

LAPORAN KELOMPPOK 3

DATA SCIENCE

“PENARIKAN DATA DARI SOCIAL MEDIA”



Disusun oleh :

- | | |
|-----------------------------|---------------|
| 1. Muhamad Ripai | (0651 21 097) |
| 2. Muhammad Defadhil Amri | (0651 21 069) |
| 3. Muhammad Ramdhan Hakim | (0651 21 088) |
| 4. Muhamad Bayu Fadayan | (0651 21 100) |
| 5. Fathur Pakapradana | (0651 21 103) |
| 6. Rafly Rahmat Amtiar | (0651 21 107) |
| 7. Zidan Al-Rasyid | (0651 21 112) |
| 8. Novi Khopipah | (0651 21 113) |
| 9. Laksana Fauzta Livepud L | (0651 21 115) |
| 10. Muhamad Yudi Aditya | (0651 21 116) |
| 11. Amalia Kartika Putri | (0651 21 118) |
| 12. Indah Cahyani | (0651 21 120) |

Dosen Pengampu :

Yusma Yanti, M.Si

Program Studi Ilmu Komputer

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Pakuan

2022

KATA PENGANTAR

Dengan menyebut nama Allah yang maha pengasih lagi maha penyayang, puji syukur kami panjatkan kehadiran Allah swt. Karena dengan rahmat dan Karunia-nya kami dapat menyelesaikan laporan tugas data science yang berjudul penarikan data dari social media tepat pada waktunya.

Kami menyadari, bahwa laporan tugas data science yang kami buat ini masih jauh dari kata sempurna baik segi penyusunan, bahasa, maupun penulisannya. Oleh karena itu, kami sangat mengharapkan kritik dan saran yang membangun dari semua pembaca guna menjadi acuan agar kami bisa menjadi lebih baik lagi di masa mendatang.

Semoga laporan tugas data science ini bisa menambah wawasan para pembaca, berguna untuk menambah pengetahuan bagi para pembaca, dan bisa bermanfaat untuk perkembangan dan peningkatan ilmu pengetahuan.

Demikian yang kami sampaikan kami berharap laporan tugas data science yang berjudul penarikan data dari social media dapat bermanfaat bagi para pembaca sekalian.

Bogor, 12 Desember 2022

Tim Penyusun

DAFTAR ISI

KATA PENGANTAR	i
DAFTAR ISI	ii
BAB 1 PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	1
1.3 Tujuan	1
BAB 2 DASAR TEORI	2
2.1 Definisi Ekstraksi Data	2
2.2 Jenis-Jenis Ekstraksi Data	2
2.3 Preprocessing Data	3
2.4 Data Cleaning	3
2.5 Transformasi Data	4
2.6 Regresi.....	4
2.7 K-Means	4
BAB 3 PEMBAHASAN	6
3.1 Proses Penarikan Data.....	6
3.2 Preprocessing Data	10
3.3 Regresi dan K-Means.....	12
BAB 4 PENUTUP	17
4.1 Kesimpulan.....	17
DAFTAR PUSTAKA	18

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Kemajuan teknologi dan perubahan terkait dalam kehidupan praktis sehari-hari telah menghasilkan perkembangan yang pesat dunia parallel konten baru, data baru, dan sumber informasi baru di sekitar kita. Terlepas dari bagaimana seseorang mendefinisikannya, fenomena atau istilah big data semakin hadir, semakin meresap, dan semakin penting.

Di sana adalah potensi nilai yang sangat besar dalam istilah yang kita kenal dengan big data termasuk seperti wawasan inovatif, pemahaman yang lebih baik tentang masalah, dan banyak lagi hal-hal lainnya. Itu juga dapat memberi peluang untuk memprediksi, dan bahkan untuk membentuk masa depan itu sendiri.

Secara umum, data science adalah sarana utama untuk menemukan dan menekankan akan potensi itu, istilah yang berarti ilmu data dalam bahasa Indonesia ini menyediakan cara untuk menangani dan memanfaatkan kumpulan data besar untuk melihat pola, untuk menemukan relasi serta untuk memahami berbagai gambar dan informasi yang memukau.

1.2 Rumusan Masalah

1. Apa itu ekstraksi data dan jenis-jenisnya?
2. Apa yang dimaksud dengan preprocessing data?
3. Apa yang dimaksud cleaning data?
4. Bagaimana tahapan proses penarikan data dan mengelolanya?

1.3 Tujuan

1. Mengetahui definisi ekstraksi data dan jenis-jenisnya
2. Mengetahui tentang preprocessing data
3. Mengetahui tentang cleaning data
4. Untuk mengetahui bagaimana tahapan proses penarikan data dan mengelolanya

BAB 2

DASAR TEORI

2.1 Definisi Ekstraksi Data

Data scraping atau yang juga sering disebut data extraction merupakan teknik atau metode otomatisasi yang memungkinkan seseorang untuk mengekstrak data dari sebuah website, database, aplikasi enterprise, atau sistem legacy yang kemudian dapat menyimpannya ke dalam sebuah file dengan format tabular atau spreadsheet. Metode mengotomatisasi proses copy paste secara manual yang dimana proses ini memakan waktu berjam-jam atau bahkan berhari-hari.

Umumnya data scraping digunakan untuk beberapa pekerjaan yang berkaitan dengan data seperti research untuk konten website, keperluan bisnis dalam komparasi harga, atau melakukan riset pasar pada sumber data publik. Kebanyakan data pada website merupakan data tidak terstruktur dalam format HTML yang kemudian diubah menjadi data dengan format terstruktur ke dalam spreadsheet atau database Anda sehingga dapat dimanipulasi. Sedangkan ada banyak cara yang digunakan dalam melakukan data scraping untuk memperoleh data dari sebuah website seperti layanan online, API tertentu atau bahkan perusahaan yang memiliki code untuk melakukan data scraping dari awal.

Cara terbaik yang bisa Anda coba adalah dengan memanfaatkan API (Application Programming Interface) yang dimiliki beberapa website besar seperti Google, Twitter, Facebook, sehingga memungkinkan Anda mengakses data mereka dengan format data terstruktur. Namun cara ini tidak berfungsi pada website lain yang tidak memiliki API atau yang tidak mengizinkan Anda untuk mengakses data dalam bentuk format terstruktur.

2.2 Jenis-Jenis Ekstraksi Data

1. Web Scraping

Web scraping memungkinkan Anda untuk mengekstrak seluruh data atau spesifik data yang Anda inginkan dari sebuah website dengan mengakses source code seperti HTML, CSS, dan Javascript ataupun menggunakan API yang disediakan pemilik website tersebut. Dengan menggunakan tools web scraping Anda dapat mengekstrak data dari website menjadi sebuah laporan yang dapat di kostumisasikan.

Web scraping membutuhkan dua bagian, yaitu crawler dan scraper dimana crawler adalah sebuah algoritma AI (Artificial Intelligence) yang melakukan pencarian data tertentu yang diperlukan dengan mengikuti link di internet. Sedangkan scraper adalah tools khusus yang dibuat untuk mengekstrak data dari website dan desain dari scraper ini dapat berbeda-beda tergantung dari tingkat kompleksitas dari pengembangnya.

2. Screen Scraping

Screen scraping merupakan tipe data scraping yang memperoleh data dari analisis visual interfaces yang dimana langsung dari tampilan website yang dapat dilihat oleh Anda. Karena tidak seperti web scraping, screen scraping tidak mengunduh dari sumber webnya melainkan melakukan scraping terhadap teks, gambar, atau konten lainnya dan membuat data tersebut ideal untuk dianalisis.

Umumnya screen scraping digunakan bagi perusahaan dan bisnis yang menggunakan cara ini untuk menyimpan data sensitif dan krusial yang merupakan merupakan data utuh dan disimpan dalam jangka waktu yang lama untuk tujuan pencatatan. Terlebih karena screen scraping sangat cocok untuk mengekstrak data tanpa mengakses source code dan tanpa API, tipe scraping ini sangat efektif untuk migrasi data karena dapat mengakses data lama dengan akurasi yang tinggi.

2.3 Preprocessing Data

Data preprocessing adalah proses yang mengubah data mentah ke dalam bentuk yang lebih mudah dipahami. Proses ini penting dilakukan karena data mentah sering kali tidak memiliki format yang teratur. Selain itu, data mining juga tidak dapat memproses data mentah, sehingga proses ini sangat penting dilakukan untuk mempermudah proses berikutnya, yakni analisis data.

2.4 Data Cleaning

Data cleaning adalah suatu prosedur untuk memastikan kebenaran, konsistensi, dan kegunaan suatu data yang ada dalam dataset. Caranya adalah dengan mendeteksi adanya error atau corrupt pada data, kemudian memperbaiki atau menghapus data jika memang diperlukan.

Alasan memakai data cleaning:

- a. Menghilangkan kesalahan dan inkonsistensi yang muncul saat beberapa data sources dikumpulkan dalam satu dataset.
- b. Meningkatkan efisiensi kerja karena proses ini akan memudahkan Anda dan tim pengolah data untuk menemukan apa yang dibutuhkan dari data.
- c. Tingkat error yang lebih rendah juga akan mendatangkan kepuasan pelanggan dan mengurangi beban kerja tim.
- d. Membantu Anda memetakan beberapa fungsi data yang berbeda. Proses ini juga akan membuat Anda lebih mengenal kegunaan data dan mempelajari asalnya.

2.5 Transformasi Data

Transformasi data merupakan suatu usaha yang ditujukan untuk mengubah skala pengukuran data asli menjadi bentuk yang lain. Dengan begitu, data tersebut bisa memenuhi asumsi yang mendasari analisis ragam yang berguna bagi penelitian. Namun, data yang akan ditampilkan pada laporan tersebut tetap menjadi data aslinya. Oleh karena itu, data transformasi tersebut dapat membantu peneliti untuk membuat data asli untuk memenuhi analisis ragam.

2.6 Regresi

Pada dasarnya, metode regresi merupakan salah satu metode analisis statistik yang bertujuan untuk menentukan hubungan sebab akibat antar variabel. Dalam pengertian lain, regresi adalah metode statistik yang digunakan untuk memperkirakan hubungan antara sebuah variabel dependen/terikat dan satu atau lebih variabel independen/bebas untuk melihat seberapa besar pengaruh variabel independen pada variabel dependen. Regresi merupakan rumus yang bisa digunakan untuk menganalisis data dari yang sederhana, sampai yang jumlahnya begitu banyak atau kompleks.

2.7 K-Means

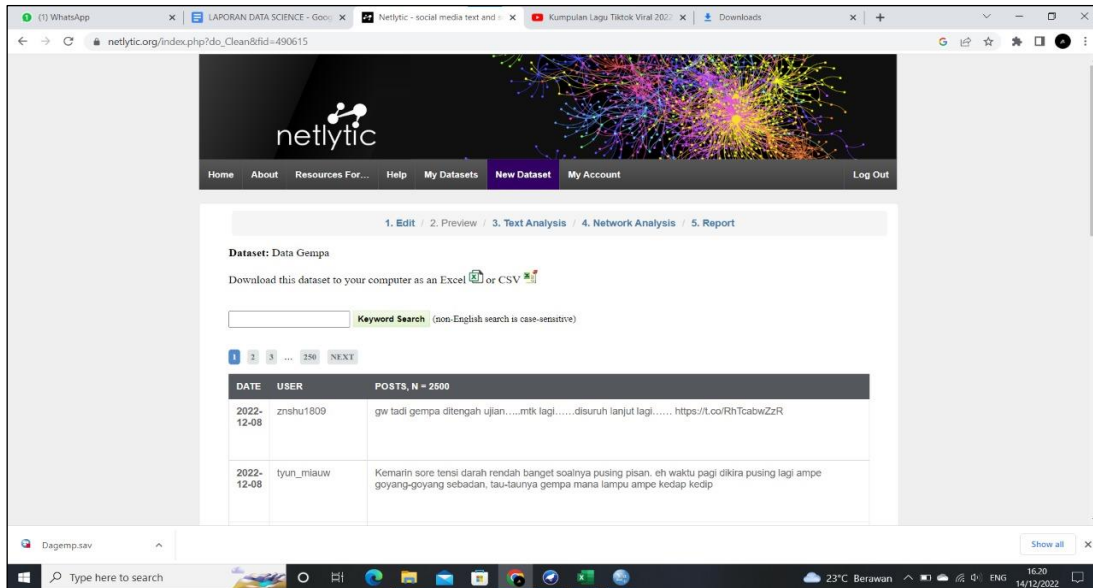
K-means merupakan salah satu algoritma yang bersifat unsupervised learning. K-Means memiliki fungsi untuk mengelompokkan data ke dalam data cluster. Algoritma ini dapat menerima data tanpa ada label kategori. K-Means Clustering Algoritma juga merupakan metode non-hierarchy. Metode Clustering Algoritma adalah mengelompokkan beberapa data ke dalam kelompok yang menjelaskan data dalam satu kelompok memiliki

karakteristik yang sama dan memiliki karakteristik yang berbeda dengan data yang ada di kelompok lain. Cluster Sampling adalah teknik pengambilan sampel di mana unit-unit populasi dipilih secara acak dari kelompok yang sudah ada yang disebut 'cluster, nah Clustering atau klasterisasi adalah salah satu masalah yang menggunakan teknik unsupervised learning.

BAB 3

PEMBAHASAN

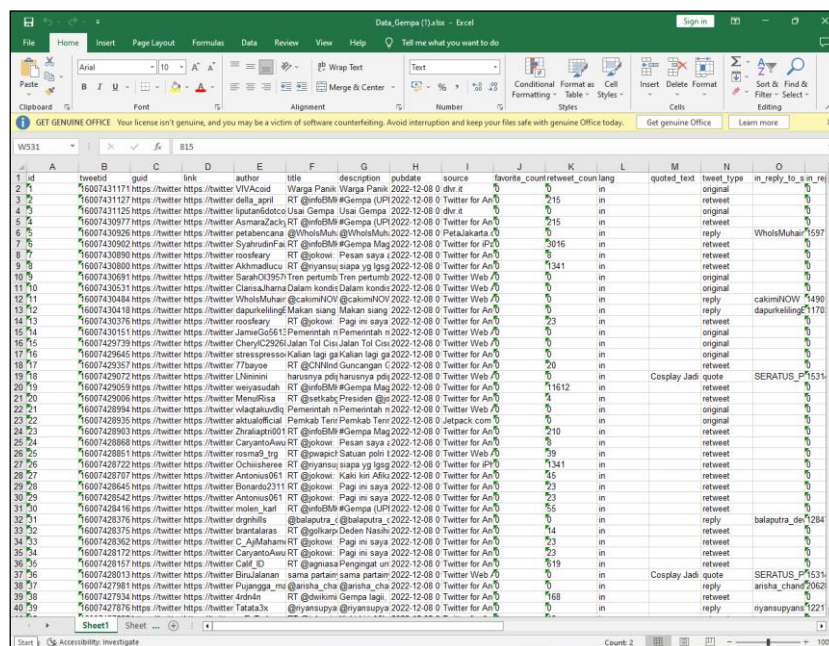
3.1 Proses Penarikan Data



The screenshot shows the Netlytic web application interface. At the top, there's a navigation bar with links like Home, About, Resources, Help, My Datasets, New Dataset, and My Account. Below this, a breadcrumb trail indicates the current view: 1. Edit / 2. Preview / 3. Text Analysis / 4. Network Analysis / 5. Report. The main content area displays the 'Dataset: Data Gempa' with a download option for Excel or CSV. A search bar is present with the text 'Keyword Search (non-English search is case-sensitive)'. Below the search bar, there's a table showing the first few rows of the dataset. The table has columns for DATE, USER, and POSTS, N = 2500. The first row shows a tweet from 'znsu1809' dated '2022-12-08' with the text 'gw tadi gempa ditengah ujian....mtk lagi....disuruh lanjut lagi..... https://t.co/RhTcbwZzR'. The second row shows a tweet from 'tyun_miauw' dated '2022-12-08' with the text 'Kemarin sore tensi darah rendah banget soalnya pusing pisan, eh waktu pagi dikira pusing lagi ampe goyang-goyang sebadan, tau-taunya gempa mana lampu ampe kedap kedip'.

Disini kita Menarik 2500 buah sample data dari sebuah platform media social yaitu Twitter menggunakan web *Netlytic.org*, yang dimana kita mengambil data berupa cuitan atau tweet yang membahas tentang gempa yang belakangan ramai di bicarakan banyak orang.

Yang dimana data yang nanti kita dapatkan sebagai berikut :



id	tweetid	guid	link	author	title	description	subdate	source	favorite_count	retweet_count	lang	quoted_text	tweet_type	in_reply_to_s
1	1600741117	https://twitter/VVAcod	Warga Panik Warga Panik	2022-12-08 0	dive it	0	in	original	0	0	0	0	0	0
2	16007431127	https://twitter/della_april	RT @inf6dM#Gempa (LPI)2022-12-08 0	Twitter for An	0	215	in	retweet	0	0	0	0	0	0
3	16007411125	https://twitter/iguanidoto	Usa Gempa	2022-12-08 0	dive 4	0	0	in	original	0	0	0	0	0
4	16007430577	https://twitter/Asmarazachy	RT @inf6dM#Gempa (LPI)2022-12-08 0	Twitter for An	0	215	in	retweet	0	0	0	0	0	0
5	16007430525	https://twitter/ptabencana	@WholsMuh	2022-12-08 0	PetaJakarta	0	0	in	reply	WholsMuh	1587	0	0	0
6	16007430565	https://twitter/SyahrulFar	RT @inf6dM#Gempa Mag	2022-12-08 0	Twitter for An	0	2016	in	retweet	0	0	0	0	0
7	16007430890	https://twitter/roosfeary	RT @okowi	Pesan saya	2022-12-08 0	Twitter for An	0	0	in	retweet	0	0	0	0
8	16007430880	https://twitter/Akhmaducy	RT @nyansu	sapa yg lgsg	2022-12-08 0	Twitter for An	0	1341	in	retweet	0	0	0	0
9	16007430891	https://twitter/SaraC3951	Tren pertumb	2022-12-08 0	Twitter Web	0	0	in	original	0	0	0	0	0
10	16007430531	https://twitter/ClassaJama	Dalam kondisi	2022-12-08 0	Twitter Web	0	0	in	original	0	0	0	0	0
11	16007430484	https://twitter/WholsMuh	@cakmNOW	2022-12-08 0	Twitter Web	0	0	in	reply	cakmNOW	1490	0	0	0
12	16007430418	https://twitter/dapurkeiling	Makan siang	2022-12-08 0	Twitter for An	0	0	in	reply	dapurkeiling	1170	0	0	0
13	16007430376	https://twitter/roosfeary	RT @okowi	Pagi ini saya	2022-12-08 0	Twitter for An	0	23	in	retweet	0	0	0	0
14	16007430151	https://twitter/JameGo5613	Pemerintah	n 2022-12-08 0	Twitter Web	0	0	in	original	0	0	0	0	0
15	16007429739	https://twitter/CherylC2908	Jalan Tol Cus	2022-12-08 0	Twitter Web	0	0	in	original	0	0	0	0	0
16	16007429645	https://twitter/stresspresso	Kalian lag	ga Kalian lag	ga 2022-12-08 0	Twitter for An	0	0	in	original	0	0	0	0
17	16007429357	https://twitter/77bayoe	RT @CINand	Guncangan C	2022-12-08 0	Twitter for An	0	20	in	retweet	0	0	0	0
18	16007429072	https://twitter/LNimini	harusnya pdj	harusnya pdj 2022-12-08 0	Twitter Web	0	0	in	Cosplay Jaki	quote	SERATUS_P	1531	0	0
19	16007429058	https://twitter/wevysadah	RT @inf6dM#Gempa Mag	2022-12-08 0	Twitter for An	0	11612	in	retweet	0	0	0	0	0
20	16007429006	https://twitter/MenuRisa	RT @setkabiz	Presiden @jo	2022-12-08 0	Twitter for An	0	4	in	retweet	0	0	0	0
21	16007428994	https://twitter/vlagakundis	Pemerintah	n 2022-12-08 0	Twitter Web	0	0	in	original	0	0	0	0	0
22	16007428935	https://twitter/abulabical	Pemkab Tem	Pemkab Tem 2022-12-08 0	Jetpack.com	0	0	in	original	0	0	0	0	0
23	16007428903	https://twitter/Zhalapri001	RT @inf6dM#Gempa Mag	2022-12-08 0	Twitter for An	0	210	in	retweet	0	0	0	0	0
24	16007428868	https://twitter/CaryantoAwi	RT @okowi	Pesan saya	2022-12-08 0	Twitter for An	0	0	in	retweet	0	0	0	0
25	16007428861	https://twitter/rosmad2	Jg RT @wspci	Saturnus poin	12-02-12-08 0	Twitter Web	0	29	in	retweet	0	0	0	0
26	16007428722	https://twitter/Ochiishere	RT @nyansu	sapa yg lgsg	2022-12-08 0	Twitter for An	0	1341	in	retweet	0	0	0	0
27	16007428707	https://twitter/Antonius061	RT @okowi	Kalo ini Abik	2022-12-08 0	Twitter for An	0	45	in	retweet	0	0	0	0
28	16007428645	https://twitter/Bonar02311	RT @okowi	Pagi ini saya	2022-12-08 0	Twitter for An	0	23	in	retweet	0	0	0	0
29	16007428642	https://twitter/Antonius061	RT @okowi	Pagi ini saya	2022-12-08 0	Twitter for An	0	23	in	retweet	0	0	0	0
30	16007428416	https://twitter/molen_karl	RT @inf6dM#Gempa (LPI)2022-12-08 0	Twitter for An	0	55	in	retweet	0	0	0	0	0	0
31	16007428376	https://twitter/dgrhills	@balaputra_c	2022-12-08 0	Twitter for An	0	0	in	reply	balaputra_de	1284	0	0	0
32	16007428375	https://twitter/bratelasas	RT @gullikan	Desden Har	2022-12-08 0	Twitter for An	0	14	in	retweet	0	0	0	0
33	16007428362	https://twitter/C_AyMaham	RT @okowi	Pagi ini saya	2022-12-08 0	Twitter for An	0	23	in	retweet	0	0	0	0
34	16007428172	https://twitter/CaryantoAwi	RT @okowi	Pagi ini saya	2022-12-08 0	Twitter for An	0	23	in	retweet	0	0	0	0
35	16007428157	https://twitter/Calli_O	RT @nyansu	Pengumpul un	2022-12-08 0	Twitter for An	0	618	in	retweet	0	0	0	0
36	16007428013	https://twitter/Birulanjan	sama partam	sama partam 2022-12-08 0	Twitter Web	0	0	in	Cosplay Jaki	quote	SERATUS_P	1531	0	0
37	16007427981	https://twitter/Pujangga_me	@arisha_cha	2022-12-08 0	Twitter for An	0	0	in	reply	arisha_cha	2062	0	0	0
38	16007427934	https://twitter/ArindaA	RT @khalim	Gempa lagi	2022-12-08 0	Twitter for An	0	168	in	retweet	0	0	0	0
39	16007427876	https://twitter/Tatata3x	@nyansu	2022-12-08 0	Twitter for An	0	0	in	reply	nyansu	1221	0	0	0

Yang dimana data awalnya berjumlah sekitar 2500 responden dan variabel atau atributnya berjumlah 29 buah.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	id	tweetid	guid	link	author	title	description	pubdate	source	favorite_count	retweet_count	lang	quoted_text	tweet_type
2	1	6007431171	https://twitter.com/VN/Accid	RT @infoBMi#Gempa	Warga Panik	Warga Panik	2022-12-08 0	0	0	0	0	in		original
3	2	6007431127	https://twitter.com/delta_april	RT @infoBMi#Gempa	Usai Gempa	Usai Gempa	2022-12-08 0	0	0	215	0	in		retweet
4	3	6007431125	https://twitter.com/liputan6dotco	Usai Gempa	Usai Gempa	Usai Gempa	2022-12-08 0	0	0	0	0	in		original
5	4	6007430977	https://twitter.com/AsmaraZack	RT @infoBMi#Gempa	Warga Panik	Warga Panik	2022-12-08 0	0	0	215	0	in		retweet
6	5	6007430926	https://twitter.com/petabencana	@WholsMuh	@WholsMuh	@WholsMuh	2022-12-08 0	0	0	0	0	in		reply
7	6	6007430902	https://twitter.com/SyahudinFai	RT @infoBMi#Gempa	Mag 2022-12-08 0	0	0	0	0	5016	0	in		retweet
8	7	6007430890	https://twitter.com/roosefary	RT @jokowi	Pesan saya	Pesan saya	2022-12-08 0	0	0	0	0	in		retweet
9	8	6007430800	https://twitter.com/Akhmaducdu	RT @nyansu	siapa yg lsgg	2022-12-08 0	0	0	0	1341	0	in		retweet
10	9	6007430691	https://twitter.com/SarahO39571	Tren pertumb	Tren pertumb	Tren pertumb	2022-12-08 0	0	0	0	0	in		original
11	10	6007430631	https://twitter.com/Clarisajama	Dalam kondisi	Dalam kondisi	Dalam kondisi	2022-12-08 0	0	0	0	0	in		original
12	11	6007430484	https://twitter.com/WholsMuhar	@cakimiNOV	@cakimiNOV	@cakimiNOV	2022-12-08 0	0	0	0	0	in		reply
13	12	6007430418	https://twitter.com/dapurkellings	Makan siang	Makan siang	Makan siang	2022-12-08 0	0	0	0	0	in		original
14	13	6007430376	https://twitter.com/roosefary	RT @jokowi	Pagi ini saya	Pagi ini saya	2022-12-08 0	0	0	23	0	in		retweet
15	14	6007430151	https://twitter.com/JamieGo5612	Pemerintah n	Pemerintah n	Pemerintah n	2022-12-08 0	0	0	0	0	in		original
16	15	6007429739	https://twitter.com/CheryC29261	Jalan Tol Cis	Jalan Tol Cis	Jalan Tol Cis	2022-12-08 0	0	0	0	0	in		original
17	16	6007429645	https://twitter.com/stresspresso	Kalian lagi ga	Kalian lagi ga	Kalian lagi ga	2022-12-08 0	0	0	0	0	in		original
18	17	6007429357	https://twitter.com/77bayoe	RT @CHNWed	Guncangan	Guncangan	2022-12-08 0	0	0	20	0	in		retweet
19	18	6007429072	https://twitter.com/LNinimi	harusnya pdj	harusnya pdj	harusnya pdj	2022-12-08 0	0	0	0	0	in	Cosplay Jadi	quote
20	19	6007429059	https://twitter.com/weiyasudat	RT @infoBMi#Gempa	Mag 2022-12-08 0	0	0	0	0	11612	0	in		retweet
21	20	6007429006	https://twitter.com/MenuRisa	RT @setkabg	Presiden @	2022-12-08 0	0	0	0	4	0	in		retweet
22	21	6007428994	https://twitter.com/vladgalkudj	Pemerintah n	Pemerintah n	Pemerintah n	2022-12-08 0	0	0	0	0	in		original
23	22	6007428935	https://twitter.com/aktualofficial	Pemkab Teri	Pemkab Teri	Pemkab Teri	2022-12-08 0	0	0	0	0	in		original
24	23	6007428903	https://twitter.com/Zhralapti001	RT @infoBMi#Gempa	Mag 2022-12-08 0	0	0	0	0	210	0	in		retweet
25	24	6007428868	https://twitter.com/CaryantoAwwu	RT @jokowi	Pesan saya	Pesan saya	2022-12-08 0	0	0	8	0	in		retweet
26	25	6007428851	https://twitter.com/rosma9_trg	RT @pwapic1	Satuan polri	2022-12-08 0	0	0	0	39	0	in		retweet
27	26	6007428722	https://twitter.com/Ochiishiene	RT @nyansu	siapa yg lsgg	2022-12-08 0	0	0	0	1341	0	in		retweet
28	27	6007428707	https://twitter.com/Antonius061	RT @jokowi	Kaki kn Adus	2022-12-08 0	0	0	0	45	0	in		retweet
29	28	6007428645	https://twitter.com/Bonardo2311	RT @jokowi	Pagi ini saya	Pagi ini saya	2022-12-08 0	0	0	23	0	in		retweet
30	29	6007428542	https://twitter.com/Antonius061	RT @jokowi	Pagi ini saya	Pagi ini saya	2022-12-08 0	0	0	23	0	in		retweet
31	30	6007428416	https://twitter.com/molen_karl	RT @infoBMi#Gempa	(UPI) 2022-12-08 0	0	0	0	0	55	0	in		retweet
32	31	6007428376	https://twitter.com/dgrhills	@balaputra_c	@balaputra_c	@balaputra_c	2022-12-08 0	0	0	0	0	in		reply
33	32	6007428375	https://twitter.com/brantalaras	RT @golikarp	Deden Naasih	2022-12-08 0	0	0	0	14	0	in		retweet
34	33	6007428362	https://twitter.com/C_AjMhamam	RT @jokowi	Pagi ini saya	Pagi ini saya	2022-12-08 0	0	0	23	0	in		retweet
35	34	6007428172	https://twitter.com/CaryantoAwwu	RT @jokowi	Pagi ini saya	Pagi ini saya	2022-12-08 0	0	0	23	0	in		retweet
36	35	6007428157	https://twitter.com/Calif_ID	RT @agniasa	Pengantar un	2022-12-08 0	0	0	0	519	0	in		retweet
37	36	6007428013	https://twitter.com/BisaJalanan	sama pertam	sama pertam	sama pertam	2022-12-08 0	0	0	0	0	in	Cosplay Jadi	quote
38	37	6007427981	https://twitter.com/Puangga_ma	@arisha_cha	@arisha_cha	@arisha_cha	2022-12-08 0	0	0	0	0	in		reply
39	38	6007427934	https://twitter.com/4rdn4n	RT @dwikimi	Gempa lagi	2022-12-08 0	0	0	0	168	0	in		retweet
40	39	6007427876	https://twitter.com/Tatata3x	@nyansupya	@nyansupya	@nyansupya	2022-12-08 0	0	0	0	0	in		reply

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
	1	in_reply_to_s	in_reply_to_u	in_reply_to_s	retweeted_sc	retweeted_u	retweeted_st	user_id	profile_image	image_user	status	user_friends	user_follower	user_created	user_bio	user_location	user_verified												
2	0	0	0	0	0	0	0	41730943	https://pbs.twimg.com/profile_images/1620514	646129	2009-05-21 22:19:27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	6006882278	234367184	https://pbs.twimg.com/profile_images/11490	621	207	2011-01-05 01:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	6006882278	234367184	https://pbs.twimg.com/profile_images/1655996	677	4485061	2009-06-16 01:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	6006882278	2773859315	https://pbs.twimg.com/profile_images/1084805	491	279	2014-09-17 15:35:53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	WholsMuhar	15971396643	600743048463282176	0	0	0	0	2726808095	https://pbs.twimg.com/profile_images/2448	311	132	2014-10-04 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	108543358	6006554329	2849711744	https://pbs.twimg.com/profile_images/25647	761	396	2010-04-09 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0	jokowi	366987179	6007409912	131370752	https://pbs.twimg.com/profile_images/303	27	11	2021-03-09 11:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	0	0	0	0	0	0	0	riyansyupani	1221719069	6006545769	1369380690	https://pbs.twimg.com/profile_images/6	14	1	2022-11-20 00:39:14	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	6539433958	https://pbs.twimg.com/profile_images/75	33	15	2022-06-21 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	0	0	0	0	0	0	0	15971396643	https://pbs.twimg.com/profile_images/14	1	2022-11-28 03:06:15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	cakimiNOW	1490116879	600341167165489153	0	0	0	0	171036907	https://pbs.twimg.com/profile_images/628	14404	2010-02-20 01:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	dapurkellings	171036907	6004702805565992	0	0	0	0	6007409683	131370752	https://pbs.twimg.com/profile_images/761	396	2010-04-09 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	6534174863	https://pbs.twimg.com/profile_images/761	396	2010-04-09 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	0	0	0	0	0	0	0	69522567688	https://pbs.twimg.com/profile_images/761	396	2010-04-09 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	11305406040	https://pbs.twimg.com/profile_images/1304	1689	2019-05-20 10:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	0	0	0	0	0	0	0	CINIndonesia	17128975	6006651343	2191962343	https://pbs.twimg.com/profile_images/1092	3083	2013-11-10 01:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	0	0	0	0	0	0	0	15843802129	https://pbs.twimg.com/profile_images/158	158	2018-05-22 10:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	6006543034	695332160391	https://pbs.twimg.com/profile_images/157	180	2019-02-12 10:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	0	0	0	0	0	0	0	6006543034	695332160391	https://pbs.twimg.com/profile_images/157	180	2019-02-12 10:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	0	0	0	0	0	0	0	6007784265	14075394689	https://pbs.twimg.com/profile_images/15510	447	334	2021-03-02 10:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	0	0	0	0	0	0	0	76384607278	https://pbs.twimg.com/profile_images/20	0	0	2016-08-11 17:14:30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	0	0	0	0	0	0	0	11421377068	https://pbs.twimg.com/profile_images/38176	61	836	2019-06-21 11:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	0	0	0	0	0	0	0	108543358	6006846273	17508653129	https://pbs.twimg.com/profile_images/210	204	2020-09-14 10:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	0	0	0	0	0	0	0	366987179	6007409912	131370752	https://pbs.twimg.com/profile_images/74542	4537	2021-10-30 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	0	0	0	0	0	0	0	pusguchuko	6008042423	60031720745	1043137419	https://pbs.twimg.com/profile_images/414	2020-08-12 23:52:52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	0	0	0	0	0	0	0	riyansyupani	1221719069	6006545769	12220025216	https://pbs.twimg.com/profile_images/4007	235	29	2020-01-27 20:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	366987179	6007359692	13278893653	https://pbs.twimg.com/profile_images/38203	14128	15651	2020-11-15 02:24:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
29	0	0	0	0	0	0	0	jokowi	366987179	6007409683	1524529085	https://pbs.twimg.com/profile_images/5125	8988	8186	2022-05-01 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	366987179	6007409683	13278893653	https://pbs.twimg.com/profile_images/38203	14128	15651	2020-11-15 02:24:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
31	0	0	0	0	0	0	0	108543358	6007204000	1030802147	https://pbs.twimg.com/profile_images/3702	1016	1016	2020-08-12 23:52:52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
32	0	0	0	0	0	0	0	6596105683	https://pbs.twimg.com/profile_images/7456	448	28	2022-11-25 10:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
33	balasputra	62847544811	600741123156500480	0	0	0	0	17218983540	6007316502	15091939086	https://pbs.twimg.com/profile_images/7007	25	21	2022-11-08 01:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
34	0	0	0	0	0	0	0	366987179	6007409683	188495265731	https://pbs.twimg.com/profile_images/70083	4221	4680	2017-07-11 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
35	0	0	0	0	0	0	0	jokowi	366987179	6007409683	1454089934	https://pbs.twimg.com/profile_images/74542	4291	3537	2021-10-30 21:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	agniasamsambal	330676201	6006653608	809151731	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
53	0	0	0	0	0	0	0	1646841800	https://pbs.twimg.com/profile_images/7495	126	408	2022-06-15 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0					

Kemudian kita ambil beberapa atribut yang sekiranya bisa kita ambil atau olah yaitu Tweet id, User id, User Friends, User follower, User Statuses count dan Tweet Type. dan juga beberapa responden yaitu sekitar 499 orang/akun.

Kemudian kita jadikan file yang tadinya berbentuk format .xlsx menjadi .csv dengan cara menyimpan file excel tersebut menjadi file CSV melalui aplikasi excel. yang dimana hasilnya sebagai berikut:

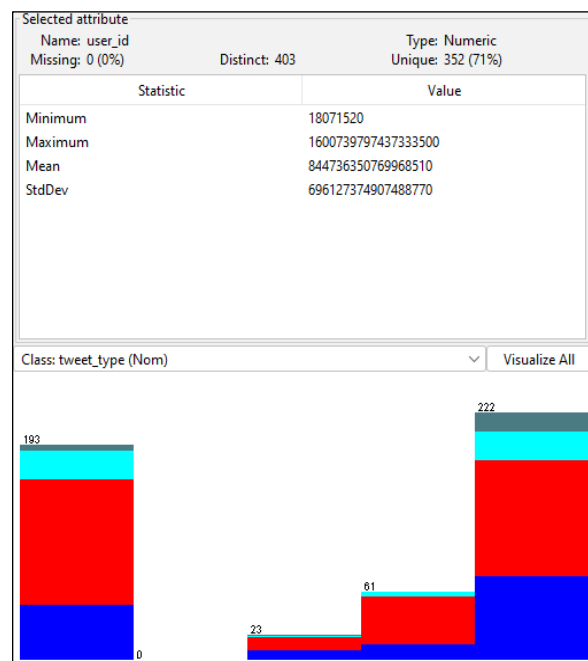
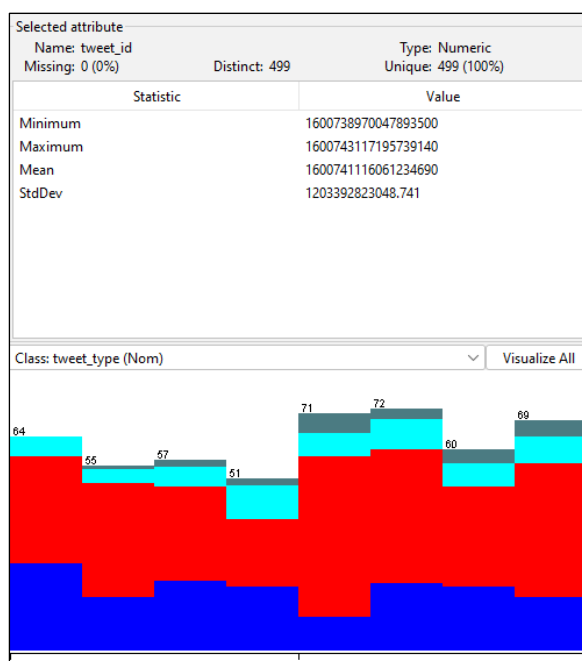
```

Data_Gempa_fa.csv - Notepad
File Edit Format View Help
tweet_id,user_id,user_friends,user_follower,user_statuses count,tweet_type
1600743117195739136,41730943,3,4646129,1620514,original
1600743112774598656,234367184,821,207,11490,retweet
1600743112560689153,47596019,677,4485061,1665996,original
160074309779667633,2773859315,491,479,3086,retweet
1600743092608708417,2276880895,331,147764,1084805,reply
1600743090293440514,2849711744,314,132,2448,retweet
1600743089035177984,131370752,761,396,25647,retweet
1600743080025415680,1369380690643611652,27,11,303,retweet
1600743069124464640,1594203599915343873,6,1,8,original
1600743053135777793,1539443995805306882,33,15,75,original
1600743048463282176,1597139664368209920,14,1,8,reply
1600743041806966784,117036907,5628,14404,84927,reply
1600743037638152192,131370752,761,396,25647,retweet
1600743015123152896,1594174365364428806,4,1,8,original
1600742973922508800,1592256768817627139,0,0,12,original
1600742964569210881,1130540064025468928,1394,1589,37237,original
1600742935728779264,2191962343,5002,3083,489456,retweet
1600742907228483584,1584380212912848898,5,3,158,quote
1600742905945419782,995332160918732000,567,190,30954,retweet
1600742900626694145,1407539468966367236,847,334,15510,retweet
1600742899431661568,763846072783351808,0,0,20,original
1600742893542686720,1142137706840875008,61,836,38176,original
1600742890350600192,1570065312917393409,20,0,18,retweet
1600742886878109696,1454608993456775171,2214,4537,74542,retweet
1600742885196173312,704317219,44,1,897,retweet
1600742872248029185,1222082251666903041,235,29,4007,retweet
1600742870738120704,1327889363560472064,14128,15651,38203,retweet
1600742864509894656,1524552908532232198,8988,8186,5125,retweet
1600742854275387392,1327889363560472064,14128,15651,38203,retweet
1600742841646338048,1038060147744989185,2016,631,13762,retweet
1600742837611507712,159616508366963713,248,28,1456,reply
1600742837569800064,159019390808249856,25,21,1007,retweet
1600742836269641728,884952657311309824,4921,3680,70083,retweet
1600742817290776064,1454608993456775171,2214,4537,74542,retweet
1600742815763308544,609115713,126,408,7495,retweet
1600742801301327873,1584377290875543552,7,1,153,quote
1600742798130483201,1514684180088569872,2215,2053,9560,reply
1600742793479344129,232390238,41,617,305779,retweet
1600742787645050882,1259643509318881280,31,3,986,reply
1600742765847277568,1219453523856781312,1984,4329,18230,retweet
1600742761409675204,766679035081824769,0,0,17,original
1600742742337892352,635221728,228,174,12793,quote
1600742739254996992,1266447850164023296,134,19,4787,retweet
1600742737157857280,1418465388015226882,1251,192,41036,retweet
1600742723916472321,1365599502510395396,499,5050,5573,original
1600742722719147649,2276880895,331,147764,1084805,original
1600742720670081025,3303234782,755,47,121877,retweet
1600742714899066881,153944191235100673,32,15,76,original
1600742710536994816,704317219,44,1,897,retweet
Ln 1, Col 1 100% Windows (CRLF) UTF-8 with BOM

```

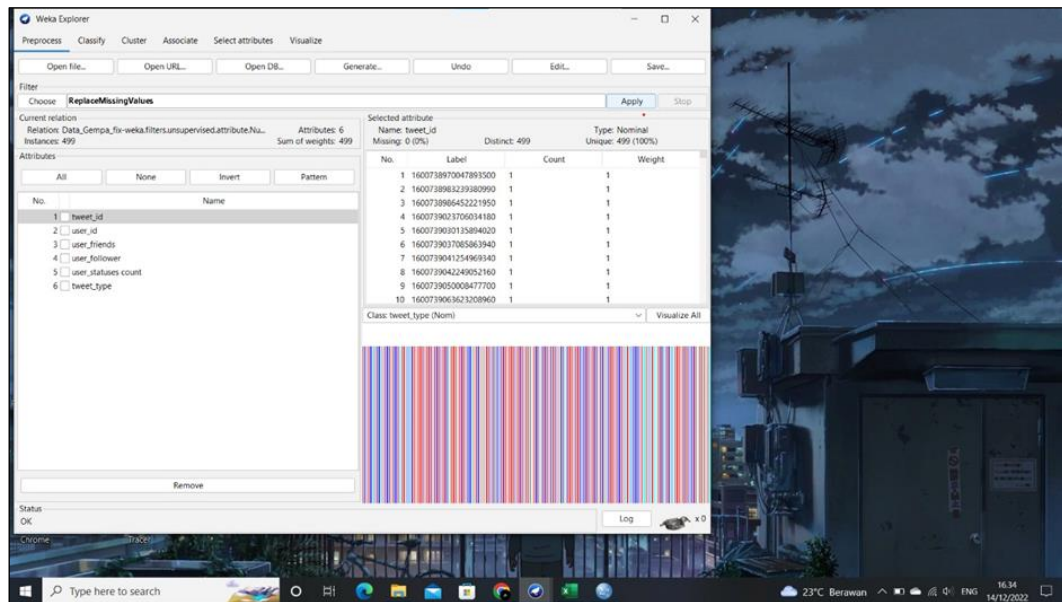
Kemudian baru data bisa kita masukkan kepada aplikasi WEKA,dan baru kita masuk kepada tahap pre-processing.

Di dalam aplikasi WEKA kita mendapatkan beberapa atribut tersebut bertipe data yakni : Tweet id, User id, User Friends, User follower, User Statuses count bertipe data numeric dan Tweet type bertipe data nominal.



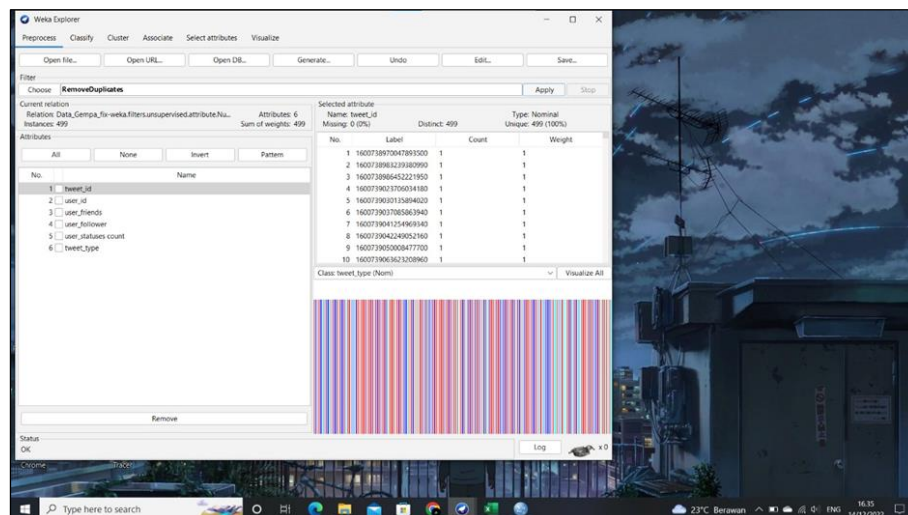
3.2 Preprocessing Data

Disini yang kita lakukan pertama kali dalam Pre- Processing Data yaitu kita membersihkan data tersebut dari missing data dan data yang mengandung duplikat (Data Cleaning). Sebagai berikut:

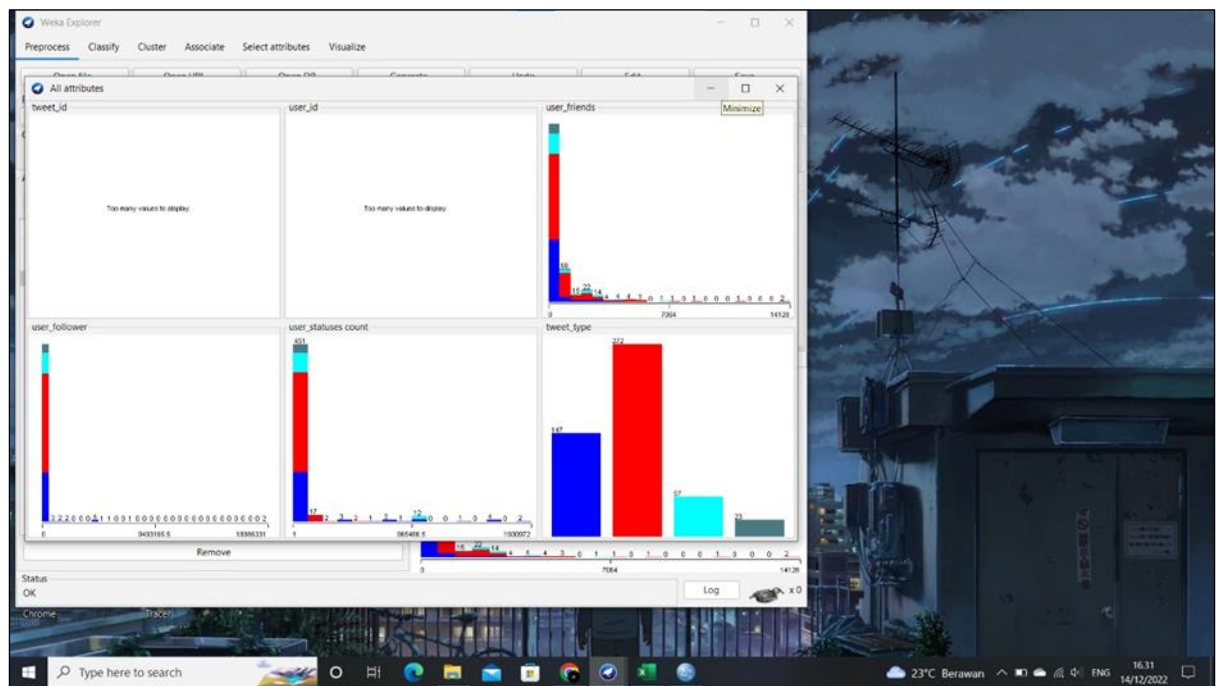


Yang dimana kita menggunakan metode ReplaceMissingValues, yang menjadi semua atribut data yang berparameter Missing menjadi 0 atau 0% yang menandakan bahwa data tidak mempunyai Missing Value.

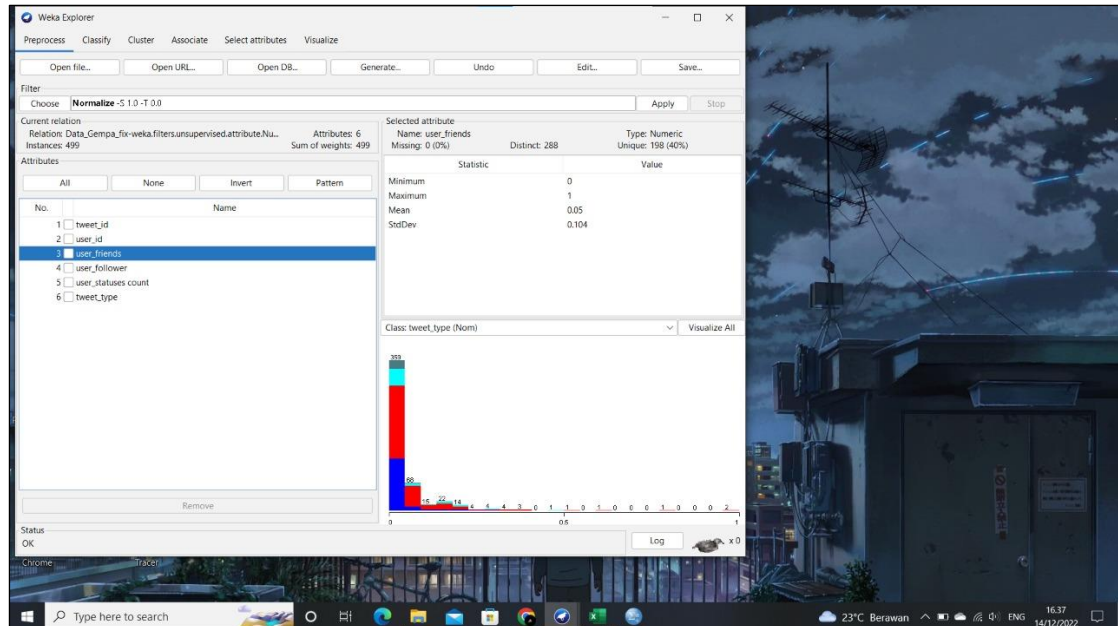
Kemudian kita menggunakan metode Remove Duplicate yang sekiranya kita memiliki data yang sama atau duplikat data akan dihapus

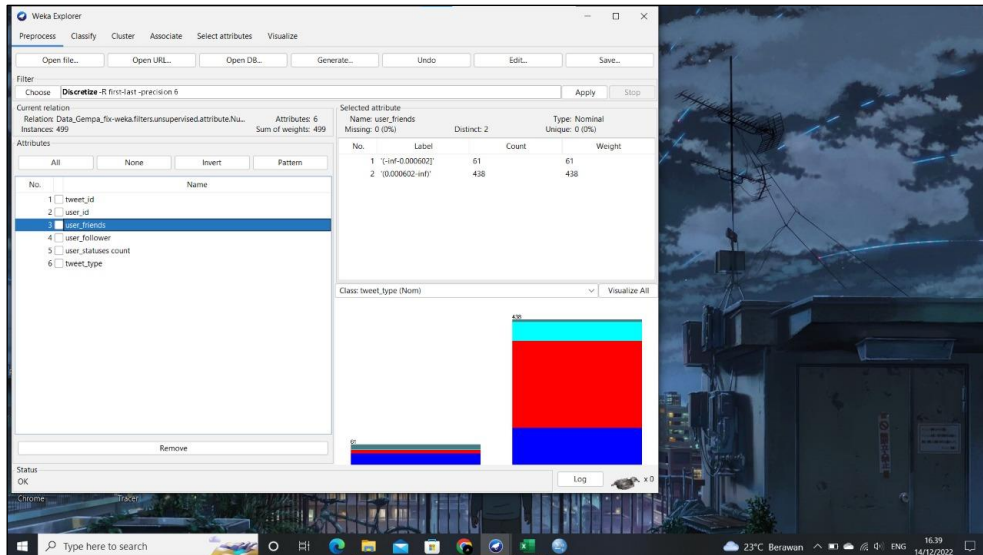


Yang dimana pada tahap ini hasilnya sebagai berikut:



Kemudian setelah itu kita masuk kepada tahap tranformasi data, yang dimana pada tahap ini kita menggunakan metode normalisasi dan diskritisasi. Sebagai berikut:

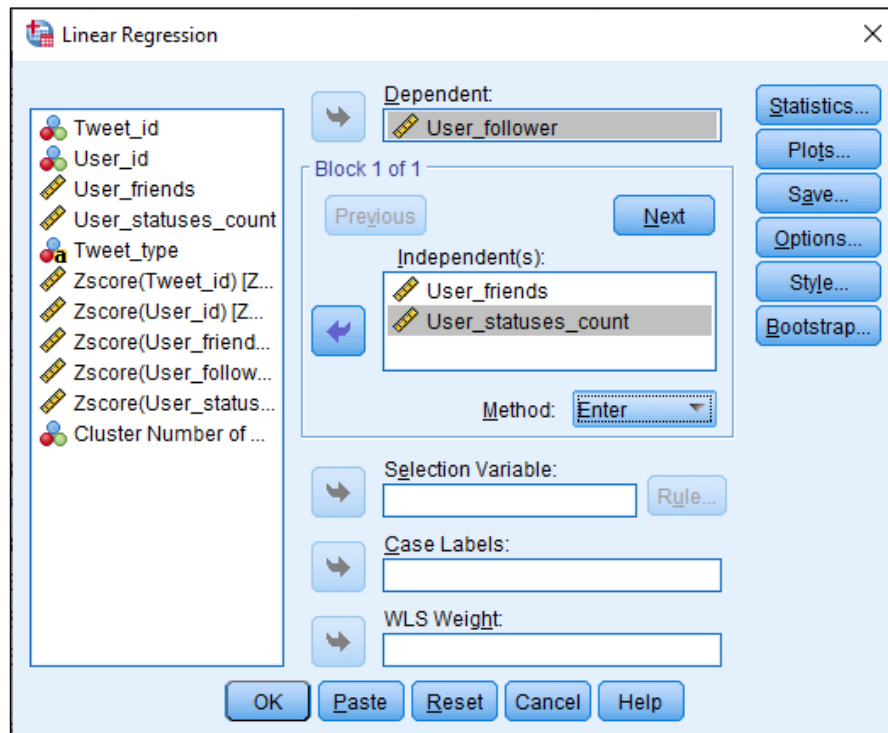




3.3 Regresi dan K-Means

Selanjutnya adalah regresi yang dimana kita disini menggunakan aplikasi SPSS yang dimana kita mengambil metode regresi linier berganda. Untuk mengetahui apakah jumlah pertemanan dan jumlah status mempengaruhi jumlah follower. Dimana datanya sebagai berikut:

	Tweet_id	User_id	User_friends	User_follower	User_statuses_count	Tweet_type
1	1600743117195739140	41730943	3.0	4646129.0	1620514.0	original
2	1600743112774598660	234367184	821.0	207.0	11490.0	retweet
3	160074311256089150	47596019	677.0	4485061.0	1665996.0	original
4	1600743097796677630	2773859315	491.0	479.0	3006.0	retweet
5	1600743092608700420	2276880895	331.0	147764.0	1084805.0	reply
6	1600743090293440510	2849711744	314.0	132.0	2448.0	retweet
7	1600743089035177980	131370752	761.0	396.0	25647.0	retweet
8	1600743080025415680	1369380690643611650	27.0	11.0	303.0	retweet
9	1600743069124464640	1594203599915343870	6.0	1.0	8.0	original
10	1600743053135777790	1539443995805306880	33.0	15.0	75.0	original
11	1600743048463282180	1597139664368209920	14.0	1.0	8.0	reply
12	1600743041806966780	117036907	6628.0	14404.0	84927.0	reply
13	1600743037638152190	131370752	761.0	396.0	25647.0	retweet
14	1600743015123152900	1594174365364428800	4.0	1.0	8.0	original
15	1600742973922508800	1592256768817627140	.0	.0	12.0	original
16	1600742964569210880	1130540064025468930	1394.0	1589.0	37237.0	original
17	1600742935728779260	2191962343	5002.0	3083.0	489456.0	retweet
18	1600742907228483580	1584380212912848900	5.0	3.0	158.0	quote
19	1600742905945419780	995332160918732800	567.0	190.0	30954.0	retweet
20	1600742900626694140	1407539468966367230	847.0	334.0	15510.0	retweet
21	1600742899431661570	763846072783351810	.0	.0	20.0	original
22	1600742893542686720	1142137706840875010	61.0	836.0	38176.0	original
23	1600742890350600190	1570065312917393410	20.0	.0	18.0	retweet
24	1600742886878109700	1454608993456775170	2214.0	4537.0	74542.0	retweet
25	1600742885196173310	704317219	44.0	1.0	897.0	retweet
26	1600742872248029180	1222002251666903040	235.0	29.0	4007.0	retweet
27	1600742870738120700	1327889365360472060	14128.0	15651.0	38203.0	retweet
28	1600742864509894660	1524552908532232190	8988.0	8186.0	5125.0	retweet
29	1600742854275387390	1327889365360472060	14128.0	15651.0	38203.0	retweet
30	1600742841646338050	1038060147744989180	2016.0	631.0	13762.0	retweet
31	1600742837611507710	1596165058365063710	248.0	28.0	1456.0	reply
32	1600742837569800060	1590193908608249860	25.0	21.0	1007.0	retweet
33	1600742836269641730	884952657311309820	4921.0	3680.0	70083.0	retweet
34	1600742817298776060	1454608993456775170	2214.0	4537.0	74542.0	retweet
35	1600742815763308540	609115173	126.0	408.0	7495.0	retweet



Kemudian kita jadikan atribut User follower sebagai variabel dependen (variabel Y) yaitu variabel yang dipengaruhi, sedangkan untuk atribut User friends dan User statuses count kita jadikan sebagai variabel independen (variabel X) yaitu variabel yang mempengaruhi.

Yang dimana hasil regresi nya sebagai berikut:

The screenshot shows the IBM SPSS Statistics Viewer window. The left sidebar contains a tree view with 'Output' expanded, showing 'Regression', 'Model Summary', 'ANOVA', and 'Coefficients'. The main area displays the following tables:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.224 ^a	.055	.051	1318219.935

a. Predictors: (Constant), User_statuses_count, User_friends

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5,015E+13	2	2,508E+13	14,431	,000 ^b
	Residual	8,619E+14	496	1,738E+12		
	Total	9,121E+14	498			

a. Dependent Variable: User_follower
b. Predictors: (Constant), User_statuses_count, User_friends

Coefficients^a

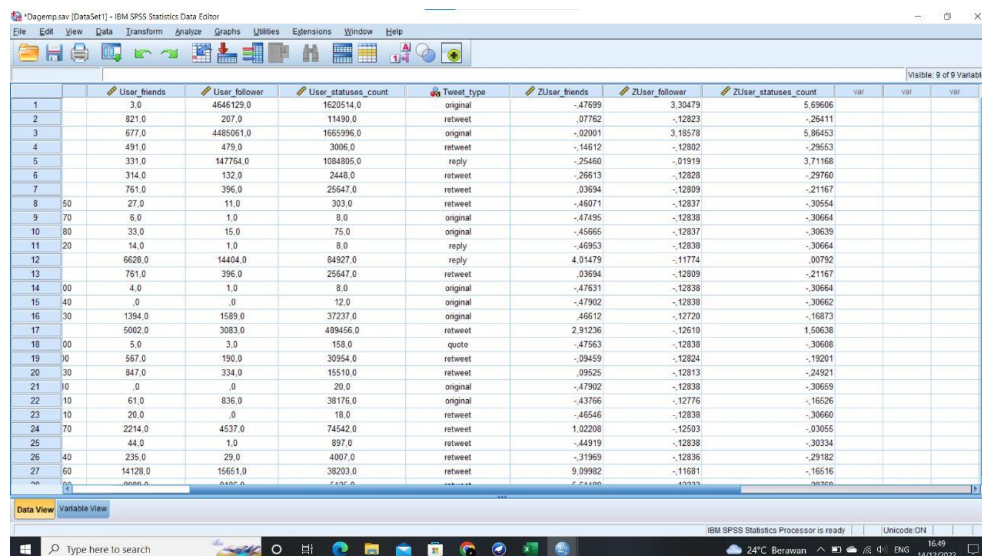
Model		Unstandardized Coefficients	Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	96952,021	67833,855	1,428	,154
	User_friends	-27,931	40,052	-.690	,486
	User_statuses_count	1,167	,219	,533	,593

a. Dependent Variable: User_follower

Kita bisa lihat nilai signifikansi dari tabel anova adalah 0, itu nilainya kurang dari 0,05, yang berarti jumlah pertemanan dan jumlah status mempengaruhi jumlah follower

Kemudian kita bisa lihat pada tabel coefficient. Pada jumlah pertemanan kita bisa melihat nilai signifikansi nya yaitu 0,486 yang dimana nilainya itu lebih dari 0,05, yang artinya jumlah pertemanan tidak mempengaruhi jumlah follower. Tapi pada jumlah status, Kita bisa lihat nilai signifikansi nya adalah 0 yang dimana kurang dari 0,05, yang mengartikan bahwa Jumlah status mempengaruhi jumlah follower.

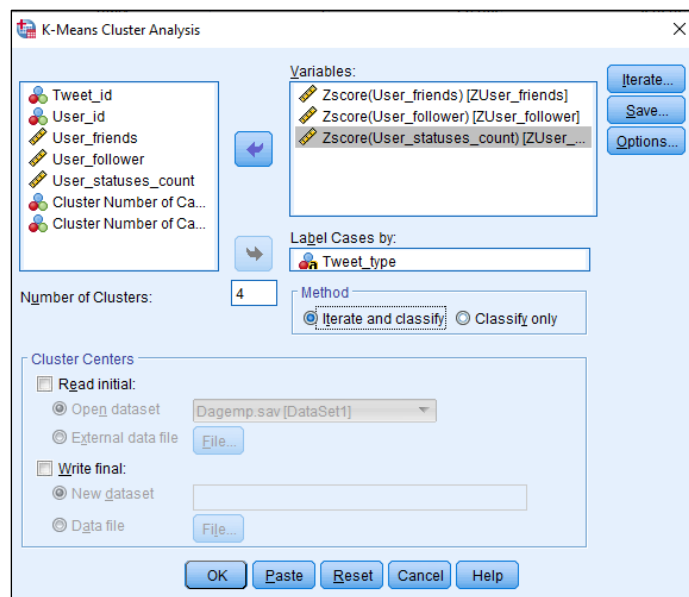
Kemudian kita masuk Ke tahap clustering K- mean Yang dimana kita menggunakan 2 metode yaitu dengan aplikasi SPSS dan WEKA. Pada aplikasi SPS Kita terlebih dahulu menormalisasi kan data yang ada. Yang tujuannya adalah menyamakan semua satuan atribut menjadi Zscore



	User_friends	User_follower	User_statuses_count	Tweet_type	ZUser_friends	ZUser_follower	ZUser_statuses_count			
1	3.0	4645129.0	1620514.0	original	-.47699	3.30479	5.69606			
2	821.0	207.0	11490.0	retweet	.07782	-.12823	-.26411			
3	677.0	4485061.0	1665396.0	original	-.02001	3.18578	5.86453			
4	491.0	479.0	3006.0	retweet	-.14612	-.12802	-.29653			
5	331.0	147764.0	1084905.0	reply	-.25460	-.01919	3.71168			
6	314.0	132.0	2448.0	retweet	-.26613	-.12828	-.29760			
7	761.0	396.0	25647.0	retweet	.03694	-.12899	-.21167			
8	27.0	11.0	303.0	retweet	-.46071	-.12837	-.30554			
9	6.0	1.0	8.0	original	-.47495	-.12838	-.30664			
10	33.0	15.0	75.0	original	-.45665	-.12837	-.30639			
11	14.0	1.0	8.0	reply	-.46953	-.12838	-.30664			
12	6620.0	14404.0	84927.0	reply	4.01479	-.11774	.00792			
13	761.0	396.0	25647.0	retweet	.03694	-.12899	-.21167			
14	4.0	1.0	8.0	original	-.47631	-.12838	-.30664			
15	0	0	12.0	original	-.47902	-.12838	-.30662			
16	1394.0	1589.0	37237.0	original	.46612	-.12720	-.16873			
17	5002.0	3083.0	489456.0	retweet	2.91236	-.12610	1.50638			
18	5.0	3.0	158.0	quote	-.47563	-.12838	-.30608			
19	567.0	190.0	30954.0	retweet	-.09459	-.12824	-.19201			
20	847.0	334.0	15519.0	retweet	.09525	-.12813	-.24921			
21	0	0	20.0	original	-.47902	-.12838	-.30669			
22	61.0	836.0	38176.0	original	-.43766	-.12776	-.16526			
23	20.0	0	18.0	retweet	-.46546	-.12838	-.30660			
24	2214.0	4537.0	74542.0	retweet	1.02208	-.12593	.03055			
25	44.0	1.0	897.0	retweet	-.44919	-.12838	-.30334			
26	235.0	29.0	4007.0	retweet	-.31969	-.12836	-.29182			
27	14128.0	15651.0	38203.0	retweet	9.09982	-.11681	-.16516			

Kemudian kita bisa mengolah datanya menggunakan data Zscore tersebut untuk proses clustering K-mean.

Pada proses clustering kita membagikan data data tersebut menjadi 4 kelompok yang dikelompokkan berdasarkan Tweet type.



Yang didapatkan hasilnya sebagai berikut :

Final Cluster Centers				
	Cluster			
	1	2	3	4
Zscore(User_friends)	-.15719	3.86238	-.15139	-.43970
Zscore(User_follower)	.91948	-.12252	-.10472	13.90123
Zscore (User_statuses_count)	4.13066	-.02470	-.21606	-.28792

Distances between Final Cluster Centers				
Cluster	1	2	3	4
1		5.874	4.466	13.716
2	5.874		4.018	14.671
3	4.466	4.018		14.009
4	13.716	14.671	14.009	

Number of Cases in each Cluster		
Cluster	1	
1	24.000	
2	19.000	
3	454.000	
4	2.000	
Valid	499.000	
Missing	.000	

Bisa kita lihat di dalam kluster pertama terdapat jumlah data sebanyak 24 buah, sedangkan di dalam kluster kedua terdapat 19 buah data, di kluster keempat terdapat 454 buah data dan kluster keempat terdapat 2 buah data. Yang dimana terdapat jarak diantara beberapa kluster data sebagaimana yang tercantum di dalam tabel “Distances between Final Cluster Centers”, dan juga terdapat nilai tengah data pada masing masing kluster sebagaimana yang tercantum pada table “Final Cluster Centers”

Kemudian kita menggunakan metode Aplikasi WEKA yang dimana kurang lebih dengan metode yang sama seperti SPSS membagi 4 kelompok berdasarkan label tweet type. yang dimana kita mendapatkan hasilnya sebagai berikut :

Final cluster centroids:						
Attribute	Full Data (499.0)	Cluster# 0 (431.0)	1 (4.0)	2 (53.0)	3 (11.0)	
tweet_id	'All'	'All'	'All'	'All'		'All'
user_id	'(0-0.959547]'	'(0-0.959547]'	'(-inf-0]'	'(0.993576-inf)'	'(-inf-0]'	'(-inf-0]'
user_friends	'(0.000602-inf)'	'(0.000602-inf)'	'(0.000602-inf)'	'(-inf-0.000602]'	'(0.000602-inf)'	'(0.000602-inf)'
user_follower	'(-inf-0.017191]'	'(-inf-0.017191]'	'(0.017191-inf)'	'(-inf-0.017191]'	'(0.017191-inf)'	'(0.017191-inf)'
user_statuses count	'(0.000145-0.318573]'	'(0.000145-0.318573]'	'(0.000145-0.318573]'	'(0.000002-0.000046]'	'(0.318573-inf)'	'(0.318573-inf)'

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	431 (86%)
1	4 (1%)
2	53 (11%)
3	11 (2%)

Class attribute: tweet_type

Classes to Clusters:

0	1	2	3	<-- assigned to cluster
102	4	30	11	original
262	0	10	0	retweet
54	0	3	0	reply
13	0	10	0	quote

Cluster 0 <-- retweet

Cluster 1 <-- No class

Cluster 2 <-- original

Cluster 3 <-- No class

Incorrectly clustered instances : 207.0 41.483 %

Sebagaimana yang tercantum pada gambar, kita membagi 4 klaster data dan setiap klaster mempunyai rentang nilai yang unik tersendiri, yang dimana klaster 0 terdapat 100 buah tweet bertipe original, 262 buah tweet bertipe retweet, 54 buah tweet bertipe reply dan 13 buah tweet bertipe quote

Untuk klaster 1 di dalamnya hanya terdapat 4 buah tweet original, untuk klaster 2 di dalamnya terdapat 30 buah tweet bertipe original, 10 tweet bertipe retweet, 3 tweet bertipe reply , dan 10 tweet bertipe quote.

Sedangkan untuk klaster terakhir yaitu klaster 3 di dalamnya hanya mengandung 11 tweet bertipe original

BAB 4

PENUTUP

4.1 Kesimpulan

Dalam pengumpulan data digunakan web *Netlytic.org*, yang dimana mengambil data berupa cuitan atau tweet yang membahas tentang gempa. Selanjutnya dalam tahap preprocessing data menggunakan aplikasi WEKA dalam peroperasiannya. Dalam tahap regresi data menggunakan aplikasi SPSS mengambil metode regresi linier berganda. Untuk mengetahui apakah jumlah pertemanan dan jumlah status mempengaruhi jumlah follower. Dimana hasil dari tabel anova adalah 0, dimana nilainya kurang dari 0,05, yang berarti jumlah pertemanan dan jumlah status mempengaruhi jumlah follower. Tahap selanjutnya menggunakan aplikasi SPSS dan WEKA untuk mengelola K-Means yang dimana hasilnya terbagi dalam 3 kluster.

DAFTAR PUSTAKA

- Bryan Orleans, E. P. (2022, Januari 31). *Clustering Algoritma (K-Means)*. Diambil kembali dari Binus University: <https://sis.binus.ac.id/>
- By Sekolah Stata. (t.thn.). *Regresi adalah Metode Analisis Statistik, Manfaat, dan Rumus*. Diambil kembali dari By Sekolah Stata: <https://sekolahstata.com/>
- Data Scraping : Definisi, Cara Kerja dan 2 Tipe/Jenisnya*. (2022 , Januari 13). Diambil kembali dari IDCloudHost: <https://idcloudhost.com/>
- Haryanto, A. (2021, Februari 17). *Data Cleansing: Pengertian, Manfaat, Tahapan dan Caranya*. Diambil kembali dari Jojonomic: <https://www.jojonomic.com/>
- Konsultan Data Penelitian & ArcGIS. (2020, Agustus 19). *Ragam Jenis Transformasi Data yang Wajib Diketahui*. Diambil kembali dari patrastatistika: <https://patrastatistika.com/>
- Muchamad Taufiq Anwar, L. H. (2021). Model Prediksi Dropout Mahasiswa. *JURNAL INFORMATIKA UPGRIS*.
- Pentingnya Data Cleaning Dalam Data Science*. (2022, Februari 10). Diambil kembali dari algoritman: <https://algorit.ma/>