Randy Zakya Suchrady
13519061

SUPERVISED LEARNING

a. Apa itu supervised learning?
   Supervised learning merupakan metode pelatihan machine learning menggunakan data yang telah diberi label.
   Contoh : Misal ada sebuah dataset bayi dengan beberapa fitur seperti berat badan, panjang badan, dan lainnya. Untuk fitur dengan nilai tertentu bayi dikategorikan sehat sedangkan untuk nilai fitur lainnya dikatakan bayi tidak sehat. Sehat/tidak sehat merupakan label dari data tersebut.

b. Jelaskan bagaimana cara kerja dan algoritma yang anda implementasikan!
   - K-Nearest Neighbor:
     1. Tentukan fitur apa saja dari dataset yang akan diperhitungkan
     2. Siapkan sebuah data yang ingin diprediksi kelas datanya
     3. Untuk setiap data pada dataset, lakukan perhitungan jarak euclideannya dengan data yang ingin diprediksi ($\sqrt{\sum_{i}^{n}(x_p - x_i)^2}$, xp : fitur data prediksi, xi fitur data pada dataset)
     4. Urutkan dataset sesuai dengan jarak euclideannya dengan data prediksi mulai dari jarak terdekat hingga terjauh
     5. Pilih k data pertama (otomatis k data terdekat) dan periksan kelas dengan jumlah data terbanyak di k data pertama tersebut
     6. Prediksinya adalah kelas yang paling banyak di k data pertama tersebut
   - Logistic Regression:
     1. Tentukan fitur apa saja dari dataset yang akan dilatih
     2. Siapkan nilai alpha (learning rate) dan epoch (jumlah iterasi)
     3. Siapkan nilai vektor **b** yang setiap elemen pada vektor akan menjadi koefisien pada persamaan regresi logistik, inisialisasi dengan **b** = <1,0,0..,0> (berpadanan dengan urutan b0, b1, .., bn)
     4. Lakukan pengulangan sebanyak jumlah epoch, setiap pengulangan lakukan iterasi setiap baris pada dataset
     5. Pada setiap iterasi baris data set, lakukan perhitungan dot product antara fitur pada baris tersebut dengan vektor **b**, akan dihasilkan sebuah nilai z
     6. Lakukan perhitungan z dengan fungsi sigmoid, dimana fungsi sigmoid didefinisikan sebagai $\sigma(z) = \frac{1}{1+e^{-z}}$
     7. Ubah nilai-nilai pada vektor **b** sebagai berikut:
        $$bi = bi + alpha \times (y - \sigma(z)) \times \sigma(z) \times (1 - \sigma(z)) \times xi$$
        bi : elemen ke-i pada vektor **b**
        alpha : learning rate
        y : 1 atau 0, nilai fitur yang akan diprediksi pada baris tersebut
        $\sigma(z)$: prediksi

xi : fitur ke-i pada baris terkait

8. Lakukan perhitungan vektor **b** sampai selesai
9. Akan dihasilkan vektor **b** yang akan menjadi koefisien-koefisien persamaan untuk prediksi

   Persamaan untuk prediksi : $\sigma(z) = \frac{1}{1+e^{-z}}$ dengan z merupakan dot product dari vektor **b** dengan vektor data yang akan diprediksi

- ID3 Decision Tree:
  1. Tentukan fitur apa saja yang akan digunakan dari dataset latihan
  2. Untuk setiap fitur yang ada pada dataset, hitung nilai entropi, average information, dan gainnya

     Entropy = $-\frac{p}{p+n}log_2(\frac{p}{p+n}) - \frac{n}{p+n}log_2(\frac{n}{p+n})$; jika p = 0 atau n = 0, maka entropi = 0
     p : jumlah baris di dataset dengan label positif
     n : jumlah baris di dataset dengan label negatif

     Average Information = $\sum_{i}^{k}\frac{pi+ni}{p+n}Entropy(C = i)$
     i : kelas pada fitur ke-C
     pi : jumlah data yang dilabeli positif untuk data dengan fitur ke-C bernilai i
     ni : jumlah data yang dilabeli negatif untuk data dengan fitur ke-C bernilai i
     p : jumlah seluruh data yang dilabeli positif
     n : jumlah seluruh data yang dilabeli negatif
     k : jumlah fitur
     Entropy(C = i) : perhitungan nilai entropi untuk seluruh data dengan fitur C bernilai i

     Gain = Entropy - Average Information
  3. Pilih fitur/kolom yang memiliki nilai Gain terbesar untuk menjadi akar dari pohon yang akan kita buat
  4. Untuk setiap dataset yang telah dipecah sesuai kelas-kelas yang ada pada kolom terpilih, lakukan langkah 2 kembali dengan menghitung fitur/kolom yang lainnya (siapkan simpul daun untuk akar yang kita miliki sebanyak jumlah nilai kelas pada fitur/kolom terpilih)
  5. Lakukanlah langkah rekursif di atas sampai mencapai kasus dasar sebagai berikut:
     - Seluruh data pada dataset yang diproses memiliki label yang sama semua, maka langsung jadikan label tersebut sebagai value di node/leaf yang sedang dikerjakan
     - Data belum memiliki label yang sama semua, namun fitur yang digunakan untuk pelatihan telah habis, maka pilih label dengan

jumlah terbanyak di dataset pada rekursi tersebut sebagai value pada node/leaf

c. Bandingkan ketiga algoritma tersebut, kemudian tuliskan kelebihan dan kelemahannya!

- K-Nearest Neighbor
  Kelebihan:
  Bisa mengklasifikasikan multi-class data (lebih dari 2)
  - Efektif menghadapi data dengan noise yang besar
  - Semakin baik jika data semakin banyak

  Kekurangan:
  - Keakuratan tergantung sebaran data dan jumlah data yang tersebar di dekat data yang ingin diprediksi, bisa menyebabkan bias pada nilai K tertentu
  - Sulit untuk fitur pada data yang bersifat kategorik
- Logistic Regression
  Kelebihan:
  - Tidak memiliki asumsi normalitas pada variabel independen
  - Variabel independen bisa campuran berupa bentuk kontinu, diskrit maupun dikotomis
  - Tidak membutuhkan keterbatasan dari variabel independennya
  - Variabel bebasnya tidak harus berbentuk interval

  Kekurangan:
  - Hanya dapat melakukan klasifikasi dua kelas
  - Jika ingin melakukan klasifikasi lebih dari satu kelas, dapat disiasati dengan bentuk model one vs the rest, namun akan muncul bias
- ID3 Decision Tree
  Kelebihan:
  - Dapat membuat klasifikasi dengan jumlah fitur/variabel yang banyak menjadi sederhana
  - Memiliki tingkat akurasi yang baik

  Kekurangan:
  - Hanya dapat melakukan klasifikasi dua kelas (positif dan negatif)
  - Kualitas prediksi tergantung dengan bagaimana pohon dirancang
  - Data bersifat kontinu atau diskrit dapat disiasati dengan membaginya ke dalam kelas interval tertentu namun tingkat akurasinya tidak akan begitu baik