LAPORAN
DATA ANALYTICS COMPETITION
FIND IT! 2022

Tim Sipending

Disusun Oleh :

Ang, Johan Nicholas	**16521232**
Habiburrohman	**16021161**

Kota Bandung

2022

1. LATAR BELAKANG

Era digital ditandai dengan pesatnya perkembangan teknologi informasi. Berbagai data dan informasi, seperti lowongan pekerjaan, dapat diperoleh dengan mudah melalui internet. Hal ini menjadi peluang baik bagi para pencari kerja. Seiring berjalannya waktu, jumlah data semakin banyak baik dari segi jumlahnya maupun fiturnya. Tidak menutup kemungkinan seseorang akan bingung dalam memilih pekerjaan karena data yang dipertimbangkan sangat banyak. Selain itu, mungkin terdapat ketidaklengkapan data. Akibatnya, informasi penting dalam data tersebut menjadi sulit ditangkap sehingga perlu waktu yang lebih lama untuk mengambil keputusan. Oleh karena itu, dibutuhkan suatu metode analisis yang mampu mengatasi masalah tersebut.

Dengan memanfaatkan perkembangan teknologi, masalah sangat mungkin dipecahkan dengan menggunakan metode *predictive analytics*. Data lowongan pekerjaan dapat dianalisis dan diolah secara mendalam sehingga mampu memprediksi gaji yang ditawarkan. Dengan demikian, diharapkan seseorang dapat memilih pekerjaan yang cocok dengan lebih mudah.

Berdasarkan latar belakang tersebut, kami menawarkan solusi berupa analisis data lowongan pekerjaan dengan metode *ensemble*. Solusi ini diyakini akan sangat bermanfaat bagi para pencari kerja. Selain itu, seseorang dapat mempertimbangkan dan memilih berbagai lowongan pekerjaan secara lebih efektif dan efisien.

2. TUJUAN DAN MANFAAT

Berdasarkan latar belakang permasalahan di atas, tujuan dari penelitian ini adalah sebagai berikut.

a. Membersihkan dan mengolah data lowongan pekerjaan.
b. Melakukan visualisasi data untuk memperoleh fitur penting pada data lowongan pekerjaan.
c. Menganalisis makna dan hubungan antar fitur pada data lowongan pekerjaan.
d. Membuat model untuk memprediksi gaji berdasarkan data lowongan pekerjaan.

e. Melatih dan mengoptimasi model yang terbaik dalam memprediksi gaji lowongan pekerjaan.

Adapun manfaat dari penelitian ini adalah sebagai berikut.

a. Memberi rekomendasi lowongan pekerjaan yang tepat bagi para pencari kerja.
b. Membantu para pencari kerja dalam mengambil keputusan untuk melamar kerja melalui hasil analisis data lowongan pekerjaan.

3. METODE ANALISIS DATA

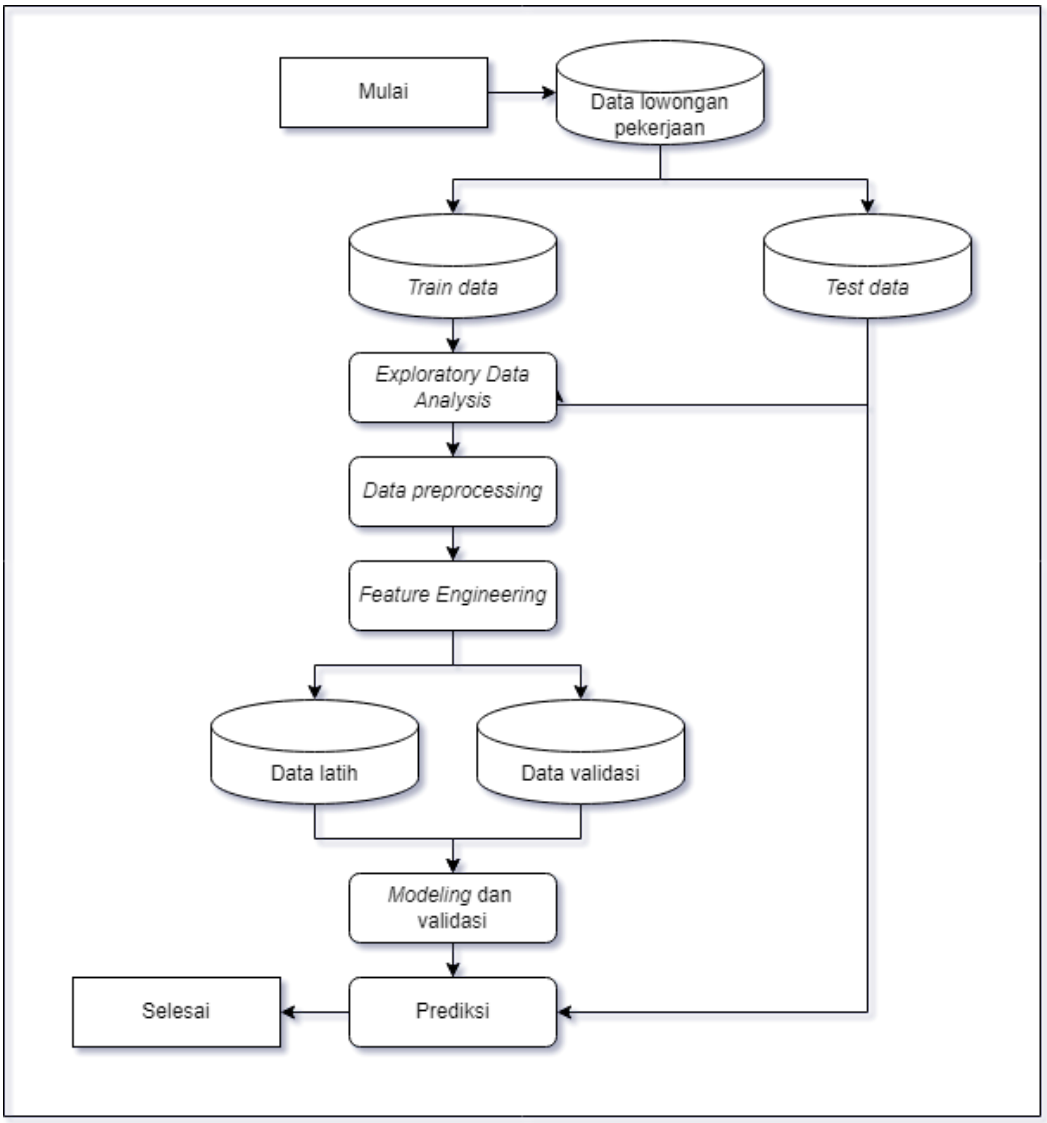
Analisis data dilakukan dengan menggunakan data lowongan pekerjaan yang disediakan di laman Kaggle oleh panitia FIND IT! 2022 (https://kaggle.com/competitions/findit2022/data). Data tersebut menjelaskan informasi-informasi lowongan pekerjaan di Indonesia yang terdiri atas 13 fitur independen, sebuah fitur *primary key* 'id', dan sebuah fitur target 'salary'.

Tabel 3.1 Penjelasan Singkat Data

Fitur	**Tipe Data**	**Deskripsi**
id	*Numerical*	*Primary key*
job_title	*Categorical*	Nama pekerjaan
location	*Categorical*	Lokasi perusahaan tempat bekerja
salary_currency	*Categorical*	Mata uang dari salary
career_level	*Categorical*	Tingkat jabatan dari pekerjaan yang ditawarkan
experience_level	*Categorical*	Lama pengalaman pelamar yang diminta perusahaan
education_level	*Categorical*	Syarat tingkat pendidikan bagi pelamar pekerjaan
employment_type	*Categorical*	Tipe kontrak
job_function	*Categorical*	Kategorisasi jenis pekerjaan
job_benefits	*Categorical*	Manfaat yang bisa diberikan perusahaan
company_process_time	*Categorical*	Rata-rata lama waktu perusahaan untuk merespon pelamar
company_size	*Categorical*	Jumlah karyawan yang bekerja di perusahaan tersebut
company_industry	*Categorical*	Jenis industri yang menjadi lini bisnis perusahaan

job_description	*Categorical*	Deskripsi pekerjaan
salary	*Numerical*	Jumlah gaji yang ditawarkan setiap bulan

Metode analisis data yang dilakukan terdiri atas empat tahapan utama, yaitu *exploratory data analysis* (EDA), *data preprocessing*, *modeling* dan validasi, serta prediksi data uji.



Gambar 3.1 Alur Kerja Metode Analisis Data

Seluruh alur dilakukan melalui media Kaggle Notebooks dan menggunakan bahasa pemrograman Python beserta *modules* dan *libraries* yang tersedia di dalamnya.

Tahap pertama, *exploratory data analysis* (EDA), bertujuan untuk memahami isi dan karakteristik data serta hubungan antar fitur. EDA dilakukan dengan menggunakan beberapa metode, seperti eksplorasi *missing values*, eksplorasi data duplikat, memvisualisasikan data, mengecek distribusi data, dan melakukan analisis secara univariat, bivariat, maupun multivariat. *Libraries* utama yang digunakan dalam tahap ini adalah Pandas, NumPy, SciPy, Matplotlib, dan Seaborn.

Tahap kedua, *data preprocessing*, bertujuan untuk memproses data latih dan uji sehingga diperoleh data yang lebih berkualitas. *Data preprocessing* dilakukan dengan menggunakan beberapa metode, seperti menangani *missing values*, menangani data duplikat, dan menghapus pencilan dan *noise*. *Libraries* utama yang digunakan dalam tahap ini adalah Pandas dan NumPy.

Tahap ketiga, *feature engineering*, bertujuan untuk merekayasa fitur sehingga dihasilkan informasi yang lebih berdampak terhadap proses prediksi oleh model. *Feature engineering* dilakukan dengan menggunakan beberapa metode, seperti menambah atau menggabungkan fitur, membuang fitur, maupun *encoding*. *Libraries* utama yang digunakan dalam tahap ini adalah Pandas dan scikit-learn. Selain itu, dilakukan pembagian train data menjadi data latih dan validasi dengan proporsi 90 : 10.

Tahap keempat, *modeling* dan validasi, bertujuan untuk membangun model yang mampu memprediksi gaji lowongan pekerjaan berdasarkan data yang tersedia. Dilakukan eksperimen terhadap berbagai model untuk memperoleh model-model dengan hasil terbaik. Setelahnya, dilakukan *hyperparameter tuning* terhadap model-model terbaik dengan menggunakan *library* Hyperopt yang menerapkan algoritma optimasi Bayesian dalam mencari kombinasi parameter terbaik. Sesudah itu, model-model tersebut digabung menggunakan metode *voting* sehingga diperoleh hasil yang lebih akurat.

Model-model yang diuji dalam tahap eksperimen adalah model berbasis *tree* dan *boosting*. Digunakan dua jenis model berbasis *tree*, yaitu Decision Tree dan Random Forest. Digunakan pula dua jenis model berbasis *boosting*, yaitu XGBoost (eXtreme Gradient Boosting) dan CatBoost (Categorical Boosting).

Model berbasis *tree* merupakan model yang menggunakan struktur pohon dalam memprediksi data. Decision Tree yang bermakna pohon keputusan menggunakan konsep mengubah data menjadi keputusan. Dengan memecah keputusan yang kompleks menjadi lebih sederhana, proses pengambilan keputusan menjadi lebih mudah.

Di sisi lain, Random Forest secara konsep merupakan kumpulan dari Decision Tree. Random Forest bekerja dengan prinsip memproses kembali hasil keputusan atau prediksi yang dibuat oleh Decision Tree. Dengan demikian, Random Forest termasuk dalam kategori algoritma *ensemble*.

Model berbasis *boosting* XGBoost termasuk dalam kategori algoritma *ensemble* yang menerapkan konsep *tree* sekaligus *gradient boosting*. Algoritma XGBoost didasarkan pada *greedy algorithm* yang memprediksi dengan cara memodifikasi *split* atau percabangan pada pohon. Percabangan ini terus-menerus dilakukan hingga nilai parameter 'max_depth' tercapai.

Model CatBoost memiliki kemiripan dengan XGBoost. Namun, terdapat sedikit perbedaan dalam pembentukan pohon keputusan. CatBoost menggunakan pohon dengan struktur *memoryless* yang berarti pohon tidak akan menyimpan memori dari pohon lainnya. Selain itu, sebelum dilakukan pelatihan, CatBoost secara otomatis membagi data-data dari setiap fitur ke dalam sebuah *bucket* atau penyimpanan yang diatur oleh *threshold value* yang bertujuan untuk membagi pasangan fitur.

Sesudah itu, model dievaluasi menggunakan data validasi sehingga diketahui performanya. Digunakan metrik *root-mean-squared error* (RMSE) yang memiliki formula

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}.$$

keterangan: n = frekuensi data, i = urutan data, $\hat{y}$ = nilai prediksi, y = nilai aktual.

Formula di atas menghitung selisih antara nilai prediksi dan nilai aktual, kemudian mengkuadratkan dan mencari rata-ratanya. Hasil perhitungan tersebut lalu diakar kuadrat sehingga diperoleh nilai RMSE.

Validasi dilakukan dengan menggunakan metode K-Fold Cross-Validation dengan $K = 10$. Metode K-Fold Cross-Validation bekerja dengan prinsip membagi data secara acak menjadi K partisi dan melakukan prediksi terhadap masing-masing partisi.

Tahap kelima, prediksi data uji, bertujuan untuk memprediksi data uji lowongan pekerjaan menggunakan model terbaik yang telah dilatih. Setelahnya, dilakukan *submission* hasil prediksi tersebut ke Kaggle.

4. ANALISIS

Data lowongan pekerjaan yang disediakan terdiri atas dua *file*, yaitu train.csv dan predict_case.csv. Data latih (train.csv) memiliki 31,746 baris dan 15 fitur, sedangkan data uji (predict_case.csv) memiliki 3,000 baris dan 14 fitur.

Tabel 4.1 Sampel Data Latih

	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...

Tabel 4.2 Sampel Data Uji

	job_title	job_function	education_level
0	Sous Chef	Hotel/Restoran,Makanan/Minuman/Pelayanan Restoran	Sertifikat Professional, D3 (Diploma), D4 (Dip...
1	Bancassurance Officer (Area: Bali, Sulawesi Ut...	Penjualan / Pemasaran,Penjualan - Jasa Keuangan	Sertifikat Professional, D3 (Diploma), D4 (Dip...
2	Marketing Staff	Penjualan / Pemasaran,Pemasaran/Pengembangan B...	SMA, SMU/SMK/STM, Sertifikat Professional, D3 ...
3	Section Head Commercials	Penjualan / Pemasaran,Penjualan Ritel	SMA, SMU/SMK/STM, Sertifikat Professional, D3 ...
4	Social Media HEAD	Penjualan / Pemasaran,Digital Marketing	SMA, SMU/SMK/STM, Sertifikat Professional, D3 ...

Melalui observasi, terdapat huruf kapital dan kecil pada masing-masing data. Oleh karena itu, dilakukan pernyeragaman menjadi huruf kecil sehingga data menjadi lebih konsisten dan seragam. Setelahnya, diperiksa tipe, frekuensi kelas unik, dan persentase *missing values* pada masing-masing data.

Tabel 4.3 Hasil Pemeriksaan Data Latih

	dtypes	nunique	nan %
id	int64	31746	0
job_title	object	19260	0
location	object	199	0
salary_currency	object	2	0
career_level	object	6	0
experience_level	object	20	14
education_level	object	21	0
employment_type	object	9	4
job_function	object	68	0
job_benefits	object	2978	21
company_process_time	object	30	29
company_size	object	7	16
company_industry	object	58	5
job_description	object	26981	0
salary	float64	477	80

Tabel 4.4 Hasil Pemeriksaan Data Uji

	dtypes	nunique	nan %
id	int64	3000	0
job_title	object	2224	0
location	object	130	0
salary_currency	object	1	0
career_level	object	5	0
experience_level	object	13	8
education_level	object	15	0
employment_type	object	8	0
job_function	object	67	0
job_benefits	object	826	25
company_process_time	object	30	35
company_size	object	7	16
company_industry	object	56	3
job_description	object	2688	0

Hasil pemeriksaan menunjukkan bahwa data latih dan uji bersifat kotor. Terlebih lagi, fitur target ‘salary’ yang menjadi pembeda antara data latih dan uji memiliki *missing values* dengan persentase 80%. Hal ini membuat data tersebut tidak dapat dijadikan sebagai data latih bagi model.

Oleh karena itu, semua *missing values* pada fitur target ‘salary’ dihapus sehingga hanya menyisakan baris dengan fitur target ‘salary’ yang tidak kosong. Setelah penghapusan tersebut, diperoleh dimensi data latih yang baru, yaitu 6,532 baris.

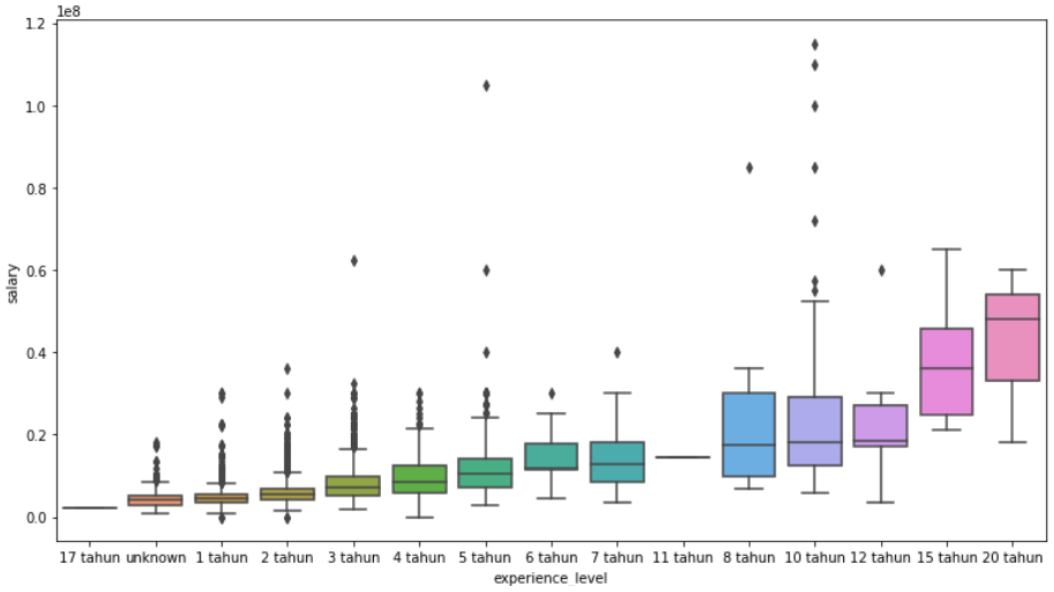
Setelahnya, dilakukan pemeriksaan keberadaan data duplikat. Hasil pemeriksaan menunjukkan terdapat data duplikat sebanyak 1,189 baris. Keberadaan data duplikat berdampak buruk sebab akan mengganggu performa model.

Oleh karena itu, semua data duplikat dihapus sehingga hanya menyisakan baris yang unik. Setelah pembuangan tersebut, diperoleh dimensi data latih yang baru, yaitu 5,343 baris. Sesudah itu, dilakukan konkatenasi antara data latih dan uji untuk diperiksa persentase *missing value*.

Tabel 4.5 Hasil Pemeriksaan Missing Values pada Data Konkatenasi

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

Hasil pemeriksaan menunjukkan bahwa terdapat lima fitur dengan *missing values*. Pertama, dilakukan analisis bivariat antara fitur ‘experience_level’ dan ‘salary’ dengan menggunakan diagram *boxplot* yang telah dimodifikasi sehingga terurut berdasarkan median gaji.



Gambar 4.1 Diagram *Boxplot* antara Fitur ‘experience_level’ dan ‘salary’

Hasil analisis menunjukkan bahwa lama pengalaman pelamar yang diminta perusahaan berbanding lurus terhadap median gaji yang ditawarkan. Kelas ‘17 tahun’ dan kelas ‘11 tahun’ dianggap sebagai pencilan dan dihapus sebab frekuensi masing-masing kelas hanya satu.

Kelas ‘unknown’ pada sumbu-x merupakan representasi dari *missing values*. Berdasarkan informasi yang diperoleh, diputuskan bahwa *missing values* ditangani dengan menjadikannya sebagai kelas tersendiri, yaitu ‘kurang dari 1 tahun’.

Kedua, dilakukan analisis univariat terhadap fitur ‘job_benefits’ dengan memeriksa frekuensi masing-masing kelas.

Tabel 4.6 Frekuensi Kelas pada Fitur ‘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

Hasil analisis menunjukkan bahwa kelas ‘unknown’ atau *missing values* memiliki frekuensi terbanyak pada fitur ‘job_benefits’. Selain itu, terdapat 1,558 kelas unik pada fitur ini. Dua informasi tersebut membuat proses *imputing missing values* menjadi sulit. Ditambah lagi, fitur dengan 1,558 variasi kelas dirasa terlalu beragam dan berpotensi mengganggu performa model. Oleh karena itu, diputuskan untuk menghapus fitur ‘job_benefits’.

Ketiga, dilakukan analisis univariat terhadap fitur ‘company_process_time’ dengan memeriksa frekuensi masing-masing kelas.

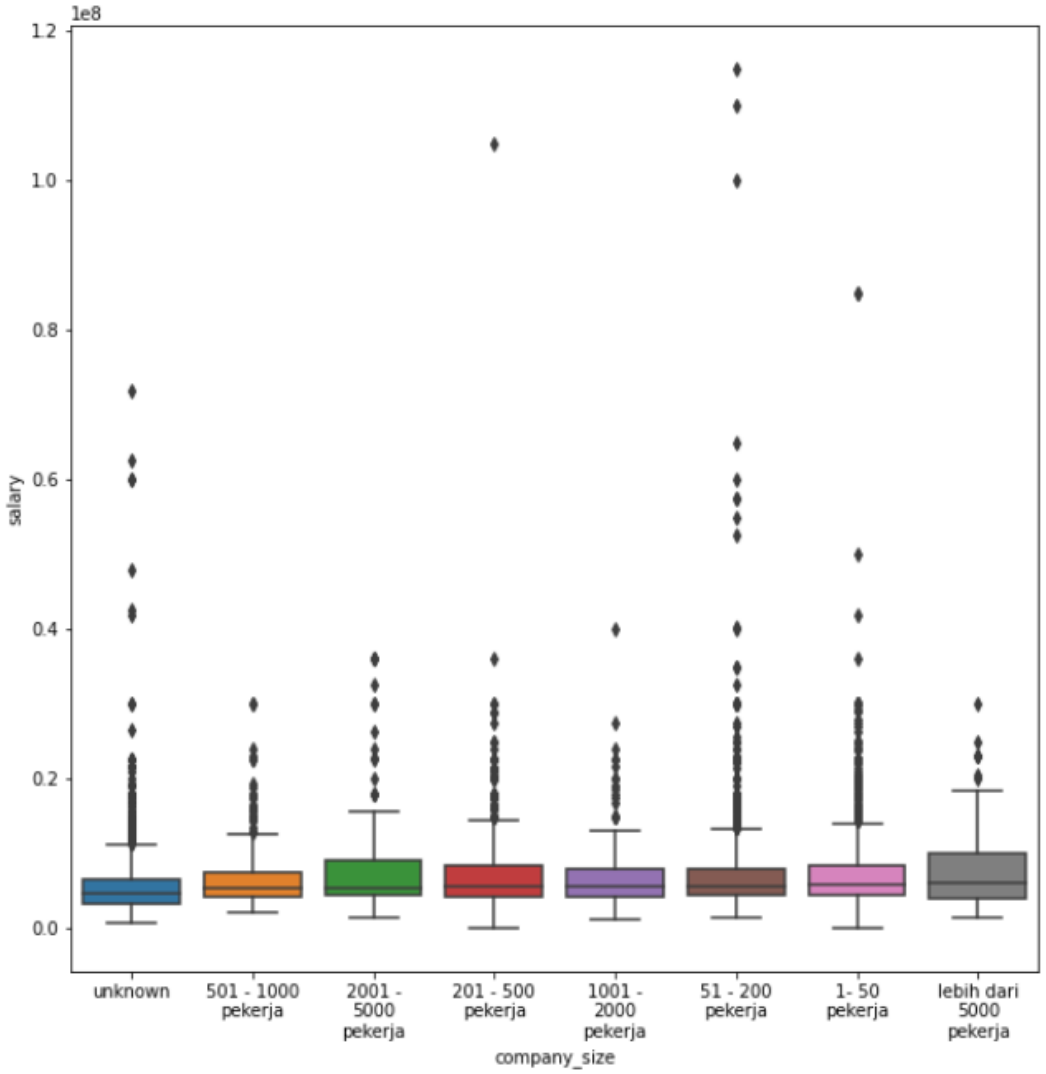
Tabel 4.7 Frekuensi Kelas pada Fitur ‘company_process_time’

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

26	13 days	77
27	8 days	66
28	6 days	53
29	11 days	41
30	10 days	33

Hasil analisis menunjukkan bahwa kelas ‘unknown’ atau *missing values* memiliki frekuensi terbanyak pada fitur ‘company_process_time’. Melalui observasi lebih lanjut, ditemukan bahwa semua kelas pada fitur ini berada dalam rentang kurang dari atau sama dengan 30 hari. Dengan kata lain, tidak terdapat rata-rata lama waktu perusahaan untuk merespon pelamar lebih dari 30 hari. Oleh karena itu, diputuskan bahwa *missing values* ditangani dengan menjadikannya sebagai kelas tersendiri, yaitu ‘lebih dari 30 hari’.

Keempat, dilakukan analisis bivariat antara fitur ‘company_size’ dan ‘salary’ dengan menggunakan diagram *boxplot* yang telah dimodifikasi sehingga terurut berdasarkan median gaji.



Gambar 4.2 Diagram *Boxplot* antara Fitur ‘company_size’ dan ‘salary’

Hasil analisis menunjukkan bahwa tidak terdapat hubungan kuat antara ukuran perusahaan dan median gaji. Ditambah lagi, tidak ada perbedaan signifikan antara median gaji satu kelas dengan kelas lainnya. Selain itu, diperoleh informasi bahwa semua kelas mewakili semua rentang ukuran perusahaan. Oleh karena itu, diputuskan bahwa *missing values* diganti dengan modus pada fitur ini, yaitu kelas ‘1 - 50 pekerja’.

Kelima, dilakukan analisis univariat terhadap fitur ‘company_industry’ dengan memeriksa frekuensi masing-masing kelas.

Tabel 4.8 Frekuensi Kelas pada Fitur ‘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

Hasil analisis menunjukkan bahwa kelas ‘unknown’ atau *missing values* memiliki frekuensi terbanyak ke-10 pada fitur ‘company_industry’. Selain itu, terdapat 59 kelas unik pada fitur ini. Melalui observasi lebih lanjut, ditemukan bahwa terdapat kelas ‘lainnya’ dengan frekuensi terbanyak ke-8 pada fitur ini. Oleh karena itu, diputuskan bahwa *missing values* diganti dengan kelas ‘lainnya’.

Setelah dilakukan analisis terhadap fitur yang memiliki *missing values*, dilakukan analisis terhadap fitur yang tidak memiliki *missing values*. Pertama, dilakukan analisis univariat terhadap fitur ‘job_title’ dengan memeriksa frekuensi masing-masing kelas.

Tabel 4.9 Frekuensi Kelas pada Fitur ‘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

Hasil analisis menunjukkan bahwa kelas ‘sales executive’ memiliki frekuensi terbanyak pada fitur ‘job_title’. Selain itu, terdapat 5,567 kelas unik pada fitur ini yang dirasa terlalu beragam dan berpotensi mengganggu performa model. Oleh karena itu, diputuskan untuk menghapus fitur ‘job_title’.

Kedua, dilakukan analisis univariat terhadap fitur ‘location’ dengan memeriksa frekuensi tiap kelas.

Tabel 4.10 Frekuensi Kelas pada Fitur ‘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

Hasil analisis menunjukkan bahwa kelas ‘jakarta raya’ memiliki frekuensi terbanyak pada fitur ‘location’. Selain itu, terdapat 173 kelas unik pada fitur ini. Melalui observasi lebih lanjut, ditemukan bahwa terdapat kelas yang merupakan nama provinsi dan nama kabupaten/kota. Indonesia mengenal konsep upah minimum provinsi (UMP), yaitu upah minimum pekerjaan yang berlaku dalam suatu provinsi. Lewat informasi tersebut, diputuskan untuk dilakukan generalisasi kelas dengan mengelompokkan kabupaten/kota berdasarkan provinsi.

Ketiga, dilakukan analisis univariat terhadap fitur ‘salary_currency’ dengan memeriksa frekuensi tiap kelas.

Tabel 4.11 Frekuensi Kelas pada Fitur ‘salary_currency’

	salary_currency	frekuensi
0	idr	8533
1	usd	2

Hasil analisis menunjukkan bahwa kelas ‘idr’ memiliki frekuensi terbanyak pada fitur ‘salary_currency’. Selain itu, terdapat satu kelas lainnya pada fitur ini, yaitu ‘usd’ yang hanya berjumlah dua. Melalui observasi lebih lanjut, ditemukan bahwa kelas ‘usd’ hanya terdapat pada data latih. Oleh karena itu, baris dengan kelas ‘usd’ dihapus. Setelahnya, fitur ‘salary_currency’ dihapus juga sebab hanya memiliki satu jenis kelas yang tidak akan membantu proses prediksi.

Keempat, dilakukan analisis univariat terhadap fitur ‘career_level’ dengan memeriksa frekuensi tiap kelas.

Tabel 4.12 Frekuensi Kelas pada Fitur ‘career_level’

	career_level	frekuensi
0	pegawai (non-manajemen & non-supervisor)	5233
1	supervisor/koordinator	1438
2	manajer/asisten manajer	1125
3	lulusan baru/pengalaman kerja kurang dari 1 tahun	612
4	ceo/gm/direktur/manajer senior	127

Hasil analisis menunjukkan bahwa pada fitur ‘career_level’, kelas ‘pegawai (non-manajemen & non-supervisor)’ memiliki frekuensi terbanyak. Selain itu, terdapat 5 kelas unik pada fitur ini.

Kelima, dilakukan analisis univariat terhadap fitur ‘education_level’ dengan memeriksa frekuensi tiap kelas.

Tabel 4.13 Frekuensi Kelas pada Fitur ‘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
10	sarjana (s1), diploma pascasarjana, gelar professional, magister (s2), doktor (s3)	41
11	diploma pascasarjana, gelar professional, magister (s2)	17
12	sma, smu/smk/stm, sarjana (s1), diploma pascasarjana, gelar professional, magister (s2)	12
13	sertifikat professional, d3 (diploma), d4 (diploma), diploma pascasarjana, gelar professional, magister (s2)	9
14	diploma pascasarjana, gelar professional, magister (s2), doktor (s3)	6
15	sma, smu/smk/stm, sertifikat professional, d3 (diploma), d4 (diploma), diploma pascasarjana, gelar professional, magister (s2)	3
16	doktor (s3)	2
17	sarjana (s1), doktor (s3)	1
18	sertifikat professional, d3 (diploma), d4 (diploma), sarjana (s1), doktor (s3)	1

Hasil analisis menunjukkan bahwa kelas ‘sarjana (s1)’ memiliki frekuensi terbanyak pada fitur ‘education_level’. Melalui observasi lebih lanjut, ditemukan bahwa jenjang pendidikan pada masing-masing kelas terurut dari yang terendah. Oleh karena itu, dilakukan generalisasi dengan mengelompokkan kelas berdasarkan jenjang pendidikan terendah. Hasil generalisasi dapat dilihat pada tabel 4.14.

Tabel 4.14 Hasil Generalisasi Kelas pada Fitur ‘education_level’

	education_level	frekuensi
0	sarjana (s1)	2944
1	sertifikat professional	2313
2	sma	1726
3	tidak terspesifikasi	1527
4	diploma pascasarjana	23
5	doktor (s3)	2

Selain itu, ditemukan bahwa kelas ‘doktor (s3)’ hanya terdapat pada data latih, sehingga baris data yang memiliki kelas ‘doktor (s3)’ pada fitur ini dapat dihapus agar tidak mengganggu performa.

Keenam, dilakukan analisis univariat terhadap fitur ‘employment_type’ dengan memeriksa frekuensi tiap kelas.

Tabel 4.15 Frekuensi Kelas pada Fitur ‘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

Hasil analisis menunjukkan bahwa kelas ‘penuh waktu’ memiliki frekuensi terbanyak pada fitur ‘employment_type’. Selain itu, terdapat sembilan kelas unik pada fitur ini. Terlihat bahwa ada beberapa kelas yang merupakan gabungan dari kelas lain sehingga dapat dikelompokkan menjadi kelas yang lebih besar sebelum tanda koma.

Ketujuh, dilakukan analisis univariat terhadap fitur ‘job_function’ dengan memeriksa frekuensi tiap kelas.

Tabel 4.16 Frekuensi Kelas pada Fitur ‘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
...	...	...
62	seni/media/komunikasi,hiburan	6
63	manufaktur,kontrol proses	5
64	pelayanan,layanan sosial/konseling	4
65	lainnya,penerbitan	3
66	sains,bioteknologi	2

Hasil analisis menunjukkan bahwa kelas ‘penjualan / pemasaran,penjualan ritel’ memiliki frekuensi terbanyak pada fitur ‘job_function’. Selain itu, terdapat 67 kelas unik pada fitur ini. Melalui observasi lebih lanjut, ditemukan bahwa kata sebelum tanda koma merupakan bidang utama, sedangkan setelah tanda koma menunjukkan subbidang. Oleh karena itu, fitur ini dapat dikelompokkan menjadi bidang utama.

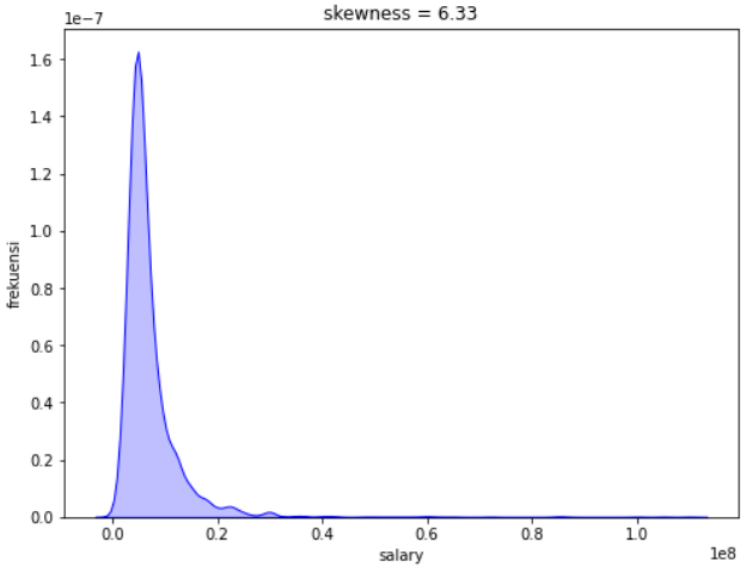
Kedelapan, dilakukan analisis univariat terhadap fitur ‘job_description’ dengan memeriksa frekuensi tiap kelas.

Tabel 4.17 Frekuensi Kelas pada Fitur ‘job_description’

	job_description	counts
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

Hasil analisis menunjukkan bahwa terdapat 7776 kelas unik pada fitur ‘job_description’. Oleh karena itu, dapat dikatakan fitur ini sangat bervariasi sehingga akan dihapus.

Setelah penanganan terhadap masing-masing fitur, dilakukan pemeriksaan data duplikat terhadap data latih yang telah diproses. Hasil pemeriksaan menunjukkan bahwa terdapat 684 data dengan fitur-fitur independen yang terduplikasi, tetapi memiliki fitur target yang berbeda. Dengan kata lain, data-data tersebut adalah data duplikat dengan ‘salary’ yang berbeda. Oleh karena itu, dilakukan penanganan dengan melakukan *group by* terhadap data duplikat dan mengambil rata-rata ‘salary’ dari data duplikat tersebut. Sesudah itu, dilakukan analisis univariat terhadap fitur target ‘salary’.



Gambar 4.3 Diagram Distribusi Fitur ‘salary’

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

Gambar 4.4 Deskripsi Nilai Kuartil Fitur ‘salary’

Hasil analisis menunjukkan bahwa distribusi fitur ‘salary’ memiliki kecondongan atau *skewness* yang sangat positif. Hal ini sejalan dengan hasil analisis kuartil yang menunjukkan bahwa terdapat interval yang besar antara nilai Q3 dan maksimum. Dengan kata lain, terdapat pencilan dengan nilai yang jauh lebih besar dari nilai pusat persebaran.

Umumnya, pencilan bersifat buruk sebab kebanyakan model pembelajaran mesin tidak dapat bekerja dengan baik dengan keberadaan pencilan. Salah satu praktik penanganan yang umum adalah menghapus pencilan tersebut dengan metode interquartile range (IQR) atau z-score. Namun, jika memperhatikan konteks data, keberadaan pencilan menjadi masuk akal sebab pada faktanya terdapat beberapa pekerjaan yang memang memberikan penghasilan fantastis. Dengan kata lain, keberadaan pencilan pada fitur ‘salary’ terjadi secara alamiah dan tidak perlu dihapus.

Di sisi lain, ditemukan *noise* atau kesalahan data pada fitur ‘salary’, yaitu gaji kurang dari 500,000. Keberadaan *noise* ditangani dengan cara dihapus agar tidak mengganggu performa model.

Tabel 4.18 Jumlah Noise pada Fitur ‘salary

id	salary
1077	10
4034	10
4265	5300

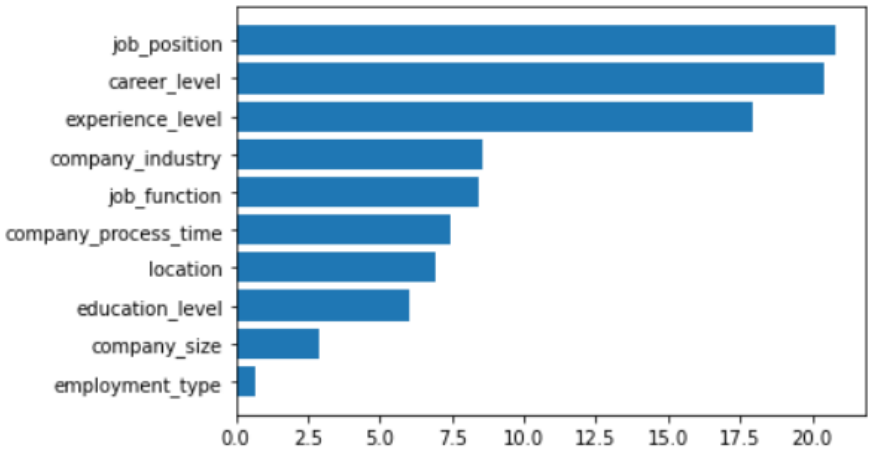
Setelah tahap EDA dan *data preprocessing* selesai, dilakukan *feature engineering* untuk memodifikasi fitur-fitur yang ada sehingga dapat meningkatkan performa model. Diputuskan untuk dibuat fitur baru bernama ‘job_position’ yang merupakan hasil penggabungan antara fitur ‘career_level’ dan ‘job_function’. Fitur baru ini memberi informasi mengenai nama atau posisi pekerjaan.

Tabel 4.19 Sampel Fitur ‘job_position’

	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

Setelahnya, dilakukan *encoding* atau konversi fitur *categorical* menjadi *numerical*. Digunakan metode *ordinal encoding* yang mengubah *categorical data* menjadi bilangan bulat.

Setelah itu, dilakukan analisis tingkat kepentingan fitur dengan *base* model CatBoost.



Gambar 4.5 Diagram Batang Tingkat Kepentingan Fitur

Hasil analisis menunjukkan bahwa fitur baru ‘job_posititon’ menjadi fitur dengan pengaruh tertinggi. Selain itu, dilakukan eksperimen untuk mengetahui kombinasi fitur-fitur yang memberi hasil prediksi terbaik.

Hasil eksperimen menunjukkan bahwa penghapusan fitur ‘employment_type’ yang merupakan fitur dengan pengaruh terendah memberi hasil prediksi terbaik. Oleh karena itu, diputuskan untuk menghapus fitur ‘employment_type’.

Setelah tahap *preprocessing* dan *feature engineering selesai,* dilakukan *modeling* dan validasi untuk membangun model dan mengevaluasi performanya menggunakan metode K-Fold Cross-Validation dengan K=10.

Tabel 4.20 Hasil Evaluasi Awal Model

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

Hasil evaluasi menunjukkan bahwa model *boosting* berbasis *tree*, CatBoost dan XGBoost, memiliki hasil yang paling baik, disusul oleh model *ensemble tree* Random Forest, dan model *singular tree* Decision Tree. Secara spesifik, model CatBoost merupakan model terbaik sebab memiliki RMSE paling

rendah di antara model lainnya. Dengan demikian, diputuskan untuk menggunakan CatBoost dan XGBoost sebagai model-model final yang akan dioptimasi dengan metode *hyperparameter tuning* dan *voting*.

Tabel 4.21 Hasil *Hyperparameter Tuning*

Parameter CatBoost	**Parameter XGBoost**
depth = 6 iterations = 1600 l2_leaf_reg = 13.48 learning_rate = 0.015 random_strength = 9.25	colsample_bylevel = 0.4 colsample_bynode = 0.6 colsample_bytree = 0.4 gamma = 2.99 learning_rate = 0.073 max_bin = 97 max_depth = 21 max_leaves = 2673 min_child_weight = 3.84 reg_alpha = 2.72 reg_lambda = 1.91 seed = 44 subsample = 0.6

Hasil evaluasi model CatBoost dan XGBoost sebelum dan sesudah dilakukan *hyperparameter tuning* ditunjukkan pada tabel 4.22. Hasil evaluasi menunjukkan bahwa penyesuaian parameter memberikan skor yang lebih baik secara cukup signifikan.

Tabel 4.22 Hasil Evaluasi Sebelum dan Sesudah *Hyperparameter Tuning*

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

Setelah itu, kedua model digabung menggunakan metode *voting* yang memanfaatkan algoritma Voting Regressor. Hasil evaluasi menunjukkan bahwa

model *ensemble* berbasis *voting* memberikan skor sedikit lebih baik daripada model individu terbaik, XGBoost.

Tabel 4.23 Hasil Evaluasi Model *Ensemble* Berbasis *Voting*

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

Model *ensemble* berbasis *voting* antara XGBoost dan CatBoost yang telah dilatih, dioptimasi, dan dievaluasi, digunakan untuk memprediksi data uji. Hasil prediksi menunjukkan bahwa model memiliki nilai RMSE sebesar 3,780,258 pada *cross-validation* dan 36,599,964 pada *leaderboard* Kaggle.

5. KESIMPULAN

Berdasarkan hasil analisis data yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut.

a. Data lowongan pekerjaan yang telah dieksplorasi dan dianalisis dengan metode-metode statistik menghasilkan informasi penting mengenai karakteristik dan hubungan antar fitur.
b. Data lowongan pekerjaan yang telah melewati tahap *preprocessing* dengan menangani *missing values*, menangani data duplikat, dan menghapus *noise*, memiliki kualitas yang jauh lebih baik.
c. Dilakukan *feature engineering* terhadap data sehingga diperoleh informasi-informasi tambahan yang membantu meningkatkan akurasi prediksi.
d. Eksperimen *screening* model menghasilkan model-model terbaik, yaitu CatBoost dan XGBoost dengan RMSE berturut-turut sebesar 4,014,913 dan 4,061,446.
e. Model *ensemble* berbasis *voting* antara CatBoost dan XGBoost yang telah dioptimasi melakukan prediksi dengan tingkat akurasi yang lebih baik, yaitu dengan nilai RMSE sebesar 3,780,258.

6. DAFTAR PUSTAKA

Goyal, C., 2021. Why You Shouldn't Just Delete Outliers - Analytics Vidhya. [Online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/05/why-you-shouldnt-just-delete-outliers/> [Accessed 16 May 2022].

Guo, R., Zhao, Z., Wang, T., Liu, G., Zhao, J. and Gao, D. (2020). Degradation State Recognition of Piston Pump Based on ICEEMDAN and XGBoost. Applied Sciences, 10(18), p.6593. doi:10.3390/app10186593.

Pham, S.T., Vo, P.S. and Nguyen, D.N. (2021). Effective Electrical Submersible Pump Management using Machine Learning. Open Journal of Civil Engineering, 11(01), pp.70–80. doi:10.4236/ojce.2021.111005.

Ren, Q., Li, M. and Han, S. (2019). Tectonic Discrimination of Olivine in Basalt using Data Mining Techniques based on Major Elements: A Comparative Study from Multiple Perspectives. Big Earth Data, 3(1), pp.8–25. doi:10.1080/20964471.2019.1572452.

Suryana, S.E., Warsito, B. and Suparti, S. (2021). Penerapan Gradient Boosting dengan Hyperopt untuk Memprediksi Keberhasilan Telemarketing Bank. Jurnal Gaussian, 10(4), pp.617–623. doi:10.14710/j.gauss.v10i4.31335.