Classification Model Report

Kevin Olind Hasudungan Nainggolan/1103210140

Exploratory Data Analysis (EDA)

- **Dataset Summary**:
    - Dataset awal memiliki jumlah 253680 dengan 22 kolom
    - Setelah dilakukan scan duplikasi dengan ***df.duplicated().sum()*** dan penghapusan duplikasi dataset memiliki 229781 baris dan 22 kolom.
    - Dataset ini juga terdiri dari data numerik tipe float64
    - Setelah dicheck menggunakan ***f.isnull().sum()*** untuk mengecek nilai NaN/NULL didapati tidak ada nilai kosong di dalam dataset.
- **Distribusi Target Variable**:
    - Kategori 0 (Tidak Diabetes) memiliki data terbanyak dengan nilai sekitar 190000 data.
    - Kategori 1 (Prediabetes) memiliki data terendah dengan sekitar 7000 data.
    - Kategori 2 (Diabetes) memiliki sekitar 40000 data.
- **Korelasi Antar Fitur (Heatmap)**:
    - Korelasi antar fitur umumnya lemah. Tidak ada hubungan linear yang kuat.
    - Hubungan yang memiliki nilai korelasi yang lumayan seperti GenHlth dan PhyHlth dengan 52% nilai korelasi positif dan 33% nilai korelasi negative pada GenHlth dengan Income
    - Nilai 1 yang ada merupakan nilai yang saling membandingi dengan kolomnya sendiri

2. Visualisasi

- **Distribusi BMI terhadap Target**:
    - Orang dengan diabetes (2) cenderung memiliki BMI yang lebih tinggi dibanding kategori lainnya.
- **Distribusi Usia terhadap Target**:
    - Usia median meningkat dari kategori 0 ke kategori 2, menunjukkan bahwa usia adalah faktor penting untuk diabetes. Orang dengan diabetes (kategori 2) cenderung lebih tua.

3. Preprocessing

- **SMOTE**:
    - Digunakan untuk menangani ketidakseimbangan kelas dengan menghasilkan data sintetis untuk kelas minoritas.
    - Setelah SMOTE, dataset menjadi seimbang dengan total 456,132 sampel di data training dan 114,033 sampel di data testing.
- **Standarisasi**:
    - Menggunakan StandardScaler untuk standarisasi fitur numerik agar semua fitur memiliki distribusi serupa.

4. Modeling dan Evaluasi

A. Logistic Regression

- **Evaluasi Awal**:
    - Akurasi: 53%
    - Weighted F1-Score: 52%
    - Model memiliki kesulitan mengenali kelas 1 (Prediabetes) dengan F1-Score 38%.
- **Hyperparameter Tuning**:
    - Hasil terbaik ditemukan dengan:
        - C = 1
        - Penalty = 'l1'
        - Solver = 'saga'
    - Weighted F1-Score terbaik setelah tuning: 52.5%

B. Decision Tree

- **Evaluasi Awal**:
    - Akurasi: 85%
    - Weighted F1-Score: 85%
    - Model memiliki keunggulan mengenali kelas 1 (Prediabetes) dengan F1-Score 91% berbeda jauh dengan nilai logistic regression.
- **Hyperparameter Tuning**:
    - Maksimum kedalaman pohon (`max_depth`): [5, 10, 20, None].
    - Minimum jumlah sampel untuk memisahkan node (`min_samples_split`): [2, 5, 10].

- Minimum jumlah sampel di daun (`min_samples_leaf`): [1, 2, 4].
- max_depth: None.
- min_samples_split: 5.
- min_samples_leaf: 1.
- Weighted F1-Score: 0.8491**.**

C. K-NNeighbor

- **Evaluasi Awal**:
  - Akurasi: 86%
  - Weighted F1-Score: 86%
  - Model ini juga memiliki keunggulan mengenali kelas 1 (Prediabetes) dengan F1-Score 94% sedikit naik dibandingkan dengan model Decision Tree.
- **Hyperparameter Tuning**:
  - Jumlah tetangga terdekat (n_neighbors): [3, 5, 7, 9].
  - Bobot (weights): ['uniform', 'distance'].
  - Metode pengukuran jarak (metric): ['euclidean', 'manhattan'].
  - n_neighbors: 7.
  - weights: Distance.
  - metric: Manhattan.
  - Weighted F1-Score: (belum diberikan hasil, estimasi sekitar 0.86 dari performa awal)**.**

Kesimpulan

k-NN memberikan akurasi terbaik (86%) dan kemampuan mengenali kelas minoritas (kategori 1 - Prediabetes) dengan F1-Score 94%, menjadikannya model paling efektif secara keseluruhan. Decision Tree memberikan hasil yang kompetitif dengan akurasi 85% dan F1-Score 91% untuk kategori 1. Logistic Regression menunjukkan performa yang lebih rendah (53%) dibandingkan model lainnya dan kurang mampu menangani kelas minoritas. Model **k-NN** adalah model terbaik untuk dataset ini dalam hal akurasi dan pengenalan kelas minoritas, meskipun begitu ada juga kelemahan yang dimiliki seperti waktu compiling yang lebih lama. Sementara Decision Tree menawarkan interpretasi yang lebih sederhana dengan performa yang kompetitif.