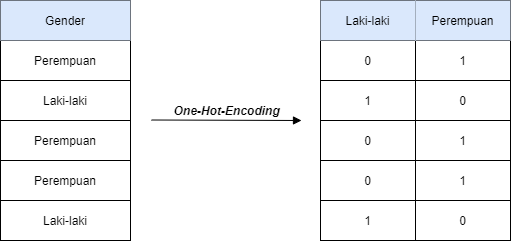
Setelah dataset dibersihkan, masih ada beberapa tahap yang perlu dilakukan agar dataset benar-benar siap untuk diproses oleh model machine learning. Biasanya, dataset Anda akan terdiri dari dua jenis data: kategorik dan numerik. Contoh data numerik adalah: ukuran panjang, suhu, nilai uang, hitungan dalam bentuk angka, dll, yang terdiri dari bilangan integer (seperti -1, 0, 1, 2, 3, dan seterusnya) atau bilangan float (seperti -1.0, 2.5, 39.99, dan seterusnya). Setiap nilai dari data dapat diasumsikan memiliki hubungan dengan data lain karena data numerik dapat dibandingkan dan memiliki ukuran yang jelas. Misal, Anda dapat mengatakan bahwa panjang 39 m lebih besar dibanding 21 m. Jenis data ini terdefinisi dengan baik, dapat dioperasikan dengan metode statistik, dan mudah dipahami oleh komputer.

Jenis data lain yang sering kita temui adalah data kategorik. Data kategorik adalah data yang berupa kategori dan berjenis string, tidak dapat diukur atau didefinisikan dengan angka atau bilangan. Contoh data kategorik adalah sebuah kolom pada dataset yang berisi perkiraan cuaca seperti cerah, berawan, hujan, atau berkabut. Contoh lain dari data kategorik adalah jenis buah misalnya apel, pisang, semangka, dan jeruk. Pada jenis data ini, kita tidak bisa mendefinisikan operasi perbandingan seperti lebih besar dari, sama dengan, dan lebih kecil dari. Dan dengan demikian, kita juga tidak dapat mengurutkan dan melakukan operasi statistik terhadap data jenis ini.

Umumnya, model machine learning tidak dapat mengolah data kategorik, sehingga kita perlu melakukan konversi data kategorik menjadi data numerik. Banyak model machine learning seperti Regresi Linear dan Support Vector Machine (kedua model ini akan dibahas pada modul-modul selanjutnya) yang hanya menerima input numerik sehingga tidak bisa memproses data kategorik. Salah satu teknik untuk mengubah data kategorik menjadi data numerik adalah dengan menggunakan One Hot Encoding atau yang juga dikenal sebagai dummy variables. One Hot Encoding mengubah data kategorik dengan membuat kolom baru untuk setiap kategori seperti gambar di bawah.

[](https://www.dicoding.com/academies/184/tutorials/8347?from=8334)

Pada modul latihan nanti, Anda akan berlatih bagaimana menerapkan teknik one-hot-encoding pada dataset. Sudah siap untuk lanjut?