PRAKTIKUM DATA MINING

Nama : Hawa Andini Hadi

NIM : 23051214238

Kelas/Angkatan : DTMG1/2023

Algoritma : Support Vector Machine, Artificial Neural Network, dan Ensamble

Jenis Analisis : Classification

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. Missing Value.

2. SMOTE (Synthetic Minority Over-sampling Technique) untuk menyeimbangkan distribusi kelas pada data.
3. Split datset: training (80%) dan test (20%).
4. Konversi Data ke PyTorch

Metode Evaluasi Model :

1. Akurasi (Accuracy Score): Mengukur seberapa sering model membuat prediksi yang benar.
2. Classification Report: Menampilkan metrik precision, recall, dan f1-score untuk setiap kategori obesitas.
3. GridSearchCV digunakan untuk optimasi hyperparameter
4. Eksperimen dengan Fungsi Aktivasi di Output Layer
    - Sigmoid: Menghasilkan output dalam rentang (0,1) (umum untuk klasifikasi biner).
    - Tanh: Menghasilkan output dalam rentang (-1,1).
    - Linear: Output bisa bernilai apa saja (tidak ada aktivasi).
    - Sign: Output hanya bisa bernilai -1 atau 1.
    - SoftPlus: Output selalu positif (mirip ReLU tetapi lebih halus).
5. Loss Function

Hasil Evaluasi :

1. Setiap metode diuji, lalu dibandingkan untuk melihat mana yang paling baik dalam membuat prediksi.
2. Hasil akhir menunjukkan metode mana yang paling akurat dalam mengenali pola dalam data.
3. Hasil Eksperimen dengan Fungsi Aktivasi di Output Layer
    - Sigmoid Output: 0.4949 (optimal)
    - Tanh Output : 0.4826
    - Linear Output: 0.4956
    - Sign Output : -1.0000
    - SoftPlus Output: 0.4735
4. Hasil akurasi:
    - Decision Tree (0.81)
    - Naïve Bayes (0.56)
    - KNN (0.69)
    - Logistic Regression (0.58)
    - Support Vector Machine (0.59)
    - Artificial Neural Network (0.63)
    - Bagging (DT) (0.95)
    - Random Forest (0.95)
    - Gradient Boosting (0.94)
    - Voting (0.94)
    - Extra Trees (0.93)

- Stacking (0.92)
- AdaBoost (0.42)

Kesimpulan :

1. Decision Tree adalah model paling akurat (0.81), tetapi rentan overfitting.
2. KNN (0.69) dan ANN (0.63) bisa menjadi alternatif, terutama jika dilakukan optimasi lebih lanjut.
3. Naïve Bayes, Logistic Regression, dan SVM kurang cocok karena akurasinya rendah.
4. Beberapa metode lebih akurat daripada yang lain, tergantung jenis datanya.
5. Dalam ANN fungsi aktivasi:
   - Sigmoid adalah pilihan terbaik karena memberikan hasil antara 0 dan 1, cocok untuk klasifikasi biner.
   - Tanh bisa digunakan, tetapi lebih cocok untuk data yang memiliki nilai negatif.
   - Linear dan SoftPlus tidak ideal karena hasilnya tidak terbatas.
   - Sign terlalu ekstrem, hanya cocok untuk kondisi tertentu.
6. Random Forest (Bagging) memiliki akurasi tertinggi karena kestabilannya terhadap overfitting.
7. Gradient Boosting lebih baik dari AdaBoost karena menangani fitur yang lebih kompleks.
8. AdaBoost memiliki akurasi terendah karena lebih sensitif terhadap noise dalam data.

Lampiran:

1. Link google colab
   https://colab.research.google.com/drive/15OVLqGqgvycprh5djtX03TTDDn7sSmAp?usp=sharing
   https://colab.research.google.com/drive/1ICORIi_qTgSE25CZ33C_cAc0IQ9Oj7ub?usp=sharing
   https://colab.research.google.com/drive/1nG3X2bRa8hFuKNl7i595M256HPjSOyjc?usp=sharing