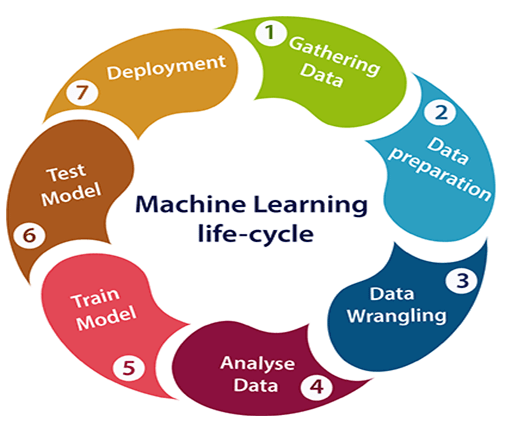
1. **Machine Learning Stages**



Dalam machine learning stages terdapat 7 langkah penting yang dapat dilakukan.

* 1. Gathering data

Pengumpulan data merupakan Langkah awal dalam machine learning. Tujuan dari langkah ini adalah untuk mengidentifikasi dan mendapatkan semua masalah terkait data.

Pada langkah ini, kita perlu mengidentifikasi sumber data yang berbeda, karena data dapat dikumpulkan dari berbagai sumber seperti file, basis data, internet, atau perangkat seluler. Ini adalah salah satu langkah terpenting dari siklus hidup. Kuantitas dan kualitas data yang dikumpulkan akan menentukan efisiensi output. Semakin banyak data, semakin akurat prediksi.

Pada langkah ini terdapat beberapa task:

* Mengidentifikasi berbagai sumber data
* Mengumpulkan data
* Mengintegrasikan berbagai data yang didapat dari sumber yang berbeda
  1. Data preparation

Data preparation merupakan langkah untuk menempatkan data yang sudah diperoleh ke tempat yang sesuai dan mempersiapkannya untuk digunakan dalam pelatihan machine learning.

* 1. Data wrangling

Data wrangling merupakan langkah untuk merapikan data mentah yang kita dapatkan ke dalam dataset yang dapat diolah sedemikian rupa untuk menjalankan machine learning.

* 1. Analyse data

Pada langkah ini dilakukan pembangunan model untuk pembelajaran machine learning. Terdapat berbagai jenis pembelajaran data yang dapat dilakukan seperti, klasifikasi, regresi, analisis cluster, dan lainnya.

* 1. Train model

Langkash selanjutnya adalah pelatihan model. Dalam langkah ini kita melatih model yang sudah disiapkan pada langkah-langkah sebelumnya untuk mendapatkan performa yang baik dalam menyelesaikan masalah yang ada. Dataset digunakan sebagai alat untuk pelatihan dengan menggunakan berbagai algoritma machine learning.

* 1. Test model

Setelah kita melatih model dengan berbagai dataset, selanjutnya kita menguji model tersebut. Pada langkah ini kita melihat akurasi yang didapatkan dengan dataset yang ada. Pengujian model menentukan presentase akurasi model sesuai dengaan kebutuhan proyek.

* 1. Deployment

Deployment menrupakan langkah di mana kita merilis model yang sudah diuji sebelumnya. Deployment hanya bisa dilakukan jika pengujian menghasilkan hasil yang akurat sesuai kebutuhan proyek. Tetapi sebelum merilisnya, petlu diperiksa apakah model tersebut meningkatkan kinerja terhadap projek atau tidak.

1. **Machine Learning Model Training**

Dalam machine learning terdapat berbagai teknik untuk training data.

* 1. Super Vector Machine

Support Vector Machine (SVM) merupakan salah satu metode dalam supervised learning yang biasanya digunakan untuk klasifikasi (seperti Support Vector Classification) dan regresi (Support Vector Regression). Dalam pemodelan klasifikasi, SVM memiliki konsep yang lebih matang dan lebih jelas secara matematis dibandingkan dengan teknik-teknik klasifikasi lainnya. SVM juga dapat mengatasi masalah klasifikasi dan regresi dengan linear maupun non-linear.

SVM digunakan untuk mencari hyperplane terbaik dengan memaksimalkan jarak antar kelas. Hyperplane adalah sebuah fungsi yang dapat digunakan untuk pemisah antar kelas. Dalam 2-D fungsi yang digunakan untuk klasifikasi antar kelas disebut sebagai line whereas, fungsi yang digunakan untuk klasifikasi antas kelas dalam 3-D disebut plane similarly, sedangan fungsi yang digunakan untuk klasifikasi di dalam ruang kelas dimensi yang lebih tinggi di sebut hyperplane.

* 1. K-nearest neighbor (KNN)

K-nearest neighbors atau knn adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (train data sets), yang diambil dari k tetangga terdekatnya (nearest neighbors). Dengan k merupakan banyaknya tetangga terdekat.

K-nearest neighbors melakukan klasifikasi dengan proyeksi data pembelajaran pada ruang berdimensi banyak. Ruang ini dibagi menjadi bagian-bagian yang merepresentasikan kriteria data pembelajaran. Setiap data pembelajaran direpresentasikan menjadi titik-titik c pada ruang dimensi banyak.

Data baru yang diklasifikasi selanjutnya diproyeksikan pada ruang dimensi banyak yang telah memuat titik-titik c data pembelajaran. Proses klasifikasi dilakukan dengan mencari titik c terdekat dari c-baru (nearest neighbor). Teknik pencarian tetangga terdekat yang umum dilakukan dengan menggunakan formula jarak euclidean. Berikut beberapa formula yang digunakan dalam algoritma knn.

* Euclidean Distance

Jarak Euclidean adalah formula untuk mencari jarak antara 2 titik dalam ruang dua dimensi.



* Hamming Distance

Jarak Hamming adalah cara mencari jarak antar 2 titik yang dihitung dengan panjang vektor biner yang dibentuk oleh dua titik tersebut dalam block kode biner.

* Manhattan Distance

Manhattan Distance atau Taxicab Geometri adalah formula untuk mencari jarak d antar 2 vektor p,q pada ruang dimensi n.

* Minkowski Distance

Minkowski distance adalah formula pengukuran antar 2 titik pada ruang vektor normal yang merupakan hibridisasi yang mengeneralisasi euclidean distance dan mahattan distance.

Teknik pencarian tetangga terdekat disesuaikan dengan dimensi data, proyeksi, dan kemudahan implementasi oleh pengguna.

Untuk menggunakan algoritma k nearest neighbors, perlu ditentukan banyaknya k tetangga terdekat yang digunakan untuk melakukan klasifikasi data baru. Banyaknya k, sebaiknya merupakan angka ganjil, misalnya k = 1, 2, 3, dan seterusnya. Penentuan nilai k dipertimbangkan berdasarkan banyaknya data yang ada dan ukuran dimensi yang dibentuk oleh data. Semakin banyak data yang ada, angka k yang dipilih sebaiknya semakin rendah. Namun, semakin besar ukuran dimensi data, angka k yang dipilih sebaiknya semakin tinggi.

* 1. K-means clustering

K-means Clustering adalah salah satu “unsupervised machine learning algorithms” yang paling sederhana dan populer .Tujuan dari algoritma ini adalah untuk menemukan grup dalam data, dengan jumlah grup yang diwakili oleh variabel K. Variabel K sendiri adalah jumlah cluster yang kita inginkan.

Untuk memproses data algoritma K-means Clustering , data dimulai dengan kelompok pertama centroid yang dipilih secara acak, yang digunakan sebagai titik awal untuk setiap cluster, dan kemudian melakukan perhitungan berulang (berulang) untuk mengoptimalkan posisi centroid.

Proses ini berhenti atau telah selesai dalam mengoptimalkan cluster ketika:

* Centroid telah stabil — tidak ada perubahan dalam nilai-nilai mereka karena pengelompokan telah berhasil.
* Jumlah iterasi yang ditentukan telah tercapai.

Hasil dari K-Mean Clustering adalah:

Centroid dari cluster K, yang dapat digunakan untuk memberi label data baru

Label untuk data pelatihan (setiap titik data ditugaskan ke satu clusters)

1. **Evaluation in Machine Learning**

Dalam machine learning terdapat berbagai teknik evaluasi.

* 1. Confusion matrix

Confusion matrix juga sering disebut error matrix. Pada dasarnya confusion matrix memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi sebenarnya. Confusion matrix berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. Terdapat 4 istilah sebagai representasi hasil proses klasifikasi pada confusion matrix. Keempat istilah tersebut adalah True Positive (TP), True Negative (TN), False Positive (FP) dan False Negative (FN). Hal ini sangat berguna karena hasil dari klasifikasi umumnya tidak dapat diekspresikan dengan baik dalam satu angka saja.

Berikut adalah beberapa manfaat dari confusion matrix:

1. Menunjukkan bagaimana model ketika membuat prediksi.
2. Tidak hanya memberi informasi tentang kesalahan yang dibuat oleh model tetapi juga jenis kesalahan yang dibuat.
3. Setiap kolom dari confusion matrix merepresentasikan instance dari kelas prediksi.
4. Setiap baris dari confusion matrix mewakili instance dari kelas aktual.
   1. Area under ROC

Area di bawah ROC Curve adalah metrik kinerja untuk mengukur kemampuan classifier biner untuk membedakan antara kelas positif dan negatif.

* 1. Regression matrix

Dalam regression matrix terdapat dua hal penting yang menjadi pengukuran. Mean Absolute Error (atau MAE) adalah jumlah dari perbedaan absolut antara prediksi dan nilai aktual. Di sisi lain, Root Mean Squared Error (RMSE) mengukur besarnya rata-rata kesalahan dengan mengambil akar kuadrat dari rata-rata perbedaan kuadrat antara prediksi dan pengamatan aktual.

* 1. Cross validation

Cross-validation adalah teknik yang melibatkan mempartisi dataset observasi asli ke dalam set pelatihan, digunakan untuk melatih model, dan set independen yang digunakan untuk mengevaluasi analisis.

Teknik cross validation yang paling umum adalah k-fold cross-validation, di mana dataset asli dipartisi menjadi k dengan ukuran sampel yang sama, yang disebut lipatan. K adalah angka yang ditentukan pengguna, biasanya dengan 5 atau 10 sebagai nilai pilihannya. Ini diulang k kali, sehingga setiap kali, salah satu himpunan bagian k digunakan sebagai set uji / set validasi dan himpunan k-1 lainnya disatukan untuk membentuk set pelatihan. Estimasi kesalahan dirata-rata pada semua percobaan k untuk mendapatkan efektivitas total dari model.