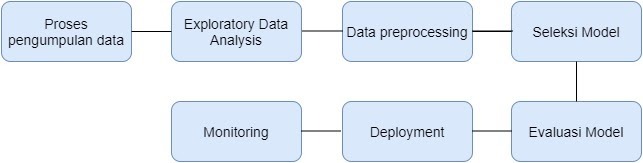
Dalam sebuah project machine learning ada tahapan-tahapan yang perlu dilalui sebelum project tersebut bisa diimplementasi di tahap produksi. Tahapan-tahapan yang dimaksud menurut buku *Hands on Machine Learning* karya Geron [4] dapat dirangkum dalam diagram berikut.

[](https://www.dicoding.com/academies/184/tutorials/8472)

Tahapan dalam diagram ini bersifat iteratif yang berarti prosesnya bisa berulang sesuai kebutuhan. Anda mungkin perlu untuk mengevaluasi ulang proses yang Anda jalankan dan kembali ke langkah sebelumnya, kapan saja dibutuhkan selama prosesnya.

**Proses pengumpulan data.**

Pada modul Data Introduction Anda telah mengetahui beberapa sumber dataset untuk model machine learning. Dari sumber-sumber tersebut, Anda dapat memilih, mengunduh, dan menggunakan dataset yang sesuai dengan kebutuhan Anda kapan saja. Proses ini relatif mudah. Tantangannya adalah memilih dataset yang tepat untuk model Anda.

Tetapi jika Anda adalah seorang Machine Learning Engineer pada sebuah perusahaan yang bertugas untuk membangun model ML untuk keperluan perusahaan, tentu proses pengumpulan datanya tidak semudah Anda mengunduh dataset yang sudah jadi. Anda perlu mengumpulkan dan mengekstrak sendiri data dari berbagai sumber misalnya database, file, data sensor, dan sumber lainnya. Pada tahap ini Anda juga perlu berurusan dengan berbagai jenis tipe data dari mulai structured data (seperti excel file atau database SQL), sampai unstructured data (seperti text file, email, video, audio, gambar, data sensor, dan lainnya). Menurut [Gartner](https://analyticsengines.com/resources/news-events/large-language-models-unstructured-datas-new-best-friend/) lebih dari 80% data perusahaan adalah unstructured data.

**Exploratory Data Analysis**

Exploratory data analysis atau EDA bertujuan sebagai analisa awal terhadap data dan melihat bagaimana kualitas data untuk meminimalkan potensi kesalahan di kemudian hari. Pada proses ini dilakukan investigasi awal pada data untuk menemukan pola, anomali, menguji hipotesis , memahami distribusi, frekuensi, hubungan antar variabel, dan memeriksa asumsi dengan teknik statistik dan representasi grafik.

Pada umumnya EDA dilakukan dengan dua cara, yaitu u*nivariate analysis,*dan *multivariate analysis.* Univariate analysis adalah analisis deskriptif yang memeriksa pola dengan satu variabel pada modelnya. Multivariate analysis merupakan analisis deskriptif yang memeriksa pola dalam data multidimensi dengan membertimbangkan dua atau lebih variabel. Jika terdapat dua variabel yang akan dianalisis, maka disebut *bivariate analysis.*

Karena multivariate analysis mempertimbangkan lebih banyak variabel, ia dapat memeriksa fenomena yang lebih kompleks dan menemukan pola data yang mewakili dunia nyata dengan lebih akurat. Jika tertarik untuk belajar lebih lanjut mengenai exploratory data analysis, Anda dapat membaca tautan-tautan berikut: [tautan 1](https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf), [tautan 2](https://www.kite.com/blog/python/data-analysis-visualization-python/), [tautan 3](https://datascienceguide.github.io/exploratory-data-analysis), [tautan 4](https://www.youtube.com/watch?v=zHcQPKP6NpM).

**Data preprocessing**

Data preprocessing adalah tahap di mana data diolah lebih lanjut sehingga menjadi siap dipakai dalam pengembangan model ML. Dengan kata lain, proses ini mengubah dan mentransformasi fitur-fitur data ke dalam bentuk yang mudah diinterpretasikan dan diproses oleh algoritma machine learning. Termasuk di dalam data preprocessing adalah proses data cleaning dan data transformation.

Beberapa hal yang bisa dilakukan dalam proses data cleaning adalah: penanganan missing value, data yang tidak konsisten, duplikasi data, ketidakseimbangan data, dll. Sementara beberapa hal yang bisa dilakukan untuk proses transformasi data adalah: *scaling*atau merubah skala data agar sesuai dengan skala tertentu, standarisasi, normalisasi, [mengonversi](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html) data menjadi variabel kategori, dan sebagainya. Tidak lupa pula, proses train-test split yang pernah kita pelajari pada modul sebelumnya juga merupakan bagian dari data preprocessing.

**Model selection**

K. P. Murphy dalam bukunya yang berjudul Machine Learning: a Probabilistic Perspective [16] menuliskan kalimat berikut: “when we have a variety of models of different complexity (e.g., linear or logistic regression models with different degree polynomials, or KNN classifiers with different values of K), how should we pick the right one?”. Berangkat dari pertanyaan ini, menentukan model yang sesuai dengan data merupakan tahapan yang penting dalam machine learning workflow.

Jie Ding, et al dalam tulisannya “Model Selection Techniques -An Overview” [17] menyatakan bahwa tidak ada model yang cocok secara universal untuk data dan tujuan apa pun. Pilihan model atau metode yang tidak tepat dapat menyebabkan kesimpulan yang menyesatkan atau performa prediksi yang mengecewakan. Sebagai contoh, saat kita memiliki kasus klasifikasi biner kita perlu mempertimbangkan apa model terbaik untuk data kita, apakah logistic regression atau SVM classifier.

Setelah kita menentukan metode yang cocok untuk data yang ada, kita perlu mengubah hyperparameter untuk mendapatkan performa terbaik dari model. Mengubah nilai hyperparameter saat kita menjalankan algoritma ML akan memberikan hasil atau performa model yang berbeda. Proses menemukan performa terbaik model dengan pengaturan hyperparameter yang berbeda ini juga disebut model selection.

Dengan demikian, dalam konteks machine learning, model selection bisa berarti dua hal: pemilihan *learning method*atau algoritma ML dan pemilihan *hyperparameter* terbaik untuk metode machine learning yang dipilih.

**Model Evaluation**

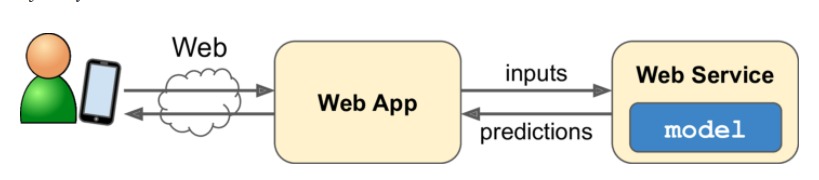
Setelah mengutak-atik model Anda dengan hyperparameter yang berbeda, akhirnya Anda mendapatkan model yang kinerjanya cukup baik. Langkah selanjutnya adalah mengevaluasi model akhir pada data uji. Sederhananya, langkah evaluasi model dapat dijabarkan sebagai berikut: memprediksi label pada data uji, menghitung jumlah prediksi yang salah (eror) kemudian membandingkannya dengan data label yang kita miliki. Dari data perbandingan ini kita dapat menghitung akurasi atau performa model.

Pada prinsipnya proses model evaluation adalah menilai kinerja model ML pada data baru, yaitu data yang belum pernah “dilihat” oleh model sebelumnya. Evaluasi model bertujuan untuk membuat estimasi generalisasi eror dari model yang dipilih, yaitu, seberapa baik kinerja model tersebut pada data baru. Idealnya, model machine learning yang baik adalah model yang tidak hanya bekerja dengan baik pada data training, tapi juga pada data baru. Oleh karena itu, sebelum mengirimkan model ke tahap produksi, Anda harus cukup yakin bahwa performa model akan tetap baik dan tidak menurun saat dihadapkan dengan data baru.

**Deployment**

Ketika model dievaluasi, model siap untuk dipakai pada tahap produksi. Caranya adalah dengan menyimpan model yang telah dilatih dari tahap preprocessing hingga pipeline prediksi. Kemudian deploy model tersebut ke tahap produksi untuk membuat prediksi dengan memanggil kode predict()-nya,

Geron [4] memberikan contoh ilustrasi model deployment seperti tampak dalam gambar berikut.

[](https://www.dicoding.com/academies/184/tutorials/8472)

Misalnya sebuah model regresi untuk menentukan harga rumah akan digunakan dalam situs web. Pengguna akan mengetik beberapa data tentang lokasi yang diinginkan dan mengeklik tombol “Prediksi Harga”. Proses ini akan mengirimkan *query*yang berisi data ke server, kemudian meneruskannya ke aplikasi web Anda. Terakhir, kode akan memanggil fungsi predict() untuk memberikan hasil prediksi pada Anda.

**Monitoring**

Model yang telah dipakai dalam tahap produksi masih harus tetap dimonitor untuk menjaga kualitasnya. Pada tahap produksi model bisa saja menemukan data yang tidak dikenali sehingga performa model dapat menurun. Contoh kasus misalnya, jika model Anda merupakan sistem rekomendasi yang menyarankan produk untuk pengguna, maka untuk memantau performa model bisa dengan cara menghitung jumlah produk rekomendasi yang terjual tiap hari. Jika angka ini turun (dibandingkan dengan produk yang tidak direkomendasikan), maka kemungkinan model kita perlu dilatih ulang menggunakan data-data baru.

Jika Anda bekerja dengan model machine learning yang datanya terus berubah, Anda perlu melakukan update pada dataset dan melatih ulang model Anda secara reguler. Atau, Anda perlu membuat sistem yang dapat membuat proses update ini berjalan secara otomatis.

Bagaimana, semoga cukup jelas ya?

Machine learning melibatkan cukup banyak proses dan infrastruktur dalam membangun modelnya. Memahami algoritma machine learning memang penting, tapi memahami keseluruhan proses machine learning hingga ke tahap produksi juga tak kalah pentingnya. Pada kelas Belajar Pengembangan Machine Learning Anda akan belajar tentang bagaimana men-deploy model machine learning Anda ke tahap produksi. Tetap semangat belajarnya, ya!