AI4Jobs | Kampus Merdeka Batch 3

Nama : Siti Nur Fadhilah

Kelas : Jumatec

Resume Materi

Data Processing

A. EDA (Exploration Data Analysis)

Exploration Data Analysis (EDA) bagaikan jiwa bagi semua proses analisis data. Kemampuan untuk melakukan EDA dengan baik adalah syarat dasar utama bagi seluruh profesi yang terkait dengan pengolahan data, baik itu business intelligence, data analyst, data scientist, dan sebagainya. EDA juga menjadi tahapan awal dari kebanyakan proses analisis data dan menjadi suatu tahapan yang amat menentukan seberapa baik Analisa data selanjutnya akan di akan dihasilkan. Diperkenalkan oleh John Turkey 1961: "Procedures for analyzing data, techniques for interpreting the result of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and result of (mathematical) statistic which apply to analyzing data". Komponen EDA meliputi preprocessing, perhitungan berbagai nilai statistic dasar (e.g. ukuran pusat dan penyebaran data), visualisasi, penyusunan hipotesis(dugaan awal), pemerikasaan asumsi, hingga story-telling dan reporting. Di dalamnya juga termasuk proses penanganan missing values, outlier, reduksi dimensi, pengelompokkan, transformasi dan distribusi data.
Beberapa tujuan dari EDA yaitu :

1. menyarankan hipotesis
2. menilai asumsi
3. pemilihan teknik statistik yang tepat
4. pengumpulan data lebih lanjut

EDA yang termasuk Exploratory yaitu :

1. Preprocessing
2. Visualisasi
3. Basic Statistic
   a. Rata – rata (mean),
   b. Nilai tengah (median),
   c. Modus, dll.
4. Clustering
   Clustering bukan inferensia, namun termasuk exploratory data. Saat melakukan clustering, tidak pernah dipisahkan antara training data dan testing data. Ini dikarenakan clustering merupakan bagian dari exploratory data analysis. EDA memiliki tujuan yaitu mengexplore data yang dimiliki. EDA tidak melakukan generalisasi, namun akan mendapatkan gambaran utuh dari data yang dimiliki lalu menyiapkannya.

B. Data (Set)

Data (set) adalah koleksi beberapa entitas/objek data dan attribute nya. Attribut adalah sifat atau karakteristik dari sebuah objek. Di dunia nyata, data yang ada tidak hanya terdiri dari satu tabel saja. Biasanya para data scientist akan melakukan join dari beberapa tabel dan menggabungkannya menjadi satu tabel. Karena data engineer dan data administrator menyimpan data dalam keadaan normalisasi. Hal ini dimaksudkan agar lebih efisien dan lebih cepat dalam me-manage data dan model yang akan dibuat hanya bisa menerima data dari satu tabel.

Contoh pada objek dari manusia yaitu :

1. umur
2. berat badan
3. tinggi badan
4. jenis kelamin

Setiap attribute memiliki beberapa kemungkinan “state”, sebagai contoh: pria/Wanita. Koleksi attribute mendefinisikan suatu objek.

C. Preprocessing

Preprocessing adalah kunci utama dalam menadapatkan model data yang valid dan reliable. Preprocessing yang berbeda, akan menghasilkan kesimpilan atau insight yang berbeda pula. Jika melakukan proses data preprocessing tidak baik, maka akan menghasilkan model dengan kesimpulan atau insight yang salah atau model dengan performa yang buruk. Jika tidak melakukan preprocessing yang baik, bahkan sekedar visualisasi saja, model tidak akan bekerja. Setiap model yang dibuat, memerlukan proses preprocessing yang berbeda pula.

Proses preprocessing ini bisa dilakukan dalam pemrograman atau dalam query – query. Bahkan, jika dilakukan dalam query – query, proses ini akan menjadi lebih cepat. Hal ini dikarenakan struktur data dalam database, terdapat beberapa index yang dapat dimanfaatkan sehingga proses menjadi lebih cepat dibandingkan lewat pemrograman. Namun, hal tersebut membutuhkan kompetensi data yang lebih dalam.

Beberapa proses dasar yaitu :

1. Seleksi variable dan “join”
2. Data cleaning: duplikasi, noise, dan outliers
3. Transformasi data jika diperlukan, tergantung juga asumsi dari modelnya seperti apa
4. Dimensional reduction

Sebelum melakukan preprocessing, terdapat beberapa hal yang harus dilakukan, yaitu business understanding dan data understanding.

Mengapa perlu adanya preprocessing?

1. Data di dunia ini biasanya tidak sebersih atau seindah data di buku akademik
   a. Noise = Gaji bernilai negative
   b. Outlier = Seseorang dengan penghasilan > 500 juta / bulan
   c. Duplikasi = Banyak di sosial media
   d. Encodings = Banyak di big data, karena masalah bagaimana data disimpan / join
2. Tidak lengkap (hanya agregrat, kurang variable penting, dsb).
3. Analisa pada data yang tidak di preprocessing biasanya menghasilkan insight yang tidak atau kurang tepat.

Langkah- langkah Preprocessing yaitu:

1. Data gathering
   a. Data warehouse, database, web crawling
   b. Identifikasi extraksi, dan integrasi data
2. Data cleaning
3. Transformasi data (misal encoding var kategorik)
4. Normalisasi/standarisasi
5. Data reduction:
   a. Variable selection (domain knowledge/automatic)
   b. Feature engineering
   c. Variable reduction

Sumber Materi :
https://www.youtube.com/watch?v=OzxmCTPpbN8&ab_channel=taudataAnalytics

source code :
https://colab.research.google.com/github/taudata-indonesia/eLearning/blob/master/eda-01_Codes.ipynb#scrollTo=AjB2Qe4ahzk-