

Laporan Ujian Tengah Semester

Machine Learning

Disusun untuk memenuhi Ujian Tengah Semester mata kuliah Machine Learning



Oleh:

HAPIZ ILHAM MAULANA 222310041

Kelas:

TI-21-KA

PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS INFORMATIKA DAN PARIWISATA
INSTITUT BISNIS DAN INFORMATIKA KESATUAN
BOGOR
2024

DAFTAR ISI

DAFTAR ISI.....	i
ATRIBUT	1
1.1 Prediktor	1
1.2 Label	1
STATISTIK DESKRIPTIF	2
2.1 Sebelum Pengisian Missing Values dan Standarisasi	2
2.2 Setelah Pengisian Missing Values dan Standarisasi	3
Model Klasifikasi Decision Tree.....	5
Model Clustering K-Means.....	6

ATRIBUT

1.1 Prediktor

Berikut adalah daftar atribut beserta jenis atribut yang terdapat dalam dataset:

No.	Nama Atribut	Jenis Atribut	Tipe Atribut	Deskripsi
1	battery_power	Numerik	<i>Continuous</i>	-
2	blue	Kategorik	<i>Binary</i>	Memiliki Nilai 0 dan 1
3	clock_speed	Numerik	<i>Continuous</i>	-
4	dual_sim	Kategorik	<i>Binary</i>	Memiliki Nilai 0 dan 1
5	fc	Numerik	<i>Continuous</i>	-
6	four-g	Kategorik	<i>Binary</i>	Memiliki Nilai 0 dan 1
7	int_memory	Numerik	<i>Continuous</i>	-
8	m_dep	Numerik	<i>Continuous</i>	-
9	mobile_wt	Numerik	<i>Continuous</i>	-
10	n_cores	Numerik	<i>Continuous</i>	-
11	pc	Numerik	<i>Continuous</i>	-
12	px_height	Numerik	<i>Continuous</i>	-
13	px_width	Numerik	<i>Continuous</i>	-
14	ram	Numerik	<i>Continuous</i>	-
15	sc_h	Numerik	<i>Continuous</i>	-
16	sc_w	Numerik	<i>Continuous</i>	-
17	talk_time	Numerik	<i>Continuous</i>	-
18	three_g	Kategorik	<i>Binary</i>	Memiliki Nilai 0 dan 1
19	touch_screen	Kategorik	<i>Binary</i>	Memiliki Nilai 0 dan 1
20	wifi	Kategorik	<i>Binary</i>	Memiliki Nilai 0 dan 1

1.2 Label

Kerja sama merupakan tindakan kolaboratif antara dua atau lebih entitas, baik individu, perusahaan, maupun negara, untuk mencapai tujuan yang sama. Dalam konteks bisnis,

No.	Nama Atribut	Jenis Atribut	Tipe Atribut	Deskripsi
1	price_range	Kategorik		Memiliki Nilai 0, 1, 2, 3

STATISTIK DESKRIPTIF

2.1 Sebelum Pengisian Missing Values dan Standarisasi

Statistik deskriptif merupakan cabang dari statistik untuk mendeskripsikan dan merangkum data. Contoh hal umum yang biasa dilakukan di dalam tipe statistik ini seperti pembuatan graph, dan menghitung berbagai macam pengukuran data seperti Mean. Sebelum dilakukan praproses data, pada dataset masih terdapat missing values yang perlu diisi

Statistik Deskriptif Sebelum Preprocessing:

	battery_power	blue	clock_speed	dual_sim	fc \
count	1990.000000	2000.0000	2000.000000	2000.000000	2000.000000
mean	1237.867839	0.4950	1.522250	0.509500	4.309500
std	439.676025	0.5001	0.816004	0.500035	4.341444
min	501.000000	0.0000	0.500000	0.000000	0.000000
25%	850.250000	0.0000	0.700000	0.000000	1.000000
50%	1225.000000	0.0000	1.500000	1.000000	3.000000
75%	1615.000000	1.0000	2.200000	1.000000	7.000000
max	1998.000000	1.0000	3.000000	1.000000	19.000000

	four_g	int_memory	m_dep	mobile_wt	n_cores \
count	2000.000000	1990.000000	2000.000000	1990.000000	2000.000000
mean	0.521500	31.987940	0.501750	140.344221	4.520500
std	0.499662	18.136427	0.288416	35.407114	2.287837
min	0.000000	2.000000	0.100000	80.000000	1.000000
25%	0.000000	16.000000	0.200000	109.000000	3.000000
50%	1.000000	32.000000	0.500000	141.000000	4.000000
75%	1.000000	48.000000	0.800000	170.000000	7.000000
max	1.000000	64.000000	1.000000	200.000000	8.000000

	pc	px_height	px_width	ram	sc_h \
count	2000.000000	2000.000000	2000.000000	1990.000000	2000.000000
mean	9.916500	645.108000	1251.515500	2124.991960	12.306500
std	6.064315	443.780811	432.199447	1084.885362	4.213245
min	0.000000	0.000000	500.000000	256.000000	5.000000
25%	5.000000	282.750000	874.750000	1208.250000	9.000000
50%	10.000000	564.000000	1247.000000	2146.500000	12.000000
75%	15.000000	947.250000	1633.000000	3065.500000	16.000000
max	20.000000	1960.000000	1998.000000	3998.000000	19.000000

	sc_w	talk_time	three_g	touch_screen	wifi
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	5.767000	11.011000	0.761500	0.503000	0.507000
std	4.356398	5.463955	0.426273	0.500116	0.500076
min	0.000000	2.000000	0.000000	0.000000	0.000000
25%	2.000000	6.000000	1.000000	0.000000	0.000000
50%	5.000000	11.000000	1.000000	1.000000	1.000000
75%	9.000000	16.000000	1.000000	1.000000	1.000000
max	18.000000	20.000000	1.000000	1.000000	1.000000

2.2 Setelah Pengisian Missing Values dan Standarisasi

Pada praproses data yang dilakukan, yaitu memisahkan bagian atribut prediktor dengan label data menjadi dua variabel terpisah, dengan memisahkan bagian atribut prediktor dilakukan dengan pertama-tama membuat dua variabel, misalnya variabel X merupakan variabel yang berisi atribut prediktor yaitu atribut selain atribut price_range sehingga kita menghapus kolom price range pada variabel x. Untuk variabel Y ditetapkan kolom price_range yang digunakan sebagai atribut label.

Dari dataset yang didapatkan terdapat beberapa missing values. Untuk mengisi missing values dilakukan dengan menggunakan SimpleImputer dan nilai mean sebagai strategi pengisian nilai dengan inisialisasi SimpleImputer dan menetapkan strategi mean yang kemudian melakukan imputasi nilai dengan SimpleImputer yang telah diinisialisasikan yang dilanjutkan dengan menyeragamkan nilai dari masing – masing atribut dengan menggunakan metode StandardScaler

Statistik Deskriptif Setelah Pengisian Missing Values dan Standarisasi:

	battery_power	blue	clock_speed	dual_sim	fc
count	2.000000e+03	2.000000e+03	2.000000e+03	2.000000e+03	2.000000e+03
mean	-1.563194e-16	-1.243450e-17	-1.545430e-16	8.082424e-17	5.861978e-17
std	1.000250e+00	1.000250e+00	1.000250e+00	1.000250e+00	1.000250e+00
min	-1.680562e+00	-9.900495e-01	-1.253064e+00	-1.019184e+00	-9.928904e-01
25%	-8.806122e-01	-9.900495e-01	-1.007906e+00	-1.019184e+00	-7.624947e-01
50%	-1.680370e-02	-9.900495e-01	-2.727384e-02	9.811771e-01	-3.017032e-01
75%	8.578380e-01	1.010051e+00	8.307794e-01	9.811771e-01	6.198797e-01
max	1.733620e+00	1.010051e+00	1.811412e+00	9.811771e-01	3.384628e+00

	four_g	int_memory	m_dep	mobile_wt	n_cores
count	2.000000e+03	2.000000e+03	2.000000e+03	2.000000e+03	2.000000e+03
mean	1.048051e-16	5.684342e-17	-1.030287e-16	-3.765876e-16	-7.727152e-17
std	1.000250e+00	1.000250e+00	1.000250e+00	1.000250e+00	1.000250e+00
min	-1.043966e+00	-1.658030e+00	-1.393304e+00	-1.709003e+00	-1.539175e+00
25%	-1.043966e+00	-8.839717e-01	-1.046495e+00	-8.876966e-01	-6.647678e-01
50%	9.578860e-01	0.000000e+00	-6.069151e-03	1.857225e-02	-2.275644e-01
75%	9.578860e-01	8.853054e-01	1.034357e+00	8.398784e-01	1.084046e+00
max	9.578860e-01	1.769944e+00	1.727974e+00	1.689505e+00	1.521249e+00

	pc	px_height	px_width	ram	sc_h
count	2.000000e+03	2.000000e+03	2.000000e+03	2.000000e+03	2.000000e+03
mean	1.403322e-16	1.181277e-16	6.084022e-17	-1.492140e-16	4.884981e-17
std	1.000250e+00	1.000250e+00	1.000250e+00	1.000250e+00	1.000250e+00
min	-1.635631e+00	-1.454027e+00	-1.739251e+00	-1.727512e+00	-1.734608e+00
25%	-8.109291e-01	-8.167289e-01	-8.719579e-01	-8.450354e-01	-7.849833e-01
50%	1.377252e-02	-1.828116e-01	-1.045034e-02	5.091091e-03	-7.276497e-02
75%	8.384742e-01	6.810064e-01	8.828792e-01	8.642295e-01	8.768595e-01
max	1.663176e+00	2.963672e+00	1.727608e+00	1.731225e+00	1.589078e+00

	sc_w	talk_time	three_g	touch_screen	wifi
count	2.000000e+03	2.000000e+03	2.000000e+03	2.000000e+03	2.000000e+03
mean	-5.506706e-17	1.421085e-16	1.421085e-17	-5.417888e-17	1.421085e-17
std	1.000250e+00	1.000250e+00	1.000250e+00	1.000250e+00	1.000250e+00
min	-1.324131e+00	-1.649584e+00	-1.786861e+00	-1.006018e+00	-1.014099e+00
25%	-8.649215e-01	-9.173306e-01	5.596406e-01	-1.006018e+00	-1.014099e+00
50%	-1.761069e-01	-2.013697e-03	5.596406e-01	9.940179e-01	9.860966e-01
75%	7.423125e-01	9.133032e-01	5.596406e-01	9.940179e-01	9.860966e-01
max	2.808756e+00	1.645557e+00	5.596406e-01	9.940179e-01	9.860966e-01

Model Klasifikasi Decision Tree

Pada Analisa ini menggunakan algoritma Decision Tree sebagai model klasifikasi untuk analisis ini. Algoritma Decision Tree merupakan metode pembelajaran mesin yang digunakan untuk mengelompokkan data dengan mengandalkan serangkaian keputusan hierarkis yang terbentuk dalam struktur pohon. Keputusan-keputusan ini dibuat berdasarkan atribut dataset, yang membantu memprediksi label atau kategori tertentu.

Model Decision Tree dilatih menggunakan data pelatihan sebesar 85% dari dataset, dengan pengaturan `random_state = 17` untuk memastikan hasil yang dapat direproduksi. Selanjutnya, model tersebut digunakan untuk melakukan prediksi pada data pengujian (15%). Akurasi model dihitung untuk mengukur tingkat keberhasilan dalam memprediksi kategori harga smartphone. Hasil akurasi yang diperoleh adalah sebesar 85.7%. Akurasi merupakan metrik yang mengukur sejauh mana model mampu memprediksi kategori yang benar.

```
Accuracy: 0.8566666666666667
Confusion Matrix:
[[66  8  0  0]
 [ 7 61  5  0]
 [ 1  8 63  4]
 [ 0  0 10 67]]
Classification Report:

```

	precision	recall	f1-score	support
0	0.89	0.89	0.89	74
1	0.79	0.84	0.81	73
2	0.81	0.83	0.82	76
3	0.94	0.87	0.91	77
accuracy			0.86	300
macro avg	0.86	0.86	0.86	300
weighted avg	0.86	0.86	0.86	300

Model Clustering K-Means

Dalam analisis ini, menggunakan algoritma K-Means untuk mengelompokkan data dalam analisis ini. Algoritma ini adalah salah satu metode clustering yang dimaksudkan untuk menggabungkan data ke dalam berbagai kelompok berdasarkan kesamaan karakteristik tertentu. Silhouette Coefficients digunakan untuk mengevaluasi kualitas pengelompokan data dalam cluster dengan berbagai jumlah cluster yang berbeda. Hasil evaluasi dengan Silhouette Coefficients adalah sebagai berikut :

Silhouette Score : 0.05746542779025056, dengan jumlah cluster 4