**CASE WEEK 1 – MACHINE LEARNING (THEORY) ANSWERS**

**DODY HARIANTO**

**NO 1**

Salah satu case machine learning adalah memprediksi apakah seseorang akan lulus ujian atau tidak. Untuk bisa memprediksi, diperlukan fitur-fitur yang mendukung seperti apakah seseorang ada les atau tidak, pendidikan orang tua, durasi belajar, dll.

**NO 2**

* Supervised Learning

Tipe learning yang membuat model untuk memprediksi data yang sudah memiliki label menggunakan fitur-fitur yang tersedia.

Contoh: Memprediksi apakah email termasuk spam atau tidak.

Dalam kasus ini, target variable sudah memiliki label (kemungkinan prediksi yang kita sudah ketahui), yaitu spam atau bukan spam.

Cabang: Classification, Regression

* Unsupervised Learning

Tipe learning yang menggunakan data yang tidak memiliki label untuk menemukan insights. Data bisa tidak memiliki label dikarenakan

Contoh: Memprediksi customer segmentation.

Karena tidak memiliki target variable, kita biasanya mencari pola-pola dalam dataset untuk bisa mengelompokkan customer ke dalam beberapa cluster berdasarkan purchases yang sudah dilakukan customer.

Cabang: Clustering, Anomaly Detection

* Reinforcement Learning

Tipe learning yang menggunakan sebuah agen yang belajar dari pengalaman dan mengambil keputusan.

Contoh: Menentukan langkah selanjutnya dalam permainan catur.

**NO 3**

Validation adalah proses untuk melakukan hyperparameter tuning yang bertujuan untuk memilih model yang paling bagus. Sedangkan testing adalah proses untuk menguji model terhadap data baru. Secara default untuk validation dan testing dataset dibagi dengan rasio 80%-20%.

**NO 4**

* Classification

Sebuah teknik supervised learning untuk memprediksi data yang bersifat kategorik dengan cara mengelompokkan hasil prediksi ke salah satu grup.

* Regression

Sebuah teknik supervised learning untuk memprediksi data yang numerik. Hasil dari prediksinya berupa nilai / angka.

**NO 5**

Overfitting bisa terjadi apabila trainin tidak bisa memprediksi data test karena terlalu hebat memprediksi data yang digunakan untuk training. Sedangkan underfitting bisa terjadi apabila model yang digunakan kurang bisa memprediksi data baru yang belum pernah dilihat sebelumnya, biasanya karena kekurangan fitur relevan sehingga prediksinya akan bias.

**NO 6**

Sebuah cara untuk mengecek performa model machine learning jika dataset nya kecil. Teknik ini sering dipakai karena bisa membantu mengurangi kemungkinan terjadinya overfitting atau underfitting, serta untuk menentukan hyperparameter.

**NO 7**

Discrete random variable adalah sebuah data yang countable, sebagai contoh jumlah anak dalam sebuah keluarga. Sedangkan continuous random variable adalah variable yang mana jika ada 2 nilai, maka terdapat tak terhingga nilai yang bisa diambil di antara 2 nilai tersebut, contohnya harga rumah.

**NO 8**

* Normal Distribution

Normal distribution adalah salah satu distribusi yang merepresentasikan banyak data. Ciri dari normal distribution adalah bentuknya yang menyerupai bell curve, yang artinya persebaran data tidak banyak tertumpuk pada salah satu sisi kiri atau kanan, melainkan di tengah. Jika sebuah data terdistribusi normal, maka nilai mean dan median data tersebut dekat atau hampir sama.

Contoh: Distribusi tinggi badan, IQ, atau catatan waktu pelari di Olimpiade.

* Poisson Distribution

Distribusi yang menggambarkan kejadian yang terjadi setiap beberapa saat / waktu tertentu.

Contoh: Jumlah pembeli setiap toko per jam.

* Binomial Distribution

Binomial Distribution adalah distribusi kejadian yang hanya memiliki dua kemungkinan.

Contoh: Distribusi pelemparan koin, distribusi jenis kelamin anak-anak yang baru lahir.

**NO 9**

Mean adalah sebuah metode statistik untuk menghitung measure of center atau ukuran pemusatan sebuah data. Sedangkan variance merupakan sebuah metode statistik untuk menghitung measure of variability. Yang membedakan mean dan variance adalah tujuan perhitungannya. Nilai mean memberikan informasi tentang rata-rata sebuah data. Sedangkan nilai variance adalah ukuran persebaran setiap data terhadap nilai rata-rata.

**NO 10**

Dalam parametric modelling, jumlah parameter akan tetap terhadap ukuran data sample. Parametric model bisa memprediksi data dengan hanya menggunakan parameter yang ada. Sedangkan dalam non parametric modelling, jumlah parameter bisa berubah seiring dengan penambahan atau perubahan ukuran data. Biasanya non parametric model memerlukan dataset yang lebih besar.

**NO 11**

Loss function adalah sebuah function untuk menghitung seberapa jauh nilai yang kita prediksi dengan value aslinya. Jika dalam Linear Regression, biasanya dilihat dari jarak vertikal sebuah titik dengan garis yang terbentuk. Loss function sangat penting karena dapat memberikan informasi tentang seberapa mampu model kita memprediksi terhadap data aslinya.

**NO 12**

Regularization adalah sebuah cara untuk mem-penalize koefisien yang besar dari fitur-fitur. Cara kerja regularization adalah dengan mengurangi variance / ragam data tanpa meningkatkan bias, sehingga penting digunakan untuk menghindari overfitting.