LAPORAN
DATA ANALYTICS COMPETITION
FIND IT! 2022

Tim Akhirnya Belajar

Disusun Oleh :

Melvin Putra
Evan Eka Wijaya
Maxwell Thomson

DKI Jakarta

2022

Daftar Isi

BAB I

Pendahuluan

1.1 Latar Belakang

Di Indonesia, terdapat tujuh juta orang yang sedang mencari kerja setiap tahunnya. Portal lowongan kerja menjadi sarana penting bagi mereka untuk mencari pekerjaan yang diinginkan. Namun, banyak sekali portal lowongan kerja seperti *JobStreet* dan *LinkedIn* yang tidak menyertakan informasi lengkap mengenai suatu lowongan kerja. Salah satu informasi yang sering kali tidak ditampilkan adalah informasi mengenai gaji lowongan kerja.

Menurut survei "What Workers Want" yang diselanggarakan Hays, 61% pencari kerja ingin mengetahui gaji suatu lowongan kerja. Namun, hanya 46% dari lowongan kerja yang menyediakan informasi tersebut.

Tanpa adanya informasi mengenai gaji, calon pekerja tidak dapat mengetahui kisaran gaji dari posisi yang dilamar. Hal ini menjadi tantangan baru bagi calon pekerja dalam proses menegosiasi gaji dengan HR, di mana tahap negosiasi gaji ini merupakan tahap kunci yang menentukan keberhasilan seseorang dalam memperoleh pekerjaan. Jika gaji yang diajukan calon pekerja melebihi standar yang telah ditetapkan perusahaan, maka hal ini dapat memperkecil peluang terjadinya kesepakatan antara calon pekerja dan HR, sebaliknya, angka gaji yang terlalu rendah justru akan merugikan calon pekerja tersebut di kemudian hari. Oleh sebab itu, tersedianya informasi gaji pada laman lowongan kerja pun menjadi hal yang esensial bagi para calon pekerja.

Salah satu solusi dari masalah ini adalah memprediksi gaji sebuah lowongan kerja berdasarkan informasi yang ada pada laman tersebut dengan metode *machine learning*. Penulis berharap bahwa dengan dilakukan analisis terhadap informasi yang ada tentang suatu lowongan kerja, dapat ditemukan berbagai macam informasi/fitur yang dapat digunakan untuk memprediksi gaji sebuah lowongan kerja.

1.2 Tujuan

Tujuan dari penilitian ini adalah sebagai berikut:

1. Membangun sebuah model yang dapat memprediksi gaji suatu lowongan kerja berdasarkan informasi yang tersedia.
2. Mendapatkan berbagai macam *insight* baru mengenai informasi yang ada pada suatu lowongan kerja.

1.3 Manfaat

Manfaat dari penilitian ini adalah sebagai berikut:

1. Membantu warga Indonesia mengetahui gaji dari suatu lowongan kerja yang tidak menyediakan informasi mengenai gaji.
2. Membantu warga Indonesia memperoleh gaji yang sesuai dengan pekerjaan yang diinginkan.

BAB II

Metode Analisis Data

2.1 Perangkat Lunak

Perangkat lunak yang diguanakn dalam proses pengolahan dan analisis data adalah layanan *cloud service* Google Colab dengan bahasa pemrogaman Python.

Python adalah bahasa pemograman tingkat tinggi yang dibuat oleh Guido Van Rossum pada tahun 1991. Python banyak digunakan dalam berbagai bidang, seperti *software development, GUI development,* dan pembelajaran mesin.

2.2 Deskripsi Data

Dataset yang diberikan oleh panitia DAC FIND-IT! 2022 adalah data mengenai informasi lowongan kerja. Terdapat tiga dataset yang diberikan, yaitu ‘train.csv’ dengan jumlah baris sebanyak 31746,‘predict-case.csv’ dan ‘sample_submission.csv’ dengan jumlah baris sebanyak 3000. Dataset ‘train.csv’ digunakan untuk membangun model, dataset ‘predict-case.csv’ adalah dataset yang akan diprediksi model yang telah dilatih dimana hasilnya nanti akan dievaluasi di website *Kaggle* dan ‘sample_submission.csv’ adalah contoh bentuk submisi yang benar. Dataset ‘train.csv’ memiliki 15 kolom dengan keterangan sebagai berikut:

No	Nama Kolom	Keterangan
1	id	ID dari lowongan kerja
2	job_title	Nama pekerjaan
3	location	Lokasi perusahaan tempat bekerja
4	salary_currency	Mata uang dari gaji
5	career_level	Tingkat jabatan dari pekerjaan yang ditawarkan
6	experience_level	Lama pengalaman pelamar yang diminta perusahaan

7	education_level	Syarat tingkat pendidikan bagi pelamar pekerjaan
8	employment_type	Tipe kontrak
9	job_function	Kategorisasi jenis pekerjaan
10	job_benefits	Manfaat yang bisa diberikan perusahaan
11	company_process_time	Rata rata lama waktu perusahaan untuk merespon pelamar
12	company_size	Jumlah karyawan yang bekerja di perusahaan tersebut
13	company_industry	Jenis industry yang menjadi lini bisnis perusahaan
14	job_description	Deskripsi pekerjaan yang ditawarkan
15	salary	Jumlah gaji yang ditawarkan

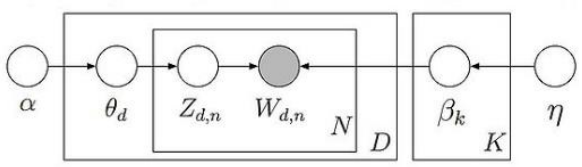
Tabel 2.1 Keterangan Variabel Dataset

Dataset ‘predict-case.csv’ kolom yang sama dengan ‘train.csv’tanpa adanya kolom ‘salary’. Dataset ‘submission_sample.csv’ hanya memiliki 2 baris, yaitu itu ‘id’ dan ‘salary’.

2.3 Algoritma dan Metode

2.3.1 LDA

Latent Dirichlet Allocation (LDA) adalah salah satu metode pemodelan topik yang memanfaatkan model probabilitas generatif dan distribusi dirichlet. Perhatikan graf algoritma LDA di bawah ini:



Gambar 2.1 Algoritma LDA

Di mana:

N : jumlah kata di dalam seluruh dokumen (ukuran kosakata)

D : jumlah dokumen

K : jumlah topik

θ_d : *topic mix* untuk dokumen d (distribusi untuk K topik pada dokumen d)

$Z_{d,n}$: penetepan/penentuan topik untuk kata n dalam dokumen d

$W_{d,n}$: kata n dalam dokumen d

β_k : distribusi kata untuk topik k

Dalam pemodelan topik menggunakan LDA, jumlah topik (K) ditentukan oleh penulis. Selain itu, LDA menggunakan dua distribusi dirichlet di dalam algoritmanya. Distribusi pertama, yaitu distribusi dirichlet untuk persebaran topik di dalam dokumen mempunyai parameter alpha (α), dan distribusi kedua, yaitu distribusi dirichlet untuk persebaran kata di dalam topik mempunyai parameter eta (η). Kedua parameter berperan sebagai parameter 'konsentrasi', di mana nilai dari kedua parameter akan memengaruhi ukuran pemusatan distribusi.

2.3.2 Random Forest

Random Forest adalah salah satu jenis ensemble-based learning yang merupakan pengembangan dari metode CART (Classification and Regression Tree). Pada gugus data yang terdiri dari n observasi dan p peubah penjelas, proses algoritma Random Forest adalah sebagai berikut (Breiman 2001; Breiman & Cutler 2003):

1. Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data (bootstrap).
2. Berdasarkan hasil *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas acak (random variable), dengan $m < p$. Pemilah terbaik dipilih dari m peubah penjelas tersebut (*random feature selection*).

3. Langkah 1 dan 2 diulangi sebanyak T kali sehingga membentuk hutan yang terdiri dari T pohon.

Penentuan hasil prediksi variabel target dapat diperoleh dari agregasi hasil prediksi k pohon yang telah dilakukan sebelumnya. Jika variabel target yang ingin diprediksi adalah kategorik (kasus klasifikasi), maka hasil prediksi akan didasarkan pada perhitungan suara terbanyak (*majority vote*)

2.3.3 XGBoost

XGBoost atau eXtreme Gradient Boosting adalah algoritma berbasis pohon yang ditemukan oleh Tianqi Chen pada tahun 2016 yang merupakan pengembangan dari algoritma Gradient Boosting Decision Tree.

Parameter yang digunakan dalam algoritma XGBoost adalah sebagai berikut:

1. Colsample_bytree, digunakan untuk memilih jumlah sampel kolom yang akan digunakan.
2. Eta, parameter learning rate yang berfungsi untuk mencegah terjadinya overfitting.
3. Gamma, parameter yang berfungsi untuk menentukan pemangkasan node pada pohon yang dibuat.
4. Max_depth, parameter yang berfungsi untuk menentukan kedalaman pohon yang dibangun.
5. Min_child_weight, parameter untuk mengatur batasan berat minimum pada sebuah node.

2.3.4 Light Gradient Boosting Machine

Light Gradient Boosting Machine atau LGBM adalah struktur hasil perkembangan yang didasari oleh Decision Tree. LGBM secara garis besar merupakan algoritma Decision Tree yang berlandaskan *gradient-based one-side sampling* (GOSS), *exclusive feature bundling* (EFB), dan strategi perkembangan histogram serta *leaf-wise* yang memiliki limit kedalaman. GOSS bisa menyimpan sampel dengan gradien besar dan melakukan pengambilan sampel secara acak pada data dengan

gradien kecil bergantung pada suatu proporsi. EFB membagi fitur menjadi kumpulan *mutually exclusive* yang berjumlah lebih kecil dengan menggunakan suatu aproksimasi. Ide dari algoritma histogram adalah untuk mendiskritkan titik fitur kontinu menjadi sebanyak *k* bilangan bulat, dan Menyusun histogram dengan lebar *k* di saat yang sama. Kemudian dalam proses pembelajarannya, *Decision Tree* pada LGBM digenerasikan dengan metode pertumbuhan secara *level-wise*.

2.3.5 CatBoost

CatBoost adalah adalah algoritma gradient boosting pada decision tree dan merupakan perkembangan dari algoritma XGBoost serta LGBM. Perkembangan pertama yang dilakukan adalah metode ini dapat mengolah data dengan variable kategorik berkardinalitas tinggi. Ketika tidak menggunakan metode *one hot encoding*, metode yang digunakan oleh CatBoost dalam melakukan encoding terhadap variabel kategorik disebut dengan "*Ordered Target Statistic*". Target statistic adalah nilai yang dapat dihitung dari nilai sebenarnya yang berkaitan dengan nilai-nilai dari variable kategorik tertentu.

2.3.6 Randomized Search

Randomized Search adalah salah satu metode *hyperparameter tuning*. Parameter dari model yang akan dilakukan *hyperparameter tuning* akan dioptimasi dengan *crossvalidated search* pada seluruh parameter yang telah didefinisikan.

Pada metode ini, tidak seluruh parameter dicoba, tetapi hanya beberapa parameter saja yang disampel dari sebuah distribusi yang ditentukan. Parameter disampel tanpa pengembalian.

2.4 Metrik Evaluasi

Dalam proses mengevaluasi model, akan digunakan metode pengukuran Root Mean Square Error (RMSE). Formula RMSE adalah sebagai berikut:

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

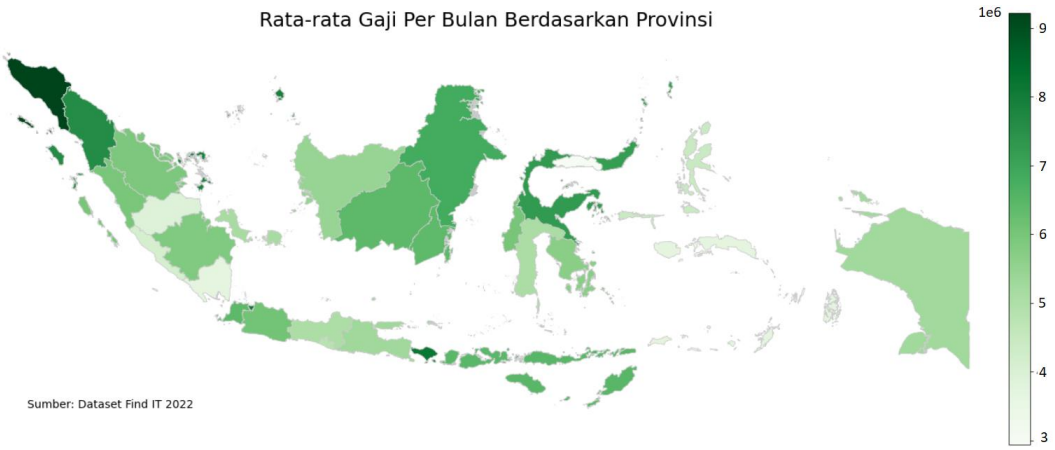
BAB III

Analisis Data

3.1 Analisis Data Eksploratori

Analisis Data Eksploratori (EDA) adalah tahap awal dalam pembentukan model untuk mengenal data yang dimiliki lebih dalam lagi. Dalam tahap ini, akan diringkaskan karakteristik utama data yang dimiliki dengan menggunakan grafik atau metode visualisasi data lainnya. EDA adalah tahap penting sebelum memasuki tahap tahap lainnya karena kita perlu memastikan terlebih dahulu data seperti apa yang kita punya dan apa yang dapat kita lakukan kepada data tersebut.

3.1.1 Lokasi



Gambar 3.1 Rata-rata Gaji Per Bulan berdasarkan Provinsi

Gambar 3.1 menunjukkan rata-rata gaji per bulan berdasarkan 34 provinsi Indonesia. Rata-rata gaji berkisar di antara angka 2.9 juta hingga 9.2 juta, di mana angka gaji yang rendah ditandai dengan warna hijau yang terang mendekati putih, sementara itu gaji yang tinggi ditandai dengan warna hijau yang semakin gelap. Lima provinsi dengan rata-rata gaji tertinggi dapat dilihat pada tabel berikut beserta dengan industri perusahaan yang mendominasi pada provinsi tersebut:

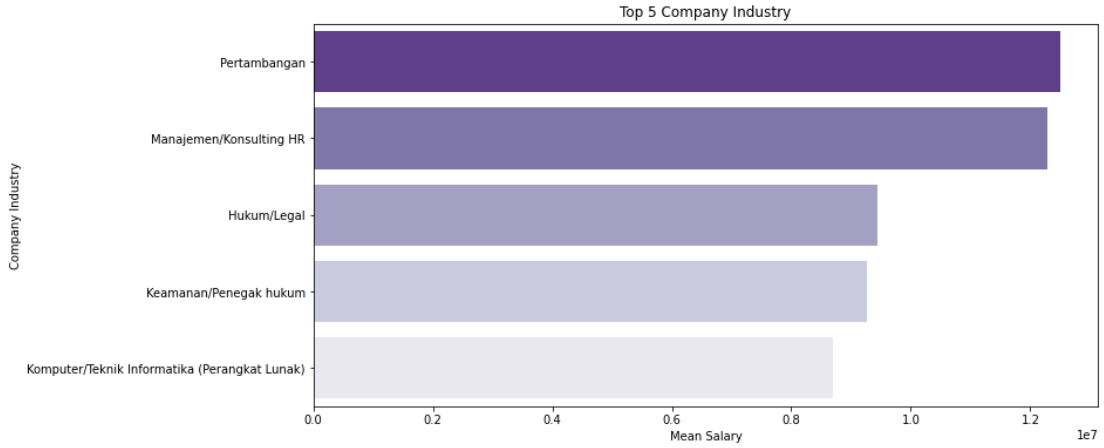
Tabel 3.1 Provinsi dengan Rata-rata Gaji Tertinggi beserta Industri Perusahaan

No	Provinsi	Rata-rata Gaji	Industri Perusahaan	Persentase
1	Aceh	Rp9.222.500	Makanan & Minuman/Katering/Restoran	81,45%
			Kesehatan/Medis	8,06%
2	DKI Jakarta	Rp8.235.869	Umum & Grosir	8,21%
			Retail/Merchandise	8,14%
3	Bali	Rp8.188.959	Komputer/Teknik Informatika (Perangkat Lunak)	19,13%
			Manajemen/Konsulting HR	10,38%
4	Kepulauan Riau	Rp7.884.166	Manufaktur/Produksi	17,24%
			Manajemen/Konsulting HR	10,34%
5	Sumatera Utara	Rp7.616.092	Makanan & Minuman/Katering/Restoran	36,93%
			Manufaktur/Produksi	10,80%

Daerah Provinsi Aceh menjadi provinsi dengan rata-rata gaji per bulan tertinggi sebesar 9.222.500 juta rupiah. Di antaranya, 81.45% jumlah lowongan pekerjaan berpusat pada bidang Makanan & Minuman/Katering/Restoran. Hal ini sejalan dengan fakta bahwa sektor usaha kuliner memang telah menjadi salah satu penyumbang pertumbuhan ekonomi tertinggi Provinsi Aceh pada tahun 2018 berdasarkan keterangan Kepala BPS Aceh Wahyudin. Selain Provinsi Aceh, Provinsi Sumatera Utara sebagai provinsi dengan rata-rata gaji tertinggi ke-5 juga cukup didominasi oleh perusahaan pada sektor Makanan & Minuman/Katering/Restoran.

Sementara itu, Provinsi Bali didominasi oleh perusahaan IT yang hampir mencapai 20% dari total perusahaan yang ada, dan Provinsi Kepulauan Riau didominasi oleh perusahaan yang bergerak di industri manufaktur & produksi sebesar 17.24%. Terakhir, industri perusahaan di Provinsi DKI Jakarta cenderung terdistribusi lebih beragam, di mana sektor perusahaan yang tertinggi yaitu industri umum & grosir hanya mencapai angka 8.21%.

3.1.2 Industri perusahaan dan Pekerjaan



Gambar 3.2 5 Industri Perusahaan dengan Rata-Rata Gaji Tertinggi

Industri Perusahaan	Rata-Rata Gaji
Pertambangan	Rp 12.510.530
Manajemen/Konsulting HR	Rp 12.281.300
Hukum/Legal	Rp 9.450.000
Keamanan/Penegak hukum	Rp 9.265.556
Komputer/Teknik Informatika (Perangkat Lunak)	Rp 8.688.701

Tabel 3.2 Data Gaji untuk 5 Industri Perusahaan dengan Rata-rata Gaji Tertinggi

Gambar 3.2 menunjukkan 5 industri perusahaan dengan rata-rata gaji tertinggi untuk industri-industri perusahaan yang memiliki lebih dari 5 observasi yang berlabel pada dataset yang diberikan. Secara rinci. rata-rata gaji dari kelima industri tersebut diberikan pada tabel 3.2.

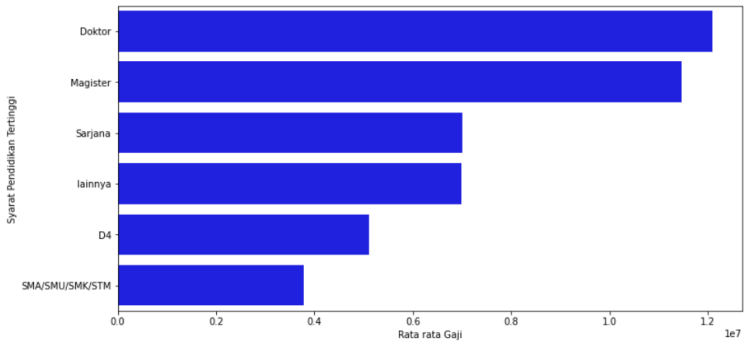
Tabel 3.3 Tiga Pekerjaan dengan Rata-Rata Gaji Tertinggi pada 5 Industri Teratas

No	Industri Perusahaan	Pekerjaan	Persentase
1	Pertambangan	Pemasaran/Pengembangan Bisnis	1.75%
		Pengacara/Asisten Legal	5.26%
		Pemeliharaan	1.75%
2	Manajemen/Konsulting HR	Keuangan/Investasi perusahaan	0.25%
		Top Management	0.25%
		Properti Real Estate	0.74%
3	Hukum/Legal	IT-Perangkat Lunak	41.67%
		Penjualan Ritel	8.33%
		Pengacara/Asisten Legal	25.00%
4	Keamanan/Penegak hukum	Angkatan Bersenjata	33.33%
		Teknik Sipil/Konstruksi Bangunan	11.11%
		IT-Perangkat Lunak	11.11%
5	Komputer/Teknik Informatika	Pelatihan & Pengembangan	0.76%
		Merchandising	0.25%
		Digital Marketing	3.80%

Pekerjaan dengan rata-rata gaji tertinggi pada kelima industri perusahaan tersebut dapat dilihat pada Tabel 3.3. Secara umum, pekerjaan-pekerjaan ini bukanlah pekerjaan yang secara umum berkaitan dengan industrinya. Contohnya, pemasaran/pengembangan bisnis dan pengacara/asisten legal bukanlah perkerjaan yang umum dikaitkan dengan industri pertambangan, namun ternyata memiliki rata-rata gaji yang paling tinggi dalam industri tersebut. Dapat diperhatikan pula, walaupun IT-perangkat lunak merupakan pekerjaan yang memiliki persentase cukup besar dalam industri Hukum/legal, pekerjaan tersebut tidak secara umum berkaitan erat dengan industrinya dibandingkan dengan pengacara/asisten legal, namun memiliki rata-rata gaji yang lebih tinggi.

3.1.3 Education level

Pada bagian ini, akan dibahas fitur education_level, yaitu syarat tingkat pendidikan pelamar untuk lowongan kerja tersebut. Untuk menganalisis lebih lanjut, akan dilihat syarat pendidikan tertinggi yang diperlukan untuk lowongan pekerjaan tersebut.



Gambar 3.3 Rata rata gaji lowongan kerja Berdasarkan Syarat Pendidikan Tertinggi

Gambar 3.3 menunjukan rata rata gaji lowongan pekerjaan dengan syarat pendidikan tertinggi Doktor, Magister, Sarjana, D4, dan SMA/SMU/SMK/STM. Lowongan pekerjaan yang tidak menyatakan syarat pendidikan yang jelas diberikan label ‘lainnya’. Dapat dilihat bahwa lowongan pekerjaan dengan syarat pendidikan Doktor memiliki rata rata gaji paling tinggi yaitu Rp12.094.444 juta, sementara lowongan pekerjaan dengan syarat pendidikan SMA/SMU/SMK/STM memiliki rata rata gaji paling rendah yaitu Rp3,788,580 juta.

Selain itu, dapat dilihat bahwa lowongan pekerjaan dengan syarat pendidikan tertinggi sarjana memiliki rata rata gaji Rp7.009.770 juta yaitu 85% lebih tinggi dari rata rata lowongan pekerjaan dengan syarat pendidikan tertinggi SMA/SMU/SMK/STM. Sementara D4 memiliki rata rata gaji Rp5.117.123 yaitu 35% lebih tinggi dari SMA/SMU/SMK/STM. Hal ini menunjukan bahwa sebenarnya lebih baik untuk melanjutkan pendidikan ke jenjang yang lebih tinggi terlebih dahulu sebelum mencari lowongan pekerjaan.

Tabel 3.4 Jabatan Berdasarkan Syarat Pendidikan Tertinggi Lowongan Kerja

Pendidikan Tertinggi	Jabatan	Rata rata gaji
SMA/SMU/SMK/STM	Lulusan Baru	Rp3.304.538
	Pegawai	Rp3.743.305
	Manajer/Asisten Manajer	Rp5.483.333
	Supervisor/Koordinator	Rp5.508.888
Sarjana	Lulusan Baru	Rp4.875.225
	Pegawai	Rp5.602.281
	Manajer/Asisten Manajer	Rp12.902.439
	Supervisor/Koordinator	Rp7.594.015

	CEO/GM/Direktur/Manajer Senior	Rp19.642.622

Tabel 3.4 mendukung dugaan ini, dapat dilihat bahwa untuk lowongan pekerjaan dengan syarat pendidikan tertinggi SMA/SMU/SMK/STM memiliki rata rata gaji lebih rendah dibandingkan syarat pendidikan tertinggi sarjana pada seluruh jabatan. Selain itu, lulusan SMA/SMU/SMK/STM hanya dapat memiliki jabatan tertinggi supervisor atau koordinator dengan rata rata gaji Rp5.508.888, sementara seorang lulusan sarjana dapat memiliki jabatan CEO/GM/Direktur/Manajer Senior dengan rata rata gaji Rp19.642.622.

Tabel 3.5 Industri Perusahaan Berdasarkan Syarat Pendidikan Tertinggi

Pendidikan Tertinggi	Jenis Industri	Rata rata gaji
SMA/SMU/SMK/STM	Akunting/Audit/Layanan Pajak	Rp6.360.000
	Pertambangan	Rp6.350.000
	Konstruksi	Rp6.057.881
	Asuransi	Rp5.750.000
	Manufaktur/Produksi	Rp5.059.442
Sarjana	Manajemen/Konsulting HR	Rp10.818.088
	Pertambangan	Rp10.782.000
	Layanan Umum	Rp9.450.000
	Keamanan/Penegak Hukum	Rp9.236.250
	Olahraga	Rp8.750.000

Tabel 3.5 menunjukan 5 industri yang memiliki rata rata gaji tertinggi untuk lowongan kerja denga syarat pendidikan tertinggi SMA/SMU/SMK/STM dan sarjana. Untuk SMA/SMU/SMK/STM, industri dengan gaji paling tinggi adalah akunting/audit/layanan pajak dengan rata rata gaji Rp6.360.000. Untuk sarjana, industri dengan gaji paling tinggi adalah manajemen/consulting HR dengan rata rata gaji Rp10.818.088.

Tabel 3.6 Jenis Pekerjaan Berdasarkan Syarat Pendidikan Tertinggi

Pendidikan Tertinggi	Jenis Pekerjaan	Rata rata gaji
SMA/SMU/SMK/STM	Praktisi/Asisten Medis	Rp10.500.000
	Pertanian	Rp7.750.000
	Teknik	Rp7.016.666
	Jurnalis/Editor	Rp6.500.000
	Sekretaris	Rp5.975.000

Sarjana	Manajemen Tingkat Atas	Rp26.166.666
	Angkatan Bersenjata	Rp13.600.000
	Teknik Industri	Rp10.493.000
	Biomedis	Rp9.793.750
	IT-Perangkat Lunak	Rp9.624.192

Tabel 3.6 menunjukan 5 jenis pekerjaan yang memiliki rata rata gaji tertinggi untuk lowongan kerja dengan syarat pendidikan tertinggi SMA/SMU/SMK/STM dan sarjana. Untuk SMA/SMU/SMK/STM, jenis pekerjaan dengan gaji paling tinggi adalah praktisi/asisten medis dengan rata rata gaji Rp10.500.000. Untuk sarjana, jenis pekerjaan dengan gaji paling tinggi adalah manajemen tingkat atas dengan rata rata gaji Rp26.166.666.

3.2 Pra-Pengolahan

Data latih (‘train.csv’) memiliki 31746 observasi yang terdiri dari 6352 observasi yang memiliki data gaji dan 25394 observasi lainnya tidak memiliki data gaji. Selain itu, dari 6352 observasi tersebut, hanya ada 2 observasi dengan data gaji dalam USD. Untuk itu, data latih yang akan digunakan pada langkah-langkah analisis selanjutnya hanya terdiri 6350 observasi yang memiliki data gaji dalam rupiah.

3.2.1 Job_title

Pada dataset yang diberikan, variabel ‘Job title’ terdiri dari 4314 nilai unik di antara 6346 observasi yang ada, sehingga akan menimbulkan masalah *curse of dimensionality* saat melakukan pemodelan pada variabel yang telah dilakukan *encoding*. Oleh karena itu, pemodelan topik (*topic modelling*) merupakan salah satu cara yang tepat dan efektif untuk membentuk topik yang nantinya akan dijadikan input dalam proses pemodelan data secara keseluruhan. Dengan cara ini, informasi bermacam-macam pekerjaan dapat direpresentasikan dalam vektor berdimensi rendah, tetapi tetap dalam diinterpretasikan dan dipahami. Dalam proses pengolahan fitur ini, metode yang digunakan adalah Latent Dirichlet Allocation (LDA).

Sebelum dilakukan pemodelan topik dengan LDA, teks nama pekerjaan akan diolah (text processing) terlebih dahulu. Langkah pengolahan yang dilakukan adalah sebagai berikut:

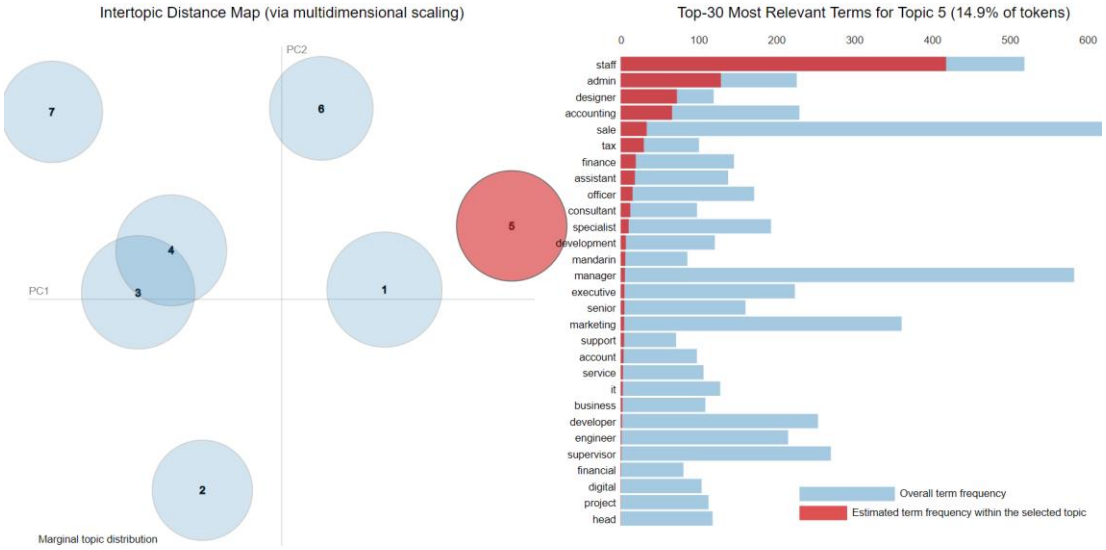
1. Menerjemah nama pekerjaan ke dalam bahasa Inggris dengan bantuan *package googletrans*
2. Melakukan filter dari kata-kata yang merupakan *stopword* serta memilih kata yang memiliki lebih dari 1 huruf
3. Mengubah kata kembali menjadi kata dasar dan menghapus imbuhan/afiks yang terdapat pada kata (*lemmatizing*)
4. Mengubah semua huruf menjadi huruf kecil (*lowercase*)

Setelah kata dalam data teks diolah dan dibersihkan, kata-kata tersebut disimpan di dalam *dictionary* dengan ketentuan sebagai berikut:

a. Menyimpan kata yang terdapat di paling tidak 90 observasi (dokumen)
b. Menyimpan kata yang berada tidak lebih dari 70% dari keseluruhan observasi (korpus)

Kata-kata yang tersimpan dalam dictionary kemudian direpresentasikan dalam ruang vektor berdasarkan frekuensi kemunculannya (bag-of-word). Setelah itu, metode Term Frequency-Inverse Document Frequency (TF-IDF) diterapkan untuk memberi bobot pada setiap kata untuk setiap vektor berdasarkan kemunculan kata dan jumlah observasi/dokumen yang mengandung kata tersebut dalam keseluruhan korpus.

Pada proses pemodelan dengan LDA, jumlah topik yang dipilih adalah sebanyak 7 berdasarkan skor *perplexity* terendah hasil percobaan, dengan parameter optimal sebagai berikut: Alpha : ‘Symmetric’, Eta : ‘0.01’. Visualisasi ketujuh topik yang telah diproyeksikan ke dua dimensi (dua sumbu utama) beserta distribusi frekuensi *term* pada salah satu topik dapat dilihat pada gambar dibawah ini:



Gambar 3.4 Visualisasi Tujuh Topik Hasil LDA pada Dua Dimensi

Topik-topik yang dihasilkan kemudian dikonversi menjadi vektor-vektor, di mana vektor-vektor merepresentasikan masing-masing topik. Pada setiap baris observasi/dokumen, jumlah probabilitas keseluruhan topik adalah sebesar 1, dengan distribusi probabilitas yang berbeda untuk setiap observasi. Vektor-vektor ini akan kemudian dijadikan input untuk proses pemodelan secara keseluruhan.

3.2.2 Location

Proses pengolahan pertama yang dilakukan pada fitur ‘location’ adalah mengekstrak nama provinsi masing-masing daerah dengan bantuan *package geopy*. Setelah itu, dilakukan proses *data binning* terhadap 34 provinsi menjadi 8 tingkat berdasarkan rata-rata gaji pada tiap provinsi tersebut, misalnya, Provinsi Aceh dengan rata-rata gaji 9 juta digolongkan ke dalam tingkat 1, Provinsi Bali dan DKI Jakarta dengan rata-rata gaji 8 juta digolongkan ke dalam tingkat 2, dan sebagainya.

3.2.3 Experience_level

Fitur ‘experience_level’ berisi pengalaman kerja pelamar yang diminta oleh perusahaan. Pada fitur ini, dilakukan *data binning* menjadi 4 kategori. Observasi yang memiliki memiliki ‘experience_level’ kurang dari 3 tahun dikategorikan sebagai ‘fresh’, ‘experience_level’ antara 3 dan 5 tahun dikategorikan sebagai ‘rendah’, ‘experience_level’ antara 6 dan 15 tahun dikategorikan sebagai ‘sedang’ dan

‘experience_level’ lebih dari 15 tahun dikategorikan sebagai ‘tinggi’. *Missing value* diimputasi dengan modus sebelum *data binning* dilakukan.

3.2.4 Education_level

Fitur ‘education_level’ berisi syarat tingkat pendidikan pelamar yang diminta oleh perusahaan. Pada fitur ini, diambil kata paling akhir dari *string* yang ada. *String* paling akhir ini adalah syarat pendidikan tertinggi yang diminta oleh perusahaan. *Missing value* diimputasi dengan modus sebelum *data binning* dilakukan.

3.2.5 Employment_type

Fitur ‘employment_type’ menunjukkan tipe pekerjaan yang dibuka untuk lowongan tersebut. Tipe pekerjaan yang tersedia terdiri dari ‘penuh waktu’, ‘paruh waktu’, ‘kontrak’, ‘temporer’, ‘magang’, dan gabungan dari tipe-tipe tersebut. Untuk mengatasi data lowongan yang memiliki gabungan dari beberapa tipe pekerjaan tersebut, pada fitur ini hanya akan diambil tipe pertama yang tersedia pada data.

3.2.6 Job_function

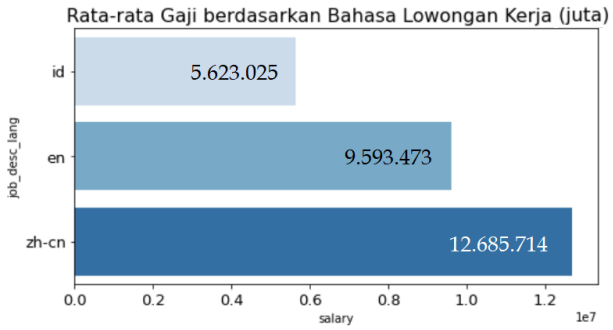
Fitur ‘job_function’ berisi kategorisasi jenis pekerjaan lowongan pekerjaan. Pada fitur ini, dilakukan pemisahan *string* yang ada menjadi 2 fitur baru. *String* pertama disebut sebagai ‘function1’, dan *string* kedua disebut sebagai ‘function2’. *Missing value* diimputasi dengan modus sebelum pemisahan dilakukan.

3.2.7 Company_industry

Fitur ‘company_indsutry’ berisi 58 industri perusahaan yang tersedia pada data. Fitur ini akan disederhanakan dengan melakukan *data binning* menjadi 8 kelompok berdasarkan rata-rata gaji dalam setiap industri. Kelompok pertama terdiri dari industri yang memiliki rata-rata gaji pada rentang 3-4 juta, kelompok kedua untuk industri dengan rata-rata gaji pada rentang 4-5 juta, kelompok ketiga untuk industri dengan rata-rata gaji pada rentang 5-6 juta, dst. Kelompok ke-8 adalah untuk industri dengan rata-rata gaji lebih dari 10 juta. *Missing value* diimputasi dengan modus setelah *data binning* dilakukan.

3.2.8 Job_description

Fitur ‘job_description’ mengandung penjelasan mengenai lowongan pekerjaan yang ditawarkan pada suatu perusahaan. Berdasarkan pengelompokkan gaji pekerjaan terhadap sumber bahasa deskripsi pekerjaan, diperoleh *insight* bahwa lowongan pekerjaan dengan deskripsi berbahasa Mandarin cenderung memiliki gaji yang lebih tinggi, diikuti dengan deskripsi pekerjaan berbahasa Inggris, dan kemudian yang berbahasa Indonesia.



Gambar 3.5

Gambar 3.5 Rata-rata Gaji berdasarkan Bahasa Lowongan Kerja

Oleh karena itu, berdasarkan data di atas, sumber bahasa pada variabel ‘job_description’ dideteksi dengan bantuan *package langdetect*, dan dibuatlah suatu fitur baru (‘job_desc_lang’) yang berisi sumber bahasa lowongan pekerjaan tersebut.

3.2.9 Fitur akhir

Setelah melakukan berbagai macam proses pra-pengolahan (feature engineering), fitur yang akhirnya akan digunakan dalam model adalah sebagai berikut:

Tabel 3.7 Fitur Akhir yang Digunakan pada Model

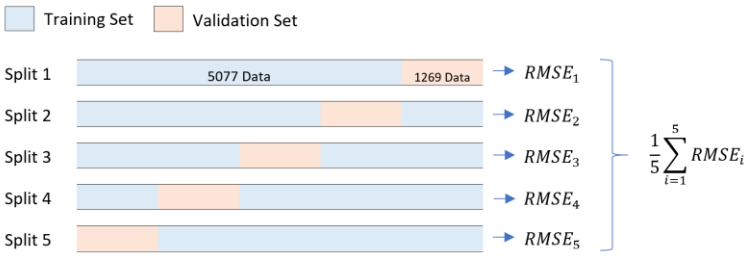
No	Nama Fitur	Keterangan
1	career_level	Tingkat jabatan dari pekerjaan yang ditawarkan
2	‘bin_exp_level’	Fitur ‘experience_level’ setelah dilakukan *data binning*

3	‘highest_edu’	Fitur ‘education_level’ setelah diambil *string* terakhir
4	‘function1’ , ‘function2’	Fitur ‘job_function’ setelah dipisah
5	‘bin_comp_ind’	Fitur ‘company_industry’ setelah dilakukan *data binning*
6	‘bin_employ_type’	Fitur ‘employment_type’ setelah dilakukan *data binning*
7	‘bin_will’	Fitur ‘lokasi’ setelah dilakukan *data binning*
8	‘topic_0’,‘topic_1’, ‘topic_2’,‘topic_3’, ‘topic_4’,‘topic_5’, ‘topic_6’	Fitur ‘job_title’ setelah dilakukan *topic modelling* dengan LDA

3.3 Pemodelan

Setelah data melalui tahap *preprocessing*, selanjutnya akan dilakukan proses *training* model dan pengujian dengan menggunakan metrik yang sudah ditentukan. Tahapan ini dibagi menjadi dua, pemilihan model dan *hyperparameter tuning.*

Pada tahap pemilihan model, calon calon model akan di-*train* dan dibandingkan performanya. Evaluasi dilakukan dengan teknik validasi silang (*cross validation*) dengan lipatan (*folds*) sebesar lima berdasarkan nilai RMSE. Skema validasi model adalah sebagai berikut:



Gambar 3.6 Skema Evaluasi *5-folds Cross Validation*

Model dengan evaluasi terbaik akan dipilih dan akan kembali dikembangkan dalam tahapan *hyperparameter tuning*.

3.3.1 Pemilihan Model

Pada penelitian ini, penulis memutuskan untuk membandingkan performa *tree-based model*, yaitu Random Forest (RF), XGBoost (XGB), Light GBM (LGBM) dan CatBoost (CB), hal ini karena performa *tree-based model* bagus ketika terdapat banyak fitur, dapat mengurangi *overfitting*, dan *robust* terhadap *outliers*. Selain itu, akan digunakan juga model *Voting Regressor* (VR), yang menggabungkan hasil prediksi dari model RF, XGB, LGBM, dan CB dengan menghitung rata-ratanya sebagai hasil akhir prediksi. Tabel 3.8 menunjukkan perbandingan performa model *baseline* berdasarkan metrik RMSE dengan metode 5-*folds cross validation*:

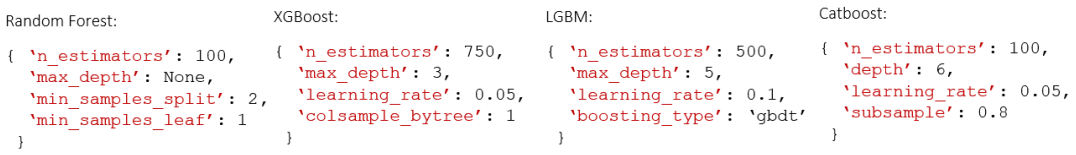
Tabel 3.8 Performa Model *Baseline*

No	Model	RMSE
1	RF	3421615.51
2	XGB	3358319.26
3	LGBM	3338357.84
4	CB	3257923.38
5	VR	**3224748.85**

Dapat dilihat bahwa model *Voting Regressor* memiliki performa terbaik. Karena performa *voting regressor* dipengaruhi oleh performa model RF, XGB, LGBM dan CB, akan dilakukan *hyperparameter tuning* pada model-model tersebut.

3.3.2 Hyperparameter Tuning

Pada tahap ini, akan dilakukan *hyperparameter tuning* pada empat model yang digunakan dalam *voting regressor*. *Hyperparameter tuning* dilakukan dengan metode Randomnized Search. Setelah melakukan *tuning*, didapatkan *hyperparameter* terbaik sebagai berikut:



Gambar 3.7 Hasil *Hyperparameter* Optimal setelah *Hyperparameter Tuning*

Tabel 3.9 menunjukan performa model setelah dilakukan *hyperparameter tuning.* Dapat dilihat bahwa performa model RF dan CB tidak memiliki perubahan, hal ini karena pada saat *tuning*, didapatkan bahwa parameter *default* menghasilkan model dengan performa yang paling baik. Selain itu, model *voting* yang dibangun berdasarkan model RF, XGB, LightGBM, CB yang telah di-*tuning* mengalami penurunan RMSE menjadi 3217929.45, sehingga model *voting regressor* tersebut kemudian dipilih sebagai model final dalam memprediksi gaji.

Tabel 3.9 Performa Model setelah Hyperparameter Tuning

No	Model	RMSE
1	RF	3421615.51
2	XGB	3321827.87
3	LGBM	3332512.63
4	CB	3257923.38
5	VR	**3217929.45**

3.4 Implementasi Model

Model ini dapat digunakan untuk membantu para calon pekerja/pencari kerja mengetahui perkiraan gaji yang akan mereka peroleh pada suatu lowongan pekerjaan. Dengan menggunakan model ini, dapat dibangun suatu *website*, aplikasi, ataupun *extension* pada *browser* yang menerima *input* berupa data lowongan kerja dengan fitur-fitur yang ada pada model dan menerima *output* berupa hasil prediksi gaji pada lowongan tersebut.

BAB IV

Kesimpulan

4.1 Kesimpulan

Berdasarkan hasil analisis dari penelitian kami, dapat disimpulkan:

1. Industri perusahaan dengan rata-rata gaji tertinggi adalah Pertambangan, diikuti dengan Manajemen/Konsulting HR dan Hukum/Legal.
2. Provinsi Aceh, DKI Jakarta, dan Bali menjadi 3 provinsi dengan rata-rata gaji lowongan kerja yang tertinggi.
3. Rata rata gaji lowongan kerja dengan syarat pendidikan tertinggi sarjana lebih tinggi 85% dibandingkan SMA/SMU/SMK/STM. Sehingga dapat dilihat bahwa perusahaan lebih mengapresiasi calon pekerja dengan tingkat pendidikan yang lebih tinggi.
4. Model akhir yang digunakan adalah model *voting regressor* berdasarkan model RF, XGB, LightGBM, dan Catboost dengan nilai RMSE sebesar 3217929.45 pada data validasi.

4.2 Saran

Beberapa saran untuk penelitian kedepannya adalah sebagai berikut:

1. Menggunakan model *semi supervised learning* dalam pembangunan model sehingga data *training* dengan label yang hilang dapat digunakan.
2. Menggunakan data eksternal untuk memperoleh fitur lain yang dapat meningkatkan performa model.

Daftar Pustaka

Setijohatmo, U., Rachmat, S., Susilawati, T. dan Rahman, Y. (2020). Analisis Metode Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik. Jurnal IRWNS, vol. 11, no. 1, pp. 402-408.

Dewi, N., Syafitri, U. dan Mulyadi, S. (2011). The Application of Random Forest in Driver Analysis. Forum Statistika dan Komputasi, vol. 16, no. 1, pp. 35- 43.

Hendro, G., Adji, T. dan Setiawan, N. (2012). Penggunaan Metodologi Analisa Komponen Utama (PCA) untuk Mereduksi Faktor-Faktor yang Mempengaruhi Penyakit Jantung Koroner. Malang: Universitas Brawijaya.

Maarif, A. (2015). Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah. Semarang: Fakultas Ilmu Komputer Universitas Dian Nuswantoro