Nama : David Vincent Gurning
NIM : 11S21001
Mata Kuliah : Pembelajaran Mesin Learning

1. Mengapa persiapan data menjadi hal yang kritikal pada machine learning?
   ➔ Persiapan data menjadi hal yang kritikal pada pembelajaran mesin dikaranakan kualitas data secara langsung memengaruhi kinerja dan hasil dari model yang akan dibangun. Melalui Data Science Tutorial yang telah saya baca bahwa ada dua peran Dimana yang pertama ialah center manager Dimana hal yang pertama yang dilakukan oleh center manager ialah mengumpulkan data, dengan beberapa pertanyaan : • Where is it? • What will you use to analyze it? • How accurate it is? • How complete is it? • Is it too big to easily read? Dan yang saya tangkap hal ini membuktikan bahwasanya data merupakan hal yang sangat krusial dikarenakan 5 pertanyaan ini dapat dijawab harus melalui data, jadi jika data kita salah maka yang akan keluar juga nantinya merupakan data yang salah , namun jika data yang kita gunakan benar, maka data yang akan dikeluarkan oleh mesin juga nantinya merupakan pengetahuan yang benar. Jadi urutan dari seorang center manager itu bekerja ialah: get data, clean data, EDA, Visualization, Interpretation, dan action. Dapat disimpulkan bahwa data sangat krusial karena :
   1. Kualitas Model:
   Model machine learning hanya sebagus data yang digunakan untuk melatihnya. Jika data yang digunakan bermasalah, seperti adanya outlier, noise, atau data yang hilang, model cenderung menghasilkan prediksi yang tidak akurat.

   2. Pembersihan Data (Data Cleaning):
   Data dapat mengandung kesalahan, nilai yang hilang, atau outlier yang dapat memengaruhi kualitas model. Persiapan data melibatkan pembersihan data untuk mengatasi masalah ini agar model dapat belajar dari data yang bersih dan akurat.

   3. Penanganan Missing Values:
   Banyak model machine learning tidak dapat menangani nilai yang hilang dalam dataset. Oleh karena itu, persiapan data melibatkan strategi untuk menangani nilai yang hilang, seperti pengisian nilai rata-rata atau interpolasi.

   4. Pemilihan Fitur (Feature Selection):
   Tidak semua fitur dalam dataset mungkin relevan atau memberikan kontribusi signifikan terhadap prediksi. Proses persiapan data melibatkan pemilihan fitur yang paling relevan dan memiliki dampak positif pada kinerja model.

   5. Normalisasi dan Skalabilitas:
   Beberapa algoritma machine learning sangat sensitif terhadap skala dan distribusi data. Persiapan data mencakup normalisasi atau standarisasi fitur agar model dapat belajar secara efisien dan menghasilkan hasil yang stabil.

   Dengan meningkatnya kesadaran tentang keamanan dan privasi data, persiapan data juga mencakup langkah-langkah untuk melindungi data sensitif dan memastikan kepatuhan terhadap regulasi privasi.

Dengan melakukan persiapan data dengan baik, kita dapat meningkatkan keakuratan, stabilitas, dan interpretabilitas model machine learning, serta mengurangi risiko terjadinya bias atau overfitting. Oleh karena itu, persiapan data merupakan langkah penting dalam memastikan keberhasilan implementasi model machine learning.

2. Berikan tahapan persiapan data dan contoh data sebelum dan sesudah tahapan

➔ Tahapan persiapan data (data preparation) melibatkan sejumlah langkah untuk memastikan bahwa data yang digunakan dalam proses machine learning atau analisis lainnya bersih, relevan, dan siap untuk digunakan. Berikut adalah beberapa tahapan umum dalam persiapan data beserta contoh data sebelum dan sesudah setiap tahapan:

- Pemahaman Data (Data Understanding):

Sebelum: Data mentah dari berbagai sumber, mungkin dalam format yang tidak terstruktur atau tercampur.
Sesudah: Identifikasi dan dokumentasikan sumber data, pemahaman awal tentang struktur data, dan definisi variabel.

- Pembersihan Data (Data Cleaning):

Sebelum: Data yang mengandung nilai yang hilang, outlier, atau kesalahan entri.
Sesudah: Mengisi atau menghapus nilai yang hilang, menangani outlier, dan mengoreksi kesalahan entri.

- Pemilihan Fitur (Feature Selection):

Sebelum: Dataset dengan banyak fitur, termasuk beberapa yang mungkin tidak relevan atau memiliki korelasi tinggi.
Sesudah: Identifikasi fitur yang paling relevan dan memiliki dampak signifikan terhadap target. Hapus fitur yang tidak diperlukan.

- Transformasi Data:

Sebelum: Data dalam skala atau distribusi yang tidak merata.
Sesudah: Normalisasi atau standarisasi data, transformasi logaritma, atau penggunaan teknik lain yang diperlukan untuk membuat data sesuai dengan persyaratan model.

- Penanganan Data Tidak Seimbang:

Sebelum: Jumlah observasi antar kelas target tidak seimbang.
Sesudah: Terapkan teknik oversampling atau undersampling untuk menyeimbangkan kelas target.

- Penanganan Missing Values:

Sebelum: Data dengan nilai yang hilang atau tidak lengkap.
Sesudah: Isi nilai yang hilang dengan nilai yang sesuai, menggunakan metode seperti imputasi dengan nilai rata-rata atau median.

- Pemisahan Data (Data Splitting):

Sebelum: Dataset tunggal.
Sesudah: Pisahkan data menjadi set pelatihan (training set), set validasi, dan set pengujian untuk evaluasi model.

- Keamanan dan Privasi:

Sebelum: Data mungkin mengandung informasi sensitif tanpa langkah-langkah keamanan.
Sesudah: Enkripsi data sensitif, anonimisasi atau penyamaran data, serta implementasi langkah-langkah keamanan untuk melindungi privasi.

Contoh ini memberikan gambaran umum tentang bagaimana data dapat berubah selama proses persiapan. Namun, setiap proyek dapat memiliki tahapan tambahan atau berbeda tergantung pada tujuan dan karakteristik data yang dimiliki.