PRAKTIKUM DATA MINING

Nama : Hawa Andini Hadi

NIM : 23051214238

Kelas/Angkatan : DTMG1/2023

Algoritma : KNN, Naive Bayes, Regresi Logistik

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. **Mengubah Data Kategorik:** Kategori diubah menjadi angka dengan One-Hot Encoding agar bisa diproses oleh model.
2. **Normalisasi Data Numerik:** Angka-angka dengan skala berbeda diseragamkan menggunakan StandardScaler agar model bekerja lebih akurat.
3. **Menyeimbangkan Data:** Jika jumlah kategori tidak seimbang, SMOTE menambah sampel sintetis agar model tidak bias.
4. **Seleksi Fitur Penting:** SelectKBest dengan ANOVA F-test digunakan untuk memilih fitur yang paling berpengaruh terhadap obesitas.

Metode Evaluasi Model :

1. **Akurasi:** Mengukur seberapa sering model memprediksi dengan benar.
2. **F1-Score:** Menyeimbangkan presisi dan recall, terutama untuk data yang tidak seimbang.
3. **ROC AUC Score:** Menilai kemampuan model dalam membedakan kelas positif dan negatif.
4. **Cross-validation:** Menggunakan Stratified K-Fold (5 lipatan) untuk hasil evaluasi yang lebih stabil.
5. **Overfitting Check**: Membandingkan akurasi train dan test untuk mendeteksi kompleksitas model.
6. **Tuning Hyperparameter:** Menggunakan GridSearchCV untuk mencari kombinasi parameter terbaik agar model bekerja optimal.

Hasil Evaluasi :

Algoritma	Train Accuracy	Test Accuracy	AUC Score
Decision Tree	0.89	0.66	0.84
KNN	0.89	0.67	0.87
Naive Bayes	0.58	0.59	0.84
Regresi Logistik	0.61	0.60	0.85

Kesimpulan :

Algoritma **KNN** memberikan hasil paling baik dengan akurasi 67% dan skor AUC 0.87. Ini berarti model tersebut cukup andal dalam mengidentifikasi tingkat obesitas dibandingkan dengan metode lainnya.

Namun, algoritma **Decision Tree** menunjukkan akurasi yang sangat tinggi (89%), tetapi kinerjanya menurun saat diuji dengan data baru (66%). Hal ini berarti model tersebut mungkin terlalu menyesuaikan diri dengan data dan kurang fleksibel terhadap data baru.

Sementara itu, algoritma **Naive Bayes dan Regresi Logistik** memiliki akurasi yang lebih rendah, tetapi tetap cukup baik dalam membedakan kategori obesitas, dengan skor AUC di atas 0.80.

Dari hasil ini, algoritma yang paling aman dan stabil untuk digunakan dalam memprediksi obesitas adalah **KNN atau Regresi Logistik**:

KNN bekerja dengan baik dalam mengenali pola obesitas, meskipun bisa lebih lambat jika data yang digunakan sangat besar dan **Regresi Logistik** lebih mudah dipahami dan memberikan hasil yang lebih stabil, sehingga cocok digunakan dalam bidang kesehatan atau penelitian medis.

Lampiran:

1. Link data set dan google colab: https://colab.research.google.com/drive/1R2hulzte7voIYHetHQTdztlyaNJQCTn_?usp=sharing