PRAKTIKUM DATA MINING

Nama : Hawa Andini Hadi

NIM : 23051214238

Kelas/Angkatan : DTMG1/2023

Algoritma : Stochastic Gradient Descent Classifier (SGDClassifier)

Jenis Analisis : Incremental Learning dan Sampling Method (Sliding Window)

Dataset : Estimation of Obesity Levels Based On Eating Habits and Physical Condition

Keterangan Dataset :

1. Gender: Jenis Kelamin
2. Age: Umur
3. Height: Tinggi
4. Weight: Berat
5. Family_history_with_overweight: Apakah ada anggota keluarga yang menderita atau menderita kelebihan berat badan?
6. FAVC: Apakah Anda sering mengonsumsi makanan berkalori tinggi?
7. FCVC: Apakah Anda biasanya mengonsumsi sayuran dalam makanan Anda?
8. NCP: Berapa banyak makanan utama yang Anda konsumsi setiap hari?
9. CAEC: Apakah Anda makan makanan apa pun di antara waktu makan?
10. SMOKE: Apakah kamu merokok?
11. CH2O: Berapa banyak air yang Anda minum setiap hari?
12. SCC: Apakah Anda memantau kalori yang Anda makan setiap hari?
13. FAF: Seberapa sering Anda melakukan aktivitas fisik?
14. TUE: Berapa banyak waktu yang Anda gunakan pada perangkat teknologi seperti telepon seluler, permainan video, televisi, komputer dan lainnya?
15. CALC: Seberapa sering Anda minum alkohol?
16. MTRANS: Transportasi apa yang biasanya Anda gunakan?
17. NObeyesdad: Tingkat obesitas

Pembahasan : Dataset ini berisi data tentang obesitas berdasarkan berbagai faktor seperti usia, berat badan, tinggi badan, kebiasaan makan, dan pola aktivitas. Variabel target adalah NObeyesdad, yang menunjukkan kategori obesitas seseorang.

Metode Preprocessing :

1. Penghapusan Data Kosong: Menghapus baris data yang memiliki nilai kosong menggunakan dropna().
2. Encoding Kategorikal: Mengubah data kategori menjadi angka dengan LabelEncoder.

3. Standarisasi Fitur Numerik: Mengubah fitur numerik agar memiliki rentang yang seragam menggunakan StandardScaler.

Metode Evaluasi Model :

1. Sampling Method (Sliding Window): Metode ini membagi data menjadi beberapa bagian kecil yang disebut "window" atau jendela. Setiap jendela berisi sebagian data yang digunakan untuk melatih model dan menguji hasilnya. Biasanya, 80% data digunakan untuk pelatihan, dan 20% sisanya untuk pengujian. Setelah setiap jendela, kita mengukur akurasi untuk melihat seberapa baik model bekerja seiring data bergerak maju.
2. Incremental Learning: Setelah model mempelajari data sedikit demi sedikit, kita menguji model dengan 20% data terakhir untuk melihat seberapa baik hasil pelatihan tersebut. Jadi, model dievaluasi setelah seluruh proses pelatihan selesai, dan menggunakan bagian data yang belum digunakan selama pelatihan.

Hasil Evaluasi :

1. Sampling Method (Sliding Window)

```
Window 0-100: Accuracy = 0.40
Window 50-150: Accuracy = 0.35
Window 100-200: Accuracy = 0.60
Window 150-250: Accuracy = 0.45
Window 200-300: Accuracy = 0.45
Window 250-350: Accuracy = 0.55
Window 300-400: Accuracy = 0.45
Window 350-450: Accuracy = 0.30
Window 400-500: Accuracy = 0.65
Window 450-550: Accuracy = 0.65
Window 500-600: Accuracy = 1.00
Window 550-650: Accuracy = 1.00
Window 600-700: Accuracy = 1.00
Window 650-750: Accuracy = 0.75
Window 700-800: Accuracy = 0.70
Window 750-850: Accuracy = 1.00
Window 800-900: Accuracy = 1.00
Window 850-950: Accuracy = 0.80
Window 900-1000: Accuracy = 0.80
Window 950-1050: Accuracy = 0.90
Window 1000-1100: Accuracy = 1.00
Window 1050-1150: Accuracy = 0.80
Window 1100-1200: Accuracy = 0.65
Window 1150-1250: Accuracy = 1.00
Window 1200-1300: Accuracy = 1.00
Window 1250-1350: Accuracy = 1.00
Window 1300-1400: Accuracy = 0.60
Window 1350-1450: Accuracy = 0.80
Window 1400-1500: Accuracy = 1.00
Window 1450-1550: Accuracy = 1.00
```

```
Window 1400-1500: Accuracy = 1.00
Window 1450-1550: Accuracy = 1.00
Window 1500-1600: Accuracy = 1.00
Window 1550-1650: Accuracy = 1.00
Window 1600-1700: Accuracy = 1.00
Window 1650-1750: Accuracy = 1.00
Window 1700-1800: Accuracy = 1.00
Window 1750-1850: Accuracy = 1.00
Window 1800-1900: Accuracy = 1.00
Window 1850-1950: Accuracy = 1.00
Sliding Window Training Completed.
```

Pada awalnya, akurasi model cenderung rendah dan bervariasi, antara 0.30 hingga 0.65. Namun, setelah beberapa jendela (sekitar jendela 500-600), akurasi meningkat secara signifikan dan mencapai nilai 1.00, yang tetap konsisten pada jendela berikutnya. Ini menunjukkan bahwa model membutuhkan waktu untuk beradaptasi, tetapi setelah beberapa iterasi, model mulai memberikan prediksi yang sangat akurat dan stabil.

2. Incremental Learning

```
Batch 1/16 trained
Batch 2/16 trained
Batch 3/16 trained
Batch 4/16 trained
Batch 5/16 trained
Batch 6/16 trained
Batch 7/16 trained
Batch 8/16 trained
Batch 9/16 trained
Batch 10/16 trained
Batch 11/16 trained
Batch 12/16 trained
Batch 13/16 trained
Batch 14/16 trained
Batch 15/16 trained
Batch 16/16 trained
```

```
Final Accuracy: 0.42
```

Model telah dilatih melalui 16 batch, dan setelah selesai, akurasi akhirnya adalah 0.42. Ini menunjukkan bahwa meskipun model sudah dipelajari dengan seluruh data, hasil akurasinya masih rendah. Hal ini bisa terjadi karena beberapa faktor, seperti model yang belum optimal, kualitas data yang kurang, atau parameter yang perlu disesuaikan. Model mungkin membutuhkan pelatihan tambahan atau perubahan pada data agar bisa mencapai akurasi yang lebih tinggi.

Kesimpulan :

1. Model SGDClassifier Belajar Secara Bertahap
   - Sliding Window (Jendela Geser): Model belajar dari sebagian data, lalu menguji hasilnya setiap kali data baru masuk.

- Incremental Learning: Model belajar dari beberapa bagian data kecil, lalu diperbarui dengan data baru.

2. Persiapan Data yang Sederhana
   - Label Encoding: Mengubah data kategori (misalnya jenis kelamin) jadi angka.
   - Standarisasi: Menyamakan ukuran data supaya model lebih mudah memahaminya.
3. Mengukur Akurasi Secara Langsung
   Setiap kali model beradaptasi, kita langsung melihat seberapa akurat hasilnya, sehingga bisa tahu apakah model bekerja dengan baik atau tidak.
4. Cocok untuk Data yang Terus Masuk
   Model ini sangat berguna jika data datang terus-menerus, seperti di sistem online learning, di mana model bisa terus belajar tanpa harus mulai dari awal.

Lampiran:

1. Link data set dan google colab:
   https://colab.research.google.com/drive/18Ijg3sltq8SJ_dqNnstvnSf2D4GuL8rw?usp=sharing