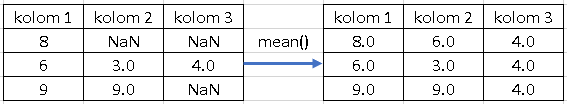
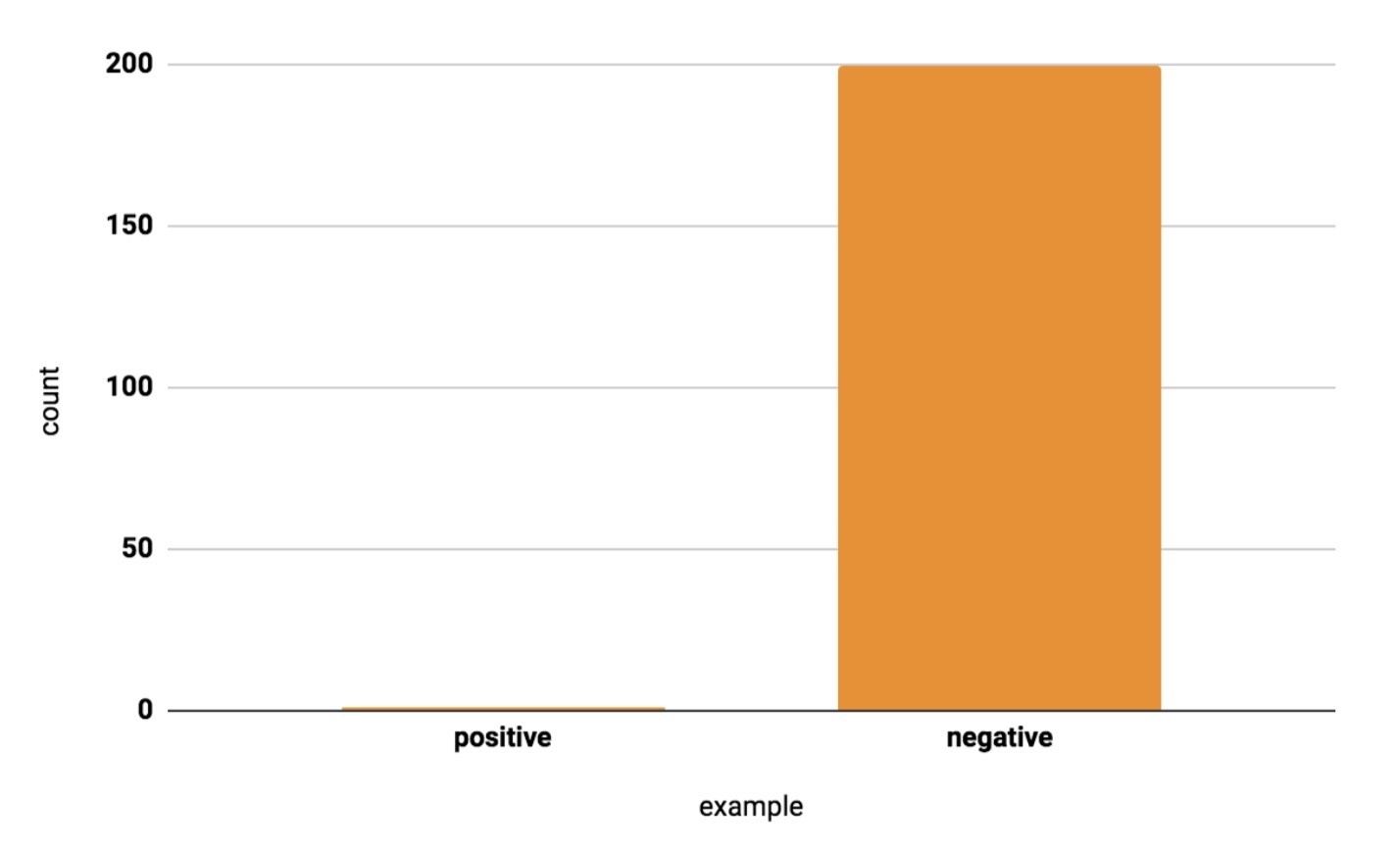
**Data Cleaning**

Data cleaning merupakan tahapan penting yang tidak boleh Anda lewatkan. Berikut adalah beberapa hal umum yang harus diperhatikan dalam proses data cleaning:

1. Konsistensi Format  
   Sebuah variabel mungkin tidak memiliki format yang konsisten seperti penulisan tanggal 10-Okt-2020 versus 10/10/20. Format jam yang berbeda seperti 17.10 versus 5.10 pm. Penulisan uang seperti 17000 versus Rp 17.000. Data dengan format berbeda tidak akan bisa diolah oleh model machine learning. Solusinya, format data harus disamakan dan dibuat konsisten terlebih dahulu.
2. Skala Data  
   Jika sebuah variabel memiliki jangka dari 1 sampai 100, pastikan tidak ada data yang lebih dari 100. Untuk data numerik, jika sebuah variabel merupakan bilangan positif, maka pastikan tidak ada bilangan negatif.
3. Duplikasi data  
   Data yang memiliki duplikat akan mempengaruhi model machine learning, apalagi jika data duplikat tersebut besar jumlahnya. Untuk itu kita harus memastikan tidak ada data yang terduplikasi.
4. Missing Value  
   Missing value terjadi ketika data dari sebuah record tidak lengkap. Missing value sangat mempengaruhi performa model machine learning. Ada 2 (dua) opsi untuk mengatasi missing value, yaitu menghilangkan data missing value atau mengganti nilai yang hilang dengan nilai lain, seperti rata-rata dari kolom tersebut (mean) atau nilai yang paling sering muncul (modus), atau nilai tengah (median).

Berikut adalah contoh handling missing value dengan nilai rata-rata (mean).  
[](https://www.dicoding.com/academies/184/tutorials/8342?from=11946)

1. Skewness Distribution  
   Skewness adalah kondisi di mana dataset cenderung memiliki distribusi data yang tidak seimbang. Skewness akan mempengaruhi data dengan menciptakan bias terhadap model. Apa itu bias? Sebuah model cenderung memprediksi sesuatu karena ia lebih sering mempelajari hal tersebut. Misalkan ada sebuah model untuk pengenalan buah di mana jumlah jeruk 92 buah dan apel 8 buah. Distribusi yang tidak imbang ini akan mengakibatkan model lebih cenderung memprediksi jeruk daripada apel.  
   [](https://www.dicoding.com/academies/184/tutorials/8342?from=11946)  
   Cara paling simpel untuk mengatasi skewness adalah dengan menyamakan proporsi kelas mayoritas dengan kelas minoritas. Untuk teknik lebih lanjut dalam mengatasi skewness atau imbalance data, Anda bisa membacanya di [tautan](https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data" \t "_blank) ini.

Refrence

<https://www.kdnuggets.com/2019/06/7-steps-mastering-data-preparation-python.html>

<https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4>

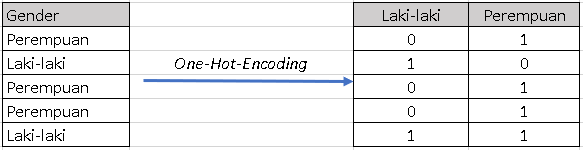
<https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data>

Selain Google Collab ada IBM Watson

<https://imam.digmi.id/post/google-colab-gratis-untuk-belajar-deep-learning/>

**Data Preparation dengan Teknik One-Hot-Encoding**

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)

**Data Preparation - Normalization dan Standardization**

**Outlier Removal**

**Decision Tree**

*Decision tree* atau pohon keputusan adalah salah satu algoritma supervised learning yang dapat dipakai untuk masalah klasifikasi dan regresi. Decision tree merupakan algoritma yang powerful alias mampu dipakai dalam masalah yang kompleks. Decision tree juga merupakan komponen pembangun utama algoritma Random Forest, yang merupakan salah satu algoritma paling powerful saat ini.

Decision tree memprediksi sebuah kelas (klasifikasi) atau nilai (regresi) berdasarkan aturan-aturan yang dibentuk setelah mempelajari data.

Misalnya kita memiliki data seperti di bawah. Data berisi informasi mengenai kondisi cuaca pada hari tertentu dan apakah cocok untuk bermain golf di kondisi cuaca tersebut.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outlook | Temperature | Humidity | Windy | Play Golf |
| Rainy | Hot | High | False | No |
| Rainy | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Sunny | Mild | High | False | Yes |
| Sunny | Cold | Normal | False | Yes |
| Sunny | Cold | Normal | True | No |
| Overcast | Cold | Normal | True | Yes |
| Rainy | Mild | High | False | No |
| Rainy | Cold | Normal | False | Yes |
| Rainy | Mild | Normal | False | Yes |

## **Supervised : Regression**

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.

Refrence:

<https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>