Mini Project - Penggalian Data

Nama Kelompok :

Anggota :

202210370311046 - Muhammad Syahrul Bachtiar

202210370311058 - Wempy Aditya Wiryawan

202210370311063 - Andika Nur Islamy

Tahap 0 (poin: 25): Business Objective

Dalam era digital, pasar kerja semakin dinamis dengan persaingan yang ketat. Banyak pencari kerja mengalami kesulitan dalam menemukan pekerjaan yang sesuai dengan keterampilan dan pengalaman mereka, sementara perusahaan juga menghadapi tantangan dalam menemukan kandidat yang tepat.

Sistem rekomendasi pekerjaan berbasis Graph-based Representation bertujuan untuk:

- Membantu pencari kerja menemukan pekerjaan yang paling sesuai berdasarkan keterampilan dan pengalaman mereka.
- Meningkatkan efisiensi perekrutan dengan memberikan rekomendasi yang lebih akurat kepada perusahaan.
- Menggunakan pendekatan graph untuk memahami hubungan kompleks antara keterampilan, pengalaman, dan lowongan pekerjaan.

Dengan pendekatan berbasis graph, sistem ini dapat menangkap hubungan tidak langsung antara pengguna dan pekerjaan, sehingga memberikan rekomendasi yang lebih relevan dibandingkan pendekatan tradisional.

Tahap 1 (poin: 25): Original Data

Tabel 1. 1 dataset informasi profil pengguna platform linkedin sebelum dilakukan seleksi fitur manual

timestamp	id	name	city	country_code	region
2023-01-08	irma-arevalo-ccp-phr-shrm-cp-722950a0	Irma Arevalo, CCP, PHR, SHRM CP	Houston, Texas, United States	US	
2021-07-07	anabel-pangalangan-6496095a	ANABEL PANGALANGAN	Rome, Latium, Italy	IT	EU
2023-01-06	kevinjcain	Kevin Cain	Miami, Florida, United States	US	
2023-01-13	dana-evans-6054a051	Dana Evans	Portland, Oregon, United States	US	
2021-06-08	marcgrgic	Marc Grgić	Innsbruck, Tyrol, Austria	AT	EU

current_company:company_id	current_company:name	position	following	about	posts
metropolitan-transit-authority-of-harris-county	Metropolitan Transit Authority of Harris County	Director Compensation at Metropolitan Transit Authority of Harris County		CCP, PHR, SHRM-CP	[{"attribution":"Liked by Irma Arevalo, CCP, PHR, SHRM CP","link":"https://www.linkedin.com/posts/un
	GIANBEL ITALIA	Titolare d'Azienda presso GIANBEL ITALIA	81.0		
	Foodservice	Operations		Operations	[{"attribution":"Like

	Industry	Consultant for Multi-Unit Restaurants and Retail Chains		Engineering Consultant for Multi-Unit Restaurants and Retail Kevin Cain was part of the t	d by Kevin Cain","img":"https://media.licdn.com/dms/image/sync/C5627AQEdt-Y9eBU
	Evans & Associates	Inventory Manager and Business Analyst at Nordstrom			[{"attribution":"Liked by Dana Evans","img":"https://static.licdn.com/sc/h/51xdp0u9qm8nhegwpxoxgd0mb
ihw	IHW Ingenieurbüro Huber GmbH	Branch Manager bei IHW Ingenieurbüro Huber GmbH	131.0	As a fire safety engineer, I try to save peoples lifes with strategic planning and I apply this mind	

groups	current_company	experience	url	people_also_viewed	educations_details
	{"company_id":"metropolitan-transit-authority-of-harris-county","industry":"Truck Transportation","l	[{"company":"Metropolitan Transit Authority of Harris County","company_id":"metropolitan-transit-aut	https://www.linkedin.com/in/irma-arevalo-ccp-phr-shrm-cp-722950a0	[{"profile_link":"https://www.linkedin.com/in/janice-gambrell-mba-cpm-cps-a8731064"},{"profile_link"	Houston International University
[{"img":null,"subtitle":"-","title":"AAA	{"link":null,"name":"GIANBEL	[{"company":"GIANBEL	https://www.linkedin.com/in/anabel-pang		Tarlac State University

gents group"},{"img":null,"subtitle":"-","title":"Hayden Jam	ITALIA"}	ITALIA","company_id":null,"location":"Roma, Italia","positions":[{"description"	alangan-6496095a		
	{"link":null,"name":"Foodservice Industry"}	[{"company":"Food service Industry","location":"Miami/Fort Lauderdale Area","positions":[{"descriptio	https://www.linkedin.com/in/kevinjcain	[{"profile_link":"https://do.linkedin.com/in/daysi-tavarez-82842b124"},{"profile_link":"https://www.	Florida International University
	{"link":null,"name":"Evans & Associates"}	[{"company":"Evans & Associates","location":"","positions":[{"description":"","duration":"Jun 2020 -	https://www.linkedin.com/in/dana-evans-6054a051	[{"profile_link":"https://www.linkedin.com/in/andria-ghahramani-7a315a117"},{"profile_link":"https:/	
	{"company_id":"ihw","industry":"Architecture and Planning","link":"https://www.linkedin.com/company/	[{"company":"IHW Ingenieurbüro Huber GmbH","company_id":"ihw","industry":"Architecture and Planning"	https://www.linkedin.com/in/marcgrgic	[{"profile_link":"https://de.linkedin.com/in/peter-zoder-44026261"},{"profile_link":"https://at.link	Donau-Universität Krems Education Donau-Universität Krems Donau-Universität Krems MScFire Safety Man

education	avatar	languages	certifications	recommendat	recommendations_co	volunteer_experience	courses

				ions	unt		
[{"degree":"Public Administration","end_year":"1989","field":"","meta":"1988 - 1989 Public Administr		[{"subtitle":"-","title":"English"}]	[{"meta":"Issued Dec 1997 Expires Jun 2025 See credential","subtitle":"HR Certification Institute -			[{"cause":"","duration":"Jan 2011","duration_short":"","end_date":"","info":"Volunteer Finance Counc	
[{"degree":"","end_year":"","field":"","meta":"","start_year":"","title":"Tarlac State University","	https://static-exp1.licdn.com/sc/h/244xhbkr7g40x6bsu4gi6q4ry						
[{"degree":"B.S.","end_year":"2002","field":"Industrial and Systems Engineering","meta":"1998 - 2002	https://static.licdn.com/sc/h/244xhbkr7g40x6bsu4gi6q4ry						
	https://media.licdn.com/dms/image/C5103AQH5-S4aXTWTUw/profile-displayphoto-shrink_800_800/0/15165123						
[{"degree":"MSc"	https://media-exp	[{"subtitle":"					

,"end_year":"2021","field":"Fire Safety Management","meta":"2019 - 2021","start_year	1.licdn.com/dms/image/C5603AQGT8Bis8XsO-w/profile-displayphoto-shrink_200_200/0/156	Native or bilingual proficiency","title":"Deutsch"},{"subtitle":"Full professional pro					

Tabel 1.2 dataset list postingan job yang ada dalam platform linkedin sebelum dilakukan seleksi fitur manual

job_id	company_name	title	description	max_salary	pay_period
3902944011	Current Power	Senior Automation Engineer - Power Systems	The Senior Automation / Power Systems Engineer will primarily be responsible for the conception, des		
3901960222	DISH Network	DISH Installation Technician - Field	Company Summary DISH, an EchoStar Company, has been reimagining the future of connectivity for more	19.75	HOURLY
3900944095	Coca-Cola Bottling Company UNITED, Inc.	Order Builder	Division: North Alabama Department : Oxford Warehouse Loading		

			Work Schedule: M-F 8 am UNTIL finish		
3903878594	Denver7 (KMGH-TV)	Mountain Multimedia Journalist, KMGH	KMGH, the E.W. Scripps Company ABC affiliate in Denver, Colorado is looking for a Multimedia Journal		
3905670593	BAYADA Home Health Care	Licensed Practical Nurse (LPN)	Come for the Flexibility, Stay for the Culture Needing more ‘life’ in your work-life balance? Apply	35.0	HOURLY

location	company_id	views	med_salary	min_salary	formatted_work_type
Houston, TX	760913.0	22.0			Full-time
Orange, TX	4296.0	5.0		19.75	Full-time
Oxford, AL	136791.0	4.0			Full-time
Denver, CO	11500365.0	4.0			Full-time
Teterboro, NJ	19472.0	4.0		30.0	Full-time

applies	original_listed_time	remote_allowed	job_posting_url	application_url	application_type
	1713279681000.0		https://www.linkedin.com/jobs/view/3902944011/?trk=jobs_biz_	https://fa-elpm-saasfaprod1.fa.ocs.oraclecloud.com/hc	OffsiteApply

			prem_srch	mUI/CandidateExperience/en/sites/CX/job/1883/?u	
	1713476974000.0		https://www.linkedin.com/jobs/view/3901960222/?trk=jobs_biz_prem_srch	https://jobs.dish.com/jobs/82101?lang=en-us&utm_source=Linkedin	OffsiteApply
	1713387988000.0		https://www.linkedin.com/jobs/view/3900944095/?trk=jobs_biz_prem_srch	https://careers.cokeonena.com/united/job/Oxford-Order-Builder-AL-36203/1156817700/?utm_source=LINKED	OffsiteApply
	1713495697000.0		https://www.linkedin.com/jobs/view/3903878594/?trk=jobs_biz_prem_srch	https://scripps.wd5.myworkdayjobs.com/Scripps_Careers/job/Denver-CO---KMGHKCDOKSBSKZCO/Mountain-Mult	OffsiteApply
	1713398400000.0		https://www.linkedin.com/jobs/view/3905670593/?trk=jobs_biz_prem_srch	https://jsv3.recruitics.com/redirect?rx_cid=3588&rx_url=https%3A%2F%2Fjobs.bayada.com%2Fen%2Fjobs%2F	OffsiteApply

expiry	closed_time	formatted_experience_level	skills_desc	listed_time	posting_domain	sponsored
1715871681000.0		Mid-Senior level		1713279681000.0		0
1716069723000.0				1713477723000.0	jobs.dish.com	0
1715980700000.0		Entry level		1713388700000.0	careers.cokeonena.com	0
1716088348000.0		Entry level		1713496348000.0	scripps.wd5.myworkdayjobs.com	0
1716112542000.0		Entry level		1713520542000.0	jsv3.recruitics.com	0

work_type	currency	compensation_type	normalized_salary	zip_code	fips
FULL_TIME				77002.0	48201.0
FULL_TIME	USD	BASE_SALARY	41080.0	77630.0	48361.0
FULL_TIME				36203.0	1015.0
FULL_TIME				80202.0	8031.0
FULL_TIME	USD	BASE_SALARY	67600.0	7608.0	34003.0

- **Urgensi topik/kasus yang dipilih.**

  Dalam era digital saat ini, jumlah pencari kerja semakin meningkat, sementara persaingan dalam dunia kerja semakin ketat. Platform profesional seperti LinkedIn menyediakan banyak informasi terkait lowongan pekerjaan dan profil pencari kerja, namun tantangan utama adalah bagaimana menghubungkan pencari kerja dengan peluang yang paling sesuai berdasarkan keterampilan dan pengalaman mereka.

  Pendekatan tradisional dalam sistem rekomendasi pekerjaan seringkali berbasis text-matching atau filtering berbasis aturan, yang kurang fleksibel dalam menangkap hubungan kompleks antara keterampilan, pengalaman, dan persyaratan pekerjaan. Oleh karena itu, pendekatan Graph-based Job Recommendation System dapat menjadi solusi yang lebih efektif dengan merepresentasikan pencari kerja, pekerjaan, dan keterampilan dalam bentuk graph, memungkinkan pemodelan hubungan yang lebih kompleks dan akurat.

- **Data yang digunakan.**

  Penelitian ini menggunakan dua dataset dari Kaggle, yaitu:

  1. LinkedIn Professional Profiles Dataset

     Dataset ini berisi informasi tentang berbagai profil profesional di LinkedIn, termasuk informasi pribadi, industri, ukuran perusahaan, spesialisasi, dan lokasi. Dataset ini digunakan untuk merepresentasikan pencari kerja dalam Graph-based Recommendation System, dengan keterampilan dan pengalaman mereka sebagai node dan edge dalam graph.
  2. LinkedIn Job Postings Dataset (2023-2024)

     Dataset ini berisi daftar pekerjaan yang diposting di LinkedIn beserta detailnya, seperti deskripsi pekerjaan, keterampilan yang dibutuhkan, gaji, dan lokasi. Dataset ini digunakan untuk merepresentasikan pekerjaan sebagai node dalam graph, serta membangun koneksi dengan pencari kerja berdasarkan keterampilan dan pengalaman yang relevan.

  Deskripsi Atribut dalam Dataset

  1. LinkedIn Professional Profiles Dataset

     Dataset ini memiliki 26 atribut, namun hanya beberapa yang relevan dalam penelitian ini. Atribut yang digunakan adalah:

     - timestamp: Tanggal dan waktu saat data dikumpulkan
     - id: Identifikasi unik pengguna, digunakan sebagai node "User" dalam graph.
     - name: Nama pengguna (opsional, hanya untuk identifikasi manusiawi, tidak digunakan dalam proses pemodelan).
     - about: Deskripsi pribadi dari pengguna.
     - city: Lokasi pengguna, penting untuk mencocokkan dengan lokasi pekerjaan.
     - country_code: Kode negara pengguna, sebagai informasi tambahan.
     - region: Wilayah pengguna secara lebih spesifik dibanding country_code
     - position: Posisi/jabatan saat ini yang dipegang pengguna, memberikan indikasi pengalaman kerja.
     - current_company:company_id: ID unik perusahaan saat ini (berbeda dari nama)
     - current_company:name: Nama perusahaan tempat pengguna bekerja (alternatif dari kolom current_company).

- current_company: Nama perusahaan tempat pengguna bekerja saat ini.
- experience: Rangkuman pengalaman kerja pengguna (dalam bentuk string JSON-like), yang dapat diekstrak menjadi node pengalaman atau keterampilan.
- following: Jumlah akun atau entitas yang diikuti oleh pengguna
- posts: Konten atau aktivitas post pengguna di LinkedIn
- groups: Kelompok atau komunitas yang diikuti pengguna
- url: Tautan profil LinkedIn pengguna
- people_also_viewed: Daftar pengguna lain yang juga dilihat oleh orang yang melihat profil ini.
- avatar: Tautan ke gambar profil pengguna.
- certifications: Daftar sertifikasi yang dimiliki pengguna
- recommendations: Isi rekomendasi dari atau untuk pengguna.
- recommendations_count: Jumlah rekomendasi yang diterima pengguna.
- education: Rangkuman pendidikan pengguna
- educations_details: Informasi detail pendidikan, berpotensi digunakan untuk profiling pengguna.
- volunteer_experience: Pengalaman relawan dari pengguna.
- courses: Daftar kursus yang pernah diikuti pengguna.
- languages: Bahasa yang dikuasai pengguna, bisa digunakan sebagai fitur tambahan jika dibutuhkan.

2. LinkedIn Job Postings Dataset

   Dataset ini memiliki 31 atribut, namun hanya beberapa yang digunakan dalam penelitian ini.

- job_id: Identifikasi unik pekerjaan.
- company_id: Identifikasi perusahaan yang menawarkan pekerjaan.
- title: Nama posisi pekerjaan.
- description: Deskripsi pekerjaan.
- location: Lokasi pekerjaan.
- job_posting_url: URL ke postingan pekerjaan.
- company_name: Nama perusahaan yang menawarkan pekerjaan.
- max_salary: Gaji maksimum yang ditawarkan untuk posisi tersebut.
- pay_period: Periode pembayaran gaji (misalnya, per jam, per bulan, atau per tahun).

- views: Jumlah tampilan (views) yang diterima oleh postingan pekerjaan.
- med_salary: Gaji median untuk posisi tersebut.
- min_salary : Gaji minimum yang ditawarkan untuk posisi tersebut.
- formatted_work_type : Jenis pekerjaan (misalnya, full-time, part-time, kontrak).
- applies : Jumlah aplikasi yang telah diajukan untuk pekerjaan tersebut
- original_listed_time : Waktu pertama kali pekerjaan diposting.
- remote_allowed : Indikator apakah pekerjaan memungkinkan kerja jarak jauh (remote)
- application_url : URL untuk mengajukan aplikasi pekerjaan.
- application_type : Jenis proses aplikasi (misalnya, aplikasi offsite, onsite yang lebih kompleks atau sederhana).
- expiry : Waktu kedaluwarsa postingan pekerjaan.
- closed_time: Waktu ketika pekerjaan dihapus atau ditutup di platform.
- formatted_experience_level : Tingkat pengalaman yang dibutuhkan untuk pekerjaan (misalnya, entry-level, associate, executive).
- listed_time : Waktu ketika pekerjaan pertama kali terdaftar
- posting_domain : Domain dari situs web tempat pekerjaan diposting.
- sponsored : Menandakan apakah pekerjaan tersebut adalah iklan berbayar atau dipromosikan
- work_type: Jenis pekerjaan yang ditawarkan (misalnya, full-time, part-time, kontrak).
- currency: Mata uang yang digunakan untuk gaji pekerjaan.
- compensation_type: Jenis kompensasi untuk pekerjaan (misalnya, gaji tetap, bonus).
- normalized_salary: Gaji yang telah dinormalisasi berdasarkan standar tertentu.
- zip_code: Kode pos lokasi pekerjaan.
- fips: FIPS (Federal Information Processing Standards) code, yang digunakan untuk menggambarkan lokasi geografis berdasarkan kode tertentu.

- Berdasarkan dataset yang digunakan, berikut beberapa task data mining yang akan digunakan:
    - Graph-based Recommendation (Link **Prediction**):

      Tujuan: Memprediksi hubungan antara pencari kerja dan pekerjaan berdasarkan keterampilan yang sesuai.

      Metode: Menggunakan Link Prediction dalam Graph Neural Networks (GNN) untuk merekomendasikan pekerjaan berdasarkan hubungan eksplisit dan implisit dalam graph.

- **Sumber Dataset**
    - **LinkedIn Professional Profiles Dataset**
    - **LinkedIn Job Postings (2023 - 2024)**

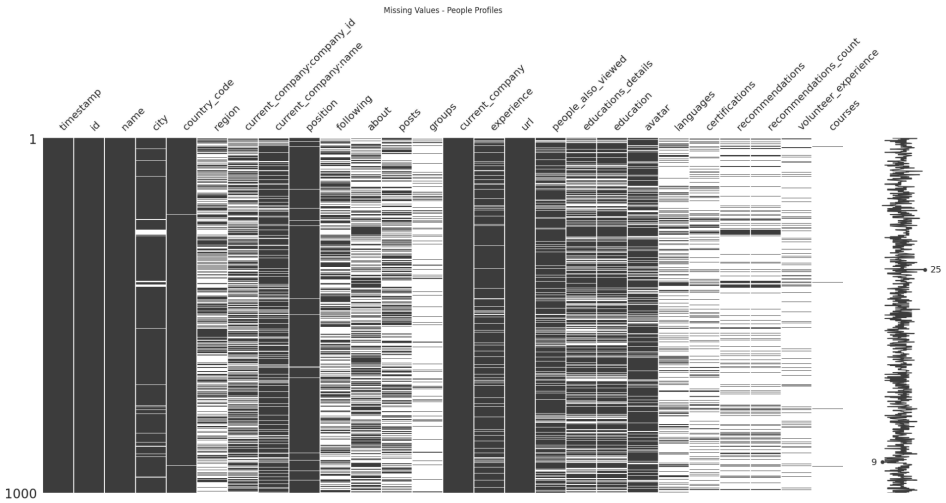
- **EDA (Exploratory Data Analysis)**

  Setelah dilakukan seleksi atribut dari kedua dataset, langkah selanjutnya adalah eksplorasi data untuk memahami struktur, distribusi, dan hubungan antar entitas yang terdapat dalam data. EDA dilakukan baik pada dataset LinkedIn Professional Profiles maupun LinkedIn Job Postings untuk mengidentifikasi pola awal serta mendukung proses pemodelan graph secara efektif.

  1. **Analisis Dataset LinkedIn Professional Profiles**

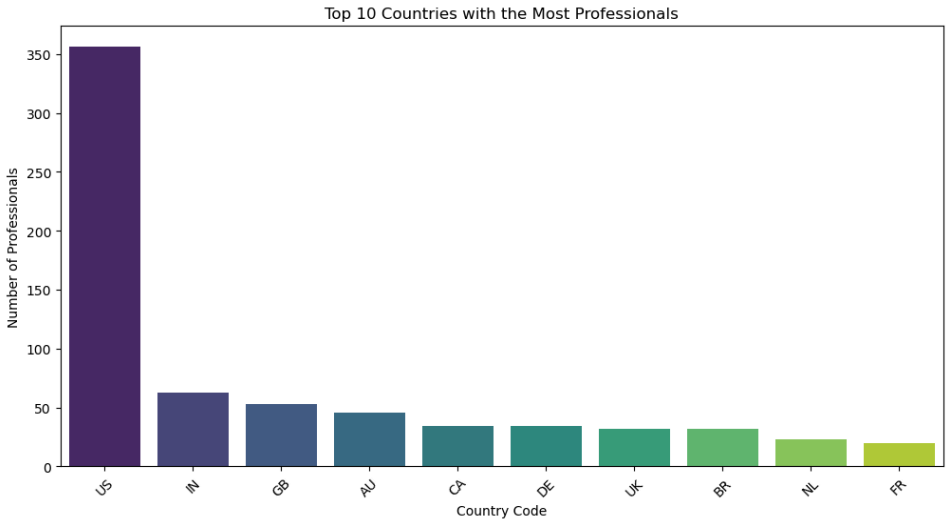
     Beberapa fokus utama dalam eksplorasi dataset pengguna meliputi:

     - Missing Value



Gambar 1.1.1 visualisasi nilai yang hilang pada dataset

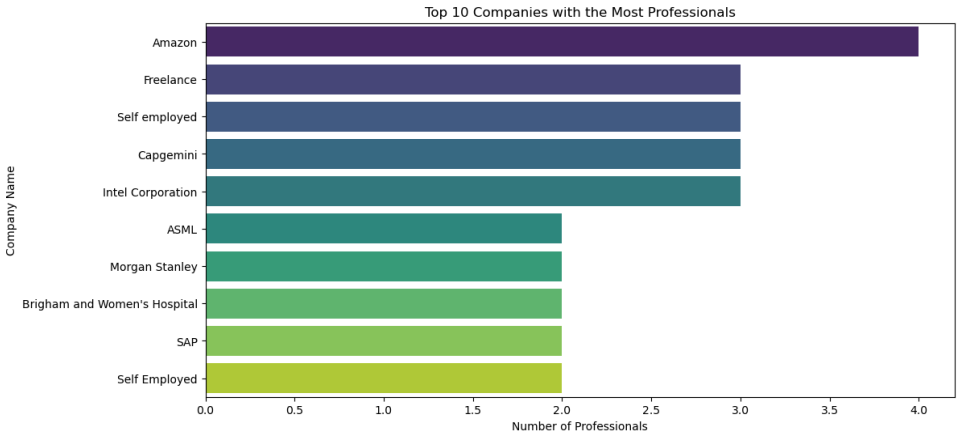
- Distribusi Lokasi dan Negara



Gambar 1.1.2 diagram batang jumlah pengguna di 10 negara teratas

Grafik di atas menunjukkan bahwa Amerika Serikat (US) memiliki jumlah profesional yang jauh lebih banyak dibandingkan negara lainnya, dengan lebih dari 350 professional. Negara-negara seperti India (IN), Inggris (GB), Australia (AU), dan Kanada (CA) berada jauh di bawahnya, masing-masing dengan jumlah yang lebih kecil, kurang dari 100 profesional. Secara keseluruhan, grafik ini menggambarkan konsentrasi profesional yang sangat tinggi di AS, sementara negara-negara lain memiliki distribusi yang lebih merata namun dengan jumlah yang jauh lebih rendah.

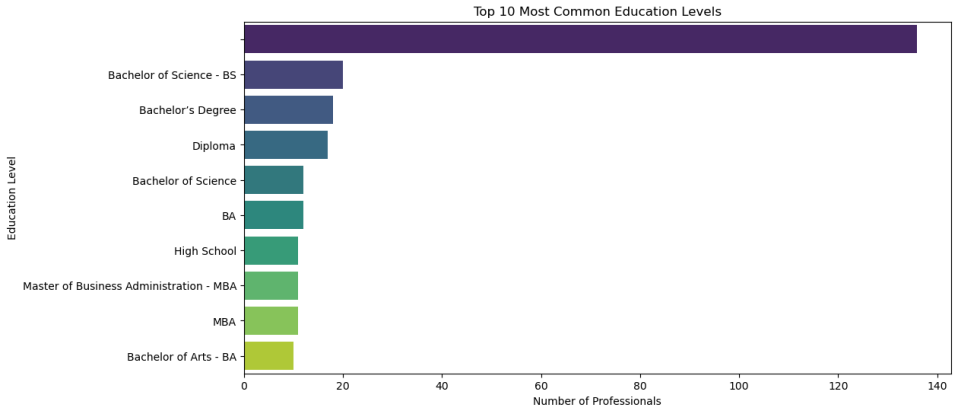
- 10 Perusahaan Teratas dengan Tenaga Profesional Terbanyak



Gambar 1.1.3 diagram batang perusahaan teratas dengan tenaga profesional terbanyak

Grafik di atas menunjukkan perusahaan-perusahaan dengan jumlah profesional terbanyak. Amazon menempati posisi teratas dengan jumlah yang sangat signifikan, jauh lebih banyak dibandingkan perusahaan lain. Di bawahnya, Freelance dan Self Employed menempati posisi kedua dan ketiga, masing-masing dengan jumlah profesional yang hampir sama. Beberapa perusahaan besar lainnya seperti Capgemini, Intel Corporation, dan ASML juga berada dalam daftar, meskipun dengan jumlah yang jauh lebih sedikit. Grafik ini menggambarkan dominasi Amazon dalam hal jumlah profesional, sementara perusahaan-perusahaan lainnya memiliki distribusi yang lebih merata dan lebih kecil.

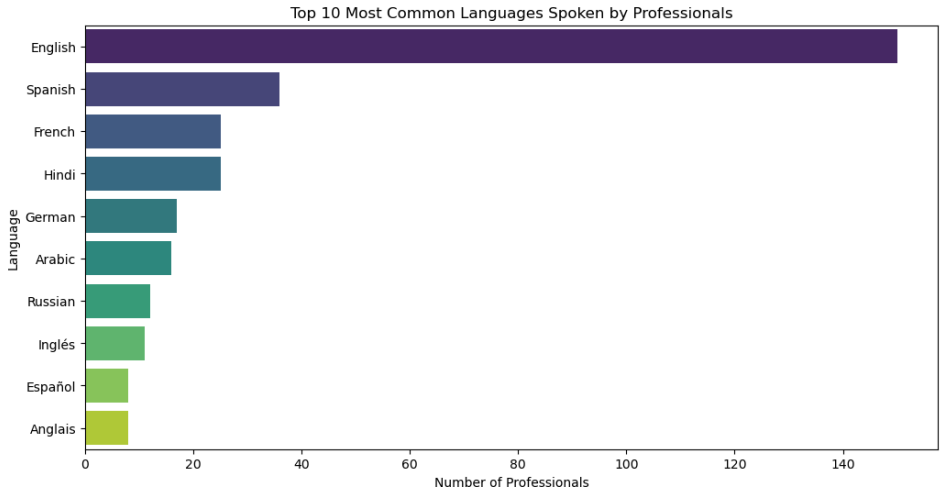
- 10 Tingkat Pendidikan Paling Umum



Gambar 1.1.4 diagram batang 10 tingkat pendidikan teratas

Grafik di atas menunjukkan tingkat pendidikan yang paling umum di antara profesional dalam data tersebut. Degree tanpa pendidikan tertentu sangat dominan, mencakup sebagian besar profesional. Selain itu, gelar Bachelor of Science (BS) juga cukup sering ditemukan, diikuti dengan Bachelor's Degree dan Diploma. Gelar pendidikan lainnya yang tercatat termasuk Master of Business Administration (MBA), Registered Nurse, serta High School. Grafik ini menggambarkan bahwa gelar sarjana (terutama Bachelor of Science) adalah yang paling umum, diikuti oleh berbagai tingkat pendidikan lainnya dengan jumlah yang jauh lebih sedikit.

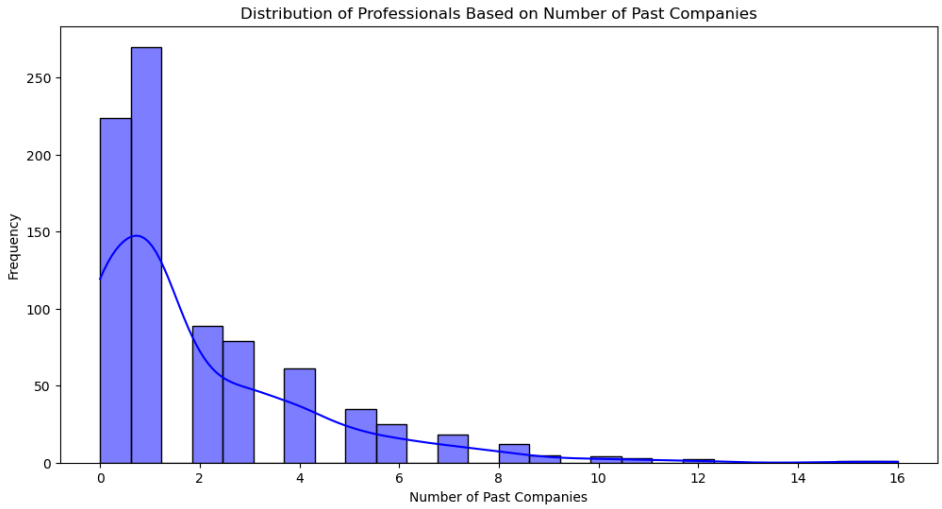
- 10 Bahasa Paling Umum yang Digunakan oleh Para Profesional



Gambar 1.1.5 diagram batang bahasa teratas yang digunakan pengguna.

Grafik di atas menunjukkan bahasa yang paling sering digunakan oleh para profesional. English mendominasi, diikuti oleh Spanish dan French dalam urutan berikutnya. Bahasa-bahasa lain seperti Hindi, German, dan Arabic juga cukup sering digunakan, namun jumlahnya lebih sedikit dibandingkan dengan tiga bahasa teratas.

- Distribusi Profesional Berdasarkan Jumlah Perusahaan Sebelumnya



Gambar 1.1.6 distribusi pengguna berdasarkan jumlah perusahaan sebelumnya

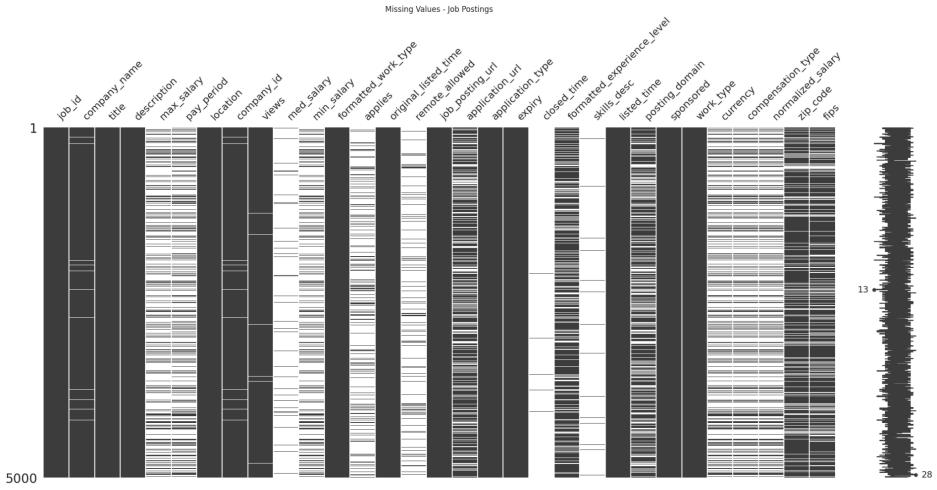
Grafik di atas menggambarkan distribusi jumlah perusahaan tempat para profesional bekerja sebelumnya. Sebagian besar profesional bekerja di 1 atau 2 perusahaan sebelumnya, yang ditunjukkan oleh frekuensi tinggi di bagian kiri grafik. Semakin besar jumlah perusahaan

sebelumnya, semakin menurun frekuensinya, dengan beberapa profesional yang memiliki pengalaman bekerja di lebih dari 10 perusahaan. Secara keseluruhan, distribusi ini menunjukkan bahwa mayoritas profesional memiliki pengalaman terbatas di beberapa perusahaan, sementara hanya sedikit yang memiliki pengalaman bekerja di banyak perusahaan.

2. Analisis Dataset LinkedIn Job Postings

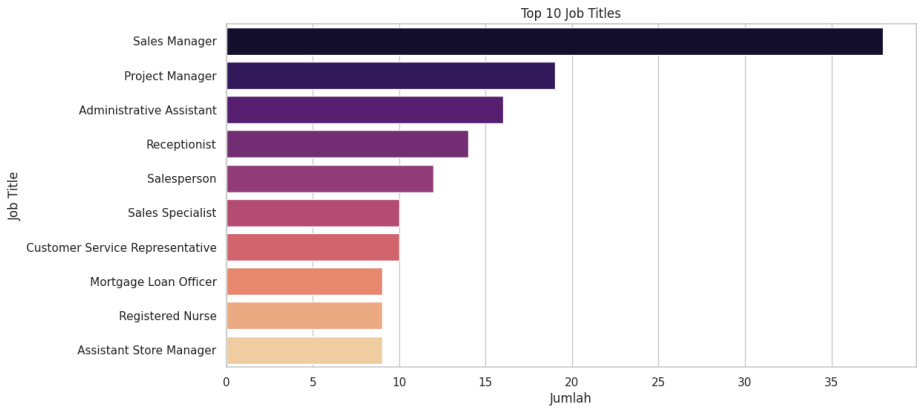
Eksplorasi pada data lowongan pekerjaan dilakukan untuk mendapatkan wawasan terkait pasar kerja dan kebutuhan industri, seperti:

- Missing Value



Gambar 1.2.1 visualisasi nilai yang hilang pada dataset

- 10 Jabatan Teratas

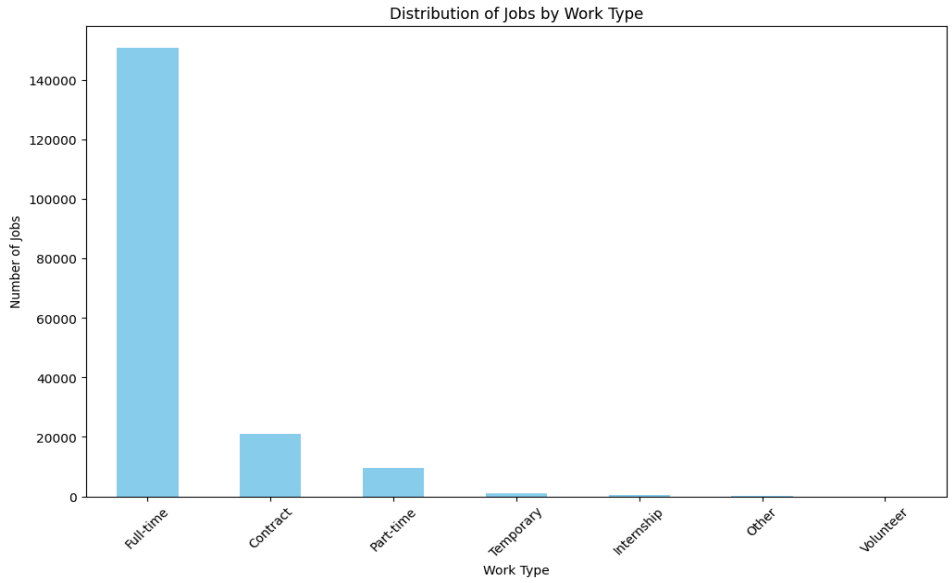


Gambar 1.2.2 diagram batang jabatan teratas berdasarkan jumlahnya

Grafik di atas menunjukkan sepuluh jabatan yang paling banyak ditemukan dalam data. Sales Manager menempati posisi teratas dengan

jumlah yang jauh lebih tinggi dibandingkan jabatan lainnya. Project Manager dan Administrative Assistant berada di urutan berikutnya dengan jumlah yang cukup besar, namun tetap lebih sedikit dari Sales Manager. Jabatan-jabatan lain seperti Receptionist, Salesperson, dan Sales Specialist juga masuk dalam daftar sepuluh besar, tetapi dengan jumlah yang relatif lebih kecil. Secara keseluruhan, grafik ini menunjukkan bahwa jabatan yang terkait dengan manajemen penjualan dan proyek lebih dominan, sedangkan posisi lain seperti Customer Service Representative, Mortgage Loan Officer, dan Registered Nurse memiliki jumlah yang lebih terbatas.

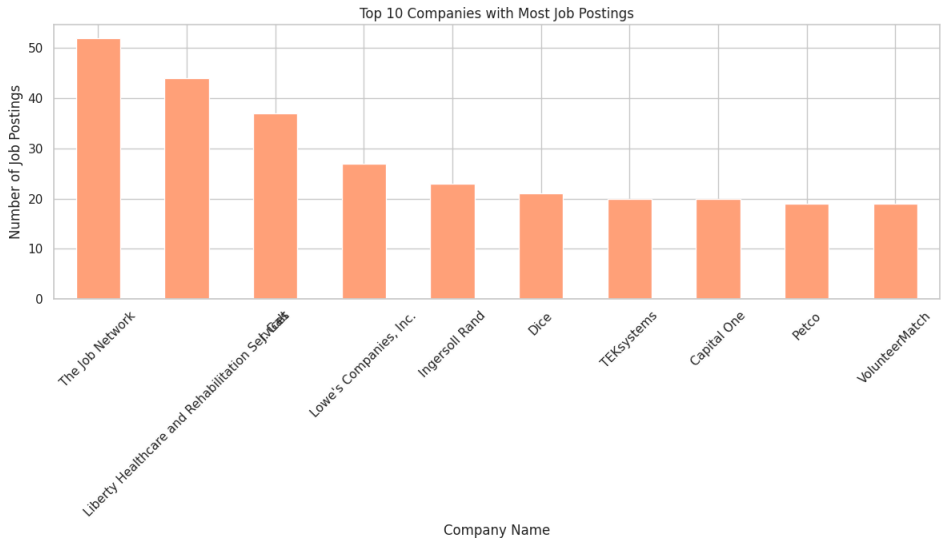
- Distribusi pekerjaan berdasarkan jenis pekerjaan



Gambar 1.2.3 diagram batang distribusi pekerjaan berdasarkan jenis pekerjaan

Posisi penuh waktu mendominasi iklan pekerjaan, membentuk sebagian besar dari total posisi yang tersedia. Posisi paruh waktu dan sementara jauh lebih jarang dibandingkan. Posisi kontrak, magang, dan jenis posisi lainnya bahkan lebih sedikit frekuensinya.

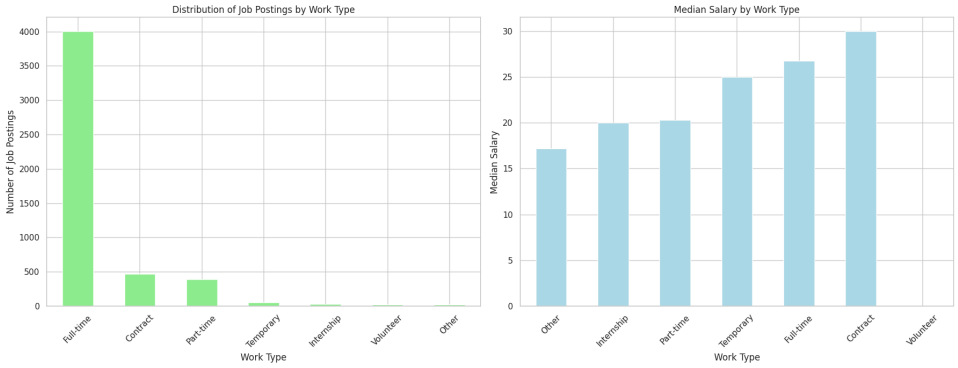
- Perusahaan teratas dengan lowongan pekerjaan terbanyak



Gambar 1.2.4 diagram batang perusahaan dengan lowongan pekerjaan terbanyak

Grafik di atas menunjukkan perusahaan-perusahaan dengan jumlah iklan pekerjaan terbanyak. The Job Network memimpin dengan jumlah iklan pekerjaan yang jauh lebih tinggi dibandingkan perusahaan lainnya, diikuti oleh Liberty Healthcare and Rehabilitation Services dan Lowe's Companies, Inc. yang memiliki jumlah iklan yang cukup besar. Perusahaan lainnya yang menonjol dalam daftar 10 besar adalah Ingersoll Rand, Dice, dan TEKsystems. Grafik ini memberikan gambaran tentang perusahaan-perusahaan aktif yang sedang merekrut di berbagai bidang.

- Distribusi Lowongan Pekerjaan Berdasarkan Jenis Pekerjaan & Gaji Rata-rata Berdasarkan Jenis Pekerjaan

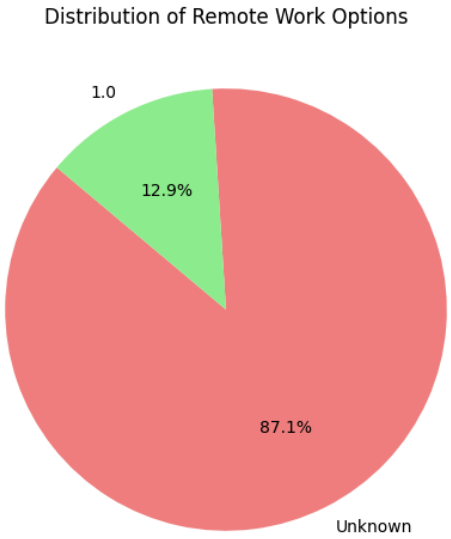


Gambar 1.2.5 diagram batang distribusi lowongan pekerjaan berdasarkan jenis pekerjaan & gaji rata-rata berdasarkan jenis pekerjaan

Posisi penuh waktu mendominasi dataset, yang menunjukkan bahwa sebagian besar iklan pekerjaan di LinkedIn adalah untuk posisi penuh waktu. Posisi kontrak, paruh waktu, dan sementara lebih jarang ditemukan.

Posisi kontrak memiliki gaji median tertinggi, yang diharapkan karena posisi penuh waktu sering kali datang dengan tanggung jawab dan manfaat yang lebih banyak. Posisi penuh waktu mengikuti dengan gaji yang cukup tinggi, menunjukkan bahwa beberapa posisi kontrak dapat menawarkan gaji yang kompetitif. Posisi paruh waktu dan sementara memiliki gaji median terendah, mencerminkan jam kerja yang lebih sedikit dan sifat pekerjaan yang sering kali bersifat sementara.

- Distribusi opsi kerja jarak jauh

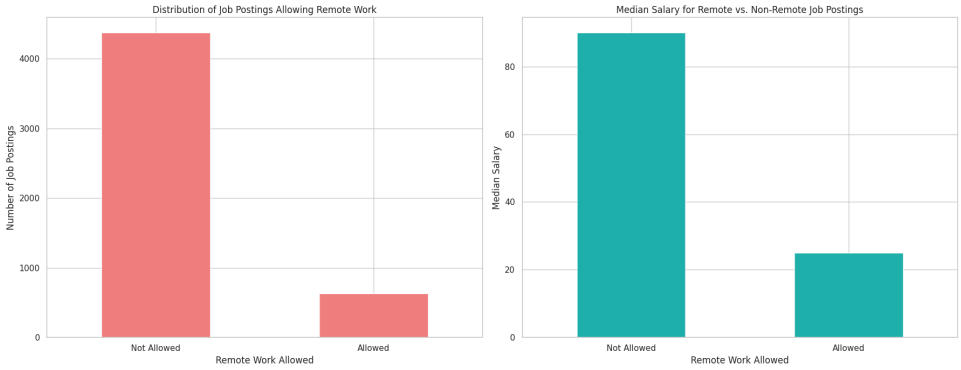


Gambar 1.2.6 plot pie chart distribusi dari pilihan kerja remote

Tidak Diizinkan Bekerja Jarak Jauh: Sebagian besar lowongan pekerjaan (sekitar 80,8%) tidak mengizinkan bekerja jarak jauh. Diizinkan Bekerja Jarak Jauh: Sekitar 13,7% lowongan pekerjaan mengizinkan bekerja jarak jauh, yang menunjukkan fleksibilitas lokasi kerja. Tidak diketahui: Sekitar 5,5% lowongan pekerjaan tidak menyebutkan apakah bekerja jarak jauh diizinkan atau tidak.

Hal ini memberikan wawasan tentang fleksibilitas lokasi kerja di pasar kerja saat ini. Sebagian besar pekerjaan yang tidak mengizinkan bekerja jarak jauh menunjukkan bahwa banyak peran masih memerlukan kehadiran di tempat.

- Distribusi lowongan pekerjaan yang memperbolehkan kerja jarak jauh



Gambar 1.2.7 diagram batang distribusi lowongan pekerjaan yang memperbolehkan secara remote

Distribusi berdasarkan Tunjangan Kerja Jarak Jauh: Mayoritas lowongan kerja tidak memperbolehkan kerja jarak jauh. Namun, sejumlah besar lowongan mengizinkan kerja jarak jauh, yang mencerminkan tren model kerja jarak jauh dan hibrida yang sedang berlangsung.

Gaji rata-rata untuk lowongan kerja yang memperbolehkan kerja jarak jauh sedikit lebih tinggi daripada yang tidak memperbolehkan. Hal ini dapat disebabkan oleh berbagai faktor, termasuk sifat pekerjaan yang menawarkan kerja jarak jauh (misalnya, peran teknologi yang cenderung memiliki gaji lebih tinggi) atau perusahaan yang menawarkan kompensasi lebih tinggi untuk peran jarak jauh guna menarik bakat.

Tahap 2 (poin: 10): Target Data (Optional)

Tidak semua atribut dalam dataset digunakan untuk membangun sistem rekomendasi berbasis graph. Oleh karena itu, dilakukan seleksi atribut yang relevan agar sesuai dengan kebutuhan analisis.

1. Dataset LinkedIn Professional Profiles

Tabel 2.1 dataset informasi profil pengguna platform linkedin setelah dilakukan seleksi fitur

id	about	city	country_code	position	experience	education
irma-arevalo-ccp-phr-shrm-cp-722950a0	CCP, PHR, SHRM-CP	Houston, Texas, United States	US	Director Compensation at Metropolitan Transit Authority of Harris County	[{"company":" Metropolitan Transit Authority of Harris County","company_id":"metropolitan-transit-aut	[{"degree":"Public Administration","end_year":"1989","field":"","meta":"1988 - 1989 Public Administr
anabel-pangalangan-6496095a		Rome, Latium, Italy	IT	Titolare d'Azienda presso GIANBEL ITALIA	[{"company":" GIANBEL ITALIA","company_id":null,"location":"Roma, Italia","positions":[{"description"	[{"degree":"","end_year":"","field":"","meta":"","start_year":"","title":"Tarlac State University","
kevinjcain	Operations Engineering Consultant for Multi-Unit Restaurants and Retail Kevin Cain was part of the t	Miami, Florida, United States	US	Operations Consultant for Multi-Unit Restaurants and Retail Chains	[{"company":" Foodservice Industry","location":"Miami/ Fort Lauderdale Area","positions":[{"descriptio	[{"degree":"B.S.","end_year":"2002","field":"Industrial and Systems Engineering","meta":"1998 - 2002
dana-evans-6054a051		Portland, Oregon,	US	Inventory Manager and	[{"company":" Evans & Associates","l	

		United States		Business Analyst at Nordstrom	ocation":"","positions":[{"description":"","duration":"Jun 2020 -	
marcgrgic	As a fire safety engineer, I try to save peoples lifes with strategic planning and I apply this mind	Innsbruck, Tyrol, Austria	AT	Branch Manager bei IHW Ingenieurbüro Huber GmbH	[{"company":"IHW Ingenieurbüro Huber GmbH","company_id":"ihw","industry":"Architecture and Planning"	[{"degree":"MSc","end_year":"2021","field":"Fire Safety Management","meta":"2019 - 2021","start_year

Atribut yang digunakan:

- id: Identitas unik pengguna, akan digunakan sebagai node utama "User".
- about: deskripsi diri atau ringkasan profesional yang ditulis langsung oleh pengguna
- city: Lokasi pengguna yang akan digunakan untuk mencocokkan preferensi lokasi pekerjaan.
- country_code: Informasi tambahan mengenai lokasi geografis pengguna.
- position: Posisi/jabatan saat ini yang dapat menjadi indikator pengalaman atau spesialisasi pengguna.
- experience: Rangkuman pengalaman kerja dalam format semi-terstruktur, digunakan untuk mengekstraksi informasi keterampilan, jabatan sebelumnya.
- education: Informasi pendidikan yang dapat memperkaya profil pengguna dalam sistem rekomendasi.

Atribut yang diabaikan:

Atribut lainnya yang tersedia dalam dataset seperti url, avatar, groups, posts, recommendations, certifications, dan courses tidak digunakan karena tidak relevan langsung dengan sistem rekomendasi.

2. Dataset LinkedIn Job Posting

Tabel 2.2 dataset list postingan job yang ada dalam platform linkedin sebelum dilakukan seleksi fitur manual

job_id	company_id	title	description	location
3902944011	760913.0	Senior Automation Engineer - Power Systems	The Senior Automation / Power Systems Engineer will primarily be responsible for the conception, des	Houston, TX
3901960222	4296.0	DISH Installation Technician - Field	Company Summary DISH, an EchoStar Company, has been reimagining the future of connectivity for more	Orange, TX
3900944095	136791.0	Order Builder	Division: North Alabama Department : Oxford Warehouse Loading Work Schedule: M-F 8 am UNTIL finish	Oxford, AL
3903878594	11500365.0	Mountain Multimedia Journalist, KMGH	KMGH, the E.W. Scripps Company ABC affiliate in Denver, Colorado is looking for a Multimedia Journal	Denver, CO
3905670593	19472.0	Licensed Practical Nurse (LPN)	Come for the Flexibility, Stay for the Culture Needing more ‘life’ in your work-life balance? Apply	Teterboro, NJ

formatted_experience_level	skills desc	formatted_work_type	job_posting_url
Mid-Senior level		Full-time	https://www.linkedin.com/jobs/view/3902944011/?trk=jobs_biz_prem_srch
		Full-time	https://www.linkedin.com/jobs/view/3901960222/?trk=jobs_biz_prem_srch
Entry level		Full-time	https://www.linkedin.com/jobs/view/3900944095/?trk=jobs_biz_prem_srch
Entry level		Full-time	https://www.linkedin.com/jobs/view/3903878594/?trk=jobs_biz_prem_srch
Entry level		Full-time	https://www.linkedin.com/jobs/view/3905670593/?trk=jobs_biz_prem_srch

Atribut yang digunakan:

- job_id: Sebagai identitas unik lowongan pekerjaan.
- company_id: Menghubungkan pekerjaan dengan perusahaan.
- title: Untuk referensi jenis pekerjaan.
- description: Berisi informasi pekerjaan (opsional).

- location: Menghubungkan pekerjaan dengan lokasi.
- formatted_experience_level: Menentukan tingkat pengalaman yang diperlukan.
- skills_desc: Berisi keterampilan yang dibutuhkan, digunakan untuk mencocokkan dengan keterampilan pengguna.
- work_type: Menentukan apakah pekerjaan full-time, part-time, remote, dll.

Atribut yang diabaikan:

pay_period, salary, application_url, application_type, posting_domain, sponsored, currency, compensation_type, expiry, closed_time → Tidak berkontribusi langsung dalam sistem rekomendasi berbasis graph.

Tahap 3-4 (poin: 25): Data Pre-processing & Transformation

Agar sistem rekomendasi berbasis Graph Neural Network dapat bekerja secara optimal, dilakukan serangkaian proses preprocessing terhadap data yang digunakan. Dataset yang digunakan berasal dari dua sumber berbeda, yaitu data profil pengguna LinkedIn dan data lowongan pekerjaan LinkedIn. Berikut adalah tahapan-tahapan preprocessing yang dilakukan:

1. Data Cleaning & Correction

   Proses pembersihan data bertujuan untuk memastikan bahwa data yang digunakan bebas dari kesalahan, inkonsistensi, serta nilai yang hilang. Langkah-langkah yang dilakukan adalah sebagai berikut:

   - Missing Value Handling: Untuk menghindari error saat pemrosesan teks, nilai kosong (NaN) pada kolom-kolom seperti about, experience, position, dan description diubah menjadi string kosong "".
   - Remove Duplicates Data: Setelah proses ekstraksi skill dari berbagai kolom teks, dilakukan penghapusan duplikasi skill agar hasil lebih bersih dan representatif.
2. Data Integration

   Karena penelitian ini menggunakan dua sumber data yang berbeda, yaitu dataset profil pengguna LinkedIn dan dataset lowongan pekerjaan LinkedIn, perlu dilakukan integrasi data agar kedua dataset dapat digunakan secara bersamaan. Langkah-langkah yang diterapkan adalah:

   - Data merging (konseptual, via similarity & graph building): Data tidak digabung secara langsung, namun dihubungkan berdasarkan tingkat kemiripan antara pengguna dan pekerjaan melalui fitur-fitur seperti skill, lokasi, tingkat pendidikan, dan pengalaman kerja. Hasil integrasi ini membentuk edge dalam struktur graph.
3. Data Transformation

   Data yang telah diintegrasikan selanjutnya melalui proses transformasi untuk memastikan kompatibilitasnya dalam analisis lebih lanjut. Beberapa teknik yang diterapkan adalah:

- Text Lowercasing & Regex Cleaning: Semua teks yang akan diekstraksi skill-nya diubah ke huruf kecil (lowercase) untuk memastikan pencocokan kata kunci lebih konsisten dan mengurangi variasi tak perlu akibat kapitalisasi.
- Text Vectorization (Text → Embedding): Kolom combined_text, yang merupakan gabungan beberapa kolom teks (position, about, experience, dan description), diubah menjadi vektor numerik menggunakan model pre-trained SentenceTransformer. Representasi vektor ini sangat penting sebagai fitur node dalam graph yang akan dibentuk.

4. Data Reduction

Untuk mengurangi kompleksitas data dan meningkatkan efisiensi analisis, dilakukan proses reduksi data dengan metode berikut:

- Penghapusan Atribut yang Tidak Relevan: Untuk mengurangi kompleksitas data dan meningkatkan efisiensi pemrosesan, dilakukan penghapusan atribut-atribut yang tidak relevan. Hanya kolom-kolom penting yang digunakan dalam tahap berikutnya, seperti id, position, about, experience, education, dan hasil transformasi lainnya.

5. Data Discretization

Tidak digunakan dalam penelitian ini. Tidak ada proses konversi atribut numerik menjadi kategori. Semua data numerik digunakan dalam bentuk aslinya atau telah dinormalisasi melalui teknik lain.

6. Data Normalization

Tidak diterapkan secara eksplisit dalam kode. Karena embedding teks yang digunakan telah berada dalam bentuk vektor terstandardisasi, tidak dilakukan normalisasi tambahan terhadap fitur numerik lainnya.

7. Feature Selection/Feature Engineering

Agar analisis lebih fokus dan efisien, dilakukan pemilihan fitur yang paling relevan dengan rekomendasi pekerjaan. Beberapa metode yang digunakan adalah:

- Feature Extraction, Beberapa informasi penting diekstrak dari data mentah:
  - Skill: Diekstraksi dari kolom about, position, experience, dan description menggunakan daftar keyword teknis yang telah disiapkan.
  - Tingkat Pendidikan: Diambil dari kolom education, lalu diklasifikasikan menjadi kategori seperti high_school, bachelors, masters, dan phd.
  - Pengalaman Kerja: Diperkirakan dari isi kolom experience, baik melalui jumlah entri maupun pencarian pola jumlah tahun dalam teks.
- Feature Construction: Fitur baru dibentuk dengan menggabungkan beberapa kolom:
  - combined_text digunakan untuk keperluan embedding.
  - location dibentuk dari gabungan kolom city dan country_code untuk mencocokkan lokasi pengguna dan lokasi pekerjaan.

Tahap 5 (poin: 25): Data Mining

- **Algoritma data mining yang digunakan.**

  Dalam proyek ini, dua tugas utama data mining diselesaikan menggunakan pendekatan berbasis graf. Untuk pengelompokan (clustering) lowongan pekerjaan, pertama-tama Cosine Similarity digunakan untuk mengukur kemiripan antar pekerjaan berdasarkan vektor keterampilan mereka, yang kemudian membentuk sebuah graf pekerjaan. Selanjutnya, model Graph Convolutional Network (GCN), sebuah jenis Graph Neural Network (GNN), diterapkan pada graf ini untuk mempelajari representasi vektor (embedding) dari setiap pekerjaan. Akhirnya, algoritma K-Means digunakan untuk mengelompokkan pekerjaan berdasarkan kedekatan embedding yang dihasilkan oleh GCN.

  Untuk tugas rekomendasi (link prediction) antara profil pencari kerja dan pekerjaan, sebuah graf bipartite dibangun. Model Graph Sample and Aggregate (GraphSAGE), jenis GNN lain, digunakan untuk mempelajari embedding dari node profil dan pekerjaan dalam graf bipartite ini. Kemungkinan adanya hubungan atau kecocokan antara profil dan pekerjaan diprediksi menggunakan Inner Product Decoder pada embedding yang dihasilkan, dan model dilatih menggunakan Binary Cross-Entropy Loss untuk membedakan antara hubungan yang relevan dan tidak relevan.

  - Graph-based Clustering:
    - Cosine Similarity: Digunakan untuk mengukur kemiripan antar pekerjaan berdasarkan vektor keterampilan mereka.
    - Model GNN/GAT: Graph Convolutional Network (GNN) atau Graph Attention Network (GAT) diterapkan pada graf pekerjaan untuk mempelajari representasi vektor (embedding) dari setiap pekerjaan.
    - K-Means: Algoritma ini digunakan untuk mengelompokkan pekerjaan berdasarkan kedekatan embedding yang dihasilkan oleh model GNN/GAT.
  - Graph-based Recommendation (Link Prediction):
    - Model GNN/GAT: GraphSAGE (sebagai implementasi GNN) atau Graph Attention Network (GAT) digunakan untuk mempelajari embedding dari node profil pencari kerja dan node pekerjaan dalam graf bipartite.
    - Inner Product Decoder: Digunakan untuk memprediksi kemungkinan adanya hubungan atau kecocokan antara profil dan pekerjaan.
    - Binary Cross-Entropy Loss: Digunakan sebagai fungsi kerugian untuk melatih model.
- **Langkah Kerja Eksperimen**
  - Graph-based Clustering:
    - Ekstraksi Keterampilan: Keterampilan diekstraksi dari deskripsi pekerjaan yang telah dibersihkan.
    - Pembuatan Graf Kemiripan Pekerjaan: Matriks biner pekerjaan vs. keterampilan dibuat. Kemiripan antar pekerjaan dihitung menggunakan

cosine similarity. Graf pekerjaan (G_jobs) dibuat dimana node adalah pekerjaan dan edge menunjukkan kemiripan di atas ambang batas tertentu.
        - Implementasi GNN untuk Clustering: Graf dikonversi ke format PyTorch Geometric. Model GNN (misalnya GNNCluster dengan lapisan GCNConv) digunakan untuk menghasilkan embedding pekerjaan. Model dilatih untuk memaksimalkan kemiripan embedding node yang terhubung dan meminimalkan kemiripan node yang tidak terhubung.
        - Clustering dengan KMeans: Embedding pekerjaan dari GNN menjadi input untuk KMeans. Jumlah cluster ditentukan (misalnya dengan metode elbow).
    - Graph-based Recommendation (Link Prediction):
        - Pembuatan Graf Bipartite: Graf (G_bipartite) dibuat dengan node profil dan node pekerjaan. Edge ditambahkan jika ada tumpang tindih keterampilan di atas ambang batas.
        - Pembagian Data: Edge dibagi menjadi set pelatihan dan pengujian.
        - Implementasi GNN untuk Link Prediction: Graf pelatihan dikonversi ke format PyTorch Geometric. Fitur node dibuat (misalnya one-hot encoding tipe node). Model GNN (misalnya GNNLinkPredictor dengan lapisan SAGEConv) digunakan untuk menghasilkan embedding profil dan pekerjaan. Inner product decoder digunakan untuk prediksi link. Model dilatih dengan Binary Cross Entropy Loss.
        - Evaluasi: Model dievaluasi pada set pengujian menggunakan metrik seperti AUC (Area Under the ROC Curve) dan AP (Average Precision).
        - Generasi Rekomendasi: Fungsi dibuat untuk menghasilkan rekomendasi pekerjaan teratas untuk profil tertentu berdasarkan skor kecocokan dari model.

- **Skenario Eksperimen**

  Masalah utama yang ingin dipecahkan adalah bagaimana mencocokkan profil pencari kerja dengan lowongan pekerjaan yang relevan secara efektif. Pendekatan yang diambil adalah dengan memanfaatkan teknik Graph Neural Network (GNN) untuk mempelajari representasi (embeddings) dari setiap lowongan pekerjaan berdasarkan keterampilannya. Embeddings ini kemudian digunakan untuk melakukan clustering pada lowongan pekerjaan.

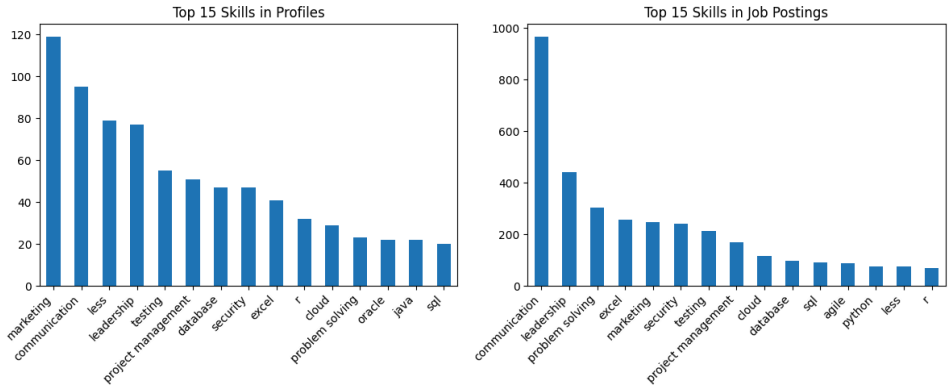
  Dilakukan total 6 eksperimen utama untuk mengevaluasi performa model clustering dengan kombinasi model GNN (GCN dan GAT) dan berbagai teknik seleksi fitur. Eksperimen tersebut adalah:

  - Model GNN (GCN) tanpa seleksi fitur.
  - Model GAT tanpa seleksi fitur.
  - Model GNN (GCN) dengan seleksi fitur Chi-Square.

- Model GAT dengan seleksi fitur Chi-Square.
- Model GNN (GCN) dengan Sequential Feature Selection (SFS).
- Model GAT dengan Sequential Feature Selection (SFS).

Setiap eksperimen mengikuti langkah-langkah umum sebagai berikut:

- Preprocessing:
  - Ekstraksi Keterampilan: Keterampilan diekstrak dari deskripsi pekerjaan. Sebuah daftar keterampilan umum (common skills) telah didefinisikan untuk membantu proses ini.



Gambar 5.1 Top 15 Skills in Profiles & Top 15 in Job Postings

  - Pembersihan Teks: Deskripsi pekerjaan dibersihkan dengan mengubah teks menjadi huruf kecil, menghapus karakter spesial, melakukan tokenisasi, dan menghilangkan stopwords.
  - Pembuatan Fitur TF-IDF: Fitur numerik dibuat dari teks keterampilan yang telah dibersihkan menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer. Dalam eksperimen, digunakan max_features=50.
- Persiapan Data: Dataset lowongan pekerjaan diambil sampelnya (head(2000)). Lowongan pekerjaan yang tidak memiliki keterampilan hasil ekstraksi akan difilter, namun jika jumlahnya sangat sedikit, semua sampel pekerjaan akan digunakan dengan memberikan keterampilan dummy ['general']. Untuk eksperimen ini, digunakan 1605 lowongan pekerjaan yang memiliki keterampilan.
- Konstruksi Graf:
  - Matriks similaritas kosinus dihitung antar lowongan pekerjaan berdasarkan fitur TF-IDF.
  - Sebuah graf dibangun dimana node merepresentasikan lowongan pekerjaan dan edge menghubungkan lowongan yang memiliki similaritas di atas ambang batas tertentu (threshold = 0.1). Jika tidak ada edge yang terbentuk, setiap node dihubungkan ke tetangga terdekatnya.

- Pemodelan (GNN/GAT):
  - Model GNN (GCN atau GAT) diinisialisasi.
  - Model dilatih menggunakan fitur node (TF-IDF) dan struktur graf (edge index dan edge weights). Proses pelatihan bertujuan untuk meminimalkan Mean Squared Error (MSE) antara matriks similaritas yang direkonstruksi dari embeddings dengan matriks similaritas target (matriks similaritas kosinus awal). Pelatihan dilakukan selama 50 epoch.
  - Setelah pelatihan, embeddings akhir untuk setiap lowongan pekerjaan diperoleh.
- Clustering:
  - Algoritma K-Means diterapkan pada embeddings yang dihasilkan oleh model GNN/GAT untuk mengelompokkan lowongan pekerjaan ke dalam 5 cluster.
- Evaluasi: Kualitas clustering dievaluasi menggunakan metrik:
  - Silhouette Score
  - Calinski-Harabasz Score
  - Davies-Bouldin Score

- **Hasil dan Analisis Perbandingan**

Dua jenis model Graph Neural Network utama digunakan dalam eksperimen ini: Graph Convolutional Network (GCN) dan Graph Attention Network (GAT).

1. Graph Convolutional Network (GCN)

GCN adalah jenis jaringan saraf yang beroperasi secara langsung pada graf. GCN menghasilkan representasi node dengan mengagregasi informasi dari tetangga-tetangganya. Proses agregasi ini mirip dengan operasi konvolusi pada gambar, tetapi diadaptasi untuk struktur graf. Model GCN yang digunakan memiliki dua lapisan GCNConv dengan fungsi aktivasi ReLU dan lapisan dropout.

GCN dipilih karena kemampuannya yang efektif dalam mempelajari representasi node dalam graf dan relatif sederhana untuk diimplementasikan. Model ini baik untuk menangkap informasi struktural dan fitur dari graf lowongan pekerjaan. Dan berikut adalah hasil evaluasi model:

Tabel 5.1 hasil evaluasi model GCN

Kriteria Evaluasi	**Tanpa Seleksi Fitur**	**Dengan Seleksi Fitur Chi-Square (k=20)**	**Dengan Seleksi Fitur SFS (n_features_to_select=15)**
Silhouette Score	0.437164	0.732855	0.801085
Calinski-Harabasz Score	872.259583	16236.781250	7772.004395

Davies-Bouldin Score	1.056675	0.535081	0.401580
Feature Shape	(1605, 50)	(1605, 20)	(1605, 15)

2. Graph Attention Network (GAT)

GAT juga merupakan jenis GNN, namun GAT menggunakan mekanisme perhatian (attention mechanism) untuk mempelajari bobot kepentingan yang berbeda untuk tetangga yang berbeda saat mengagregasi informasi. Hal ini memungkinkan model untuk secara dinamis fokus pada node tetangga yang paling relevan. Model GAT yang digunakan memiliki dua lapisan GATConv, dimana lapisan pertama menggunakan multi-head attention.

GAT dipilih karena kemampuannya untuk memberikan bobot yang berbeda pada node tetangga, yang berpotensi menghasilkan representasi yang lebih kaya dan akurat, terutama pada graf dimana hubungan antar node tidak seragam.

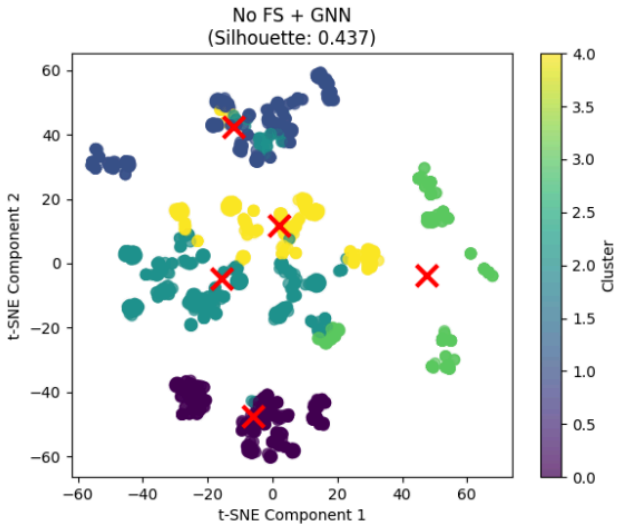
Tabel 5.2 hasil evaluasi model GAT

Kriteria Evaluasi	**Tanpa Seleksi Fitur**	**Dengan Seleksi Fitur Chi-Square (k=20)**	**Dengan Seleksi Fitur SFS (n_features_to_select=15)**
Silhouette Score	0.484394	0.757058	0.873675
Calinski-Harabasz Score	4974.860840	18356.222656	11954.293945
Davies-Bouldin Score	0.755110	0.502403	0.472430
Feature Shape	(1605, 50)	(1605, 20)	(1605, 15)

- **Analisis Visual Kualitas Cluster (t-SNE)**

Visualisasi t-SNE embeddings dari keenam skenario eksperimen memberikan gambaran kualitatif mengenai seberapa baik masing-masing model dan konfigurasi fitur mampu memisahkan lowongan pekerjaan ke dalam cluster-cluster yang berbeda.

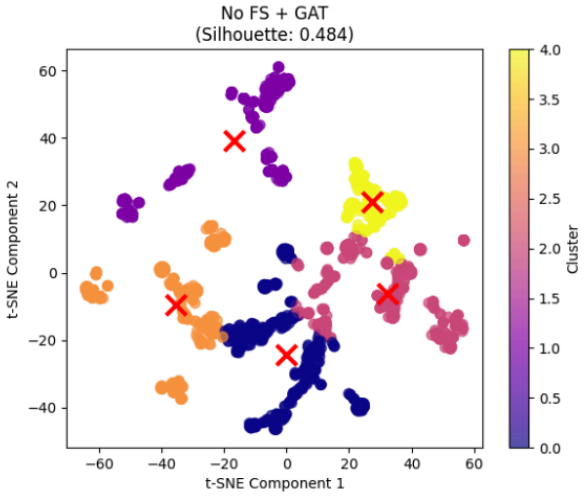
1. No FS + GNN (Silhouette: 0.437)



Gambar 5.2 No FS + GNN

Visualisasi untuk model GNN tanpa seleksi fitur menunjukkan cluster yang cenderung tumpang tindih dan kurang terdefinisi dengan jelas. Titik-titik data dari cluster yang berbeda tampak bercampur, terutama di bagian tengah plot. Hal ini sejalan dengan skor Silhouette yang relatif rendah (0.437), yang mengindikasikan bahwa beberapa sampel mungkin berada di antara cluster atau bahkan salah diklasifikasikan.

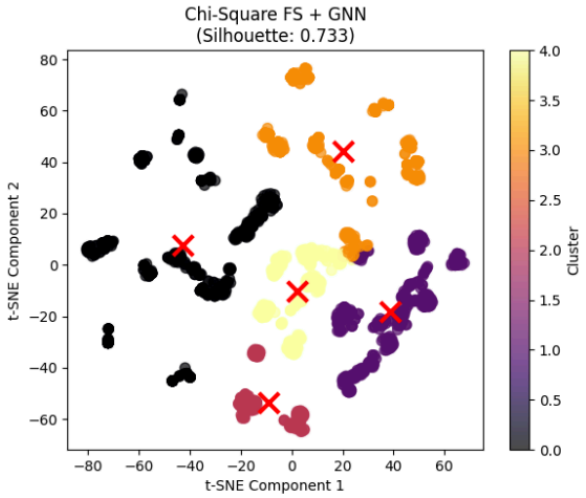
2. No FS + GAT (Silhouette: 0.484)



Gambar 5.3 No FS + GAT

Mirip dengan GNN tanpa seleksi fitur, model GAT tanpa seleksi fitur juga menunjukkan cluster yang belum terpisah dengan baik, meskipun secara visual mungkin tampak sedikit lebih baik daripada GNN. Masih terdapat area tumpang tindih yang signifikan antar cluster. Skor Silhouette (0.484) yang sedikit lebih tinggi dari GNN tanpa FS juga mendukung observasi ini.

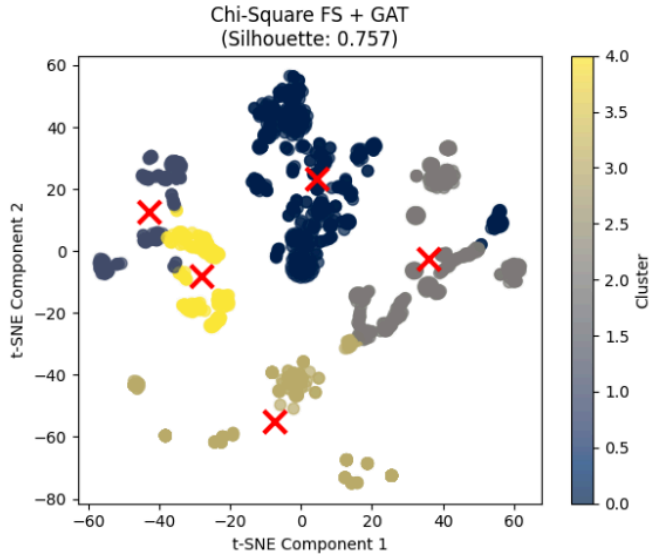
3. Chi-Square FS + GNN (Silhouette: 0.733)



Gambar 5.4 Chi_Square FS + GNN

Dengan penerapan seleksi fitur Chi-Square, kualitas visual cluster pada model GNN meningkat drastis. Cluster-cluster tampak lebih padat dan lebih jelas terpisah satu sama lain, meskipun masih ada beberapa titik yang berada di perbatasan atau sedikit tercampur. Peningkatan ini sangat sesuai dengan lonjakan signifikan pada skor Silhouette menjadi 0.733.

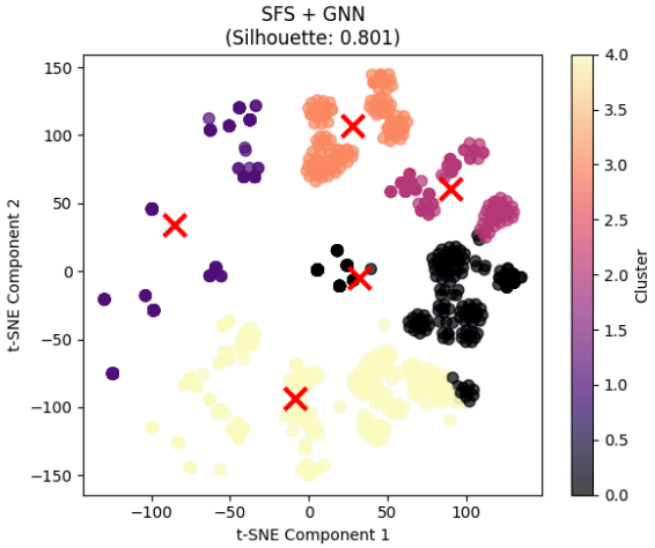
4. Chi-Square FS + GAT (Silhouette: 0.757)



Gambar 5.5 Chi_Square FS + GAT

Model GAT dengan seleksi fitur Chi-Square juga menunjukkan pemisahan cluster yang jauh lebih baik dibandingkan tanpa seleksi fitur. Secara visual, pemisahannya mirip atau sedikit lebih baik dari GNN dengan Chi-Square, dengan cluster yang lebih kompak. Ini didukung oleh skor Silhouette yang juga tinggi (0.757).

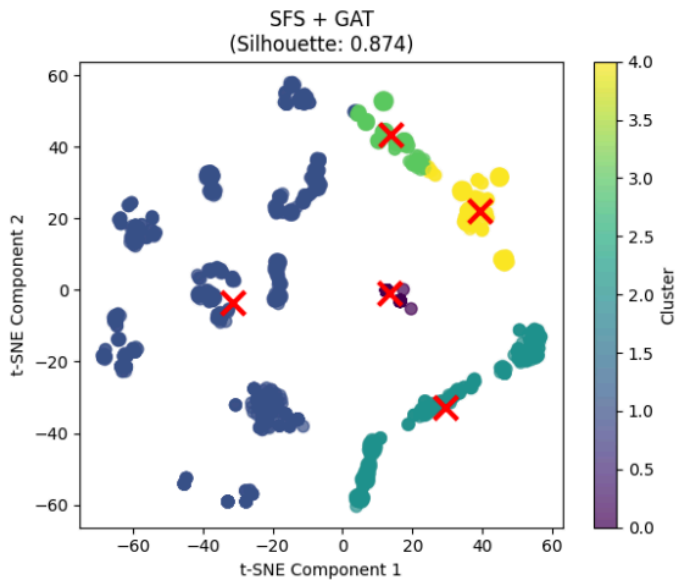
5. SFS + GNN (Silhouette: 0.801)



Gambar 5.6 SFS + GNN

Visualisasi untuk GNN dengan Sequential Feature Selection (SFS) menunjukkan pemisahan cluster yang sangat baik. Cluster-cluster terlihat jelas berbeda, padat, dan memiliki batas yang relatif tegas. Ini mencerminkan skor Silhouette yang tinggi (0.801), yang mengindikasikan cluster yang padat dan terpisah dengan baik.

6. SFS + GAT (Silhouette: 0.874)



Gambar 5.7 SFS + GAT

Ini adalah visualisasi dengan pemisahan cluster terbaik di antara keenam skenario. Cluster-cluster tampak sangat berbeda, masing-masing membentuk kelompok yang jelas dan terisolasi dari yang lain. Hanya sedikit sekali titik yang tampak berada di antara cluster. Kualitas visual ini sangat mendukung skor Silhouette tertinggi yang dicapai (0.874).

Secara umum, visualisasi t-SNE ini secara konsisten mendukung hasil metrik kuantitatif. Semakin tinggi skor Silhouette, semakin baik pula pemisahan visual antar cluster yang terlihat pada plot t-SNE. Penggunaan seleksi fitur (baik Chi-Square maupun SFS) secara signifikan meningkatkan kemampuan model GNN dan GAT dalam membentuk cluster lowongan pekerjaan yang lebih bermakna dan terstruktur. Model SFS + GAT secara visual menunjukkan pengelompokan yang paling ideal.

Tahap 6 (poin: 20): Knowledge Interpretation

- **Perbandingan Umum Antar Skenario dan Kesimpulan Pemilihan Model**

Berdasarkan hasil evaluasi yang telah dijelaskan pada tahap 5, terdapat beberapa informasi penting yang dapat diinterpretasikan. Yang pertama adalah perbandingan asil evaluasi dari semua skenario eksperimen yang akan disajikan dalam tabel perbandingan berikut:

Tabel 6.1 Perbandingan Kinerja Clustering untuk Model GNN dan GAT dengan Metode Seleksi Fitur yang Berbeda.

Metode	N_Clusters	Silhouette _Score	Calinski_ Harabasz	Davies_ Bouldin	Feature _Shape	N_Edges
Tanpa Feature Selection						
GNN	5	0.437164	872.259	1.056675	(1605, 50)	1152352
GAT	5	0.484394	4974.860	0.755110	(1605, 50)	1152352
Chi-Square						
GNN	5	0.732855	16236.781	0.535081	(1605, 20)	1894124
GAT	5	0.757058	18356.222	0.502403	(1605, 20)	1894124
SFS						
GNN	5	0.801085	7772.004	0.401580	(1605, 15)	993108

GAT	5	0.873675	11954.293	0.472430	(1605, 15)	993108

- **Insight dan Interpretasi (Pola-pola useful yang telah ditemukan)**
  - Model dengan Performa Terbaik, Terburuk, dan yang Meningkat Signifikan:

    Berdasarkan evaluasi, model SFS + GAT menunjukkan performa terbaik secara keseluruhan. Hal ini terutama terlihat pada Silhouette Score yang mencapai 0.873675, mengindikasikan kualitas pemisahan cluster yang baik, serta Davies-Bouldin Score yang rendah yaitu 0.472430, yang juga menunjukkan hasil yang baik. Sebaliknya, model No FS + GNN memiliki performa terburuk, dengan Silhouette Score terendah sebesar 0.437164 dan Davies-Bouldin Score tertinggi mencapai 1.056675. Secara umum, semua model menunjukkan peningkatan performa yang signifikan ketika seleksi fitur diterapkan. Peningkatan yang paling dramatis terlihat ketika beralih dari "No FS" ke "Chi-Square FS" atau "SFS" untuk kedua model (GNN dan GAT). Sebagai contoh, Silhouette Score untuk GNN meningkat dari 0.437 menjadi 0.732 dengan seleksi fitur Chi-Square, dan lebih lanjut meningkat menjadi 0.801 dengan seleksi fitur SFS.

  - Pengaruh Feature Selection terhadap Performa Masing-masing Model

    Seleksi fitur menunjukkan peningkatan performa yang signifikan pada model GNN (GCN). Metode SFS memberikan peningkatan yang lebih baik pada Silhouette Score dan Davies-Bouldin Score dibandingkan dengan Chi-Square, meskipun Calinski-Harabasz Score lebih tinggi dengan Chi-Square. Pengurangan dimensi fitur dari 50 menjadi 20 dengan Chi-Square, dan selanjutnya menjadi 15 dengan SFS, tampaknya membantu model GCN untuk lebih fokus pada fitur-fitur yang informatif, mengurangi noise, serta menghasilkan cluster yang lebih kohesif dan terpisah.

    Serupa dengan GNN, model GAT juga mendapatkan manfaat besar dari penerapan seleksi fitur. Secara khusus, SFS menghasilkan Silhouette Score tertinggi, yaitu 0.873675, di antara semua eksperimen yang dilakukan. Hal ini menunjukkan bahwa GAT, meskipun telah memiliki mekanisme perhatian (attention mechanism), tetap mendapatkan keuntungan dari pengurangan dimensi fitur pada tahap awal untuk meningkatkan kualitas hasil clustering.

  - Model Mana yang Cocok Tanpa Feature Selection, Mana yang Lebih Baik Dengan Feature Selection

    Berdasarkan hasil yang diperoleh, tidak ada model yang cocok digunakan tanpa melakukan seleksi fitur. Baik GNN maupun GAT menunjukkan performa

yang jauh lebih rendah ketika seleksi fitur tidak diterapkan. Namun, semua model menunjukkan peningkatan signifikan ketika seleksi fitur diterapkan. Hal ini mengindikasikan bahwa fitur TF-IDF awal (50 fitur) kemungkinan mengandung noise atau fitur yang kurang relevan, yang dapat mengganggu proses pembelajaran embedding dan clustering.

- Insight dari Pola Hasil Klasifikasi (Cluster Characteristics)

  Analisis karakteristik cluster dilakukan menggunakan hasil dari experiment1_results (No FS + GNN), yang bukan merupakan model terbaik. Namun, kita masih bisa mendapatkan gambaran umum:

  - Cluster 0 (Ukuran: 303): Didominasi oleh keterampilan seperti "problem solving", "communication", dan "leadership". Judul pekerjaan bervariasi, namun terdapat beberapa posisi layanan pelanggan dan manajemen proyek.

    Tabel 6.2 Distribusi Keahlian dan Frekuensi Cluster 0

Keahlian	Jumlah
problem solving	303
communication	223
leadership	89
project management	57
excel	55
testing	43
security	42
marketing	36
database	28
critical thinking	28

  - Cluster 1 (Ukuran: 318): Keterampilan utama adalah "leadership" dan "communication", dengan "marketing" juga muncul. Banyak judul pekerjaan terkait manajemen (misalnya, Assistant Store Manager, Production Manager).

Tabel 6.3 Distribusi Keahlian dan Frekuensi Cluster 1

Keahlian	Jumlah
leadership	318
communication	169
marketing	50
security	47
project management	35
excel	31
testing	22
less	17
cloud	14
networking	12

- Cluster 2 (Ukuran: 433): Menunjukkan campuran keterampilan teknis seperti "security", "testing", "cloud", "sql", "python", dan "database" bersama dengan "communication". Judul pekerjaan mencakup Software Developer, System Administrator.

  Tabel 6.4 Distribusi Keahlian dan Frekuensi Cluster 2

Keahlian	Jumlah
security	143
testing	142
communication	136
cloud	73
sql	60
python	53
database	52

project management	51
agile	46
excel	35

- Cluster 3 (Ukuran: 277): "Communication" sangat dominan. Keterampilan teknis lain seperti "networking" dan "windows" juga ada. Judul pekerjaan banyak terkait penjualan (Sales Specialist).

  Tabel 6.5 Distribusi Keahlian dan Frekuensi Cluster 3

Keahlian	Jumlah
communication	277
less	7
networking	6
classification	6
windows	6
ai	5
database	5
full stack	4
swift	4
r	4

- Cluster 4 (Ukuran: 274): "Communication", "marketing", dan "excel" adalah keterampilan utama. Judul pekerjaan seringkali terkait penjualan, pemasaran, dan administrasi (Sales Manager, Marketing Intern, Administrative Assistant).

  Tabel 6.6 Distribusi Keahlian dalam Cluster 4 dan Jumlahnya

Keahlian	Jumlah
communication	162

marketing	155
excel	137
project management	27
classification	25
networking	12
less	11
security	9
windows	8
seo	8

Clustering berhasil mengelompokkan pekerjaan dengan profil keterampilan yang berbeda. Ada cluster yang lebih berfokus pada soft skills dan manajemen (Cluster 0, 1, 4), cluster dengan fokus teknis (Cluster 2), dan cluster dengan fokus penjualan/komunikasi (Cluster 3). Analisis ini akan lebih bermakna jika dilakukan pada hasil model terbaik (SFS + GAT).

- **Analisis Kesalahan (Error Analysis)**

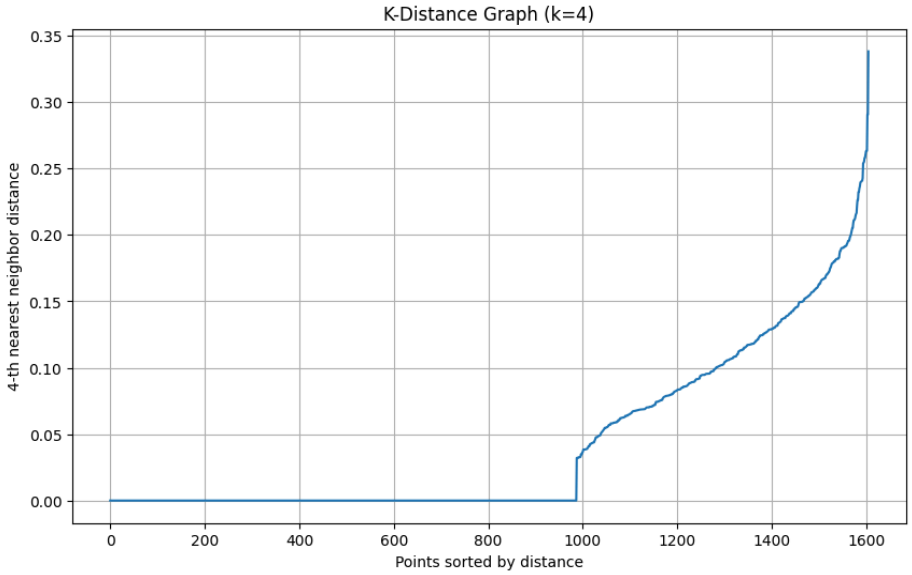
Analisis kesalahan penting untuk memahami mengapa model membuat prediksi yang salah dan bagaimana meningkatkannya.

1. Outlier/Noise dalam Clustering

Analisis kesalahan dalam konteks clustering ini dapat difokuskan pada identifikasi noise points atau outliers yang tidak cocok dengan baik ke dalam cluster mana pun, serta potensi masalah dalam pembentukan cluster. Projek ini menyertakan analisis tambahan menggunakan DBSCAN pada embeddings dari *experiment1_results (No FS + GNN)*.

Deteksi Outlier/Noise dengan DBSCAN:

- Menggunakan k-distance graph (dengan k=4), nilai epsilon optimal untuk DBSCAN disarankan sekitar 0.1754.



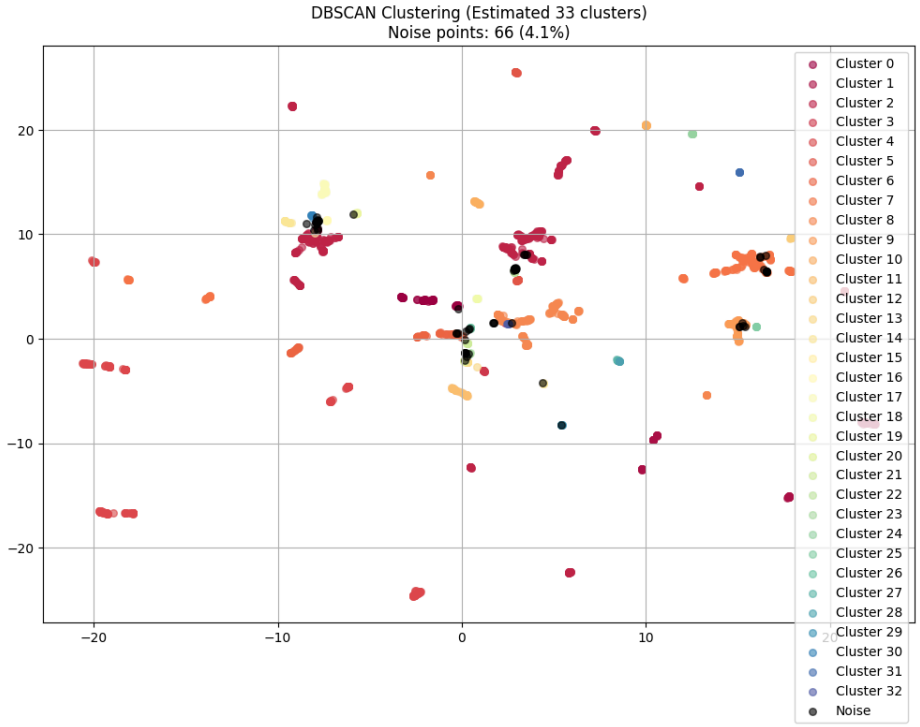
Gambar 6.1 Grafik Jarak K (k=4)

Dengan parameter ini dan min_samples=5, DBSCAN mengidentifikasi 33 cluster dan 66 noise points (sekitar 4.11% dari data).

Tabel 6.7 Hasil Parameter

Number of clusters	33
Number of noise points	66
Percentage of noise	4.11%

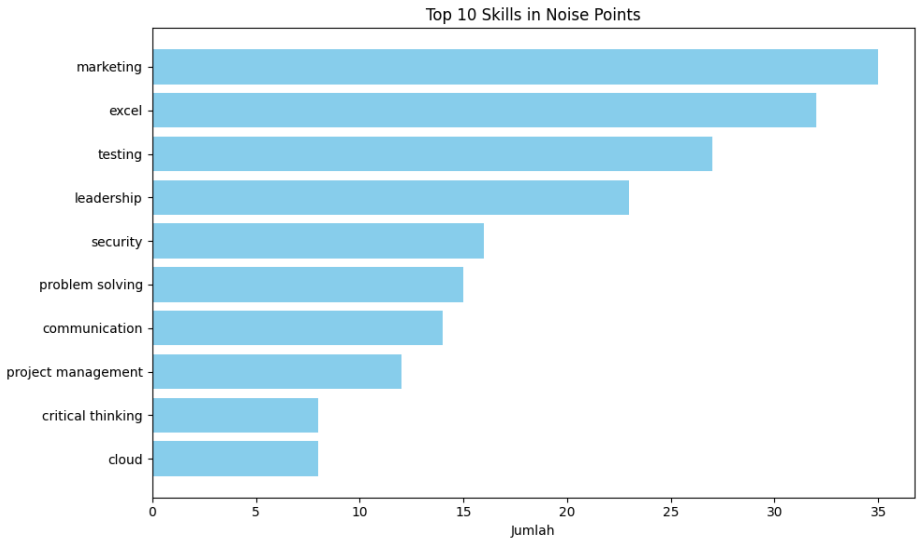
- Visualisasi menggunakan UMAP menunjukkan sebaran cluster dan noise points.



Gambar 6.2 Hasil Clustering DBSCAN

Karakteristik Noise Points:

- Noise points ini awalnya terdistribusi ke berbagai cluster oleh K-Means pada experiment1_results (misalnya, 15 ke Cluster 0, 23 ke Cluster 2, dst.).
- Keterampilan yang dominan pada noise points meliputi "marketing", "excel", "testing", "leadership", dan "security".



Gambar 6.3 Distribusi 10 keterampilan teratas

- Contoh noise jobs termasuk "Clinical Paramedic", "VP, Strategy & Insights", dan "Software Engineering Manager".

Instances yang salah Cluster

1. Job 6: Clinical Paramedic: Peran ini sangat spesifik di bidang medis dengan potensi ekstraksi keterampilan ("less", "testing") yang kurang relevan, membuatnya berbeda dari mayoritas pekerjaan lain dalam dataset.

   Tabel 6.8 Informasi dan Detail Cluster

Informasi	**Detail**
Job	6
Title	Clinical Paramedic
Key Skills	problem solving, less, testing
Description snippet	An exciting career awaits you At MPC, we're committed to being a great place to work – one that welcomes new ideas, encourages diverse perspectives, develops our people, and fosters a collaborative….

2. Job 18: VP, Strategy & Insights: Posisi ini memiliki cakupan tanggung jawab yang sangat luas dan beragam antar-disiplin (strategi, branding, sosial), sehingga sulit dikelompokkan dengan peran yang lebih terfokus.

   Tabel 6.9 Informasi dan Detail Cluster

Informasi	**Detail**
Job	18
Title	VP, Strategy & Insights
Key Skills	agile, leadership, testing, marketing
Description snippet	What to knowWe're looking for a seasoned leader with a record of success in developing the strategy that underpins brand engagement programs, supporter experience, and creative development. You'll….

3. Job 54: Software Engineering Manager: Merupakan peran hibrida yang menggabungkan keterampilan teknis tinggi dengan aspek manajerial, sehingga bisa jadi tidak cukup padat dengan cluster teknis murni maupun manajerial murni.

   Tabel 6.10 Informasi dan Detail Cluster

Informasi	**Detail**
Job	54
Title	Software Engineering Manager
Key Skills	javascript, ai, docker, kubernetes, agile, security, leadership, problem solving, testing
Description snippet	Job Description Summary As a software manager at GE Healthcare, you will be responsible for designing and programming a small module and designing a feature, set of features, or whole feature….

Interpretasi Penyebab Noise/Kesalahan:

- Kualitas Data dan Ekstraksi Fitur: Seperti yang disebutkan dalam analisis kesalahan di notebook, hanya sekitar 80.2% pekerjaan yang memiliki keterampilan yang diekstrak. Pekerjaan dengan deskripsi yang ambigu, keterampilan yang tidak umum, atau deskripsi yang sangat singkat mungkin sulit untuk direpresentasikan dengan baik oleh fitur TF-IDF dan menghasilkan embeddings yang terisolasi. Tingkat sparsity fitur TF-IDF (11.52 sebelum normalisasi untuk 0) juga bisa menjadi faktor.
- Keterbatasan Model: Embeddings yang dihasilkan oleh model (terutama model dengan performa lebih rendah seperti "No FS + GNN" yang digunakan untuk analisis DBSCAN) mungkin tidak sepenuhnya menangkap nuansa dari semua jenis pekerjaan. Beberapa pekerjaan mungkin memiliki kombinasi keterampilan yang unik sehingga sulit dikelompokkan.
- Ambiguitas Pekerjaan: Beberapa pekerjaan mungkin secara inheren berada di persimpangan beberapa domain, sehingga sulit untuk dimasukkan secara tegas ke dalam satu cluster. DBSCAN sensitif terhadap densitas, jadi pekerjaan di area dengan densitas rendah akan dianggap sebagai noise.

- Parameter DBSCAN: Pilihan parameter eps dan min_samples sangat mempengaruhi hasil DBSCAN. Nilai yang dipilih mungkin tidak optimal untuk semua area dalam ruang fitur.

Dengan menganalisis noise points, kita dapat memperoleh pemahaman yang lebih baik tentang jenis-jenis lowongan pekerjaan yang sulit untuk diklasifikasikan oleh model saat ini dan area mana yang memerlukan perbaikan, baik dari sisi preprocessing, pemilihan fitur, maupun arsitektur model itu sendiri. Penggunaan DBSCAN pada embeddings dari model dengan performa terbaik (SFS + GAT) akan memberikan analisis error yang lebih relevan.

2. Untuk Clustering:
    - Heterogenitas Cluster: Beberapa cluster masih bersifat heterogen, yang berarti pekerjaan di dalamnya tidak sepenuhnya serupa. Misalnya, Cluster 1 dalam analisis awal menunjukkan pekerjaan dari bidang kesehatan, hukum, dan manajemen restoran tanpa adanya keterampilan umum yang signifikan, yang mungkin disebabkan oleh keterampilan yang sangat spesifik atau deskripsi pekerjaan yang kurang detail.
    - Kualitas Ekstraksi Keterampilan: Akurasi clustering sangat bergantung pada kualitas ekstraksi keterampilan. Keterampilan yang tidak terdeteksi atau salah diekstraksi dapat memengaruhi pembentukan graf kemiripan dan embedding yang dihasilkan.
    - Pemilihan Jumlah Cluster (K): Meskipun metode elbow digunakan, penentuan jumlah cluster yang optimal bisa subjektif dan mungkin perlu dieksplorasi lebih lanjut dengan metrik evaluasi clustering lainnya (misalnya Silhouette Score).
3. Untuk Link Prediction (Rekomendasi):
    - Kinerja AUC/AP yang Moderat: Meskipun Skenario 2 menunjukkan peningkatan, nilai AUC (0.74040) dan AP (0.69034) masih menunjukkan ruang untuk perbaikan. Model belum sempurna dalam membedakan semua pasangan profil-pekerjaan yang relevan dan tidak relevan.
    - Representasi Fitur Node yang Sederhana: Seperti yang disebutkan dalam dokumen Anda, penggunaan fitur node yang sederhana (misalnya hanya tipe node, tanpa informasi keterampilan langsung dalam fitur GNN) bisa menjadi salah satu penyebab kinerja yang belum maksimal.
    - Model GNN yang Relatif Sederhana: Arsitektur model GNN yang digunakan (jumlah layer, jenis agregasi) mungkin perlu disesuaikan atau dieksplorasi lebih lanjut untuk menangkap pola yang lebih kompleks.
    - Sparsity Data: Kemungkinan tidak semua hubungan ideal antara pencari kerja dan pekerjaan ada dalam dataset (data bersifat *sparse*), yang dapat menyulitkan model untuk belajar.

- Cold Start Problem: Untuk profil pencari kerja baru dengan sedikit atau tanpa data keterampilan, model mungkin kesulitan memberikan rekomendasi yang relevan dan terpersonalisasi. Rekomendasi mungkin cenderung generik.

Tahap 7 (poin: 15): Reporting

- Academic Poster.
- Report.
- Notebook (Google Colab.)
- Dashboard