RANDOM FOREST

Kelompok 5

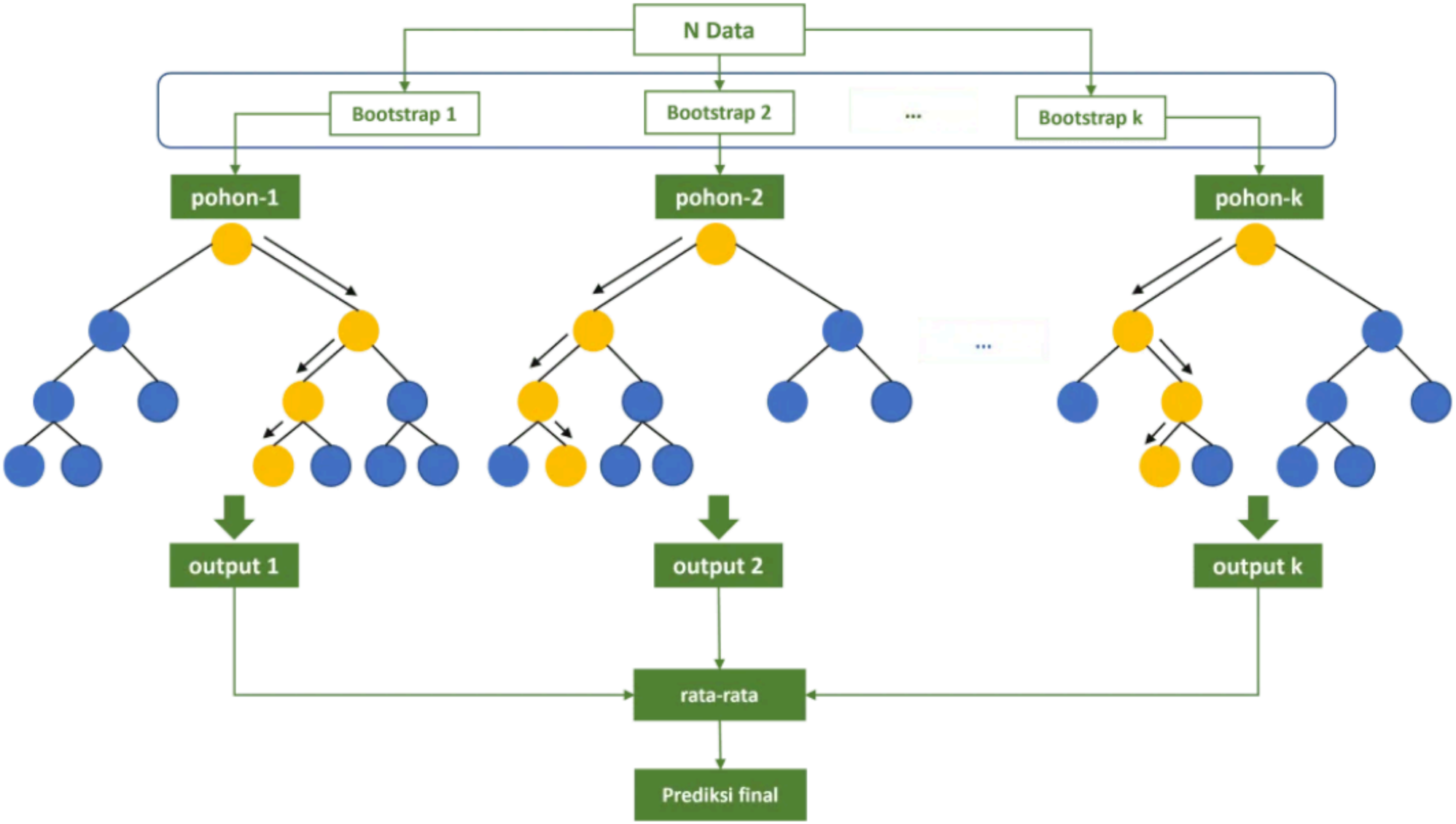
- Arenta Putri (233307034)
- Erlina Putri (233307046)
- Nailah Adyan (233307055)

Apa itu Random Forest?

Random Forest adalah teknik machine learning yang populer dalam bidang data mining. RF bekerja menggunakan pendekatan berbasis kelompok (ensemble). RF dibangun dari banyak decision tree secara acak (random) yang bekerja sama. Setiap decision tree memberikan hasil prediksi, lalu RF akan mengambil vote terbanyak (klasifikasi) / nilai rata-rata (regresi)

Keunggulan RF adalah akurasinya yang tinggi dibanding metode lain

Gambar Random Forest



HIPOTESA FUNCTION

Hipotesa Random Forest, atau Fungsi Hipotesis Akhir adalah cara model menggabungkan (mengagregasi) prediksi dari semua pohon keputusan individual yang ada untuk menghasilkan satu prediksi tunggal yang lebih akurat dan stabil.

Random Forest menggunakan dua metode hipotesis agregasi, tergantung pada jenis masalahnya:

1. Classification
2. Regression

CLASSIFICATION

untuk memprediksi label kategori diskrit (seperti A, B, C atau Ya dan Tidak) menggunakan voting mayoritas atau kelas suara yang paling sering muncul akan menjadi hasil akhir.

$$H_{\text{Klasifikasi}}(x) = \text{Mode}\{h_1(x), h_2(x), \ldots, h_K(x)\}$$

- Hklasifikasi(x): Prediksi akhir yang diberikan oleh Random Forest untuk satu data input.
- x (Data Input): Berbentuk vektor fitur yang ingin diprediksi
- Mode: Fungsi Modus yang digunakan untuk menentukan kelas mayoritas, yaitu nilai yang paling sering muncul.
- hk(x): Prediksi yang dihasilkan oleh pohon keputusan ke-k (dari total K pohon) untuk datax. Setiap pohon memberikan satu label kelas.
- K: Total keseluruhan pohon keputusan yang ada dalam Random Forest untuk membuat prediksi.

PROSES CLASSIFICATION

Tahap Training

- Bootstrap Sampling

  Dataset asli diambil ulang secara acak dengan pengembalian untuk menghasilkan beberapa dataset baru. Setiap dataset ini melatih satu pohon sehingga menghasilkan variasi antar pohon dan model lebih stabil.

- Random Feature Selection

  Pada setiap node, hanya sebagian fitur yang dipilih secara acak sebagai kandidat pemisahan. Hal ini membuat struktur tiap pohon berbeda dan mengurangi korelasi antar pohon.

- Pembangunan Pohon Keputusan

  Setiap pohon dibangun secara independen menggunakan dataset bootstrap. Pohon memilih split terbaik berdasarkan impurity (Gini/Entropy) hingga mencapai batas tertentu.

PROSES CLASSIFICATION

Tahap Prediksi

- Prediksi Individu (hk(x))

  Setiap pohon memproses data baru dan memberikan satu prediksi kelas. Setiap pohon akan memberikan satu hasil prediksi kelas. Perbedaan prediksi terjadi karena tiap pohon dilatih pada data dan fitur acak.

- Pengumpulan Suara

  Setelah semua pohon memberikan prediksinya, semua prediksi dari K pohon dikumpulkan sebagai suara untuk setiap kelas.

- Majority Voting

  Kelas dengan suara terbanyak ditetapkan sebagai hasil akhir prediksi Random Forest.

REGRESSION

untuk memprediksi nilai numerik (seperti harga, suhu, atau biaya) menggunakan rata-rata atau averaging dari semua prediksi pohon.

$$H(x) = \frac{1}{K} \sum_{k=1}^{K} h_k(x)$$

- Hregresi(x): prediksi nilai numerik akhir oleh Random Forest untuk data x.
- K: Jumlah total pohon regresi dalam Random Forest.
- $\sum$: Penjumlahan dari semua prediksi pohon.
- hk(x): Nilai prediksi dari pohon regresi ke-k untuk data x.
- 1/K: Dibagi jumlah pohon atau rata-ratanya.

PROSES REGRESSION

Tahap Training

- Bootstrap Sampling

  Dataset asli diambil ulang secara acak dengan pengembalian untuk membentuk beberapa dataset baru. Setiap dataset digunakan untuk melatih satu pohon regresi agar tiap pohon memiliki keragaman dan mencegah overfitting.

- Random Feature Selection

  Pada setiap node, hanya sebagian fitur yang dipilih secara acak sebagai kandidat pemisahan. Hal ini membuat struktur tiap pohon berbeda dan mengurangi korelasi antar pohon.

- Pembangunan Pohon Keputusan

  Setiap pohon dilatih secara mandiri hingga mencapai batas tertentu (misalnya kedalaman maksimum atau jumlah sampel minimum). Pada setiap split, pohon memilih pemisahan terbaik dengan meminimalkan error, menggunakan Mean Squared Error (MSE) atau Mean Absolute Error (MAE).

PROSES REGRESSION

Tahap Prediksi

- Prediksi oleh Setiap Pohon

  Ketika data baru masuk, setiap pohon menghasilkan satu nilai prediksi berdasarkan leaf node tempat data tersebut berakhir.

- Pengumpulan Nilai Prediksi

  Seluruh prediksi numerik dari K pohon dikumpulkan menjadi satu kumpulan nilai.

- Averaging (Rata-rata)

  Menghasilkan output akhir dengan menghitung rata-rata dari seluruh prediksi pohon, sehingga menghasilkan prediksi yang lebih stabil dan akurat.

COST FUNCTION (BIAYA)

Random Forest tidak ada cost function tunggal yang digunakan untuk seluruh model. Dikarenakan RF adalah ensemble dari banyak decision tree, dan setiap decision tree memiliki cost function-nya sendiri, tergantung jenis tugas yang dilakukan. Cost function digunakan pada setiap node untuk memilih fitur dan titik pemisahan terbaik.

- Klasifikasi → menggunakan Impurity Function
- Regresi → menggunakan Loss Function berbasis error

KLASIFIKASI → IMPURITY FUNCTION

Pada klasifikasi, tujuan pohon keputusan adalah membuat node yang semurni mungkin. Untuk mengukur kemurnian node, ada 2 metrik:

Gini Impurity

Digunakan untuk mengukur seberapa "bercampur" kelas-kelas dalam satu node

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

1 → konstanta
C → jumlah kelas
i → indeks kelas
Pi → probabilitas data (kelas ke-i)
(Pi)2 → probabilitas dipangkatkan 2

Contoh:

- Kelas A 100%, Gini = 0 (paling bagus)
- Kelas A 50%, Kelas B 50%, Gini 0.5 (campur)

Entropy

Digunakan untuk mengukur tingkat ketidakpastian dalam satu node

$$Entropi = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

- → supaya nilai entropy tidak negatif
C → jumlah kelas
i → indeks kelas
Pi → probabilitas kelas ke-i
log2 → logaritma basis 2

Artinya:

- Entropi tinggi (data sangat acak, kelas tercampur)
- Entropi rendah (data lebih murni)

REGRESI → LOSS FUNCTION

Pada Random Forest Regresi, metrik eror untuk mengukur seberapa jauh prediksi pohon terhadap nilai sebenarnya.

Mean Squared Error (MSE)

Digunakan untuk mengukur rata-rata kesalahan kuadrat antara nilai prediksi dan nilai sebenarnya

$$Varians = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

n: jumlah sampel/data
yi: nilai sebenarnya dari data ke-i
y: nilai prediksi dari model

Apabila terdapat outlier, MSE menjadi besar karena error dikuadratkan, sehingga sangat sensitif terhadap nilai ekstream.

Mean Absolute Error (MAE)

Digunakan untuk mengukur rata-rata kesalahan absolute, yaitu selisih mutlak antara nilai prediksi dan nilai sebenarnya

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

n: jumlah sampel/data
yi: nilai sebenarnya dari data ke-i
y: nilai prediksi dari model

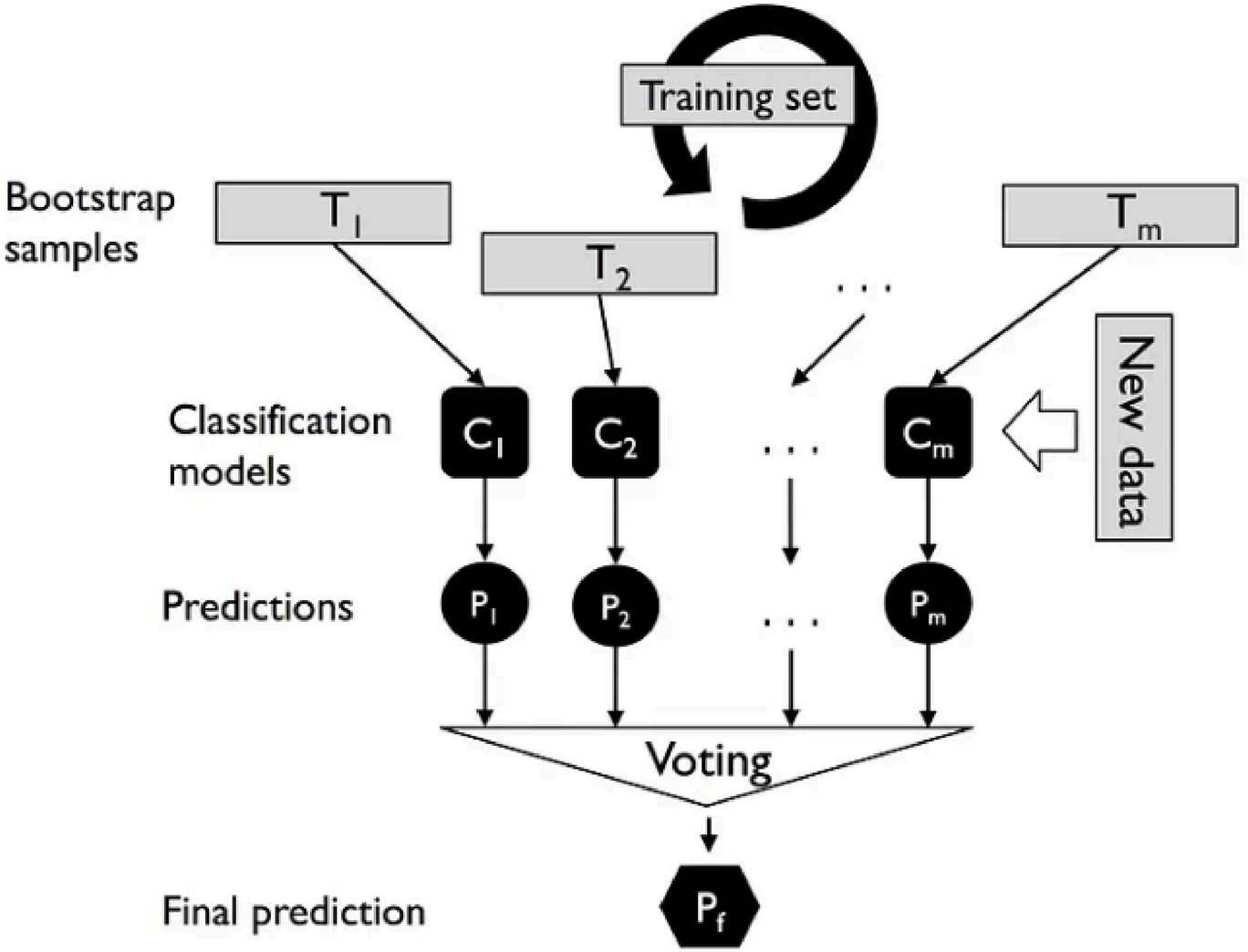
Lebih tahan terhadap outlier dibanding MSE, dan tidak mengukur error besar secara ekstream.

PERHITUNGAN RATA RATA

Impurity	Task	Formula	Description
Gini impurity	Classification	$\sum_{i=1}^{C} f_i(1-f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Entropy	Classification	$\sum_{i=1}^{C} -f_i \log(f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Variance / Mean Square Error (MSE)	Regression	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$
Variance / Mean Absolute Error (MAE) (Scikit-learn only)	Regression	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\mu\rvert$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$

Impurity Formulas used by Scikit-learn and Spark

PERHITUNGAN VOGING (BAGGING)



CONTOH KASUS

Memprediksi harga smartphone berdasarkan 25 fitur spesifikasi, seperti kapasitas baterai, ukuran layar, prosesor, kamera, RAM, dan memori internal. Memiliki nilai harga yang menjadi target prediksi. Studi kasus ini menunjukkan bagaimana informasi spesifikasi dapat dimanfaatkan untuk memperkirakan harga smartphone.

Link Kode :
https://colab.research.google.com/drive/1HKQuNwvNcVQXPb6ZbuhDQeKSUrzY9WW5?usp=sharing

Link Dataset :
https://www.kaggle.com/datasets/chaudharisanika/smartphones-dataset?resource=download

DATASET

brand_name	model	price	rating	has_5g	has_nfc	has_ir_blaster	processor_brand	num_cores	processor_speed	battery_capacity	fast_charging_available	fast_charging	ram_capacity	internal_memory	screen_size
oneplus	OnePlus 11 5G	54999	89	TRUE	TRUE	FALSE	snapdragon	8	3.2	5000	1	100	12	256	6.7
oneplus	OnePlus Nord CE 2 Lite 5G	19989	81	TRUE	FALSE	FALSE	snapdragon	8	2.2	5000	1	33	6	128	6.59
samsung	Samsung Galaxy A14 5G	16499	75	TRUE	FALSE	FALSE	exynos	8	2.4	5000	1	15	4	64	6.6
motorola	Motorola Moto G62 5G	14999	81	TRUE	FALSE	FALSE	snapdragon	8	2.2	5000	1		6	128	6.55
realme	Realme 10 Pro Plus	24999	82	TRUE	FALSE	FALSE	dimensity	8	2.6	5000	1	67	6	128	6.7
samsung	Samsung Galaxy F23 5G (6GB RAM + 128GB)	16999	80	TRUE	TRUE	FALSE	snapdragon	8	2.2	5000	1	25	6	128	6.6
apple	Apple iPhone 14	65999	81	TRUE	TRUE	FALSE	bionic	6	3.22	3279	1		6	128	6.1
xiaomi	Xiaomi Redmi Note 12 Pro Plus	29999	86	TRUE	FALSE	TRUE	dimensity	8	2.6	4980	1	120	8	256	6.67
nothing	Nothing Phone 1	26749	85	TRUE	TRUE	FALSE	snapdragon	8	2.5	4500	1	33	8	128	6.55
oneplus	OnePlus Nord 2T 5G	28999	84	TRUE	TRUE	FALSE	dimensity	8	3	4500	1	80	8	128	6.43
realme	Realme 10 Pro	18999	82	TRUE	FALSE	FALSE	snapdragon	8	2.2	5000	1	33	6	128	6.72
oppo	Oppo A78	18999	79	TRUE	TRUE	FALSE	dimensity	8	2.2	5000	1	33	8	128	6.56
xiaomi	Xiaomi Redmi Note 12 Pro 5G	24762	79	TRUE	FALSE	TRUE	dimensity	8	2.6	5000	1	67	6	128	6.67
vivo	Vivo T1 5G (6GB RAM + 128GB)	16990	80	TRUE	FALSE	FALSE	snapdragon	8	2.2	5000	1	18	6	128	6.58
samsung	Samsung Galaxy S23 Ultra 5G	114990		TRUE	TRUE	FALSE	snapdragon	8	3.2	5000	1	45	8	256	6.8
apple	Apple iPhone 13	62999	79	TRUE	TRUE	FALSE	bionic	6	3.22	3240	1		4	128	6.1
vivo	Vivo Y16	9999	65	FALSE	FALSE	FALSE	helio	8	2.3	5000	1	10	3	32	6.51
oppo	OPPO Reno 9 Pro Plus	45999	86	TRUE	TRUE	FALSE	snapdragon	8	3.2	4700	1	80	16	256	6.7
oneplus	OnePlus 10R 5G	32999	86	TRUE	TRUE	FALSE	dimensity	8	2.85	5000	1	80	8	128	6.7
vivo	Vivo Y22	14499	72	FALSE	FALSE	FALSE	helio	8	2	5000	1	18	4	64	6.55
oneplus	OnePlus 11R	39999	85	TRUE	TRUE	FALSE	snapdragon	8	3.2	5000	1	100	8	128	6.7
vivo	Vivo V25 Pro 5G	35999	85	TRUE	FALSE	FALSE	dimensity	8	3	4830	1	66	8	128	6.56
poco	Poco X4 Pro 5G	14999	80	TRUE	FALSE	TRUE	snapdragon	8	2.2	5000	1	67	6	64	6.67
xiaomi	Xiaomi Redmi Note 12	17859	76	TRUE	FALSE	TRUE	snapdragon	8	2	5000	1	33	4	128	6.67