Jenis kategori selanjutnya adalah regression. Submodul sebelumnya telah membahas sekelumit tentang regresi. Regresi adalah salah satu teknik ML yang mirip dengan klasifikasi. Bedanya pada klasifikasi, sebuah model ML memprediksi sebuah kelas, sedangkan model regresi memprediksi bilangan kontinu. Bilangan kontinu adalah bilangan numerik.

Jadi model klasifikasi memprediksi kelas atau kategori dan model regresi memprediksi sebuah nilai berdasarkan atribut yang tersedia. Agar lebih paham, perhatikan contoh di bawah.

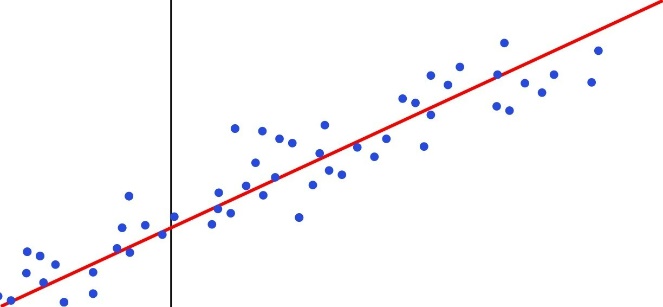
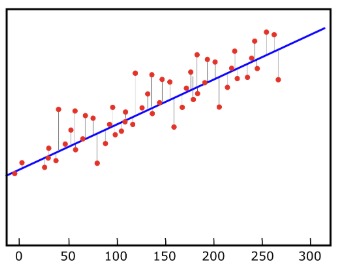
| **Lama Bekerja** | **Industri** | **Tingkat Pendidikan** | **Gaji** |
| --- | --- | --- | --- |
| 6 tahun | Marketing | SMA | 8.000.000 |
| 12 tahun | IT | S1 | 16.000.000 |
| 8 tahun | Kesehatan | S2 | 20.000.000 |
| 5 tahun | IT | SMK | ? |
| 6 tahun | Marketing | S2 | 14.000.000 |
| 21 tahun | Perbankan | S3 | 35.000.000 |
| 3 tahun | IT | S1 | 10.000.000 |

Pada contoh data di atas, model regresi akan memprediksi gaji berdasarkan atribut lama bekerja, industri, dan tingkat pendidikan. Gaji adalah contoh dari bilangan kontinu, di mana gaji tak memiliki kategori-kategori yang terbatas.

Pada submodul ini jenis regresi yang akan dibahas adalah regresi linier. Selain regresi linier terdapat juga jenis regresi lain seperti regresi polinomial, lasso regression, stepwise regression dan sebagainya. Untuk penjelasan dari jenis-jenis regression yang ada, kunjungi [tautan](https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/) berikut.

**Linear Regression**

Regresi linier adalah salah satu metode supervised yang masuk dalam golongan regression, sesuai namanya. Contoh paling terkenal dari regresi linier adalah memperkirakan harga rumah berdasarkan fitur yang terdapat pada rumah seperti luas rumah, jumlah kamar tidur, lokasi dan sebagainya. Ini adalah model paling sederhana yang perlu diketahui guna memahami metode machine learning lain yang lebih kompleks. Regresi linier cocok dipakai ketika terdapat hubungan linear pada data. Namun untuk implementasi pada kebanyakan kasus, ia kurang direkomendasikan. Sebabnya, regresi linier selalu mengasumsikan ada hubungan linier pada data, padahal tidak.

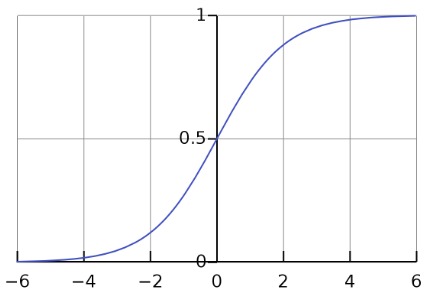
1. Secara sederhana regresi linear adalah teknik untuk memprediksi sebuah nilai dari variable Y (variabel dependen) berdasarkan beberapa variabel tertentu X (variabel independen) jika terdapat hubungan linier antara X dan Y.
2. Hubungan antara hubungan linier dapat direpresentasikan dengan sebuah garis lurus (disebut garis regresi). Ilustrasi hubungan linier dapat dilihat pada gambar di mana data-data cenderung memiliki pola garis lurus.  
   [](https://www.dicoding.com/academies/184/tutorials/8392?from=8397)
3. Ketika sebuah garis regresi digambar, beberapa data akan berada pada garis regresi dan beberapa yang lainnya akan berada di dekat garis tersebut. Sebabnya, garis regresi adalah sebuah model probabilistik dan prediksi kita adalah perkiraan. Jadi tentu akan ada eror/penyimpangan terhadap nilai asli dari variabel Y. Pada gambar di bawah, garis abu-abu (garis dari titik merah) yang menghubungkan data-data ke garis regresi (garis biru) merupakan error. Semakin banyak eror artinya model regresi itu belum optimal.  
   [](https://www.dicoding.com/academies/184/tutorials/8392?from=8397)

**Logistic Regression**

Setelah sebelumnya Anda mengenal regresi linier untuk masalah regresi, ada juga model seperti logistic regression, terlepas dari namanya merupakan sebuah model yang dapat digunakan untuk klasifikasi.

Logistic regression dikenal juga sebagai logit regression, maximum-entropy classification, dan log-linear classification merupakan salah satu metode yang umum digunakan untuk klasifikasi. Pada kasus klasifikasi, logistic regression bekerja dengan menghitung probabilitas kelas dari sebuah sampel.

Sesuai namanya, logistic regression menggunakan fungsi logistik seperti di bawah untuk menghitung probabilitas kelas dari sebuah sampel. Contohnya sebuah email memiliki probabilitas 78% merupakan spam maka email tersebut termasuk dalam kelas spam. Dan jika sebuah email memiliki <50% probabilitas merupakan spam, maka email tersebut diklasifikasikan bukan spam.

[](https://www.dicoding.com/academies/184/tutorials/8392?from=8397)