BAB II

LANDASAN TEORI

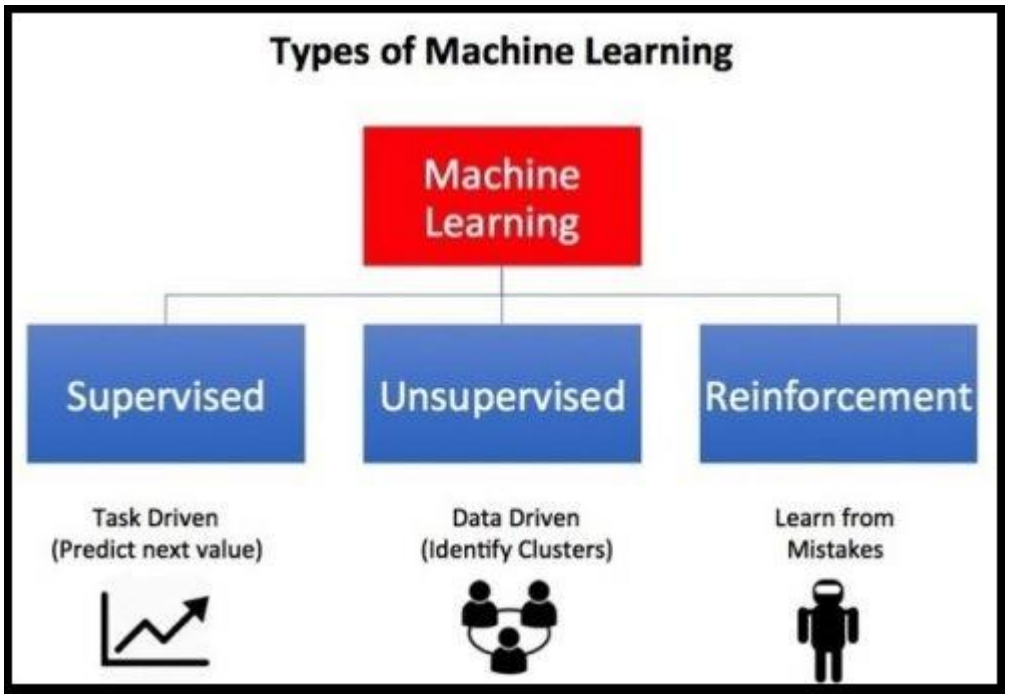
2.1 Penjurusan/Prodi

Prodi atau Program Studi merupakan kesatuan rencana belajar yang digunakan sebagai pedoman jalannya pendidikan akademik yang penyelenggaraannya berdasarkan suatu kurikulum. Adanya prodi bertujuan supaya mahasiswa bisa menguasai suatu pengetahuan, keterampilan, dan sikap sesuai dengan target kurikulum pendidikan yang digunakan. Biasanya anda akan diminta memilih Prodi ketika sudah di semester 3 atau 4, tergantung dari kebijakan universitas masing-masing.

2.2 Pembelajaran Mesin

Pembelajaran mesin adalah disiplin ilmu yang mencakup perancangan dan pengembangan algoritma yang memungkinkan komputer untuk mengembangkan perilaku yang didasarkan pada data empiris, seperti dari sensor data basis data**.** Sistem pembelajar dapat memanfaatkan contoh (data) untuk menangkap ciri yang diperlukan dari probabilitas yang mendasarinya (yang tidak diketahui) (Politan, 2016). Data dapat dilihat sebagai contoh yang menggambarkan hubungan antara variabel yang diamati. Fokus besar penelitian pemelajaran mesin adalah bagaimana mengenali secara otomatis pola kompleks dan membuat keputusan cerdas berdasarkan data.

2.2.1 Tipe Pembelajaran Mesin



Gambar 2. 1 Tipe Pembelajaran Mesin

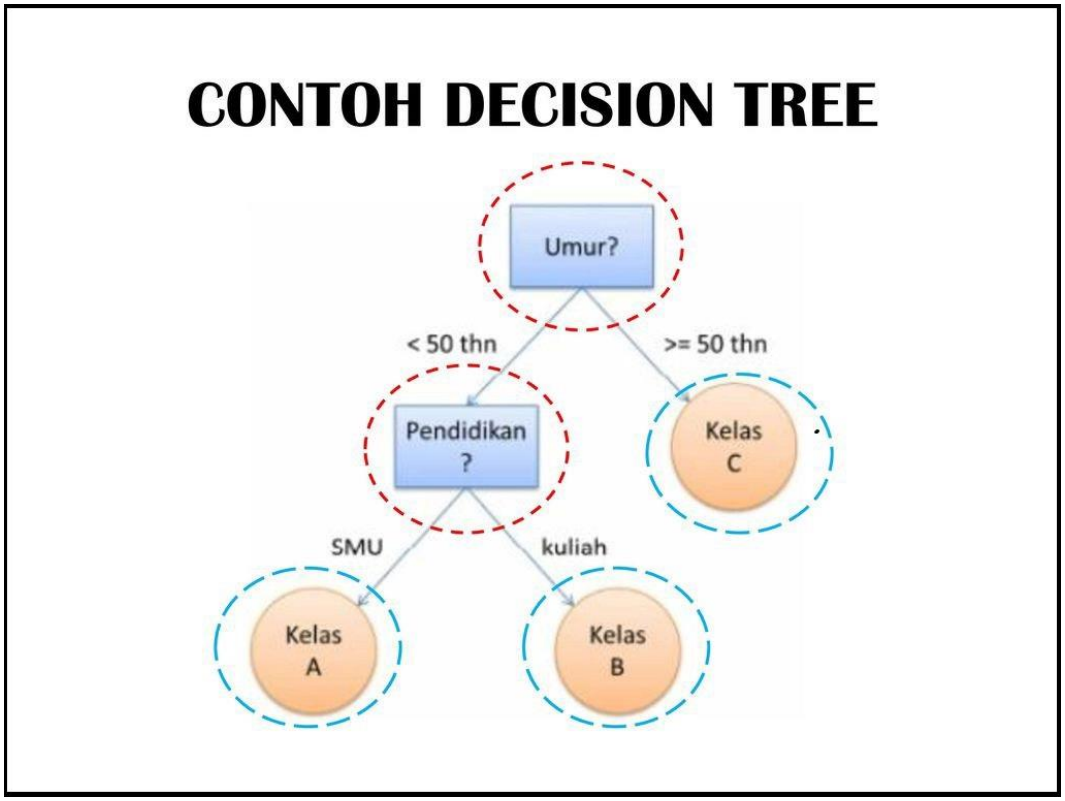
Pada pembelajaran mesin terdapat 3 pembagian jenis pembelajaran yaitu (Hendrikarisma, 2013):

a. *Supervised Learning* memiliki karakteristik masalah yang diselesaikan biasanya berupa klasifikasi. Biasanya selain memiliki atribut untuk setiap *instance*, namun juga sudah memiliki kelas yang jelas, sehingga *task* selanjutnya dari hipotesis atau model yang ditemukan adalah melakukan klasifikasi terhadap *instance* yang baru dan belum memiliki label (belum diklasifikasi).
b. *Unsupervised Learning* biasanya memiliki kata kunci *clustering* atau melakukan pengklusteran terhadap sekelompok data atau sekelompok *instances* yang tidak memiliki label, sehingga memiliki informasi bahwa terdapat sekumpulan data yang membentuk cluster, namun kita belum tahu apa pengetahuan atau hipotesis yang membuat *instances* tersebut saling berkumpul (membuat kelompok) menjadi satu cluster atau lebih.
c. *Reinforcement Learning* biasanya berupa permasalah yang membutuhkan aktifitas eksplorasi, sehingga cukup sesusai jika digunakan untuk membangun suatu intelijen pada suatau game (terutama *puzzle*).

2.3 Decision Tree Learning

Decision tree learning (DTL) adalah salah satu pendekatan pemodelan prediktif yang digunakan dalam statistik, *data mining*, dan pembelajaran mesin ini menggunakan pohon keputusan (sebagai model prediksi) untuk beralih dari pengamatan tentang suatu barang (diwakili di cabang-cabang) ke kesimpulan tentang nilai target barang (diwakili dalam dedaunan). Model pohon di mana variabel target dapat mengambil nilai diskrit disebut pohon klasifikasi. Dalam struktur pohon ini, daun

mewakili label kelas dan cabang mewakili konjungsi fitur yang mengarah ke label kelas tersebut. Pohon keputusan tempat variabel target dapat mengambil nilai kontinu (biasanya bilangan real) disebut pohon regresi (Iykra, 2018).



Gambar 2. 2 Decison Tree

Pada gambar 2.2 memiliki root yaitu umur yang memiliki leaf pendidikan dan kelas. Jika umur lebih dari sama dengan 50 tahun maka akan masuk ke kelas C, tetapi jika umur kurang dari 50 tahun akan masuk ke pendidikan. Pendidikan pada *decision tree* ini adalah node induk yang memiliki node kelas A dan kelas B, jika SMU maka akan masuk ke node kelas A dan kalau kuliah akan masuk node kelas B.

2.4 Random Forest

Algoritma *Random Forest* (RF) merupakan pengembangan dari metode *Classification and Regression Tree* (CART) dengan menerapkan metode bootstrap aggregating (bagging) dan random feature selection (Breiman 2001). Metode ini merupakan metode pohon gabungan (ensemble tree). Dalam RF, banyak pohon ditumbuhkan sehingga terbentuk suatu hutan (forest), kemudian analisis dilakukan pada kumpulan pohon tersebut. Penggunaan bagging pada RF berguna dalam mengatasi sifat ketidakstabilan dari metode klasifikasi tunggal.

Pada RF pembentukan tree dilakukan dengan cara melakukan training sampel data. Cara yang digunakan untuk mengambil sampel data adalah dengan *sampling with replacement*. Variabel yang digunakan sebagai split dipilih secara acak. Proses klasifikasi dilakukan setelah semua tree terbentuk dan penentuan hasil klasifikasi diambil berdasarkan vote dari masing-masing tree. Vote terbanyak yang akan menjadi pemenangnya (Meliana, 2016).

Berikut ini adalah prosedur atau algoritma untuk membangun Random Forest pada gugus data yang terdiri dari n amatan dan p peubah penjelas (Breiman, 2001; Breiman dan Cutler, 2003):

1. Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data. Langkah ini dinamakan dengan bootstrap (bag).
2. Dengan menggunakan contoh bootstrap, pohon dibangun sampai mencapai ukuran maksimum yaitu tanpa pemangkasan (pruning). Pembangunan pohon dilakukan dengan menerapkan random feature selection yaitu m peubah penjelas dipilih secara acak dengan $m << p$, selanjutnya pemilah terbaik dipilih berdasarkan m peubah penjelas.

Langkah 1 dan 2 diulangi sebanyak k kali untuk membuat sebuah forest yang terdiri dari k pohon.

Tahapan pembuatan model klasifikasi menggunakan algoritma *Random Forest* dilakukan setelah membuat pemodelan data latih menggunakan package *Random Forest* salah satunya di R.

Pada Random Forest pembentukan tree dilakukan dengan menerapkan motode CART dimana split yang akan dijadikan root dihitung dengan menggunakan nilai Gini Index. Gini index memiliki persamaan sebagai berikut (Han et al. 2012) :

$$Gini\ (S) = 1 - \sum_{i=1}^{k} {P_i}^2 \quad \ldots (1)$$

$$Gini\ Gain\ (S) = Gini\ (S) - Gini\ (A,S) = Gini\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|}\ Gini\ (S_i) \quad \ldots (2)$$

dengan Pi adalah probabilitas dari S milik kelas i, sedangkan Si adalah partisi dari dataset S yang memiliki atribut A.

Metode CART menghasilkan suatu pohon klasifikasi jika peubah responnya kategorik, dan menghasilkan pohon regresi jika peubah responnya kontinu. Untuk peubah kontinu x_j penyekatan yang diperbolehkan adalah $x_j \leq c$ dan $x_j \geq c$ dimana c adalah nilai tengah antara dua nilai amatan peubah x_j secara berurutan (Breiman et al. 1993).