Nama : Dena Cahya Setia Putri
Kelas : JumaTec

EDA - DATA PREPROCESSING

Exploratory Data Analysis adalah bagian awal sebelum seorang data science dan data analis membuat pemodelan yang lebih komplex memberikan insight melalui visualisasi dan statistik dasar. EDA diperkenalkan oleh John Suci thun 1961 dan terus berkembang seiring berkembangnya ilmu dan teknologi. Yang dilakukan di sini adalah data preparation untuk menangani mentransformsi data lalu menyesuaikan tipe data apabila ada yang kurang tepat, menangani *noise*, *missing value* dan *outliers*.

Pada EDA ini akan mendapatkan hipotesis yang akan kita verifikasi melalui pemodelan. EDA memang powerfull namun memilliki kelemahan yang bisa ditutupi oleh pemodelan. EDA bisa memberikan rekomendasi terkait data yang diperlukan atau variabel lain yang dibutuhkan, dapat melihat performa model seperti apa.

Dataset

- Koleksi entitas/objek data dan atributnya
- Atribut adalah sifat atau karkteristik dari objek
- Contoh pada objek manusia: umur, berat badan, tinggi badan, jenis kelamin, dsb.
- Setiap atribut memiliki beberapa kemungkinan "state", sebagai contoh: pria/wanita.
- koleksi atribut mendefinisikan suatu objek.

Data preprocessing adalah teknik yang digunakan untuk mengubah data mentah menjadi format yang berguna dan efisien yang juga merupakan kunci utama dalam mendapatkan model Data Science yang valid & reliable. Hal ini diperlukan untuk:

1. Data di dunia nyata biasanya tidak sebersih/indah data di buku akademik.
   Noise: Misal gaji bernilai negatif
   Ouliers: Misal seseorang dengan penghasilan >500 juta/bulan.
   Duplikasi: Banyak di media sosial
   Encodings, dsb: Banyak di Big Data, karena masalah bagaimana data disimpan/join.
2. Tidak lengkap: hanya agregat, kurang variabel penting
3. Analisa pada data yang tidak di preprocess biasanya menghasilkan insight yang tidak/kurang tepat.

1. Data Gathering:
   Data warehouse, database, web crawling/scrapping/streaming.
   Identifikasi, ekstraksi, dan integrasi data

2. Data Cleaning

   Proses memperbaiki atau menghapus data yang salah, rusak, salah format, duplikat, atau tidak lengkap dalam kumpulan data.

   Data cleaning: redudant sample, missing value, outliers.

   Outliers: nilai yang tdk normal/ekstrem. Untuk menangani: remove (hapus: untuk datanya yang banyak), imputasi (jika datanya sedikit/ tdk terlalu banyk)

   (menginput nilai perkiraan (mean, median, modus).
3. Transformasi data (misal encoding var kategorik)

   Proses mengubah format, struktur, atau nilai pada data.

   - agrerasi: meringkas data yang ada

   - Normalization: min max normalization

   Keuntungan data transformation:

   - data diubah menjadi lebih terorganisir, sehingga lebih mudah digunakan oleh manusia dan komputer

   - Data yang diformat dan divalidasi dengan benar dapat meningkatkan kualitas datanya

   - Transformasi data memfasilitasi kompabilitas anyara app, sistem, tipe data

   - Mempercepat proses komputasi

   Jenis encoding (https://towardsdatascience.com/categorical-feature-encoding-547707acf4e5)

   1. One hot encoding (key: menambah kolom sebanyak jumlah kategori): merubah menjadi numerik (mengubah warna). (-) Tidak cocok jika jumlah kategorinya banyak.

   2. Label encoding: untuk tipe data nominal, kategorikal yg tidak memiliki jenjang, (+)tidak menambah kolom tp akan langsung otomatis, lebih cepat, kolom tidak banyak. (-) bisa kehilangan informasi/uniqe value

   3. Ordinal encoding: tipe data yang memiliki jenjang. Tentukan dulu urutannya dari yang terbagus sampai cuting yang terburuk (hrs sesuai potongan yang terbaik urutannya).

   4. Target encoding: harus cari dulu rata-rata. Misal ada kolom warna trs ada harga, lalu dikelompokan, dan ketauan harga warna merah itu brp dan dihitung rata2nya, itu yg akan menjadi pengganti dari labelnya. (-) kemungkinan menimbulkan overfiting.
4. Normalisasi/standarisasi
5. Data reduction: menghasilkan representasi dari kumpulan data menjadi ukuran volume yang lebih kecil tetapi tidak mengurangi informasi/fitur penting pada data sehingga akan menghasilkan analisis yang sama.

   variable selection (domain knowledge/automatic)

   - Feature Engineering
   - Variable reduction

Noisy Data

Dapat terjadi karena:

- Kesalahan instrumen pengukuran: Misal di alat IoT pada saat cuaca buruk/baterai yang lemah.
- Kesalahan input/entry
- Transmisi yang tidak sempurna
- inkonsistensi penamaan

Tipe missing value:

1. Missing completely at random (MCAR)

   Data hilang secara acak, dan tidak berkaitan dengan variabel tertentu
2. Missing at random (MAR)

   Data di suatu variabel hilang hanya berkaitan dengan variabel respon/pengamatan. Sebagai contoh, orang yang memiliki rasa was-was tinggi (x) cenderung tidak melaporkan pendapatan (y) mereka, walaupun missing value bergantung pada berapa nilai x, tapi seberapa besar nilai y yang missing tersebut masih tetap acak
3. Missing not at random (MNAR)

   Data di suatu variabel y berkaitan dengan variabel itu sendiri, tidak terdistribusi secara acak. Sebagai contoh, orang yang pendapatannya rendah cenderung tidak melaporkan pendapatannya. Tipe missing value ini yang relatif paling sulit untuk di handle

   Pada MCAR dan MAR, kita boleh menghilangkan data dengan missing value ataupun mengimputasinya. Namun pada kasus MNAR, menghilangkan data dengan missing value akan menghasilkan bias pada data. mengimputasinya pun tidak selalu memberikan hasil yang baik.

   Source: https://youtu.be/OzxmCTPpbN8