Nama/NIM: Vito Ghifari/13520153

1. Jelaskan yang dimaksud dengan supervised learning dan cakupannya!
2. Jelaskan cara kerja algoritma yang telah diimplementasikan!
3. Bandingkan ketiga algoritma tersebut, lalu tuliskan kelebihan dan kelemahannya!
4. Jelaskan penerapan dari algoritma supervised di berbagai bidang (misalnya industri atau kesehatan)!

Jawaban:

1. Supervised learning merupakan salah satu jenis kategori dari pembelajaran mesin/*machine learning* yang membutuhkan data dan hasil/*output* dari data tersebut. Supervised learning akan mendapatkan pola pada data yang berkorelasi pada hasil sehingga terbentuk aturan untuk menerjemahkan data menjadi hasil. Cakupan dari supervised learning ada dua, yaitu untuk *task* klasifikasi dan *task* regresi.
2. Cara kerja algoritma:
   i. Logistic Regression
   Pada tahap training, X sebagai data train untuk model ini distandardisasi terlebih dahulu. Setelah itu, X akan menghasilkan z dengan fungsi $z = wx + b$, dengan *w* dan *b* adalah weight dan bias. Z dimasukkan ke dalam fungsi sigmoid $(\sigma(x) = \frac{1}{1+e^{-z}})$ untuk mendapatkan probabilitas untuk label pada setiap data X. Hasil dari prediksi y, atau y_hat, dimasukkan ke rumus gradien untuk logistic regression, yaitu $grad = \frac{1}{m}\sum_{i=1}^{m} x * (y_{hat} - y)$. Weight yang baru akan di update dengan rumus $w = w_{last} - \alpha * grad$. Nilai α dapat dipilih sebagai learning rate. Weight akan diperbarui terus-menerus sebanyak num_iter.
   Untuk tahap prediksi, data test akan dimasukkan ke fungsi z, lalu z ke sigmoid sehingga didapatkan probabilitasnya.
   ii. KNN
   Pada KNN, training hanya bertujuan untuk menyimpan *data train*, tidak seperti algoritma logistic regression. Pada prediksi, setiap *data point* akan dibandingkan jaraknya dengan seluruh *data train*, dan didapatkan label berdasarkan mayoritas label dari tetangga terdekat sebanyak K (dapat diatur). Perhitungan jarak pada KNN yang diimplementasikan adalah menggunakan jarak Euclidean.
   iii. Decision Tree ID3
   Algoritma ID3 pada bagian training akan melakukan fit data terhadap tree yang awalnya kosong. Setiap kolom pada data train dihitung information gain-nya terlebih dahulu. Kolom dengan information gain terbesar akan dipilih menjadi node pada tree. Node berikutnya akan didasarkan pada kategori yang terdapat pada kolom tersebut, dan begitu seterusnya secara rekursif hingga kolom yang dipilih adalah *pure class*

(hanya mempunyai 1 kategori). Implementasi tree yang telah dibuat dibantu dengan struktur data dictionary pada Python, dengan kategori pada suatu kolom menjadi key dari dictionary tersebut.

3. Perbandingan:
   i. Logistic Regression:
      - Kelebihan:
        1. Tidak mudah untuk overfit.
        2. Korelasi masing-masing atribut/kolom terhadap hasil dapat diketahui.
      - Kelemahan:
        1. Hanya dapat memisahkan data yang *linearly separable*.
        2. Butuh data yang lebih banyak daripada daripada decision tree.
        3. Interpretability lebih rendah.

   ii. KNN:
      - Kelebihan:
        1. Tidak membutuhkan data yang *linearly separable*.
        2. Interpretability tinggi.
      - Kelemahan:
        1. Lambat jika jumlah data train sangat besar.
        2. Sensitif terhadap outlier.

   iii. Decision Tree:
      - Kelebihan:
        1. Tidak membutuhkan data yang linearly separable.
        2. Interpretability tinggi.
      - Kelemahan:
        1. Sedikit perubahan pada data train dapat membentuk tree yang jauh berbeda.
        2. Mudah overfit.

4. Contoh penggunaan supervised learning adalah pada industri perbankan. Dengan algoritma supervised, transaksi yang merupakan fraud atau penipuan dapat dideteksi. Selain itu, supervised learning digunakan untuk industri yang menghasilkan produk, salah satunya adalah memperkirakan nilai penjualan. Pada kesehatan, supervised learning digunakan salah satunya untuk memprediksi keadaan paru-paru.