Tahukah kamu?

5.83%

Tingkat pengangguran di Indonesia*

Faktor: Keterbatasan akses informasi lowongan pekerjaan

*Februari 2022, sumber: BPS

LAPORAN
DATA ANALYTICS COMPETITION
FIND IT 2022

TIM SIPENDING

Ang, Johan Nicholas
Habiburrohman

Latar Belakang

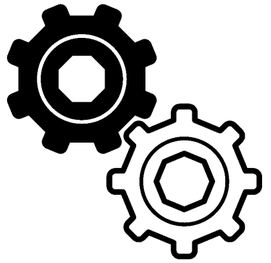
Tujuan dan Manfaat

Metodologi

Data Lowongan Pekerjaan

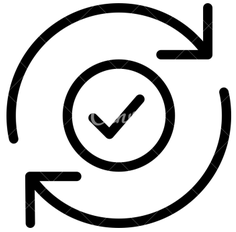
Exploratory Data Analysis (EDA)

Data Preprocessing



Feature Engineering

Modeling



Validation

EDA

01 Eksplorasi Missing Values

02 Eksplorasi Data Duplikat

03 Visualisasi dan Analisis Statistik

Data Preprocessing

01 Menangani Missing Values

02 Menangani Data Duplikat

03 Menghapus Pencilan dan Noise

Feature Engineering

01 Menambah Fitur

02 Ordinal Encoding

03 Membuang Fitur Berdasarkan Tingkat Kepentingan

Modeling dan Validasi

01 Validation & Model Screening

02 Hyper-parameter Tuning

03 Model Ensembling

Analisis

Data Latih

31,746 baris
15 fitur

Data Uji

3,000 baris
14 fitur

	job_title	job_function	education_level
0	Facility Maintenance & Smart Warehouse Manager	Manufaktur,Pemeliharaan	Sertifikat Professional, D3 (Diploma), D4 (Dip...
1	Procurement Department Head	Manufaktur,Pembelian/Manajemen Material	Sarjana (S1), Diploma Pascasarjana, Gelar Prof...
2	SALES ADMIN	Penjualan / Pemasaran,Penjualan Ritel	Sarjana (S1)
3	City Operation Lead Shopee Express (Cirebon)	Pelayanan,Logistik/Rantai Pasokan	Sarjana (S1), Diploma Pascasarjana, Gelar Prof...
4	Japanese Interpreter	Lainnya,Jurnalis/Editor	Sertifikat Professional, D3 (Diploma), D4 (Dip...

Missing pada Fitur Target

80%

Missing values pada fitur 'salary'

Setelah pembersihan

6,532 baris

Data Duplikat

1,189 baris

Frekuensi data duplikat

Setelah pembersihan

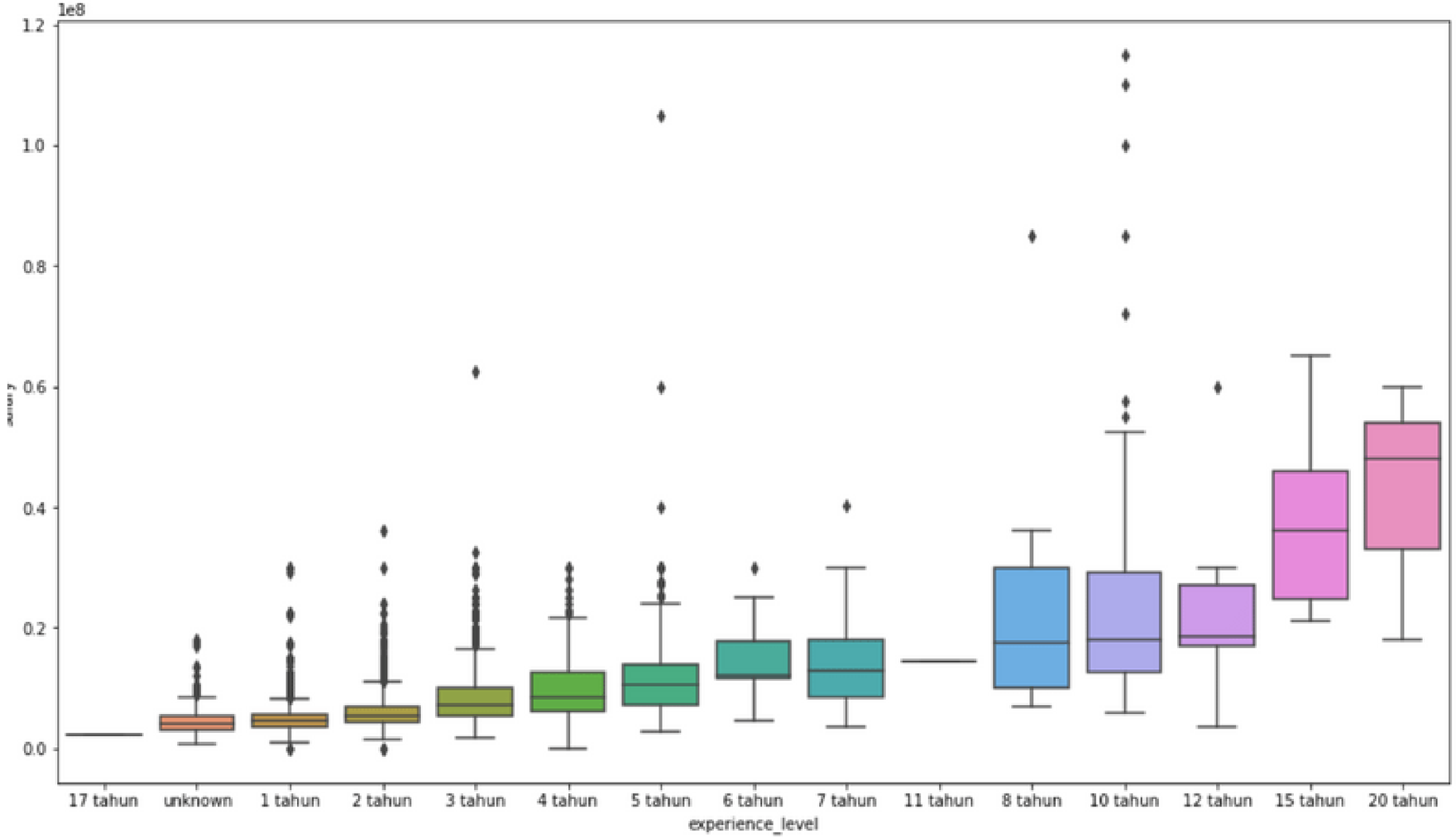
5,343 baris

Fitur dengan Missing Values

	nan %
experience_level	8
job_benefits	24
company_process_time	34
company_size	16
company_industry	3

5 fitur dengan missing values

experience_level



job_benefits

	job_benefits	frekuensi
0	unknown	2043
1	asuransi kesehatan;waktu regular, senin - juma...	458
2	asuransi kesehatan;waktu regular, senin - juma...	252
3	waktu regular, senin - jumat;bisnis (contoh: k...	231
4	tip;asuransi kesehatan;waktu regular, senin - ...	148
...	...	...
1553	asuransi gigi;tunjangan pendidikan;tip;asurans...	1
1554	tip;senin - sabtu	1
1555	asuransi gigi;waktu regular, senin - jumat;kas...	1
1556	tip;asuransi kesehatan;pinjaman;parkir;penglih...	1
1557	asuransi kesehatan;waktu regular, senin - juma...	1

Kendala...

Frekuensi kelas unik yang terlalu beragam dan missing values yang merupakan mayoritas membuat proses imputing menjadi sulit.

Solusi: Drop!

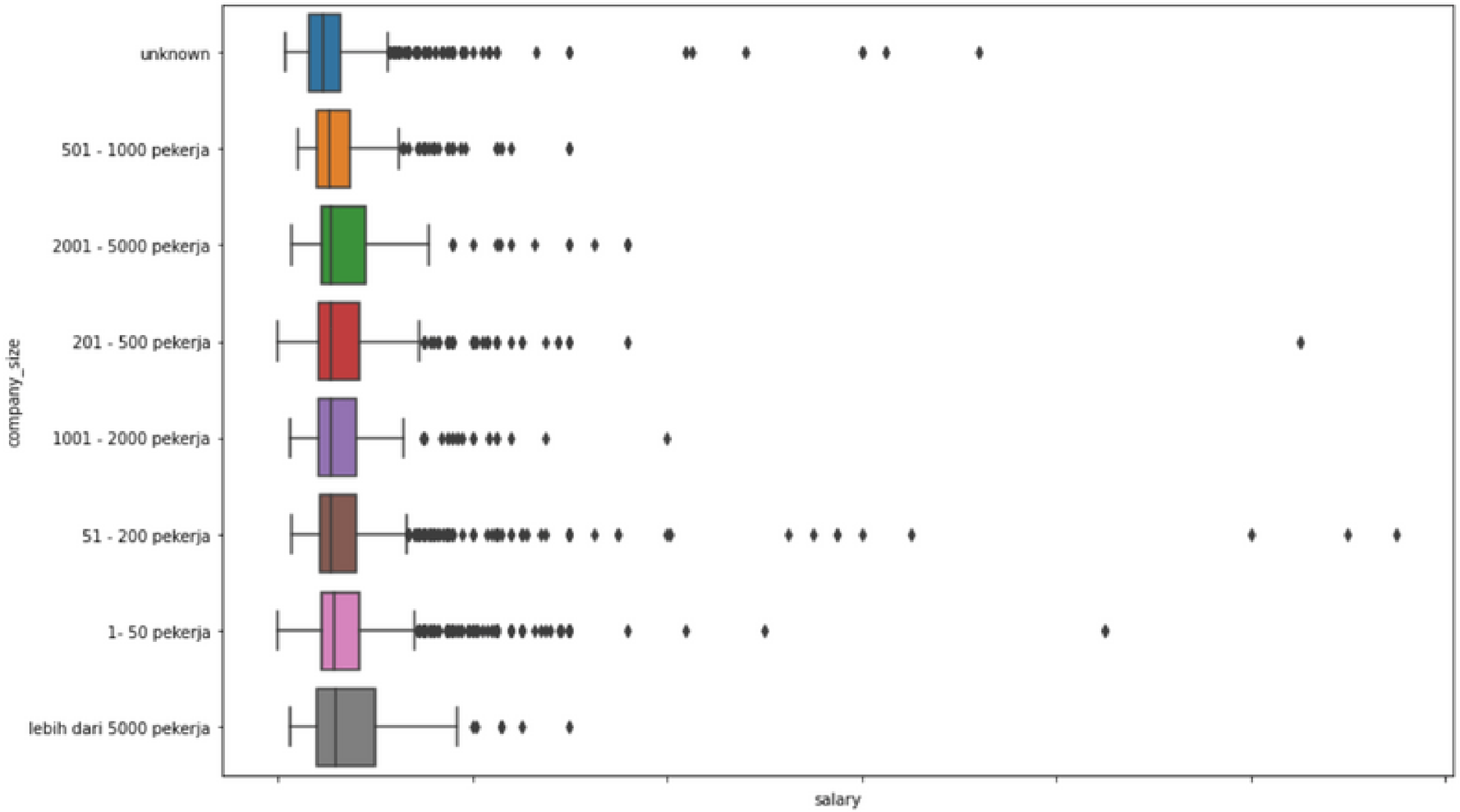
company_process_time

	company_process_time	frekuensi
0	unknown	2943
1	29 days	993
2	28 days	586
3	27 days	431
4	26 days	312

Insight

- Missing values merupakan mayoritas
- 31 kelas unik
- Kelas-kelas mewakili semua hari dalam sebulan
- Tidak ada kelas lebih dari 30 hari

company_size



company_industry

	company_industry	frekuensi
0	manufaktur/produksi	704
1	umum & grosir	596
2	retail/merchandise	581
3	makanan & minuman/katering/restoran	580
4	manajemen/konsulting hr	535
5	komputer/teknik informatika (perangkat lunak)	507
6	perbankan/pelayanan keuangan	422
7	lainnya	298
8	konstruksi/bangunan/teknik	291
9	unknown	274
56	jurnalisme	2
57	r&d	1
58	tembakau	1

Insight

- Missing values terbanyak ke-10
- 59 kelas unik
- Terdapat kelas 'lainnya'

Analisis Fitur Tanpa Missing Values

job_title
location
salary_currency
career_level
education_level
employment_type
job_function
job_description

8 fitur tanpa missing values

job_title

	job_title	frekuensi
0	sales executive	142
1	sales	72
2	digital marketing	52
3	graphic designer	51
4	sales engineer	40
...	...	...
5562	distribution & collection staff - jakarta	1
5563	marketing - sales	1
5564	e-commerce merchandiser	1
5565	legal assistant manager - pontianak	1
5566	credit marketing officer (cmo) - tangerang & c...	1

Kendala...

Frekuensi kelas unik yang terlalu beragam: 5567 kelas

Solusi: Drop!

location

	location	frekuensi
0	jakarta raya	1644
1	jakarta selatan	703
2	jakarta barat	656
3	tangerang	543
4	jakarta utara	531
...	...	...
168	minahasa	1
169	batu	1
170	kepulauan riau	1
171	purworejo	1
172	papua barat	1

Insight

- 173 kelas unik
- Terdapat kelas yang mewakili provinsi
- Terdapat kelas yang mewakili kabupaten/kota

salary_currency

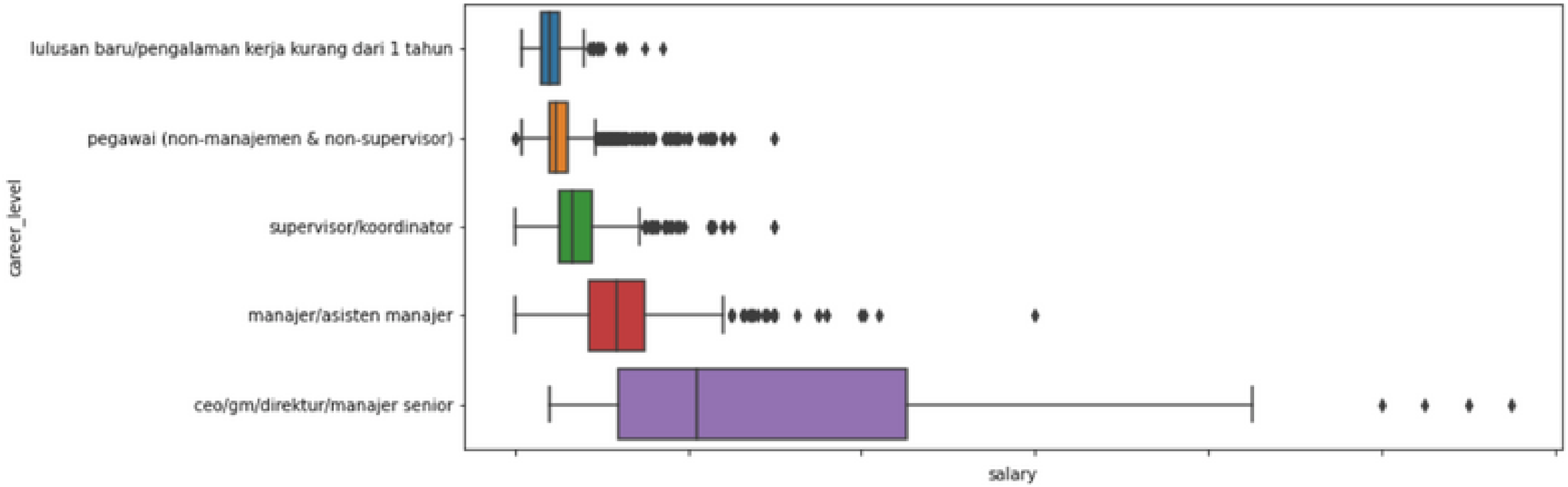
	salary_currency	frekuensi
0	idr	8533
1	usd	2

Insight

- 2 kelas unik
- Kelas 'usd' hanya ada di data latih

Solusi: Drop 'usd', lalu drop fitur

career_level



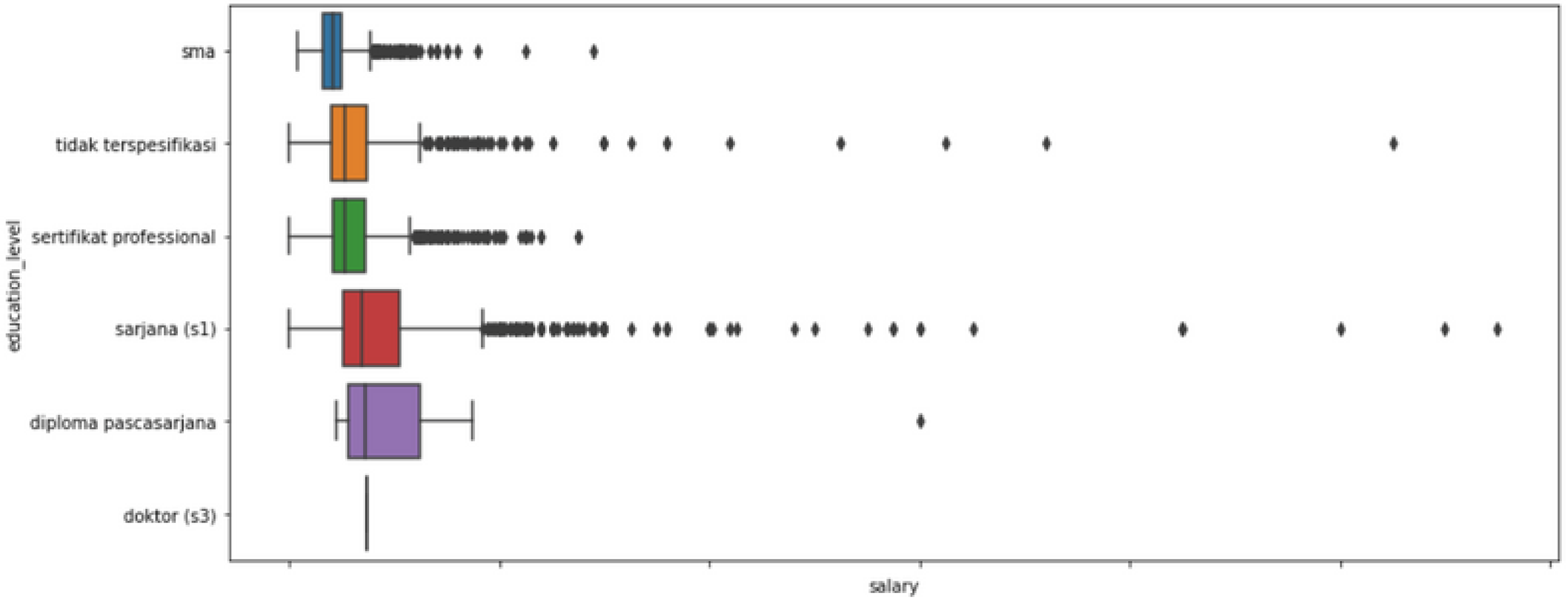
education_level

	education_level	frekuensi
0	sarjana (s1)	2470
1	tidak terspesifikasi	1527
2	sertifikat professional, d3 (diploma), d4 (diploma), sarjana (s1)	1518
3	sma, smu/smk/stm, sertifikat professional, d3 (diploma), d4 (diploma), sarjana (s1)	898
4	sma, smu/smk/stm	562
5	sertifikat professional, d3 (diploma), d4 (diploma)	441
6	sarjana (s1), diploma pascasarjana, gelar professional, magister (s2)	432
7	sertifikat professional, d3 (diploma), d4 (diploma), sarjana (s1), diploma pascasarjana, gelar professional, magister (s2)	344
8	sma, smu/smk/stm, sertifikat professional, d3 (diploma), d4 (diploma)	184
9	sma, smu/smk/stm, sarjana (s1)	67

Insight

- Perhatikan jenjang pendidikan paling awal pada tiap kelas

after splitting



Hasil

- Dari 19 kelas unik menjadi 6 kelas unik dengan 'sarjana (s1)' sebagai modus

employment_type

	employment_type	frekuensi
0	penuh waktu	7291
1	kontrak	1100
2	paruh waktu	74
3	magang	35
4	temporer	24
5	penuh waktu, kontrak	6
6	penuh waktu, paruh waktu	2
7	kontrak, temporer	2
8	penuh waktu, magang	1

Solusi: Generalisasi!

job_function

	job_function	frekuensi
0	penjualan / pemasaran,penjualan ritel	885
1	komputer/teknologi informasi,it-perangkat lunak	741
2	akuntansi / keuangan,akuntansi umum / pembiayaan	633
3	penjualan / pemasaran,pemasaran/pengembangan b...	526
4	sumber daya manusia/personalia,sumber daya man...	365
5	sumber daya manusia/personalia,staf / administ...	321
6	penjualan / pemasaran,penjualan - korporasi	320
7	penjualan / pemasaran,digital marketing	315
8	seni/media/komunikasi,seni / desain kreatif	280
9	hotel/restoran,makanan/minuman/pelayanan restoran	265
10	manufaktur,pembelian/manajemen material	219
11	penjualan / pemasaran,penjualan - jasa keuangan	198

Insight

- Perhatikan bagian sebelum koma pada tiap kelas

after splitting

	job_function	frekuensi
0	penjualan / pemasaran	2707
1	akuntansi / keuangan	984
2	komputer/teknologi informasi	968
3	sumber daya manusia/personalia	785
4	manufaktur	552
5	pelayanan	451
6	seni/media/komunikasi	398
7	teknik	382
8	bangunan/konstruksi	354
9	hotel/restoran	290
10	lainnya	217
11	pendidikan/pelatihan	175
12	layanan kesehatan	171
13	sains	96

Hasil

- Dari 67 kelas unik menjadi 14 kelas unik

job_description

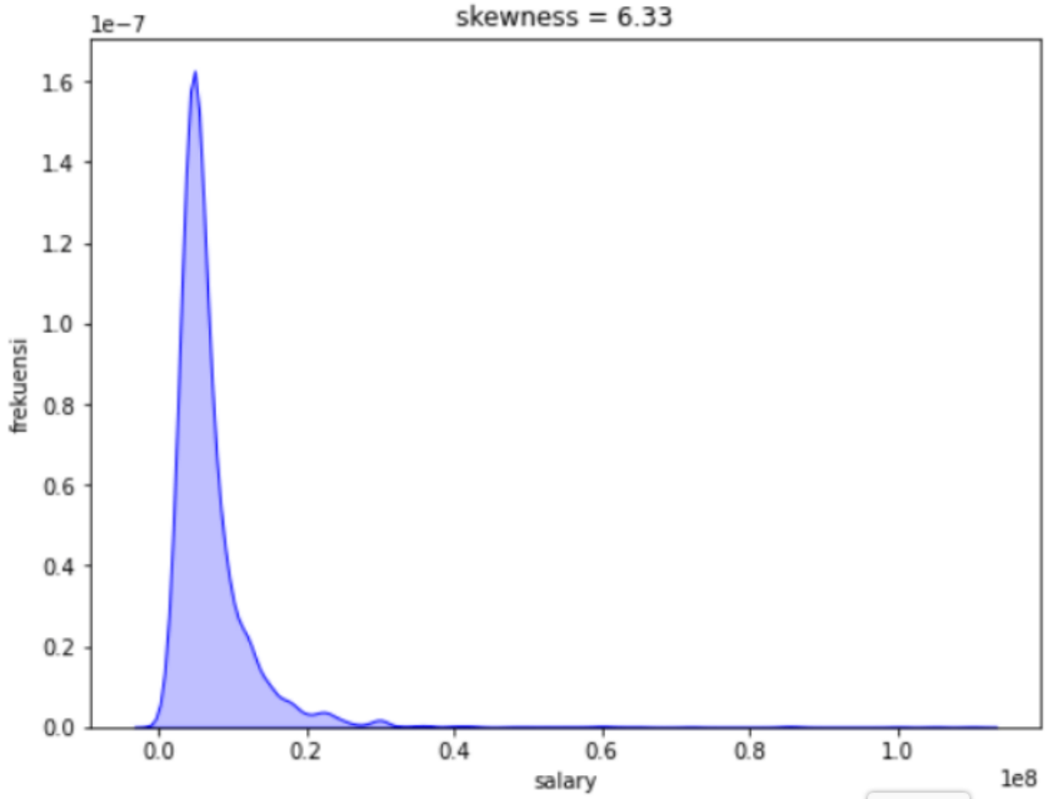
	job_description	frekuensi
0	kualifikasi:berpenampilan menarik & rapihbisa ...	21
1	# terapis atau beautician berpengalaman# magan...	13
2	deskripsi pekerjaan melakukan kunjungan langsu...	13
3	# must managerial skills in above field# must ...	12
4	summaryas an area business development associa...	10
...	...	...
7772	kualifikasipendidikan min.d3/s1 (semua jurusan...	1
7773	performs thorough maintenance and repair works...	1
7774	kualifikasi:- memiliki pendidikan minimal d3/s...	1
7775	posisi pekerjaan :field collectionkualifikasi ...	1
7776	cmo motor barumelakukan penjualan produk pembi...	1

Kendala...

Frekuensi kelas unik yang terlalu beragam: 7777 kelas

Solusi: Drop!

Analisis Fitur Target



Deskripsi Kuartil

min	10
25%	4,250,000
50%	5,500,000
75%	8,000,000
max	110,000,000

Analisis Fitur Target

id	salary
1077	10
4034	10
4265	5300

Insight

- Terdapat noise pada fitur target
- 'salary' kurang dari 500,000

Solusi: Drop!

Feature Engineering

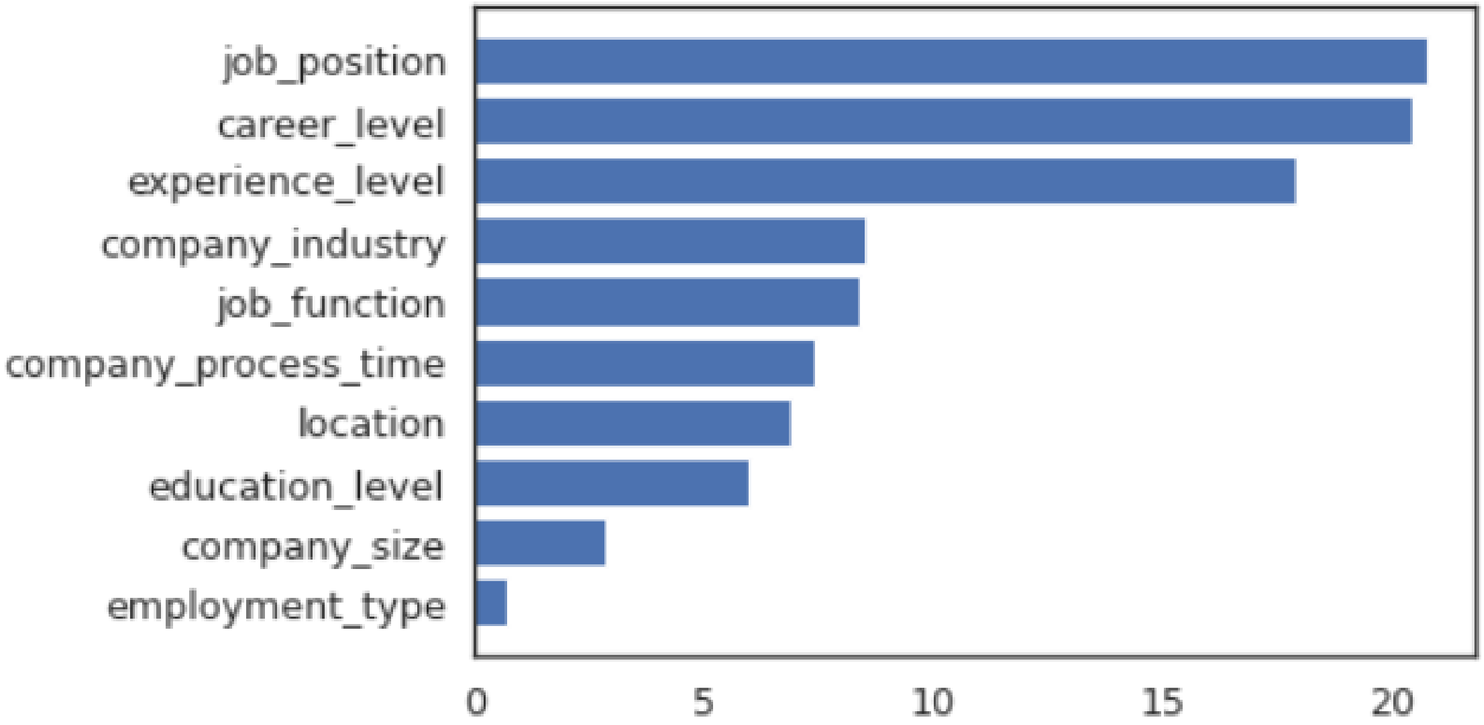
	job_position	frekuensi
0	pegawai (non-manajemen & non-supervisor) penjualan / pemasaran	1664
1	pegawai (non-manajemen & non-supervisor) komputer/teknologi informasi	697
2	pegawai (non-manajemen & non-supervisor) akuntansi / keuangan	549
3	pegawai (non-manajemen & non-supervisor) sumber daya manusia/personalia	492
4	manajer/asisten manajer penjualan / pemasaran	406

Fitur baru 'job_position'

- Gabungan antara 'career_level' dan 'job_function'
- Memberi informasi mengenai nama atau posisi pekerjaan

Feature Engineering

Encoding menggunakan ordinal encoding, lalu analisis tingkat kepentingan fitur



Eksperimen

Diperoleh bahwa data dengan top 9 fitur menghasilkan akurasi terbaik

Hasil Preprocessing dan Feature Engineering

Dimensi Data Latih

5100 baris

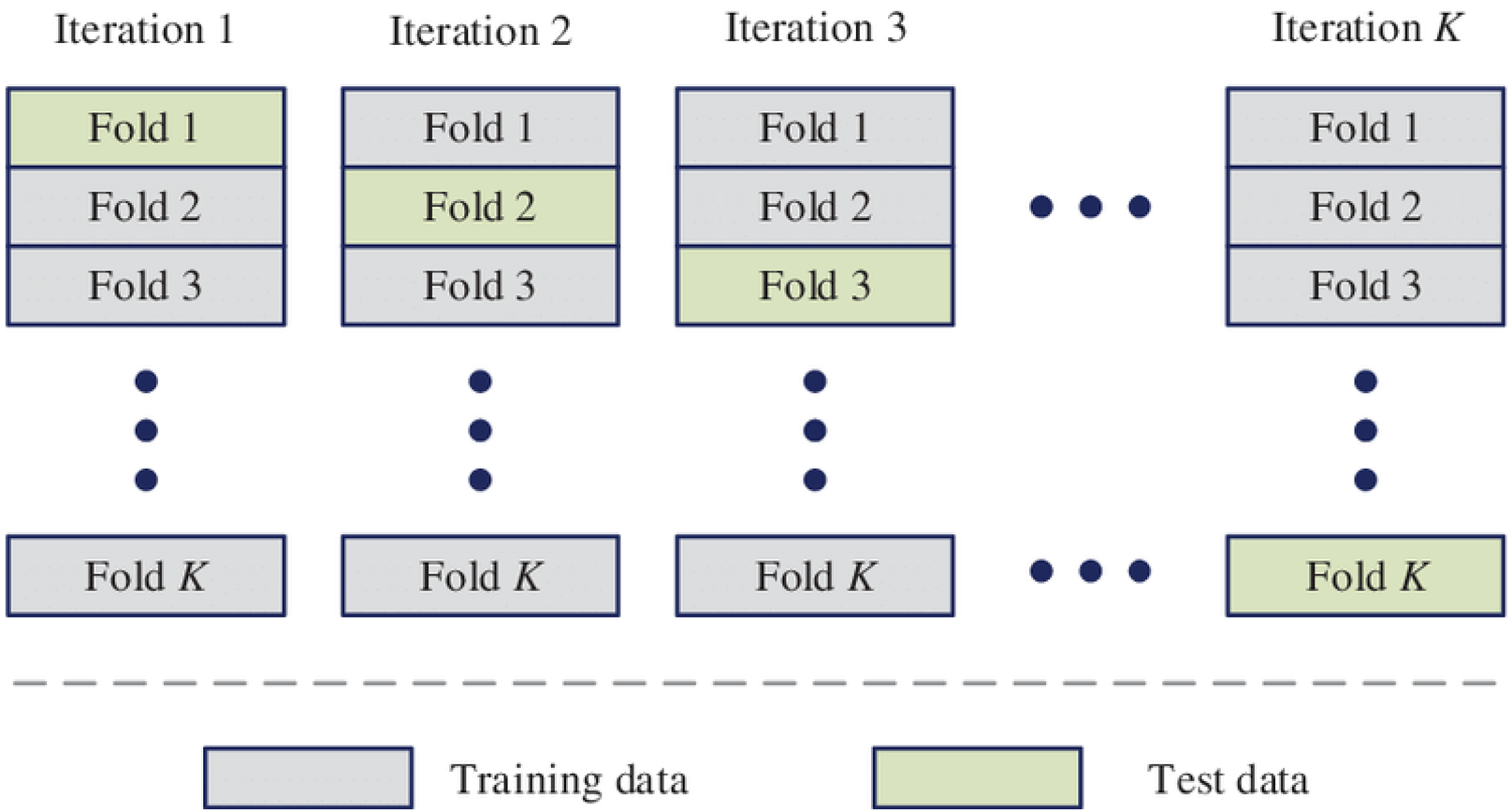
9 fitur independen

1 fitur target

Modeling dan Validasi

Validasi

K-Fold Cross-Validation dengan K = 10



sumber: Li 2019

Screening

Metrik root-mean-squared error (RMSE)

semakin mendekati 0 semakin baik

Model	RMSE	
CatBoost	4,014,913	✓
XGBoost	4,061,446	✓
Random Forest	4,251,466	
Decision Tree	5,479,650	

Hyperparameter Tuning

mencari kombinasi parameter terbaik bagi model-model pilihan

terjadi peningkatan akurasi!

Kondisi	Model	RMSE
Sebelum *tuning*	CatBoost	4,014,913
	XGBoost	4,061,446
Sesudah *tuning*	CatBoost	3,898,224
	XGBoost	3,783,793

Ensembling

Menggunakan metode voting untuk menggabungkan model XGBoost dan CatBoost

Terjadi peningkatan akurasi!

Model	RMSE
Voting	3,780,258
XGBoost	3,783,793
CatBoost	3,898,224

Kesimpulan

Model ensemble berbasis voting antara XGBoost dan Catboost mampu memprediksi gaji data lowongan pekerjaan dengan nilai RMSE sebesar 3,780,258 pada cross-validation dan 36,599,964 pada leaderboard.

Terima kasih!