BAB 2

TINJAUAN PUSTAKA

2.1 Machine Learning

Machine Learning merupakan aplikasi atau bagian dari kecerdasan buatan (*Artificial Inteligence*) yang membuat sistem memiliki kemampuan belajar secara otomatis dan meningkatkan kemampuannya berdasarkan pengalaman tanpa diprogram secara eksplisit. Dalam hal ini, program komputer tidak ditulis secara statis.

Menurut definisi yang lain, machine learning adalah ilmu yang memungkinkan komputer berperilaku seperti manusia, dimana komputer dapat meningkatkan pemahamannya melalui pengalaman atau dengan berjalannya waktu secara otomatis (Fraggella, 2020). Kemampuan belajar tersebut diperoleh dengan mempelajari sejumlah data (learn from data) sehingga dapat menghasilkan suatu model untuk melakukan proses inputoutput tanpa menggunakan kode program yang dibuat secara eksplisit.

Pembelajaran mesin berfokus pada pengembangan program computer yang bisa mengakses data dan menggunakannya untuk belajar sendiri. Proses pembelajaran dimulai dengan observasi data seperti berdasarkan pada pengalaman langsung atau intruksi atau mencari pola data dan membuat keputusan yang lebih baik dimasa depan berdasarkan contoh tersebut. Tujuan utamanya adalah membiarkan komputer belajar secara otomatis tanpa intervensi atau bantuan manusia dan menyesuaikan aktifitas yang sesuai. Karena teknologi komputasi, machine learning saat ini tidak seperti machine learning di masa lalu.

Proses belajar tersebut menggunakan algoritma khusus yang disebut machine learning algorithms. Terdapat banyak algoritma machine learning dengan efisiensi dan spesifikasi kasus yang berbeda-beda. Secara garis besar terdapat 2 teknik dalam machine learning, yaitu belajar dengan pengawasan (*supervised learning*) dan belajar

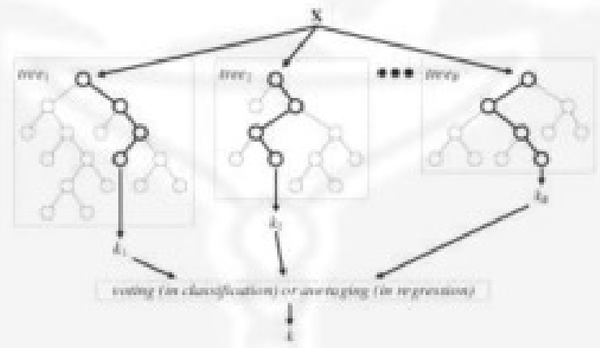
tanpa pengawasan (*unsupervised learning*). *Supervised learning* adalah sebuah pendekatan dimana sudah terdapat data yang dilatih, dan terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengelompokan suatu data ke data yang sudah ada. Lain halnya dengan unsupervised learning, unsurpervised learning tidak memiliki data latih, sehingga dari data yang ada, kita mengelompokan data tersebut menjadi 2 bagian atau 3 bagian dan seterusnya.

Pada penulisan ini, teknik yang dipakai adalah supervised learning dengan data latih yang sudah tersedia. Algoritma pembelajar mesin yang digunakan adalah pohon keputusan dan regresi yang merupakan bagian dari supervised learning atau belajar dengan pengawasan. Permasalahan yang dapat diselesaikan dengan menggunakan pembelajaran mesin dengan pengawasan (supervised machine learning) dapat selanjutnya dikelompokkan kedalam dua kategori masalah yaitu Classification dan Regression.

Classification adalah suatu permasalahan dimana variable keluarannya adalah berupa kategori ("biru", "hitam", "sakit", atau "tidak sakit") sementara Regression adalah suatu permasalahan dimana variabel keluaarannya adalah berupa nilai riil ("dolar", "berat", atau "harga tiket") dimana beberpa masalah terbentuk dari kombinasi Classification maupun Regression, seperti masalah rekomendasi dan prediksi terhadap data deret waktu (time series predicition) secara umum (Jason Brownlee, 2016). Beberapa contoh algoritma pembelajaran mesin dengan pengawasan adalah: (1) Linear Regression untuk masalah regresi; (2) Random Forest untuk masalah klasifikasi dan regresi; (3) Support Vector Machine untuk masalah klasifikasi. Pada tulisan ini, yang digunakan adalah Random Forest untuk mengklasifikasi untuk mendeteksi apakah seseorang mengidap penyakit jantung atau tidak .

2.2 Random Forest Classifier

Random forest merupakan salah satu metode yang digunakan untuk klasifikasi dan regresi. Metode ini merupakan sebuah *ensemble* (kumpulan) metode pembelajaran menggunakan pohon keputusan sebagai *base classifier* yang dibangun dan dikombinasikan. *Random forest* adalah pengklasifikasi yang terdiri dari kumpulan pengklasifikasi pohon terstruktur dimana masingmasing pohon melemparkan unit suara untuk kelas paling popular di input x (Breiman, 2001). *Random forest* merupakan metode klasifikasi yang *supervised*. Sesuai dengan namanya, metode ini menciptakan sebuah hutan (*forest)* dengan sejumlah pohon (*tree*). Secara umum, semakin banyak pohon (*tree*) pada sebuah hutan (*forest*) maka semakin kuat juga hutan tersebut terlihat. Banyaknya pohon yang akan dibentuk sangat berpengaruh terhadap tingkat akurasi hasil klasifikasi. Semakin banyak pohon, semakin akurat hasil klasifikasinya.



Gambar 2.1 Random Forest Classifier

Metode ini sama dengan metode *decision forest* yang akan menggunakan *information gain* dan *gini index* untuk perhitungannya dalam membangun *tree*, perbedaannya random forest akan membangun lebih dari satu *tree*. Masing-masing *tree* dibangun menggunakan set data dengan atribut yang diambil secara acak dari *data training*.

2.3 Confusion Matrix

Confusion matrix merupakan salah satu teknik yang dapat digunakan untuk mengukur kinerja suatu model khususnya kasus klasifikasi (*supervised learning*) pada *machine learning*. *Confusion matrix* sering disebut dengan *error matrix*. Pada dasarnya *confusion matric* memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi sebenarnya. Confusion matrix berbentuk table matrix yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. Gambar berikut merupakan confusion matrix dengan 4 kombinasi nilai prediksi dan nilai aktual yang berbeda.

Tabel 2.1 Confusion Matrix

	Predicted Positive Class	*Predicted Negative Class*
Actual Positive Class	*TP (True Positive)*	*FN (False Negative)*
Actual Positive Class	*FP (False Positive)*	*TN (True Negative)*

Terdapat 4 istilah sebagai representasi hasil proses klasifikasi confusion matrix, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Supaya dapat lebih dipahami, penulis mengimplementasikan dengan contoh kasus sederhana untuk memprediksi seorang pasien menderita penyakit jantung atau tidak.

- *True Positive* (TP) : Jumlah observasi positif yang tepat prediksi.
- *True Negative* (TN) : Jumlah observasi negatif yang tepat prediksi
- *False Positive* (FP) : Jumlah observasi positif yang salah diprediksi sebagai negatif.
- *False Negative* (FN) : Jumlah observasi negative yang salah diprediksi sebagai positif.

Teknik ini sangat berguna karena *confusion matrix* akan memberi tahu seberapa baik model yang kita buat. Secara khusus *confusion matrix* juga memberikan informasi tentang TP, FP, TN, dan 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. Setiap kolom dari *confusion matrix* merepresentasikan *instance* dari kelas prediksi.
3. Setiap baris dari *confusion matrix* mewakili *instance* dari kelas actual.
4. Tidak hanya memberi informasi tentang kesalahan yang dibuat oleh model tetapi juga jenis kesalahan yang dibuat.

2.2.1 *Accuracy*

Accuracy menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. Maka, accuracy merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Dengan kata lain, accuracy merupakan tingkat kedekatan nilai prediksi dengan nilai aktual (sebenarnya). Nilai accuracy dapat diperoleh dengan persamaan berikut.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.1)$$

2.2.2 *Precision* (*Positive Predictive Value*)

Precision menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Maka, precision merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi postif. Dari semua kelas positif yang telah di prediksi dengan benar, berapa banyak data yang benar-benar positif. Nilai precision dapat diperolah dengan persamaan berikut.