

Pemodelan Gaji pada Lowongan Kerja dengan *Tree-Based Model*

Akhirnya
Belajar

Evan Eka Wijaya
Maxwell Thomson
Melvin Putra

Daftar Isi

1. Pendahuluan

2. Exploratory Data Analysis

3. Pemodelan

4. Kesimpulan dan Saran





Pendahuluan

Keinginan Pencari Kerja yang Tak Terjawab

PT Cemerlang Majaya

Beberapa lokasi kerja ▼

IDR 5.000.000 - IDR 7.000.000

Ditayangkan pada 14-May-22



PT Alfabeta Indonesia

Cirebon

Ditayangkan pada 13-May-22



PT Saison Modern Finance

Jakarta Selatan

Ditayangkan pada 21 jam yang lalu



Berdasarkan Survey

Oleh HAYS - What Workers Want 2018

- **61%** pencari kerja **ingin melihat gaji** suatu lowongan kerja
- Tetapi, hanya **46%** lowongan kerja yang **memiliki informasi gaji**





Exploratory Data Analysis

Dataset Overview

Persentase Missing Value - Target

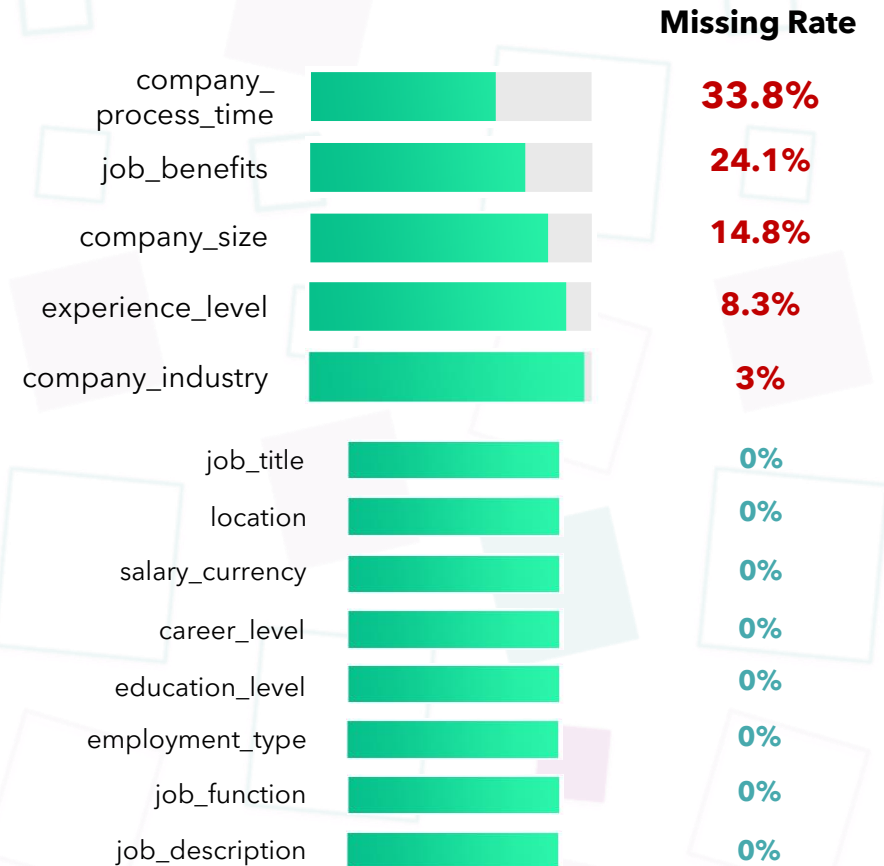


Distribusi Gaji



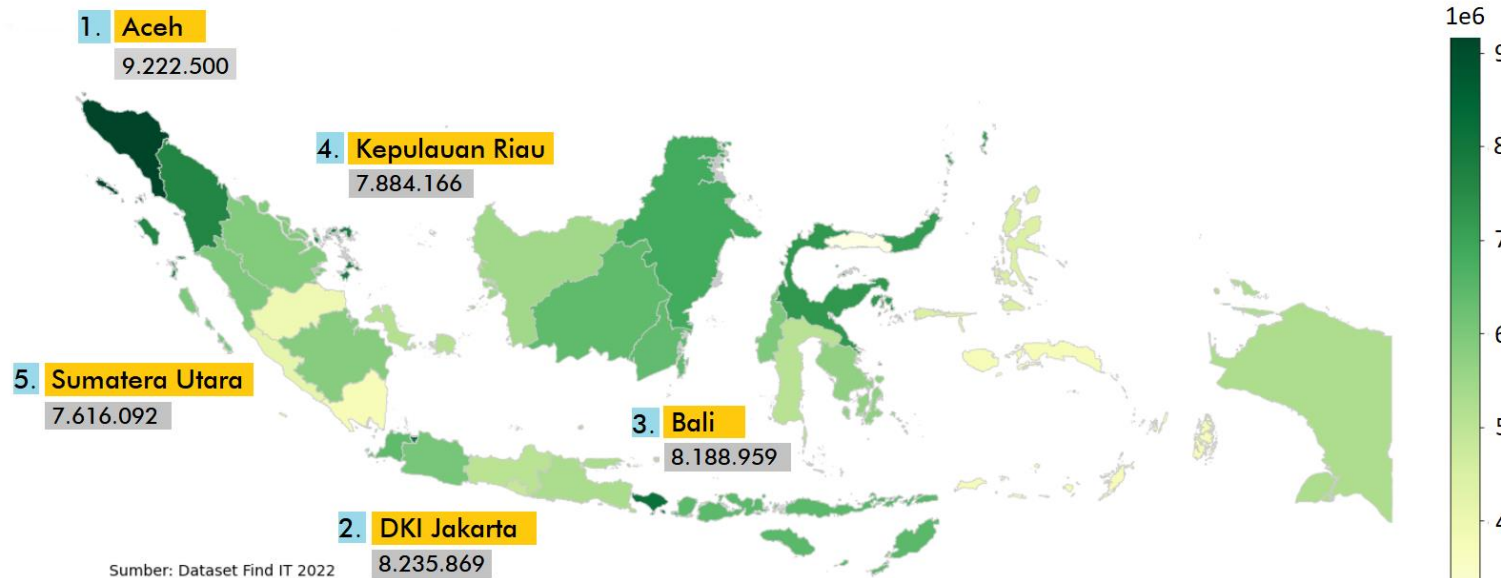
Right Skewed Distributed
(Skewness = 6.26)

Persentase Missing Value - Features



Rata-Rata Gaji: Lokasi

Rata rata Gaji Berdasarkan Provinsi (rupiah)



Persentase Industri Top 5 Provinsi

1. Aceh

1. Makanan & Minuman **81.4%**
2. Kesehatan/Medis 8.06%

2. DKI Jakarta

1. Umum & Grosir **8.21%**
2. Retail/Merchandise 8.14%

3. Bali

1. Komputer/IT **19.13%**
2. Manajemen/Konsulting HR 10.38%

4. Kepulauan Riau

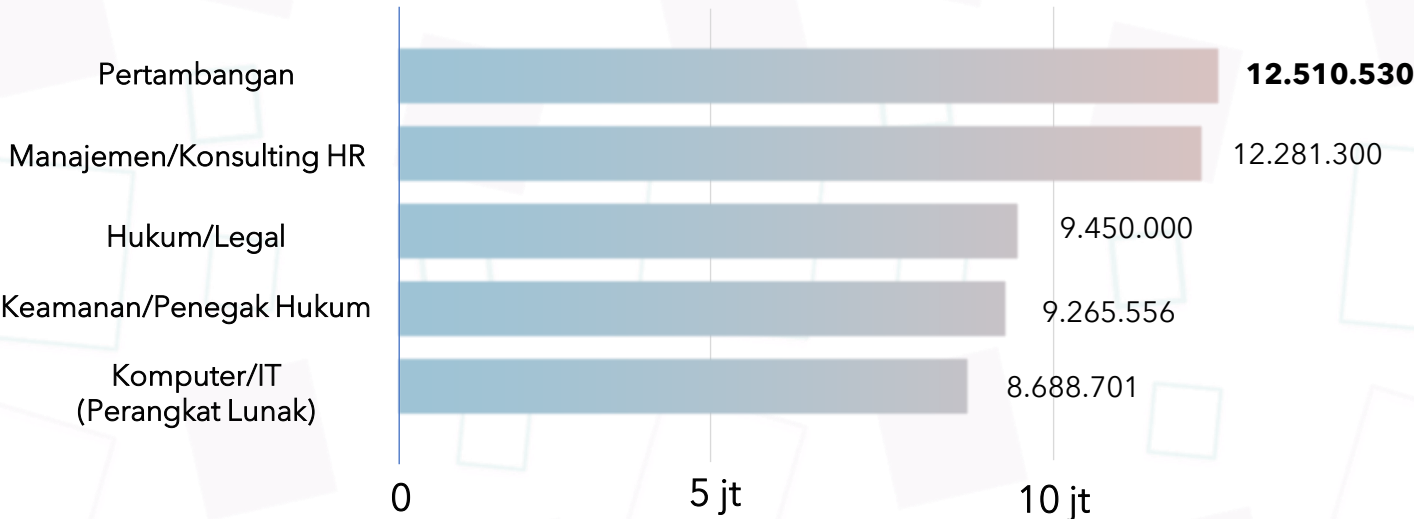
1. Manufaktur/Produksi **17.24%**
2. Manajemen/Konsulting HR 10.34%

5. Sumatera Utara

1. Makanan & Minuman **36.93%**
2. Manufaktur/Produksi 10.8%

Rata-Rata Gaji: Industri Perusahaan

Rata rata Gaji Berdasarkan Jenis Industri (rupiah)



Distribusi Top 3 Pekerjaan Gaji Tertinggi (berdasarkan Jenis Industri)

Pertambangan:

- 1. Pemasaran/Pengembangan Bisnis 1.75%
- 2. Pengacara/Asisten Legal 5.26%
- 3. Pemeliharaan 1.75%

Manajemen/Konsulting HR:

- 1. Keuangan/Investasi Perusahaan 0.25%
- 2. Top Management 0.25%
- 3. Properti Real Estate 0.74%

Hukum/Legal:

- 1. IT-Perangkat Lunak 41.67%
- 2. Penjualan Ritel 8.33%
- 3. Pengacara/Asisten 25%

Keamanan/Penegak Hukum:

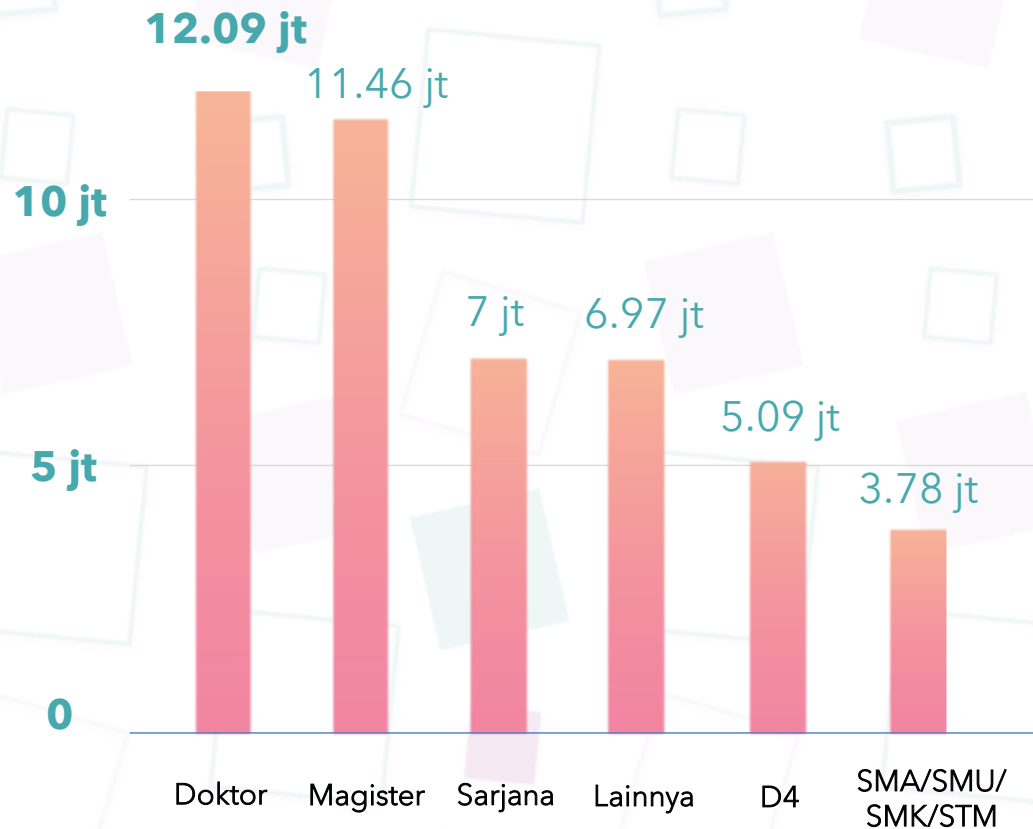
- 1. Angkatan Bersenjata 33.33%
- 2. Teknik Sipil/Konstruksi Bangunan 11.11%
- 3. IT-Perangkat Lunak 11.11%

Komputer/IT:

- 1. Pelatihan & Pengembangan 0.76%
- 2. Merchandising 0.25%
- 3. Digital Marketing 3.80%

Rata-Rata Gaji: Syarat Pendidikan Tertinggi

Rata rata Gaji Berdasarkan Syarat Pendidikan Tertinggi (rupiah)

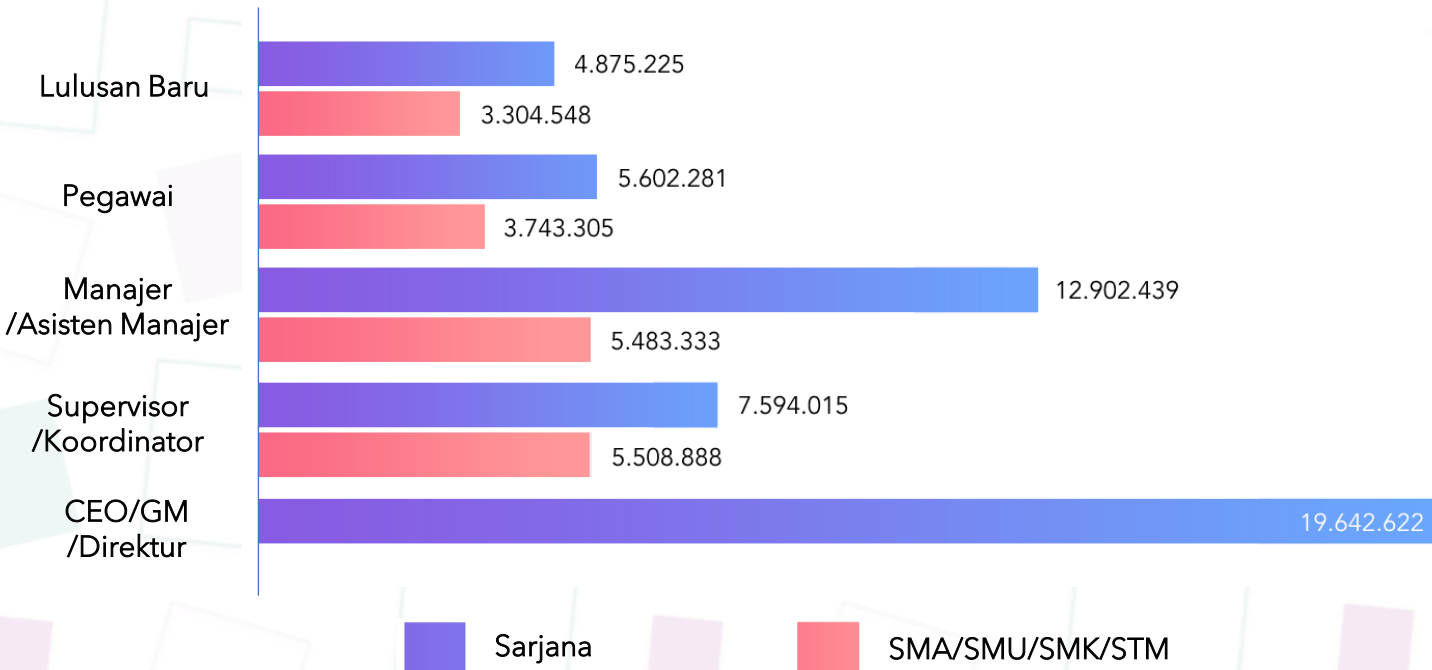


Lowongan kerja dengan syarat pendidikan **Doktor** memiliki rata rata gaji tertinggi. Sementara itu **SMA/SMU/SMK/STM** memiliki gaji rata rata terendah.

Syarat pendidikan sarjana meningkatkan **85%** rata-rata gaji dibandingkan SMA/SMU/SMK/ STM

Rata-Rata Gaji: Syarat Pendidikan Tertinggi

Rata rata Gaji Berdasarkan Jabatan (rupiah)



Persentase Jenis Jabatan

Sarjana

1. Pegawai 61.71%
2. Supervisor 17.88%

SMA/SMU/SMK/STM

1. Pegawai 71%
2. Lulusan Baru 21.14%

Lowongan kerja dengan syarat pendidikan **sarjana** memiliki rata rata gaji **lebih tinggi** pada seluruh jabatan.

Selain itu, **tidak** terdapat lowongan dengan jabatan CEO/Direktur bagi lulusan SMA/SMU/SMK/STM.

Rata-Rata Gaji: Syarat Pendidikan Tertinggi

Top 5 Industri Gaji Tertinggi - Sarjana



Top 5 Industri Gaji Tertinggi - SMA/SMU/SMK/STM



Persentase Industri Pekerjaan

Sarjana

1. Manufaktur/Produksi 10.7%
2. Komputer 7%

SMA/SMU/SMK/STM

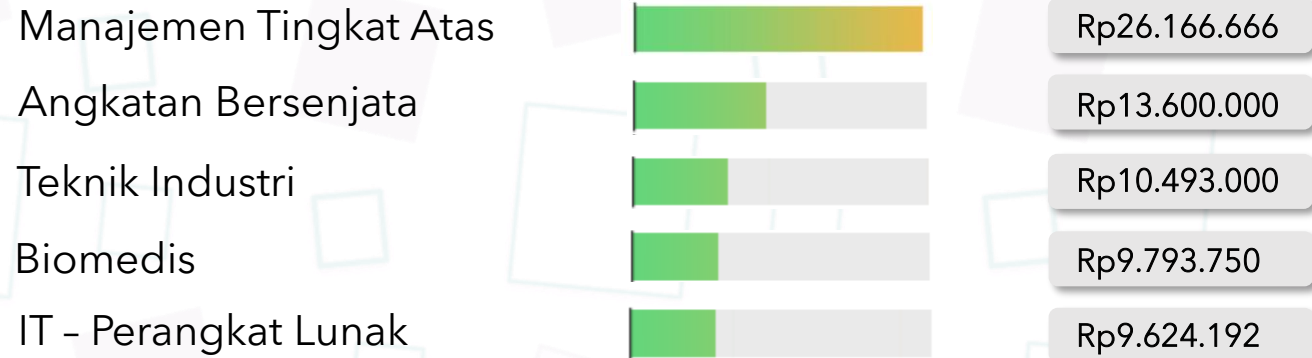
1. Makanan & Minuman 15.2%
2. Retail/Merchandise 10.45%

Sektor Pertambangan menjadi industri dengan gaji yang tinggi baik bagi lulusan Sarjana maupun SMA/SMU/SMK/STM

15% Industri bagi lulusan SMA/SMU/SMK/STM didominasi oleh sektor Makanan & Minuman

Rata-Rata Gaji: Syarat Pendidikan Tertinggi

Top 5 Pekerjaan Gaji Tertinggi - Sarjana



Top 5 Pekerjaan Gaji Tertinggi - SMA/SMU/SMK/STM



Persentase Lowongan Pekerjaan

Sarjana

1. Penjualan Ritel 11.31%
2. IT-Software 10.25%

SMA/SMU/SMK/STM

1. Penjualan Ritel 23.75%
2. Makanan dan Restoran 14.96%

Lulusan SMA/SMU/SMK/STM mendapatkan rata-rata gaji paling tinggi dengan bekerja sebagai seorang **praktisi**

Penjualan ritel menjadi bidang pekerjaan paling umum bagi para lulusan Sarjana dan SMA/SMU/SMK/STM



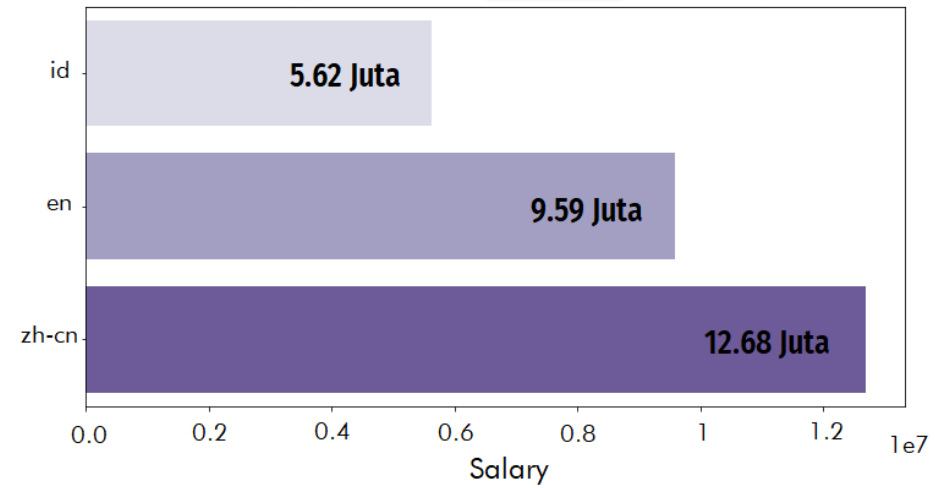
Pemodelan/Modelling

Job_description

Dengan package *langdetect*, membuat kolom baru yang berisi bahasa dari 'job_description'

Job_description	Job_desc_lang
Tanggung jawab :Mengumpulkan, menganalisa....	id
Your responsibilities:The job....	en
年龄：22-27岁精通中文，印尼文（HSK5级以上）...	zh-cn

Rata-rata Gaji berdasarkan Sumber Bahasa



Lowongan pekerjaan dengan deskripsi **Bahasa Mandarin** memiliki rata rata gaji yang lebih tinggi, diikuti oleh Bahasa Inggris, lalu Bahasa Indonesia.

Education_level

Mengambil *string* terakhir (kecuali 'Tidak Terspesifikasi') untuk mendapatkan syarat pendidikan tertinggi dari lowongan pekerjaan.

'Tidak terspesifikasi'

'SMA, SMU/SMK/STM'

'SMA, SMU/SMK/STM, Sertifikat Professional, D3 (Diploma), **D4 (Diploma)**'

'Sertifikat Professional, D3 (Diploma), D4 (Diploma), **Sarjana (S1)**'

'Sarjana (S1), Diploma Pascasarjana, Gelar Professional, **Magister (S2)**'

'Sarjana (S1), **Doktor (S3)**'



education_level	highest_edu
Tidak Terspesifikasi	Lainnya
SMA, SMU/SMK/STM	SMA/SMU/SMK/STM
..... D4 (Diploma)	D4
..... Sarjana (S1)	Sarjana
..... Magister (S2)	Magister
..... Doktor (S3)	Doktor

Job_function

Memisahkan string menjadi dua bagian. Pemisahan dilakukan pada tanda baca koma

'Manufaktur,Pembelian/Manajemen Material'
'Komputer/Teknologi Informasi,IT-Perangkat Lunak'
'Sumber Daya Manusia/Personalia,Sekretaris'
'Sains,Aktuaria/Statistik',
'Lainnya,Jurnalis/Editor'

job_function	function1	function2
Manufaktur,Pembelian/Manajemen Material	Manufaktur	Pembelian/Manajemen Material
Komputer/Teknologi Informasi,IT-Perangkat Lunak	Komputer/Teknologi Informasi	IT-Perangkat Lunak
Sumber Daya Manusia/Personalia,Sekretaris	Sumber Daya Manusia/Personalia	Sekretaris
Sains,Aktuaria/Statistik	Sains	Aktuaria/Statistik
Lainnya,Jurnalis/Editor	Lainnya	Jurnalis/Editor

Data Binning

Fresh : < 3 tahun

Rendah : 3 - 5 tahun

Sedang : 6 - 15 tahun

Tinggi : > 15 tahun

Experience Level

'Penuh Waktu, Magang'

'Penuh Waktu, Kontrak' → **'Penuh Waktu'**

'Penuh Waktu, Paruh Waktu'

Reduced to 5 levels (37.5%)

Employment Type

Location

Tier 1 : > 9 juta*

Tier 2 : 8 - 9 juta

...

Tier 8 : 2 - 3 juta

Reduced to 8 levels (76.4%)

*rata-rata gaji berdasarkan provinsi

Company Industry

Tier 1 : 3 - 4 juta*

Tier 2 : 4 - 5 juta

...

Tier 8 : > 10 juta

Reduced to 8 levels (85.9%)

*rata-rata gaji berdasarkan jenis industri

Job Title – Topic Model

Memperoleh vektor topik dengan (Latent Dirichlet Allocation) **LDA** sebagai input pemodelan

Data awal

4314 unique job titles

job_title
Social Media Specialist
Tim Creative
ELECTRONIC ENGINEER
KOKI / JURU MASAK
Web Developer

Tahap 1

- Menerjemah kata ke dalam **Bahasa Inggris** dengan *googletrans*
- Mengubah kata menjadi **lowercase**

job_trans
social media specialist
tim creative
electronic engineer
chef / cook
web developer

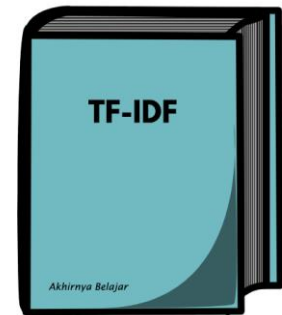
Tahap 2

- Menghapus karakter **non-alphabetic**
- Tokenisasi

```
[['social', 'medium', 'specialist'],  
 ['tim', 'creative'],  
 ['electronic', 'engineer'],  
 ['chef', 'cook'],  
 ['web', 'developer']]
```

Tahap 3

- Menyimpan teks dalam *dictionary*
- Merepresentasikan teks dalam bentuk vektor dengan **TF-IDF**



Tahap Akhir

- Memperoleh vector akhir dengan pemodelan topik **LDA**

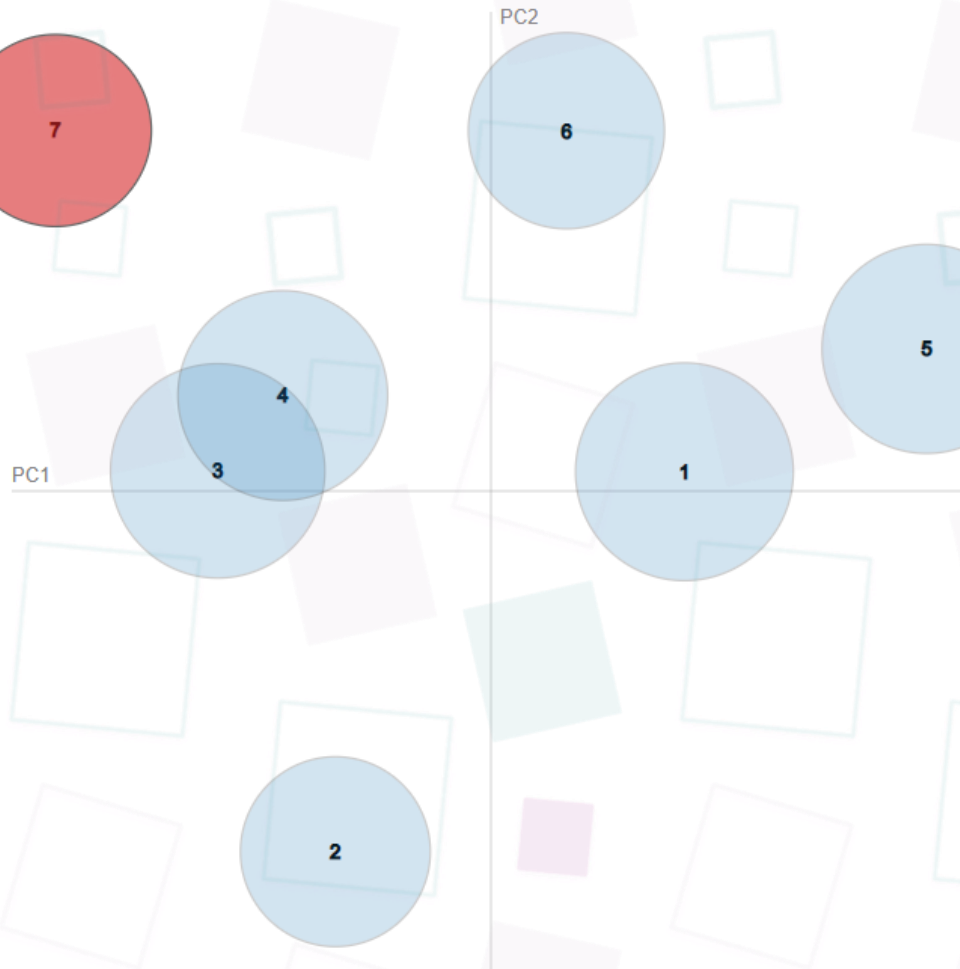
Optimal Parameter*: {**num_topics**:7, **alpha**:‘symmetric’, **eta**:0.01}

**berdasarkan perplexity score terendah*

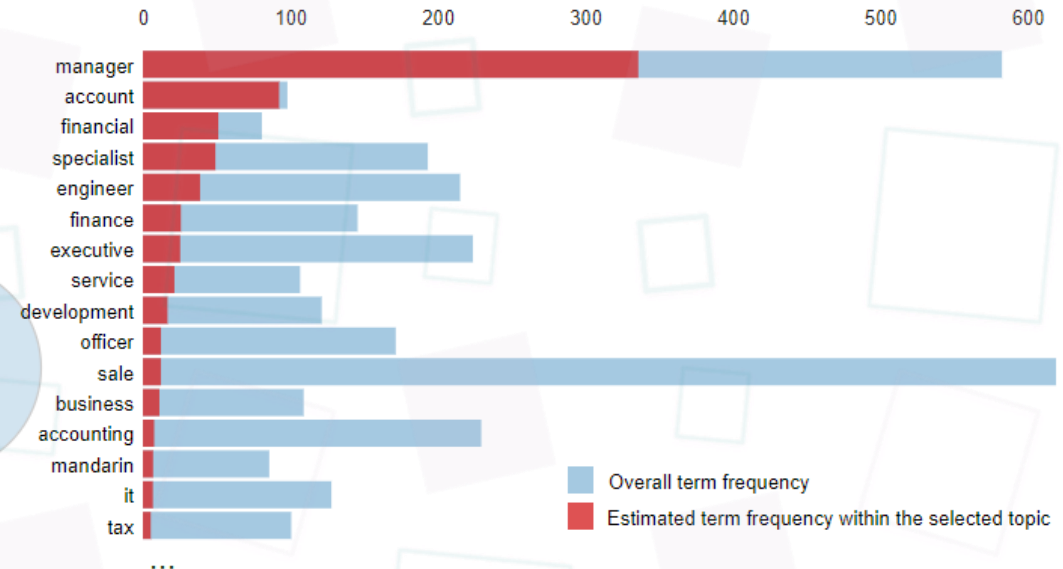


Job Title – Topic Model

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 7 (12.6% of tokens)



Final Features

Nama Fitur	Keterangan
'bin_comp_ind'	Fitur 'company_industry' setelah dilakukan <i>data binning</i>
'bin_employ_type'	Fitur 'employment_type' setelah dilakukan <i>data binning</i>
'bin_wil'	Fitur 'lokasi' setelah dilakukan <i>data binning</i>
'topic_0', 'topic_1', 'topic_2', 'topic_3', 'topic_4', 'topic_5', 'topic_6'	Fitur 'job_title' setelah dilakukan <i>topic modelling</i> dengan LDA

Nama Fitur	Keterangan
'career_level'	Tingkat jabatan dari pekerjaan yang ditawarkan
'bin_exp_level'	Fitur 'experience_level' setelah <i>data binning</i>
'highest_edu'	Fitur 'education_level' setelah diambil <i>string</i> terakhir
'function1', 'function2'	Fitur 'job_function' setelah dipisah

Teknik Evaluasi

Metrik

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

Dimana:

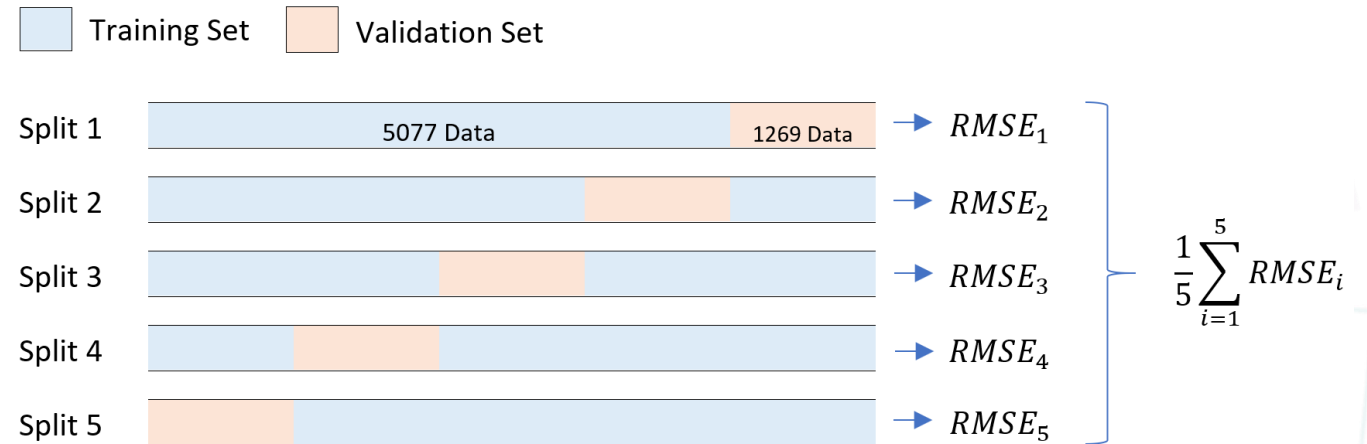
y_i : actual value

\hat{y}_i : predicted value

n : jumlah observasi

Metode Evaluasi

5-Fold Cross Validation



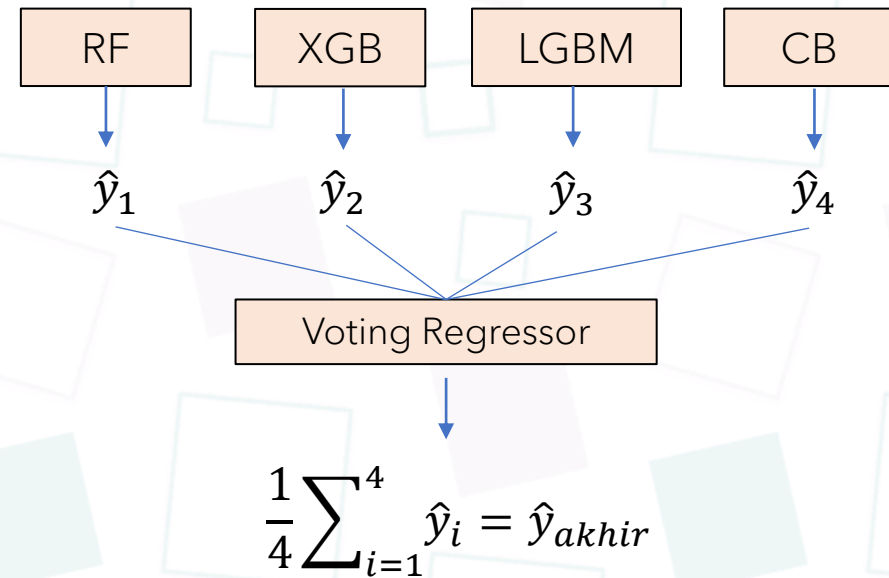
Performa Base Model

Hasil CV-5 Folds

Model	RMSE
RandomForest	3421615.51
XGBoost	3358319.26
LightGBM	3338357.84
CatBoost	3257923.38
Voting Regressor	3224748.85

Model **Voting Regressor** menghasilkan RMSE paling kecil dari antara model lainnya.

Ilustrasi Voting Regressor



Hyperparameter Tuning

Hyperparameter Optimal:

Random Forest:

```
{ 'n_estimators': 100,  
  'max_depth': None,  
  'min_samples_split': 2,  
  'min_samples_leaf': 1  
}
```

XGBoost:

```
{ 'n_estimators': 750,  
  'max_depth': 3,  
  'learning_rate': 0.05,  
  'colsample_bytree': 1  
}
```

LGBM:

```
{ 'n_estimators': 500,  
  'max_depth': 5,  
  'learning_rate': 0.1,  
  'boosting_type': 'gbdt'  
}
```

Catboost:

```
{ 'n_estimators': 100,  
  'depth': 6,  
  'learning_rate': 0.05,  
  'subsample': 0.8  
}
```

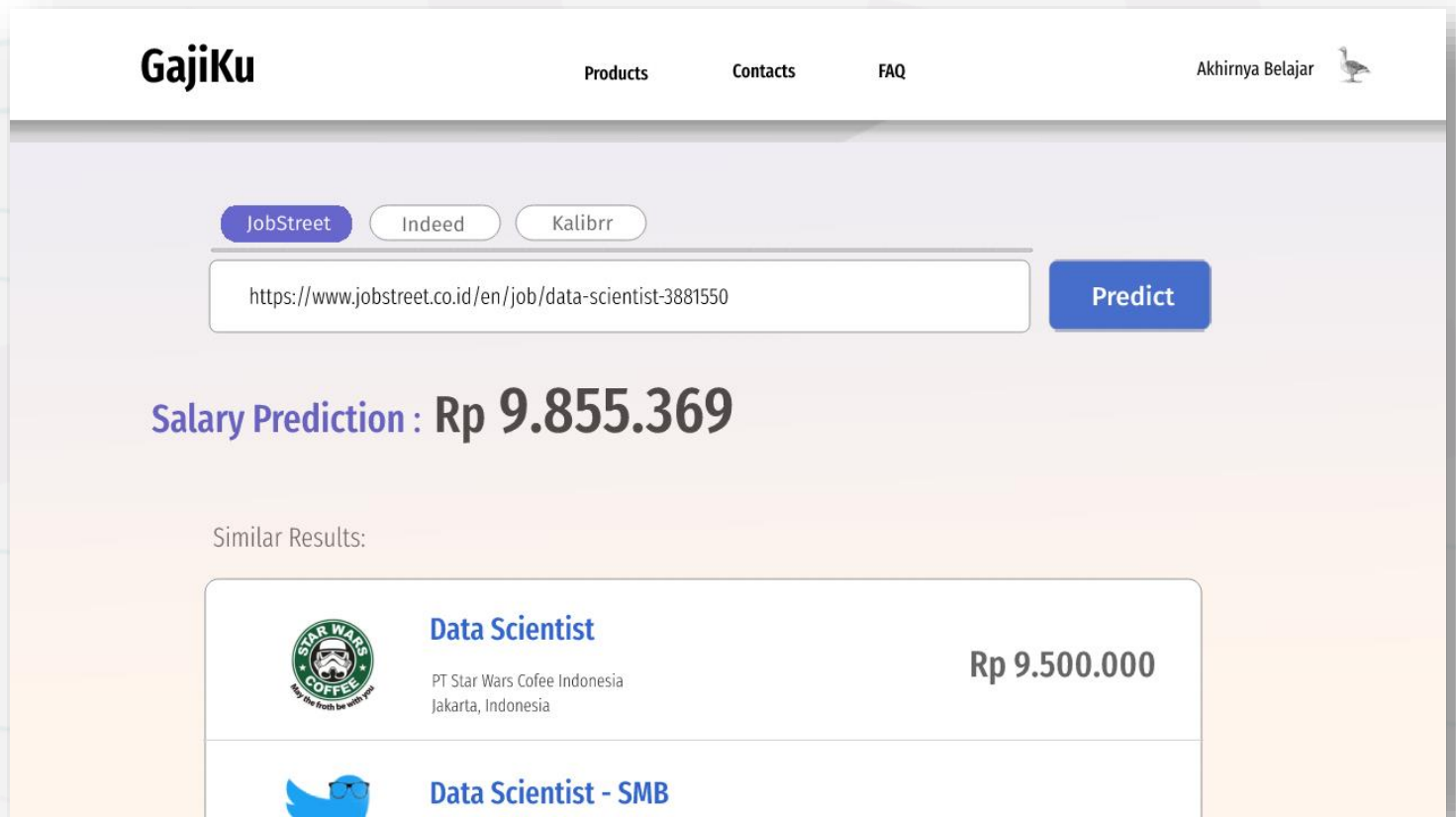
Model	RMSE
RandomForest	3421615.51
XGBoost	3321827.87
LightGBM	3332512.63
CatBoost	3257923.38
Voting Regressor	3217929.45

Setelah Hyperparameter Tuning, model baseline RandomForest dan CatBoost memiliki performa paling baik.



Implementasi Model

GajiKu

Situs prediksi gaji secara *real-time* berbasis *machine-learning*



The image shows a website mockup for 'GajiKu'. The header includes the 'GajiKu' logo, navigation links for 'Products', 'Contacts', and 'FAQ', and a user profile 'Akhirnya Belajar' with a bird icon. Below the header, there are three tabs: 'JobStreet' (selected), 'Indeed', and 'Kalibrr'. A text input field contains the URL 'https://www.jobstreet.co.id/en/job/data-scientist-3881550', and a blue 'Predict' button is to its right. The main content area displays the 'Salary Prediction : Rp 9.855.369'. Below this, a section titled 'Similar Results:' contains two job listings. The first listing is for 'Data Scientist' at 'PT Star Wars Cofee Indonesia' in 'Jakarta, Indonesia', with a salary of 'Rp 9.500.000' and a 'STAR WARS COFFEE' logo. The second listing is for 'Data Scientist - SMB' with a Twitter bird icon.

Logo	Job Title	Company	Location	Salary
	Data Scientist	PT Star Wars Cofee Indonesia	Jakarta, Indonesia	Rp 9.500.000
	Data Scientist - SMB			

Website Mockup



Kesimpulan & Saran

Kesimpulan

1. Industri perusahaan dengan rata rata gaji tertinggi adalah **Pertambangan**, diikuti dengan **Manajamen/Konsulting HR** dan **Hukum/Legal**.
2. Provinsi **Aceh, DKI Jakarta** dan **Bali** menjadi 3 provinsi dengan rata rata gaji lowongan kerja 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 **Random Forest, XGBoost, LightGBM** dan **Catboost** dengan nilai RMSE sebesar 3217929.45 pada data validasi.

Saran

1. Menggunakan model **semi supervised learning** dalam pembangunan model sehingga data *training* dengan label 'salary' yang *missing* 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

Terima Kasih

