PRAKTIKUM DATA MINING

Nama : Hawa Andini Hadi

NIM : 23051214238

Kelas/Angkatan : DTMG1/2023

Algoritma : Decision Tree

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. Handling Missing Values: Nilai yang kosong (NaN) diganti dengan rata-rata untuk fitur numerik.
2. Label Encoding: Fitur kategorikal diubah menjadi nilai numerik menggunakan LabelEncoder.
3. Feature Selection: Fitur target (NObeyesdad) dipisahkan dari fitur prediktor (X).
4. Data Splitting: Dataset dibagi menjadi training set (80%) dan testing set (20%) menggunakan train_test_split() dan cross_validation.

Metode Evaluasi Model :

1. Akurasi (Accuracy Score): Mengukur seberapa sering model membuat prediksi yang benar.
2. Confusion Matrix: Untuk melihat distribusi prediksi benar dan salah pada setiap kelas obesitas.
3. Classification Report: Menampilkan metrik precision, recall, dan f1-score untuk setiap kategori obesitas.
4. Perbandingan Gini dan Entropy: Perbedaan akurasi antara kedua metode dibandingkan untuk menentukan mana yang lebih optimal dalam mengklasifikasikan tingkat obesitas.
5. Train-Test Split: Model dilatih menggunakan 80% data dan diuji pada 20% data test.
6. Cross-Validation (10-Fold CV): Model dilatih 10 kali, dengan 9 bagian untuk training dan 1 bagian untuk testing secara bergantian.
7. Hyperparameter Tuning: Parameter max_depth dan min_samples_split dicari optimalnya menggunakan GridSearchCV untuk meningkatkan performa model.

Hasil Evaluasi :

1. Akurasi model Decision Tree cukup baik, dengan skor yang dapat bervariasi tergantung pada kedalaman pohon dan parameter lainnya.
2. Perbandingan Gini vs Entropy:
   - Gini memberikan akurasi yang lebih tinggi dan pemrosesan yang lebih cepat pada dataset ini.
   - Entropy menghasilkan model yang sedikit lebih dalam, tetapi terkadang dapat meningkatkan kemampuan model dalam menangani data dengan distribusi kelas yang lebih kompleks.
3. Classification report menunjukkan bahwa model bekerja lebih baik pada beberapa kelas obesitas dibandingkan yang lain, yang menunjukkan adanya ketidakseimbangan dalam dataset.
4. Confusion Matrix mengungkapkan bahwa model memiliki beberapa kesalahan klasifikasi, terutama pada kelas yang lebih jarang muncul.
5. Train-Test Split
   - Model dilatih menggunakan 80% data dan diuji pada 20% data test.
   - Akurasi hasil train-test split: 79.34%.
6. Cross-Validation (10-Fold CV)
   - Dataset dibagi menjadi 10 bagian.
   - Model dilatih 10 kali dengan skema 9 bagian training dan 1 bagian testing secara bergantian.

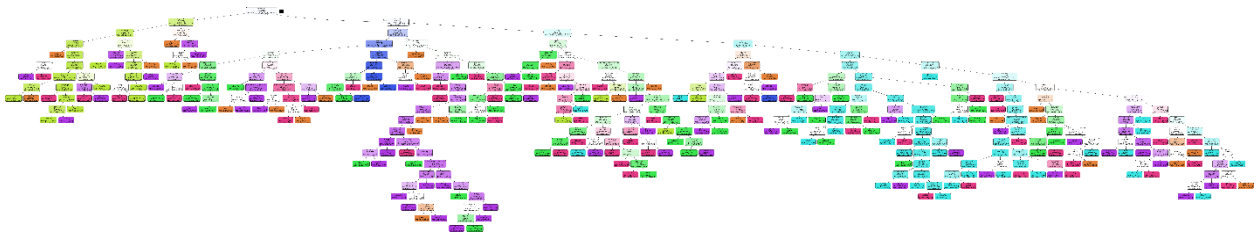
- Rata-rata akurasi dari Cross-Validation: 82.52%, lebih stabil dibandingkan train-test split.

Kesimpulan :

1. Decision Tree Classifier dapat digunakan untuk klasifikasi obesitas dengan akurasi yang cukup baik.
2. Hyperparameter tuning meningkatkan kinerja model dengan mencari parameter optimal.
3. Cross-validation (82.52%) lebih baik dibandingkan train-test split (79.34%), karena:
   - Memberikan evaluasi yang lebih stabil.
   - Mengurangi risiko overfitting atau underfitting.
4. Menggunakan Cross-Validation lebih akurat dibandingkan Train-Test Split.

Lampiran:

1. Gini



2. Entropy

3. Link data set dan google colab
   https://drive.google.com/drive/folders/1uCqCh7-mt9cx9uwGaQw2aBNAR4nWHW_M?usp=drive_link