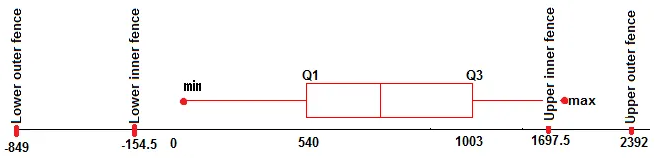
**Teknik dalam Machine Learning (ML/AI) - Classification/Prediction**

1. **EDA - Data Preparation**
2. Mengenali semua arti variable/kolom
3. Melihat informasi data
4. Dimensi Dataset
5. Tipe Data dari tiap kolom/atribut/variable
6. Cek kelompok variabel numerik dan kelompok variabel categorical . Jika kategorikal nanti pada preprocessing di lakukan penanganan menggunakan label encoder/ binary encoder (untuk 2 jenis class)
7. Jika ada data date gunakan metode endcoding untuk memisahkan nilai dd/mm/yyyy menjadi kolom baru yaitu dd | mm | yyyy
8. Mengetahui summary dataset
9. Mencari Korelasi variable (Heatmap - Scatter Plot)
10. Distribusi Data/Histogram Plot (Melihat Persebaran Data)
11. Mencari outlier/Mendeteksi outlier (Distribusi Data/Histogram Plot, Boxplot, Ballpoint plot) : https://pacmann.io/blog/cara-mendeteksi-dan-menangani-outlier-saat-melakukan-data-analysis
    1. Jika Hasil Plot Histogramnya itu normal distribusi -> Std Standard Deviasi (Plot Histogram
    2. Jika Hasil Plotnya itu Skewed Distribution/Bukan Normal Distribusi -> Interquartile Range (IQR)
12. Melihat missing value
13. **Data Preprocessing**
14. **Data Cleansing :**
15. Mengisi nilai kosong NAN pada kelompok variabel numerik dengan mean/median/modus

Mean/Median cocok tipe data numerik

Modus cocok untuk tipe data kategorikal

Mengisi nilai kosong dengan mean pada setiap class label

1. Cek Kardinalitas setiap variable menggunakan method .nunique(), total variabel nilai unik yg tertinggi harus di hapus variabelnya/ tidak digunakan atau bisa di encoding untuk tipe data date dimana diubah ke kolom baru dan di gabung dengan dataframe
2. Menangani Data kategorikal dengan label encoder
3. Mengisi nilai kosong pada kelompok variabel kategorikal
4. Menangani Outlier : https://www.kaggle.com/code/aimack/how-to-handle-outliers / https://wellsr.com/python/outlier-data-handling-with-python/
   * Trimming adalah metode memotong atau menghilangkan outlier pada akhir dataset. Ini lebih mudah daripada semua metode lainnya. : https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/
   * Capping adalah pengaturan batas untuk fitur dan mengatur nilai semua outlier yang melebihi batas ke nilai batas. Jadi dalam contoh siswa, kami akan menetapkan batas skor dan mengubah skor siswa 2% menjadi itu. Misalnya, 75 adalah batas skor maksimal yang kita tetapkan. Skor siswa outlier 2% akan ditetapkan menjadi 75.
   * Metode persentil adalah persentil yang sama di kedua sisi untuk mendeteksi outlier. Setelah Anda mendeteksi outlier ini, Anda dapat menggunakan pemangkasan atau pembatasan untuk menghilangkannya.
5. **Data Split Train Test :**
6. **Scalling Data**
7. **Data Classification**
8. **Cross Validation (Teknik Evaluasi Kinerja Model)**

Cross validation adalah suatu metode tambahan dari teknik data mining yang bertujuan untuk memperoleh hasil akurasi yang maksimal. Metode ini sering juga disebut dengan k-fold cross validation dimana percobaan sebanyak k kali untuk satu model dengan parameter yang sama.

Setelah kita selesai melatih model kita, kita tidak bisa berasumsi bahwa itu akan bekerja dengan baik pada data yang belum pernah dilihat sebelumnya. Dengan kata lain, kami tidak dapat memastikan bahwa model tersebut akan memiliki akurasi dan variasi yang diinginkan dalam lingkungan produksi. Kami membutuhkan semacam jaminan keakuratan prediksi yang dibuat oleh model kami. Untuk ini, kita perlu memvalidasi model kita. Proses memutuskan apakah hasil numerik yang mengkuantifikasi hubungan yang dihipotesiskan antara variabel, dapat diterima sebagai deskripsi data, dikenal sebagai validasi.

Untuk mengevaluasi kinerja model pembelajaran mesin apa pun, kami perlu mengujinya pada beberapa data yang tidak terlihat. Berdasarkan kinerja model pada data yang tidak terlihat, kami dapat mengatakan bahwa cuaca model kami Under-fitting/Over-fitting/Well generalized. Validasi silang (CV) adalah salah satu teknik yang digunakan untuk menguji keefektifan model pembelajaran mesin, juga merupakan prosedur pengambilan sampel ulang yang digunakan untuk mengevaluasi model jika kita memiliki data yang terbatas.Untuk melakukan CV kita perlu menyisihkan sampel/bagian dari data yang tidak digunakan untuk melatih model, nanti gunakan sampel ini untuk pengujian/validasi.

1. **Confusion Matrix**
2. **Model Evaluation : Data Overfitting/Underfitting/ROC**

**Overfitting,** Overfitting adalah suatu kedaan dimana data yang digunakan pada data training memiliki prediksi yang terlalu baik, namun prediksinya buruk pada data testing. Ketika sebuah model Overfit, model tidak dapat melakukan generalisasi dengan baik sehingga apabila dilakukan tes dengan menggunakan data yang berbeda dapat mengurangi akurasi (hasil yang dibuat tidak sesuai yang diharapkan).

**Underfittin,** Underfitting adalah keadaan dimana model pelatihan data yang dibuat tidak mewakilkan keseluruhan data yang akan digunakan nantinya. Underfit dapat terjadi ketika model terlalu sederhana dan tidak mampu untuk menyesuaikan pola yang terdapat pada data training, sehingga menghasilkan peforma yang buruk dalam data training.

1. **Data Testing**
2. **Analisa Hasil**