**Apa itu Regresi, Klasifikasi, dan Clustering (Klasterisasi)?**

Belajar tentang *data science*, pasti kita akan menemukan tiga istilah ini, regresi, klasfikasi, dan *clustering* atau klasterisasi. Ketiga istilah tersebut merupakan metode-metode yang populer digunakan di dunia *data science* dan *machine learning*. Nah, jadi apa itu regresi, klasifikasi, dan *clustering*?

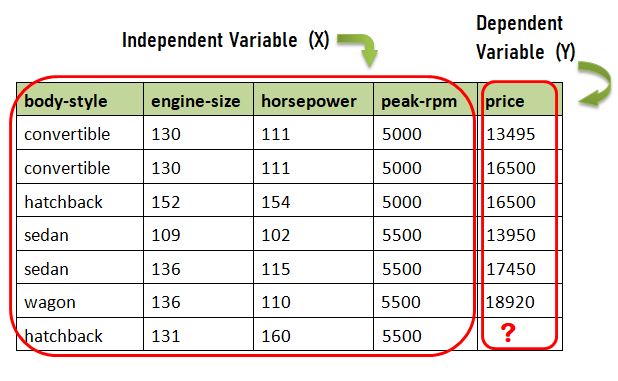
Dalam menganalisis data, tentunya kita berharap akan menemukan informasi dari data tersebut bukan? kita pasti akan memikirkan, mau diapakan data ini? Metode apa yang sesuai digunakan dengan data ini  untuk dapat menyelesaikan masalah?

Regresi, klasifikasi, dan *clustering* merupakan tiga metode yang sering digunakan untuk analisis data. Selain tiga teknik tersebut, masih ada beberapa teknik atau metode lainnya yang dapat digunakan yang telah saya singgung sedikit di artikel [Pengenalan Machine Learning](https://ilmudatapy.com/pengenalan-machine-learning/). Sementara itu di artikel ini, saya hanya akan membahas tentang apa itu regresi, klasifikasi, dan *clustering*.

**Regresi (Regression)**

Regresi adalah suatu teknik analisis untuk mengidentifikasi relasi atau hubungan diantara dua variabel atau lebih. Regresi bertujuan untuk menemukan suatu fungsi yang memodelkan data dengan meminimalkan *error* atau selisih antara nilai prediksi dengan nilai sebenarnya.

Regresi termasuk ke dalam *supervised learning* yang digunakan untuk memprediksi nilai kontinu. Untuk lebih jelasnya, perhatikan potongan data dari dataset automobile. Dataset lengkapnya dapat di akses [disini](https://archive.ics.uci.edu/ml/datasets/Automobile).



Data di atas terdiri dari beberapa variabel di antaranya *body-style*, *engine-size*, *horsepower*, *peak-rpm*, dan *price*. Saya sengaja mengambil beberapa variabel saja sebagai contoh, sementara dataset automobile yang sesungguhnya terdiri dari 26 variabel.

Dalam regresi, ada dua jenis variabel yaitu *Dependent variable*dan*Independent variable.* Variabel dependen adalah variabel yang akan kita prediksi atau pelajari, sedangkan variabel independen adalah variabel yang menjelaskan atau menyebabkan nilai target di variabel dependen.

Variabel independen dinotasikan dengan X, sementara variabel dependen dinotasikan dengan Y. Yang harus diperhatikan dalam kasus regresi adalah nilai dari variabel dependen (Y) harus berupa nilai kontinu, bukan diskrit. Sementara itu untuk variabel independen (X) bisa berupa nilai kontinu maupun kategori, misalnya sedan, hatchback, wagon, convertible.

Untuk memprediksi harga mobil, kita harus membuat model regresi dari data sebelumnya. Setelah model selesai dibuat, kita dapat menggunakannya untuk memprediksi harga mobil menggunakan data baru.

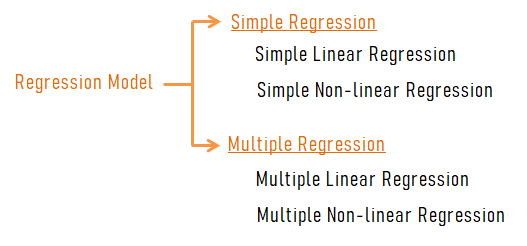
**Tipe model regresi**

Pada dasarnya, ada dua tipe model regresi yaitu *simple regression* (regresi sederhana), dan *multiple regression* (regresi berganda).

*Simple regression* adalah ketika hanya satu variabel independen yang digunakan untuk memprediksi dependen variabel, bisa berupa linear (*simple linear regression*) maupun non-linear (*simple non-linear regression*). Misalnya memprediksi harga mobil hanya dengan berdasarkan *engine-size* saja.

*Multiple regression* adalah ketika ada lebih dari satu variabel independen yang digunakan untuk memprediksi variabel dependen. Misalnya memprediksi harga mobil berdasarkan *engine-size*, *body-style*, *horsepowe*r, dan sebagainya.

Linearitas regresi ditentukan berdasarkan sifat hubungan antara variabel independen dan dependen. Sama seperti *simple regression*, *multiple regression* juga bisa berupa linear maupun non-linear.



**Algoritma regresi**

*Linear regression* merupakan salah satu algoritma regresi yang paling populer. Kenyataannya, bukan hanya *linear regression* saja yang dapat kita gunakan untuk pemodelan regresi, tetapi ada banyak algoritma lainnya yang dapat kita coba, di antaranya:

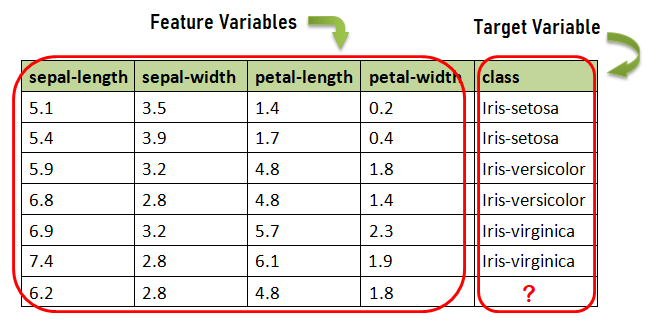
* Linear Regression
* Polynomial Regression
* Lasso Regression
* Ridge Regression
* Bayesian Linear Regression
* Decision Tree Regression
* Neural Network Regression
* dan lain-lain.

Algoritma-algoritma yang disebutkan di atas hanya sebagian dari teknik yang dapat digunakan untuk kasus regresi. Seiring dengan proses belajar, mungkin kita akan menemukan serta dapat mengeksplorasi teknik-teknik regresi lainnya.

**Klasifikasi (Classification)**

Sama halnya dengan regresi, klasifikasi juga termasuk ke dalam *supervised learning*. Klasifikasi adalah sebuah teknik untuk mengklasifikasikan atau mengkategorikan beberapa *item* yang belum berlabel ke dalam sebuah set kelas diskrit.

Klasifikasi mencoba mempelajari hubungan antara kumpulan variabel fitur dan variabel target. Dalam klasifikasi, variabel targetnya bertipe kategori.

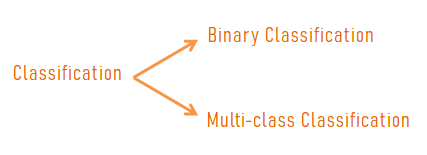


Dapat dilihat pada contoh gambar tabel di atas, di dalam klasifikasi ada dua jenis variabel yaitu *Target variable* dan *Feature variabel*. Ini sebenarnya mirip dengan regresi, hanya penggunaan istilahnya saja yang berbeda karena disesuaikan dengan konsep masing-masing.

Perbedaan klasifikasi dan regresi yang dapat kita lihat dengan jelas adalah nilai dari variabel targetnya. Di klasifikasi, variabel target harus berupa kategori atau nilai diskrit. Data baru yang berlabel nantinya akan dikelompokkan ke dalam salah satu dari kategori yang ada variabel target.

**Tipe model klasifikasi**

Di dalam machine learning, klasifikasi dibagi menjadi dua jenis yaitu *Binary classification* dan *Multi-class classification.*



*Binary classification* adalah jika kategori dalam target variabel hanya ada dua, misalnya 0 dan 1, *Yes*dan *No*, X dan Y, dan sebagainya. Misalnya melihat kemungkinan nasabah bank akan mengambil pinjaman atau tidak.

Sebaliknya, *Multi-class classification* memiliki lebih dari dua kategori pada variabel targetnya. Contoh dataset Iris di atas termasuk jenis *multi-class classification* karena memiliki tiga kategori yaitu Iris-setosa, Iris-versicolor, dan Iris-virginica.

**Algoritma klasifikasi**

Untuk menyelesaikan masalah klasifikasi, kita dapat menggunakan algoritma-algoritma berikut ini, di antaranya:

* Decision Tree
* Naïve Bayes
* K-Nearest Neighbor (KNN)
* Logistic Regression
* Support Vector Machines (SVM)
* Linear Discriminant Analysis
* Neural Network
* dan lain-lain.

Perlu dicatat bahwa, tidak seperti namanya, *Logistic regression* bukan merupakan algoritma untuk menyesaikan masalah regresi, melainkan klasifikasi.

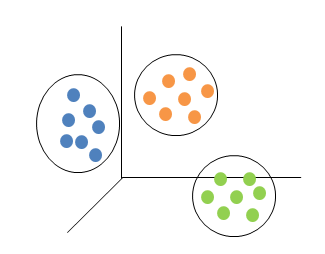
**Klasterisasi (Clustering)**

*Clustering* atau klasterisasi termasuk *unsupervised learning*. *Clustering* dilakukan jika kita ingin menemukan klaster dari sebuah dataset.

Jadi, apa itu klaster? Klaster adalah sebuah kumpulan data atau objek yang memiliki kemiripan satu sama lain di dalam kumpulan atau kelompok tersebut, dan berbeda dengan objek di kelompok lain.

Tidak seperti klasifikasi dimana setiap data latih sudah memiliki label alias sudah ditentukan kelasnya, dalam clustering, data tidak berlabel.

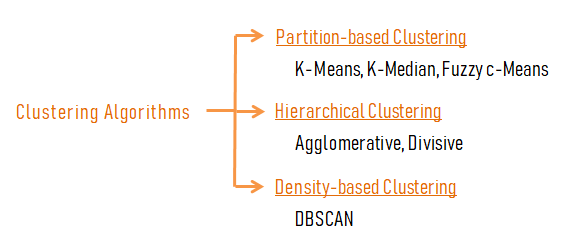
*Clustering* umumnya banyak digunakan untuk segmentasi. Proses segmentasi biasanya tidak dapat dikerjakan dengan mudah untuk data dengan jumlah yang sangat besar, oleh karena itu dibutuhkan pendekatan analitis untuk membagi data tersebut menjadi bagian-bagian yang lebih kecil berdasarkan kemiripan karakteristiknya. Perhatikan gambar berikut.



Gambar di atas adalah ilustrasi bagaimana data dibagi ke dalam kelompok-kelompok yang lebih kecil, dalam contoh ini data dibagi menjadi tiga kelompok. Setelah dilakukan pengelompokkan seperti di atas, selanjutnya adalah memberi nama yang sesuai dengan karakteristik masing-masing kelompok.

**Algoritma clustering**

Algoritma *clustering* terbagi menjadi tiga jenis, yaitu *Partition-based clustering*, *Hierarchical clustering*, dan *Density-based clustering*.



*Partition-based clustering* merupakan kelompok algoritma *clustering* yang menghasilkan bagian-bagian yang lebih kecil atau klaster-klaster seperti gambar ilustrasi di atas. Contoh algoritma jenis ini di antaranya K-Means, K-Median, atau Fuzzy c-Means. Algoritma-algoritma ini relatif efisien dan digunakan untuk data dengan jumlah sedang dan besar.

Jenis algoritma berikutnya adalah *hierarchical clustering*. Algoritma-algoritma jenis ini menghasilkan pohon-pohon klaster, misalnya algoritma Agglomerative dan Divisive. Algoritma ini sangat intuitif dan secara umum baik digunakan untuk dataset berukuran kecil alias tidak terlalu mengandung banyak data.

Yang terakhir yaitu *density-based clustering*. Algoritma *density-based clustering* menghasilkan klaster-klaster sesuai kehendaknya. Karena membentuk klaster sesuai kehendaknya, algoritma ini bekerja dengan baik ketika berurusan dengan *noise* yang ada di dataset atau ketika ada data yang tidak masuk ke klaster manapun. Contoh algoritma ini adalah DBSCAN.