**Rangkuman Petimbangan dalam Pengolahan Data**

**Pengantar Petimbangan dalam Pengolahan Data**

Mungkin sebagian dari Anda pernah mendengar tentang beberapa dataset akademik yang umum digunakan sebagai bahan latihan para calon praktisi data masa depan, seperti *iris dataset*, *titanic dataset*, dll. Seluruh dataset tersebut memang dikhususkan sebagai bahan latihan dan *benchmark* dalam penelitian sehingga bentuknya sangat sederhana. Selain itu, dataset seperti ini umumnya sudah dalam keadaan bersih dan siap untuk digunakan dalam proses analisis data ataupun proses lanjutan lainnya.

Data yang kita jumpai di industri sangatlah kompleks dan umumnya masih sangat kotor sehingga perlu dilakukan tahap pembersihan data (data cleaning) terlebih dahulu sebelum digunakan. Selain itu, pada proses analisis data, kita sering kali harus menggunakan lebih dari satu dataset dengan ukuran data yang sangat besar untuk menjawab sebuah masalah bisnis.

Tantangan lain yang akan Anda jumpai di industri salah satunya adalah kompleksitas dari kebutuhan bisnis. Kebutuhan bisnis di industri umumnya dipengaruhi banyak hal, mulai dari keadaan pasar, keberadaan kompetitor, hingga keadaan makro ekonomi dan tentunya masih banyak lagi faktor yang lain. Semua faktor tersebut membuat kebutuhan bisnis di industri menjadi sangat dinamis.

Selain tantangan terkait kebutuhan bisnis dan data, seorang praktisi data juga dihadapkan dengan banyak tantangan lain seperti privasi dan etika dalam pengolahan data serta keamanan dari data itu sendiri. Saat ini, ada banyak regulasi yang mengatur kedua hal tersebut dan sebagai calon praktisi data yang Andal, kita harus mematuhi semua regulasi tersebut.

**Bias dalam Data**

*By definition*, bias merupakan sebuah prasangka dan kecenderungan yang mendukung atau menentang suatu hal, individu, atau kelompok lain dengan cara yang kurang adil. Ia bisa terjadi secara sengaja maupun tidak sengaja. Bias merupakan salah satu masalah yang sering dijumpai di kehidupan sehari-hari termasuk dalam sebuah data.

Bias dalam data umumnya didefinisikan sebagai eror atau galat yang secara sistematis mengarahkan hasil menuju arah tertentu. Bias dalam data umumnya muncul disebabkan banyak faktor salah satunya ialah kesalahan dalam proses sampling.

Secara umum bias dalam data dapat dikategorikan dalam empat tipe, yaitu *sampling bias*, *observer bias*,*interpretation bias*, dan *confirmation bias*.

* **Sampling bias**  
  Sampling bias terjadi ketika sampel tidak mewakili populasi secara keseluruhan. Ini merupakan salah satu tipe bias yang paling sering terjadi. Bias seperti ini dapat dihindari menggunakan teknik *random sampling* ketika mengumpulkan sebuah data. Teknik ini memungkinkan semua bagian dari populasi mendapat kesempatan yang sama untuk terpilih.
* **Observer bias**  
  Tipe bias selanjutnya ialah observer bias atau sering juga disebut sebagai *experimenter bias*. Bias tipe ini muncul karena adanya kecenderungan yang berbeda dari setiap individu ketika melakukan observasi. Hal ini bisa dihindari dengan melakukan observasi berulang. Sebagai contoh, untuk menghindari observasi bias ketika melakukan pengukuran, kita bisa melakukan proses pengukuran secara berulang dan mengambil nilai rata-ratanya sebagai hasil akhir.
* **Interpretation bias**  
  Interpretation bias merupakan bias yang terjadi karena adanya kecenderungan kita dalam menginterpretasikan situasi ambigu hanya ke dalam dua keadaan (hitam dan putih). Hal ini berpotensi melahirkan bias karena setiap individu memiliki kecenderungan yang berbeda dalam menginterpretasi situasi yang ambigu. Hal ini dapat dihindari dengan cara melibatkan orang lain untuk mengonfirmasi hasil interpretasi yang kita buat.
* **Confirmation bias**  
  Tipe bias terakhir adalah confirmation bias. Bias ini muncul karena adanya kecenderungan kita dalam mencari atau menafsirkan informasi untuk mengonfirmasi keyakinan yang sudah ada sebelumnya. Tipe bias ini akan sangat berbahaya karena dapat menjebak kita dalam *echo chamber*. Oleh karena itu, sebagai calon praktisi data yang andal, kita harus selalu menempatkan diri senetral mungkin dalam setiap proses analisis data.

**Data Credibility**

Kredibilitas data merupakan bagian dari *data quality assessment* yang digunakan sebagai sebuah tolok ukur untuk menilai tingkat kepercayaan terhadap suatu data.

Tingkat kepercayaan dari sebuah data akan sangat berdampak terhadap kualitas dan keakuratan hasil analisis yang akan dibuat. Tingkat kepercayaan ini tentunya dipengaruhi oleh banyak faktor, seperti sumber data, kelengkapan data, keakuratan data, dll.

Berikut merupakan beberapa poin sebagai patokan untuk memastikan data yang digunakan berkualitas dan dapat dipercaya.

* **Reliabilitas data**  
  Reliabilitas merupakan salah satu faktor penting untuk menilai kualitas dan kredibilitas dari sebuah data. Data disebut reliabel jika memuat informasi yang akurat, lengkap, dan tidak bias.
* **Orisinalitas data**  
  Selain reliabilitas, orisinalitas dari sebuah data juga merupakan komponen kunci dalam menilai kualitas dan kredibilitas dari sebuah data. Hal ini erat kaitannya dengan *ownership* (kepemilikan) dan sumber data.
* **Komprehensif**  
  Poin selanjutnya yang tidak kalah penting ialah komprehensif. Poin ini sangat penting untuk memastikan data yang digunakan memiliki semua informasi yang dibutuhkan untuk menjawab pertanyaan dan tantangan bisnis.
* **Aktualitas data**  
  *Last*,*but not least* adalah aktualitas data. Seperti yang telah kita bahas sebelumnya, data di industri umumnya bersifat dinamis atau terus berubah. Jika kita menggunakan data lama, informasi di dalamnya mungkin akan *outdated* sehingga berdampak terhadap hasil analisis yang dibuat. Oleh karena itu, sangat penting bagi kita untuk memastikan proses analisis dilakukan menggunakan data terbaru.

**Privasi dan Etika dalam Pengolahan Data**

Pada beberapa dekade terakhir, kedua topik tersebut telah menjadi salah satu pembahasan hangat dalam dunia data. Hal ini merupakan imbas dari maraknya pemanfaatan data di industri sehingga mulai banyak yang menyadari potensi bahaya dalam praktik tersebut seperti pelanggaran privasi.

Untuk mencegah bahaya tersebut, mulai muncul banyak studi terkait *data ethics*. Ia merupakan kumpulan acuan standar dalam menilai benar dan salah sebuah proses pengolahan data. Berikut merupakan beberapa aspek penting dalam data ethics.

* **Ownership**  
  Aspek ini menyatakan bahwa kepemilikan data berapa pada setiap individu yang menghasilkan data tersebut. Sebagai pemilik data, individu memiliki kontrol penuh akan seluruh proses dan penggunaan data tersebut.
* **Transparency**  
  Seluruh tahapan dan algoritma dalam proses pengolahan data harus transparan dan dipahami oleh setiap individu yang menghasilkan data tersebut.
* **Consent**  
  Setiap individu harus mengetahui secara eksplisit terkait penggunaan dan alasan suatu data dikumpulkan. Hal ini dilakukan sebelum individu tersebut menyetujui pengumpulan data tersebut.
* **Currency**  
  Jika data digunakan untuk transaksi keuangan, setiap individu yang menghasilkan data harus mengetahui dengan jelas detail terkait transaksi tersebut.
* **Privacy & Openness**  
  Organisasi yang mengumpulkan data bertanggung jawab dalam menjaga setiap informasi yang bersifat pribadi pada setiap proses pengolahan data. Selain itu, organisasi tersebut harus menyediakan akses terhadap penggunaan dan pemanfaatan data (non private) secara bebas.

Berikut merupakan beberapa regulasi yang dibuat untuk menjaga data privasi milik pengguna dan memastikan penerapan data ethics.

* General Data Protection Regulation (GDPR)
* Peraturan Menteri (Permen) No 20 Tahun 2016 tentang Perlindungan Data Pribadi (PDP)

**Keamanan Data**

Data security dapat didefinisikan sebagai upaya dalam menjaga dan melindungi informasi digital (data) dari berbagai pihak yang tidak bertanggung jawab dalam seluruh proses pengolahannya. Hal ini mencakup keamanan fisik hardware (storage device) serta keamanan dari software atau aplikasi yang bersangkutan. Selain itu, ia juga mencakup seluruh kebijakan dan prosedur dalam sebuah organisasi.

Berikut merupakan dua teknik atau metode A.

* **Encryption**  
  Encryption merupakan teknik yang memanfaatkan algoritma tertentu untuk mengubah data menjadi bentuk *unusable* bagi individu atau aplikasi yang tidak memiliki akses terhadap algoritma tersebut. Pada penerapnnya, kita bisa menganggap algoritma tersebut sebagai *key* (kunci) untuk membalikkan (*reverse*) proses encryption.
* **Tokenization**  
  Metode ini memungkinkan kita untuk mengubah elemen tertentu dalam sebuah data menjadi sebuah data random yang berperan sebagai token. Pada metode ini, data asli beserta *token mapping* akan disimpan dalam lokasi yang berbeda. Hal ini memungkinkan kita membatasi akses terhadap data asli dan memastikan keamanannya dengan menyimpan data tersebut di dalam lokasi yang berbeda.