengumpulan data yang masih mentah atau belum diproses, data scientist dapat mengumpukan data berupa Email, log, catatan medis, Survei, Darah diambil, catatan Olimpiade, halaman web NYT (artikel). Ketika data diproses , data scientist menggambarkan data sebagai Saluran pipa, Mengikis web, Membersihkan, Mengayun, Bergabung, Berperang. Ketika proses pemilahan data atau pembersihan data , data scientist mengharapkan dari data sebelumnya digunakan oleh data scientist agar data Bersih, Pencilan, Nilai tidak ada, Debugging, Tabel.

**c. Berikan contoh penyelesaian masalah di kehidupan sehari-hari dengan memperhatikan langkah-langkah inti dalam proses Data Science**

**5. Sebutkan dan jelaskan langkah-langkah inti dalam proses Data Mining !**

Langkah 1

Mengintegrasikan data : Untuk mengumpulkan dan menggabungkan data dari semua sumber yang berbeda.

Memilih data : Tidak semua data yang dikumpulkan berguna, jadi pada langkah berikutnya, hanya memilih data yang berguna untuk penggalian data.

Pembersihan data : Data yang dipilih mungkin mengandung kesalahan, nilai yang hilang, dan inkonsistensi yang perlu dibersihkan. Dibutuhkan teknik dan alat yang berbeda dalam proses ini.

Langkah 2

Transformasi data : Perataan, agregasi, normalisasi, dll merupakan beberapa teknik yang digunakan untuk mengubah data menjadi format yang dapat dimengerti.

Langkah 3

Penambangan data : Sekarang adalah waktu untuk menerapkan teknik seperti pengelompokan dan analisis asosiasi untuk proses penambangan data dan untuk menemukan pola yang menarik.

Langkah 4

Evaluasi pola : Menghapus pola yang berlebihan untuk menghindari kebingungan dan menganalisis pola yang tersisa adalah langkah penting dalam proses ini.

Langkah 5  
Menggunakan pengetahuan yang ditemukan : Langkah terakhir dalam proses ini adalah memanfaatkan pengetahuan yang ditemukan dalam proses ini untuk membuat keputusan.

**6. Temukan permasalahan di kehidupan sehari-hari yang dapat diselesaikan dengan implementasi Top-K dan Skyline (dalam lingkup Data Science, bukan Data Engineering).**

**a. Jelaskan permasalahan dan data yang anda gunakan !**

**b. Selesaikan dengan menggunakan Top-K dan Skyline, jelaskan proses atau langkah-langkah penyelesaiannya secara detail !**

**c. Berikan analisis hasilnya**