Laporan Analisa Datasat Cereal

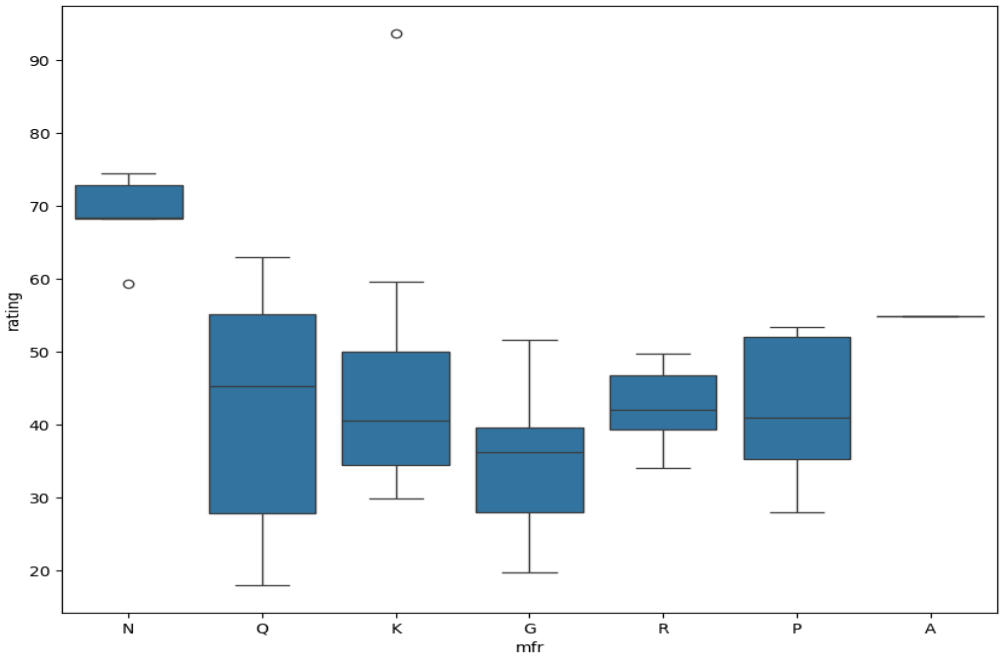
Daniel Christopher Juwono - C14220006
Harvey Kristandi - C14220244

1. Pendahuluan
    a. Tujuan : Menganalisa faktor nutrisi apa saja yang paling mempengaruhi Rating (penilaian) konsumen terhadap produk sereal & Memprediksi kategori tingkat kesehatan dari setiap cereal
    b. Variabel : name, manufatures, type, calories, protein, fat, sodium, fiber, carbohidrat, sugars, potassium, vitamins, display shelf level, weight, cups per serving, rating
    c. Data Cleaning :

```
# Hilangkan - 1
data = data[(data.carbo >= 0) & (data.sugars >= 0) & (data.potass >= 0)]
data[data == -1].count(axis=0)
```

- 1 menunjukan bahwa data tersebut NaN sehingga kami menghapus semua nilai -1.

2. Analisis Eksploratory Data Analysis (EDA)
    a. Visualisasi Perbandingan Rating dan Manufactures



Berdasarkan visualisasi perbandingan Manufaktur terhadap Rating, ditemukan bahwa Nabisco (N) memimpin pasar dalam hal kepuasan konsumen (rating rata-rata tertinggi), meskipun jumlah produknya lebih sedikit dibandingkan Kellogg's (K) dan General Mills (G). Hal ini mengindikasikan bahwa konsumen sereal cenderung memberikan penilaian positif pada kualitas nutrisi (kesehatan) dibandingkan sekadar variasi rasa yang ditawarkan oleh produsen massal.

    b. Urutan sereal dengan rating tertinggi dan terendah

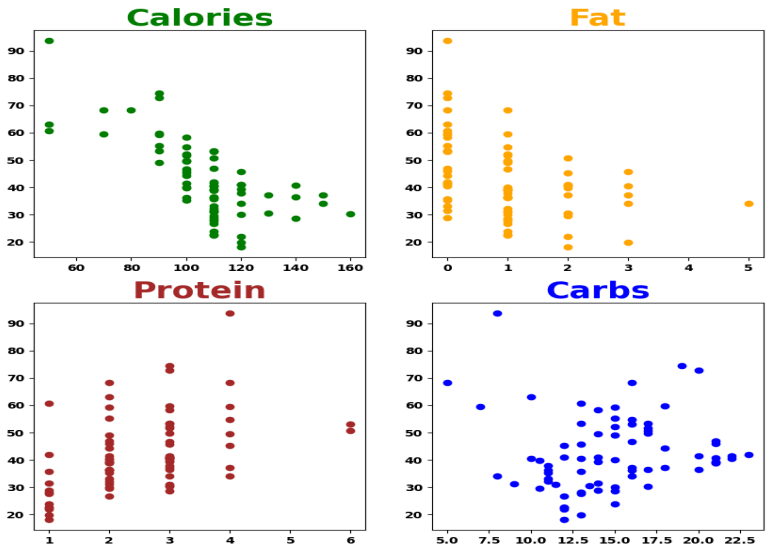
10 Sereal dengan Rating Tertinggi:

	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating	Health_Grade
3	All-Bran_with_Extra_Fiber	2	0	50	4	0	140	14.0	8.0	0.0	330.0	25	3	1.00	0.50	93.704912	Healthy
64	Shredded_Wheat_'n'Bran	3	0	90	3	0	0	4.0	19.0	0.0	140.0	0	1	1.00	0.67	74.472949	Healthy
65	Shredded_Wheat_spoon_size	3	0	90	3	0	0	3.0	20.0	0.0	120.0	0	1	1.00	0.67	72.801787	Healthy
0	100%_Bran	3	0	70	4	1	130	10.0	5.0	6.0	280.0	25	3	1.00	0.33	68.402973	Healthy
63	Shredded_Wheat	3	0	80	2	0	0	3.0	16.0	0.0	95.0	0	1	0.83	1.00	68.235885	Healthy
55	Puffed_Wheat	5	0	50	2	0	0	1.0	10.0	0.0	50.0	0	3	0.50	1.00	63.005645	Healthy
54	Puffed_Rice	5	0	50	1	0	0	0.0	13.0	0.0	15.0	0	3	0.50	1.00	60.756112	Healthy
50	Nutri-grain_Wheat	2	0	90	3	0	170	3.0	18.0	2.0	90.0	25	3	1.00	1.00	59.642837	Healthy
2	All-Bran	2	0	70	4	1	260	9.0	7.0	5.0	320.0	25	3	1.00	0.33	59.425505	Healthy
68	Strawberry_Fruit_Wheats	3	0	90	2	0	15	3.0	15.0	5.0	90.0	25	2	1.00	1.00	59.363993	Healthy

10 Sereal dengan Rating Terendah:

	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating	Health_Grade
70	Total_Raisin_Bran	1	0	140	3	1	190	4.0	15.0	14.0	230.0	100	3	1.5	1.00	28.592785	Unhealthy
29	Fruity_Pebbles	4	0	110	1	1	135	0.0	13.0	12.0	25.0	25	2	1.0	0.75	28.025765	Unhealthy
73	Trix	1	0	110	1	1	140	0.0	13.0	12.0	25.0	25	2	1.0	1.00	27.753301	Unhealthy
42	Lucky_Charms	1	0	110	2	1	180	0.0	12.0	12.0	55.0	25	2	1.0	1.00	26.734515	Unhealthy
31	Golden_Grahams	1	0	110	1	1	280	0.0	15.0	9.0	45.0	25	2	1.0	0.75	23.804043	Unhealthy
14	Cocoa_Puffs	1	0	110	1	1	180	0.0	12.0	13.0	55.0	25	2	1.0	1.00	22.736446	Unhealthy
18	Count_Chocula	1	0	110	1	1	180	0.0	12.0	13.0	65.0	25	2	1.0	1.00	22.396513	Unhealthy
35	Honey_Graham_Ohs	5	0	120	1	2	220	1.0	12.0	11.0	45.0	25	2	1.0	1.00	21.871292	Unhealthy
12	Cinnamon_Toast_Crunch	1	0	120	1	3	210	0.0	13.0	9.0	45.0	25	2	1.0	0.75	19.823573	Unhealthy
10	Cap'n'Crunch	5	0	120	1	2	220	0.0	12.0	12.0	35.0	25	2	1.0	0.75	18.042851	Unhealthy

c. Pengaruh Macronutrients dan Kalori terhadap Ratings



i. Karbohidrat dan Kalori adalah penghancur rating terbesar. Semakin tinggi kandungan karbohidrat atau kalori dalam satu porsi sereal, semakin rendah rating yang diberikan. Ini menunjukkan bahwa algoritma atau persepsi rating ini sangat memprioritaskan aspek kesehatan dan 'menghukum' sereal yang terlalu manis atau padat energi kosong.

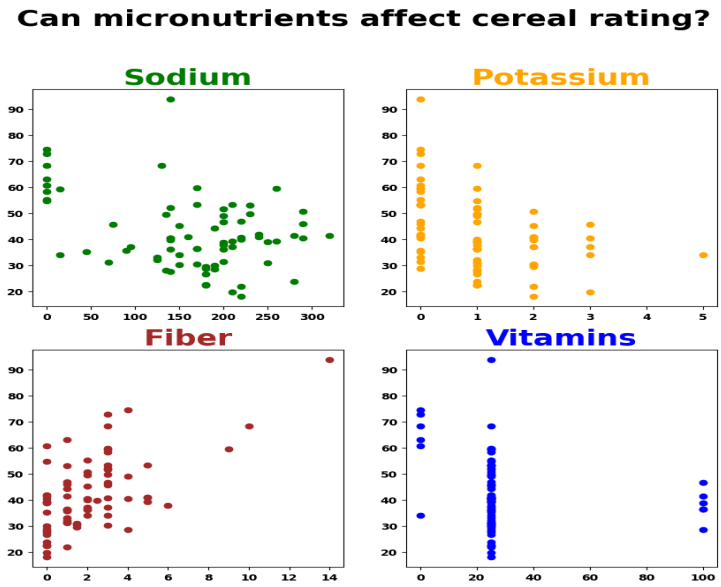
ii. Serat adalah faktor yang mendapatkan rating tinggi. Sereal dengan kandungan serat tinggi (seperti *All-Bran* atau sereal gandum utuh) hampir dipastikan

memiliki skor di atas rata-rata. Jika produsen ingin menaikkan rating produknya secara instan, cara paling efektif adalah meningkatkan kadar serat.

iii. Protein berkontribusi positif terhadap rating, namun peranannya adalah sebagai pendukung. Produk tinggi protein *dan* tinggi serat akan menjadi juara, tetapi protein saja tanpa serat mungkin tidak cukup untuk mendongkrak rating secara drastis jika gulanya masih tinggi.

iv. Meskipun lemak dan sodium (garam) berpengaruh negatif (lebih baik rendah), dampaknya tidak sedestruktif gula. Artinya, konsumen atau sistem rating ini lebih 'memaafkan' sedikit lemak dibandingkan kelebihan gula.

d. Pengaruh Micronutrients terhadap Rating



i. Pada grafik sodium, titik-titik data tersebar tanpa pola yang jelas, namun terdapat kecenderungan bahwa cereal dengan kandungan sodium lebih tinggi cenderung memperoleh rating lebih rendah. Banyak produk dengan sodium di atas 150 mg justru berada pada rentang rating 20 hingga 40, sedangkan cereal dengan sodium rendah lebih sering mencapai rating di atas 60.

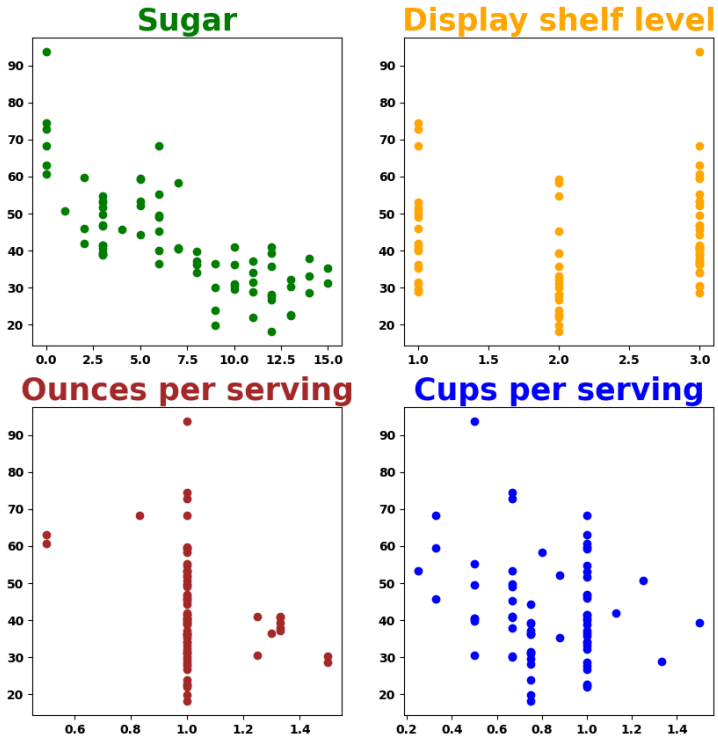
ii. Pola yang lebih tegas tampak pada grafik potassium, di mana peningkatan kandungan potassium berasosiasi dengan penurunan rating. Produk dengan potassium rendah masih menunjukkan variasi rating yang luas, namun ketika kandungan potassium meningkat hingga angka 3–5, ratingnya konsisten berada di kisaran rendah.

iii. Berbeda dengan sodium dan potassium, fiber justru menunjukkan kecenderungan positif terhadap rating. Cereal dengan kandungan fiber rendah (0–2 gram) umumnya memperoleh rating pada tingkat menengah ke bawah, sementara cereal dengan kandungan fiber yang tinggi—bahkan sampai 8 hingga 14 gram—cenderung mendapatkan rating lebih tinggi, beberapa bahkan mendekati

skor maksimal. Hal ini mengindikasikan bahwa serat merupakan salah satu komponen yang diapresiasi dalam penentuan rating cereal.

iv. grafik vitamins tidak menunjukkan pola hubungan yang berarti. Meski vitamin dikelompokkan dalam beberapa persentase (seperti 0%, 25%, hingga 100%), semua kelompok tersebut menghasilkan rating yang bervariasi dan tidak menunjukkan tren yang konsisten. Cereal dengan kandungan vitamin tinggi tidak secara otomatis mendapatkan rating yang lebih baik.

e. Pengaruh Sugar, Display Shelf Level, Weight dan Cups terhadap Rating



i. Pada grafik sugar tampak kecenderungan yang cukup jelas bahwa semakin tinggi kandungan gula dalam sebuah cereal, rating-nya justru menurun. Titik-titik pada kadar gula rendah masih menunjukkan variasi rating yang luas—mulai dari rendah hingga sangat tinggi—namun saat kandungan gula meningkat ke rentang 10 hingga 15 gram, hampir semua produk terakumulasi pada rating yang lebih rendah, umumnya di bawah 40. Hal ini menunjukkan bahwa gula merupakan faktor yang berdampak negatif terhadap persepsi kualitas cereal.

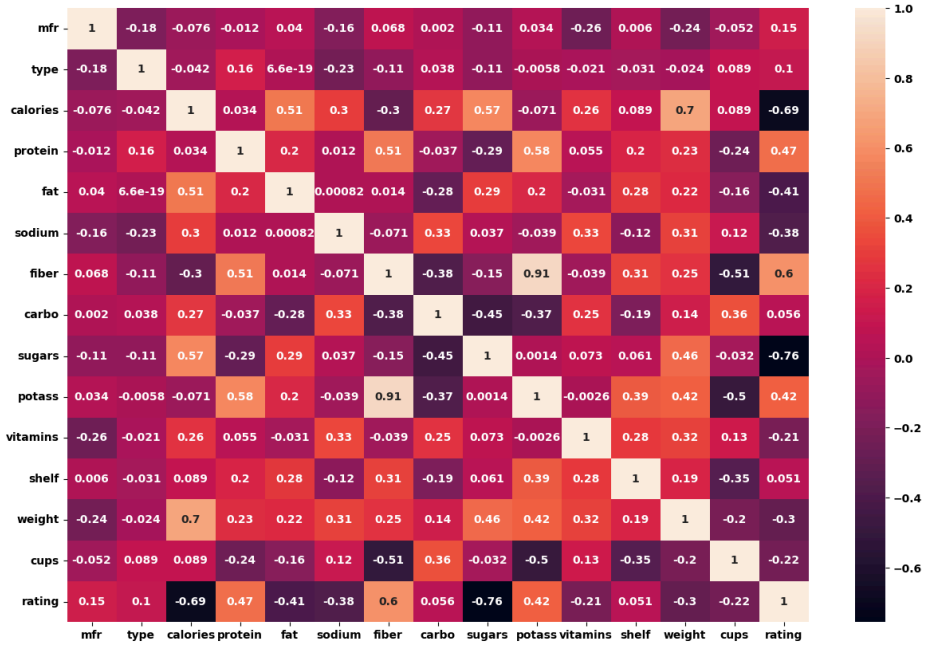
ii. Pada grafik display shelf level, tidak tampak pola yang kuat antara posisi produk di rak dan rating yang diperoleh. Setiap level (1, 2, dan 3) memiliki penyebaran rating yang serupa, baik untuk nilai rendah maupun tinggi. Ini mengindikasikan bahwa posisi penempatan cereal di rak tampaknya tidak menjadi faktor penentu yang memengaruhi kualitas atau penilaian produk.

iii. Berbeda dengan itu, grafik ounces per serving tidak memperlihatkan hubungan nyata antara jumlah ons per porsi dengan rating. Hampir semua data terkonsentrasi pada rentang ons yang sangat kecil, dan rating tersebar luas tanpa

pola meningkat atau menurun. Dengan demikian, ukuran porsi dalam bentuk ons tidak memberikan pengaruh berarti terhadap penilaian cereal.

iv. Hal yang sama juga terlihat pada grafik cups per serving. Meski variasi bentuk porsi lebih terlihat dibandingkan grafik ounces, pola hubungan tetap tidak menunjukkan arah tertentu. Cereal dengan jumlah cup yang rendah hingga tinggi memiliki rating yang sangat bervariasi, menunjukkan tidak ada korelasi yang kuat atau konsisten.

f. Heatmap Korelasi antara variabel



Dari heatmap korelasi tersebut terlihat bahwa hubungan antarvariabel dalam dataset cereal menunjukkan beberapa pola yang cukup kuat dan relevan untuk menjelaskan faktor-faktor yang mempengaruhi rating. Salah satu temuan paling mencolok adalah korelasi negatif yang sangat kuat antara sugars dan rating (–0.76), yang berarti semakin tinggi kandungan gula, semakin rendah rating cereal. Temuan ini konsisten dengan grafik sebelumnya dan menunjukkan bahwa gula merupakan faktor utama yang menurunkan persepsi kualitas. Selain itu, calories juga memiliki korelasi negatif cukup besar dengan rating (–0.69), menunjukkan bahwa cereal dengan kalori lebih tinggi cenderung dinilai kurang baik. Hal ini bisa jadi karena kalori yang tinggi sering muncul bersama kandungan gula dan karbohidrat yang lebih besar.

Sebaliknya, beberapa variabel menunjukkan hubungan positif dengan rating. Fiber memiliki korelasi positif moderat (0.60), yang mengindikasikan bahwa cereal tinggi serat lebih disukai atau dianggap lebih sehat oleh konsumen maupun penilai. Protein juga berhubungan positif dengan rating (0.47), sehingga kandungan protein dapat dianggap sebagai salah satu indikator kualitas. Kandungan potassium juga menunjukkan korelasi positif dengan rating (0.42), meski tidak sekuat variabel sebelumnya. Ketiga variabel ini

menggambarkan karakteristik cereal yang lebih sehat dan bergizi, dan konsisten berasosiasi dengan penilaian yang lebih baik.

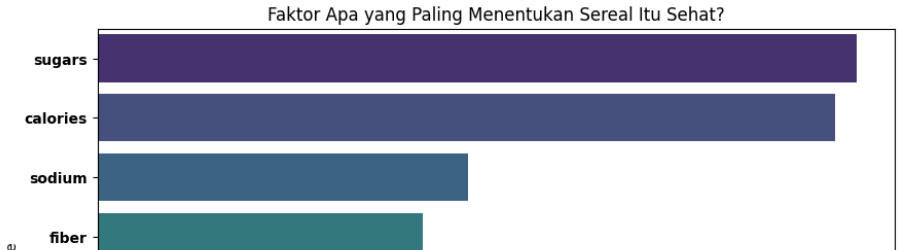
Selain hubungan dengan rating, heatmap ini juga menunjukkan beberapa korelasi antarvariabel yang cukup kuat dan perlu diperhatikan. Misalnya, fiber dan potassium memiliki korelasi sangat tinggi (0.91), yang menunjukkan bahwa produk tinggi serat biasanya juga tinggi potassium. Hubungan ini bisa mengarah pada multikolinearitas jika analisis regresi dilakukan. Kandungan weight dan calories juga berkorelasi tinggi (0.70), yang cukup logis karena porsi yang lebih berat cenderung mengandung lebih banyak kalori. Sebaliknya, beberapa variabel seperti vitamins, shelf level, dan cups tidak menunjukkan korelasi kuat dengan rating, sehingga perannya dalam memengaruhi kualitas cereal relatif kecil.

3. Model Klasifikasi
    a. Classification Report

```
Classification Report:
              precision    recall  f1-score   support

     Healthy       0.86      1.00      0.92         6
   Unhealthy       1.00      0.94      0.97        17

    accuracy                           0.96        23
   macro avg       0.93      0.97      0.95        23
weighted avg       0.96      0.96      0.96        23


Confusion Matrix:
[[ 6  0]
 [ 1 16]]
```

Hasil classification report menunjukkan bahwa model memiliki performa prediksi yang sangat baik dengan akurasi 96%. Model mampu mengenali seluruh sereal yang tergolong sehat dengan sempurna, ditunjukkan oleh nilai recall 1.00 pada kelas “Healthy”. Selain itu, model juga sangat akurat dalam memprediksi sereal tidak sehat, terbukti dari nilai precision 1.00 pada kelas “Unhealthy”. Satu-satunya kesalahan model terjadi pada kasus ketika satu sereal tidak sehat diprediksi sebagai sehat, sebagaimana terlihat pada confusion matrix. Secara keseluruhan, model menunjukkan kemampuan yang sangat kuat dan stabil dalam membedakan sereal sehat dan tidak sehat, terutama dengan penekanan kuat pada kandungan gula, kalori, dan sodium sebagai faktor utama. Dengan demikian, model ini dapat diandalkan untuk melakukan klasifikasi kesehatan sereal secara akurat dan konsisten.

    b. Feature Importance



Berdasarkan hasil analisis feature importance, faktor yang paling menentukan apakah suatu sereal dikategorikan sebagai sehat adalah kandungan gula, diikuti oleh kalori dan sodium. Hal ini menunjukkan bahwa model sangat sensitif terhadap tingkat kemanisan dan jumlah energi per porsi, yang memang secara nutrisi sering dijadikan indikator utama untuk menilai kesehatan sebuah produk makanan. Sereal dengan kandungan gula dan kalori yang tinggi cenderung diklasifikasikan sebagai tidak sehat, sementara sereal dengan serat lebih tinggi lebih sering diidentifikasi sebagai sehat. Faktor-faktor seperti lemak, karbohidrat, dan brand memberikan kontribusi yang jauh lebih kecil terhadap keputusan model, sehingga dapat disimpulkan bahwa evaluasi kesehatan sereal dalam model ini lebih berbasis kandungan nutrisi inti daripada aspek merek atau komposisi minor lainnya.

4. Kesimpulan
    a. Determinan Utama
        i. Gula (*Sugars*) dan Kalori adalah faktor yang paling merusak rating. Analisis korelasi dan *Feature Importance* dari Random Forest mengonfirmasi bahwa semakin tinggi kandungan gula, semakin rendah drastis skor rating produk tersebut.
        ii. Serat (*Fiber*) memiliki korelasi positif terkuat. Produk dengan kandungan serat tinggi secara konsisten menempati kuadran rating teratas.
    b. Manufactures Insight
        i. Meskipun Kellogg's dan General Mills mendominasi pasar dari sisi volume varian produk, rata-rata rating mereka cenderung moderat karena portofolio produk yang "campur aduk" (antara sereal sehat dan sereal tinggi gula).
        ii. Nabisco mencatatkan performa rating rata-rata tertinggi. Hal ini menunjukkan keberhasilan strategi produk mereka yang fokus pada sereal berbasis gandum utuh (*shredded wheat*) yang minim gula.
    c. Kinerja Model
        i. Pendekatan *Machine Learning* menggunakan Random Forest terbukti efektif dalam mengklasifikasikan sereal. Model berhasil memvalidasi bahwa keputusan

klasifikasi (baik/buruk) sangat bergantung pada fitur Gula dan Kalori sebagai indikator utama, mengungguli faktor lain seperti lemak atau sodium.

Secara keseluruhan, analisis ini membuktikan bahwa algoritma kepuasan/rating sereal pada dataset ini sangat bias terhadap kesehatan. Kunci keberhasilan produk dalam metrik ini bukan terletak pada kompleksitas vitamin atau mineral mikro, melainkan pada keseimbangan makronutrisi dasar: Rendah Gula, Rendah Kalori, dan Tinggi Serat.