ANALISIS PENGARUH KONSUMSI KOPI TERHADAP KESEHATAN MENGGUNAKAN PENDEKATAN MACHINE LEARNING RANDOM FOREST DENGAN STUDI DATA GLOBAL

Dipresentasikan oleh Kelompok 4

Dosen Pengampu :
Ir. Satrio Hadi Wijoyo, S.Si., S.Pd., M.Kom.

Data Sains

NAMA ANGGOTA KELOMPOK:

1. Acik Imtia Chana - 235150701111038
2. Hanifa Syifa Safitri - 235150707111031
3. Mutiara Rosida Sholihat - 235150707111035
4. Lutfiah Nailil Izzah - 235150707111038
5. Sylvasisca Andini Faradyan - 235150707111040

PENDAHULUAN

Konsumsi kopi penting diteliti karena kopi merupakan minuman yang sangat populer dan menjadi bagian dari gaya hidup masyarakat, sehingga potensi dampaknya terhadap kesehatan perlu dipahami secara ilmiah. Dataset Global Coffee Health digunakan untuk melihat hubungan konsumsi kopi, tidur, stres, BMI, dan kebiasaan lainnya. Oleh karena itu, machine learning, khususnya Random Forest digunakan karena mampu memproses dataset besar, menangani hubungan non-linear, serta mengidentifikasi fitur paling berpengaruh sehingga dapat membantu memberikan gambaran yang lebih akurat mengenai pengaruh konsumsi kopi terhadap kesehatan.

TUJUAN PENELITIAN

- Memprediksi kategori Health_Issues berdasarkan fitur gaya hidup.
- Mengetahui fitur yang paling berpengaruh terhadap kondisi kesehatan.
- Menggunakan algoritma Random Forest untuk klasifikasi multiclass.

DATASET OVERVIEW

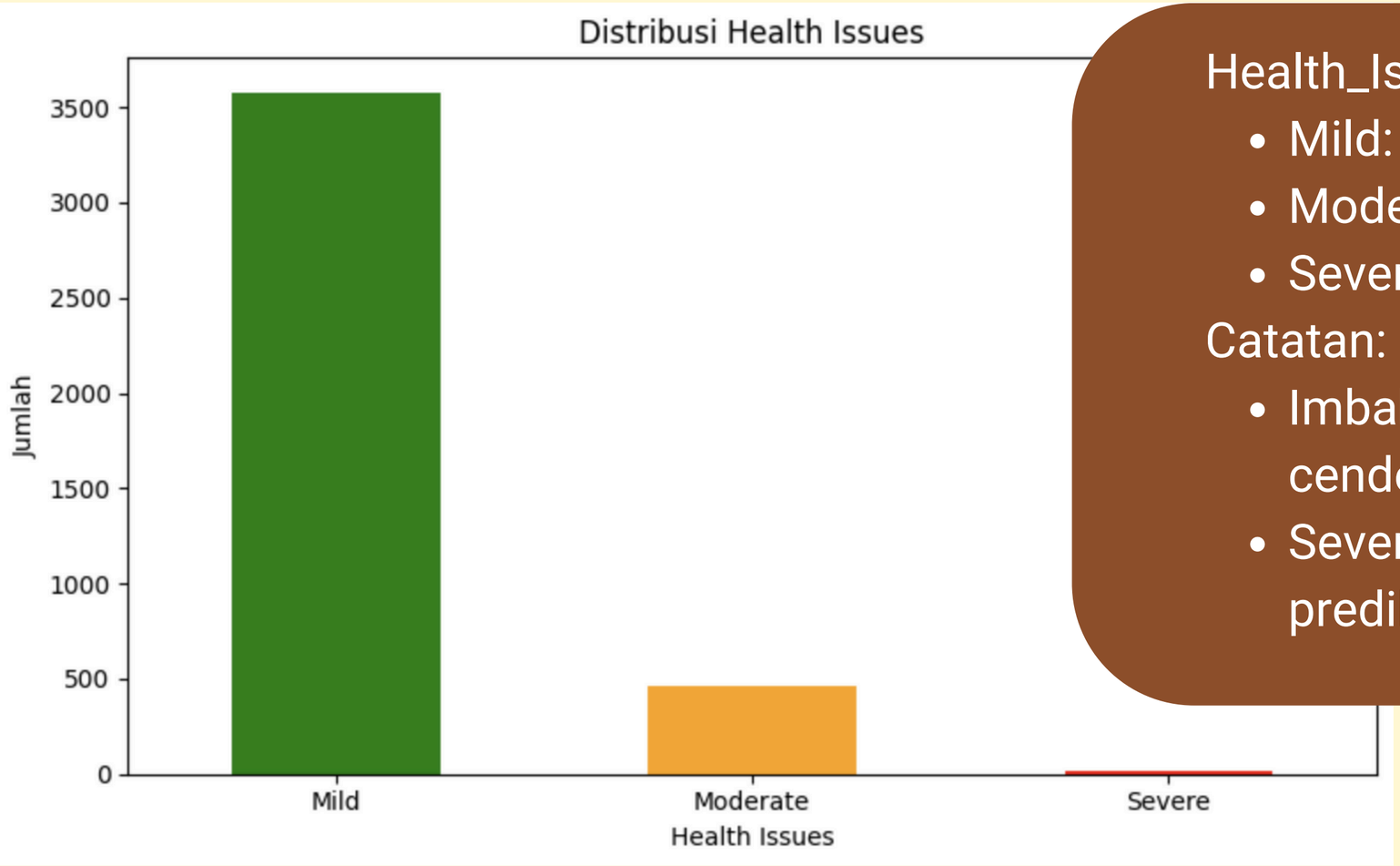
	ID	Age	Gender	Country	Coffee_Intake	Caffeine_mg	Sleep_Hours	Sleep_Quality	BMI	Heart_Rate	Stress_Level	Physical_Activity_Hours	Health_Issues	Occupation	Smoking	Alcohol_Consumption
0	1	40	Male	Germany	3.5	328.1	7.5	Good	24.9	78	Low	14.5	NaN	Other	0	0
1	2	33	Male	Germany	1.0	94.1	6.2	Good	20.0	67	Low	11.0	NaN	Service	0	0
2	3	42	Male	Brazil	5.3	503.7	5.9	Fair	22.7	59	Medium	11.2	Mild	Office	0	0
3	4	53	Male	Germany	2.6	249.2	7.3	Good	24.7	71	Low	6.6	Mild	Other	0	0
4	5	32	Female	Spain	3.1	298.0	5.3	Fair	24.1	76	Medium	8.5	Mild	Student	0	1

Dataset GlobalCoffeeHealth berisi 10.000 catatan sintetis yang mencerminkan pola nyata konsumsi kopi, perilaku tidur, dan hasil kesehatan di 20 negara. Dataset ini mencakup demografi, konsumsi kopi harian, kadar kafein, durasi dan kualitas tidur, BMI, detak jantung, stres, aktivitas fisik, masalah kesehatan, pekerjaan, merokok, dan konsumsi alkohol.

EKSPLORASI DATA

- Total data awal: 10.000 observasi, 16 fitur
- Setelah pembersihan (drop missing target) → 4.059 observasi
- Distribusi kelas:
  - Mild: 3.579
  - Moderate: 463
  - Severe: 17
- Tantangan utama: class imbalance (Severe hanya 0.4%)

DISTRIBUSI TARGET



Health_Issues Distribution:

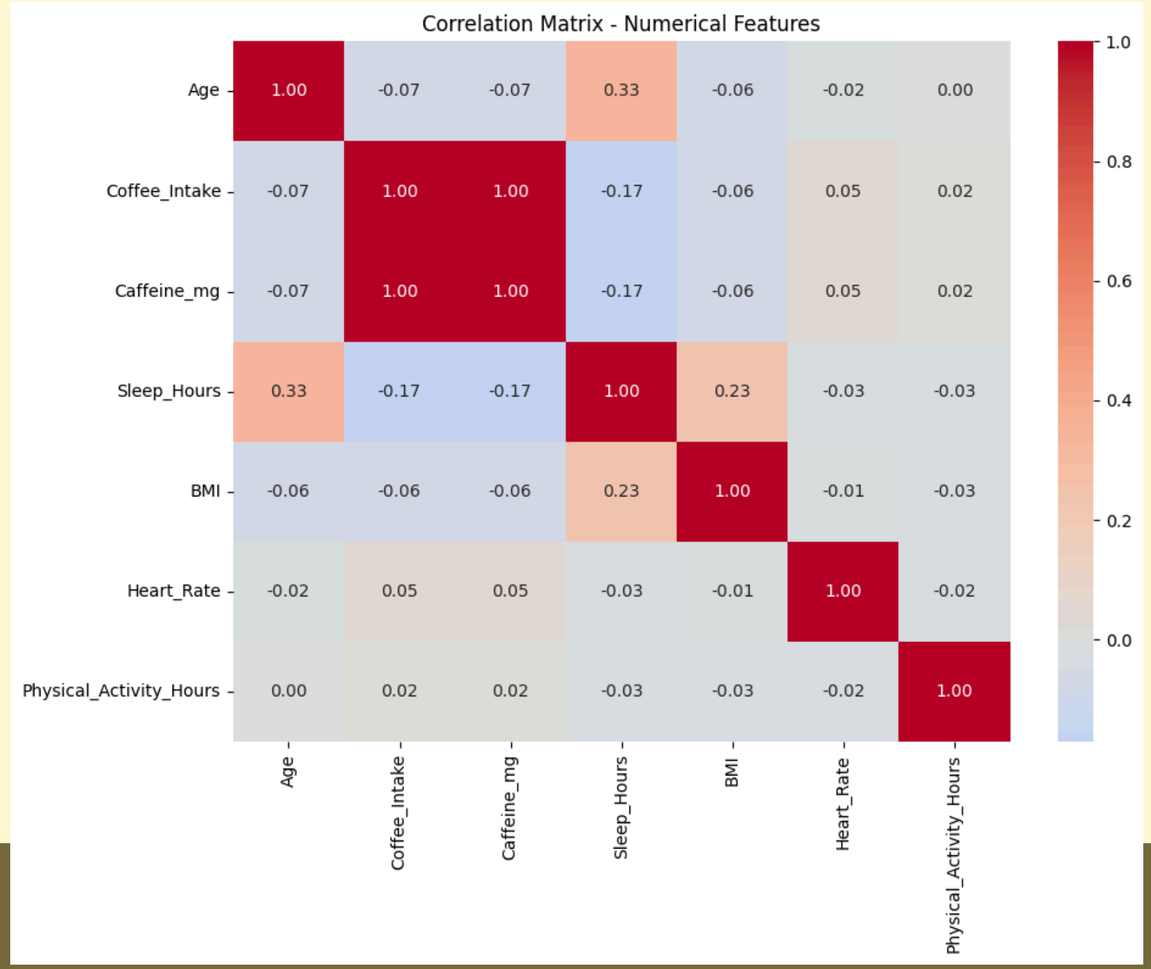
- Mild: 88.2%
- Moderate: 11.4%
- Severe: 0.4%

Catatan:

- Imbalanced sangat parah → model cenderung bias ke Mild
- Severe hampir tidak terwakili → risiko prediksi gagal

KORELASI FITUR NUMERIK

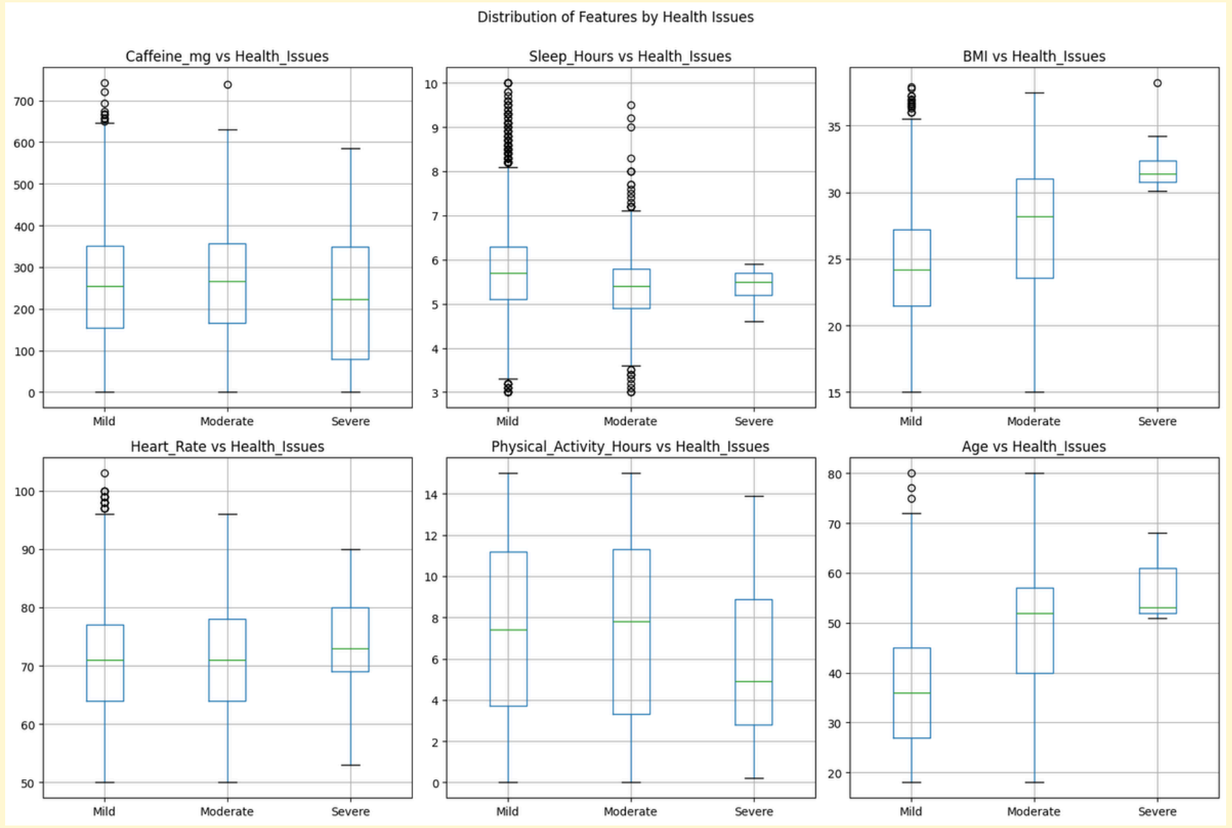
- Coffee_Intake ↔ Caffeine_mg: korelasi sempurna (r = 1.00) → fitur redundant → dihapus
- Korelasi lain lemah–sedang:
  - Age ↔ Sleep_Hours (0.33)
  - BMI ↔ Sleep_Hours (0.23)



DISTRIBUSI FITUR PER KELAS

Temuan utama dari boxplot:

- BMI: meningkat seiring tingkat masalah kesehatan (Mild ≈24 → Moderate ≈28 → Severe ≈32)
- Sleep_Hours: Mild lebih tinggi (7–10 jam), Severe rendah (3–6 jam)
- Age: Severe lebih tua (55–60), Mild (30–40)
- Caffeine_mg: Severity lebih rendah konsumsi → kemungkinan karena pembatasan medis



TAHAP PREPROCESSING

```
Sleep_Quality encoded: {0: 961, 1: 2050, 2: 826, 3: 222}
Stress_Level encoded: {0: 1048, 1: 2050, 2: 961}
Health_Issues encoded: {0: 3579, 1: 463, 2: 17}
Gender encoded
Smoking encoded
Country one-hot encoded into 19 columns
Occupation one-hot encoded into 4 columns
Alcohol_Consumption one-hot encoded into 1 columns

Dataset shape after encoding: (4059, 35)
Missing values setelah encoding: 0
```

Langkah utama:

1. Drop missing values pada target (5.941 baris)
2. Drop fitur redundant: Coffee_Intake
3. Drop kolom ID
4. Encoding:
    - Ordinal: Sleep_Quality, Stress_Level, Health_Issues
    - Label: Gender, Smoking
    - One-Hot: Country, Occupation, Alcohol

Hasil akhir:

Fitur meningkat dari 15 → ±30–40 fitur setelah one-hot encoding.

PEMBAGIAN DATA

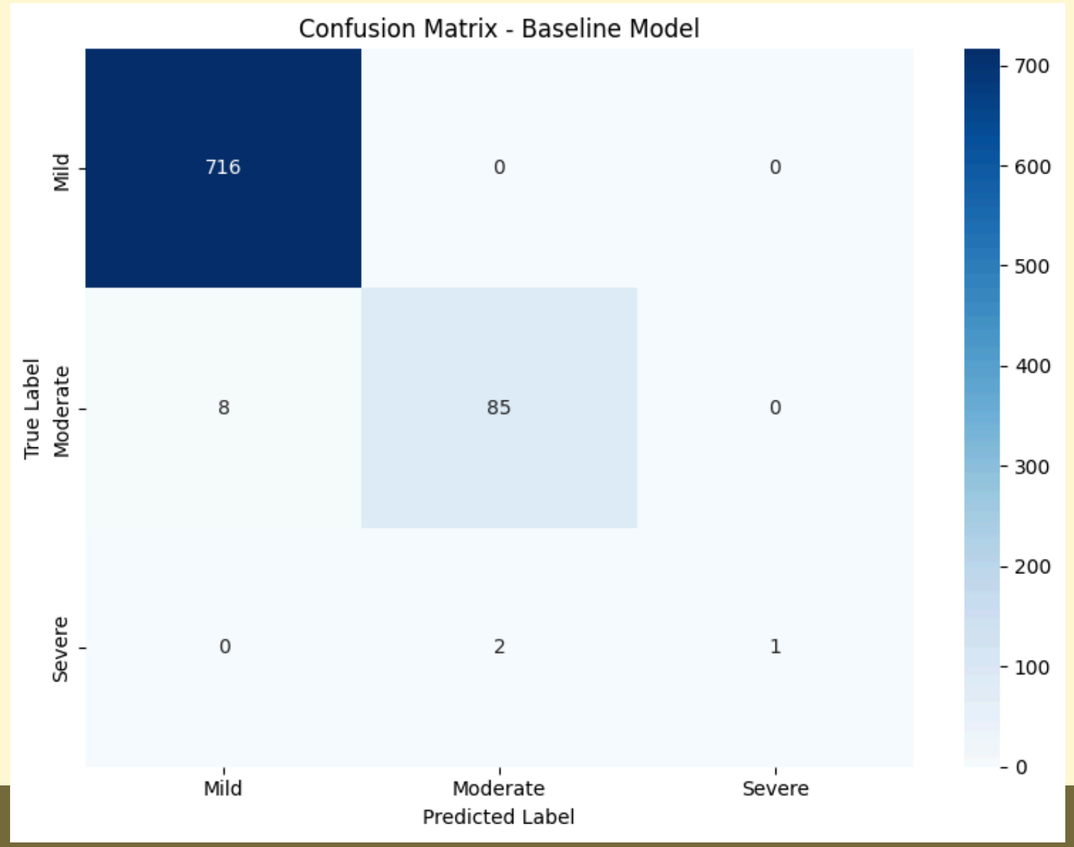
```
Target distribution in training set:
Health_Issues
0    2863
1     370
2      14
Name: count, dtype: int64
...
0    716
1     93
2      3
Name: count, dtype: int64
```

Train-Test Split (80:20, Stratified):

- Train: Mild 2.863 | Moderate 370 | Severe 14
- Test: Mild 716 | Moderate 93 | Severe 3

Alasan stratified: menjaga proporsi kelas pada kedua subset.

MODEL BASELINE



Algoritma: Random Forest (default param)

- n_estimators = 100

Hasil baseline:

- Accuracy: 0.9840
- F1-score (weighted): 0.9820
- Severe: F1 = 0.00 (gagal total)

Insight:

- Kelas Mild diprediksi sangat baik (715/716 benar)
- Moderate masih ada misclass → diprediksi Mild
- Severe tidak terdeteksi karena sampel sangat kecil

HYPERPARAMETER TUNING

GridSearchCV (48 kombinasi, 5-fold CV)

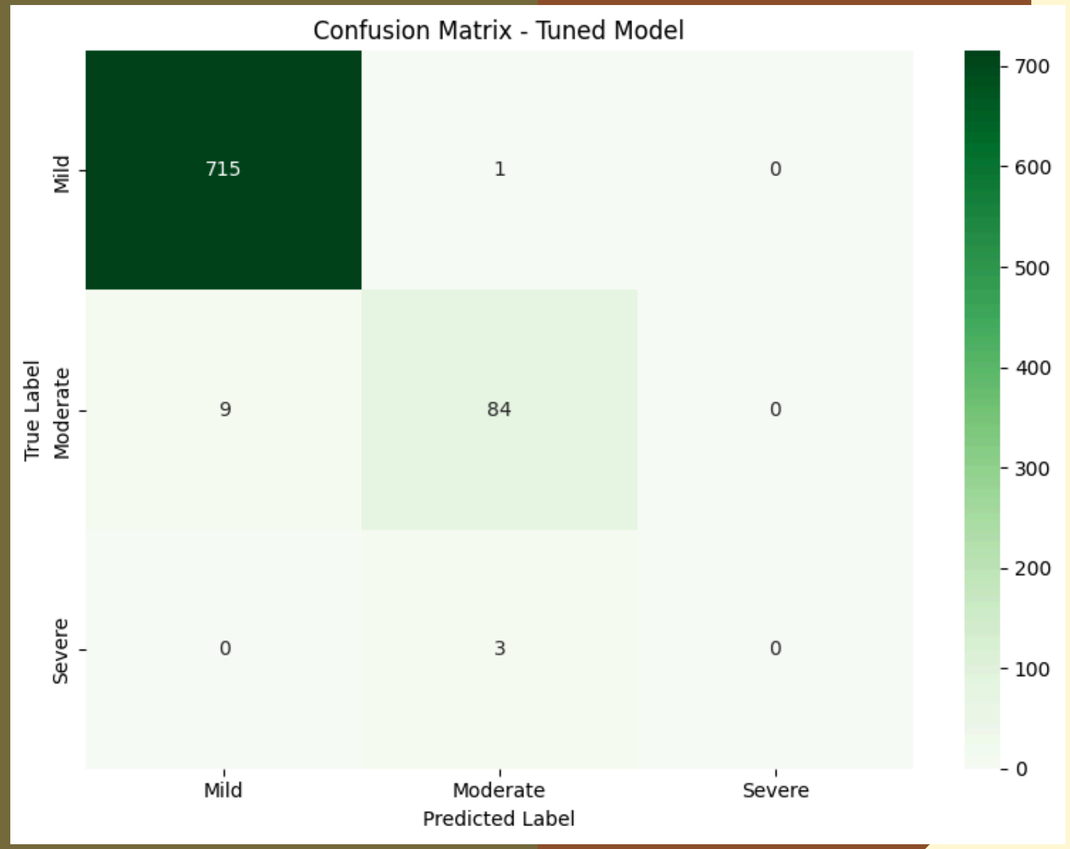
Parameter terbaik:

- class_weight = balanced
- max_depth = 20
- min_samples_leaf = 2
- min_samples_split = 5
- n_estimators = 100

CV F1-score: 0.9872

Catatan:

class_weight balanced → sangat penting mengatasi imbalance.



EVALUASI MODEL FINAL

Hasil setelah tuning:

- Accuracy: 0.9840
- F1-score (weighted): 0.9820
- F1-score (macro): 0.6404

Per kelas:

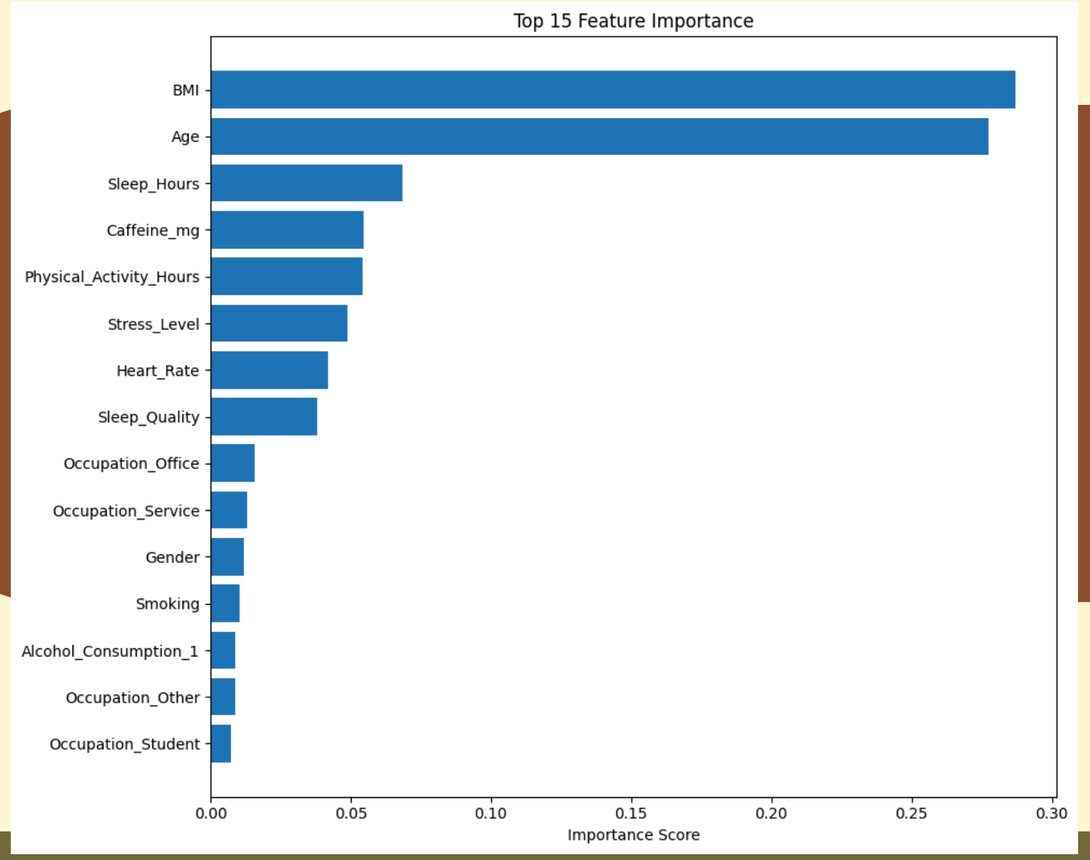
- Mild → Precision 0.99 | Recall 1.00 | F1 0.99
- Moderate → Precision 0.95 | Recall 0.90 | F1 0.93
- Severe → F1 = 0.00 (3 sampel test → semua salah)

FEATURE IMPORTANCE

Top fitur paling berpengaruh:

1. BMI – 0.29
2. Age – 0.27
3. Sleep_Hours – 0.07
4. Caffeine_mg – 0.05
5. Physical_Activity_Hours – 0.05

Fitur lain (Stress_Level, Sleep_Quality, Heart_Rate) memiliki pengaruh sedang, sementara demografi (Gender, Smoking, Occupation) sangat kecil.



KESIMPULAN PRESENTASI

Dataset penelitian menunjukkan class imbalance yang sangat ekstrem, khususnya pada kelas Severe yang hanya memiliki sedikit sampel sehingga tidak cukup mewakili pola yang valid. Dari hasil analisis fitur, BMI, Age, dan Sleep_Hours muncul sebagai prediktor kesehatan yang paling kuat, karena ketiganya menunjukkan perbedaan distribusi yang jelas antar kategori Mild, Moderate, dan Severe. Pemodelan menggunakan Random Forest memberikan performa yang sangat baik untuk memprediksi kelas Mild dan Moderate, namun tetap mengalami kegagalan dalam memprediksi kelas Severe karena ketidakseimbangan data yang terlalu besar. Proses hyperparameter tuning memang meningkatkan akurasi dan stabilitas model, tetapi perbaikan tersebut tidak mampu sepenuhnya mengatasi masalah imbalance berat, sehingga kelas Severe tetap sulit diprediksi secara akurat.

TERIMA KASIH

Kelompok 4