

# PERBANDINGAN TEKNIK PENGOLAHAN DATA DARI SUDUT PANDANG STATISTIKA DAN *MACHINE LEARNING*

Ariska Kurnia Rachmawati  
Matematika Fakultas Sains dan Teknologi UIN Walisongo Semarang  
[ariskakurnia@walisongo.ac.id](mailto:ariskakurnia@walisongo.ac.id)

## Abstract

Kemajuan teknologi berkembang begitu pesat, bisa dibayangkan saat ini dunia sedang mengalami era dimana transformasi digital berlangsung, mulai dari teknologi berukuran besar hingga teknologi berukuran mikro yang tidak pernah dibayangkan manusia sebelumnya. Seluruh informasi atau data mengenai apapun yang ingin diketahui dapat diakses melalui internet baik informasi mengenai industri, politik, hiburan, apapun itu dapat ditemukan melalui internet. Statistika mempunyai peranan penting sebagai katalis perkembangan ilmu-ilmu lain, baik ilmu alam ataupun ilmu social. Statistika dipakai oleh disiplin ilmu lain sebagai alat untuk mengambil kesimpulan, menguji hipotesis/teori, memahami fenomena, menganalisis eksperimen, menentukan keputusan, dsb. *Machine learning* merupakan topik yang hangat di ranah teknologi beberapa tahun terakhir, karena dengan sistem ini diyakini akan mengubah serta mempermudah cara hidup dan bekerja manusia. *Machine learning* dan statistika adalah dua disiplin ilmu yang saling berkaitan dalam mempelajari ilmu tentang data dimana keduanya mempunyai dasar teori dan model yang sama. Perbedaan keduanya terletak pada fokus yang berbeda. Statistika lebih fokus ke arah pengambilan kesimpulan, sedangkan *machine learning* fokus ke prediksi data baru dengan tingkat keakuratan yang tinggi. Dari persamaan dan perbedaan tersebut, tidak salah kalau statistika dan *machine learning* disebut sebagai dua wajah berbeda dari satu kesatuan disiplin ilmu.

**Kata Kunci** : *statistika, machine learning, pengolahan data.*

## 1. PENDAHULUAN

Hampir semua orang yang bergelut di dunia ilmu pengetahuan sedikit banyak pasti pernah berinteraksi dengan statistika (ilmu statistik). Tidak dapat dipungkiri bahwa statistika mempunyai peranan penting sebagai katalis perkembangan ilmu-ilmu lain, baik ilmu alam (seperti astronomi dan biologi) ataupun ilmu sosial (seperti ekonomi, demografi, sosiologi, dsb.). Statistika dipakai oleh disiplin ilmu lain sebagai alat untuk mengambil kesimpulan, menguji hipotesis/teori, memahami fenomena, menganalisis eksperimen, menentukan keputusan, dsb.

*Machine learning* saat ini menjadi cabang ilmu pengetahuan yang populer dibicarakan di media. Didapati sebagai salah satu cabang dari ilmu kecerdasan buatan (*artificial intelligence*), hampir semua orang pernah berinteraksi, memakai ataupun mendengar sistem komputer yang dibangun memakai teknik *machine learning*. Mulai dari melihat tag otomatis foto di Facebook, menggunakan rekomendasi pencarian di Google, meng-klik rekomendasi produk sejenis di *online shopping*, dan menikmati servis *email* tanpa spam.

Pondasi dasar dari statistika dan *machine learning* adalah ilmu teori peluang. Semua teknik-teknik dalam statistika dan *machine learning* dibangun di atas teori peluang, yang merupakan bahasa matematika untuk mengukur derajat ketidakpastian. Ilmu-ilmu dasar yang juga penting bagi statistika dan *machine learning* diantaranya adalah aljabar linear, kalkulus dan teknik optimisasi. Ilmu dasar tersebut menjadi pondasi dalam Teknik dalam pengumpulan, pengolahan data serta analisis data serta mengetahui tantangan akan pentingnya *machine learning* di masa yang akan datang.

## 2. LANDASAN TEORI

### 2.1 Probabilitas

Di dunia ini, ada banyak hal yang tidak pasti (*uncertain*). Sesungguhnya *machine learning* berurusan dengan ketidakpastian (*uncertainty*). Dengan begitu *machine learning* memiliki

kaitan yang sangat erat dengan statistika. Probabilitas menyediakan *framework* untuk kuantifikasi dan manipulasi ketidakpastian.

Secara definisi, probabilitas merupakan kemungkinan – kemungkinan yang terjadi pada suatu peristiwa dengan memperkirakan kejadian apa yang akan terjadi maka dari itu probabilitas hanya memiliki rentang nilai antara 0 sampai 1. Nilai 0 menunjukkan bahwa peristiwa tidak terjadi, sedangkan nilai 1 menunjukkan bahwa peristiwa tersebut terjadi. Probabilitas dapat digunakan untuk meramalkan kejadian yang akan terjadi sehingga dengan ini kita bisa menentukan keputusan yang akan diambil (Murphy, 2012).

## 2.2 Statistika

Statistika merupakan ilmu dasar dari *machine learning*, dimana statistika dapat dilakukan pemahaman fenomena, penentuan keputusan, menguji hipotesis berdasarkan data. Statistik menyelesaikan masalah atau menentukan hasil dengan mencari pola yang memiliki nilai *error* terkecil, sehingga hasil yang diterukan akurat.

Statistika adalah bagian dari matematika yang secara khusus membicarakan cara-cara pengumpulan, analisis dan penafsiran data. Dengan kata lain, istilah statistika di sini digunakan untuk menunjukkan tubuh pengetahuan (*body of knowledge*) tentang cara-cara penarikan sampel (pengumpulan data), serta analisis dan penafsiran data. (Furqon, 2009).

Statistik berfungsi hanya sebagai alat bantu. Peranan statistik dalam penelitian tetap diletakkan sebagai alat. Artinya, statistik bukan menjadi tujuan yang menentukan komponen penelitian lain. Oleh sebab itu, yang berperan menentukan masalah yang dicari jawabannya dan tujuan penelitian itu sendiri. Statistik dapat berguna dalam penyusunan model, perumusan hipotesis, pengembangan alat pengambil data, penyusunan rancangan penelitian, penentuan sampel, dan analisis data, yang kemudian data tersebut diinterpretasikan sehingga bermakna.

Hampir semua penelitian ilmiah dilakukan terhadap sampel kejadian, dan atas dasar sampel itu ditarik suatu *generalization*. Suatu generalisasi pasti mengalami *error*. Disinilah salah satu tugas statistik bekerja atas dasar sampel bukan populasi. Dengan demikian pengujian hipotesis dapat kita lakukan dengan teknik-teknik statistik. Dari hasil analisis statistik yang diperoleh berdasarkan perhitungan yang berbentuk angka-angka tersebut, sebenarnya belum mempunyai arti apa-apa tanpa dideskripsikan dalam bentuk kalimat atau kata-kata di dalam penarikan kesimpulan. Jika tidak, maka hasil analisis tersebut tidak akan bermakna dan hanya tinggal angka-angka yang tidak "berbunyi" (Sugiyono, 2014).

## 3. PEMBAHASAN

Perkembangan ilmu statistika dan *machine learning* tentunya tidak terlepas dari faktor utamanya, yaitu data. *Machine learning* mempunyai setidaknya dua tujuan utama, yaitu pemecahan masalah dalam memprediksi masa depan (*unobserved event*) dan/atau memperoleh ilmu pengetahuan (*knowledge discovery*). *Statistical machine learning* mengacu pada teknik untuk memprediksi masa depan dan mendapatkan pengetahuan dari data secara rasional. Untuk dapat mendapatkan tujuan tersebut, *statistical machine learning* dapat menjadi alat atau metode yang tepat. Statistika berperan sebagai basis pembelajaran yang memanfaatkan teori statistik untuk melakukan inferensi dan interpretasi dari model, sedangkan *machine learning* fokus pada penggunaan model untuk memprediksi data baru.

Topik-topik yang dipelajari statistika dan *machine learning* kurang lebih sama, ilmu-ilmu dasar yang dibutuhkan-pun juga sama, lalu di mana letak perbedaan statistika dan *machine learning*. Perbedaan utamanya terletak pada fokus yang berbeda dari kedua disiplin ilmu tersebut. Statistika lebih fokus ke pengambilan kesimpulan dan interpretasi dari model, sedangkan *machine learning* lebih fokus ke penggunaan model untuk prediksi data baru. Pada dasarnya, *machine learning* memungkinkan komputer untuk mempelajari dan membedakan pola tanpa harus diprogram secara eksplisit. Ketika teknik statistik dan *machine learning* digabungkan bersama-sama keduanya adalah alat yang kuat untuk menganalisis berbagai jenis data di banyak bidang ilmu komputer termasuk, pengolahan gambar, pemrosesan ucapan, pemrosesan bahasa alami, kontrol robot, serta dalam ilmu dasar seperti biologi, kedokteran, astronomi, fisika dan material.

Sebelum lanjut ke tahap pengolahan data, perlu dipahami terlebih dahulu tentang jenis data, karena data menjadi salah satu hal krusial dalam suatu pemodelan. Pada *statistical machine learning*, pemilihan sampel data (atau *training data*) menjadi hal yang sangat penting. Apabila *training data* tidak mampu merepresentasikan populasi, maka model yang dihasilkan pembelajaran (*training*) tidak bagus. Untuk itu, biasanya terdapat juga *test data*. Mesin dilatih menggunakan *training data*, kemudian diuji kinerjanya menggunakan *test data*. Representasi populasi dapat diketahui dengan menyelidiki fitur-fitur (*features* atau karakteristik) yang dimiliki oleh data. Secara ringkas, istilah *training* adalah proses membangun model dan testing adalah proses menguji kriteria model pembelajaran. *Dataset* adalah kumpulan data (sampel dalam statistik). Sampel ini adalah data yang kita gunakan untuk membuat model maupun mengevaluasi model *machine learning*.

### 3.1 Pengolahan Data pada *Machine Learning*

Terdapat dua istilah penting dalam pembangunan model *machine learning* yaitu : *Training* adalah proses konstruksi model dan *Testing* adalah proses menguji kinerja model *learning*. *Dataset* adalah kumpulan data (sampel dalam statistic). Sampel ini adalah data yang digunakan untuk membuat model maupun mengevaluasi model *machine learning*. Umumnya, data set dibagi menjadi tiga jenis yang tidak beririsan (suatu sampel pada himpunan tertentu tidak muncul pada himpunan lainnya) :

- a. *Training set* adalah himpunan data yang digunakan untuk melatih atau membangun model.
- b. *Development set* atau *validation set* adalah himpunan data yang digunakan untuk mengoptimasi saat melatih model. Model dilatih menggunakan *training set* dan pada umumnya kinerja saat Latihan diuji dengan *validation set*. Hal ini berguna untuk generalisasi agar model mampu mengenali pola secara generic.
- c. *Testing set* adalah himpunan data yang digunakan untuk menguji model setelah proses Latihan selesai.

Suatu sampel pada himpunan data disebut sebagai *data point* yang merepresentasikan suatu kejadian statistic (*event*). Proses *training*, *validation* dan *testing data* secara ideal diambil (*sampled*) dari distribusi yang sama dan memiliki karakteristik yang sama (*independent and identically distributed*).

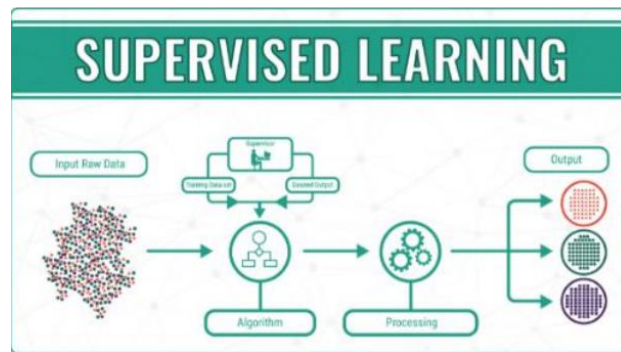
*Machine Learning* memiliki 3 jenis yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*.

#### a. *Supervised Learning*

Pada algoritma ini sistem diberikan *training data set* berupa informasi *input* dan *output* yang diinginkan sehingga sistem akan mempelajari berdasarkan data yang telah ada, sistem akan mencari pola dari data set sehingga pola tersebut akan dijadikannya sebagai acuan untuk data-data berikutnya, dibawah ini adalah cara kerja dari *supervised learning*.

Model *supervised learning* merupakan model yang digunakan untuk memprediksi hasil masa depan dengan berdasarkan data historis yang ada. Model ini diberi instruksi di awal untuk mempelajari sesuatu dan bagaimana cara mempelajarinya. Sebagai contoh, pada model *supervised learning* algoritma yang ada digunakan untuk memprediksi kemungkinan terjadinya bahaya bencana alam, seperti gempa bumi dan tsunami.

Model ini juga bisa diartikan sebagai suatu pendekatan sebuah data yang sudah terlatih. Selain itu, *supervised learning* juga sudah memiliki variabel yang dilabelkan guna mengelompokkan suatu data ke dalam data yang sudah ada.

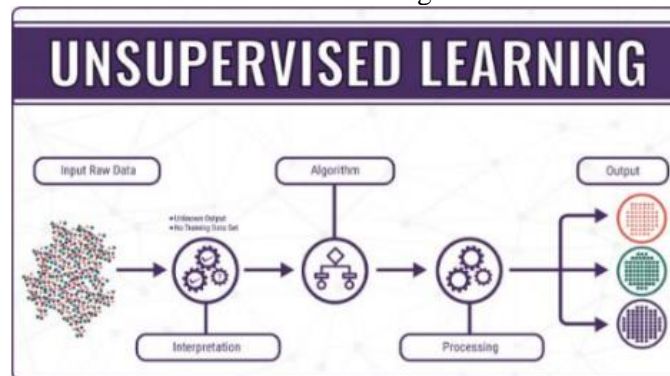


Gambar 1. Cara kerja dari *supervised learning*

b. *Unsupervised Learning*

Berlawanan dengan *supervised*, algoritma ini tidak mendapatkan training data set dikarenakan algoritma ini bukan bersifat prediktif sehingga membutuhkan pembelajaran dari data yang telah ada, akan tetapi algoritma ini bersifat deskriptif, akan berguna untuk mengelompokkan atau mengkategorikan data, berikut adalah cara kerja algoritma *unsupervised learning*.

Model *unsupervised learning* tidak memiliki target atau variabel yang harus ditetapkan. Seorang praktisi data juga tidak memerlukan label khusus untuk memprediksi suatu data. Selain itu, algoritma pada model *unsupervised learning* mampu menemukan pola tersembunyi dalam data itu sendiri. Contoh dari model ini adalah membuat segmentasi pasar untuk melakukan campaign secara efektif berdasarkan clustering.

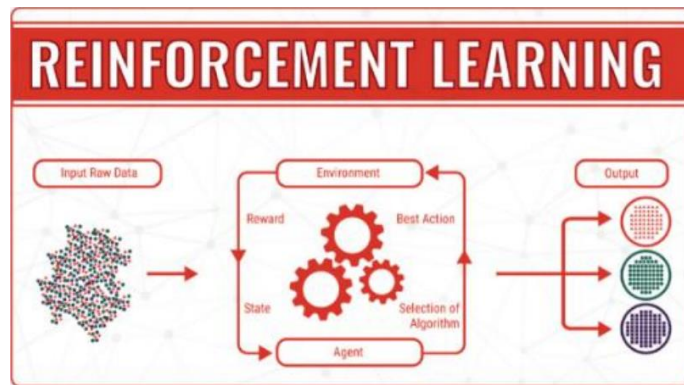


Gambar 2. Cara kerja dari *unsupervised learning*

c. *Reinforcement Learning*

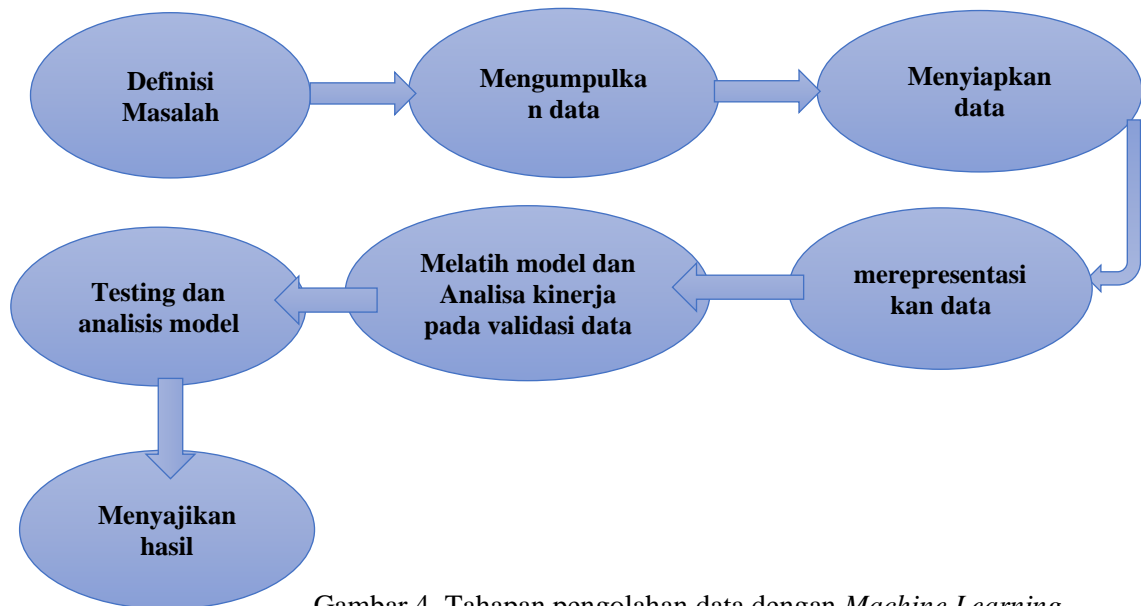
Konsep dasar dari algoritma ini ialah terdapat agent yang ditempatkan di sebuah *environment* yang tidak diketahui tempatnya, kemudian agent akan mengeksplorasi keseluruhan dari *environment* sehingga menemukan istilah reward dan error, apabila agent menemukan sebuah error maka ia akan terus mencari jalan hingga mendapatkan *reward*. Eksplorasi ini beroperasi seperti trial and error, agent akan terus melakukan percobaan akan tetapi agent tidak akan melakukan error pada satu masalah yang sama.

Pada model *reinforcement learning*, mesin dilatih untuk mengambil suatu keputusan secara spesifik berdasarkan kebutuhan bisnis yang bertujuan untuk memaksimalkan kinerjanya. Model ini membuat suatu perangkat lunak atau mesinnya melatih dirinya secara terus menerus dengan berdasarkan lingkungan yang dipengaruhinya. Selain itu, model ini juga diterapkan untuk mampu memecahkan permasalahan pada suatu bisnis.



Gambar 3. Cara kerja dari *Reinforcement learning*

Teknik dan prosedur dalam penerapan *machine learning* sangat bervariasi, namun secara garis besar sebuah project *machine learning* memiliki tahapan sebagai berikut:



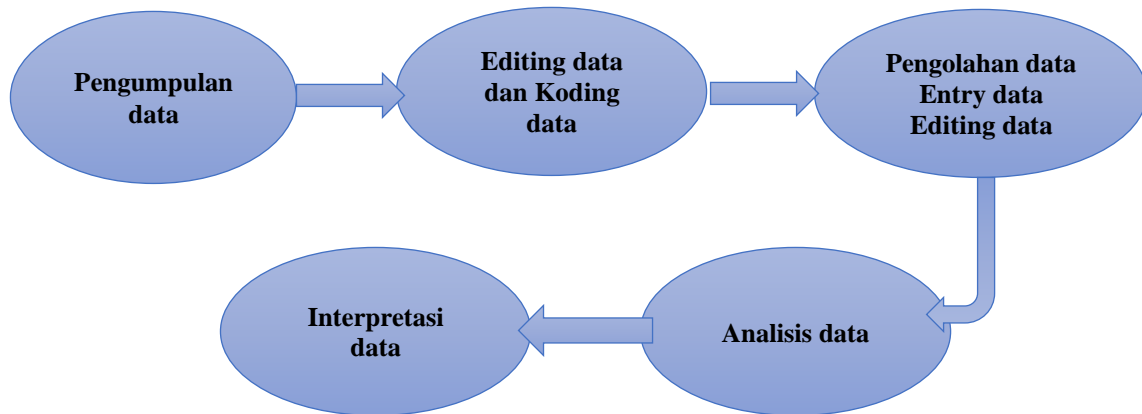
Gambar 4. Tahapan pengolahan data dengan *Machine Learning*

- Definisi masalah  
Masalah adalah Ketika tujuan yang diinginkan tidak tercapai (*current state* bukanlah *desired state*). Agar *current state* menjadi *desired state*, perlu dilakukan kegiatan yang disebut *problem solving*. Tiap bidang (domain) mendefinisikan permasalahan secara berbeda. Oleh karena itu, mengetahui Teknik *machine learning* tanpa mengetahui domain aplikasi adalah sesuatu yang kurang baik.
- Mengumpulkan data  
Terbentuk dari simpanan pada beberapa file yang berisikan data yang dapat dipelajari komputer, dan setelah itu dipisah antara fitur masukan (input) dan keluaran (output).
- Menyiapkan data.  
Penentuan kualitas data pada *preprocessing* data sehingga hasil yang didapat juga optimal/baik
- Merepresentasikan data sebagai *feature vector*.  
Langkah ini dilakukan dengan pemilihan algoritma dan representasi data yang tepat dalam bentuk model.
- Melatih model dan menganalisis kinerjanya pada *validation (development)* data.
- Melakukan *testing* dan analisis model baik kuantitatif dan kualitatif.  
Langkah ini menguji keakuratan hasil berdasarkan bagian *test dataset*.
- Menyajikan hasil (presentasi).

### 3.2 Pengolahan Data pada Statistika

Dalam Undang-undang Nomor 16 Tahun 1997 tentang Statistik, pengertian statistik memiliki tiga dimensi yaitu data atau informasi yang berupa angka, sistem yang memadukan penyelenggaraan statistik, serta ilmu atau metode yang mempelajari cara pengumpulan, pengolahan, penyajian, dan analisis data. Sedangkan statistika adalah ilmu pengaturan data dan probabilitas. Statistika mengajar kita mengatur data dan menghitung kemungkinan-kemungkinan di masa depan.

Tujuan pokok dilaksanakan suatu penelitian adalah untuk menjawab pertanyaan – pertanyaan penelitian. Untuk mencapai tujuan tersebut antara lain harus melalui proses pengolahan dan analisis data. Alur kerja dimulai dari pengumpulan hingga interpretasi data dalam ilustrasi berikut :



Gambar 5. Tahapan pengolahan data dengan Statistika

Prosedur dalam pengolahan data statistic adalah :

a. Pengumpulan Data

Sebelum melakukan pengolahan data, ada beberapa tahap yang harus dilakukan. Sedangkan setelah analisis data yaitu suatu proses penyederhanaan data, maka dapat dilakukan interpretasi data dengan mudah. Kuesioner merupakan salah satu alat pengumpul data yang digunakan untuk survai, guna memudahkan proses selanjutnya, sebaiknya dalam kuesioner telah tersedia kolom untuk koding.

b. Editing Data.

Data lapangan yang ada misalkan dalam kuesioner perlu diedit, tujuannya adalah untuk melihat lengkap atau tidaknya pengisian kuesioner, melihat logis atau tidaknya jawaban dan melihat konsistensi antar pertanyaan.

c. Koding Data

Langkah ini dilakukan untuk jenis pertanyaan tertutup (dilakukan pengkodean sebelum ke lapangan), setengah terbuka (pengkodean sebelum dan sesudah dari lapangan), dan terbuka (pengkodean sepenuhnya dilakukan setelah dari lapangan).

d. Pengolahan Data

Terdapat dua hal yang perlu dilakukan Ketika melakukan pengolahan data yaitu *entry data* atau memasukkan data dalam proses tabulasi , dan melakukan editing ulang terhadap data yang telah ditabulasi untuk mencegah terjadinya kekeliruan memasukkan data, atau salah penempatan dalam kolom maupun baris table.

e. Analisis dan Interpretasi Data

Hal penting yang perlu diingat dalam melakukan analisis data adalah mengetahui dengan tepat penggunaan alat analisis, sebab jika tidak memenuhi prinsip-prinsip dari pemakaian alat analisis, walaupun alat analisis tersebut sangat canggih, maka hasil yang diperoleh akan salah diinterpretasikan dan menjadi tidak manfaat untuk mengambil suatu kesimpulan. Model – model statistika untuk keperluan analisis data telah berkembang, dari model statistika deskriptif hingga statistika inferensial non parametrik.

Ketika diputuskan untuk melakukan analisis data menggunakan alat statistika, ada beberapa hal yang perlu diperhatikan antara lain :

a. Dari mana data diperoleh, apakah berasal dari sampel (melalui proses sampling) atau dari populasi (dengan cara sensus).

- b. Jika berasal dari sampel apa teknik sampling yang digunakan, apakah termasuk kelompok sampling probabilitas atau non probabilitas.
- c. Memakai skala apa data diukur, apakah menggunakan skala nominal, ordinal, interval, atau rasio.
- d. Bagaimana hipotesis yang dibuat apakah perlu dilakukan pengujian satu arah atau dua arah kalau memakai statistika inferensial.

### 3.3 Perbedaan Teknik pengolahan data dengan statistika dan *machine learning*

Salah satu ilustrasi sederhana tentang *statistical machine learning* dapat digambarkan dengan suatu variabel *dependen* kategorik  $y$  yang ingin dianalisis dengan beberapa variabel independen  $x_1, \dots, x_m$ . Contoh aplikasinya misalnya analisis tentang kelulusan mahasiswa dimana  $y = 1$  bermakna lulus dan  $y = 0$  bermakna tidak lulus. Variabel-variabel  $x_i$  bisa berupa profil mahasiswa, total jam belajar, nilai mata kuliah, keaktifan di kelas, dan variabel-variabel lain yang berhubungan dengan kelulusan. Teknik yang sering dipakai untuk memodelkan permasalahan seperti ini adalah menggunakan model Regresi Logistik.

Statistika dan *machine learning* membentuk satu konsep yang disebut *Statistical machine learning* dengan menggunakan model regresi logistik. Meskipun peneliti statistika rata-rata memulai belajar dari latar belakang matematika, sedangkan peneliti *machine learning* rata-rata memulai dari latar belakang algoritma. Namun statistika dan *machine learning* menggunakan teori dan rumus yang sama dalam menyelesaikan suatu permasalahan. Keduanya sama-sama membahas konsep random variable, distribusi-distribusi statistik, expected value, variansi, sampai pada konsep distribusi prior dan posterior.

Alur kerja praktisi statistika dalam membangun model regresi logistik biasanya dimulai dengan pengecekan asumsi, menjalankan estimasi parameter (maximum likelihood) dengan software, mengecek nilai parameter dan signifikansinya, dan memilih variabel yang signifikan. Untuk mengecek dan membandingkan model mana yang lebih bagus (variabel mana saja yang sebaiknya dimasukkan ke model), ukuran seperti Akaike Information Criterion (AIC) akan digunakan. Setelah modelnya fix, praktisi statistika akan menganalisis interpretasi dari parameter yang dihasilkan, seperti variabel apa yang mempengaruhi kelulusan, apakah mempengaruhi secara positif/negatif, dan seberapa besar pengaruhnya (bisa dicek dari nilai parameter untuk variabel tersebut). Pada akhirnya akan praktisi statistika akan menarik kesimpulan tentang hubungan variabel-variabel tersebut dengan kelulusan mahasiswa.

Contoh permasalahan yang dihadapi praktisi *machine learning* dengan tipe yang mirip, satu variabel *dependen* kategorik  $y$  dan beberapa variabel independen  $x_1; \dots; x_m$ . Contoh aplikasinya misalnya untuk mendeteksi email spam. Nilai variabel  $y = 1$  berarti email spam,  $y = 0$  berarti bukan email spam. Variabel-variabel  $x_j$  berisi karakteristik dari email tersebut, misalnya jumlah kata di email, apakah email mengandung attachment, apakah ada gambar di email itu, atau apakah email mengandung kata-kata tertentu, misalnya kata 'promo' atau 'obat'. Model regresi logistik juga termasuk salah satu model yang sering dipakai praktisi *machine learning*. Mereka juga melakukan estimasi parameter dengan maximum likelihood via *software*.

Dari dua alur kerja diatas, walaupun sama-sama menggunakan model regresi logistik, terlihat jelas perbedaan fokus dari praktisi statistika dan *machine learning*. Selain dari fokus yang berbeda (interpretasi versus prediksi), ada beberapa perbedaan yang terlihat. Diantaranya adalah bagaimana memperlakukan parameter hasil estimasi. Tidak seperti praktisi statistika, praktisi *machine learning* kurang peduli tentang nilai dari parameter-parameter tersebut. Ada preferensi di mana praktisi *machine learning* ingin agar nilai parameter-parameter tersebut relatif kecil dengan menambahkan regularisasi, namun berapapun nilainya tidak menjadi permasalahan, asalkan model tersebut bisa memprediksi data baru dengan akurat. Perbedaan yang lain ada dalam cara pemilihan model. Praktisi statistika cenderung memilih model berdasarkan teori analitik seperti AIC, sedangkan praktisi *machine learning* cenderung memilih model berdasarkan performa empirikal di tahap *cross validation*.

#### 4. KESIMPULAN

Disiplin ilmu statistika dan *machine learning* mempunyai banyak persamaan dan juga perbedaan. Kedua disiplin ilmu sama-sama berdasarkan teori peluang dan membahas dasar-dasar teori dan model yang sama. Perbedaan keduanya terletak pada fokus yang berbeda. Statistika lebih fokus ke arah pengambilan kesimpulan, sedangkan *machine learning* fokus ke prediksi data baru. *Machine learning* dengan fungsi untuk mencari pola pada data, lalu apabila pola dari data sudah diketahui *machine learning* dapat mengambil kesimpulan berupa pola yang telah terbentuk sehingga keputusan dapat ditentukan setelah kesimpulan telah ditentukan, *machine learning* memiliki sifat prediktif dan deskriptif yang membuktikan bahwa *machine learning* akan menjadi suatu problematika yang amat penting untuk masa yang akan mendatang, karena dengan *machine learning* kita dapat menentukan keputusan dengan tepat. Dalam menghadapi suatu masalah, peneliti statistika lebih melihat dari sisi formulasi matematika untuk memodelkan masalah. Masing – masing Teknik baik statistika dan *machine learning* mempunyai keunggulan , dimana kedua disiplin ilmu saling melengkapi khususnya dalam perkembangan teknologi *Big Data*.

#### 5. DAFTAR PUSTAKA

- Furqon, 2009. Statistika Terapan Untuk Penelitian. Bandung : Alfabeta
- Harrell, Frank. 2015. Regression Modeling Strategies: With Applications To Linear Models, Logistic And Ordinal Regression, And Survival Analysis. Springer, 2015.
- K. P. Murphy. 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.
- P. Dangeti, 2017. Statistics for Machine Learning. Mumbai:Packt.
- Putra, Jan Wira Gotama. 2020. Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4, Tokyo Jepang.
- Sugiyono. 2014. Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta.